BMC Proceedings

Proceedings

A mixture model approach to multiple testing for the genetic analysis of gene expression

Cyril Dalmasso*[1], Joseph Pickrell[2,3], Marianne Tuefferd[1], Emmanuelle Génin[2], Catherine Bourgain[2] and Philippe Broët[1]

Address: [1]JE 2492 Universite Paris-Sud, Hôpital Paul Brousse – Batiment 15/16, 16 Avenue Paul Vaillant Couturier, Villejuif CEDEX 94807, France, [2]INSERM UMR-S 535, Universite Paris-Sud, Villejuif F94807, France and [3]Department of Human Genetics, The University of Chicago, 920 East 58th Street, Chicago, Illinois 60637, USA

Email: Cyril Dalmasso* - dalmasso@vjf.inserm.fr; Joseph Pickrell - pickrell@uchicago.edu; Marianne Tuefferd - tuefferd@vjf.inserm.fr; Emmanuelle Génin - genin@vjf.inserm.fr; Catherine Bourgain - bourgain@vjf.inserm.fr; Philippe Broët - broet@vjf.inserm.fr

* Corresponding author

from Genetic Analysis Workshop 15
St. Pete Beach, Florida, USA. 11–15 November 2006

Published: 18 December 2007

BMC Proceedings 2007, **1**(Suppl 1):S141

This article is available from: http://www.biomedcentral.com/1753-6561/1/S1/S141

Abstract

With the availability of very dense genome-wide maps of markers, multiple testing has become a major difficulty for genetic studies. In this context, the false-discovery rate (FDR) and related criteria are widely used. Here, we propose a finite mixture model to estimate the local FDR (lFDR), the FDR, and the false non-discovery rate (FNR) in variance-component linkage analysis. Our parametric approach allows empirical estimation of an appropriate null distribution. The contribution of our model to estimation of FDR and related criteria is illustrated on the microarray expression profiles data set provided by the Genetic Analysis Workshop 15 Problem 1.

Background

In the context of genetic studies for which high-density genetic maps are now widely available, a major multiple testing problem arises due to the large number of statistical tests that are performed simultaneously. In a recent study, Morley et al. [1] analysed microarray gene-expression data together with genome-wide single nucleotide-polymorphism (SNP) genotyping in 14 three-generation families to localize the genetic determinants underlying gene-expression variability (data provided for Genetic Analysis Workshop 15 (GAW 15) Problem 1). For the genome-wide linkage analysis, the authors calculated a non-parametric Haseman-Elston statistic and used the genome-wide significance thresholds proposed by Lander and Kruglyak [2] to identify linked loci. Thus, they controlled the classical family-wise error rate (FWER), i.e., the probability of falsely rejecting at least one null hypothesis.

Although the FWER is the oldest extension of the classical type I error rate, FWER-based procedures are often too conservative, particularly when numerous hypotheses are being tested [3]. As an alternative and less stringent error criterion, Benjamini and Hochberg introduced, in their seminal paper published in 1995 [4], the false-discovery rate (FDR), which is defined as the expected proportion of false discoveries among all discoveries (here, a discovery refers to a rejected null hypothesis). The opposing criterion, the false non-discovery rate (FNR), corresponds to the expected proportion of false non-discoveries among all the non-rejected null hypotheses [5].

More recently, Efron et al. introduced the local FDR (lFDR) [6], which can be interpreted as a variant of the Benjamini and Hochberg's FDR that gives each tested null hypothesis its own "measure of significance". While the FDR is defined for a whole rejection region, the lFDR is defined as the probability that a null hypothesis is true conditional on a particular value of the test statistic. As discussed by Efron [7], the local nature of the lFDR is advantageous for interpreting results from individual test statistics. Moreover, the FDR can be estimated directly from the lFDR [6].

Efron proposed an empirical Bayes' procedure [7,8] to estimate the lFDR without any assumption about the distribution under the alternative hypothesis. From this procedure, only an upper bound estimate can be obtained for the lFDR and, indirectly, a lower bound for the FNR. One important feature of this approach is that it considers an empirical rather than theoretical null distribution. Indeed, as noted by Efron, these distributions may be very different and strong arguments support using the empirical one in genetic studies for which extensive data are available [5].

