Transferring Annotations between Single Cell Sequencing Datasets

Xingfan Huang
Paul G. Allen School of Computer Science and Engineering and Department of Genome Sciences
Final Project for CSE546, Autumn 2018, Seattle, WA

Introduction

Cells are the main building blocks of complex life forms. Within the human body, each of the billions of cells has the same DNA, but can have vastly different functions and characteristics. A comprehensive understanding of the various types of cells would help us understand biological processes and diagnose and treat diseases.

As new experimental methods to measure distinct cellular modalities like gene expression and chromatin accessibility at single cell level are developed, a key analytic challenge is to leverage knowledge from existing datasets to accelerate biological discovery in future datasets.

In this project we explore machine learning methods that can learn a model on existing annotated datasets, apply to a new dataset and transfer annotations to the new dataset. We show that we can detect novel cell types in new datasets and transfer cell type annotations cell by cell to new datasets.

Data

We used single-cell (sc) RNA-seq and scATAC-seq datasets from adult mouse kidney and brain, retrieved from Han et al., 2018, and Cusanovich and Hill et al., 2018. We also used scRNA-seq datasets from human pancreas produced by different scRNA-seq technologies, retrieved from Stuart and Butler et al., 2018.

We preprocessed the datasets to extract high quality samples and useful features. For each test, we use two matrices, a reference and a query.

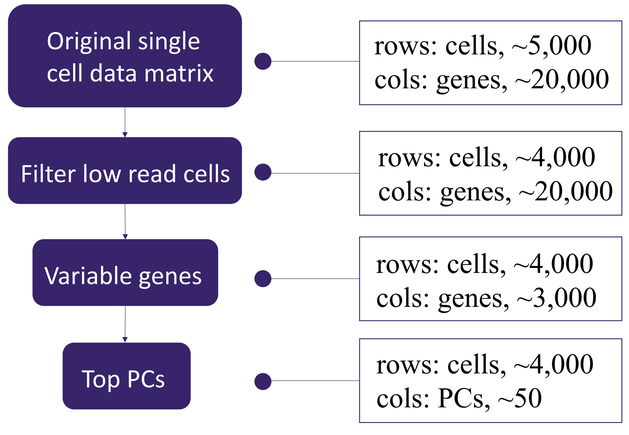


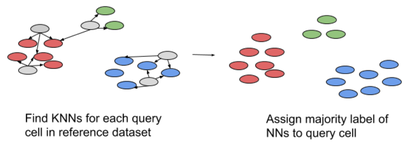
Diagram for preprocessing scATAC-seq data matrix from Cusanovich and Hill et al., 2018

Methods

Novelty Detection

In the query dataset there may be cells that are from a population that does not exist in the reference dataset. We tested one-class SVM with the RBF kernel to detect these novel cells. We reserved one cell type as "novel", and treated the remaining data as "normal". We split the "normal" data into training and test sets and use the "novel" data as test set.

Unsupervised Learning



For each sample (cell) in the query dataset, we find k-nearest neighbors in the reference dataset and transfer the most common label among the nearest reference neighbors to the query cell.

Supervised Learning

We take the reference dataset and train a multi-class classification model on the dataset. We then apply this model to the query dataset to predict the cell type of each cell, and compare the predictions with true annotations. We tested various models for classification, including logistic regression, KNN classification and XGBoost.

