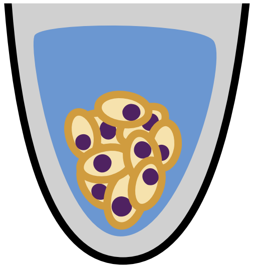
Comparing single-cell RNA-seq datasets

Ken Lau, Assistant Professor of Cell and Developmental Biology
ken.s.lau@vanderbilt.edu

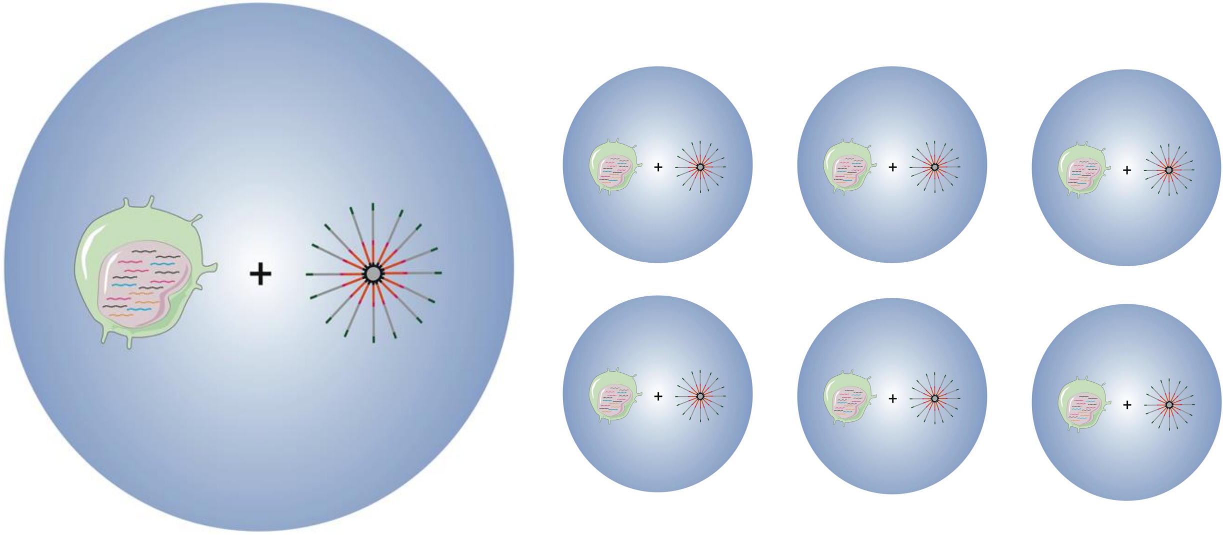
CQS Summer Academy (8/14/2018)

http://www.mc.vanderbilt.edu/vumcdept/cellbio/laulab/index.html
Twitter: @KenLauLab

Single cell suspension***



Single-cell encapsulation/
Library preparation



Sequencing and alignment
(Bioinformatics I)

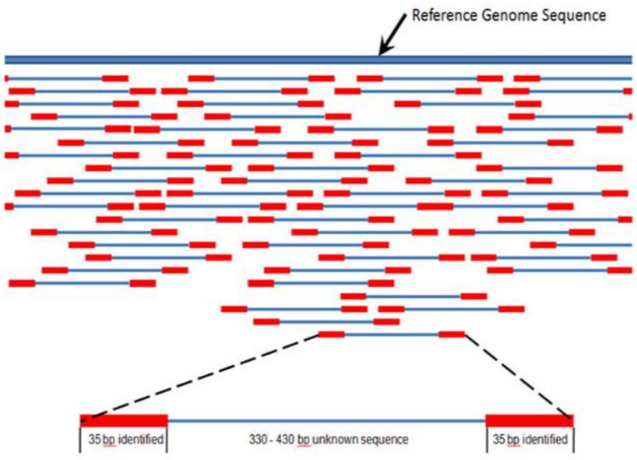
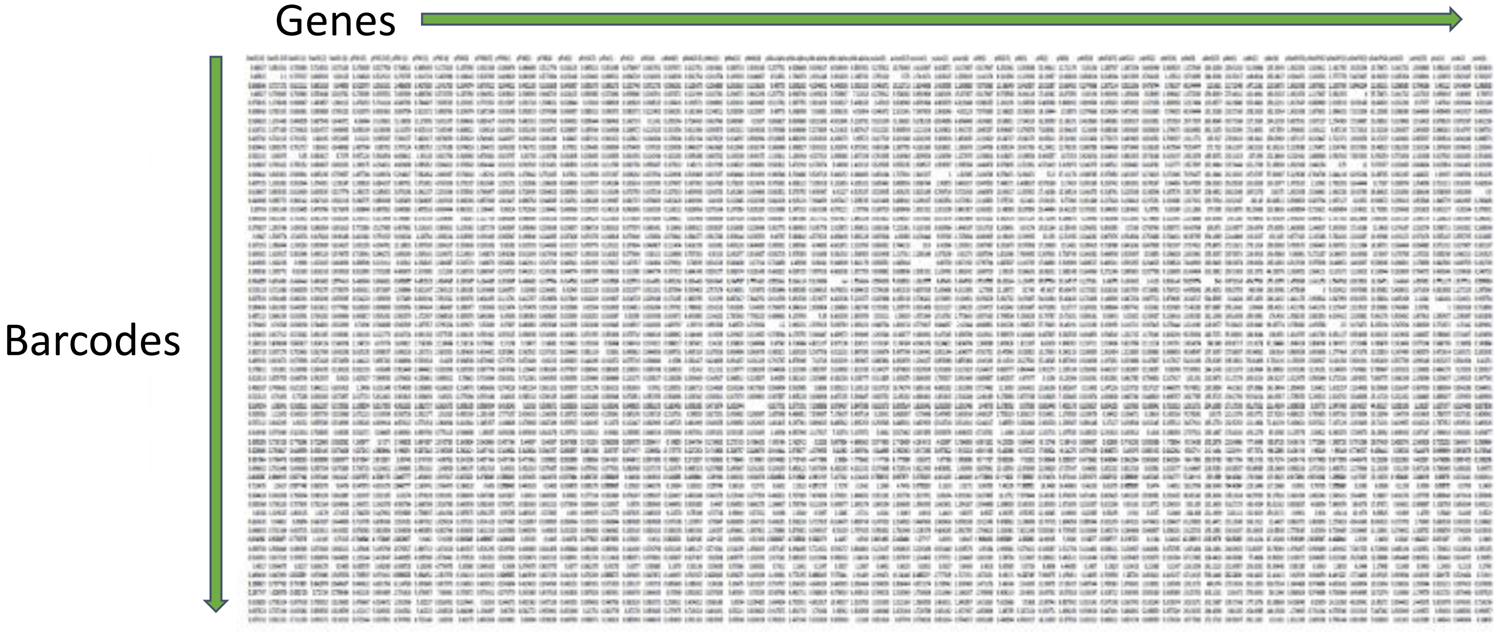


Table of genes and barcodes

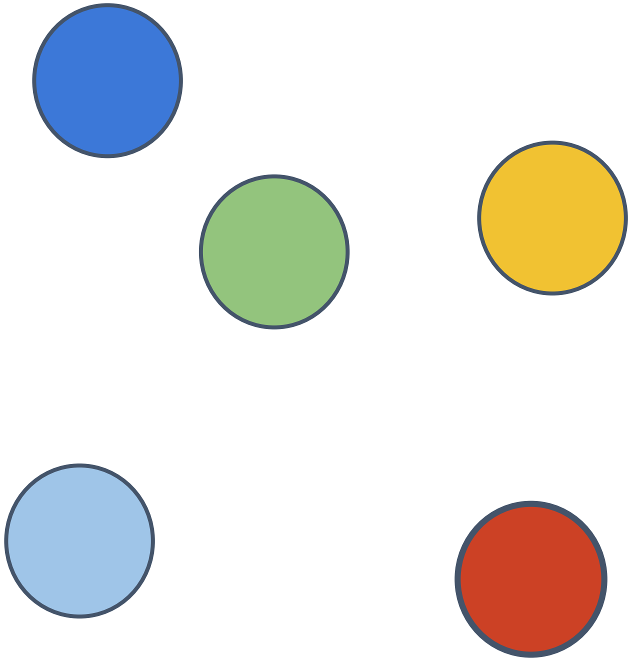


Genes

Barcodes

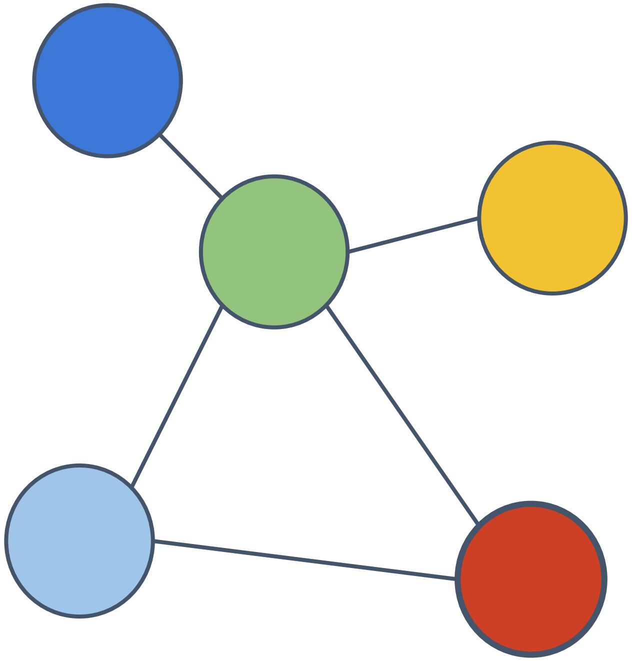
Common graph terms

- Common terms:
  - nodes are cells



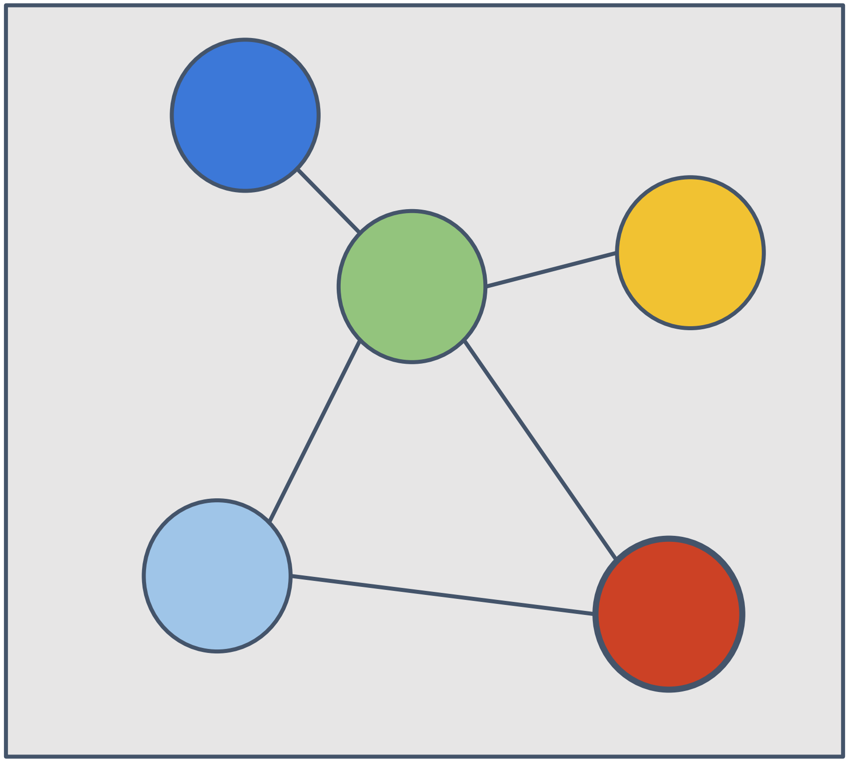
Common graph terms

- Common terms:
  - nodes are cells
  - edges are connections between nodes



Common graph terms

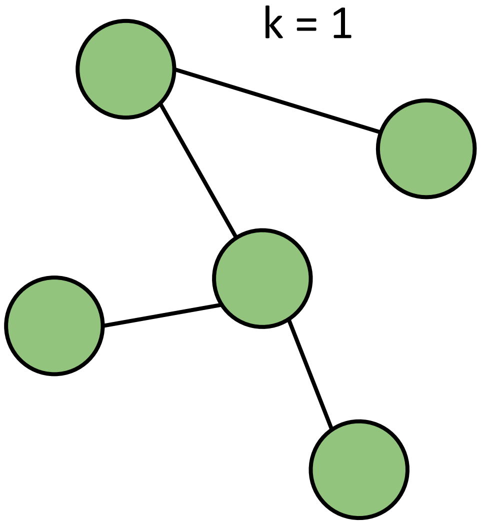
- Common terms:
    - nodes are cells
    - edges are connections between nodes
    - topology refers to overall structure of the graph

