SOFTWARE

MicroScope: ChIP-seq and RNA-seq software analysis suite for gene expression heatmaps

Bohdan B. Khomtchouk[1*], James R. Hennessy[2], Vytas Dargis-Robinson[3] and Claes Wahlestedt[1]

*Correspondence:
b.khomtchouk@med.miami.edu
[1]Center for Therapeutic Innovation and Department of Psychiatry and Behavioral Sciences, University of Miami Miller School of Medicine, 1120 NW 14th ST, Miami, FL, USA 33136
Full list of author information is available at the end of the article

Abstract

We propose a user-friendly ChIP-seq and RNA-seq software suite for the interactive visualization and analysis of gene expression heatmaps, including integrated features to support: principal component analysis, differential expression analysis, gene ontology analysis, and dynamic network visualization of genes directly from a heatmap.

MicroScope is hosted online as an R Shiny web application based on the D3 JavaScript library: http://microscopebioinformatics.org/. The methods are implemented in R, and are available as part of the MicroScope project at: https://github.com/Bohdan-Khomtchouk/Microscope.

Background

Most currently existing heatmap software produce static heatmaps (Saeed et al. 2003, Reich et al. 2006, Verhaak et al. 2006, Qlucore, GENE-E, Chu et al. 2008, Khomtchouk et al. 2014), without features that would allow the user to dynamically interact with, explore, and analyze the landscape of a heatmap via integrated tools supporting user-friendly analyses in differential expression, principal components, gene ontologies, and networks. Such features would allow the user to engage the heatmap data in a visual and analytical manner while in real-time, thereby allowing for a deeper, quicker, and more comprehensive data exploration experience.

An interactive, albeit non-reproducible heatmap tool was previously employed in the study of the transcriptome of the *Xenopus tropicalis* genome (Tan et al. 2013). Likewise, manual clustering of dot plots depicting RNA expression is an integral part of the Caleydo data exploration environment (Turkay et al., 2014). Chemoinformatic-driven clustering can also be toggled in the user interface of Molecular Property Explorer (Kibbey and Calvet, 2005). Furthermore, an interactive heatmap software suite was previously developed with a focus on cancer genomics analysis and data import from external bioinformatics resources (Perez-Llamas & Lopez-Bigas, 2011). Most recently, a general-purpose heatmap software providing support for transcriptomic, proteomic and metabolomic experiments was developed using the R Shiny framework (Babicki et al. 2016).

Moreover, an interactive cluster heatmap library, InCHlib, was previously proposed for cluster heatmap exploration (Škuta et al. 2014), but did not provide built-in support for gene ontology, principal component, or network analysis. However, InCHlib concentrates primarily in chemoinformatic and biochemical data clustering analysis, including the visualization of microarray and protein data. On the contrary, MicroScope is designed specifically for ChIP-seq and RNA-seq data visualization

and analysis in the differential expression, principal component, gene ontology, and network analysis domains. In general, prior software has concentrated primarily in hierarchical clustering, searching gene texts for substrings, and serial analysis of genomic data, with no integrated features to support the aforementioned built-in features (Saldanha 2004, Caraux and Pinloche 2005, Wu et al. 2010).

As of yet, no free, open-source heatmap software has been proposed to explore heatmaps at such multiple levels of genomic analysis and interactive visualization capacity. Here we propose a user-friendly genome software suite designed to handle dynamic, on-the-fly JavaScript visualizations of gene expression heatmaps as well as their respective differential expression analysis, principal component analysis, gene ontology analysis, and network analysis of genes.

Implementation

MicroScope is hosted online as an R Shiny web server application. MicroScope may also be run locally from within R Studio, as shown here: https://github.com/Bohdan-Khomtchouk/Microscope. MicroScope leverages the cumulative utility of R's d3heatmap (Cheng et al. 2015), shiny (Chang et al. 2015), stats (R Core Team, 2015), htmlwidgets (Vaidyanathan et al. 2015), RColorBrewer (Neuwirth, 2014), dplyr (Wickham et al. 2015), data.table (Dowle et al. 2015), goseq (Young et al. 2010), GO.db (Carlson, 2016a), and networkD3 (Gandrud et al. 2015) libraries to create an integrative web browser-based software experience requiring absolutely no programming experience from the user, or even the need to download R on a local computer.

