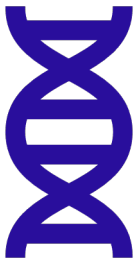
Analysis and interpretation of single-cell RNA-seq data Part I



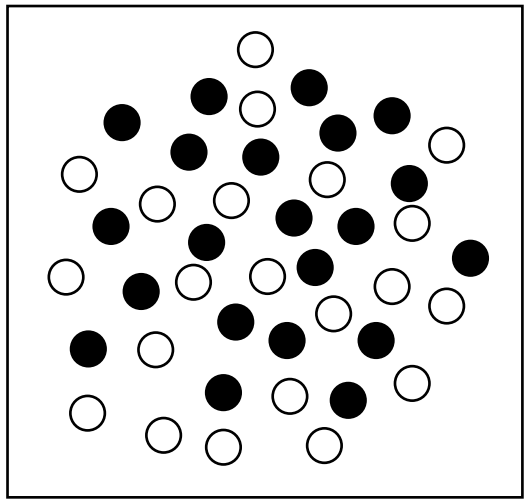
Bioinformatics Workshop
February 15, 2021

Allegra Petti, PhD
Departments of Neurosurgery, Medicine, and Genetics
allegra.petti@wustl.edu

Single-cell RNA-seq captures expression heterogeneity

Identifies and counts unique transcripts in each cell

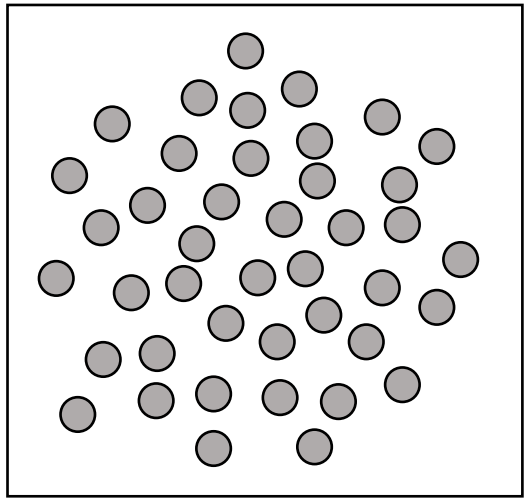
For one gene:

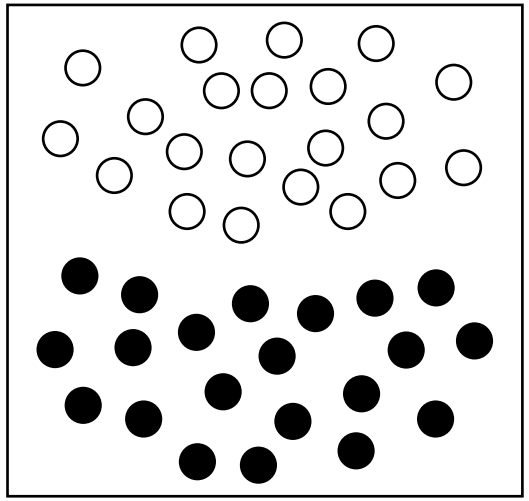


Gene X ON
Gene X OFF

Bulk

Single-cell





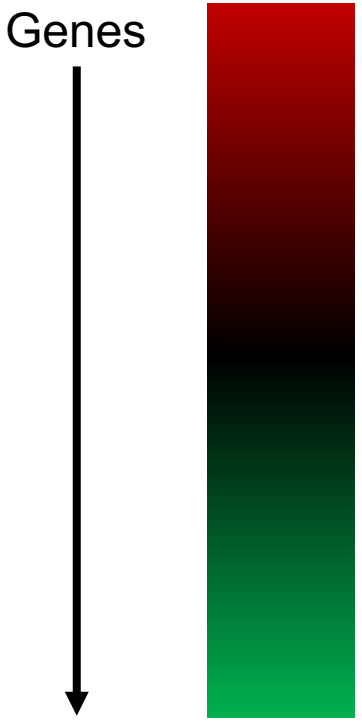
Bulk RNA-seq averages across the population

scRNA-seq reports per-cell expression and enables computational "sorting"

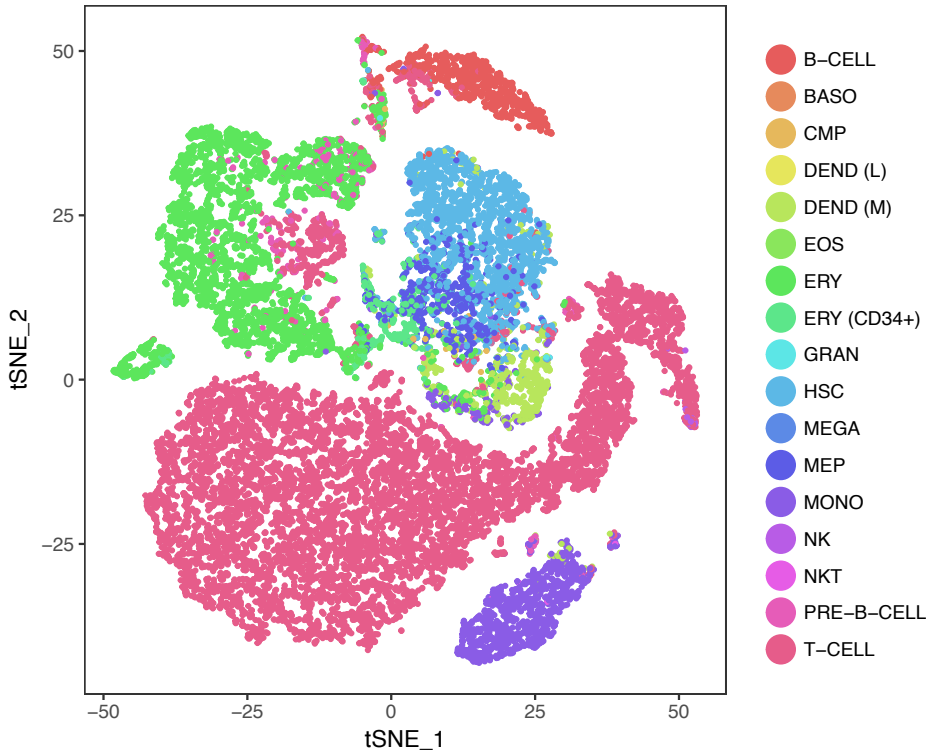
Single-cell RNA-seq captures expression (and genetic) heterogeneity

Identifies and counts unique transcripts in each cell

Bulk RNA-seq

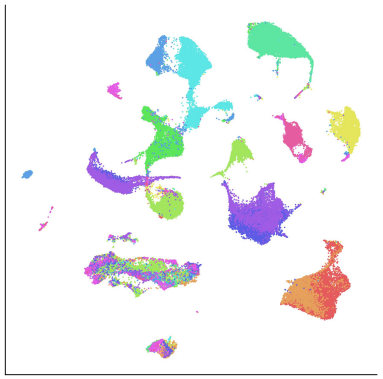


Single-cell RNA-seq
(tSNE/UMAP plot)

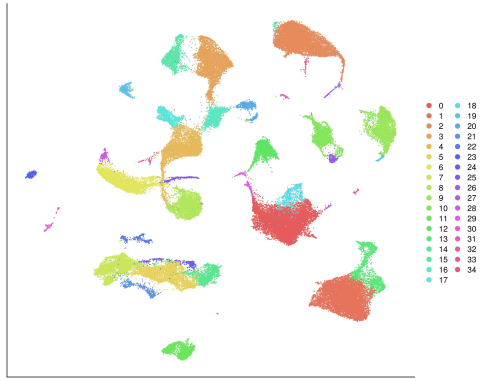


For many genes, multiple samples:

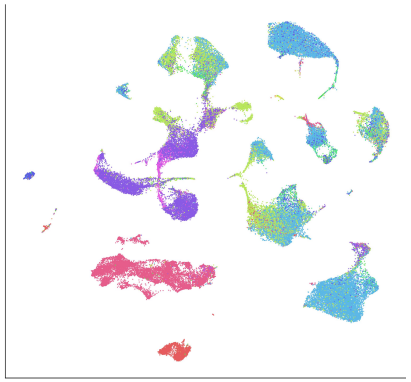
By Sample:



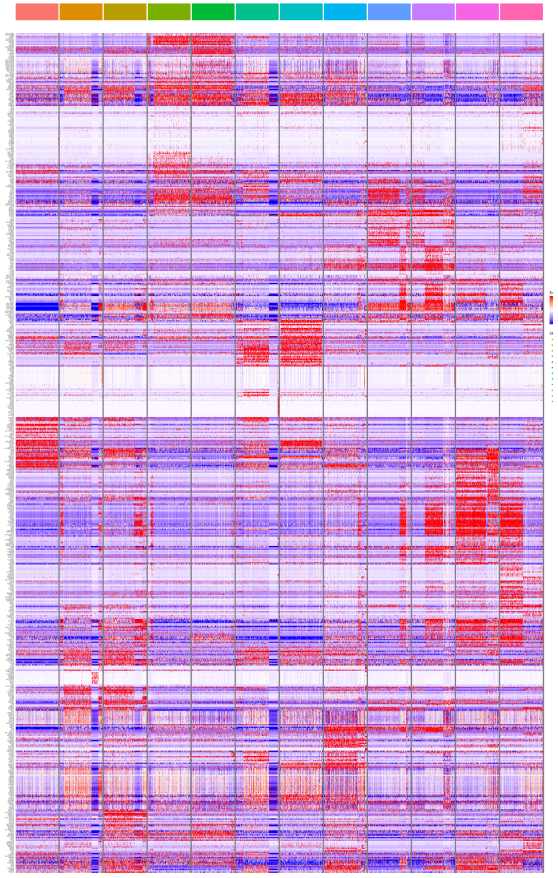
By Cluster:



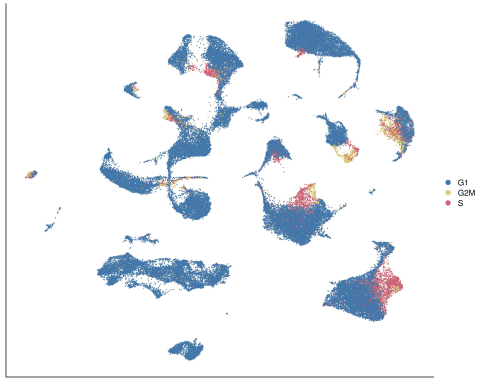
By Cell Type:



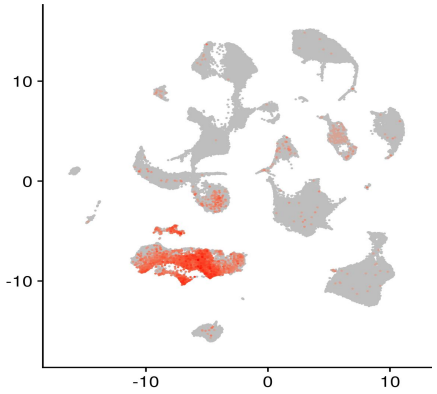
By Gene, Cell, and Sample:



By Cell Cycle Phase:



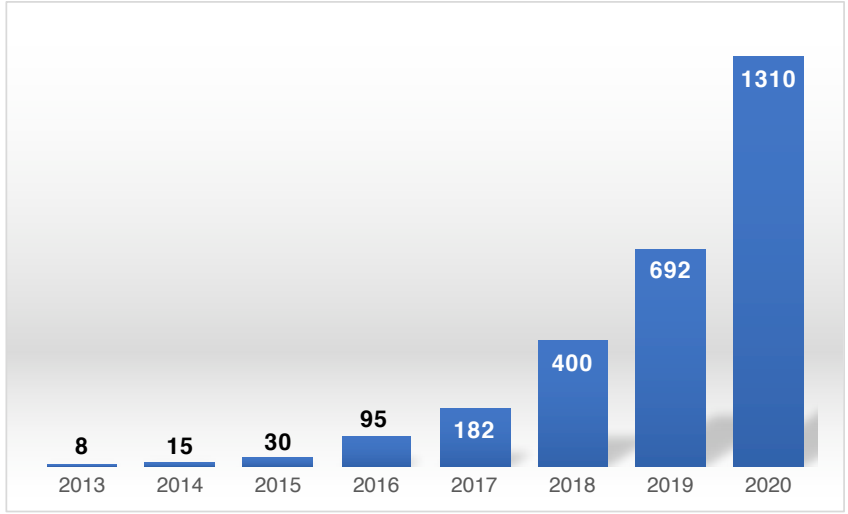
By Gene:



A new era in biology and medicine?

