J. R. Statist. Soc. B (2002)
64, *Part* 4, *pp.* 695–715

Assessing population differentiation and isolation from single-nucleotide polymorphism data

George Nicholson,

University of Oxford, UK

Albert V. Smith, Frosti Jónsson, Ómar Gústafsson and Kári Stefánsson

deCODE genetics, Reykjavik, Iceland

and Peter Donnelly

University of Oxford, UK

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on 'Statistical modelling and analysis of genetic data' on Wednesday, May 22nd, 2002*, Professor D. Firth and Professor R. A. Bailey *in the Chair*]

Summary. We introduce a new, hierarchical, model for single-nucleotide polymorphism allele frequencies in a structured population, which is naturally fitted via Markov chain Monte Carlo methods. There is one parameter for each population, closely analogous to a population-specific version of Wright's F_{ST}, which can be interpreted as measuring how isolated the relevant population has been. Our model includes the effects of single-nucleotide polymorphism ascertainment and is motivated by population genetics considerations, explicitly in the transient setting after divergence of populations, rather than as the equilibrium of a stochastic model, as is traditionally the case. For the sizes of data set that we consider the method provides good parameter estimates and considerably outperforms estimation methods analogous to those currently used in practice. We apply the method to one new and one existing human data set, each with rather different characteristics—the first consisting of three rather close European populations; the second of four populations taken from across the globe. A novelty of our framework is that the fit of the underlying model can be assessed easily, and these results are encouraging for both data sets analysed. Our analysis suggests that Iceland is more differentiated than the other two European populations (France and Utah), a finding which is consistent with the historical record, but not obvious from comparisons of simple summary statistics.

Keywords: Bottleneck; Coalescent; Demography; Fixation indices; Markov chain Monte Carlo methods; Population histories

1. Introduction

Populations of the same species in different geographical regions tend to differ genetically. Such differences may in part reflect their adaptation to different environments, or they may simply be the result of chance events in the evolutionary histories of the populations since their divergence. Any exchange of genes between the populations, through migration or interbreeding, will act to reduce the differences between them.

It has thus long been of interest in population genetics to describe the nature of genetic dif-

Address for correspondence: George Nicholson, Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG, UK.
E-mail: nicholson@stats.ox.ac.uk

 1369–7412/02/64695

ferentiation that is observed in real populations, and if possible to understand the mechanisms which are responsible for it. Knowledge of the extent of geographical population structure is important for a range of applications, including conservation biology, the mapping of disease genes, inference of population histories and forensic deoxyribonucleic acid (DNA) profiling. It is also vital to the correct interpretation of patterns of genetic diversity and plays a central role in the shifting balance theory of evolution (Wright, 1932, 1982).

Consider a collection of populations, from each of which is available a sample of individuals whose genetic type is assessed at a number of genetic markers, or *loci*. (Throughout, we shall conceptualize a population as a group of individuals that is at least approximately genetically homogeneous, the members of which mate at random with respect to genetic type at the markers of interest. Whether this is a reasonable description of any specific group of individuals, taken at a particular level of geographical stratification, is obviously an empirical question.) There is a vast, and in places confusing, literature on quantitative approaches to describing genetic differentiation, and for inferring from it various underlying evolutionary or demographic parameters of interest. These have been applied to classical genetic markers, restriction fragment length polymorphisms, to DNA sequence data and to data at tandem repeat loci. For background, see the reviews by Excoffier (2001) and Rousset (2001) and references therein.

Here we consider the coming generation of genetic data, namely single-nucleotide polymorphisms (SNPs). An SNP is a single position in the DNA at which there is variation between different individuals within a species. (Virtually all SNPs only exhibit two variants, and we shall make this assumption in what follows.) Most positions in the genome are identical across all members of a population. The idea behind the use of SNPs is that an initial study or studies will search for them—so-called SNP discovery—by a variety of experimental methods, and then subsequent studies can assess genetic variability only at sites known in advance to exhibit polymorphism. This greatly reduces the experimental cost of assessing variability. The promised advent of cheap high throughput SNP genotyping technology means that it is likely soon to be routine to collect data on large numbers of SNP loci. (As an example, in the most characterized organism for SNPs, namely humans, well over a million SNPs have been identified.) Although SNPs promise previously undreamt of volumes of data, they come at some analytical cost. The process of SNP discovery introduces an ascertainment effect—informally, the more polymorphic an SNP is (by which we mean the larger the relative frequency of the less common, or *minor*, variant) the more likely it is to be found during the SNP discovery process. A failure to allow properly for this ascertainment effect in subsequent analyses can lead to misleading conclusions.

In general the patterns of spatial variation at a particular marker will depend on the demographic history of the populations, the mutation process at the marker and the effects of selection at or close to the marker. It has thus been typical to quantify spatial structure separately for each of the loci in the study. In contrast, a single summary, or estimator, is usually used across all populations.

For SNPs we take a different perspective. This paper introduces a new statistical model, motivated by population genetics considerations explicitly in a non-equilibrium setting, for SNP data. For reasons given in the next section, the model has the same probabilistic structure at each SNP locus, with population-specific parameters aiming to capture historical demographic differences between populations. An important feature of the model is that it explicitly includes (and hence corrects appropriately for) the SNP ascertainment process. We fit the model by using Markov chain Monte Carlo (MCMC) methods, here in a Bayesian framework, and show that this outperforms estimation methods analogous to those which are currently popular in

this context. We apply our new method to two human data sets: one from three European populations; the other from a worldwide sample.

The availability of data from many independent loci is rather novel in population genetics. It has two important consequences. The first is that, whereas individual SNP loci carry very little information about demographic processes, replication over many such loci promises much more precise inferences than have been possible from other kinds of data. Secondly, and perhaps even more importantly, such data allow for an assessment of model fit. We regard the discussion of such diagnostics as an important aspect of this paper.

In contrast with our approach, much of the literature in this area is concerned with the equilibrium of various stochastic models for the evolution of a spatially structured collection of populations. For some simple models, especially the so-called *island model* (e.g. Hudson (1998) and Rousset (2001)), expressions are available relating certain quantities at equilibrium to parameters such as migration and mutation rates. It is common to use such relationships, notably via estimates of F_{ST} (see Section 2.3), to estimate migration rates at stationarity, via the method of moments. The estimation of migration rates through this route is fraught with difficulties (e.g. Whitlock and McCauley (1999)). The stationarity assumption is rarely likely to obtain and is effectively impossible to check; the procedure can be exquisitely sensitive to the (uncheckable) population dynamics which are assumed; and the precision of the resulting moment estimators is in any case disappointing. None-the-less the approach is much used in practice.

Our model is motivated as an approximation to one arising when populations split and evolve in subsequent isolation (sometimes called the *pure drift* setting). Exact, though complicated, expressions are available for likelihoods under this model for the simplest demographic scenarios, and full likelihood inference methods (reviewed in Beaumont (2001)) have been developed for parameter estimation, although practicability may be limited for large numbers of loci.

Perhaps surprisingly, a model with similar structure to ours arises (again as an approximation, and for specific assumptions about migration) in the 'opposite' demographic setting of equilibrium under migration and genetic drift. It was developed by Balding and Nichols (e.g. Balding and Nichols (1995)) and has been widely applied in the context of the loci used for forensic DNA profiling (e.g. Balding and Nichols (1997), Roeder *et al.* (1998) and Foreman *et al.* (1997)). In the Balding–Nichols (BN) model, allele frequencies at a particular locus in each population are assumed to vary independently about 'mean' values, according to a symmetric Dirichlet distribution (beta for biallelic systems like SNPs), with an additional parameter for each locus–population pair specifying the variance structure. We believe that the different perspective here (explicitly transient) is appropriate for the different genetic systems (SNPs) on which we focus. There is a literature (reviewed in Beaumont (2001)) on distinguishing between the BN model and the exact pure drift model for real data. Although the model that is introduced in this paper agrees with the BN model to first and second moments, an important difference in principle is that because it has mass outside [0, 1] the use of a normal, but not of a beta, distribution allows for variation to be lost (population allele frequencies of 0 or 1) in some contemporaneous populations—a feature that is apparently important in the second data set below. Other differences also follow from the different setting: the use here of a common parameter across loci, within populations; large numbers of SNP loci allow an assessment of model fit, whereas this is difficult and has not been done for forensic microsatellite loci; we incorporate ascertainment effects which can be important for SNP loci; and we have an explicit prior on ancestral population frequencies whereas the BN model estimates these empirically.

Recently, Wakeley *et al.* (2001) considered human SNP data with a slightly different structure from that considered here. They used a full likelihood approach, incorporating SNP ascertain-

ment and implemented by MCMC methods, in the context of a particular (generalized island) stochastic model for subdivided populations (e.g. Wakeley (2001) and references therein) both to estimate migration rates and to test for possible changes in human effective population sizes. There, as here, it would have been possible to assess whether their particular model fits the data, and an interpretation of their conclusions would be easier in the light of such model checking.

2. Modelling

We consider the setting in which we have a collection of P populations, from each of which we have data at each of L SNP loci. For the applications in which we are primarily interested, L is large (say 50 to many thousands) whereas P would range from 2 to 20. Write n_{ij} for the number of chromosomes typed at the ith SNP in the jth population. For each SNP we arbitrarily pick (and then fix) one of the two variants, and write x_{ij} for the number of copies of the chosen variant at locus i in the sample from population j. We shall denote by α_{ij}, $0 \leqslant \alpha_{ij} \leqslant 1$, the (unobserved) frequency in population j of the chosen variant at locus i.

