JOHNS HOPKINS
MEDICINE

McKUSICK-NATHANS
Institute of
Genetic Medicine

Introduction to scRNA-Seq analysis

Loyal A. Goff, Ph.D.
Assistant Professor, Johns Hopkins University
Quantitative Neurogenomics
10/17/2022

The Solomon H. Snyder
Department of Neuroscience

Learning Objectives

- Understand the basic steps of single cell RNA-Sequencing analysis workflows
- Develop a baseline awareness of cellular heterogeneity both between and within cell 'types'.
- Learn to identify and examine cell state transitions via pseudotime analysis
- Understanding the application of dimensionality reduction to visualization and high-dimensional sequencing data analysis.

Cellular context matters in development & disease

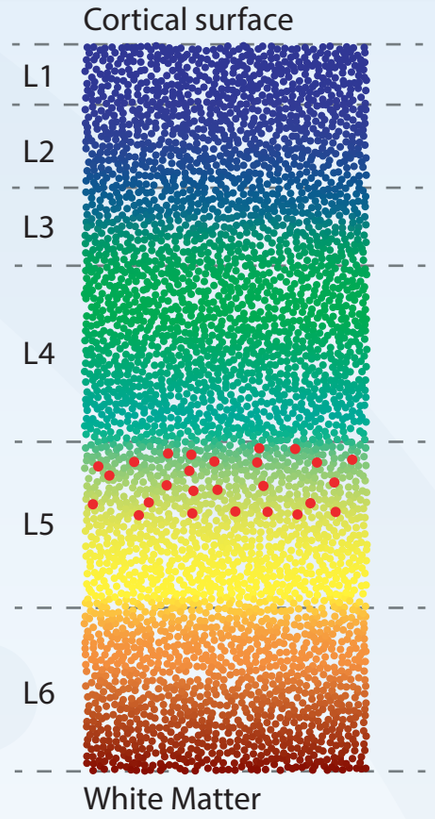
- ~3.0-4.0 x 10^{13} cells in the human body

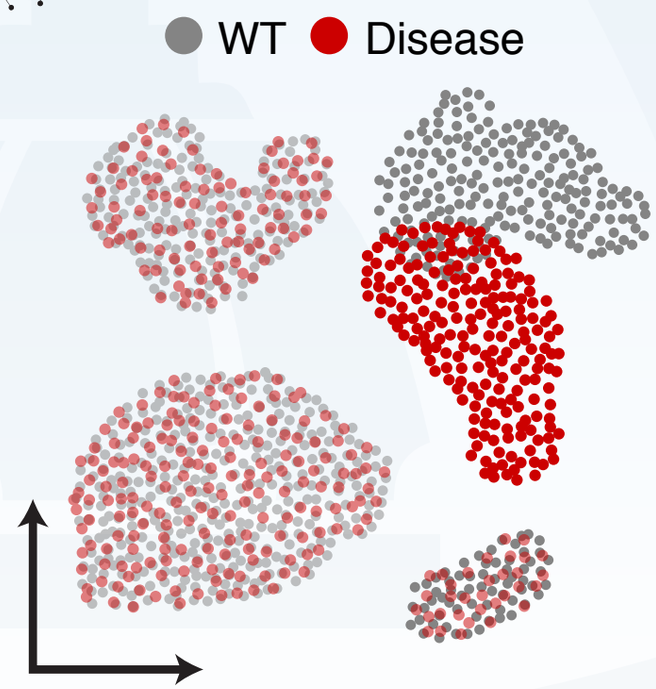
- We know relatively little about human cellular diversity
  - True number of ‘finite’ cell types
  - Diverse cellular states
  - Intrinsic and extrinsic influences of cellular behavior
  - Cellular input → phenotype

- All disease involves cells:
  - Off script (Cancer)
  - Missing or mis-programmed (Genetic disorders)
  - Damaged cells (Inflammation)

- Different cells respond differently to developmental cues and disease-causing insults

- Understanding ***cellular*** biology is foundational to modeling higher order properties in human health and disease.



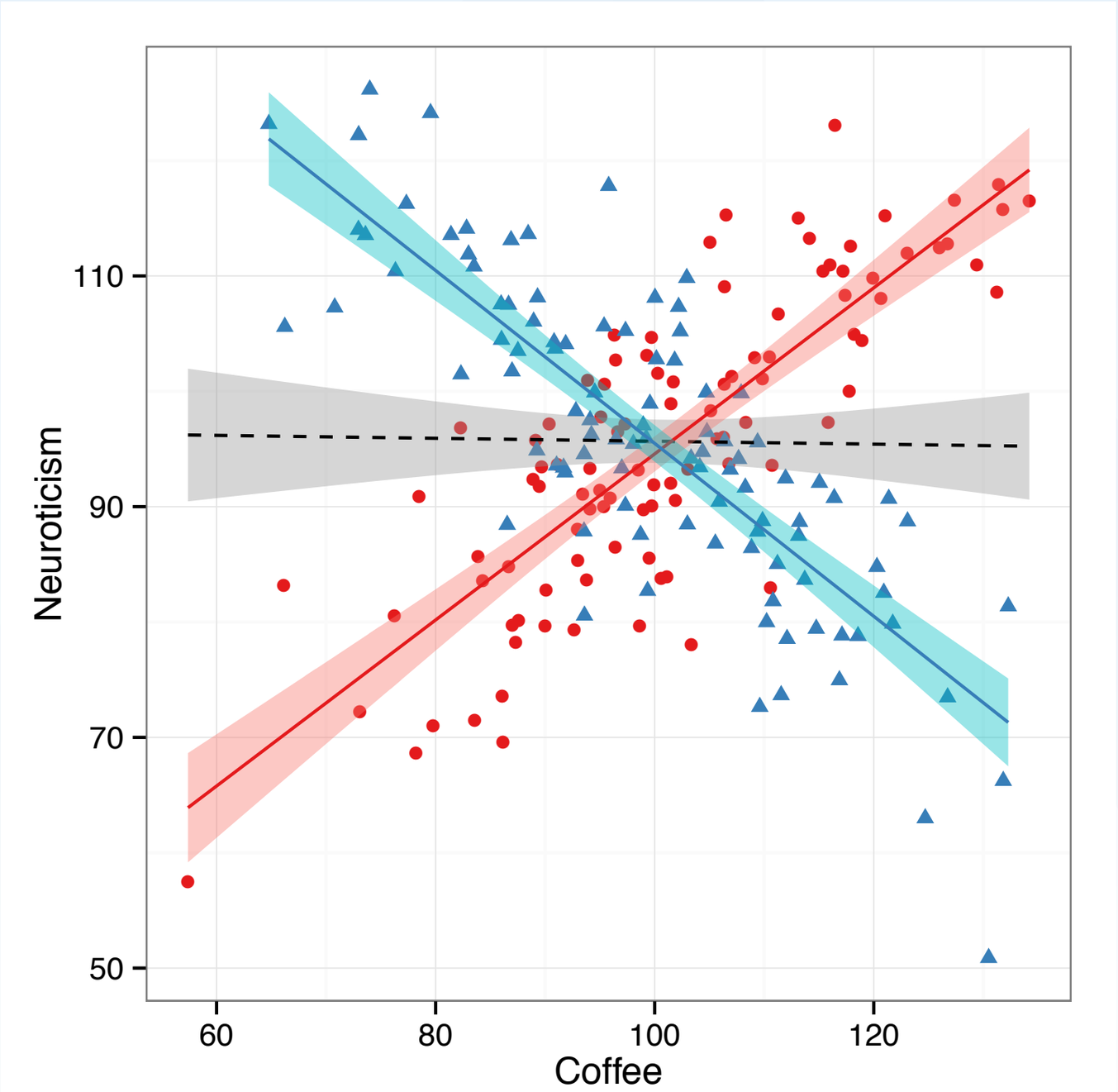


Simpson's Paradox

- Historically, molecular biology assays involve bulk sampling of heterogeneous tissue/ cells.

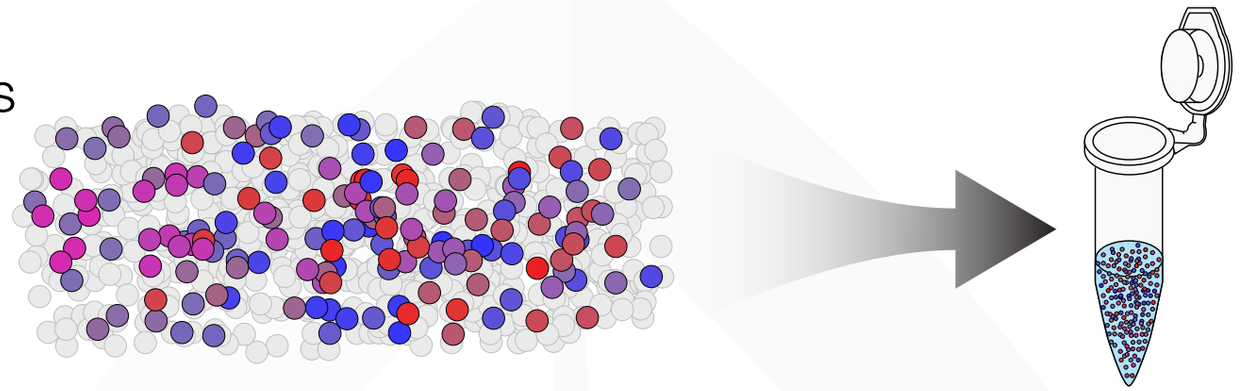
- Weighted averages can lead to reversals/ masking of meaningful relationships.

- Phenomenon applies to complex systems such as gene expression and cellular response



The case for single cell analysis

- Cells are the fundamental unit of life
  - Variation between and within cell populations is an important ***phenotype***

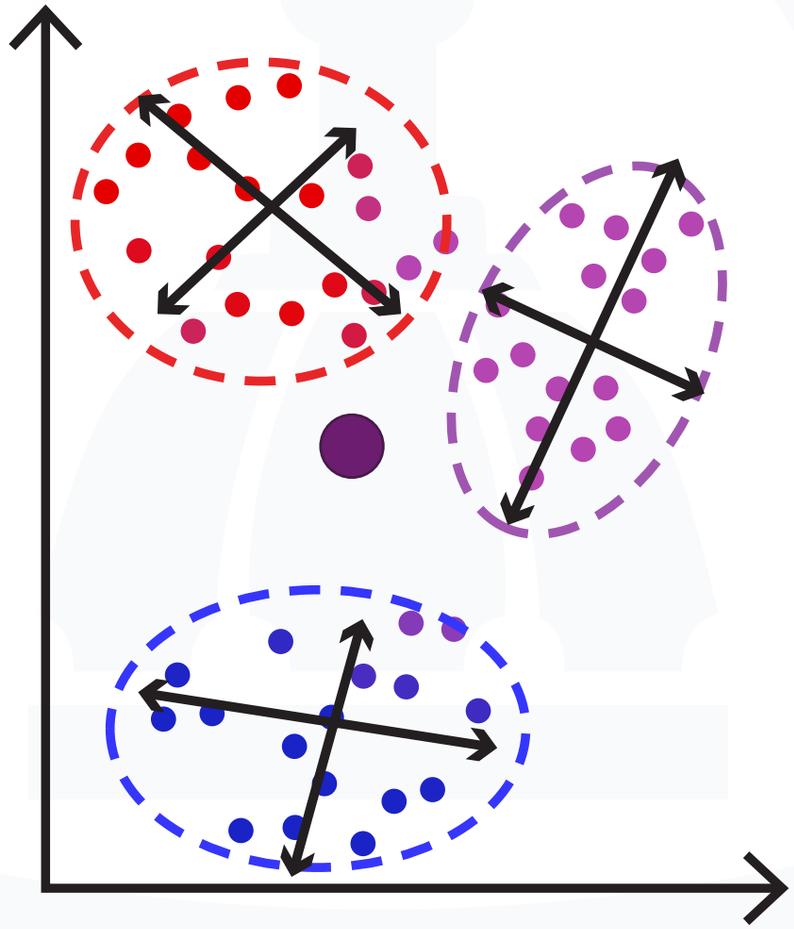


- Aggregate expression profiles of mixed-cellular samples do not accurately reflect the expression profile of any one cell type.

- Bulk measures of cellular responses to treatments/ insults/timecourse may hide variable responses:
  - Tissue-level
  - Mixed cell types
  - Contaminating cell type
  - Intrinsic variable response to insult

- Estimates of variability are compressed or worse, not accounted for, which makes predicting response difficult.

- Single cell analysis allows for unbiased estimation of sample heterogeneity and distinct celltype responses



Single cell RNA-Seq

Single cell RNA-Seq

Bulk RNA-Seq

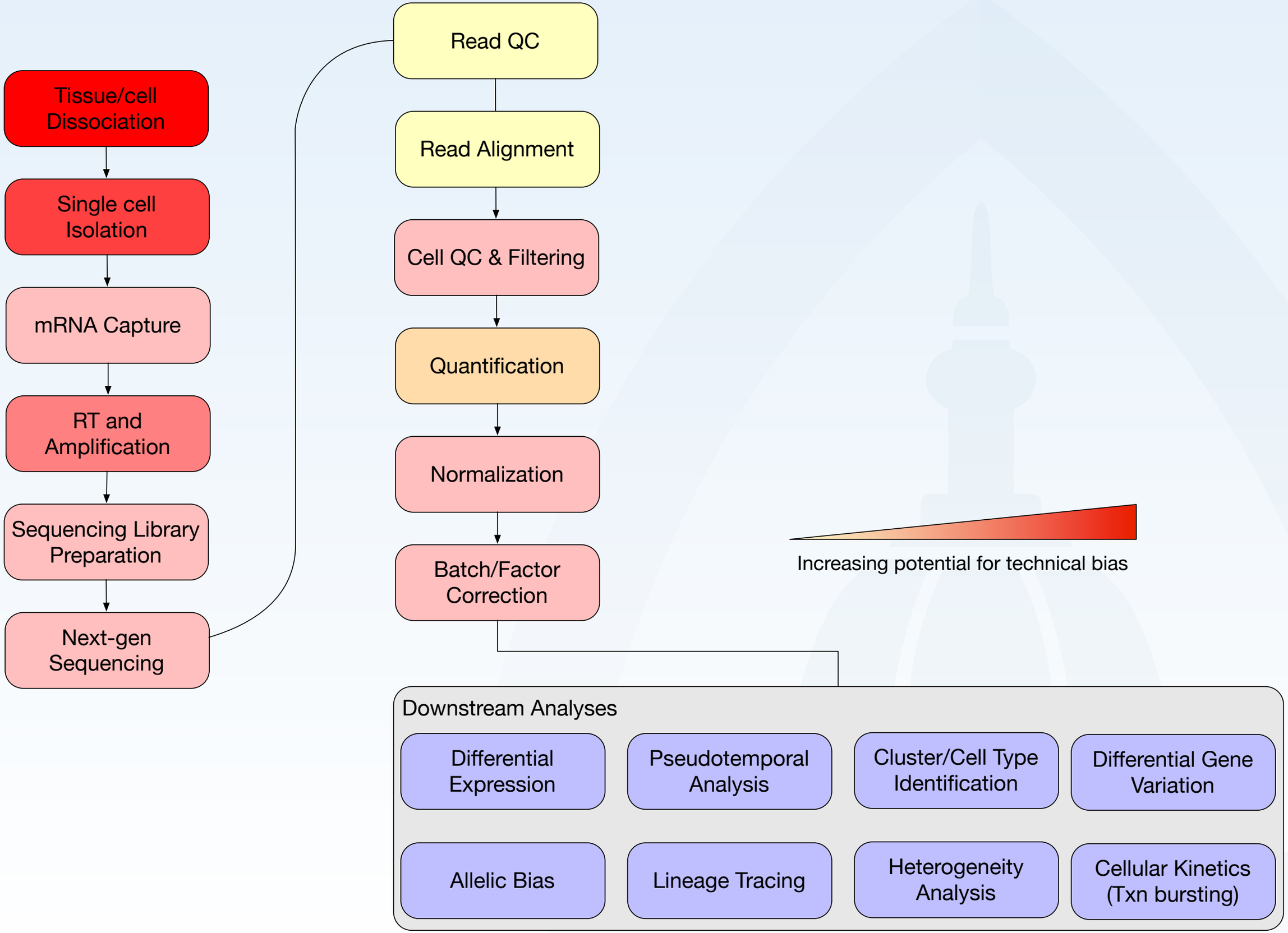
Principle differences from bulk RNA-Seq

- scRNA-Seq is ***not*** just low-input RNA-Seq
  - Fundamental differences in sample handling, alignment approaches, & data processing

- All of the standard biases of RNA-Seq…
  - Sample handling effects
  - RNA-quality effects
  - Library batch effects
  - Sequencing lane effects

- Plus a few unique to the technology:
  - Cell dissociation
  - Cell isolation/capture techniques
  - Low-input library prep
  - Amplification bias
  - Gene ‘dropout’
  - Specificity of canonical ‘marker genes’

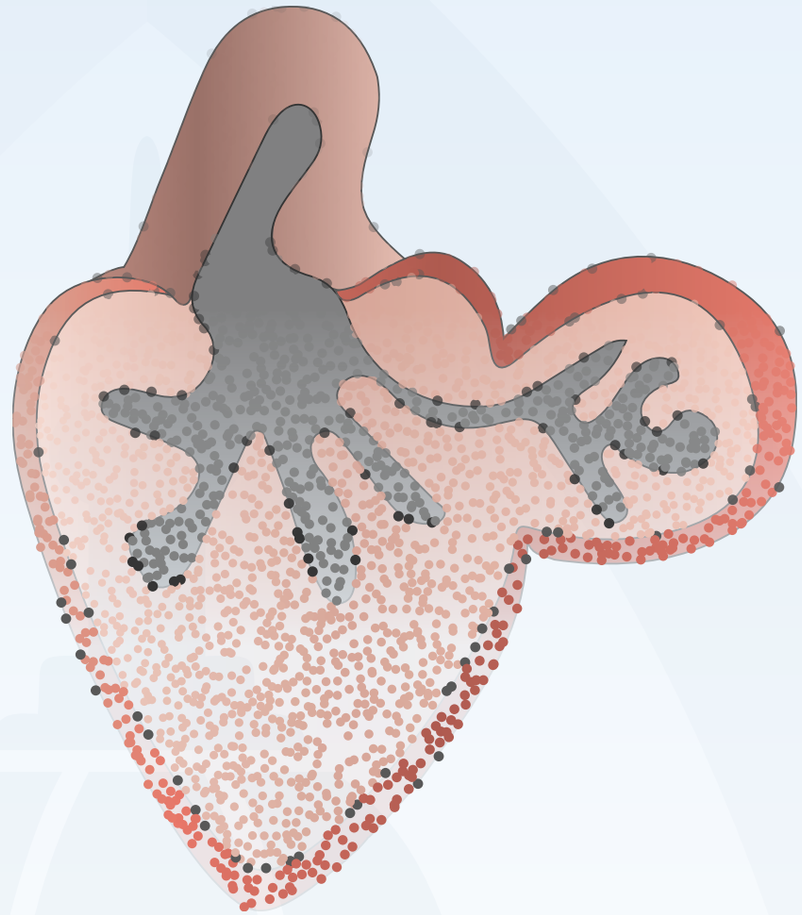
Typical scRNA-Seq experimental workflow



Tissue Dissociation

Tissue/cell Dissociation

- Most common tissue dissociations are enzymatic
- ***Must*** be optimized for each experimental system:
  - Choice of protease
  - Maximize viability and yield
  - Minimize cellular stress
    - Time
    - Hypoxic conditions
    - Physical manipulations
    - Temperature (can sometimes use lower than optimal)
  - Minimize potential for RNA-degradation
    - Nuclease-free solutions
    - RNase-free workspace
    - RNase inhibitors
- Alternatives:
  - Single *nuclei* analysis via hypotonic swelling
  - Newer psychrophilic proteases enable dissociation at lower temperatures
  - Methanol fixation protocols exist to stabilize cells after dissociation
  - Nuclear isolation



Cell Dissociation

Tissue/cell Dissociation

- *In vitro* culture dissociation is generally more amenable than tissue dissociation

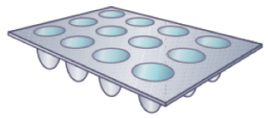
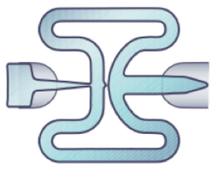
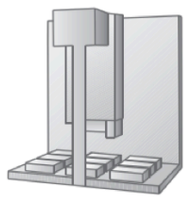
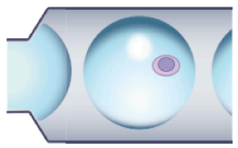
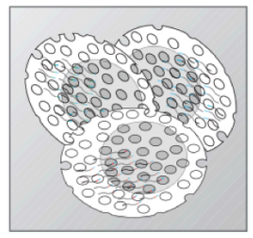
- Ensure consistency to minimize batch effects:
  - Method
  - Duration
  - Conditions

- Ideally, all replicates should be processed concurrently to avoid batch effects

- ***Caution:*** Not all cells in a heterogeneous mixture will respond the same way to dissociation methods!
  - Differential viability, release, transcriptional response
  - Even centrifugation time/intensity will bias cell type composition!
  - This step ***will*** introduce both a selection bias ***and*** a cell type-specific response bias that may not be of experimental interest
  - Replication and balance can help minimize these effects.