Community Goals

- Redefine cell “type”
- Redefine relationships among cell types
- Catalog all cell types in all diseased and normal tissues
- Discover/define new cell types



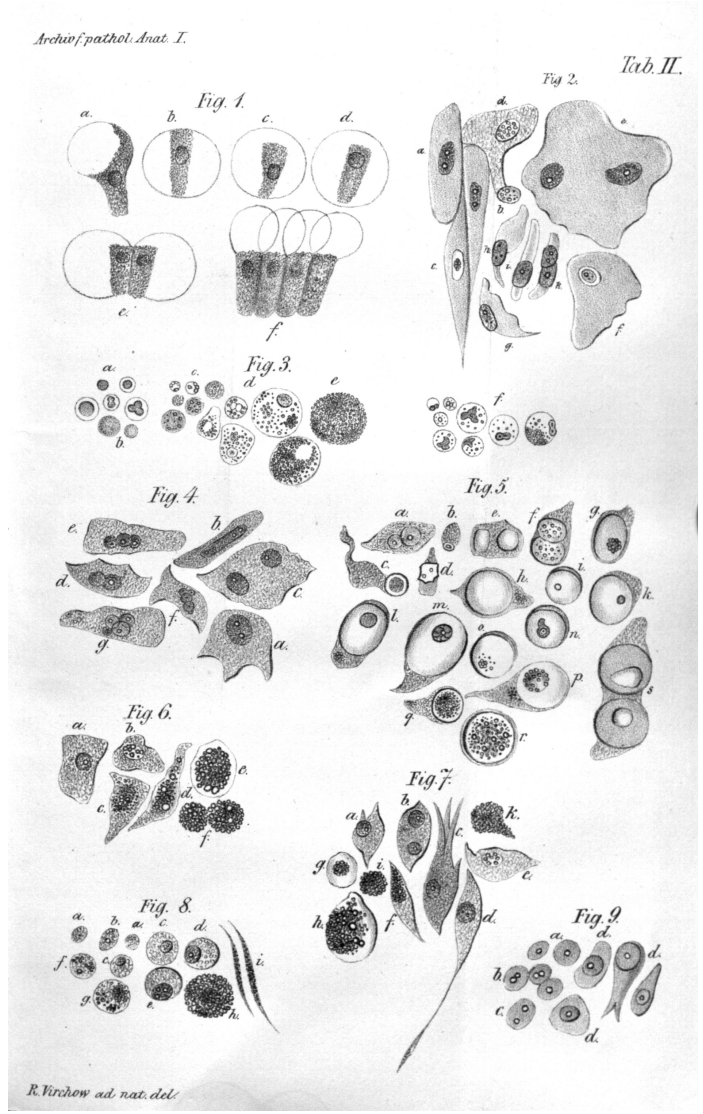
Personalized medicine

- Variation at the level of the individual - *between* individuals
- scRNA-seq: Variation at the level of the cell - *within* AND *between* individuals
  - High-resolution variation in diseased and normal cell types and states
  - Enables cross-patient correlations to be made at the level of individual cells

scRNA-seq in historical context of cell characterization

Timeline of histology/pathology:

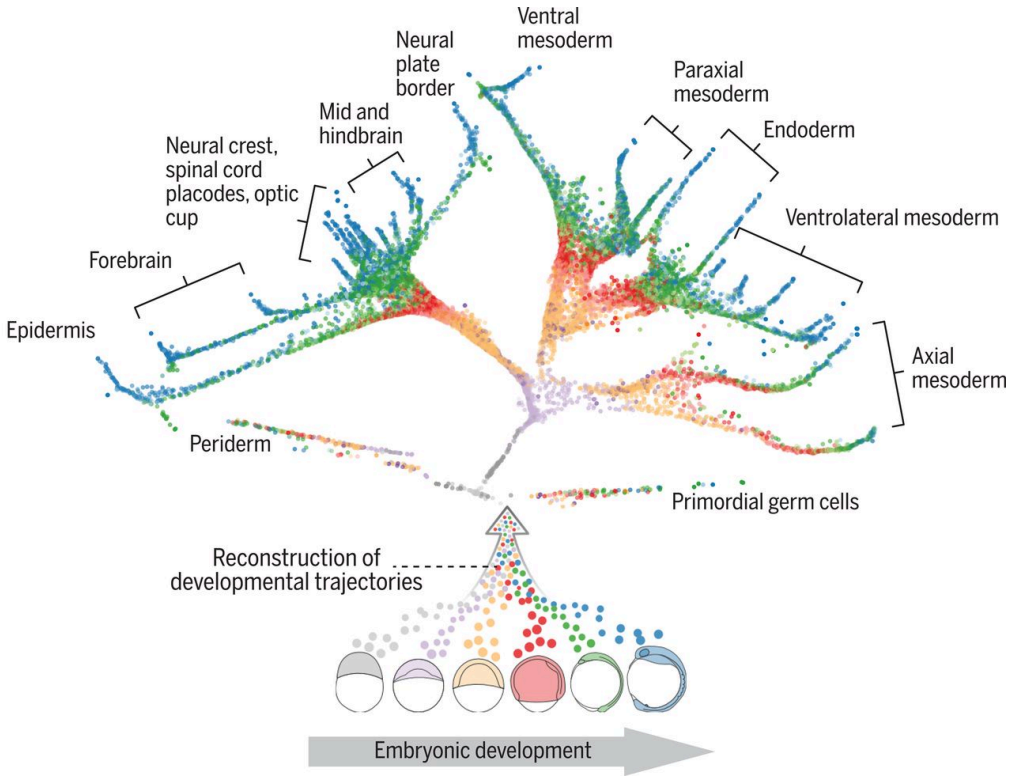
1665	Hooke	Coined ”cell”
mid-1800s	assorted; dye industry	Histological stains
1855	Virchow	Cellular theory
1941	Coons	Immunohistochemistry
1994	Chalfie	Individual cells with GFP
1953 1968	Coulter Fulwyler	Flow cytometry: 17-18 features/cell
2009	U. Toronto, DVS	Mass cytometry (CyTOF): ~100 features/cell
2009-2015	Tang, Klein, Macosko	single-cell RNA-seq: 2-6K features/cell (~20K/sample)



scRNAseq adds nuance to concept of “cell type”

- Cell type (stable, “hard-wired,” e.g. by transcription factors)
- Lineage (often continuous)
- State
  - Variable and continuous
  - reprogrammable, “soft-wired” (e.g. by environment)
  - Normal range of cell states vs. pathological range

High dimensional scRNA-seq data permits detailed analysis and reconstruction of cell lineage and state:

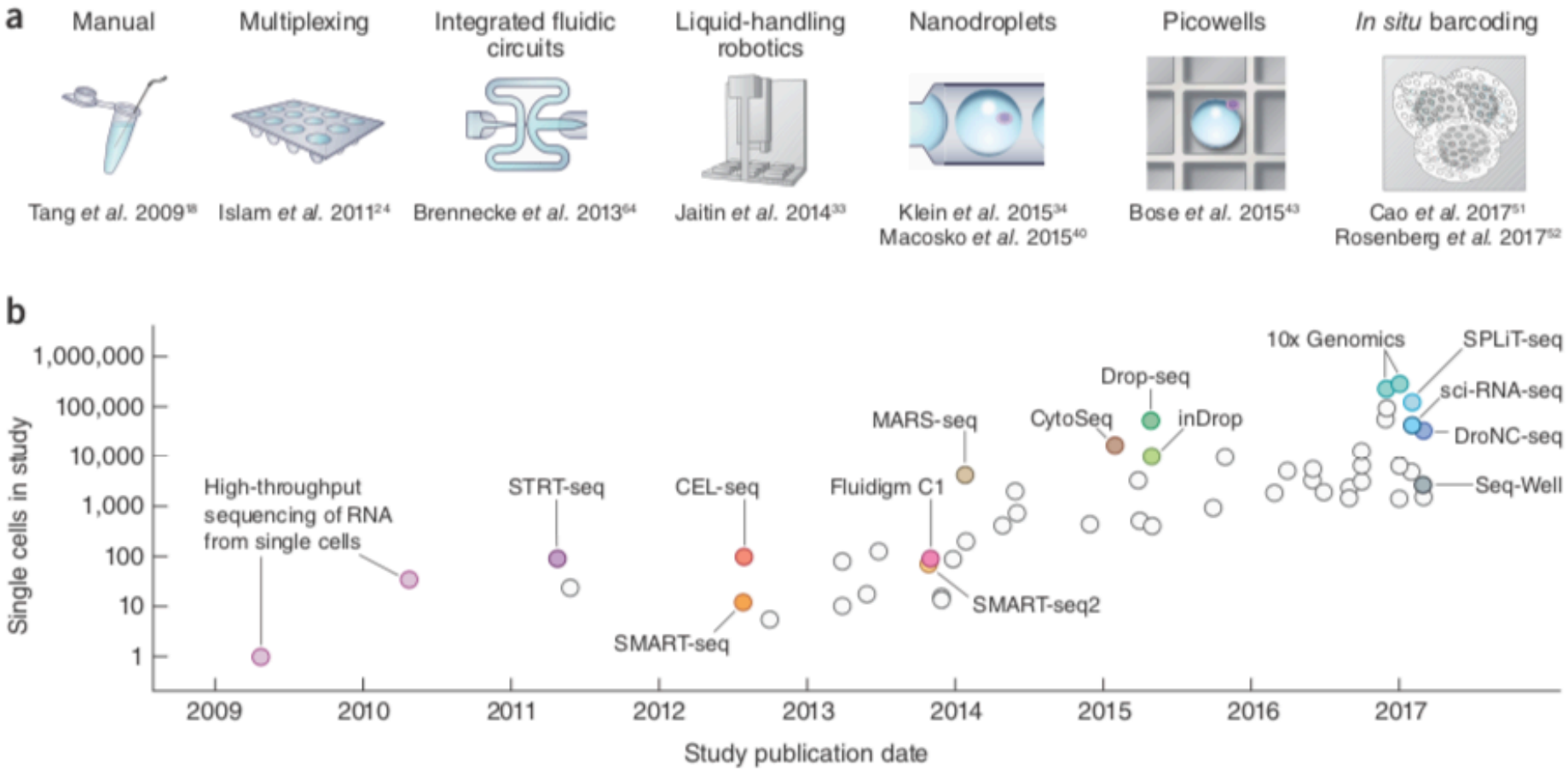


Developmental tree of early zebrafish embryogenesis

Farrell et al. Science 2018

Camp et al, *Science* 2019; S.A. Morris, *Company of Biologists* 2019

Technology Development



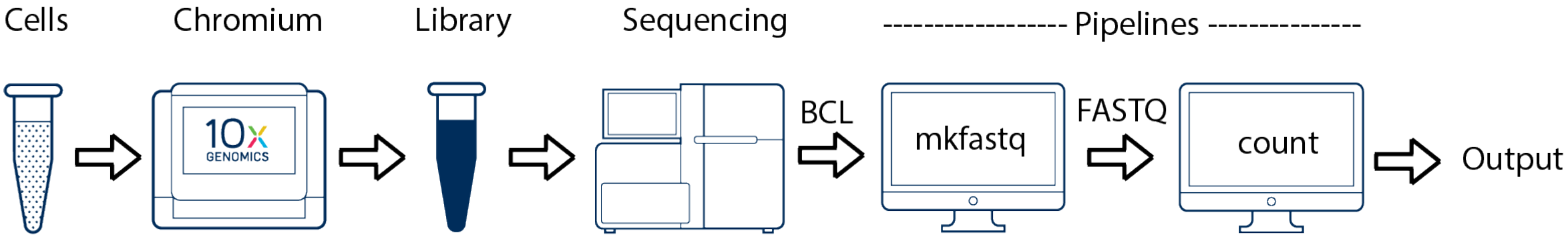
Popular commercial platforms for single-cell/single-nucleus RNA-seq:

