Bioinformatics, 34(14), 2018, 2510–2512
doi: 10.1093/bioinformatics/bty122
Advance Access Publication Date: 1 March 2018
Applications Note

Gene expression

anexVis: visual analytics framework for analysis of RNA expression

Diem-Trang Tran[1,2], Tian Zhang[2], Ryan Stutsman[2], Matthew Might[3], Umesh R. Desai[4] and Balagurunathan Kuberan[1,5,*]

[1]Department of Medicinal Chemistry and [2]School of Computing, University of Utah, Salt Lake City, UT 84112, USA, [3]Hugh Kaul Personalized Medicine Institute, University of Alabama at Birmingham, Birmingham, AL 35294, USA, [4]Department of Medicinal Chemistry, Virginia Commonwealth University, Richmond, VA 23298, USA and [5]Department of Biology, University of Utah, Salt Lake City, UT 84112, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on October 13, 2017; revised on December 13, 2017; editorial decision on February 27, 2018; accepted on February 28, 2018

Abstract

Summary: Although RNA expression data are accumulating at a remarkable speed, gaining insights from them still requires laborious analyses, which hinder many biological and biomedical researchers. This report introduces a visual analytics framework that applies several well-known visualization techniques to leverage understanding of an RNA expression dataset. Our analyses on glycosaminoglycan-related genes have demonstrated the broad application of this tool, anexVis (analysis of RNA expression), to advance the understanding of tissue-specific glycosaminoglycan regulation and functions, and potentially other biological pathways.

Availability and implementation: The application is accessible at https://anexvis.chpc.utah.edu/, source codes deposited on GitHub.

Contact: kuby.balagurunathan@utah.edu

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

1 Introduction

Among the ‘omics’ techniques that have advanced extensively over the last decade, RNA-sequencing provide both comprehensive and direct view into functional states of cells and tissues. Transcriptomic profiles from RNA-seq experiments organized or generated a new by initiative such as The Cancer Genome Atlas (TCGA), Encyclopedia of DNA Element (ENCODE) and genotype-tissue expression (GTEx) have equipped the research community with a large enough dataset for more rigorous analyses. The challenge has thus gradually shifted from consolidating to understanding these data. The appearance of recent web-based RNA-seq analysis tools reflects an increasing need and technical capability for visual analytics. For example, the conventional heatmap representation of *genes* x *samples* expression matrix is improved in *Shinyheatmap* (Khomtchouk *et al.*, 2017) and *NG-CHM* (Broom *et al.*, 2017), differential expression analysis is enhanced with interactive plots in *DEIVA* (Harshbarger *et al.*, 2017), and the processing and analyzing workflow are ported into graphical user interface in *QuickRNASeq* (Zhao *et al.*, 2016) and *ASAP* (Gardeux *et al.*, 2017). Although these tools have significantly enhanced the representation of various data types in an RNA-seq workflow, the integrated visual analysis of multiple data types and views remains inadequate. Such integration is critical for further insights, such as how gene expression profiles account for different phenotypes, or whether a co-expression measure could capture meaningful co-expression relations. *anexVis* has been designed to facilitate such analyses. In particular, the gene-centric parallel coordinate plot (PCP) not only overlays multi-gene expression profiles but is also synchronized with sample metadata and phenotypes to reveal potential gene expression-phenotype relations. Similarly, the adjacency matrix not only presents multi-gene co-expression profiles but is also coupled with scatter diagrams to put forth the raw data underlying each correlation value.

We provided three distinct examples to demonstrate the utility of this framework in the study of glycosaminoglycan (GAG). The first two closely follow the workflows suggested by the framework to explore (1) the gene-based tissue-specific signature of proteoglycans and (2) the correlation-based tissue signature of heparan sulfate biosynthesis. The third example illustrates an advanced use of the available functionalities to (3) understand a human genetic disease related to defective GAG catabolic pathways.

2 Materials and methods

The framework was implemented as a web application to maximize accessibility to biological researchers (Supplementary Fig. S1). In this set-up, users only need a working internet browser, ideally Chrome, to use the application. During an interactive session, R function calls are sent to an OpenCPU server (Ooms, 2014) as HTTP requests. These functions are then invoked on this R-powered server to perform data queries or computations and return JSON-formatted responses. A key engineering challenge in anexVis is dealing with a massive dataset while maintaining access rates fast enough for interactive use. To achieve this, we used redis as an in-memory store on the server. On the client side, visualization components and interactivity are implemented in Javascript, using the visualization library d3js as the primary groundwork.

2.1 Datasets

The RNA-seq data built-in with this framework were generated by the GTEx project (Melé *et al.*, 2015). Three versions of this dataset resulting from three different processing workflows (Vivian *et al.*, 2017) are included and can be selected to analyze separately. The metadata including sample-based annotations and selected subject phenotypes were downloaded from the database of genotypes and phenotypes (dbGaP), under phs000424.v6.p1. The selected subject phenotypes were de-identified and included with the approval of NIH.

2.2 Visualization components and interactive scheme

The application aims to support the exploration of two types of patterns in RNA expression data: the *gene-based signature* composed of the expression of individual genes in a gene set and the *co-expression signature* composed of co-expression between gene pairs. Each signature depicts a characteristic aspect of a tissue type.

