Molecular Ecology (2004) **13**, 969–980 doi: 10.1111/j.1365-294X.2004.02125.x

Identifying adaptive genetic divergence among populations from genome scans

MARK A. BEAUMONT and DAVID J. BALDING

School of Animal and Microbial Sciences, The University of Reading, Whiteknights, PO Box 228, Reading RG6 6AJ, UK Department of Epidemiology and Public Health, Imperial College, St Mary's Campus, Norfolk Place, London W2 1PG, UK

Abstract

The identification of signatures of natural selection in genomic surveys has become an area of intense research, stimulated by the increasing ease with which genetic markers can be typed. Loci identified as subject to selection may be functionally important, and hence (weak) candidates for involvement in disease causation. They can also be useful in determining the adaptive differentiation of populations, and exploring hypotheses about speciation. Adaptive differentiation has traditionally been identified from differences in allele frequencies among different populations, summarised by an estimate of F_{ST}. Low outliers relative to an appropriate neutral population-genetics model indicate loci subject to balancing selection, whereas high outliers suggest adaptive (directional) selection. However, the problem of identifying statistically significant departures from neutrality is complicated by confounding effects on the distribution of F_{ST} estimates, and current methods have not yet been tested in large-scale simulation experiments. Here, we simulate data from a structured population at many unlinked, diallelic loci that are predominantly neutral but with some loci subject to adaptive or balancing selection. We develop a hierarchical-Bayesian method, implemented via Markov chain Monte Carlo (MCMC), and assess its performance in distinguishing the loci simulated under selection from the neutral loci. We also compare this performance with that of a frequentist method, based on moment-based estimates of F_{ST}. We find that both methods can identify loci subject to adaptive selection when the selection coefficient is at least five times the migration rate. Neither method could reliably distinguish loci under balancing selection in our simulations, even when the selection coefficient is twenty times the migration rate.

Keywords: adaptation, beta-binonical, gene flow, Lewontin–Krakauer test, population structure, selection

Received 21 October 2003; revision received 19 December 2003; accepted 19 December 2003

Introduction

A key problem of evolutionary biology concerns the number and location of genes involved in adaptation and speciation. Evidence from molecular population genetics indicates that speciation generally occurs in the face of gene flow and is adaptively driven (Wu 2001). Populations in different environments will initially differ genetically at a few key sites in their genomes, and the surrounding DNA will also differ due to linkage disequilibrium. Divergent selection reduces immigration locally in the vicinity of the selected locus and, as the populations become more diverged, gene flow is progressively reduced and eventually ceases. The challenge is to identify the key genes involved in this process.

Hitherto, the degree of adaptive divergence between populations has been determined genetically by some measure of distinctiveness — for example the possession of reciprocal monophyly in mitochondrial sequences. However, this distinctiveness may, particularly if only based on mtDNA or a few nuclear markers, largely reflect the vagaries of demographic history. What is needed is to be able to quantify the distinctiveness of populations in terms of their local adaptation, which may also only involve a few genes, but genes with key functional roles (Black *et al.* 2001; Luikart *et al.* 2003).

Correspondence: Mark Beaumont. Fax: 0118 931 0180; E-mail: m.a.beaumont@reading.ac.uk

An approach to using genetic information to address these problems was suggested by Lewontin and Krakauer (1973). Differential adaptation, or artificial selection on traits, can lead to large between-population allele frequency differences at the loci that control the traits involved. These differences will occur at a small number of DNA sites, but are potentially identifiable because linkage will lead to 'islands' of differentiation around the selected sites, and any markers sampled within an island should also show differentiation. F_{ST} provides an appropriate scale on which to quantify this differentiation, and hence identify outlier loci.

It is now becoming possible to implement such an approach relatively cheaply on a genome-wide scale. In particular the development of methods to genotype large numbers of individuals for many types of marker such as SNPs, AFLPs, CATs, and ESTs, greatly enlarges the scope of methods based on the original idea of Lewontin and Krakauer (1973). There has been recent interest in using these general ideas to identify loci that might be targets for selection, in humans (e.g. Akey *et al.* 2002 using SNPs; Kayser *et al.* 2003 using microsatellites) and in other populations of evolutionary interest (Wilding *et al.* 2001; Storz and Nachman 2003).

A variety of different statistical methods have been developed from the original idea of Lewontin and Krakauer (Bowcock *et al.* 1991; Beaumont and Nichols 1996; Vitalis *et al.* 2001; Schlotterer 2002; Porter 2003). However, the statistical problem remains challenging because of the multiple testing of many genomic locations, and because of the confounding effects of demographic factors on F_{ST}. In this study we develop simulations to assess the performance of F_{ST}-based methods in distinguishing loci under balancing or adaptive selection from unlinked, neutral loci. We also develop a likelihood-based method, implemented via Markov chain Monte Carlo (MCMC), that exploits a hierarchical-Bayesian model similar to that of Balding *et al.* (1996). The model has two levels: a lower level in which the likelihood for the allele-frequency counts is expressed as a function of F_{ST}, and a higher-level model for the F_{ST} values. We compare the performance of the new method with that of the FDIST program of Beaumont and Nichols (1996) on both the simulated datasets and a real dataset.

Methods

Likelihood for allele counts

A likelihood formula expresses the probability of a set of observed data given the parameters of an assumed model. Likelihood-based methods are typically preferred to methods based on summary statistics because they can exploit all of the information in the data, whereas typically there is some loss of information in extracting a summary statistic from raw data. This benefit only accrues if the modelling assumptions are reasonable: methods based on summary statistics may be more robust to deviations from the modelling assumptions. However, likelihood-based methods have further advantages, such as flexibility to accommodate missing data, and quantitative model comparison, see for example Sorensen & Gianola (2002).

We adopt the multinomial-Dirichlet likelihood for the allele frequency counts at a locus within a population. This likelihood arises in a wide range of neutral population genetics models, in particular the infinite-island model of Wright (1943); for assumptions and derivations, see Balding and Nichols (1995), Rannala and Hartigan (1996), and Balding (2003).

Introducing Γ to denote the gamma function defined by $\Gamma(y + 1) = y\Gamma(y)$ (with $\Gamma(1) = 1$), the multinomial-Dirichlet likelihood can be conveniently expressed in the form:

$$\begin{aligned} L_{ij} &\equiv P(a_{ij1}, \dots, a_{ijK_i} | \lambda_{ij}, x_{i1}, \dots, x_{iK_i}) \\ &= \frac{\Gamma(\lambda_{ij})}{\Gamma(n_{ij} + \lambda_{ij})} \prod_{k=1}^{K_i} \frac{\Gamma(a_{ijk} + \lambda_{ij} x_{ik})}{\Gamma(\lambda_{ij} x_{ik})}, \end{aligned} \tag{1}$$

in which a_{ijk} denotes the count of allele k in population j at locus i, and n_{ij} denotes the sample size, $n_{ij} = \sum_{k=1}^{K_i} a_{ijk}$. The scaling parameter λ_{ij} in (1) controls, together with the n_{ij}, the variance of the sample allele proportions at locus i away from the baseline values x_{ik}. Under a neutral island model, λ_{ij} corresponds to the number of migrants into population j, which hence should be constant over loci i. Outlier values of λ_{ij} for one or more populations j may indicate that locus i is subject to selection. We will focus below on the case $K_i = 2$ of diallelic loci, in which case the multinomial-Dirichlet is also called the beta-binomial.

