Theory and Application of Path Probability Functionals in Population Genetics

By

Joshua G. Schraiber

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Integrative Biology

and the Designated Emphasis

in

Computational and Genomic Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Montgomery Slatkin, Chair
Professor Rasmus Nielsen
Professor Rachel Brem

Fall 2014

Theory and Application of Path Probability Functionals in Population Genetics

Abstract

Theory and Application of Path Probability Functionals in Population Genetics

by

Joshua G. Schraiber

Doctor of Philosophy in Integrative Biology

and the Designated Emphasis in Computational and Genomic Biology

University of California, Berkeley

Professor Montgomery Slatkin, Chair

Calculations based on the Wright-Fisher process and its limits are the primary tools of population genetics theory. Decades of theoretical work have elucidated much about the properties of the neutral Wright-Fisher model, in which different mutations have the exact same Darwinian fitness. The model becomes significantly more complicated when the action of natural selection is taken into account, and we are only just beginning to understand the details of evolution subject to natural selection. In this thesis, I take a path integral approach to understanding the Wright-Fisher diffusion with selection, in contrast to the typical approach, using partial differential equations. This allows me to use powerful machinery from quantum physics and mathematical finance to come up with novel ways to perform difficult calculations. The work here is composed of three parts. First, I develop a rejection sampling approach to obtaining Wright-Fisher diffusion paths when the allele frequency trajectory is conditioned to start and end at certain points (such paths are called bridges). Next, I use perturbation theory to calculate the transition densities of the Wright-Fisher process with genic selection, assuming weak selection. Finally, I implemented a Markov chain Monte Carlo approach to estimating selection coefficients from allele frequency time series that makes use of aspects of both of the previous chapters.

To Crystal Perreira, without whom I wouldn't have made it this far.

Contents

List of Figures

List of Tables

Acknowledgments

I owe a tremendous debt to all the awesome people who've helped me get to this point in my life. Might as well start at the beginning! Despite being separated since before I can remember, my mom and dad were instrumental in making sure I grew up to be a pretty okay person who works hard. And this Ph.D. was definitely hard work! I can't forget Mr. Craig Fox, a high school teacher who sparked my interest in science and turned from a straight-C student into someone who actually cared about learning. Despite my best web sleuthing skills I've never been able to find out what happened to him, so if you're reading this, Mr. Fox, send me an e-mail!

Without the generosity of Doug Cook in actually *paying me* to work in his lab with the amazing Brendan Riely, I doubt I would have gotten the awesome research experience I did as an undergrad. While I didn't end up pipetting full-time (or really at all) in grad school, that exposure to the process of science was invaluable. Carole Hom, Rick Grosberg and Sebastian Schreiber and everyone I worked with in CLIMB really helped me build my skills as a mathematical biologist (the first programming I ever did was in CLIMB!) and showed me just how damn hard it is to come up with your own research questions. And of course, without the encouragement, knowledge and general awesomeness of Michael Turelli, I'd probably have never realized what an awesome field population genetics really is.

I've met so many cool people who have helped me along the way in grad school it's impossible to name you all without forgetting someone, but I really want to thank my awesome current and former officemates, Mel Yang, Ben Peter, Fernando Racimo, Kelley Harris, Mason Liang, Eric Durand, Anna-Sapfo Malasinas, Flora Jay, Nick Matzke, Vera Kaiser, all the awesome visitors we've had, and probably some people I've forgotten. Everyone else in IB, CTEG, Comp Bio or if I just know you randomly: you all deserve to be in here, too, but I'm lazy.

Finally, I owe a tremendous debt to my committee, Monty Slatkin, Rasmus Nielsen and Rachel Brem, along with "shadow committee" member Steve Evans. I've gotten the chance to work on some amazing stuff because of them and I'll always remember my time here at Berkeley. And while I didn't end up including any of the work I did with Rachel in my thesis, she can rest assured that she shaped the direction I'm going with my postdoc so it's not a total loss!

This dissertation was typeset using the ucastrothesis LaTeX template.

Chapter 1

Introduction

From its origins as a primarily theoretical discipline by the hands of R.A. Fisher, Sewall Wright and J.B.S. Haldane, population genetics has grown into an extensive field encompassing both applied and theoretical aspects. For the past several decades, the primary lens through which both theoretical and empirical observations were made was the neutral theory (Kimura 1984). From a theoretical point-of-view, the neutral theory provides simple calculations and clear interpretations. The Wright-Fisher Markov chain, and its infinite-population limits, serve as primary tools in the theoretician's toolbox (Ewens 2004). For neutral alleles, several results can be obtained, including the stationary distributions (Wright 1931a), the complete transition density (Kimura 1955a), and certain sampling distributions (Ewens 1972).

Moreover, in the neutral case, it is possible to consider a simple genealogical process, *the coalescent*, introduced by Kingman (1982). In this model, first the genealogy of the individuals in a sample is constructed, and then neutral mutations are laid on top of that genealogy. The probability of the genetic data obtained can then be computed and used to fit complicated models of demographic history and population structure (Kuhner et al. 1995, 1998; Drummond et al. 2005; Li & Durbin 2011; Harris & Nielsen 2013).

Because it is easy to determine properties of the data under the neutral model, it can be used as a null model for interpreting empirical genetic data (Watterson 1978; Tajima 1989; Slatkin 1994; Fay & Wu 2000). In many cases, the goal of such *neutrality tests* is to determine whether the observed genetic data have been shaped by natural selection. A wide array of tests have been developed to detect the signature of selection in linked neutral variation and applied successfully across numerous species (Nielsen et al. 2005a; Voight et al. 2006; Pickrell et al. 2009). However, going beyond merely rejecting neutrality to estimating the strength and direction of natural selection has proved difficult.

Part of this difficulty stems from the added complexity entailed by natural selection. Unlike the neutral case, there is no simple solution for the transition probabilities of the diffusion limit of the Wright-Fisher Markov chain with selection. Kimura (1955b) made an early attempt using a separation of variables approach; however, he was unable to come up with simple expressions for the eigenvalues of the associated differential operator. More

recently, the work of Song & Steinrücken (2012) (see also Steinrücken et al. (2012)) developed a computational approach to approximating the transition density, relying on the theory of orthogonal functions. When Neuhauser & Krone (1997) (see also Krone & Neuhauser (1997)) introduced a genealogical approach—known as the ancestral selection graph—to modeling natural selection, many hoped that it would be a golden age for inferring natural selection from population genomic data. Unfortunately, the same difficulties that plagued earlier approaches to modeling natural selection have ultimately shown that the ancestral selection graph was not as powerful as initially hoped.

In this thesis, I present a different way of thinking about the Wright-Fisher diffusion. As opposed to the partial differential equations (PDE) approach that is traditionally used (Ewens 2004), I use an approach inspired by the path integral formulation of quantum mechanics, first laid out in Feynman (1948) (but note the influence of Dirac (1933)). In the setting of quantum physics, each possible path a particle can take between points A and B is assigned a phase in the complex plane, and by integrating over the (uncountably infinite) space of paths, one can calculate quantities of interest. As recognized by Kac (1949), this approach could be extended to diffusion processes by using the theory of Wiener (1921) related to integrals over the space of Brownian motion. In the context of allele frequency change, we consider assigning a probability to every possible path that an allele can take in going from frequency x to frequency y.

While this probability is well defined for the *discrete-time* Wright-Fisher model, it's necessary to be very precise when considering the infinite-population limit. Whereas in quantum systems, transitioning between the PDE viewpoint and the path integral viewpoint corresponds to switching between a Hamiltonian and Lagrangian viewpoint, it is not so simple. In particular one cannot simply compute the Onsager-Machlup functional (the equivalent of a Lagrangian) corresponding to the diffusion generator (the equivalent of a Hamiltonian). This is because many diffusion have a space-dependent diffusion coefficient; thus, they are equivalent to models on curved spaces (Graham 1977; Dürr & Bach 1978).

To overcome this difficulty, it is necessary to use ideas from stochastic analysis to come up with appropriate *path probability functionals* (such objects are functionals because they are maps from a functional space to $\mathbb{R}$). The key idea, explored throughout this thesis, is to use Girsanov's theorem (Girsanov 1960) to compute path densities relative to an appropriate dominating measure. Because there are many such choices of possible dominating measures, different choices are suitable for different applications.

In Chapter 2, I explore the utility of path probability functionals for rejection sampling Wright-Fisher diffusion bridges. This work builds off of the foundation laid by Beskos & Roberts (2005) and was done in collaboration with Robert C. Griffiths and Steven N. Evans. I developed an efficient rejection sampler, implemented in R, and used to understand certain path properties of the Wright-Fisher process conditioned on hitting a specific frequency at a specific time.

In Chapter 3, I attempt to come up with a simple, analytic approach to compute Wright-Fisher diffusion transition densities when natural selection is incorporated. I do this by taking the perturbative approach outlined in Feynman & Hibbs (2012), taking a neutral

Wright-Fisher diffusion to be a "free propagator" and treating natural selection as a kind of potential energy.

Finally, in Chapeter 4, I combine these two directions into a Markov chain Monte Carlo method for estimating natural selection from allele frequency time series. In this work, joint with Montgomery Slatkin and Steven N. Evans, I implemented C++ software to determine the posterior distribution of the selection intensity, dominance coefficient and allele age. I applied this method to empirical data and found striking signals of selection.

Chapter 2

Rejection Sampling Wright-Fisher Bridges

2.1 Introduction

The Wright-Fisher Markov chain is of central importance in population genetics and has contributed greatly to the understanding of the patterns of genetic variation seen in natural populations. Much recent work has focused on developing sampling theory for neutral sites linked to sites under selection (Smith & Haigh 1974; Kaplan et al. 1989; Nielsen et al. 2005b; Etheridge et al. 2006). Typically, the site under selection is assumed to have dynamics governed by the diffusion process limit of the Wright-Fisher chain, in which case the genealogy of linked neutral sites can be constructed using the framework of Hudson & Kaplan (1988). However, due to the complicated nature of this model, analytical theory is necessarily approximate and the main focus is on simulation methods. In particular, a number of simulation programs, including mbs (Teshima & Innan 2009) and msms (Ewing & Hermisson 2010) have recently appeared to help facilitate the simulation of neutral genealogies linked to sites undergoing a Wright-Fisher diffusion with selection.

Simulations of Wright-Fisher paths under selection can be easily carried out using standard techniques for simulating diffusions. Frequently, however, it is necessary to simulate a Wright-Fisher path conditioned on some particular outcome. For example, to simulate the path of an allele under selection that is currently at frequency x, a time-reversal argument shows that it is possible to simulate a path starting at x conditioned to hit 0 eventually (Maruyama 1974). However, more complicated scenarios, including the action of natural selection on standing genetic variation, require more elaborate simulation methods (Peter et al. 2012).

The stochastic process describing an allele that starts at frequency x at time 0 and is conditioned to end at frequency y at time T is called a bridge between x and y in time T or a bridge between x and y over the time interval $[0, T]$. Wright-Fisher diffusion bridges appear naturally in the study of selection acting on standing variation because it is necessary to know the path taken by an allele at current frequency y that fell under the influence of

natural selection at a time T generations in the past when it was segregating neutrally at frequency x. Wright-Fisher diffusion bridges are also of interest for their application to inference of selection from allele frequency time series (Bollback et al. 2008a; Malaspinas et al. 2012; Mathieson & McVean 2013; Feder et al. 2013). In particular, analysis of bridges can help determine the extent to which more signal is gained by adding further intermediate time points.

In addition to their applied interest, there are interesting theoretical questions surrounding Wright-Fisher diffusion bridges. For alleles conditioned to eventually fix, Maruyama (1974) showed that the distribution of the trajectory does not depend on the sign of the selection coefficient; that is, both positively and negatively selected alleles with the same absolute value of the selection coefficient exhibit the same dynamics conditioned on eventual fixation. It is natural to inquire whether the analogous result holds for a bridge between any two interior points. Moreover, the degree to which a Wright-Fisher bridge with selection will differ from a Wright-Fisher bridge under neutrality is not known (in connection with this question, we recall the well-known fact that the distribution of a bridge for a Brownian motion with drift does not depend on the drift parameter, and so it is conceivable that the presence of selection has little or no effect on the behavior of Wright-Fisher bridges). Lastly, the characteristics of the sample paths of the frequency of alleles destined to be lost in a fixed amount of time are not only interesting theoretically but may also have applications to geographically structured populations (Slatkin & Excoffier 2012).

Here we investigate various features of Wright-Fisher diffusion bridges. The paper is structured as follows. First, we establish analytical results for neutral Wright-Fisher bridges. Then, we derive a novel rejection sampler for Wright-Fisher bridges with selection and use it to study the properties of such processes. For example, we estimate the distribution of the maximum of a bridge from 0 to 0 under selection and investigate how this distribution depends on the strength of selection.

2.2 Background

A Wright-Fisher diffusion with genic selection is a diffusion process $\{X_t,\, t \geq 0\}$ with state space $[0, 1]$ and infinitesimal generator

$$\mathcal{L} = \gamma x(1-x)\frac{\partial}{\partial x} + \frac{1}{2}x(1-x)\frac{\partial^2}{\partial x^2}. \tag{2.1}$$

When $\gamma = 0$, the diffusion is said to be neutral; otherwise, the drift term captures the strength and direction of natural selection.

The corresponding Wright-Fisher diffusion bridge, $\{X_t^{x,z,[0,T]},\, 0 \leq t \leq T\}$ is the stochastic process that results from conditioning the Wright-Fisher diffusion to start with value x at time 0 and end with value z at time T. Denote by $f(x, y; t)$ the transition density of the diffusion corresponding to (2.1). By the Markov property of the Wright-Fisher diffusion, the bridge is a time-inhomogeneous diffusion and the transition density for the bridge going

from state u at time s to state v at time t is

$$f_{x,z,[0,T]}(u,v;s,t) = \frac{f(u,v;t-s)f(v,z;T-t)}{f(u,z;T-s)}. \tag{2.2}$$

The time-inhomogeneous infinitesimal generator of the bridge acting on a test function g at time s is

$$\begin{aligned}\mathcal{L}_{x,z,[0,T];s}g(u) &= \lim_{t\downarrow s}\frac{\mathbb{E}[g(X_t)\,|\,X_0=x, X_s=u, X_T=z]-g(u)}{t-s}\\ &= u(1-u)\left(\gamma+\frac{\partial}{\partial u}\log f(u,z;T-s)\right)\frac{\partial g}{\partial u}(u)\\ &\quad+\frac{1}{2}u(1-u)\frac{\partial^2 g}{\partial u^2}(u).\end{aligned} \tag{2.3}$$

An obvious method for simulating a Wright-Fisher bridge would be to simulate the stochastic differential equation (SDE) corresponding to this infinitesimal generator. There are two obstacles to this approach. Firstly, analytic expressions for the transition density f are only known for the neutral case, and even there they are in the form of infinite series. Secondly, note that the first order coefficient in the infinitesimal generator becomes increasing singular as $s \uparrow T$; consequently, an attempt to simulate the bridge by simulating the SDE would be quite unstable because the drift term in the SDE would explode at times close to the terminal time T. It is because this naive approach is infeasible that we need to consider the more sophisticated simulation methods explored in this paper.

In addition to conditioning the process to obtain a particular value at a particular time, it is possible to condition a process's long term behavior. The transition densities of the conditioned process, $f_h(x,y;t)$ are related to to the transition densities of the unconditioned process by the usual Doob h-transform formula,

$$f_h(x,y;t) := h(x)^{-1}f(x,y;t)h(y).$$

The h-transformed process has infinitesimal generator

$$\mathcal{L}^h := x(1-x)\left(\gamma+\frac{h'(x)}{h(x)}\right)\frac{\partial}{\partial x}+\frac{x(1-x)}{2}\frac{\partial^2}{\partial x^2}. \tag{2.4}$$

Note that the finite dimensional marginal distribution at times $0 \le t_1 \le \ldots \le t_n \le T$ of the Wright-Fisher diffusion bridge starting at x at time 0 and ending at y at time T has density

$$\frac{f(x,v_1;t_1)f(v_1,v_2;t_2-t_2)\cdots f(v_n,y;T-t_n)}{f(x,y;T)}$$

whereas the analogous density for the corresponding bridge of the h-transformed process is

$$\begin{aligned}&\frac{h(x)^{-1}f(x,v_1;t_1)h(v_1)h(v_1)^{-1}f(v_1,v_2;t_2-t_1)h(v_2)\cdots h(v_n)^{-1}f(v_n,y;T-t_n)h(y)}{h(x)^{-1}f(x,y;T)h(y)}\\ &= \frac{f(x,v_1;t_1)f(v_1,v_2;t_2-t_1)\cdots f(v_n,y;T-t_n)}{f(x,y;T)}.\end{aligned}$$

Thus, the the bridges for the two processes have the same distribution.

Typical h-transforms include the conditioning a process to eventually hit a particular value, and for the sake of future reference we recall from standard diffusion theory (Rogers & Williams 2000a) that the probability that the Wright-Fisher diffusion started from x eventually hits y is

$$p_{xy} = \begin{cases} \frac{S(x)-S(0)}{S(y)-S(0)}, & \text{if } y > x, \\ \frac{S(1)-S(y)}{S(1)-S(x)}, & \text{if } y < x, \end{cases} \tag{2.5}$$

where S is the scale function given by

$$S(x) = \begin{cases} \frac{1-e^{-2\gamma x}}{1-e^{-2\gamma}}, & \text{if } \gamma \neq 0, \\ x, & \text{if } \gamma = 0. \end{cases}$$

Thus,

$$p_{xy} = \begin{cases} \frac{1-e^{-2\gamma x}}{1-e^{-2\gamma y}}, & \text{if } y > x, \\ \frac{e^{-2\gamma y}-e^{-2\gamma}}{e^{-2\gamma x}-e^{-2\gamma}}, & \text{if } y < x, \end{cases} \tag{2.6}$$

when $\gamma \neq 0$ and

$$p_{xy} = \begin{cases} \frac{x}{y}, & \text{if } y > x, \\ \frac{1-y}{1-x}, & \text{if } y < x. \end{cases} \tag{2.7}$$

2.3 Rejection sampling Wright-Fisher bridge paths

2.3.1 General framework

When selection is incorporated into the Wright-Fisher model, there is no known series formula for the transition density (but see Kimura (1955b) and Kimura (1957) for attempts using perturbation theory, as well as Song & Steinrücken (2012) and Steinrücken et al. (2012) for methods of approximating an eigenfunction expansion computationally). Therefore, we develop a rejection sampling method that can sample paths of Wright-Fisher diffusion bridges with genic selection efficiently for the purpose of investigating their properties. In this work, we focus on a diffusion with genic selection, instead of general diploid selection, for analytical convenience. The following approach would apply even in the more general case.

Before we explain how rejection sampling can be used to sample paths of a Wright-Fisher bridge, we first describe the analogous, but simpler, method for sampling paths of diffusion bridges that have distributions which are absolutely continuous with respect to that of a Brownian bridge. Fix $x, z \in \mathbb{R}$ and $T > 0$. Let $\mathbb{W}$ be the distribution of Brownian bridge from x to z over the time interval $[0, T]$, and let $\mathbb{P}$ be the distribution of the path of a bridge from x to z over the time interval $[0, T]$ for a diffusion with infinitesimal generator

$$\mathcal{G} = a(x)\frac{\partial}{\partial x} + \frac{1}{2}\frac{\partial^2}{\partial x^2}. \tag{2.8}$$

It follows from Girsanov's theorem (see, for example, Rogers & Williams (2000a)) that the probability measure $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{W}$ with Radon-Nikodym derivative (that is, density)

$$\frac{d\mathbb{P}}{d\mathbb{W}}(\omega) = \exp\left\{\int_0^T a(\omega_t)\, d\omega_t - \frac{1}{2}\int_0^T a^2(\omega_t)\, dt\right\} \tag{2.9}$$

for the path ω, where the first integral in (2.9) is an Itô integral – see Beskos & Roberts (2005) for the details of the disintegration argument that concludes this fact about Radon-Nikodym derivatives with respect to the Brownian bridge distribution from the usual statement of Girsanov's theorem, which is about Radon-Nikodym derivatives with respect to the distribution of Brownian motion. Because a Brownian bridge can be constructed using a simple transformation of a Brownian motion (namely, if B is a standard Brownian motion, then the process $\{x + (B_t - \frac{t}{T}B_T) + \frac{t}{T}(z - x),\ 0 \le t \le T\}$ has the distribution $\mathbb{W}$), it is computationally feasible to obtain fine-grained samples of the Brownian bridge. Once we have a sequence of Brownian bridge paths, (2.9) can be used to compute a likelihood ratio, and a standard rejection sampling scheme can then be utilized to obtain realizations of diffusion bridge paths; see Beskos & Roberts (2005) for examples of extensions to this approach.

2.3.2 Application to Wright-Fisher process

The above method is not immediately applicable to the Wright-Fisher bridge because its infinitesimal generator is not of the form (2.8). However, it was shown on pp 119-120 of Wright (1931b) that if X is the Wright-Fisher process with infinitesimal generator (2.1), then the transformation

$$Y_t := \arccos(1 - 2X_t) \tag{2.10}$$

suggested in Fisher (1922a) produces a diffusion process Y on the state space $[0, \pi]$ with infinitesimal generator

$$\mathcal{L}_Y = \frac{1}{2}(\gamma \sin(y) - \cot(y))\frac{\partial}{\partial y} + \frac{1}{2}\frac{\partial^2}{\partial y^2}.$$

Because Y has absorbing boundaries at 0 and π, sampling paths of bridges for Y by sampling Brownian bridges can involve extremely high rejection rates. More specifically,

$$\frac{1}{2}(\gamma \sin(y) - \cot(y)) \approx -\frac{1}{2y}, \quad \text{as } y \downarrow 0,$$

and so the likelihood ratio (2.9) becomes extremely small for paths that spend a significant amount of time near 0. A similar phenomenon occurs near π.

To overcome the difficulty near 0, we develop a rejection sampling scheme where the proposals are realizations of a process other than the Brownian bridge.

