

6.874, 6.802, 20.390, 20.490, HST.506
Computational Systems Biology
Deep Learning in the Life Sciences

Lecture 15: Deep Learning for single-cell genomics

Prof. Manolis Kellis

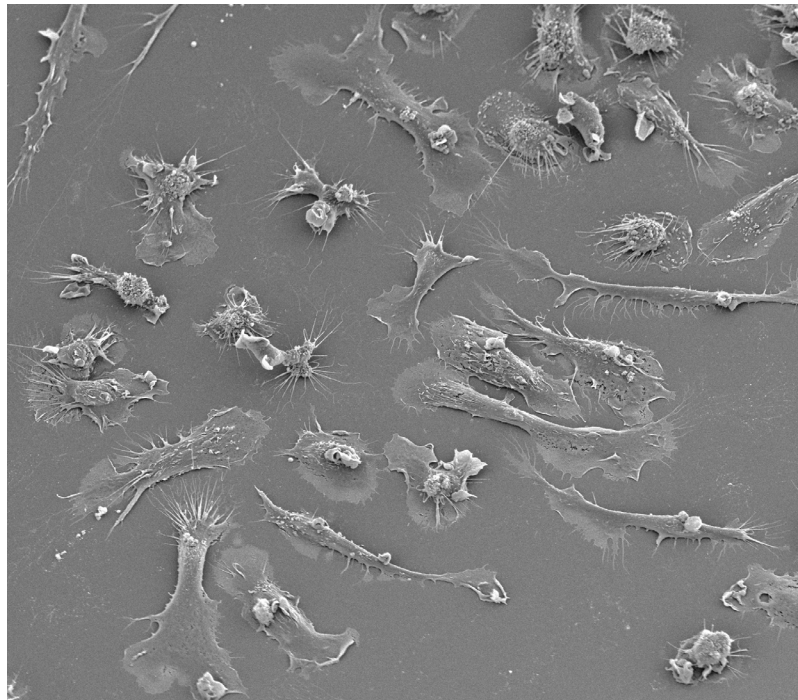
Today: Deep Learning for Single-cell Genomics

1. Why single cells, traditional approaches, scRNA-seq
2. Scaling up single-cell technologies: evolution of scRNA-seq
3. Dealing with noise, doublets, and other scRNA-seq issues
4. Computational challenges in scRNA-seq data analysis
5. Deep learning methods for scRNA-seq data analysis
6. Beyond scRNA-seq: scATAC-seq, multi-omics

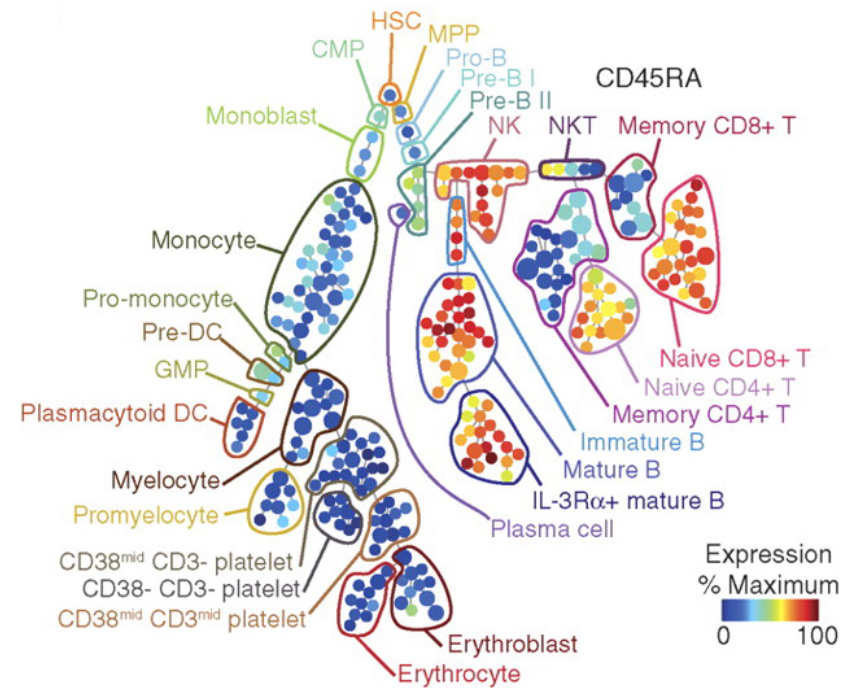
1. Foundations: Why single-cell profiling

Why single cells

Cellular heterogeneity

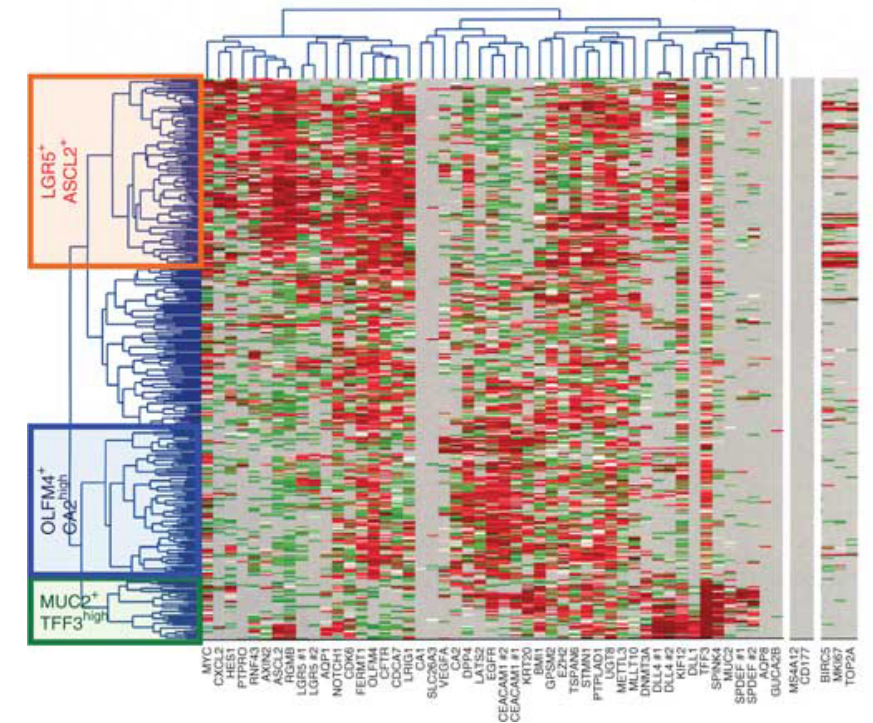


Differentiation trajectories



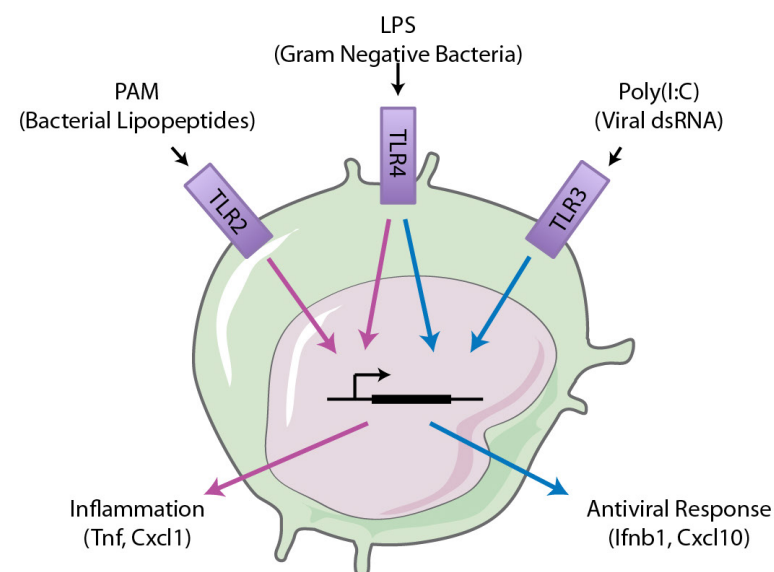
Bendall et al. (2011), Science

Within-cell-type differences

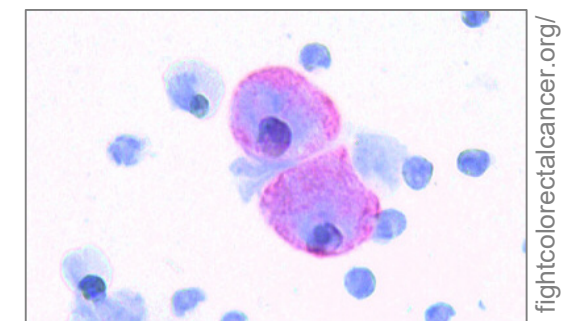
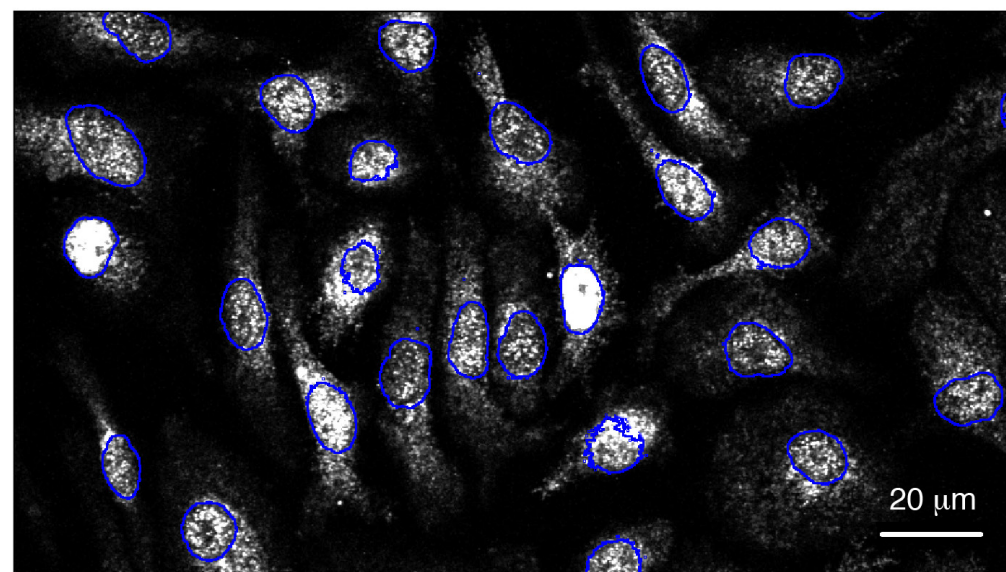


Dalerba et al. (2011), Nature Biotech

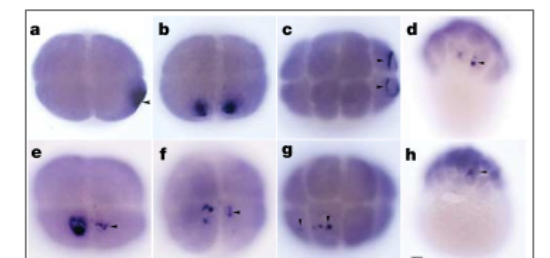
TLR Signaling



IRF3 Protein Levels - 4h LPS



Circulating Tumor Cells

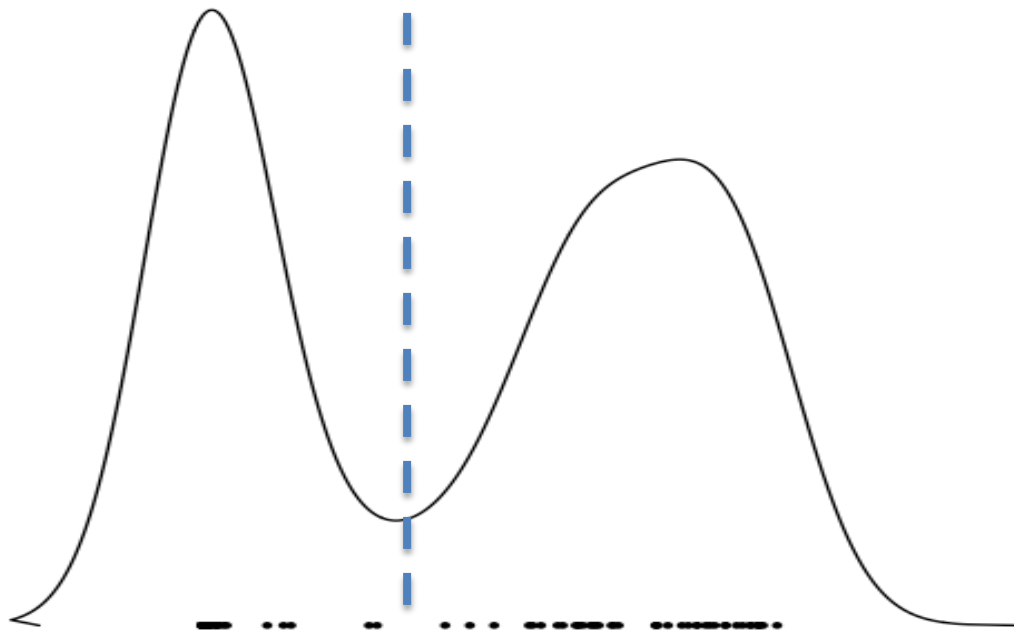


Zebrafish early embryo

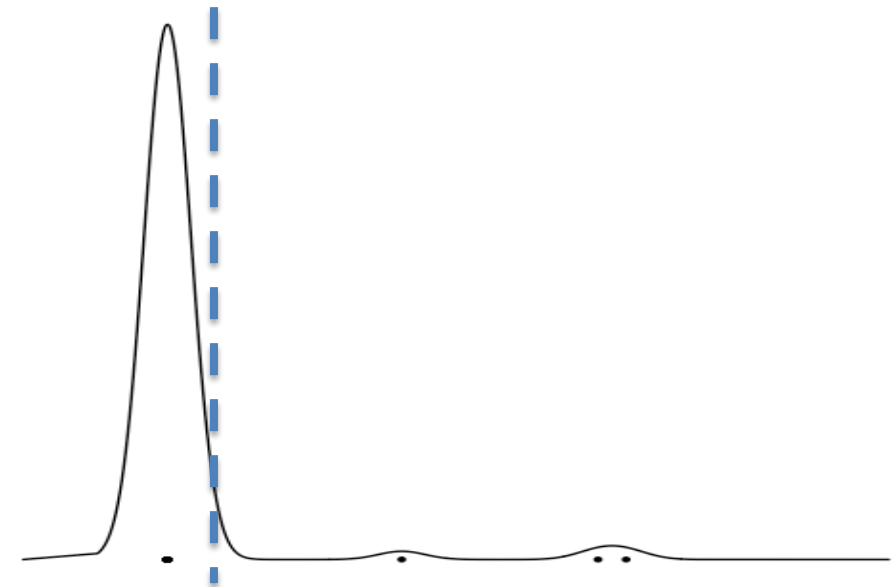
Cellular responses can vary substantially between “identical” cells.

Overcome low input

Whole-sample analysis can lead to misleading views



The average may not represent the population



Rare events can be lost ...

Traditional technologies for single-cell analysis

In living cells

Method	Species?	Endogenous?	Real-Time?	# probes?	Amplification	Advantages	Disadvantages
MS2 or Spinach	RNA	No	Yes	<5	No	Spatial Info	Need long UTRs
Molecular Beacons or SmartFlares	RNA	No	Yes	<5	No	Spatial Info Lots of cells	Requires microinjection

Pros: we can watch **dynamics** and get spatial information!

Cons: we can only look at a **few** different **known** things at once!

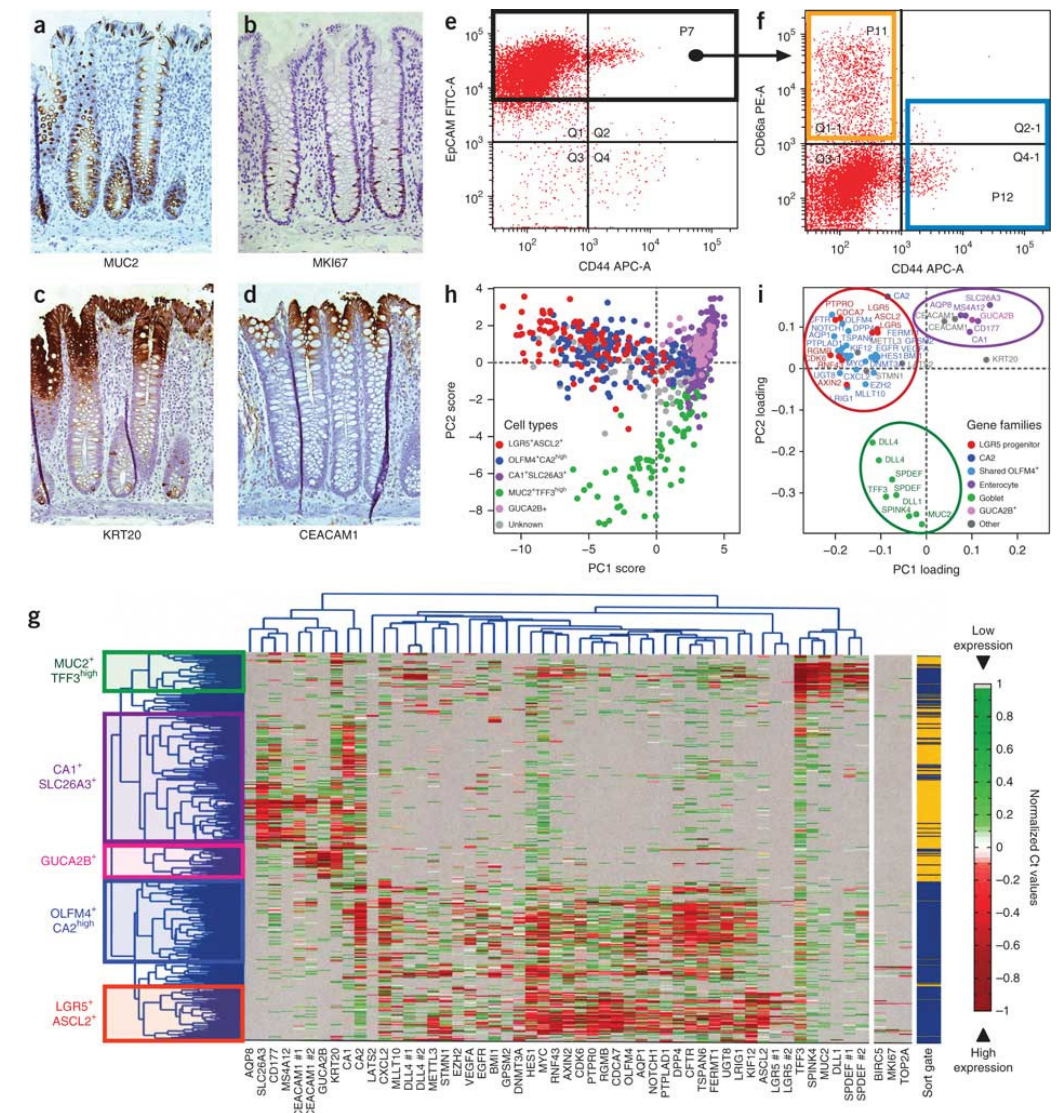
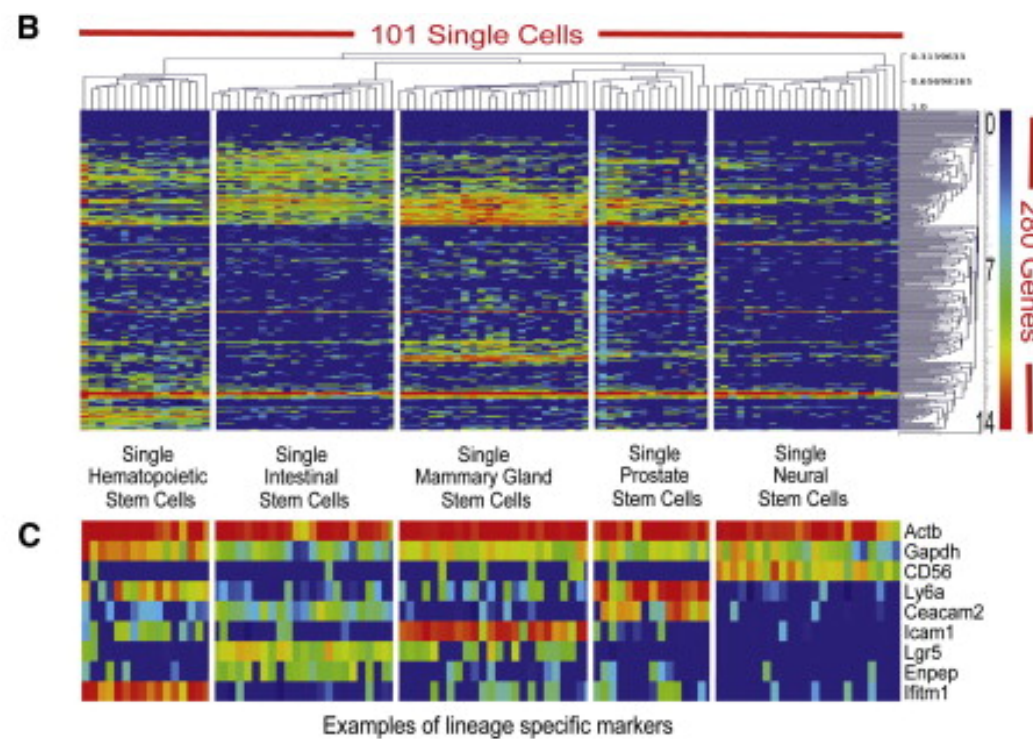
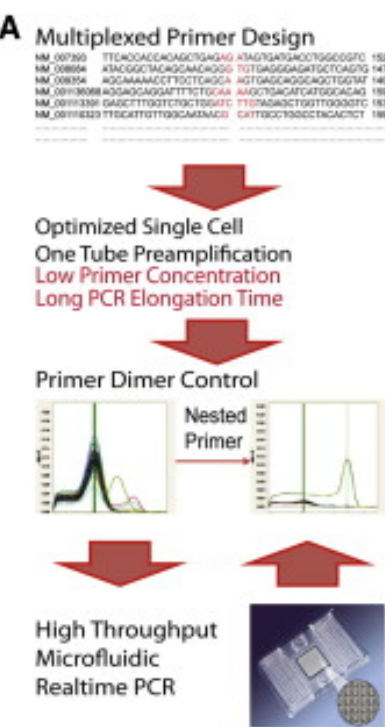
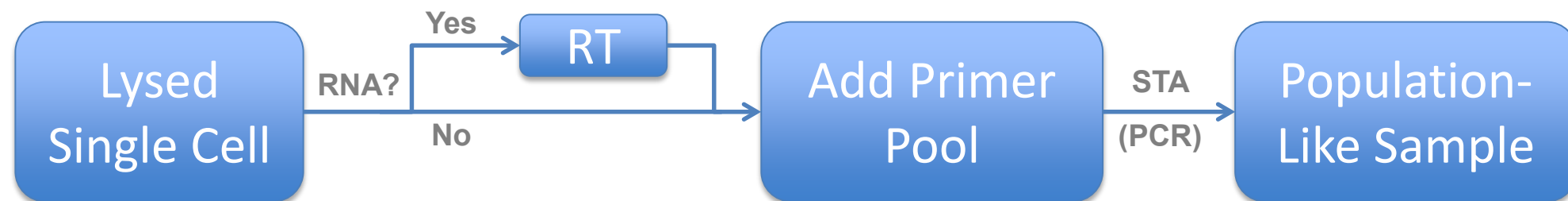
In dead cells

Method	Species?	Endogenous?	Real-Time?	# probes?	Amplification	Advantages	Disadvantages
FISH	DNA, RNA	Yes	No	<5	No	Spatial Info Lots of cells	Expensive
In-Situ Sequencing	DNA, RNA	Yes	No	Many	Yes	Spatial Info Lots of cells	Bias Slow
Single-cell (RT)-PCR	DNA, RNA	Yes	No	<500	Yes	Simple	Need to know targets Bias
Single-cell Sequencing	DNA, RNA	Yes	No	Many	Yes	Genome wide	Bias Expensive

Pros: we can increase our multiplexing dramatically!

Cons: no dynamics!

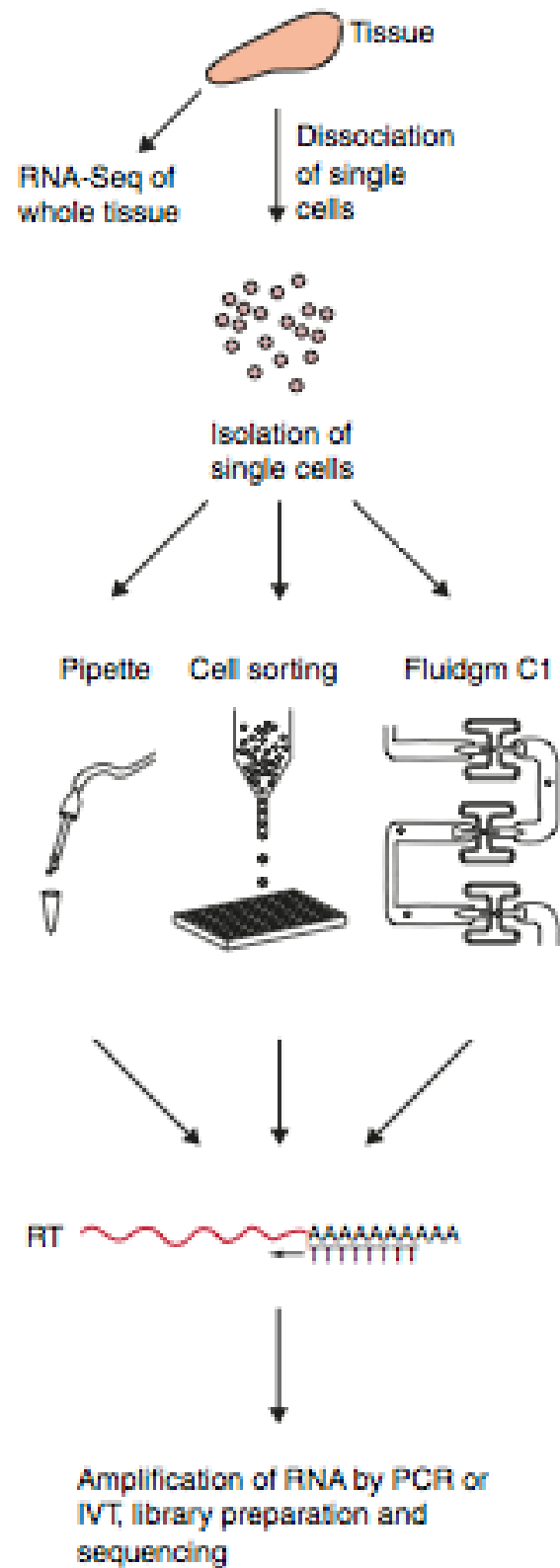
Foundational technology: (RT)-PCR



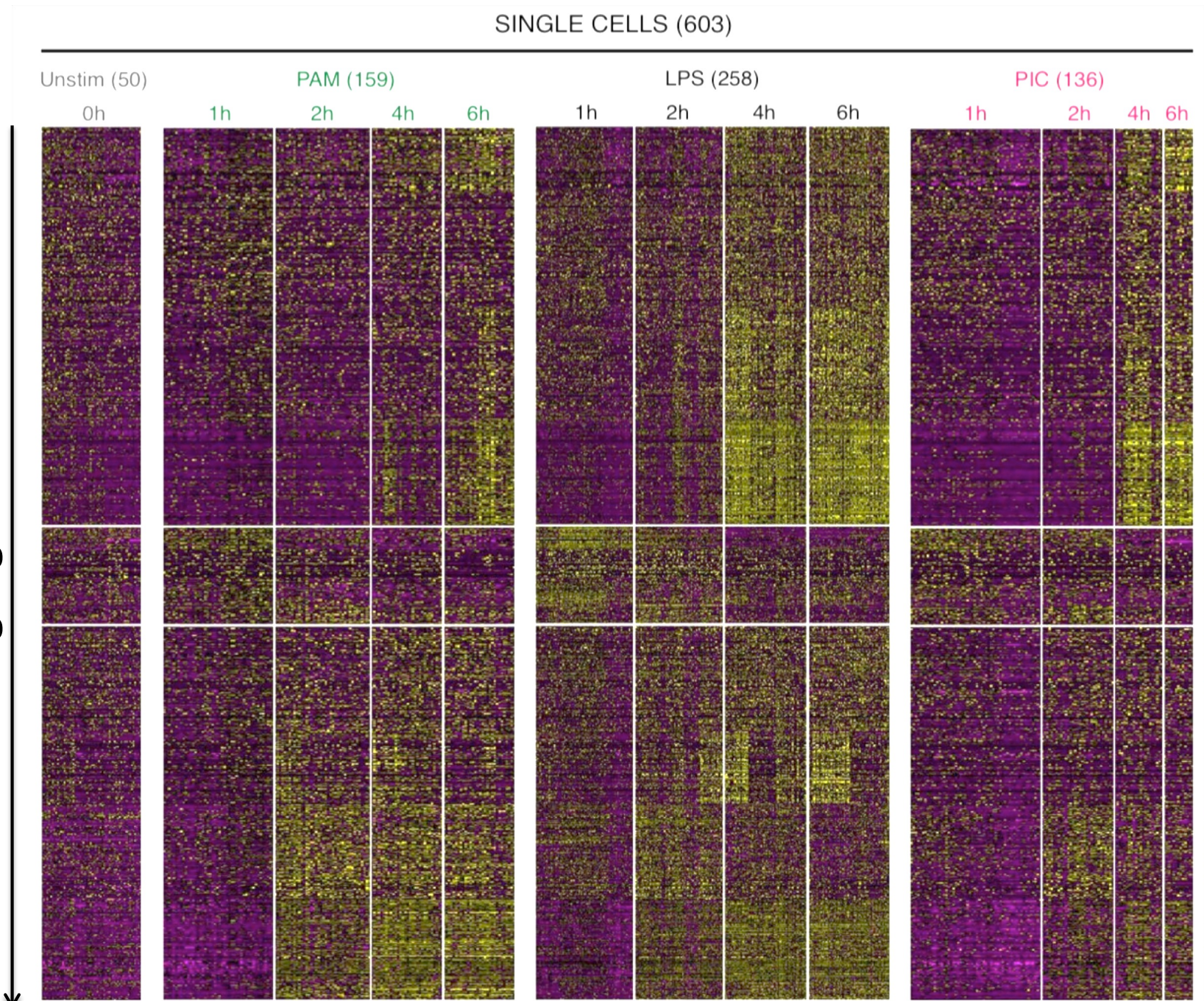
G Guo et al, Cell Stem Cell, 2013; 13: 492-505.

P Dalerba et al, Nature Biotech, 2011; 29: 1120-1127.

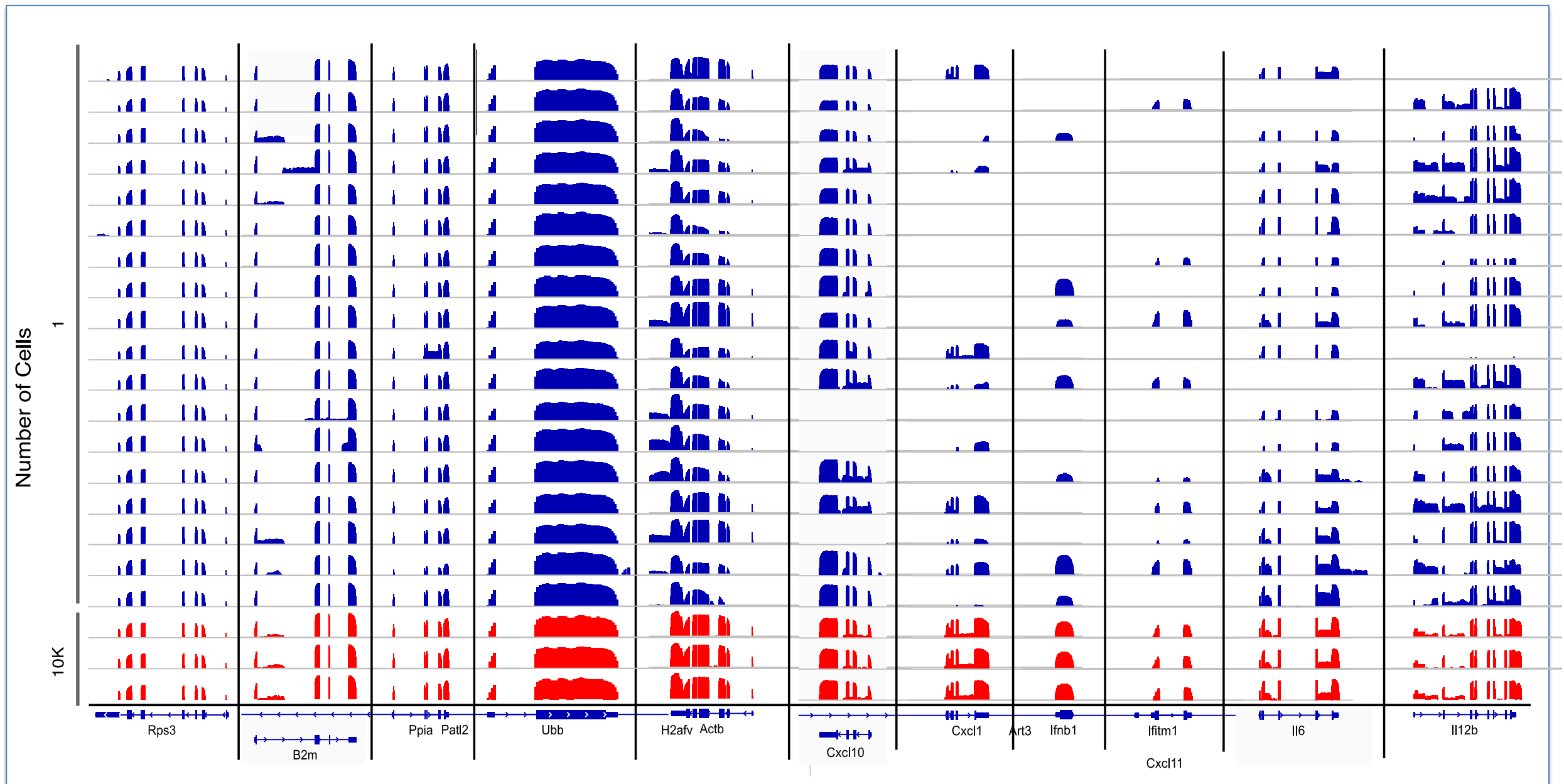
Scaling up: Single-cell RNA-Seq



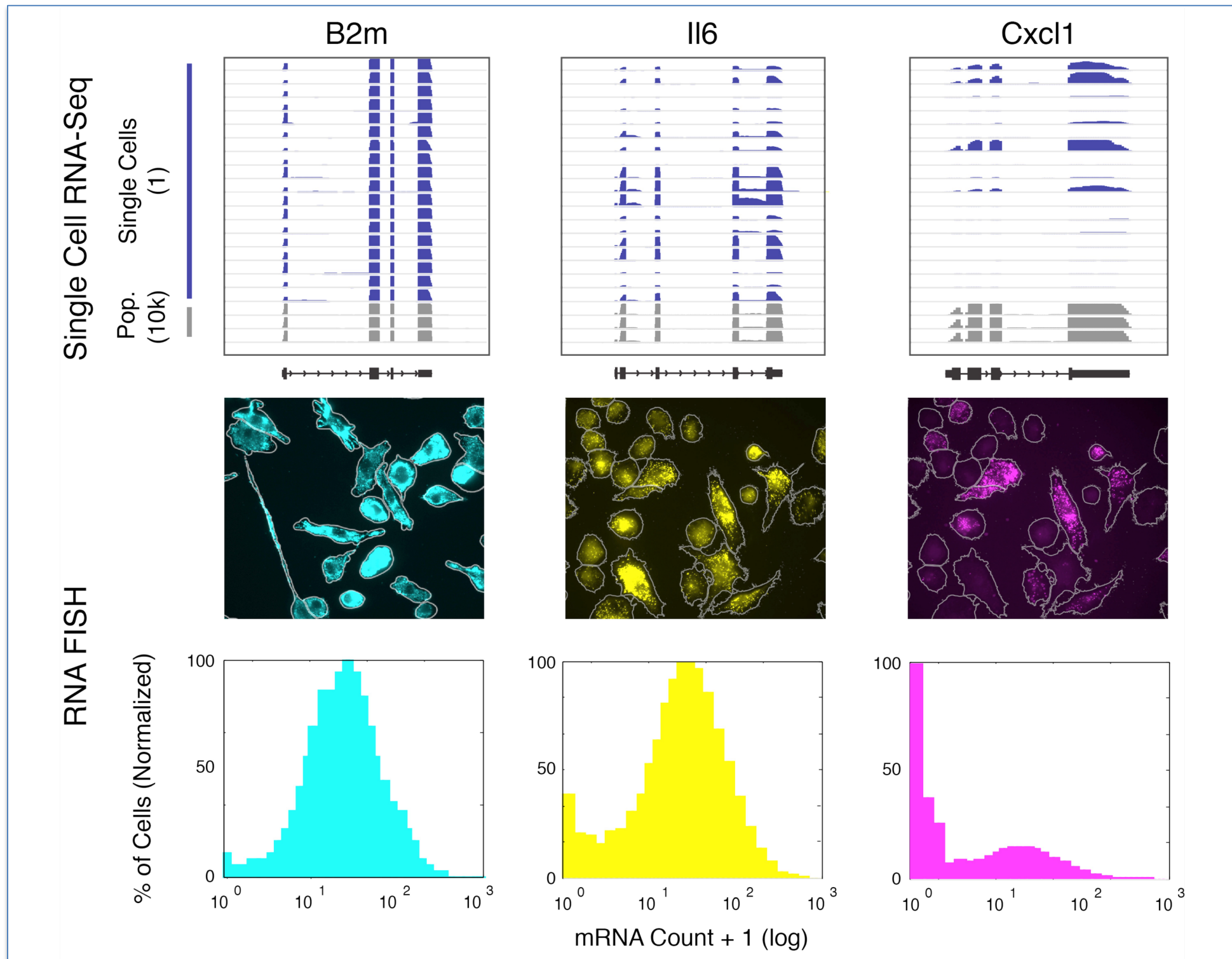
Single genes



scRNA data looks like RNA-seq



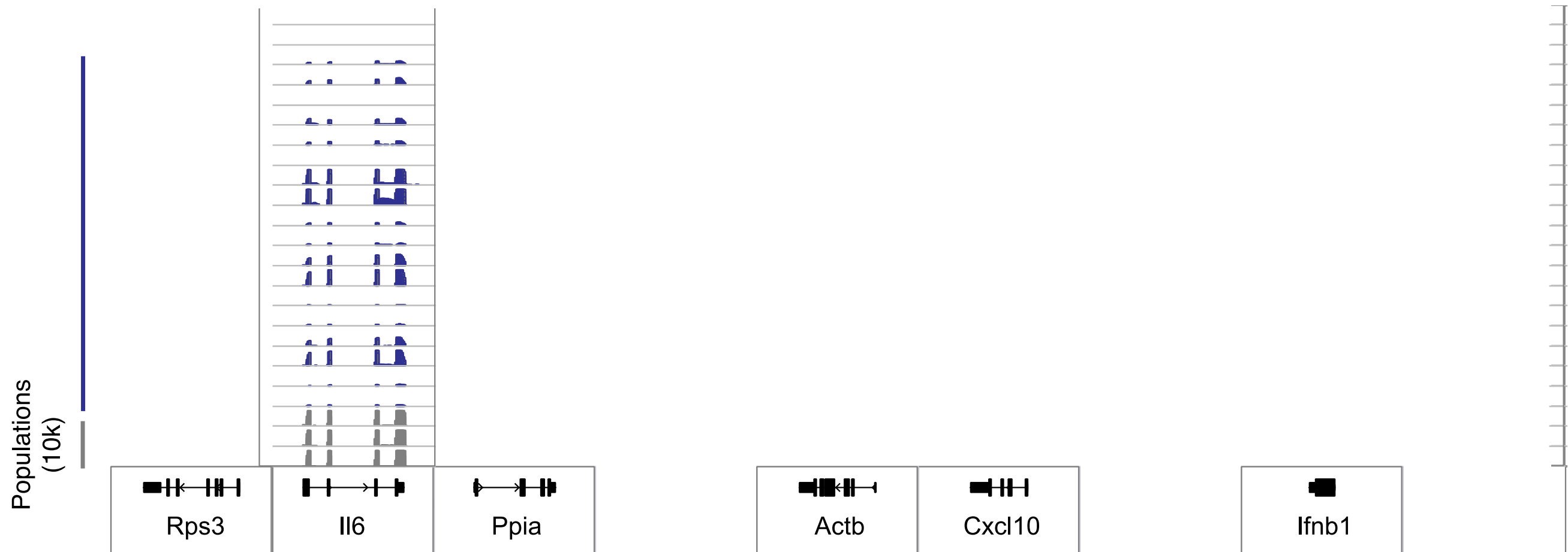
RNA FISH Validates Single-Cell RNA-Seq



Key single cell results can be validated using amplification-free methods.

Single-Cell RNA-Seq captures inter-cellular variability

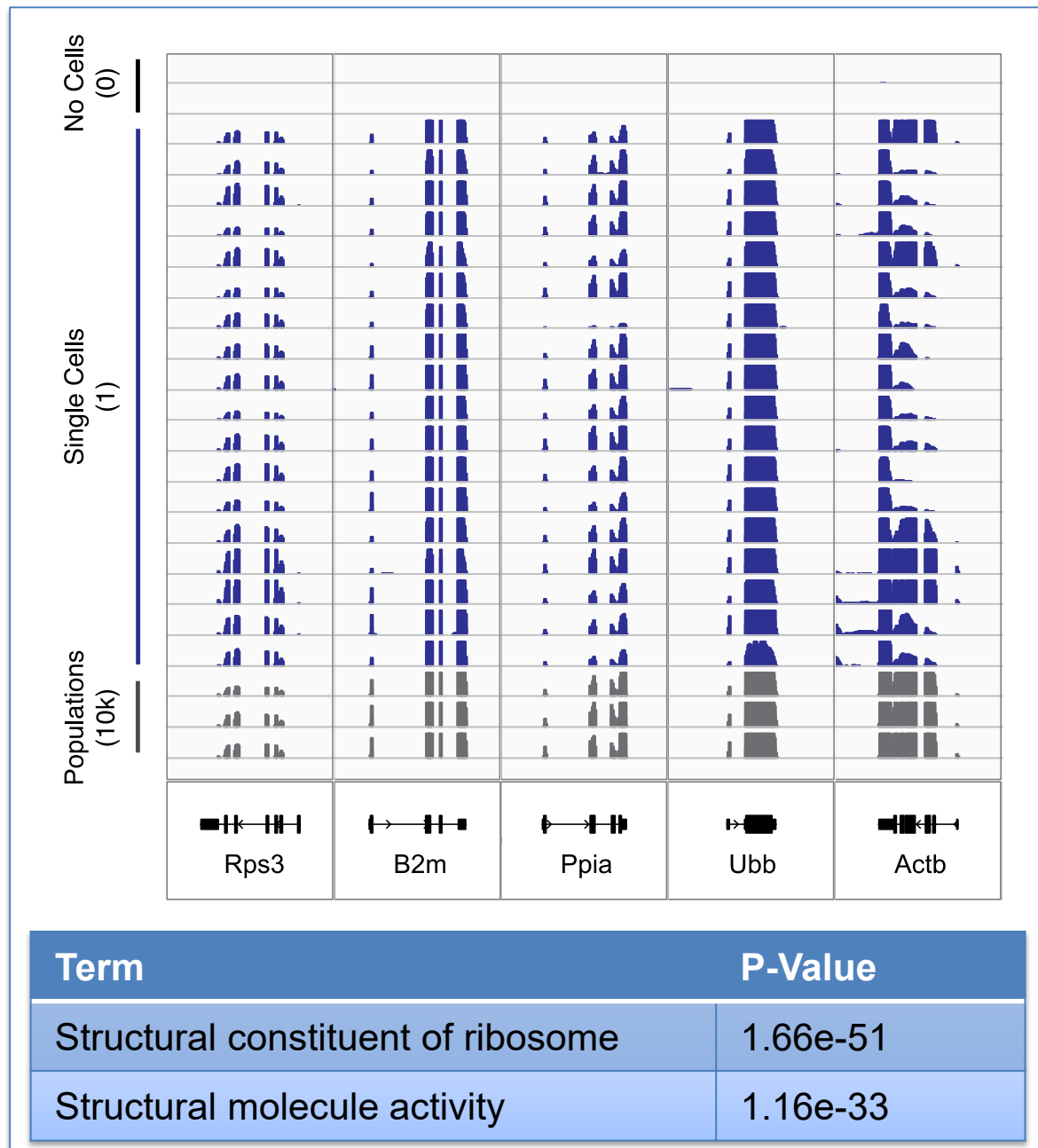
18 Single Dendritic Cells Stimulated with LPS



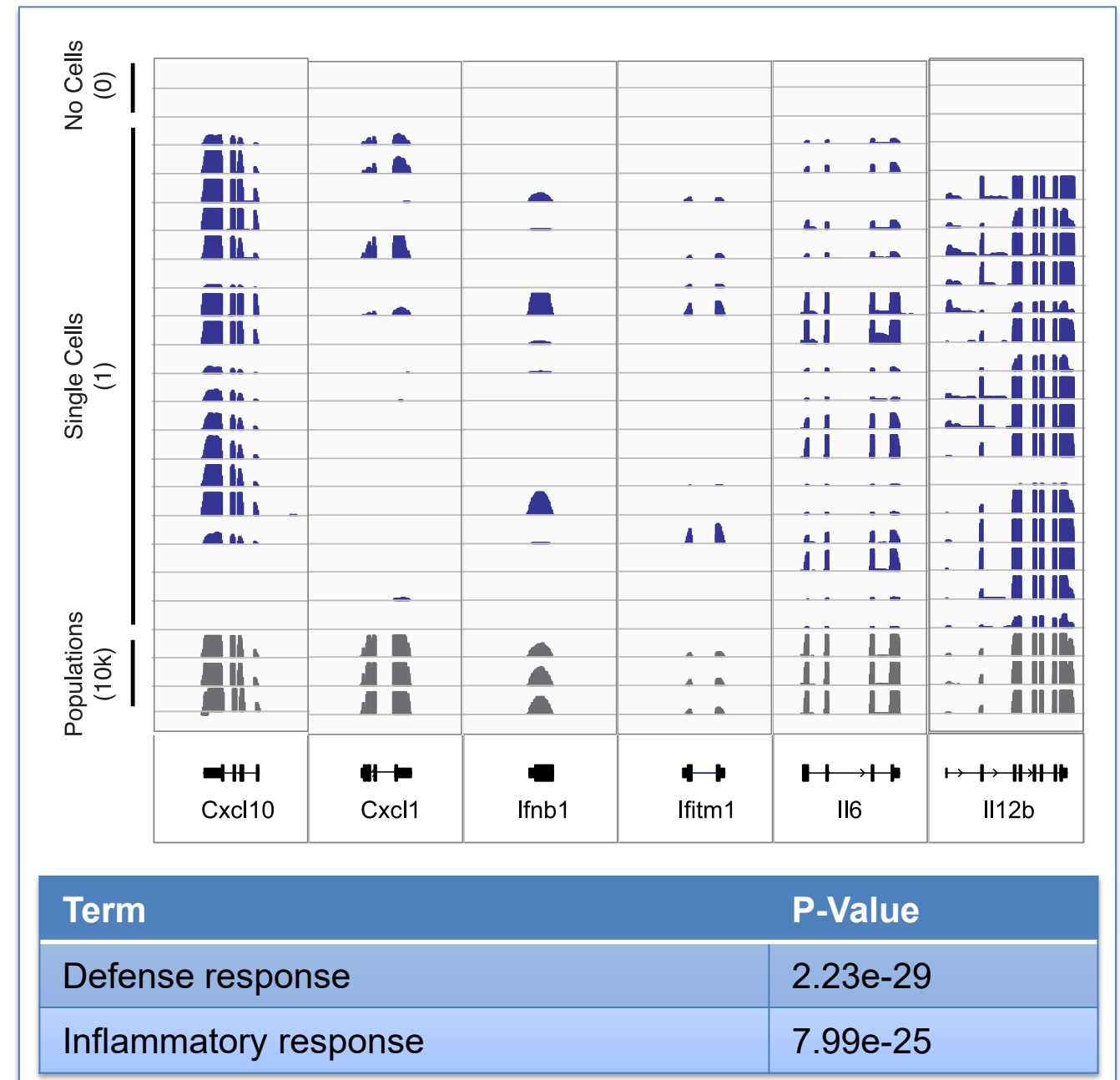
We can profile the full transcriptomes of cells with SMART-Seq.

Housekeeping vs. variable genes

Least Variable Genes



Most Variable Genes



Screenshots are log scale in IGV

Housekeeping & ribosomal genes are among the least variable, while immune response elements are among the most variable.

Shalek*, Satija* et al, *Nature*, **498**, (2013)

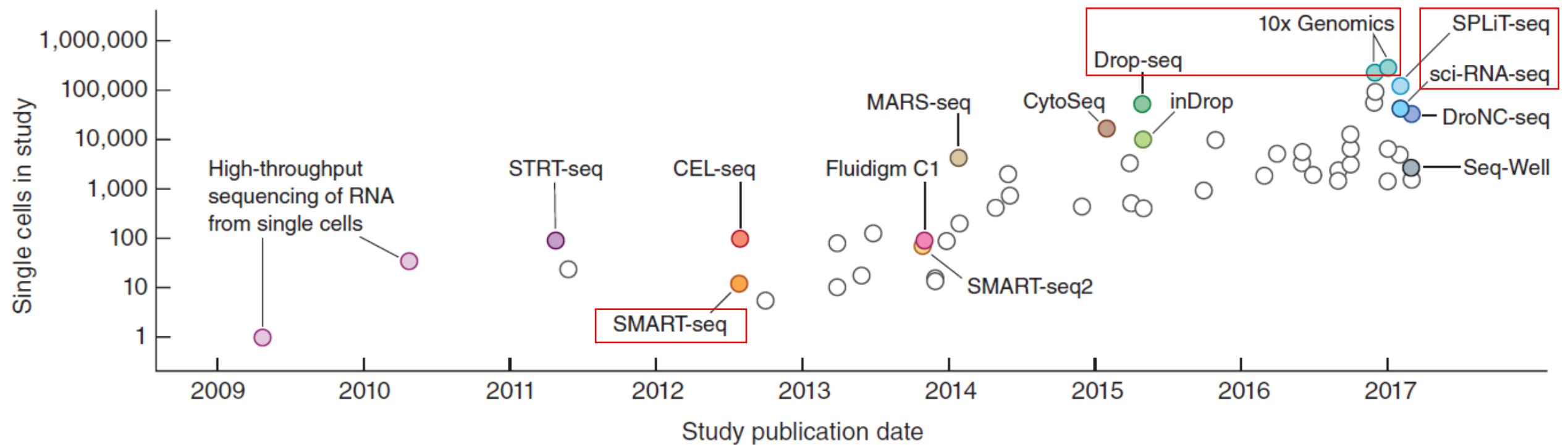
Today: Deep Learning for Single-cell Genomics

1. Why single cells, traditional approaches, scRNA-seq
2. Scaling up single-cell technologies: evolution of scRNA-seq
3. Dealing with noise, doublets, and other scRNA-seq issues
4. Computational challenges in scRNA-seq data analysis
5. Deep learning methods for scRNA-seq data analysis
6. Beyond scRNA-seq: scATAC-seq, multi-omics

2. Scaling up scRNA-seq technology

Exponential scaling of single-cell RNA-seq in the past decade

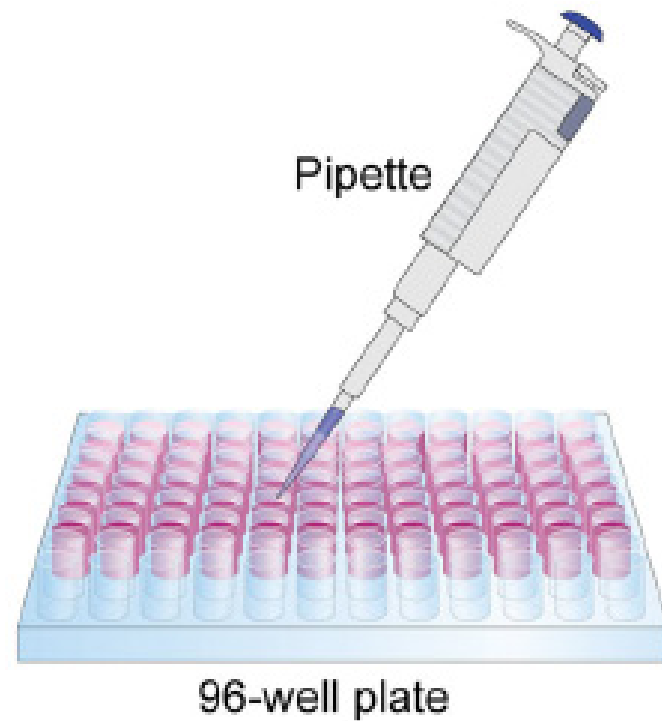
Cell numbers reported in representative publications by publication date. Key technologies are indicated.



