BMC Genomics

RESEARCH Open Access

Simultaneous inferences based on empirical Bayes methods and false discovery rates ineQTL data analysis

Arindom Chakraborty[1,2], Guanglong Jiang[1,2], Malaz Boustani[3], Yunlong Liu[1,2], Todd Skaar[4], Lang Li[1,2,4*]

From The International Conference on Intelligent Biology and Medicine (ICIBM 2013)
Nashville, TN, USA. 11-13 August 2013

Abstract

Background: Genome-wide association studies (GWAS) have identified hundreds of genetic variants associated with complex human diseases, clinical conditions and traits. Genetic mapping of expression quantitative trait loci (eQTLs) is providing us with novel functional effects of thousands of single nucleotide polymorphisms (SNPs). In a classical quantitative trail loci (QTL) mapping problem multiple tests are done to assess whether one trait is associated with a number of loci. In contrast to QTL studies, thousands of traits are measured alongwith thousands of gene expressions in an eQTL study. For such a study, a huge number of tests have to be performed ($\sim 10^6$). This extreme multiplicity gives rise to many computational and statistical problems. In this paper we have tried to address these issues using two closely related inferential approaches: an empirical Bayes method that bears the Bayesian flavor without having much *a priori* knowledge and the frequentist method of false discovery rates. A three-component t-mixture model has been used for the parametric empirical Bayes (PEB) method. Inferences have been obtained using Expectation/Conditional Maximization Either (ECME) algorithm. A simulation study has also been performed and has been compared with a nonparametric empirical Bayes (NPEB) alternative.

Results: The results show that PEB has an edge over NPEB. The proposed methodology has been applied to human liver cohort (LHC) data. Our method enables to discover more significant SNPs with FDR<10% compared to the previous study done by Yang et al. (*Genome Research*, 2010).

Conclusions: In contrast to previously available methods based on p-values, the empirical Bayes method uses local false discovery rate (lfdr) as the threshold. This method controls false positive rate.

Introduction

Genome-wide association studies (GWASs) have done a remarkable progress in searching for susceptibility genes. In GWAS, instead of one gene at a time, variation across the entire genome is tested for association with disease risk. GWASs exploit the linkage disequilibrium (LD) relationships among single nucleotide polymorphisms (SNPs), making it possible to assay genome by testing a finite number of SNPs. Till date, the signals that can be discovered through GWAS has not been reported exhaustively. It is important to annotate SNPs information on expression for the better understanding of the genes and mechanisms driving the association. In many situations, there are more common variants truly associated with disease. These variants are highly likely to be expression quantitative trait loci (eQTLs). eQTLs are derived from polymorphisms in the genome that result in differential measurable transcript levels. Microarrays are used to measure gene expression levels across genetic mapping populations. For at least a subset of complex disorders, gene expression levels could be used as a surrogate/biomarker for classical phenotypes. The gene underlying the eQTL is considered to be an excellent candidate for phenotypic QTL.

* Correspondence: lali@iu.edu
[1] Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, Indiana 46202, USA
Full list of author information is available at the end of the article

eQTL mapping is a statistical technique to locate genomic intervals, that are likely to regulate the expression of each transcript, by correlating quantitative measurements of mRNA expression with genetic polymorphisms segregating in a population. In a GWAS, millions of SNPs are tested at once. Associations that initially appear to be significant must be statistically adjusted to account for the large number of tests being performed. A large number of false positives will result in if this correction is ignored. The multiple-testing correction, however, sets a very high threshold for genome-wide significance, on the order of 5×10^{-8} when a million SNPs are tested. In the vast majority cases, however, association studies have achieved only limited success. Large sample sizes are needed to achieve sufficient statistical power to detect risk alleles with effects weak enough to have escaped detection in the past; the disease risk alleles identified by GWASs so far do have weak effects, each with odds ratios of 1.1 or 1.2 [1].

Two closely related inferential procedures for multiple testing have been discussed in this work-afrequentist approach based on Benjamini and Hochberg's ([2]) false discovery rate procedure, and an empirical Bayes methodology developed in Efron et al. [3,4]. These two methods are not only very closely related, they can be used to support each other. In a classic two-sample problem in a microarray experiment, these approaches have been discussed by Efron and Tibshirani[5]. However, they have considered nonparametric empirical Bayes (NPEB) model. Parametric Bayesian modeling has been considered by Newton et al. [6], Lee et al. [7], Kendziroski et *al.* [8-10], Gelfond et *al.* [11]. Hierarchical models like gamma-gamma [6] or lognormal-normal [8] are used quite often in PEB procedures. These models suffer from a serious drawback that the variation is constant among genes. An extension has been done to these models by considering gene specific variations[12]. The application of empirical Bayes has been somehow not very common in literature. The obvious reason is that, experimenters have not brought us many data sets having the parallel structure necessary for empirical Bayes to do its stuff. Because of the recent surge in high-throughput ([13]) technologies and genome projects, many genome studies are now underway. These studies have become a major data generator in the post-genomics era. Empirical Bayes procedures seem to be particularly well-suited for combining information in expression data.

One of the fundamental statistical problems in microarray gene expression analysis is the need to reduce dimensionality of the transcripts. This can be achieved by identifying differentially expressed (DE) genes under different conditions or groups. Regulatory network can be obtained by associating differential expressions with the genotype of molecular markers. It is possible to have a large number of DE genes that influences a certain phenotype while their relative proportion is very small. It is very important to identify these DE genes from among the number of recorded genes [6,7,9,14,15]. Empirical Bayes methods provide a natural approach to reduce the dimensionality significantly [16,17]. Following the empirical Bayes approach DE genes are identified using the posterior probability for differential expression. EB approaches detect a DE gene by sharing information across the whole genome.

