tspex: a tissue-specificity calculator for gene expression data

Antonio P. Camargo
University of Campinas https://orcid.org/0000-0003-3913-2484
Adrielle A. Vasconcelos
University of Campinas https://orcid.org/0000-0001-8145-4669
Mateus B. Fiamenghi
University of Campinas https://orcid.org/0000-0003-4535-8594
Gonçalo A. G. Pereira
University of Campinas https://orcid.org/0000-0003-4140-3482
Marcelo F. Carazzolle (mcarazzo@lge.ibi.unicamp.br)
University of Campinas https://orcid.org/0000-0002-5474-2830

Short Report

Keywords: RNA-Seq, transcription, quantification, software, bioinformatics, program

Posted Date: August 4th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-51998/v1

Abstract

When comparing gene expression data of different tissues it is often interesting to identify tissue-specific genes or transcripts. Even though there are several metrics to measure tissue-specificity, a user-friendly tool that facilitates this analysis is not available yet. We present tspex, a software that allows easy computation of a comprehensive set of different tissue-specificity metrics from gene expression data. tspex can be used through a web interface, command-line or the Python API. Its package version also provides visualization functions that facilitate inspection of results. The documentation and the source code of tspex are available at https://apcamargo.github.io/tspex/ and the web application can be accessed at https://tspex.lge.ibi.unicamp.br/

Introduction

High-throughput sequencing technologies have allowed the quantification of gene expres- sion of several tissues from different species, generating huge amounts of data that can be analyzed for discovering novel expression profiles. Regarding expression patterns across tissues, genes can be positioned within a continuous scale that goes from housekeeping genes to tissue-specific genes. Housekeeping genes are responsible for critical cell functions and are ubiquitously expressed[1], while tissue-specific genes are expressed in a single or a small subset of tissues, suggesting specialized functions.

Detection of tissue-specific genes can be important, for instance, for gene discovery, evolutionary comparisons[2,3], drug target identification[4], association with diseases[5], and cancer studies[6,7]. In this context, there are several initiatives to sequence various tissues of an organism, such as the Genotype-Tissue Expression (GTEx) project[8], which offers expression data of 53 human tissues sampled from nearly 1,000 individuals, providing a foundation for the identification of new tissue-specific genes.

Although methods to quantify gene tissue-specificity have been extensively used in the literature, there is no available tool that allows easy measurement of tissue-specificity from gene expression data, forcing users to develop their own one-time use solutions. Herein, we present tspex, a tool that allows easy computation of twelve different tissue-specificity metrics from gene expression data. It provides visualization functions that facilitate exploration of results and can be used through different interfaces, including a web version. Additionally, tspex is also useful to measure expression specificity among distinct biological conditions, such as developmental stages, time points, genetic varieties, etc.

Implementation

tspex is implemented as a Python package and it can be used locally through a Python Application Programming Interface (API), command-line interface or web version. Local installation of tspex is as easy as calling it with pip or conda and requires few dependencies. Refer to the tspex GitHub repository for the most up-to-date source code, dependency details and instructions. An open source web interface

(Figure 1A), built with Flask and deployed using Docker containers, is also available at https://tspex.lge.ibi.unicamp. br/.

tspex provides twelve distinct tissue-specificity metrics, which differ in their assump- tions, scale and properties. Broadly, these metrics can be divided into two groups[9]: (1) general scoring metrics, that summarize in a single value how tissue-specific or ubiquitous is a gene across all tissues and (2) individualized scoring metrics that quantify how specific is the expression of each gene to each tissue.

The general scoring metrics provided by tspex are: Counts[3], Tau[10], Gini coefficient[11], Simpson index[12], Shannon entropy specificity[13], ROKU specificity[14], Specificity measure dispersion (SPM DPM)[15], and Jensen-Shannon specificity dispersion (JSS DPM)[16]. As for individualized scoring metrics, tspex includes: Tissue-specificity index (TSI)[17], Z- score[18], Specificity measure (SPM)[19], and Jensen-Shannon specificity (JSS)[16]. Each metric provides values that range within different scales, thus tspex includes an option to transform tissue-specificity values so that they fall within 0 (ubiquitous expression) and 1 (tissue-specific expression). The equations for all provided metrics as well as their transformations can be found in the Supplementary Material or at https://apcamargo. github.io/tspex/metrics/.

