Bioinformatics, 31(18), 2015, 3060–3062
doi: 10.1093/bioinformatics/btv297
Advance Access Publication Date: 13 May 2015
Applications Note

OXFORD

Systems Biology

GEO2Enrichr: browser extension and server app to extract gene sets from GEO and analyze them for biological functions

Gregory W. Gundersen[1,2,3], Matthew R. Jones[1,2,3], Andrew D. Rouillard[1,2,3], Yan Kou[1,2,3], Caroline D. Monteiro[4], Axel S. Feldmann[1,2,3], Kevin S. Hu[1,2,3] and Avi Ma'ayan[1,2,3,*]

[1]Department of Pharmacology and Systems Therapeutics, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA, [2]BD2K-LINCS Data Coordination and Integration Center, [3]Mount Sinai Knowledge Management Center for Illuminating the Druggable Genome and [4]367 Airport Sector, Goiania, Goias 74075, Brazil

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on January 12, 2015; revised on May 1, 2015; accepted on May 4, 2015

Abstract

Motivation: Identification of differentially expressed genes is an important step in extracting knowledge from gene expression profiling studies. The raw expression data from microarray and other high-throughput technologies is deposited into the Gene Expression Omnibus (GEO) and served as Simple Omnibus Format in Text (SOFT) files. However, to extract and analyze differentially expressed genes from GEO requires significant computational skills.

Results: Here we introduce GEO2Enrichr, a browser extension for extracting differentially expressed gene sets from GEO and analyzing those sets with Enrichr, an independent gene set enrichment analysis tool containing over 70 000 annotated gene sets organized into 75 gene-set libraries. GEO2Enrichr adds JavaScript code to GEO web-pages; this code scrapes user selected accession numbers and metadata, and then, with one click, users can submit this information to a web-server application that downloads the SOFT files, parses, cleans and normalizes the data, identifies the differentially expressed genes, and then pipes the resulting gene lists to Enrichr for downstream functional analysis. GEO2Enrichr opens a new avenue for adding functionality to major bioinformatics resources such GEO by integrating tools and resources without the need for a plug-in architecture. Importantly, GEO2Enrichr helps researchers to quickly explore hypotheses with little technical overhead, lowering the barrier of entry for biologists by automating data processing steps needed for knowledge extraction from the major repository GEO.

Availability and implementation: GEO2Enrichr is an open source tool, freely available for installation as browser extensions at the Chrome Web Store and FireFox Add-ons. Documentation and a browser independent web application can be found at http://amp.pharm.mssm.edu/g2e/.

Contact: avi.maayan@mssm.edu

1 Introduction

Gene expression profiling data are widely and openly available with many computational and statistical techniques developed in the past decade to process and analyze this type of data. The identification of differentially expressed genes and their analysis through gene set enrichment are two important steps in extracting knowledge from expression data. However, there is still a need for creating interoperable tools to improve the reuse of the collective rapidly

accumulating expression datasets for extracting more knowledge from such data.

One of the major repositories for gene expression data is the Gene Expression Omnibus (GEO), a public database maintained by the National Center for Biotechnology Information (NCBI) for depositing and serving functional genomics data, mostly from gene expression profiling studies (Edgar *et al.*, 2002). GEO contains a search engine and each dataset is provided with metadata that includes the publication associated with each dataset, the assay platform, and a description of the experiment and conditions for each sample. GEO stores data in tab-delimited ASCII text files called Simple Omnibus Format in Text (SOFT) files. GEO organizes data into four primary types: GEO sample (GSM), GEO series (GSE), GEO dataset (GDS) and, GEO platform (GPL), which are accessioned with unique and constant identifiers. GSMs contain expression measurements from a single experiment where each sample consists of pairs of probe IDs and their expression values. GSEs organize samples into sets which make up a complete study. GDSs contain more curated and annotated sets of samples from related GSEs. GPLs contain the lists of probe IDs along with mappings of probes to genes for the various expression platforms that are supported. While the interface that provides access to GDSs has functionality to identify differentially expressed genes, it is difficult for a user to obtain a coherent view of the results, or to have the ability to analyze the differentially expressed genes with prior biological knowledge for functional characterization and link to external tools. GSEs provide the data for processing with R using the GEO2R format and this creates a barrier to entry for biologists without computational expertise in R. Here we present GEO2Enrichr, a browser extension and a web-based client/server application that adds functionality to the GEO site for easier extraction of differentially expressed genes, and for piping such differentially expressed genes to the external gene set enrichment analysis tool Enrichr (Chen *et al.*, 2013). GEO2Enrichr's front-end adds buttons and a modal dialog box to GEO's GDS and GSE pages via the extension; whereas the back-end of GEO2Enrichr handles the downloading, processing and analyzing of the expression data from GEO specified by the user input.

2 Methods

GEO2Enrichr is comprised of two parts: A browser extension that embeds functionality into GEO web pages, and a web-based independent application for data processing and storage. The front-end of the browser extension and web-application are written in JavaScript, HTML and CSS. The back-end server-side is written in Python and uses Flask, a framework for serving Python web applications. The deployed project consists of an Apache web server and the Web Server Gateway Interface (WSGI)-compliant application Flask (Grinberg, 2014). The web browser loads the extension when the user navigates to a GEO web page. Upon loading of a GDS or a GSE web-page, GEO2Enrichr embeds itself onto the page as a single button. Once clicked, after sample selection, the button opens a modal dialog box (Fig. 1) that allows users to add metadata about the datasets they selected for analysis, and choose settings for identifying the differentially expressed genes. Once the samples are submitted for an analysis, GEO2Enrichr checks if the values maximum range is greater than 100 to determine if the original data was already transformed to log-scale. Based on this assumption, if the data was not log-transformed, GEO2Enrichr log2 transforms the data. Similarly, GEO2Enrichr checks if any row's median or standard deviation deviates from the

