J. R. Statist. Soc. B (2002)
64, *Part* 4, *pp.* 641–656

A model selection approach for the identification of quantitative trait loci in experimental crosses

Karl W. Broman

Johns Hopkins University, Baltimore, USA

and Terence P. Speed

University of California, Berkeley, USA, and Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia

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

Summary. We consider the problem of identifying the genetic loci (called quantitative trait loci (QTLs)) contributing to variation in a quantitative trait, with data on an experimental cross. A large number of different statistical approaches to this problem have been described; most make use of multiple tests of hypotheses, and many consider models allowing only a single QTL. We feel that the problem is best viewed as one of model selection. We discuss the use of model selection ideas to identify QTLs in experimental crosses. We focus on a back-cross experiment, with strictly additive QTLs, and concentrate on identifying QTLs, considering the estimation of their effects and precise locations of secondary importance. We present the results of a simulation study to compare the performances of the more prominent methods.

Keywords: Bayesian information criterion; Composite interval mapping; Markov chain Monte Carlo methods; Model selection; Quantitative trait loci; Regression

1. Introduction

The identification of the genetic loci that are responsible for variation in traits that are quantitative in nature (such as the yield from an agricultural crop, the number of abdominal bristles on a fruit-fly and the survival time of a mouse following an infection) is a problem of great importance to biologists. The number and effects of such loci (called quantitative trait loci (QTLs)) help us to understand the biochemical basis of these traits, and of their evolution in populations over time. Moreover, knowledge of these loci may aid in the design of selection experiments to improve these traits.

Repeated sibling mating (or, in plants, selfing) of experimental organisms has led to the establishment of panels of well-defined strains. The process of inbreeding has fixed a large number of biomedically (or agriculturally) relevant traits in these strains. If two strains, raised in a common environment, show consistent differences in a trait, we may be confident that the difference has a genetic basis. The genetic loci contributing to such a trait difference may be revealed by performing a series of experimental crosses, of which the simplest is the back-cross.

Address for correspondence: Karl W. Broman, Department of Biostatistics, Johns Hopkins University, 615 North Wolfe Street, Baltimore, MD 21205, USA.
E-mail: kbroman@jhsph.edu

 1369–7412/02/64641

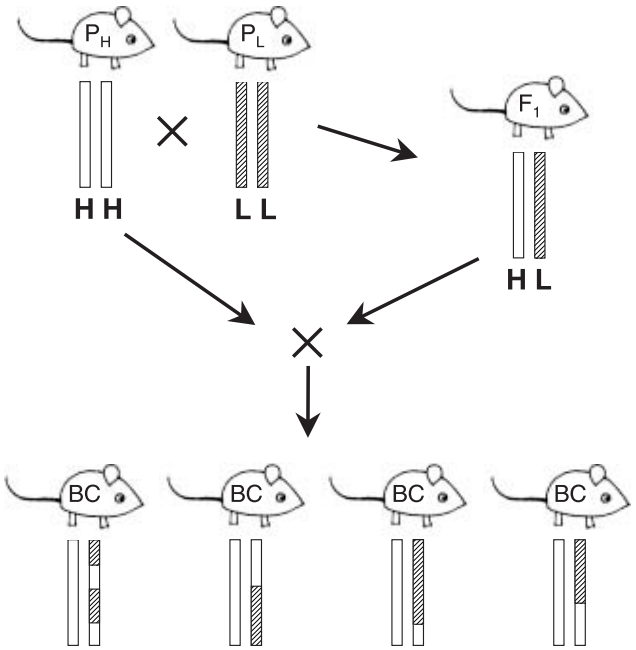

Fig. 1. A back-cross experiment begins with two inbred strains that differ in the trait of interest: the two strains are crossed to produce the F_1-generation, which is then crossed back to one of the parental strains to obtain the back-cross generation, the back-cross generation exhibits genetic variation

In a back-cross (Fig. 1), an investigator chooses two inbred strains that differ in the trait of interest (we shall call these the high (H) and low (L) parental strains). All individuals within an inbred strain are genetically identical and are homozygous at all loci. The two parental strains are crossed to form the first filial (F_1-) generation. The F_1-individuals are also genetically identical, and are heterozygous at loci at which the parental strains differ. The F_1-individuals are crossed to one of the two parental strains (e.g. the H-strain) to obtain the back-cross generation. The back-cross individuals receive one chromosome from the H-strain and one from the F_1. Thus, at each locus, they have genotype either HL or HH. The chromosome received from the F_1-parent is a mosaic of the two grandparental chromosomes, as a result of recombination during meiosis.

The investigator produces a number of back-cross progeny and determines the quantitative phenotype for each individual. Each individual is genotyped at a number of genetic markers (generally 100–300), chosen to cover the genome uniformly. At each marker and for each individual, it is observed whether the F_1-parent transmitted the H- or the L-allele. A genetic map for the marker loci will either be known or estimated on the basis of the current experiment. Such a map specifies the linear order of the marker loci along each chromosome and the distances between markers, measured in genetic distance. The genetic distance between two loci is d centimorgans (cM), if d is the average number of crossovers (points of exchange) in the intervening interval, in 100 products of meiosis.

The objective of the experiment is to identify the genomic regions for which there is an association between the phenotype of a back-cross individual and whether it received the H- or L-allele from the F_1-parent in the region. Although other experiments, such as the intercross, are more commonly used in practice, we focus on the back-cross for simplicity.

Consider a back-cross with n individuals. Let y_i denote the phenotype (trait value) of individual i, and let $x_{ij} = 1$ or $x_{ij} = 0$, according to whether individual i has genotype HH or HL respectively at marker j.

The locations of crossovers in meiosis are often modelled as a Poisson process (an assumption of *no crossover interference*). In this case, the x_{ij} for each chromosome form a Markov chain, with transition probabilities $\Pr(x_{i,j+1} = 1|x_{ij} = 0) = \Pr(x_{i,j+1} = 0|x_{ij} = 1) = r_j$, where r_j is called the *recombination fraction* between markers j and $j + 1$. We further assume that $\Pr(x_{ij} = 1) = \Pr(x_{ij} = 0) = \frac{1}{2}$, in accordance with Mendel's rules.

Imagine that there is a reasonably small number, p, of genetic loci (QTLs) that influence the trait. Let us temporarily suspend the index i for individuals and consider the relationship between an individual's genotypes at the QTLs and its phenotype (trait value). Let $z = (z_1, \ldots, z_p)$, with $z_j = 1$ or $z_j = 0$, according to whether the individual has genotype HH or HL respectively at the jth QTL. In principle, $E(y|z) = \mu_z$ and $\text{var}(y|z) = \sigma_z^2$ are arbitrary functions of z. Generally, we assume that the trait is homoscedastic—that the variance is constant within genotype groups, $\text{var}(y|z) = \sigma^2$. It is often further assumed that the residual variation is normally distributed, that $y|z \sim N(\mu_z, \sigma^2)$.

There remains the possibility that each of the 2^p possible genotypes has a distinct trait mean. However, often it is assumed that the QTLs act additively; we imagine that $E(y|z) = \mu + \Sigma_{j=1}^{p} \beta_j z_j$. Deviation from additivity (i.e. interactions between the QTLs) is called *epistasis* (Frankel and Schork, 1996). Many studies have provided strong evidence for the presence of interactions between QTLs (e.g. Shrimpton and Robertson (1988), Roberts *et al.* (1999) and Shimomura *et al.* (2001)). In this paper, however, we shall focus on the case of strict additivity. This is not because we feel that it is the best approach, but rather because this simple case is still not well solved.

With the assumption of additivity, the aim of QTL mapping is to identify the number and locations of the QTLs. One may further seek interval estimates of QTL locations and estimates of QTL effects; although these are both clearly important, we consider them of secondary interest and focus on the identification of QTLs. In the following section, we describe the current approaches to this problem. In Section 3, we frame the problem as one of model selection and describe an approach for QTL mapping that makes use of a modified version of the Bayesian information criterion BIC (Schwarz, 1978). In Section 4, we present the results of a large computer simulation to assess the performance of several major approaches to QTL mapping.

2. Current approaches

In this section, we describe the commonly used approaches for QTL mapping. For a more extensive review of the statistical methods for QTL mapping, see Doerge *et al.* (1997), Lynch and Walsh (1998) or Broman and Speed (1999).

The simplest approach to identifying QTLs, with data on an experimental cross, is to perform analysis of variance (ANOVA) at each of the marker loci (see Soller *et al.* (1976)). At each genetic marker, we split the back-cross progeny into two groups, according to their genotypes at the marker, and compare the two group phenotype means, by a t-test. Geneticists often prefer to report a LOD score, defined as the (base 10) log-likelihood ratio comparing the hypotheses

(a) the phenotypes in the two groups are normally distributed with distinct means but a common variance and
(b) the phenotypes for all individuals follow a common normal distribution, independent of genotype.

Marker loci giving large LOD scores are indicated to be linked to a QTL.

This approach has several weaknesses. First, if a QTL is not located exactly at a marker, its

effect will be attenuated as a result of recombination between the marker and the QTL. Second, at each genetic marker, we must discard individuals whose genotypes are missing. Third, when the markers are widely spaced, a QTL may be quite far from all markers, and so the power for QTL detection will decrease. Fourth, the approach considers only one locus at a time; in the presence of several QTLs, approaches that model multiple QTLs will give greater power for QTL detection, better separate linked QTLs and allow the examination of interactions between QTLs (though such interactions will not be considered here).

Lander and Botstein (1989) developed *interval mapping*, which overcomes the first three weaknesses of ANOVA at marker loci, described above. The method, which continues to be the most popular approach for QTL mapping, makes use of a genetic map of the typed markers, and, like ANOVA, assumes the presence of a single QTL. Each location in the genome is posited, one at a time, as the location of the putative QTL.

Given the marker genotype data (and assuming no crossover interference), one may calculate the probability that an individual has genotype HH (or HL) at a putative QTL. These QTL probabilities depend only on the genotypes at the flanking markers and may be found in Table 2 of Doerge *et al.* (1997). In interval mapping, one assumes that, given the QTL genotype, the phenotype follows a normal distribution with mean μ_{H} or μ_{L}, according to whether the QTL genotype is HH or HL respectively, and common standard deviation σ. Given the genotypes at the markers flanking the QTL, the conditional phenotype distribution is then a mixture of the two normal distributions, with the conditional QTL genotype probabilities, given the marker genotype data, as mixing proportions. At each position in the genome (or, in practice, at steps of 0.5 cM), one may use a version of the EM algorithm (Dempster *et al.*, 1977) to estimate the three parameters, μ_{H}, μ_{L} and σ, and may calculate a LOD score: the (base 10) log-likelihood ratio, comparing the hypothesis that there is a single QTL at the given location with the hypothesis that there is no QTL anywhere in the genome. The LOD score, as a function of chromosome position, forms a profile log-likelihood. Genomic regions for which the LOD score is large are indicated as harbouring QTLs.

The advantages of interval mapping, over ANOVA at marker loci, are that it makes more complete use of the marker genotype data (making proper allowance for missing data), and it considers positions between markers as putative locations for a QTL, thus providing increased power in the case of widely spaced markers, as well as improved estimates of QTL effects. However in the case of dense genetic markers and relatively complete marker genotype data, interval mapping provides little advantage over ANOVA. Moreover, interval mapping, like ANOVA, makes use of a single-QTL model and so is not ideal in the presence of multiple (especially linked) QTLs.

Both ANOVA at marker loci and interval mapping make use of multiple tests of hypotheses and so require some adjustment for test multiplicity. Much effort has been expended on this problem, the aim being to obtain an approximate genome-wide LOD threshold, defined as the 95th percentile of the distribution of the maximum LOD score, genome wide, under the hypothesis that there are no QTLs (i.e. that the phenotypes are simply normally distributed, independent of the marker data). Lander and Botstein (1989) performed extensive computer simulations to estimate the appropriate LOD threshold for various genome sizes and marker densities, and gave analytical calculations for the case of a very dense marker map. Another approach is to perform a permutation test (Churchill and Doerge, 1994).

As mentioned above, methods that make use of multiple-QTL models can provide increased sensitivity, better separate linked QTLs and allow the examination of interactions between QTLs. The simplest multiple-QTL method is multiple regression, the obvious extension of ANOVA at marker loci. Cowen (1989) appears to be the first to have recommended the use

of multiple regression in this context (see also Whittaker *et al.* (1995)). We shall defer further discussion of this approach to the next section.

Jansen and Zeng independently developed a method which attempts to reduce the multi-dimensional search for identifying multiple QTLs to a one-dimensional search (Jansen, 1993; Jansen and Stam, 1994; Zeng, 1993, 1994). This is done using a hybrid of interval mapping and multiple regression on marker genotypes. One includes other markers (on the same chromosome and on different chromosomes) as regressors while performing interval mapping, in an effort to control for the effects of QTLs in other intervals, so that there will be greater power for QTL detection, and so that the effects of the QTLs will be estimated more precisely. Zeng called this approach composite interval mapping (CIM).

The method is performed as follows. We choose a subset of markers, S, to control for background genetic variation. Then, we perform a genome scan, as in interval mapping. At each locus in the genome, we hypothesize the presence of a QTL and write $y = \mu + \beta z + \Sigma_{j \in S^*} \beta_j x_j + \varepsilon$ where y is the phenotype, $z = 1$ or $z = 0$ according to whether the genotype at the putative QTL is HH or HL, $x_j = 1$ or $x_j = 0$ according to whether the genotype at the jth marker is HH or HL and S^* is a subset of the marker regressors, S, where we exclude any markers that are within, say, 10 cM of the putative QTL. The residual, ε, is assumed to be distributed $N(0, \sigma^2)$.

As in interval mapping, at each locus, a LOD score is calculated, comparing the hypothesis that there is a QTL at the putative locus with the hypothesis that there is not a QTL there, in which case we imagine that all progeny have phenotypes which are normally distributed with mean $\mu + \Sigma_{j \in S^*} \beta_j x_j$ and variance σ^2. The LOD score is plotted as a function of genome position and compared with a genome-wide threshold. (Such a threshold should take into account the selection of the set of marker regressors, S.) Areas of the genome for which the LOD score exceeds a genome-wide threshold are said to contain a QTL.

The key problem with CIM is the choice of the set of markers to use as regressors: using too many markers will increase the variance of the LOD score and thus will decrease the power for QTL detection. Jansen (1993) and Jansen and Stam (1994) used backward elimination with Akaike's information criterion (Akaike, 1969), or a slight variant, to pick the subset of markers. Basten *et al.* (2000), in a manual for the program QTL Cartographer, recommended using forward selection up to a fixed number of markers, and then dropping any markers that are within 10 cM of the putative QTL.

More recently, Kao *et al.* (1999) proposed multiple-interval mapping (see also Zeng *et al.* (1999)), which is much like CIM, but the additional regressors are not required to reside at marker loci. In multiple-interval mapping, Kao *et al.* (1999) have adopted a more standard model selection approach, making use of stepwise selection.

Several other methods have been described, including Bayesian methods (e.g. Satagopan *et al.* (1996), Sillanpää and Arjas (1998), Ball (2001) and Sen and Churchill (2001)) and the use of genetic algorithms (e.g. Carlborg *et al.* (2000)). These approaches are more in line with our view that QTL mapping is a model selection problem.

3. Model selection

We consider a back-cross and assume that the genotype data are complete, and that the genetic markers are sufficiently dense, so that we may dispense with interval mapping, considering only the marker loci as putative locations for QTLs. Let y_i denote the phenotype of individual i, and let $x_{ij} = 1$ or $x_{ij} = 0$ according to whether individual i has genotype HH or HL respectively, at

marker j. We assume the linear model $y_i = \mu + \Sigma_{j=1}^{M} \beta_j x_{ij} + \varepsilon_i$, where the ε_i are independent and identically distributed $N(0, \sigma^2)$.

The problem of identifying QTLs in an experimental cross is one of model selection: in the above linear model, we seek to identify the subset of markers for which $\beta_j \neq 0$. By viewing the problem in this way, we may hope to take advantage of the extensive literature on subset selection in regression. However, much of the model selection literature has focused on the minimization of prediction error, whereas we are not so much interested in prediction as in the identification of an appropriate model.

We split the model selection problem into four distinct parts:

(a) select a class of models,
(b) compare models,
(c) search through the space of models and
(d) assess the performance of a model selection procedure.

We focus here on the class of additive models, though one might also consider linear models with pairwise interactions, or regression trees. The inclusion of the assessment of a procedure's performance as part of the model selection problem may be viewed as unusual but is clearly integral to the problem. Whether we choose to minimize the prediction error or to maximize the number of correctly identified QTLs while controlling the rate of inclusion of extraneous loci at a fixed level, a clearly stated objective is a prerequisite for making informed choices on a model selection procedure.

3.1. Model comparison

Consider the case of a linear model with normally distributed residual variation. Let Γ denote the set of models, with $\gamma \in \Gamma$ written as an M-vector with jth element 1 or 0 according to whether the jth marker is included in the model. Let $|\gamma|$ denote the number of markers in model γ, and let $\mathrm{RSS}(\gamma)$ denote the residual sum of squares after fitting γ by least squares. Imagine that we can fit all possible models.

For models with the same number of regressors k, we choose that with the smallest RSS. We write $\gamma_k = \arg\min_{\gamma:|\gamma|=k}\{\mathrm{RSS}(\gamma)\}$. Thus γ_M is the full model, with all markers included, and γ_0 is the model including no markers. $\mathrm{RSS}(\gamma_k)$ must be non-increasing in k. The key problem is to determine the decrease in RSS that must accompany the inclusion of an additional regressor. Our aim is to balance the errors of excluding important loci and of including extraneous loci.

Classical criteria for choosing the appropriate size of the model include Mallows's C_p and adjusted R^2 (Miller, 1990). In our experience, these criteria tend to include a large number of extraneous regressors and so are unsatisfactory for our purposes.

Two more modern approaches for choosing subsets of regressor variables include cross-validation and the bootstrap. In both of these approaches, an estimate of the mean-squared error of prediction is obtained. The chosen model has the smallest estimated mean-squared error of prediction. Because we are interested in identifying a reasonable model rather than minimizing the prediction error, we have not studied the performance of these approaches.

An additional approach for model comparison is the use of sequential permutation tests, appropriate in the context of a nested sequence of models, such as would be obtained by forward selection (see Doerge and Churchill (1996)). One works from the null model γ_0 to the full model γ_M, performing a permutation test at each step, testing whether the inclusion of an additional regressor is accompanied by a statistically significant decrease in the RSS. The first time that the null hypothesis is not rejected, one stops.

The approach that we favour is to minimize a criterion of the form

$$\Phi(\gamma) = \log\{\mathrm{RSS}(\gamma)\} + |\gamma| D(n)/n$$

where $D(n)$ is some function of the sample size n. (This is equivalent to maximum likelihood with a penalty on the model complexity, since in the case of normally distributed residuals $-(n/2)\log\{\mathrm{RSS}(\gamma)\}$ is the log-likelihood for the model γ.) The choice $D(n) = 2$ gives Akaike's information criterion (Akaike, 1969), whereas $D(n) = \log(n)$ gives BIC (Schwarz, 1978), and $D(n) = \log\{\log(n)\}$ gives the criterion of Hannan and Quinn (1979).

Minimization of $\Phi(\gamma)$ is approximately equivalent to the use of a threshold on the conditional LOD score $(n/2)\log_{10}\{\mathrm{RSS}(\gamma_{k-1})/\mathrm{RSS}(\gamma_k)\}$, the threshold being $D(n)/2\log(10)$. Consider our sequence of models $\gamma_0, \gamma_1, \ldots, \gamma_M$. In the case that $\mathrm{RSS}(\gamma_k)/\mathrm{RSS}(\gamma_{k-1})$ is strictly increasing in k, minimization of $\Phi(\gamma)$ is equivalent to choosing the largest value of k for which $\mathrm{RSS}(\gamma_k)/\mathrm{RSS}(\gamma_{k-1})$ is greater than $-\exp\{D(n)/n\}$. Note that it is sufficient, but not necessary, that the ratios $\mathrm{RSS}(\gamma_k)/\mathrm{RSS}(\gamma_{k-1})$ be strictly increasing, for this equivalence.

When viewed in this way, the criterion appears quite reasonable. Further support lies in the consistency of the resulting procedures. With a fixed number of possible regressors (i.e. genetic markers), and provided that $D(n)/n \to 0$ and $D(n)/\log\{\log(n)\} \to \infty$, the criterion $\Phi(\gamma)$ gives a consistent estimate of the underlying model, meaning that, as the sample size increases, the probability that the correct model is chosen converges to 1 (Rao and Wu, 1989).

We have concentrated on the case $D(n) = \delta \log(n)$, which we call BIC_δ:

$$\mathrm{BIC}_\delta(\gamma) = \log\{\mathrm{RSS}(\gamma)\} + \delta|\gamma| \log(n)/n.$$

Letting $\delta = 1$, this gives BIC. We have found that $\delta = 1$ performs poorly, including far too many extraneous regressor variables. A larger value of δ can give improved results, as a greater penalty on the size of the model leads to the inclusion of fewer extraneous regressors. We shall discuss the choice of δ in Section 3.3.

A further approach to the model selection problem is to place prior probabilities on each of the possible models, as well as on the model parameters, and to use Bayes's theorem to calculate the posterior distribution of the models given the data. If the goal were to pick out just one model, we could choose that which gives the largest posterior probability.

As an example, consider the priors discussed in Smith (1996). We let $y|\gamma, \beta_\gamma, \sigma^2 = X_\gamma \beta_\gamma + \varepsilon$ where $\varepsilon \sim N(0, \sigma^2)$, and use the prior $\beta_\gamma \sim N\{0, c\sigma^2(X_\gamma' X_\gamma)^{-1}\}$, $p(\sigma^2|\gamma) \propto 1/\sigma^2$, $p(\gamma) \propto (c/d)^{|\gamma|/2}$. Let $c \to \infty$, resulting in a diffuse improper prior, and integrate out β_γ and σ^2. Smith (1996) showed that the resulting posterior for γ gives $-(2/n)\log\{p(\gamma|y)\} = \log\{\mathrm{RSS}(\gamma)\} + |\gamma| \log(d)/n$. Taking $D(n) = \log(d)$, we see that the model with maximum posterior is that which minimizes the above-described criterion, $\Phi(\gamma)$. We may consider this as further support for the use of the criterion $\Phi(\gamma)$. The only real justification for a criterion, however, is its performance. We shall study the performance of this criterion in Section 4.

3.2. Search of model space

The number of possible additive models is very large. If there are more than around 40 genetic markers, it will be infeasible to fit each of the $2^{40} \approx 10^{12}$ possible models. Thus, we must form a strategy for searching this large space of models, hopefully so that we may identify the good ones—those that would have been chosen if we could fit all possible models.

In the case that the number of markers is only marginally large, we may use a branch-and-bound procedure to pick out the best subsets of each size, without actually fitting all possible subsets (Miller, 1990), thus gaining considerable savings in computation over an exhaustive

search. However, with many markers, this type of procedure is still not feasible. We are thus led to techniques such as forward selection and backward elimination.

In forward selection, one begins with the null model and builds a nested sequence of models of increasing size; at each step, one adds the marker that gives the greatest decrease in the RSS. In backward elimination, one begins with the full model and builds a nested sequence of models of decreasing size; at each step, one drops the marker that gives the smallest increase in the RSS. These two sequences of models may be quite different.

Forward selection and backward elimination provide great savings in computation, since only a small fraction of the possible models are fitted. This saving is also a cost, however: we see only a fraction of the possible models, and we might not see the good ones. With forward selection, once a regressor has been included, it will be retained in all further models. With backward elimination, once a regressor has been dropped, it will be excluded from all further models.

Stepwise selection procedures, which iteratively add or subtract regressors, are commonly used for subset selection in regression. In such procedures, the 'stopping rule', for choosing the appropriate model size, is generally intertwined with the search through the model space. We prefer to keep separate the criteria for model comparison and the procedures for model search.

Forward selection has a particularly bad reputation. One can find quite simple situations in which forward selection will miss the correct model, even when the sample size is extremely large. This occurs as a result of collinearity in the regressor variables, where a regressor that does not belong in the model mimics a set of regressors that do. Backward elimination does not suffer from this problem, at least with large samples. An and Gu (1985) showed that, when using BIC, and in the case of a fixed number of regressors, the backward elimination procedure is consistent, meaning that, as the sample size increases, the probability of choosing the correct model converges to 1. The result also applies to BIC_δ. Forward selection, however, is *overconsistent*; in the limit, the selected model will contain the true model, but may also include additional, extraneous, regressors.

However, in the situation considered here, the regressors are genetic markers that, under the assumption of no crossover interference, form a Markov chain. Given the genotypes at any one marker, the genotypes at markers to its left are conditionally independent of the genotypes at markers to its right. This suggests that the sort of collinearity among regressors that may cause forward selection to include extraneous regressors, even with large samples, will not be a problem in the context of QTL mapping. Indeed, Broman (1997) showed that, in the case of a strictly additive QTL model, forward selection with BIC_δ is consistent. In computer simulations, Broman (1997) found that forward selection also worked reasonably well in samples of more typical size. We shall see below, however, that forward selection can still suffer from the inclusion of extraneous loci.

A different approach to searching the space of models is to use a randomized algorithm, such as a Markov chain Monte Carlo (MCMC), simulated annealing or a genetic algorithm. We shall consider only the MCMC method, in which one places a prior on each model and on the model parameters, and then forms a Markov chain whose stationary distribution is the posterior distribution of the models given the data. Simulations of the Markov chain give a sequence of models (a sort of walk through the space of models) which will, eventually, spend more time at models that have a high posterior probability. Whereas this method is usually used to obtain an approximation of the posterior distribution, and especially to find the region with highest posterior, here we consider it simply as a method for searching the space of models.

There are several standard ways to form a Markov chain with the desired stationary distribution. With the prior discussed above (Section 4.1), Smith (1996) used a Gibbs sampler to obtain

a Markov chain whose stationary distribution satisfies $-(2/n)\log\{p(\gamma|y)\} = \log\{\mathrm{RSS}(\gamma)\} + |\gamma|\log(d)/n$. The method, which is much like stepwise selection, is as follows. First, pick an initial model $\gamma^{(0)}$ (e.g. the null model or the model obtained by forward selection). Then, at step t, cycle through the M different markers; for each $j = 1, \ldots, M$, draw $\gamma_j^{(t)}$ from the distribution $p(\gamma_j|\gamma_{-j}^{(t)}, y)$ where $\gamma_{-j}^{(t)}$ is composed of all the elements of γ, except for γ_j, at their current values. For $i < j$, it contains the γ_i for the current step t and, for $i > j$, it contains the γ_i for the previous step $t - 1$. For the posterior written above,

$$\Pr(\gamma_j = 1|\gamma_{-j}, y) = \frac{\mathrm{RSS}(\gamma_1, \ldots, \gamma_{j-1}, 1, \gamma_j, \ldots, \gamma_M)^{-n/2}}{\mathrm{RSS}(\ldots, 1, \ldots)^{-n/2} + \sqrt{d}\,\mathrm{RSS}(\ldots, 0, \ldots)^{-n/2}}.$$

The most important characteristic for the Markov chain is that it mixes well—that it travels through the space of models with relative ease, not becoming stuck in local modes. We have implemented the above MCMC sampler and have found that it works well. In 1000 steps of the chain, it will visit around 300–500 distinct models and will almost always visit the best of those models (i.e. that giving the largest posterior probability) within the first 100 steps.

