GENESORTER: FEATURE RANKING IN CLUSTERED SINGLE CELL DATA

A PREPRINT

Mahmoud M Ibrahim
Faculty of Medicine
RWTH Aachen University
Pauwelstr. 30, 52074 Aachen
mmibrahim@pm.me

Rafael Kramann
Faculty of Medicine
RWTH Aachen University
Pauwelstr. 30, 52074 Aachen
rkramann@ukaachen.de

June 25, 2019

ABSTRACT

Marker genes identified in single cell experiments are expected to be highly specific to a certain cell type and highly expressed in that cell type. Detecting a gene by model-based differential expression analysis does not necessarily satisfy those two conditions and is typically computationally expensive for large cell numbers. We present genesorteR, an R package that ranks features in single cell data in a manner consistent with the expected definition of marker genes. genesorteR is orders of magnitude faster than current implementations of differential expression analysis methods and can operate on data containing millions of cells. genesorteR is available at `https://github.com/mahmoudibrahim/genesorteR`.

1 Introduction

A main task in single cell gene expression data analysis (scRNA-Seq) is selecting marker genes after cell types are identified via unsupervised clustering of single cells. This is usually accomplished by differential expression analysis [1, 2]. On the experimental side, detecting a cell type's marker gene by immunostaining or in a FACS experiment is often used as a proxy for identifying all cells comprising that cell type. Therefore, marker genes are expected to be highly expressed in a specific cell type and only in that cell type (ie. highly expressed and highly specific). Detecting a gene by differential expression analysis does not necessarily satisfy those two conditions.

Typically only a small fraction of expressed genes is captured in any single cell. The detection rate of a gene, defined as the proportion of cells in which a gene is detected, relates to its expression level (Supplementary Figures S1-S2), and can be used to identify highly variable genes [3]. We propose that empirical statistics combining gene detection rate with cell clustering information can be used to reliably estimate the specificity of genes in cell clusters.

Here we present genesorteR, an R package that provides cell type specificity gene ranking metrics that are directly related to the expected "marker gene" definition in Biology. genesorteR can analyze millions of cells in a few minutes and provides various methods to select sets of marker genes and to assess cell clustering quality.

2 Approach

2.1 Feature Ranking by Specificity Scores

We start with a sparse $m \times n$ matrix of gene expression data for m genes and n cells and a partitioning of n cells into k types $t \in \{t_1, \cdots, t_k\}$ where $k > 1$ (Figure 1a). The posterior probability of observing a cell type t_i given that a gene g_j is detected is

