# A model selection approach for the identification of quantitative trait loci in experimental crosses, allowing epistasis

University of Wisconsin–Madison

Department of Biostatistics & Medical Informatics

Technical Report # 205

Ani Manichaikul[*], Jee Young Moon[†], Śaunak Sen[**],

Brian S. Yandell[†,‡], Karl W. Broman[§,1]

[*]Department of Biomedical Engineering, University of Virginia, Charlottesville, Virginia 22908, [†]Department of Statistics, [‡]Department of Horticulture and [§]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, Wisconsin 53706, [**]Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, California 94143

[1]Corresponding author, kbroman@biostat.wisc.edu

29 July 2008

ABSTRACT

The identification of quantitative trait loci (QTL) and their interactions is a crucial step toward the discovery of genes responsible for variation in experimental crosses. The problem is best viewed as one of model selection, and the most important aspect of the problem is the comparison of models of different sizes. We present a penalized likelihood approach, with penalties on QTL and pairwise interactions chosen to control false positive rates. We extend the work of Broman and Speed to allow for pairwise interactions. A conservative version of our penalized LOD score provides strict control over the rate of extraneous QTL and interactions; a more liberal criterion is more lenient on interactions but seeks to maintain control over the rate of inclusion of false loci. We illustrate the use of our model selection criteria as exploratory tools in the analysis of backcross and intercross data. Simulation studies demonstrate reasonable power to detect QTL. Our liberal criterion is comparable in power to two Bayesian approaches.

# INTRODUCTION

Quantitative traits, such as blood pressure and fasting glucose, are often affected by multiple genetic factors, called quantitative trait loci (QTL). The identification of such QTL can lead to improved understanding of molecular mechanisms behind such traits, and is the central goal of many experimental crosses involving inbred lines.

We will focus on the case of a backcross or intercross derived from two inbred parental lines, and on a continuously varying quantitative trait with normally distributed residual variation. The goals of a QTL mapping experiment include the identification of QTL and epistatic interactions, the derivation of interval estimates for the locations of the QTL, and the estimation of QTL effects. We will focus strictly on the identification of QTL and their interactions.

The simplest and most commonly used approach for QTL mapping is interval mapping (LANDER and BOTSTEIN 1989). One posits the presence of a single QTL and considers each genomic location, one at a time, as the putative location for the QTL. At each location, a LOD score is calculated, comparing the hypothesis of a single QTL at the given position to the null hypothesis of no QTL. Much of the focus has been on statistical significance correcting for the genome scan (that is, for the multiplicity of statistical tests that are performed). A significance threshold or corrected p-values are derived based on the distribution of the genome-wide maximum LOD score, under the null hypothesis of no QTL. This null distribution is commonly estimated via a permutation test (CHURCHILL and DOERGE 1994).

Despite being based on a single-QTL model, interval mapping has been remarkably useful. However, the consideration of multiple-QTL models has a number of advantages. First, by

controlling for a QTL with large effect, one may reduce the residual variation and so increase power to detect additional QTL of more modest effect. (This has been especially useful for organisms, such as Drosophila, in which initial selection experiments can result in traits of high heritability.) Second, one can better separate linked QTL: to decide whether there might be multiple QTL on a chromosome, one should first compare the fit of a two-QTL model to the best single-QTL model. Finally, the investigation of potential epistatic interactions among QTL requires the fit of multiple-QTL models. In this regard, one may be particularly concerned about QTL with limited marginal effects (so that they would not likely be identified through interval mapping) but important interactions.

The exploration of multiple-QTL models does not fit well into a hypothesis testing framework. The task is better viewed as one of *model selection* (also known as variable selection): one seeks to identify the set of QTL (and epistatic interactions) that are best supported by the data. There are a number of special features to the model selection problem in QTL mapping that make it unique. First, while much work on model selection has focused on building a model for prediction, and so on minimizing prediction error, our primary goal is to identify a good model; this involves balancing the two possible errors, of missing important loci (false negatives) and of including extraneous loci (false positives). Second, we are faced with a continuum of ordinal-valued covariates (the genetic loci). Finally, the covariates have a simple structure: loci on different chromosomes are independent, and along a chromosome there is a very simple (and effectively known) correlation structure. With the assumption of no crossover interference, genotypes at loci along a chromosome form a Markov chain. In the presence of crossover interference, the Markov property does not hold, but the correlation

structure remains relatively simple.

We approach the model selection problem by splitting it into four distinct parts: choice of a class of models, model fit, model search, and model comparison. For the class of models, one may consider strictly additive QTL models or may include pairwise interactions or also higher-order interactions. Model fit concerns the fit of a defined model to data; with complete genotype data at the putative QTL, one may use linear regression, but if QTL are allowed to reside between markers, one needs a method for using genotype data at linked markers to accommodate the missing data at the putative QTL. Multiple methods are available, including Multiple Interval Mapping (MIM, KAO *et al.* 1999), Haley-Knott regression (HALEY and KNOTT 1992), and multiple imputation (SEN and CHURCHILL 2001). For any class of models, the number of possible models will be too large for them to be considered exhaustively, and so we require a procedure for searching the space of models to identify good ones, recognizing that we will only be able to consider small slices through the model space. Possible procedures include forward selection, backward elimination, stepwise selection and randomized algorithms including simulated annealing and Markov chain Monte Carlo (MCMC).

Model comparison is arguably the most important aspect of the model selection problem. For models of the same size, one would choose that giving the best fit to the data (i.e., the maximum likelihood). However, the fit can always be improved by adding an additional QTL to a model. Thus, in considering models of varying size, we must seek some balance between the fit of a model (represented by the likelihood) and the complexity of the model (the number of QTL and interactions). A common approach is to use a penalized likelihood criterion, with a penalty on the number of terms in the model. Many classical criteria have this form,

including the Akaike Information Criterion (AIC, AKAIKE 1969) and the Bayesian Information Criterion (BIC, SCHWARTZ 1978).

In their original forms, neither AIC nor BIC is ideal for model selection in QTL mapping, due to the large numbers of potential covariates. However, numerous modifications to BIC have been proposed. BALL (2001), BOGDAN *et al.* (2004) and BAIERL *et al.* (2006) suggested incorporating a prior on the model size, while SIEGMUND (2004) arrived at a similar criterion by formulating QTL mapping as a change-point problem to account for the correlation between linked markers. BROMAN and SPEED (2002) modified the penalty term in order to control the false positive rate.

Another important stream of methods for multiple-QTL mapping makes use of a Bayesian analysis via MCMC (e.g., SATAGOPAN *et al.* 1996; YI *et al.* 2003, 2005). An advantage of the Bayesian methods is that an expression of uncertainty in the final inference is integral to the framework. However, the Bayesian framework requires a specification of prior distributions on all aspects of the model (including the number of QTL and the number of interactions), implementing the MCMC algorithm requires great care, and the interpretation of the MCMC results can be difficult.

In this paper, we extend the work of BROMAN and SPEED (2002) to allow for pairwise interactions among QTL. BROMAN and SPEED (2002) had considered strictly additive QTL models, and used the null distribution of the genome-wide maximum LOD score to derive a penalty on the number of QTL that provided appropriate control on the rate of inclusion of extraneous (i.e., false positive) loci. We apply the same logic, considering the results of a two-dimensional, two-QTL scan, to derive a penalty for the interaction terms. We focus on a

class of models that includes pairwise interactions, but with an imposed hierarchical structure in which the inclusion of an interaction term requires the inclusion of the corresponding main effects.

Our goal is to develop a model selection procedure that identifies as many true QTL as possible, while controlling the rate of inclusion of extraneous loci. We generally view the inclusion of an extraneous interaction (or the failure to identify an interaction) as less severe than the inclusion of an extraneous locus (or the failure to identify a locus). This target is most appropriate for biomedical research, and may not be appropriate for agricultural or evolutionary studies.

We illustrate our methods through application to two real data sets, and validate the performance of these methods through computer simulation under several realistic models.

## METHODS

We consider the case of a backcross or intercross derived from two inbred lines and of a continuously varying quantitative trait with normally distributed residual variation. We consider QTL models with possible pairwise interactions among QTL, and impose a hierarchy on the models, with the inclusion of a pairwise interaction requiring the inclusion of both corresponding main effects. In the case of an intercross, we always include both the additive and dominance terms for a QTL, and a pairwise interaction requires the inclusion of all four parameters.

We extend the $\text{BIC}_\delta$ criterion of BROMAN and SPEED (2002) to allow pairwise interactions among QTL, defining two new criteria, $\text{pLOD}_H$ and $\text{pLOD}_L$, with heavy and light penalties on the pairwise interactions. We rename the $\text{BIC}_\delta$ criterion as $\text{pLOD}_a$, for additive QTL models.

We begin by reviewing the $\text{BIC}_\delta$ criterion, focusing particularly on the logic used to choose the penalty. We then turn to our extensions to allow for pairwise interactions. Finally, we describe model search strategies.

**The $\text{pLOD}_a$ criterion:** Consider strictly additive QTL models, of the form

$$y = \mu + \textstyle\sum_i \beta_i q_i + \epsilon,$$

where $y$ is the quantitative phenotype and $q_i$ are the genotypes at a set of QTL (coded as 0 or 1 in a backcross). The residuals, $\epsilon$, are assumed to be independent and to follow a normal distribution with constant variance.

In the original Bayesian Information Criterion (BIC, SCHWARTZ 1978), one chooses the

model that minimizes

$$\text{BIC}(\gamma) = -2l(\gamma) + \log n \cdot |\gamma|,$$

where $\gamma$ denotes a model, $l$ is the log likelihood, $n$ is the sample size, and $|\gamma|$ denotes the number of terms in the model. While BIC is often viewed as conservative, it can give a high false positive rate in the presence of numerous potential covariates (as in QTL mapping).

BROMAN and SPEED (2002) proposed to include a larger penalty, $\delta$, to control the rate of inclusion of extraneous loci:

$$\text{BIC}_\delta(\gamma) = -2l(\gamma) + \delta \cdot \log n \cdot |\gamma|.$$

For $\delta > 1$, minimizing $\text{BIC}_\delta$ will lead to smaller models and so a lower false positive rate. While BROMAN and SPEED (2002) had restricted the model space so that putative QTL were located precisely at marker loci and assumed complete genotype data at the markers (thus allowing model fit to be accomplished by linear regression), the approach can be immediately extended to allow QTL to reside between markers by the use of MIM (KAO *et al.* 1999), multiple imputation (SEN and CHURCHILL 2001), or similar approaches, to calculate the log likelihood for a model.

We prefer to rewrite the $\text{BIC}_\delta$ criterion as follows.

$$\text{pLOD}_a(\gamma) = \text{LOD}(\gamma) - T|\gamma| \tag{1}$$

where $\text{LOD}(\gamma)$ is the $\log_{10}$ likelihood ratio for the model $\gamma$ relative to the null model (of no QTL), and $T = (\delta/2)\log_{10} n$ is the penalty. We seek the model for which $\text{pLOD}_a$ achieves its

maximum.

