

Lab 8



Single-cell RNA seq

Maarten de Vries

slides adapted from Qing Zhang, Aviv Regev

exercises adapted from Harvard Chan Bioinformatics Core

Outline

- HW5 on scRNA-seq released, due April 11th
 - scRNA analysis with Seurat
- Today:
 - Brief review of scRNA-seq
 - Hands-on practice with Seurat
- Office hours
 - Maarten: Saturday 3/27 & 4/3 @ 10:30 am + Thursday @ 8 pm
 - Philip: Friday 3/26 & 4/2 @ 4:30 pm

Why scRNA-seq



Bulk genomics

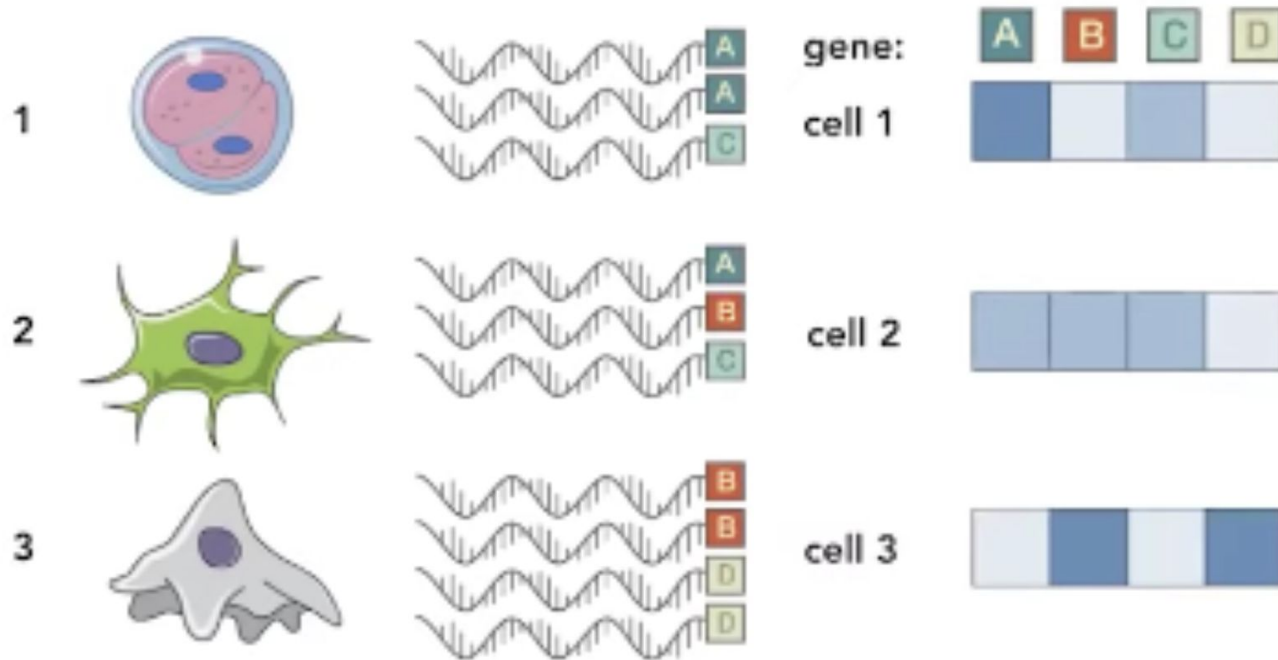


Single cell genomics

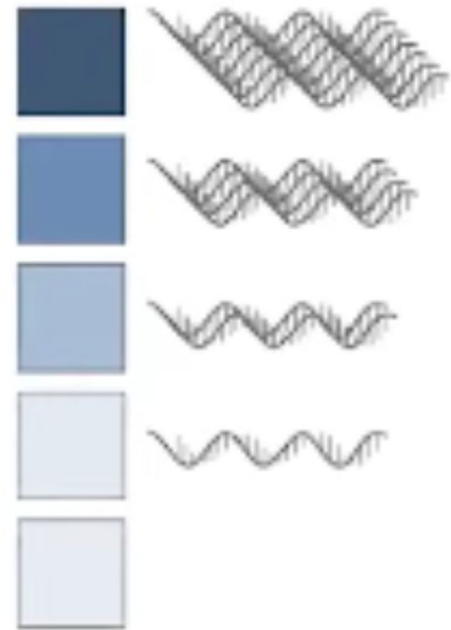
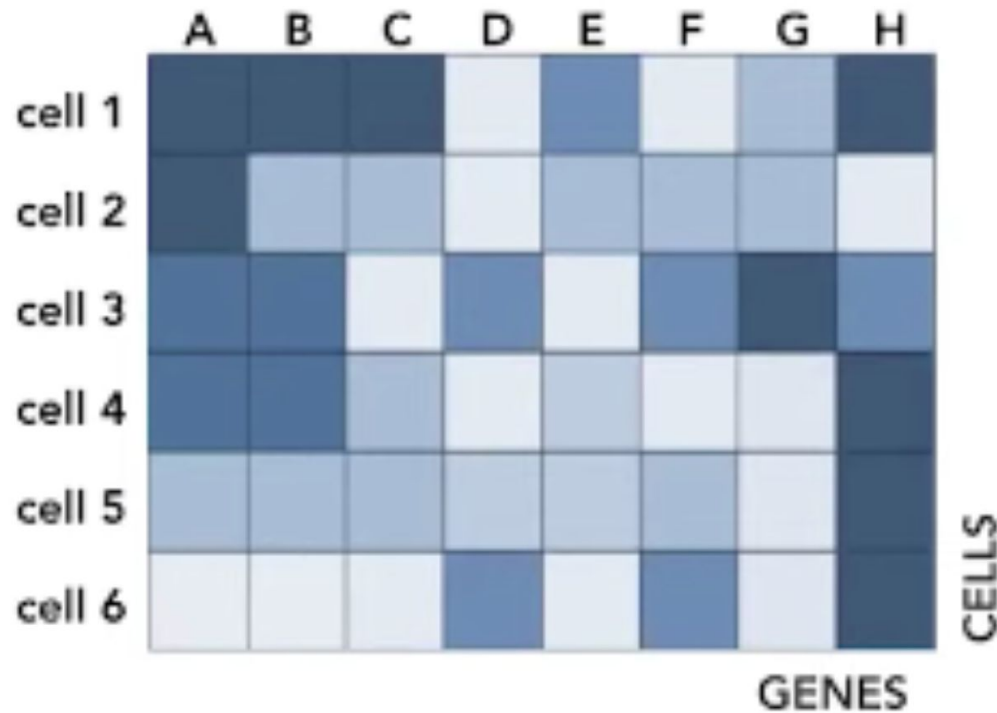


