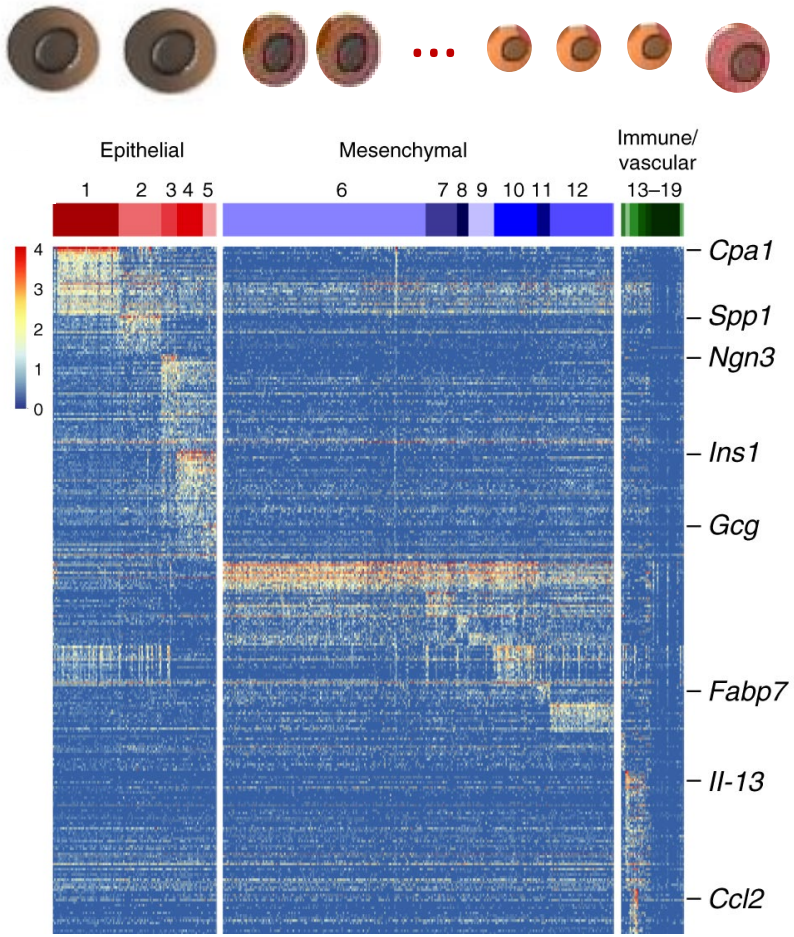
Single Cell RNA-sequencing:

visualization, alignment, denoising and transfer learning

Jingshu Wang
University of Chicago
ISMB, July 21st, 2019

Part I
Visualization and dimension reduction

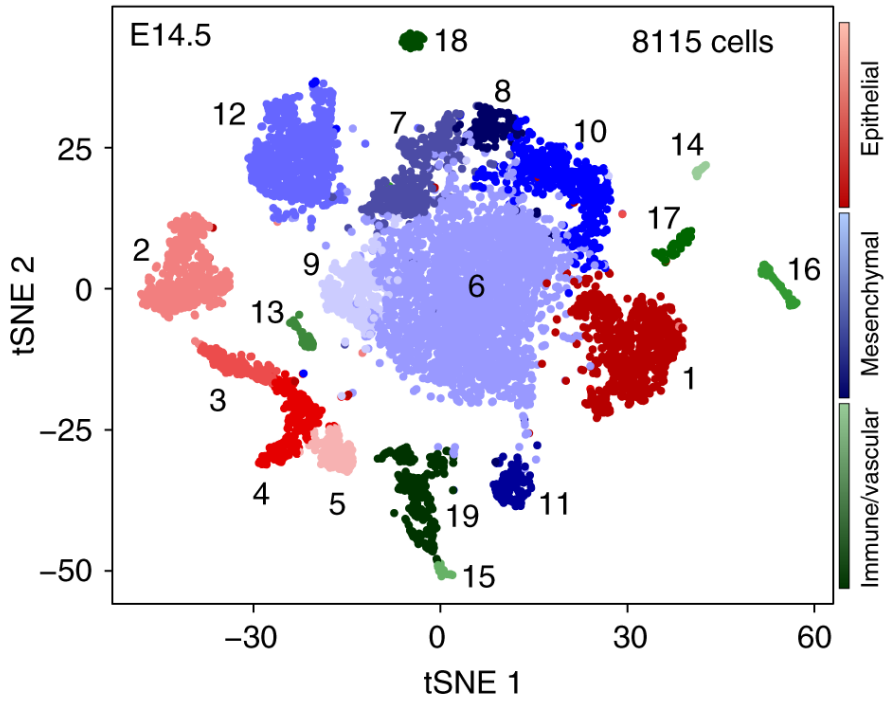
Single cell RNA-seq data



Genes

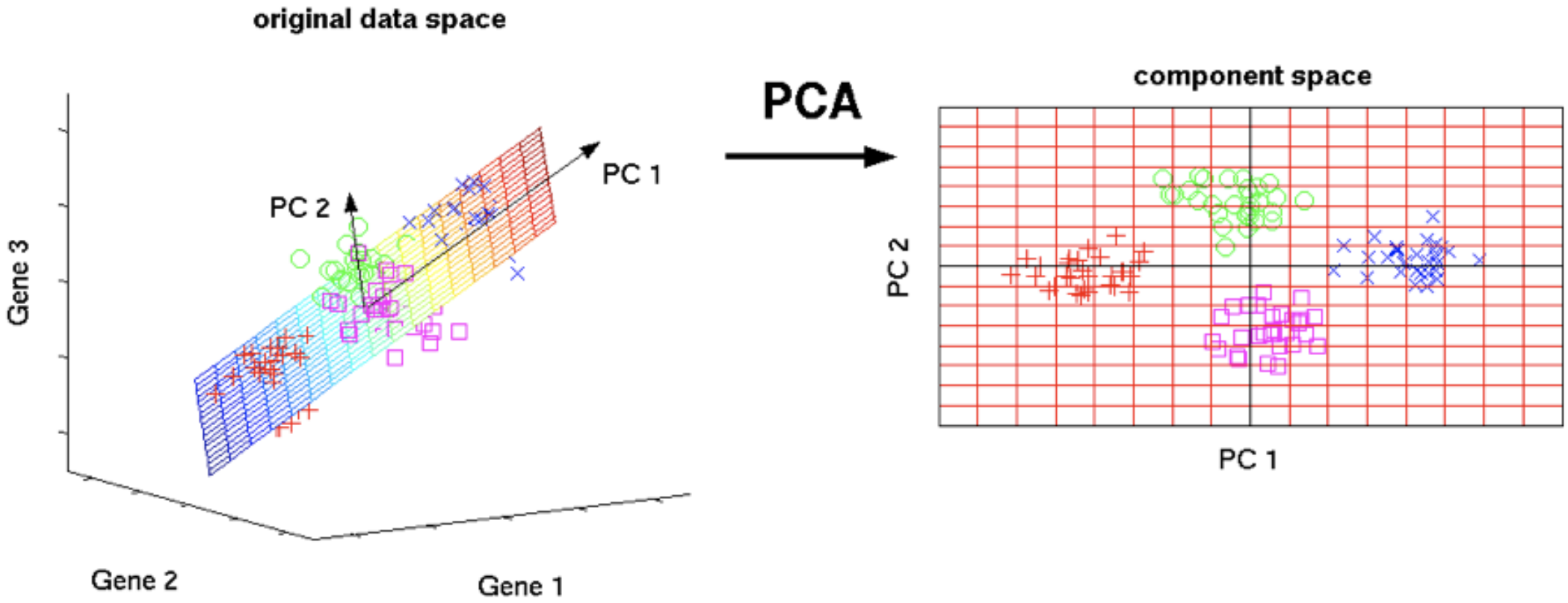
Cells

(Colors are determined by separate clustering methods!!)



Lineage dynamics of murine pancreatic development at single-cell resolution, Byrnes et. al. *Nature Comm.* 2018

Linear dimension reduction: PCA



Interpretation: Find direction of the data that has the largest variation

Non-linear dimension reduction: t-SNE & UMAP

- t-SNE: t-Distributed Stochastic Neighbor Embedding

  Paper: https://lvdmaaten.github.io/publications/papers/JMLR_2008.pdf
  Presentation: https://www.youtube.com/watch?v=RJVL80Gg3lA&list=UUtXKDgv1AVoG88PLl8nGXmw

- UMAP: Uniform Manifold Approximation and Projection

  Paper: https://arxiv.org/pdf/1802.03426.pdf
  Benchmark paper on scRNA-seq: https://www.nature.com/articles/nbt.4314
  Presentation: https://www.youtube.com/watch?v=nq6iPZVUxZU

The idea of t-SNE

SNE (stochastic neighbor embedding)

- Preserve the similarity of high-dimensional points in low-dimensional points
- Measure similarity by Gaussian density

Determined by perplexity

Original space:

$$p_{ij} = \frac{\exp\left(-\|x_i - x_j\|^2/2\sigma^2\right)}{\sum_{k\neq l}\exp\left(-\|x_k - x_l\|^2/2\sigma^2\right)}$$

Low-dimensional space:

$$q_{ij} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k\neq l}\exp\left(-\|y_k - y_l\|^2\right)}$$

Find $\{y_i\}$ to minimize:

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

The idea of t-SNE

SNE (stochastic neighbor embedding)

- Preserve the similarity of high-dimensional points in low-dimensional points
- Measure similarity by Gaussian density

Determined by perplexity

Original space:

$$p_{ij} = \frac{\exp\left(-\|x_i - x_j\|^2/2\sigma^2\right)}{\sum_{k\neq l}\exp\left(-\|x_k - x_l\|^2/2\sigma^2\right)}$$

Low-dimensional space:

$$q_{ij} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k\neq l}\exp\left(-\|y_k - y_l\|^2\right)}$$

Find $\{y_i\}$ to minimize:

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

t-SNE (t-distribution density)

$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k\neq l}\left(1 + \|y_k - y_l\|^2\right)^{-1}}$$

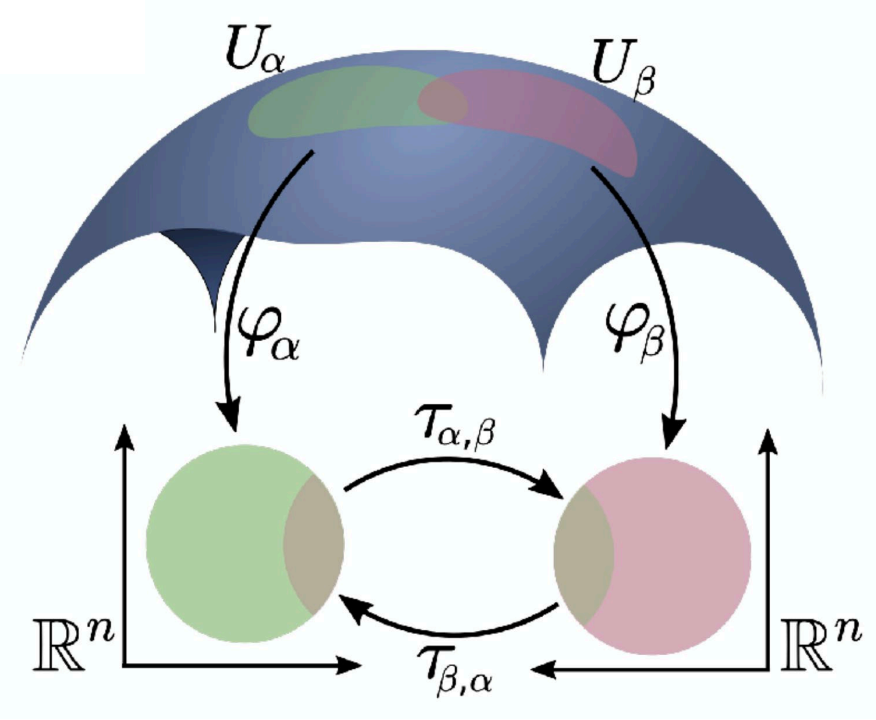
Represent high-dimensional points better and keep moderately far-away points not too close