Each x_{ik} in (1) is a (nuisance) parameter representing the frequency of allele k at locus i in the migrant gene pool. In an infinite-island model, x_{ik} is the mean allele proportion over the islands; equivalently, it is the allele proportion in the continent of an island-continent model. In some settings there may be background information leading to point estimates of the x_{ik}. Below, we treat them as unknown a priori, and eliminate them via integration with respect to a (multivariate) uniform prior distribution. We often restrict attention to diallelic loci, in which case we can drop the subscript k and assume a single allele proportion x_i which has a uniform prior distribution.

Balding (2003) gives a recursive expression for L_{ij} that is useful for a genealogical interpretation in which

$$\lambda_{ij} = 1/F_{ij} - 1 \tag{2}$$

where, for locus i in population j, F_{ij} is the probability that two randomly chosen chromosomes in a population have

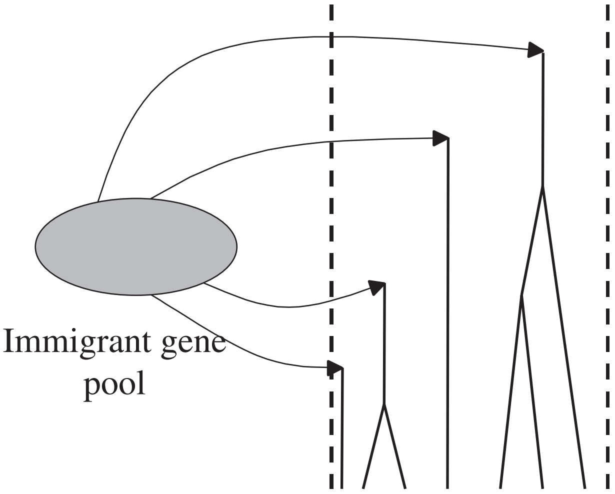

Fig. 1 An example of the genealogical structure leading to equation (1). Lineages either coalesce backwards in time, or are immigrants. Eventually all lineages have an immigrant ancestor. The allelic type of the immigrants is drawn at random from the immigrant gene pool. Dropping this type for all immigrants forwards in time down to the sample gives the allele frequency in the sample.

a common ancestor within that population, without there having been any intervening migration or mutation. This corresponds to the original view of F_{ST} as an inbreeding coefficient (Crow and Kimura, 1970), and we will follow that usage here. Genealogically, the migration process underlying (eqn. 1) corresponds to the 'scattering phase' of Wakeley (1999). The only difference between the description here (see Figure 1) and that of Wakeley (1999) is that in the latter the genealogy of the immigrants is explicitly modelled (the 'collecting phase').

Combining information across loci and populations

Assuming that the allele frequencies at distinct loci and/or populations are mutually independent, given the λ_{ij}, the joint likelihood for the λ_{ij} and x_{ik} is:

$$L = \prod_{i=1}^{I} \prod_{j=1}^{J} L_{ij}. \tag{3}$$

We could work directly with L to estimate the λ_{ij} and hence the F_{ij}, in which case we would effectively be estimating each F_{ij} based only on the allele counts at locus i in population j, with the allele counts at locus i in the other populations only being used to estimate the x_{ik}. This would lead to poor estimates for the F_{ij}, particularly at diallelic loci, because each allele count in effect represents a sample of size one (Balding 2003).

To achieve more precise estimates, we exploit the fact that each F_{ij} value reflects contributions from locus-specific effects, such as mutation and some forms of selection, and population-specific effects, such as effective population sizes, migration rates, and population-specific mating patterns. To model these locus and population effects, we adopt a regression approach and hence seek a regression equation for F_{ij}. Balding *et al.* (1996) chose the inverse linear regression model:

$$1/F_{ij} = 1 + \alpha_i + \beta_j + \gamma_{ij}, \tag{4}$$

in which α_i is a locus effect, β_j is a population effect and γ_{ij} is an 'error' term corresponding to a specific locus-by-population effect. Equation (4) has the property that if β_j is large for some population j, then F_{ij} is small in population j for all loci i. This property is not suitable for our present application of identifying outlying loci, and instead we adopt the more familiar logistic regression model in which

$$\log\left(\frac{F_{ij}}{1 - F_{ij}}\right) = \alpha_i + \beta_j + \gamma_{ij}. \tag{5}$$

Equivalently, (5) can be written

$$F_{ij} = \frac{\exp(\alpha_i + \beta_j + \gamma_{ij})}{1 + \exp(\alpha_i + \beta_j + \gamma_{ij})}. \tag{6}$$

There is substantial flexibility in the choice of prior distributions for the regression parameters α_i, β_j, and γ_{ij}. Gaussian priors are natural for the coefficients of regression models, and we selected their means and variances to achieve an implied prior distribution for each F_{ij} that has non-negligible density over almost the whole interval from zero to one. For the population effects β_j, the prior mean and standard deviation (sd) adopted below are, respectively, –2 and 1.8. We assigned prior mean zero to the α_i and γ_{ij}, so that the model in which F_{ij} is constant over i is favoured a priori. The prior sd values for the α_i and γ_{ij} then control the amount of 'shrinkage' towards this model. Below, the prior sd of each of the α_i is 1, while the γ_{ij} each have a prior sd of 0.5. The resulting prior for the F_{ij} has mode 0.5%, median 12%, mean 23%, and 95% equal-tailed interval from 0.2% to 90%.

Implementation

The Bayesian statistical model specified by equations (1), (2), and (6), together with the prior assumptions described above, has been implemented in a Metropolis-Hastings MCMC algorithm (see, for example, Gilks *et al.* 1996). At each step, either a locus i is chosen, and a mean-zero Gaussian update is added to α_i and each of the γ_{ij}, or a population j is chosen and β_j is updated similarly. In either case, the update is accepted or rejected in the usual way for a Metropolis algorithm: if the posterior density (proportional to the product of likelihood and prior) is increased, the update is always accepted; if it decreases, the update is accepted with probability equal to the ratio of the new posterior density to its current value.

The baseline frequencies are also updated, one locus at a time. In this case the proposed values are chosen from a Dirichlet (or beta in the diallelic case) distribution, with mean equal to the current values. Since the Dirichlet is not symmetric, a Metropolis-Hastings update is required, in which the posterior density ratio is weighted by the ratio of the forward and reverse proposal densities. The intuition is that if a particular update is proposed frequently, its probability of acceptance must be decreased proportionately; see Gilks *et al.* (1996) for further details. Dirichlet updates can be problematic if a frequency value becomes very close to zero, and we avoided this by imposing a minimum allele frequency of 10^{-3}.

