BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 18 2008, pages 2015–2022
doi:10.1093/bioinformatics/btn373

Gene expression

Considering dependence among genes and markers for false discovery control in eQTL mapping

Liang Chen[1,*], Tiejun Tong[2] and Hongyu Zhao[3,4,*]

[1]Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, Los Angeles, CA, [2]Department of Applied Mathematics, University of Colorado, Boulder, CO, [3]Department of Epidemiology and Public Health and [4]Department of Genetics, Yale University, New Haven, CT, USA

Received on January 15, 2008; revised on July 4, 2008; accepted on July 15, 2008

Advance Access publication July 17, 2008

Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: Multiple comparison adjustment is a significant and challenging statistical issue in large-scale biological studies. In previous studies, dependence among genes is largely ignored. However, such dependence may be strong for some genomic-scale studies such as genetical genomics [also called expression quantitative trait loci (eQTL) mapping] in which thousands of genes are treated as quantitative traits and mapped to different genetical markers. Besides the dependence among markers, the dependence among the expression levels of genes can also have a significant impact on data analysis and interpretation.

Results: In this article, we propose to consider both the mean as well as the variance of false discovery number for multiple comparison adjustment to handle dependence among hypotheses. This is achieved by developing a variance estimator for false discovery number, and using the upper bound of false discovery proportion (uFDP) for false discovery control. More importantly, we introduce a weighted version of uFDP (wuFDP) control to improve the statistical power of eQTL identification. In addition, the wuFDP approach can better control false positives than false discovery rate (FDR) and uFDP approaches when markers are in linkage disequilibrium. The relative performance of uFDP control and wuFDP control is illustrated through simulation studies and real data analysis.

Contacts: liang.chen@usc.edu; hongyu.zhao@yale.edu

Supplementary information: Supplementary figures, tables and appendices are available at *Bioinformatics* online.

1 INTRODUCTION

Advanced chip technologies such as microarrays facilitate biological discoveries by studying thousands of genes simultaneously. However, false positive control presents a challenging statistical problem because a large number of hypotheses are tested in such studies. Family-wise error rate (FWER) control is one approach for multiple comparison correction and is defined as the probability of at least one false positive occurring. For example, when FWER is controlled at 0.01, the probability of identifying one or more false positives is less than or equal to 0.01. However, it is generally agreed that FWER control is conservative in genomic studies when there may be many signals and the primary goal is discovery. We may relax our criteria for more discoveries by tolerating more false positives. Consequently, alternative procedures, e.g. false discovery rate [FDR, which is the expectation of false discovery proportion (FDP)] controls (Benjamini and Hochberg, 1995; Storey and Tibshirani, 2003), are widely used for multiple comparison correction in high-dimensional genomic studies. However, the dependence among hypotheses is largely ignored in the existing methods based on FDR control, despite the facts that correlations among hypotheses may be high for genomics studies.

In this article, we focus on the analysis and interpretation of data arising from genetical genomics [expression quantitative trait loci (eQTL) mapping] studies, whose goal is to search for genetic loci associated with gene expression variations in a study population, e.g. samples from experimental crosses or an outbred population. Genetical genomics allows us to systematically study transcriptional regulation through sequence variations across study subjects. In this context, sequence variations can be considered as natural perturbations that can affect gene expressions. This approach has been successfully applied to yeast, fly, maize, mice, rat, human and other organisms (Brem *et al.*, 2002; Bystrykh *et al.*, 2005; Chesler *et al.*, 2005; Hubner *et al.*, 2005; Morley *et al.*, 2004; Schadt *et al.*, 2003; Spielman *et al.*, 2007; Stranger *et al.*, 2005). In these studies, gene expressions can vary significantly across individuals and genes often exhibit a complicated correlation structure among them. For example, genes sharing biological functions or in the same chromosomal domains (Cohen *et al.*, 2000) may be correlated. In addition, markers in close physical proximity may be in linkage disequilibrium and they are highly correlated. Therefore, there is a need to develop statistical methods to address the issue of dependence among hypotheses in such studies. Recently, several papers have addressed the multiple comparison problem for correlated hypothesis tests. Lehmann and Romano (2005) proposed to control the probability of k or more false rejections as a generalized family-wise error rate control without making any assumptions about the dependence structure among different hypotheses. Efron (2007) addressed this problem from a different perspective by proposing to use the expectation of false discovery number conditioning on a correlation effect parameter.

In this article, we propose to control the upper bound of FDP (uFDP) to handle multiple comparisons for dependent hypotheses. It is similar to the generalized family-wise error rate control because we control the probability of false rejection proportion larger than

*To whom correspondence should be addressed.

a given threshold. More importantly, we introduce a weighted version to control the uFDP (wuFDP) to improve the statistical power of eQTL identification. These weights are related to the correlation structure of the hypotheses. Thus, in contrast to previous studies, we not only consider the dependence among hypotheses, but also utilize it to improve statistical power. In addition, the wuFDP approach can better control false positives than FDR and uFDP approaches when markers are in linkage disequilibrium. We, therefore, recommend using wuFDP control as a practical approach to identify significant marker–gene pairs in eQTL studies.

In the following sections, we will illustrate the effect of dependence on false discovery control using an eQTL dataset. Then, we will introduce uFDP and wuFDP controls followed by simulations and real data analysis. Finally, we conclude this article in Section 4.

2 METHODS

2.1 Impact of dependence among hypotheses

As discussed earlier, the objective of an eQTL study is to identify chromosomal regions affecting the expression levels of the genes measured on microarrays. For each individual, both gene expression levels and marker genotypes are collected and their associations are investigated. If each gene is treated as a quantitative trait, traditional QTL mapping methods can be applied to identify markers associated with each gene. Much work has been done in the literature to consider the dependence among markers in QTL mapping when only one trait is studied. However, the expression levels of many genes are highly dependent, and appropriate statistical methods are needed to take into account such dependence. The complexity of gene expression pattern in eQTL mapping makes the dependence among hypotheses complicated. Supplementary Figure 1 illustrates the non-trivial dependence among genes in an eQTL expression dataset. The dataset contains 1000 genes for 60 individuals. The details of the dataset are described subsequently in Section 3.3. This figure shows the histograms of pairwise correlations among genes. It clearly indicates a strong dependence pattern among genes. When genes are highly correlated with each other, if one of them is falsely declared significant, other correlated genes are also likely to be falsely declared significant. In the presence of correlation among hypotheses, the variance of false positive number will increase and correspondingly FDP may deviate much more from the mean, i.e. FDR, than when hypotheses are independent of each other. This phenomenon was also explored by Owen (2005) for the identification of differentially expressed genes under two conditions.

