Identifying spatial expression trends in scRNA-seq data with trendsceek

Article #8: Edsgärd, D., Johnsson, P. & Sandberg, R. Identification of spatial expression trends in single-cell gene expression data. Nat Methods 15, 339–342 (2018).

https://doi.org/10.1038/nmeth.4634

Elaine Lee, lee02326@umn.edu, 5541154

Csci 5461

Spring 2021

Introduction

Spatial transcriptomics gives information about the location of gene activity in a tissue sample, and there are many existing ways to measure spatial gene expression. It is necessary to develop and test computational tools to analyze spatial transcriptomics data and reveal trends about the gene expression and location of cells that can help improve understanding of biological processes and diseases. These tools need to be tested for accuracy and reproducibility to be sure they give reliable results.

Trendsceek is a set of computational tools developed by the authors of the original study to analyze different types of spatial gene expression data[1]. The purpose of this project was to use their analysis method to reproduce the results from one of the datasets used in the study[2], and to apply that method to another open source scRNA-seq dataset on mouse dorsal root ganglion[3]. The motivation was to determine whether the trendsceek algorithm and their analysis method yields reproducible results and to see whether the method can be applied to a new dataset, as well as to analyze spatial expression patterns in the new dataset.

Data preparation

This project used two publicly available scRNA-seq datasets. The first dataset on mouse gastrulation was used in the original study and consists of gene expression counts from 481 mouse epiblast cells. The second is from a 2014 study on mouse lumbar dorsal root ganglion, consisting of 862 cells. Both datasets were filtered to exclude genes with low expression across all cells, defined as having less than 3 cells with at least 5 counts. After filtering, there were 17625 genes in the mouse gastrulation dataset and 15788 genes in the dorsal root ganglion

dataset. All zero expression values were set to 1 to allow the log function to be applied as part of the analysis.

Methods

To select the genes with the most variation in their expression levels for analysis, a gene-variability statistic was calculated for each gene. It was assumed that technical variability accounted for most of the variability in gene expression and that the expression of a gene followed a negative binomial distribution where the variance depended on the mean read count (m) based on the equation: $v = m + m^2/r$, where r is the overdispersion. The squared coefficient of variance was: $cv^2 = v/m^2 = 1/m + 1/r$. A generalized linear model was fitted to the cv^2 and average read counts of all the genes to estimate the technical variability, and this was used to normalize the gene-variability statistic.

To assign the 2D positions of the cells based on variations in gene expression, PCA and t-SNE were used to reduce dimensionality so that the spatial expression can be easily visualized. PCA is a linear technique to reduce dimensions and maximize variance, which preserves large distances. The initial number of dimensions were reduced using PCA before calculating more precise positional information with t-SNE. t-SNE focuses on small pairwise distances and is more accurate than PCA on non-linear datasets[4].

Trendsceek was the main algorithm developed to determine whether a gene has significant spatial expression trends. It takes a point pattern as input and calculates four different summary statistics for each pair of points: mark-correlation, mean-mark, variance-mark, and mark-variogram, and tests for significance against a null distribution. The main summary

statistics used were mark-correlation, defined as: $\frac{E(m_1 m_2)_P(r)}{\overline{m}^2}$, and mark-variogram, defined as: $E[\frac{(m_1-m_2)^2}{2}]_P(r)$ where m_1, m_2 = marks and r = distance between them. The null distribution is calculated by permuting the marks of the point pattern many times to obtain a distribution that is independent of spatial location.

Experiment design

After data preparation, the lowest 10% of genes based on expression level were dropped. The remaining genes were used to calculate normalized gene-variability statistics and the 500 most variable genes were selected. The variance vs average read counts of the selected genes were plotted to check for linearity. The read counts were normalized by size factors using existing algorithms in the DESeq2 library.

Then t-SNE was used to assign position information based on expression levels. The parameters used were 100 initial PCA dimensions, perplexity fraction of 0.2, and 300 iterations to reach convergence. A 2D marked point pattern was created using the t-SNE positions and normalized read counts.

The top 10 genes with the highest variation were selected as input into the trendsceek algorithm to calculate their summary statistics and determine whether they have a significant spatial trend. Due to limited computing power, the parameter for calculating the null distribution was set to 50 permutations. The significant genes were selected and their spatial expression patterns were plotted.