Our primary focus is on relatively closely related populations. We assume that, at most or all SNPs analysed, the variation currently observed was present in the population ancestral to those under study. This is likely to be the case for SNPs which are present in a range of contemporaneous populations, since it is widely believed that SNPs are the result of a single chance mutational event, rather than recurrent mutation. Current SNP ascertainment methods, which preferentially select more polymorphic, and hence more likely older, SNPs, will also act to choose SNPs which were present in ancestral populations. The differences in allele frequencies between populations will be the result largely or exclusively of demographic events: the chance fluctuations, often called *genetic drift*, that are inherent in Mendelian segregation and the fact that individuals have different numbers of offspring, mitigated by the exchange of genetic material through migration and interbreeding. Ignoring loci under selection the probabilistic mechanisms affecting each SNP are thus the same, though these may differ across populations. Our methods should work well provided that most SNPs under study have not been affected by selection. With the availability of large numbers of SNPs, many outside the regions which have any known functional effect, it should be possible to achieve this. We thus ignore selection in the subsequent discussion, but we note in passing that, in the same spirit as Lewontin and Krakauer (1973), the fitting of a model like that developed here allows the possibility of detecting loci at which selection might be acting by looking for outliers.

Unless stated otherwise, throughout by 'frequency' we shall mean relative frequency. To avoid cumbersome terminology, we shall refer to the 'allele frequency' or allele count at a locus, rather than 'the frequency (or count) of the chosen variant' at that locus. The omission of subscripts will denote the entire collection of quantities, so for example x will refer to the collection $x_{ij}, i = 1, 2, \ldots, L, j = 1, 2, \ldots, P$. The arbitrary choice of variant means that relevant distributions, e.g. of allele frequencies, need to be symmetric. An alternative would be to count for example the less common of the two variants.

2.1. A hierarchical model

We now introduce a hierarchical model for the data. We give a population genetics motivation in the next subsection, but it is not unnatural on purely statistical grounds.

At the lowest level, we have binomial data: given n and α,

$$x_{ij} \sim \text{binomial}(n_{ij}, \alpha_{ij}). \tag{1}$$

We model the dependence structure among the population allele frequencies α as follows. First introduce another collection of unobserved quantities, one for each locus: π_i, $i = 1, 2, \ldots, L$. In the population genetics discussion below these will play the role of the allele frequencies in a population ancestral to those sampled. In addition, introduce parameters c_j, $j = 1, 2, \ldots, P$, one for each population. These will be the parameters that we aim to estimate. They specify how far (in the sense of variance) each population's allele frequencies tend to be from typical values. Formally, conditionally on π, c,

$$\alpha_{ij} \sim \text{normal}\{\pi_i, c_j \pi_i (1 - \pi_i)\}. \tag{2}$$

Here and throughout we shall sometimes specify continuous distributions for quantities which are restricted to [0, 1]. By this we mean the distribution whose density on (0, 1) is the relevant density, with atoms at 0 and 1 whose size is the total mass of the relevant distribution on $(-\infty, 0)$ and $(1, \infty)$ respectively.

To complete the hierarchy we put independent priors on π and c:

$$\pi_1, \ldots, \pi_L \quad \text{are independent and identically distributed with density } f; \tag{3}$$

$$c_1, \ldots, c_P \quad \text{are independent and identically distributed with density } g. \tag{4}$$

Where appropriate, we write a and b for hyperparameters in the distributions f and g respectively. Sometimes we might actually wish to learn about f from data, and this is in principle easily done by extending the hierarchy up to an additional level.

The general rationale for having the $\alpha_{\cdot j}$ (conditionally) independent and identically distributed and for having a separate parameter c_j for each population was described at the beginning of this section. That the variance should be parameterized as a multiple of $\pi_i(1 - \pi_i)$ is standard in this setting, and natural, for reasons that should become clearer in the next subsection, but in effect are concerned with an inherent binomial structure to genetic evolution.

Finally we need to model SNP ascertainment. We do so by conceptualizing a large collection of SNP loci, each of which might happen to be examined (with a common probability which does not depend on allele frequency at the locus) during the SNP discovery process. The exact ascertainment process may differ between studies, but typically a small sample of chromosomes is examined and the locus will be ascertained if they display (either some or enough, depending on the design of the study) variability.

For the common ascertainment strategy in which a sample of m_j chromosomes from population j is examined, and the SNP is ascertained if the sampled chromosomes exhibit any variation, we would thus have

$$P(\text{locus } i \text{ ascertained}|\alpha) \propto 1 - \prod_{j \in A} \alpha_{ij}^{m_j} - \prod_{j \in A} (1 - \alpha_{ij})^{m_j}, \tag{5}$$

where A is the collection of populations used in the ascertainment process.

It might be the case that some of the populations used during the ascertainment process are not among the P that are currently under study, for which SNP genotype data are available. (This is indeed the case in one of our examples.) This is naturally handled in our modelling framework simply by increasing P appropriately to include any such additional populations, but then treating the relevant $x_{\cdot j}$ as missing data. There is almost no information in the data from which to learn about the values of c_j for such populations. In practice then, we might wish to fix the relevant c_j on the basis of background knowledge about the populations involved, or at least to use a different prior for those c_js. We shall see from a sensitivity analysis of the example in question that in this case the exact way in which ascertainment is modelled, and so

in particular the way in which this issue is handled, does not greatly affect the posteriors for quantities of interest.

2.2. Population genetics motivation

We now describe one theoretical motivation for the model of Section 2.1. The setting is one in which the populations have each diverged from an ancestral population relatively recently. As is usual in population genetics, the relevant timescales are measured in units of the order of the (here, local) population size, so 'relatively recently' here may still refer to many hundreds or possibly thousands of generations. Write π_i for the allele frequency at the ith locus in the ancestral population.

Fix attention on a particular autosomal locus and on a particular descendant population. In a slight abuse of our notation, write π for the allele frequency at this locus in the ancestral population and α for its frequency in the descendant population. As noted above, we assume that $\pi \neq 0, 1$, or in other words that the SNP in question also exhibited variation in the ancestral population. Suppose that at the locus in question the descendant population is formed by first sampling N_0 chromosomes from the ancestral population, and that it then evolves in discrete generations with sizes $N_1, N_2, \ldots, N_t$ chromosomes, until the present. If the initial size N_0 is small relative to that of the ancestral population, the descendant population is said to have undergone a *bottleneck* at its formation. More generally, any sudden decline in population size is referred to as a bottleneck.

Under rather general assumptions about the demography of the population, the evolution of the allele frequency at the locus in question, from separation until the present, will be well approximated by that of the so-called neutral Wright–Fisher diffusion (with no mutation), with initial value π, run for a period τ satisfying

$$\tau = N_0^{-1} + \sigma^2 \sum_{k=1}^{t} N_k^{-1} \tag{6}$$

(e.g. Ewens (1979) or Ethier and Kurtz (1986)). Strictly this is a limiting result as the population sizes go to ∞, so the approximation will be more accurate if the N_k are large. In practice it works well unless some of the N_k are very small (Nordborg, 2001). The details of the demography enter only through the multiplicative constant σ^2 in equation (6).

The version of the Wright–Fisher diffusion in question has infinitesimal mean 0 and infinitesimal variance at z given by $z(1 - z)$. Now, provided that τ is small, the increment in the diffusion, from starting value π, over time τ, is approximately normally distributed, in this case with mean 0 and variance $\tau\pi(1 - \pi)$. Thus we have, approximately,

$$\alpha \sim \text{normal}\{\pi, \tau\pi(1 - \pi)\}, \tag{7}$$

which is consistent with the marginal distribution implied by equation (2). Both the natural Markov model and the diffusion approximation have absorbing boundaries corresponding to either all or none of the chromosomes in the population carrying the variant in question. Provided that the distributional statement (7) is interpreted as described after expression (2), then expression (7) remains a reasonable approximation. See Beaumont (2001) and references therein for an exact treatment.

In this setting, the parameter c_j for population j has a very natural interpretation as the time, on the diffusion timescale, for which the population has undergone genetic drift. In the special case in which the major effect is the bottleneck at founding, then c_j can be interpreted as the inverse of the size of the bottleneck (and expression (7) is effectively just the normal ap-

proximation to the binomial distribution for the number of copies of the allele of interest after the bottleneck). In somewhat more general settings, c_j^{-1} might be thought of as an effective size of bottleneck. Thus for example, for a population with Wright–Fisher demography which experiences a bottleneck of size N_B for t generations, distribution (2) would hold approximately with $c = t/N_B$.

Now consider the idealized setting of loci evolving independently (which will typically obtain unless they are very close on the same chromosome—say within a few centimorgans for humans), in populations which simultaneously diverge from an ancestral population and subsequently evolve independently. Then conditionally on the ancestral allele frequencies at each locus, and ignoring selection, the same model will apply, independently, to all autosomal loci within a population (with a simple correction to the variance structure for X- or Y-linked loci), with independence also across populations, leading to the full model (2).

We would thus expect the model to fit best in settings where divergence occurs because an ancestral population splits to fill new niches, which simultaneously become available. Violations might occur if the populations diverged from the ancestral population at different times—but only if the ancestral allele frequencies change over the periods between divergence, so this will be less serious for large ancestral populations in which allele frequencies change very slowly. A more likely concern would be migrations or admixture between populations, which induce correlations, a topic to which we return in Section 5.