The (very high-level) idea of UMAP

- Construct a weighted k-nearest neighbor graph
  - Assume that the data points uniformly lie on a low-dimensional manifold
  - Define local distance by k-nearest neighbors

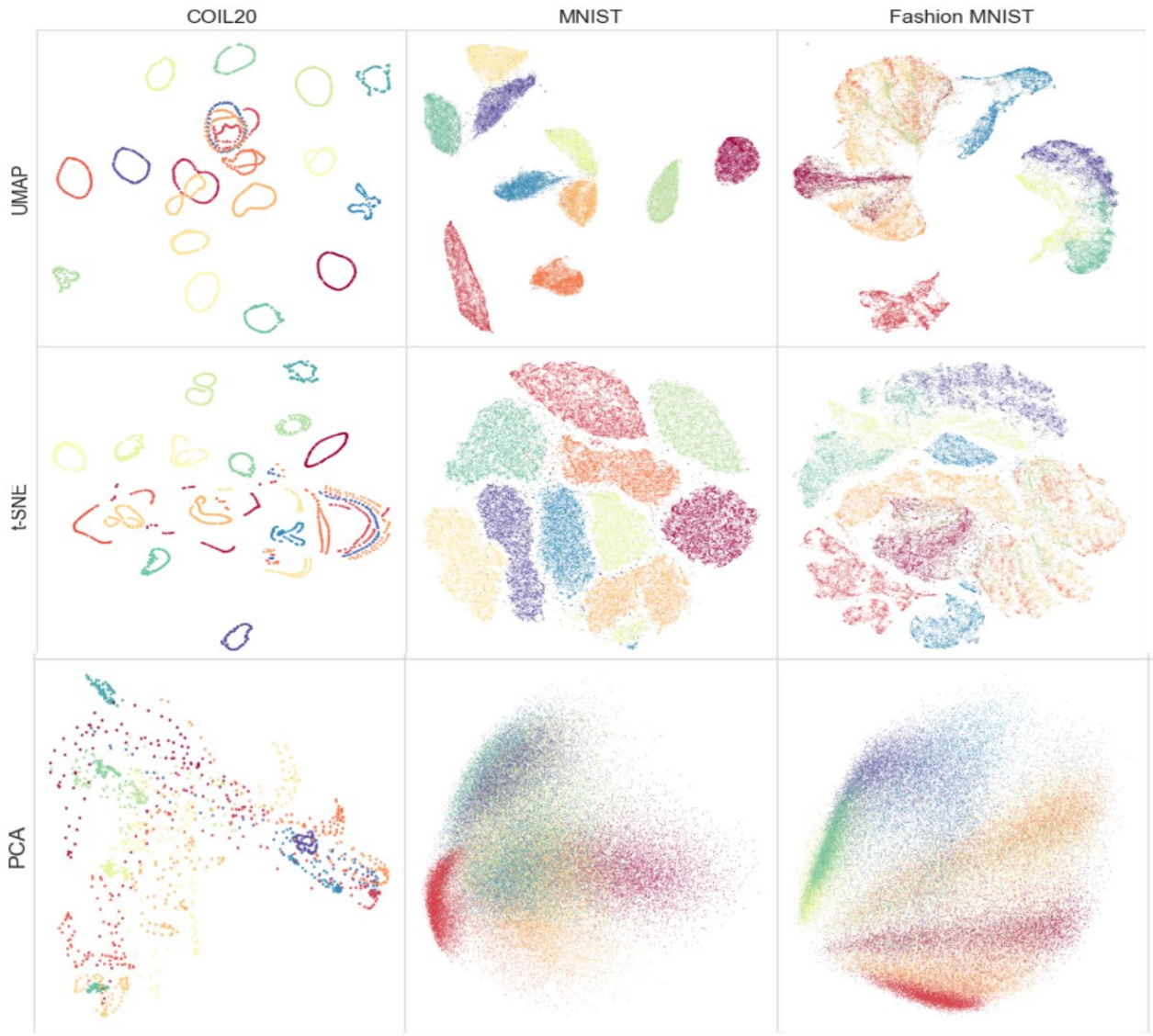
- Represent the manifold by low-dimensional points
  - Based on the theory of fuzzy topological representation
  - Similarity in the low-dimensional space defined as

$$w_{ij} = \left(1 + a \left\| y_i - y_j \right\|_2^{2b}\right)^{-1}$$



https://www.youtube.com/watch?v=nq6iPZVUxZU

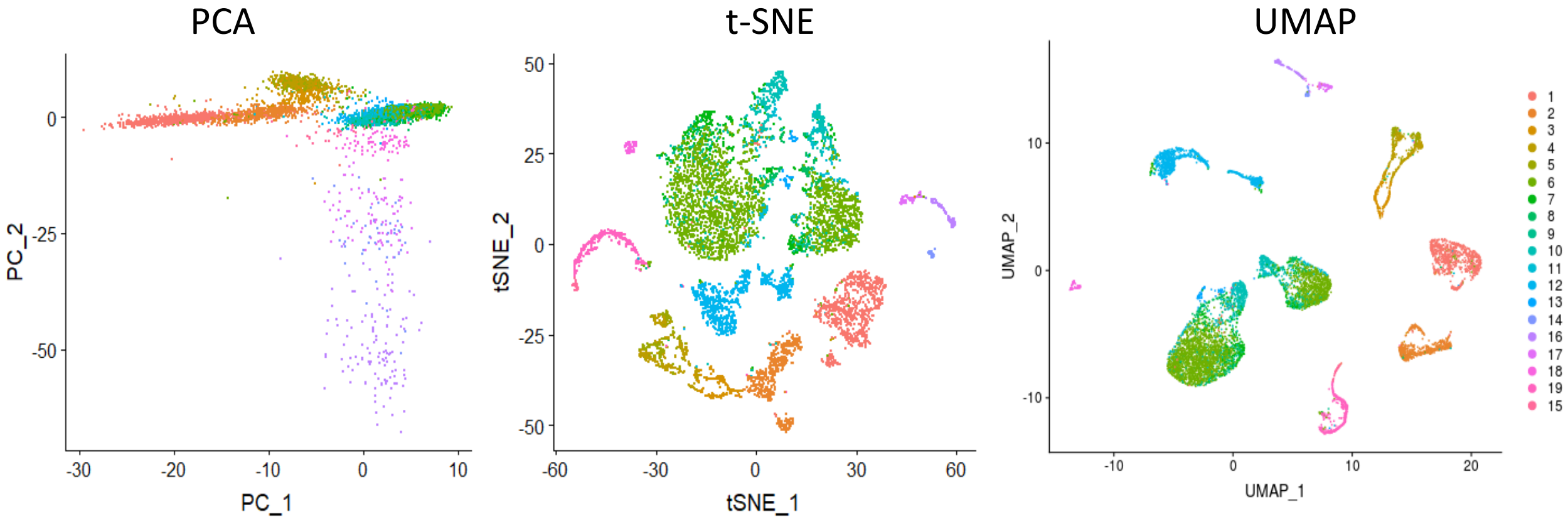
Compare PCA, t-SNE, UMAP



- PCA: keep global distance
- T-SNE: focus on local distance
- UMAP: focus on local distance, but may keep more global distance features

https://arxiv.org/pdf/1802.03426.pdf

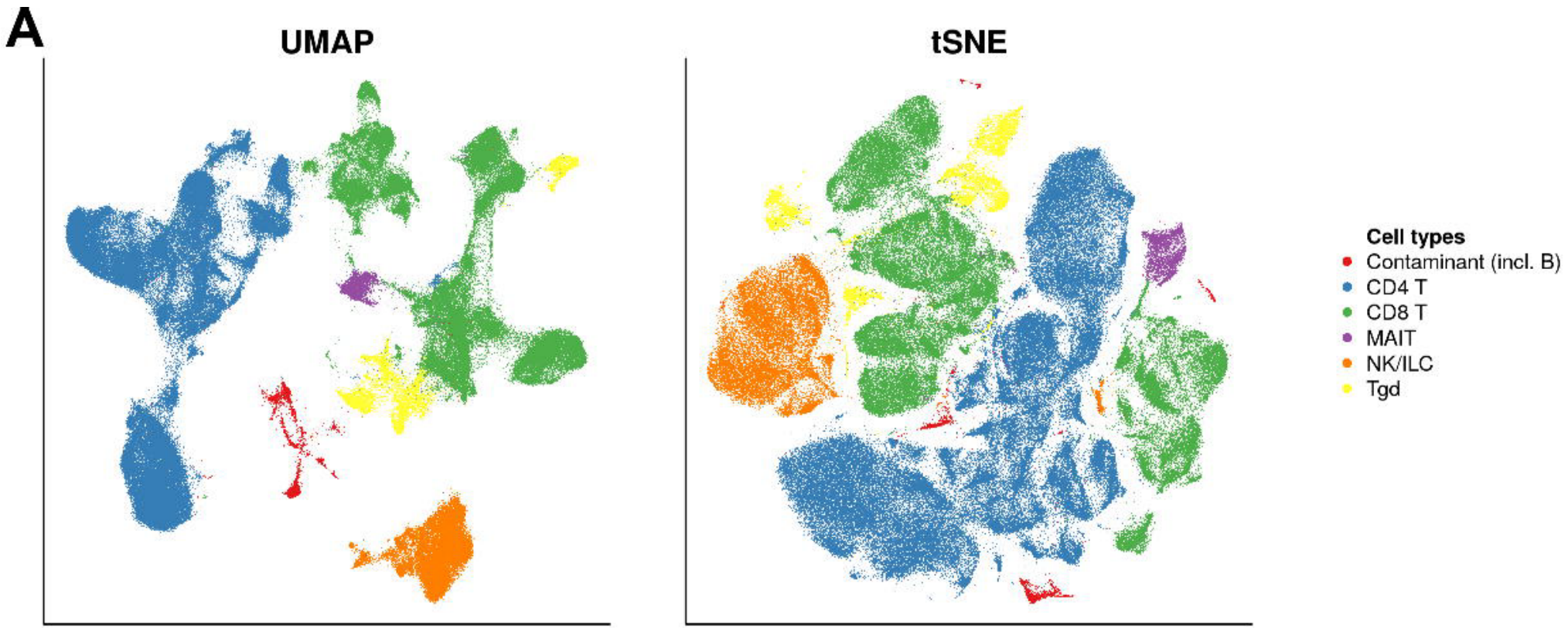
Visualize scRNA-seq using PCA, t-SNE, UMAP