Results

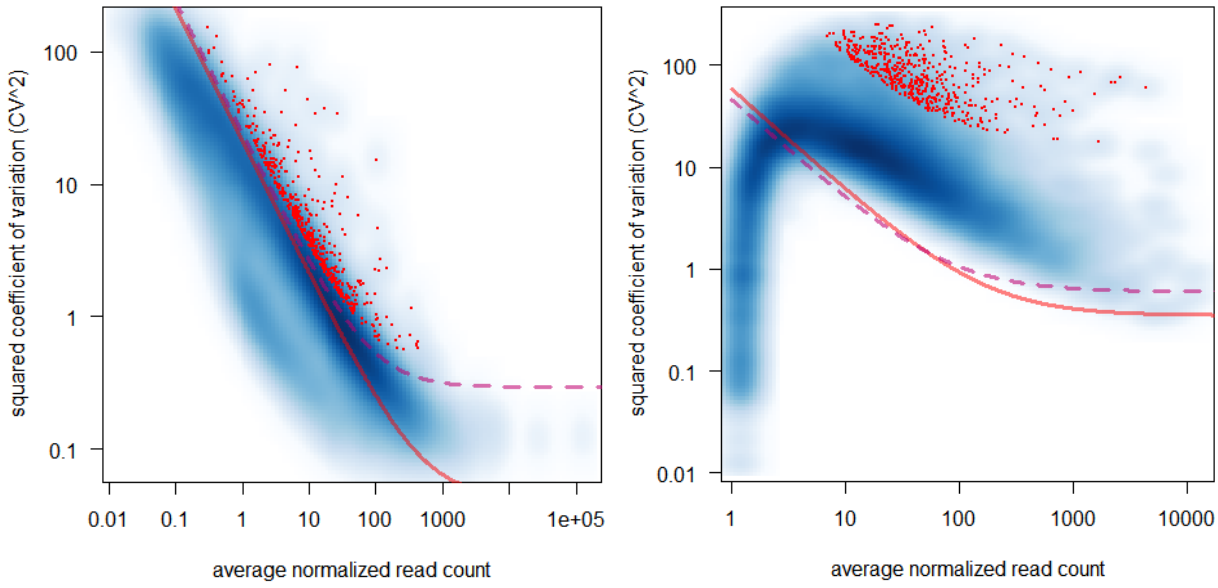

Figure 1.1, 1.2: Variability vs average read count of top 500 most variant genes in each dataset