As a first step, consider the Wright-Fisher diffusion conditioned to be eventually absorbed at 1. By the argument given in Section 2, this process has the same bridges as the

unconditional process. It follows from (2.6) and (2.7) with $y = 1$ that the probability the process starting from x is absorbed at 1 is

$$h(x) := \begin{cases} \frac{1-e^{-2\gamma x}}{1-e^{-2\gamma}}, & \gamma \neq 0, \\ x, & \gamma = 0. \end{cases}$$

The transition densities of the conditioned process, $f_h(x, y; t)$, are related to the unconditional transition densities by the usual Doob h-transform formula

$$f_h(x, y; t) := h(x)^{-1} f(x, y; t) h(y).$$

The corresponding infinitesimal generator is

$$\mathcal{L}^h := \begin{cases} \gamma x(1-x)\coth(\gamma x)\frac{\partial}{\partial x} + \frac{1}{2}x(1-x)\frac{\partial^2}{\partial x^2}, & \gamma \neq 0, \\ (1-x)\frac{\partial}{\partial x} + \frac{1}{2}x(1-x)\frac{\partial^2}{\partial x^2}, & \gamma = 0. \end{cases} \tag{2.11}$$

Applying the transformation (2.10) to the process with infinitesimal generator (2.11) results in a process with infinitesimal generator

$$\mathcal{L}^h_Y := \begin{cases} \frac{1}{2}\left(\gamma \sin(y)\coth(\gamma \sin^2(y/2)) - \cot(y)\right)\frac{\partial}{\partial y} + \frac{1}{2}\frac{\partial^2}{\partial y^2}, & \gamma \neq 0, \\ \frac{1}{2}\left(\sin(y)\csc^2(y/2) - \cot(y)\right)\frac{\partial}{\partial y} + \frac{1}{2}\frac{\partial^2}{\partial y^2}, & \gamma = 0. \end{cases} \tag{2.12}$$

Note that

$$\frac{1}{2}\left(\gamma \sin(y)\coth(\gamma \sin^2(y/2)) - \cot(y)\right) \approx \frac{3}{2y} \quad \text{as } y \downarrow 0 \tag{2.13}$$

and

$$\frac{1}{2}\left(\sin(y)\csc^2(y/2) - \cot(y)\right) \approx \frac{3}{2y} \quad \text{as } y \downarrow 0. \tag{2.14}$$

Moreover, if $\mathbb{Q}$ is the distribution of a bridge from x to z over the time interval $[0, T]$ for some diffusion with infinitesimal generator

$$\mathcal{G} = b(x)\frac{\partial}{\partial x} + \frac{1}{2}\frac{\partial^2}{\partial x^2}$$

and $\mathbb{P}$ is the distribution of a bridge from x to z over the time interval $[0, T]$ for the diffusion with infinitesimal generator (2.8), then

$$\begin{aligned} \frac{d\mathbb{P}}{d\mathbb{Q}}(\omega) &= \frac{d\mathbb{P}}{d\mathbb{W}}(\omega)\frac{d\mathbb{W}}{d\mathbb{Q}}(\omega) \\ &= \frac{d\mathbb{P}}{d\mathbb{W}}(\omega) \Big/ \frac{d\mathbb{Q}}{d\mathbb{W}}(\omega) \\ &= \exp\left\{ \int_0^T a(\omega_t) - b(\omega_t)\, d\omega_t - \frac{1}{2}\int_0^T a^2(\omega_t) - b^2(\omega_t)\, dt \right\}. \end{aligned}$$

This suggests that a better rejection sampling scheme for bridges of the process Y with end points close to zero will result when the proposals come from a diffusion with an infinitesimal generator having a first order coefficient with a singularity at zero matching the one appearing in both (2.13) and (2.14).

For such a modified scheme to be feasible, it is necessary to work with a proposal diffusion for which it is easy to simulate the associated bridges. We now introduce such a process. The 4-dimensional Bessel process is the radial part of a 4-dimensional Brownian motion. That is, if $\{B_t = (B_t^{(i)})_{i=1}^4,\, t \geq 0\}$ is a vector of 4 independent one-dimensional Brownian motions, then

$$\beta_t := |B_t| = \sqrt{(B_t^{(1)})^2 + (B_t^{(2)})^2 + (B_t^{(3)})^2 + (B_t^{(4)})^2}, \quad t \geq 0,$$

is a 4-dimensional Bessel process (see Revuz & Yor (1999, Section XI.1) for a thorough discussion of Bessel processes). The 4-dimensional Bessel process is a diffusion with infinitesimal generator

$$\mathcal{B} := \frac{3}{2}\frac{1}{x}\frac{\partial}{\partial x} + \frac{1}{2}\frac{\partial^2}{\partial x^2}.$$

Letting $\mathbb{P}$ (resp. $\mathbb{B}$) be the distribution of the bridge for the process with infinitesimal generator (2.12), and hence the distribution of the transformed Wright-Fisher diffusion Y, (resp. the 4-dimensional Bessel bridge) from x to z over the time interval $[0, T]$, we have

$$\begin{aligned}
\frac{d\mathbb{P}}{d\mathbb{B}}(\omega) &= \frac{d\mathbb{P}}{d\mathbb{W}}(\omega)\frac{d\mathbb{W}}{d\mathbb{B}}(\omega) \\
&= \exp\left\{\int_0^T \frac{1}{2}\left(\gamma\sin(\omega_t)\coth(\alpha\sin^2(\omega_t/2)) - \cot(\omega_t) - \frac{3}{\omega_t}\right) d\omega_t \right. \\
&\quad \left. - \frac{1}{2}\int_0^T \frac{1}{4}\left(\left(\gamma\sin(\omega_t)\coth(\alpha\sin^2(\omega_t/2)) - \cot(\omega_t)\right)^2 - \frac{9}{\omega_t^2}\right) dt\right\}. \quad (2.15)
\end{aligned}$$

We next explain how to simulate a 4-dimensional Bessel bridge. We can construct the bridge from $u \in \mathbb{R}^4$ to $v \in \mathbb{R}^4$ over the time interval $[0, T]$ for the 4-dimensional Brownian motion as

$$C_t := \left(1 - \frac{t}{T}\right) u + \frac{t}{T} v + \left(B_t - \frac{t}{T} B_T\right),$$

where $B_0 = 0$. The distribution of $u + B_T$ conditional on $|u + B_T| = z$ has density proportional to $w \mapsto \exp(w \cdot u/T)$ with respect to the normalized surface measure on the sphere centered at the origin with radius y, where $w \cdot u$ is the usual scalar product of the two vectors $w, u \in \mathbb{R}^4$. Hence, a 4-dimensional Bessel bridge from x to z over the time interval $[0, T]$ is given by

$$\gamma_t := \left|\left(1 - \frac{t}{T}\right) u + \frac{t}{T} V + \left(B_t - \frac{t}{T} B_T\right)\right|,$$

where $B_0 = 0$, $u \in \mathbb{R}^4$ is any vector with $|u| = x$, and V is random vector taking values on the sphere centered at the origin with radius z that is independent of B and has a density with

respect to the normalized surface measure that is proportional to $w \mapsto \exp(w \cdot u/T)$. Note that the random vector V/z, which takes values on the unit sphere centered at the origin, has a Fisher – von Mises distribution with mean vector u/x and concentration parameter xz/T (see, for example, Mardia et al. (1979, Ch. 15)).

Increasing the strength of natural selection causes the Wright-Fisher bridge to move faster for intermediate frequencies, but the method proposed above uses the same 4-dimensional Bessel bridge regardless of the value of the selection parameter γ, and so the rejection rate can become very high for large values of γ. To deal with this phenomenon, we introduce the following further refinement to the proposal process.

With $\mathbb{P}$ the distribution of the transformed Wright-Fisher bridge from x to z over the time interval $[0, T]$ as above, let $\omega^\epsilon : [0, T] \to [0, \pi]$, $\epsilon > 0$, be the path with $\omega_0^\epsilon = x$ and $\omega_T^\epsilon = z$ that maximizes

$$\omega \mapsto \mathbb{P}\left\{\omega' : \sup_{0 \le t \le T} |\omega'_t - \omega_t| \le \epsilon\right\}.$$

Then, ω^ϵ converges as $\epsilon \downarrow 0$ to a path ω^*. Heuristically, we can think of ω^* as the path that has "maximum probability" or is "modal" for $\mathbb{P}$. This path is sometimes called an Onsager-Machlup function and it can be found by solving a certain variational problem – see, for example, Ikeda & Watanabe (1989). For the transformed Wright-Fisher bridge, an analysis of the variational problem shows that the maximum probability path satisfies the second order ordinary differential equation

$$\ddot{\omega}^* = \frac{\gamma^2}{8} \sin \omega^* - \frac{3}{4} \cot \omega^* \csc^2 \omega^* \tag{2.16}$$

with boundary conditions $\omega_0^* = x$ and $\omega_T^* = z$.

With a solution to (2.16) in hand, it is possible to construct a better proposal distribution by linking together bridges that are "close" to the maximum probability path. First, choose a number of discretization points N and take times $0 < t_1 < \ldots < t_N < T$. Then, sample independent random variables $U_1, U_2, \ldots, U_N$ with densities $g_1, g_2, \ldots, g_N$ to be specified later. Put $t_0 = 0$, $t_{N+1} = T$, $U_0 = x$ and $U_{N+1} = z$. Build conditionally independent 4-dimensional Bessel bridges from U_i to U_{i+1} over the time intervals $[t_i, t_{i+1}]$. The distribution of U_i should be chosen so that U_i is close to the maximum probability path at time t_i; we choose re-scaled Beta distributions with mode at the solution of (2.16) at time t_i. More specifically, we set $U_i = \pi X_i$, where X_i has the Beta distribution with parameters

$$\left(\frac{1 + \frac{x^*_{t_i}}{\pi}(\theta - 2)}{1 - \frac{x^*_{t_i}}{\pi}}, \theta\right).$$

for some free parameter θ. We used the particular value $\theta = 50$ for the examples in this paper, but other value of θ could be used in a given situation in an attempt to optimize the frequency of rejection.

By stringing these bridges together, we get a path going from x to z over the time interval $[0, T]$. However, the distribution of this path is certainly not that of the 4-dimensional Bessel

bridge because of the manner in which we have chosen the endpoints of the component bridges. Therefore, we can't simply use the Radon-Nikodym derivative (2.15) as it stands to construct a rejection sampling procedure. Rather, if we let $\mathbb{Q}$ be the distribution of the path built by stringing the bridges together, then we must accept a path ω with probability proportional to

$$\frac{d\mathbb{P}}{d\mathbb{B}}(\omega)\frac{d\mathbb{B}}{d\mathbb{Q}}(\omega). \tag{2.17}$$

Note that

$$\frac{d\mathbb{B}}{d\mathbb{Q}}(\omega) = \frac{\prod_{i=0}^{N}\rho(\omega_{t_i},\omega_{t_{i+1}};t_{i+1}-t_i)}{\rho(x,z;T)\prod_{i=1}^{N}g_i(\omega_{t_i})}, \tag{2.18}$$

where

$$\rho(x,z;t) := I_1\left(\frac{xy}{t}\right)\frac{y^2}{xt}e^{-\frac{x^2+z^2}{2t}} \tag{2.19}$$

is the transition density of the 4-dimensional Bessel process with I_ν the modified Bessel function of the first kind.

2.3.3 Simulation results

To demonstrate the effectiveness of the rejection sampling scheme, Figure 2.1 shows Q-Q plots of the one-dimensional marginal at time t of a Wright-Fisher bridge with genic selection as estimated using the rejection sampler compared to an approximation that uses the method of Song & Steinrücken (2012) to compute the cumulative distribution function of the marginal. For both rows, the bridge goes from $x = .2$ to $z = 0.7$ over the time interval $[0, T] = [0, 0.1]$. The left panels correspond to $t = 0.03$ and the right panels correspond to $t = 0.07$. The top row corresponds to $\gamma = 10$ and the bottom row to $\gamma = 50$, demonstrating the effectiveness of the rejection sampling scheme over a wide range of selection coefficients.

Figure 2.2 demonstrates the behavior of a Wright-Fisher diffusion bridge as the selection coefficient increases. A bridge from $x = 0.01$ to $z = 0.8$ over the time interval $[0, T] = [0, 0.1]$ is shown for $\gamma = 0$, $\gamma = 50$ and $\gamma = 100$. As the selection coefficient increases, the proportion of time the bridge spends near the boundary also increases, because the Wright-Fisher diffusion moves faster when it is away from the boundaries. In addition, the paths that the bridge takes become more tightly centered around the most probable path as the selection coefficient increases.

Being able to sample Wright-Fisher bridge paths makes it very easy to numerically approximate the distribution and expectation of various functionals of the path. As an example, Figure 2.3 shows the density of the maximum in a bridge from $x = 0$ to $z = 0$ over the time interval $[0, T] = [0, 0.1]$ for $\gamma = 0$, $\gamma = 50$ and $\gamma = 100$. Note that the maximum in the bridge decreases as the strength of selection increases, and also becomes more tightly concentrated around its expectation.

To gain a more quantitative understanding of the extent to which a bridge for an allele experiencing natural selection looks different from the bridge for a neutral allele, it is possible to compute the Radon-Nikodym derivative (i.e. the likelihood ratio) of the distribution under

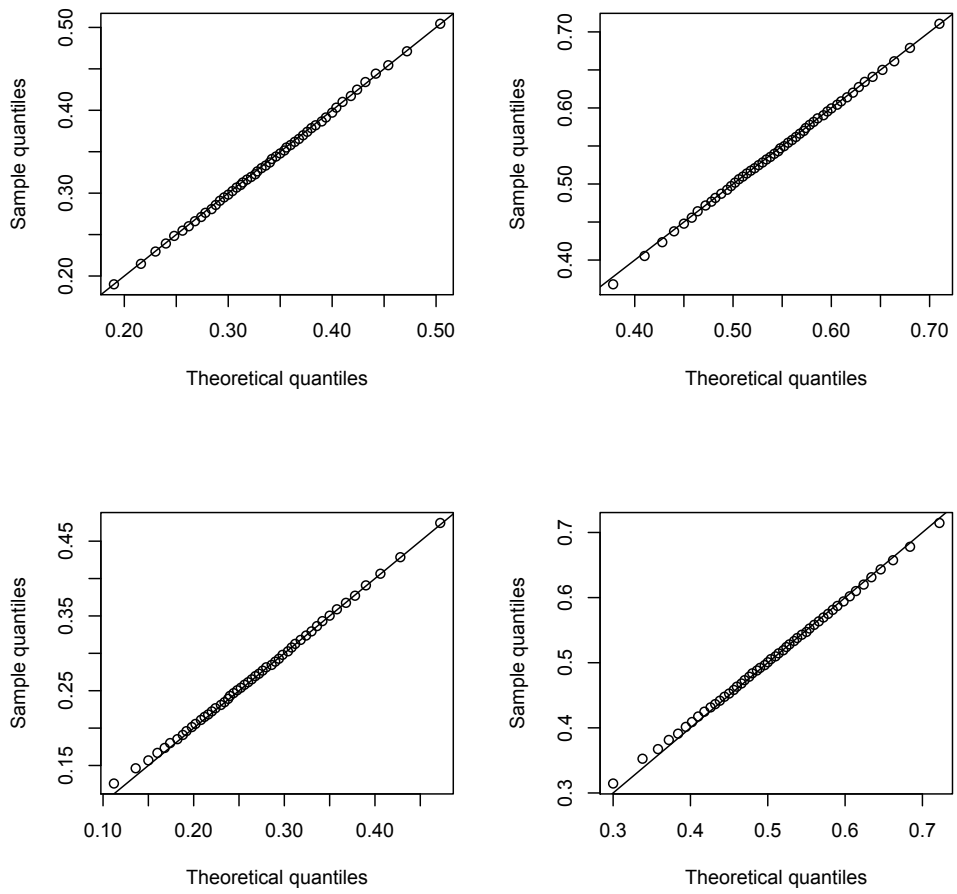

Figure 2.1: Q-Q plot showing the accuracy of the rejection sampling scheme. Theoretical quantiles were calculated using the method of Song & Steinrücken (2012) and sample quantiles are determined from 1000 bridges simulated using the method described in the text. The bridge goes from $x = 0.2$ to $z = 0.7$ over the time interval $[0, T] = [0, 0.1]$. The left panels correspond to $t = 0.03$ and the right panels correspond to $t = 0.07$. The top row corresponds to $\gamma = 10$ and the bottom row to $\gamma = 50$.

selection against the distribution under neutrality. Using an argument similar to that which led to (2.15), the likelihood ratio is

$$\frac{d\mathbb{P}_\gamma}{d\mathbb{P}_0}(\omega) \propto \exp\left\{-\frac{1}{8}\int_0^T \gamma^2 \sin^2(\omega_t)\, dt\right\}, \tag{2.20}$$

where the constant of proportionality only depends on the endpoints. A few things are immediately evident from (2.20). First of all, the likelihood ratio does not depend on the sign of the selection coefficient, only the magnitude. This is analogous to the result Maruyama

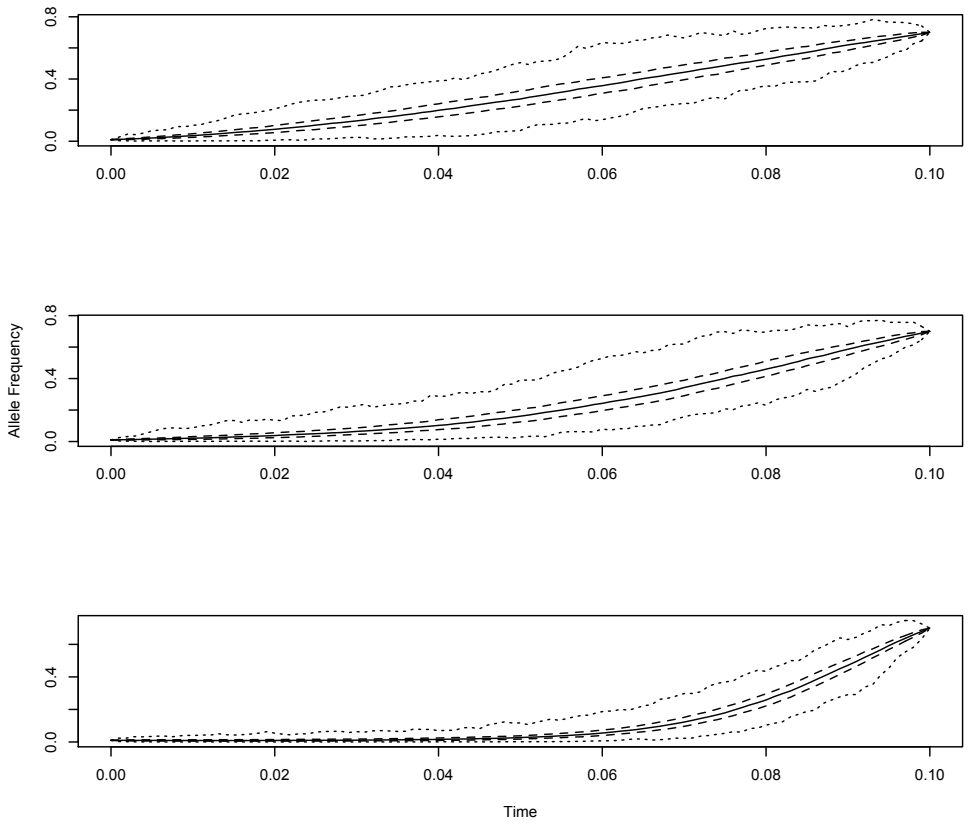

Figure 2.2: Plot showing the properties of bridge paths as the strength of selection increases. Each bridge is from $x = 0.01$ to $z = 0.8$ over the time interval $[0, T] = [0, 0.1]$. The successive selection coefficients are $\gamma = 0$, $\gamma = 50$ and $\gamma = 100$. For each selection coefficient, pointwise 0%, 25%, 50%, 75% and 100% quantiles are calculated. Solid line is the 50% quantile, dashed line indicates 25% and 75% quantiles, and the dotted line indicates 0% and 100% quantiles.

(1974) that, conditioned on eventual fixation, the sign of the selection coefficient is irrelevant to the distribution of the Wright-Fisher diffusion path. Also apparent is that bridges with strong natural selection will be more likely to be found near the boundary than bridges under neutrality. Finally, because $0 \leq \sin^2(x) \leq 1$, we see that, very loosely, a bridge will look approximately neutral if

$$\frac{1}{8}\gamma^2 T \approx 0. \tag{2.21}$$

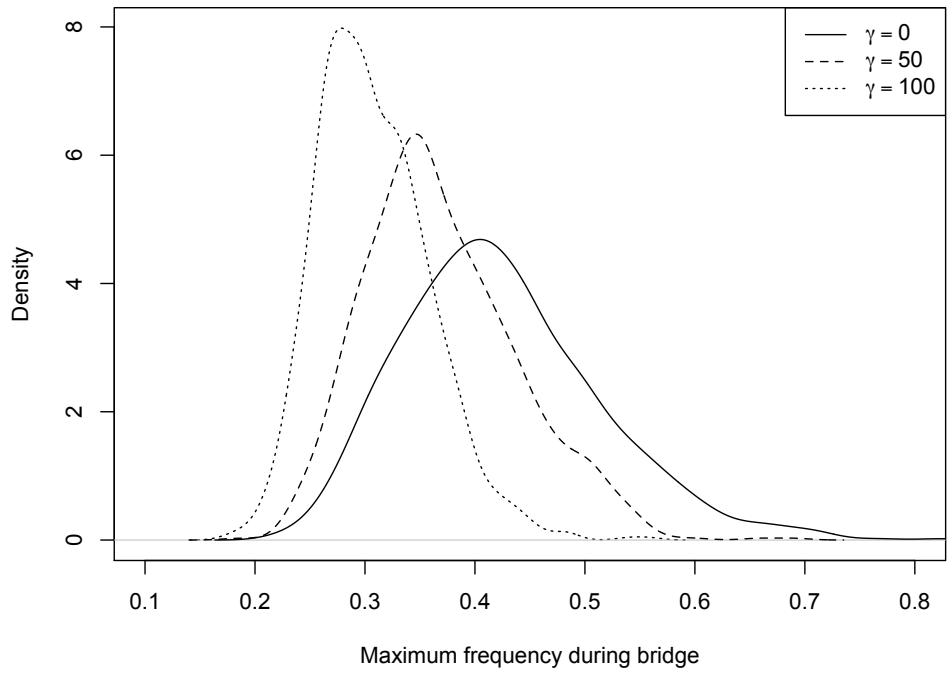

Figure 2.3: Densities of the maximum in a 0 to 0 bridge over the time interval $[0, T] = [0, 0.1]$ for the selection strengths $\gamma = 0$, $\gamma = 50$ and $\gamma = 100$.

2.4 Discussion

We have examined the behavior of Wright-Fisher diffusion bridges under both neutral models and models with genic selection. Although various conditioned Wright-Fisher diffusions have been studied in the past, Wright-Fisher diffusions conditioned to obtain a specific value at a predetermined time have not been studied extensively. We have elucidated some of the properties of Wright-Fisher bridges using a combination of analytical theory and simulations.