10x Genomics vs Fluidigm SMART-Seq (2, 3)

- **Drop-Seq: 10x Genomics**
  - 3’ Gene expression
  - 5’ Gene expression (more sequence; less-biased coverage), TCR/BCR sequencing

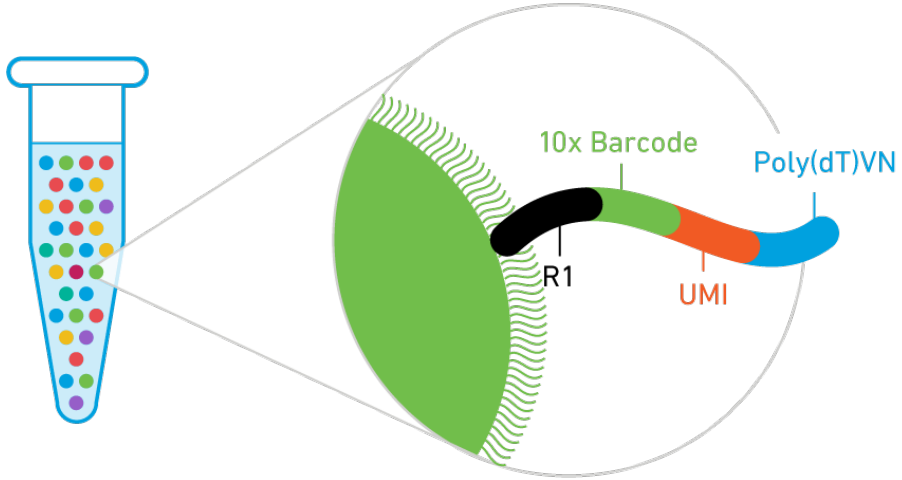
- Plate-based: Fluidigm
  - C1 SMART-Seq2: lower-throughput, more genes/cell, longer cDNAs, no UMIs
  - SMART-Seq3*: lower-throughput, more genes/cell, longer cDNAs (uses UMIs)

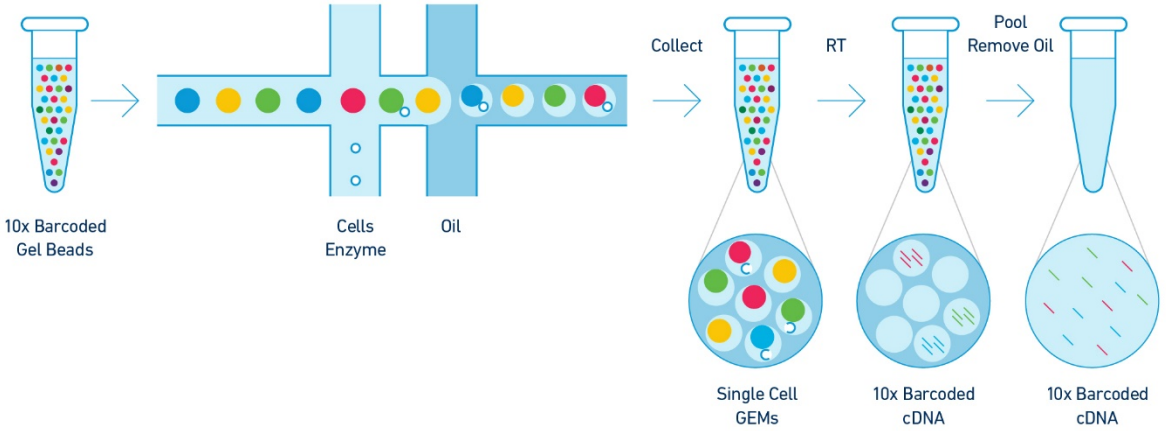
- All have limitations: must choose technology best suited to application
  - Lafzi et al, *Nature Protocols* 13:2742-2757



“Single Cell 3’ Solution” (10x Genomics)

Barcoded bead + cell = bar-coded cDNA library





Pool
(sample index)
Amplify
Fragment*
Add primers

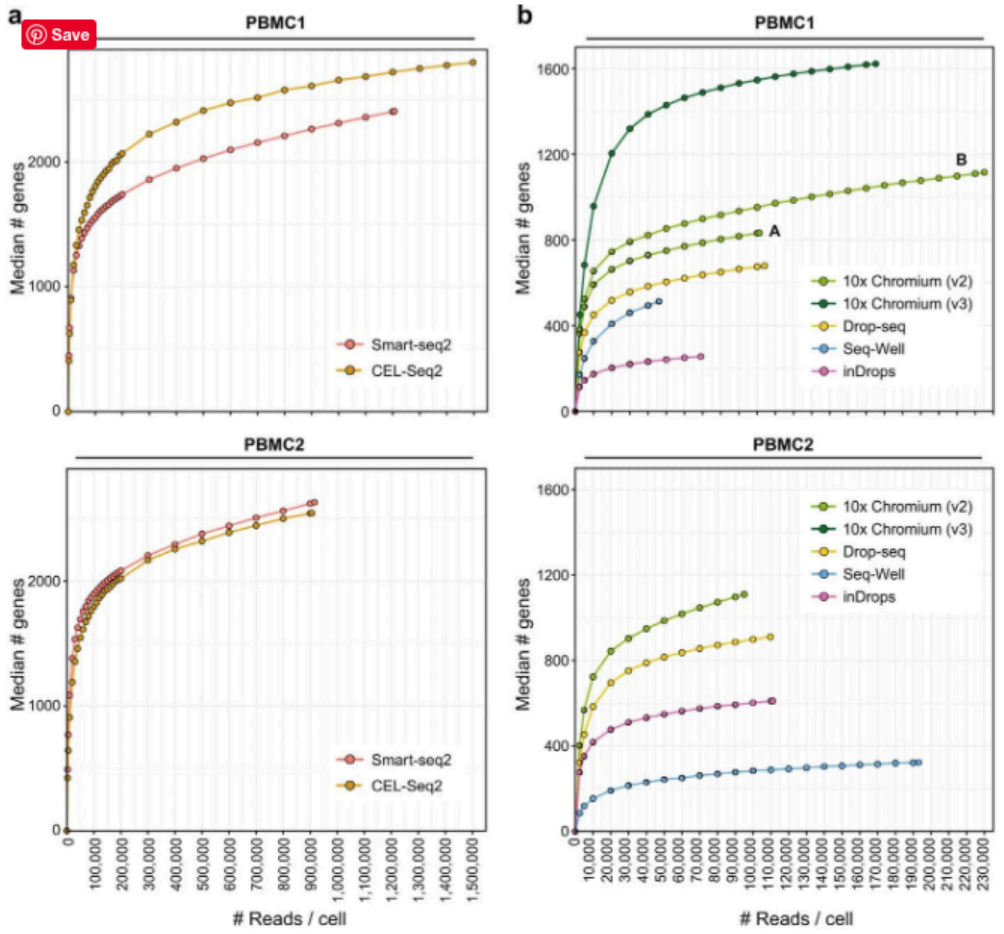
Sequence

How deeply do you need to sequence?

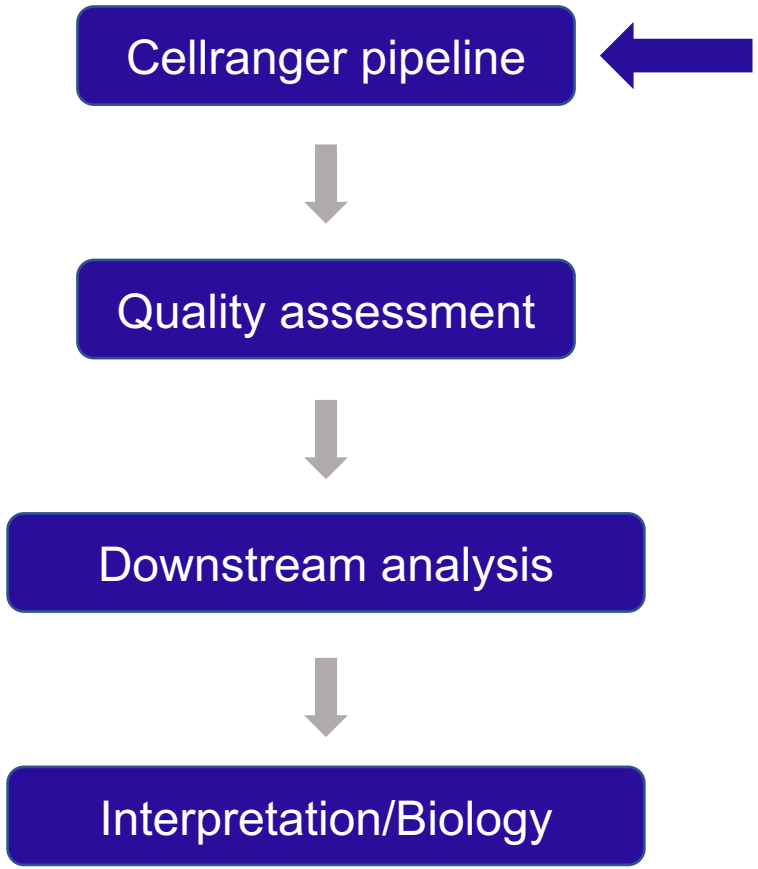
Rule of thumb:
Achieve 90% saturation

Official Recommendations (reads/cell):

3’ V3: 20K
3’ V2: 50K
5’ V1.1 20K
5’ with variant identification: 200K
5’ V(D)J: 5K
Higher for cell lines



