JOHNS HOPKINS
MEDICINE

McKUSICK-NATHANS
Institute of
Genetic Medicine

Single Cell RNA-Seq: NeuroCog II Lab

Loyal A. Goff, Ph.D.

February 18th, 2019

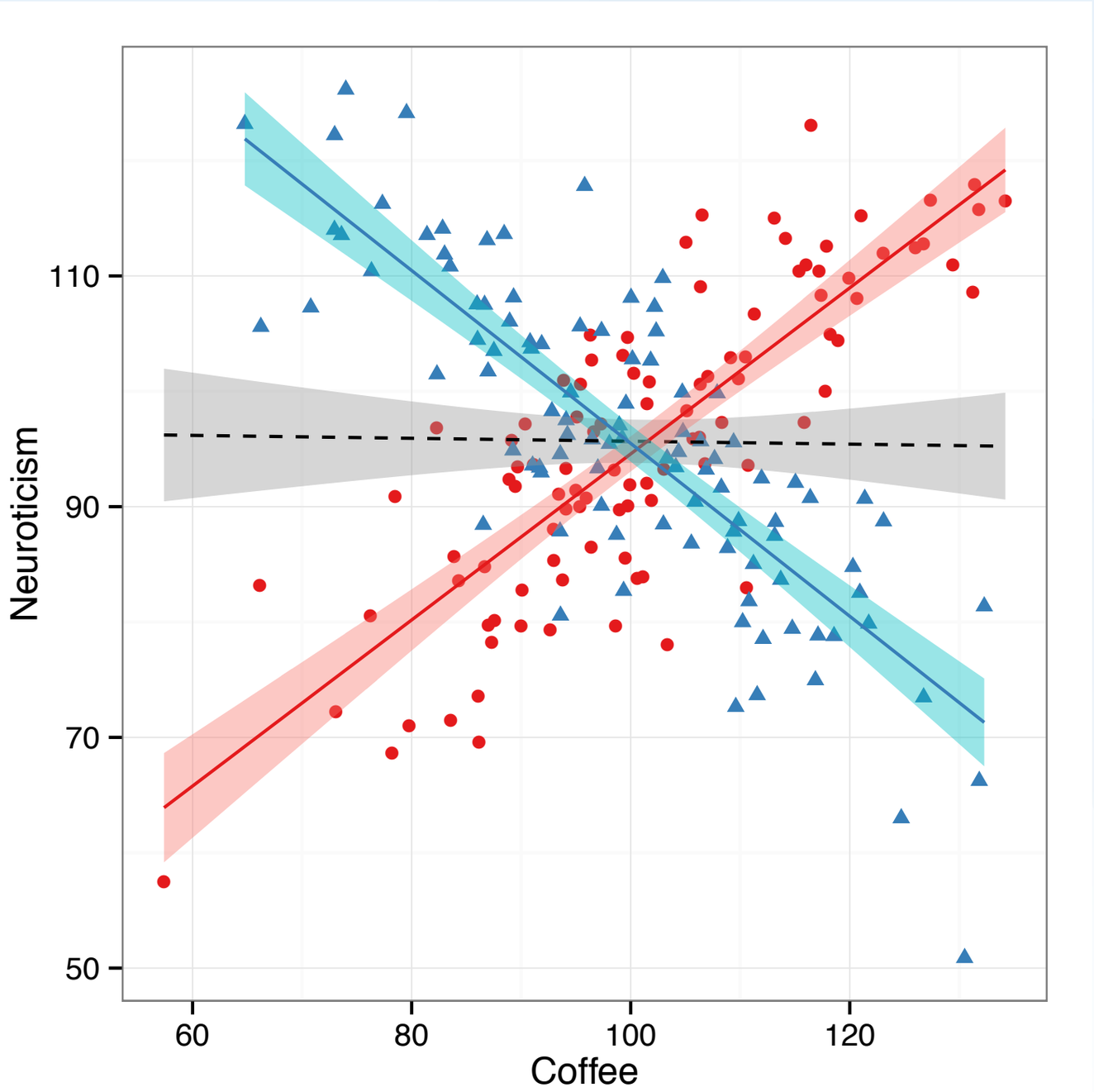
The Solomon H. Snyder
Department of Neuroscience

Simpson's Paradox

- The majority of molecular biology assays involve bulk sampling of heterogeneous tissue/ cells.

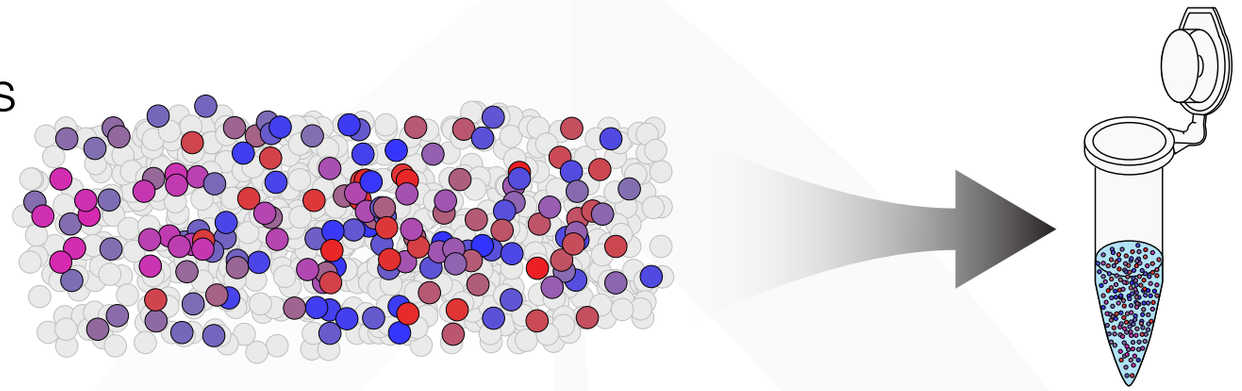
- Weighted averages can lead to reversals/ masking of meaningful relationships.

- 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***

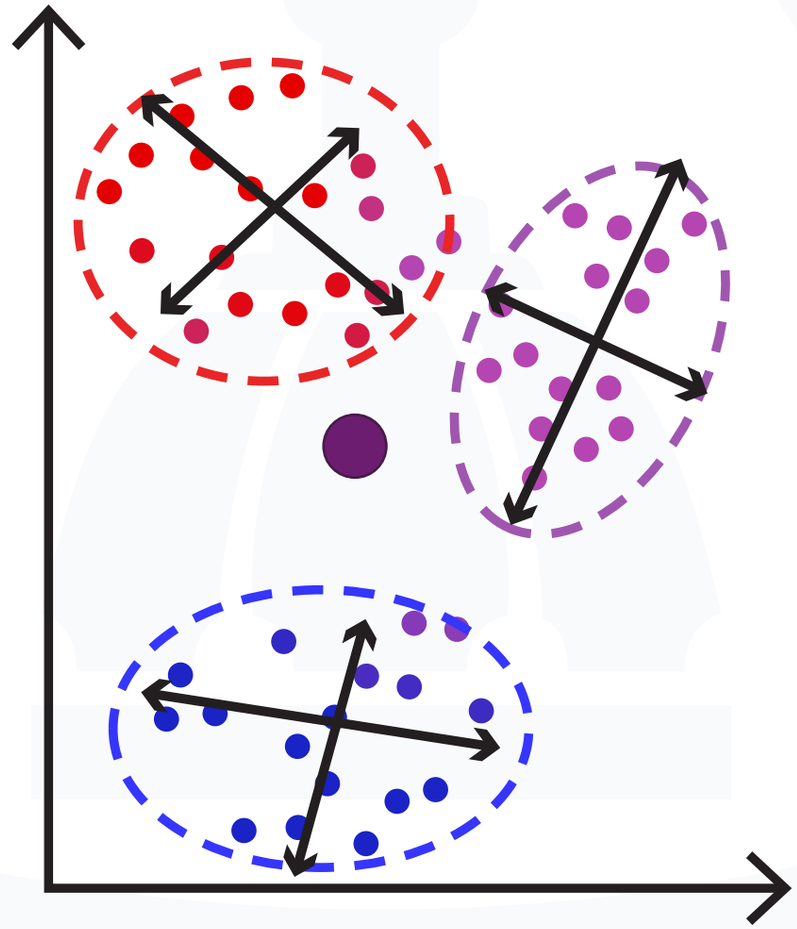


- Aggregate expression profiles of mixed-cellular samples do not accurately reflect the expression profile of any one cell type.

- 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

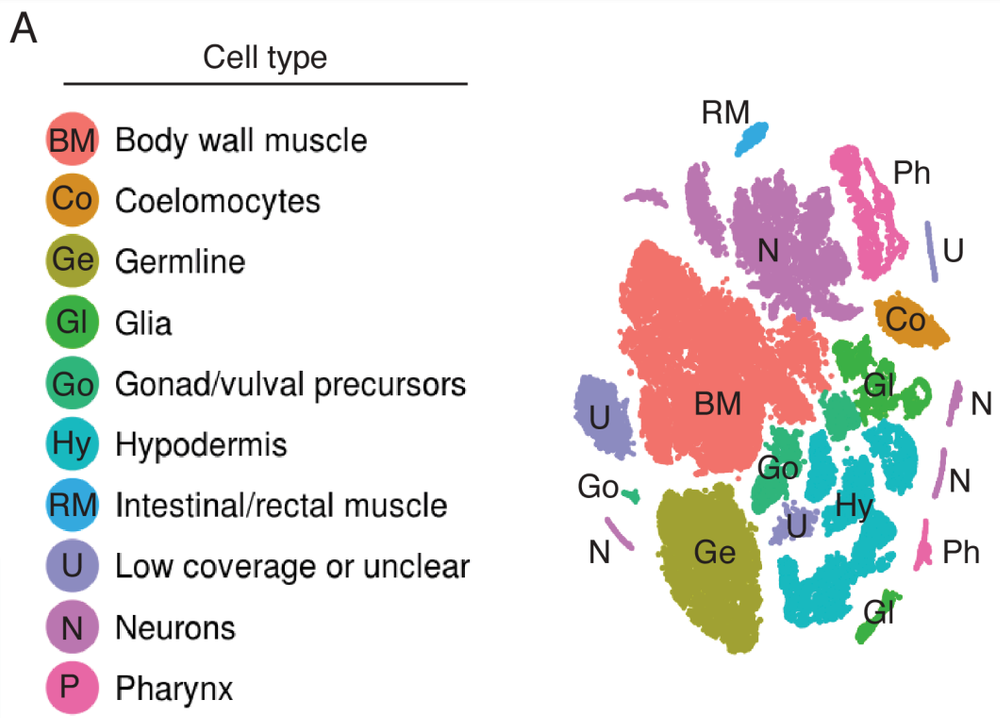
- scRNA-Seq is ***not*** just low-input RNA-Seq
  - Fundamental differences in sample handling and data processing

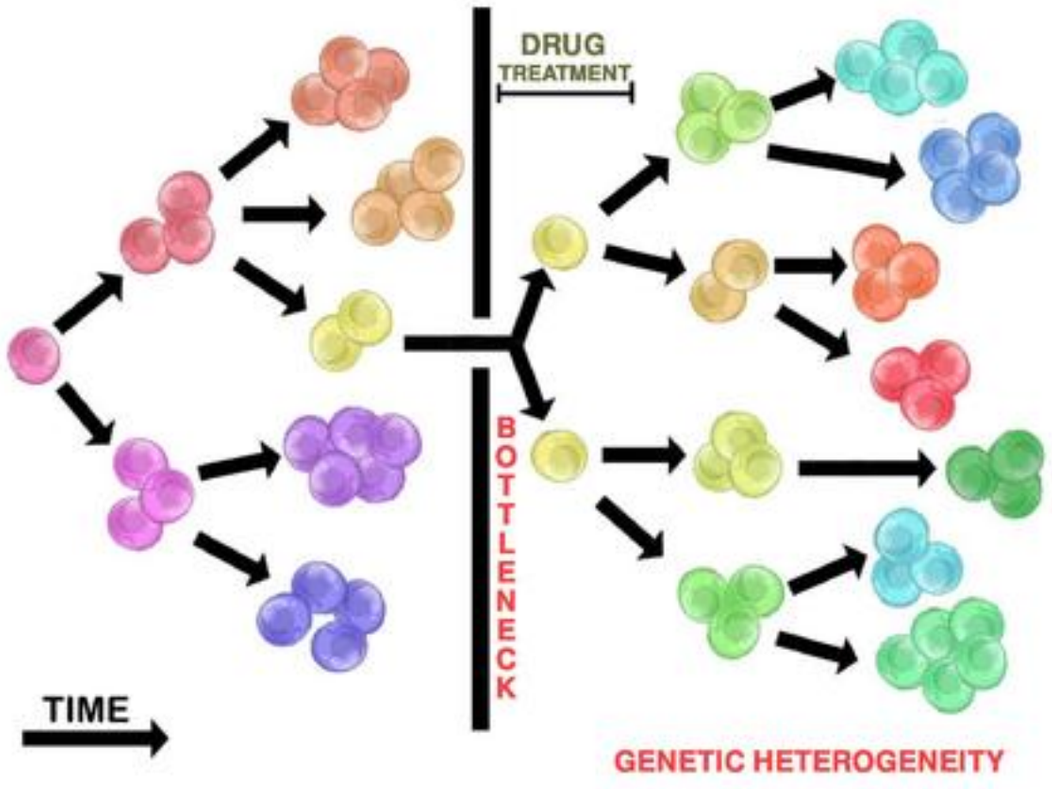
- All of the standard biases of RNA-Seq…
  - Sample handling effects
  - RNA-quality effects
  - Library batch effects
  - Sequencing lane effects