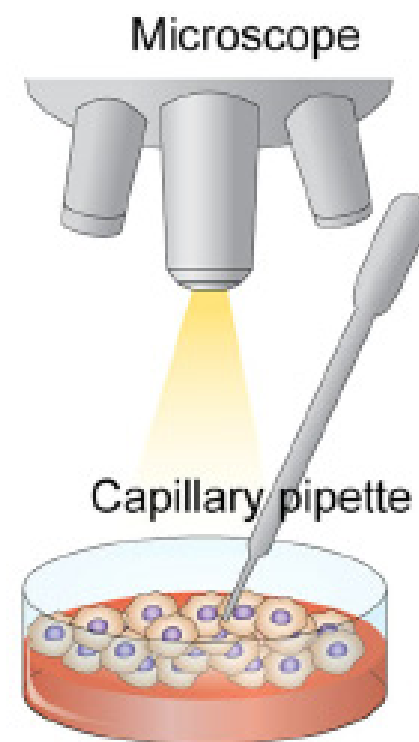
Svensson V, Vento-Tormo R, Teichmann SA. Nat Protoc. 2018 Apr;13(4):599-604.

Evolution of methods for isolating single cells for profiling

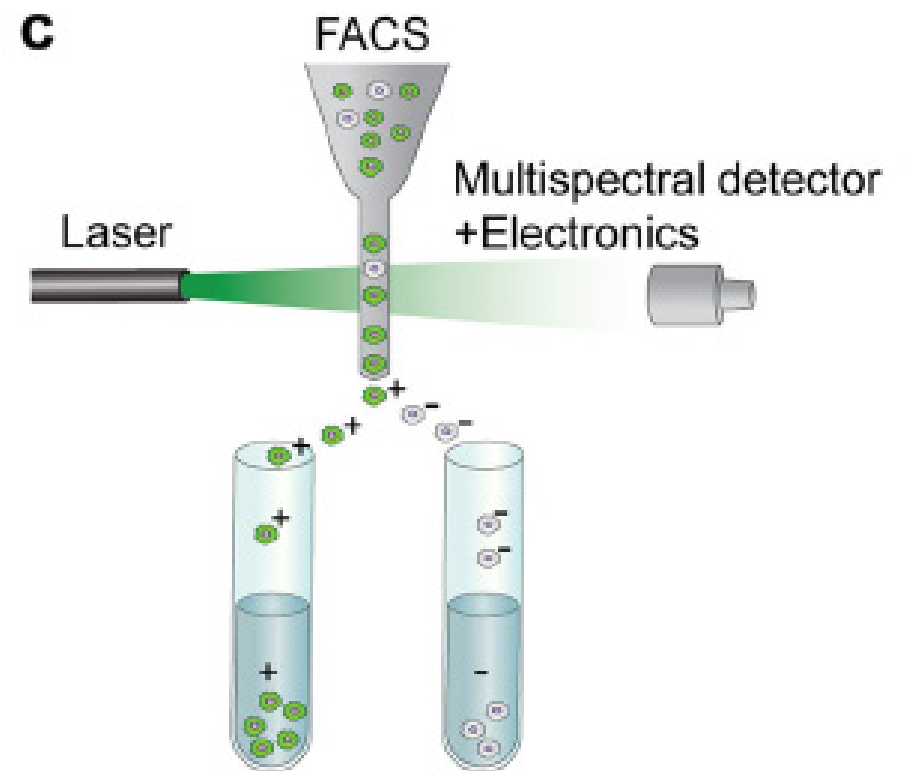
a



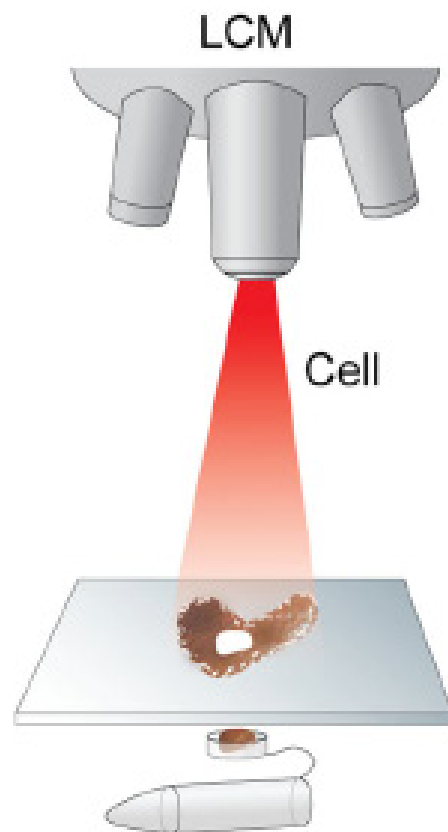
b



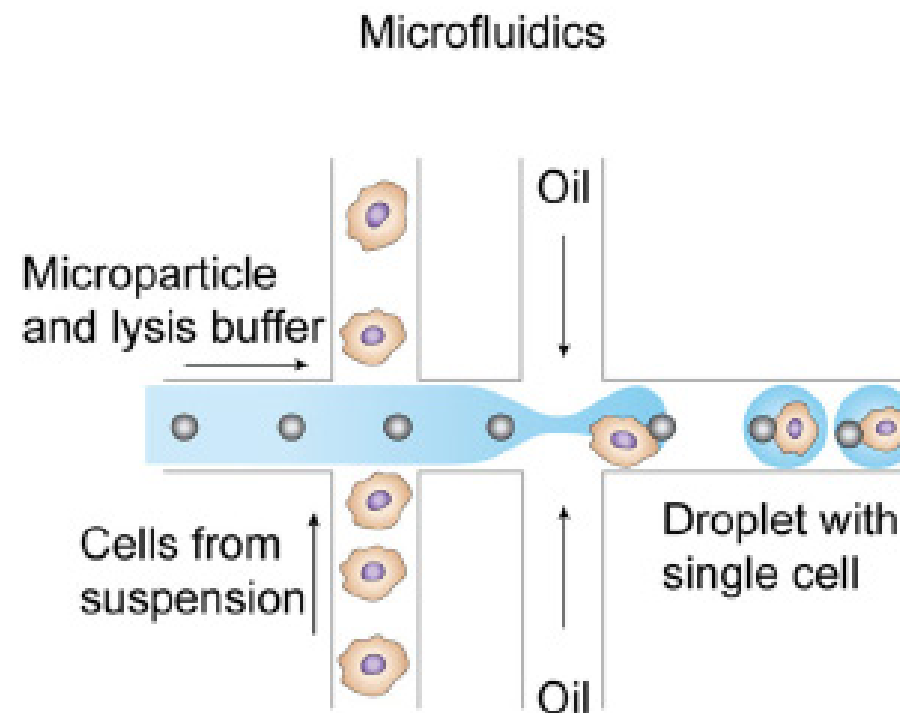
c



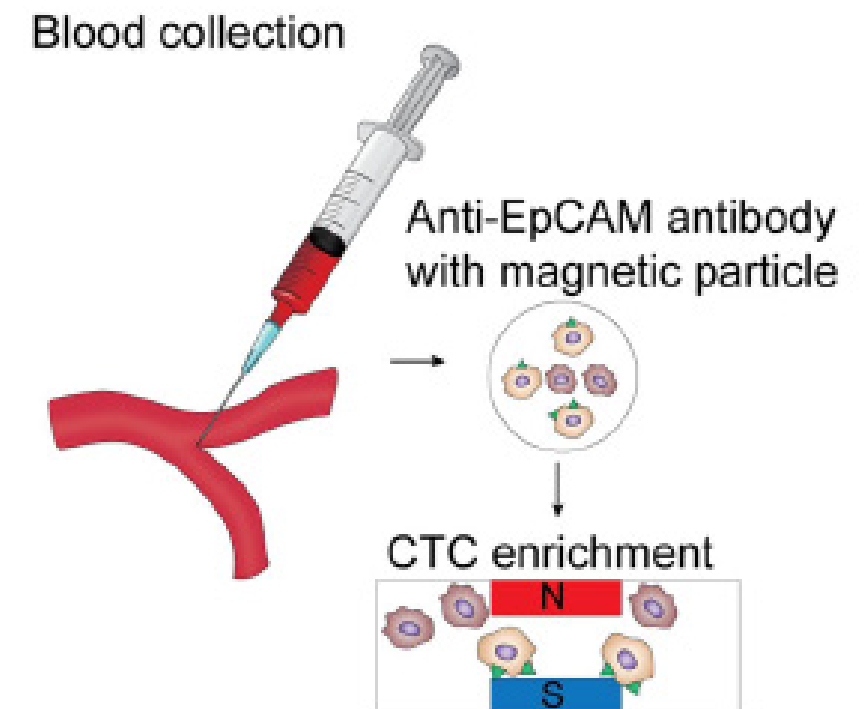
d



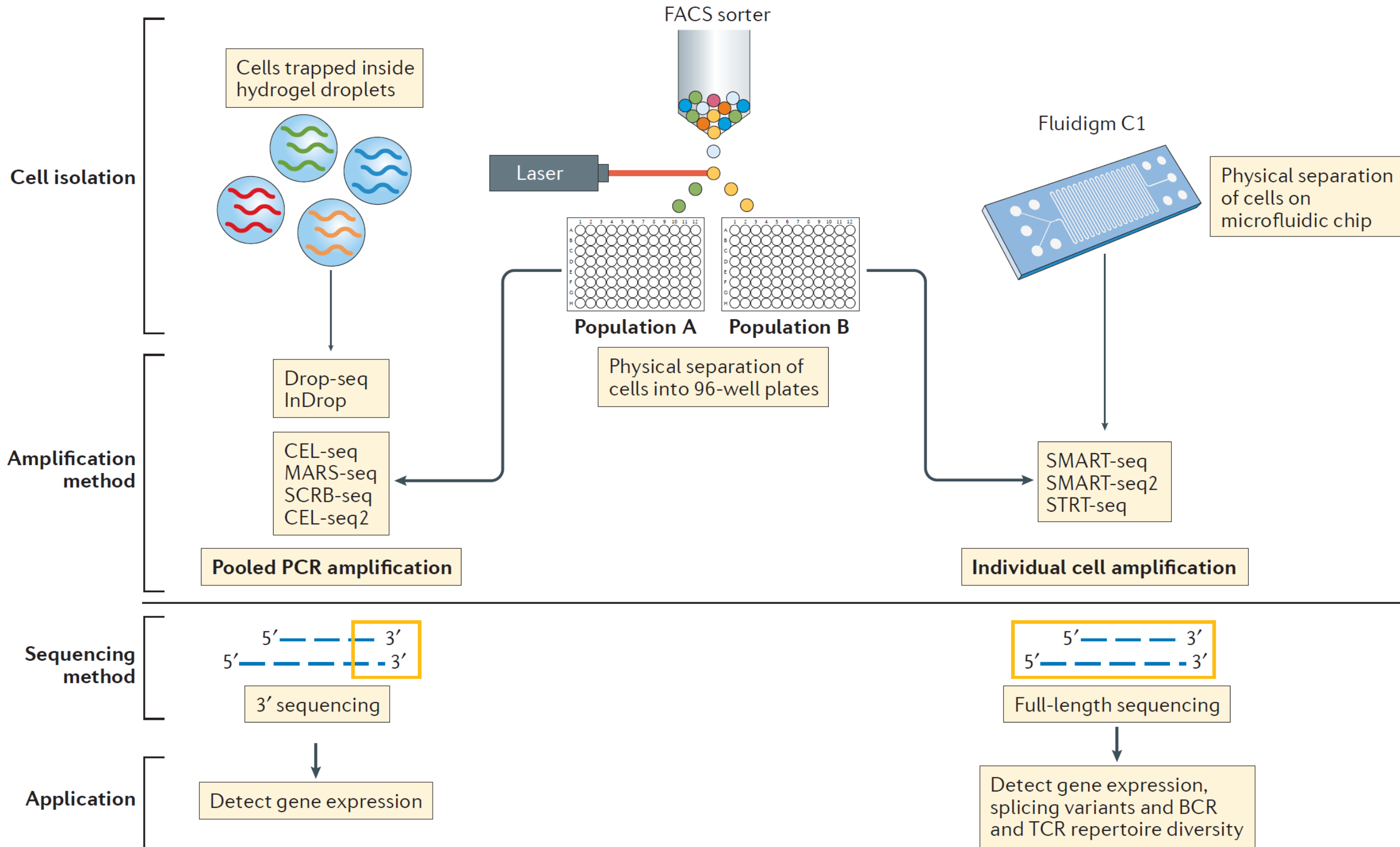
e



f

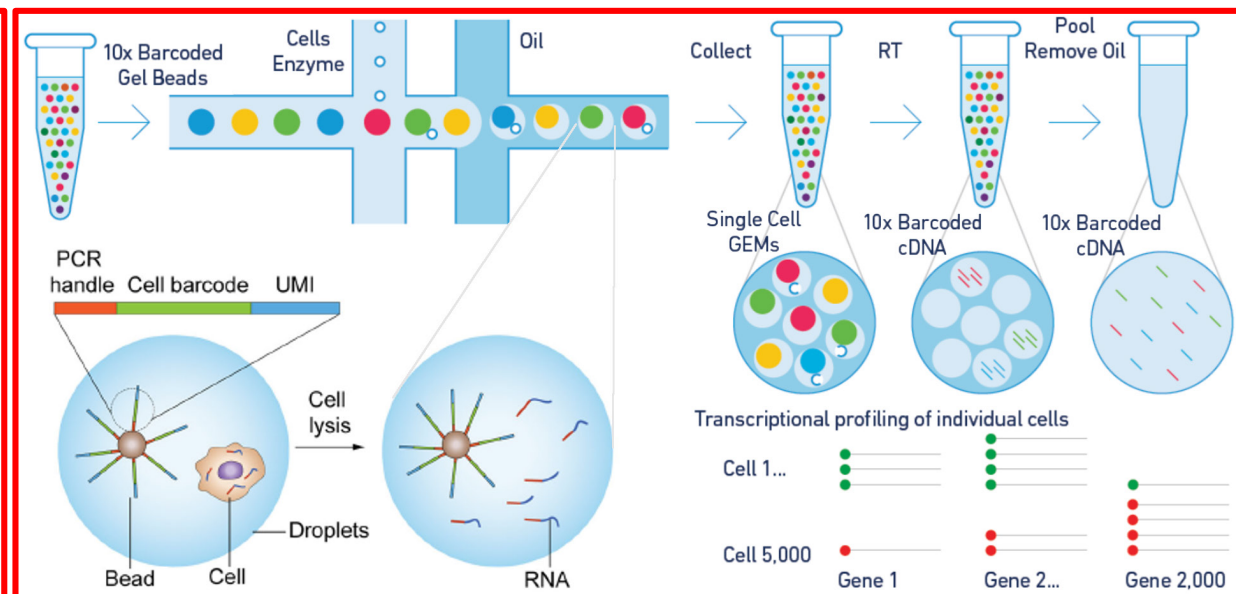
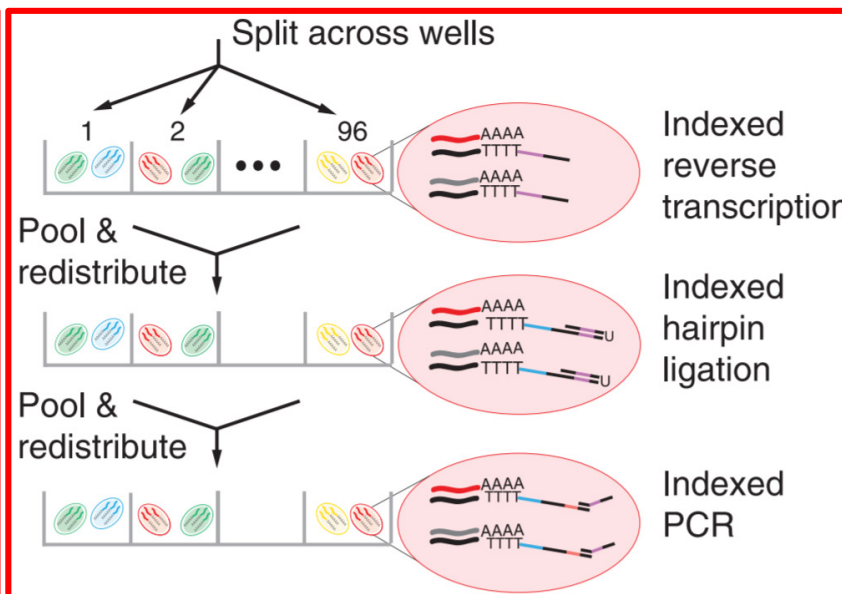
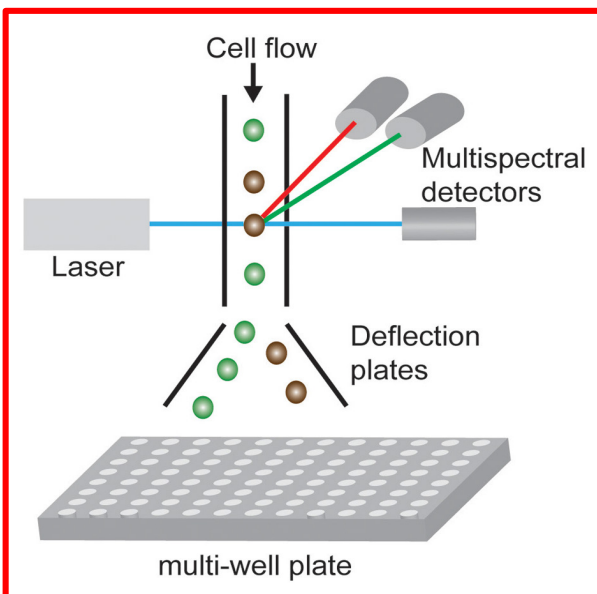


All methods seek to: separate cells, amplify RNA, sequence



scRNA-seq technologies vary in cost and sensitivity

Paplexi et al. 2017	✗ FACS	✗ CyTOF	✗ qPCR	↓ Plate-based protocols (STRT-seq, SMART-seq, SMART-seq2)	✗ Fluidigm C1	↓ Pooled approaches (CEL-seq, MARS-seq, SCRB-seq, CEL-seq2)	↓ Massively parallel approaches (Drop-seq, InDrop)
Cell capture method	Laser	Mass cytometry	Micropipettes	FACS	Microfluidics	FACS	Microdroplets
Number of cells per experiment	Millions	Millions	300–1,000	50–500 ★	48–96 ✗	500–2,000 ★	5,000–10,000 ★
Cost	\$0.05 per cell	\$35 per cell	\$1 per cell	\$3–6 per well	\$35 per cell	\$3–6 per well	\$0.05 per cell
Sensitivity	Up to 17 markers ✗	Up to 40 markers ✗	10–30 genes per cell ✗	7,000–10,000 genes per cell for cell lines; 2,000–6,000 genes per cell for primary cells	6,000–9,000 genes per cell for cell lines; 1,000–5,000 genes per cell for primary cells	7,000–10,000 genes per cell for cell lines; 2,000–6,000 genes per cell for primary cells	5,000 genes per cell for cell lines; 1,000–3,000 genes per cell for primary cells



CEL-seq, cell expression by linear amplification and sequencing; CyTOF, cytometry by time of flight (mass cytometry); FACS, fluorescence-activated cell sorting; InDrop, indexing droplets sequencing; MARS-seq, massively parallel single-cell RNA sequencing; qPCR, quantitative PCR; SCRB-seq, single-cell RNA barcoding and sequencing; STRT-seq, single-cell tagged reverse transcription sequencing.

Single-Cell Expression Profiling Pipeline

1. Cell Harvest

- Harvest cells in media
- Pre-enrich (FACS)

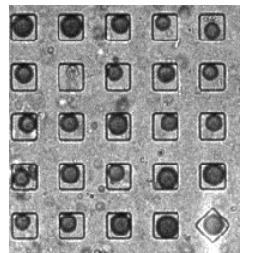
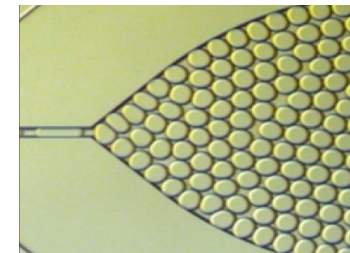
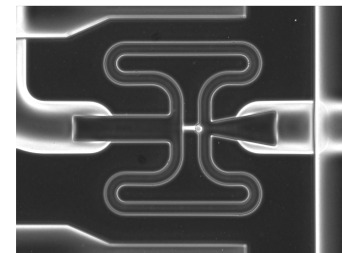
Cell work



2. Single Cell Preparation

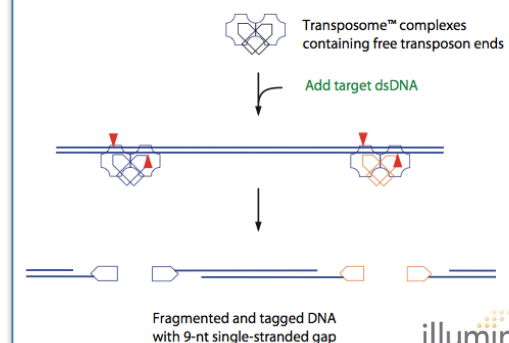
- (a) C_1 : Cells $\rightarrow$ Whole Transcriptome Amplification
- (b) Multiwell Plates
- (c) Next Gen Technologies

Single cell processing



3. Expression Profiling

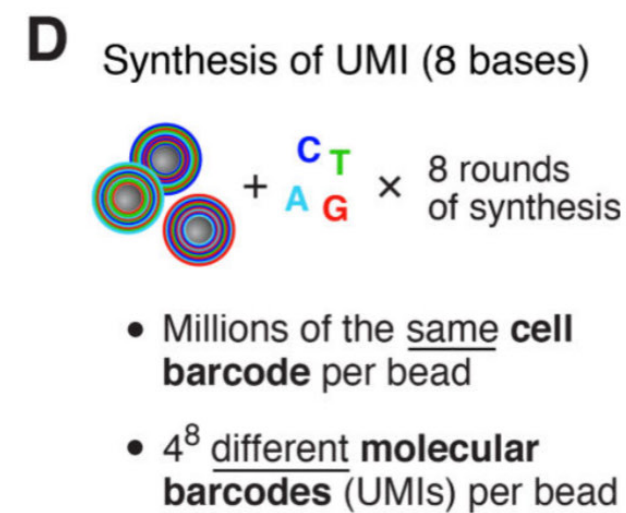
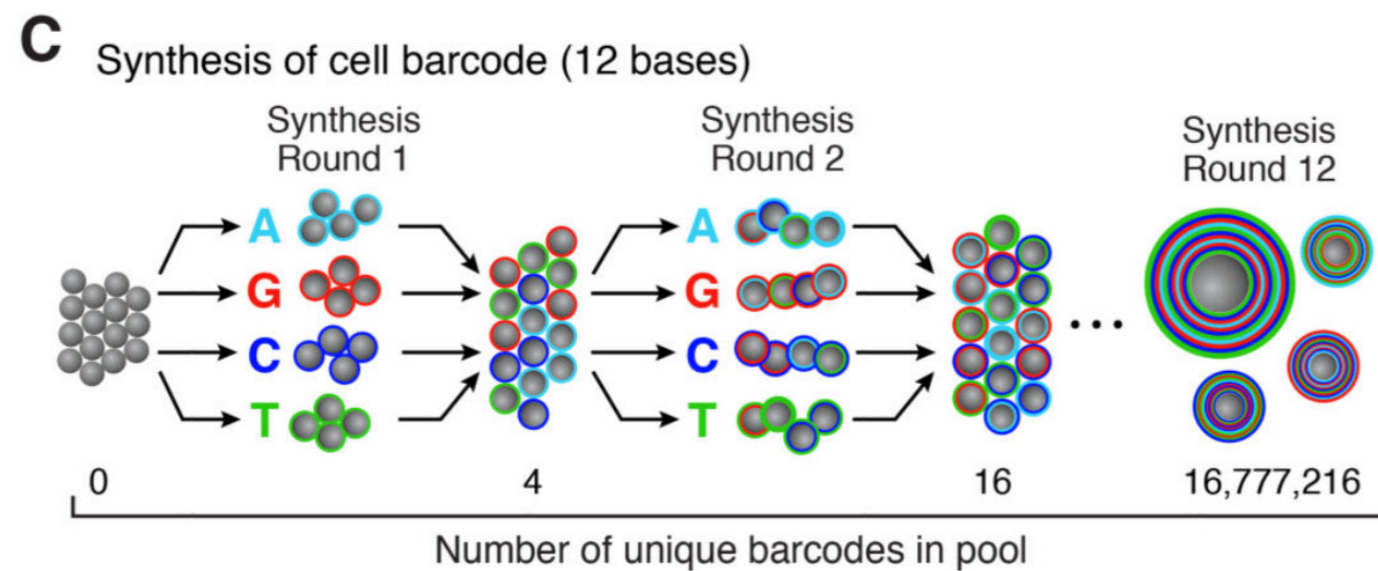
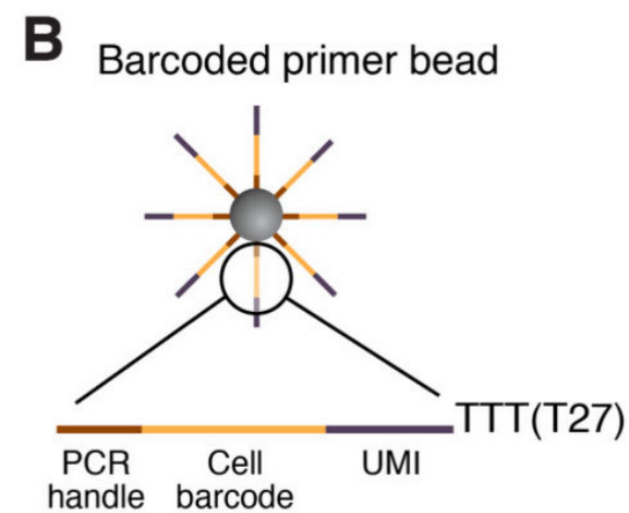
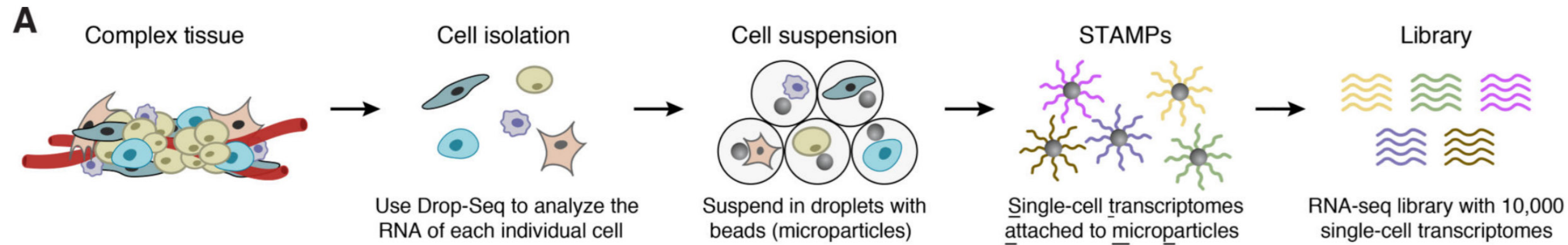
Library construction, validation, HiSeq



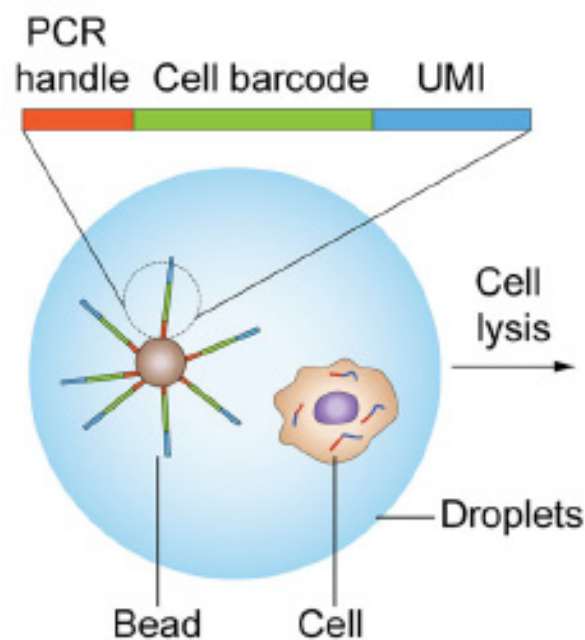
We can go from single cells to aligned reads in less than a day

w/ Fluidigm, E. Macosko, S. McCarroll, A. Regev, D. Weitz, C. Love, T. Gierahn, & Others

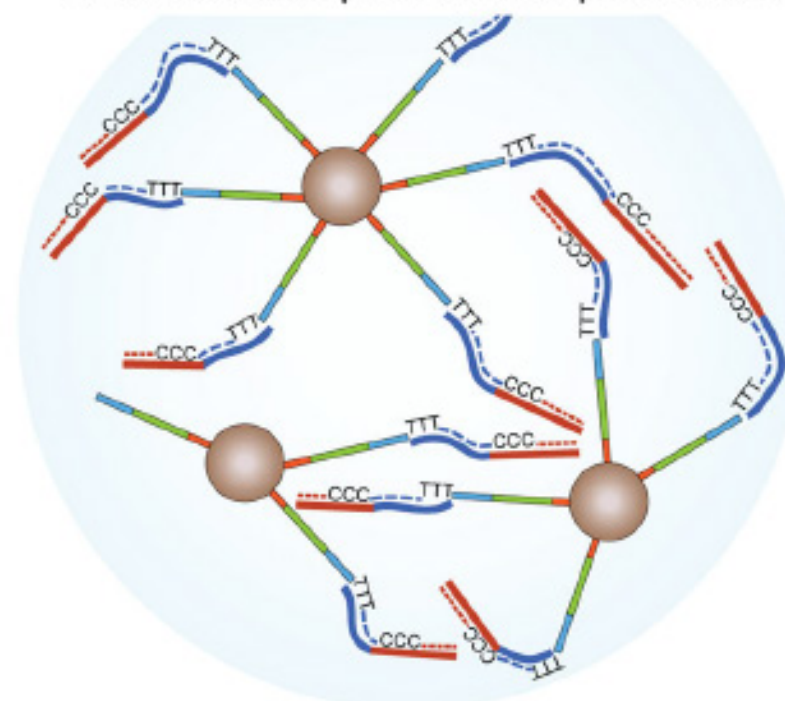
Drop-seq: Droplets as reaction chambers (10x)



9 Structure of the barcode primer bead



Reverse transcription with template switching



SPLIT-seq/sci-RNA-seq: Sequential combinatorial barcoding



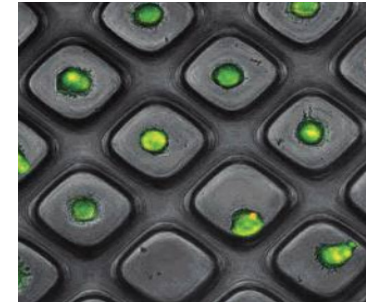
<https://sites.google.com/uw.edu/splitseq/home>

- Single cells never individually isolated
- Instead: fixed, and mRNA is manipulated in situ inside each cell
- Split cells into ~100 wells (e.g. 96 or 384-well plate) with unique barcodes in each well
- Labels all cells with a first barcode, for that well. Chance of same barcode: **1/100**
- Pool cells, shuffle, split again, **randomly** re-assorting into same set of ~100 wells
- Add second barcode. Chance of same 2 barcodes: **1/10,000**.
- Repeat: pool, shuffle, split, add 3rd barcode. Chance of same 3 barcodes: **1/1,000,000**
- Can scale number of cells exponentially by number of barcoding rounds

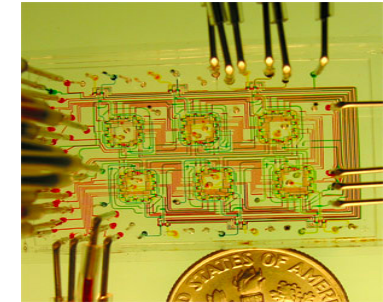
Single-cell Profiling technologies

1. Cells in wells, traps, and valves (nanowell, Flow sorting, Fluidigm C1)

- Screen for and retrieve single cells of interest
- Enrich for rare cells with desired properties
- Control the cellular microenvironment
- Monitor or control cell-cell interactions
- Precise/extensive manipulation of single cells



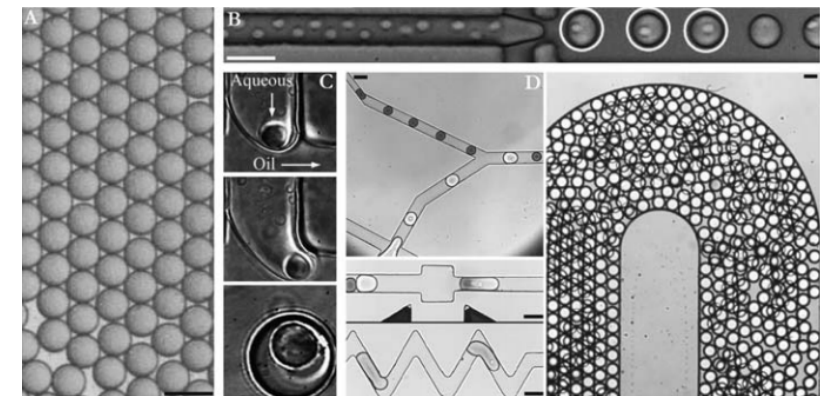
Passive wells



Active pumps and valves

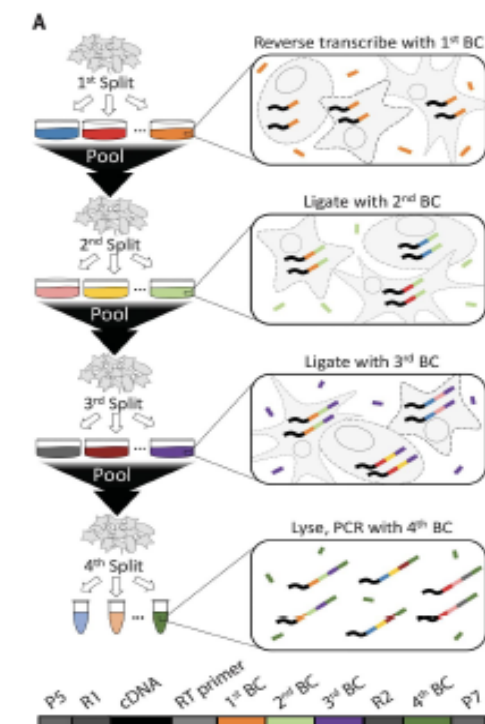
2. Droplets (Drop-seq, ddPCR)

- Introduce distinct “packets” of reagents to single cells
 - e.g., primers, barcodes
- Perform amplification on individual cells
- Sort large populations of single cells



3. Combinatorial indexing (SCI-seq, SPLiT-seq)

- Economic use of reagents for cell separation
- Efficiency of handling larger populations than Drop-seq
- Maintain complexities of population without bias from droplet or well.



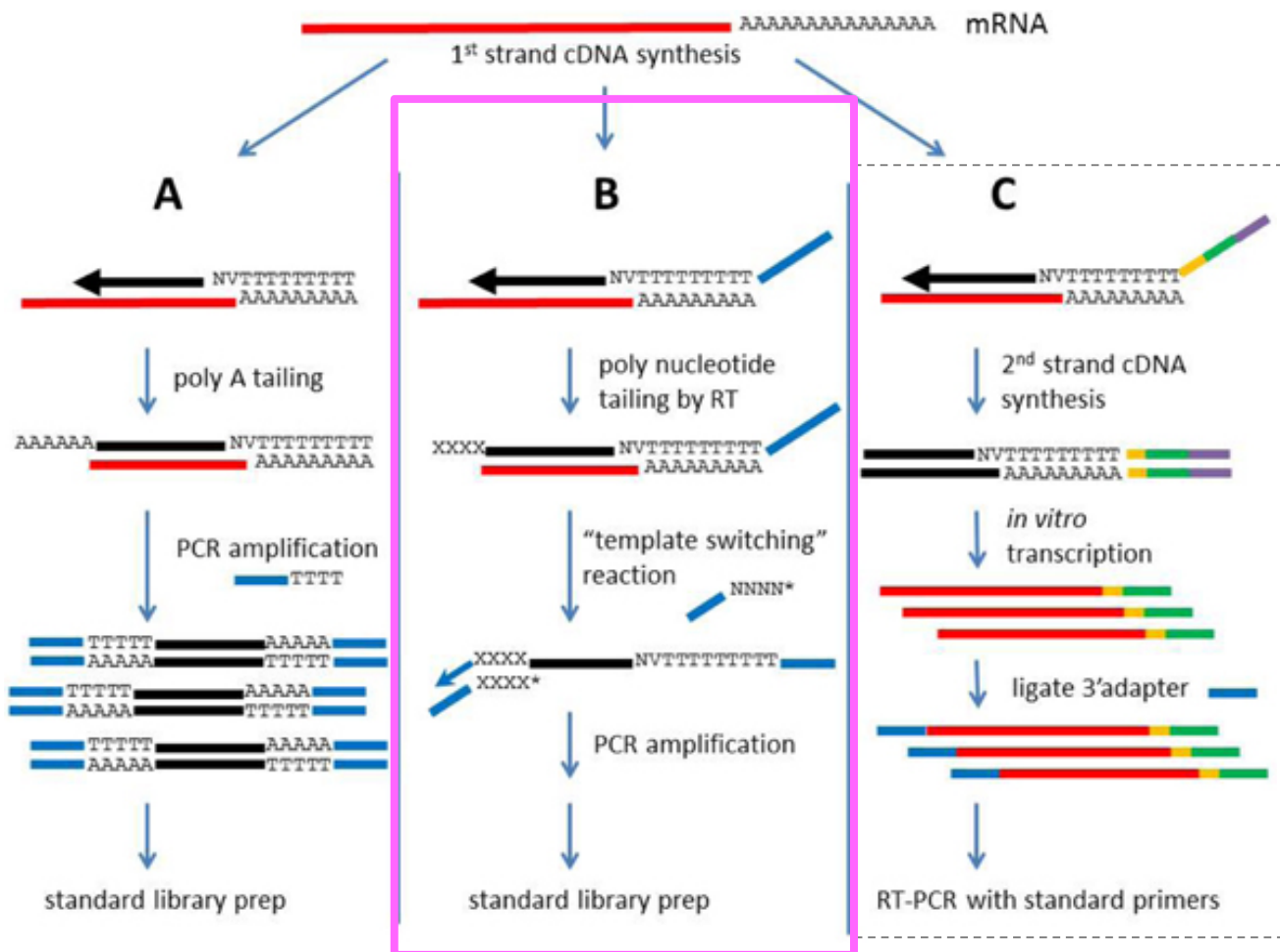
Today: Deep Learning for Single-cell Genomics

1. Why single cells, traditional approaches, scRNA-seq
2. Scaling up single-cell technologies: evolution of scRNA-seq
3. Dealing with noise, doublets, and other scRNA-seq issues
4. Computational challenges in scRNA-seq data analysis
5. Deep learning methods for scRNA-seq data analysis
6. Beyond scRNA-seq: scATAC-seq, multi-omics

3. Dealing with noise and doublets in snRNA-seq data

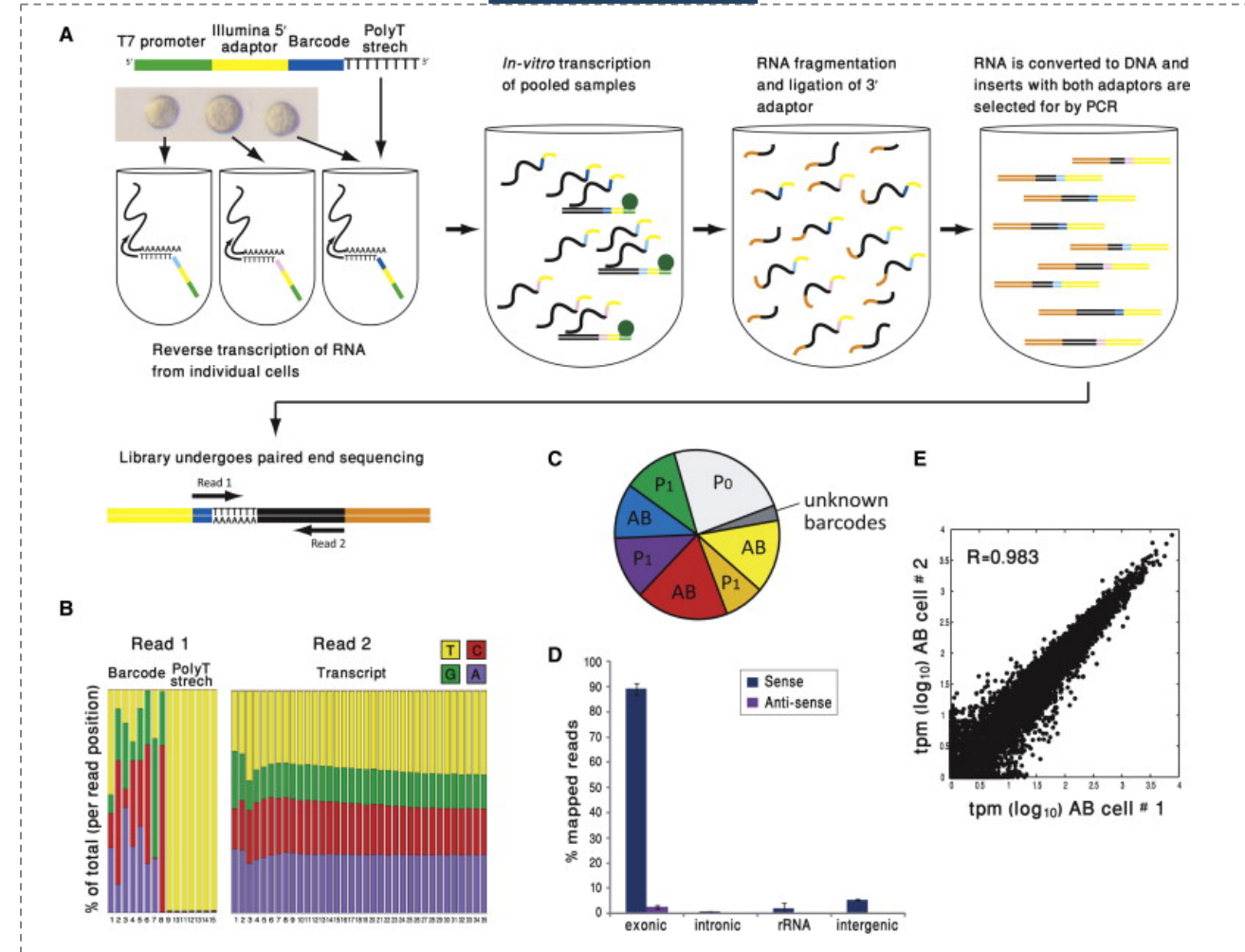
Dealing with rRNA contamination

Tang/Quartz-Seq | SMART-Seq | CEL-Seq



From www.biotechniques.com

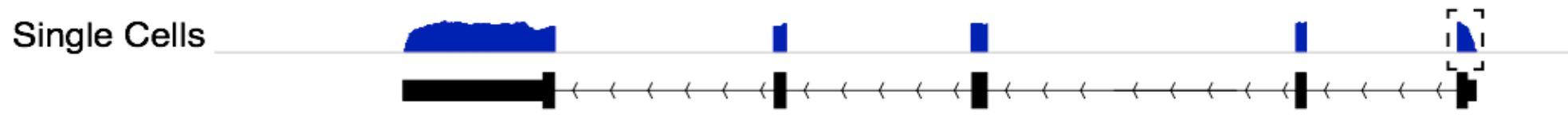
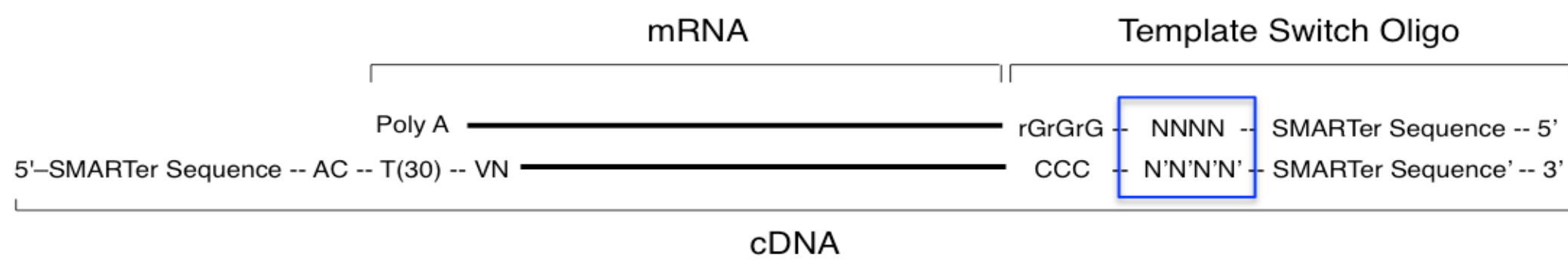
CEL-Seq



T Hashimshony et al, Cell Rep, 2012; 2: 666-673.

