The Stationary Distribution of Allele Frequencies when Selection acts at Unlinked Loci

Paul Fearnhead[1]

1. Department of Mathematics and Statistics, Lancaster University, Lancaster, LA1 4YF, UK
(e-mail: p.fearnhead@lancs.ac.uk).

Summary: We consider population genetics models where selection acts at a set of unlinked loci. It is known that if the fitness of an individual is multiplicative across loci, then these loci are independent. We consider general selection models, but assume parent-independent mutation at each locus. For such a model, the joint stationary distribution of allele frequencies is proportional to the stationary distribution under neutrality multiplied by a known function of the mean fitness of the population. We further show how knowledge of this stationary distribution enables direct simulation of the genealogy of a sample at a single locus. For a specific selection model appropriate for complex disease genes, we use simulation to determine what features of the genealogy differ between our general selection model and a multiplicative model.

Keywords: *Ancestral Selection Graph, Complex Diseases, Genealogies, Wright-Fisher Model*

1 Introduction

Consider selection acting on a set of unlinked loci. Under a multiplicative model for fitness, these loci will evolve independently (Risch, 1990). However, for complex diseases there may be interactions between unlinked genes, and a simple multiplicative model will be innappropriate (Wiesch *et al.*, 1999; Niu *et al.*, 1999; Cordell *et al.*, 2001). Fearnhead (2003) introduced a coalescent-type ancestral process for such models of selection, where the fitness depends in a non-multiplicative way on the alleles at a set of unlinked loci. Here we show that if parent-independent mutations occur at each loci, then the joint stationary distribution of the allele frequencies at the unlinked loci is proportional to this stationary distribution under neutrality multiplied by the exponential of the mean fitness of the population. This result is a natural extension of the result for selection acting at a single loci (Wright, 1949; Donnelly *et al.*, 2001).

Central to our proof of this result is the following result for Langevin diffusions (see Roberts and Stramer, 2002; Kent, 1978)

Theorem 1 *Consider a diffusion in d-dimensions, specified by the following stochastic differential equation*

$$dX(t) = b(X(t))dt + \sigma(X(t))dB(t) \tag{1}$$

where B is a d-dimensional Brownian Motion, $b(\cdot)$ is a d-dimensional vector and $\sigma(\cdot)$ a $d \times d$ matrix. Assume $X = (X_1, \ldots, X_n)$ is not explosive, define the $d \times d$ matrix $a(x) = \sigma(x)\sigma^T(x)$, and $\delta(x)$ as the determinant of $a(x)$. Further assume that there exists a density function $\pi(x)$ such that for all j

$$b_i(x) = \frac{1}{2}\sum_{j=1}^{d} a_{ij}(x)\partial \log \pi(x)/\partial x_j + \delta^{1/2}(x)\sum_{j=1}^{d}\frac{\partial}{\partial x_j}(a_{ij}(x)\delta^{-1/2}(x)). \tag{2}$$

Then $\pi(x)$ is the unique stationary distribution of X.

The interpretation of the d- dimensional drift vector $b(x)$ and the $d \times d$ matrix $a(x)$, are that

$$\lim_{h \to 0} \left(\frac{\mathrm{E}(X_i(t+h) - X_i(t) | X(t) = x)}{h} \right) = b_i(x), \tag{3}$$

represents the expected infinitesimal change in $X(t)$ and

$$\lim_{h \to 0} \left(\frac{\mathrm{Cov}(X_i(t+h), X_j(t+h) | X(t) = x)}{h} \right) = a_{ij}(x) \tag{4}$$

represents the instantaneous covariation of $X(t)$.

We introduce our general selection model in Section 2; this is equivalent to the model considered in Fearnhead (2003). Our result for the stationary distribution of the allele frequencies is then given in Section 3. In the following two Sections we compare features of our general selection model with multiplicative selection models where loci are independent. Firstly we compare the stationary distribution of each model; and we then look at features of the distribution of the genealogy at a single locus under each model. For the latter, we extend existing approaches to sample genealogies at a non-neutral locus. The paper ends with a discussion.

2 Model and Notation

We consider the diffusion limit of the following Wright-Fisher model. Assume a population of N diploid individuals. Each individual is charaterised by its two haplotypes at a set of L independent loci, and we assume K_i possible alleles at loci i (for $i = 1, 2, \ldots, L$). Thus one haplotype of an individual will be a vector of L alleles, describing the allele carried by the individual at each of its L loci on one of its two copies of its genome. We characterise the state of the population by the list of haplotypes (and their multiplicities) in the current generation.

The fitness of an individual with haplotypes (α, β) is $1 + s_{\alpha,\beta}$. We proceed from one generation to another as follows. We choose $2N$ pairs of haplotypes from the current generation. These choices are independent, and the chance that a particular pair, (α, β) is chosen at each choice is proportional to $1 + s_{\alpha,\beta}$. For each of these $2N$ pairs we then produce a new haplotype by choosing an allele uniformly at random from the pair of alleles at each loci (with the choices at each loci being independent of all other choices), and for locus $l = 1, \ldots, L$, we have a probability u_l of mutating the chosen allele. If a mutation occurs at locus l the mutant will be of type i with probability $\nu_i^{(l)}$, which is independent of the parent allele. (Note, that for this and all our notation that follows, we subscript by the allele and superscript by the locus.)

We consider the diffusion limit of this model as $N \to \infty$; time is measured in units of $2N$ generations; and $\theta_i^{(l)} = 4Nu_l\nu_i^{(l)}$ (population scaled mutation rates) and $\sigma_{\alpha,\beta} = 4Ns_{\alpha,\beta}$ (population scaled selection rates) are kept fixed. We let $\theta^{(l)} = \sum_i \theta_i^{(l)}$, the mutation rate at locus l. Note that this diffusion limit applies to a wide range of population models and hence our results apply more widely than for just the Wright-Fisher model.

In this diffusion limit, the loci are unlinked, and the population frequency of a haplotype α, $\Pr(\alpha)$, is obtained as the product of the population frequencies of the alleles of α across the L loci. The population can thus be characterised by the population frequencies of the alleles at each loci. We denote the frequency of allele i at loci l by $x_i^{(l)}$, and let $x = (x_1^{(1)}, \ldots, x_{K_L-1}^{(L)})$ be the set of allele frequencies at the L loci. Note that for locus l we record only the frequencies of alleles $1, \ldots, K_l - 1$ as the final allele frequency is defined by the fact that allele frequencies sum to 1.

Conditional on x we can define the mean selection rate of the population

$$\bar{\sigma} = \sum_{\alpha} \sum_{\beta} \Pr(\alpha) \Pr(\beta) \sigma_{\alpha,\beta}. \tag{5}$$

We can also define the conditional mean selection rate of an individual of genotype (i, j) at locus l as

$$\bar{\sigma}_{ij}^{(l)} = \sum_{\alpha} \sum_{\beta} \Pr_l(\alpha|i) \Pr_l(\beta|j) \sigma_{\alpha} \beta,$$

where $\Pr_l(\alpha|i)$ is the conditional probability of haplotype α given an allele i at locus l. If haplotype α has allele i at locus l then this is the product of the marginal frequencies of the alleles at other loci, otherwise this is 0. Note that the $\bar{\sigma}_{ij}^{(l)}$s are independent of the allele frequencies at locus l, and that

$$\bar{\sigma} = \sum_{i=1}^{K_l} \sum_{j=1}^{K_l} x_i^{(l)} x_j^{(l)} \bar{\sigma}_{ij}^{(l)}. \tag{6}$$

Finally we can define a mean selection rate for an allele i at locus l as

$$\bar{\sigma}_i^{(l)} = \sum_{j=1}^{K_l} x_j^{(l)} \sigma_{ij}^{(l)}.$$

By standard arguments the stochastic differential equation which describes the evolution of x is of the form (1). Using (3) and (4) we can obtain the drift function of $x_i^{(l)}$ as,

$$b_i^{(l)}(x) = \frac{1}{2}(\theta_i^{(l)} - \theta^{(l)} x_i^{(l)}) + x_i^{(l)} (\bar{\sigma}_i^{(l)} - \bar{\sigma}).$$

and that the infinitessimal covariance matrix is block diagonal (due to the sampling independence across loci). The block diagonal entries $A^{(1)}, \ldots, A^{(L)}$ correspond to the covariance matrices for loci $1, \ldots, L$ respectively, with $A^{(l)}$ having entries

$$a_{ij}^{(l)} = \begin{cases} x_i^{(l)}(1 - x_i^{(l)}) & \text{if } i = j \\ -x_i^{(l)} x_j^{(l)} & \text{otherwise.} \end{cases}$$

Note that the form of this covariance matrix comes directly from the covariance of multinomial random variables.

