BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 4 2006, pages 507–508
doi:10.1093/bioinformatics/btk005

Gene expression

EDGE: extraction and analysis of differential gene expression

Jeffrey T. Leek*, Eva Monsen, Alan R. Dabney and John D. Storey

Department of Biostatistics, University of Washington, Seattle 98195, USA

Received on October 18, 2005; revised on December 10, 2005; accepted on December 11, 2005
Advance Access publication December 15, 2005
Associate Editor: John Quackenbush

ABSTRACT

Summary: EDGE (Extraction of Differential Gene Expression) is an open source, point-and-click software program for the significance analysis of DNA microarray experiments. EDGE can perform both standard and time course differential expression analysis. The functions are based on newly developed statistical theory and methods. This document introduces the EDGE software package.

Availability: EDGE is freely available for non-commercial users. EDGE can be downloaded for Windows, Macintosh and Linux/UNIX from http://faculty.washington.edu/jstorey/edge

Contact: jtleek@u.washington.edu

1 INTRODUCTION

DNA microarrays have become a standard tool used in identifying and characterizing gene expression variation across differing biological conditions. A variety of software packages are available for the significance analysis of microarray experiments. Many of these packages are closed source, difficult to use or available for only one operating system. Most are unable to analyze data from time course microarray experiments. EDGE is a user friendly software package that includes functions for missing data imputation, data transformation and visualization, eigen-genes/eigen-array analysis, hierarchical clustering, differential expression analysis (static and time course) and automatic internet-based NCBI queries of user chosen genes. EDGE can be used to analyze microarray data across all platforms, although interpretation of the results may depend on the experimental design. The EDGE interface is multithreaded, and reports real time updates for the time remaining in lengthy calculations. Many of these calculations are performed through C++ extensions for `R` that dramatically reduce computation time. Differential expression analyses in EDGE are based on newly developed statistical methodology, including the Optimal Discovery Procedure for static differential expression (Storey, 2005, http://www.bepress.com/uwbiostat/paper259). EDGE is open source and is available for Windows, Macintosh and Linux/UNIX operating systems.

2 EDGE

EDGE runs on top of the statistical software package `R` (R Development Core Team, 2005, http://www.R-project.org). Detailed downloading and installation instructions are available from the EDGE website. At the beginning of each EDGE session, the main menu should appear as in Figure 1. The first step in an EDGE analysis is to load the pre-normalized expression data and covariate files using the `Load/Save Expression Data and Covariates` menu. (The covariate file contains information about the experimental design, such as which biological group from which each array comes.) If the expression matrix has missing values, they can be imputed using the KNN imputation algorithm from the `Impute Missing Data` menu (Troyanskaya *et al.*, 2001). After loading expression data and covariate information, the covariates can be checked for accuracy using the `View Covariates` menu. It is also possible to center, scale or log transform the expression values using the `Transform Data` menu.

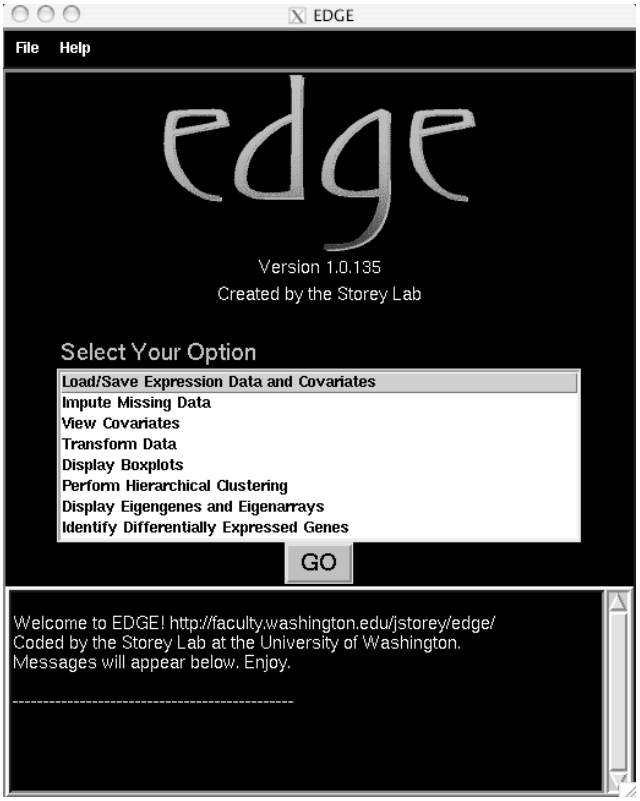

Fig. 1. The main menu of EDGE.

Several tools for visual exploratory analysis are included in the EDGE interface. Boxplots and eigengenes (Alter *et al.*, 2000) can be displayed for each array, or stratified by a covariate using the `Display Boxplots` option and `Display Eigengenes and Eigenarrays` options, respectively. EDGE also allows the user to plot clusters of genes with similar expression patterns

*To whom correspondence should be addressed.

