Genet. Res., Camb. (2001), **78**, *pp.* 59–77. *With* 3 *figures.* © 2001 Cambridge University Press
DOI: 10.1017/S001667230100502x Printed in the United Kingdom

A Bayesian approach to the identification of panmictic populations and the assignment of individuals

KEVIN J. DAWSON[1,2]* AND KHALID BELKHIR[1]

[1] *Laboratoire Génome, Populations et Interactions, CNRS UMR 5000, Université de Montpellier II, Place Eugène Bataillon, 34095 Montpellier Cedex 5, France*

[2] *IACR, Long Ashton Research Station, Department of Agricultural Science, University of Bristol, Bristol BS41 9AF, UK*

(*Received 11 September 2000 and in revised form 22 December 2000*)

Summary

We present likelihood-based methods for assigning the individuals in a sample to source populations, on the basis of their genotypes at co-dominant marker loci. The source populations are assumed to be at Hardy–Weinberg and linkage equilibrium, but the allelic composition of these source populations and even the number of source populations represented in the sample are treated as uncertain. The parameter of interest is the partition of the set of sampled individuals, induced by the assignment of individuals to source populations. We present a maximum likelihood method, and then a more powerful Bayesian approach for estimating this sample partition. In general, it will not be feasible to evaluate the evidence supporting each possible partition of the sample. Furthermore, when the number of individuals in the sample is large, it may not even be feasible to evaluate the evidence supporting, individually, each of the most plausible partitions because there may be many individuals which are difficult to assign. To overcome these problems, we use low-dimensional marginals (the 'co-assignment probabilities') of the posterior distribution of the sample partition as measures of 'similarity', and then apply a hierarchical clustering algorithm to identify clusters of individuals whose assignment together is well supported by the posterior distribution. A binary tree provides a visual representation of how well the posterior distribution supports each cluster in the hierarchy. These methods are applicable to other problems where the parameter of interest is a partition of a set. Because the co-assignment probabilities are independent of the arbitrary labelling of source populations, we avoid the label-switching problem of previous Bayesian methods.

1. Introduction

The population genetic analysis of an outcrossing wild species often has to begin with an attempt to identify populations – 'evolutionary units' through which advantageous alleles can spread in response to selection. The problem is to identify complete or partial barriers to gene flow. These may be geographic barriers, or they may result from differences in habitat preferences within the same geographic range. Barriers to gene flow may also be maintained by assortative mating, or selection against hybrid genotypes (Barton, 1979; Barton & Hewitt, 1989). It is very difficult to deduce population structure by direct observation of migration or mating behaviour.

Identification of panmictic populations can provide a preliminary indication of the relevant evolutionary units. A population is said to be panmictic if mating is at random, in the sense that mating pairs are formed as if by choosing the male parent and the female parent at random from the population. Clearly, this is an idealization, which is at best a useful approximation to the behaviour of an outcrossing natural population. Under random mating, a large population reaches Hardy–Weinberg equilibrium in a single generation and approaches linkage equilibrium at a geometric rate. So, in the absence of recent immigration, it is reasonable to assume that unlinked, and loosely linked, loci will be close to linkage equilibrium. The genotypes of sampled individuals, at polymorphic co-

* Corresponding author. Tel: +44 (0)1275 549209. Fax: +44 (0)1275 394007. e-mail: Kevin.Dawson@bbsrc.ac.uk

dominant loci (such as microsatellites), can be used to identify population units which are close to Hardy–Weinberg and linkage equilibrium.

The inference problem is to assign each of the individuals in the sample to a panmictic source population. These source populations are not defined *a priori*. We will typically have some prior information about the number of source populations (for example, from the locations of known breeding grounds). Our prior information about the allele frequencies in these source populations may be more limited. The assignment of individuals to source populations induces a partition of the set of sampled individuals into non-empty disjoint subsets. This partition of the sample is the parameter of interest and, ideally, we would like to evaluate the evidence supporting each possible partition of the sample. In the Bayesian formulation of this inference problem, the evidence in favour of each possible sample partition is represented by the posterior distribution of this parameter.

Here, we present a Markov chain Monte Carlo method, based on the Metropolis–Hastings algorithm, for generating this posterior distribution. The output from the Markov chain is a large sample of these sample partitions, generated under the posterior distribution.

Pritchard *et al.* (2000) have already introduced a similar Bayesian formulation of the assignment problem. They generate the posterior distribution using a Markov chain Monte Carlo method based on Gibbs sampling. Our Bayesian analysis differs more fundamentally from that of Pritchard *et al.* (2000) in how we treat the output from the Markov chain. Pritchard *et al.* assign individuals to discrete clusters using the posterior probability that a particular individual is assigned to a population associated with some particular label. This makes sense only when we have reference samples, or some other information which characterizes specific source populations *a priori*, because in the absence of such information, the labels attributed to populations do not refer to any fixed entity. The posterior probability that a particular individual is assigned to a population associated with some particular label should be a constant which (by symmetry) is the same for all individuals and for all population labels. Indeed, Pritchard *et al.* freely admit that the only reason they can use such a procedure to assign individuals to discrete clusters is that their Markov chain fails to mix properly. According to Pritchard *et al.*, their Markov chain samples in the vicinity of a single mode of the posterior distribution. This should still give some idea of the uncertainty associated with the assignment of each individual. Other Bayesian, or partially Bayesian, formulations of the assignment problem have also appeared recently (Rannala & Mountain, 1997; Cornuet *et al.*, 1999).

The fact that the parameter of interest is a partition, presents particular difficulties. When the sample is large, it will not be feasible to evaluate the evidence supporting every possible partition of the sample. When many individuals are difficult to assign there will be many plausible partitions, each one having an individually low posterior probability. Given these problems, we believe that the most promising approach to the problem of making inferences about an unobserved partition of a large sample is a synthesis between Bayesian computations (based on explicit models) and clustering algorithms, which can identify clusters of individuals whose assignment together is well supported by the posterior distribution. We use low-dimensional marginals ('co-assignment probabilities') of the posterior distribution of the sample partition as measures of 'similarity', and then apply the 'furthest neighbour' (complete linkage) hierarchical clustering algorithm or its generalizations.

When the source populations are sufficiently distinct that individuals can be easily assigned, a point estimate of the sample partition is useful. The maximum likelihood estimation procedure, described below, provides a point estimate of the sample partition, for a chosen number of source populations.

2. The model

The point of departure for the present Bayesian analysis (like that of Pritchard *et al.*, 2000) is a reformulation of the underlying model. In the traditional formulation (Milner *et al.*, 1985; Smouse *et al.*, 1990), the sample is assumed to be from a mixed population, composed of unknown proportions, $\pi_1, \pi_2, \ldots$, from an unknown, or partially known, set of panmictic source populations, $1, 2, \ldots$, respectively. The proportions, $\pi_1, \pi_2, \ldots$, are treated as parameters of the model, which have to be inferred. Here, however, it is the assignment of individuals to source populations (and thus the partition which this induces on the set of sampled individuals) that is the parameter of interest. When the model is formulated in this way, the proportions, $\pi_1, \pi_2, \ldots$, no longer enter into the likelihood function.

A sample of n diploid individuals, labelled $1, 2, \ldots, n$, have been genotyped at a set of m marker loci. These n diploid genotypes constitute the data $\boldsymbol{X}$. An assignment of the individuals in the sample to source populations induces a partition of the set $S = \{1, 2, \ldots, n\}$ into non-empty disjoint subsets. Let $\tau = \{S_1, S_2, \ldots, S_\kappa\}$ denote the partition of S which corresponds to the true assignment of individuals to source populations, and let κ denote the number of subsets in the partition τ.

The distinct allele types represented in the sample from source population i at locus a are labelled $1, 2, \ldots, r_{i,a}$. Additional alleles $\ldots, r_{i,a}+1, r_{i,a}+2, \ldots$, may be present in the population. The set of distinct

https://doi.org/10.1017/S001667230100502X Published online by Cambridge University Press

diploid m-locus genotypes represented in the sample is denoted by $R(\boldsymbol{X}) = \{\boldsymbol{G}, \boldsymbol{G}', \boldsymbol{G}'', \ldots\}$.

In addition to the partition τ of S into κ parts, the only other parameters are the allele frequencies in each of the source populations at each of the marker loci. Let $p_{i,a}(s)$ denote the frequency in source population i of allele type s, at locus a. Let $\boldsymbol{p}_{i,a} = (p_{i,a}(1), \ldots, p_{i,a}(r_{i,a}), \ldots)$, $\boldsymbol{p}_i = (\boldsymbol{p}_{i,1}, \ldots, \boldsymbol{p}_{i,m})$ and $\boldsymbol{p} = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_\kappa)$.

Let n_i denote the number of individuals in the sample that are assigned to source population i. Let $d_{i,a}(s)$ denote the count (the number of copies) of the allele type s, at locus a among the individuals assigned to source population i. Let $\boldsymbol{d}_i = (\boldsymbol{d}_{i,1}, \ldots, \boldsymbol{d}_{i,m})$ and $\boldsymbol{d} = (\boldsymbol{d}_1, \ldots, \boldsymbol{d}_\kappa)$. Let $D_i(\boldsymbol{G})$ denote the count of the diploid m-locus genotype $\boldsymbol{G}$, among the individuals assigned to source population i. Let $\boldsymbol{D}_i = (D_i(\boldsymbol{G}), D_i(\boldsymbol{G}'), D_i(\boldsymbol{G}''), \ldots)$ and $\boldsymbol{D} = (\boldsymbol{D}_1, \ldots, \boldsymbol{D}_\kappa)$.

It is assumed that, at every marker locus, the allele compositions of the separate source populations are mutually independent. So, the likelihood function $P(\boldsymbol{X} \mid \tau, \boldsymbol{p})$ is of the form

$$P(\boldsymbol{X} \mid \tau, \boldsymbol{p}) = \prod_{i=1}^{\kappa} P(\boldsymbol{D}_i \mid \boldsymbol{p}_i). \tag{1}$$

The probabilities $P(\boldsymbol{D}_i \mid \boldsymbol{p}_i)$ can be factorized by a standard argument, to yield

$$\begin{aligned} P(\boldsymbol{D}_i \mid \boldsymbol{p}_i) &= P(\boldsymbol{D}_i \mid \boldsymbol{d}_i) P(\boldsymbol{d}_i \mid \boldsymbol{p}_i) \\ &= P(\boldsymbol{D}_i \mid \boldsymbol{d}_i) \left(\prod_{a=1}^{m} P(\boldsymbol{d}_{i,a} \mid \boldsymbol{p}_{i,a}) \right), \end{aligned} \tag{2}$$

where

$$P(\boldsymbol{d}_{i,a} \mid \boldsymbol{p}_{i,a}) = \frac{(2n_i)!}{\prod_{s=1}^{r_{i,a}} d_{i,a}(s)!} \prod_{s=1}^{r_{i,a}} p_{i,a}(s)^{d_{i,a}(s)} \tag{3}$$

is a multinomial distribution for sampling alleles at locus a, and

$$P(\boldsymbol{D}_i \mid \boldsymbol{d}_i) = 2^{h_i} \frac{n_i!}{\prod_{G \in R(X)} D_i(G)!} \left(\prod_{a=1}^{m} \left(\frac{\prod_{s=1}^{r_{i,a}} d_{i,a}(s)!}{(2n_i)!} \right) \right) \tag{4}$$

is the distribution of genotypes in a sample when alleles are permuted at random among individuals within the sample. Here h_i denotes the total number of heterozygous loci, among those individuals assigned to source population i.

3. Maximum likelihood estimation of the sample partition

Smouse *et al.* (1990) used maximum likelihood estimation to infer the proportions, $\pi_1, \pi_2, \ldots$, of different panmictic source populations, present in a mixed population, together with the allele frequencies in these source populations. Here, we use maximum likelihood estimation to infer the sample partition τ, together with the allele frequencies in the source populations. This is closely related to the approach of Belkhir & Bonhomme (2001).