In the setting of this subsection, the natural prior f for the π_i is just the distribution of SNP allele frequencies in the ancestral population. In practice this will of course not be known and will depend on the demographic history of the ancestral population. A natural starting-point would be the simple, so-called standard neutral, model for which it is known that, in a population of size N at a segregating site (one at which there is variability) with low mutation rate, the probability that the mutant variant is present in y copies, $y = 1, 2, \ldots, N-1$, is proportional to $1/y$ (Ewens, 1979). In practice we do not usually know which of the two current variants is the mutant, so the relevant distribution is the symmetrized version of this:

$$f(\pi) \propto \{\pi(1-\pi)\}^{-1} \qquad \pi \in [\varepsilon, 1-\varepsilon] \tag{8}$$

for small positive ε.

2.3. Relationship with F_{ST}

There is a commonly used measure in population genetics for quantifying the extent of differentiation among populations. It originated with Sewall Wright and is usually denoted by F_{ST}. Again we refer the reader to the reviews by Rousset (2001) and Excoffier (2001) for a full account; we provide only the briefest sketch. Our understanding has benefited from access to unpublished notes by David Balding.

Wright (1951) described F_{ST} as 'the correlation between random gametes, drawn from the same subpopulation, relative to the total'. Unfortunately this definition is not precise, and some of the subsequent confusion in the literature stems from different interpretations. One conceptual dichotomy between approaches arises from differences (usually implicit rather than explicit) in what is being conditioned on.

A common definition is

$$F_{ST} = \frac{Q_2 - Q_3}{1 - Q_3}, \tag{9}$$

where Q_2 and Q_3 are respectively the probability that two copies of the region on different

chromosomes sampled from within and between populations are the same. In what Balding has called the *descriptive* approach, often associated with Nei and colleagues, these probabilities are thought of as relating only to the sampling process by which the chromosomes are chosen. Equivalently, they are defined conditionally on the population allele frequencies. As a consequence, F_{ST} is a function of the population allele frequencies. In practice it is often 'evaluated' simply by replacing these by the obvious estimates from the sample data.

Other approaches (*model based* in Balding's terminology) interpret the probabilities in equation (9) as relating to repetitions of the entire evolutionary process, rather than simply over repetitions of sampling from the extant populations. In this case, F_{ST} would be thought of as a statistical parameter, and the goal is to estimate it from data, and/or to relate it (by probability calculations) to parameters which directly specify the evolutionary model. The most common estimation procedure (see for example Weir (1996)) is often formulated by analogy with analysis of variance. It is equivalent (Rousset, 2001) to a method-of-moments approach in which the probabilities Q_2 and Q_3 in equation (9) are estimated by the frequencies of identical pairs of chromosomes at the locus, within and between populations respectively, in the sample, and the estimates substituted into equation (9).

There is no guidance in Wright's definition on how to handle differing correlations for different alleles at multiallelic loci, and in practice these are often just averaged. (The issue is moot for SNP data.) As noted above, most existing approaches follow Wright in having one value of F_{ST} for the collection of populations (often with an implicit assumption of exchangeability across populations), though sometimes different values for distinct loci. We think it more appropriate to have population-specific parameters, common across SNP loci.

To see the relationship with our setting, note that for an SNP locus, writing α for the allele frequency in a population, a rearrangement of equation (9) gives

$$F_{ST} = \frac{\mathrm{var}(\alpha)}{E(\alpha)\{1 - E(\alpha)\}}. \tag{10}$$

In the descriptive framework the randomness relates simply to which population may happen to be chosen. In a model-based framework the expectations are over the evolutionary model. Some 'model-based' approaches, notably that of Cockerham and Weir (1987), only actually model the first two moments of allele frequencies. Note the standard multiplicative parameterization of $\mathrm{var}(\alpha)$ relative to $E(\alpha)\{1 - E(\alpha)\}$ to which we referred earlier.

Formula (10) bears a close similarity to the marginal variance structure implied by our model (2). Indeed, if we were to *insist* on a common value of c across populations, then equation (10) would obtain, with F_{ST} replaced by c, provided that it was interpreted as being conditional on π. In this sense, particularly as different approaches involve different conditioning anyway, our parameters c_j might be thought of as analogous to F_{ST}-values, but with one for each population.

3. Inference

In this section we discuss implementation of an MCMC method (here Gibbs sampling) to fit the model, and then compare the resulting Bayes estimates with those which would be obtained by methods that are analogous to those currently in use.

3.1. Implementation

We complete the specification of the model from Section 2.1, by taking the prior f on π to be

a symmetric beta(a, a) distribution. In our simulation studies and data analyses we considered the two cases $a = 0.1$ and $a = 1$. The former is a close approximation to the prior (8) suggested by the standard neutral model, whereas the latter is just the uniform distribution. Sensitivity analyses in our simulations and the two data analyses show that conclusions do not depend on which choice of a is made. We use a uniform prior g for the c_j. In simulation studies we have used a model for ascertainment that is analogous to that for the data in Section 4.1 and described there. We shall see that results are insensitive to the inclusion or exclusion of the ascertainment effect.

We implemented a Gibbs sampler to study the posterior distribution of quantities of interest (most notably of the c). To sample from the joint posterior, we successively update α, π and c by sampling from their current full conditional distribution.

Step 1: sample α^t from $p(\alpha \mid c^{t-1}, \pi^{t-1}, x)$.
Step 2: sample π^t from $p(\pi \mid c^{t-1}, \alpha^t, a)$.
Step 3: sample c^t from $p(c \mid \pi^t, \alpha^t)$.

Simple rejection sampling is adequately efficient to produce samples from each of the three full conditional distributions.

We assessed the convergence characteristics of the Gibbs sampler by simulating realizations of the Markov chain, starting from different points in the parameter space. Methods such as

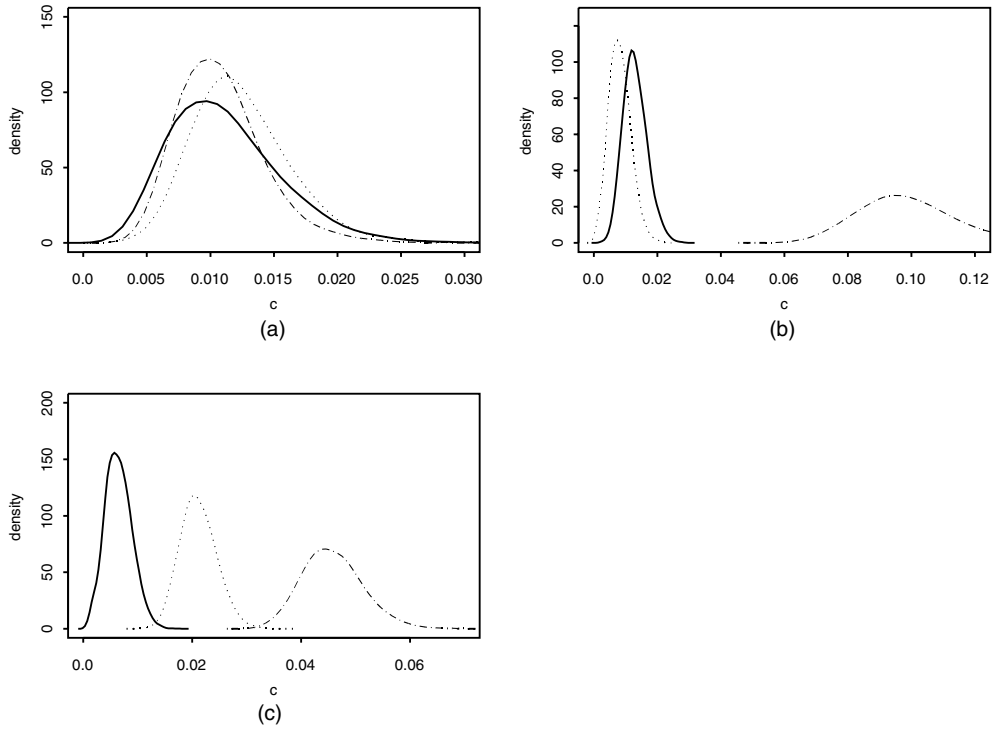

Fig. 1. Plots of the marginal posterior densities for c_1 (———), c_2 (·····) and c_3 (-·-·-·) for three simulated data sets: (a) simulated with parameter values of $c_1 = c_2 = c_3 = 0.01$ and $L = 50$; (b) simulated with parameter values of $c_1 = c_2 = 0.01, c_3 = 0.1$ and $L = 100$; (c) simulated with parameter values of $c_1 = 0.005$, $c_2 = 0.02$, $c_3 = 0.05$ and $L = 200$ (we used $P = 3$ and $a = 0.1$ throughout; the Gibbs sampler was run for 10 000 iterations subsequently to a burn-in period of 1000 iterations)

the variance ratio method of Gelman and Rubin (1992) also provided evidence that the Gibbs sampler achieved convergence well within the run length that is typically used (a burn-in period of 1000 iterations, followed by a run length of 10000 iterations). For certain data sets the chain can move only slowly through regions of parameter space when one of the c_j is very small (and $\pi_1 \ldots \pi_L$ close to $\alpha_{1j} \ldots \alpha_{Lj}$). Notwithstanding this, mixing was satisfactory for the run lengths that we used.

To give some sense of the amount of information in data sets of different sizes, Fig. 1 shows typical marginal posteriors for c for data simulated under the model under several different scenarios.

