BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 20 2008, pages 2407–2408
doi:10.1093/bioinformatics/btn379

Genetics and population analysis

ExactFDR: exact computation of false discovery rate estimate in case-control association studies

Jérôme Wojcik* and Karl Forner

Department of Bioinformatics, Merck Serono Geneva Research Center, 1202 Geneva, Switzerland

Received on February 26, 2008; revised on June 27, 2008; accepted on July 18, 2008

Advance Access publication July 28, 2008

Associate Editor: Alex Bateman

ABSTRACT

Summary: Genome-wide association studies require accurate and fast statistical methods to identify relevant signals from the background noise generated by a huge number of simultaneously tested hypotheses. It is now commonly accepted that exact computations of association probability value (*P*-value) are preferred to χ^2 and permutation-based approximations. Following the same principle, the *ExactFDR* software package improves speed and accuracy of the permutation-based false discovery rate (FDR) estimation method by replacing the permutation-based estimation of the null distribution by the generalization of the algorithm used for computing individual exact *P*-values. It provides a quick and accurate non-conservative estimator of the proportion of false positives in a given selection of markers, and is therefore an efficient and pragmatic tool for the analysis of genome-wide association studies.

Availability: A Java 1.6 (1.5-compatible) version is available on SourceForge: http://sourceforge.net/projects/exactfdr.

Contact: Jerome.wojcik@merckserono.net

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Genome-wide case-control association analyses are becoming a routine tool in studies of genetic risk factors for complex diseases. Recent technologies now allow the genotyping of hundreds of thousands of single nucleotide polymorphisms (SNPs) on chips. These screening campaigns generally result in a selection of markers that require further validation (both genotyping confirmation, ideally replication, and ultimately functional validation). Because validation steps can only be performed at low throughput, selecting the most relevant set of markers from the primary screen is critical. Most currently used methods are based on setting a *P*-value cutoff, but this does not control the rate of false positives generated by the multiple hypothesis testing problem.

Recently, a methodology for the estimation of the false discovery rate (FDR) was proposed (Forner *et al.*, 2008), less conservative than the pioneering Benjamini and Hochberg initial control procedure (Benjamini *et al.*, 1995) and applicable to any study design, using any association statistic. This algorithm estimates accurately the proportion of false discovery *V/R*, where *V* and *R* are the numbers of false positives and positives at a given *P*-value level, respectively. In Forner *et al.* (2008), a permutation-based implementation of this algorithm was proposed, similarly to other methods developed for differential gene expression studies (Benjamini *et al.*, 2001; Ge *et al.*, 2003; Storey *et al.*, 2003). In this implemented method, called here 'permutation-based FDR estimation', the distribution of *P*-values under the null hypothesis is estimated by computing all *P*-values after random shuffling of case/control labels. The precision of the distribution estimation depends on the number of rounds of shuffling, and therefore the method suffers from long execution times. In order to tackle this issue and make this algorithm easily usable, we have developed the *ExactFDR* software package. *ExactFDR* is a user-friendly tool, available for several common platforms, which implements an exact computation of the FDR estimate based on exact computations of allelic or genotypic *P*-values (Balding , 2006). Its execution time is short and it estimates *V/R* at least as accurately as the permutation-based FDR estimator proposed in Forner *et al.* (2008).

2 ALGORITHM PRINCIPLE

The principle of the algorithm is to estimate the global (experiment-wise) null distribution of the test statistic and then globally adjust exact *P*-values in order to have a perfectly uniform null distribution despite the discreteness and dependency of data. For a given SNP, represented by its genotypic counts $(x_1, \ldots, x_6)$ in a 2×3 contingency table representing two samples of which genotyping distribution is compared (Table 1), the exact *P*-value computation relies on the enumeration of all contingency tables having the same margins as the observed one (dubbed *compatible* tables). The exact *P*-value is the sum of the multiple hypergeometric probabilities of all compatible tables having a statistic as extreme as the observed one (Guedj *et al.*, 2006). The *ExactFDR* algorithm is based on a similar principle: the distribution of *P*-values under the null hypothesis is simulated by comprehensively computing and storing exact *P*-values of all the compatible tables of all the SNPs in the study. Then the type I error rate for a statistic test value α is the proportion of stored *P*-values equal or smaller than α. This ratio is computed using all observed individual exact *P*-values as α thresholds (Forner *et al.*, 2008).

3 SOFTWARE OVERVIEW

ExactFDR requires an input file listing an identifier and the six genotypic counts $(x_1, \ldots, x_6)$ for every SNP in the study. The algorithm is implemented in a Java multithreaded package

*To whom correspondence should be addressed.

Table 1. A genotypic 2 × 3 contingency table

	Homozygote1	Heterozygote	Homozygote2
Sample set #1	x_1	x_2	x_3
Sample set #2	x_4	x_5	x_6

allowing multiprocessors parallel computing. Since the number of individual statistics and multiple hypergeometric probability computations is large [$\sim O(mn^2)$, where n is the total sample size and m the number of SNPs], the software program has been extensively and thoroughly optimized, both at the programming and algorithmic levels. Interested readers can refer to the code documentation and previous publications for details (Forner *et al.*, 2008; Guedj *et al.*, 2006).

