BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 17 2012, pages 2272–2273
doi:10.1093/bioinformatics/bts430

Gene expression

Advance Access publication July 10, 2012

Gemma: a resource for the reuse, sharing and meta-analysis of expression profiling data

Anton Zoubarev[1,†], Kelsey M. Hamer[1,†], Kiran D. Keshav[2,†], E. Luke McCarthy[1], Joseph Roy C. Santos[1], Thea Van Rossum[1], Cameron McDonald[1], Adam Hall[3], Xiang Wan[1], Raymond Lim[1,3], Jesse Gillis[1] and Paul Pavlidis[1,*]

[1]Department of Psychiatry and Centre for High-throughput Biology, 2185 East Mall, University of British Columbia, Vancouver, British Columbia, V6T 1Z4 Canada, [2]Department of Biomedical Informatics, Columbia University, New York, NY, 10032 USA and [3]Graduate Program in Bioinformatics, University of British Columbia, Vancouver, British Columbia, V5Z 4S6 Canada

Associate Editor: Trey Ideker

ABSTRACT

Summary: Gemma is a database, analysis software system and web site for genomics data re-use and meta-analysis. Currently, Gemma contains analyzed data from over 3300 expression profiling studies, yielding hundreds of millions of differential expression results and coexpression patterns (correlated expression) for retrieval and visualization. With optional registration users can save their own data and securely share it with other users. Web services and integration with third-party resources further increase the scope of the tools, which include a Cytoscape plugin.

Availability: http://chibi.ubc.ca/Gemma, Apache 2.0 license.

Contact: paul@chibi.ubc.ca

Received on April 2, 2012; revised on June 28, 2012; accepted on July 3, 2012

1 SCOPE AND DATA SOURCES

The goal of Gemma is to enable the rapid exploration and analysis of large quantities of genomics data, leveraging the extensive data available from other public bioinformatics resources such as the Gene Expression Omnibus (Barrett *et al*., 2007).

Currently, Gemma contains nearly 4000 expression profiling studies ('datasets'; in total over 170 000 assays, from eight taxa). Multiple technology types are supported, such as array-based platforms and RNA sequencing. To enable comparisons across platforms, we perform sequence analysis and gene assignment based on the current genome annotations (Barnes *et al*., 2005). Each public dataset undergoes automated (French *et al*., 2009) and manual annotation using controlled vocabularies such as the Disease Ontology (Schriml *et al*., 2012), adding information about the experimental design to allow group comparisons. Additional quality control steps to detect outlier samples or datasets with large batch effects are also performed.

Each dataset is then analyzed for differential expression (e.g. between conditions or tissues) and coexpression (correlation of expression levels across samples). Differential expression is computed using a standard multivariate linear modeling approach (Pavlidis and Noble, 2001) comparing each condition in a dataset with baseline, accommodating complex factorial designs and continuous covariates. Coexpression is computed for each dataset and stored as a set of 'coexpression links' that meet stringent statistical criteria (Lee *et al*., 2004). The results of these analysis are stored in the system for user search and retrieval.

2 FUNCTIONALITY

A main entry point for Gemma is a form that allows users to search for differential expression or coexpression results. The search facilities enable analysis of selected genes [by symbols, key words or Gene Ontology terms (Ashburner *et al*., 2000)] and experiments (based on free text or our annotations of disease, treatment, tissue, etc.). Users can flexibly organize genes or datasets into groups. With optional registration, these groups persist across sessions and can be securely shared with other users.

Differential expression results are presented in a matrix visualization displaying the genes in rows and individual conditions across studies in columns (Fig. 1). For each gene, a 'meta-*P*-value' is provided. For each condition's differentially expressed genes, Gemma provides information on the enrichment of the user's selected genes in that pattern. The data view can be filtered, sorted and exported as an image. Visualizations of the underlying expression patterns are also readily obtained (Fig. 1).

For coexpression searches, Gemma applies a user-settable threshold for how many datasets a link must be observed in before it is displayed. Coexpression results are shown in a tabular format and as an interactive network view (Fig. 1) (Lopes *et al*., 2010). Gemma uses the concept of node degree (how many links a gene has; i.e. 'hubbiness') to assist the user in gauging the importance of an observation (Gillis and Pavlidis, 2011). This is important because in a query-driven network view, only a tiny subset of the network is displayed. For each gene, Gemma estimates the overall node degree and indicates low-node degree genes in darker shades (Fig. 1).

Gemma offers many other features for exploration and analysis. For each gene, an overview page shows the datasets in which the gene is differentially expressed, genes with which the

