Statistical Applications in Genetics and Molecular Biology

Volume 10, *Issue* 1 2011 *Article* 15

Application of the Lasso to Expression Quantitative Trait Loci Mapping

Andrew Anand Brown, *Oslo University Hospital and University of Oslo*
Sylvia Richardson, *Imperial College London*
John Whittaker, *GlaxoSmithKline and London School of Hygiene & Tropical Medicine*

Recommended Citation:
Brown, Andrew Anand; Richardson, Sylvia; and Whittaker, John (2011) "Application of the Lasso to Expression Quantitative Trait Loci Mapping," *Statistical Applications in Genetics and Molecular Biology*: Vol. 10: Iss. 1, Article 15.
DOI: 10.2202/1544-6115.1606

Application of the Lasso to Expression Quantitative Trait Loci Mapping

Andrew Anand Brown, Sylvia Richardson, and John Whittaker

Abstract

Univariate methods have frequently been used to discover Quantitative Trait Loci for gene expression measurements, often with much success. However, correlations caused by Linkage Disequilibrium as well as chance correlations, which are functions of the large number of markers typically used in such studies, mean that causative regions can often cause multiple signals. Traditional investigations into the number of QTL for a given phenotype, such as visual inspection of likelihood plots, are not feasible when considering thousands of phenotypes. Stepwise methods have been suggested to counter this, but these are known to produce unstable models and there are difficulties in deriving significance estimates. The Lasso is a shrinkage method which has often been employed to discover true signals when the number of variables exceeds the number of observations. We propose a test statistic based on the threshold at which variables enter the Lasso model, prove analytic properties of this statistic which demonstrate parallels with univariate methods and demonstrate its utility in proposing candidate QTL. We show that this method controls for LD structure, and the estimates of statistical significance produced have superior properties when compared to those derived by stepwise methods. We study the performance of our method using simulation studies. These simulations find that the ratio of true discoveries to false positives is often superior for our method compared to univariate and stepwise approaches. Finally, we apply the derived method to data from a previous eQTL mapping experiment to investigate the nature of genetic regulation in this population.

KEYWORDS: expression quantitative trait loci mapping, gene expression, Recombinant Inbred Lines, lasso, penalised regression, hypothesis testing

Author Notes: Andrew Anand Brown, Division of Mental Health and Addiction, Oslo University Hospital, Oslo, Norway. Institute of Basic Medical Sciences, Department of Biostatistics, University of Oslo, Norway. Sylvia Richardson, Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, UK. John Whittaker, Genetics, GlaxoSmithKline, Stevenage, UK. Department of Non-communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK. We would like to thank Professor Tim Aitman and Dr. Enrico Petretto for providing access to the dataset this paper is mainly based on, and guidance for how it should be interpreted. We would also like to thank Statistics for Innovation, Oslo and the Royal Society Wolfson Awards for supporting the work of Professor Richardson and the Medical Research Council for funding the PhD studies of Dr.

Brown. Professors Richardson and Whittaker contributed equally to this paper and should be regarded as co final authors.

1 Introduction

The development of microarrays and other genomic technologies has allowed the collection of a wealth of data on processes at the cellular level. Animal models have provided a particularly fruitful avenue for the exploration of these molecular phenotypes: environmental variation and genetic pedigree can be strictly controlled, and experiments can be designed to test specific hypotheses. There have been a number of successes from this approach. For example, the use of rat models to discover genes under genetic control has led to the identification of candidate genes implicated in conditions such as neuroinflammation, glomerulonephritis, mammary cancer, left ventricular mass and heart failure (Aitman et al., 2008).

Typically expression Quantitative Trait Loci (eQTL) mapping studies have borrowed both their design and their analysis from the field of quantitative trait loci mapping: this was originally the search for loci associated with *any* continuous phenotype. The principal difference is that traditional QTL studies have investigated relatively few phenotypes, at most hundreds compared to the thousands studied in eQTL experiments.

These eQTL studies typically searched for transcripts where expression segregated at a known marker (Hübner et al., 2005); techniques such as Interval Mapping (Lander and Botstein, 1989), which considered associations with unobserved loci, were not implemented. This is because as genotyping costs fell and denser marker maps with better coverage of the genome became available, the need to consider such loci was no longer obvious. These denser marker maps have however brought new challenges: QTL can now be in Linkage Disequilibrium (LD) with several makers and it can be difficult to differentiate between two significant markers correlated with a single causal locus, or two significant markers which represent two separate causal mechanisms.

This problem is exacerbated in the case of studying Recombinant Inbred Lines. Such populations have a short breeding history, typically extending over around 60 generations and as such there remains considerable LD. This means that any causal locus is likely to be linked to several of the genotyped markers, and it can be problematic to disentangle multiple signals caused in this manner from multiple signals that are related to multiple distinct loci which regulate a given gene. Simply using the location of signals to judge the number of causal loci can give misleading results: the large numbers of markers typed in these experiments means that markers in distinct regions of the genome can be highly correlated simply by chance; in addition there have also been other studies which have discovered multiple sources of regulation in the same genomic region (Casola et al., 2001).

Doerge and Churchill (1996) extended Interval Mapping to search for multiple QTL for a given phenotype: they suggest a stepwise procedure, known as

Residual Empirical Threshold (RET) method, which iteratively searches for further QTL, conditioning on all those previously discovered. These methods rely on resampling methods to estimate the significance of QTL, such methods have a computational cost. For this reason Zou et al. (2004) suggested a less computationally intensive method to approximate the necessary distributions. However, modern computing power is easily sufficient for problems involving thousands of markers (as is typically the case in these animal experiments) and millions of permutations. When only considering potential eQTL at marker locations, such approximations are no longer necessary for typical animal eQTL analyses.

Litvin et al. (2009) have suggested a stepwise procedure for discovering sets of two or three loci where there is evidence of interaction on a given transcript. As a final stage the authors partition the genes into groups which show similar linkages. Other authors have suggested more sophisticated models which provide a wider search of the model space: in particular Bayesian Variable Selection schemes based around Reversible Jump Markov Chain Monte Carlo (Stephens and Fisch, 1998, Yi et al., 2003, Jia and Xu, 2007, Bottolo and Richardson, 2010). We will not discuss such methods here, or related Bayesian shrinkage regression methods such as Xiaohong and Shizhong (2010) which do not produce sparse solutions but rather significance estimates through resampling, as they involve a high computational cost which limits their application to expression QTL mapping on a large scale. Zhang et al. (2010) have proposed a Bayesian partition model to discover sets of genes linked to common markers. Another approach to eQTL mapping has been suggested by Kendziorski et al. (2006), that of Mixture over Markers. However this method has an inbuilt limitation that each expression can be regulated by no more than one marker.

More recently, shrinkage methods have been applied to expression QTL mapping, for example Pan (2009) proposes a penalised regression method. The network penalty suggested shares properties with a widespread method known as the Lasso (Least Angle Shrinkage Selection Operator) which, as with Pan (2009), gives sparse solutions. However, it is also able to incorporate information on which genes are associated within a given pathway, potentially increasing power to discover common eQTL. Chun and Kele (2009) describe an approach based on another shrinkage method: partial least squares. The authors name their method the M-SPLS eQTL method, and show that it proposes blocks of markers in LD. They argue that this is superior to arbitrarily choosing one marker from a given region to represent the group. This property is also shared by a method known as the Elastic Net: Zou and Hastie (2003) demonstrate that this grouping of related variables can lead to greater prediction accuracy, illustrating this by using gene expression to classify leukemia. We would argue that focussing on groups of correlated predictors is not necessarily of direct interest in eQTL mapping: the central point of interest is the causal loci, not the numbers of markers correlated with causal loci.

This reasoning has led us to focus on the Lasso. We use it to suggest a more consistent method for eQTL discovery than stepwise methods, which also accounts for multiple signals produced by LD. In contrast with Pan (2009), we do not attempt to exploit the dependencies between transcripts within the same pathway. While coregulation is an important question into which there has been much recent research, we wish to focus on uncovering eQTL specific to individual transcripts and to place the Lasso in the context of hypothesis testing. In this work we are interested in producing estimates of statistical significance, in addition to lists of eQTL produced by the Lasso.

The Lasso is an example of a penalised regression method; mentioned above this class of methods is similar to ordinary least squares (OLS) regression but with a penalty on the size of the coefficients to prevent overfitting and produce identifiable models when the number of variables exceeds the number of observations. Versions of the Lasso have been applied in many contexts, including considering functional connectivity between brain regions (Valdés-Sosa et al., 2005), inferring gene regulatory networks (Meinshausen and Bühlmann, 2006), and survival analyses of patients with liver disease (Zhang and Lu, 2007). There are several reasons for its ubiquity. This class of model produces sparse solutions, frequently a desirable property; there is an intuitive geometrical interpretation to the solutions proposed (Tibshirani, 1996), and simulation studies have expanded on cases in which its predictive ability is superior to other methods such as ridge regression or partial least squares (Tibshirani, 1996).

Taking y_{ij} the phenotype and X_{ik} the genotype, where $i = 1, \ldots, n$ indexes the observation, j the transcript and $k = 1, \ldots, p$ the marker, the Lasso solution for a threshold t is defined as:

$$\underset{\beta_{jk}(t)}{\text{argmin}} \sum_i (y_i - \sum_k X_{ik}\beta_{jk}(t))^2 \text{ subject to } \sum_k |\beta_{jk}(t)| < t$$

There is an equivalent formulation, which we shall concentrate on in this paper:

$$\underset{\beta_{jk}(\lambda)}{\text{argmin}} (\sum_i (y_i - \sum_k X_{ik}\beta_{jk}(\lambda))^2 + \lambda \sum_k |\beta_{jk}(\lambda)|)$$

The parameter λ controls the amount of shrinkage, high values of λ will generally produce parsimonious models. Our contribution is to propose a different perspective on the parameter λ and to use it to define a test statistic to estimate the significance of each marker as a candidate eQTL. To be precise, let us consider statistics given by the threshold at which markers first enter the model:

$$\lambda_{jk} = \begin{cases} 0 & \text{if } \beta_{jk}(\lambda) = 0 \forall \lambda > 0 \\ \max \lambda \text{ such that } \beta_{jk}(\lambda) \neq 0 & \text{otherwise} \end{cases}$$

The first case in this formula is necessary because the solution at $\lambda = 0$ is unidentifiable when p is greater than n: without this λ_{jk} would not be well defined for all j. We propose to combine this test statistic with a method of controlling the Family Wise Error Rate (FWER) which was described in Westfall and Young (1993), and we derive analytic properties for p values produced in this way. We show that, for a given transcript, the most significant eQTL discovered by the Lasso is identical to the most significant eQTL discovered by univariate methods, and that both methods assign identical significance to this marker. Subsequent eQTL proposed by the Lasso are given p values which control for previously discovered eQTL, removing false positive signals caused by LD. Unlike stepwise methods, the p values form a increasing sequence in order of discovery: Occam's razor suggests this is a desirable property.

We will use simulated data to investigate the behaviour of this approach in a biological reasonable setting. We find that in many situations the average proportion of true discoveries amongst proposed eQTL is greater for this Lasso approach than the univariate approach and the RET approach of Doerge and Churchill (1996). Finally we will consider the data from the experiment presented in Hübner et al. (2005), and investigate the nature of genetic regulation of gene expression within this animal population.