Single cell isolation methods

Single cell Isolation

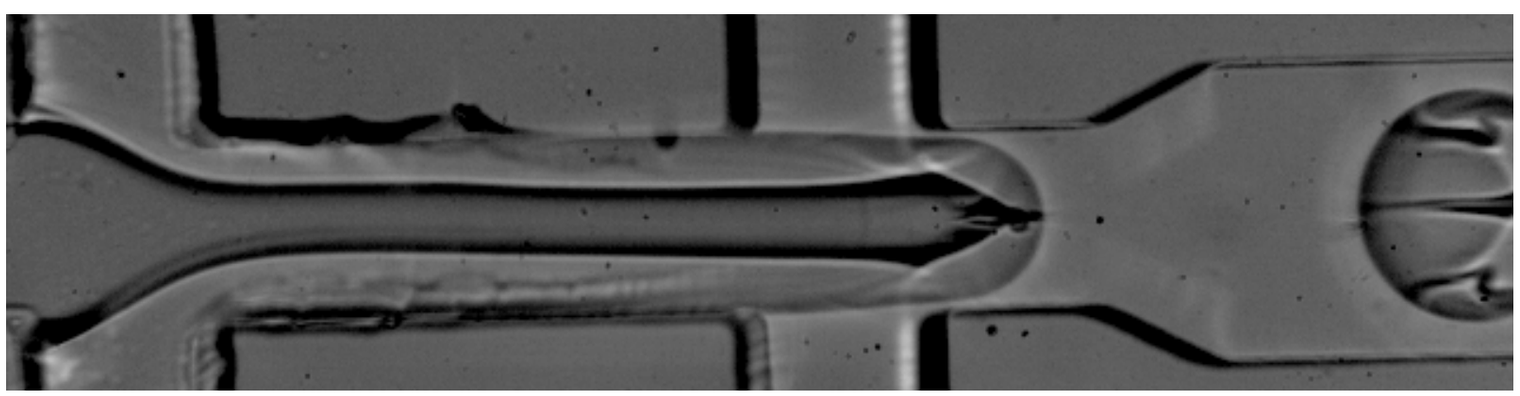
Manual	Multiplexing	Integrated fluidic circuits	Liquid-handling robotics	Nanodroplets	Picowells	*In situ* barcoding
						
Tang *et al.* 2009[18]	Islam *et al.* 2011[24]	Brennecke *et al.* 2013[64]	Jaitin *et al.* 2014[33]	Klein *et al.* 2015[34] Macosko *et al.* 2015[40]	Bose *et al.* 2015[43]	Cao *et al.* 2017[51] Rosenberg *et al.* 2017[52]

Svensson et al Arxiv 2017

Number of single cells

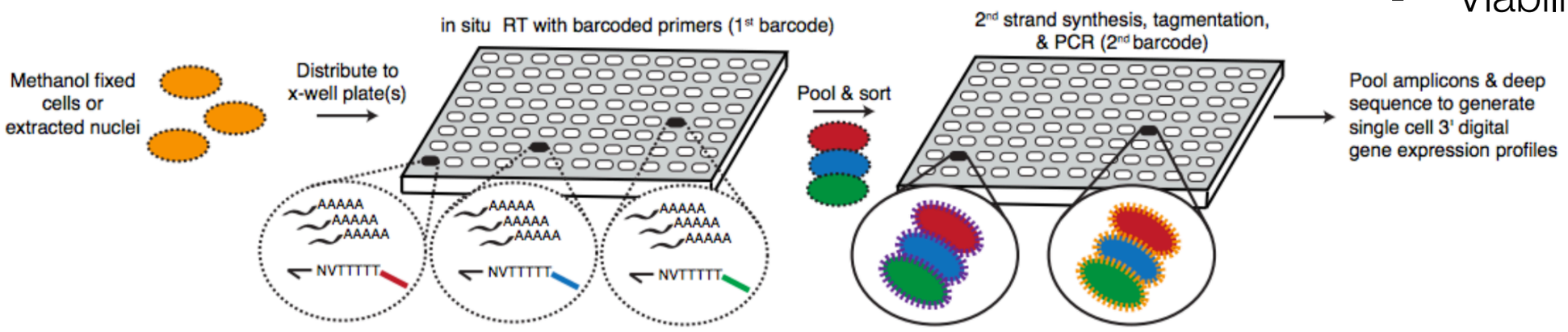
Information content per cell

Microfluidics - Drop Seq



- Trend is towards increasing # of cells
- Several require specialized equipment
- Most methods can be combined with up-front enrichment methods (FACS)
  - Viability may be impacted

In situ barcoding - sci-RNA-Seq



Isolation methods dictate sample preprocessing approaches

- Single-cell isolation methods
  - Cell becomes individual ***sample*** and can be preprocessed in ways similar to bulk ***samples***
  - Standard RNA-Seq aligners can/should be used but may require intense parallelization
    - e.g. 100's-1000's of kallisto/HISAT2 alignments must be performed (1 per cell)

- Emulsion/Droplet-based methods
  - Multiple cells are tagged (cell barcode) & processed in the same tube.
  - Resulting library must be sequenced in a way to assign individual reads to both genes ***and cell of origin***
  - Read pairs must be aligned to genome/transcriptome AND have their cell barcode 'demultiplexed'
  - CellRanger (10x) or kallisto + bustools can be used on paired end sample (contains multiple cells worth of reads) to demultiplex

Methods for mRNA Capture and Amplification

mRNA Capture | RT and Amplification

PT | IVT | TS | RamDA

Development of fundamental molecular biological techniques

1976 [39] 3' end labeling method using terminal deoxytinucleotidyl transferase

1982 [40] Okayama-Berg method for cDNA cloning using termial deoxynucleotidyl transferase

1983 [49] RNase H-DNA polymerase I mediated 2nd strand synthesis strategy

1988 -1989 [41-44] Establishment of Poly-A tagging (PT) strategy

1990 [45] Single-cell transcript analysis | [50] Establishment of *In vitro* transcription (IVT) strategy with RNase H-DNA polymerase I mediated 2nd strand synthesis

1992 [51] Single-cell transcript analysis

1995 [46,47] Single-cell transcript analysis

1999 [48] Establishment of Template-switching (TS) strategy

Single-cell microarray

2001 [1] Eberwine et al.

2002 -2003 [2] Iscove et al. [3] Tietjen et al. | [4] Kamme et al.

2006 [5] Kurimoto et al. | [6] Subkhankulova and Livesey

Single-cell RNA-seq

2009 [7] Tang et al.

2011 [8] STRT †

2012 [9] CEL-seq † | [10] Smart-Seq

2013 [11] Quartz-Seq | [12] STRT2 †,# [13] C1-Smart-Seq § [14] Smart-Seq2

2014 [15] Streets et al. § | [16] MARS-seq †,# | [17] SCRB-seq †,#

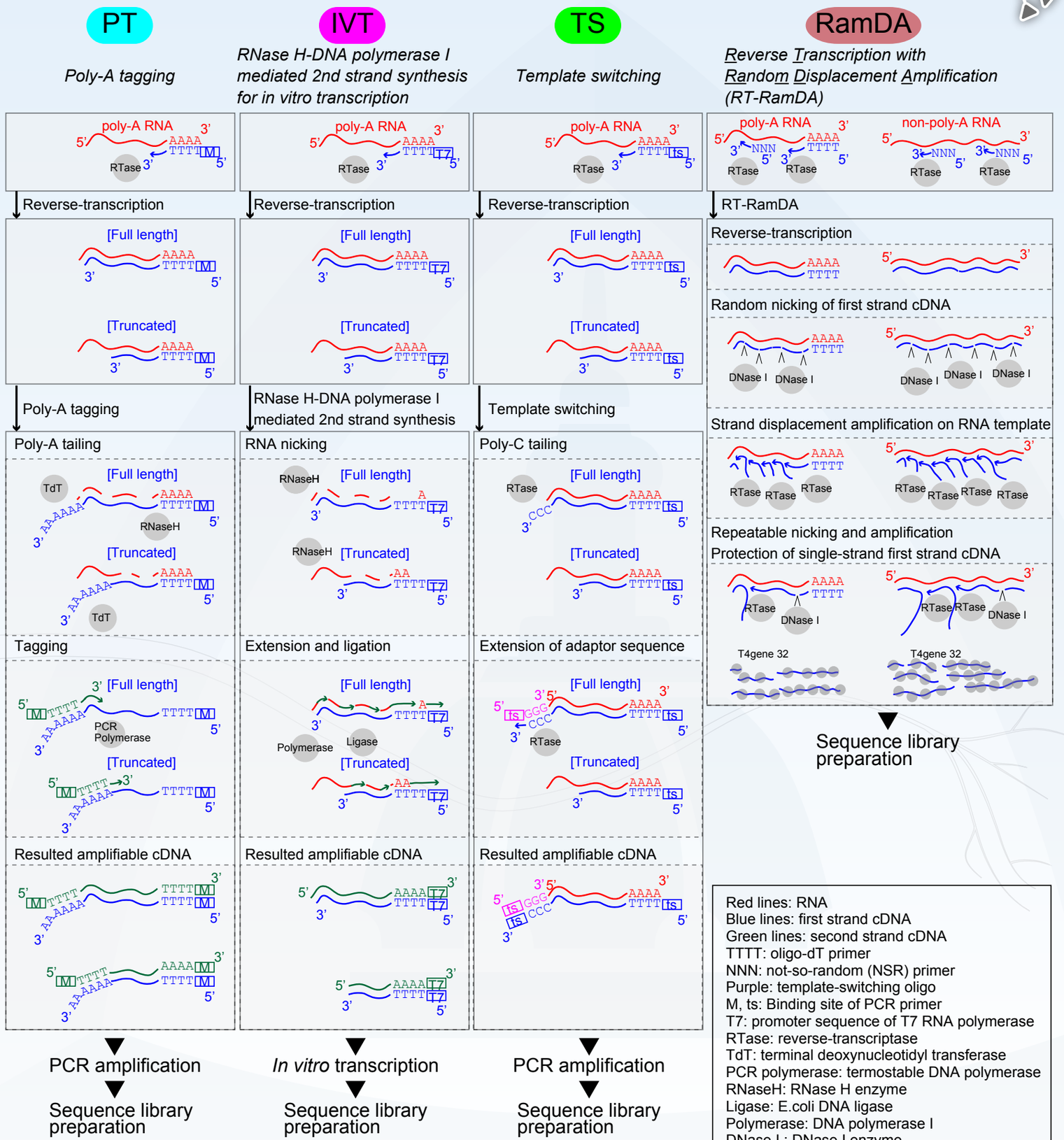
2015 [18] SUPeR-seq [19] SC3-seq [20] Beads-seq | [21] inDrop †,#,§ [22] Bose et al. †,#,§ | [23] Drop-seq †,#,§

2016 | [24] CEL-seq2 †,#,§ [25] SORT-seq †,# | [38] Establishment of RT-RamDA

2017 [26] MATQ-seq # [27] MR-seq [28] scDual-seq †,# | [29] STRT-seq-2i †,#,§ [30] Chromium †,#,§ [31] Seq-Well †,#,§ [32] Nx1-seq †,#,§

2018 [33] Quartz-Seq2 †,# | [34] mCEL-seq2 †,# | [35] SPLiT-seq †,#,‡ [36] Microwell-Seq †,#,§ | [37] RamDA-seq

†: using cell barcoding technology
#: using molecular barcoding technology
§: using microfluidics/microwell device
‡: using combinatorial cell barcoding technology



Red lines: RNA
Blue lines: first strand cDNA
Green lines: second strand cDNA
TTTT: oligo-dT primer
NNN: not-so-random (NSR) primer
Purple: template-switching oligo
M, ts: Binding site of PCR primer
T7: promoter sequence of T7 RNA polymerase
RTase: reverse-transcriptase
TdT: terminal deoxynucleotidyl transferase
PCR polymerase: termostable DNA polymerase
RNaseH: RNase H enzyme
Ligase: E.coli DNA ligase
Polymerase: DNA polymerase I
DNase I : DNase I enzyme
T4 gene 32: T4 gene 32 protein

mRNA capture, RT & Amplification

mRNA Capture | RT and Amplification

	SMART-seq2	CEL-seq2	STRT-seq	Quartz-seq2	MARS-seq	Drop-seq	inDrop	Chromium	Seq-Well	sci-RNA-seq	SPLiT-seq
Single-cell isolation	FACS, microfluidics	FACS, microfluidics	FACS, microfluidics, nanowells	FACS	FACS	Droplet	Droplet	Droplet	Nanowells	Not needed	Not needed
Second strand synthesis	TSO	RNase H and DNA pol I	TSO	PolyA tailing and primer ligation	RNase H and DNA pol I	TSO	RNase H and DNA pol I	TSO	TSO	RNase H and DNA pol I	TSO
Full-length cDNA synthesis?	Yes	No	Yes	Yes	No	Yes	No	Yes	Yes	No	Yes
Barcode addition	Library PCR with barcoded primers	Barcoded RT primers	Barcoded TSOs	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers and library PCR with barcoded primers	Ligation of barcoded RT primers
Pooling before library?	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Library amplification	PCR	In vitro transcription	PCR	PCR	In vitro transcription	PCR	In vitro transcription	PCR	PCR	PCR	PCR
Gene coverage	Full-length	3′	5′	3′	3′	3′	3′	3′	3′	3′	3′
Number of cells per assay	10^5 / 10^4 / 10^3 / 10^2										

- Most popular methods exploit polyA-tail hybridization for mRNA capture
  - Oligo-dT priming can contribute to lower capture efficiency (~5-25%)
- Reverse transcription to cDNA to stabilize RNA
- Amplification of cDNA required for library prep
  - Most methods still require PCR amplification

Chen et al. 2018 Ann. Rev. Biodatasci

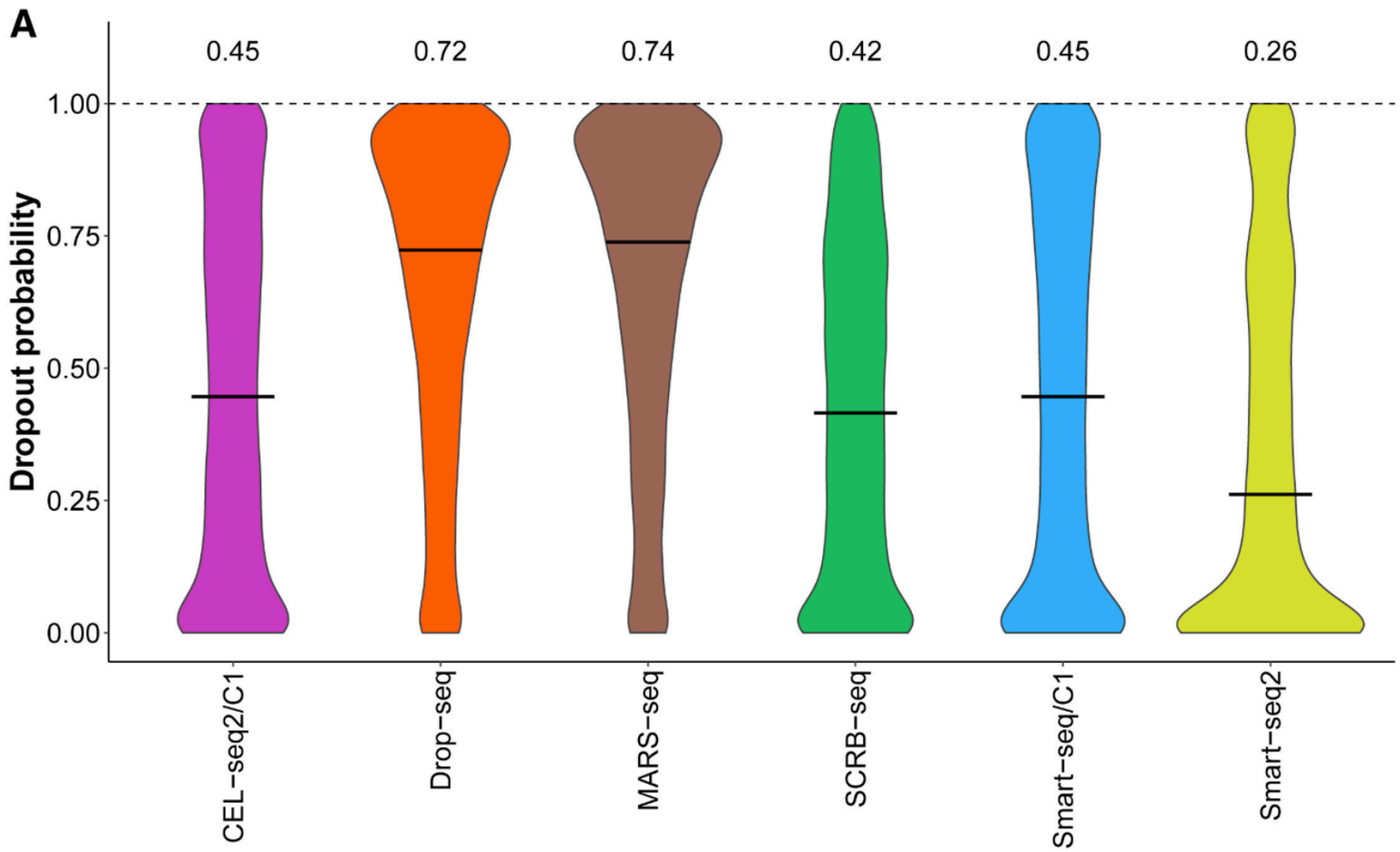
Capture efficiency - Dropout rate

- *An event in which a transcript is not detected owing to a failure to capture or amplify it*

- Impacts lower-abundance genes more severely

- Different capture efficiencies for different isolation methods

- Factors influencing dropout rate:
  - *Ex vivo time*
  - Primer choice
  - ***TSO choice***
  - Molar availability of primer
  - Constrained diffusion (e.g. primers bound to bead)
  - Cell viability
  - ***Intrinsic noise***



Ziegenhain et al. 2017

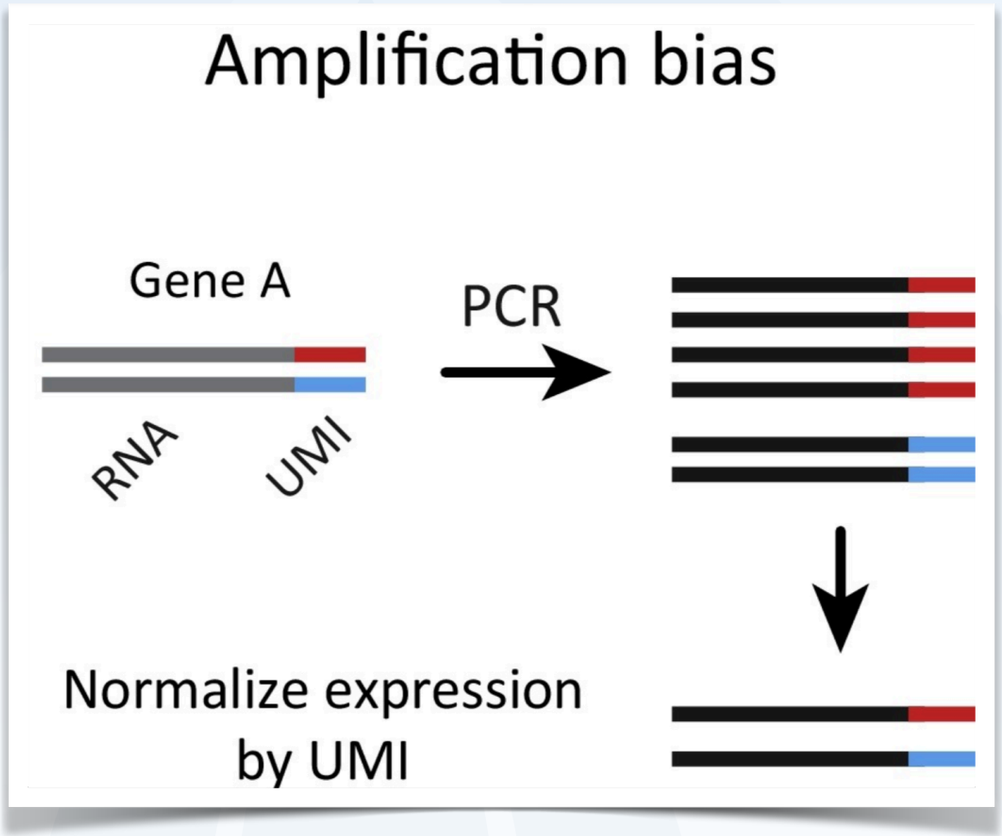
Amplification bias

- Most methods require amplification to produce sufficient material for library prep and sequencing
  - No amplification method is without selection bias, but some are worse than others
  - This is a trade-off between sufficient material and library complexity.

