Perfect Simulation from Non-neutral Population Genetic Models: Variable Population Size and Population sub-division

Paul Fearnhead[1]

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

Summary: We show how the idea of monotone coupling from the past can produce simple algorithms for simulating samples at a non-neutral locus under a range of demographic models. We specifically consider a biallelic locus, and either a general variable population size mode, or a general migration model for population subdivision. We investigate the effect of demography on the efficacy of selection, and the effect of selection on genetic divergence between populations.
Keywords: *Ancestral Selection Graph, Coupling from the Past, Demes, Frequency Spectrum, Migration model, Stepping-stone model*

1 Introduction

While simulating from neutral population genetics models is straightforward using the coalescent (Kingman, 1982), even for complex demographic models (see e.g. Donnelly and Tavaré, 1995; Hudson, 2002), simulation from non-neutral models is much more difficult. There are currently numerous approaches to simulation for non-neutral models. These can be split into two cases, simulating from the stationary distribution of the population, or simulating from the population at a specfic time. Examples of the latter include simulating samples a specific time after the fixation of a beneficial allele (e.g. Przeworski, 2003). In this paper we focus on the former.

For non-neutral models which assume parent-independent mutation, constant population size and random-mating, the stationary distribution of allele frequencies is known and can be simulated from directly using rejection sampling (Donnelly *et al.*, 2001) or numerical integration methods (Fearnhead and Meligkotsidou, 2004; Joyce and Genz, 2006). For such models it is also possible to simulate the genealogy of a sample, and linked neutral variation. This is possible by simulating and conditioning on the frequency of the non-neutral allele in the history of the population(Spencer and Coop, 2004; Coop and Griffiths, 2004; Nordborg, 2001; Nordborg and Innan, 2003), or by simulating from the conditional distribution of the ancestral selection graph given the allelic type of the sample (Slade, 2000; Stephens and Donnelly, 2003; Fearnhead, 2006)

For models which do not assume parent-independent mutations, it is possible to use coupling-from-the-past (CFTP) (Propp and Wilson, 1996; Kendall, 2005) to simulate samples from the ancestral selection graph (Fearnhead, 2001). Here we extend this idea. Firstly we introduce a simpler implementation of CFTP, which is obtained by having a state-space of unordered rather than ordered samples. This means that we characterise a sample by the number of alleles of each type, rather than by the allelic type of each of an ordered set of chromosomes. This has two advantages, firstly that as we work on a smaller dimensional space, coupling should be quicker. Secondly, for the models we consider, by working with unordered samples we obtain a monotonicity of the sample space, which makes it more straightforwad to detect when we have obtained a sample from the stationary distribution of interest, and thus provides a much simpler algorithm to implement.

Our second extension is to allow for a variety of demographic models, including arbitrary variable population size models, and models with multiple sub-populations (demes). As far as we are aware, the method we propose is the only current method for simulating

from such multiple deme coalescent models in the presence of selection. Our method is computationally efficient, with for example 1,000 samples of size 100 being simulated in less than 40s for a 10 deme stepping-stone model under a variety of selection models (see RESULTS).

2 Methods

Monotone Coupling From The Past

Consider an ergodic Markov chain X_t with state space $\{0, 1, \ldots, K\}$. Coupling from the past (CFTP; Propp and Wilson, 1996) gives a method for simulating from the stationary distribution of X_t. The idea is based on the fact that if we simulated the Markov chain from time $-\infty$ to time 0, then regardless of the initial value of the chain, $X_{-\infty}$, we would have that X_0 is a draw from the stationary distribution of the chain. The idea of CFTP is that it enables us to perform such simulation in finite computing time.

To do this we first introduce some extra simulation, in that we consider simulating the value of X_{t+1} for all possible values of X_t. To simplify notation we will define a function $F_t(\cdot)$ which specifies all these transitions. So if we are told that $X_t = x$, then we have $X_{t+1} = F_t(x)$. Note that the function F_t is a realisation of a random variable; the stochasticity of the Markov chain is now encompassed in the randomness of this function, but given a set of realisations $F_{-T}, F_{-T+1}, \ldots, F_{-1}$ we have a deterministic relationship between X_{-T} and X_0.

We now have an idealised simulation algorithm:

(a) For $t = -1, -2, \ldots, -\infty$ simulate the value of X_{t+1} for all possible values of X_t; and hence the function $F_t(\cdot)$.

(b) Arbitrarily specify $X_{-\infty}$; and for $t = -\infty, \ldots, -1$ recursively apply $x_{t+1} = F_t(x_t)$ to obtain x_0, a draw from the stationary distribution of the Markov chain.

This algorithm is obviously impracticable, as it involves infinite computing time. However we can perform this simulation in finite computing time by doing this simulation a bit at a time. This gives the following CFTP algorithm:

(a) Arbitrarily choose a negative integer T_0; set $T = T_0$ and $S = 0$.

(b) For $t = S - 1, S - 2, \ldots, T$, simulate F_t.

(c) For each possible starting value $x = 0, 1, \ldots, K$, set $X_T = x$ and recursively apply $x_{t+1} = F_t(x_t)$ to obtain x_0. If the value of x_0 is identical for all starting values of X_t then output x_0, a draw from the stationary distribution of the Markov Chain; otherwise set $S = T$, $T = 2T$ and return to (b).

The idea here is that we imagine that we are doing the idealised simulation algorithm above. However, we first simulate the dynamics of the chain from time T_0 to time 0; then the extra dynamics from time $2T_0$ to time T_0; then the extra dynamics from time $4T_0$ to times $2T_0$ and so on (step b). The *coupling* condition in step (c) enables us to determine with certainty what the value of X_0 would be if we had continue to simulated the dynamics of the chain back to time $-\infty$ (as in the idealised algorithm). For example, imagine that the condition in step (c) is satisfied after simulating back to time $4T_0$; this condition says that regardless of the value of X_{4T_0}, the value of X_0 will be the same (x_0). Thus we do not need to know the realisation of the chain from time $-\infty$ to time $4T_0$, as regardless of this we know that continuing the realisation on to time 0 will produce $X_0 = x_0$. Thus the value we output in step (c) is the same as the value we would obtain in step (b) of the idealised simulation algorithm, and is thus a draw from the stationary distribution of the Markov chain.