3.3. Recommended approach

It is best to consider model comparison and model search separately. One should devote the greatest effort to the formulation of a criterion for model comparison, as this is the most difficult aspect of model selection. It is helpful to imagine that we could examine all possible models. In choosing between them, we must balance the errors of excluding important regressors and including extraneous ones. The appropriate balance of these errors will vary according to the goals of the experiment, and so the appropriate criterion for comparing models should also vary.

We prefer the BIC_δ criterion, for its simplicity and its reasonable interpretability. One approach for choosing an appropriate δ is through the connection between BIC_δ and conditional LOD scores: we may choose the value of δ that corresponds to a genome-wide LOD threshold for interval mapping or ANOVA at marker loci. Let L denote such a threshold (the 95th percentile of the maximum LOD score, genome wide, under the hypothesis that there are no QTLs); then we may let $\delta = 2L/\log_{10}(n)$. Use of the derived BIC_δ criterion should, in the case of no QTLs, result in the selection of one or more extraneous loci, approximately 5% of the time. In the presence of QTLs, the rate at which extraneous loci are included is not necessarily under control, though we show in the next section, through computer simulations, that it performs adequately. Of course, such a choice of δ results in a procedure that is not consistent, as the rate of inclusion of extraneous loci will continue to be 5%, in spite of increasing sample size. If one desires a smaller false positive rate, a larger value of δ should be chosen.

The search of model space is a matter of exhausting or repetitive work. More extensive searches are better, though the improvement may not be sufficient to compensate for the increased computation. Forward selection and backward elimination are quick and simple to implement. The MCMC sampler described above is also simple to implement, and the increase in computation may be sufficiently small to justify its use.

4. Simulations

Computer simulation studies are crucial for understanding the relative performance of different model selection procedures, because such procedures aretoo complex to be assessed by analy-

tical means, at least in the situations in which they would be used in practice. It is unfortunate that large scale computer simulations are not routinely included in statistical methodological papers on QTL mapping. Many researchers have used simulations to illustrate methods for finding QTLs, but most have either presented the results on a single simulation replicate or data set or considered only very simple situations. In some cases, the value of a new approach has simply been declared on the basis of increased complexity.

Any simulation study is necessarily incomplete and artificial. Real QTL experiments do not have equally spaced markers and exhibit complex patterns of missing genotype data. The number, effects and locations of QTLs are not known; the QTLs have effects of varying size, and the QTLs may interact in complex ways. The simulation study reported here includes a small number of additively acting QTLs located exactly at marker loci and having equal-sized effects; the genetic markers were equally spaced and the genotype data were complete. Although this study may be criticized as not being sufficiently realistic, we believe that it is among the most complete and realistic such studies, and that the results are of considerable value for the assessment of the performance of the QTL mapping methods included.

4.1. Methods

We simulated a back-cross obtained from inbred lines, composed of 100, 250 or 500 progeny, with nine chromosomes, each of length 100 cM and having 11 equally spaced markers (at a spacing of 10 cM). The recombination process was assumed to exhibit no crossover interference. The marker data were complete and without errors. For each sample size, we performed 2000 simulation replicates.

We considered a model with seven QTLs of equal effect, 0.76, with all QTLs positioned exactly at marker loci. Two QTLs were located at markers 4 and 8 on chromosome 1 (separated by 40 cM), linked in *coupling* (i.e. their effects had the same sign). Two QTLs were located at markers 4 and 8 on chromosome 2, linked in *repulsion* (i.e. their effects had opposite signs). Three further QTLs were located at markers 6, 4 and 1, on chromosomes 3, 4 and 5 respectively. Four chromosomes contained no QTLs. The environmental variation followed a normal distribution with standard deviation $\sigma = 1$. As a result, the *heritability* of the trait (the proportion of the phenotypic variance attributable to the QTLs) was 50%.

We compared seven methods for identifying QTLs: ANOVA at marker loci, a simplified version of CIM, forward selection with permutation tests and the BIC_δ criterion with forward selection, backward elimination, forward selection followed by backward elimination, and MCMC sampling. Interval mapping was not considered, because it provides little improvement in power over simple ANOVA in the case of a relatively dense marker map and a moderate number of progeny, and because it would require a great increase in computation time.

For CIM, we used forward selection up to either 3, 5, 7, 9 or 11 markers to obtain the set of regressors, and we limited the search for QTLs to marker loci. With both ANOVA and CIM, we obtained genome-wide LOD thresholds (specific for the case of nine chromosomes of length 100 cM with 11 equally spaced markers on each chromosome) by performing 50000 simulations under the null hypothesis of no QTLs. The estimated thresholds were obtained as the 95th percentile of the maximum LOD score across all markers and appear in Table 1. In addition, for these methods, we required that the LOD score dropped by at least 1.5 between 'peaks' before we declared that two QTLs were identified. This value was obtained empirically and may not be ideal. Note that this prevents these methods from identifying adjacent markers as QTLs.

The value of δ for the BIC_δ criterion was chosen to correspond to the LOD threshold for ANOVA in Table 1: $\delta = 2\ \mathrm{LOD}/\log_{10}(n)$. For $n = 100, 250, 500$, the value of δ was 2.56, 2.10

Table 1. Estimated LOD thresholds, based on 50 000 simulation replicates, for a back-cross with nine chromosomes, each 100 cM long and containing 11 equally spaced markers†

n	*ANOVA*	*Thresholds from CIM for the following numbers of markers:*				
		3	*5*	*7*	*9*	*11*
100	2.56	3.50	4.12	4.64	5.13	5.60
250	2.52	3.23	3.56	3.77	3.95	4.09
500	2.50	3.15	3.38	3.51	3.60	3.67

†Standard errors are approximately 0.01.

and 1.85 respectively. The permutation tests used 1000 replicates with $\alpha = 0.05$. In the use of forward selection, a maximum of 25 markers were considered. Backward elimination was begun at the full model. We further applied forward selection up to a model with 25 markers followed by backward elimination; the model with the minimum value of BIC_δ, among all fitted models, was chosen. For the MCMC method, we used 1000 steps of the sampler described in Section 3.2 and chose the model giving the minimum BIC_δ value. In the first 1000 of the 2000 simulation replicates performed, the MCMC sampler was started at the null model; in the second 1000 replicates, the sampler was started at the model obtained by forward selection with BIC_δ. The results were indistinguishable and thus were pooled.

The result of the application of each method was a set of marker loci indicated to be at or near QTLs. In assessing the results, we defined a chosen marker to be correctly identifying a QTL if it was within 10 cM of a QTL (i.e. if the marker was at or adjacent to the QTL); otherwise it was deemed extraneous. If more than one chosen marker were within 10 cM of the same QTL, one was called correct and the others were called extraneous.

4.2. Results

The results of the simulations are displayed in Figs 2 and 3. In terms of the number of QTLs correctly identified (upper panels in Fig. 2), MCMC sampling with the BIC_δ criterion performed best, though it was only slightly better than forward selection, and it was essentially indistinguishable from forward selection followed by backward elimination. Forward selection with BIC_δ was slightly better than with permutation tests. Backward elimination performed poorly at the smallest sample size. CIM performed slightly worse than forward selection with BIC_δ. CIM performed best when the number of markers used as regressors was 7, the number of simulated QTLs; a considerable attenuation of power was accompanied by a choice of too many or too few markers to serve as regressors in CIM. ANOVA, as might be expected, performed rather poorly for this model of multiple QTLs.

Fig. 3 provides greater detail on the number of QTLs that are correctly identified, giving separate results on the QTLs linked in coupling (upper panels), the QTLs linked in repulsion (centre panels) and the three other QTLs (lower panels). The inferior performance of ANOVA and of CIM with three or five markers serving as regressors, in the cases $n = 250$ or $n = 500$, was due largely to their poor ability to detect the QTLs linked in repulsion. In the case $n = 250$, forward selection with permutation tests also performed poorly on the QTLs linked in repulsion, because, for this method, forward selection was stopped when the first test in the

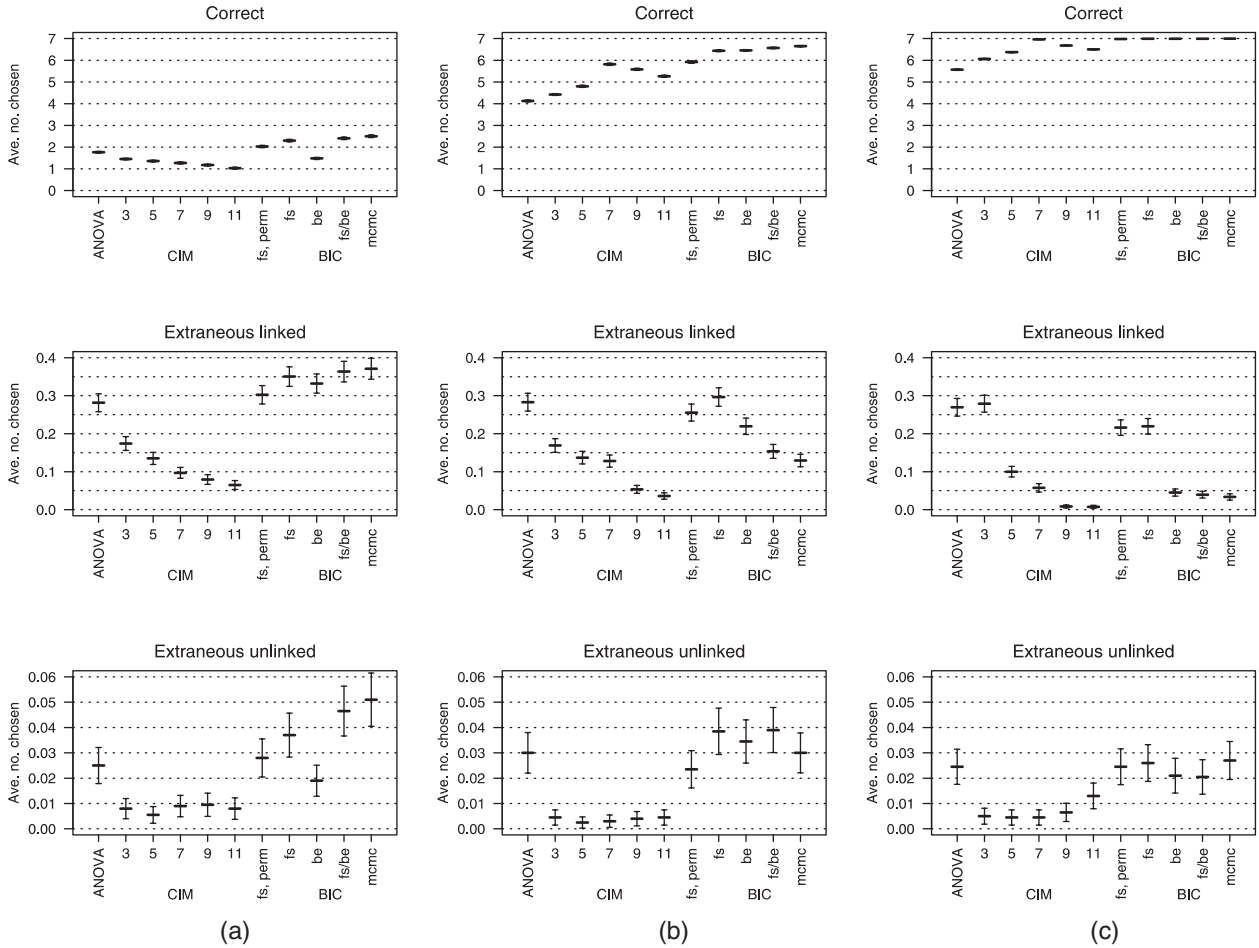

Fig. 2. Results of simulations of a back-cross with (a) $n = 100$, (b) $n = 250$ or (c) $n = 500$ individuals, under a seven-QTL model, including two QTLs linked in coupling (effects of the same sign) and two QTLs linked in repulsion (effects of opposite sign) (the upper panels indicate the average number of QTLs that are correctly identified; the centre panels indicate the average number of extraneous loci that were linked to a QTL; the lower panels indicate the average number of extraneous loci that were not linked to a QTL): the methods considered were ANOVA, CIM, preceded by forward selection up to a fixed number of loci, forward selection with permutation tests and the BIC_δ criterion with forward selection, backward elimination, forward selection followed by backward elimination, and MCMC sampling

sequence was not rejected, while the loci in repulsion appear important only when considered jointly.

All the methods except CIM chose a rather high proportion of extraneous loci linked to a QTL (centre panels in Fig. 2). For the MCMC sampling, backward elimination, and forward selection followed by backward elimination, this effect went away at high sample sizes, but ANOVA and forward selection continued to include a high proportion of extraneous linked loci even at $n = 500$. For the case $n = 100$, these extraneous linked loci were largely imprecisely localized (but correctly identified) QTLs. If a QTL was considered to be correctly identified when a marker within 20 cM was chosen (*versus* the 10 cM criterion used to create Fig. 2), the proportion of extraneous linked loci was reduced from around 30% to around 10%. For the case $n = 500$, however, these loci were truly extraneous. Forward selection identified all the QTLs but also included additional marker loci; if a more complete search of the model space was undertaken (as in the MCMC method or by following forward selection with backward elimination), these additional loci were excluded.

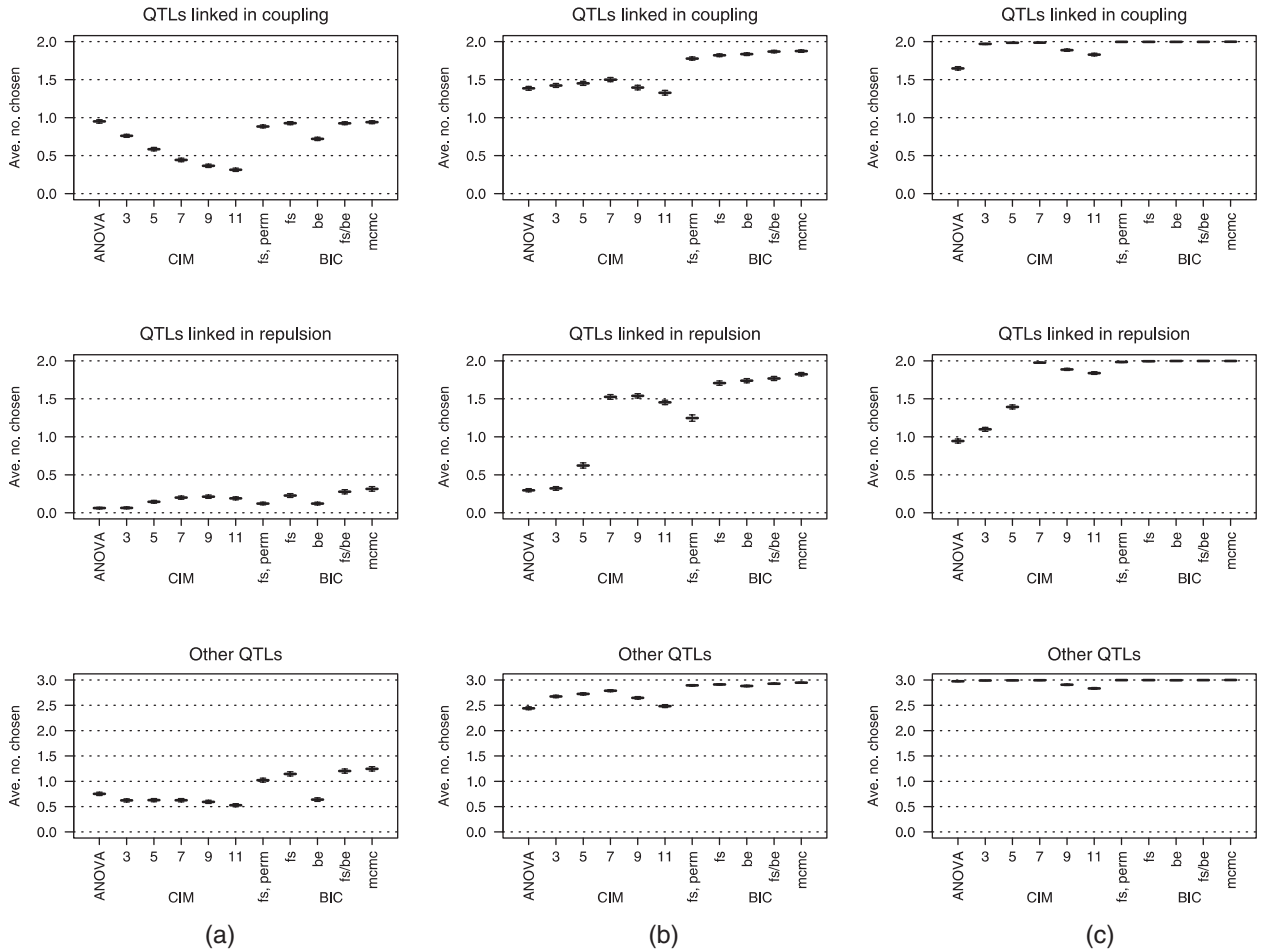

Fig. 3. Detailed results on the upper panels of Fig. 2, displaying the average number of QTLs correctly identified in a seven-QTL model, including two QTLs linked in coupling and two QTLs linked in repulsion (the upper, centre and lower panels indicate the average number of QTLs correctly identified, for the two QTLs linked in coupling, the two QTLs linked in repulsion and the three other QTLs respectively): (a) $n = 100$; (b) $n = 250$; (c) $n = 500$

The lower panels in Fig. 2 display the proportion of extraneous loci not linked to a QTL. CIM was quite conservative, delivering only about 0.5% of such extraneous unlinked loci, whereas the other methods all delivered approximately 2–3% of such extraneous unlinked loci (which might be expected, given that a 5% genome-wide threshold was used, and four out of the nine chromosomes contained no QTLs). Under the null hypothesis of no QTLs, all these methods will identify at least one extraneous QTL, 5% of the time. These results illustrate that the performance of a procedure in the presence of QTLs may be rather different from what might be expected, given its behaviour under the null hypothesis of no QTLs.

In summary, MCMC sampling with the BIC_δ criterion performed best. The key advantage of MCMC sampling over forward selection was the elimination of extraneous linked loci, which forward selection included at a reasonably high rate at $n = 500$. The same benefit could be obtained by following forward selection with backward elimination. CIM performed only slightly worse than forward selection and did not suffer from the inclusion of extraneous loci, but a correct choice of the number of markers included as regressors was extremely important; the use of too few or too many such marker regressors was accompanied by a loss of power.

5. Discussion

There are four key points that we wish to make in this paper. First, QTL mapping is best viewed as a problem of model selection. Second, the comparison of models is the most difficult part of the model selection problem. Third, large scale computer simulation studies are important for understanding the relative performance of different model selection procedures, and they should be routinely included in papers describing new approaches for QTL mapping. Fourth, more refined procedures will not necessarily provide sufficiently improved results to justify their added complexities and increased computational requirements; the choice of adopting such procedures should be based on honest estimates of the gains in performance that they provide.

We have focused on the problem of identifying QTLs. Although we have not considered the precise localization of QTLs and the estimation of QTL effects, it cannot be denied that these are important (and not always straightforward) problems in practice. However, the selection of the number and approximate locations of QTLs is a prerequisite, and so we are justified in considering only this essential part of the problem. In the case of inbred strains of mice, the step following localization of a QTL is frequently the creation of a congenic strain incorporating the relevant region onto a desired background. This is done in the hope of recovering the phenotype in this strain; only when this happens will the search for actual genes begin.

We have discussed back-cross designs because they are the simplest, but of course our study could and should be repeated for other common designs such as the intercross. Although we expect the broad conclusions to be similar to the case considered here, the work needs to be done.

We have also focused on the unrealistic situation in which QTLs are located exactly at marker loci, and in which markers are densely and regularly spaced and exhibit no missing genotype data. This was done to make plain the essence of the QTL mapping problem, and to illustrate the performance of several approaches to the problem. It may happen that our quantitative conclusions change if the markers are not densely and regularly spaced, though we do not expect this. In the case of missing genotype data and/or gaps between markers, the multiple-regression approach that we have considered will not be appropriate. One may confront this missing data problem by multiple-interval mapping (Kao *et al.*, 1999; Zeng *et al.*, 1999) or multiple imputation (Ball, 2001; Sen and Churchill, 2001); the model selection issues that we have discussed remain the essence of the problem.

We considered the case of QTLs acting strictly additively. Of course, one cannot know in advance that this will be appropriate, and a growing number of experiments provide strong evidence for the presence of interactions between QTLs. Thus we recommend that, in practice, one pursues the possibility of interactions. This may be done by the inclusion of pairwise interactions in a linear model, or the consideration of tree-based models. The BIC_δ criterion will probably remain useful in this situation, though a larger value for δ (a larger penalty for model complexity) may be required, and one may wish to place different penalties on main effects and interactions. More complex, randomized search algorithms, such as an MCMC sampler, may be especially valuable for the search of these expanded spaces of models.

We have recommended the use of the BIC_δ criterion, with the value of δ chosen by the approximate correspondence between BIC_δ and a genome-wide threshold on the LOD score. We hope that this is not interpreted as a recommendation for strict adherence to thresholds. In particular, 5% significance thresholds may not be in accordance with the goals of the experimenter. A consideration of the models selected with larger and smaller values of δ provides valuable information regarding the strength of evidence for QTLs.

Our computer simulations demonstrate the value of the BIC_δ criterion. The MCMC sampler performed best. Forward selection was nearly as good, and its tendency to include extraneous loci could be alleviated by following forward selection with backward elimination. CIM performed reasonably well, though it has the disadvantage of requiring a choice of the number of markers to serve as regressors. The sensitivity of the results of CIM to this choice suggests that, although its conversion of a multidimensional into a single-dimensional search is enviable, the approach should not be recommended. There are various schemes for selecting variables in CIM, and it may be true that one of these, different from the one that we have used, gives generally better results and invalidates this conclusion.

The improved performance of these multiple-QTL approaches, over ANOVA at marker loci, is clear but is not nearly as fantastic as we might have hoped. It is difficult to deny that a genome scan by interval mapping can give quite reasonable results. The advantages of multiple-QTL methods are the better separation of linked QTLs and the ability to examine interactions between QTLs.

Acknowledgements

Dursun Bulutoglu and Śaunak Sen generously provided comments to improve the manuscript.

References

Akaike, H. (1969) Fitting autoregressive models for prediction. *Ann. Inst. Statist. Math.*, **21**, 243–247.
An, H. and Gu, L. (1985) On the selection of regression variables. *Acta Math. Appl. Sin.*, **2**, 27–36.
Ball, R. D. (2001) Bayesian methods for quantitative trait loci mapping based on model selection: approximate analysis using the Bayesian information criterion. *Genetics*, **159**, 1351–1364.
Basten, C. J., Weir, B. S. and Zeng, Z.-B. (2000) *QTL Cartographer, Version 1.14.* Raleigh: North Carolina State University.
Broman, K. W. (1997) Identifying quantitative trait loci in experimental crosses. *PhD Dissertation*. Department of Statistics, University of California, Berkeley.
Broman, K. W. and Speed, T. P. (1999) A review of methods for identifying QTLs in experimental crosses. *IMS Lect. Notes Monogr. Ser.*, **33**, 114–142.
Carlborg, O., Andersson, L. and Kinghorn, B. (2000) The use of a genetic algorithm for simultaneous mapping of multiple interacting quantitative trait loci. *Genetics*, **155**, 2003–2010.
Churchill, G. A. and Doerge, R. W. (1994) Empirical threshold values for quantitative trait mapping. *Genetics*, **138**, 963–971.
Cowen, N. M. (1989) Multiple linear regression analysis of RFLP data sets used in mapping QTLs. In *Development and Application of Molecular Markers to Problems in Plant Genetics* (eds T. Helentjaris and B. Burr), pp. 113–116. Cold Spring Harbor: Cold Spring Harbor Laboratory.
Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc.* B, **39**, 1–38.
Doerge, R. W. and Churchill, G. A. (1996) Permutation tests for multiple loci affecting a quantitative character. *Genetics*, **142**, 285–294.
Doerge, R. W., Zeng, Z.-B. and Weir, B. S. (1997) Statistical issues in the search for genes affecting quantitative traits in experimental populations. *Statist. Sci.*, **12**, 195–219.
Frankel, W. N. and Schork, N. J. (1996) Who's afraid of epistasis? *Nat. Genet.*, **14**, 371–373.
Hannan, E. J. and Quinn, B. G. (1979) The determination of the order of an autoregression. *J. R. Statist. Soc.* B, **41**, 190–195.
Jansen, R. C. (1993) Interval mapping of multiple quantitative trait loci. *Genetics*, **135**, 205–211.
Jansen, R. C. and Stam, P. (1994) High resolution of quantitative traits into multiple loci via interval mapping. *Genetics*, **136**, 1447–1455.
Kao, C.-H., Zeng, Z.-B. and Teasdale, R. D. (1999) Multiple interval mapping for quantitative trait loci. *Genetics*, **152**, 1203–1216.
Lander, E. S. and Botstein, D. (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, **121**, 185–199.
Lynch, M. and Walsh, B. (1998) *Genetics and Analysis of Quantitative Traits*, ch. 15. Sunderland: Sinauer.
Miller, A. J. (1990) *Subset Selection in Regression*. New York: Chapman and Hall.