- PCR amplification (Smart-Seq2, Chromium, Drop-Seq)
  - Amplification is exponential ***and*** biased based on amplicon features
    - Excess # of cycles will 'jackpot' easily amplifiable sequences
  - ***Must*** optimize number of cycles for each experimental system.
  - Use of UMIs can minimize effect of PCR-based amplification.

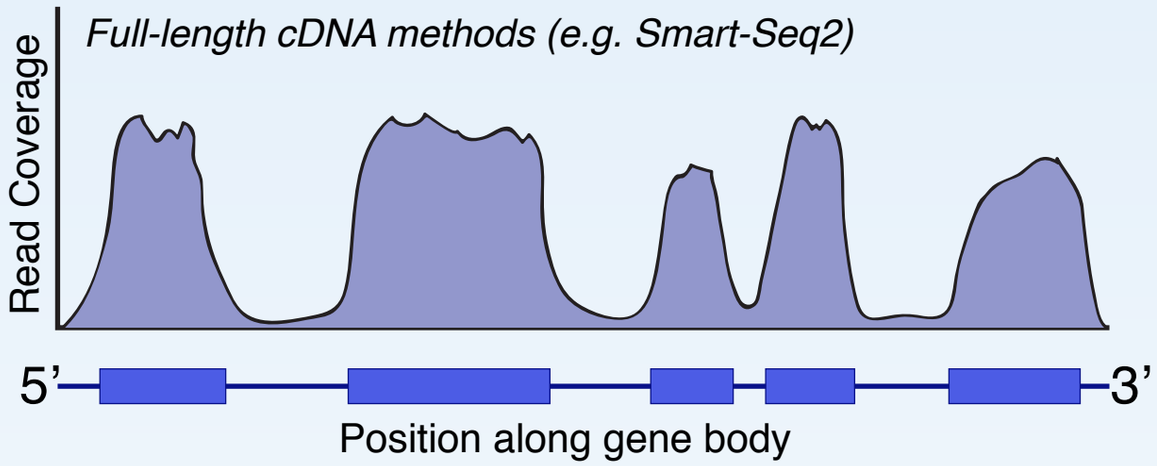
- IVT amplification (CEL-Seq, InDrop)
  - Linear amplification of RNA from ds-cDNA
    - Less prone to selective enrichment
  - Longer protocol
  - Amplified RNA still prone to degradation relative to DNA.

- RT-RamDA
  - Reverse Transcription with RAndoM Displacement Amplification
  - Efficient linear amplification

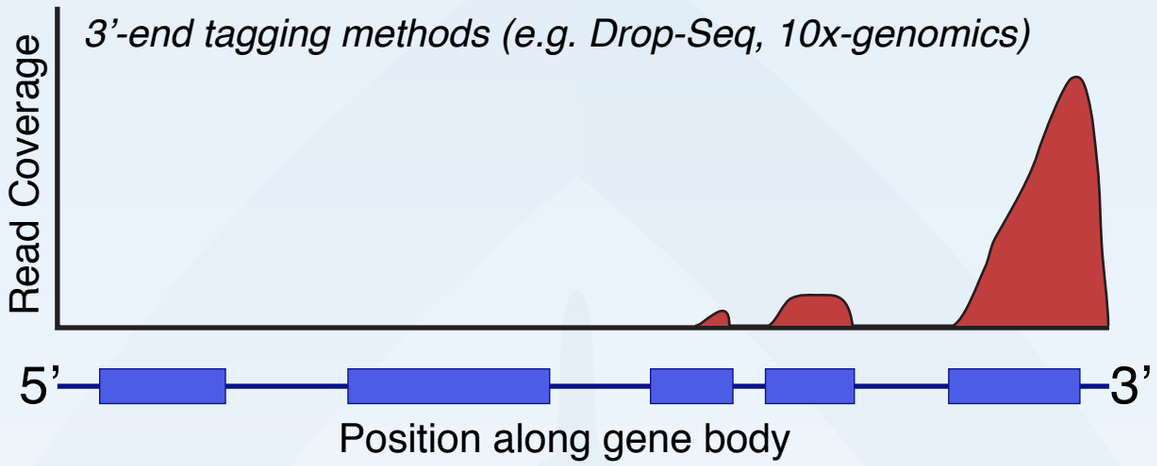


Two types of quantification strategies for scRNA-Seq

Sequencing Library Preparation



- Attempts to produce uniform read coverage of each transcript
  - Often there are biases in this coverage
- Better mappability
- Can potentially:
  - distinguish isoforms
  - alternative TSS
- Accurate quantification requires more reads per cell



- Captures only the 3’ (or more recently 5’) end of a given transcript
- Can be used in conjunction with UMIs to improve quantification and remove impact of amp. bias
- Theoretically unbiased by gene length.’
- Requires fewer reads per cells

Choice of library prep method significantly affects throughput, sensitivity, and types of questions that can be asked of the final data.

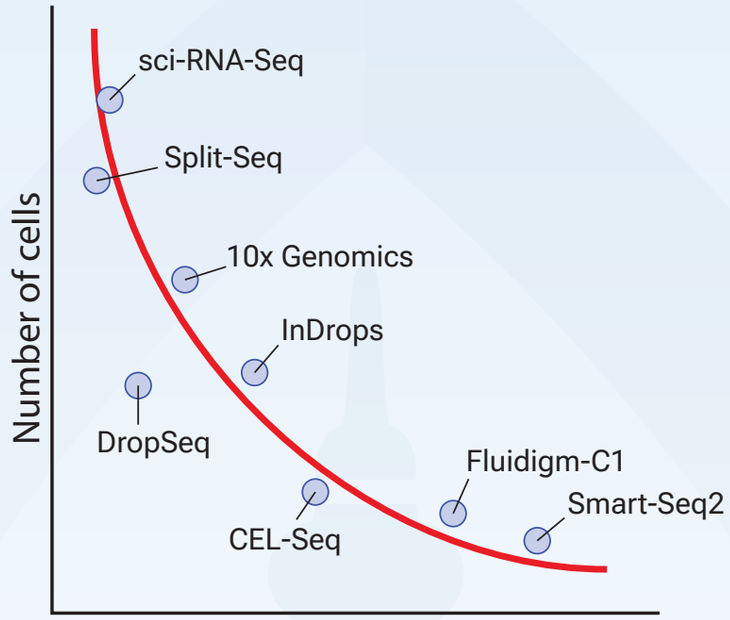
How many cells do I need?

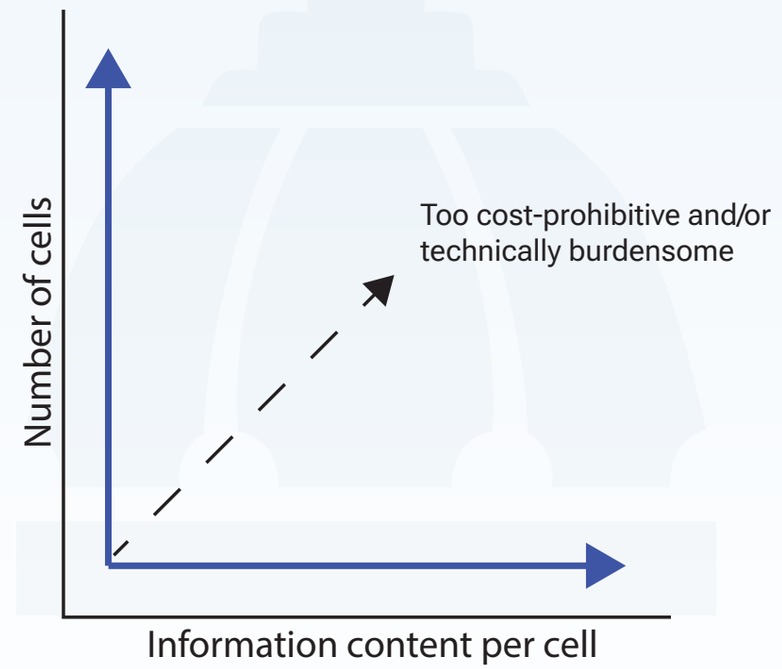
- The choice of cell number is based on experimental question(s).

- Choice of cell number should guide cell isolation and library prep method selection

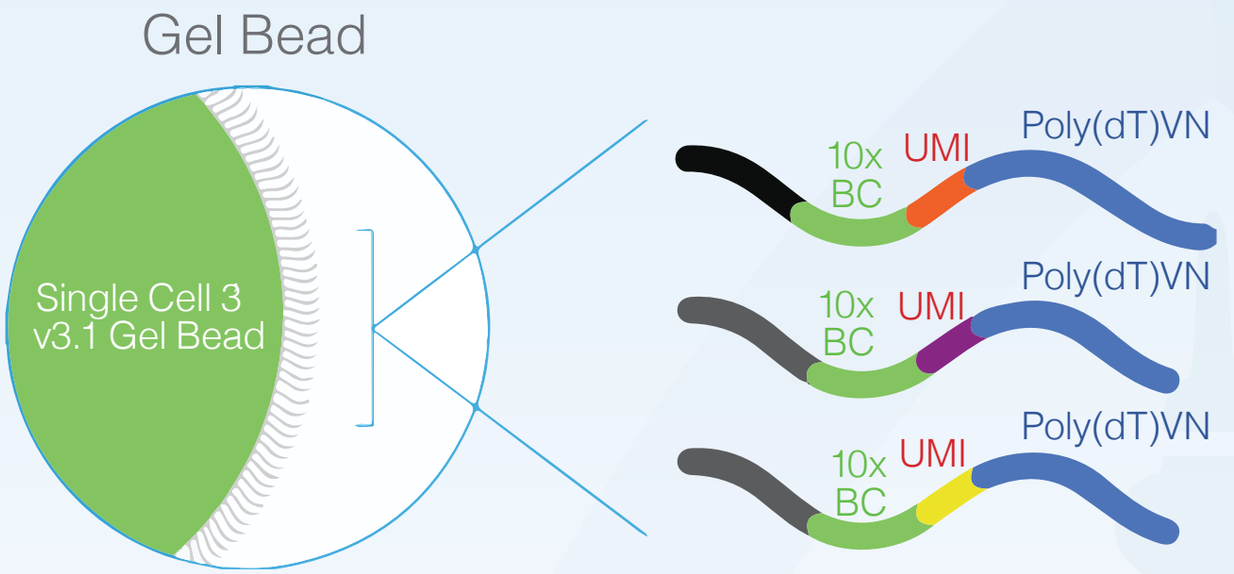
- ‘Discovery’ experiments generally require larger numbers of cells

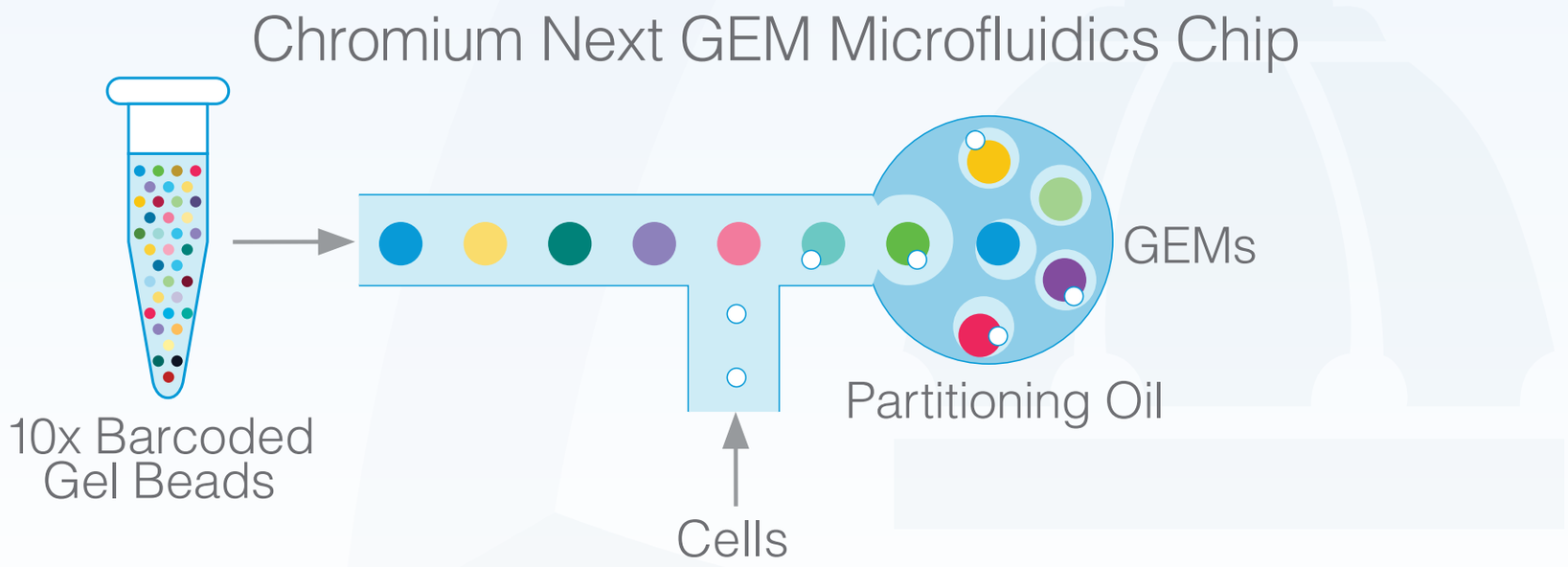
- To identify subtle differences between cell types/states or when comparing cells with fewer differences more information per cell is optimal.





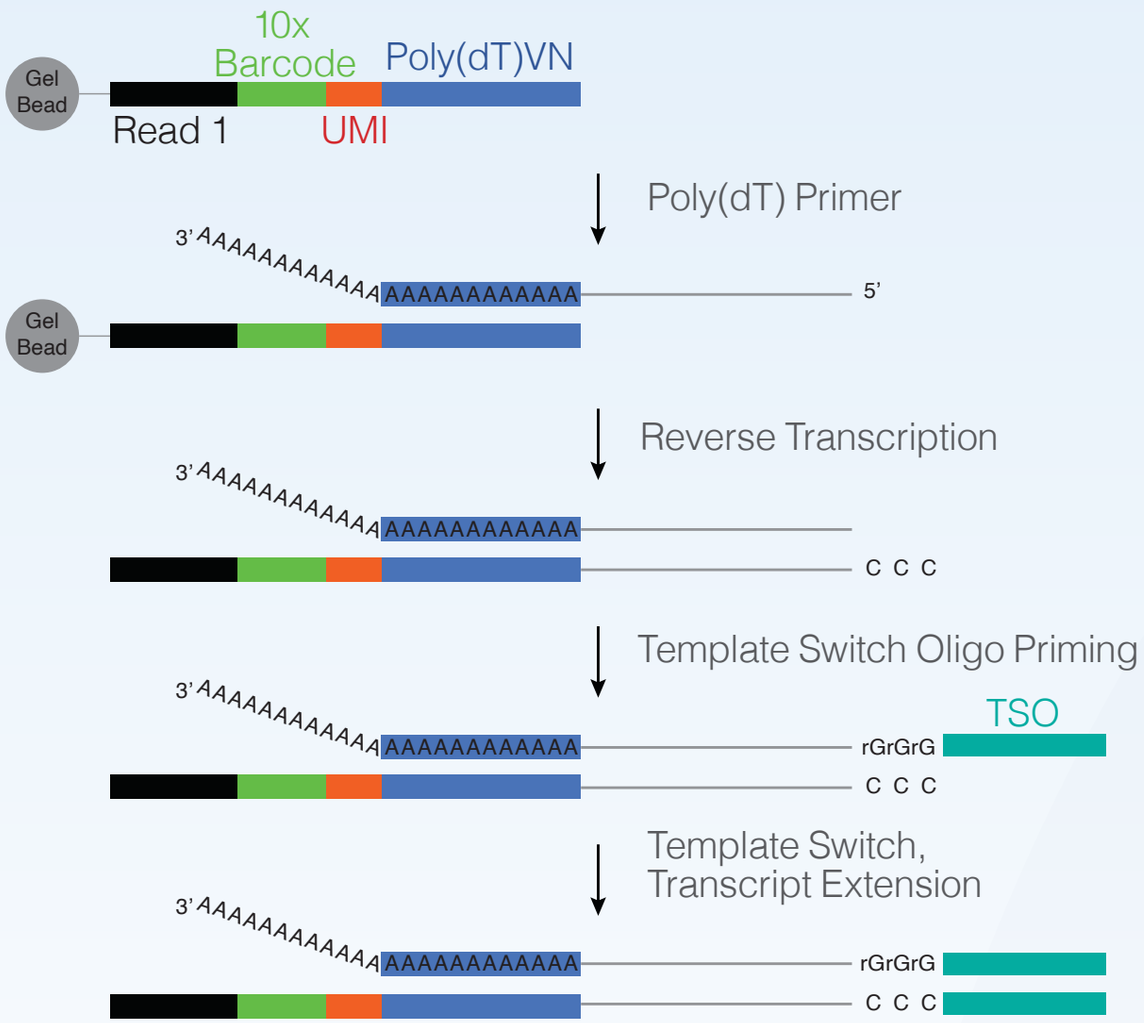
10x single cell RNA-Seq chemistry & library construction





