Genet. Res., Camb. (2010), **92**, *pp.* 141–155.
doi:10.1017/S0016672310000121

A two-step method for detecting selection signatures using genetic markers

DANIEL GIANOLA[1,2,3]*, HENNER SIMIANER[3] AND SABER QANBARI[3]
[1] *Department of Animal Sciences and Department of Dairy Science, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA*
[2] *Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, N-1432 Ås, Norway*
[3] *Department of Animal Sciences, Georg-August-Universität, Göttingen, Germany*

(*Received 4 September 2009 and in revised form 12 December 2009 and 8 March 2010*)

Summary

A two-step procedure is presented for analysis of θ (F_{ST}) statistics obtained for a battery of loci, which eventually leads to a clustered structure of values. The first step uses a simple Bayesian model for drawing samples from posterior distributions of θ-parameters, but without constructing Markov chains. This step assigns a weakly informative prior to allelic frequencies and does not make any assumptions about evolutionary models. The second step regards samples from these posterior distributions as 'data' and fits a sequence of finite mixture models, with the aim of identifying clusters of θ-statistics. Hopefully, these would reflect different types of processes and would assist in interpreting results. Procedures are illustrated with hypothetical data, and with published allelic frequency data for type II diabetes in three human populations, and for 12 isozyme loci in 12 populations of the argan tree in Morocco.

1. Introduction

The discovery of a massive number of single nucleotide polymorphisms (SNPs) in the genome of several species has enabled exploration of genome-wide signatures of selection via an assessment of variation in marker allele frequencies among populations (e.g. Holsinger & Weir, 2009). Several methods have been proposed for doing this, such as site frequency spectrum, linkage disequilibrium and population differentiation (Sabeti *et al.*, 2006; Akey, 2009). Concerning population differentiation, a parameter $\theta = F_{ST}$, measuring relatedness between pairs of alleles within a sub-population relative to that in an entire population, has been used for this purpose (Wright, 1951; Cockerham, 1969; Weir & Hill, 2002); Lewontin & Krakauer (1973) and Robertson (1975) discuss related approaches. Equivalently, θ can be interpreted as a measure of dispersion of gene frequencies among groups relative to the variation expected in the population from which such groups derived. For example, Akey *et al.* (2002) analysed over 26 500 SNPs for which allele frequencies were available in three populations of humans. The θ parameter was estimated for every marker locus and the distribution of estimates over the entire genome, and by chromosome, was examined. By referring these estimates to their empirical genome-wide distribution, 174 candidate genes were identified as possible targets of selection.

Holsinger & Weir (2009) provide an account of the logic of the procedure. Briefly, given a set of loci in a given species, a reasonable assumption is that all share the same demographic history and patterns of migration. If these loci are neutral and have similar mutation rates, members of this set can be conceivably regarded as exchangeable realizations of the same evolutionary process. Loci showing departures from the resulting distribution may serve as flags of genomic regions that have been under the influence of selection. Under the hypothesis of selective neutrality, the distribution (over loci) of estimates of θ is expected to be driven by genetic drift, assumed to affect all loci in a similar fashion. On the other hand, when selection operates on one or several loci (as in a multifactorial model for complex traits), markers that are within genes or in nearby locations will display large or small values of θ, the latter occurring when some sort of balancing selection takes place (Cavalli-Sforza, 1966). This opens an avenue for identification of regions associated with population differentiation,

* Corresponding author. e-mail: gianola@ansci.wisc.edu

e.g., high versus low producing breeds of dairy cattle. Knowledge of such regions may be useful for enhancing the effectiveness of breeding programs via marker-assisted selection, or for tagging variants associated with disease or quantitative traits. While unusual values of θ may point to genomic locations where selection may have operated, there is arbitrariness with respect to characterizing the type of selection that might have occurred. Typically, loci are classified as either neutral, or subject to balancing selection (low values of θ), or favoured by selection within some specific population or environment (large population differentiation, thus leading to large values of θ). If the values of θ arise from different evolutionary or artificial (such as in plant and animal breeding) processes, one would expect to observe a mixture of distributions leading to clusters representing the different kinds of mechanisms operating. There is no apparent reason why there should be only two or three such clusters; there may be several clusters harbouring loci undergoing different types of selection processes. On the other hand, if θ values vary completely at random due to genetic drift, a single cluster is to be expected.

Statistical issues associated with inferring θ-statistics have been discussed, e.g., by Weir & Cockerham (1984) and Weir & Hill (2002), with emphasis on methods of moments estimation; by Balding (2003) using maximum likelihood for beta-binomial and Dirichlet-multinomial distributions; and by Holsinger (1999), Beaumont & Balding (2004) and Guo *et al.* (2009) employing Bayesian procedures. None of these treatments have addressed the possible existence of a clustered structure.

The objective of this paper is to present a two-step procedure eventually leading to clusters of θ values. The first step, along the lines of Holsinger (1999), Balding (2003) and Beaumont & Balding (2004), uses a simple Bayesian structure for drawing samples from the posterior distributions of θ-parameters, but without constructing Markov chains. This step assigns a weakly informative prior to allelic frequencies and does not make any assumptions about evolutionary models. The second step regards samples from these posterior distributions as 'data' and fits a sequence of finite mixture models, with the aim of identifying clusters of θ-statistics. Hopefully, these would reflect different types of processes and would assist in interpreting results.

The paper is organized as follows. Section 2 reviews basic concepts. In section 3, the first step of the procedure is presented, contrasted with maximum likelihood, and illustrated with a hypothetical dataset and with data on type II diabetes in three populations. Section 4 describes the second step of the procedure, and illustrates it with a dataset containing allelic frequencies for 12 polymorphic isozyme loci in 12 populations of the argan tree (*Argania spinosa* L. Skeels) of Morocco presented in Petit *et al.* (1998) and analysed by Holsinger (1999). The paper concludes with a discussion of the proposed methodology.

2. Background

(i) *Basic concepts*

The stage is set by reviewing essentials of a random effects treatment proposed by Cockerham (1969, 1973). Suppose that genetic markers (e.g. SNPs) are screened in a set of individuals in each of R groups or populations, the latter viewed as drawn at random from some conceptual hyper-population from which such groups derive. Consider a bi-allelic locus (developments carry to multiple alleles as well) and let A_l and a_l be the two alleles at locus l ($l=1, 2, \ldots, L$); define $p_l=\Pr(A_l)$ to be the true, unobserved, frequency of allele A_l in the hyper-population, with $1-p_l=\Pr(a_l)$ being the frequency of a_l. Cockerham (1969) defines a_l as any allele other than A_l and uses an indicator variable x to denote allelic state ('content'), such that

$$x_{rij,l}=\begin{cases}1, & \text{if an allele is } A_l,\\ 0, & \text{otherwise.}\end{cases}$$

Here, $r=1, 2, \ldots, R$ denotes group or replicate, i indicates an individual, j is an index for a within-individual deviation and $l=1, 2, \ldots, L$ is an indicator for locus. Even though $x_{rij,l}$ is a binary variable (so a linear model is questionable), Cockerham (1969) uses the linear decomposition

$$x_{rij,l}=p_l+a_{r,l}+b_{ri,l}+w_{rij,l}, \tag{1}$$

where p_l is as before and $a_{r,l}\sim(0, \sigma^2_{a,l})$, $b_{ri,l}\sim(0, \sigma^2_{b,l})$ and $w_{rij,l}\sim(0, \sigma^2_{w,l})$ are mutually uncorrelated zero-mean random deviates, specific to locus l; the σ^2's are variance components. Under the assumption that all alleles at locus l have the same marginal distribution,

$$E(x_{rij,l})=p_l$$

and

$$\mathrm{Var}(x_{rij,l})=p_l(1-p_l)=\sigma^2_{a,l}+\sigma^2_{b,l}+\sigma^2_{w,l}=\sigma^2_l$$

for $l=1, 2, \ldots, L$. Decomposition (1) induces the following covariance structure between allelic content variables:

$$\mathrm{Cov}(x_{rij,l}, x_{r'i'j',l})=\begin{cases}\sigma^2_l, & \text{if } r=r', i=i', j=j',\\ \sigma^2_a, & \text{if } r=r', i\neq i', j\neq j',\\ \sigma^2_{a,l}+\sigma^2_{b,l}, & \text{if } r=r', i=i', j\neq j',\\ \mathrm{Cov}(a_r, a_{r'}), & \text{if replicates are correlated somehow.}\end{cases}$$

https://doi.org/10.1017/S0016672310000121 Published online by Cambridge University Press

A standard assumption is $\mathrm{Cov}(a_r, a_{r'})=0$. The following correlations (all positive) follow.

- Pairs of alleles drawn at random from different individuals in the same group are correlated as

$$\rho_{a,l}=\frac{\sigma^2_{a,l}}{\sigma^2_{a,l}+\sigma^2_{b,l}+\sigma^2_{w,l}}=\theta_l=F_{\mathrm{ST},l}, \tag{2}$$

  so $0\leqslant\theta_l\leqslant 1$ for all l.
- Pairs of alleles drawn within individuals over all replicates bear a correlation equal to

$$\rho_{ab,l}=\frac{\sigma^2_{a,l}+\sigma^2_{b,l}}{\sigma^2_{a,l}+\sigma^2_{b,l}+\sigma^2_{w,l}}=F_l=F_{\mathrm{IT},l},$$

  where F is the total inbreeding coefficient, also known as F_{IT} (e.g. Weir & Hill, 2002).
- The correlation between alleles within individuals within the same replicate is

$$\rho_{b,l}=\frac{\sigma^2_{b,l}}{\sigma^2_{b,l}+\sigma^2_{w,l}}=f_l=F_{\mathrm{IS},l},$$

  which is the within sub-population inbreeding coefficient f.

