Applications of topological data analysis to single-cell genomics

Keshav Motwani

November 29, 2020

Background

- Single-cell RNA sequencing (scRNA-seq) allows us to measure gene expression in thousands of cells at once
- Previously, only bulk RNA-seq was possible, meaning the observed gene expression was the result of summing across all cells within a sample
- scRNA-seq is scientifically useful as it allows us to understand what role specific cell types play in biological processes
- Resulting data is in the form of a cells by genes matrix (approximately 30,000 genes) per sample

Application

- Is it possible to detect differences in gene expression caused by treating blood cells with Interferon-γ?
    - Interferon-γ is known to induce a variety of immune responses.
- What cell types does Interferon-γ modulate?

- Dataset:
    - Kang et al. 2018 published scRNA-seq data from blood cells pre- and post-treatment for a total of 8 patients
    - Every cell is annotated with a cell type label

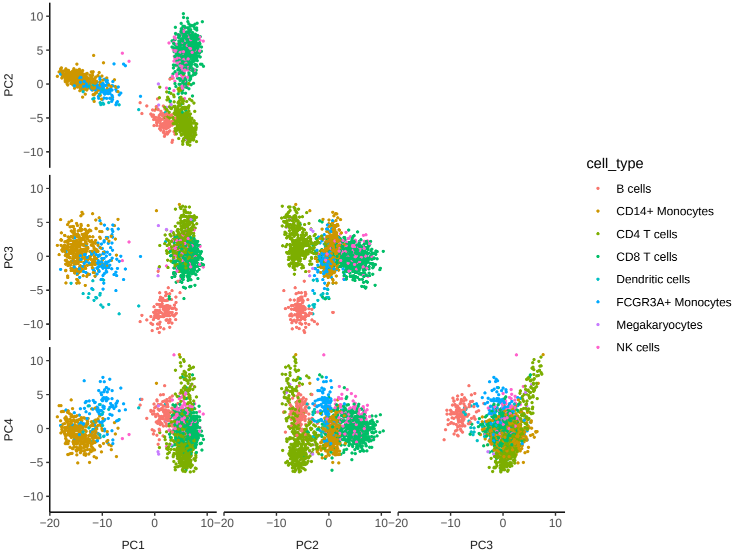
Motivation for using TDA

- We want to understand differences in the distribution of cells in gene expression space that is caused by a treatment, which fits perfectly into the TDA workflow described in class for point clouds
- Currently, there exist no published methods to classify entire scRNA-seq samples other than to simply average gene expression over all cells in the sample, and applying standard classification algorithms on the averaged data
  - No benefit over older bulk RNA-seq technology with that method

Data preprocessing

- Filter out dead cells and doublets
- Represent each sample based on it’s top 50 principal components
    - For computational feasibility in computing pairwise Euclidian distances

Data example (one patient, pre-treatment)



Simplicial complex construction

- Vietoris-Rips complex with varying radius
- 200 values of radius equally spaced from 0 to R, where R is chosen to be the 0.1 quantile of the values inside the pairwise distance matrices

Persistent homology computations

Let A be one scRNA-seq sample. We consider $p \in \{0, 1\}$.

- We have

$$\mathrm{VR}_1(A) \subset \mathrm{VR}_2(A) \subset \cdots \subset \mathrm{VR}_{200}(A).$$