Results

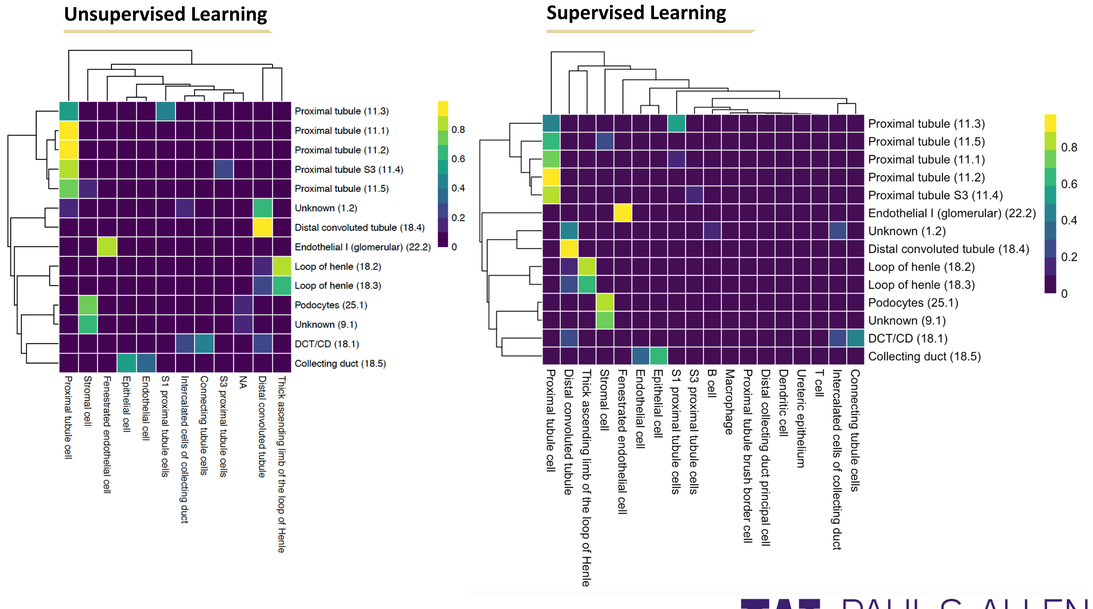
Novelty Detection: scATAC-seq Mouse Kidney Dataset

Cell Type	Collecting duct DCT/CD	Distal convoluted tubule	Endothelial I (glomerular)	Loop of henle	Podocytes	Proximal tubule	Proximal tubule S3	Unknown
Cell count	117	434	256	192	673	317	2128	588	116
FP rate	17.74%	18.45%	20.48%	16.52%	16.62%	13.65%	24.12%	19.13%	17.43%
FN rate	14.53%	17.97%	23.05%	0	14.71%	1.26%	4.46%	7.14%	43.10%

Cell type annotation transfer: Fluidigm C1 ←→ SMART-Seq2, Human Pancreas Dataset

Reference Dataset	Cell Count	Accuracy (Unsupervised Learning)	Accuracy (Stuart and Butler et al., 2018)
Fluidigm C1	638	80%	97%
SMART-Seq2	2394	95%	97%

Cell type annotation transfer: scRNA-seq→ scATAC-seq, Mouse Kidney Dataset



Discussion

- Our methods work with single cell data from different organisms, tissues, cell modalities and RNA-seq technologies.
- Our methods could have immediate practical use because usually annotation happens on a cluster, and we can accurately transfer annotations on whole clusters by taking majority annotations in the cluster.
- A limitation is that we are unable to extract biological meaning from our models, because our features are PCs instead of genes. We will further development our methods to improve performance in general, and to provide more interpretable results.

References

- Cusanovich, D.A. *, Hill, A.J. *, Aghamirzaie, D., Daza, R.M., Pliner, H.A., Berletch, J.B., Filippova, G.N., Huang, X., Christiansen, L., DeWitt, W.S., Lee, C., Regalado, S.G., Read, David F., Steemers, Frank J., Disteche, C.M., Trapnell, C., Shendure, J. (2018). A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell* 174(5), 1309-1324.e18.
- Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., et al. (2018). Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* 172, 1091–1107.
- Pliner H.A. *, Packer J.S. *, McFaline-Figueroa J.L., Cusanovich D.A., Daza R.M., Aghamirzaie D., Srivatsan S., Qiu X., Jackson D., Minkina A., Adey A.C., Steemers F.J., Shendure J. **, Trapnell C. ** (2018). Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. {\it Mol Cell} 71(5):858-871.e8. doi: 10.1016/j.molcel.2018.06.044.
- Stuart, T. *, Butler, A. *, Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Stoeckius, M., Smibert, P., Satija, R. (2018). Comprehensive integration of single cell data. *bioRxiv*. doi: http://dx.doi.org/10.1101/460147.

PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING

Genome Sciences UNIVERSITY OF WASHINGTON