Data from paper: Lineage dynamics of murine pancreatic development at single-cell resolution, Byrnes et. al. *Nature Comm.* 2018

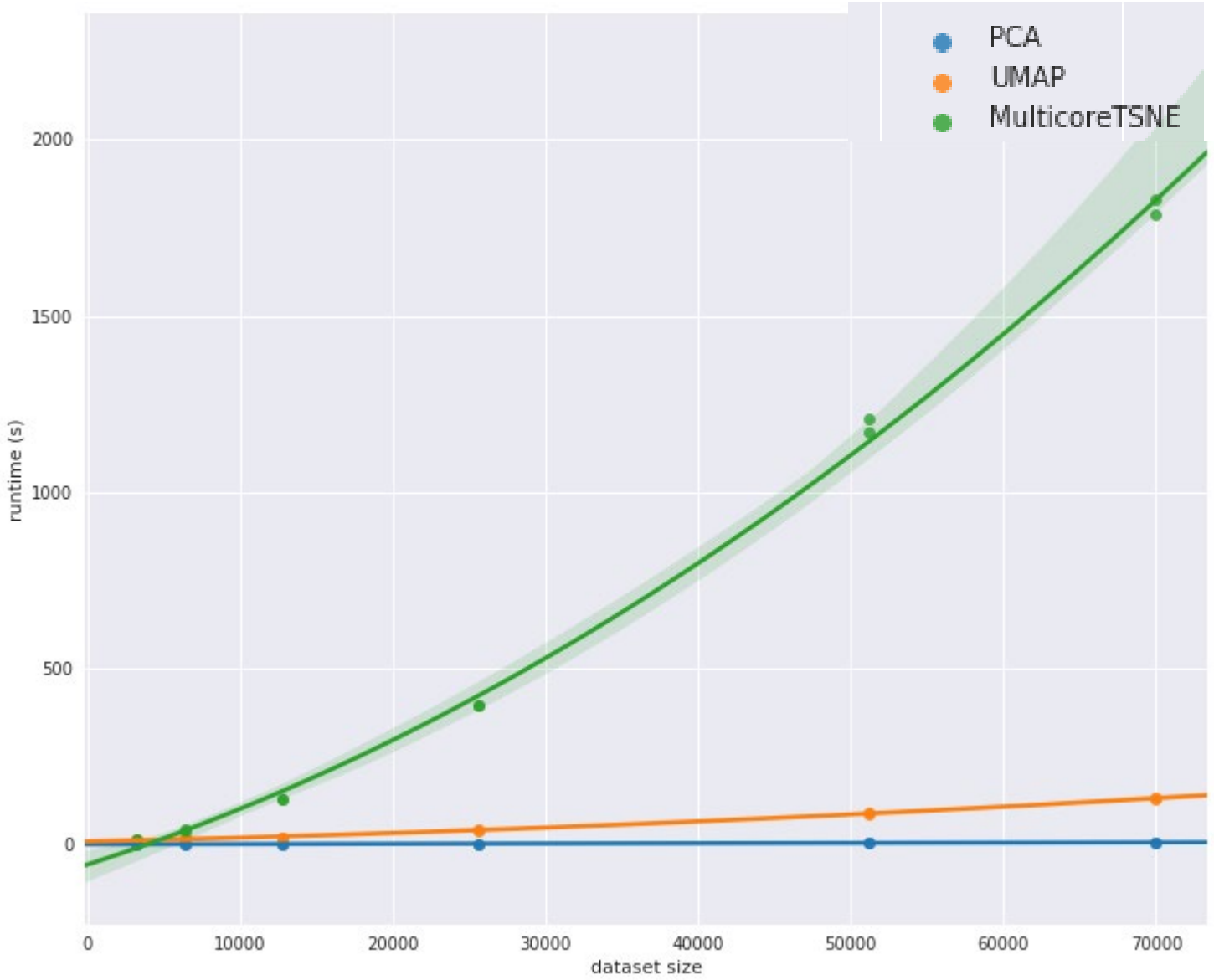
Analysis pipeline see Seurat tutorial: https://satijalab.org/seurat/v3.0/pbmc3k_tutorial.html

Compare tSNE and UMAP



Dimensionality reduction for visualizing single-cell data using UMAP, Becht et. al., *Nature Biotech*, 2018

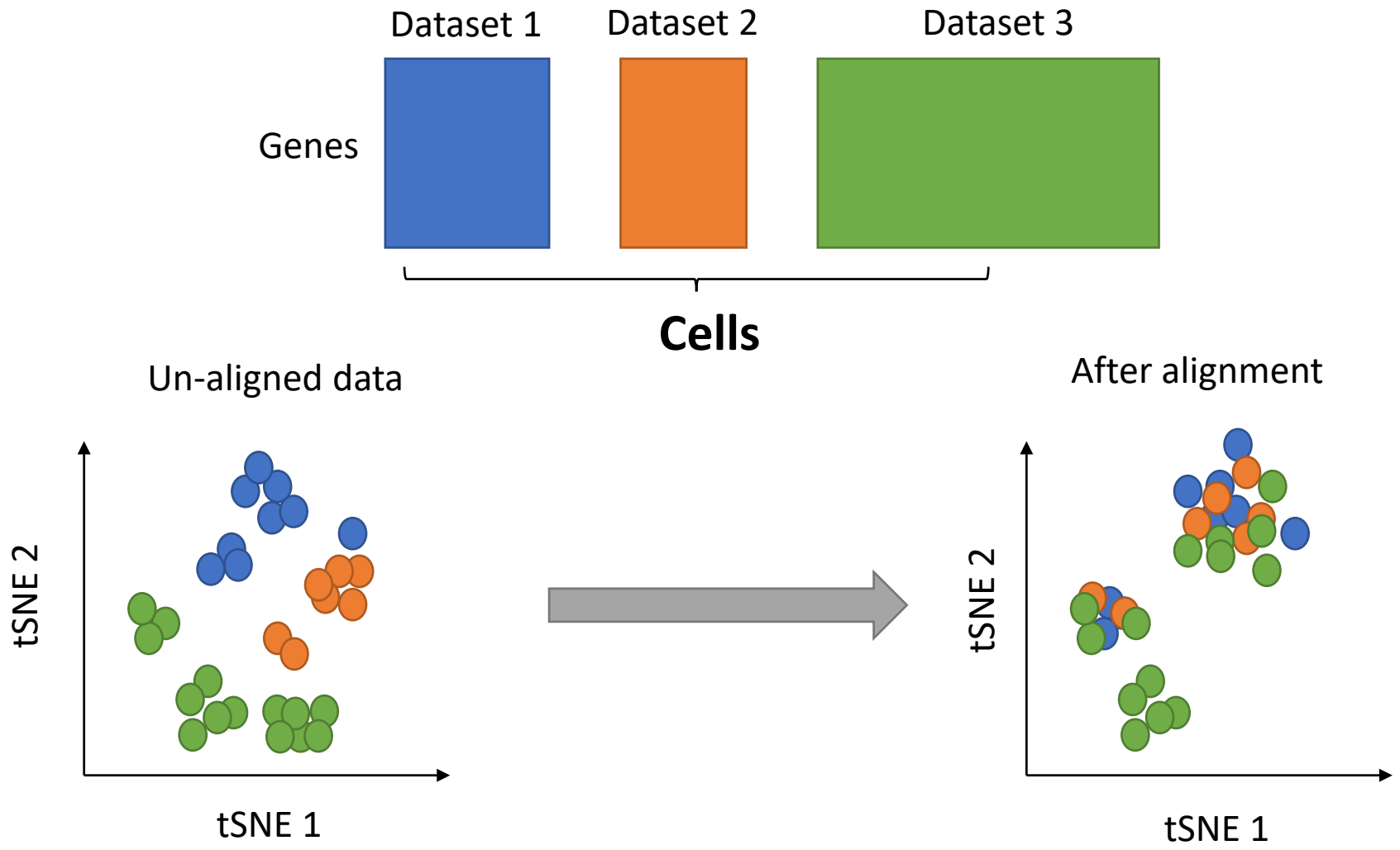
Running time comparison



https://umap-learn.readthedocs.io/en/latest/benchmarking.html

Part II
Data Integration/alignment

What is data integration/alignment?



Un-alignment between datasets

- Biological differences:
  - Different cell population (tissue, individual, species)
  - Different cell types

- Technical differences:
  - batch effects
  - different sequencing depth

Integration/Alignment = batch correction?

- Jointly analyze of multiple datasets
  - Remove batch effects
  - Remove unwanted/not interesting biological differences

    ‘uninteresting’ differences between individuals, species

- Confounding between batches and unknown cell types

Current methods

Methods	Group	Publication	Year	Can handle Multiple datasets	Align Multi-modal data	Feature
MNNcorrect	J.C. Marioni	Nature Biotech	2018	No	No	The first using KNN
MultiCCA	R. Satijia	Nature Biotech	2018	Yes	No	The first using CCA
MultiCCA V2	R. Satijia	Cell	2019	Yes	Yes	Combine CCA and KNN
LIGER	E. Macosko	BioRxiv	2018	Yes	Yes	Clear model-based CCA
Harmony	S. Raychaudhuri	BioRxiv	2018	Yes	No	Fast
Scanorama	B. Berger	BioRxiv	2018	Yes	No	Fast
BBKNN	S.A. Teichmann	BioRxiv	2018	Yes	No	Fast
Conos	P.V. Kharchenko	BioRxiv	2018	Yes	No	
Scmap	M. Hemberg	NM	2018	No	No	align target to reference
scVI	N. Yosef	NM	2018	Yes	No	Deep learning