The key to the computational efficiency of the algorithm is that the likelihood can be decomposed in two ways: as a product over loci; and as a product over populations. Thus each update requires only the re-computation of the relevant terms of the likelihood. The analyses of the 1000-gene simulations described below, took approximately 6 hours on a 2.4 Ghz Pentium processor running Linux to achieve good convergence, assessed using standard diagnostic checks, and 2000 approximately uncorrelated outputs.

Interpretation

For the purposes of identifying loci subject to selection, interest primarily focusses on the posterior distribution of the locus-effect parameters: a positive value of α_i suggests that locus i is subject to adaptive selection, whereas a negative value suggests balancing selection that tends to homogenise allele frequencies over populations. Ideally within a Bayesian framework, we would wish to assign a posterior probability to each hypothesis of the form $\alpha_i = 0$. However, this adds to the computational burden and poses the problem of specifying appropriate alternative hypotheses. Instead, we adopt a simple informal criterion for classifying a posterior distribution for α_i that is 'significant': we define α_i to be 'significant at level P' if its equal-tailed $100(1 - P)\%$ posterior interval excludes zero. For example, if $P = 5\%$ then α_i is significantly positive if its 2.5% quantile is positive, and is significantly negative if its 97.5% quantile is negative.

The γ_{ij} also have an interpretation in terms of selection: a large positive or negative value indicates an individual locus and population at which the allele frequencies are unexpected given the α_i, β_j and x_{ik} values. For example, a large positive γ_{ij} could indicate a population in which selection has driven one allele to fixation, possibly due to local climate, whereas selection effects are weak or absent at that locus in the environment experienced by the other populations. Because inferences about the γ_{ij} do not, by definition, benefit from sharing information across populations or loci, and because of the very large number of γ_{ij} values in genome-wide settings, only extremely marked selection effects are likely to be distinguishable via the γ_{ij}.

Simulation Model

In order to compare the behaviour of the different methods, we simulated gene frequency data from a Wright-Fisher model with migration. The model allows for an arbitrary number d of constant-size demes. In our simulations all demes have the same size, N chromosomes, although this could easily be changed. To allow for the possibility of adaptive selection, for a particular locus the demes are arbitrarily given attributes: 'neutral', 'red', or 'blue'. These are assigned at random, independently for each selected locus, so that the number of populations in which selection applies at a locus under selection is random. As described in more detail below, in 'neutral' demes all alleles have the same fitness, whereas in the others it is possible for alleles to be 'blue'-adapted or 'red'-adapted. The random allocation of labels implies that each deme has different attributes to which a locus can adapt — for example, 'hot' versus 'cold', and 'wet' versus 'dry', and 'crabs present' versus 'crabs absent'. In effect, each locus is specialised to be able to adapt to a particular attribute independently of other loci. This lack of correlation is unlikely to be true in general (i.e. there will be a number of loci that affect, for example, adaptation to heat), but is biologically more plausible than the alternative in which a deme has the same attribute for all loci.

The immigration rate into a deme is chosen at random by drawing individual Fs from a beta distribution, and then substituting $m = (1 - F)/(2NF)$. In each generation, a binomially-distributed number of chromosomes in each deme is replaced by immigrants chosen at random from all other demes. Each immigrant chromosome replaces a randomly chosen resident chromosome.

Two mutation models are employed. In one case (two-locus) there is a 'marker' locus completely linked with a 'selected' locus. The marker locus evolves according to an infinite allele model. The selected locus evolves according to a parent-independent K-allele model with up to three alleles: 'blue', 'red', and neutral. In the other case (marker-selected), the marker locus is itself susceptible to selection and evolves according to an infinite allele model, but each mutation, as well as being distinct, has a selective effect that is either 'blue', 'red', or neutral. In both models the selective effect of each realised allele is drawn with probabilities P_{neut}, P_{red}, P_{blue}, that sum to 1. We will use 'gene' or 'locus' to refer to one realisation from either the two-locus or the marker-selected models. A multilocus data set is made up from independent realisations and are either all two-locus, or all marker-selected.

In the marker-selected case, the initial parametric frequencies are simulated from a Dirichlet distribution with

K_{max} alleles, and all parameters equal to Θ_M/K_{max}. This distribution is expected under a K-allele model, where each mutation has probability $1/K_{max}$ of being to a particular allele independent of its current type, and where Θ_M is the scaled mutation rate. From this distribution a sample of size Nd is drawn, and allocated among the d demes. The realised number of alleles, K, is generally much lower than K_{max} so that it is well approximated by the infinite allele model. Mutations subsequently occur at a rate μ_M.

Initialisation in the two-locus case is based on an urn-scheme simulation of the coalescent (Donnelly and Tavaré 1995) in which the mutations are laid down for each locus, with rates Θ_M for the marker locus and Θ_S for the selected locus, using the mutation models described above. The sample is allocated among demes as for the marker-selected case. Mutations subsequently occur at frequency μ_M for the marker locus and μ_S for the selected locus. The frequencies may or may not be consistent with values chosen for Θ_M and Θ_S.

The genes are unlinked, and each may be either neutral, directional-selected, or balancing-selected. (We use 'directional-selected' as a synonym for adaptively-selected.) For neutral genes $P_{neut} = 1$. For directional-selected genes we assume a diploid selection model in which relative fitness in a blue deme is $1 + s$ for blue homozygotes, and $1 + s/2$ for blue heterozygotes, and 1 for all other genotypes. The same selective effects pertain for red alleles in red demes. For balancing-selected genes, in either red or blue demes, blue-red heterozygotes have fitness $1 + s$, all other heterozygotes and all homozygotes have fitness 1. In the current model the selection coefficients can be different for balancing-selected and neutral-selected genes, but within a run are the same for all genes in the same class. In neutral demes all genotypes have fitness 1. The mean fitness is calculated from these selection coefficients assuming Hardy-Weinberg equilibrium. The proportions of each allele in the gamete pool are then calculated.

Chromosomes in the current generation are subject to mutation, then some are replaced with immigrants, and then the next generation is sampled according to the selection coefficients specified above. Thus, selection occurs *after* mutation and immigration. In the analyses described here, we used $T = 50\,000$ generations, which is not intended to lead to any equilibrium, but is chosen to be sufficiently long that the allele frequencies reflect the selection coefficients. Biologically, the simulations correspond to a situation in which a species with no loci under selection colonises new habitats T generations ago, and selection (both directional and balancing) occurs under this changed regime. Certainly for the neutral loci, the chosen value of T is large enough that it would be reasonable to assume that the MRCA for most loci would be contained within this simulation period.

