GeoDiver: Differential Gene Expression Analysis & Gene-Set Analysis for GEO Datasets

Ismail Moghul[1], Suresh Hewapathirana[1], Nazrath Nawaz[1], Anisatu Rashid[1], Marian Priebe[1], Fabrizio Smeraldi[2] and Conrad Bessant[1,*]

[1]School of Biological and Chemical Sciences, Queen Mary University of London, United Kingdom

[2]School of Electronic Engineering and Computer Science, Queen Mary University of London, United Kingdom

*To whom correspondence should be addressed.

Abstract

Summary: GeoDiver is an online web application for performing Differential Gene Expression Analysis (DGEA) and Generally Applicable Gene-set Enrichment Analysis (GAGE) on gene expression datasets. Users can either analyse publically available data from the Gene Expression Omnibus (GEO) or upload their own datasets. The output produced includes numerous high-quality interactive graphics; allowing users to easily explore and examine complex datasets instantly. Furthermore, the results produced can be reviewed at a later date and share with collaborators.

Availability: GeoDiver is freely available online at http://www.geodiver.co.uk. The source code is available under the GNU AGPL license on Github: https://github.com/GeoDiver/GEODiver.

1 Introduction

Gene expression analysis is a powerful methodology by which differences in gene expression profiles can be identified between of sample populations. As such, these analyses are routinely used in a variety of research areas including cancer research (e.g. Arpino et al. 2013), human disease research (e.g. Emilsson et al. 2008), and drug development (e.g. Bai et al. 2013).

The recent exponential decrease in sequencing cost and advancement in microarray technology has resulted in an accumulation of large gene expression datasets; many of which are publically available on the Gene Expression Omnibus (GEO) (Barett et al. 2013). Despite having access to numerous datasets, analysing them can be challenging and time-consuming.

Typical analyses on gene-expression data includes differential gene expression analysis (DGEA) and Gene-set Analysis (GSA), which are used to find statistically significant differences in expression for specific genes or gene-sets between sample populations. There are a wide variety of R packages that can be used to analyse gene expression data (e.g. Robinson et al., 2010; Ritchie et al., 2015). However, the use of these R packages and the subsequent generation of visualisations require extensive coding proficiency in the R programming language (R Core Team, 2013).

Existing tools, such as the Geo2R (Barrett et al., 2012) and GSEA (Subramanian et al., 2005), can be used to carry out gene expression analysis, but do not produce graphics for visual analysis.

GeoDiver is an online web application that analyses gene expression datasets using DGEA and GSA. Users can upload their own datasets or retrieve existing datasets from GEO (supporting both GEO Datasets and GEO Data Series). Interactive graphics are produced to visualise the results of both analyses as well as the overall data. These high-quality graphics include interactive (two- and three-dimensional) principal component analysis plots, heatmaps, interaction networks and volcano plots. The results of the two analyses are presented within interactive tables that support fuzzy searching and sorting. These results and graphics can be downloaded. Alternatively, users can also login in order to store their results online, review past analyses and share these with collaborators.

2 Approach

Over the recent years, a fairly standard workflow for analysing and reporting differential gene expression experiments has finally evolved, which is what has been implemented in GeoDiver.

At the first instance, users are required to either enter the accession number of the GEO Dataset or Data Series to be analysed or upload their own dataset (in the NCBI GEO Standard format).

Upon importing the data, GeoDiver initially carries out a preliminary analysis to determine whether the data needs to be scaled and automatically log-transforms the data if necessary. GeoDiver additionally examines the data for missing values and imputes any missing values using KNN imputation.

A principal component analysis is carried out showing the presence of any discrete clusters in the data. Users can explore these clusters in two- or three-dimensions. Moreover, the individual samples are analysed to determine the likelihood for it being an outlier.

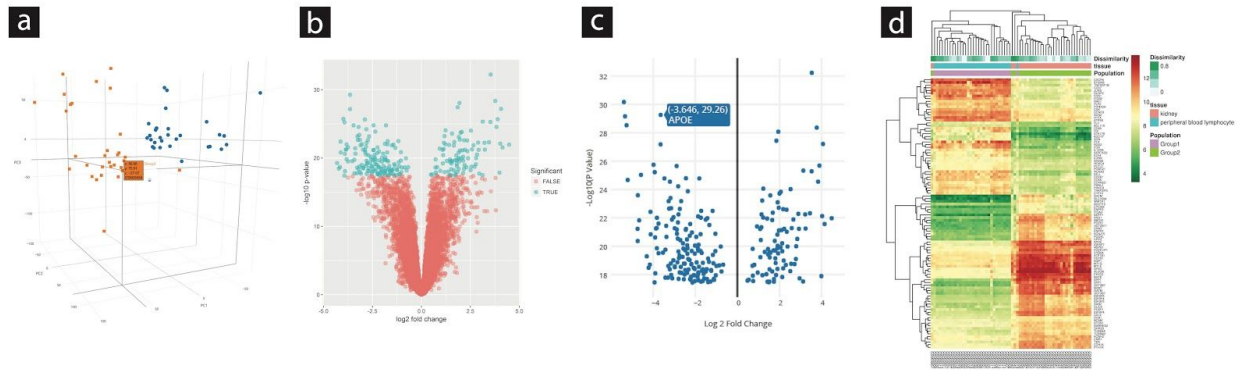

Fig. 1. Exemplar GeoDiver Visualisations: Above are a few exemplar visualisations produced by GeoDiver. a) An Interactive 3D PCA plot showing data points for the first three components, b) a Volcano plot, c) an interactive volcano plot showing significant genes and d) a heatmap of the top differentially expressed genes. Other visualisations produced by GeoDiver can be seen in the Supplementary.

