BIOINFORMATICS APPLICATIONS NOTE

Vol. 25 no. 11 2009, pages 1461–1462
doi:10.1093/bioinformatics/btp176

Systems biology

FDR made easy in differential feature discovery and correlation analyses

Xuefeng Bruce Ling*, Harvey Cohen, Joseph Jin, Irwin Lau and James Schilling*

Department of Pediatrics, Stanford University School of Medicine, Stanford University, Stanford, CA 94305, USA

Received on December 14, 2008; revised on February 26, 2009; accepted on March 24, 2009

Advance Access publication April 17, 2009

Associate Editor: Trey Ideker

ABSTRACT

Summary: Rapid progress in technology, particularly in high-throughput biology, allows the analysis of thousands of genes or proteins simultaneously, where the multiple comparison problems occurs. Global false discovery rate (gFDR) analysis statistically controls this error, computing the ratio of the number of false positives over the total number of rejections. Local FDR (lFDR) method can associate the corrected significance measure with each hypothesis testing for its feature-by-feature interpretation. Given the large feature number and sample size in any genomics or proteomics analysis, FDR computation, albeit critical, is both beyond the regular biologists' specialty and computationally expensive, easily exceeding the capacity of desktop computers. To overcome this digital divide, a web portal has been developed that provides bench-side biologists easy access to the server-side computing capabilities to analyze for FDR, differential expressed genes or proteins, and for the correlation between molecular data and clinical measurements.

Availability: http://translationalmedicine.stanford.edu/Mass-Conductor/FDR.html

Contacts: xuefeng_ling@yahoo.com; jschill@stanford.edu

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

1 INTRODUCTION

The ability to simultaneously analyze vast number of genes or proteins in genomics, proteomics and imaging offers both unprecedented scientific opportunities and significant statistical challenges. The concurrent statistical test of thousands of null hypotheses leads to the multiple-testing problem, demanding that the derived test statistics be adjusted to control the expected proportion of false discoveries among all discoveries. This can be achieved either by the overly conservative Bonferroni correction or the analysis of the global false discovery rate (gFDR) (Benjamini and Hochberg, 1995). After determining the gFDR test threshold for significance, the lFDR (lFDR) analysis can compute and assign significance measures to all features. The lFDR analysis (Efron *et al*., 2001) addresses one drawback of the gFDR, statistically distinguishing features that are close to the threshold and therefore more likely to be falsely positive from those that are not.

Although important, multiple testing corrections do not always enter into practitioners' considerations. The FDR computation usually requires some programming with either commercial packages (e.g. SAS, S-Plus) or programs (e.g. C, FORTRAN or R). In reality, few biologists or clinicians are proficient in software coding, scripting or are well informed of packages' computational performance (Westfall *et al*., 1999) and potential memory allocation problems. Due to the large feature number and sample size, the FDR analysis can also be computationally expensive and most likely exceeds the capacity of desktop or notebook computers. To overcome these barriers, a web portal has been developed for the bench-side biologists such that they can easily access the server-side computing tools and generate graphic output for meaningful interpretation. Permutation-based gFDR analysis was implemented as previously described (Tusher *et al*., 2001) and the lFDR analysis integrated previous implementations of the FDR estimators (Aubert *et al*., 2004). The analysis results are summarized in plots and excel tables online, and are also e-mailed to the user from the server.

2 ANALYSIS GUIDELINES

There are two applications deployed at the web site: (i) differential analysis by Student's t-test or non-parametric Mann–Whitney U-test and (ii) correlation analysis between molecular data and measurements of a clinical parameter, including the calculation of Pearson product–moment correlation coefficient, Kendall tau correlation coefficient and Spearman's rank correlation coefficient. Detailed instructions can be found on the web site. The applications can potentially be applied to various molecular datasets such as gene expression analysis, proteomics profiling, kPCR profiling and multiplex ELISA. For the correlation analysis the clinical parameter should be a continuous or categorical variable of any clinical measurements.

In the gFDR analysis, features, e.g. genes, with p-values of the feature-specific t-test, U-test or the correlation tests lower than a threshold deemed statistically significant and are defined as total discoveries. Given the same threshold, permutations of the molecular measurements construct false positive feature sets of varied sizes. To assess the overall quality of the discoveries, thresholds ranging from the minimum single test P-value up to 1 are surveyed comprehensively. The box–whisker graphs, shown in Supplementary Figure 1 (P-value threshold range, left panel: minimum to 1.0; right panel: minimum to 0.05), illustrate the spread of the distribution of the sizes of the false discovery (FD) sets, using

*To whom correspondence should be addressed.