Ribosomal RNA contamination
rRNA overwhelms mRNA (~98%)
polyA-selection is too inefficient

Dealing w/PCR bias: Unique molecular identifiers (UMIs)



SMARTer Sequence	NNNN	GGG	Read
AAGCAGTGGTATCAACGCAGAGT	GCGG	GGG	GAAAAGGGGCTGTTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCC
AAGCAGTGGTATCAACGCAGAGT	AGAG	GGG	GAGGGCTGTTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCCTGTG
AAGCAGTGGTATCAACGCAGAGT	ACGA	GGG	GGGCTGTTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCCTGTGAA
AAGCAGTGGTATCAACGCAGAGT	CGGG	GGG	GAAAAGGGGCTGTTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCC
AAGCAGTGGTATCAACGCAGAGT	AGCG	GGG	AGGGCTGTTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCCTGTGA
AAGCAGTGGTATCAACGCAGAGT	CCGG	GGG	GCAAAGGGCTGTTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCCT
AAGCAGTGGTATCAACGCAGAGT	CGAG	GGG	GGCTGTTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCCTGTGAAG
AAGCAGTGGTATCAACGCAGAGT	CCCG	GGG	GAAAAGGGGCTGTTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCC
AAGCAGTGGTATCAACGCAGAGT	AGCG	GGG	AGGGCTGCTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCCTGTGA
AAGCAGTGGTATCAACGCAGAGT	AGAG	GGG	GAGGGCTGTTTCAGGGCTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCCTGTG
AAGCAGTGGTATCAACGCAGAGT	AGCG	GGG	AGGGCTGTTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCCTGTGA
AAGCAGTGGTATCAACGCAGAGT	CCCG	GGG	GAAAAGGGGCTGTTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCC
AAGCAGTGGTATCAACGCAGAGT	GTAT	GGG	CAACGCAGAGTAATCGGGGAGGGCTGTTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTG
AAGCAGTGGTATCAACGCAGAGT	AAGG	GGG	GAGGGCTGTTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCCTGTG
AAGCAGTGGTATCAACGCAGAGT	AAGG	GGG	GAGGGCTGTTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCCTGTG
AAGCAGTGGTATCAACGCAGAGT	GTCT	GGG	GGGAGGGCTGTTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCCTG
AAGCAGTGGTATCAACGCAGAGT	GTAT	GGG	CAACGCAGAGTAATCGGGGAGGGCTGTTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTG
AAGCAGTGGTATCAACGCAGAGT	AGCG	GGG	GAGGGCTGTTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCCTGTG
AAGCAGTGGTATCAACGCAGAGT	CCCG	GGG	GAAAAGGGGCTGTTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCC
AAGCAGTGGTATCAACGCAGAGT	AAGG	GGG	GAGGGCTGTTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCCTGTG
AAGCAGTGGTATCAACGCAGAGT	AGCG	GGG	AGGGCTGTTTGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCCTGTGA

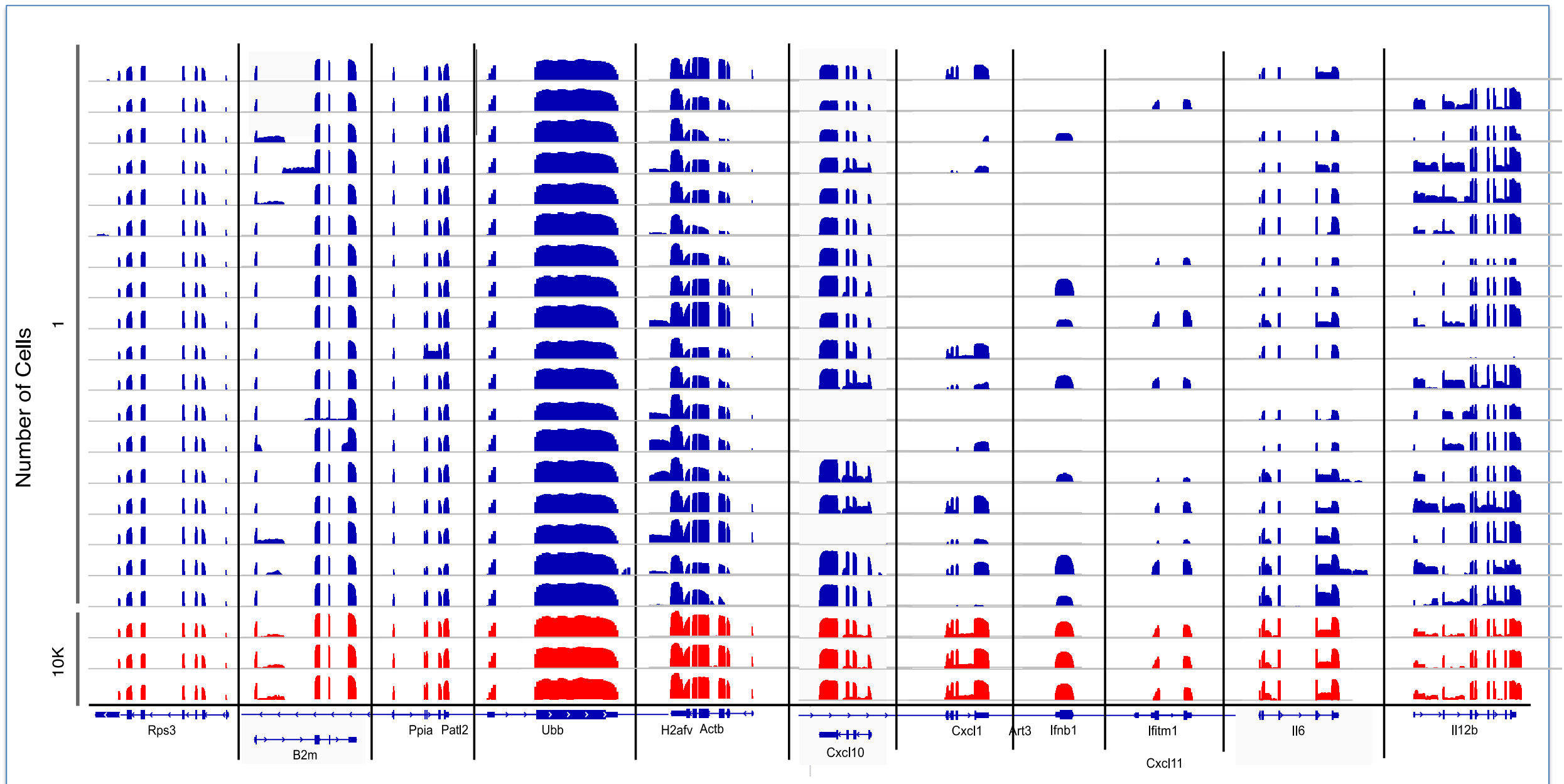
Ultra-low input RNA-seq is problematic

Bias in early PCR stages when using random priming

“PCR Jackpotting” of some RNA molecules

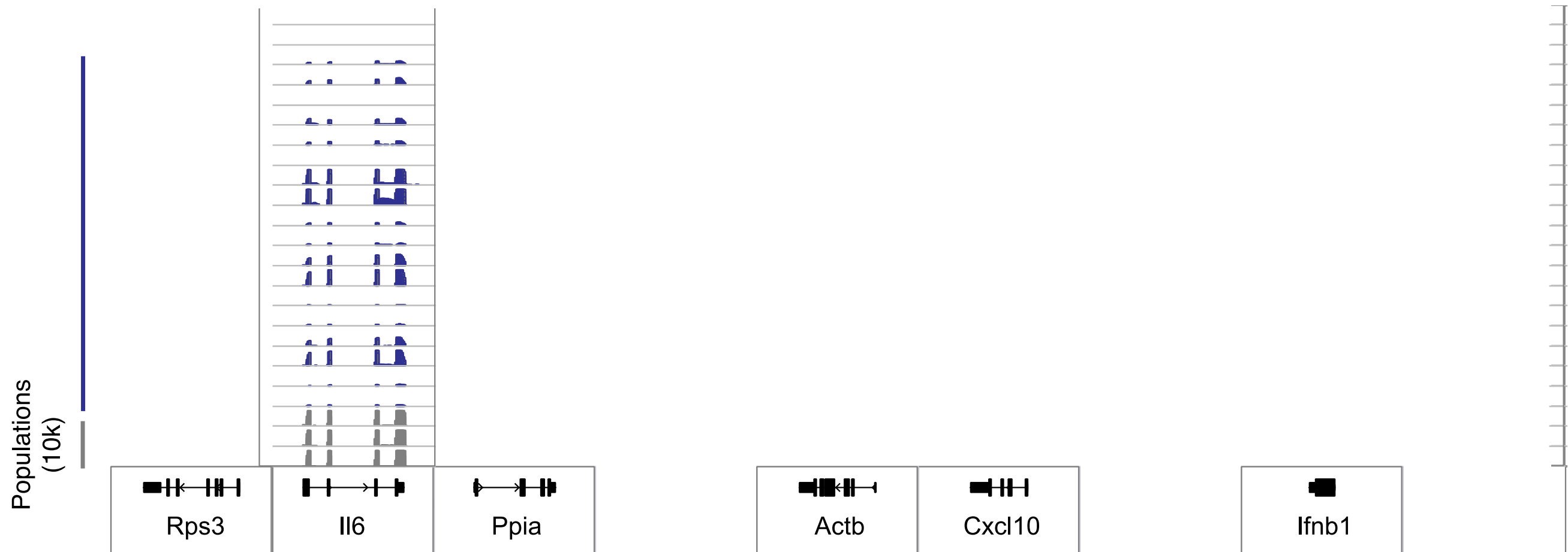
Suppression of some RNA molecules

scRNA data looks like RNA-seq



Single-Cell RNA-Seq captures inter-cellular variability

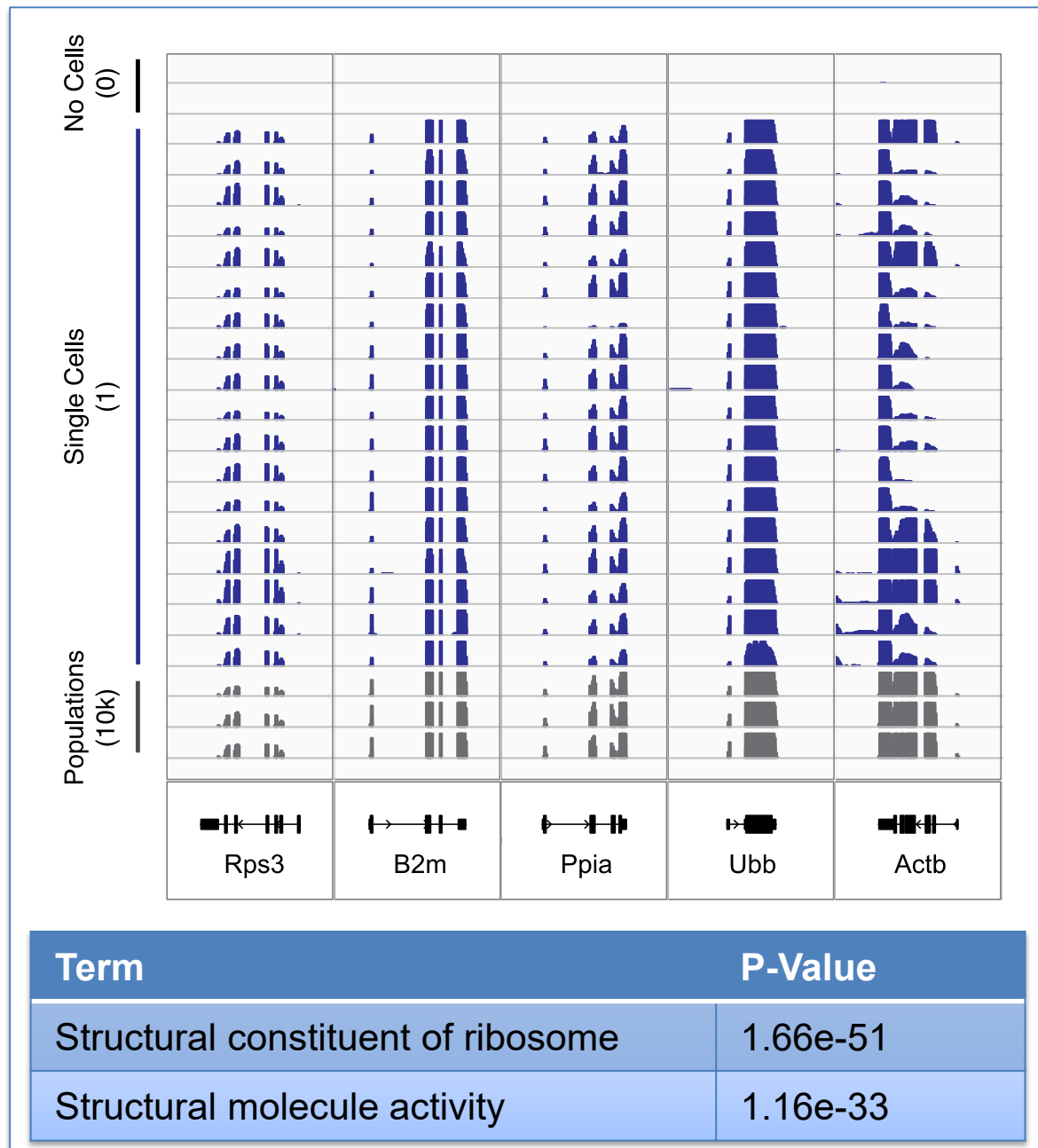
18 Single Dendritic Cells Stimulated with LPS



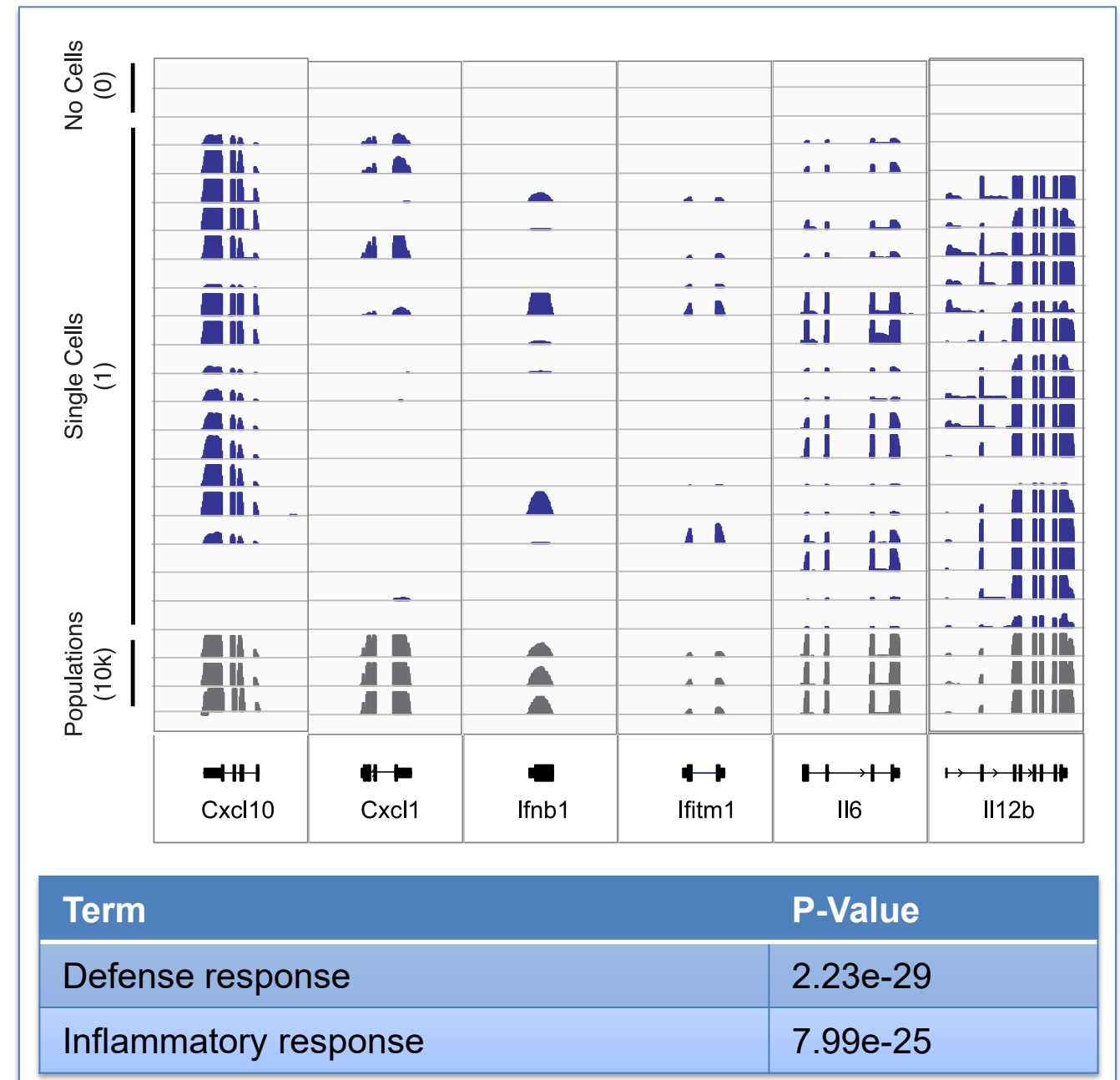
We can profile the full transcriptomes of cells with SMART-Seq.

Housekeeping vs. variable genes

Least Variable Genes



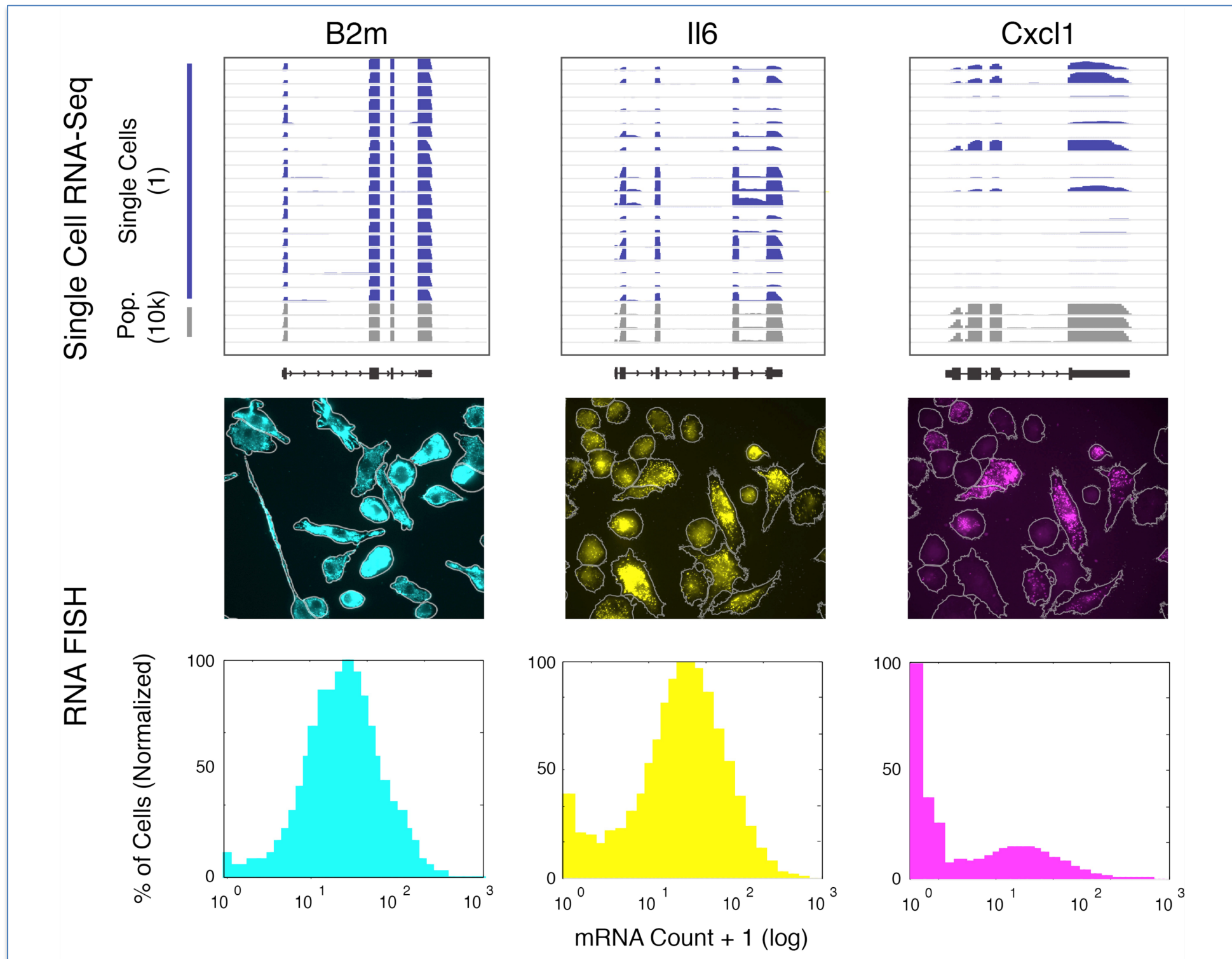
Most Variable Genes



Screenshots are log scale in IGV

Housekeeping & ribosomal genes are among the least variable, while immune response elements are among the most variable.

RNA FISH Validates Single-Cell RNA-Seq



Key single cell results can be validated using amplification-free methods.

Quality Control

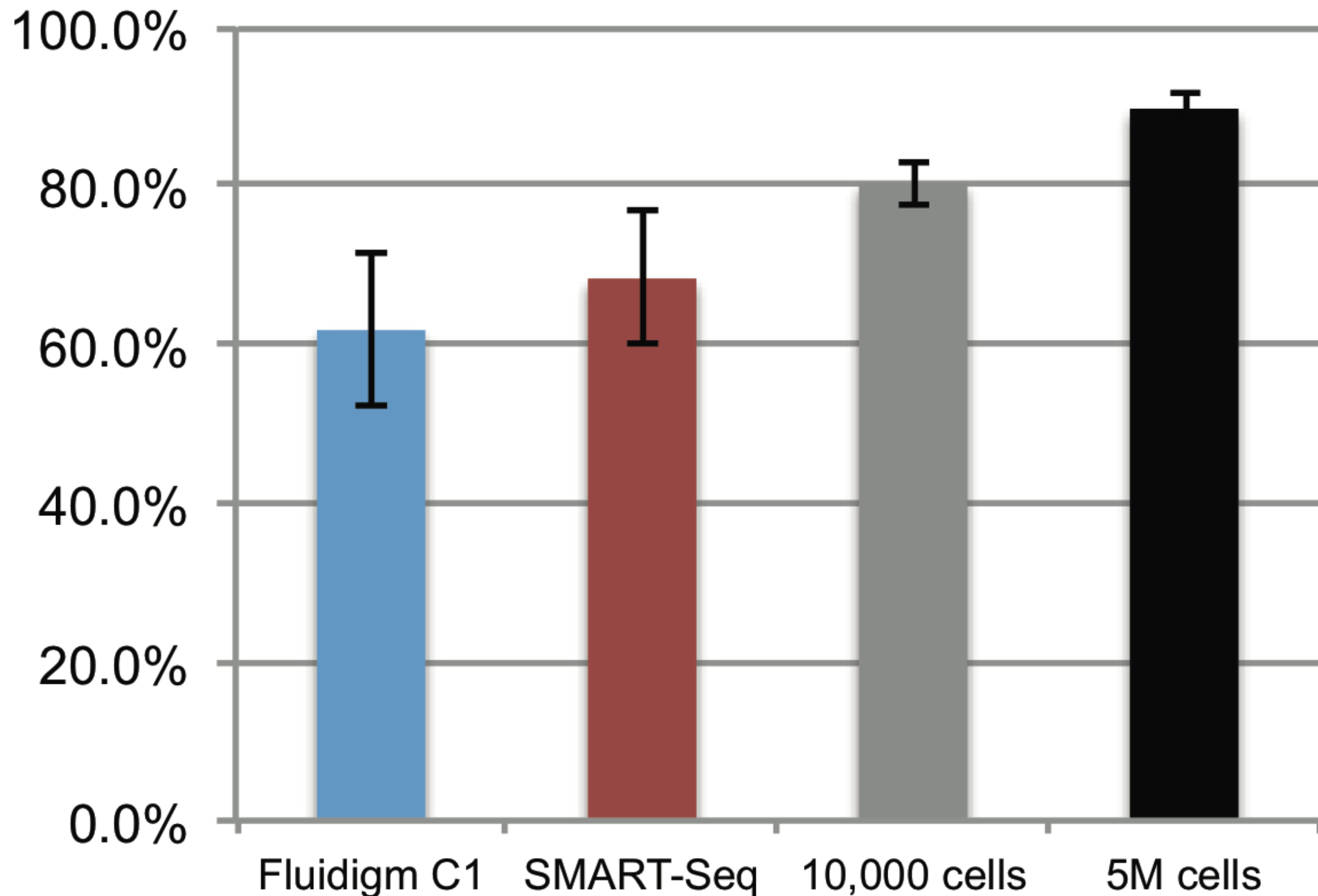
Standard metrics:

- Alignment rates
- Transcript coverage
- rRNA contamination
- ~~Duplication rate~~

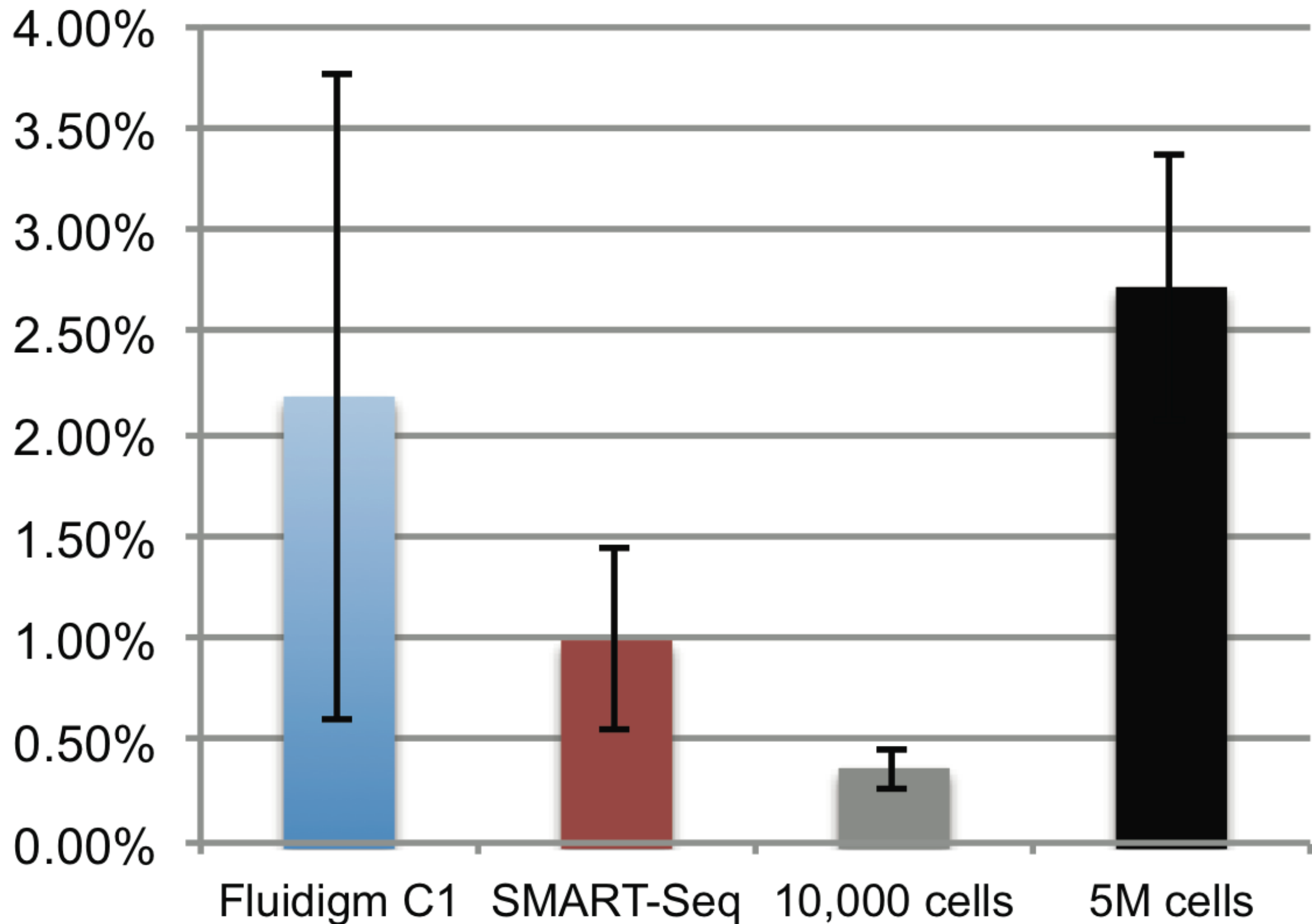
Single cell metrics:

- Complexity
- Probability of transcript detection
 - Relationship with gene properties

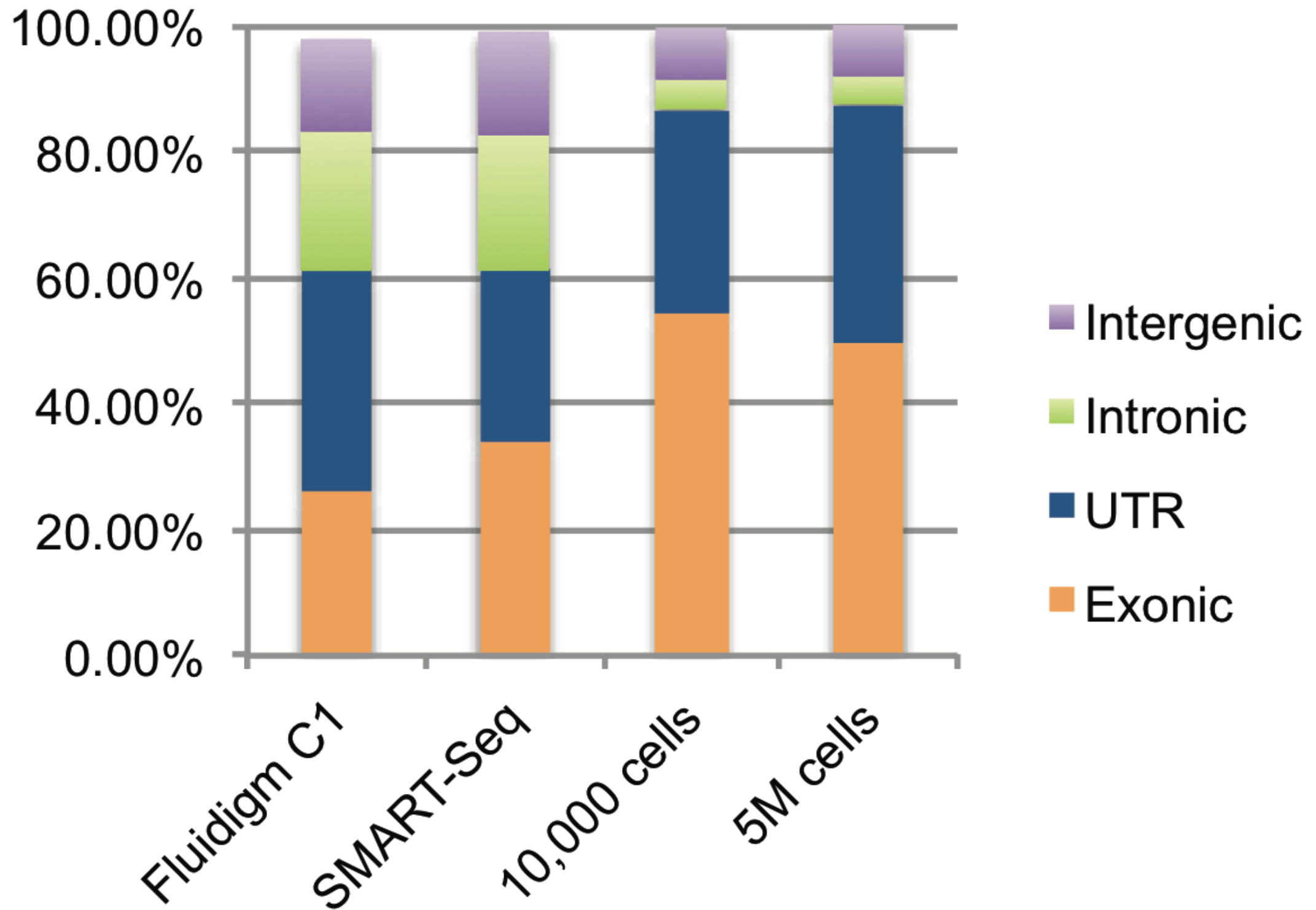
Genomic alignment rates



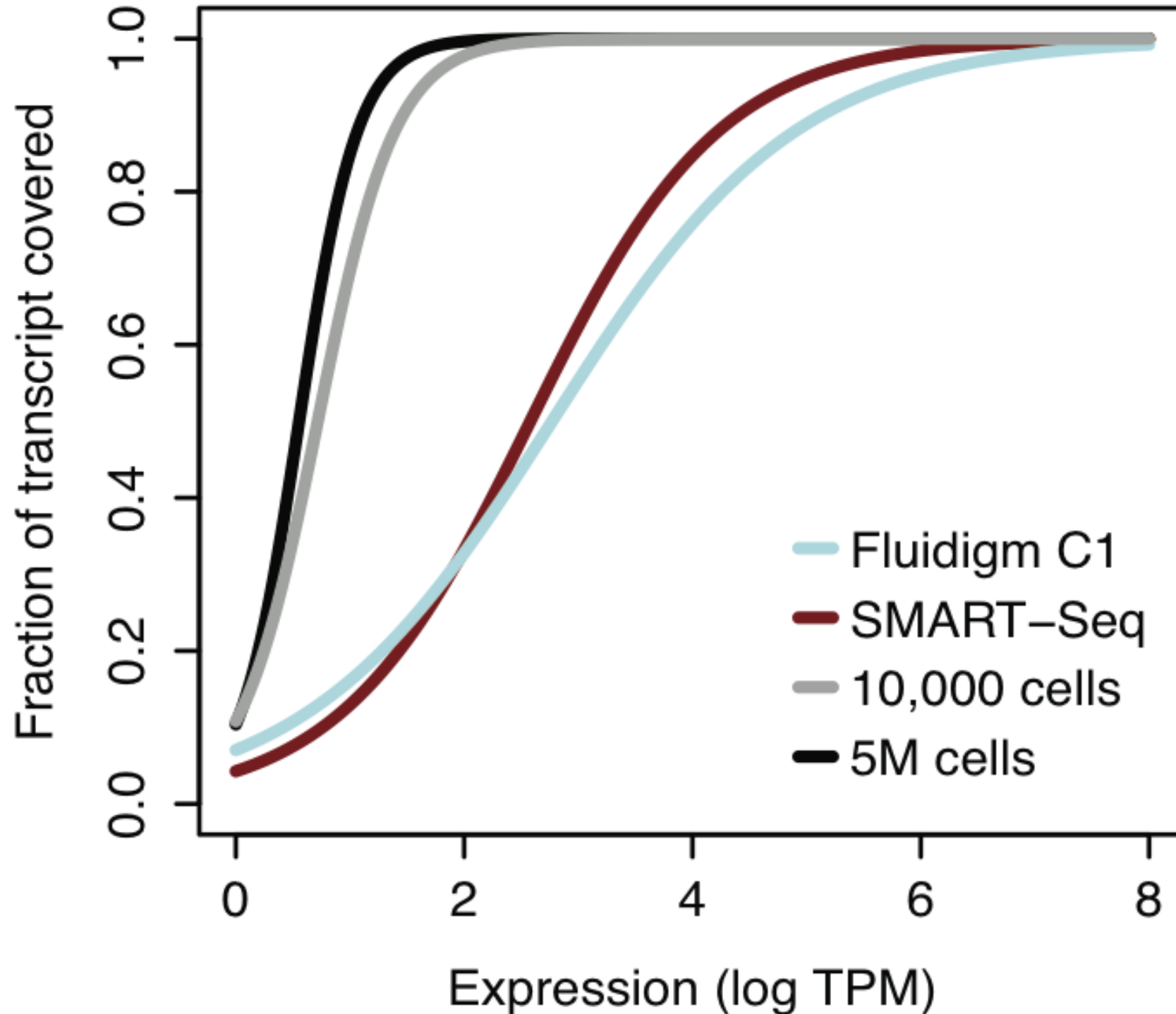
rRNA alignment rates



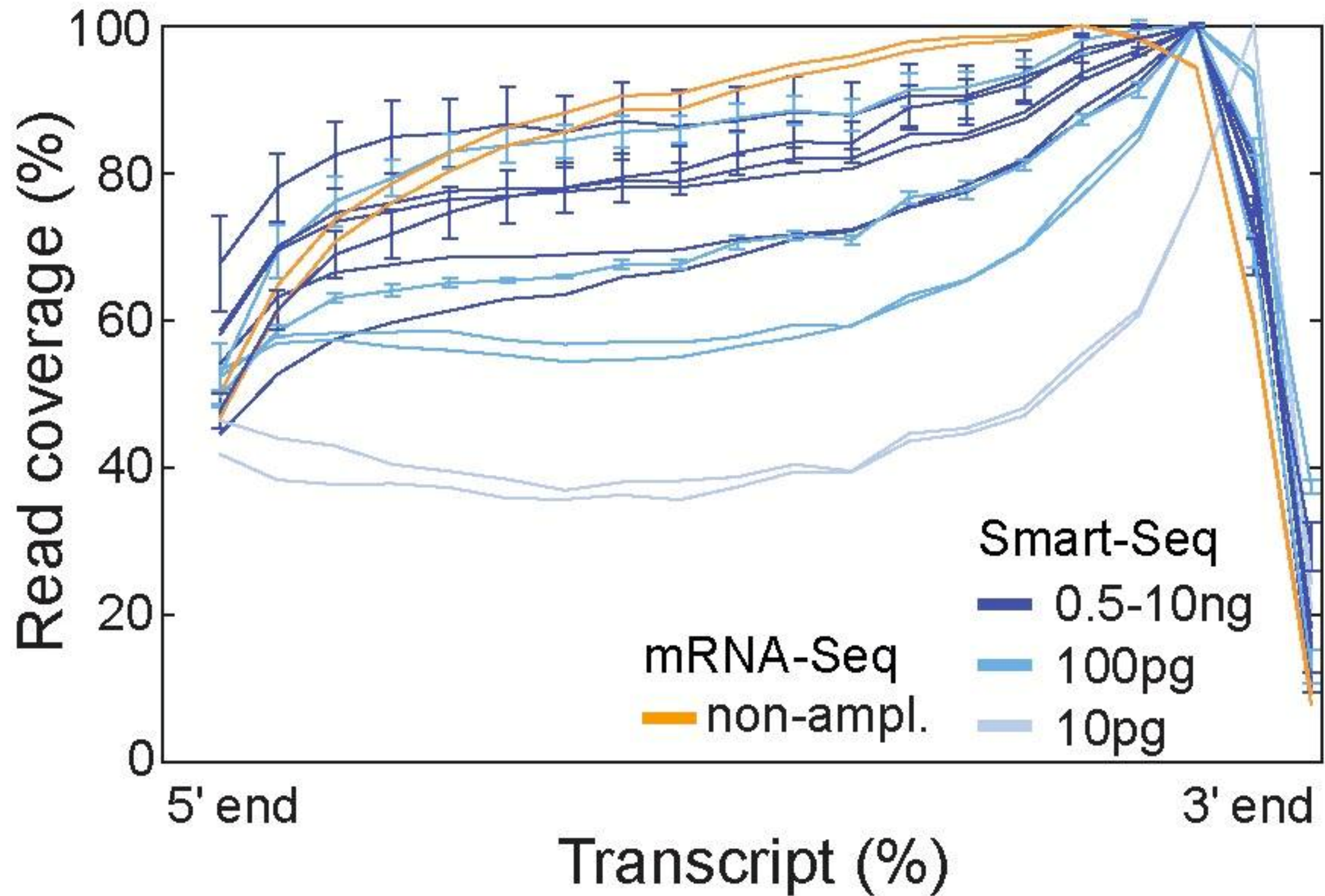
Where do the reads map?



