Aust. N. Z. J. Stat. **45**(4), 2003, 395–430

THE INAUGURAL EDITOR'S INVITED PAPER FOR 2002

ANCESTRAL INFERENCE IN POPULATION GENETICS MODELS WITH SELECTION
(with Discussion)

Matthew Stephens[1] and Peter Donnelly[2]*

University of Washington and University of Oxford

Summary

A new algorithm is presented for exact simulation from the conditional distribution of the genealogical history of a sample, given the composition of the sample, for population genetics models with general diploid selection. The method applies to the usual diffusion approximation of evolution at a single locus, in a randomly mating population of constant size, for mutation models in which the distribution of the type of a mutant does not depend on the type of the progenitor allele; this includes any model with only two alleles. The new method is applied to ancestral inference for the two-allele case, both with genic selection and heterozygote advantage and disadvantage, where one of the alleles is assumed to have resulted from a unique mutation event. The paper describes how the method could be used for inference when data are also available at neutral markers linked to the locus under selection. It also informally describes and constructs the non-neutral Fleming–Viot measure-valued diffusion.

Key words: ancestral selection graph; coalescent; Fleming–Viot diffusion; importance sampling; intra-allelic genealogy; perfect simulation.

1. Introduction

Following the human and other genome projects, and recent advances in experimental technology, there are now large and growing amounts of data that document genetic variation, at the DNA sequence level, within natural populations. Such data are shaped by the interaction of the genetic forces — mutation, recombination and natural selection — and by the demographic history of the population. In most contexts the natural models for the data are stochastic. Randomness enters in several ways: chance plays a role in mutation, selection and recombination, and in the lottery of reproduction, through which certain segments of DNA happen to have more copies than others in succeeding generations.

The patterns in molecular population genetics data thus carry information about the history of the population, and about the underlying genetic mechanisms. But the structure of the

Received April 2002; revised May 2002; accepted July 2002.

* Author to whom correspondence should be addressed.

[1] Dept of Statistics, Box 354322, University of Washington, Seattle, WA 98195-4322, USA.

[2] Dept of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG, UK.
e-mail: donnelly@stats.ox.ac.uk

Acknowledgments. The authors thank P. Fearnhead, R.C. Griffiths, P. Slade and C. Wiuf for helpful discussions, and an associate editor for comments on the submitted version. The first author was supported by a grant from the University of Oxford and by a Wellcome Trust Research Fellowship (reference 057416); the second author received support from EPSRC Grant number GR/M14197 and BBSRC grant number 43/MMI09788.

data is complicated and its analysis is far from straightforward. Data from a single region (or locus) in the genome carry limited information, so there is a premium on methods that utilize all the information in the data. This situation has led to considerable recent interest in the development of full-likelihood-based methods for inference from population genetics data — a challenging statistical problem. Recent progress has relied on the application of modern computationally-intensive approaches, but the field is still in its early stages, with the size of available datasets currently threatening to outstrip the development of faster and more efficient methods, for more realistic evolutionary models.

Throughout this paper we focus on settings in which genetic data are available from a sample of chromosomes from the population at a single, relatively small, region of the genome. There are two ways of thinking about the inference problem in this setting, and they give some sense of the challenges involved. The first is in terms of inference for stochastic processes and the second is in terms of a missing data formulation.

Consider the first approach. There are natural stochastic models for population genetics data, which are now relatively well understood. A starting point is to use a discrete-time Markov chain to capture the way in which genetic types in the population evolve from one generation to the next. These discrete-time processes are typically intractable, but their behaviour is simpler in the limit as the population size N increases, with time measured in units of N generations. Formally, weak convergence results establish convergence to a particular limiting process, which is naturally formulated as a measure-valued diffusion. For a genetics model, the state of the population at a single time-point can be represented as a probability distribution — in most cases, this can simply be thought of as a list of the genetic types present, with their respective frequencies. The discrete, pre-limiting, processes can thus be thought of as taking values in the set of probability measures on the collection E of genetic types at the locus of interest. The limiting process, which is called the Fleming–Viot (measure-valued) diffusion, also takes values in this set of probability measures. Its sample paths move smoothly (in an appropriate metric) through this state space, hence the 'diffusion' aspect. The Fleming–Viot process can be thought of as describing the evolution of a hypothetically infinite population. In practice the convergence results justify approximating the discrete models, with large population size, by the Fleming–Viot process. (As it happens, the approximation is typically good even for moderate population sizes, say in the hundreds.) If data were available for the whole population, one way of thinking about the inference problem, then, would be as inference for a particular measure-valued diffusion, based on an observation of its value at a single time-point. In practice, we have at best partial information on the value of the process at the observation point, through data from a sample from the population.

The second way of thinking about the inference problem is in terms of a missing data formulation. At the locus in question, sampled chromosomes are related because they share an evolutionary history. For example, for neutral models without recombination, one way to simulate such samples is first to simulate the underlying genealogical tree that relates the sampled chromosomes, and then, conditional on the tree, to superimpose the effects of mutation along the branches of the tree. (This is an early example of the process that has since come to be called 'perfect simulation' — exact simulation from a complicated stationary distribution with finite computational effort.) The (random) tree relating sampled sequences is called the 'coalescent'. It has a particularly simple structure in the same limit as the one leading to the Fleming–Viot process. Analogous approaches are possible in the presence of recombination and natural selection, but the genealogical object is more complicated than a tree. If the underlying genealogy were known, inference would be straightforward. In practice it is not known, but can be thought of as missing data. The actual likelihood thus involves an average over the unobserved genealogy. While missing data problems have attracted much recent attention in statistics, those arising in population genetics are particularly challenging, effectively because tree-space, the home of the missing data, is enormous.

In this paper we focus on inference in the context of genetics models that incorporate natural selection, and in particular on 'ancestral inference'; that is, inferring the past history of the types (alleles) at the locus of interest. In particular, for a restricted class of mutation models (so-called 'parent-independent' mutation), we develop a new method for generating independent replicates from the conditional distribution of the genealogical history of a sample, given the sample composition, under general diploid selection. Our approach exploits recent advances in modelling (the ancestral selection graph of Neuhauser & Krone (1997)) and in inference in molecular population genetics (Stephens & Donnelly, 2000).

Ancestral inference for models with selection has attracted considerable recent attention. For example Wiuf (2001a) develops an approximation to the conditional genealogy of an allele in a two-allele model with selection, given the current frequency of the allele in the population. The approximation is simple and works well provided the allele is rare, in the sense that its population frequency is less than 10%. Wiuf (2001a) also reviews and contrasts earlier approaches to the problem. Griffiths (2003) (see also references therein) gives a theoretical treatment for the diffusion limit of various aspects of the genealogy of a mutant allele, under selection, including, for example, expressions for its mean age under various selective schemes. Slatkin (2001) introduces a two-stage process for importance sampling to approximate the conditional distribution of the genealogy of an allele in a two-allele model, given either its sample or its population frequency. First, importance sampling is used to simulate from the past allele frequency trajectory, and then, conditional on this, samples of the relevant conditional genealogy are obtained. Wiuf, Griffiths and Slatkin are concerned with the genealogy of only one of the alleles present. In contrast, we describe methods for inferring the genealogy of the entire sample, which in general contains several alleles. (Samples consisting of just one allele are a special case of our analysis.)

Our work here is closest to that of Slade (2000a), who developed, in the context of a two-allele model, a computational method for approximating properties of the conditional distribution of the genealogy, given a sample. This method is based on extending the method of Griffiths & Tavaré (1995) to models with selection and, like the Griffiths–Tavaré method, Slade's approach can be viewed as importance sampling — conditional genealogies are simulated from one distribution (the 'proposal distribution'), and they are given weights depending on how likely they are under another. (See Felsenstein *et al.* (1999) and Stephens & Donnelly (2000) for further discussion on the link between the Griffiths–Tavaré method and importance sampling.) Efficiency, and hence practicality, of importance sampling approaches depends critically on choice of proposal distribution. Here we describe how to obtain independent samples from the optimal proposal distribution: that is, the conditional distribution of the genealogy, given a sample. In addition to improving efficiency, we hope that this framework also makes the approach more transparent.

All the studies mentioned immediately above focus on models with only two alleles. Our method applies whenever the distribution of a mutant type does not depend on the type of the progenitor allele — so-called parent-independent mutation — regardless of the number of alleles. Any two-allele model can be rewritten as an equivalent parent-independent mutation model, and so our approach applies to all two-allele models. Although approximate, the approaches of Wiuf (2001a) (and several of the papers referenced therein) and Slatkin (2001) handle changes in population size.

Aside from inherent interest, an understanding of the structure of genealogical trees at loci under selection is important for several reasons. The shape of the tree affects the patterns of molecular variation around selected sites, and selection of a particular type may or may not have a major effect on this local variation. Therefore, the extent to which we can detect natural selection, and the most powerful methods for doing so, depend on the way in which selection changes the shape and depth of the tree. Many of the current methods for uncovering the genetic basis of common complex diseases in humans aim to exploit linkage disequilibrium (LD) — the non-independent association of alleles at loci close together on the same chromosome. Patterns of LD depend crucially on the shape of the genealogical trees at the loci involved, so that there is considerable interest in understanding how these would be affected by selection.

In the next section we describe the stochastic models that underlie inference in this context, and give the properties we need to use here. We also aim to provide an accessible introduction to the Fleming–Viot process with selection, and in particular to a recent discrete construction of it. Although our new method for ancestral inference is best understood in the context of this construction of the Fleming–Viot process, the method can be applied without any knowledge of the Fleming–Viot diffusion, and readers concerned primarily with applications could thus skip much of Section 2. Section 3 develops our method for ancestral inference, which we then illustrate by considering properties (including allele age and depth of the subtree) of the genealogy of an allele in a two-allele model under either genic selection or heterozygote advantage (or disadvantage). The final section discusses inference (either for genetic parameters or for ancestral inference) when the data document molecular genetic variation, both at the locus under selection and at linked neutral markers.

2. Stochastic models

2.1. Discrete models

We now describe the models to which our analysis applies. Consider a single locus in a randomly-mating population of constant size N diploid individuals (i.e. $2N$ chromosomes). For definiteness we describe one of the most commonly used models, the so-called Wright–Fisher model with selection, but our results apply to any model that converges (after suitable rescaling) to the Fleming–Viot process with selection. In particular this also includes the Moran model with selection. We proceed informally here. For a more detailed treatment, see for example Donnelly & Kurtz (1999).

Denote by E the set of genetic types in the population. (We need to assume some topological conditions on E, for example that it is a complete separable metric space, but these are satisfied for all genetics applications of interest.) Individuals in the population carry two chromosomes at the locus of interest, and we write (A_i, A_j) for the pair of types (genotype) of an individual (the genotype is an unordered pair of types) and $w_N(A_i, A_j)$ $(= w_N(A_j, A_i))$ for the fitness of an individual of this genotype.

The Wright–Fisher model is for a population that evolves in non-overlapping generations. The random process by which the $2N$ types for the next generation are chosen from the current generation is as follows. First, $2N$ pairs of types are chosen, independently and with replacement, from the current generation; the probability that a particular pair (A_i, A_j) is chosen is proportional to its fitness, $w_N(A_i, A_j)$. Next, one of the types within each pair is chosen uniformly, independently across pairs. With probability $1-u$, a copy of that type is included in the next generation. Otherwise, with probability u, a mutation occurs, with the mutant type chosen randomly from a prespecified distribution ν on E. (We allow the possibility that the type chosen from ν may be the same as the current type, in which case the effect of the mutation is that no actual change in type occurs.) This selection and possible mutation within each pair is independent across pairs.

It should be clear that the population evolves in a Markov way, with transition probabilities depending on the current collection of types, and on the distribution ν of mutant types. As just described, natural selection acts by making some genotypes more likely than others to have offspring. This is known as fecundity selection. There are other models with so-called viability selection, in which some genotypes tend to survive longer than others. They behave identically in the limit we consider.

Note the implicit assumption that the (random) type of a mutant does not depend on the type of the progenitor allele. This is often called parent-independent mutation. Although the treatment below is more general than that of previous authors, this assumption about mutation does restrict applicability. We describe below several settings to which the analysis applies.

Example 1. Two-allele models. It is a straightforward exercise to show that if the type-space E consists of only two alleles, then the model can always be rewritten in an equivalent form in such a way that the type of a mutant does not depend on the type of its parent. As a consequence, our analysis applies to any two-allele model. As noted above, most previous authors have considered only that case.

Example 2. K-allele models. More generally, suppose there are K possible alleles, $E = \{A_1, A_2, \dots, A_K\}$, with the probability that a mutant allele is of type A_i being ν_i, $i = 1, 2, \dots, K$. A special case of this model was used by Hartl & Sawyer (1991) to estimate the amount of selection needed to explain amino acid replacement polymorphism.

Example 3. Infinite alleles models. It is sometimes convenient to model the type space as consisting of a hypothetically infinite number of types, with mutations always resulting in types that have never previously been observed in the population. One way of achieving this is to take $E = [0, 1]$, with ν the uniform distribution on the unit interval.

We consider the usual (so-called diffusion) limit of the Wright–Fisher process, in which the population size is large, and the forces of mutation, selection and so-called genetic drift (the randomness inherent in the demography) are comparable in size, and all of order N^{-1}. As is usual, we scale the mutation and selection parameters, and write $\theta = 4Nu$ and $\sigma(A_i, A_j) = 4N(w_N(A_i, A_j) - 1)$. Formally, with this scaling, and time measured in units of $2N$ generations, the Wright–Fisher model converges, as $N \to \infty$, with θ and $\sigma(\cdot,\cdot)$ fixed, to a process called the Fleming–Viot measure-valued diffusion.

It turns out that in the limit we consider, changing all fitnesses by adding a constant (i.e. using $\sigma(x, y) + a$, for any $a \neq 0$, instead of $\sigma(x, y)$) does not change the process in any interesting respect (e.g. Donnelly & Kurtz, 1999). Throughout, we restrict attention to the case of genetic interest, in which the fitness function σ is bounded. By the preceding comment we can also assume, without loss of generality, that it is non-negative. Thus:

$$0 \le \sigma(x, y) \le \sigma_{\max} \quad \text{for all } x, y \in E\,.$$

Other than this, and the assumption that the function σ is symmetric, there are no assumptions on (scaled) fitnesses, though it is usually convenient to set $\min_{x,y} \sigma(x, y) = 0$. Some special cases have attracted particular attention.