The development of the empirical Bayes methodologies that improve the power to detect DE genes essentially reduces to the choice of whether gene-specific effects should be modeled as fixed or random [18]. Both mean and error variance can be of either of these two: fixed or random. Fixed mean and random error variance has been considered by Wright and Simon [19] and Cui et al. [20] whereas Lonnstedt et al. [21], Tai and Speed [22], Lonnstedt and Speed [23] have considered both the parameters to be random. Random mean effect with homogeneous fixed error variance has been considered by Newton et al. [6,24], Kendziroski et al. [9] and Kendziroski et al. [10]. However an extension to this fixed error variance has been considered by Gelfond et al. [11]. They have considered discrete uniform prior for the variance component.

The paper is organized as follows. In the Methods section we introduce the necessary notations for our additive genetic model along with the notions of false discovery rate (fdr). In this section we have tried to establish the relationship between fdr and empirical Bayes. Methods section also describes, the proposed Expectation/Conditional Maximization Either (ECME) (Liu and Rubin [25]) in details. This algorithm generalizes the Expectation-Maximization algorithm with better convergence rate. A simulation study has been performed and described in the Results section. We show that proposed parametric empirical Bayes performs better compared to nonparametric empirical Bayes in terms of controlled fdr. In the Results section, as an application, we have applied the proposed methodology to the Liver Cohort (LHC) dataset. We conclude the article the Discussion section.

Methods

In a microarray experiment, we obtain several thousand expression values, one or many for each gene. These studies offer an unprecedented ability to do large-scale studies of gene expression. Let us define $G_i i = 1.....l$ as the genomic marker(i.e. SNP), and $T_j(j = 1......J)$ as the transcripts. The identified eQTLs refer to the significant Gs that are associated with Ts. These associations can be found using a test statistics based on all n samples. The genetic model for this association can be one of the three models: dominant, recessive and additive. Under the dominance model, we can have two genotypes for each of the SNPs. However for an additive model, three genotype groups are available. A transcript T_j is assumed to be associated with marker G_i if the mean level of expression of transcript T_j for one genotype group is different from that of the other genotype group corresponding to that marker. Let $\mu_{T,G}^{(1)}$ and $\mu_{T,G}^{(0)}$ be the group means corresponding to the genotypes G_i. To test the hypothesis $H_0 : \mu_{T,G}^{(1)} = \mu_{T,G}^{(0)}$, a few test statistics are proposed for microarray data analysis[26]. The present work is based on the statistic proposed by Efron et *al.* [4]. The test statistic is defined as

$$Z_{ij} = \frac{\bar{x}_{T,G}^{(1)} - \bar{x}_{T,G}^{(0)}}{\left(a_{0i} + S_{ij}\right)} \tag{1}$$

where S_{ij} is the usual standard deviations and a_{0i} is defined to minimize the difference in the coefficient of variation of Z_{ij} within classes of genes with approximately equal variance. A drawback of calculating a_{0i} is the computational cost. Note that if $a_{0i} = 0$, this reduces to usual t-statistic. Here a_{0i} is considered to be 90^{th} percentile of all S_{ij} values (Efron el *al.* [4]).

When expression measurements between two groups are compared for any transcript, the observations are partitioned into two user defined groups of sizes n_1 and n_2 with $n_1 + n_2 = n$. If there is no significant difference between the group means, the transcript is assumed to be equivalently expressed (EE). On the contrary, if significant difference is observed, the transcript is termed as differentially expressed (DE). For any transcript T_j and SNP G_i it may be either of these two: DE or EE. This uncertainty can be modeled by a mixture of two distributions as follows:

$$f\left(Z_{ij}|\boldsymbol{\theta}\right) = \pi_o f_0\left(Z_{ij}|\boldsymbol{\theta}\right) + \pi_1 f_1\left(Z_{ij}|\boldsymbol{\theta}\right) \tag{2}$$

where π_0 is the mxining proportion of EE transcripts and $\pi_1 = (1 - \pi_0)$ is the proportion of DE transcripts, $\boldsymbol{\theta}$ is a vector parameters involved to characterize the distributions. Let F_i be the minor allele frequency of the ith SNP then we model the distribution of Z_{ij} as a mixture model of the form:

$$\Pr(Z_{ij}|F_i) \propto \left[f_0\left(Z_{ij}|F_i\right)\right]^{1-\delta_{ij}}\left[f_1\left(Z_{ij}|F_i\right)\right]^{\delta_{ij}} \tag{3}$$

where $f_1(.)$ denotes the distribution of Z_{ij} for nonzero associations between G_i and T_j and $f_0(.)$ denotes the distribution of Z_{ij} for the zero associations. δ_{ij} isdefined as

$$\delta_{ij} = \begin{cases} 1 \; \textit{if nonzero association is present} \\ 0 \; \textit{if zero association is present} \end{cases}$$

For any transcript and any SNP there may be three possible relations: no association, positive association and negative association. Extending the idea of two component mixture model, the distribution of the test statistics is modeled by the following mixture model:

$$f\left(Z_{ij}|\psi_i, F_i\right) = \sum_{k=0}^{2} \pi_{ik} f_k\left(Z_{ij}; \mu_k, \tau_k^2, \nu_k\right) \tag{4}$$

Where

$$\psi_i = \left(\boldsymbol{\pi}'_i, \boldsymbol{\theta}'_i, \boldsymbol{\nu}'_i\right)$$

$$\boldsymbol{\pi}_i = (\pi_{0i}, \pi_{1i}) \quad \boldsymbol{\theta}_i = \left(\mu_{1i}, \mu_{2i}, \tau_{1i}^2, \tau_{2i}^2\right) \quad \boldsymbol{\nu}_i = (\nu_{1i}, \nu_{2i})$$

with $\mu_{0i} = 0, \tau_{0i}^2 = 1$. Mixing proportions π_{ik} are non-negative constantsand sum to one for fixed i. $f_0(.)$ corresponds to distribution for no associationwhereas $f_1(.)$ and $f_2(.)$ correspond to distributions related to positive and negativeassociation respectively. In a recent work, Noma and Matsui [27], have used semiparametric hierarchical mixture model where the distribution of mean expression level of a transcript is considered to be a three-component mixture distribution.