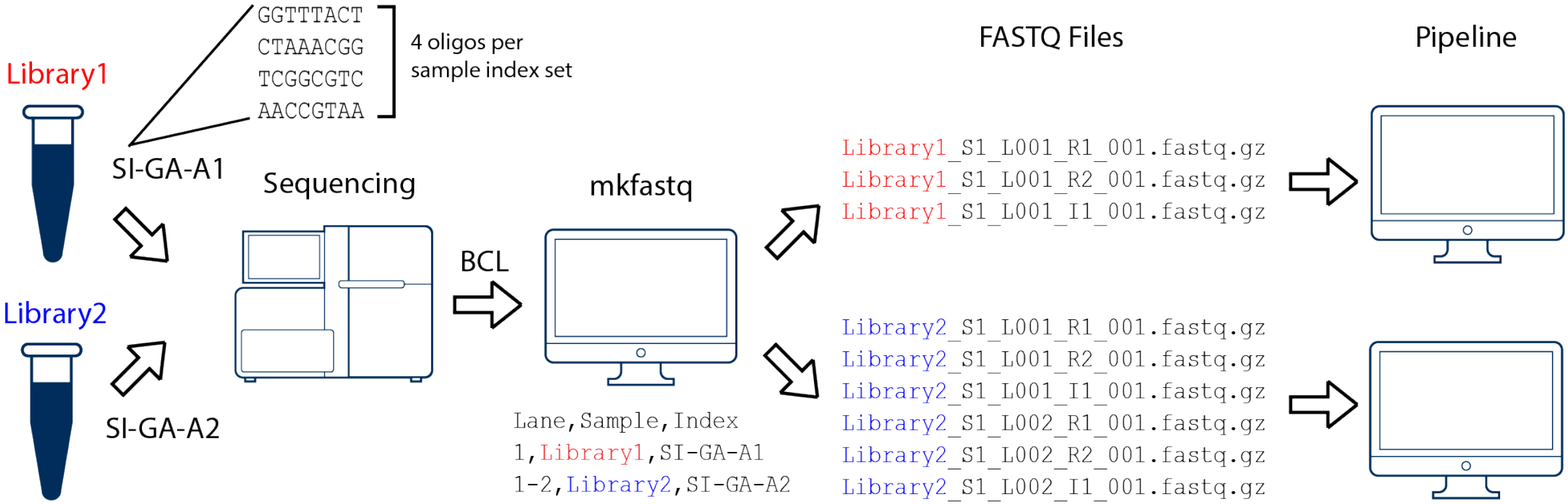
Post-sequencing workflow



Alternative pipelines:

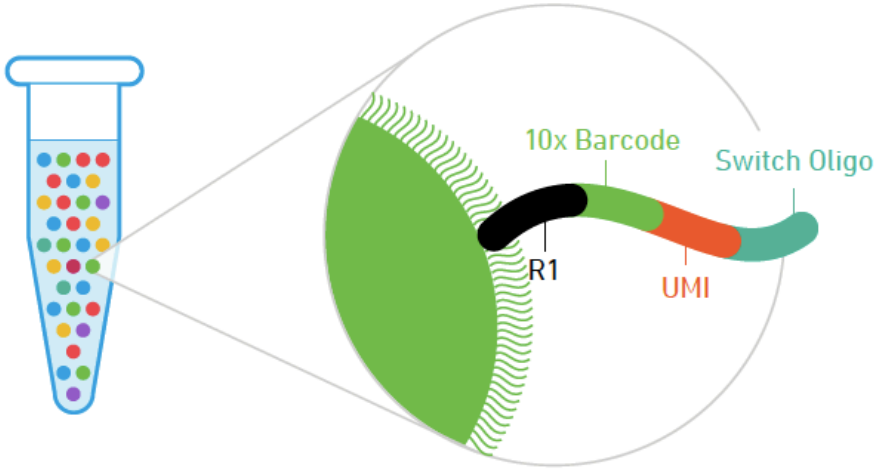
- kallisto bustools: https://www.kallistobus.tools/
- scumi: https://bitbucket.org/jerry00/scumi-dev/src/master/

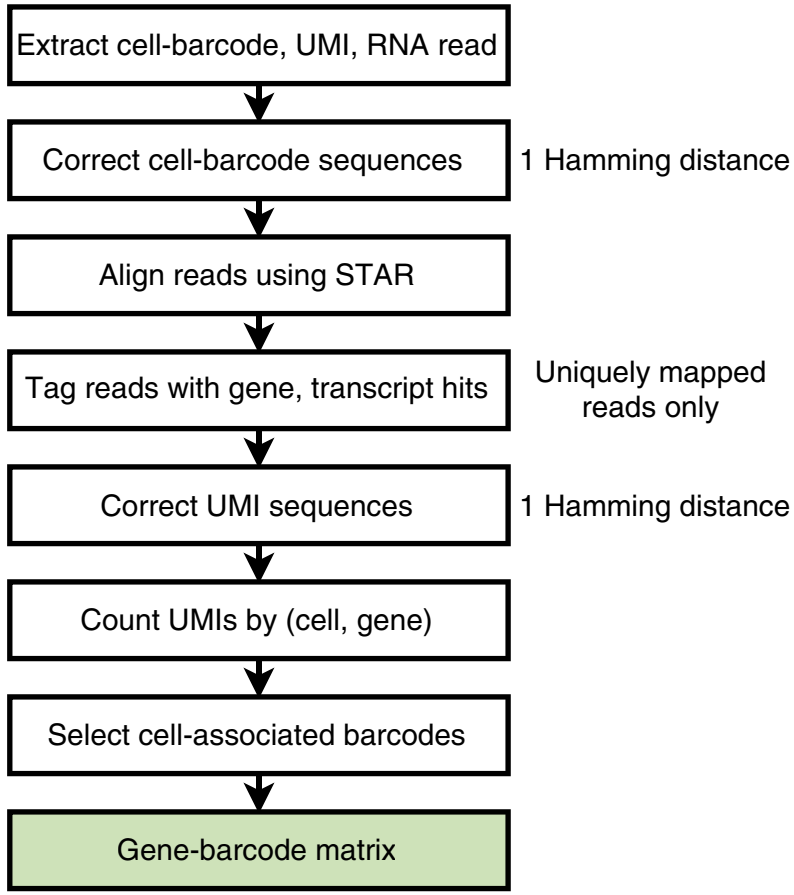
Cellranger Step 0: Sample Demultiplexing



https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/using/mkfastq

Cellranger Overview

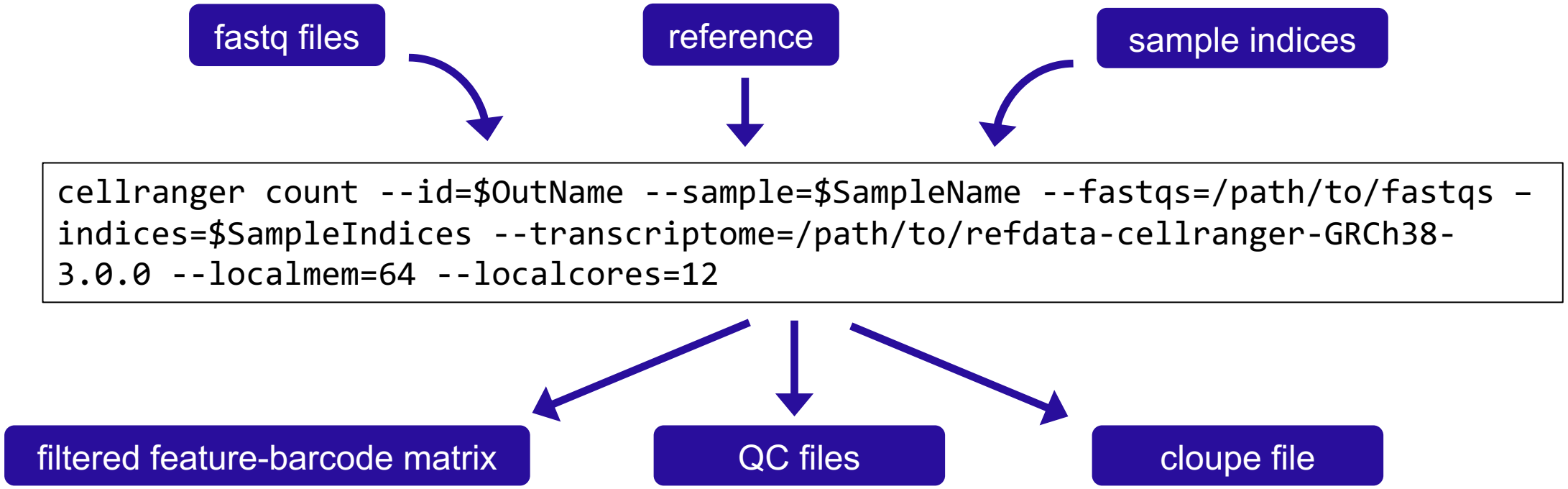




Zheng, et al. 2016. Nature Communications 8:14049

Running cellranger using the command line

https://support.10xgenomics.com/single-cell-gene-expression/software/overview/welcome



$OutName = what you want the output directory to be called (using the sample name works well)
$SampleName = sample name provided to the sequencer; in fastq file name, e.g. SampleName_S1_L003_R1_001.fastq.gz
$SampleIndices = Set of four oligos, such as CAGTACTG,AGTAGTCT,GCAGTAGA,TTCCCGAC, *OR* a code like SI-GA-A2