It is easy to show that

$$\theta_l=\frac{F_{\mathrm{IT},l}-F_{\mathrm{IS},l}}{1-F_{\mathrm{IS},l}}=F_{\mathrm{ST},l}.$$

This expression satisfies

$$1-F_{\mathrm{IT},l}=(1-F_{\mathrm{IS},l})(1-F_{\mathrm{ST},l}),$$

indicating that a reduction in heterozygosity has two sources: one that is due to population sub-division or Wahlund's effect, $(1-F_{\mathrm{ST},l})$, and a reduction within subpopulation or group caused by 'local' inbreeding, $(1-F_{\mathrm{IS},l})$.

Note that parameter F_{ST} given in (2) can also be written as

$$\theta_l=\frac{\sigma^2_{a,l}}{\sigma^2_{a,l}+\sigma^2_{b,l}+\sigma^2_{w,l}}=\frac{\sigma^2_{a,l}}{p_l(1-p_l)}.$$

In Cockerham (1969), the variance $\sigma^2_{a,l}$ arises by drawing alleles from a random sample of populations. Under conceptual repeated sampling, this generates a distribution having such variance. However, in many applications, the R groups under study are targeted (as opposed to randomly sampled) populations, e.g., Myles *et al.* (2007); this defines a 'fixed effects' model. Now, since $\sigma^2_{a,l}$ is the between-group variance in allelic content as per model (1), an alternative parametric definition of θ_l in terms of the unknown gene frequencies of the R groups is

$$\theta_l=\frac{\frac{\sum_{r=1}^{R}(p_{r,l}-\bar{p}_l)^2}{R}}{\bar{p}_l(1-\bar{p}_l)}, \tag{3}$$

where $\bar{p}_l=\sum_{r=1}^{R}p_{r,l}/R$ is the average (over groups) of the frequencies of allele A_l at locus l. Note that $\bar{p}_l$ is taken as an unweighted average; it does not seem sensible to express a parameter in terms of sample size (unless weights assigned to samples reflect true population sizes). Expressing θ_l explicitly in terms of the locus-specific gene frequencies yields

$$\theta_l=\frac{\sum_{r=1}^{R}p^2_{r,l}-\frac{\left(\sum_{r=1}^{R}p_{r,l}\right)^2}{R}}{\left(\sum_{r=1}^{R}p_{r,l}-\frac{\left(\sum_{r=1}^{R}p_{r,l}\right)^2}{R}\right)}, \tag{4}$$

providing a mapping from the joint space of R allelic frequencies to the single-dimensional space of θ_l, which resides in (0, 1). If allelic frequencies for the different loci are driven from the same stochastic evolutionary process (e.g. as generated by random drift), this defines a distribution of values of θ expected under neutrality assumptions. From a Bayesian perspective, every unknown is a random variable and, since allelic frequencies are unknown, θ as given in (3) will posses a distribution, both *a priori* and *a posteriori*. In the first step of the method proposed in this paper, the posterior distribution of θ_l will result from assigning a vague prior to all allelic frequencies, corresponding in some sense to what could be termed as a fixed effects treatment from a frequentist perspective. The second step addresses the question of whether or not all θ_l stem from the same distribution or from different distributions resulting from heterogeneity of the underlying stochastic processes. This makes the treatment proposed here different from those in, e.g., Holsinger (1999) or Balding (2003).

3. Estimation of parameters

(i) *Inferring gene frequencies*

Gene frequencies can be inferred using a simple Bayesian approach. Suppose that n_r individuals are genotyped in population r, so that the number of alleles screened at locus l is $2n_r=n_{r,Al}+n_{r,al}$, where $n_{r,Al}$ and $n_{r,al}$ are the observed numbers of copies of A_l and a_l, respectively.

A convenient assumption is that of mutual independence between the distributions of alleles at different loci (stronger than that of pairwise linkage equilibrium). Linkage disequilibrium is pervasive but the assumption made above facilitates matters and is widely used, e.g., by Corander *et al.* (2003). Let $\mathbf{p}=(\mathbf{p_1}, \mathbf{p_2}, \ldots, \mathbf{p}_R)'$ be an $RL\times 1$ vector of allelic frequencies for all R groups, where $\mathbf{p}_r=(p_{r,1},p_{r,2}, \ldots, p_{r,L})'$ has order $L\times 1$. Under the mutual independence assumption, the likelihood conferred by the observed number of copies of alleles to the gene frequencies is

$$l(\mathbf{p}|\mathrm{DATA})=\prod_{r=1}^{R}\prod_{l=1}^{L}p_{r,l}^{n_{r,A_l}}(1-p_{r,l})^{n_{r,a_l}}. \tag{5}$$

https://doi.org/10.1017/S0016672310000121 Published online by Cambridge University Press

The maximum likelihood estimator of $p_{r,l}$ is $\hat{p}_{r,l}=n_{r,A_l}/2n_r$ and its empirical variance is $\widehat{\mathrm{Var}}(\hat{p}_{r,l})=\hat{p}_{r,l}(1-\hat{p}_{r,l})/2n_r$. The maximum likelihood estimator is unbiased but unstable, and may take values at the boundaries of the parameter space in small samples.

In a Bayesian treatment, allelic frequencies are assigned a prior distribution that might be homogeneous or heterogeneous over populations, chromosomes or genomic regions (e.g. coding versus non-coding regions). For example, Holsinger (1999, 2006) adopts a prior beta distribution, $Beta\,(p_l|a_l, b_l)$ (and interprets it as describing variation over populations) with parameters

$$a_l=\frac{1-\theta}{\theta}x_l$$

and

$$b_l=\frac{1-\theta}{\theta}(1-x_l).$$

Here θ is common to all loci (i.e. the hypothesis of neutrality) and x_l is the mean allelic frequency at locus l (averaged over populations). Using properties of the beta distribution in the parametric definition of θ leads to

$$\frac{\mathrm{Var}(p_l)}{E(p_l)[1-E(p_l)]}=\frac{\frac{a_l b_l}{(a_l+b_l)^2(a_l+b_l+1)}}{\frac{a_l}{a_l+b_l}\cdot\frac{b_l}{a_l+b_l}}=\theta.$$

Then, the joint posterior distribution of all unknowns (allelic frequencies, θ and vector $\mathbf{x}=\{x_l\}$) is

$$g(\mathbf{p},\theta,\mathbf{x}|\mathrm{DATA})$$
$$\propto\prod_{r=1}^{R}\prod_{l=1}^{L}p_{r,l}^{n_{r,A_l}+((1-\theta)/\theta)x_l-1}$$
$$(1-p_{r,l})^{n_{r,a_l}+((1-\theta)/\theta)(1-x_l)-1}g(\theta)g(\mathbf{x}).$$

Holsinger (1999) took $g(\theta)=\mathrm{Beta}(1, 2)$ distribution as prior for θ, and assumed that all x_l were identically distributed according to the uniform process $g(x_l)=U(0, 1)$. Given θ and $\mathbf{x}$, the allelic frequencies are conditionally independent with conditional posterior distributions:

$$g(p_{r,l}|\mathrm{ELSE})=\mathrm{Beta}\left(n_{r,A_l}+\frac{1-\theta}{\theta}x_l, n_{r,a_l}+\frac{1-\theta}{\theta}(1-x_l)\right),$$
$$r=1,\ 2,\ \ldots,\ R\quad l=1,\ 2,\ \ldots,\ L,$$

where ELSE means all parameters other than $p_{r,l}$ and the data observed. However, the conditional posterior distributions of θ and $\mathbf{x}$ are not recognizable, so an elaborate sampling scheme, e.g., one based on Markov chain Monte Carlo (MCMC) methods, must be tailored. Holsinger (1999) found that inferences were insensitive with respect to the choices of beta and uniform prior distributions for θ and elements of $\mathbf{x}$, respectively. However, it was assumed (as in a neutral model) that all loci share the same θ parameter. This produces a mutual borrowing of information among loci (shrinking $p_{r,l}$ towards a common value), but the procedures are not explicit with respect to the existence of heterogeneity over loci due to forces such as differential mutation or selective sweeps. As proposed by Beaumont & Balding (2004), one could estimate locus specific θ values and refer these estimates to the posterior distribution of θ under the homogeneity value. In this manner, outliers could be found with respect to the 'neutral' distribution, but this would not inform about the structure of any latent heterogeneity.

Here, an alternative approach is used. Jeffreys rule (Bernardo & Smith, 1994; Sorensen & Gianola, 2002) is used to produce a reference prior, which is a $\mathrm{Beta}\left(\frac{1}{2},\frac{1}{2}\right)$ distribution assigned to all loci in all populations. This reference prior distribution is minimally informative in a well defined sense (Bernardo and Smith, 1994). Using Bayes theorem, the joint posterior density of all allelic frequencies is now

$$g(\mathbf{p}|\mathrm{DATA})\propto\prod_{r=1}^{R}\prod_{l=1}^{L}p_{r,l}^{n_{r,A_l}+(1/2)-1}(1-p_{r,l})^{n_{r,a_l}+(1/2)-1}$$
$$=\prod_{r=1}^{R}\prod_{l=1}^{L}\mathrm{Beta}\left(n_{r,A_l}+\tfrac{1}{2}, n_{r,a_l}+\tfrac{1}{2}\right). \qquad (6)$$

Thus, allelic frequencies at different loci are mutually independent, a posteriori, with $p_{r,l}$ following a beta distribution with parameters $\alpha_{rl}=n_{r,A_L}+\frac{1}{2}$ and $\beta_{rl}=n_{r,a_l}+\frac{1}{2}$. Possible point estimates of allelic frequencies are the posterior mean

$$\overleftrightarrow{p}_{r,l}=\frac{n_{r,A_l}+\frac{1}{2}}{2n_r+1}, \qquad (7)$$

and the posterior mode

$$\tilde{p}_{r,l}=\frac{n_{r,A_l}-\frac{1}{2}}{2n_r-1},\quad \text{for } n_{r,A_l}\geqslant 1. \qquad (8)$$

The variance of the posterior distribution of $p_{r,l}$ is

$$\mathrm{Var}(p_{r,l}|\mathrm{DATA})=\frac{\left(n_{r,A_l}+\frac{1}{2}\right)\left(n_{r,a_l}+\frac{1}{2}\right)}{(2n_r+1)^2(2n_r+2)}. \qquad (9)$$

