BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 5 2012, pages 726–728
doi:10.1093/bioinformatics/bts011

Gene expression

Advance Access publication January 13, 2012

nEASE: a method for gene ontology subclassification of high-throughput gene expression data

Thomas W. Chittenden[1,2,3,4,†], Eleanor A. Howe[1,4,†], Jennifer M. Taylor[5,†], Jessica C. Mar[1,2,3], Martin J. Aryee[6,7], Harold Gómez[1], Razvan Sultana[1], John Braisted[8], Sarita J. Nair[1], John Quackenbush[1,2,3] and Chris Holmes[4,9,*]

[1]Department of Biostatistics and Computational Biology, [2]Department of Cancer Biology, Dana-Farber Cancer Institute, [3]Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA, [4]Department of Statistics, University of Oxford, Oxford, UK, [5]Bioinformatics Group, CSIRO Plant Industry, Canberra ACT, Australia, [6]Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, [7]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, [8]J. Craig Venter Institute, Rockville, MD, USA and [9]MRC Mammalian Genetics Unit, Harwell, UK

Associate Editor: Janet kelso

ABSTRACT

Summary: High-throughput technologies can identify genes whose expression profiles correlate with specific phenotypes; however, placing these genes into a biological context remains challenging. To help address this issue, we developed nested Expression Analysis Systematic Explorer (nEASE). nEASE complements traditional gene ontology enrichment approaches by determining statistically enriched gene ontology subterms within a list of genes based on co-annotation. Here, we overview an open-source software version of the nEASE algorithm. nEASE can be used either stand-alone or as part of a pathway discovery pipeline.

Availability: nEASE is implemented within the Multiple Experiment Viewer software package available at http://www.tm4.org/mev.

Contact: cholmes@stats.ox.ac.uk

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

Received on September 1, 2011; revised on January 4, 2012; accepted on January 5, 2012

1 INTRODUCTION

Numerous studies indicate that high-throughput measurement of gene transcription provides the means to examine how organisms respond on a genome-wide scale to experimental perturbations or to the development of pathological conditions. Analysis of the data produced by these technologies typically results in lists of genes whose expression patterns correlate with the phenotypes under study. However, placing these lists into a useful biological context remains a significant challenge.

A common approach to this problem is to perform statistical assessment of categorical assignments of functional attributes (Ashburner *et al.*, 2000) associated with genes identified in a particular analysis. Current functional enrichment algorithms can be classified into three main classes: singular enrichment analysis (SEA); gene set enrichment analysis (GSEA); and modular enrichment analysis (MEA) (Huang da *et al.*, 2009).

Traditional SEA approaches, such as EASE (Hosack *et al.*, 2003) and Onto-express (Draghici *et al.*, 2003a, b; Khatri *et al.*, 2002, 2004), use a predefined list of differentially expressed genes to linearly assess the enrichment of individual gene ontology (GO) terms (Ashburner *et al.*, 2000) in a term-by-term fashion with statistical methods that include chi-square, Fisher's exact test and Binomial probability. The GSEA class, on the other hand, takes a 'no-cutoff' approach that utilizes expression data and associated experimental values from all genes in a microarray experiment to derive enrichment scores for annotation terms. In the original GSEA method, gene members of each functional category are rank ordered based on expression values to determine a maximum term enrichment score; *P*-values are then assigned to each term by a Kolmogorov–Smirnov-like statistic (Subramanian *et al.*, 2005). The more recent MEA class of algorithms use varied statistical methods combined with network discovery approaches that improve enrichment sensitivity and specificity by evaluating the relationships among categorical assignments (i.e. parents, children, siblings) (Alexa *et al.*, 2006; Grossmann *et al.*, 2007; Zhang *et al.*, 2010).