MicroScope employs the Bioconductor package edgeR (Robinson et al. 2010) to create a one-click, built-in, user-friendly differential expression analysis feature that provides differential expression analysis of gene expression data based on Fischer's exact test and the Benjamini & Hochberg correction. This supplies the user with rank-based information about nominal p-value, false discovery rate, fold change, and counts per million in order to establish which specific genes in the heatmap are differentially expressed with a high degree of statistical significance. This information, in turn, is used to investigate the top gene ontology categories of differentially expressed genes, which can then be conveniently visualized as interactive network graphics. Finally, MicroScope provides user-friendly support for principal component analysis via the generation of biplots, screeplots, and summary tables. PCA is supported for both covariance and correlation matrices via R's prcomp() function in the stats package.

Results & Discussion

Figure 1 shows the MicroScope user interface (UI) in action. MicroScope allows the user to magnify any portion of a heatmap by a simple click-and-drag feature to zoom in, and a click-once feature to zoom out. MicroScope is designed with large gene expression heatmaps in mind, where individual gene labels overlap and render the text unreadable. However, MicroScope allows the user to repeatedly zoom in to any sector of the heatmap to investigate a region, cluster, or even a single gene. MicroScope also allows the user to hover the mouse pointer over any specific gene to show gene name, expression level, and column ID.

One of the user-friendly features within MicroScope is that it is responsive to the demands asked of it by the user. For example, gene ontology analysis buttons are not provided in the UI until a user both generates a heatmap and runs differential expression analysis on its contents, both of which constitute prerequisite steps required prior to conducting a successful gene ontology analysis. In other words, MicroScope is user-responsive in the sense that it automatically unlocks new features only as they become needed when the user progresses through successive stages in the software. Furthermore, MicroScope automatically provides short and convenient written guidelines directly in the UI to guide the user on the next steps in the usage of the software. As such, complex analytical operations can be performed by the user in a friendly, step-by-step fashion, each time facilitated by the help of the MicroScope software suite, which adjusts to the needs of the user and provides written guidelines on the next steps to pursue.

Some of MicroScope's initial login features include:

- Sample input file download widget
- User-specified file input widget
- Buffer size widget
- log_2 data transformation widget
- Multiple heatmap color schemes widget
- Hierarchical clustering widget
- Row/column dendrogram branch coloring widget
- Row/column font size widget
- Heatmap download widget

After a user inputs an RNA-seq/ChIP-seq data file containing read counts per gene per sample, a heatmap is automatically produced in the Heatmap panel and a PCA suite (Figure 2) automatically appears, as well as a differential expression analysis suite. Details about the 'Specify Control Samples' widget and the ensuing differential expression analysis (Figure 3) are provided in-depth for users in the Instructions panel of the MicroScope software. It should be noted that the differential expression analysis (Fischer's exact test and Benjamini & Hochberg correction) is broadly applicable to be run on any ChIP-seq or RNA-seq data inputted by the user. However, since the differential expression analysis feature of MicroScope is based on raw read counts from RNA-seq/ChIP-seq experiments (not raw/normalized fluorescence intensity values from microarray experiments), users with microarray data are advised to only use MicroScope's heatmap and PCA utilities. Likewise, prior to visualizing heatmaps in MicroScope, experiment-specific data normalization procedures are left to the discretion of the user (Conesa et al. 2016, Soneson & Delorenzi 2013, Bailey et al. 2013, Shin et al. 2013), depending on whether the user wants to visualize differences in magnitude among genes or see differences among samples.