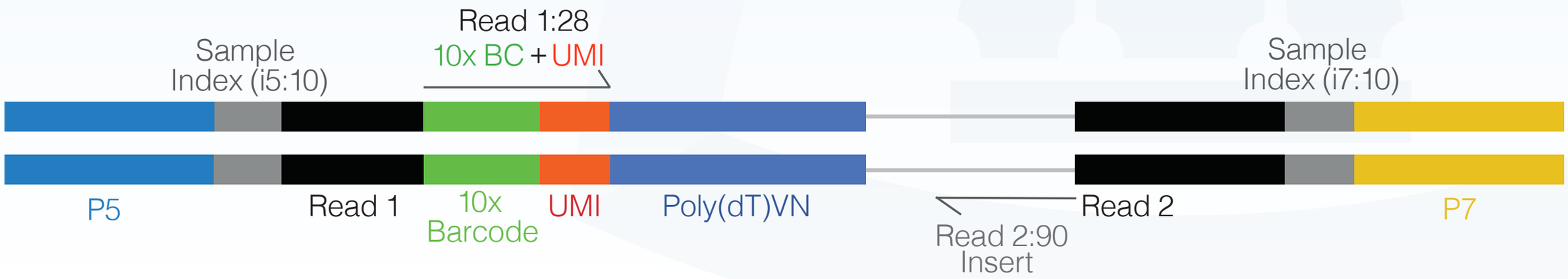
10x single cell RNA-Seq chemistry & library construction

Inside individual GEMs



- A 'paired end library' where one read contains the cell barcode + UMI & the other contains the gene info
- Specific alignment/ mapping algorithms required to simultaneously assign read to:
  - Barcode
  - UMI
  - Sequence (Gene)

10x Chromium Single Cell 3' Gene Expression Dual Index Library



Cellranger (10x Aligner)

- Analysis pipeline to process (exclusively) 10x Chromium single cell/nucleus data
  - Make fastq files from raw sequencing run
  - Align reads to reference
  - Generate count matrices
  - Perform some standard post-processing analysis

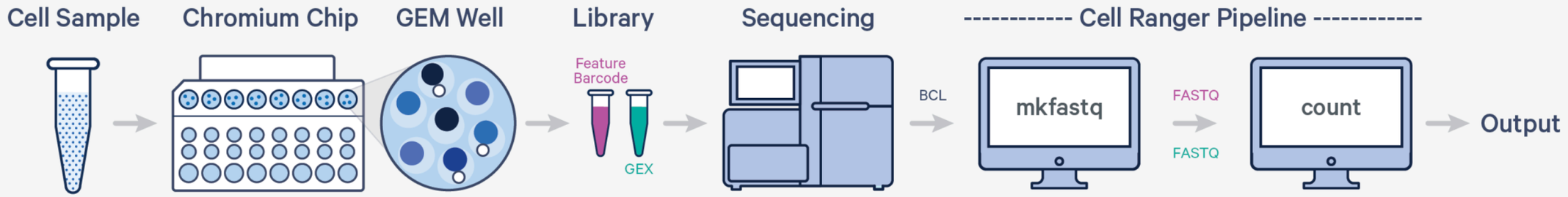
- Workflow information
  - https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger

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

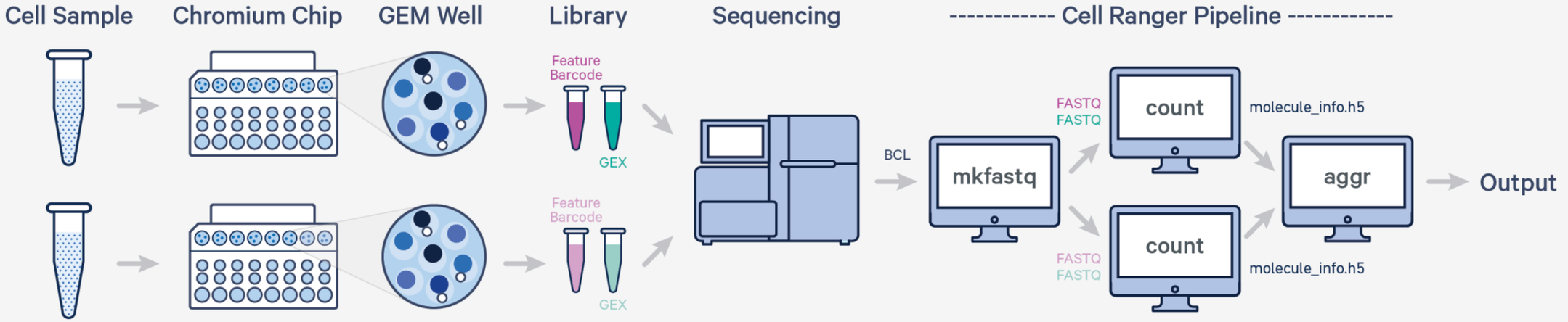
- Availability
  - Only available on linux
  - We will make this available on Rockfish for your use

Cellranger workflows

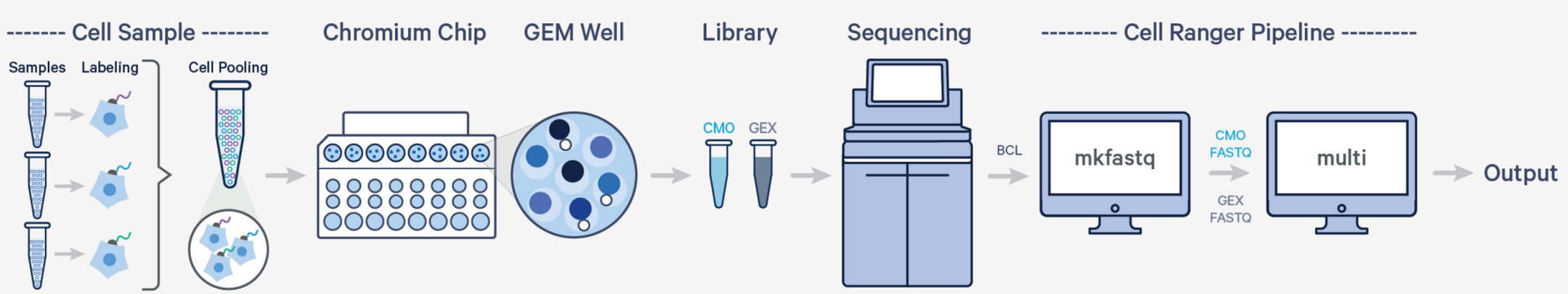
One sample, one GEM well, one flow cell



Multiple samples, multiple GEM wells, one flow cell



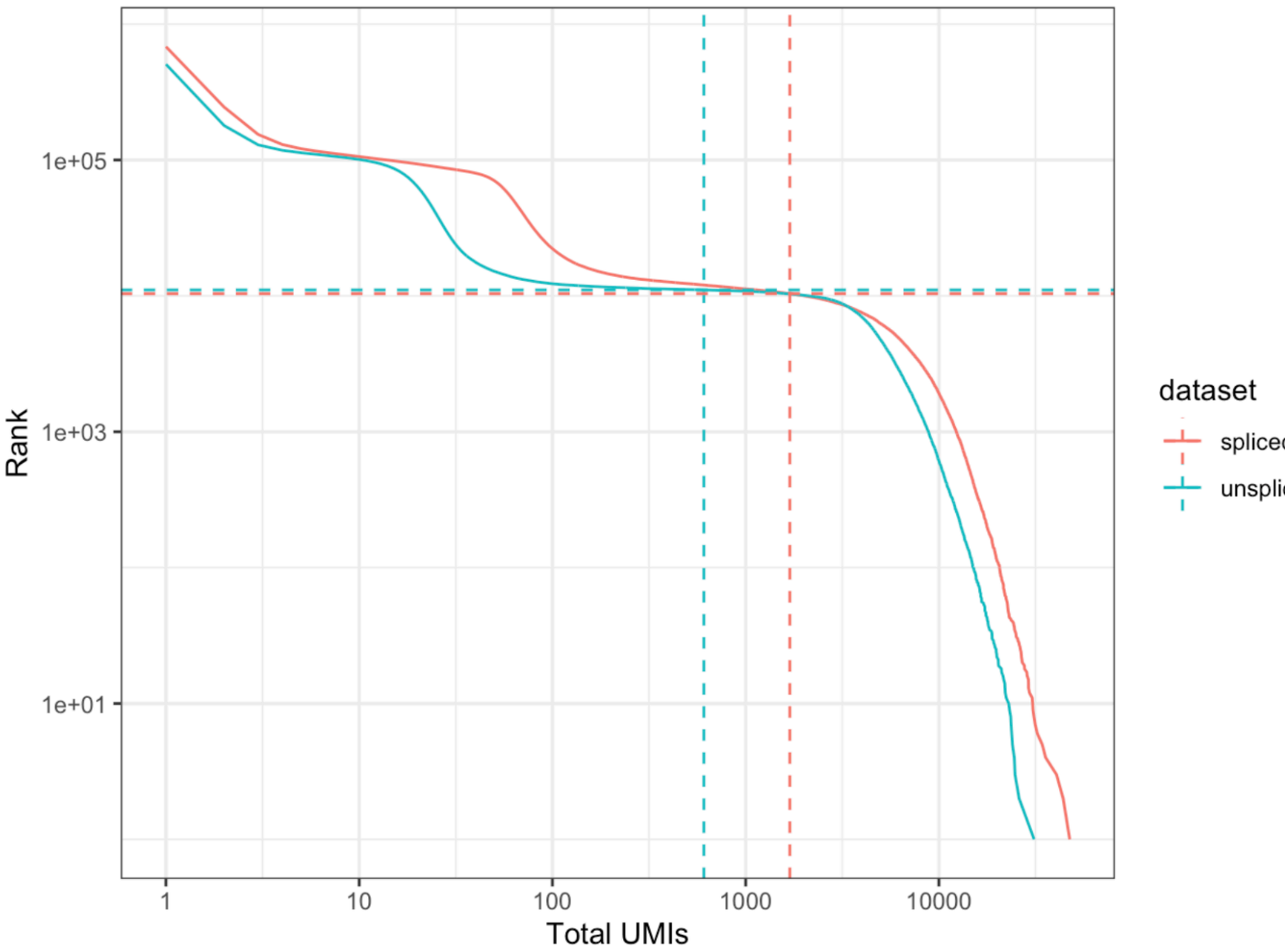
Multiple samples, one GEM well, one flow cell (Cell Multiplexing)



Kallisto + bustools

- A more generic single cell/nucleus preprocessing workflow based on kallisto alignment
  - https://www.kallistobus.tools/
  - https://www.nature.com/articles/s41587-021-00870-2
- Wrapped into a convenient tool called `kb` which can be installed as part of the `kb-python` package
  - `pip install kb-python`
- Performs:
  - Index retrieval/building
  - Sample pseudo alignment
  - UMI collapsing
  - Barcode correction
  - Cell demultiplexing
  - Droplet selection
- Sample aggregation, normalization and differential expression analysis can be performed with the accompanying `sleuth` R-package
  - Isoform- or gene-level analysis available
  - DE analysis with `sleuth` uses standard R model definitions.

Knee plot



- A filter to separate out ‘true’ beads that contain cells vs those of lesser quality and/or containing ‘background’ RNA
- Inflection point of curve used as a UMI threshold cutoff for beads containing cells
- In vivo analysis often shows multiple ‘knees’, why?

Main sources of variability in scRNA-Seq

- Two types:
  - Extrinsic (technical)
    - Some are manageable, others are unavoidable.
  - Intrinsic (biological)
    - These are the interesting bits

Source of bias	Type	Effect	Current solutions
RNA capture and RT efficiency	Technical	Stochastic zeroes	Spike-ins, statistical modelling
cDNA amplification	Technical	Loss of quantification accuracy	UMIs, statistical modelling
Batch effects	Technical	Introduce a signal different from the true biological signal	Statistical modelling
HVGs, transcriptional burst	Biological	Increase variance in the data	Statistical modelling
Cell-cycle stage, differentiation state, etc.	Biological	Confuse the true biological signal	Cell visualization, statistical modelling

- Solutions:
  - Proper controls, replication, and managing your logistical workflow(s)
  - Some biases can be either be corrected (fixed) or modeled (fit) in downstream analyses.
  - Document as many features/parameters for each sample as possible
    - Can be useful in identifying *where* bias arose and facilitate statistical modeling

Modeling some aspects of technical variation in scRNA-Seq data

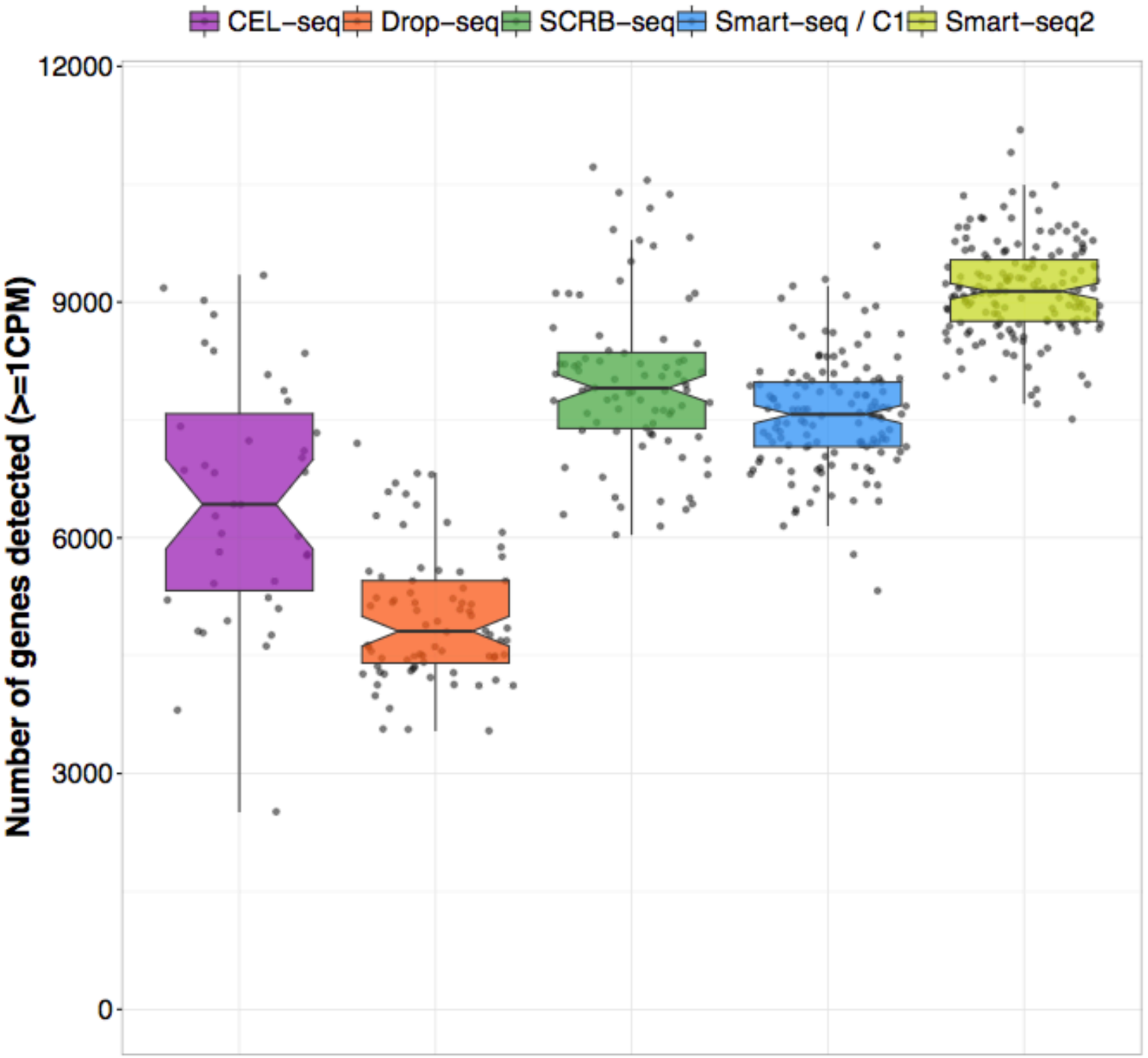
- Variation is not consistent with the mean
  - heteroscedastic

- Most analysis approaches model the mean:variance relationship
  - Assumption is that this fit represents the technical variance associated with scRNA-Seq
    - Capture efficiency

- One mitigation strategy is to focus analyses on genes with high-residuals to this fit.

- Caution: Different conditions (cell types, treatments, batches, etc) can result in significantly different fits.

‘Number of detected genes’ as a proxy for sensitivity

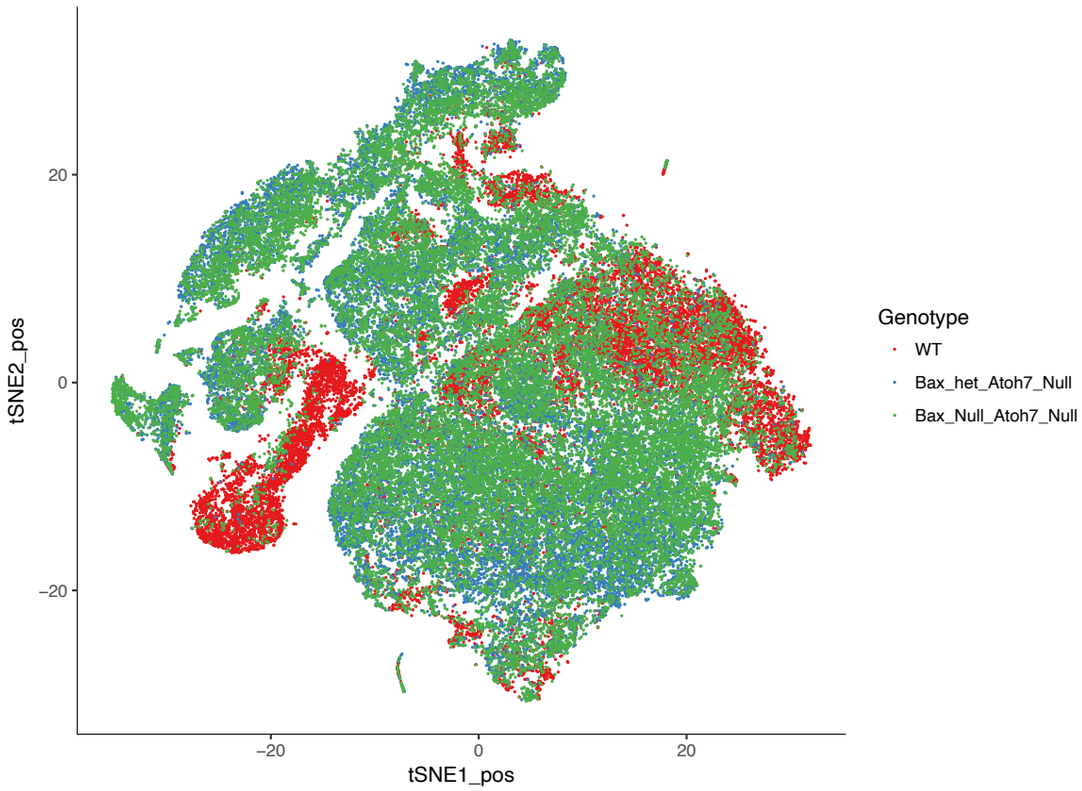


Ziegenhain et al. 2017

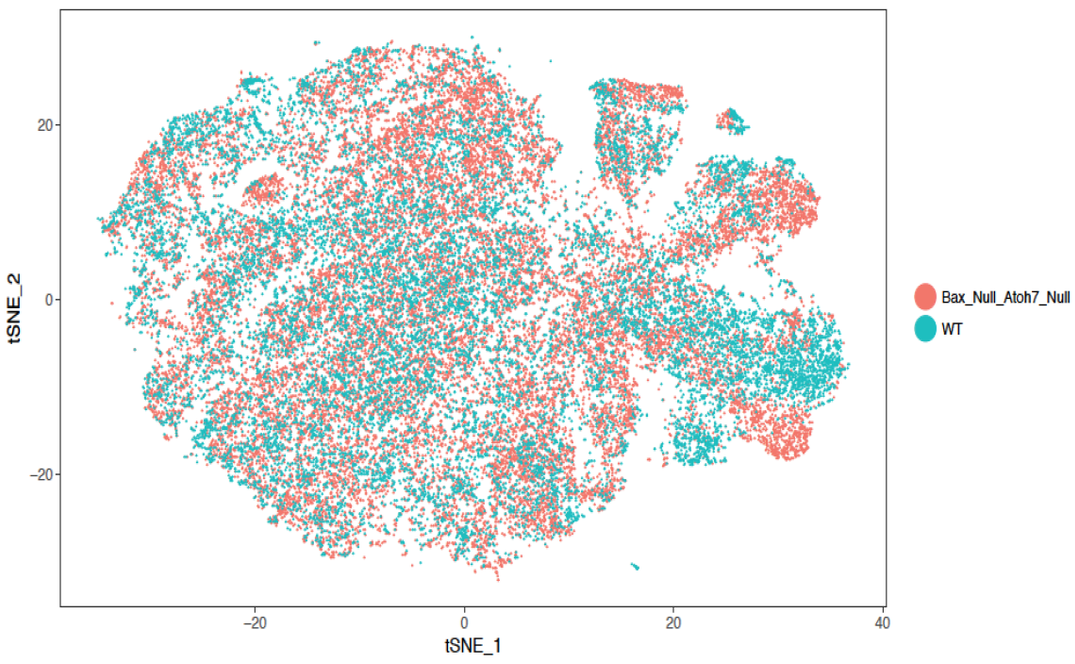
- nGenes is widely used to estimate sensitivity and as a proxy measure for capture efficiency
  - Per sample/run
  - Per cell
- Full-length cDNA Template-switching methods generally detect more genes (with lower throughput)
  - Usually sequenced to a greater depth per cell as well.
- Difficult to standardize this metric across different experiments
  - Different cell types have different # of expressed genes
- Capture efficiency and ∴ # genes is also a function of cell size

‘Fixing’ batch effects in single cell RNA-Seq

Before alignment



After alignment



- The Seurat R-package enables one approach to ‘correcting’ batch effects
- Many other methods available
  - ComBat
  - SVA
- Aligns conditions to a shared manifold in reduced dimensional space
- A significant consequence is reduced power for differential analysis
- Community remains divided as to ‘fixing’ or ‘fitting’ these effects.