Any choice for how to decrease T when coupling does not occur would produce a valid algorithm, but it has been argued that doubling T each time is optimum (by its relationship with a binary search; see e.g. Kendall, 2005). Note that the validity of the approach requires that we do not resimulate any of the dynamics of the Markov chain if coupling does not occur in step (c). Furthermore, the algorithm requires only that the functions $F_t(\cdot)$ are simulated in such a way that marginally the dynamics of the Markov chain are correct, and it is possible to have any amount of dependence in terms of the value of X_{t+1} for different values of X_t. In some cases it is possible to utilise this to obtain a more efficient way of determining whether coupling occurs in step (c). If we can find a distribution on F_t, which marginally has the dynamics of the Markov chain, and that staisfies the monotonicty condition that $x \geq y \Rightarrow F(x) \geq F(y)$ then we can replace step (c) with:

(c) For both $x = 0$ and $x = K$, set $X_T = x$ and recursively apply $x_{t+1} = F_t(x_t)$ to obtain x_0. If the value of x_0 is identical for both starting values of X_t then output x_0, a draw from the stationary distribution of the Markov Chain; otherwise set $S = T$, $T = 2T$ and return to (b).

We call the resulting algorithm monotone CFTP. The only difference is that we only run

the Markov chain forward in time for the minimal and maximal elements of the state-space. The monotonicity ensures that if these two realisations couple, then all realisations from other starting points (which lie between them) will also couple.

While in this example we have assumed a totally ordered state-space, monotone CFTP will still apply if we have only a partial order: if starting the chain at each of the maximal and minimal elements of our state-space produce the same x_0 value, then so must all other starting values of the chain (which lie between the minimal and maximal elements). We will use this extension to partial orderings in the multi-deme models we consider below.

An example of the CFTP algorithms is given in Figure 1.

Models

We consider coalescent models which include selection, variable population size and population structure, and focus on a single biallelic locus. Details of how selection is incorporated within a coalescent-type process can be found in work on the Ancestral Selection Graph (Krone and Neuhauser, 1997; Neuhauser and Krone, 1997) or the Ancestral Influence Graph (Donnelly and Kurtz, 1999). For background on incorporating population growth or population structure see for example Donnelly and Tavaré (1995) and Hudson (1990) and references therein.

The coalescent model we consider is obtained in the large population size limit to a range of forward population genetics model. We briefly describe how the parameters are defined in the case of the Wright-Fisher model. Firstly define an effective population size of N_0 diploid individuals, or equivalently $2N_0$ chromosomes. We measure time in units of $2N_0$ generations, and as is common define time in terms of time before the present. We consider a population consisting of D demes, with random mating within each deme. At time t in the past, the population size of deme d is $N_d(t) = N_0\lambda_d(t)$ diploid individuals. For the case $D > 1$ we further define a migration matrix by $M_{ij} = 4N_0m_{ij}$ for $i \neq j$, where m_{ij} is the proportion of the population of deme i in a generation that have migrated there from deme j in the previous generation. We define $M_i = \sum_{j\neq i} M_{ij}$.

We denote the alleles at our locus by 1 and 2. As with any biallelic model, mathematically we can describe the mutation process in terms of parent-independent mutations. We let $\theta = 4N_0u$, where u is the probability of mutation per chromosome per generation, and let $\nu = (\nu_1, \nu_2)$ be the probability distribution of the mutant allele. (Note that this model allows for "silent" mutations, which do not change the allele at the locus; thus the effective mutation rate of allele 2 to allele 1 is $\theta\nu_1$, and of allele 1 to allele 2 is $\theta\nu_2$.)

Finally we define the selection process. We assume that the fitness of an ij genotype is

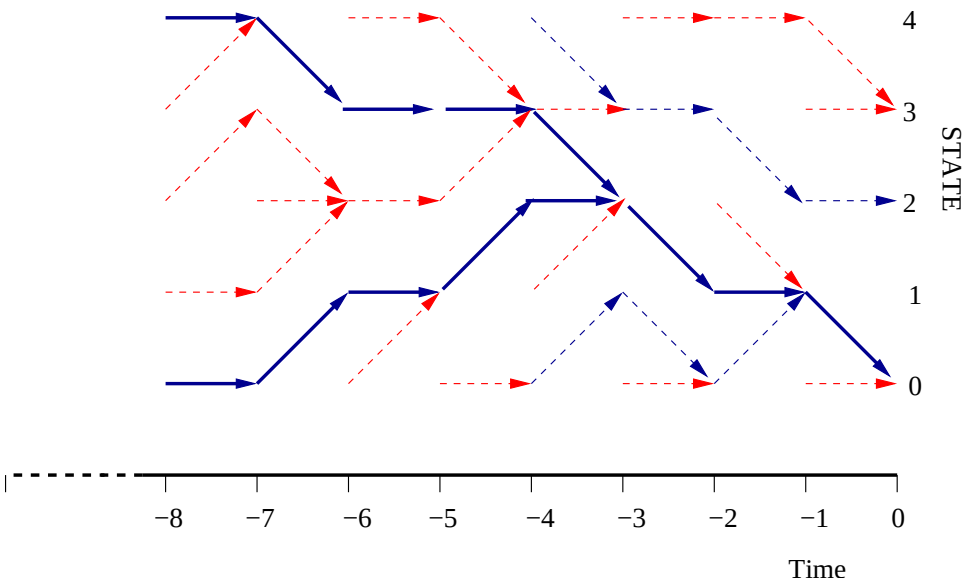

Figure 1: Example of monotone CFTP for a 5-state Markov chain. For each state at each time we imagine simulating the transitions of the Markov chain (red dashed lines). (Informally the monotonicity can be seen by the fact that no pair of transitions cross each other.) Initially in this example we actually simulate these from times $t = -1, -2, \ldots, -4$. We then simulate forward in time from the maximal (4) and minimal (0) states at $t = -4$ (blue dashed lines). As simulating forward these states produce different samples, we simulate the transitions for times $t = -5, \ldots, -8$, and repeat simulating forward in time from the maximal and minimal states at $t = -8$. These both give the value $X_0 = 0$, and thus the Markov chain has coupled and we have a draw from the stationary distribution of the chain. (Notice that regardless of what value X_{-8} is, forward simulation produces the value $X_0 = 0$, and thus this is the value we would have obtained if we had simulated the Markov chain from time $t = -\infty$.)