Following the successful completion of the differential expression analysis, a user is automatically supplied with five more UI widgets:

- Genome database widget
- Gene identifier widget
- Number of top gene ontologies widget
- Gene ontology stratification widget
- Gene ontology p-value cutoff widget

- Gene ontology FDR cutoff widget

Specifying values for these features and clicking the Do Gene Ontology Analysis button returns a list of the top gene ontology (GO) categories according to these exact specifications set by the user (Figure 4). Supported organisms for GO categories enrichment analysis include: human (Carlson, 2016b), mouse (Carlson, 2016c), rat (Carlson, 2016d), zebrafish (Carlson, 2016e), worm (Carlson, 2016f), chimpanzee (Carlson, 2016g), fly (Carlson, 2016h), yeast (Carlson, 2016i), pig (Carlson, 2016j), bovine (Carlson, 2016k), and canine (Carlson, 2016l).

The successful completion of this step can be followed up by running a network analysis on the top GO categories, thereby generating network graphics corresponding to the number of top gene ontology categories previously requested by the user (Figure 5). Nodes represent either gene names or gene ontology identifiers, and links represent direct associations between the two entities. In addition to serving as a visualization tool, this network analysis capability automatically identifies differentially expressed genes that are present within each top gene ontology, which is a level of detail not readily available by running gene ontology analysis alone. By immediately extracting the respective gene names from each top gene ontology category, MicroScope's network analysis features serve to aid the biologist in identifying the top differentially expressed genes in the top respective gene ontology categories. Figure 6 compares interactive network visualizations of the top two gene ontologies, thereby demonstrating the immediate responsiveness of MicroScope's network graphics to user-specified settings (e.g., number of top gene ontologies to display widget).

Conclusion

We provide access to a user-friendly web application designed to visualize and analyze dynamically interactive heatmaps within the R programming environment, without any prerequisite programming skills required of the user. Our software tool aims to enrich the genomic data exploration experience by providing a variety of complex visualization and analysis features to investigate gene expression heatmaps. Coupled with a built-in analytics platform to pinpoint statistically significant differentially expressed genes, a principal component analysis platform to investigate variation and patterns in gene expression, a gene ontology platform to categorize the top gene ontology categories, and a network analysis platform to dynamically visualize gene ontology categories at the gene-specific level, MicroScope presents a significant advance in heatmap technology over currently available software.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

BBK conceived of the study. BBK, JRH, and VDR wrote the code. CW participated in the management of the source code and its coordination. BBK wrote the paper. All authors read and approved the final manuscript.

Acknowledgements

BBK wishes to acknowledge the financial support of the United States Department of Defense (DoD) through the National Defense Science and Engineering Graduate Fellowship (NDSEG) Program: this research was conducted with Government support under and awarded by DoD, Army Research Office (ARO), National Defense Science and Engineering Graduate (NDSEG) Fellowship, 32 CFR 168a.

Author details

[1]Center for Therapeutic Innovation and Department of Psychiatry and Behavioral Sciences, University of Miami Miller School of Medicine, 1120 NW 14th ST, Miami, FL, USA 33136. [2]Department of Mathematics, University of Miami, 1365 Memorial Drive, Coral Gables, FL, USA 33146. [3]Department of Microbiology and Immunology, University of Miami Miller School of Medicine, 1600 NW 10th Ave, Miami, FL, USA 33136.

