BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 1 2014, pages 127–128
doi:10.1093/bioinformatics/btt592

Gene expression Advance Access publication October 15, 2013

gCMAP: user-friendly connectivity mapping with R

Thomas Sandmann, Sarah K. Kummerfeld, Robert Gentleman and Richard Bourgon*
Department of Bioinformatics and Computational Biology, Genentech Inc., South San Francisco, CA 94080, USA

Associate Editor: Ziv Bar-Joseph

ABSTRACT

Connections between disease phenotypes and drug effects can be made by identifying commonalities in the associated patterns of differential gene expression. Searchable databases that record the impacts of chemical or genetic perturbations on the transcriptome—here referred to as 'connectivity maps'—permit discovery of such commonalities. We describe two R packages, gCMAP and gCMAPWeb, which provide a complete framework to construct and query connectivity maps assembled from user-defined collections of differential gene expression data. Microarray or RNAseq data are processed in a standardized way, and results can be interrogated using various well-established gene set enrichment methods. The packages also feature an easy-to-deploy web application that facilitates reproducible research through automatic generation of graphical and tabular reports.

Availability and implementation: The *gCMAP* and *gCMAPWeb* R packages are freely available for UNIX, Windows and Mac OS X operating systems at Bioconductor (http://www.bioconductor.org).

Contact: bourgon.richard@gene.com

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

Received on May 8, 2013; revised on August 12, 2013; accepted on October 10, 2013

1 INTRODUCTION

Microarray and RNAseq technologies enable the study of transcriptional changes triggered by experimental perturbations. Gene expression changes observed after one perturbation often resemble those produced by others, suggesting a shared biological mechanism (Lamb *et al*., 2006). This strategy for establishing links between perturbations can then be used, for example, to identify bioactive compounds (Kunkel *et al*., 2011; Stumpel *et al*., 2012), elucidate their mode of action (Coombs *et al*., 2012) or reposition approved pharmaceuticals for use in new indications (Dudley *et al*., 2011). Standalone tools (Subramanian *et al*., 2005), modular software libraries (Pacini *et al*., 2013) or hosted web applications are available to query the original Connectivity Map (Lamb *et al*., 2006), but tools that allow users to easily compile and search their own reference datasets are lacking.

The *gCMAP* package provides utilities and memory-efficient structures to create, store and query large experimental datasets. *gCMAP* also provides a unified R command-line interface to multiple gene set enrichment (GSE) approaches, permitting their use for querying user-created connectivity maps as well as more traditional gene set collections—e.g. Gene Ontology, Reactome or WikiPathways (Jupe *et al*., 2012; Kelder *et al*., 2012). In addition, the companion *gCMAPWeb* package provides a customizable browser-based graphical user interface, and can leverage R's built-in web server or be integrated into a production-scale server.

2 FEATURES

To compile a new connectivity map, users supply expression data in the form of standard Bioconductor *ExpressionSet* objects. *gCMAP* provides convenience functions to split large studies into individual perturbation instances and process them in batch, taking advantage of the widely used *limma* (Smyth, 2004) and *DESeq* (Anders and Huber, 2010) Bioconductor packages. QC metrics permit removal of unsuitable experiments, e.g. those suffering from normalization artifacts (Supplementary Material). To allow memory-efficient access to large datasets, results are stored on disk and data subsets are loaded on demand.

gCMAP represents gene set membership and (optionally) direction of expression change via sparse numerical matrices, enabling efficient GSE analysis using matrix operations. Analysis of complete differential expression profiles, which retains more information than gene sets, is also supported, permitting permutation-based assessment of statistical significance.

gCMAP implements several well-established GSE methods, including Fisher's exact test, the original GSEA statistic (Lamb *et al*., 2006), the JG score (Jiang and Gentleman, 2007), mgsa (Bauer *et al*., 2011) and Roast (Wu *et al*., 2010). All *gCMAP* implementations use common input and output formats, enabling direct comparisons among the supported statistics.

The *gCMAPWeb* package complements *gCMAP* by providing a graphical user interface through a distributable web application. *gCMAPWeb* leverages R's built-in web server, permitting single-user access on any computer running R, and can also be integrated into an Apache web server in a multiuser environment. To submit queries, users can paste gene identifiers directly into a web form or upload text files (Supplementary File S1). *gCMAPWeb* returns results in graphical and tabular form, with detail at the gene set and single-gene levels (Fig. 1). Every report, including graphs, tables and a binary R object, is available for download. The *gCMAPWeb* user interface is automatically configured to match reference datasets but can be easily customized.

3 EXAMPLES

We used *gCMAP* to reconstruct the original Connectivity Map Lamb *et al*. (2006), containing 453 individual perturbations in five human cell lines, from raw microarray data (ArrayExpress E-GEOD-5258, Supplementary File S1). We then used *gCMAPWeb* to query this dataset with genes found to be

*To whom correspondence should be addressed.