After T generations, chromosomes are sampled, with replacement, in some or all of the demes. A simulated data set may then be discarded if an allele has global frequency greater than a given threshold, or if non-neutral genes have no non-neutral alleles (this is only relevant if mutations can give rise to neutral alleles). If SNP-like data are desired then datasets with more than two alleles can also be discarded.

Simulated data sets

Twelve data sets have been generated using the Wright-Fisher model described above, three of which consist of 1000 genes, and the remainder 500 genes. For the 500-gene data sets half were marker-selected and half two-locus. In each case we assume six populations, each of size 500 chromosomes, from each of which 100 chromosomes are sampled. Each population is red with probability 0.4, blue with probability 0.4, and neutral with probability 0.2, independently at each locus. For initialisation $K_{max} = 100$, $\Theta_M = 0.06$, and $\Theta_S = 0.6$. For non-neutral genes $P_{neut} = 0$, $P_{red} = 0.5$, $P_{blue} = 0.5$. The mutation rates are $\mu_M = 0.00001$ and $\mu_S = 0.0001$. The same mutation parameters for the marker locus were used in the two-locus and marker-selected models. Within a data set the selection coefficients for balancing-selected and directional-selected loci are the same. The value of F was allowed to vary among populations and this was chosen by simulating values from a beta distribution with parameters 0.25 and 2.25. This produces generally about a three to five fold difference in F between the most inbred and most outbred populations, with an expected mean of 0.1. Only diallelic markers with minor allele frequency > 0.05 were kept. The details that distinguish the different data sets are given in Table 1.

Analysis with FDIST

In order to compare the results from the Bayesian analyses with a frequentist method based on summary statistics, we use the program FDIST described in Beaumont and Nichols (1996). This is a method in which θ, Weir and Cockerham's (1984) estimator of F_{ST}, is calculated for each locus in the sample. Coalescent simulations are then performed to generate data sets with a distribution of θ close to the empirical distribution. Based on this simulated distribution it is possible to calculate quantiles or P-values for loci of interest. Loci with unusually high or low values of θ are regarded as potentially under selection. The currently distributed version (http://www.rubic.reading.ac.uk/~mab/software/fdist2.zip) differs slightly from that originally described in that it generates a roughly uniform distribution of heterozygosities from (0,1), by generating data with different mutation rates. In addition it allows the user to specify the total number of demes in the system, whereas in the original version this was fixed at 100. For this study FDIST 2 has additionally been modified so that a) the user can specify the maximum number of alleles (here, two), b)

Table 1 Parameters of simulated data sets. Details of mutation models are given in the text

Identifier	Selection coefficient	Mutation model	Number of neutral loci	Number of direc.-sel. loci	Number of bal.-sel. loci
M-2L-02	0.02	2-locus	900	80	20
M-2L-05	0.05	2-locus	900	80	20
M-2L-10	0.1	2-locus	900	80	20
D-2L-02	0.02	2-locus	450	40	10
D-2L-05	0.05	2-locus	450	40	10
D-2L-10	0.1	2-locus	450	40	10
D-2L-20	0.2	2-locus	450	40	10
D-MS-02	0.02	marker-selected	450	40	10
D-MS-05	0.05	marker-selected	450	40	10
D-MS-10	0.1	marker-selected	450	40	10
D-MS-20	0.2	marker-selected	450	40	10
D-NEUTR	0.0	—	500	0	0

a single mutation rate can be specified, c) a maximum global allele frequency can be specified. For the original version of FDIST a trial-and-error procedure was used for finding the best-fitting value of F_{ST} to use in the simulations. To simplify the current analyses we calculate the median estimate of F_{ST} (using θ of Weir and Cockerham, 1984) for loci with heterozygosity > 0.1. This value is then used to simulate data. We simulate 20 000 points and then use these to obtain approximate conditional densities and calculate approximate P-values for each data sample locus, using the procedure in Beaumont and Nichols (1996). For the simulated data sets we assumed six sampled populations, with a metapopulation size of 100, and with sample sizes 100. Simulated loci with more than two alleles, or with a minor allele frequency < 0.05 were rejected. We assumed a value of $\theta = 0.06$. For one test data set we investigated the effect of assuming an almost 20-fold higher value of $\theta = 1.0$, and, as illustrated below, the results are remarkably insensitive to this. This lack of sensitivity to θ has already been shown in Beaumont and Nichols (1996).

Results

Bayesian regression method

In interpreting the results from the simulated data, recall that the effective size of each of the six demes is just 500. Since selection coefficients scale with effective populations size, the equivalent selection coefficients will be lower in larger populations (e.g. humans). For selection coefficients of 2%, the genes correctly detected as under directional selection are almost balanced by the false positives (Table 2). At higher selection coefficients, the results are encouraging in respect of directionally-selected loci, but none of the balancing-selected loci are detected at the

Table 2 Numbers of genes simulated under ('true') balancing selection, neutrality, and directional selection that were classified in each category at level 5% (10%) by the Bayesian regression analysis. At the 100P% level, gene i was classified as 'directional' if the $P/2$ quantile of the posterior distribution of α_i was positive, 'balancing' if the $(1 - P/2)$ quantile was negative, and otherwise was classified as 'neutral'

True:	Balancing selection			Neutral			Directional selection		
Classified:	bal.	neut.	direc.	bal.	neut.	direc.	bal.	neut.	direc.
M-2L-02	0 (0)	20 (20)	0 (0)	0 (5)	893 (877)	7 (18)	0 (0)	72 (70)	8 (10)
M-2L-05	0 (0)	20 (20)	0 (0)	0 (2)	897 (889)	3 (9)	0 (0)	51 (41)	29 (39)
M-2L-10	0 (3)	20 (17)	0 (0)	1 (7)	897 (887)	2 (6)	0 (0)	20 (15)	60 (65)
D-2L-02	0 (0)	10 (10)	0 (0)	0 (3)	448 (439)	2 (9)	0 (0)	37 (33)	3 (7)
D-2L-05	0 (0)	10 (10)	0 (0)	0 (3)	446 (438)	4 (9)	0 (0)	25 (21)	15 (19)
D-2L-10	0 (1)	10 (9)	0 (0)	0 (0)	448 (447)	2 (3)	0 (0)	10 (5)	30 (35)
D-2L-20	0 (1)	10 (9)	0 (0)	0 (2)	450 (447)	0 (1)	0 (0)	1 (1)	39 (39)
D-MS-02	0 (0)	10 (10)	0 (0)	0 (0)	449 (447)	1 (3)	0 (0)	39 (37)	1 (3)
D-MS-05	0 (1)	10 (9)	0 (0)	0 (4)	450 (491)	0 (5)	0 (0)	15 (11)	25 (29)
D-MS-10	0 (0)	10 (10)	0 (0)	0 (3)	450 (444)	0 (3)	0 (0)	11 (2)	29 (38)
D-MS-20	0 (0)	10 (10)	0 (0)	0 (3)	450 (445)	0 (3)	0 (0)	0 (0)	40 (40)
D-NEUT				0 (2)	496 (487)	4 (12)			