Finally, we note that the neutral model is a special case of this model. Selection only effects the drift function, and if $c(x)$ is the drift function under neutrality we have

$$b_i^{(l)}(x) = c_i^{(l)}(x) + x_i^{(l)}(\bar{\sigma}_i^{(l)} - \bar{\sigma}). \tag{7}$$

Furthermore, in the neutral case the allele frequencies at each locus are independent. We denote the marginal distribution of allele frequencies at locus l in this case by

$$\pi_N^{(l)}(x^{(l)}) \propto \prod_{i=1}^{K_l} (x_i^{(l)})^{\theta_i^{(l)}-1}, \tag{8}$$

which is a Dirichlet distribution with parameters $(\theta_1^{(l)}, \ldots, \theta_{K_l}^{(l)})$.

3 Stationary Distribution

Our main result is the following, which gives the stationary distribution of the multi-locus selection model introduced in Section 2.

Theorem 2 *Consider the multi-locus model of Section 2. Let $\pi_N(x)$ be the stationary distribution of the neutral model ($\sigma_{\alpha,\beta} = 0$ for all α and β),*

$$\pi_N(x) = \prod_{l=1}^{L} \pi_N^{(l)}(x^{(l)}),$$

where $\pi_N^{(l)}(x^{(l)})$ is defined by (8). The stationary distribution of the general selection model is

$$\pi(x) \propto \pi_N(x) \exp\left\{\frac{1}{2}\bar{\sigma}\right\}, \tag{9}$$

where $\bar{\sigma}$ is the mean population selection rate, defined by (5).

Proof: See Appendix A. □

This result is a multilocus extension of the result presented in Donnelly *et al.* (2001). For the single locus case, (9) differs from equations 1–4 of Donnelly *et al.* (2001) by the factor of 1/2 in the exponent; and this is because we chose a different scaling for our population-scaled selection rates (of $4N$ as opposed to $2N$).

A special case of the general selection model of Section 2 is *genic* selection, where $\sigma_{\alpha,\beta} = \sigma^*_\alpha + \sigma^*_\beta$. In this case, we can define a haplotype mean selection rate $\bar{\sigma}^* = \sum \Pr(\alpha)\sigma^*_\alpha$, and we get $\bar{\sigma} = 2\bar{\sigma}^*$. Thus for genic selection

$$\pi(x) \propto \pi_N(x) \exp\left\{\bar{\sigma}^*\right\}.$$

Finally the conditional distribution of the population allele frequencies at locus l, given the population frequencies at all other loci satisfies

$$\pi(x^{(l)}|x^{(-l)}) \propto \pi_N^{(l)}(x^{(l)}) \exp\left\{\sum_i \sum_j x_i^{(l)} x_j^{(l)} \sigma_{ij}^{(l)}\right\}.$$

Which is the stationary distribution of allele frequencies for a locus with the same mutation parameters as locus l, and selection rates $\sigma_{ij}^{(l)}$.

4 Properties of the Stationary distribution

We now consider properties of the joint stationary distribution of allele frequences at different loci, and how this distribution differs from that of a multiplicative selection model for which the allele frequencies are independent across loci. There are a large range of possible multi-locus models and parameter values we could consider. For simplicity we concentrate on 2-locus models with genic selection, and we choose parameter values suitable for modelling complex disease genes. The properties of the marginal and joint distributions of allele frequencies in the multiplicative model have been studied by Pritchard (2001) and Pritchard and Cox (2002) to address the question of whether common complex diseases are caused by common or rare variants. We will focus on one such model for complex diseases, and address to what extent the conclusions of this work are robust to the assumption of a multiplicative selection model.

We consider a 2 locus model with alleles 1 and 2 at each locus. The loci represent 2 unlinked genes, and we let 1 denote the susceptible allele at each gene (the allele that marginally increases the risk of a disease), and 2 the normal allele. Selection is in favour of the normal allele, whereas the mutation rate from normal to susceptible allele is much greater than for the reverse mutation, modelling the fact that there may be many ways to impair the function of a gene, but mutations which repair a gene must be more specific (see Pritchard, 2001, for more discussion)

We simplify notation from Section 3 by dropping the $*$ superscript on genic selection rates. We fix mutation rates at $\theta_1^{(l)} = 1.5$ and $\theta_2^{(l)} = 0.1$. We assume the selection rates for the haplotype across the two loci satisfy $\sigma_{11} = 0$, $\sigma_{12} = \sigma_{21} \geq 0$ and $\sigma_{22} \geq \sigma_{12}$. Under the multiplicative model of Pritchard (2001), $\sigma_{12} = \sigma_{21} = \sigma$ and $\sigma_{22} = 2\sigma$ for some parameter σ. We consider two general selection models, which are the two models most different from the multiplicative model: (A) $\sigma_{12} = \sigma_{21} = 0$; and (B) $\sigma_{22} = \sigma_{12}$.

For models A and B we choose the (single) selection rate in the model to be 12. For an appropriate comparison with a multiplicative selection model we found the value of σ in the multiplicative selection model which gives the same mean population frequency of the susceptible alleles at each loci. These mean frequencies are 49% and 52% for models A and B respectively; and we require $\sigma = 5.6$ and $\sigma = 5.3$ in the multiplicative model to attain the same mean frequencies. We denote these two multiplicative models as models MA and MB respectively.

A comparison of the joint distribution of allele frequencies at the two loci is shown in Figure 1, and a comparison of the marginal allele frequencies (calculated via numerical integration, using ideas from Joyce, 2005) at a single locus are shown in Figure 2. The joint distribution for model A has a similar, though more highly peaked, mode at low allele frequencies at both loci as compared to model MA. The main difference between these two models is that model A has a further mode where both allele frequencies are large; whereas the model MA has further modes where only one of the allele frequencies is large. The distribution of marginal allele frequencies has similar modes for both models A and MA, with the main difference in the two distributions being that the variance of the allele frequencies is larger for model A, with less mass at intermediate frequencies (20%-80%).

There is a greater difference in the joint distribution of allele frequencies for models B and MB. Model B has modes for a high allele frequency at just one loci; whereas the main mode for model MB is for low allele frequencies at both loci. Again the main difference in the marginal distribution of allele frequencies at a single locus is that there is greater variance for model B, with less probability mass at intermediate frequencies.

We further looked at the distribution of the larger and smaller minor allele frequencies at the two loci for each model. These are shown in Figure 3, and are of interest as they show the probability of either one of both loci having a high minor allele frequency (and hence being a common variant responsible for the complex disease we are modelling). The distribution of the larger minor allele frequency is concentrated on smaller values for the general models A and B then for the comparative multiplicative models, while there is little difference in the smaller minor allele frequency in each case. This shows that the probability of a common variant for the complex disease under either general model is smaller than under the corresponding multiplicative model, and is consistent with the smaller probability mass at intermediate frequencies (see Figure 2) under the general models. Quantitatively, the difference between general and multiplicative models is small for Model B (probability of larger minor allele frequency greater than 10% is 85% and 94% for the general and multiplicative models), but much more substantial for model A (probabilities of 67% and 94% for the general and multiplicative models).