Rao, C. R. and Wu, Y. (1989) A strongly consistent procedure for model selection in a regression problem. *Biometrika*, **76**, 369–374.
Roberts, L. J., Baldwin, T. M., Speed, T. P., Handman, E. and Foote, S. J. (1999) Chromosomes X, 9, and the H2 locus interact epistatically to control Leishmania major infection. *Eur. J. Immunol.*, **29**, 3047–3050.
Satagopan, J. M., Yandell, B. S., Newton, M. A. and Osborn, T. C. (1996) A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. *Genetics*, **144**, 805–816.
Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.
Sen, Ś. and Churchill, G. A. (2001) A statistical framework for quantitative trait mapping. *Genetics*, **159**, 371–387.
Shimomura, K., Low-Zeddies, S. S., King, D. P., Steeves, T. D., Whiteley, A., Kushla, J., Zemenides, P. D., Lin, A., Vitaterna, M. H., Churchill, G. A. and Takahashi, J. S. (2001) Genome-wide epistatic interaction analysis reveals complex genetic determinants of circadian behavior in mice. *Genome Res.*, **11**, 959–980.
Shrimpton, A. E. and Robertson, A. (1988) The isolation of polygenic factors controlling bristle score in *Drosophila melanogaster*: I, Allocation of third chromosome sternopleural bristle effects to chromosome sections. *Genetics*, **118**, 437–443.
Sillanpää, M. J. and Arjas, E. (1998) Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics*, **148**, 1373–1388.
Smith, M. S. (1996) Nonparametric regression: a Markov chain Monte Carlo approach. *PhD Dissertation*. University of New South Wales, Sydney.
Soller, M., Brody, T. and Genizi, A. (1976) On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. *Theoret. Appl. Genet.*, **47**, 35–39.
Whittaker, J. C., Curnow, R. N., Haley, C. S. and Thompson, R. (1995) Using marker-maps in marker-assisted selection. *Genet. Res.*, **66**, 255–265.
Zeng, Z.-B. (1993) Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proc. Natn. Acad. Sci. USA*, **90**, 10972–10976.
———(1994) Precision mapping of quantitative trait loci. *Genetics*, **136**, 1457–1468.
Zeng, Z.-B., Kao, C.-H. and Basten, C. J. (1999) Estimating the genetic architecture of quantitative traits. *Genet. Res.*, **74**, 279–289.

J. R. Statist. Soc. B (2002)
64, *Part* 4, *pp.* 737–775

Discussion on the meeting on 'Statistical modelling and analysis of genetic data'

David J. Balding (*Imperial College School of Medicine, London*)

I extend my apologies to the authors that an unavoidable commitment arising unexpectedly in the 18 hours before the meeting robbed me of my final preparation, so that my comments at the meeting were poorly presented. I shall try to do a better job in this written version, and to leave enough space I shall omit my introductory comments about the role of statisticians in bioinformatics.

Broman and Speed

I have no special expertise in this area, and so the authors should feel more than usually free to ignore my own views. However, I did consult widely among statistical researchers who have worked in quantitative tract locus (QTL) mapping, and I found little enthusiasm for the paper. There is not much new here for the general statistician and, as far as I could detect, there is also little of interest for the applied researcher.

I have no major quibble with the paper's analyses. I would have liked to have seen more attention paid to the multiple-QTL mapping method of Jansen (1994) rather than, or in addition to, composite interval mapping (CIM), since it attempts to tackle the 'key problem with CIM'. Given the recent interest in Bayesian methods for QTL mapping it would have been useful to examine them. However, my main complaint is with the paper's emphasis, summarized by the four 'key points that [the authors] wish to make in this paper', listed at the start of Section 5. The authors are prominent statisticians who, I believe, do not work principally in QTL mapping. In emphasizing model selection, model comparison and simulation studies, as if they were novel ideas in the field, and drawing four pedestrian conclusions, they seem to me to condemn the field as a statistical backwater. No doubt there is bad statistical practice in this field as in many others, and academic researchers certainly should not shy away from criticism where it is due. However, I do not think that the implied criticism is fair here.

The authors display little awareness of recent developments in the field: there are few citations appearing since Broman and Speed (1999) that covered similar ground and came to similar conclusions. I have been made aware of numerous recent papers that were not cited by the authors that deal extensively with model selection and comparison, and/or which describe substantial simulation studies. I believe that others with more expertise than I have will contribute the key missing references; I note Piepho and Gauch (2001) and the review of Jansen (2001). If there are weaknesses in these analyses then I would have been very happy for the present authors to criticize them, but instead they ignore these and other papers and create the false impression that the model selection viewpoint is novel in this setting, and that thorough simulation studies are rare.

The authors tackle a simplified version of the QTL mapping problem, for example assuming the backcross design and QTLs at markers. I agree that such simplifications are often worthwhile, but inevitably it is not clear whether or not conclusions reached in this setting will carry over to the questions of direct interest: the advantageous features of more sophisticated methods may not appear in the simpler setting.

In short I would have preferred that the authors had tackled a more challenging problem (weakening the additivity assumption, Bayesian methods, varying QTL effects and measures of confidence) and drew more substantial conclusions about methods for QTL mapping.

Fearnhead and Donnelly

I congratulate the authors on their excellent paper which makes great progress on an important problem. The intuition underlying the validity of the approximations is well explained, but I wonder whether the authors can give more explanation of the reasons that the approximations lead to substantial computational savings. In particular, I invite them to speculate on whether their approximations are likely to be helpful for approaches to the problem based on Markov chain Monte Carlo sampling.

I have just one query relating to Section 5.1. The CpG dinucleotides mutate more frequently than other dinucleotides, and hence they are substantially under-represented in much of the genome. The wording '... we identified CpG doublets in the data and allowed these to mutate at a rate that was 10 times higher

 1369–7412/02/64737

. . .' suggests that sites which are *currently* CpG are assigned a higher mutation rate. I do not believe that this is a sensible modelling assumption, and I would welcome the authors' reassurance that my reading of this sentence is incorrect.

Larget, Simon and Kadane

The idea of using the information in mitochondrial DNA gene order to help to infer phylogenies is a good one; it has been raised before, but the present paper seems to give by far the most serious attempt to implement the idea in a statistical modelling framework.

The paper is disappointingly brief, and the most obvious lack is of results. The 'movie' shown at the meeting was an excellent innovation. However, in the paper the authors have barely attempted to tackle the challenge of reporting useful summaries of posterior distributions over tree space. Possibly the reason is that what results are available are disappointing: posterior probabilities for inferred relationships are generally low, and there are the awkward results for the squid.

It seems to me that, not surprisingly, the authors' enthusiasm for their new approach leads them to overplay its strengths and to downplay its drawbacks. Mitochondrial DNA gene order is essentially only one character, whereas sequence data offer replication over sites and hence vastly more information, as well as the potential for investigating for example the effects of variable rates and/or selection. The assumption that selection is not an issue for mitochondrial DNA gene order seems highly questionable to this non-expert. Do we have any idea whether a non-observed gene order is viable?

Nevertheless, I agree that the authors' simple model for the evolution of gene order is a good place to start. The paper overall represents, in my view, a substantial advance and a good start to what I hope will be a continuing research programme.

Nicholson, Smith, Jónsson, Gústafsson, Stefánsson and Donnelly

There is much good work in this paper but its presentation seems to me inadequate. The authors have replaced the beta distribution, in an established model of subpopulation variation, with a truncated normal distribution. The authors rightly criticize the standard justification for the beta–binomial model, in terms of an equilibrium maintained by migration, yet their own restrictive model seems no less implausible *a priori*, including as it does a random-mating homogeneous ancestral population and isolation since subdivision. The authors fail to give a serious discussion of the relative merits of the beta–binomial and the normal–binomial models, and their suitabilities for different data sets which might arise, and astonishingly they do not include this most natural comparison in their simulation study. The two models are very similar in many settings of practical interest but differ when π is close to 0 or 1, or c is large. The latter rarely applies for human populations and the former is rare under typical single-nucleotide polymorphism ascertainment scenarios. One qualitative difference between the models is that the normal–binomial model allows point masses at 0 and 1 whereas the beta–binomial model allows instead an infinite density at the boundaries: the practical implications of this difference are not obvious and deserve exploration.

The four novelties claimed for the authors' analyses do not fully stand up to scrutiny. In particular the assumption of c constant across loci is effectively a special case of the hierarchical model of Balding and Nichols (1997); in their general setting, outlier loci subject to diversifying selection can be accommodated, and outlier locus–population combinations detected.

I welcome the authors' goodness-of-fit checks. They discuss a robustness check of removing one population at a time, but it may have perhaps been of more interest to replace an existing population with another, or to add a population: it is an unfortunate feature of the model that inferences depend on which subpopulations are studied and the sample sizes.

Parmigiani, Garrett, Anbazhagan and Gabrielson

I found this the hardest to read of the five papers: in part this reflects my lack of expertise in the subject area, and also the excellence of the competition—the four other papers are very clearly written. However, to try to get to grips with the paper I organized a journal club meeting, and the group collectively had some difficulty sorting out definitions and understanding figure legends. We also felt that some modelling assumptions were far from obvious and needed more motivating discussion. In particular, for the three-component mixture introduced in Section 2.2, the use of two uniform components on abutting intervals to represent underexpression and overexpression did not seem natural. The authors are commendably honest in showing Fig. 3: the fit of the model to data seems very poor to a casual inspection, casting further doubt on the choice of model.

Our questions were admirably answered in the oral presentation. Not only was the motivation for the models clearly discussed, but also Dr Parmigiani effectively disarmed potential critics with his impressive

frankness about possible weaknesses of the approach and the need for further work to assess its usefulness fully.

I was disappointed that the authors were downbeat about the contributions of statisticians to the interpretation of gene expression data. I had been hopeful that this area would provide an entry point for statisticians into bioinformatics, leading to a wider appreciation of statistical ideas. I continue to hope that the authors are overly pessimistic. And, although I welcome their cautious assessment of their own contribution, as far as I am aware nobody has done better in tackling their stated goals. I am hopeful that the paper will initiate a fruitful avenue of research.

I hope that some of the criticisms that I have made will be useful to the authors in their future work. I also hope that they do not detract unduly from the merits of the authors' achievements. I cannot give due credit in the space allocated to me, but I hope that the breadth, depth and clarity of the papers will be clear to readers.

I am greatly honoured to be allowed to propose the vote of thanks for all five papers.

Andrew D. Carothers (*Medical Research Council Human Genetics Unit, Edinburgh*)

Mr Chairman, Fellows, members: I must confess that before coming here today I felt slightly apprehensive that my comments might be construed as somewhat overly critical. However, after listening to Professor Balding's remarks, I see that my fears were groundless!

The advances in molecular genetics that have taken place over the past decade or so have led to the production of large amounts of new kinds of data, the interpretation of which often requires novel analytical approaches. The process of developing such techniques typically occurs in several stages. In the earliest, the practitioners tend to adopt *ad hoc*, though often ingenious, approaches that require no great statistical sophistication. Sooner or later, however, comes the realization that probability-based models and inference are required to extract the full potential of the new technology, and statisticians become increasingly involved. Finally, there is a stage, which never truly ends, of increasing refinement as the models are tested against reality and improved. In my view, these papers belong firmly in the middle stages of this process—i.e. they provide new and interesting ways to look at molecular genetic data without, at present, the extensive checks that would allow them to be 'sold' as going concerns to clinicians or biological scientists. This is not to denigrate them in any way, since they will no doubt go on to contribute to the definitive analytical methods of the future. Indeed, these middle stages are in many ways the most interesting and challenging, and therefore the most suitable for meetings such as this.

The analysis of quantitative trait loci (QTLs) using experimental crosses has a long history, and Broman and Speed have provided a useful comparison of several widely used computational methods using simulated examples. Their multi-QTL model is almost the simplest possible—a straightforward back-cross with strict additivity, no epistasis, no missing data, no crossover interference, equally spaced markers and equal effect QTLs situated at marker loci. Yet, as they demonstrate, even in this situation it is difficult to draw firm conclusions about the relative merits of different methods. Their results are chastening for the researcher in human genetics, such as myself, who does not even have the luxury of arranged matings.

The problem of estimating recombination rates on a scale of kilobases, or tens thereof, increases in importance as population-based association studies become more widely used for mapping complex disease loci, as in the proposed UK Biobank project (`http://www.mrc.ac.uk/index/public_interest/public-news/public-biobank_uk.htm`). Fearnhead and Donnelly provide ways of approximating the full likelihood for a coalescent-based model of recombination that are both computationally and informationally efficient. Their approach has the potential to be generalized in many ways and promises to feature prominently in the development of this area. Concerning the 'checks' that I referred to earlier, a simple way to examine the robustness of their methods might be to compare inferences from k subsets of the data, each generated by selecting every kth base.

The paper by Larget and colleagues looks at whole-gene rearrangements in mitochondria with a view to inferring relationships at the level of phyla. This is a novel and promising approach, although I would question their assumption that the probability of an inversion is independent of its length. Since their method treats genes as indivisible units in the process of differentiation, it would seem to imply that the inversion of incomplete gene sequences is cell lethal, so that what we observe today are merely the survivors of many such events. If this is so it would have implications for modelling the inversion process, though no doubt their general approach could be modified accordingly. It would also be interesting to compare this method with one based on DNA sequence data. Admittedly, as the authors point out, the latter has limitations for inferring distant evolutionary relationships but perhaps there is a degree of relationship, somewhere between the levels of species and phyla, where both methods are valid and could be compared.

Interestingly juxtaposed with this paper is that by Nicholson and colleagues, which uses data on single-nucleotide polymorphisms to look at 'drift' of subpopulations within a single species. The two papers are therefore both concerned with inferring relationships but with very different timescales (millions as against tens or hundreds of generations) and types of DNA (contiguous blocks as against single base pairs). The power of their approach arises from the large redundancy in the degrees of freedom. In their notation, where P is the number of subpopulations and L is the DNA sequence length in base pairs, they essentially fit of the order of $P + L$ parameters to a data set with of the order of PL observations. Clearly, there is much scope for refining their models, subject of course to computing constraints.

In Section 1 of the final paper, Parmigiani and colleagues have provided an excellent overview of the problems of inferring levels of gene expression in tissue samples (tumours in this case). As they point out, the basic difficulty from a statistical point of view is the very high gene:tumour ratio, typically 100 or more, leading to a problem of unsupervised clustering in high dimensional space—a notorious mine-field for the unwary. This is a question of logistics that is unlikely ever to be resolved by purely statistical fixes, at least in the medical field where the availability of tissue samples is the limiting factor. The particular model that they present is ingenious and plausible, but it needs to be validated against real data. That of course begs the difficult question of what constitutes 'validation' in this context.

In conclusion, this audience will need no reminding of the historic and productive interplay between statistics and genetics over the past 100 years or so. These presentations should convince us that, in the modern era of molecular genetics, that interplay continues to the mutual benefit of both fields. I am grateful to all the authors for providing us with such a stimulating programme, and I am very happy to second the vote of thanks.

The vote of thanks was passed by acclamation.

Jonathan L. Marchini and Lon R. Cardon (*University of Oxford*)

A main component of the paper by Nicholson and his colleagues is the use of a truncated normal distribution for the population allele frequencies. This model allows for fixation in individual subpopulations and is based on approximations to well-understood models of how the genetic material within populations evolves through time. The authors also acknowledge an alternative model that replaces the truncated normal distribution with the conjugate beta distribution (Balding and Nichols, 1995). We agree with Box (1976) that 'all models are wrong, but some are useful' and find it interesting to compare the performance of these two models on real data sets. The beta–binomial model is written as

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

$$\alpha_{ij} \sim \text{beta}\{\pi_i d_j, (1-\pi_i)d_j\}, \qquad d_j = (1-c_j)/c_j. \tag{2}$$

The conjugacy allows each α_{ij} to be integrated out of the likelihood to give a marginal likelihood for $\boldsymbol{\pi}$ and $\mathbf{c}$, i.e.

$$L(\boldsymbol{\pi}, \mathbf{c}) = \int_{\alpha} L(\boldsymbol{\alpha}, \boldsymbol{\pi}, \mathbf{c}) \, \mathrm{d}\boldsymbol{\alpha} \tag{3}$$

$$\propto \prod_{i=1}^{L} \prod_{j=1}^{P} \frac{\beta\{x_{ij} + \pi_i d_j, n_{ij} - x_{ij} + (1-\pi_i)d_j\}}{\beta\{\pi_i d_j, (1-\pi_i)d_j\}}. \tag{4}$$

Placing uniform priors on each component of $\boldsymbol{\pi}$ and $\mathbf{c}$ we can use a Metropolis–Hastings algorithm to simulate approximately from the posterior distribution. This avoids having to sample $\boldsymbol{\alpha}$.

We have applied the beta–binomial model to the two real data sets that were used in the paper. Figs 1(a) and 1(b) show the estimated posterior distributions of $\mathbf{c}$ for the European and global data sets respectively. Comparing these figures with Figs 6(a) and 6(b) in the paper we see that there is some non-trivial disagreement between the two models.

To compare the models more formally we used the deviance information criterion DIC (Spiegelhalter *et al.*, 2002) which has recently been proposed as a method of comparing models formulated in the Bayesian framework. The results are shown in Table 1. As a rule of thumb Spiegelhalter *et al.* (2002) suggest that a difference in DIC of 3 or greater should be considered significant. Using these guidelines we conclude that the normal–binomial model provides a better fit to the European data set and that the two models do equally well for the global data set.

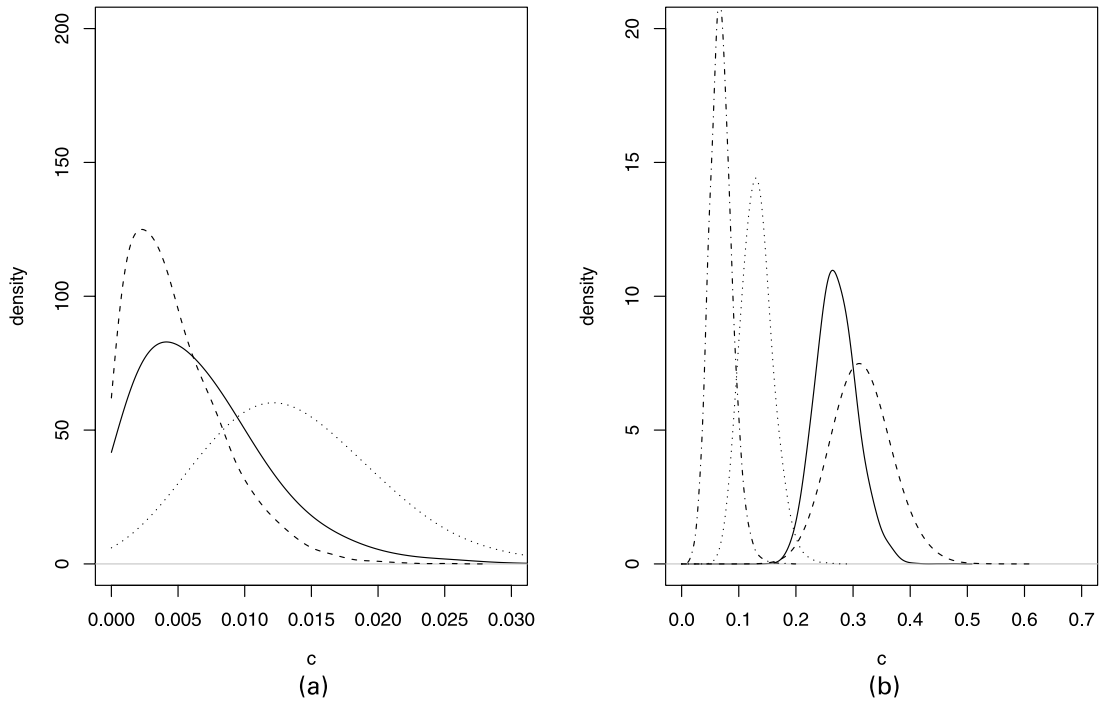

Fig. 1. Estimated posterior densities of the c_js for (a) the European data set using the beta–binomial model (———, France; -------, Utah; · · · · · · · , Iceland) and (b) the global data set using the beta–binomial data set (———, Mbuti; -------, Nasioi; · · · · · · · , Chinese; ----------, European)

Table 1. DIC for the normal–binomial and beta–binomial models applied to the European and global single-nucleotide polymorphism data set

	DIC for the following models:	
	European	*Global*
Normal–binomial	1248.578	1886.559
Beta–binomial	1255.301	1884.265

It is also interesting to consider how these models can be extended to incorporate unknown population structure. Much work has already been achieved in this regard within the context of forensic identification (Roeder *et al.*, 1997; Foreman *et al.*, 1997; Dawid and Pueschel, 1999). More recently Pritchard *et al.* (2000) extended this approach to account for admixture between individuals. This model is easily extended to model population-specific variation by using the natural Dirichlet generalization of the beta model used above, i.e.

$$\alpha_{ij.} \sim \mathcal{D}\left(\frac{1-c_j}{c_j}\pi^i_1, \ldots, \frac{1-c_j}{c_j}\pi^i_{J_i}\right). \tag{5}$$

A Metropolis–Hastings algorithm can be used to sample from the approximate posterior distribution of $\boldsymbol{\pi}$ and $\mathbf{c}$. Initial experiments suggest that this model can uncover subtle population subdivisions that are missed by the model of Pritchard *et al.* (2000).

Atam Vetta (*Oxford*)

Fisher (1918) hypothesized that a quantitative trait (QT) is polygenic. Thompson (1975) hypothesized a

few genes plus environment. I did not agree (Vetta, 1976). Now, it can be one gene plus environment! I would like to make five points.

(a) A QT is discovered almost every month (e.g. divorce, wealth or watching television!). With so many QTs, the number of human genes should be in millions. There are only 30 000–35 000 human genes. Venter (2001) said 'The idea that there is a simple deterministic explanation—that is: we are the sum total of our genes—makes me as a scientist, want to laugh and cry'. Perhaps laugh at their follies and cry that they call themselves 'scientists'.
(b) Broman and Speed use back-crosses of two inbred strains of mouse. There are about 30 000 genes in the mouse genome. Hubbard (BBC News, May 6th, 2002) said that the two 'genomes are so similar that you can just compare the two directly'. With a few thousand more genes Micky Mouse could become the master of the universe, Adam! The difference lies in the 'control' genes and interactions between genes. We can now study the two genomes directly.
(c) A gene is propagated through progeny. If you have no progeny, your genes are dead. A QT is inherited only if it is positively correlated with 'fitness'. A negatively correlated trait like the intelligence quotient or homosexuality has no future.
(d) What can we say about 'environment'? Our evolutionary history consists of two epochs. The first is when our ancestors tried desperately to adapt to the environment. We are the progeny of those who succeeded. The second is when they tried to adapt the environment to 'us'. The two epochs overlapped. Our genetic behavioural patterns were moulded in the first epoch. The two effects cannot be separated. Galton's 19th-century idea of a gene–environment separation has no place in the 21st century.
(e) QT advocates do not mention the brain. I hypothesized that the brain evolved by 'problem solving' and not by natural selection (Vetta and Capron, 1999). The human brain doubles in size during the first two years, and it doubles again by adulthood. We are born with most of the neurons that we need but the *connections* between them change. The extensions that grow out of neurons 'are constantly strengthened by experience or atrophy through lack of it' (Greenfield, 2000). Which experiences strengthen the extensions of human neurons? Here is the alternative to QT research.

Bob Griffiths (*University of Oxford*)

An alternative to the normal approximation for gene frequencies that is used in the paper by Nicholson and his colleagues is a coalescent model incorporating variable population size which holds for a divergence time that is not necessarily small (in a diffusion timescale).

Model

Our model comprises P populations of sizes $N_1(t), \ldots, N_P(t)$ at time t back from the present. The current total population size is N, with size ratios $b_1 = N_1(0)/N, \ldots, b_P = N_P(0)/N$. Time t is measured in units of N generations. Relative population sizes at time t back from the present are denoted by $\lambda_p(t) = N_p(0)/N_p(t)$. For example with exponential growth $\lambda_p(t) = \exp(\beta_p t)$, $\beta_p > 0$.

The approach of Nicholson and colleagues is to suppose that the time T since divergence from a common population is small; then the distribution of the single-nucleotide polymorphism (SNP) frequency $\alpha_{sp}(t)$, $p = 1, \ldots, P$, with initial frequency π_s, $s = 1, \ldots, L$, in the founding population in a diffusion model can be approximated by a normal distribution with mean π_s and variance $\lambda_p(T)Tb_p^{-1}\pi_s(1-\pi_s)$. Interest focuses on the divergence parameters $c_p = \lambda_p(T)Tb_p^{-1}$. The exact variance from the diffusion model, valid for any T, is actually

$$\left[1 - \exp\left\{-\int_0^T \lambda_p(u)\, b_p^{-1}\, \mathrm{d}u\right\}\right] \pi_s(1-\pi_s).$$

The distribution of $\alpha_{sp}(T)$, and sampling distributions, with time not necessarily small, can be expressed in terms of the lines-of-descent distribution from the coalescent process. The distribution of $X = X_{sp}$, the number of sample values equal to the given SNP type in a sample of $n = n_{sp}$ genes, is

$$\sum_{l=1}^{n} q_l^{p;n}(T) \sum_{k=0}^{l} \binom{l}{k} \pi_s^k (1-\pi_s)^{l-k} \binom{n}{x} \frac{k_{(x)}(l-k)_{(n-x)}}{l_{(n)}},$$

$x = 0, \ldots, n$, where $\{q_l^{p;n}(T)\}$ is the distribution of ancestor lines in the sample, and $a_{(b)} = a(a+1) \ldots a(a+b-1)$. There is a time coupling with a variable population size model and a constant size model with

line-of-descent distribution $\{q_l^{0;n}(T)\}$ so that

$$q_l^{p;n}(T) = q_l^{0;n}\left\{\int_0^T \lambda_p(u)\, b_p^{-1}\, \mathrm{d}u\right\}.$$

The only divergence parameters in this model are $c_p = \int_0^T \lambda_p(u)\, b_p^{-1}\, \mathrm{d}u$, $p = 1, \ldots, P$. The form of the sampling distribution suggests that computation would be effective with an improper prior $\pi_s^{-1}(1-\pi_s)^{-1}$ for SNPs with both types in the combined sample. There are easy asymptotic forms for the line-of-descent distributions for small T that are convenient for computation (Griffiths, 1984). As $T \to 0$, the number of ancestor lines $L^{0;n}(T)$ tends to n, the sample size. Limit distributions as $n \to \infty$ are normal or Poisson. If $T \to 0$ while nT is bounded then there is a normal limit, whereas if $T \to 0$ more rapidly $n - L^{0;n}(T)$ has a Poisson limit.

B. S. Weir (*North Carolina State University, Raleigh*) **and W. G. Hill** (*University of Edinburgh*)
Nicholson and his colleagues have presented an interesting Bayesian analysis of population structure that allows different populations to have different levels of relatedness or coancestry. We present moment estimators that are different from theirs.