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

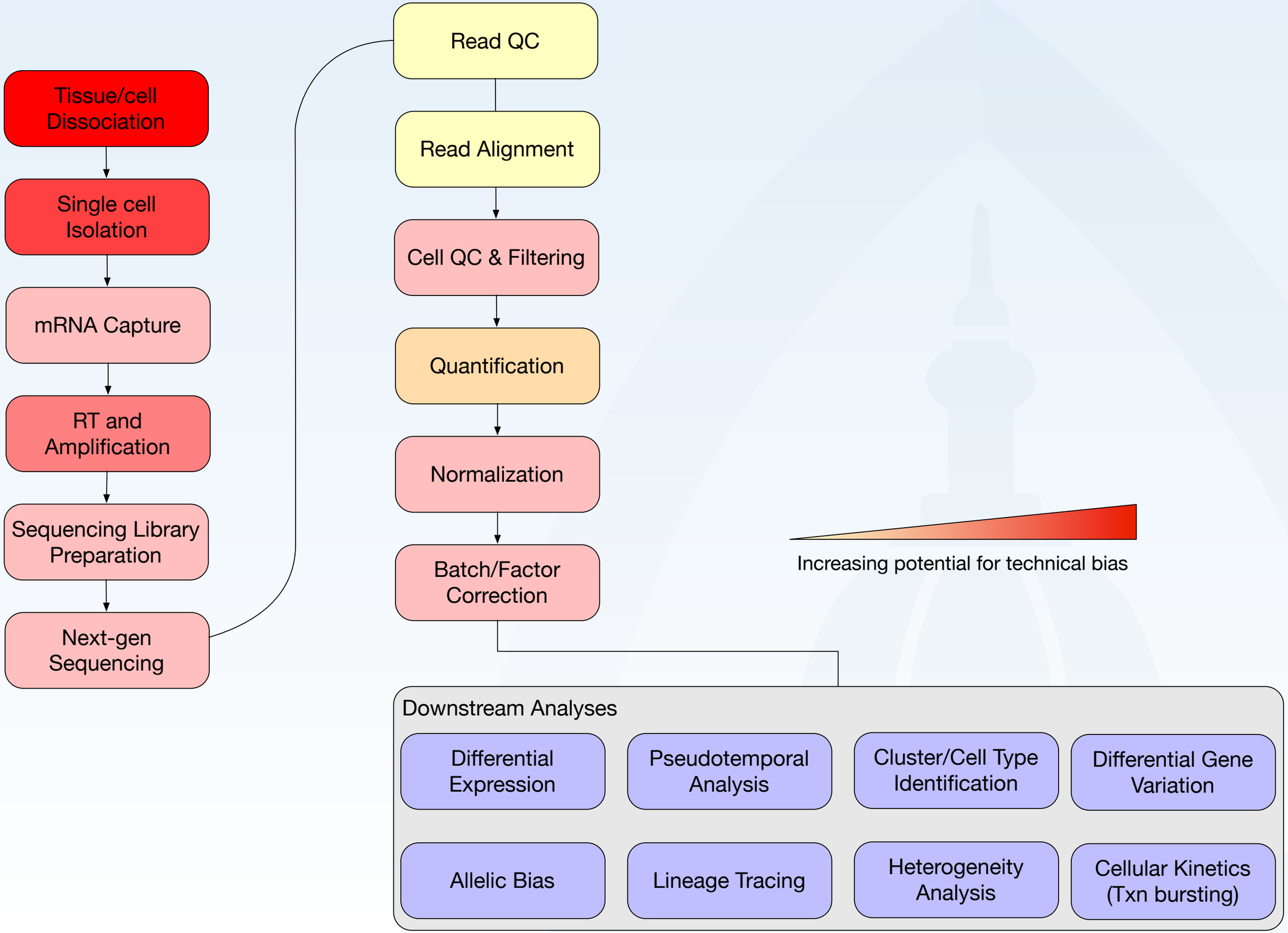
A non-comprehensive list of applications of single cell sequencing

- Cellular heterogeneity
  - Tumors
  - Mosaicism (SNPs, CNVs)
  - Epigenetic variation (DNA methylation)
- Microbiome characterization
- Identification of novel cell types/subtypes/states
  - Circulating tumor cells
  - Stable intermediate progenitors
- Lineage tracing
  - Cumulative mutation loads
  - Pseudotemporal ordering
- Tissue organization and development
- Reconstruction of continuous/ergodic biological processes
- Mechanisms of drug/treatment resistance
- Kinetics of gene regulation
  - Transcriptional ‘bursting’
- Variability of intrinsic responses to extrinsic cues
  - Dynamics/kinetics of response to morphogen gradients
  - Paracrine signaling / cell-cell communication





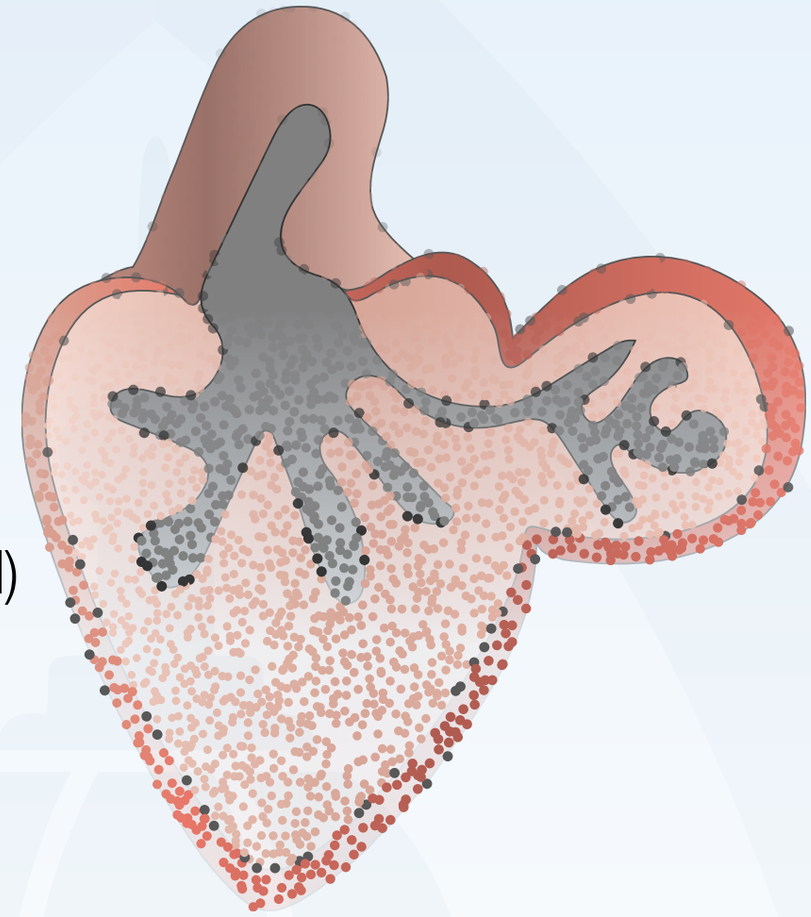
Typical scRNA-Seq experimental workflow



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:
  - Newer psychrophilic proteases enable dissociation at lower temperatures
  - Methanol fixation protocols exist to stabilize cells after dissociation
  - Nuclear isolation