To illustrate the magnitude of the variation of false positive number in the presence of correlated hypotheses in eQTL studies, Supplementary Figure 2 shows the histogram of false positive number from permuted datasets, where all the null hypotheses are true. The permuted data were generated from the original data as follows. In the original dataset, the observations for each individual consist of two components, gene expression data and genotype data. For each permuted dataset, the association between these two sets of observations was randomly paired across all the individuals, whereas the gene expression data vector and the genotype vector were kept intact individually. That is, we permuted the whole transcriptome together instead of permuting every gene separately, so the dependence among genes and the dependence among markers were kept but any association between genes and markers was destroyed. A linear regression model was used to test the association between gene expression and marker genotype. For permuted datasets, all the discoveries are false positives. For 1000 simulations, using the P-value threshold of 1.0×10^{-5}, although the average false positive number 258.1 was close to 202.8 (the expected false positive number for 1000 gene by 20 281 marker comparisons under the independence situation), the false positive number was very high for some permuted datasets and the maximum false positive number was 482. As can be seen from Supplementary Figure 2, the distribution is right-skewed and the false positive number tends to deviate much from the mean for some permutations. The SD of the false positive numbers across 1000 permutations was about 31.6, much larger than 14.2, the expected SD when all the hypotheses were independent.

As we mentioned that much work has been done to consider marker dependence in QTL mapping. Most of them focus on the FWER control to avoid any single false positive across the whole genome scan (Churchill and Doerge, 1994; Doerge and Churchill, 1996). It has been reported that the applicability of traditional FDR approach for the linkage analysis of a single trait is dubious (Chen and Storey, 2006). The dependence among markers makes the interpretation of FDR problematic. By genotyping more markers around the causal marker, we can identify more true positives. These true positives cannot provide us additional information because they represent the same signal from the causal marker. However, the FDR will be decreased by adding these markers. Correspondingly, more false positives will be included to achieve the specified FDR level. To overcome the above limitations of FDR control, we propose a wuFDP control. First, we define uFDP control.

2.2 uFDP and wuFDP control

As shown in Section 2.1, because FDR control only considers the expected false discovery proportion, the FDP for each realization may differ from the desired FDR level, an issue that may become much more severe when the hypotheses are dependent on each other. To remedy this problem, we propose to control FDP at a certain level so that we are confident about our results. FDP is formally defined as (Lehmann and Romano, 2005)

$$FDP=\begin{cases} V/R & R>0 \\ 0 & R=0 \end{cases},$$

where V is the false positive number and R is the total discovery number. V can be written as $\sum_{i=1}^{h} v_i$, where h is the total number of hypotheses and v_i is the indicator function of whether hypothesis i is falsely rejected. R can be written as $\sum_{i=1}^{h} r_i$ and r_i is the indicator function of whether hypothesis i is rejected.

If $R>0$, we define the uFDP as

$$uFDP=\frac{E(V)+z_{1-\alpha}\sqrt{\mathrm{Var}(V)}}{R},$$

where $z_{1-\alpha}$ is the $100(1-\alpha)$-th percentile of the standard normal distribution. According to the central limit theorem, if $v_1,...,v_h$ are independent of each other, when $h\to\infty$, $(V-E(V))/\sqrt{\mathrm{Var}(V)}$ converges to the standard normal distribution. For a given dataset, R is fixed, if we approximate the distribution of V as a normal distribution and if both the mean and variance of V can be calculated,

$$Pr(\frac{V}{R}\geq uFDP)=\alpha.$$

Therefore, for a given α, the probability that FDP is larger than uFDP can be controlled at α. It will have a better statistical power than other procedures which control $Pr(V/R\geq uFDP)\leq\alpha$ with the equal sign only achieved under certain situations. However, if $v_1,...,v_h$ are dependent on each other, the approximation of V as a normal distribution may be inappropriate. Therefore, we consider a weighting scheme to make the dependence among hypotheses smaller. Specifically, if a gene (or marker) is highly correlated with other genes (or markers) , it will be assigned a smaller weight. If $\sum_{i=1}^{h} w_i r_i>0$, the weighted version of the uFDP is defined as

$$wuFDP=\frac{E(\sum_{i=1}^{h} w_i v_i)+z_{1-\alpha}\sqrt{\mathrm{Var}(\sum_{i=1}^{h} w_i v_i)}}{\sum_{i=1}^{h} w_i r_i}.$$

If we approximate the distribution of $\sum_{i=1}^{h} w_i v_i$ as a normal distribution, we can find wuFDP such that

$$Pr(\frac{\sum_{i=1}^{h} w_i v_i}{\sum_{i=1}^{h} w_i r_i}\geq wuFDP)=\alpha.$$

If we treat V and $\sum_{i=1}^{h} w_i v_i$ as continuous random variables with a unimodal probability density function, according to Vysochanskiĭ−Petunin inequality (Vysochanskiĭ and Petunin, 1980), for $z_{1-\alpha} > \sqrt{8/3}$, we have,

$$Pr\left(\frac{V}{R} \geq uFDP\right) \leq \frac{4}{9z_{1-\alpha}^2},$$

$$Pr\left(\frac{\sum_{i=1}^{h} w_i v_i}{\sum_{i=1}^{h} w_i r_i} \geq wuFDP\right) \leq \frac{4}{9z_{1-\alpha}^2}.$$

If $z_{1-\alpha}$ is chosen to be 2.33, the FDP or weighted FDP ($\sum_{i=1}^{h} w_i v_i / \sum_{i=1}^{h} w_i r_i$) are controlled at 0.01 level according to the normal distribution approximation and 0.08 level according to Vysochanskiĭ−Petunin inequality. If R or $\sum_{i=1}^{h} w_i r_i$ is equal to 0, FDP or weighted FDP is defined as 0, which can also be controlled at α level.