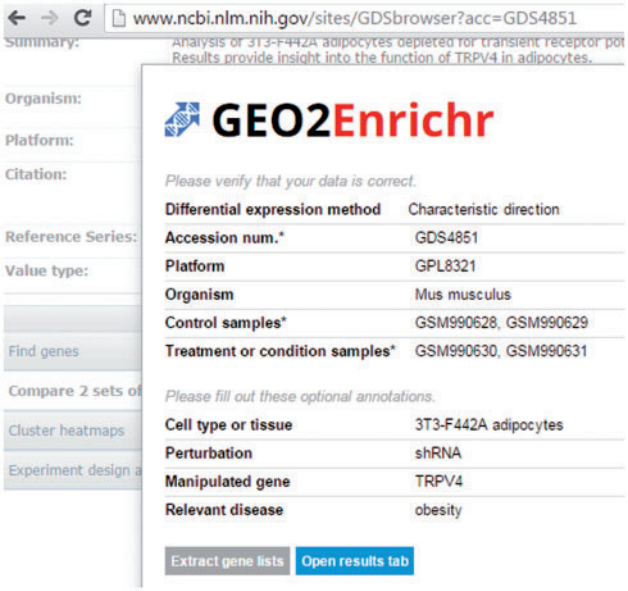

Fig. 1. Screenshot from the GEO2Enrichr modal dialog after the processing of the GDS4851 was completed. In this study TRPV4 was depleted in 3T3-F442A cells

average median or standard deviation by a magnitude greater than 4. If either is true, GEO2Enrichr quantile normalizes the data. Differential expression is then computed using the Characteristic Direction method (Clark *et al.*, 2014). The Characteristic Direction uses a Linear Discriminant Analysis (LDA) classifier, in which the probability that a sample x derives from each of the classes G is modeled with the Bayes' rule. By making the assumption that the class-conditional density is a multivariate Gaussian, and that the covariance matrix for each class is the same, the class boundary is a hyperplane with a normal vector. This vector defines the orientation of the classification boundary hyperplane and the Characteristic Direction approach uses this vector to characterize the differential expression between the two classes. Recently, we demonstrated that the Characteristic Direction method significantly outperforms all the leading methods in the field including: limma, DESeq and SAM when benchmarked these methods with data from transcription factors knockdowns followed by expression and ChIP-seq data collected for the same transcription factors. GEO2Enrichr also provides the T-test as an alternative method to compute differential expression. Future implementations of GEO2Enrichr will add other statistical tests including the ones mentioned above.

To programmatically download files from GEO and access the metadata from each study, GEO2Enrichr queries GEO with the Entrez Programming Utilities (E-Utils) application programming interface (API) (Sayers and Wheeler, 2004). When parsing SOFT files GEO2Enrichr discards lines with null values. It then converts probe IDs to their appropriate gene symbols by using a persistent key-value store that maps probe IDs to gene symbols for many platforms. The key-value store was created from the GPL files on GEO. Probes that cannot map to gene symbols are dropped. Next, the expression values of probes that map to the same gene are averaged. At this stage the differentially expressed genes are identified by the Characteristic Direction method (Clark *et al.*, 2014). Finally, GEO2Enrichr submits the up-, down- and combined-gene-sets to Enrichr for gene set enrichment analysis using the Enrichr API. A results page is opened in a new tab with links to download a cleaned data table, the gene sets and the links to Enrichr.

3 Results

GEO2Enrichr adds useful functionality to the GEO database, lowering the barrier to entry for biologists to extract gene sets from GEO. So far the application was used to quickly create three types of gene set libraries: one for extracting single gene perturbations (2800 signatures), one for extracting drug induced signatures, and one for extracting disease signatures. These three gene-set libraries were created by students from the Coursera (Adams, 2012) course Network Analysis in Systems Biology https://www.coursera.org/course/netsysbio. The students from the course collectively used the tool in a crowdsourcing microtask (Good and Su, 2013) to build these gene gene-set libraries. To submit gene expression signatures, the students utilized a crowdsourcing portal we developed for the BD2K-LINCS Data Coordination and Integration Center available at http://maayanlab.net/crowdsourcing. Comparing single gene knockdowns, drug-induced and disease signatures can be used to identify novel relationships between genes and diseases, and suggest novel drug targets, as well as provide hypotheses for drug repurposing. So far, the GEO2Enrichr Chrome extension was installed by over 700 users. The systematic use of GEO2Enrichr is generating a searchable resource for gene expression signatures that will be used in the future by the broad scientific community including the users of Enrichr.

Funding

This work was supported by NIH grants: R01GM098316, U54HG008230 and U54CA189201 (to A.M.).

Conflict of Interest: none declared.

References

Adams,S. (2012) Is coursera the beginning of the end for traditional higher education? Forbes. October 4.

Chen,E.Y. *et al.* (2013) Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, **14**, 128.

Clark,N.R. *et al.* (2014) The characteristic direction: a geometrical approach to identify differentially expressed genes. *BMC Bioinformatics*, **15**, 79.

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

Good,B.M. and Su,A.I. (2013) Crowdsourcing for bioinformatics. *Bioinformatics*, **29**, 1925–1933.

Grinberg,M. (2014) *Flask Web Development: Developing Web Applications with Python*. O'Reilly Media, Inc, Sebastopol, CA.

Sayers,E. and Wheeler,D. (2004) Building customized data pipelines using the entrez programming utilities (eUtils) NCBI Online books, http://www.ncbi.nlm.nih.gov/books/NBK1058/.