2 Data

This investigation was conducted principally in the context of an experiment in RI rats mentioned previously, reported in Hübner et al. (2005). This experiment used 120 rats, from 30 Recombinant Inbred lines derived from the Brown Norway (BN) and Spontaneously Hypertensive Rat (SHR) inbred lines (the SHR is especially susceptive to metabolic syndrome, this has been a much studied strain (Aitman et al., 2008)). We have data on the expression of 15,923 transcripts in fat cells, and marker data at 800 loci. Of these 800 loci, 34 were in perfect LD with another marker: these were removed leaving 766. More detail on this experiment is available in Hübner et al. (2005). We use this data both for substantial analysis and as a basis for our first simulation experiment.

We repeated the simulation studies in two other datasets, to investigate how our conclusions generalise to other populations. We used genotype data from an experiment in barley using double haploid lines derived from the Steptoe and Morex

DOI: 10.2202/1544-6115.1606

lines and presented in Kleinhofs et al. (1993). There were 150 strains, each was genotyped at 495 markers. We removed 3 markers as they were in perfect LD with another marker, leaving 492 markers. The other dataset consisted of genotype data from an experiment into RI mice, presented in Taylor et al. (1999). We had genotype data at 3796 markers from 93 strains: we removed 3 strains as marker data was completely missing and 1549 markers as they were in perfect LD with another marker. This left 90 strains and 2247 markers. Both of these datasets are available at http://genenetwork.org/. In all three datasets the genotypes are coded 0 or 1, depending on the parental progenitor; we impute missing data by averaging the genotypes of the flanking markers.

We used these data in two ways: firstly we used the genotype data from all three datasets to simulate biologically reasonable expression profiles, to which we applied the methods under investigation. To illustrate our method on eQTL data, we also analysed the gene expression data generated by the experiment from Hübner et al. (2005), using the considered methods to draw conclusions on the nature of genetic regulation in this population.

For our simulation study based on the marker and gene expression data from Hübner et al. (2005) we have generated datasets of 120 independent observations per transcript. We cannot however consider the experimental data as 120 independent observations: there will be dependencies between the four observations within each line caused by unobserved genetic variation or possibly differing environmental conditions which must be considered. Instead we pooled data, considering the means within lines. Belknap (1998) and Zou et al. (2006) both contain discussions of this. There is a further complication that one line was found to be genetically contaminated (Petretto et al., 2006). We removed data from this line, leaving 29 observations.

We use the notation described above and, to simplify the algebra, scale and center the data so that $\sum_i y_{ij} = \sum_i X_{ik} = 0$ and $\sum_i y_{ij}^2 = \sum_i X_{ik}^2 = 1$.

3 General model

With the assumption that the chosen set of markers captures all of the genetic information we can formulate a general model for a transcript j which simultaneously considers all genetic effects:

$$y_{ij} = \sum_k X_{ik}\beta_{jk} + \varepsilon_{ik}, \quad \varepsilon_{ik} \sim N(0, \sigma_k^2)$$

For each locus k we wish to know whether expression depends on genotype, i.e. we wish to test the hypothesis $H_{jk}^0 : \beta_{jk} = 0$. In most eQTL mapping experiments there are many more markers than individuals, this means that the general model is unidentifiable. Univariate methods, such as t tests for equality of mean or likelihood ratio (LR) tests of nested univariate models, investigate each marker separately: this means that they ignore effects of confounding markers and this can generate false positives. Stepwise methods, in contrast, use an iterative procedure to produce a sub model of the general model. This choice of sub model can be unstable, to the extent that omission of a single datum can completely change the chosen model (Breiman, 1995).

4 Multiple testing

Application of any of these methods produces a test statistic summarising evidence of differential expression for each transcript at each marker (in the case of stepwise selection many of these statistics will be zero as the associated marker does not enter the model before the algorithm terminates). These statistics are converted into a measure of significance, which allows a set of eQTL to be declared, accompanied with a statement on the statistical significance of this list. It is well established that p values are poor guides to significance when testing more than one hypothesis: uncritically adopting a 0.05 cut off leads to unacceptable numbers of false positive results (Benjamini and Hochberg, 1995). This has led authors to suggest extensions to the concept of the p value, which take the number of tests into consideration.

One such extension is the Family Wise Error Rate (FWER). While the p value estimates the probability of falsely rejecting a single true hypothesis, procedures for controlling the FWER are designed to place asymptotic bounds on the probability that at least one hypothesis is falsely rejected. A procedure for controlling the FWER, known as the max T procedure, is outlined in Westfall and Young (1993). To test hypotheses $H_1,...,H_n$, it compares test statistics $\hat{t}_1,...,\hat{t}_n$ to the distribution of the maximum test statistic under the complete null hypothesis, $T = \max(t_1,...,t_n|H_1,...H_n$ are true). The distribution of T has typically been estimated by permutation methods: datasets are randomly permuted and the maximum test statistic calculated over all hypotheses is taken as a realisation of T. Adjusted p values are calculated as $p_i = P(T > t_i)$. Westfall and Young (1993) show that a decision rule based on rejecting any hypothesis with $p_i < \alpha$ offers weak control of the FWER at the level α; they go on to specify a condition under which this procedure offers strong control. Weak control means that the procedure controls the FWER under the complete null hypothesis, strong control offers control under any combination of true and false null hypotheses.

Doerge and Churchill (1996) present the RET method, which combines stepwise regression with an adapted max T procedure to estimate the significance of the discovered eQTL. This begins with the null model and proceeds iteratively. The first stage compares the maximum of the univariate test statistics across all loci under consideration to the distribution of this statistic under the complete null, as estimated using permutations. If this loci exceeds a given significance threshold it is declared an eQTL, and the procedure continues iteratively; if not the procedure stops, having discovered no evidence of genetic control of the phenotype. At each subsequent iteration the method redefines the phenotype as the residuals of the original phenotype, after controlling for all the currently postulated eQTL. Then permutation methods are used to assess significance of the remaining loci; if one is deemed significant it is added to the list and the process continues, else it halts.

In this manner a list of loci $L_1,..,L_r$, together with p values for significance $p_1,...,p_r$ is produced, with $p_j < \alpha$ for all j (α is the predefined cut off for significance). This approach has drawbacks, as discussed in Storey (2003). Firstly, while at each stage multiple testing across the markers is controlled for, the number of stages is not. Secondly, each of the p values is conditioned on a different model; this makes direct comparison problematic. The calculated p values depend on the order in which the eQTL are discovered, the correlations between eQTL and the sign of the effects. In particular, the p values no longer necessarily form an increasing sequence. Subsequently discovered eQTL can be attached greater significance than the first eQTL discovered due to the correlations between these markers and the nature of their effects.

The FWER has been widely adopted as a measure of significance in traditional QTL studies (for example Bottger et al. (1998)) but with expression QTL studies, in which the number of tests can be greater by factors of tens of thousands, this approach has been typically seen as too conservative: numerous interesting results would be discarded for not reaching the appropriate significance threshold. As such, the less conservative False Discovery Rate (FDR) measure has often been used (Benjamini and Hochberg, 1995). One commonly applied approach is to use the max T procedure to control the FWE across the markers, and subsequently apply an FDR procedure to control for multiple testing across expressions, e.g. Hübner et al. (2005). This is the approach we apply in this paper.

5 Constructing a test statistic from the Lasso model

Our intention is to use the Lasso solution to the general model to suggest eQTL while avoiding the pitfalls of the univariate and stepwise approaches. We define a test statistic for each locus, which assesses evidence of differential expression, as follows:

$$\lambda_{jk} = \begin{cases} 0 & \text{if } \beta_{jk}(\lambda) = 0 \forall \lambda > 0 \\ \max \lambda \text{ such that } \beta_{jk}(\lambda) \neq 0 & \text{otherwise} \end{cases}$$

This is the threshold at which marker k first enters the Lasso model. We compare this statistic to the distribution of the maximum over all markers of this statistic under the complete null hypothesis ($H_j : \beta_{j1} = ... = \beta_{jp} = 0$); this distribution is estimated using permutation methods.

In Appendix A.1 we consider the application of permutations to calculate p values adjusted for multiple testing by Westfall and Young (1993); either to univariate statistics (equivalently the t test or a likelihood ratio test) or the Lasso statistic described above. We show that, if the same set of permutations are used, then identical minimum p values across all markers will be calculated by this Lasso method or either of the univariate approaches. We believe this is an important property of our approach, when declaring only one QTL for a given transcript our results exactly coincide with those of a univariate analysis (this is not necessarily true for stepwise methods). However, when declaring multiple QTL our approach diverges from the univariate analysis: confounding between markers is considered.

We can see this from the lars algorithm, proposed in Efron et al. (2004) to efficiently calculate the Lasso solution for all values of λ. This algorithm is based around the insight that the coefficients are piecewise linear functions of λ. For any given value of λ, the absolute covariance of a variable with the residuals of the model is the same for all variables in the Lasso model. Using Efron et al. (2004), we can write an expression for λ_{jk} in terms of the Lasso solution at this value $\beta(\lambda_{jk})$. Given l such that $\hat{\beta}_l(\lambda_{jk}) \neq 0$, and denoting the covariance function, cov:

$$\lambda_{jk} = |\text{cov}(X_l, y - X\hat{\beta}(\lambda_{jk}))| = |\sum_i X_{il}(y_{ij} - \sum_m X_{im}\hat{\beta}_{jm}(\lambda_{jk}))|$$

We read directly from this formula that, when proposing eQTL, the Lasso accounts for multiple signals and LD: the threshold is dependent on the covariance of the marker with the phenotype, after controlling for previously discovered eQTL.

Markers enter the model at a series of thresholds $\lambda_{j1} \geq \lambda_{j2} \geq .. \geq \lambda_{jp}$. The p values are calculated by comparing these thresholds to the same distribution, that of $\max_k \lambda_{jk}$ under the complete null hypothesis. This means that the associated measures of significance will form an increasing sequence in the order of which markers enter the model, $p_{j1} \leq p_{j2} \leq ... \leq p_{jr}$. In contrast, p values calculated using RET no longer necessarily form an increasing sequence. An appeal to Occam's razor persuades us that this is a desirable property: the Lasso automatically favours simple explanations of the genetic variation.

6 Simulation study

We compared the results of the application of our method to those produced by two other methods: the univariate approach suggested by Lander and Botstein (1989) (equivalently this could be based upon t tests or LR tests) and RET, the stepwise approach described in Doerge and Churchill (1996). We also investigated whether the use of tag markers could automatically correct for the correlations between signals, by repeating the analysis on a pre selected subset of markers which captured most of the genetic variation.

The methods were compared by simulating datasets of 10,000 gene expression transcripts using the genotype data from the three experiments described in Section 2. We simulated data based on two genetic models, in each case we simulated varying numbers of simple additive QTL located at marker positions with fixed effect size. Under one genetic model each transcript had between zero and three QTL of effect size 0.5 (we will refer to this as Model A), in the other there were between zero and ten QTL of effect size 0.25 (Model B). Each transcript included a random gaussian noise term with variance 1. Explicitly we proceeded as follows: first a subset of markers C was chosen, with the number of elements equal to the number of eQTL to be simulated. Then y_{ij} was sampled from the following distribution:

$$y_{ij} \sim N(0.5 \sum_{k \varepsilon C} X_{ik}, 1)$$

where X_{ik} are the unscaled genotypes, equal to zero or one depending on parental progenitor in all three populations. Once generated, X_{ik} and y_{ij} were centred and scaled to have mean zero and variance one.