- Compute homology in degree p

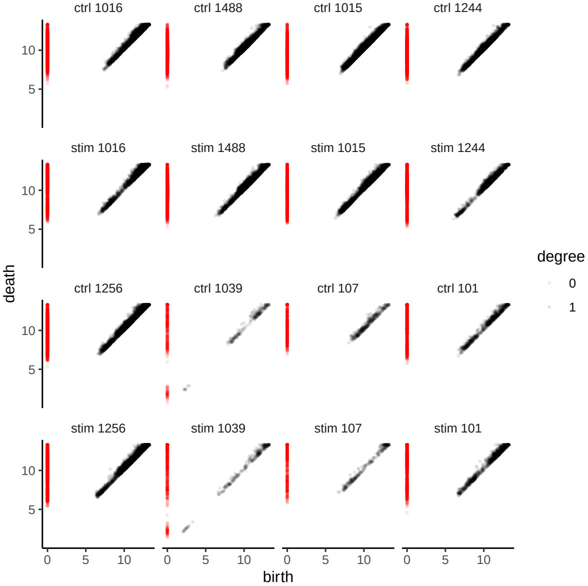
$$H_p(\mathrm{VR}_j(A)) = Z_p(\mathrm{VR}_j(A))/B_p(\mathrm{VR}_j(A))$$

- We get

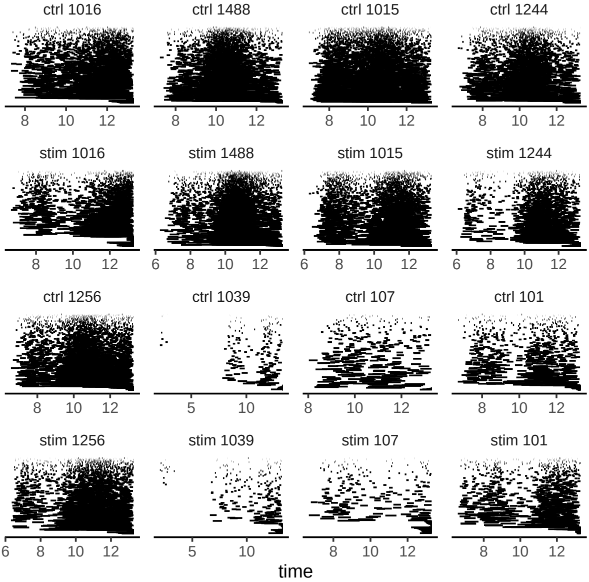
$$H_p(\mathrm{VR}_1(A)) \mapsto H_p(\mathrm{VR}_2(A)) \mapsto \cdots \mapsto H_p(\mathrm{VR}_{200}(A)).$$

- Then we compute barcodes and persistence landscapes for statistics and machine learning

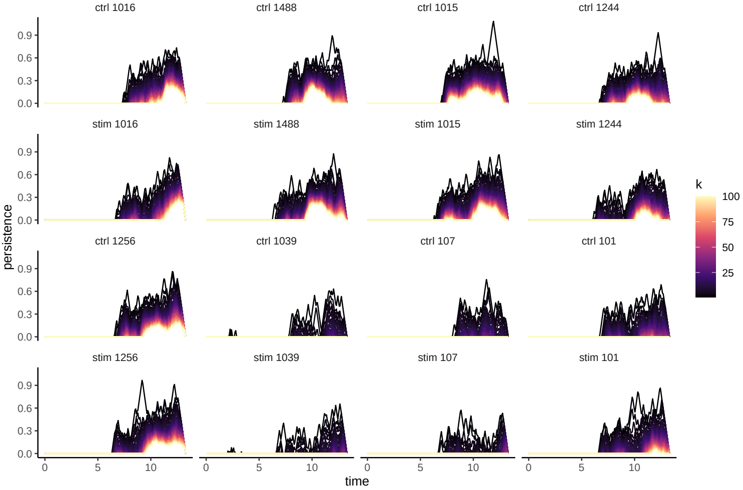
Features in H_0 and H_1



H_1 barcodes


ctrl 1016
8 10 12
ctrl 1488
8 10 12
ctrl 1015
8 10 12
ctrl 1244
8 10 12
stim 1016
8 10 12
stim 1488
6 8 10 12
stim 1015
6 8 10 12
stim 1244
6 8 10 12
ctrl 1256
8 10 12
ctrl 1039
5 10
ctrl 107
8 10 12
ctrl 101
8 10 12
stim 1256
6 8 10 12
stim 1039
5 10
stim 107
8 10 12
stim 101
8 10 12
time