Analysis courtesy of Justin Brodie-Kommit, Brian Clark, Samer Hattar, and Seth Blackshaw

Popular Frameworks for scRNA-Seq analysis

- R
  - ***SingleCellExperiment*** - https://doi.org/doi:10.18129/B9.bioc.SingleCellExperiment
  - ***Monocle3*** - https://cole-trapnell-lab.github.io/monocle3/
  - ***Seurat*** - https://satijalab.org/seurat/
  - ***Scater*** - https://github.com/davismcc/scater
  - ***Scran*** - https://bioconductor.statistik.tu-dortmund.de/packages/3.4/bioc/html/scran.html

- Python
  - ***Scanpy*** - https://doi.org/10.1186/s13059-017-1382-0
  - ***Loompy*** - https://linnarssonlab.org/loompy/
  - ***scVI*** - https://www.nature.com/articles/s41592-018-0229-2

- Other
  - ***Granatum*** - https://gitlab.com/uhcclxgg/granatum

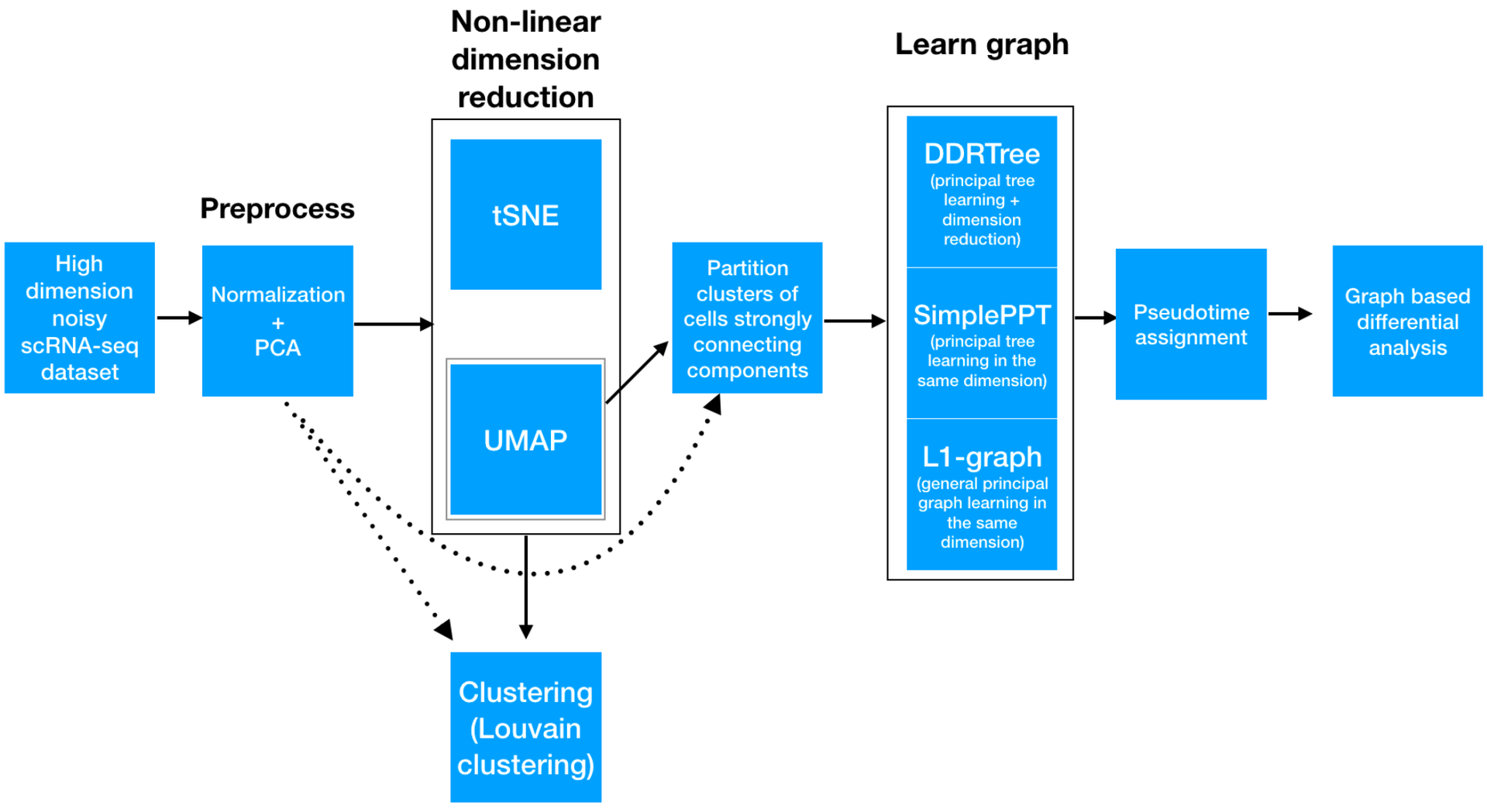
The Dataset

- Dissociated E18 mouse cortex
- 10x Genomics 3’ Gene Expression analysis
  - ~10,000 cells targeted

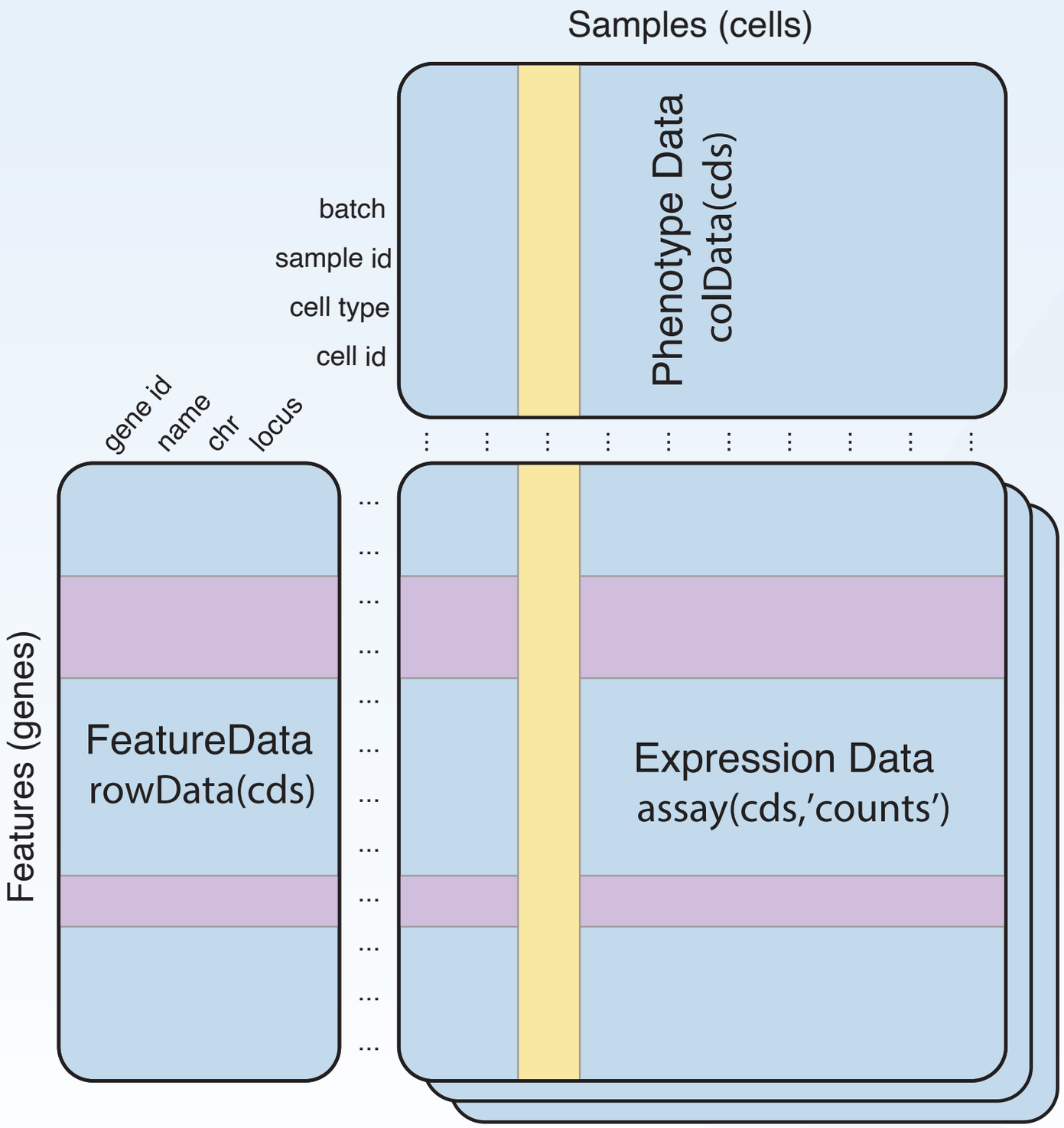
- Objectives for this week:
  - Import single cell count matrix and associated metadata into Monocle3 Framework (R/Bioconductor)
  - Learn methods for QC analysis of cell and gene quality from scRNA-Seq
  - Understand principles of dimensionality reduction and how to interpret cells in a UMAP embedding
  - Evaluate clustering and cell type annotation for scRNA-Seq data
  - Learn to perform differential gene expression between conditions and across pseudotime trajectories
  - Identify ways to learn patterns of co-regulated gene expression in single cell RNA-Seq data

Monocle3 Framework & Workflow

- Originally designed for ‘pseudotime’ analysis
- Expanded into a fully-functional scRNA-Seq framework
-

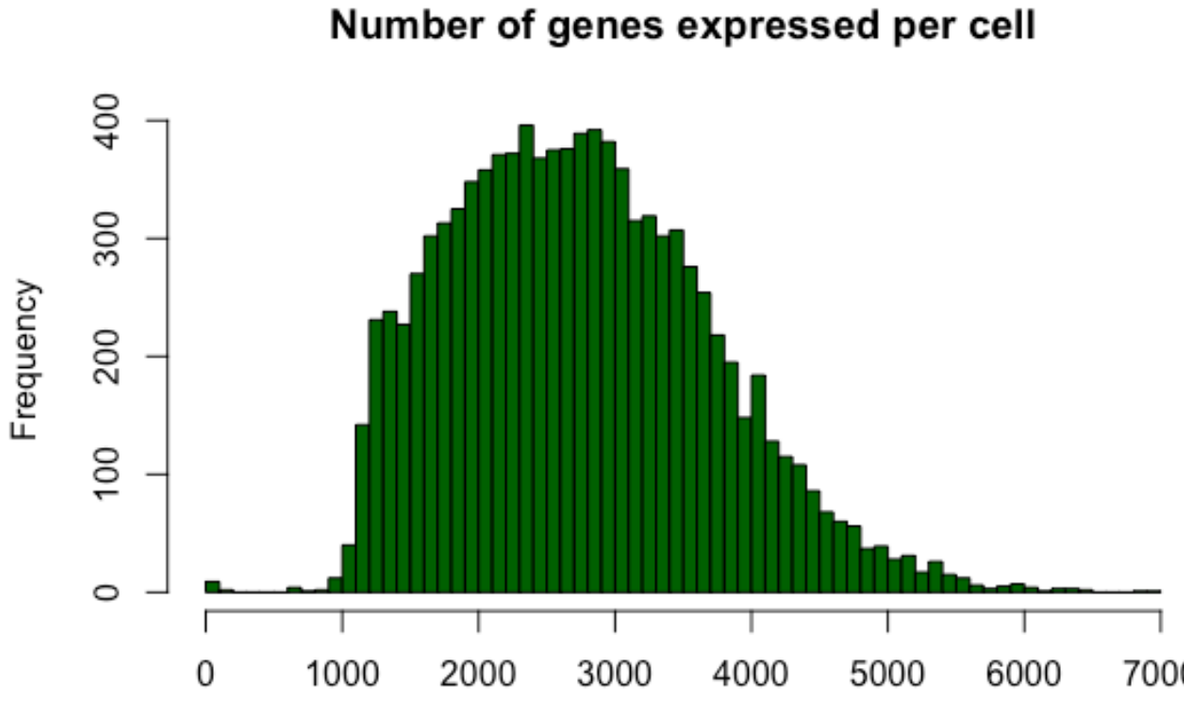


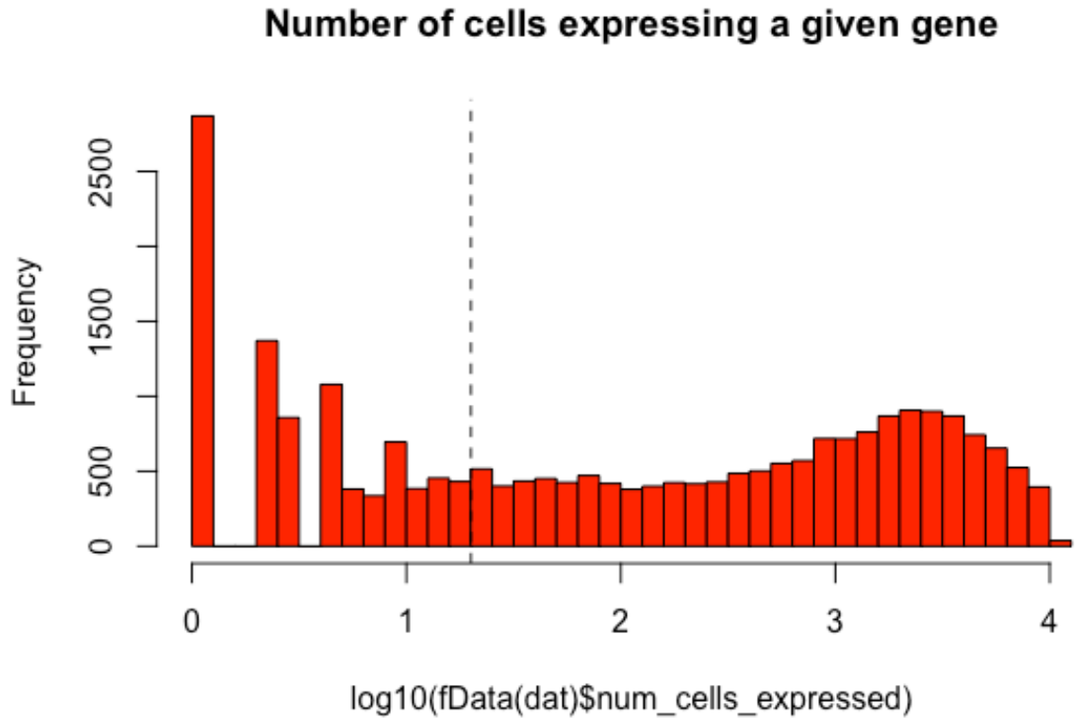
The SingleCellExperiment class (Bioconductor)

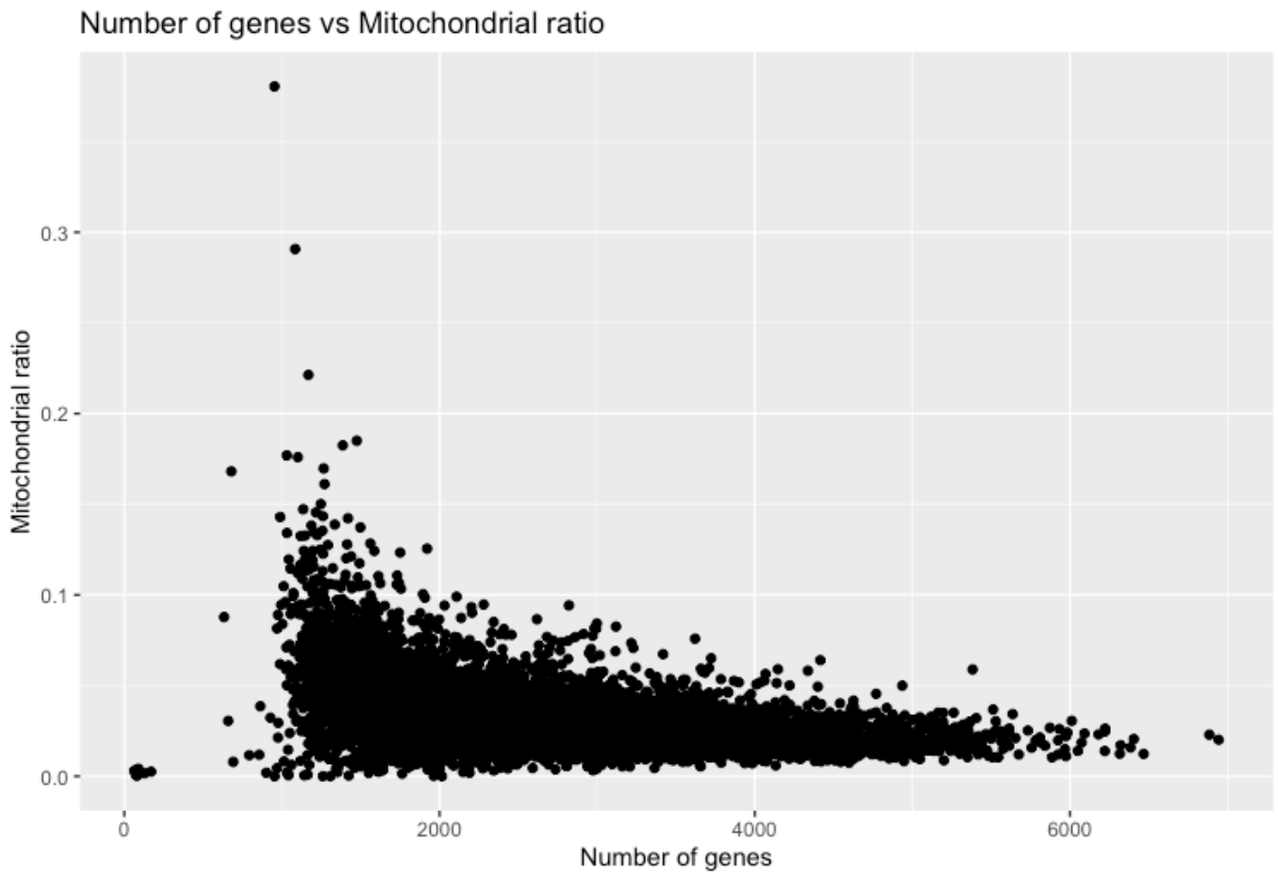


- Extends the RangedSummarizedExperiment class
- Many other classes (like Monocle3 cell_data_set) further extend this standard class
- Most frameworks define similar object structures to hold gene expression data.
- Gene features and sample (phenotype) information are stored and indexable with expression matrix.
- Enables slicing of the data to preserve annotations on both dimensions
- Other analysis features are stored in accessory slots (not show)
  - Reduced dimensions
  - Graph representations
  - Etc.