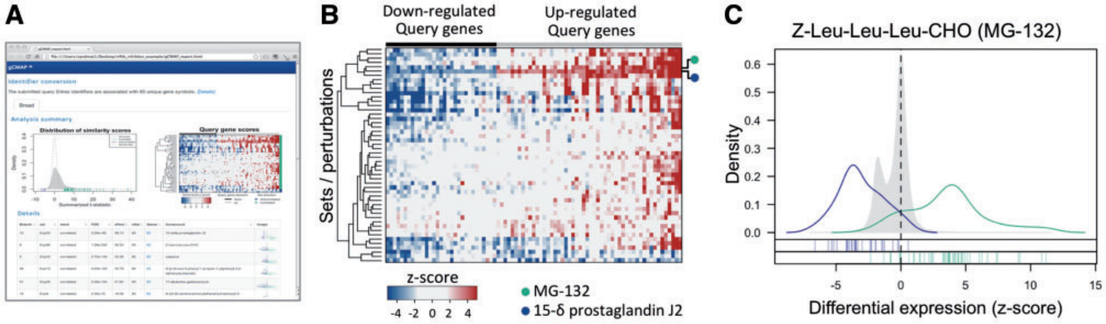

Fig. 1. Output returned by *gCMAPWeb* for a directional query with genes up- or downregulated in MCF7 cells after treatment with 15-δ prostaglandin J2 for 6 h. **(A)** *gCMAPWeb*'s main result page. **(B)** Heatmap displaying differential expression scores for genes (columns) from the top 50 most similar experiments in the Broad Connectivity Map (rows). Treatment with MG-132 (green dot) received the highest JG summary score and was clustered next to the experiment corresponding to the query itself (blue dot). **(C)** Distribution of Z-scores indicating differential expression following MG-132 treatment for all assayed genes (gray) and up- (green) or downregulated (blue) query genes. Query gene scores are shifted toward positive and negative scores

up- or downregulated in MCF7 cells treated with the NFκB inhibitor 15-δ prostaglandin J2. Top hits included MG-132/Z-Leu-Leu-Leu-CHO and celastrol, two known inhibitors of the same pathway (Fig. 1).

To assess differential expression from count data, *gCMAP* leverages the widely used *DESeq* package (Anders and Huber, 2010). To re-examine the transcriptional response of HepG2 cells to the carcinogen Benzo[a]pyrene [van Delft *et al.* (2012), ENA accession SRP011233, Supplementary File S2], we constructed a local gene set collection from WikiPathways (Kelder *et al.*, 2012) and queried it with genes significantly up- or downregulated in response to Benzo[a]pyrene. In concordance with the original study, *gCMAPWeb* reported significant overlap between the query and the Benzo[a]pyrene metabolism and Nrf2/Keaf1 pathways [van Delft *et al.* (2012), Supplementary File S1].

4 CONCLUSION

The *gCMAP* Bioconductor packages combine powerful command-line functionality with a convenient portable web application. The efficient handling of large datasets empowers users to assemble large connectivity maps from private or public data, query them programmatically or through an interactive user interface and store queries and results in a reproducible report.

ACKNOWLEDGEMENTS

The authors thank David Dornan and Harinder Singh for helpful discussions, and Michael Lawrence, Kyongryun Lee and Qixin Bei for support in developing *gCMAPWeb*.

Conflict of Interest: none declared.

REFERENCES

Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.

Bauer,S. *et al.* (2011) Model-based gene set analysis for bioconductor. *Bioinformatics*, **27**, 1882–1883.

Coombs,G.S. *et al.* (2012) Modulation of wnt/-catenin signaling and proliferation by a ferrous iron chelator with therapeutic efficacy in genetically engineered mouse models of cancer. *Oncogene*, **31**, 213–225.

Dudley,J.T. *et al.* (2011) Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Sci Transl Med*, **3**, 96ra76.

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

Jupe,S. *et al.* (2012) Reactome—a curated knowledgebase of biological pathways: megakaryocytes and platelets. *J. Thromb. Haemost.*, [Epub ahead of print, doi: 10.1111/j.1538-7836.2012.04930.x., September 17, 2012].

Kelder,T. *et al.* (2012) Wikipathways: building research communities on biological pathways. *Nucleic Acids Res.*, **40**, D1301–D1307.

Kunkel,S.D. *et al.* (2011) mrna expression signatures of human skeletal muscle atrophy identify a natural compound that increases muscle mass. *Cell Metab.*, **13**, 627–638.

Lamb,J. *et al.* (2006) The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929–1935.

Pacini,C. *et al.* (2013) Dvd: an r/cytoscape pipeline for drug repurposing using public repositories of gene expression data. *Bioinformatics*, **29**, 132–134.

Smyth,G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3** Article3, [Epub ahead of print, February 12, 2004].

Stumpel,D.J. *et al.* (2012) Connectivity mapping identifies hdac inhibitors for the treatment of t(4;11)-positive infant acute lymphoblastic leukemia. *Leukemia*, **26**, 682–692.

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.

van Delft,J. *et al.* (2012) Rna-seq provides new insights in the transcriptome responses induced by the carcinogen benzo[a]pyrene. *Toxicol. Sci.*, **130**, 427–439.

Wu,D. *et al.* (2010) Roast: rotation gene set tests for complex microarray experiments. *Bioinformatics*, **26**, 2176–2182.