Even though a weakly informative prior is used, differences exist with respect to maximum likelihood. To illustrate this point, consider a hypothetical example with two groups, M and N. Suppose that 100 individuals are genotyped in group M and that the observed number of A_l alleles is 199, i.e. the locus is nearly fixed. The maximum likelihood estimate of $p_{M,l}$ is 0·995 and its estimated standard error is $4{\cdot}99\times10^{-3}$; a calculation based on asymptotic normality (without truncation) yields that the probability of obtaining estimates larger than 1 is close

https://doi.org/10.1017/S0016672310000121 Published online by Cambridge University Press

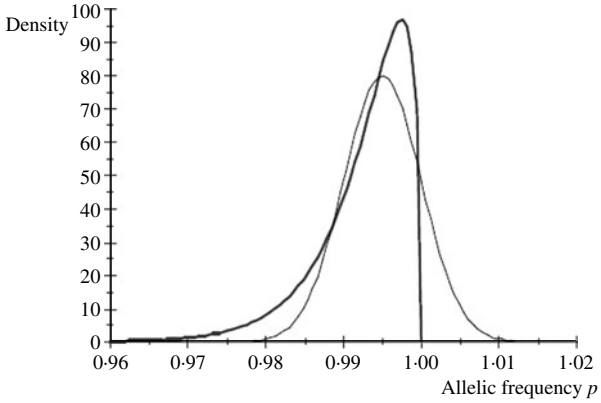

Fig. 1. Posterior density (thick line) of the allelic frequency p at a locus for which 199 copies have been observed out of 200 alleles counted in hypothetical population M; the posterior distribution is Beta$\left(199+\frac{1}{2}, 1+\frac{1}{2}\right)$. The thin line is the density of a normal approximation to the sampling distribution of the maximum likelihood estimator.

to 0·16! Further, the probability of obtaining estimates between 0·9 and 0·995 is close to $\frac{1}{2}$. On the other hand, the posterior distribution of $p_{M,l}$ is Beta$\left(199+\frac{1}{2}, 1+\frac{1}{2}\right)$. The posterior mean and posterior standard deviation are 0·993 (note some shrinkage away from the edge of the parameter space) and $6{\cdot}06\times10^{-3}$, respectively; the posterior probability of the frequency being larger than 1 is exactly zero, and the probability that $p_{M,l}$ takes values between 0·9 and 0·995 is about 0·57. Figure 1 displays the posterior distribution of the allelic frequency obtained with Jeffreys prior, overlaid against the normal approximation to the distribution of the maximum likelihood estimates. Clearly, the approach used makes a difference, even in a situation where allelic frequencies are estimated with reasonable precision, as indicated by the small standard error of the maximum likelihood estimate and the small posterior standard deviation in the Bayesian analysis (the coefficient of variation of the posterior distribution is less than 1%).

In the second population, N, 30 individuals are genotyped and 10 alleles are of the type A_l; the maximum likelihood estimate of $p_{N,l}$ is then $\frac{1}{6}$, much lower than in M, and its sampling variance is $2{\cdot}31\times10^{-3}$. The posterior distribution of $p_{N,l}$ is Beta$\left(10+\frac{1}{2},\ 50+\frac{1}{2}\right)$. In N, the posterior density of $p_{N,l}$ and the normal approximation to the density of the distribution of the maximum likelihood estimator are very similar (not shown here).

Differences in allelic frequencies between populations M and N at the locus in question may be due to random drift or may suggest a signature of selection.

(ii) *Inferring* θ *by maximum likelihood*

A likelihood-based estimate of θ can be obtained by replacing in (3) or (4) the unknown allelic frequencies by their maximum likelihood estimates. For the example of populations M and N above, the estimate is

$$\hat{\theta}_l=\frac{\sum_{r=1}^{2}(\hat{p}_{r,l}-\bar{\hat{p}}_l)^2}{2\bar{\hat{p}}_l(1-\bar{\hat{p}}_l)}\approx 0{\cdot}7046.$$

The sampling variance of the maximum likelihood estimator of θ_l can be approximated using a Taylor series expansion. As shown in Appendix A, the first derivative of θ_l with respect to the allelic frequency at locus l in group r is

$$\frac{\partial}{\partial p_{r,l}}\theta_l=\left[\frac{2(p_{r,l}-\bar{p}_l)}{\overline{p_l^2}-\bar{p}_l^2}-\frac{(1-2\bar{p}_l)}{\bar{p}_l(1-\bar{p}_l)}\right]\frac{\theta_l}{R},$$

for $r=1,2,\ldots,R$; $l=1,2,\ldots,L$, where $\bar{p}_l=\sum_{r=1}^{R}p_{r,l}/R$ is as before and $\bar{p}_l^2=\sum_{r=1}^{R}p_{r,l}^2/R$. Further, let

$$\hat{\nabla}=\left\{\frac{\partial}{\partial p_{r,l}}\theta_l\right\}_{p_{r,l}=\hat{p}_{r,l}},$$

be an $RL\times1$ vector of first derivatives evaluated at the maximum likelihood estimates of the allelic frequencies. Then, approximately

$$\begin{aligned}\widehat{\mathrm{Var}}(\hat{\theta}_l)&\approx\hat{\nabla}'\,\widehat{\mathrm{Var}}(\hat{\mathbf{p}})\hat{\nabla}\\&=\sum_{r=1}^{R}\left\{\left[\frac{2(p_{r,l}-\bar{p}_l)}{\overline{p_l^2}-\bar{p}_l^2}-\frac{(1-2\bar{p}_l)}{\bar{p}_l(1-\bar{p}_l)}\right]\frac{\theta_l}{R}\right\}^2_{p_{r,l}=\hat{p}_{r,l}}\\&\quad\times\frac{\hat{p}_{r,l}(1-\hat{p}_{r,l})}{2n_r},\end{aligned}$$

where $\widehat{\mathrm{Var}}(\hat{\mathbf{p}})=\mathrm{Diag}(\hat{p}_{r,l}(1-\hat{p}_{r,l})/2n_r)$ is a diagonal matrix containing the estimates of the sampling variances of the maximum likelihood estimates of allelic frequencies $p_{r,l}$. For the hypothetical example, $\widehat{\mathrm{Var}}(\hat{\theta}_l)\approx9{\cdot}8265\times10^{-5}$. The asymptotic normal approximation to the distribution of the estimates assigns nil probability to 'estimates' outside of (0,1); the probability of obtaining estimates of θ between 0·67 and 0·74 for this two-population situation is 0·9996.

(iii) *Bayesian inference of* θ

Consider now finding the posterior distribution of θ_l as defined in (4) and without making the assumption that the θs are realizations from the same stochastic process, i.e. without borrowing information across loci over and above the shrinkage of allelic frequencies produced by Jeffreys prior. The posterior distribution is analytically difficult to arrive at because θ_l is a non-linear function of gene frequencies in all R groups. However, since it is easy to obtain independent samples from each of the Beta$(n_{r,A_l}+\frac{1}{2}$,

https://doi.org/10.1017/S0016672310000121 Published online by Cambridge University Press

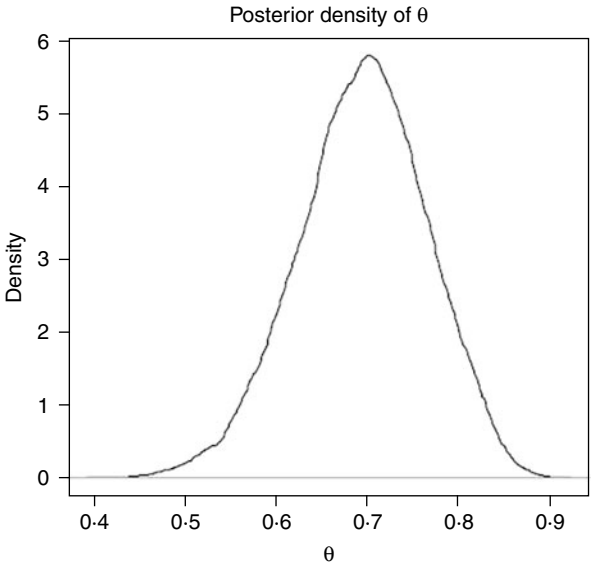

Fig. 2. Posterior density of θ_l for the hypothetical example of populations M and N.

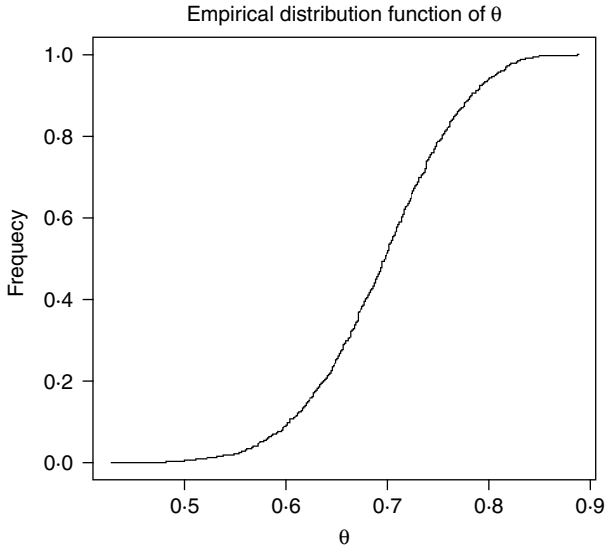

Fig. 3. Empirical cumulative distribution function of θ_l for the hypothetical example of populations M and N.