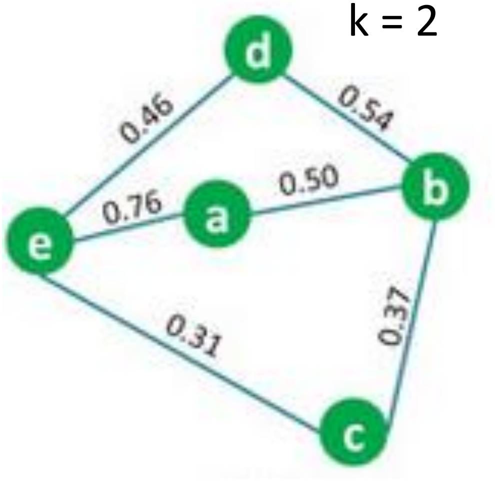


Introduction of k Nearest Neighbor (kNN) networks

Rules

- Connect each node to its k nearest neighbors
- Nearest Neighbor refers to "closest" cell in distance (can be in expression space, not in necessarily in physical space)
- Robust when k reaches a certain number





scRNA-seq processing steps

1. Cell identification from barcodes – doublet and low quality cell discrimination

2. Normalize and transform

3. Feature selection or data imputation

4. Batch correction

5. Dimension reduction

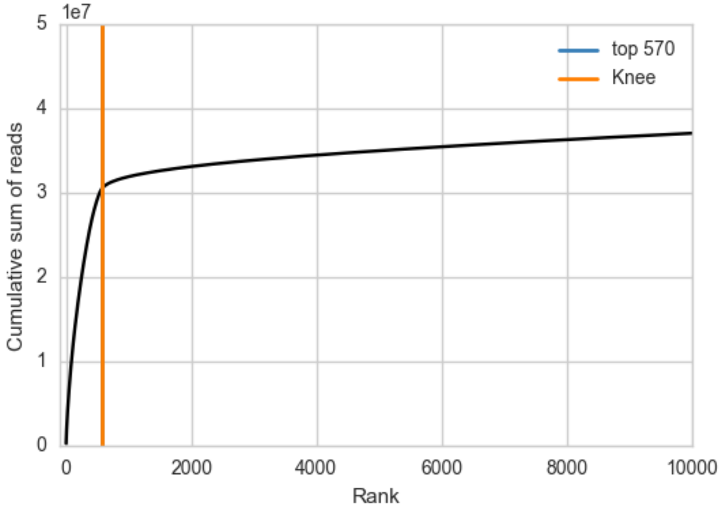
6. Clustering

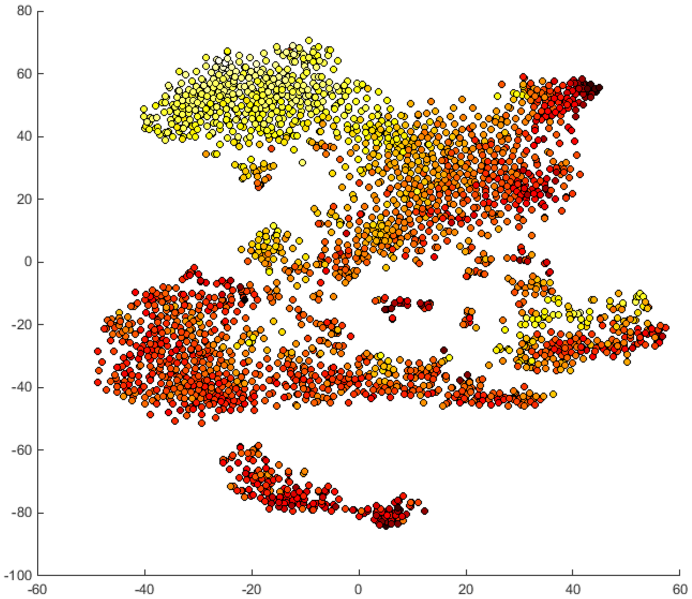
7. Trajectory analysis

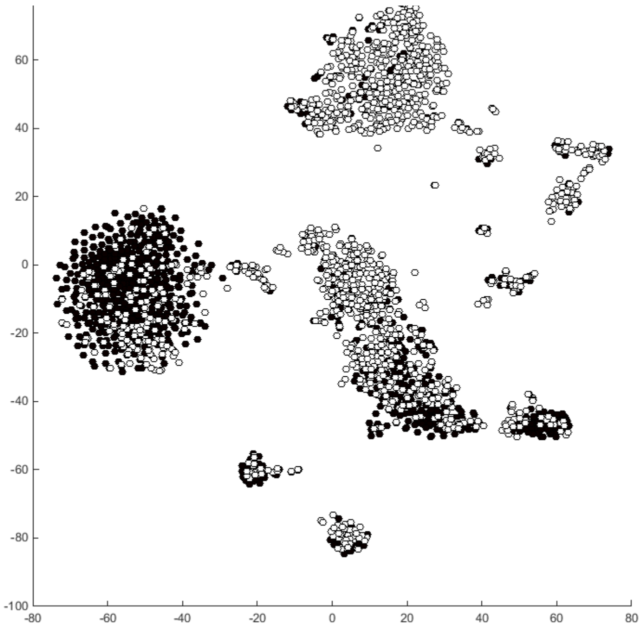
8. Data alignment

9. Differential expressed genes analysis

1. Cell identification

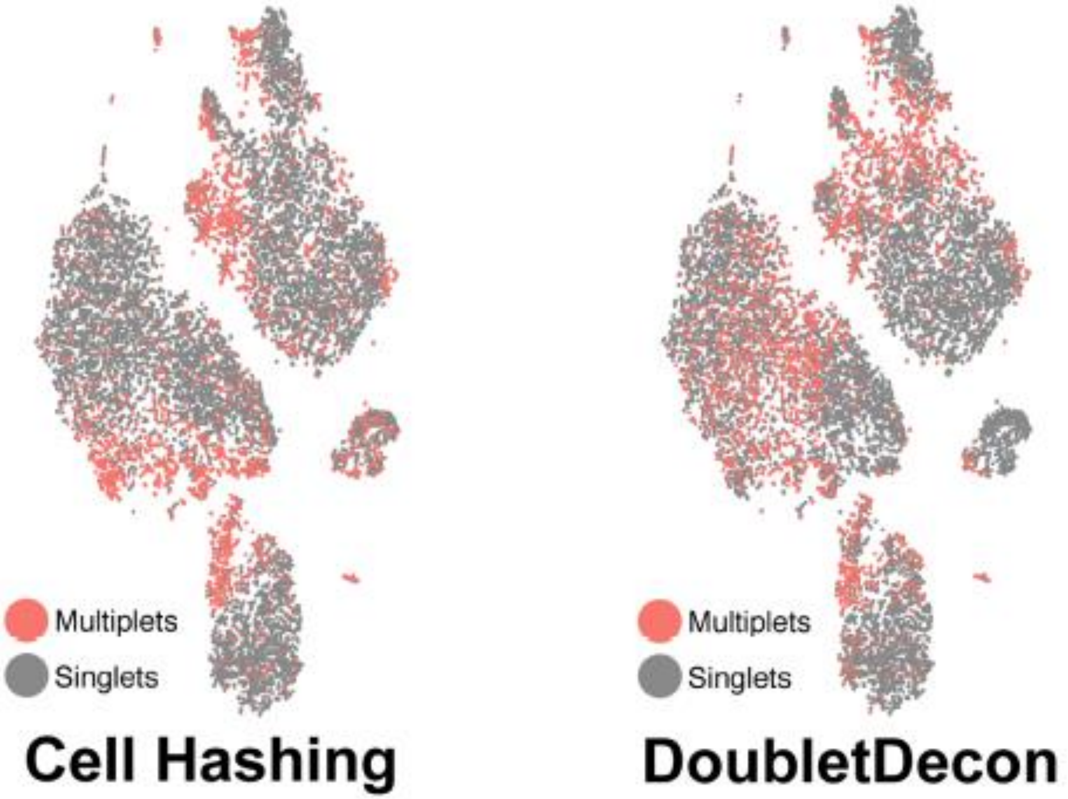






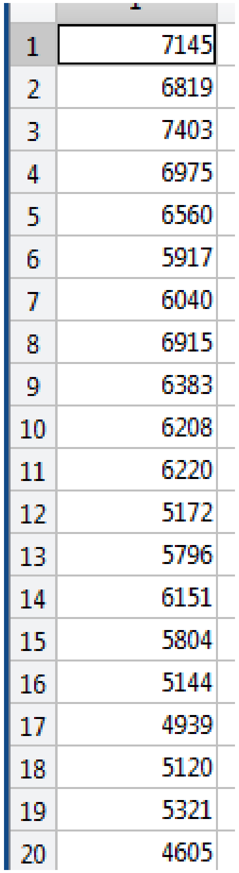
1a. Doublet discrimination – DoubletDecon (DePasquale et al.)

- Identify cell clusters, and then generate synthetic doublets by mixing reference cell profiles. Match cells to synthetic doublets.
- “Rescue” cells with unique expression profiles to the reference clusters.
- Performance not super great.



