Unsupervised Learning for Single-Cell Transcriptomic Data

Anthony Degleris, Spencer Guo, Clara Kelley

CS 221 Final Project

Stanford

Background

- Single-cell RNA sequencing has led to a rapid increase in the **size and complexity** of gene expression data.
- RNA transcriptome profiles are **high-dimensional**, i.e. many more genes (features) than cells (examples).
- Some genes are more influential in governing cell type.
- Some cell types are well described by transcriptome archetypes. Archetypes allow us to determine gene expression clusters.

- **Problem Statement:** We seek to reduce the high-dimensional gene expression data in two ways:
  1. By identifying the *most influential genes* in explaining the transcriptome variance.
  2. By identifying *transcriptome archetypes* that naturally describe the common cell types.

Dataset

- This project leveraged a Gene Expression Omnibus dataset of zebrafish embryo cells over **7 time steps in a 24 hour period**.
- Each timestep contains the RNA transcriptome of between **3000 and 5000 cells**. Each transcriptome contains the count of unique molecular identifiers for all of **30,000+ gene markers**.
- Information known of particular genes was taken from NCBI's gene lookup database.

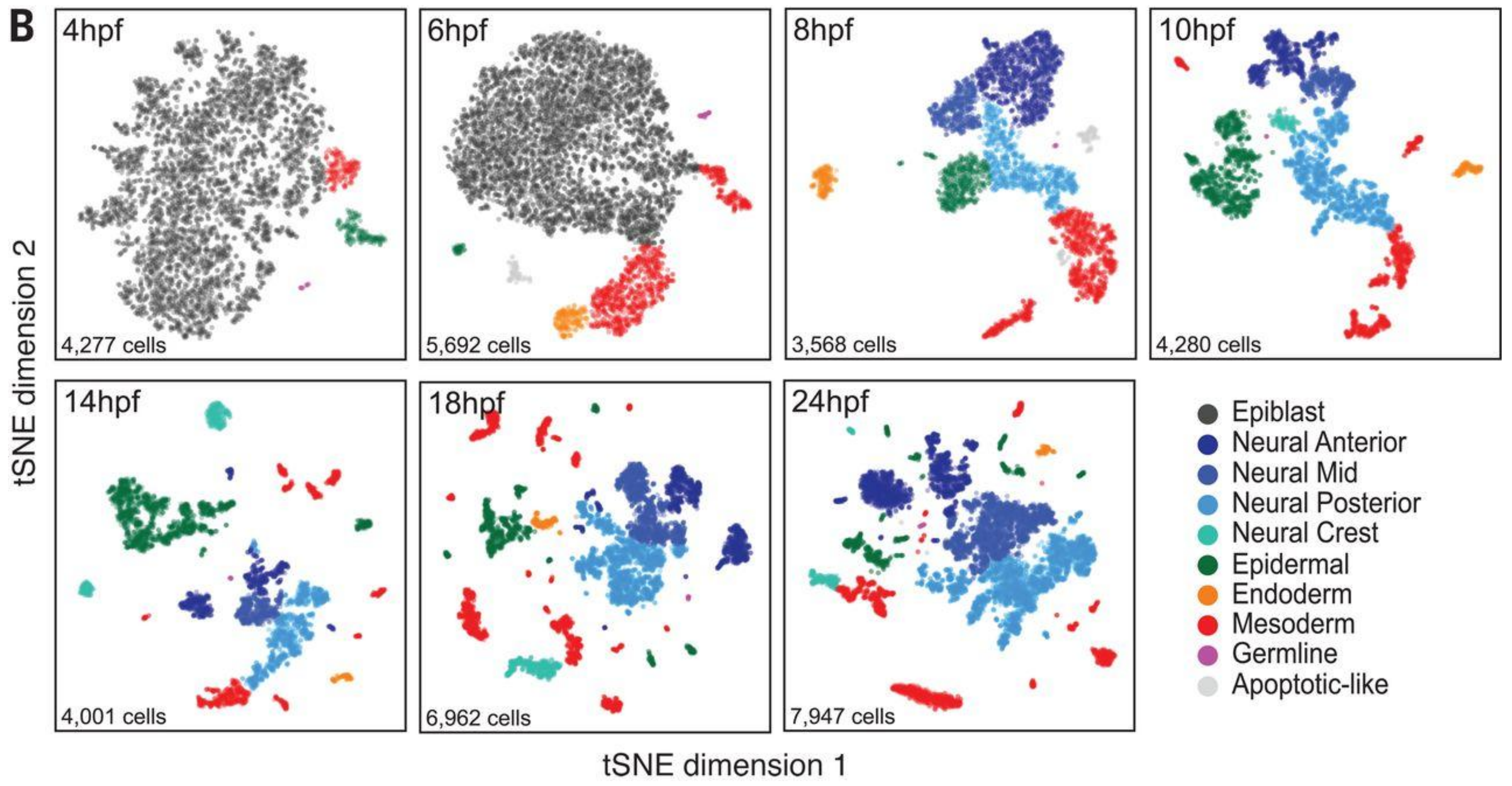

Figure 1: Examples colored by known cell types inferred from genetic markers. Figure adapted from Wagner et. al.