It is important to check that the number of eQTL and the proportion of variance explained by genetics can be justified as biologically reasonable. The mean heritabilities, calculated on the 10,000 simulated datasets, are displayed in Table 1. As a comparison, there have been 40 loci found to be associated with height, and the heritability of this trait is around 80% in certain populations (Manolio et al., 2009). These values are also consistent with the experimental data as analysed in Petretto et al. (2006).

We applied the methods under investigation to produce a matrix of p values, where p_{jk} captures the evidence that transcript j is regulated by marker k, controlling for multiple testing across the markers using the Westfall and Young permutation procedure (2000 permutations were used). A set of candidate eQTL was declared at a given threshold α as all marker/transcript pairs such that $p_{jk} < \alpha$. Figure 1 shows the number of eQTL declared at various thresholds when the methods were applied to the data from Hübner et al. (2005) with eQTL simulated under

	Model A	Model B
Rat	0.090	0.112
Barley	0.093	0.116
Mouse	0.093	0.123

Table 1: Mean heritability of the simulated datasets under the two genetic models.

Model A. We see that univariate methods declare many more transcripts at a given significance threshold than either the Lasso or stepwise methods; we expect many of these eQTL to be "ghost eQTL", induced by marker correlations. Stepwise regression is more conservative than the Lasso: we can view this in the context of Efron et al. (2004) who showed parallels between the Lasso and stepwise regression, describing the former as a less greedy relation to the latter. When considering the overlap between eQTL declared by various methods, the line describing the behaviour of the Lasso, and the line for the overlap between the t test and the Lasso appear indistinguishable; in fact there are 36 eQTL which are proposed by the Lasso and not by univariate methods at a given threshold. We also see that few of the eQTL declared by stepwise regression are unique to that method.

To assess the effectiveness of the procedure for declaring eQTL, we calculated two quantities which correspond to realisations of the FDR and the power. For a measure analogous to the power, we considered the proportion of all simulated eQTL that were in the candidate set; for a measure related to the FDR, the proportion of false positives.

The results are displayed in Figures 2, 3 and 4 for the three datasets. The range of the x axis differs between graphs. This is because the efficacy of the methods differs between simulations. In the case of the mouse data, where there are many markers and high LD, relatively few candidate eQTL are declared before the FDR first increases steeply and then plateaus. Therefore we have concentrated in this case on the region containing relatively strong candidates. The full graphs, where the threshold is allowed to vary from declaring no eQTL to declaring every marker an eQTL, can be found in Section A.3. Both the Lasso and RET procedures use stopping rules at which point no more eQTL are included in the model: the RET procedure terminates at some predefined α as discussed in Section 4, the Lasso is limited to including at most the rank of the design matrix variables in the model. For our simulations we set α=0.4, and restrict the Lasso to including at most 10 variables in the model: the threshold at which these limits are reached can be seen by the straight line segments in the graphs. These are arbitrary choices, but in each of the graphs the performance of the two methods appears to have diverged by the time this point is reached, thus we do not believe more lenient thresholds would change our conclusions. The extreme left hand side of the graph represents an average over few candidate eQTL, as such results in this region show more random variation. This is especially visible in the case of the mouse dataset with data simulated under Model B; the more complicated simulation strategy and the high degree of LD in the data leads to unstable models. In a comparison of the Lasso to stepwise methods, we see similar performance when few candidate eQTL are declared (the left hand side of these graphs) in all datasets and under all genetic models; in the barley dataset there is some evidence that stepwise methods discover a higher proportion of genuine eQTL when data is simulated under Model A while the Lasso is superior when data is simulated under Model B. At more moderate levels of significance, the Lasso shows superior performance in all three datasets. The shift from adopting FWER corrections to FDR corrections represents a realisation that concentrating on few, highly significant results risks discarding many interesting findings: it is in this window the Lasso appears to show superior properties. There are greater differences between univariate methods and stepwise or Lasso methods (which we will refer to together as multiple regression methods). When considering Model A we see that univariate methods perform worse than multiple regression methods in the rat and barley datasets. There is some evidence that univariate methods discover a higher proportion of genuine QTL at moderate levels of significance in the mouse dataset. We believe this is because the higher levels of LD in this dataset mean that multiple regression methods produce unstable models.

We have described how high LD in the data means univariate methods will frequently declare blocks of high correlated markers. This has severe implications for univariate methods: if a genuine eQTL is proposed as a candidate, often all its neighbours will also be attached a high level of significance; these neighbouring markers will contribute to the FDR but not the power. We investigated whether standard methods could alleviate this problem. One such method is that of tag markers (Chapman et al., 2003), in which a single marker is chosen to represent an LD block. We used the program Tagger (de Bakker et al., 2005) to chose a set of markers such that every marker was in LD with at least one member of this set ($R^2 = 0.8$). This required 548 markers in the rat dataset, 109 in the barley dataset and 633 in the mouse dataset. We repeated the simulation studies on these datasets, with eQTL simulated under Model A. These results are displayed in Figure 5. We see that accounting for LD reduces the difference between univariate and multiple regression methods in the rat and barley datasets: however, these methods maintain their lower ratio of false positives. We also see that any advantage univariate methods had in the mouse dataset is now removed, lower LD in this new dataset means more stable multiple regression models. When comparing multiple regression methods, we draw similar conclusions to the original analysis on the full set of markers, though differences in the barley dataset are now exaggerated.

7 Application to experimental data

We used univariate and the Lasso methods to analyse the experimental data from Hübner et al. (2005); the Westfall and Young approach was used to correct for multiple testing across the markers (the number of permutations varied between 1,000 and 100,000, according to how accurate an estimate of the p value was necessary). We used the approach described in Storey (2002) to estimate the FDR, controlling for multiple testing across transcripts.

We find that univariate methods suggest that many of the transcripts are regulated by multiple loci. However, most of these signals disappear once confounding between markers is controlled for using the Lasso. For instance, we find 159 transcripts with evidence of genetic regulation with an FDR of 0.01: univariate methods suggest a total of 227 eQTL responsible for this regulation, but after controlling for marker confounding using the Lasso there is no evidence of multiple regulation for *any* of these transcripts. We find 259 transcripts with an associated FDR of 0.05. Univariate methods suggest 455 loci responsible for this regulation, the Lasso declares that 254 are regulated by one locus, and five are regulated by two. Figure 6 displays the number of transcripts for which there are at least n eQTL for the two methods. The line corresponding to at least one eQTL is identical for the Lasso and the univariate approach. This is because of the result derived in Section 5: the most significant eQTL for a given transcript has the same p value using both univariate methods and the Lasso method. For larger numbers of eQTL for a given transcript we see that many of the multiple eQTL suggested by the univariate methods disappear once marker correlations are considered.

We list the transcripts for which there is the most significant evidence for regulation in Table 2 in Section A.2 (FDR=0.01, which implies the expected number of false positives is approximately 1.6). For the 112 transcripts for which an unambiguous location is available, 110 are on the same chromosome as the marker which regulates them, suggesting that most of the discovered regulation is *cis* regulation; this was also the conclusion reached in Hübner et al. (2005).

8 Discussion

This paper has provided a discussion of the Lasso method in the context of eQTL mapping, and a comparison of the Lasso to univariate and stepwise methods commonly applied to such studies. In comparison with a univariate approach we show that, in the absence of very high LD, controlling for the marker structure brings tangible benefits in the search for eQTL. Our simulation studies suggest that both the Lasso and RET show similar performance. However, there are additional arguments which support the Lasso as a superior alternative. In terms of hypothesis testing we have shown that the Lasso can account for confounding variables while weakly controlling the FWER. This is not true of RET. At each stage this procedure produces *p* values for the existence of further sources of regulation, corrected for multiple testing across markers but not corrected for the multistage process, and conditional on previously discovered effects. In addition, these *p* values no longer form a increasing sequence in the order in which eQTL are proposed, which we believe to be an attractive property.

We have shown in the methods section that the *p* value for the most significant eQTL for a given transcript is identical using the Lasso and univariate methods. There has also recently been an interest in other penalised regression methods which offer sparse solutions while claiming further superior properties, notably the Dantzig selector proposed in Candes and Tao (2007) (though the desirable properties of this estimator have now been shown to apply to the Lasso as well (Bickel et al., 2009)) and the Smoothly Clipped Absolute Deviation penalty (SCAD), proposed in Fan and Li (2001). James et al. (2009) show that the Lasso and Dantzig solutions coincide when $p = 2$, for SCAD the solutions are the same for one variable when $|\sum_i X_{ik} y_{ij}| < 2\lambda$. This means that the threshold at which the first variable enters these models is identical for the Lasso, Dantzig selector, and the SCAD methods; thus significance testing derived from all these models is equivalent to significance testing based on the maximum univariate test statistic. We see this as an important validation of this class of methods: when proposing a single association, the results agree with those of a univariate analysis (this is not true when choosing the most significant eQTL proposed by stepwise methods).

The Westfall and Young method achieves weak control of the FWER, i.e. control of the FWER given the complete null hypothesis. Strong control, which is control of the FWER under any combination of null and alternative hypotheses, requires a property known as subset pivotality for the test statistics. This requires that the individual test statistics are not functions of realisations of data related to more than one hypothesis, a property which cannot hold for any test statistic which measures significance of one eQTL, controlling for another. Univariate methods do fulfill the criterion of subset pivotality and, as these *p* values are calculated by reference to the same distribution under the complete null as the Lasso *p* values, the Lasso *p* values will be more conservative if the test statistics are smaller: this implies that if the inequality $|\sum_i X_{ik}(y_{ij} - \sum_l X_{il}\hat{\beta}_{jl}(\lambda_{jk}))| \leq |\sum_i X_{ik}(y_{ij})|$ (the absolute covariance of the variable with the residuals of this Lasso solution is less than the absolute covariance of the variable with the phenotype) would always be fulfilled, then strong control would follow. We find this inequality to hold for 98.4% of the eQTL suggested by application of the Lasso to the experimental data; furthermore the inequality holds for all eQTL whose *p* value after correcting for multiple testing across the markers is less than 0.3.

We have presented a method for analysing eQTL mapping experiments which uses penalised regression methods to remove "ghost" QTL, while still allowing traditional inference. If we were to restrict ourselves to searching for evidence of genetic variation for a given transcript, then the Lasso proves equivalent to the LR test when combined with permutation based correction, and it is unclear whether more sophisticated statistical methods would add more statistical power. The Neymann Pearson Lemma states that the LR test is the uniformly most powerful test, and permutation methods are less conservative than other multiple testing corrections commonly applied in genetic association studies, such as the Bonferroni or Holm corrections (though methods which borrow information across transcripts rather than treating them as independent could improve inference, hence the interest in Kendziorski et al. (2006), Litvin et al. (2009), Pan (2009), Zhang et al. (2010)). It is when investigating the numbers and locations of eQTL that the benefits of adopting a multivariate strategy appear. This paper has shown that the Lasso produces fewer false positives by eliminating those caused by high correlations between markers.