In practice, the threshold for rejecting hypotheses is chosen to let uFDP or wuFDP equal to our predefined level. uFDP and wuFDP can also be written as

$$uFDP = \frac{\pi_0 E(\sum_{i=1}^{h} I(|T_i^0| \geq t)) + z_{1-\alpha}\pi_0\sqrt{\mathrm{Var}(\sum_{i=1}^{h} I(|T_i^0| \geq t))}}{\sum_{i=1}^{h} I(|T_i| \geq t)},$$

$$wuFDP = \frac{\pi_0 E(\sum_{i=1}^{h} w_i I(|T_i^0| \geq t)) + z_{1-\alpha}\pi_0\sqrt{\mathrm{Var}(\sum_{i=1}^{h} w_i I(|T_i^0| \geq t))}}{\sum_{i=1}^{h} w_i I(|T_i| \geq t)},$$

where π_0 is the proportion of true null hypotheses among all the hypotheses. T_i^0 is the test statistic random variable under the null and T_i is the observed test statistic. $I(\cdot)$ is the indicator function and t is the threshold. In this article, we estimate π_0 as 1, which means only a very small proportion of true marker–gene associations among all of the possible marker–gene pairs. This is a reasonable assumption in genetical genomics studies. We note that a better estimator of π_0 will improve the results in other settings.

2.3 Variance estimation of V

In this section, we discuss how to estimate the variance of V in the context of eQTL analysis. To detect eQTL, we fit a linear regression model relating the expression of gene g to the genotype of marker m (coded as 0, 1 or 2 for homozygous rare, heterozygous and homozygous common alleles):

$$y_{gk} = \beta_{gm} x_{mk} + \varepsilon_{gk},\ k = 1, \ldots, n,$$

y_{gk} is the expression level for gene g and individual k, x_{mk} is the number of common alleles for marker m and individual k, the ε_{gk} ($k = 1, \ldots, n$) are independent normal random variables with mean 0 and variance σ_g^2. Y's and X's are standardized so that $\sum_{k=1}^{n} y_{gk} = 0$, $\sum_{k=1}^{n} y_{gk}^2 = 1$, $\sum_{k=1}^{n} x_{mk} = 0$ and $\sum_{k=1}^{n} x_{mk}^2 = 1$. Similarly, we can fit a regression between gene g' and marker m'. Although ε_{gk} are independent for different individuals and $\varepsilon_{g'k}$ are independent for different individuals, ε_{gk} and $\varepsilon_{g'k}$ may be related to each other for the same individual k. That is, we have the following covariance structure: $\mathrm{Cov}(\varepsilon_{gk}, \varepsilon_{g'k'}) = \varrho_{gg'}$ for $k = k'$ and $\mathrm{Cov}(\varepsilon_{gk}, \varepsilon_{g'k'}) = 0$ for $k \neq k'$. Under the above setup, the least squares estimates for β_{gm} and $\beta_{g'm'}$ are

$$\hat{\beta}_{gm} = \sum_{k=1}^{n} y_{gk} x_{mk} \sim Norm(\beta_{gm}, \sigma_g^2),$$

$$\hat{\beta}_{g'm'} = \sum_{k=1}^{n} y_{g'k} x_{m'k} \sim Norm(\beta_{g'm'}, \sigma_{g'}^2),$$

where σ_g^2 and $\sigma_{g'}^2$ can be estimated by the residual sum of squares divided by $(n-2)$, and their estimates are denoted as $\hat{\sigma}_g^2$ and $\hat{\sigma}_{g'}^2$.

Under the null hypotheses that $\beta_{gm} = 0$ and $\beta_{g'm'} = 0$,

$$\frac{\hat{\beta}_{gm}}{\hat{\sigma}_g} \sim t_{n-2},$$

$$\frac{\hat{\beta}_{g'm'}}{\hat{\sigma}_{g'}} \sim t_{n-2}.$$

We define the test statistics as follows:

$$T_{gm} = \Phi^{-1}\left(Pt_{n-2}\left(\frac{\hat{\beta}_{gm}}{\hat{\sigma}_g}\right)\right),$$

$$T_{g'm'} = \Phi^{-1}\left(Pt_{n-2}\left(\frac{\hat{\beta}_{g'm'}}{\hat{\sigma}_{g'}}\right)\right),$$

where Pt_{n-2} is the cdf of t distribution with $n-2$ degrees of freedom, Φ^{-1} is the inverse function of the cdf of standard normal. Under the null hypotheses, $(T_{gm}^0, T_{g'm'}^0)^T$ can be approximated by a bivariate normal distribution with mean $(0,0)^T$ and variance–covariance matrix $\begin{pmatrix} 1, \rho_{gg'mm'} \\ \rho_{gg'mm'}, 1 \end{pmatrix}$. $\rho_{gg'mm'}$ can be approximated as the correlation between $\hat{\beta}_{gm}/\hat{\sigma}_g$ and $\hat{\beta}_{g'm'}/\hat{\sigma}_{g'}$. With a large sample size, the latter one can be further approximated as the correlation between $\hat{\beta}_{gm}$ and $\hat{\beta}_{g'm'}$, that is $\frac{\varrho_{gg'}}{\sigma_g \sigma_{g'}} \sum_{k=1}^{n} x_{mk} x_{m'k}$. The correlation consists of two parts: $\varrho_{gg'}/\sigma_g\sigma_{g'}$ is the correlation between expression-level residuals which can be estimated as the Pearson's correlation between $\hat{\varepsilon}_g = Y_g - \hat{\beta}_{gm} X_m$ and $\hat{\varepsilon}_{g'} = Y_{g'} - \hat{\beta}_{g'm'} X_{m'}$; $\sum_{k=1}^{n} x_{mk} x_{m'k}$ can be treated as the correlation between markers. Denote the estimator for $\rho_{gg'mm'}$ as $\hat{\rho}_{gg'mm'}$.

The covariance between v_{gm} and $v_{g'm'}$ can be estimated as:

$$\mathrm{Cov}(v_{gm}, v_{g'm'}) = Pr(|T_{gm}^0| \geq t, |T_{g'm'}^0| \geq t | \hat{\rho}_{gg'mm'}) - Pr(|T_{gm}^0| \geq t)^2.$$

Therefore, for a threshold t and when π_0 is set as 1, we can calculate the FDR, uFDP and wuFDP as:

$$FDR = \frac{E(\sum_{i=1}^{h} I(|T_i^0| \geq t))}{\sum_{i=1}^{h} I(|T_i| \geq t)},$$

$$uFDP = \frac{E(\sum_{i=1}^{h} I(|T_i^0| \geq t)) + z_{1-\alpha}\sqrt{\mathrm{Var}(\sum_{i=1}^{h} I(|T_i^0| \geq t))}}{\sum_{i=1}^{h} I(|T_i| \geq t)},$$

$$wuFDP = \frac{E(\sum_{i=1}^{h} w_i I(|T_i^0| \geq t)) + z_{1-\alpha}\sqrt{\mathrm{Var}(\sum_{i=1}^{h} w_i I(|T_i^0| \geq t))}}{\sum_{i=1}^{h} w_i I(|T_i| \geq t)}.$$

For each threshold t, we can count the total number of rejections R and calculate FDR, uFDP and wuFDP. We stop at the largest R whose corresponding FDR, uFDP and wuFDP are less than our predefined levels (e.g. 0.1). This FDR is similar to Storey's FDR (Storey and Tibshirani, 2003) except that we estimate π_0 as 1.