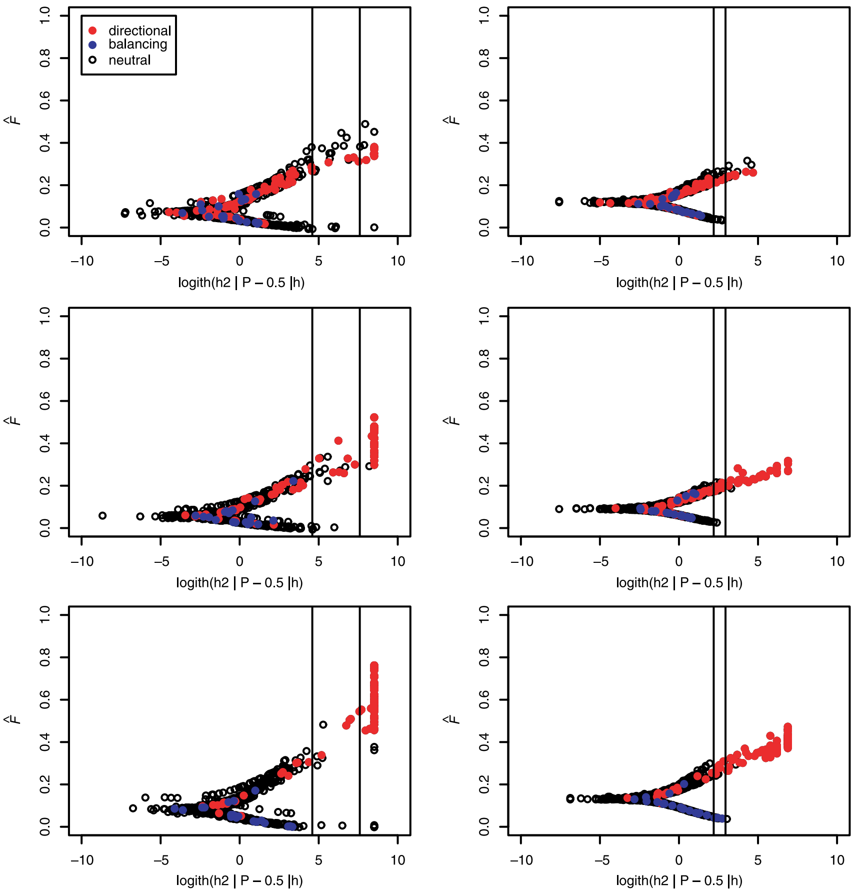

Fig. 2 Summary of the results of analyses of the three 1000-locus data sets (top: M-2L-02, middle: M-2L-05, bottom: M-2L-10). The results from FDIST are shown on the left, and those from the Bayesian regression method are shown on the right. An estimate of F_{ST} is plotted against empirical P-values for each locus. The vertical bars show the critical P-values used for identifying outlier loci, as described in the text. Because of sample-size effects the minimum two-tailed P-value was set at 0.001 for the Bayesian regression method and 0.0002 for FDIST

5% level, and only six (4.3%) are detected over all the simulations at the 10% level. Of the 6800 neutral loci in all 12 datasets, 25 (0.4%) were mis-classified as directionally-selected and one (0.01%) as under balancing selection at the 5% level. At the 10% level, the rates were 1.2% for directional false positives, and 0.5% for balancing false positives.

Although directionally-selected loci are more often detected under the marker-selected model than under the two-locus model, the difference is not as marked as might have been expected: 95 versus 87 at the 5% level, 110 vs 100 at the 10% level.

The distribution of empirical 'P-values' in the three 1000-loci datasets, together with their associated posterior mean F_{ST} values, are illustrated in Figure 2 (right panels). Recall that 'P-value' here means P ($\alpha_i < 0$), and empirically it is the proportion of negative values among the MCMC outputs for α_i. In the figure, for improved visualisation, the P-value has been transformed via logit(2|p – 0.5|), where logit(x) ≡ log(x/(1 – x)). In general, except when the selection coefficient is 2% (top right panel), the directionally-selected (red) and balancing-selected (blue) loci tend to be located towards the appropriate tails of the distribution. However, for the latter to achieve significance, F_{ST} has to be very close to zero, which is only occasionally realised, and only when the selection coefficient is 10% (bottom right panel). The simulations were carried out with a mean F_{ST} of 0.1. In populations with a higher F_{ST} it may be easier to detect loci under balancing selection. However, it is also the case that typical values of F_{ST} for many populations (including humans) are generally around 0.1 or lower, which implies that it may often be difficult to detect balancing selection from gene frequency data.

Summary-statistic method

The results from the analyses with FDIST are presented in Table 3. In order to compare these with the Bayesian regression method we chose critical p-values such that the two-tailed false positive rate for all 6800 neutral loci matched as closely as possible for the two methods. In the absence of information about the true levels of selection coefficients, matching the false-positive rate seems a pragmatic way of comparing the two methods. Thus, we assigned a P-value in FDIST of 0.0005 to compare with the Bayesian 5%, and a P-value of 0.01 to compare with the Bayesian 10% level.

Table 3 Numbers of genes simulated under ('true') balancing selecting, neutrality, and directional selection that were classified in each category at level 0.05% (1%) by the FDIST program. The entry marked with an asterisk (*) is a replicate analysis using scaled mutation rate of 1.0 rather than 0.06, as described in the text. Other details are as for Table 2

True:	balancing selection			neutral			directional selection		
Classified:	bal.	neut.	direc.	bal.	neut.	direc.	bal.	neut.	direc.
M-2L-02	0 (0)	20 (20)	0 (0)	1 (5)	895 (880)	4 (15)	0 (0)	72 (68)	8 (12)
M-2L-05	0 (0)	20 (20)	0 (0)	0 (4)	898 (886)	2 (10)	0 (0)	47 (40)	33 (40)
M-2L-05*	0 (0)	20 (20)	0 (0)	2 (5)	896 (885)	2 (10)	0 (0)	47 (40)	33 (40)
M-2L-10	0 (0)	20 (20)	0 (0)	3 (5)	895 (889)	2 (6)	0 (0)	22 (17)	58 (63)
D-2L-02	0 (0)	10 (10)	0 (0)	2 (2)	446 (443)	2 (5)	0 (0)	39 (33)	1 (7)
D-2L-05	0 (0)	10 (10)	0 (0)	5 (6)	444 (433)	1 (11)	0 (0)	23 (14)	17 (26)
D-2L-10	0 (1)	10 (9)	0 (0)	0 (2)	449 (445)	1 (3)	0 (0)	10 (8)	30 (32)
D-2L-20	0 (1)	10 (9)	0 (0)	0 (0)	450 (445)	0 (5)	0 (0)	1 (1)	39 (39)
D-MS-02	0 (0)	10 (10)	0 (0)	0 (1)	450 (447)	0 (2)	0 (0)	38 (35)	2 (5)
D-MS-05	0 (1)	10 (9)	0 (0)	2 (5)	448 (442)	0 (3)	0 (0)	20 (5)	20 (35)
D-MS-10	0 (1)	10 (9)	0 (0)	0 (0)	450 (447)	0 (3)	0 (0)	6 (1)	34 (39)
D-MS-20	0 (0)	10 (10)	0 (0)	0 (3)	450 (443)	0 (4)	0 (0)	1 (0)	39 (40)
D-NEUT				0 (2)	499 (494)	1 (4)			