In contrast to Brownian motion with drift, for which the distribution of a bridge does not depend on the magnitude of the drift coefficient, the distribution of a Wright-Fisher bridge does depend on the magnitude of the selection coefficient. As one might expect, bridges under strong selection are more constrained than neutral bridges. This can clearly be seen in Figure 2.2, in which the bridge with $\gamma = 0$ has a broad range, but when $\gamma = 100$ the paths of the bridge are highly likely to be confined near the boundary at 0 until quite late in the bridge. A similar conclusion can be drawn from Figure 2.3 which shows the density of the maximum in a bridge from 0 to 0 over the time interval $[0, T] = [0, 0.1]$. The expected maximum of a neutral bridge is much higher than one with strong selection, and there is

significantly more variance about that maximum under neutrality.

Much of the behavior of Wright-Fisher bridges under selection can be understood in terms of the likelihood ratio (2.20). Because $\sin(x)$ takes its smallest values for $x \approx 0$ and $x \approx \pi$, very strong selection will confine a bridge of the transformed process Y to near these boundaries. Intuitively, this is because the Wright-Fisher diffusion has the largest magnitude of drift and diffusion coefficients at $x = 0.5$, and thus the diffusion moves "faster" when it is away from the boundaries 0 and 1. In order for a diffusion with a large selection coefficient to reach an interior point after a large amount of time, it must spend most of that time near the boundary.

However, these differences between selection and neutrality are mostly apparent in cases of extreme selection coefficients or very long times. This has important implications for maximum likelihood inference of selection coefficients from allele frequency time series. Because the realizations are likely to be quite similar for a selected allele and a neutral allele when the selection coefficient is moderate, most of the information about the selection coefficient comes from the end-points. This is consistent with the work of Watterson (1979), who showed that even with the whole sample path, it is difficult to reject neutrality when selection is weak. Therefore, in many cases increasing the time-density of samples may not provide much additional information about the selection coefficient. Because many allelic time-series are obtained via costly ancient DNA techniques, this is an important consideration for the many researchers who are interested in the history of selection acting on a particular allele.

In addition to results directly concerning bridges, we have made several technical advances in the analysis of the Wright-Fisher diffusion. We have developed the theory of first passage times of a neutral Wright-Fisher diffusion starting from low frequency and we were able to provide a closed-form for the density of the maximum in a neutral bridge that goes from 0 to 0.

While our rejection sampling scheme is similar to that of Beskos & Roberts (2005) in some regards, there are several differences. Primarily, we do not provide exact samples, in the sense that Beskos & Roberts (2005) does. Because we store a discrete representation of our proposal bridges in computer memory, the calculation of (2.15) is necessarily an approximation, and hence the samples are only approximate. However, Figure 2.1 shows that they are extremely accurate. Also, because we are concerned with a specific model, we used 4-dimensional Bessel bridges, instead of Brownian bridges, in our proposal mechanism. This choice is superior for the Wright-Fisher diffusion because both the Bessel bridge and the Wright-Fisher bridge have boundaries at 0 with asymptotically equivalent singularities in the drift coefficient, while the Brownian bridge can assume negative values and hence result an unacceptably high rejection rate when it is used as a proposal distribution. Ideally, we would sample from a proposal distribution that describes a diffusion that was also bounded above and had a suitable singularity in its drift coefficient at the upper boundary; however, we have not yet discovered an appropriate diffusion for which it is easy to sample the corresponding bridges. Finally, we make use of the "most likely" bridge path as a means of guiding samples of bridges that are likely to be extremely different from those generated by the 4-dimensional Bessel bridge proposal distribution. This modification is akin to shifting the mean of a

proposal distribution when doing rejection sampling of a 1-dimensional random variable, and it greatly increases the efficiency of sampling.

Chapter 3

A Path Integral Formulation of the Wright-Fisher Process with Genic Selection

3.1 Introduction

Modern population genetics theory can be broken down into two broad subclasses: forward-in-time, in which the generation-to-generation allele frequency dynamics are tracked, and backward-in-time, in which genealogical relationships are modeled. While forward-in-time models were developed first, the introduction of the coalescent by Kingman (1982) ushered in a revolution in our understanding of neutral genetic variation. The success of the coalescent in providing a simple framework for analyzing neutral loci has inspired a number of attempts to construct a genealogical representation of models with natural selection (Krone & Neuhauser 1997; Neuhauser & Krone 1997; Donnelly & Kurtz 1999). However these models have not been particularly amenable to analysis due to their complicated structure.

The forward-in-time approach remains the most straight-forward method for analyzing genetic variation under the combined effects of genetic drift and natural selection. This approach is characterized by the diffusion approximation to the Wright-Fisher model (Ewens 2004). For many important quantities (such as ultimate fixation probabilities), the diffusion approximation provides a concise, exact analytic expression. These formulas, in terms of common parameters such as the population scaled selection coefficient α, allow for an understanding of how different evolutionary forces impact the dynamics of allele frequency change. Assuming a constant population size, exact analytic results from the diffusion approximation can even be used to estimate the distribution of selection coefficients in the genome (Boyko et al. 2008; Torgerson et al. 2009).

Unfortunately, when both selection and genetic drift affect allele frequency dynamics, there is no simple analytic expression for the transition density of the diffusion (that is, the probability that an allele currently at frequency x is at frequency y after t time units have passed). Recently, interest in the transition density has been fueled by advances in

experimental evolution (Kawecki et al. 2012) and ancient DNA (Wall & Slatkin 2012), leading to the development of numerous methods for estimating the population scaled selection coefficient from allele frequency time series data (Bollback et al. 2008b; Malaspinas et al. 2012; Mathieson & McVean 2013; Feder et al. 2013). Moreover, because the transition density fully characterizes the allele frequency dynamics, many interesting quantities, such as the time-dependent fixation probability, could be calculated once the transition density is known.

While the diffusion approximation allows one to write down a partial differential equation (PDE) that the transition density must satisfy, it has proved challenging to solve in a robust manner either analytically or numerically. Numerical solution of the PDE is, in principle, straightforward by discretization techniques (see Zhao et al. (2013) for a recent approach that accounts for fixations and losses of alleles). However, because the relative importance of drift and selection depend on the allele frequency, the discretization scheme must be chosen wisely. Another drawback of numerical methods is that they can be quite time consuming; in particular, this is what limits the method of Gutenkunst et al. (2009) to 3 populations while using a diffusion approximation to find the site frequency spectrum for demographic inference.

Kimura (1955b) provided an analytical solution to the transitional density with selection, in the form of an eigenfunction decomposition with oblate spheroid wave functions. However, he was unable to compute the eigenvalues exactly, instead resorting to perturbation theory. Motivated by the fact that the eigenfunction decomposition of the model with no selection is known, Song & Steinrücken (2012) developed a novel computational method for approximating the transition density analytically. Their method, based on the theory of Hilbert spaces spanned by orthogonal polynomials, is a significant advance and represents the state-of-the-art in terms of finding the transition density with selection. This method still has several limitations, as it needs to be recomputed if a new selection coefficient is chosen; moreover, for certain values of the selection coefficient and dominance parameter, computation times can be long because they were required to use high-precision arithmetic.

In this paper, I present a novel method for approximating the transition density of the Wright-Fisher diffusion with genic selection. This method is based on the theory of path integration, which was introduced by Wiener (1921) for Brownian motion and has found substantial success in applications in quantum mechanics (Feynman 1948; Feynman & Hibbs 2012) and quantum field theory (Zee 2010). The key insight of of this approach is to associate every path from x at time 0 to y at time t with a probability, and then integrate over all possible paths to find the transition density. While computing this integral exactly is only possible in the neutral case, I develop a perturbation scheme to approximate it as a power series in α for the case with genic selection. To facilitate computation of this perturbation expansion, I demonstrate the use of a mnemonic, called Feynman diagrams, to compute the transition density to arbitrary accuracy.

3.2 Methods

3.2.1 Partial differential equation formulation

Here I review some preliminaries about the Wright-Fisher diffusion that will prove useful in the following. Denoting by $\phi_\alpha(x, y; t)$ the transition density with genic selection and population-scaled selection coefficient α, standard theory shows that ϕ satisfies the PDE

$$\frac{\partial}{\partial t}\phi_\alpha(x, y; t) = \frac{1}{2}\frac{\partial^2}{\partial y^2}\left\{y(1-y)\phi_\alpha(x, y; t)\right\} - \alpha\frac{\partial}{\partial y}\left\{y(1-y)\phi_\alpha(x, y; t)\right\}, \tag{3.1}$$

with the initial condition $\phi_\alpha(x, y, 0) = \delta(x - y)$ where $\delta(\cdot)$ is the usual Dirac delta function (Ewens 2004).

Kimura (1955a) found that the for the case $\alpha = 0$, the transition density admits an eigenfunction decomposition,

$$4x(1-x)\sum_{i=1}^{\infty}\frac{2i+1}{i(i+1)}C_{i-1}^{(3/2)}(1-2x)C_{i-1}^{(3/2)}(1-2y)e^{-\frac{1}{2}i(i+1)t}, \tag{3.2}$$

where the $C_i^\lambda(z)$ are the Gegenbauer polynomials.

3.2.2 Path integral formulation

The path integral formulation begins by defining a probability density functional, which assigns a probability density to any path from x to y. Then, the total transition probability from x to y is computed by integrating this density over all paths from x to y.

This probability density functional can be developed intuitively by considering the "short-time transition densities". Standard theory for diffusion processes shows that when $\delta t \ll 1$, we can approximate

$$\phi_\alpha(x, y; \delta t) \approx \frac{1}{\sqrt{2\pi x(1-x)\delta t}}\exp\left\{\frac{(y - (x + \alpha x(1-x)))^2}{2x(1-x)\delta t}\right\}$$

A naive approach might be to attempt to approximate the probability density of a path by dividing the interval $[0, t]$ into n intervals of length δt. Then we approximate with the probability of the so-called "zig-zag path",

$$\mathcal{P}[z] \approx \prod_{i=1}^{n}\phi_\alpha(z_{i-1}, z_i; \delta t)dz_i,$$

with $z_i = z(i\delta t)$. However, this fails for a variety of reasons, in particular the dependence of the diffusion coefficient on the current allele frequency (Graham 1977; Dürr & Bach 1978).

Instead, I compute the relative probability density function for a path with selection compared to a neutral path. This functional, which can be rigorously derived using Girsanov's theorem (Rogers & Williams 2000b) can be intuitively developed as

$$\begin{aligned}\mathcal{G}[z] &\approx \frac{\prod_{i=1}^{n}\phi_\alpha(z_{i-1},z_i;\delta t)dz_i}{\prod_{i=1}^{n}\phi_0(z_{i-1},z_i;\delta t)dz_i}\\ &\approx \frac{\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi z_{i-1}(1-z_{i-1})\delta t}}\exp\left\{\frac{(z_i-(z_{i-1}+\alpha z_{i-1}(1-z_{i-1})))^2}{2z_{i-1}(1-z_{i-1})\delta t}\right\}dz_i}{\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi z_{i-1}(1-z_{i-1})\delta t}}\exp\left\{\frac{(z_i-z_{i-1})^2}{2z_{i-1}(1-z_{i-1})\delta t}\right\}dz_i}\\ &= \exp\left\{\alpha\sum_{i=1}^{n}(z_i-z_{i-1})-\frac{\alpha^2}{2}\sum_{i=1}^{n}z_i(1-z_i)\delta t\right\}.\end{aligned}$$

Thus, as $\delta t \downarrow 0$ and $n \uparrow \infty$ such that $n\delta t = t$, we have

$$\mathcal{G}[z] = \exp\left\{\alpha(y-x)-\frac{\alpha^2}{2}\int_0^t z(1-z)ds\right\}, \tag{3.3}$$

in which the time dependence of z is suppressed for notational convenience. Now, we can write the transition density as the integral over all *neutral* Wright-Fisher paths of the relative probability of that path with selection,

$$\phi_\alpha(x,y;t) = \int_{(0,x)}^{(t,y)} e^{\alpha(y-x)-\frac{\alpha^2}{2}\int_0^t z(1-z)ds}\mathcal{D}z \tag{3.4}$$

where $\mathcal{D}z$ is the measure on path space induced by the neutral Wright-Fisher process.

The path integral (3.4) can be understood as depicted in Figure 3.1. Paths from x to y can be drawn from the neutral Wright-Fisher path measure, $\mathcal{D}z$. For each path, the functional $\mathcal{G}[\cdot]$ is evaluated (panel a). This results in a one-dimensional probability distribution for values of $\mathcal{G}$ (panel b). Then, the mean value of of the distribution of $\mathcal{G}$ can be computed, which is equal to the transition density.

3.2.3 Perturbation approximation

I now show how to approximate the transition density using a perturbation expansion. Note that the first term in the exponential of (3.4) is independent of the path, and hence we focus on the path integral

$$\int_{(0,x)}^{(t,y)} e^{-\frac{\alpha^2}{2}\int_0^t z(1-z)ds}\mathcal{D}z.$$

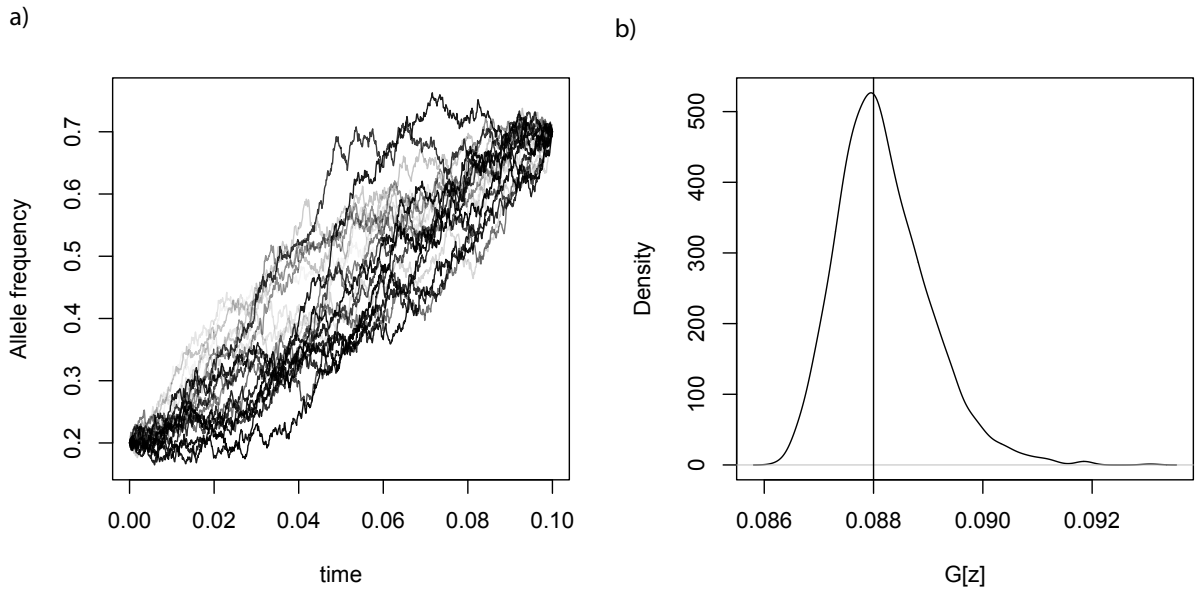

Figure 3.1: Path integrals. In panel a, neutral Wright-Fisher paths starting at $x = 0.2$ and ending at $y = 0.7$ after 0.1 time units have passed are sampled using the rejection-sampling method of Schraiber et al. (2013). Paths are colored according to $\mathcal{G}[z]$ with $\alpha = 5$, with darker paths corresponding to larger $\mathcal{G}[z]$ values. In panel b, the density of $\mathcal{G}[z]$ is plotted. The path integral estimate of the transition density for a Wright-Fisher process with $\alpha = 5$ to go from 0.2 to 0.7 in 0.1 time units is the mean value of the density in panel b, indicated by the vertical line.

We begin by expanding the exponential in a Taylor series about $\alpha = 0$,

$$\begin{aligned}\int_{(0,x)}^{(t,y)} e^{-\frac{\alpha^2}{2}\int_0^t z(1-z)ds}\mathcal{D}z &= \int_{(0,x)}^{(t,y)} \sum_{k=0}^{\infty}(-1)^k \frac{\alpha^{2k}}{2^k}\frac{1}{k!}\left(\int_0^t z(1-z)ds\right)^k \mathcal{D}z \\ &= \sum_{k=0}^{\infty}(-1)^k \frac{\alpha^{2k}}{2^k}\frac{1}{k!}\int_{(0,x)}^{(t,y)}\left(\int_0^t z(1-z)ds\right)^k \mathcal{D}z. \end{aligned} \tag{3.5}$$

In the Appendix, I show that the exchange of the summation and the integral is justified by Fubini's theorem. This is in stark contrast to the case in quantum physics, in which the exchange of the sum and integral is often not justified, leading to zero radius of convergence in the perturbation parameter.

Thus, the task of approximating the transition density with selection is reduced to the task of computing the functional integrals

$$\int_{(0,x)}^{(t,y)}\left(\int_0^t z(1-z)ds\right)^k \mathcal{D}z.$$

Integrals of this form were considered by Nagylaki (1974) and Watterson (1979) although they were focused on the case where the allele is eventually fixed or lost, whereas here we need to consider only those paths that go from x to y in time t. To compute these integrals, it is useful to introduce a diagrammatic method, known as a Feynman diagram (Feynman & Hibbs 2012; Chorin & Hald 2006). Borrowing from the language of physics momentarily, we can regard $V(x) = x(1-x)$ as a *potential energy*, and we can consider the allele frequency being *scattered* by the potential.

The idea can be seen in Figure 3.2. When the integrand is raised to the kth power, we imagine that the allele frequency changes neutrally until some time s_1, at which point it interacts with the potential and is scattered. Then, it evolves neutral until time s_2, at which point it again interacts with the potential and is scattered. This proceeds until the scattering at time s_k, after which the allele frequency evolves neutrally to y at time t. Because the interaction times s_i could have happened at any time between 0 and t and the allele frequency z_i at time s_i is random, we integrate over all times and allele frequencies. For example, we can compute

$$\int_{(0,x)}^{(t,y)} \left(\int_0^t z(1-z)ds \right) \mathcal{D}z = \int_0^t \int_0^1 \phi_0(x, z_1; s_1) z_1(1-z_1) \phi_0(z_1, y; t-s_1) dz_1 ds_1 \tag{3.6}$$

and

$$\int_{(0,x)}^{(t,y)} \left(\int_0^t z(1-z)ds \right)^2 \mathcal{D}z =$$
$$2 \int_0^t \int_0^{s_2} \int_0^1 \int_0^1 \phi_0(x, z_1; s_1) z_1(1-z_1) \phi_0(z_1, z_2; s_2 - s_1) z_2(1-z_2) \phi_0(z_2, y; t-s_2) dz_1 dz_2 ds_1 ds_2,$$

where the factor of 2 comes from the two orderings in which the scatterings happened. In general, the kth order Feynman diagram will come with a factor of $k!$ to count the number of orderings of the scattering events.

Because we know the neutral transition density, computing the integrals that arise from Feynman diagrams is straightforward. Unfortunately, the neutral transition density is only known as an infinite series and in practice computing the integrals is more difficult. In the Appendix, I show how to achieve efficient computation of these integrals for arbitrary k.

3.3 Results

3.3.1 Accuracy of the perturbation expansion

A simple error bound can be derived for perturbation expansions (Harlow 2009). For the kth-order perturbation expansion, $\phi_\alpha^{(k)}(x, y; t)$, this bound is

$$\left| \phi_\alpha(x, y; t) - \phi_\alpha^{(k)}(x, y; t) \right| \leq \left| \frac{\alpha^{2(k+1)}}{2^{k+1}(k+1)!} \int_{(0,x)}^{(t,y)} \left(\int_0^t z(1-z)ds. \right)^{k+1} \mathcal{D}z \right|. \tag{3.7}$$

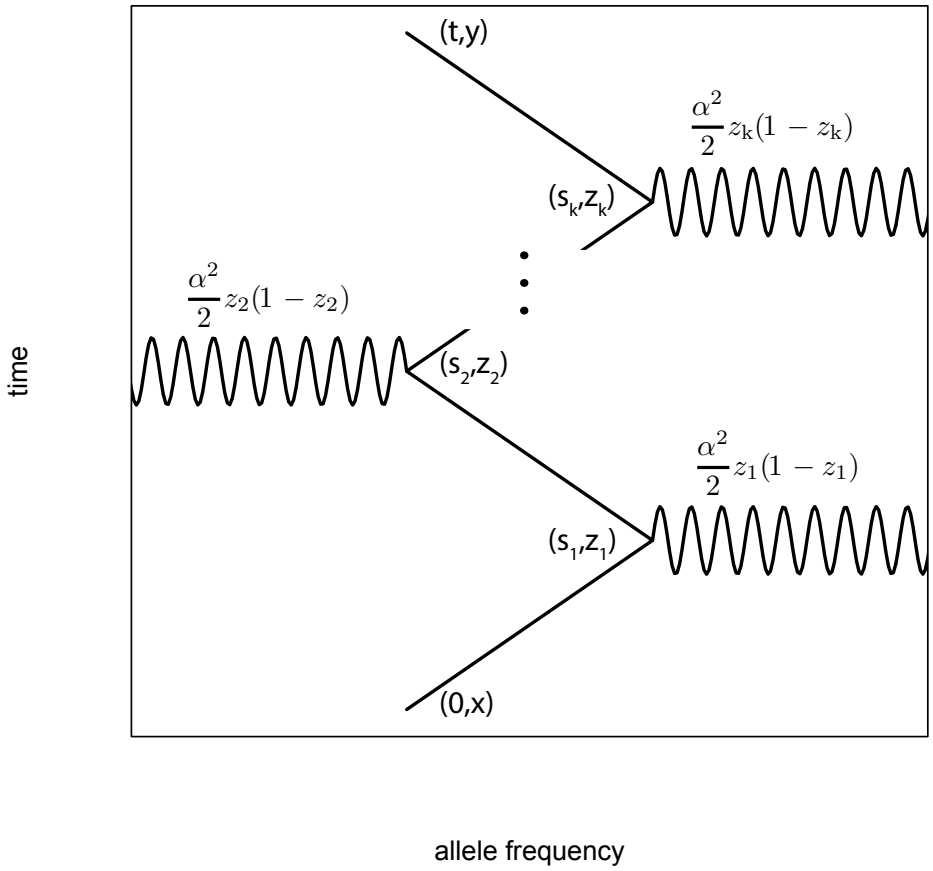

Figure 3.2: Feynman diagrams. Feynman diagrams are used to evaluate the integrals that show up in the perturbation expansion. The allele starts at time 0 and frequency x, evolving neutrally until time s_1, when it has frequency z_1 and is perturbed by natural selection. It then evolves to time s_2 and allele frequency z_2, at which point it is again perturbed by natural selection. This continues until the final perturbation at time s_k and frequency z_k, after which it evolves neutrally to time t and frequency y.