In our data analysis we have found few genes to have multiple sources of regulation. While this could be a genuine result, or the result of low power due to small sample size, it could also be due to limitations in the experimental design we have investigated. In the particular case of RI lines, the limited number of recombinations that have occurred over the sixty generation breeding history means that considerable LD remains, limiting the potential for fine mapping of loci. Thus, distinguishing between distinct loci independently responsible for *cis* regulation is not possible. However, the methods we have proposed can easily be adapted to outbred populations with finer scale LD structure. Both univariate and stepwise methods are used in human association studies (Orozco et al., 2009); the limitations of these approaches are, if anything, exacerbated in this setting. The outbred nature of these populations means that many more markers are needed to capture the genetic information, and advances in technology have meant that these denser marker maps are already available (both the Affymetrix and Illumina genechips measure on the order of half a million markers). As the number of markers under consideration rapidly increases we will have more highly correlated markers, thus the removal of these "ghost" eQTL will become ever more important. We have shown the Lasso to be well adapted to this task. The code necessary to implement this method in R is available at http://www.bgx.org.uk/software.html.

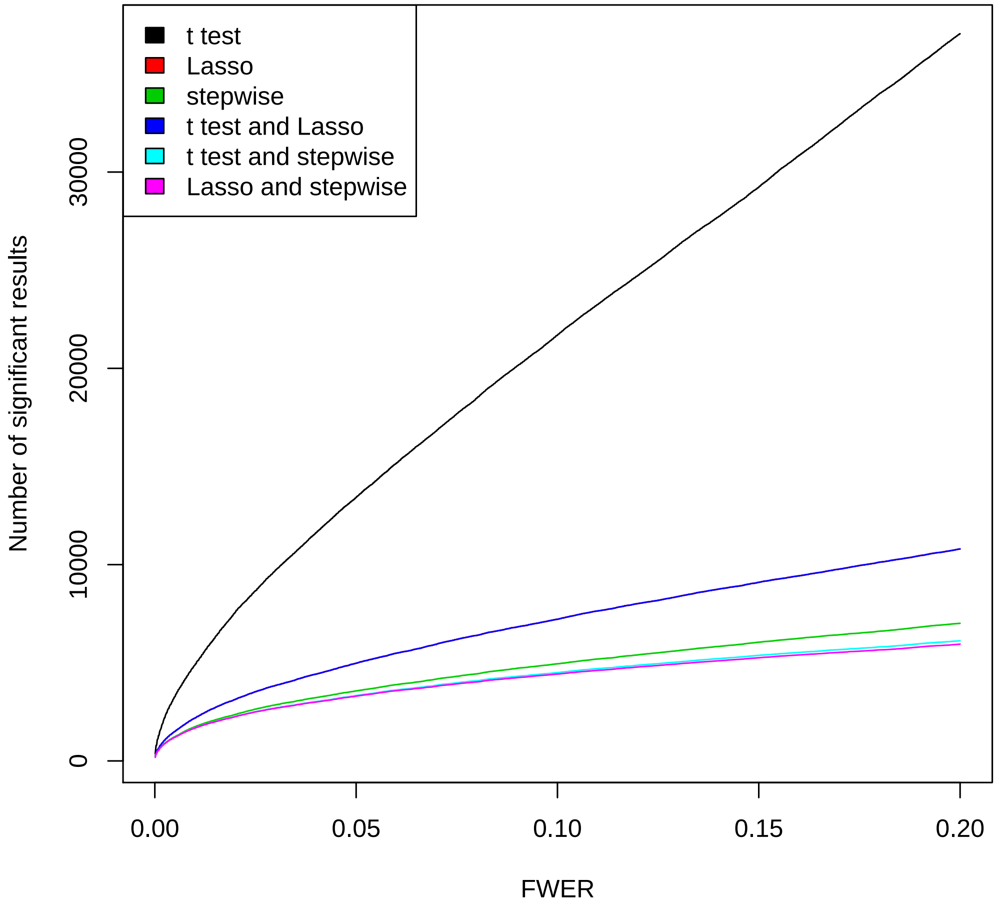
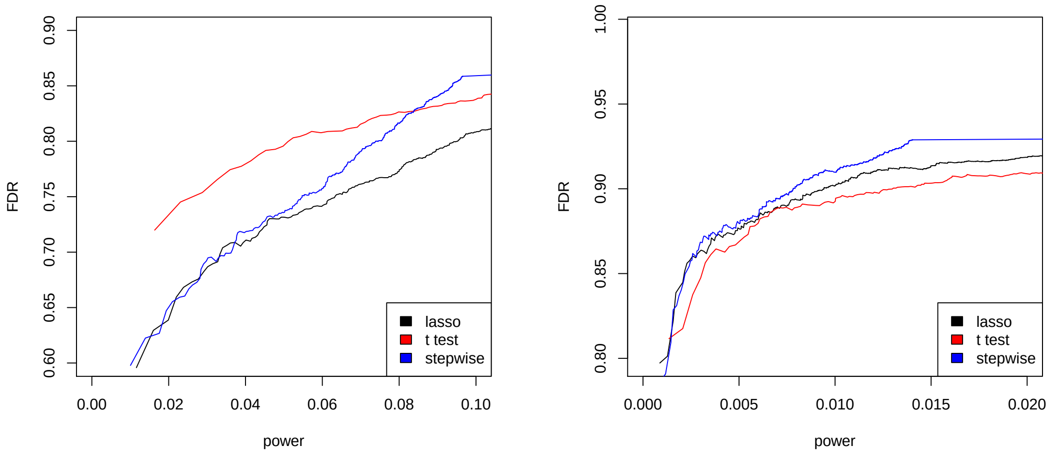


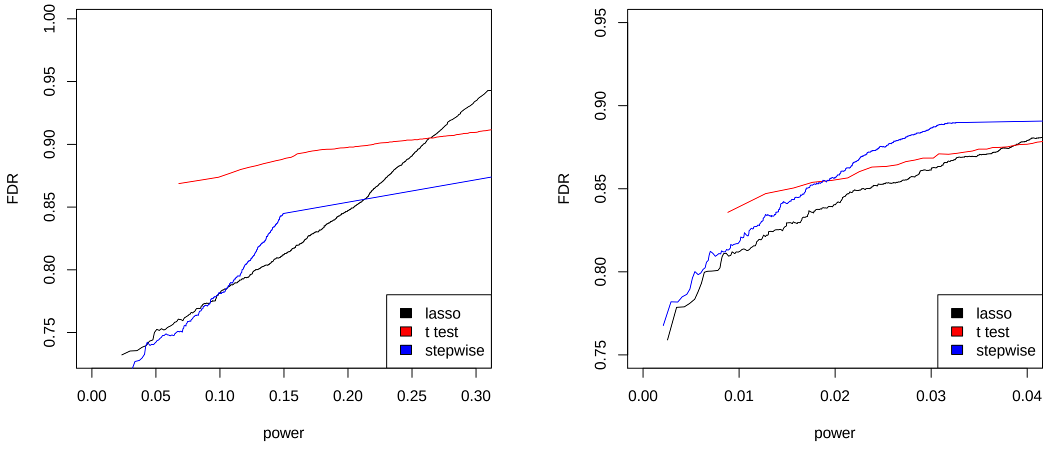
Figure 1: We display the number of significant eQTL declared at various significance thresholds by the three methods applied to 10,000 simulated datasets, as well as the intersections between these choices. Univariate methods declare the most eQTL, though many of these will form correlated blocks and relate to a single signal; stepwise methods are the most conservative. The line corresponding to the Lasso is almost completely obscured by the line corresponding to the intersection of the Lasso and *t* methods, implying that the Lasso proposes few unique eQTL.



(a) Genetic Model A

(b) Genetic Model B

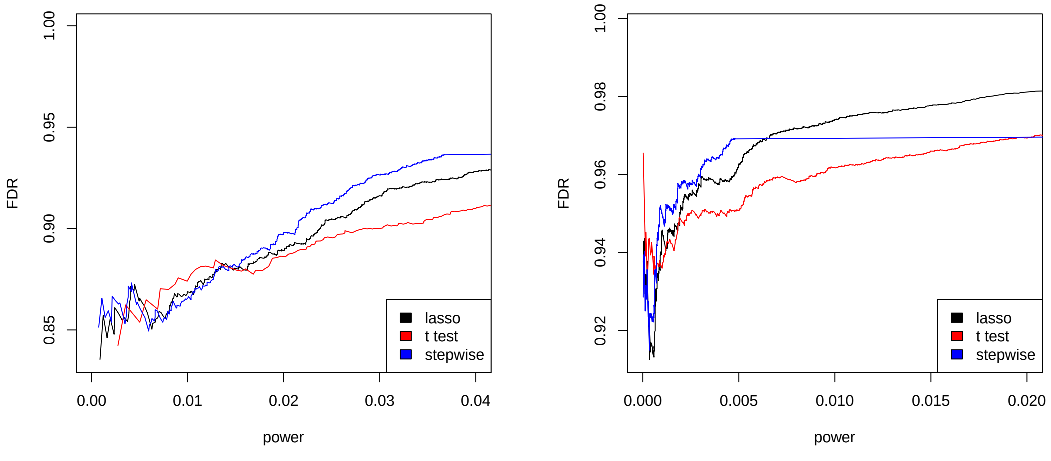
Figure 2: We plot the proportion of genuine eQTL discovered against the proportion of false positives at various thresholds when the three methods are applied to the datasets simulated from the RI rat data presented in Hübner et al. (2005) under genetic Model A (Figure (a)) and Model B (Figure (b)).



(a) Genetic Model A

(b) Genetic Model B

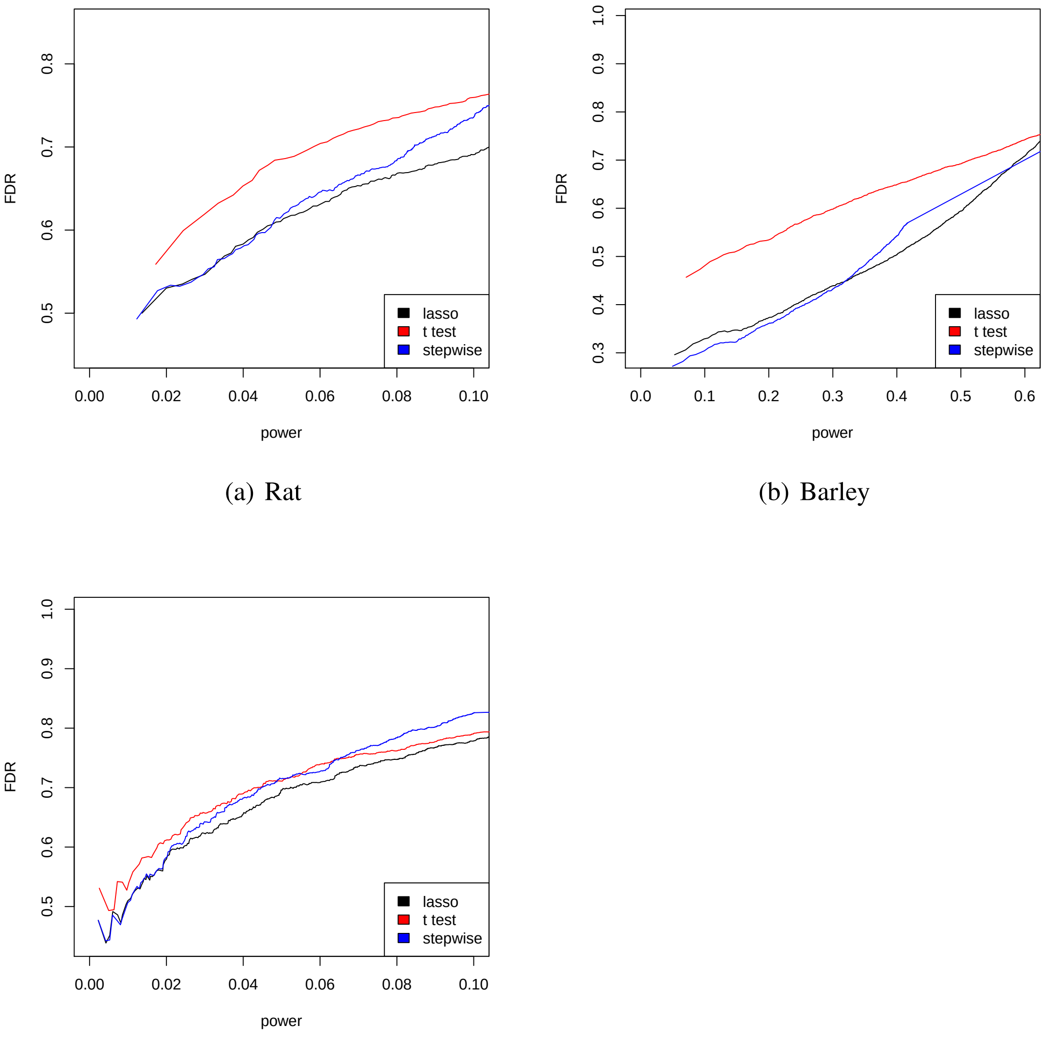
Figure 3: We repeat the simulation studies presented in Figure 2, this time gene expression profiles are simulated from the barley genotype data from Kleinhofs et al. (1993).