In this work, and for variance-component linkage analysis, we introduced a two-component mixture model based approach that allows estimation of lFDR, FDR, and FNR. We illustrate the contribution of our model to the analysis of real GAW15 data. Our results highlight the importance of correctly estimating the null distribution through the proposed mixture model based approach.

Methods

Consider the variance-component linkage analysis between a particular phenotype (here, the expression level of a defined gene) and a specific marker. The null hypothesis of no linkage (additive genetic variance due to the studied quantitative trait locus (QTL) equals zero) is tested by comparing the likelihood of this restricted model with that of a model in which the variance is estimated. Under the null hypothesis, the theoretical asymptotic distribution of the likelihood-ratio statistic X is a 50:50 mixture of a χ^2 and a point mass at 0 [9]. When testing n markers, n likelihood-ratio statistics X_i, $(i = 1,...,n)$ are available, with each X_i following either the null or the alternative distribution.

For modeling of the marginal distribution of X, we consider the following two-component mixture model, in which the marginal cumulative distribution F_X of X is:

$$F_X(x) = \omega_1 \times \{\theta \times 1_{\{X=0\}} + (1 - \theta) \times F_1(x|\alpha_1, \beta_1)\} + \omega_2 \times F_2(x|\alpha_2, \beta_2),$$

where ω_c is the mixing proportions for the c components ($c = 1, 2$; $\omega_c \in [0, 1]$; $\omega_1 + \omega_2 = 1$). Here, $c = 1$ corresponds to the null hypothesis component and $c = 2$ to the alternative hypothesis component, respectively. The parameter $\theta \in [0, 1]$ is the weight of the point mass at 0 for the null hypothesis component.

In this model, the conditional distributions $F_c(x|\alpha_c, \beta_c)$ are gamma distributions with parameters α_c and β_c, where α_c is the mean and α_c/β_c the variance of the distribution. Here, we impose that $\alpha_1 < \alpha_2$.

As discussed in the Background, the empirical distribution under the null hypothesis can be very different from the theoretical distribution [8]. Therefore, we decided to not consider theoretical values ($\theta = 1/2$, $\alpha_1 = 1$, and $\beta_1 = 1/2$) for the first component distribution parameters but rather to estimate them. For the second component, we used a gamma distribution, which represents a convenient and parsimonious way to model the non-null distribution.

Parameters of interest are inferred by sampling from their joint posterior distributions using Monte Carlo Markov chain (MCMC) samplers implemented in WinBUGS software [9]. All results presented correspond to 25,000

Table 1: Estimated parameters of the two-component mixture model for each of the ten genes analyzed

Gene	θ	α_1	β_1	α_2	β_2
CHI3L2	0.41	1.72	0.41	40.40	0.47
DDX17	0.47	2.96	0.24	33.90	0.32
PSPHL	0.39	2.76	0.25	63.04	0.73
IL16	0.50	1.08	0.90	5.14	1.85
HOMER1	0.45	1.02	0.92	3.36	1.25
ALG6	0.53	0.79	1.58	9.71	2.77
CBR1	0.41	0.74	1.37	2.42	2.15
TNFRSF11A	0.52	1.27	0.68	7.34	2.97
TGIF	0.54	0.70	1.35	3.96	1.47
DSCR2	0.52	0.83	1.17	3.57	1.38

sweeps of MCMC algorithms following a burn-in period of 25,000 sweeps (period required to achieve algorithm stability). Convergence is checked by visual inspection of the curve of the plots for the different parameters of the mixtures.

For each marker, the posterior probabilities of belonging to the null hypothesis can be estimated directly from the algorithm output, using empirical averages. These probabilities are natural estimates of the lFDR for each marker. They can be used to compute model-based estimates of the observed FDR and FNR (conditionally to the data) [10,11].