As argued in the Appendix, this bound is less than

$$\frac{\alpha^{2(k+1)} t^{k+1}}{8^{k+1}(k+1)!} \phi_0(x, y; t),$$

and when $t < 4$, it approaches 0 as $k \to \infty$ for any α. Thus, the perturbation expansion convergences to the true transition density for any α, provided that t is small enough.

The error bound presented above is rather crude. To get a more informative picture of the

accuracy of the perturbation expansion, I compared to simulations. An interesting quantity that sums up the overall accuracy of the perturbation approximation is the time-dependent probability of absorption. This quantity can be calculated analytically as

$$\int_0^1 \phi_\alpha(x, y; t)dy$$

and is easily estimated from simulations. The perturbation method proves to be increasing accurate as more terms are added to the expansion (Figure 3.3). However, even for moderate values of α, a large number of terms are required for an accurate approximation.

3.4 Discussion

The Wright-Fisher process with selection is a primary tool for elucidating the impact of natural selection on genetic variation. However, the transient behavior of the process has been difficult to study, with much work focusing on equilibrium aspects, such as the stationary distribution (Wright 1931a) or the site frequency spectrum (Sawyer & Hartl 1992). Nonetheless, transient dynamics have an important impact on natural variation and are critical to forming a complete understanding of how natural selection shapes genomes. In this paper, I presented a novel path integral formulation of the Wright-Fisher process with genic selection. This led naturally to a simple perturbation scheme for computing the transition density with weak selection.

The perturbation expansion of the transition density can be understood by using Feynman diagrams (Figure 2). Although the traditional motivation for Feynman diagrams comes from quantum physics (Feynman & Hibbs 2012), they can be interpreted in a population genetic context. For instance, in the first-order term of the perturbation expansion, an allele begins drifting neutrally. At a time when the allele frequency is z, there is a rate $\frac{\alpha^2}{2}z(1-z)$ of a selective event occurring, which has a natural interpretation as an individual of one allelic type encountering an individual of the other allelic type, weighted by the strength of selection. The strength of selection enters as α^2 because the rate of selective events must be positive. After the selective event occurs, the allele again drifts neutrally. Higher-order terms in the perturbation expansion include more selective events. However, the impact of selective events on an overall neutral trajectory is not sufficient to model the dynamics. The additional factor of $e^{\alpha(y-x)}$ multiplying the entire perturbation expansion arises to account for the deterministic effects of selection (Baibuz et al. 1984); intuitively, it acts to increase the probability that an allele increases in frequency under positive selection or decreases in frequency under negative selection.

The perturbation scheme described in this paper works best for weak selection. For the weak selection scheme examined here, this method and the method of Song & Steinrücken (2012) perform comparably in terms of time taken to obtain an estimate of the transition density. For stronger selection, other approximation methods, including the Gaussian diffusion approximation (Nagylaki 1990; Feder et al. 2013) may be useful. Also, the model

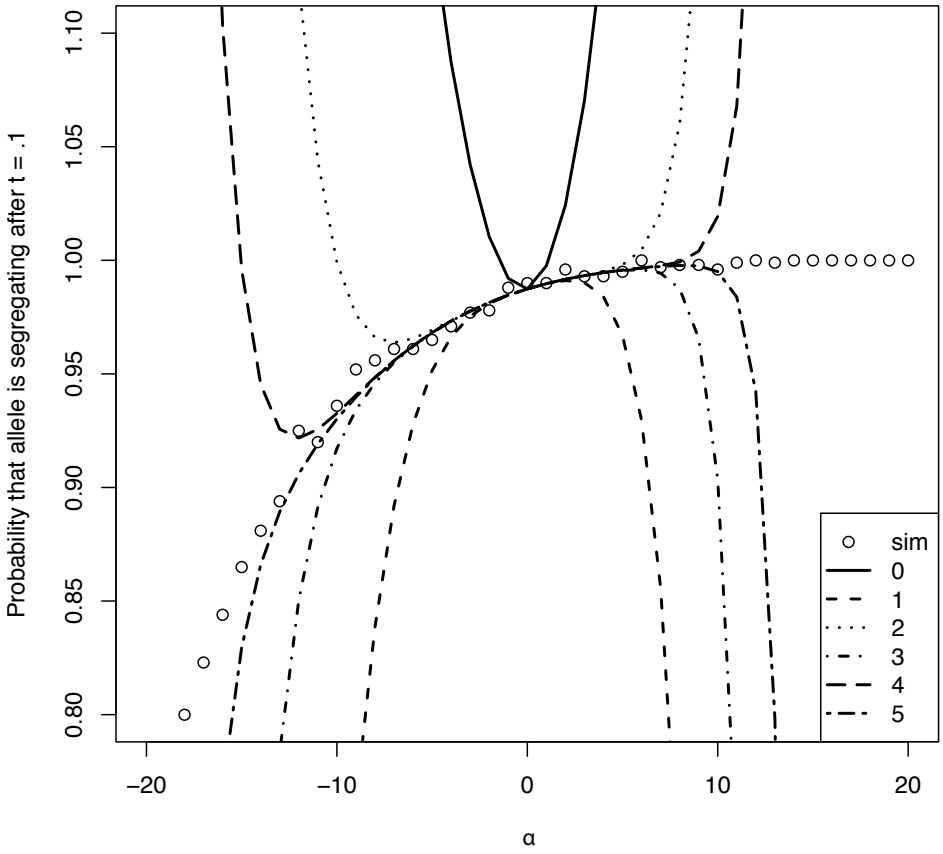

Figure 3.3: Accuracy of the perturbation expansion. The perturbation expansion is compared to simulations of the Wright-Fisher diffusion. An allele with starting frequency $x = .2$ evolves under genetic drift and natural selection for $t = .1$ with a variety of selection coefficients. The probability that the allele was not absorbed is then computed. Dots show the values from simulations while lines indicate successively higher orders of perturbation expansion.

considered in this paper does not have fully general diploid selection. The path integral approach applies in an equally straightforward fashion to diploid selection, but the details of the mathematics become significantly more complicated. In that case, the orthogonal polynomial method of Song & Steinrücken (2012) may be better suited.

Path integral formulations have been used successfully in population genetics in the past. Rouhani & Barton (1987) made use of a path integral to approximate the probability of shifting between selective optima in the context of quantitative trait evolution. Their approximation scheme, however, was quite different than one that I explored. They assumed

relatively strong selection and expanded the path integral around the most likely path between the two selective optima, in contrast to the weak selection perturbative approach taken in this paper. More recently, path integrals have been used to examine fitness flux (Mustonen & Lässig 2010) and Muller's ratchet (Neher & Shraiman 2012).

A significant strength of the path integral approach is its adaptability to evolution of a locus with a large number of alleles, each of which corresponds to a phenotypic value. In previous contexts, such a model has been used to model quantitative trait evolution, called a continuum-of-alleles model (Kimura 1965). Earlier approaches to incorporate genetic drift into a continuum-of-alleles model using the theory of measure-valued diffusions (Fleming & Viot 1979; Ethier & Kurtz 1987) have made significant advances in understanding the neutral dynamics of such processes (Dawson & Hochberg 1982; Ethier & Griffiths 1993; Ethier & Kurtz 1993; Donnelly & Kurtz 1996). However, incorporating selection greatly increases the difficulty of obtaining analytical results (but see Donnelly & Kurtz (1999); Dawson & Feng (2001)). It is possible that a path integral formulation of such a process could lead to a perturbative approach to incorporating selection, in much the same way as the path integral approach has been successful in perturbative quantum field theory (Zee 2010).

3.5 Appendix

3.5.1 Exchanging the order of integration and summation in (3.5) is justified

To establish this fact, we first need a simple bound on the functional

$$\mathcal{F}_k[X_s] = \frac{\alpha^{2k}}{2^k k!}\left(\int_0^t z(1-z)ds.\right)^k$$

Note that, because z represents a frequency, we know that z is bounded between 0 and 1 for all s. Thus,

$$\begin{aligned}\int_0^t z(1-z)ds &\leq \int_0^t \frac{1}{2}\left(1-\frac{1}{2}\right)ds \\ &= \frac{1}{4}t.\end{aligned}$$

Therefore,

$$\mathcal{F}_k[z] \leq \frac{\alpha^{2k}t^k}{8^k k!}.$$

Now, to apply the general version of Fubini's theorem, which allows the exchange of the order of integration in general measure spaces (Bogachev 2007), I must show that

$$\int_{(0,x)}^{(t,y)}\left(\sum_{k=0}^{\infty}\mathcal{F}_k[z]\right)\mathcal{D}z < \infty$$

and

$$\sum_{k=0}^{\infty} \int_{(0,x)}^{(t,y)} \mathcal{F}_k[z]\mathcal{D}z < \infty$$

For the first case, observe that

$$\begin{aligned}
\int_{(0,x)}^{(t,y)} \left(\sum_{k=0}^{\infty} \mathcal{F}_k[z] \right) \mathcal{D}z &\leq \int_{(0,x)}^{(t,y)} \left(\sum_{k=0}^{\infty} \frac{\alpha^{2k} t^k}{8^k k!} \right) \mathcal{D}z \\
&= \int_{(0,x)}^{(t,y)} e^{\frac{\alpha^2}{8} t} \mathcal{D}z \\
&= e^{\frac{\alpha^2 t}{8}} \phi_0(x, y; t) \\
&< \infty.
\end{aligned}$$

An extremely similar calculation shows that the second case is true as well.

3.5.2 Computation of Feynman diagrams

The integrals arising from the Feynman diagrams can only be expressed properly as infinite sums. In practice, it is necessary to truncate these sums after a finite number of terms. In this section, I develop an approach efficiently compute the sums. From Kimura's spectral representation of the transition density, equation (3.2), and the general form of each term in the perturbation expansion, for example equation (3.6), it is clear that the kth order Feynman diagram results in a term

$$\begin{aligned}
4x(1-x) \sum_{i_1, i_2, \ldots, i_{k+1}} & C_{i_1-1}^{(3/2)}(1-2x) C_{i_{k+1}-1}^{(3/2)}(1-2y) \prod_{j=1}^{k+1} \frac{2i_j + 1}{i_j(i_j+1)} \\
& \times \prod_{j=1}^{k-1} \int_0^1 z_j^2 (1-z_j)^2 C_{i_j}^{(3/2)}(1-2z_j) C_{i_{j+1}}^{(3/2)}(1-2z_j) dz_j \qquad (3.8) \\
& \times \int_0^t \int_0^{s_k} \cdots \int_0^{s_1} e^{-(\lambda_{i_1} s_1 + \sum_{j=2}^{k} \lambda_{i_j}(s_j - s_{j-1}) + \lambda_{i_{k+1}}(t - s_k))} ds_1 \cdots ds_{k-1} ds_k,
\end{aligned}$$

with $\lambda_i = i(i+1)/2$, the eigenvalues of Kimura's transition density.

The integrals over allele frequencies can be done exactly by using the properties of the Gegenbauer polynomials. First,

$$\int_0^1 x(1-x) C_i^{(3/2)}(1-2x) C_j^{(3/2)}(1-2x) dx = -\frac{1}{32} \int_{-1}^1 (1-z^2)^2 C_i^{(3/2)}(z) C_j^{(3/2)}(z) dz \qquad (3.9)$$

after making the substitution $z = 1 - 2x$. This puts the Gegenbauer polynomials on their natural domain, $[-1, 1]$. Now, multiplying through by one of the factors of $(1 - z^2)$,

$$\int_{-1}^1 (1-z^2)^2 C_i^{(3/2)}(z) C_j^{(3/2)}(z) dz = \int_{-1}^1 (1-z^2) C_i^{(3/2)}(z) C_j^{(3/2)}(z) dz - \int_{-1}^1 (1-z^2) z C_i^{(3/2)}(z) z C_j^{(3/2)}(z) dz.$$

The first term can be recognized as the orthogonality condition for the Gegenbauer polynomials and hence,

$$\int_{-1}^{1}(1-z^2)C_i^{(3/2)}(z)C_j^{(3/2)}(z)dz = \delta_{i,j}\frac{2(i+1)(i+2)}{3+2i}.$$

To simplify the second term, use the recurrence relation for the Gegenbauer polynomials to find that,

$$zC_i^{(3/2)}(z) = \frac{1}{3+2i}\left((i+1)C_{i+1}^{(3/2)}(z) + (i+2)C_{i-1}^{(3/2)}(z)\right).$$

Substituting and multiplying through yields

$$\begin{aligned}\int_{-1}^{1}(1-z^2)zC_i^{(3/2)}(z)zC_j^{(3/2)}(z)dz = &\frac{1}{(2i+3)(2j+3)}\\ &\times\Bigg((i+1)(j+1)\int_{-1}^{1}(1-z^2)C_{i+1}^{(3/2)}(z)C_{j+1}^{(3/2)}(z)dz\\ &\times(i+1)(j+2)\int_{-1}^{1}(1-z^2)C_{i+1}^{(3/2)}(z)C_{j-1}^{(3/2)}(z)dz\\ &\times(i+2)(j+1)\int_{-1}^{1}(1-z^2)C_{j-1}^{(3/2)}(z)C_{j+1}^{(3/2)}(z)dz\\ &\times(i+2)(j+2)\int_{-1}^{1}(1-z^2)C_{i-1}^{(3/2)}(z)C_{j-1}^{(3/2)}(z)dz\Bigg).\end{aligned}$$

Again, these integrals can be simplified using the orthogonality of the Gegenbauer polynomials to finally see that the integral in (3.9) equals

$$\begin{aligned}-\frac{1}{32}\Bigg(&\delta_{i,j}\frac{2(i+1)(i+2)}{3+2i} + \frac{1}{(2i+3)(2j+3)}\Bigg(\delta_{i,j}(i+1)^2\frac{2(i+2)(i+3)}{5+2i}\\ &+\delta_{i,j-2}(i+1)(i+4)\frac{2(i+2)(i+3)}{5+2i} + \delta_{i,j+2}(i+2)(i-1)\frac{i(i+1)}{1+2i}\\ &+\delta_{i,j}(i+2)^2\frac{i(i+1)}{1+2i}\Bigg)\Bigg).\end{aligned} \tag{3.10}$$

An important consequence of this fact is that the many of the terms in the sum (3.8) are equal to zero.

The integrals over the intermediate times can also be evaluated exactly, although I have not been able to find a general formula. In this case, it is straight-forward to precompute the integral for all possible sets of equal indices and then substitute into the sum.

Chapter 4

Bayesian inference of natural selection from allele frequency time series

4.1 Introduction

The ability to obtain high-quality genetic data from ancient samples is revolutionizing the way that we understand the evolutionary history of populations. One of the most powerful applications of ancient DNA (aDNA) is to study the action of natural selection. While methods making use of only modern DNA sequences have successfully identified loci evolving subject to natural selection (Nielsen et al. 2005a; Voight et al. 2006; Pickrell et al. 2009), they are inherently limited because they look indirectly for selection, finding its signature in nearby neutral variation. In contrast, by sequencing ancient individuals, it is possible to directly track the change in allele frequency that is characteristic of the action of natural selection.

Several studies have obtained aDNA at the lactase locus (LCT) in humans. While the allele of LCT that confers lactase persistence is at high frequency in much of Europe, aDNA of individuals ranging from 4000-8000 BCE shows that the lactase persistence allele had an extremely low frequency during that time period (Burger et al. 2007; Malmström et al. 2010; Lacan et al. 2011; Plantinga et al. 2012). This is interpreted as evidence that the lactase persistence allele has been under extremely strong selection due to the onset of dairy farming in Europe.

To infer the action of natural selection more rigorously, several methods have been developed to explicitly fit a population genetic model to a time series of allele frequencies obtained via aDNA. Bollback et al. (2008b) extended an approach devised by Williamson & Slatkin (1999) to estimate the population-scaled selection coefficient, $\alpha = 2N_e s$, along with the effective size, N_e. To incorporate natural selection, Bollback et al. (2008b) used the continuous diffusion approximation to the Wright-Fisher model. This required them to use numerical techniques to solve the partial differential equation (PDE) associated with transition densities of the Wright-Fisher diffusion to calculate the probabilities of the population allele frequencies at each time point. Ludwig et al. (2009) obtained an aDNA time series

from 6 coat-color-related loci in horses and applied the method of Bollback et al. (2008b) to find that 2 of them, ASIP and MC1R, showed evidence of strong positive selection.

Recently, a number of methods have been proposed to extend the generality of the Bollback et al. (2008b) framework. To define their HMM, Bollback et al. (2008b) were required to posit a prior distribution on the allele frequency at the first time point. They chose to use a uniform prior on the initial frequency; however, in truth the initial allele frequency is dictated by the fact that the allele at some point arose as a new mutation. Using this information, Malaspinas et al. (2012) developed a method that also infers allele age. They also extended the selection model of Bollback et al. (2008b) to include fully recessive fitness effects. A more general selective model was implemented by Steinrücken et al. (2013), who model general diploid selection, and hence are able to fit data where selection acts in an over- or under-dominant fashion, although they assumed recurrent mutation and hence could not estimate allele age. The work of Mathieson & McVean (2013) is designed for inference of metapopulations over short time scales, due to using a discrete approximation to the Wright-Fisher diffusion. Finally, the approach of Feder et al. (2013) is ideally suited to experimental evolution studies, because they work in a strong selection, weak drift limit that is common in evolving microbial populations.

One key way that these methods differ from each other is in how they compute the probability of the underlying allele frequency changes. For instance, Malaspinas et al. (2012) approximate the diffusion with a one-step Markov chain while Steinrücken et al. (2013) calculate the likelihood analytically using a spectral representation of the diffusion discovered by Song & Steinrücken (2012). These different computational strategies are necessary because of the inherent difficulty in solving the Wright-Fisher PDE. A different approach, used by Mathieson & McVean (2013) in the context of a densely-sampled discrete Wright-Fisher model, is to instead compute the probability of the entire allele frequency trajectory in between sampling times.

In this work, we develop a novel approach for inference of general diploid selection and allele age from allele frequency time series obtained from aDNA. The key innovation of our approach is that we impute the allele frequency trajectory between sampled points when they are sparsely-sampled. This approach to inferring parameters from a sparsely-sampled diffusion is known as high-frequency path augmentation, and has been successfully applied in a number of contexts (Roberts & Stramer 2001; Golightly & Wilkinson 2005, 2008; Sørensen 2009; Fuchs 2013). The Wright-Fisher diffusion, however, has several features that are atypical in the context of high-frequency path augmentation, including a time-dependent diffusion coefficient and a bounded path-space. We then apply this new method to several datasets and find that we have power to estimate parameters of interest from real data.

4.2 Model

4.2.1 Generative model

We assume a randomly mating diploid population that is size $N(t)$ at time t, where t is measured in units of $2N_0$ generations for some arbitrary N_0. At the locus of interest, the ancestral allele, A_0, was fixed until some time t_0 when the derived alelle, A_1, arose with diploid fitnesses as given in Table 4.1.

Genotype	A_1A_1	A_1A_0	A_0A_0
Fitness	$1+s$	$1+hs$	1

Table 4.1: Fitness scheme assumed in the text.

The frequency of A_1 at time t, X_t, is modeled by the Wright-Fisher diffusion. While many treatments of the Wright-Fisher diffusion take a PDE approach (e.g. Ewens (2004)), we instead frame this in terms of a stochastic differential equation (SDE). X_t satisfies the SDE

$$dX_t = \alpha X_t(1-X_t)(X_t + h(1-2X_t))dt + \sqrt{\frac{X_t(1-X_t)}{\rho(t)}}dB_t \tag{4.1}$$

where $\alpha = 2N_0 s$ and $\rho(t) = N(t)/N_0$. Intuitively, this SDE says that for small δt,

$$X_{t+\delta t} \overset{\text{approx}}{\sim} \mathcal{N}\left(X_t + X_t(1-X_t)(X_t + h(1-2X_t))\delta t, \frac{X_t(1-X_t)}{\rho(t)}\delta t\right)$$

where $\mathcal{N}(\mu, \sigma^2)$ is a normal distribution with mean μ and variance σ^2.

We assume that the at times $t_1, t_2, \ldots, t_k$ samples of size $n_1, n_2, \ldots, n_k$ chromosomes are taken, and $c_1, c_2, \ldots, c_k$ copies of the derive allele are found at each time point (Figure 4.1). Note that it is possible that some of the sampling times are more ancient than t_0, the age of the allele.

