Title

Phitest for analyzing the homogeneity of single-cell populations.

Permalink

https://escholarship.org/uc/item/5hm9c9vz

Journal

Bioinformatics, 38(9)

ISSN

1367-4803

Author

Li, Wei Vivian

Publication Date

2022-04-28

DOI

10.1093/bioinformatics/btac130

Peer reviewed

Bioinformatics, 38(9), 2022, 2639–2641
https://doi.org/10.1093/bioinformatics/btac130
Advance Access Publication Date: 3 March 2022
Applications Note

Gene expression

Phitest for analyzing the homogeneity of single-cell populations

Wei Vivian Li *

Department of Biostatistics and Epidemiology, Rutgers School of Public Health, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on August 3, 2021; revised on January 24, 2022; editorial decision on February 25, 2022; accepted on February 28, 2022

Abstract

Motivation: Single-cell RNA sequencing technologies facilitate the characterization of transcriptomic landscapes in diverse species, tissues and cell types with unprecedented molecular resolution. In order to better understand animal development, physiology, and pathology, unsupervised clustering analysis is often used to identify relevant cell populations. Although considerable progress has been made in terms of clustering algorithms in recent years, it remains challenging to evaluate the quality of the inferred single-cell clusters, which can greatly impact downstream analysis and interpretation.

Results: We propose a bioinformatics tool named Phitest to analyze the homogeneity of single-cell populations. Phitest is able to distinguish between homogeneous and heterogeneous cell populations, providing an objective and automatic method to optimize the performance of single-cell clustering analysis.

Availability and implementation: The PhitestR package is freely available on both Github (https://github.com/Vivianstats/PhitestR) and the Comprehensive R Archive Network (CRAN). There is no new genomic data associated with this article. Published data used in the analysis are described in detail in the Supplementary Data.

Contact: vivian.li@rutgers.edu

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

1 Introduction

Single-cell RNA sequencing (scRNA-seq) experiments enable gene expression measurement at a single-cell resolution, and provide an opportunity to characterize the molecular signatures of diverse cell types, states and structures (Haque *et al.*, 2017; Li, 2019). Because of these advantages, scRNA-seq technologies have been widely used in various biological disciplines, including developmental biology, neurology, immunology and cancer research.

In order to provide detailed catalogs of cells found in a sample and to enable convenient comparison in downstream analysis, unsupervised clustering is often used to group cells with similar transcriptome profiles into the same cluster (Duò *et al.*, 2018; Sheng and Li, 2021). To date, more than 20 clustering methods have been developed for scRNA-seq data to address challenges caused by high dimensionality, high sparsity, and technical noises. However, the evaluation and interpretation of cell clusters have been hampered by the scarcity of methods to analyze the heterogeneity of single-cell populations. The resolution of the inferred clusters often depends on clustering algorithms and software parameters. The inferred clusters may group cells of distinct cell types into one heterogeneous population or partition a homogeneous cell population into several distinct groups. If not being carefully evaluated, these errors will be propagated to downstream analyses including cell type annotation and differential expression analysis.

In this article, we propose a new method named Phitest to analyze the heterogeneity of single-cell populations, providing an objective and automatic method to evaluate the performance of clustering and quality of cell clusters. To the best of our knowledge, the only existing method available for a similar purpose is ROGUE, which calculates an entropy-based score (between 0 and 1) to quantify the purity of cell clusters (Liu *et al.*, 2020). In contrast, Phitest uses the Negative Binomial (NB) distribution to model unique molecular identifier (UMI) counts from scRNA-seq experiments. It then evaluates the homogeneity of a single-cell population based on the dispersion of the genes, and calculates a *P* value to guide decisions.