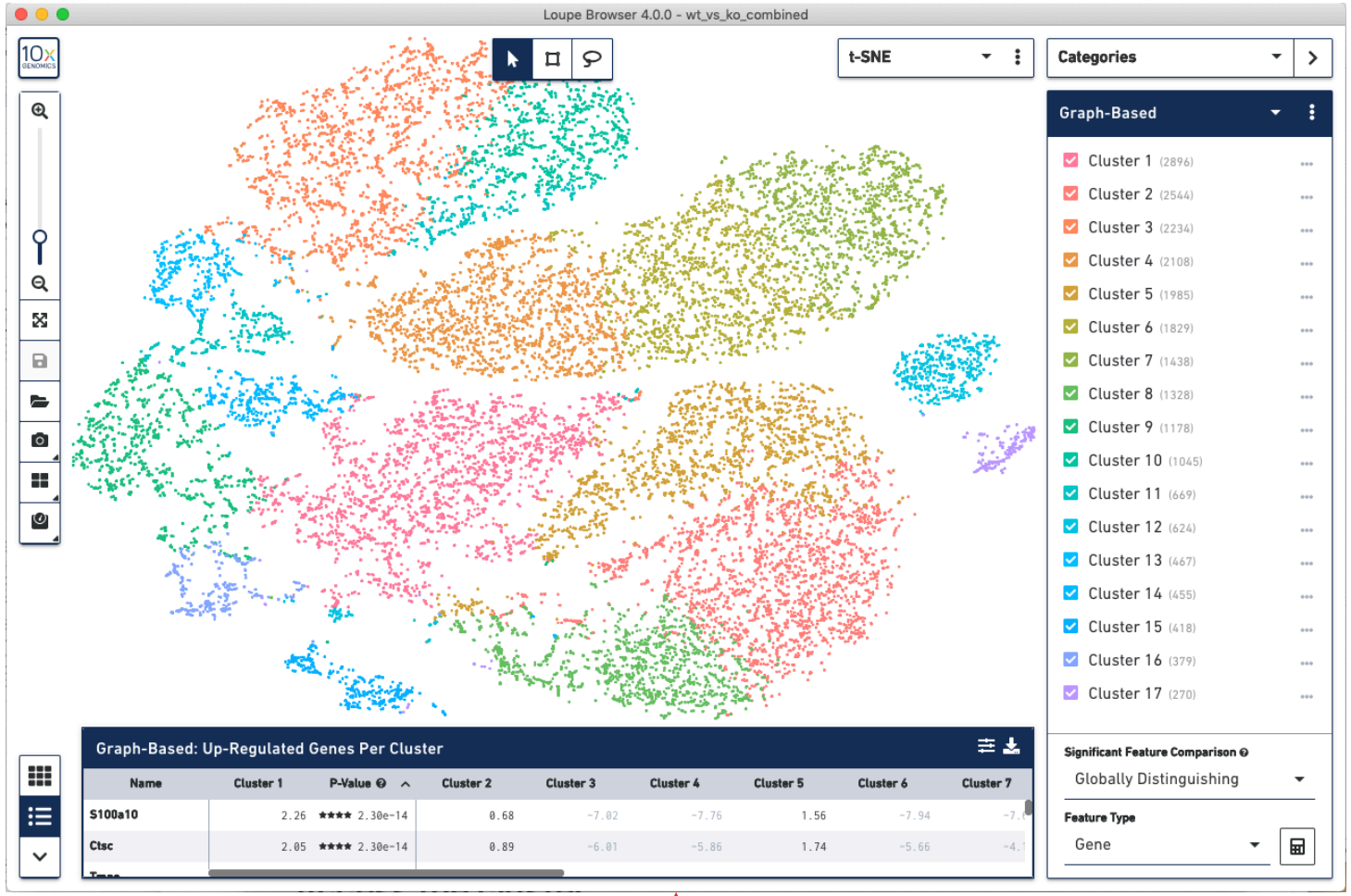
Values from some real experiments

Metric	Human – Cryo. Bone Marrow	Human – Fresh GBM Cell lines	Mouse – Cryo. Bone Marrow
Estimated Cells	7000	8500	5000
Target Reads/Cell	50K (expression), 200K (variants)		
Median Genes/Cell	2000	5600	2100
% Transcriptome mapping	>50%*	70	>70%
% Antisense Reads	~3%	~5	~3%
Fraction reads in cells	80-90%	80-90%	80-90%
Total Genes Detected	20,000	25,000	16,500
Median UMIs/Cell	5000-6000	25000	7000-8000

QC: Possible reasons for low quality

Metric	Human
Estimated Cells (>7000)	Low viability, lysed cells
Target Reads/Cell	Rarely problematic
% Transcriptome mapping (>50%)	Wrong transcriptome, low sequence quality
% Antisense Reads (<5%)	Wrong chemistry, low sequence quality
Fraction reads in cells (80-90%)	Lysed cells, extracellular RNA

The Loupe browser is a useful GUI for preliminary analysis



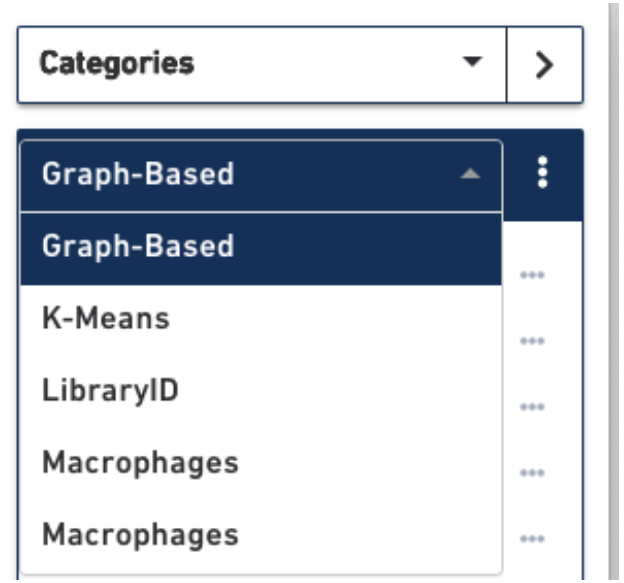


Table of differentially expressed genes in each cluster

Methods Galore

Number of single cell tools ~ 842 (https://www.scrna-tools.org/)

Point and Click:
Loupe Browser (free)
Partek Flow ($$$)
Flow-Jo ($$$)

Large data sets:
scSVA
SAUCIE

Pseudotemporal Ordering:
STREAM
Monocle 2 & 3 (R)
Slingshot (R)
PAGA (Python)
pCreode (Python)

General-purpose:
Seurat V3, V4
scanPy
LIGER (NMF)
CellHarmony
scAlign
Scanorama
Monocle V3

Predicting the future:
RNAvelocity
scVelo

Cell type assignment:
SingleR
Cellassign
CellHarmony
scPred
Moana
Garnet

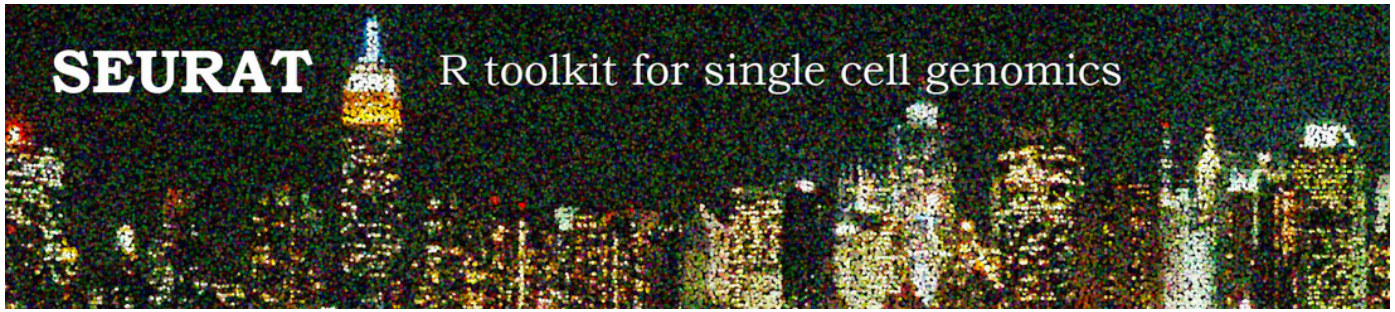
Mutation Detection:
CONICSmat (CNV)
HoneyBadger (CNV, LOH)
cb_sniffer (SNVs, Indels)
Vartrix (SNVs, Indels)

→ Need thoughtful, creative application of existing tools to extract new biology

Analysis of the gene-by-cell matrix: Overview

https://rnabio.org/module-08-scrna/0008/02/01/scRNA/

- For multiple samples, optionally subsample to achieve comparable sequencing depth using ‘cellranger aggr’ function
- Read in data and perform initial gene and cell filtering:
  - Retain genes present in >= x cells
  - Retain cells containing >= y genes
- Merge samples if not done in cellranger
- Batch correction, if necessary (cellranger does not do this properly)
- Data quality overview for subsequent filtering
  - Plot distribution of Genes/cell, UMIs/cell, mitochondrial percentage per cell, ribosomal percentage per cell
- Secondary cell filtering (depends on data set, questions):
  - Genes and/or UMIs
  - Mitochondrial transcript %
  - Ribosomal transcript %
- Calculate G1/S and G2/M scores for each cell
- Normalize expression, Find variable genes, Scale Data
- Remove unwanted sources of variation (part of scaling)
  - Cell cycle (total cell cycle, not “Cell cycle difference”)
  - Mitochondrial percentage
  - Ribosomal percentage
  - Combinations of these variables
- Principal Component Analysis (PCA) on variable genes
- Retain and plot key information about each principal component (PC):
  - Percentage of standard deviation explained
  - P-value (obtained from bootstrapping “Jackstraw”)
  - Plot gene expression heatmaps for each of the top ~12 principal components
- Choose Principal Components:
  - Purpose: choose relative importance of minor expression signatures
  - Discontinuity in elbow plot (of standard deviation explained by each PC)
  - All PCs that explain >= 2% of SD
  - P-value from JackStraw analysis < $1x10^{-100}$
  - Clarity of PC heatmaps
- Compute t-SNE and UMAP layouts on n Principal Components (NB: not raw data)
  - 5-50
  - Cellranger default: 10
  - Partek default: 50
- Clustering
  - A tool for finding patterns in the data
  - Graph-based (unsupervised, must specify resolution (0.7))
  - Alternative: k-means (supervised, must specify k)
- Characterizing Clusters in terms of individual genes
  - Differentially expressed genes (numerous methods)
  - PC-perspective
    - Choose genes that contribute heavily to top principal components
    - Plot heatmaps of these genes in each cluster
    - Independent of clustering
    - Shows relationships of clusters to each other
- Cell type discovery
  - Markers
  - Reference data set
  - DEGs




```
scrna.counts <- Read10X(data.dir = "/yourpath/outs/filtered_feature_bc_matrix")
scrna <- CreateSeuratObject(counts = scrna.counts)
scrna <- NormalizeData(object = scrna)
scrna <- FindVariableFeatures(object = scrna)
scrna <- ScaleData(object = scrna)
scrna <- RunPCA(object = scrna) # Principal Component Analysis
scrna <- FindNeighbors(object = scrna) # build K-Nearest Neighbor network
scrna <- FindClusters(object = scrna) # cluster the data
scrna <- RunTSNE(object = scrna)
scrna <- RunUMAP(object = scrna)
DimPlot(object = scrna, reduction = "tsne")
UMAPPlot(object = scrna)
```

picky

Features = Genes (and/or proteins if using CITE-seq)
Counts = UMIs
Barcodes = Cells

Analysis governed by two main principles

1. Single-cell RNA-seq data is very high-dimensional
2. And very sparse:
   Fraction of transcripts captured per cell:
   10x 3' V2: 14-15%
   10x 3' V3: 30-32%

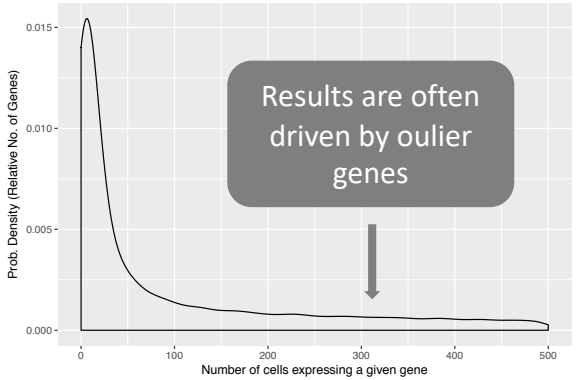
Additional Quirks:

- Transcripts encoding ribosomal proteins can comprise 30-50% of reads
- Top 100 transcripts often comprise ~50% of reads
- Low sensitivity
- Works best for cell type classification, more subtle signatures may get lost
- Results may favor highly expressed genes (e.g. *VIM*)

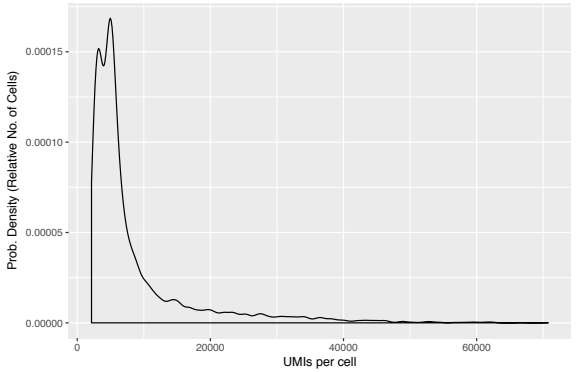
Properties of the data influence interpretation and analysis

Key variables have *wide* ranges, multimodal distributions

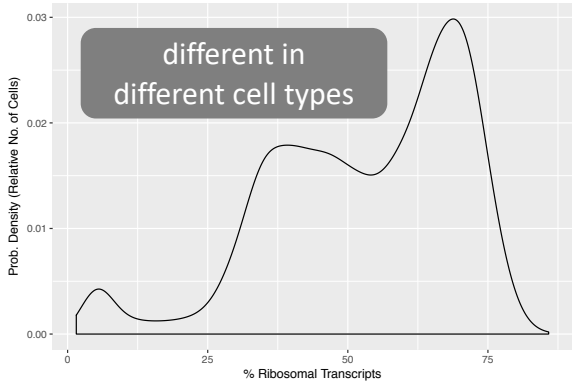
Most genes measured in a few cells!