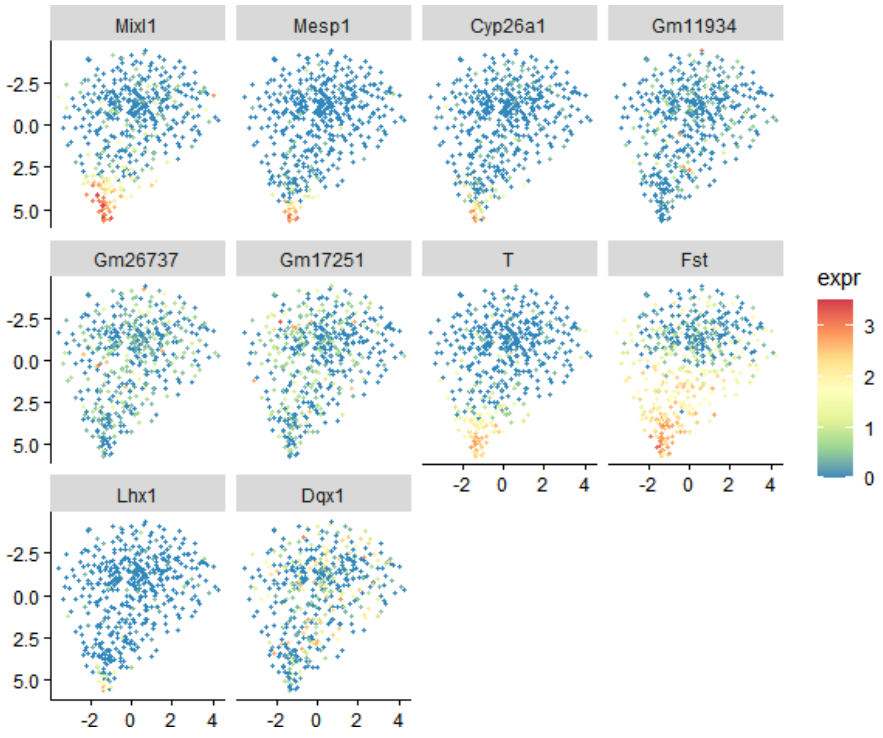

Figure 2.1: Spatial expression of top 10 most variable genes, dataset 1

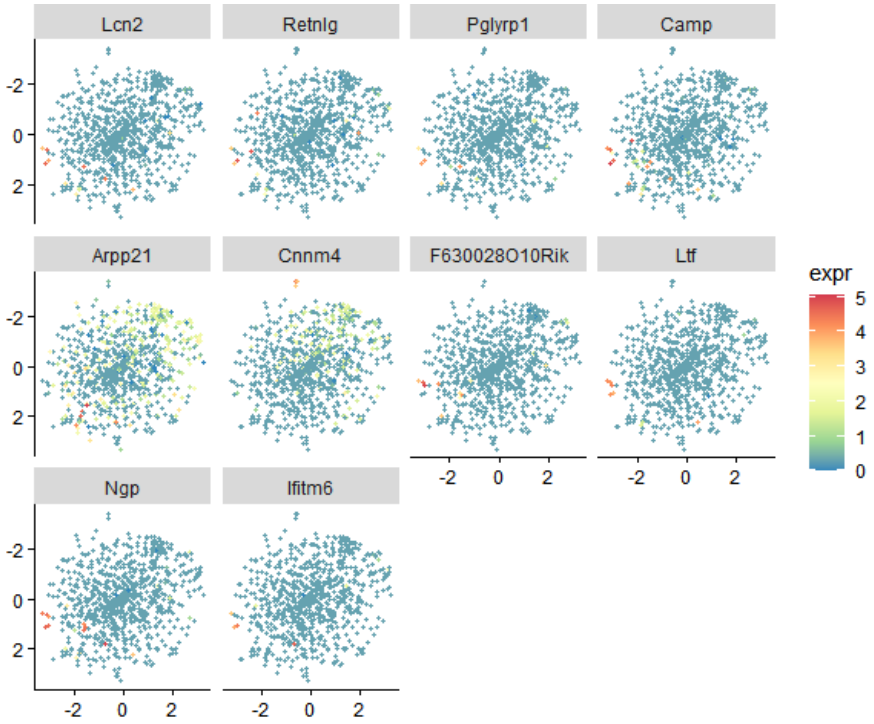

Figure 2.2: Spatial expression of top 10 most variable genes, dataset 2

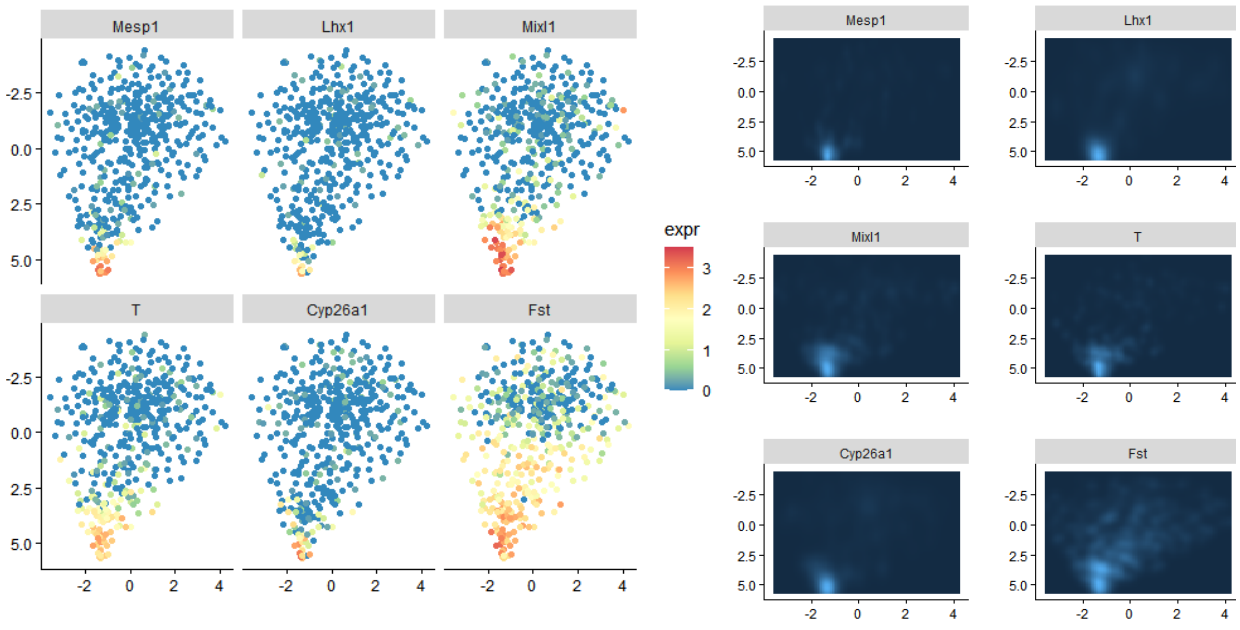

