`singleCellHaystack`: Finding surprising genes in 2-dimensional representations of single cell transcriptome data

Alexis Vandenbon[1,2,*] and Diego Diez[3]

[1] Institute for Frontier Life and Medical Sciences, Kyoto University, Japan

[2] Institute for Liberal Arts and Sciences, Kyoto University, Japan

[3] Immunology Frontier Research Center, Osaka University, Japan

* To whom correspondence should be addressed.

Abstract

Summary: Single-cell sequencing data is often visualized in 2-dimensional plots, including t-SNE plots. However, it is not straightforward to extract biological knowledge, such as differentially expressed genes, from these plots. Here we introduce `singleCellHaystack`, a methodology that addresses this problem. `singleCellHaystack` uses Kullback-Leibler Divergence to find genes that are expressed in subsets of cells that are non-randomly positioned on a 2D plot. We illustrate the usage of `singleCellHaystack` through applications on several single-cell datasets. `singleCellHaystack` is implemented as an R package, and includes additional functions for clustering and visualization of genes with interesting expression patterns.

Availability and implementation: https://github.com/alexisvdb/singleCellHaystack

Contact: alexisvdb@infront.kyoto-u.ac.jp

1 Introduction

The parallel sequencing of transcriptomes of single cells (scRNA-seq) has revealed considerable heterogeneity in gene expression among single cells. In recent years, a myriad of bioinformatics and machine learning tools have become available for processing, analyzing and interpreting scRNA-seq data (Zappia *et al.*, 2018).

The high dimensionality of scRNA-seq data makes interpretation and visualization difficult. The standard way of dealing with this problem is to apply t-distributed stochastic neighbor embedding (t-SNE), and represent the data in fewer dimensions (i.e. a 2D plot) (van der Maaten and Hinton, 2008). Arguably, t-SNE is the most widely used method in the analysis of scRNA-seq data.

However, because of the high-dimensional raw data and the heterogeneity it contains, t-SNE often results in plots where it's hard to define boundaries between groups of cells. In some cases, there are exorbitant numbers of clusters, while in other cases there are

large agglomerates formed by many loosely connected subsets of cells. As a result, it is difficult to find interesting genes (i.e. genes that are detected in some subset of cells but not in others) in these plots, because it is unclear what subsets of cells to compare with each other.

Here, we present `singleCellHaystack`, a methodology that addresses this problem. `singleCellHaystack` uses Kullback-Leibler Divergence (D_{KL}; also called relative entropy) to find genes that are expressed in subsets of cells that are non-randomly positioned on a 2D plot (e.g. a t-SNE plot or plot of principal components) (Kullback and Leibler, 1951). The D_{KL} of each gene is compared with randomized data to evaluate its significance and estimate a p-value. `singleCellHaystack` does not rely on clustering of cells, and can detect any non-random pattern of expression in a 2D plot. An R package for running `singleCellHaystack` analysis and additional functions for visualization and clustering of genes is available at https://github.com/alexisvdb/singleCellHaystack.

2 Materials and methods

More details are given in Supplementary Material.

2.1 `singleCellHaystack` methodology

The main function, `haystack`, uses D_{KL} to estimate the difference between a reference distribution of all cells on a 2D plot (distribution Q) and the distributions of the cells in which a gene G was detected (distribution $P(G = T)$) and not detected (distribution $P(G = F)$).

To do so, first the 2D plot is divided into a grid along both axes. Next, a Gaussian kernel is used to estimate the density of cells at each grid point. Summing the contributions of all cells gives us Q; the subset of cells in which G is detected $P(G = T)$; and the subset of cells in which G was not detected $P(G = F)$. A small pseudo count is added to each grid point, and each distribution is normalized to sum to 1.

The divergence of gene G, $D_{KL}(G)$, is calculated as follows:

$$D_{KL}(G) = \sum_{s\in\{T,F\}} \sum_{x\in grid\ points} P(G = s, x) \log\left(\frac{P(G=s,x)}{Q(x)}\right) \quad \text{Eq. 1}$$

where $P(G = s, x)$ and $Q(x)$ are the values of $P(G = s)$ and Q at grid point x, respectively.

Finally, the significance of $D_{KL}(G)$ is evaluated using randomizations, in which the expression levels of G are randomly shuffled over all cells. The mean and standard deviation of $D_{KL}(G)$ in randomized datasets follow a clear pattern in function of the number of cells in which a gene was detected (see Supplementary Fig. S1 examples), which is modeled using B-splines (Schoenberg, 1946). P-values are calculated by comparing the observed $D_{KL}(G)$ to predicted mean and standard deviations (log values).

2.2 `singleCellHaystack` advanced options

The distribution Q and the randomizations described above ignore the fact that some cells have more detected genes than others. `singleCellHaystack` can be run in an advanced mode, in which both the calculation of Q and the randomizations are done by weighting cells by their number of detected genes (see Supplementary Material for more details).

In addition, `singleCellHaystack` includes functions for visualization and clustering gene expression patterns in the 2D plot.

2.3 scRNA-seq datasets and processing

We downloaded processed data (read counts or unique molecular identifiers) of the Tabula Muris project (Smart-seq2: 20 sets; Microfluidic droplets: 28 sets), the Mouse Cell Atlas (Microwell-seq: 87 sets) and a dataset of several hematopoietic progenitor cell types (Schaum *et al.*, 2018; Han *et al.*, 2018; Nestorowa *et al.*, 2016). For each dataset, cells and genes were filtered, the 1,000 most variable genes were selected, and principal component analysis (PCA) and t-SNE analysis were conducted, following the recommendations by Kobak and Berens (Kobak and Berens, 2018). Finally, `singleCellHaystack` was run on each dataset to find interesting genes in its t-SNE plot.

3 Results

We applied `singleCellHaystack` on 136 scRNA-seq datasets of varying sizes (149 to 19,693 cells). Median runtimes were and 75 and 84 seconds using the simple and advanced mode, respectively. For datasets containing < 5,000 cells, runtimes follow an approximately linear function of the number of cells in each dataset (Supplementary Fig. S2).

In all datasets, large numbers of genes were found to have significantly biased distributions on the t-SNE plot. This observation in itself is not surprising, since t-SNE reduces the distance between cells with similar gene expression profiles. Rather than interpreting `singleCellHaystack` p-values in the conventional definition, the ranking of genes is more relevant.

Figure 1 summarizes the result of the Tabula Muris fat tissue dataset (Smart-Seq2). 5,604 cells and 15,337 genes were used as input, and the `singleCellHaystack` run took 221s in the advanced mode. The t-SNE plot shows a typical mixture of clearly separated as well as loosely connected groups of cells, with considerable variety in the number of genes detected (Fig. 1A). The gene with the most significantly biased expression was Cd74, which is detected only in a few subsets of cells (Fig. 1B). To illustrate the variety in patterns, we grouped biased genes into 5 clusters based on hierarchical clustering of their density plot (Supplementary Fig. S3). Fig. 1C-F show the most significantly biased genes of the other 4 groups.

Results for two other example datasets are shown in Supplementary Fig. S4 and S5.

Limitations of `singleCellHaystack`

As noted above, `singleCellHaystack` returns inflated p-values because of the fundamental properties of PCA and t-SNE plots. In future updates we hope to address this issue.

4 Implementation

`singleCellHaystack` is implemented as an R package, available at https://github.com/alexisvdb/singleCellHaystack. The repository includes additional instructions for installation in R.

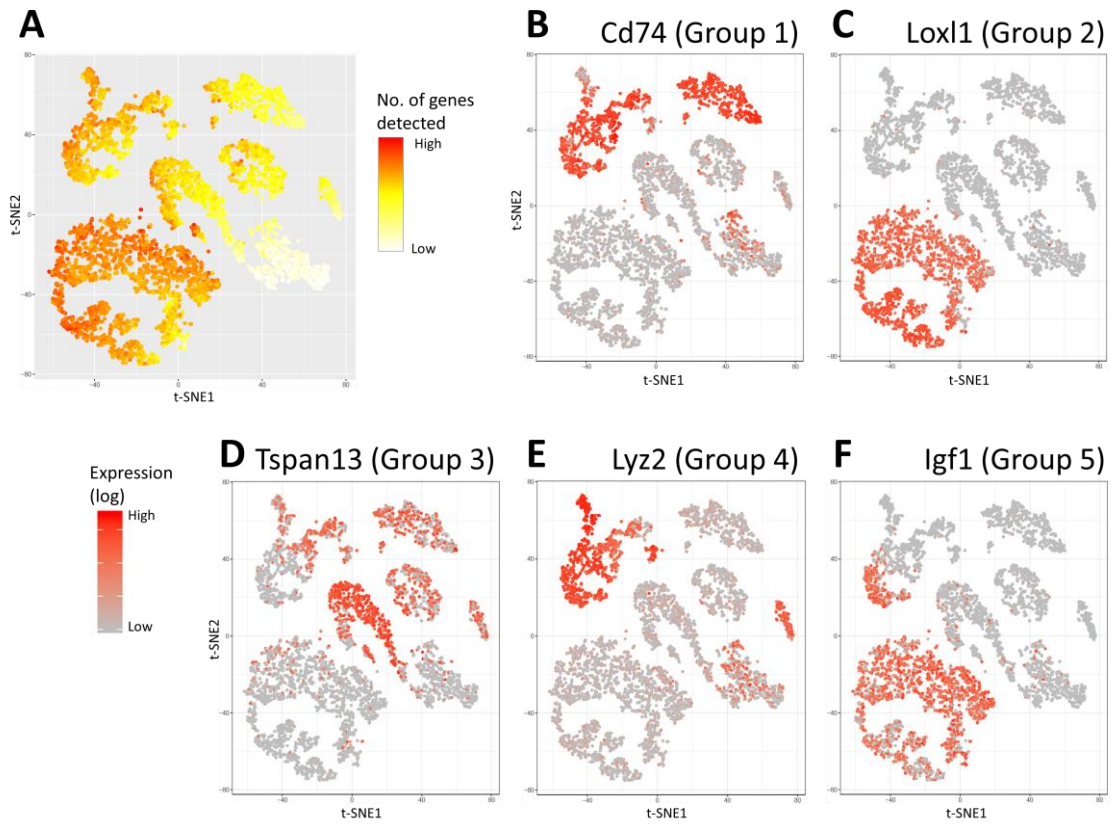

Figure 1: Application of `singleCellHaystack` on fat tissue dataset. (A) t-SNE plot of the 5604 cells. The color scale shows the number of genes detected in each cell. **(B-F)** Expression pattern of five highly biased genes, representative of the five groups in which the genes were clustered.

Author contributions

A.V. conceived of the project and methodology and ran the analyses. A.V. and D.D. implemented the methods and wrote the manuscript.

Conflict of Interest: none declared.

References

Han,X. *et al.* (2018) Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell*, **172**, 1091–1097.

Kobak,D. and Berens,P. (2018) The art of using t-SNE for single-cell transcriptomics. *bioRxiv*, 1–25.

Kullback,S. and Leibler,R.A. (1951) On Information and Sufficiency. *Ann. Math. Stat.*, **22**, 79–86.

van der Maaten,L. and Hinton,G. (2008) Visualizing Data using t-SNE. *J. Mach. Learn. Res.*, **9**, 2579–2605.

Nestorowa,S. *et al.* (2016) A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood*, **128**, e20–e31.

Schaum,N. *et al.* (2018) Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, **562**, 367–372.

Schoenberg,I.J. (1946) Contributions to the problem of approximation of equidistant data by analytic functions. *Q. Appl. Math.*, **4**, 45–99.

Zappia,L. *et al.* (2018) Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLoS Comput. Biol.*, **14**.