Spatial genomics

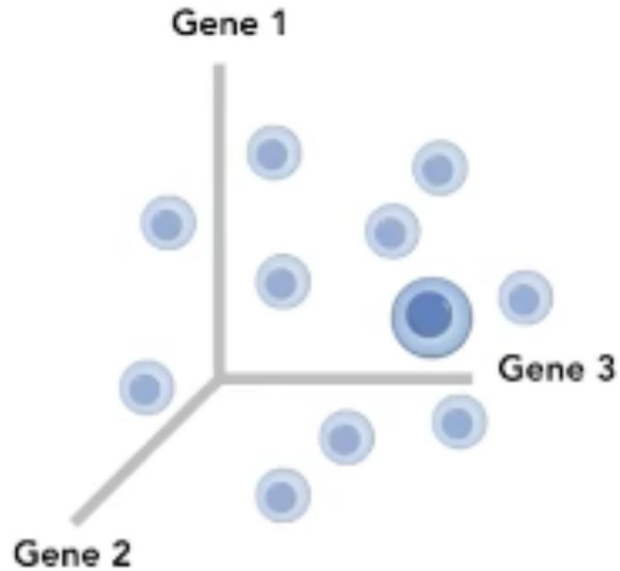
Cells express the genes they need at the right quantities

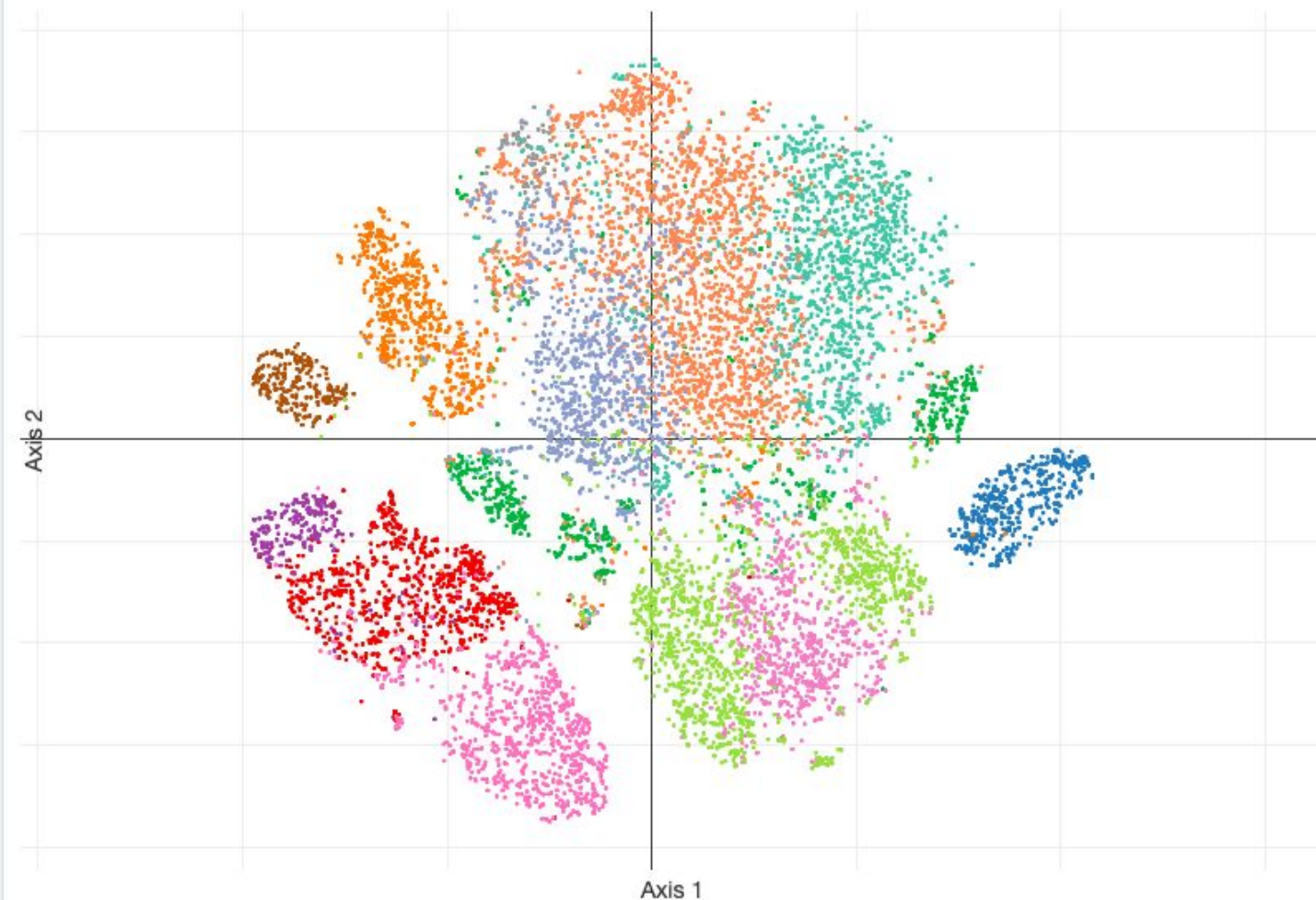


Gene expression profile is the calling card of a cell



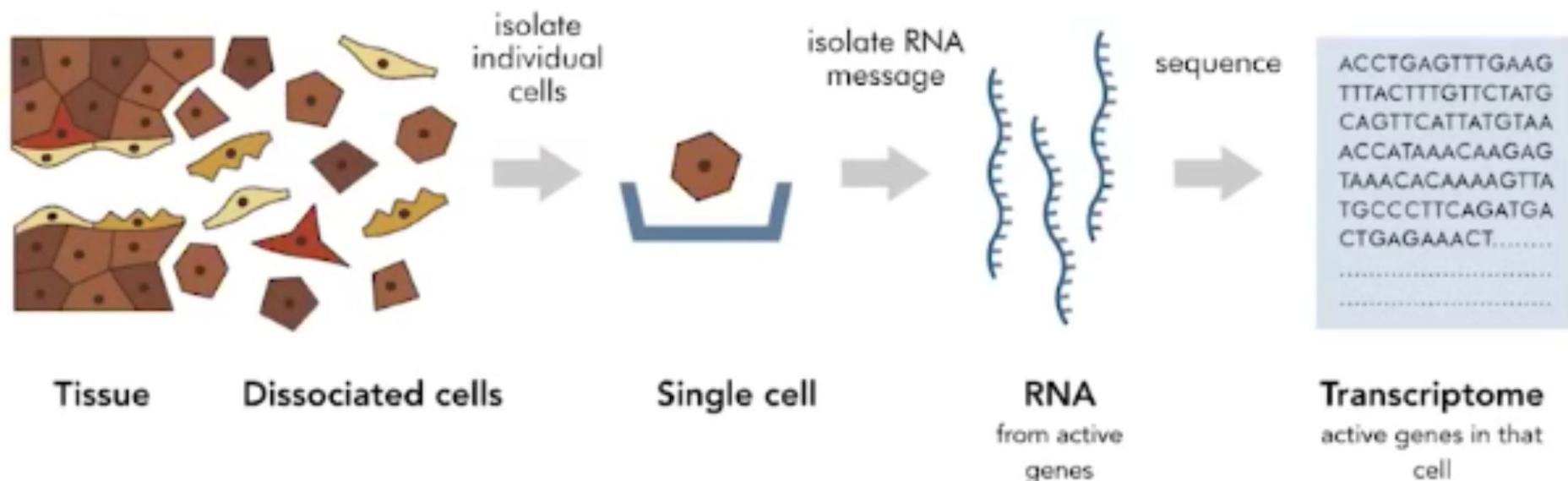
Cell as a point in 20,000+ dimensional gene expression space



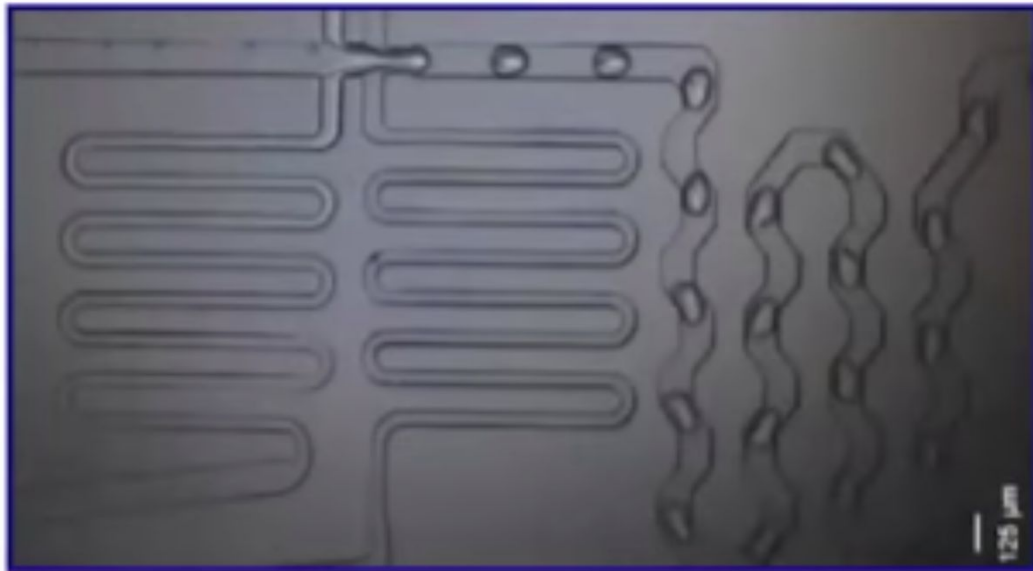


- Endothelial (773 points)
- Glia (415 points)
- Inflammatory Fibroblasts (715 points)
- Microvascular (208 points)
- Myofibroblasts (606 points)
- Pericytes (226 points)
- Post-capillary Venules (756 points)
- RSPO3+ (118 points)
- WNT2B+ Fos-hi (1349 points)
- WNT2B+ Fos-lo 1 (2008 points)
- WNT2B+ Fos-lo 2 (952 points)
- WNT5B+ 1 (736 points)
- WNT5B+ 2 (1138 points)