2.4 Weighting scheme

In this section, we discuss wuFDP control as a way to increase statistical power. The rationale of our approach is based on minimizing the variance of weighted false discoveries. Let $\Sigma = (\tau_{ij})_{i,j=1,\ldots,h}$ denote the estimated variance–covariance matrix of $v_1, \ldots, v_h$. For the threshold t, we can get a corresponding P-value threshold p, and $\tau_{ij} = p - p^2$ for $i = j$ and $\tau_{ij} = p_{ij} - p^2$ for $i \neq j$ where $p_{ij} = Pr(|T_{gm}^0| \geq t, |T_{g'm'}^0| \geq t | \hat{\rho}_{gg'mm'})$. Let hypothesis i correspond to gene g and marker m and hypothesis j correspond to gene g' and marker m'. Given $\sum_{i=1}^{h} w_i = h$, the optimal weights minimizing $\mathrm{Var}(\sum_{i=1}^{h} w_i v_i)$ are given as

$$w_{opt} = \frac{h\Sigma^{-1}\mathbf{1}}{\mathbf{1}^T\Sigma^{-1}\mathbf{1}}, \quad (1)$$

where $\mathbf{1} = (1, \ldots, 1)^T$ with size h. Note that the optimal weights in (1) are not guaranteed to be non-negative, which makes the interpretation difficult. The constraint that $w_i \geq 0$ in variance minimization is commonly used in portfolio optimization. But the computation is prohibitive for our studies where there are thousands of hypotheses and the weights need to be updated many times according to different threshold values. In this article, we propose to use the following weight for hypothesis i:

$$w_i = \frac{h/s_i}{\sum_{i=1}^{h} 1/s_i},\ i = 1, \ldots, h, \quad (2)$$

where $s_i = \sum_{j=1}^{h} \tau_{ij}$. In Appendix 1 in Supplementary Material, we proved that the weights in (2) are all positive under the setting that the tests are two-sided.

If the hypotheses are independent of each other (i.e. Σ is diagonal), the optimal weights defined in (1) are

$$w_{opt} = \left(\frac{h/\tau_1}{\sum_{i=1}^{h} 1/\tau_i}, \ldots, \frac{h/\tau_h}{\sum_{i=1}^{h} 1/\tau_i} \right),$$

which are equal to our proposed weights in (2).

More generally, the optimality of our proposed weights holds when the variance–covariance matrix Σ is a block diagonal matrix with compound symmetric matrix in each block. The proof is in Appendix 2, in Supplementary Material.

Though the proposed weights is not optimal for general Σ, simulations (data not shown) indicate that the weighted variance is smaller than the original one in most situations. Appendix 3 in Supplementary Material provides some theoretical justifications under certain conditions. Specifically, if $\tau_{ij} \le \tau_{kl}$ implies that $f_{ij}(\tau_{ij}) \ge f_{kl}(\tau_{kl})$, where $f_{ij}(\tau_{ij}) = 1/s_i s_j$, we have

$$\mathrm{Var}\left(\sum_{i=1}^{h} w_i v_i \right) \le \mathrm{Var}\left(\sum_{i=1}^{h} v_i \right).$$

Because $\tau_{11} = \cdots = \tau_{hh}$ in our study, we have

$$\sum_{i=1}^{h} \mathrm{Var}(w_i v_i) = \tau_{11} \sum_{i=1}^{h} w_i^2 \ge \frac{\tau_{11}}{h} (\sum_{i=1}^{h} w_i)^2 = \sum_{i=1}^{h} \mathrm{Var}(v_i).$$

The sum of diagonal elements of the weighted variance is always larger than the original one. This implies that the reduction of variance using the proposed weights (or the optimal weights) are obtained in the off-diagonal elements. $v_1, \ldots, v_h$ are more likely to be uncorrelated.

3 RESULTS

3.1 Simulations for gene dependence

We focus on the dependence among genes first. We consider a study consisting of 100 individuals, with each one typed at one marker and measured expression levels for 1000 genes. As for genotype data, among these individuals, 9 have genotype aa (coded as 0), 42 have genotype Aa (coded as 1) and 49 have genotype AA (coded as 2). Thus the major allele frequency is 0.7 and the minor allele frequency is 0.3.

As for the expression levels of the 1000 genes, we assume that 50 genes have different expression levels across different genotype groups. Four scenarios are considered.

(1) All of the non-differentially expressed genes are independent.

(2) Fifty non-differentially expressed genes are correlated with each other (pairwise correlation is 0.6, 0.7, 0.8 or 0.9).

(3) Hundred non-differentially expressed genes are correlated with each other (pairwise correlation is 0.6, 0.7, 0.8 or 0.9).

(4) Two hundred non-differentially expressed genes are correlated with each other (pairwise correlation is 0.6, 0.7, 0.8 or 0.9).

For all of the scenarios, we assume that the differentially expressed genes and the non-differentially expressed genes are independent. For each individual, expression data was simulated using a multivariate normal distribution with mean 0 and corresponding covariance matrix defined by the above correlation structures. For the 50 differentially expressed genes, differential signal β was varied from 0.3, 0.4, 0.5, ..., to 0.9. The simulations were repeated 1000 times. Each gene expression and each marker genotype data were standardized to have sample mean 0 and sample variance $1/(n-1)$.