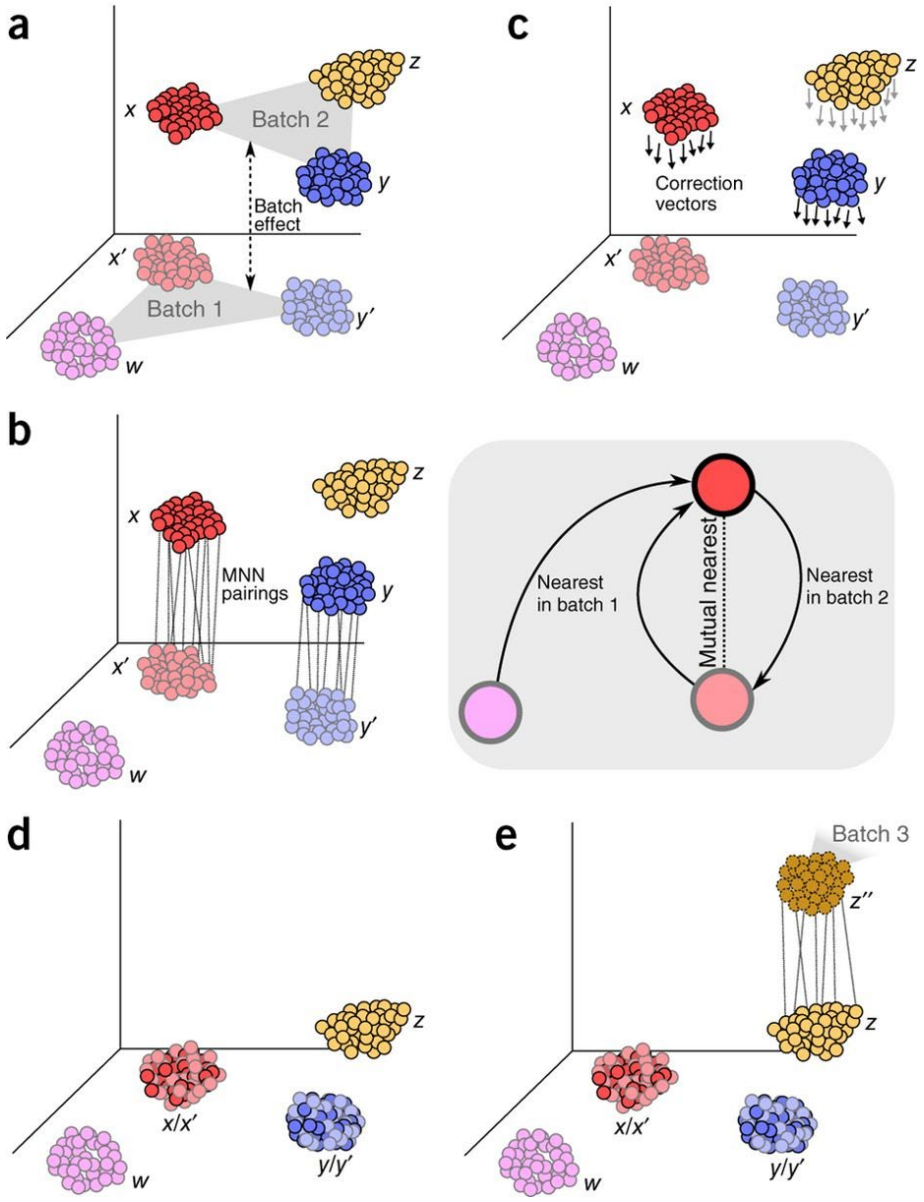
MNN correct

- Steps:

a) Measure cell similarity (Euclidean distance)

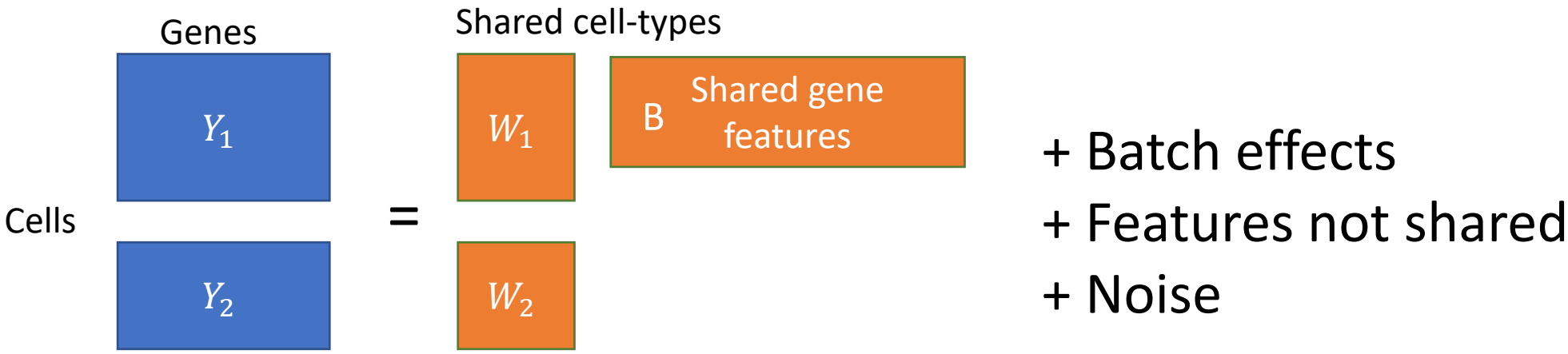
b) Identify KNN in the other batch
Find paired cells from two batches

c) Calculate pair-specific and cell-specific correction vector using Kernel methods



Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors, Hadhverdi L. et. al., *Nature Biotech*, 2018

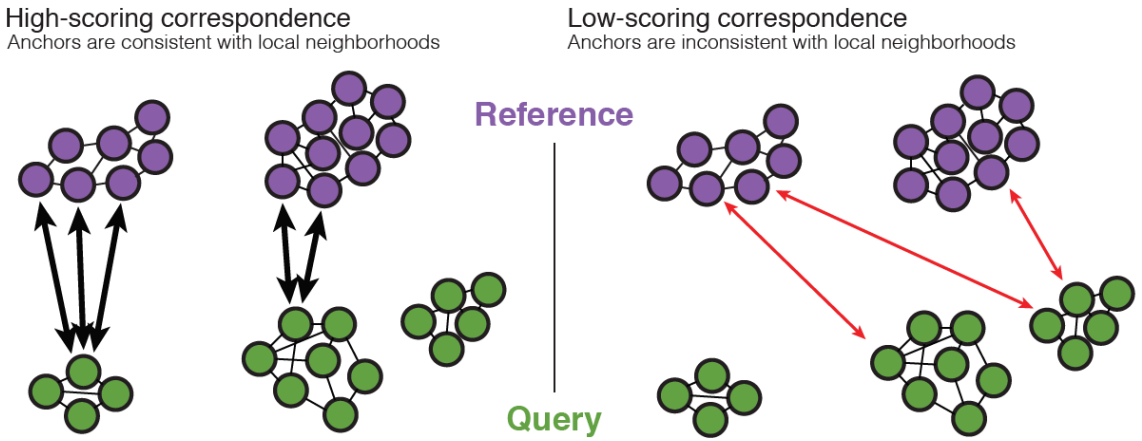
The idea of CCA (MultiCCA v1)



- CCA: Calculate the shared gene feature space B
- Align the cells on the projected space W_1 and W_2

Integrating single-cell transcriptomic data across different conditions, technologies, and species, Butler A. et. al., *Nature Biotech*, 2018

MultiCCA (v2)

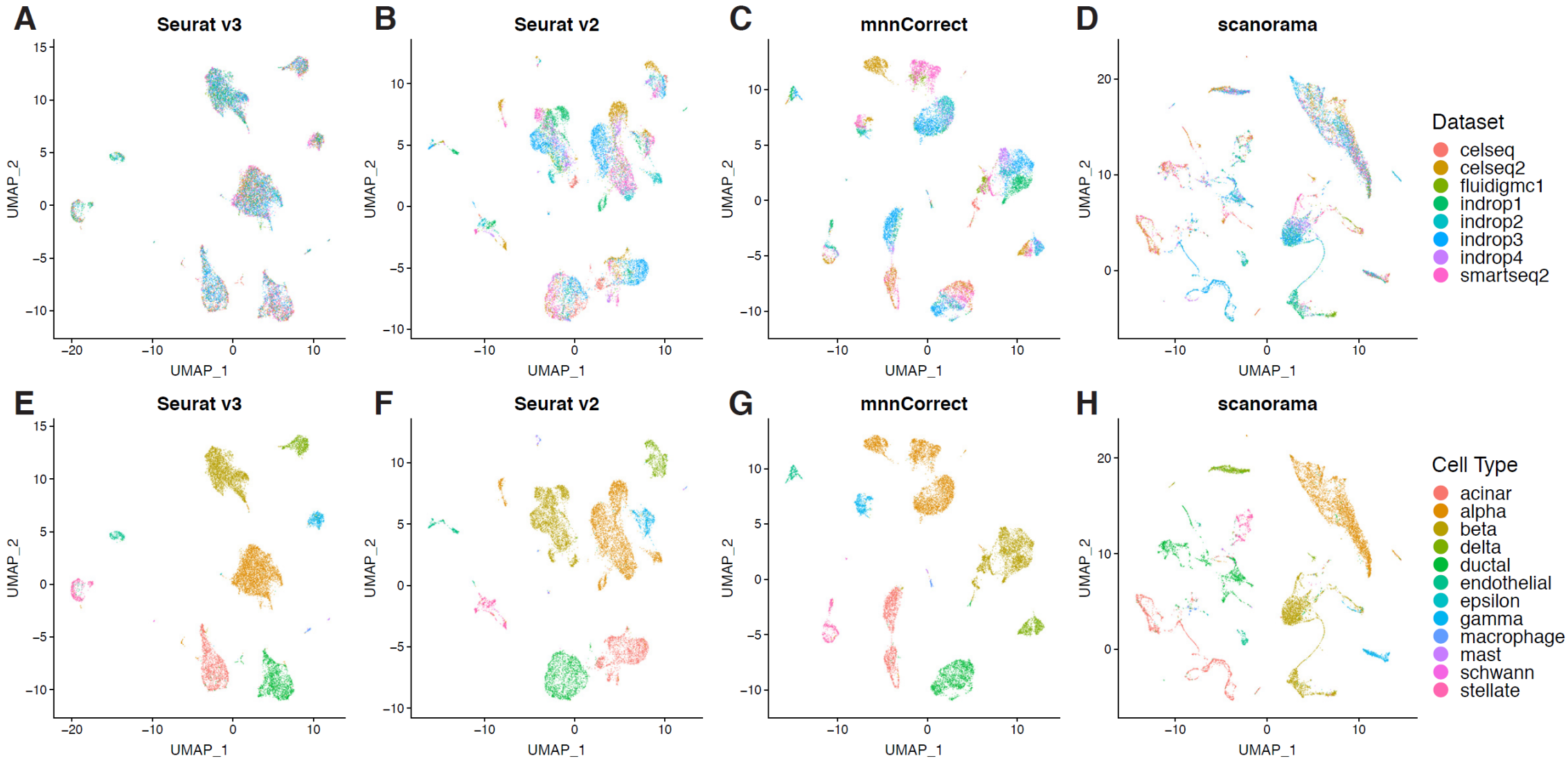


- A hybrid of multiCCA (v1) and MNNcorrect

- Steps
  - CCA
  - Identify anchor cells using MNN
    - Give each cell an anchor score
      - Check MNN also in the original space
      - **Anchors scoring**: find consistency of KNNs within each dataset and with other datasets
      - **Anchor weighting W**: a matrix of anchors by cells in Y_2
      - **Alignment**: $\hat{Y}_2 = Y_2 + (Y_{1,\mathrm{A}} - Y_{2,A})W$
      - Multiple datasets: align sequentially
  - Label transfer and feature imputation

Comprehensive Integration of Single-Cell Data, Stuart T. et. al., *Cell*, 2019

MultiCCA (V2)



Comprehensive Integration of Single-Cell Data, Stuart T. et. al., *Cell*, 2019

Part III
Denoising

How noisy is scRNA-seq?