4.2.2 Path likelihoods

It is possible to compute the likelihood of a diffusion sample path; however, there are some complications compared to the finite-dimensional random variables that are usually considered in population genetics inference. Suppose that a diffusion satisfies the SDE

$$dX_t = a(X_t, t)dt + dB_t \tag{4.2}$$

and let $\mathbb{P}$ be the distribution on paths satisfying (4.2). If we denote by $\mathbb{W}$ the distribution on Brownian motion paths, then Girsanov's theorem (Girsanov 1960) gives the likelihood of the path $\{X_s, t_0 \le s \le t\}$ under $\mathbb{P}$ relative to $\mathbb{W}$ as

$$\frac{d\mathbb{P}}{d\mathbb{W}}(X) = \exp\left\{\int_{t_0}^{t} a(X_s, s)dX_s - \frac{1}{2}\int_{t_0}^{t} a^2(X_s, s)ds\right\} \tag{4.3}$$

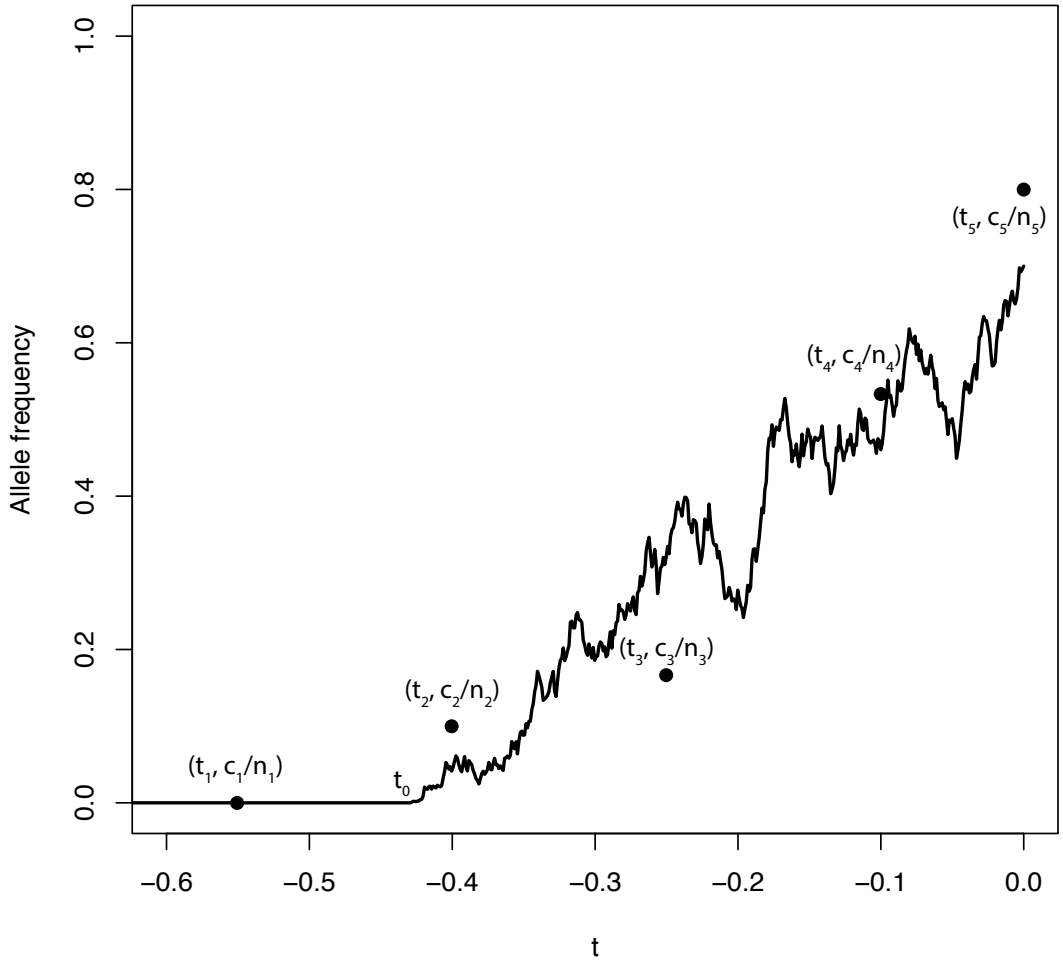

Figure 4.1: Taking samples from an allele frequency trajectory. An allele frequency trajectory is simulated from the Wright-Fisher diffusion (solid line). At each time, t_i, a sample of size n_i chromosomes is taken and c_i copies of the derived allele are observed. Each point corresponds to the observed allele frequency of sample i. Note that t_1 is more ancient than the allele age, t_0.

where the first integral in the exponentiand is an Ito integral.

However, the Wright-Fisher SDE (4.1) is not in the form (4.2). In particular, factor multiplying dB_t depends on both space and time. To deal with this issue, first applying the time transformation $\tau = f(t)$ with

$$f(t) = \int_0^t \frac{1}{\rho(s)} ds$$

to obtain a new SDE with time-independent diffusion coefficient,

$$dX_\tau = \alpha\rho(f^{-1}(\tau))X_\tau(1-X_\tau)(X_\tau + h(1-2X_\tau))d\tau + \sqrt{X_\tau(1-X_\tau)}dB_\tau. \tag{4.4}$$

Next, apply an angular transform first suggested by Fisher (1922b), $Y_\tau = \arccos(1-2X_\tau)$. Applying Ito's lemma (Itô 1944) shows that Y_t is a diffusion satisfying the SDE

$$dY_\tau = \frac{1}{4}\left(\alpha\rho(f^{-1}(\tau))\sin(Y_\tau)(1+(2h-1)\cos(Y_\tau)) - 2\cot(Y_\tau)\right)d\tau + dB_\tau. \tag{4.5}$$

Hence, we can write the likelihood of the transformed path relative to its density under a Brownian motion.

However, the drift coefficient of (4.5) has singularities at the boundaries 0 and π (corresponding to the boundaries 0 and 1 for allele frequencies). While this is acceptable if paths are confined away from the boundaries, it would be impossible estimate allele age because any path starting from 0 would have a likelihood of 0. We can accommodate this by letting $\mathbb{Q}$ be the distribution of a diffusion satisfying the SDE

$$dX_t = b(X_t, t)dt + dB_t \tag{4.6}$$

then we can compute the likelihood of a distribution a path under $\mathbb{P}$ (i.e. satisfying (4.2)) relative to $\mathbb{Q}$ (i.e. satisfying (4.6)) by

$$\begin{aligned}\frac{d\mathbb{P}}{d\mathbb{Q}}(X) &= \frac{d\mathbb{P}}{d\mathbb{W}}(X)\Big/\frac{d\mathbb{Q}}{d\mathbb{W}}(X) \\ &= \exp\left\{\int_{t_0}^{t}\left(a(X_s,s)-b(X_s,s)\right)dX_s - \frac{1}{2}\int_{t_0}^{t}\left(a^2(X_s,s)-b^2(X_s,s)\right)dt\right\}\end{aligned} \tag{4.7}$$

So, because

$$\frac{1}{4}\left(\alpha\rho(f^{-1}(\tau))\sin(Y_\tau)(1+(2h-1)\cos(Y_\tau)) - 2\cot(Y_\tau)\right) = -\frac{1}{2Y_\tau} + O(Y_\tau)$$

when Y_τ is small, a good choice for $\mathbb{Q}$ would be one where $b(x,t) \approx -1/(2x)$ as $x \downarrow 0$. An appropriate choice, along the lines of suggestions by (Schraiber et al. 2013) and (Jenkins 2013), is the Bessel(0) process, which satisfies

$$dX_t = -\frac{1}{2X_t}dt + dB_t. \tag{4.8}$$

This choice for $\mathbb{Q}$ not only ensures that there is no singularity at 0 but has desirable properties that will be exploited in the development of the Markov chain Monte Carlo algorithm. The Bessel(0) process is a natural choice to match the Wright-Fisher diffusion at small allele frequencies, as it can be derived from a branching process approximation to the allele frequency change (Haldane 1927; Feller 1951).

4.2.3 The joint likelihood of the data and the path

To write down down the full likelihood of the observations and the path, we make the assumption that $\rho(t)$ is continuously differentiable except at finitely many times $d_1 < d_2 < \ldots < d_M$, and we require that the population size function is such that $\rho(d_i^+) = \lim_{t \downarrow d_i} \rho(t)$ exists and is equal to $\rho(d_i)$ while $\rho(d_i^-) = \lim_{t \uparrow d_i} \rho(t)$ also exists (though it may not necessarily equal $\rho(d_i)$). That is, we assume that ρ is right continuous with left limits.

In the Appendix, we show that the likelihood of the data and the path, given α, h and t_0 can be written

$$\begin{aligned} L(D, Y|\alpha, h, t_0) = \exp \Bigg\{ & A(Y_{f(t_k)}, t_k^-) + A(Y_{f(d_m)}, d_m^-) - (A(Y_{f(d_K)}, d_K) + A(Y_{f(t_0)}, t_0)) \\ & + \sum_{i=m}^{K} \left[A(Y_{f(d_{i+1})}, d_{i+1}^-) - A(Y_{f(d_i)}, d_i) \right] \\ & - \int_{t_0}^{t_k} B(Y_{f(s)}, s) ds - \frac{1}{2} \int_{t_0}^{t_k} C(Y_{f(s)}, s) ds - \frac{1}{2} \int_{t_0}^{t_k} D(Y_{f(s)}, s) ds \Bigg\} \\ & \times \prod_{i=1}^{k} \binom{n_i}{c_i} \left(\frac{1 - \cos(Y_{f(t_i)})}{2} \right)^{c_i} \left(\frac{1 + \cos(Y_{f(t_i)})}{2} \right)^{n_i - c_i} \end{aligned} \tag{4.9}$$

where $m = \min\{i : d_i > t_0\}$ and $K = \max\{i : d_i > t_k\}$, and

$$\begin{aligned} A(y,t) &= \frac{\log(y)}{2} - \frac{1}{8} \left(\alpha \rho(t) \cos(y)(2 + (2h-1)) \cos(y)) + 4 \log(\sin(y)) \right) \\ B(y,t) &= -\frac{1}{8} \alpha \frac{d\rho}{dt}(t) \cos(y)(2 + (2h-1)\cos(y)) \\ C(y,t) &= \frac{1}{4} \left(\alpha(\cos(y) + (2h-1)\cos(y)) + 2\frac{\csc(y)^2}{\rho(t)} \right) - \frac{1}{2y^2 \rho(t)} \\ D(y,t) &= \frac{1}{16\rho(t)} \left(\alpha \rho(t) \sin(y)(1 + (2h-1)\cos(y)) - 2\cot(y) \right)^2 - \frac{1}{4y^2 \rho(t)}. \end{aligned}$$

4.3 Method

We developed a Markov chain Monte Carlo method for Bayesian inference of the parameters α, h and t_0. While updates to α and h do not require updating the path, updating t_0 requires proposing updates to the path. Additionally, we developed proposals to update small sections of the path without updating any parameters, as well as to update the allele frequency at the most recent sample time.

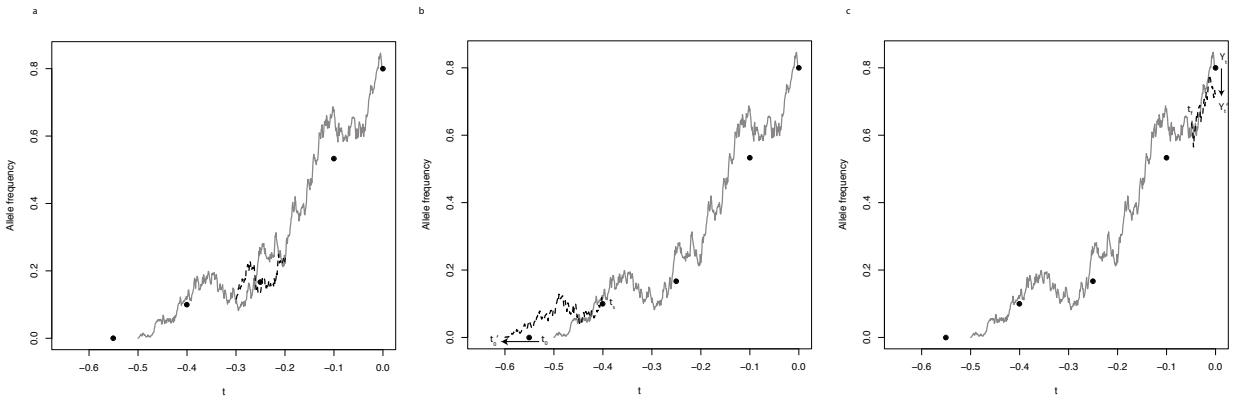

Figure 4.2: Illustration of path updates. Filled circles correspond to the same sample frequencies as in Figure 4.1. The solid gray line in each panel is the current allele frequency trajectory and the dashed black lines are the proposed updates. In panel a, an interior section of path is proposed between points s_1 and s_2. In panel b, a new allele age, t_0' is proposed and a new path is drawn between t_0' and t_s. In panel c, a new most recent allele frequency Y_t' is proposed and a new path is drawn between t_f and t.

4.3.1 Interior path updates

To update a section of the allele frequency, we first choose random times $s_1, s_2 \in (t_0, t_k)$, and then propose a new path from s_1 to s_2 while keeping the values $Y_{f(s_1)}$ and $Y_{f(s_2)}$ fixed (Figure **??**a). Such a path is called a bridge. Note that bridges must be sampled against the *transformed* time scale. The best bridges would be realizations of Wright-Fisher bridges themselves. However, sampling Wright-Fisher bridges is challenging (but see Schraiber et al. (2013)), so we instead opt to sample bridges from the Bessel(0) process. Sampling Bessel(0) bridges can be accomplished by first sampling Bessel(4) bridges (as described in Schraiber et al. (2013)) and then recognizing that a Bessel(4) process is the same as a Bessel(0) process conditioned to never hit 0 and hence has the same bridges. We denote by Y' the path with the proposed bridge spliced in between s_1 and s_2.

Because we are calculating path likelihoods relative to the distribution of Bessel(0) paths, the proposal probability is simply the inverse of the probability that a Bessel(0) process goes from $y_1 = Y_{f(s_1)}$ to $y_2 = Y_{f(s_2)}$ in time $\tau = f(s_2) - f(s_1)$,

$$\begin{aligned} q(Y'|Y) &= \frac{1}{\phi(Y_{f(s_1)}, Y_{f(s_2)}; s_2 - s_1)} \\ &= \frac{\tau}{y_1} \exp\left\{\frac{y_1^2 + y_2^2}{2\tau}\right\} \frac{1}{I_1\left(\frac{y_1 y_2}{\tau}\right)} \end{aligned}$$

where $I_1(\cdot)$ is the modified Bessel function of the first kind with index 1. Thus, we accept

the proposed update to the path with probability

$$\min\left\{1, \frac{L(D, Y'|\alpha, h, t_0)}{L(D, Y|\alpha, h, t_0)} \frac{q(Y|Y')}{q(Y'|Y)}\right\}$$

but note that $q(Y|Y') = q(Y'|Y)$ and that we only need to compute the likelihood ratio for the bit of path that changed between s_1 and s_2.

4.3.2 Allele age updates

Allele age updates proceed in two steps: first, a new allele age, t_0', is proposed, and then a new section of path is proposed starting from 0 at time t_0' and ending at $Y_{f(t_s)}$ where $s = \min\{i : c_i > 0\}$, i.e. the first sample time with a non-zero count of the derived allele (Figure **??**b). Denote by $p(t_0'|t_0)$ the proposal density for t_0' given t_0 and let $q(t_0', Y'|t_0, Y)$ be the proposal probability for the update to both the age and the path. Then the proposal ratio for this update is

$$\begin{aligned}\frac{q(t_0, Y|t_0', Y')}{q(t_0', Y'|t_0, Y)} &= \lim_{y_0 \downarrow 0} \frac{\phi(y_0, y, \tau')}{\phi(y_0, y, \tau)} \frac{p(t_0|t_0')}{p(t_0'|t_0)} \\ &= \exp\left\{-\frac{1}{2} y^2 \left(\frac{1}{\tau'} - \frac{1}{\tau}\right)\right\} \left(\frac{\tau}{\tau'}\right)^2 \frac{p(t_0|t_0')}{p(t_0'|t_0)}\end{aligned}$$

where $\tau = f(t_s) - f(t_0)$, $\tau' = f(t_s) - f(t_0')$ and $y = Y_{f(t_s)}$. This procedure can be rigorously justified by considering the entrance law of the Bessel(0) process. Note that the implicit prior on allele age is $\pi(t_0) = \rho(t_0)$ (Slatkin 2001). With this in hand, we accept a newly proposed allele age with probability

$$\min\left\{1, \frac{L(D, Y'|\alpha, h, t_0)}{L(D, Y|\alpha, h, t_0)} \frac{q(Y|Y')}{q(Y'|Y)} \frac{\rho(t_0')}{\rho(t_0)}\right\}$$

where is only necessary to compute the path likelihood $L(D, Y|\alpha, h, t_0)$ over the interval $[t_0, t_s]$ and the likelihood $L(D, Y'|\alpha, h, t_0')$ over the interval $[t_0', t_s]$.

4.3.3 Most recent allele frequency update

While the allele frequency at sample times $t_1, t_2, \ldots, t_{k-1}$ are updated implicitly by the interior path update, the allele frequency at t_k must be updated separately. We do this by first proposing a new allele frequency $Y'_{f(t_k)}$ and then proposing a new bridge from $Y_{f(t_f)}$ to $Y'_{f(t_k)}$ where $t_f \in (t_{k-1}, t_k)$ is a fixed time (Figure **??**c). If $p(Y'_{f(t_k)}|Y_{f(t_k)})$ is the proposal density for $Y'_{f(t_k)}$ given $Y_{f(t_k)}$, then we accept this update with probability

$$\min\left\{1, \frac{L(D, Y'|\alpha, h, t_0)}{L(D, Y|\alpha, h, t_0)} \frac{p(Y_{f(t_k)}|Y'_{f(t_k)})}{p(Y'_{f(t_k)}|Y_{f(t_k)})} \frac{\phi(Y_{f(t_f)}, Y'_{f(t_k)}; \tau)}{\phi(Y_{f(t_f)}, Y_{f(t_k)}; \tau)}\right\}$$

where $\tau = f(t_k) - f(t_s)$ and it is only necessary to compute the likelihood ratio between t_f and t_k.

4.3.4 Updates to α and h

Updates to α and h are conventional scalar parameter updates. Letting θ be either α or ρ, and $q(\theta'|\theta)$ be the proposal density for the new value of θ, we accept the new proposal with probability

$$\min\left\{1, \frac{L(D,Y|\theta',t_0)}{L(D,Y|\theta,t_0)}\frac{q(\theta|\theta')}{q(\theta|\theta')}\frac{\pi(\theta')}{\pi(\theta)}\right\}.$$

Here, it is necessary to compute the likelihood across the whole path. For α, we use a Cauchy(0,100) prior and for h we use a Cauchy(.5,.5) prior, indicating that our prior belief favors genic selection ($h = .5$).

4.4 Results

We apply our method to simulated data to assess its performance and then apply it to several real datasets from humans and horses.

4.4.1 Simulation performance

To both test the accuracy of our MCMC approach and to determine how different sampling strategies impact inference for different allele ages, we simulated data in which alleles arose at two different times: 0.3 time units ago and 0.7 time units ago. Additionally, we simulated $\alpha \in \{0, 10, 50\}$ and $h \in \{0, 0.5, 1, 2\}$. Setting time 0 as the present, we then sampled 20 chromosomes at each time point according to four, progressively more dense, schemes, as outlined in Table 4.2.

k	Additional sample times
3	-1.0, -0.5, 0
5	$-0.75 -0.25$
9	-0.875 -0.625 -0.375 -0.125
17	-0.9375, -0.8125, -0.6875, -0.5625, -0.4375, -0.3125, -0.1875, -0.0625

Table 4.2: Sampled schemes for simulations. For each k, the additional sample times augment those of the previous k.

Our MCMC analysis results in a full posterior distribution on parameter values, and we summarized the accuracy of the estimation procedure in two ways. First, we used the maximum *a posteriori* (MAP) value of a parameter as a point estimate and computed the difference between the MAP estimate and the true value of the parameter for each simulation. Because the distribution of MAP estimates can be highly skewed (Supplemental Figure 4.1), we measured the typical bias of the MAP by taking the mode of the sampling distribution, rather than the mean. Figure 4.3, top panel, shows the behavior of the bias when $h = 0.5$ and $t_0 = 0.3$ as the number of samples increases. The solid, dashed, and dotted lines correspond

to $\alpha = 0, 10$, and 50, respectively. As expected, as the number of data points increases, the bias in estimating α is reduced substantially. However, bias in h is relatively unaffected by increased sampling density; this makes sense as the signal for h is relatively subtle, especially for a relatively short trajectory. Additionally, bias in estimation of t_0 is relatively insensitive to increased sampling density, likely due to the fact that the bias is already quite small. Similar patterns can be seen in other parameter regimes (Supplementary Figures 2-8, top panels).

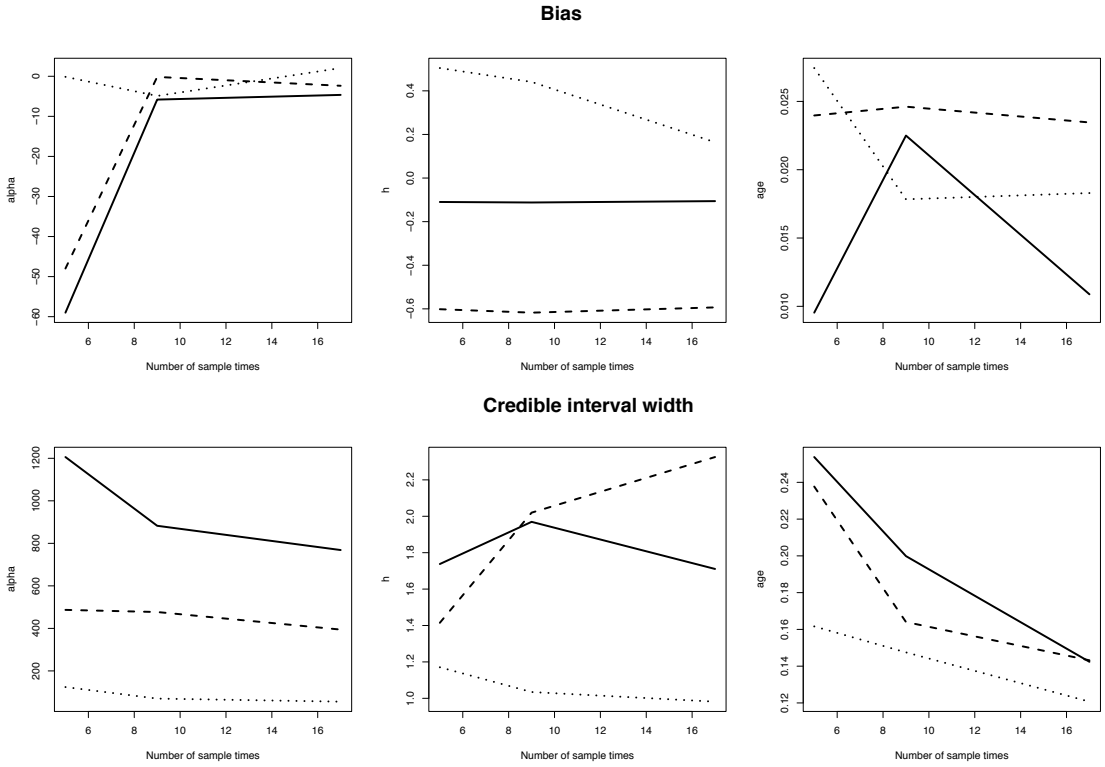

Figure 4.3: Bias of maximum *a posteriori* estimates and credible interval width. Top panel reports the modal bias of the sampling distribution of MAP estimates, bottom panel reports the modal credible interval width. Solid, dashed and dotted lines correspond to $\alpha = 0, 10$ and 50, respectively. Here, $h = 0.5$ and $t_0 = 0.3$.