For any given partition τ, the likelihood function $P(\boldsymbol{X} \mid \tau, \boldsymbol{p})$ is at a maximum when the allele frequencies, in each population, are equated to their observed values under the sample partition τ, here denoted by $\hat{\boldsymbol{p}}(\tau)$. So, given the number of source populations κ, the problem of finding the point $(\hat{\tau}_\kappa, \hat{\boldsymbol{p}})$ where the likelihood function $P(\boldsymbol{X} \mid \tau, \boldsymbol{p})$ is at its global maximum, reduces to the problem of finding the sample partition $\hat{\tau}_\kappa$, where $P(\boldsymbol{X} \mid \tau, \hat{\boldsymbol{p}}(\tau))$ is maximum. ($\hat{\tau}_1$ is the trivial sample partition where all individuals are assigned to a single population.) The maximum likelihood estimate $\hat{\tau}_\kappa$ can be found using a simulated annealing algorithm (Kirkpatrick *et al.*, 1983), in which the Metropolis–Hastings algorithm is used to simulate the Boltzmann distribution, at different 'temperatures' for a system whose 'state' is the sample partition τ, and whose 'energy' is proportional to the natural logarithm of the likelihood function $P(\boldsymbol{X} \mid \tau, \hat{\boldsymbol{p}}(\tau))$. To begin with, the Boltzmann distribution at a high temperature is simulated, then the temperature is lowered in a series of steps. This is referred to as the 'cooling schedule'. The Metropolis–Hastings algorithm used here is similar to that described in Section 5 and the Appendix, but has only two types of proposals (similar to the **exchange** and **transfer** proposals described in the Appendix). We found that a simple exponential cooling schedule performs well.

We now turn to the problem of choosing between the different maximum likelihood estimates $\hat{\tau}_1, \hat{\tau}_2, \hat{\tau}_3, \ldots$ Each time another source population is added to the model, the number of parameters increases. (We must specify the allele frequencies in the new source population.) The model having more parameters will never have a lower likelihood maximum than the model having fewer parameters. So, likelihood maximization is in conflict with 'Occam's razor'. Likelihood ratio tests are supposed to compensate for the fact that likelihood maximization intrinsically favours models with more parameters, because we accept a model with more parameters only if the increase in likelihood is 'unusually high'. (Unusually high, under what is often an extremely questionable null hypothesis.) Other procedures have been proposed for achieving the same result (see, for example, Burman & Nolan, 1995).

The relative simplicity or complexity of a hypothesis is at least to some extent a matter of subjective judgement. Can this be reduced to a matter of counting parameters or degrees of freedom? In the Bayesian approach our preference for simpler models can be incorporated into the prior probability dis-

https://doi.org/10.1017/S001667230100502X Published online by Cambridge University Press

tribution. The Bayesian methodology is a very flexible approach for extracting statistical inferences from the likelihood function.

One possible measure of the evidence supporting a particular maximum likelihood estimate, say $\hat{\tau}_2$, over another, say $\hat{\tau}_1$, is the posterior odds ratio $\pi(\hat{\tau}_2|\boldsymbol{X})/\pi(\hat{\tau}_1\,|\,\boldsymbol{X})$. If we want to reduce dependence on the prior, we may prefer to use the corresponding Bayes factor $(\pi(\hat{\tau}_2\,|\,\boldsymbol{X})/\pi(\hat{\tau}_1\,|\,\boldsymbol{X}))\ (\pi_T(\hat{\tau}_1)/\pi_T(\hat{\tau}_2))$. This is independent of the prior distribution $\pi_T(\tau)$, on the sample partition, but does depend on the prior distribution of the allele frequencies. However, it is only possible to obtain a reliable estimate of this Bayes factor when both $\hat{\tau}_1$ and $\hat{\tau}_2$ have sufficiently high posterior probabilities to be estimated by sampling the posterior distribution. Alternatively, we could infer κ from the posterior distribution, before seeking the corresponding maximum likelihood estimate $\hat{\tau}_\kappa$. In the next two sections, the Bayesian calculation for our model is described.

4. The choice of prior

The prior distribution on the partition τ of S is assumed to be of the form

$$\pi_T(\tau) = \pi_{T\,|\,K}(\tau\,|\,\kappa)\pi_K(\kappa). \tag{5}$$

The number of source populations κ, represented in the sample, can have any value from 1, up to a chosen maximum $\nu \leqslant n$. The prior distribution of the number of source populations κ is of the form

$$\pi_K(\kappa) = Au^\kappa, \tag{6}$$

for $\kappa = 1, 2, \ldots, \nu$. The parameter u can be chosen to lie anywhere in the interval $0 < u \leqslant 1$. For $u = 1$, the prior on κ is uniform. All possible partitions of the sample into a given number of parts κ, are assumed to have equal probability. Therefore

$$\pi_{T\,|\,K}(\tau\,|\,\kappa) = \frac{1}{S_n^{(\kappa)}}, \tag{7}$$

where $S_n^{(\kappa)}$ is a Stirling number of the second kind, which is equal to the number of distinct ways of partitioning a set of n distinct elements into κ non-empty disjoint (and unlabelled) subsets (Berge, 1971, pp. 40–41).

The likelihood function, $P(\boldsymbol{X}\,|\,\tau, \boldsymbol{p})$, is the same regardless of whether or not allele s at locus a in population i is really the same allele as allele s at locus a in some other population j. The allele labels carry no information whatsoever about the nature of these alleles. We further assume prior independence across loci, and across populations, so that

$$\pi_{\boldsymbol{P}}(\boldsymbol{p}) = \prod_{i=1}^{\kappa}\prod_{a=1}^{m}\pi_P(\boldsymbol{p}_{i,a}). \tag{8}$$

The search for an appropriate prior distribution for the composition of a population has provoked much controversy in the Bayesian literature. (See for example Walley, 1996, and the discussion which follows that paper.) A popular choice is the symmetric Dirichlet distribution. In its most general form, the Dirichlet distribution has r parameters $\alpha_1, \alpha_2, \ldots, \alpha_r$, and density function

$$\frac{\Gamma(\alpha_1+\alpha_2+\cdots+\alpha_r)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_r)}p_1^{\alpha_1-1}p_2^{\alpha_2-1}\ldots p_r^{\alpha_r-1} \tag{9}$$

over the simplex where $0 \leqslant p_s$ and $p_1+p_2+\cdots+p_r = 1$. If we take the limit where $\alpha_1+\alpha_2+\cdots+\alpha_r \to \theta$ as $r \to \infty$, while $\max\{\alpha_1, \alpha_2, \ldots, \alpha_r\} \to 0$, then the marginal distribution of the s highest allele frequencies always converges to a non-degenerate limit, and the joint distribution of ordered allele frequencies is the Poisson–Dirichlet distribution with parameter θ (see Kingman, 1975; Watterson, 1976; Kingman, 1980, pp. 40–42.)

We have chosen the prior distribution of the allele frequencies in each population i, at each locus a, to be a Poisson–Dirichlet distribution, with parameter $\theta_{i,a}$. The values of the parameters $\theta_{i,a}$ may be fixed; or alternatively, they may have a prior distribution $\pi_\Theta(\theta_{i,a})$, in which case

$$\pi_P(\boldsymbol{p}_{i,a}) = \int_{\theta_{i,a}}\pi_{P\,|\,\Theta}(\boldsymbol{p}_{i,a}\,|\,\theta)\pi_\Theta(\theta_{i,a})\,d\theta_{i,a}. \tag{10}$$

Let $\theta_i = (\theta_{i,1}, \ldots, \theta_{i,m})$ and $\theta = (\theta_1, \ldots, \theta_\kappa)$. We can now write the joint posterior distribution in the form

$$\pi(\tau, \boldsymbol{p}, \theta\,|\,\boldsymbol{X}) = C^{-1}P(\boldsymbol{X}\,|\,\tau, \theta) \times\left(\prod_{i=1}^{\kappa}\prod_{a=1}^{m}\pi_{P\,|\,\Theta}(\boldsymbol{p}_{i,a}\,|\,\theta)\pi_\Theta(\theta_{i,a})\right)\pi_T(\tau). \tag{11}$$

The Poisson–Dirichlet prior has the convenient consequence that the allele frequencies can be integrated out analytically (see Appendix for details), leaving a posterior distribution of the form

$$\pi(\tau, \theta\,|\,\boldsymbol{X}) = C^{-1}P(\boldsymbol{X}\,|\,\tau, \theta)\pi(\tau, \theta), \tag{12}$$

where $P(\boldsymbol{X}\,|\,\tau, \theta)$ is the likelihood function for the model parameterized by θ (see equation A4 of the Appendix); and the joint prior is of the form

$$\pi(\tau, \theta) = \left(\prod_{i=1}^{\kappa}\prod_{a=1}^{m}\pi_\Theta(\theta_{i,a})\right)\pi_T(\tau), \tag{13}$$

where $\pi_\Theta(\theta_{i,a})$ may be degenerate.

5. The Markov chain Monte Carlo computation

In the previous section, we derived an explicit expression (12) for the posterior distribution $\pi(\tau, \theta\,|\,\boldsymbol{X})$, up to an unknown normalizing constant. However, what we are really interested in is the

https://doi.org/10.1017/S001667230100502X Published online by Cambridge University Press

marginal posterior distribution of τ. Formally, this marginal is given by integrating over θ in (12). Even if we choose to fix the value of the parameter θ (the prior $\pi_\Theta(\theta_{i,a})$ is degenerate), so that the integration over the parameter θ does not need to be performed, we still need a method for collapsing the posterior distribution $\pi(\tau \mid \boldsymbol{X})$, down to marginals of much lower dimension.

Rather than trying to compute the posterior distribution (by first computing the likelihood function), we can generate a large sample of observations from the posterior distribution. This is a very convenient form in which to store information about the joint distribution of many random variables, because we can easily extract any low-dimensional marginal of this joint distribution, simply by taking each observation in the sequence, and discarding everything except the component which is of interest. The resulting sequence of low-dimensional observations can then be represented graphically (for example, as a low-dimensional histogram, or other density estimate).

The Metropolis–Hastings algorithm (Hastings, 1970; see also Besag *et al.*, 1995; Gilks *et al.*, 1996) can be used to generate a sample of observations of any random variable whose probability distribution is known up to a normalizing constant. This algorithm simulates a time-reversible Markov chain whose equilibrium distribution is the required 'target' distribution. The transition process of this Markov process consists of two steps: a proposal step, and an acceptance/rejection step.

If we propose a change in some, or all, of the variables from (τ, θ) to new values (τ', θ'), with proposal probability $q((\tau, \theta) \rightarrow (\tau', \theta'))$, then, at the acceptance/rejection step, we have to compute the ratio

$$\begin{aligned} R((\tau, \theta) &\rightarrow (\tau', \theta')) \\ &= \frac{\pi(\tau', \theta' \mid \boldsymbol{X}) q((\tau', \theta') \rightarrow (\tau, \theta))}{\pi(\tau, \theta \mid \boldsymbol{X}) q((\tau, \theta) \rightarrow (\tau', \theta'))} \\ &= \frac{P(\boldsymbol{X} \mid \tau', \theta') \pi(\tau', \theta') q((\tau', \theta') \rightarrow (\tau, \theta))}{P(\boldsymbol{X} \mid \tau, \theta) \pi(\tau, \theta) q((\tau, \theta) \rightarrow (\tau', \theta'))}. \quad (14) \end{aligned}$$

We draw a random variable, Z, from a uniform distribution on the interval [0, 1]. If Z is less than the ratio $R((\tau, \theta) \rightarrow (\tau', \theta'))$, then we accept the proposal, and the state of the Markov chain is updated to (τ', θ'). Otherwise, the state of the Markov chain remains as (τ, θ).

We are free to choose any proposal process, provided it is compatible with the requirement that the resulting Markov chain is 'irreducible' (every state can be reached from every other state, in a finite number of steps). If we use a number of separate, carefully chosen, proposal processes, each of which is more restricted in the changes that it can make, then 'massive cancellations' may occur in this probability ratio, which can speed up the calculation. We have made use of two highly constrained proposal processes, **exchange** and **transfer**, which allow massive cancellations in the probability ratio, and thus speed up the Bayesian computation. An additional proposal process, **change** κ, can change the number of subsets into which the sample is partitioned, either by splitting an existing subset in two, or by fusing two subsets together. For details of these proposal processes, see the Appendix.

6. Processing the output from the Bayesian computation

The immediate output from a run of the Markov chain is a sequence of partitions, $\tau_1, \tau_2, \ldots$, which should resemble a random sample from the marginal posterior distribution $\pi_T(\tau \mid \boldsymbol{X})$. We can easily collapse the posterior distribution $\pi_T(\tau \mid \boldsymbol{X})$, to obtain the marginal posterior distribution $\pi_K(\kappa \mid \boldsymbol{X})$ of the number of source populations, κ, represented in the sample.