Quality Control Metrics







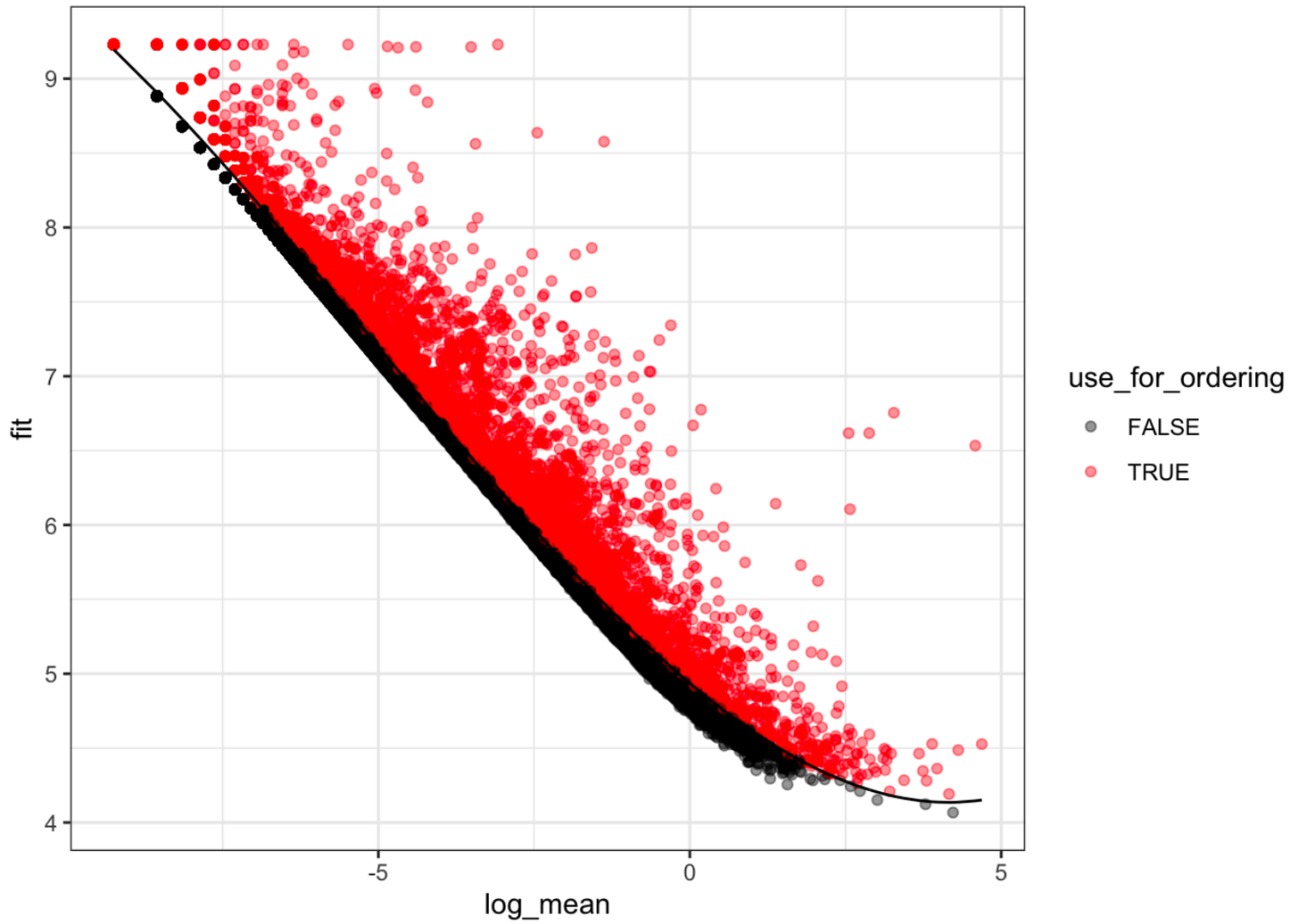
- Low quality cells:
  - Total mRNA count per cell
  - Number of expressed genes per cell
  - % of counts derived from mitochondrial genes
- Gene QC:
  - Mean expression level
  - Minimum # of cells with detectable expression

High-variance genes

- Per-gene technical variance is well modeled by a negative binomial fit.

- Genes with excess variation (overdispersion) *should* have high biological variation in addition.

- We select this subset to highlight differences between cell types and states for downstream analyses

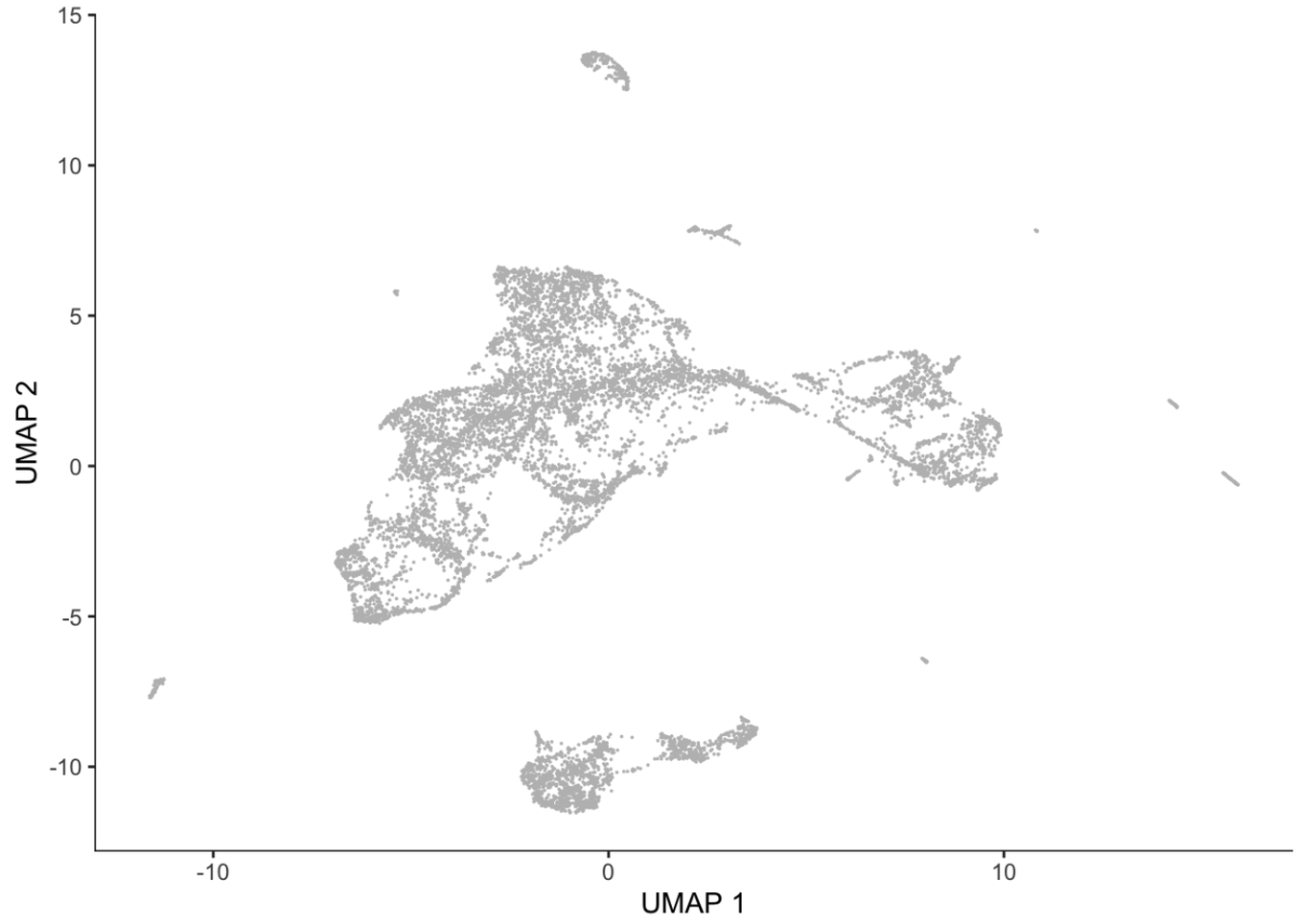


Dimensionality Reduction

- Summarizes *n* principal components into a visually interpretable 2-dimensions
  - Could also be 3 as we will see

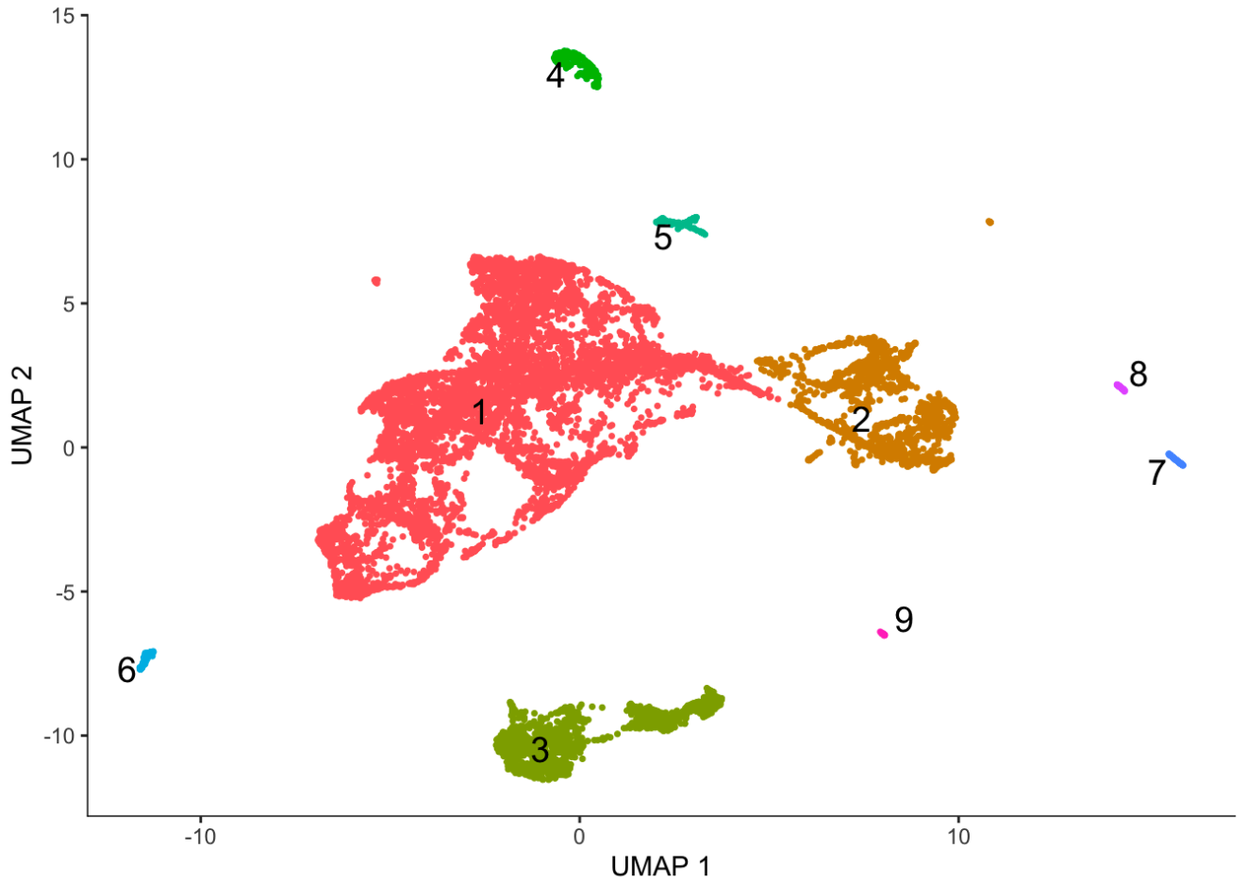
- Popular non-linear algorithms include t-SNE and UMAP

- Important to remember that there are an infinite number of ‘embeddings’ (views) of a given dataset.



Partioning into subtypes

Leiden community detection



- Multiple algorithms available to identify groups of cells with similar expression profiles
  - K-Means
  - Gaussian Clustering
  - KNN
  - Community Detection (Louvain/Leiden)

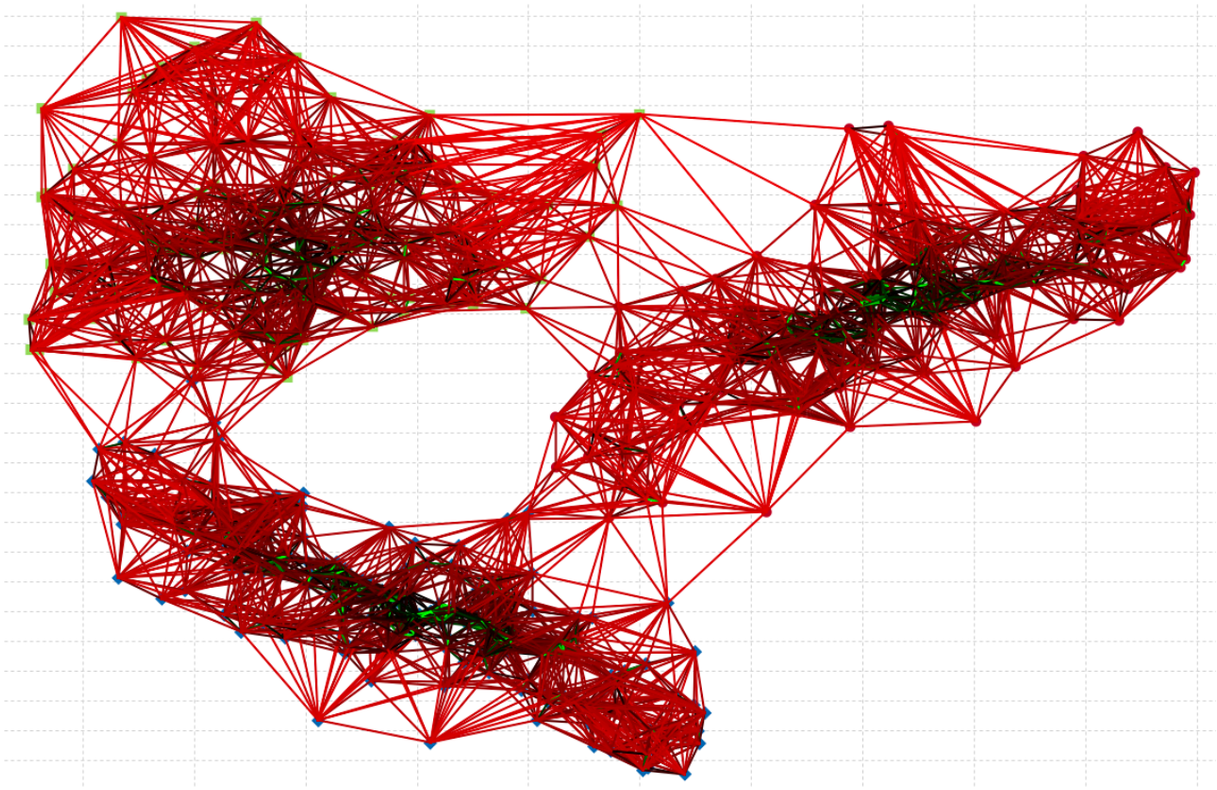
- Clustering is often an arbitrary decision-making process.