Table 4 Summary of the results in Tables 2 and 3

	FDIST		Bayesian regression	
	$P = 0.0005$	$P = 0.01$	$P = 0.05$	$P = 0.1$
Total false positive	26	106	26	115
total dir. detected	281	338	279	324
total bal. detected	0	4	0	6

Comparing Tables 2 and 3 (see also the summary Table 4), the most striking feature of the results is the similarity of the performance of the two methods, but FDIST has detected slightly more adaptively-selected loci than the Bayesian method, whereas the Bayesian method has detected two more loci under balancing selection. The difference in inferences obtained using θ of 1.0 rather than 0.06 are very small, affecting only the false positive rate, and giving a maximum difference of two out of 900 loci.

An interesting question is to what extent the same loci are identified by both methods and whether there is any advantage in combining the results from both methods. A preliminary examination suggests that there is no advantage in doing so. For example for M-2L-02, at the less stringent level, the number of loci under directional selection that are detected to be so by FDIST is 12. The Bayesian method finds 10, but these include only one locus not detected by FDIST. For the M-2L-05 data set FDIST detects 40 directionally-selected genes, whereas the Bayesian method finds 39 and only five of these are not found by FDIST. In each case, the false positives generated by the two methods have less than 50% overlap, and so it seems likely that any attempt to combine the results of the two methods will find few novel true positives, insufficient to compensate for the additional false positives generated.

Further simulations

The results presented above suggested a further set of simulations to investigate the effects of some of the parameters in the simulation model. We have performed seven additional simulations based on the data set M-2L-05.

There are a number of sources of variation among the replicates: a) the values of F in each population; b) the sampling of population types with respect to selection; c) the evolutionary processes across loci. We first simulated a replicate M-2L-05 data set, with parameter settings unchanged. We then noted the values of subpopulation F obtained, and retained these for subsequent simulations. The changes we made and their outcome are summarised in Tables 5 and 6.

For changes not involving mean population F we retained the same set of values of F, and for the two cases with mean population $F = 0.2$ we used the same set of simulated values of F. In the case where each subpopulation had different N we simply modified N rather than m to obtain the required value of F in each subpopulation.

In this more wide-ranging set of simulations the Bayesian method seems clearly superior to FDIST at the 5% level. In all seven additional simulations the Bayesian method detects more directionally selected loci than does FDIST, and in total it detects 278 directionally selected loci, 50 more than were detected by FDIST, despite a larger false positive rate (0.67% for FDIST versus 0.44% for the Bayesian method). In contrast, FDIST performs very slightly better than the Bayesian method at the (Bayesian) 10% level.

Table 5 Results from the Bayesian regression analysis of the additional simulations, investigating the effects of variations in the parameters underlying dataset M-2L-05. The details are as for Table 2, with the exception that the loci classified as neutral are not reported: they can be inferred from the fact that there are 1000 genes in each simulated dataset

True:	balancing selection		neutral		directional selection	
Classified:	bal.	direc.	bal.	direc.	bal.	direc.
Original	0 (0)	0 (0)	0 (2)	3 (9)	0 (0)	29 (39)
Replicate	0 (0)	0 (0)	0 (7)	10 (16)	0 (0)	35 (42)
Sample size 40	0 (0)	0 (0)	0 (0)	4 (11)	0 (0)	29 (40)
$N = 5 \times 10^3$	0 (0)	0 (0)	1 (14)	1 (7)	0 (0)	37 (38)
$N = 5 \times 10^3$, $T = 5 \times 10^5$	0 (0)	0 (0)	1 (11)	3 (10)	0 (0)	55 (59)
Variable N (instead of m)	0 (0)	0 (0)	0 (4)	2 (5)	0 (0)	37 (45)
$F = 0.2$	1 (3)	0 (0)	1 (10)	0 (3)	0 (0)	31 (40)
$F = 0.2$, $s = 0.1$	1 (3)	0 (0)	4 (16)	1 (5)	0 (0)	54 (58)
Total (exc. original)	2 (6)	0 (0)	7 (62)	21 (57)	0 (0)	278 (322)

Table 6 Results from the FDIST analysis of the additional simulations. The details are as for Table 5

True:	balancing selection		neutral		directional selection	
Classified:	bal.	direc.	bal.	direc.	bal.	direc.
Original	0 (0)	0 (0)	0 (4)	2 (10)	0 (0)	33 (40)
Replicate	0 (0)	0 (0)	1 (4)	4 (15)	0 (0)	29 (50)
Sample size 40	0 (1)	0 (0)	4 (10)	2 (9)	0 (0)	22 (43)
$N = 5 \times 10^3$	0 (0)	0 (0)	4 (8)	1 (12)	0 (0)	29 (36)
$N = 5 \times 10^3$, $T = 5 \times 10^5$	0 (0)	0 (0)	8 (13)	1 (7)	0 (0)	49 (54)
Variable N (instead of m)	0 (0)	0 (0)	1 (6)	4 (10)	0 (0)	25 (39)
$F = 0.2$	0 (0)	0 (0)	6 (6)	0 (3)	0 (0)	27 (43)
$F = 0.2$, $s = 0.1$	0 (1)	0 (0)	5 (8)	1 (7)	0 (0)	47 (62)
Total (exc. original)	0 (2)	0 (0)	29 (55)	13 (63)	0 (0)	228 (327)

Reducing the sample sizes from 100 chromosomes to 40 leads to a smaller reduction for the Bayesian method than for FDIST in the number of directionally-selected genes detected. This supports the intuition that a likelihood-based method should be relatively better for small sample sizes, but a more extensive study is required to obtain conclusive results.

The variability among independent simulations makes it difficult to make strong statements about the effect of particular parameter settings on the overall success of the methods. Allowing the deme size to vary, rather than the immigration rate, has no clear effect. Increasing F_{ST} to 0.2 does increase the number of balancing loci selected, but the increase is small and the false positive rate is also higher. An increased selection coefficient has little effect on the number of balancing-selected loci that are detected, but does increase the number of directionally-selected loci detected. Two simulations lead to a marked increase in the proportion of detected directionally-selected loci: when both the population size and the duration of the simulation is increased 10-fold, and where F_{ST} and the selection coefficients are both doubled. Neither of these results is surprising. In particular, a 10-fold increase in population size means that m is decreased 10-fold, to maintain $F_{ST} = 0.1$, and hence s/m is increased 10-fold. However, the results suggest that this depends on the time over which the simulation is allowed to equilibrate.

