

Gene expression

Methodological aspects of the genetic dissection of gene expression

Ö. Carlborg¹, D. J. De Koning¹, K. F. Manly², E. Chesler², R. W. Williams² and C. S. Haley^{1,*}¹Roslin Institute, Roslin, Midlothian EH25 9 PS, UK and ²University of Tennessee Health Science Center, 855 Monroe Avenue, Memphis, TN 38163, USA

Received on November 16, 2004; accepted on December 17, 2004

Advance Access publication December 21, 2004

ABSTRACT

Motivation: Dissection of the genetics underlying gene expression utilizes techniques from microarray analyses as well as quantitative trait loci (QTL) mapping. Available QTL mapping methods are not tailored for the highly automated analyses required to deal with the thousands of gene transcripts encountered in the mapping of QTL affecting gene expression (sometimes referred to as eQTL). This report focuses on the adaptation of QTL mapping methodology to perform automated mapping of QTL affecting gene expression.

Results: The analyses of expression data on >12 000 gene transcripts in BXD recombinant inbred mice found, on average, 629 QTL exceeding the genome-wide 5% threshold. Using additional information on trait repeatabilities and QTL location, 168 of these were classified as 'high confidence' QTL. Current sample sizes of genetical genomics studies make it possible to detect a reasonable number of QTL using simple genetic models, but considerably larger studies are needed to evaluate more complex genetic models. After extensive analyses of real data and additional simulated data (altogether >300 000 genome scans) we make the following recommendations for detection of QTL for gene expression: (1) For populations with an unbalanced number of replicates on each genotype, weighted least squares should be preferred above ordinary least squares. Weights can be based on the repeatability of the trait and the number of replicates. (2) A genome scan based on multiple marker information but analysing only at marker locations is a good approximation to a full interval mapping procedure. (3) Significance testing should be based on empirical genome-wide significance thresholds that are derived for each trait separately. (4) The significant QTL can be separated into high and low confidence QTL using a false discovery rate that incorporates prior information such as transcript repeatabilities and co-localization of gene-transcripts and QTL. (5) Including observations on the founder lines in the QTL analysis should be avoided as it inflates the test statistic and increases the Type I error. (6) To increase the computational efficiency of the study, use of parallel computing is advised. These recommendations are summarized in a possible strategy for mapping of QTL in a least squares framework.

Availability: The software used for this study is available on request from the authors.

Contact: Chris.Haley@bbsrc.ac.uk

1 INTRODUCTION

With the emergence of genome-wide gene expression arrays in the late 1990s it has become possible to consider genome-wide studies aimed at dissecting the genetic regulation of gene expression. Jansen and Nap (2001) published the formal description of this new research area and coined it genetical genomics. The concept is based on a segregating population where for each individual the level of mRNA transcript abundance is measured for a large number of genes and genome-wide genotypes are collected. The expression levels for the individual genes are measured on a continuous scale and can be treated as a quantitative phenotype affected by multiple genes and environmental factors. By combining these quantitative phenotypes and the genetic marker data, quantitative trait loci (QTL) mapping algorithms can be used to dissect the transcriptional regulation for the entire transcriptome and identify the effects of some of the individual QTL affecting gene expression. Klose *et al.* (2002) described the first experimental genetical genomics results about the genetic regulation of the mouse brain proteome. Subsequently, Brem *et al.* (2002) described the genetics of gene expression in budding yeast followed by Schadt *et al.* (2003) who reported on genetical genomics in maize, mouse and man.

QTL mapping algorithms have been described for many types of experimental crosses and natural populations. In the genetical genomics studies performed so far, standard QTL mapping packages, such as MapMaker QTL (Lander *et al.*, 1987), QTL Cartographer (Wang *et al.*, 2001–2003, <http://statgen.ncsu.edu/qtlcart/WQTLCart.htm>) and WebQTL (Wang *et al.*, 2003), have been used. For future studies, the QTL mapping algorithms need to be reviewed in the light of the unique properties of mapping QTL affecting gene expression. Such mapping studies need to be largely automated because they involve analyses of potentially many thousands of gene transcripts. In traditional QTL mapping studies, the number of traits rarely exceeds a hundred. Due to this huge increase in number of phenotypes, the computational aspects of the algorithms that are employed also need to be carefully considered. The process of automating the analyses is, however, not trivial.

An automated QTL mapping strategy needs to rely strictly on statistical measures to highlight candidate regions because manual inspection of QTL results across the genome for individual traits, which is common in standard QTL mapping, is not feasible for every individual gene transcript. In this study, we will apply various

*To whom correspondence should be addressed.

standard QTL mapping scenarios to analyse data from one of the first publicly available genetical genomics datasets (Chesler *et al.*, 2005). Based on these results, we will outline a potential automated strategy for detecting QTL affecting gene expression. This report focuses on the use of gene-expression data as a quantitative phenotype for QTL analysis rather than on the extraction and normalization of gene expression data.

2 SYSTEMS AND METHODS

In this report we have evaluated a wide range of components of a streamlined and semi-automated QTL mapping strategy. This strategy allows high throughput of thousands of phenotypes, in parallel on supercomputers if available, and the results are robust enough to use directly for post-QTL analyses aiming to clarify the biological relevance of the QTL. For this study we have used an experimental recombinant inbred (RI) mouse dataset as well as some limited simulated data to assess specific components of the methodology.

2.1 The BXD mouse data

We have based the bulk of our study on a mouse model system, where mRNA expression levels and genome-wide genotypes were collected in a population of BXD RI mice that are derived from a cross between C57BL/6J (B) and DBA/2J (D). The results of this study are reported in depth by Chesler *et al.* (2005). Expression profiles were collected for 12 422 transcripts using Affymetrix U74Av2 chips for 78 pools from 29 recombinant inbred lines (RIL). The transcript data were processed with the Affymetrix M.A.S. 5 software. Expression was measured in one to four replicates for each line and each replicate contained pooled tissue samples from three individual mice of the same sex. For additional information on the experimental procedures see the description at <http://www.webqtl.org/search.html> for the data set UTHSC Brain mRNA U74Av2 (May03) MAS5.

2.2 Components of an automated QTL mapping strategy

In this section we outline analytical and inferential aspects of QTL mapping for gene-expression data. Subsequently some aspects of the analysis were varied and the entire gene-expression data set was reanalysed for QTL. This resulted in 23 scenarios, each containing a unique combination of analytical components, totalling more than 300 000 genome scans (with additional simulations).

We used a range of statistics to compare the results for these scenarios. For simplicity, we assume that, for a given trait, all significant QTL on a chromosome represent the same QTL when comparing different scenarios.

First to compare the statistical power of different scenarios, we estimated pairwise correlations between the scenarios based on (1) the traits with significant QTL and (2) the number of significant QTL for each trait. Second, we tested the degree of similarity between the QTL curves obtained by the different methods by estimating the pairwise correlations for the highest obtained *F*-statistics as well as the point estimates for the location of the QTL.

2.2.1 The QTL model The QTL analyses were based on modifications of the least squares QTL mapping approach of Haley and Knott (1992). In short, the Haley and Knott algorithm contains the following steps. First, QTL genotype probabilities are estimated conditional on marker genotypes at selected grid points in the genome. Then, regression indicator variables are calculated for the genetic effect(s) of the QTL using the QTL genotype probabilities. These indicator variables are used in a least squares framework to estimate the genetic effect(s) of the QTL. In the RI population, the marker genotypes were used to estimate the probability (P_{BB} and P_{DD}) for every RIL being each of the two possible QTL genotypes (*BB* and *DD*) at the grid nodes along the genome. A putative QTL with a marginal (additive) effect is modelled at every grid location:

$$y_i = \beta_0 + FZ + \beta_j a_j + \varepsilon_j, \quad (1)$$

where y_i is a vector of phenotypes, β_0 is the mean, F is a vector of additive effects for earlier detected QTL, Z is a vector of regression variables for earlier detected QTL, β_j is the regression coefficient for additive effect for a putative QTL at genomic location j , a_j is the regression indicator variable for the additive effect of QTL k at genomic location j ($a_j = P_{BBj} - P_{DDj}$) and ε_j is the residual error.