(a) Genetic Model A

(b) Genetic Model B

Figure 4: We repeat the simulation studies presented in Figure 2, this time gene expression profiles are simulated from the RI mouse genotype data from Taylor et al. (1999).



(a) Rat

(b) Barley

(c) Mouse

Figure 5: We repeat the simulation studies presented in Figures 2(a), 3(a) and 4(a) on a subset of tag markers from each dataset.

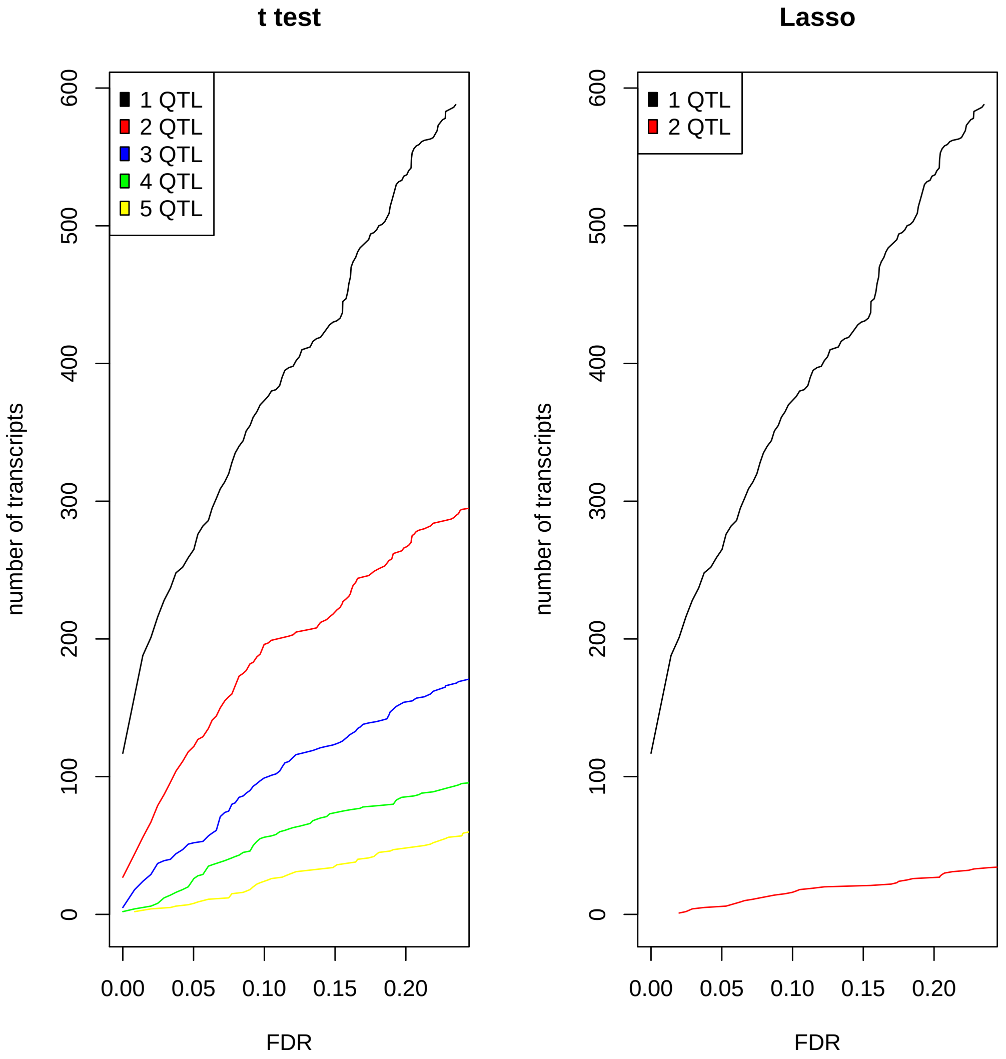


Figure 6: We show the number of eQTL found to regulate individual transcripts at various significance thresholds using the experimental data from Hübner et al. (2005). We see that univariate methods suggest that many transcripts are regulated by up to 5 eQTL. However, by using the Lasso method to correct for correlations between markers, we explain away the majority of this multiple regulation.

A Supplementary materials

A.1 Properties of the Lasso statistic

In this section we derive the result that when considering at most one QTL per transcript, our approach derives identical estimates of significance to univariate methods. We proceed by first deriving an explicit formula for $\lambda_j = \max_k \lambda_{jk}$ and showing that there are simple increasing functions which map the maximum t statistic, $t_j = \max_k t_{jk}$ and the maximum LR statistic, $\Delta_j = \max_k \Delta_{jk}$, onto λ_j. As these functions preserve the order comparisons using which such p values are calculated, we have the following immediate corollary: if the same set of permutations are used to calculate p values using the Lasso and both univariate methods, then the minimum p value across all markers will be the same for all three methods.

Proposition A.1 $\lambda_j = \max_k |\lambda_{jk}| = \max_k |\sum_i X_{ik} y_{ij}|$

Proof The Least Angle Regression (lars) algorithm (Efron et al., 2004) demonstrates that, as λ varies, variables enter and leave the model singly (with exceptions that can be solved by small perturbations of the data). This paper shows that the coefficients of the Lasso solution are piecewise linear functions in λ, the algorithm follows from this fact. Figure 7 gives an example of the solution the Lasso problem applied to randomly generated data, plotting the size of the coefficients against the value of λ. Because variables enter the model singly, there is an interval $[\lambda_j - \varepsilon, \infty)$ and a marker K for which $\hat{\beta}_{jk}(\lambda) = 0$ for all $k \neq K, \lambda \in [\lambda_j - \varepsilon, \infty)$. On this interval, the Lasso problem reduces to:

$$\underset{\beta_j}{\text{argmin}} \sum_i (y_{ij} - X_{iK}\beta_{jK})^2/2 + \lambda|\beta_{jK}|$$

as all other coefficients of $\hat{\beta}_j$ are zero. This means that we do not have to consider the variables jointly: we know that in the neighbourhood of λ_j only one coefficient is non zero. Therefore we investigate each marker k individually.

For each marker k we consider the Lasso solution when X_{ik} is the only explanatory variable, and denote the associated coefficient $\beta^*_{jk}(\lambda)$. We define λ^*_{jk} as the minimum value of λ for which $\hat{\beta}^*_{jk}(\lambda)$ is zero: i.e. $\lambda^*_{jk} = \min \lambda$ such that $\underset{\beta}{\text{argmin}} \sum_i (y_{ij} - X_{ik}\beta^2 + \lambda|\beta| = \hat{\beta}^*_{jk}(\lambda) = 0$. Then, $\lambda_j = \max_k \lambda^*_{jk}$, the threshold at which the *first* marker enters the model.

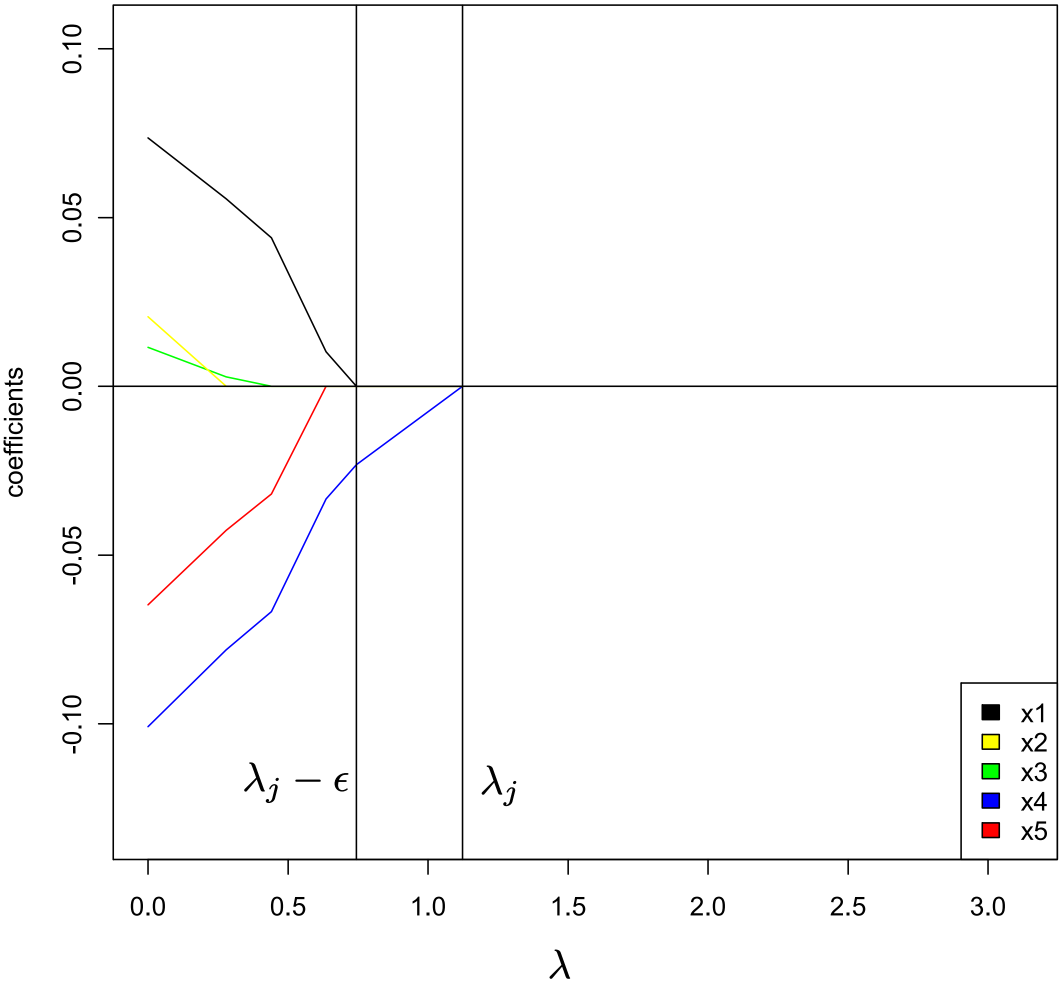


Figure 7: We show, as an example, the Lasso solution for a randomly generated dataset at various values of λ. This illustrates that the coefficients are piecewise linear functions in λ and that there exists an interval $(\lambda_j - \varepsilon, \lambda_j)$ on which only one coefficient is non zero.