FISH experiment setup

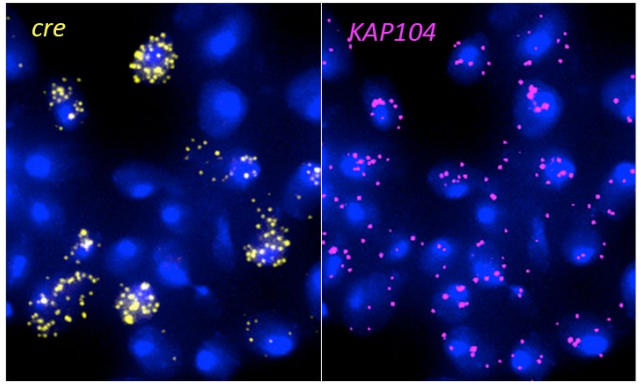


Photo courtesy of Anne Dodson and Professor Jasper Rine

Single cell RNA sequencing (Drop-seq)

~8,000 cells
~12,000 genes

Melanoma Cell Line

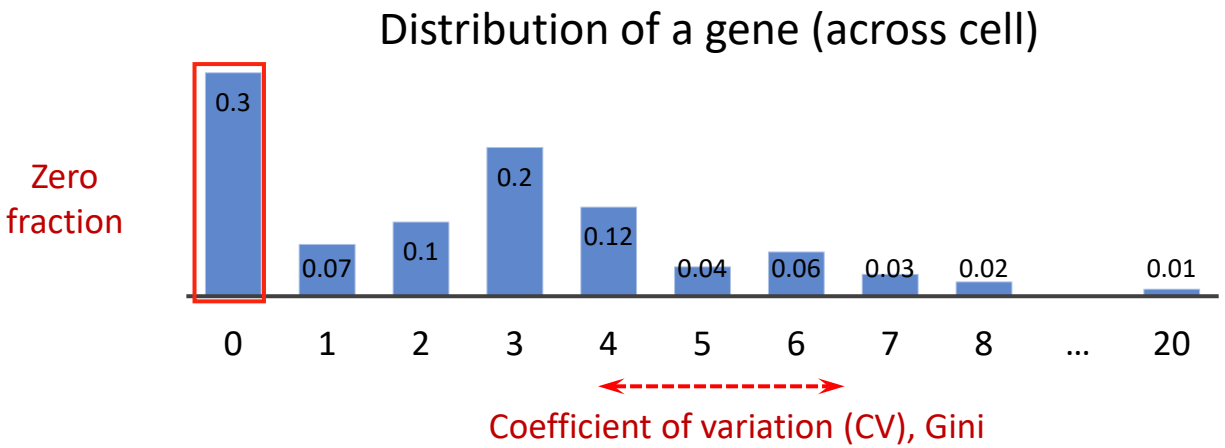
Fluorescence in-situ hybridization (RNA FISH)

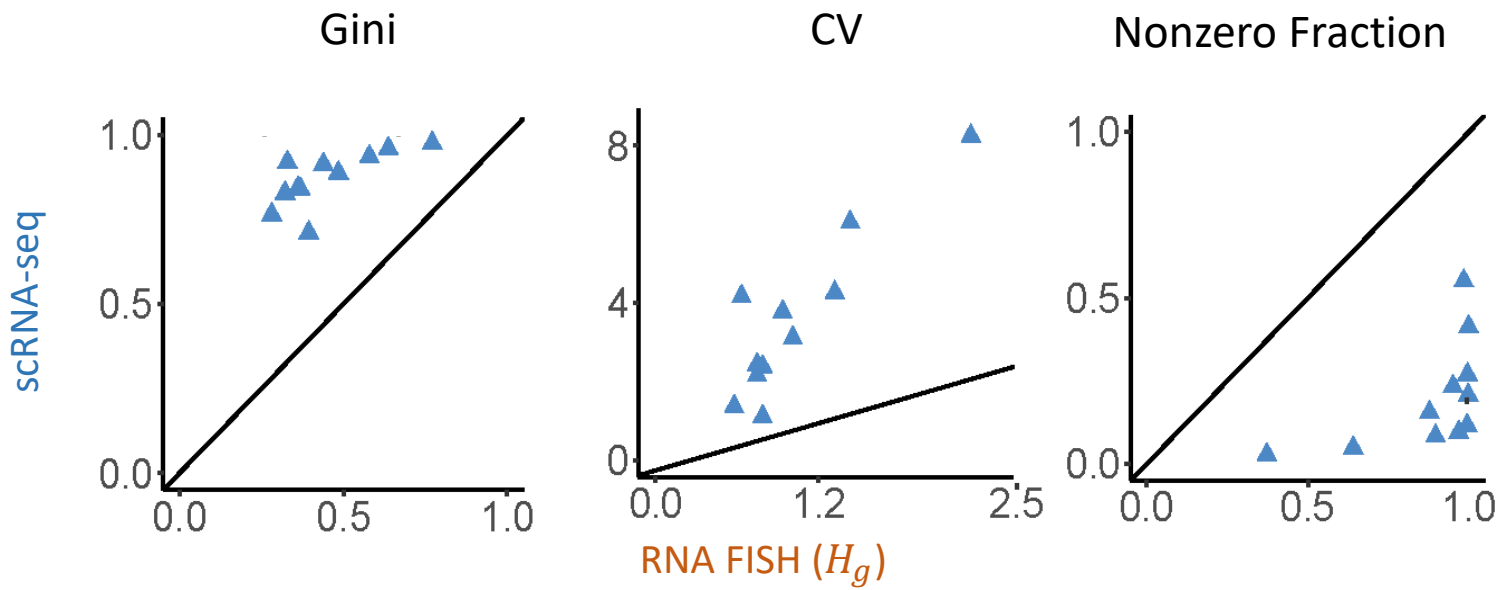
~80,000 cells
26 genes

11 genes overlap between FISH and Drop-seq

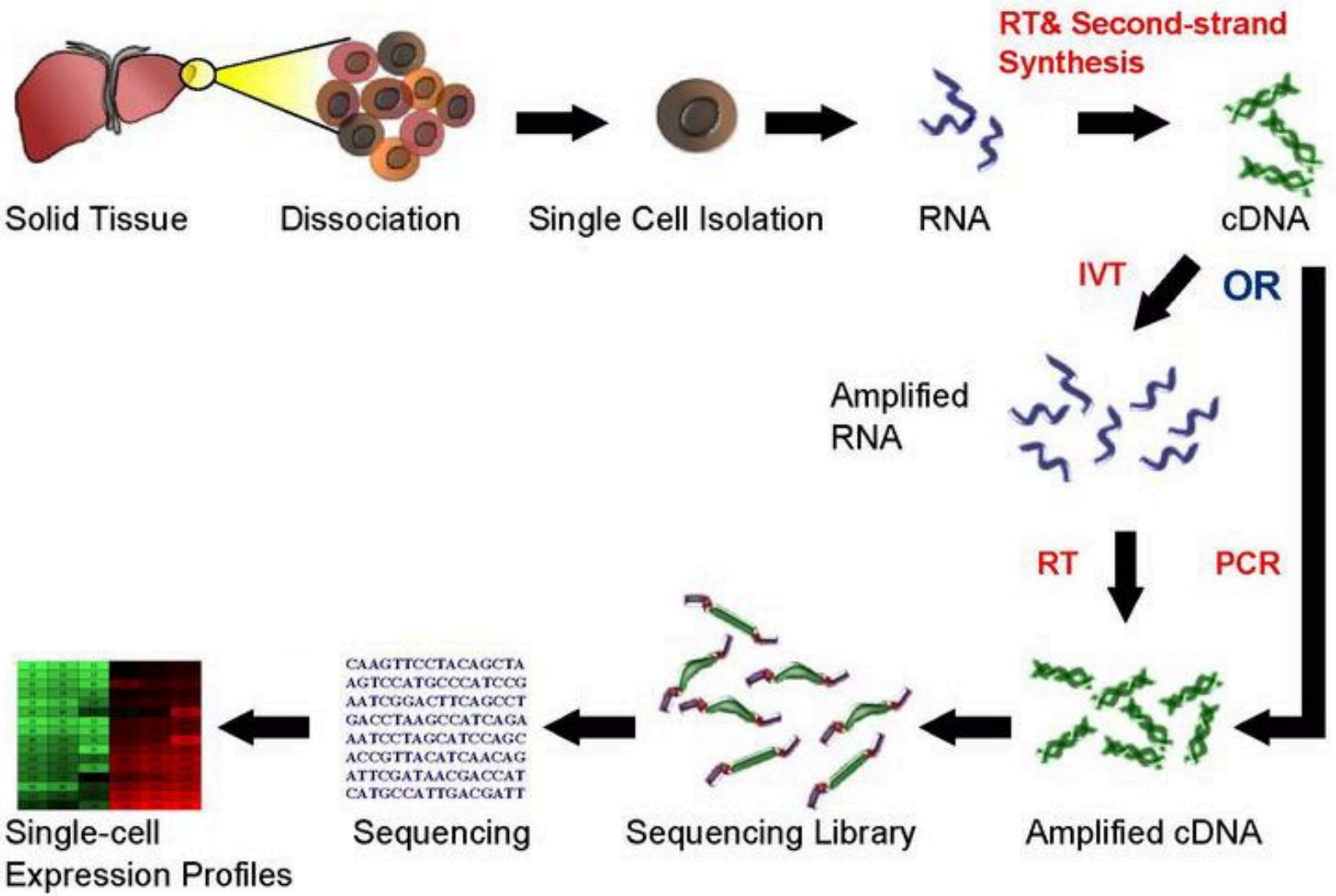
Rare Cell Detection by Single-Cell RNA Sequencing as Guided by Single-Molecule RNA FISH, Torre et. al., *Cell Systems* 2018

How noisy is scRNA-seq?