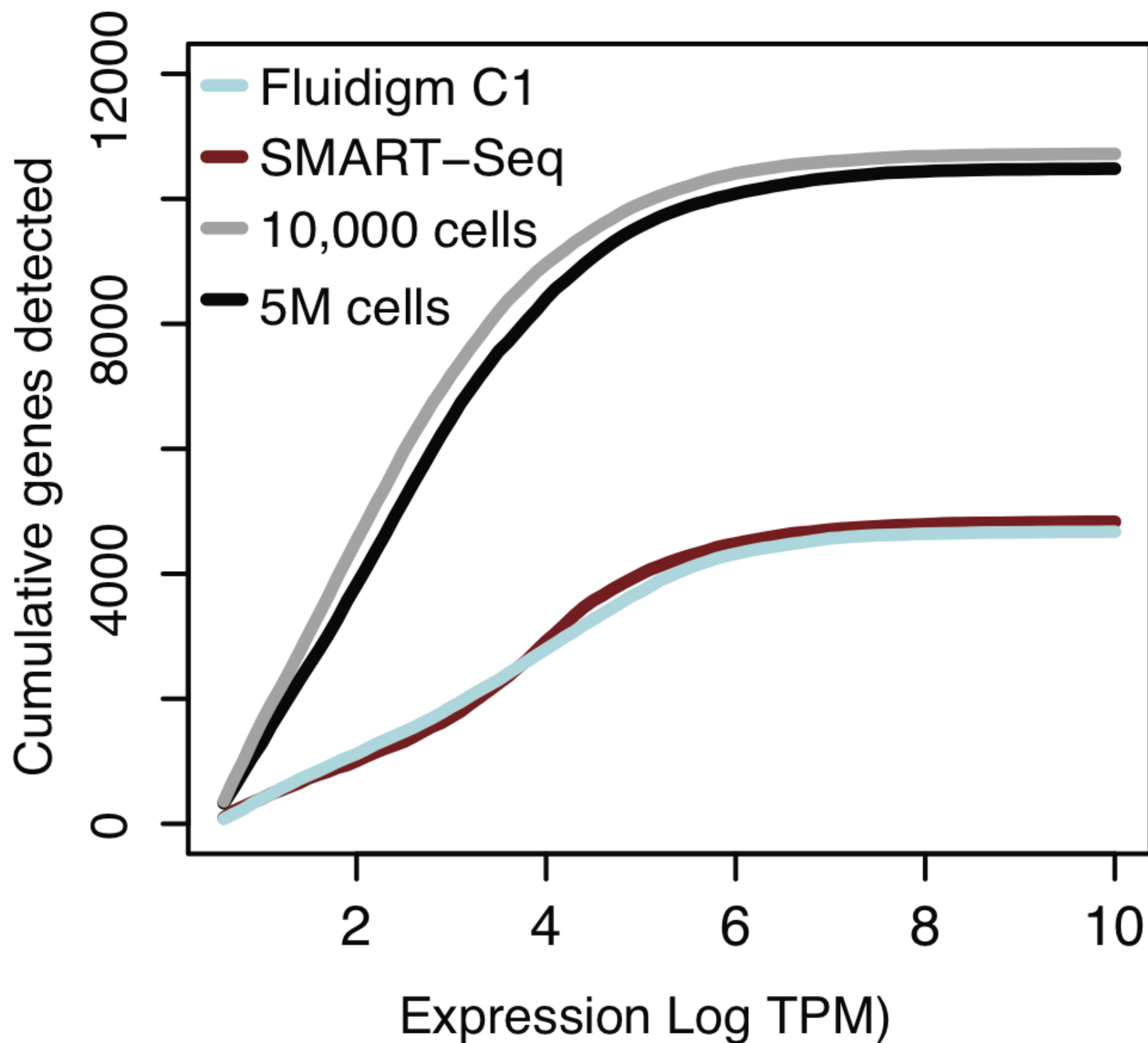
Transcript coverage



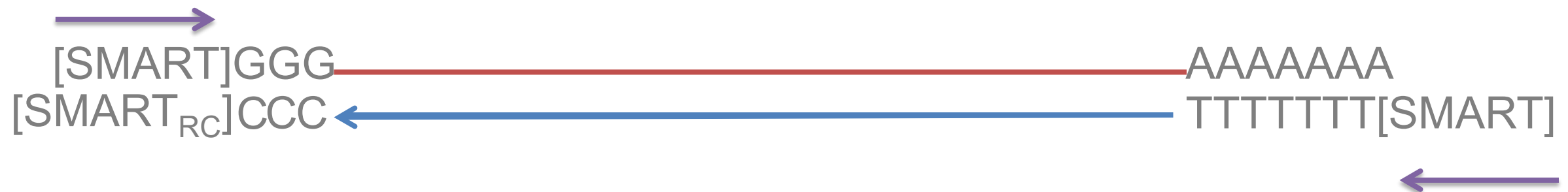
Transcript coverage



Complexity

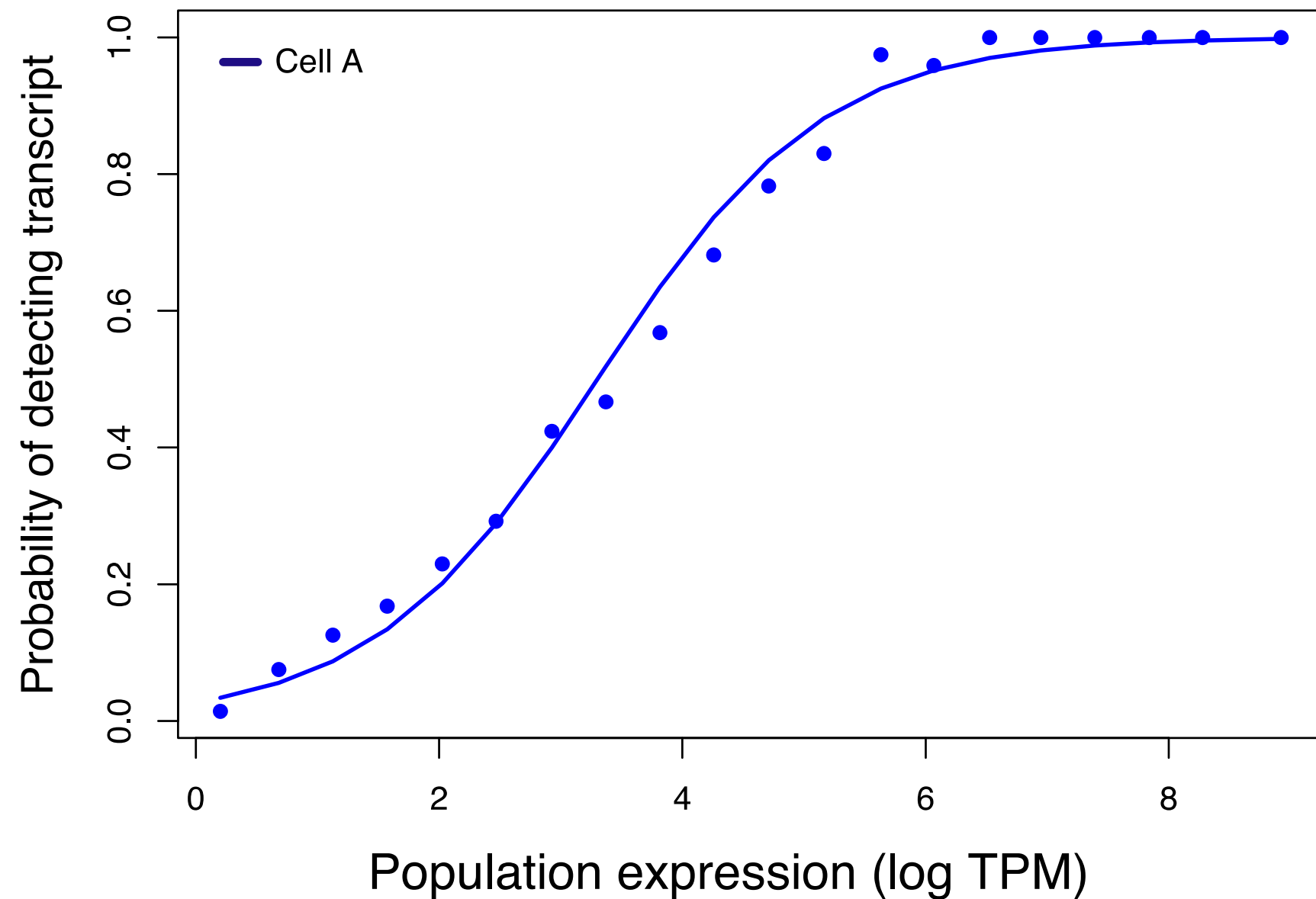


~~Duplication rate~~

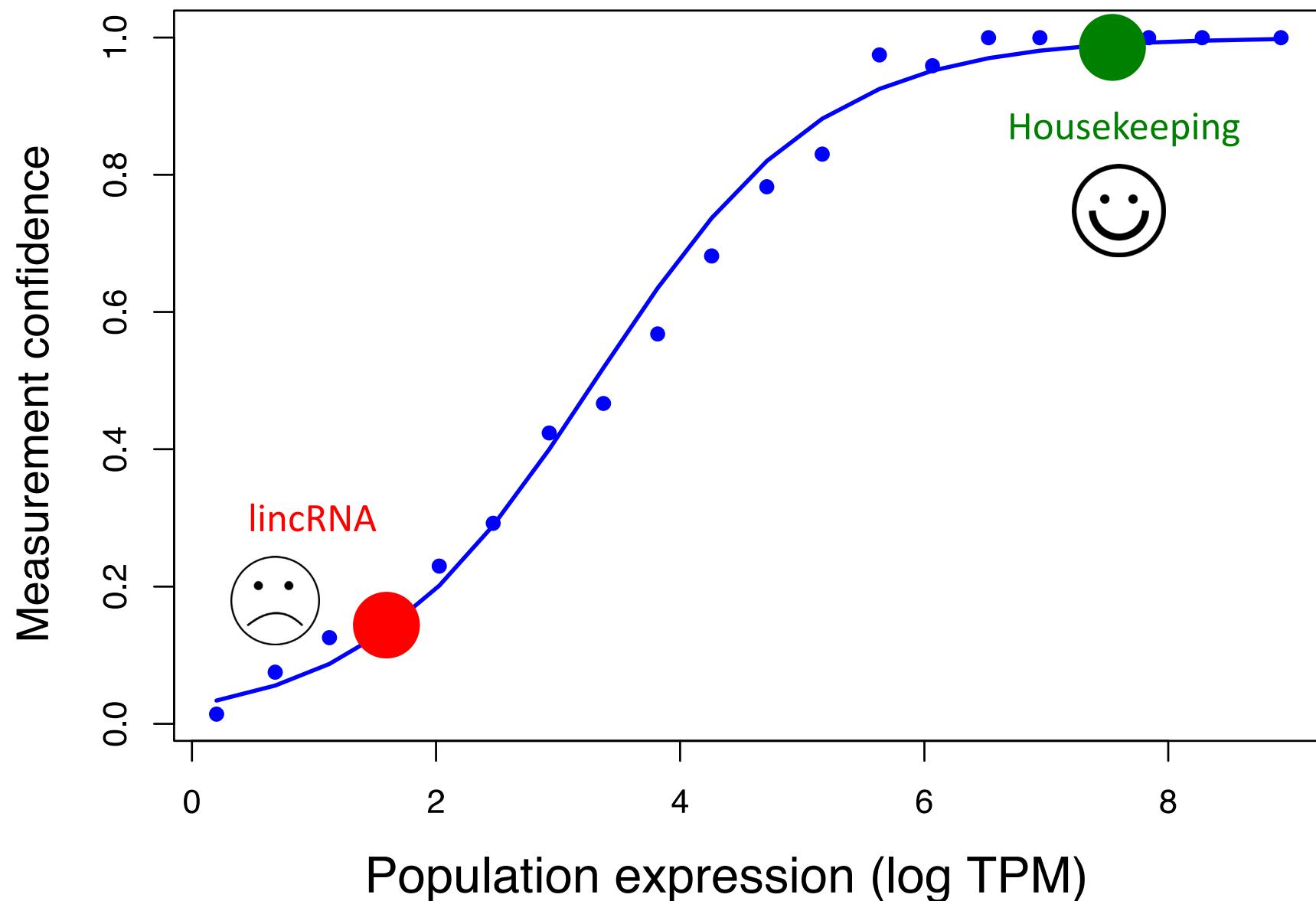


First round of full-length cDNA means that PCR duplicates cannot be identified

Technical noise is dependent on expression

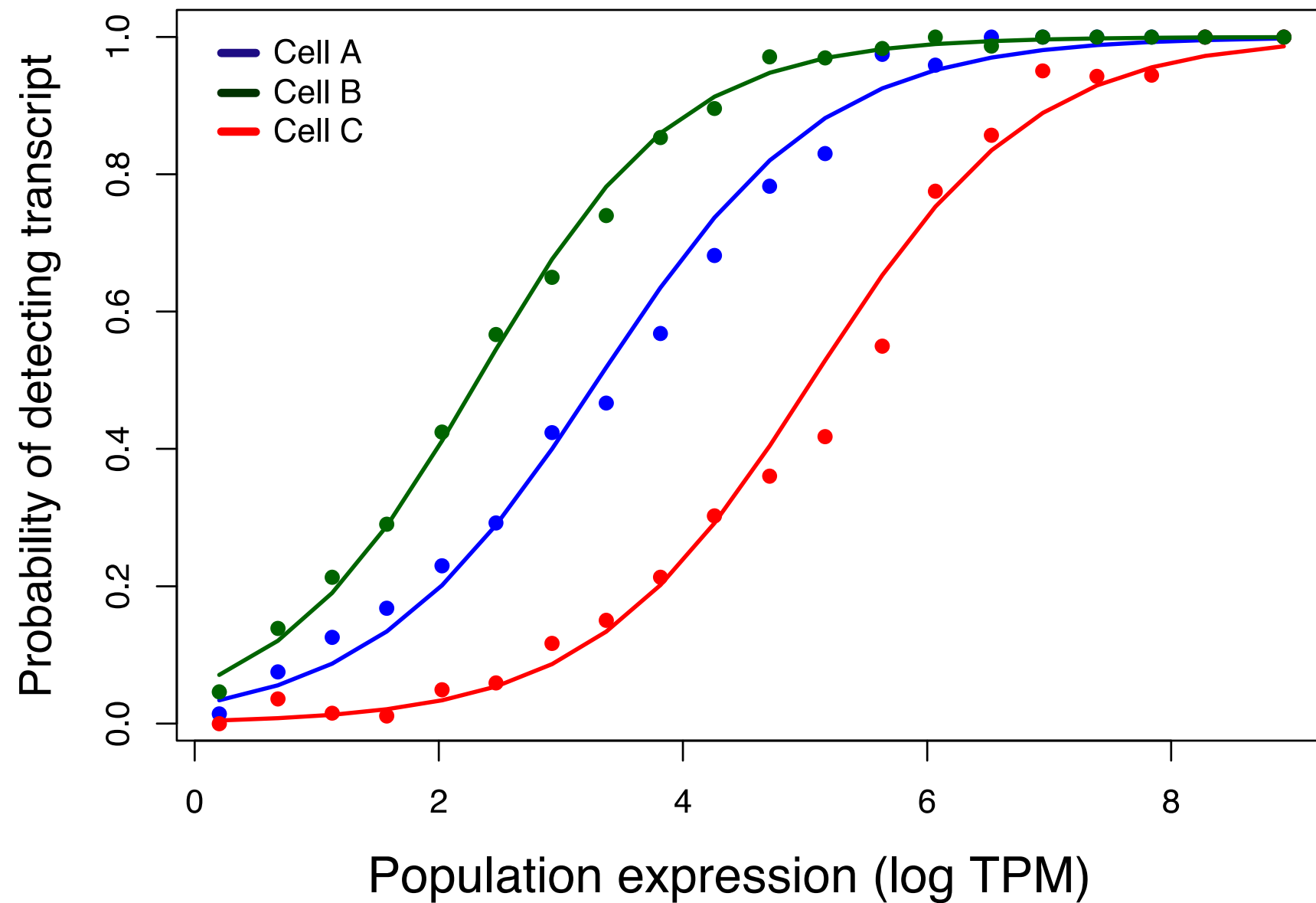


Calculating uncertainty in single cell data

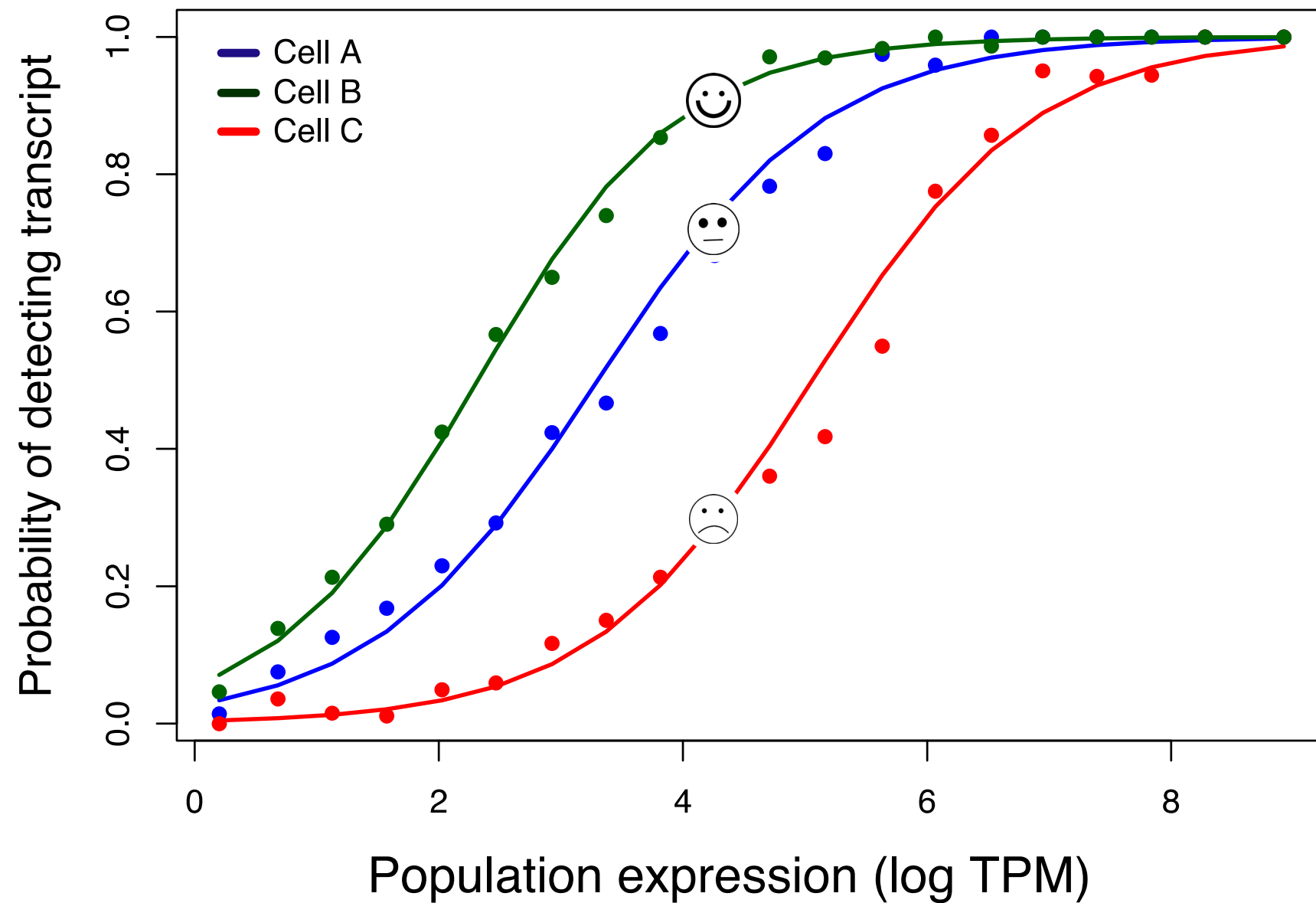


Expression level, gene length, GC content all influence technical noise

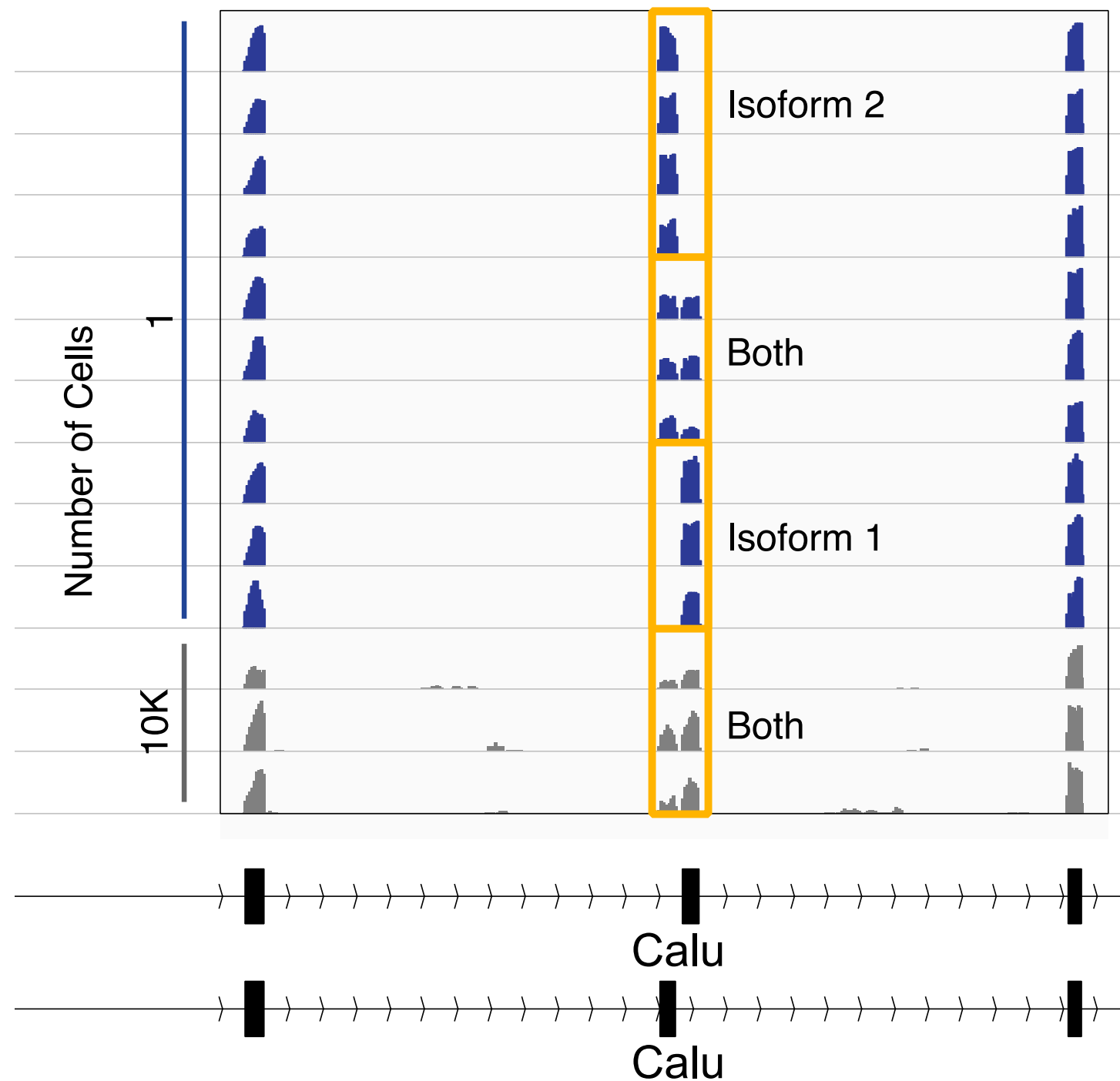
Cells have distinct technical qualities



Cells have distinct technical qualities



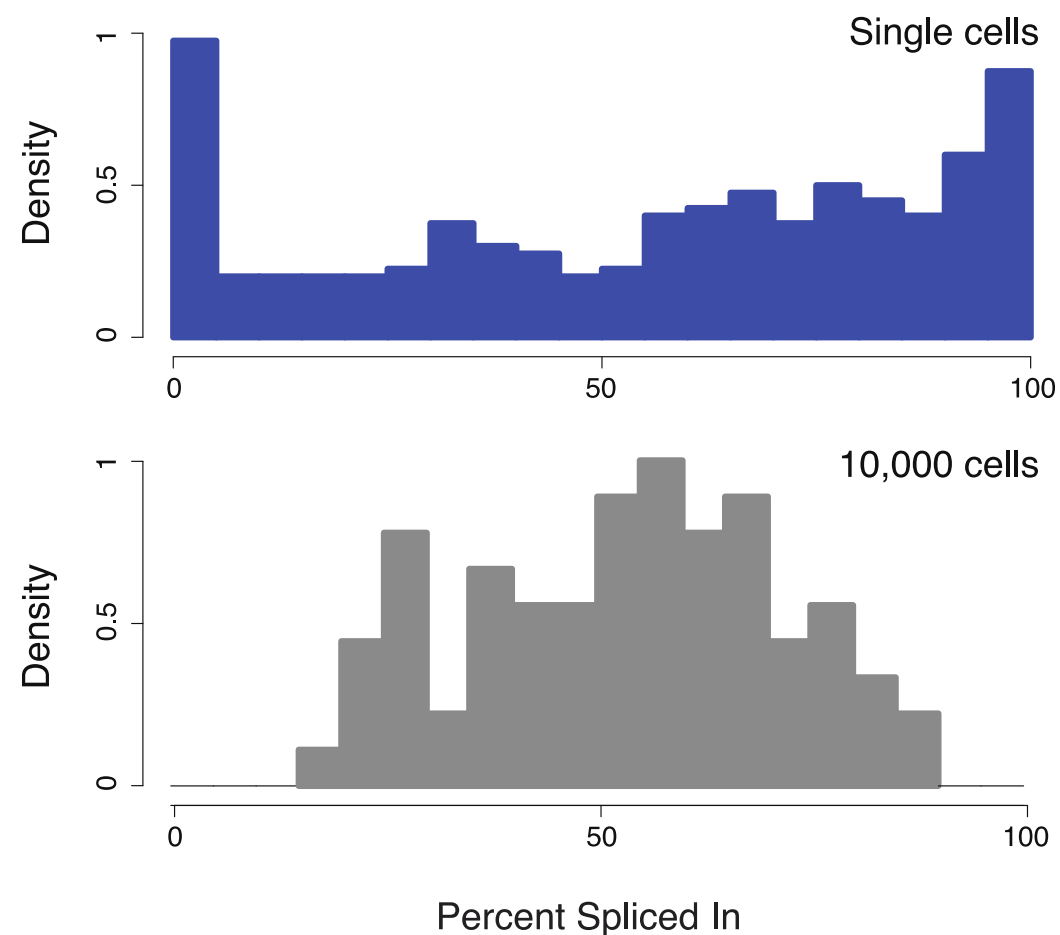
Two sources of noise in single cell data



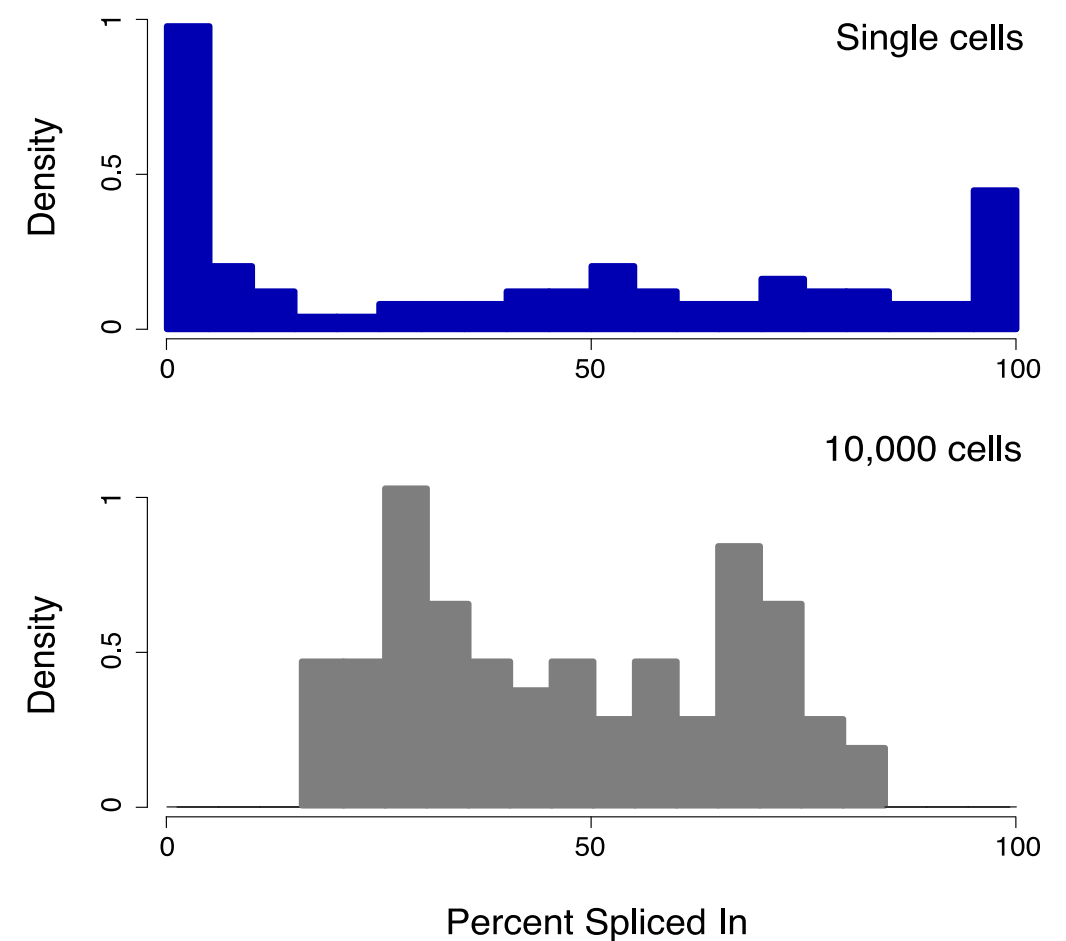
Both sampling noise, and PCR bias could contribute to this result

Two sources of noise in single cell data

TPM > 250



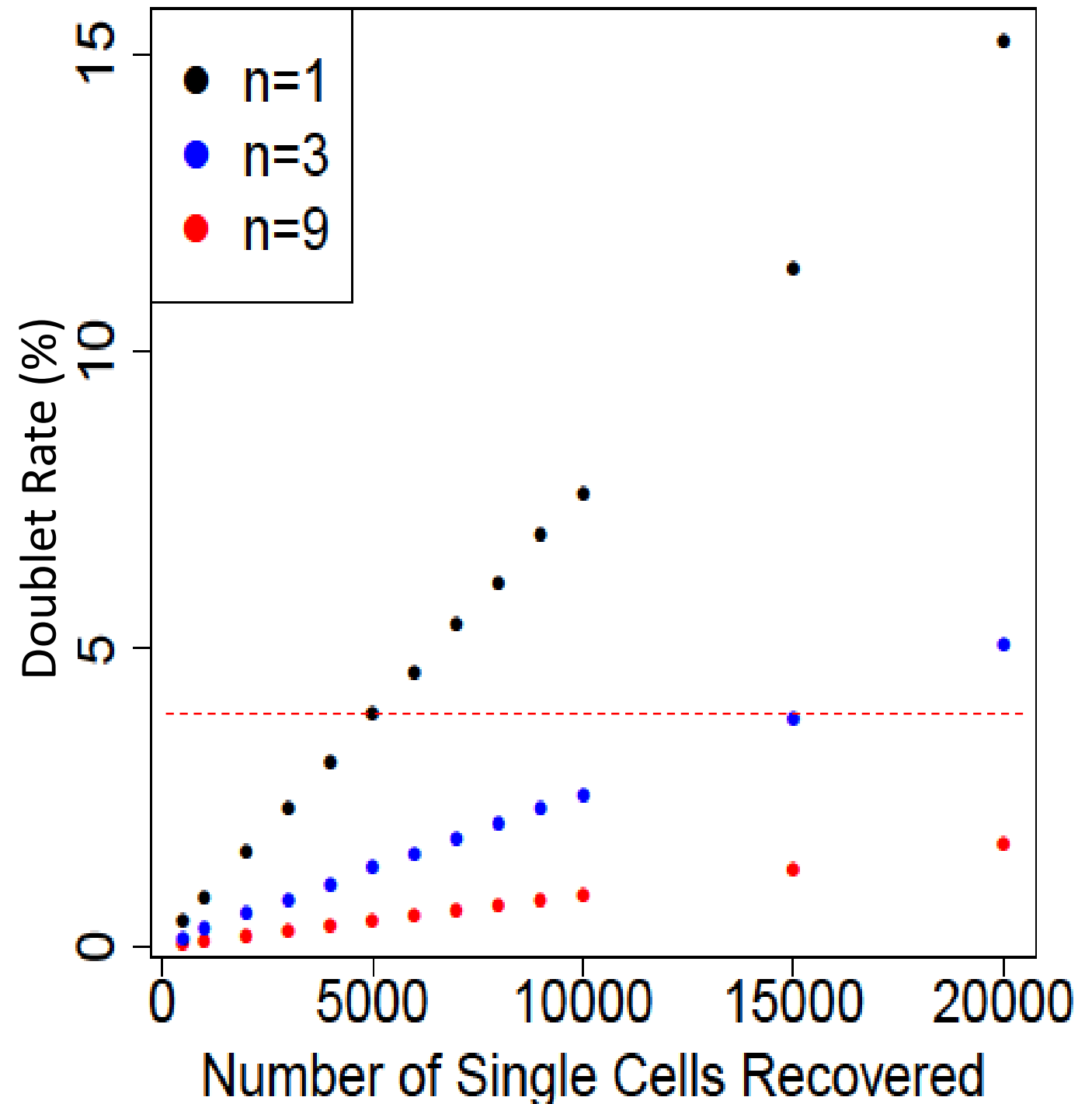
UMB > 15



One way to overcome this – focus on highly expressed genes

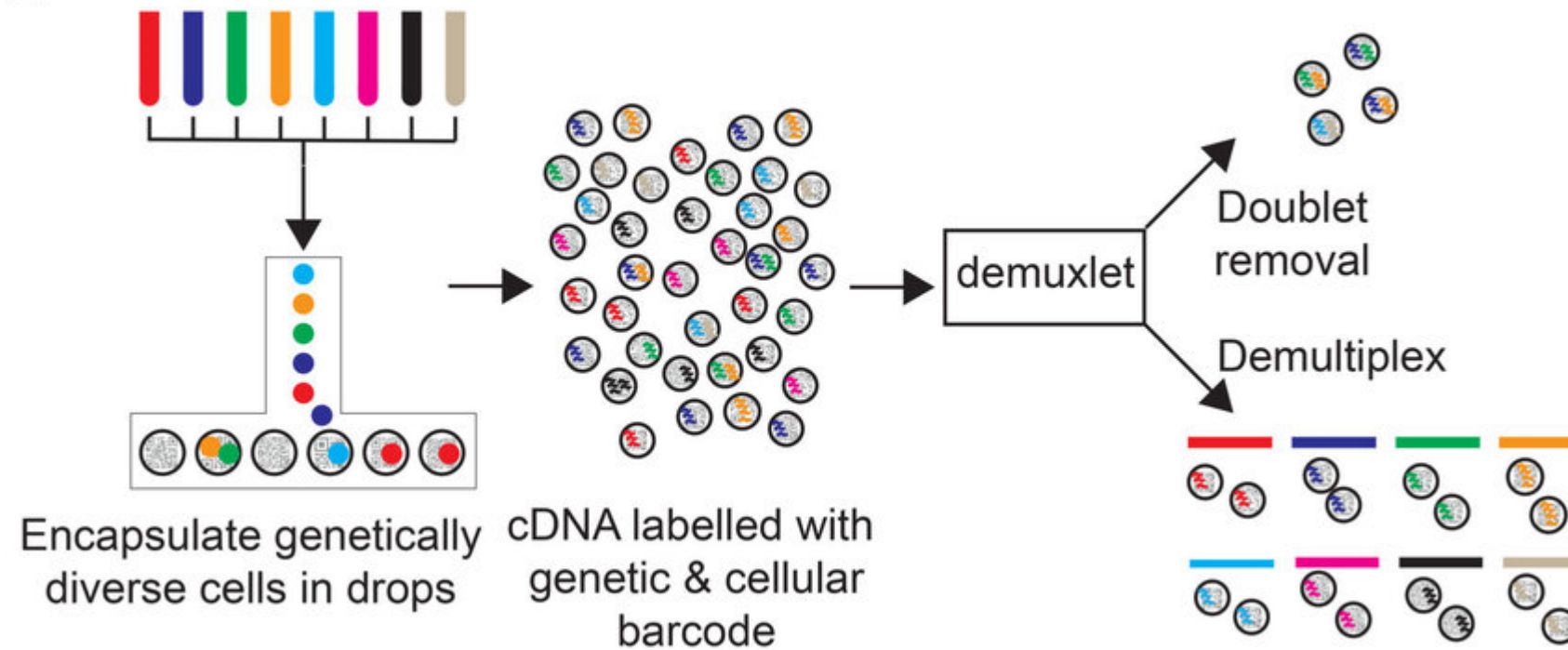
Doublets increase with numbers of cells profiled

- Prohibitive cost (~\$2,000 per sample)
- Prominent batch effects
- Limitations to the number of cells that can be captured
 - As more cells are captured more doublets appear in the data



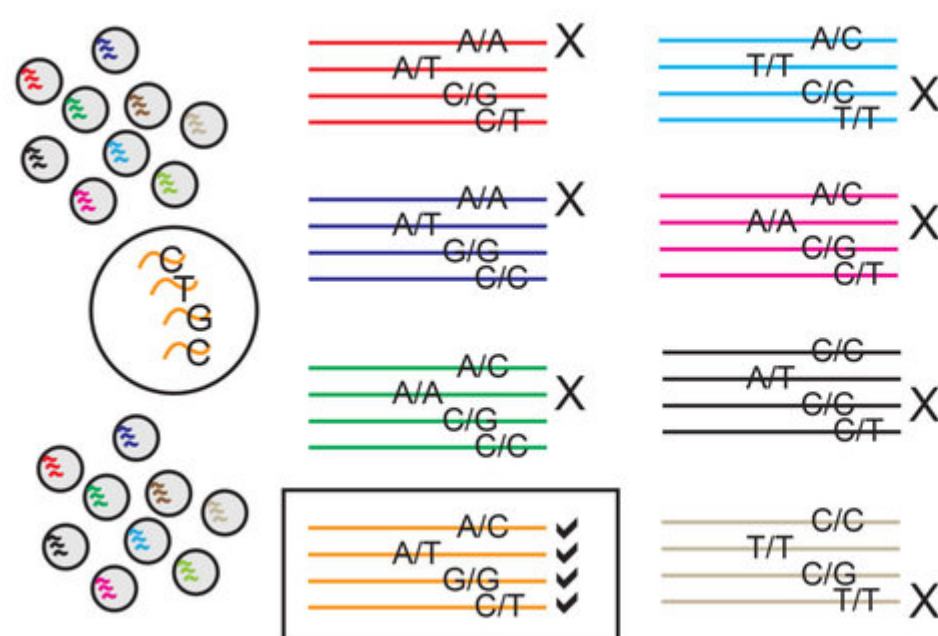
Recognizing doublets with cellular barcodes

A.



Recognizing doublets with genetic 'barcodes'

B.

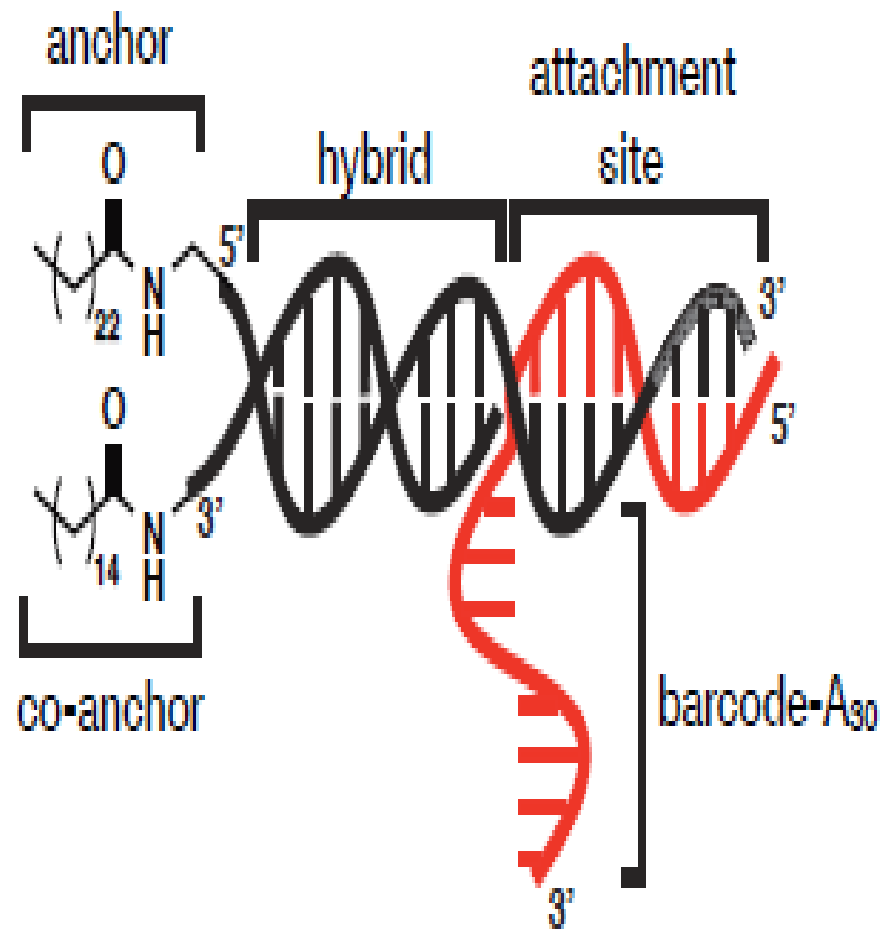


C.

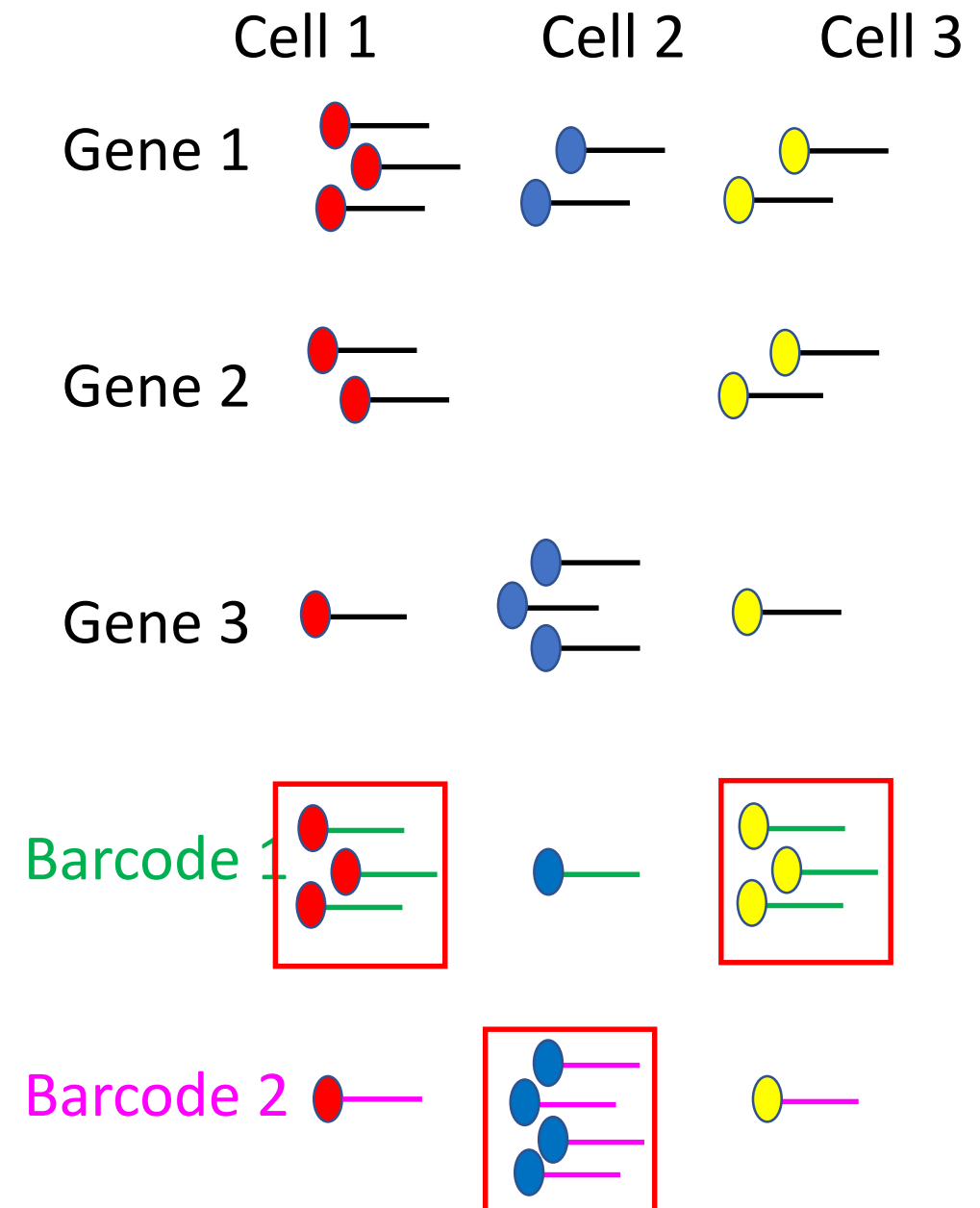


Multiplexing Using Lipid-Tagged Indices

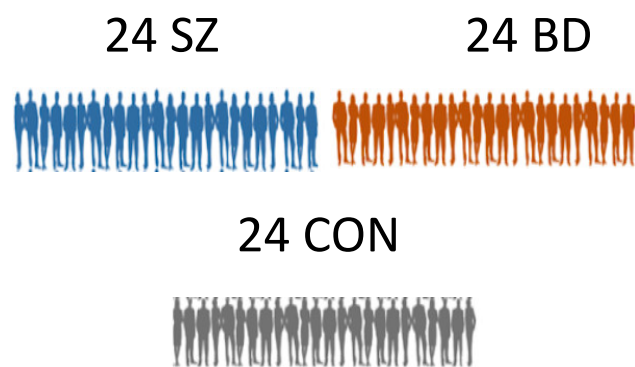
MULTI-Seq



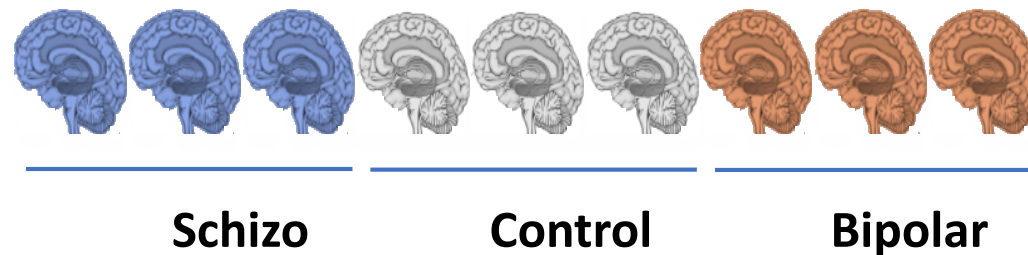
McGinnis et al. Nature Methods 2019



24 Sz + 24 BD + 24 Ctrl = Multiplexing (batches of 9)

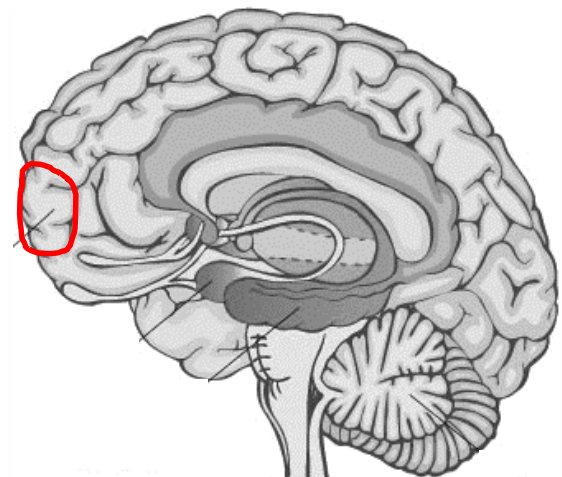


8 batches, 9 individuals each, mix phenotypes

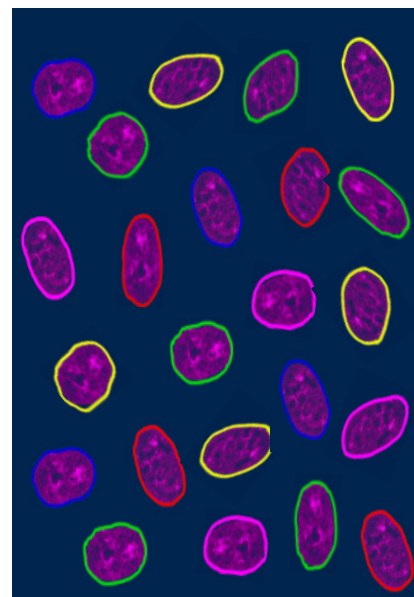


Isolate nuclei and label each sample
with a unique oligo hashtag

Pool all material into a single nuclear suspension



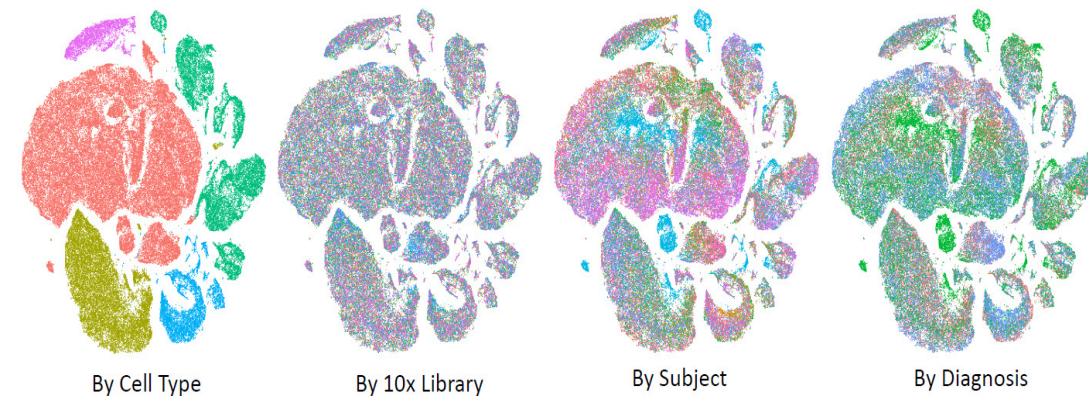
72 individuals total



Pooled nuclear suspension
spread across eight 10x channels



Eight 10X libraries from each batch of 9
(64 total libraries, 72 individuals)

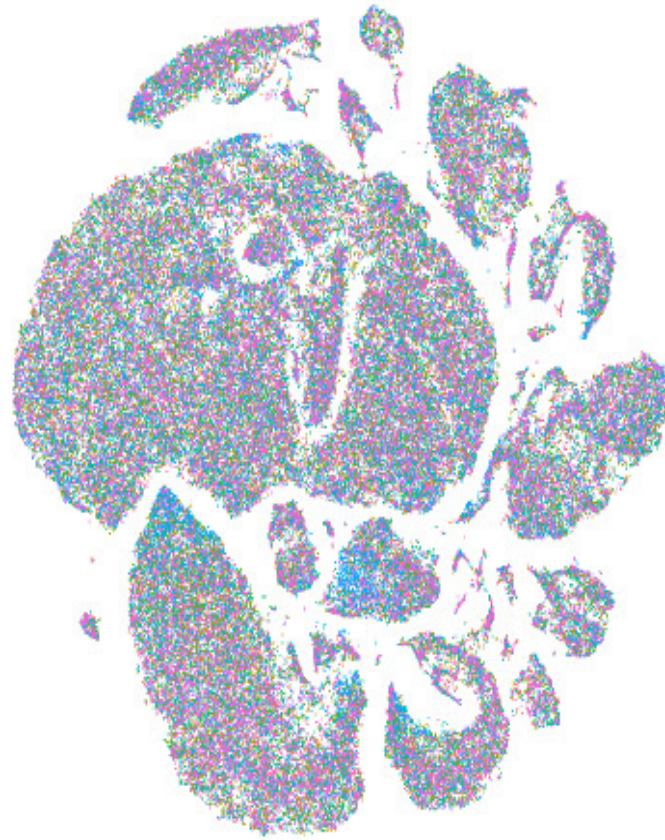


Integrate all libraries and batches into one
dataset of 800,000 single cells

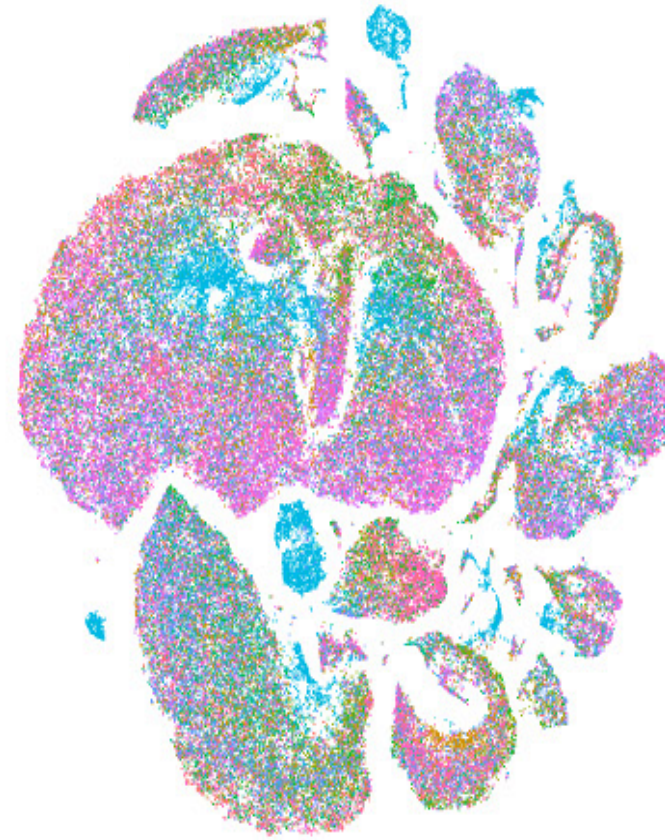
Multiplexing helps with doublets, and with batch correction



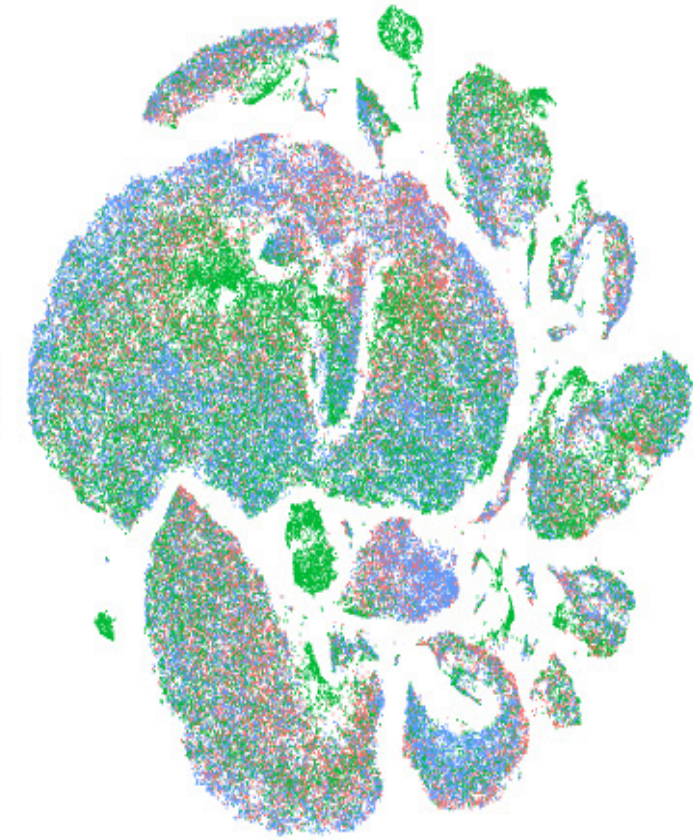
By Cell Type



By 10x Library



By Subject



By Diagnosis

Brad Ruzicka



Today: Deep Learning for Single-cell Genomics

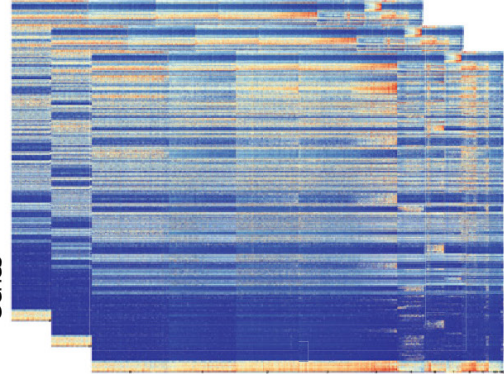
1. Why single cells, traditional approaches, scRNA-seq
2. Scaling up single-cell technologies: evolution of scRNA-seq
3. Dealing with noise, doublets, and other scRNA-seq issues
4. Computational challenges in scRNA-seq data analysis
5. Deep learning methods for scRNA-seq data analysis
6. Beyond scRNA-seq: scATAC-seq, multi-omics

4. Computational challenges in scRNA-seq data analysis

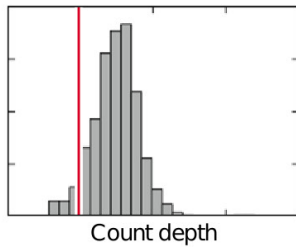
PRE-PROCESSING

Raw data processing

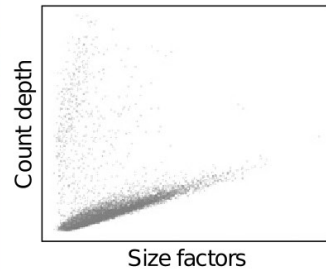
Cells



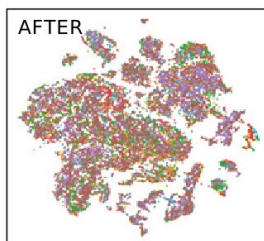
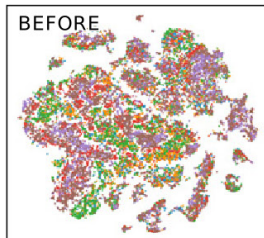
Quality control



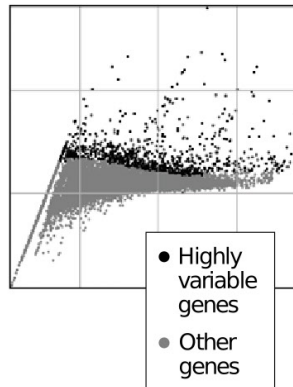
Normalization



Data correction (e.g. batch)



Feature selection



Visualization



DOWNSTREAM ANALYSIS

Cell-level analyses

Clustering

1

Marker identification

Cluster annotation

Multi-resolution analysis

Tuft cells
EEC
TA
EP (early)
EP (late)
Goblet cells
Paneth cells
Stem cells
Enterocytes

Clustering similar genes

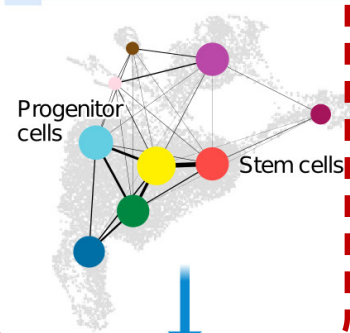
Dimensionality Reduction

Missing data imputation

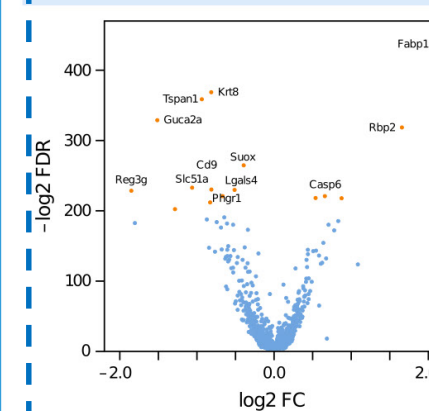
Deconvolution of bulk RNA-seq

Gene-level analyses

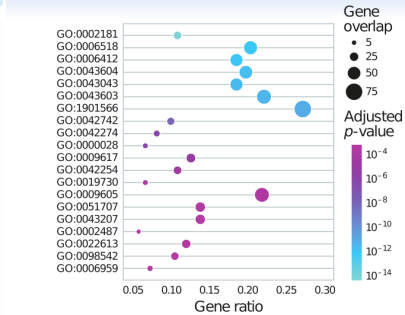
Trajectory inference



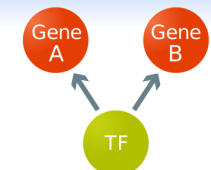
Differential expression



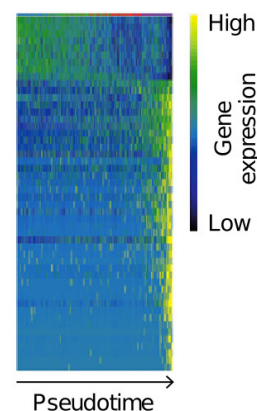
Gene set analysis



Gene regulatory networks



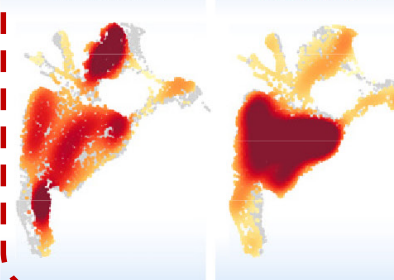
Gene dynamics



Compositional analysis

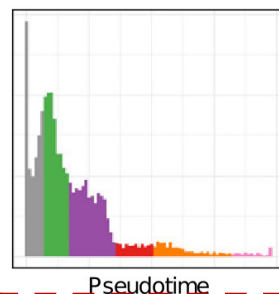
Condition 1

Condition 2



Multi-Omics analyses

Metastable states



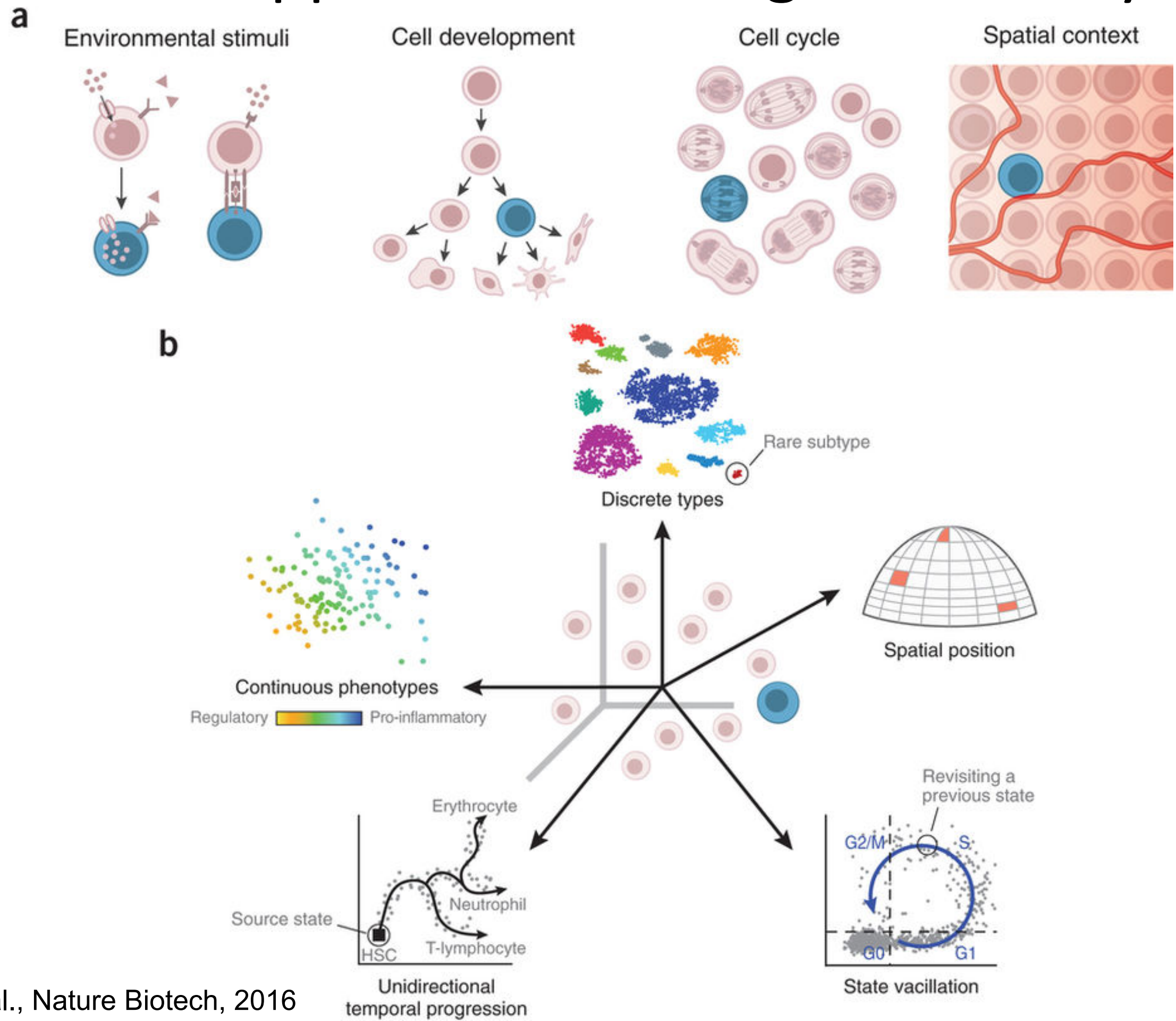
Luecken, M. D., & Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6).