2 METHODS AND IMPLEMENTATION

The nEASE algorithm allows users to discover biological subclassifications for biological process, molecular function and cellular component GO terms. nEASE is a unique MEA method that combines SEA statistics with an iterative discovery approach based on gene co-annotation. The algorithm identifies enriched, sublevel GO terms based on co-annotating genes from adaptive gene tallies generated from statistically enriched EASE terms. That is, the program first runs EASE to identify enriched primary terms. Then within each enriched primary term, a second nested EASE is run, restricted to elements of the primary term, to see if there is interesting biological functionality that distinguishes the expressed from the non-expressed genes within the primary term. The algorithm performs a Fisher's exact test on each enriched EASE term with the primary GO term as background, while controlling for familywise error rate as described in Supplementary Material. Figure 1 provides an example from our Supplementary Material

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

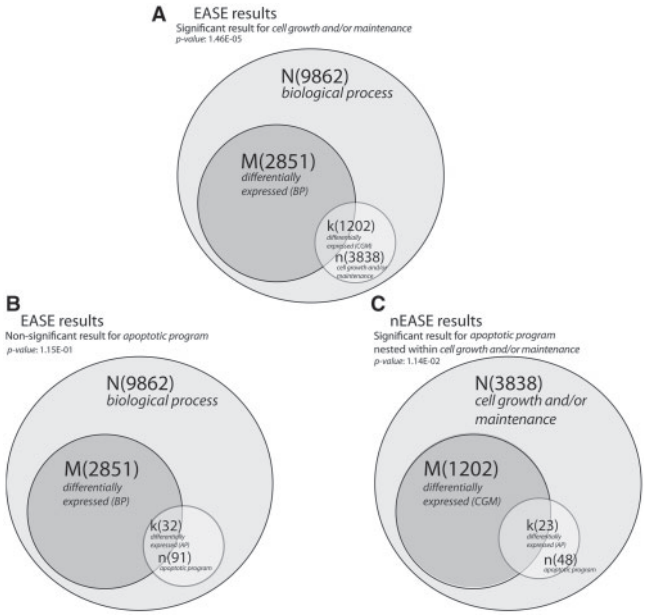

Fig. 1. Euler diagrams indicating the background adjustment used in nEASE. (**A**) The enriched EASE result for the GO biological process term, cell growth and/or maintenance, relative to estrogen receptor (ER) status of our Supplementary Material analysis of the Miller *et al.* study (2005). (**B**) The non-enriched EASE result for the GO biological process term, apoptotic program. (**C**) The enriched nEASE apoptotic program GO biological process term when nested within the enriched EASE GO cell growth and/or maintenance. N represents the total number of annotated genes associated with either total GO biological process annotation of the Affymetrix HG-U133A chip (A and B) or a given enriched EASE GO term (C). The annotated genes for each enriched EASE GO term are used as individual background distributions for subsequent nEASE analysis. M is either the number of differentially expressed annotated biological process genes of the Affymetrix HG-U133A chip (A and B) relative to ER status of the Miller *et al.* study or the number of differentially expressed annotated biological process genes associated with the same enriched EASE GO term as N (C). *n* is the total number of annotated genes within N that are also annotated for a specific EASE (A and B) or sublevel nEASE (C) GO term. *k* is the total number of differentially expressed genes within M that are annotated for the same EASE (A and B) or sublevel nEASE (C) GO term as *n*.

analysis of the background distribution adjustment performed by the nEASE algorithm.

In our Supplementary Material analysis, we applied nEASE to large, publicly available human breast, lung, prostate and renal cancer DNA microarray datasets. We uncovered differences among differentially expressed gene lists relative to these human cancers not identified with the standard SEA GO analysis. We also provide experimental support for the usefulness of nEASE to generate unique working hypothesizes relative to a given condition under study. While one should be careful of such *post hoc* analysis, we believe it does add further evidence as to the expected utility of our method. Thus, we find that nEASE provides additional, complementary biological information based on gene co-annotation of four human cancers that would otherwise be overlooked by existing methods.