Gene-based signature is visualized by a PCP, a well-known technique for high-dimensional data. Considering that an expression profile of *n* genes in a biological sample is an *n*-dimension data point, the PCP is created by plotting the expression level of each gene on an axis, resulting in each polyline representing a biological sample (Supplementary Fig. S2B). Based on this view, the pattern of polyline distribution across *n* axes defines the expression pattern of a group of samples in the *n*-gene space. Two types of features can be identified from the PCP (i) the segregation of samples into clusters along with the segregating gene and (ii) the correlation of genes on adjacent axes by the presence of parallel (positive correlation) or cross-over (negative correlation) lines between them (Wegman, 1990). Since such correlation is only visible for adjacent axes, manual axis re-ordering is supported to allow users to interrogate any pair of interest. The second component of the view is a data table enlisting all the samples included in the plots, with the pre-selected metadata, including organ system, tissue type, ontology term, age, gender and race. The last component is a set of bar charts drawn on categorical dimensions of the metadata, summarizing the distribution of the selected samples in these partitions. Sample selection is synchronized across all three elements.

The correlation-based signature is represented by a square matrix in which the entry (*i*, *j*) is colored by the co-expression measure of the two genes *i* and *j* (Supplementary Fig. S3B). In the current implementation, co-expression is measured by Pearson correlation coefficient. Although the choice of co-expression measure is beyond the scope of this article, the framework can be easily extended to include other choices of co-expression measures. By clicking on a square of interest, user can select the gene pair to generate a scatter diagram visualizing the relation between them. This scatter diagram supports interactive coloring based on various attributes of a data point: organ system, ontology term, gender and race. Such feature is especially useful when there are potential structures in the data, for example, when studying a highly heterogeneous organ such as brain (Supplementary Fig. S5).

3 Results

We used *anexVis* to analyze the tissue-specific expression patterns of GAG genes (Example 1–3, see Supplementary Materials). Previous observations can now be summarized and a number of undiscovered patterns could be revealed at a glance. Although the tissue-specific expression have been observed for many of these genes, patterns and relationships have been difficult to identify. With our tool, a systematic survey is now available at the disposal of every researcher. In addition, the design of *anexVis* has helped to make technical limitations more visible, among which are the false negative readings of lowly expressed genes (Supplementary Fig. S4), and the problematic correlation measures in a heterogeneous group of samples (Supplementary Fig. S5). Such insights undoubtedly probe more cautious interpretation of the datasets and greater efforts to address the technical limitations.

The number of genes included in the web service is limited by the infrastructure available in our hands, and the challenge of fitting this large dataset on an in-memory data store. Since it is open-source, users can implement a custom version for their own dataset with their choice of genes. Future versions of the framework aim at including all detected genes in the human genome. A number of enhancements are also planned, including a better automatic arrangement of axes in PCP to optimize pattern recognition, additional options of co-expression measures, and more user control over plotting parameters.

In conclusion, *anexVis* facilitates the exploration of RNA expression data of multiple genes across multiple tissue types by integrating different data types and views and presenting them in a highly interactive manner. Despite our focus on GAGs, the application is fully customizable to include other genes and thus, would be useful in many other biological fields.

Acknowledgements

We thank the CloudLab and Center for High Performance Computing at the University of Utah for the computing resources in developing, testing and provisioning the application, Jie Shi Chua for proofreading, and the anonymous reviewer for providing valuable suggestions to improve the manuscript.

Funding

This work was supported by the National Heart, Lung, and Blood Institute (NHLBI) – NIH sponsored Programs of Excellence in Glycosciences [Grant number HL107152 to B.K.].

Conflict of Interest: none declared.

References

Broom,B.M. *et al*. (2017) A galaxy implementation of next-generation clustered heatmaps for interactive exploration of molecular profiling data. *Cancer Res*., **77**, e23–e26.

Gardeux,V. *et al*. (2017) ASAP: a web-based platform for the analysis and interactive visualization of single-cell RNA-seq data. *Bioinformatics*, **33**, 3123–3125.

Harshbarger,J. *et al*. (2017) DEIVA: a web application for interactive visual analysis of differential gene expression profiles. *BMC Genomics*, **18**, 47.

Khomtchouk,B.B. *et al*. (2017) Shinyheatmap: ultra fast low memory heatmap web interface for big data genomics. *PLos One*, **12**, e0176334.

Melé,M. *et al*. (2015) The human transcriptome across tissues and individuals. *Science*, **348**, 660–665.

Ooms,J. (2014). The OpenCPU System: towards a universal interface for scientific computing through separation of concerns. *arXiv: 1406.4806 [stat.CO]*.

Vivian,J. *et al*. (2017). Toil enables reproducible, open source, big biomedical data analyses. *Nat. Biotechnol*., **35**, 314–316.

Wegman,E.J. (1990) Hyperdimensional data analysis using parallel coordinates. *J. Am. Stat. Assoc*., **85**, 664–675.

Zhao,S. *et al*. (2016) QuickRNASeq lifts large-scale RNA-seq data analyses to the next level of automation and interactive visualization. *BMC Genomics*, **17**, 39.