Single cell isolation methods

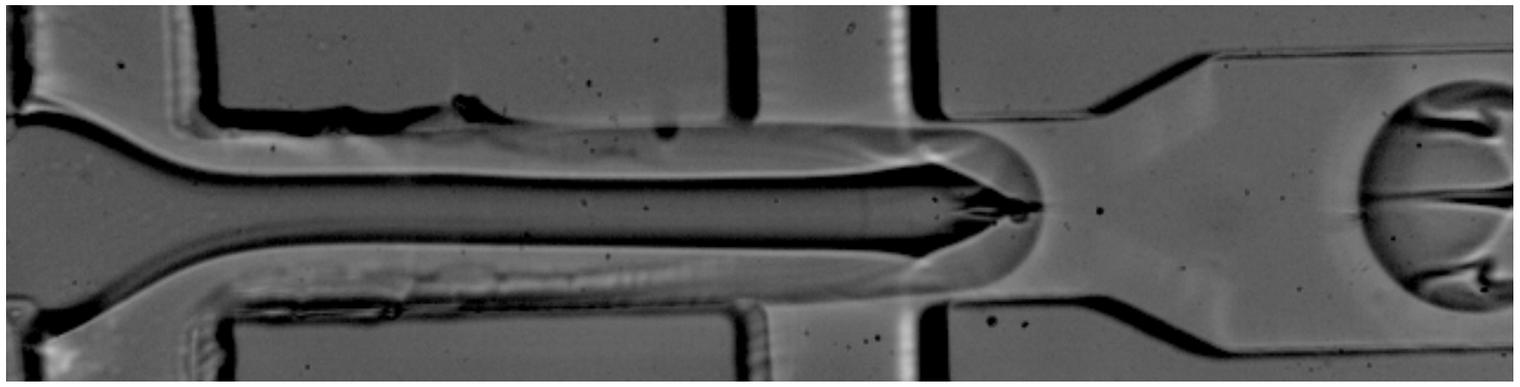
Single cell Isolation

Svensson et al Arxiv 2017

Number of single cells

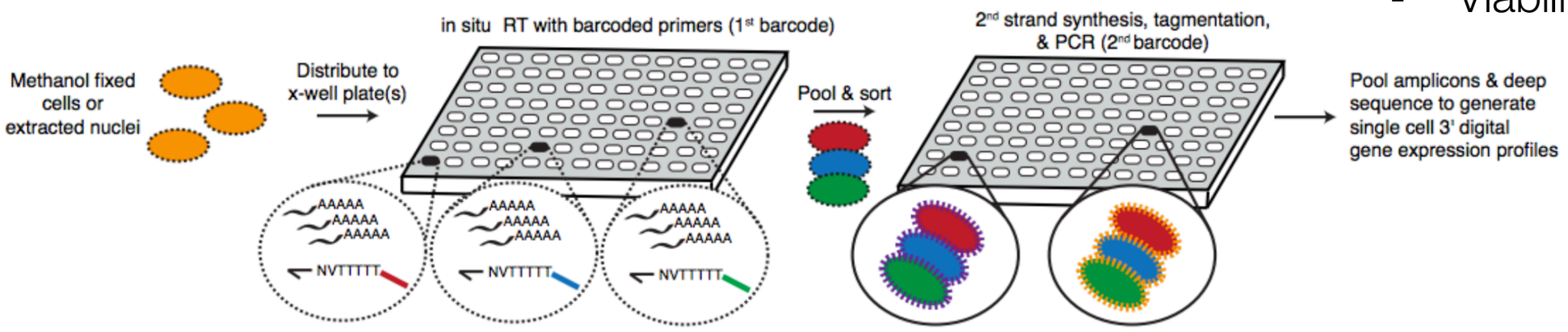
Information content per cell

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



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 †,#

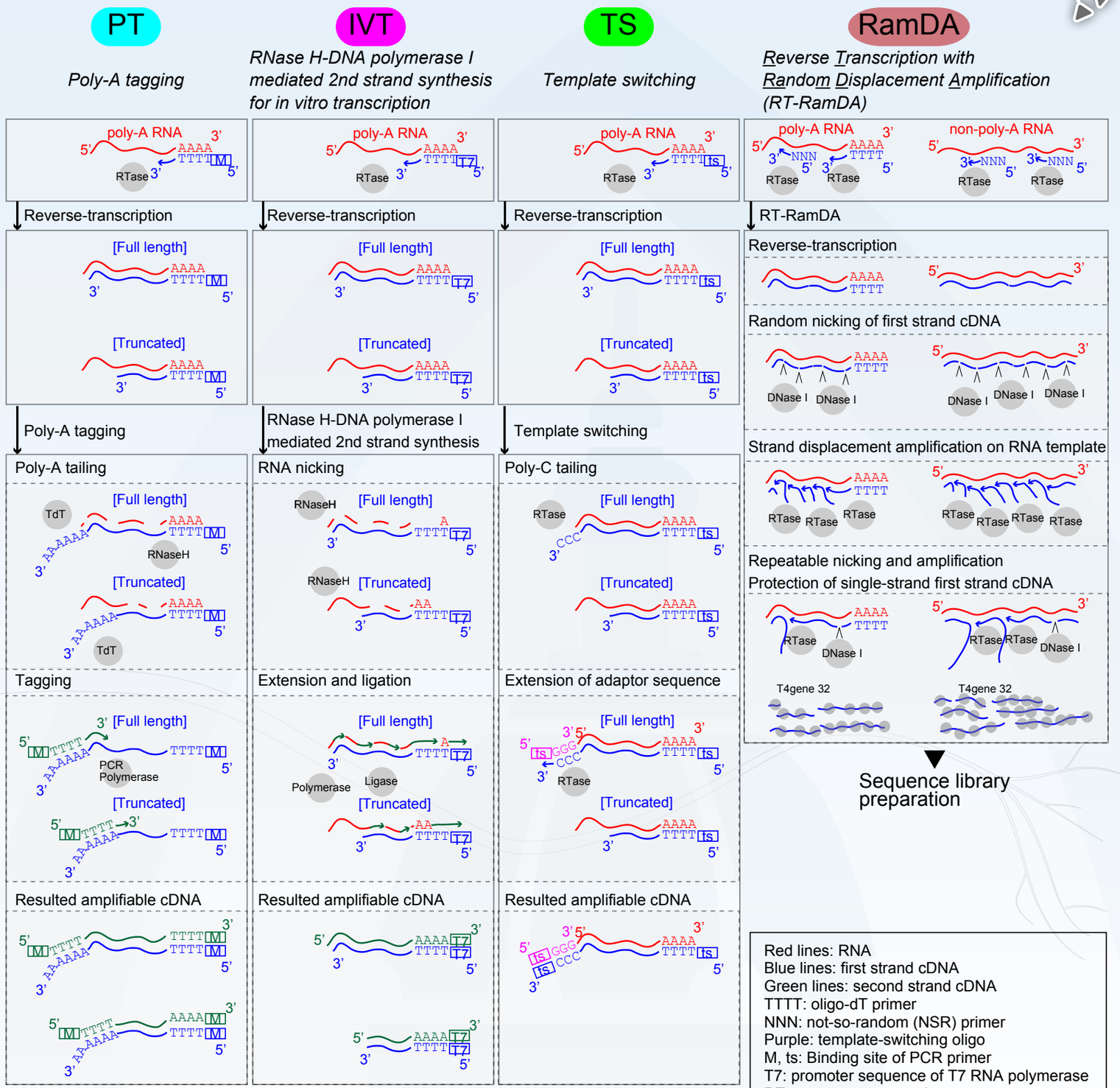
2015 [18] SUPeR-seq [19] SC3-seq [20] Beads-seq | [21] inDrop †,#,§ [22] Bose et al. †,#,§ | [23] Drop-seq †,#,§

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

- 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

Method	Reference	RT primers	cDNA synthesis	Amp method	UMIs	Transcript coverage	Sample pooling
Tang *et al.*	Tang *et al.* [11]	Poly(T) + poly(A)	Poly(A) tailing	PCR	No	Nearly full-length	No
Quartz-seq2	Sasagawa *et al.* [12]	Poly(T) + poly(A)	Poly(A) tailing	PCR	Yes	3′-end	Yes
STRT-seq	Islam *et al.* [18]	Poly(T)	Template switching	PCR	Yes	5′-end	Yes
SMART-seq2	Picelli *et al.* [13]	Poly(T)	Template switching	PCR	No	Full-length	No
SCRB-seq	Soumillon *et al.* [14]	Poly(T)	Template switching	PCR	Yes	3′-end	No
Drop-seq	Macosko *et al.* [9]	Oligo-dT	Template switching	PCR	Yes	3′-end	Yes
Seq-Well	Gierahn *et al.*[15]	Poly(T)	Template switching	PCR	Yes	3′-end	Yes
SPLiT-seq	Rosenberg, Roco *et al.* [16]	Poly(T) (*in situ*)	Template switching	PCR	Yes	3′-end	Yes
CEL-seq2	Hashimshony *et al.* [17]	Poly(T)	IVT	IVT	Yes	3′-end	Yes

Dal Molin et al. 2018

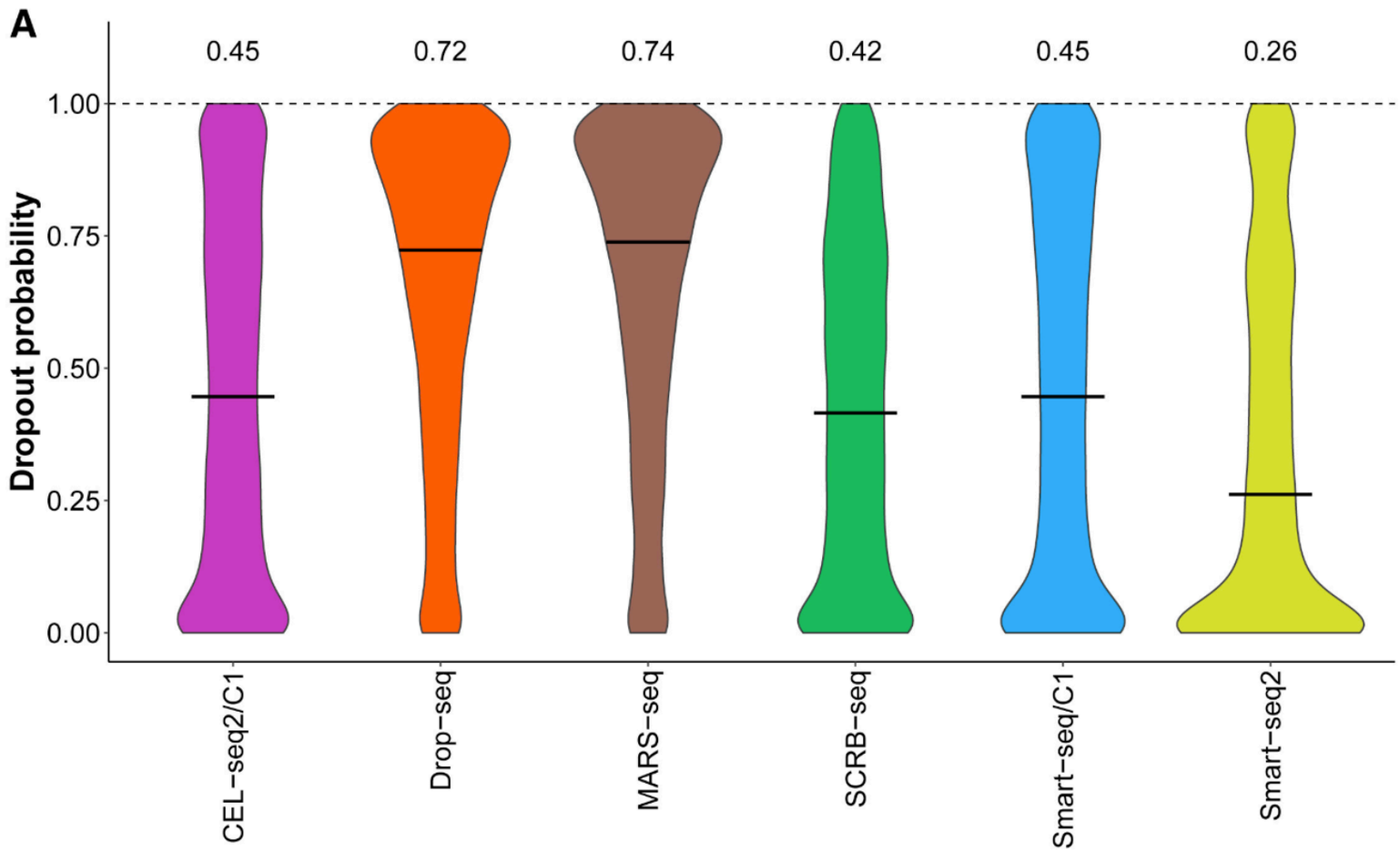
Capture efficiency - Dropout rate

- *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

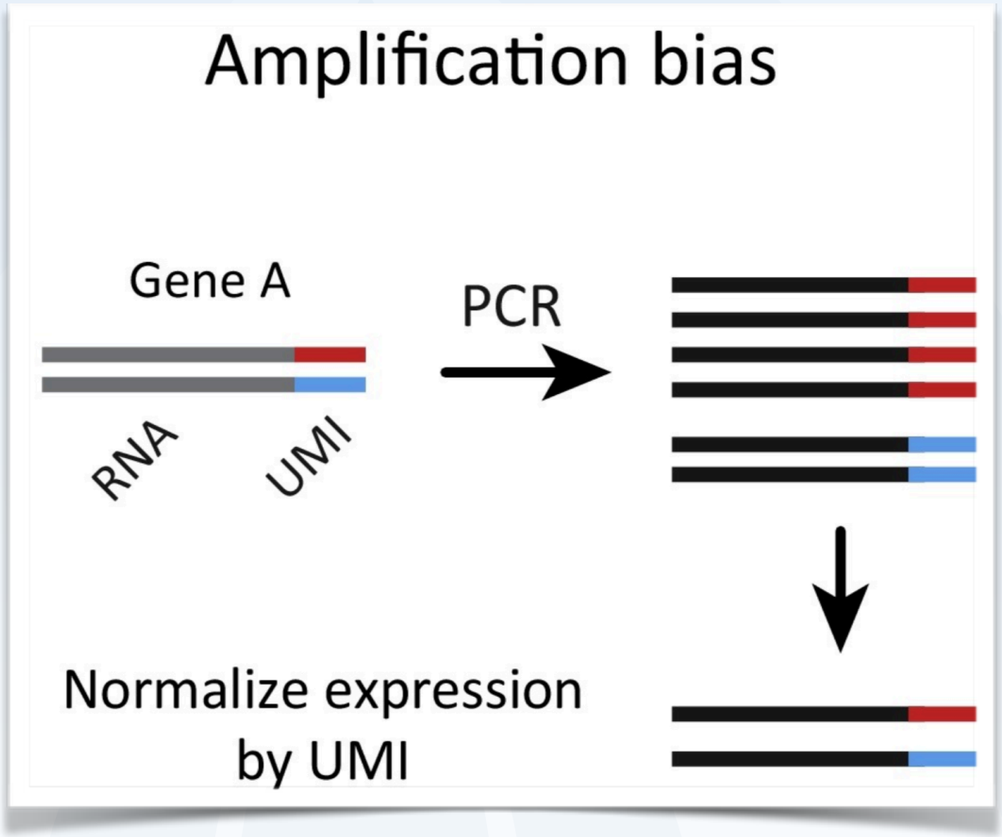
Amplification bias

- 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.