$n_{r,a_l}+\frac{1}{2}$) processes, Monte Carlo estimates of features of the posterior distribution of θ_l can be obtained without using MCMC methods at all. Let $p_{r,}^{(s)}$, $s=1,2,\ldots,S$, be samples from the posterior (beta) distribution of $p_{r,l}$, the frequency of allele A_l at locus l. Then, a draw from the posterior distribution of θ_l is given by

$$\theta_l^{(s)}=\frac{\sum_{r=1}^{R}(p_{r,l}^{(s)})^2-\frac{\left(\sum_{r=1}^{R}p_{r,l}^{(s)}\right)^2}{R}}{\left(\frac{R\sum_{r=1}^{R}p_{r,l}^{(s)}-\left(\sum_{r=1}^{R}p_{r,l}^{(s)}\right)^2}{R}\right)}, \tag{10}$$

which is a random variable with support in (0,1) (Holsinger, 2006). Then, from S samples, the mean, median, variance, etc., of the posterior distribution of θ_l can be estimated. Each θ_l ($l=1,2,\ldots,L$) will have a point estimate and an assessment of uncertainty, e.g., a credibility interval of size 95% given by the 2·5% and 97·5% percentiles of the corresponding posterior distribution estimated either from samples or from the normal theory approximation given in Appendix B.

In the hypothetical populations M and N, the posterior distributions of the frequency of A_l are Beta(199·5,1·5) and Beta(10·5,50·5), respectively. With draws denoted as $B^{(s)}$ (.,.), S samples from the posterior distribution of θ_l can be obtained as

$$\theta_l^{(s)}=\frac{[B^{(s)}(199{\cdot}5,1{\cdot}5)]^2+[B^{(s)}(10{\cdot}5,50{\cdot}5)]^2-\frac{\{[B^{(s)}(199{\cdot}5,1{\cdot}5)]+[B^{(s)}(10{\cdot}5,50{\cdot}5)]\}^2}{2}}{[B^{(s)}(199{\cdot}5,1{\cdot}5)]+[B^{(s)}(10{\cdot}5,50{\cdot}5)]-\frac{\{[B^{(s)}(199{\cdot}5,1{\cdot}5)]+[B^{(s)}(10{\cdot}5,50{\cdot}5)]\}^2}{2}},\quad s=1,2,\ldots,S.$$

To illustrate, 5000 samples were drawn from each of the two beta distributions, to form $S=5000$ corresponding draws from the posterior distribution of θ_l. The mean and median were 0·6966 and 0·6972, respectively; the standard deviation was 0·070 and the range of values samples spanned from 0·4268 to 0·8883. The posterior density of θ_l and the empirical cumulative distribution function are in Figs 2 and 3, respectively. Values of θ_l appearing with appreciable density range from about 0·5 to 0·9 (Fig. 2), with small posterior probability assigned to values smaller than 0·6 (Fig. 3).

(iv) *A Bayesian 'null' distribution for assessing sampling variation uncertainty*

It is important to check whether or not posterior estimates of θ_l depart from what would be expected by chance alone. A posterior distribution consistent with expectations under a 'null' model is formulated next. The θ_l statistics calculated from the 'full' model above can then be referred to this null distribution. Note that the 'null' distribution given below describes the uncertainty to be expected from drawing random samples from the same population, but not the variability to be expected due to genetic drift. If estimates of θ_l fall in this null distribution, this would indicate that the study lacks power to answer evolutionary questions in any meaningful manner.

https://doi.org/10.1017/S0016672310000121 Published online by Cambridge University Press

A 'null' distribution is arrived at by stating that $p_{r,l}=p_l$ is the same random variable for all R populations. Under this assumption, the posterior distribution of the vector of gene frequencies (now of dimension $L\times 1$) under the 'null' model is

$$g(\mathbf{p}|\text{DATA, Null}) \propto \left[\prod_{r=1}^{R}\prod_{l=1}^{L} p_l^{n_{r,A_l}}(1-p_l)^{n_{r,a_l}}\right]\prod_{l=1}^{R} p_l^{\frac{1}{2}-1}(1-p_l)^{\frac{1}{2}-1} = \prod_{l=1}^{L}\text{Beta}\left(\sum_{r=1}^{R} n_{r,A_l}+\frac{1}{2},\ \sum_{r=1}^{R} n_{r,a_l}+\frac{1}{2}\right). \quad (11)$$

Hence, allelic frequencies p_l are mutually independent, a posteriori, with $p_l|$DATA, Null being a beta distribution with parameters $\alpha_l=\sum_{r=1}^{R} n_{r,A_l}+\frac{1}{2}$ and $\beta_r=\sum_{r=1}^{R} n_{r,a_l}+\frac{1}{2}$. A draw from the posterior distribution of the F_{ST} statistic under this model takes the form

$$\theta_{l,\text{Null}}^{(s)}=\frac{\frac{\sum_{r=1}^{R}(p_l^{(r,s)}-\bar{p}_l^{(s)})^2}{R}}{\bar{p}_l^{(s)}(1-\bar{p}_l^{(s)})}, \quad (12)$$

where $p_l^{(r,s)}$ is a draw from $\text{Beta}\left(\sum_{r=1}^{R} n_{r,A_l}+\frac{1}{2}, \sum_{r=1}^{R} n_{r,a_l}+\frac{1}{2}\right)$, with R such draws involved in a realization of $\theta_l^{(s)}$, and $\bar{p}_l^{(s)}$ is the average of the R draws. A set of samples from the posterior distribution of θ_l under the null model is obtained by repeating the sampling process S times. This distribution serves as a reference against which the θ_l statistics calculated from the 'full' model can be compared. If the posterior mean of θ_l obtained from the 'full' model falls outside of a high density area of the posterior distribution of θ in the null model, then the divergence between populations would be probably due to drift or selection (assuming mutation rates are constant over populations), but not due to chance alone.

For the example of populations M and N, $\sum_{r=1}^{R} n_{r,A_l}=209$ and $\sum_{r=1}^{R} n_{r,a_l}=51$. Figure 4 depicts the Beta(209·5,51·5) distribution of the allelic frequency under the 'null' model. Note that the maximum likelihood estimates of the allelic frequencies in the M and N populations, of 0·995 and $\frac{1}{6}$, respectively, are not assigned any appreciable density under this model. Upon drawing 5000 independent samples from the beta distribution of the allelic frequency under the null model, 5000 draws for $\theta_{l,\text{Null}}^{(s)}$ were obtained by evaluating (12) for each of the samples. Draws ranged from $8{\cdot}24\times 10^{-13}$ to 0·0503; the mean (standard deviation) was 0·0038(0·0053) and the posterior median was 0·0017. The posterior density of $\theta_{l,\text{Null}}$ was very sharp as shown in Fig. 5. In the full model, the estimated posterior mean (standard deviation) of θ_l was 0·6966, which is unlikely to have been generated under the null distribution. This would make the locus a reasonable candidate for further examination.

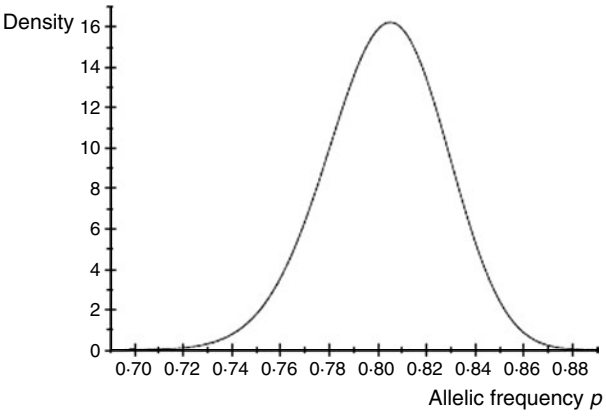

Fig. 4. Posterior density of the allelic frequency p under the 'null' model for hypothetical populations M and N: 209 copies of A_l are observed out of 260 alleles screened.

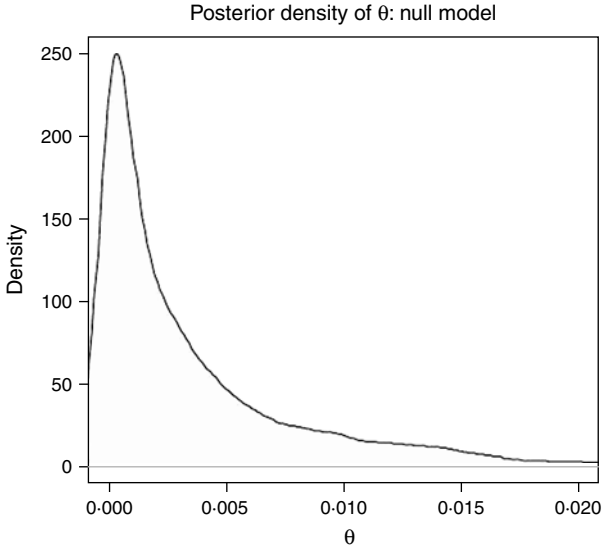

Fig. 5. Posterior density of θ_l under the null model for the hypothetical example of populations M and N.

(v) *Illustration of sampling variation with candidate genes for type-II diabetes*

The Bayesian method was applied to data pertaining to identification of candidate gene variants for type II diabetes in Polynesians (Myles *et al.*, 2007). Prevalence of this disease is high in several Pacific populations, e.g. 40% of adults living in the island of Nauru. DNA samples were obtained from 23 Polynesians, 23 New Guineans and 19 Han Chinese from Beijing. Type II diabetes-associated alleles were from 10 SNP loci having evidence of association. Estimated frequencies and θ_l statistics are shown in page 587 of Myles *et al.* (2007). To illustrate the Bayesian procedure, data for the KCNJ11 locus was used, and

https://doi.org/10.1017/S0016672310000121 Published online by Cambridge University Press

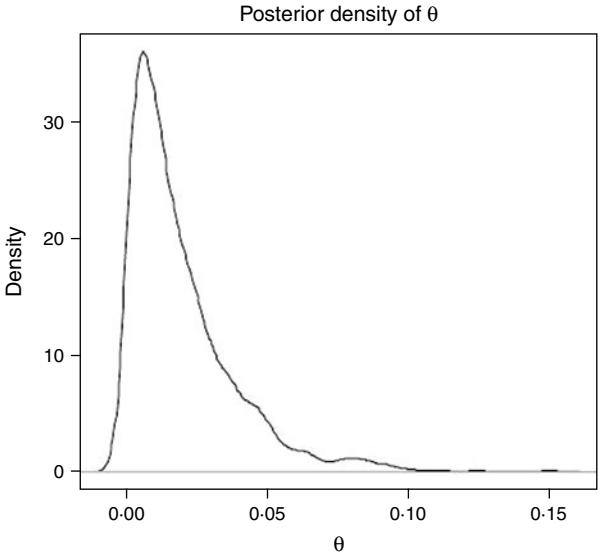

Fig. 6. Density of the posterior distribution of θ_{KCNJ11} obtained from allelic frequencies in Myles *et al.* (2007).

susceptibility allele frequencies (A_l in our notation) were 0·30, 0·25 and 0·34 in the three populations above, respectively. Their figures do not lead to an integer number of alleles, due to rounding error, so the number of observed A_l alleles used here was set to 14 (Polynesians), 12 (New Guineans) and 13 (Han Chinese). Myles *et al.* (2007) employed an 'unbiased estimator' of θ_l for calculating population pairwise differences, and their estimates were 0·003 (New Guinea–China), −0·024 (China–Polynesia) and −0·017 (New Guinea–Polynesia). Note the two negative estimates of a parameter that resides in (0,1); standard errors or significance levels were not provided. Their analysis suggests that this locus is not associated with prevalence of the disease.