Table 1. The frequency of FDP or weighted FDP larger than 0.1 among 1000 simulations for FDR, uFDP and wuFDP controls

N	Corr	FDR: $Pr(\mathrm{V/R} \ge 0.1)$	uFDP: $Pr(\mathrm{V/R} \ge 0.1)$	wuFDP: $Pr\left(\frac{\sum w_i v_i}{\sum w_i r_i} \ge 0.1\right)$	wuFDP: $Pr(\mathrm{V/R} \ge 0.1)$
0		0.44	0.05	0.04	0.04
50	0.6	0.44	0.05	0.04	0.05
	0.7	0.42	0.05	0.05	0.06
	0.8	0.44	0.05	0.05	0.06
	0.9	0.42	0.04	0.05	0.06
100	0.6	0.41	0.03	0.04	0.05
	0.7	0.37	0.04	0.05	0.06
	0.8	0.40	0.02	0.04	0.05
	0.9	0.37	0.02	0.04	0.05
200	0.6	0.35	0.02	0.04	0.05
	0.7	0.34	0.02	0.04	0.06
	0.8	0.32	0.01	0.04	0.05
	0.9	0.29	0.01	0.04	0.04

The differential signal $\beta = 0.5$. The threshold for FDR, uFDP and wuFDP is 0.1. $z_{1-\alpha}$ is 1.65 to control $Pr(V/R \ge 0.1)$ and $Pr(\sum w_i v_i / \sum w_i r_i \ge 0.1) \le 0.05$. The number of correlated non-differentially expressed genes varies from 0 to 200. The correlation among these genes varies from 0.6 to 0.9. For wuFDP control, $Pr(V/R \ge 0.1)$ is also listed. Note that if R or $\sum w_i r_i$ is equal to 0, V/R or $\sum w_i v_i / \sum w_i r_i$ is treated as 0.

Table 1 summarizes the frequency of FDP or weighted FDP larger than 0.1 among those 1000 simulations for FDR, uFDP and wuFDP controls. The differential signal β is 0.5 and FDR, uFDP and wuFDP cutoff values are 0.1. $z_{1-\alpha} = 1.65$ so that $Pr(V/R \ge 0.1)$ and $Pr(\sum w_i v_i / \sum w_i r_i \ge 0.1)$ can be controlled at 0.05 level according to the normal distribution approximation. For wuFDP control, all of the estimated probabilities of weighted FDP larger than 0.1 are about 0.04∼0.05 which are close to our significant level 0.05. Even for the FDP instead of the weighted one, wuFDP still performs well with $Pr(FDP \ge 0.1) \le 0.06$. For uFDP control, the probabilities of FDP larger than 0.1 are much less than 0.05 when the dependence is high. It suggests that uFDP control may be too conservative in the presence of dependence. On the contrary, for FDR control, there is a chance of 29%∼44% to have FDP larger than 0.1. This table indicates that wuFDP and uFDP control FDP and weighted FDP almost at the predefined significant level. However, FDP may be much larger than the controlled average level for regular FDR control. Figure 1 shows the boxplots of FDP or weighted FDP for FDR, uFDP and wuFDP controls. If the dependence among hypotheses is strong, FDP is more likely to deviate from the mean for FDR control. However, uFDP and wuFDP can control FDP not to deviate much from the mean.

Statistical power of uFDP and wuFDP is compared using these simulation results. Power is defined by the average of ratios between truly declared positives and total positives for these 1000 simulations. In Figure 2, power is plotted for each differential signal (0.3–0.9). The triangle symbol line is for wuFDP control and the cross symbol line is for uFDP control. When the correlation among non-differentially expressed genes increases or the number of correlated non-differentially expressed genes increases, wuFDP control has a better power than uFDP control. If all of the non-differentially expressed genes are independent of each other, Supplementary Figure 3 shows that wuFDP control performs almost the same as uFDP control.

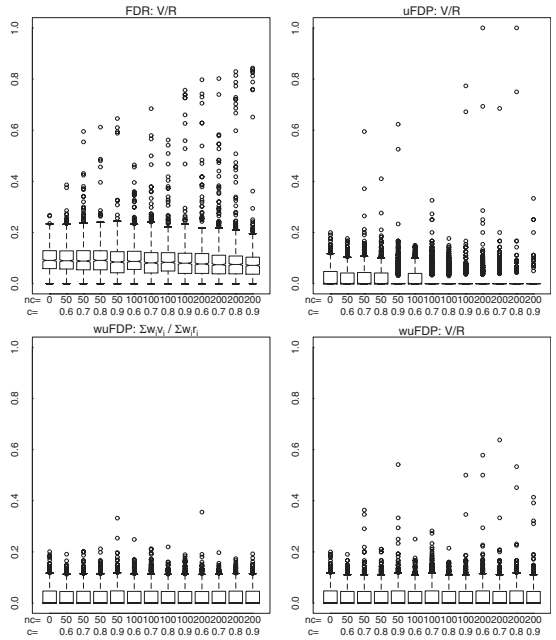

Fig. 1. Boxplots of FDP or weighted FDP among 1000 simulations for FDR, uFDP and wuFDP controls. The differential signal $\beta = 0.5$. The threshold for FDR, uFDP and wuFDP is 0.1. $z_{1-\alpha}$ is 1.65. The number of correlated non-differentially expressed genes (nc) varies from 0 to 200. The correlation among these genes (c) varies from 0.6 to 0.9.

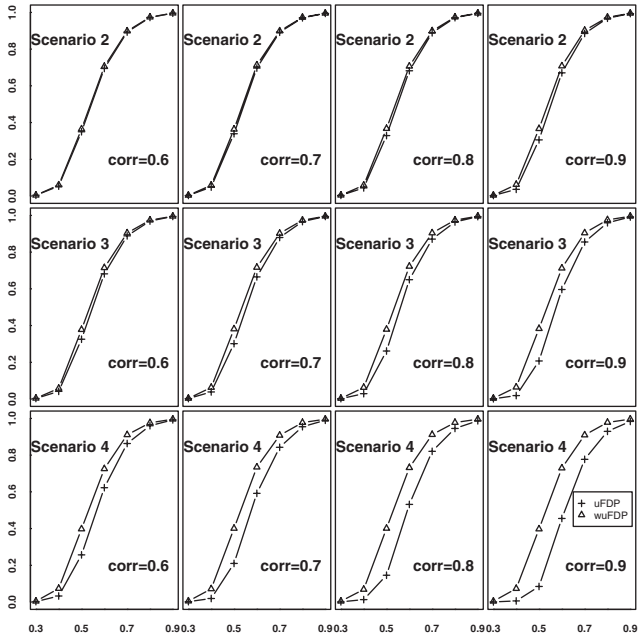

Fig. 2. Statistical power for different differential signal (0.3–0.9). The number of correlated non-differentially expressed genes varies from 50 to 200. The correlation among these genes varies from 0.6 to 0.9. The threshold for uFDP and wuFDP is 0.1. $z_{1-\alpha}$ is 1.65. Triangle symbol line is for wuFDP control and cross symbol line is for uFDP control.

Table 2 summarizes the average true positives and false positives when the signal is 0.5. Here, for wuFDP control, if one hypothesis is correctly rejected or falsely rejected, we count it as one without using its weight. wuFDP control performs well when some non-differentially expressed genes are correlated with each other. The performance of wuFDP is almost the same as that of uFDP, when all of the non-differentially expressed genes are independent of each other. uFDP controls false positives well for the dependent cases, but it has lower statistical power.