The choice of the penalty ($\delta$ or $T$) is critical. Ideally, we seek a penalty so that, no matter the true model, the false positive rate will be maintained at a chosen level. In particular, suppose that the true model is the null model (no QTL), and suppose that we restrict our search to models with no more than one QTL. For the null model, we have $\text{pLOD}_a(\emptyset) = 0$. For a single-QTL model (with the QTL at position $\lambda$), $\text{pLOD}_a(\{\lambda\}) = \text{LOD}(\lambda) - T$, where $\text{LOD}(\lambda)$ is the LOD score from a single-QTL genome scan (e.g., by interval mapping). In this context, one would falsely choose a single-QTL model over the null model when $\text{pLOD}_a(\{\lambda\}) > \text{pLOD}_a(\emptyset)$, which is precisely the case that $\text{LOD}(\lambda) > T$.

This suggests using as the penalty, $T$, the $1 - \alpha$ quantile of the genome-wide maximum LOD score under the null hypothesis of no QTL, which could be derived, for example, from a permutation test (CHURCHILL and DOERGE 1994). With this choice, the false positive rate will be maintained at the rate $\alpha$ in the case of no QTL, and with the search restricted to models with no more than one QTL. BROMAN and SPEED (2002) showed, through computer simulation, that the false positive rate is also maintained at or near $\alpha$ in the presence of QTL, and for wider searches.

The $\text{BIC}_\delta$ criterion, which we will now call $\text{pLOD}_a$, thus extends the usual interval mapping approach to larger (though strictly additive) QTL models. The key idea is to assume a penalized likelihood criterion, as in equation (1), and to choose the penalty so that the false positive rate is guaranteed to be maintained in some simple case.

Note that in BIC, the penalty increases with the sample size, on the order $\log n$, but in the $\text{BIC}_\delta$ criterion, with the penalty $\delta$ chosen as described, the penalty is approximately constant

with sample size. (The null distribution of the genome-wide maximum LOD score is approximately constant, once a moderate size sample is reached.) Thus $\text{BIC}_\delta$ might be better termed $\text{AIC}_\delta$; we will use the term $\text{pLOD}_a$ throughout the remainder of this paper.

Also note that, while one usually counts model parameters in these penalized likelihood criteria, in equation (1) we are counting QTL.

**Model comparison with epistasis:** We now turn to the case of pairwise interactions among QTL. We consider models of the form

$$y = \mu + \sum_i \beta_i q_i + \sum_{ij} \beta_{ij} q_i q_j + \epsilon.$$

We enforce a hierarchy on the models, such that if an interaction term is included, both of the corresponding main effects must also be included.

We will consider a penalized LOD score of the form

$$\text{pLOD}(\gamma) = \text{LOD}(\gamma) - T_m|\gamma|_m - T_i|\gamma|_i \tag{2}$$

where $|\gamma|_m$ and $|\gamma|_i$ are the number of QTL and the number of pairwise interactions, respectively, in the model $\gamma$. We thus allow separate penalties on the main effects and interactions, and note again that our penalties are on the number of QTL and interactions, and not on the individual degrees of freedom in such terms.

For consistency, we will use the same penalty on the main effects as in Equation (1): the significance threshold from a single-QTL genome scan. Thus, if one restricts the search to additive QTL models, this more general criterion reduces to the $\text{pLOD}_a$ criterion.

To derive the penalty for the interaction terms, we will follow the same logic as in BROMAN and SPEED (2002), but here we will consider the results of a two-dimensional, two-QTL genome scan (HALEY and KNOTT 1992; SEN and CHURCHILL 2001). This is the obvious extension of the single-QTL genome scan of interval mapping: we consider each pair of genomic positions $(\lambda_1, \lambda_2)$ as putative QTL, and fit both a full model, with the two QTL allowed to interact, and an additive model. We also consider the results of a single-QTL genome scan. We consider three sets of LOD scores. $\text{LOD}_f(\lambda_1, \lambda_2)$ is the $\log_{10}$ likelihood ratio comparing the hypothesis of two interacting QTL, with one at position $\lambda_1$ and the other at position $\lambda_2$, to the null hypothesis (no QTL). $\text{LOD}_a(\lambda_1, \lambda_2)$ is the analogous LOD score comparing the hypothesis of two additive QTL to the null hypothesis. $\text{LOD}_1(\lambda)$ is the LOD score from the single-QTL scan, at position $\lambda$.

In considering the appropriate penalty for interaction terms, imagine that there are two additive QTL, and that one performs the two-dimensional, two-QTL scan outlined above. We might then choose the penalty as follows.

$$T_i^H = (1-\alpha) \text{ quantile of } \left\{ \max_{\lambda_1,\lambda_2} \text{LOD}_f(\lambda_1, \lambda_2) - \max_{\lambda_1,\lambda_2} \text{LOD}_a(\lambda_1, \lambda_2) \right\} \tag{3}$$

With this choice, the penalized LOD score in equation (2) has the property that, if the truth is a pair of additive QTL, and one restricts the search to models with no more than two QTL, we will falsely choose an interacting pair over the additive model at the target rate, $\alpha$. It remains to be shown that the false positive rate will be maintained for larger models and for a more extensive model search. We call this the heavy penalty (to distinguish it from a lighter penalty to be described below).

The quantile in equation (3) is ideally for the distribution in the presence of two additive QTL. Such a quantile would be difficult to derive, and would likely depend on the locations and effects of the two QTL. However, it is likely to be well approximated by the quantile under the null hypothesis of no QTL, as the interaction terms are approximately orthogonal to the main effect terms. Thus, we again recommend the use of a permutation test (CHURCHILL and DOERGE 1994) to derive this penalty. As at each permutation replicate, one must perform a two-dimensional scan of the genome, the computational effort can be quite large, but it is feasible, particularly if multiple processors are used in parallel.

It is useful, at this point, to consider graphical representations of QTL models. For the class of models under consideration, allowing pairwise interactions but with our enforced hierarchy, one may represent a QTL model as an undirected graph, with nodes representing QTL and edges representing pairwise interactions. Four example models are displayed in Figure 1. In Figure 1A, there is a pair of interacting QTL. In Figure 1B, there are three QTL, with two of the three possible pairwise interactions. In Figure 1C, there are three QTL with all possible pairwise interactions. In Figure 1D, there are seven QTL, one of which shows an interaction with each of the other QTL.

The heavy penalty in equation (3) seeks to control the rate of inclusion of an extraneous interaction (an extraneous edge in the graph). With this approach, we will have low power to detect interacting loci with limited marginal effects. Moreover, the inclusion of a false interaction (or the exclusion of a true interaction) is generally not so bad as the inclusion of a false locus (or the exclusion of a true locus). Our primary goal should be to identify the major players. The correct identification of the detailed interactions can be useful (particularly for

guiding subsequent fine-mapping experiments), but is not so critical as the correct identification of the loci themselves.

With this in mind, we consider a variation on our logic for deriving the penalty on interactions. Consider the case that the truth is a single QTL, and that we perform a two-dimensional, two-QTL genome scan, and seek to control the rate of inclusion of an additional interacting locus. With the graphs in Figure 1 in mind, the idea is that the truth is a single node, and we wish to control the rate of inclusion of an extraneous "pin" (that is, an additional node with a pairwise interaction edge). An additional interacting QTL would give an additional main effect penalty ($T_m$) plus an interaction penalty ($T_i$). Thus, one could choose the interaction penalty according to the following equation.

$$T_m + T_i^L = (1-\alpha) \text{ quantile of } \left\{ \max_{\lambda_1,\lambda_2} \text{LOD}_f(\lambda_1,\lambda_2) - \max_{\lambda} \text{LOD}_1(\lambda) \right\} \tag{4}$$

This gives our light interaction penalty: the difference between the quantile specified on the right-hand side of equation (4) and the main effect penalty, $T_m$. The penalized LOD in equation (2), with the light interaction penalty, $T_i^L$, has the property that, if the truth is a single QTL and one performs a search over models with no more than two QTL, the rate of inclusion of a second (false) QTL is maintained at rate no more than $\alpha$. Again, the quantile in equation (4) is ideally for the distribution in the presence of a single QTL, but this is not likely to be too different from the distribution under the null hypothesis of no QTL, and so we recommend estimating the penalty via a permutation test.

While we would hope that the false positive rates would be maintained at their target rates irrespective of the size of the model, analysis of numerous data sets using penalized LOD

scores with the light penalty, $T_i^L$, often indicated models similar to that depicted in Figure 1D, with a single locus interacting with each of many other loci. Subsequent computer simulation experiments indicated that, in the presence of multiple QTL, use of the light interaction penalty will often lead to a model with an extraneous QTL exhibiting interactions with multiple true QTL. The light interaction penalty controls the rate of inclusion of an extraneous pin (a QTL with a single interaction), and improperly penalizes the case of a multi-pronged pin (as in the central locus in Figure 1D).

Therefore we use a compromise between the exclusive use of the heavy interaction penalty and the exclusive use of the light interaction penalty. We consider the graphical representation of a model (for example, imagine that the four panels in Figure 1 represent a single QTL model, with 15 QTL and 12 pairwise interactions). For each cluster of connected QTL, we apply one light interaction penalty, $T_i^L$, and give all other pairwise interactions the heavy interaction penalty, $T_i^M$. (Thus, for the model consisting of the entirety of Figure 1, the penalty would be $15T_m + 4T_i^L + 8T_i^H$.)

We thus define two penalized LOD scores. The first, more conservative criterion, denoted pLOD$_H$, is of the form in equation (2), with the exclusive use of the heavy interaction penalty, $T_i^H$, defined to control the rate of inclusion of a false interaction in the presence of two additive QTL. A second, more liberal criterion, denoted pLOD$_L$, does not fit the form of equation (2), but requires consideration of the clusters of QTL connected via pairwise interactions; each cluster of QTL is assigned a single light interaction penalty, $T_i^L$ (except, of course, if the cluster consists of a single QTL), with all additional interactions being assigned the heavy penalty $T_i^H$.

The interaction penalties in these criteria were derived using the same logic that BROMAN

and SPEED (2002) used to derive the penalty in the $BIC_\delta$ criterion (here denoted $pLOD_a$), by considering the case of a simple QTL model (either a single QTL or a pair of additive QTL), and a restricted search of model space (over models with no more than two QTL). The performance of the criteria in the presence of a larger number of QTL and for a more extensive model search will be explored through computer simulation, below.

**Search of the model space:** As mentioned earlier, we prefer to separate the definition of a model comparison criterion from the procedures to search through the space of models. In constructing a model comparison criterion, we imagine the case that one could consider all possible models exhaustively. In forming a model search procedure, the task is then to optimize the chosen model comparison criterion. While a more exhaustive search is likely better than a less exhaustive search, a carefully considered procedure might give good results with a great savings in computation time. We benefit, in the QTL mapping problem, from the simple correlation structure among the potential covariates: that loci on different chromosomes are independent, and that the correlation among loci along a chromosome is not complex. For example, BROMAN and SPEED (2002) found that, for additive QTL models, forward selection followed by backward elimination performed as well as a more exhaustive search via MCMC. In extending model space to include pairwise interactions, the number of possible models is increased tremendously, making the computational efficiency gains from forward selection approaches even more important.