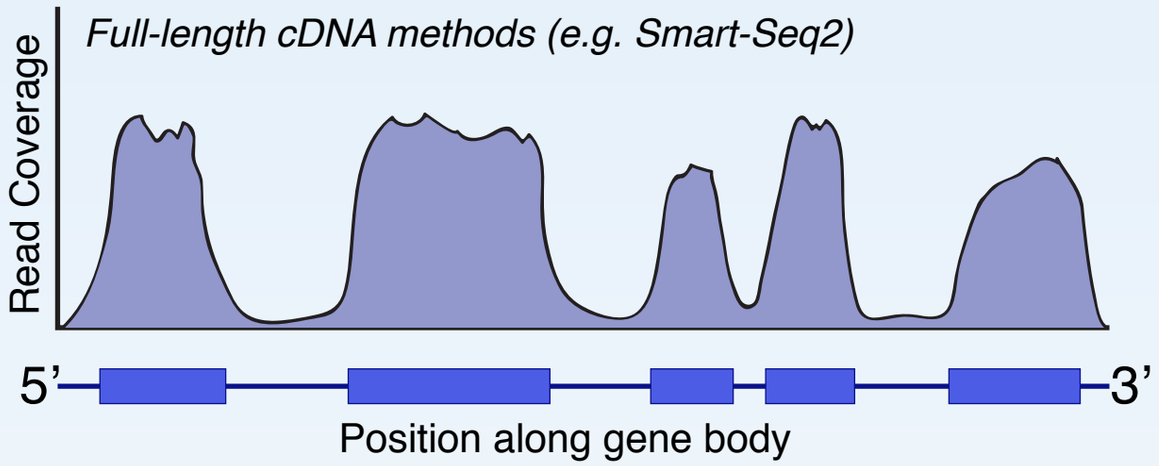
- PCR amplification (Smart-Seq2, Chromium, Drop-Seq)
  - Amplification is exponential ***and*** biased based on amplicon features
    - Excess # of cycles will 'jackpot' easily amplifiable sequences
  - ***Must*** optimize number of cycles for each experimental system.
  - Use of UMIs can minimize effect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

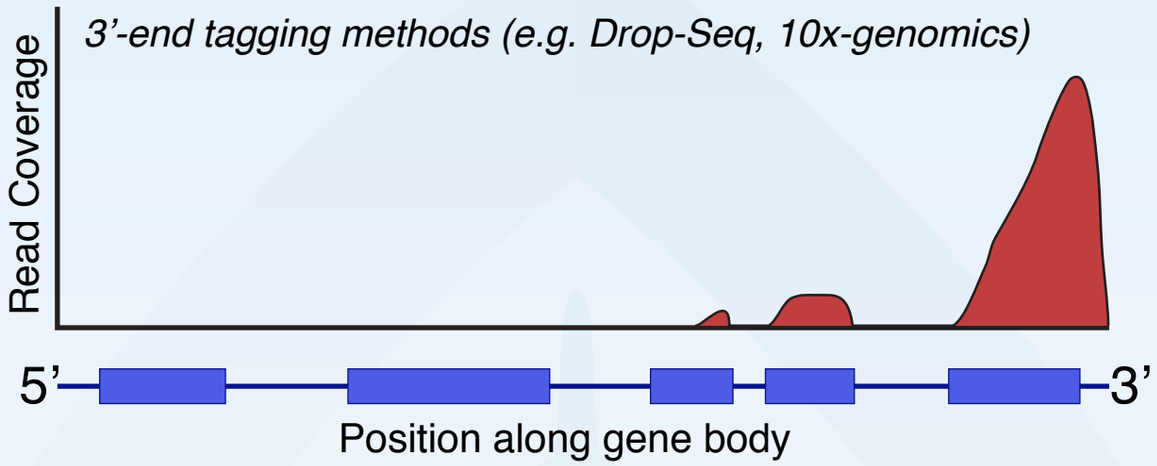
- RT-RamDA
  - Reverse Transcription with RAndoM Displacement Amplification
  - Efficient linear amplification



Two types of quantification strategies for scRNA-Seq



- 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

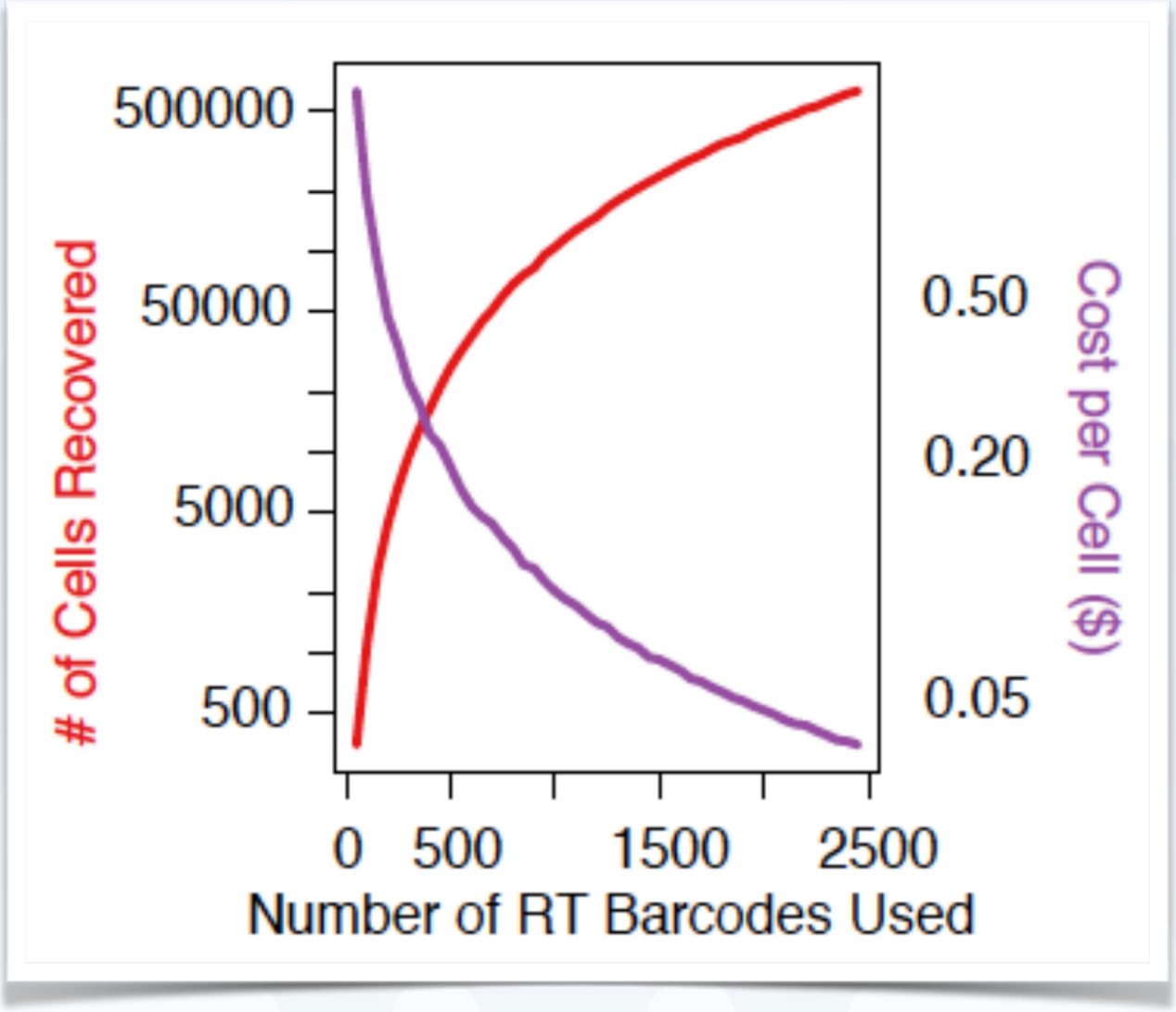
Choice of library prep method significantly affects throughput, sensitivity, and types of questions that can be asked of the final data.

Library sequencing

- Multiplexing libraries enables higher throughput and reduces sequencing batch/lane batch effects

- Recent combinatorial indexing methods can be scaled to 10^5 - 10^6 cells per library.

- Multiple barcoding schemes enables single cell & sample/library multiplexing simultaneously



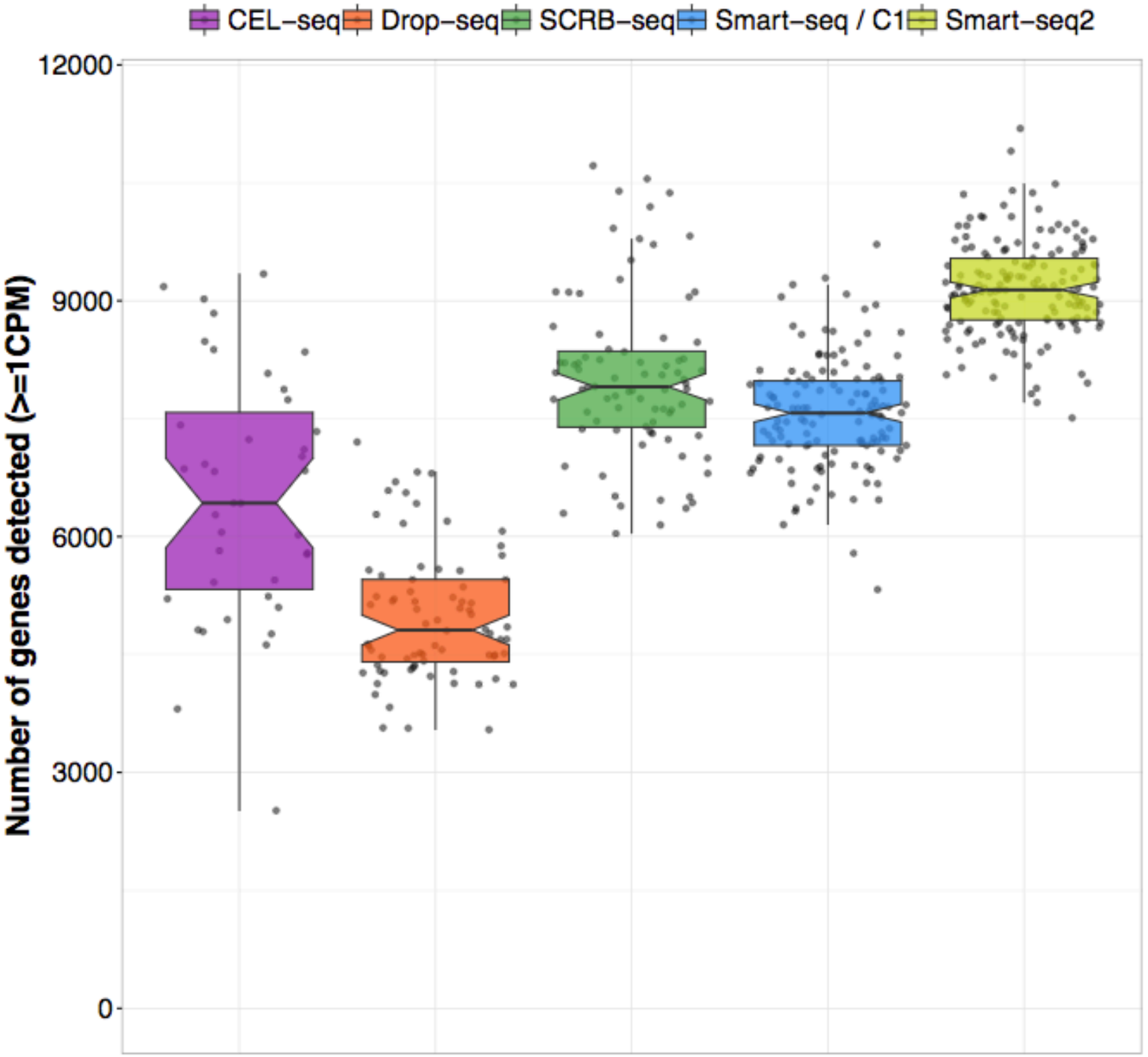
Main sources of variability in scRNA-Seq

- 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

‘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

Modeling some aspects of technical variation in scRNA-Seq data

- Variation is not consistent with the mean
  - heteroscedastic

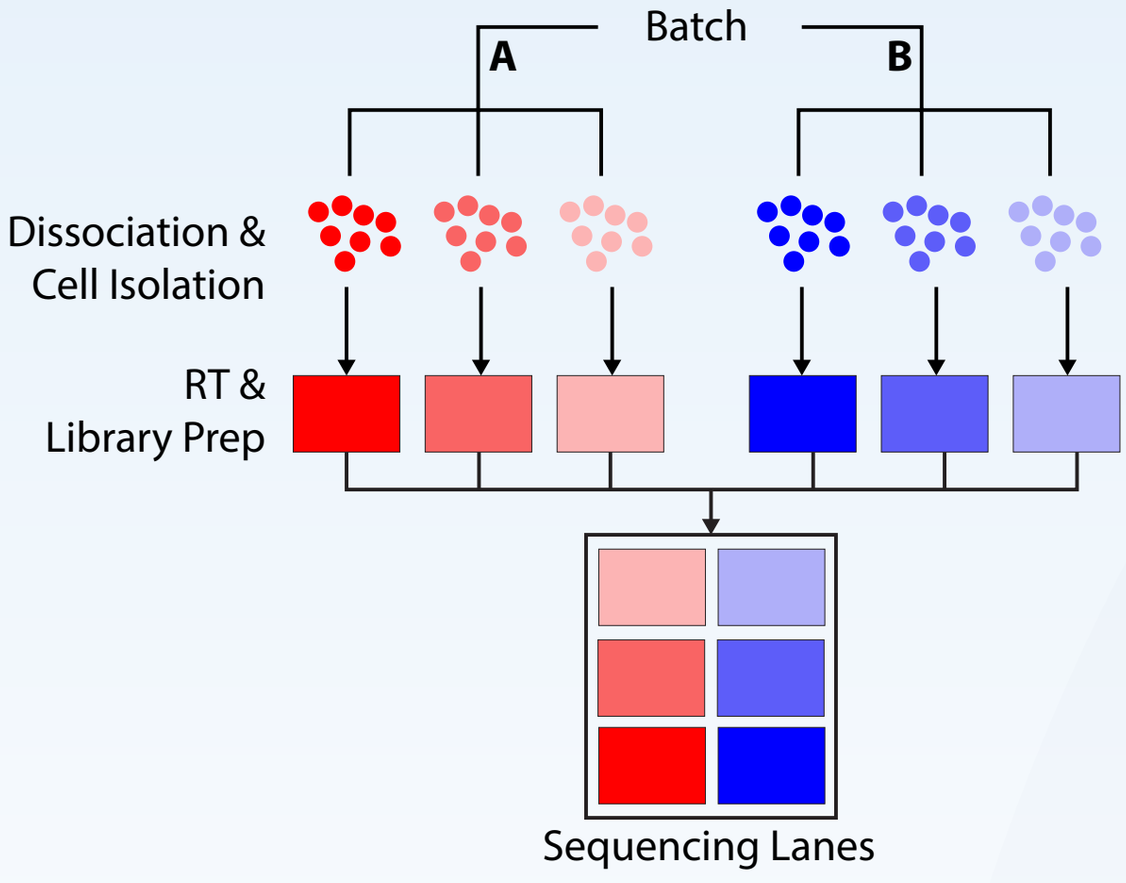
- Most analysis approaches model the mean:variance relationship
  - Assumption is that this fit represents the technical variance associated with scRNA-Seq
    - Capture efficiency