Example 4. Genic selection. The case in which (scaled) fitness is additive: $\sigma(x, y) = \gamma(x) + \gamma(y)$ for some function γ on E, is called genic selection. In this case (in the limit we consider), the model is equivalent to one in which individuals are haploid; that is they carry only a single chromosome each, with fitness depending on the type of that single chromosome.

Example 5. Selective classes. It is sometimes natural, for example in the context of infinite alleles mutation models, to break the collection of types E into two classes, with fitness depending only on the classes of the two alleles in a genotype, rather than the alleles themselves. For example, if E is partitioned into a neutral class $\mathcal{N}$ and a deleterious class $\mathcal{D}$, respectively, we can have $\sigma(x, y)$ taking the values $\sigma_{11} \geq \sigma_{12} \geq \sigma_{22}$ when x and y are both in $\mathcal{N}$, one in each class, or both in $\mathcal{D}$, respectively.

Example 6. Heterozygote advantage. For some genetic systems, fitness depends only on whether the pair of alleles within an individual are the same or different, being higher for the latter than the former configuration. In our setting, this would correspond to $\sigma(x, y)$ taking the values σ or 0, according to whether $x \neq y$ or $x = y$ respectively.

2.2. The Fleming–Viot process with selection

We next give a discrete representation of the limiting Fleming–Viot (FV) process. This material is not essential for the remainder of the paper (although we hope it is helpful). We give first the construction for fecundity selection, and then the similar construction for viability selection. The easiest way to think about the Fleming–Viot process is as the description of the evolution of the genetic types present in a hypothetically infinite population. As noted above, the collection of types present at a single time is naturally encoded as a probability measure on E. As time passes, and the population evolves, the probability measure describing the genetic composition also changes. The state space for the FV process is thus the collection of probability measures on E.

Unfortunately, the intuitive simplicity of the discrete, prelimiting processes, such as the Wright–Fisher or Moran models, does not carry over easily in the limit as the population size increases. Even questions about the existence or uniqueness of the limiting measure-valued processes are non-trivial, and for example require the modern, martingale machinery of stochastic processes. In recompense it should be added that the richness of the structure of measure-valued diffusions, and their connections with particular problems in analysis, has made them central objects of study in probability theory for about the last 20 years. See for example Etheridge (2000) and references therein.

Nonetheless, it turns out that there is a discrete representation of the FV (and in fact of a more general class of) measure-valued diffusion, in a sense, as the collection of types in a hypothetically infinite population. These discrete representations can make it easier to think about the processes, and (although this is not our principal concern here) to establish many of their properties. To see that it is not straightforward to specify the dynamics of a suitable infinite population, observe (for example) that it is impossible to choose a pair uniformly at random from an infinite collection, as required in the discrete Wright–Fisher or Moran models in constructing the next generation.

We now describe this construction. (Actually, we skim over one tricky bit and refer the reader to Donnelly & Kurtz (1999) for full details.) We give the dynamics of an evolving collection of E-valued particles, with each one of which is associated a level, indexed from $\{1, 2, \ldots\}$. Write $X_i(t)$ for the type of the particle on level i at time t. The dynamics of the discrete (or particle) construction are defined in terms of a collection of Poisson processes, which index the times of possible changes in the types of the particles at each level. This collection of Poisson processes is as follows: for each pair of levels $i < j$, L_{ji} is a Poisson process of unit rate; for each level, $i = 1, 2, \ldots,$ M_i is a Poisson process of rate $\frac{1}{2}\theta$, and S_i is a Poisson process of rate $\frac{1}{2}\sigma_{\max}$. All these Poisson processes are mutually independent.

The initial distribution of an FV process is a probability distribution on the state space for the process, which is the set of probability measures on E. To initialize the discrete construction, first choose a measure μ according to the initial distribution for the FV process, and then, conditional on μ, choose the values for $X_1(0), X_2(0), \ldots,$ independent and identically distributed (iid) with distribution μ.

From these initial values, the process evolves as follows.

Step 1. If there is an event at time s in the process L_{ji}, then at this time the type of the particle on level j is set to be equal to the type of the particle on level i: $X_j(s) = X_i(s-)$. Events of this type are called 'look downs' — think of the type of the particle on level i being inserted on level j. Also at this time, set $X_{k+1}(s) = X_k(s-)$, $k = j, j+1, \ldots$ (i.e. when the type of particle i is inserted on level j, all other particles move up one level, to make room).

Step 2. If there is an event at time s in the 'mutation' process, M_i, on level i, the type of the particle on level i is replaced by a type randomly chosen according to the distribution ν of mutant types on E.

Step 3. If there is an event at time s in the 'selective' Poisson process S_i on level i, an ordered pair of types Z_1, Z_2 is chosen independently from the population (this is the tricky bit, although we say a little more below). Conditional on the pair chosen, with probability $\sigma(Z_1, Z_2)/\sigma_{\max}$ the type on level i is set to Z_1: $X_i(s) = Z_1$; otherwise no change is made on level i at time s.

There is asymmetry in the look-down mechanism. The type on the lowest level is never changed through a look down, while on level k, look downs are happening at rate $k - 1$, so for large k this type of discontinuity is frequent. From the way in which the discrete construction is initialized, it is immediate that the collection of types $X_1(0), X_2(0), \ldots$ at time zero is exchangeable. (This is close to deFinetti's theorem — the most general exchangeable sequence has a representation in terms of first choosing a random probability distribution, and then, conditional on that choice, sampling an iid sequence from the chosen distribution.) Although not obvious because of the asymmetry in the look-down mechanism, it turns out to be true that the collection of types $X_1(t), X_2(t), \ldots$ is also exchangeable. This observation is the key to the construction. It follows, for example, that for this t there exists (almost surely) a random probability measure $Z(t)$ that is the limit of the empirical measure defined by the types on the first n levels:

$$Z(t) = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \delta_{X_i(t)}, \tag{1}$$

where δ_x is the probability measure that assigns mass 1 to x. The measure $Z(t)$ is just the deFinetti measure associated with the exchangeable collection $X_1(t), X_2(t), \ldots.$ It can be shown that there exists (almost surely) a process $Z(\cdot)$, for which (1) holds for each t.

Think of this situation in the following way. We have defined the (Markovian) dynamics of a discrete collection of particles. By exchangeability, at any time t the types of the countably infinite collection of particles specify a probability measure $Z(t)$ on E. This probability measure, 'carried' by the particles, is the value of the associated FV process. Further, again by exchangeability, for any n, the types of the first n particles $X_1(t), X_2(t), \dots, X_n(t)$ have the distribution of a sample of size n from the FV population. (This observation is crucial, because data come invariably in the form of such a sample.)

In the neutral case the FV process can be constructed in the same way, except that there is no need for the 'selection' Poisson processes S_i, $i = 1, 2, \dots$. In this case the dynamics of the first n particles are self-contained — since particles only ever look down, the collection $(X_1(\cdot), X_2(\cdot), \dots, X_n(\cdot))$ is Markov. This helpful property vanishes under selection.

We have specified the dynamics for the particular mutation models of interest in this paper. For more generality, the dynamics can be extended simply by allowing the distribution ν of a mutant type in step 2 above to depend on the current type $X_i(s)$ on level i. This level of generality is adequate for all forms of mutation in genetics. At a more abstract level, we could specify a Markov process on E (Brownian motion is one example to have attracted attention), and change step 2 above by stipulating that between discontinuities induced either by look downs or selection events, the type on each level evolves independently according to the mutation process.

How do we choose the two types Z_1 and Z_2 'randomly from the population' at the event times of the selection Poisson processes? One answer is that once we have established exchangeability, and hence the existence of $Z(s)$, choosing randomly is equivalent to choosing an independent sample of size two from $Z(s)$, which is easy. This answer is somewhat unsatisfactory in two respects. The first is that it looks a little circular: to construct the process we need to have constructed the process. In fact this difficulty is not insurmountable (Donnelly & Kurtz, 1999 Section 4). A more serious concern is that we would actually like to go backwards in time and trace the ancestry of the types on each level. If we accomplish step 3 above simply by sampling from $Z(s)$ it is meaningless to ask about the ancestors, at time $s - \epsilon$, of the two types sampled at time s, and hence it is impossible to trace genealogical history. Donnelly & Kurtz (1999) give a special construction that makes it possible to follow ancestry back beyond these selective events.

The construction of the FV process with viability selection is very similar to that just given. The only change necessary is in step 3 of the dynamics. Writing $\beta(x, y)$ for the relative mortality of a genotype (x, y), (where $\beta(\cdot, \cdot)$ is symmetric and bounded between 0 and $\beta_{\max}$) step 3 is changed to:

3′ At a time s in the 'selective' Poisson process S_i on level i, an ordered pair of types Z_1, Z_2 is chosen independently from the population. Conditional on the pair chosen, with probability $\beta(X(s), Z_2)/\beta_{\max}$ the type on level i is set to Z_1: $X_i(s) = Z_1$; otherwise no change is made on level i at time s.

In this set-up, large values of $\beta(x, y)$ for the genotype (x, y) correspond to it being relatively more likely to die, and hence less fit, in contrast to the case for fecundity selection. Putting $\sigma(x, y) = \beta_{\max} - \beta(x, y)$ shows the equivalence of the two types of selection.

2.3. Genealogy and the ancestral selection graph

Suppose the particle construction has been evolving indefinitely. (Formally this just needs the collections of Poisson processes to be defined on the doubly infinite time interval.)

Focus attention on a particular time t. We wish to ask about the ancestry of a collection of n chromosomes sampled from the population. By the exchangeability noted above, this is equivalent to asking about the ancestral history of the particles on the first n levels in the particle construction.

Consider first the neutral case (put $\sigma_{\max} = 0$), and for the moment ignore the mutation processes on each level. Fix a level k and go backwards in time tracing the ancestry of the particle on level k at time t. Write s for the time of the first event prior to t in the collection of processes L_{ji}, $i < j \leq k$. If this is an event in the process L_{ki} for some $i < k$ then the ancestry of the particle on level k jumps down to level i at time s. Further back into the past (i.e. before time s) the ancestry of the particle on level k at t and that on level i at t is the same. The two ancestral lineages or histories are said to have 'coalesced'. If the first event in the collection of processes L_{ji}, $i < j \leq k$ is an event in L_{ji} for some $i < j < k$, then at this time s the ancestry of the particle on level k at t simply moves down to level $k - 1$. (Forward in time the particle on level $k - 1$ jumps up to level k at this time.)

Now simultaneously trace the ancestry of the particles on the first n levels back from t. If we ignore the levels on which the ancestries are located the process is still Markov, with simple dynamics. There are initially n ancestral lineages. After an exponentially distributed period of time with parameter $\binom{n}{2}$ (this is just the minimum of the $\binom{n}{2}$ independent unit exponentials until an event in one of the relevant L_{ji}) a randomly chosen pair of lineages coalesces, and the number of lineages drops to $n - 1$. More generally, for $l = n, n-1, \ldots, 2$, there are l lineages for a period of time that is exponentially distributed with parameter $\binom{l}{2}$, after which a randomly chosen pair of lineages coalesces and the number of lineages drops to $l - 1$. The choices of lineages to coalesce are independent of each other and of all the waiting times between coalescences. These waiting times are also independent.

The resulting random object is called a coalescent, or to make precise the dependence on the sample size n, it is sometimes called an n-coalescent. The n-coalescent can be thought of as a random binary tree in which the branches have lengths. It starts with n branches and, writing T_l for the period of time for which there are l branches, the T_l, $l = n, n-1, \ldots, 2$, are independent exponentially distributed random variables with parameters $\binom{l}{2}$ respectively. Each time the number of branches decreases, a randomly chosen pair of branches coalesces, with successive choices independent of each other and of the times T_l, $l = n, n-1, \ldots, 2$.

Tracing genealogical history in the presence of selection is rather more complicated. The way forward, due to Krone & Neuhauser (1997), is to follow a larger than necessary collection of lineages, called the ancestral selection graph.

To see the idea, consider what happens forward in time at an event in the selection Poisson process S_k on level k at time s. At this time, two additional chromosomes are sampled from the population. Denote the type of the first chromosome sampled by Z_1 and the type of the second one sampled by Z_2. What happens after time s depends on the fitness of the genotype (Z_1, Z_2). Either the type already on level k continues across the potential selection event, with probability $1 - \sigma(Z_1, Z_2)/\sigma_{\max}$, or the chromosome on level k dies and is replaced by an offspring of the first of the chromosomes sampled, with probability $\sigma(Z_1, Z_2)/\sigma_{\max}$ (and hence is of type Z_1). To know what actually happens thus requires knowledge of the types of the three chromosomes involved (and in general additional randomization) — the one on level k before time s, and the two additional chromosomes sampled from the population.

Motivated by this, the ancestral selection graph (ASG) trifurcates each time a lineage in it crosses an event in one of the selection Poisson processes. Think of one of the branches as representing the lineage on the relevant level immediately before the selection event, and the other two as representing the lineages of the two additional chromosomes sampled at the selection event. (In the particle construction in Donnelly & Kurtz (1999), the two additional lineages in some sense 'jump up to infinity' at the selection event and then instantaneously come back down to finite levels. This is the difficult bit in the construction, but it is not relevant here. Just as the important structure of the n-coalescent does not require knowledge of which levels the lineages are on, so here the actual levels occupied by lineages in the ASG are immaterial.) In addition to the trifurcations induced in the ASG at selection events, pairs of lineages in the ASG coalesce exactly when there is a look-down event involving the pair of levels occupied by the lineages. Thus, exactly as for the coalescent, each pair of lineages in the ASG independently coalesces at rate 1. Equivalently, when there are l lineages, the number of lineages decreases by one, through the coalescence of a randomly chosen pair of lineages, at rate $\binom{l}{2}$.

In summary, the ASG for a sample of size n is a graph that starts with n edges. When there are l edges, the number of edges in graph increases by exactly two at rate $\frac{1}{2}l\sigma_{\max}$, and decreases by one at rate $\binom{l}{2}$. All choices of pairs of edges to coalesce are independent, and independent of the times between events, and the process is Markov.

An important difference between the coalescent and the ASG is that in the coalescent the number of edges is strictly decreasing, and is eventually absorbed at 1 (after a time with mean $2(1-1/n)$ for the n-coalescent), while in the ASG the number of edges can increase as well as decrease. There are times during which the ASG happens to have only one branch, and the chromosomes at these times are called ultimate ancestors of the sample. It follows, from the fact that the rate at which the number of edges decreases is quadratic whilst the rate of increase is linear, that whatever the value of $\sigma_{\max}$, and the initial sample size n, the ASG is certain to reach a state in which it has only one branch, in finite time. Such times are regeneration events for the ASG. Before having a single branch, the number of branches undergoes an excursion through values larger than 1, before again returning to 1 in finite time.