Pritchard *et al.* (2000) noted that the presence of hybrid individuals in the sample could lead to the identification of spurious 'source' populations. In such cases, the posterior distribution of κ will be misleading. In any case, the real parameter of interest is the partition of the sample induced by the assignment of individuals to source populations. When faced with complex patterns of hybridization, perhaps the best that can be done is to identify 'clusters' of individuals, whose assignment together is well supported by the posterior distribution. In such cases, the number of clusters, or populations, which are well supported by the posterior distribution, is of much more interest than the posterior distribution of the parameter κ.

Even in the absence of hybridization, the assignment problem is difficult simply because the parameter of interest is a partition. The number of distinct ways of partitioning a set of n distinct elements into κ non-empty disjoint (and unlabelled) subsets, is given by the Stirling number of the second kind $S_n^{(\kappa)}$. These numbers grow very rapidly with n. If there are many individuals that are difficult to assign, then there will be many plausible partitions, each one having an individually low posterior probability. However, even in such situations, it should be possible to identify 'clusters' of individuals, whose assignment together is well supported by the posterior distribution. We use a 'hierarchical agglomerative clustering' approach, because this corresponds to scepticism towards lumping clusters together. In contrast, 'hierarchical divisive clustering' approaches (recommended by Guénoche *et al.*, 1991) would correspond to scepticism towards splitting clusters apart.

Let $\pi(U \mid \boldsymbol{X})$ denote the posterior probability that the subset of individuals $U \subset S$, all belong to the same

https://doi.org/10.1017/S001667230100502X Published online by Cambridge University Press

source population, and that these are the only individuals in the sample which belong to that particular source population. That is

$$\pi(U \mid \boldsymbol{X}) = \sum_{\tau : U \in \tau} \pi_T(\tau \mid \boldsymbol{X}). \tag{15}$$

We also define the corresponding *cumulative* probability, as follows. Let $\Pi(U \mid \boldsymbol{X})$ denote the posterior probability that the subset of individuals $U \subset S$ all belong to the same source population. That is

$$\Pi(U \mid \boldsymbol{X}) = \sum_{V : U \subset V \subset S} \pi(V \mid \boldsymbol{X}). \tag{16}$$

A descriptive name for this probability would be the *probability of co-assignment* for the individuals belonging to the subset $U \subset S$. These co-assignment probabilities are exactly what we need for our hierarchical clustering approach. We can also use the corresponding conditional co-assignment probabilities $\Pi(U \mid \kappa = k, \boldsymbol{X})$, and $\Pi(U \mid \kappa \geqslant k, \boldsymbol{X})$.

Estimates of the co-assignment probabilities $\Pi(U \mid \boldsymbol{X})$, and the corresponding conditional co-assignment probabilities $\Pi(U \mid \kappa = k, \boldsymbol{X})$, and $\Pi(U \mid \kappa \geqslant k, \boldsymbol{X})$, are provided by the relative frequency of the corresponding events in the output of the Metropolis–Hastings Markov chain.

Let $\Pi_d(\boldsymbol{X})$ denote the array of co-assignment probabilities $\Pi(U \mid \boldsymbol{X})$, for all subsets $U \subset S$, of size $|U| = d$. Let $\Pi_d(\kappa = k, \boldsymbol{X})$ and $\Pi_d(\kappa \geqslant k, \boldsymbol{X})$ denote the corresponding arrays of conditional co-assignment probabilities, $\Pi(U \mid \kappa = k, \boldsymbol{X})$ and $\Pi(U \mid \kappa \geqslant k, \boldsymbol{X})$. The clustering algorithm can be applied to any of the arrays $\Pi_d(\boldsymbol{X})$, $\Pi_d(\kappa = k, \boldsymbol{X})$, $\Pi_d(\kappa \geqslant k, \boldsymbol{X})$, for $d = 2, 3, 4$.

Next, we construct a binary tree. Every individual in the sample is associated with a *terminal node*. Every internal node in a binary tree has two descendant nodes. A descendent may be either a terminal node or an internal node. If there are n terminal nodes, then there will be $n-1$ internal nodes when the tree is completed. The internal nodes are labelled $t = 1, 2, \ldots, n-1$, where t is the generation at which the node was created. At any generation t, a node is said to be *open* if it has not yet become the descendent of another node, and is said to be *closed* if it has become a descendent. Every internal node defines a set of terminal nodes (the terminal nodes that can be reached by descending the tree, starting from that node), and hence a set, or 'cluster', of individuals $C(t)$.

Each internal node t is associated with a *probability level*, p_t, which tells us about the 'worst' aspect of the cluster of individuals which is defined by that node. This probability level is also the height of the node. (Terminal nodes are defined to have unit height.) We could base the probability level p_t on any of the low-dimensional marginals: $\Pi_2(\boldsymbol{X})$, $\Pi_3(\boldsymbol{X})$ or $\Pi_4(\boldsymbol{X})$. To construct a binary tree, and the associated hierarchy of clusters, using the array of co-assignment probabilities $\Pi_d(\boldsymbol{X})$, of dimension $d \geqslant 2$, we must make use of all the arrays $\Pi_2(\boldsymbol{X})$, $\Pi_3(\boldsymbol{X})$,…, $\Pi_d(\boldsymbol{X})$, up to the chosen dimension d.

The algorithm for constructing the binary tree is as follows. In the first *generation*, we identify the pair of individuals (I_0, I_1) for which the probability $\pi_2(I_0, I_1 \mid \boldsymbol{X})$ is maximum, and thus construct the first *internal* node. At every subsequent generation t, we take every possible pair of open nodes, and we propose joining the pair to form a new internal node. For each of these proposed nodes, we calculate the probability level. Whichever proposed node has the maximum probability level is then accepted and added to the tree.

The probability level, p_t, associated with an internal node, and cluster C, is defined to be the lowest posterior co-assignment probability $\Pi(C' \mid \boldsymbol{X})$, of any subset $C' \subset C$, of size $|C'| = d$, which can be formed from the individuals belonging to cluster C. It can be computed as follows. If $|C| \leqslant d$, then the probability level associated with the proposed internal node is defined to be $\Pi(C \mid \boldsymbol{X})$. Otherwise, when $d < |C|$, we must enumerate every possible subset of C, which contains exactly d individuals. We then identify the subset $C' \subset C$, $|C'| = d$, for which $\Pi(C' \mid \boldsymbol{X})$ has the lowest value. The probability level associated with the proposed internal node is then equated to $\Pi(C' \mid \boldsymbol{X})$. Notice that it is only necessary to consider subsets $C' \subset C$, $|C'| = d$, which are not subsets of the clusters defined by either descendant node.

So, in this algorithm, clusters are constructed so as to maximize the minimum co-assignment probability (the 'similarity' measure) within clusters. An advantage of this max–min, or furthest neighbour, method in this Bayesian context is that each node (and the cluster which it defines) is associated with a Bayesian measure of the 'worst' aspect of the aggregation (assignment) of individuals which is under that node. Thus, a cluster is recognized on the basis of its internal cohesiveness, or 'homogeneity'. In the case $d = 2$, our clustering algorithm reduces to the classical 'furthest neighbour' (or complete linkage) algorithm (Defays, 1977). At the other extreme is the 'nearest neighbour' (or single linkage) method of hierarchical clustering (for a historical survey, see Graham & Hell, 1985; for some more recent advances see Olson, 1995), in which a cluster is recognized on the basis of its separation or isolation from all external individuals. The resulting clusters may be distinctly non-homogeneous. Other clustering methods based on centroids or average distances, introduce distance measures which can have no direct Bayesian interpretation.

We would expect higher values of d to result in systematically lower probability levels. Nevertheless, using an array of higher dimension should have the

https://doi.org/10.1017/S001667230100502X Published online by Cambridge University Press

Table 1. *The posterior distribution of the number of source populations κ, and Bayes factor B_1. In each case the prior has parameter $\nu = 4$, while u and θ have the values indicated in the table*

Data			κ				
set	u	θ	1	2	3	4	B_1
1	1	1	0·3492	0·6045	0·0427	0·0036	1·61
2	1	1	0·9759	0·0179	0·0059	0·0003	121·44
3	1	1	0·0001	0·0039	0·0544	0·9416	0·0003
3	0·75	1	0·0003	0·0026	0·0523	0·9448	0·0005
3	1	10	0·0002	0·0183	0·2632	0·7183	0·0007
4	1	1	0·0016	0·0037	0·0303	0·9644	0·005
5	1	1	0·0059	0·0039	0·0299	0·9603	0·02
5	0·75	1	0·0140	0·0021	0·0301	0·9538	0·02
5	1	10	0·3676	0·3836	0·0824	0·1663	1·74

advantage of being more sensitive to the creation of any poorly supported clusters. Ideally, we would like to set the dimension at its maximum value of $d = |S|$. This would ensure that the probability level associated with each cluster C is the cumulative probability $\Pi(C|\boldsymbol{X})$, given by (16).

Once the tree has been constructed, we can easily identify subsets of individuals whose assignment together is well supported by the marginal posterior distribution $\pi_T(\tau|\boldsymbol{X})$. Each node defines a set of individuals. The set of internal nodes which are open at generation t (including the node which is created at generation t) defines a set $\lambda_t(\boldsymbol{X})$ of clusters of individuals. $\lambda_t(\boldsymbol{X})$ is a partition of a set of individuals $\Lambda_t(\boldsymbol{X}) \subset S$. We refer to the set of individuals $\Lambda_t(\boldsymbol{X})$ as a 'core' of the sample, and the partition $\lambda_t(\boldsymbol{X})$ of $\Lambda_t(\boldsymbol{X})$ as a 'core partition'.

It is important to recognize that the probability level p_t associated with the core partition $\lambda_t(\boldsymbol{X})$ has no frequentist interpretation. It is a posterior probability, which depends both on the observed data and on the choice of prior. So it is a subjective probability. This raises an important question: How low should we allow p_t to fall, before we stop accepting further agglomeration? More experience with the method will be needed to resolve this. Provisionally, we recommend using the graph of p_t against t as a guide, paying particular attention to any values of t where the probability level falls more sharply than usual.

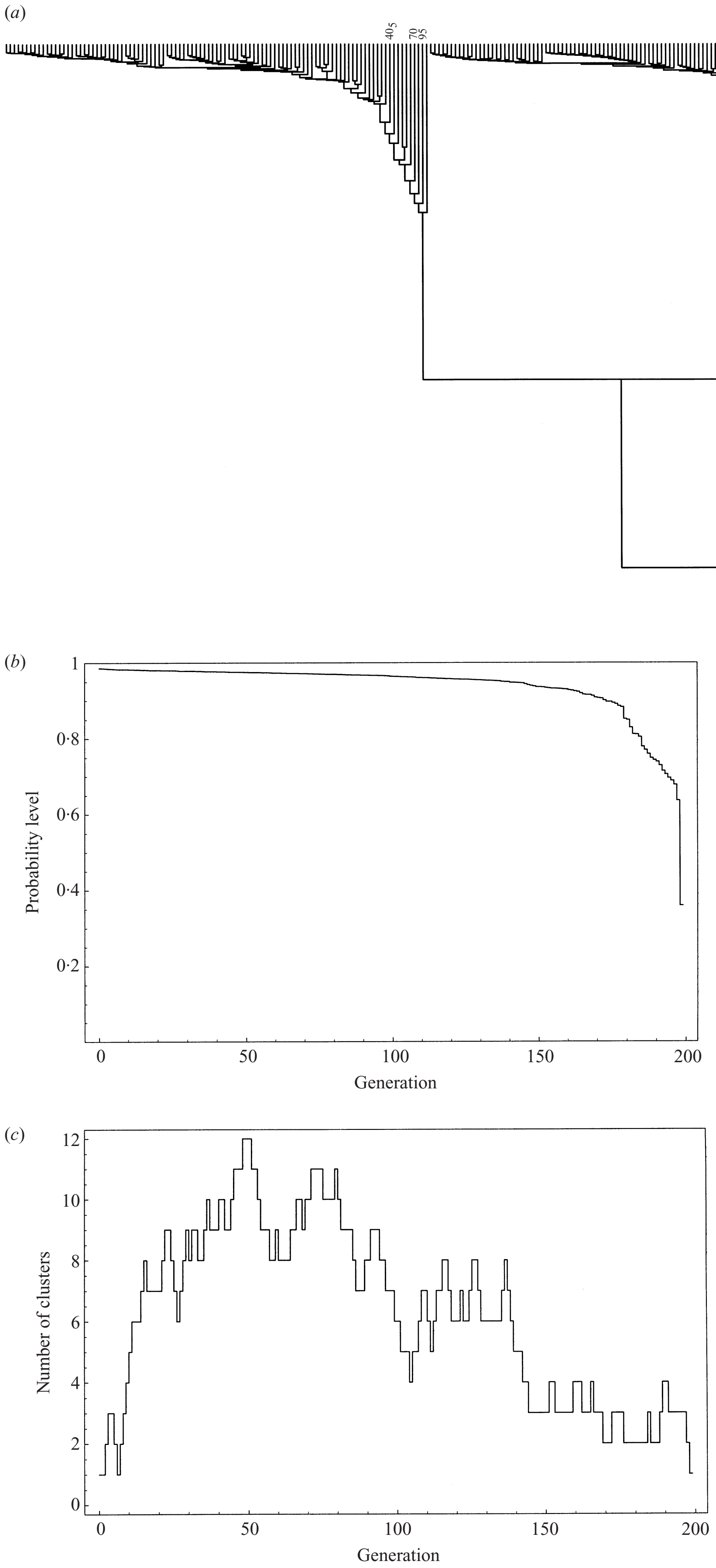
Once we are reasonably confident of the number of well-supported clusters in our sample, we could then seek a more accurate assignment of individuals, again using the furthest neighbour algorithm, but this time based on the array of conditional co-assignment probabilities $\Pi_d(\kappa = k, \boldsymbol{X})$. However, because of the residual uncertainty about the existence of additional source populations, it might still be preferable to assign individuals on the basis of the less restrictive conditional co-assignment probabilities $\Pi_d(\kappa \geqslant k, \boldsymbol{X})$.