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

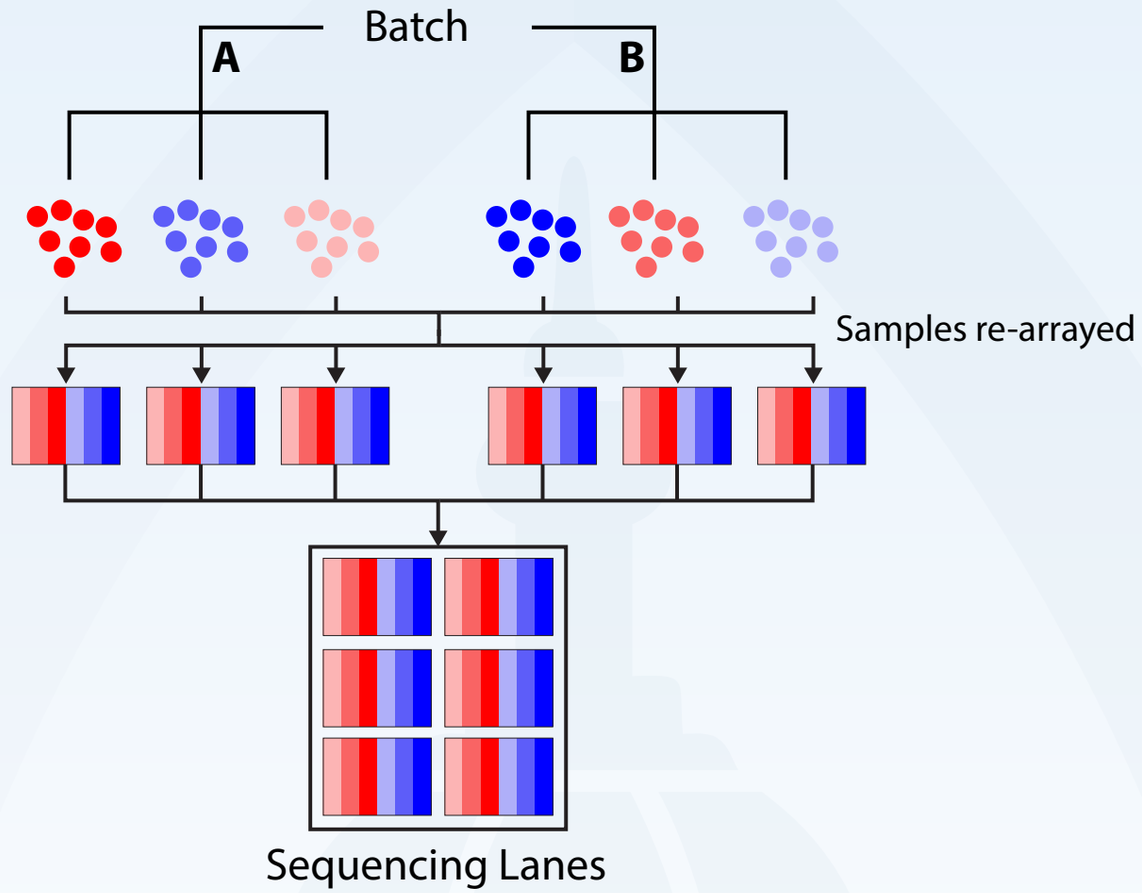
- Caution: Different conditions (cell types, treatments, batches, etc) can result in significantly different fits.

Balanced Block Design

Confounded Design



Balanced Design



- Process replicates for each condition across batches.
- Re-array cells for balanced RT, amplification, and library preparation
- Balanced plates can be planned and sorted for multi-well assays.
- For 10x or droplet-based, balance conditions across chips for encapsulation and randomize groupings for downstream processing.
- ***Failure to design experiments in a balanced manner will result in artefacts!***

Popular Frameworks for scRNA-Seq analysis

- R
  - Monocle - http://cole-trapnell-lab.github.io/monocle-release/
  - 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

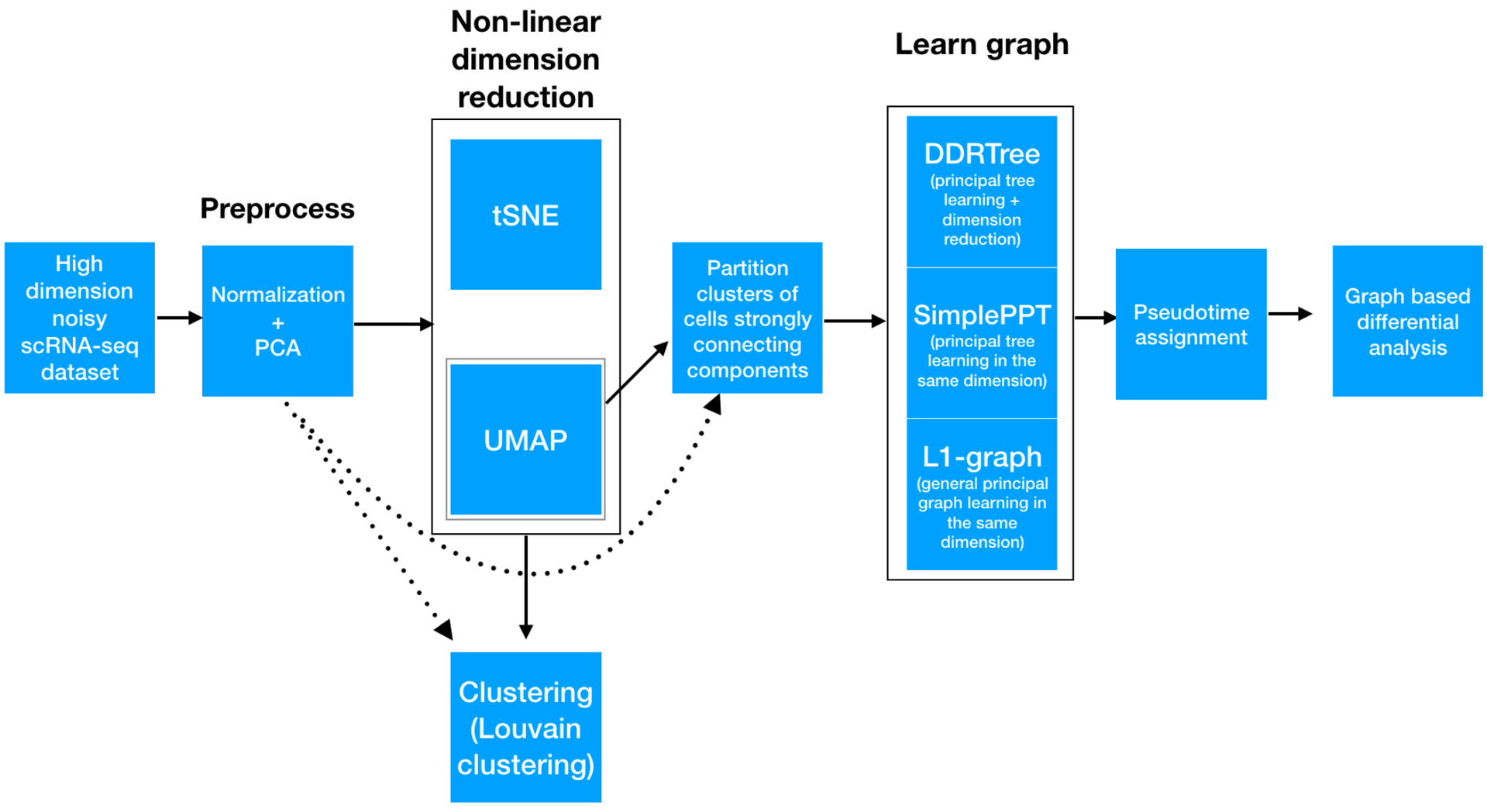
The Dataset

- Dissociated P6-7 motor cortex
- n=5 replicates each of wild type and a human *C9orf72* pathogenic repeat BAC transgenic line (C9-500-BAC)
- 10x Genomics 3’ Gene Expression analysis
  - ~8000 cells targeted per replicate
- 10 samples sequenced on 1 NovaSeq 6000 S1 Flowcell
  - ~1.6B reads
- Standard demuxing and preprocessing with 10x genomics cellranger (v2.2)

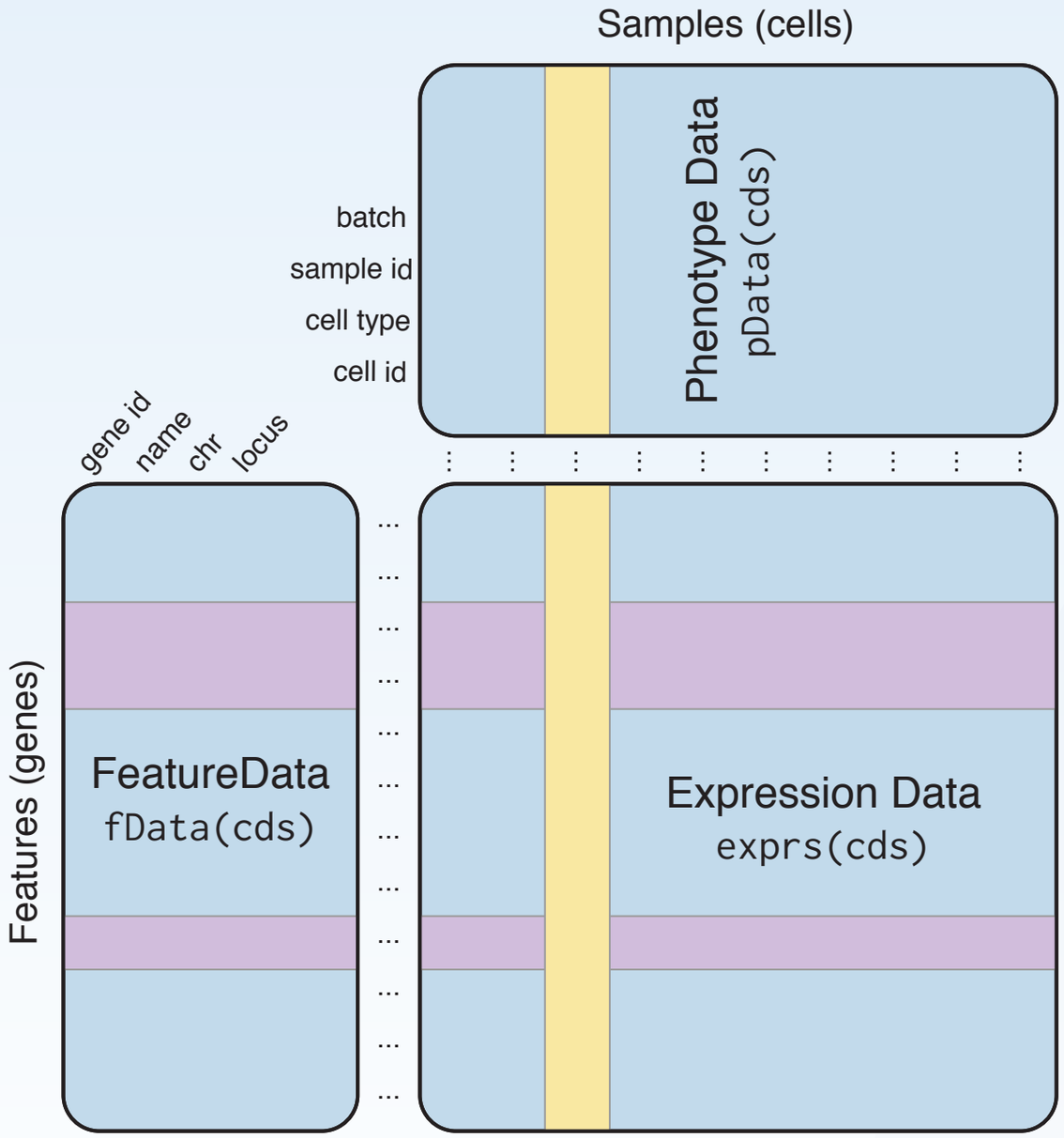
- Objectives for today:
  - Import/clean preprocessed scRNA-Seq data into Monocle Framework (R/ Bioconductor)
  - QC analysis of cell and gene quality
  - Dimensionality reduction
  - Clustering
  - Cell Type Annotation
  - Intro to Differential Expression