We see below that the actual genealogy of the sample is contained in the ASG, and the most recent common ancestor (MRCA) of the sample occurs no further back in the past than the first time at which the ASG has a single branch. But the difficulty induced by selection is that to know which subset of the ASG is the actual genealogy requires knowledge of the types of the chromosomes on lineages in the ASG. With this knowledge we can resolve the selection events and decide which of two possible chromosomes before the trifurcation is actually ancestral to the chromosome after the trifurcation. It can be convenient to label the three branches that result from a trifurcation in the ASG: label one branch 'C' (for continuing), one '1' and one '2'. The branch labelled C represents the particle on the relevant level immediately before the selection event. The other two branches represent respectively the first or second additional chromosomes sampled at that time.

In the special case of genic selection (example 4 in Section 2.1) there is an equivalent and simpler version of both the FV process construction and the ASG. Adopt the notation there for the fitness function $\sigma(\cdot,\cdot)$, and write $\gamma_{\max}$ for the maximum of the genic fitnesses: $0 \le \gamma(x) \le \gamma_{\max}$ for all $x \in E$. Then it turns out that the relevant FV process can be constructed by running independent, rate $\frac{1}{2}\gamma_{\max}$, selection Poisson processes S_i on each level and replacing step 3 of the construction by

3$''$ If an event occurs at time s in the 'selective' Poisson process S_i on level i, a type Z_1 is chosen independently from the population. Conditional on the type chosen, with probability $\gamma(Z_1)/\gamma_{\max}$ the type on level i is set to Z_1: $X_i(s) = Z_1$; otherwise no change is made on level i at time s.

With genic selection the ASG has bifurcations, at rate $\gamma_{\max}$, rather than the trifurcations needed for general diploid selection.

2.4. Stationary distributions

For the class of models we are considering, explicit expressions are available for the stationary distribution of the FV process (and hence in the diffusion limit for the discrete processes) under neutrality in the case of parent-independent mutation. Under selection, with parent-independent mutation, the stationary distribution of the FV process is known up to a normalizing constant. No expressions for stationary distributions are known for more general mutation models, even under neutrality.

We give here the stationary distribution of the FV process for a K-allele model, with parent-independent mutation, as described in example 2 of Section 2.1. It is also known for infinite alleles mutation. See, for example, Donnelly, Nordborg & Joyce (2001), for the details. For a K-allele model, the state space of the FV process is just the set of probability distributions on the finite set $E = \{A_1, A_2, \ldots, A_K\}$. The state space is thus equivalent to the $K - 1$ dimensional unit simplex

$$\Delta_K = \{(x_1, x_2, \ldots, x_K) \colon x_i \geq 0, i = 1, 2, \ldots, K,\ x_1 + x_2 + \cdots + x_K = 1\},$$

in which we can think of the component x_i as describing the proportion of the population of type A_i, $i = 1, 2, \ldots, K$. Recall that the probability that a mutant allele is of type A_i is ν_i, $i = 1, 2, \ldots, K$.

Suppressing the dependence on θ, we write π_σ for the stationary distribution of the FV process under selection, with π_0 denoting its neutral stationary distribution, and write $\pi_\sigma^{(n)}$ ($\pi_0^{(n)}$) for the distribution of an ordered sample of size n at stationarity under selection (neutrality).

Under neutrality, π_0, the stationary distribution of the FV process is just the Dirichlet distribution on Δ_K with parameters $(\theta\nu_1, \theta\nu_2, \ldots, \theta\nu_K)$ (Wright, 1949). The distribution of an ordered sample from a Dirichlet distribution is well known (e.g. Bernardo & Smith, 1994 Appendix A). It follows that under neutrality the probability $\pi_0^{(n)}(y_1, y_2, \ldots, y_n)$ that in a sample of size n the jth chromosome sampled is of type y_j, $j = 1, 2, \ldots, n$, is

$$\pi_0^{(n)}(y_1, y_2, \ldots, y_n) = \frac{(\theta\nu_1)_{(n_1)}(\theta\nu_2)_{(n_2)} \cdots (\theta\nu_K)_{(n_K)}}{\theta_{(n)}}, \tag{2}$$

where n_i denotes the number of y_j taking the value A_i, $i = 1, 2, \ldots, K$, and $z_{(n)}$ denotes the rising factorial $z(z+1) \cdots (z+n-1)$.

To describe the relevant stationary distributions under selection we introduce one further piece of notation. Write $\sigma^*(\mu)$ for the mean fitness of a population whose composition is described by the measure μ on E. Think of $\sigma^*(\mu)$ as the fitness of a randomly chosen genotype (X, Y) where X and Y are independently distributed according to μ. Formally:

$$\sigma^*(\mu) = \int_{E \times E} \sigma(x, y)\, \mu(dx)\, \mu(dy)\,.$$

For K-allele models we can identify each such μ with a point $(x_1, x_2, \dots, x_K) \in \Delta_K$. Then

$$\sigma^*(x_1, x_2, \dots, x_K) = \sum_{i,j} \sigma(A_i, A_j) x_i x_j .$$

The stationary distribution of an FV process with selection is absolutely continuous with respect to the equivalent model under neutrality (Donnelly & Kurtz, 1999). For parent-independent mutation, the form of the Radon–Nikodým derivative is known exactly (e.g. Ethier & Kurtz, 1993). This situation means in practice that for any model with parent-independent mutation, we can relate π_σ, the stationary distribution of the FV population under selection, with π_0, the stationary distribution of the relevant (by which we mean the same mutation mechanism) neutral FV process

$$\pi_\sigma(\mu) = C\pi_0(\mu) \exp\left(\tfrac{1}{2}\sigma^*(\mu)\right), \tag{3}$$

where C is a normalizing constant chosen to ensure that the distribution (3) integrates to 1. Equation (3) applies to any FV process with parent-independent mutation. It is not restricted to K-allele models.

There is an analogue of (3) that relates the distributions of samples of size n under selection with those under neutrality. The general version is given, for example, by Donnelly *et al.* (2001 Equation (4)). We give it here for K-allele models

$$\begin{aligned}\pi_\sigma^{(n)}(y_1, y_2, \dots, y_n) &= C\pi_0^{(n)}(y_1, y_2, \dots, y_n) \frac{\Gamma(n+\theta)}{\Gamma(\theta\nu_1 + n_1)\dots\Gamma(\theta\nu_K + n_K)} \\ &\times \int_{\Delta_K} x_1^{\theta\nu_1+n_1-1} \cdots x_K^{\theta\nu_K+n_K-1} e^{\sigma^*(x_1,\dots,x_K)/2}\, dx_1 \dots dx_{K-1} ,\end{aligned} \tag{4}$$

where $x_K = 1 - (x_1 + \dots + x_{K-1})$, and n_i denotes the number of y_j taking the value A_i, $i = 1, 2, \dots, K$ (Donnelly *et al.*, 2001). The normalizing constant C in (4) is the same as the one defined implicitly in (3); in particular it does not depend on n, a fact that we use later.

The normalizing constant C in (3) and (4) is not known in closed form. Its value depends on the mutation parameters as well as the selection intensities $\sigma(\cdot,\cdot)$. In some special cases there are efficient numerical schemes for approximating it, but in general (Donnelly *et al.*, 2001) this is a challenging problem. In contrast, it is often reasonably straightforward to numerically approximate the integral on the right-hand side of (4). An important feature of the method we describe below is that it is not necessary to calculate the normalizing constant C.

2.5. Simulating samples via genealogy

For the inference questions in the sequel it is important to understand how the coalescent and the ASG can be used to simulate samples from population genetics models.

If the primary goal is simply to simulate samples from the population then it is not actually necessary to use genealogy for models with parent-independent mutation. In the neutral case, samples can be simulated directly from (2), or by first simulating the population according to the relevant Dirichlet distribution and then taking a multinomial sample. Under selection, we can use importance sampling, for example via (4). See Donnelly *et al.* (2001) for details.

We first consider the neutral case. Simulation of a sample of size n via the coalescent involves three steps.

Step 1. Choose a realization of the n-coalescent.

Step 2. Independently of the previous step, choose the type of the MRCA of the sample. In general the type is distributed according to the stationary distribution of the mutation process. For parent-independent mutation, it has the distribution ν.

Step 3. Conditional on the previous two steps, go forward through the genealogical tree specified by the realization of the coalescent, from the MRCA, and superimpose the effects of mutation, independently along distinct branches of the tree.

For neutral models with general mutation mechanisms, coalescent-based simulation is extremely efficient. If it is possible to simulate exactly from the stationary distribution of the mutation process, which is often the case, then it also provides perfect simulation, in the sense that finite simulation effort gives an exact sample from the relevant stationary distribution. Recall that no explicit expressions are available for that distribution.

In the case of parent-independent mutation, the fact that the scheme just described simulates samples at stationarity follows directly from the particle construction of the FV process described in Section 2.2. Recall that if we traced back the ancestry of the particles on each of the first n levels, and then ignored the information about the actual levels occupied by ancestral lineages, the resulting genealogical tree was a version of the n-coalescent. The MRCA of the sample occurs the first time all lineages are traced back to the bottom level. The time at which this occurs is independent of the type on that level at that time. (This follows here because the time in question is a function of the look-down processes L_{ji}, and the type on the bottom level is a function of the process M_1 and random choices from ν, all of which are independent of the look-down processes.) It follows that the type of the MRCA has distribution ν, independent of the realization of the n-coalescent. Finally, again by the independence of the mutation and look-down processes, it follows that as we go forward from the MRCA through the tree, along each branch mutations will occur independently at rate $\frac{1}{2}\theta$, at which time the new type will be a random draw from ν.

Now consider models with selection. The extension of the simulation scheme described in the neutral case is as follows.

1. Choose a realization of the ASG for the sample of size n, back to the first time at which there is a single branch.
2. Independently of this, choose the type of the ultimate ancestor at that time. This has the distribution of a sample of size 1 from the relevant FV process with selection.
3. Conditional on steps 1 and 2, go forward through the ASG from the ultimate ancestor superimposing the effects of mutation independently on each branch. Resolve each trifurcation event independently of everything else as follows:
   (a) write x, z_1 and z_2 for the types on the branches labelled C, 1 and 2 respectively at the trifurcation event;
   (b) with probability $\sigma(z_1, z_2)/\sigma_{\max}$, the type after the trifurcation event (by which we mean closer to the present) is z_1; otherwise the type after the trifurcation event is x.

That this scheme gives a sample with the required distribution for parent-independent mutation follows as in the neutral case from the particle construction of the process given in Section 2.2. The ultimate ancestor is of the same type as a sample of size 1 from the FV process so long as the type and the time at which the ASG first has a single branch are both independent. This follows again from the independence of the various components that make up the construction, and the independent increments property of the relevant Poisson processes. Again, all we are doing here is tracking back through the particle construction. In contrast to the neutral case, the type of the ultimate ancestor is no longer just the stationary distribution of the mutation process (in our case, no longer just ν).

As has been noted elsewhere (Donnelly *et al.*, 2001), under parent-independent mutation there is no point in using the ASG to simulate samples under selection: implementation of step 2 requires simulation of a sample of size 1, for example from (4), involving as much computational effort as simulation of a sample of size n. Where the ASG is useful in this setting is in simulating genealogies under selection. If we implement the three steps of the simulation algorithm just described, then amongst other things the resolution of the trifurcations associated with selective events means that we know which branch at each trifurcation carries an actual ancestor of the sample. Having undertaken the simulation as described, we can specify the subset of the ASG that contains the actual genealogy of the sample. (An alternative approach to simulating unconditional genealogies, which may be more efficient in some settings, is first to simulate a sample, according to (4), and then to use the approach of this paper to simulate the genealogy conditional on the simulated sample.)

In more general mutational settings the distribution of a sample of size 1 from the FV process is not known, and hence step 2 above cannot be implemented. Fearnhead (2001) describes an ingenious way to circumvent this problem, allowing samples and genealogies to be simulated under more general mutation models, and reducing the computational burden even in the parent-independent case. However, the inferential procedure we describe in the next section relies on the availability (up to a normalizing constant) of the stationary probabilities (4), and so cannot currently be applied in these more general settings.

3. Statistical inference

3.1. Background

Suppose we have data in the form of a sample of n chromosomes from a single locus taken at stationarity from a Fleming–Viot model. It is helpful to separate three different sorts of inference question that can be asked.

1. What can be said about the genetic parameters (in our context, mutation rates and mechanisms and selective intensities) governing the dynamics of the process?
2. What can be said about the past history of the genetic types observed in the sample?
3. What can be said about the demographic history of the population from which the data are sampled?

The first type of question, inference about genetic parameters, is in principle straightforward for parent-independent mutation in the neutral case. Equation (2) gives the likelihood for the data. For parent-independent mutation under selection, progress is possible, at computational cost, by exploiting the absolute continuity (3); see Donnelly *et al.* (2001). Likelihood-based inference for general mutation models under neutrality is a challenging problem which has attracted much recent attention, and several computationally-intensive statistical methods have been developed, using either Markov chain Monte Carlo or importance sampling. See Stephens (2001) for a review. We are not aware of published methods for likelihood-based parameter estimation for general mutation models under selection.

In Section 3.2 we develop a new method for the second type of question, namely inference about the past history of sampled types (often called ancestral inference) for parent-independent mutation models under selection. We see in subsequent sections that this kind of ancestral inference is also an important building block for answering more general inference questions in genetics.

Inferring population demographic history from genetic data is an enormous field. It turns out that information from a single locus is limited, and often confounded with genetic effects. (For example, certain sorts of selection produce patterns in the data that are similar to the patterns generated by population expansion.) Since demographic history affects all loci in the genome, inference about demographic history is better undertaken from multi-locus data. For an example of recent applications, see Frisse *et al.* (2001) and references therein, and the reviews by Excoffier (2001) and Rousset (2001).

3.2. Ancestral inference

Write X_n for a sample of size n from the population at stationarity. We adopt the usual diffusion limit of the standard population genetics models, as described in Section 2. In this section we show how to simulate exactly from the conditional distribution of the ASG given the sample configuration, and hence how to simulate exactly from the genealogy of the sample given its configuration. In Section 3.3 we show how this technique can be applied to several other inference questions in genetics.