We have implemented an open-source version of nEASE within the Multiple Experiment Viewer (MeV) software package. MeV is the main data analysis and visualization tool of the TM4 software suite (Saeed *et al.*, 2006). MeV consists of open-source tools for data management and reporting, data normalization and pipeline control, and data mining and visualization. An integrated MIAME-compliant MySQL database is also included in the package. Within the MeV data mining environment, users can load raw or normalized data from a variety of input file types. A user-friendly, integrated scripting interface and XML-based format allows users to access a broad range of established high-throughput data analysis tools.

Further details of the nEASE algorithm are provided in Supplementary Material. We also provide comprehensive MeV user support, including a detailed tutorial describing how to use the nEASE algorithm, a Frequently Asked Questions (FAQ) page and the MeV Sourceforge help forum at http://www.tm4.org/mev/?q=support.

3 DISCUSSION

There are many GO classification tools in the literature, including but not limited to GOTermFinder (Boyle *et al.*, 2004), GOMiner (Zeeberg *et al.*, 2003), Onto-express (Draghici *et al.*, 2003a, b; Khatri *et al.*, 2002, 2004), Gene Set Enrichment Analysis (GSEA) (Jiang and Gentleman, 2007; Subramanian *et al.*, 2005), FatiScan (Al-Shahrour *et al.*, 2007), Ontologizer (Bauer *et al.*, 2008; Grossmann *et al.*, 2007) and GO-Bayes (Zhang *et al.*, 2010). These tools provide the user with an effective means to biologically interpret genomic data findings. While these methods take different approaches to determining overrepresentation of functional attributes, to our knowledge none considers the relationships among GO terms based on gene co-annotation. To complement the existing SEA methods, we developed nEASE. Conditioning on enriched upper-level EASE terms dramatically reduces the dimensionality of the search space, increasing the power to detect significant GO subterm enrichment. Unlike contemporary MEA algorithms, nEASE evaluates the relationships among all categorical assignments of the GO directed acyclic graph.

Funding: C.H. is supported by the Medical Research Council, UK.

Conflict of Interest: none declared.

REFERENCES

Al-Shahrour,F. *et al.* (2007) From genes to functional classes in the study of biological systems. *BMC Bioinformatics*, **8**, 114.

Alexa,A. *et al.* (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, **22**, 1600–1607.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Bauer,S. *et al.* (2008) Ontologizer 2.0–a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics*, **24**, 1650–1651.

Boyle,E.I. *et al.* (2004) GO: TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, **20**, 3710–3715.

Draghici,S. *et al.* (2003a) Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Res.*, **31**, 3775–3781.

Draghici,S. *et al.* (2003b) Global functional profiling of gene expression. *Genomics*, **81**, 98–104.

Grossmann,S. *et al.* (2007) Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis. *Bioinformatics*, **23**, 3024–3031.

Hosack,D.A. *et al.* (2003) Identifying biological themes within lists of genes with EASE. *Genome Biol.*, **4**, R70.

Huang da,W. *et al.* (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.

Jiang,Z. and Gentleman,R. (2007) Extensions to gene set enrichment. *Bioinformatics*, **23**, 306–313.

Khatri,P. *et al.* (2002) Profiling gene expression using onto-express. *Genomics*, **79**, 266–270.

Khatri,P. *et al.* (2004) Onto-Tools: an ensemble of web-accessible, ontology-based tools for the functional design and interpretation of high-throughput gene expression experiments. *Nucleic Acids Res.*, **32**, W449–W456.

Miller,L.D. *et al.* (2005) An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc. Natl Acad. Sci. USA*, **102**, 13550–13555.

Saeed,A.I. *et al.* (2006) TM4 microarray software suite. *Methods Enzymol.*, **411**, 134–193.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

Zeeberg,B.R. *et al.* (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.*, **4**, R28.

Zhang,S. *et al.* (2010) GO-Bayes: Gene Ontology-based overrepresentation analysis using a Bayesian approach. *Bioinformatics*, **26**, 905–911.