The posterior distributions of A_l were $\mathrm{Beta}_{\mathrm{Polynesians}}(14{\cdot}5,32{\cdot}5)$, $\mathrm{Beta}_{\mathrm{NewGuineans}}(12{\cdot}5,34{\cdot}5)$ and $\mathrm{Beta}_{\mathrm{HanChinese}}(13{\cdot}5,25{\cdot}5)$. The number of samples drawn from each of these 3 posterior distributions was $S=1000$, and 1000 draws from the posterior distribution of θ_{KCNJ11} were obtained by evaluation of (10). Values of θ_{KCNJ11} ranged from $2{\cdot}423\times10^{-5}$ to 0·1500, with an estimated posterior mean (standard deviation) of 0·019 (0·0193); this estimate is higher than that of Myles *et al.* (2007). The non-parametric estimate of the posterior density of θ_{KCNJ11} is shown in Fig. 6, illustrating that the true value of the F_{ST} parameter is very likely below 0·10. The posterior inter-correlation structure between allelic frequencies and θ_{KCNJ11} in the full model was examined and, as expected, draws from the posterior distributions of allelic frequencies in the three populations were uncorrelated. Samples of θ_{KCNJ11} were positively correlated (0·55) with those for allelic frequency in Chinese Han, and the 95% confidence interval for the correlation was 0·51–0·60. However, draws for θ_{KCNJ11} were negatively correlated with allele frequencies in Polynesians (−0·07) and New Guineans (−0·39); the confidence intervals for these two correlations were (−0·13, −0·01) and (−0·44, −0·34), respectively.

For the 'null' model, the 1000 samples from the posterior distribution of $\theta_{\mathrm{KCNJ11,Null}}$ ranged from $3{\cdot}62\times10^{-6}$ to 0·1460, with the posterior mean (standard deviation) estimated at 0·002 (0·002); the posterior median was 0·002 as well. The posterior mean (standard deviation) estimate of θ_{KCNJ11} under the 'full' model was 0·019, and it did not enter with high density in the 'null' model (not shown). Although variation in allelic frequency at locus KCNJ11 among the three populations departs from what would be expected from chance alone (statistical sampling), the observed θ value is very small. This may support the hypothesis that this locus may not be associated with differences in prevalence of type II diabetes, in agreement with Myles *et al.* (2007). Allelic frequencies were uncorrelated, as it should be, given that the three replicates were drawn from the same Beta(209·5,51·5) distribution. The $\theta_{\mathrm{KCNJ11,Null}}$ statistic was uncorrelated with allelic frequencies, and the correlations were −0·08, −0·11 and 0·03 in the three replicates, with all confidence intervals including 0.

4. Clustering of θ-parameters

The second step of the procedure consists of clustering a set of estimates of θ values (in this case, posterior means) from a multi-locus analysis into data driven groups. The expectation is that these clusters might be representative of different processes taking place in the populations such as balancing or directional selection, neutrality or anything else.

The method is illustrated with data from a study of Petit *et al.* (1998) in which alleles were sampled for 12 isozyme loci of the *Argania* genus tree in each of 12 areas (populations) of Morocco. The data, given in page 847 of Petit *et al.* (1998), were modified as shown in Table 1. The modification consisted of treating all loci as bi-allelic by lumping alleles for loci with more than two variants into two classes. The number of individuals sampled per population ranged between 20 and 50, and the number of alleles per locus varied originally between 2 and 5. Note that, at some loci, one of the alleles was fixed in almost all populations. For example, for locus 3, the only population in which segregation was observed was TA.

For each locus, 2000 samples were drawn from the beta posterior distributions of allelic frequencies. For example, the posterior distribution of $p_{AB,1}$ was Beta(21·5,19·5). From these samples, 2000 draws from the posterior distribution of θ for each locus

https://doi.org/10.1017/S0016672310000121 Published online by Cambridge University Press

Table 1. *Allelic frequencies at 12 isozyme loci in each of 12 Argan tree populations, adapted from Petit* et al. *(1998) by making all loci bi-allelic. A1–A12 represent frequencies of the 'A' allele at loci 1–12; No. A1–No. A12 are the observed number of copies of the alleles. The number of 'a' alleles can be calculated from the number of individuals samples and the number of 'A' alleles observed*

Population	AB	AD	AR	BS	GO	MI	OG	S1	TA	TE	TM	TT
No. Individuals	20	40	20	30	32	20	30	20	30	20	20	50
A1	0·525	0·512	0·475	0·467	0·047	0·475	0·517	0·575	0·517	0·425	0·55	0·52
No. A1	21	41	19	28	3	19	31	23	31	17	22	52
A2	0·4	0·438	0·55	0·917	0·688	0·525	0·467	0·825	0·483	0·925	0·475	0·51
No. A2	16	35	22	55	44	21	28	33	29	37	19	51
A3	1	1	1	0	1	1	1	1	0·75	1	1	1
No. A3	40	80	40	0	64	40	60	40	45	40	40	100
A4	0·525	0·375	0·45	0·517	0·922	0·525	1	0·7	0·467	0·575	0·5	0·52
No. A4	21	30	18	31	59	21	60	28	28	23	20	52
A5	0·475	0·463	0·475	1	1	1	1	1	0·817	1	1	0·51
No. A5	19	37	19	60	64	40	60	40	49	40	40	51
A6	0·85	0·538	0·9	0·533	0·922	0·575	0·55	0·75	0·517	0·525	0·55	0·53
No. A6	34	43	36	32	59	23	33	30	31	21	22	53
A7	1	1	1	0·567	0·922	0·9	1	1	0·967	1	1	1
No. A7	40	80	40	34	59	36	60	40	58	40	40	100
A8	1	1	1	1	1	1	1	1	1	1	0·575	0·97
No. A8	40	80	40	60	64	40	60	40	60	40	23	97
A9	1	0·937	1	1	0·312	1	1	1	1	1	1	1
No. A9	40	75	40	60	20	40	60	40	60	40	40	100
A10	0·925	0·5	0·525	0·625	0·475	0·5	0·55	0·4	0·575	0·5	0·475	0·5
No. A10	37	40	21	38	30	20	33	16	35	20	19	50
A11	0·6	0·7	0·575	0·5	0·6	0·525	1	0·375	0·625	0·475	0·55	0·47
No. A11	24	56	23	30	38	21	60	15	38	19	22	47
A12	1	1	0·85	0·6	0·875	0·775	1	0·875	1	1	1	0·87
No. A12	40	80	34	36	56	31	60	35	60	40	40	87

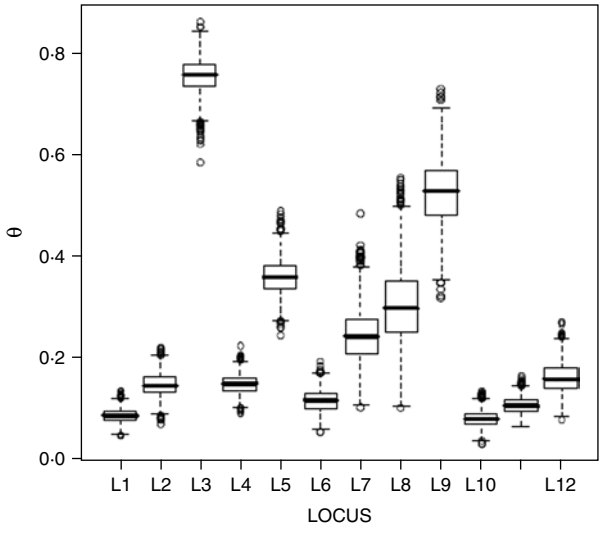

Fig. 7. Box plot of the posterior distributions of θ-parameters in 12 isozyme loci of the argan tree in Morocco (data originally from Petit *et al.* 1998).

were formed as in (10). The posterior means were as follows:

θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{10}	θ_{11}	θ_{12}
0·098	0·168	0·791	0·166	0·393	0·137	0·299	0·382	0·593	0·095	0·122	0·190

so estimates of θ varied over loci from about 0·095 (locus 10) to 0·791 (locus 3); all these estimates did not enter into the corresponding 'null' distributions. Box plots of the posterior distributions of the θ parameters are in Fig. 7. Visually, it is tempting to suggest four clusters: the first one would include locus 3, with the posterior mean of θ close to 0·79; the second cluster would include locus 9, with an estimate of θ of 0·59. The third cluster would include loci 5, 7 and 8 with estimates ranging between 0·30 and 0·39, and the fourth cluster would be represented by loci 1, 2, 4, 6, 10, 11 and 12 having the lowest estimates of θ.

The existence of an underlying structure is suggested by the distribution of all 24 000 samples, presented in Fig. 8. In the left panel, a non-parametric density estimate was obtained from these samples treated as if all draws (2000 for each of the 12 loci) had been made from the same process; the densities in the middle and right panels correspond to the logit, i.e., $\log(\theta/(1-\theta))$, and Gompit, $-\log(-\log(\theta))$, transforms of the sampled θ values, respectively. The three densities suggest that θ values cluster around 3, perhaps 4, modes.

https://doi.org/10.1017/S0016672310000121 Published online by Cambridge University Press

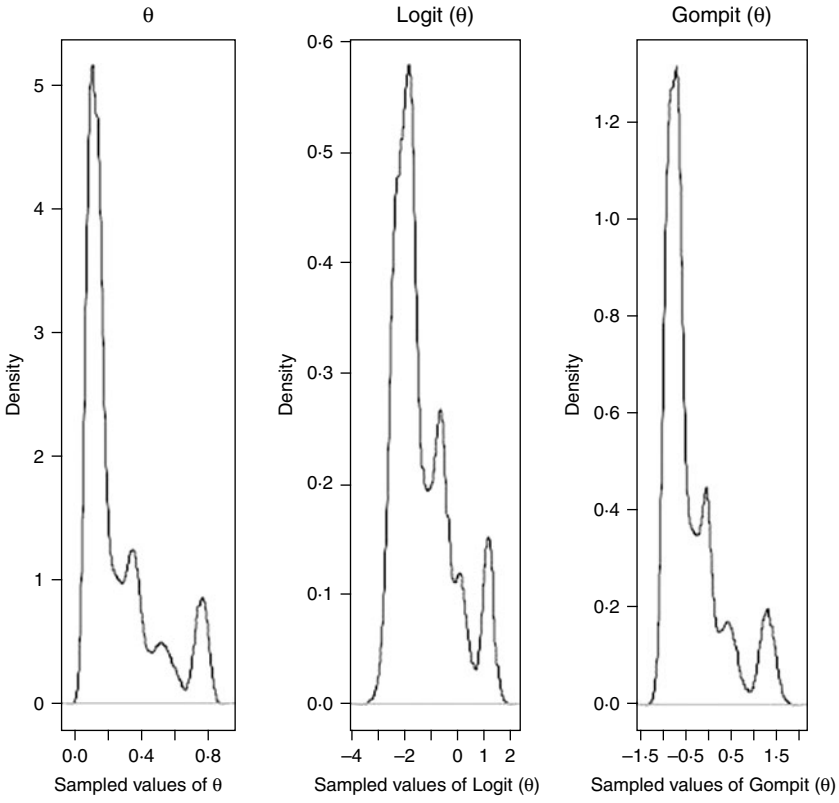

Fig. 8. Non-parametric density estimates of θ values (based on 2000 sample for each of 12 loci), logit(θ) and Gompit(θ). All samples treated as homogeneous, i.e. as generated from the same stochastic process.

The structure was explored more formally by fitting a sequence of finite mixture models to the means of the posterior distribution of the θ values for each of the 12 loci. These posterior means are independent (under the assumptions made for the allelic frequency models), but not identically distributed, since they are estimated with different precision, due to unequal numbers of individuals sampled and varying allelic frequencies. The distributions of θ values among loci are not normal (the logit and Gompit transforms would be expected to be more nearly so). This should not be an issue because the mixture model was not used for testing hypotheses; its objective, rather, was to explore a clustered structure. Since there are only 12 posterior means, the mixture models must have less than 12 parameters; otherwise, a perfect fit would be obtained. The mixture model fitted to the posterior mean estimates $\bar{\theta}_l$ postulated that

$$\bar{\theta}_l \text{ or } \log\left(\frac{\bar{\theta}_l}{1-\bar{\theta}_l}\right) \text{ or } -\log(-\log(\bar{\theta}_l)) \sim \sum_{k=1}^{K} \pi_k N(\bar{\theta}_l|\mu_k, \sigma_k^2),$$

where K is the number of components of the mixture (clusters of posterior means of θ values or transforms thereof), π_k is the probability that $\bar{\theta}_l$ belongs to cluster k (subject to $\sum_{k=1}^{K}\pi_k = 1$), and μ_k and σ_k^2 are the mean and variance, respectively, of component k. For example, if $k=2$, there are 5 'free' parameters in the mixture; if $k=4$, there are 11 such parameters, so it is not sensible to fit a model with more than 4 components. Mixture model parameters were estimated by maximum likelihood via the expectation–maximization algorithm as implemented in the FlexMix package (Leisch, 2004) in the R project (R development core team, 2008). Upon convergence (assuming the stationary point was a global maximum), the conditional probability that $\bar{\theta}_l$ (or its transformation) belongs to cluster k is calculated as

$$\Pr(\text{locus } l \in \text{cluster } k|\text{parameter estimates}) = \frac{\hat{\pi}_k N(\bar{\theta}_l|\hat{\mu}_k, \hat{\sigma}_k^2)}{\sum_{k=1}^{K}\hat{\pi}_k N(\bar{\theta}_l|\hat{\mu}_k, \hat{\sigma}_k^2)}.$$

The locus was assigned to the cluster with the largest conditional probability. Models with different values of $k=1$, 2, 3 and 4 were compared using Akaike's information criterion (AIC), that is

$$\text{AIC}(K) = 2\left[p_K - \sum_{l=1}^{12}\log\left(\sum_{k=1}^{K}\hat{\pi}_k N(\bar{\theta}_l|\hat{\mu}_k, \hat{\sigma}_k^2)\right)\right],$$

where p_K is the number of parameters for a model with K components (McLachlan & Peel, 2000). Models with the smallest AIC values are preferred. It is known that this criterion tends to overstate the number of components due to violation of regularity conditions in mixture models (Celeux & Soromenho, 1996).

https://doi.org/10.1017/S0016672310000121 Published online by Cambridge University Press

Table 2. *Comparison of mixture models with 2, 3 or 4 components fitted to the 12 posterior means of* θ*-parameters and their logit or Gompit transforms in the argan tree data of Petit* et al. (*1998*). *AIC* (*models with smallest values are favoured and indicated in boldface*)

Variable	No. of components (k)	Iterations to convergence	AIC
θ	$k=1$	2	−0·651
	$k=2$	16	**−6·299**
	$k=3$	36	−2·921
	$k=4$	39	3·079
$\log\frac{\theta}{1-\theta}$	$k=1$	2	**39·100**
	$k=2$	28	40·102
	$k=3$	77	44·392
	$k=4$	94	50·392
$-\log[-\log(\theta)]$	$k=1$	2	26·909
	$k=2$	36	**24·328**
	$k=3$	41	27·742
	$k=4$	48	33·742

Results of the mixture model analysis, by number of components fitted, are shown in Table 2. The AIC favoured a mixture with two clusters when the response was either θ or its Gompit transform, and a single component when the logit transformation was used. Clearly, with data from only 12 loci, the analyses did not have enough power to resolve heterogeneity in a finer manner. This would certainly not be the case with SNP data, where the number of marker loci typically oscillates between a few thousands in some animal species to close to a million in humans. Classification probabilities using $K=2$ and estimates of cluster mean and standard deviation are shown in Table 3. Irrespective of whether θ values were transformed or not, loci were clustered into two groups, one consisting of loci 3, 5, 7, 8 and 9, possibly reflecting a selection signature, and the other one including the remaining loci, presumably representing neutral loci. The maximum likelihood estimates of the mean and variance of θ values in the cluster with loci 3, 5, 7, 8 and 9 were $0{\cdot}41 \pm 0{\cdot}21$, whereas the corresponding estimates in the other cluster were $0{\cdot}12 \pm 0{\cdot}03$. This assignment into clusters is consistent with the picture emerging from visual consideration of the box plots in Fig. 7.

The two-step procedure is simple and does not require the tailoring of problem-specific software. However, it has the drawback of not taking into account the uncertainty associated with the posterior distributions of the θ-parameters, inferred in the first step. In principle, a better approach is to feed the entire set of posterior samples to the clustering procedure, such that not only the location of the posterior distributions of the θs is considered but also their uncertainty as well. Although this is very appealing conceptually, it may create difficulties with the EM algorithm, leading to convergence failure. For instance Qanbari *et al.* (personal communication) employed the procedure with posterior means (each calculated with 1 million samples from the corresponding posterior distribution) with about 35 000 SNPs in Hereford and Simmental cattle. When posterior means were used as data, the mixture model approach revealed the existence of 4–5 clusters. However, when the 35 million samples were used as data points, the Expectation–Maximization (EM) algorithm, as implemented in FlexMix, failed to converge. An alternative to using the entire collection of samples is to feed a selected set of percentiles of the posterior distribution of each θ_l, so that a proxy for the dispersion of the individual posterior distributions enters into the analysis.

5. Discussion

The use of *F*-statistics for the study of genetic divergence between population dates back to Wright (1931). Holsinger & Weir (2009) have provided a justification for their usefulness, e.g., in association mapping and in detecting genomic regions affected by evolutionary processes, such as selection. These authors also reviewed different types of statistical methods for inferring F_{ST}, including Bayesian procedures. Method of moments estimation was prompted by the linear model formalism of Cockerham (1969, 1973), and a review is in Weir & Hill (2002). There has been an increased interest in Bayesian methods, and important contributions in this front have been made by Holsinger (1999, 2006), Beaumont & Balding (2004) and Guo *et al.* (2009).