Extracting biological insights from scRNA-seq data

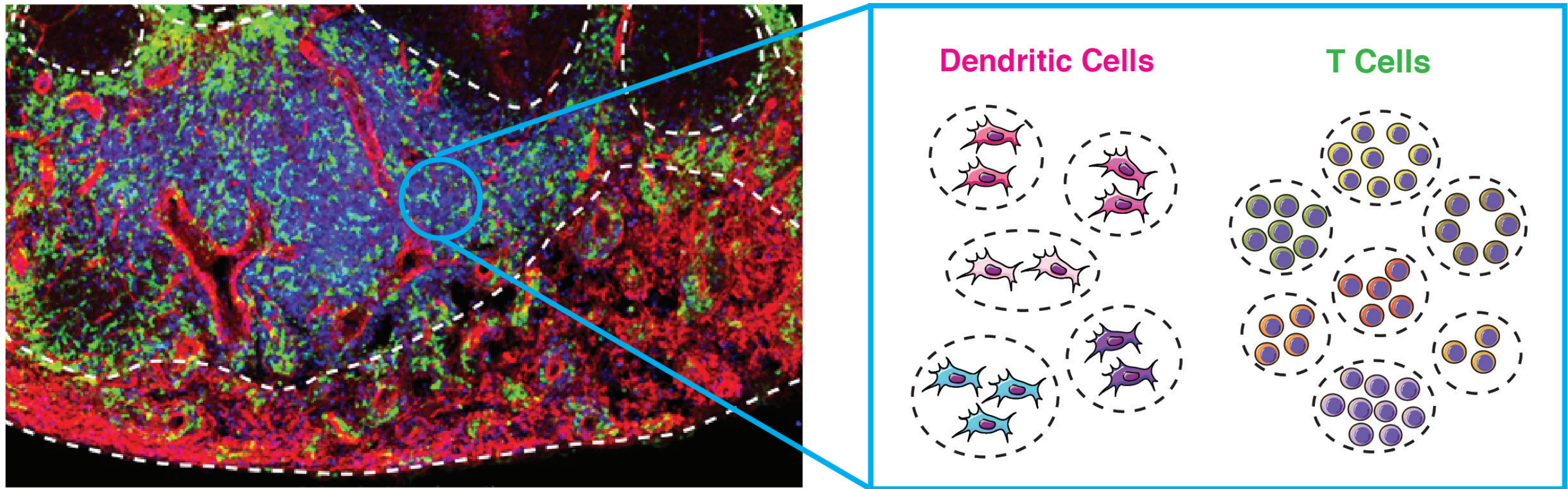
- Cell-to-cell correlation
- Gene-to-gene correlation
- Imputation of missing values
- Cellular trajectories and differentiation

Clustering similar cells

Methods + applications of single-cell analysis

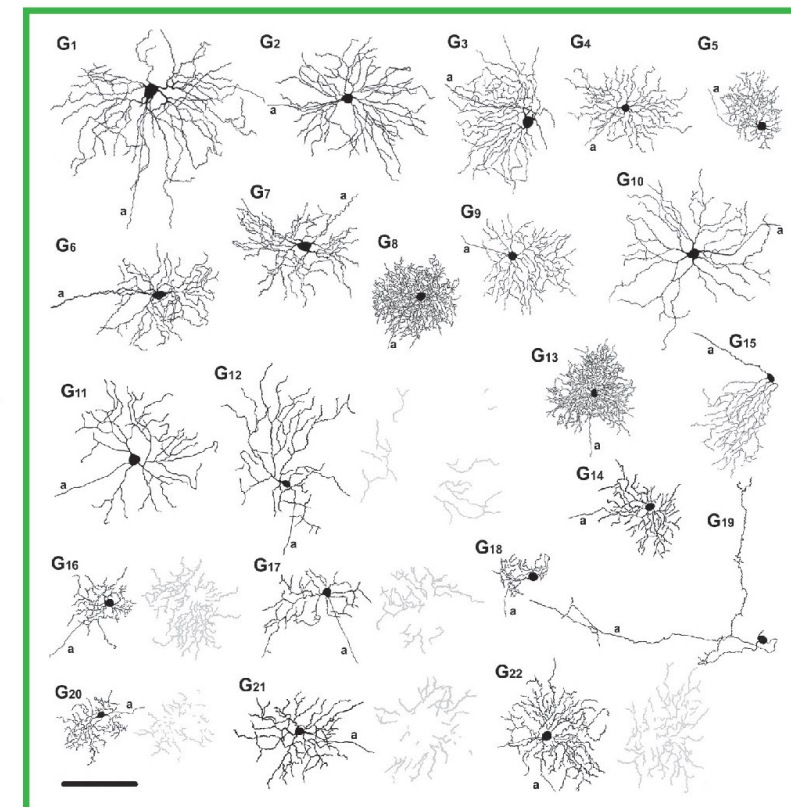
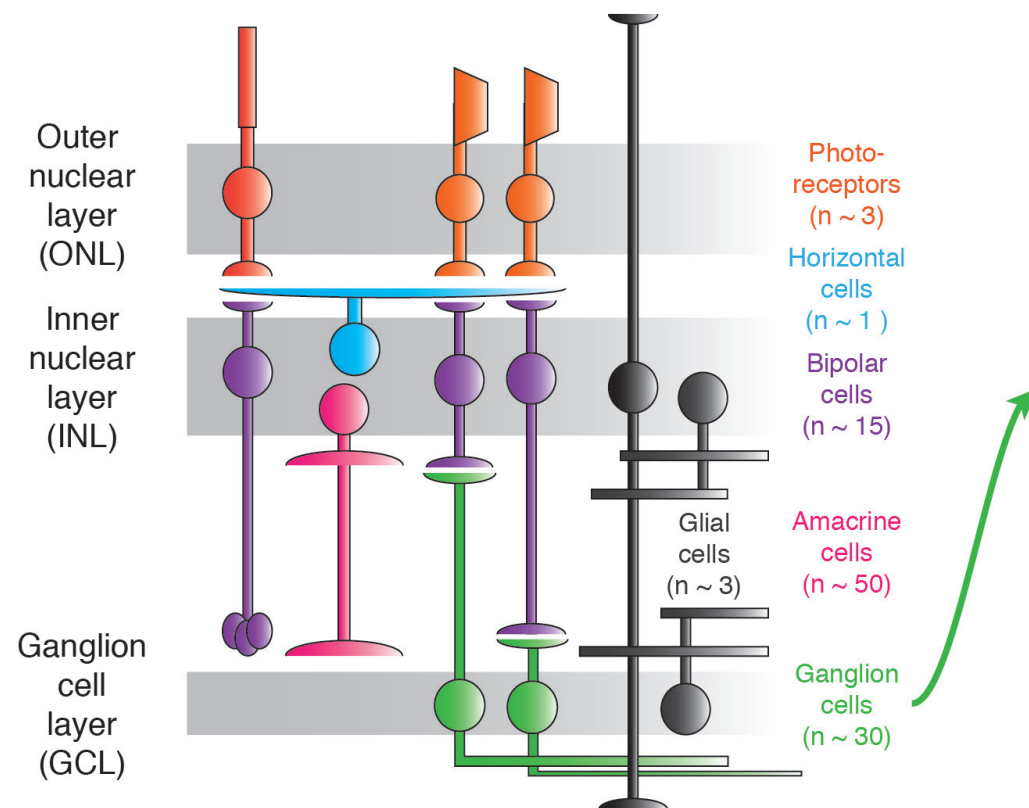
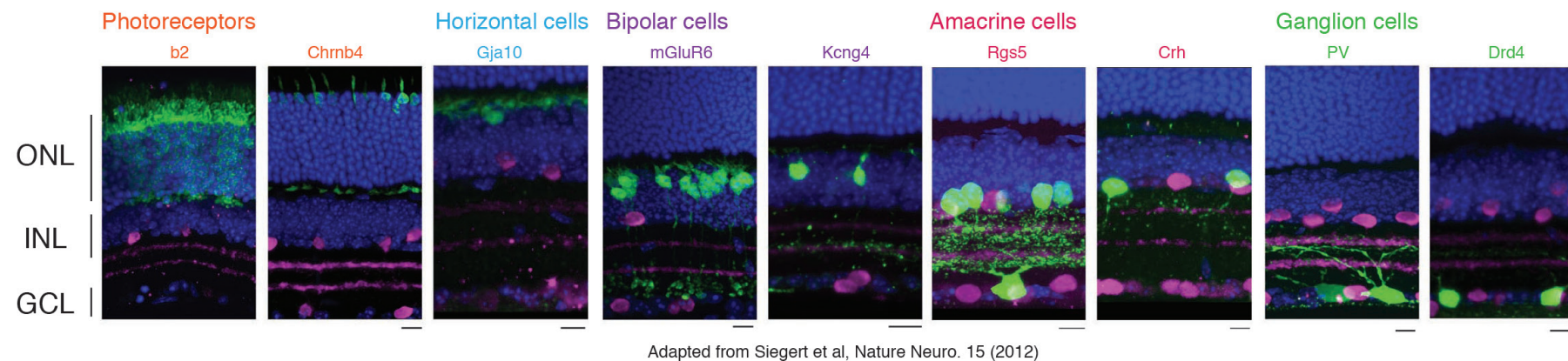


From Complex Tissues to individual cell types



Can we identify the different cell types/states in a complex tissue?

Brain Case Study: The Mouse Retina

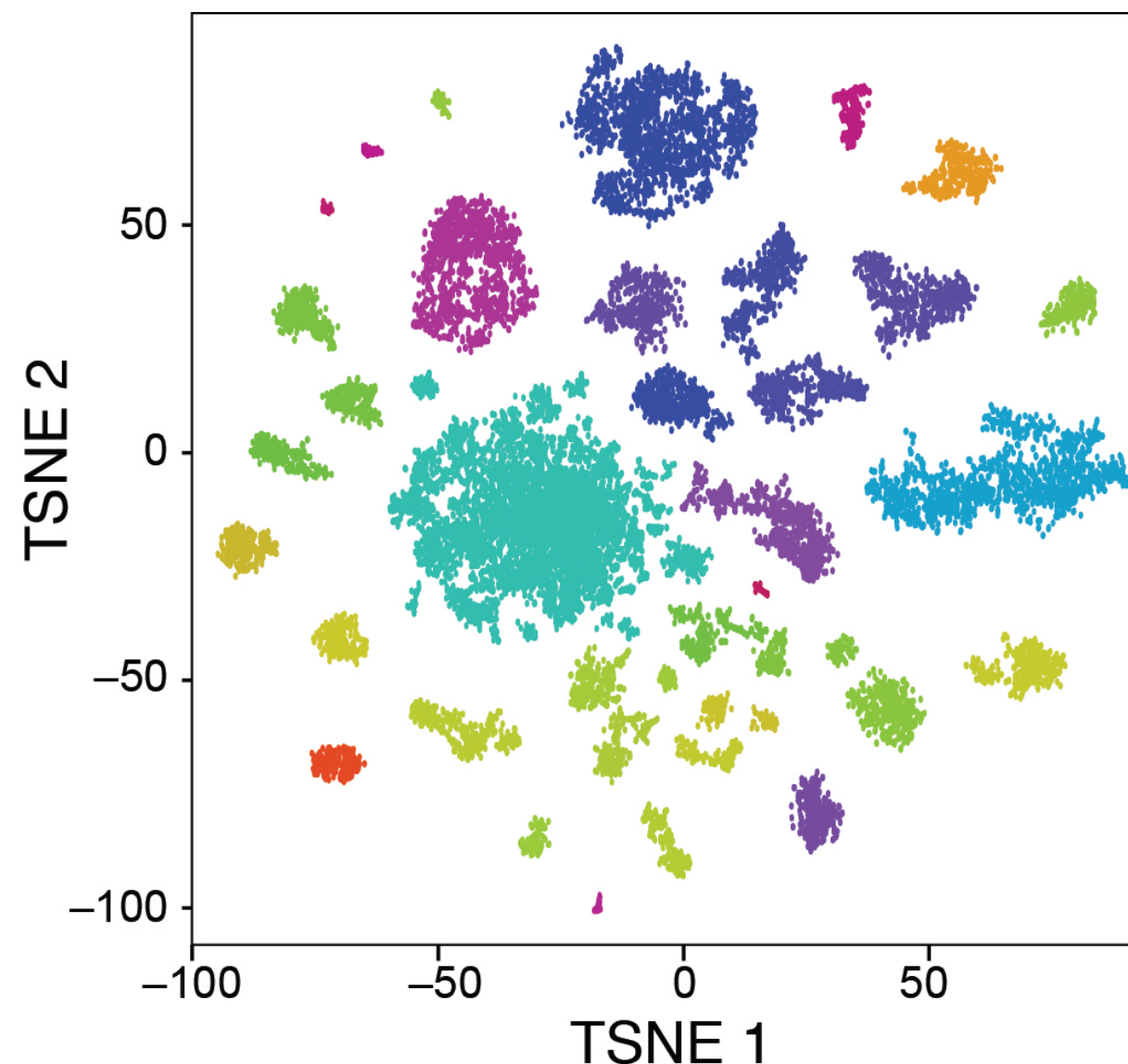


Adapted from Volgyi et al, J. Comparative Neurology 512 (2009)

~100 cell subtypes, only some with molecular markers

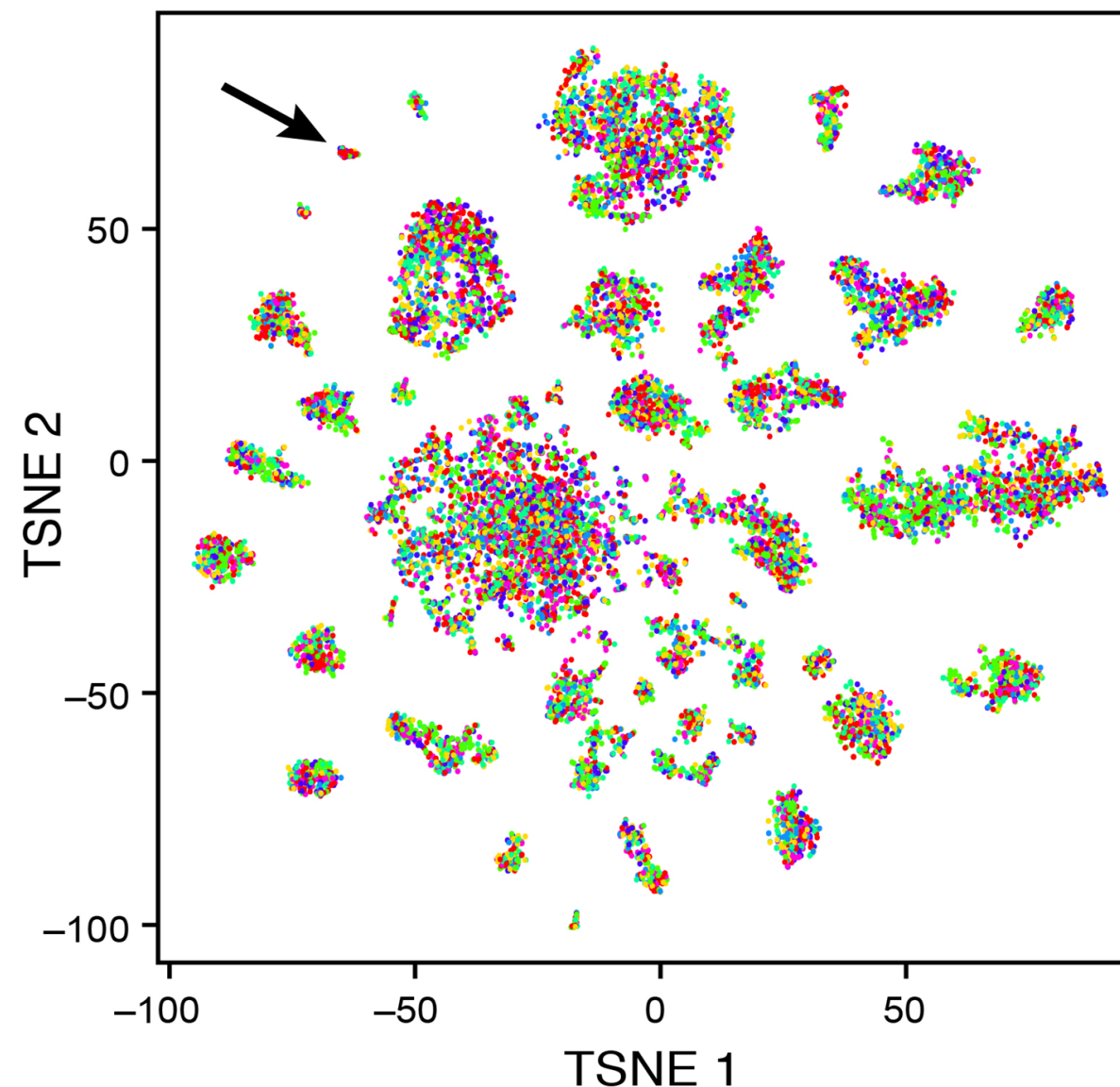
49,300 Retina Cells Grouped Into 39 Clusters

- **Drop-Seq:** 49,300 cells from dissociated mouse retina (P14) (~15k reads per cell)
- **Computational pipeline:** Select 25% best coverage cells, Dimensionality reduction (PCA+tSNE), Project remaining cells, Identify cell types (density clustering), Refine clusters (differential expression)

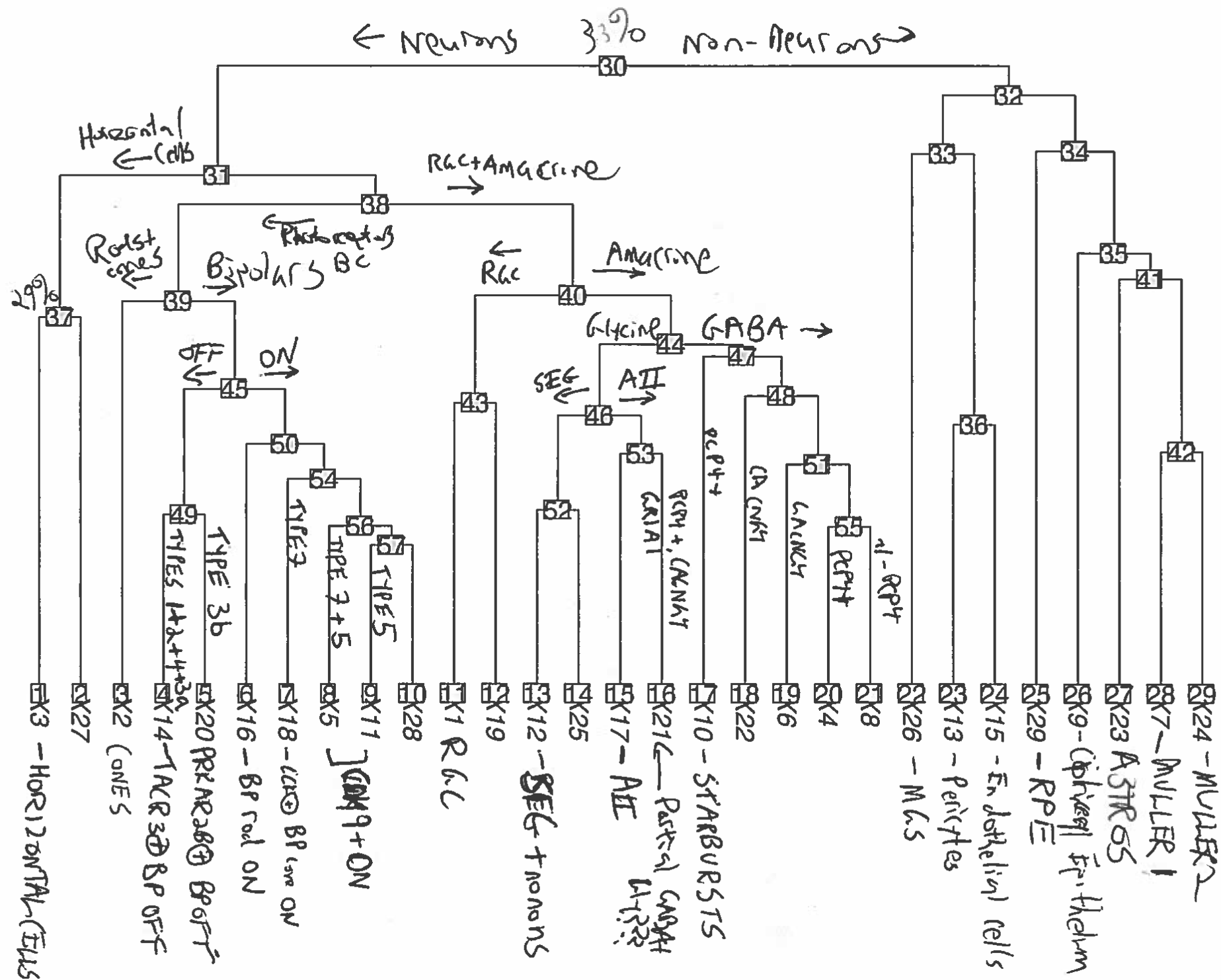


49,300 Retina Cells Grouped Into 39 Clusters

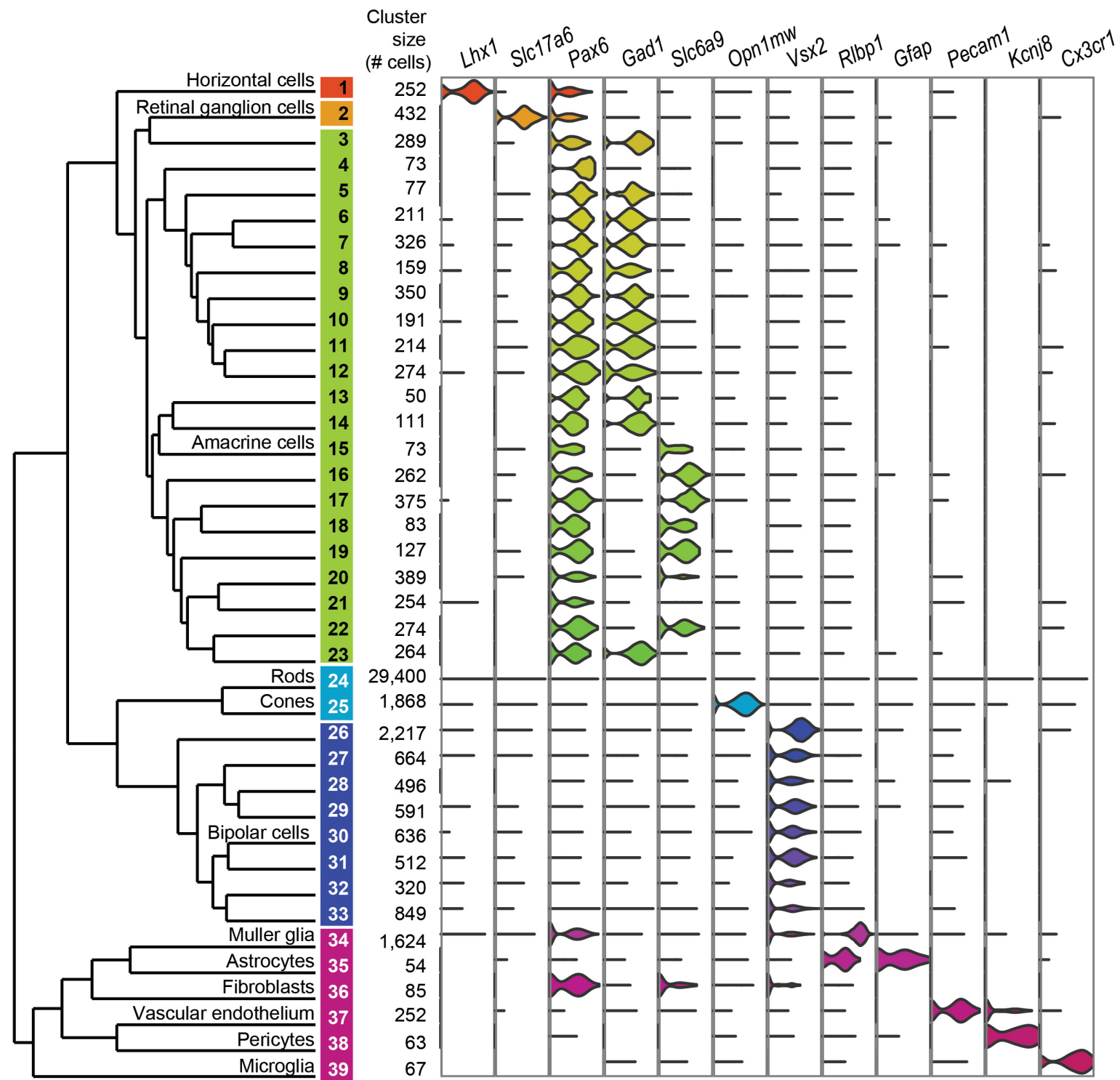
- **Drop-Seq:** 49,300 cells from dissociated mouse retina (P14) (~15k reads per cell)
- **Computational pipeline:** Select 25% best coverage cells, Dimensionality reduction (PCA+tSNE), Project remaining cells, Identify cell types (density clustering), Refine clusters (differential expression)



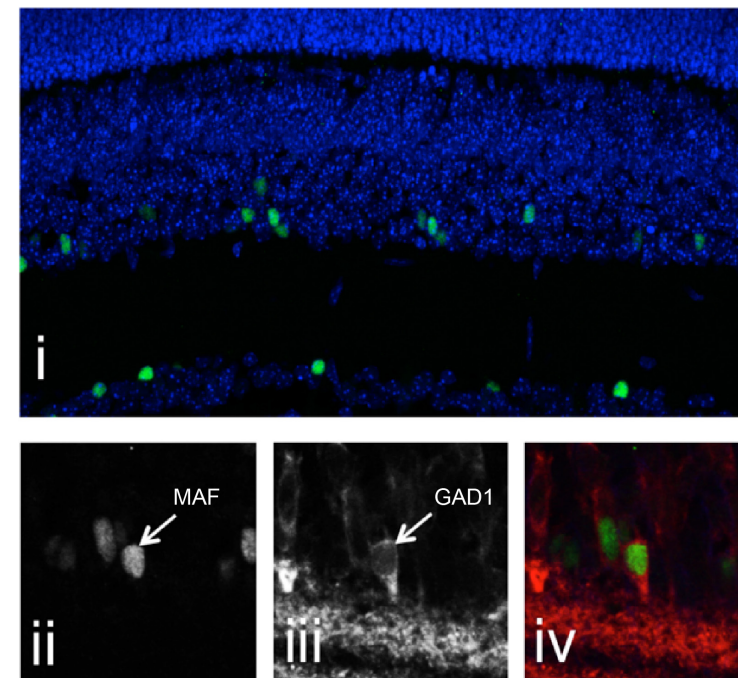
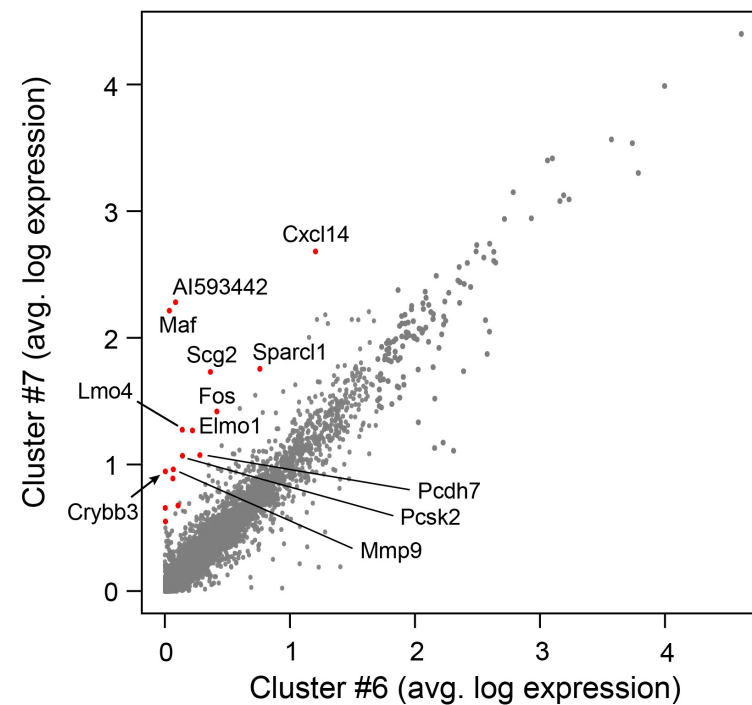
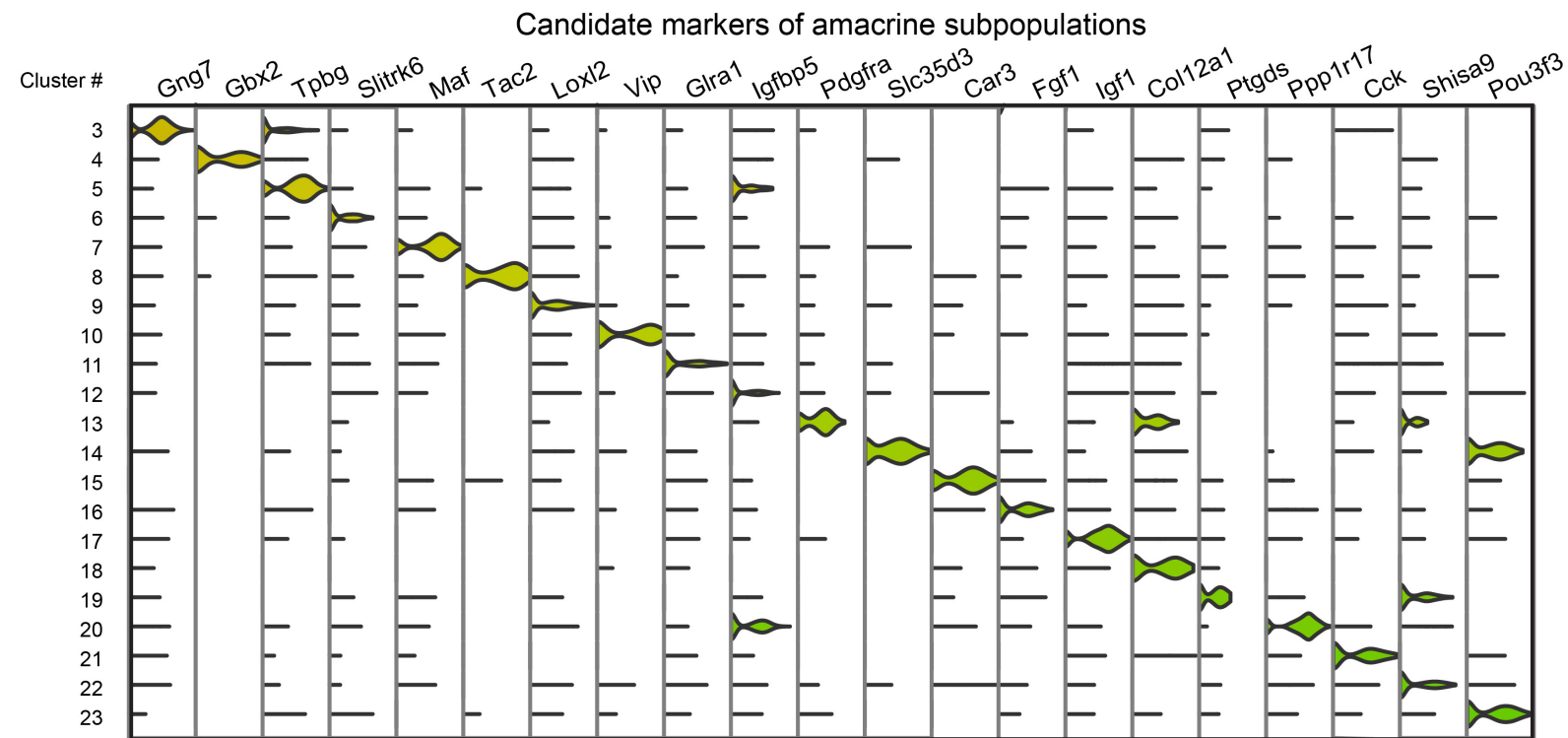
Annotating A Cell Atlas



39 Clusters: Known Cell Types & Relationships

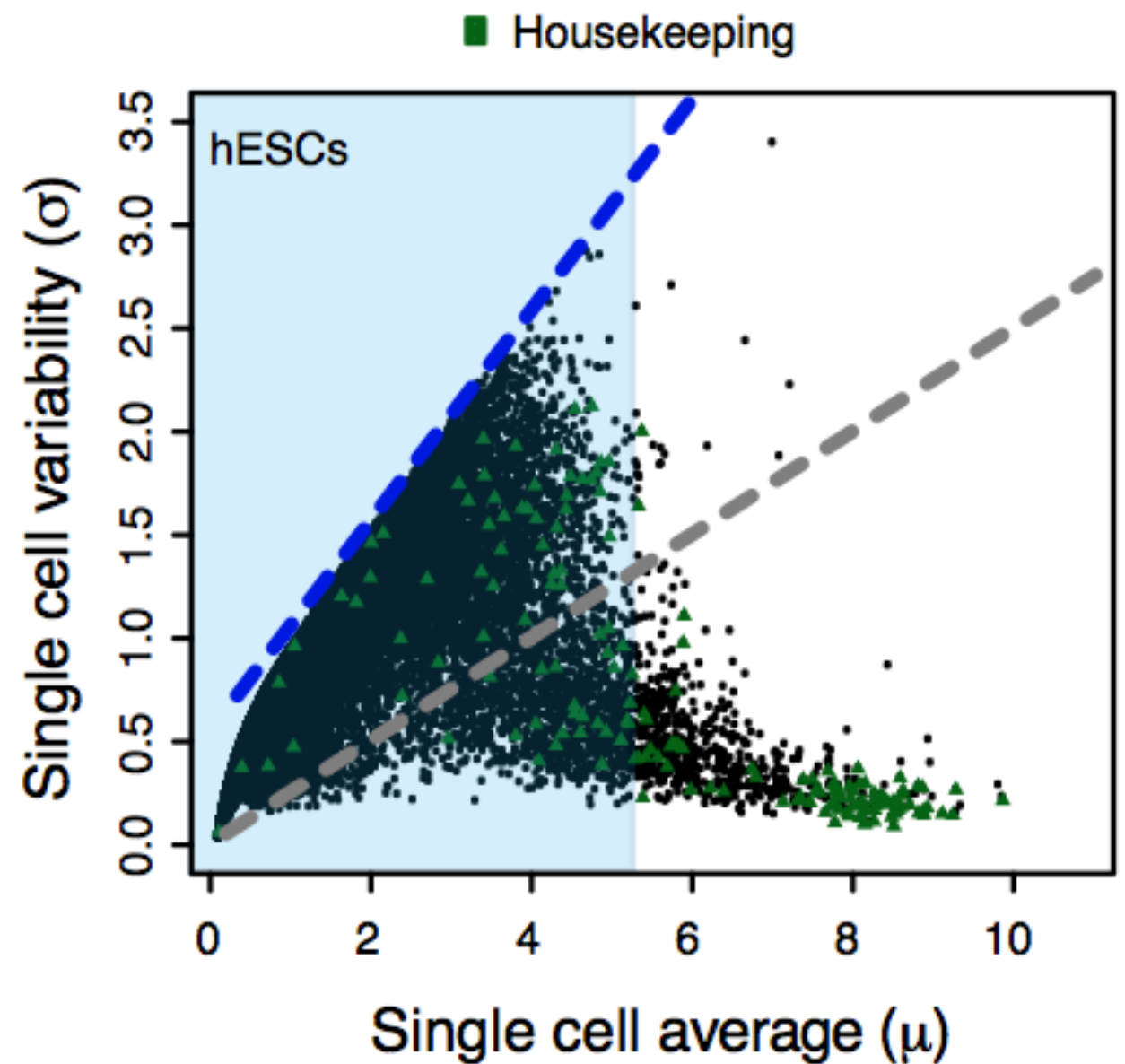
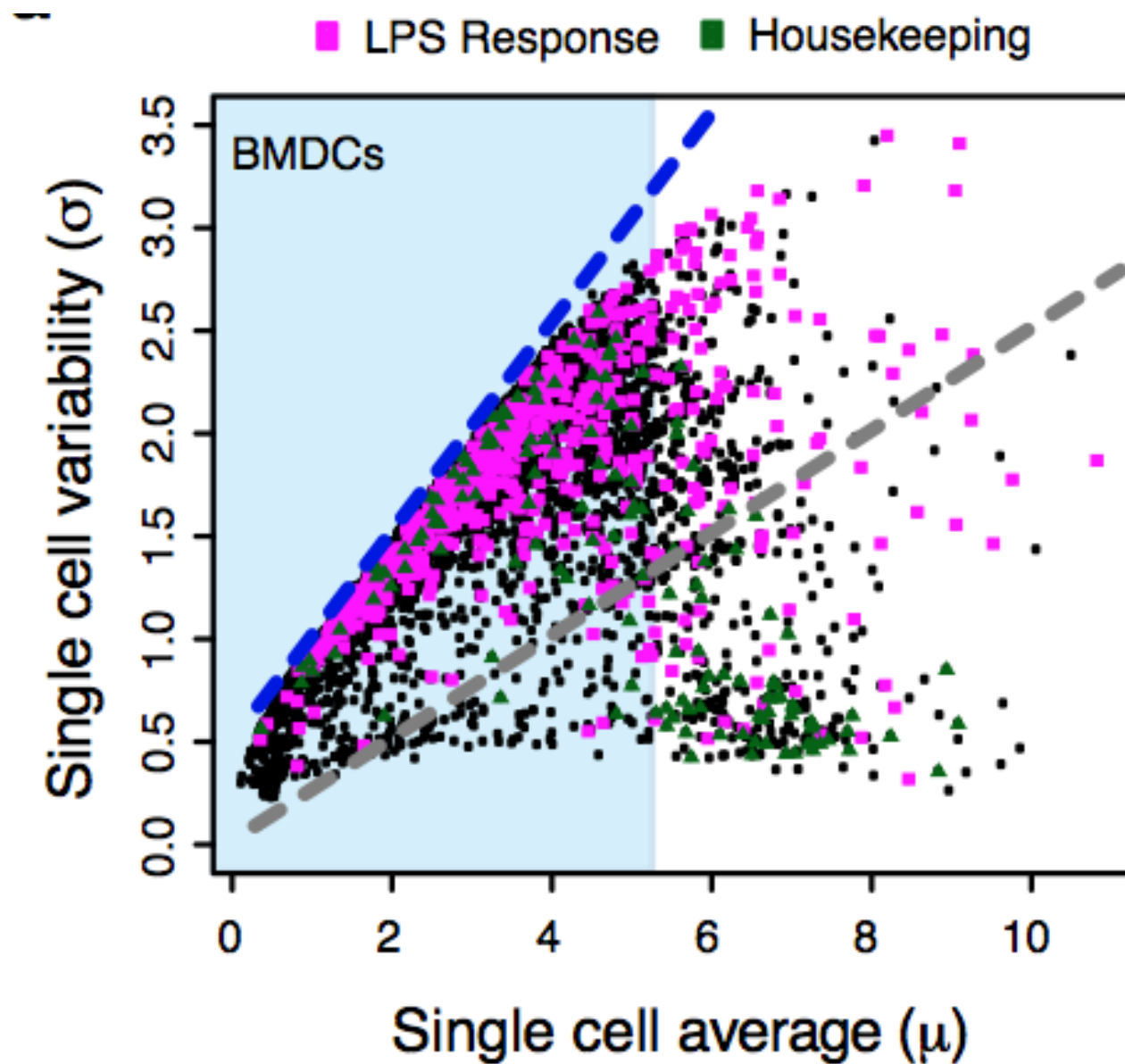


39 Clusters: New Markers That Can Be Validated!



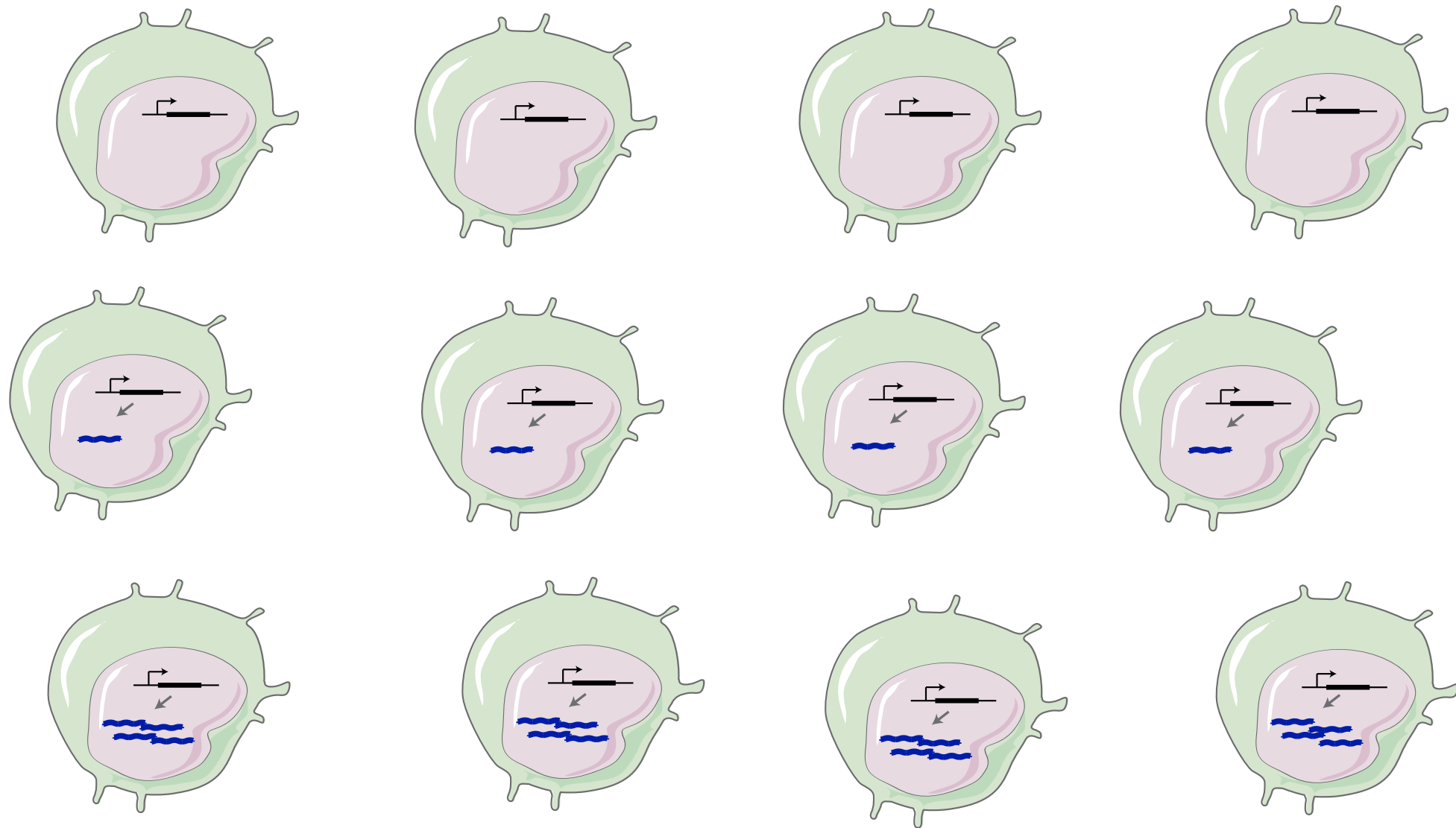
Clustering similar genes

Identifying 'variable' genes

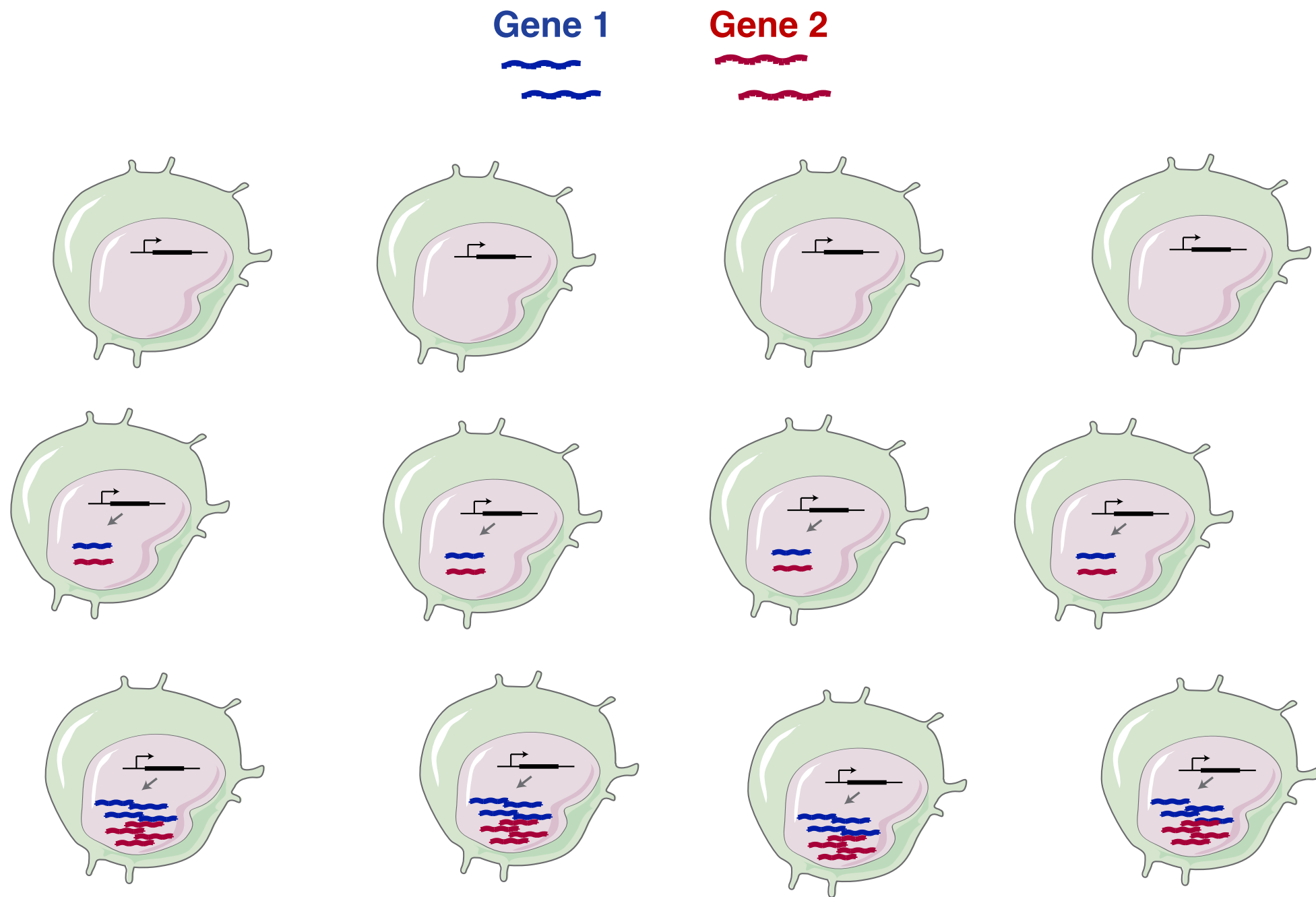


Variation is interesting

Gene 1



Co-variation implies co-regulation



Sparsity-based gene network inference

Hematopoietic Stem Cell (HSC)

Innate Immunity

NK-cells

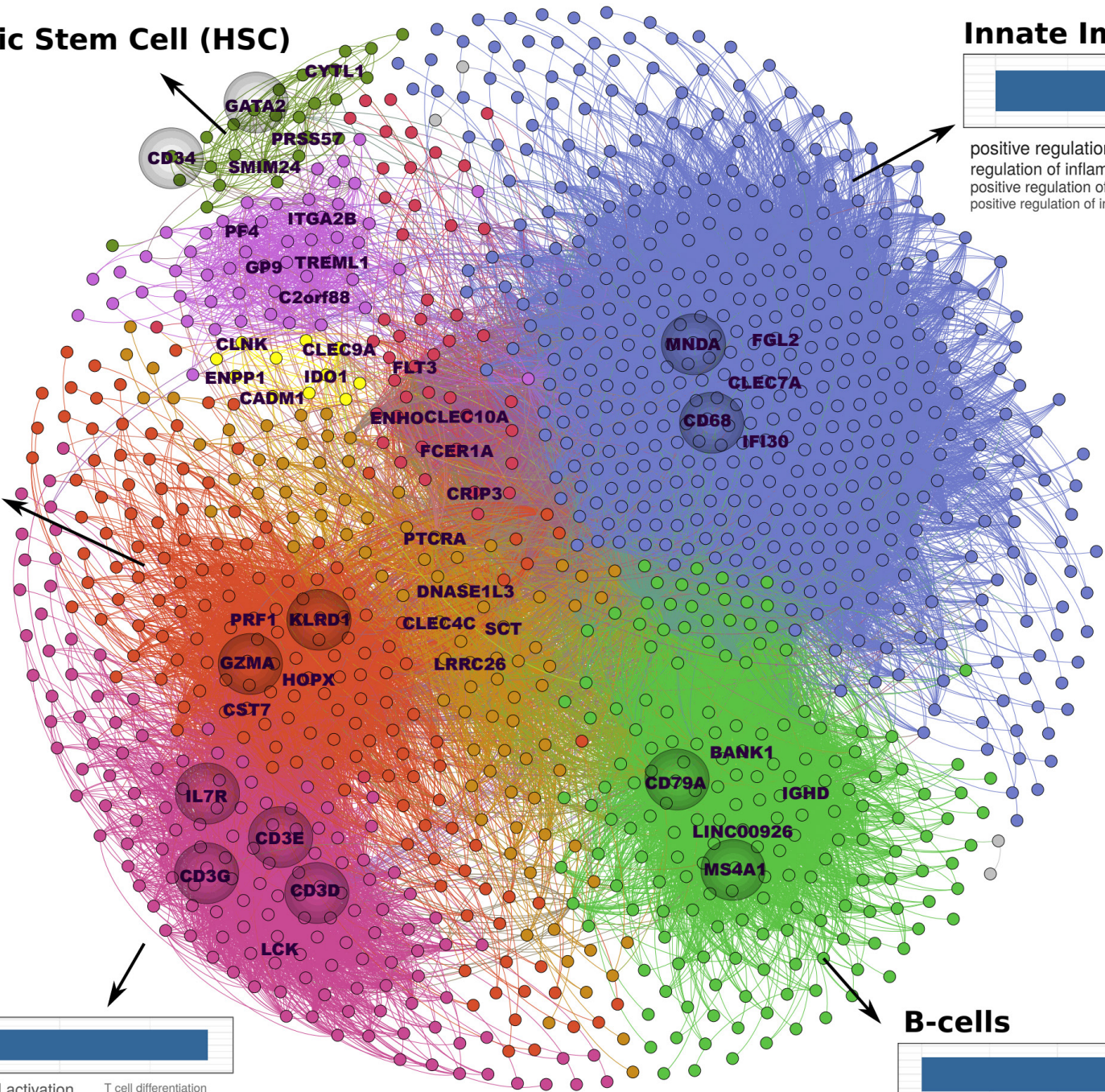
T cell activation
cell killing
regulation of alpha-beta T cell ...
cellular defense response
positive regulation of cytokine ...
natural killer cell mediated imm...
leukocyte cell-cell adhesion

T-cells

T cell activation
regulation of lymphocyte activation
positive regulation of T cell ac...
regulation of T cell activation
positive regulation of leukocyte...
antigen receptor-mediated signal...