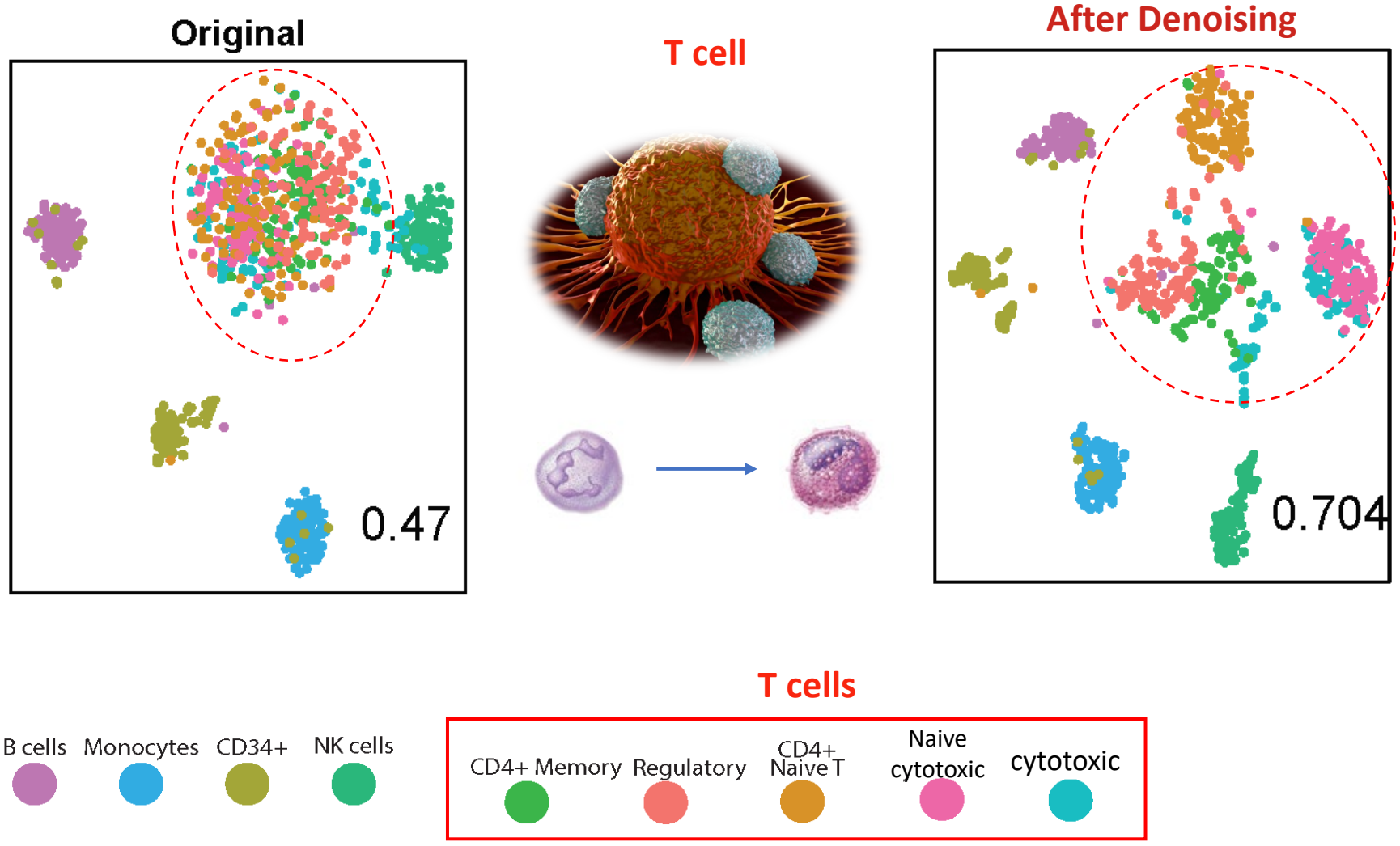
scRNA-seq workflow



(Image taken from http://hemberg-lab.github.io/scRNA.seq.course/)

How can denoising help?

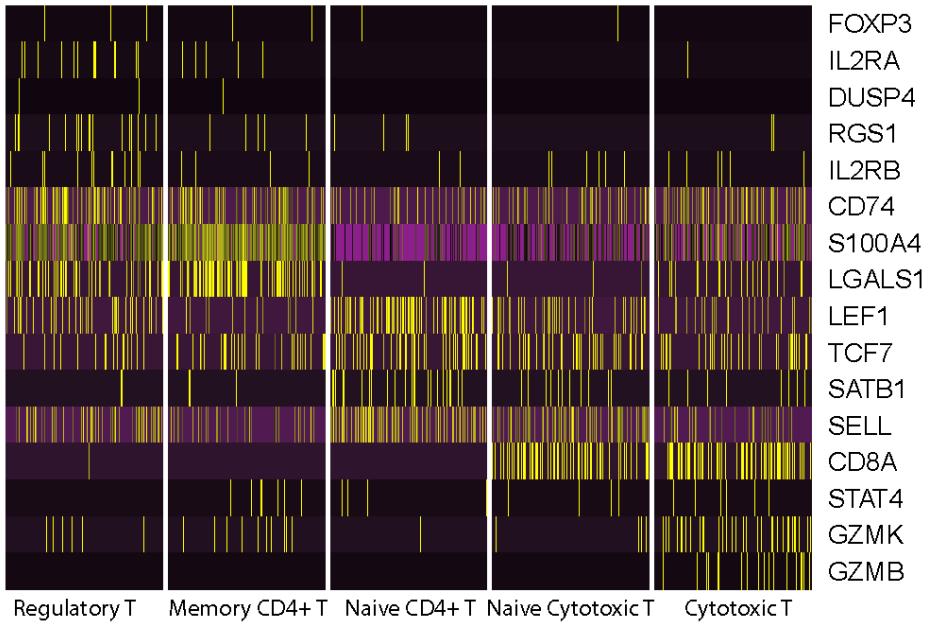
900 PBMC cells (immune cells in peripheral blood) with labels [Zheng et. al., 2017]



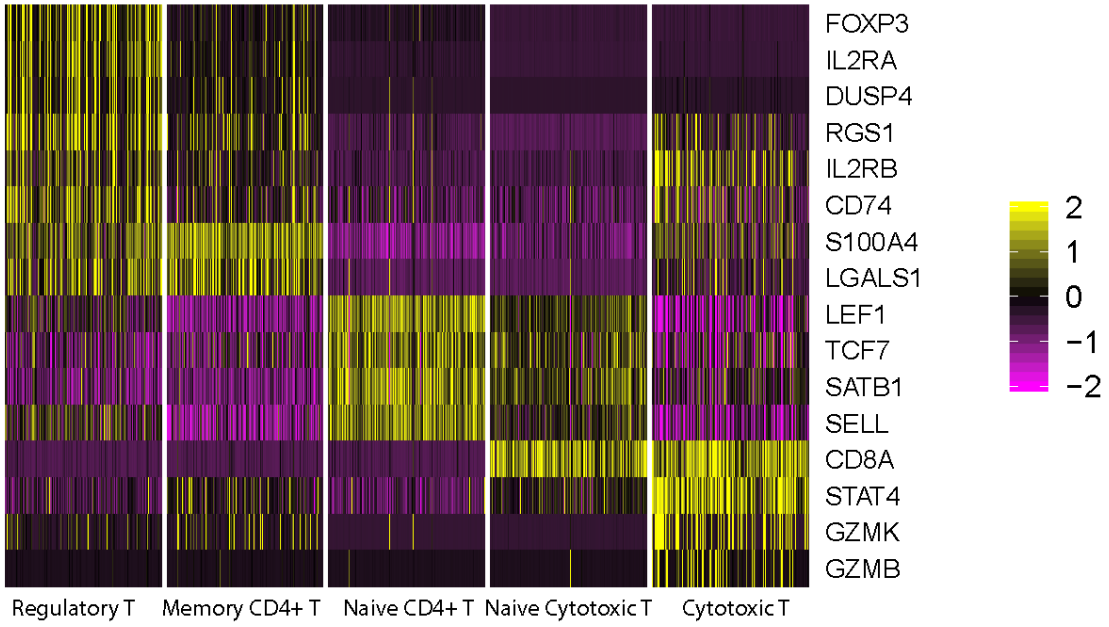
Improve recovering gene expression patterns

Identify the marker genes in each cell type

Original



After Denoising



Current methods

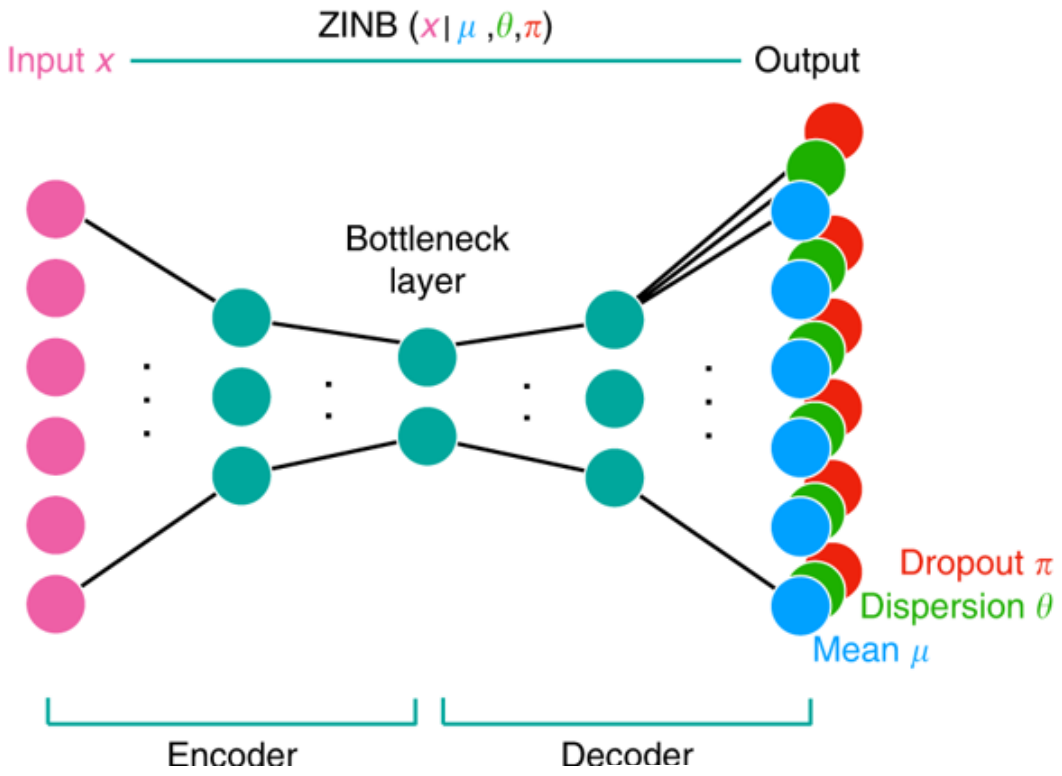
- **scImpute** [Li, W.V. et. al. (2018) *Nature Communications*]
- **MAGIC** [Van Dijk, D. et. al. (2018) *Cell]*
- **SAVER** [Huang, M. et. al. (2018) *Nature Methods*]
- **KNNsmooth** [Wagner, F. et. al. (2018) *BioRXiv*]
- **DrImpute** [Gong et. al. (2018) *BMC bioinformatics*]
- **ALRA** [Lopez, R. et. al. (2018) *BioRXiv*]

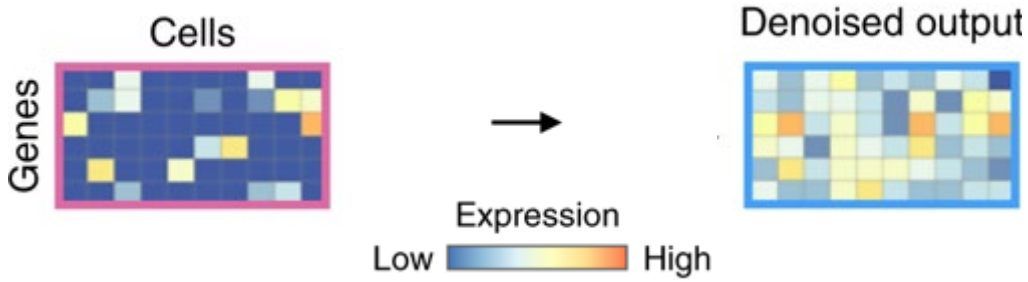
- **DCA** [Eraslan, G. et. al. (2018) *Nature Communications*]
- **scVI** [Lopez, R. et. al. (2018) *Nature Methods*]
- **TRANSLATE** [Badsha M. B. et. al. (2018) *BioRXiv*]
- **SAVER-X** [Wang, J. et. al. (2018) *BioRXiv*]