5 Simulating the Genealogy

We now consider simulating from the genealogy at a selective locus under our multi-locus selection model. We first describe the ancestral process for these models, and then how knowledge of the stationary distribution of the population allele frequency enables us to perform exact simulation of the genealogy (at one locus) of a sample conditional on the type of the sample. Furthermore, this enables us to sample from the unconditional distribution of the genealogy by (i) simulating the type of the sample (using Equation 9 to first simulate the population allele frequencies); and (ii) simulating the genealogy conditional on the type of the sample.

The idea of conditional simulation of genealogies under (single-locus) selection was first considered by Slade (2000), and exact simulation was introduced for these models by Stephens and Donnelly (2003). We use the approach of Stephens and Donnelly (2003), together with a simplification from Fearnhead (2002) which can greatly reduce the computational burden, to perform exact conditional simulation for our multi-locus selection models.

For simplicity and ease of exposition, we solely consider 2-locus, 2-allele genic selection models. We denote the alleles at each locus by 1 and 2. As described

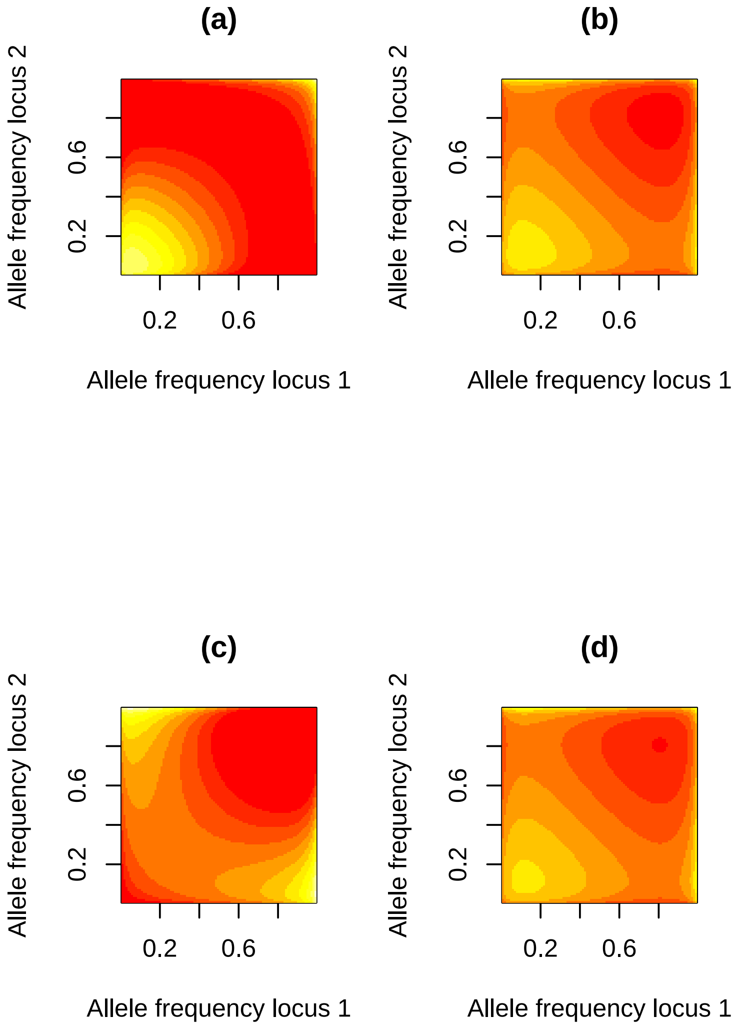

Figure 1: Comparison of the joint distribution of allele frequencies at two loci. (a) General model A, $\sigma_{12} = \sigma_{21} = 0$ and $\sigma_{22} = 12$ (b) Multiplicative model MA $\sigma = 5.6$; (b) General model B, $\sigma_{12} = \sigma_{21} = \sigma_{22} = 12$; and (d) Multiplicative model MB $\sigma = 5.3$. The mean marginal allele frequencies are the same in (a) and (b); and in (c) and (d). In each plot lighter colours imply higher density.

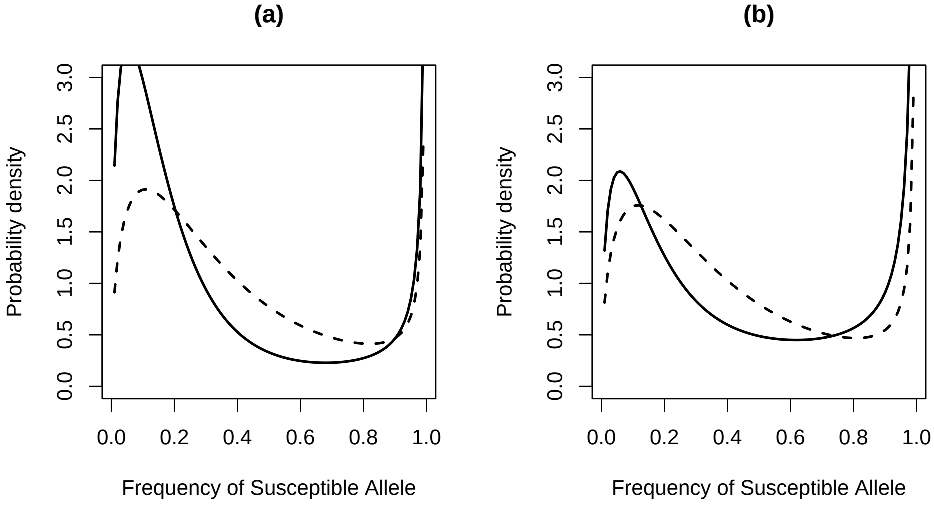

Figure 2: Comparison of the marginal allele frequency of two-locus models. (a) General model A, $\sigma_{12} = \sigma_{21} = 0$ and $\sigma_{22} = 12$ (full-line) and multiplicative model MA, $\sigma = 5.6$ (dashed-line); (b) General model B, $\sigma_{12} = \sigma_{21} = \sigma_{22} = 12$ (full line) and multiplicative model MB, $\sigma = 5.3$ (dashed-line).

