Bioinformatics, 32(23), 2016, 3679–3681
doi: 10.1093/bioinformatics/btw519
Advance Access Publication Date: 8 August 2016
Applications Note

Gene expression

shinyGEO: a web-based application for analyzing gene expression omnibus datasets

Jasmine Dumas[1], Michael A. Gargano[2] and Garrett M. Dancik[2,*]

[1]College of Computing and Digital Media, DePaul University, Chicago, IL, USA and [2]Department of Mathematics and Computer Science, Eastern Connecticut State University, Willimantic, CT, USA

*To whom correspondence should be addressed.

Associate editor: Ziv Bar-Joseph

Received on April 28, 2016; revised on July 19, 2016; accepted on August 3, 2016

Abstract

Summary: The Gene Expression Omnibus (GEO) is a public repository of gene expression data. Although GEO has its own tool, GEO2R, for data analysis, evaluation of single genes is not straightforward and survival analysis in specific GEO datasets is not possible without bioinformatics expertise. We describe a web application, *shinyGEO*, that allows a user to download gene expression data sets directly from GEO in order to perform differential expression and survival analysis for a gene of interest. In addition, *shinyGEO* supports customized graphics, sample selection, data export and *R* code generation so that all analyses are reproducible. The availability of *shinyGEO* makes GEO datasets more accessible to non-bioinformaticians, promising to lead to better understanding of biological processes and genetic diseases such as cancer.

Availability and Implementation: Web application and source code are available from http://gdancik.github.io/shinyGEO/.

Contact: dancikg@easternct.edu

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

1 Introduction

The Gene Expression Omnibus (GEO) is a public repository of genomic data (Barrett *et al.*, 2013; Edgar *et al.*, 2002) that currently hosts >50 000 gene expression datasets containing >1 million samples. GEO is a valuable resource for identifying biomarkers of biological processes and disease. Its datasets have been used to identify biomarkers of HRAS and MYC pathway activity (Bild *et al.*, 2006), biomarkers of monocyte differentiation in response to viral infection (Hou *et al.*, 2012), and prognostic biomarkers in breast (Grinchuk *et al.*, 2015; Ma *et al.*, 2004), bladder (Kim *et al.*, 2010; Lindgren *et al.*, 2010) and lung cancer (Lee *et al.*, 2008; Tomida *et al.*, 2009).

Biomarker evaluation in GEO datasets typically involves data retrieval and gene (probe) selection followed by an appropriate statistical analysis. A differential expression analysis determines whether the expression of a gene differs significantly across two or more groups of samples. A survival analysis determines whether gene expression is significantly associated with an event (such as death from disease), through comparison of Kaplain–Meier curves based on censored time to event data (Bland and Altman, 1998). To facilitate data analysis, GEO provides an interactive web tool [GEO2R; http://www.ncbi.nlm.nih.gov/geo/geo2r/], which allows users to identify differentially expressed genes in GEO datasets. However, GEO2R does not support survival analysis and is designed for gene discovery rather than for evaluation of individual genes. In addition, although specialized tools have been developed to facilitate gene expression analysis by non-bioinformaticians (Dancik, 2015; Győrffy *et al.*, 2013; Rhodes *et al.*, 2007), there are no available tools that allow a user to evaluate whether a specific gene is differentially expressed or associated with survival in a specific GEO dataset.

In this paper, we describe *shinyGEO*, a web-based tool for performing differential expression and survival analysis on gene expression datasets in GEO. In addition, *shinyGEO* produces publication-ready and customizable graphics, allows for sample selection, data correction, data export for custom analyses and *R* code generation for reproducibility.

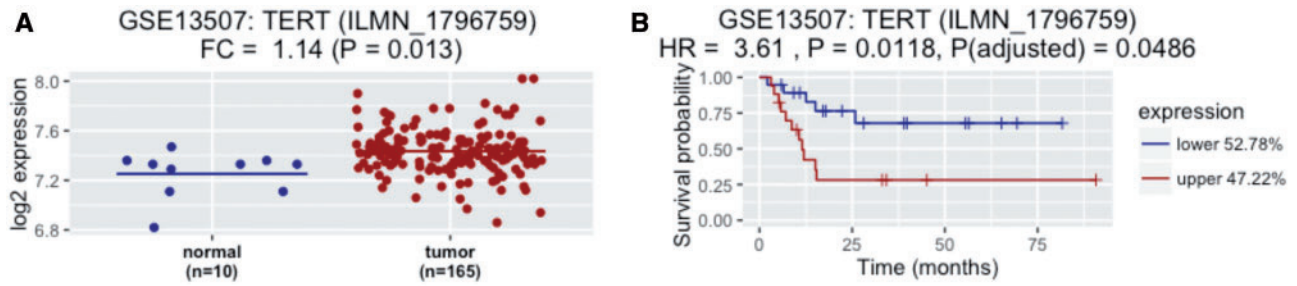

Fig. 1. Example *shinyGEO* analysis. **(A)** Evaluation of *TERT* expression in normal bladder and tumor samples, with fold change (FC) and *P*-value calculated by two-sample *t*-test. **(B)** Kaplan–Meier cancer-specific survival curves for bladder cancer patients with low (lower 52.78%) and high (upper 47.22%) *TERT* expression, with groups chosen based on the optimal cut point. All patients have muscle-invasive tumors and did not receive chemotherapy. The hazard-ratio (HR) and log-rank and adjusted *P*-values are reported (Color version of this figure is available at *Bioinformatics* online.)