For ease of reference, we give here an explicit algorithm for simulation from the ASG unconditional on the observed sample.

Algorithm 3.1. To simulate from the ASG for a sample of size n taken at time $t = 0$, back to a fixed time $t = t_0$:

Step 1. Start with n lineages at $t = 0$.

Step 2. Going backwards in time, when there are k lineages, simulate the time s to the next event from an exponential distribution with rate parameter $\frac{1}{2}k(k - 1 + \theta + \sigma_{\max})$. If $t + s > t_0$, go to step 5.

Step 3. Generate the next event backwards in time (a coalescence, a mutation, or a trifurcation event) according to the following rules:

- with probability $(k - 1)/(k - 1 + \theta + \sigma_{\max})$, coalesce (merge) a pair of lineages chosen uniformly at random;
- with probability $\theta/(k - 1 + \theta + \sigma_{\max})$, place a mutation on a lineage chosen uniformly at random;
- with probability $\sigma_{\max}/(k - 1 + \theta + \sigma_{\max})$, trifurcate (split) a lineage chosen uniformly at random into three lineages labelled C, 1 and 2.

Step 4. Return to step 2.

Step 5. Simulate the genetic types on the ASG at time t_0 from $\pi_S(\cdot)$.

Step 6. Follow types forward through the ASG from $t = t_0$ to $t = 0$, resolving the mutation and selection events independently as follows:

- when a mutation occurs on a lineage the type on that lineage changes to type A_i with probability ν_i $(i = 1, 2, \dots, K)$;
- at a node corresponding to a trifurcation event, write x, z_1 and z_2 respectively for the types on the branches labelled C, 1 and 2. With probability $\sigma(z_1, z_2)/\sigma_{\max}$ the type after the trifurcation (i.e. closer to time $t = 0$) is z_1; otherwise it is x.

As written, Algorithm 3.1 specifies how to simulate the ASG over a fixed time period. Alternatively, the ASG can be simulated until it has a particular form, for example the first time it reaches an ultimate ancestor. This variation requires a change to the stopping rule. It follows from Donnelly & Kurtz (1999 Lemma 8.1) that provided the time at which the ASG is stopped is a stopping rule for the ASG (as defined above Lemma 8.1 in Donnelly & Kurtz, 1999), then the types on the branches of the ASG at the time of stopping are still given by a random sample from the appropriate FV process, so the analogue of step 5 is still valid.

Now consider the ASG conditional on the types in a sample. We show that backwards in time the ASG with the types on the lineages is still a Markov process, and give its transition rates. Suppose that at time t the types in the ASG are given by $X_k(t) = (x_1, \dots, x_k)$. The possible events that could occur in the ASG between times t and $t + \delta$, going backwards in time are as follows:

(a) events of type $\mathcal{C}_i$ $(i = 1, 2, \dots, K)$, in which two lineages carrying alleles of type A_i coalesce;

(b) events of type $\mathcal{M}_{ij}$ $(i, j = 1, 2, \dots, K)$, in which an allele of type A_i being carried on a particular lineage arose through a mutation occurring on that lineage, where type A_j was carried on that lineage before the mutation;

(c) events of type $\mathcal{B}^1_{ijl}$ $(i, j, l = 1, 2, \dots, K)$, in which an allele of type A_i on a particular lineage arose through a selection event involving types A_i, A_j and A_l on the branches labelled 1, C and 2 respectively, and the type on branch 1 is transmitted;

(d) events of type $\mathcal{B}^C_{jil}$ $(i, j, l = 1, 2, \dots, K)$, in which an allele of type A_i on a particular lineage arose through a selection event involving types A_j, A_i and A_l on the branches labelled 1, C and 2 respectively, and the allele on branch C is transmitted;

(e) nothing happens.

If n_i denotes the number of lineages carrying allele A_i, then there are $\binom{n_i}{2}$ possible events of type $\mathcal{C}_i$, and n_i events of each of the types $\mathcal{M}_{ij}$, $\mathcal{B}^1_{ijl}$ and $\mathcal{B}^C_{jil}$ $(i, j, l = 1, 2, \dots, K)$.

The following proposition (which is an analogue of Theorem 1 in Stephens & Donnelly, 2000) characterizes the process consisting of the ASG with types on lineages, backwards in time. In combination with (4) it allows us to compute the rates at which each event of types (a)–(d) occurs. Denote by $\Gamma(\cdot)$ the process consisting of the ASG with the types on the lineages of the graph. (We think of the edges of the ASG as being labelled. How this is done is not important. In the particle construction these branches carry the labels of the levels on which they find themselves. Alternatively, we could label the tips of the ASG, at the sampling time, from $\{1, 2, \dots, n\}$, and at a coalescence the branch resulting from the join would carry the label which would be the concatenation of the labels of the two branches which coalesce, and at a trifurcation event the three branches would carry the label of the branch which trifurcates, augmented by one of C, 1 and 2.) The process includes information about the locations of mutation and selection events, and that we can (and do) choose to retain in Γ the information about what happens at the randomization associated with each selection event (i.e. whether the type on the continuing branch C is or is not replaced by the type on branch 1) and hence that we know from Γ which branches represent the actual ancestors.

Proposition 3.1. *Viewed backwards in time, the process Γ is Markov. Suppose that in the current state, the types on the n lineages are written in an ordered list as x. Write $x - i$ for any ordering of the same types with one copy of the allele A_i removed, $x + j + l$ for any ordering of the same types with an additional copy of each of A_j and A_l added, and $x - i + j$ for any ordering of the same types with one copy of A_i removed and a copy of A_j added.*

Then the transition rates for the possible events considered backwards in time are as follows:

$$\textit{Each event of type } \mathcal{C}_i \textit{ occurs at rate } \quad \frac{\pi_\sigma^{(n-1)}(x-i)}{\pi_\sigma^{(n)}(x)}\,. \tag{5}$$

$$\textit{Each event of type } \mathcal{M}_{ij} \textit{ occurs at rate } \quad \frac{\theta\nu_i}{2}\,\frac{\pi_\sigma^{(n)}(x-i+j)}{\pi_\sigma^{(n)}(x)}\,. \tag{6}$$

$$\textit{Each event of type } \mathcal{B}^1_{ijl} \textit{ occurs at rate } \quad \frac{\sigma_{\max}}{2}\,\frac{\sigma(A_i,A_l)}{\sigma_{\max}}\,\frac{\pi_\sigma^{(n+2)}(x+j+l)}{\pi_\sigma^{(n)}(x)}\,. \tag{7}$$

$$\textit{Each event of type } \mathcal{B}^C_{jil} \textit{ occurs at rate } \quad \frac{\sigma_{\max}}{2}\Big(1-\frac{\sigma(A_j,A_l)}{\sigma_{\max}}\Big)\,\frac{\pi_\sigma^{(n+2)}(x+j+l)}{\pi_\sigma^{(n)}(x)}\,. \tag{8}$$

Remark 1. All the rates in the process Γ depend on ratios of the sampling distribution $\pi_\sigma^{(n)}$ given at (4). As noted previously, the unknown normalizing constant C does not depend on n, and cancels from these ratios, so that, happily, simulation of Γ does not require knowledge of C.

Proof. The result mimics the standard result about the time reversal of a continuous time Markov chain at stationarity, for a non-time-reversible process. See for example Kelly (1979 Theorem 1.12) for a statement and proof. For a proof of an analogous result in a setting close to the one here, see Stephens & Donnelly (2000 Theorem 1). We proceed somewhat informally here.

Donnelly & Kurtz (1999 Lemma 8.1) establish that at any fixed time t, conditional on the number of lineages, k, in the ASG, the types on these lineages are those of a sample of size k at stationarity from the relevant FV process, i.e. they have distribution $\pi_\sigma^{(k)}$ given by (4). In fact, that lemma shows that this situation remains true for any time t which is a stopping time, in the sense defined by Donnelly & Kurtz (1999) for the ASG.

For definiteness, imagine following the ASG back from time 0, and tracing the types on the lineages. (The distribution of Γ does not depend on the choice of time from which we look backwards.) Condition on $\Gamma(t)=x$, and let n denote the number of lineages in $\Gamma(t)$.

Consider a particular lineage carrying an allele of type A_i in $\Gamma(t)$. Write $\mathcal{B}$ for the event that an event of type $\mathcal{B}^1_{ijl}$ occurred on this lineage over the interval $(t,t+\delta)$ (with time measured into the past from time 0). For $\mathcal{B}$ to occur, leaving the configuration on the edges of the ASG as x, we require:

- that a selection event occurs on the particular lineage of interest;
- that forward in time in the particle construction the types on the n levels in $\Gamma(t+\delta)$ are given by $x-i+j$;
- that the first and second additional types sampled from the population at the event are A_i and A_l respectively; and
- that the type on branch 1 replaces the type (A_j) on the particular lineage.

Conditional on the number of lineages in $\Gamma(t)$ the probability of this collection of events is

$$\tfrac{1}{2}\sigma_{\max}\,\delta\pi_\sigma^{(n+2)}(x+j+l)\,\frac{\sigma(A_i,A_l)}{\sigma_{\max}}\,, \tag{9}$$

since the types on the n levels and the additional two types sampled from the population have the distribution of a sample of size $n+2$ from the population.

Thus, using (9), conditional on n,

$$\Pr(\mathcal{B} \mid \Gamma(t) = x) = \frac{\Pr(\mathcal{B} \text{ and } \Gamma(t) = x)}{\Pr(\Gamma(t) = x)} = \tfrac{1}{2}\sigma_{\max}\, \delta\, \frac{\sigma(A_i, A_l)}{\sigma_{\max}}\, \frac{\pi_\sigma^{(n+2)}(x + j + l)}{\pi_\sigma^{(n)}(x)},$$

as required for (7). By symmetry, this probability is true for all events of this type. Analogous arguments give (5), (6) and (8). That the process Γ is Markov follows from the independent increments properties of the Poisson processes used in the particle construction of the FV process, and Donnelly & Kurtz (1999 Lemma 8.1).

Amongst other things, Proposition 3.1 characterizes the probabilistic structure of the ASG back from a sample whose composition is known. Since the process in the proposition also includes the information about which branches represent the actual ancestors at selection (trifurcation) events, the proposition allows us to simulate from the actual genealogy of a sample whose composition is specified. However, this simulation still requires simulation of the full ASG — some of the lineages in the ASG do not represent actual ancestors of the sample, but we still need to keep track of them because the rates in the conditional process depend on the types on all the lineages, not just those which are ancestral (though see Remark 4 below for some possible shortcuts). Slade (2000a) distinguishes between lineages that are ancestral and those that are not, and calls the former ‘real’ and the latter ‘virtual’.

The following algorithm allows simulation of sample genealogy conditional on the types in the sample. In fact, it provides the genealogy together with the genetic types of all ancestors back to and including the MRCA of the sample.

Algorithm 3.2. To simulate the sample genealogy from the genealogy of a sample whose composition is known:

Step 1. Start with the n observed types, X_n, at $t = 0$. Initially all branches of the ASG carry ancestors of the sample.

Step 2. Calculate the rate of each possible event backwards in time in the ASG, as given in Proposition 3.1 above. Label the possible events $E_1, \dots, E_R$, and let $\lambda_1, \dots, \lambda_R$ denote their respective rates.

Step 3. Simulate the time s to the next event from an exponential distribution with rate parameter $\lambda = \lambda_1 + \cdots + \lambda_R$.

Step 4. Simulate which event occurs, with the probability of E_r occurring being λ_r/λ. At trifurcation events on a lineage carrying an ancestor of the sample, keep track of which branch carries the actual ancestor of the sample.

Step 5. If the sample has more than one actual ancestor, return to step 2; otherwise stop.

Remark 2. The times, sampled in step 3, can actually be sampled after the topology has been sampled, and in general it may be most efficient to sample several sets of these times per genealogy sampled (see Stephens (2000) for further discussion of this in the neutral case).

Remark 3. In cases where only the most recent part of the total genealogy is of interest (e.g. Section 4), the algorithm can be stopped early, before the sample has reached a single actual ancestor.

Remark 4. In fact we can get away with not following some of the branches in the ASG backwards in time. See Fearnhead (2002) for details. Our implementations did not take advantage of that additional simplification.

We now consider some practical issues in implementing Algorithm 3.2. For a K-allele model there are K possible mutation events and $2K^2$ possible selection events which could occur on each level. If the number of current levels in the ASG is k, the number R of possible events in step 2 of the algorithm is $k(K+2K^2)$ plus the number of possible coalescence events.

The number of events going backwards in time in the ASG increases rapidly with $\sigma_{\max}$. The fact that the algorithm stops when the sample has only one real ancestor, and need not continue all the way back to the ultimate ancestor, mitigates this problem somewhat. Nevertheless, for large values of $\sigma_{\max}$, simulation of even one genealogy from the conditional distribution can become computationally impractical. The additional simplification noted in Remark 4 above can reduce these difficulties.

While calculation of the transition rates in Algorithm 3.2 does not require knowledge of the normalizing constant C in (4), it does require evaluation of the integral in (4). Inspection shows these integrals (including the ratio of the Gamma functions, which precedes them) to be expectations of exponential functionals of Dirichlet random variables. In general the integrals can be approximated by simulation, or by numerical integration; the difficulty in approximating closely increases with increasing $\sigma_{\max}$. In many cases it may be more efficient to quickly approximate these integrals rather roughly, foregoing the luxury of simulating from essentially the exact conditional distribution, and to correct for this roughness by using importance sampling, as in Stephens & Donnelly (2000). In the commonly studied case of a two-allele model with genic selection, where without loss of generality we take the scaled fitness of one of the alleles to be 0, the integral in question reduces to the form

$$M(\alpha,\gamma;x)=\frac{\Gamma(\gamma)}{\Gamma(\alpha)\Gamma(\gamma-\alpha)}\int_0^1 f^{\alpha-1}(1-f)^{\gamma-\alpha-1}\exp(fx)\,df=\sum_{j=0}^{\infty}\frac{\alpha_{(j)}}{\gamma_{(j)}}\frac{x^j}{j!}\,. \qquad (10)$$

Here $M(\alpha,\gamma;x)$ is the confluent hypergeometric function. This infinite sum converges sufficiently rapidly to be computed very accurately (for the values of α, γ and x we consider) by summing the first 10 000 terms (Slade, 2000b, makes use of the same method). By tabulating results for suitable values of α, γ and x, fast simulation from essentially the exact conditional distribution can be achieved for moderate values of $\sigma_{\max}$.