$1 + s_{ij}$ for $i, j = 1, 2$, with $s_{ij} \geq 0$ and $s_{12} = s_{21}$, We can thus define selection rates $\sigma^*_{ij} = 4N_0 s_{ij}$. We choose a different parameterisation of these selection rates where $\sigma^*_{11} = \sigma_{11}$, $\sigma^*_{12} = \sigma_g + \sigma_{12}$ and $\sigma^*_{22} = 2\sigma_g + \sigma_{22}$. This is an over-parameterisation of the selection process, and any choice of $\sigma_g \geq 0$ and $\sigma_{ij} \geq 0$ for $i, j = 1, 2$ which gives the correct values of σ^*_{ij} could be used. We then define $\sigma = \max\{\sigma_{11}, \sigma_{12}, \sigma_{22}\}$.

The choices of defining the mutation process in terms of a parent-independent mutation process, and the parameterisation of selection rates, have been made in order to make the simulation algorithm detailed below as efficient as possible. In particular the parameterisation of the selection rates is in terms of a genic component (given by the σ_g terms) and a non-genic component. The efficiency of the simulation algorithm is increased by choosing σ_g to be as large as possible subject to the constraints on the positivity of the other selection rates. (To do this we should label the alleles so that $\sigma^*_{22} \geq \sigma^*_{11}$.)

For ease of presentation we have assumed that the selection rates are constant through time and across demes; though it is straightforward to generalise to the case where these rates vary. The coalsescent process for our above model can be described in terms of a backward process for the history of our sample (which is independent of the alleles in the sample) and conditional on this backward process some forward dynamics for the allelic types of the branches. If we simulated the backward process until time $t = \infty$, then for any initial choice for the population frequency of the alleles at that time, when we simulated the process forward we would obtain a sample at the present time which is drawn from the stationary distribution of the population. We now describe these backward and forward processes and how we can appply monotone CFTP.

Backward Process

The Backward Process is a continuous time Markov Chain, whose state $\mathbf{N(t)} = (\mathbf{N_1(t)}, \ldots, \mathbf{N_D(t)})$ is the number of branches in the underlying coalscent-type process at time t in the past that come from each deme. The transitions correspond to different possible events in this coalescent-type process. To ease notation, let the state at current time t be $(n_1, n_2, \ldots, n_D)$, then the rates of the possible events, and the corresponding transitions, are:

(i) **Coalescence, deme** d occurs at rate $n_d(n_d-1)/(2\lambda_d(t))$; with transition $n_d = n_d - 1$.

(ii) **Migration, deme** d **to deme** i occurs at rate $n_d M_{di}/2$;with transition $n_d = n_d - 1$ and $n_i = n_i + 1$.

(iii) **Mutation deme** d occurs at rate $n_d\theta/2$; no change to state.

(iv) **Genic Selection, deme** d occurs at rate $n_d\sigma_g/2$; with transition $n_d = n_d + 1$.

(v) **Diploid Selection, deme** d occurs at rate $n_d\sigma/2$; with transition $n_d = n_d + 2$.

(When describing the transition we have listed only the elements of the state that change.) The initial value of the state, $\mathbf{N}(\mathbf{0})$ is given by the number of chromosomes sampled from each deme.

Forward Dynamics and Monotone CFTP

Assume we have simulated and stored T events in the backward process above. We now consider the forward in time dynamics; however in order to apply CFTP we need to make these forward dynamics deterministic, and thus for each event that we simulate in the backward process we also simulate and store a realisation of an independent continuous uniform $[0, 1]$ random variable. These realisations will then determine the specific forward transition at each event.

Forward in time the state of our system is given by the number of each type of allele in each deme in our ancestral process. As we have stored the number of branches in each deme at each event, we can define this state solely in terms of the number of type 1 alleles in each deme; which we denote by $\mathbf{n}^{(1)} = (n_1^{(1)}, \ldots, n_D^{(1)})$. Our forward in time dynamics are determined by specifing an initial value for the state $\mathbf{n}^{(1)}$. Then for the Tth event, $T-1$th event, ..., 1st event in turn we update this state as follows.

Consider the jth event. Let u denote the realisation of the uniform random variable associated with this event. Assume the current state is $\mathbf{n}^{(1)}$, and the number of branches in each deme is $\mathbf{n} = (n_1, \ldots, n_D)$. Then the dynamics depend on the type of the jth event as follows:

(i) **Coalescence, deme** d; if $u < n_d^{(1)}/n_d$ then let $n_d^{(1)} = n_d^{(1)} + 1$.

(ii) **Migration, deme** d **to deme** i; if $u < n_i^{(1)}/n_i$ then let $n_d^{(1)} = n_d^{(1)} + 1$ and $n_i^{(1)} = n_i^{(1)} - 1$.

(iii) **Mutation deme** d; if $u < (n_d - n_d^{(1)})\nu_1/n_d$ then let $n_d^{(1)} = n_d^{(1)} + 1$; if $u > 1 - n_d^{(1)}\nu_2/n_d$ then let $n_d^{(1)} = n_d^{(1)} - 1$.

(iv) **Genic Selection, deme** d; if $u < 1 - (n_d - n_d^{(1)})(n_d - n_d^{(1)} - 1)/(n_d(n_d - 1))$ then let $n_d^{(1)} = n_d^{(1)} - 1$.

(v) **Diploid Selection, deme** d; if

$$u < \frac{n_d^{(1)}(n_d^{(1)}-1)(n_d^{(1)}-2+(\sigma-\sigma_{11}+\sigma_{12})(n_d-n_d^{(1)})/\sigma)}{n_d(n_d-1)(n_d-2)}, \tag{1}$$

then let $n_d^{(1)} = n_d^{(1)} - 2$; otherwise if

$$u < 1 - \frac{(n_d-n_d^{(1)})(n_d-n_d^{(1)}-1)(n_d-n_d^{(1)}-2+(\sigma-\sigma_{22}+\sigma_{12})n_d^{(1)}/\sigma)}{n_d(n_d-1)(n_d-2)}, \tag{2}$$

then let $n_d^{(1)} = n_d^{(1)} - 1$.