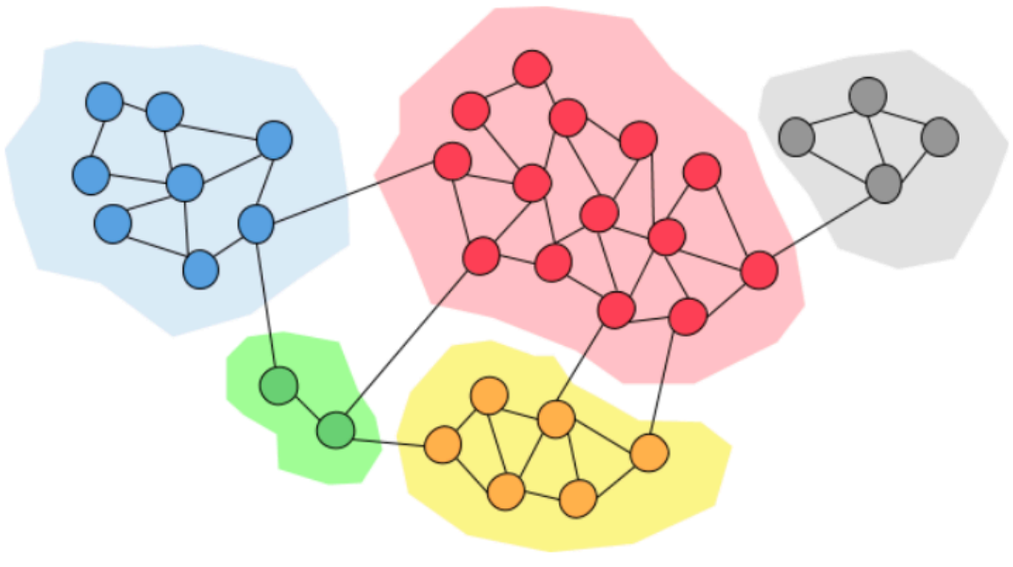
- Parameter sweep to identify ideal clustering solutions

- Clusters should NOT be treated as definitive cell types.
  - Requires validation

- Never a good idea to cluster in low dimensional representations.
  - All low-dim representations are ***approximations*** with ***distortions*** of the true high-dim relationships

Graph-based methods in scRNA-Seq





- Increasingly, graph-based algorithms are used to represent relationships between cells (or genes) in high-dimensional space.
  - Speed
  - Reduced memory data structures

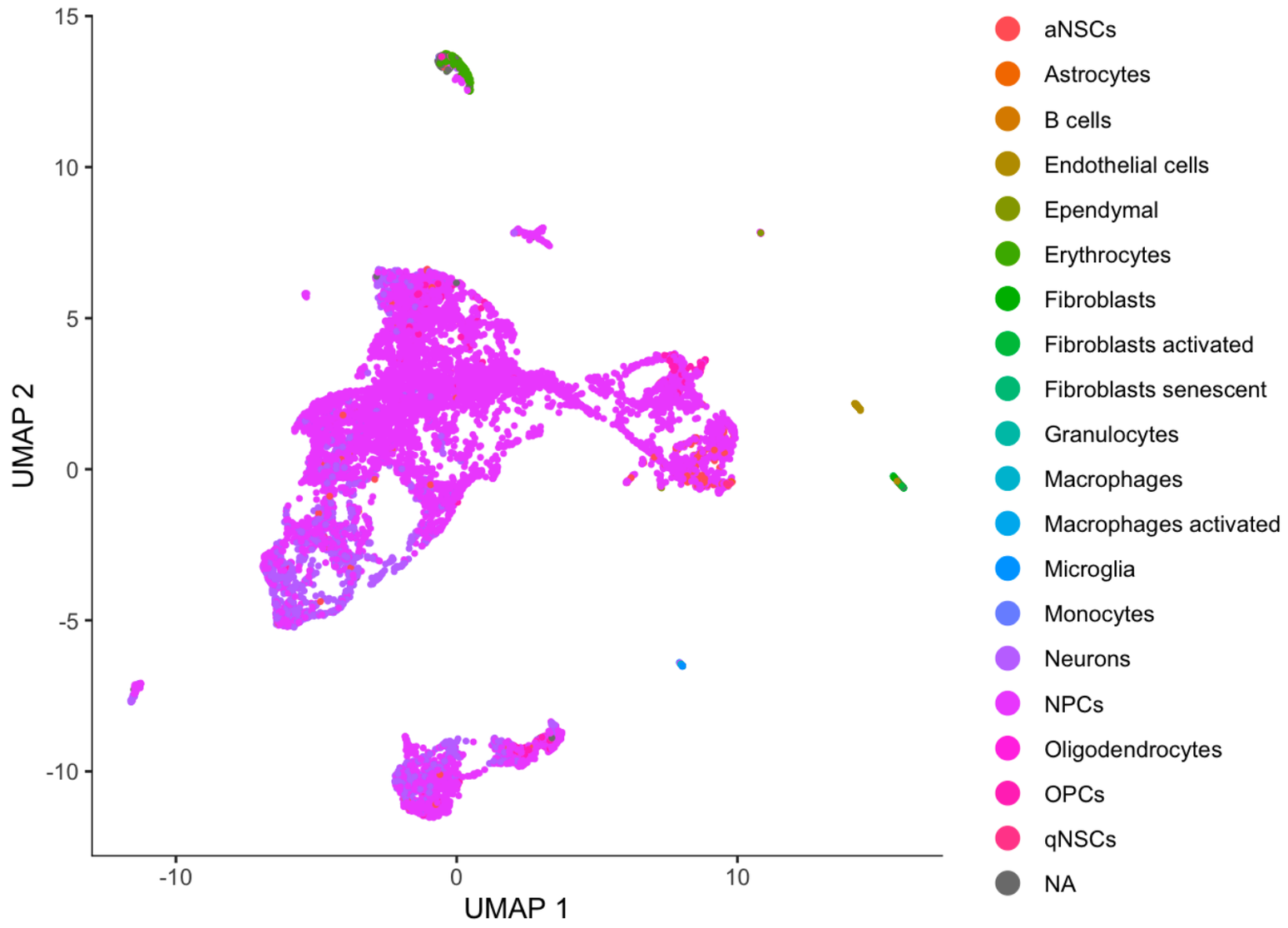
- Nodes (cells) are connected to each other (via edges) by their relative similarities

- Often used for modeling interactions across complex systems

- Graphs can contain ‘communities’ of nodes that are more tightly connected to each other.

- Many algorithms to perform ‘community detection’
  - Graph-equivalent of clustering

Cell Type Annotation



- Can be done using marker gene expression, identifying patterns of co-regulated genes, and/or transfer learning using existing annotations to train a classifier.
  - Each approach has drawbacks and advantages.
- SingleR is an R/BioC package that annotates clusters of single cells based on their similarity to established ‘reference’ transcriptome profiles.

Differential Expression - Likelihood ratio test

- Compares two linear models (full vs reduced) for a given gene to determine whether the parameters lost in the reduced model explain a significant amount of variance in the data

Full model (m1)

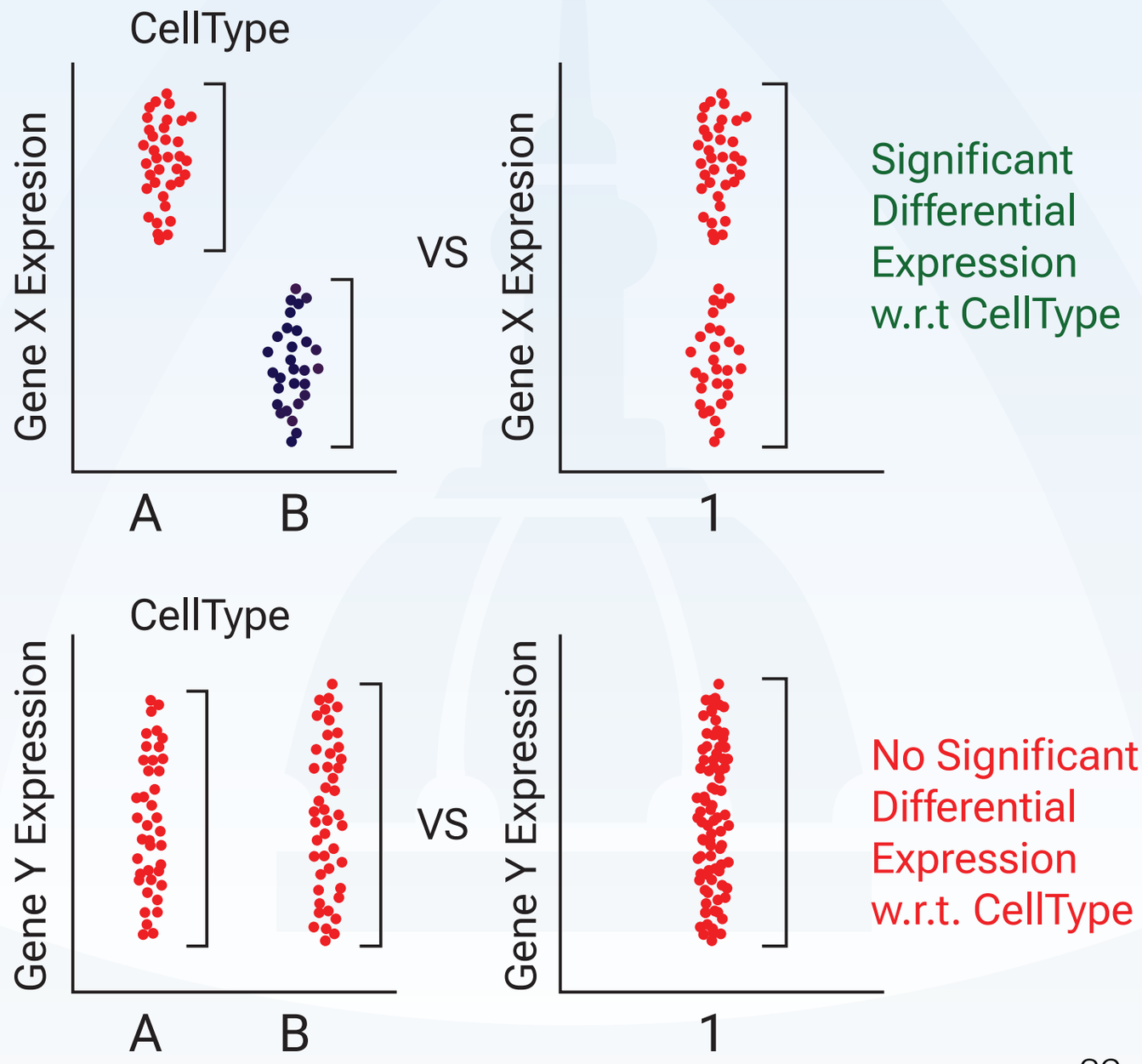
$$y_{gbt} \sim nGenes + Batch_b + CellType_t$$

Reduced model (m2)

$$y_{gbt} \sim nGenes + Batch_b$$

Likelihood Ratio Test

$$lr = -2\frac{ln(L_{m_1})}{L_{m_2}}$$

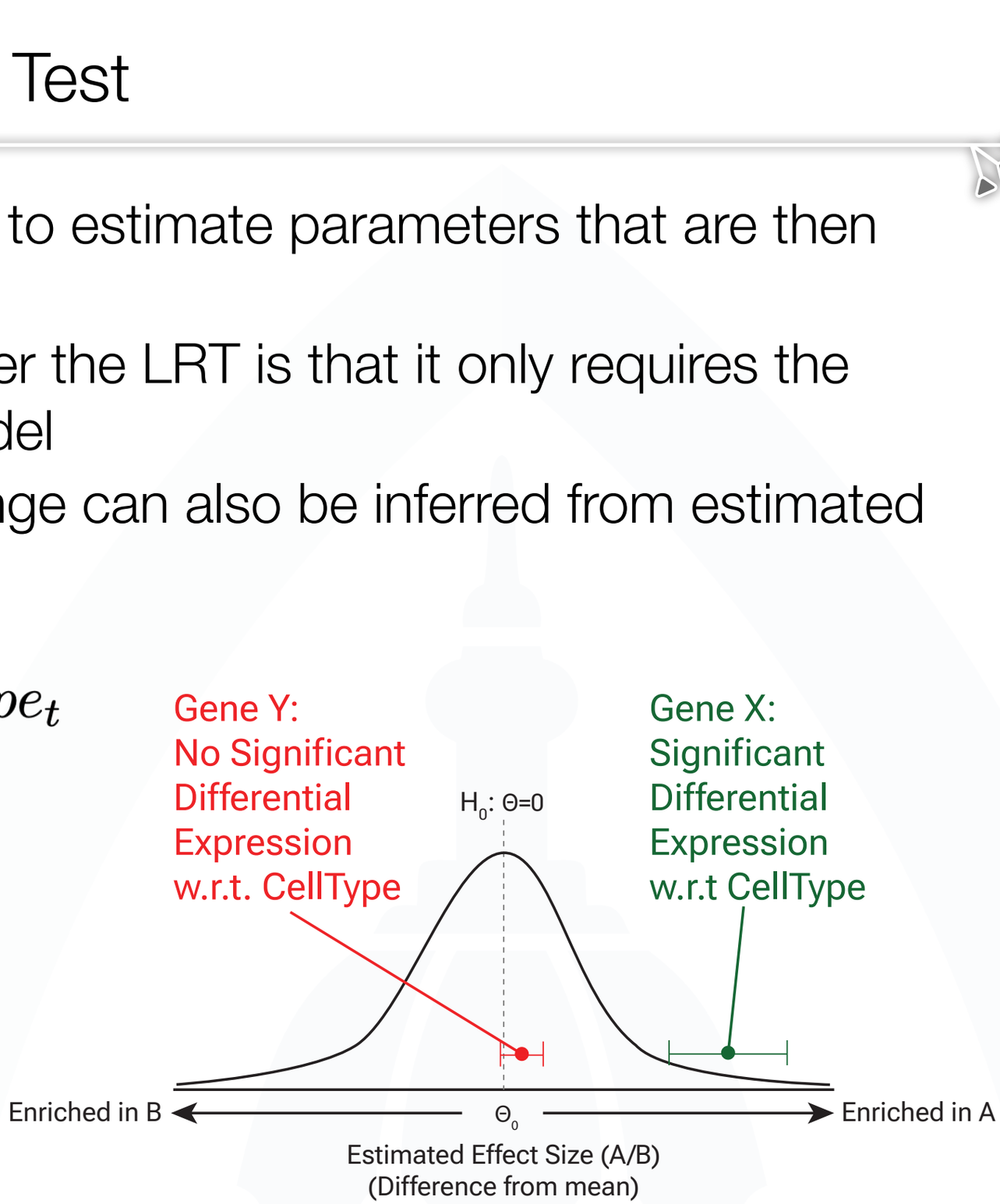


Differential Expression - Wald Test

- Uses a ‘reasonably well fit’ model to estimate parameters that are then tested
- An advantage of the Wald test over the LRT is that it only requires the estimation of the unrestricted model
- Direction of gene expression change can also be inferred from estimated effect

$$y_{gbt} \sim nGenes + Batch_b + CellType_t$$

$$W = \frac{(\hat{\theta} - \theta_0)^2}{\mathrm{var}(\hat{\theta})}$$



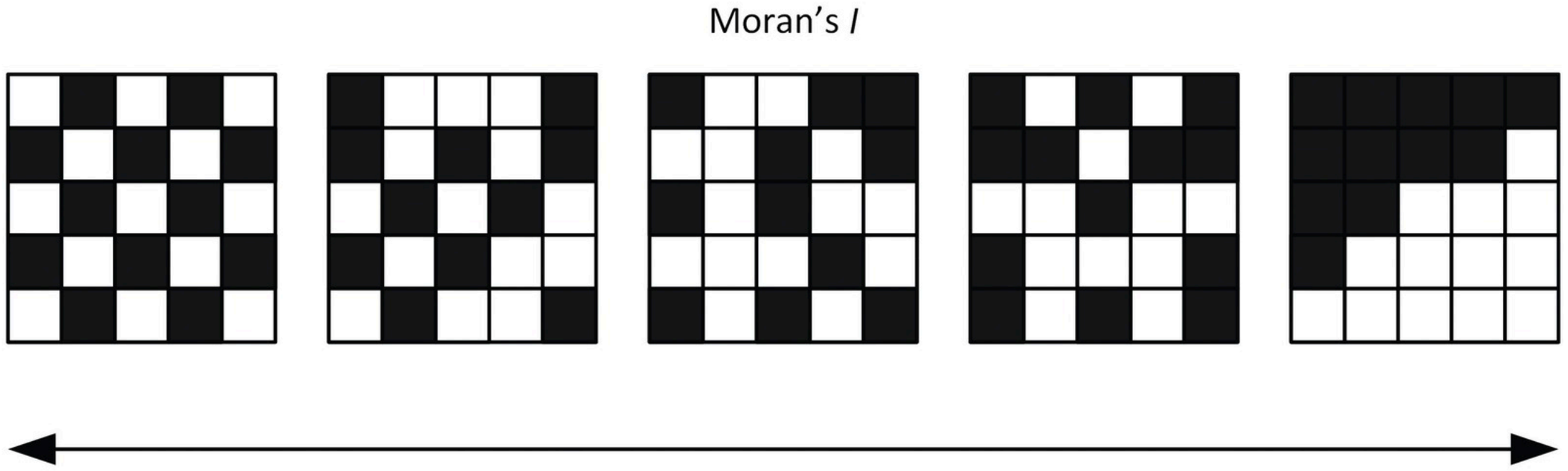
- Tests whether the learned effect for a given parameter deviates significantly from the null hypothesis H0: θ=0

Differential Testing - Principal Graph Test

- Moran's I test (spatial [auto]correlation)

- Tests whether features are randomly distributed across a spatial landscape
  - In practice, this is performed on the learned principal graph

- Can also determine whether features (genes) are correlated with a path/ trajectory across a spatial embedding as well
  - Test for differential gene expression as a function of pseudo time embedding.

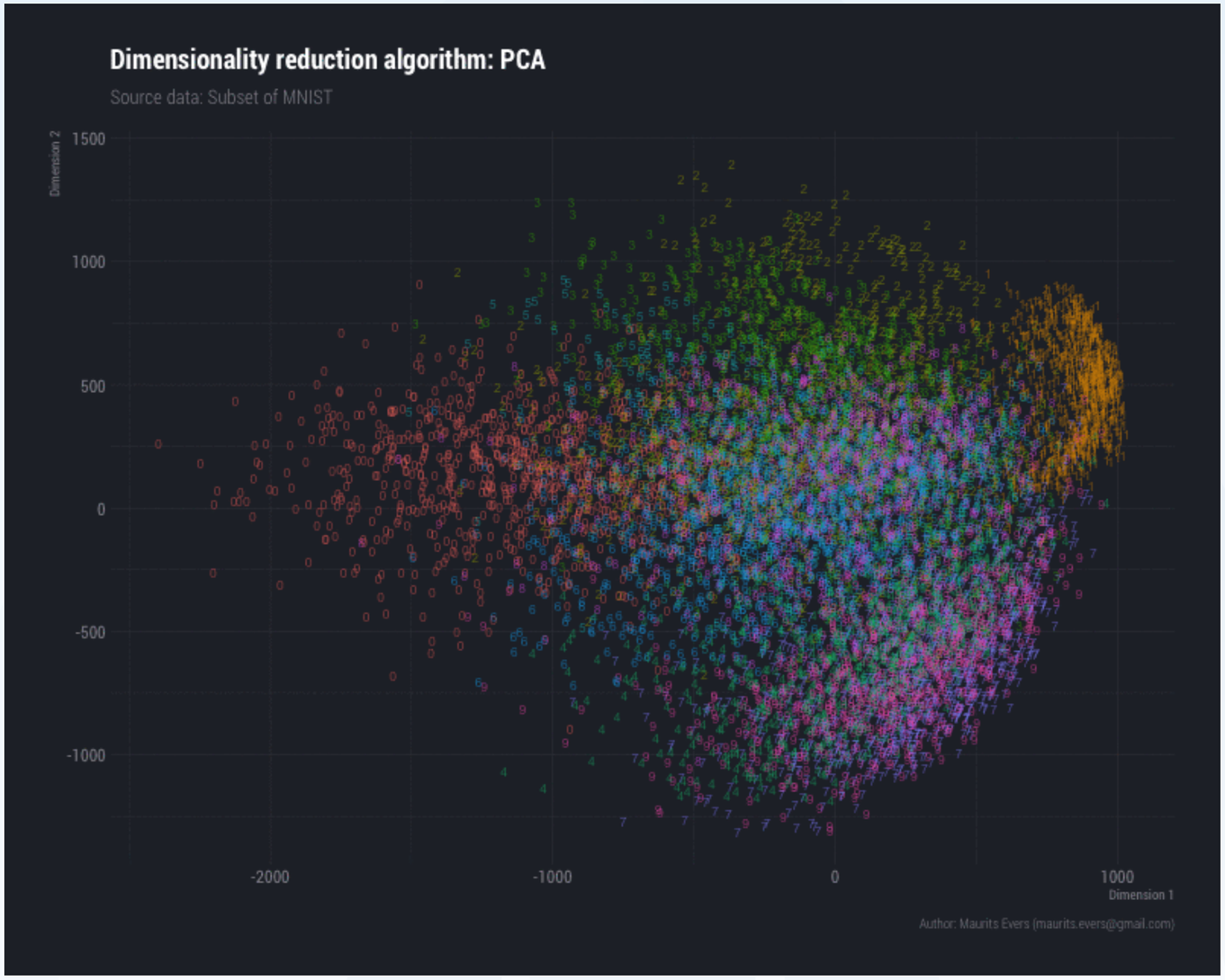


Dimensionality reduction is a cornerstone of single cell data analysis

- Diversity of linear and non-linear methods to identify ‘hidden’ latent spaces and reduce ‘complexity’ of HD data

- Usually, analysis focuses on ***outcomes*** of dimensionality reduction
  - Cluster identification
  - Cell States
  - Trajectories