3.3. Conditioning on a unique mutation event

In some settings a natural model for the variation present at a locus is that a single mutation event occurred in the ancestral history of the locus since its MRCA, to introduce a second type. In this context the allele that was present before the mutation is called the 'wild type' and the one that resulted from the mutation is called the 'derived type'. This framework is natural for example for most instances of single nucleotide polymorphisms (SNPs) in humans — sites in the human genome which exhibit variation across different chromosomes — typically because each sampled chromosome carries one of two possible nucleotides at the site in question. Sequence data from a closely related species (called an 'outgroup') can determine which of the two nucleotides at an SNP is the wild type, and which is derived.

Adopt a two-allele model and consider conditioning on the event $\mathcal{U}$ that a unique mutation, from the wild type to the derived type, occurred in the ancestry of the sample. Provided there is variation present in the sample configuration X_n, in the limit $\theta \to 0$ all (and in practice, for the value $\theta = 0.0001$ we used, almost all) the sampled ASGs from $\Pr(\Gamma \mid X_n)$ have a single mutation. Some sampled ASGs have a mutation from the wild type to the derived type, while others have a mutation from the derived type to the wild type. To condition on there being a unique mutation from the wild type to the derived type we can discard those simulated ASGs in which the mutation occurs in the opposite direction, or in which there is more than one mutation. In some cases (e.g. where the derived type is much fitter than the wild type, but is at very low frequency in the sample) the proportion of samples discarded approaches 1, and the method becomes impractical. The (long run) proportion of samples discarded is just the conditional probability, given the data and the parameter values, that there was indeed a unique mutation event from the wild to derived type. In settings where the approach becomes impractical we may wish to re-evaluate whether there was indeed a unique mutation of the type hypothesized. (Initially it might seem that the proportion of sampled Γ that must be discarded will depend on the values of ν_1 and ν_2, and that the method could be made more efficient by taking the limit $\nu_1 \to 0$ for example. However, careful analysis reveals that this is not the case. Indeed, computing the rates (5)–(8) in the limit $\theta \to 0$ shows that, in this limit, the conditional distribution $\Pr(\Gamma \mid X_n)$ does not depend on (ν_1, ν_2).)

4. Application: The age of a non-neutral allele

In a celebrated paper, Kimura & Ohta (1973) derived the formula

$$\mathrm{E}(A) = -\frac{2f \log f}{1-f}$$

for the expected value of the age A of a neutral mutation at current frequency f in a finite, constant-size population. More recently several authors, including Griffiths & Tavaré (1998), Wiuf & Donnelly (1999) and Stephens (2000), have used ideas from coalescent theory to obtain results for the full distribution of A, in the neutral case, under a variety of demographic models; the most general formula appears in Griffiths & Tavaré (1998). Griffiths (2003) has obtained some analogous results under selection. Here we consider the distribution of A for a 'non-neutral' mutation, and investigate how it is affected by the intensity and direction of selective pressures. We also consider the height H of the subtree relating chromosomes that share the mutation (Figure 1), and the total length of this tree, which are often of more interest than the age, as they have a more direct effect on the extent of linkage disequilibrium around the mutation, and thus for example on the efficacy of association methods for gene mapping, the interpretation of linked variation, and detection of the selection.

Throughout, we describe the selection model in the most convenient way. For example to contrast genic selection for and against the mutant allele, we have its genic fitness σ being positive or negative, with the wild type having genic fitness 0. In practice we simulate the case $\sigma < 0$ by adding $-\sigma$ to both selection coefficients.

4.1. Genic selection

We follow previous authors (including Wiuf, 2001a) in addressing the question in the context of a two-allele model, with genic selection, and assuming that $\theta \to 0$ and that the allele of interest arose as the result of a unique mutation event in the history of the sample. Then the ancestral tree $\mathcal{T}$ relating the sampled individuals can be represented as in Figure 1. Denote the wild-type allele by 0 and the derived type by 1, and assume that the fitness of an individual carrying k copies of the derived type is $k\sigma$ $(k = 0, 1, 2)$; or, equivalently,

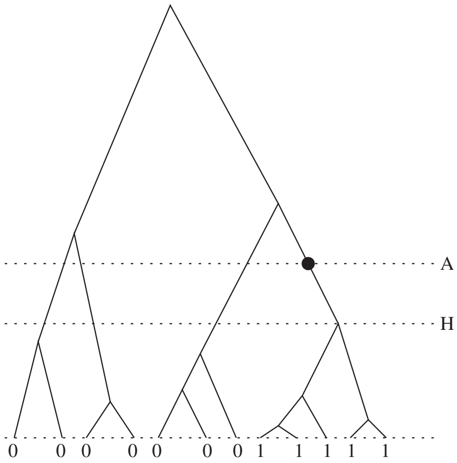

Figure 1. An example of an ancestral tree relating $n = 12$ randomly sampled individuals who are represented by the tips at the bottom of the diagram. The sample shown contains $n_0 = 7$ chromosomes of the wild type, type 0, and $n_1 = 5$ chromosomes of type 1 which share a single mutation (shown as a black circle). Measuring time backwards from the present, the age A of the mutation, and the height H of the subtree relating the individuals carrying the mutation are as indicated.

assume that the fitness of an individual carrying k copies of the wild type is $-k\sigma$. For various values of $|\sigma|$ (up to 500), and various sample configurations, we use Algorithm 3.2 to generate $M = 50\,000$ independent samples from the conditional distribution of the ASG, back to the time of the mutation. This time is typically much shorter than the time to the sample MRCA, making the study of large values of $|\sigma|$ computationally tractable. We use a small value of θ $(= 0.0001)$ to mimic the limit $\theta \to 0$ (see the discussion in Section 3.3). Conditional on the order of events in each sampled ASG, we generate 10 independent sets of times between these events (in a way analogous to that outlined in Stephens, 2000), giving $10M$ (dependent) samples from the conditional distribution of A and H given the sample configuration, and produce histograms of the results. To give some idea of the computational complexity, the simulation of $M = 50\,000$ independent samples, for $n = 1000$ chromosomes of which $n_1 = 10$ are of type 1, and $\sigma = -5$, takes about 40 minutes on a desktop machine with an 800 MHz processor. More challenging configurations (larger n_1 and larger absolute value for σ) can take up to a day.

Maruyama (1974) showed that, somewhat surprisingly, for genic selection in a population of constant size, in the diffusion limit we are considering, the distribution of the age of the mutation, conditional on its population frequency, does not depend on the sign of the selection coefficient σ. (However, the distribution of the age conditional on its frequency in a sample *does* depend on the sign of σ.) Not surprisingly then, for the large sample sizes we are using, our simulation results for $\sigma > 0$ are essentially indistinguishable from the corresponding results for $-\sigma$, although for large values of $|\sigma|$, and low frequencies for the derived type, the simulation takes considerably more computer time for $\sigma > 0$ (for reasons discussed in Section 3.3). To aid clarity the following figures are based only on simulations with $\sigma < 0$.

Our results (Figures 2–4) show that increasing the absolute value of the selection coefficient stochastically decreases the conditional distribution of the age of the allele. For large (absolute) values, there is also a sharp decrease in the spread of the distribution of allele age.

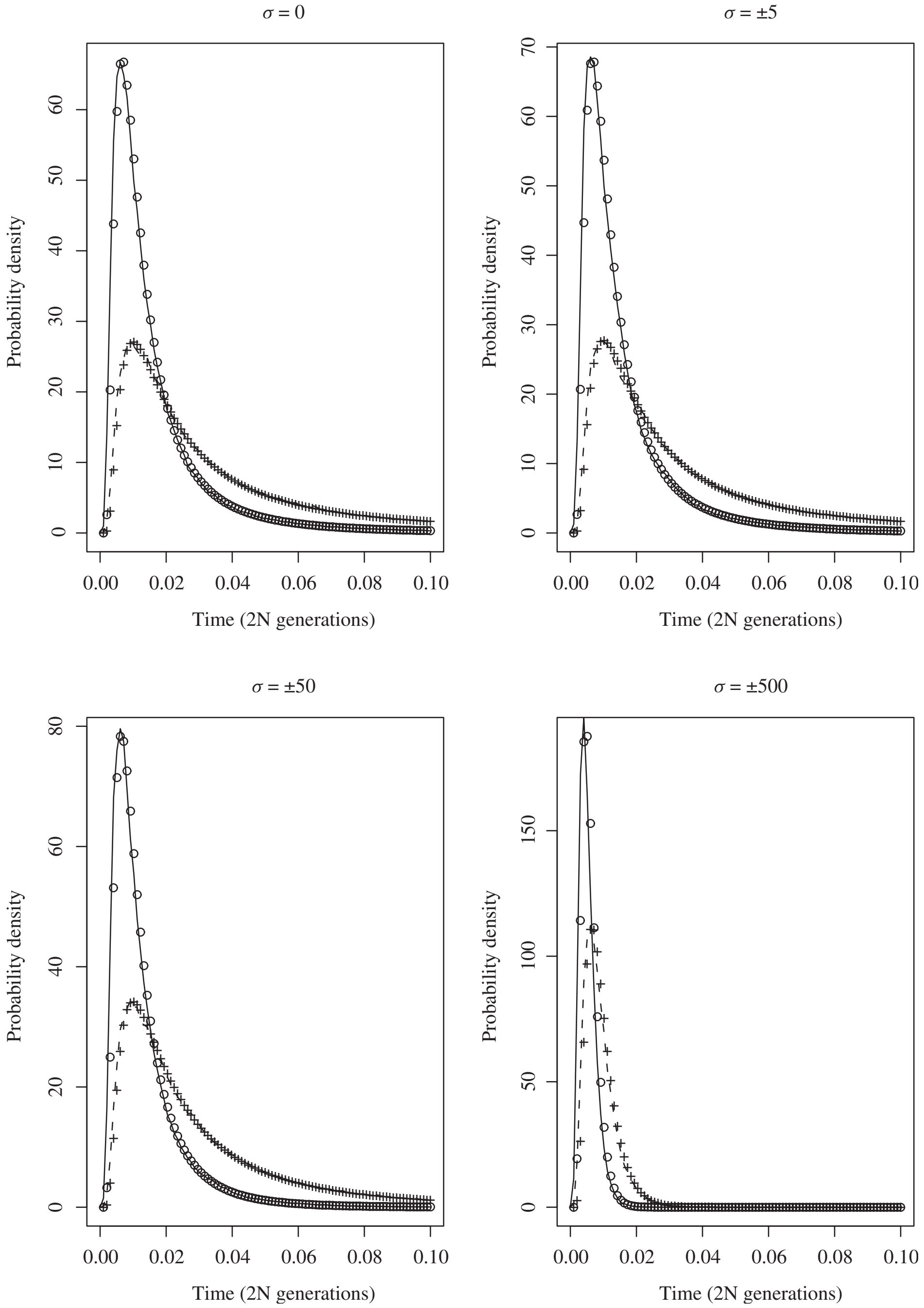

Figure 2. Histograms of samples from the conditional distributions of H (solid line) and A (dashed line), given a sample of $n = 5000$ chromosomes containing $n_1 = 50$ derived alleles. The histograms are overlain with the corresponding approximate distributions for H (circles) and A (crosses) obtained using equation (14) from Wiuf (2001a) to condition on the derived allele being at frequency 1% in the population.

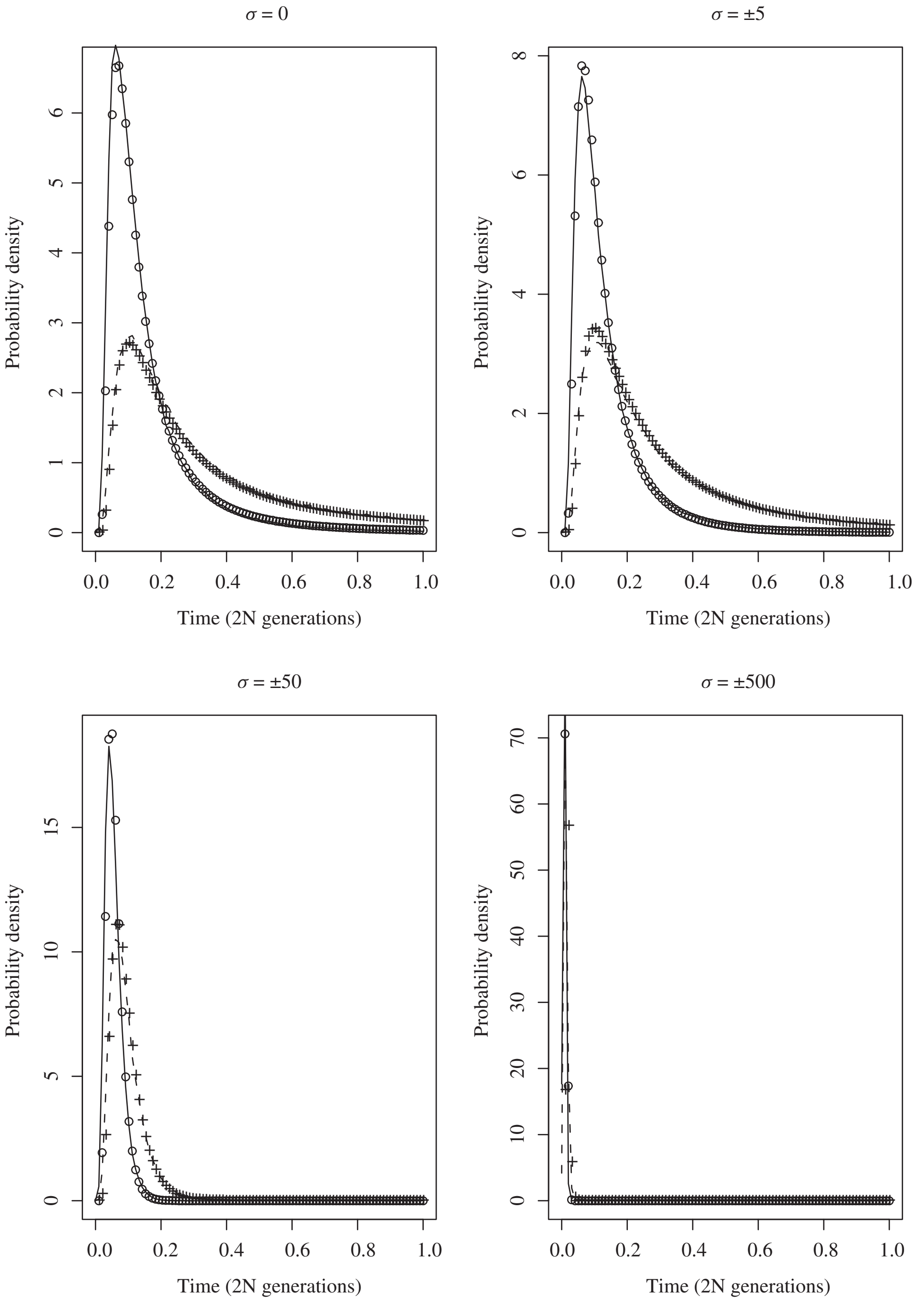

Figure 3. Histograms of samples from the conditional distributions of H (solid line) and A (dashed line), given a sample of $n = 1000$ chromosomes containing $n_1 = 100$ derived alleles. The histograms are overlain with the corresponding approximate distributions for H (circles) and A (crosses) obtained using equation (14) from Wiuf (2001a) to condition on the derived allele being at frequency 10% in the population.