We have described the dynamics purely in terms of changes in $\mathbf{n}^{(1)}$, the changes in the number of branches in each deme is given by the reverse of the dynamics of the backward process. These dynamics come from the different possible events in the coalescent process that would affect $\mathbf{n}^{(1)}$. For (i) this is a coalescent event to branch of allele 1; for (ii) a migration of an allele 1 branch from population i to d; for (iii) a mutation of an allele 2 branch to allele 1, or vice-versa; for (iv) a selection event at which an allele 1 branch is non-ancestral (which occurs unless both incoming and continuing branches have allele 2); and for (v) a selection event at which both non-ancestral branches have allele 1, or only one has. (See Appendix A for the calculation of the probabilities in this case).

We can now use CFTP to simulate a sample from the stationary distribution of the population. The dynamics specified above satisfy a monotonicity condition (see Appendix B) if $2\sigma_{21} \leq \sigma + \sigma_{11} + \sigma_{22}$. (This can always be achieved by, if necessary, choosing $\sigma > \max\{\sigma_{11}, \sigma_{12}, \sigma_{22}\}$; for example if $\sigma_{11} = \sigma_{22} = 0$ as in heterozygote advantage then we choose $\sigma = 2\sigma_{12}$.) The partial ordering of this monotonicity is that $(n_1, \ldots, n_D) \geq (n'_1, \ldots, n'_D)$ if and only if $n_d \geq n'_d$ for $d = 1 \ldots, D$.

Thus the monotone CFTP algorithm described above can be applied, whereby to check coupling we need only run the process forward in time from two values of the state: $\mathbf{n}^{(1)} = (n_1, \ldots, n_D)$, all branches in all demes carry allele 1, and $\mathbf{n}^{(1)} = (0, \ldots, 0)$, no branches in any deme carry allele 1. If we obtain the same sample from each of these initial conditions, then that sample is drawn from the population at stationarity. If not we have to simulate the backward process further into the past and repeat until coupling occurs.

Programs implementing CFTP for both single-population variable population size and for multiple-deme constant-population size models were written in a combination of R and C. These are available from `www.maths.ac.uk/~fearnhea`.

Verifying the CFTP Results

To check the validity of the programs implementing this monotone CFTP algorithm we ran the programs under two special cases. Firstly we considered the neutral case, and compared the results of our program with those obtained by the program `ms` (Hudson, 2002). Note that this program defines time in units of N_0 generations, rather than the $2N_0$ used here. It assumes an infinite sites mutation model - this output can be converted into a 2-allele model with $\nu = (0.5, 0.5)$ by (i) setting the mutation rate it $\theta/2$, and (ii) assuming each mutation changes the allele of the chromosome so that the allele of a chromosome is determined by whether it has an odd or an even number of mutations. For the constant population size and a single panmictic population ($D = 1$) we compared results with those based on simulating from the known stationary distribution of the population frequency of allele 1 using rejection sampling (see Donnelly *et al.*, 2001).

3 Results

Single-Population Model

First we used the CFTP algorithm to simulated from a series of single population models. There are many demographic models that have been suggested or inferred for human or other populations (e.g. Wakeley *et al.*, 2001; Wall *et al.*, 2002; Marth *et al.*, 2004; Schaffner *et al.*, 2005). We considered four different scenarios for the variation in population size, and four different selection models. In each case we set $N_0 = 10,000$ diploid individuals, $\theta = 1$ and $\nu = (0.9, 0.1)$ (based loosely on appropriate mutation models for disease genes; see Pritchard, 2001). The four population size scenarios are:

Constant A constant population size.

Growth An exponentially growing population with $\lambda(t) = \exp(-0.7t)$ (based on an inferred model for beta-globin, see Harding *et al.*, 1997).

Bottleneck A population bottleneck from $t = 0.15$ to $t = 0.175$. During the bottleneck the effective population size is 1,000, and prior to it it is 5,000. (based on a model from Marth *et al.*, 2004).

Complex A more complicated scenario based loosely on the population-size of a non-African population in the model of Schaffner *et al.* (2005). It includes recent exponential growth and a bottleneck.

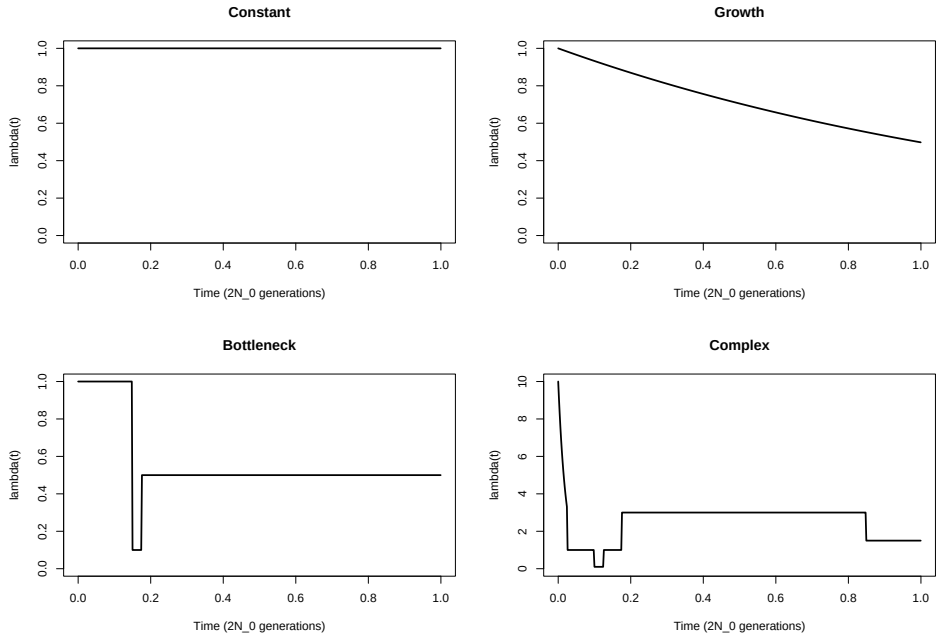

Figure 2: Plots of $\lambda(t) = N(t)\mathrm{N}_0$ for the four population-size scenarios.(Note the diffent scale for the Complex scenario.)