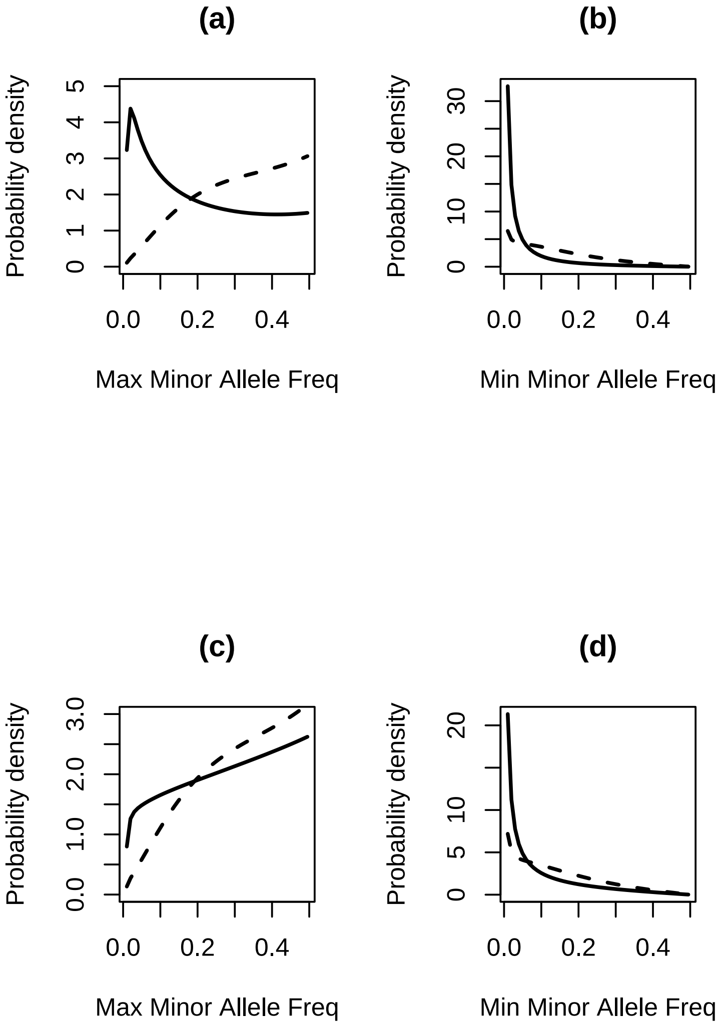

Figure 3: Comparison of the marginal distribution of the larger and smaller minor allele frequencies. (a) larger minor allele frequencies; and (b) smaller minor allele frequencies for general model A (full-line) and multiplicative model MA (dashed-line); (c) larger minor allele frequencies; and (d) smaller minor allele frequencies for general model B (full-line) and multiplicative model MB (dashed-line).

above, genic selection models are parameterised by a set of selection rates, one for each possible haplotype. We denote these rates σ_{ij} for a haplotype with allele i at the first locus, and allele j at the second locus (again we have dropped the $*$ superscript used in Section 3 to simplify notation). Without loss of generality we assume $\sigma_{11} = 0$. For concreteness, from Section 5.2 onwards we further assume that $\sigma_{12} = \sigma_{21} = \sigma_1$ and $\sigma_{22} = \sigma_1 + \sigma_2$, with $\sigma_1, \sigma_2 \geq 0$. However, generalisation of our approach to multi-locus, multi-allele and general selection models is possible.

5.1 Ancestral Processes

The ancestral process for our multi-locus model was derived in Fearnhead (2003), and is called the complex selection graph (CSG). This processes is an extension of the Ancestral Selection Graph (Krone and Neuhauser, 1997) and the Ancestral Influence Graph (Donnelly and Kurtz, 1999) and produces supra-genealogies, that is graphs in which the genealogies are embedded, of samples at each of the loci. For a sample consisting of $n^{(l)}$ chromosomes at locus l, the CSG is started at time 0 with $n^{(l)}$ branches at locus l, for $l = 1, 2$. The CSG is a continuous time Markov process, which simulates events in the history of the sample. The possible events are coalescent, mutation, and selection events. If at time t (in the past) there are $n(t)^{(l)}$ branches at locus l, then (backwards in time) coalescent and mutation events occur at the same rates as for the coalescent $(n(t)^{(l)}(n(t)^{(l)}-1)/2$ and $n(t)^{(l)}\theta^{(l)}/2$ respectively), independent of the state or events at the other locus. Selection events jointly affect both supra-genealogies. A selection event occurs at rate $\sigma/2$ to each branch at each locus, where $\sigma = \max\{\sigma_{ij}\}$ is the maximum selection rate. At a selection event a new branch is added to the supra-genealogy at each locus. The branch to which the selection event occured becomes the *continuing* branch, the new branch at the same locus the *incoming* branch, and the new branch at the other locus the *linked-incoming* branch.

If the CSG is simulated until a time T in the past, and the alleles on each of the branches in the CSG at time T are simulated, then the alleles on the branches in the CSG for all $t < T$ can be simulated forward in time, and the genealogy of the sample at each locus can be recovered. Simulation forward in time is the same as for the coalescent except for at the selection events. Consider a selection event to a branch at locus 1. At such a selection event, if the alleles on the incoming and linked-incoming branches are i and j respectively, then with probability σ_{ij}/σ the incoming branch will be parental, otherwise the continuing branch will be. This determines the allele of the branch to which the selection event occurred, and by resolving which branches are parental at each selection event we can obtain the genealogy at each locus.

For full details of the CSG, and its generalisations to multi-locus, multi-allele, and general selection models see Fearnhead (2003). Whilst the above procedure can be used to simulate genealogies under our multi-locus model, it can be inefficient. In particular, the CSG has to be simulated back until a time T that is sufficiently large that there is negligible probability that the time to the most recent common ancestor (TMRCA) at each locus will be larger than T. A more efficient approach to simulating genealogies is to first simulate the type of the sample, and then simulate the CSG back in time conditional on this.

5.2 Conditional Simulation

We now assume that the type of our sample is known. We focus on the case where the genealogy at a single locus is of interest. Without loss of generality we assume that this is the first locus, and consider a sample of size $n^{(1)}$ at locus 1, and of size 0 at locus 2. We further assume the haploid selection model where $\sigma_{12} = \sigma_{21} = \sigma_1$ and $\sigma_{22} = \sigma_1 + \sigma_2$, with $\sigma_1, \sigma_2 \geq 0$. We initially know the alleles of the $n^{(1)}$ branches at locus 1 in the CSG, and we will simulate the CSG backwards in time such that we always know the allele on each branch at each locus. At time t in the past we will let $n(t)_i^{(l)}$ denote the number of branches at locus l which have allele i. Where the meaning is clear, we write $n_i^{(l)}$ for $n(t)_i^{(l)}$ in the following.

Backward Simulation

Stephens and Donnelly (2000) show how to calculate the rates of events in the coalescent backwards in time, conditional on knowing the alleles on the branches. These calculations were extended to the ancestral selection graph in Stephens and Donnelly (2003). We briefly describe the general form of these calculations, which can be intuitively viewed as an application of Bayes formula, or of time-reversing a Markov process.

Let $\pi(x)$ denote the stationary distribution of population allele frequencies (at a single locus). We can define the stationary probability of an ordered sample A (which consists of n_i branches carrying allele i, for $i = 1, \ldots, K$) by

$$\pi(A) = \int \left(\prod_{i=1}^{K} x_i^{n_i} \right) \pi(x) \mathrm{d}x.$$

Now assume that the current state of our ancestral process, defined as the alleles on each of the branches, is denoted by A; the unconditional rate of a specific event is λ and the new state of the process after this event is A'. Then the conditional rate of such an event is just

$$\lambda \pi(A')/\pi(A). \tag{10}$$

These results generalise to the multi-locus case we consider, and a list of the conditional rates are given in Table 1. In this case A will represent the type of branches at both loci. Our numerical method for calculating of $\pi(A)$ is described in Appendix B.

Virtual and Real Branches

Event	Rate
Coalescence: two branches type i	$n_i^{(1)}(n_i^{(l)}-1)\pi(A-i)/(2\pi(A))$
Mutation: branch i to j	$n_i^{(1)}\theta_i^{(l)}\pi(A-i+j)/(2\pi(A))$
Selection: to branch type 1:	
S1a: addition of $(1,1)$	$n_1^{(1)}(\sigma_1+\sigma_2)\pi(A+1^{(1)}+1^{(2)})/(2\pi(A))$
S1b: addition of $(1,2)$	$n_1^{(1)}(\sigma_1+\sigma_2)\pi(A+1^{(1)}+2^{(2)})/(2\pi(A))$
S1c: addition of $(2,1)$	$n_1^{(1)}\sigma_2\pi(A+2^{(1)}+1^{(2)})/(2\pi(A))$
S1d: addition of $(2,2)$	$n_1^{(1)}\sigma_1\pi(A+2^{(1)}+2^{(2)})/(2\pi(A))$
Selection: to branch type 2:	
S2a: addition of $(1,1)$	$n_2^{(1)}(2\sigma_1+\sigma_2)\pi(A+1^{(1)}+1^{(2)})/(2\pi(A))$
S2b: addition of $(1,2)$	$n_2^{(1)}(\sigma_1+2\sigma_2)\pi(A+1^{(1)}+2^{(2)})/(2\pi(A))$
S2c: addition of $(2,1)$	$n_2^{(1)}(\sigma_1+\sigma_2)\pi(A+2^{(1)}+1^{(2)})/(2\pi(A))$
S2d: addition of $(2,2)$	$n_2^{(1)}(\sigma_1+\sigma_2)\pi(A+2^{(1)}+2^{(2)})/(2\pi(A))$