4 EXAMPLE APPLICATION

The *ExactFDR* package has been run on experimental data of 313/351 cases/controls from a multiple sclerosis whole genome association study using Affymetrix 500K Genechip® technology. After quality control filtering, 350 000 SNPs have been analyzed. The FDR based on exact allelic and genotypic tests is estimated in 35.0 and 76.5 min, respectively, on a 1.5 GHz Itanium single-processor computer. Execution time drops to 2.1 and 4.5 min, respectively, with 32 processors, demonstrating the efficiency of the multithreaded implementation. The FDR curve corresponding to the genotypic test is illustrated in Figure 1: it overlaps almost perfectly the FDR estimates obtained with the previously published permutation-based estimator, thus proving that *ExactFDR* is an accurate estimator of the actual proportion of false discoveries *V/R*. Differences between the two estimators (Fig. 1b) are most frequently negatives (for 70% of estimates), showing that the *ExactFDR* is on average less conservative than the permutation-based estimator. In addition, *ExactFDR* is about 400 times faster than the permutation-based FDR estimator using 10 000 permutations.

5 DISCUSSION

The identification of genetic risk factors in complex diseases requires efficient statistical tools to analyze the data and address the so-called multiple-testing problem (the number of tested hypotheses is much greater than the sample size) in genome-wide association studies. A pragmatic and accurate methodology has been proposed for estimating the FDR, applicable to any study design (Forner *et al.*, 2008). *ExactFDR* is an exact implementation of this methodology. It combines the accuracy of false-positive proportion estimation with enhanced speed of execution. It requires no arbitrary parameters to set. The software program is available on most common platforms and is simple to use. In conclusion, *ExactFDR* is a useful tool in practice as it permits to estimate, after a genome scan, the proportion of false positives amongst the selection of SNPs that will enter further validation stages.

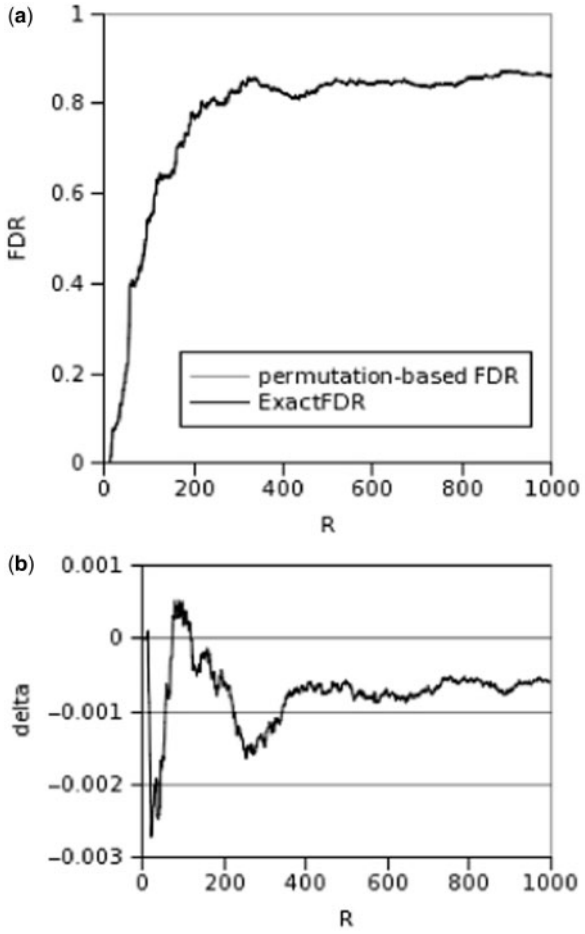

Fig. 1. (**a**) On an experimental dataset (see text), *ExactFDR* estimates are compared with permutation-based estimates (using 10 000 permutations); both are plotted against the number of positives R (for the first 1000). The two curves overlap almost perfectly. Differences (exact—permutation-based estimates) are plotted in (**b**): average difference over the 10 000 first positives is $-1.4e^{-4}$ (variance $2.1e^{-6}$).

Conflict of Interest: none declared.

REFERENCES

Balding,D.J. (2006) A tutorial on statistical methods for population association studies. *Nat. Rev. Genet.*, **7**, 781–791.

Benjamini,Y. *et al.* (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*,289–300.

Benjamini,Y. *et al.* (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, **29**, 1165–1188.

Forner,K. *et al.* (2008) Universal false discovery rate estimation methodology for genome-wide association studies. *Hum. Hered.*, **65**, 183–194.

Ge,Y. *et al.* (2003) Resampling-based multiple testing for microarray data analysis. *Test*, **12**, 1–77.

Guedj,M. *et al.* (2006) A fast, unbiased and exact allelic test for case-control association studies. *Hum. Hered.*, **61**, 210–221.

Storey,J.D. *et al.* (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.