B-cells

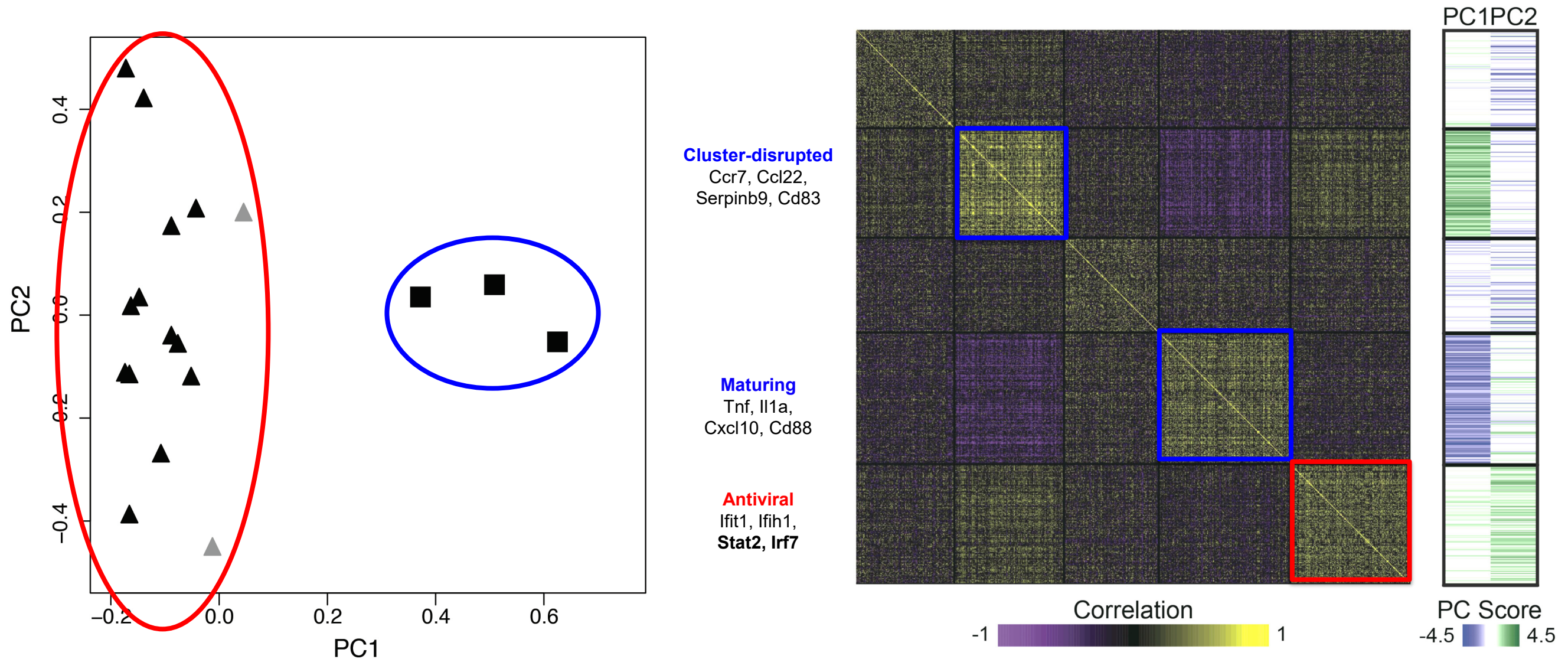
B cell activation
antigen receptor-mediated signal...
regulation of lymphocyte activation
positive regulation of leukocyte...
humoral immune response
antigen processing and presentat...
B cell differentiation
T cell activation
T cell costimulation



positive regulation of cytokine ...
regulation of inflammatory response
positive regulation of innate im...
positive regulation of inflammat...
cytokine biosynthetic process
regulation of interleukin-8 prod...
chemokine production
myeloid leukocyte migration
response to molecule of bacteria...
cell chemotaxis
detection of biotic stimulus
interleukin-1 beta secretion
regulation of interleukin-6 prod...
defense response to other organism
L-kappaB kinase/NF-kappaB signaling
positive regulation of cytokine ...
cytokine secretion
heterotypic cell-cell adhesion
cellular response to molecule of...
positive regulation of tumor nec...

CO-Dependency network of genes

Genes Co-Vary Across Single Cells

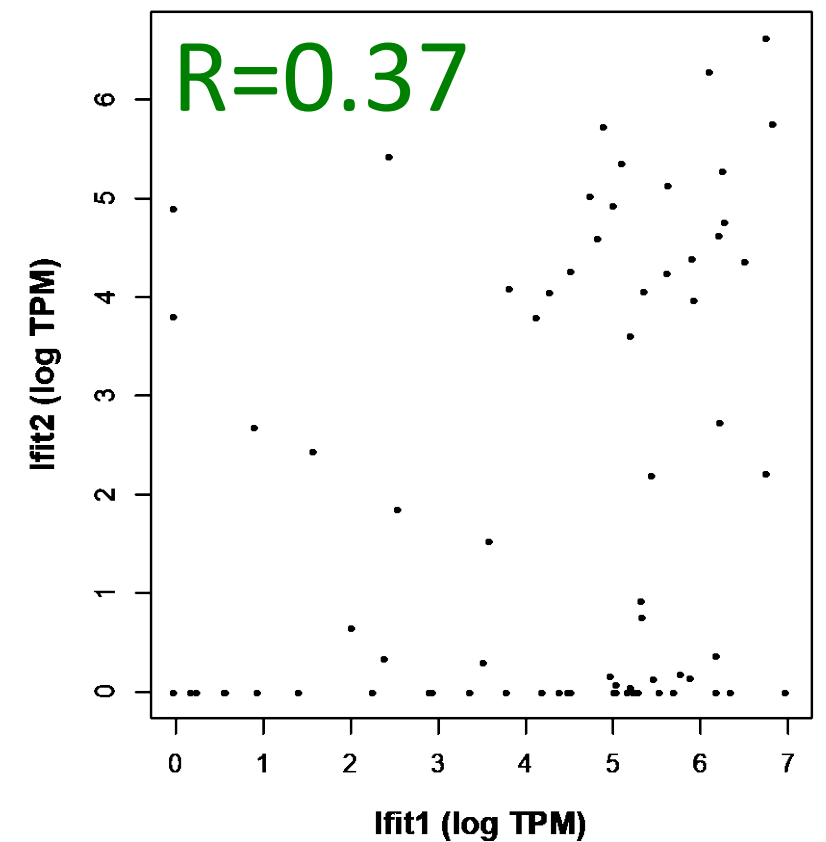
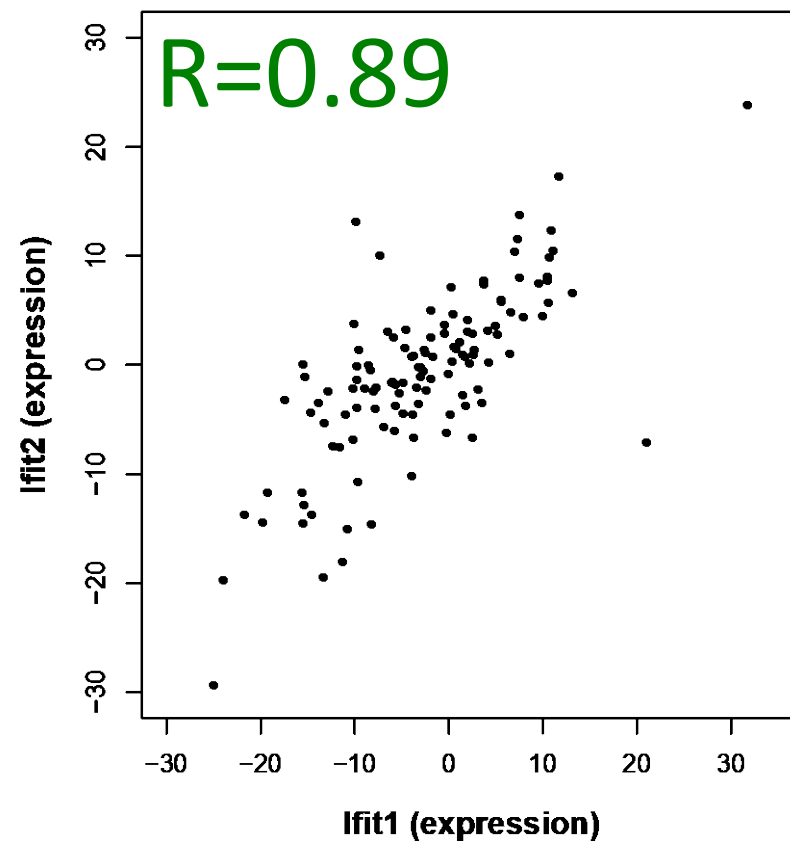
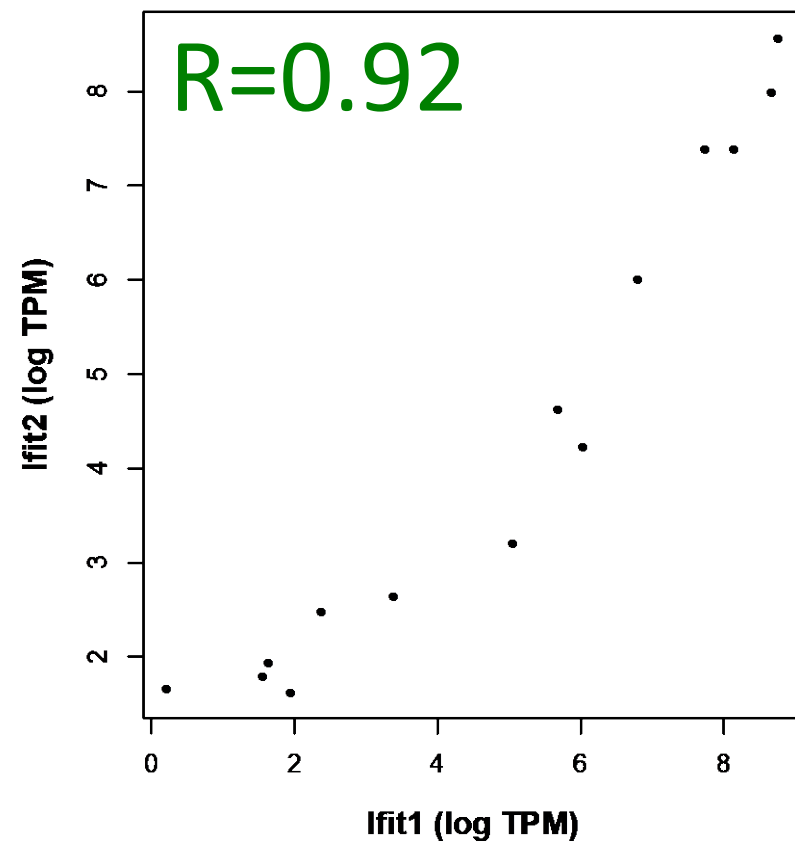


We can uncover cell states and circuits, as well as their markers and drivers, from structures in cell-to-cell variation

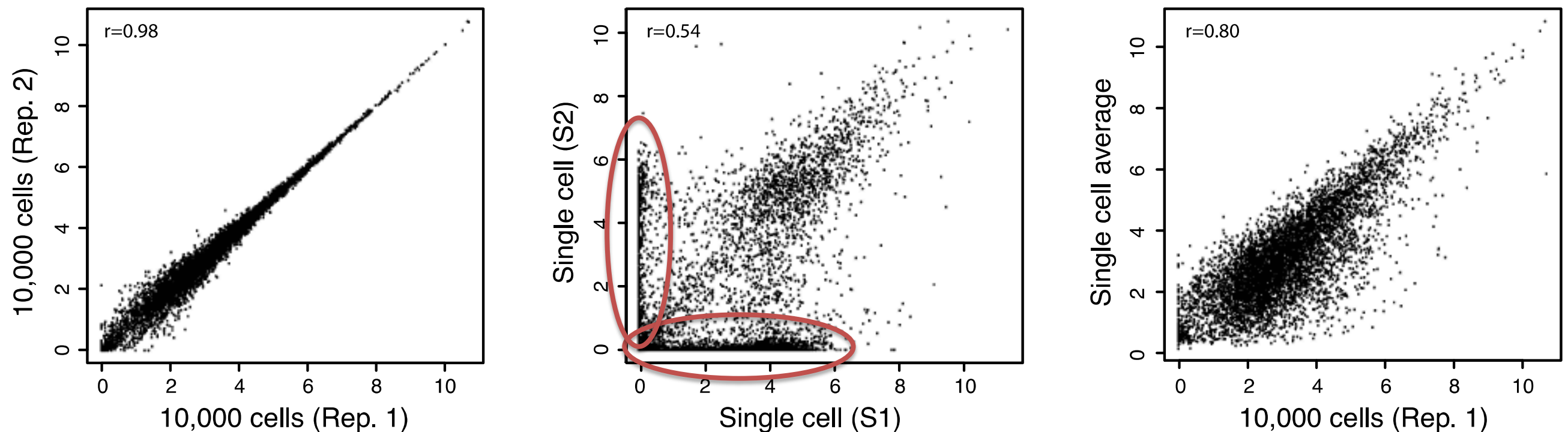
Correlation is not well-suited for single-cell analysis

POPULATIONS

SINGLE CELLS



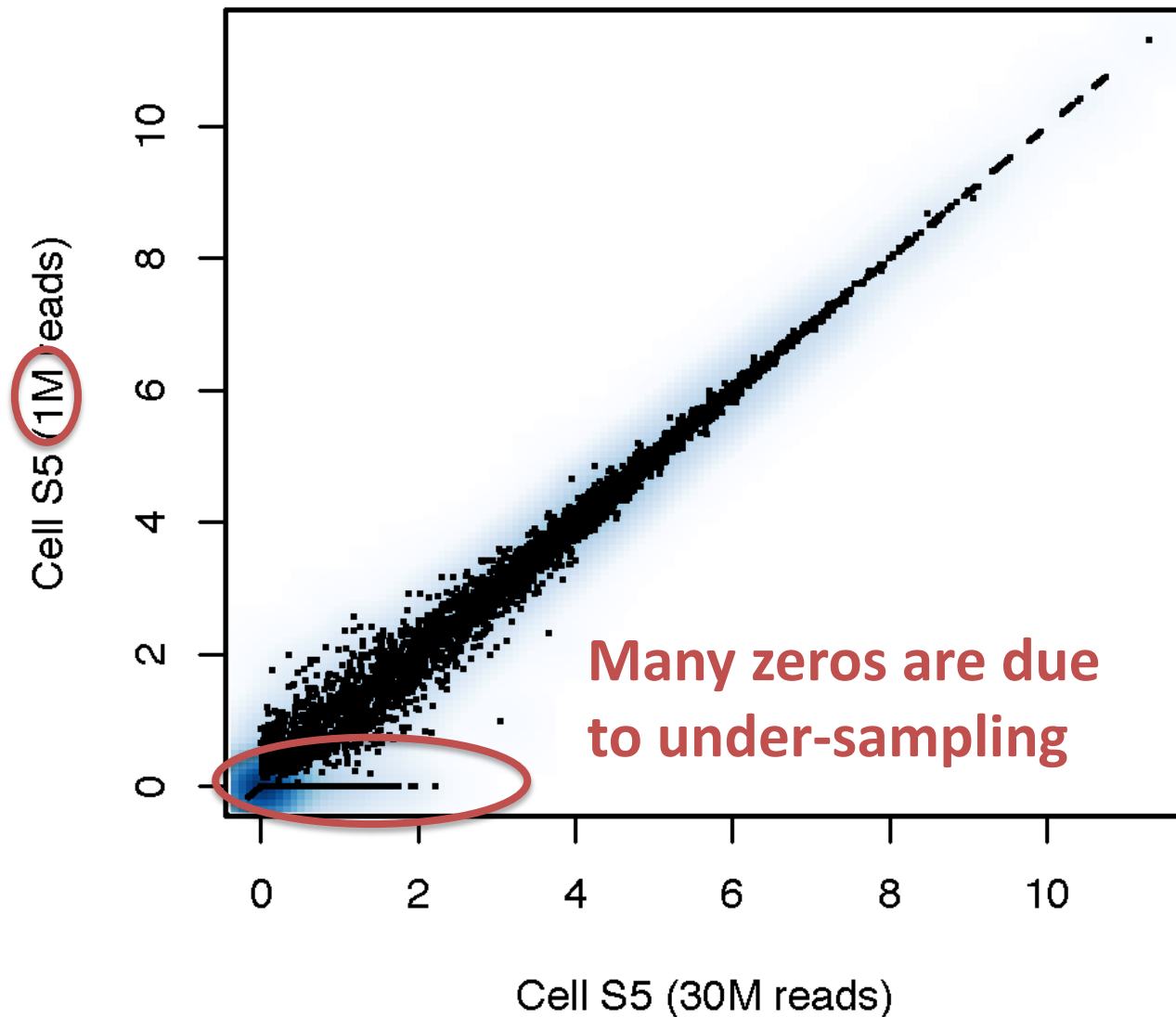
scRNA-Seq data has many many zeros



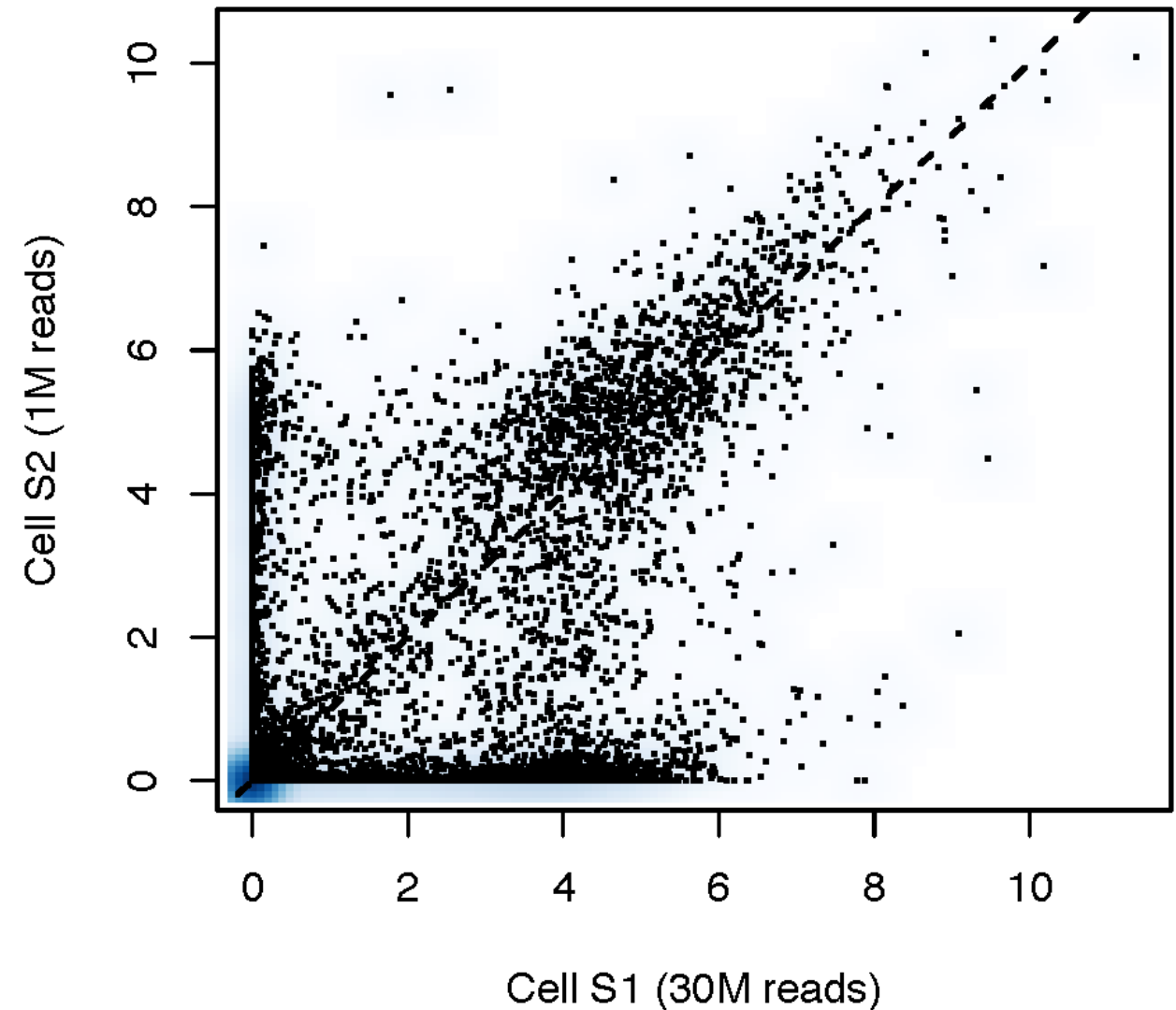
Transcriptome-wide, single cells are very different.

Variability due to sampling vs. biology

Sampling Bias

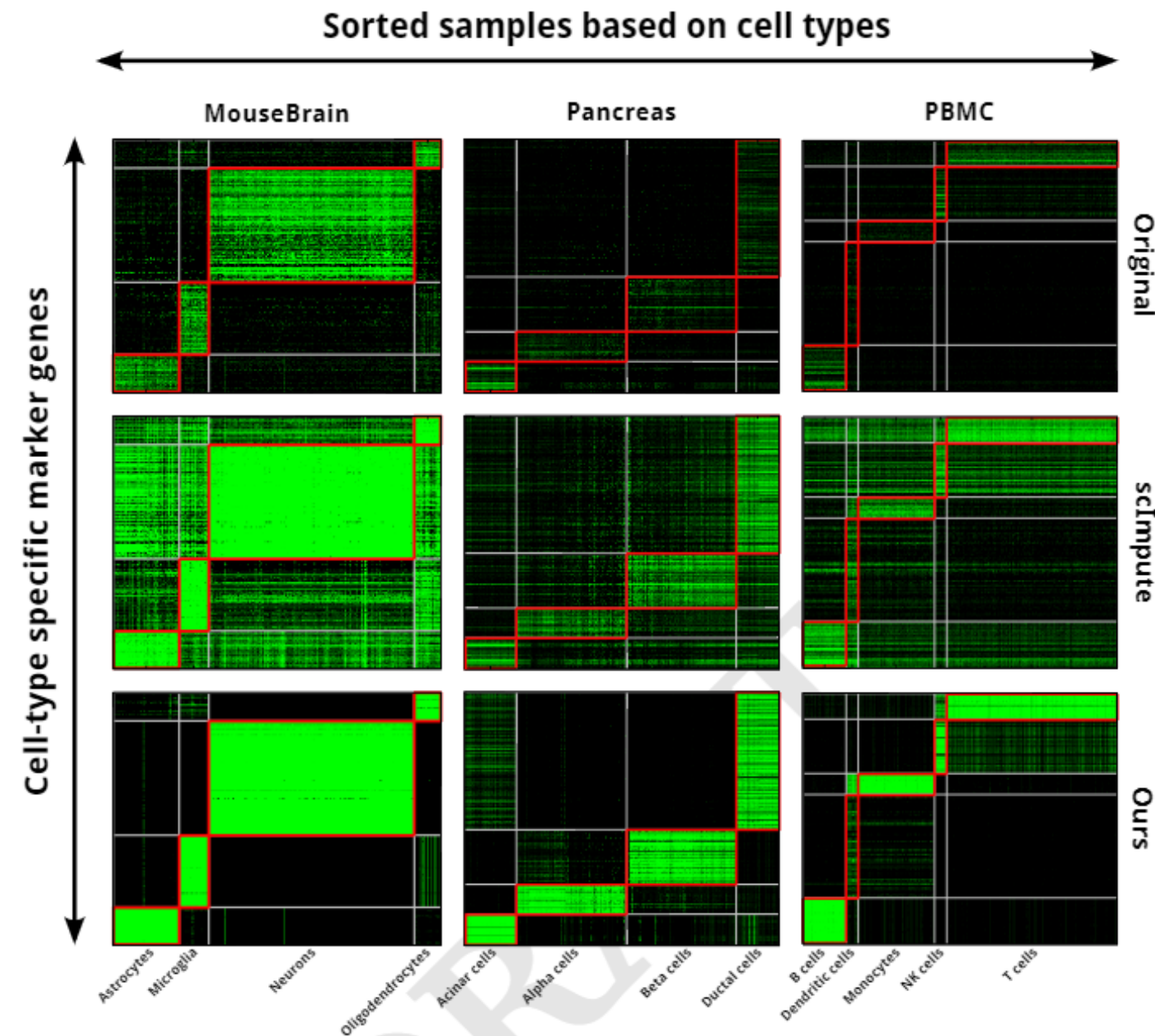


Cellular Variation



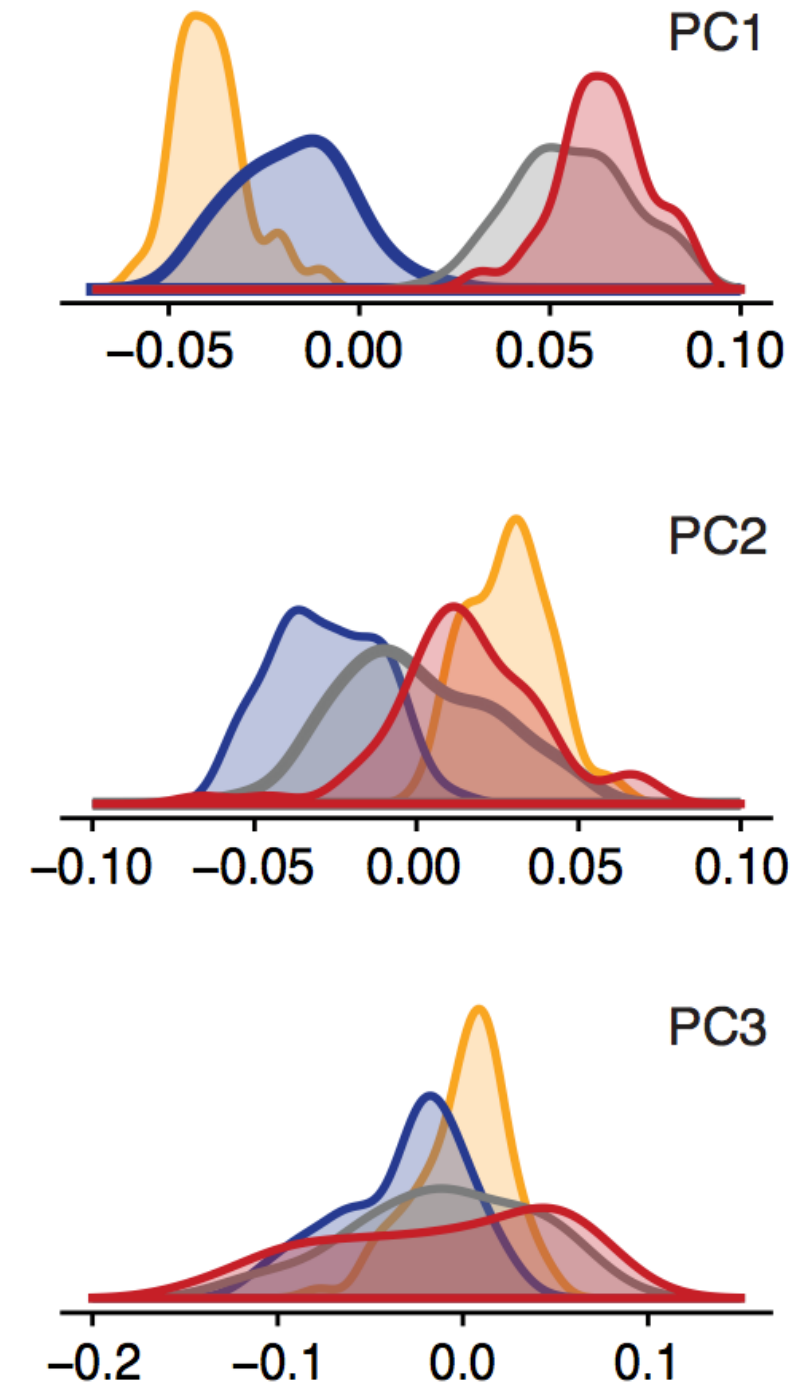
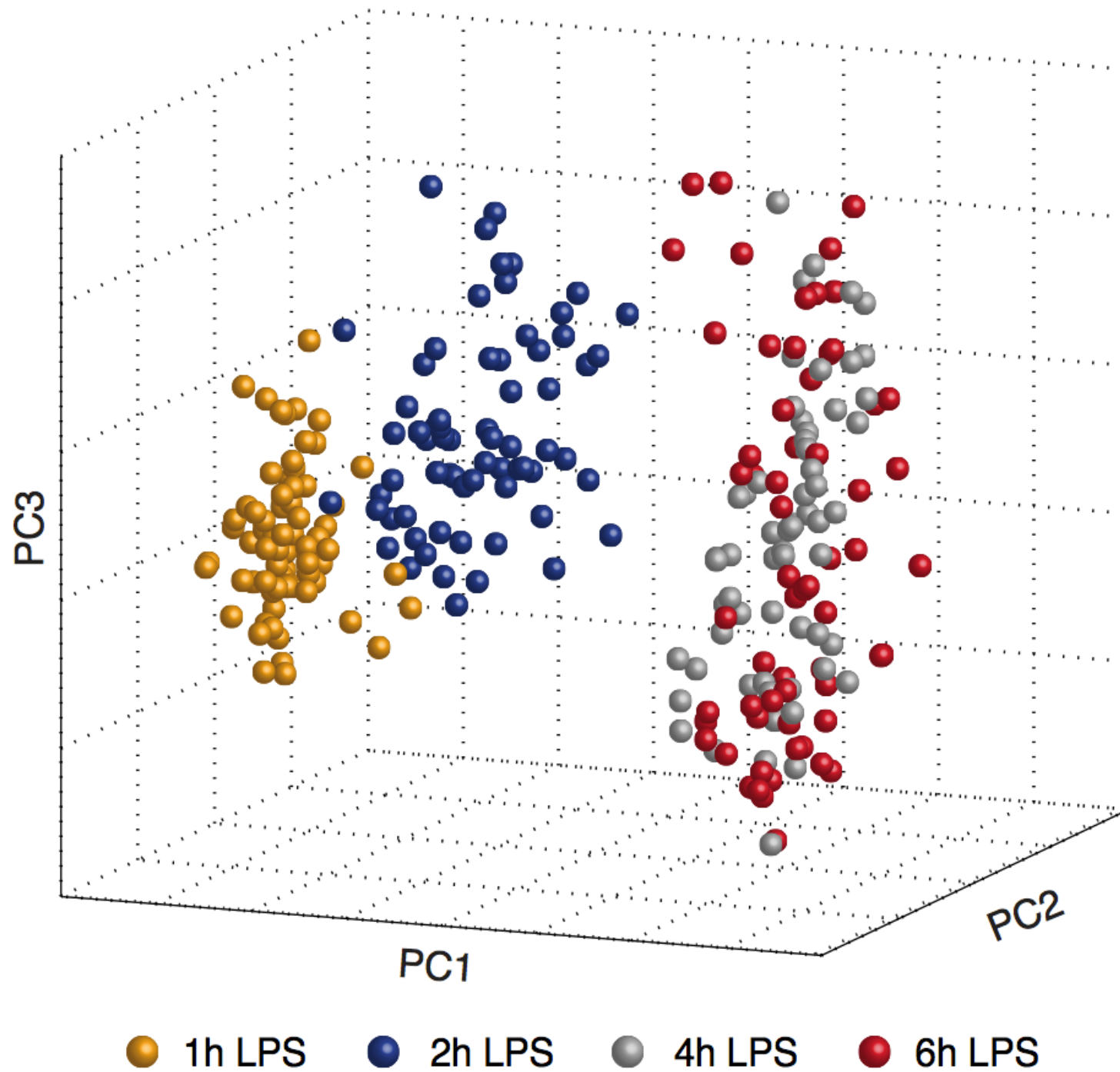
Dimensionality reduction

Dimensionality Reduction



- Curse of dimensionality
- Easier to visualize/process
- Reduce noise
- Linear methods: *PCA*
 - Identifying batch/cell cycle effects
- Nonlinear methods: t-distributed stochastic neighbor embedding (*t-SNE*)
 - Exploratory data analysis

PCA – 300 cell dataset

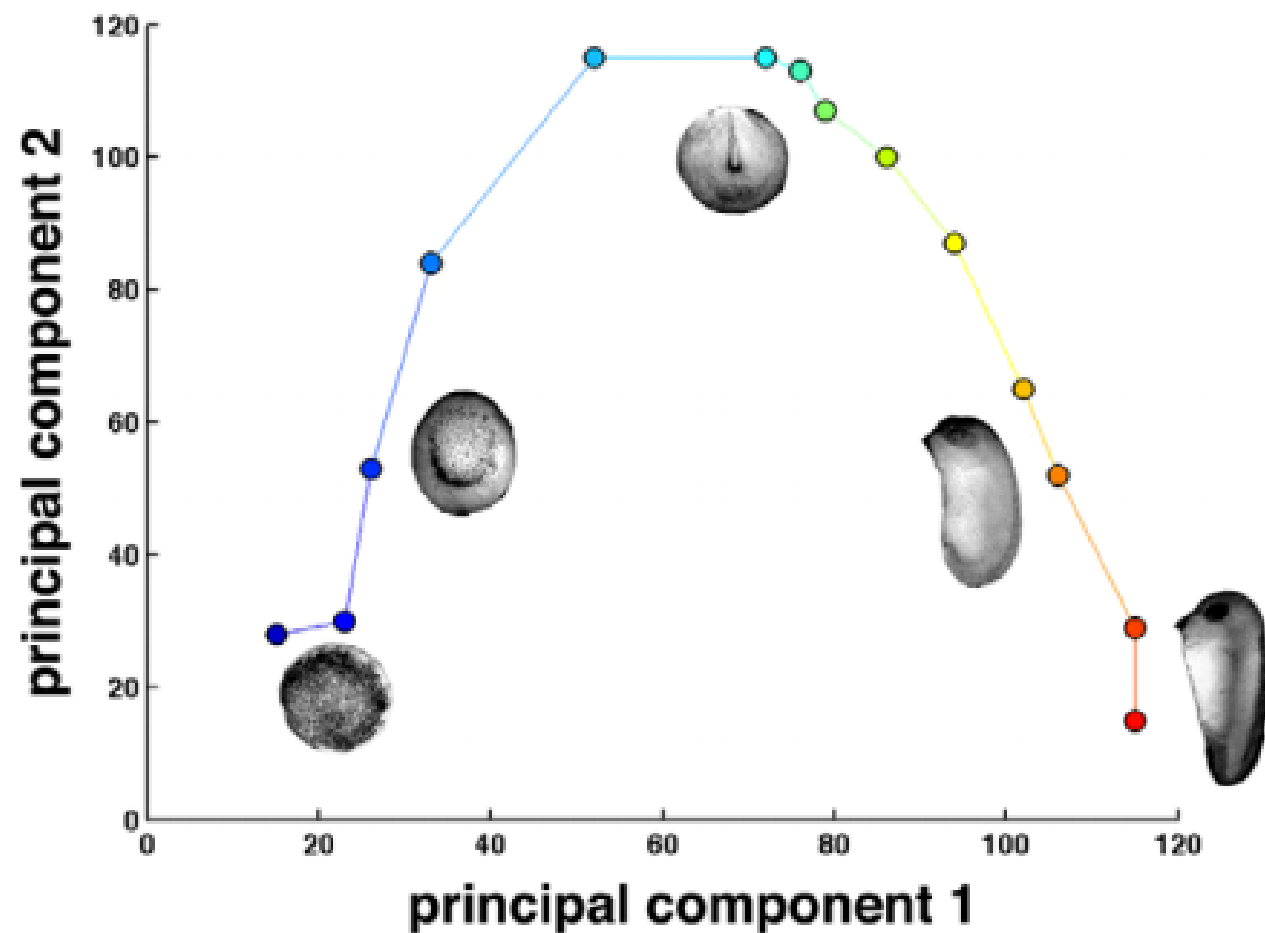


Important consideration for PCA

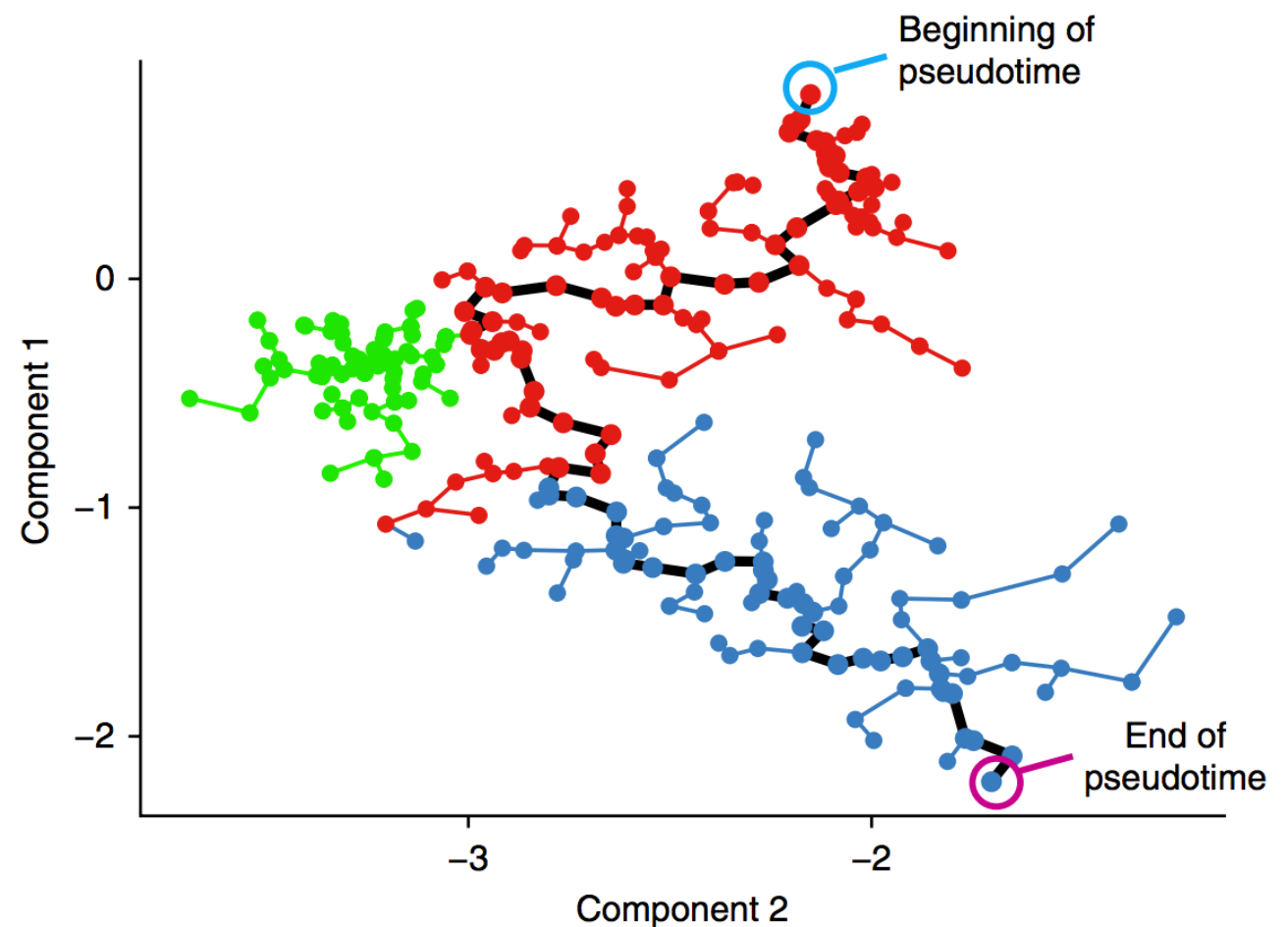
- **Input gene list**
 - Can dramatically alter output
- **Interpretation:**
 - ‘Assigning ‘biology’ or function requires prior knowledge
 - PCs often correlate with technical quality
 - Not all PCs are significant (Chung, Storey, arXiv.org)
- **Limitations/extensions:**
 - PCs represent **linear** combination of individual features

Interpreting dimensionality reduction

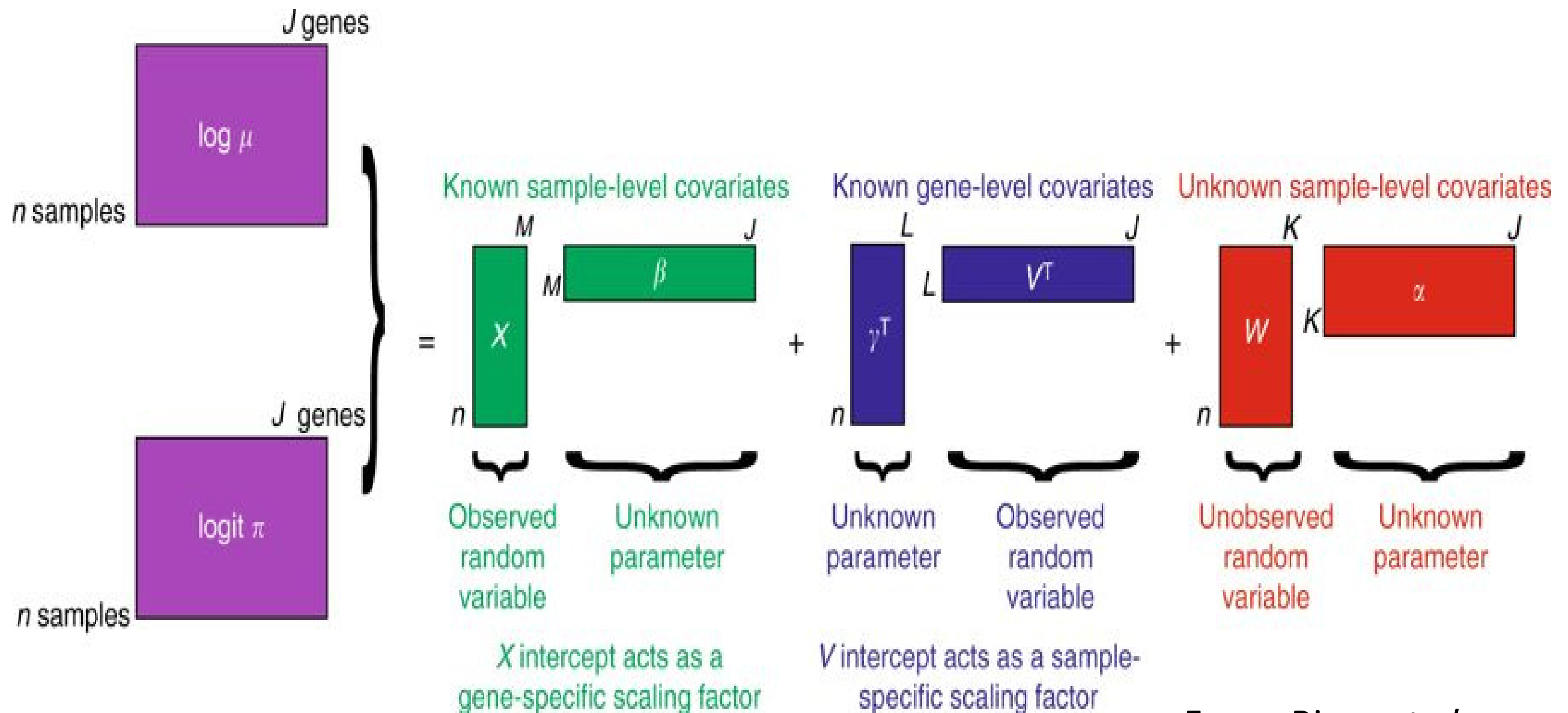
Anavy et al, Development, 2014



Trapnell et al, Nat. Biotech., 2014



Zero-inflated negative binomial model (ZINB-WaVE)



From: Risso *et al.*,
2018

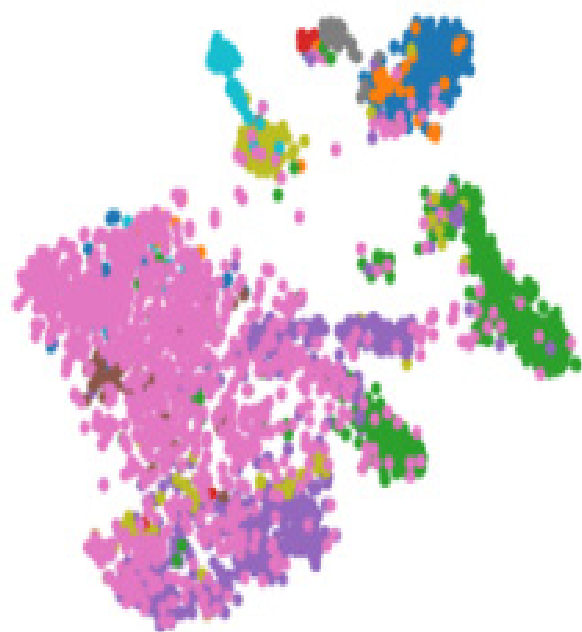
- A generalized linear factor analysis model

Dropping factor analysis in favor of deep autoencoders

- Single-cell Variational Inference (scVI)
- Deep count autoencoder (DCA)

tSNE of low dimensional representation

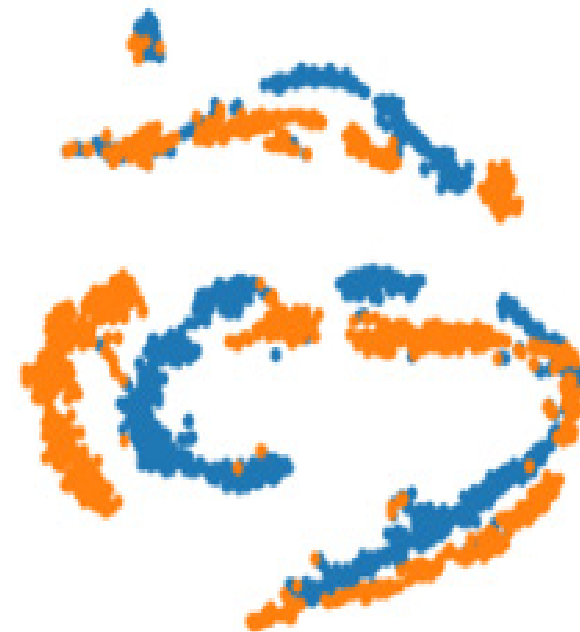
DCA - Rosenberg 2018



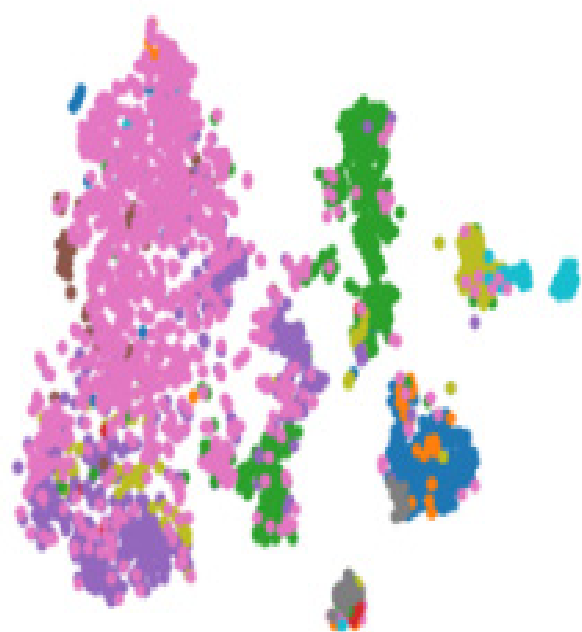
DCA - Zeisel 2018



DCA - Lukassen 2018



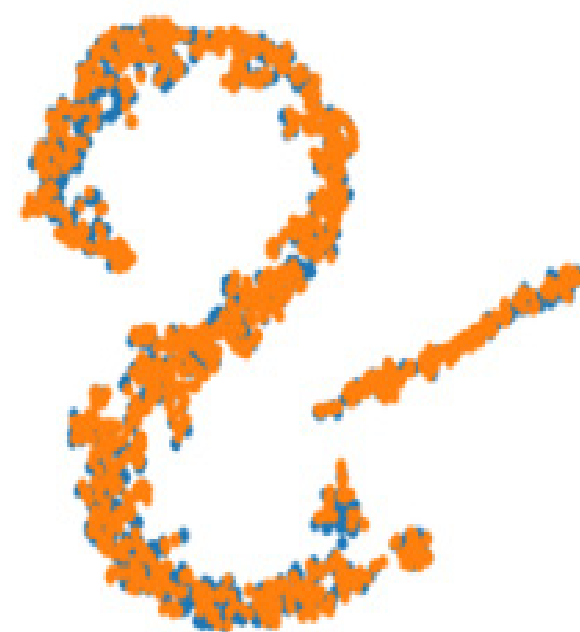
scVI - Rosenberg 2018



scVI - Zeisel 2018



scVI - Lukassen 2018



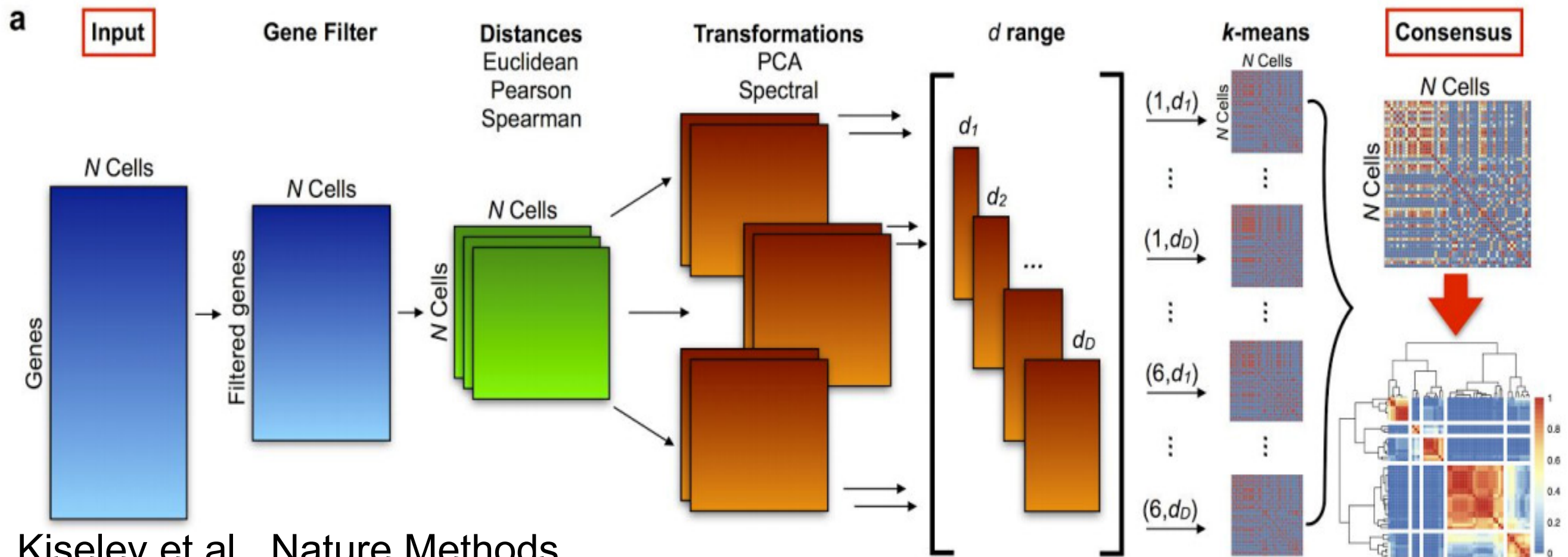
**Dinstinguishing
different cell types**