For the purpose of identifying interactions, we propose searching the model space using a two-at-a-time version of forward selection. At each step of forward selection, we consider adding up to two QTL plus their pairwise interaction, and we choose the model with the

maximum penalized LOD (even if it is lower than the penalized LOD for the current model). This two-at-a-time version of forward selection provides a more exhaustive search than standard forward selection and can be particularly useful in identifying QTL linked in repulsion (that is, having effects of opposite size) and interacting QTL with weak main effects. After a predetermined number of steps forward, we proceed with backward elimination as usual. The final model chosen is the one with the highest penalized LOD among all models visited in the search space.

We use an additional model refinement step, based closely on the search algorithm described by ZENG *et al.* (1999). After each step of forward selection (and also, potentially, after each backward elimination step), we iteratively refine the location of each QTL along a chromosome, keeping positions of all other QTL fixed. In the case of multiple QTL on a chromosome, the position for a given QTL is scanned between the nearest flanking QTL (or ends of the chromosome), so that the order of the QTL is not modified. This model refinement procedure has two advantages. First, as new terms are added to the model, the positions of other model terms can be updated to find the best positions for the updated model. Second, model refinement allows us to avoid searching the full two-dimensional grid at each step of two-at-a-time forward selection by focusing on a subset of that grid. For example, when using a search grid every 2.5 cM, we may look only at the marker positions to select which terms to add to the model. We then use conditional search model refinement to localize the positions to the 2.5 cM grid. In this way, model refinement allows us to save computational effort in simulations.

**Summary:** Let us briefly summarize our new method. Calculations are performed at the

genetic markers and on an evenly spaced grid (e.g., at 2.5 cM), with Haley-Knott regression, multiple interval mapping, or multiple imputation used to deal with the missing genotype information. A permutation test with a two-dimensional, two-QTL scan is used to derive the penalties in the pLOD criteria. We use forward selection to a model with a predetermined number of parameters, followed by backward elimination to the null model. At each step of forward selection, one may add an additional additive QTL, an additional QTL interacting with one of the QTL in the current model, an interaction between two QTL in the current model, an additional pair of additive QTL, or an additional pair of interacting QTL. After each step of forward selection, we perform a model refinement step to improve the locations of the QTL in the current model. The final chosen model is that with maximum penalized LOD, among all models visited.

## APPLICATIONS

We illustrate our multiple-QTL mapping methods through application to two mouse QTL mapping experiments: one a backcross and the other an intercross.

**Backcross with blood pressure phenotypes:** We first consider the data of SUGIYAMA *et al.* (2001), on salt-induced hypertension in 250 backcross mice. A selective genotyping strategy was used (for most markers, only 92 individuals with extreme phenotypes were genotyped), and so we chose not to use Haley-Knott regression (HALEY and KNOTT 1992), which is known to produce inflated evidence of linkage in this context (FEENSTRA *et al.* 2006). Instead, we used multiple imputation (SEN and CHURCHILL 2001).

We begin by calculating appropriate significance thresholds for this data set. To account for selective genotyping, we perform stratified permutations, preserving the relationship between the missing genotype pattern and the phenotypes (MANICHAIKUL *et al.* 2007). The appropriate LOD penalties of $T_m = 2.56$, $T_i^H = 2.39$, and $T_i^L = 1.01$ were obtained as 95% quantiles of distributions from 10,000 permutations with a two-dimensional, two-QTL scan on a 2.5 cM grid with 64 imputed sets of genotypes.

Previous analysis by SUGIYAMA *et al.* (2001) suggested a six QTL model with a pair of linked loci on chromosome 1; unlinked QTL on chromosomes 4, 5, 6, and 15; and a pairwise interaction between the QTL on chromosomes 6 and 15. We use this model as a starting point and proceed to explore neighboring models, to see how well the data support models with additional or fewer QTL. In accordance with our two-at-a-time forward selection strategy, we considered neighboring models of the following types: (1) adding an interaction between an existing pair of QTL, (2) an additional main effect term, (3) a new main effect interacting with

a QTL already in the model, (4) a pair of additional main effect QTL, and (5) a pair of additional QTL with a pairwise interaction.

Data analysis was performed on a 2.5 cM grid with 256 imputed sets of genotypes. The best model of each type is summarized in Figure 2. An interaction between the loci on distal chromosome 1 and chromosome 5 from the original model increased the model LOD score by 0.76, missing the light interaction penalty of 1.01. An additional locus on proximal chromosome 5 improved the LOD by 1.56, missing the main effect threshold of 2.56, and decreasing the penalized LOD score by 1.0. Fitting an interaction between this new locus on chromosome 5 and the existing locus on proximal chromosome 1 further increased the model LOD score by 1.53, exceeding the corresponding light interaction threshold of 1.01, and increasing the value of $\text{pLOD}_L$ by 0.52 compared to the model without this interaction. However, the extended model with a second locus on chromosome 5 interacting with proximal chromosome 1 still had a lower penalized LOD than the original model of SUGIYAMA *et al.* (2001) because the improvement due to interaction did not outweigh the decrease from addition of the main effect on chromosome 5. In considering whether to add a pair of loci to the model, the best pair we identified consisted of a new locus on chromosome 2, together with the aforementioned additional locus on chromosome 5. However, adding this pair of loci did not yield the required LOD increase of twice the main effect penalty, or $2 \times 2.56$. The best pair of interacting loci contained a new QTL on chromosome 4, and the same locus on proximal chromosome 5. The increase in LOD from the addition of the new term on chromosome 4 and its interaction with proximal chromosome 5 was 2.61, missing the corresponding penalty of $2.56 + 1.01 = 3.57$. In summary, our search through neighboring

models did not identify any additional model terms which improved the penalized LOD beyond that of SUGIYAMA *et al.* (2001).

We also performed backward elimination to explore the possibility of a smaller model. We found that removing the loci on proximal chromosome 1 and chromosome 5 improved the penalized LOD score from 7.40 to 9.38. So, our $\text{pLOD}_L$ criterion would support removing these two QTL from the model, resulting in a model with QTL on chromosomes 1, 4, 6 and 15, and including the $6\times15$ interaction. This more parsimonious model was also chosen by application our fully automated search algorithm. It reflects the strict control over false positives in our approach to QTL model selection. But note that we are not limited to choosing a single model; our proposed penalized LOD scores can be used as an exploratory tool to compare relative support of the data for different models.

**Intercross with measurements on TgB:** To demonstrate use of our automated model selection approach, we consider the data of SMITH RICHARDS *et al.* (2002), on self-selected diet intake. The data set available to us consisted of 502 intercross mice generated from C57BL/6J and CAST/EiJ parental strains. A selective genotyping strategy was used, based on extreme values of total kilocaloric intake during the study. We consider the TgB phenotype (total food intake weight, in grams, adjusted for post-diet body weight), and to simplify our analysis, we focus on the 165 individuals with complete genotype data. The TgB phenotype was not strongly associated with the phenotype on which the selective genotyping was based (Pearson correlation of -0.09).

To obtain thresholds for model selection, we first performed a permutation using Haley-Knott regression (HALEY and KNOTT 1992) in a two-dimensional, two-QTL genome

scan on a 2.5 cM grid. It was particularly important to obtain thresholds specific for this data set since the phenotype distribution was right skewed. For exploratory purposes, we decided to use 92.5% quantiles in our analysis. The appropriate penalties estimated from 10,000 permutations were: $T_m = 3.22$, $T_i^H = 4.12$, and $T_i^L = 2.74$. We proceeded to analyze the data using $\text{pLOD}_L$ as the model selection criterion, searching the model space using the automated approach described above: two-at-a-time forward selection to 50 model degrees of freedom, followed by backward elimination, with models fit by Haley-Knott regression. The model with the highest penalized LOD score, among all models visited, was selected as the final model.

Our automated search through the model space is summarized in Table 1. In the first step of forward selection, a single QTL on chromosome 8 at 16 cM was added to the model, and the resulting penalized LOD score was 1.80. In the second step of forward selection, a pair of interacting QTL on chromosomes 7 and 10 were added to the model, decreasing the penalized LOD score to 1.77. In the third step, two additional QTL were added on chromosome 17 and 18, with no interactions added. The position of the locus on chromosome 10 also shifted from 8.5 cM to 6 cM. The shift in estimated position demonstrates that addition of terms to the model can affect localization of QTL on other chromosomes. The resulting penalized LOD score was 2.01, suggesting this five QTL model was a better fit to the data than the single QTL model we saw in the first step of forward selection. In the fourth step of forward selection, no new QTL were added to the model, but an interaction was added between the QTL on chromosomes 8 and 18, slightly altering the position of best fit for the QTL on chromosome 8 from 16 cM to 14 cM. The resulting penalized LOD score of 1.98 was a bit lower than in the previous step of forward selection. In the fifth step of forward selection, interacting QTL were

added on chromosomes 5 and 16, improving the penalized LOD to its maximum observed value of 2.34. In the sixth step, an additional QTL was added on chromosome 19, shifting the position of the QTL on chromosome 16 from 48.2 cM to 45.7 cM, and the penalized LOD score decreased to a low value of 1.03. After ten steps of two-at-a-time forward selection, the maximum QTL model degrees of freedom was reached. The search of the model space continued with backward elimination to the null model. Through the course of forward selection and backward elimination, the model with the highest penalized LOD score was that from step 5 of forward selection, so this seven QTL model with three interactions was chosen as the final model by $\text{pLOD}_L$. The QTL model positions for the final model are summarized in Table 1.

The search through model space for this particular data set nearly covers the full set of possible moves in two-at-a-time forward selection. To the null model, we can add a single QTL, two QTL, or two QTL plus an interaction. Once forward selection has gotten started, we can also consider adding an interaction between two existing QTL, or a new QTL plus an interaction with an already present QTL. This flexibility of two-at-a-time forward selection is useful in looking for interactions, especially when the interacting QTL have weak main effects.

Our analysis of the TgB phenotype with this intercross data also highlights the importance of incorporating model refinement during the model search. As we added more terms to the model during forward selection, the newly added model terms affected our inference about the positions of already existing QTL model terms. Updating QTL positions at each step of forward selection, as described by ZENG *et al.* (1999), provides flexibility to shift QTL to their optimal positions at each step of forward selection. For example, the penalized LOD score at

the third step of forward selection was only 1.75 before refinement, suggesting this model was not as good of a fit as the model seen at the second step, which had a penalized LOD of 1.77. After model refinement, the penalized LOD increased to 2.01, with the suggestion that the refined model with the additional two QTL was a better fit than the model identified by the coarser search before refinement.

For comparison with $\text{pLOD}_L$, we also performed automated model selection using the $\text{pLOD}_H$ and $\text{pLOD}_a$ criteria, both with two-at-a-time forward selection to 50 model degrees of freedom. For each of these model selection criteria, the selected final model consisted of a single QTL on chromosome 8. These results demonstrate that the stringent $\text{pLOD}_H$ criterion can give very different results from its liberal counterpart, $\text{pLOD}_L$. With $\text{pLOD}_L$, we focus on controlling the inclusion of extraneous loci and are not concerned by extraneous interactions, while with $\text{pLOD}_H$, we also seek to control the inclusion of extraneous interaction. This more strict control with $\text{pLOD}_H$ is accompanied by reduced power.

## SIMULATIONS

While the application of a QTL mapping method to specific data sets can be informative, a more complete exploration of the performance of a method requires computer simulation experiments. We performed simulations, under various QTL model scenarios and using different model selection strategies, to characterize the performance of our proposed penalized LOD criteria, and to compare them to other approaches.