In the Bayesian approaches that have been suggested, e.g., Holsinger (1999), the model poses a product binomial (or product multinomial in the case of multiple alleles) likelihood function for allelic frequencies, with conjugate prior distributions, such as beta or Dirichlet processes. Marginalizing over the allelic frequencies yields the beta binomial or Dirichlet-multinomial distributions used by Balding (2003) for likelihood-based inference. Holsinger (1999) matched the mean and variance of, e.g., the beta distribution, to the definition of θ, and obtained a joint posterior distribution which is a function of the unknown allelic frequencies, of θ (assumed exchangeable over all loci) and of the mean allelic frequencies in an undivided population. The implementation, as well as those of Beaumont & Balding (2004) and of Guo *et al.* (2009) requires MCMC. While the power and flexibility of hierarchical models coupled with MCMC are well known (Sorensen & Gianola, 2002), implementations are not trivial and monitoring of convergence to the equilibrium distribution is a

https://doi.org/10.1017/S0016672310000121 Published online by Cambridge University Press

Table 3. *Conditional probabilities of membership to one of two clusters for mixture models fitted to the posterior means of* θ *for the 12 loci in the argan tree, and their logit,* $log(\theta/1-\theta)$*, and Gompit,* $-log(-log(\theta))$*, transformations* (*boldfaced probability indicates the cluster with largest probability of membership*)

	θ means		logit(θ)		Gompit(θ)	
Locus	Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 1	Cluster 2
1	**0·93**	0·07	**0·91**	0·09	**0·91**	0·09
2	**0·92**	0·08	**0·83**	0·17	**0·89**	0·11
3	0·00	**1·00**	0·00	**1·00**	0·00	**1·00**
4	**0·92**	0·08	**0·82**	0·18	**0·88**	0·12
5	0·00	**1·00**	0·00	**1·00**	0·00	**1·00**
6	**0·95**	0·05	**0·91**	0·09	**0·93**	0·07
7	0·00	**1·00**	0·08	**0·92**	0·04	0·96
8	0·00	**1·00**	0·00	**1·00**	0·00	**1·00**
9	0·00	**1·00**	0·00	**1·00**	0·00	**1·00**
10	**0·92**	0·08	**0·89**	0·11	**0·89**	0·11
11	**0·95**	0·05	**0·92**	0·08	**0·93**	0·07
12	**0·87**	0·13	**0·76**	0·24	**0·83**	0·17
Cluster mean	0·12	0·41	−2·03	−0·52	−0·11	0·76
Cluster standard deviation	0·03	0·21	0·32	1·02	0·67	0·13

delicate matter (Cowles & Carlin, 1996). The idea in these methods is that, under a neutral model, all θ (over loci) should be realizations of the same stochastic process. Outlying θ values may be suggestive of genomic regions affected by selection. Typically, it is argued that loci are either neutral, subject to balancing selection or to directional selection favouring alleles in specific environments, e.g. Akey *et al*. (2002). However, the assignment of loci to specific types of processes is often arbitrary.

The present paper follows ideas of Holsinger (1999), but it differs in two important respects. The proposed method has two steps. First, allelic frequencies are assigned a non-informative prior, so that the mutual borrowing of information between loci is limited, leading to less shrinkage of frequencies towards a common value; in maximum likelihood there is no shrinkage at all, an issue criticized by Haldane (1948). Samples of allelic frequencies can be obtained directly (actually, their posterior distributions are tractable, analytically), and these draws are used to form draws from the posterior distribution of locus-specific θ-parameters, using the parametric definition of F_{ST} as a function of allelic frequencies. The first step was illustrated with hypothetical data and with type II diabetes data in Myles *et al*. (2007). The step leads to estimates of the posterior distribution of the θs, which can be used to explore underlying structure, presumably caused by different evolutionary forces. In the second step, the structure is explored by using features of the posterior distribution of the θs (posterior means or transformations thereof) as response variables in a mixture model. Data from Petit *et al*. (1998) on 12 isozyme loci in 12 populations of the argan tree in Morocco were used to illustrate the second step. Here, the posterior means of θ are treated as belonging to a mixture of normal distributions, which is then resolved into data-supported components. Since the final objective is that of clustering loci according to their similarity in θ values, departures from normality are arguably of little consequence. Here, logit and Gompit transformations were examined, and the clustering procedure produced exactly the same results. Using AIC as a gauge for model comparison, it was suggested that the 12 estimates of θ clustered into two groups, one representing putatively neutral loci (provided that this group reflects variation due to drift), and another one possibly corresponding to genomic regions affected by selection. With 12 loci only, it is unreasonable to expect a finer clustering structure. An ongoing study is applying the two-step procedure to large-scale SNP data in an animal population and this will be reported in a future communication.

As mentioned earlier in the paper, the two-step procedure has the disadvantage of not incorporating the uncertainty about the posterior distributions inferred in the first step. Although this can be remedied by using all posterior samples as input into the mixture model analysis, it can create numerical difficulties with the EM algorithm. This is an issue that needs additional research.

The method proposed here extends naturally to multiple alleles. In this case the likelihood is product multinomial, and the beta prior distribution is replaced by a Dirichlet distribution with minimum information content. The posterior distribution of the allelic frequencies is product Dirichlet, which is simple to sample from. Then, samples from the posterior distribution of θ_l would be drawn by evaluation of formulae similar to those in Nei (1973) where θ-parameters are averaged over alleles. For example,

https://doi.org/10.1017/S0016672310000121 Published online by Cambridge University Press

one could define

$$\theta_l = \sum_{m=1}^{M} \bar{p}_{l,m} \frac{\frac{\sum_{r=1}^{R}(p_{r,l,m}-\bar{p}_{l,m})^2}{R}}{\bar{p}_{l,m}(1-\bar{p}_{l,m})} = \sum_{m=1}^{M} \frac{\sum_{r=1}^{R}(p_{r,l,m}-\bar{p}_{l,m})^2}{R(1-\bar{p}_{l,m})},$$

where $p_{r,l,m}$ is the frequency of allele m at locus l in population r and $\bar{p}_{l,m}$ is the unweighted average over the R populations.

In common with the studies of Holsinger (1999), Beaumont & Balding (2004), Weir *et al.* (2005) and Guo *et al.* (2009), the procedure presented here assumes that allelic frequencies are in linkage equilibrium, so that the likelihood of all allelic frequencies is either product binomial or product multinomial. Accommodating linkage disequilibrium, especially with dense batteries of marker loci, represents a formidable task and it is a challenge for future research. For example, Akey *et al.* (2002) and Weir *et al.* (2005) reported that θ values of loci in regions of high linkage disequilibrium were similar. Guo *et al.* (2009) address correlations due to linkage, but not due to linkage disequilibrium, and do so by introducing a spatial structure for loci located in the same chromosome. Specifically, they proposed an autoregressive model in which logit transforms of θ values are correlated according to physical distance. The model is quite involved and requires MCMC computations. However, loci may be in linkage disequilibrium even though not being physically linked (Crow & Kimura, 1970), and such disequilibrium is very common in animal populations (Sandor *et al.*, 2006; de Roos *et al.*, 2008; Lipkin *et al.*, 2009; Qanbari *et al.*, 2009), where finite size and selection under epistasis are factors in building up linkage disequilibrium. The two-step approach considered here could be enhanced by exploring algorithms alternative to EM as well as by consideration of different types of mixtures, e.g., of beta distributions, which are more appropriate for random variables taking values in (0,1).

Appendix A: First derivatives of θ with respect to allelic frequencies

Let $\bar{p}_{.,l} = \frac{\sum_{r=1}^{R} p_{r,l}}{R}$. From (4), the derivative is

$$\begin{aligned}
\frac{\partial}{\partial p_{r,l}}\theta_l &= \frac{1}{\left(\frac{R\sum_{r=1}^{R} p_{r,l} - \left(\sum_{r=1}^{R} p_{r,l}\right)^2}{R}\right)}\left(2p_{r,l} - \frac{2\sum_{r=1}^{R} p_{r,l}}{R}\right) \\
&\quad - \frac{\sum_{r=1}^{R} p_{r,l}^2 - \frac{\left(\sum_{r=1}^{R} p_{r,l}\right)^2}{R}}{\left(\frac{R\sum_{r=1}^{R} p_{r,l} - \left(\sum_{r=1}^{R} p_{r,l}\right)^2}{R}\right)^2}\left(\frac{R-2\sum_{r=1}^{R} p_{r,l}}{R}\right) \\
&= \frac{2\theta_l}{\sum_{r=1}^{R} p_{r,l}^2 - \frac{\left(\sum_{r=1}^{R} p_{r,l}\right)^2}{R}}\left(p_{r,l} - \frac{\sum_{r=1}^{R} p_{r,l}}{R}\right) \\
&\quad - \frac{\theta_l}{\left(\frac{R\sum_{r=1}^{R} p_{r,l} - \left(\sum_{r=1}^{R} p_{r,l}\right)^2}{R}\right)}\left(\frac{R-2\sum_{r=1}^{R} p_{r,l}}{R}\right) \\
&= \left[\frac{2\left(p_{r,l} - \frac{\sum_{r=1}^{R} p_{r,l}}{R}\right)}{\sum_{r=1}^{R} p_{r,l}^2 - \frac{\left(\sum_{r=1}^{R} p_{r,l}\right)^2}{R}} - \frac{\left(1 - \frac{2\sum_{r=1}^{R} p_{r,l}}{R}\right)}{\sum_{r=1}^{R} p_{r,l} - \frac{\left(\sum_{r=1}^{R} p_{r,l}\right)^2}{R}}\right]\theta_l \\
&= \left[\frac{2\left(p_{r,l} - \bar{p}_{.,l}\right)}{\sum_{r=1}^{R} p_{r,l}^2 - \frac{\left(R\bar{p}_{.,l}\right)^2}{R}} - \frac{\left(1-2\bar{p}_{.,l}\right)}{\sum_{r=1}^{R} p_{r,l} - \frac{\left(R\bar{p}_{.,l}\right)^2}{R}}\right]\theta_l \\
&= \left[\frac{2\left(p_{r,l} - \bar{p}_{.,l}\right)}{R\left(\overline{p^2}_{.,l} - \bar{p}^2_{.,l}\right)} - \frac{\left(1-2\bar{p}_{.,l}\right)}{R\left(\bar{p}_{.,l} - \bar{p}^2_{.,l}\right)}\right]\theta_l \\
&= \left[\frac{2\left(p_{r,l} - \bar{p}_{.,l}\right)}{\overline{p^2}_{.,l} - \bar{p}^2_{.,l}} - \frac{\left(1-2\bar{p}_{.,l}\right)}{\bar{p}_{.,l}\left(1-\bar{p}_{.,l}\right)}\right]\frac{\theta_l}{R} \qquad \text{(A.1)}
\end{aligned}$$