2. Processing – standard

- Total transcript (UMI) different even among same cell type due to stochastic sampling



1	7145
2	6819
3	7403
4	6975
5	6560
6	5917
7	6040
8	6915
9	6383
10	6208
11	6220
12	5172
13	5796
14	6151
15	5804
16	5144
17	4939
18	5120
19	5321
20	4605

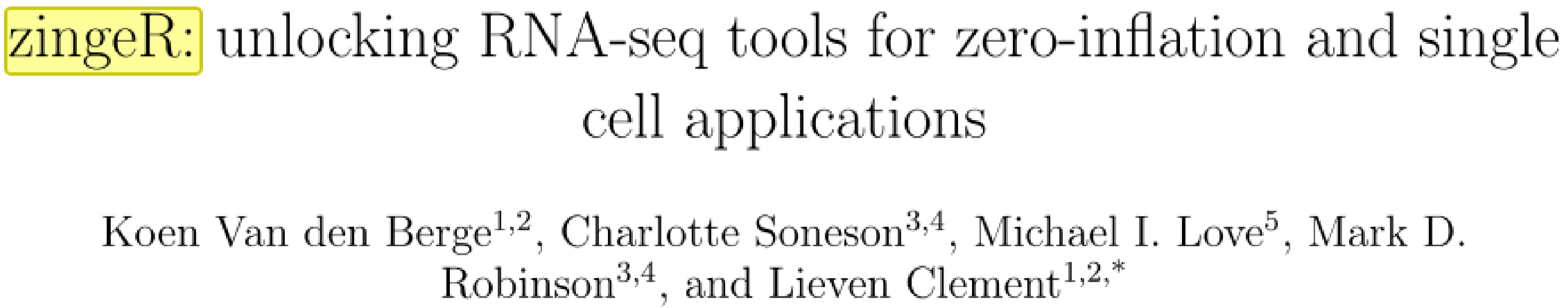
Data processing:

1. Separate mouse and human cells (if spike in)

2. Normalize UMI counts to total UMI counts (to get a fraction expression for each gene)

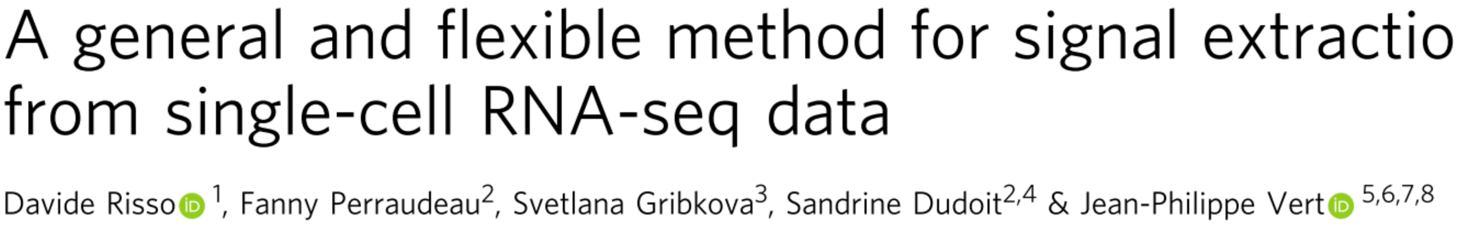
3. Variance normalization (most people use log+1 transform or asinh)

Ways to work with zero-inflated data without normalization also available (mostly using fits of negative binomial distributions and then breaking out the zero and negative binomial components)



zingeR: unlocking RNA-seq tools for zero-inflation and single cell applications

Koen Van den Berge[1,2], Charlotte Soneson[3,4], Michael I. Love[5], Mark D. Robinson[3,4], and Lieven Clement[1,2,*]



A general and flexible method for signal extraction from single-cell RNA-seq data

Davide Risso[1], Fanny Perraudeau[2], Svetlana Gribkova[3], Sandrine Dudoit[2,4] & Jean-Philippe Vert[5,6,7,8]

3. Feature selection

What is feature selection?

- Feature selection – from ~25000 genes, select a subset for further analysis
- This is inherently done in candidate-based approaches
- We want to do this in a unsupervised way (aka instead of us handpicking genes)

Why feature select?

- Not all features reliable (drop out/noisy)
- Not all features relate to process of study

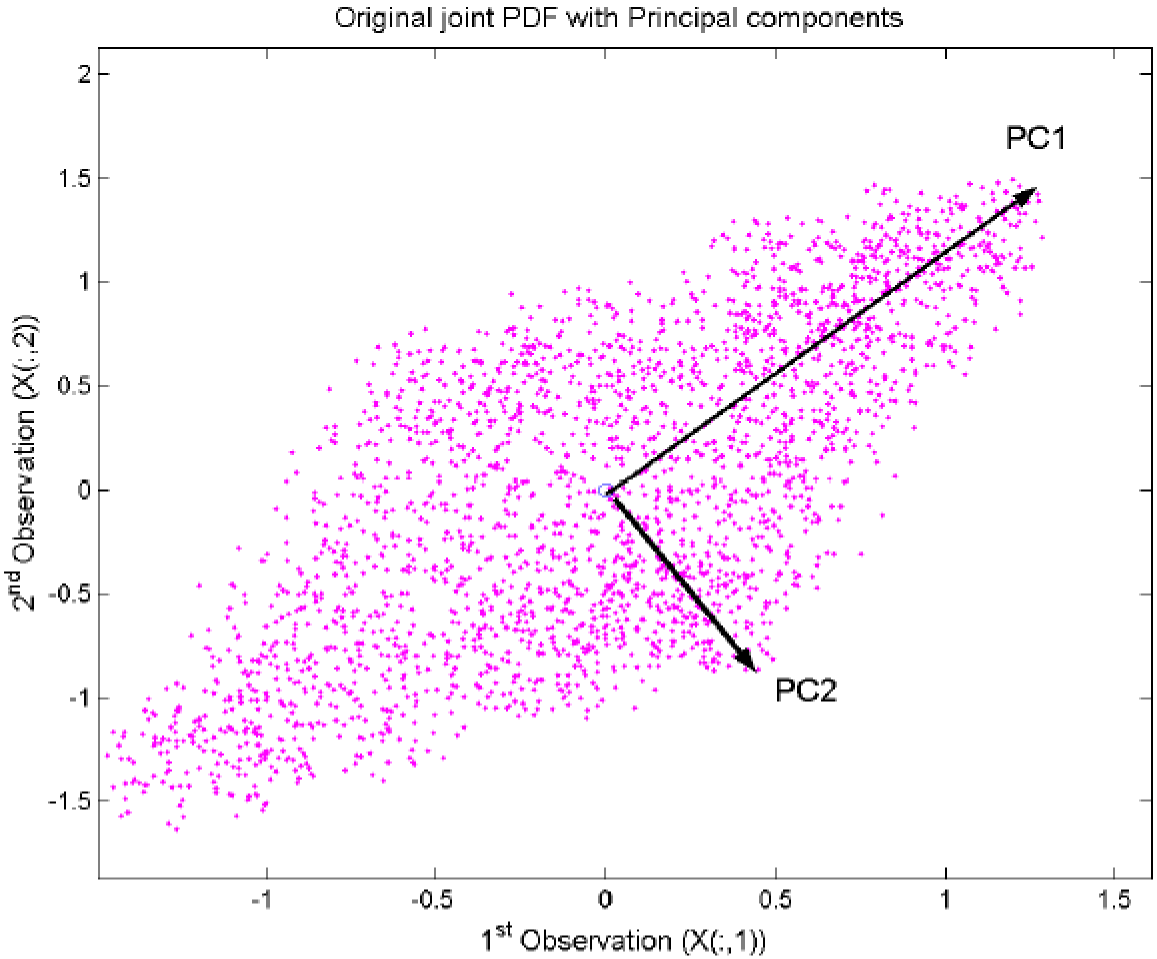
Unsupervised approaches for feature selection relies on specific patterns of gene expression over entire ensemble of cells.

Simple solution – pick features with highest variances

PCA (Principal Component Analysis)

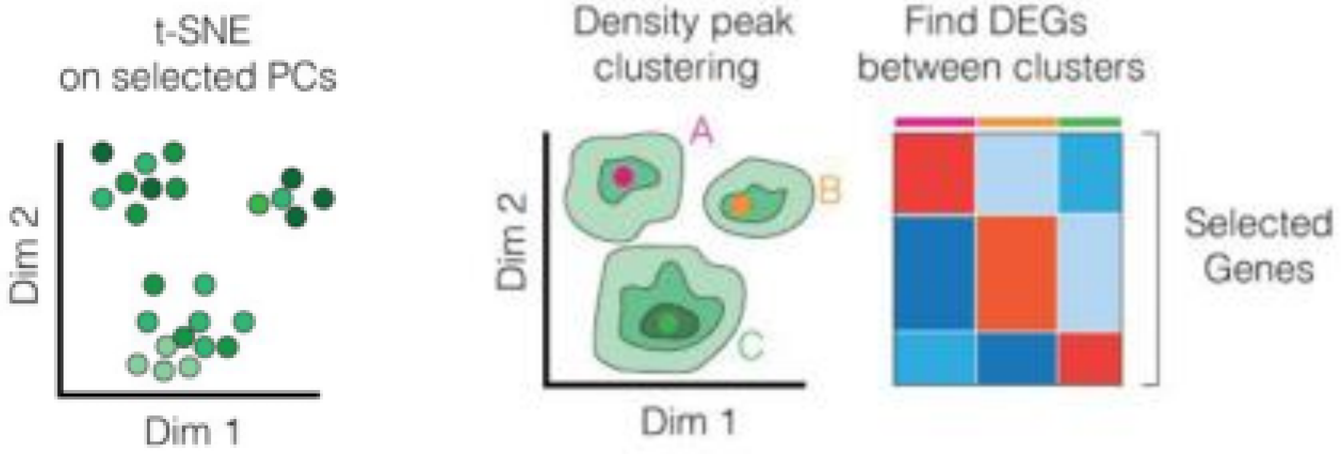
- Principle of PCA is to maximize the Variance of X with the least amount of principal components (latent variables)
- What is variance? Spread of the data, information content, change etc.
- Variance is the covariance of a dataset with itself, i.e. Var(X) = Cov(X,X) → Maximize
- What are principal components? Linear combinations of original variables