As input, tspex requires an expression matrix (TSV, CSV or Excel formats) in any appropriate unit, such as TPM, FPKM or CPM. Optionally, tspex allows the expression values to be log-transformed before computation of tissue-specificity, which reduces the dependency between expression variance and expression level, improving the reliability of tissue-specificity measurements[9]. Internally, expression data and the tissue-specificity values are stored in a Python object and can be easily accessed for further investigation through the Python API.

Finally, the tspex package provides built-in functions for data visualization. Specifically, the user can plot histograms of tissue-specificity values (Figure 1B) and heatmaps of the expression of genes whose tissue-specificity is above a chosen value (Figure 1C). These visualizations allow quick inspection of the results and can be helpful for deciding threshold values.

Results

To address the lack of tools for calculating tissue-specificity metrics from gene expression data, we developed tspex, a tool for easy computation of twelve different tissue-specificity scoring metrics. It is available to be used through three different interfaces: Python API, command-line, and a web version. Detailed tutorials on the usage of tspex interfaces can be found in the tspex documentation.

On the tspex website, calculation is easily done by simply uploading the input file, choosing the metric in the drop-down menu, and submitting it for calculation (Figure 1a). For guidance in choosing a tissue-specific metric for your gene expression data refer to this benchmark[9]. When it is done, the results can be downloaded and/or viewed it on the results page (only for data with up to 5000 genes). This can also be achieved by using tspex locally as a command line tool or in a Python API. Running tspex on a command

line doesn't require prior knowledge in Python and can be useful when adding it into an automated analysis pipelines or to run on multiple files at once. The Python API offers the advantages of tight integration with popular data analysis libraries, such as NumPy[20], SciPy[21], and pandas[22], as well as visualization functions to create publication quality figures that aid the inspection of results.

Here, we demonstrate the features of tspex by running the tool in the Python API with real gene expression data from the Genotype-Tissue Expression (GTEx) project[8], which provides a large catalogue of gene expression across 54 human tissues. To showcase an example analysis, we used gene expression data from only five tissues: Bladder, Liver, Lung, Pancreas and Stomach. After removing genes that are not expressed in any of these tissues, expression values in transcripts per million (TPM) of 31872 genes were used as input for tspex. In order to obtain tissue-specificity values for these genes in the sampled human tissues, we calculated the general scoring metric Tau, which results in a single tissue-specificity score per gene. By running the visualization functions on the TissueSpecificity object, we can have an overview of the tissue-specificity results. The histogram plotting function can be used to verify the distribution of the tissue-specificity values, which is helpful for deciding thresholds for selecting genes (Fig 1b). Whereas, the heatmap plotting function is useful to visualize the amount of genes that are specific for each tissue above a given threshold (Fig 1c). In this heatmap, it is possible to see that lung has the largest number of tissue-specific genes among the five tissues.

The user can further explore and manipulate the tissue-specificity results obtained with tspex depending on what is being investigated. For example, searching for the top most tissue-specific gene for each tissue, generating lists of tissue-specific genes above a threshold for Gene Ontology term enrichment analysis, or filtering it to look at only certain class of genes. It is relevant to note that tissue-specificity values rely and will change according to the number of tissues being analysed and how different they are. Besides its application for finding tissue-specific expression, tspex is also useful to measure expression specificity among distinct biological conditions, such as developmental stages, time points, genetic varieties, etc.

Conclusion

tspex is a software for calculating a variety of tissue-specificity metrics from gene expression data, addressing the lack of tools that perform this important task. It is entirely developed in Python to provide integration with an extensive library of data analysis packages. Finally, tspex can be used through three different interfaces, including a web version, providing solutions for different use cases.

Declarations

Availability of data and materials

The datasets used for the analyses described in this manuscript are publicly available and were obtained from the GTEx Portal and/or dbGaP accession number phs000424.vN.pN.