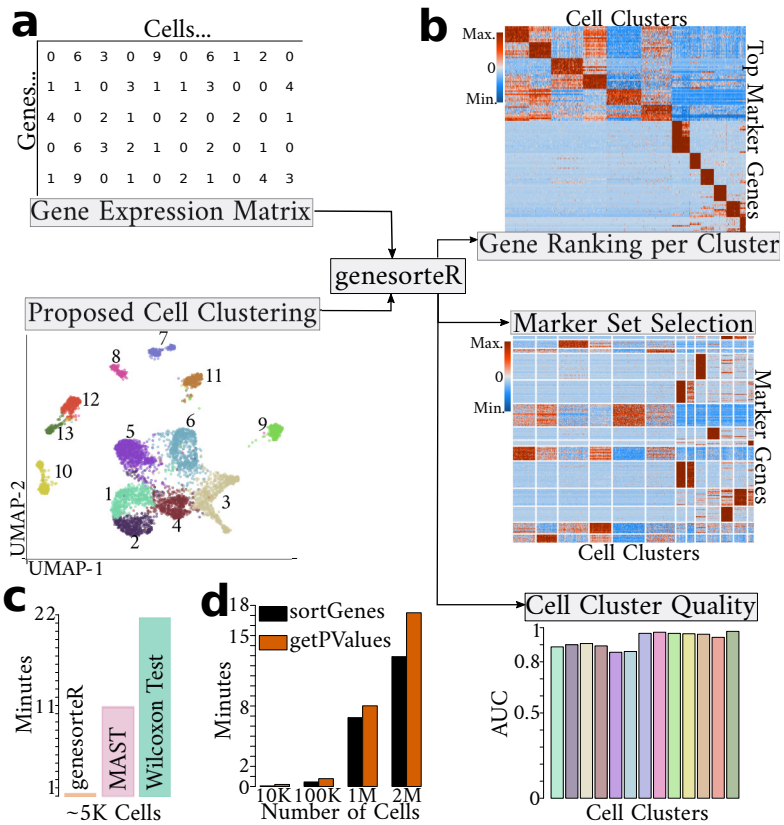

Figure 1: [A] The input to genesorteR is a gene expression matrix and a corresponding vector of cluster labels. [B] genesorteR, demonstrated here using 10x embryonic mouse heart data [4], can output accurate ranking of features for each cluster (top), select small sets of marker genes for each cluster (middle) and provide metrics for clustering quality (bottom, Supplementary Figure S7). [C] Running time of genesorteR (running sortGenes+getPValues sequentially) compared to MAST [5] and Wilcoxon Rank-sum test [6] as implemented in the Seurat R package finalAllMarkers function [7] (same data as [B]). [D] Running time of two main genesorteR functions on large datasets.

$$P(t_i|g_j) = \frac{P(g_j|t_i) \cdot P(t_i)}{P(g_j)} \quad , \quad \sum_{i=1}^{k} P(t_i|g_j) = 1 \tag{1}$$

where $P(g_j|t_i)$ is the fraction of cells assigned to t_i in which g_j was detected, $P(t_i)$ is the fraction of all n cells that were assigned to the partition t_i and $P(g_j)$ is the fraction of all n cells where g_j was detected. We define a Specificity Score for gene g_j in cell type t_i:

$$s_{ij} = \frac{P(g_j|t_i)^2 \cdot P(t_i)}{P(g_j)} \quad , \quad 0 \leq s_{ij} \leq 1 \tag{2}$$

$s_{ij} = 1$ iff g_j is detected in all cells assigned to t_i and in none of the cells assigned to $t_{x \neq i}$. The genesorteR function *sortGenes* returns an $m \times k$ matrix S of specificity scores that can be used to rank all genes for each cell type by their specificity scores (Figure 1b top, Supplementary Figures S3-S4).

2.2 Selecting Sets of Marker Genes

A marker gene will ideally have high s score in only one cell type. We define an entropy-based metric for each gene g_j as the Gene Specificity Shannon Index:

$$si_j = -\sum_{i=1}^{k} s_{ij}.log(s_{ij}) \tag{3}$$

Marker genes can be selected as those with high s scores and low si scores [8]. The *getMarkers* function selects a small set of marker genes (Figure 1b, middle) by selecting the top x percentile of genes from S and partitioning them into low si and high si genes (Supplementary Figure S5, Supplementary Text 1).

We may also be interested in an extended set of highly variable genes. The *getPValues* function performs a permutation test where the matrix S is recomputed after cluster labels are randomly permuted to obtain an empirical null distribution of specificity scores. *getPValues* returns an $m \times k$ matrix P of *p*-values expressing the probabilities of observing the scores in S if cells were randomly allocated to k partitions (Supplementary Figure S6, Supplementary Text 1).

2.3 Visualization

plotMarkerHeat (Figure 1b, middle) makes use of gene clustering and cell averaging to quickly summarize the expression of up to tens of thousands of genes over thousands to millions of cells. *plotMarkerHeat* clusters genes using the k-means algorithm applied to the Euclidean norm-scaled gene expression vectors (see Supplementary Text 1).

2.4 Implementation and Speed

genesorteR operates on sparse Matrices using the Matrix R package [9], and can process data containing thousands to millions of cells in a few seconds to a few minutes (Figure 1c-d).

3 Results

We tested genesorteR on various datasets including 10x chromium v3 data (~5k cells, Figure 1a,b, Supplementary Figures S3-S7)[4], Cel-Seq2 data (~50 cells, Supplementary Figure S8) [10], Smart-Seq2 data (~45k cells, Supplementary Figure S9)[11] and SPLiT-Seq data (~100k cells, Supplementary Figure S10)[12]. genesorteR can provide accurate gene ranking in all these datasets. Importantly, genesorteR's *getPValues* function provides well-calibrated *p*-values that better enrich for genes with higher expression fold-change ratio compared to genes selected by the Wilcoxon test [6] and MAST which also uses gene detection rate information [5] (Supplementary Figure S11, Supplementary Text 2).