Results

We started from the cell intensity files (*.CEL) obtained from the GeneChip® Human Genome Focus Array Hgfocus [12] that provide gene-expression measurements of 8794 probe sets for 276 samples. We normalized and summarized those measurements using the robust multi-array average (RMA) method proposed by Irizarry et al. [13]. A multipoint variance-component linkage analysis was performed with MERLIN [14] on the normalized phenotypes using all 194 individuals from the 14 Centre d'Etude du Polymorphisme Humain (CEPH) families and the 2819 autosomal SNP data. Using the proposed mixture model, we then estimated the lFDR at each marker, and FDR and FNR. Here, we present only the results obtained for the following 10 genes discussed in the article by Morley et al. [1]: *CHI3L2*, *DDX17*, *PSPHL*, *IL16*, *HOMER1*, *ALG6*, *CBR1*, *TNFRSF11A*, *TGIF*, and *DSCR2*.

Table 1 gives the estimated parameters of the two-component mixture model for the expression of each of the 10 genes (phenotypes). The estimated values of the null distribution parameters differed markedly from the theoretical values. For the 10 selected genes, the maximal differences between the theoretical and empirical values were: 0.11 for θ (*PSPHL*), 1.96 for α_1 (*DDX17*), and 1.08 for β_1 (*ALG6*). For example, Figure 1 illustrates the histogram distribution of the (non-null) observed likelihood-ratio statistic *X*, and superimposed theoretical null hypothesis, marginal and null hypothesis densities estimated from the mixture model for the *DDX17* gene. The marked difference between the theoretical and estimated null distributions strongly supports the use of the estimated null distribution rather than the theoretical one. As noted by Efron [8], these differences can substantially affect any simultaneous inference (including FDR estimation and FWER control). It is worth noting that when the FWER is controlled at 5% with a classical Bonferroni procedure, the *p*-values for the *DDX17* gene calculated from the theoretical null distribution yielded 52 significant results, while the *p*-values calculated from the estimated null dis-

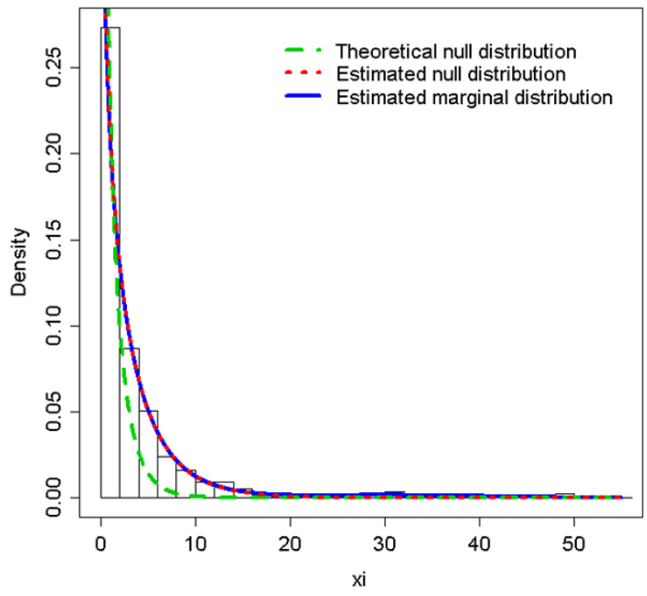

Figure 1
Histogram distribution of the (non-null) observed likelihood ratio statistic, theoretical null hypothesis density, and marginal and null hypothesis densities estimated from the mixture model for the *DDX17* gene.