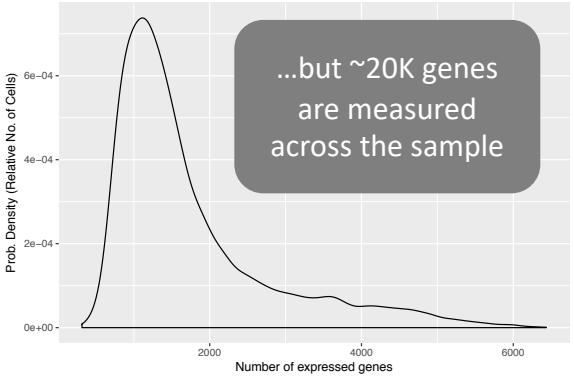
Distribution of UMIs/cell



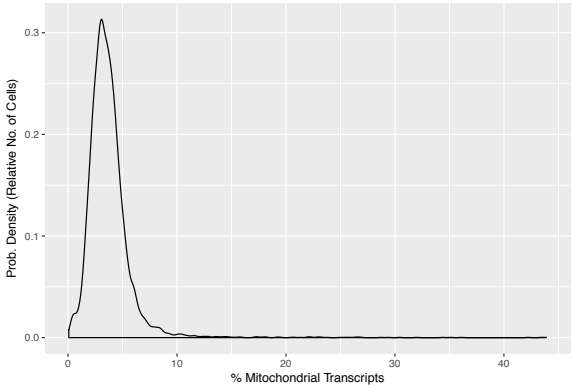
Distribution of Ribosomal transcripts/cell



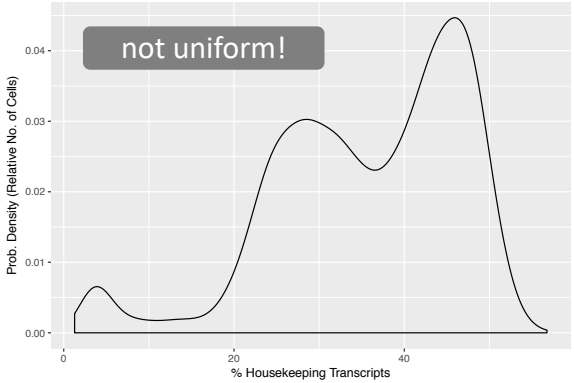
Most cells: ~1200 genes/cell



Distribution of MC transcripts/cell



Distribution of HK transcripts/cell



Some numbers

(50K reads/cell, 3' V2 kit, cellranger V2, circa 2017)

The average gene is detected in ~150 cells

~30-50% of the reads are from transcripts that encode ribosomal proteins

When detected, a gene is represented by one read on average, but the range is huge!

Metric	Sample1	Sample2	Sample3
gene.total	21342	21036	22377
gene.per.cell.mean	1499	1454	1751
gene.per.cell.med	1381	1438	1395
gene.per.cell.min	431	383	317
gene.per.cell.max	4447	4059	6435
gene.per.cell.sd	524	478	1052
cell.per.gene	168	123	123
umi.per.cell.mean	4186	4361	8412
umi.per.cell.med	3570	4144	5296
umi.per.cell.min	1655	1278	2143
umi.per.cell.max	21273	18114	70759
umi.per.cell.sd	2253	2049	8574
umi.per.gene.mean	1	1	1
umi.per.gene.max	1150	2397	17067
umi.per.gene.sd	2	2	5

More numbers (50K reads/cell, 3’ kit)

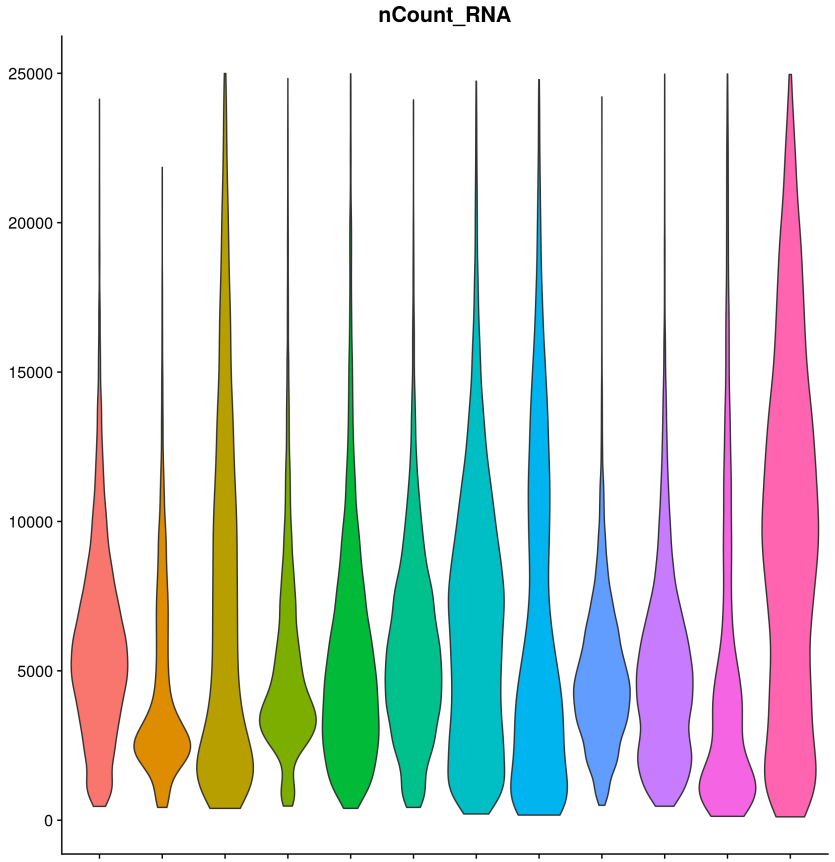
~50% of the reads come from just 100 genes

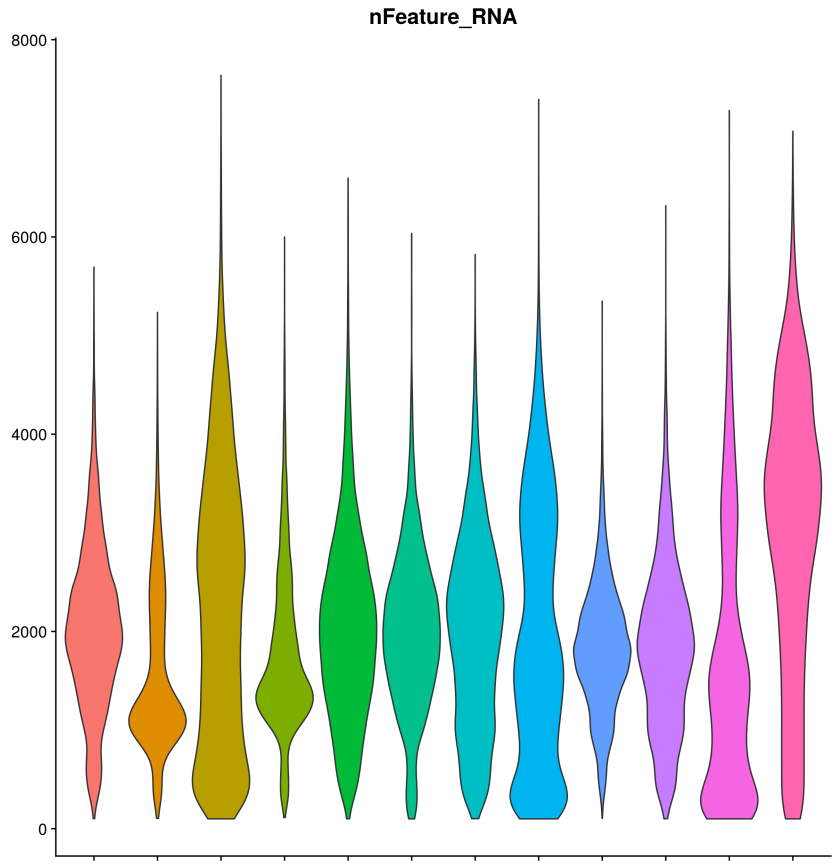
~30-50% of the reads are from transcripts that encode ribosomal proteins

~20-40% of the reads are from housekeeping genes, which are not uniformly expressed

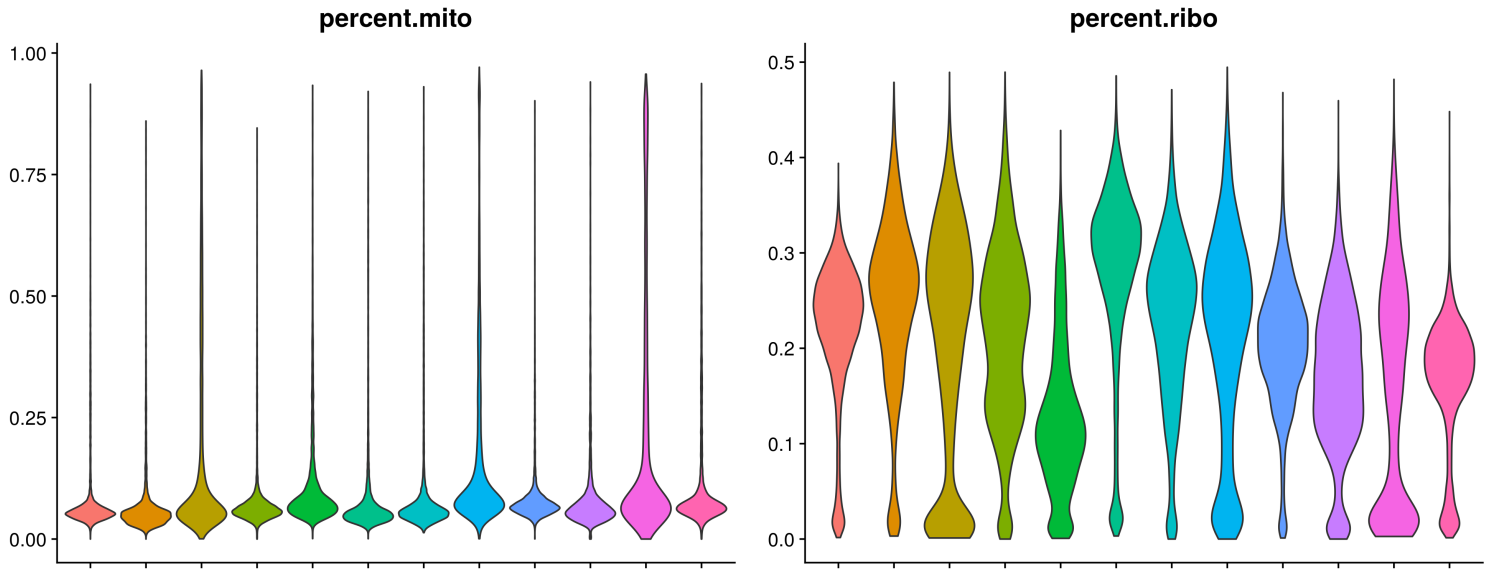
Metric	Sample1	Sample2	Sample3
Topgene %	48	56	57
Ribogene %	32	53	51
ribo.mean %	30	51	53
ribo.med %	29	53	56
ribo.min %	12	14	2
ribo.max %	68	73	86
ribo.sd %	8	8	17
mt.mean %	7	8	4
mt.med %	6	7	3
mt.min %	2	2	0
mt.max %	56	51	44
mt.sd %	2	4	2
hkgene %	22	35	34
hk.mean %	10	16	31
hk.med %	8	15	20
hk.min %	2	2	1
hk.max %	71	72	232
hk.sd %	7	9	31