Although data are informative for the c_j, there is still sensitivity to the prior g, especially for relatively small values of L. We did not observe much dependence of posteriors on the choice of prior f for π, especially for data that were consistent with small values of the c_j. In this setting, especially with a reasonable number of populations, there is quite good information about plausible ranges for the relevant π_i from the observed range of values of x_{ij}/n_{ij}, and these ranges are small. Over the small ranges that were consistent with the data, our two priors (and most sensible priors) are thus effectively flat, and in particular very similar.

For the data sets that we examined, an incorporation of the ascertainment effect did not affect posteriors of interest. For intuition, consider ascertainment in a single population, in the most extreme case in which SNP discovery is based on examining only two chromosomes. For the locus to be ascertained, it must be the case that exactly one of the two chromosomes in the ascertainment sample carried the allele of interest. Suppose that the study data on that locus consist of x copies of the allele of interest in a sample of n chromosomes. Then, by an exchangeability argument, the full (study and ascertainment) data are equivalent to a sample of size $n+2$ of which $x+1$ carry the allele of interest. This changes estimates of the allele frequency towards 0.5, but the effect will be small unless either n is small or x/n is small. For ascertainment schemes in which $k > 2$ chromosomes are examined the effect will be weaker: think of the ascertainment process as augmenting the study sample with an additional k chromosomes, but only learning that of these not all are the same type. Allele frequencies across populations are correlated, so ascertainment in a different population is still informative for the population under consideration, but less so (to an extent which decreases with decreasing correlation) than for ascertainment directly in that population. Thus, provided that the study sample sizes are reasonable, or large, the ascertainment process will not greatly change posteriors on allele frequencies, except possibly for rare alleles.

3.2. Comparison with other estimators

Two other simple estimation methods are analogous to the heterozygosity-based, and method-of-moments (analysis-of-variance), approaches to estimating F_{ST} in widespread current use. For simplicity, the statistics are described as functions of α_{ij} (which in turn must be estimated by x_{ij}/n_{ij}). The first of two summary statistic methods for estimating the c_j involves an estimate of the heterozygosity. It follows from the prescribed model that

$$E\{\alpha_{ij}(1-\alpha_{ij})\} = E\{\pi(1-\pi)\}(1-c_j).$$

A natural ratio estimator of $(1-c_j)/(1-c_k)$ is therefore

$$\sum_{i=1}^{L} \alpha_{ij}(1-\alpha_{ij}) \Big/ \sum_{i=1}^{L} \alpha_{ik}(1-\alpha_{ik}). \tag{11}$$

Note that there is an identifiability issue. The use of this approach to estimate each of the cs separately is impossible without additional assumptions.

The method-of-moments estimator that we examine is defined as

$$\hat{c}_j = \frac{1}{L} \sum_{i=1}^{L} \frac{(\alpha_{ij} - \bar{\alpha}_i)^2}{\bar{\alpha}_i(1 - \bar{\alpha}_i)}, \tag{12}$$

where $\bar{\alpha}_i$ denotes the mean of the α_{ij} at locus i across populations. The intuition is that $\bar{\alpha}_i$ estimates π_i, and so the estimator defined at equation (12) aims to estimate the sample variance of the deviations around π_i. (More sophisticated versions might correct for finite sample effects.) We would expect, on general grounds, that the estimator would perform reasonably well unless the c_j differ substantially, and indeed we shall find this to be so below. Note that the average across populations of the $\hat{c}_j$ is a natural estimator of F_{ST}.

There are other possible approaches to estimation. We could treat the ancestral population allele frequencies π as nuisance parameters in a classical statistical framework, and for example consider maximum likelihood estimation, but the likelihood tends to ∞ at points where $c_j = 0$

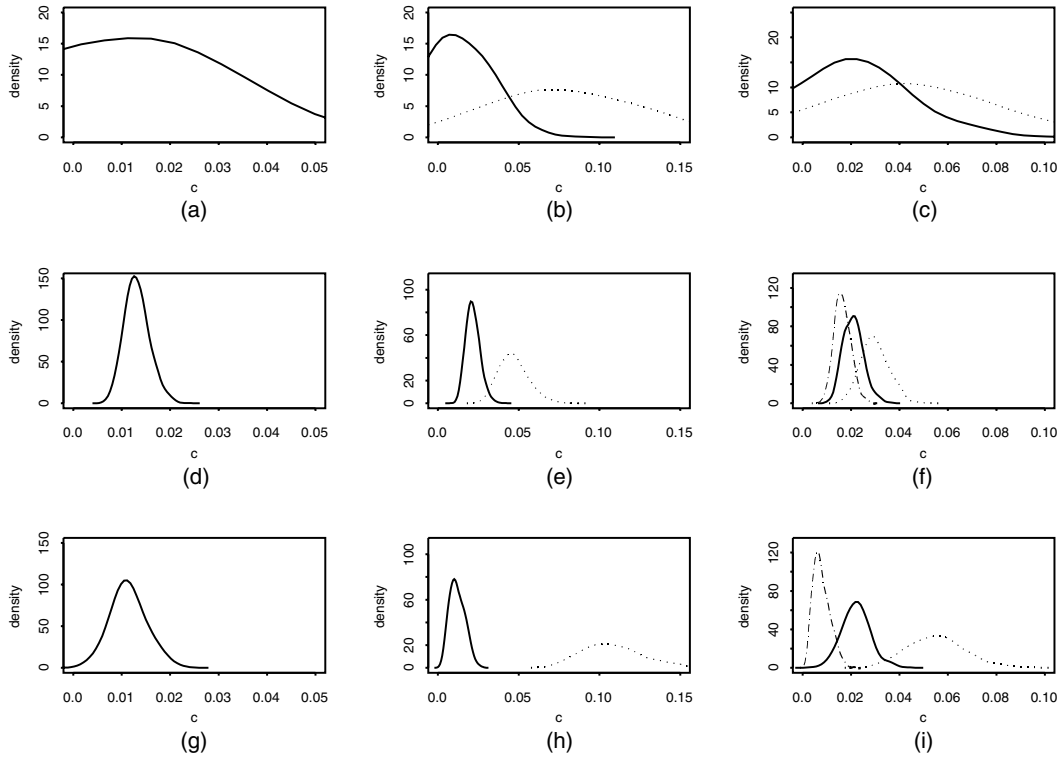

Fig. 2. Frequentist sampling properties of the heterozygosity-based, method-of-moments and Bayes estimators for c applied to data sets simulated under the model with three distinct sets of parameters (A, B and C) (parameter set A has $c_1 = c_2 = c_3 = 0.01$, set B has $c_1 = c_2 = 0.01$ and $c_3 = 0.1$ and set C has $c_1 = 0.005$, $c_2 = 0.02$ and $c_3 = 0.05$; in all simulations $P = 3$, $L = 50$ and $a = 1.0$): (a), (b), (c) marginal sampling distribution of the heterozygosity-based estimator for c_2 (———) and c_3 (· · · · · · ·) (with c_1 set to the true value) for sets A, B and C respectively; (d), (e), (f) marginal sampling distribution of the moments estimator for c_1 (- · - · - ·), c_2 (———) and c_3 (· · · · · · ·) for sets A, B and C respectively; (g), (h), (i) marginal sampling distribution of the Bayes estimator for c_1 (- · - · - ·), c_2 (———) and c_3 (· · · · · · ·) for sets A, B and C respectively (each density estimate is based on 500 simulated data sets)

and $\pi_i = \alpha_{ij}$, for some j. Alternatively, we might regard either the cs or the πs as random effects in a frequentist framework, but we have not pursued this.

Next, we compare the (frequentist) sampling properties of point estimates given by the three approaches, using the posterior mean as our Bayes point estimate. We simulated data sets under the specified model and examined the sampling distribution of the point estimates, for three different scenarios concerning the cs:

(a) all cs equal,
(b) all except one of the cs equal, with the other an order of magnitude larger, and
(c) the cs spread over an order of magnitude,

each with $P = 3$ and $P = 12$ populations. We simulated values for the population frequencies α, and then drew a binomial sample of size 100 at each locus from each population, using this as the data from which to estimate. Figs 2 and 3 give estimates of sampling distributions, with Table 1 summarizing these. In many of the scenarios that we consider, several of the cs take the same value. By symmetry the estimator of each of such c will have the same marginal sampling distribution, and we show only one version.