Deep learning approaches using autoencoder

Autoencoder: non-linear dimension reduction

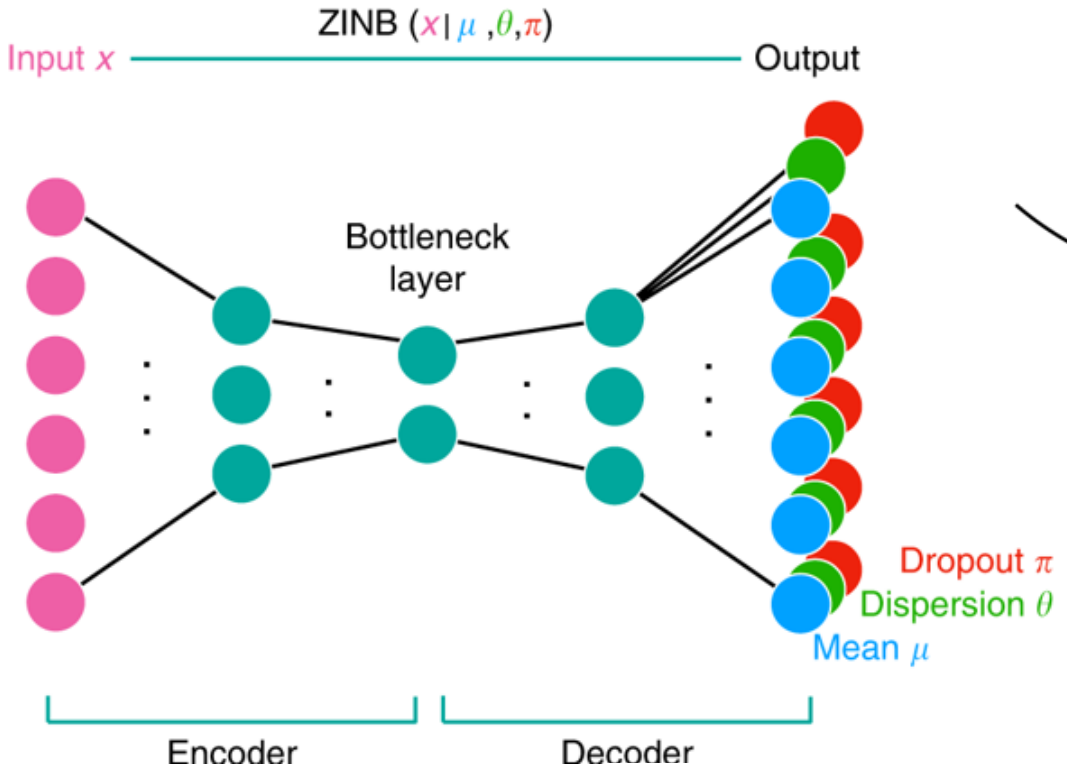
DCA model:

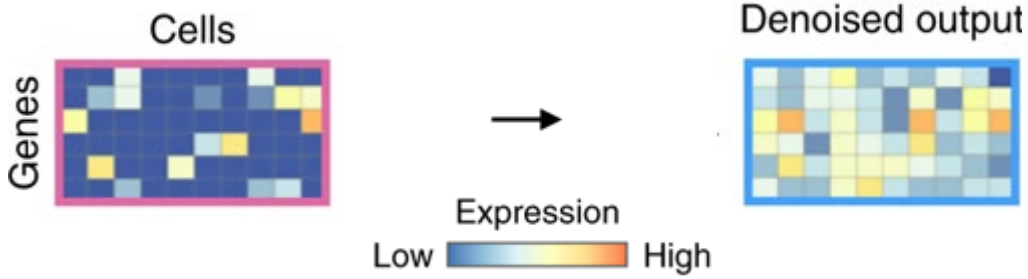


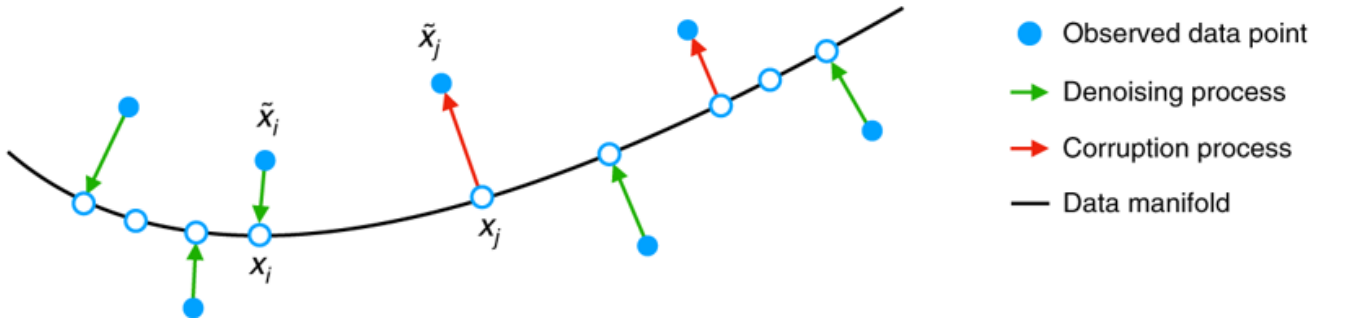


Autoencoder: non-linear dimension reduction

DCA model:





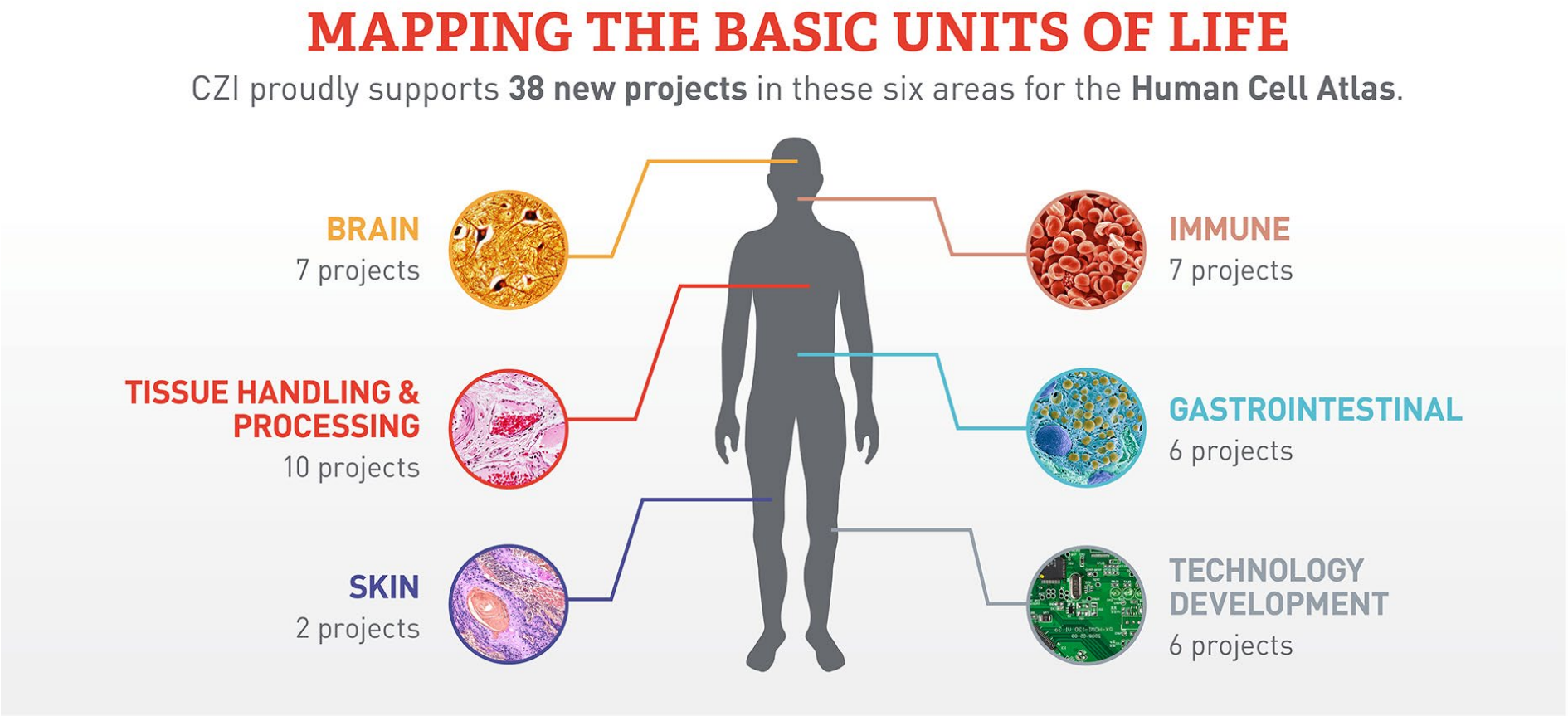


- **DCA:**
  assumes that data without noise lies in a low-dimensional manifold
- **scVI:**
  Variational autoencoder
- **SAVER-X:**
  separates biological randomness from technical noise and preserves biological randomness

Part IV
Transfer learning

Human Cell Atlas

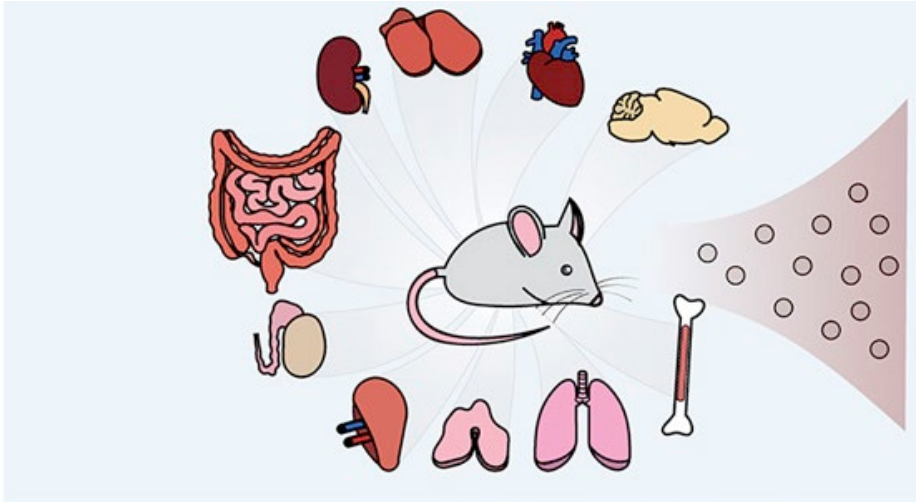
HCA: global collaboration to map all cells in a human body