2.1 Differential Gene Expression Analysis

GeoDiver identifies differentially expressed genes by fitting a linear model to each gene which estimates the fold change in expression while accounting for standard errors by applying empirical Bayes smoothing. Genes are then ordered according to the difference in the expression values between the two sets of samples selected.

This information is presented as an interactive table, a heatmap and a volcano plot. Upon clicking on a gene, users are provided with an interactive bar chart displaying the gene expression levels for each sample expressing the gene. The Volcano plot has added interactivity showing the gene name, fold-change and p-value of each data point.

2.2 Generally Applicable Gene-Set Enrichment Analysis

Generally Applicable Gene-set Enrichment for Pathway Analysis (GAGE) is a variation of gene set enrichment analysis. Instead of sample randomization, it uses gene randomization, making it able to carry out accurate analyses of smaller datasets (i.e. datasets with few samples) (Luo et al., 2009).

GeoDiver utilises the KEGG (Kanehisa and Goto, 2000) and Gene Ontology (Gene Ontology Consortium, 2004) databases to identify the pathways each gene in the data is associated with. Using this information, GeoDiver is able to identify KEGG or GO pathways that are significantly differentially expressed between the sample populations

selected.

This information is presented as an interactive table and a heatmap. For analyses based on the KEGG database, a colour-coded interaction network for the pathway is additionally produced.

3 Usage & Implementation

GeoDiver provides users with a minimalistic graphical interface designed in accordance with Google's Material Design Specifications.

Initially, users are either required to upload their own gene expression dataset or enter a GEO accession number to retrieve a publicly available Dataset or Data Series. Users can then select the sample populations they wish to compare and immediately start their analysis. Users are also provided with several options to customise and fine-tune their analysis by being able to change over 20 parameters. Hovering the cursor over these options provides the user with a short explanation.

A standard analysis takes a few minutes to complete, and the results are displayed on the same page. Amongst the results produced for each analysis are high-quality vector-based visualisations, tab delimited tables and R data objects that can be downloaded.

Additionally, users are provided with the option to log in (using a Google Account), which allows them to review as well as share their previous analyses via a URL. GeoDiver's web application has been written in the Ruby programming language (largely based on GeneValidator (Dragan et al., 2016)) while the DGEA and GAGE analyses in are in R (see supplementary figure 1).

4 Discussion

Without adjusting any advanced parameters users can easily run both DGEA and GSA within a few minutes of accessing the GeoDiver web application. For users who are familiar with gene-expression data, the option to adapt the advanced parameters of the analyses mirrors the flexibility that one might expect from writing a custom script. This flexibility along with the helpful parameter explanations enables users, without the ability to write their own code, to still perform powerful analyses.

GeoDiver allows researchers to fully take advantage of growing online resources, such as GEO, without the need for downloading or installing additional software, learning command line skills or having prior programming knowledge.

Several of the graphics produced are interactive; this helps users to interpret and understand the data they have analysed. Other graphics such as the heatmaps are high-resolution, information rich and can be easily exported for use in publications. GeoDiver is therefore not only designed to facilitate the analysis of gene-expression data but also to ensures that users are able to fully explore the results of their analysis. This is important as the ability to use such powerful analytical tools has to be paired with the corresponding ability to interpret the output. GeoDiver's utility is extended by allowing users to view and rerun (perhaps changing some parameters) their past analyses.

Acknowledgements

We would like to take this opportunity to thank all academics on the MSc Bioinformatics course at QMUL (http://www.qmul.ac.uk/postgraduate/taught/coursefinder/courses/121410.html) - the six-week Software Development Group Project in the Master's course resulted in the conception of GeoDiver. We also want to thank Elena Chatzimichali for her useful feedback.

Funding

This research utilised Queen Mary's MidPlus computational facilities, supported by QMUL Research-IT and funded by EPSRC grant EP/K000128/1.

References

Barrett, T. et al. (2012). Ncbi geo: archive for functional genomics data sets–update.

Arpino, G. et al. (2013). Gene expression profiling in breast cancer: A clinical perspective. The Breast, 22(2), 109–120.

Bai, J. P. F. et al. (2013). Strategic Applications of Gene Expression: From Drug Discovery/Development to Bedside. The AAPS Journal, 15(2), 427–437.

Barrett, T. et al. (2013). NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Research, 41(D1), D991–D995.

Dragan, M.-A. et al. (2016). GeneValidator: identify problems with protein-coding gene predictions. Bioinformatics (Oxford, England), pages btw015.

Emilsson, V. et al. (2008). Genetics of gene expression and its effect on disease. Nature, 452(7186), 423–428.

Gene Ontology Consortium (2004). The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research, 32(90001), 258D–261.

Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes.Nucleic acids research, 28(1), 27–30.

Luo, W. et al. (2009). GAGE: generally applicable gene set enrichment for pathway analysis. BMC bioinformatics, 10, 161.

R Core Team (2013). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ritchie, M. E. et al. (2015). limma powers differential expression analyses for RNA sequencing and microarray studies. Nucleic Acids Research, 43(7), e47–.

Robinson, M. D. et al. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics (Oxford, England), 26(1), 139–40.

Subramanian, A. et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences, 102(43), 15545–15550.