7. Accuracy of Bayesian inferences

In order to test the performance of the Bayesian inference procedures described above, we applied these procedures to artificial data sets, generated using coalescent simulations (Hudson, 1991), of the following simple *divergence model*. Two isolated populations, with (haploid) effective population sizes N_1, N_2, respectively, separated t generations before the present. The common ancestral population was perturbed away from mutation–drift equilibrium a further t_0 generations before this split, by a change in population size from N_e to N_0. The mutation process follows the infinite-alleles model (IAM), with mutation rate U. The diversity of the common ancestral population at mutation–drift equilibrium is determined by the parameter $\Theta = 2N_eU$. The divergence of the two populations from their common ancestral populations is determined principally by the parameters $\tau_1 = t/N_1$, $\tau_2 = t/N_2$, respectively.

Throughout, the parameter values in the divergence model where: $U = 5\times 10^{-5}$ and $N_e = 10^4$, so that $\Theta = 1$; $N_0 = 2\times 10^4$ and $t_0 = 2500$, so that $\tau_0 = 0{\cdot}125$; and $N_1 = N_2 = 10^4$. However, the parameter t varied. In the *strong divergence model*, $t = 2500$, so that $\tau_1 = \tau_2 = 0{\cdot}25$. In the *weak divergence model*, $t = 625$, so that $\tau_1 = \tau_2 = 0{\cdot}0625$. Finally, in the *single population model*, two divergent populations (of equal size) are generated with $t = 5000$ (so that $\tau_1 = \tau_2 = 0{\cdot}5$) and then pooled, and brought to Hardy–Weinberg and linkage equilibrium. In practice, this was achieved by random assignment of alleles to individuals in the pooled sample. The motivation for this was to increase the difficulty of the inference problem, by creating substantial departures from Ewens' sampling distribution (which is favoured by the prior on the allele frequencies).

Each artificial data set contained 10 loci, unless stated otherwise. For each Bayesian computation, the



https://doi.org/10.1017/S001667230100502X Published online by Cambridge University Press

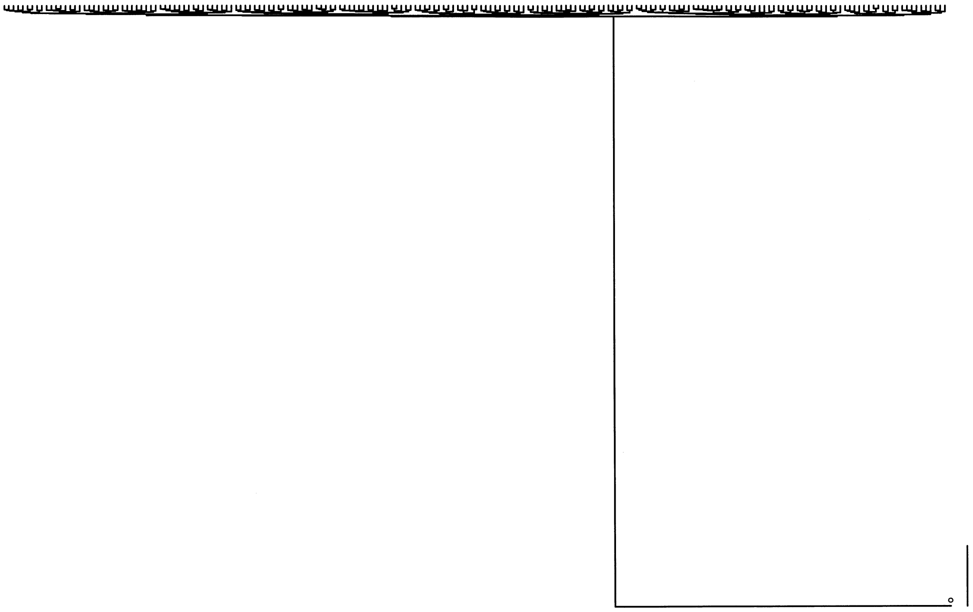


Fig. 2. Tree plot for data set 2, with dimension $d = 2$. All individuals 1–200 were drawn from a single population 1. The tree offers no evidence of population subdivision, and is consistent with the existence of a single source population.

choice of prior, and the parameters of the Markov chain sampler, were as follows, except where explicitly stated otherwise. The prior $\pi_K(\kappa)$, on the number of source populations κ, was flat ($u = 1$), with $\nu = 4$. The prior $\pi_\Theta(\theta)$ was degenerate, with a fixed value of $\theta = 1$. The Markov chain sampler was run for 100 000 iterations, to generate 10 000 observations with a period of 10. After inspecting the sample path of the number of populations κ, and the log likelihood, we concluded that it was sufficient to exclude the first 1000 observations (10 000 iterations) as burn-in.

Binary trees were constructed using the algorithm described in Section 6, with dimension $d = 2$, unless stated otherwise. The height of each node corresponds to the probability level of the cluster defined by that node. All terminal nodes are defined to have height $p = 1$. The trees were viewed using the software package TreeView (Page, 1996). For larger samples, the software package NJplot (Perriere & Gouy, 1996) is particularly useful.

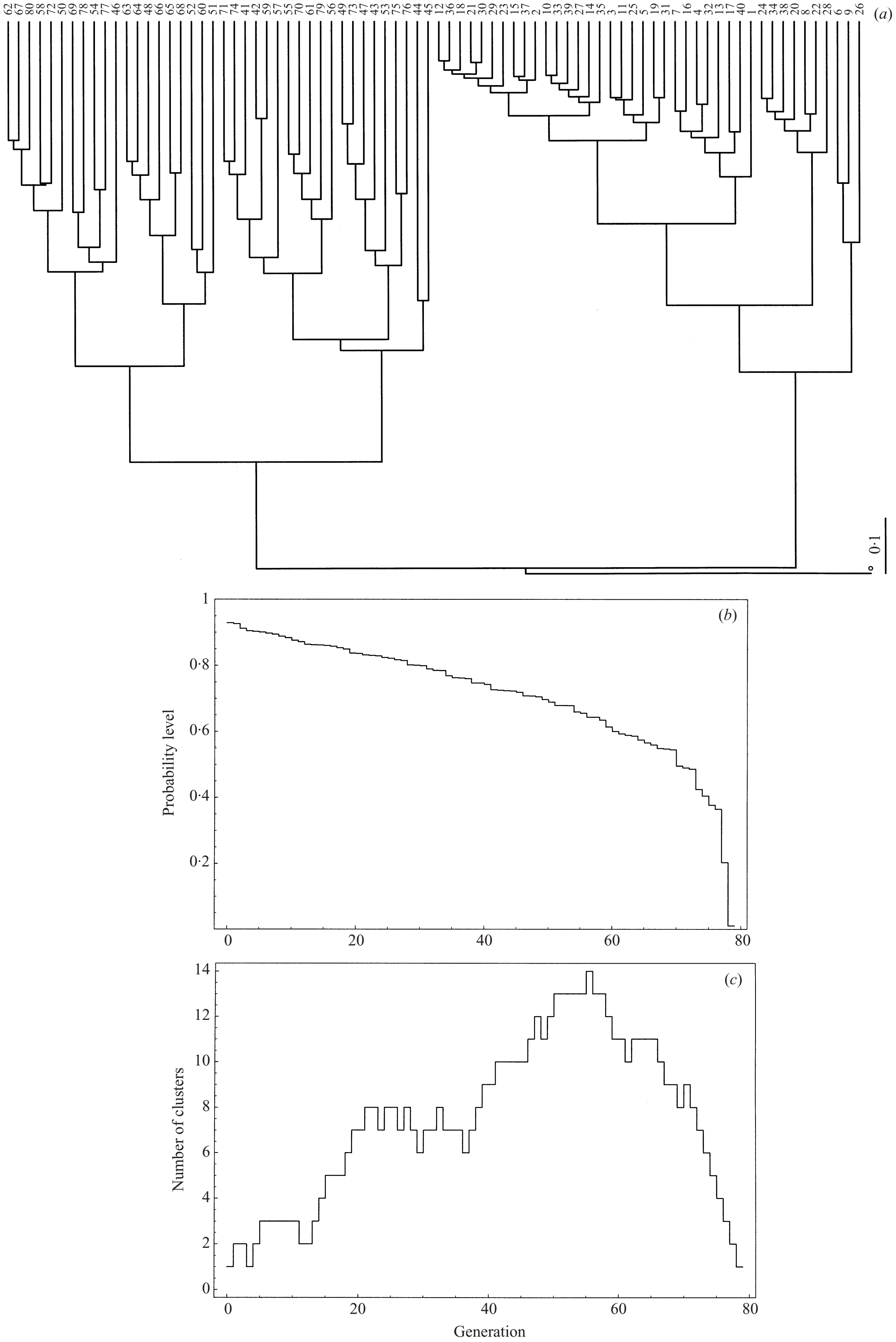
Data set 1 was generated under the *weak divergence model*, with sample sizes of $n_1 = n_2 = 100$ diploid individuals from each of the two source populations. The posterior distribution of the number of source populations κ (Table 1) clearly favours the true value of $\kappa = 2$. The tree (Fig. 1*a*) also indicates a well-supported bi-partition of the sample. The Bayesian probability level plot (Fig. 1*b*) shows a gradual decline, until the lumping of the two remaining clusters into a single cluster precipitates a dramatic fall in probability level, indicating that this last agglomeration is not supported by the data. For comparison, we provide the plot of the number of clusters (defined by open internal nodes) against the generation t (Fig. 1*c*).

So, the Bayesian analysis reveals a clear bi-partition of the sample. But is this the 'true' bi-partition? For the bi-partition of the sample specified by the model, $F_{ST} = 0{\cdot}0699$ and $F_{IS} = 0{\cdot}0008$ (using Weir & Cockerham's (1984) multi-locus estimators). The bi-partition identified by the tree reassigns 5 individuals. This bi-partition coincides exactly with the maximum likelihood partition for $\kappa = 2$, obtained by the simulated annealing procedure described in Section 3. For this partition, $F_{ST} = 0{\cdot}0716$ and $F_{IS} = -0{\cdot}0001$, suggesting that this bi-partition provides a better fit to this particular data set than the bi-partition specified by the underlying model. We also applied the program *Structure*, of Pritchard *et al.* (2000) to this data set. It did not yield a consistent result. In some runs, it identified bi-partitions that were very close to the maximum likelihood bi-partition. In another run, it identified a tri-partition.