Table 1: Backward rates conditional on the current state: the set of alleles on each branch in the CSG. Rates are given for events at locus 1 (the rates at locus 2 can be calculated by symmetry). The current state is denoted by A. For coalesence and mutation events we have used the shorthand, whereby we denote by $A-i$ and $A+j$, states which differ from A by the removal of an allele i and the addition of allele j at locus 1 respectively. For selection events we use the notation $A+1^{(1)}+2^{(2)}$ to denote a state which differs from A by the addition of a branch of type 1 at locus 1 and of type 2 at locus 2. At selection events two virtual branches are added to CSG; for example the event denoted S1b adds a branch of type 1 at locus 1 and a branch of type 2 at locus 2. For derivation of the selection rates see Appendix C.

By simulating selection events conditional on the alleles on the branches, we are able to determine which of the incoming and continuing branches are ancestral (see Table 1). As noted by Slade (2000), this enables us to keep track of which branches in the CSG are ancestral to the sample, and hence in the genealogy of the sample, and which are not. We call the branches which are in the genealogy *real* branches, and those which are not, *virtual* branches.

There are two practical advantages of keeping track of which branches are real and which are virtual. Firstly it enables us to determine the first point at which we can stop our conditional simulation of the genealogy, as this will be the first time at which there is just one real branch at the first locus in our CSG.

Secondly, it enables us to simplify the conditional simulation of the CSG, as it is only events which effect the real branches that we are interested in. We do need to include virtual branches in our CSG, however it is shown in Fearnhead (2002) that some virtual branches can be removed. The intuition behind this idea is as follows.

A key feature of the CSG (and other ancestral processes) is that the distribution of the population allele frequencies at a time t in the past, conditional on the events in the CSG up to time t, is equal to the conditional distribution of the allele frequencies given the state of the CSG at time t. This is the condition that means that (10) is the correct rate for simulating events in the CSG (and that these rates only depend on the current state). The reason that virtual branches cannot be removed from the CSG is that this condition would no longer hold: that is that the distribution of the population allele frequencies at a time t in the past, conditional on the events in the CSG up to time t, is not necessarily equal to the conditional distribution of the allele frequencies given the alleles on the real branches at time t.

However it is possible to remove certain virtual branches in the CSG, such that this key feature of the CSG is maintained. The condition for removing branches is that the distribution of the allele on the virtual branch must be the conditional distribution of an allele given the alleles on all branches in the CSG.

Removal of Virtual Branches

In practice, virtual branches can be removed at (i) mutation events to virtual branches; and (ii) selection events. We first describe (i) in detail, and then more briefly cover the calculations for (ii).

Consider a mutation event at locus l (backwards in time) to a virtual branch of type i. Let the current state of the CSG be A, and denote by $A-i+j$ the state obtained by removing an allele of type i and adding an allele of type j to the current set of alleles at locus l. (We have suppressed the dependence on l in our notation to simplify notation.) The backward rate at which such an event occurs is $\theta_i^{(l)}\pi(A-i+j)/(2\pi(A))$. Note the mutation rate depends on i and not j, as the actual mutation forward in time is from allele j to i.

By summing over j we get that the rate of a mutation at locus l to a virtual branch of type i is just $\theta_i^{(l)}\pi(A-i)/(2\pi(A))$. So we could simulate a mutation event by (a) simulating a mutation event with rate $\theta_i^{(l)}\pi(A-i)/(2\pi(A))$; and (b) conditional on a mutation occuring, simulate the new type from $\pi(A-i+j)/\pi(A-i) = \pi(j|A-i)$, the conditional distribution of an allele given the alleles on all the other branches in the CSG. As shown by Theorem 1 of Fearnhead (2002), and described above, we can thus remove this branch.

We now describe the calculations for selection events in the case $\sigma_2 \geq \sigma_1$. The calculations for $\sigma_2 < \sigma_1$ proceed similarly. The resulting rates of selection event are summarised in Table 2.

First consider selection to branches of type 1, and sum the rate of the events S1a–S1d. Using the same notation as in Table 1, the total rate of selection events to branches of type 1 becomes

$$\frac{n_1^{(1)}}{2\pi(A)}\left(\sigma_1(\pi(A+1^{(1)}+1^{(2)})+\pi(A+1^{(1)}+2^{(2)})+\pi(A+2^{(1)}+1^{(2)})+\pi(A+2^{(1)}+2^{(2)}))\right.$$
$$\left.+\sigma_2(\pi(A+1^{(1)}+1^{(2)})+\pi(A+1^{(1)}+2^{(2)}))+(\sigma_2-\sigma_1)\pi(A+2^{(1)}+1^{(2)})\right),$$

which simplifies to

$$\frac{n_1^{(1)}}{2\pi(A)}\left(\sigma_1\pi(A)+\sigma_2\pi(A+1^{(1)})+(\sigma_2-\sigma_1)\pi(A+2^{(1)}+1^{(2)})\right).$$

Thus we can split simulating an selection event to a specfic branch of type 1 into three, each refering to a term in this expression:

(1) At rate $n_1^{(1)}\sigma_1/2$ simulate the types of new branches at both loci from $\pi(i^{(1)}, j^{(2)}, |A)$.

(2) At rate $n_1^{(1)}\sigma_2\pi(A+1^{(1)})/(2\pi(A))$ add a branch of type 1 to locus 1. Simulate the type of a new branch at locus 2 from $\pi(i^{(2)}|A+1^{(1)})$.

(3) At rate $n_1^{(1)}(\sigma_2-\sigma_1)\pi(A+2^{(1)}+1^{(2)})/(2\pi(A))$ add a branch of type 2 to locus 1 and of type 1 to locus 2.

Thus by Theorem 1 of Fearnhead (2002), we do not need to add the branch at locus 2 for event (2) or either new branch for event (1). Thus in total we add a new virtual branch of type 1 to locus 1 with rate $n_1^{(1)}\sigma_2\pi(A+1^{(1)})/(2\pi(A))$ and two virtual branches of types 2 at locus 1 and type 1 at locus 2 with rate $n_1^{(1)}(\sigma_2-\sigma_1)\pi(A+2^{(1)}+1^{(2)})/(2\pi(A))$.

By similar argument for selection events to branches of type 2, we just need to add a new virtual branch of type 1 to locus 1 with rate $n_2^{(1)}\sigma_2\pi(A+1^{(1)})/(2\pi(A))$, and two new virtual branches, of types 1 at locus 1 and 2 at locus 2 with rate $n_2^{(1)}(\sigma_2-\sigma_1)\pi(A+2^{(1)}+1^{(2)})/(2\pi(A))$.