We also measured the typical width of the 95% credible interval across our simulations. Again, because the sampling distribution of credible intervals is highly skewed, we assessed the typical credible interval by reporting the modal credible interval width. Figure 4.3, bottom panels shows the behavior of the credible interval for the same parameters as in the top panel. Again, the width of the credible interval is typically reduced as the the sample size increases. Noteworthy is the extremely large credible intervals associated with estimation of α from neutral trajectories. This is because many neutral trajectories stay very close to the boundaries, at which point genetic drift dominates and there is a significant amount of

uncertainty in the estimate of the selection coefficient. Again, similar patterns can be seen for other parameter regimes (Supplementary Figures 2-8, bottom panels).

4.4.2 Application to ancient DNA

We began by reanalyzing the ASIP and MC1R data from Ludwig et al. (2009). We assumed that $N_0 = 2500$, a generation time of 5 years and that population size has been constant through time. Table 4.3 shows the sample configurations and sampling times corresponding to each locus.

Sample time (years BCE)	20,000	13,100	3,700	2,800	1,100	500
Sample time (diffusion units)	0.8	0.524	0.148	0.112	0.044	0.020
Sample size	10	22	20	20	36	38
Count of ASIP alleles	0	1	15	12	15	18
Count of MC1R alleles	0	0	1	6	13	24

Table 4.3: Sample information for horse data. Diffusion time units are calculated assuming $N_0 = 2500$ and a generation time of 5 years.

For the ASIP locus, we find that the most likely selective mechanism is overdominance ($\hat{h} = 2.02$, $\hat{\alpha} = 10.23$; Figure 4.4B,C), in agreement with the conclusion reached by Steinrücken et al. (2013). This inference makes sense in light of the allele frequency trajectory inferred ASIP: the allele quickly rises to intermediate frequency and then stays at an approximately constant frequency, a hallmark of overdominance (Figure 4.4A). An interesting feature of the posterior distribution of h is that there is a second mode for $h < 0$; this corresponds to the fact that when $h < 0$ and $\alpha < 0$ is also overdominant. The interplay between h and α can clearly be seen in Supplemental Figure 9, which shows the joint posterior of α and h. The posterior is concentrated in two places; in one case, $h > 1$ and $\alpha > 0$ and in the other case, $\alpha < 0$ and $h < 0$. We also find that the allele almost certainly arose more recently than the most ancient time point, at which time zero copies of the derived allele were found ($\hat{t}_0 = -0.53$, approximately $13,700$ years BCE; Figure 4.4A), concordant with the analysis of Malaspinas et al. (2012).

The results of analysis of the MC1R locus are again concordant with previous analyses; the most likely selective regime is positive selection ($\hat{h} = .63$, $\hat{\alpha} = 26.27$; Figure 4.5A,B). However, there is again substantial evidence for overdominance and the same pattern can be seen in the joint posterior of α and h for MC1R as could be seen for ASIP (Supplementary Figure 10). Although the inferred allele frequency trajectory for MC1R seems most consistent with positive selection (Figure 4.5A), this result is again concordant with the inference of Steinrücken et al. (2013) who found that MC1R may have evolved under a regime of overdominance. Similar to ASIP, we infer that the MC1R allele is almost certainly much younger than either of the two time points at which it appeared with 0 allele frequency ($\hat{t}_0 = -0.16$, approximately $4,400$ years BCE; Figure 4.5A).

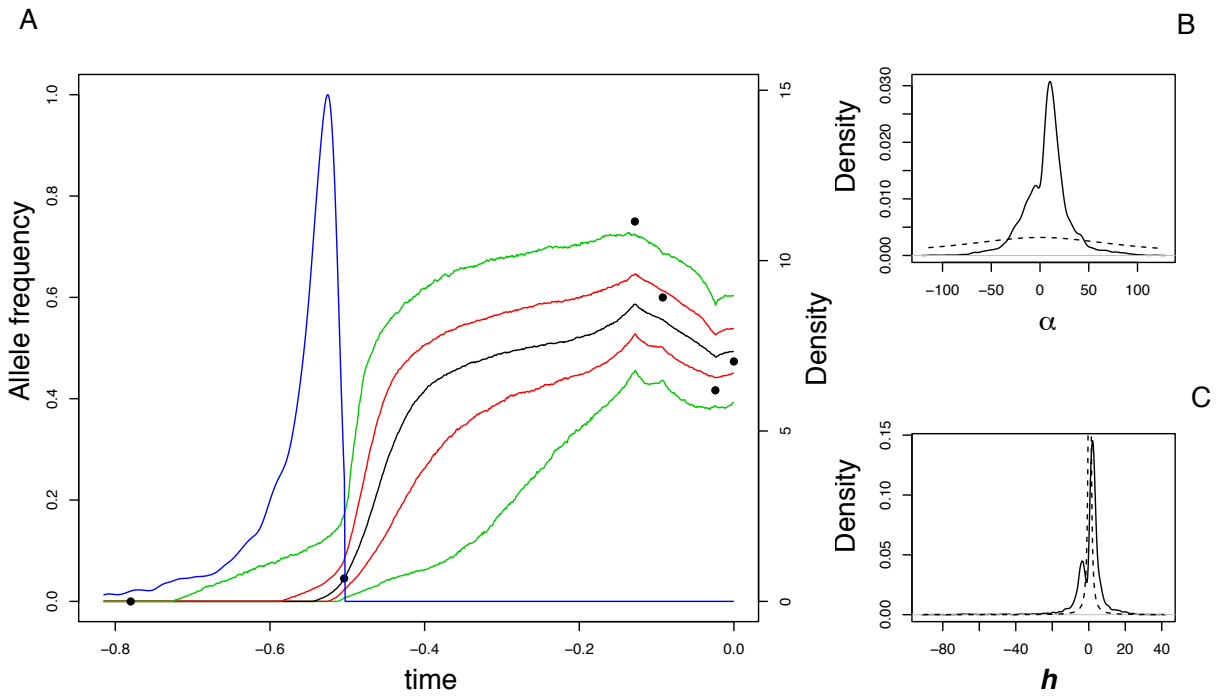

Figure 4.4: Summary of results for the ASIP locus in horses. Panel A shows the posterior distribution of paths as well as the posterior distribution of allele age. Filled circles are the sample allele frequencies, while the solid black, red and green lines show the median, interquartile, and 95% credible intervals of the path, respectively. The blue curve shows the posterior distribution of the allele age. Time is measured in diffusion units relative to the most recent sample (so that 0.0 corresponds to 500 years BCE). Panel B and C show the posterior distribution of α and h, respectively. In both, solid lines are the posterior while dashed lines show the prior.

We next analyzed the Lactase persistence allele (LCT) in humans. Because no previous study has formally analyzed the frequency dynamics of LCT, we gathered data from two sources. Plantinga et al. (2012) obtained LCT sequence from 26 individuals in Basque Country ranging from 5000 to 4500 years before present. Out of the 52 haplotypes sampled, 12 of them carried the lactase persistence allele. As our source for recent data, Bersaglieri et al. (2004) obtained LCT sequence for 45 French Basque haplotypes, and found that 32 of them carried the lactase persistence allele. To account for recent demography of Europe, we used the demographic history inferred by Tennessen et al. (2012). This model includes two epochs of recent exponential growth along with a history of ancestral bottlenecks.

Our analysis of LCT shows that lactase persistence has evolved under extremely strong positive selection in Basques ($\hat{h} = .46$, $\hat{\alpha} = 70$; Figure 4.6B,C), consistent with other analyses of LCT that do not include ancient DNA. This is unsurprising given the posterior distribution on allele frequency trajectories. A striking feature of our inference is the multimodal distribution of allele age (Figure 4.6A). The reason for this can be seen in Supplementary

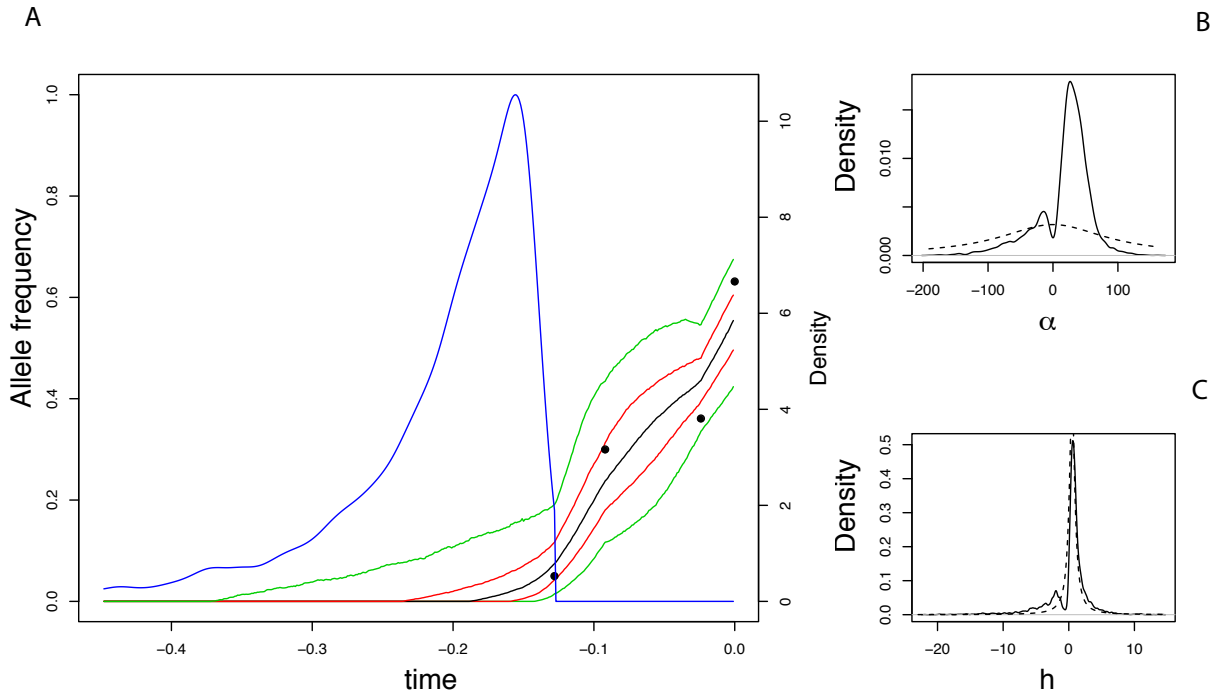

Figure 4.5: Summary of results for the MC1R locus in horses. Panels are as in Figure 4.4.

Figure 11, which shows the Tennessen et al. (2012) population size history along with posterior of allele age. The first mode, around $t = -0.1$ corresponds to the allele arising during a population bottleneck, while the second mode, around $t = -0.2$ is during a period of much larger population size. Thus, the fact that the allele is much more likely to have arisen during a time of a larger population size makes it more probable that the allele arose more anciently than would have been inferred under a model of constant population size.

4.5 Discussion

Using DNA from ancient specimens, we have obtained a number of insights into evolutionary processes that were previously inaccessible. One of the most interesting aspects of ancient DNA is that it can provide a *temporal* component to evolution that has long been impossible to study. In particular, instead of making inferences about the allele frequencies, we can directly measure these quantities. To take advantage of this new data, we developed a novel Bayesian method for inferring the intensity and direction of natural selection from allele frequency time series. In order to circumvent the difficulties inherent in calculating the transition probabilities under the standard Wright-Fisher process of selection and drift, we used a data augmentation approach in which we learn the posterior distribution on allele frequency paths. Doing this not only allows us to efficiently calculate likelihoods, but

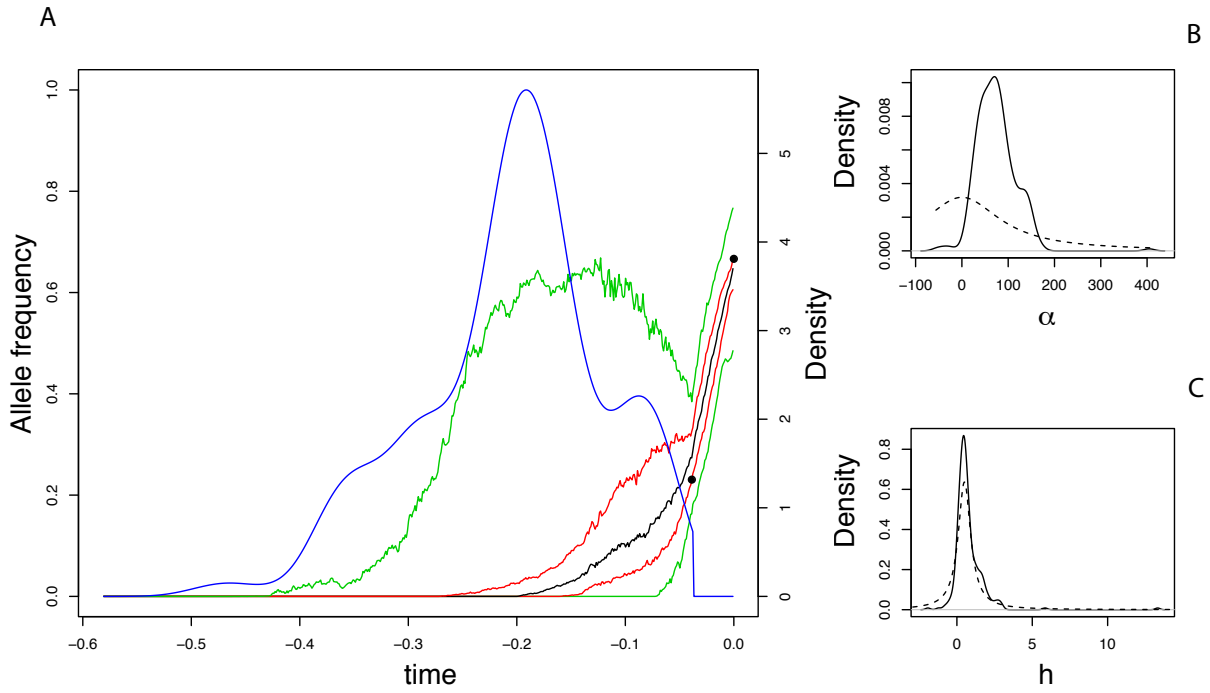

Figure 4.6: Summary of results for the MC1R locus in horses. Panels are as in Figure 4.4. Here, time is measured in diffusion time units assuming a generation time of 25 years and $N_0 = 7,310$.

provides an unprecedented glimpse at the historical allele frequency dynamics.

The key innovation of our method is to apply high-frequency path augmentation methods (Roberts & Stramer 2001) to analyze genetic time series. The logic of the method is similar to the logic of a path integral, in which we average over all possible allele frequency trajectories that are consistent with the data (Schraiber 2014). By choosing a suitable probability distribution against which to compute likelihood ratios, we were able to adapt these methods to infer the age of alleles and properly account for variable population sizes through time. Moreover, because of the computational advantages of the path augmentation approach, we were able to infer a model of general diploid selection. To our knowledge, ours is the first work that can estimate both allele age and general diploid selection while accounting for demography.

Using simulations, we showed that our method performs well for strong selection and densely sampled time series. However, when selection is weak, even densely sampled time series result in very large credible intervals. This is unsurprising in light of the work of Watterson (1979), who showed that even knowledge of the full trajectory results in very flat likelihood surfaces when selection is not strong. This is because for weak selection, the trajectory is extremely stochastic and it is difficult to disentangle the effects of drift and selection.

We then applied our method to real data from horses and humans. In the horses, we

recapitulated the results of Steinrücken et al. (2013), suggesting that the ASIP locus has evolved under overdominant selection and that the MC1R locus evolved while experiencing positive selection (although it is possible that MC1R, too, evolved under overdominance). The prevalence of overdominance in these data may at least partially reflect the suggestion of Sellis et al. (2011), that adaptation in diploids is expected to proceed through periods of heterozygote advantage. In humans, we analyzed the lactase persistence allele in the Basque population. Consistent with expectations, we found a signature of extremely strong positive selection affecting lactase persistence. Interestingly, because the allele frequency of Lactase was low during a population bottleneck in Europeans, it is possible that the lactase allele is significantly older than the onset of agriculture in Europe, suggesting that it may have arisen from standing variation.

One key limitation of this method is that it assumes that the aDNA samples all come from the same, continuous population. If there is in fact a discontinuity in the populations from which alleles have been sampled, this could cause rapid allele frequency change and create spurious signals of natural selection. Several methods have been devised to test this hypothesis (Sjödin et al. 2014), and one possibility would be to apply these methods to putatively neutral loci sampled from the same individuals, thus determining which samples form a continuous population. Alternatively, if our method is applied to a number of loci throughout the genome and an extremely large portion of the genome is determined to be evolving under selection, this could be evidence for model misspecification and suggest that the samples do not come from a continuous population.

An advantage of the method that we introduced is that it may be possible to extend it to incorporate information from linked neutral diversity. In general, computing the likelihood of neutral diversity linked to a selected site is difficult and many have used Monte Carlo simulation and importance sampling (Slatkin 2001; Coop & Griffiths 2004; Chen & Slatkin 2013). These approaches average over allele frequency trajectories in much same way as our method; however, each trajectory is drawn completely independently of the previous trajectories. Using a Markov chain Monte Carlo approach, as we do here, has the potential to ensure that only trajectories with a high posterior probability are explored and hence greatly increase the efficiency of such approaches.

4.6 Appendix

4.6.1 The likelihood of the data and the path

Using equation (4.7), the likelihood of the path can be written

$$\exp\left\{\int_{\tau_0}^{\tau_k}\left(\mu_1(Y_r,r)-\mu_2(Y_r)\right)dY_r-\frac{1}{2}\int_{\tau_0}^{\tau_k}\left(\mu_1^2(Y_r,r)-\mu_2^2(Y_r)\right)dr\right\} \tag{4.10}$$

where

$$\mu_1(y,\tau)=\frac{1}{4}\left(\alpha\rho(f^{-1}(\tau))\sin(Y_\tau)(1+(2h-1)\cos(Y_\tau))-2\cot(Y_\tau)\right)$$

is the infinitesimal mean of the transformed Wright-Fisher process and

$$\mu_2(y) = -\frac{1}{2y}$$

is the infinitesimal mean of the Bessel(0) process. However, as shown by Sermaidis et al. (2012), attempting to approximate the Ito integral in (4.10) using a finite representation of the path can lead to biased estimates of the posterior distribution. Instead, consider the potential functions

$$\begin{aligned} H_1(y,\tau) &= \int^y \mu_1(\xi,\tau)d\xi \\ &= -\frac{1}{8}\left(\alpha\rho(f^{-1}(\tau))\cos(y)(2+(2h-1)\cos(y)) + 4\log(\sin(y))\right) \end{aligned}$$

and

$$\begin{aligned} H_2(y) &= \int^y \mu_2(\xi,\tau)d\xi \\ &= -\frac{\log(y)}{2}. \end{aligned}$$

If we assume that ρ is continuous (not merely right continuous with left limits) then Ito's lemma shows that we can write

$$\begin{aligned} \int_{\tau_0}^{\tau_k} (\mu_1(Y_r,r) - \mu_2(Y_r))\, dY_r = H_1(Y_{\tau_k},\tau_k) - H_2(Y_{\tau_k}) - (H_1(Y_{\tau_0},\tau_0) - H_2(Y_{\tau_0})) \\ - \int_{\tau_0}^{\tau_k} \left(\frac{\partial H_1}{\partial \tau}(Y_r,r) - \frac{\partial H_2}{\partial \tau}(Y_r)\right) dr \\ - \int_{\tau_0}^{\tau_k} \left(\frac{\partial^2 H_1}{\partial y^2}(Y_r,r) - \frac{\partial^2 H_2}{\partial y^2}(Y_r)\right) dr. \end{aligned}$$

To generalize this to the case where ρ is right continuous with left limits, write

$$\int_{\tau_0}^{\tau_k} (\mu_1(Y_r,r) - \mu_2(Y_r))\, dY_r = I_0 + \sum_{i=m}^{K} I_i$$

where m and K are defined in the main text,

$$I_0 = \lim_{\tau \uparrow f(d_m)} \int_{\tau_0}^{\tau} (\mu_1(Y_r,r) - \mu_2(Y_r))\, dY_r,$$

for $m < i < K$,

$$I_i = \lim_{\tau \uparrow f(d_{i+1})} \int_{f(d_i)}^{\tau} (\mu_1(Y_r,r) - \mu_2(Y_r))\, dY_r,$$

and

$$I_K = \lim_{\tau \uparrow \tau_k} \int_{f(d_K)}^{\tau} (\mu_1(Y_r, r) - \mu_2(Y_r))\, dY_r.$$

Ito's lemma can then be applied to each segment in turn. Following the conversion of the Ito integrals into ordinary Lebesgue integrals, making the substitution $s = f^{-1}(r)$ results in the path likelihood from (4.9).