Suppose that data are available from r populations. For the ith sample, the sample size is n_i alleles and the sample frequency of allele A_u at some locus is $\tilde{p}_{iu}$. The weighted average sample frequency is

$$\bar{p}_u = \sum_i n_i \tilde{p}_{iu} \sum_i n_i.$$

If each population has the same expected allele frequency p_u, the second moments of interest are

$$\mathrm{var}(\tilde{p}_{iu}) = \frac{1}{n_i} p_u(1-p_u)\{1 + (n_i - 1)\,\theta_i\},$$

$$\mathrm{cov}(\tilde{p}_{iu}, \tilde{p}_{i'u}) = p_u(1-p_u)\,\pi_u\theta_{ii'}, \qquad i \neq i',$$

where the coancestry coefficients θ_i and $\theta_{ii'}$ are for alleles within and between populations. The expectation process here is over samples and over populations, and the variance is equivalent to that of Nicholson and his colleagues, who assumed independent populations, $\theta_{ii'} = 0$.

A method-of-moments approach to estimating the θs needs only the second moments and places no constraints on higher order moments. Nicholson and his colleagues assumed a normal distribution for allele frequencies over populations, which is equivalent to assuming that there are no n-allele dependences for $n > 2$. The normal distribution has higher order moments that can be expressed in terms of only two-allele dependences. If

$$\theta_{\mathrm{A}} = \sum_{i,i'} n_i n_{i'} \theta_{ii'} \Big/ \sum_{i,i'} n_i n_{i'},$$

then moment estimators for θ_i relative to θ_{A} are

$$\frac{\theta_i - \theta_{\mathrm{A}}}{1 - \theta_{\mathrm{A}}} \hat{=} 1 - \frac{\left(\sum_i n_{ic}\right)\{n_i/(n_i-1)\}\sum_u \tilde{p}_{iu}(1-\tilde{p}_{iu})}{\sum_i \sum_u \{n_i(\tilde{p}_{iu} - \bar{p}_u)^2 + n_{ic}\tilde{p}_{iu}(1-\tilde{p}_{iu})\}}$$

where $n_{ic} = n_i - n_i^2/\Sigma_i n_i$. It is assumed that the θs do not depend on the allelic state u.

Although these estimates allow the different values of the θ_is to be compared, it is not possible to estimate them individually for a set of populations unless there are data from another set of populations that is unrelated to the first. If the populations in a set are assumed to be unrelated, then the different within-population values can be estimated, and for a large number of large samples

$$\hat{\theta}_i = \sum_u (\tilde{p}_{iu}^2 - \bar{p}_u^2) \Big/ \sum_u \bar{p}_u(1-\bar{p}_u).$$

Averaging over populations leads to the classical moment estimator for the single θ characterizing a set of populations,

$$\hat{\theta} = \sum_i \sum_u (\tilde{p}_{iu} - \bar{p}_u)^2 \Big/ r \sum_u \bar{p}_u(1-\bar{p}_u).$$

It may be that this last result suggested the moment estimator used by Nicholson and his colleagues:

$$\hat{\theta}_i^* = \sum_u (\tilde{p}_{iu} - \bar{p}_u)^2 \Big/ \sum_u \bar{p}_u(1 - \bar{p}_u) \tag{6}$$

but this is quite biased for small numbers of populations, since it has expectation

$$\mathcal{E}(\hat{\theta}_i^*) \approx \frac{(r-2)\theta_i + \bar{\theta}}{r - \bar{\theta}}$$

where $\bar{\theta} = \Sigma_i \theta_i / r$. (Nicholson and his colleagues consider the case of two alleles per locus, and then it is not necessary to sum over u in equation (6).) It is not surprising that they found that the estimate in equation (6) did not perform well for $r = 3$. We note that their Bayesian estimator did well for data simulated from the normal distribution, but this may not indicate its performance for data simulated under a drift model.

Darlene Goldstein (*École Polytechnique Fédérale de Lausanne and Institut Suisse de Recherche Experimentale sur le Cancer, Epalinges*)

Broman and Speed

Broman and Speed should be thanked for presenting this large study of model selection procedures for identifying quantitative trait loci (QTLs). It is particularly welcome to note the care that was taken to set an appropriate threshold to facilitate fair comparisons (although it appears to have turned out to be slightly conservative). This practice is seldom seen in the genetics, as distinct from statistics, literature, which contains many examples of comparisons using faulty criteria, e.g. power comparisons of procedures with differing false positive rates. It will be easier to draw reliable conclusions when it becomes standard in genetics to determine thresholds that are specific to the problem under consideration rather than to rely on asymptotics or on a rule of thumb that may not apply.

The authors focus attention on the identification of QTLs. But, although they consider effect size estimation to be secondary, it can affect judgments of the priority given to loci for follow-up (further localization, elucidation of function, etc.). The practical importance of this aspect might also be acknowledged and addressed.

Data are simulated under the unrealistic assumption of no crossover interference. The presence of crossover interference may be expected to induce some collinearity in the prediction variables. I wonder how forward selection–backward elimination, the procedure identified as 'best', might perform in this case.

The authors further confined the simulations to the class of additive models, which may also be considered to be somewhat unrealistic. It is perhaps also worth mentioning, then, that a smaller scale simulation study carried out by Fridlyand (2001) found that forward selection with Bayes information criterion $\delta = 1.25$ and $\delta = 1.5$ performed well compared with tree-based methods at detecting QTLs in the presence of low order interactions or linked interacting loci.

Parmigiani, Garrett, Anbazhagan and Gabrielson

Parmigiani and his colleagues should be thanked for presenting an approach that could be quite useful in overcoming some of the problems in the analysis of unclassified tumours based on their gene expression profiles, putting clustering into a statistical context and thereby allowing probabilistic statements and inferences to be made. Their framework is not unrelated to the usual implementations of model-based clustering, which most commonly assume a mixture of normal distributions. One can quibble over the precise modelling details, but the fairly novel idea here of categorizing genes as overexpressed, underexpressed or normally expressed as a prelude to classification provides a potentially powerful simplification. At a grosser level than using the measured expression level, it may therefore be less sensitive to the inherent unreliability in the measurements while not incurring too large a loss of information.

The model as presented is straightforwardly extensible to a larger number of more refined gene categories, if that should become desirable. It was also demonstrated how it can be extended to incorporate interactions through a cross-classification of sets of genes, although higher order interactions than pairwise would not seem advisable.

An important related issue concerns the subsequent use of genes identified. If the purpose is to predict a response to therapy or to aid in prognosis, say, then outcome information for the set of samples would need to be incorporated. In that case, supervised methods of classification would seem more appropriate. It was not clear to me how the method would be used for these purposes.

Korbinian Strimmer (*University of Munich*)
I congratulate Larget, Simon and Kadane on an innovative and stimulating paper. I shall restrict myself to three fairly independent comments.

First, I would like to emphasize the novelty and usefulness of the new algorithm. This is the first fully probabilistic approach to modelling gene inversion data that can be used practically to infer evolutionary relationships. Previous related work has either been impractical or not been based on a stochastic model. Now that we are realizing that DNA sequence data do also have their limits (e.g. to infer deep nodes in phylogenetic trees reliably) the development of models and computational techniques for genetic data other than DNA sequences (Felsenstein, 1981a) or continuous characters (Felsenstein, 1973) is greatly appreciated.

My second point concerns the construction of confidence intervals (or sets) for evolutionary trees. In Sections 3.1 and 3.2 the authors look at posterior probabilities for various hypothesized clades. As an alternative they could also test whether the trees investigated are significantly different. A review of relevant likelihood-based approaches has recently been given by Goldman *et al*. (2000). However, we need to be very careful about misspecifications. If neither the trees analysed nor the underlying substitution or rearrangement model is correct, undue weight may be given to the trees or clades with the largest likelihood or posterior (Strimmer and Rambaut, 2002). My impression is that this may be the case for example 1 in the paper.

Finally, I would like to discuss the relative performance of the Bayesian Markov chain Monte Carlo tree search in comparison with more traditional likelihood-based approaches. Inferring gene trees with a large number of leaves is an intrinsically difficult task for any kind of method. This is due to the astronomically large number of possible tree topologies. To optimize the likelihood efficiently over the tree space an array of specialized strategies, most of them heuristic, has developed in the last 20 years (e.g. Swofford *et al*. (1996) and Strimmer and von Haeseler (1996)). From looking into the recent biological literature one can easily gain the impression that the stochastic search that is implicit in the Bayesian Markov chain Monte Carlo approach is somehow a magic tool that can beat the complexity of the tree space. I do not think that this is generally true. On the contrary, it is in fact rather difficult to set up a Markov chain that mixes well and still covers the whole tree space. It is on this that the authors of the present paper should be congratulated. However, particularly when inferring large trees, serious mixing problems can probably not be avoided.

Simon Myers (*University of Oxford*)
Where substantial recombination may have occurred, inference on the recombination rate is difficult owing to the enormous number of different possible histories for a given data set. Under certain assumptions, though, the data are directly informative about past recombination in the sense that we need a certain number of past recombination events to construct *any* history for the sample (Hudson and Kaplan, 1985). Templeton *et al*. (2000) first suggested a recombination hot spot for the lipoprotein lipase (LPL) data set, along with widespread repeat mutation. In contrast Fearnhead and Donnelly here

Table 2. Minimum number of recombination events for the three data sets in the different site ranges†

Region	*Results for the following site ranges:*			
	106–2987	*2987–4872*	*4872–9721*	*Full region*
Jackson	10 (0.00347)	9 (0.00477)	13 (0.00268)	36 (0.00374)
Finland	2 (0.00069)	13 (0.00690)	11 (0.00227)	27 (0.00281)
Rochester	1 (0.00035)	13 (0.00690)	7 (0.00144)	21 (0.00218)
Combined	12 (0.00417)	22 (0.01167)	28 (0.00577)	70 (0.00728)

†The pairs of entries give the number of detections and (in parentheses) the number of detections per site for the relevant region. The middle interval (sites 2987–4872) corresponds to the suggested recombination hot spot.

find that only a few repeat mutations are likely to have occurred. However, the analysis of Clark *et al.* (1998) using incompatible site pairs (assuming no repeat mutation) did not find a clustering of recombination events.

We have developed a statistic (Myers and Griffiths, 2002) which gives a minimum number of recombination events in a sample history, under the assumption of no repeat mutation. Briefly, the method uses an optimization approach to bound the number of recombination events. If we have a collection of bounds B_{ij} on the number of recombination events between the site pairs i and j, we can view the minimum number of recombinations subject to these bounds as the solution, which may be obtained by using a simple algorithm, to a linear programming problem. One method to obtain B_{ij} uses the fact that, if we have H observed types in the sample, all of these types are created by mutation, recombination or ancestral. Then, if S is the number of segregating sites and R is the number of past recombinations, $R \geqslant H - S - 1$. Applying this bound to subsets of the original segregating sites gives a collection of bounds for different regions.

This method was applied to the LPL data set and detected many more recombination events than previous analyses (though many recombination events will still go undetected). The recombination events detected also showed clustering in the central region (Table 2). This is much less evident from the data for Jackson; however, pairs flanking the suggested hot spot show a somewhat raised level of detection. It is thus possible that using information over a longer distance would recover the signal for a hot spot here. Perhaps the LPL data set simply has so much pairwise incompatibility (primarily due to recombination and not repeat mutation) that the hot spot signal is swamped if we use only incompatibility to detect recombinations. When we use more detailed information about the history, the detection picture agrees well with the findings of Fearnhead and Donnelly.

Mark A. Beaumont (*University of Reading*)

By harnessing the modern machinery of computational statistics Nicholson and his colleagues look afresh at some of the earliest problems of population genetic inference. Particularly innovative aspects are the modelling of ascertainment, and statistical model checking. There appear to have been two strands of development of non-mutational models of population divergence—those based on population splitting, such as that considered by Cavalli-Sforza and Edwards (1967), which are transient models, and equilibrium models based on the island model of Wright (1931). In the island model the likelihood is given by a multinomial–Dirichlet distribution (Balding and Nichols, 1995, 1997; Rannala and Hartigan, 1996). The likelihood in the transient model is difficult to compute, which has led to the Brownian motion approx-

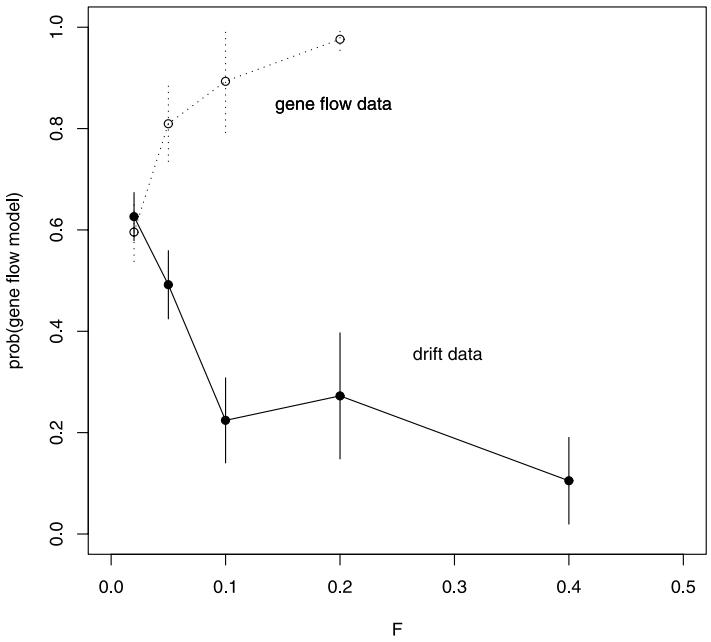

Fig. 2. $p(I|D)$ plotted against F: the points are means based on five replicates and the vertical lines show estimates of the standard errors (∘, gene flow model; •, transient model)

imation to drift considered by Cavalli-Sforza and Edwards (1967) and Felsenstein (1981b), and also the approximation considered by Nicholson and Donnelly. However, the likelihood can be evaluated by using sampling methods, as reviewed in Beaumont (2001).

It is possible to parameterize the gene flow and transient models on the same footing by the probability F_i that two lineages in the ith subpopulation coalesce before being drawn from a base-line gene frequency distribution $\mathbf{x}$, as described in Ciofi *et al.* (1999) (see the mathematical appendix at `http://www.pubs.royalsoc.ac.uk/publish/pro_bs/rpb1435.htm`). Thus we can estimate the likelihood $p(D|F, \mathbf{x}, I)$, where D is the data, as considered by Nicholson and his colleagues, F includes all F_i, and I is an indicator variable that takes for example 0 for the transient model and 1 for the gene flow model. Using Markov chain Monte Carlo methods it is possible to integrate over all F_i and $\mathbf{x}$ to estimate $p(I|D)$ (Ciofi *et al.*, 1999).

Performance has been assessed with data simulated according to the assumptions of either model. Samples of size 50 chromosomes taken from five populations have been simulated. The data consist of allele frequencies at five loci, each with 10 alleles in the base-line gene pool. Although 10 alleles are present in the base-line frequency distribution, owing to drift and sampling generally fewer are present in the samples. Groups of five replicates from each model were simulated, where F had values 0.02, 0.05, 0.1 and 0.2 in both models and additionally 0.4 in the transient model (F_i identical). The results are illustrated in Fig. 2. There is more accurate inference with data simulated from the gene flow model. The explanation is that the inferences have been made conditional on the number of alleles in the sample, rather than those in the base-line gene pool. As discussed by Nicholson and colleagues, there is a higher tendency for fixation in the transient model, and this effect is also observed in the simulated data. By conditioning on the number of alleles observed in the data one biases against acceptance of the transient model. This effect should also be observed with single-nucleotide polymorphisms and suggests that ascertainment may be important when attempting to distinguish between the two models.

C. A. Glasbey and C. D. Mayer (*Biomathematics and Statistics Scotland, Edinburgh*)

Microarrays are an important new technology, introduced well by Parmigiani and his colleagues. There are several stages in the analysis of such data, including

(a) image analysis to extract spot intensities (see, for example, Glasbey and Ghazal (2002)),
(b) normalization (Yang *et al.*, 2002),
(c) the identification of differential expression,
(d) clustering and
(e) modelling dynamic interactions (Friedman *et al.*, 2000).

Parmigiani and his colleagues focus on one approach to steps (c) and (d).

Our main concern is the ambiguity in the definition of *differential expression*. The data used for illustration are obtained from radioactive labelling. We are more familiar with fluorescence labelling, where a sample is compared against a control on a single microarray, thus simplifying key aspects of the analysis. We can remove ambiguity about what is meant by differential expression, whereas they cannot. So, for gene 1753, data shown in their Fig. 3 are bimodally distributed, and the choice is arbitrary about which, if either, of the modes is labelled as normally expressed. This leads to problems with the Markov chain Monte Carlo method, which is sensitive to initial values and is unlikely to explore the parameter space fully, and inconsistencies between Figs 4(a) and 4(b). Although the method has reduced sampling variability, new, man-made noise appears to have been added! It is possible that their mixture model could be modified to rectify this problem partially, by introducing a constraint such as $\pi_g < 0.5$.

The authors emphasize that their method fits mixture distributions *across arrays*, rather than *across genes* as considered by Lee *et al.* (2000). We find the across-arrays approach problematic as it relies heavily on a proper normalization, whereas the across-genes method is simpler and has intrinsic normalizing properties. In addition, Hoyle *et al.* (2002) show that there is biological information in the across-gene distribution.

We note that the estimates of profiling probabilities, as they are products of terms $\hat{p}_{gt}$, will have poor sampling properties. Also, it is not clear from the paper how the authors knew that the grey points in their Fig. 1 were due to a *loss of signal* rather than to differential expression, and conversely for the outlying points in several of the plots in their Fig. 2. Finally, in their Fig. 3 the presence of outlying data distorts the whole of a normal probability plot, making it uninterpretable, so only the subset of data values classified as being normal should be plotted.

Sylvia Richardson and Clare Marshall (*Imperial College School of Medicine, London*)

The paper by Parmigiani and his colleagues explores many important issues from proposing a categorical model to describe the profile across tumours to using derived summaries to mine for a subset of genes that would be used for a classification of tumours. As such, the authors are to be congratulated for providing a 'mine of ideas'. Yet, we feel that some issues are not addressed in a fully convincing manner.

We shall concentrate on the hypotheses behind the mixture model for the expression of a gene across tumours. This model relies heavily on the assumption that there is a modal class of so-called normal expression even though all tissues are diseased and it is departures from this modal class that are used to discriminate sets of tumours in a symmetric fashion. We find this hypothesis quite restrictive with respect to the wealth of patterns of expression profiles that could be exhibited such as many small peaks and bimodality, a small sample of which is shown in their Fig. 3. Are the authors claiming that the implied similarity of profile shapes for all the genes is generally appropriate in many tumour classification problems? To convince the reader that this hypothesis is appropriate for this data set, it would have been useful to present a comprehensive analysis of residuals and summary statistics for the fit of all the profiles.

A further remark concerns the use of uniform distributions with end points at the centre of the normal distribution for modelling underexpression or overexpression. By imposing arbitrary constraints on the end points, the authors reduce the overlap between these uniform distributions and the central normal distribution. Nevertheless, there is a less clear interpretation of the p_{gt} than if separated distributions were used. Would it not have been better to use a mixture of three well-separated normal distributions (and not half-normal distributions as was suggested as an alternative in the paper)? Did the authors think that there would be too many parameters to estimate (i.e. six instead of four)? It would be possible to link some of these parameters, say the variances, to reduce this set and to increase the stability of the fit. The authors claim that the bad fit of the uniform distributions is of no consequence, but our intuition is that it will influence the estimate of the p_{gt}*s* that are used subsequently in the data mining exercise.

In general, one could think of using a mixture of Gaussian distributions as a set of basis functions to model flexibly the diversity of the (standardized) profiles and use the corresponding allocations for clustering the tumours.

As a final point, could the authors provide a statistical rationale for discarding the genes with the presumed loss of signal. How is the distinction made between a loss of signal and underexpression?

Richard Durrett and Rasmus Nielsen (*Cornell University, Ithaca*)

We congratulate Larget and his colleagues for a stimulating paper on mitochondrial evolution. Previous approaches have concentrated on calculating the minimum number of inversions, and a statistical approach to the problem has been long overdue. We have, independently of them, used similar methods to estimate the number of inversions that have occurred between homologous chromosomes in two species (York *et al*., 2002). We compared chromosome arm 3R of *Drosophila melanogaster* with chromosome 2 of *Drosophila replete*. In this part of the *Drosophila* genome translocations and pericentric inversions are rare or absent. A consideration of a data set with 79 markers due to Ranz *et al*. (2001) yields a maximum posterior probability estimate of 87 inversions and a 95% credible interval (71, 118) that does not include the parsimony number of inversions (53). This shows that the method can handle a large number of markers even when a large number of inversions have occurred and suggest that the traditional parsimony estimator may not have desirable statistical properties.

At the conclusion of their paper, Larget and his colleagues mention the problem of extending their work to cover translocations and chromosome fissions and fusions. In fact this extension is easy to accomplish with existing technology. Hannenhalli and Pevzner (1995a) devised algorithms to compute the most parsimonious sequence of inversions between two chromosomes. Hannenhalli and Pevzner (1995b) extended their first result to compute the most parsimonious sequence of inversions, translocations and chromosome fissions and fusions to relate two genomes. There were two steps in their solution:

(a) observe that, in the case that all of the chromosome ends were the same, one could reduce the general problem to that of inversions by concatenating the chromosomes and
(b) reduce the general case to the co-tailed case by introducing extra labels for chromosome ends and empty chromosomes with two artificially labelled ends to compensate for the difference in chromosome number.

Using these ideas it is straightforward to extend the methods developed for inversions to the more complex situation. This is being done for the comparison of two genomes by Durrett, York and Nielsen in unpublished work. The methods described in the appendix to the paper by Larget and his colleagues should allow us also to treat multiple genomes in the context of a known phylogeny.

P. M. Visscher and S. A. Knott (*University of Edinburgh*) **and C. S. Haley** (*Roslin Institute, Edinburgh*)
We congratulate Dr Broman and Professor Speed on revisiting a well-known problem in quantitative trait locus (QTL) mapping in simple and complex pedigrees, i.e. how to detect multiple QTLs from a sample of individuals with phenotypes and marker genotypes on multiple chromosomes. The authors find that, as expected, no model selection method is superior in controlling both the type I and the type II error rates, and that their proposed Markov chain Monte Carlo approach performs similarly to previously proposed stepwise selection methods. Our contribution focuses on five aspects.

(a) The approach favoured by the authors is a standard statistical application of model selection and does not explicitly utilize the biological nature of the data, i.e. that markers on the same chromosome are linked whereas markers on separate chromosomes segregate independently. Flanking markers absorb all the variation due to a QTL in the interval and QTLs in adjacent intervals cannot be mapped unambiguously (Whittaker *et al*., 1996), and these properties could be used in a model selection procedure.
(b) Interval mapping methods based on least squares methods have been developed for line crosses and other population structures, and these perform virtually identically with the more computationally intensive maximum likelihood methods in terms of type I errors, statistical power and estimation of effects (Haley and Knott, 1992; Knott *et al*., 1996).
(c) An alternative strategy to the 'black box' model selection is to test specific biological hypotheses, e.g. 'is genetic variation associated with a particular chromosome?' (De Koning *et al*., 1998; Visscher and Haley, 1998; Visscher *et al*., 2000) or 'are the data consistent with the segregation of many QTLs in coupling?' (Visscher and Haley, 1996; De Koning *et al*., 1998).
(d) Line crosses are performed to map QTLs underlying the genetic differences between the lines, and we know *a priori* that there is genetic variation in the cross-bred population. An alternative model selection procedure is to detect chromosome regions that explain more variation than would be expected if the genetic variance was equally distributed across chromosomes (Visscher and Haley, 1996; De Koning *et al*., 1998).
(e) The authors make a plea for more large scale computer simulation studies in methodological papers on QTL mapping but appear to have overlooked many relevant references. Many studies, in particular those based on the computationally efficient least squares methods, have used and advocated simulation approaches, and these include permutation tests to set significance thresholds (Churchill and Doerge, 1994), bootstrap methods to estimate confidence intervals of QTL locations (Visscher *et al*., 1996) and comparisons of model selection approaches (Visscher *et al*., 2000). Least-squares-based QTL mapping software that offers permutation tests and bootstrapping techniques for a variety of population structures is available (Seaton *et al*., 2002).

The following contributions were received in writing after the meeting.

Roderick D. Ball (*New Zealand Forest Research Institute, Rotorua*)
I would like to thank Broman and Speed for some interesting results of comparative simulations of quantitative trait locus (QTL) mapping methods. Not suprisingly, single-marker methods did not do so well for linked QTLs in repulsion, and the number of unlinked 'extraneous markers' was much larger than the nominal type I error rate, for 'detecting' QTLs to within for example 10 centimorgans. The alternative hypothesis can be technically true (there is a QTL somewhere) while being false in practice (there is no QTL close by). This should lead us to question the appropriateness of hypothesis testing for this problem.

By assuming that QTLs are at marker locations, the QTL mapping problem reduces (approximately) to a problem of model selection. For a review see Sillanpää and Corander (2002). The authors use the Bayes information criterion–δ-criterion to find a single 'best' model. However, unless the sample sizes are very large, inference is problematic and estimates of effects are affected by selection bias (Miller, 1990), unless an independent verification population is used. Ball (2001) used the criterion to estimate approximate posterior probabilities for each of a range of models. This approach enables us to give approximate probabilities for the presence of QTLs in a region and to correct for selection bias.

Carson *et al*. (2002a, b) simulated several genetic architectures, and thresholds, with a range of sample sizes, selection criteria and experimental designs applied to both 'detection' and 'validation' populations, with simulations of the genetic gain obtained from various strategies. One possible strategy is to select markers that are statistically significant in both populations. For a given threshold, this will not be optimal across a range of genetic architectures and sample sizes, and using an independent validation population

can even result in *less* gain. So we question the choice of δ (page 649), to correspond to a given P-value—this seems to be a retrograde step. An alternative suggestion is to choose δ so that the Bayes factors for comparison of a limited number of models with the null model match the corresponding values estimated using the Bayes information criterion–δ-method.

Finally a philosophical question: the authors are interested in finding 'an appropriate model' but, we ask, appropriate for what? They consider that estimates of QTL location and magnitude of effects are of secondary importance. For applications, e.g. marker-based selection, the size and location of QTL effects are critical. Could the authors elaborate on how they envisage the application of the selected model?

Christine A. Hackett (*Biomathematics and Statistics Scotland, Dundee*)

Broman and Speed have given an interesting paper on model selection for quantitative trait locus (QTL) identification. This problem also interested us for a doubled haploid barley population, where the genetic model is almost identical with the back-cross. Our model was based on the regression mapping approach of Haley and Knott (1992), in which the expected trait value at a putative QTL location can be expressed as a linear function

$$E(y) = \mu + \beta\, f(r) \tag{7}$$

where μ is the trait mean, β is the QTL effect and $f(r)$ is a function of the distance to the markers flanking the location and their genotypes. The functions $f(r)$ can be evaluated at a set of positions through the genome and used as explanatory variables in a model selection step. Our interest was in simultaneously mapping QTLs responsible for six yield characters (Hackett, Meyer and Thomas, 2001), so equation (7) was extended to a multivariate regression model and forward selection was used to optimize the modified (multivariate) Akaike information criterion of Bedrick and Tsai (1994). We also concluded that forward selection was inclined to include extraneous QTL locations and attributed this to influential individuals in the mapping population. Such extraneous locations were removed by bootstrapping the data and retaining only the QTL locations that were significant in 95% of the bootstrap samples.