Add	$\sigma_1 \leq \sigma_2$	$\sigma_1 > \sigma_2$
$1^{(1)}$	$(n_2^{(1)}\sigma_1 + n_1^{(1)}\sigma_2)\pi(1^{(1)}\|A)/2$	$(n_2^{(1)}\sigma_2 + n_1^{(1)}\sigma_1)\pi(1^{(1)}\|A)/2$
$1^{(2)}$	$(n_2^{(2)}\sigma_1 + n_1^{(2)}\sigma_2)\pi(1^{(2)}\|A)/2$	$(n_2^{(2)}\sigma_2 + n_1^{(2)}\sigma_1)\pi(1^{(2)}\|A)/2$
$1^{(1)}, 1^{(2)}$	0	$(n_2^{(1)} + n_2^{(2)})(\sigma_1 - \sigma_2)\pi(1^{(1)} + 1^{(2)}\|A)/2$
$1^{(1)}, 2^{(2)}$	$(n_2^{(1)} + n_1^{(2)})(\sigma_2 - \sigma_1)\pi(1^{(1)} + 2^{(2)}\|A)/2$	0
$2^{(1)}, 1^{(2)}$	$(n_1^{(1)} + n_2^{(2)})(\sigma_2 - \sigma_1)\pi(2^{(1)} + 1^{(2)}\|A)/2$	0
$2^{(1)}, 2^{(2)}$	0	$(n_1^{(1)} + n_1^{(2)})(\sigma_1 - \sigma_2)\pi(2^{(1)} + 2^{(2)}\|A)/2$

Table 2: Simplified rates for addition of virtual branches due to selection events. For notational convenience we write $\pi(1^{(1)}|A)$ for $\pi(A+1^{(1)})/\pi(A)$ etc. See text for details of calculation of these rates; rates are calculated by summing rates of adding branches of specific type at selection events for both loci.

5.3 Results

We simulated genealogies at the first locus under Models A and B from Section 4. For each model we considered two cases (i) the distribution of the genealogy at a locus conditional on a sample of 50 alleles of type 1 and 50 alleles of type 2 at that locus; and (ii) the unconditional genealogy of a sample of size 100. For comparison, we also simulated genealogies under models MA and MB (see Section 4) in each case. We simulated 1000 genealogies for each of these eight cases, and looked at the distribution of summaries of these genealogies and the mutations on the genealogies.

Firstly we looked at three features of the genealogies, the time to the most recent common ancestor (TMRCA), the total length of the branches, and the total length of the exterior branches. We found little difference in the distribution of these between the general and multiplicative models in each to the four cases above. For example, Figure 4 shows the distributions of TMRCA and length of the tree for Model A.

Secondly we looked at features of mutations on the genealogies. We motivated models A and B as suitable for complex disease genes, and the age and frequency of susceptible mutations under such models are important factors underlying the power of studies to detect such genes (see Pritchard, 2001). Again we aim to get some insight into the robustness of the results in Pritchard (2001) to the assumption of a multiplicative model for selection.

We studied the number of mutations from a Normal to a Susceptible allele, and the age of the most common Susceptible allele. For each of our four scenarios the mean number of mutations is smaller under the general model (2.2, 3.6, 2.1 and 3.5 for scenarios A(i), A(ii), B(i) and B(ii) respectively) than the multiplicative model (2.8, 3.9, 2.5 and 3.9 respectively for the four scenarios). The distribution of the age of the most common mutation is concentrated on smaller values for the general models (see Figure 5). These results suggest smaller allelic heterogeneity under the general models, which means greater power for association studies; and younger mutations under the general models, which means larger regions around the mutations that are identical by descent.

6 Discussion

We have considered a population genetics model for unlinked loci, where parent-independent mutations occur at each locus, but where the fitness of an individual depends on the alleles at each locus. If the fitness of an individual depended in a multiplicative way on individual fitnesses of genotypes at each locus, then the distribution of the allele frequencies would be independent across loci. We have calculated the stationary distribution of the allele frequencies for a general selection model. This distribution is proportional to the distribution under neutrality

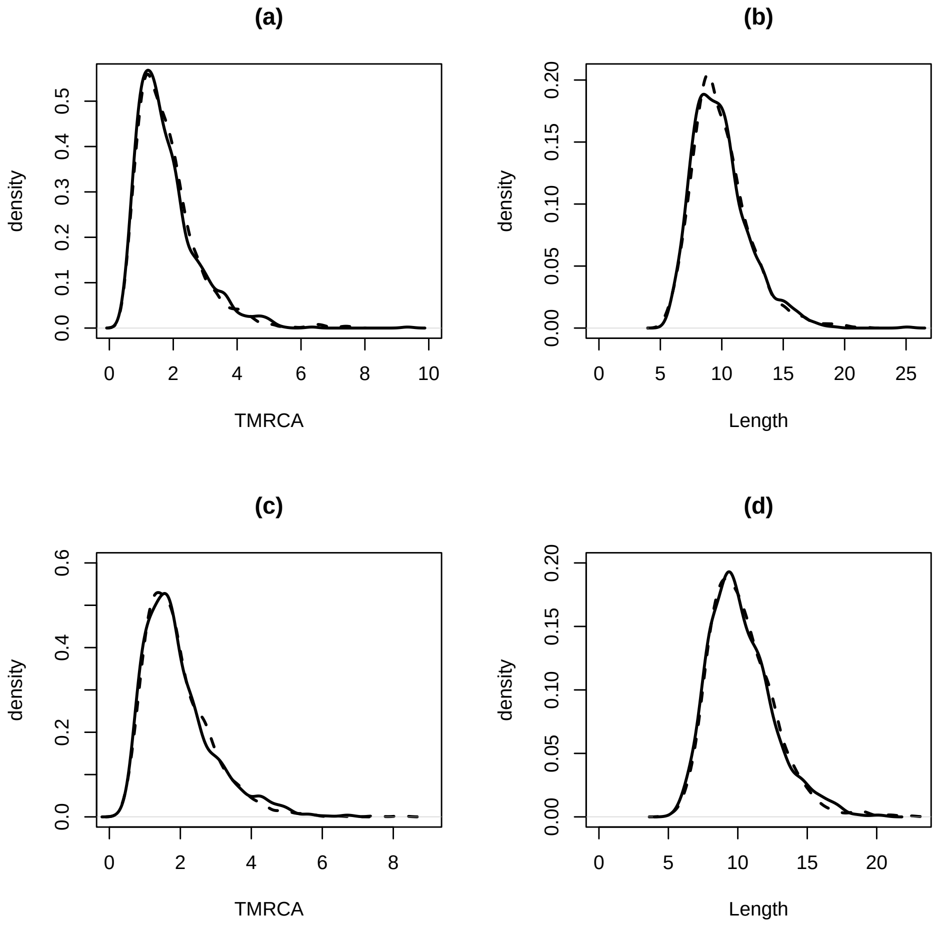

Figure 4: Comparison of the distribution of TMRCA and length of the tree for a sample of size 100 for Model A (full-line) and a model MA (dashed-line). Plots (a) and (b): unconditional distributions; plots (c) and (d): conditional on 50 alleles of each type.