PCA

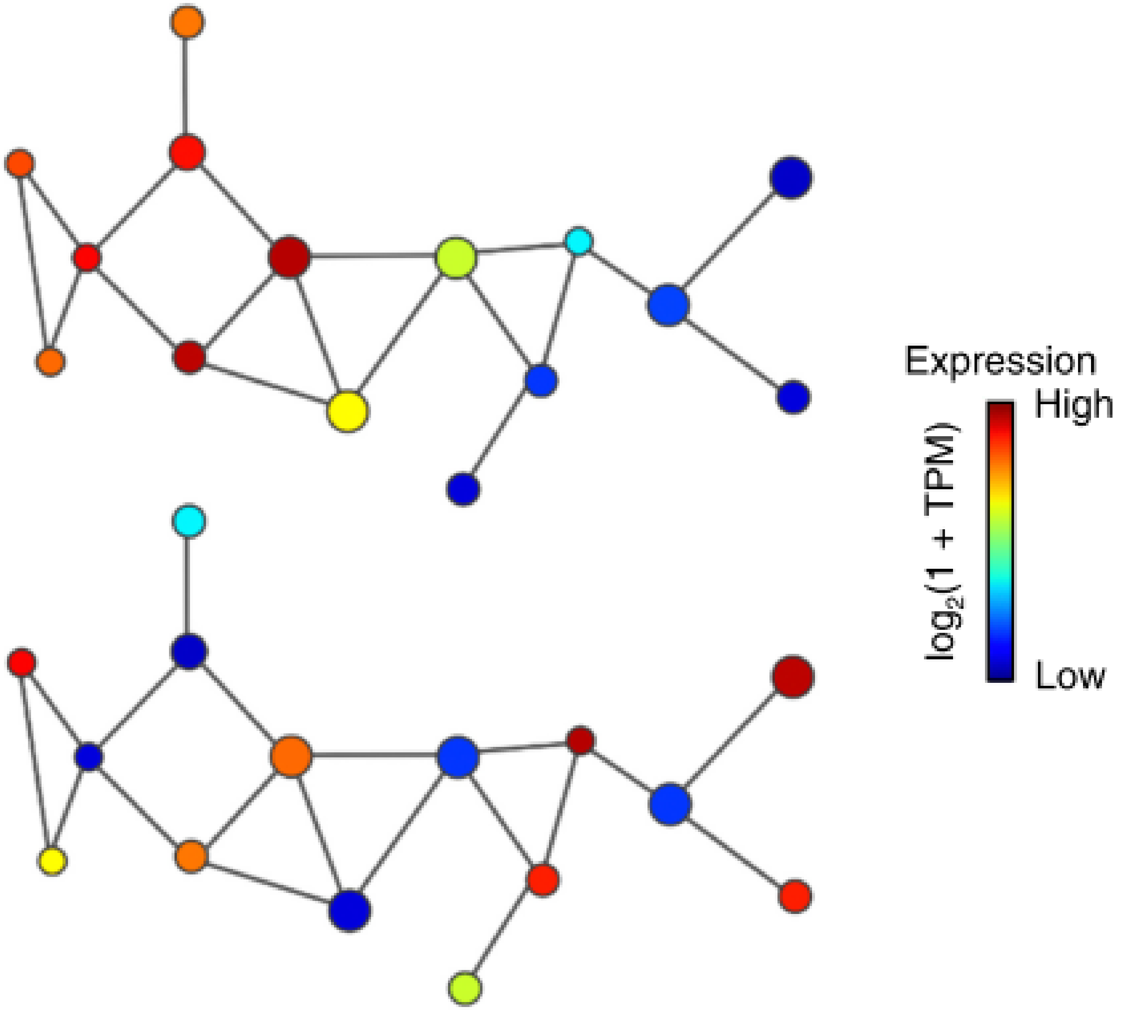


Problem: Technical variance may be bigger than real meaningful variance

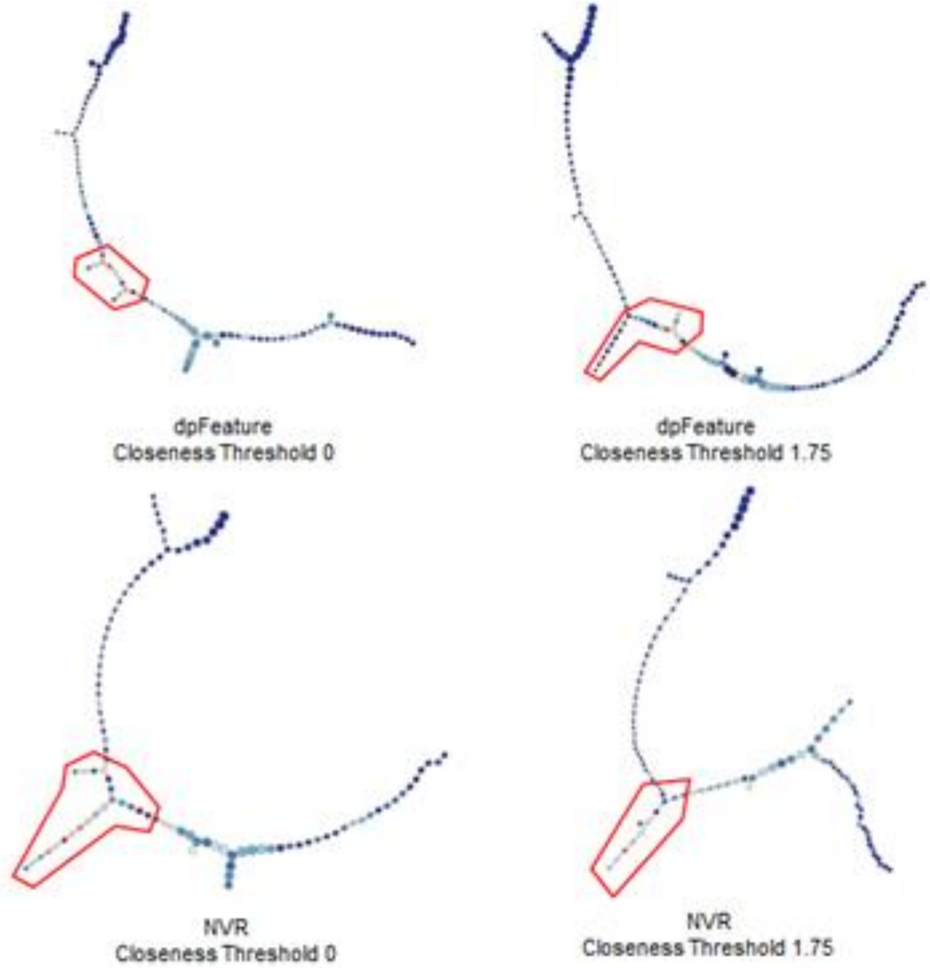
Cluster-based feature selection (dpFeature)



Neighborhood variance feature selection



Feature selection performance depends upon distribution of the data



dpFeature
Closeness Threshold 0

dpFeature
Closeness Threshold 1.75

NVR
Closeness Threshold 0

NVR
Closeness Threshold 1.75

3a. Data Imputation

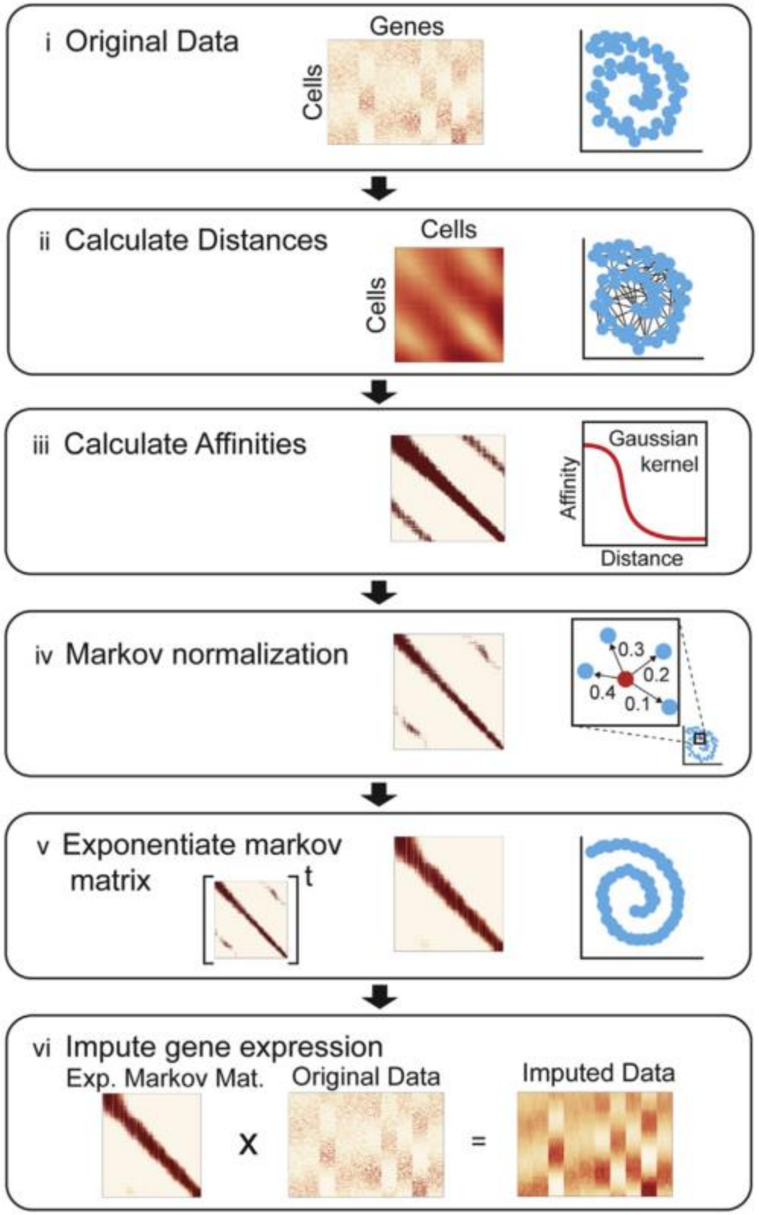
What is imputation?

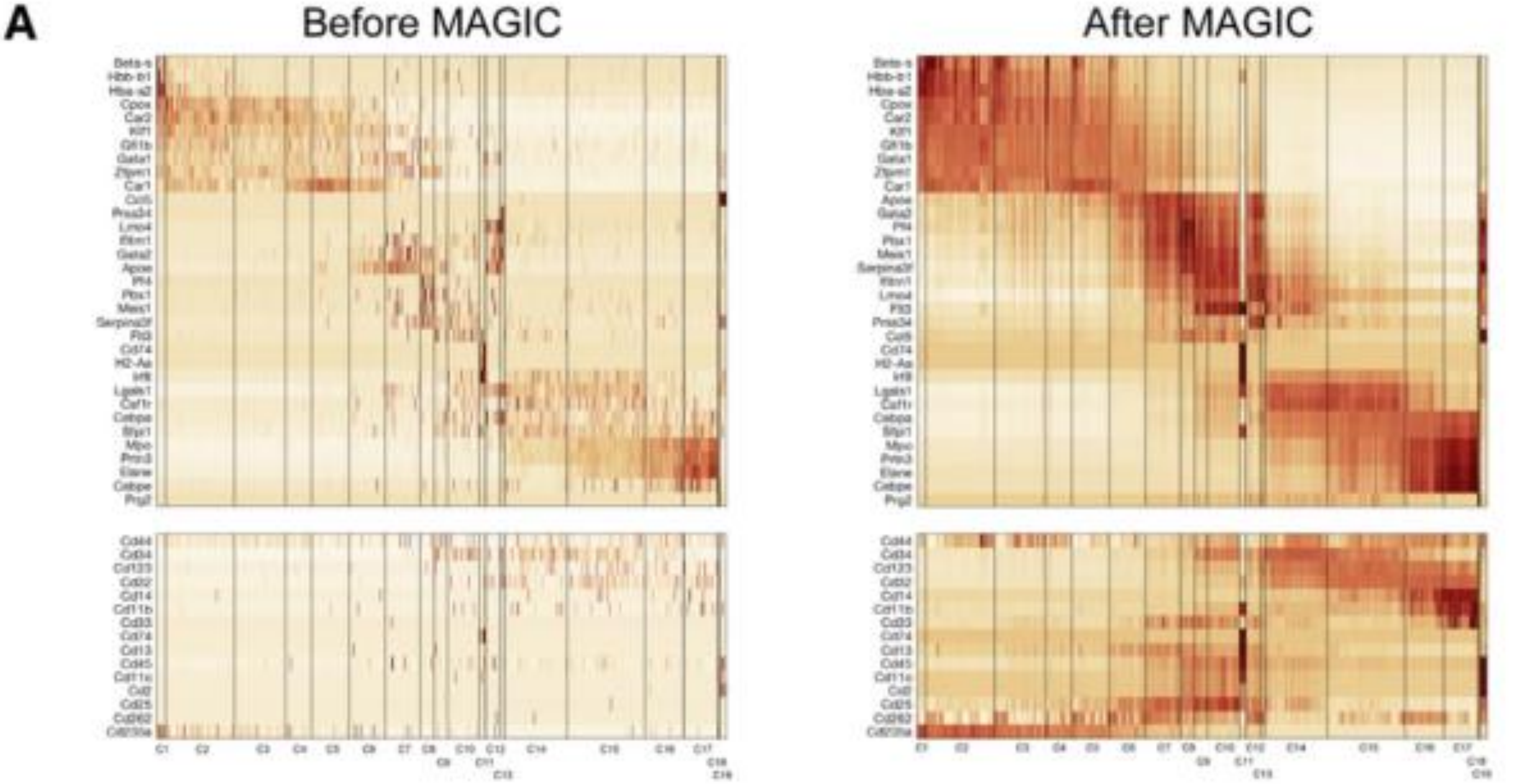
- Fill in zero or low expression entries with values
- Rely on what a cell neighbor is expressing

Why imputation?