4 Discussion

One challenge in single cell data analysis is to find differentially expressed genes despite data sparsity. Here we show that empirical statistics on gene detection rate can in fact provide accurate estimates of gene-cell type specificity. This is consistent with recent observations that droplet scRNA-Seq data is not zero-inflated [13].

Methods to evaluate cell clustering quality are desirable since clustering results will vary based on the selection of clustering parameters. genesorteR provides metrics for cell clustering quality evaluation (Figure 1b, bottom, Supplementary Text 1), based on the assumption that a well-defined cell cluster can be uniquely identified using relatively few marker genes. Indeed, a cell type is often defined by the expression of one or few marker genes that uniquely identify it. genesorteR ranks genes in a manner consistent with this marker gene definition.

5 Methods

genesorteR v0.1.6 was used for all analyses in this manuscript. Mouse heart data (Figure 1) was obtained as a filtered gene expression matrix from the 10x website [4]. Data was clustered as described in Supplementary Text 2. For Figure 1c, Seurat's *finalAllMarkers* function (v3.0.1) [7] was used to obtain results from the Wilcoxon Rank-sum test [6] and MAST [5] setting only.pos to TRUE. For Figure 1d, cells were sampled from SPLiT-Seq data [12] and clusters were randomly assigned as follows (cell number/cluster number): 0.1K/9 1K/18 10K/35 100K/75 1M/159 2M/177. Running time was evaluated using a single compute core. For further methods details, see Supplementary Text 2.

Acknowledgements. We would like to thank Ivan G. Costa for critical reading of the manuscript.

Funding. This research project is supported by the START-Program of the Faculty of Medicine, RWTH Aachen to MMI, and by a grant from the European Research Council (EERC-StG 677448) to RK.

References

[1] Mark D. Robinson and Gordon K. Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–2887, 2007.

[2] Tianyu Wang, Boyang Li, Craig E. Nelson, and Sheida Nabavi. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinformatics*, 20(1), 2019.

[3] Tallulah S Andrews and Martin Hemberg. M3drop: dropout-based feature selection for scRNASeq. *Bioinformatics*, 2018.

[4] 10x Genomics. 10k Heart Cells from an E18 Mouse (v3 chemistry). This dataset is licensed under the Creative Commons Attribution license (https://creativecommons.org/licenses/by/4.0/)., 2018.

[5] Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K. Shalek, Chloe K. Slichter, Hannah W. Miller, M. Juliana McElrath, Martin Prlic, Peter S. Linsley, and Raphael Gottardo. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology*, 16(1), December 2015.

[6] Frank Wilcoxon. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80, December 1945.

[7] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck, III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive Integration of Single-Cell Data. *Cell*, 177(7):1888–1902.e21, June 2019.

[8] A. Schliep, C. Steinhoff, and A. Schonhuth. Robust inference of groups in gene expression time-courses using mixtures of HMMs. *Bioinformatics*, 20:i283–i289, 2004.

[9] Douglas Bates and Martin Maechler. Matrix: Sparse and dense matrix classes and methods. *R package version 1.2-17*, 2019.

[10] Tamar Hashimshony, Naftalie Senderovich, Gal Avital, Agnes Klochendler, Yaron de Leeuw, Leon Anavy, Dave Gennert, Shuqiang Li, Kenneth J. Livak, Orit Rozenblatt-Rosen, Yuval Dor, Aviv Regev, and Itai Yanai. CEL-seq2: sensitive highly-multiplexed single-cell RNA-seq. *Genome Biology*, 17(1), 2016.

[11] The Tabula Muris Consortium, Overall coordination, Logistical coordination, Organ collection and processing, Library preparation and sequencing, Computational data analysis, Cell type annotation, Writing group, Supplemental text writing group, and Principal investigators. Single-cell transcriptomics of 20 mouse organs creates a tabula muris. *Nature*, 562(7727):367–372, 2018.

[12] Alexander B. Rosenberg, Charles M. Roco, Richard A. Muscat, Anna Kuchina, Paul Sample, Zizhen Yao, Lucas T. Graybuck, David J. Peeler, Sumit Mukherjee, Wei Chen, Suzie H. Pun, Drew L. Sellers, Bosiljka Tasic, and Georg Seelig. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*, 360(6385):176–182, 2018.

[13] Valentine Svensson. Droplet scRNA-seq is not zero-inflated. *bioRxiv 10.1101/582064*, 2019.