Full Bayesian analysis of (4) will require prior specifications of $\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\nu}, f_0(Z)$ and $f_1(Z)$. However, one can use the massively parallel structure of microarray data to estimate an empirical Bayes estimate of the posterior probability. These huge data motivates to be quite empirical rather than specifying *a-priori* models in favor of data-based investigations [27].

Empirical Bayes, false discovery rates (fdr) and local false discovery rate (lfdr)

False discovery rate (fdr) is defined as the expected proportion of errors committed by falsely rejecting null hypotheses. Benjamini and Hochberg's [2]fdr criterion has very close relation with the empirical Bayes analysis. This relation improved the connection between Bayesian and frequentist testing theory. The close connection between fdr and the empirical Bayes methodology follows directly from Bayes theorem and this has been established by the "Equivalence theorem"[28]. Tail area rejection regions like $\{Z_{ij} < z\}$ are common in the frequentist framework. According to this theorem, if the tail area rejection region is taken to be as large as possible subject to the constraint that the estimated Bayes proportions of false discoveries is less than α, then the frequentist expected proportion of false discoveries is also less than α.

The empirical Bayes approach suggests a local version of the fdr called local false discovery rate (*lfdr*). The Bayes probability that a transcript T_j for SNP G_i is "EE" given the test statistic Z_{ij}, is known as $lfdr(Z_{ij})$ and it is defined as

$$lfdr\left(Z_{ij}\right) \equiv \Pr\left(T_j \textit{ is } \mathrm{EE}|Z_{ij}\right) = \pi_{i0} f_0(Z_{ij})/f(Z_{ij})$$

Analytically, *fdr* is a conditional expectation of *lfdr* defined as

$$fdr\left(Z_{ij}\right) = \int_{-\infty}^{Z_{ij}} lfdr\,(Z)\,f\,(Z)\,dZ / \int_{-\infty}^{Z_{ij}} f\,(Z)\,dZ = E_f\{lfdr(Z|Z \leq Z_{ij})\}$$

For the above set up in (3), $1 - \delta_{ij}$ represents the local false discovery rate (lfdr) and fdr can be estimated:

$$\hat{\delta}_{ij} = \frac{\pi_{i1} f_1\left(Z_{ij}\right)}{(1 - \pi_{i0}) f_0\left(Z_{ij}\right) + \pi_{i1} f_1\left(Z_{ij}\right)};\, lfdr\left(Z_{ij}\right) = 1 - \hat{\delta}_{ij} \quad (5)$$

and hence

$$fdr\left(Z_{ij}\right) = \frac{(1 - \pi_{1i}) \int_{|Z_{ij}|}^{\infty} f_0\,(x)\,dx}{(1 - \pi_{1i}) \int_{|Z_{ij}|}^{\infty} f_0\,(x)\,dx + \pi_{1i} \int_{|Z_{ij}|}^{\infty} f_1\,(x)\,dx} \quad (6)$$

ν_{0i} is estimated bypermutation method (Efron et *al.* [4]) and p_{oi} is estimated from the nonnegative constraint

$$p_{0i} \leq \min_Z \frac{f_i(Z)}{f_{i0}(Z)}$$

All other parameters will be estimated by EM algorithm assuming $f_{i0}(.)$ to be known. There are some practical difficulties with the lfdr that relies on densities. The estimation of null becomes more problematic in the far tails. It is relatively easier to work with cumulative distribution function than work with densities. Identification of discoveries by lfdr may not be reproducible for a new data. Therefore, even in empirical Bayes framework, fdr should be preferred.

Nonparametric empirical Bayes (NPEB)

The main difference between parametric empirical Bayes (PEB) and nonparametric empirical Bayes (NPEB) is the way in which $f_1(.)$and$f_2(.)$ are treated. In PEB model, the functional form of $f_1(.)$and$f_2(.)$ are known, i.e., we have a parametric family of priors. In contrast, the NPEB does not assume the functional form to be known. Though NPEB methods are quite powerful, these are more suitable for large sample analyses. To compute the fdr under NPEB setup, we have followed the algorithm proposed by Efron et al. [4].

ECME algorithm

To fit a mixture model, EM algorithm is widely used. In case of t distribution the mean parameter μ and variance component τ^2 can easily be estimated by EM algorithm assuming that degrees of freedom ν is known. However when ν is unknown EM still can be used as demonstrated by Lange, Little and Taylor [29]. But this method appears to be very slow (Liu and Rubin [30]) and an extension has been proposed by Meng and Rubin [31] as ECM algorithm. This is a generalization of EM algorithm where the E step remains the same butthe M step is replaced by CM (constrained or conditional maximization) step. ECM algorithm is basically a generalized EM (GEM) as shown by Meng and Rubin [31]. Incidentally, the rate of convergence, in terms of iterations, for this ECM algorithm is slower compared to EM. To overcome this computational problem, Liu and Rubin [30] propose an efficient algorithm ECME which is again an extension of ECM algorithm. Though this is not a GEM, it converges faster.

For the i -th SNP, the complete data is defined as

$$D_{iC} = (Z_{ij}, \delta_{ijk1}, \delta_{ijk2} \ldots\ldots\ldots \delta_{ijkn}, U_{i1}, U_{i2} \ldots\ldots U_{in})$$

where

$$\delta_{ijks} = \begin{cases} 1 \ \textit{if s th observation of } Z_{ij} \in \textit{kth component} \\ 0 \ \textit{otherwise} \end{cases}$$

and U_i s are independently distributed gamma variables.