References

1. Babicki S, Arndt D, Marcu A, Liang Y, Grant JR, Maciejewski A, Wishart DS: *Heatmapper: web-enabled heat mapping for all*. Nucleic Acids Research 2016, pii: gkw419. [Epub ahead of print].
2. Bailey T, Krajewski P, Ladunga I, Lefebvre C, Li Q, Liu T, Madrigal P, Taslim C, Zhang J: *Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data*. PLoS Computational Biology 2013, 9(11): e1003326.
3. Caraux G, Pinloche S: *Permutmatrix: A Graphical Environment to Arrange Gene Expression Profiles in Optimal Linear Order*. Bioinformatics. 2005, 21: 1280–1281.
4. Carlson M: *GO.db: A set of annotation maps describing the entire Gene Ontology*. 2015. R package version 3.3.0.
5. Carlson M: *org.Hs.eg.db: Genome wide annotation for Human*. 2016. R package version 3.3.0.
6. Carlson M: *org.Mm.eg.db: Genome wide annotation for Mouse*. 2016. R package version 3.3.0.
7. Carlson M: *org.Rn.eg.db: Genome wide annotation for Rat*. 2016. R package version 3.3.0.
8. Carlson M: *org.Dr.eg.db: Genome wide annotation for Zebrafish*. 2016. R package version 3.3.0.
9. Carlson M: *org.Ce.eg.db: Genome wide annotation for Worm*. 2016. R package version 3.3.0.
10. Carlson M: *org.Pt.eg.db: Genome wide annotation for Chimp*. 2016. R package version 3.3.0.
11. Carlson M: *org.Dm.eg.db: Genome wide annotation for Fly*. 2016. R package version 3.3.0.
12. Carlson M: *org.Sc.sgd.db: Genome wide annotation for Yeast*. 2016. R package version 3.3.0.
13. Carlson M: *org.Ss.eg.db: Genome wide annotation for Pig*. 2016. R package version 3.3.0.
14. Carlson M: *org.Bt.eg.db: Genome wide annotation for Bovine*. 2016. R package version 3.3.0.
15. Carlson M: *org.Cf.eg.db: Genome wide annotation for Canine*. 2016. R package version 3.3.0.
16. Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J, RStudio, jQuery Foundation, jQuery contributors, jQuery UI contributors, Otto M, Thornton J, Bootstrap contributors, Twitter Inc, Farkas A, Jehl S, Petre S, Rowls A, Gandy D, Reavis B, Kowal KM, es5-shim contributors, Ineshin D, Samhuri S, SpryMedia Limited, Fraser J, Gruber J, Sagalaev I, R Core Team: *shiny: Web Application Framework for R*. 2015. R package version 0.12.2.
17. Cheng J, Galili T, RStudio Inc, Bostock M, Palmer J: *d3heatmap: Interactive Heat Maps Using 'htmlwidgets' and 'D3.js'*. 2015. R package version 0.6.1.
18. Chu VT, Gottardo R, Raftery AE, Bumgarner RE, Yeung KY: *MeV+R: using MeV as a graphical user interface for Bioconductor applications in microarray analysis*. Genome Biology. 2008, 9: R118.
19. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A: *A survey of best practices for RNA-seq data analysis*. Genome Biology. 2016, 17(13): 1–19.
20. Dowle M, Srinivasan A, Short T, Lianoglou S, Saporta R, Antonyan E: *data.table: Extension of Data.frame*. 2015. R package version 1.9.6.
21. Gandrud C, Allaire JJ, Russell K, Lewis BW, Kuo K, Sese C, Ellis P, Owen J, Rogers J: *networkD3: D3 JavaScript Network Graphs from R*. R package version 0.2.8.
22. Gould J: GENE-E software hosted at the Broad Institute. http://www.broadinstitute.org/cancer/software/GENE-E/.
23. Khomtchouk BB, Van Booven DJ, Wahlestedt C: *HeatmapGenerator: high performance RNAseq and microarray visualization software suite to examine differential gene expression levels using an R and C++ hybrid computational pipeline*. Source Code for Biology and Medicine. 2014, 9(1): 1–6.
24. Kibbey C, Calvet A: *Molecular Property eXplorer: a novel approach to visualizing SAR using tree-maps and heatmaps*. J Chem Inf Model. 2005, 45(2): 523–532.
25. Neuwirth E: *RColorBrewer: ColorBrewer Palettes*. 2014. R package version 1.1-2.
26. Perez-Llamas C, Lopez-Bigas N: *Gitools: analysis and visualisation of genomic data using interactive heat-maps*. PLoS One. 2011, 6: e19541.
27. Qlucore Omics Explorer: The D.I.Y Bioinformatics Software. http://www.qlucore.com.
28. R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
29. Robinson MD, McCarthy DJ, Smyth GK: *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data*. Bioinformatics. 2010, 26: 139–140.
30. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP: *GenePattern 2.0*. Nat Genet. 2006, 38(5): 500–501. 10.1038/ng0506-500.
31. Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M, Sturn A, Snuffin M, Rezantsev A, Popov D, Ryltsov A, Kostukovich E, Borisovsky I, Liu Z, Vinsavich A, Trush V, Quackenbush J: *TM4: a free, open-source system for microarray data management and analysis*. Biotechniques. 2003, 34(2): 374–378.
32. Saldanha AJ: *Java Treeview – extensive visualization of microarray data*. Bioinformatics. 2004, 20(17): 3246–3248.
33. Shin H, Liu T, Duan X, Zhang Y, Liu XS: *Computational methodology for ChIP-seq analysis*. Quantitative Biology. 2013, 1(1): 54–70.
34. Škuta C, Bartůněk P, Svozil D: *InCHlib — interactive cluster heatmap for web applications* Journal of Cheminformatics. 2014, 6(44): 1–9.
35. Soneson C, Delorenzi M: *A comparison of methods for differential expression analysis of RNA-seq data* BMC Bioinformatics. 2013, 14:91.
36. Tan MH, Au KF, Yablonovitch AL, Wills AE, Chuang J, Baker JC, Wong WH, Li JB: *RNA sequencing reveals a diverse and dynamic repertoire of the Xenopus tropicalis transcriptome over development*. Genome Research.