Data set 2 was generated under the *single population*

Fig. 1. (*a*) Tree plot for data set 1, using dimension $d = 2$. Individuals 1–100 were drawn from population 1, and individuals 101–200 were drawn from population 2. The bi-partition identified by the tree assigns individual 152 to population 1, and individuals 5, 40, 70 and 95 to population 2. This bi-partition coincides exactly with the maximum likelihood partition for $\kappa = 2$. (*b*) Plot of the probability level p_t against the generation t, for data set 1, with $d = 2$. (*c*) Plot of the number of clusters $|\lambda_t(\boldsymbol{X})|$ in the core partition $\lambda_t(\boldsymbol{X})$ against the generation t, for data set 1, with $d = 2$.

https://doi.org/10.1017/S001667230100502X Published online by Cambridge University Press


62 67 80 58 72 50 69 78 54 77 46 63 64 48 66 65 68 52 60 51 71 74 41 42 59 57 55 70 61 79 56 49 73 47 43 53 75 76 44 45 12 36 18 21 30 29 23 15 37 2 10 33 39 27 14 35 3 11 25 5 19 31 7 16 4 32 13 17 40 1 24 34 38 20 8 22 28 6 9 26
(a)
0·1
0
(b)
Probability level
1
0·8
0·6
0·4
0·2
0
20
40
60
80
(c)
Number of clusters
14
12
10
8
6
4
2
0
0
20
40
60
80
Generation

model, with a sample size of $n = 200$ diploid individuals. The posterior distribution of κ (Table 1) clearly favours the true value of $\kappa = 1$. The tree (Fig. 2) strongly suggests the existence of a single source population. The probability level plot shows a slight, and very gradual, decline, down to the final agglomeration of all individuals into a single cluster.

Smaller samples pose more difficult inference problems. As the data provide less information, the choice of prior has a greater influence on the posterior distribution. Data set 3 was generated under the *strong divergence model*, with sample sizes of $n_1 = n_2 = 40$ diploid individuals from each of the two source populations. Data set 4 was generated under the *weak divergence model*, with the same sample sizes. Data set 5 was generated under the *single population model*, with a sample size of $n = 80$ diploid individuals.

Despite the relatively strong population subdivision in data set 3 ($F_{ST} = 0{\cdot}215$), the posterior distribution of κ (Table 1) is concentrated at the maximum value of $\nu = 4$. This is a poor inference for κ, with a greatly exaggerated impression of our confidence in this inference. However, the tree (Fig. 3*a*), together with the probability level plot, did suggest a bi-partition, or possibly a tri-partition, of the sample. There was an unusually large fall in probability level when the number of clusters was reduced from three to two, and an even greater fall when the two remaining clusters were lumped into one (Fig. 3*b*). Individuals 1–40 were drawn from population 1, and individuals 41–80 were drawn from population 2. The bi-partition identified by the tree coincides exactly with this, as does the maximum likelihood bi-partition.

For data set 4, the posterior distribution of κ (Table 1) is again concentrated at the maximum value of $\nu = 4$. In this case, the tree, and the probability level plot, did not offer clear evidence in support of a bi-partition. However, the bi-partition induced by the tree coincided with that specified by the model (1–40:41–80), except that individual 56 was assigned to population 1. The maximum likelihood bi-partition returns individual 56 to population 2, and assigns individual 23 to population 2.

Despite the absence of appreciable subdivision in data set 5, the posterior distribution of κ (Table 1) is again concentrated at the maximum value of $\nu = 4$. The tree, and the probability level plot, offered no evidence of subdivision. The probability level declined gradually all the way down to the final agglomeration of all individuals. For this data set, we also constructed a tree using the clustering algorithm with dimension $d = 3$. Again, the tree, and the probability level plot, offered no evidence of subdivision.

The strong tendency for the posterior distribution of κ to be concentrated at its maximum value ν whenever the sample is small is a serious concern. We believe that the reason for this is that the prior on the allele composition used in these examples is too restrictive, and that this prior becomes very influential when the sample size is small. This issue is explored in more detail in the next section. Even when the sample size is small, the tree, together with the probability level plot, provides robust inferences about the source populations present in the sample.

8. Sensitivity analysis

In order to understand the strong upward bias in the posterior distribution of κ, when the sample is small, we investigated the sensitivity of the posterior distribution to the choice of prior.

To see the influence of the parameter ν, of the prior on κ, we reanalysed data set 3, with $u = 1$ and $\nu = 8$. Again, the posterior distribution of κ was concentrated at its maximum value, now $\nu = 8$. The tree provided less evidence of a bi-partition than before. But despite this, the bi-partition defined by the tree remained the same as before.

In view of this extreme dependence of $\pi_K(\kappa \mid \boldsymbol{X})$ on ν, when the sample size is 80, we also reanalysed data set 1 (sample size 200) with the same prior. However, in this case the change of prior had almost no effect on $\pi_K(\kappa \mid \boldsymbol{X})$. The tree revealed the same bi-partition, and showed much similarity on a finer scale.

The other parameter of the prior on κ is u. We reanalysed data sets 3 and 5, with $u = 0{\cdot}75$ and $\nu = 4$. Surprisingly, this had almost no effect on the posterior distribution of κ (Table 1). The new tree for data set 3 provided stronger evidence for the maximum likelihood bi-partition. The new tree for data set 5 again offered no evidence of population subdivision.

The prior distribution of the allele composition influences the posterior distribution of κ. In all the above examples, we have used a Poisson–Dirichlet distribution, with $\theta = 1$. In this case, the prior (or hyper-prior) on θ is degenerate. We suspect that this choice has favoured the identification of many source populations with lower allelic diversity, over the alternative of fewer source populations with higher

Fig. 3. (*a*) Tree plot for data set 3, with dimension $d = 2$. Individuals 1–40 were drawn from population 1, and individuals 41–80 were drawn from population 2. The bi-partition identified by the tree coincides exactly with this, as does the maximum likelihood bi-partition. (*b*) Plot of the probability level p_t against the generation t, for data set 3, with $d = 2$. (*c*) Plot of the number of clusters $|\lambda_t(\boldsymbol{X})|$ in the core partition $\lambda_t(\boldsymbol{X})$ against the generation t, for data set 3, with $d = 2$.

https://doi.org/10.1017/S001667230100502X Published online by Cambridge University Press

allelic diversity. To test this idea, we reanalysed data sets 3 and 5, with $\theta = 10$. Increasing θ leads to a substantial shift in the posterior distribution of κ, away from its maximum value (Table 1). The new tree for data set 3 provided stronger evidence for the maximum likelihood bi-partition. The new tree for data set 5 provided stronger evidence for a single source population.

These preliminary results suggest that the solution to the small-sample bias in the posterior distribution of κ, may be to choose a suitable diffuse prior on the parameter θ, so that the data can influence the joint posterior distribution of θ and κ. We hope to resolve this question in the near future, by experimenting with different diffuse priors on θ. Progress is impeded by the long run time for the Bayesian computation.

For a sample of 80 individuals, genotyped at 10 marker loci, a run of 100000 iterations (10000 observations, with a period of 10) takes about 8 hours on a Pentium 300 processor. For a sample of 200 individuals, genotyped at 10 marker loci, 100000 iterations takes about 24 hours. The run time of the clustering algorithm grows rapidly with the sample size. For a sample of 80 individuals, it takes about 10 minutes to construct a tree, using dimension 2, and 40 minutes using dimension 3. For a sample of 200 individuals, it takes about 3 hours using dimension 2.

9. The general problem of making inferences about partitions

The problem of assigning the individuals in a sample to panmictic source populations is formulated here as a problem of inferring the partition of the sample induced by the assignment of individuals to source populations. This is in contrast to the alternative formulation, as a problem of inferring the proportions of the different source populations in a mixture.

The data in the assignment problem can be presented in the form of a multi-dimensional contingency table, where the factors are the loci. We are looking for the evidence of populations within which there are no statistical associations among factors (linkage disequilibrium). Mixture models, assignment problems and multi-dimensional contingency tables are all amenable to a similar Bayesian analysis.

Pritchard *et al.* (2000) also introduced a Bayesian method for assigning individual allele copies to source populations, where the source populations are again assumed to be panmictic, but are not identified *a priori*. The objective of this analysis is to reveal patterns of hybridization and introgression. Here, the parameter of interest is the proportion of each individual's genome which is derived from a particular source population. This is a very challenging inference problem, and it appears that useful inferences can only be made when there is strong prior information about the number of source populations, or the distribution of the hybrid index, or preferably strong prior information about both of these.

The problem of assigning allele copies to source populations is really an alternative formulation of the problem of inferring the distribution of the hybrid index (or, in the case of multiple source populations, its multivariate generalization). Barton (2000) has recently addressed this problem. The ultimate objective of his analysis was to estimate measures of linkage disequilibrium within a single spatially localized population. This was achieved by first estimating the distribution of the hybrid index, jointly with the allele frequencies and their divergence between source populations. Barton's analysis incorporates stronger prior information in that it assumes that all linkage disequilibrium was ultimately generated by admixture of two source populations, each of which was at linkage equilibrium. The additional assumption that the geographically localized population has reached a migration–recombination balance (a quasi-equilibrium) is not strictly necessary. Barton used a maximum likelihood approach, but the same model could usefully be analysed from a Bayesian point of view.

When the work presented above was near completion, our attention was drawn to a number of important recent references, in addition to Pritchard *et al.* (2000). Green (1995) developed a Markov chain Monte Carlo sampler for performing Bayesian computations in situations where one of the parameters of the model is a partition of a set. Our **change** κ proposal process, with 'unify' and 'divide' proposals, is essentially identical to the 'birth' and 'death' proposal processes of Green (1995; re-named 'split' and 'combine' in Richardson & Green, 1997). Green (1995) also chose a flat prior on the number of subsets (κ in our notation).

Assignment problems and mixture problems are examples of *model choice* problems where the alternative models have different numbers of parameters. In such problems, the posterior distribution is neither a discrete distribution nor a joint probability density. For this reason, Green (1995) presented the underlying Markov chain Monte Carlo algorithm as a new generalization of the Metropolis–Hastings algorithm, which is now commonly referred to as the 'reversible jump sampler', or sometimes the 'Metropolis–Hastings–Green algorithm'. However, any Markov chain which is simulated on a computer is necessarily a discrete Markov chain, where all probability densities are replaced by discrete approximations. From this point of view, we are still using the Metropolis–Hastings algorithm.

Green (1995) focused on the problem of evaluating the evidence in support of a particular preconceived partition of the sample, and did not confront the

https://doi.org/10.1017/S001667230100502X Published online by Cambridge University Press

general problems associated with inferring an unknown partition of a set. Richardson & Green (1997) tackled this more challenging problem, and recognized the problems associated with the arbitrary labelling of the subsets in the partition (or equivalently, the components of the mixture distribution), and the consequent 'label switching' in the output from the Markov chain sampler. The solution which they recommended was processing of the output to rank the components of each observation in a consistent order. In his contribution to the published discussion of that paper, Stephens (1997) pointed out the shortcomings of this approach. These points are developed further in Stephens (2000*b*), where an alternative type of post-processing is recommended. The solution proposed by Stephens (2000*b*) is to apply a k-means type clustering algorithm to the set of observations from the Markov chain sampler. Here, the problem is presented as one of assigning the observations from the Markov chain sampler to the distinct symmetric modes of an exchangeable posterior distribution. We avoided this label-switching problem by basing our Bayesian inferences about the sample partition on the posterior co-assignment probabilities. These are entirely independent of any labelling of the source populations.

Besides the technical problem of label-switching, there is a more fundamental problem of what to do when the mode of the posterior distribution of partitions $\pi_T(\tau \mid \boldsymbol{X})$ is rather flat. To put it another way, there may be many individuals which are difficult to assign to source populations. This problem has already been discussed in some detail above. Our solution is to identify clusters of individuals whose assignment together is well supported by the posterior distribution. This is achieved by applying simple hierarchical clustering algorithms to the arrays of posterior co-assignment probabilities.

Similar considerations apply to any inference problem where the partition of a set (or the composition of a finite mixture) is the parameter of interest. This is not the case when mixture models are used for density estimation (Richardson & Green, 1997; Stephens, 2000*a*, *b*), where the ordinates of the density, together with credibility intervals, are of principal interest.