- “Rescue” unreliable genes instead of throwing them out
- Low expressing genes are important, such as transcription factors

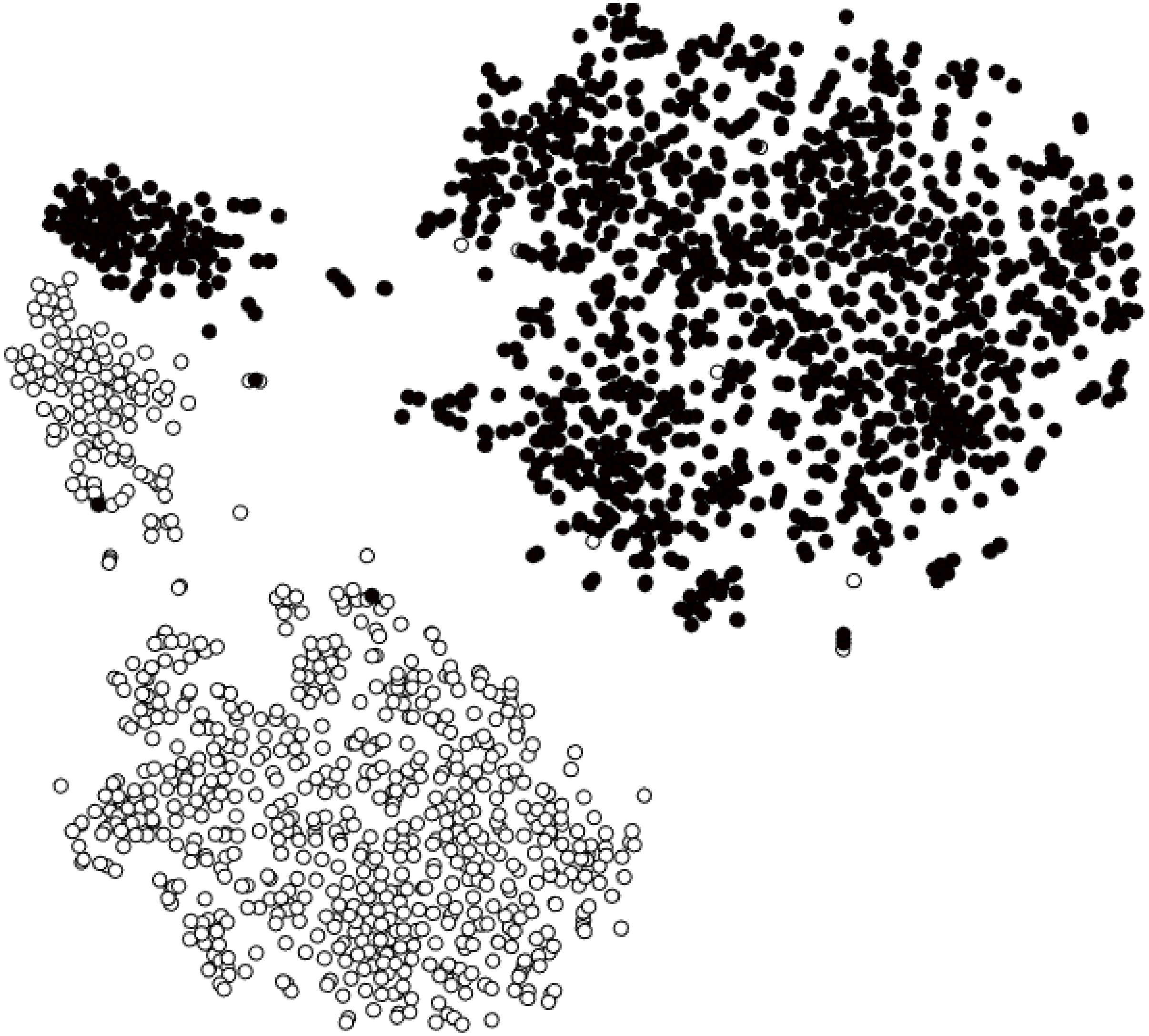
MAGIC – data manifold imputation (van Djik et al., 2018)





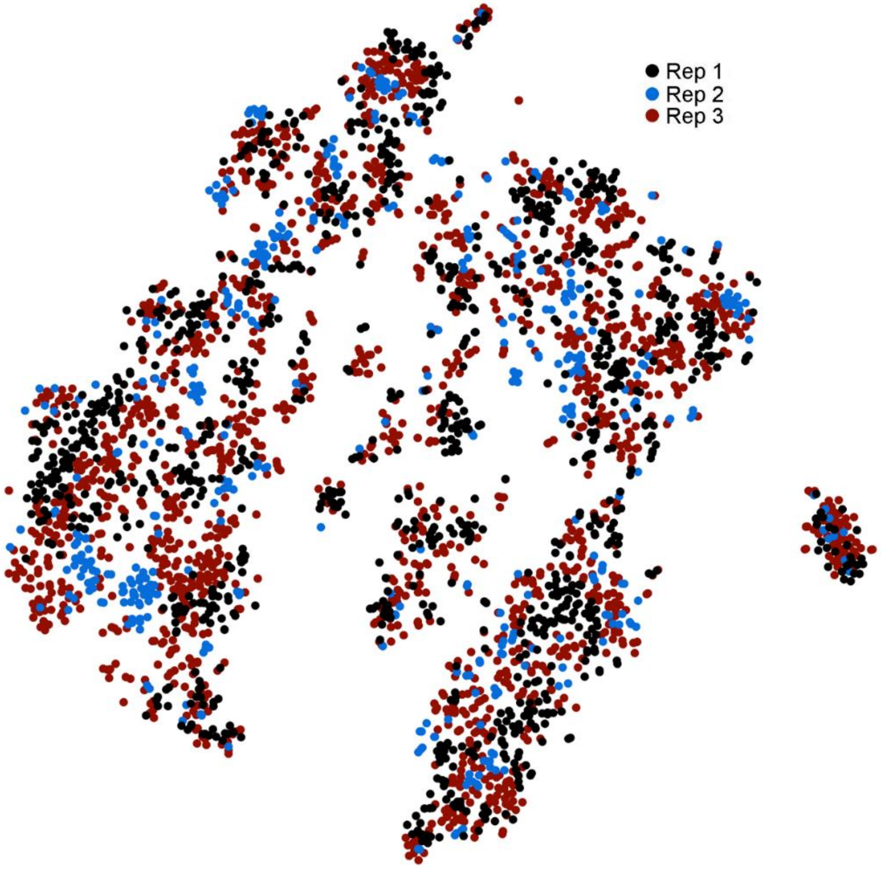
- Dangerous (but useful) as data is being “made up”.

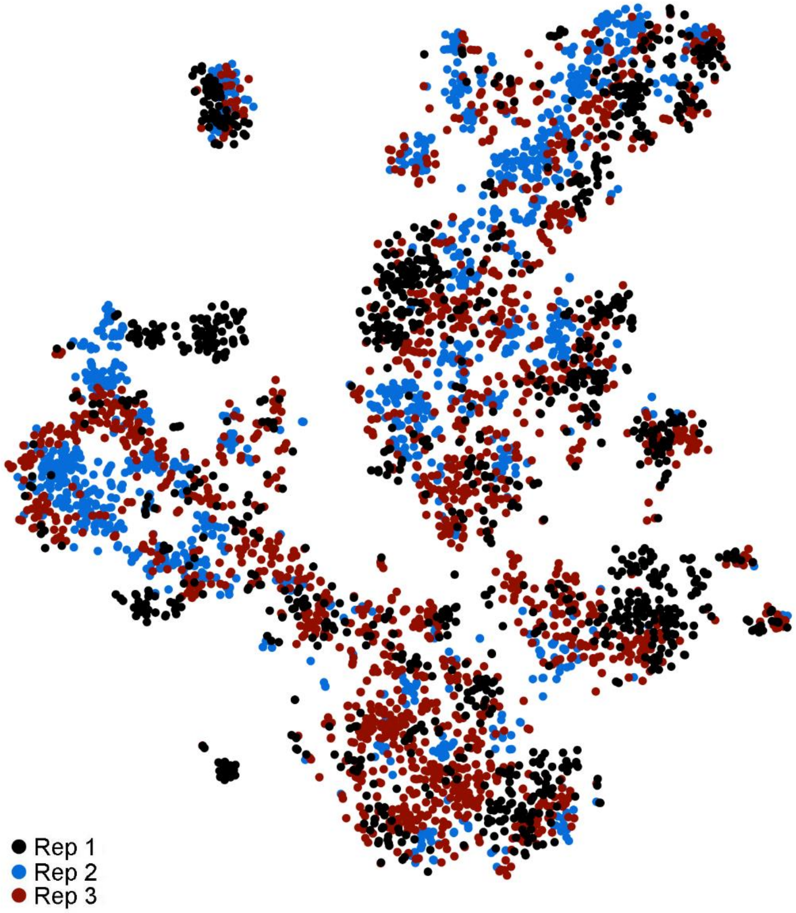
4 (8). Batch Correction (Alignment)



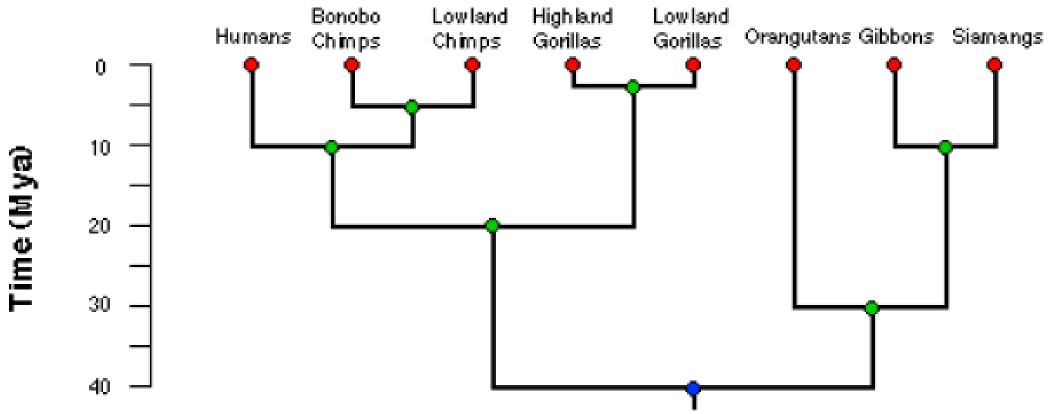
Replicates grouping by samples/runs indicate technical effects