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

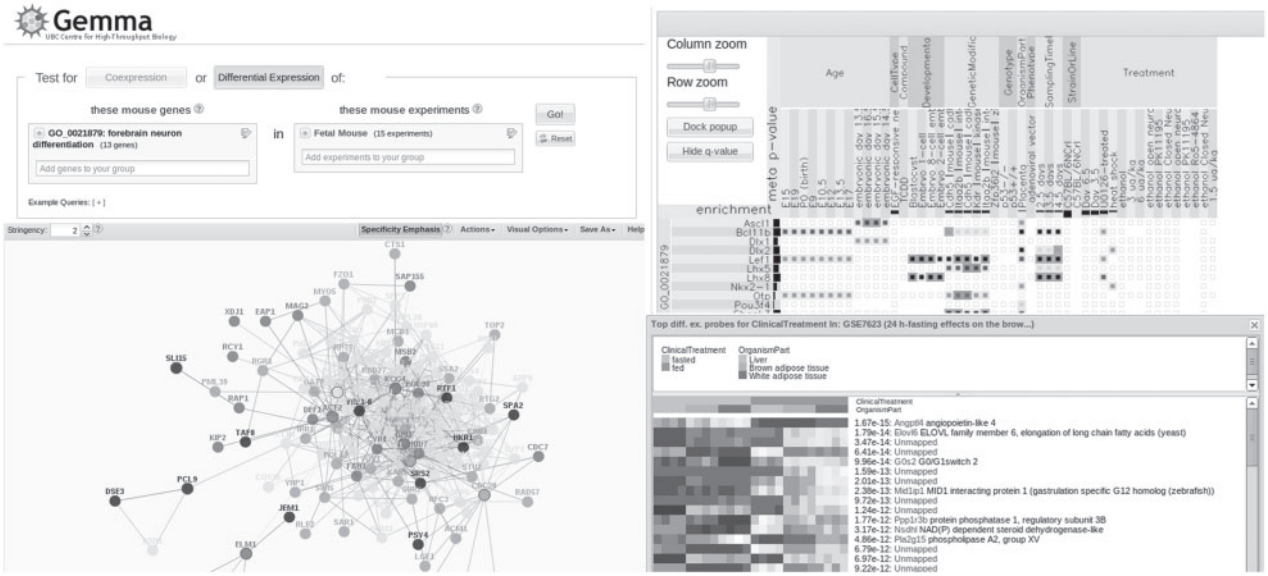

Fig. 1. Screen shots from Gemma illustrating (clockwise from top left) the main search form, the output of query for differential expression, a heatmap view of expression profiles and coexpression query results.

gene is reproducibly coexpressed and expression platforms on which the gene is represented (e.g. http://www.chibi.ubc.ca/Gemma/g/?id=14676). Similarly, for each dataset, Gemma provides annotations, summaries of the analyses and visualizations (e.g. http://www.chibi.ubc.ca/Gemma/ee/?id=1570). Registered users can also upload their own expression datasets to be included for meta-analysis and are provided with an extensive suite of administration tools for data management.

Gemma was designed with data reuse and extensibility in mind. Results can be downloaded in tab-delimited formats for external analysis, and web services are available to access Gemma programmatically. As an example of such integration and to allow more advanced visualization and analysis of coexpression data from Gemma, we have developed 'GemScape,' a plugin for the popular network analysis tool Cytoscape (Kohl *et al.*, 2011). Data from Gemma are also currently available through the Neuroscience Information Framework (Gupta *et al.*, 2008) (differential expression results) and inSilicoDb (Taminau *et al.*, 2011) (experimental design annotations).

Tutorials, in-line help and a wiki with additional user manuals and system information are available through the Gemma web site.

ACKNOWLEDGEMENTS

We thank Suzanne Lane, Artemis Lai, Willie Kwok, Celia Siu, Cathy Kwok, Yiqi Chen, Roland Au, Lydia Xu, Tamryn Loo, Olivia Marais and Tianna Koreman for curation and testing. We also thank Louise Donnison, Vaneet Lotay, Gavin Ha, Meeta Mistry, Leon French and David Quigley for code contributions and Elodie Portales-Casamar and Sanja Rogic for comments on this article. Finally, we thank many investigators who have made their data publicly available.

Funding: National Institutes of Health (GM076990); Canadian Foundation for Innovation; Michael Smith Foundation for Health Research and Canadian Institutes for Health Research.

Conflict of Interest: none declared.

REFERENCES

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Barnes,M. *et al.* (2005) Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. *Nucleic Acids Res.*, **33**, 5914–5923.

Barrett,T. *et al.* (2007) NCBI GEO: mining tens of millions of expression profiles—database and tools update. *Nucleic Acids Res.*, **35**, D760–D765.

French,L. *et al.* (2009) Application and evaluation of automated semantic annotation of gene expression experiments. *Bioinformatics (Oxford, England)*, **25**, 1543–1549.

Gillis,J. and Pavlidis,P. (2011) The impact of multifunctional genes on "guilt by association" analysis. *PLoS One*, **6**, e17258.

Gupta,A. *et al.* (2008) Federated access to heterogeneous information resources in the Neuroscience Information Framework (NIF). *Neuroinformatics*, **6**, 205–217.

Kohl,M., Wiese,S. and Warscheid,B. (2011) Cytoscape: software for visualization and analysis of biological networks. *Methods Mol. Biol.*, **696**, 291–303.

Lee,H.K. *et al.* (2004) Coexpression analysis of human genes across many microarray data sets. *Genome Res.*, **14**, 1085–1094.

Lopes,C.T. *et al.* (2010) Cytoscape Web: an interactive web-based network browser. *Bioinformatics (Oxford, England)*, **26**, 2347–2348.

Pavlidis,P. and Noble,W.S. (2001) Analysis of strain and regional variation in gene expression in mouse brain. *Genome Biol.*, **2**, RESEARCH0042.

Schriml,L.M. *et al.* (2012) Disease ontology: a backbone for disease semantic integration. *Nucleic Acids Res.*, **40**, D940–D946.

Taminau,J. *et al.* (2011) inSilicoDb: an R/Bioconductor package for accessing human Affymetrix expert-curated datasets from GEO. *Bioinformatics (Oxford, England)*, **27**, 3204–3205.