The case of a single variable, scaled such that $\sum_i X_{ij}^2 = 1$, is a trivial example of an orthonormal design matrix, this was examined in more depth in Fan and Li (2001). They show that the Lasso solution reduces to the ordinary least squares coefficients translated towards the x axis by a constant (this is known as "soft thresholding", also referred to in Tibshirani (1996)), and give an explicit formula for $\beta_{jk}^*(\lambda)$:

$$\beta_{jk}^*(\lambda) = \begin{cases} \text{sgn}(\beta_{jk}^{OLS})(|\beta_{jk}^{OLS}| - \lambda) & \text{if } |\beta_{jk}^{OLS}| > \lambda \\ 0 & \text{otherwise} \end{cases}$$

β_{jk}^{OLS} is the ordinary least squares estimate when y_{ij} is regressed on X_{ik}. From this expression we deduce that $\lambda_{jk}^* = |\beta_{jk}^{OLS}| = |\sum_i X_{ik} y_{ij}|$. Taking the maximum over k we have $\lambda_j = \max_k |\sum_i X_{ik} y_{ij}|$. $\square$

We now deduce a number of results from this formula, finally proving that p values for the hypothesis $H_0 : \beta_{j1} = ... = \beta_{jp} = 0$ calculated on this statistic will be equal to p values testing the same hypothesis calculated using the maximum of the t test or LR test statistics. For this we will need the following Lemma:

Lemma A.2 *Let $T(y)$ be a test statistic calculated on a dataset y and f be a strictly increasing function on the range of T. Then, for a given set of permutations of y, $\{y_{\sigma_1}, ..., y_{\sigma_M}\}$, permutation p values, which are defined $p_T = \frac{\#T(y_{\sigma_m}) > T(y)}{M}$ and $p_f = \frac{\#f(T(y_{\sigma_m})) > f(T(y))}{M}$, are equal.*

Proof Because f is a strictly increasing function it preserves order relations: $T(y_{\sigma_k}) > T(y) \Leftrightarrow f(T(y_{\sigma_k})) > f(T(y))$; therefore p values based on these order relations will be equal. $\square$

It remains to show that the t statistic $t_j = \max_k t_{jk}$ and the LR test statistic $\Delta_j = \max_k \Delta_{jk}$ are strictly increasing functions of λ_j. First we calculate the range of λ_j:

Lemma A.3 $0 \leq \lambda_j \leq 1$

Proof The Cauchy Schwarz inequality gives us $|\sum_i X_{ik} y_{ij}| \leq \sqrt{\sum_i X_{ik}^2 \sum_i y_{ij}^2} = 1$ as $\mathbf{X}$ and $\mathbf{y}$ are scaled such that $\sum_i X_{ik}^2 = \sum_i y_{ij}^2 = 1$. $\square$

Lemma A.4 *The t statistic $t_j = \max_k t_{jk}$ is a strictly increasing function of λ_j.*

DOI: 10.2202/1544-6115.1606

Proof We can show that λ_j is the maximum of the absolute Pearson product moment correlations between the gene expression and the marker variables:

$$\begin{aligned} r_j = \max_k |r_{jk}| &= \max_k |\frac{n\sum_i X_{ik}y_{ij} - \sum_i X_{ik}\sum_i y_{ij}}{\sqrt{n\sum_i X_{ik}^2 - (\sum_i X_{ik})^2}\sqrt{n\sum_i y_{ij}^2 - (\sum_i y_{ij})^2}}| \\ &= \max_k |\sum_i X_{ik}y_{ij}| = \lambda_j \end{aligned}$$

because $\sum_i X_{ik} = \sum_i y_{ij} = 0$ and $\sum_i X_{ik}^2 = \sum_i y_{ij}^2 = 1$.

There is a well known transformation between the Pearson product moment correlation and the Student's *t* (Andrés et al., 1995):

$$\begin{aligned} t_{jk} &= |\frac{r_{jk}\sqrt{n-2}}{\sqrt{1-r_{jk}^2}}| \\ \Rightarrow t_j &= \max_k \frac{|r_{jk}|\sqrt{n-2}}{\sqrt{1-r_{jk}^2}} \\ &= \frac{\max_k |r_{jk}|\sqrt{n-2}}{\sqrt{1-\max|r_{jk}|^2}} \\ &= \frac{\lambda_j\sqrt{n-2}}{\sqrt{1-\lambda_j^2}} \end{aligned}$$

$f(x) = \frac{x\sqrt{n-2}}{\sqrt{1-x^2}}$ is a strictly increasing function on the range $0 < x < 1$. □

Lemma A.5 *The LR test statistic $\Delta_j = \max_k \Delta_{jk}$ is a strictly increasing function of λ_j.*

Proof The LR test for transcript *j* and marker *k* compares two nested models: $y_{ij} - \alpha X_{ik} \sim N(0, \sigma_1^2)$ (model 1) and $y_{ij} \sim N(0, \sigma_2^2)$ (model 2) and the test statistic is -2 times the difference in the maximum log likelihood: $\Delta_{jk} = -2(\Delta_{2jk} - \Delta_{1jk})$. We derive expressions for the maximum likelihood estimates of the parameters in the model:

$$
\begin{aligned}
\hat{\alpha} &= \frac{\sum_i X_{ik} y_{ij}}{\sum_i X_{ik}^2} = \sum_i X_{ik} y_{ij} \\
\hat{\sigma}_1^2 &= \frac{1}{n} \sum_i (y_{ij} - \hat{y_{ij}})^2 \\
&= \frac{1}{n} \sum_i (y_{ij} - \hat{\alpha} X_{ik})^2 \\
&= \frac{1 - (\sum_i X_{ik} y_{ij})^2}{n} \\
\hat{\sigma}_2^2 &= \frac{1}{n} \sum_i y_{ij}^2 = 1/n
\end{aligned}
$$

Substituting these values into the likelihood functions for the two models we have:

$$
\begin{aligned}
\Delta_{1jk} &= -n/2 \log(2\pi) - n/2 \log(\hat{\sigma}_1^2) - \frac{\sum_i (y_{ij} - \hat{y_{ij}})^2}{2\hat{\sigma}_1^2} \\
&= -n/2 \log(2\pi) - n/2 \log(\frac{1 - (\sum_i X_{ik} y_{ij})^2}{n}) - n/2 \\
\Delta_{2jk} &= -n/2 \log(2\pi) + n/2 \log(n) - n/2 \\
\Rightarrow \Delta_{jk} &= -2(\Delta_{2jk} - \Delta_{1jk}) = -n \log(1 - (\sum_i X_{ik} y_{ij})^2)
\end{aligned}
$$

Now, the maximum LR statistic is:

$$
\Delta_j = \max_k \Delta_{jk} = -n \log(1 - \max_k (\sum_i X_{ik} y_{ij})^2) = -n \log(1 - \lambda_j^2)
$$

This is a strictly increasing function on the domain $0 < \lambda_j < 1$. □

Combining Lemmas A.2, A.4 and A.5 we have our result:

Theorem A.6 *The minimum p values across transcripts, adjusted for multiple testing using the max T approach (Westfall and Young, 1993), produced by the comparison of a Lasso, t and LR test statistic calculated on a set of gene expressions y_{ij} to statistics calculated on a set of permutations of y_{ij}, $\{y_{\sigma_1(i)j}, \ldots, y_{\sigma_M(i)j}\}$, $i = 1, \ldots, n$, are equal.*

A.2 eQTL discovered in RI Rat population.

Table 2: List of eQTL discovered by the Lasso with an FDR of 0.01

Marker	Transcript	Gene symbol
Cd36	1367689_a_at	Cd36
D14Rat77	1367758_at	Afp
D5Rat93	1367965_at	Slc9a1
D10Wox13	1368285_at	Shbg
D13Mit3	1368304_at	Fmo3
D6Cebrp424s2	1368440_at	Slc3a1
D1Rat327	1368526_at	Pex3
D20Rat4	1368597_at	Snf1lk
D19Rat103	1368890_at	Gnpat
D19Utr7	1368891_at	Gnpat
Tnfa	1369110_x_at	RT1-Aw2
D8Rat202	1369171_at	Mst1
D11Cebr87s2	1369202_at	Mx2
D9Rat104	1369313_at	Fhl2
Tnfa	1369667_at	Vps52
D14Rat8	1369784_at	Tpo
D2Rat236	1369866_at	LOC56825
D10Rat145	1369989_at	Pnpo
D19Rat72	1370148_at	Hp
D9Rat19	1370176_at	Trak2
D9Rat93	1370193_at	Ptp4a1
D3Utr6	1370195_at	Snap23
Tnfa	1370290_at	Tubb5
D19Rat14	1370352_at	Cesl1 / Es22 / rCG_44273
Tnfa	1370428_x_at	RT1-A2 / RT1-A3 / RT1-Aw2
Tpm1	1370539_at	Rab8b
D5Rat140	1370597_at	Stx17
D4Rat35	1370806_at	Retsat
Tnfa	1370882_at	Hla-dmb
D10Wox13	1370930_at	-
Tnfa	1371033_at	RT1-Bb
D1Arb17	1371085_at	Ascl3
Tnfa	1371119_at	LOC360231
Tnfa	1371213_at	RT1-A3
D1Rat327	1371324_at	Sf3b5

Continued on next page

Table 2 – continued from previous page

Marker	Transcript	Gene symbol
Scnb2	1371442_at	Hyou1
Cryga	1371494_at	-
D13Rat131	1371732_at	Dpt
D7Rat129	1371749_at	RGD1306001
D10Rat215	1371803_at	Gm2a
D5Rat79	1371879_at	Lrrc42
D12Rat28	1371905_at	MGC94190
D9Rat104	1371951_at	Fhl2
D12Rat36	1372415_at	Gtf2h3
D8Cebr46s6	1372478_at	Cmtm7
D12Rat36	1372532_at	Pitpnm2
D9Rat104	1372646_at	RGD1305645
D2Rat236	1372728_at	Sort1
D1Rat293	1372846_at	Cybasc3
D2Cebr104s1	1373243_at	Pmvk
D2Mit7	1373389_at	Acad9
D4Rat66	1373510_at	Vamp1
D3Cebr26s1	1373537_at	-
Lca	1373661_a_at	Cxcr4
D1Rat212	1373663_at	Dmkn
D12Rat53	1373682_at	Ddx51
D1Cebr103s1	1373833_at	RGD1305713
D2Rat66	1374006_at	Kat3
D17Rat50	1374126_at	-
Cryga	1374196_at	Lancl1
D14Cebrp136s2	1374539_at	Atp10d
D16Utr1	1374553_at	Fam32a
D20Utr2	1374558_at	-
D18Rat55	1374560_at	Prrc1
Kcnj1	1374583_at	-
D1Rat287	1374606_at	-
D11Cebr15s1	1374651_at	Dopey2
D3Cebr204s4	1374653_at	Fam73b
D10Rat145	1374888_at	Ccdc49
Rt6	1374907_at	Arhgef17
D20Utr2	1374916_at	-
D17Mit2	1374959_at	Nqo2

