Bioinformatics, 33(21), 2017, 3500–3501
doi: 10.1093/bioinformatics/btx452
Advance Access Publication Date: 14 July 2017
Applications Note

Data and text mining

ScanGEO: parallel mining of high-throughput gene expression data

Katja Koeppen*, Bruce A. Stanton and Thomas H. Hampton

Department of Microbiology and Immunology, The Geisel School of Medicine at Dartmouth, Hanover, NH 03755, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on February 15, 2017; revised on June 20, 2017; editorial decision on July 9, 2017; accepted on July 11, 2017

Abstract

Summary: Current options to mine publicly available gene expression data deposited in NCBI's gene expression omnibus (GEO), such as the GEO web portal and related applications, are optimized to reanalyze a single study, or search for a single gene, and therefore require manual intervention to reanalyze multiple studies for user-specified gene sets. ScanGEO is a simple, user-friendly Shiny web application designed to identify differentially expressed genes across all GEO studies matching user-specified criteria, for a flexible set of genes, visualize results and provide summary statistics and other reports using a single command.

Availability and implementation: The ScanGEO source code is written in R and implemented as a Shiny app that can be freely accessed at http://scangeo.dartmouth.edu/ScanGEO/. For users who would like to run a local instantiation of the app, the R source code is available under a GNU GPLv3 license at https://github.com/StantonLabDartmouth/AppScanGEO.

Contact: katja.koeppen@dartmouth.edu

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

1 Introduction

The NCBI gene expression omnibus (GEO, Edgar *et al.*, 2002) is a repository of high-throughput data containing millions of significant results, a tiny percent of which have been reported in the literature. While the NCBI GEO web portal permits reanalysis of gene expression one experiment at a time using GEO2R (Barrett *et al.*, 2013) and allows the identification of scientifically relevant studies through a database search, it does not connect these two functionalities. Related web applications like MeV (Saeed *et al.*, 2003) or shinyGEO (Dumas *et al.*, 2016) also do not allow for differential gene expression analysis of more than one study at a time (a detailed comparison to existing applications is provided in Supplementary Section S1). The reanalysis of multiple experiments therefore requires either an extremely time-consuming manual process or custom programming, barring investigators from engaging with multiple GEO data sets in a systematic way.

ScanGEO, by contrast, uses a database query to identify relevant GEO data sets, then reanalyzes each qualifying data set for differential expression of a custom list of genes or all genes in a Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway (Kanehisa and Goto, 2000; Kanehisa *et al.*, 2016, 2017). We currently support access to curated gene expression array data sets from the top 20 organisms, which covers essentially all GEO DataSets (GDS). Scan output includes summary tables with the number of significant studies per gene and number of significant genes per study as well as downloadable files with descriptions of selected studies, p-values for user specified genes, maximum absolute $\log_2$ fold changes for user genes, as well as additional files for differentially expressed genes including dot plots and CSV files of raw data to provide interoperability with other analysis software.

In summary, ScanGEO is an easy-to-use, interactive and interoperable Shiny web application that significantly accelerates the analysis of publicly available high-throughput data for hypothesis generation and validation of experimental data.

2 Materials and methods

The source code for ScanGEO is written in R (R Core Team, 2016) and implemented as a Shiny App using the 'shiny' package (Chang *et al.*, 2017). For users wishing to create their own local instance of

ScanGEO, the source code and installation instructions are available on Github https://github.com/StantonLabDartmouth/AppScanGEO. Please note that the installation requires ~20 GB of disc space and at least 8 GB of RAM.

The most convenient way to use ScanGEO is through our web server at http://scangeo.dartmouth.edu/ScanGEO/ as follows. A ScanGEO search starts with the selection of an organism in the 'Select Studies' tab. Next, users have the option to scan only studies whose title or description contains a specific word. Clicking on the 'Find GEO data sets' button prompts a database search that uses the 'GEOmetadb' package (Zhu *et al.*, 2008). The number of GEO studies that match the search parameters is then displayed in the 'Table' tab together with summary information for each study including Title, GDS, Pubmed ID, study type, platform organism and date.