Capturing a single cell

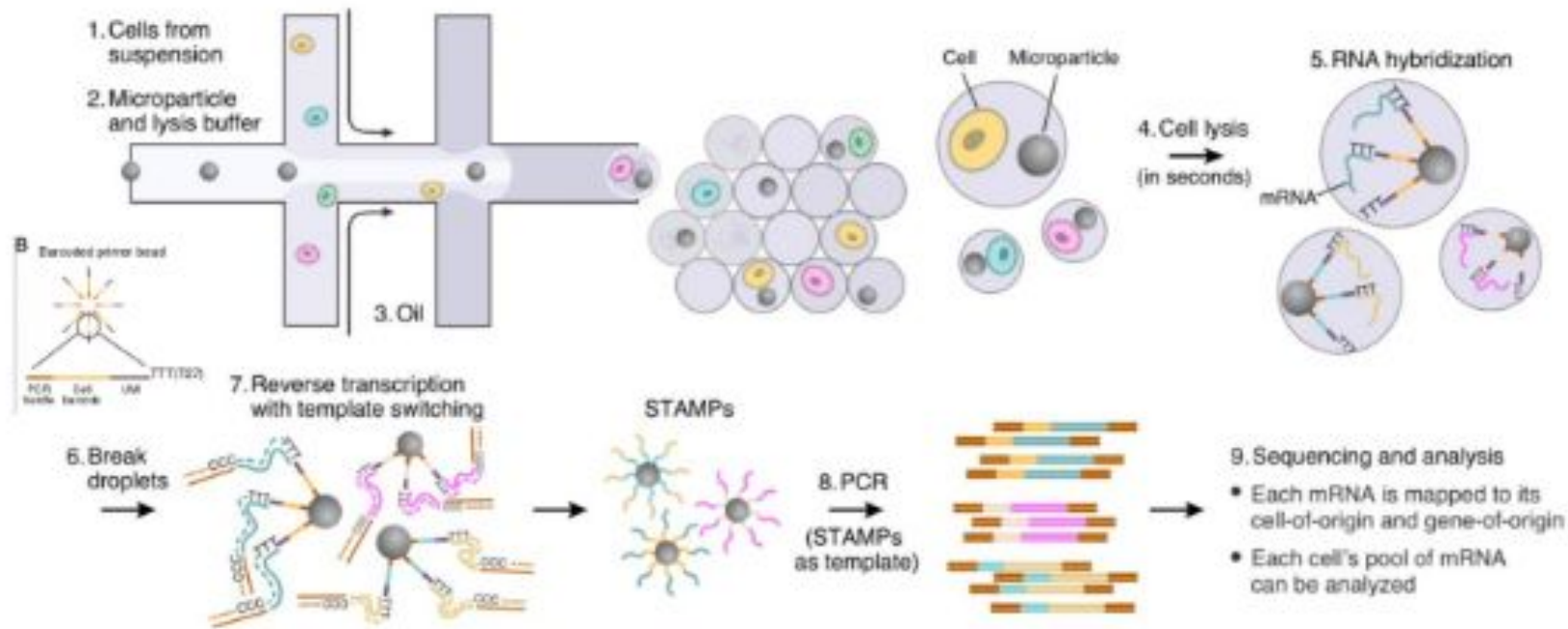


Droplet sequencing

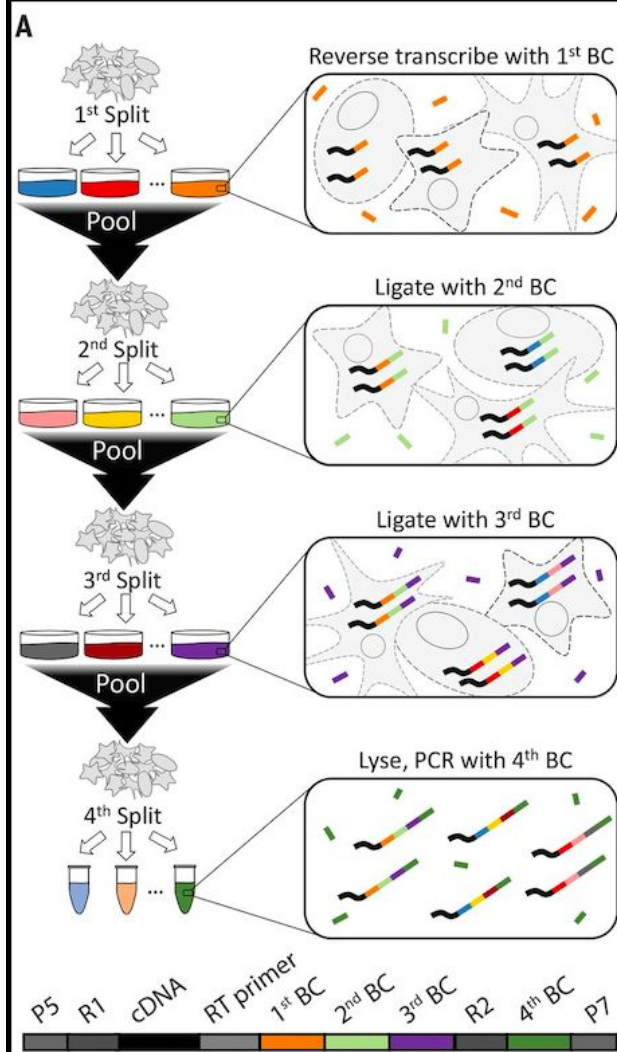


5,000 cells/second

Droplet sequencing



SPLIT-seq



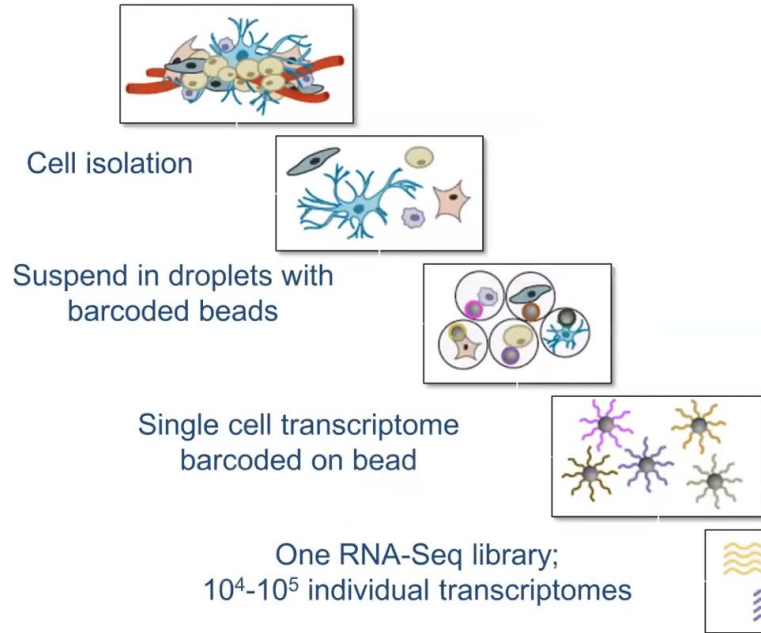
SPLIT-seq: the movie



and prepare more containers with