2.2.2 Assessment of phenotypic data and power of the experiment

Phenotypic outliers. Individual observations with phenotypic values that are outliers in the distribution can be influential on the outcome of a statistical analysis. Good statistical practice dictates that outliers are identified and dealt with before any further analyses. In the BXD data, 7743 (0.8%) of the observations could be classified as distributional outliers (>3 phenotypic SD from the mean). We compared the effect of omitting outliers versus shrinking the phenotypic value of outliers back to 3 SD from the mean versus retaining them at their original values in the analyses, but no clear 'best strategy' emerged from these comparisons (data not shown).

Estimating the repeatabilities of the gene transcripts. The level of similarity between repeated measurements of gene expression in samples with the same genotype (e.g. from the same RIL) is an indication of the balance between genetic control and environmental and technical error. The repeatability (r) is the ratio between line variance and the total variance and is also the upper limit of the heritability of the transcript of interest (Falconer and Mackay 1996):

$$r = \frac{\sigma_b^2}{(\sigma_w^2 + \sigma_b^2)}$$

where σ_b^2 is the variance between BXD lines and σ_w^2 is the variance within BXD lines. The repeatabilities for each transcript were estimated using a mixed linear model in GENSTAT with BXD line as a random effect and sex as a fixed effect. Age at sampling was evaluated but not included as no consistent effect for this effect was observed.

Statistical power to detect QTL. The power to detect QTL for a given experiment depends on the size of the experiment, the size of the QTL effect, the variability of the trait and the threshold that is used to claim a significant QTL. While the distribution of the QTL effects is largely unknown (especially when it comes to QTL for gene expression), we can predict the statistical power to detect a given QTL effect (Lynch and Walsh, 1998). We adapted the prediction for an F_2 population (Lynch and Walsh, 1998) to RI lines by taking the homozygosity of RI lines into account. For the BXD design we have calculated the projected power of detection for various standardized QTL effects (0.25, 0.4, 0.5, 0.6 and 0.75 phenotypic SD). We calculated the projected power for 1, 3 and 10 replicates per BXD line, because it was possible for RI lines to do multiple measurements in the same genotype.

2.2.3 Use of expression data in QTL analysis For populations with multiple observations on each experimental unit, e.g. the genotypes for the BXDRIL, there are several options for parameter estimation in QTL mapping. We have evaluated alternative regression methods which use the information from the 78 individual measurements in different ways.

Ordinary least squares. First the raw phenotypic means of each RIL (i.e. 29 values) were used as phenotypes in ordinary least squares interval mapping. In our example, each phenotype in the QTL analysis is obtained by averaging the gene-expression level over several repeated measurements. This analysis does not account for the fact that the number of replicates varies from one per line to four per line.

Weighted least squares. Weighted regression can be used to account for the differences in number of replicates between different lines. We have evaluated weighted regression using two alternative weights. The first weight evaluated was the inverse of the estimated within-line variance. In this case lines exhibiting small within line-variances get large weights and vice versa. Due to the small number of replicates, the within line variances were poorly estimated in the present study. As a second option, the weights were based on the repeatability for the trait (r) and the number of measurements (n)

for each RIL as $w = n/[1 + r(n - 1)]$. This weight is proportional to the expected reduction in variance for multiple observations rather than the estimated variance for each line from the actual data.

Use of Parental and F₁ observations in QTL analyses. Jansen and Stam (1994) suggest the use of phenotypic information on parental lines and F₁ to increase power of QTL detection in a multiple QTL mapping analysis. This was accommodated in WebQTL until June 2004 and has been used in a number of studies (e.g. Risinger *et al.*, 2000; Grisel *et al.*, 2002). By including parental and F₁ data in the QTL analysis for RIL, an increase in the test statistic and hence the significance level is expected due to an increased number of degrees of freedom in the statistical tests. However, these observations cannot be treated as ordinary RIL. The full phenotypic difference between the parental lines will be associated with any marker discriminating between the two lines, not just the effect of any QTL linked to the marker. The standard significance tests used in QTL mapping (e.g. randomization tests or multiple testing corrected analytical thresholds) do not account for this. We included the expression data from F₁ and parental lines (15 additional measurements) for the 50 best QTL and compared the test-statistics to the analysis based on BXD data only. To evaluate the effect of including parental lines and F₁ in the absence of QTL effects, we also simulated 100 replicates of 35 lines for each of the three following scenarios: (1) a population of 35 RIL with no QTL segregating (all phenotypes sampled from a normal distribution with mean 0 and SD 4). (2) A population with 32 RIL sampled from the same distribution as in Equation (1), with, in addition, two parental lines, on average, one phenotypic SD difference between them, and an F₁ line with mean 0. (3) As in (2) but with two phenotypic SD difference between the two parental lines. Genotypes of the existing BXD lines were used throughout, with founder lines and F₁ being *BB*, *DD* and *BD*, for all loci, respectively. We calculated the increase in the test-statistic and the significance threshold for all three scenarios to evaluate the effect of line differences not specifically associated with any individual marker on the significance test for QTL.

2.2.4 Density of the genetic grid in QTL analysis The computational demand of QTL mapping can be decreased by using a sparser genetic grid for a genome scan. Most of the currently used QTL mapping strategies are based on interval mapping where QTL are evaluated at regular intervals (e.g. 1 cM) on the genetic map. In a situation where markers are fully informative Coffman *et al.* (2003) suggest that a genome scan using single marker information can be equally or even more powerful than analyses based on flanking markers. We evaluated three alternatives. The first grid was based on genetic marker locations (single marker mapping) where at the marker positions the BXD lines were classified as either of the two possible homozygote genotypes (*BB* or *DD*). Lines with missing marker genotypes were given uninformative status. The second grid was the same density and also based on genetic marker locations, but markers with missing data were assigned the interval mapping genotypic probabilities calculated from information on adjacent markers (Haley and Knott, 1992). This was termed marker position mapping. The third grid was based on the genetic map (interval mapping). The probability of each line being either of the two genotypes was estimated at 1 cM intervals throughout the genome (Haley and Knott, 1992). Following Lynch and Walsh (1998) the expected map expansion in an RIL was accounted for.

2.2.5 Significance testing

Randomization testing. Randomization testing is widely used in QTL mapping to obtain empirical significance thresholds for QTL detection. However, the method increases the computational demand of the analyses by at least a factor of 1000 (i.e. the number of permutations used). It is therefore tempting not to derive trait specific significance thresholds, but instead use analytical significance thresholds or a universal randomization threshold (an average threshold from randomization tests for a subset of traits in the study) in QTL mapping (Schadt *et al.*, 2003). In the present study, randomization testing ($n = 1000$) was used to derive trait specific genome-wide

significance thresholds (Churchill and Doerge, 1994; Doerge and Churchill, 1996).

2.2.6 Multiple testing and post-hoc inferences Performing genome scans for all transcripts that are represented on a microarray introduces two levels of multiple testing. A total of 12442 gene transcripts are tested for association with markers covering the entire mouse genome. We choose to control the genome-wide error rate by permutation tests as described above and explore alternative strategies to deal with the large number of transcripts evaluated.