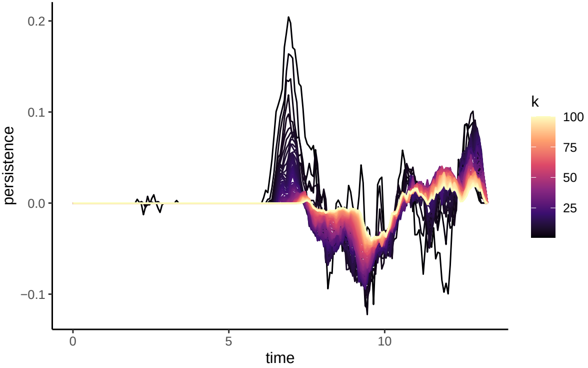
Persistence landscapes



Average persistence landscape difference

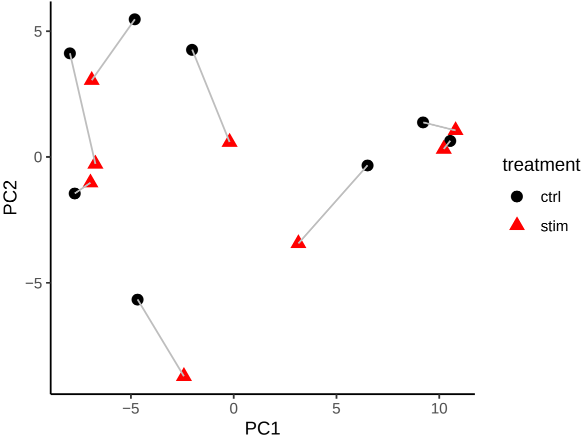
Some features that:

- persist longer in earlier timepoints post-treatment
- persist longer in middle timepoints pre-treatment
- persist longer in later timepoints post-treatment



PCA on persistence landscapes

There is separation by treatment status in PC2, generally post-treatment has lower PC2 values



Paired sample permutation test

Let $X_i \sim F(\text{PL}(X_i))$ and $Y_i \sim F(\text{PL}(Y_i))$ denote the pre- and post-treatment sample from the ith patient respectively, and $\text{PL}(\cdot)$ be the true persistence landscape vector.

- Want to test the null hypothesis that $\text{PL}(X_i) = \text{PL}(Y_i)$

Test statistic:

- $T(X,Y) = ||\frac{1}{N}\sum_{i=1}^{N} \hat{\text{PL}}(X_i) - \hat{\text{PL}}(Y_i)||_2$
- Same as two-sample test statistic

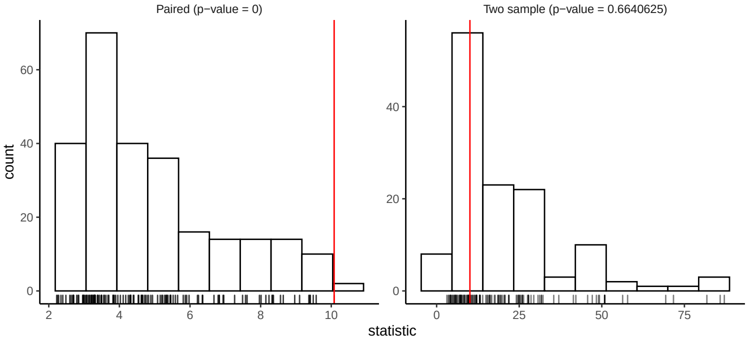
Null distribution:

- Construct permutations that permute treatment status only within the same patient, let X^*, Y^* denote the permutations where