I question the authors' implementation of composite interval mapping with the number of regressors fixed in advance. In practice, it would be more usual to select as a first set of regressors those on each side of the peaks obtained by interval mapping. There would then be a genome scan using this set to control background genetic variation, and further locations would be identified and included in the set of regressors. This would continue until a suitable stopping rule is reached.

The authors have focused on a back-cross population, where the probability of a QTL genotype conditional on the genotypes at its two flanking markers is independent of the genotypes at more distant markers. In many species inbred parental lines are not available, markers will segregate in different ratios depending on the parental genotypes and an analysis of variance at each marker locus is less informative than QTL interval mapping. Maliepaard and Van Ooijen (1994) have shown how all markers on a chromosome can be used to infer the QTL genotype in interval mapping in such species. This approach has been extended to QTL interval mapping in tetraploid species by Hackett, Bradshaw and McNicol (2001).

Susan Holmes (*Stanford University*)

Larget and his colleagues justify their preference for a likelihood-based approach because they feel that this is the only way to do statistics properly for phylogenetic trees and they rightly point to the lively debate between the proponents of the various methods for estimating phylogenetic trees. However, staging the methods in clear statistical terms definitely clarifies the tensions. It is not true that 'only likelihood methods provide a clear framework for assessing uncertainty', unless the authors wish to write off all inference in the nonparametric paradigm, ignoring 20 years of contemporary research.

Parsimony in particular is a nonparametric method (see Holmes (2002) for a detailed description of this argument); distance-based methods are what statisticians would call 'semiparametric' models. There is actually a nice continuum between nonparametric methods and parametric methods illustrated by Tuffley and Steel's (1997) observation that parsimony and likelihood estimates coincide when rates are allowed to differ on different edges of the tree. Thus, as the number of parameters to estimate increases, maximum likelihood coincides with parsimony.

This is important, because biologists have become polarized on 'the best method'. But, as working statisticians know, a robust nonparametric approach is very useful in the exploratory stages, before the correct models are available, and for the gene rearrangement situation we may still be in such an early stage. There is rarely a unique 'best method' and the choice of estimator must rely on prior information.

A model of genome rearrangement

The authors must choose a prior distribution on the set of possible trees, which is exponentially large (the actual formula for the number of trees is due to Schröder (1870), a 19th-century German combinatorialist). It does not seem realistic to take all trees as equally likely; as is well known in the practice of rooting trees, there is always actual prior information on the relationships between species. Other sensible prior distributions are detailed in Aldous (1996, 2001).

It would be even more satisfactory to be able to use the geometry of the space of trees (Billera *et al.*, 2001) to devise more realistic nonparametric priors on tree space by neighbourhoods.

Finally, this is a very interesting step in the direction of a clear Bayesian analysis of the phylogenetic estimation problem given gene arrangement data. It would be wonderful to have the computer programs available so that we could calibrate the methodology on cases where the tree is already well agreed on.

Dirk Husmeier (*Biomathematics and Statistics Scotland, Edinburgh*)

The important contribution of the paper by Parmigiani and his colleagues is the development of a systematic statistical framework for the discretization of continuous gene expression levels. Such a discretization is required for modelling non-linear regulatory interactions between genes with Bayesian networks, as discussed, for instance, by Friedman *et al.* (2000). The discretization in Friedman *et al.* (2000), however, is done in a heuristic *ad hoc* manner and can be considerably improved on by adopting the systematic statistical framework proposed by the authors. A few questions need to be addressed, though. Modelling expression levels with a mixture of a Gaussian distribution and two finite support uniform distributions is ill specified, since the infinite tails of the Gaussian distribution imply that genes with abnormally high or low expression levels will be classified as *not* differently expressed. Although the heuristic constraint $\kappa > 5\sigma$ might be a sufficient fix for this shortcoming for the particular data set under consideration, we may wonder why the authors did not apply the constrained mixture of Gaussian distributions mentioned at the end of Section 2.2, which would overcome this ill specification in principle. The divergence measures $q(g, g')$ and $q(t, t')$ of their equations (3) and (5) are unstable. If for any tumour t and genes g and g', for instance, the expression probabilities $P(e_{gt}|a_{gt})$ and $P(e_{g't}|a_{g't})$ are very different such that

$$P(e_{gt} = k|a_{gt})\ P(e_{g't} = k|a_{g't}) \approx 0 \qquad \forall k \in \{1, 0, -1\},$$

then $q(g, g') \approx 0$. This leads to the undesirable consequence that $q(g, g')$ may be dominated by a few tumours (or, equivalently, that $q(t, t')$ may be dominated by a few genes), which renders it very susceptible to noise. A more robust measure is the Sibson divergence:

$$q_{\mathrm{S}}(g, g') = \sum_t \sum_k \left[P(e_{gt} = k|a_{gt})\ \ln \left\{ \frac{P(e_{gt} = k|a_{gt})}{\overline{P_t(k)}} \right\} + P(e_{g't} = k|a_{g't}) \ln \left\{ \frac{P(e_{g't} = k|a_{g't})}{\overline{P_t(k)}} \right\} \right]$$

where

$$\overline{P_t(k)} := \frac{P(e_{gt} = k|a_{gt}) + P(e_{g't} = k|a_{g't})}{2}.$$

The support of $\overline{P_t(k)}$ contains both the supports of $P(e_{gt}|a_{gt})$ and $P(e_{g't}|a_{g't})$ as subsets, and this overcomes the instability.

Ritsert C. Jansen (*University of Groningen*), **Cajo J. F. ter Braak, Chris A. Maliepaard and Martin P. Boer** (*Biometris, Wageningen*)

Many selection procedures can be represented as special cases of minimizing the expression $D/\phi + \alpha k$, where D is the deviance function, ϕ is the dispersion parameter, k is the number of parameters and α is a constant or function of the population size n (McCullagh and Nelder, 1989). The range of $\alpha = 2$–6 usually provides plausible models, but Broman and Speed use the more stringent value $\alpha = 4.6\,\mathrm{LOD} \sim 11.5$ in their simulations. It is important to distinguish at this point between selection and testing (see Jansen (2001), pages 585–586). Aiming at higher resolution for disentangling the effects of closely linked quantitative trait loci (QTLs), Jansen (1994) used $\alpha = 6$ in selection, and $\alpha = 4.6$ LOD in subsequent conditional testing for QTLs. The values of α should depend on the need for controlling false positive (which is important in gene cloning) *versus* false negative results (which is important in marker-assisted breeding). Broman and Speed have combined selection and testing, and it remains unclear how their method can deal with these different experimental goals. It is also important to distinguish between maximum likelihood and analysis of deviance (Jansen (2001), pages 582–584). Jansen's (1994) analysis-of-deviance approach

uses a single and unbiased estimate of ϕ to prevent overfitting due to selection bias in ϕ, whereas the maximum likelihood approach of Broman and Speed uses different, biased, ϕs when comparing models. A consequence is that Broman and Speed's method is likely to produce inferior results for $k > 2\sqrt{n}$.

Broman and Speed's method of randomly searching through all models may be better than existing methods in searching the model space, but we argue that only restricting the search space in a sensible way will bring real progress. In the paper, models receive informative priors, whereas regression parameters have diffuse improper priors. An attractive alternative is to specify informative priors for the regression parameters. Narrow priors for marker cofactors correspond to interval mapping of a single QTL only, whereas flat priors correspond to fitting a QTL plus all marker cofactors. Thus, the form of the prior determines the effective dimension d_{eff} of the model. The expression $D/\phi + \alpha k$ can be formulated as $D/\phi + \alpha d_{\text{eff}}$, where both D and d_{eff} depend on the dispersion parameter λ of the prior. Given λ, obtaining estimates of the regression parameters is standard (Hastie and Tibshirani, 1990). Next, the criterion can be optimized for λ. Putting different priors on parameters for main effects and interaction effects can extend this approach. Thus, penalized (ridge) regression methods offer a promising, more continuous approach to model selection.

Paul Joyce (*University of Idaho, Moscow*)

Statistical models describing genealogies involving recombination have been known since the 1980s (Kaplan and Hudson, 1985). However, the recent interest in understanding the way in which recombination rates vary across the human genome has made these models increasingly important. It is interesting to note that, although the underlying mathematics is well understood, the statistical inference problem is still a computationally difficult problem even with ever increasingly more powerful computers. Fearnhead and Donnelly have been leaders in tackling this difficult problem. In Fearnhead and Donnelly (2001) they showed that a full likelihood method is possible, but not practical for the large data sets of interest. The approximations proposed seem to work nearly as well as the full likelihood method but are much more computationally efficient.

The marginal likelihood approach is based on the likelihood of summary statistics. I wonder whether it is possible to develop a theory for 'approximately sufficient statistics' which could serve as a guide for how to pick summary statistics without losing much information on the parameters of interest. I suspect that this would be a difficult theory to develop in this present context.

I found the composite likelihood approach particularly appealing. This approach seems to work quite well, indicating that ignoring some of the long-range dependences has little cost with respect to estimating the recombination rate. It is interesting to note that the maximum likelihood estimator for the recombination rate $\hat{\rho}$ appears to have an approximate normal distribution. However, the χ^2-distribution does not appear to be a good approximation to the log-likelihood ratio. Have the authors investigated the log-likelihood ratio in the case where $\rho = 0$? In this case the true parameter lies on the boundary and according to Self *et al.* (1987) the asymptotic distribution should be distributed according to a mixture of χ^2-distributions. What do simulations suggest about the distribution of the log-likelihood ratio in this case?

I thank for authors for a very well-written and enjoyable paper as well as an important contribution to an important problem.

Na Li and Matthew Stephens (*University of Washington, Seattle*)

We congratulate Fearnhead and Donnelly on their stimulating paper and admire their efforts to overcome the considerable computational challenges that arise in this context. The approach of multiplying likelihoods obtained from managable subregions seems promising for estimating recombination rates that are assumed to be constant along the sequence, but it is less well suited to investigating recombination rate heterogeneity, particularly the large variation over small scales suggested by some recent data (e.g. Jeffreys *et al.* (2001)).

We have been investigating an alternative approach to exploring recombination rate heterogeneity (and, more generally, to modelling linkage disequilibrium across the genome), based on the fact that the likelihood for n observed haplotypes $h = (h_1, \ldots, h_n)$ can be written as

$$L(\rho, \theta) = \Pr(h_1|\rho, \theta) \prod_{i=2}^{n} \Pr(h_i|h_1, \ldots, h_{i-1}, \rho, \theta) . \tag{8}$$

The conditional distributions $\Pr(h_i|h_1, \ldots, h_{i-1}, \rho, \theta)$ are unknown, but Fearnhead and Donnelly (2001) proposed an approximation which they used to develop the importance sampling procedures employed in this paper. Our suggestion is to exploit this approximation (or, in fact, a slight modification that eliminates

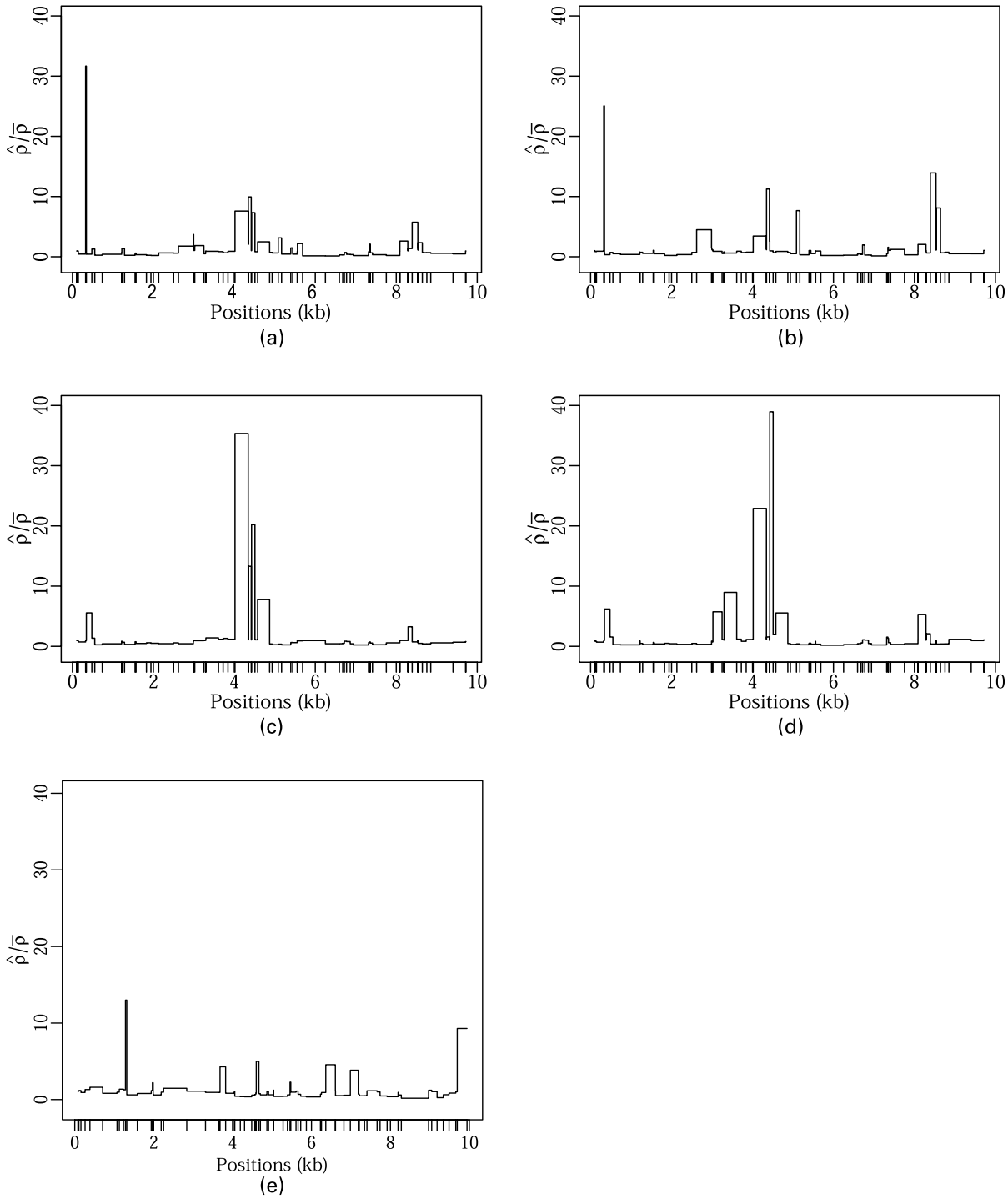

Fig. 3. Estimated maximum PACL estimates of ρ_i for the lipoprotein lipase data sets (a) combined, (b) Jackson, (c) Rochester and (d) Finland, and (e) for a data set simulated under constant recombination ρ: the simulated data were generated by using the program `mksamples` (Hudson, 2002), assuming a constant-sized population, with the number of chromosomes, segregating sites and ρ chosen to match the Jackson population sample; tick marks on the *x*-axes show positions of segregating sites; the computations for each subpopulation took less than 30 min of central processer unit time

θ by conditioning on the polymorphic sites being segregated) in another way: substitute it into equation (8) to approximate the likelihood directly. The resulting 'product of approximate conditionals' likelihood (PACL) combines information from all sites simultaneously and can be computed quickly for huge regions, even when ρ varies along the region.

Although the theoretical properties of the PACL are unclear (for example, it depends on the order in which the haplotypes are considered; below, we average over five random orders), we have obtained encouraging results for real and simulated data (Li and Stephens, 2002). To investigate potential recombination rate heterogeneity in lipoprotein lipase, we first analysed the four (Jackson, Finland, Rochester and combined) data sets, assuming constant ρ, using the estimated haplotypes from Nickerson *et al*. (1998). The maximum PACL estimates $\bar{\rho}$ were 8.4, 2.3, 2.9 and 11 per kilobase respectively. We then analysed each data set assuming the scaled recombination rate between sites i and $i+1$ to be $\rho_i d_i$, where d_i is the known physical distance in kilobases between sites i and $i+1$, and with $\log(\rho_i)$ assumed *a priori* independent and identically distributed $N\{\log(\bar{\rho}), 2.3^2\}$, which gives approximately 95% probability that ρ_i is within a factor of 100 of $\bar{\rho}$. The joint maximum *a posteriori* estimates for $(\rho_1/\bar{\rho}, \ldots, \rho_{S-1}/\bar{\rho})$ (Figs 3(a)–3(d)) show reasonable agreement across subpopulations, and considerably more variation, and spatial correlation, than the corresponding estimates for a data set simulated with constant ρ (Fig. 3(e)). The results suggest a more subtle pattern of recombination rate variation than a single hot spot near the centre. It would be interesting to compare these results with maximum *a posteriori* estimates for ρ obtained by using the authors' approximate marginal likelihood, for a sliding window along the sequence.

George A. Marcoulides and Zvi Drezner (*California State University, Fullerton*)

We congratulate Broman and Speed on an impressive thought-provoking paper. We are certain that it will stimulate much further research into the use of model selection ideas to identify quantitative trait loci in experimental crosses. In our discussion, we would like to draw attention to some recent research on the identification of an appropriate model that was not mentioned by them. Our aim is to shift attention from stepwise-type model selection methods to more efficient but extensive space searching strategies.

Drezner *et al*. (1999) introduced the tabu search model selection procedure and demonstrated its superiority over other model selection procedures and its comparability with the all-possible regression approach. Commonly used regression searches examine the neighbourhood of a set of models and add or remove regressors as long as the criterion of the set improves. The tabu search procedure does not restrict the search to improving moves. The search may move to inferior solutions in the neighbourhood of the current solution and, as a result, it allows the process possibly to exit local optima when taking uphill moves. An addition of a certain regressor may lead to a set with an inferior criterion, but additional changes of regressors may lead the search to a set with a better criterion. To avoid cycling, tabu search imposes a tabu (prohibited) status to the parameters recently involved in the choice of the new solution.

The flow of the tabu search procedure is as follows (for more details see Drezner *et al*. (1999)).

Step 1: an initial subset K of selected regressors is generated.
Step 2: the best current subset is $K_{\text{best}} = K$.
Step 3: the iteration counter is set to iter $= 0$ (the current iteration).
Step 4: the neighbourhood $N(K)$ of the subset K is created.
Step 5: the criteria $\text{crit}(K')$ for all $K' \in N(K)$ are evaluated.
Step 6: if $\text{crit}(K') < \text{crit}(K_{\text{best}})$ for any $K' \in N(K)$, set $K_{\text{best}} = K'$. Go to step 8.
Step 7: if, for all $K' \in N(K)$, $\text{crit}(K') \geqslant \text{crit}(K_{\text{best}})$, choose the best admissible subset $K' \in N(K)$.
Step 8: set $K = K'$ and iter $=$ iter $+ 1$.
Step 9: the tabu list is updated. Go to step 4 unless the stopping criterion has been met.

Contrary to the authors' belief that 'more refined procedures will not necessarily provide sufficiently improved results to justify their added complexities and increased computation requirements', we found that the implementation of our tabu search model selection procedure is quite simple and the results outstanding. For all problems considered tabu search found the best model, whereas other techniques often failed to do so. The capability and efficiency of the tabu search procedure has been demonstrated with as many as $2^{80} = 1.208\,928\,196\,1 \times 10^{24}$ possible models which are solved in just a few seconds.

Kanti Mardia (*University of Leeds*)

All the authors should be congratulated on giving solutions to so many exciting new challenges in genomics. Another area of interest is proteomics and structural genomics where one of the most

challenging problems for many decades is how to deduce or predict the three-dimensional structure of protein from the one-dimensional amino-acid sequence. Also, there is a problem in resolving functions of unknown proteins and to design unknown enzymes, for example. Thanks to advancements in information technology, large protein databases are now available. In understanding the protein structure, I believe, here stochastic geometry would play a significant role, especially to understand the 'similarity' of unlabelled sites of unequal forms in three dimensions in a large database.

Of course, any new statistical methodology must incorporate physicochemical properties such as that of different amino-acids. Substantial advancements have been made in analysing amino-acid sequences in a similar fashion to that of DNA sequences. But the statistical work related to the interplay of amino-acid sequences and three-dimensional protein structures is still in its infancy; see, for examples, the recent reviews by Baker and Sali (2001) and Taylor *et al.* (2001).

Another challenging problem is to obtain the configurational entropy of a given molecular configurational state of a protein; the entropy helps to understand factors that are involved in the protein's stability; instability leads to misfolding of the protein. This entropy depends on the distribution of large numbers of conformational angles and some initial interdisciplinary work is described in Demchuk and Singh (2001) and Demchuk *et al.* (2001). Markov random fields on a torus have a great potential for such problems.

In all these advancements, interdisciplinary research will again be of vital importance and it will be helpful to learn from the authors their thoughts on future directions in bioinformatics.

Gilean McVean (*University of Oxford*)

The results of applying the composite likelihood method of Fearnhead and Donnelly to empirical data can be compared with those derived from using alternative *ad hoc* likelihood-based methods. Hudson's (2001) estimator (see also McVean *et al.* (2002)) works by combining the coalescent likelihood for all pairs of segregating sites in the sample and finding the value of $\rho = 4N_e r$ that maximizes the composite likelihood.

For the lipoprotein lipase data collected from three populations (Jackson, Finland and Rochester) (Nickerson *et al.*, 1998) we estimate $4N_e r$ to be 2.98 per kilobase, 2.77 per kilobase and 1.39 per kilobase respectively. The comparable figures from Fearnhead and Donnelly are 14.4 per kilobase, 7 per kilobase and 3 per kilobase respectively. Thus the pairwise likelihood and composite likelihood methods give very different estimates of the amount of recombination in lipoprotein lipase, although we cannot be sure that the confidence intervals of the estimates do not overlap. Given that neither estimator shows strong bias when applied to simulated data sets (Hudson (2001), McVean *et al.* (2002) and Fearnhead and Donnelly), the discrepancy is most likely to be the result of inadequate modelling. One biological complication that could lead to such a difference is gene conversion (Frisse *et al.*, 2001). Gene conversion, like reciprocal crossing-over, breaks down associations between alleles, but its effect is largely independent of the distance between sites. Consequently, for closely situated sites, gene conversion may be the major factor affecting patterns of allelic association. Because the composite likelihood method of Fearnhead and Donnelly works by combining the likelihood of short disjoint regions, and ignores the contribution of associations between regions, it is likely to overestimate the rate of reciprocal crossing-over under the presence of unacknowledged gene conversion. Using the method of Frisse *et al.* (2001) to estimate the relative contribution of

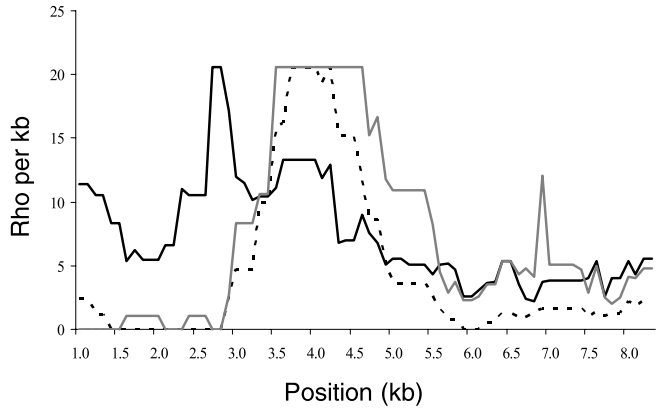

Fig. 4. Variation in the recombination rate along the lipoprotein lipase gene as estimated by the pairwise likelihood method: estimates based on overlapping 2 kb windows with a maximum value of ρ of 20.0 per kilobase: ———, Jackson; - - - - - - - , Rochester; ———, Finland

gene conversion to reciprocal crossing-over for the Jackson data suggests that over half all recombination events are resolved as gene conversions (assuming that gene conversion tract lengths are exponentially distributed with a mean of 350 base pairs (Hilliker *et al.*, 1994), the estimated ratio is 0.61).

Variation in local recombination rates can also be addressed by using the pairwise likelihood approach. Fig. 4 shows a series of sliding window estimates for the population recombination rate for the three population samples, where each window spans a region of 2 kb. As with the composite likelihood analysis of Fearnhead and Donnelly we find evidence for considerable variation along the chromosome in recombination rate, with an elevated rate in the region between 3 and 5 kb from the start of the sequence. The signal of variation in recombination rate does, however, vary between population samples, with the signal being weakest (though still present) in the Jackson population.

Xiao-Li Meng (*Harvard University, Cambridge, and University of Chicago*)

Whereas many may consider both Fearnhead and Donnelly's methods as cases of *pseudolikelihood*, I would call their 'marginal likelihood' a *partial data* (not *partial*) likelihood to highlight the fact that it is a *real* likelihood, although it uses only a part of the data. This will avoid the general inconsistency in the meaning of 'marginal' between 'marginal likelihood' (referring to marginal data) and 'marginal posterior' (referring to marginal parameter). More importantly, since 'partial' is equivalent to 'incomplete', the more precise naming could provide a beneficial link to the literature on incomplete or missing data. For example, an insight derived from that literature is that inference based on the partial data can also be viewed as a semiparametric approach, leaving the 'unused or unobserved' part of the full likelihood unspecified. This suggests investigating whether and how their partial data likelihood is more robust to a misspecification of population evolution models.

As another example, Fearnhead and Donnelly investigated the issue of lost information by plotting their partial data likelihood against the full data likelihood (their Fig. 3). Although this visual inspection is very appealing, it needs to be complemented by quantitative results on how the loss of information varies with simulation configurations and with data sets. The measure of 'fraction of missing information' (FMI) in the incomplete–missing data literature (Dempster *et al.*, 1977; Meng and Rubin, 1991) is readily computable in Fearnhead and Donnelly's simulation, I believe. Although the FMI is a large sample measure, a histogram of the FMI and/or an analysis of variance of it can nevertheless provide users with a more complete glimpse of what to expect in practice than Fig. 3 can.

Fearnhead and Donnelly's 'composite likelihood', however, is a *real* pseudolikelihood! I hope that they will not be annoyed by my 'name picking', for the prefix 'pseudo' can remind us that this function should not be treated as a real likelihood. Apparently the undercoverages in their Table 1 are due to

(a) the pseudo-Fisher information's overestimating the information in the pseudomaximum likelihood estimator and
(b) the failure of the normal approximation.