False Discovery Rate (FDR). A competing approach to the traditional statistical approach of introducing penalties to account for multiple testing (e.g. using Bonferroni corrections) are methods based on controlling the false discovery rate (FDR) (Hochberg and Benjamini, 1990). Instead of trying to control the Type I error rate of the study by increasing the significance threshold for each result, this method is aimed at estimating the proportion of the significant QTL that are likely to be Type I errors. The FDR is defined as the proportion of results where the null hypothesis is true among those results that were called significant. We used genome-wide empirical *P*-values to estimate the FDR across all QTL transcripts or strategically chosen subsets of data. Applications of the FDR often assume a priori that all tests represent true null hypotheses. We also evaluated an alternative technique for FDR calculation that includes empirical estimation of the proportion of true null hypotheses using the QVALUE package (<http://faculty.washington.edu/~jstorey/qvalue/index.html>) (Storey and Tibshirani 2003).

The effect of more stringent thresholds on the FDR. While 1000 permutations are sufficient to obtain a reasonable estimate of the 5% genome-wide thresholds, additional permutation is needed to estimate smaller *P*-values. For the 256 most significant QTL, we performed up to 1 000 000 permutations to obtain more precise estimates of their genome-wide *P*-values and reevaluated the FDR for these gene transcripts.

Censoring based on the repeatability of the data. As outlined, the repeatability is a measure of the degree of genetic control of a given transcript level. As such, the repeatability could be used to identify genes that have a higher a priori expectation for detection of QTL. Focussing the analyses on only those transcripts that exceed a certain repeatability not only decreases the overall multiple testing problem, but will also give a set of transcripts that should be enriched for QTL compared to the whole set of transcripts. In this study, we analysed all gene transcripts for eQTL and demonstrated the potential effects of censoring by repeatability on the FDR retrospectively. We verified by simulation that variation in the repeatability in the absence of any QTL did not affect the distribution of the test statistic (data not shown). Therefore, we do not expect the estimation of the FDR to be affected by this type of censoring.

Use of location information in evaluating putative QTL. When mapping QTL affecting gene expression, the location of the mapped transcript is often known a priori. The probability of spuriously finding a QTL mapping to the location of the gene transcript (*cis*-acting QTL: the QTL maps to the same genomic bin as the transcript) can be postulated to be:

$$P_{cis} = \frac{\alpha}{nb},$$

where α is the genome-wide significance threshold and nb is the number of genomic bins. The size and number of genomic bins depends on which confidence interval is used for the QTL. If the abundance of *cis*-acting QTL is greater than that expected by chance, this additional source of information can be used to increase the number of significant *cis*-acting QTL. Across the 12422 traits in our model system, about 17 *cis*-acting QTL are expected by chance when the definition of a *cis*-acting QTL is based on a 40 cM genomic bin, up to 20 cM either side of the transcript location, where the transcript location is ($P_{cis} = 0.05/37 \approx 0.0014$). When 20 and 10 cM windows are used, the expected number reduces to 10 and 4, respectively. This prior expectation can be used to calculate a FDR of *cis*-acting QTL in the genome scan.

3 ALGORITHM

3.1 Genomic search algorithms

A genome scan involves fitting a statistical model at multiple locations in the genomic grid with the objective of finding the location(s) in the genome with significant statistical support for one QTL or several QTL. The mapping procedure can be multidimensional when searching for multiple QTL. Two search algorithms were used to select the locations for the QTL.

3.1.1 Forward selection We used a forward selection search algorithm to reduce the multidimensional search for the QTL to a series of one-dimensional (1D) searches for marginal effects of individual QTL. The most significant QTL from a series of successive 1D genome scans are sequentially added to the multiple QTL model. Good performance is expected when the QTL are independent (i.e. non-interacting and non-linked) and the algorithm has been widely used for this purpose previously. To be included in the model, a QTL needed to exceed a 5% genome-wide significance threshold as derived from permutation (Doerge and Churchill, 1996). For multiple linked QTL, we imposed the restriction that at least one marker interval between the QTL should be below the significance threshold.

3.1.2 Exhaustive search An exhaustive (enumerative) search involves fitting the statistical model at all nodes in the (1D or multidimensional) grid. The best location in the grid, at the given resolution, will be found when all locations are evaluated, but at a high computational cost (>1500 tests for a genome scan at 1 cM intervals). We used an exhaustive search in the repeated 1D genome searches for QTL with marginal effects included in the forward selection procedure described in Section 3.1.1. Although the analytical software was equipped to perform two-dimensional (2D) genome scans for epistatic QTL (Carlborg, 2002), the current experiment was too small to detect epistasis and we will only report on the analyses for main (additive) effects.

3.2 Parallel algorithm for QTL mapping

QTL mapping is very suitable for parallel computing (Carlborg *et al.*, 2001; Carlborg, 2002) and substantial reduction in elapsed time for computations can be obtained by parallelization of genome scans and randomization tests. When large numbers of traits are analysed, e.g. in QTL mapping using expression data, parallelization can be efficiently implemented across traits as well. Here, we used a parallel algorithm where input of data is done in a serial part of the code. The analyses of individual traits are then distributed to n individual processors (up to 512 for the present analyses) and when the analyses are complete, the results are collected from the individual processors to a master node, which writes the output to disk. Output can be restricted to traits for which QTL exceeding a preidentified threshold have been detected.

4 IMPLEMENTATION

The analysis program used for these analyses has been developed by the authors (Carlborg, 2002), written in Fortran90 and adapted for parallel computation using MPI. The software is available from the authors on request. The code is highly optimized and the parallel algorithm used gives a nearly linear speedup for the 256–512 processors used in these analyses. The analyses were performed on a 512 processor SGI Origin 3000 at CSAR, Manchester, UK.

5 RESULTS

The results will be presented in three parts. In the first section we will describe two properties of the BXD data: the power to detect QTL for this experiment and the repeatabilities of the gene transcripts. In the second section we will focus on comparisons between different scenarios and computational aspects of the analysis. In the third section we will elaborate on various strategies to improve the FDR of the BXD experiment.

5.1 Distribution of repeatabilities for the BXD phenotypes and statistical power to detect eQTL

The distribution of the repeatabilities for the 12 422 expression profiles showed that approximately half of the transcripts had repeatabilities <5% and only a small proportion (<1%) had repeatabilities >45% (Fig. 1). The observations with repeatability >0.35 ($n = 390$) were treated separately with regard to statistical inferences (FDR) and comparisons between methods as the high repeatability (HR) data set. The statistical power to detect QTL with 29 BXD lines is presented in Table 1 for a trait where 50% of the variance is attributed to BXD line and 50% to environmental and technical noise. For the present experiment with, on average, 3 replicates per line, there was reasonable power to detect QTL with an effect >0.5 phenotypic SD. For smaller QTL effects, there was little power of detection, even if the number of replicates per RIL was increased to 10 (Table 1).

5.2 Analytical components of the QTL mapping strategy

All results reported are, unless otherwise stated, the marginal effect of a specific component of the QTL mapping strategy across all other components. To prevent the potentially meaningless comparison between many false positive results, we focus on the results for the HR dataset which we expect to be enriched for true QTL.