2013, 23: 201–216.
37. Turkay C, Lex A, Streit M, Pfister H, Hauser H: *Characterizing cancer subtypes using dual analysis in Caleydo StratomeX*. IEEE Comput Graph Appl. 2014, 34(2): 38–47.
38. Vaidyanathan R, Xie Y, Allaire JJ, Cheng J, Russell K, RStudio: *htmlwidgets: HTML Widgets for R*. 2015. R package version 0.5.
39. Verhaak RGW, Sanders MA, Bijl MA, Delwel R, Horsman S, Moorhouse MJ, van der Spek PJ, Lowenberg B, Valk PJM: *HeatMapper: powerful combined visualization of gene expression profile correlations, genotypes, phenotypes and sample characteristics*. BMC Bioinformatics. 2006, 7:337.
40. Wickham H, Francois R, RStudio: *dplyr: A Grammar of Data Manipulation*. 2015. R package version 0.4.3.
41. Wu HM, Tien YJ, Chen CH: *GAP: A Graphical Environment for Matrix Visualization and Cluster Analysis*. Computational Statistics and Data Analysis. 2010, 54: 767–778.
42. Young MD, Wakefield MJ, Smyth GK, Oshlack A: *Gene ontology analysis for RNA-seq: accounting for selection bias*. Genome Biology. 2010, 11: pp. R14.

Figures

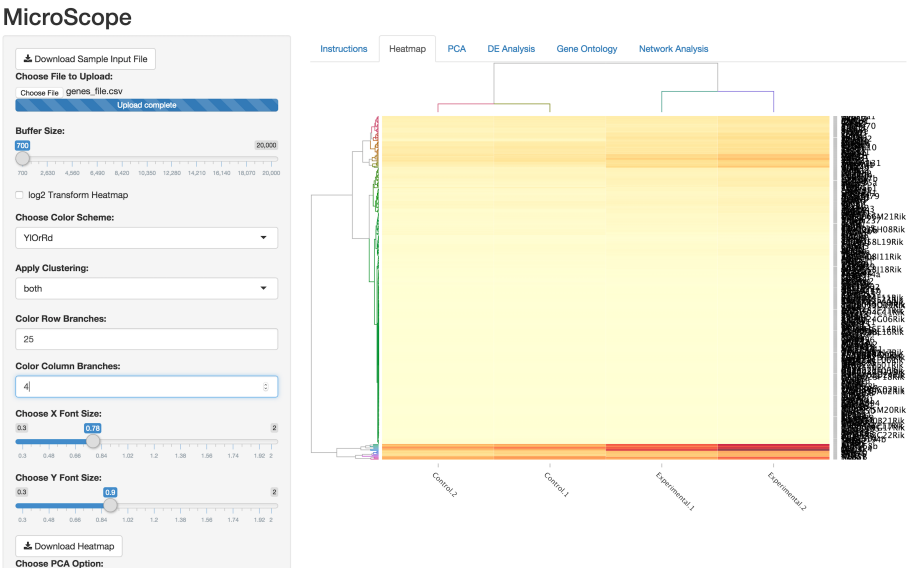

Figure 1 MicroScope user interface. MicroScope UI showcasing the magnification feature as well as features such as: sample file download, file input, buffer size, log_2 data transformation, multiple heatmap color schemes, hierarchical clustering, row/column dendrogram branch coloring, row/column font size, and heatmap download button. Additional UI features are sequentially unlocked as the user progresses through the MicroScope software suite.