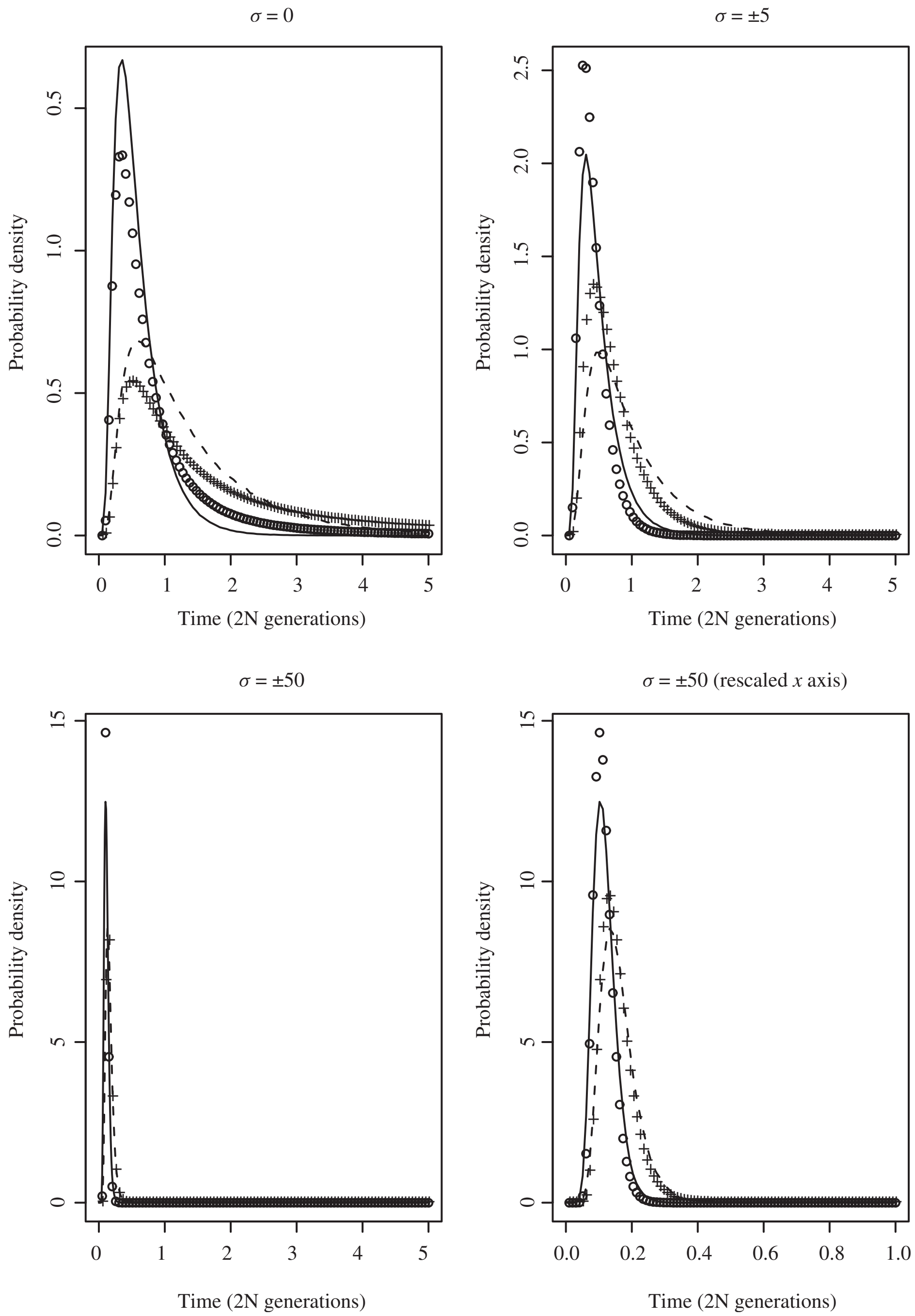

Figure 4. Histograms of samples from the conditional distributions of H (solid line) and A (dashed line), given a sample of $n = 100$ chromosomes containing $n_1 = 50$ derived alleles. The histograms are overlain with the corresponding approximate distributions for H (circles) and A (crosses) obtained using equation (14) from Wiuf (2001a) to condition on the derived allele being at frequency 50% in the population.

TABLE 1

Means and variances (in parentheses) of the age A of the mutation, and the height H, and total length of the subtree relating the mutant alleles, as a function of selection intensity and sample frequency of the mutant; see captions to Figures 2–4 for sample sizes

Frequency	$\|\sigma\|$	Age, A		Subtree height, H		Subtree length	
0.01	0	0.095	(0.073)	0.019	(0.00082)	0.088	(0.0048)
	5	0.075	(0.029)	0.018	(0.00074)	0.087	(0.00447)
	50	0.032	(0.0010)	0.014	(0.00013)	0.074	(0.0013)
	500	0.0091	(2.1×10^{-5})	0.0057	(6.3×10^{-6})	0.045	(0.00015)
0.1	0	0.51	(0.45)	0.16	(0.026)	0.95	(0.20)
	5	0.37	(0.16)	0.14	(0.016)	0.89	(0.14)
	50	0.097	(0.0024)	0.061	(0.00071)	0.59	(0.018)
	500	0.017	(2.6×10^{-5})	0.013	(9.9×10^{-6})	0.22	(0.00087)
0.5	0	1.4	(1.0)	0.60	(0.12)	3.5	(1.0)
	5	0.97	(0.41)	0.51	(0.082)	3.2	(0.90)
	50	0.16	(0.0030)	0.12	(0.0012)	1.3	(0.054)
	500	0.018	(2.7×10^{-5})	0.014	(1.0×10^{-5})	0.19	(0.00066)

The same effect occurs when studying the height H of the tree linking the derived alleles, although both effects are less marked. Other authors (see Wiuf, 2001a and references therein) have developed relatively simple analytic approximations for aspects of the genealogy under selection, including allele age, conditional on the current population frequency of the allele (which is slightly different from conditioning on the observed frequency in a sample of size n, as we do here). Comparison of our results with Wiuf's (2001a) approximations (Figures 2–4) shows that, at least in the case of a constant-size population, Wiuf's approximate distributions, which were obtained assuming that the derived allele is rare, are remarkably accurate for frequencies of the derived allele up to 10%. (In addition to the slightly different conditioning, the formula we use from Wiuf (2001a) relates to the subtree linking all chromosomes in the population of type 1, rather than a sample of these individuals — the close similarity between the results suggests that these differences are negligible for the sample sizes n used to produce the figures; unsurprisingly, comparisons using smaller sample sizes n display greater discrepancies.) As expected, the approximation is less good, though still informative, for a mutant with population frequency 0.5 (Figure 4).

Table 1 presents the mean and variance of the age, and the height and the total length of the subtree relating the mutant alleles. Selection has a proportionately greater effect in reducing the height of the subtree than in reducing its total length, suggesting that, as expected, selection tends to make the (conditional) tree more star-shaped.

4.2. Heterozygote advantage/disadvantage

We now consider the case where the two homozygous genotypes (00 and 11) each have scaled fitness 0, and the heterozygote has scaled fitness σ, which may be positive or negative. The case where the heterozygote has a selective advantage ($\sigma > 0$) is sometimes referred to as 'balancing selection', as it tends to preserve the existence of more than one type in the population. Thus, variants that are acted on by balancing selection tend to remain polymorphic in the population for longer than neutral variants, and so intuitively we might expect balancing selection to stochastically increase A and H. Here we use simulation to confirm this intuition.

In this case the integral in (4), which must be computed to find the transition rates in Algorithm 3.2, does not reduce to the convenient form (10), so we choose to approximate the integral very quickly and simply, and then to correct the sampled genealogies using importance sampling, as in Stephens & Donnelly (2000). Specifically, we approximate the integral in (4) as follows. The integral can be written as $\mathrm{E}(\exp(\frac{1}{2}\sigma^*(X_0, 1 - X_0)))$, where the expectation is over a Beta distribution for X_0. We approximate this by $\exp(\frac{1}{2}\sigma^*(\mathrm{E}(X_0), 1 - \mathrm{E}(X_0)))$. Although rather rough, the importance sampling procedure produced by this approximation is efficient for the samples and values of σ we are examining, producing Effective Sample Sizes (Kong, Liu & Wong, 1994) of around 0.8–1.0 times the actual sample size we use, which is $M = 100\,000$. In particular, this importance sampling approach appears very much more efficient than the one used by Slade (2000b).

As in the genic case we simulate 10 sets of times for each sampled genealogy, to produce (importance weighted) samples from the conditional distribution of A and H. Weighted histograms (Figure 5) confirm the intuition that the effect of balancing selection is to stochastically increase both A and H, through a considerable lengthening of the tail, with little change in the mode. In addition, the results show that, unlike the genic selection case, the conditional distributions depend on the sign of the selection coefficient, with A and H being stochastically decreased under heterozygote disadvantage. This appears to contradict the argument in Wiuf (2001b) that for low frequency variants the fitness of the rare homozygote can be ignored. So, conditional on a variant being at low frequency, the conditional distribution of A and H under heterozygote advantage or disadvantage should look very similar to the genic selection case, and in particular should not depend on the sign of σ. However, theoretical confirmation of this would be welcome.

5. Extension to intra-allelic variation

Molecular genetic data often document variation that is linked to (which for our purposes means physically very close to, on the DNA sequence of the chromosome) the locus under selection, in addition to the alleles present at the locus itself. For example, typically there may be two alleles present at the locus under selection, and information on DNA sequence variation, or microsatellite markers, close to the locus where it is thought that variation at these other positions has no effect on fitness. This type of variation has been called ‘intra-allelic’, because it relates to variation within the two alleles under selection.

Intra-allelic variation can be used in two ways. First, it can be incorporated in making inferences about genetic parameters, such as the selection intensities at the selected locus: for example, what is the likelihood of the full data (alleles at the selected locus and the linked neutral variation) as a function of the genetic parameter, or in a Bayesian analysis the posterior on these parameters, given the data? Second, given selection intensities, intra-allelic variation can be incorporated into ancestral inference: for example, given the sample frequency of the derived allele, and the pattern of linked neutral variation on chromosomes carrying the derived allele, what is the conditional distribution of the age of the mutation that created the allele? For recent applications of this kind, see e.g. Wiuf (2001b) and Slatkin (2001), and references therein.

We now show how to use the methods developed above in this context. (The full system, consisting of the selected site and the linked neutral sites, does not undergo parent-independent mutation.) Write the data as $\mathcal{D} = (\mathcal{D}_S, \mathcal{D}_N)$ where $\mathcal{D}_S$ denotes the data at the selected locus, and $\mathcal{D}_N$ the linked neutral variation. (Data on linked neutral variation are often only available for one of the alleles, but this poses no difficulty below.) Then if $\mathcal{T}$ denotes the genealogy at

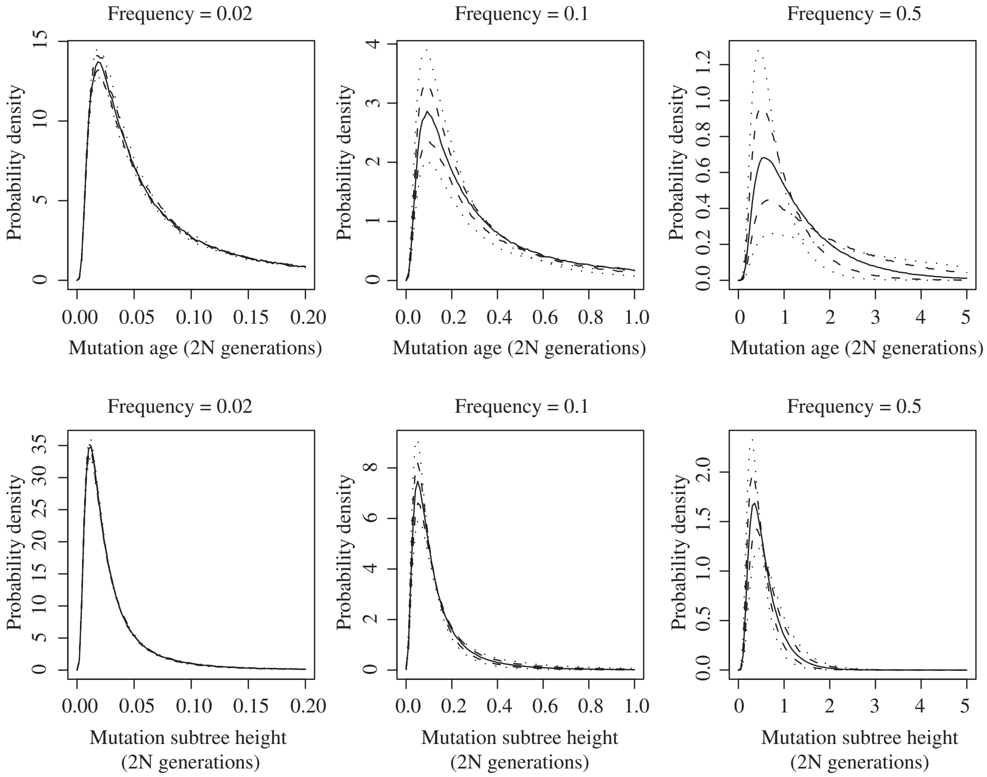

Figure 5. Curves of importance weighted samples from the conditional distributions of A (top panels) and H (bottom panels), given a sample of: (left) 20 derived alleles out of 1000 sampled chromosomes; (centre) 10 derived/100 sampled; (right) 50 derived/100 sampled. Reading from top to bottom at the mode, the lines in each figure correspond to $\sigma = -10, -5, 0, 5, 10$.

the selected locus, we can write the likelihood as

$$\Pr(\mathcal{D}) = \Pr(\mathcal{D}_S) \int_{\mathcal{T}} \Pr(\mathcal{T} \mid \mathcal{D}_S) \Pr(\mathcal{D}_N \mid \mathcal{T}) \,. \tag{11}$$

The first factor in (11) can be obtained, as in Donnelly *et al.* (2001), from (4), while the integral can be evaluated by Monte Carlo methods: simulate genealogies $\mathcal{T}^{(i)}$, $i = 1, 2, \ldots, M$, conditional on $\mathcal{D}_S$ according to Algorithm 3.2; for each $\mathcal{T}^{(i)}$ calculate $\Pr(\mathcal{D}_N \mid \mathcal{T}^{(i)})$; then approximate the integral by $M^{-1} \sum_i \Pr(\mathcal{D}_N \mid \mathcal{T}^{(i)})$. In calculating $\Pr(\mathcal{D}_N \mid \mathcal{T}^{(i)})$, it is necessary to model mutation for the linked neutral markers, and possibly also to model recombination for these and the selected locus.