Table 2. The average true positives (TP) and false positives (FP) among 1000 simulations for FDR, uFDP and wuFDP controls

N	Corr	FDR: TP	uFDP: TP	wuFDP: TP	FDR: FP	uFDP: FP	wuFDP: FP
0		29.4	17.9	17.9	3.2	0.5	0.5
50	0.6	29.3	17.5	18.2	3.2	0.4	0.5
	0.7	29.3	17.0	18.2	3.3	0.5	0.5
	0.8	29.6	16.5	18.4	3.3	0.4	0.5
	0.9	29.4	15.3	18.3	3.2	0.4	0.6
100	0.6	29.6	16.3	18.9	3.3	0.3	0.5
	0.7	29.4	15.1	19.1	3.2	0.3	0.6
	0.8	29.6	13.1	19.0	3.2	0.2	0.5
	0.9	29.7	10.4	19.2	3.6	0.2	0.6
200	0.6	29.2	12.8	19.9	3.5	0.2	0.6
	0.7	29.3	10.5	20.0	3.4	0.2	0.7
	0.8	29.3	7.3	20.1	4.0	0.1	0.6
	0.9	29.3	4.2	19.9	3.9	0.0	0.6

The differential signal $\beta = 0.5$. The threshold for FDR, uFDP and wuFDP is 0.1. $z_{1-\alpha}$ is 1.65. The number of correlated non-differentially expressed genes varies from 0 to 200. The correlation among these genes varies from 0.6 to 0.9. For wuFDP control, if one hypothesis is rejected, we count it as 1 when we calculate the TP or the FP.

Average sensitivity is defined as the average of ratios between truly declared positives and total positives (i.e. 50). Average specificity is defined as the average of ratios between truly declared negatives and total negatives (i.e. 950). ROC curves are plotted according to the average sensitivity and 1−average specificity for wuFDP, uFDP and FDR controls. The ROC curves shown in Figure 3 also suggest that wuFDP control performs better than uFDP control under the dependence situation. Because the ROC curves correspond to the average sensitivity and the average specificity, FDR control also performs well. However, as we discussed before, FDP for a specific experiment may be much larger than the average value for FDR control. When the dependence is strong, the performance of wuFDP is as good as or even better than FDR control as shown in Figure 3.

3.2 Simulations for marker dependence

The above simulations demonstrate that the uFDP and the wuFDP approaches can control the probability of $Pr(FDP \geq$ a specified value) at a specified level. In addition, the wuFDP control has a better power than the uFDP control. We next consider the dependence among markers.

We simulate marker genotype data for 100 F_2 offsprings from intercross experiments using R package qtl (Broman *et al.*, 2003). For 100 independent genes, 50 of them are differentially expressed. Two scenarios are considered.

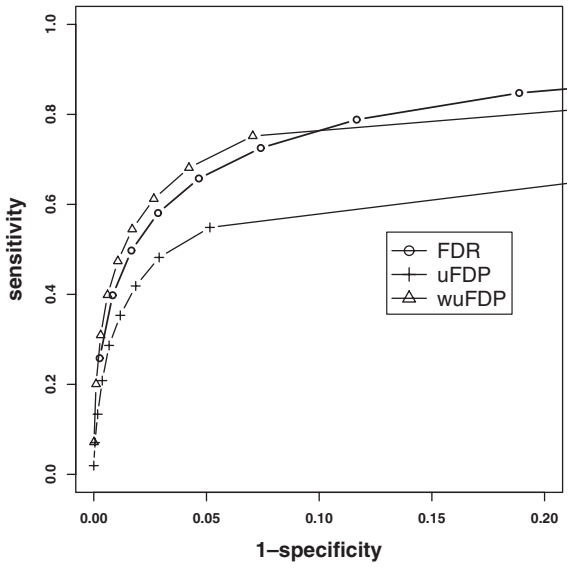

Fig. 3. ROC curves for FDR, uFDP and wuFDP controls. The differential signal $\beta = 0.4$. Two hundred non-differentially expressed gene are correlated with correlation 0.7. $z_{1-\alpha}$ is 1.65 to control $Pr(V/R \geq uFDP)$ and $Pr(\sum w_i v_i / \sum w_i r_i \geq wuFDP)$ less than or equal to 0.05. Triangle symbol line is for wuFDP control, cross symbol line is for uFDP control and circle symbol line is for FDR.

1. Ten markers are on 10 different chromosomes. For marker 1 on chromosome 1, 50 genes are differentially expressed according to the genotypes of marker 1.
2. Compared with scenario 1, we add additional nine markers on chromosome 1. Thus, 10 markers are equally spaced on chromosome 1 with 1 cM distance. Another nine markers are on chromosome 2–10. Marker 1 is associated with 50 genes.

For scenario 2, markers on chromosome 1 are tightly linked. Therefore, any marker called significant on chromosome 1 is a true positive. Otherwise it is a false positive. Table 3 summarizes the average true positives and false positives for 1000 simulations. Compared with scenario 1, many more true positives are identified in scenario 2. However, these additional true discoveries are purely due to the linkage disequilibrium and they cannot provide more information about the association between gene and marker. On the other hand, we include many more false positives in scenario 2. According to the definition of FDR: $FDR = \frac{E(\text{false positives})}{\text{false positives+true positives}} = \frac{E(\sum I(|T_i^0| \geq t))}{\sum I(|T_i| \geq t)}$, given a test statistic threshold t, the increased number of true positives will make the FDR smaller. Correspondingly, a smaller threshold t will be chosen for a particular FDR level (e.g. 0.1). Thus, an arbitrary small threshold t can be chosen to control FDR at a specific level by adding more markers around the casual marker. The small threshold t will result in more false positives (40.6 versus 7.4). According to the definition of uFDP, the increased number of true positives will make the denominator of uFDP bigger and the uFDP will be smaller correspondingly. More false positives are included for scenario 2 (22.2 versus 3.1). But compared to FDR control, uFDP control is a more stringent approach and the number of false positives are smaller (22.2 versus 40.6). For wuFDP, smaller weights will be assigned to those highly correlated true positives. It penalizes the effect of additional markers and the denominator of wuFDP changes little. Correspondingly, the false positive number changes little (6.9 versus 3.5). Note that the correlation between two hypotheses under the null consists of two parts: the correlation between expression-level residuals and the correlation between markers (see Section 2.3). If the markers are close to each other, the correlation between hypotheses will be high. If two genes are associated with the same tested marker, the correlation between hypotheses may not necessarily be high because we use the correlation between expression residuals instead of expression levels themselves.