5.2.1 Selection of the genetic grid for QTL analysis There was a high degree of similarity between the results from using all three genetic grids using the correlation measures (correlations 0.93–1.00). There were high correlations between the highest F -statistics (0.94–0.95) for interval mapping and marker position mapping. The average number of QTL detected by these methods was also very similar (92 versus 91 in the HR dataset). The individual scenarios indicate that in some cases interval mapping detects more QTL and in others marker position mapping. The set of detected QTL between scenarios was fairly similar (correlation 0.92). On average, the interval mapping procedure gave higher maximum F -statistics than the single marker and marker position mapping, but this was counteracted by a tendency for higher thresholds for interval mapping.

5.2.2 Statistical method for parameter estimation Preliminary analysis using inverse variance weighted least squares for parameter estimation gave markedly different results than all other analysis methods in the test data. This is probably due to the wide variation in weights due to the poor estimation of within-line variances. Due to this, the method was not evaluated further using the full dataset. In the HR dataset, there was a very high correlation between the set of QTL identified by using ordinary least squares on raw means or repeatability weighted least squares (0.98). The agreement between the maximum test statistics was also very high (1.00). The correlation between the estimates of QTL locations was 0.90. This strong correlation between results from using weighted least squares

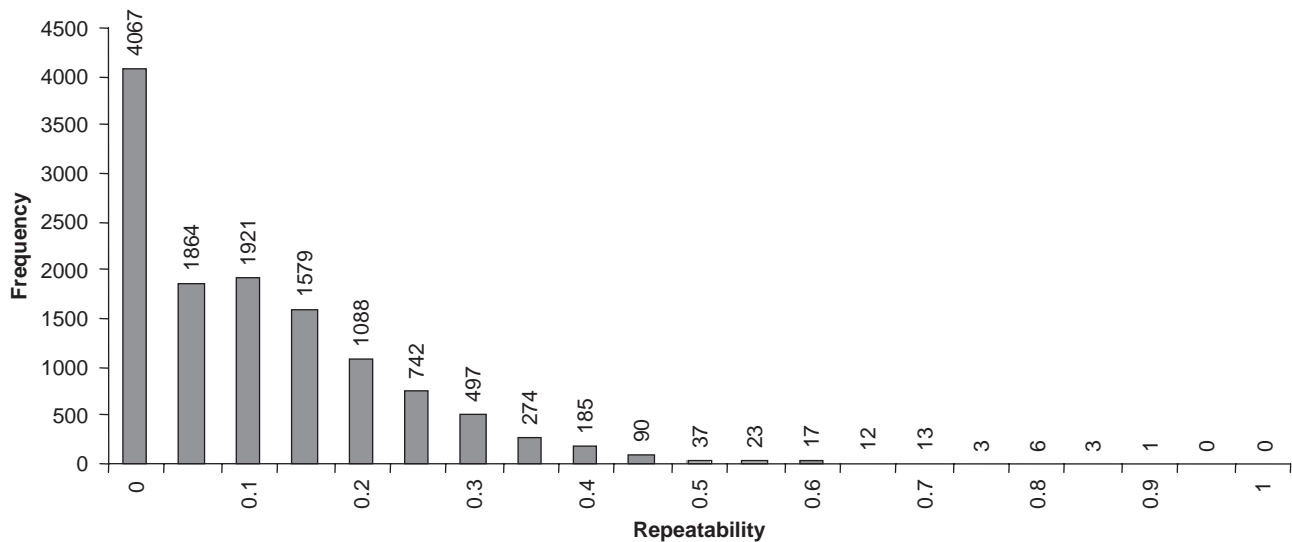


Fig. 1. Distribution of the repeatabilities for the 12 422 gene transcripts in the BXD study.

Table 1. The statistical power to detect QTL using 29 BXD lines, given 50% heritability of the trait and a nominal threshold of $P < 0.001$

| QTL effect (Phenotypic SD) | Number of replicate measurements for each line | | |
|-------------------------------|--|------|------|
| | 1 | 3 | 10 |
| 0.25 | 0.02 | 0.04 | 0.06 |
| 0.40 | 0.11 | 0.23 | 0.32 |
| 0.50 | 0.24 | 0.50 | 0.68 |
| 0.60 | 0.50 | 0.84 | 0.96 |
| 0.75 | 0.87 | 0.99 | 0.99 |

or ordinary least squares suggests that the analyses are fairly robust against variance heterogeneity due to differences in the number of replicates.

5.2.3 Significance testing Randomization testing was used to derive empirical genome-wide significance thresholds for each analysed trait. Figure 2 shows a representative distribution of the genome-wide 5% significance thresholds (F -statistics) obtained by randomization testing for the 12 422 traits in the full dataset. The frequency distribution of the thresholds is nearly normal with a mean of 19.5, a variance of 7.0 and a range from 9.7 to 28.1. Additional permutation tests ($n = 256$) for the traits with the most extreme thresholds showed that the thresholds for the individual traits are very stable with repeated permutation (Fig. 2). This rather broad distribution of threshold levels across traits implies that thresholds should be calculated for individual traits rather than using consensus thresholds from averaging or literature.

The analyses also revealed a positive correlation between the maximum F -statistic in the actual data and the empirical significance threshold (0.22) and between the maximum F -statistic and the repeatability of the trait (0.20). This indicates that the significance thresholds obtained in the real data are higher for traits with high maximum F -statistics (which also happen to be the traits with the highest repeatabilities).

5.2.4 Use of Parental and F_1 observations in QTL analyses The results from the simulation study evaluating the effect of including the parental and F_1 individuals are reported in Table 2. The LRT test-statistic increases in the populations where a parental difference but no QTL was simulated, but there was no corresponding increase in the significance threshold derived empirically by permutation. The same can be observed for the 50 transcripts with the most significant QTL (Table 2).

To evaluate the potential impact of including the parental lines in analysing the BXD data, we plotted the distribution of the observed differences between the parental lines (in phenotypic SD) for the 12 422 traits in the BXD data (Fig. 3A). It was observed that ~ 2400 traits had a difference of >1 SD between the parental lines and 81 traits had a difference of >2 SD. Figure 3B shows the distribution of these differences between the parental lines for the 50 traits in the data where the lowest P -values for detected QTL were observed when the parental means were included in the analysis. Clearly, when the parental lines are included as though they are additional RI lines, a very high proportion of the highest significance values are observed for traits with large parental phenotypic differences. Our simulation results suggest that these values may be inflated by the parental difference even if no QTL is present. For this reason all other results we report are for data from the RI lines alone, without the inclusion of the parental line data.

5.3 Inferences on the eQTL results

5.3.1 Applications of the FDR The number of QTL detected in the full dataset ranges from 608 to 663 across the analysis scenarios with a mean of 629. The use of a 5% genome-wide significance level for each trait implies that $12\,422 \times 0.05 = 621$ false positive results were expected. The number of significant QTL across all transcripts just exceeded the number expected by chance. However, application of the FDR combined with other sources of information can be used to make more informed inferences. The estimated proportion of true null hypotheses for this dataset was very close to unity. This means that the method of Storey and Tibshirani (2003) and the standard FDR calculations will provide similar results.

