Identification of conditions associated with gene expression variation using GENEVA

Zicheng Hu (✉ hzc363@gmail.com)

UCSF

Navchetan Kaur

UCSF

Short Report

Keywords:

Posted Date: February 7th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1284968/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Motivation: While the normal expression patterns of human genes have been well characterized, how the expression varies under different conditions, such as diseases, drug treatments, and genetic perturbations, is not fully understood. Information related to gene expression variation is critical for understanding the regulation of human genes and identifying approaches for targeting disease-related genes.

Results: We previously developed a method named GENEVA to mine a massive amount of RNA-sequencing data to identify previously studied conditions associated with the expression variation of ACE2. Here, we provide an introduction to GENEVA and its web application. We demonstrate that GENEVA can be used to study the expression of previously understudied genes.

Availability: GENEVA is implemented as a web application and is available at <https://genevatool.org/>

Introduction

The expression of human genes in healthy human tissues has been well characterized by existing resources such as the GTEx, the Human Protein Atlas, and the Human Cell Atlas (Lonsdale et al., 2013; Regev et al., 2017; Uhlén et al., 2015). However, how the expression of these genes varies with different conditions, such as diseases, drug treatments, and genetic mutations, is not fully characterized. Currently, more than 200,000 human RNA-seq samples are publicly available, providing an unprecedented opportunity for us to examine the expression of human genes under a variety of conditions. In addition, data harmonization efforts such as ARCHS4 have uniformly preprocessed the RNA-seq data, making them readily available for analysis (Lachmann et al., 2018). However, fully automated analysis of these datasets is hindered by the non-standardized metadata in the RNA-seq dataset.

We previously developed a computational tool named GENEVA (GENe Expression Variance Analysis), a semi-automated framework for exploring public RNA-seq datasets without manually curating the metadata. Using GENEVA, we identified multiple conditions that modulate the expression of ACE2, the cell entry receptor of the SARS-CoV-2 virus (Kaur et al., 2020). In this article, we demonstrate that GENEVA can be used to study the expression of any human gene. We provide an introduction to the GENEVA web app and show that GENEVA can be used to study previously understudied genes.

Results

GENEVA first concatenates all the metadata fields in an RNA-seq dataset and calculates the pairwise Levenshtein distance between the metadata of the samples. The distance between the metadata is used to approximate the similarity of the experimental conditions of the samples. Next, GENEVA embedded the metadata into a two-dimensional space using multidimensional scaling (MDS). The embedding is used in the GENEVA web app to visualize the experimental groups in a dataset (Figure 1). GENEVA tests if the expression of a given gene is associated with the experimental groups using a regression model

(expression ~ embedding dimension 1 = embedding dimension 2). The variance explained by the metadata embedding is used to measure the association between the expression and the experimental conditions.

In many cases, a single RNAseq dataset contains samples from different tissue or cell types. GENEVA controls for the overall heterogeneity of samples in the study to prioritize datasets in which the gene of interest is modulated by experimental conditions rather than a result of cell-type differences. Specifically, the GENEVA score is defined as $VARE / VARm$. VARE represents the variance of the gene of interest explained by the metadata embedding. VARm represents the average variance of all genes (VARm) in the dataset. GENEVA web application displays the RNA-seq datasets from high to low GENEVA scores.

We used the GENEVA to study the expression of KCNAB2, a regulatory beta subunit of the potassium voltage-gated channels. Mutations of the gene are associated with Chromosome 1P36 Deletion Syndrome and Brugada Syndrome (Heilstedt et al., 2001; Portero et al., 2016). However, conditions that modulate its expression are not fully characterized. GENEVA identified multiple datasets in which the expression of KCNAB2 are associated with experimental conditions. Figure 1 shows one example (GSE103114) in which KCNAB2 is downregulated by Zika virus infection in monocytes.

Declarations

Software, code, and data availability: The GENEVA app is freely available at <https://genevatool.org/>. The code is available at <https://github.com/NavchetanKaur/geneva-webtool>.

Acknowledgments: We thank all researchers who have contributed RNA-seq datasets to the Gene Expression Omnibus. We thank Dr. Alexander Lachmann, Dr. Avi Ma'ayan, and other co-authors for creating ARCHS4, which enabled our study. We also thank Dr. Douglas Arneson and Sanchita Bhattacharya for their helpful discussions during the execution of this study.

Funding: This work was supported by the National Institute of Allergy and Infectious Diseases (bioinformatics support contract HHSN272201200028C and research grant UH2AI153016) and the common fund of the National Institute of Health (U24DK116214).

Competing interests: The authors declare no competing interests.

References

- Heilstedt, H.A. *et al.* (2001) Loss of the potassium channel beta-subunit gene, KCNAB2, is associated with epilepsy in patients with 1p36 deletion syndrome. *Epilepsia*, **42**, 1103–1111.
- Kaur, N. *et al.* (2020) Mining transcriptomics and clinical data reveals ACE2 expression modulators and identifies cardiomyopathy as a risk factor for mortality in COVID-19 patients.

Lachmann,A. *et al.* (2018) Massive mining of publicly available RNA-seq data from human and mouse. *Nat Commun*, **9**, 1366.

Lonsdale,J. *et al.* (2013) The Genotype-Tissue Expression (GTEx) project. *Nat Genet*, **45**, 580–585.

Portero,V. *et al.* (2016) Dysfunction of the Voltage-Gated K⁺ Channel β 2 Subunit in a Familial Case of Brugada Syndrome. *J Am Heart Assoc*, **5**, e003122.

Regev,A. *et al.* (2017) The Human Cell Atlas. *eLife*, **6**, e27041.

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

Figures

Figure 1. An example from GENEVA shows that the Zika virus regulates the expression of KCNAB2 in monocytes. (A) A bar plot showing the expression level of KCNAB2 in different samples within the GSE103114 dataset. (B) A scatter plot showing the multidimensional scaling embedding of the metadata of the GSE103114 dataset. The metadata embedding identifies the experimental groups within the dataset without manual curation. The color code represents the expression level of

Figure 1

See image above for figure legend.