2 Materials and methods

As demonstrated in multiple studies (Sun *et al.*, 2021; Svensson, 2020; Townes *et al.*, 2019), the UMI counts of a gene in single cells sequenced in the same experiment can be characterized using an NB model. We use $X_{n\times m}$ to denote a UMI count matrix with n genes and m cells. Then, the NB model assumes that

$$X_{ij} \sim \mathrm{NB}(\mu_i, \phi_i), \quad i = 1, \ldots, n;\ j = 1, \ldots, m\ , \tag{1}$$

where μ_i is the mean and ϕ_i is the dispersion of gene i. The variance of gene i depends on its mean through a quadratic function: $\mu_i + \phi_i\mu_i^2$.

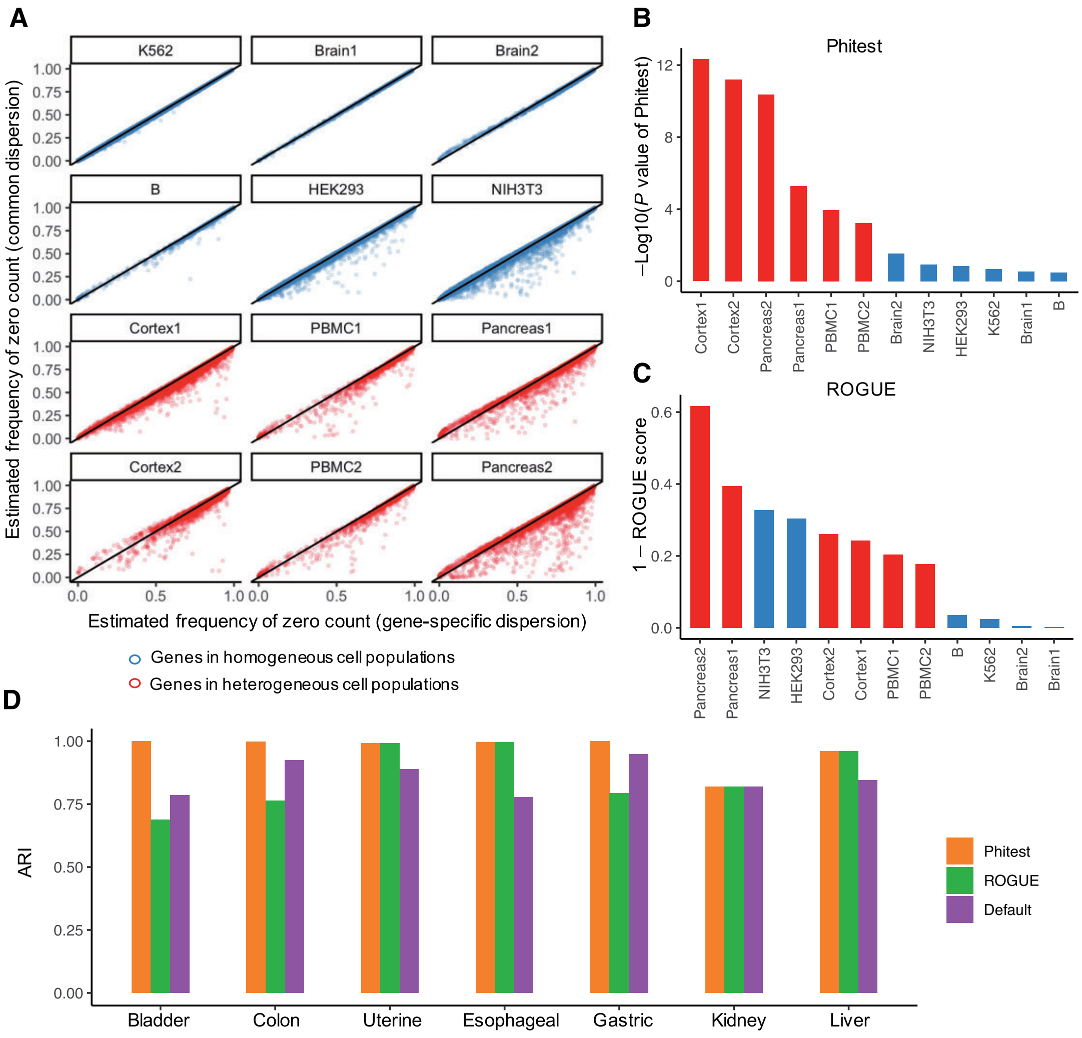

Fig. 1. Evaluation of the Phitest method. (**A**) Estimated gene-wise frequency of zero count using common or gene-specific dispersion parameters. (**B**) −Log10(*P* value of Phitest) for the homogeneous (blue) and heterogeneous (red) datasets. (**C**) (1 − ROGUE score) for the homogeneous (blue) and heterogeneous (red) datasets. (**D**) ARI of Seurat-inferred clusters using resolution parameters selected by Phitest, ROGUE or default setting

If the single cells in this count matrix come from a homogeneous population, then we expect a common dispersion parameter for all genes ($\phi_i = \phi_c, i = 1, \ldots, n$), since the genes are subject to similar sampling process and technical variation during sequencing. To estimate this common dispersion ϕ_c, we first calculate the sample mean and variance of each gene, denoted as $\bar{x}_i$ and s_i^2, respectively. Then, a linear regression model ($s_i^2 = \bar{x}_i + \phi_c \bar{x}_i^2$) is fitted to obtain $\hat{\phi}_c$ as the estimated coefficient, using genes whose mean and variance are below the 99th percentiles, respectively (Supplementary Fig. S1). In contrast, if the single cells come from a mixture of biologically different cell populations, then each gene is subject to different biological variation in addition to technical variation. Therefore, a common dispersion parameter would not be sufficient to characterize gene expression distribution. Phitest fits an NB distribution for each gene to estimate the gene-specific dispersion parameters. The estimated dispersion and mean of gene i are denoted as $\hat{\phi}_i$ and $\hat{\mu}_i$, respectively.