Monocle Framework & Workflow

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

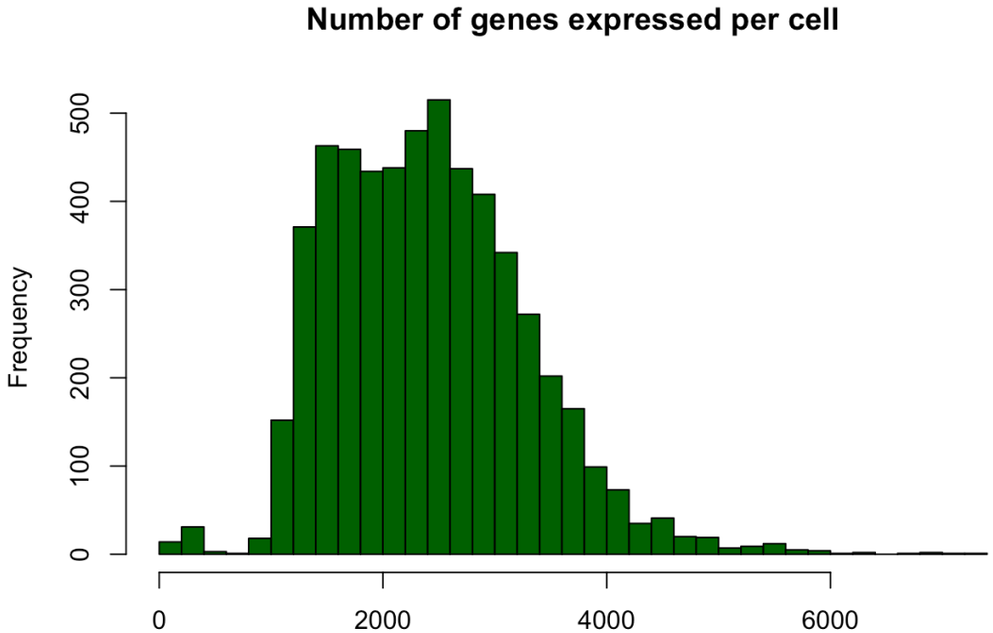


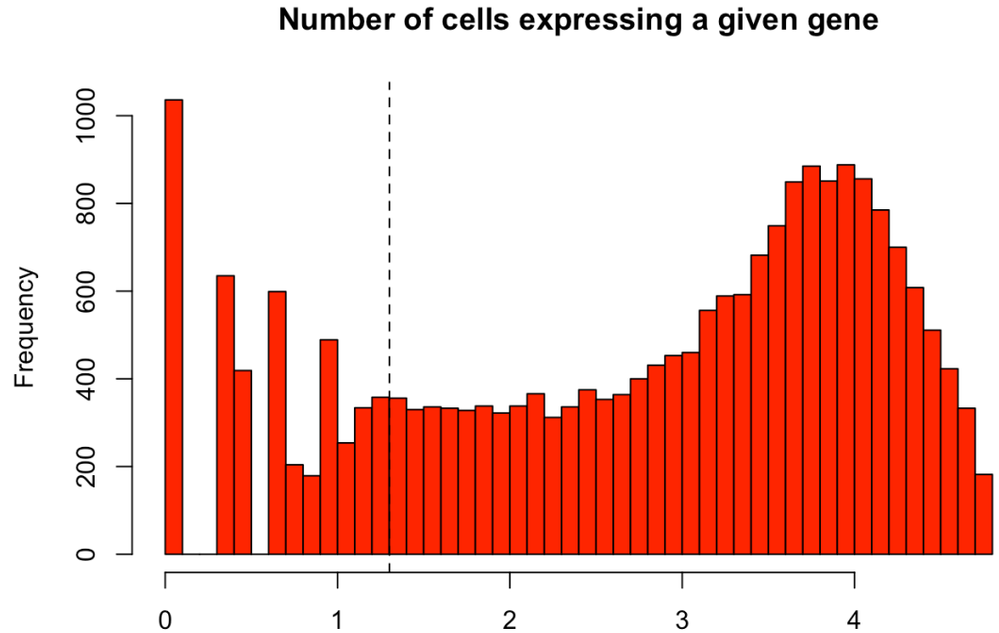
The CellDataSet class

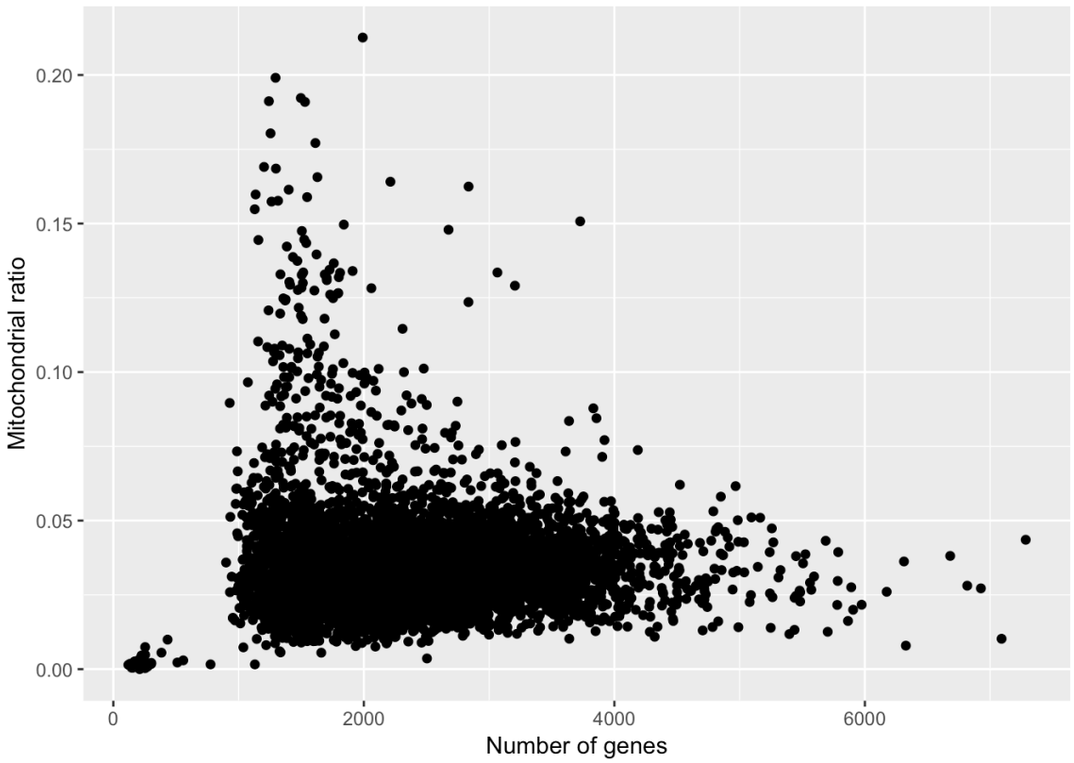


- Most frameworks define similar object structures to hold gene expression data.
- Gene features and sample (phenotype) information are stored and indexable with expression matrix.

Quality Control Metrics







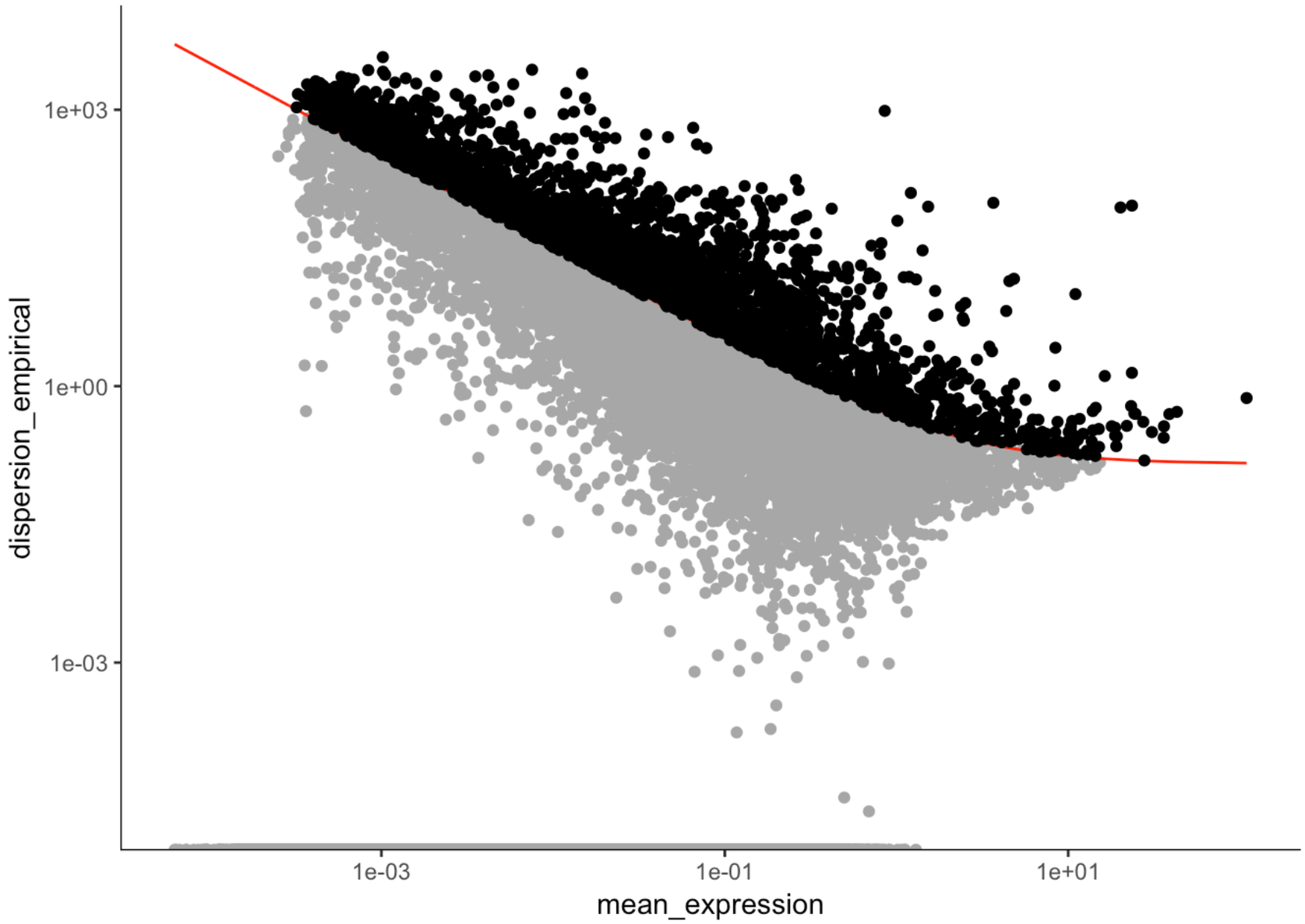
- Low quality cells:
  - Total mRNA count per cell
  - Number of expressed genes per cell
  - % of counts derived from mitochondrial genes
- Gene QC:
  - Mean expression level
  - Minimum # of cells with detectable expression

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.