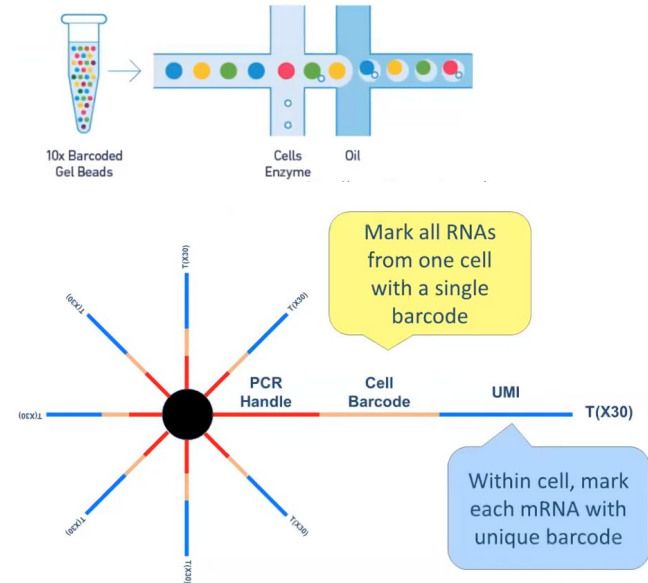


Recap: Single cell RNA seq



Macosko et al, *Cell*, 161, 2015

10x Genomics - Microfluidic droplet-based scRNA-Seq



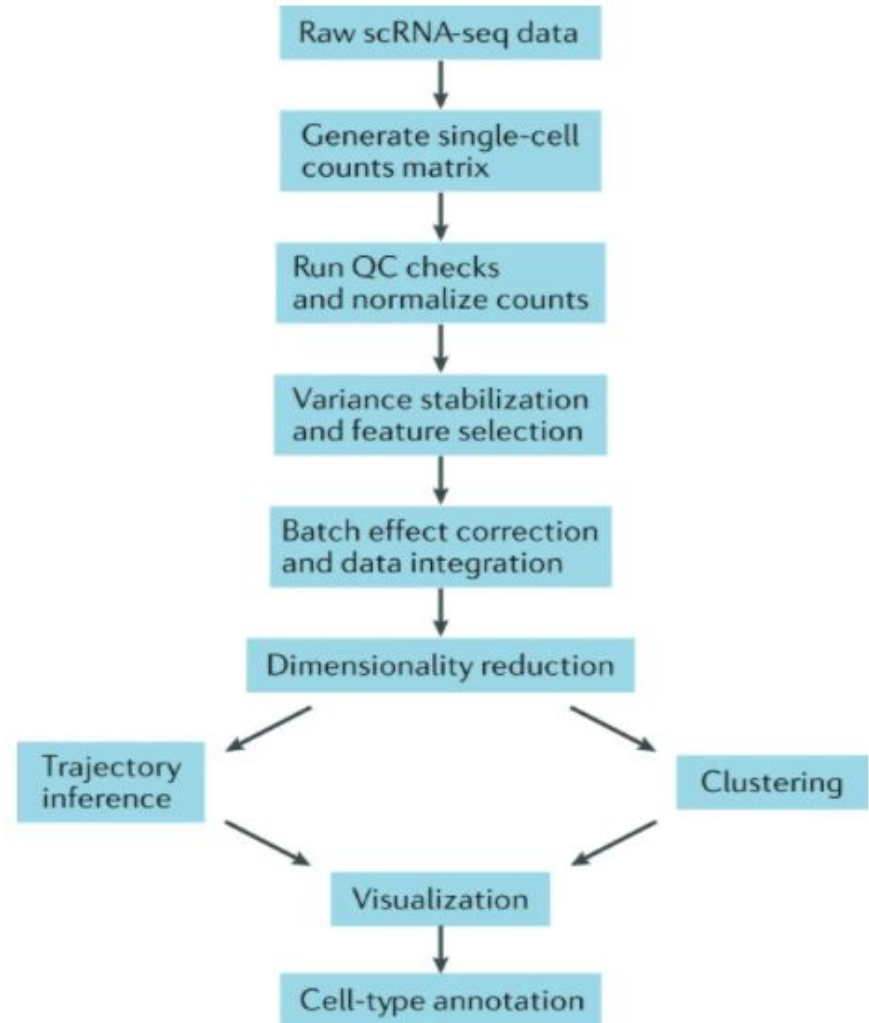
UMI: distinguish multiple copies of a transcript vs PCR artifacts