An important generalization of the assignment problem treated above is the assignment of individuals to source populations, where reference samples are available from certain source populations. The inclusion of reference samples in the data changes substantially the nature of the inference problem. When we have reference samples from particular source populations, these populations can be associated with fixed labels, so that we can use the posterior probability that an individual is assigned to such a population as an assignment criterion. However, if there are individuals which have a low probability of being assigned to the reference populations, then we will also need to be able to assign these individuals to alternative clusters, using methods like those introduced above. We will treat this problem in a forthcoming paper.

10. Prospects for Bayesian analysis of Hardy–Weinberg and linkage disequilibrium

In many situations, the first question of interest is: do we have a (random, or non-stratified) sample from a single panmictic population (at, or at least close to, Hardy–Weinberg and linkage equilibrium), or do we have something more complicated? One possible measure of the evidence in support of a single panmictic population is the Bayes factor for $\kappa = 1$ against the alternative of $\kappa > 1$, which is given by

$$B_1 = \left\{ \frac{\pi(\kappa = 1 \mid \boldsymbol{X})}{1 - \pi(\kappa = 1 \mid \boldsymbol{X})} \right\} \left\{ \frac{1 - \pi_K(\kappa = 1)}{\pi_K(\kappa = 1)} \right\}.$$

This is independent of the prior distribution $\pi_K(\kappa)$, on the number of source populations κ. However, it does still depend on the prior distribution $\pi_{T \mid K}(\tau \mid \kappa)$, of sample partitions conditional upon κ. It also depends on the prior distribution on the allele frequencies. Reporting this Bayes factors is a possible likelihood-based alternative to classical tests of Hardy–Weinberg and linkage equilibrium (such as those of Guo & Thompson, 1992; Zykin *et al.*, 1995; and Rousset & Raymond, 1995). However, if inbreeding, or stratification of the sample with respect to kinship, are among the alternative hypotheses under consideration, then further likelihood-based comparisons should be performed (see, for example, Ayres & Balding, 1998; and Zhivotovsky, 1999). Arguably, this particular Bayes factor, B_1, is only relevant when the mode of the posterior distribution $\pi_K(\kappa \mid \boldsymbol{X})$ is at $\kappa = 1$.

This Bayesian approach to evaluating the evidence supporting Hardy–Weinberg and linkage equilibrium (against the alternative of population subdivision) is not without its problems. When samples are small, the posterior distribution of κ, and hence the Bayes factors B_1, can greatly exaggerate the evidence against a single population at Hardy–Weinberg and linkage equilibrium. This is a potentially serious problem, since the number of sampled individuals belonging to each source population is part of what we are trying to infer. We hope that this problem can be overcome by choosing an appropriate diffuse prior on the allele frequencies.

Both the Bayesian and maximum likelihood procedures for inferring the sample partition, described above, have been incorporated into a software package, *Partition*, available at http://www.univ-montp2.fr/~genetix/partition.htm.

https://doi.org/10.1017/S001667230100502X Published online by Cambridge University Press

Appendix

(i) *Integrating out the allele frequencies*

The posterior distribution $\pi(\tau, \theta \mid \boldsymbol{X})$ is obtained from (11) by integrating out the allele frequencies as follows:

$$\pi(\tau, \theta \mid \boldsymbol{X}) = \int_{\boldsymbol{p}} \pi(\tau, \boldsymbol{p}, \theta \mid \boldsymbol{X})\, \boldsymbol{dp}$$

$$= C^{-1} \left(\prod_{i=1}^{\kappa} P(\boldsymbol{D}_i \mid \boldsymbol{d}_i) \right) \left(\prod_{i=1}^{\kappa} \prod_{a=1}^{m} \left(\int_{\boldsymbol{p}_{i,a}} P(\boldsymbol{d}_{i,a} \mid \boldsymbol{p}_{i,a}) \pi_{P|\Theta}(\boldsymbol{p}_{i,a} \mid \theta)\, \boldsymbol{dp}_{i,a} \right) \pi_{\Theta}(\theta_{i,a}) \right) \pi_T(\tau). \quad \text{(A1)}$$

From the allele counts $\boldsymbol{d}_{i,a} = (d_{i,a}(1), \ldots, d_{i,a}(r_{i,a}))$, we can construct the *sample configuration* $\boldsymbol{k}_{i,a} = (k_{i,a}(1), \ldots, k_{i,a}(2n_{i,a}))$, where $k_{i,a}(d)$ denotes the number of distinct allele types that are represented with an allele count of exactly d copies in the sample from population i, at locus a. So, $\boldsymbol{k}_{i,a}$ is a 'partition' of the integer $2n_i$. Let $\boldsymbol{k}_i = (\boldsymbol{k}_{i,1}, \ldots, \boldsymbol{k}_{i,m})$ and $\boldsymbol{k} = (\boldsymbol{k}_1, \ldots, \boldsymbol{k}_\kappa)$. The total number of distinct allele types at locus a represented among the individuals assigned to source population i is $k_{i,a}(1) + \cdots + k_{i,a}(2n_{i,a}) = K_{i,a}$. If the ordered allele frequencies in the population have the Poisson–Dirichlet distribution, then the integral

$$\int_{\boldsymbol{p}_{i,a}} P(\boldsymbol{d}_{i,a} \mid \boldsymbol{p}_{i,a}) \pi_{P|\Theta}(\boldsymbol{p}_{i,a} \mid \theta_{i,a})\, \boldsymbol{dp}_{i,a} = P(\boldsymbol{k}_{i,a} \mid \theta_{i,a}) \quad \text{(A2)}$$

is given by the Ewens sampling distribution (Ewens, 1972; Karlin & McGregor, 1972):

$$P(\boldsymbol{k} \mid \theta) = P_r(\boldsymbol{k} \mid \theta) = \frac{r!}{\theta(\theta+1)\cdots(\theta+r-1)} \frac{\theta^K}{\left(\prod_{d=1}^{r} k(d)!\, d^{k(d)} \right)}, \quad \text{(A3)}$$

where r is the total number of allele copies, and K is the total number of distinct allele types represented in the sample. So the factor

$$P(\boldsymbol{X} \mid \tau, \theta) = \prod_{i=1}^{\kappa} P(\boldsymbol{D}_i \mid \boldsymbol{d}_i) \left(\prod_{a=1}^{m} P(\boldsymbol{k}_{i,a} \mid \theta_{i,a}) \right) \quad \text{(A4)}$$

in (12) is the likelihood function for the model, parameterized by $\theta = (\theta_1, \ldots, \theta_\kappa)$ (rather than by the allele frequencies).

The relationship (A2) between the Ewens sampling distribution and the Poisson–Dirichlet distribution was derived by Watterson (1976), and then more rigorously by Kingman (1977). Hoppe (1987) obtained a more direct proof using the relationship between the GEM (Griffiths–Engen–McClowskey) distribution (Engen, 1975) and the Poisson–Dirichlet distribution (established by Patil & Taillie, 1977).

(ii) *The proposal processes of the Metropolis–Hastings algorithm*

The proposal process **exchange** can change the composition, but not the size, of the existing subsets of the sample, while the proposal process **transfer** can change both the size and the composition of the existing subsets of the sample. The proposal process **change** κ can change the number of subsets into which the sample is partitioned, as well as the composition and size of these subsets. The proposal process **jiggle** θ simply adjusts the parameters of the prior on the allele frequencies in the populations (or equivalently, the prior on the allelic 'configuration' of the samples from these populations).

Each cycle of the Markov chain begins with a **change** κ proposal, followed by the sequence of proposal processes: **transfer**, **exchange**; repeated n times (where n is the number of individuals in the sample); and the cycle ends with the **jiggle** θ proposal process. In the case where the prior $\pi_\Theta(\theta_{i,a})$ is degenerate (so that the values of the parameters $\theta_{i,a}$ are all fixed at the same value θ), the proposal process **jiggle** θ is omitted.

In general, the posterior probability ratio for a proposed change from state (τ, θ) to (τ', θ') is

$$\frac{\pi(\tau', \theta' \mid \boldsymbol{X})}{\pi(\tau, \theta \mid \boldsymbol{X})} = \frac{\prod_{i=1}^{\kappa'} P(\boldsymbol{D}'_i \mid \boldsymbol{d}'_i) \left(\prod_{a=1}^{m} P(\boldsymbol{k}'_{i,a} \mid \theta'_{i,a}) \pi_\Theta(\theta'_{i,a}) \right)}{\prod_{i=1}^{\kappa} P(\boldsymbol{D}_i \mid \boldsymbol{d}_i) \left(\prod_{a=1}^{m} P(\boldsymbol{k}_{i,a} \mid \theta_{i,a}) \pi_\Theta(\theta_{i,a}) \right)} \frac{\pi_T(\tau')}{\pi_T(\tau)}. \quad \text{(A5)}$$

https://doi.org/10.1017/S001667230100502X Published online by Cambridge University Press

Changes in the partition τ or the parameter θ have no effect on the total number of heterozygotes $h_{1,a}+\cdots+h_{\kappa,a}$, at individual loci, and the total number of heterozygotes $h_1+\cdots+h_\kappa$. So the factor $2^{h_1+\cdots+h_\kappa}$ cancels out of the likelihood ratio. From (3) and (A3), the factor $\prod_{a=1}^{m}(2n_i)!$ cancels out of the likelihood for each population, leaving

$$P(\boldsymbol{D}_i\,|\,\boldsymbol{d}_i)\left(\prod_{a=1}^{m}P(\boldsymbol{k}_{i,a}\,|\,\theta_{i,a})\right)=2^{h_i}\frac{n_i!}{\prod\limits_{G\in R(\boldsymbol{X})}D_i(\boldsymbol{G})!}\left(\prod_{a=1}^{m}\left(\prod_{d=1}^{n}\frac{((d-1)!)^{k_{i,a}(d)}}{k_{i,a}(d)!}\right)\frac{\theta^{K_{i,a}}}{\theta_{i,a}(\theta_{i,a}+1)\cdots(\theta_{i,a}+2n_i-1)}\right). \tag{A6}$$

The proposal process **jiggle** θ was not used in the Bayesian computations presented in Sections 7 and 8. In this proposal process, we choose a source population i, at random, and change the values of the parameters $\theta_{i,a}$, at every locus a. The proposal distribution is of the form

$$q(\theta_i\rightarrow\theta_i^{'})=\prod_{a=1}^{m}Q(\theta_{i,a}^{'}\,|\,\theta_{i,a},\,\delta), \tag{A7}$$

where $Q(\theta'\pm\theta,\,\delta)$ is a distribution which corresponds to the following proposal process. First, we choose whether to decrease or increase $\theta_{i,j}(\theta'<\theta$ or $\theta<\theta')$. Both outcomes have probability 1/2. Second, we choose the value of θ' from a uniform distribution on a certain open interval. If we have chosen to decrease $\theta_{i,a}(\theta'<\theta)$, then θ' is chosen from a uniform distribution on the interval $\langle\theta-\delta,\,\theta\rangle$ (when $\delta<\theta$), or on the interval $\langle 0,\,\theta\rangle$ (when $0<\theta<\delta$). If we have chosen to increase $\theta_{i,a}(\theta'<\theta)$, then θ' is chosen from a uniform distribution on the interval $\langle\theta,\,\theta+\delta\rangle$. Therefore, this proposal process results in a probability density $Q(\theta'\,|\,\theta,\,\delta)$, which is given by

$$Q(\theta'\,|\,\theta,\,\delta)=\begin{cases}\dfrac{1}{2\delta}, & \delta<\theta\\[2ex] \dfrac{1}{2\theta}, & 0<\theta<\delta\end{cases}. \tag{A8}$$

This proposal process ensures that $\theta_{i,a}$ remains in the open interval $\langle 0,\,\infty\rangle$.