McLachlan and Krishnan [32] have already discussed the application of the EM algorithm for ML estimation in case of single component t distribution. In ECME algorithm, this result has been extended to cover the present set up of a 3-component mixture of t distribution. For the sake of brevity, in this section we omit the suffix *ij* for all the variables. To define t distribution with mean μ, variance τ^2 and degrees of freedom ν, we proceed as follows:

$$\textit{If } Z|U = u, \delta_{ks} = 1 \sim N\left(\mu, \frac{\tau^2}{u}\right) \textit{and } U \sim \Gamma\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$$

then marginally, $Z \sim t(\mu, \tau^2, \nu)$.

Following the above definition, the complete data likelihood L_{iC} can be factorized a product of three terms-marginal densities of δ s, the conditional densities of $U|\delta$, and conditional densities of $Z|U = u, \delta$. In notation, the log-likelihood of the complete-data can be expressed as

$$\log L_C(\boldsymbol{\psi}) = \log L_{1C}(\boldsymbol{\pi}) + \log L_{2C}(\boldsymbol{\nu}) + \log L_{3C}(\boldsymbol{\theta}) \quad (7)$$

where

$$\log L_{1C}\,(\boldsymbol{\pi}) = \textstyle\sum_{\mathrm{k=0}}^{2} \cdot \sum_{\mathrm{s=1}}^{\mathrm{n}} \delta_{ks} \log \pi_k \quad (8)$$

$$\log L_{2C}\,(\boldsymbol{\nu}) = \textstyle\sum_{\mathrm{k=0}}^{2} \cdot \sum_{\mathrm{s=1}}^{\mathrm{n}} \delta_{ks}\{-\log \Gamma\left(\frac{\nu_k}{2}\right) + \frac{1}{2}\nu_k \log \Gamma\left(\frac{\nu_\mathrm{k}}{2}\right) + \frac{1}{2}\nu_k\left(\log u_s - u_s\right) - \log u_s\} \quad (9)$$

and

$$\log L_{3C}\,(\boldsymbol{\theta}) = \textstyle\sum_{\mathrm{k=0}}^{2} \cdot \sum_{\mathrm{s=1}}^{\mathrm{n}} \delta_{ks}\{-\frac{1}{2}\pi_k \log(2\pi) - \frac{1}{2}\tau_k^2 - \frac{1}{2}\frac{u_s(z - \mu_k)^2}{\tau_k^2}\} \quad (10)$$

E-Step

To compute the E-step of the proposed algorithm, at (t+1) th step we need to calculate $Q(\boldsymbol{\psi}; \boldsymbol{\psi}^{(t)})$, the current conditional expectation of the complete-data log likelihood function $\log L_C\,(\boldsymbol{\psi})$. From equation (4) to (7), we can write

$$Q\left(\boldsymbol{\psi}; \boldsymbol{\psi}^{(t)}\right) = Q_1\left(\boldsymbol{\pi}; \boldsymbol{\psi}^{(t)}\right) + Q_2\left(\boldsymbol{\nu}; \boldsymbol{\psi}^{(t)}\right) + Q_3\left(\boldsymbol{\theta}; \boldsymbol{\psi}^{(t)}\right) (11)$$

where

$$Q_1\left(\boldsymbol{\pi};\boldsymbol{\psi}^{(t)}\right) = \sum_{k=0}^{2}\cdot\sum_{s=1}^{n} E_{\psi^{(t)}}\left(\delta_{ks}—z_s\right) = \sum_{k=0}^{2}\cdot\sum_{s=1}^{n}\xi_{ks}^{(t)}\log\pi_k \tag{12}$$

and

$$\xi_{ks}^{(t)} = \frac{\pi_k^{(t)} f\left(Z_s;\mu_k^{(t+1)},\tau^2{}_k^{(t+1)},\nu_k^{(t+1)}\right)}{f(Z;\psi^{(t+1)})} \tag{13}$$

which is the posterior probability that Z belongs to the k-th component of the mixture based on current fit $\boldsymbol{\psi}^{(t)}$.

Similarly,

$$Q_2\left(\boldsymbol{\nu};\boldsymbol{\psi}^{(t)}\right) = \sum_{k=0}^{2}\cdot\sum_{s=1}^{n}\xi_{ks}^{(t)}[-\log\Gamma\left(\frac{\nu_k}{2}\right)+\frac{1}{2}\nu_k\log\Gamma\left(\frac{\nu_k}{2}\right)+\frac{1}{2}\nu_k\{\sum_{s=1}^{n}\left(\log u_{ks}{}^{(t)}-u_{ks}^{(t)}\right)+ \psi\left(\frac{\nu_k^{(t)}+1}{2}\right)-\log(\frac{\nu_k^{(t)}+1}{2})\}] \tag{14}$$

Where

$$u_{ks}^{(t)} = \frac{\nu_k^{(t)}+1}{\nu_k^{(t)}+\left(Z_s-\mu_k^{(t)}\right)^2/\tau_k^{(t)}} \tag{15}$$

$\psi(.)$ is a digamma function and

$$Q_3\left(\boldsymbol{\theta};\boldsymbol{\psi}^{(t)}\right) = \sum_{k=0}^{2}\cdot\sum_{s=1}^{n}\xi_{ks}^{(t)}[-\frac{1}{2}\log(2\pi)+\frac{1}{2}\log u_{ks}{}^{(t)}-\frac{1}{2}u_{ks}^{(t)}((Z_s-\mu_k)/\tau_k)^2\}] \tag{16}$$