scRNA-seq analysis

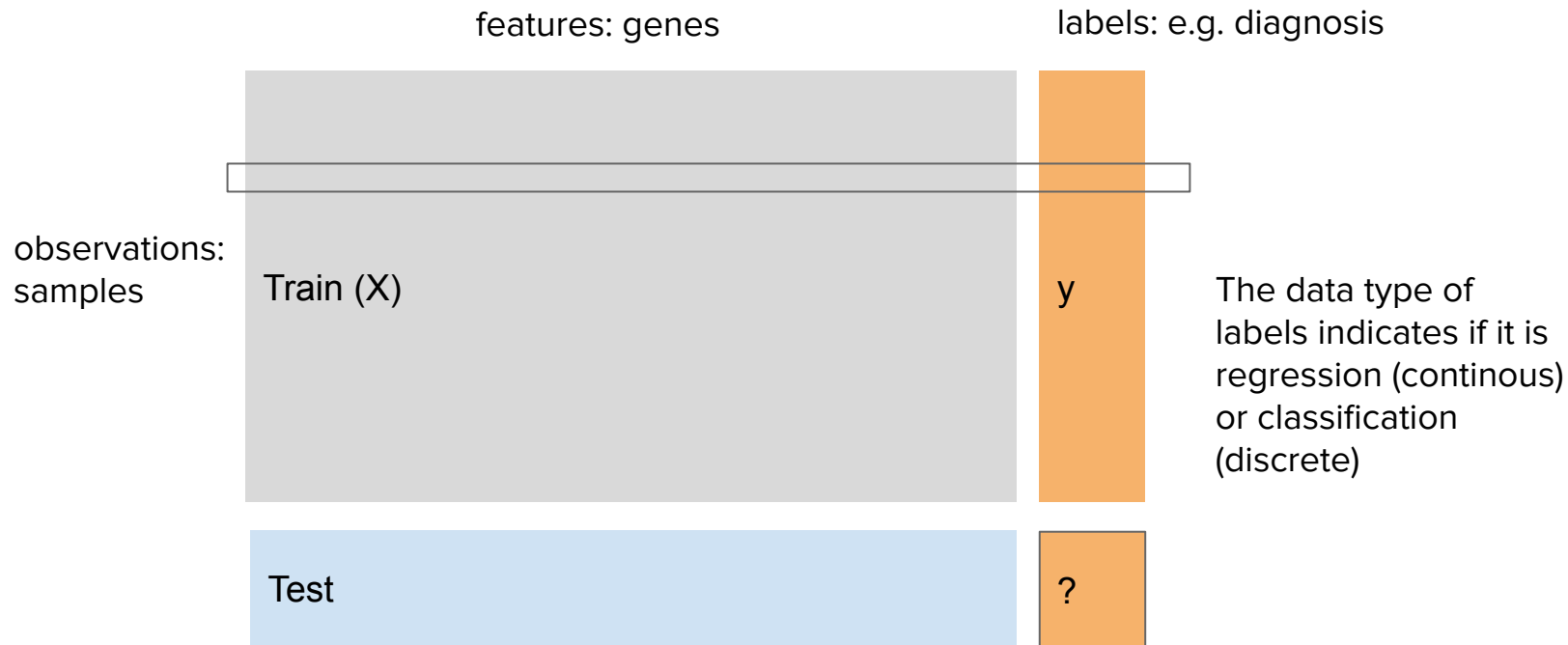
Wu & Zhang

Nature Rev Neph 2020

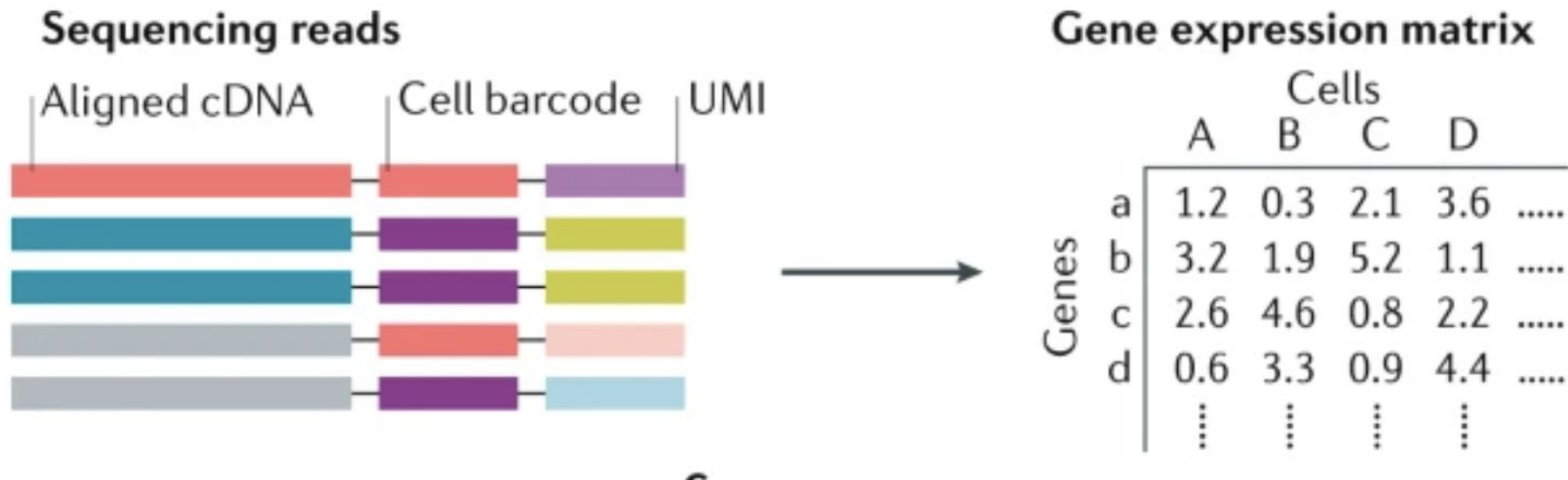
Fig. 1: Overview of the single-cell RNA sequencing analysis pipeline.