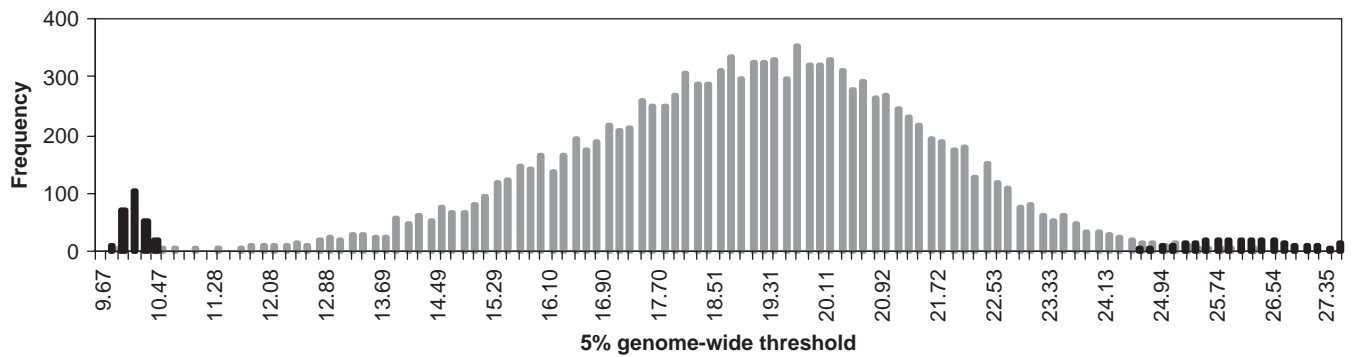


Fig. 2. Frequency distribution of the 5% genome-wide thresholds (F-statistic) derived by randomization testing for QTL affecting the 12 422 expression profiles (gray bars). The black bars on the left and the right show the frequency distribution of 256 additional randomization tests for the gene transcripts with the highest and lowest significance thresholds, respectively.

Table 2. The effects of including parental RI lines (P and F₁) in the QTL analysis on the test-statistic and the empirically derived significance thresholds

| Scenario | Relative test-statistic | Relative significance threshold |
|---|-------------------------|---------------------------------|
| No P or F ₁ , simulation | 1.00 ± 0.01 | 1.00 ± 0.01 |
| P and F ₁ included, 1 SD between P means, simulation | 1.13 ± 0.01 | 1.00 ± 0.01 |
| P and F ₁ included, 1 SD between P means, simulation | 1.15 ± 0.01 | 1.00 ± 0.01 |
| P and F ₁ included, 50 genes with most significant QTL from BXD data | 1.10 ± 0.005 | 0.98 ± 0.004 |

More stringent genome-wide thresholds. When using 1000 permutations, we could not estimate *P*-values <0.05 very accurately. By increasing the number of permutations we can obtain reasonable estimates of *P*-values <0.05. By using up to 1 000 000 permutations, 52 QTL could be identified at a 30% FDR (*P*-values <0.00125).

Post-hoc censoring by repeatability. There was a positive correlation between the number of detected QTL and the repeatability. The analyses have very low power to detect QTL for traits with repeatability <35%. Figure 4 shows how the proportion of traits with significant QTL increased with the repeatability in analyses of the full dataset. By censoring the data into datasets with a minimum repeatability, it is possible to identify groups of traits where the number of significant QTL is much greater than the expected number of false positives. Figure 5 gives an example of this, where the FDR decreases from ~100% in the full dataset to ~5% in the dataset containing the traits with repeatability >70%. By using this approach, a set of 112 QTL could be identified using a 30% FDR in the groups of transcripts with a repeatability >0.30 and a set of 12 QTL could be identified using a 5% FDR in the groups of transcripts with a repeatability >0.70.

Transcript location. The transcript location can be used as posterior information to increase the confidence for apparently *cis*-acting QTL. Of the 663 putative QTL detected using a given scenario, 155 mapped to the same chromosome as the gene transcript studied. Eighty-six QTL mapped within ±5 cM of the gene, an additional 37 mapped within ±10 cM (total 123) and 20 more mapped within ±20 cM of the gene (total 143). Given that the expected number of spurious *cis*-acting QTL are 4, 10 and 17, respectively, and without taking account of the confidence intervals, these results indicate that nearly all *cis*-acting QTL among the results are expected to represent true effects. This is true even for QTL falling into the band of 10–20 cM from the gene and probably represents the wide confidence interval on an estimated location in a study of this size (i.e. even if the QTL location is actually coincident with the gene, mapping accuracy is such that its estimated location will often be some distance away).

Figure 6 shows the distribution of repeatabilities for the QTL that map within 20 cM of the gene (i.e. *cis*-acting QTL) as well as the proportion of total number of genes in each repeatability bin for which *cis*-acting QTL were detected. It can be seen that the majority of the *cis*-acting QTL are for genes with repeatabilities >0.35. Even though quite a large number of the QTL were detected in the low repeatability groups, this merely reflects the very large number of genes in those groups, as the proportion of genes for which a QTL was found is very low. It also shows that for the HR classes the probability of detecting at least one significant *cis*-acting QTL gets rather large. Furthermore, looking at all detected QTL we find that when the trait repeatability >0.40, about 70–80% of the detected QTL are *cis*-acting (data not shown).

6 DISCUSSION

The major challenges for implementation of genetical genomics are the integration of the available technology and the scaling up of both microarray analyses and QTL mapping methods. The main focus of this study has been to thoroughly evaluate the adaptation of existing QTL mapping strategy to the mapping of QTL affecting gene expression. A real dataset of BXD mice has been used for this because there is not enough information about the genetics of gene regulation in the literature to design a realistic simulation study. Some simulations were used to test performance of methods in the absence of QTL effects. It should be noted that the BXD data originated from

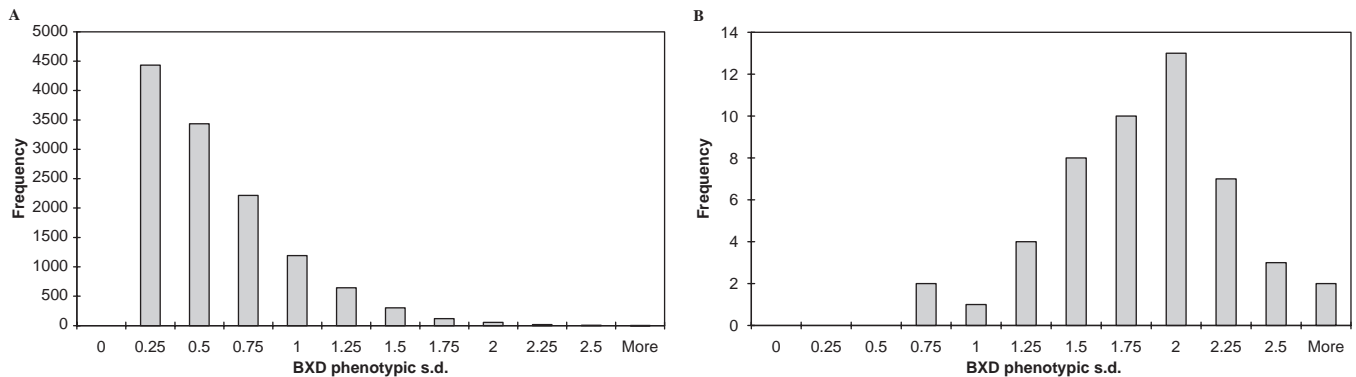


Fig. 3. Histograms of mean gene-expression differences between founder lines for all transcripts (A) and the transcripts for which the 50 most significant QTL were found (B).