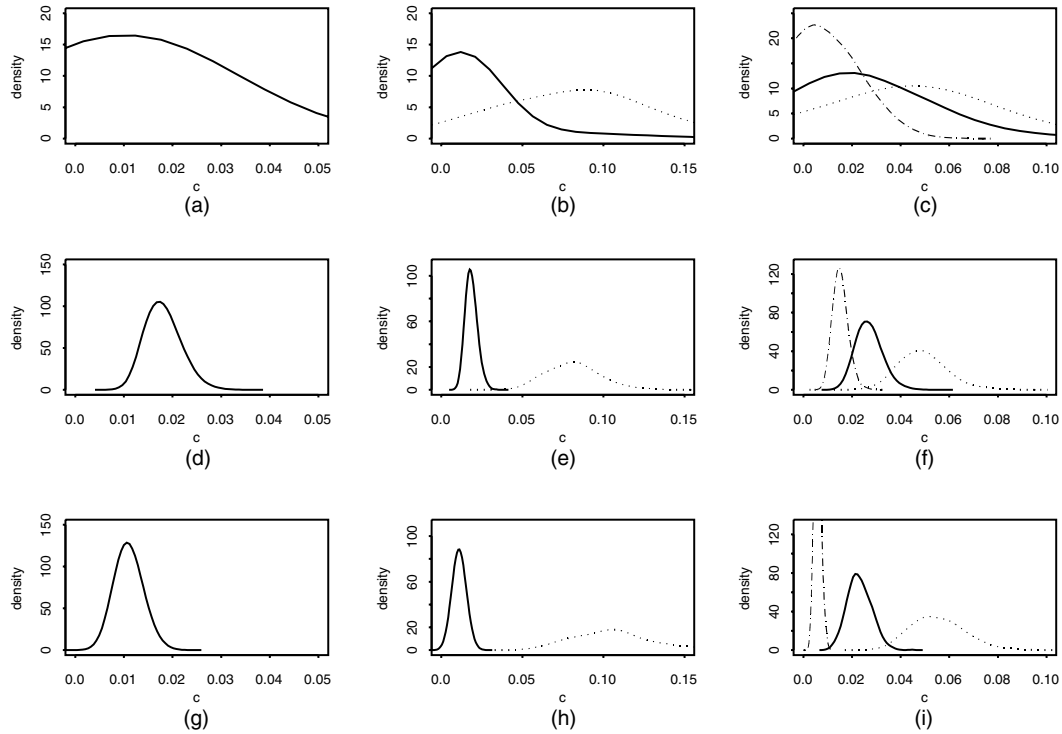

Fig. 3. Frequentist sampling properties of the heterozygosity-based, method-of-moments and Bayes estimators for c applied to data sets simulated under the model with three distinct sets of parameters (A, B and C) (parameter set A has $c_1 = \ldots = c_{12} = 0.01$, set B has $c_1 = \ldots = c_{11} = 0.01$ and $c_{12} = 0.1$ and set C has $c_1 = \ldots = c_4 = 0.005$, $c_5 = \ldots = c_8 = 0.02$ and $c_9 = \ldots = c_{12} = 0.05$; in all simulations $P = 12$, $L = 50$ and $a = 1.0$): (a), (b), (c) marginal sampling distribution of the heterozygosity-based estimator for sets A, B and C respectively; (d), (e), (f), marginal sampling distribution of the moments estimator for sets A, B and C respectively; (g), (h), (i) marginal sampling distribution of the Bayes estimator (c_2 (- · · · · ·), c_5 (———), c_{12} (· · · · · · ·)) for sets A, B and C respectively (to calculate the heterozygosity-based estimator in each case, we set c_1 to its actual value to avoid identifiability problems; each density estimate is based on 500 simulated data sets)

Table 1. Performance of each of the three estimators on simulated data†

P	*Parameter*	*Actual value*	*Results for the heterozygosity estimator*			*Results for the moments estimator*			*Results for the Bayes estimator*		
			Mean	*SE*	*RMSE*	*Mean*	*SE*	*RMSE*	*Mean*	*SE*	*RMSE*
Set A											
3	c_2	0.01	0.0087	0.0238	0.0238	0.0131	0.0026	0.0040	0.0114	0.0035	0.0037
12	c_2	0.01	0.0091	0.0235	0.0236	0.0180	0.0037	0.0088	0.0110	0.0024	0.0026
Set B											
3	c_2	0.01	0.0098	0.0236	0.0236	0.0217	0.0044	0.0124	0.0120	0.0046	0.0050
	c_3	0.1	0.0788	0.0503	0.0546	0.0468	0.0090	0.0540	0.1076	0.0180	0.0195
12	c_2	0.01	0.0099	0.0237	0.0237	0.0185	0.0037	0.0093	0.0110	0.0025	0.0027
	c_{12}	0.1	0.0801	0.0495	0.0533	0.0816	0.0161	0.0245	0.1049	0.0225	0.0229
Set C											
3	c_1	0.005	—	—	—	0.0165	0.0034	0.0119	0.0075	0.0035	0.0043
	c_2	0.02	0.0186	0.0256	0.0256	0.0209	0.0042	0.0043	0.0206	0.0057	0.0057
	c_3	0.05	0.0437	0.0359	0.0364	0.0299	0.0057	0.0209	0.0520	0.0105	0.0107
12	c_2	0.005	0.0052	0.0172	0.0172	0.0152	0.0031	0.0107	0.0056	0.0014	0.0015
	c_5	0.02	0.0182	0.0260	0.0260	0.0270	0.0055	0.0089	0.0219	0.0044	0.0048
	c_{12}	0.05	0.0432	0.0380	0.0386	0.0490	0.0097	0.0097	0.0529	0.0104	0.0108

†Means, standard errors SE and root-mean-squared errors RMSE are displayed for each of the three estimators for three distinct parameter sets and for two values of P (3 and 12). For parameter set A all cs are set to 0.01; for set B all cs except one are set to 0.01 and the final c is 0.1; for set C a third of all cs are set to 0.005, a third to 0.02 and a third to 0.05. For all simulations $L = 50$ and $a = 1.0$.

The general conclusions are that the heterozygosity-based estimate always performs relatively poorly compared with the other two (in addition to its identifiability problems). The Bayes estimator tended to outperform the other two estimators in all the analyses of simulated data that were conducted. As might be expected, the moment estimator performed poorly when there was substantial variation among the c_j, though this effect decreases as the number of populations increased: the bias arises because the unweighted mean $\bar{\alpha}_i$ will be a poor estimate of π_i when the cs differ substantially, though less so when there are more populations. For example, when only one of the cs, say c_k, is much larger, $\bar{\alpha}_i$ will tend to be closer to α_{ik}, and on average further from the other $\alpha_{i\cdot}$, than will π_i, so c_k will tend to be underestimated, and the other cs to be overestimated by the moment estimator.

Also of interest is the relationship between the amount of data available (L and P) and the performance of the Bayes estimator. Fig. 4 displays the frequentist sampling properties of the Bayes estimator (i.e. the posterior mean) on simulated data for various combinations of L and P. As we would expect, increasing either L or P leads to a lower sampling variance around the actual value of c. It is interesting that increasing L from 100 to 200 leads to a greater reduction in the variance than does increasing P from 6 to 12. The substantial increase in efficiency with increasing L is encouraging and suggests that with the imminent availability of huge amounts of SNP data much more precise estimation will be feasible in the near future.

4. Applications

4.1. Three European populations

We present and analyse a new SNP data set for three European populations, from France,

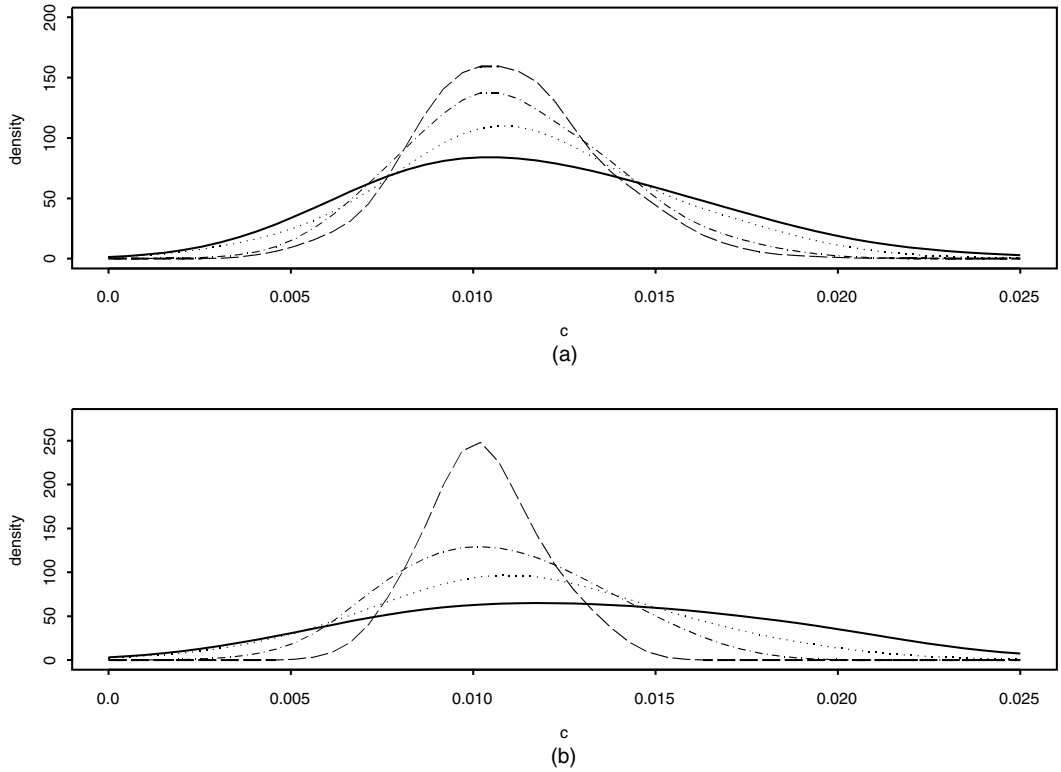

Fig. 4. Frequentist sampling properties of the Bayes estimator for c applied to data sets simulated under the model with varying values of L and P (in all simulations the value of c is set to 0.01 in all populations and $a = 1.0$): (a) marginal sampling distribution for c_1 with $L = 50$ for the cases when $P = 2$ (———), $P = 3$ (· · · · · · ·), $P = 6$ (- - - - - -) and $P = 12$ (— —); (b) marginal sampling distribution for c_1 with $P = 3$ for the cases when $L = 25$ (———), $L = 50$ (· · · · · · ·), $L = 100$ (- - - - - -) and $L = 200$ (— —) (each density estimate is based on 200 simulated data sets)

Utah and Iceland. The data consist of 83 SNP loci, with the number of chromosomes sampled from the three populations at each locus varying around 30, 60 and 70 respectively. The Icelandic chromosomes were sampled from unrelated Icelandic individuals. Those for the other two populations were taken from unrelated individuals in the collection of pedigrees from the Centre d'Etude du Polymorphism Humain, Paris, France. SNPs were selected from a set originally discovered by Wang *et al.* (1998) and genotyped using standard dye terminator sequencing (Applied Biosystems). The region surrounding the polymorphic base was sequenced in both directions, when possible. Resulting chromatographs were scored at the polymorphic position. Fig. 5 illustrates the data by showing the three two-way scatterplots of allele frequencies at each locus.