In the case of the proposal process **jiggle** θ, the posterior probability ratio reduces to

$$\frac{\pi(\tau,\,\theta'\,|\,\boldsymbol{X})}{\pi(\tau,\,\theta\,|\,\boldsymbol{X})}=\prod_{a=1}^{m}\frac{P(\boldsymbol{k}_{i,a}\,|\,\theta_{i,a}^{'})}{P(\boldsymbol{k}_{i,a}\,|\,\theta_{i,a})}\frac{\pi_\Theta(\theta_{i,a}^{'})}{\pi_\Theta(\theta_{i,a})}. \tag{A9}$$

For a proposed change in the parameter $\theta_{i,a}$, from θ to θ', the likelihood ratio is

$$\frac{P(\boldsymbol{k}_{i,a}\,|\,\theta')}{P(\boldsymbol{k}_{i,a}\,|\,\theta)}=\left(\frac{\theta'}{\theta}\right)^{K_{i,a}}\frac{\theta(\theta+1)\cdots(\theta+2n_i-1)}{\theta'(\theta'+1)\cdots(\theta'+2n_i-1)}. \tag{A10}$$

When the prior $\pi_\Theta(\theta)$, on θ is a gamma distribution

$$\pi_\Theta(\theta)=\frac{1}{\beta^\alpha\Gamma(\alpha)}\theta^{\alpha-1}\exp\left(-\frac{\theta}{\beta}\right), \tag{A11}$$

the prior probability ratio reduces to

$$\frac{\pi_\Theta(\theta')}{\pi_\Theta(\theta)}=\left(\frac{\theta'}{\theta}\right)^{\alpha-1}\exp\left(-\frac{(\theta'-\theta)}{\beta}\right). \tag{A12}$$

In the proposal process **exchange**, a pair of populations is chosen at random, and an individual is chosen at random from within each population. These two individuals then exchange their assignments, so that each individual is reassigned to the population to which the other was previously assigned.

If $\kappa=1$, then the state $(\tau,\,\theta)$ of the Markov chain remains unchanged. If $\kappa>1$, a pair of distinct populations $i,j,\ (i\neq j)$ is chosen at random. Each of the $\binom{\kappa}{2}$ possible pairs have the same probability, $\binom{\kappa}{2}^{-1}$, of being chosen. And within each of the two subsets, each individual has the same probability of being chosen. Therefore, the probability of proposing an exchange between populations i and j is

$$q_{Ex}(\tau\rightarrow\tau')=\binom{\kappa}{2}^{-1}\frac{1}{n_in_j}. \tag{A13}$$

https://doi.org/10.1017/S001667230100502X Published online by Cambridge University Press

But since an exchange does not change the number of subsets, or the number of individuals in each subset, the probability of proposing the reverse of this exchange is

$$q_{Ex}(\tau' \rightarrow \tau) = \binom{\kappa}{2}^{-1} \frac{1}{n_i' n_j'} = \binom{\kappa}{2}^{-1} \frac{1}{n_i n_j}. \tag{A14}$$

Hence

$$\frac{q_{Ex}(\tau' \rightarrow \tau)}{q_{Ex}(\tau \rightarrow \tau')} = 1. \tag{A15}$$

In the case of an exchange, the posterior probability ratio reduces to

$$\begin{aligned}\frac{\pi(\tau', \theta \,|\, \boldsymbol{X})}{\pi(\tau, \theta \,|\, \boldsymbol{X})} &= \left(\prod_{i=1}^{\kappa} \frac{P(\boldsymbol{D}_i' \,|\, \boldsymbol{d}_i')}{P(\boldsymbol{D}_i \,|\, \boldsymbol{d}_i)} \left(\prod_{a=1}^{m} \frac{P(\boldsymbol{k}_{i,a}' \,|\, \theta_{i,a})}{P(\boldsymbol{k}_{i,a} \,|\, \theta_{i,a})} \right) \right) \frac{\pi_T(\tau')}{\pi_T(\tau)} \\ &= \left(\frac{P(\boldsymbol{D}_i' \,|\, \boldsymbol{d}_i')}{P(\boldsymbol{D}_i \,|\, \boldsymbol{d}_i)} \left(\prod_{a=1}^{m} \frac{P(\boldsymbol{k}_{i,a}' \,|\, \theta_{i,a})}{P(\boldsymbol{k}_{i,a} \,|\, \theta_{i,a})} \right) \right) \left(\frac{P(\boldsymbol{D}_j' \,|\, \boldsymbol{d}_j')}{P(\boldsymbol{D}_j \,|\, \boldsymbol{d}_j)} \left(\prod_{a=1}^{m} \frac{P(\boldsymbol{k}_{j,a}' \,|\, \theta_{j,a})}{P(\boldsymbol{k}_{j,a} \,|\, \theta_{j,a})} \right) \right) \frac{\pi_T(\tau')}{\pi_T(\tau)}.\end{aligned} \tag{A16}$$

Let $\boldsymbol{X}(I)$ denote the genotype of individual I. For each individual $I \in S$, at each marker locus a, we have a pair of allele types, $X_{a,1}(I)$, $X_{a,2}(I)$. By convention, the *ordering labels*, 1, 2, are chosen such that $X_{a,1}(I) \leqslant X_{a,2}(I)$.

Before the exchange, individual I is assigned to population i, and individual J is assigned to population j. After the exchange, individual I is assigned to population j, and individual J is assigned to population i. Note the massive cancellations in the ratio

$$\begin{aligned}&\frac{P(\boldsymbol{D}_i' \,|\, \boldsymbol{d}_i')}{P(\boldsymbol{D}_i \,|\, \boldsymbol{d}_i)} \left(\prod_{a=1}^{m} \frac{P(\boldsymbol{k}_{i,a}' \,|\, \theta_{i,a})}{P(\boldsymbol{k}_{i,a} \,|\, \theta_{i,a})} \right) \\ &\quad = \frac{n_i'!}{n_i!} \left(\prod_{G \in \mathscr{R}} \frac{D_i(\boldsymbol{G})!}{D_i'(\boldsymbol{G})!} \right) \left(\prod_{a=1}^{m} \left(\frac{\theta_{i,a}(\theta_{i,a}+1)\cdots(\theta_{i,a}+2n_i-1)}{\theta_{i,a}(\theta_{i,a}+1)\cdots(\theta_{i,a}+2n_i'-1)} \right) \prod_{d \in \mathscr{I}} \left(\frac{k_{i,a}(d)!}{k_{i,a}'(d)!} ((d-1)!\theta_{i,a})^{k_{i,a}'(d)-k_{i,a}(d)} \right) \right)\end{aligned} \tag{A17}$$

where $\mathscr{R}$ is the set of *distinct* genotypes among: $\boldsymbol{X}(I)$, $\boldsymbol{X}(J)$, and $\mathscr{I}$ is the set of all *distinct* values among: $d_{i,a}(X_{a,1}(I))$, $d_{i,a}(X_{a,2}(I))$, $d_{i,a}(X_{a,1}(J))$, $d_{i,a}(X_{a,2}(J))$, $d_{i,a}'(X_{a,1}(I))$, $d_{i,a}'(X_{a,2}(I))$, $d_{i,a}'(X_{a,1}(J))$, $d_{i,a}'(X_{a,2}(J))$, *excluding* zero.

In the proposal process **transfer**, a population is chosen at random and an individual is chosen at random from within this population, and reassigned to a different population. A transfer is not allowed to empty a subset, or to create a new subset. So, we must choose a population at random from among those populations for which the corresponding subset contains *more than one* individual. Let $\eta(\tau)$ denote the number of subsets of the sample (elements of the partition τ) which contain *more than* one individual.

If $\kappa = 1$, or $\kappa = \nu$, then the state (τ, θ) of the Markov chain remains unchanged. If $1 < \kappa < \nu$, first a 'donor' population, i, is chosen at random from among the $\eta(\tau)$ populations for which the corresponding subset contains *more than one* individual. (The constraint $\nu < n$ ensures that $\eta(\tau) > 0$.) Each of these populations has the same probability, $1/\eta(\tau)$, of being chosen. Second, a 'recipient' population, j, is chosen at random from among all the κ populations excluding population i. (Each of these populations has the same probability, $1/(\kappa-1)$, of being chosen.) Finally, an individual I is chosen at random from the donor population, and reassigned to the recipient population. (Each individual in the donor population has the same probability, $1/n_i$, of being chosen.) Therefore, the probability of proposing a transfer from population i to population j is

$$q_{Trans}(\tau \rightarrow \tau') = \frac{1}{\eta(\tau)(\kappa-1)n_i}. \tag{A18}$$

The probability of proposing the reverse of this exchange is

$$q_{Trans}(\tau' \rightarrow \tau) = \frac{1}{\eta(\tau')(\kappa'-1)n_j'} = \frac{1}{\eta(\tau')(\kappa-1)(n_j+1)}. \tag{A19}$$

Hence

$$\frac{q_{Trans}(\tau' \rightarrow \tau)}{q_{Trans}(\tau \rightarrow \tau')} = \frac{\eta(\tau)n_i}{\eta(\tau')n_j'} = \frac{\eta(\tau)n_i}{\eta(\tau')(n_j+1)}. \tag{A20}$$

https://doi.org/10.1017/S001667230100502X Published online by Cambridge University Press

In the case of a transfer, the posterior probability ratio again reduces to the formula given in (A16). But this time the massive cancellations leave

$$\frac{P(\boldsymbol{D}_i'\,|\,\boldsymbol{d}_i')}{P(\boldsymbol{D}_i\,|\,\boldsymbol{d}_i)}\left(\prod_{a=1}^{m}\frac{P(\boldsymbol{k}_{i,a}'\,|\,\theta_{i,a})}{P(\boldsymbol{k}_{i,a}\,|\,\theta_{i,a})}\right)$$

$$=\frac{n_i'!}{n_i!}\left(\frac{D_i(\boldsymbol{X}(I))!}{D_i'(\boldsymbol{X}(I))!}\right)\left(\prod_{a=1}^{m}\left(\frac{\theta_{i,a}(\theta_{i,a}+1)\cdots(\theta_{i,a}+2n_i-1)}{\theta_{i,a}(\theta_{i,a}+1)\cdots(\theta_{i,a}+2n_i'-1)}\right)\prod_{d\in\mathscr{I}}\left(\frac{k_{i,a}(d)!}{k_{i,a}'(d)!}((d-1)!\theta_{i,a})^{k_{i,a}'(d)-k_{i,a}(d)}\right)\right) \tag{A21}$$

where $\mathscr{I}$ is now the set of all the *distinct* values among: $d_{i,a}(X_{a,1}(I))$, $d_{i,a}(X_{a,2}(I))$, $d_{j,a}'(X_{a,1}(I))$, $d_{j,a}'(X_{a,2}(I))$, *excluding* zero.

The proposal process **change** κ is composed of two subprocesses: 'unify' and 'divide'. In the process 'unify', a pair of populations is chosen at random and the corresponding subsets of individuals (assigned to these populations) are unified to form a single subset of individuals, which is assigned to a new population. The process 'unify' can only be applied when $\kappa > 1$. A pair of distinct populations $i, j (i \neq j)$ is chosen at random, and their members are assigned to a single new population $i' (= \min\{i, j\})$. Each of the $\binom{\kappa}{2}$ possible pairs has the same probability, $\binom{\kappa}{2}^{-1}$, of being chosen. Therefore, the probability of proposing a 'unification' of populations i and j is

$$q_{\smile}(\tau \rightarrow \tau') = \binom{\kappa}{2}^{-1}. \tag{A22}$$

Whenever we create a new population, i', we also have to specify the values of the parameters $\theta_{i'} = (\theta_{i',1}, \ldots, \theta_{i',m})$. These values are chosen using a special proposal distribution:

$$q_{\smile}(\theta_i \rightarrow \theta_i') = \prod_{a=1}^{m} Q\left(\theta_{i',a}'\,\middle|\left(\frac{\theta_{i,a}+\theta_{j,a}}{2}\right), \delta\right), \tag{A23}$$

where $Q(\theta'\,|\,\theta, \delta)$ is the proposal distribution introduced in the explanation of **jiggle** θ, above. So, a unification is in fact a transition from (τ, θ) to (τ', θ'), which occurs with probability

$$q_{\smile}((\tau, \theta) \rightarrow (\tau', \theta')) = q_{\smile}(\tau \rightarrow \tau')q_{\smile}(\theta_i \rightarrow \theta_i'). \tag{A24}$$