Continued on next page

DOI: 10.2202/1544-6115.1606

Table 2 – continued from previous page

Marker	Transcript	Gene symbol
D3Rat194	1375068_at	Med22
D8Utr5	1375119_at	Nedd4
D14Rat52	1375194_at	-
D18Rat13	1375343_at	-
Rt6	1375516_at	Ndufc2
D9Rat15	1375655_at	-
Scnn1g	1375664_at	Tnrc6a
D3Rat166	1375676_at	-
D3Rat53	1375687_at	-
D4Mgh2	1375724_at	RGD1563612
D14Rat36	1375790_at	-
D1Rat212	1375927_at	-
D12Rat10	1375958_at	Tsc22d4
Cyp11b2	1376021_at	-
Pthlh	1376200_at	MGC72974
D1Rat327	1376249_at	Fuca2
D2Cebr4s8	1376398_at	-
D5Rat144	1376405_at	-
Rt2	1376453_at	-
D12Mit5	1376550_at	-
D5Rat140	1376628_at	Zfp189
D3Rat53	1376796_at	Rab14
D12Cebrp97s9	1376840_at	-
D5Rat140	1376859_at	-
D3Rat53	1377007_at	-
D3Mit3	1377040_a_at	-
D14Utr1	1377200_at	-
D12Cebr4s3	1377212_at	-
D12Cebrp97s9	1377244_at	Zkscan5
D6Cebrp97s14	1377329_at	-
Tnfa	1377334_at	RT1-Ba
D8Cebr16s5	1377452_at	-
D8Rat68	1377501_at	Zfp317
D16Utr1	1377959_at	Fam32a
D18Rat13	1380293_at	LOC361346
D2Rat236	1380404_at	-
D14Rat37	1382667_at	Mfsd10

Continued on next page

Table 2 – continued from previous page

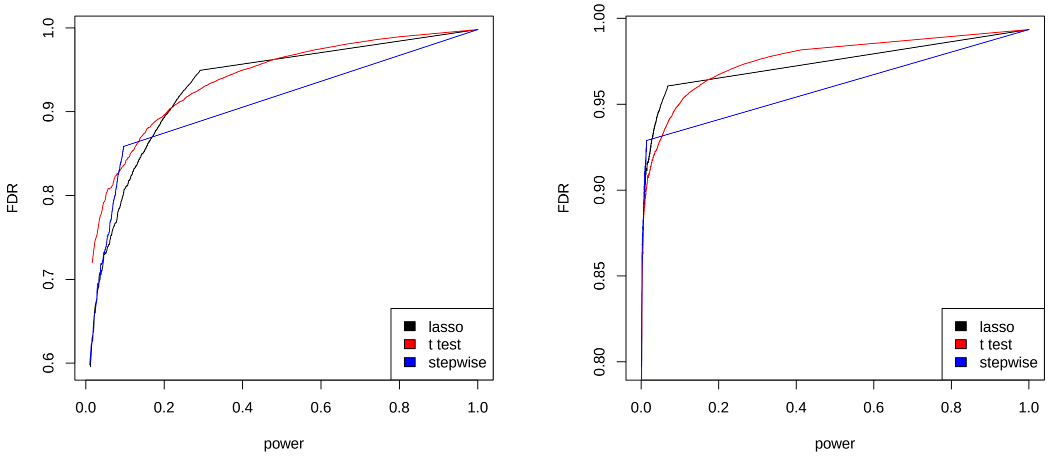
Marker	Transcript	Gene symbol
D3Rat35	1383440_at	Accs / LOC690470
D9Rat156	1384309_at	-
D7Rat17	1384717_at	-
D3Rat53	1385314_at	-
Cd36	1386901_at	Cd36
Pai1	1386936_at	Grifin
D2Rat201	1387144_at	Itga1
D6Cebrp424s2	1387222_at	Cript
D8Rat68	1387366_at	Ilf3
D9Rat93	1387376_at	Aox1
D13Utr5	1387671_at	Sctr
D11Cebr87s2	1387789_at	Erg
D15Rat6	1387808_at	Slc7a7
D3Rat132	1387906_a_at	Gnas
Tnfa	1388202_at	RT1-Aw2
Cd36	1388223_at	Gnat3
Tnfa	1388236_x_at	RT1-CE12
D8Rat68	1388366_at	Mrpl4
D7Cebr77s1	1388437_at	-
D16Utr1	1388508_at	Fam32a
D17Ucsf2	1388603_a_at	Isca1
D17Rat144	1388617_at	Bphl
D18Rat13	1388630_at	-
D15Rat6	1388654_at	Mrpl52
Tnfa	1388694_at	H2-T24
D3Cebr204s4	1388912_at	Rexo4
Lca	1389244_x_at	Cxcr4
D7Rat129	1389264_at	Ankrd54
Cacna1s	1389293_at	Cpsf2
D15Cebr7s13	1389352_at	-
Bzrp	1389511_s_at	-
D1Rat42	1389650_at	-
Tnfa	1389734_x_at	H2-T24
D16Utr1	1389793_at	-
D3Cebr204s4	1389816_at	Endog
D17Rat62	1389867_at	-
Rt2	1390012_at	Man2b1

Continued on next page

Table 2 – continued from previous page

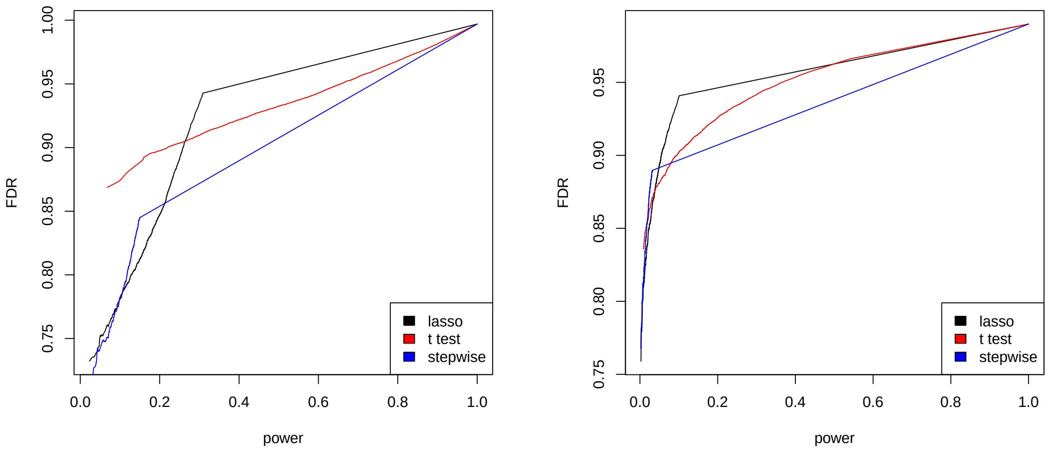
Marker	Transcript	Gene symbol
D18Rat61	1390128_at	Chmp1b
D11Cebr15s1	1390154_at	LOC686326
Kcnj1	1390185_at	Dcps
D11Mit2	1390364_at	-
D16Mit2	1390435_at	Snhg8
D11Rat20	1390443_at	RGD1563888
Tnfa	1390562_s_at	-
D13Rat75	1390722_at	-
D7Mit17	1392720_at	Cyp4f17
Tnfa	1394386_s_at	Vps52
D12Cebr1s1	1398960_at	Cct6a / LOC316484 / LOC688183
D16Utr1	1399080_at	LOC688495
D1Rat27	1399157_at	RGD1310358

A.3 Extended graphs for simulation studies



(a) Genetic Model A (b) Genetic Model B

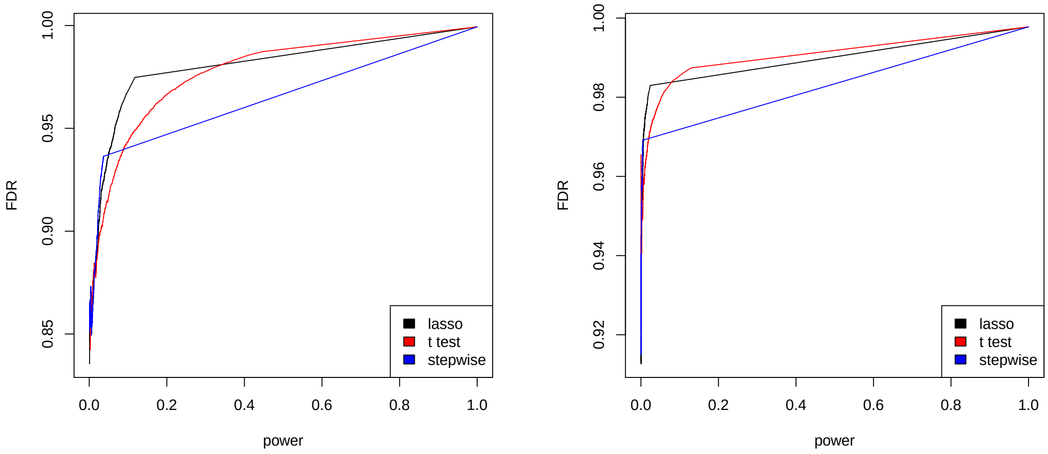
Figure 8: We plot the proportion of genuine eQTL discovered against the proportion of false positives at various thresholds when the three methods are applied to the datasets simulated from the RI rat data presented in Hübner et al. (2005) under genetic Model A (Figure (a)) and Model B (Figure (b)) as the threshold is altered from declaring no candidate eQTL to declaring all markers as eQTL..