The ascertainment procedure for discovery of the SNPs used is known (Wang *et al.*, 1998). A pool of chromosomes, four from the Amish population, four from Utah and six from Venezuela, was examined at many locations in the genome, and a site was 'discovered' as an SNP if it exhibited any variability across the chromosomes in the pool. We thus use the ascertainment model (5) including the method for handling the use of chromosomes from a population other than those for which data are available. We choose and fix the value of c for Venezuela and the Amish at 0.02, but we note that the conclusions are not sensitive to whether or not an ascer-

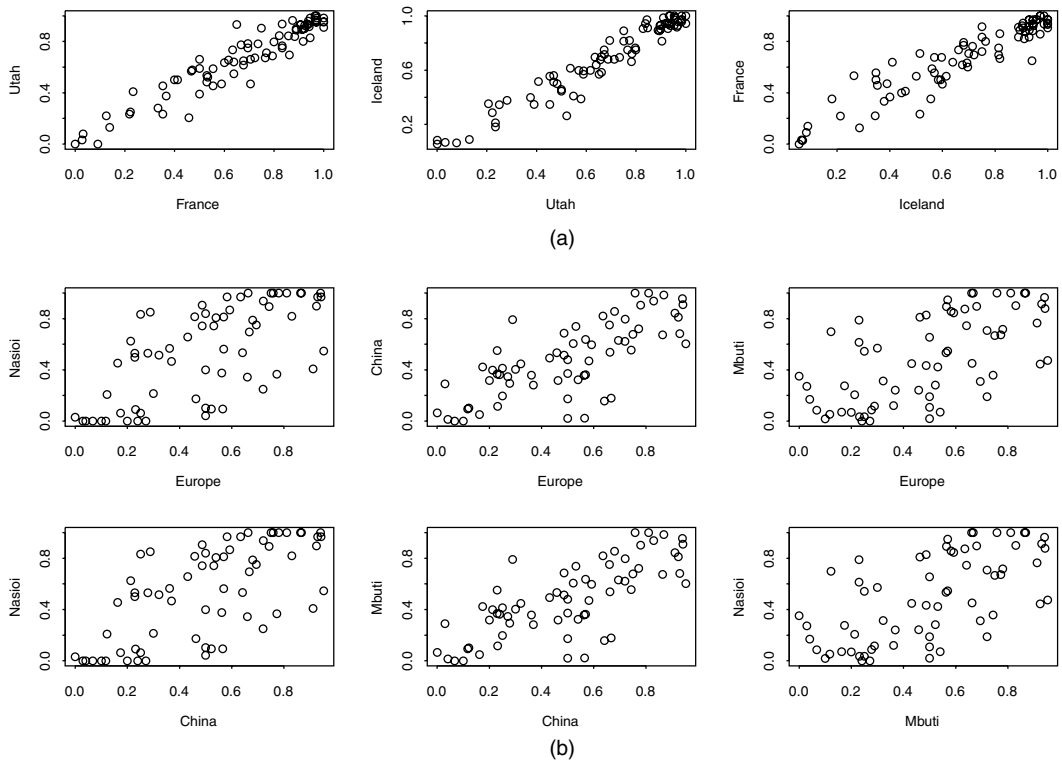

Fig. 5. Two-way scatterplots of allele frequencies at each locus for (a) the European data set and (b) the global data set

tainment correction is used, much less to the details of the way in which the use of Amish and Venezuelan chromosomes is handled.

The estimated heterozygosities for France, Utah and Iceland are 0.293, 0.289 and 0.281 respectively. The smaller value in Iceland hints at increased genetic divergence, but there does not seem to be a natural way to asses the extent of this, nor even to judge whether the observed differences in heterozygosities are 'significant'. For interest we calculate an estimate of F_{ST}. The model-based estimator of Weir (1996) gives 0.005.

Fig. 6 plots marginal posteriors for the cs for each population. Those for France and Utah are concentrated on similar values, whereas that for Iceland has considerable support at larger values. The Bayes estimates are 0.006, 0.003 and 0.013 for France, Utah and Iceland respectively. Our analysis thus suggests that Iceland has indeed diverged more genetically than the other two populations.

4.2. Global populations

The second data set that we consider is from 66 biallelic loci surveyed in five human populations, spread over four continents. The samples are from the Biaka and the Mbuti (two groups of pygmies), the Nasioi (a Melanesian population from Bouganville, off Papua), a mixed European and a mixed Chinese population. The data are originally from Bowcock *et al.* (1991a, b). A preliminary analysis (see the next subsection) suggested that there may be substantial correlation between allele frequencies for the two pygmy groups, in violation of the assumptions of the model. To avoid this difficulty, we consider only the four population groups Mbuti, Nasioi,

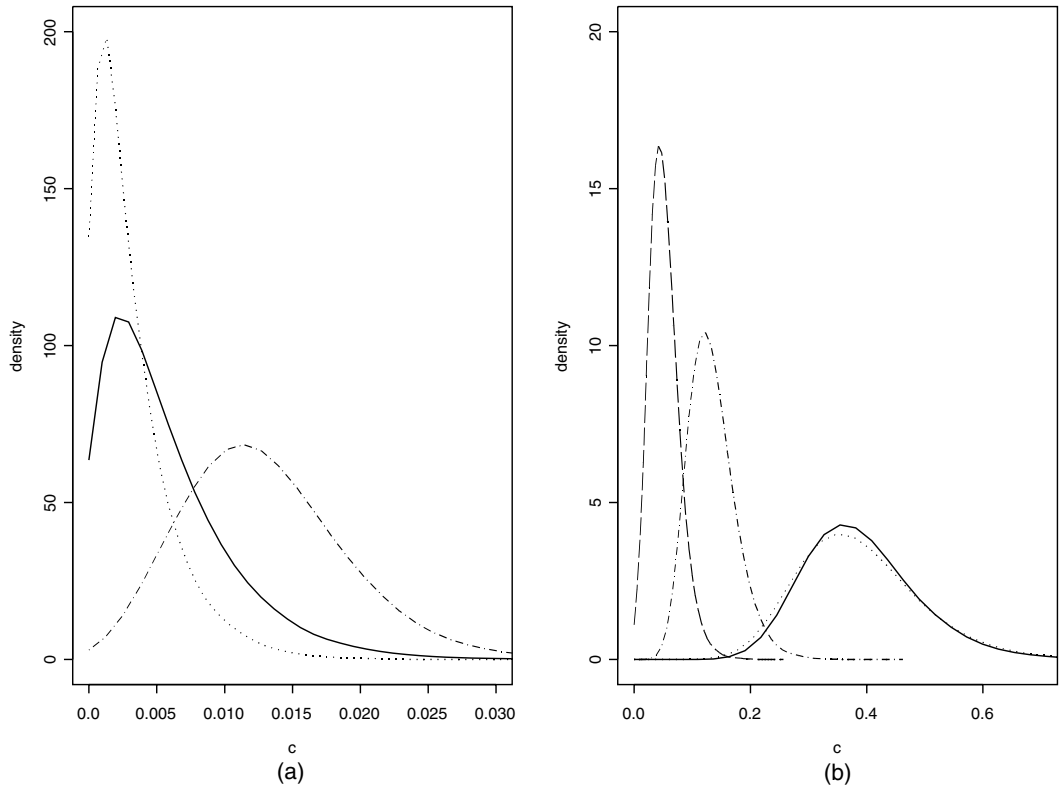

Fig. 6. Marginal posterior density plots obtained for c for the two data sets (a) European (———, France; · · · · · · · , Utah; - - - - - -, Iceland) and (b) global (———, Mbuti; · · · · · · · , Nasioi, - - - - - -, Chinese; — —, European): in both cases, $a = 1.0$ and the ascertainment procedure (outlined in Sections 4.1 and 4.2) is corrected for; the results are insensitive both to the value of a and to whether or not ascertainment is corrected for; the Gibbs sampler was run for 50 000 iterations subsequently to a burn-in of 5000 iterations

European and Chinese. As one illustration of the data, Fig. 5 shows the six possible pairwise scatterplots of allele frequencies.

The estimated population heterozygosities for the Mbuti, Nasioi, Chinese and Europeans are 0.28, 0.25, 0.34 and 0.36 respectively. An interpretation of the fact that heterozygosity is high in Europe is somewhat complicated by the ascertainment effect. The estimates of F_{ST} for this data set are 0.16 and 0.17 for the descriptive and model-based estimators respectively.

The loci were ascertained in a European sample, but the exact protocol is not known. We approximated it from model (5) with the set A corresponding to the European population only, and varying assumptions about its size. Here, the inclusion of ascertainment does not have an effect on inference within our model.