Plots of $\lambda(t) = N(t)/N_0$ for each of these models are given in Figure 2
The selection models are defined in terms of the selection rates $\sigma^* = (\sigma^*_{11}, \sigma^*_{12}, \sigma^*_{22})$. Our four selection models are (i) Neutral $\sigma^* = (0, 0, 0)$; (ii) Genic $\sigma^* = (0, 5, 10)$; (iii) Heterozygote advantage $\sigma^* = (0, 10, 0)$; and (iv) Heterozygote overdominance $\sigma^* = (0, 20, 10)$. For each combination of population size scenario and selection model we simulated 10,000 samples. The CPU cost varied across the 16 pairs we considered. In the constant population-size case, on a 3.4GHz Laptop, it took 0.3s, 1s, 15s and 25s to simulate 1,000 samples for the four selection models respectively. CPU times were reduced by around a factor of 2 for the Growth and Bottleneck scenarios, and were increased by around a factor of 1.5 for the Complex scenario.
Histograms of the frequency of allele 1 in a sample of size 50 (conditional on the allele segregating) are given in Figure 3. These histograms agree with other simulation results for the cases which include constant population size (see Verifying the CFTP Results). The different population size scenarios have little effect on the frequency spectrum in the neutral case, but quite a noticeable effect for each of other selection scenarios. The effect is

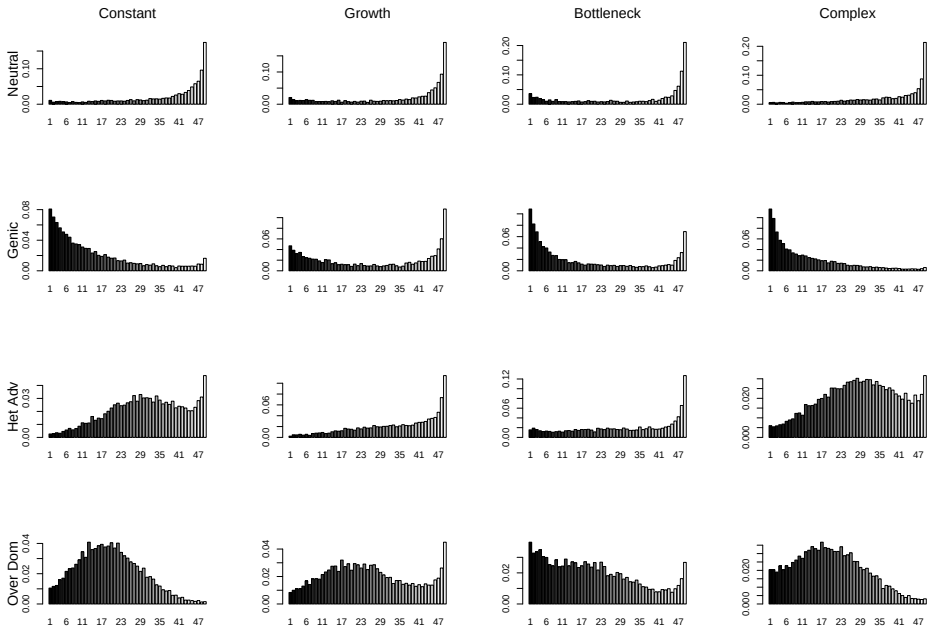

Figure 3: Histograms of the frequency of allele 1 in a sample of size 50 (conditional on the allele segregating) for each pair of variable population size scenario and selection model. See text for full details of the models.

most pronounced for the Heterozygote Advantage case where the Growth and Bottleneck scenarios have substantially reduced any effect of selection.

Two Deme Model

We now consider a model based on two demes each containing a constant-sized panmictic populations with migration between the demes. For simplicity we assume each deme has the same population size, and idential migration rate from deme 1 to deme 2 and vice versa. We assume a total sample of size 100, with 50 chromosomes sampled from each deme. We assume $\nu = (0.5, 0.5)$ and present results for $\theta = 0.1$. We obtain very similar results for smaller values of θ except that the probability of the allele segregating in the population is reduced.

We present results for four selection models and four migration rates. Again we summarise the selection models in terms of $\sigma^* = (\sigma^*_{11}, \sigma^*_{12}, \sigma^*_{22})$. Our four selection models are (i) Neutral $\sigma^* = (0, 0, 0)$; (ii) Genic $\sigma^* = (0, 5, 10)$; (iii) Heterozygote advantage $\sigma^* = (0, 10, 0)$; and (iv) Recessive $\sigma^* = (0, 10, 10)$. The four migration rates (the values of $M_{12} = M_{21}$) are 10, 2, 0.5 and 0.1.

We simulated 10,000 samples for each of the 16 combinations of selection model and migration rate. We verified the results for the neutral model with other simulation methods (see Verifying the CFTP Results). The CPU cost of the simulation varied little with migration, and was approximately 0.4s, 0.7s, 10s and 4s per 1,000 samples for the four selection models respectively (CPU times for a 3.4GHz Laptop). The frequency spectrum for samples of size 50 within a single deme, conditional on the allele segregating are shown in Figure 4. The amount of migration appears to have little effect except in the case of heterozygote advantage, when smaller migration rates appear to reduce the effect of selection.

We also studied the effect of selection and the degree of divergence in gene frequency between the demes. Table 1 gives the mean Fst values for samples under a SNP ascertainment model which requires 2 randomly chosen chromosomes (across both demes) to segregate. If the observed frequency of type 1 alleles in the demes were p_1 and p_2 respectively, and $p = (p_1 + p_2)/2$, then the Fst value for that sample was $(p_1 - p)^2/(p(1-p))$ (see e.g Section 2.3 of Nicholson *et al.*, 2002). The weight given to a sample is proportional to $p(1-p)$, and gives more importance to samples whose minor allele frequency is close to 0.5.

As noted by Hughes *et al.* (2005), the effect of selection is to reduce the amount of variation in allele frequencies, and hence Fst values, across demes. The effect is most pronounced as migration rates are decreased. (Measuring the variation in allele frequencies using the d-statistic of Nei (1987) produced similar results.)