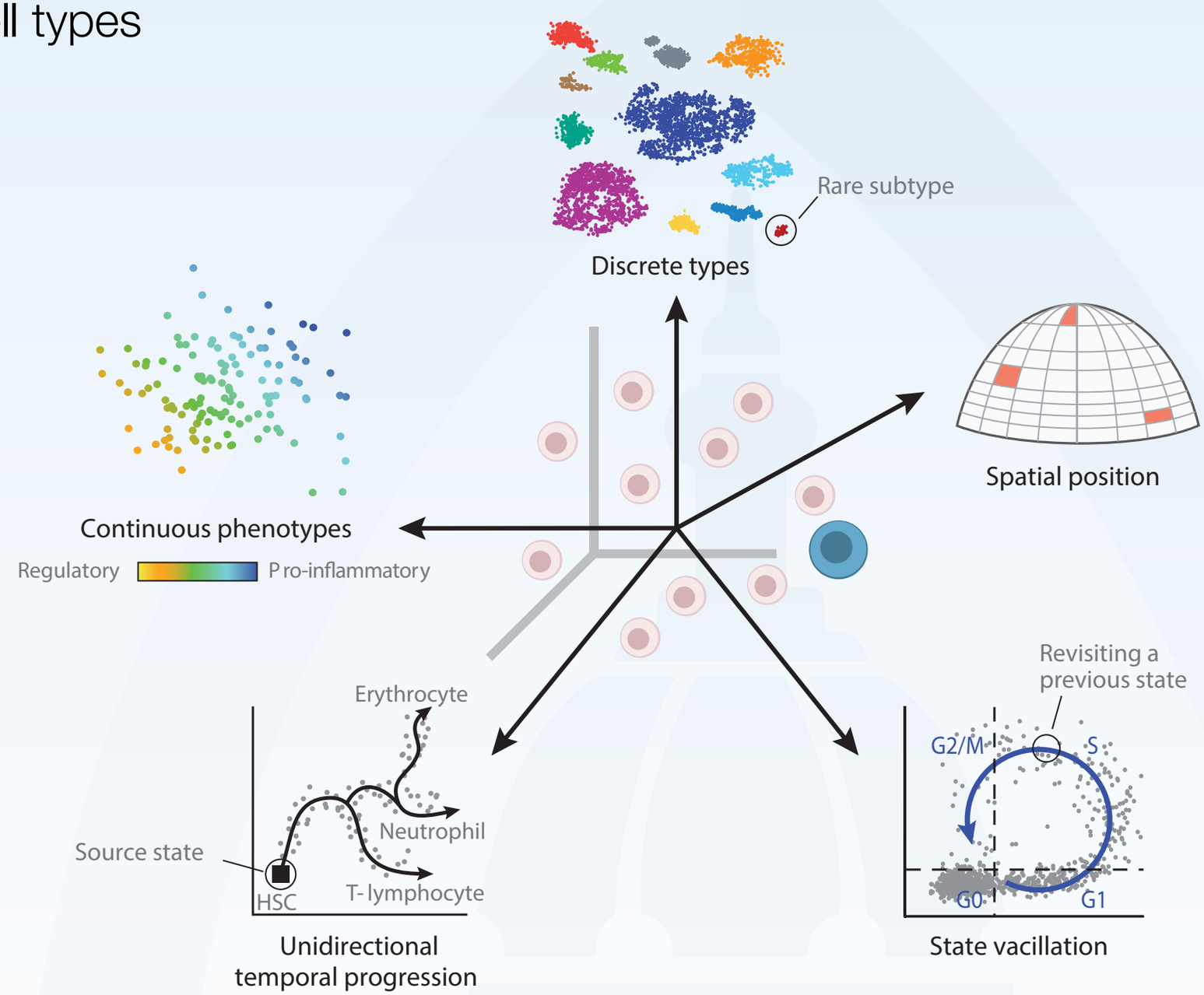
- Latent spaces ***themselves*** often encode biological processes or technical features



https://github.com/mevers/animated_dimensionality_reduction

Latent spaces of gene expression reveal axes of cellular identity

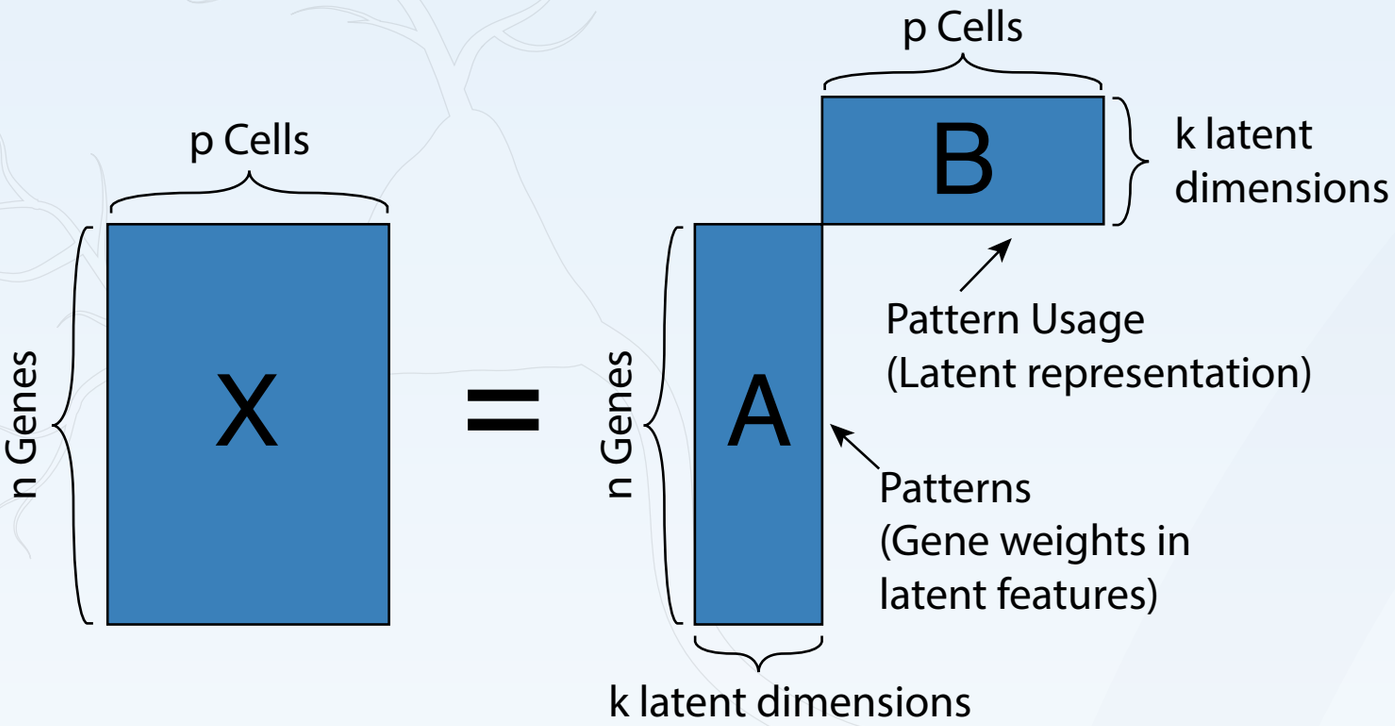
- Taxonomy
  - Discrete (irreducible?) cell types
- Organization
  - Spatial position
  - Tissue composition
  - Cell-cell communication
- Stimulus response
  - O_2/nutrient sensing
  - Cell signaling
- Physiology
  - Cell cycle
  - Metabolic activity
- State Transitions
  - Development
  - Disease onset/progression



Wagner et al, Nat. Biotech. 2016

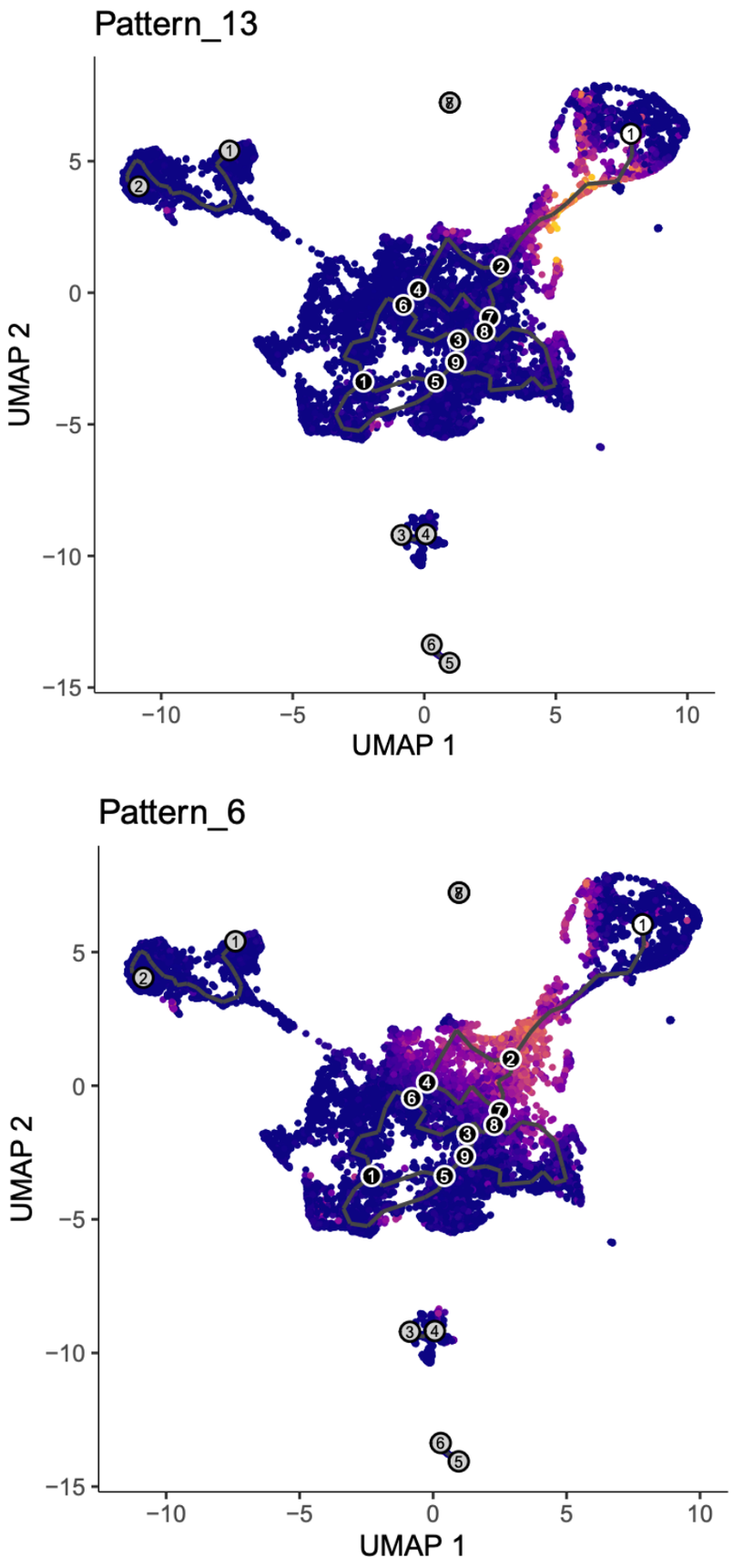
Latent space identification via Non-negative Matrix Factorization (NMF)

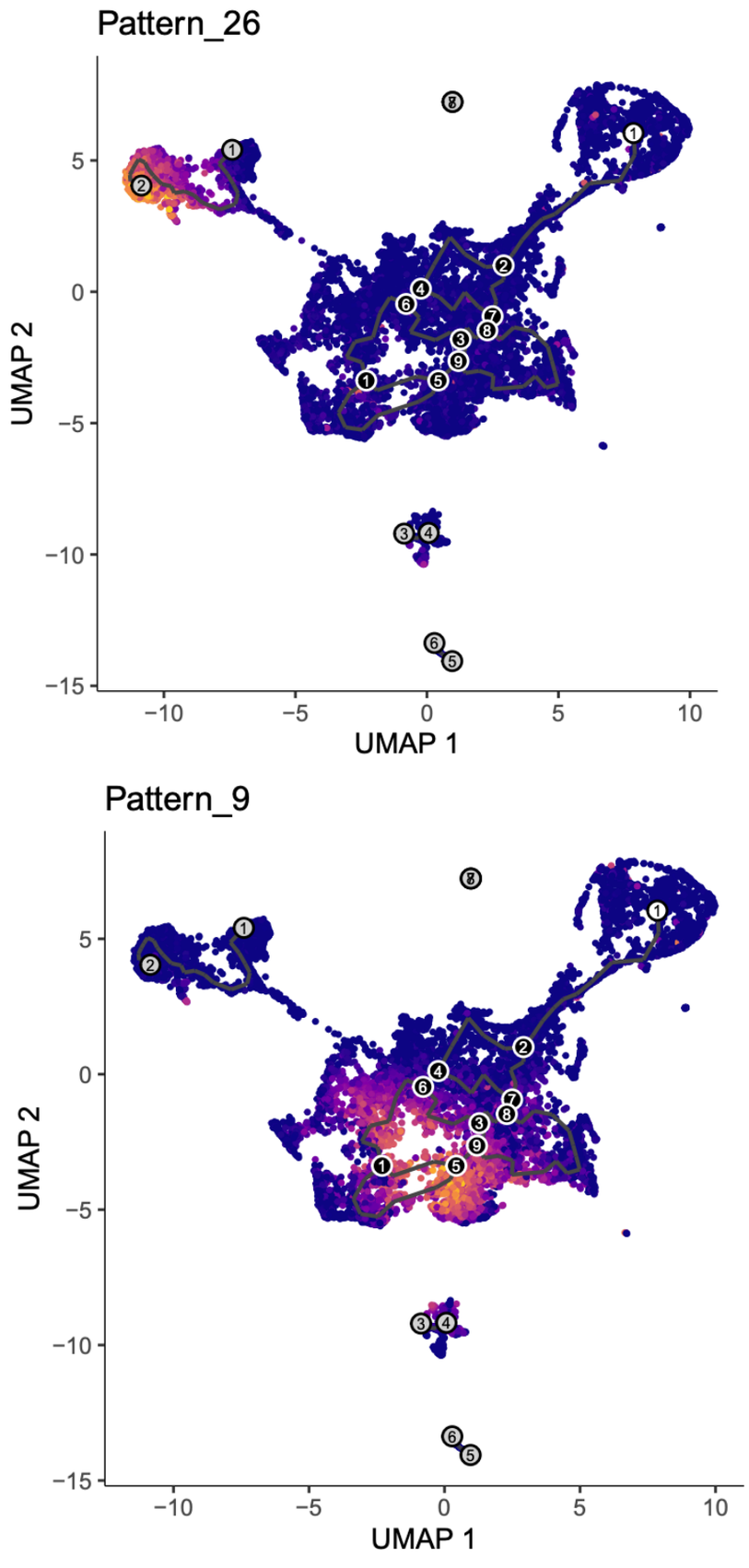
$$X = AB$$
$$n \times p \qquad n \times k \quad k \times p$$



- Matrix decomposition methods (including nonnegative matrix factorization) can be used to identify ‘patterns’ of co-regulated gene expression
- For NMF, solutions constrained to non-negative values for each matrix
  - Consistent with non-negative values for gene expression
- Columns of (**A**) reflect weighted relationships between co-expressed genes
- Rows of (**B**) reflect relationships between samples/cells/ conditions
- Learned patterns may correspond to:
  - Cell type identities
  - Biological processes
  - Spatial gradients
  - Other cellular features

Axes of variation are identified as (latent space) modules of coordinated gene expression

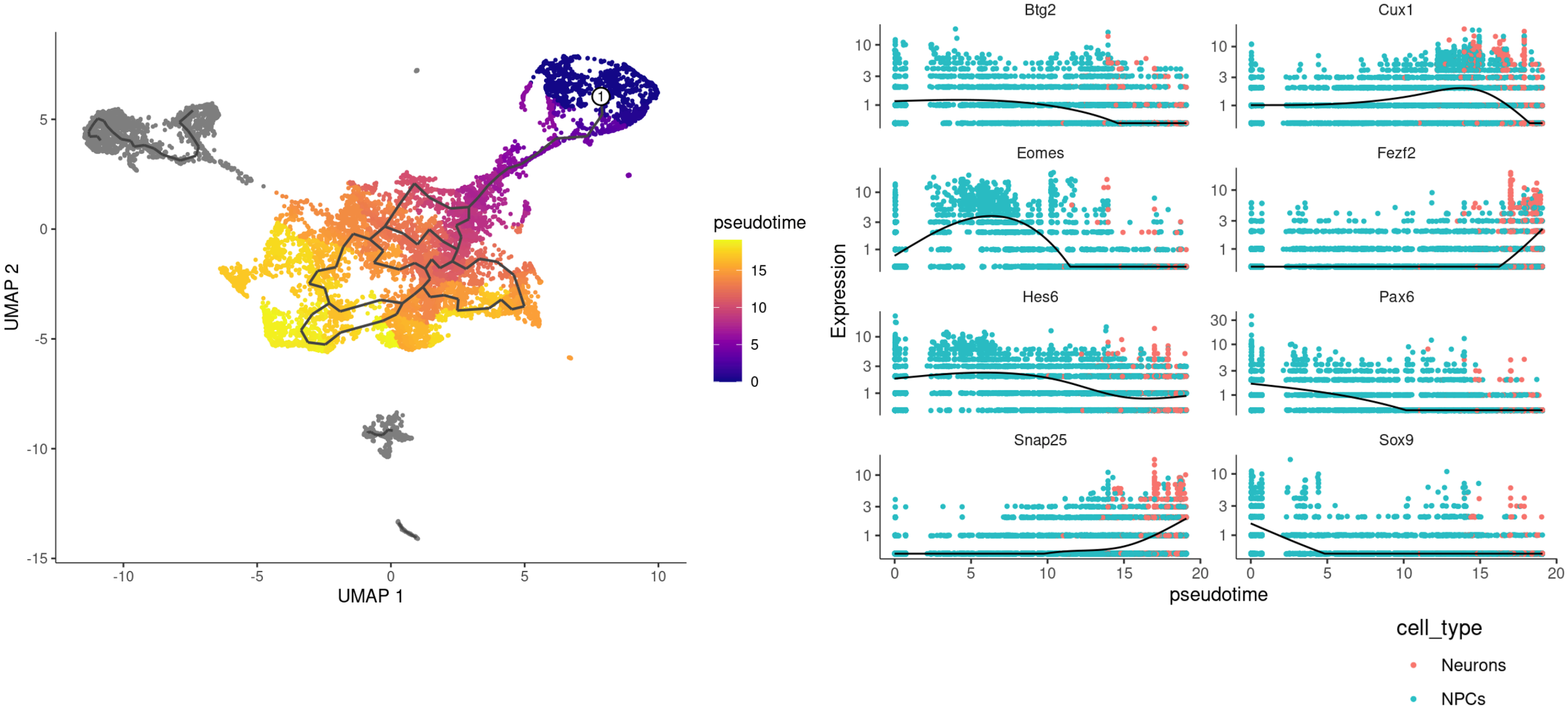




Pseudotemporal analysis to characterize continuous biological processes

Pseudotemporal analysis to characterize continuous biological processes

Pseudotemporal analysis resolves continuous biological processes



- Fitting a principal curve or graph to a partition as a ‘proxy’ for progression along a biological process such as differentiation

- Regression testing along this fit can identify genes whose expression changes along this process

Additional resources

- Monocle3 Tutorials:
  - https://cole-trapnell-lab.github.io/monocle3/

- scRNA-Seq tutorials:
  - https://hemberg-lab.github.io/scRNA.seq.course/index.html
  - https://satijalab.org/scgd18/

- Databases of scRNA-Seq tools:
  - https://github.com/seandavi/awesome-single-cell
  - https://github.com/Oshlack/scRNA-tools

- Publicly available datasets:
  - https://github.com/czi-hca-comp-tools/easy-data
  - https://preview.data.humancellatlas.org/

- Orchestrating Single Cell Analysis - Book (Hicks)
  - https://bioconductor.org/books/release/OSCA/