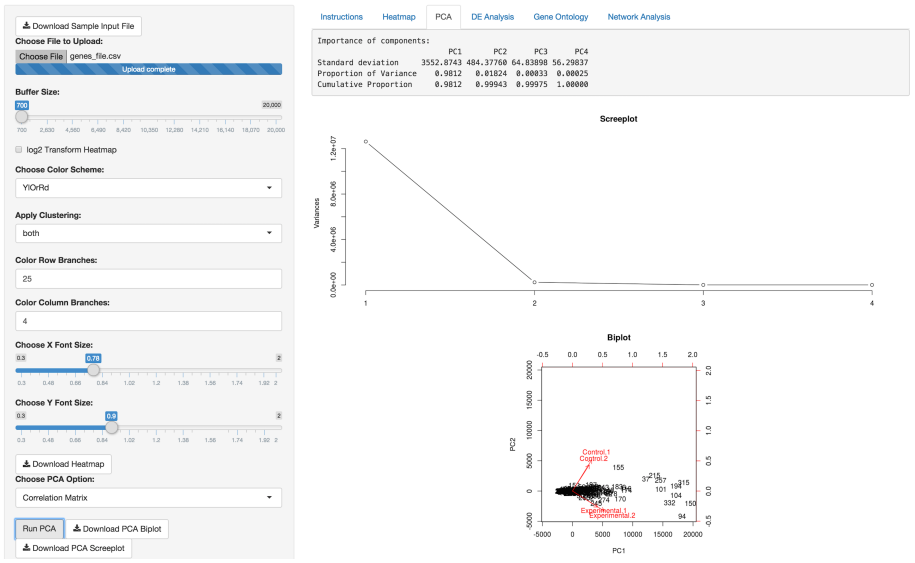

Figure 2 Principal component analysis. Tabulated summary table of importance of principal components, as well as biplot and screeplot graphics visualizations, are produced to investigate variation and patterns in gene expression.

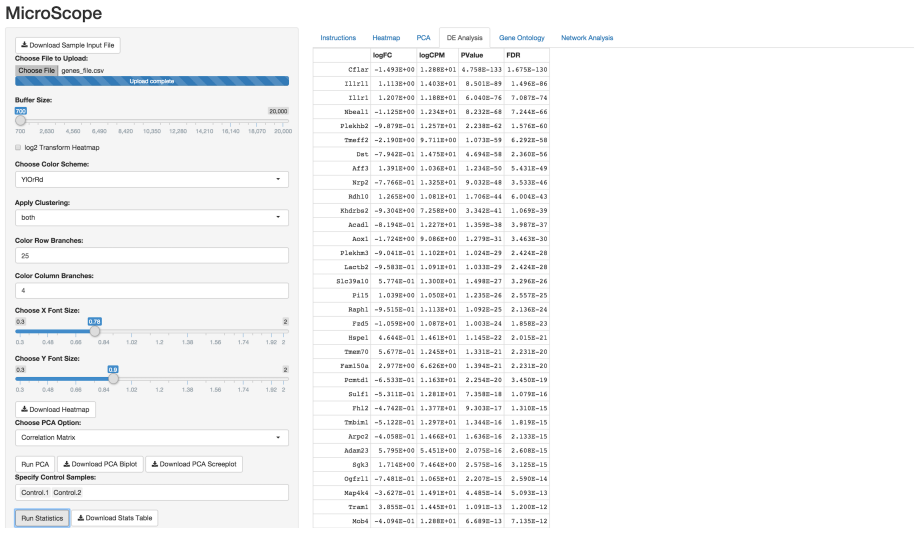

Figure 3 Differential expression analysis tabulated results. Fischer's exact test and the Benjamini-Hochberg correction are used to supply the user with information about the nominal p-value, false discovery rate, fold change, and counts per million calculations for differentially expressed genes in the heatmap.