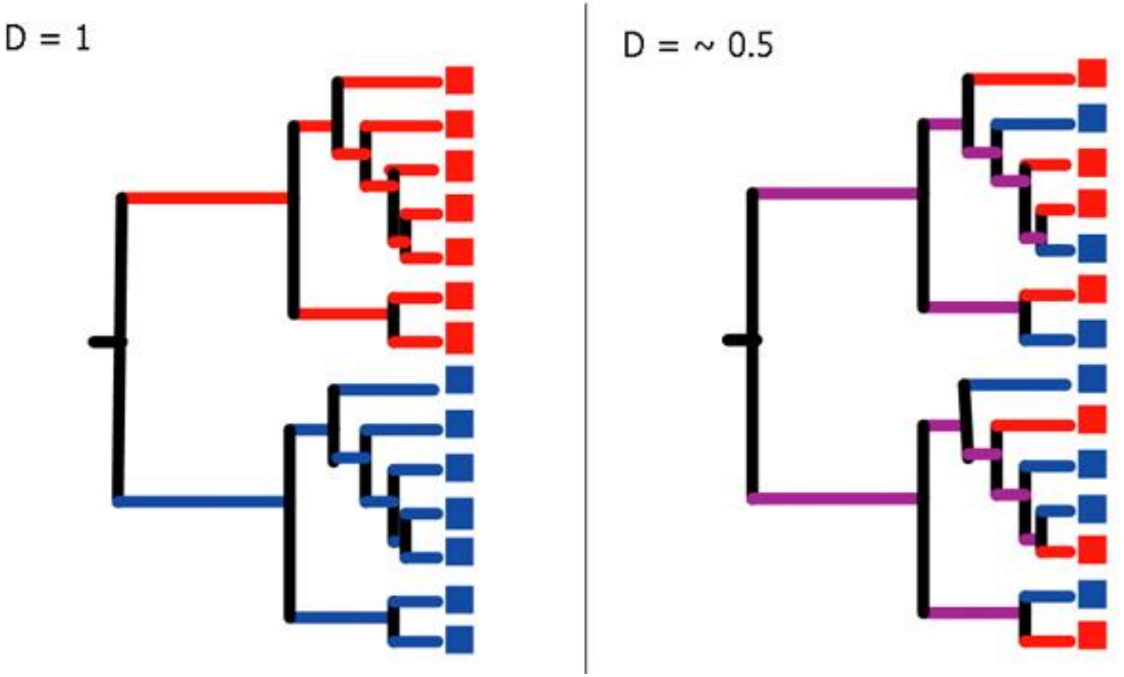
“Mixing” of data points from multiple replicates on t-SNE plot





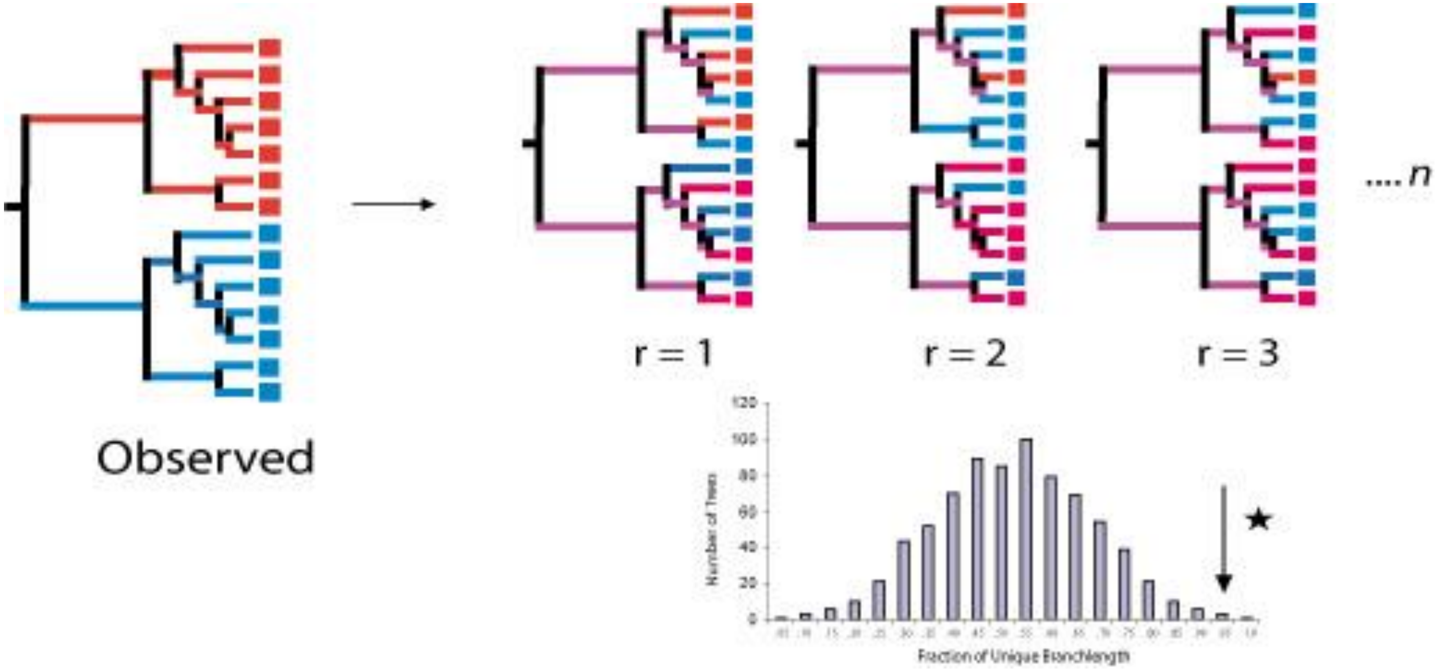
UniFrac – comparing two evolutionary trees (ecology/microbiome)



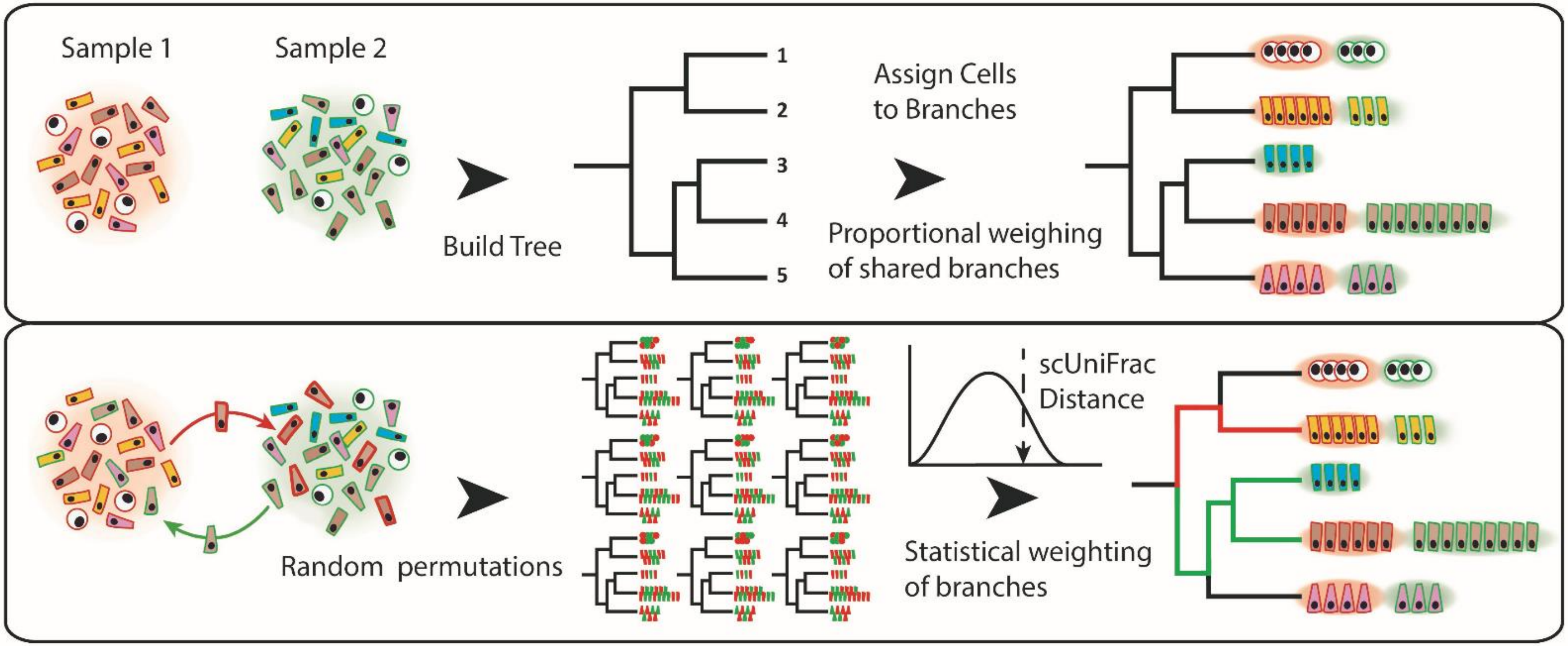


$$D= \frac{\text{the sum of "unshared" branch lengths)}}{\text{the sum of all tree branch lengths}}$$

UniFrac is a statistically robust way to compare trees between two samples with differing membership

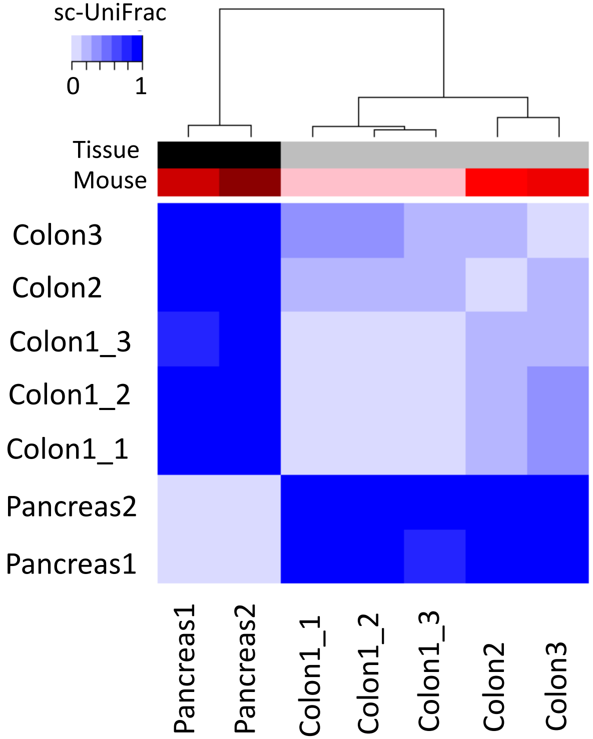


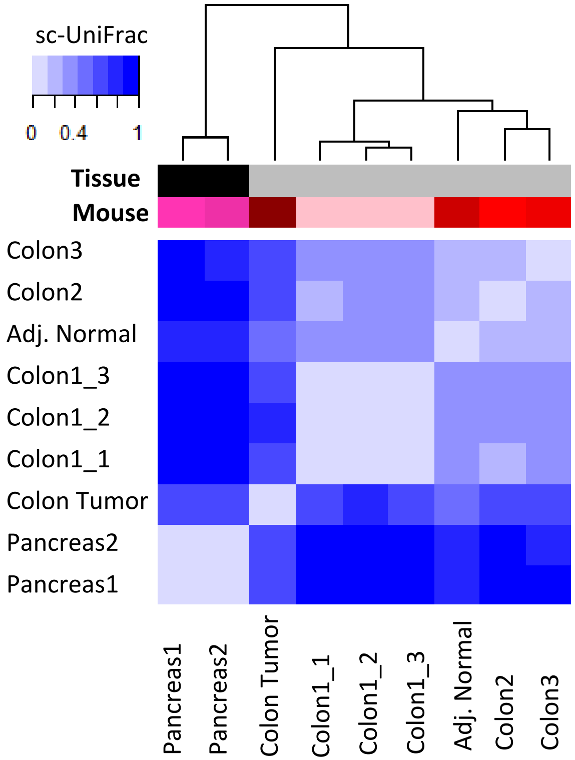
sc-UniFrac to quantify local and global "mixing" between samples