Dataset

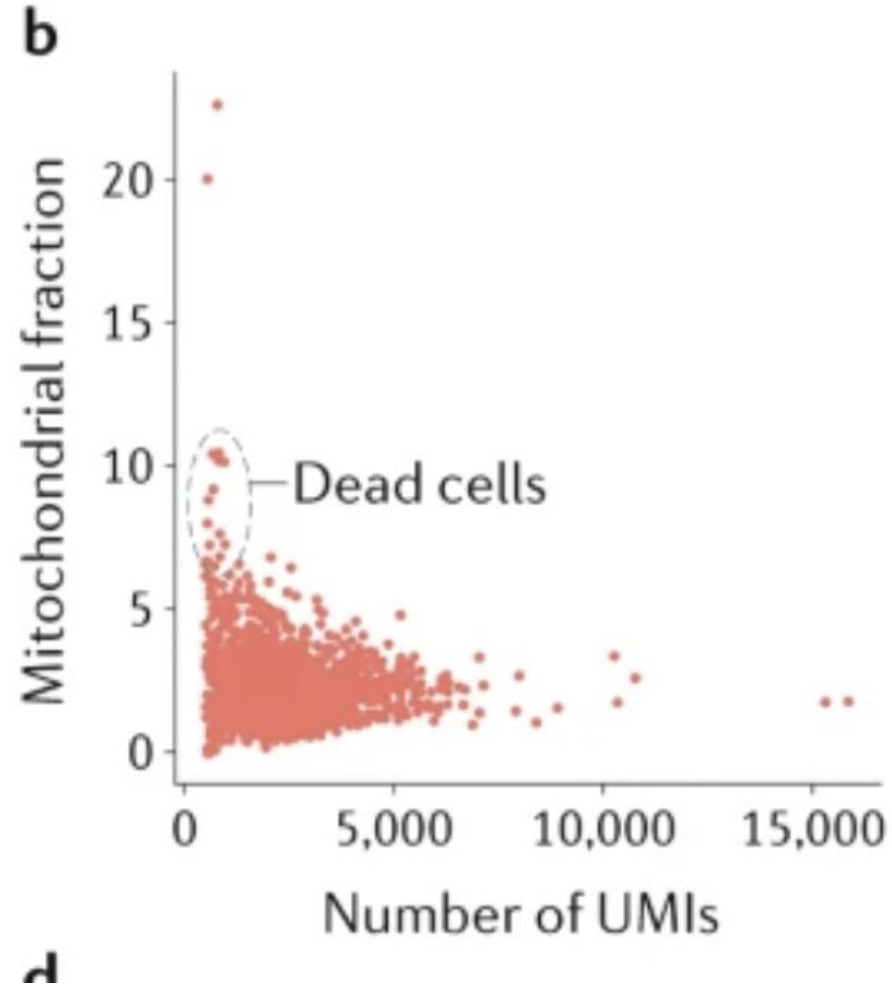


Generating the counts matrix

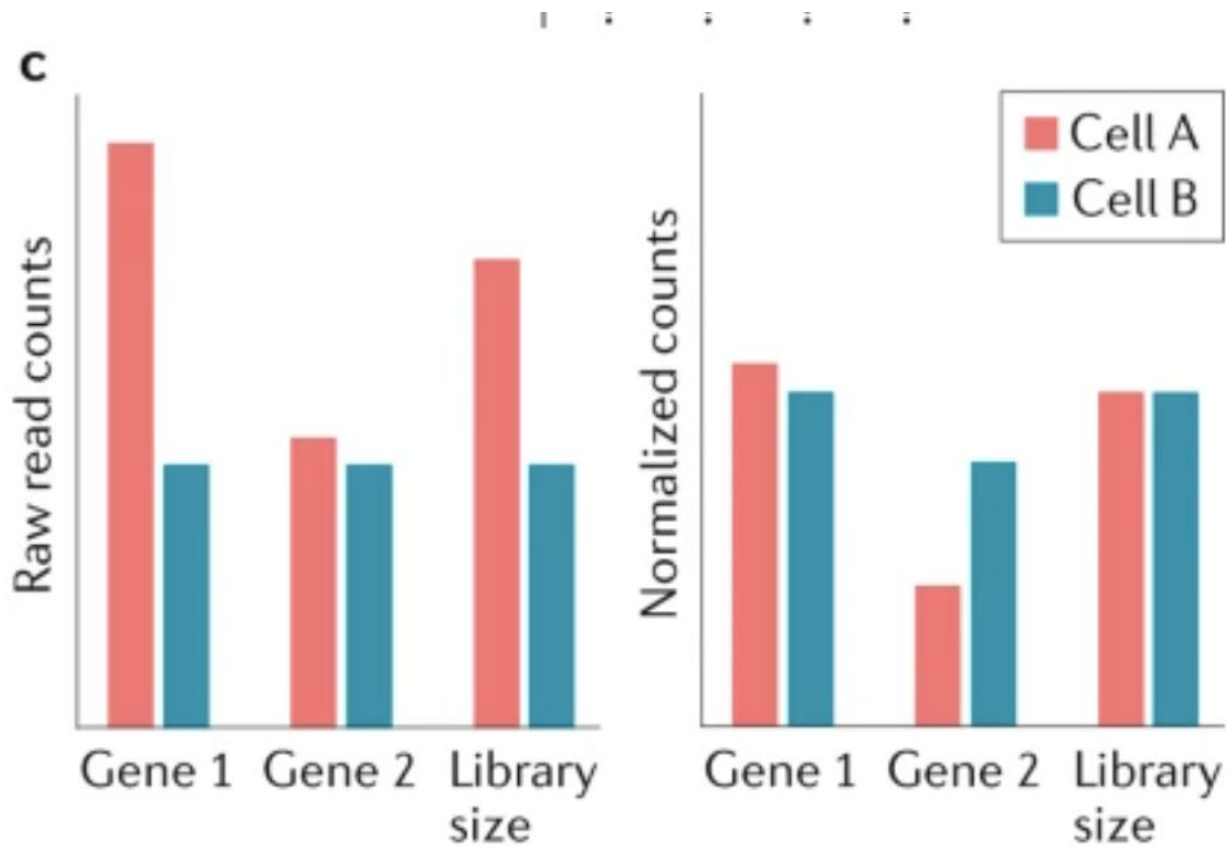




- Detect cells with
 - Low read count
 - Low % of mapped reads
 - Too few/many genes
 - High mitochondrial fraction



Normalization



Dimensionality reduction

- Dimensionality reduction
 - Curse of dimensionality
 - Linear: Principal Component Analysis (PCA)
 - Finds linear combinations of genes that best capture the variance in data
 - Zero-inflated factor analysis (ZIFA) is a version of PCA designed to explicitly model the high expected count of zero values in scRNA-seq data.
 - Nonlinear: t-SNE, UMAP, Deep neural networks
- Imputation of zero-values
 - Technical limitations in RNA capture lead to zero-inflation
 - In part biological variance, in part noise
 - Imputation methods are available
 - MAGIC uses information from neighboring cells
 - Lots of deep learning methods recently developed

Clustering & Visualization