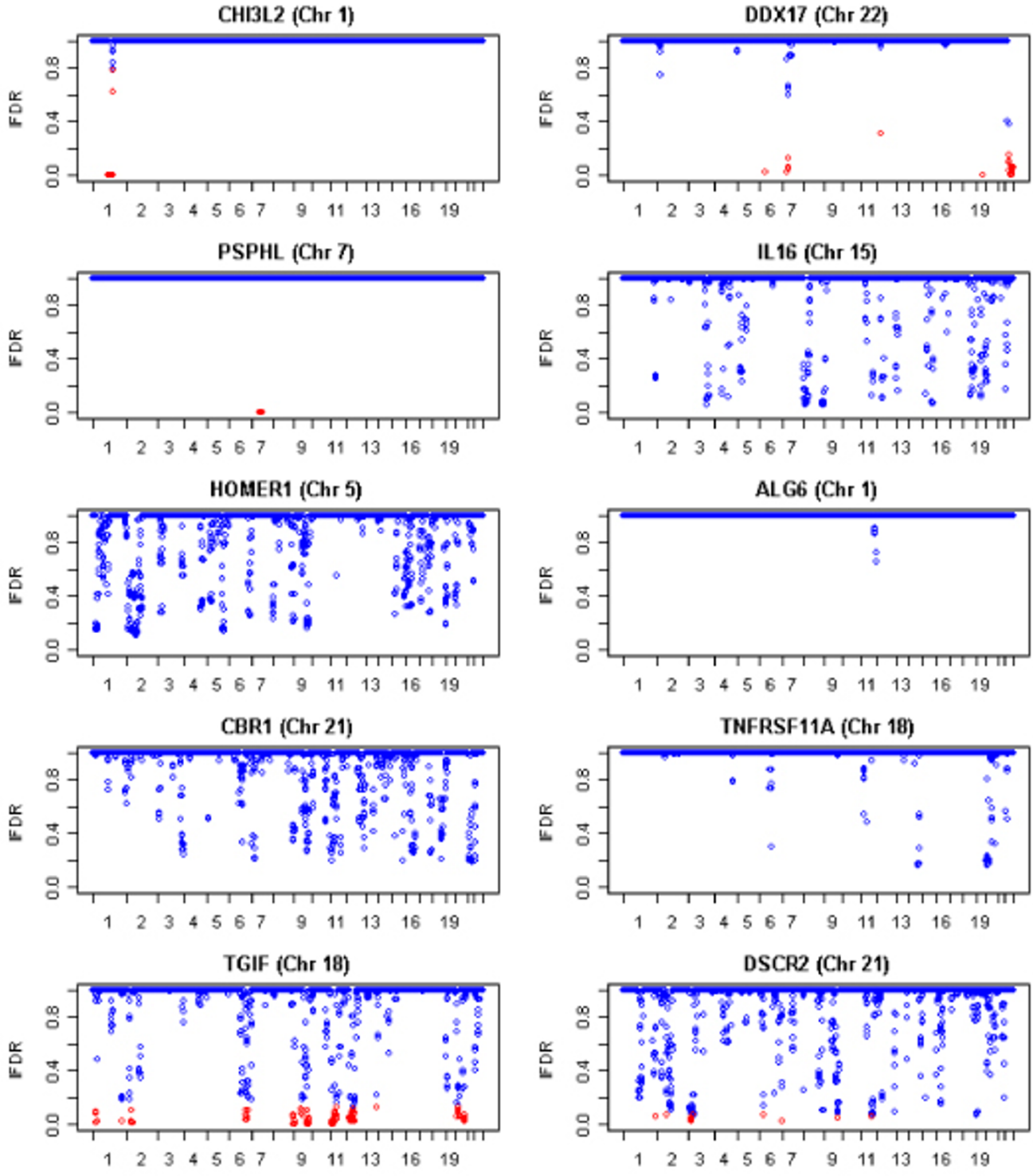

Figure 2
Estimated posterior probabilities (IFDR) for the 10 selected genes along the 22 chromosomes. Significant results at FDR threshold 0.05 are plotted in red.

tribution gave only 13 significant results. In this example, considering the theoretical null distribution clearly tended to overestimate the number of significant results.

Summary statistics calculated from the full output of the MCMC algorithm (after discarding the burn-in samples) provide information on the posterior probabilities of belonging to the null hypothesis component. Using these estimates, probabilistic classification of the data (in terms of discoveries and non-discoveries) can be obtained concomitantly with the estimations of FDR and FNR [10,11]. Herein, we decided to consider as discoveries (linkage) the markers with posterior probabilities below a threshold value, which can be different for each phenotype and was chosen to ensure 5% FDR. Figure 2 shows the estimated posterior probabilities (equivalent to the lFDR) along the 22 chromosomes for the 10 phenotypes. Meanwhile, the estimated FNR ranged from 23% (*PSPHL*) to 28% (*HOMER1*) (data not shown). The selected markers with an lFDR estimate below the defined threshold are plotted in red. These selected markers differed substantially from those obtained by Morley et al. [1]. For example, we found multiple *cis*-acting and *trans*-acting regulators for *DDX7* and *IL16*, while Morley et al. [1] found only *cis*-acting regulators for these genes.