https://github.com/liuqivandy/scUnifrac

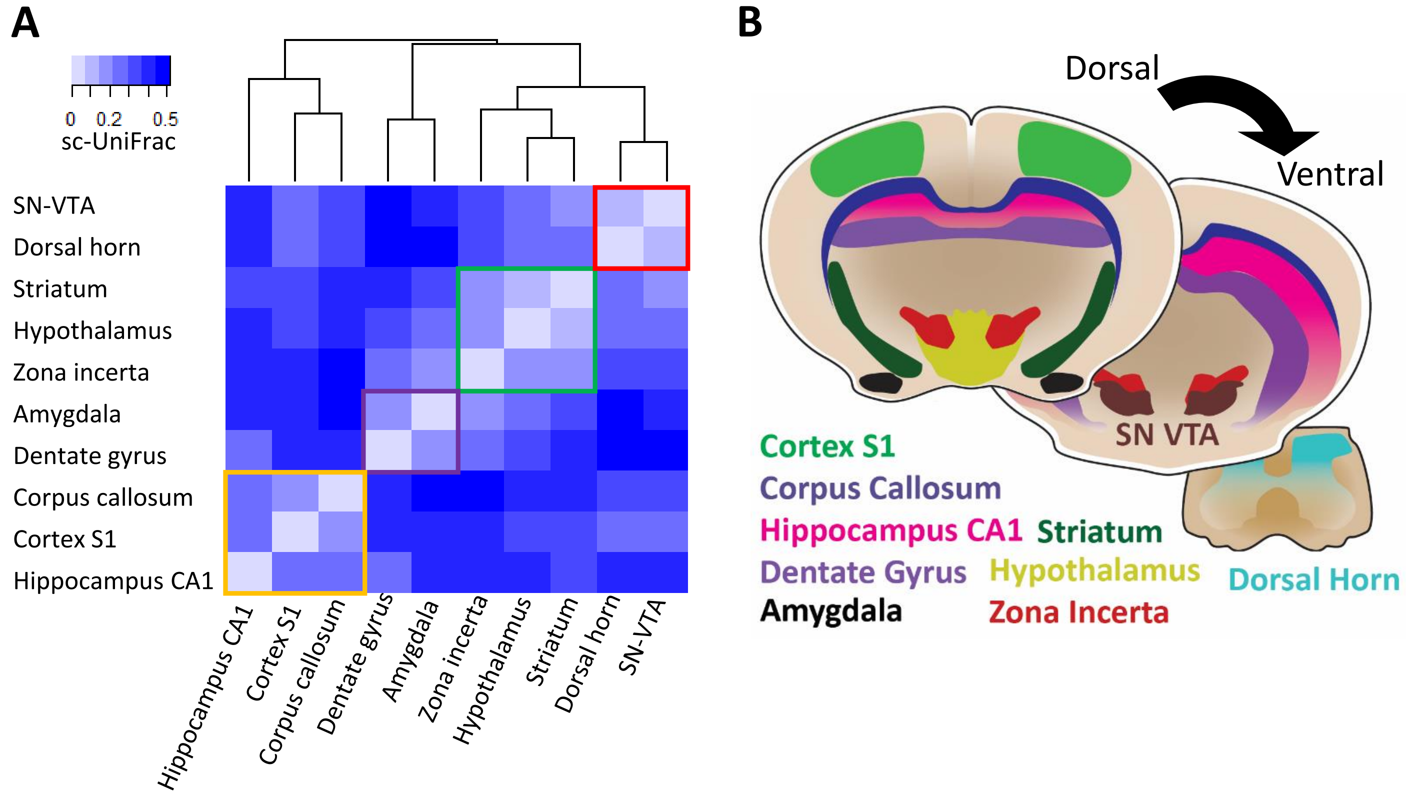
Degree of similarity between replicates and outgroup organ quantified





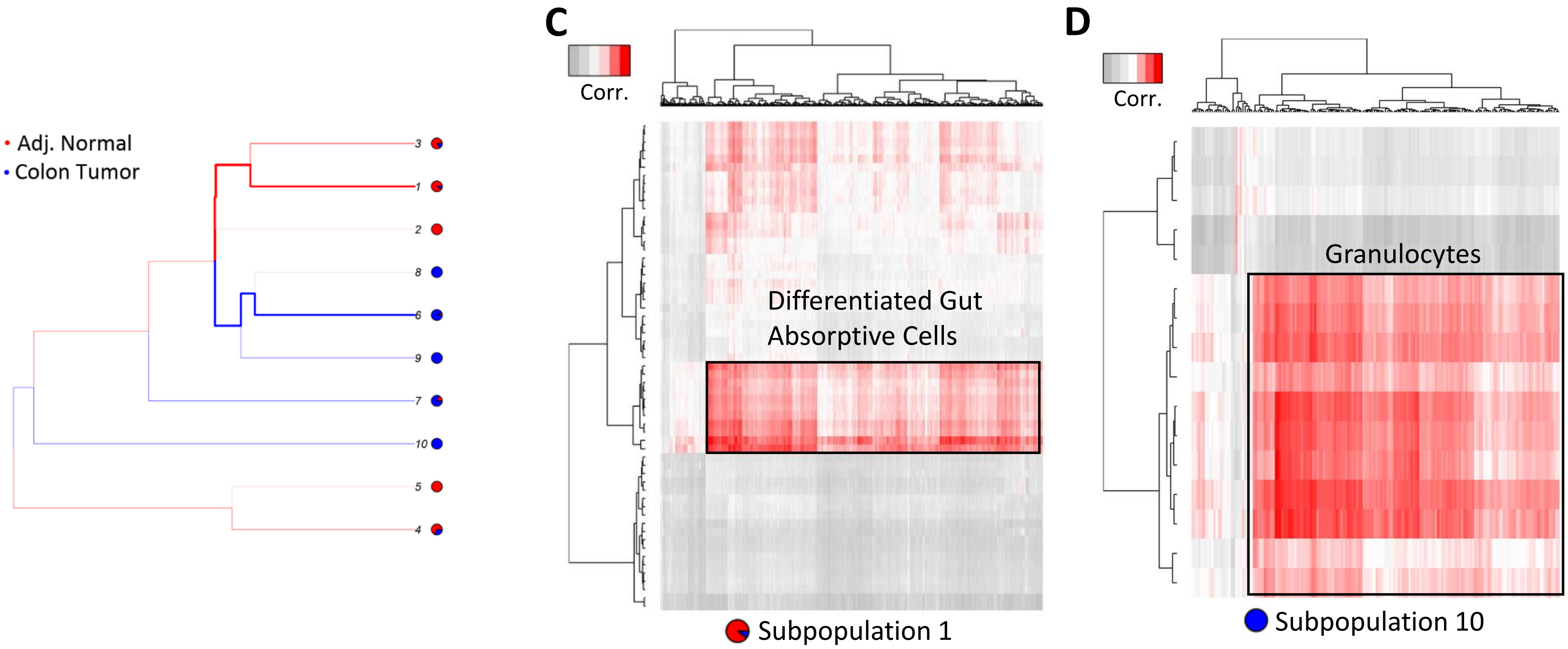
- Adding samples without redoing entire analysis

sc-UniFrac can be used to quantitatively group different samples (brain regions by developmental origins – Marques et al.)



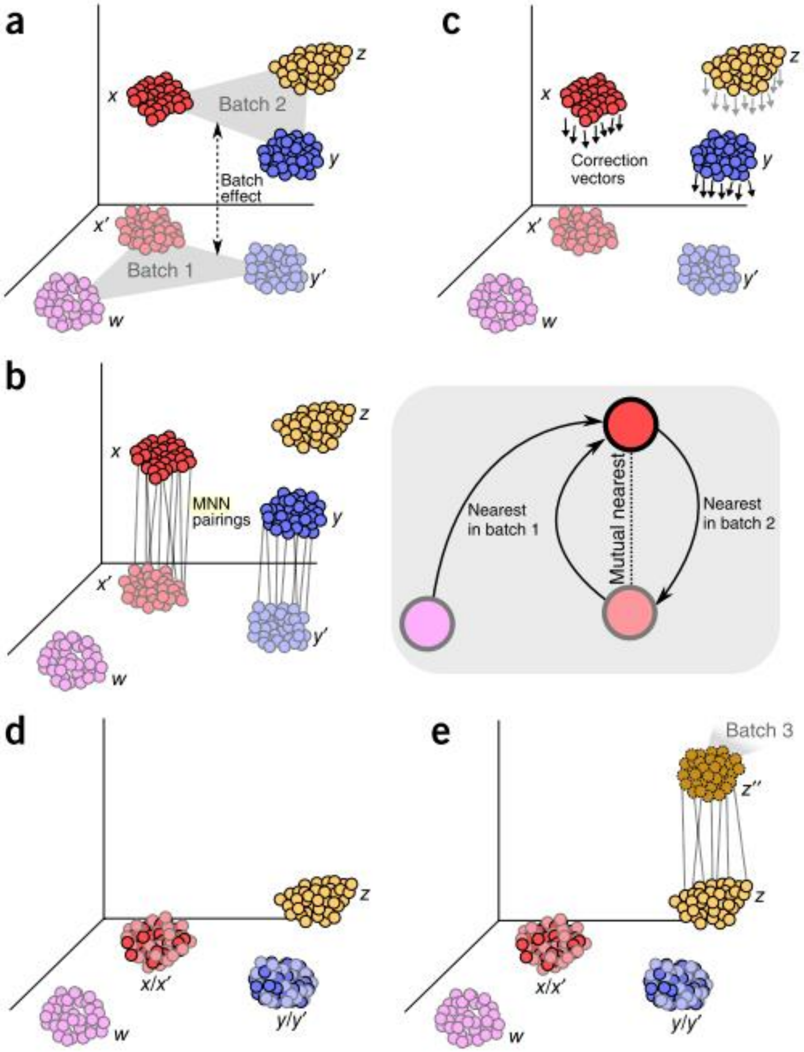
(c/o Dr. Rebecca Ihrie)

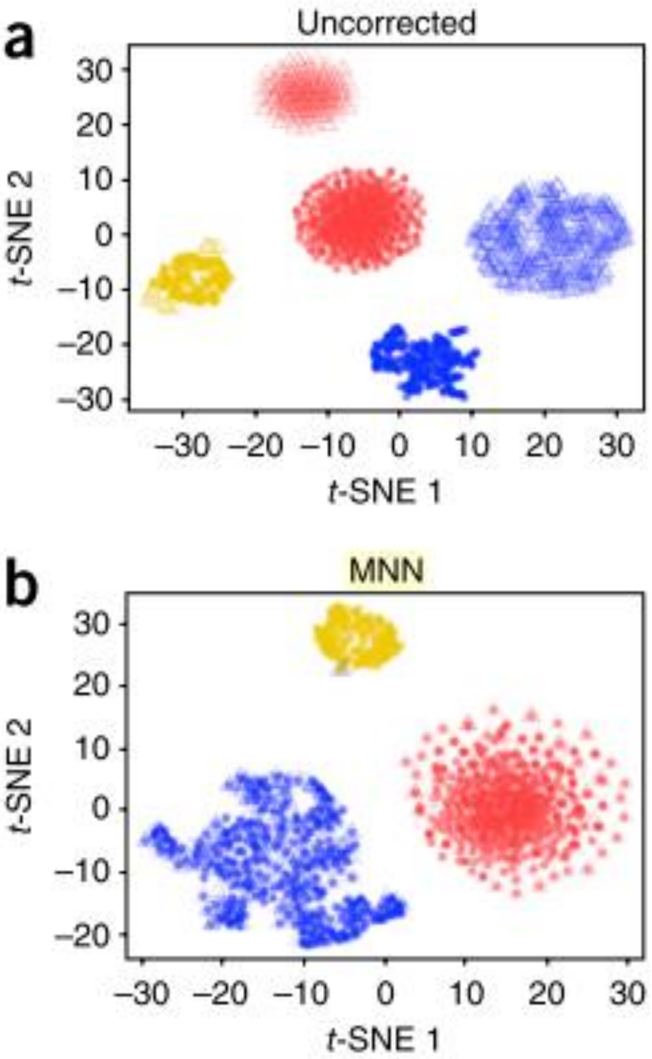
Automatically blasting against reference cell atlases based on shifting branch structures