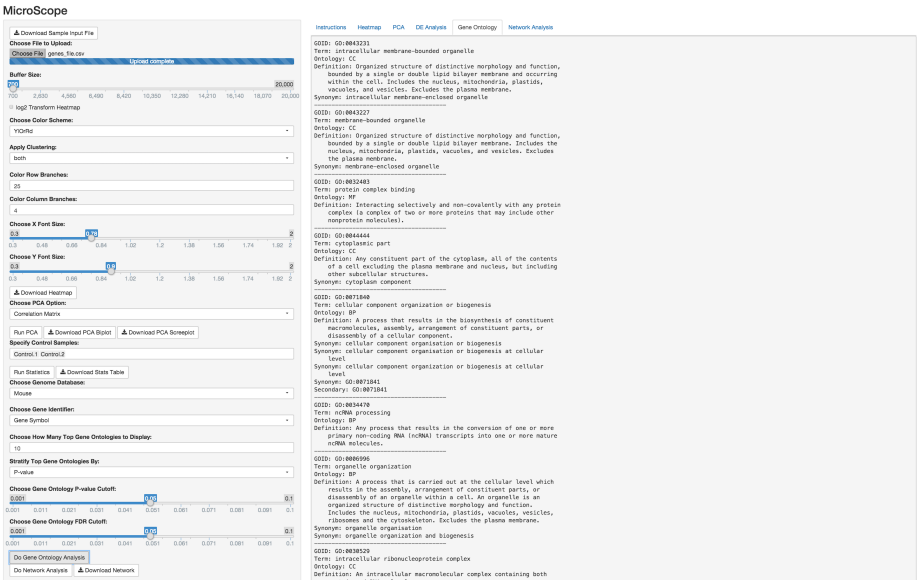

Figure 4 Gene ontology analysis tabulated results. Top gene ontology categories are automatically calculated and returned as a ranked list in the UI.

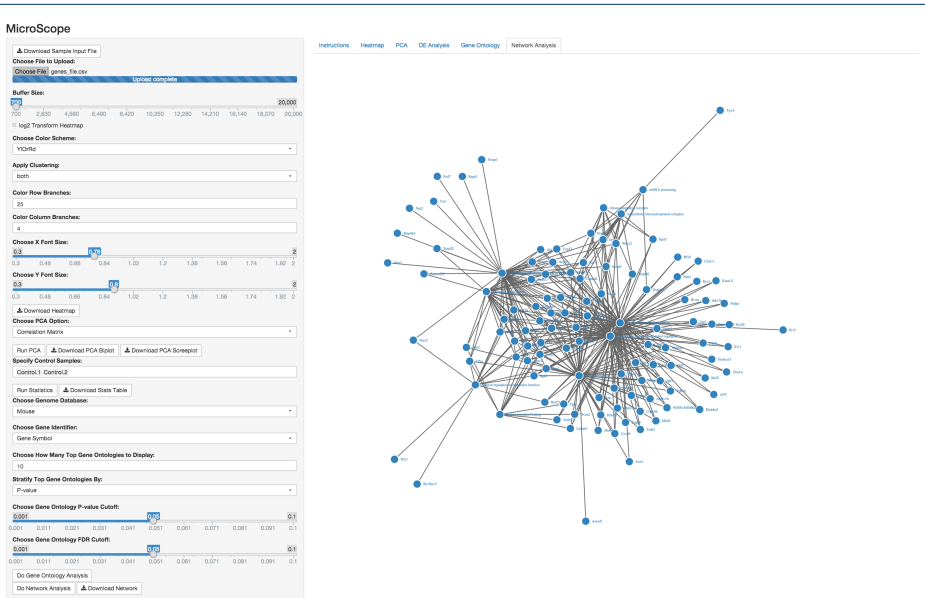

Figure 5 Network graphics visualizations of top gene ontology categories. Differentially expressed genes belonging to the respective gene ontology categories are automatically displayed during the network analysis of the data. Networks are zoomable and dynamically interactive, allowing the user to manually drag nodes across the screen to explore gene_name–gene_ontology interconnectivity and network architecture.