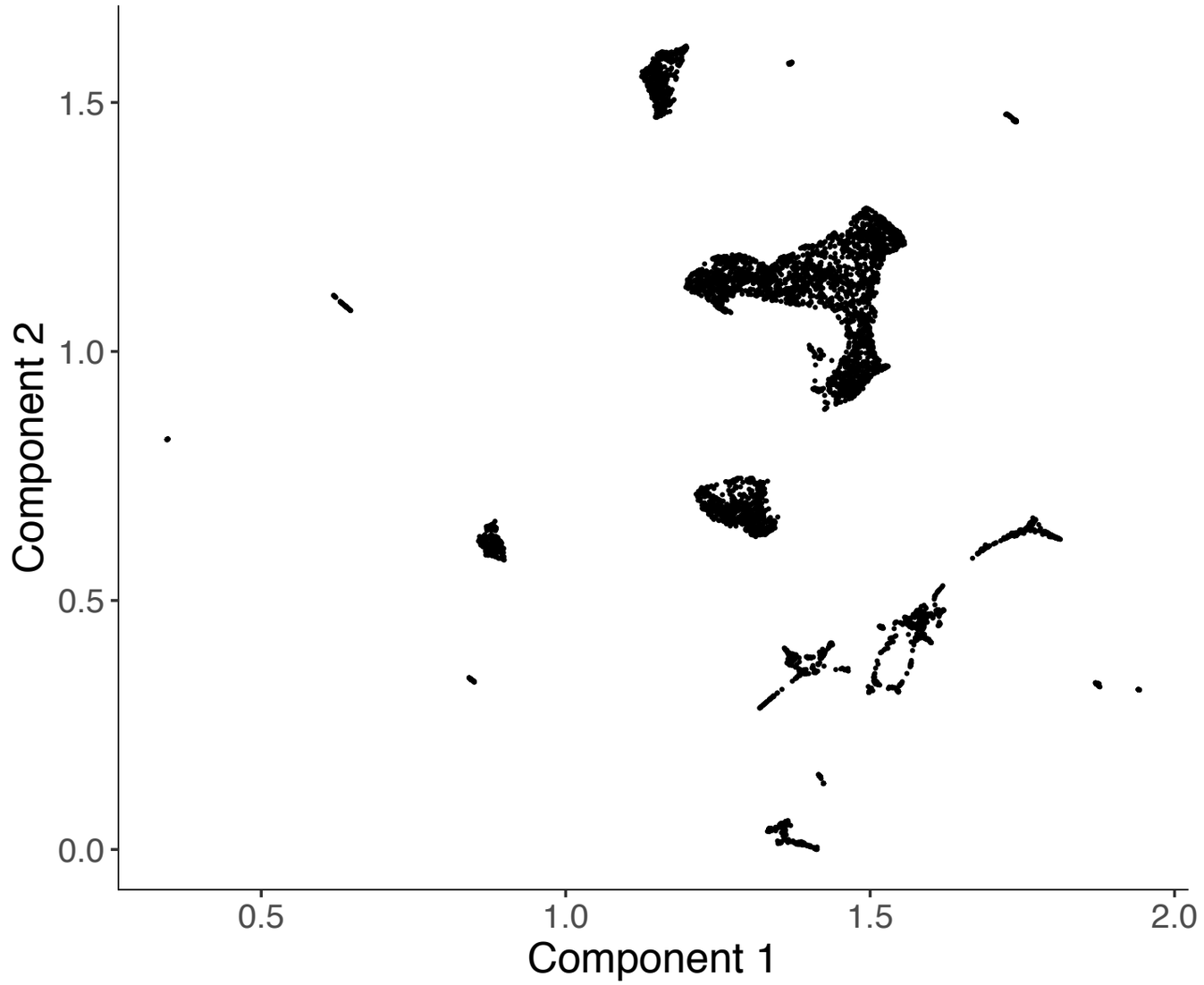
- We select this subset to highlight differences between cell types and states for downstream analyses



Dimensionality Reduction

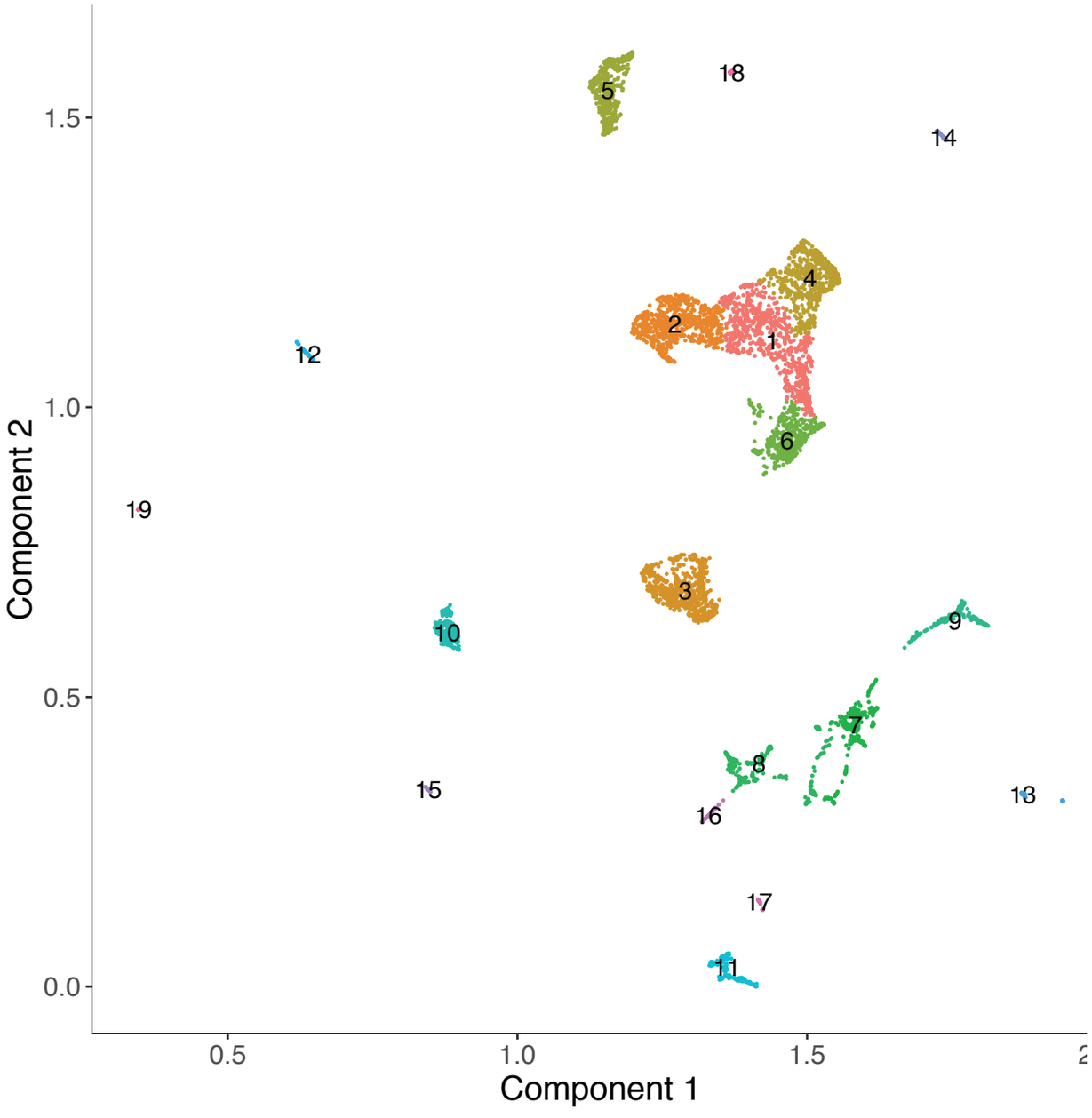
- Summarizes *n* principal components into a visually interpretable 2-dimensions
  - Could also be 3

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

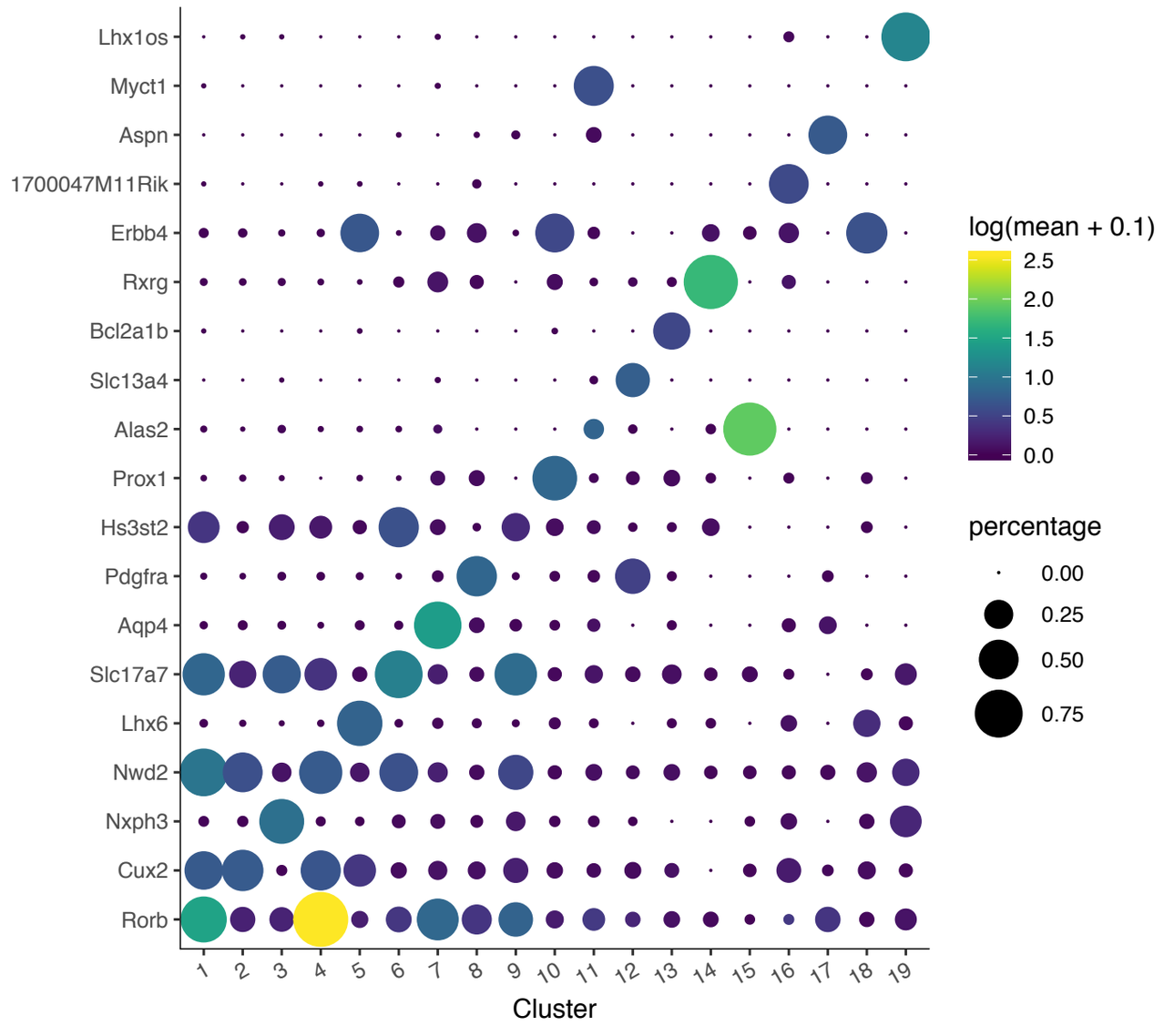


Clustering into subtypes

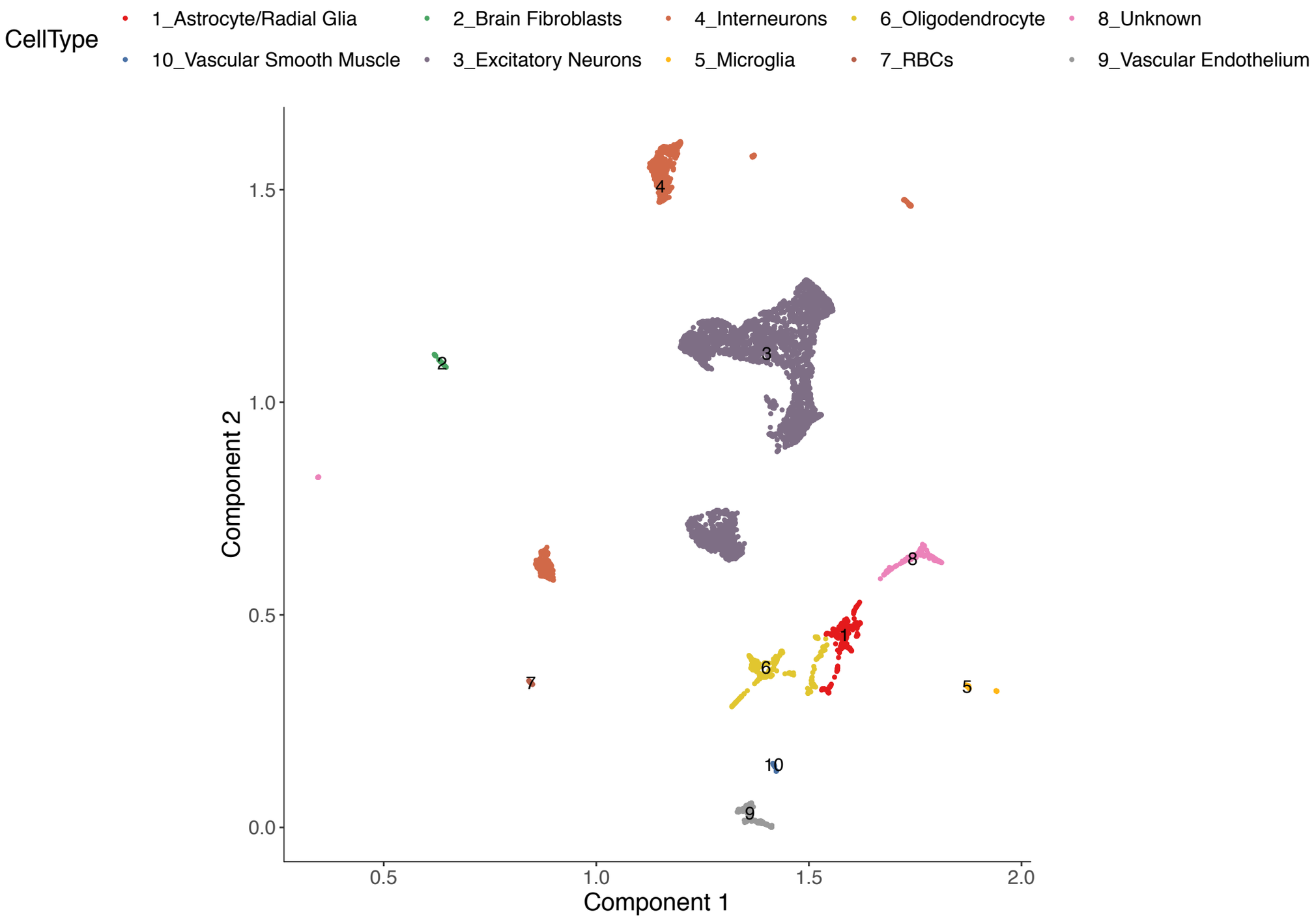
Louvain clustering



Top cluster-specific genes



Cell type annotation



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)