**QTL models:** We considered a variety of QTL models. First, to investigate pure error rates, we performed 2000 simulation replicates under the null model (no QTL) with backcrosses and intercrosses of 250 individuals each, using a genetic map based on the mouse genome with 19 autosomes and markers about every 10 cM, as specified by the `map10` data included in the R/qtl package (BROMAN *et al.* 2003).

To assess performance in the absence of epistatic interactions, we performed 2000 simulation replicates using the seven-QTL model previously studied in BROMAN and SPEED (2002). Here, we simulated backcrosses of 250 individuals with nine chromosomes each of length 100 cM and with 11 evenly spaced markers. Two QTL were placed on each of chromosomes 1 and 2 at positions 30 and 60 cM, with QTL on chromosome 1 linked in coupling and QTL on chromosome 2 linked in repulsion. One QTL was placed on each of chromosomes 3, 4 and 5 at positions 50, 30 and 0 cM respectively. The magnitude of all QTL effects was 0.76.

In addition to the models above, we performed simulations involving epistatic interactions using models derived from the two data sets considered in the APPLICATIONS section, above. For these simulations, we continued using `map10` with autosomes only.

Backcrosses of 250 individuals were generated in each of 4000 simulation replicates, with a model based on that presented for the hypertension data in SUGIYAMA *et al.* (2001). The model included two QTL on chromosome 1; a single QTL on each of chromosomes 4, 5, 6, and 15; and an interaction between the loci on chromosomes 6 and 15. Using the QTL positions given by SUGIYAMA *et al.* (2001), we estimated QTL effects for the simulation study using 64 sets of imputed genotypes for the hypertension data set available in R/qtl (BROMAN *et al.* 2003). The resulting heritability values for each of the model terms are displayed in Figure 4B.

Intercrosses of 500 individuals were generated in each of 4000 simulation replicates using the final model we found by $\text{pLOD}_L$ for the TgB data of SMITH RICHARDS *et al.* (2002). The model included QTL on chromosomes 5, 7, 8, 10, 16, 17, 18, and 19, and interactions between QTL on the chromosome pairs $7\times10$, $8\times18$, and $5\times16$. QTL effect parameters (estimated by multiple imputation with 64 sets of imputed genotypes) were reduced by a factor of 0.6 to more clearly differentiate the relative performance of the different methods. The resulting heritability values for each of the model terms are displayed in Figure 6B. As seen in the figure, the heritability corresponding to main effect terms of QTL on chromosomes 5 and 16 is relatively weak, so we would expect our power to detect these QTL without allowing for epistasis to be markedly reduced.

**Model selection strategies:** For each of the simulated data sets described above, we analyzed data using each of the following methods to choose a final model. These analyses allow us to compare method performance in a range of QTL model scenarios.

*$pLOD_L$*: We used $\text{pLOD}_L$ as the criterion to choose a final model. Search of the model space was performed using two-at-a-time forward selection, followed by backward

elimination. For the 19 chromosome simulations, the search was performed on a 2.5 cM grid, and models were fitted using Haley-Knott regression (HALEY and KNOTT 1992). For the 9 chromosome simulations, the search was performed on markers only for consistency with the work of BROMAN and SPEED (2002). Penalized LOD scores were calculated using 95% thresholds of the appropriate $\log_{10}$ likelihood ratio statistics, obtained according to the map and search space used for the particular simulation (see Table 2). The maximum number of model parameters (excluding the grand mean and error variance) was limited to 50. Hence, forward selection was stopped upon reaching the first model with at least 50 degrees of freedom. Further, we required at least two markers within the interval between any two linked QTL fitted at non-marker positions, and at least one marker between linked QTL where one QTL was at a marker and the other was not. This restriction is justified because essentially all information in a QTL position in between markers can be summarized by genotype data at the flanking markers (WHITTAKER *et al.* 1996), and so we should not fit linked QTL at non-marker positions with flanking markers in common.

*pLOD*$_H$: Same as for pLOD$_L$ above, but using heavy interaction penalties only.

*pLOD*$_a$ (BROMAN and SPEED 2002): Simulations performed using the pLOD$_a$ criterion were the same as those described for the pLOD$_L$, except that interactions were excluded from the search space.

*Bayesian model selection* (YI *et al.* 2005): Bayesian model fitting was performed, with priors specified on number of QTL, QTL locations, genetic effects, overall phenotypic mean, and residual variance as described in YI *et al.* (2005). The prior on the number of QTL in the model was chosen as a Poisson distribution with mean 6, and the upper bound on the number

of QTL allowed in the model was 13.

Inclusion of QTL main effects was assumed independent of epistatic effects, and QTL locations were assumed independent and uniformly distributed. Genetic effects of included main and epistatic terms were assumed to be normally distributed. The prior on the overall phenotypic mean was specified as normally distributed with mean and variance taken as their observed sample values. Finally, residual variance was assigned the improper non-informative prior, $p(\sigma^2) \propto 1/\sigma^2$.

For the 19 chromosome map, pseudomarkers were spaced approximately every 2.5 cM. The MCMC sampler was run for 404,000 steps, with 4,000 burn-in steps removed. The posterior distribution was estimated based on the results from every twentieth iteration, and so contained 20,000 MCMC samples.

Model selection was then performed by choosing the pattern of chromosomes and interactions with the highest posterior probability. Estimated positions of the QTL were obtained as the median positions from among all visited models whose pattern either matched the selected pattern exactly, or was a superset of the selected pattern.

*Modified BIC with empirical Bayes (*BAIERL et al. *2006)*: For the intercross simulations based on the data of SMITH RICHARDS *et al.* (2002), we also considered the modified BIC criterion proposed by BAIERL *et al.* (2006). To control the problem of model overfitting encountered in allowing for epistasis, mBIC (BOGDAN *et al.* 2004) incorporates a prior distribution to account for the large number of model parameters considered. Extending this work to deal with both backcrosses and intercrosses, BAIERL *et al.* (2006) proposed QTL

model selection by minimizing the following general criterion:

$$\text{mBIC} = n \log \text{RSS} + (p+q) \log n + 2p \log(l-1) + 2q \log(u-1). \tag{5}$$

Here $p$ is the number of main effect parameters and $q$ is the number of epistatic parameters in the model. The mBIC criterion is very similar to the original BIC, with the addition of two separate penalty terms on main effects and interactions to account for the large number of possible predictors of these two types. Specifically, $l := m_v/\text{EN}_v$ and $u := m_e/\text{EN}_e$, where $m_v$ and $m_e$ are the number of main and epistatic parameters in the model, while $\text{EN}_v$ and $\text{EN}_e$ are the expected number of main and epistatic parameters, according to the chosen prior. Performing model selection on the $m$ marker positions only, we take $m_v = 2m$ and $m_e = 2m(m-1)$ and $\text{EN}_v = \text{EN}_e = 2.2$, which gives mBIC obtained as an direct extension of that presented in BOGDAN *et al.* (2004).

As an alternative to fixed priors, BAIERL *et al.* (2006) presented an empirical Bayes type strategy for setting the expected number of true QTL model terms. This practical approach begins with an initial scan using mBIC as described above. The observed numbers of main and epistatic terms, $\hat{N}_v$ and $\hat{N}_e$ are then used to set the expected number of QTL parameters as $\text{EN}_v = \max(2.2, \hat{N}_v)$ and $\text{EN}_e = \max(2.2, \hat{N}_e)$. Plugging these new expected values into the general version of mBIC in equation (5) yields an empirical Bayes version of the criterion, called $\text{mBIC}_1$. This modified version always sets the expected number of model parameters to be as large or larger than specified by mBIC, and so should be more powerful in detecting QTL effects.

*Modified BIC with additional scan for interactions*: A second extension of mBIC was

proposed in BAIERL *et al.* (2006) to improve power to detect interactions. This modification, called $\text{mBIC}_2$, is designed to perform a focused scan for interactions involving QTL terms already identified by $\text{mBIC}_1$. The idea is that $\text{mBIC}_1$ could miss real interactions because it allows interactions to be added to a model without corresponding main effects, making the model space more noisy, and requiring the search criterion to be more stringent. By performing a scan targeted at a reduced set of interactions, the extended criterion, $\text{mBIC}_2$, is designed to improve power to pick up epistatic terms.

For simulations, each of $\text{mBIC}_1$ and $\text{mBIC}_2$ are implemented to search models with QTL at marker positions. Also, note that BOGDAN *et al.* (2004) and BAIERL *et al.* (2006) do not impose the same hierarchy as we do on the model space: they allow interactions to be included in a model without corresponding main effects. They also allow additive effects in an intercross model without corresponding dominance terms and vice versa.

**Software:** Simulations were performed using the statistical software R (IHAKA and GENTLEMAN 1996), and the add-on package R/qtl (BROMAN *et al.* 2003). Some computationally intensive functions were coded in C or Matlab, and called from R to improve speed. The Bayesian analysis was performed using R/qtlbim (YANDELL *et al.* 2007). Analysis with the mBIC criteria was performed with Matlab code obtained from Andreas Baierl.

**Results:** *Backcross under the null:* In our 2000 simulations of backcross data under the null model, $\text{pLOD}_L$, $\text{pLOD}_H$ and $\text{pLOD}_a$ identified one or more false positives 6.8%, 6.1% and 6.1% of the time, respectively. Under the null hypothesis, using the lighter penalty system of $\text{pLOD}_L$ did not increase error rates much beyond those obtained by more conservative methods, $\text{pLOD}_H$ and $\text{pLOD}_a$.

*Intercross under the null:* In our 2000 simulations of intercross data under the null model, $\text{pLOD}_L$, $\text{pLOD}_H$ and $\text{pLOD}_a$ identified one or more false positives 5.6%, 5.3% and 5.3% of the time, respectively. Again, the error rates for $\text{pLOD}_L$ were not much higher than for $\text{pLOD}_H$ and $\text{pLOD}_a$.

*Backcross with seven QTL additive model:* In each of our 2000 simulation replicates, a QTL was considered to be identified successfully by a particular method if the final model chosen by that method contained a QTL within 10 cM of the true QTL position. The mean numbers of QTL identified correctly by $\text{pLOD}_L$, $\text{pLOD}_H$ and $\text{pLOD}_a$ were 6.23, 6.24 and 6.26, respectively, with one or more extraneous unlinked QTL in 9.6%, 3.6% and 3.2% of simulation replicates. For each of the three penalized LOD approaches, there was higher power to detect QTL linked in coupling compared to QTL linked in repulsion. Specifically, the average numbers of QTL identified on chromosome 1 (with two QTL linked in coupling) in each of the simulation replicates by $\text{pLOD}_L$, $\text{pLOD}_H$ and $\text{pLOD}_a$ were 1.80, 1.80 and 1.81, compared to 1.57, 1.58 and 1.57 on chromosome 2 (with two QTL linked in repulsion).

