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

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

*To whom correspondence should be addressed.

Fig. 1. The main menu of EDGE.

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.

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.
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


Corrigendum
EDGE: extraction and analysis of differential gene expression

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

Bioinformatics (2006) 22(4), 507–508

The authors would like to apologise for the omission of John D. Storey’s email address, as he is a corresponding author for the article. He can be contacted by jstorey@u.washington.edu.

The authors would also like to apologise for an incorrect reference given in the article. The reference cited by Storey et al. (2005b) should be as follows: