`singleCellHaystack`: Finding surprising genes in 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: A typical analysis of single-cell sequencing data includes dimensionality reduction using Principal Component Analysis (PCA) and visualization in 2-dimensional plots, using t-SNE or UMAP. However, identifying differentially expressed genes or extracting biological knowledge remains challenging, even after reducing dimensionality. Here we introduce `singleCellHaystack`, a methodology that uses Kullback-Leibler Divergence to find genes that are expressed in subsets of cells that are non-randomly positioned in a multi-dimensional space. Critically, `singleCellHaystack` does not rely on arbitrary clustering of cells. We illustrate the usage of `singleCellHaystack` through applications on several single-cell datasets. Genes with highly biased expression profiles often include known cell type marker genes. `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 use dimensionality reduction methods, such as PCA, t-distributed stochastic neighbor embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) to represent the data in fewer (typically 2) dimensions (van der Maaten and Hinton, 2008; McInnes *et al.*, 2018).

Unfortunately, the interpretation of t-SNE or UMAP plots is not straightforward either. Cells often form clusters, but it is often 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 in a multi-dimensional (≥2D) space (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 expression pattern of expression in a multi-dimensional space. 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

Our approach contains two main functions: `haystack_2D` and `haystack_highD`, for 2D and multi-dimensional (≥2D) input spaces, respectively. The concept of both functions is the same, and is briefly explained below. More details are given in Supplementary Material. We refer to Supplementary Fig. S1 for an overview of the workflow.

2.1 `singleCellHaystack` methodology

`singleCellHaystack` uses D_{KL} to estimate the difference between a reference distribution of all cells in a multi-dimensional space (distribution Q) and the distributions of the cells in which a gene G was detected (distribution $P(G = T)$) and not detected (distribution $P(G = F)$).

For 2D spaces (such as typical t-SNE or UMAP plots), `haystack_2D` divides the 2D space into a grid along both axes. For multi-dimensional spaces (such as the first several principal components), `haystack_highD` defines a set of grid points covering the subspace in which the cells are located (see Supplementary Material).

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 $P(G = F)$. Each distribution is normalized to sum to 1.

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

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

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. S2 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 multi-dimensional space.

2.3 scRNA-seq datasets and processing

We downloaded processed data (read counts or unique molecular identifiers) of the Tabula Muris project (FACS-sorted cells: 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 dimensionality reduced using PCA. Subsequently, the first 50 Principal Components (PCs) were used as input for t-SNE and UMAP analysis, following the recommendations by Kobak and Berens (Kobak and Berens, 2018). Finally, `singleCellHaystack` was run on both 2D (2D t-SNE and UMAP coordinates) and multi-dimensional (5, 10, 15, 25, or 50 PCs) representations of each dataset to find genes with biased expression patterns.

2.4 Known cell type marker genes

We downloaded marker gene data from the CellMarker database (Zhang *et al.*, 2019). A total of 7,852 unique mouse gene symbols are included as marker genes, which we split into 630 “high-confidence” markers (reported in ≥5 publications), and 7,222 “low-confidence” markers (reported in 1 to 4 publications). Other genes we regarded as “non-marker” genes.

3 Results

We applied `singleCellHaystack` on 136 scRNA-seq datasets of varying sizes (149 to 19,693 cells). Median runtimes of `haystack_highD` with 5 PC inputs were and 104 and 113 seconds using the simple and advanced mode, respectively. Runtimes followed an approximately linear function of the number of cells in each dataset

(Supplementary Fig. S3A-B). No difference in run times was seen for 10, 15, 25, or 50 PC input (not shown). Median runtimes for `haystack_2D` on 2D t-SNE coordinates were 75 and 84 seconds using the simple and advanced mode, respectively (Supplementary Fig. S3C-D).

In all datasets, large numbers of genes were found to have significantly biased distributions in the input spaces. This observation in itself is not surprising, since samples typically include a variety of different cell types, often forming loose clusters in the PC space, which can be further strengthened by t-SNE and UMAP. Rather than interpreting `singleCellHaystack` p-values in the conventional definition, the ranking of genes is more relevant.

As an illustration of the usage of `singleCellHaystack`, we here present 3 example results of datasets based on different sequencing technologies. In all three cases, the coordinates of cells in the first 5 PCs was used as input, along with the detection levels of all genes in all cells.

Figure 1 summarizes the result of the Tabula Muris marrow tissue dataset (FACS-sorted data). 5,250 cells and 13,756 genes were used as input, and the `singleCellHaystack` run took 225s in the default 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 *Ifitm2*, which is detected only in a 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. S4). 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. S5 and S6.

Known cell type marker genes often have biased expression patterns

Focusing on results of `haystack_highD` applied on the first 5 PCs of each dataset, we investigated whether genes with strongly biased expression are often known marker genes. For each dataset, we ranked genes by their p-value, and counted how often the gene at each rank was a high-confidence marker, a low-confidence marker, or a non-marker gene (see Materials and methods). High-confidence marker genes (such as *Cd45* and *Kit*) were strongly enriched among top biased genes: although they comprise only 2.2% of all genes, on average 32.7% of the top 50 ranked genes were high-confidence cell type markers (Supplementary Fig. S7).

The advanced mode takes into account general gene detection levels

In scRNA-seq data, there can be considerable variation in the number of detected genes in each cell. In some datasets this results in clusters of cells with higher or lower general detection levels. The “advanced” mode of `singleCellHaystack` can be used to find genes that have biased expression patterns that are contrary to the general pattern of detected genes. Supplementary Fig. S8 shows three example results, comparing the “advanced” mode to the “default” mode.

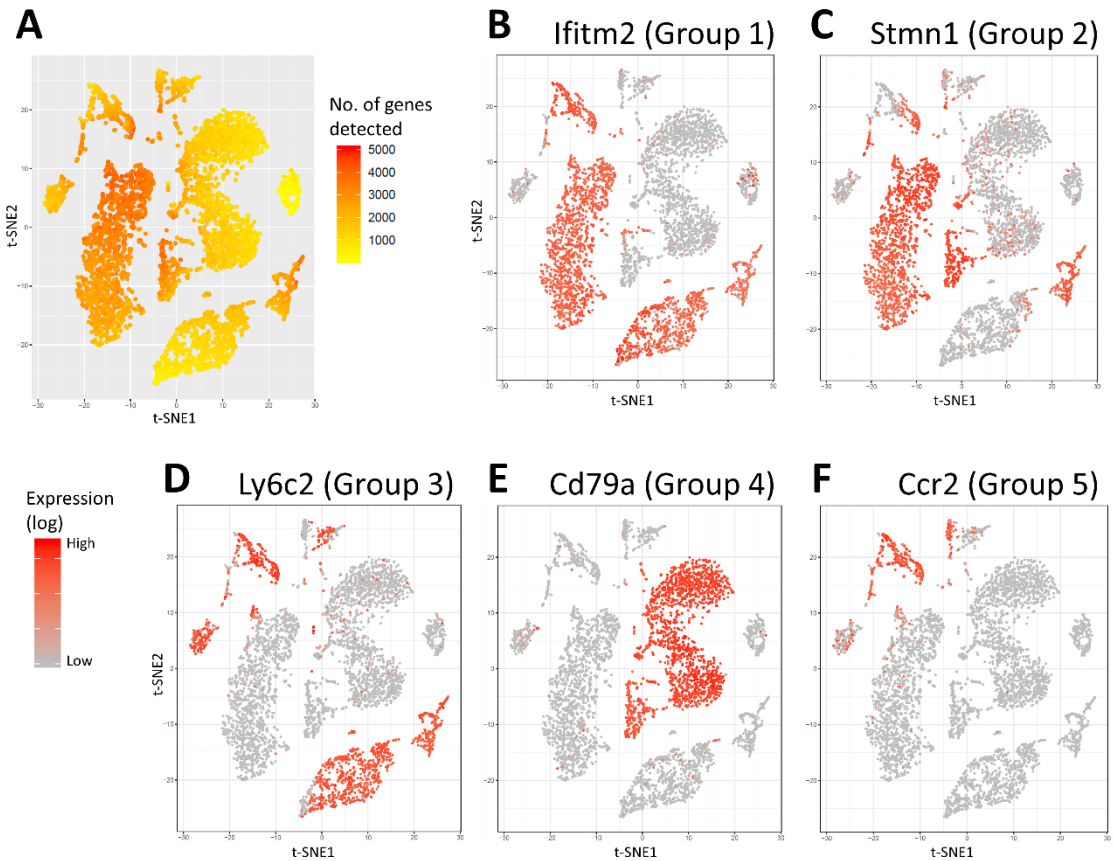

Figure 1: Application of `singleCellHaystack` on marrow tissue dataset. (A) t-SNE plot of the 5,250 cells. The color scale shows the number of genes detected in each cell. **(B-F)** Expression pattern of five highly biased genes, representative of the five groups in which the genes were clustered.

Consistency of results

The definition of the grid points in `haystack_highD` is not deterministic (see Supplementary Material). As a result, grid points differ between each run. To evaluate how much this impacts results, we ran `haystack_highD` ten times on each dataset using the first 5 PCs as input. For each dataset, we calculated the mean rank of each gene in the results and compared the ranking in individual runs to the mean ranking. In general, there was a high consistency in the ranking of genes, suggesting that the differences in grid points have only a limited impact on results (Supplementary Fig. S9A).

In addition, we observed that results of applications on 5 PCs were in general consistent with applications of `haystack_2D` on t-SNE and UMAP coordinates (both based on 50 PCs, Supplementary Fig. S9B-C), confirming that part of the information contained in the first 50 PCs is indeed captured in the t-SNE and UMAP coordinates. In contrast, there was larger discrepancy with results using only the 2 first PC coordinates as input (Supplementary Fig. S9D). Discrepancies increased as the difference in input

dimensionality increases (10 PCs < 15 PCs < 25 PCs < 50 PCs; Supplementary Fig. S9E-H).

Limitations of `singleCellHaystack`

As noted above, `singleCellHaystack` returns inflated p-values, because the input coordinates (PCs, t-SNE or UMAP coordinates) are dimensions which contain a large proportion of the variability in the original data. In future updates we hope to address this issue.

4 Implementation

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

5 Conclusions

`singleCellHaystack` is a generally applicable method for finding genes with biased expression patterns in single-cell transcriptome data. `singleCellHaystack` does not rely on clustering of cells, thus avoiding arbitrary clustering of cells in t-SNE or UMAP plots. It can detect any non-random pattern of expression, and can be a useful tool for finding new marker genes. The `singleCellHaystack` R package includes additional functions for clustering and visualization of genes with interesting expression patterns.

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.

Acknowledgements

We thank the members of the Lab. of Systems Virology (Kyoto University), the Lab. of Functional Analysis *in silico* (Tokyo University), and Dr. Yutaro Kumagai for helpful discussions and advice.

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.

McInnes,L. *et al.* (2018) UMAP : Uniform Manifold Approximation and Projection for Dimension Reduction. *arxiv*, 1–51.

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**.

Zhang,X. *et al.* (2019) CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res.*, **47**, D721–D728.