2 Methods

shinyGEO is implemented using *R* [https://www.r-project.org/], a free, open-source software environment for statistical computing and graphics; and *shiny* [http://shiny.rstudio.com/], an *R* package for building interactive web applications. In *shinyGEO*, processed gene expression datasets are downloaded from GEO using the *GEOquery* package (Davis and Meltzer, 2007). *P*-values are calculated using a two-sample *t*-test or one-way analysis of variance (ANOVA) for differential expression. For survival analysis, the log-rank test is used to compare samples with high expression (e.g. top 50%) to samples with low expression (e.g. bottom 50%). Optimal cut points by *P*-value are identified and adjusted *P*-values calculated using the *survMisc* package (Contal and O'Quigley, 1999).

3 Software usage

An overview of *shinyGEO* is provided in Supplementary Figures S1 and S2. First, the user selects the GEO accession number for a desired dataset, followed by the platform. The user can carry out a differential expression or survival analysis, and selects the gene and probe of interest. For a differential expression analysis, the user selects the two or more groups to compare. Dot plots are dynamically generated to visualize expression differences across groups, and results from the appropriate statistical analysis are displayed. For survival analysis, the user selects the survival information (i.e. the columns from the sample data containing the time to event and outcome information), and whether to use the median or optimal cut point to define high and low expression groups. The corresponding Kaplan–Meier curves are then dynamically generated and the statistical results displayed. For both analyses, *shinyGEO* supports sample selection and data correction, and graphs can be formatted with respect to color and axis labels. The user can also export the data and generate the *R* code used for each analysis, following best practices in scientific computing (Wilson *et al.*, 2014) to ensure reproducibility.

4 Example application

We demonstrate *shinyGEO* by repeating a recent analysis which found that high telomerase reverse transcriptase (*TERT*) expression is associated with poor disease-specific survival in bladder cancer patients (Borah *et al.*, 2015). We analyze the dataset GSE13507, which contains expression profiles for 165 primary bladder tumors and 10 normal bladder samples obtained from the Chungbuk National University Hospital (Kim *et al.*, 2010). A differential expression analysis finds that *TERT* expression is up-regulated in primary tumors compared to normal bladder samples, with a low fold-change (FC = 1.14, $P = 0.013$, Fig. 1A).

Following Borah *et al.* (2015), we limit the survival analysis to only those patients with muscle-invasive tumors who did not receive chemotherapy or additional treatment (Supplementary Fig. S3). For our selected patients, a cancer-specific survival analysis (using the outcome column 'characteristics_ch1.9') finds that high *TERT* expression is associated with poor outcomes (HR = 3.61, adjusted $P = 0.049$, Fig. 1B), consistent with the findings of Borah *et al.* (2015). Raw data is exported (Supplementary File S1), and for reproducibility we 'Save R Code'. Within *R*, a notebook integrating *R* code and output can then be generated using the *rmarkdown* library (Supplementary File S2) (Stodden *et al.*, 2014).

Funding

This work is supported, in part, by Google Summer of Code 2015 funding to JD.

Conflict of Interest: none declared.

References

Barrett,T. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.

Bild,A.H. *et al.* (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, **439**, 353–357.

Bland,J.M. and Altman,D.G. (1998) Survival probabilities (the Kaplan-Meier method). *BMJ*, **317**, 1572–1580.

Borah,S. *et al.* (2015) TERT promoter mutations and telomerase reactivation in urothelial cancer. *Science*, **347**, 1006–1010.

Contal,C. and O'Quigley,J. (1999) An application of changepoint methods in studying the effect of age on survival in breast cancer. *Comput. Stat. Data Anal.*, **30**, 253–270.

Dancik,G.M. (2015) An online tool for evaluating diagnostic and prognostic gene expression biomarkers in bladder cancer. *BMC Urol.*, **15**, 59.

Davis,S. and Meltzer,P.S. (2007) GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinform. Oxf. Engl.*, **23**, 1846–1847.

Edgar,R. *et al.* (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.

Grinchuk,O.V. *et al.* (2015) Sense-antisense gene-pairs in breast cancer and associated pathological pathways. *Oncotarget*, **6**, 42197–42221.

Győrffy,B. *et al.* (2013) Online survival analysis software to assess the prognostic value of biomarkers using transcriptomic data in non-small-cell lung cancer. *PLoS One*, **8**, e82241.

Hou,W. *et al.* (2012) Viral infection triggers rapid differentiation of human blood monocytes into dendritic cells. *Blood*, **119**, 3128–3131.

Kim,W.J. *et al*. (2010) Predictive value of progression-related gene classifier in primary non-muscle invasive bladder cancer. *Mol. Cancer*, **9**, 3.

Lee,E.S. *et al*. (2008) Prediction of recurrence-free survival in postoperative non-small cell lung cancer patients by using an integrated model of clinical information and gene expression. *Clin. Cancer Res*., **14**, 7397–7404.

Lindgren,D. *et al*. (2010) Combined gene expression and genomic profiling define two intrinsic molecular subtypes of urothelial carcinoma and gene signatures for molecular grading and outcome. *Cancer Res*, **70**, 3463–3472.

Ma,X.J. *et al*. (2004) A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer Cell*, **5**, 607–616.

Rhodes,D.R. *et al*. (2007) Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia N. Y. N*, **9**, 166–180.

Stodden,V. *et al*. eds. (2014) *Implementing Reproducible Research*. Chapman and Hall/CRC, Boca Raton.

Tomida,S. *et al*. (2009) Relapse-related molecular signature in lung adenocarcinomas identifies patients with dismal prognosis. *J. Clin. Oncol*., **27**, 2793–2799.

Wilson,G. *et al*. (2014) Best practices for scientific computing. *PLoS Biol*., **12**, e1001745.