1. Feature Identification

- Of the 30000+ genetic features in the dataset only a small percentage contribute to the majority of variability in the dataset.
- To pre-process the data for efficient analysis, **genetic features comprising less than 1% of the variance in the data were removed**.
- Features with similar relations between samples were combined using **recursive feature agglomeration**.
- Finally, outlier examples were then eliminated using a **Local Outlier Factor algorithm with 10% contamination**.

2. Transcriptome Archetypes

- We fit the RNA transcriptomes at all seven timesteps with three low-rank models to identify transcriptome archetypes.
  - **Singular Value Decomposition (SVD)**
    Only the first 50 components were computed.
  - **Nonnegative Matrix Factorization (NMF)**
    Only the first 50 components were computed.
  - **Latent Dirichlet Allocation (LDA)**
    LDA was run multiple times with between 3 and 7 topics, and the best number of topics (minimum perplexity) was selected.

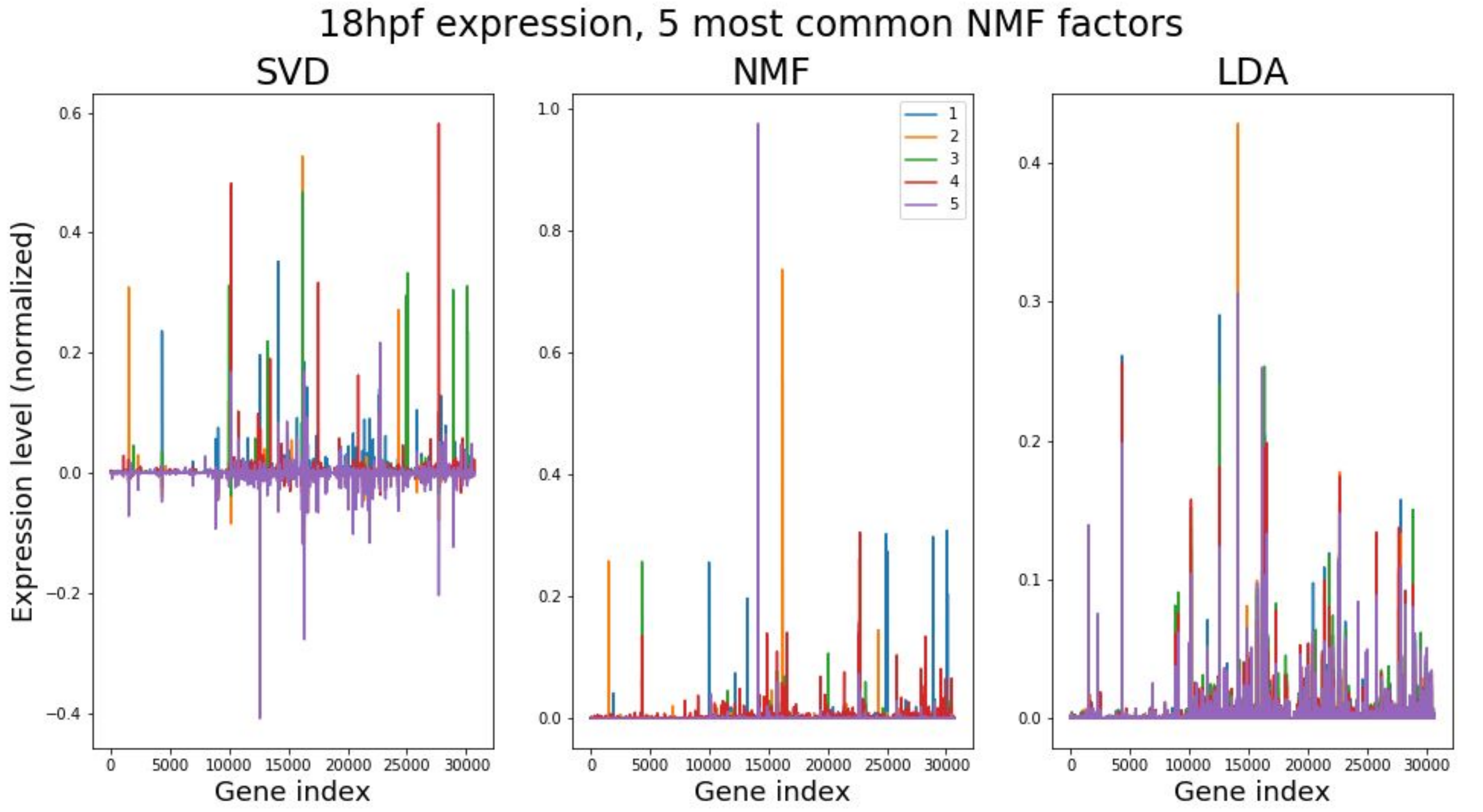

Figure 2: Transcriptome archetypes produced by each low rank model (SVD, NMF, LDA). Note that NMF and LDA produce strictly nonnegative archetypes, similar to the actual data.