For ancestral inference with intra-allelic variation, again simulate genealogies $\mathcal{T}^{(i)}$, $i = 1, 2, \ldots, M$, conditional on $\mathcal{D}_S$ according to Algorithm 3.2, and for each $\mathcal{T}^{(i)}$ calculate $w_i = \Pr(\mathcal{D}_N \mid \mathcal{T}^{(i)})$. Then the w_i can be interpreted as importance weights and the sample of trees $\mathcal{T}^{(i)}$, $i = 1, 2, \ldots, M$, as an importance sample from the conditional distribution of the genealogies given both $\mathcal{D}_S$ and $\mathcal{D}_N$.

In either setting, if we wished to condition on the event that the derived allele resulted from a unique mutation event, we would simply ignore genealogies $\mathcal{T}^{(i)}$ for which this is not the case.

Because our method gives exact samples from the conditional distribution of the genealogy given the data at the selected locus, it has the potential to be more efficient for analyses of intra-allelic variation than other existing approaches. However, the efficiency of any approach along these lines (including the one just described) will depend on how informative the linked neutral variation is about genealogy. If, as has been the case in applications to date, $\mathcal{D}_N$ consists of summary statistics (such as the number of mutations, or length of the region shared around the derived mutation) then this approach can work quite well. Developing practicable methods for handling more informative types of linked variation is a challenging problem. The difficulty with the methods in the literature, and the method just described, is analogous to the difficulty that makes full-likelihood coalescent-based inference challenging for mutation models that are more general than parent-independent mutation: informally, if genealogies are sampled from $\Pr(\mathcal{T} \mid \mathcal{D}_S)$, effectively no genealogies are seen for which the linked neutral data are relatively likely.

References

BERNARDO, J.M. & SMITH, A.F.M. (1994). *Bayesian Theory.* Chichester: Wiley.

DONNELLY, P. & KURTZ, T.G. (1999). Genealogical processes for Fleming–Viot models with selection and recombination. *Ann. Appl. Probab.* **9**, 1091–1148.

DONNELLY, P., NORDBORG, M. & JOYCE, P. (2001). Likelihoods and simulation methods for a class of non-neutral population genetics models. *Genetics* **159**, 853–867.

ETHERIDGE, A.M. (2000). *An Introduction to Superprocesses.* Providence Rhode Island: AMS.

ETHIER, S.N. & KURTZ, T. (1993). Fleming–Viot processes in population genetics. *SIAM J. Control Optim.* **31**, 345–386.

EXCOFFIER, L. (2001). Analysis of population subdivision. In *Handbook of Statistical Genetics*, eds D.J. Balding, M. Bishop & C. Cannings, pp. 271–307. Chichester: Wiley.

FEARNHEAD, P. (2002). The common ancestor at a non-neutral locus. *J. Appl. Probab.* **39**, 1–17.

FEARNHEAD, P.N. (2001). Perfect simulation from population genetic models with selection. *Theoret. Popul. Biol.* **59**, 263–279.

FELSENSTEIN, J., KUHNER, M.K., YAMATO, J. & BEERLI, P. (1999). Likelihoods on coalescents: a Monte Carlo sampling approach to inferring parameters from population samples of molecular data. In *Statistics in Molecular Biology and Genetics, Volume 33 of IMS Lecture Notes — Monograph Series*, ed. F. Seillier-Moiseiwitsch, pp. 163–185. Hayward, California: Institute of Mathematical Statistics and American Mathematical Society.

FRISSE, L., HUDSON, R.R., BARTOSZEWICZ, A., WALL, J.D., DONFACK, J. & RIENZO, A.D. (2001). Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *Amer. J. Human Genetics* **69**, 831–843.

GRIFFITHS, R.C. (2003). The frequency spectrum of a mutation and its age, in a general diffusion model. *Theoret. Popul. Biol.* **64**, 241–251.

GRIFFITHS, R.C. & TAVARÉ, S. (1995). Unrooted tree probabilities in the infinitely-many-sites model. *Math. Biosci.* **127**, 77–98.

GRIFFITHS, R.C. & TAVARÉ, S. (1998). The age of a mutation in a general coalescent tree. *Stochastic Models* **14**, 273–295.

HARTL, D.L. & SAWYER, S.A. (1991). Inference of selection and recombination from nucleotide sequence data. *J. Evol. Biol.* **4**, 519–532.

KELLY, F.P. (1979). *Reversibility and Stochastic Networks.* Chichester: Wiley.

KIMURA, M. & OHTA, T. (1973). The age of a neutral mutation persisting in a finite population. *Genetics* **75**, 199–212.

KONG, A., LIU, J.S. & WONG, W.H. (1994). Sequential imputation and Bayesian missing data problems. *J. Amer. Statist. Assoc.* **89**(425), 278–288.

KRONE, S.M. & NEUHAUSER, C. (1997). Ancestral processes with selection. *Theoret. Popul. Biol.* **51**, 210–237.

MARUYAMA, T. (1974). The age of an allele in a finite population. *Genetical Research* **23**, 137–143.

NEUHAUSER, C. & KRONE, S.M. (1997). The genealogy of samples in models with selection. *Genetics* **145**, 519–534.

ROUSSET, F. (2001). Inferences from spatial population genetics. In *Handbook of Statistical Genetics*, eds D.J. Balding, M. Bishop & C. Cannings, pp. 239–269. Chichester: Wiley.

SLADE, P. (2000a). Simulation of selected genealogies. *Theoret. Popul. Biol.* **57**, 35–49.

SLADE, P. (2000b). Most recent common ancestor probability distributions in gene genealogies under selection. *Theoret. Popul. Biol.* **58**, 291–305.

SLATKIN, M. (2001). Simulating genealogies of selected alleles in a population of variable size. *Genetical Research (Cambr.)* **78**, 49–57.

STEPHENS, M. (2000). Times on trees and the age of an allele. *Theoret. Popul. Biol.* **57**, 109–119.

STEPHENS, M. (2001). Inference under the coalescent. In *Handbook of Statistical Genetics*, eds D.J. Balding, M. Bishop & C. Cannings, pp. 213–238. Chichester: Wiley.

STEPHENS, M. & DONNELLY. P. (2000). Inference in molecular population genetics. *J. Roy. Statist. Soc. Ser. B* **62**, 605–655.

WIUF, C. (2001a). Rare alleles and selection. *Theoret. Popul. Biol.* **59**, 287–296.

WIUF, C. (2001b). Do $\Delta F508$ heterozygotes have a selective advantage? *Genetical Research (Cambr.)* **78**, 41–47.

WIUF, C. & DONNELLY, P.J. (1999). Conditional genealogies and the age of a neutral mutant. *Theoret. Popul. Biol.* **56**, 183–201.

WRIGHT, S. (1949). Adaptation and selection. In *Genetics, Palaeontology, and Evolution,* eds G.L. Jepson, G.G. Simpson & E. Mayr, pp. 365–389. Princeton, NJ: Princeton University Press.

6. Discussion by Susan R. Wilson

(*Centre for Mathematics & its Applications and Centre for Bioinformation Science, Mathematical Sciences Institute, The Australian National University*)

Accommodating the genealogical history of a sample, given the sample composition, has taken its rightful place of importance in theoretical population genetic modelling and is the basis of this interesting paper by Stephens & Donnelly (henceforth referred to as S&D). Allowing for this unknown structure is also important in many other areas of genetic data analysis, and some examples are given below. In S&D, concern is with inference under a selection model, and essentially their approach is based on determining an appropriate probability model for the unknown genealogical tree. This line of research has strong Australian roots, and so it is most appropriate that S&D's paper was selected for the Editor's Invited Paper session at a Statistical Society of Australia conference.

The term 'coalescent' (Section 2.3) was coined by Kingman in 1982, and in recounting the origins of this useful tool, Kingman (2000) traces the story back to his travels in Australia during 1974 when he was enthused by Warren Ewens in Melbourne and Pat Moran in Canberra who were working on models for neutral evolution in finite populations. Since Kingman's pioneering publications, many researchers have advanced these ideas that have been evolving in tandem with progress in computer technology. A large proportion of researchers who work on the development of these ideas are Australian.

Coalescent methods are used to understand the pattern of shared ancestry among alleles; that is, the shape of the genealogy or evolutionary tree. The shape contains information about the effective population size and demographic history of the population. The standard coalescent model is based on the assumptions of neutrality and panmixia, that the population size is large and constant, and that recombination is absent (or negligible). A broad class of mutation models can be incorporated into the coalescent. The basic assumptions can all be weakened, but at some computational cost. Competing software is available (via the web) and approaches taken by different developers can vary in terms of both output and efficiency, dependent on, say, whether the mutation rate is low or high, or whether the migration rate is high or low.

Here S&D have used the ancestral selection graph to develop algorithmic approaches to enable the important effect of (reasonably general diploid) selection to be evaluated. Their focus is population genetic modelling. We can ask obvious questions about generalizations to other forms of selection, and about weakening of assumptions to allow for variable population size, non-random mating, population subdivision and recombination. Further, if several loci are interacting in such a way as to affect selection, then what might appear to be happening if only one of these loci is considered? What if the selection differs between subpopulations or over time, perhaps due to environmental changes? More importantly, in general when analysing real, as opposed to simulated, data, how do we know if the locus is neutral or under selection of one form or another and whether the selection is constant or varying? Also, in the simulations, additive selection (Section 4.1) and balancing selection (Section 4.2) are assumed. What is the likely consequence if the selection is multiplicative or of some other, possibly more realistic, form?

In Section 4, S&D give us the results from simulations concerning the age of a non-neutral allele, as well as the height H of the subtree relating chromosomes that share the mutation and the subtree length. Is there an intuitive reason why the coefficient of variation (CV) for the height H is about double the CV for the subtree length (Table 1)? My own interests lie in statistical analyses of genetic data. As S&D note, H is of potential applied interest as it relates to the efficacy of association methods for finding disease genes as well as 'the interpretation of linked variation'. Certainly simulated data are very important for understanding theoretical population genetics results that in turn can give some insight into the results from real data analyses. However, the real world seems to always be much more complicated, and later I give an example of an actual problem interpreting the association results for a complex disease (Crohn's disease).

For me, the first sentence of the paper was tantalizing: 'there are now large and growing amounts of data' from the human and other genome projects and from experiments. Unfortunately no such data appear or are considered by S&D. So, for those who are not familiar with this research area, it is useful to briefly describe two areas of genetic data analysis where coalescent-type methods are proving useful.

First, in anthropology, genetic data from contemporary populations provide some evidence concerning ancestral generations. For example (recognizing the limitations of single locus studies) Harding & Liu (2003) sampled β-globin gene data from Papua New Guinea (PNG) Highlands and Kenya. Their primary question concerned the major direction of gene flow between Africa and PNG. They studied several population genetic models in the coalescent framework and found that models that incorporated population subdivision, regardless of the pattern of migration, provided a better explanation than models that assume random mating. They found that a population subdivision model with symmetric migration had a higher likelihood than a model with migration exclusively out of Africa, with the best model being subdivision with high rates of migration into Africa out of PNG. However, it is quite likely that natural selection has affected patterns of diversity for these data, although the relationship between the site Harding and Liu considered in the β-globin gene and selection in the region is not clear*. Generally it is preferable to use neutral loci to infer population demographic history, but how do we know if the chosen loci are really neutral, not only now but also in the past that is of interest? Also, there may well have been differential selection between the sampled populations.

Second, in human genetics, several applications of coalescent-type methods are starting to appear. One recent example concerned the way to represent human population genetic structure in the analysis of drug safety and efficacy, and how to relate this structure to drug response. In this context Wilson *et al.* (2001) use coalescent models to show that the commonly used ethnic labels can be both insufficient and inaccurate representations of the inferred genetic clusters. Another example is the analysis of case-control genetic association studies in structured populations. Pritchard *et al.* (2000) propose the use of a set of unlinked genetic markers to infer details of population structure, and to estimate the ancestry of the sampled individuals, and they then use this information to test for associations for a candidate marker within subpopulations. Morris, Whittaker & Balding (2002) develop a multipoint model for complete marker haplotypes, conditional on the genealogy underlying a sample of case chromosomes. Uncertainty about ancestry is accommodated in a Bayesian MCMC framework, by simulating over the distribution of ancestral marker haplotypes and genealogical trees. Morris *et al.* (2002) apply their 'shattered coalescent' approach, which has the advantage of allowing for multiple founding mutations at the disease locus as well as for sporadic cases of disease, to case-control cystic fibrosis (CF) data, and find that it performs well compared with other methods based on alternative models. As they note, the simple models so far used cannot 'fully capture the reality underlying the actual genealogy, which will be further complicated by factors including selection, population substructure, and ascertainment'. Another complicating factor is epistasis (gene–gene interaction) and gene–environment interactions. In fact it can be shown that all methods of association analysis encounter problems in the presence of epistasis (Wilson, 2001).