Batch Alignment

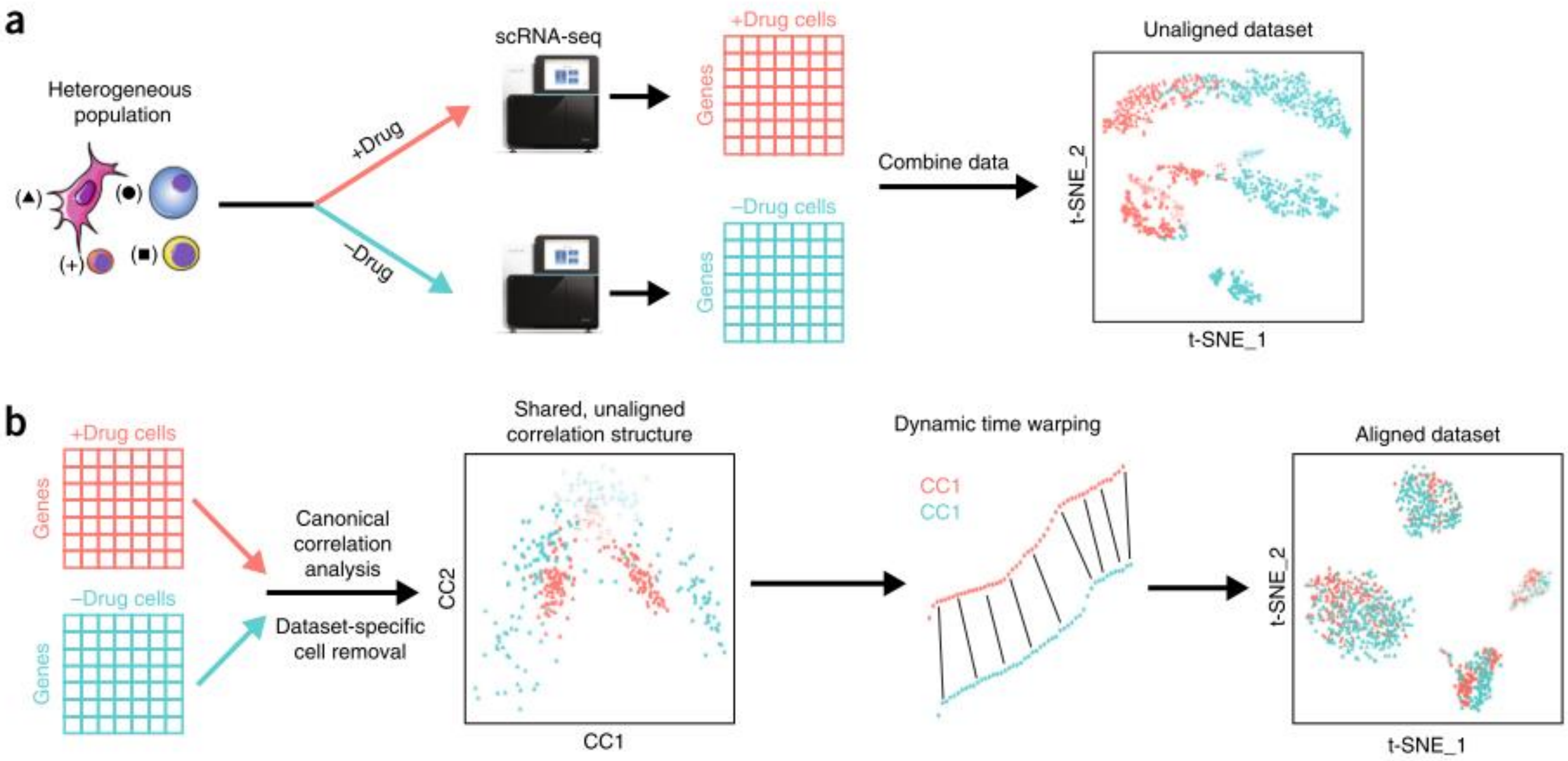
MNN (Haghverdi et al.)





Using mutual nearest neighbors to correct for batch

Seurat-CCA (Butler et al.)



Canonical Correlation Analysis

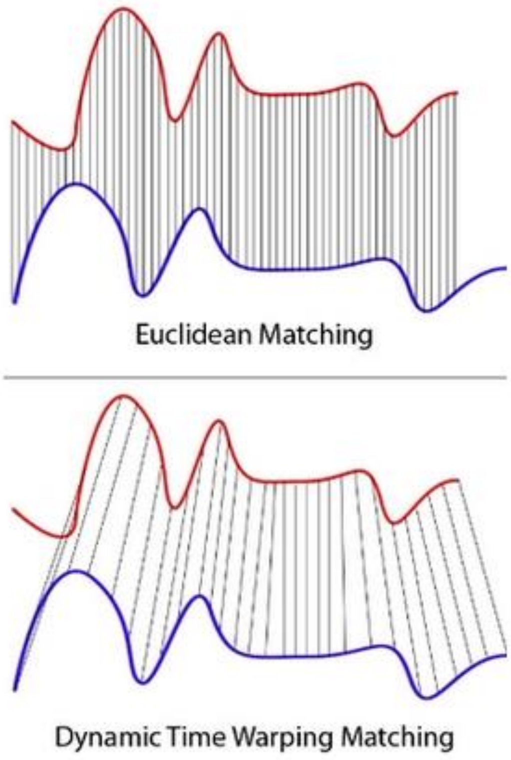
- Instead of looking at mutually neighboring cells, look for gene patterns that are maximally correlated between data sets (basically identifying a set of marker genes for populations that do not change over batches)

- Use these genes for data embedding (like PCA, orthogonal axes etc.)

Alignment via dynamic time warping

- Linear transform the CCA axes initially, then scale according to dynamic time warping (DTW)

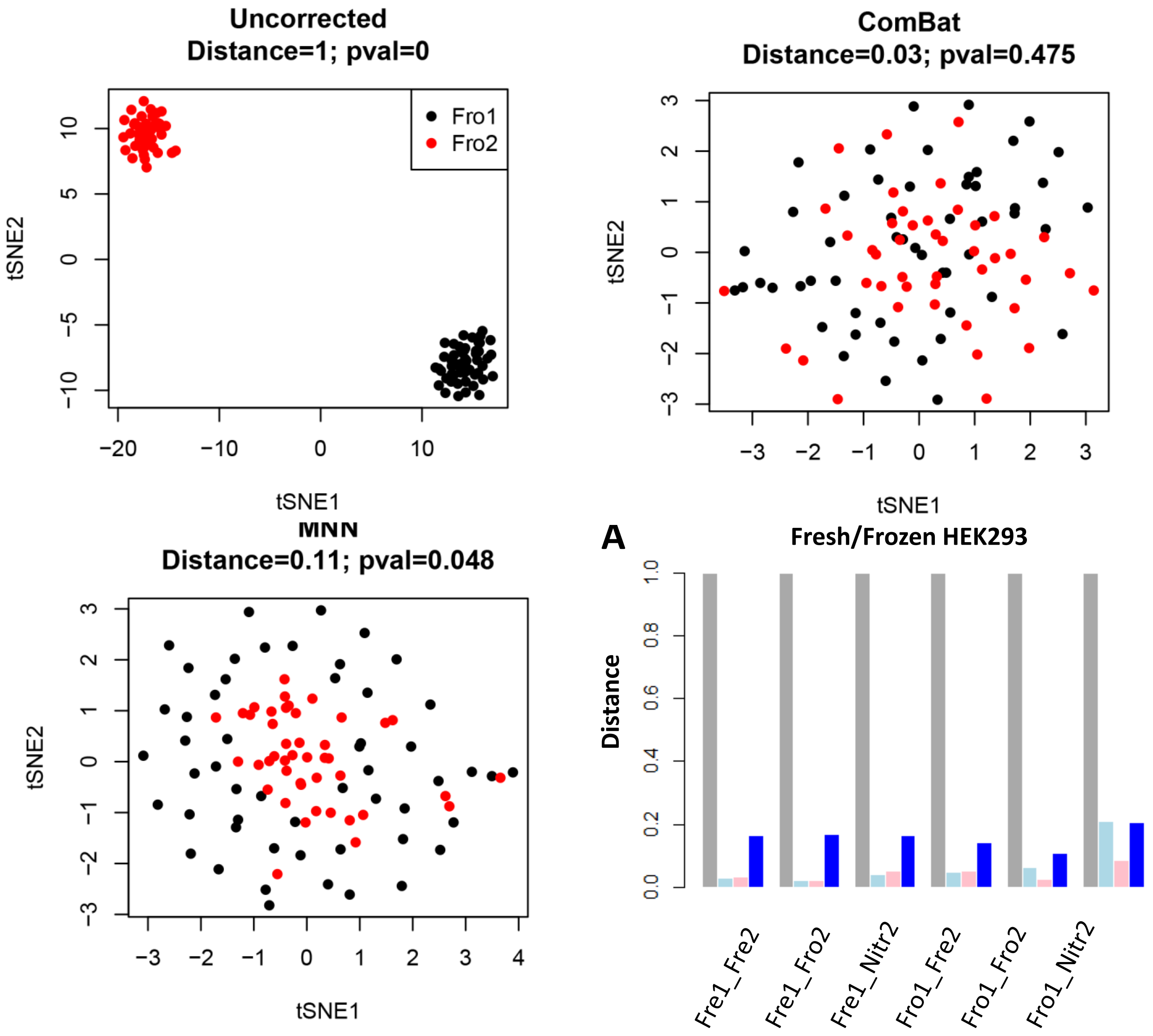
- DTW takes two sequences of signals that maybe out of sync with different "acceleration" and matches them, producing a new aligned axis



Euclidean Matching

Dynamic Time Warping Matching

Assessing alignment/mixing using sc-UniFrac



Uncorrected
Distance=1; pval=0

ComBat
Distance=0.03; pval=0.475

MNN
Distance=0.11; pval=0.048

A
Fresh/Frozen HEK293

Summary

- Significant batch effects exist in scRNA-seq data
  - Differences in cell isolation
  - Differences in library preparation
  - Differences in sequencing

- Alignment can be used for correcting batch effects

- Alignment can be extended to different experimental conditions (untreated and treated experimental samples)

- One must be able to quantitatively assess quality of alignment. AKA to address the question of “Would anything align to anything”?