Fig. 6 shows marginal posteriors for the cs for each population. With ascertainment accounted for (as described above with $m = 10$), the posterior means are 0.39, 0.39, 0.13 and 0.05 for the Mbuti, Nasioi, Chinese and Europeans respectively. They show that the African and Melanesian populations have been relatively more differentiated than the European and Chinese. The major feature of this data set which we wish to emphasize is the fact that (as discussed in the next subsection) it still appears that our model is providing a reasonable fit, even though the population genetics assumptions on which it was motivated seem very unlikely to apply.

4.3. Assessing model fit

Here we describe and discuss four diagnostic tools which aim to assess differing aspects of model fit. If the assumption

$$\alpha_{ij} \sim N\{\pi_i, c_j\pi_i(1-\pi_i)\}$$

is reasonable, then the set of standardized residuals,

$$\left\{ \frac{x_{ij}/n_{ij} - \hat{\pi}_i}{[\{\hat{c}_j + (1-\hat{c}_j)/n_{ij}\}\hat{\pi}_i(1-\hat{\pi}_i)]^{1/2}};\ i = 1, \ldots, L,\ j = 1, \ldots, P \right\} \tag{13}$$

(where $\hat{\pi}_i$ and $\hat{c}_j$ denote the posterior mean of π_i and c_j respectively), ought to resemble a (correlated—see below) sample from a standard normal distribution.

A comparison of these residuals with a standard normal distribution allows an assessment of the distributional assumption, whereas plots of the residuals against fitted π-values allows an assessment of the variance structure. Fig. 7 shows Q–Q-plots of the standardized residuals, and plots against fitted πs, for the two data sets that we have just analysed. Each is encouraging. That the variance of the residuals is less than 1 is due to the inherent negative correlation between the P residuals for each locus. It is analogous to the need for a divisor of $n-1$ rather than n for unbiased estimation of the variance from independent and

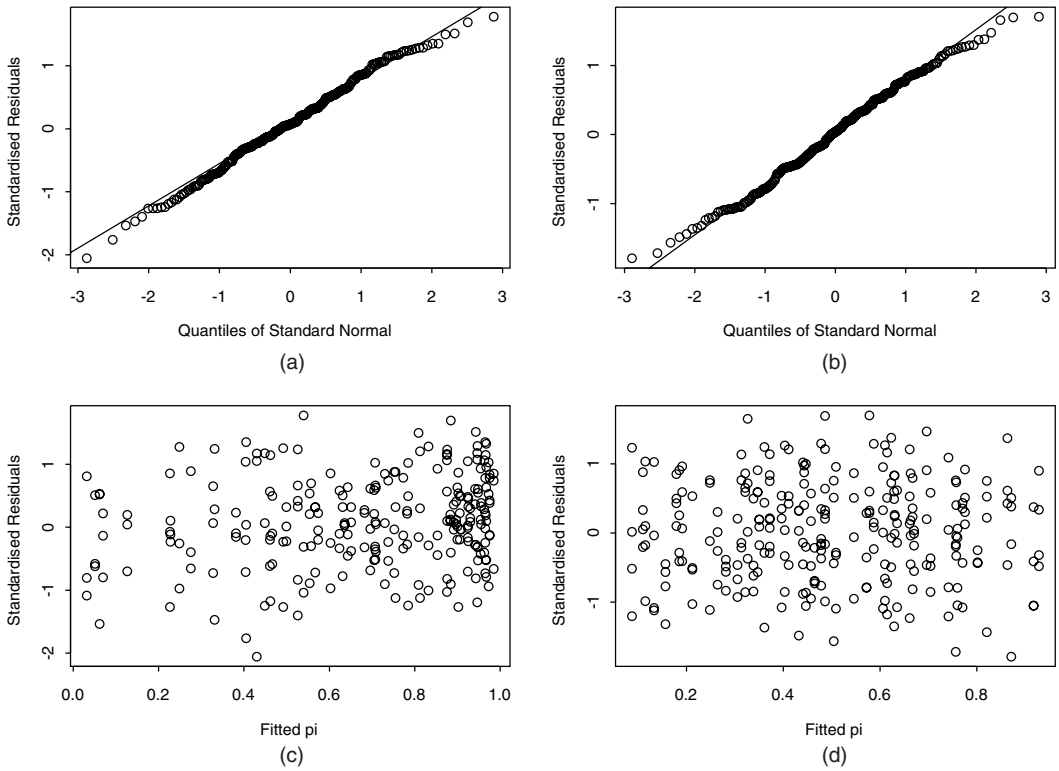

Fig. 7. Residual diagnostic plots for the two data sets: (a) Q–Q-plot of standardized residuals, European populations; (b) Q–Q-plot of standardized residuals, global populations; (c) scatterplot of standardized residuals at locus i against the mean of the marginal posterior for π_i (i.e. the fitted value for π_i), European populations; (d) scatterplot of standardized residuals, global populations

identically distributed observations with unknown mean, and we see the same behaviour in residuals from data simulated under the model. With $P = 3$ for the first data set, the estimated variance of the residuals is 0.550, whereas for $P = 4$ for the second data set it is 0.615.

One of the most likely deviations from the model for real data would be via correlations in allele frequencies (conditional on π) between some populations, induced either by gene flow or shared history. One tool for assessing this is to estimate the correlation structure of the residuals from fitting the model with no such correlation. (An alternative, to which we return in Section 5, is to fit a model which explicitly incorporates such correlations.) Simulations suggest that this has power to detect major departures provided that the number of populations is not too small (essentially because it relies on having decent estimates of the π). For example, consider five populations which follow the model except that for two populations the correlation in α-values (conditional on π) is 0.5. Estimates of correlations of residuals for this pair averaged 0.2 and were always positive and there was virtually no overlap between the distribution of these estimates and that of estimated correlations for other pairs of populations. In the world data set with all five populations, the estimated correlation of residuals between the two African populations was 0.72, which is striking evidence that the independence assumptions in the model do not apply. When the Biaka population was removed, all estimated correlations were within the range that is consistent with the model. For data from only three populations, simulations suggest that there is little power to detect correlations from residuals.

Another informal approach to assessing the fit of the model arises from the observation that ignoring all the data from one (or several) populations does not change the model for the remaining populations. Of course ignoring data in this way reduces the amount of information about each π_i, and hence the precision with which the c_j may be estimated, but, under the model, it should not systematically shift the location of the posteriors for the cs. In both our data sets, leaving out populations tends to widen the relevant posteriors, but in ways that are consistent with simulations under the model (the data are not shown), which is also encouraging. (But, for the world data set which included both African populations, this 'leave-one-out' diagnostic displayed very unstable estimates of the cs for several of the populations.)

5. Discussion

The imminent explosion of SNP data, initially for humans, and subsequently for other species, offers an unprecedented opportunity to increase our understanding of population structure and population histories. Whereas each locus carries limited information, it is the prospect of data from very large numbers of independent loci which offers the potential for powerful statistical inference.

We have developed a new model for such data. The model is motivated by population genetics considerations and is explicitly non-equilibrium. There is a single parameter for each population, closely related to Wright's F_{ST}, which effectively measures the extent of that population's differentiation. In the population genetics setting that we consider, the parameters have a direct interpretation as the amount of genetic drift to which the population has been subjected. We stress, though, that our approach will be useful *whenever* the underlying model captures important features of real data, whether or not the population genetics assumptions under which it was derived apply. Our setting allows an assessment of the fit of the model (itself an important novelty in this area) and there is encouraging evidence from the two data sets that we examined that the model does indeed fit under quite different scenarios. Although inferring population history is a common aim from this kind of data, there are other situations in which a good model for the joint distribution of allele frequencies across populations will be useful, e.g. in

assessing the importance in other populations of disease-predisposing mutations discovered in one population. Assessing population structure is also of central importance in the context of population-based association studies for common human diseases.

We fitted the model by using a fully Bayesian approach, implemented via MCMC sampling. This has several advantages, and in particular we have shown that, if point estimation of our F_{ST}-like parameters was of central interest, then the Bayesian approach can substantially outperform the method-of-moments and heterozygosity methods which are analogous to those that are widely used in current practice. The method also handles differing sample sizes across loci and/or populations easily (and appropriately). Our analysis of three European populations (the sort of setting in which we would expect our model to apply) shows that our new approach reveals features of the data which are not apparent from a comparison of simple summary statistics.

Our method allows for the ascertainment process by which SNPs are discovered. We argued in Section 3.1, and saw in practice, that provided that the sample sizes are reasonable this does not typically affect inference under our model. Note that this conclusion will not necessarily hold more widely. For some questions much of the information in data will be at loci with rare alleles, and it is for these that a failure to allow for ascertainment will have most dramatic effects.