4.7 Supplementary Figures

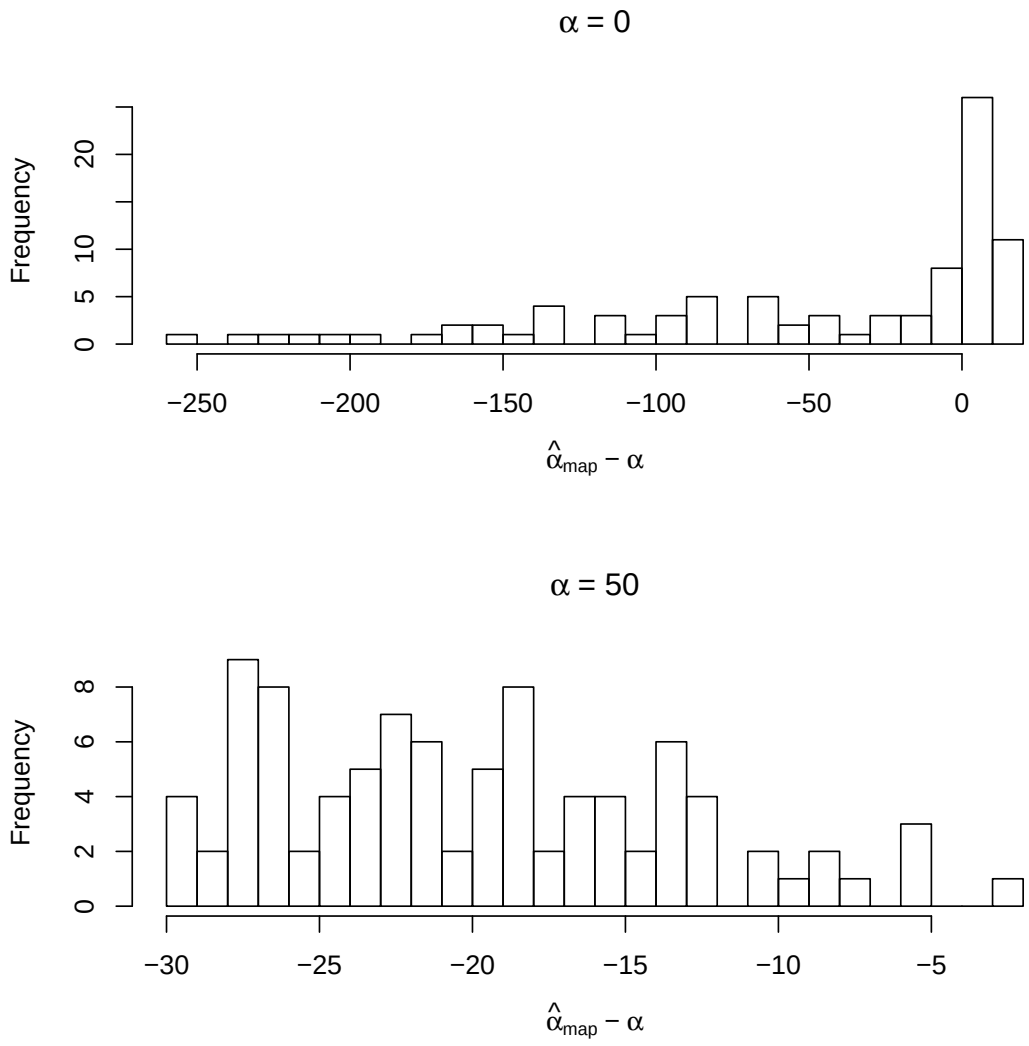

Figure 4.7: **Supplementary Figure 1.** Distribution of maximum *a posteriorio* estimates of α for two cases. In the top panel, the true $\alpha = 0$ and there is a substantial bias toward negative $\hat{\alpha}$. In the bottom panel, the true $\alpha = 50$ and the bias is largely removed, although there still tends to be underestimation of α.

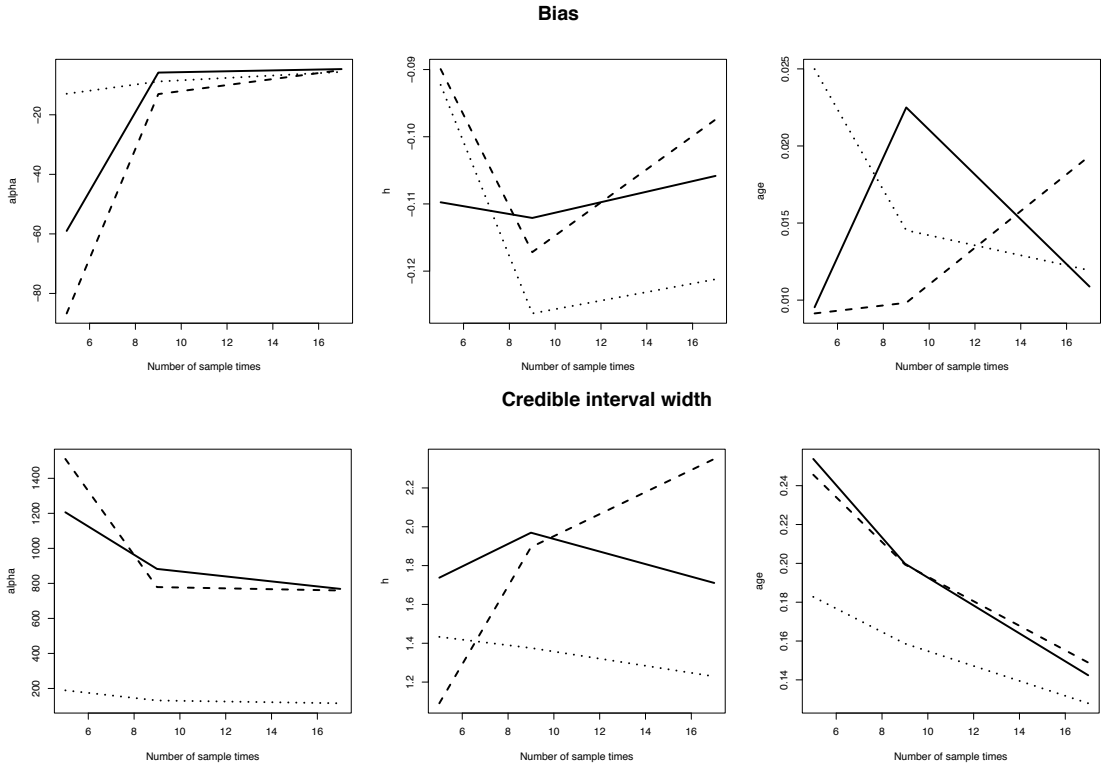

Figure 4.8: **Supplementary Figure 2.** Bias of maximum *a posteriori* estimates and credible interval width. Top panel reports the modal bias of the sampling distribution of MAP estimates, bottom panel reports the modal credible interval width. Solid, dashed and dotted lines correspond to $\alpha = 0, 10$ and 50, respectively. Here, $h = 0$ and $t_0 = 0.3$.

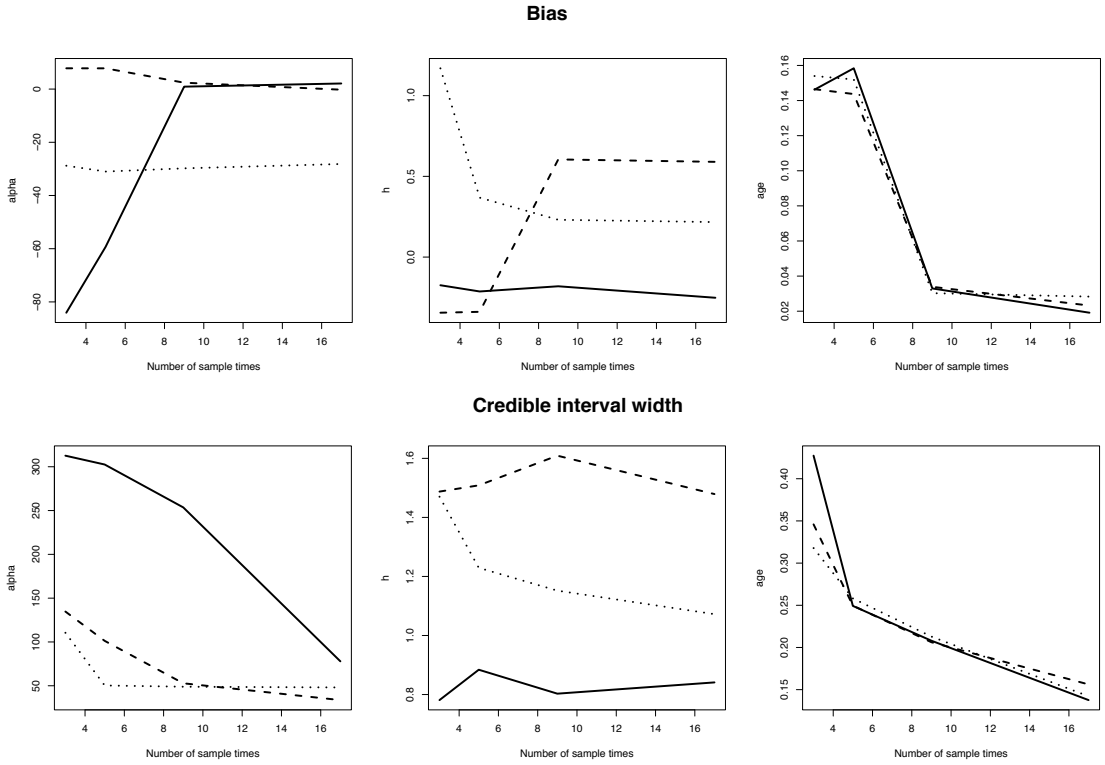

Figure 4.9: **Supplementary Figure 3.** Bias of maximum *a posteriori* estimates and credible interval width. Top panel reports the modal bias of the sampling distribution of MAP estimates, bottom panel reports the modal credible interval width. Solid, dashed and dotted lines correspond to $\alpha = 0, 10$ and 50, respectively. Here, $h = 0$ and $t_0 = 0.7$.

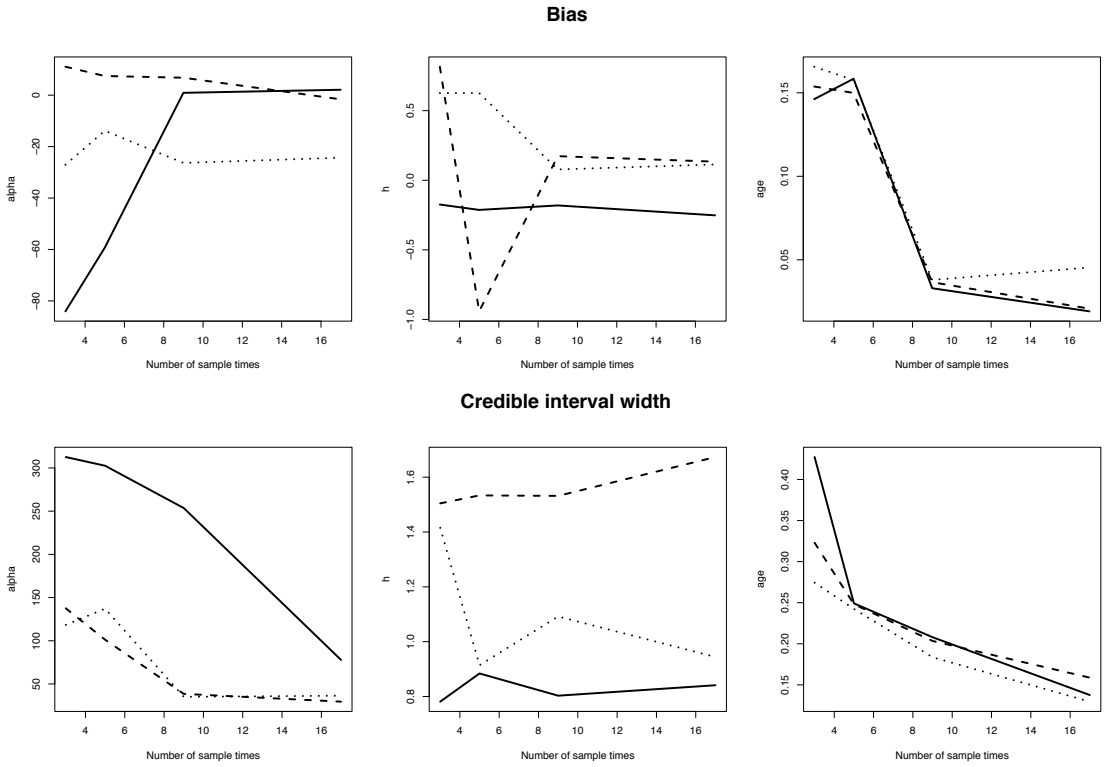

Figure 4.10: **Supplementary Figure 4.** Bias of maximum *a posteriori* estimates and credible interval width. Top panel reports the modal bias of the sampling distribution of MAP estimates, bottom panel reports the modal credible interval width. Solid, dashed and dotted lines correspond to $\alpha = 0, 10$ and 50, respectively. Here, $h = 0.5$ and $t_0 = 0.7$.

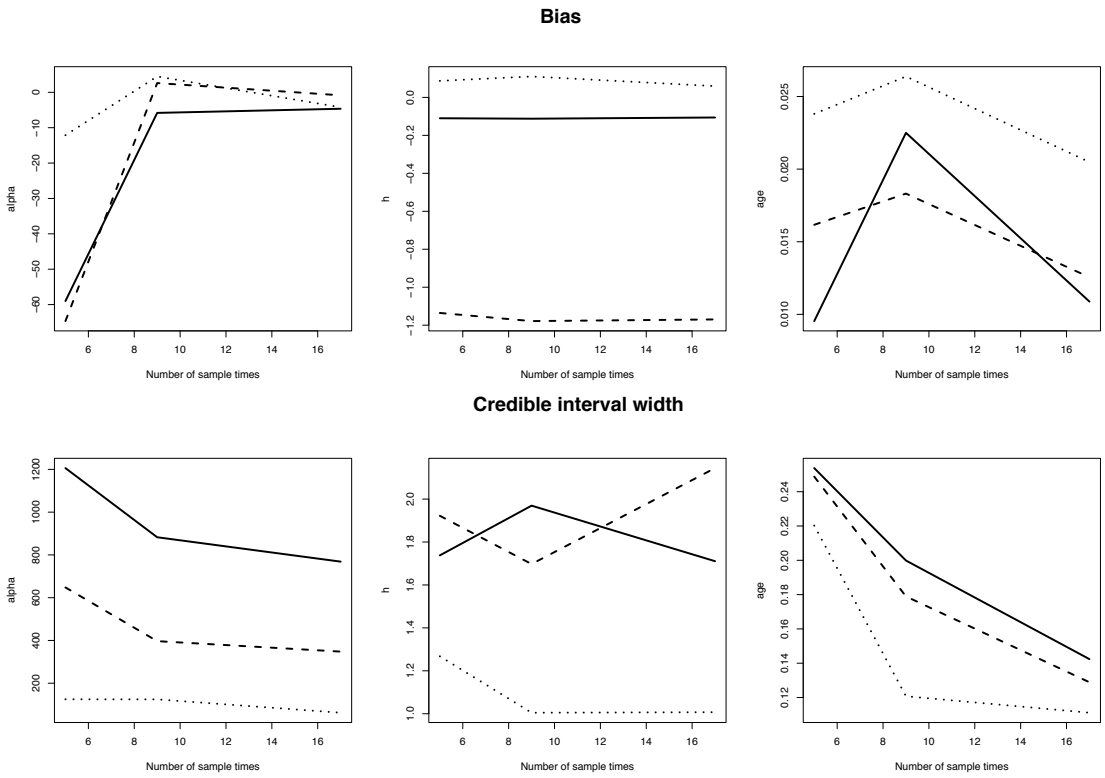

Figure 4.11: **Supplementary Figure 5.** Bias of maximum *a posteriori* estimates and credible interval width. Top panel reports the modal bias of the sampling distribution of MAP estimates, bottom panel reports the modal credible interval width. Solid, dashed and dotted lines correspond to $\alpha = 0, 10$ and 50, respectively. Here, $h = 1$ and $t_0 = 0.3$.

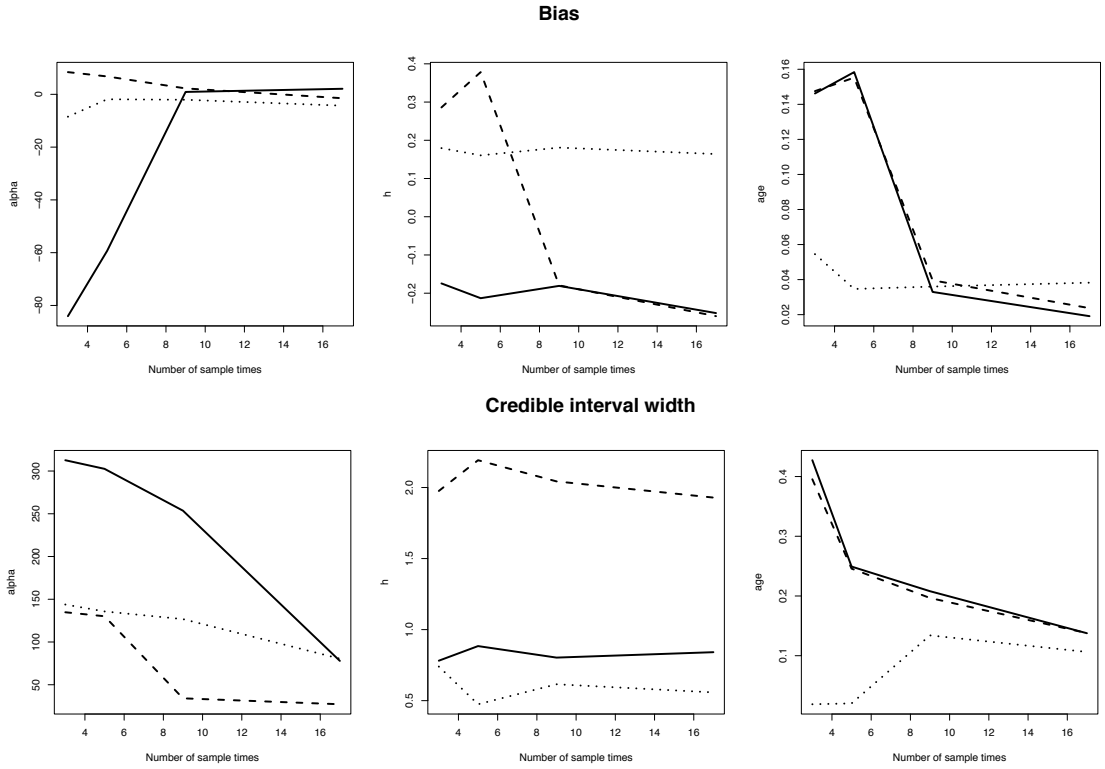

Figure 4.12: **Supplementary Figure 6.** Bias of maximum *a posteriori* estimates and credible interval width. Top panel reports the modal bias of the sampling distribution of MAP estimates, bottom panel reports the modal credible interval width. Solid, dashed and dotted lines correspond to $\alpha = 0, 10$ and 50, respectively. Here, $h = 1$ and $t_0 = 0.7$.

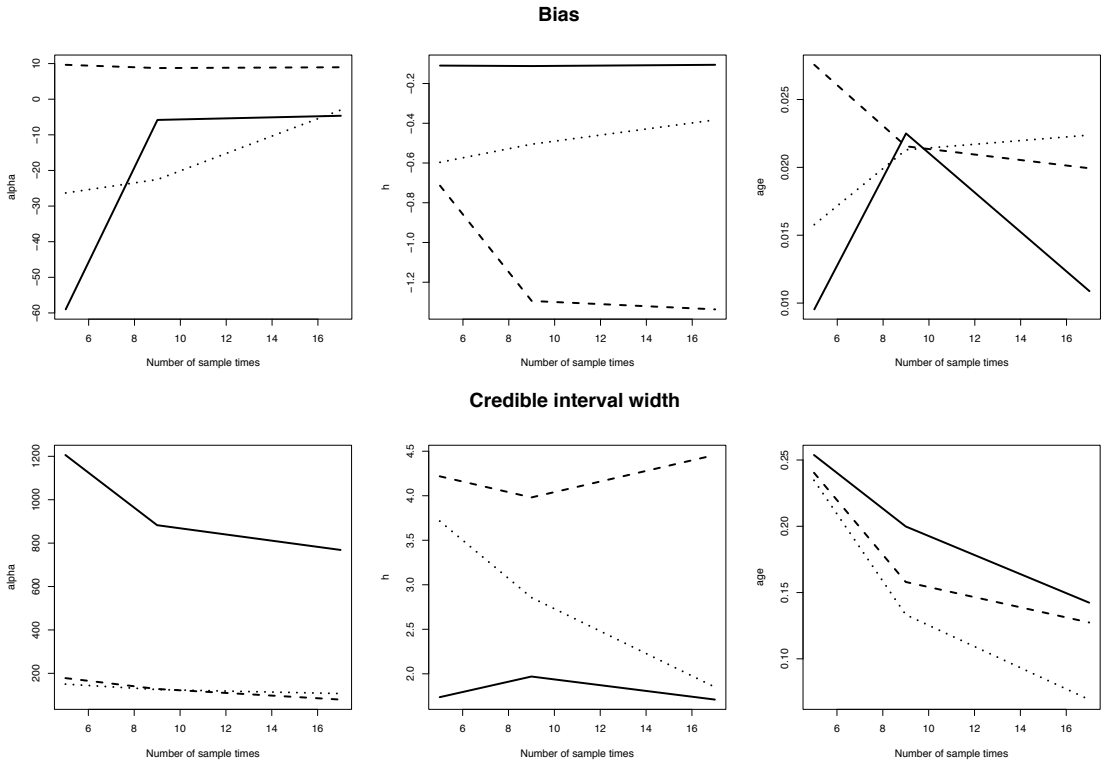

Figure 4.13: **Supplementary Figure 7.** Bias of maximum *a posteriori* estimates and credible interval width. Top panel reports the modal bias of the sampling distribution of MAP estimates, bottom panel reports the modal credible interval width. Solid, dashed and dotted lines correspond to $\alpha = 0, 10$ and 50, respectively. Here, $h = 2$ and $t_0 = 0.3$.

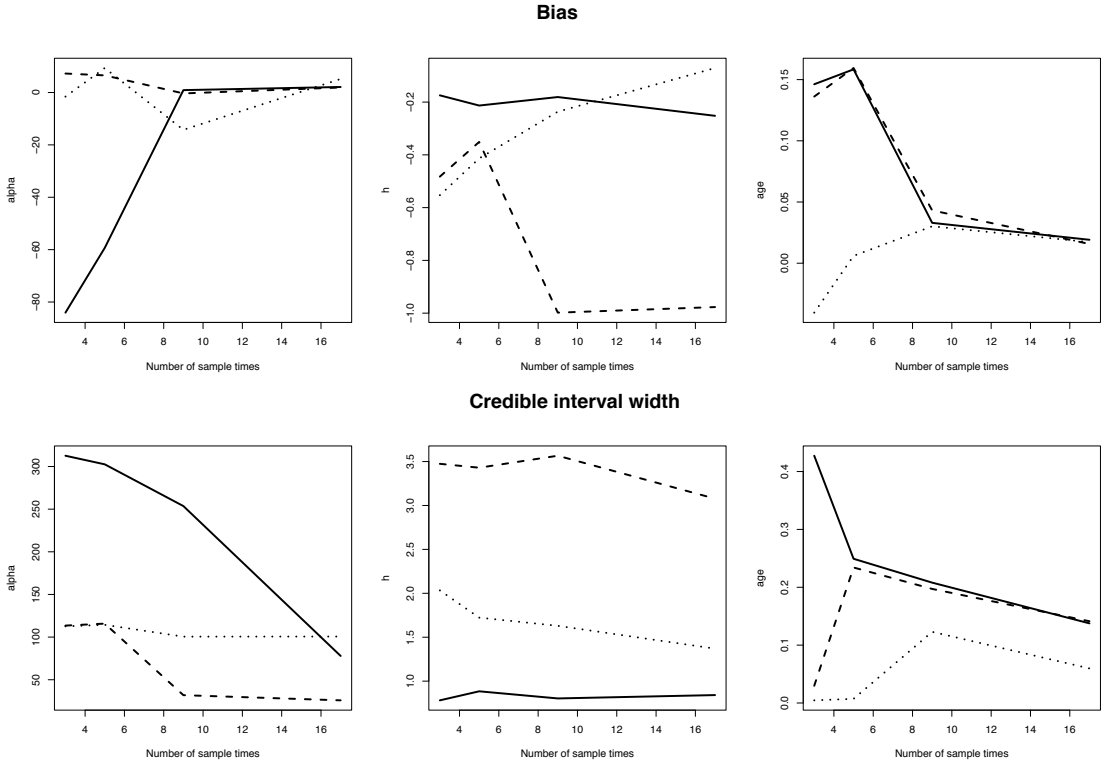

Figure 4.14: **Supplementary Figure 8.** Bias of maximum *a posteriori* estimates and credible interval width. Top panel reports the modal bias of the sampling distribution of MAP estimates, bottom panel reports the modal credible interval width. Solid, dashed and dotted lines correspond to $\alpha = 0, 10$ and 50, respectively. Here, $h = 2$ and $t_0 = 0.7$.