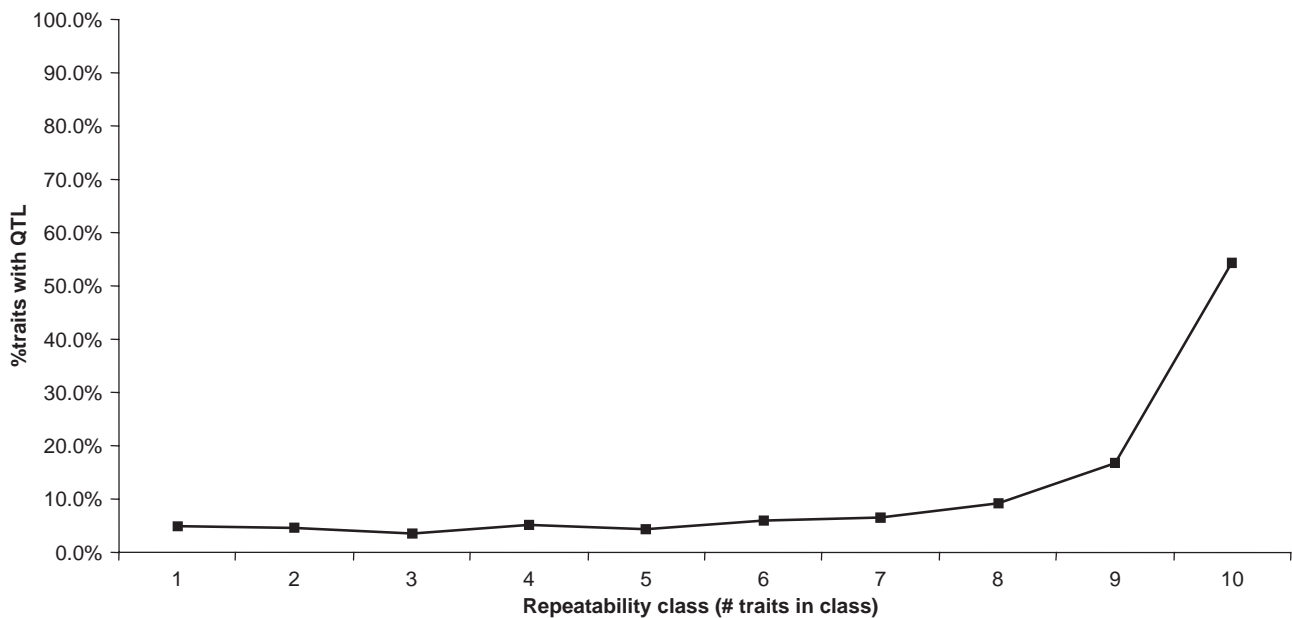


Fig. 4. The percentage of gene transcripts with at least a single QTL as a function of the repeatability for the analysed trait.

an earlier extraction of the expression data using MAS 5.0 than that reported by Chesler *et al.* (2005), who also used alternative methods to extract the expression data such as robust multichip average (RMA). In this report we described extensive analyses, which simultaneously address several of the key issues in QTL mapping and based on the results we have compiled a proposed strategy for automated mapping of QTL affecting gene expression. This strategy is outlined in Figure 7 and aspects of the individual components are discussed below.

The repeatability describes the proportion of the variance of a single measurement that is due to genetic and permanent environmental differences between lines. It is also the upper boundary of the heritability and it can be used to estimate the 'effective' heritability in a population when multiple measures are available. The expectation of the repeatability is not affected by the number of replicates, but the 'effective' heritability of the mean of a number of measures (which is a function of the repeatability) increases as the number

of replicates increases (Falconer and Mackay, 1996). As shown by Knapp and Bridges (1990), replication in a QTL mapping experiment only increases power when; (1) all genetic variation between lines is explained by QTL parameters and (2) when the total number of phenotypic measurements that can be collected is unlimited. This suggests that for QTL mapping the use of multiple microarrays per line is recommended when the number of available lines is limited as it is in the case of RIL. Otherwise, if the main concern is improving the power to detect QTLs a better option would be increasing the number of lines rather than the number of individuals per line. The repeatability can also be used as a means of censoring the data prior to QTL analyses. To reduce the environmental error the replicates should be sampled as uniform or balanced as possible with regard to environmental and biological factors (time-point, gender, etc.)

The average number of QTL detected across all scenarios (624) was very close to the number expected by chance (621) when 12 422 traits were analysed using a 5% genome-wide significance threshold.

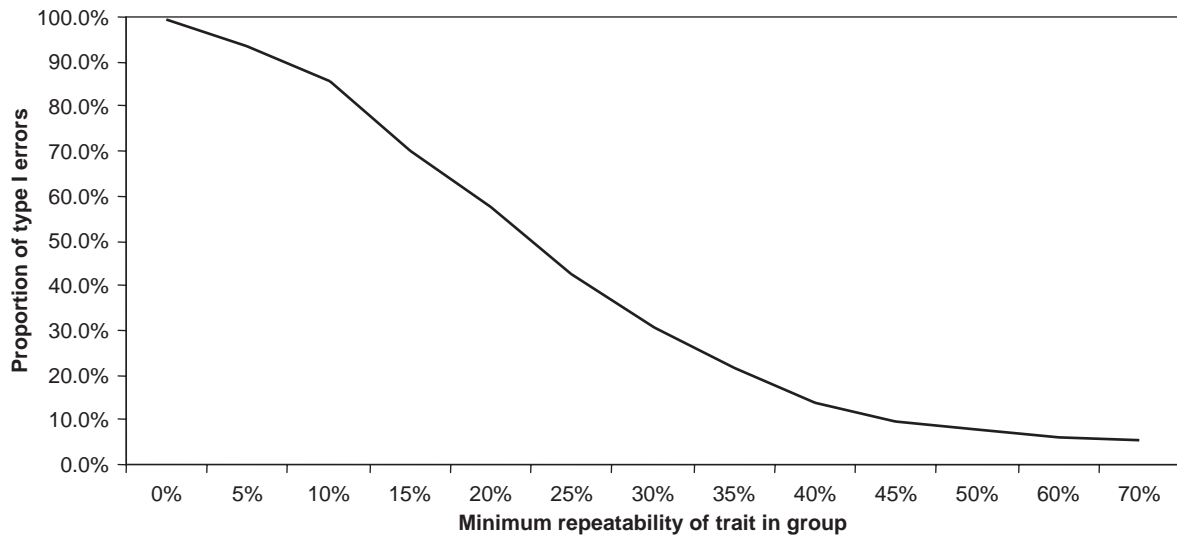


Fig. 5. The FDR as a function of the minimum level of repeatability for inclusion in QTL analysis.

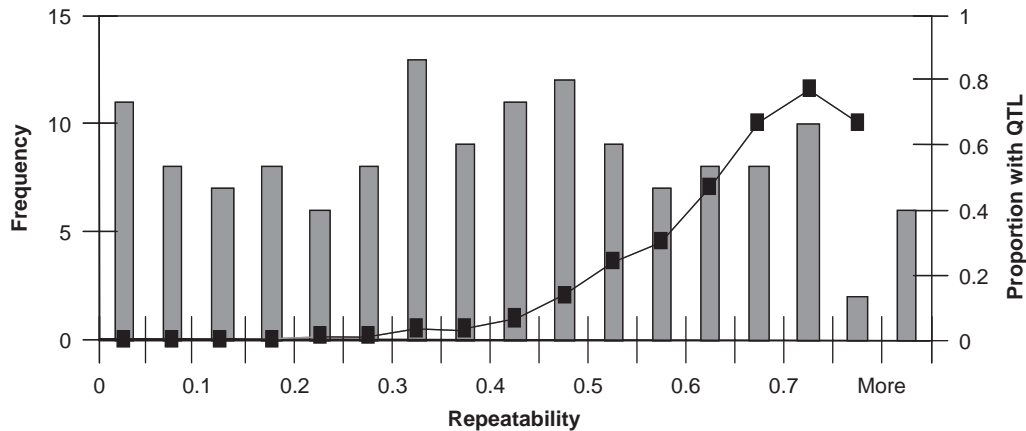


Fig. 6. Distribution of detected *cis*-acting QTL (within 20 cM proximal or distal from the location of the gene) across different repeatabilities. The line indicates the proportion of total number of genes in each repeatability bin for which *cis*-acting QTL were detected.