For (a), a well-known remedy is to use the robust 'sandwich' variance estimate, since the pseudoscore is just an estimating equation. For (b), using a normal approximation to $\log(\rho) - \log(\hat{\rho})$, rather than to $\rho - \hat{\rho}$, can go a long way in dealing with the non-normal tails displayed in the authors' Q–Q-plots (their Fig. 6). Common experience suggests that these two remedies together may bring the coverages in Table 1 much closer to their nominal levels.

Michael F. Ochs (*Fox Chase Cancer Center, Philadelphia*)

Parmigiani and his colleagues consider an important problem in the classification of tumours by using microarray data. The need for a proper statistical framework for analysing these data sets arises from many issues, including the limited number of tumour samples that are available, the high levels of noise in expression data and large differences in the natural variation of expression across genes. Despite these obstacles, some progress has been made in classifying tumours from microarrays, as noted in the paper. Nevertheless, the need for a solid statistical basis for classification, including proper handling of noise, cannot be overstated. This paper is an important step in the development of that basis and provides probabilistic statements concerning tumour assignments, which is highly desirable and not provided in the examples cited from previous work.

Of special interest in this paper is the development of the probability measure of two genes having identical patterns over all tumours, $q(g, g')$. This provides a noise-adjusted estimate for clustering of the data, which could be of great utility. In addition, the value of $q(g, g)$ provides an internal measure for

the data, especially when considered over all g. Also of interest is the potential for a biological interpretation of the molecular profiles. However, the example given in Section 4 illustrates the difficulty in the approach, perhaps because a single consistent role for all genes within a group was sought. Any true biological response will inevitably involve genes with multiple functions as defined in the authors' Table 1; therefore an interpretation of a profile in biological terms will result, not in all genes within a group having consistent assignments, but in a set of functions that together provide the necessary components for a complex response.

One concern with the approach outlined arises when gene interaction, or more generally co-regulation of genes, is considered. The mathematical development of Section 2 depends on the e_gs being independent across genes, but it is not clear how this independence can be maintained with co-regulation. In addition, since the true distributions of gene expression are unknown, it would be good to know how much the results in Section 4 depend on the assumption of normal and uniform distributions. The ability to 'plug in' different distributions in the algorithm and to reanalyse data would be useful in this regard, allowing users to investigate how these assumptions affect their results.

Mark Pagel (*University of Reading*)

The idea that gene arrangements may be useful for inferring phylogenetic relationships has been proposed before, but Larget and his colleagues are the first to propose a formal statistical model of the process. Their implementation of this model in a Markov chain Monte Carlo framework is welcome and promising.

The utility of gene arrangements as phylogenetic markers rests on the assumptions that they are rare and that gene order does not affect fitness. If the first assumption is correct, then all species that share a particular gene order are likely to form a monophyletic group. The second assumption is required to construct plausible probability densities of all possible gene orders.

A difficulty that may arise in this context is that gene arrangement very likely is strongly subject to selection. Most gene rearrangements will probably reduce fitness, if not be fatal because the spatial relationships of gene promoters and enhancers to the genes on which they act are crucial—gene order rearrangements will change these relationships. Equally, in organisms that do not have large stretches of non-coding DNA, removing a random section of the DNA for rearrangement is likely to cut active genes into two pieces, neither of which will work. What this may mean for the model of Larget and his colleagues is that the expected probability densities for all possible gene rearrangements substantially misrepresent the actual pool of possible rearrangements. It may also mean that only a small subset of possible rearrangements can ever reach fixation (owing to selection against deleterious rearrangements). If this is correct, then the probability of two lineages independently suffering the same rearrangement is far higher than anticipated by the model, and the utility of the character (a particular rearrangement) for phylogenetic inference is correspondingly reduced. This may explain why the authors' models could not always conclusively assign high posterior probabilities to deep nodes in their test cases.

If the probability model has been written to assign the same probability densities to all branches with a given number of rearrangements, then the model may lose some specificity. The probability of observing one branch with n rearrangements may differ substantially from that of another branch with n rearrangements, depending on which rearrangements have occurred. Much phylogenetic information may reside in the particular rearrangement or rearrangements that have occurred in a branch.

These considerations aside, there is much to recommend the approach of Larget and his colleagues as a method for inferring deep phylogenetic relationships. The explosive increase in whole-genome data, from which gene order can be extracted, will mean that a growing and captive audience awaits future refinements of their work.

Naijun Sha and Marina Vannucci (*Texas A&M University, College Station*)

We would like to congratulate Parmigiani and his colleagues for this well-executed paper and for their insightful discussion of the many issues concerning microarray experiments.

The authors mention the use of mixture distributions when they suggest the alternative specification of $f_{-1,g}(\cdot)$ and $f_{1,g}(\cdot)$ as half-normal distributions on the negative and positive line respectively. Mixture distributions have been extensively used for variable selection in regression, following the work of George and McCulloch (1993). In Sha *et al.* (2002) we show how mixtures of normal priors can be employed for gene selection in supervised microarray experiments.

Let (Z, X) indicate the observed data, with $X_{n\times p}$ being the predictor matrix of expression levels and $Z_{n\times 1}$ being a categorical response vector coded as $0, \ldots, J-1$, for J classes. Data can be modelled via a multinomial probit model and latent variables can be used to transform the model into a normal linear

regression on the latent responses, as in Albert and Chib (1993). Mixture priors can then be used to select important genes, following the framework used by Brown *et al*. (1998) for variable selection in multivariate regression. These priors assign a probability to each possible subset of genes and are extremely flexible. Information on the size of models as well as on interactions between the genes can be easily included. *A posteriori*, Markov chain Monte Carlo methods to sample from the marginal distribution of single models can be combined with truncated sampling techniques on the unobserved latent variables. The selection of genes is done via the inspection of the best visited models or of the marginal probabilities of inclusion of single variables. The method also allows a classification of future samples.

In Sha *et al*. (2002) we describe the methodology for both binary and multiple responses and provide applications in chemometrics and in functional genomics. Genes that we identify in bench-mark data sets for cancer classification based on microarrays appear to be all biologically relevant.

Mikko J. Sillanpää (*Rolf Nevanlinna Institute, Helsinki*)

Broman and Speed nicely present a new model selection criterion (BIC_δ) for selecting putative markers which are physically close to the influential genes (quantitative trait loci) and are therefore taken as locations contributing to the genetic variation of the considered quantitative character (trait) in a back-cross design of inbred lines. The proposed BIC_δ criterion has many attractive properties including

(a) consistency,
(b) that the criterion selects the model which has the highest posterior probability and
(c) that the choice of δ has a direct connection to the genome-wide LOD-score threshold (the false positive rate).

The authors illustrated the potential of their method by using a large simulation study of 2000 replicates. In the simulation study many alternative search methods to scan through the large discrete model space were compared. The method based on Markov chain Monte Carlo sampling seemed to offer arguably the best performance.

As also stated by the authors, the use of Markov chain Monte Carlo sampling allows the illustration and estimation of the posterior probabilities associated with each putative model and that way provides an uncertainty measure of the model chosen. Alternatively, from BIC or BIC_δ we can derive estimates for the model-specific posterior probabilities based on asymptotics (Ball, 2001). The posterior distribution provides additional information on the other models which may have been—in the light of the data—almost as good as the model chosen. Furthermore, model-averaged estimation based on several different models (that are compatible with the data) is also possible if model-specific posterior probabilities are available. In such an estimation, each model-specific estimate is weighted with the probability of the corresponding model. This practice has been shown to have many useful properties including robust effect estimates (Ball, 2001; Madigan and Raftery, 1994; Raftery *et al.*, 1997). More discussion on this topic can be found in Sillanpää and Corander (2002).

Another issue which I would like to comment on briefly is the testing of the new method on the basis of a large number of simulation replicates, as was proposed in the paper. We must proceed with caution here because taking an average is not always a good idea. For example, we should expect problems if we average different LOD-score curves over 2000 replicates. In the paper, however, an average was taken in a safe manner over the quantities of interest (e.g. the number of correctly identified quantitative trait loci).

Scott Sisson (*University of Puerto Rico*)

I commend the efforts of Broman and Speed in providing a detailed comparison of model selection techniques when applied to the identification of quantitative trait loci (QTLs). It is revealing to contrast the performances of the more prominent methods, and I look forward to seeing how they compare in future studies applied to more detailed models, or different selection criteria. My comment relates to the use of Markov chain Monte Carlo (MCMC) sampling to traverse and sample from the model space.

Powerful simulation methods have become invaluable in the ever-widening field of biostatistics as the abundance of sequenced data allows increasingly complex models to be addressed (e.g. Thompson (2001)). However, problems in efficiently traversing model space or state space via MCMC samplers are well known in the genetics literature, even in relatively simple circumstances (Hoeschele, 2001; Sisson, 2002). Although the MCMC search through model space presented here is perfectly adequate for the given simulation study whereby each marker contributes independently of all other markers to a purely additive phenotype, it is difficult to recommend this simplified procedure for searching through the model space in more general situations without a thorough appreciation of the problems that may occur.

For example, one concern is that a major QTL is only major when it is considered in combination with an epistatic locus with otherwise no direct contribution to the phenotype. Thus the QTL will go undetected in an analysis excluding QTL interaction terms. Although the authors note that the MCMC model space search can be extended to include pairwise interactions in linear models or tree-based models, it is essential that the MCMC sampler can efficiently move to such models. In the above situation, unless both loci are proposed simultaneously the major QTL may be missed as there will be little or no chance of accepting moves to models including either locus individually.

Further, as it is not known for a given circumstance how many jointly epistatic loci may exist for a given trait, it is clear that more general movements through the model space should be considered rather than sequentially proposing moves to structurally adjacent models. One such method might proceed by additionally proposing the inclusion or exclusion of blocks of markers of random size in the manner of Hurn *et al*. (1999).

Brian S. Yandell and Chunfang Jin (*University of Wisconsin—Madison*), **Jaya M. Satagopan** (*Memorial Sloan Kettering Cancer Center, New York*) **and Patrick J. Gaffney** (*Lubrizol, Cleveland*)

The balance of model fit and complexity central to assessment is captured in Broman and Speed's BIC_δ, but complementary instruments have great value. Empirical studies of complex traits may detect 'major quantitative trait loci (QTLs)', overlooking modifier genes that cannot be localized. Model assessment guides the discovery process, selecting 'better' models without fixating on one 'best' model.

We find closely related Bayes factors effective despite recent criticism. A judicious choice of robust priors and empirical Bayes methods reduces the influence of priors on Bayes factors (Gaffney, 2001). Averaging over nuisance parameters can stabilize Bayes factors (Satagopan *et al*., 2000). Semilogarithmic plots of the posterior or prior against model identifier (number and chromosome pattern of the QTL) provide useful graphical guides.

We differ on prediction. Model-averaged posteriors over the 'better' models of QTLs and effects along the genome reveal genetic architecture (see Ball (2001)). Further, agricultural breeding studies use predicted breeding values to 'select' individuals for future crosses. Marker-assisted selection alone ignores important modifiers that become fixed in a few generations (Edwards and Page, 1994).

Model search on a simulated framework map with a spacing of 10 centimorgans is revealing, but markers are clearly in a model or not. A practical QTL search spans a genome continuum, with two closely linked loci easily confused as one, depending on the sample size and marker spacing. Model search with reversible jump Markov chain Monte Carlo (MCMC) sampling allows joint sampling in the neighbourhood to distinguish them. Whole-genome reversible jump MCMC sampling differs from the authors' marker regression-based MCMC method (see Satagopan and Yandell (1996), Sillanpää and Arjas (1998), Stephens and Fisch (1998), Gaffney (2001) and Yi and Xu (2002)).

Forward selection is biased, and backward elimination is impossible on the whole genome. Gaffney (2001) used a 'pre-burn-in' phase beginning with no QTL and a high prior mean to build large initial models aggressively for subsequent sampling with the reversible jump MCMC method. This combined with block updates of effects and long-range position updates improves the efficacy of multiple QTL searches (Gaffney, 2001).

Finally, thresholds for model assessment criteria should be used with extreme caution. Thresholds were developed to test a single QTL against no QTL. Simulations (Goffinet and Mangin, 1998) show an empirical dependence on the size of other linked QTLs.

Zhao-Bang Zeng (*North Carolina State University, Raleigh*)

Model selection has been used in quantitative trait locus (QTL) mapping analysis. Kao *et al.* (1999) used the model selection approach in developing multiple-interval mapping (MIM), a maximum-likelihood-based method to map multiple QTLs in a genome. Given any positions for multiple QTLs, the method treats QTL genotypes as missing data and uses an EM algorithm to evaluate the likelihood of the mixture model. Thus, rather than restricting the analysis to markers, which is the case considered by Broman and Speed, MIM can search the positions of multiple QTLs at any place in the genome and offers a more precise way to estimate the positions and effects of QTL. MIM can also take QTL epistasis into account in mapping analysis and offers a simultaneous estimation of complex interaction effects between multiple QTLs. On model selection, we adopted an interactive stepwise search procedure aided by a residual permutation test (Zeng *et al.*, 2000), which takes the experimental design, data structure and specific models into account. MIM methods and procedures have been implemented in QTL Cartographer (Basten *et al*., 2002), particularly the easy-to-use version of Windows QTL Cartographer (Wang *et al*., 2002). It has been successfully

applied to several data sets (Zeng *et al.*, 2000; Weber *et al.*, 1999, 2001) with ample demonstration of its advantages.

Broman and Speed suggest BIC_δ to aid the model selection. However, it is not clear how to relate the δ-value to experimental design, data structure and some biological parameters, such as heritability and genome size, and on the basis of what considerations. The final recommendation of Broman and Speed is to choose the δ-value on the basis of the genome-wide LOD threshold of Lander and Botstein (1989) for interval mapping, back to the previous criterion justified for the likelihood ratio test of one *versus* no QTL in a genome. For model selection of multiple QTLs, this is insufficient. For example, in a simulation study of Wang (2000), it was found that heritability and genome size, along with other factors such as marker density, play important roles for setting up an appropriate criterion for model selection, on the basis of controlling a defined, say 5%, false positive error rate in QTL detection. In contrast, the sample size in the range of 150 to 1000, which is typical for QTL mapping experiments, has relatively little effect on the criterion. The simulation of Broman and Speed considered only one genetic make-up with heritability 0.5 and missed probably the most important factor influencing the model selection criterion in this application. Although it is appropriate to emphasize that QTL mapping analysis is better to be viewed as a model selection problem, the study of Broman and Speed does not deliver the result, how to set up the model selection criterion properly for QTL mapping analysis.

The **authors** replied later, in writing, as follows.

Karl W. Broman and Terence P. Speed

We thank the discussants for their comments. They have given us much to consider. We shall do our best to respond to the points raised, though we shall focus on those that we consider most important.

Choice of δ

Several discussants commented on our method for choosing the penalty δ in the BIC_δ criterion for comparing models. We chose δ with reference to a 95% genome-wide LOD threshold, L. With $\delta = 2L/\log_{10}(n)$, we thus seek the model γ for which $\log_{10}\{\mathrm{RSS}(\gamma)\} + |\gamma|(2L/n)$ is minimized. Note that the LOD threshold depends primarily on the size of the genome and very little on the sample size. Thus this criterion is actually more like Akaike's information criterion than the Bayes information criterion.

With δ chosen in this way, we can expect that, under the null model of no quantitative trait loci (QTLs), the rate of inclusion of loci will be under control. The behaviour of the criterion when the null model is not true is, of course, a matter of investigation. We have shown that it performs well in one particular situation, a back-cross with 50% heritability and seven QTLs of equal size effect.

Yandell and his colleagues point out that the performance of the procedure may depend on the sizes of the effects of the QTLs. Zeng states that the heritability of the trait can greatly affect the false positive rate.

In response, we briefly investigated the influence of heritability on the performance of the BIC_δ criterion, by repeating our simulations for the case of low (10%) and high (90%) heritability, with the structure of the cross and the number and locations of QTLs as before. We again performed 2000 simulation replicates but considered only the performance of BIC_δ (with the values of δ as before) with forward selection to 25 markers, followed by backward elimination.

The results appear in Fig. 5. (Note that the simulations for the case of 50% heritability were not repeated.) The rate of inclusion of extraneous unlinked loci is approximately constant with heritability. The rate of inclusion of extraneous linked loci is stable with the sample size in the case of 90% heritability, increases with the sample size in the case of 10% heritability and decreases with the sample size in the case of 50% heritability. The observed high rates of inclusion of extraneous linked loci are largely due to correctly identified but imprecisely localized QTLs. In terms of the rate of inclusion of extraneous loci, the performance of the BIC_δ criterion, with δ chosen as described above, does *not* appear to be greatly affected by the heritability of the trait.

Jansen and his colleagues point out that the appropriate penalty depends on the goals of the experimenter. We completely agree and are surprised that our emphasis of this point was not perfectly clear. Our particular choice of penalty cannot be *generally* recommended, but it served here to allow the proper comparison of different model selection procedures.

The point made by Jansen and his colleagues concerning maximum likelihood *versus* deviance, and the magic $2\sqrt{n}$, could be translated into an alternative criterion for model comparison. Perhaps they could study the performance of such a criterion relative to those considered here.

Bayesian methods

Bayesian methods for QTL mapping have generated much interest. The principal advantages of such

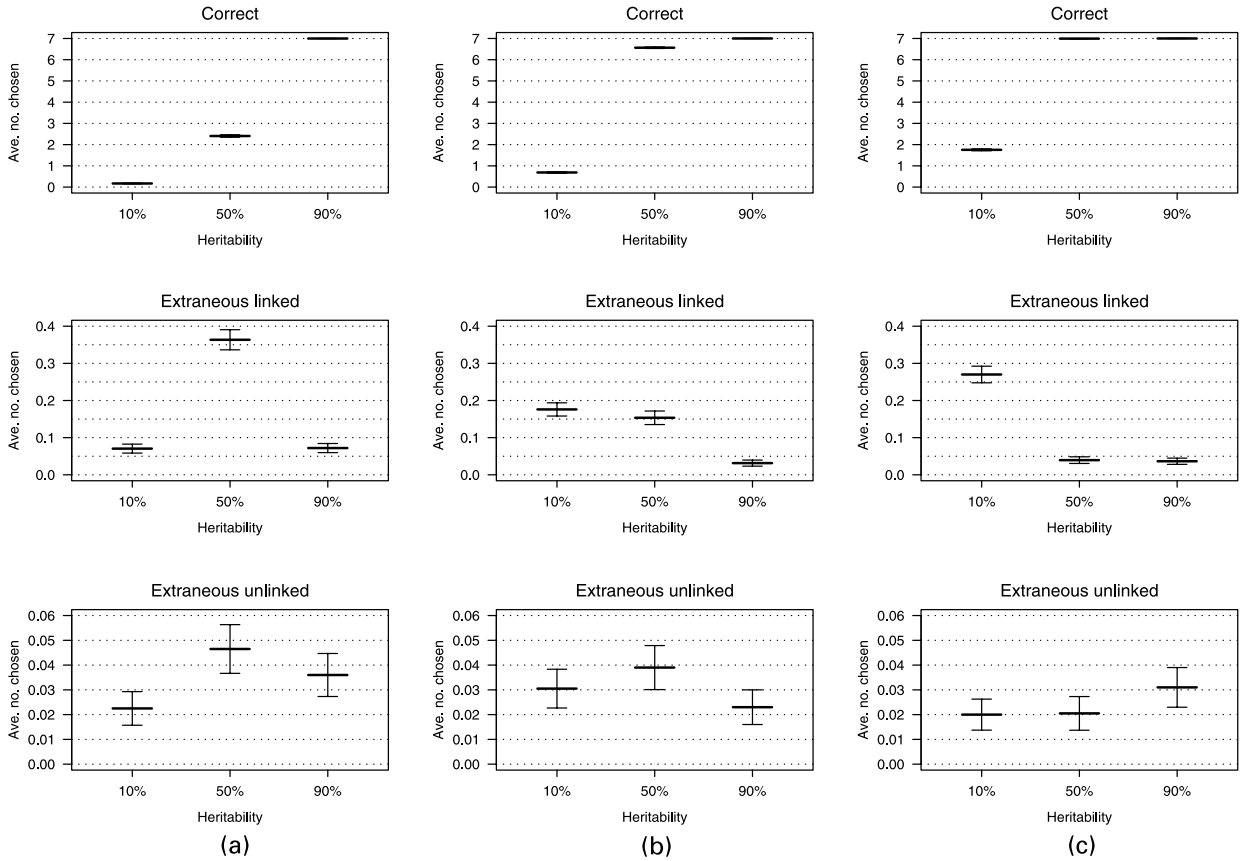

Fig. 5. Effect of heritability on the performance of the BIC_δ criterion, using forward selection followed by backward elimination, in simulations of a back-cross with (a) $n = 100$, (b) $n = 250$ or (c) $n = 500$ individuals under a seven-QTL model

methods are in the expression of model uncertainty (through the posterior distribution on models), and improved estimates of QTL effects (through model averaging). These problems can also be addressed through non-Bayesian methods. For example, regarding model uncertainty, in a traditional model selection setting one can and should inspect models that are close competitors of the model selected.

We have focused on the problem of identifying QTLs, and it is here that Bayesian statisticians can become slippery. The outcome of a Bayesian method is a posterior distribution on the space of QTL models, and it is generally not clear how this may be turned into a decision about which loci should be called QTLs. Several discussants have referred to Bayes factors, but without precise criteria a comparison of methods is not possible.

It should further be pointed out that the choice of prior on QTL models, and of criteria for dealing with Bayes factors, corresponds exactly to the choice of model comparison criteria in traditional model selection. Bayesian methods have no advantage in this, the essential aspect of the problem.

Large scale simulations

We have emphasized the importance of large scale computer simulations for the comparison of QTL mapping procedures. That such simulations are rare should be obvious to those who are familiar with the QTL mapping literature. (We are not speaking here of permutation tests and the bootstrap; such simulation-based analysis methods are in common use, as they should be.)

The studies of Piepho and Gauch (2001) and Visscher *et al.* (2000) are important exceptions. We hope that such studies will become more common, so that we may gain a better understanding of the relative performance of different QTL mapping procedures, in the presence of multiple QTLs.

Composite interval mapping and multiple quantitative trait locus mapping

Composite interval mapping (Zeng 1993, 1994) and multiple-QTL mapping (Jansen, 1993; Jansen and Stam, 1994) are essentially the same method: an initial step of selection of marker covariates, followed by a testing step via a one-dimensional genome scan. There are numerous approaches to the initial selection of

marker covariates. Although the particular method that we considered for comparison has been criticized, the approach appears to be commonly used in practice, as it is implemented in the popular computer software QTL Cartographer (Basten *et al.*, 2000).

We are in favour of discarding this two-step procedure and focusing directly on model comparison and selection. Those who disagree are challenged to study the performance of their favourite version of composite interval mapping or multiple-QTL mapping, in relation to the procedures considered herein.

Other comments

Ball asks, regarding the choice of an appropriate model, 'appropriate for what?'. Our principal goal was to identify regions of the genome that contribute to variation in the quantitative trait, with the ultimate aim of identifying the underlying gene or genes that are responsible. We would like to create a list with as many such loci as possible, while controlling the inclusion of extraneous loci, so that the investigator will not follow too many false leads.

Goldstein suggests that the presence of crossover interference might be expected to induce some collinearity in the prediction variables. We did not expect this; nor did we expect interference to make an important difference in any other way, but we have not carried out a systematic study of the matter.

Visscher and his colleagues state that our approach 'does not explicity utilize the biological nature of the data'. On the contrary, our recommendation of forward selection, as a valuable model search strategy in the case of additive QTL models, relies heavily on the simple correlation structure of marker genotype data, induced by the linear order of markers on chromosomes.

Jansen and his colleagues, in consideration of the problem of model search, state that 'only restricting the search space in a sensible way will bring real progress'. We completely disagree. If computation time were not a factor, a more complete search would always be at least marginally better, provided that the criteria for model comparison were chosen appropriately.

That QTL mapping is a model selection problem has been widely accepted but has not been widely adopted, and so it deserves further emphasis. Numerous QTL mapping procedures have been proposed, but their relative performance, in the presence of multiple QTLs, deserves further study.

Paul Fearnhead and Peter Donnelly

We are grateful to the discussants for their interesting and thought-provoking comments.

The new approximation to the likelihood, the product of approximate conditionals likelihood, described by Li and Stephens is very interesting, and we look forward to better understanding its properties. The idea of using good proposal distributions *directly* as approximations to likelihoods in complex problems has wider applicability (within and outside genetics) which could also profitably be explored. Both the product of approximate conditionals likelihood and the pairwise likelihood method described by McVean typically involve substantially less computational effort than our composite likelihood approach. Significantly, the existence of several fundamentally different approximate methods is very encouraging. We recommend that data be analysed by *each* of these methods. If all methods point to the same conclusions, this is reassuring. When methods differ in their answers to a particular inference question, care is obviously needed.

We welcome the reanalyses by Li and Stephens, by McVean and by Myers of the lipoprotein lipase (LPL) data. Fig. 6 presents an analysis of the LPL data analogous to that in the discussions by Li and Stephens, and McVean. We calculated the approximate marginal likelihood separately for 17 1.5 kb regions and a final 1.25 kb region, each starting 0.5 kb apart, covering the 9.75 kb of the gene. The general pattern of apparent rate variation is similar to that in the analogous analyses by these discussants. Our analysis suggests a hot spot in the Finnish and Rochester data, in the region from 2.5 to 5 kb, with much less evidence of a hot spot in the Jackson data. Our estimates exhibit less variability across subregions than do those of Li and Stephens, which we believe may be due to our averaging of the variability over 1.5 kb regions, rather than over regions between each pair of consecutive segregating sites. In contrast with the general suggestion made by Li and Stephens, this kind of analysis shows that the methods developed in our paper are not necessarily unsuited to studying local variation in recombination rates.

As noted in the paper, in the presence of gene conversion our approach will be estimating a different quantity from some other methods. McVean presents convincing evidence that gene conversion is important in the LPL data, a suggestion which is also consistent with informal analyses that we have done using 2 kb subregions, which suggest lower maximum likelihood estimates (MLEs) for ρ per kilobase.