Figure 3.1: Mark correlation of significant genes found by trendsceek ($p<0.05$), dataset 1

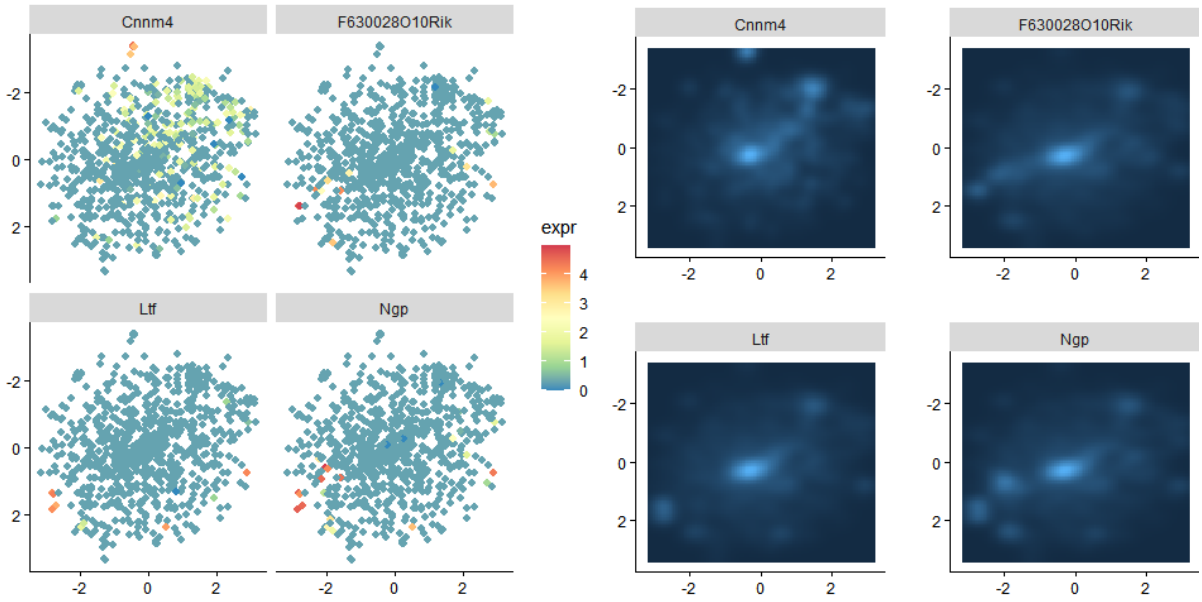

Figure 3.2: Mark correlation of spatially variable genes (p<0.10); no significant genes were found by trendsceek, dataset 2