Fig. 4: Cell clustering in datasets with discrete cell types.

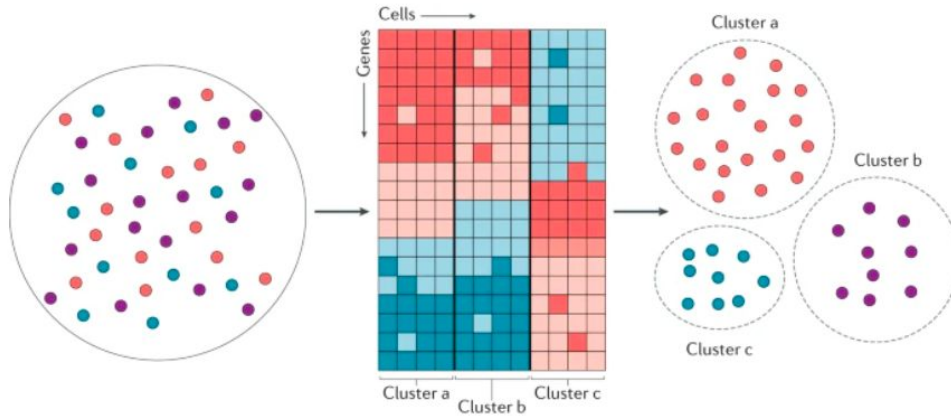
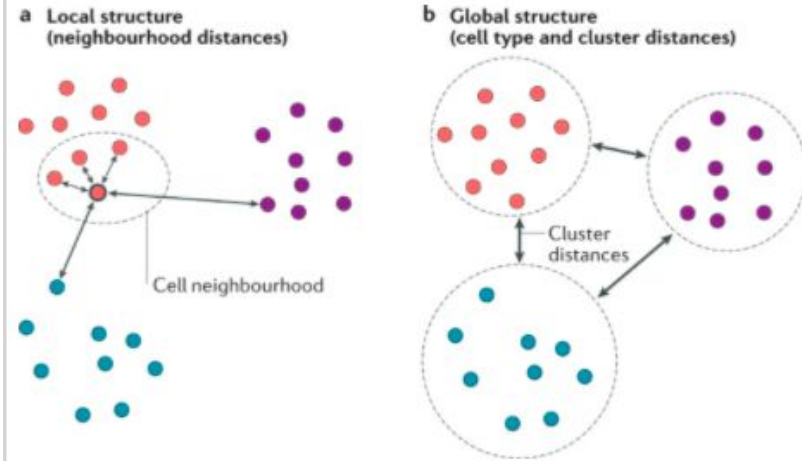


Fig. 6: Local and global structure in a dataset.



Cell-type annotation

- Time-consuming if done manually
- Find genes that are uniquely expressed in each cluster
- Match those genes to lists of canonical cell-type markers
- To accelerate this, use functional pathways and gene ontology terms
- New (semi-)automated cell-type classification methods are being developed.
- Novel cell types and states still need to be manually annotated, or do they?

Workflows

Conveniently, there exist toolkits that enable all of the aforementioned steps within a single workflow.

- Seurat (in R)
- SCANPY (in Python)
- MAESTRO (Philip's Lab next week)

Seurat [v4.0]

<https://satijalab.org/seurat/>

R package for single cell analysis

Great vignettes!

Make sure you have Seurat 4.0 and
R version >4.0 installed



Further reading

https://hbctraining.github.io/scRNA-seq_online/schedule/