Now CF is a single locus Mendelian disease, and localizations for many Mendelian diseases, including CF, have been realized, without needing coalescent methods. On the other hand, finding and understanding the genetic basis of complex disease where many genetic loci as well as environmental interactions may be involved in disease susceptibility is proving to be much more challenging. Arguably the best localization result to date for a complex disease is for Crohn's disease (CD). CD is one of the two main types of inflammatory bowel disease, and occurs primarily in young adults, with an estimated prevalence of about 1 in 1000 in western countries. Its incidence has increased markedly over the past 50 years, and there is familial aggregation of the disease. The first susceptibility localization for IBD1 on chromosome 16 was found using multipoint linkage methods in 1996, followed by variable support for this localization, with convincing replication being provided by an international collaboration (IBD International Genetics Consortium, 2001). What insight might the results in S&D give concerning 'the interpretation of linked variation' for the following results?

Last year the IBD1 gene was identified in back-to-back publications, (i) by a positional cloning approach, based on linkage analysis followed by linkage disequilibrium mapping in European CD families, and (ii) by recognizing NOD2 as a gene that encodes a protein with

*There are many amino acid substitutions and deletions in the globin chain leading to haemoglobinophathies, blood disease. One of the most famous in the β chain is sickle cell anaemia. Those who are heterozygous for the abnormal gene are apparently at a selective advantage against malaria.

homology to plant disease resistance gene products, followed by linkage disequilibrium mapping in American CD families. The first study found three alleles; the second, only one of these three. Another study (iii) found this common allele in a second European CD population, and an Australian study (iv) has identified all three; details on all four studies can be found in Cavanaugh *et al.* (2002). Functional evidence supports the characterization of NOD2 as the IBD1 locus, especially the allele $980_{fs}981$ that is common to all studies. Now for IBD1, by far the highest multipoint score for linkage occurs in the Australian CD families, although our sample sizes are small compared with other studies. However, the association results (for all three alleles) are not statistically significant in the Australian data, although they are highly significant in all the other studies. Also, considering allele $980_{fs}981$, in the above-identified four CD populations the frequencies are (i) 0.12, (ii) 0.08, (iii) 0.16, (iv) 0.07, respectively. The allele frequencies of $980_{fs}981$ in the controls are (i) 0.02, (ii) 0.04, (iii) 0.04, (iv) 0.01, respectively. The value for the Australian population of controls (iv) is significantly lower than the other three control population values that are not statistically different. Taken together, these results are intriguing. Certainly the linkage evidence suggests that there are very likely to be additional IBD1 susceptibility alleles to be found in Australian CD. Also implicated are differences between Australia and elsewhere in environmental effects on development of CD. None of the above analyses used coalescent-type methods. To do so, how important is it to know whether the NOD2 locus is neutral or undergoing selection? Also, informally we can make some adjustment of summary statistics to take account of the non-random sampling of the families, but it would be helpful to have a more precise indication of the extent to which they need to be adjusted.

In conclusion, there are significant statistical and computational challenges in this niche of statistical genetics, and this is certainly an exciting area for current and future research.

References

Cavanaugh, J.A., Adams, K.E., Quak, E.J., Bryce, M.E., O'Callaghan, N.J., Beck, N.R., Rodgers, H.J., Magarry, G.R., Butler, J.R., Eaden, J.A., Roberts-Thomson, I.C., Pavli, P., Wilson, S.R. & Callen, D.F. (2003). NOD2 risk alleles in the development of Crohn's disease in the Australian population. *Ann. Human Genetics* **67**, 35–41.

Harding, R.M. & Liu, Y. (2003). Time scales for genetic diversity in Melanesia: a look at some evidence for estimates of 100,000 years or more. In *Papuan Pasts: Investigations into the Cultural, Linguistic and Biological History of the Papuan-speaking Peoples*, eds A. Pawley, R. Attenborough, J. Golson & R. Hide. (To appear.)

IBD International Genetics Consortium (2001). International collaboration provides convincing linkage replication in complex disease through analysis of a large pooled data set: Crohn disease and chromosome 16. *Amer. J. Human Genetics* **68**, 1165–1171.

Kingman, J.F.C. (2000). Origins of the coalescent: 1974–1982. *Genetics* **156**, 1461–1463.

Morris, A.P., Whittaker, J.C. & Balding, D.J. (2002). Fine-scale mapping of disease loci via shattered coalescent modelling of genealogies. *Amer. J. Human Genetics* **70**, 686–707.

Pritchard, J.K., Stephens, M., Rosenberg, N.A. & Donnelly, P. (2000). Association mapping in structured populations. *Amer. J. Human Genetics* **67**, 170–181.

Wilson, J.F., Weale, M.E., Smith, A.C., Gratrix, F., Fletcher, B., Thomas, M.G., Bradman, N. & Goldstein, D.B. (2001). Population genetic structure of variable drug response. *Nature Genetics* **29**, 265–269.

Wilson, S.R. (2001). Epistasis and its possible effects on transmission disequilibrium tests. *Ann. Human Genetics* **62**, 565–575.

7. Discussion by Melanie Bahlo, Russell Thomson and Terry Speed

(*Division of Genetics & Bioinformatics, Walter & Eliza Hall Institute of Medical Research*)

First, we would like to congratulate the authors for their fine contribution to the field, in which they extend a standard model to permit a more diverse range of selection models. The paper presents the algorithms that allow full likelihood calculations for selection under a wider range of parent-independent mutation models. Most previous studies have considered the two-allele mutation models only.

The authors show how their method could be used, in theory, in conjunction with a number of fully linked loci, and recombination is not included in the model. Possible extensions, to allow ancestral inference with informative markers that are in linkage disequilibrium (LD) with a causal mutation, are mentioned.

The methods of this paper require an assumption of a constant (and sufficiently large) population size over time. References (Wiuf, 2001; Slatkin, 2001) are made to approximate approaches that handle changes in population size.

The authors forgo much of the complex stochastic process theory in this expository paper. They provide an introduction to the underlying Fleming–Viot diffusion process with selection. This introduction presents understanding of their new method, but the complex recursive relationships for the model are not shown. Instead the results are presented as simple algorithms, ready to implement.

After our initial reading of the paper, we were not entirely clear whether the paper was presenting algorithms for use in practical gene finding, as these authors have done in the past, or whether it was a more theoretical contribution, intending to provide qualitative insights relating to the different forms of selection assumed in the simulations. Our uncertainty was removed in the authors' rejoinder, but below we offer the questions we posed.

Our major concern was with the possible use of the results in the paper in gene-finding projects. We feel that there are a number of issues which continue to prevent such methods from being applied to real data, falling under the following headings.

Computational complexity. The large state space (or possible missing data as the authors view it) will always present a challenging computational problem. This is attacked by using a form of importance sampling, as in this paper, or by MCMC methods (Felsenstein *et al.*, 1999).

Linkage disequilibrium. To model LD in even the simplest cases demands the inclusion of recombination. This has already been addressed for certain restricted mutation models. To make it appropriate and useful for real-world LD studies requires mutation models appropriate for the markers currently used: microsatellite short tandem repeats (STRs) and single nucleotide polymorphisms (SNPs).

Selection. The paper is based on a model for selection. How generally applicable is this model? It obviously has some pleasing mathematical properties which ensure that the tree will have a finite time until a single ancestor is reached, but what evidence is there to support such a model for any real-world populations?

Cost. The approach presented requires haplotypic data, which are currently costly to obtain for much of the human genome. Mitochondrial DNA and the non-recombining region of the Y chromosome offer two regions of the genome where haplotypic data can be derived cheaply since these data are not diploid and do not recombine.

Comparison of methods. Wiuf (2001) and the authors' results show remarkable agreement for those cases where the simpler method is appropriate. What computational efforts are necessary for both these methods?

References

FELSENSTEIN, J., KUHNER, M.K., YAMATO, J. & BEERLI, P. (1999). Likelihoods on coalescents: a Monte Carlo sampling approach to inferring parameters from population samples of molecular data. In *Statistics in Molecular Biology and Genetics. Volume 33 of IMS Lecture Notes — Monograph Series*, ed. F. Seillier-Moiseiwitsch, pp. 163–185. Hayward, California: Institute of Mathematical Statistics and American Mathematical Society.

SLATKIN, M. (2001). Simulating genealogies of selected alleles in a population of variable size. *Genetical Research (Cambr.)* **78**, 49–57.

WIUF, C. (2001). Rare alleles and selection. *Theoret. Popul. Biol.* **59**, 287–296.

8. Authors' rejoinder

It is a pleasure to thank the discussants for their interesting comments. They point to the excitement of the science of modern genetics, the potential for central contributions by statisticians, and some of the achievements to date in that direction.

Perhaps we should have been more explicit, but our primary motivation (here as elsewhere) is the science: a desire to improve our understanding of the context and roles of the main evolutionary forces; of the history of populations and their genomes; and of the genetics of common human diseases.

Wilson summarized both the history (including some of the Australian influences) and the pervasiveness in modern population genetics, of coalescent approaches. Selective neutrality is both a reasonable and a common assumption in many contexts. Under the assumption of neutrality, the mathematical theory of the coalescent is mature and well understood, and its use for statistical inference is becoming well developed. Wilson was kind enough to reference some of our own work directly on applications to human genetics (Pritchard *et al.*, 2000), and the application by others of statistical tools we have developed (Wilson *et al.*, 2001), but as she noted, use of the (neutral) coalescent is now widespread in the literature.

This is much less true when the complications of natural selection are included in the models. Indeed, until a few years ago, there was virtually no available mathematical theory, much less inference methods, under selection. In a sense, progress on non-neutral coalescents is perhaps a decade behind that in the neutral case. We agree with the discussants, and do not see our paper, nor indeed this part of the field, as yet providing off-the-shelf methods for application in human genetics. Instead, we would like to hope that the work we have described (our own, and others') will provide useful steps along a road leading eventually to solutions to the pressing applied problems referred to in the discussion. Our philosophy for stochastic modelling is to weaken assumptions in well-understood models step by step, gaining insights at each stage of this process, rather than to leap immediately to the most realistic scenario. In this, we may differ from the discussants. We do not know the answers to the questions we were asked about extending our analyses to more complicated demographic scenarios, and suspect that they are hard. For those not familiar with this area, it is worth noting that even the standard neutral coalescent involves many assumptions that seem from the outside implausible for, say, human populations. Nonetheless, the application of the simple model has proved immensely helpful, with important consequences in many applied problems. An old statistical aphorism refers to all models as being false, but some being useful. The coalescent now seems well entrenched in the 'useful' category, and while we would welcome further progress, we are perhaps less concerned than the discussants that our analysis does not include all the complications of the real world.

So have we learnt anything that might be practically useful? We think so, on several fronts. First, as Speed and his colleagues have noted, there is value in qualitative insights: armed with such insights about the way in which natural selection can affect genealogies, we can ask how existing methods, which assume neutrality, might be misleading were selection to be acting. Next, we have described how to do efficient ancestral inference for a particular class of models. (In answer to a question posed by both discussants, the model of selection is pretty general; it is the restriction on the mutation model that we see as being much more limiting in practice.) Knowing the 'right' answer for these models, even for relatively small data sets, provides an extremely helpful yardstick for assessing other, more *ad hoc* or approximate approaches — without knowing the truth, comparison of different approximations may not be especially enlightening. In addition, our results could be helpful to those who seek to make qualitative inferences about the selective pressures acting on a variant by estimating its age (e.g. from patterns of linked variation). At a minimum here, it should be possible to determine whether older or younger ages are evidence for or against particular selective regimes: a task on which intuition and informal arguments can be unreliable.

Finally, for complicated inference problems even under neutrality, one approach that is proving useful when full-likelihood methods are impractical is the use of likelihood for low dimensional summaries of the data, where simulation methods are used to approximate these likelihoods (see for example Beaumont, Zhang & Balding (2002) and references therein). The key idea, of simulating genealogies under the model and from that approximating the distribution of the summary statistics, can now also be implemented for the class of non-neutral models we have considered. This seems an interesting direction for further work (G. McVean, pers. comm.), and a promising alternative when the full likelihood inference approaches described at the end of the paper are impractical, for example in incorporating recombination and linkage disequilibrium, as Speed and colleagues have challenged. Both the problem of deciding whether natural selection *has* been acting in a particular genomic region, and gene mapping (either indirectly through inferring a target for selection, or directly) may be amenable to this kind of approach. (In response to one of Wilson's questions, a likelihood-based approach to detecting selection at a locus from data at that locus is developed by Donnelly, Nordborg & Joyce (2001), for the class of models considered here. For a more general review of this question, see Nielsen (2001).)

We turn now to some of the particular issues raised in the discussion. Speed and his colleagues noted that we did not give the 'complex recursive relationships for the model'. This was not an accidental omission. As Slade (2000a,b) has shown, in applying an approach initially due to Griffiths and Tavaré, these recursive relationships can be used to derive an importance sampling approach to ancestral inference under selection. Not only is this route not transparent, it often leads to very inefficient importance sampling schemes. We have shown elsewhere (Stephens & Donnelly, 2000) that it is usually much more efficient to develop importance sampling approaches directly from properties of the underlying stochastic model, for example via the discrete construction of the measure-valued process. For the models of this paper, this approach potentially allows exact sampling from the conditional distribution of the genealogy given the sample data. Not only are the recursive relationships an unnecessary complication, they actively lead in the wrong direction.

Speed and colleagues pointed out that our approach requires haplotype information. This may be available even for diploid regions of the genome, either experimentally or from data from pedigrees. If not, there are now statistical approaches for inferring haplotypes from genotype information that are surprisingly accurate (e.g. Stephens & Donnelly, 2003). They also asked for a comparison of the computational cost of our approach compared to Wiuf's (2001a) approximation. Wiuf's result is analytic, so involves essentially no computational cost (though the accuracy of the approximation needs to be checked computationally). For our approach, computational cost depends greatly on sample size, mutation frequency, and strength and type of selection, but to give an idea, most of the figures in the paper were based on results that took hours, but not days, to produce (CPU time on our desktop machine with an 800 MHz processor).

Wilson presented examples of some of the complications in disease studies of the different genetic and environmental backgrounds at work in different populations. Again this seems a very challenging problem. It may be that the current international HapMap project, which aims to characterize patterns of haplotypic diversity in distinct human populations, will be helpful here, as will some of the very large proposed prospective population studies (for example, the UK 'Biobank' will be a prospective study of 500 000 UK adults aimed at unravelling some of the gene–environment interactions relevant to common human disease).

Additional references

BEAUMONT, M., ZHANG, W. & BALDING, D.J. (2002). Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025–2035.

NIELSEN, R. (2001). Statistical tests of selective neutrality in the age of genomics. *Heredity* **86**, 641–647.

STEPHENS, M. & DONNELLY, P. (2003). A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Amer. J. Human Genetics* (to appear).