Based on the above results, Phitest infers if a single-cell population is homogeneous by comparing the common and gene-specific dispersion parameters. For gene i, the expected frequency of zero count is $(1 + \hat{\phi}_c \hat{\mu}_i)^{-1/\hat{\phi}_c}$ with the common dispersion and $(1 + \hat{\phi}_i \hat{\mu}_i)^{-1/\hat{\phi}_i}$ with the gene-specific dispersion. Phitest then compares the zero frequencies with a two-sample t test. If the single cells are homogeneous, there should not be significant difference between the two sets of zero frequencies. If the single cells contain multiple cell types, the common dispersion tends to under-estimate the zero frequencies of biologically variable genes compared with the gene-specific dispersion. Therefore, the P value from the t test can be used to detect heterogeneous single-cell populations.

3 Results

To evaluate the performance of Phitest, we first applied it to 12 scRNA-seq datasets with gold standard information (Supplementary Table S1). Six of the datasets (B, K562, Brain1, Brain2, HEK293 and NIH3T3) are known to contain homogeneous cell populations. The other six datasets (Cortex1, Cortex2, PBMC1, PBMC2, Pancreas1, Pancreas2) contain heterogeneous populations (Supplementary Methods).

We first confirmed that the gene-specific NB models fit the UMI counts well for both homogeneous and heterogeneous datasets (Supplementary Fig. S2). Next, we compared the estimated zero count frequencies based on the common or gene-specific dispersion parameters, and observed an apparent distinction between the homogeneous and heterogeneous datasets (Fig. 1A). In the heterogeneous datasets, the common dispersion under-estimates the gene-wise frequency of zero count, compared with those based on the gene-specific dispersion. Ordering the datasets based on P values calculated by Phitest, we could obtain a clear separation with the six heterogeneous datasets having the smallest *P* values (Fig. 1B). We also applied the ROGUE method (Liu *et al.*, 2020) to the 12 datasets. The heterogeneous datasets on average have smaller ROGUE scores, but ROGUE cannot distinguish two homogeneous populations (NIH3T3 and HEK293) from the heterogeneous populations (Fig. 1C).