Discrete cell type identification

- **Based on traditional clustering approaches:**
k-means, hierarchical, and graph-based clustering techniques
- **tSNE + k-means (traditional)**
- **SINCERA (Guo et al. 2015)**
 - Based on hierarchical clustering
 - Data is converted to z-scores before clustering
- **SNNCliq (C. Xu and Su 2015)**
 - Identifies the k-nearest-neighbours of each cell according to the distance measure.
 - Clusters are defined as groups of cells with many edges between them using a “clique” method.
- **PCAReduce (žurauskienė and Yau 2016)**
 - Combines PCA, k-means and “iterative” hierarchical clustering.
 - Starting from a large number of clusters pcaReduce iteratively merges similar clusters
 - After each merging event it removes the principle component explaining the least variance in the data.
- **SC3 (Kiselev et al. 2017)**
 - Based on PCA and spectral dimensionality reductions
 - Utilises k-means
 - Additionally performs the consensus clustering

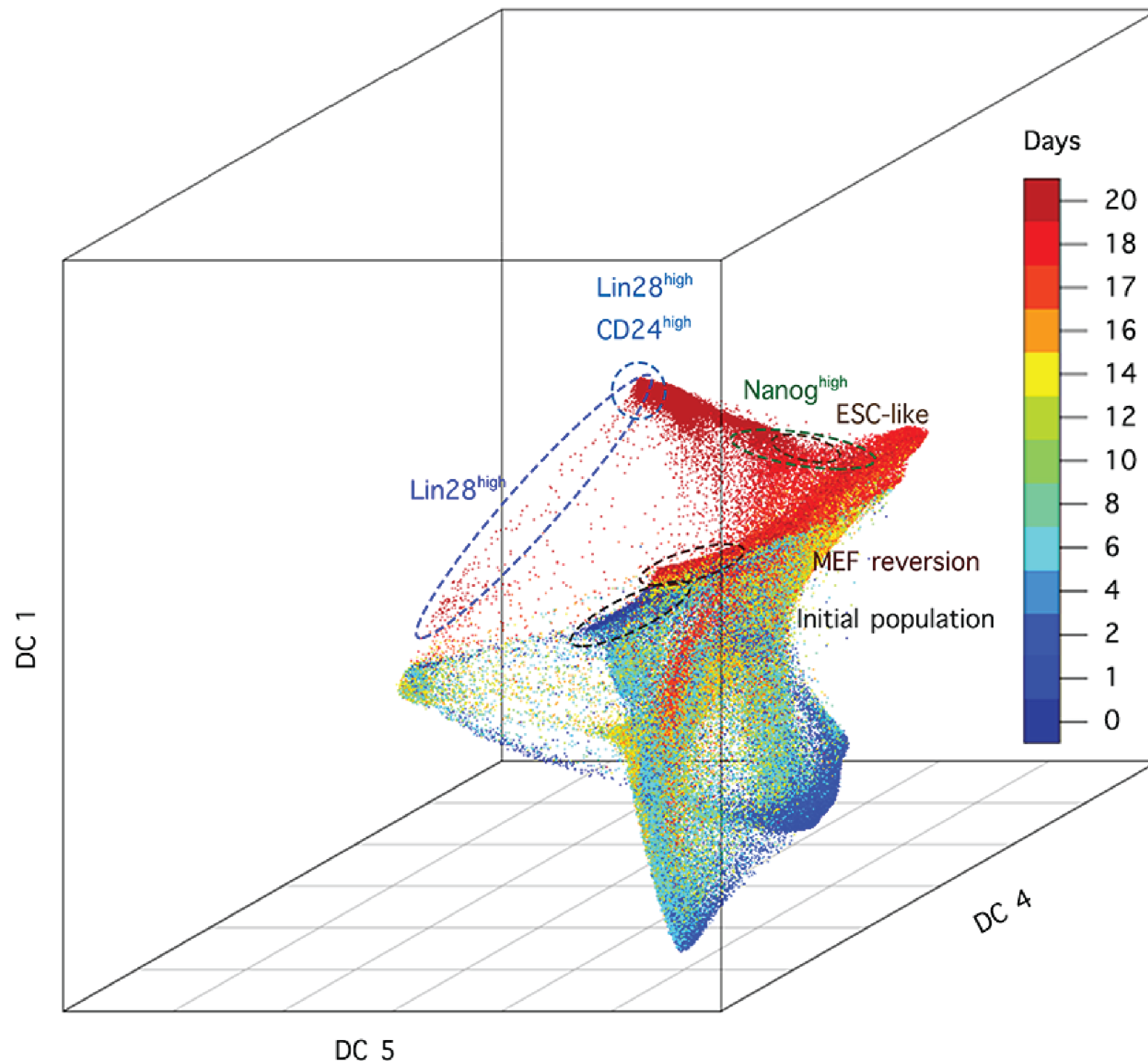


Single-Cell Consensus Clustering (SC3)

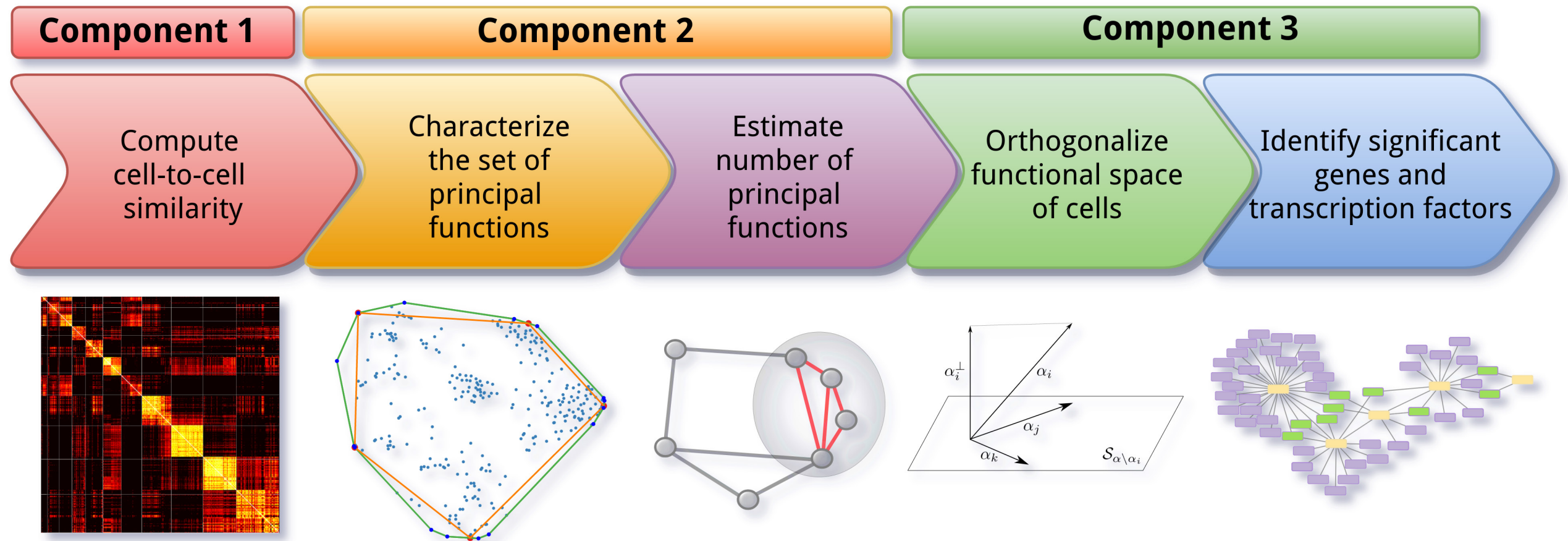


Kiselev et al., Nature Methods,
2017

Continuous cell states: diffusion map



Archetypal-analysis for Cell type indentificaTION (ACTION)



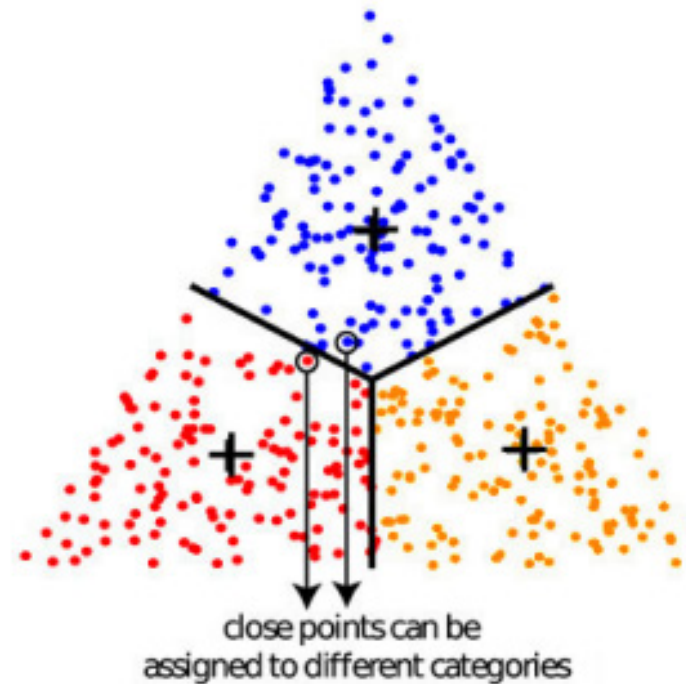
Mohammadi *et al.*, BioRxiv 2016, Nature Communications, under review

Combine discrete + continuous: archetype analysis

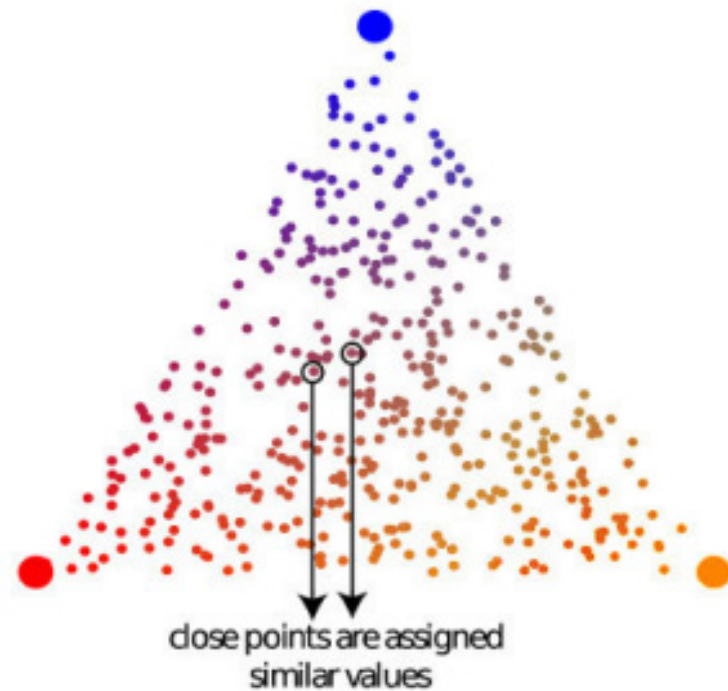
A Clustering works well when data is in discrete groups



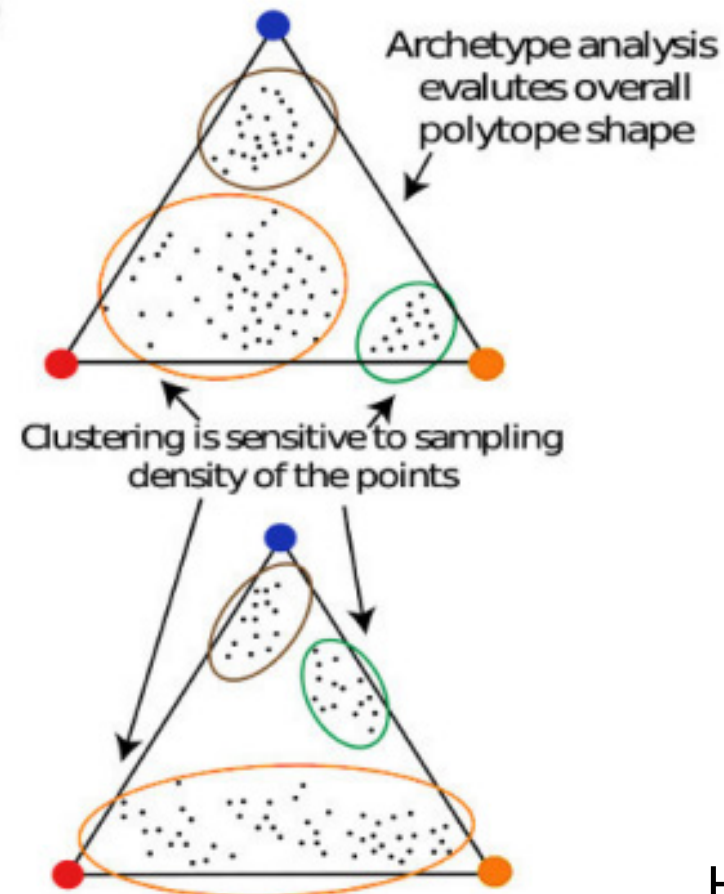
B But gives arbitrary grouping when data is continuous



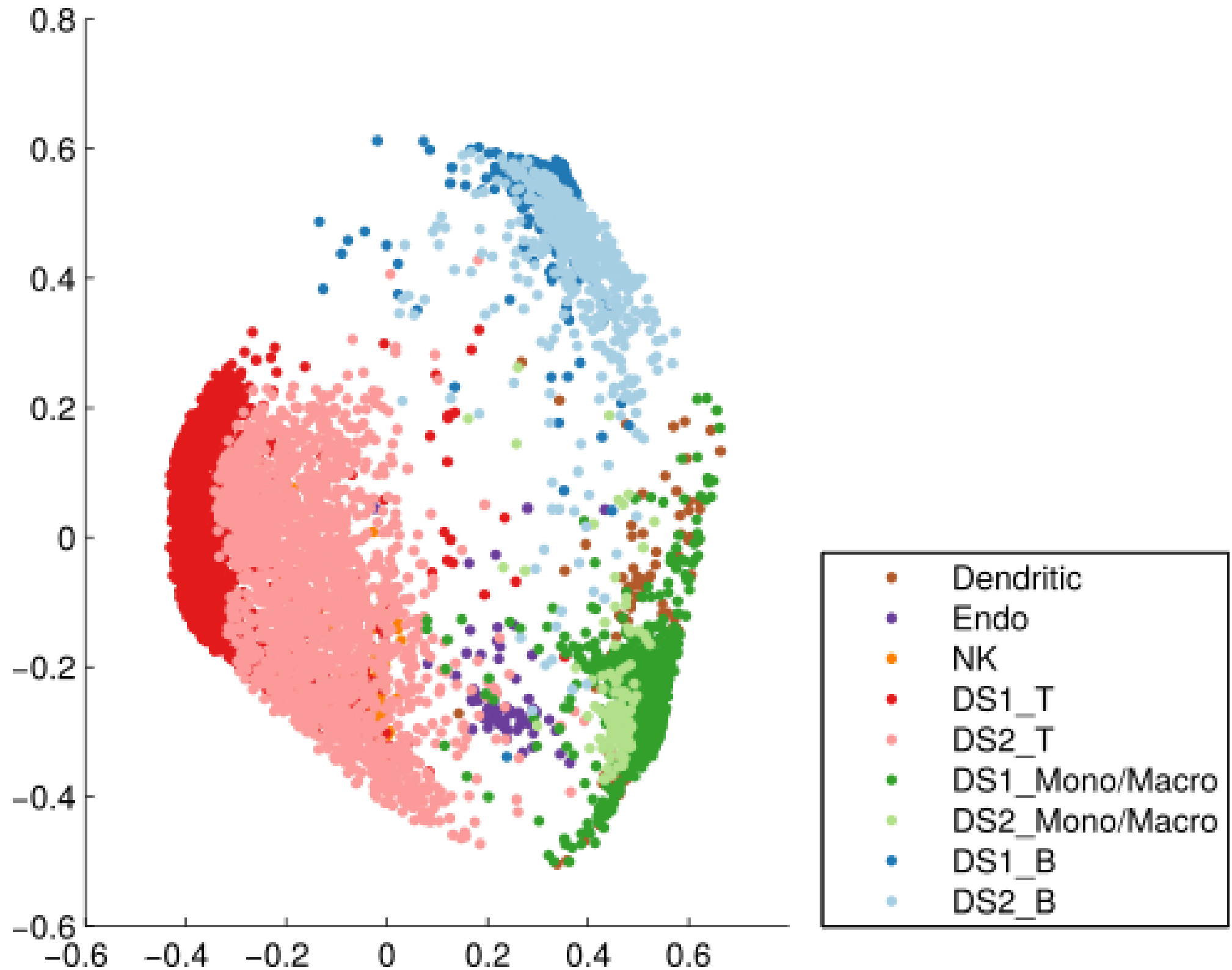
C Archetype analysis characterizes points by distance from the vertices



D



Matching cell types across datasets



Alignment of PBMC vs. Tumor scRNA

Multi-resolution analysis

ACTIONet

Main issues with parametric methods

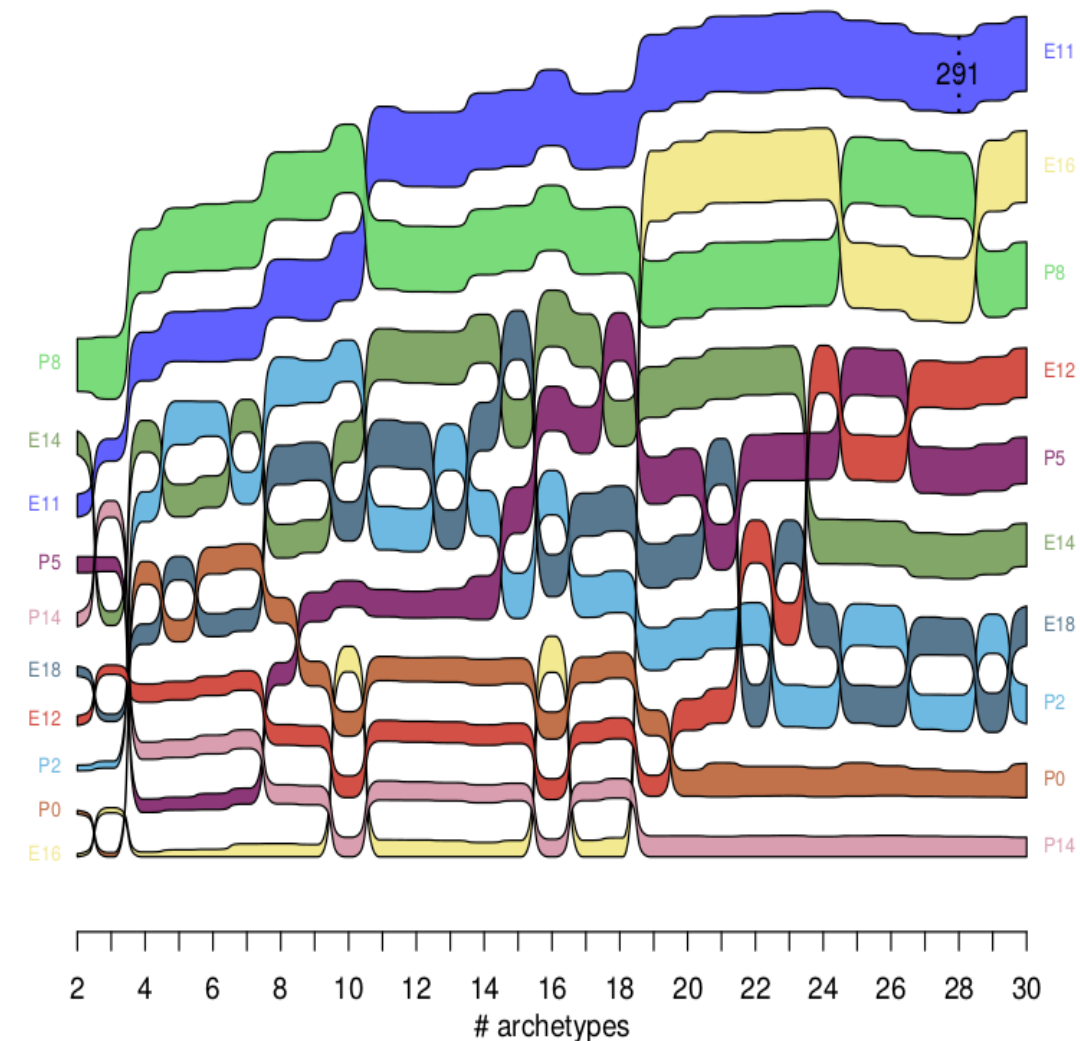
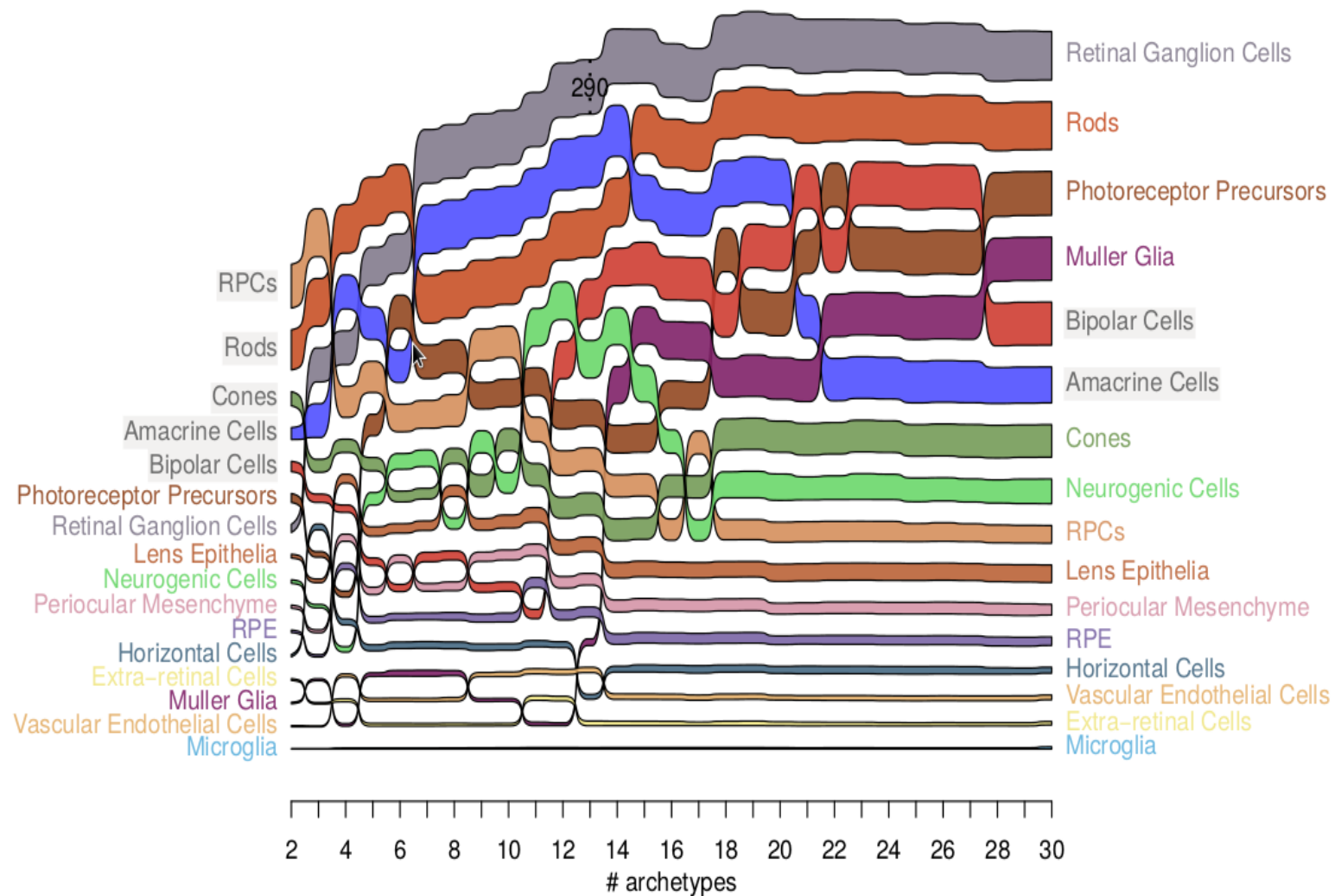
How many archetypes?

How many factors?

How many clusters?

Optimal number of factors differs by celltype/age

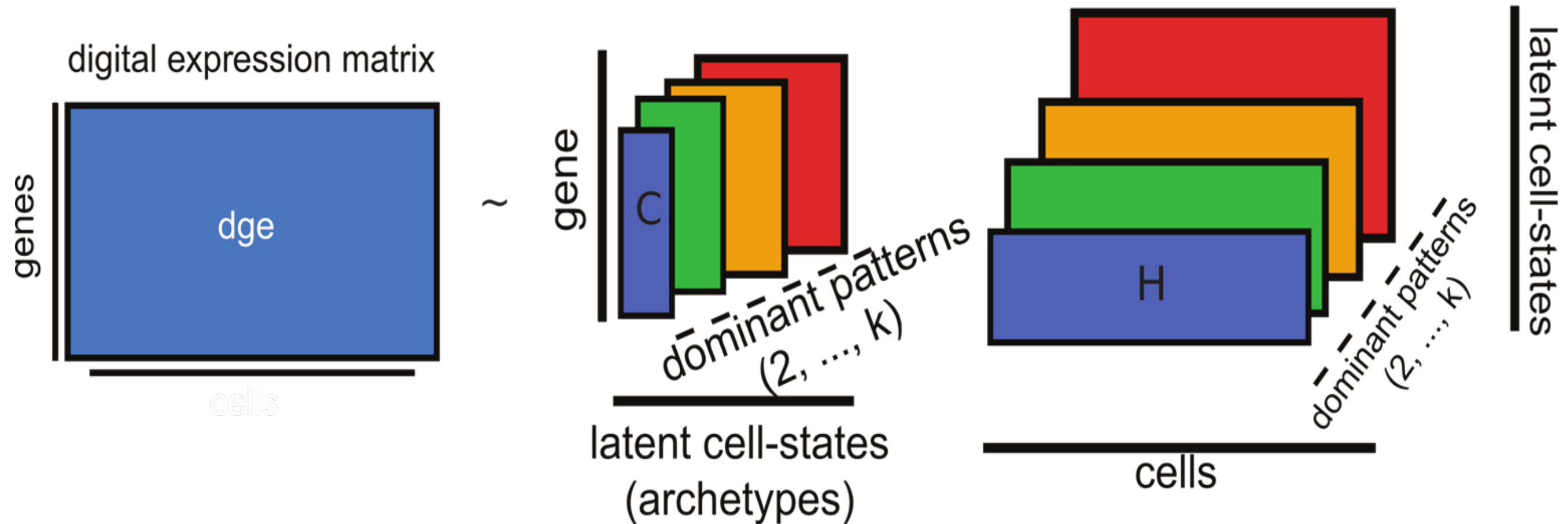
(Ex.: Mouse retina development -- similar results with other species/tissues)



Choosing one “optimal” k is dominated by the major cell type
(defeating the whole purpose of single-cell analysis)

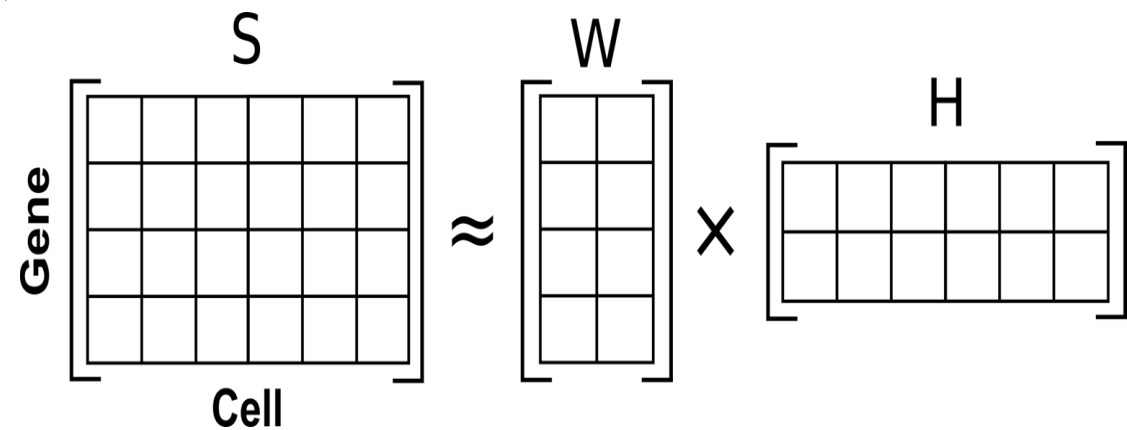
Recently developed method: ACTIONet

ACTION multiresolution decompositions



Complementary approaches

Factor analysis



- Identifies hidden cell states/gene programs
 - Biological
 - Technical
- Deconvolves complex biological processes
- Uncovers discriminating/marker genes

Network analysis



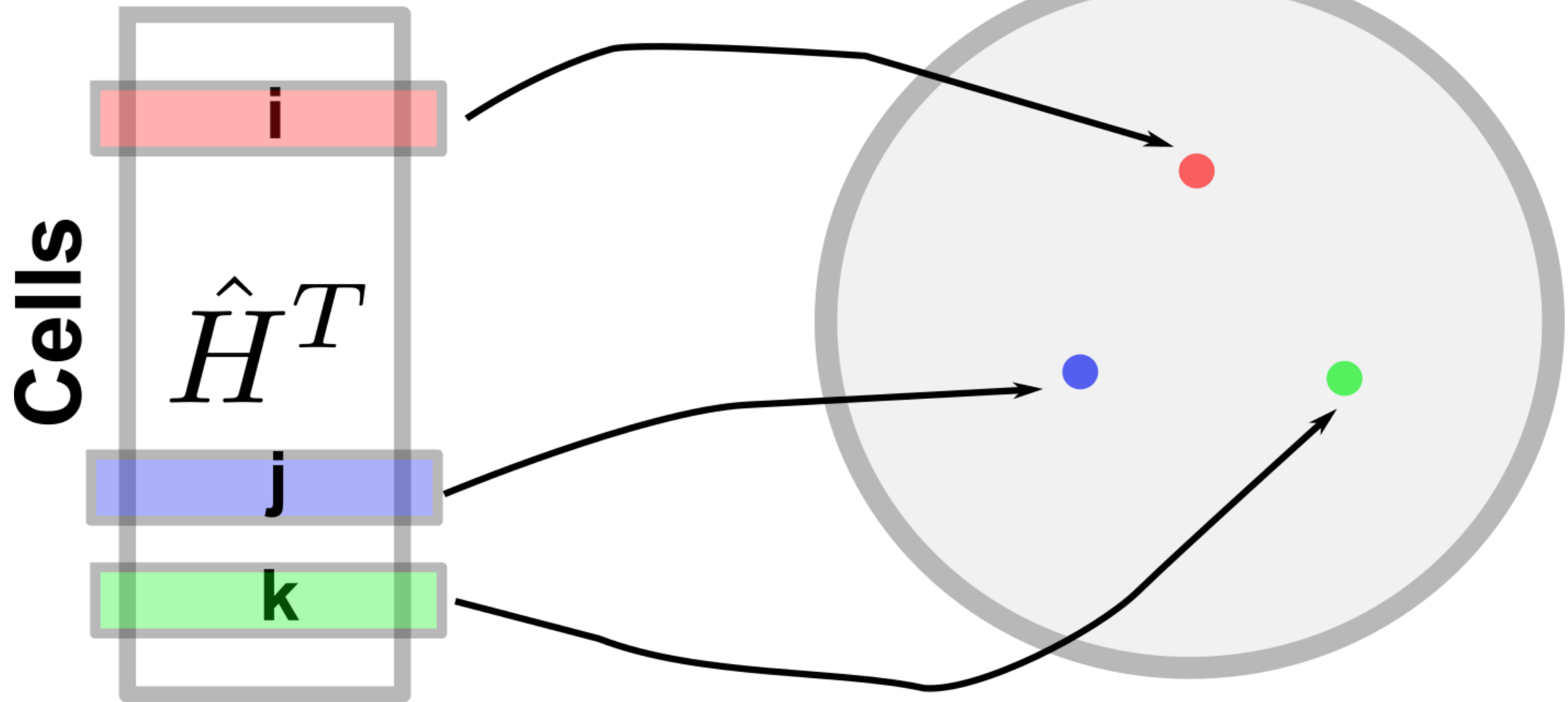
- Reconstruct the topography of cell space
- Rich set of graph-based algorithms
 - Visualization (UMAP)
 - Clustering (Louvain/Leiden)
 - Imputation (PageRank)

Step 1: Define a metric cell space

Metric cell space

$$\delta(h_i, h_j) \leq \delta(h_i, h_k) + \delta(h_k, h_j)$$

Cell states



$$\delta(h_i, h_j) = \sqrt{\mathbf{JS}(\hat{h}_i, \hat{h}_j)}$$

How?

Step 1: Define a metric cell space

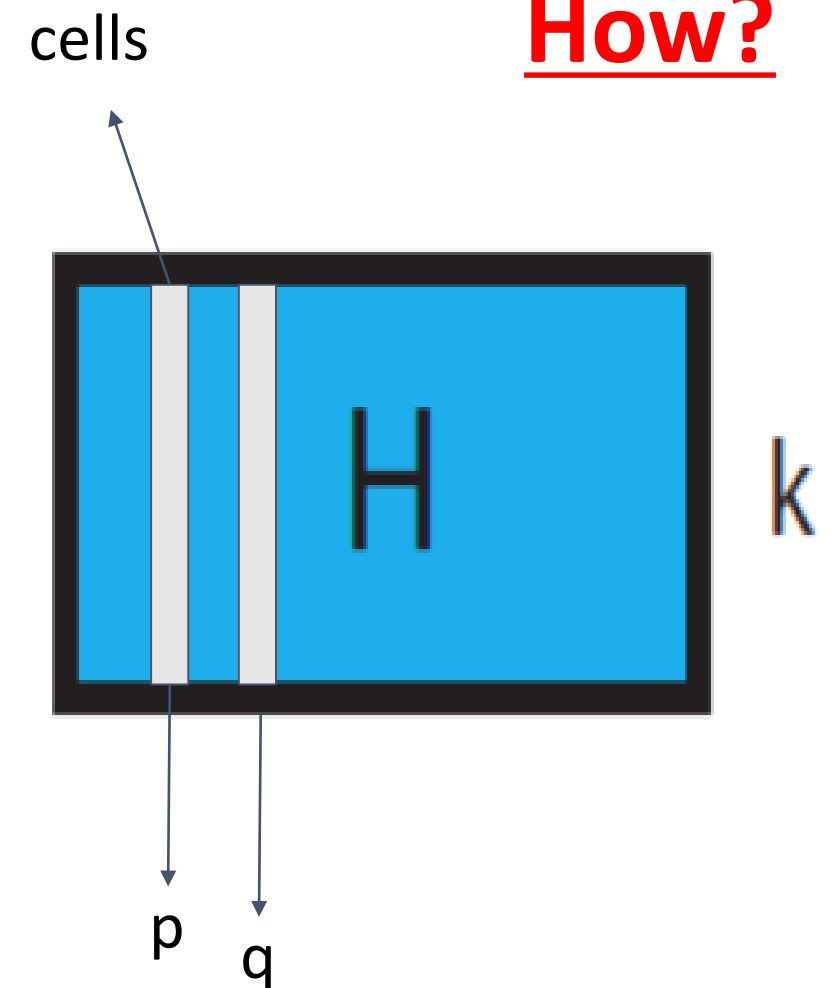
Kullback–Leibler divergence

$$\mathbf{KLD}(p \parallel q) = \sum x_x p(x) \log\left(\frac{q(x)}{p(x)}\right)$$

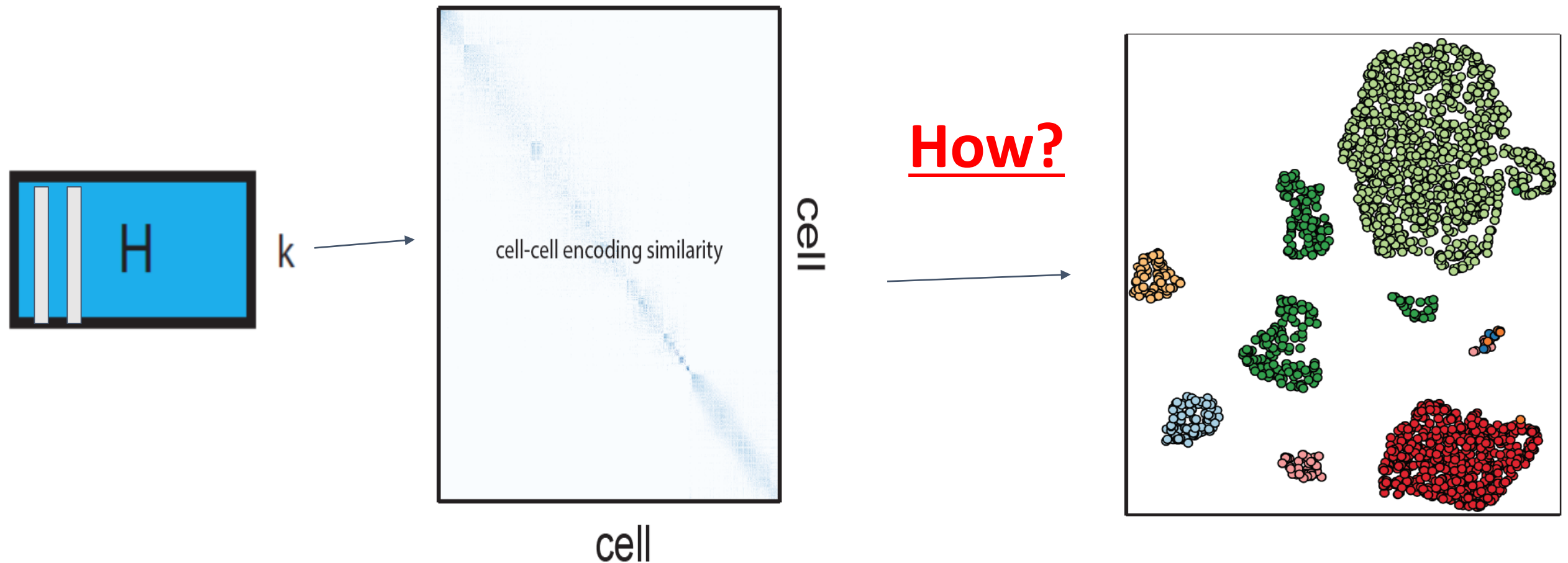
Jensen-Shannon divergence

$$\mathbf{JSD}(p \parallel q) = \frac{1}{2} \{ \mathbf{KLD}(p \parallel m) + \mathbf{KLD}(q \parallel m) \}, m = \frac{1}{2}(p + q)$$

Square-root of JSD is a metric (we love metric space ... Triangle inequality rocks! => Efficient proximity search)



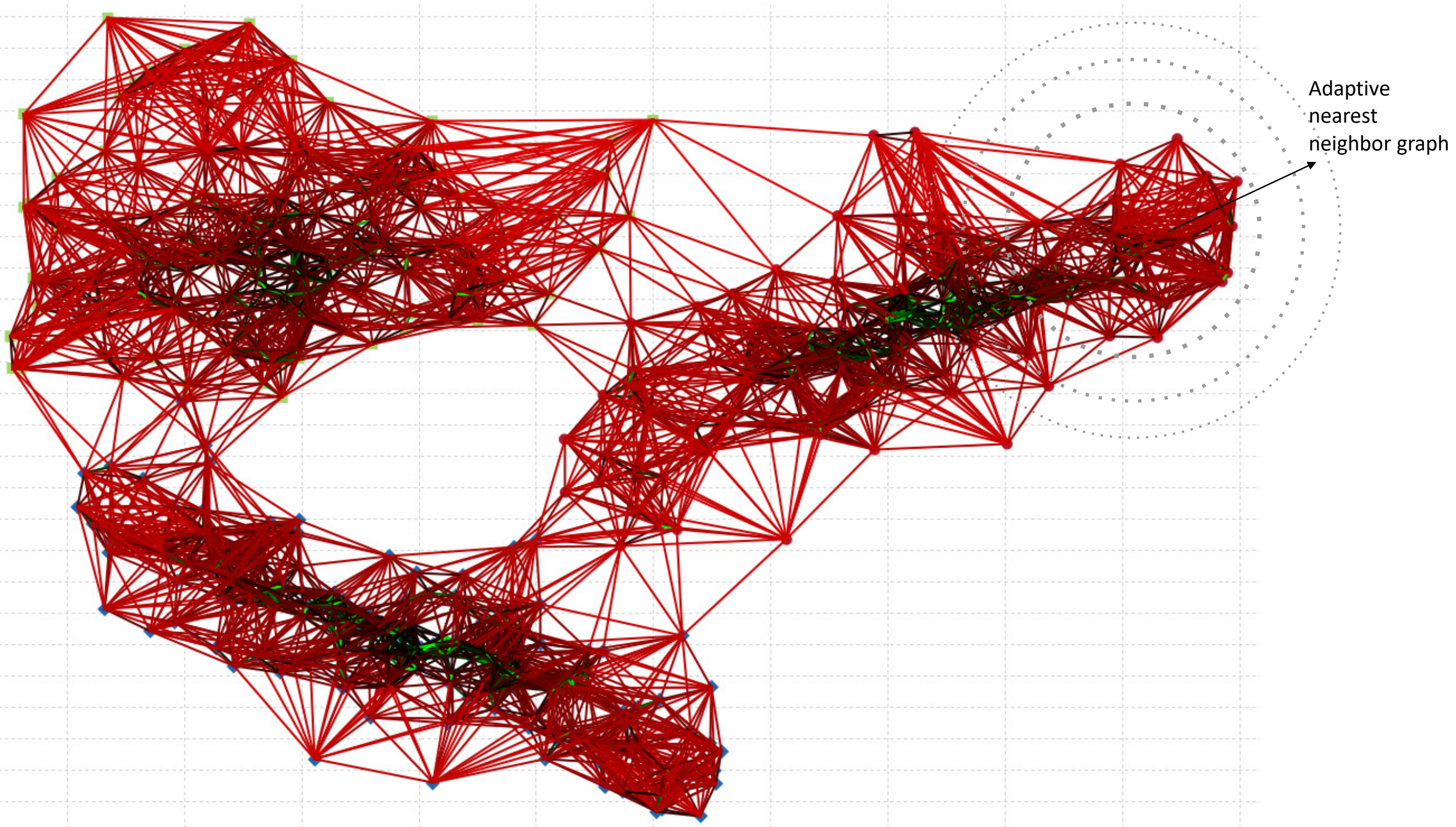
Step 2: Construct a network representation of the cell space



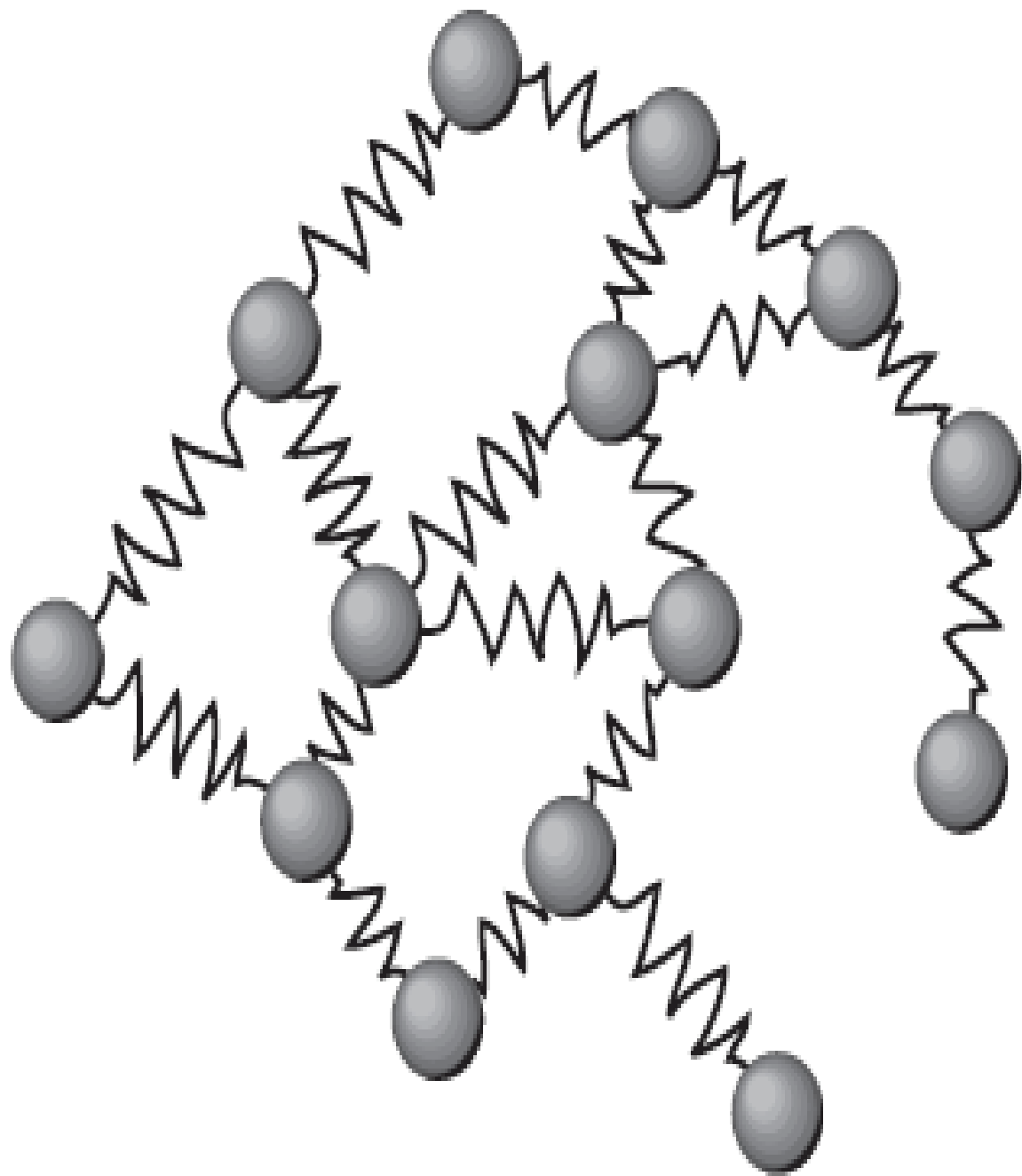
Density-dependent adaptive nearest neighbor graph

- Uses k^* -nearest neighbor algorithm
- Automatically identifies an optimal number of nearest neighbors for each cell
 - Depends on the heterogeneity of the neighbors

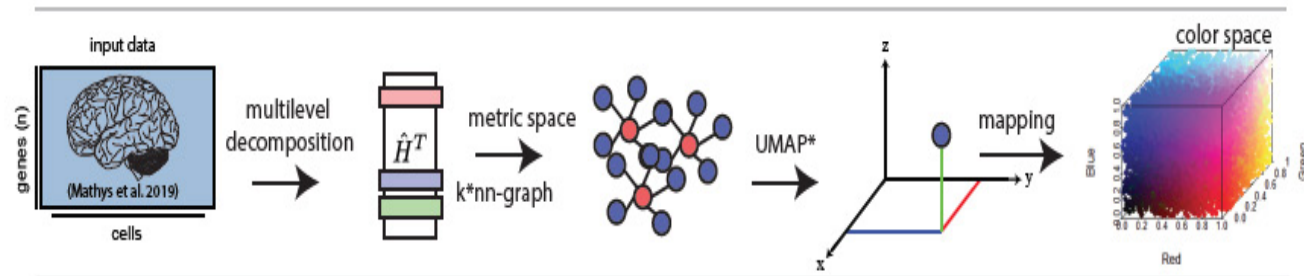
Step 2: Construct a network representation of the cell space



Step 3: Visualize cell-cell network (layout)

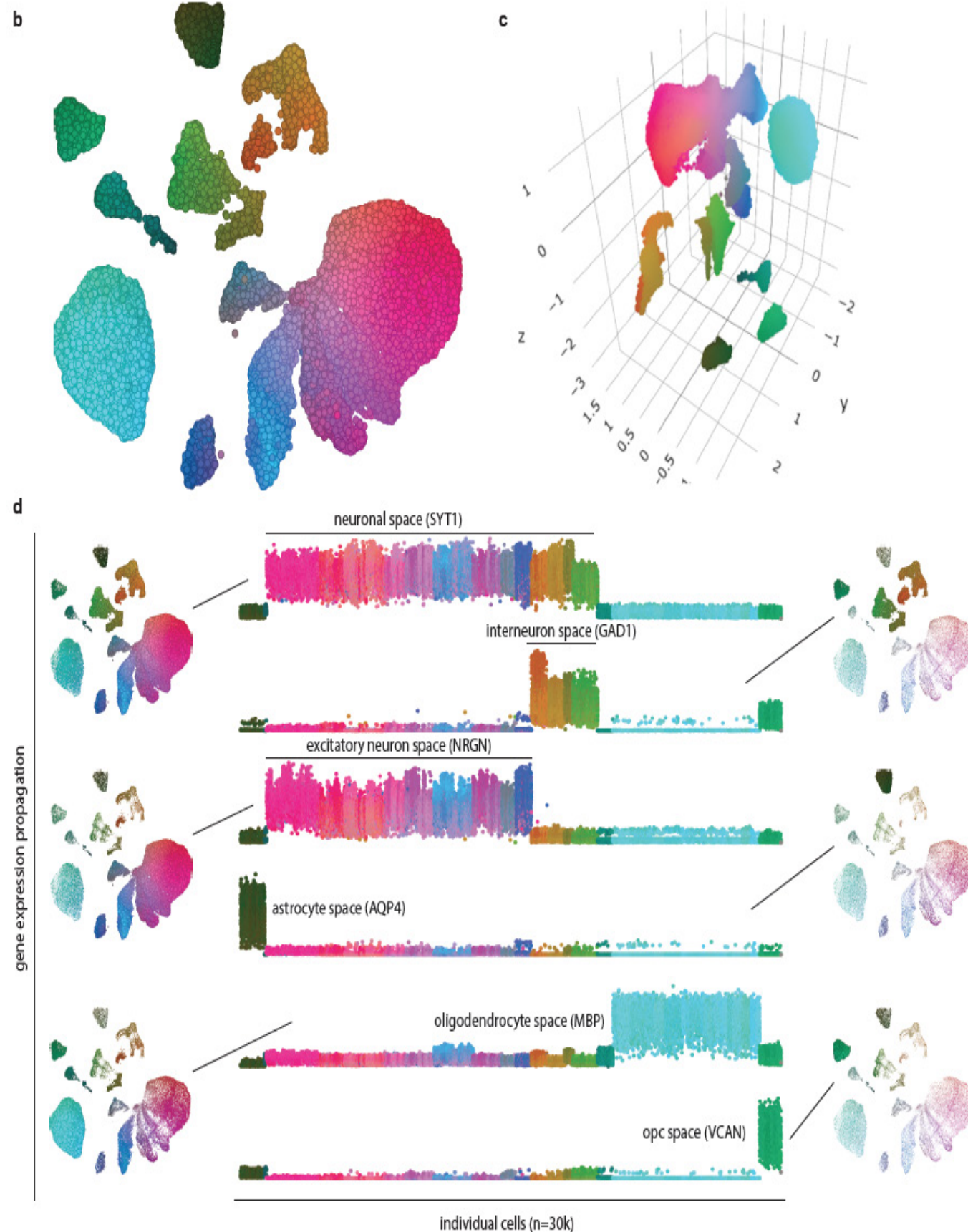


- Adopted from **UMAP** and reimplemented to work with the ACTIONet graph
- **Force-directed** layout
 - Stochastic-gradient descent (**SGD**)-based