Table 3. A comparison of FDR, uFDP and wuFDP controls when markers are dependent

Scenario	FDR: TP	uFDP: TP	wuFDP: TP	FDR: FP	uFDP: FP	wuFDP: FP
1	49.2	48.3	48.4	7.4	3.1	3.5
2	488.7	480.3	452	40.6	22.2	6.9

The average true positives (TP) and false positives (FP) are calculated based on 1000 simulations. The differential signal $\beta = 0.5$. The threshold for FDR, uFDP and wuFDP is 0.1. $z_{1-\alpha}$ is 1.65. For wuFDP control, if one hypothesis is rejected, we count it as 1 when we calculate the TP or the FP.

3.3 Real data analysis

We use a human eQTL dataset to demonstrate the usefulness of our proposed methods. Stranger *et al.* (2007) used Illumina's Sentrix Human-6 Expression BeadChips to measure gene expression in B-lymphocyte cells of CEPH Utah individuals. In total, we consider 60 unrelated individuals. There are four replicates for each individual. In the original paper, the raw datasets were background corrected and quantile normalized across replicates of a single individual followed by a median normalization across different individuals. We downloaded the normalized data from the Sanger Institute website (ftp://ftp.sanger.ac.uk/pub/genevar/). We chose the 1000 most variable probes among the total of 47 293 probes for further analysis after excluding probes with extreme outliers. These individuals are the subjects in the HapMap project (Consortium, 2003). Genotypes of 20 281 autosomal SNPs which are in the 5′ UTR region or 3′ UTR region and have minor allele frequency ≥ 0.1 were downloaded from the HapMap website (http://www.hapmap.org/). UTR is an important region for gene transcription regulation. Therefore, we prioritize these SNPs. The genotype is coded as 0, 1 or 2 for homozygous rare, heterozygous and homozygous common alleles. Gene expression and marker genotype data were standardized to have mean 0 and sample variance $1/(n-1)$.

As we know, markers far away on the same chromosome or on different chromosomes are in linkage equilibrium in a homogeneous population, as is the case for the samples considered here. Therefore, $\sum_{k=1}^{n} x_{mk} x_{m'k}$ is very close to 0 for such markers. As a result, the variance–covariance matrix of $v_i's$ can be simplified as a block diagonal matrix with each block i corresponding to hypotheses between those 1000 genes and marker i and its neighboring markers which are within 10 Mb region of marker i. FDR, uFDP and wuFDP procedures as mentioned before were applied to perform the analysis.

Table 4. Gene–marker pairs identified by wuFDP or FDR but not uFDP

Probe	Gene	Marker	*P*-val	*Cis*/*Trans*
GI_21328454−S	HIST2H2AA	rs3761026	1.2×10^{-7}	*trans*
GI_34147330−S	C20orf22	rs1046073	1.3×10^{-7}	*cis*
GI_34147330−S	C20orf22	rs761025	1.3×10^{-7}	*cis*
GI_34147330−S	C20orf22	rs6050626	1.4×10^{-7}	*cis*
GI_34147330−S	C20orf22	rs12428	1.4×10^{-7}	*cis*
GI_13259530−A	SMN2	rs2523454	1.4×10^{-7}	*trans*
GI_11095446−S	HLA-DQA2	rs482194	1.6×10^{-7}	*cis*
GI_11321616−S	DPYSL4	rs7096307	2.1×10^{-7}	*cis*
GI_42661257−S	MGC19764	rs1055636	2.8×10^{-7}	*cis*
GI_38683865−S	RNASET2	rs11787880	3.0×10^{-7}	*trans*
GI_4504212−S	GUCY1A3	rs999917	3.1×10^{-7}	*trans*
GI_27484056−S	LOC284120	rs17199242	3.2×10^{-7}	*trans*
GI_27484056−S	LOC284120	rs17199249	3.2×10^{-7}	*trans*
GI_27484056−S	LOC284120	rs6705406	3.2×10^{-7}	*trans*
GI_27484056−S	LOC284120	rs7600694	3.2×10^{-7}	*trans*
GI_5803140−S	RBPMS	rs16843614	4.1×10^{-7}	*trans*
GI_5803140−S	RBPMS	rs16843618	4.1×10^{-7}	*trans*
GI_16753224−S	RPL14	rs14306	4.3×10^{-7}	*trans*
GI_16753224−S	RPL14	rs7198524	4.3×10^{-7}	*trans*
GI_16753224−S	RPL14	rs4783941	4.3×10^{-7}	*trans*
GI_29029571−I	CD86	rs377298	4.6×10^{-7}	*trans*
GI_42657060−S	LOC401118	rs3131283	4.6×10^{-7}	*trans*

The first seven gene–marker pairs were identified by wuFDP and FDR but not identified by uFDP control. The next 15 gene–marker pairs were identified by FDR but not identified by wuFDP and uFDP controls. If the distance between gene probe and marker is <1 Mb, the gene–marker pair is called *cis*-regulation pair. Otherwise, the pair is called *trans*-regulation pair. Note that another 74 significant gene–marker pairs can be identified by all of the three methods. They are listed in Supplementary Table 1.

Using wuFDP 0.1 as a cutoff and at significance level 0.01 ($z_{1-\alpha}=2.33$), there are 81 significant gene–marker pairs corresponding to 29 genes and 79 markers. Using uFDP control, we can identify 74 eQTLs corresponding to 27 genes and 73 markers. Using FDR = 0.1 as a cutoff, 96 eQTLs corresponding to 36 genes and 94 markers can be identified. As we mentioned in Section 2.3, the FDR approach is Storey's approach (Storey and Tibshirani, 2003) with $\hat{\pi}_0=1$.

The 74 significant gene–marker pairs resulting from uFDP control can be identified by all of the three methods: uFDP, wuFDP and FDR (see Supplementary Table 1). If the distance between the middle point of gene probe and marker is <1 Mb, we call this pair as *cis*-regulation pair. Otherwise, we call the pair as *trans*-regulation pair. Among the 74 pairs, 57 pairs are *cis*-regulation pairs. It was reported that signal of *cis*-regulation eQTL is more abundant and more stable than distal and *trans*-regulation eQTL across statistical methodologies (Stranger *et al.*, 2005). Previous studies also found a significant proportion of *cis*-regulation eQTL pairs (Schadt *et al.*, 2003). Table 4 lists the gene–marker pairs identified by wuFDP and FDR but not identified by uFDP control and the gene–marker pairs identified by FDR but not identified by wuFDP and uFDP controls. Compared with uFDP control, wuFDP control identified seven more eQTL pairs with five of them being *cis*-regulation pairs. Compared with FDR control, wuFDP control missed 15 eQTL pairs with only two of them being *cis*-regulation pairs.