We further evaluated the ability of Phitest in improving single-cell clustering analysis, using scRNA-seq datasets of seven cancer types with ground truth information (Supplementary Table S2 and Supplementary Methods). To cluster the cells, we selected the most widely used method Seurat, which has a resolution parameter to control the number of clusters (Stuart *et al.*, 2019). A larger resolution tends to result in more clusters and the default parameter is 0.8. To investigate if Phitest can help select the optimal parameter in unsupervised analysis, for each dataset, we applied Seurat to perform clustering with different resolution parameters (0.001, 0.01, 0.1, 0.4, 0.8 and 1.2), and then used Phitest to evaluate the results. For each parameter, we used 0.05 as a threshold on Phitest's *P* values to determine the homogeneity of single-cell clusters, and selected the parameter that led to the largest number of cells in homogeneous clusters. We also used ROGUE to assess the clusters. Since 0.82 is the largest ROGUE score of heterogeneous datasets in Figure 1C, we used 0.83 as a threshold to determine the homogeneity of single-cell clusters.

We compared the adjusted Rand index (ARI) of clusters identified with the parameters selected by Phitest, ROGUE, and default (Fig. 1D). The results show that Phitest leads to the best clustering performance, and parameters selected by Phitest outperform the default value. For example, the bladder Cancer dataset contains six cell types, but the default parameter leads to eight clusters, and ROGUE finds a resolution that leads to nine clusters. In contrast, Phitest finds a resolution that leads to accurate clustering (Supplementary Fig. S3).

4 Conclusions

In this work, we propose a bioinformatics and statistical method named Phitest to analyze the heterogeneity of single-cell populations. By evaluating if a reduced model with a common dispersion parameter can sufficiently explain the observed gene counts compared with a full model with gene-specific dispersion parameters, Phitest is able to detect heterogeneous single-cell populations. In real data applications, we suggest users to also investigate plots like Figure 1 and Supplementary Figure S2 in addition to the *P* values. We have demonstrated the accuracy of Phitest on 12 scRNA-seq datasets (six homogeneous and six heterogeneous) with a gold standard. In addition, we have also shown that Phitest can be used to select the optimal parameter in single-cell clustering analysis, leading to higher clustering accuracy than the default parameter. In summary, the Phitest method (implemented in the PhitestR package) will improve unsupervised analysis of scRNA-seq data by providing an objective and automatic tool to help evaluate the quality of cell clusters and the performance of clustering.

Acknowledgements

The author would like to thank the two anonymous reviewers for their insightful comments and suggestions.

Funding

This work was supported by the National Institutes of Health [R35GM142702] and Rutgers Busch Biomedical Grant to W.V.L.

Conflict of Interest: none declared.

References

Duò,A. *et al.* (2018) A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research*, **7**, 1141.

Haque,A. *et al.* (2017) A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.*, **9**, 1–12.

Li,W.V. (2019) *Statistical Methods for Bulk and Single-Cell RNA Sequencing Data*. University of California, Los Angeles, CA.

Liu,B. *et al.* (2020) An entropy-based metric for assessing the purity of single cell populations. *Nat. Commun.*, **11**, 1–13.

Sheng,J. and Li,W.V. (2021) Selecting gene features for unsupervised analysis of single-cell gene expression data. *Brief. Bioinform.*, **22**, bbab295.

Stuart,T. *et al.* (2019) Comprehensive integration of single-cell data. *Cell*, **177**, 1888–1902.

Sun,T. *et al.* (2021) scdesign2: a transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured. *Genome Biol.*, **22**, 1–37.

Svensson,V. (2020) Droplet scRNA-seq is not zero-inflated. *Nat. Biotechnol.*, **38**, 147–150.

Townes,F.W. *et al.* (2019) Feature selection and dimension reduction for single-cell RNA-seq based on a multinomial model. *Genome Biol.*, **20**, 1–16.