In the process 'divide', a population i is chosen at random and the subset of individuals assigned to that population is bi-partitioned at random, to create two new non-empty disjoint subsets, $i' (= i)$, $j' (= \kappa' = \kappa + 1)$, which are assigned respectively to two new populations. The process 'divide' can only be applied when $\kappa < \nu$. A population, i, is chosen at random from among the $\eta(\tau)$ populations for which the corresponding subset contains *more than one* individual. (Each of these populations has the same probability, $1/\eta(\tau)$, of being chosen.) A bi-partition of this subset is then chosen at random from among all the possible bi-partitions of this subset (excluding the trivial 'bi-partition' into one part). (Each of the $S_{n_i}^{(2)} = 2^{n_i-1} - 1$ non-trivial bi-partitions has the same probability, $1/(2^{n_i-1} - 1)$, of being chosen.) Therefore, the probability of proposing a particular 'division' of population i is

$$q_{+}(\tau \rightarrow \tau') = \frac{1}{\eta(\tau)(2^{n_i-1} - 1)}. \tag{A25}$$

Since we have created two new populations, i', j', we also have to specify the values of the parameters $\theta_{i'} = (\theta_{i',1}, \ldots, \theta_{i',m})$, $\theta_{j'} = (\theta_{j',1}, \ldots, \theta_{j',m})$. This time, these values are chosen using the proposal distribution $q(\theta_i \rightarrow \theta_i')$ described above in the explanation of **jiggle** θ. So, a division is a transition from (τ, θ) to (τ', θ'), which occurs with probability

$$q_{\div}((\tau, \theta) \rightarrow (\tau', \theta')) = q_{\div}(\tau \rightarrow \tau')q(\theta_i \rightarrow \theta_{i'}')q(\theta_i \rightarrow \theta_{j'}'). \tag{A26}$$

If $\kappa = 1$, then 'divide' is applied. If $\kappa = \nu$, then 'unify' is applied. Otherwise, for $1 < \kappa < \nu$, with probability $1/2$ 'unify' is applied, and with probability $1/2$ 'divide' is applied. So, for $1 < \kappa < \nu$,

$$q_{Change\ \kappa}((\tau, \theta) \rightarrow (\tau', \theta')) = \begin{cases} \tfrac{1}{2}q_{\smile}((\tau, \theta) \rightarrow (\tau', \theta')), \text{ when } \tau \rightarrow \tau' \text{ is a `unification'}, \\ \tfrac{1}{2}q_{\div}((\tau, \theta) \rightarrow (\tau', \theta')), \text{ when } \tau \rightarrow \tau' \text{ is a `division'}. \end{cases} \tag{A27}$$

https://doi.org/10.1017/S001667230100502X Published online by Cambridge University Press

In the case of a 'division', the posterior probability ratio reduces to

$$\frac{\pi(\tau',\theta'\,|\,\boldsymbol{X})}{\pi(\tau,\theta\,|\,\boldsymbol{X})}=\frac{\prod_{i=1}^{\kappa'}P(\boldsymbol{D}'_i\,|\,\boldsymbol{d}'_i)\left(\prod_{a=1}^{m}P(\boldsymbol{k}'_{i,a}\,|\,\theta'_{i,a})\pi_\Theta(\theta'_{i,a})\right)}{\prod_{i=1}^{\kappa}P(\boldsymbol{D}_i\,|\,\boldsymbol{d}_i)\left(\prod_{a=1}^{m}P(\boldsymbol{k}_{i,a}\,|\,\theta_{i,a})\pi_\Theta(\theta_{i,a})\right)}\frac{\pi_T(\tau')}{\pi_T(\tau)}$$

$$=\frac{\left(P(\boldsymbol{D}'_{i'}\,|\,\boldsymbol{d}'_{i'})\left(\prod_{a=1}^{m}P(\boldsymbol{k}'_{i',a}\,|\,\theta'_{i',a})\pi_\Theta(\theta'_{i',a})\right)\right)\left(P(\boldsymbol{D}'_{j'}\,|\,\boldsymbol{d}'_{j'})\left(\prod_{a=1}^{m}P(\boldsymbol{k}'_{j',a}\,|\,\theta'_{j',a})\right)\right)}{\left(P(\boldsymbol{D}_i\,|\,\boldsymbol{d}_i)\left(\prod_{a=1}^{m}P(\boldsymbol{k}_{i,a}\,|\,\theta_{i,a})\pi_\Theta(\theta_{i,a})\right)\right)}\frac{\pi_T(\tau')}{\pi_T(\tau)} \tag{A28}$$

In the case of a 'unification', the posterior probability ratio reduces to

$$\frac{\pi(\tau',\theta'\,|\,\boldsymbol{X})}{\pi(\tau,\theta\,|\,\boldsymbol{X})}=\frac{\prod_{i=1}^{\kappa'}P(\boldsymbol{D}'_i\,|\,\boldsymbol{d}'_i)\left(\prod_{a=1}^{m}P(\boldsymbol{k}'_{i,a}\,|\,\theta'_{i,a})\pi_\Theta(\theta'_{i,a})\right)}{\prod_{i=1}^{\kappa}P(\boldsymbol{D}_i\,|\,\boldsymbol{d}_i)\left(\prod_{a=1}^{m}P(\boldsymbol{k}_{i,a}\,|\,\theta_{i,a})\pi_\Theta(\theta_{i,a})\right)}\frac{\pi_T(\tau')}{\pi_T(\tau)}$$

$$=\frac{\left(P(\boldsymbol{D}'_{i'}\,|\,\boldsymbol{d}'_{i'})\left(\prod_{a=1}^{m}P(\boldsymbol{k}'_{i',a}\,|\,\theta'_{i',a})\pi_\Theta(\theta'_{i',a})\right)\right)}{\left(P(\boldsymbol{D}_i\,|\,\boldsymbol{d}_i)\left(\prod_{a=1}^{m}P(\boldsymbol{k}_{i,a}\,|\,\theta_{i,a})\pi_\Theta(\theta_{i,a})\right)\right)\left(P(\boldsymbol{D}_j\,|\,\boldsymbol{d}_j)\left(\prod_{a=1}^{m}P(\boldsymbol{k}_{j,a}\,|\,\theta_{j,a})\pi_\Theta(\theta_{j,a})\right)\right)}\frac{\pi_T(\tau')}{\pi_T(\tau)}. \tag{A29}$$

When we began this work, we were supported by a European Community grant (BI04 CT 961189) to Pierre Boursot. Some of the ideas presented in this article emerged from conversations with Francois Bonnhome, and we thank him for his continuing encouragement and interest in this work. We also thank Pierre Boursot, Nathalie Raufaste and Angela Karp for useful discussion, and Rosie Filipiak of *Genetical Research* for suggestions on the presentation. We particularly thank Nicolas Galtier, Nick Barton, Bill Hill and an anonymous referee for their many helpful comments on an earlier version of this manuscript. Kevin Dawson would also like to thank Ian Wilson and Mark Beaumont for some very helpful discussions about the Metropolis–Hastings algorithm in 1998.

References

Ayres, K. L. & Balding, D. J. (1998). Measuring departures from Hardy–Weinberg: a Markov chain Monte Carlo method for estimating the inbreeding coefficient. *Heredity* **80**, 769–777.

Barton, N. H. (1979). Gene flow past a cline. *Heredity* **43**, 333–339.

Barton, N. H. (2000). Estimating multilocus linkage disequilibria. *Heredity* **84**, 373–389.

Barton, N. H. & Hewitt, G. M. (1989). Adaptation, speciation and hybrid zones. *Nature* **341**, 497–503.

Belkhir, K. & Bonhomme, F. (2001). PartitionML, a maximum likelihood programme to estimate the best possible partition of a sample into independent panmictic units. (Submitted to *Bioinformatics*.).

Berge, C. (1971). Principles of Combinatories. New York: Academic Press (First published 1968, "Principes de Combinatoire", Paris: Dunod).

Besag, J., Green, P., Higdon, D. & Mengersen, K. (1995). Bayesian computation and stochastic systems. *Statistical Science* **10**, 3–66.

Burman, P. & Nolan, D. (1995). A general Akaike-type criterion for model selection in robust regression. *Biometrika* **82**, 877–886.

Cornuet, J.-M., Piry, S., Luikart, G., Estoup, A. & Solignac, M. (1999). New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics* **153**, 1989–2000.

Defays, D. (1977). An efficient algorithm for a complete link method. *Computing Journal* **20**, 364–366.

Engen, S. (1975). A note on the geometric series as a species frequency model. *Biometrika* **62**, 694–699.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3**, 87–112.

Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. (eds.) (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Graham, R. L. & Hell, P. (1985). On the history of the minimum spanning tree problem. *Annals of the History of Computing* **7**, 43–57.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Guénoche, A., Hansen, P. & Jaumard, B. (1991). Efficient algorithms for divisive hierarchical clustering with the diameter criterion. *Journal of Classification* **8**, 5–30.

Guo, S. W. & Thompson, E. A. (1992). Performing the exact test of Hardy–Weinberg proportion for multiple alleles. *Biometrika* **48**, 361–372.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Hoppe, F. M. (1987). The sampling theory of neutral alleles and an urn model in population genetics. *Journal of Mathematical Biology* **25**, 123–159.

Hudson, R. R. (1991). Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology*, vol. 7 (ed. D. Futuyama & J. Antonovics), pp. 1–44. Oxford: Oxford University Press.

Karlin, S. & McGregor, J. L. (1972). Addendum to a paper of W. Ewens. *Theoretical Population Biology* **3**, 113–116.

Kingman, J. F. C. (1975). Random discrete distributions. *Journal of the Royal Statistical Society* B **37**, 1–15, with discussion 16–22.

https://doi.org/10.1017/S001667230100502X Published online by Cambridge University Press

Kingman, J. F. C. (1977). The population structure associated with the Ewens sampling formula. *Theoretical Population Biology* **11**, 274–283.

Kingman, J. F. C. (1980). *The Mathematics of Genetic Diversity*. SIAM.

Kirkpatrick, S., Gelatt, C. D. Jr & Vecchi, M. P. (1983). Optimisation by simulated annealing. *Science* **220**, 671–680.

Milner, G. B., Teel, D. J., Utter, F. M. & Winans, G. A. (1985). A genetic method of stock identification in mixed populations of pacific salmon, *Oncorhynchus* spp. *Marine Fisheries Review* **47**, 1–8.

Olson, C. F. (1995). Parallel algorithms for hierarchical clustering. *Parallel Computing* **21**, 1313–1325.

Page, R. D. M. (1996). TreeView: an application to display phylogenetic trees on personal computers. *Computer Applications in the Biosciences* **12**, 357–358.

Patil, G. P. & Taillie, C. (1977). Diversity as a concept and its implications for random communities. *Bulletin of the International Statistical Institute* **47**, 497–515.

Perriere, G. & Gouy, M. (1996). WWW-Query: an on-line retrieval system for biological sequence banks. *Biochimie* **78**, 364–369.

Pritchard, J. K., Stephens, M. & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959.

Rannala, B. & Mountain, J. L. (1997). Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of Sciences of the USA* **94**, 9197–9221.

Richardson, S. & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society* B **59**, 731–758, with discussion 758–792.

Rousset, F. & Raymond, M. (1995). Testing heterozygote excess and deficiency. *Genetics* **140**, 1413–1419.

Smouse, P. E., Waples, R. S. & Tworek, J. A. (1990). A genetic mixture analysis for use with incomplete source population data. *Canadian Journal of Fisheries and Aquatic Science* **47**, 620–634.

Stephens, M. (1997). Discussion of paper by Richardson and Green. *Journal of the Royal Statistical Society* B **59**, 768–769.

Stephens, M. (2000*a*). Bayesian analysis of mixtures with an unknown number of components – an alternative to reversible jump methods. *Annals of Statistics* **28**, 40–74.

Stephens, M. (2000*b*). Dealing with label-switching in mixture models. *Journal of the Royal Statistical Society* B **62**, 795–809.

Walley, P. (1996). Inferences from multinomial data: learning from a bag of marbles. *Journal of the Royal Statistical Society* B **58**, 1–34, with discussion 34–57.

Watterson, G. A. (1976). The stationary distribution of the infinitely-many neutral alleles diffusion model. *Journal of Applied Probability* **13**, 639–651.

Weir, B. S. & Cockerham, C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370.

Zhivotovsky, L. A. (1999). Estimating population structure in diploids with multilocus dominant DNA markers. *Molecular Ecology* **8**, 907–913.

Zykin, D., Zhivotovsky, L. & Weir, B. S. (1995). Exact tests for association between alleles at arbitrary numbers of loci. *Genetica* **96**, 169–178.

https://doi.org/10.1017/S001667230100502X Published online by Cambridge University Press