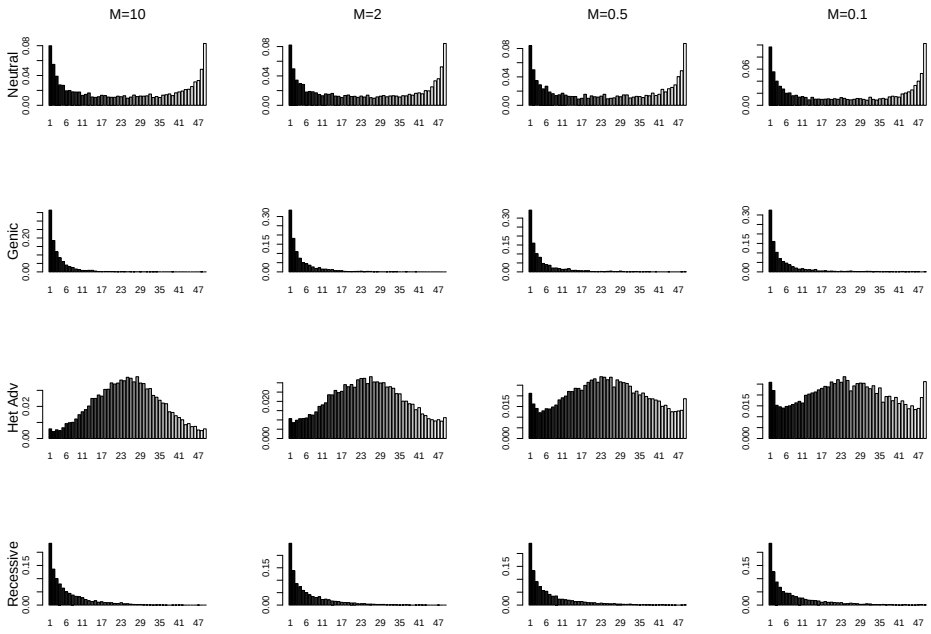

Figure 4: Histograms of the frequency of allele 1 in a sample of size 50 from a single deme (conditional on the allele segregating) for each pair of migration rate and selection model. See text for full details of the models.

	$M = 10$	$M = 5$	$M = 0.5$	$M = 0.1$
Neutral	0.035	0.123	0.326	0.650
Genic	0.031	0.070	0.127	0.218
Het Adv	0.031	0.080	0.159	0.295
Recessive	0.033	0.083	0.149	0.199

Table 1: Mean Fst values (based on 10,000 samples) under a 2-deme model for each combination of selection model and migration rate (M). A weighted mean was calculated with the Fst value from each sample being weighted by the probability that 2 randomly chosen chromosomes carry different alleles. See text for full details of the models.

	$M = 10$	$M = 5$	$M = 0.5$	$M = 0.1$
Genic	0.036	0.122	0.367	0.706
Het Adv	0.034	0.110	0.253	0.439
Recessive	0.033	0.115	0.330	0.654

Table 2: Mean Fst values (based on 10,000 samples) under a 2-deme model with different selection regimes within each deme. We assumed neutrality within deme 1, and either the genic, heterozygote advantage or recessive selection model for deme 2. Results are given for different migration rates (M). A weighted mean was calculated with the Fst value from each sample being weighted by the probability that 2 randomly chosen chromosomes carry different alleles. See text for full details of the models.

We further tested the effect of different selection regimes within the two demes. For simplicity we assumed neutrality within deme 1, and the three selection models above within deme 2. The mean Fst values in this case are given in Table 2. As expected Fst values increased as compared to the case of the same selection regime within each population, though only in the genic selection case are Fst values consistently greater than in the completely neutral case.

Stepping Stone Model

We now consider a linear (circular) stepping stone model . We assume 10 ordered demes, each of equal population size. Migration events are possible between neighbouring demes, with $M_{i,i+1} = M_{i+1,i} = 1$ for $i = 1, \ldots, 9$ and $M_{1,10} = M_{10,1} = 1$. We consider the same mutation model and each of the four selection models used in the 2 deme case, and simulate samples of size 100, with 50 chromosomes sampled from deme 1, and 50 from deme $1 + c$ for $c = 1$, 2, 3, and 4. We are interested in the degree of correlation in allele frequencies for differing degrees of physical separation of the demes (c) and differing selection models. Again we simulated 10,000 samples for each combination of selection model and c. CPU costs were affected only slightly by c and were 0.5s, 1,4s, 40s and 20s per 1,000 samples for the neutral, genic, heterozygous advantage, and recessive selection models respectively (on a 3.4GHz Laptop). We calculated Fst values in the same way as for the 2 deme case (above), once again under a SNP ascertainment scheme that requires two randomly chosen chromosomes from our sample of size 100 to be carrying different alleles.

Results are given in Table 3. Again the effect of selection is to reduce the amount of variation in allele frequencies between demes. Most strikingly, for these selection models

	$c=1$	$c=2$	$c=3$	$c=4$
Neutral	0.19	0.27	0.32	0.34
Genic	0.09	0.10	0.10	0.10
Het Adv	0.09	0.11	0.11	0.12
Recessive	0.11	0.12	0.13	0.14

Table 3: Mean Fst values (based on 10,000 samples) under a stepping-stone model for each combination of selection model and spatial separation of the two demes (c). A weighted mean was calculated with the Fst value from each sample being weighted by the probability that 2 randomly chosen chromosomes carry different alleles. See text for full details of the models.

	$c=1$	$c=2$	$c=3$	$c=4$
$\sigma_g=1$	0.16	0.22	0.23	0.25
$\sigma_g=2$	0.13	0.17	0.19	0.17
$\sigma_g=3$	0.11	0.13	0.14	0.14
$\sigma_g=4$	0.10	0.11	0.12	0.11

Table 4: Mean Fst values (based on 10,000 samples) under a stepping-stone model for genic selection and different spatial separation of the two demes (c). A weighted mean was calculated with the Fst value from each sample being weighted by the probability that 2 randomly chosen chromosomes carry different alleles. See text for full details of the models.

the amount of variation in allele frequencies depends little on the spatial separation of the demes, which is not the case for the neutral case. We further investigated this effect by simulating samples under the genic model for a range of smaller selection values; $\sigma_g = 1$, 2, 3 and 4 (see Table 4). The effect that Fst values depend little on spatial separation is noticeable for $\sigma_g \geq 2$.