The ability to detect QTL is thus limited in this experiment and it is difficult to separate potentially true QTL from Type I errors. However, in Section 5.3 we outlined the use of prior genetic information that can be utilized in a genetical genomics study to improve the FDR and to separate the detected QTL to ‘high-confidence’ and ‘low-confidence’ significant QTL.

The power to detect QTL was shown to be greater for HR traits than for traits with low-repeatabilities. By separating the results *post hoc* according to the repeatability of the trait whilst monitoring the FDR for high and low-repeatability traits, a reduced set of HR traits can be identified where the FDR is considerably smaller than that in the entire dataset. If the computational resources are limited, we recommend that the traits be analysed in order of decreasing repeatabilities to obtain the most reliable results first. To reduce the multiple testing problem, one can also choose to analyse only transcripts that exceed a predefined level of repeatability.

A further source of prior genetic information when mapping QTL affecting gene expression is the location of the gene whose transcript

is analysed. By comparing the expected colocalization of QTL and gene-transcripts with that observed for the genome-wide significant QTL, it is possible to calculate an FDR for the potentially *cis*-acting QTL. In the BXD data, there was about 80% overlap between *cis*-acting QTL and the high-confidence QTL based on the repeatability calculations. This coincidence of *cis*-acting and high-confidence QTL suggests either a majority of QTL controlling expressions are *cis*-acting in this study, or perhaps that *cis*-acting QTL can have a large effect and thus tend to have higher repeatability.

In a standard FDR calculation, very large numbers of permutations (up to 1 000 000) are used to obtain an experiment-wide FDR based on very stringent *P*-values, resulting in 52 QTL detected at FDR of 30% in the present study. The joint use of trait repeatabilities and gene transcript location led to the identification of 168 high-confidence QTL using an FDR of 30% for the repeatability based tests and an FDR of 12% for 40 cM bins for the *cis*-acting QTL, which is a marked improvement compared to the 52 QTL detected when using an FDR based on *P*-values only. There is also a saving on

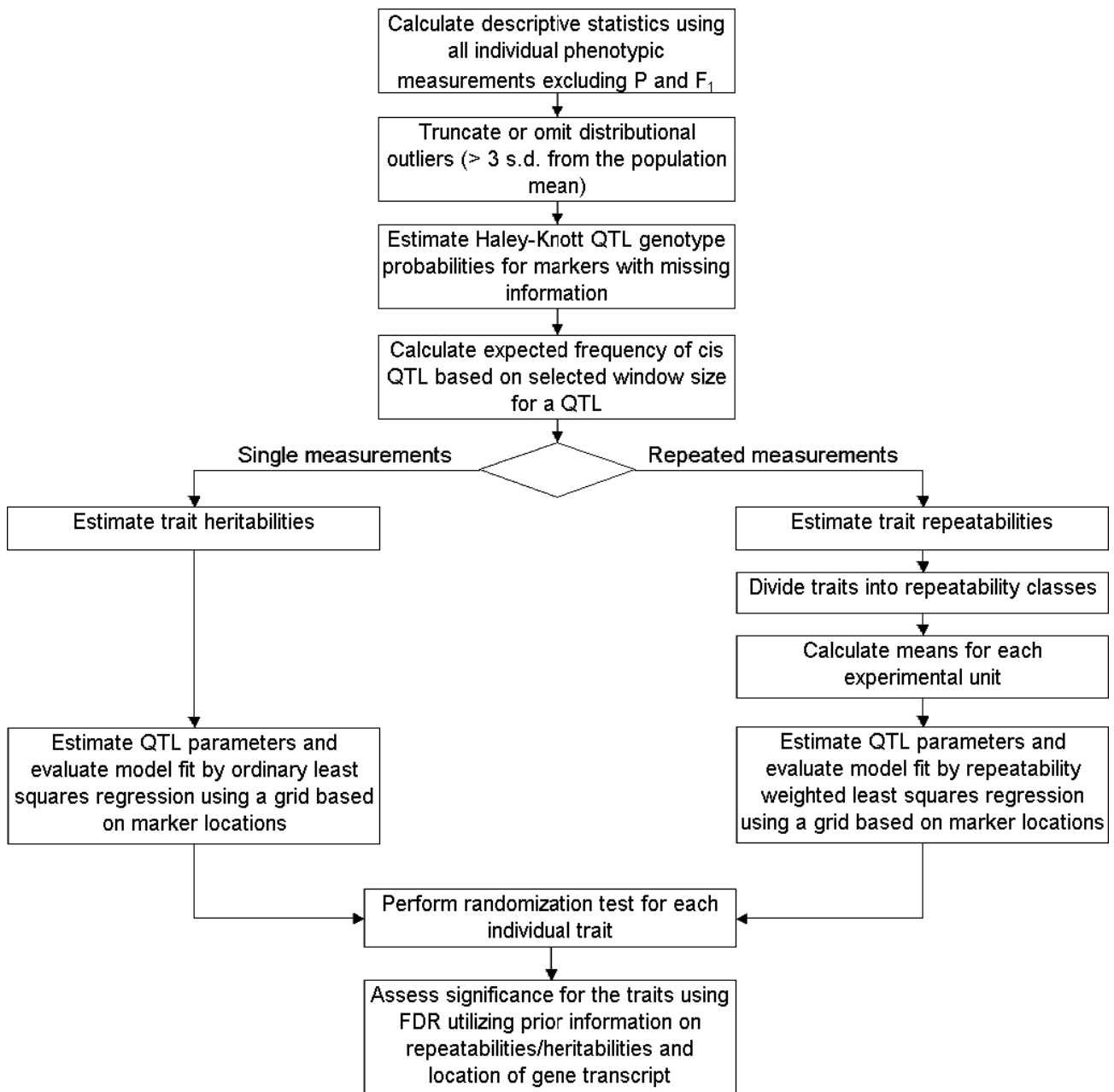


Fig. 7. An automated strategy for mapping QTL affecting gene expression. Further explanations of the components in the figure can be found in the text.

computing time because 1000 permutations are sufficient to estimate genome-wide P -values of 0.05. This clearly demonstrates how genetic information can be incorporated in a statistical framework for significance testing and the procedure is highly recommended when mapping QTL affecting gene expression.

When gene-expression data is analysed there are often groups of transcripts that are highly correlated with one another. In this dataset, for example, there exist cohorts of up to 250 transcripts, where all transcripts have a correlation >0.7 with each other. This correlation structure implies that most of the phenotypic information could be extracted by analysing a smaller number of phenotypes, i.e. only one representative transcript or a few principal components representing

the expression profile of the highly correlated transcripts in a cohort. Lan *et al.* (2003) present a method to use principal components or hierarchical clustering to reduce the dimensionality of QTL detection for gene-expression data. This would reduce the number of actual tests in these QTL analyses and potentially improve the FDR. An analysis based on these so-called supergenes would be most appropriate for the detection of *trans*-acting QTL that affect multiple gene transcripts. However, such analysis may fail to detect many *cis*-acting effects, especially if the *cis*-QTL do not have any direct or indirect *trans*-effects on other genes. Therefore, it may be desirable to supplement an analysis based on reduced data (i.e. principal components) with an additional analysis for *cis*-effects on individual

genes. Such an analysis can be computationally realistic as, for each gene only a very small proportion of the genome needs to be analysed.