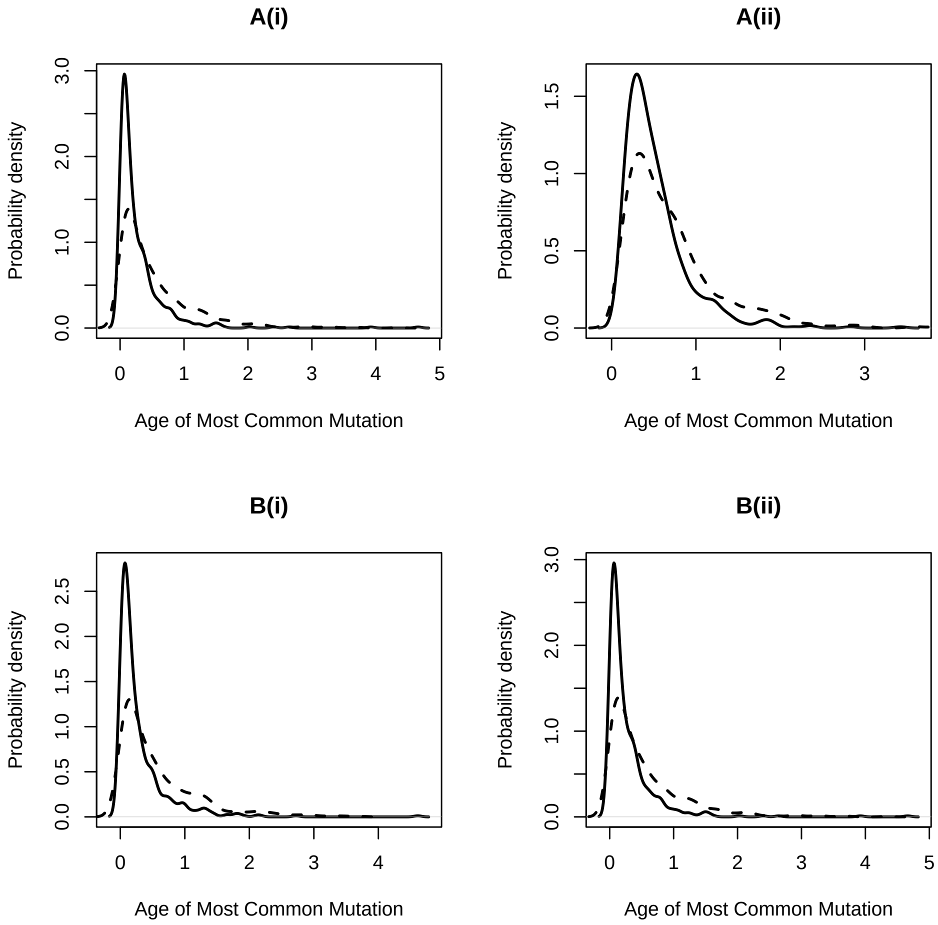

Figure 5: Comparison of the age of the most common susceptible mutation. Plots are labelled according to the Model (A or B) and whether unconditional (i) or conditional (ii). In each plot the full-line is the density under the general model, and the dashed-line the density under the multiplicative model.

multiplied by the exponential of half the mean population fitness. This result is a multi-locus extension of the result in Donnelly *et al.* (2001).

We have also shown how knowledge of this stationary distribution allows independent samples from the distribution of genealogies at a single locus to be simulated. Our simulation method is based on the idea of Stephens and Donnelly (2003), but additionally uses the simplification of Fearnhead (2003). This simplification, and the resulting method for simulation applies more general to simulation of genealogies at selected loci. For example, it trivially applies to single locus models (as they are a special case of our multi-locus models); and this simplification can substantially reduce the computational burden of the approach of Stephens and Donnelly (2003) for some selection models.

We have presented a few results comparing the features of our general selection model and the multiplicative model. Our focus has been on suitable models for complex diseases. For the parameter values considered we found that most important features of the data and genealogy under the multiplicative model were very similar to those under the general model. The two exceptions appear to be that (i) mutations which increase susceptibility to the disease (or equivalently reduce fitness) tend to be younger under the general model; (ii) there are fewer such mutations in the history of a sample under the general model; and (iii) there is less chance of obtaining alleles at intermediate frequency under the general model.

Acknowledgement I would like to thank Gareth Roberts for helpful discussions about Langevin Diffusions. This work was funded by EPSRC grant GR/R91724/01.

References

Cordell, H. J., Todd, J. A., Hill, N. J., Lyons, P. A., Peterson, L. B., Wicker, L. S. and Clayton, D. G. (2001). Statistical modeling of interlocus interactions in a complex disease: Rejection of the multiplicative model of epistasis in type 1 diabetes. *Genetics* **158**, 357–367.

Donnelly, P. and Kurtz, T. (1999). Genealogical processes for Fleming-Viot models with selection and recombination. *Annals of Applied Probability* **9**, 1091–1148.

Donnelly, P., Nordborg, M. and Joyce, P. (2001). Likelihoods and simulation methods for a class of non-neutral population genetics models. *Genetics* **159**, 853–867.

Fearnhead, P. (2002). The common ancestor at a non-neutral locus. *Journal of Applied Probability* **39**, 38–54.

Fearnhead, P. (2003). Ancestral processes for non-neutral models of complex diseases. *Theoretical Population Biology* **63**, 115–130.

Fearnhead, P. and Meligkotsidou, L. (2004). Exact filtering for partially-observed continuous-time Markov models. *Journal of the Royal Statistical Society, series B* **66**, 771–789.

Joyce, P. (2005). Efficient simulation methods for a class of nonneutral population genetics models. *Theoretical Population Biology* **to appear**.

Kent, J. (1978). Time-reversible diffusions. *Advances in Applied Probability* **10**, 819–835.

Krone, S. M. and Neuhauser, C. (1997). Ancestral processes with selection. *Theoretical Population Biology* **51**, 210–237.

Niu, T. H., Xu, X. P., Cordell, H. J., Rogus, J., Zhou, Y. S., Fang, Z. and Lindpaintner, K. (1999). Linkage analysis of candidate genes and gene-gene interactions in Chinese hypertensive sib pairs. *Hypertension* **33**, 1332–1337.

Pritchard, J. K. (2001). Are rare variants responsible for susceptibility to complex diseases? *American Journal of Human Genetics* **69**, 124–137.

Pritchard, J. K. and Cox, N. J. (2002). The allelic architecture of human disease genes: common disease–common variant...or not? *Human Molecular Genetics* **11**, 2417–2423.

Risch, N. (1990). Linkage strategies for genetically complex traits. I. Multilocus models. *American Journal of Human Genetics* **46**, 222–228.

Roberts, G. O. and Stramer, O. (2002). Tempered Langevin diffusions and algorithms .

Slade, P. F. (2000). Simulation of selected genealogies. *Theoretical Population Biology* **57**, 35–49.

Stephens, M. and Donnelly, P. (2000). Inference in molecular population genetics (with discussion). *Journal of the Royal Statistical Society, Series B* **62**, 605–655.

Stephens, M. and Donnelly, P. (2003). Ancestral inference in population genetics models with selection. *Australian and New Zealand Journal of Statistics* **45**, 395–423.

Wiesch, D. G., Meyers, D. A. and Bleecker, E. R. (1999). Genetics of asthma. *Journal of allergy and clinical immunology* **104**, 895–901.

Wright, S. (1949). Adaption and selection. In: *Genetics, Paleontology and Evolution* (eds. G. L. Jepson, G. G. Simpson and E. Mayr), Princeton University Press, Princeton, NJ, 365–389.

Appendix A: Proof of Thoerem 2

An immediate corollary of Theorem 1 is the following. Assume we have a d-dimensional diffusion with drift function $c(x)$ and covariance matrix $a(x)$ which satisfies Equation 2 (with $c(x)$ in place for $b(x)$) for a given probability density $\pi(x)$. Further assume a process X' satisfies is second d-dimensional diffusion which satisifes stochastic differential equation (1) with drift function

$$b(x) = c(x) + d(x)$$

for some d-dimensional function $d(x)$, but the same covariance matrix $a(x)$. Then if there exists a positive d-dimensional function $\lambda(x)$ such that for $i = 1, \dots, d$,

$$d_i(x) = \frac{1}{2} \sum_{j=1}^{d} a_{ij}(x) \partial \log \lambda(x) / \partial x_j \tag{11}$$

then the stationary distribution, $\tilde{\pi}(x)$ of X' satisfies

$$\tilde{\pi}(x) \propto \pi(x)\lambda(x).$$

This corrolary can be applied to the multi-locus selection model of Section 2 by letting the first diffusion be that for the neutral model, and the X' diffusion be that for the non-neutral model. The relationship between the drifts is given by (7), so

$$d_i^{(l)}(x) = x_i^{(l)}(\bar{\sigma}_i^{(l)} - \bar{\sigma}).$$

We thus only need to show that $\log \lambda(x) = \frac{1}{2}\bar{\sigma}$ satisfies (11) in order to prove Theorem 2. We further note that the if we consider the drift for $x_{(i)}^{(l)}$ we need only consider the sum over j for the allele frequencies at locus l on the right-hand side of (11), as the a_{ij} terms are 0 for allele frequencies at two distinct loci.