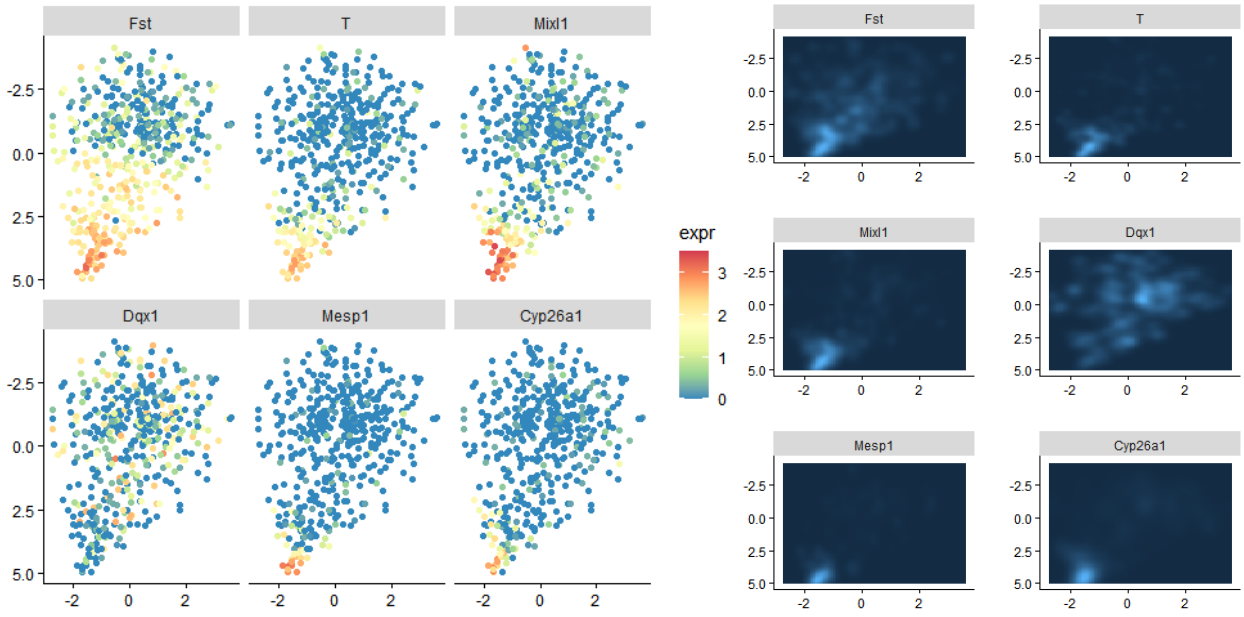

Figure 4.1: Mark variogram of significant genes found by trendsceek (p<0.05), dataset 1

Discussion

Figure 3.1 shows the mark-correlation statistics of the significant genes from the first dataset and it matches the expected results from the study, meaning the method gives reproducible results. Figure 4.1 shows the mark-variograms of significant genes, and it mostly matches the expected results except two genes, Dqx1 and Mesp, that were not expected to be significant. No significant spatial trends were found for the second dataset using the same method, but it was possible to identify genes with high variability and visualize their spatial expression patterns.

Comparing Figures 1.1 and 1.2, the second dataset had a less linear relationship between variability and mean read count, so the linear model was likely not a good estimate of the variability in the dataset. This may have led to errors when selecting the top genes and affected later parts of the analysis. To improve this, it would be necessary to change some parameters or find a different model of gene variability for this dataset.

Calculating the null distribution is a crucial part of the trendsceek algorithm, however its runtime increases with the number of permutations and sample size. The study originally used 10,000 permutations to obtain the null distribution but in order to reduce the amount of computing power needed, only 50 permutations were used for each dataset in this experiment. This may have led to errors in the null distribution, which would lead to inaccurate p-values when determining significant spatial trends, and this is likely the cause of the difference in Figure 4.1. This problem could be addressed by limiting the number of datapoints input into the trendsceek algorithm, however selecting a smaller sample can introduce other sources of error.

Conclusion

Trendsceek is a helpful tool for analyzing spatial gene expression data and can be applied to many data types. This project used it to analyze two scRNA-seq datasets and was able to reproduce most of the results from the first dataset even after changing the parameters for the null distribution, which may have negatively affected its accuracy. The analysis method was also applied to a new dataset to identify highly variable genes and visualize their spatial expression patterns. Areas of improvement include optimizing the accuracy and runtime based on the existing parameters and finding a better way to calculate variability for data that doesn't fit the linear model, and this project can also be extended by using trendsceek on other data types.

References

1. Edsgärd D. et al., Identification of spatial expression trends in single-cell gene expression data, Nature Methods, 2018. doi:10.1038/nmeth.4634

2. Scialdone, A., Tanaka, Y., Jawaid, W. *et al.* Resolving early mesoderm diversification through single-cell expression profiling. *Nature,* 2016. 535, 289–293. https://doi.org/10.1038/nature18633

3. Usoskin D, Furlan A, Islam S, Abdo H et al. Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nat Neurosci,* 2015 Jan;18(1):145-53. PMID: 25420068

4. An Introduction to t-SNE with Python Example. *Medium,* 2018. Retrieved 15 May 2021, from https://towardsdatascience.com/an-introduction-to-t-sne-with-python-example-5a3a293108d1#:~:text=t%2DDistributed%2