CM-step

In usual M-step parameters $\boldsymbol{\pi}$, $\boldsymbol{\nu}$, $\boldsymbol{\theta}$ can be estimated by considering equations (10) - (12) independently. The new updates for $\boldsymbol{\pi},\boldsymbol{\theta}$ can be obtained as a closed form solution whereas for $\boldsymbol{\nu}$ an iterative procedure may be used using the following equations:

$$\pi_k^{(t+1)} = \sum_{s=1}^{n}\frac{\xi_{ks}^{(t)}}{n} \tag{17}$$

$$\mu_{\mathrm{k}}^{(t+1)} = \left(\sum_{s=1}^{n}\xi_{ks}^{(t)}u_{ks}^{(t)}Z_s\right)\Big/\sum_{s=1}^{n}\xi_{ks}^{(t)}u_{ks}^{(t)} \tag{18}$$

$$\tau_{\mathrm{k}}^{(t+1)} = \sum_{s=1}^{n}\xi_{ks}^{(t)}u_{ks}^{(t)}\left(\frac{Z_s-\mu_k}{\tau_k}\right)^2\Bigg/\sum_{s=1}^{n}\xi_{ks}^{(t)} \tag{19}$$

and $\nu_{\mathrm{k}}^{(t+1)}$ is the solution of the following equation

$$\{-\psi\left(\frac{\nu_{\mathrm{k}}}{2}\right)+\log\left(\frac{\nu_{\mathrm{k}}}{2}\right)+1+\frac{1}{n_k^{(t)}}\sum_{s=1}^{n}\xi_{ks}^{(t)}\log\Gamma\left(\frac{\nu_{\mathrm{k}}}{2}\right)+\frac{1}{2}\nu_k\{\sum_{s=1}^{n}\left(\log u_{ks}{}^{(t)}-u_{ks}^{(t)}\right)+ \psi\left(\frac{\nu_{\mathrm{k}}^{(t)}+1}{2}\right)-\log(\frac{\nu_{\mathrm{k}}^{(t)}+1}{2})\}=0 \tag{20}$$

To get an efficient algorithm, let us partition $\boldsymbol{\psi}$ as $(\boldsymbol{\psi}'_1,\boldsymbol{\psi}'_2)'$ where $\boldsymbol{\psi}_1$ contains all the parameters except parameters corresponding to degree of freedom of t-distributions. The above M-step is replaced by two CM-steps, as follows.

CM-Step 1. Keeping $\boldsymbol{\psi}_2$ fixed, i.e. $\boldsymbol{\nu}$ is fixed at $\boldsymbol{\nu}^{(t)}$, maximize $\mathrm{Q}\left(\boldsymbol{\psi};\boldsymbol{\psi}^{(t)}\right)$ to get $\boldsymbol{\psi}_1^{(t+1)}$

CM-Step 2. Now fix $\boldsymbol{\psi}_1$ at $\boldsymbol{\psi}_1^{(t+1)}$ and calculate $\boldsymbol{\psi}_2^{(t+1)}$ by maximizing $\mathrm{Q}\left(\boldsymbol{\psi};\boldsymbol{\psi}^{(t)}\right)$

Furthermore to make the algorithm more efficient, after the first CM-step, we replace the E-step with $\boldsymbol{\psi}=(\boldsymbol{\psi}_1^{(t+1)'},\boldsymbol{\psi}_2^{(t)'})'$ instead of $\boldsymbol{\psi}=(\boldsymbol{\psi}_1^{(t)'},\boldsymbol{\psi}_2^{(t)'})'$.

Results

Simulation study

To assess the proposed methodology, a small sample simulation study has been performed. This gives an idea whether or not the parameters are well estimated and most importantly, they provide information of false discovery rates.

First we simulated a dominant model with 10,000 transcripts and 10 SNPs. The equivalently expressed (EE) transcripts are generated from N(0,1) after log-transformation. We have simulated the data under three choices of proportions of differentially expressed (DE) transcripts (p_1). We have taken p_1 to be (0.01, 0.05, 0.10). If the transcript is DE, it has to be generated from N(4,0.5) after log-transformation. The controlled fdr are also assumed to be (0.01, 0.05, 0.10) for these data sets. For p_1 = 0.05, the simulated data is given in Figure 1.

The impact of minor allele frequency (MAF) on the distributions under null has also been studied. Under null, for a t-distribution, the only parameter to be estimated is its degrees of freedom. The comparison has been made by computing different quantiles for six choices of MAFs. For the lower quantiles, they almost overlapped with each other. Very small deviations are observed for upper quantiles (Figure 2).

For the 10 SNPs, we fitted the null distribution using permutation method in a balanced way. From each group, randomly selecterd 35 samples are shifted from one group to the other and the value of the statistic is noted. This process is repeated 40 times and histograms are plotted. From the histograms, the degrees of freedom corresponding to the null distribution for each SNP is estimates. To get an idea about the goodness-of-fit, Q-Q plots are done (Figure 3). These plots show that the null distribution is well approximated by the standardized t-distribution with appropriate degrees of freedom.

Parameters related to the mixture model (4) are estimated using proposed ECME algorithm after estimating the null distribution using permutation method. Then FDR is computed under both proposed parametric empirical Bayes and nonparmetic empirical Bayes setup and the result is given in Table 1.

It is evident from the above table that the nonparmateric empirical Bayes is much conservative compared to its parametric alternative. For parametric set up, the true FDR is very much close to the controlled one,