Plot counts (UMIs) and features (genes) for every data set





Plot Mitochondrial and Ribosomal Protein content for every data set



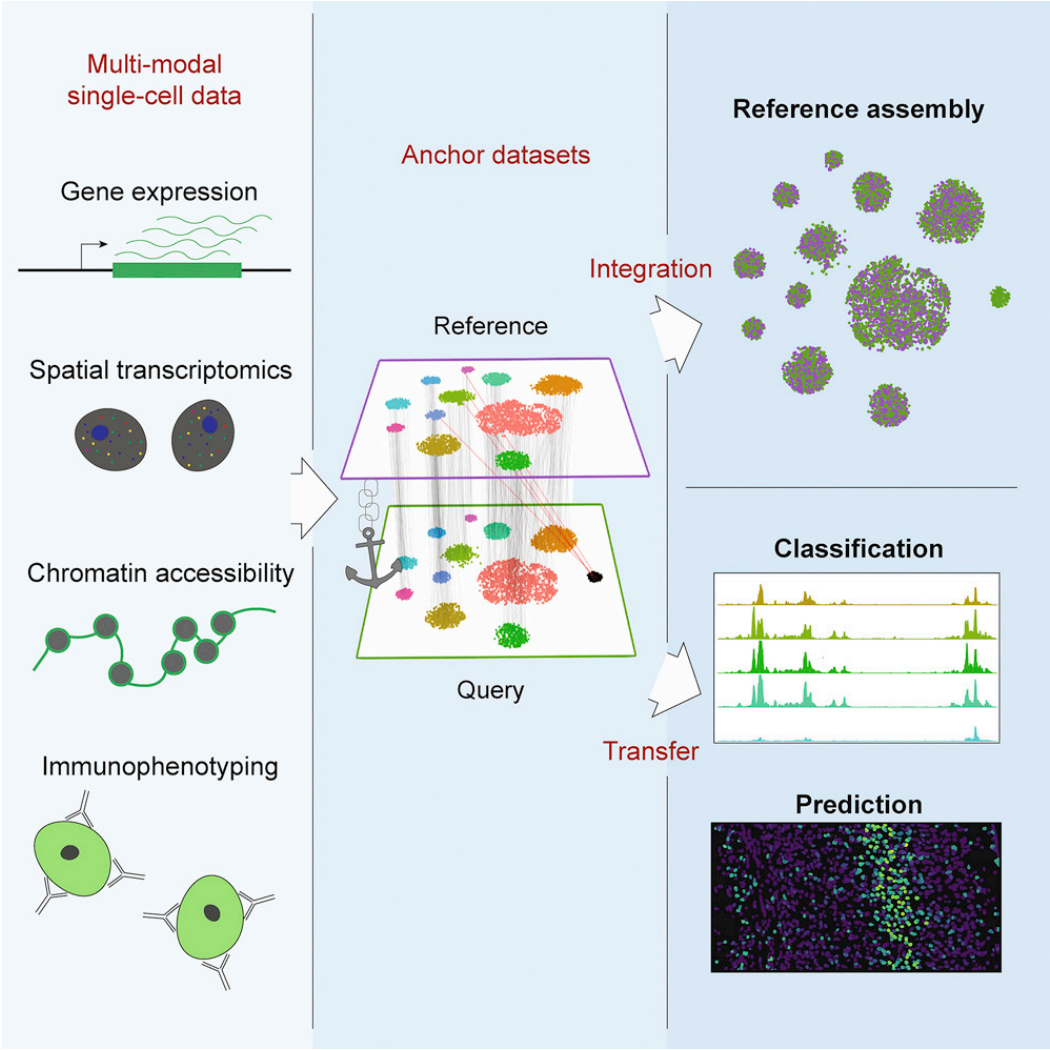
- High MC content implies cellular damage (resulting in loss of cytoplasmic transcripts) or impending apoptosis
  - Ilicic, et al. Classification of low quality cells from single-cell RNA-seq data. (2016) Genome Biology 17:29
  - Marquez-Jurado et al. Mitochondrial levels determine variability in cell death by modulating apoptotic gene expression. (2018) Nat. Comm. 9:389
- Ribosomal content: Reflects transcriptional diversity, and probably proliferative potential. (I do not recommend filtering for ribosomal content.)

Batch correction

Use batch correction only if there's a clear need for it

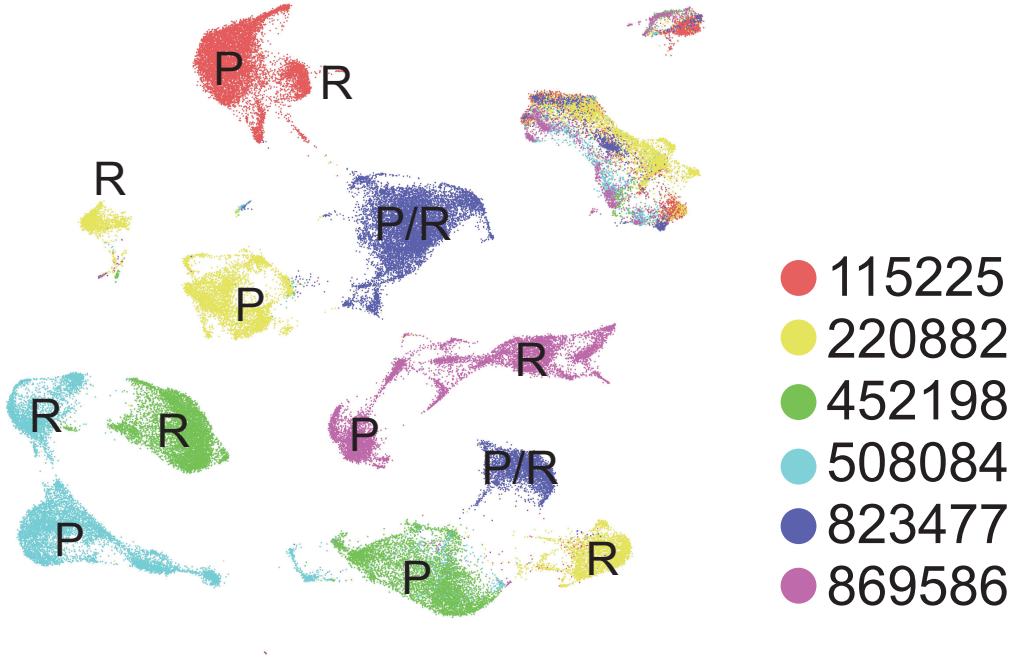
When to use it:

- Correct for different technologies (e.g. 3' and 5')
- Correct for grossly different batches
- Discover conserved biology by finding corresponding cells across different data sets
- Combining data of different types (e.g. scRNA-seq, ATAC-seq)

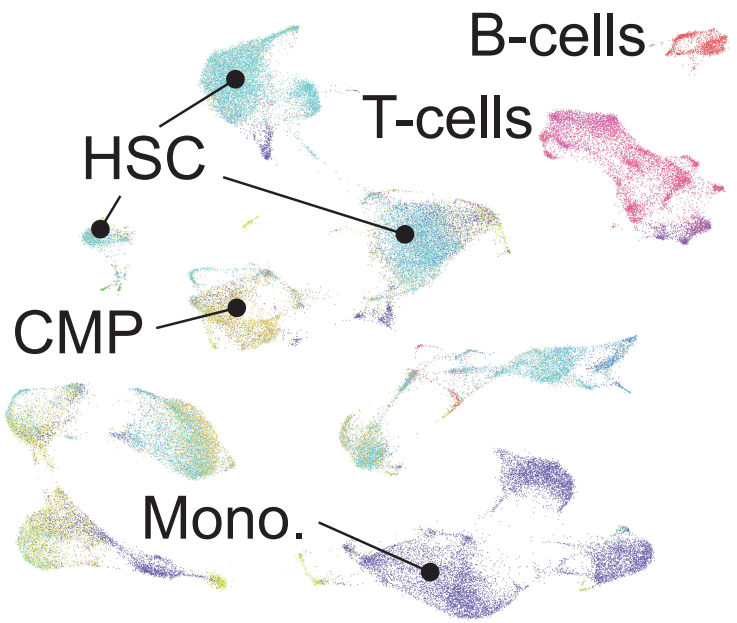


Paired presentation/relapse samples from 6 AML patients

- Every patient a different batch
- No batch correction
- Malignant cells from all samples are unique
- Normal cells co-cluster