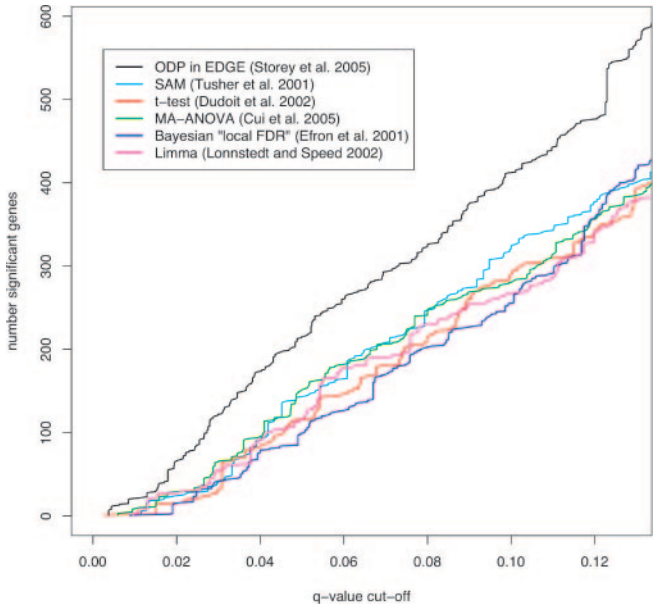

Fig. 2. A comparison between EDGE and five leading procedures for identifying differentially expressed genes applied to the Hedenfalk *et al*., 2001 study. For each *Q*-value (false discovery rate) cutoff, the number of genes found to be significant is plotted for each procedure. See Storey *et al*. (2005a) for comparisons based on a 3-sample analysis, where improvements are even greater.

(Eisen *et al*., 1998) from the `Display Hierarchical Clustering` menu. Clustering can be performed on the entire set of genes, or only on the significant genes from a differential expression analysis. A variety of plotting options are available for visualizing the clusters.

The `Identify Differentially Expressed Genes` menu allows users to set options for performing both static and time course differential expression analyses. For a static analysis, the user should select a class variable indicating the biological group assignment, or the option `None (within class differential expression)` to identify differentially expressed genes in a single biological sample. In the static setting, significance calculations are based on the Optimal Discovery Procedure (Storey, 2005), which estimates the optimal rule for identifying differentially expressed genes (Storey *et al*., 2005a, http://www.bepress.com/uwbiostat/paper260). For time course data, the user can perform either a 'between class' analysis by selecting a variable distinguishing biological groups, or a 'within class' analysis by selecting `None (within class differential expression)` for the class variable. A 'between class' analysis assesses the evidence for a difference in expression over time between two or more biological groups, while a 'within class' analysis looks for any differential expression over time within a single group. The user must specify a covariate for the time points, and if necessary, should also specify a covariate corresponding to which individuals were sampled. EDGE implements statistical methodology specifically designed for time course experiments (Storey *et al*., 2005b).

For either type of analysis, the user should specify the number of permutations to be used in the significance calculations and, in some cases, set a seed for reproducible results. For time course analyses, the user can also specify the type of spline used in fitting the longitudinal model, the dimension of the basis for the spline model and whether to include the baseline expression level in the time course analysis. If the baseline level is included, EDGE will not only identify genes showing different patterns of expression over time, but will also identify genes with different baseline levels of expression.

Once the appropriate options have been selected and the user clicks `GO`, the expression analysis is performed and the `Differential Expression Results` menu is displayed. A significance measure is assigned to each gene via the *Q*-value methodology (Storey and Tibshirani, 2003). The user can select a *Q*- or *P*-value cutoff to display the genes that meet that significance threshold. For advanced users, optional *Q*-value arguments can also be adjusted. The user can plot a histogram of the *P*-values from all significance tests, create a *Q*-plot, or cluster significant genes based on similarities in their expression profiles. If the EDGE session is being performed on a computer with internet access, the user can select a significant gene in the results window, and access NCBI information for that gene name. Results of differential expression analyses can be saved for further analysis or reporting.

3 RESULTS

Figure 2 shows the results of a differential expression analysis on a subset of 3170 genes on 15 arrays from the Hedenfalk *et al*. (2001) study. The analysis compared expression levels for BRCA1 and BRCA2 tumors. EDGE shows substantial improvements over five leading methodologies.

ACKNOWLEDGEMENTS

This software development was supported in part by NIH grant R01 HG002913-01.

Conflict of Interest: none declared.

REFERENCES

Alter,O. *et al*. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl Acad. Sci*., **97**, 10101–10106.

Cui,X. *et al*. (2005) Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics*, **6**, 59–75.

Dudoit,S. *et al*. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc*., **97**, 77–87.

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

Eisen,M. B. *et al*. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci*., **95**, 14863–14868.

Hedenfalk,I. *et al*. (2002) Gene-expression profiles in hereditary breast cancer. *N. Engl. J. Med*., **344**, 539–548.

Lonnstedt,I. and Speed,T. (2002) Replicated microarray data. *Stat. Sinica*, **12**, 31–46.

R Development Core Team (2005) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

Storey,J.D. (2005) The optimal discovery procedure: a new approach to simultaneous significance testing. *UW Biostatistics Working Paper Series Working Paper*, **259**.

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genome-wide studies. *Proc. Natl Acad. Sci*., **100**, 9440–9445.

Storey,J.D., Dai,J.Y. and Leek,J.T. (2005a) The Optimal Discovery Procedure for Large-Scale Significance Testing, with Applications to Comparative Microarray Experiments. *UW Biostatistics Working Paper Series*. Working Paper 260.

Storey,J.D. *et al*. (2005b) Significance analysis of time course microarray experiments. *Proc. Natl Acad. Sci*., **36**, 12837–12842.

Troyanskaya,O. *et al*. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.