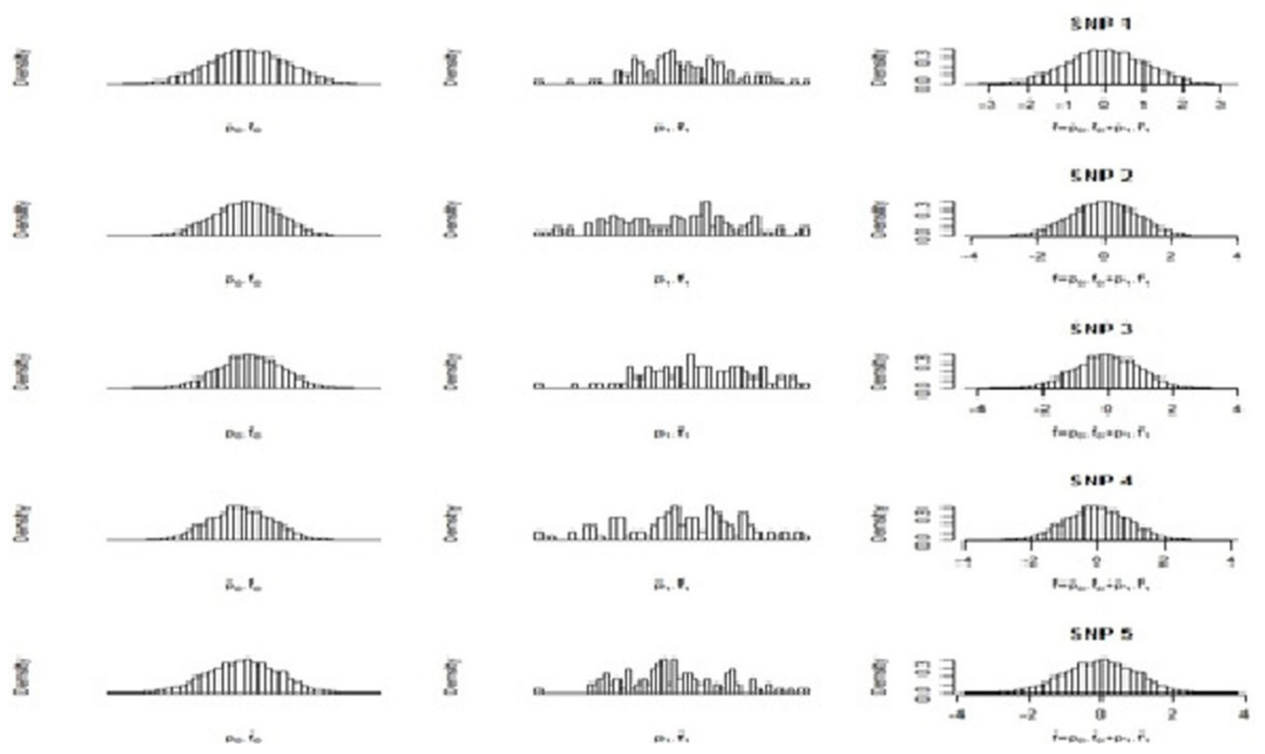

Figure 1 A part of the simulated data for $p_1 = 0.05$.

whereas, for nonparametric empirical Bayes these values are not so close as the true fraction of DE transcripts increases.

HLC data analysis

We applied the empirical Bayes model to analyze a sequencing data publicly available. In the current study, we have started with liver tissue data of 213 Caucasian samples from apreviously described human liver cohort (LHC) (Yang et al. [33]). To get the genotypes and gene expression profiles, DNA and RNA have been isolated. Illumina platform is used to get the expressions. After putting some filtration (MAF>5%, HWE<10^{-5},) we are left with 173 samples, 472,000 SNPs and 30,000 expressions.

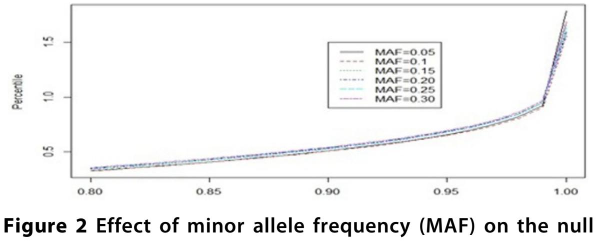

Figure 2 Effect of minor allele frequency (MAF) on the null distribution. Only upper quantiles (from 80%) have been considered as lower quantiles showing almost no difference.

The distribution of minor allele frequency (MAF) over SNPs is given in the histogram (Figure 4). For all possible SNP-transcript combinations, test statistic, Z_{ij} s are computed. We fit the mixture model using the ECME algorithm in R 2.15.1 after estimating the null distribution using permutation method. However, due to high dimension data, it becomes very difficult to fit a mixture model using the proposed algorithm. For the sake of parsimony, we further filtered the data and ECME algorithm is used for only top SNPs with $p - value < 10^{-3}$. For these top SNPs, the mixture model is fitted and estimates are obtained. To compute lfdr and FDR from (5) and (6) respectively, these estimates are used.

Conclusion

To compare our result with [33], we focus on 18 of the 54 P450 genes used in the study. These are CYP3A5, CYP2D6, CYP4F12, CYP2E1, CYP2U1, CYP1B1, CYP2C18, CYP4F11, CYP4V2, CYP2F1, CYP39A1, CYP26C1, CYP2C19, CYP2C9, CYP2S1, CYP46A1, CYP4A11 and CYP4X1.However our method fails to identify a single SNP with FDR<10% for CYP2R1 and that gene symbol has been excluded from the table (Table 2). It can be seen from the table (Table 2) that for a threshold of 10% FDR number of significant eQTL pairsis4916.Since we have considered only top SNPs, this may be an overestimate. SNPs which are within <1-Mb distance from gene location are defined as