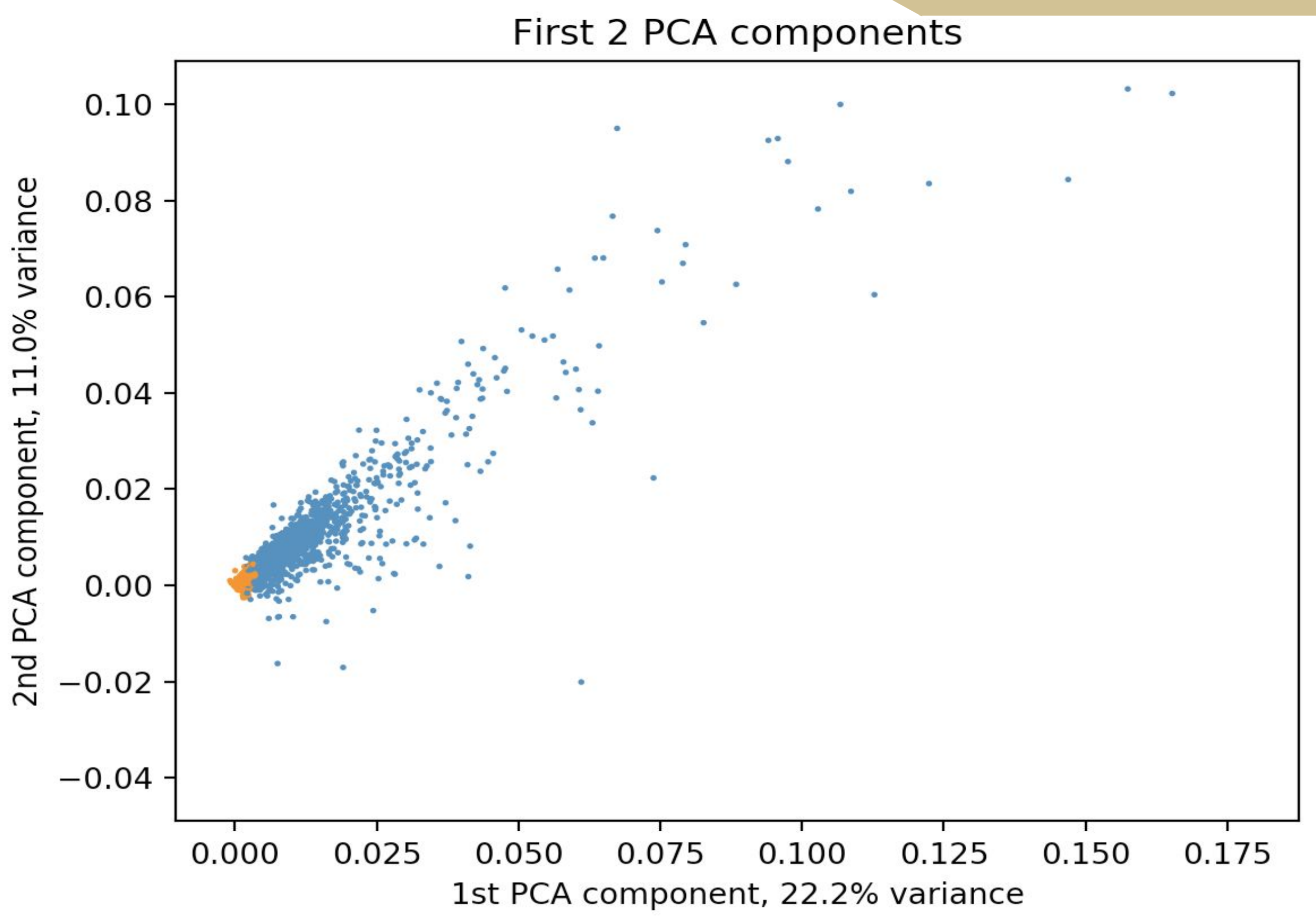

Figure 3: Features distributed by variance (first two PC). 9.7% of total features kept (blue) while low variance features (orange) were removed.

Discussion

Feature Identification

- **Variance Threshold:** may have removed valuable features.
- **LOF:** contamination factor depends on RNA-seq accuracy.

Transcriptome Archetypes

- **SVD**: unsuitable archetypes, negative expression and orthogonality have no physical interpretation.
- **NMF**: sparse, interpretable archetypes, no cost to MSE.
- **LDA**: probabilistic interpretation, at expense of sparsity.

Future Work

- Fitting a low-rank models to all time points simultaneously. Standard tensor factorizations will fail because different cells are sampled at each time point.
- Tf-idf (term frequency-inverse document frequency transformation of data).
- Exploring connection between archetypes and developmental processes.

References

1. Wagner et al. (2018). Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science*.