The number of pairs identified depends on the threshold used. Using wuFDP or uFDP 0.05 as a cutoff and at significance level 0.05 ($z_{1-\alpha}=1.65$), there are 72 or 65 eQTL pairs. Using FDR = 0.05 as a cutoff, 81 eQTLs pairs can be identified.

4 DISCUSSION

False discovery control is widely used in genomics and proteomics studies involving large-scale hypothesis tests. Most of the approaches assume independence among genes. However, the dependence among hypotheses may lead to misleading interpretation of FDP. In this article, we proposed a method that not only considers the correlation effect in multiple comparison adjustment, but also utilizes such correlations among hypotheses to improve the power of identifying eQTL. We treat different hypotheses differently according to their correlations with other hypotheses. The effective size of each hypothesis for false discovery control is reflected in its weight. If the dataset shows non-trivial dependence among hypotheses such as in Supplementary Figure 1, we recommend to use the wuFDP control. Otherwise, either the wuFDP or uFDP control can be used since they perform almost the same (as shown in Supplementary Figure 3).

We note that weighted analysis has been proposed in the literature in multiple comparisons. Roeder *et al.* (2006) proposed to assign different weights to *P*-values for association tests between markers and complex diseases. The weights are predefined and based on prior information such as results from linkage analysis. If the linkage study is informative, which means we choose the correct weights, the weighted FDR procedure for the association study will improve power significantly. On the contrary, the power loss is small if the linkage study is uninformative. However, their weights do not address the dependence problem. Cheverud (2001) proposed a simple correction for multiple comparisons in interval mapping. Hypothesis tests are not independent if makers are in linkage disequilibrium. The effective number of independent tests for these tests is used in Bonferroni correction. The effective number is measured by the variance of eigenvalues derived from the observed marker correlation matrix.

In this article, instead of only focusing on marker correlations, we also consider gene correlations. Weights based on the variance–covariance matrix of false discovery number V are considered when we perform multiple comparison adjustment. Genes (or markers) which have lower correlations with other genes (or markers) are assigned with larger weights. On the contrary, genes (or markers) which have high correlations with others are penalized by assigning smaller weights. Through these approaches, we can improve statistical power. More importantly, by assigning smaller weights to markers highly correlated with each other, we can handle the situation where many true positives actually come from the same signal source. The markers are declared significant only because they are in linkage disequilibrium with the causal marker. The augmented true positive number will lead to a smaller test statistic threshold for a predefined FDR level. It will include much more false discoveries. However, wuFDP control can penalize markers highly correlated with each other and the weighted number of true positives can represent the number of true signal sources. Therefore, the wuFDP approach can better control false positives.

uFDP and wuFDP controls can be easily applied to other types of data analysis, such as identifying differentially expressed genes across different conditions which sometimes can be treated as expression biomarkers for diseases. Besides the large-scale gene

expression data, now it is feasible to do high-throughput and high-quality genotyping (e.g. the Affymetrix 500k array set covers about 500k SNPs and costs about $250 per sample). How to handle this large amount of data computationally is still a significant issue. In this article, we choose a computationally feasible weighting scheme. However, the optimality properties for those weights still need to be studied. And eQTL mapping only provides us the directed genetic interaction links from markers to genes. The detailed transcriptional regulatory path needs to be recovered in combination with other types of data.

ACKNOWLEDGEMENTS

We thank two reviewers for their insightful comments.

Funding: This research was supported in part by NIH grants R01 GM59507, N01 HV28286, P30 DA018343, U24 NS051869, P50 HG 002790, and NSF grant DMS 0714817.

Conflict of Interest: none declared.

REFERENCES

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate - a practical and powerful appraoch to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **57**, 289–300.

Brem,R. *et al*. (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science*, **296**, 752–755.

Broman,K. *et al*. (2003) R/qtl: Qtl mapping in experimental crosses. *Bioinformatics*, **19**, 889–890.

Bystrykh,L. *et al*. (2005) Uncovering regulatory pathways that affect hematopoietic stem cell function using "genetical genomics". *Nat. Genet.*, **37**, 225–232.

Chen,L. and Storey,J. (2006) Relaxed significance criteria for linkage analysis. *Genetics*, **173**, 2371–2381.

Chesler,E. *et al*. (2005) Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nat. Genet.*, **37**, 233–242.

Cheverud,J. (2001) A simple correction for multiple comparisons in interval mapping genome scans. *Heredity*, **87**, 52–58.

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

Cohen,B. *et al*. (2000) A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat. Genet.*, **26**, 183–186.

Consortium,T.I.H. (2003) The international hapmap project. *Nature*, **426**, 789–796.

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

Efron,B. (2007) Correlation and large-scale simultaneous significance testing. *J. Am. Stat. Assoc.*, **102**, 93–103.

Hubner,N. *et al*. (2005) Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nat. Genet.*, **37**, 243–253.

Lehmann,E. and Romano,J. (2005) Generalizations of the familywise error rate. *Ann. Stat.*, **33**, 1138–1154.

Morley,M. *et al*. (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature*, **430**, 743–747.

Owen,A. (2005) Variance of the number of false discoveries. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **67**, 411–426.

Roeder,K. *et al*. (2006) Using linkage genome scans to improve power of association in genome scans. *Am. J. Hum. Genet.*, **78**, 243–252.

Schadt,E. *et al*. (2003) Genetics of gene expression surveyed in maize and mouse and man. *Nature*, **422**, 297–302.

Spielman,R. *et al*. (2007) Common genetic variants account for differences in gene expression among ethnic groups. *Nat. Genet.*, **39**, 226–231.

Storey,J. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.

Stranger,B. *et al*. (2005) Genome-wide associations of gene expression variation in humans. *PLoS Genet.*, **1**, e78.

Stranger,B. *et al*. (2007) Population genomics of human gene expression. *Nat. Genet.*, **39**, 1217–1224.

Vysochanskiĭ,D. and Petunin,Y. (1980) Justification of the 3 σ rule for unimodal distributions. *Theor. Probab. Math. Stat.*, **21**, 22–36.