*Backcross based on hypertension data:* Our 4000 simulation replicates based on hypertension data from SUGIYAMA *et al.* (2001) allowed us to compare the performance of penalized LOD scores $\text{pLOD}_L$, $\text{pLOD}_H$ and the original $\text{pLOD}_a$ (BROMAN and SPEED 2002) with the Bayesian approach (YI *et al.* 2005) in the presence of epistatic interaction. QTL main effects were considered to be identified correctly if the final selected model contained a QTL within 50 cM of the true QTL position. Likewise, the interaction between chromosomes 6 and 15 was considered to be identified correctly if the final model contained a pairwise interaction between chromosomes 6 and 15, with each of these terms localized to within 50 cM of their

correct positions. The large window for picking up true QTL main effects gave us a generous view of power, but we complemented this measure with summaries of precision in identified QTL positions. Additional QTL identified on the same chromosome as correctly identified true QTL, as well as QTL identified on chromosomes with neither true main effects nor interaction terms, were considered to be extraneous QTL. Additional interactions beyond the first correctly identified interaction were counted as extraneous interactions.

While the overall error rate for pLOD$_H$ (6.8%) closely matched that of pLOD$_a$ (6.3%), the rate of including one or more extraneous terms was about doubled for pLOD$_L$ (12.9%). The overall error rate for the Bayesian approach (19.0%) was higher than for the penalized LOD scores. Similarly, the rate of including two or more extraneous model terms was notably higher for pLOD$_L$ and the Bayesian approach compared to the more conservative pLOD$_H$ and pLOD$_a$, but all of these methods had a very small probability of including three or more extraneous terms (Figure 3A).

The pLOD$_L$ and Bayesian approaches had high rates of extraneous interactions (Figure 3B); they selected models with one or more false interaction terms in 12.7% and 8.3% of simulation replicates, respectively, compared to 1% for pLOD$_H$, and 0% for pLOD$_a$ (since this method does not allow interactions).

Among extraneous interactions reported in the final models, a portion of these carried in false main effect terms (Figure 3C). For the Bayesian approach, a majority of extraneous interactions also introduced extraneous main effects to the model. In 4.9% and 7.2% of simulation replicates, respectively, pLOD$_L$ and the Bayesian approach identified models with one or more false interaction terms involving extraneous main effects. Indeed, the pLOD$_L$

criterion was designed to allow interactions with a false main effect about 5% of the time, so the criterion performed appropriately in this regard.

It is also interesting to note that the rate of including extraneous model terms increased with the total number of model terms identified (Figure 3D). As the penalized LOD criteria identified more true main effects, there were more model terms with which extraneous loci could interact. This problem appeared most notably for pLOD$_L$, which allowed interactions to be added more liberally in the presence of strong main effects. In contrast, the Bayesian error rates were relatively fixed as a function of model size.

The precision with which model terms were identified closely followed the power to identify the corresponding model terms (Figure 4C). Each of the methods had similar root mean squared error (RMSE) of identified non-interacting main effects. For the interacting loci on chromosomes 6 and 15, pLOD$_L$, pLOD$_H$ and Bayesian model selection had notably lower RMSE than pLOD$_a$, which did not incorporate the interaction term.

*Intercross based on TgB data:* Simulations based on the TgB data of SMITH RICHARDS *et al.* (2002) allowed us to see the additional power gained by allowing interactions into our model search space for an intercross. Model summaries identical to those described for the hypertension simulations are shown for these intercross simulations in Figures 5 and 6, with the addition of results for the mBIC methods (BAIERL *et al.* 2006).

As with the hypertension simulations, pLOD$_L$ detected more true QTL than pLOD$_H$ and pLOD$_a$ (Figure 6D), with notably higher power to detect the loci on chromosomes 5 and 16, which had strong interaction but weak main effects. At the same time, pLOD$_L$ also had notably higher error rates than pLOD$_H$ and pLOD$_a$ (Figure 5A).

The Bayesian approach detected correct main effects approximately in the same range as $\text{pLOD}_L$ (Figure 6D). Similarly, power for individual model terms was comparable for these two methods, although $\text{pLOD}_L$ performed noticeably better in picking up the interaction terms (Figure 6A). Extraneous main effects were included in models selected by the Bayesian approach at a rate comparable to that seen for $\text{pLOD}_L$ (Figure 5A).

The detection of correct main effects for the mBIC methods was roughly in between that of $\text{pLOD}_H$ and $\text{pLOD}_a$ (Figure 6D), with $\text{mBIC}_2$ having the higher power of the two mBIC approaches. The power to detect interactions was quite varied across the pLOD and mBIC methods. The $\text{pLOD}_L$ and $\text{pLOD}_H$ criteria had the highest power to detect interactions, followed by $\text{mBIC}_2$, with $\text{mBIC}_1$ having the lowest power among methods designed to detect epistasis (Figure 6A). In terms of overall rates of including extraneous main effects, $\text{mBIC}_2$ was slightly higher than $\text{pLOD}_L$, while $\text{mBIC}_1$ was a bit higher than $\text{pLOD}_H$ (Figure 5A). Since the mBIC approaches do not constrain additive and dominance terms to be at the same marker positions, we adjusted our reported error rates by dropping one of each pair of neighboring markers from our count of extraneous main effects, but the rates displayed for mBIC may still be somewhat inflated.

Looking at the error rates stratified by the number of correctly identified main effects, we saw that $\text{pLOD}_L$ tended to add more extraneous terms when more QTL were identified correctly (Figure 5D). The same pattern was also noticeable to a lesser extent for $\text{mBIC}_2$. This phenomenon makes sense for $\text{mBIC}_2$ and $\text{pLOD}_L$ because both of these criteria were designed to allow more interaction terms into the model in the presence of already identified main effects, and these interaction terms often carry in extraneous main effects (Figure 5C). The

Bayesian approach seemed to show a slight decrease in error rate as the number of correctly identified main effects increased, but this may simply reflect the default prior of six main effect QTLs. Further, the small number of realizations with two or fewer QTL identified correctly by the Bayesian method make it difficult to characterize this relationship completely.

# DISCUSSION

QTL mapping methods that consider models with multiple QTL have a number of advantages over methods, such as interval mapping, based on single-QTL models. By controlling for QTL with large effects, one may increase power to detect QTL of more modest effect; the presence of two linked QTL is best assessed by comparing the fit of a two-QTL model to the best single-QTL model; and the assessment of epistatic interactions between QTL requires the fit of a multiple-QTL model. However, it can be difficult to establish the support for QTL in the context of multiple-QTL models.

Bayesian methods provide the most natural expression of model uncertainty: the posterior distribution on QTL models and, as a corollary, the posterior probability that an individual locus or interaction term is involved in the phenotype. Bayesian methods have the further advantages of more completely capturing uncertainty in the inference and of tying together all aspects of the problem.

However, the Bayesian approach requires a specification of prior distributions; particularly difficult is the specification of the prior on QTL models, including the number of QTL and the number of interactions. The posterior can be quite sensitive to the choice of prior. Additional difficulties include the construction of efficient MCMC samplers and the diagnosis of deficiencies in a sampler. Moreover, making sense of the MCMC results can be difficult, particularly for the novice.

We have described a penalized likelihood approach to the problem of comparing multiple QTL models; the approach allows one to weigh the support for QTL through penalties on QTL and on epistatic interactions. We present simple approaches for deriving appropriate penalties,

which make use of the results of a permutation test to control particular false positive rates in the presence of one or two QTL and with a single- or two-dimensional search. Our approach extends the work of BROMAN and SPEED (2002) to allow for epistatic interactions. We also allow QTL to reside at non-marker locations. Likelihood calculations may be performed by any of a number of fitting algorithms, including MIM (KAO *et al.* 1999), multiple imputation (SEN and CHURCHILL 2001), or Haley-Knott regression (HALEY and KNOTT 1992).

The penalties on interactions in the two criteria, $\text{pLOD}_H$ and $\text{pLOD}_L$, are readily derived by permutation (CHURCHILL and DOERGE 1994) and are often calculated as LOD thresholds in routine QTL analysis. Thus, the derivation of these penalties requires essentially no more work than is typically done in standard analyses.

The choice of a prior on the number of QTL and interactions in a Bayesian approach and in the modified BIC criterion is essentially equivalent to a choice of penalties on model complexity; modification of the prior will affect the performance characteristics of the methods. In QTL mapping for biomedical research, control of the false positive rate is particularly important. We view our system for selecting penalties, which considers such false positive rates directly, as more natural than the specification of a prior. It may be possible to derive priors for which our penalized LOD criteria correspond approximately to the log posterior for a model. This deserves further investigation, as it would enable a simple system for specifying the prior in Bayesian QTL analysis.

While we have illustrated our methods in two applications, an understanding of method performance requires large-scale computer simulations. Our two main simulations were based on models derived from the two applications, and included comparisons of our penalized LOD

criteria to a Bayesian approach and to the modified BIC criteria of BAIERL *et al.* (2006).

As expected, we observed notably different frequencies of extraneous QTL using $\text{pLOD}_L$ versus $\text{pLOD}_H$. In fact, the two extensions to $\text{pLOD}_a$ are useful toward different goals. With the interaction penalty in the $\text{pLOD}_H$ criterion, we seek to control the rate of any extraneous interaction, while with $\text{pLOD}_L$ we seek to control the inclusion of an extraneous interacting locus and are relatively permissive of extraneous interactions, providing that they do not bring in a false locus. For both simulations, the penalized LOD criteria had false positive rates exceeding their targets, and given that there are two separate penalties, it should probably not be surprising that the error rates are on the order of $2\alpha$, when the target is $\alpha$. Moreover, the error rate for the $\text{pLOD}_L$ criterion was seen to grow with the number of correctly identified QTL. This again is not unexpected, as with a larger number of true QTL, there are more ways in which an extraneous interacting locus may enter the model. (That is, with a larger model, there are more ways in which to attach an extraneous "pin".) This behavior is likely to be acceptable; one should be less concerned about a single extraneous QTL among six correctly identified QTL than about an extraneous QTL when only one QTL was correctly identified.

The performance of the Bayesian method that we considered was seen to be comparable to the $\text{pLOD}_L$ criterion. It should be noted that we considered a single possible model prior, and a single approach for choosing a model. Also, a Bayesian would generally not be satisfied with choosing a single model.

Unlike BOGDAN *et al.* (2004) and BAIERL *et al.* (2006), we have required the hierarchy that interactions are considered together with their corresponding main effects, and we have also restricted additive and dominance terms of an intercross to join any model as a pair. This

makes a comparison between our methods and the modified BIC criterion somewhat difficult. As with any choice, our restriction of model space has both advantages and disadvantages. On the one hand, we have reduced flexibility in terms of the types of models we consider, and so should expect lower power in certain situations (such as the case of pure interactions). On the other hand, the model space we consider is somewhat less noisy. Furthermore, we find results with additive and dominance terms at the same position to be easier to interpret and more biologically plausible than models with additive and dominance terms at neighboring positions, which often arise in the method of BAIERL *et al.* (2006), and so required special consideration in our summaries of the simulation results.

While we have limited the model space to pairwise interactions, it may be worthwhile to consider higher-order interactions in future studies, particularly for backcrosses. Doing so will result in a vastly increased model space, and so will present a major challenge in controlling the false positive rates. The trend, for our penalized LOD criteria, towards increasing error rates for larger models, will likely be more severe if higher-order interactions are pursued.