By using (6), and noting $\sigma_{ij}^{(l)} = \sigma_{ji}^{(l)}$ we can see that

$$\partial\bar{\sigma}/\partial x_j^{(l)} = 2\sum_{k=1}^{K_l}(\sigma_{kj}^{(l)} - \sigma_{kK_l}^{(l)}).$$

So if $\lambda(x) = \exp\{\bar{\sigma}/2\}$ the right-hand side of (11) becomes

$$x_i^{(l)}\left[\sum_{k=1}^{K_l} x_k^{(l)}\sigma_{kj}^{(l)} - \sum_{k=1}^{K_l} x_k^{(l)}\sigma_{kK_l}^{(l)} - \sum_{j=1}^{K_l} x_j^{(l)}\sum_{k=1}^{K_l} x_k^{(l)}\sigma_{kj}^{(l)} + \sum_{j=1}^{K_l} x_j^{(l)}\sum_{k=1}^{K_l} x_k^{(l)}\sigma_{kK_l}^{(l)}\right]. \tag{12}$$

The first two sums on the left-hand side of this equation come from the product of $\partial\bar{\sigma}/\partial x_i^{(l)}$ and the $x_i^{(l)}$ term in a_{ii} and the third and fourth terms come from the product of $\partial\bar{\sigma}/\partial x_j^{(l)}$ and the $-x_i^{(l)}x_j^{(l)}$ terms in the a_{ij} (remembering $a_{ij} = 0$ for allele frequencies at separate loci).

Finally we note that the second and fourth sums in this equation cancel; the first simplifies to $\bar{\sigma}_i^{(l)}$ and the third simplifies to $\bar{\sigma}$. Thus (12) simplifies to $d_i^{(l)}(x)$ as required. □

Appendix B: Evaluating $\pi(A)$

Calculating $\pi(A)$ requires the calculation of an integral of the form

$$\int_0^1\int_0^1 x^{a-1}(1-x)^{b-1}y^{c-1}(1-y)^{d-1}\exp\{\sigma_1(x(1-y)+y(1-x))+(\sigma_1+\sigma_2)xy\}\mathrm{d}x\mathrm{d}y, \tag{13}$$

where in the integrand x and y represent the frequency of the advantageous alleles at the first and second loci, and the known constants a, b, c and d depend on the mutation rates and sample configuration.

Our approach to calculating integrals of this form is based on ideas from Fearnhead and Meligkotsidou (2004) and depends on whether or not $\sigma_1 > \sigma_2$. For each case we introduce the following notation

$$R_x[i,j] = \int_0^1 x^{a+i-1}(1-x)^{b+j-1}\mathrm{d}x, \text{and}$$

$$R_y[i,j] = \int_0^1 y^{c+i-1}(1-y)^{d+j-1}\mathrm{d}y,$$

which are standard Beta constants.

Case 1: $\sigma_2 \geq \sigma_1$

In this case we can simplify and then expand the exponential in (13) as

$$\exp(\sigma_1(x+y) + (\sigma_2 - \sigma_1)xy) = \sum_{k=0}^{\infty} \frac{(\sigma_2 - \sigma_1)^k}{k!} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \frac{\sigma_1^{i+j}}{i!j!} x^{i+k} y^{j+k}.$$

Thus (13) can be written as

$$\sum_{k=0}^{\infty} \frac{(\sigma_2 - \sigma_1)^k}{k!} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \frac{\sigma_1^{i+j}}{i!j!} R_x[i+k, 0] R_y[j+k, 0].$$

Finally writing

$$A_k = \sum_{i=0}^{\infty} \frac{\sigma_1^i}{i!} R_x[i+k, 0], \text{and } B_k = \sum_{j=0}^{\infty} \frac{\sigma_1^j}{j!} R_y[j+k, 0],$$

we get that (13) is equal to

$$\sum_{k=0}^{\infty} \frac{(\sigma_2 - \sigma_1)^k}{k!} A_k B_k. \tag{14}$$

We first evaluate A_k and B_k by truncating their infinite sums, and then truncate this infinite sum to evaluate (13). All sums are sums of positive terms, and are thus stable to evaluate, and the terms in the sums decay exponentially for sufficiently large values of the sums. Furthermore a look-up table of the A_ks and B_ks can be constructed to speed up the calculation of (14).

Case 2: $\sigma_1 > \sigma_2$

We now write and expand the exponential in (13) as

$$\exp(\sigma_1 y + (\sigma_1 - \sigma_2)x + \sigma_2 x(1-y)) = \sum_{k=0}^{\infty} \frac{\sigma_2^k}{k!} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \frac{(\sigma_1 - \sigma_2)^i \sigma_1^j}{i!j!} x^{i+k} y^j (1-y)^k.$$

Thus (13) can be written as

$$\sum_{k=0}^{\infty} \frac{\sigma_2^k}{k!} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \frac{(\sigma_1 - \sigma_2)^i \sigma_1^j}{i!j!} R_x[i+k, 0] R_y[j, k].$$

Finally writing

$$C_k = \sum_{i=0}^{\infty} \frac{(\sigma_1 - \sigma_2)^i}{i!} R_x[i+k, 0], \text{and } D_k = \sum_{j=0}^{\infty} \frac{\sigma_1^j}{j!} R_y[j, k],$$

we get that (13) is equal to

$$\sum_{k=0}^{\infty} \frac{\sigma_2^k}{k!} C_k D_k.$$

This can be evaluate in an equivalent way to (14), and again is stable as it requires sums of positive terms which decay exponentially. Calculation of this sum can be made efficient by constructing a look-up table for the C_k and D_k terms.

Appendix C: Backward Rates at Selection events

Consider a selection event forward in time. There are 8 possible configurations of continuing, incoming and linked-incoming branch, and for each configuration we can write down the probability of the new branch being of type 1 or 2. These probabilities multiplied by the unconditional rate of selection events per branch $(\sigma_1 + \sigma_2)/2$ are summaried in Table 3

To get conditional backward rates involves (i) summing up the rates of selection events per branch that produce the correct offsrping, and have the correct type of virtual branch at each locus; and (ii) multiplying this by the number of branches of the correct type and the ratio of the stationary probabilities of the new and old states of the CSG.

So to obtain the rate of event S2a (selection to branch of type 2, producing virtuals of type 1 at each locus) in Table 1, for (i) we get contributions of $\sigma_1/2$ and $(\sigma_1 + \sigma_2)/2$ from respectively lines 3 and 5 of the above table; and for (ii) we get a factor of $n_2^{(1)} \pi(n_1^{(1)} + 1, n_2^{(1)}, n_1^{(2)} + 1, n_2^{(2)})/\pi(A)$. The rates in (i) are added and their sum multiplied by the factor in (ii) to obtain the required rate. Other entries in Table 1 are obtained similarly.

Continuing	Incoming	Linked Incoming	1	2
1	1	1	$(\sigma_1 + \sigma_2)/2$	0
1	1	2	$(\sigma_1 + \sigma_2)/2$	0
1	2	1	$\sigma_2/2$	$\sigma_1/2$
1	2	2	0	$(\sigma_1 + \sigma_2)/2$
2	1	1	0	$(\sigma_1 + \sigma_2)/2$
2	1	2	$\sigma_1/2$	$\sigma_2/2$
2	2	1	0	$(\sigma_1 + \sigma_2)/2$
2	2	2	0	$(\sigma_1 + \sigma_2)/2$

Table 3: Rates of specific selection events producing branches of type 1 and 2 (forward in time)