The equations for all provided metrics as well as their transformations can be found in the Supplementary information. Source code and documentation is freely available on GitHub.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This work was supported by The S˜ao Paulo Research Foundation (FAPESP) grants 2013/08293-7 and 2016/23218-0, FAPESP fellowships to A.P.C (2018/04240-0) and
A.A.V (2017/13015-7) and a fellowship from Coordenac¸a˜o de Aperfeic¸oamento de Pessoal de N´ıvel Superior – Brasil (CAPES) to M.B.F.

Authors' contributions

AC conceived the project, implemented source code, constructed the web application, and wrote documentation. AV contributed with the source code, writing documentation and testing the software. MF contributed writing code for the web application. AC and AV wrote the manuscript. GP and MC supervised and reviewed the manuscript. All authors read and approved the final manuscript.

References

1. Goldman, in *Encyclopedia of Genetics* 978 (Elsevier, 2001). isbn: 9780122270802. https://linkinghub.elsevier.com/retrieve/pii/B012227080000639X.
2. Kryuchkova-Mostacci, N. & Robinson-Rechavi, M. Tissue-specificity of gene ex- pression diverges slowly between orthologs, and rapidly between paralogs. *PLoS computational biology* 12, e1005274 (2016).
3. Duret, L. & Mouchiroud, D. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Molecular biology and evolution* 17, 68–070 (2000).
4. Yang, Y. *et al.* A systematic survey of human tissue-specific gene expression and splicing reveals new opportunities for therapeutic target identification and evaluation. *bioRxiv,* 311563 (2018).
5. Gamazon, E. R. *et al.* Using an atlas of gene regulation across 44 human tissues to inform complex disease-and trait-associated variation. *Nature Genetics* 50, 956 (2018).
6. Haigis, K. M., Cichowski, K. & Elledge, S. J. Tissue-specificity in cancer: The rule, not the exception. *Science* 363, 1150–1151 (2019).
7. Kim, *et al.* TissGDB: tissue-specific gene database in cancer. *Nucleic acids research* 46, D1031–D1038 (2017).
8. Lonsdale, J. *et al.* The genotype-tissue expression (GTEx) project. *Nature genetics* 45, 580 (2013).

9. Kryuchkova-Mostacci, & Robinson-Rechavi, M. A benchmark of gene expression tissue-specificity metrics. *Briefings in bioinformatics* 18, 205–214 (2017).
10. Yanai, *et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* 21, 650–659 (2004).
11. Ceriani, & Verme, P. The origins of the Gini index: extracts from Variabilit*aeMutabilita* (1912) by Corrado Gini. *The Journal of Economic Inequality* 10, 421–443 (2012).
12. Simpson, H. Measurement of diversity. *Nature* 163, 688 (1949).
13. Schug, *et al.* Promoter features related to tissue specificity as measured by Shannon entropy. *Genome biology* 6, R33 (2005).
14. Kadota, K., Ye, , Nakai, Y., Terada, T. & Shimizu, K. ROKU: a novel method for identification of tissue-specific genes. *BMC bioinformatics* 7, 294 (2006).
15. Pan, J.-B., Hu, S.-C., Wang, , Zou, Q. & Ji, Z.-L. PaGeFinder: quantitative identification of spatiotemporal pattern genes. *Bioinformatics* 28, 1544–1545 (2012).
16. Cabili, M. N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & development* 25, 1915–1927 (2011).
17. Julien, *et al.* Mechanisms and evolutionary patterns of mammalian and avian dosage compensation. *PLoS biology* 10, e1001328 (2012).
18. Vandenbon, A. & Nakai, K. Modeling tissue-specific structural patterns in human and mouse promoters. *Nucleic acids research* 38, 17–25 (2009).
19. Xiao, S.-J., Zhang, C., Zou, Q. & Ji, Z.-L. TiSGeD: a database for tissue-specific genes. *Bioinformatics* 26, 1273–1275 (2010).
20. Walt, v. d., Colbert, S. C. & Varoquaux, G. The NumPy array: a structure for efficient numerical computation. *Computing in science & engineering* 13, 22–30 (2011).
21. Virtanen, *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods* 17, 261–272 (2020).
22. Pandas development team, *pandas-dev/pandas: Pandas* version latest. Feb. 2020. https://doi.org/10.5281/zenodo.3509134.