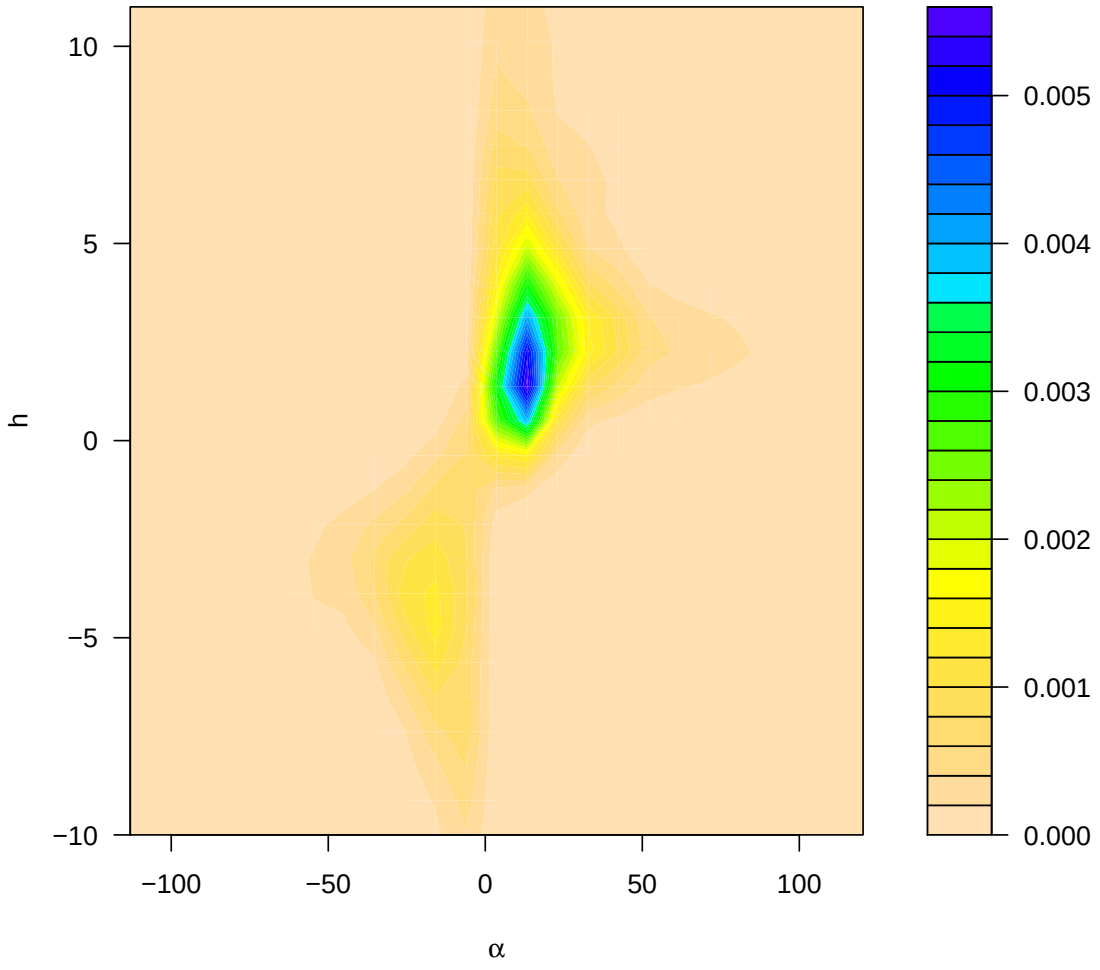

Figure 4.15: **Supplementary Figure 9.** Joint posterior density of α and h for the ASIP locus in horses. A filled contour plot with the x-axis representing α and the y-axis representing h. Regions of highest posterior density are shown in blue.

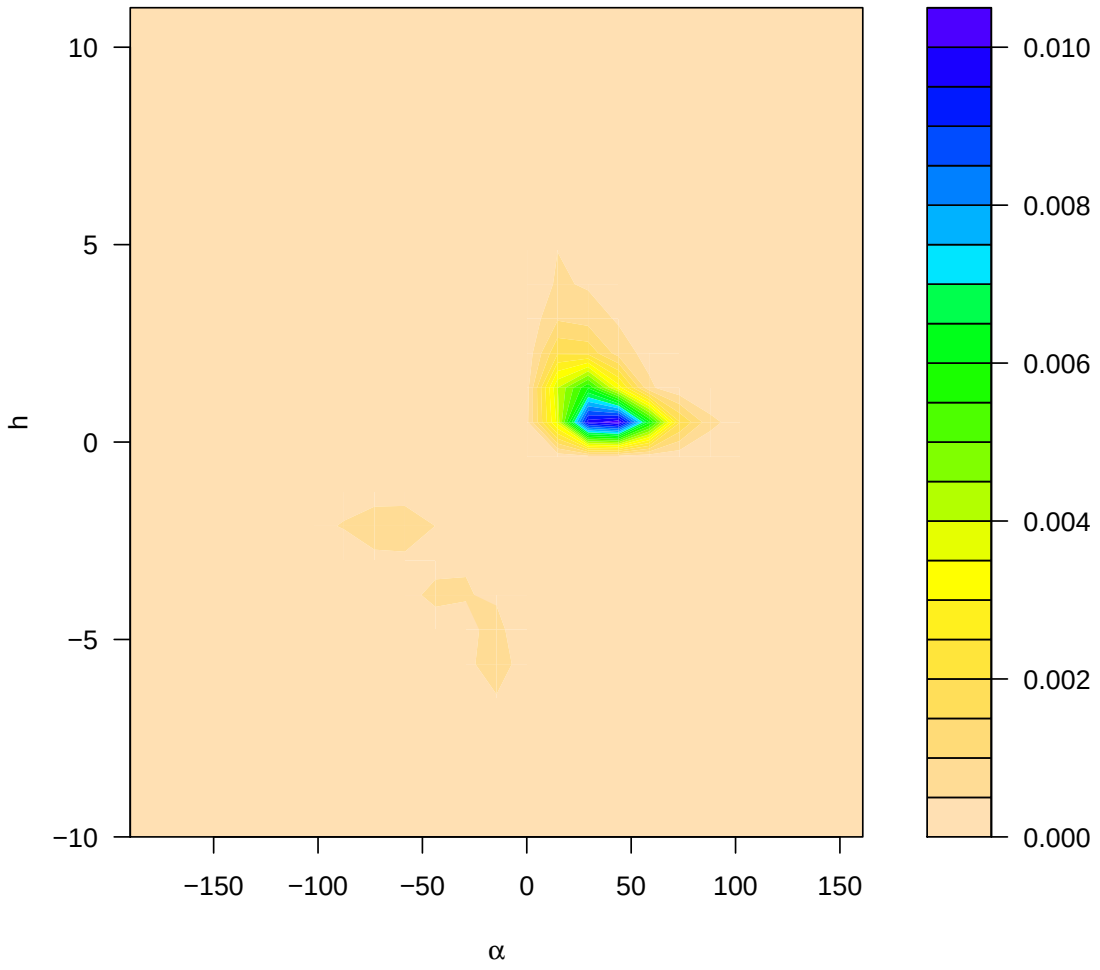

Figure 4.16: **Supplementary Figure 10.** Joint posterior density of α and h for the MC1R locus in horses. A filled contour plot with the x-axis representing α and the y-axis representing h. Regions of highest posterior density are shown in blue.

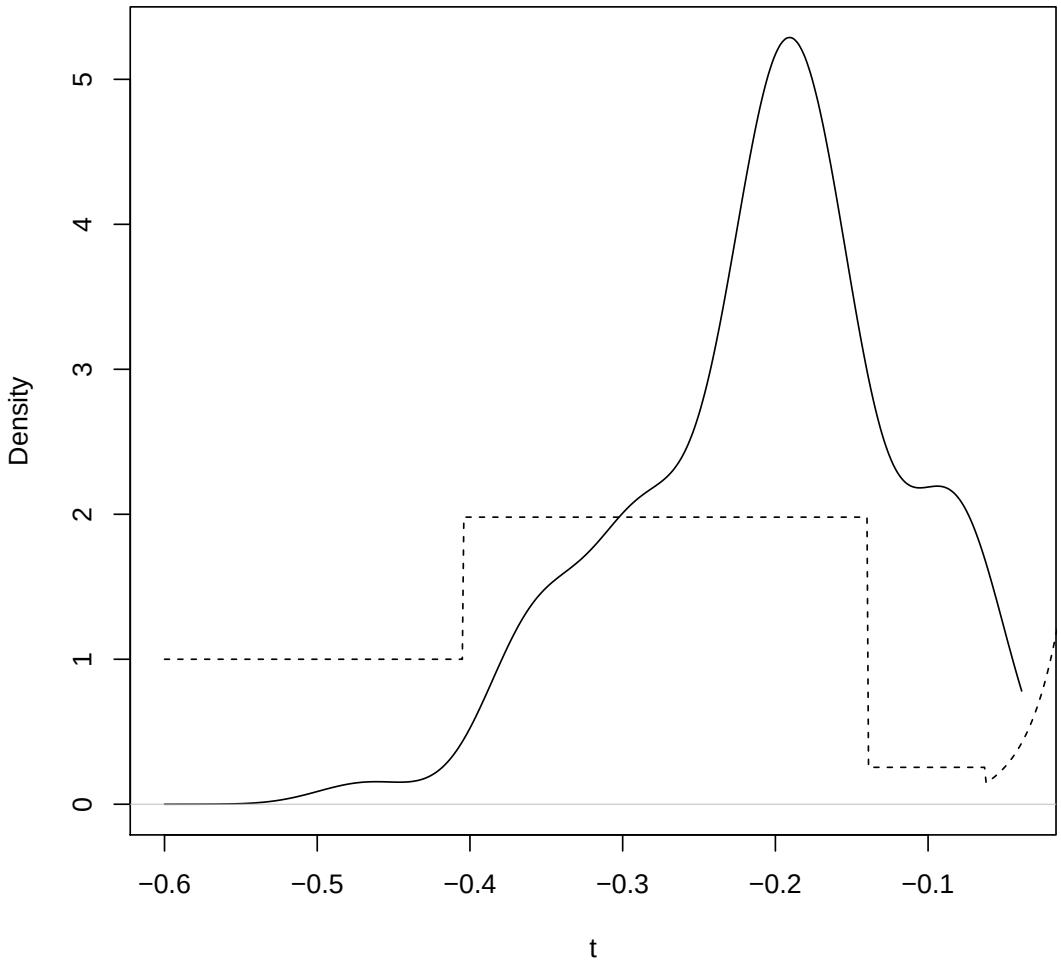

Figure 4.17: **Supplementary Figure 11.** Posterior of allele age at LCT and the demographic model. The solid lines shows the posterior for the allele age at LCT (the same as the blue solid line in Figure 4.6). The dashed line shows the assumed demographic model.

Bibliography

Baibuz, V., Zitserman, V. Y., & Drozdov, A. 1984, Physica A: Statistical Mechanics and its Applications, 127, 173
Bersaglieri, T., Sabeti, P. C., Patterson, N., et al. 2004, The American Journal of Human Genetics, 74, 1111
Beskos, A., & Roberts, G. O. 2005, Ann. Appl. Probab., 15, 2422
Bogachev, V. I. 2007, Measure theory, Vol. 1 (Springer)
Bollback, J. P., York, T. L., & Nielsen, R. 2008a, Genetics, 179, 497
—. 2008b, Genetics, 179, 497
Boyko, A. R., Williamson, S. H., Indap, A. R., et al. 2008, PLoS genetics, 4, e1000083
Burger, J., Kirchner, M., Bramanti, B., Haak, W., & Thomas, M. G. 2007, Proceedings of the National Academy of Sciences, 104, 3736
Chen, H., & Slatkin, M. 2013, G3: Genes| Genomes| Genetics, 3, 1429
Chorin, A. J., & Hald, O. H. 2006, Stochastic tools in mathematics and science, Vol. 1 (Springer)
Coop, G., & Griffiths, R. C. 2004, Theoretical population biology, 66, 219
Dawson, D. A., & Feng, S. 2001, Stochastic processes and their applications, 92, 131
Dawson, D. A., & Hochberg, K. J. 1982, The Annals of Probability, 554
Dirac, P. A. 1933, Physikalische Zeitschrift der Sowjetunion, 3, 64
Donnelly, P., & Kurtz, T. G. 1996, The Annals of Probability, 24, 698
—. 1999, Annals of Applied Probability, 1091
Drummond, A. J., Rambaut, A., Shapiro, B., & Pybus, O. G. 2005, Molecular biology and evolution, 22, 1185
Dürr, D., & Bach, A. 1978, Communications in Mathematical Physics, 60, 153
Etheridge, A., Pfaffelhuber, P., & Wakolbinger, A. 2006, Ann. Appl. Probab., 16, 685
Ethier, S., & Kurtz, T. 1987, in Stochastic methods in biology (Springer), 72
Ethier, S. N., & Griffiths, R. 1993, The Annals of Probability, 1571
Ethier, S. N., & Kurtz, T. G. 1993, SIAM Journal on Control and Optimization, 31, 345
Ewens, W. J. 1972, Theoretical population biology, 3, 87
—. 2004, Mathematical population genetics: I. Theoretical introduction, Vol. 27 (Springer)
Ewing, G., & Hermisson, J. 2010, Bioinformatics (Oxford, England), 26, 2064
Fay, J. C., & Wu, C.-I. 2000, Genetics, 155, 1405
Feder, A., Kryazhimskiy, S., & Plotkin, J. B. 2013, arXiv preprint arXiv:1302.0452

Feller, W. 1951, in Proc. Second Berkeley Symp. Math. Statist. Prob, Vol. 227, 246
Feynman, R. P. 1948, Reviews of Modern Physics, 20, 367
Feynman, R. P., & Hibbs, A. R. 2012, Quantum mechanics and path integrals: Emended edition (DoverPublications. com)
Fisher, R. 1922a, Proceedings of the Royal Society of Edinburgh, 42, 321
Fisher, R. A. 1922b, Proceedings of the royal society of Edinburgh, 42, 321
Fleming, W. H., & Viot, M. 1979, Indiana Univ. Math. J, 28, 817
Fuchs, C. 2013, Inference for Diffusion Processes: With Applications in Life Sciences (Springer)
Girsanov, I. 1960, Theory of Probability & Its Applications, 5, 285
Golightly, A., & Wilkinson, D. J. 2005, Biometrics, 61, 781
—. 2008, Computational Statistics & Data Analysis, 52, 1674
Graham, R. 1977, Zeitschrift für Physik B Condensed Matter, 26, 281
Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. 2009, PLoS genetics, 5, e1000695
Haldane, J. B. S. 1927, Mathematical Proceedings of the Cambridge Philosophical Society, 23, 838
Harlow, D. 2009, arXiv preprint arXiv:0905.2466
Harris, K., & Nielsen, R. 2013, PLoS genetics, 9, e1003521
Hudson, R. R., & Kaplan, N. L. 1988, Genetics, 120, 831
Ikeda, N., & Watanabe, S. 1989, North-Holland Mathematical Library, Vol. 24, Stochastic differential equations and diffusion processes, 2nd edn. (Amsterdam: North-Holland Publishing Co.), xvi+555
Itô, K. 1944, Proceedings of the Japan Academy, Series A, Mathematical Sciences, 20, 519
Jenkins, P. A. 2013, arXiv preprint arXiv:1311.5777
Kac, M. 1949, Teans. Amer. Math. Soc., 65, 1
Kaplan, N. L., Hudson, R. R., & Langley, C. H. 1989, Genetics, 123, 887
Kawecki, T. J., Lenski, R. E., Ebert, D., et al. 2012, Trends in ecology & evolution
Kimura, M. 1955a, Proceedings of the National Academy of Sciences of the United States of America, 41, 144
Kimura, M. 1955b, in Cold Spring Harbor Symposia on Quantitative Biology, Vol. 20, Cold Spring Harbor Laboratory Press, 33
—. 1957, The Annals of Mathematical Statistics, 882
—. 1965, Proceedings of the National Academy of Sciences of the United States of America, 54, 731
—. 1984, The neutral theory of molecular evolution (Cambridge University Press)
Kingman, J. F. 1982, Stochastic processes and their applications, 13, 235
Krone, S. M., & Neuhauser, C. 1997, Theoretical population biology, 51, 210
Kuhner, M. K., Yamato, J., & Felsenstein, J. 1995, Genetics, 140, 1421
—. 1998, Genetics, 149, 429
Lacan, M., Keyser, C., Ricaut, F.-X., et al. 2011, Proceedings of the National Academy of Sciences, 108, 9788

Li, H., & Durbin, R. 2011, Nature, 475, 493
Ludwig, A., Pruvost, M., Reissmann, M., et al. 2009, Science, 324, 485
Malaspinas, A.-S., Malaspinas, O., Evans, S. N., & Slatkin, M. 2012, Genetics, 192, 599
Malmström, H., Linderholm, A., Lidén, K., et al. 2010, BMC evolutionary biology, 10, 89
Mardia, K. V., Kent, J. T., & Bibby, J. M. 1979, Multivariate analysis (London: Academic Press [Harcourt Brace Jovanovich Publishers]), xv+521, probability and Mathematical Statistics: A Series of Monographs and Textbooks
Maruyama, T. 1974, Genetical Research, 23, 137
Mathieson, I., & McVean, G. 2013, Genetics, 193, 973
Mustonen, V., & Lässig, M. 2010, Proceedings of the National Academy of Sciences, 107, 4248
Nagylaki, T. 1974, Proceedings of the National Academy of Sciences, 71, 746
—. 1990, Theoretical Population Biology, 37, 192
Neher, R. A., & Shraiman, B. I. 2012, Genetics, 191, 1283
Neuhauser, C., & Krone, S. M. 1997, Genetics, 145, 519
Nielsen, R., Williamson, S., Kim, Y., et al. 2005a, Genome research, 15, 1566
—. 2005b, Genome Research, 15, 1566
Peter, B. M., Huerta-Sanchez, E., & Nielsen, R. 2012, PLoS Genetics, 8, e1003011
Pickrell, J. K., Coop, G., Novembre, J., et al. 2009, Genome research, 19, 826
Plantinga, T. S., Alonso, S., Izagirre, N., et al. 2012, European journal of human genetics, 20, 778
Revuz, D., & Yor, M. 1999, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], Vol. 293, Continuous martingales and Brownian motion, 3rd edn. (Berlin: Springer-Verlag), xiv+602
Roberts, G. O., & Stramer, O. 2001, Biometrika, 88, 603
Rogers, L. C. G., & Williams, D. 2000a, Diffusions, Markov processes, and martingales. Vol. 2, Cambridge Mathematical Library (Cambridge: Cambridge University Press), xiv+480, itô calculus, Reprint of the second (1994) edition
—. 2000b, Diffusions, Markov processes and martingales: Volume 2, Itô calculus, Vol. 2 (Cambridge university press)
Rouhani, S., & Barton, N. 1987, Theoretical Population Biology, 31, 465
Sawyer, S. A., & Hartl, D. L. 1992, Genetics, 132, 1161
Schraiber, J. G. 2014, Theoretical population biology, 92, 30
Schraiber, J. G., Griffiths, R. C., & Evans, S. N. 2013, Theoretical Population Biology, 89, 64
Sellis, D., Callahan, B. J., Petrov, D. A., & Messer, P. W. 2011, Proceedings of the National Academy of Sciences, 108, 20666
Sermaidis, G., Papaspiliopoulos, O., Roberts, G. O., Beskos, A., & Fearnhead, P. 2012, Scandinavian Journal of Statistics
Sjödin, P., Skoglund, P., & Jakobsson, M. 2014, Molecular biology and evolution, msu059
Slatkin, M. 1994, Genetical research, 64, 71
—. 2001, Genetical research, 78, 49

Slatkin, M., & Excoffier, L. 2012, Genetics, 191, 171
Smith, J. M., & Haigh, J. 1974, Genetical Research, 23, 23
Song, Y. S., & Steinrücken, M. 2012, Genetics, 190, 1117
Song, Y. S., & Steinrücken, M. 2012, Genetics, 190, 1117
Sørensen, M. 2009, in Handbook of financial time series (Springer), 531
Steinrücken, M., Bhaskar, A., & Song, Y. S. 2013, arXiv preprint arXiv:1310.1068
Steinrücken, M., Wang, Y., & Song, Y. S. 2012, Theoretical Population Biology
Tajima, F. 1989, Genetics, 123, 585
Tennessen, J. A., Bigham, A. W., OÕConnor, T. D., et al. 2012, Science, 337, 64
Teshima, K. M., & Innan, H. 2009, BMC Bioinformatics, 10, 166
Torgerson, D. G., Boyko, A. R., Hernandez, R. D., et al. 2009, PLoS genetics, 5, e1000592
Voight, B. F., Kudaravalli, S., Wen, X., & Pritchard, J. K. 2006, PLoS biology, 4, e72
Wall, J. D., & Slatkin, M. 2012, Annual Review of Genetics, 46, 635
Watterson, G. 1978, Genetics, 88, 405
—. 1979, Advances in Applied Probability, 14
Wiener, N. 1921, Proceedings of the National Academy of Sciences of the United States of America, 7, 294
Williamson, E. G., & Slatkin, M. 1999, Genetics, 152, 755
Wright, S. 1931a, Genetics, 16, 97
—. 1931b, Genetics, 16, 97
Zee, A. 2010, Quantum field theory in a nutshell (Princeton University Press)
Zhao, L., Yue, X., & Waxman, D. 2013, Genetics