Appendix B: Approximate Bayesian analysis

An approximate Bayesian analysis without sampling from the posterior distribution is also possible. An approximation to the mean and variance of the posterior distribution of θ_l can be obtained using a Taylor series expansion about the modes $\tilde{p}_{r,l}$ of the allelic frequencies. Let now $\widetilde{\nabla} = \left\{\frac{\partial}{\partial p_{r,l}}\theta_l\right\}_{p_{r,l} = \tilde{p}_{r,l}}$ be an $R \times 1$ vector of first derivatives evaluated at the posterior mode estimates (8) of the allelic frequencies. Then, approximately\tab

$$\theta_l \approx \widetilde{\theta}_l + \widetilde{\nabla}'(\mathbf{p}_l - \widetilde{\mathbf{p}}_l),$$

where $\tilde{\mathbf{p}}_l$ is the vector of posterior mode estimates of allele frequencies in the R groups. Then, approximately

$$\begin{aligned}
&E(\theta_l|\text{DATA}) \\
&\approx \widetilde{\theta}_l + \sum_{r=1}^{R}\left\{\frac{\partial}{\partial p_{r,l}}\theta_l\right\}_{p_{r,l}=\tilde{p}_{r,l}}\left(\frac{n_{r,A_l}+\frac{1}{2}}{2n_r+1} - \frac{n_{r,A_l}-\frac{1}{2}}{2n_r-1}\right) \\
&= \widetilde{\theta}_l + \sum_{r=1}^{R}\left\{\left[\frac{2\left(p_{r,l}-\bar{p}_{.,l}\right)}{\overline{p^2}_{.,l} - \bar{p}^2_{.,l}} - \frac{\left(1-2\bar{p}_{.,l}\right)}{\bar{p}_{.,l}\left(1-\bar{p}_{.,l}\right)}\right]\frac{\theta_l}{R}\right\}_{p_{r,l}=\bar{p}_{r,l}} \\
&\quad \times \left(\frac{n_{r,A_l}+\frac{1}{2}}{2n_r+1} - \frac{n_{r,A_l}-\frac{1}{2}}{2n_r-1}\right). \qquad \text{(B.1)}
\end{aligned}$$

Likewise

$$\text{Var}(\theta_l|\text{DATA}) \approx \widetilde{\nabla}'\,\text{Var}(\mathbf{p}_l|\text{DATA})\widetilde{\nabla}.$$

Since allelic frequencies have mutually independent distributions, the $R \times R$ variance–covariance matrix Var ($\mathbf{p}_l$|DATA) is diagonal with elements given by (9). Thus

$$\text{Var}(\theta_l|\text{DATA}) \approx \widetilde{\nabla}'\,\text{Diag}\left[\frac{\left(n_{r,A_l}+\frac{1}{2}\right)\left(n_{r,a_l}+\frac{1}{2}\right)}{\left(2n_r+1\right)^2\left(2n_r+2\right)}\right]\widetilde{\nabla}'. \qquad \text{(B.2)}$$

In short, each θ_l ($l = 1, 2, \ldots, L$) statistic will have a point estimate and an assessment of uncertainty, e.g., a credibility interval of size 95% given by the 2·5% and 97·5% percentiles of the corresponding posterior

https://doi.org/10.1017/S0016672310000121 Published online by Cambridge University Press

distribution estimated from samples, or from using a normal theory approximation, e.g.,

$$\widetilde{\theta}_l + 2\sum_{r=1}^{R}\left\{\left[\frac{2(p_{r,l}-\bar{p}_l)}{\overline{p_l^2}-\bar{p}_l^2}-\frac{(1-2\bar{p}_l)}{\bar{p}_l(1-\bar{p}_l)}\right]\frac{\theta_l}{R}\right\}_{p_{r,l}=\hat{p}_{r,l}}\left(\frac{n_{r,A_l}+\frac{1}{2}}{2n_r+1}-\frac{n_{r,A_l}-\frac{1}{2}}{2n_r-1}\right)$$

$$\pm 1{\cdot}96\sqrt{\sum_{r=1}^{R}\left\{\left[\frac{2(p_{r,l}-\bar{p}_l)}{\overline{p_l^2}-\bar{p}_l^2}-\frac{(1-2\bar{p}_l)}{\bar{p}_l(1-\bar{p}_l)}\right]\frac{\theta_l}{R}\right\}^2_{p_{r,l}=\hat{p}_{r,l}}\frac{\left(n_{r,A_l}+\frac{1}{2}\right)\left(n_{r,a_l}+\frac{1}{2}\right)}{(2n_r+1)^2(2n_r+2)}}.$$

Part of this work was carried out while the senior author was a Visiting Professor at Georg-August-Universität, Göttingen (Alexander von Humboldt Foundation Senior Researcher Award). Support by the Wisconsin Agriculture Experiment Station, and by grant NSF DMS-NSF DMS-044371 is acknowledged. This study is part of the project FUGATO-plus GenoTrack and was financially supported by the German Ministry of Education and Research, BMBF, the Föderverein Biotechnologieforschung e.V. (FBF), Bonn, and Lohmann Tierzucht GmbH, Cuxhaven. S. Qanbari thanks the H. Wilhelm Schaumann Stiftung Hamburg for financial support. Professor W. G. Hill and three anonymous reviewers are thanked for useful comments.

References

Akey, J. M. (2009). Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Research* **19**, 711–722.

Akey, J. M., Zhang, G., Zhang, K., Jin, L. & Shriver, M. D. (2002). Interrogating a high-density SNP map for signatures of natural selection. *Genome Research* **12**, 1805–1814.

Balding, D. J. (2003). Likelihood-based inference for genetic correlations. *Theoretical Population Biology* **63**, 221–230.

Beaumont, M. A. & Balding, D. J. (2004). Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology* **13**, 969–980.

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

Cavalli-Sforza, L. L. (1966). Population structure and human evolution. *Proccedings of the Royal Society London B. Biological Sciences* **164**, 362–379.

Celeux, G. & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Classification Journal* **13**, 195–212.

Cockerham, C. C. (1969). Variance of gene frequencies. *Evolution* **23**, 72–84.

Cockerham, C. C. (1973). Analyses of gene frequencies. *Genetics* **74**, 679–700.

Corander, J., Waldmann, P. & Sillanpää, M. (2003). Bayesian analysis of genetic differentiation between populations. *Genetics* **163**, 367–374.

Cowles, M. K. & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association* **91**, 883–904.

Crow, J. F. & Kimura, M. (1970). *An Introduction to Population Genetics Theory*. Caldwell: Blackburn Press.

de Roos, A. P. W., Hayes, B. J., Spelman, R. J. & Goddard, M. E. (2008). Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. *Genetics* **179**, 1503–1512.

Guo, F., Dey, D. K. & Holsinger, K. E. (2009). A Bayesian hierarchical model for analysis of single nucleotide polymorphisms, diversity in multilocus, multipopulation samples. *Journal of the American Statistical Association* **104**, 142–154.

Haldane, J. B. S. (1948). The precision of observed values of small frequencies. *Biometrika* **35**, 297–303.

Holsinger, K. E. (1999). Analysis of genetic diversity in geographically structured populations: a Bayesian perspective. *Hereditas* **130**, 245–255.

Holsinger, K. E. (2006). Bayesian hierarchical models in geographical genetics. In: *Applications of Computational Statistics in the Environmental Sciences* (ed. J. S. Clark & A. E. Gelfand), pp. 25–37. New York: Oxford University Press.

Holsinger, K. E. & Weir, B. S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. *Nature Review Genetics* **10**, 639–650.

Leisch, F. (2004). FlexMix: a general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software* **11**, 1–18.

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

Lipkin, E., Straus, K., Stein, T., Bagnato, A., Schiavini, F., Fontanesi, L., Russo, V., Medugorac, M., Foerster, M., Sölkner, J., Dolezal, M., Medrano, M. F., Friedmann, A. & Soller, M. (2009). Extensive long-range and non-syntenic linkage disequilibrium in livestock populations. *Genetics* **181**, 691–699.

Myles, S., Hradetzky, E., Engelken, J., Lao, O., Nürnberg, P., Trent, R. J., Wang, X., Kayser, M. & Stoneking, M. (2007). Identification of a candidate genetic variant for the high prevalence of type II diabetes in Polynesians. *European Journal of Human Genetics* **15**, 584–589.

McLachlan, G. & Peel, D. (2000). *Finite Mixture Models*. New York: Wiley.

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences of the USA* **70**, 3321–3323.

Petit, R. J., El Mousadik, A. & Pons, O. (1998). Identifying populations for conservation on the basis of genetic markers. *Conservation Biology* **12**, 844–855.

Qanbari, S., Pimentel, E. C. G., Tetens, J., Thaller, G., Lichtner, P., Sharifi, A. R. & Simianer, H. (2009). The pattern of linkage disequilibrium in German Holstein cattle. *Animal Genetics* doi: 1111/j.1365-2052.2.2009.02011.X.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0. Available at http://www.R-project.org

Robertson, A. (1975). Gene frequency distributions as a test of selective neutrality. *Genetics* **81**, 775–785.

Sabeti, P. C., Schaffner, S. F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, S., Palma, A., Mikkelsen, T. S., Altshuler, D. & Lander, E. S. (2006). Positive natural selection in the human lineage. *Science* **312**, 1614–1620.

Sandor, C., Farnir, F., Hansoul, S., Coppieters, W., Meuwissen, T. & Georges, M. (2006). Linkage disequilibrium on the bovine X chromosome: characterization and use in quantitative trait locus mapping. *Genetics* **173**, 1777–1786.

Sorensen, D. & Gianola, D. (2002). *Likelihood, Bayesian and MCMC Methods in Quantitative Genetics*. New York: Springer.

Weir, B. S., Cardon, L. R., Anderson, A. D., Nielsen, D. M. & Hill, W. G. (2005). Measures of human population structure show heterogeneity among genomic regions. *Genome Research* **15**, 1468–1476.

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

Weir, B. S. & Hill, W. G. (2002). Estimating *F*-statistics. *Annual Review of Genetics* **36**, 721–750.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics* **16**, 97–159.

Wright, S. (1951). The genetical structure of populations. *Annals of Eugenics* **15**, 323–354.