In their present form, our penalized LOD criteria treat all QTL main effects equally. However, reduced information due to the underlying correlation between linked loci suggests that linked QTL should be treated differently. Moreover, the errors incurred by declaring a pair of linked loci to be a single QTL, or vice versa, likely deserve special consideration. SIEGMUND (2004) approached this issue by deriving a BIC-like criterion that approximates the Bayes factor of a linked QTL model. A modification of our criteria, to allow separate penalties for additional QTL on a chromosome, deserves consideration. Permutation test results for a two-dimensional, two-QTL scan might still be used, by considering within- and

between-chromosome results separately.

Another important extension of our criteria concerns the X chromosome. In the present work, we have focused exclusively on the autosomes. BROMAN *et al.* (2006) recognized that the X chromosome requires a different LOD threshold due to the different degrees of freedom required to handle the X chromosome. A reasonable approach may be to penalize main effects on the X chromosome by the amount suggested as the X chromosome-specific significance threshold in BROMAN *et al.* (2006). However, further work needs to be done to choose appropriate penalties for epistatic interactions involving the X chromosome. In doing so, interactions involving two loci on the X chromosome would have to be treated differently from interactions between a locus on the X chromosome and a locus on an autosome.

We have also omitted consideration of covariates. A fixed set of additive covariates may be easily included within our scheme, but if covariates are to be selected from a set of possible covariates, or if QTL$\times$covariate interactions are to be considered, a modification of our criteria, with specific penalties for such terms, must be derived.

An important limitation in applying our penalized LOD criteria is that, in the case of extremely large models, unusual results may be obtained. In particular, we suggest that one avoids models with more than $n/5$ parameters (where $n$ is the sample size). In numerous simulations and applications (results not shown), we have found that extremely large models can give penalized LOD scores that suggest a vastly improved fit, especially in the case of an intercross. While this behavior means that use of pLOD$_L$ and pLOD$_H$ does not extend to the full set of models, in practice the limitation to models with no more than $n/5$ parameters is broad enough to include most relevant models. For a sample size of about 250 individuals, we

can still fit 50 model parameters, which allows for more than the number of QTL we would typically expect to be able to identify. Furthermore, this restriction is not unique to our criteria, but applies more generally (KASS and RAFTERY 1995).

We have employed a relatively simple model search algorithm. Improvements in this algorithm, to more exhaustively search model space and to reduce computation time, deserve additional investigation. Our two-at-a-time forward selection algorithm can be computationally intensive, and may have little advantage over one-step forward selection to a large model followed by backward elimination. A more adaptive search algorithm, in place of forward selection to a model of predetermined size, may further reduce computation time. In addition, randomized search methods, such as MCMC and simulated annealing, could allow a more exhaustive consideration of the space of interactions.

While our penalized LOD criteria are based on fixed penalties which are derived from fixed targets for the false positive rates, we generally disapprove of strict significance thresholds. Precisely defined criteria can be useful to the novice QTL mapper, and particularly for the assessment of the performance of a method in computer simulations, but one should keep in mind that there will generally be an array of similarly plausible models. Thus, in practice, it is important to explore the set of models whose penalized LOD is similar to the chosen model. The simplest approach may be to consider a single working model (such as that for which the penalized LOD score is optimized) and focus on the change in likelihood that accompanies the omission of a single term or the addition of one or two QTL or interactions. Such evidence can be usefully represented in a graphs, such as that in Figure 2.

In summary, we have presented a straightforward and versatile approach to QTL model

selection in the presence of pairwise interactions among QTL. The model selection criteria directly control false positive rates, and so the support for QTL identified through these procedures can be interpreted in a sense similar to traditional hypothesis testing. Since the model selection technique is closely tied to permutation testing, the criteria can be applied quite generally.

## ACKNOWLEDGMENTS

The authors thank Gary Churchill for inspiring conversations leading to this work, Andreas Baierl for sharing his Matlab code for mBIC which made the comparative simulations possible, Bev Paigen for making the hypertension data publicly available, and Brenda Smith Richards for the use of the TgB data. This work was supported in part by National Institutes of Health grants GM074244 (to K.W.B.) and GM069430 (to B.S.Y. and J.Y.M.) and a National Science Foundation graduate research fellowship (to A.M.).

## LITERATURE CITED


AKAIKE, H., 1969 Fitting autoregressive models for prediction. Ann. Inst. Statist. Math. **21**: 243–247.

BAIERL, A., M. BOGDAN, F. FROMMLET and A. FUTSCHIK, 2006 On locating multiple interacting quantitative trait loci in intercross designs. Genetics **173**: 1693–1703.

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.

BOGDAN, M., J. K. GHOSH and R. W. DOERGE, 2004 Modifying the Schwarz Bayesian Information Criterion to locate multiple interacting quantitative trait loci. Genetics **167**: 989–999.

BROMAN, K. W., S. SEN, S. E. OWENS, A. MANICHAIKUL, E. M. SOUTHARD-SMITH *et al.*, 2006 The X chromosome in quantitative trait locus mapping. Genetics **174**: 2151–2158.

BROMAN, K. W. and T. P. SPEED, 2002 A model selection approach for the identification of quantitative trait loci in experimental crosses. J. R. Statist. Soc. B **64**: 641–656.

BROMAN, K. W., H. WU, S. SEN and G. A. CHURCHILL, 2003 R/qtl: QTL mapping in experimental crosses. Bioinformatics **19**: 889–890.

CHURCHILL, G. A. and R. W. DOERGE, 1994 Empirical threshold values for quantitative trait mapping. Genetics **138**: 963–971.

FEENSTRA, B., I. M. SKOVGAARD and K. W. BROMAN, 2006 Mapping quantitative trait loci by an extension of the Haley-Knott regression method using estimating equations. Genetics **173**: 2269–2282.

HALEY, C. S. and S. A. KNOTT, 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity **69**: 315–324.

IHAKA, R. and R. GENTLEMAN, 1996 R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics **5**: 299–314.

KAO, C. H., Z. B. ZENG and R. D. TEASDALE, 1999 Multiple interval mapping for quantitative trait loci. Genetics **152**: 1203–1216.

KASS, R. E. and A. E. RAFTERY, 1995 Bayes factors. Journal of the American Statistical Association **90**: 773–795.

LANDER, E. S. and D. BOTSTEIN, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121**: 185–199.

MANICHAIKUL, A., A. A. PALMER, S. SEN and K. W. BROMAN, 2007 Significance thresholds for quantitative trait locus mapping under selective genotyping. Genetics **177**: 1963–1966.

SATAGOPAN, J. M., B. S. YANDELL, M. A. NEWTON and T. C. OSBORN, 1996 A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. Genetics **144**: 805–816.

SCHWARTZ, G., 1978 Estimating the dimension of a model. The Annals of Statistics **6**: 461–464.

SEN, S. and G. A. CHURCHILL, 2001 A statistical framework for quantitative trait mapping. Genetics **159**: 371–387.

SIEGMUND, D., 2004 Model selection in irregular problems: Applications to mapping quantitative trait loci. Biometrika **91**: 785–800.

SMITH RICHARDS, B. K., B. N. BELTON, A. C. POOLE, J. J. MANCUSO, G. A. CHURCHILL *et al.*, 2002 QTL analysis of self-selected macronutrient diet intake: fat, carbohydrate, and total kilocalories. Physiol Genomics **11**: 205–217.

SUGIYAMA, F., G. A. CHURCHILL, D. C. HIGGINS, C. JOHNS, K. P. MAKARITSIS *et al.*, 2001 Concordance of murine quantitative trait loci for salt-induced hypertension with rat and human loci. Genomics **71**: 70–77.

WHITTAKER, J., R. THOMPSON and P. VISSCHER, 1996 On the mapping of QTL by regression of phenotype on marker type. Heredity **77**: 23–32.

YANDELL, B. S., T. MEHTA, S. BANERJEE, D. SHRINER, R. VENKATARAMAN *et al.*, 2007 R/qtlbim: QTL with Bayesian Interval Mapping in experimental crosses. Bioinformatics **23**: 641–3.

YI, N., S. XU and D. B. ALLISON, 2003 Bayesian model choice and search strategies for mapping interacting quantitative trait loci. Genetics **165**: 867–883.

YI, N., B. S. YANDELL, G. A. CHURCHILL, D. B. ALLISON, E. J. EISEN *et al.*, 2005 Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. Genetics **170**: 1333–1344.

ZENG, Z.-B., C.-H. KAO and C. J. BASTEN, 1999 Estimating the genetic architecture of quantitative traits. Genetical Research **74**: 279–289.

Table 1: Model selection summaries for analysis of intercross data with the TgB phenotype. The first six steps of forward selection using $\text{pLOD}_L$ with 92.5% penalties are shown with positions of the fitted QTL (in cM), and the resulting penalized LOD score. After ten steps of forward selection followed by backward elimination, the best visited model was that seen in step five of forward selection. The best models chosen by $\text{pLOD}_H$ and $\text{pLOD}_a$ were both the same as single QTL model visited in the first step of forward selection for $\text{pLOD}_L$, shown above.

| | Main effects (Chr) | | | | | | | | Intxns (Chr:Chr) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | 5 | 7 | 8 | 10 | 16 | 17 | 18 | 19 | 7:10 | 8:18 | 5:16 | pLOD |
| FS 1 | | | 16 | | | | | | | | | 1.80 |
| FS 2 | | 63.5 | 16 | 8.5 | | | | | ✓ | | | 1.77 |
| FS 3 | | 63.5 | 16 | 6 | | 11.8 | 30 | | ✓ | | | 2.01 |
| FS 4 | | 63.5 | 14 | 6 | | 11.8 | 30 | | ✓ | ✓ | | 1.98 |
| FS 5 | 80 | 63.5 | 14 | 6 | 48.2 | 11.8 | 30 | | ✓ | ✓ | ✓ | 2.34 |
| FS 6 | 80 | 63.5 | 14 | 6 | 45.7 | 11.8 | 30 | 15 | ✓ | ✓ | ✓ | 1.03 |
| $\text{pLOD}_L$ | 80 | 63.5 | 14 | 6 | 48.2 | 11.8 | 30 | | ✓ | ✓ | ✓ | 2.34 |
| $\text{pLOD}_H$ | | | 16 | | | | | | | | | 1.80 |
| $\text{pLOD}_a$ | | | 16 | | | | | | | | | 1.80 |

Table 2: LOD penalties used for calculating penalized LOD scores in our simulation studies. For the 19 chromosome map, the penalties are based on a grid search every 2.5 cM using Haley-Knott regression. For the 9 chromosome map, the penalties are based on a search at markers only. LOD penalties are estimated by 10,000 simulation replicates under the null distribution.

| | 19 chromosome map | | 9 chromosome map |
|---|---|---|---|
| Penalty | Backcross | Intercross | Backcross |
| Main effect | 2.8040 | 3.5808 | 2.5100 |
| Light Interaction | 1.2705 | 2.8128 | 1.0212 |
| Heavy Interaction | 2.7162 | 4.4117 | 2.3741 |

FIGURE LEGENDS

**Figure 1.** Example graphical representations of QTL models. Nodes represent QTL and edges represent pairwise interactions. (A) A model containing two QTL and their interaction. (B) A model containing three QTL and two pairwise interactions. (C) A model containing three QTL and all possible pairwise interactions. (D) A model containing seven QTL, with one central locus interacting with the other six.