Example data set

We reanalyse the *Drosophila melanogaster* allozyme data set of Singh and Rhomberg (1987), studied also in Beaumont and Nichols (1996). The data consists of 61 polymorphic loci, many of them multi-allelic, surveyed in 15 populations around the world. We used the currently distributed version of FDIST with expected F_{ST} taken to be 0.16. The results are summarised in Figure 3 and Table 7.

Five loci are detected as under directional selection by the Bayesian method at the 5% level. At the corresponding level FDIST identifies these five plus an additional three

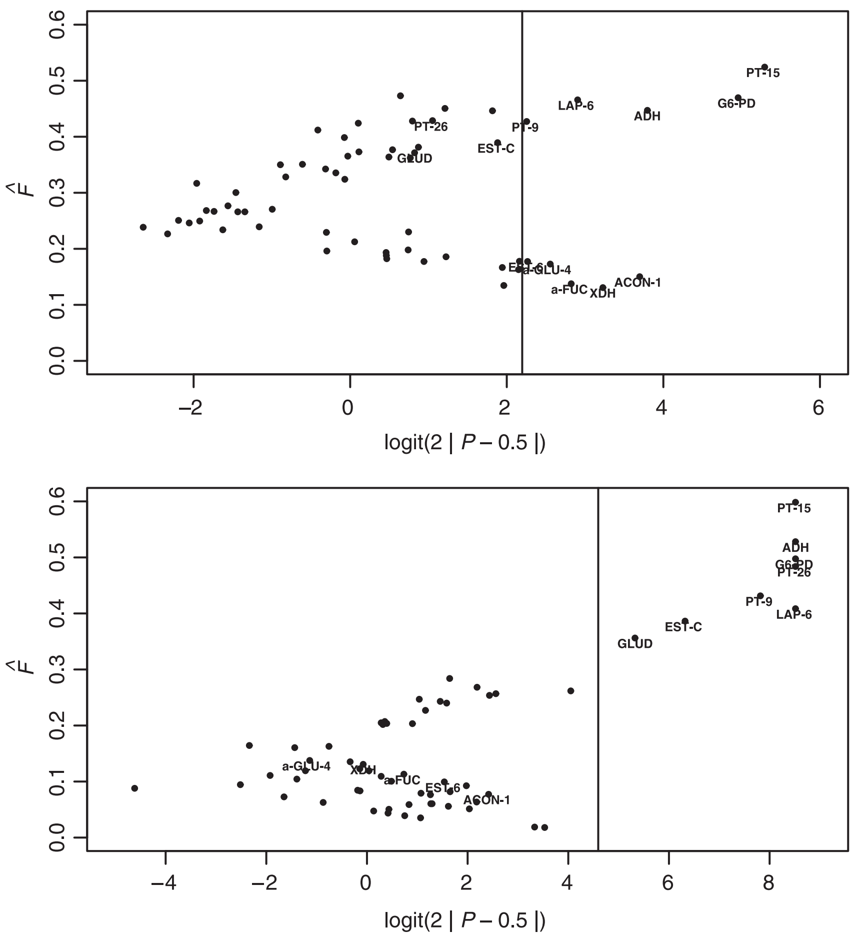

Fig. 3 Results from the analysis of the the *Drosophila* data set using the Bayesian regression method (top) and FDIST (below). An estimate of F_{ST} is plotted against empirical *P*-values for each locus. The vertical bar shows the critical *P*-value used for identifying outlier loci. The thirteen loci identified by either or both methods are labelled.

Table 7 Estimates of F_{ST} and corresponding *P*-values for loci detected as under selection by either the Bayesian regression method (5% level) or the frequentist summary-statistic method (*P*-value 0.05%). Asterisks denote a significant value, and locus labels are also marked with asterisks if both methods indicate a significant result

	F_{ST}		*p*-value	
Locus	Bayes	FDIST	Bayes	FDIST
G6-PD*	0.47*	0.50*	0.00*	0.0000*
ADH*	0.45*	0.53*	0.01*	0.0000*
EST-C	0.39	0.39*	0.07	0.0009*
EST-6	0.18*	0.10	0.95*	0.9115
PT-9*	0.43*	0.43*	0.05*	0.0002*
PT-15*	0.52*	0.60*	0.00*	0.0000*
XDH	0.13*	0.13	0.98*	0.7406
a-FUC	0.14*	0.11	0.97*	0.8374
LAP-6*	0.47*	0.41*	0.03*	0.0001*
ACON-1	0.15*	0.08	0.99*	0.9591
a-GLU-4	0.17*	0.14	0.96*	0.6211
GLUD	0.37	0.36*	0.15	0.0024*
PT-26	0.43	0.48*	0.13	0.0000*

loci. In particular, PT-26 is extremely significant according to FDIST, whereas the Bayesian method assigns little significance to this locus overall. We note below that West Africa is an outlying population at this locus, and outliers are treated very differently by the two methods. Five loci are inferred to be under balancing selection by the Bayesian regression method, whereas no such loci are inferred by FDIST. In the simulations, very few loci were identified by either method as under balancing selection, whether they be true positives or false positives. Thus, if the simulations reflect the genetic structure in the fly data reasonably well, we can be confident that nearly all these five loci have been correctly identified.

We found that there were no significant locus-by-population effects, γ_{ij}, at the 10% level. At the 20% level we detect a positive outlier for PT-26 in the West African sample. This locus has two alleles among all 15 populations and the allele that is rarest in all other populations is fixed in West Africa. Another, negative, outlier at this probability level is AO in a sample from Texas. In this case there are nine alleles among the 15 samples, many of which are absent in individual populations. In the Texan sample, however, eight of the nine alleles are present, and at frequencies that are similar to the average over all the other populations, giving an overall low value of F_{ST}.

It may be argued that the island model assumed here will not well describe the demographic structure of the flies. Beaumont and Nichols (1996) considered both island and colonisation models and showed that they produced very similar distributions of F_{ST} among loci. They also obtained very similar distributions of F_{ST} using a stepping-stone model. In Beaumont and Nichols (1996) the largest differences came when populations with widely differing values of F_{ST} were considered, such as through bottlenecks (in the case of the colonisation model) or small population size or reduced immigration in some demes. FDIST assumes the same value of F_{ST} in each deme. By contrast the Bayesian method assumes different values of F_{ST} for each subpopulation and should therefore be insensitive to this. Wakeley and Aliacar (2001) demonstrate that a metapopulation model with migration, colonization and extinction yields the same two-tier genealogical structure assumed in the present analysis. Thus it would seem that the modelling approach here encapsulates many types of population processes, although it is probably important to take into account variation in F_{ST} among demes when it is large.