In our analysis of variable mutation rate for CpG sites in the LPL data (see Balding's question), we

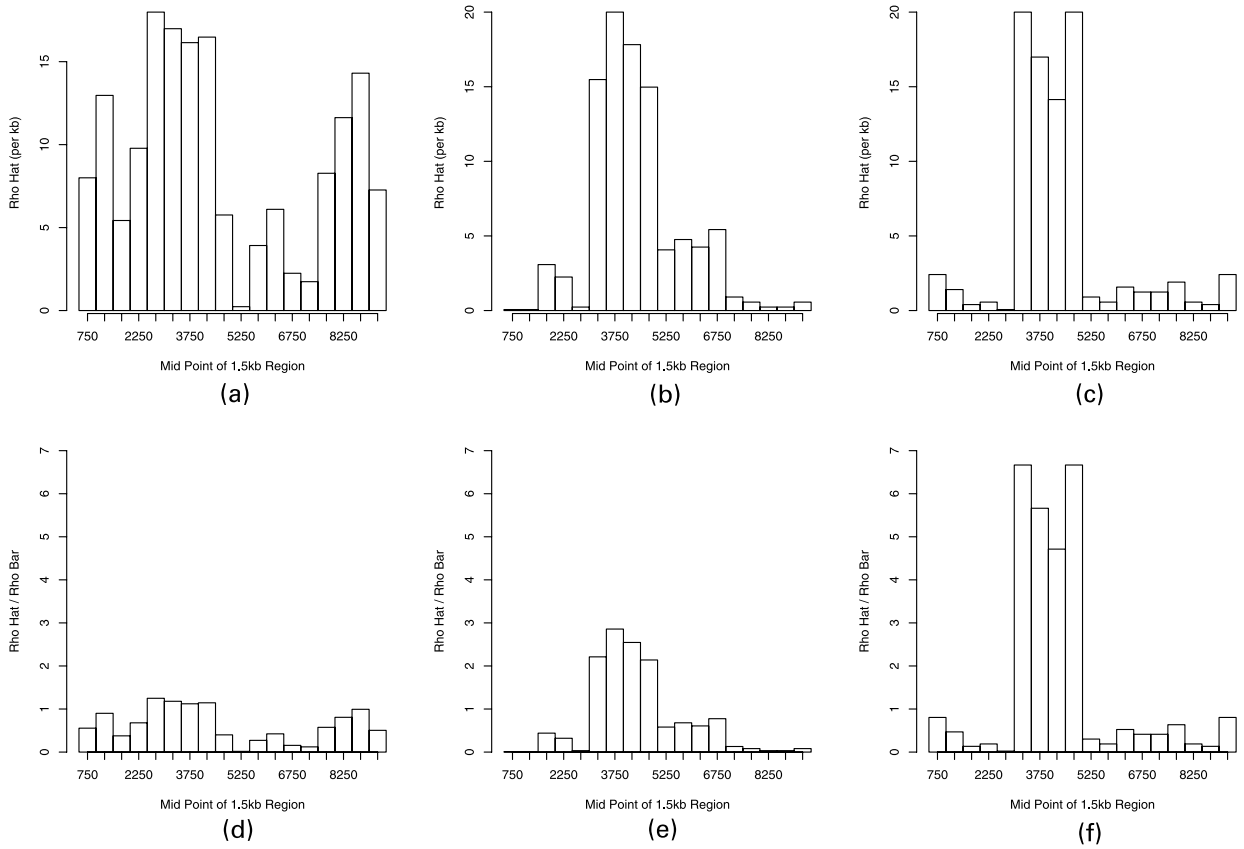

Fig. 6. Analysis of the variation in recombination rate across the LPL gene—each histogram shows either (a)–(c) estimates of the recombination rate per kilobase for 18 subregions or (d)–(f) the ratio of that estimate to the estimate of the recombination rate per kilobase, for the same population, based on all 9.75 kb of data (the subregions are all 1.5 kb long, except the last, which is only 1.25 kb long, and are spaced 0.5 kb apart, covering 9.75 kb of the gene; the estimates for each subregion are based on the approximate marginal likelihood, and the estimates for the complete sequence are based on the composite likelihood obtained from the approximate marginal likelihoods for 10 975 kb subregions): (a), (d) Jackson; (b), (e) Finland; (c), (f) Rochester

considered all pairs of sites which are currently CpG, or could have mutated from CpG sites in the history of the sample.

Myers's ingenious new approach to estimating the minimum number of recombination events in the history of a sample represents a real advance and leads to much better lower bounds than current methods do. It would be interesting to see whether it can be extended to allow for possible recurrent mutation. Our discussion of the LPL data suggests that recurrent mutation has not played a major role in shaping its patterns of genetic diversity. Assuming this to be so, Myers's analysis of the data, by a different method, reinforces our claim that there are many more recombination events in the history of the sample than were detected by the methods used by Templeton *et al*. (2000).

The reason (see Balding's question) for the substantial computational saving in our approximate marginal likelihood is that the number of recombination events in the history of the ancestral recombination graph is substantially reduced. For example, if the data were summarized by just two segregating sites, then recombinations can only occur to branches which are ancestral at both sites: if recombination rates are large, then most branches will be ancestral at just a single site. (In the infinite sites model, recombinations can still occur to branches which are ancestral at a single segregating site.) This saving is related to the size of the 'missing data' and should be equally helpful for approaches based on Markov chain Monte Carlo methods.

Carothers suggests (of all five papers) that the methods presented are not yet at the stage of being 'going concerns' for biological scientists. We think that this is harsh in general, and in the case of our methods for estimating recombination rates might have hoped that the publication and application in the mainstream biological literature of the full and approximate likelihood methods that we and others have developed

would be testament to their being rather further along the path to usefulness than his comments might suggest.

The possibility (Carothers) of checking robustness via k analyses of the data, each based on selecting every kth base, has considerable initial appeal. For typical data from humans, and moderate k, the (relatively few) segregating sites in the data will end up sparsely and perhaps unevenly spread among these subsets. The sparseness will lead to a considerable (legitimate) variation in estimated likelihood curves. An uneven spread of segregating sites among subsets may result in systematic differences. Further, the lack of underpinning theory means that there is no formal (or informal) way of comparing likelihood curves estimated from different subsets of the data to ask whether differences reflect systematic changes or just chance effects. We note that, if these problems could be overcome, there are also other ways of subdividing the data to provide robustness checks. For example one could select the whole sequence for distinct subsets of the sample, or subdivide on the basis of segregating sites.

A theory of 'approximately sufficient statistics' (Joyce) is tantalizing but we suspect inaccessible, at least in the genetics context. Our view is that the most profitable way forward will be to inform the choice of estimators and approximations by careful thought about the underlying stochastic models. We have not attempted a simulation study of the case $\rho = 0$, largely because real data are always subject to recombination (at least on current wisdom), so $\rho > 0$ in practice. As with other questions in this area, it is unclear whether the theoretical results to which Joyce refers would apply in the genetics setting.

As Meng suggests, it would be very nice to find a more quantitative measure than the comparisons of curves in Fig. 3 of our paper for the information loss through using our approximate marginal likelihood, rather than full likelihood. But this does not seem straightforward. Direct application of standard asymptotic measures such as the fraction of missing information (FMI) is problematical, as there is no relevant asymptotic theory, likelihood curves often appear far from quadratic and MLEs for one or both curves can often be on the boundary of the parameter space (notably at $\rho = 0$). An added complication is that as ρ becomes larger the inference problem becomes more difficult, and it becomes more likely that estimated full likelihood curves will underestimate the likelihood away from the peak (Stephens and Donnelly, 2000), hence overestimating the information in the full likelihood case. With all these *caveats*, we report that a simulation study (which ignored data sets in which the MLE was close to the boundary of the parameter space for either curve) suggests that, on the basis of the expected Fisher information, the FMI in using the approximate marginal likelihood relative to the full likelihood curve is small: 9% for 1 kb data ignoring singleton mutations, and 18% for 2 kb data in which only the five most common mutations, or all mutations with minor allele frequency over 30%, were retained. The FMI was larger when measured on the basis of the observed Fisher information: roughly 40% and 50% for the 1 kb and 2 kb approaches just described. For the reasons given above, we find all these numerical values, and the differences between the comparisons for observed and expected information, very difficult to interpret (although we suspect that the difficulties of accurately estimating the full likelihood curve for larger ρ are implicated in the latter comparison).

Meng's criticisms of our choice of terminology have substance. But it is a difficult area. Sustained and illuminating informal interactions with Meng and colleagues (notably Augie Kong) in their 'Chicago' days played a significant role in the way in which we (or at least PD) have thought about these questions, and we believe that the conceptualization as a missing data problem, with the unobserved genealogy as the missing data, is helpful. So the problem with referring to our 'marginal likelihood' approach as a 'partial data likelihood' is that the actual data (sequences sampled from the population) are not the 'complete data' in the missing data framework, and this may lead to more confusion. There are also different ways of viewing our 'composite likelihood'. In contrast with Meng's claim, it can be thought of as a real likelihood but for a different model in which certain long range dependences are set to zero.

It would be nice (Meng) if our marginal likelihood approach were more robust than the full likelihood to model misspecification, but this seems unclear. The log-transformation was considered by Hudson (2001) and seems sensible. The sandwich variance estimator is difficult (if not impossible) to use for our data, as there is no independence in the data—and it is unclear how to obtain the necessary consistent estimator of the correlation of the score function.

Bret Larget, Donald L. Simon and Joseph B. Kadane

We thank each of the discussants for their comments and observations. We are especially grateful to those discussants who recognize the novelty and promise of our approach, even as they point to many areas in which improvements are possible. Our rejoinders to these comments are organized by topic.

Selection

David Balding and Mark Pagel question our assumption of no selection. If it is the case that mitochondrial genome arrangement is not independent of fitness, the space of viable arrangements could be much smaller than the space of possible arrangements. Although this may greatly affect the calculated probabilities of sequences of certain gene inversions, it is not apparent that this would necessarily cause a bias in our methods. Other biologists with whom we have spoken feel that the assumption of no selection is not bad. In any case, too little is known currently about how genome arrangement might affect fitness to model it effectively. The disparity in mitochondrial genome arrangements within some phyla (e.g. Brachiopoda and Mollusca) suggests that vastly different mitochondrial genome arrangements must be viable.

Uniform distribution on inversions

Andrew Carothers questions our assumption that all possible gene inversions are equally likely and suggests that a distribution that accounts for the lengths of the segments may fit better. There is evidence that both large and small inversions do occur. Sea-star and sea-urchin mitochondrial genome arrangements differ by a single gene inversion that accounts for only three of the non-transfer ribonucleic acid (tRNA) genes, but the inverted segment includes several tRNA genes and includes about 30% of the mitochondrial genome (by base pair). For the complete data set in our paper, each gene inversion partitions the 15 non-tRNA genes into two non-empty sets. The number of genes in each part of the realized gene inversions in the simulations occur with these relative frequencies: 1/14 (13.5%), 2/13 (11.5%), 3/12 (15.9%), 4/11 (11.1%), 5/10 (12.6%), 6/9 (15.6%) and 7/8 (19.8%). These relative frequencies are fairly close, which suggests that the uniform distribution assumption might not be too bad.

Andrew Carothers and Mark Pagel also ask about gene inversions that cut genes in two. We assume that these would be fatal and not seen. If we assume that any two points in the genome could be cut with equal probability and do not assume selection, then the probability of each possible inversion will be a function of the spaces between genes, which is not uniform. Locations of the origins of replications on each strand as well as the D-loop region might have effects that we ignore.

In general, we agree in principle that the model that we assume in our paper is much simpler than the actual processes that rearrange genomes. When these processes are better understood we shall be able to develop more complex and realistic models.

The prior distribution

Susan Holmes suggests that a uniform prior on all possible tree topologies is unrealistic. This prior represents the opinion of a hypothetical individual who knows nothing or chooses to ignore all other information including previous taxonomic classifications. It is valid to consider the conclusions of such an individual and, in fact, any strong conclusions would be strong for most priors.

However, it would be better to consider other possible prior distributions as well. If we make the obvious but crude assumption that each of the eight animal phyla in our data set must remain monophyletic (and form subtrees in the unrooted tree), this essentially puts a prior probability of 0 on much of the tree space. Rerunning the simulations under this prior places posterior probabilities of 0.05, 0.03 and 0.02 on the three unrooted tree topologies that are consistent with the proposed phylogeny on the right of our Fig. 1. We do not sample a single tree in which the brachiopods and the deuterostomes form a clade, so there is no support for the traditional phylogeny on the left of our Fig. 1.

Summarizing the posterior

David Balding criticizes the inadequate summary of the posterior distribution. Truth be told, the summary in the paper is more of a function of the short time between when the simulations were completed and the final deadline for the paper than the limitations of what can be summarized. There is considerable information in the posterior distribution, but summarizing it more satisfactorily remains an open problem. The prior distribution in the paper places equal probability on the 6.3×10^{18} unrooted tree topologies relating 19 taxa. Our estimated posterior distribution puts half of the posterior probability on a set of about 1500 tree topologies and 90% on a set of about 25000 tree topologies. No single tree topology has a very substantial posterior probability, but the posterior distribution is much more concentrated than the prior distribution.

Comparison with York et al. (2002)

York *et al.* (2002) describe a Bayesian approach to reconstruct the sequence of gene inversions along a single branch. Although their example includes only two taxa compared with our 19, it has 79 markers compared with our 15. The update mechanism in York *et al.* (2002) is quite similar to that in our update 2,

except that they apply it to parts of the sequence of gene inversions on a branch whereas we always apply this update to an entire branch. For single branches with more than a few gene inversions, there may be a computational advantage to their approach.

Difficulty with large tree spaces

Korbinian Strimmer states that searching tree spaces when there are many taxa is likely to be difficult. We agree that current Markov chain Monte Carlo algorithms may need to be adapted to mix effectively in larger spaces. However, the computational complications in adapting Bayesian methods to larger spaces are likely to be much easier to overcome than in doing a similar thing with maximum likelihood because, in general, integration in large parametric spaces is more computationally efficient than optimization.

Comparisons with sequence data

David Balding, Andrew Carothers and Susan Holmes would all like to see comparisons with sequence data. We agree that it is an interesting open problem to determine how much information there is in arrangement data relative to that in sequence data. We are at present working on the problem of jointly using both sequence and arrangement data to infer phylogeny.

Comparison with parsimony

Susan Holmes takes us to task for ignoring 20 years of contemporary research of nonparametric and semiparametric methods of phylogenetic inference. She questions whether at this early stage in the development of methods for phylogenetic inference from genome arrangements nonparametric methods might not be very useful. We welcome the opportunity to clarify our opinions on this topic.

Holmes (2002) describes two goals of statistical methods to the phylogeny problem. These are to estimate the tree and to provide a confidence statement associated with the estimator. In the well-studied case of estimating phylogeny from deoxyribonucleic (DNA) sequence data, the methods based on maximum likelihood, parsimony and distance all produce estimates of the tree. Confidence statements are nearly always produced by applying the bootstrap with 'bootstrap support values' attached to clades of the estimated tree. Bayesian methods produce posterior distributions on the space of all parameters. One summary of this posterior distribution is the most probable tree topology with associated posterior probabilities on each clade.

Bayesian posterior probabilities are distinguished from bootstrap support values in the strength of existing theory to justify their interpretation. Although recognizing that model misspecification (in the prior and the likelihood) and computational difficulties associated with Markov chain Monte Carlo sampling can lead to serious errors in estimation, the Bayesian approach to phylogenetic inference provides clearly interpretable assessments of uncertainty. In contrast, there are weaknesses in the theoretical justifications for statistical interpretations of bootstrap support values associated with reconstructed phylogenies. Holmes (2002) lists several potential problems with the bootstrap. She reports,

> 'The [estimated tree topology] is a discrete statistic for which no applicable theory exists for the use of the bootstrap with reasonable amounts of data'

and

> 'The [estimated tree topology] is based on a maximum. It is well-documented that bootstrapping doesn't work for maximums of random variables.'

Despite these potential problems, the bootstrap does provide reasonable approximate p-values (Efron *et al.*, 1996) for DNA sequence data sets.

However, as David Balding points out, genome arrangement is only a single character and so cannot be naturally thought of as a realization of independent observations in the same way that sites in a DNA sequence can. With our present understanding, it is not at all obvious how the bootstrap could be applied to assess uncertainty in the context of genome arrangements.

Using genome arrangement data, Moret *et al.* (2001) and Bourque and Pevzner (2002) described two different heuristic methods to search for most parsimonious reconstructions and applied them to 'one of the most challenging genome rearrangement data sets', a data set of chloroplast genome arrangements with 105 markers from 12 Campanulaceae genera plus the outgroup tobacco. Moret *et al.* (2001) reported 216 most parsimonious trees, each of which requires 67 total gene inversions, whereas Bourque and Pevzner (2002) reported a single tree with 65 total gene inversions. Neither addressed uncertainty in the estimates. When applied to this data set, the computational approach in our paper consistently samples

the same set of 180 tree topologies that each require only 64 total gene inversions, and does so after much less computation (Larget *et al.*, 2002).

We agree that this is an early stage in the development of methods to infer phylogeny from genome arrangements and that subsequent work by us and others will result in more realistic models and ultimately a better tool for the biologist. However, even aside from our philosophical preference for the Bayesian point of view, we believe that there is little to recommend a parsimony approach to phylogenetic inference from genome arrangement data because the current state of the art provides estimates that may not be accurate, does not assess uncertainty and is computationally relatively slow.

Software

Susan Holmes wishes that we distributed software that implements the methods of this paper. It is our intention to do so in a future release of BAMBE (Simon and Larget, 2001). We are an even longer time away from distributing code to produce movies that are similar to that shown at the meeting (for which we are greatly indebted to MarkDerthick). A couple of sample movies, however, are available on the Web (`http://www.cs.cmu.edu/~sage/animations/`).

We thank all the discussants and hope that this paper and the related discussion spur further interest by statisticians in an exciting new area.

George Nicholson, Albert V. Smith, Frosti Jónsson, Ómar Gústafsson, Kárl Stefánsson and Peter Donnelly

It is a pleasure to thank the discussants for their interest, and their stimulating comments.

In one way or another much of the discussion of our paper relates to the underlying model and associated issues of parameterization. It might be helpful to list the model choices arising in the paper and the discussion. In each case we refer to the marginal model for the population allele frequencies, one of the αs in our notation, conditional on the ancestral allele frequency (the appropriate π). Data at that locus in that population consist of a binomial sample from the population. We are grateful to Beaumont for pointing out the development and antecedents of these approaches (indeed, the paper read to the Society by Edwards (1970) considered a closely related problem):

(a) the Balding–Nichols model, with a beta distribution for the α;
(b) the model of our paper, with a truncated normal distribution augmented by atoms at 0 and 1;
(c) the 'pure drift' model in which the distribution of α is given by the transition density of the Wright–Fisher diffusion, or an equivalent representation in terms of the coalescent;
(d) Griffiths's suggestion of generalizing the pure drift model to allow for variation in subpopulation sizes;
(e) Griffiths's suggestions of approximating the exact distributions from (c) and (d) above for small times;
(f) what Weir and Hill refer to as the 'drift model' by which we understand them to mean the case in which α is distributed according to the transition distribution of the discrete Wright–Fisher model (so that (c) above is an approximation to this for large population size);
(g) the model introduced by Cavalli-Sforza and Edwards in which allele frequencies evolve as Brownian motion—this will also lead to normally distributed α, but without the variance parameterized as a multiple of $\pi(1 - \pi)$.

In addition there is the tradition referred to in the paper of only specifying necessary moments for the α.

To make our position absolutely clear, we see the issues about exactly which model to fit as interesting but secondary. There will soon be large amounts of single-nucleotide polymorphism (SNP) data. Our paper has advocated and implemented

(i) a concise and easily interpretable parameterization which exploits the fact that within a population most SNP loci should be subject to the same probability model and hence that, with reasonable numbers of loci, powerful inferences can be made by combining information across loci,
(ii) a fully Bayesian approach to inference,
(iii) incorporation of SNP ascertainment and
(iv) model checking, which had previously been notable for its absence from most analyses;

and for what it is worth we believe that this combination is useful and novel. As the amount of data grows, it will be possible, and for many purposes essential, to refine models to capture correlations across populations, so that none of the models listed above will be adequate. As we noted in Section 6 of the

paper, the particular normal model that we introduced happens to be easy to extend to allow for these correlations.

Any explicit stochastic model for evolution, such as the discrete Wright–Fisher model, its diffusion or coalescent approximation or the island model, is at best a coarse approximation to the complicated reality of the evolution of real populations of humans or other organisms. It thus seems to us inappropriate to elevate one or several such models to a position of primacy. For example it seems unclear *a priori* which of the models (f), (e), (d), (c) or our humble normal model (b) above will actually fit real data better. If one of these models captures important features of real data, we would expect them all to. For one of them to fit *much* better than the others would seem surprising, but this can now be investigated empirically. In this sense we disagree with the perspective of several discussants that, even in the splitting setting, one of these models is 'right', and the others are approximations to it. We hope that all are approximations to reality. The model that we suggested is the simplest, making it easier to understand and easier to fit computationally, which may be practically relevant for data from many loci.

We share Balding's evident disappointment that our paper did not contain an extended discussion of the pros and cons of our approach as compared with that of Balding and Nichols; however, it was written under stringent constraints on length. But we did explain (albeit briefly) what we saw as the five major differences between the approaches, and the reasons for our choice of approach rather than theirs, in Section 1 of our paper, and it seems harsh to accuse this of not being 'serious'. We return to some of these in the light of Balding's comments.

The hierarchical model of Balding and Nichols (1997) introduced the most general parameterization in this setting (effectively one parameter for each data point), so of course any subsequent parameterization is a special case of this. Interestingly, Balding and Nichols (1997) explicitly rejected the possibility of a common parameter (F there; c in our notation) across loci for the data that they examined. We share Carothers's view that the power of our approach is a direct consequence of our adoption of a common c for all loci within a population. This is not to criticize Balding and Nichols. It was not appropriate for the genetic systems that they were analysing; it does seem to be appropriate, and very helpful, for SNP loci.

SNP data definitely have the property that the allele frequency can be 0 or 1 in some populations, because the variation at that locus has been lost there. Balding claims this to be 'rare' for human populations but, as our Fig. 5(b) suggests, it is far from rare for our second data set. In fact, for 20 of the 66 loci, at least one of the populations has a sample relative allele frequency of 0 or 1. We continue to take the view that, other things being equal, a model which explicitly captures this feature of the data is preferable in principle to one which does not.

Ultimately, whether the beta–binomial or the normal–binomial model is 'better' is an empirical question. We are encouraged by the analysis of Marchini and Cardon which shows that the normal model fits one of the two data sets better, and the other not 'significantly' worse, than the beta model, by one current model choice criterion. However, we share Balding's view that the issues, especially near the boundaries at 0 and 1, are subtle. In summary, we see our new parameterization as more important than the choice of a normal rather than a beta distribution and, moving forwards, we view the need for a revision of these models (to capture correlations across populations) as more important than comparisons between them.

We are muddled by two of Balding's criticisms. Any model will have the property that inferences will in general differ when different data are analysed. It is not just our approach that will give different answers when different subpopulations are studied or when sample sizes are changed! And relating to our lack of a simulation comparison between the beta and normal models—for fully Bayesian estimation, and in most settings for the almost fully Bayesian approach of Balding and Nichols, inference will be better when the inference model coincides with the simulation model. So it is not clear what we would learn by fitting the Balding–Nichols model to our simulated data, or by fitting our model to data simulated under their model. It may be that Balding had in mind a study similar to the one that Beaumont reported, concerning the power to distinguish the models on simulated data, which is very interesting.

Griffiths's demonstration that the pure drift model can be extended to allow for population growth, effectively only at the cost of some reparameterization, is elegant and interesting, as is his suggestion of easing the computational burden in this 'exact' approach by exploiting available normal or Poisson approximations to line-of-descent distributions in the coalescent.

Marchini and Cardon draw attention to the complementary problem to the one discussed in our paper, namely that of inferring population structure when it is not known which sampled individuals belong to which populations (nor in general how many populations there are). It is nice that the use of our parameterization may offer an improvement over available approaches in this context for SNP data.

Weir and Hill remind us of the traditional approach of modelling sampling distributions rather than allele frequencies. Thus for SNPs we could specify conditional probabilities of the form

$$P(n\text{th allele sampled is of type } A | j \text{ type } A \text{ and } n-1-j \text{ type } B \text{ in sample of } n-1).$$

Of course, if we did model the distribution of allele frequencies in the population, the sampling distributions are just ratios of appropriate moments of these population allele frequencies. We interpret Hill and Weir's comment about the relationship between the normal distribution and independence of samples of various sizes to mean that for our model these sampling distributions for $n > 3$ can be written in terms of those for $n = 1$ and $n = 2$. But this is simply a consequence of the fact that we have chosen a two-parameter (here c and π—Hill and Weir's comments should be interpreted as conditional on π) distribution for population allele frequencies. For *any* two-parameter model for allele frequencies (including the Balding–Nichols model and the pure drift model (c) above), all sampling distributions will be a function only of these two parameters, and so typically of the sampling distributions for samples of size 1 and 2. The exact drift model (e) does not lead to a two-parameter family for allele frequency distributions (although in a sense the additional parameters are of order N^{-2} and smaller, where N is the population size). We also welcome the new estimator suggested by Weir and Hill and look forward to understanding its performance in this context.

Giovanni Parmigiani, Elizabeth S. Garrett, Ramaswamy Anbazhagan and Edward Gabrielson

We thank all the discussants for their insightful comments and useful suggestions. There are common themes in many of the points raised, so we shall proceed by theme in our rejoinder.

Role of statistics in molecular classification

We found Professor Carothers's description of the process by which the development of novel methodological techniques interacts with technological developments in the life sciences very insightful. We agree with his categorization of our work as being in the middle stages of the process. High throughput gene expression measurements are contributing at a rapidly increasing rate to our understanding of several biological systems. There is a spectrum of biological investigations that is made possible by these technologies. At one end of this spectrum we have highly controlled and replicated comparisons such as treating *versus* not treating two groups of genetically identical mice. In these applications the signal-to-noise ratios are relatively favourable, and statistical questions, albeit difficult to address, are well defined. At the opposite end we have what we could describe as 'genome biometry', i.e. the description of the variability of genomic information in human populations. In this arena, signal-to-noise ratios are far less favourable, and the more exploratory nature of the biological investigation makes the statistical questions less well defined. Although solid statistical methods are critical at both extremes, in genome biometry, statistical methods cannot, alone, take us from data to answers, and, we believe, need to be built to enhance the exploratory endeavours rather than to replace them. Our model is an effort in this direction. Because we intend to contribute to an exploratory investigation, we see our approach as a complement, rather than as an alternative, to existing techniques. We ask the user to pay a price in terms of computing time and model checking; we may offer a pay-off in terms of data simplification, robustness against noise and the interpretability of results.

Variability and mixtures

In our molecular classification approach, variability in the array matrix is described probabilistically both across genes and across arrays. We used mixture models in both domains, in a hierarchical way. First, sample-to-sample variation is described by a three-component mixture; then the corresponding parameters are described using a continuous mixture across genes. The difference is somewhat artificial, as expression variation in populations could be smooth, and some quantitative information may be both reliably measured and biologically meaningful. We hope that the ability to define a profile in a comparatively more interpretable and reproducible way may offset any loss of information.