**Figure 2.** The final model for hypertension data in SUGIYAMA *et al.* (2001) contained a pair of QTL on chromosome 1, unlinked QTL on chromosomes 4 and 5, and a pair of interacting QTL on chromosomes 6 and 15 (solid). We explored the possibility of additional interactions between QTL in the original model of SUGIYAMA *et al.* (2001), and also considered additional QTL on chromosomes 2, 4 and 5, with possible interactions involving these loci (dashed). Each new model term is annotated with its corresponding contribution to the model LOD score. For example, an interaction between model terms 1b and 5a increased the model LOD score by 0.76 compared to the original model.

**Figure 3.** Error rates for backcross simulations under the model based on data from SUGIYAMA *et al.* (2001). Plots display observed distribution summaries of the (A) number of extraneous main effects, (B) number of extraneous interactions, (C) number of interactions associated with extraneous main effects, and (D) the probability of one or more extraneous unlinked terms as a function of the number of correct main effects, shown only when at least thirty simulation replicates had the specified number of correct main effects using a given method. 95% confidence limits in (D) are obtained by inverting an exact binomial test. Results are

shown for $\text{pLOD}_L$ (solid black circle), $\text{pLOD}_H$ (solid red square), $\text{pLOD}_a$ (solid blue triangle), and Bayes (open black diamond).

**Figure 4.** Power to detect true QTL in backcross simulations under the model based on data from SUGIYAMA *et al.* (2001). Plots display (A) observed power to detect each individual model term, (B) heritability of the QTL main effects and interactions, (C) root mean squared error of the estimated positions among correctly identified QTL, and (D) the observed distribution of the number of correct main effects identified in each simulation replicate. Panels A, C and D show results for $\text{pLOD}_L$ (solid black circle), $\text{pLOD}_H$ (solid red square), $\text{pLOD}_a$ (solid blue triangle), and Bayes (open black diamond).

**Figure 5.** Error rates for intercross simulations under the model based on data from SMITH RICHARDS *et al.* (2002). Plots display observed distribution summaries of the (A) number of extraneous main effects, (B) number of extraneous interactions, (C) number of interactions associated with extraneous main effects, and (D) the probability of one or more extraneous unlinked terms as a function of the number of correct main effects, shown only when at least thirty simulation replicates had the specified number of correct main effects using a given method. 95% confidence intervals in (D) are obtained by inverting an exact binomial test. Results are shown for $\text{pLOD}_L$ (solid black circle), $\text{pLOD}_H$ (solid red square), $\text{pLOD}_a$ (solid blue triangle), Bayes (open black diamond), $\text{mBIC}_2$ (open black circle), and $\text{mBIC}_1$ (open red square).

**Figure 6.** Power to detect true QTL in intercross simulations under the model based on data from SMITH RICHARDS *et al.* (2002). Plots display (A) observed power to detect each

individual model term, (B) heritability of the QTL main effects and interactions, (C) root mean squared error of the estimated positions among correctly identified QTL, and (D) the observed distribution of the number of correct main effects identified in each simulation replicate. Panels A, C and D show results for $\text{pLOD}_L$ (solid black circle), $\text{pLOD}_H$ (solid red square), $\text{pLOD}_a$ (solid blue triangle), Bayes (open black diamond), $\text{mBIC}_2$ (open black circle), and $\text{mBIC}_1$ (open red square).

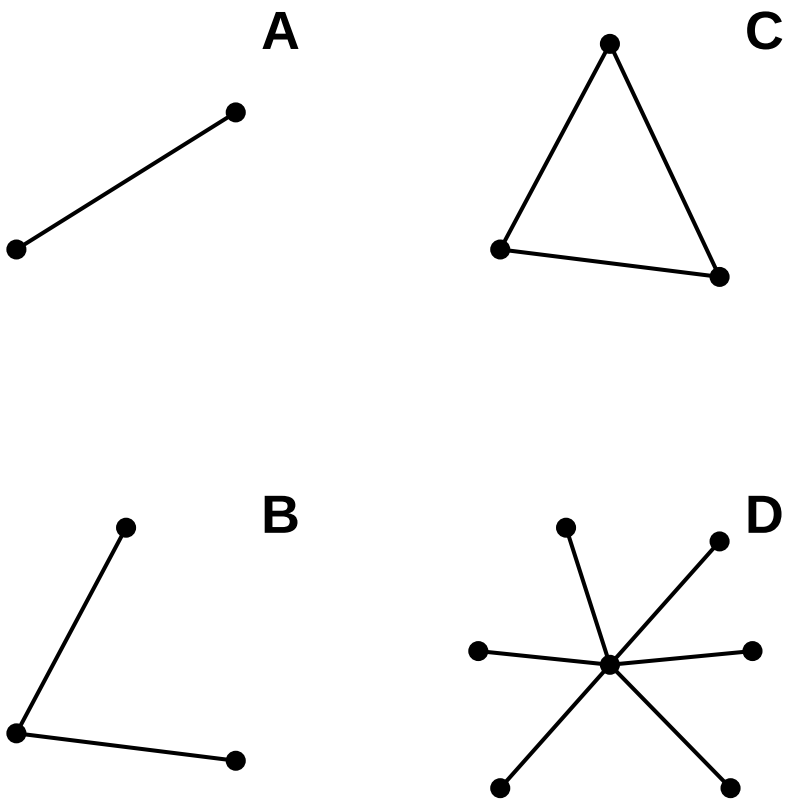


Figure 1

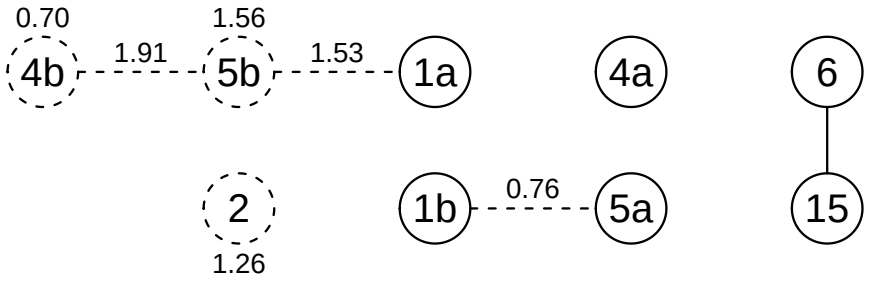


Figure 2

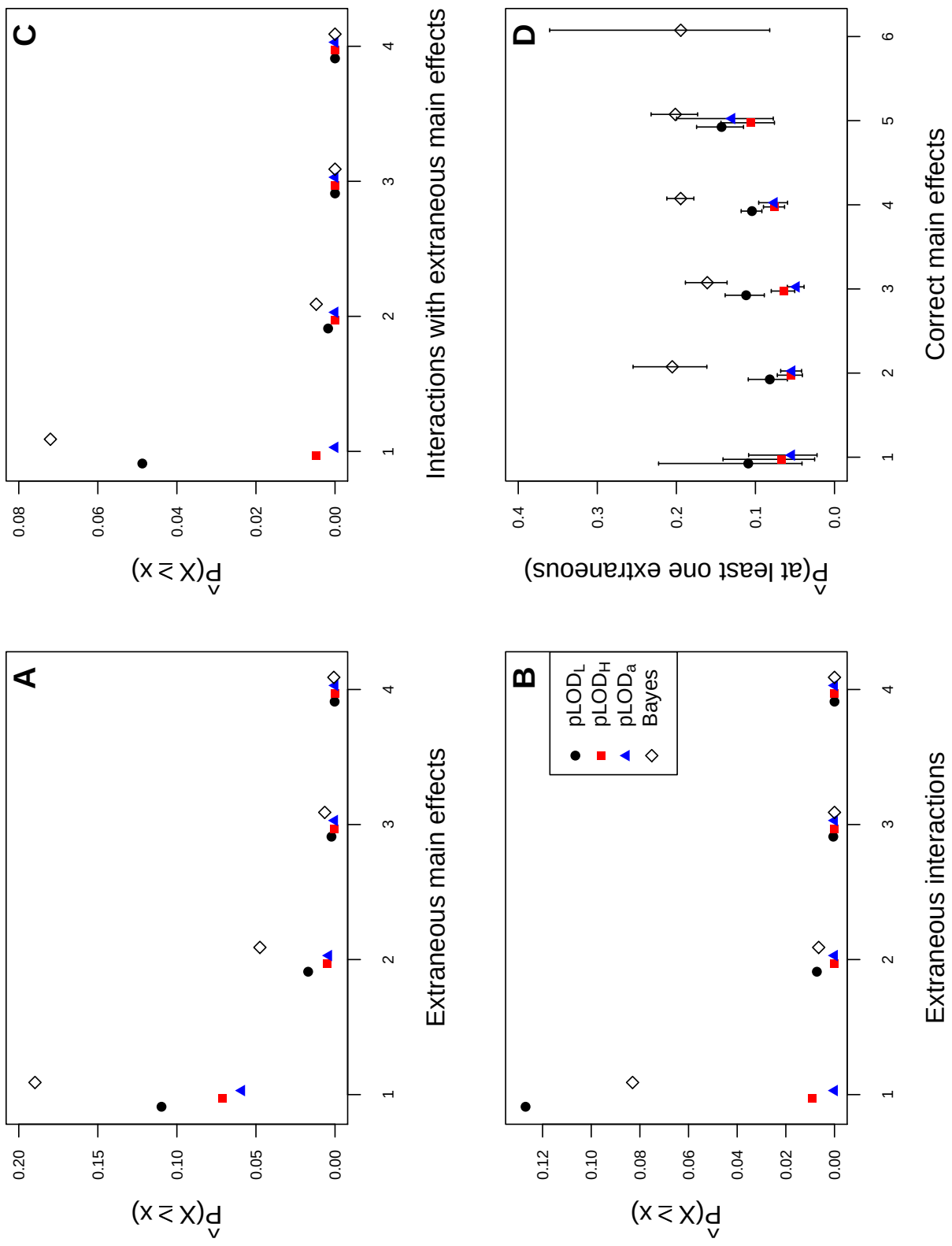
A
B
C
D
Extraneous main effects
Extraneous interactions
Interactions with extraneous main effects
Correct main effects
P>(X ≥ x)
P>(at least one extraneous)
pLOD_L
pLOD_H
pLOD_a
Bayes