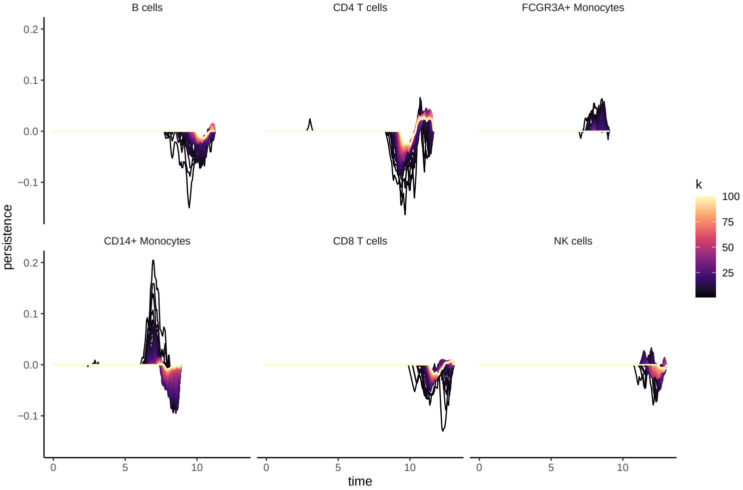
$$X_i^* = X_i \text{ or } Y_i, \quad Y_i^* = \begin{cases} X_i & \text{if } X_i^* = Y_i \\ Y_i & \text{otherwise} \end{cases}$$

- There are 2^N such permutations

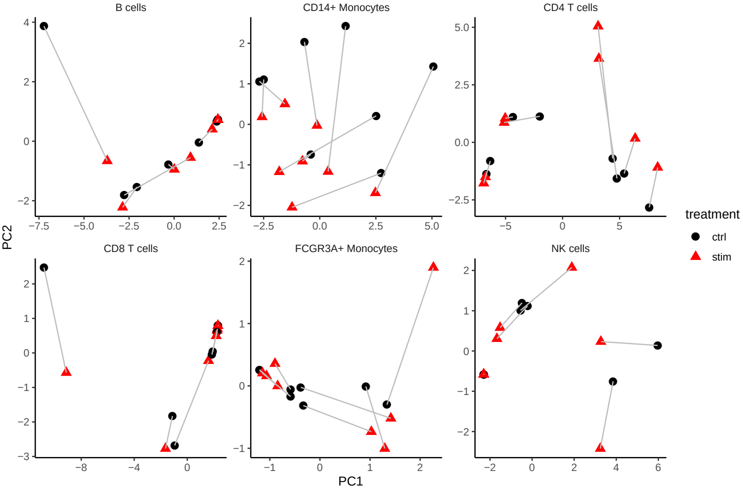
Permutation test



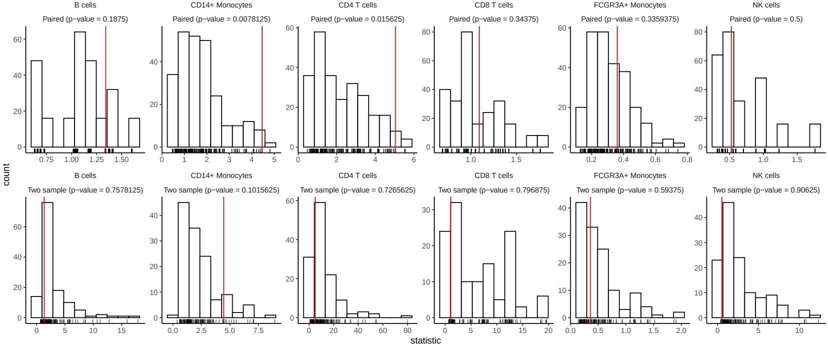
Average persistence landscape difference per cell type



PCA on persistence landscapes per cell type



Permutation test per cell type



Conclusions

- Persistence landscapes show differences pre- and post-treatment in the same patient
- CD14+ Monocytes and CD4+ T cells show the largest differences after Interferon-γ treatment
- Topological data analysis is a promising method for single-cell genomics and should be explored further