Discussion

Until now there has been no systematic examination of the ability of published methods to detect selected loci using simulations based on a model of selection. Thus, for example, the study of Beaumont and Nichols (1996) examined the effect of a number of confounding demographic, mutational, and sampling factors on the distribution of F_{ST} of *neutral* loci, but did not directly test for the ability of their method to detect loci under selection. The results of the simulation study described here are generally encouraging. Under reasonably realistic conditions it is possible to identify the majority of loci under adaptive selection. In our simulations we obtained appreciable discrimination for adaptively selected loci when $s > 5m$. This rule-of-thumb may well hold more generally, but other factors must surely affect discrimination power — for example sample sizes, and number of loci studied. The picture is also unlikely to be so simple for loci under balancing selection. The results suggest that even for very large selection coefficients it is difficult to detect such loci with the neutral mean F_{ST} of around 0.1 used here.

Although overall the Bayesian regression method performs better than the summary-statistic method, FDIST, the superiority is not as marked as we had expected. It should be noted that the Bayesian method estimates many more parameters than FDIST, and, in particular, allows for variable F_{ST} among subpopulations, whereas FDIST assumes them to be identical. It is possible that the presence of outlier loci vitiates the underlying likelihood model sufficiently badly that the information in the data cannot be efficiently used. Ideally, rather than using the multinomial-Dirichlet distribution for neutral loci and attempting to fit it to loci under selection by adjusting values of F_{ST}, it might be preferable to directly utilise diffusion approximations for loci under selection in island models with immigration (Wright, 1969). The challenge here is that this would involve fitting selection (and dominance) coefficients to each allele at each locus, as well as partitioning demes into different groups for each locus, which would substantially complicate the model, although it is, in principle, feasible within a Bayesian framework.

One simpler improvement in this direction that we have explored is to implement Bayesian model selection within the MCMC algorithm. The general idea here is that loci can be either 'neutral', and hence have a zero locus effect, or 'selected', in which case the locus effect has some posterior distribution as in the present model. It is then possible to make statements about the posterior probability that a locus is 'selected' or 'neutral'. Preliminary results suggest that this leads to a small improvement, and further development is ongoing. This approach also leads to important advantages of interpretation, since it deals, through the prior distribution, with the problem of multiple-testing. In contrast, using FDIST a P-value of 0.05 would be suggestive if only a single locus was tested, but would have little weight if 1000 loci were tested because around 50 neutral loci would be expected to have such a P-value or greater. Thus, the cut-off P-value for significance needs to be adjusted for the number of loci being tested, and may also need to be adjusted for any available prior information about particular loci. Further possible improvements to the Bayesian method include extensions to allow for multiple marker types and linked markers.

Overall, although additional testing is needed, and further modelling developments are possible, the results here provide encouragement for studies that attempt to identify loci under selection from whole genome scans.

Acknowledgements

We are grateful to Kevin Dawson for helpful discussions, and to Gordon Luikart for discussion and making available a draft manuscript. We thank Laurent Excoffier and two anonymous reviewers for useful comments. MAB is supported by a Natural Research Council Advanced Fellowship.

References

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

Balding DJ (2003) Likelihood-based inference for genetic correlation coefficients, *Theoretical Population Biology*, **63**, 221–230.

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

Balding DJ, Greenhalgh M, Nichols RA (1996) Population genetics of STR loci in Caucasians, *Intl. J. Leg. Med.*, **108**, 300–305.

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure, *Proceedings of the Royal Society of London, Series B*, **263**, 1619–1626.

Black WC, Baer CF, Antolin MF, DuTeau NM (2001) Population genomics: genome-wide sampling of insect populations. *Annual Review of Entomology* **46**, 441–469.

Bowcock AM, Kidd JR, Mountain JL, Hebert JM, Carotenuto L, Kidd KK, Cavalli-Sforza LL (1991) Drift, admixture, and selection in human evolution: A study with DNA polymorphisms. *Proceedings of the National Academy of Sciences USA*, **88**, 839–843.

Crow JF, Kimura M (1970) *An introduction to population genetics theory*. Harper and Row, New York.

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

Gilks WR, Richardson S, Spiegelhalter DJ (1996) *Markov chain Monte Carlo in practice*. Chapman and Hall, London.

Kayser M, Brauer S, Stoneking M (2003) A genome scan to detect candidate regions influenced by local natural selection in human populations. *Molecular Biology and Evolution*, **20**, 893–900.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics*, **4**, 981–994.

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

Porter AH (2003) A test for deviation from island-model population structure. *Molecular Ecology* **12**, 903–915.

Rannala B, Hartigan JA (1996) Estimating gene flow in island populations. *Genetical Research Cambridge*, **67**, 147–158.

Singh RS, Rhomberg LR (1987) A comprehensive study of genic variation in natural populations of *Drosophila melanogaster*. II. Estimates of heterozygosity and patterns of geographic variation. *Genetics*, **117**, 255–271.

Schlotterer C (2002) A microsatellite-based multilocus screen for the identification of local selective sweeps. *Genetics*, **160**, 753–763.

Sorensen D, Gianola D (2002) *Likelihood, Bayesian, and MCMC methods in quantitative genetics*. Springer, New York.

Storz JF, Nachman MW (2003) Natural selection on protein polymorphism in the rodent genus Peromyscus: evidence from interlocus contrasts. *Evolution* **57**, 2628–2635.

Storz JF, Beaumont MA (2002) Testing for genetic evidence of population expansion and contraction: an empirical analysis of microsatellite DNA variation using a hierarchical Bayesian model. *Evolution* **56**, 154–166.

Vitalis R, Dawson K, Boursot P (2001) Interpretation of variation across marker loci as evidence of selection. *Genetics*, **158**, 1811–1823.

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.

Wilding CS, Butlin RK, Grahame J (2001) Differential gene exchange between parapatric morphs of Littorina saxatilis detected using AFLP markers. *Journal of Evolutionary Biology*, **14**, 611–619.

Wakeley J (1999) Nonequilibrium migration in human history. *Genetics*, **153**, 1863–1871.

Wakeley J, Aliacar N (2001) Gene genealogies in a metapopulation. *Genetics*, **159**, 893–905.

Wright S (1943) Isolation by distance. *Genetics*, **28**, 114–138.

Wright S (1969) *Evolution and the Genetics of Populations. Volume 2: The Theory of Gene Frequencies*. University of Chicago Press, Chicago.

Wu C-I (2001) The genic view of the process of speciation. *Journal of Evolutionary Biology*, **14**, 851–865.

Mark Beaumont in an NERC Advanced Fellow at the University of Reading. His background is in ecological genetics and conservation genetics, and current interests are in the area of population genetic modelling and inference. David Balding is Professor of Statistical Genetics at Imperial College London. His educational formation was in mathematics, and he now works to apply mathematical and statistical methods in may areas of population, evolutionary, and human genetics, as well as forensic applications of DNA profiling.