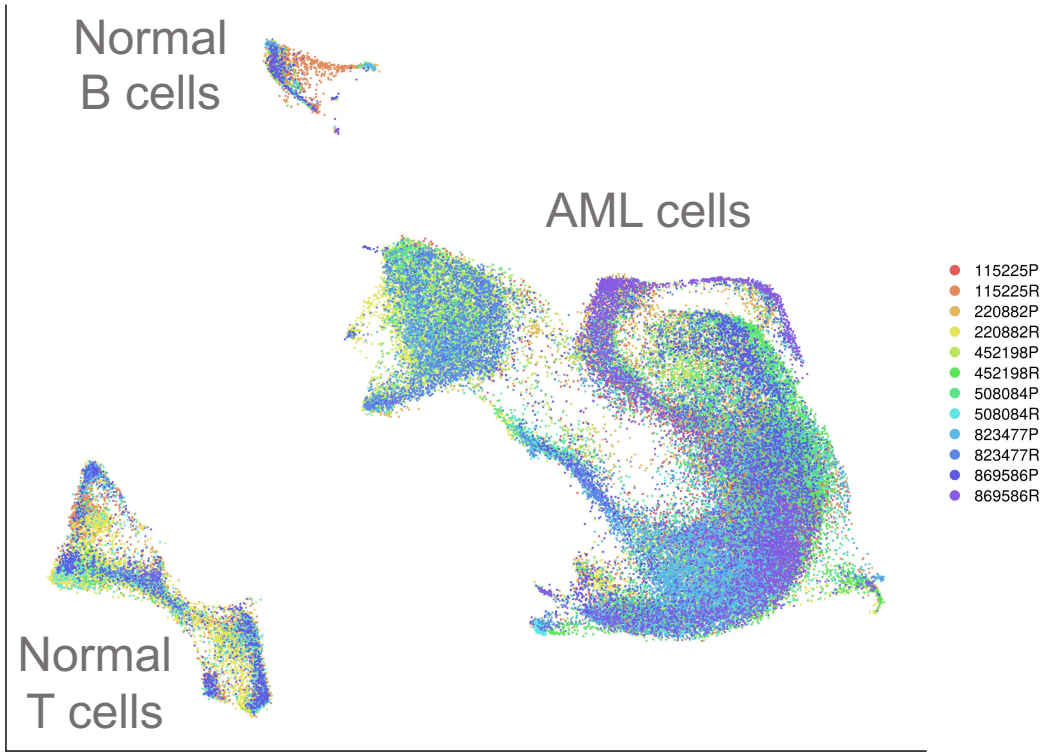


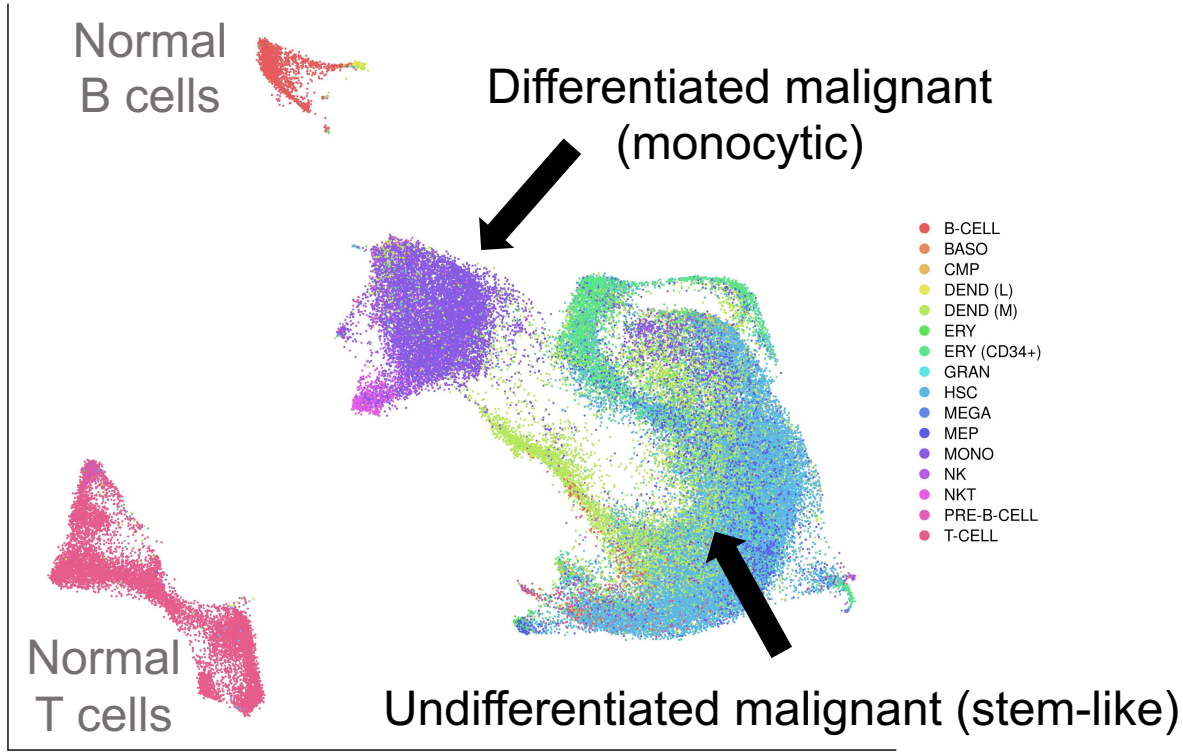
115225
220882
452198
508084
823477
869586



B-cell
BASO
CMP
DEND
EOS
ERY
GMP
GRAN
HSC
MEGA
MEP
MONO
NK
NKT
No.Prediction
T-Cell(CD4+)
T-Cell(CD8+)

Removing inter-individual differences reveals two malignant cell populations

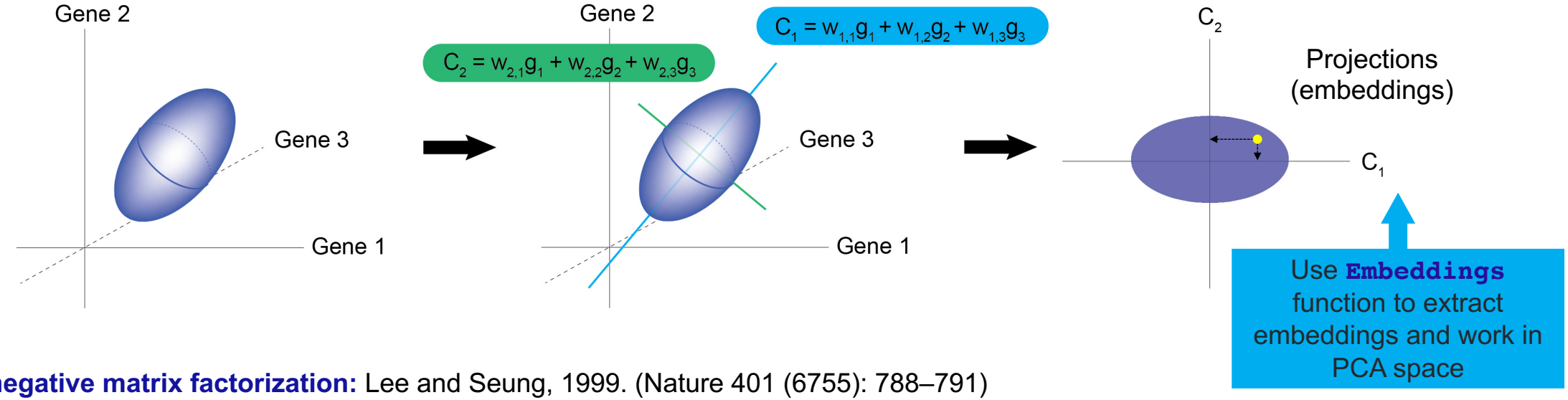




Dimensionality reduction using PCA

Purpose: Approximate original data using fewer dimensions. Define new axes that capture as much "information" as possible in as few dimensions as possible.

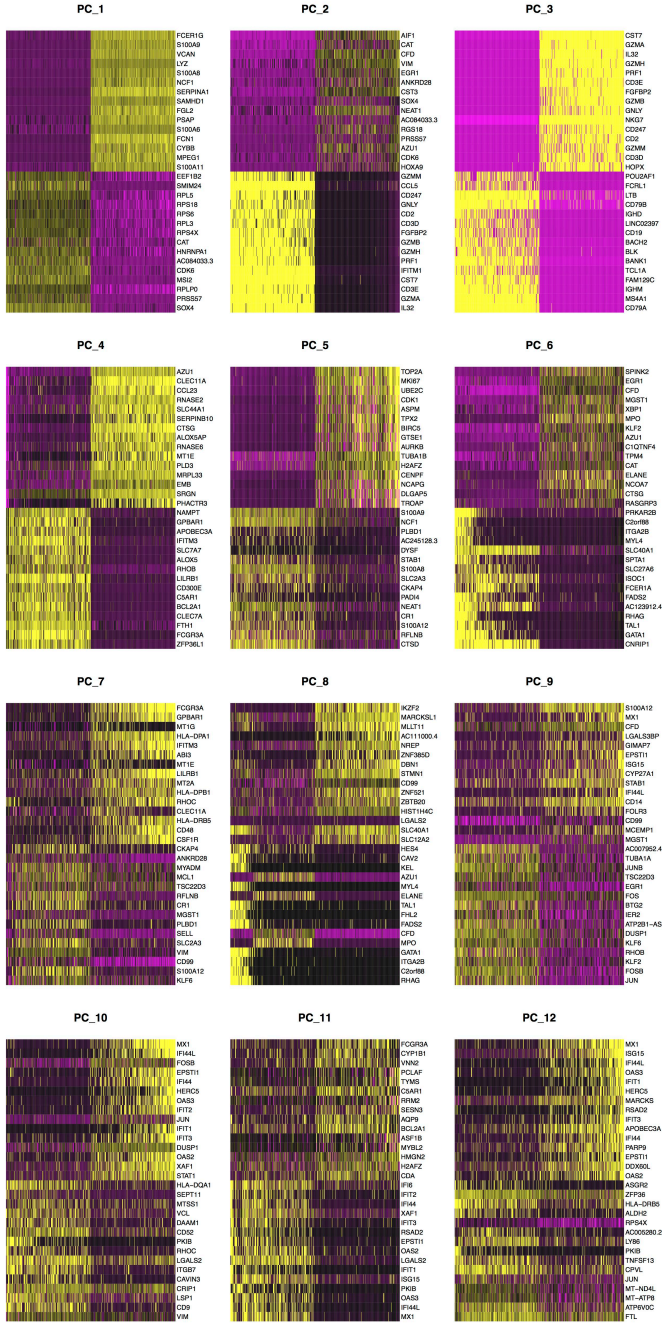
PCA = Principal Component Analysis (similar to SVD - Singular Value Decomposition)
Principal Axes, Eigen Decomposition: Euler (1751)…Cauchy (1829)
SVD: Eugenio Beltrami, 1873 (etc)
PCA: Karl Pearson, 1901
Computation: Gene Golub, Christian Reinsch, 1970
Gene Expression: Orly Alter, Patrick Brown, David Botstein, 2000 (PNAS 97 (18): 10101-10106)
Other applications: image processing, video games, math, statistics, computer science, machine learning, finance, etc.



Non-negative matrix factorization: Lee and Seung, 1999. (Nature 401 (6755): 788–791)
Similar to PCA, but axes are not mutually orthogonal, and clustering and factorization are coupled.

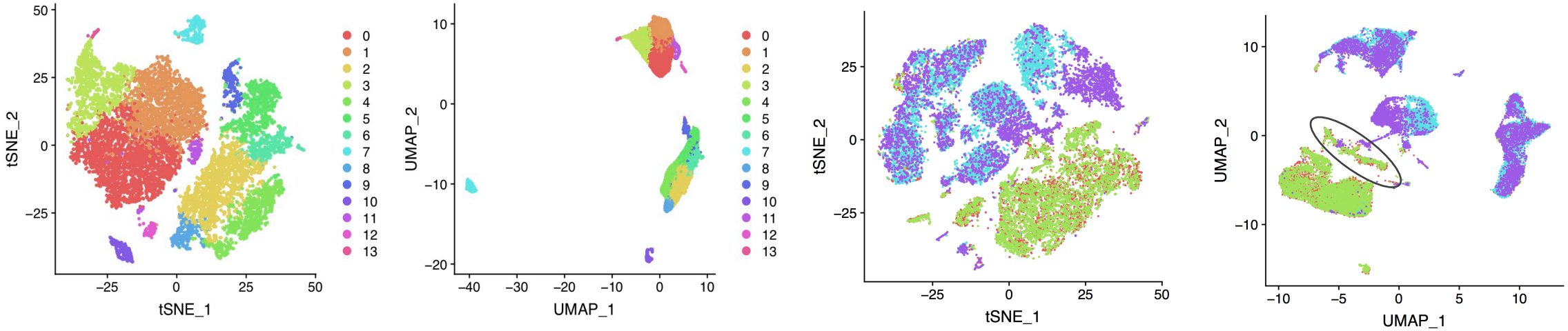
PC Tradeoff: More components → more signal, more noise

- tSNE/UMAP dominated by top components
- Minor components contain signal that is not represented in t-SNE/UMAP



Plotting using t-SNE/UMAP

t-SNE = *t*-distributed Stochastic Neighbor Embedding
UMAP = Uniform Manifold Approximation and Projection



- Algorithms & parameters matter.
- Figure out how/whether your results depend on your analysis methods.

Homework: A complete Seurat workflow

- https://rnabio.org/module-08-scrna/0008/02/01/scRNA/
- Data, files, and ancillary R script: https://wustl.box.com/s/he96swvk5gamg9f9bykqzmubngx9t8m6
- Note:

```
scrna.counts <- Read10X(data.dir = "/yourpath/samplename/filtered_feature_bc_matrix")
```