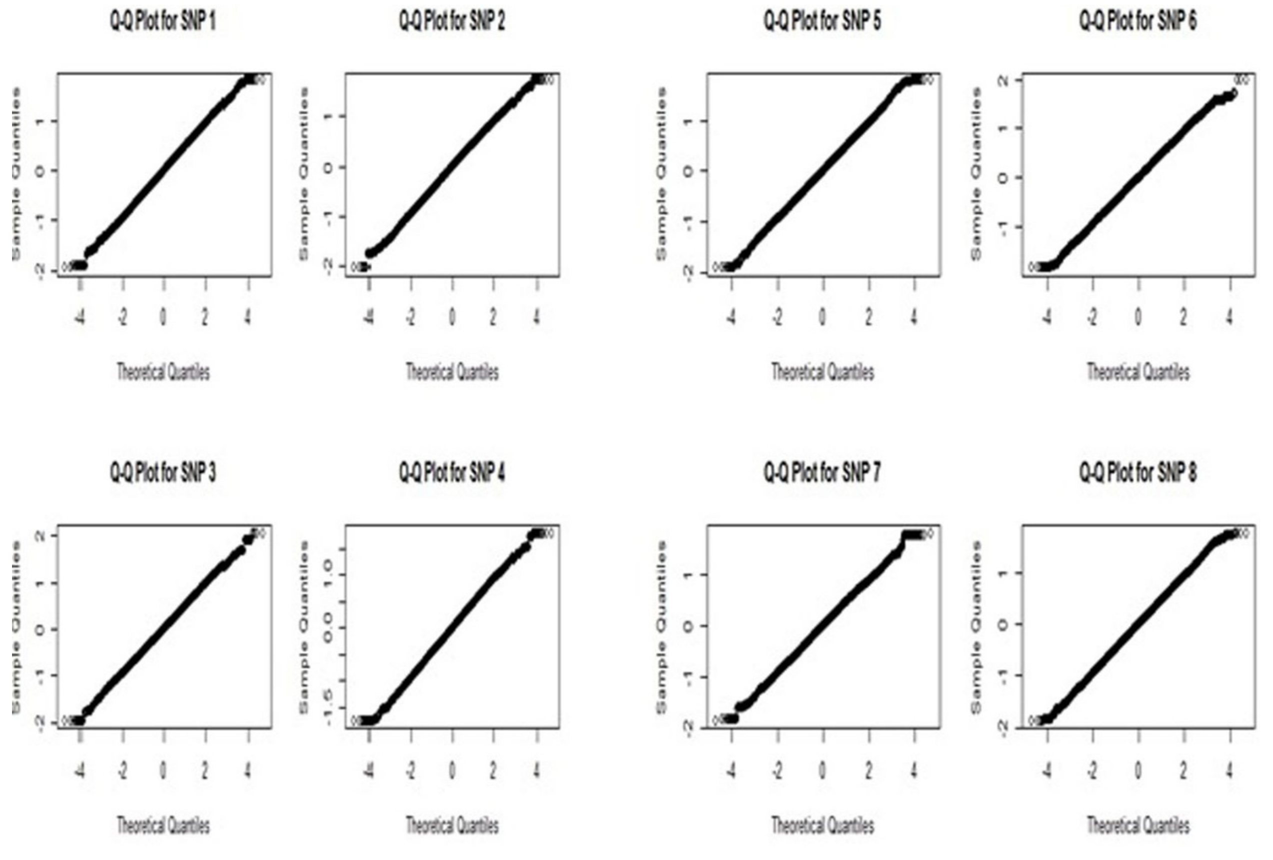

Figure 3 QQ-plot for eight SNPs.

Table 1 The True FDR Performance of Controlled FDR in EB Models

True fraction of DE	Controlled FDR					
	Nonparametric empirical Bayes			Parametric empirical Bayes		
	0.01	0.05	0.10	0.01	0.05	0.10
0.01	0.004	0.029	0.067	0.005	0.042	0.090
0.05	0.006	0.041	0.079	0.006	0.045	0.094
0.10	0.007	0.043	0.087	0.008	0.047	0.097

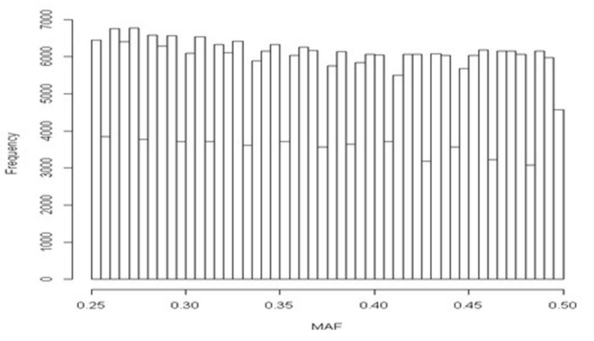

Figure 4 Minor allele frequency (MAF) distribution. X axis corresponds to minor allele frequency 25% to 50%.

cis-SNPs. It is interesting to note that, among these 18 genes, the first five (CYP3A5, CYP2D6, CYP4F12, CYP2E1 and CYP2U1) having more than 40 cis-SNPs. In all cases FDR based analysis results in identifying more cis-SNPs for these 18 genes compared to that of Yang et al. (2010) [33].

Discussion

In contrast to previously available methods based on p-values, the empirical Bayes method uses local false discovery rate (lfdr) as the threshold. This method controls false positive rate. For a particular SNP, the lfdr is computed for the site-specific evidence whereas the FDR averages over other sites with stronger evidence. There are some limitations of using FDR which may result in

Table 2 Number of eQTL pairs after crossing the threshold of FDR

Gene symbol	No. of SNPs (FDR<10%)	No. of cis-SNP	No. of cis-eSNP (FDR<10%) by Yang et al. (2010)
CYP3A5	263	62	56
CYP2D6	264	67	54
CYP4F12	392	55	46
CYP2E1	130	45	31
CYP2U1	549	45	26
CYP1B1	168	21	13
CYP2C18	90	13	9
CYP4F11	169	15	7
CYP4V2	159	25	3
CYP2F1	324	10	2
CYP39A1	448	17	2
CYP26C1	154	29	1
CYP2C19	356	7	1
CYP2C9	413	20	1
CYP2S1	319	10	1
CYP46A1	430	7	1
CYP4A11	461	4	1
CYP4X1	151	3	1

misleading inferences in genome studies. In such a situation, it is better to use lfdr which is a bit difficult to estimate compared to FDR.However there is still one computational problem which needs much attention. Due to the high dimensionality in the data, sometimes existing algorithms fail. This necessitates the need to find some more efficient algorithms. The choice of threshold FDR value is an important deciding factor in such studies. It would be interesting to see, how number of cis-SNPs vary with the change in FDR threshold. In this way FDR criterion can be used to estimate number of SNPs that we may need to consider.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This work is supported by the U.S. National Institutes of Health grants R01 GM74217 (Lang Li) and AHRQ Grant R01HS019818-01 (MalazBoustani)

Declarations

The publication costs were funded by the authors through P50 CA113001 (Huang, T.M.), R01 GM088076 (Skaar, T.), R01 HS019818 (Dexter).