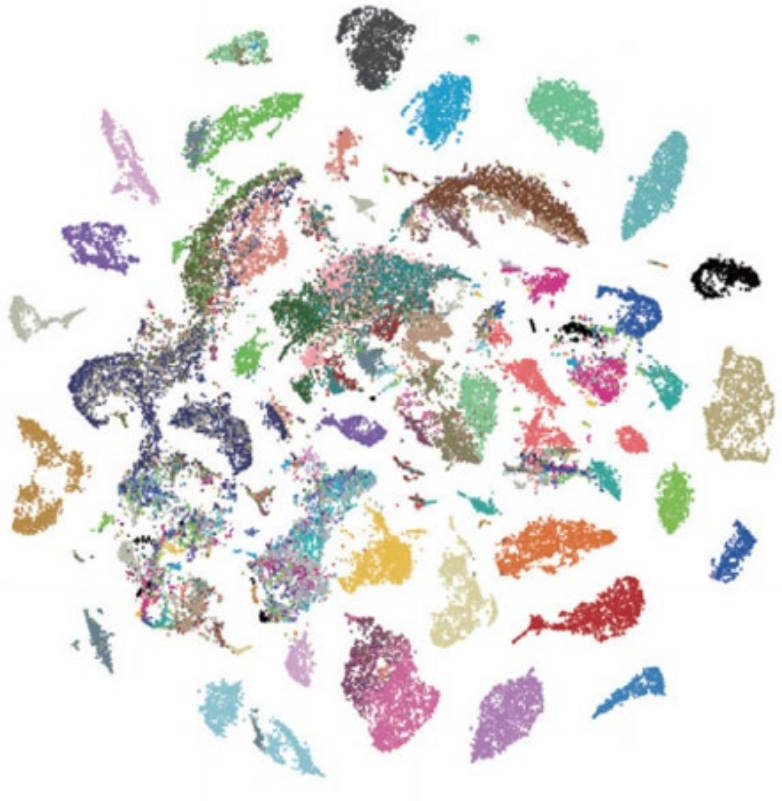
Mouse Cell Atlas

- Han et. al., Cell 2018
  $\sim 500{,}000$ cells, 40 tissues

- Tabula Muris Consortium, Nature 2018
  $\sim 100{,}000$ cells, 20 tissues



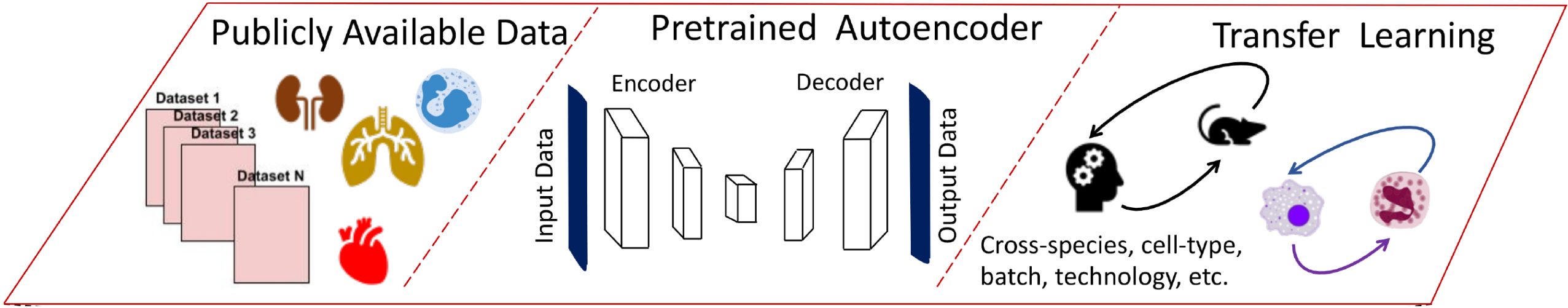
(Tabula Muris Consortium, *Nature 2018)*



(Han et. al., *Cell 2018)*

Transfer learning for denoising

- SAVER-X
  - Pretrain the autoencoder with public data
  - Modify weights with target data

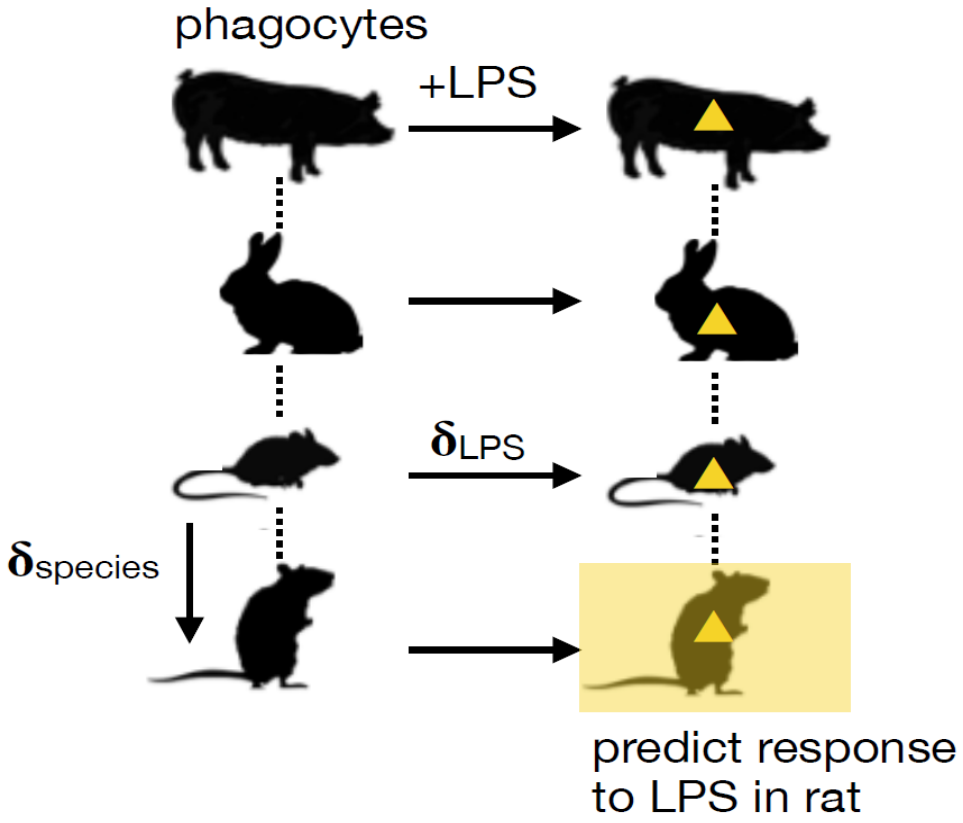


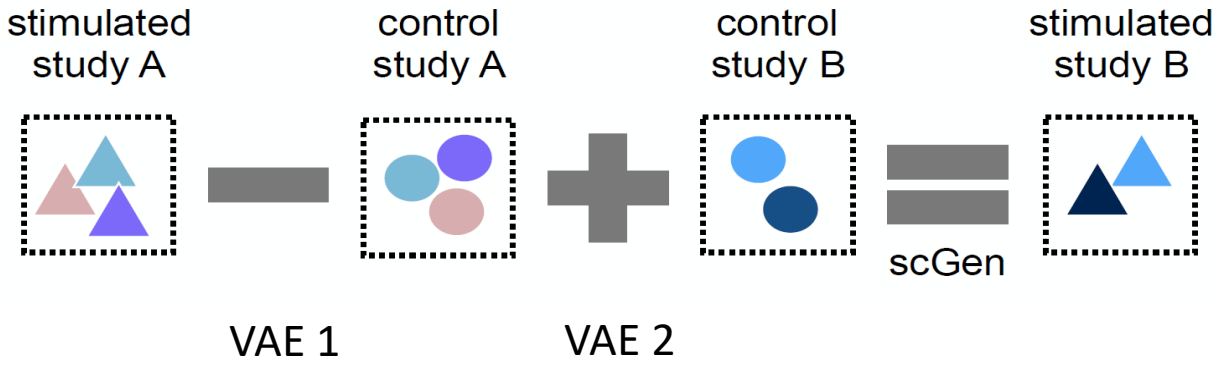
- TRANSLATE
  - Similar idea

https://github.com/jingshuw/SAVERX

Transfer learning across domain

scGen (Lotfollahi M. *BioRxiv*, 2018)

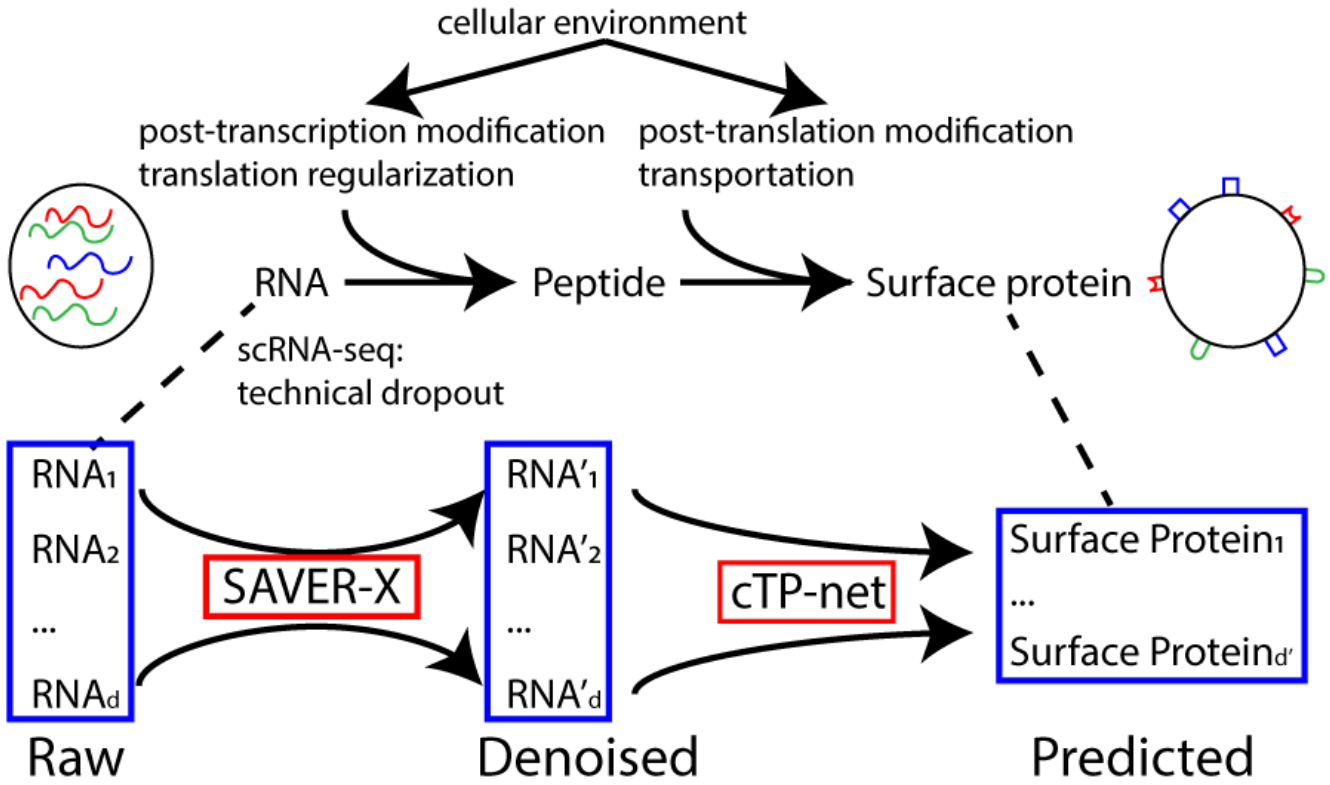




- Use the idea of style transfer variational autoencoder (VAE)
- Transfer the perturbation effect from one study/species to another study/species
- Assume that different study/species has similar response

Transfer learning for surface protein prediction

- cTP-net (Zhou Z. et. al.. *BioRxiv* 2019)



- Train a prediction model with CITE-seq/REAP-seq data
- Apply the model on new scRNA-seq datasets to predict surface protein levels

https://www.biorxiv.org/content/10.1101/671180v1?rss=1