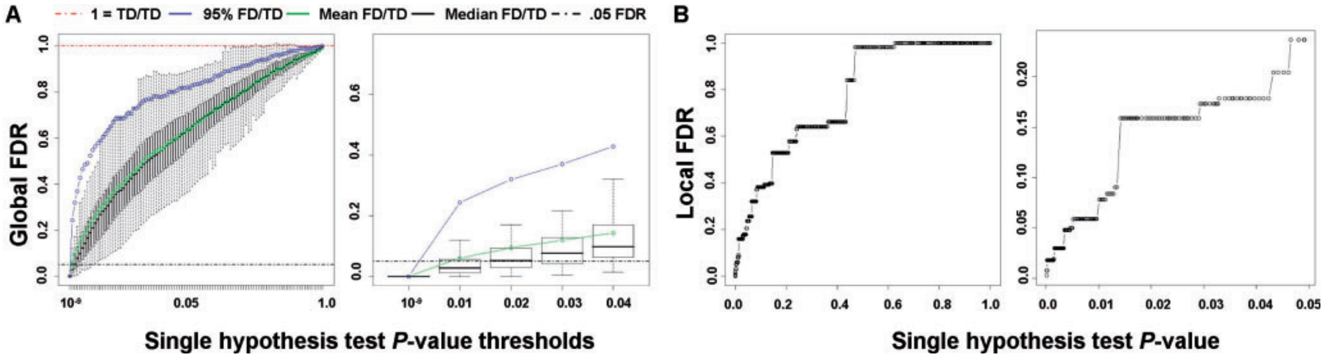

Fig. 1. FDR analyses. The dataset is from a two disease comparison proteomic experiment simultaneously analyzing 1496 proteins for differential expression. (**A**) gFDR analysis. Thresholds ranging from the minimum feature test *P*-value to 1 (left panel) or 0.04 (right panel) are surveyed comprehensively. At each threshold, discoveries in both the input dataset (total discoveries) and the permutations of the input dataset (false discoveries, distribution described by a box–whisker graph) are counted to calculate gFDR. TD: total discoveries. (**B**) lFDR analysis. Each feature-specific test *P*-value is plotted against its corresponding lFDR.

a 'box' (25–75%) and 'whiskers' to break down data by percentile. As shown in Figure 1A, the mean, median and 95th percentile of the sizes of these FD sets when divided by the total discoveries are used to estimate the mean, median and 95th percentile of the FDR, respectively. We found that the largest 5th percentile of the sizes of the FD sets are usually quite different from the remaining 95th percentile, exaggerating the totals of potential FDs. Therefore, 95 percentile of the FD distribution can serve as a good empirical upper bound for the FDR. To choose a good feature-specific *P*-value threshold, balancing the need to maximize the chance of genuine discoveries while minimizing that of the false ones, the generally accepted 5% FDR is commonly used to guide the selection of the suitable feature-specific *P*-value threshold.

To use 5% FDR as guideline to control overall false positive during multiple hypothesis testing is arbitrary. Therefore, the lFDR, which measures the significance that can be attached to each feature, is more appealing to the experimentalists because it can directly estimate the probability for the feature to be a false positive. Another utility of lFDR is to assist the exploration of the biological mechanism, as previously suggested (Aubert *et al*., 2004). The significance of any given functional class or regulatory network can be computed by summing the lFDRs of all component features. The graphic plots contrast the feature-specific test *P*-values and the corresponding lFDRs (shown in Fig. 1B, *P*-value threshold range, left panel: 0–1.0; right panel: 0–0.05). For review purposes, lFDRs are also plotted against the indices of features ordered along their feature-specific *P*-value (Supplementary Fig. 2). It has been suggested (Aubert *et al*., 2004) that the first abrupt change of the lFDR can be an indication for the determination of a good threshold to choose genuinely statistically significant features.

In addition to the graphic outputs, the results of feature-specific tests and FDRs are also summarized in excel files. For the feature-specific *t* or *U* tests, the summary table encapsulates all the feature-specific test *P*-values, and the lFDRs. For the correlation analysis between molecular data and a clinical variable, results are described in two tables: one summarizes all the feature-specific correlation *P*-values and the correlation coefficient estimates; and the other lists the feature-correlation *P*-values and the corresponding lFDRs.

3 OUTLOOK

Our FDR analysis portal is mainly designed for biologists or clinicians to analyze FDR in the hope of discovering genuinely differential features such as genes or proteins with statistical significance or correlations between molecular and clinical datasets to explore translational mechanisms. The analysis begins with a raw data upload and ends with a set of data sheets and plots for easy data drilling and visualization. With all the functions established in the Stanford server no informatics knowledge is required for the end user and computational capacity is guaranteed. This FDR portal should be of general interest to those working in the field of high-throughput biology to address the multiplicity of the statistical tests.

ACKNOWLEDGEMENTS

The authors are very grateful to Dr John Whitin, Dr Elizabeth Mellins and other members of the Stanford Pediatrics Proteomics Group for their critical discussions during the development of this work. The authors thank Mariana Wang and Susan Gao for their work in the tutorials at our FDR web site.

Conflict of Interest: none declared.

REFERENCES

Aubert,J. *et al.* (2004) Determination of the differentially expressed genes in microarray experiments using local FDR. *BMC Bioinformatics*, **5**, 125.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *JRSSB*, **57**, 289–300.

Efron,B. *et al.* (2001) Empirical Bayes analysis of microarray experiment. *J. Am. Stat. Assoc.*, **96**, 1151–1160.

Tusher,V.G. *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Westfall,P.H. *et al.* (1999) *Multiple Comparisons and Multiple Tests (Using the SAS System)*. SAS Institute.