Mapping single-cell data to reference atlases by transfer learning

Shang Gao, Ph.D candidate
Advisor: Dr. Yang Dai, Dr. Jalees Rehman
CBQB journal club, Feb. 7th
Introduction

• The authors developed the deep learning strategy scArches for mapping query datasets onto reference datasets for the purpose of data integration and the annotation or identification of cell types.
Motivation

Sample 1

Sample 2

Sample 3

Downstream analysis
Motivation

Sample 1

Sample 2

Sample 3

Downstream analysis
Motivation

Sample 1
Sample 2
Sample 3

Downstream analysis

UMAP of 3 samples
Color: Cell types

UMAP of 3 samples
Color: Batches

UMAP of 3 samples
Color: Seurat clusters
Motivation

Integration tools should be applied to eliminate the batch effect before the downstream analysis.
Motivation

• Existing approaches like the Seurat platform allow for integration of data but require that users run the complete pipeline on new datasets which requires excessive computational resources and time.

• Unlike the existing tools, scArches uses a transfer learning approach to transfer the knowledge from pre-trained model to user specific data which not only reduces run time for integration but also leverages prior knowledge included in the preference dataset for the identification of cell types in the user-specific data.
scArches Overview

Transfer learning, Architecture surgery
Deep generative model

- scArches integrates the reference data by deep generative model (DGM). It provided multiple variants of DGM including trVAE, scVI, scANVI, totalVI.
scArches trVAE

- trVAE builds upon conditional variational autoencoder (CVAE).
Autoencoder

\[ X \in \mathbb{R}^n \quad \Rightarrow \quad X' \in \mathbb{R}^n \]

\[ Z \in \mathbb{R}^l, l \ll n \]

\[ \text{Loss} = \frac{1}{M} \sum (X - X')^2 \]

- Dimension reduction methods.
- If the autoencoder has no hidden layer in Encoder and Decoder and no non-linearity activation on neurons, autoencoder is identity to PCA.
- Adding multiple hidden layers in Encoder and Decoder and the non-linearity activation on neuron will build a deep autoencoder.
Variational autoencoder

- Variational inference
  Find a variational distribution $Q(Z|X)$ which minimize,

$$D = KL(Q(Z|X)||P(Z|X))$$

$$D = -E_{Z \sim Q} [\log P(X|Z)] + KL(Q(Z|X)||P(Z)) + \log P(X)$$

Minimize $-E_{Z \sim Q} [\log P(X|Z)] + KL(Q(Z|X)||P(Z))$
Variational autoencoder

Minimize $-E_{Z \sim Q} [\log P (X | Z)] + KL(Q(Z|X)||P(Z))$

$Q(Z|X)$ is the Encoder; $P(X|Z)$ is the Decoder.

$KL(Q(Z|X)||P(Z))$ is the difference between latent space distribution $Q(Z|X)$ and prior distribution of $Z$ $P(Z)$.

$P(X|Z)$ is equivalent to $P(X|X')$,
- If the $X \sim$ Normal, then $P(X|X')$ can be written as the form like $\exp(-|X - X'|^2)$. $MSE = \frac{1}{M} \sum (X - X')^2$.
- If the $X \sim$ Bernoulli, $P(X|X')$ is equivalent to cross entropy of $X$ and $X'$.
- If we assume $X$ follow other distributions like negative binomial distribution, the output layer will be the parameter of the NB distribution instead of reconstructed data. $-E_{Z \sim Q} [\log P (X | Z)]$ is the negative loglikelihood of NB.
Variational autoencoder

Minimize \(-E_{Z \sim Q} [\log P (X|Z)] + KL(Q(Z|X)||P(Z))\)

\(Q(Z|X)\) is the Encoder; \(P(X|Z)\) is the Decoder.

\(KL(Q(Z|X)||P(Z))\) is the difference between latent space distribution \(Q(Z|X)\) and prior distribution of \(Z\) \(P(Z)\).

\(P(X|Z)\) is equivalent to \(P(X|X')\),
- If the \(X\)~ Normal, then \(P(X|X')\) can be written as the form like \(\exp(-|X - X'|^2)\). \(MSE = \frac{1}{M} \sum (X - X')^2\).
- If the \(X\)~ Bernoulli, \(P(X|X')\) is equivalent to cross entropy of \(X\) and \(X'\).
- If we assume \(X\) follow other distributions like negative binomial distribution, the output layer will be the parameter of the NB distribution instead of reconstructed data. \(-E_{Z \sim Q} [\log P (X|Z)]\) is the negative loglikelihood of NB.
VAE on healthy lung data

- UMAP on PCA
  - Color: Batches
- UMAP on VAE
  - Color: Batches
- UMAP on PCA
  - Color: Cell types
- UMAP on VAE
  - Color: Cell types
Conditional variational autoencoder (CVAE)

Minimize $-E_{Z \sim Q}[\log P(X|Z, S)] + KL(Q(Z|X, S)||P(Z|S))$
Loss of conditional variational autoencoder

Minimize $-E_{z \sim Q} [\log P(X|Z,S)] + KL(Q(Z|X,S)||P(Z|S)) - MMD(y_i, y_j)$

MMD is the Maximum Mean Discrepancy. It is the similarity of two distributions.

y is the values of the first hidden layer of Decoder. i and j are the labels of samples.
scVI

Zero inflated negative binomial distribution
Results

UMAP on trVAE
Color: Cell types

UMAP on trVAE
Color: Batches

UMAP on scVI
Color: Cell types

UMAP on scVI
Color: Batches
Transfer learning
Results: transfer learning of trVAE

- trVAE on ref data
  - Color: Batches

- Mapping of query data
  - Color: Batches

- trVAE on ref data
  - Color: Cell types

- Mapping of query data
  - Color: Cell types
Results

TL and architecture surgery allow fast and accurate reference mapping.
Results

scArches remains robust even when cell fractions in the reference data are varied.
Results

scArches enables efficient reference mapping compared existing data-integration methods.
Results

scArches discovers cell types in query data, even when they such cell types are removed from reference data.
Results

scArches resolves severity in COVID-19 query data mapped to a healthy reference and reveals emergent cell states.
Summary

• Deep generative model

“Things may shift their forms ten thousand times, but the principle remain unchanged.” *Xunzi, third century BC*

The goal of these different DGMs are trying to learn a low dimensional space of data by adding constraints on the model and modifying the learning target of the model.

Single cell studies are trying to understand the similarities between cells. By applying DGMs on single cell data, we can define the cell-cell distance on this new space. This will provide us information for clustering and trajectory building analysis.

• Transfer learning

Large number of single cell data are generated. But the analysis tools which making use of these big data is a few.
Unanswered Questions and new directions

• Learn biological meaningful latent variables

• distinguish the technical batch effect from biological effect

• Integrate more prior biological knowledge into the deep generative model
Acknowledgements

Rehman lab:  
Xinge Wang  
Mark Sanborn  
Priyanka Gajwani  
Arundhati Jana  
Young-Mee Kim  
Li Wang  
Joseph Leasure  
Shubhi Srivastava  
Zijing Ye  
Lissette Magana

Dai lab:  
Yongchao Huang  
Mehrdad Zandigohar  

Mentors:  
Yang Dai  
Jalees Rehman