Ambiguities

The potential ambiguities in the definition of the class $e = 0$ have been identified by us, as well as some of the discussants, as the most important limitation of our approach. Ambiguities arise when there are multiple genes in which samples cluster in two groups of roughly equal size. The data presented here paint a somewhat pessimistic picture in this regard. In our experience with applying our methodology to several other data sets, extensive occurrences of two groups of roughly equal size are not common.

More importantly, many biologists are now augmenting molecular profiling data on tumours with data on normal tissue; examples include Bhattacharjee *et al*. (2001), Garber *et al*. (2001) and Alizadeh *et al*. (2000). When normal samples are available, the class $e = 0$ can be defined on the basis of the expression in normal tissue. Although this will also vary across individuals, many potential ambiguities will be avoided and a direct biological interpretation will be more reliable. From this standpoint, the completely unsupervised design that was investigated here is the most challenging and may be setting the bar higher than it needs to be.

A constraint on π_g could avoid gross misbehaviour, such as very small normal components. It may be useful and it is straightforward to implement in the context of Markov chain Monte Carlo (MCMC) sampling.

Distribution assumptions

Several discussants touched on the choice of distributional assumptions. These should, as elsewhere, be specific to the data at hand and should be checked. Naturally, no distributional choice will work everywhere. Checking assumptions for genomic distributions is similar to traditional hierarchical models. We examined a small sample of *qq*-plots by using sampled values from the MCMC output. These *qq*-plots, however, depend on the distributional assumptions themselves; approaches to dampen this dependence are also available in some cases, as illustrated, for example, in Dominici *et al*. (1999).

Checking assumptions on the fs is more challenging. A typical analysis involves several thousands of genes. A gene-specific model choice and a gene-specific number of mixture components are possible but would be computationally very demanding. We experimented with various choices of distributions and found the uniform distribution to perform reliably over the large number of shapes of differentially expressed genes that arise in an array. Any distribution that has decaying tails runs the risk of classifying some of the observations based on ratios of infinitesimal densities and leads to instabilities. As our experience with new data sets expands, we look forward to validating this choice further, or to improve on it. Our current software implementation, described in Garrett and Parmigiani (2003), relies on uniform distributions, but we like the idea of user-specified fs. Providing a reasonable menu of choices may be feasible and worthwhile.

Some measure of diagnostics of goodness of fit is still possible even without looking at all the tens of thousands of gene-specific *qq*-plots, by mining for genes that may not fit well, as we discuss in the paper. We have found plotting all points in the *qq*-plot useful, even though it admittedly requires a visual inspection that is not the same as that of traditional *qq*-plots. We usually check two aspects: that the points attributed to the $e = 0$ class are linear and that the points attributed to other classes are not on the same line as those attributed to $e = 0$. When that occurs, classification probabilities are likely to be reliable. For example, the results in our Figs 3(f) and 3(g) are satisfactory. The results in Fig. 3(h) suggest that there may be two subgroups of underexpressed genes and indicate that the lowest of these two subgroups could have constituted an alternative $e = 0$ class.

The extremes of the uniform components of the mixture are unknown and distributed according to an exponential across genes. Therefore the support of the marginal distribution of the data is not bounded for any of the three classes. During MCMC sampling, however, it is possible that some of the observed residuals will be larger than the sampled κs for outlying samples and therefore assigned to the normal component. When these assignments persist, increasing the value of κ_0 solves the problem. An alternative is to add an exponential tail with small mass directly to the uniform distribution.

Measure of similarity

The measures of similarity that we proposed in our paper for profiles of pairs of genes can be sensitive to small probabilities and unstable. We see classification probabilities as building-blocks for interpretable distance measures, and we welcome the suggestions made for alternative ways of building pairwise distances.

Normalization

Preliminary normalization of the data is critical to this as well as to most microarray analyses. Normalization is a complex iterative process of identification of likely artefacts followed by a decision about whether the artefact can be corrected or calls for the removal of spots or arrays. In our case, we identified changes in gene expression that did not display the type of variability that we would expect from a biological phenomenon, because of the large number of genes involved and the marked separation of the two subgroups. Also, the few samples showing these changes were all hybridized within a relatively short interval in time.

Supervised analyses

In a supervised analysis, in addition to the expression matrix A we have phenotype information for each of the samples. The interest is then in building classification models, in which a relatively small number of genes predict the phenotype reliably. Several of the outputs of our analysis can provide potentially useful transformed predictors for supervised analyses. These include $E(\eta_{gt}|a_{gt}, \omega)$ and $\hat{p}^{+}_{gt} - \hat{p}^{-}_{gt}$ and ternary categorization built on expression probabilities. More formal integration of the mixture model into the supervised analysis would, however, be desirable and could be constructed by postulating relationships between phenotype and latent classes. These could be evaluated by using an MCMC technique similar to what we described. Using latent classes involves a trade-off between denoising and a loss of potentially relevant quantitative information. In prediction models, this trade-off would probably need to be resolved differently from that in unsupervised models, as the phenotype provides additional biological interpretability. Some researchers have started to investigate this issue, but further exploration would be needed in the context of our approach.

References in the discussion

Albert, J. H. and Chib, S. (1993) Bayesian analysis of binary and polychotomous response data. *J. Am. Statist. Ass.*, **88**, 669–679.

Aldous, D. A. (1996) *Probability Distributions on Cladograms*, pp. 1–18. New York: Springer.

———(2001) Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Statist. Sci.*, **16**, 23–34.

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C. A., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Levy, R., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O. and Staudt, L. M. (2000) Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511; comment; **403**, 491–492; **404**, 921.

Baker, D. and Sali, A. (2001) Protein structure prediction and structural genomics. *Science*, **294**, 93–96.

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

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

Ball, R. D. (2001) Bayesian methods for quantitative trait loci mapping based on model selection: approximate analysis using the Bayesian information criterion. *Genetics*, **159**, 1351–1364.

Basten, C., Weir, B. S. and Zeng, Z.-B. (2002) QTL Cartographer. Department of Statistics, North Carolina State University, Raleigh. (Available from `http://statgen.ncsu.edu/qtlcart/cartographer.html`.)

Beaumont, M. A. (2001) Conservation genetics. In *The Handbook of Statistical Genetics* (eds D. J. Balding, M. Bishop and C. Cannings). New York: Wiley.

Bedrick, E. J. and Tsai, C.-L. (1994) Model selection for multivariate regression in small samples. *Biometrics*, **50**, 226–231.

Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E. J., Lander, E. S., Wong, W., Johnson, B. E., Golub, T. R., Sugarbaker, D. J. and Meyerson, M. (2001) Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natn. Acad. Sci. USA*, **98**, 13790–13795.

Billera, L., Holmes, S. and Vogtmann, K. (2001) The geometry of tree space. *Adv. Appl. Math.*, 771–801.

Bourque, G. and Pevzner, P. (2002) Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Res.*, **12**, 26–36.

Box, G. E. P. (1976) Science and statistics. *J. Am. Statist. Ass.*, **71**, 791–799.

Broman, K. W. and Speed, T. P. (1999) A review of methods for identifying QTLs in experimental crosses. *IMS Lect. Notes Monogr. Ser.*, **33**, 114–142.

Brown, P. J., Vannucci, M. and Fearn, T. (1998) Multivariate Bayesian variable selection and prediction. *J. R. Statist. Soc.* B, **60**, 627–641.

Carson, S. D., Djorovic, N., Djorovic, A., Wilcox, P. and Ball, R. (2002a) Simulation of QTL detection and MAS for quantitative traits: I, Impact of population size, underlying genetic structure, and criteria for choosing markers. To be published.

———(2002b) Simulation of QTL detection and MAS for quantitative traits: II, Comparison of gain and selection bias for alternate experimental designs including selective genotyping and map density. To be published.

Cavalli-Sforza, L. L. and Edwards, A. W. F. (1967) Phylogenetic analysis: models and estimation procedures. *Evolution*, **32**, 550–570.

Churchill, G. A. and Doerge, R. W. (1994) Empirical threshold values for quantitative trait mapping. *Genetics*, **138**, 963–971.

Ciofi, C., Beaumont, M. A., Swingland, I. R. and Bruford, M. W. (1999) Genetic divergence and units for conservation in the Komodo Dragon *Varanus komodoensis*. *Proc. R. Soc. Lond.* B, **266**, 2269–2274.

Clark, A. G., Weiss, K. M., Nickerson, D. A., Taylor, S. L., Buchanan, A., Stengård, J., Salsmaa, V., Vartiainen, E., Perola, M., Boervinkle, E. and Sing, C. F. (1998) Haplotype structure and population genetic inferences from nucleotide sequence variation in human Lipoprotein Lipase. *Am. J. Hum. Genet.*, **63**, 595–612.

Dawid, A. P. and Pueschel, J. (1999) Hierarchical models for dna profiling using heterogeneous databases. In *Bayesian Statistics 6* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 187–212. Oxford: Oxford University Press.

De Koning, D. J., Visscher, P. M., Knott, S. A. and Haley, C. S. (1998) Strategies for QTL detection in half sib populations. *Anim. Sci.*, **67**, 257–268.

Demchuk, E. and Singh, H. (2001) Statistical thermodynamics of hindered rotation from computer simulations. *Molec. Phys.*, **99**, 627–636.

Demchuk, E., Singh, H., Hnizo, V., Mardia, K. V. and Sharp, D. S. (2001) Statistics and molecular structure of biological macromolecules. In *Proc. Functional and Spatial Data Analysis* (eds K. V. Mardia and R. G. Aykroyd), pp. 9–14. Leeds: Leeds University Press.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc.* B, **39**, 1–38.

Dominici, F., Parmigiani, G., Wolpert, R. L. and Hasselblad, V. (1999) Meta-analysis of migraine headache treatments: combining information from heterogeneous designs. *J. Am. Statist. Ass.*, **94**, 16–28.

Drezner, Z., Marcoulides, G. A. and Salhi, S. (1999) Tabu search model selection in multiple regression analysis. *Communs Statist.*, **28**, 349–367.

Edwards, A. (1970) Estimation of the branch points of a branching diffusion process (with discussion). *J. R. Statist. Soc.* B, **32**, 155–174.

Edwards, M. D. and Page, N. J. (1994) Evaluation of marker-assisted selection through computer-simulation. *Theor. Appl. Genet.*, **88**, 376–382.

Efron, B., Halloran, E. and Holmes, S. (1996) Bootstrap confidence levels for phylogenetic trees. *Proc. Natn Acad. Sci. USA*, **93**, 13429–13434.

Fearnhead, P. N. and Donnelly, P. (2001) Estimating recombination rates from population genetic data. *Genetics*, **159**, 1299–1318.

Felsenstein, J. (1973) Maximum-likelihood estimation of evolutionary trees from continuous characters. *Am. J. Hum. Genet.*, **25**, 471–492.

———(1981a) Evolutionary trees from DNA sequences: a maximum-likelihood approach. *J. Molec. Evoln*, **17**, 368–376.

———(1981b) Evolutionary trees from gene frequencies and quantitative characters: finding maximum likelihood estimates. *Evolution*, **35**, 1229–1242.

Fisher, R. A. (1918) The correlation between relatives on the supposition of Mendelian, inheritance. *Trans. R. Soc. Edinb.*, **52**, 399–433.

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

Fridlyand, Y. J. M. (2001) Resampling methods for variable selection and classification: applications to genomics. *PhD Dissertation*. Statistics Department, University of California at Berkeley, Berkeley.

Friedman, N., Linial, M., Nachman, I. and Pe'er, D. (2000) Using Bayesian networks to analyze expression data. *J. Comput. Biol.*, **7**, 601–620.

Frisse, L., Hudson, R. R., Bartoszewicz, A., Wall, J. D., Donfack, J. and Di Rienzo, A. (2001) Gene conversion and difference population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *Am. J. Hum. Genet.*, **69**, 831–843.

Gaffney, P. J. (2001) An efficient reversible jump Markov chain Monte Carlo approach to detect multiple loci and their effects in inbred crosses. *PhD Dissertation*. Department of Statistics, University of Wisconsin—Madison, Madison.

Garber, M. E., Troyanskaya, O. G., Schluens, K., Petersen, S., Thaesler, Z., Pacyna-Gangelbach, M., van de Rijn, M., Rosen, G. D., Perou, C. M., Whyte, R. I., Altman, R. B., Brown, P. O., Botstein, D. and Petersen, I. (2001) Diversity of gene expression in adenocarcinoma of the lung. *Proc. Natn. Acad. Sci. USA*, **98**, 13784–13789.

Garrett, E. S. and Parmigiani, G. (2003) POE: statistical tools for molecular profiling. In *The Analysis of Gene Expression Data: Methods and Software*. New York: Springer.

George, E. I. and McCulloch, R. E. (1993) Variable selection via Gibbs sampling. *J. Am. Statist. Ass.*, **88**, 881–889.

Glasbey, C. A. and Ghazal, P. (2002) Combinatorial image analysis of DNA microarray features. *Bioinformatics*, to be published.

Goffinet, B. and Mangin, B. (1998) Comparing methods to detect more than one QTL on a chromosome. *Theor. Appl. Genet.*, **96**, 628–633.

Goldman, N., Anderson, J. P. and Rodrigo, A. G. (2000) Likelihood-based tests of topologies in phylogenetics. *Syst. Biol.*, **49**, 652–670.

Greenfield, S. A. (2000) *The Private Life of the BRAIN*. London: Lane.

Griffiths, R. C. (1984) Asymptotic line of descent distributions. *J. Math. Biol.*, **21**, 67–75.

Hackett, C. A., Bradshaw, J. E. and McNicol, J. W. (2001) Interval mapping of QTLs in autotetraploid species. *Genetics*, **159**, 1819–1832.

Hackett, C. A., Meyer, R. C. and Thomas, W. T. B. (2001) Multitrait QTL mapping in barley using multivariate regression. *Genet. Res.*, **77**, 95–106.
Haley, C. S. and Knott, S. A. (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*, **69**, 315–324.
Hannenhalli, S. and Pevzner, P. A. (1995a) Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). *J. Ass. Comput. Mach.*, **46**, 1–27.
———(1995b) Transforming men into mice (polynomial algorithm for the genomic distance problem). In *Proc. 36th A. Symp. Foundations of Computer Science*, pp. 581–592. New York: Institute of Electrical and Electronic Engineers.
Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*. London: Chapman and Hall.
Hilliker, A. J., Harauz, G., Reaume, A. G., Gray, M., Clark, S. H. and Chovnick, A. (1994) Meiotic gene conversion tract length distribution within the *rosy* locus of *Drosophila melanogaster*. *Genetics*, **137**, 1019–1026.
Hoeschele, I. (2001) Mapping quantitative trait loci in outbred pedigrees. In *Handbook of Statistical Genetics* (eds D. J. Balding, M. Bishop and C. Cannings), pp. 599–644. Chichester: Wiley.
Holmes, S. (2002) Phylogenies: a statistician's perspective. *Theor. Popln Biol.*, **62**, in the press.
Hoyle, D. C., Rattray, M., Jupp, R. and Brass, A. (2002) Making sense of microarray data distributions. *Bioinformatics*, **18**, 576–584.
Hudson, R. R. (2001) Two-locus sampling distributions and their application. *Genetics*, **159**, 1805–1817.
———(2002) Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337–338.
Hudson, R. and Kaplan, N. (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, **11**, 147–164.
Hurn, M. A., Rue, H. and Sheehan, N. (1999) Block updating in constrained Markov chain Monte Carlo sampling. *Statist. Probab. Lett.*, **41**, 353–361.
Jansen, R. C. (1993) Interval mapping of multiple quantitative trait loci. *Genetics*, **135**, 205–211.
———(1994) Controlling the type I and type II errors in mapping quantitative trait loci. *Genetics*, **138**, 871–881.
———(2001) Quantitative trait loci in inbred lines. In *Handbook of Statistical Genetics* (eds D. J. Balding, M. Bishop and C. Cannings), pp. 567–597. Chichester: Wiley.
Jansen, R. C. and Stam, P. (1994) High resolution of quantitative traits into multiple loci via interval mapping. *Genetics*, **136**, 1447–1455.
Jeffreys, A. J., Kauppi, L. and Neumann, R. (2001) Intensely punctate meiotic recombination in the class ii region of the major histocompatibility complex. *Nat. Genet.*, **29**, 217–222.
Kao, C. H., Zeng, Z.-B. and Teasdale, R. D. (1999) Multiple interval mapping for quantitative trait loci. *Genetics*, **152**, 1203–1216.
Kaplan, N. and Hudson, R. R. (1985) The use of sample genealogies for studying a selectively neutral m-loci model with recombination. *Theor. Popln Biol.*, **28**, 382–396.
Knott, S. A., Elsen, J. M. and Haley, C. S. (1996) Methods for multiple marker mapping of quantitative trait loci in half-sib populations. *Theor. Appl. Genet.*, **93**, 71–80.
Lander, E. S. and Botstein, D. (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, **121**, 185–199.
Larget, B., Kadane, J. B. and Simon, D. L. (2002) A Markov chain Monte Carlo approach to reconstructing ancestral genome arrangements. Submitted to *Molec. Biol. Evoln.*
Lee, M. L., Kuo, F. C., Whitmore, G. A. and Sklar, J. (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natn. Acad. Sci. USA*, **97**, 9834–9839.
Li, N. and Stephens, M. (2002) A new multilocus model for linkage disequilibrium, with application to estimating recombination rates. To be published.
Madigan, D. and Raftery, A. E. (1994) Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Am. Statist. Ass.*, **89**, 1535–1546.
Maliepaard, C. and Van Ooijen, J. W. (1994) QTL mapping in a full-sib family of an outcrossing species. In *Biometrics in Plant Breeding: Applications of Molecular Markers* (eds J. W. Van Ooijen and J. Jansen).
McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.
McVean, G., Awadalla, P. and Fearnhead, P. (2002) A coalescent-based method for detecting and estimating recombination rates from gene sequences. *Genetics*, **160**, 1231–1241.
Meng, X. L. and Rubin, D. B. (1991) Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *J. Am. Statist. Ass.*, **94**, 899–909.
Miller, A. J. (1990) *Subset Selection in Regression*. New York: Chapman and Hall.
Moret, B. M. E., Wang, L., Warnow, T. and Wyman, S. (2001) New approaches for reconstructing phylogenies from gene order data. *Bioinformatics*, **17**, S165–S173.
Myers, S. R. and Griffiths, R. C. (2002) Bounds on the minimum number of recombinations in a sample history. *Genetics*, to be published.
Nikerson, D. A., Taylor, S. L., Weiss, K. M., Clark, A. G., Hutchinson, R. G., Stengard, J., Salomaa, V.,

Vartiainen, E., Boerwinkle, E. and Sing, C. F. (1998) DNA sequence diversity in a 9.7-kb region of the human Lipoprotein Lipase gene. *Nat. Genet.*, **19**, 233–240.
Piepho, H. P. and Gauch, H. G. (2001) Marker pair selection for mapping quantitative trait loci. *Genetics*, **157**, 433–444.
Pritchard, J. K., Stephens, M. and Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.
Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997) Bayesian model averaging for linear regression models. *J. Am. Statist. Ass.*, **92**, 179–191.
Rannala, B. and Hartigan, J. A. (1996) Estimating gene flow in island populations. *Genet. Res. Camb.*, **67**, 147–158.
Ranz, J. M., Casals, F. and Ruiz, A. (2001) How malleable is the eukaryotic genome?: extreme rate of chromosomal rearrangement in Drosophila. *Genome Res.*, **11**, 230–239.
Roeder, K., Escobar, M., Kadane, J. and Balazs, I. (1997) Measuring heterogeneity in forensic databases using hierarchical bayes models. *Biometrika*, **85**, 269–287.
Satagopan, J. M., Newton, M. A. and Raftery, A. E. (2000) Easy estimation of normalizing constants and Bayes factors from posterior simulation: stabilizing the harmonic mean estimator. *Technical Report 1028*. Department of Statistics, University of Wisconsin—Madison, Madison.
Satagopan, J. M. and Yandell, B. S. (1996) Estimating the number of quantitative trait loci via Bayesian model determination. *Joint Statistical Meet., Chicago*.
Schröder, E. (1870) Vier combinatorische Probleme. *Z. Math. Phys.*, **15**, 361–376.
Seaton, G., Knott, S. A., Kearsey, M. J., Haley, C. S. and Visscher, P. M. (2002) QTL Express: user-friendly software to map quantitative trait loci in outbred populations. *Bioinformatics*, **183**, 339–340.
Self, S. G. and Liang, K.-Y. (1987) Large sample properties of maximum likelihood estimator and the likelihood ratio test on the boundary of the parameter space. *J. Am. Statist. Ass.*, **82**, 605–610.
Sha, N., Vannucci, M., Brown, P. J. and Liò, P. (2002) Bayesian variable selection in multinomial probit models with application to spectral data and DNA microarrays. *Technical Report UKC/IMC/02/05*. University of Kent at Canterbury, Canterbury.
Sillanpää, M. J. and Arjas, E. (1998) Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics*, **148**, 1373–1388.
Sillanpää, M. J. and Corander, J. (2002) Model choice in gene mapping: what and why. *Trends Genet.*, **18**, 301–307.
Simon, D. and Larget, B. (2001) Bayesian analysis in molecular biology and evolution (BAMBE). (Available from `http://www.mathcs.duq.edu/larget/bambe.html`.)
Sisson, S. A. (2002) An algorithm to characterize non-communicating classes on complex genealogies. *Preprint*. University of Puerto Rico, Puerto Rico.
Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *J. R. Statist. Soc.* B, **64**, 583–639.
Stephens, D. A. and Fisch, R. D. (1998) Bayesian analysis of quantitative trait locus data using reversible jump Markov chain Monte Carlo. *Biometrics*, **54**, 1334–1347.
Stephens, M. and Donnelly, P. (2000) Inference in molecular population genetics (with discussion). *J. R. Statist. Soc.* B, **62**, 605–655.
Strimmer, K. and von Haeseler, A. (1996) Quartet-puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Molec. Biol. Evoln*, **13**, 964–969.
Strimmer, K. and Rambaut, A. (2002) Inferring confidence sets of possibly misspecified gene trees. *Proc. R. Soc. Lond.* B, **269**, 137–142.
Swofford, D., Olsen, G., Waddell, P. and Hillis, D. (1996) Phylogenetic inference. In *Molecular Systematics* (eds D. M. Hillis, C. Moritz and B. K. Mable), pp. 407–514. Sunderland: Sinauer.
Taylor, W. R., May, A. C. W., Brown, N. P. and Aszodi, A. (2001) Protein structure: geometry, topology and classification. *Rep. Prog. Phys.*, **64**, 517–590.
Templeton, A. R., Clark, A. G., Weiss, K. M., Nickerson, D. A., Boerwinkle, E. and Sing, C. F. (2000) Recombinational and mutational hotspots within the human Lipoprotein Lipase gene. *Am. J. Hum. Genet.*, **66**, 69–83.
Thompson, E. A. (2001) Monte Carlo methods on genetic structures. *Monogr. Statist. Appl. Probab.*, **87**, 175–218.
Thompson, J. N. (1975) Quantitative variation and gene numbers. *Nature*, **258**, 665–668.
Tuffley, C. and Steel, M. (1997) Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bull. Math. Biol.*, **59**, 581–607.
Venter (2001) *The Independent*, Feb. 12th.
Vetta, A. (1976) Evidence for polygenes. *Nature*, **261**, 525.
Vetta, A. and Capron, C. (1999) The mind does not work: review of "How the mind works" by Steven Pinker. *Curr. Psychol. Cogn*, **18**, 105–111.
Visscher, P. M. and Haley, C. S. (1996) Detection of putative quantitative trait loci in line crosses under infinitesimal genetic models. *Theor. Appl. Genet.*, **93**, 691–702.
———(1998) A chromosomal test to detect genetic variation using genetic markers. *Heredity*, **81**, 317–326.
Visscher, P. M., Thompson, R. and Haley, C. S. (1996) Confidence intervals for QTL locations using bootstrapping. *Genetics*, **143**, 1013–1020.

Visscher, P. M., Whittaker, J. C. and Jansen, R. C. (2000) Mapping multiple QTL of different effects: comparison of a simple sequential testing strategy and multiple QTL mapping. *Molec. Breed.*, **6**, 11–24.
Wang, S. C. (2000) Simulation study on the methods for mapping quantitative trait loci in inbred line crosses. *PhD Thesis*. Zhejiang University, Zhejiang. (Available from `http://statgen.ncsu.edu/zeng/Wang-Shengchu-Thesis.pdf.`)
Wang, S., Basten, C. and Zeng, Z.-B. (2002) WINDOWS QTL Cartographer. Department of Statistics, North Carolina State University, Raleigh. (Available from `http://statgen.ncsu.edu/qtlcart/WQTLCart.htm.`)
Weber, K., Eisman, R., Higgins, S., Kuhl, L., Patty, A., Sparks, J. and Zeng, Z.-B. (1999) An analysis of polygenes affecting wing shape on chromosome three in *Drosophila melanogaster. Genetics*, **153**, 773–786.
Weber, K., Eisman, R., Higgins, S., Morey, L., Patty, A., Tausek, M. and Zeng, Z.-B. (2001) An analysis of polygenes affecting wing shape on chromosome 2 in *Drosophila melanogaster. Genetics*, **159**, 1045–1057.
Whittaker, J. C., Thompson, R. and Visscher, P. M. (1996) On the mapping of QTL by regression of phenotype on marker-type. *Heredity*, **77**, 23–32.
Wright, S. (1931) Evolution in Mendelian populations. *Genetics*, **16**, 97–159.
Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J. and Speed, T. P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucl. Acids Res.*, **30**, e15.
Yi, N. and Xu, S. (2002) Mapping quantitative trait loci with epistatic effects. *Genet. Res.*, to be published.
York, T. L., Durrett, R. and Nielsen, R. (2002) Bayesian estimation of the number of inversions in the history of two chromosomes. *J. Comput. Biol.*, to be published.
Zeng, Z.-B. (1993) Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proc. Natn. Acad. Sci. USA*, **90**, 10972–10976.
———(1994) Precision mapping of quantitative trait loci. *Genetics*, **136**, 1457–1468.
Zeng, Z.-B., Kao, C.-H. and Basten, C. J. (1999) Estimating the genetic architecture of quantitative traits. *Genet. Res.*, **74**, 279–289.
Zeng, Z.-B., Liu, J., Stam, L. F., Kao, C.-H., Mercer, J. M. and Laurie, C. C. (2000) Genetic architecture of a morphological shape difference between two *Drosophila* species. *Genetics*, **154**, 299–310.