4 Discussion

We have presented a method for simulating samples from the stationary distribution of a class of non-neutral population genetic models. The method is simple compared to other approaches based on CFTP (Fearnhead, 2001) due to the monotonicity inherent in the

problems we consider. Thus simulating samples only requires (i) to simulate the number of branches within the ancestral selection graph back in time; and (ii) to simulate the number of these branches (within each deme) that carry allele 1 forward in time for two initial configurations: all branches initially carrying allele 1 and all branches initially carrying allele 2. The resulting algorithm has been shown to be computationally efficient, and enables simulation of a large number of samples from complex selection and demographic models within practicable CPU time. The main limitation on the simulation algorithm is the mutation rate θ, with the CPU cost appearing to increase proportional $1/\theta$ for small theta. Thus simulation for very small values of the mutation rate can be prohibitive.

The monotonicity characteristic of our models occurs because we consider only a single selective locus which carries two different alleles. For more general models, monotonicity will not necesarily apply, but simulating unordered samples as here should still be easier and more efficient than simulating ordered samples as in Fearnhead (2001).

The examples we considered either assumed a single population and variable population size, or constant population size and multiple demes. There are alternative approaches for analysing and simulating under the former class of models: in some cases direct simulation is possible by simulating the ancestral selection graph back in time until there is a constant population size and then simulating the alleles on the branches at that time from the known stationary distribution of the constant population size model. Alternatively there is recent work by Evans *et al.* (2006) which calculates the frequency spectrum under variable population size models and an infinite sites mutation model, and the method used with `SelSim` (Spencer and Coop, 2004) could easily be adapted to this situation.

However, simulating samples from non-neutral models with multiple demes is much more challenging, and we know of no other current coalescent-based simulation methods in this case. Some methods exist for special cases, for example diffusion approximations for the island model in the limit as the number of demes tends to infinity (Cherry and Wakeley, 2003; Cherry, 2003).

Acknowledgements Motivation for this work came from the ICMS workshop on Mathematical Population Genetics, March 2006. This research was supported by Engineering and Physical Sciences Research grant number C531558.

Appendix A: Diploid Selection Dynamics

Fix a deme and let n be the number of branches located in that deme and $n^{(1)}$ the number which carry allele 1. The dynamics at a diploid selection event are as follows (see Neuhauser and Krone, 1997): (i) choose 3 branches at random, call the first the incoming branch, the

secong the continuing branch and the third the checking branch; (ii) denote by ij the genotype given be the incoming and checking branch; and (iii) with probability σ_{ij}/σ the incoming branch is parental, otherwise the continuing branch is.

Firstly consider the probability of the non-ancestral branches at a diploid selection event both carrying allele 1. This corresponds to the first condition on u given by (1). This occurs if either (a) all three branches chosen in (i) carry allele 1, or (b) if two of these branches carry allele 1 and they are non-ancestral. The probability of (a) is $n^{(1)}(n^{(1)}-1)(n^{(1)}-2)/(n(n-1)(n-2))$. The probabilitiy of (b) is

$$\frac{3n^{(1)}(n^{(1)}-1)(n-n^{(1)})}{n(n-1)(n-2)}\left(\frac{\sigma_{12}}{3\sigma}+\frac{\sigma-\sigma_{11}}{3\sigma}\right),$$

where the first term is the probability of choosing 2 branches carrying allele 1 and one carrying allele 2; the second the term is the probability of the branch carrying allele 2 being the incoming branch and being ancestral; and the final term is the probability of the branch carrying allele 2 being the continuing branch and being ancestral.

Combining these probabilities gives the expression on the right-hand side of (1).

The right-hand side of (2) is one minus the probability of the two non-ancestral branches carrying allele 2, and is calculated in an identical manner. The probability of u satisfying (2) but not (1) is thus the required probability of the precisely one non-ancestral branches carrying allele 1.

Appendix B: Monotonicity

For notational simplicity we will drop the (1) superscipt for the state. Consider two values for the state $\mathbf{n}$ and $\mathbf{n}'$ which satisfy $\mathbf{n} \geq \mathbf{n}'$. To demonstrate monotonicity we need to show that for any possible event and value of u this ordering of the states is preserved.

Monotonicity at coalescence, mutation and genic selection events hold trivially. Assume one of these events occurs to a branch in deme d. Either $n_d = n'_d$ in which case the dynamics at this event are identical for both states; or $n_d \geq n'_d + 1$ in which case the ordering is preserved as the transition changes the n_d and n'_d values by either 0 or 1.

Monotonicity also follows for migration events. Consider a migration from deme i to deme d. We need only consider $n_i \neq n'_i$, as otherwise the dynamics at this event are identical for both states. However in this case the ordering is preserved in deme i by the same argument as above; and is also preserved in deme d as $n_d \geq n'_d$ and the dynamics mean that whenever n'_d increases by one then so does n_d.

Finally consider diploid selection in deme d. By the same arguments as above monotonicity trivially holds if $n_d = n'_d$ or $n_d \geq n'_d + 2$, so we focus on $n_d = n'_d + 1$. The key point is to

check that it is never possible for n'_d to be unchanged at this event at the same time as n_d being decreased by 2. Again for ease of exposition, we slightly change notation and denote the number of branches in deme d by n. For such a transition to occur we would need

$$u < \frac{n_d(n_d-1)(n_d-2+(\sigma-\sigma_{11}+\sigma_{12})(n-n_d)/\sigma)}{n(n-1)(n-2)}, \tag{3}$$

for the n_d to be decreased by 2 (see Equation 1) and

$$u > 1 - \frac{(n-n_d+1)(n-n_d)(n-n_d-1+(\sigma-\sigma_{22}+\sigma_{12})(n_d-1)/\sigma)}{n(n-1)(n-2)}, \tag{4}$$

for n'_d to be unchanged (using Equation 2 and $n'_d = n_d - 1$). Now let $a = (\sigma - \sigma_{11} + \sigma_{12})/(3\sigma)$ and $b = (\sigma - \sigma_{22} + \sigma_{12})/(3\sigma)$. Inequalities (3) and (4) can simultaneously hold for the same u if and only if

$$\frac{n_d(n_d-1)(n_d-2+3a(n-n_d))}{n(n-1)(n-2)} > 1 - \frac{(n-n_d+1)(n-n_d)(n-n_d-1+3b(n_d-1))}{n(n-1)(n-2)}.$$