This article has been published as part of *BMC Genomics* Volume 14 Supplement 8, 2013: Selected articles from the International Conference on Intelligent Biology and Medicine (ICIBM 2013): Genomics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/14/S8.

Authors' details

[1]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, Indiana 46202, USA. [2]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, Indiana 46202, USA. [3]Regenstrief Institute, Indianapolis, Indiana 46202, USA. [4]Division of Clinical Pharmacology, Department of Medicine, Indiana University School of Medicine, Indianapolis, Indiana 46202, USA.

Published: 9 December 2013

References

1. Liu , Chunyu : **Brain expression quantitative trait locus mapping informs genetic studies of psychiatric diseases.** *Neurosci Bull* 2011, **27(2)**:123-133.
2. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *Journal of the Royal Statistical Society. Series B (Methodological)* 1995, 289-300.
3. Efron B, Storey J, Tibshirani R: **Microarrays, empirical Bayes methods, and false discovery rates.** *Stanford Technical Report* 2001.
4. Efron B, Tibshirani R, Storey JD, Tusher V: **Empirical Bayes analysis of a microarray experiment.** *Journal of the American statistical association* 2001, **96(456)**:1151-1160.
5. Efron B, Tibshirani R: **Empirical Bayes methods and false discovery rates for microarrays.** *Genetic epidemiology* 2002, **23(1)**:70-86.
6. Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW: **On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data.** *Journal of Computational Biology* 2001, **8**:37-52.
7. Lee MLT, Kuo FC, Whitmore GA, Sklar J: **Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations.** *Proceedings of the National Academy of Sciences* 2000, **97(18)**:9834-9839.
8. Kendziorski CM, Zhang Y, Lan H, Attie A: **The efficiency of MRNA pooling in microarray experiments.** *Biostatistics* 2003, **4**:465-477.
9. Kendziorski CM, Newton MA, Lan H, Gould MN: **On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles.** *Stat Med* 2003, **22**:3899-33914.
10. Kendziroski CM, Chen M, Yuan M, Lan H, Attie AD: **Statistical methods for expression quantitative trait loci (eQTL) mapping.** *Biometrics* 2006, **62(1)**:19-27.
11. Gelfond JAL, Ibrahim JG, Zou F: **Proximity Model for Expression Quantitative Trait Loci (eQTL) Detection.** *Biometrics* 2007, **63**:1108-1116.
12. Lo K, Gottardo R: **Flexible empirical Bayes models for differential gene expression.** *Bioinformatics* 2007, **23(3)**:328-335.
13. Sánchez-Linares I, Pérez-Sánchez H, Cecilia JM, García JM: **High-Throughput parallel blind Virtual Screening using BINDSURF.** *BMC Bioinformatics* 2012, **13(Suppl 14)**:S13.
14. Bergemann TL, Wilson J: **Proportion statistics to detect differentially expressed genes: a comparison with log-ratio statistics.** *BMC Bioinformatics* 2011, **12**:228.

15. Ruan L, Yuan M: An Empirical Bayes' Approach to Joint Analysis of Multiple Microarray Gene Expression Studies. *Biometrics* 2011, **67(4)**:1617-1626.
16. Efron B, Morris C: Combining possibly related estimation problems (with discussion). *Journal of the Royal Statistical Society, Series B* 1973, **35**:379-421.
17. Efron B, Morris C: Stein's paradox in statistics. *Scientific American* 1977, **236**:119-127.
18. Bar H, Booth J, Schifano E, Wells MT: Laplace approximated EM microarray analysis: an empirical Bayes approach for comparative microarray experiments. *Statistical Science* 2010, **25(3)**:388-407.
19. Wright GW, Simon RM: A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics* 2003, **19**:2448-2455.
20. Cui X, Hwang JG, Qiu J, Blades NJ, Churchill GA: Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics* 2005, **6(1)**:59-75.
21. Lönnstedt I, Grant S, Begley G, Speed TP: Microarray analysis of two interacting treatments: a linear model and trends in expression over time. Technical Report, Department of Mathematics, Uppsala University, Sweden; 2001.
22. Tai YC, Speed TP: A multivariate empirical Bayes statistic for replicated microarray time course data. *The Annals of Statistics* 2006, **34(5)**:2387-2412.
23. Lonnstedt I, Speed T: Replicated microarray data. *StatisticaSinica* 2002, **12**:31-46.
24. Newton MA, Kendziorski CM: Parametric empirical Bayes methods for microarrays. *The analysis of gene expression data: methods and software* 2003, 254-271.
25. Liu C, Rubin DB: The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. *Biometrika* 1994, **81(4)**:633-648.
26. Cui X, Churchill GA: Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol* 2003, **4(4)**:210.
27. Noma H, Matsui S: The optimal discovery procedure in multiple significance testing: an empirical Bayes approach. *Statistics in Medicine* 2012, **31(2)**:165-176.
28. Efron B: Robbins, empirical Bayes and microarrays. *The annals of Statistics* 2003, **31(2)**:366-378.
29. Lange KL, Little RJ, Taylor JM: Robust statistical modeling using the t distribution. *Journal of the American Statistical Association* 1989, **84(408)**:881-896.
30. Liu C, Rubin DB: The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. *Biometrika* 1994, **81(4)**:633-648.
31. Meng XL, Rubin DB: Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* 1993, **80(2)**:267-278.
32. McLachlan G, Krishnan T: The EM Algorithm and Extensions. *Wiley Series in Probability and Statistics* 1997.
33. Yang X, Zhang B, Lum PY: Systematic genetic and genomic analysis of cytochrome P450 enzyme activities in human liver. *Genome Research* 2010, **20(8)**:1020-1036.

doi:10.1186/1471-2164-14-S8-S8
Cite this article as: Chakraborty *et al.*: Simultaneous inferences based on empirical Bayes methods and false discovery rates ineQTL data analysis. *BMC Genomics* 2013 **14**(Suppl 8):S8.