Bioinformatics, 35(11), 2019, 1966–1967
doi: 10.1093/bioinformatics/bty890
Advance Access Publication Date: 22 October 2018
Applications Note

Gene expression

TissueEnrich: Tissue-specific gene enrichment analysis

Ashish Jain[1,2] **and Geetu Tuteja**[1,2,*]

[1]Bioinformatics and Computational Biology and [2]Genetics, Development and Cell Biology, Iowa State University, Ames, IA 50011, USA

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on April 16, 2018; revised on October 11, 2018; editorial decision on October 17, 2018; accepted on October 19, 2018

Abstract

Summary: RNA-Seq data analysis results in lists of genes that may have a similar function, based on differential gene expression analysis or co-expression network analysis. While tools have been developed to identify biological processes that are enriched in the genes sets, there remains a need for tools that identify enrichment of tissue-specific genes. Therefore, we developed TissueEnrich, a tool that calculates tissue-specific gene enrichment in an input gene set. We demonstrated that TissueEnrich can assign tissue identities to single cell clusters and differentiated embryonic stem cells.

Availability and implementation: The TissueEnrich web application is freely available at http://tissueenrich.gdcb.iastate.edu/. The R package is available through Bioconductor at https://bioconductor.org/packages/TissueEnrich. Both the web application and R package are for non-profit academic use under the MIT license.

Contact: geetu@iastate.edu

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

1 Introduction

The development of RNA-Seq technology has enabled large-scale comparison of gene expression in a multitude of developmental stages, cell-types and conditions. RNA-Seq data analysis identifies genes that likely have a shared function, either through differential gene expression analysis or co-expression network analysis. Gene ontology (GO) enrichment analysis is widely used to assign function to the gene sets and to gain insights into the biological processes they are involved in. While GO analysis identifies enriched processes in sets of genes, it does not determine enrichment of tissue-specific genes. Understanding which groups of genes are tissue-specific is valuable, as tissue-specific genes are more likely to be associated with human disease (Winter *et al.*, 2004). To this end, tools have been developed that include tissue enrichment or tissue-specific enrichment calculations (http://genetics.wustl.edu/jdlab/tsea/; Komljenovic *et al.*, 2016; Watanabe *et al.*, 2017; Xu *et al.*, 2014). However, these tools are limited in the flexibility of parameters for defining tissue-specificity, or the ability to allow users to define their own tissue-specific gene sets to use for enrichment analysis (see Supplementary Material for more details). Therefore, we developed ‘TissueEnrich’, a tool to carry out tissue-specific gene enrichment. TissueEnrich is available as an interactive web application, allowing the user to visualize tissue-specific gene enrichment and visualize expression of genes from the input set that were determined to be tissue-specific. We also developed an R package that further allows users to define tissue-specific genes in custom datasets, which they can then use for tissue-specific enrichment analysis.

2 Materials and methods

2.1 RNA-Seq datasets

To calculate tissue-specificity of genes, we used RNA-Seq data from GTEx (Ardlie *et al.*, 2015), HPA (Uhlén *et al.*, 2015) (https://v18.proteinatlas.org/about/download) and the mouse ENCODE project (Shen *et al.*, 2012). We used data from tissues with >1 biological replicate. Processed GTEx data were downloaded from the Expression Atlas (E-MTAB-5214; Petryszak *et al.*, 2016), and the samples from sub-tissues were grouped and considered to be biological replicates. The one-to-one orthologous genes between human and mouse were downloaded from Ensemble (Version 91; Zerbino *et al.*, 2018). TissueEnrich only uses protein-coding genes for analysis. Tissue-specific metrics and group definitions are provided in the Supplementary Material.

2.2 Tissue-specific gene enrichment

We used the hypergeometric test to calculate the enrichment of tissue-specific genes in the input gene set (see Supplementary Material).

3 Results

3.1 Web application

We used R Shiny (Winston *et al.*, 2018) to develop a user-friendly web application to calculate tissue-specific gene enrichment in a user-provided gene set. From the 'Tissue Enrichment' tab, the user can select the organism their data is from and the RNA-Seq dataset to use for tissue-specificity information. The output is an interactive bar chart, depicting the significance of tissue-specific gene enrichment, plotted as $-\text{Log}_{10}(P\text{-adjusted})$ on the y-axis, across each tissue. The user can view the expression values of the genes that were part of a tissue-specific group in an interactive heatmap by clicking the corresponding bar for that tissue in the bar chart. Furthermore, users can search for the expression values of individual genes along with their tissue-specific groups under the 'Tissue Specific Genes' tab. The 'Help' tab provides detailed usage instructions.

3.2 R package

The TissueEnrich R package contains similar functions to those in the web application. The package includes the 'teGeneRetrieval' function, which can be used to define tissue-specific genes in any given dataset. In this function, users can adjust thresholds for calculating tissue-specificity. The resulting tissue-specific genes can be used to carry out tissue-specific gene enrichment, using the 'teEnrichmentCustom' function.