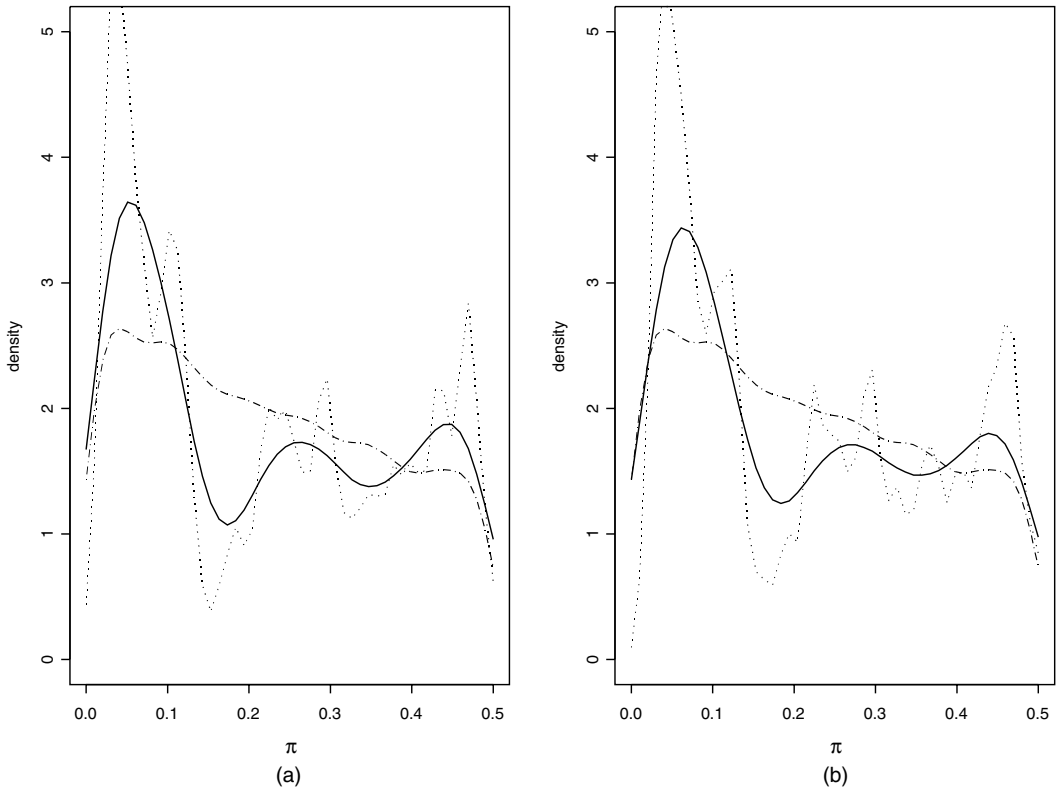

Fig. 8. Smoothed and unsmoothed density estimates for π for the European data set: (a) $a = 0.1$; (b) $a = 1.0$ (in each part we display a smoothed density estimate (———), an unsmoothed density estimate (· · · · · · ·) and an 'ascertained beta' density (- · - · - ·); the last density was obtained by sampling from a beta(0.1, 0.1) density and accepting or rejecting the value sampled according to the same ascertainment procedure that was used for the European data set (see Section 4.1))

We note for example that Wakeley *et al.* (2001) reported that the incorporation of ascertainment alters important conclusions.

There are some potentially interesting generalizations of our approach. It naturally handles missing data and could be extended to incorporate (and if necessary to estimate) correlations between loci, e.g. for studies which genotype several linked SNPs within each of a number of genomic regions. It also naturally provides a basis for the estimation of parameters describing historical population admixture.

One concern with our current approach is that it does not account for correlations across populations induced by shared history or by gene flow. For example, we cannot be certain whether increased differentiation in Iceland is due to its isolation and smaller effective population size, or to gene flow between the French and Utah populations, or between the 'source' ancestral population and either or both of France and Utah (although we would expect all to play some role). Nielsen and Wakeley (2001) developed a method for jointly estimating migration rates, times and effective population sizes since population divergence, for a pair of populations from which DNA sequences are available at a single locus. Beaumont (2001) has reviewed other methods for comparing migration with splitting models. Initial investigations show promise for the natural extension to our model in which the parameter is the variance–covariance matrix of the αs given π (deemed the same for each locus, after appropriate rescaling by $\pi_i(1 - \pi_i)$). This contrasts with some current approaches which would estimate pairwise distances between populations and then summarize these in a tree, or via multidimensional scaling. Using all data simultaneously for estimating the correlation structure has obvious advantages in principle over a separate analysis of each pair of populations.

A more ambitious extension would be to specify a tree which describes the temporal process by which populations diverged, and to use the tree in the obvious way to specify the dependence structure. Whether there is enough information in data to estimate the tree and the other additional parameters well remains to be seen. A similar procedure, for a slightly different likelihood, is available within the phylogeny estimation package PHYLIP.

Throughout we have focused on the estimation of the parameters c_j. In some settings it is of interest to estimate the distribution of allele frequencies in the ancestral population, e.g. to learn about the demographic history of that population. One way of doing so would be to extend the hierarchy to allow learning about the distribution f, or its hyperparameters. A simpler alternative in our framework would be to examine the empirical distribution of point estimates of each π_i. This is illustrated in Fig. 8 for the European populations. We note that the estimated empirical distribution of ancestral allele frequencies at ascertained loci is not greatly affected by the prior f. Further (see Fig. 8) it is similar to the curve which would be expected under the standard neutral model with this ascertainment strategy.

Acknowledgements

It is a pleasure to acknowledge helpful discussions with David Balding, Mark Beaumont, Agnar Helgasson, Richard Nichols and Peter McCullagh, and we thank Anna Di Rienzo for providing an electronic version of the second data set. PD was supported in part by UK Engineering and Physical Sciences Research Council grant GR/M14197 and Biotechnology and Biological Sciences Research Council grant 43/MMI09788, and by the Marvin and Mildred Conney faculty development fund to the Department of Human Genetics at the University of Chicago. GN was supported by an earmarked Engineering and Physical Sciences Research Council studentship.

References

Balding, D. and Nichols, R. (1995) A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, **96**, 3–12.

———(1997) Significant genetic correlations among Caucasians at forensic DNA loci. *Heredity*, **78**, 583–589.

Beaumont, M. (2001) Conservation genetics. In *Handbook of Statistical Genetics* (eds D. Balding, M. Bishop and C. Cannings), ch. 29, pp. 779–809. Chichester: Wiley.

Bowcock, A., Hebert, J., Mountain, J., Kidd, J., Rogers, J., Kidd, K. and Cavalli-Sforza, L. (1991a) Study of an additional 58 dna markers in five populations from four continents. *Gene Geogr.*, **5**, 151–173.

Bowcock, A., Kidd, J., Mountain, J., Hebert, J., Carotenuto, L., Kidd, K. and Cavalli-Sforza, L. (1991b) Drift, admixture, and selection in human evolution: a study with dna polymorphisms. *Proc. Natn. Acad. Sci. USA*, **88**, 839–843.

Cockerham, C. C. and Weir, B. S. (1987) Correlations, descent measures: drift with migration and mutation. *Proc. Natn. Acad. Sci. USA*, **84**, 8512–8514.

Ethier, S. and Kurtz, T. (1986) *Markov Processes—Characterisation and Convergence*. New York: Wiley.

Ewens, W. (1979) *Mathematical Population Genetics*. New York: Springer.

Excoffier, L. (2001) Analysis of population subdivision. In *Handbook of Statistical Genetics* (eds D. Balding, M. Bishop and C. Cannings), ch. 10, pp. 271–307. Chichester: Wiley.

Foreman, L. A., Smith, A. F. M. and Evett, I. W. (1997) Bayesian analysis of DNA profiling data in forensic identification applications (with discussion). *J. R. Statist. Soc.* A, **160**, 429–469.

Gelman, A. and Rubin, D. (1992) Inference from iterative simulation using multiple sequences. *Statist. Sci.*, **7**, 457–511.

Hudson, R. (1998) Island models and the coalescent process. *Molec. Ecol.*, **7**, 413–418.

Lewontin, R. and Krakauer, J. (1973) Distribution of gene frequencey as a test of the theory of selective neutrality of polymorphisms. *Genetics*, **74**, 175–195.

Nielsen, R. and Wakeley, J. (2001) Distinguishing migration from isolation: a markov chain monte carlo approach. *Genetics*, **158**, 885–896.

Nordborg, M. (2001) Coalescent theory. In *Handbook of Statistical Genetics* (eds D. Balding, M. Bishop and C. Cannings), ch. 7, pp. 179–212. Chichester: Wiley.

Roeder, K., Escobar, M., Kadane, J. and Balazs, I. (1998) Measuring heterogeneity in forensic databases using hierarchical Bayes models. *Biometrika*, **85**, 269–287.

Rousset, F. (2001) Inferences from spatial population genetics. In *Handbook of Statistical Genetics* (eds D. Balding, M. Bishop and C. Cannings), ch. 9, pp. 239–269. Chichester: Wiley.

Wakeley, J. (2001) The coalescent in an island model of population subdivision with variation among demes. *Theor. Popln Biol.*, **59**, 133–144.

Wakeley, J., Nielsen, R., Liu-Cordero, S. and Ardlie, K. (2001) The discovery of single-nucleotide polymorphisms—and inferences about human demographic history. *Am. J. Hum. Genet.*, **69**, 1332–1347.

Wang, D., Fan, J., Siao, C., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittman, M., Morris, M., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T. and Lander, E. (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, **280**, 1077–1082.

Weir, B. (1996) *Genetic Data Analysis II*. Sunderland: Sinauer.

Whitlock, M. and McCauley, D. (1999) Indirect measures of gene flow and migration: $f_{ST} \neq 1/(4nm + 1)$. *Heredity*, **82**, 117–125.

Wright, S. (1932) The roles of mutation, inbreeding, crossbreeding, and selection in evolution. In *Proc. 6th Int. Congr. Genetics*, *Menasha*, vol. 1, pp. 356–366.

———(1951) The genetic structure of populations. *Ann. Eugen.*, **15**, 323–354.

———(1982) The shifting balances theory and macroevolution. *A. Rev. Genet.*, **16**, 1–19.