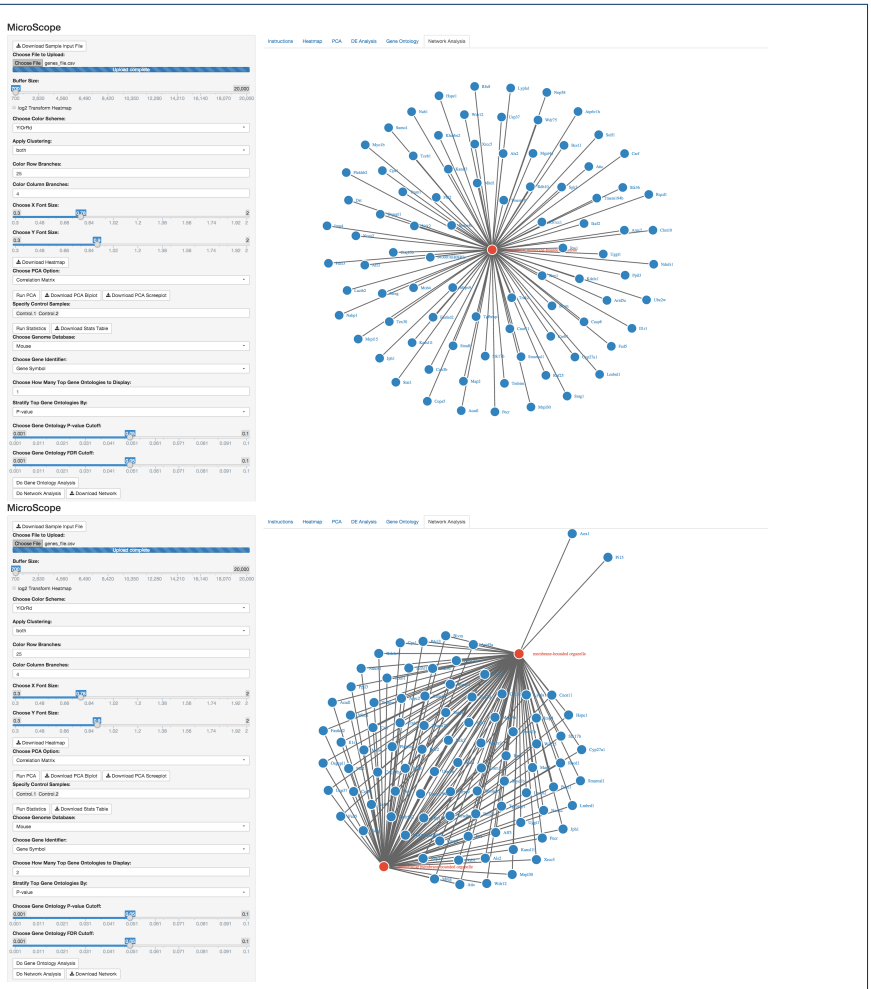

Figure 6 Network visualizations of first ranked gene ontology vs. top-two ranked gene ontologies. Comparison of dynamically interactive network graphics at various user-specified gene ontology settings (e.g., 'Choose How Many Top Gene Ontologies to Display' button in the UI). Clearly, the GO category "membrane-bounded organelle" contains two unique genes, while the rest are (perhaps unsurprisingly) shared in common with the GO category "intracellular membrane-bounded organelle".

Ethics
This study does not involve humans, human data or animals.

Abbreviations used
FDR: false discovery rate
GO: gene ontology
UI: user interface
PCA: principal component analysis
DE: differential expression

Availability of Data and Materials
Not applicable to this study.

Figures as additional files
Figures have been uploaded as additional files. Standard BioMed Central bmc_article LaTeX template has been used for production of figure captions.