However, it is difficult to directly compare the two approaches because the selection strategies rely on completely different criteria. Moreover, it is worth noting that while the Bonferroni procedure depends on the order of the *p*-values, our procedure depends on the order of the posterior probability (lFDR) values, and the two can be completely different.

Conclusion

Herein we described a mixture model based approach to estimate FDR, FNR, and lFDR in the context of variance component linkage analyses. This approach allows the selection process to take into account both false positives and false negatives. Moreover, it provides an estimate of the empirical null distribution, which is a key component for any simultaneous inference procedure.

Indeed, in many situations, the empirical null distribution deviates from the theoretical one [8], leading to incorrect statistical inferences and resulting decisions. Traditional estimating methods in linkage analysis used simulation approaches in which marker alleles were randomly dropped from the genealogies. When markers are numerous or pedigrees are complex, that method can become very burdensome, with computations requiring several days of running time. New genetic studies for which large amounts of data are available open new opportunities by allowing the estimation of appropriate null and alternative densities without resorting to simulations. Hence, our approach is much easier to handle because examination of each of the different phenotypes analysed required less than 1 hour of computer time. It is important to note that this approach can be extended by incorporating different null distribution parameters for a set of phenotypes in a single model. In conclusion, we think that new insights on linkage analysis using genome-wide technologies might emerge from a mixture model-based approach.

Competing interests

The author(s) declare that they have no competing interests.

Acknowledgements

This article has been published as part of *BMC Proceedings* Volume 1 Supplement 1, 2007: Genetic Analysis Workshop 15: Gene Expression Analysis and Approaches to Detecting Multiple Functional Loci. The full contents of the supplement are available online at http://www.biomedcentral.com/1753-6561/1?issue=S1.

References

1. Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG: **Genetic analysis of genome-wide variation in human gene expression.** *Nature* 2004, **430:**743-747.
2. Lander E, Kruglyak L: **Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results.** *Nat Genet* 1995, **11:**241-247.
3. Hochberg Y, Tamhane A: *Multiple Comparison Procedures New York: John Wiley & Sons, Inc*; 1987.
4. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J R Stat Soc Ser B* 1995, **57:**289-300.
5. Genovese CR, Wasserman L: **Operating characteristics and extensions of the false discovery rate rrocedure.** *J R Stat Soc Ser B* 2002, **64:**499-518.
6. Efron B, Tibshirani R, Storey J, Tusher V: **Empirical Bayes analysis of a microarray experiment.** *J Am Stat Assoc* 2001, **96:**1151-1160.
7. Efron B: **Local false discovery rates.** *Technical Report* 2005 [http://www-stat.stanford.edu/~brad/papers/False.pdf].
8. Efron B: **Large-scale simultaneous hypothesis testing: the choice of a null hypothesis.** *J Am Stat Assoc* 2004, **99:**96-104.
9. Spiegelhalter D, Thomas A, Best N, Lunn D: **WinBUGS User manual Version 1.4.1.** [http://www.mrc-bsu.cam.ac.uk/bugs].
10. Newton MA, Noueiry A, Sarkar D, Ahlquist P: **Detecting differential gene expression with a semiparametric hierarchical mixture method.** *Biostatistics* 2004, **5:**155-176.
11. McLachlan GJ, Bean RW, Ben-Tovim Jones L: **A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays.** *Bioinformatics* 2006, **22:**1608-1615.
12. **Affymetrix website** [http://www.affymetrix.com]
13. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data.** *Biostatistics* 2003, **4:**249-264.
14. Abecasis GR, Cherny SS, Cookson WO, Cardon LR: **Merlin – rapid analysis of dense genetic maps using sparse gene flow trees.** *Nat Genet* 2002, **30:**97-101.