3.3 Case study 1: Defining tissue-specificity of genes expressed in differentiated embryonic stem cells

We used two RNA-Seq datasets generated from differentiated embryonic stem cells (ESCs) to test TissueEnrich. The first dataset is from ESCs differentiated into cardiomyocytes (Szabo *et al.*, 2015). We ran TissueEnrich on the 1000 most highly expressed genes, processed as previously described in Jain et al. (2017) and found enrichment for heart-specific genes using both the HPA and mouse ENCODE datasets (Supplementary Fig. S1 and Supplementary Data S1). These results validate the robustness of TissueEnrich and highlight heart-specific genes that may be interesting for follow up experiments. The second RNA-Seq dataset is from ESCs differentiated into trophoblast-like cells (Yabe *et al.*, 2016). While Yabe *et al.* concluded that the differentiated cells are of trophoblast origin, the origin has been debated (Roberts et al., 2014). We previously used an approach similar to what we have now packaged into TissueEnrich to determine that the 1000 most highly expressed genes in these cells have strong enrichment for placenta-specific genes defined through the HPA (Jain *et al.*, 2017) (Supplementary Data S2). Here, we used TissueEnrich for the same genes and found that they also show enrichment for placenta-specific genes defined through the mouse ENCODE project (Supplementary Fig. S2), indicating trophoblast-like cells may have a conserved function in mouse.

3.4 Case study 2: Annotation of cell clusters from single-cell RNA-Seq data

Next, we used single-cell RNA-Seq data from mouse embryos at e6.5 (Scialdone *et al.*, 2016). Scialdone *et al.* clustered the cells using gene expression data and assigned tissue identities to the cell clusters based on marker genes that were upregulated in each cluster compared to all other clusters. Here, we ran TissueEnrich for each set of genes that were specifically up-regulated in cell clusters associated with e6.5 to determine if more information could be obtained about the cell identities. We found that cluster 1 showed strongest tissue-specific enrichment for intestine, liver and kidney (tissues derived from visceral endoderm); cluster 2 for placenta (tissue derived from extraembryonic ectoderm); and cluster 3 for brain and olfactory bulb tissue (Supplementary Fig. S3 and Supplementary Data S3). These results are in agreement with the tissue identities assigned in Scialdone *et al.*, and provide additional detail on tissue-specific gene expression that may further help understand cell specification during early development.

4 Conclusions

We developed the TissueEnrich web application and R package, and demonstrated that it provides an unbiased way to carry out tissue-specific gene enrichment in mouse and human RNA-Seq data.

Acknowledgements

The authors thank Tuteja lab members for testing TissueEnrich, Majd Abdulghani and Dr. Haninder Kaur for critically reviewing the manuscript, Levi Baber (LAS research IT director at ISU) for setting up the web application server, Anilisa Biswas and Kokulapalan Wimalanathan for fruitful discussions, Dr. Martin Morgan (Bioconductor) for help with code optimization and Per Oksvold HPA for providing pseudo-code for calculating tissue-specific genes.

Funding

This work was supported in part by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health [R00HD079545 to GT]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agency.

Conflict of Interest: none declared.

References

Ardlie,K.G. *et al.* (2015) The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, **348**, 648–660.

Jain,A. *et al.* (2017) Deciphering transcriptional regulation in human embryonic stem cells specified towards a trophoblast fate. *Sci. Rep.*, **7**, 17257.

Komljenovic,A. *et al.* (2016) BgeeDB, an R package for retrieval of curated expression datasets and for gene list expression localization enrichment tests. *F1000Research*, **5**, 2748.

Petryszak,R. *et al.* (2016) Expression atlas update—an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Res.*, **44**, D746–D752.

Roberts, R.M. *et al.* (2014) Differentiation of trophoblast cells from human embryonic stem cells: to be or not to be? *Reproduction*, **147**, D1–D12.

Scialdone,A. *et al.* (2016) Resolving early mesoderm diversification through single-cell expression profiling. *Nature*, **535**, 289–293.

Shen,Y. *et al.* (2012) A map of the cis-regulatory sequences in the mouse genome. *Nature*, **488**, 116–120.

Szabo,L. *et al.* (2015) Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development. *Genome Biol.*, **16**, 126.

Uhlén,M. *et al.* (2015) Tissue-based map of the human proteome. *Science*, **347**, 1260419–1260419.

Winston,C. *et al.* (2018) *shiny: Web Application Framework for R. R package version 1.1.0.* https://CRAN.R-project.org/package=shiny.

Watanabe,K. *et al.* (2017) Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.*, **8**, 1826.

Winter,E.E. (2003) Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. *Genome Res.*, **14**, 54–61.

Xu,X. *et al.* (2014) Cell type-specific expression analysis to identify putative cellular mechanisms for neurogenetic disorders. *J. Neurosci.*, **34**, 1420–1431.

Yabe,S. *et al.* (2016) Comparison of syncytiotrophoblast generated from human embryonic stem cells and from term placentas. *Proc. Natl. Acad. Sci. USA*, 1601630113.

Zerbino,D.R. *et al.* (2018) Ensembl 2018. *Nucleic Acids Res.*, **46**, D754–D761.