Figure 3

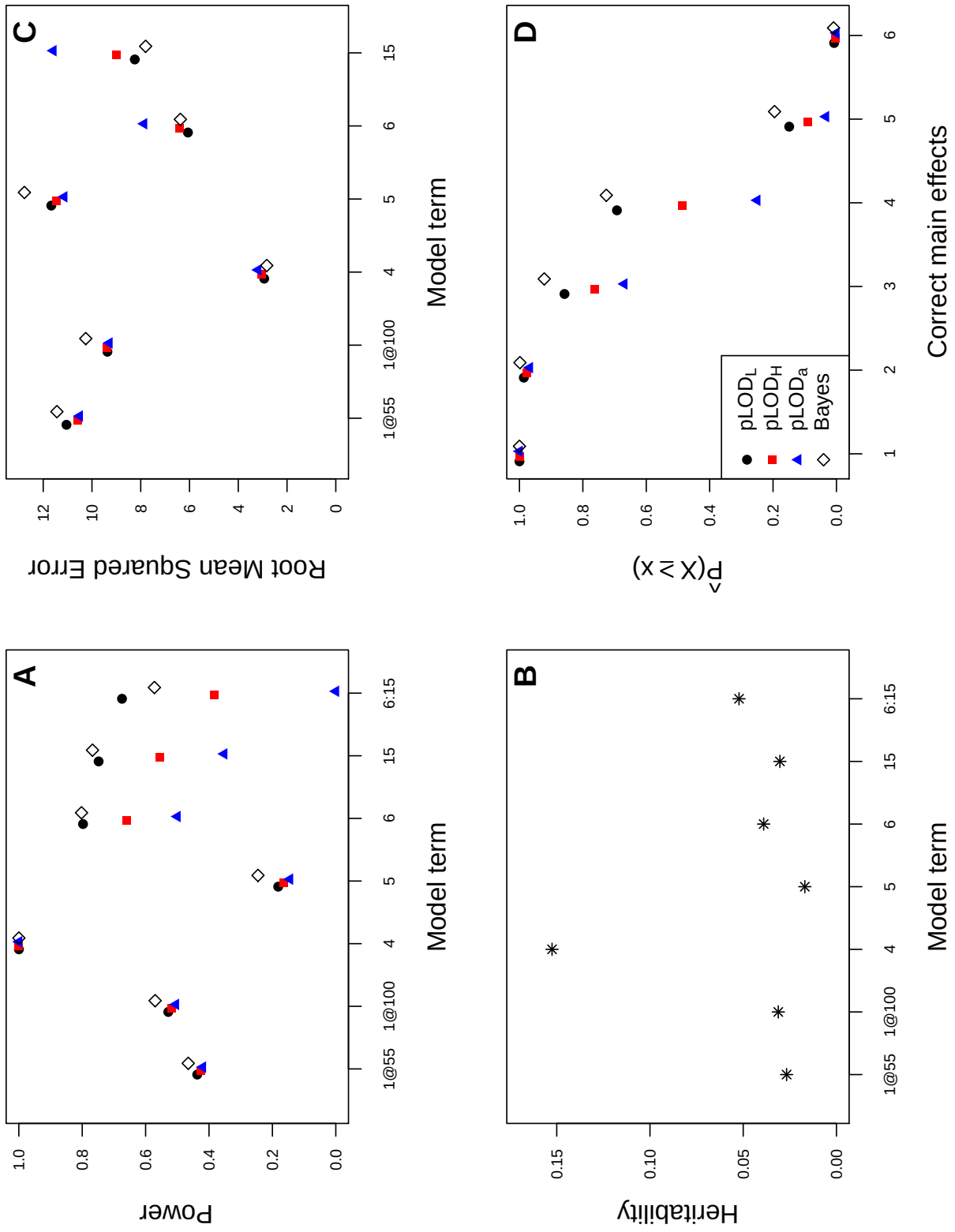
A
B
C
D
Power
Heritability
Root Mean Squared Error
P>(X ≥ x)
Model term
Model term
Model term
Correct main effects
1@55
1@100
6:15
pLOD_L
pLOD_H
pLOD_a
Bayes

Figure 4

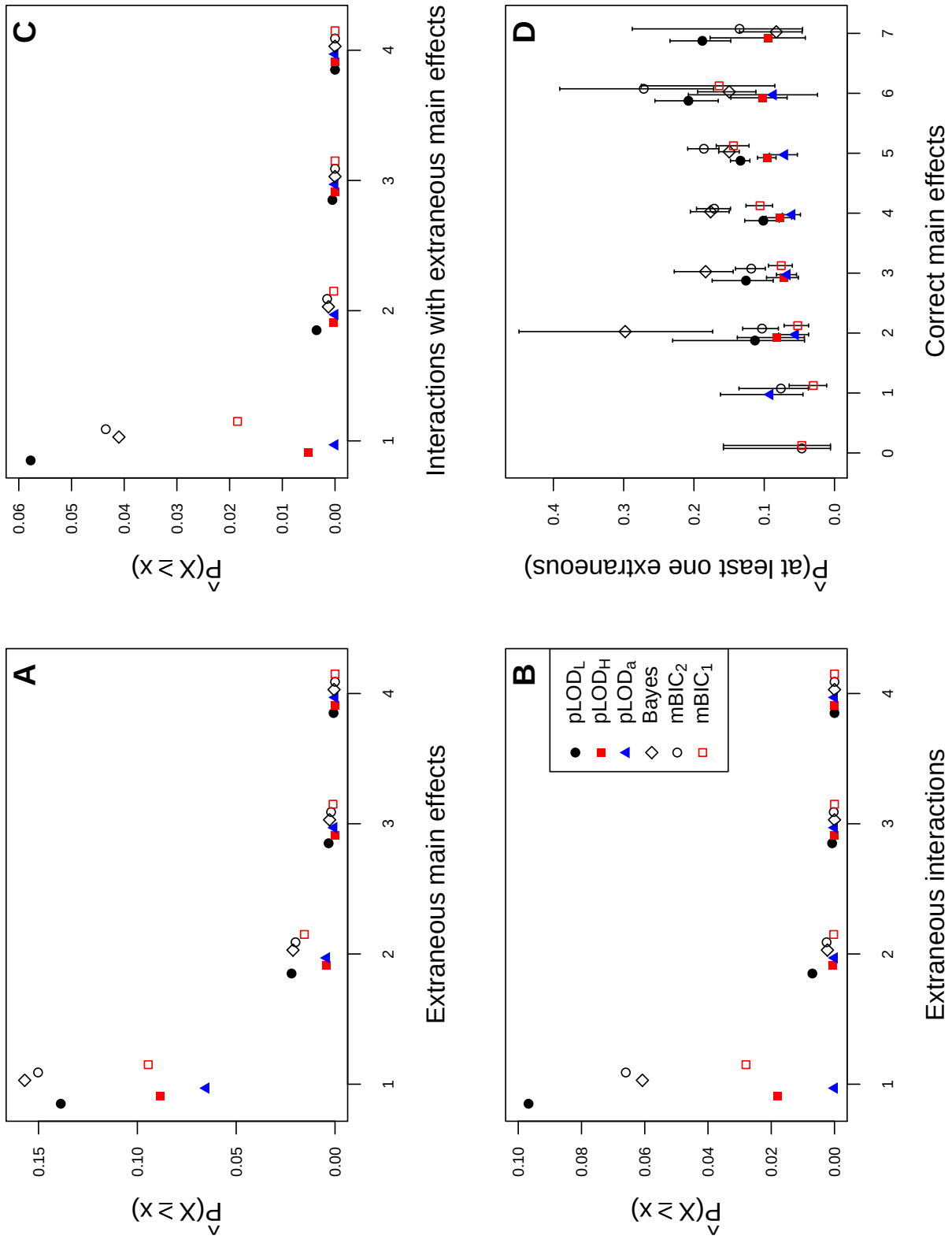
A
B
C
D
Extraneous main effects
Extraneous interactions
Interactions with extraneous main effects
Correct main effects
P>(X ≥ x)
P>(at least one extraneous)
pLOD_L
pLOD_H
pLOD_a
Bayes
mBIC_2
mBIC_1

Figure 5

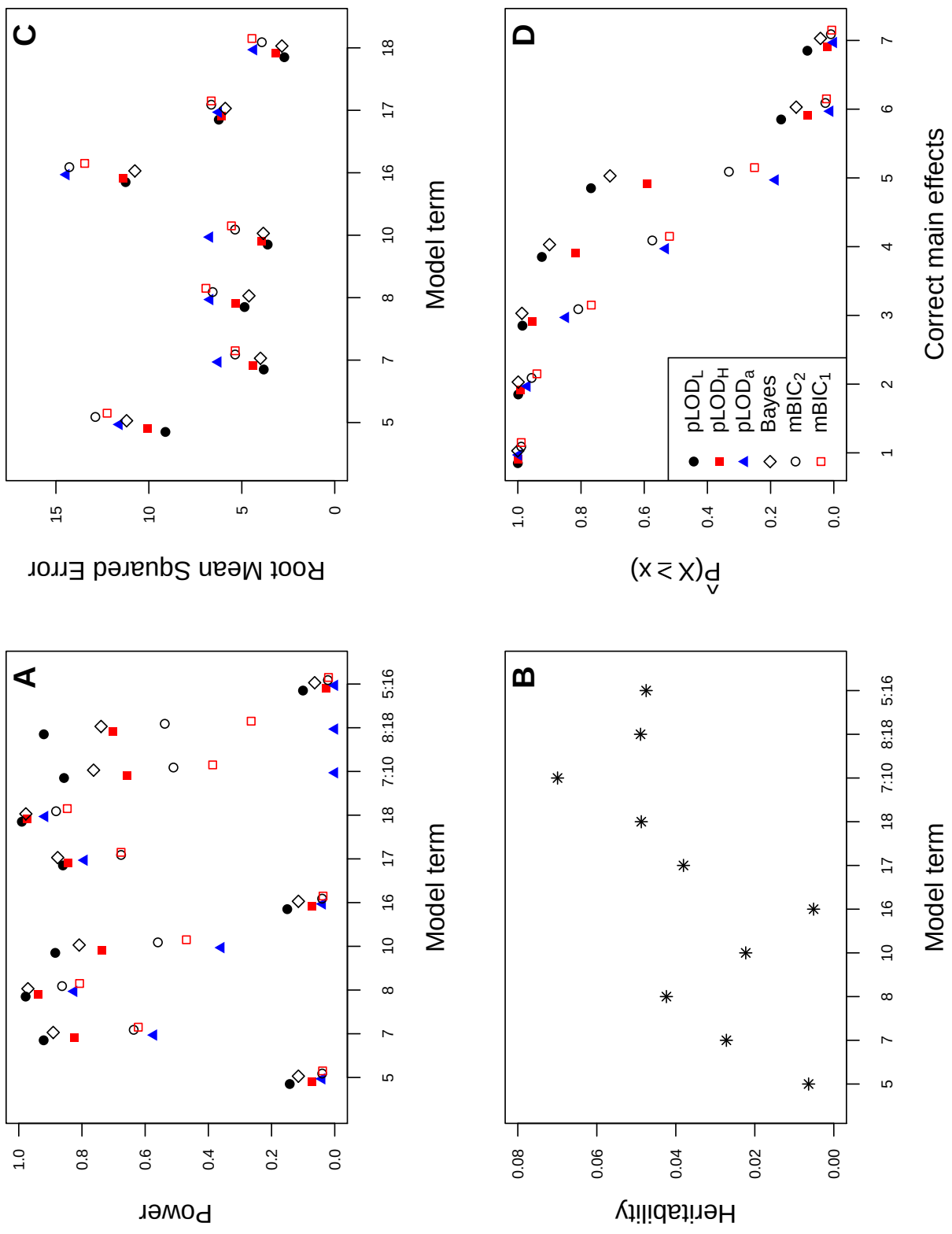
A
B
C
D
Power
Heritability
Root Mean Squared Error
$\hat{P}(X \geq x)$
Model term
Correct main effects
pLOD$_L$
pLOD$_H$
pLOD$_a$
Bayes
mBIC$_2$
mBIC$_1$

Figure 6