Thus for monotonicity we need to show that this cannot occur, and thus that for all n and n_d

$$\begin{aligned} n(n-1)(n-2) \geq\ & n_d(n_d-1)(n_d-2) + 3n_d(n_d-1)(n-n_d)a + \\ & (n-n_d+1)(n-n_d)(n-n_d-1) + 3(n-n_d+1)(n-n_d)(n_d-1)b \end{aligned}$$

Now consider the right hand side; this can be re-written as

$$\begin{aligned} & n_d(n_d-1)(n_d-2) + 3n_d(n_d-1)(n-n_d) - 3n_d(n_d-1)(n-n_d)(1-a) + \\ & \quad (n-n_d)(n-n_d-1)(n-n_d-2) + 3(n-n_d)(n-n_d-1) + 3(n-n_d)(n-n_d-1)(n_d) + \\ & \quad 3(n-n_d)(3n_d-n-1) - 3(n-n_d+1)(n-n_d)(n_d-1)(1-b) \end{aligned}$$

Now the first, second, fourth and sixth terms of this expression sum to $n(n-1)(n-2)$, so we get that the inequality we required simplifies to

$$\begin{aligned} 0 \geq\ & -3n_d(n_d-1)(n-n_d)(1-a) + 3(n-n_d)(n-n_d-1) + \\ & \quad 3(n-n_d)(3n_d-n-1) - 3(n-n_d+1)(n-n_d)(n_d-1)(1-b) \\ =\ & 3(n-n_d)((n-n_d-1+3n_d-n-1) - (1-a)n_d(n_d-1) - (1-b)(n-n_d+1)(n_d-1)) \\ =\ & 3(n-n_d)(n_d-1)[2-(1-a)n_d - (1-b)(n-n_d+1)] \end{aligned}$$

Now this inequality trivially holds for $n_d = 1$ and for $n = n_d$. For larger values of n_d and $n - n_d$ the term in the square brackets is decreasing with both n_d and $n - n_d$ (as both $a < 1$ and $b < 1$). So we need only show that the inequality holds for $n_d = 2$ and $n - n_d = 1$.The square bracket term in this case is $2(a + b - 1)$; and the inequality $a + b - 1 < 0$ holds if and only if $2\sigma_{21} \leq \sigma + \sigma_{11} + \sigma_{22}$.

References

Cherry, J. L. (2003). Selection in a subdivided population with dominance or local frequency dependence. *Genetics* **163**, 1511–1518.

Cherry, J. L. and Wakeley, J. (2003). A diffusion approximation for selection and drift in a subdivided population. *Genetics* **163**, 421–428.

Coop, G. and Griffiths, R. C. (2004). Ancestral inference on gene trees under selection. *Theoretical Population Biology* **66**, 219–232.

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. and Tavaré, S. (1995). Coalescents and genealogical structure under neutrality. *Annual Review of Genetics* **29**, 401–421.

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.

Evans, S. N., Shvets, Y. and Slatkin, M. (2006). Non-equilibrium theory of the allele frequency spectrum. *Theoretical Population Biology* , To appear-Doi:10.1016/j.tpb.2006.06.005.

Fearnhead, P. (2001). Perfect simulation from population genetic models with selection. *Theoretical Population Biology* **59**, 263–279.

Fearnhead, P. (2006). The stationary distribution of allele frequencies when selection acts at unlinked loci. *Theoretical Population Biology* , To AppearDoi:10.1016/j.tpb.2006.02.001.

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.

Harding, R. M., Fullerton, S. M., Griffiths, R. C., Bond, J., Cox, M. J., Schneider, J. A., Moulin, D. S. and Clegg, J. B. (1997). Archaic African and Asian lineages in the genetic ancestry of modern humans. *American Journal of Human Genetics* **60**, 772–789.

Hudson, R. R. (1990). Gene genealogies and the coalescent process. In: *Oxford Surveys in Evolutionary Biology* (eds. D. Futuyma and J. Antonovics), volume 7, Oxford University Press, New York, 1–44.

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338.

Hughes, A. L., Packer, B., Welch, R., Bergen, A. W., Chanock, S. J. and Yeager, M. (2005). Effects of natural selection on interpopulation divergence at polymorphic sites in human protein-coding loci. *Genetics* **170**, 1181–1187.

Joyce, P. and Genz, A. (2006). Efficient simulation methods for a class of nonneutral population genetics models. *To appear in Theoretical Population Biology* .

Kendall, W. S. (2005). Notes on perfect simulation. In: *Markov Chain Monte Carlo: Innovations and Applications* (eds. W. S. Kendall, F. Liang and J. Wang), volume 7 of *Lecture Note Series, Instiute for Mathematical Science, National University of Singapore*, 93–146.

Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and their Applications* **13**, 235–248.

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

Marth, G. T., E Czabarka, J. M. and Sherry, S. T. (2004). The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* **166**, 351–372.

Nei, M. (1987). *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Neuhauser, C. and Krone, S. M. (1997). The genealogy of samples in models with selection. *Genetics* **145**, 519–534.

Nicholson, G., Smith, A. V., Jónsson, F., Gústafsson, O., Stefánsson, K. and Donnelly, P. (2002). Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal of the Royal Statistical Society Series B* **64**, 695–715.

Nordborg, M. (2001). Coalescent theory. In: *Handbook of Statistical Genetics*, John Wiley and Sons, England, 179–212.

Nordborg, M. and Innan, H. (2003). The genealogy of sequences containing multiple sites subject to strong selection in a subdivided population. *Genetics* **163**, 1201–1213.

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

Propp, J. G. and Wilson, D. B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms* **9**, 223–252.

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

Schaffner, S. F., Foo, C., Gabriel, S., Reich, D., Daly, M. J. and Altshuler, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. *Genome Research* **15**, 1576–1583.

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

Spencer, G. C. A. and Coop, G. (2004). SelSim: a program to simulate population genetic data with natural selection and recombination. *Bioinformatics* **20**, 3673–3675.

Stephens, M. and Donnelly, P. (2003). A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *American Journal of Human Genetics* **73**, 1162–1169.

Wakeley, J., Nielsen, R., Liu-Cordero, S. N. and Ardlie, K. (2001). The discovery of Single-Nucleotide Polymorphisms – and inferences about human demographic history. *American Journal of Human Genetics* **69**, 1332–1347.

Wall, J. D., Andolfatto, P. and Przeworski, M. (2002). Testing models of selection and demography in Drosophila simulans. *Genetics* **162**, 203–216.