(a) Genetic Model A

(b) Genetic Model B

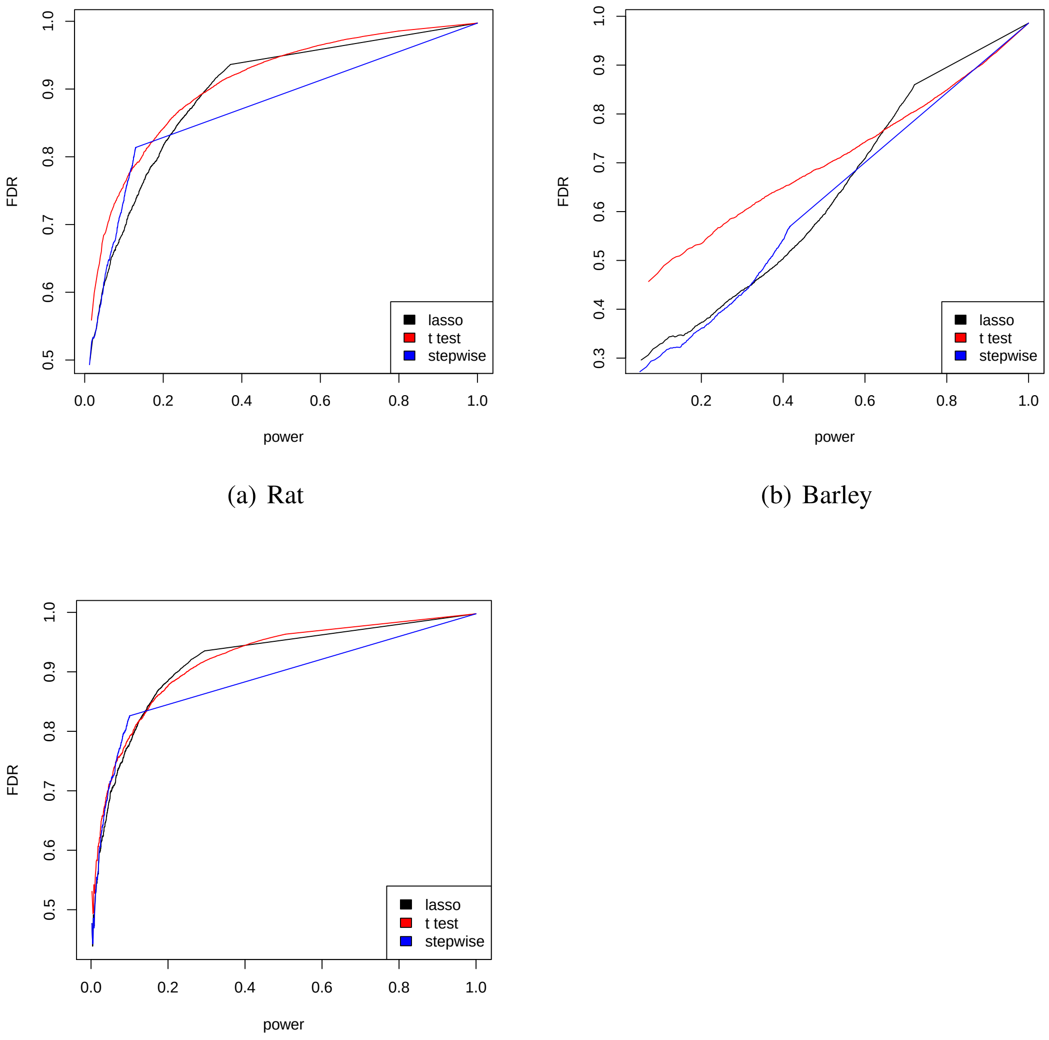
Figure 9: We repeat the simulation studies presented in Figure 8, this time gene expression profiles are simulated from the barley genotype data from Kleinhofs et al. (1993).



(a) Genetic Model A

(b) Genetic Model B

Figure 10: We repeat the simulation studies presented in Figure 8, this time gene expression profiles are simulated from the RI mouse genotype data from Taylor et al. (1999).



(a) Rat

(b) Barley

(c) Mouse

Figure 11: We repeat the simulation studies presented in Figures 8(a), 9(a) and 10(a) on a subset of tag markers from each dataset.

References

Aitman, T. J., J. K. Critser, E. Cuppen, A. Dominiczak, X. M. Fernandez-Suarez, J. Flint, D. Gauguier, A. M. Geurts, M. Gould, P. C. Harris, R. Holmdahl, N. Hubner, Z. Izsvk, H. J. Jacob, T. Kuramoto, A. E. Kwitek, A. Marrone, T. Mashimo, C. Moreno, J. Mullins, L. Mullins, T. Olsson, M. Pravenec, L. Riley, K. Saar, T. Serikawa, J. D. Shull, C. Szpirer, S. N. Twigger, B. Voigt, and K. Worley (2008): "Progress and prospects in rat genetics: a community view." *Nature Genetics*, 40, 516–522.

Andrés, A. M., I. H. Tejedor, and A. S. Mato (1995): "The Wilcoxon, Spearman, Fisher, χ^2, Student and Pearson tests and 2 x 2 tables," *The Statistician*, 44, 441–450.

Belknap, J. K. (1998): "Effect of within-strain sample size on QTL detection and mapping using recombinant inbred mouse strains." *Behavior Genetics*, 28, 29–38.

Benjamini, Y. and Y. Hochberg (1995): "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society, Series B*, 57, 289–300.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009): "Simultaneous analysis of lasso and dantzig selector," *Annals of Statistics*, 37, 1705–1732.

Bottger, A., E. Lankhorst, H. A. van Lith, L. F. van Zutphen, V. Zdek, A. Musilov, M. Simkov, R. Poledne, V. Bl, V. Ken, and M. Pravenec (1998): "A genetic and correlation analysis of liver cholesterol concentration in rat recombinant inbred strains fed a high cholesterol diet." *Biochemical and Biophysical Research Communications*, 246, 272–275.

Bottolo, L. and S. Richardson (2010): "Evolutionary Stochastic Search for Bayesian model exploration," *Bayesian Analysis*, 5, 583–618.

Breiman, L. (1995): "Better subset regression using the nonnegative garrote," *Technometrics*, 37, 373–384.

Candes, E. and T. Tao (2007): "The dantzig selector: Statistical estimation when *p* is much larger than *n*," *Annals of Statistics*, 35, 2313–2351.

Casola, A., R. P. Garofalo, H. Haeberle, T. F. Elliott, R. Lin, M. Jamaluddin, and A. R. Brasier (2001): "Multiple cis regulatory elements control RANTES promoter activity in alveolar epithelial cells infected with respiratory syncytial virus," *Journal of Virology*, 75, 6428–6439.

Chapman, J. M., J. D. Cooper, J. A. Todd, and D. G. Clayton (2003): "Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power," *Human Heredity*, 56, 18–31.

Chun, H. and S. Kele (2009): "Expression quantitative trait loci mapping with multivariate sparse partial least squares regression," *Genetics*, 182, 79–90.

de Bakker, P. I., R. Yelensky, I. Pe’er, S. B. Gabriel, M. J. Daly, and D. Altshuler (2005): “Efficiency and power in genetic association studies,” *Nature Genetics*, 37, 1217–1223.

Doerge, R. W. and G. A. Churchill (1996): “Permutation tests for multiple loci affecting a quantitative character,” *Genetics*, 142, 285–294.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004): “Least angle regression,” *Annals of Statistics*, 32, 407–499.

Fan, J. and R. Li (2001): “Variable selection via nonconcave penalized likelihood and its oracle properties,” *Journal of the American Statistical Association*, 96, 1348–1360.

Hübner, N., C. A. Wallace, H. Zimdahl, E. Petretto, H. Schulz, F. Maciver, M. Mueller, O. Hummel, J. Monti, V. Zidek, A. Musilova, V. Kren, H. Causton, L. Game, G. Born, S. Schmidt, A. Mller, S. A. Cook, T. W. Kurtz, J. Whittaker, M. Pravenec, and T. J. Aitman (2005): “Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease.” *Nature Genetics*, 37, 243–253.

James, G., P. Radchenko, and J. Lv (2009): “DASSO: connections between the Dantzig selector and lasso,” *J. Roy. Statist. Soc. Ser. B*, 71, 127–142.

Jia, Z. and S. Xu (2007): “Mapping quantitative trait loci for expression abundance.” *Genetics*, 176, 611–623.

Kendziorski, C. M., M. Chen, M. Yuan, H. Lan, and A. D. Attie (2006): “Statistical methods for expression quantitative trait loci (eQTL) mapping.” *Biometrics*, 62, 19–27.

Kleinhofs, A., A. Kilian, M. Saghai Maroof, R. Biyashev, P. Hayes, F. Chen, N. Lapitan, A. Fenwick, T. Blake, V. Kanazin, et al. (1993): “A molecular, isozyme and morphological map of the barley (Hordeum vulgare) genome,” *Theoretical and Applied Genetics*, 86, 705–712.

Lander, E. S. and D. Botstein (1989): “Mapping mendelian factors underlying quantitative traits using RFLP linkage maps.” *Genetics*, 121, 185–199.

Litvin, O., H. C. Causton, B.-J. Chen, and D. Pe’er (2009): “Modularity and interactions in the genetics of gene expression.” *Proceedings of the National Academy of Sciences*, 106, 6441–6446.

Manolio, T., F. Collins, N. Cox, D. Goldstein, L. Hindorff, D. Hunter, M. McCarthy, E. Ramos, L. Cardon, A. Chakravarti, et al. (2009): “Finding the missing heritability of complex diseases,” *Nature*, 461, 747–753.

Meinshausen, N. and P. Bühlmann (2006): “High-dimensional graphs and variable selection with the lasso,” *Annals of Statistics*, 34, 1436–1462.

Orozco, G., A. Hinks, S. Eyre, X. Ke, L. Gibbons, J. Bowes, E. Flynn, P. Martin, et al. (2009): “Combined effects of three independent SNPs greatly increase the risk estimate for RA at 6q23,” *Human Molecular Genetics*, 18, 2693.

Pan, W. (2009): "Network-based multiple locus linkage analysis of expression traits." *Bioinformatics*, 25, 1390–1396.

Petretto, E., J. Mangion, N. J. Dickens, S. A. Cook, M. K. Kumaran, H. Lu, J. Fischer, H. Maatz, V. Kren, M. Pravenec, N. Hubner, and T. J. Aitman (2006): "Heritability and tissue specificity of expression quantitative trait loci." *PLoS Genetics*, 2, e172.

Stephens, D. A. and R. D. Fisch (1998): "Bayesian analysis of quantitative trait locus data using reversible jump markov chain monte carlo," *Biometrics*, 54, 1334–1347.

Storey, J. D. (2002): "A direct approach to false discovery rates," *Journal of the Royal Statistical Society, Series B*, 64, 479–498.

Storey, J. D. (2003): "The positive false discovery rate: A Bayesian interpretation and the q-value," *The Annals of Statistics*, 31, 2013–2035.

Taylor, B. A., C. Wnek, B. S. Kotlus, N. Roemer, T. MacTaggart, and S. J. Phillips (1999): "Genotyping new BXD recombinant inbred mouse strains and comparison of BXD and consensus maps," *Mamm. Genome*, 10, 335–348.

Tibshirani, R. (1996): "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society*, 58, 267–288.

Valdés-Sosa, P., J. Sánchez-Bornot, A. Lage-Castellanos, M. Vega-Hernández, J. Bosch-Bayard, L. Melie-García, and E. Canales-Rodríguez (2005): "Estimating brain functional connectivity with sparse multivariate autoregression," *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360, 969.

Westfall, P. H. and S. Young (1993): *Resampling-based Multiple Testing*, Wiley.

Xiaohong, C. and X. Shizhong (2010): "Significance test and genome selection in Bayesian shrinkage analysis," *International Journal of Plant Genomics*, 2010.

Yi, N., V. George, and D. B. Allison (2003): "Stochastic search variable selection for identifying multiple quantitative trait loci." *Genetics*, 164, 1129–1138.

Zhang, H. and W. Lu (2007): "Adaptive Lasso for Cox's proportional hazards model," *Biometrika*, 94, 691–704.

Zhang, W., J. Zhu, E. E. Schadt, and J. S. Liu (2010): "A bayesian partition method for detecting pleiotropic and epistatic eqtl modules." *PLoS Computational Biology*, 6, e1000642.

Zou, F., J. Fine, J. Hu, and D. Lin (2004): "An efficient resampling method for assessing genome-wide statistical significance in mapping quantitative trait loci," *Genetics*, 168, 2307.

Zou, F., Z. Xu, and T. Vision (2006): "Assessing the significance of quantitative trait loci in replicable mapping populations." *Genetics*, 174, 1063–1068.

Zou, H. and H. Hastie (2003): "Regression shrinkage and selection via the elastic net, with applications to microarrays," Technical report, Stanford University, URL `http://www-stat.stanford.edu/ hastie/pub.htm`.