Figures

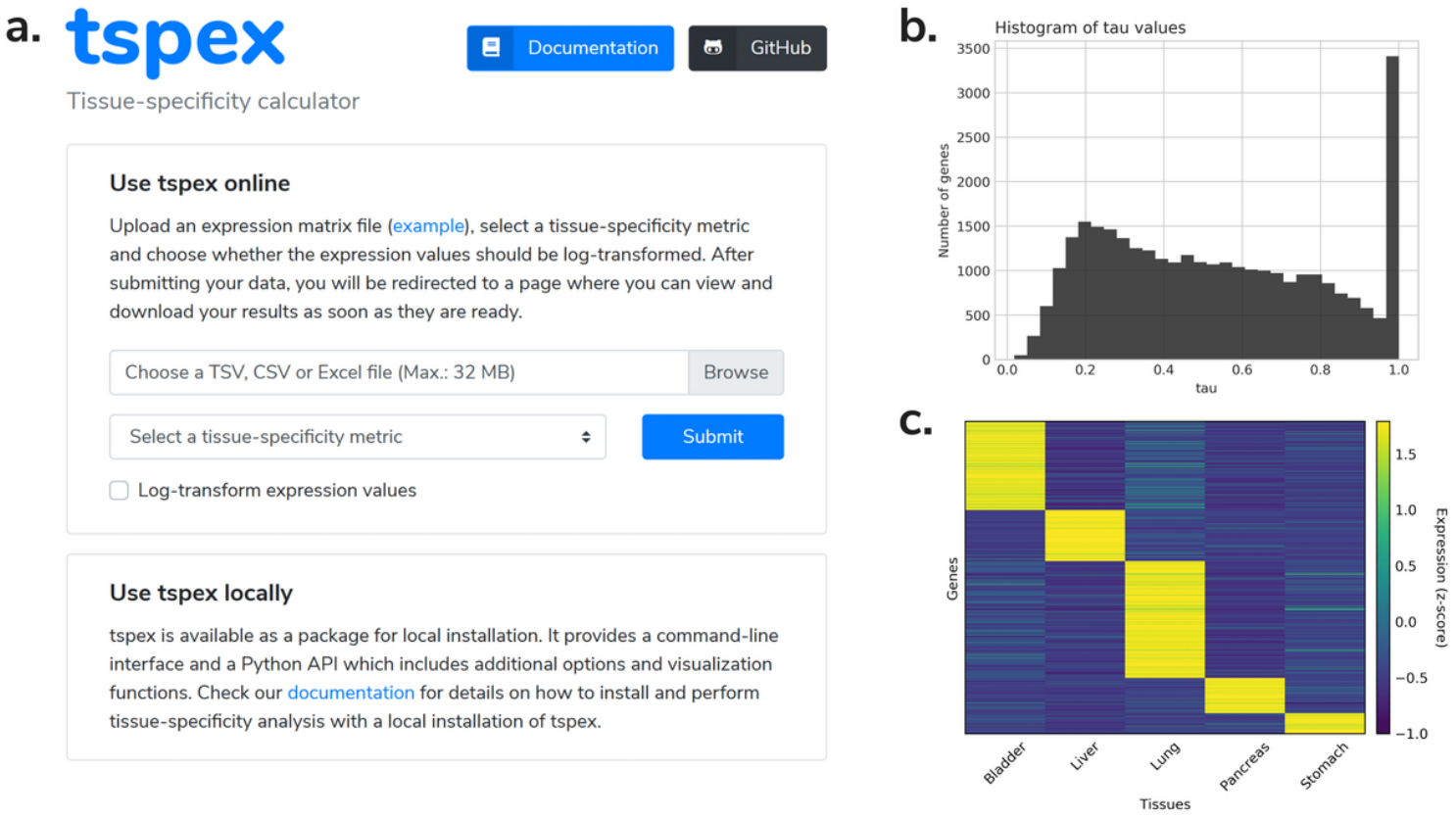

Figure 1

(a) tspex web interface home page. (b) Histogram and (c) heatmap created with the visualization functions available in the Python API.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- tspexformulas.pdf