Although inclusion of data on parental lines and F_1 has been suggested (Jansen and Stam, 1994), we have shown that this may increase the Type I error. Using simulations, we have shown that the observed increase in the test-statistic, when including parental observations, is present at all locations throughout the genome and rather than increasing the power it increases the Type I error. Jansen and Stam (1994) state that parental and F_1 observations should only be used in a multiple QTL mapping framework in order to divide the parental line difference among the cofactors and the QTL region under study. We feel that, although their approach may not inflate the Type I error, inclusion of the founder lines is unlikely to have a clear benefit in improving the resolution or power of QTL mapping.

Franken *et al.* (2001) use repeated measurements on BXD lines as independent observations in a QTL study for sleep regulation. We have shown by simulation that this approach leads to inflated test-statistics and highly increased Type I errors (data not shown). Because most QTL research is based on RIL means rather than on individual observations no further details are given here, but results from simulation and real data are available from the authors upon request.

When multiple observations are available on the same experimental unit, an analysis based on raw phenotypic means ignoring any variance heterogeneity between units resulting from differences in number of replicates per unit. We have evaluated weighted regression as an alternative and compared two different weights. Weighting based on repeatability of the trait and the number of individuals in each line was more robust than weighting based on the crudely estimated within line variance.

Interval mapping is generally expected to give a higher power than single marker mapping. The expected increase in power has, however, not resulted in a larger number of significant QTL in this study and appear to support the conclusions of Coffman *et al.* (2003) that interval mapping is not necessarily more powerful than single marker mapping when markers are highly informative as in this case. With increasing marker densities and particularly with markers that are fully informative, as they are in RI lines or other crosses between inbred lines, the difference between the two approaches is expected to diminish. In large scale mapping studies where there is a considerable computational demand, mapping of QTL at marker locations is therefore recommended to improve the computational efficiency of the study provided the marker density is fairly high.

The empirical significance thresholds, obtained from the randomization tests, vary substantially between traits in this study and it has been shown that this is not due to the sampling variance (Fig. 2). This implies that trait specific empirical thresholds should be used to draw valid conclusions from a QTL study. In contrast, Schadt *et al.* (2003) used the same consensus LOD scores across all transcripts while Brem *et al.* (2002) used nominal P -values for their linkage tests and only used permutation testing to predict the expected number of false positives under H_0 . Our findings suggest that further scrutiny of these published results using empirical thresholds is merited. Although the heterogeneity of significance thresholds may be more prominent for smaller studies, like the present example, investigators should verify whether their thresholds are homogeneous before imposing the same threshold across traits.

There is a positive correlation between the level of the genome-wide thresholds and the number of QTL detected for a gene transcript. This demonstrates that the permutation test only approximates the true null distribution, but bias of this type will make the test more conservative and this is not considered too unfavourable. However, testing for additional QTL for the same trait requires a new empirical significance threshold, which accounts appropriately for the effects of identified QTL.

In conclusion, genetical genomics is a new and exciting area but, as we have shown in this report, the adaptation of available technologies to this new framework is necessary. Here we provide more information on technical aspects regarding QTL analyses, but more work is necessary to optimize the design of the studies. This is especially true if one is to identify genetic interactions, which could potentially provide important information for reconstruction of genetic regulatory networks.

ACKNOWLEDGEMENTS

Ö.C. was funded by a fellowship from the Knut and Alice Wallenberg foundation. D.-J.K. and C.H. acknowledge funding from BBSRC. This research was supported by a BBSRC grant (csb005) for HPC at CSAR in Manchester.

REFERENCES

- Brem, R.B., Yvert, G., Clinton, R. and Kruglyak, L. (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science*, **296**, 752–755.
- Carlborg, Ö. (2002) New methods for mapping quantitative trait loci. PhD Thesis, Acta Universitatis Agriculturae Sueciae. Veterinaria 121. Swedish University of Agricultural Sciences, Uppsala, Sweden.
- Carlborg, Ö., Andersson-Eklund, L. and Andersson, L. (2001) Parallel computing in interval mapping of quantitative trait loci. *J. Hered.*, **92**, 449–451.
- Churchill, G.A. and Doerge, R.W. (1994) Empirical threshold values for quantitative trait mapping. *Genetics*, **138**, 963–971.
- Coffman, C.J., Doerge, R.W., Wayne, M.L. and McIntyre, L.M. (2003) Intersection tests for single marker QTL analysis can be more powerful than two marker QTL analysis. *BMC Genetics*, **4**, 10.
- Doerge, R.W. and Churchill, G.A. (1996) Permutation tests for multiple loci affecting a quantitative character. *Genetics*, **142**, 285–294.
- Falconer, D.S. and Mackay, T.F.C. (1996) *Introduction to Quantitative Genetics*. Longman, Harlow, UK.
- Franken, P., Chollet, D. and Tafti, M. (2001) The homeostatic regulation of sleep need is under genetic control. *J. Neurosci.*, **21**, 2610–2621.
- Grisel, J.E., Metten, P., Wenger, C.D., Merrill, C.M. and Crabbe, J.C. (2002) Mapping of quantitative trait loci underlying ethanol metabolism in BXD recombinant inbred mouse strains. *Alcohol Clin. Exp. Res.*, **25**, 610–616.
- Haley, C.S. and Knott, S.A. (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*, **69**, 315–324.
- Hochberg, Y. and Benjamini, Y. (1990) More powerful procedures for multiple significance testing. *Stat. Med.*, **9**, 811–818.
- Jansen, R.C. and Stam, P. (1994) High resolution of quantitative traits into multiple loci via interval mapping. *Genetics*, **136**, 1447–1455.
- Jansen, R.C. and Nap, J.P. (2001) Genetical genomics: the added value from segregation. *Trends Genet.*, **17**, 388–391.
- Klose, J., Nock, C., Herrmann, M., Stühler, K., Marcus, K., Blüggel, M., Krause, E., Schalkwyk, L.C., Rastan, S., Brown, S.D.M. *et al.* (2002) Genetic analysis of the mouse brain proteome. *Nat. Genet.*, **30**, 385–393.
- Knapp, S.J. and Bridges, W.C. (1990) Using molecular markers to estimate quantitative trait locus parameters: power and genetic variances for unreplicated and replicated progeny. *Genetics*, **126**, 769–777.
- Lan, H., Stoeber, J.P., Nadler, S.T., Schueler, K.L., Yandell, B.S. and Attie, A.D. (2003) Dimension reduction for mapping mRNA abundance as quantitative traits. *Genetics*, **164**, 1607–1614.
- Lander, E.S., Green, P., Abrahamson, J., Barlow, A., Daly, M.J., Lincoln, S.E. and Newburg, L. (1987) MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics*, **1**, 174–181.

- Lynch,M. and Walsh,B.I.(1998) *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Inc. Sunderland, MA.
- Risinger,F.O., Quick,E. and Belknap,J.K. (2000) Quantitative trait loci for acute behavioural sensitivity to paraoxon. *Neurotoxicol. Teratol.*, **22**, 667–674.
- Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci.*, **100**, 9440–9445.
- Schadt,E.E., Monks,S.A., Drake,T.A., Luskis,A.J., Che,N., Colinayo,V., Ruff,T.G., Milligan,S.B., Lamb,J.R., Cavet,G. *et al.* (2003) Genetics of gene expression surveyed in maize, mouse and man. *Nature*, **422**, 297–302.
- Wang,J., Williams,R.W. and Manly,K.F. (2003) WebQTL: web-based complex trait analysis. *Neuroinformatics*, **1**, 299–308.
- Wang,S., Basten,C.J. and Zeng,Z.-B. (2001–2003) Windows QTL Cartographer 2.0. Department of Statistics, North Carolina State University, Raleigh, NC.