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

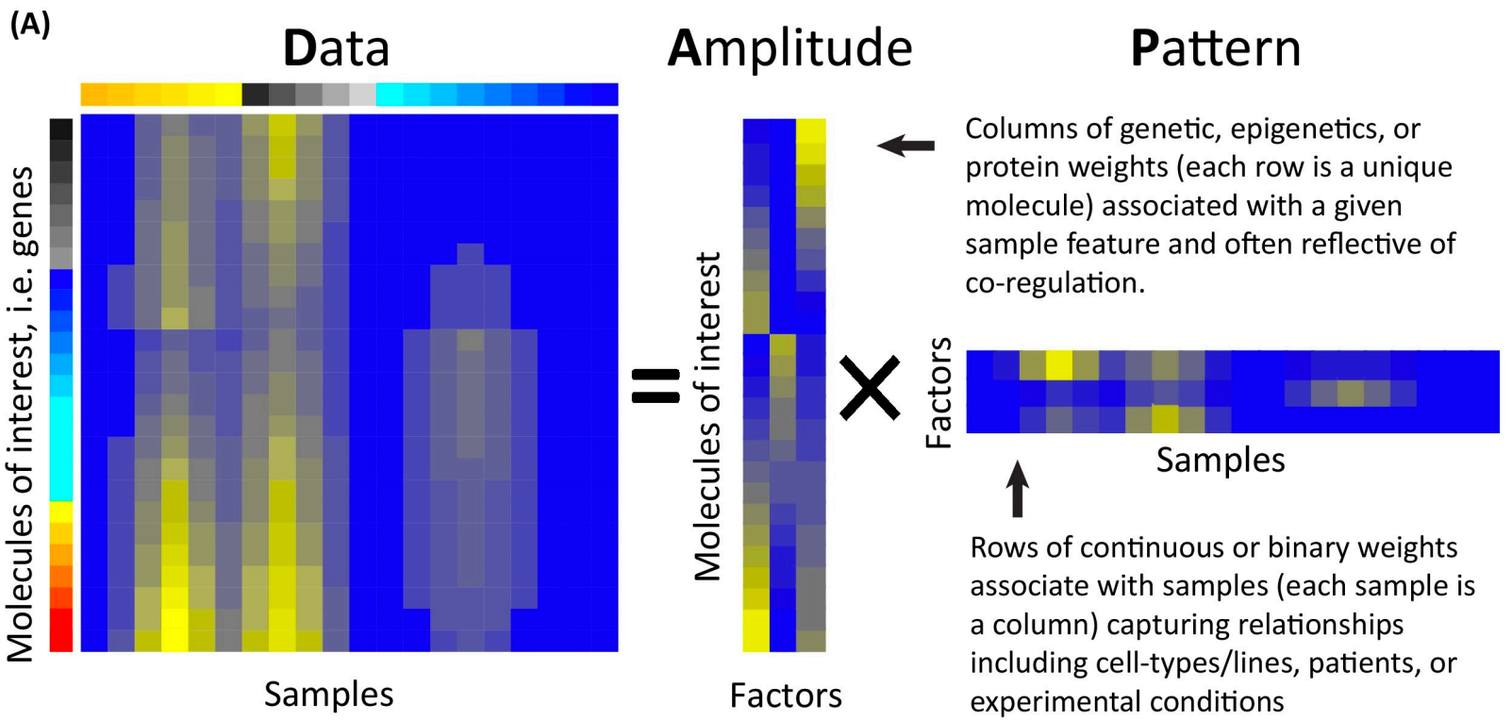
Reduced model (m2)

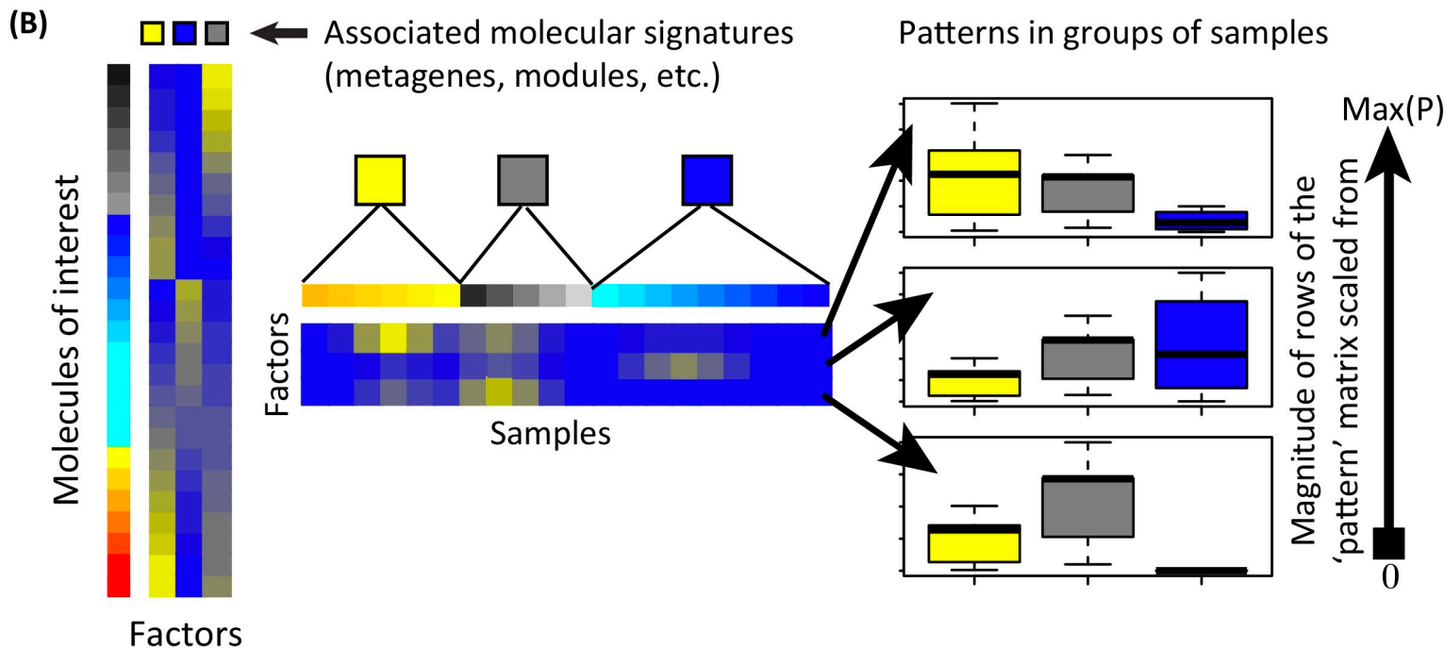
$$y_{gbt} \sim nGenes + Batch_b$$

Likelihood Ratio Test

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

Pattern Discovery - NMF

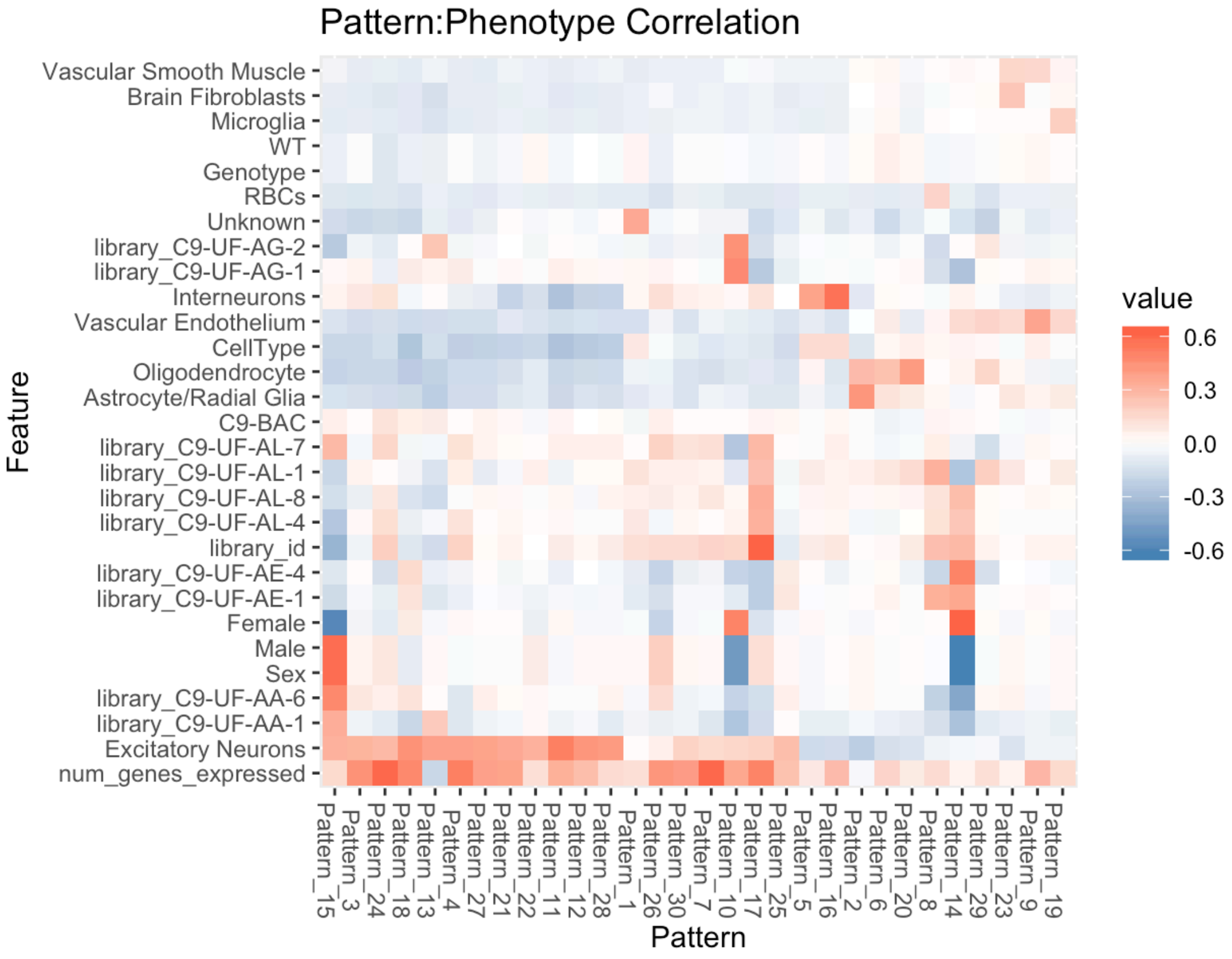




Trends in Genetics

- Matrix decomposition methods (including nonnegative matrix factorization) can be used to identify 'patterns' of co-regulated gene expression
- Patterns may correspond to:
  - Cell type identities
  - Biological processes
  - Spatial gradients
  - Other cellular features

Patterns to annotate cellular features



Additional questions

- How many distinct types of interneurons can you identify?

- Can you identify genes with differential expression between genotypes? How might you test for cell-type specific differential expression?

Additional resources

- Interactive dataset browser:
  - http://neurocoglab.gofflab.org
- Code/scripts to reproduce todays lab:
  - https://github.com/gofflab/neurocog_scRNA_seq
- Data link:
  - https://drive.google.com/open?id=1sSoBnsXhOrLTC-EsTdSk18h4IEiZHQ8D
- 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/
- Developing mouse atlas:
  - https://oncoscape.v3.sttrcancer.org/atlas.gs.washington.edu.mouse.rna/landing