Having selected target studies, users next identify target genes by selecting a KEGG pathway, entering custom genes, uploading a list of genes, or any combination of these. Genes belonging to a KEGG pathway can be selected from an organism-specific drop-down menu in the 'KEGG Pathway' tab. Lists containing all KEGG pathways for a given organism were compiled using the 'KEGGREST' package (Tenenbaum, 2016) and currently support all organisms except Oryza. To enter custom genes, the user types into the gene selection field of the 'Custom Genes' tab. This action brings up a scrollable list from which matching gene symbols can be selected. Alternatively, users can select all gene symbols starting with a set of characters (wildcard search) or upload a CSV file with up to 200 species-appropriate gene symbols.

Once target studies and genes have been selected, the estimated scan time is displayed, based on the number of selected studies and genes. For example, ScanGEO can run exploratory statistics for differential expression of a single gene on all human GDS in about half an hour, but scanning 400 genes in all human GDS would take around 2 hours. Therefore, it is advised that users select genes and study criteria with some care.

Finally, users specify a significance threshold for differentially expressed genes using radio buttons, and initiate the scan. The application then tests all selected studies for differential expression of all selected genes using ANOVA. While the algorithm is running, a progress bar appears providing feedback.

Upon scan completion, results tables will appear in the two tabs behind the 'Table' tab. 'Significant Genes' lists the number of studies in which each gene was found to be significantly differentially expressed and 'Significant Studies' shows the number of significant genes for each study. Moreover, downloadable output is available, including a table summarizing selected studies, ANOVA p-values for all user-selected genes and studies, and absolute maximum $\log_2$ fold difference between any two groups in each selected study. If any genes pass the user defined significance threshold, dot plots and data tables for each of these genes are created. Downloadable files are compressed into a ZIP file which also contains a 'README.pdf' file detailing the content of each downloadable. The ScanGEO user manual can be found under the 'Documentation' tab and in Supplementary Section S2.

3 Case study

In this worked example we identified differentially expressed genes in the context of chronic bacterial infection in patients with cystic fibrosis.

1. Start up the ScanGEO shiny app in your browser of choice by going to http://scangeo.dartmouth.edu/ScanGEO/
2. In the 'Select Studies' tab on the left choose 'Homo' from the organism list and enter 'cystic fibrosis' as a search term.
3. Click on 'Find GEO data sets' and notice the appearance of the search results message and the summary table with relevant studies in the 'Table' tab.
4. Select 'Staphylococcus aureus infection' from the 'KEGG Pathway' tab.
5. Switch to the 'Scan' tab, leaving the significance level at 0.05 and click on 'ScanGEO'.
6. After completion of the scan, results tables will display the number of studies in which each gene was differentially expressed in the 'Significant Genes' tab and the number of significant genes for each study in the 'Significant Studies' tab.
7. A ZIP file with all results can be downloaded by clicking on 'Download Results'. The first file in the unzipped results folder (01_README.pdf) describes all files in the folder and contains an example output plot that includes a description of the features of the plot.

In summary, ScanGEO rapidly mines high throughput gene expression data to identify genes and pathways of interest to answer biological questions relevant to diverse fields of study.

Acknowledgements

We thank our beta testers for trying out the app and giving us constructive feedback. We would also like to thank John Wallace for help with the Shiny Server as well as Dean Attali for consultation and advice on implementing the Shiny app.

Funding

This work was supported by the National Heart, Lung and Blood Institute (R01 HL074175 to B.A.S.), the Cystic Fibrosis Foundation (STANTO15R0 to B.A.S.) and the National Institute of Environmental Health Sciences (P42 ES007373 to B.A.S.).

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.

Chang,W. *et al.* (2017) shiny: Web Application Framework for *R*. *R package version 1.0.3*. https://CRAN.R-project.org/package=shiny

Dumas,J. *et al.* (2016) shinyGEO: a web-based application for analyzing gene expression omnibus datasets. *Bioinforma. Oxf. Engl.*, **32**, 3679–3681.

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

Kanehisa,M. *et al.* (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, **44**, D457–D462.

Kanehisa,M. *et al.* (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353–D361.

Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

R Core Team (2016) R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.

Saeed,A.I. *et al.* (2003) TM4: a free, open-source system for microarray data management and analysis. *BioTechniques*, **34**, 374–378.

Tenenbaum,D. (2016) KEGGREST: Client-side REST access to KEGG. *R package version 1.14.1*.

Zhu,Y. *et al.* (2008) GEOmetadb: powerful alternative search engine for the Gene Expression Omnibus. *Bioinforma. Oxf. Engl.*, **24**, 2798–2800.