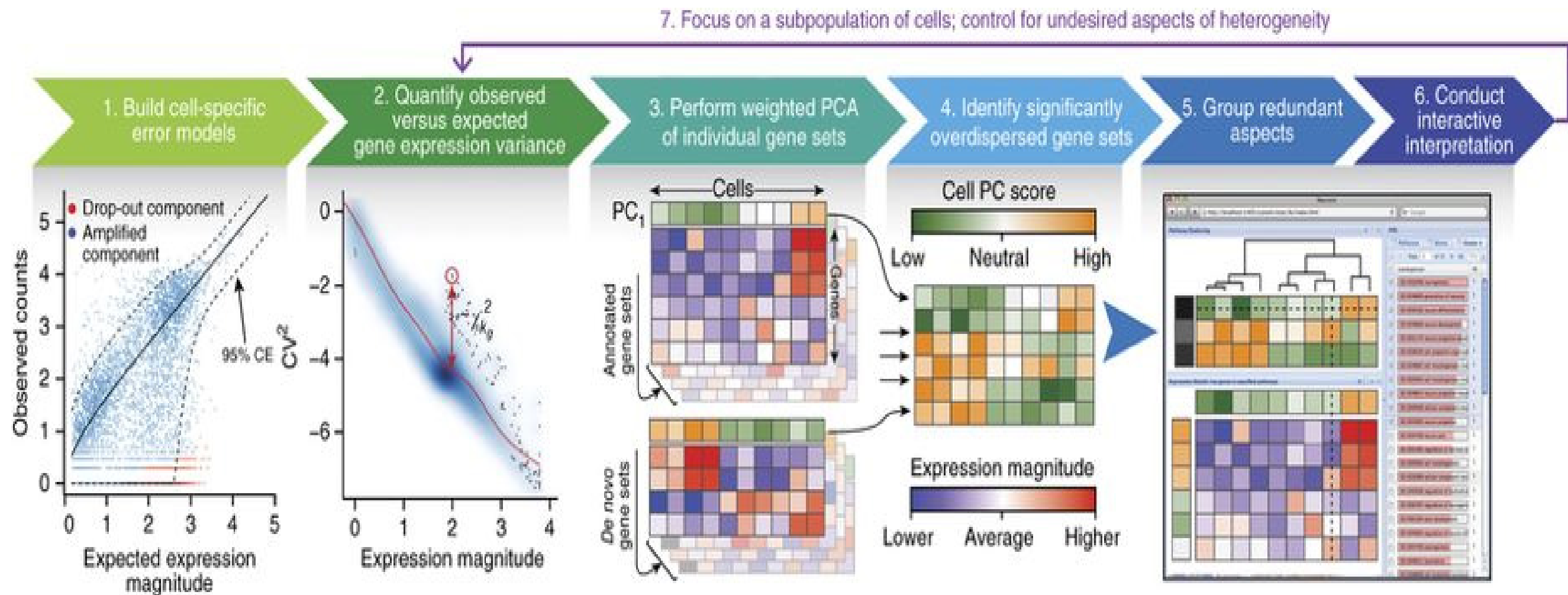
Step 4: Color-coding cells

- **Idea:** Use *de novo* coloring to fill the gap between 2D and 3D embeddings
- Projecting 3D coordinates onto a **Perceptually uniform** color space
 - CIE L*a*b*



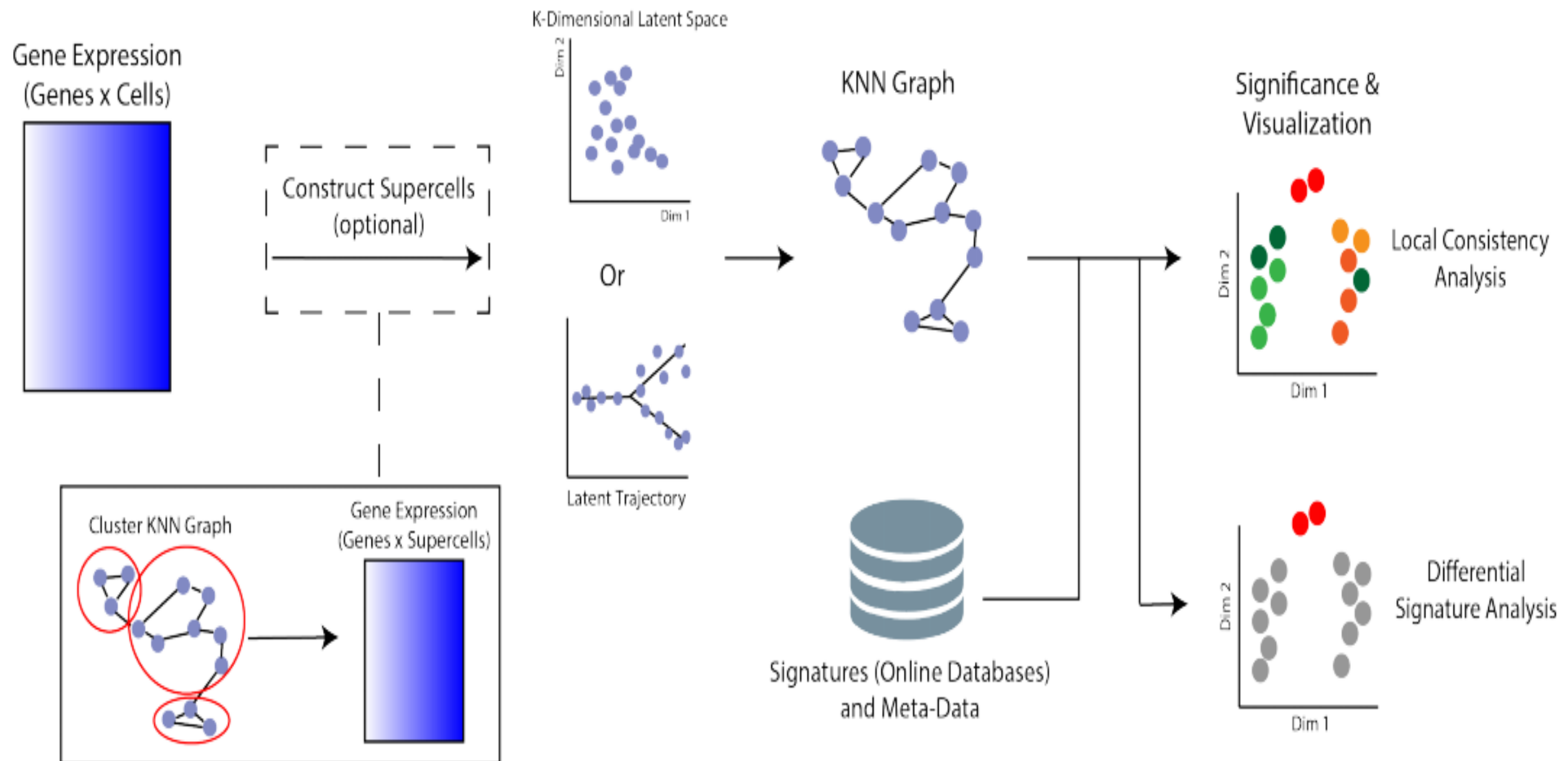
Interpreting cell-to-cell
variabilities using known
genesets/pathways

Pathway and gene set overdispersion analysis (PAGODA)



From: Fan *et al.*, 2016

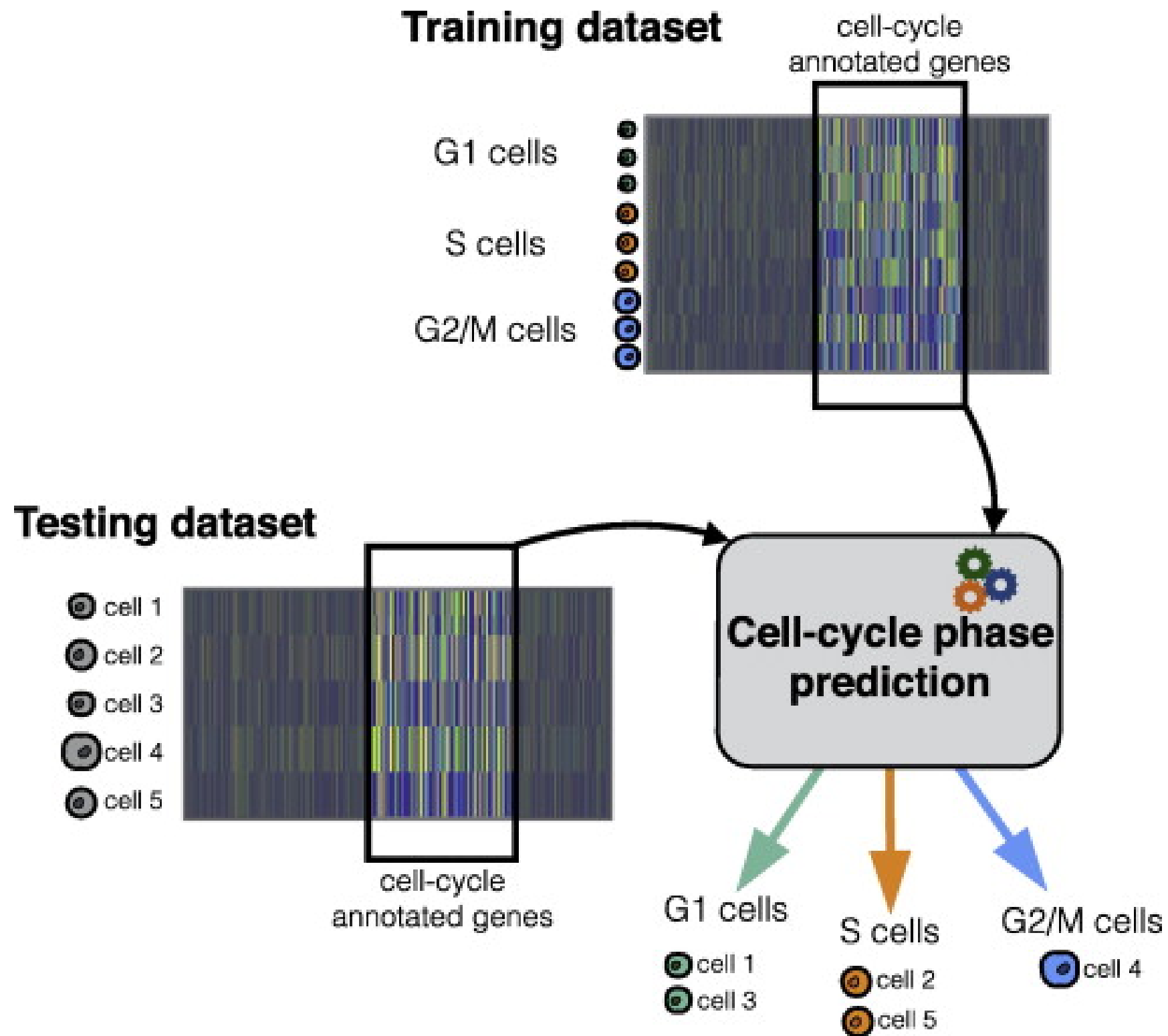
VISION method



From: DeTomaso *et al.*, 2018

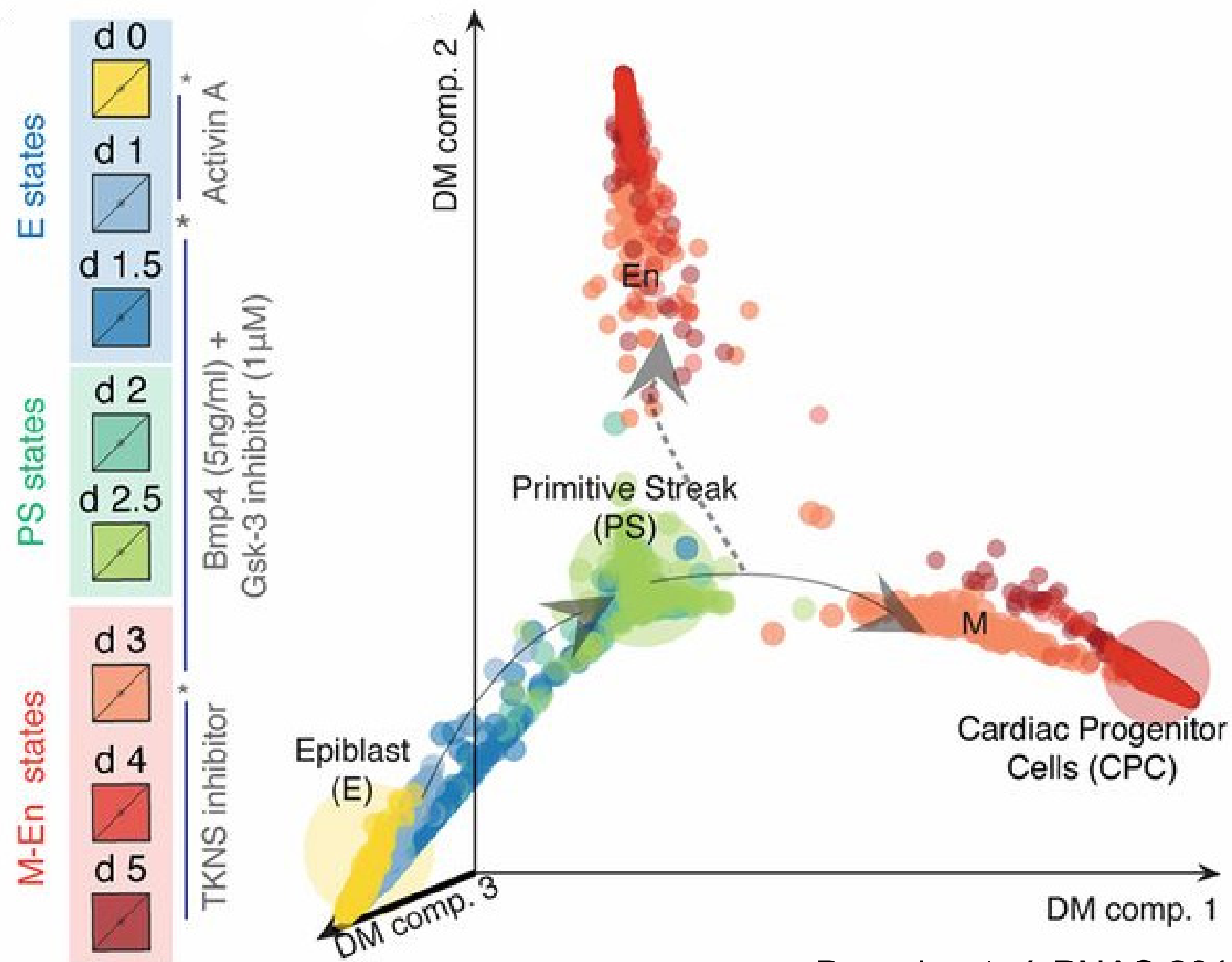
Trajectories through cell space

Cell-cycle phase prediction



Scialdone *et al.*, 2015

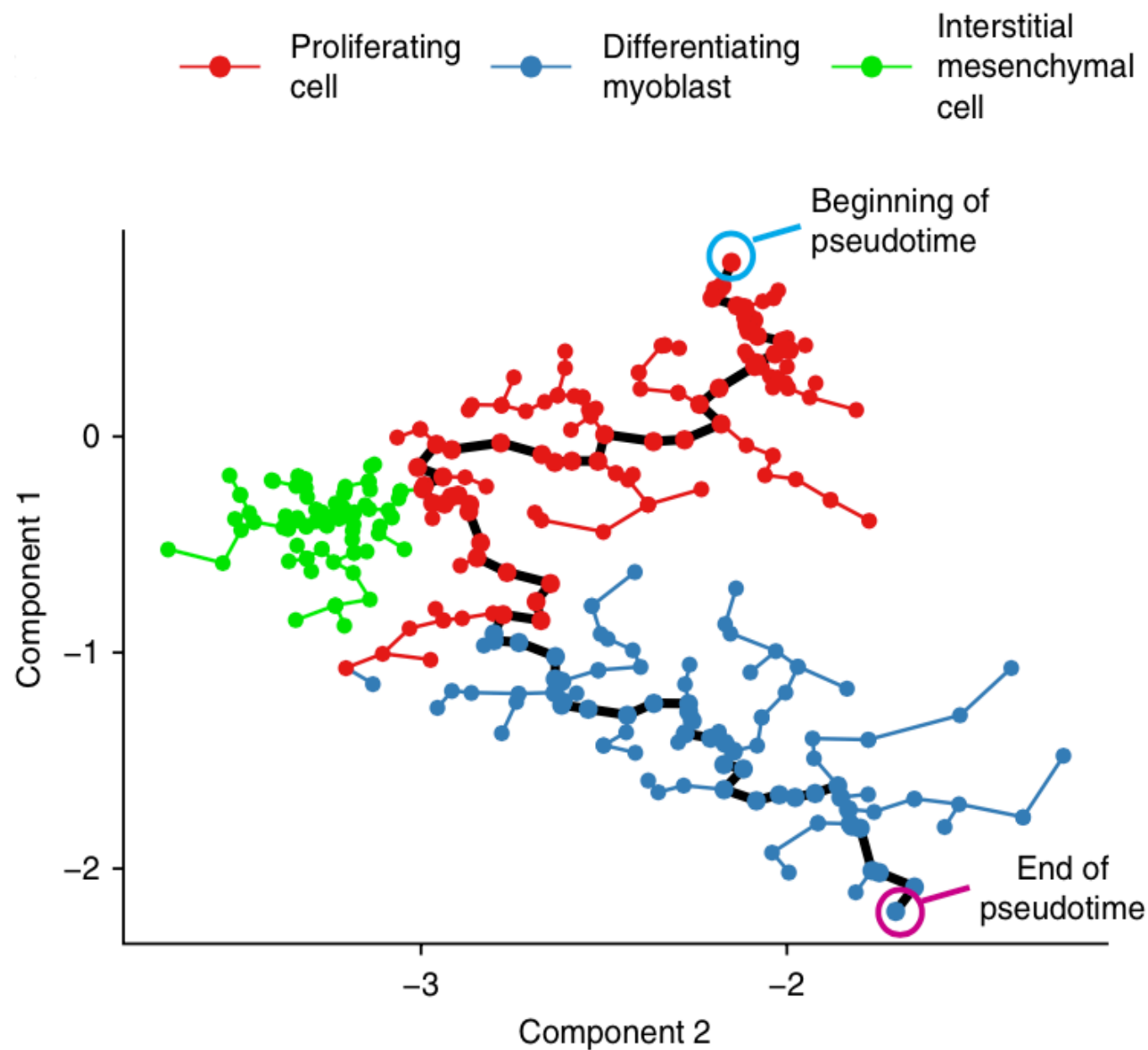
Trajectory inference



Bargaje *et al*, PNAS 2017

- Identify key branching points in development/disease
- Regulatory circuits that drive these transitions

Trajectory inference methods



- Start with dimension reduction
- Build a graph among cells/inferred cell types
 - Typically underlying structure is based on minimum spanning tree (**MST**) or k-nearest neighborhood (**kNN**) graph.
- Either infer a linear (pseudo-time) ordering, or identify branching points

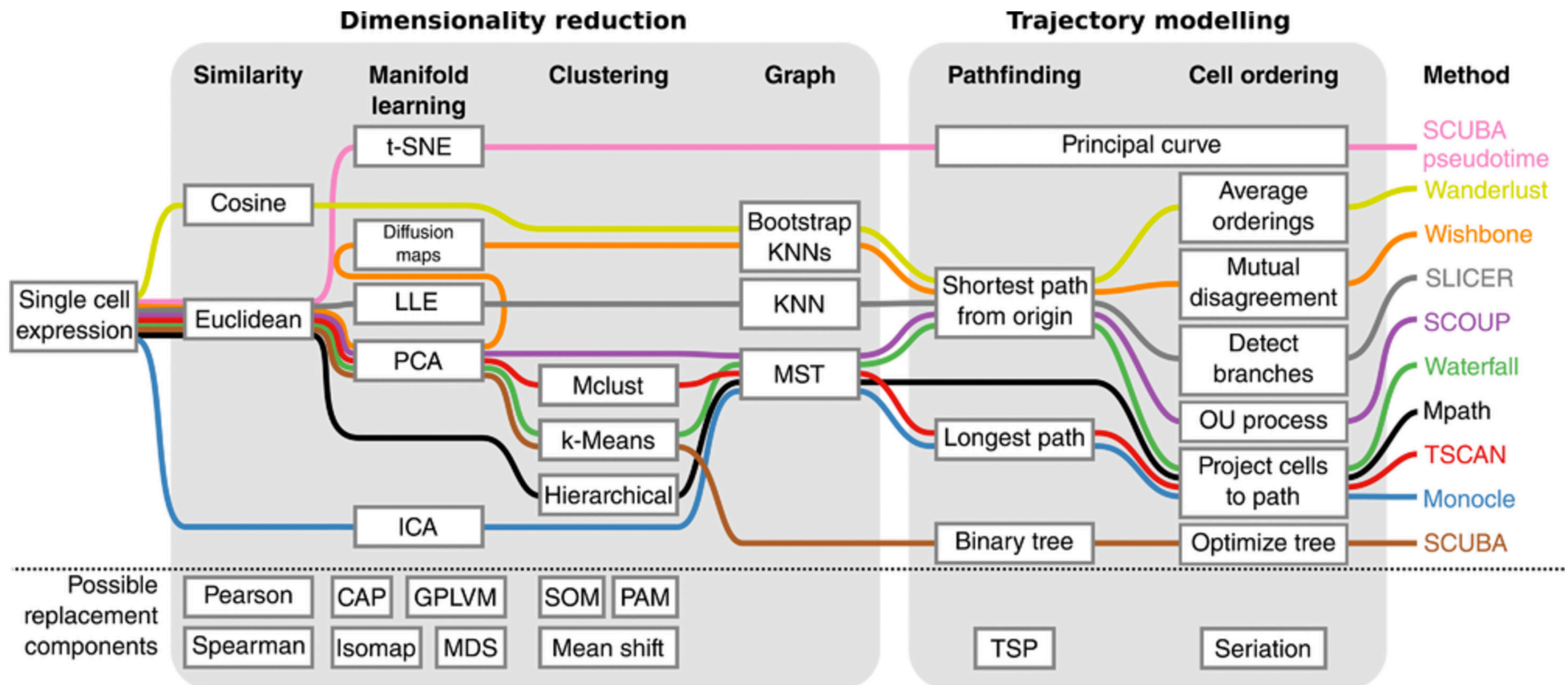
TSCAN pseudotime reconstruction with monocle

Overview of trajectory identification methods

Method	SCUBA pseudotime	Wanderlust	Wishbone	SLICER	SCOUP	Waterfall	Mpath	TSCAN	Monocle	SCUBA
Visual abstract										
Structure	Linear	Linear	Single bifurcation	Branching	Branching	Linear	Branching	Linear	Branching	Branching
Robustness strategy	Principal curves	Ensemble, starting cell	Ensemble, starting cell	Starting cell	Starting population	Clustering of cells	Clustering of cells using external labelling	Clustering of cells	Differential expression	Simple model
Extra input requirements	None	Starting cell	Starting cell	Starting cell	Starting population	None	Time points	None	Time points	Time points
Unbiased	+	±	±	±	±	+	-	+	-	-
Scalability w.r.t. cells	-	-	±	±	-	±	+	+	-	±
Scalability w.r.t. genes	+	+	+	+	-	+	±	±	±	+
Code and documentation	-	±	+	±	+	±	+	+	+	±
Parameter ease-of-use	+	+	+	+	-	±	-	+	+	+

First Author	Marco	Bendall	Setty	Welch	Matsumoto	Shin	Chen	Ji	Trapnell	Marco
Last Author	GC Yuan	Dana Pe'er	Dana Pe'er	Hartemink, Prins	Kiryu	Hongjun Song	Poidinger	Ji	Rinn	GC Yuan
Journal	PNAS	Cell	Nature Biotechnology	Genome Biology	BMC Bioinformatics	Cell Stem Cell	Nature Communications	NAR	Nature Biotechnology	PNAS
Year	2014	2014	2016	2016	2016	2015	2016	2016	2014	2014

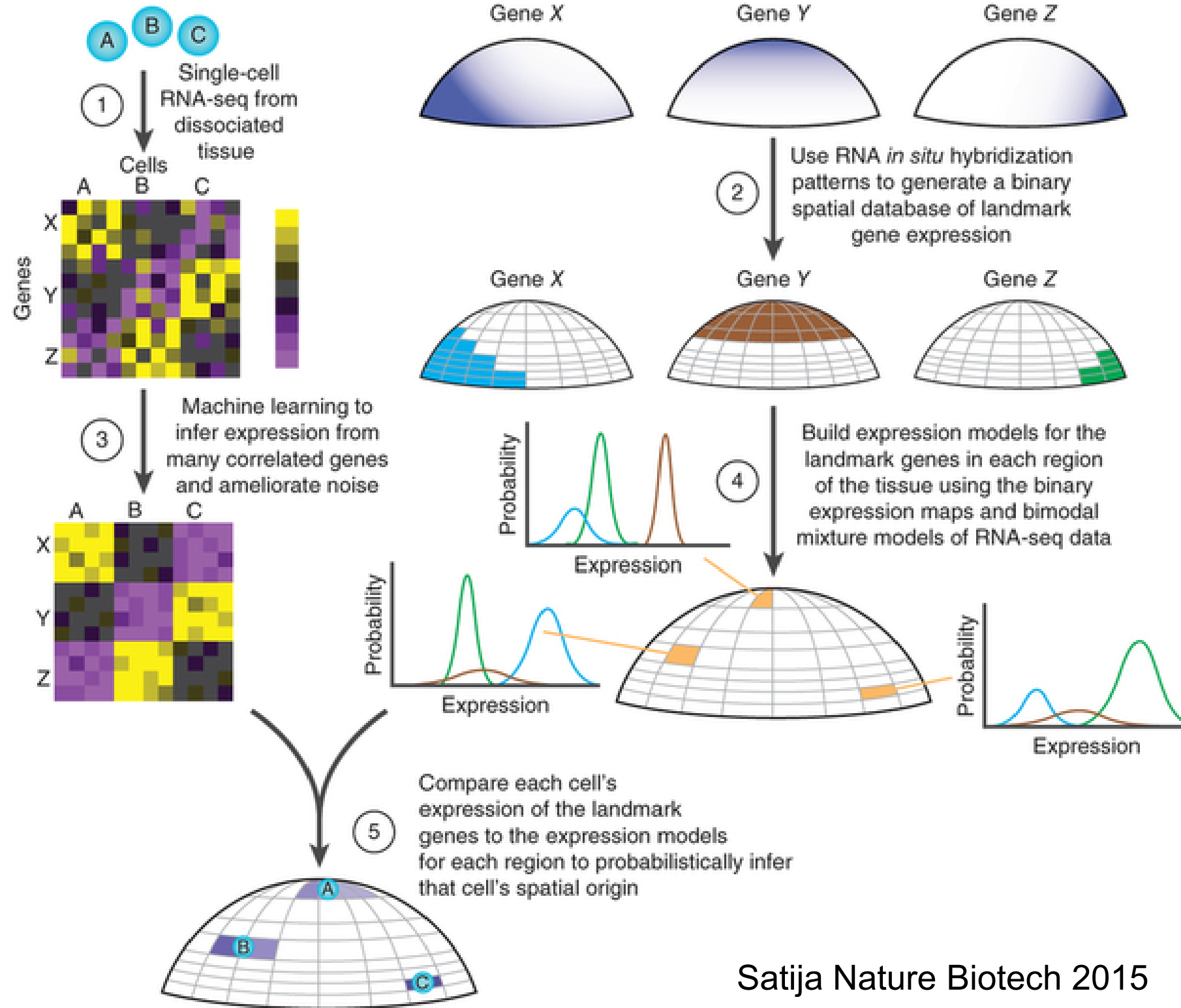
Trajectory identification: meta-method view



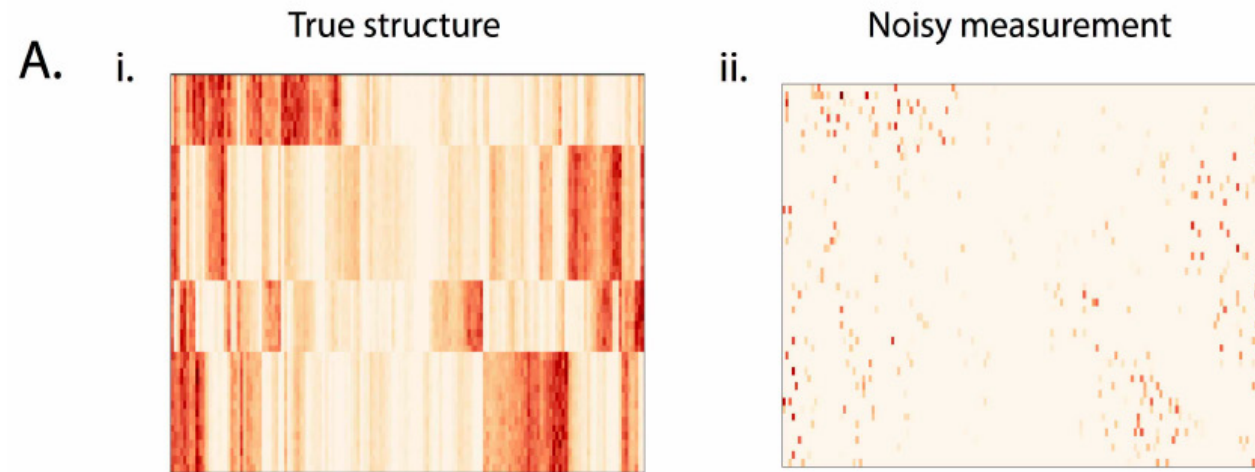
hemberg-lab.github.io/scRNA.seq.course

Dataset completion & missing data imputation

Spatial reconstruction of single-cell gene expression

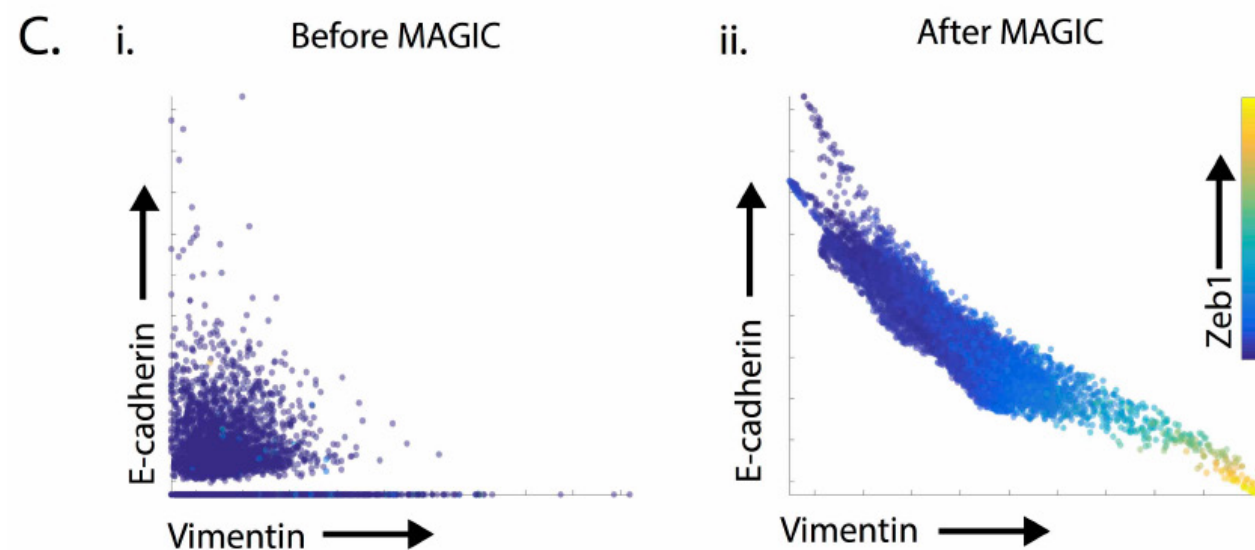
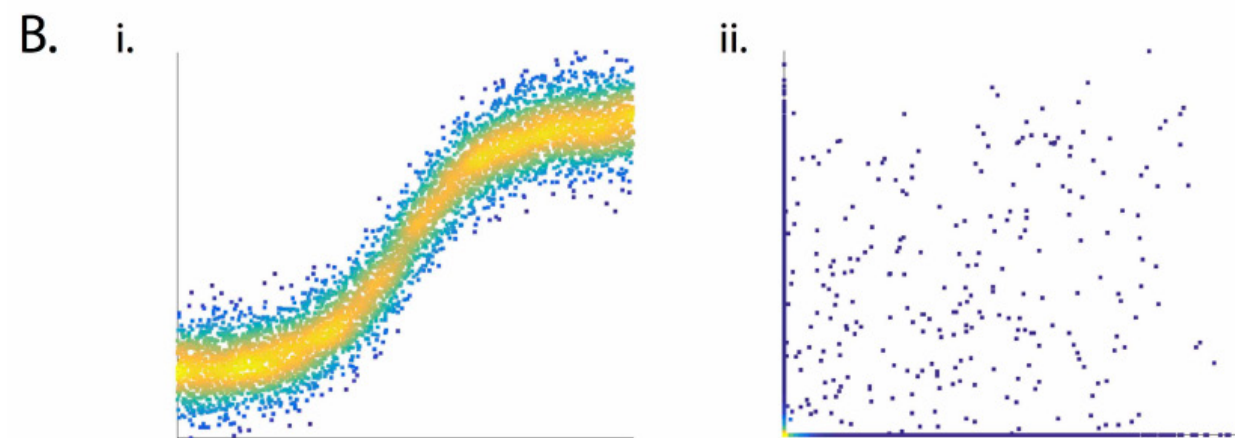


Missing value imputation with MAGIC (Markov Affinity-based Graph Imputation of Cells)



Random walk on cell-cell similarity graph

- uses neighborhood-based Markov-affinity matrix
- shares weight information across cells
- generate an imputed count matrix

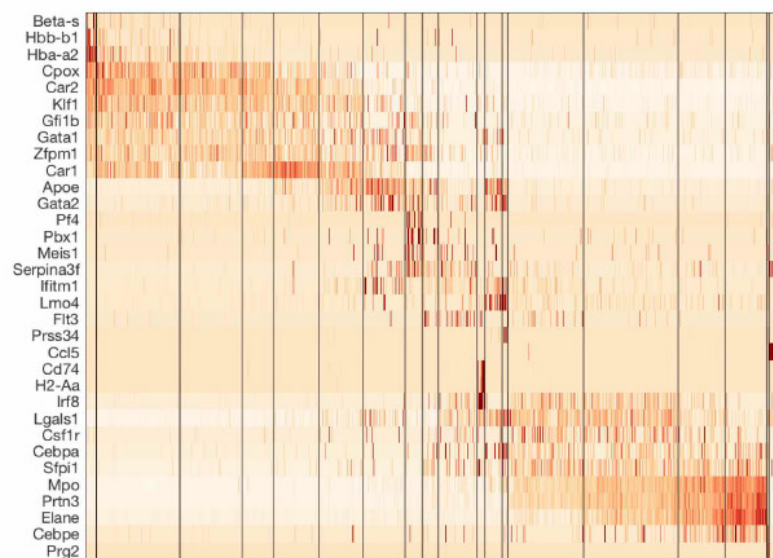


Imputation reveals gene-gene correlation patterns

A.

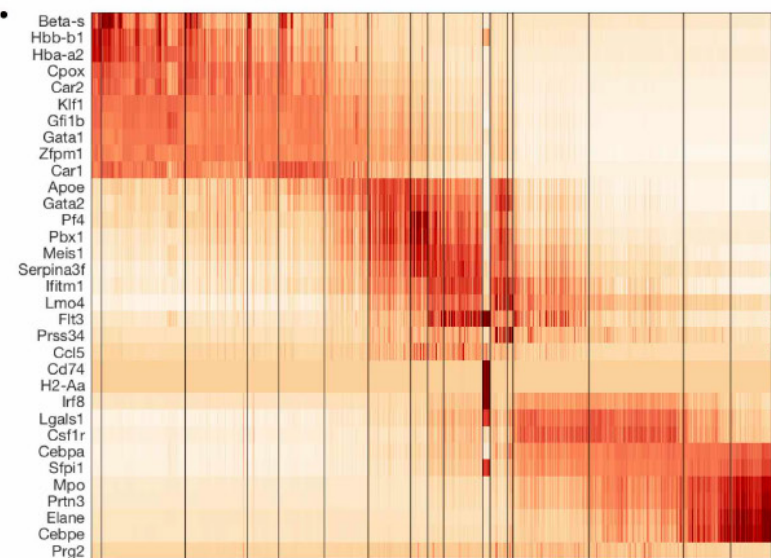
i.

Before MAGIC



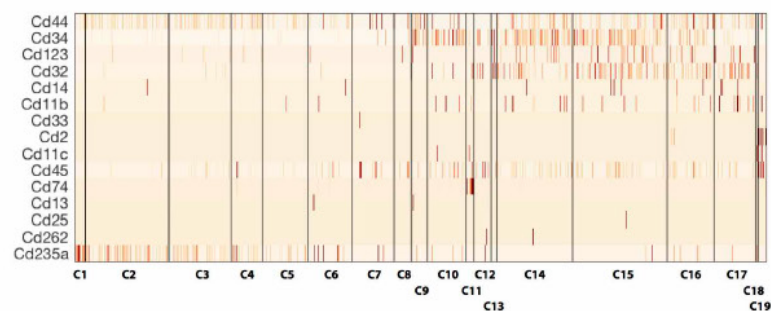
ii.

After MAGIC

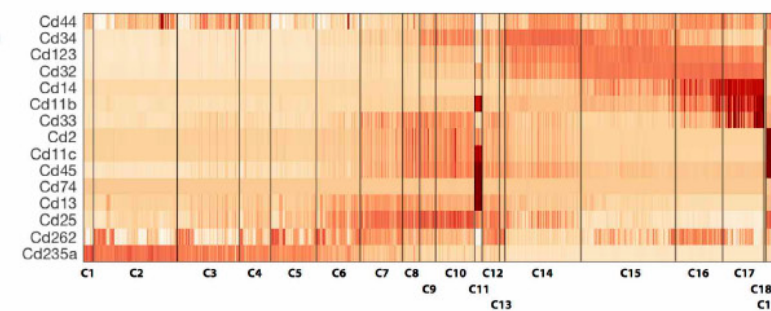


B.

i.



ii.



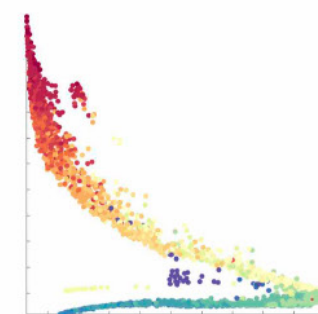
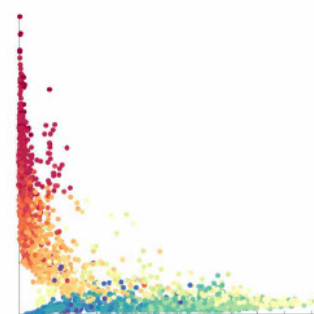
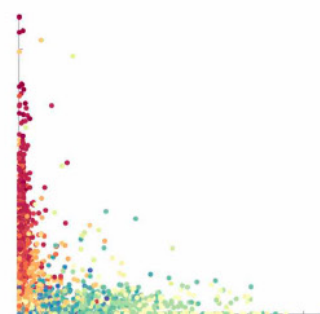
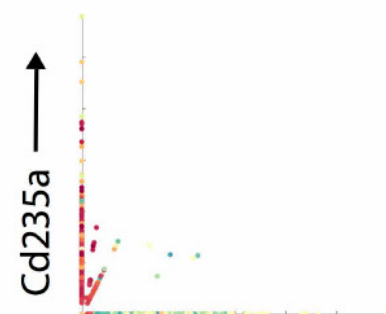
C.

t=0

t=1

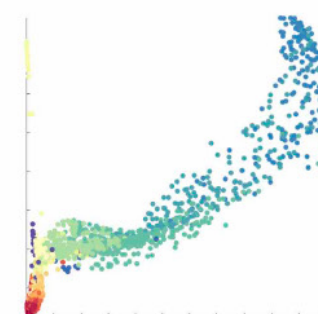
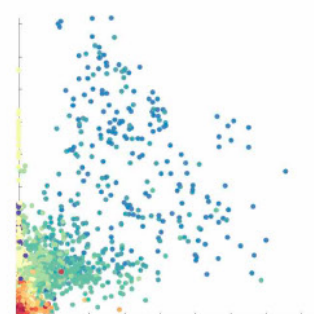
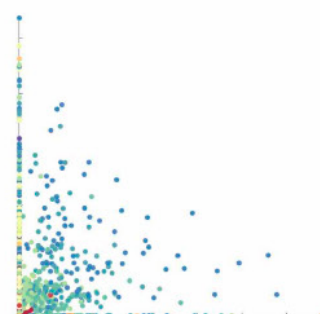
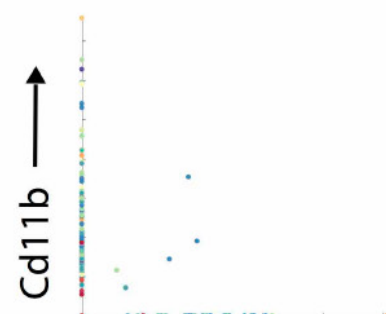
t=3

t=9



D.

Cd34



E.

Cd14

- erythrocyte C1
- erythrocyte C2
- erythrocyte C3
- erythrocyte C4
- erythrocyte C5
- erythrocyte C6
- early erythrocyte C7
- megakaryocyte C8
- early neutrophil C9
- early monocyte C10
- dendritic cells C11
- early basophil C12
- basophil C13
- monocyte C14
- monocyte C15
- neutrophil C16
- neutrophil C17
- eosinophil C18
- lymphoid progenitors (NK) C19

A: model-based imputation

bayNorm	[47]	Binomial model, empirical Bayes prior
BISCUIT	[48]	Gaussian model of log counts, cell- and cluster-specific parameters
CIDR	[49]	Decreasing logistic model (DO), non-linear least-squares regression (imp)
SAVER	[50]	NB model, Poisson LASSO regression prior
ScImpute	[51]	Mixture model (DO), non-negative least squares regression (imp)
scRecover	[52]	ZINB model (DO identification only)
VIPER	[53]	Sparse non-negative regression model

B: data smoothing

DrImpute	[54]	<i>k</i> -means clustering of PCs of correlation matrix
knn-smooth	[55]	<i>k</i> -nearest neighbor smoothing
LSImpute	[56]	Locality sensitive imputation
MAGIC	[57]	Diffusion across nearest neighbor graph
netSmooth	[58]	Diffusion across PPI network

C: data reconstruction, matrix factorization

ALRA	[59]	SVD with adaptive thresholding
ENHANCE	[60]	Denoising PCA with aggregation step
scRMD	[61]	Robust matrix decomposition
consensus NMF	[62]	Meta-analysis approach to NMF
f-scLVM	[63]	Sparse Bayesian latent variable model
GPLVM	[64]	Gaussian process latent variable model
pCMF	[65]	Probab. count matrix factorization with Poisson model
scCoGAPS	[66]	Extension of NMF
SDA	[67]	Sparse decomposition of arrays (Bayesian)
ZIFA	[68]	ZI factor analysis
ZINB-WaVE	[69]	ZINB factor model

C: data reconstruction, machine learning

AutoImpute	[70]	AE, no error back-propagation for zero counts
BERMUDA	[71]	AE for cluster batch correction (MMD and MSE loss function)
DeepImpute	[72]	AE, parallelized on gene subsets
DCA	[73]	Deep count AE (ZINB / NB model)
DUSC / DAWN	[74]	Denoising AE (PCA determines hidden layer size)
EnImpute	[75]	Ensemble learning consensus of other tools
Expression Saliency	[76]	AE (Poisson negative log-likelihood loss function)
LATE	[77]	Non-zero value AE (MSE loss function)
Lin_DAE	[78]	Denoising AE (imputation across <i>k</i> -nearest neighbor genes)
SAUCIE	[79]	AE (MMD loss function)
scScope	[80]	Iterative AE
scVAE	[81]	Gaussian-mixture VAE (NB / ZINB / ZIP model)
scVI	[82]	VAE (ZINB model)
scvis	[83]	VAE (objective function based on latent variable model and t-SNE)
VASC	[84]	VAE (denoising layer; ZI layer, double-exponential and Gumbel distribution)
Zhang_VAE	[85]	VAE (MMD loss function)

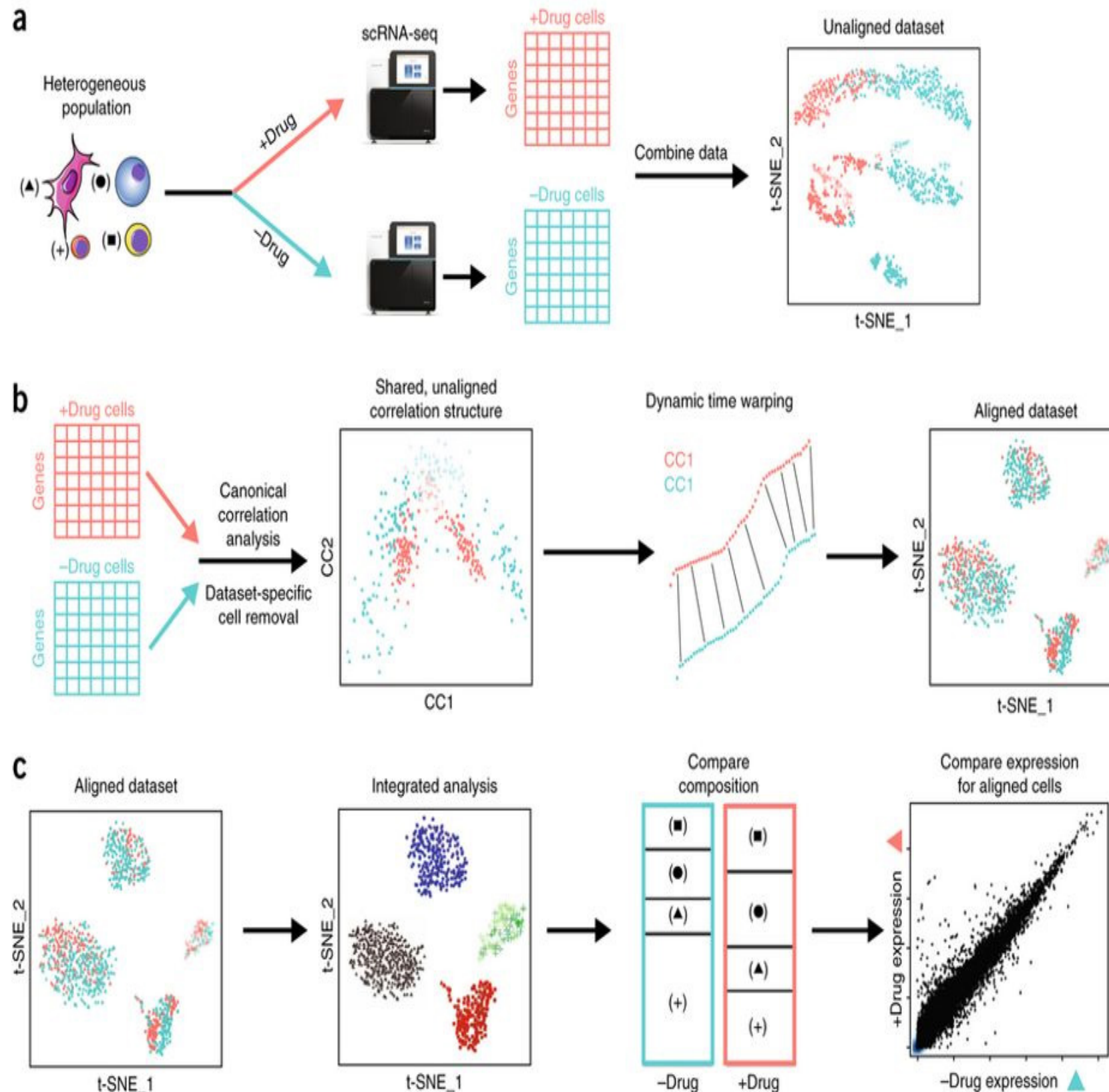
T: using external information

ADImpute	[86]	Gene regulatory network information
netSmooth	[58]	PPI network information
SAVER-X	[87]	Transfer learning with atlas-type resources
SCRABBLE	[88]	Matched bulk RNA-seq data
TRANSLATE	[77]	Transfer learning with atlas-type resources
URSM	[89]	Matched bulk RNA-seq data

47. Tang W, Bertaux F, Thomas P, Stefanelli C, Saint M, et al. bayNorm: Bayesian gene expression recovery, imputation and normalisation for single cell RNA-sequencing data. *bioRxiv*. 2018. <https://www.biorxiv.org/content/10.1101/384586v2.abstract>.
48. Azizi E, Prabhakaran S, Carr A, Pe'er D. Bayesian inference for single-cell clustering and imputing. *Genomics Comput Biol*. 2017;3(1):46. <https://doi.org/10.18547/gcb.2017.vol3.iss1.e46>. Accessed 27 Mar 2019.
49. Lin P, Troup M, Ho JWK. CIDR: ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol*. 2017;18(1):59. <https://doi.org/10.1186/s13059-017-1188-0>. Accessed 27 Mar 2019.
50. Huang M, Wang J, Torre E, Dueck H, Shaffer S, Bonasio R, Murray JL, Raj A, Li M, Zhang NR. SAVER: gene expression recovery for single-cell RNA sequencing. *Nat Methods*. 2018;15(7):539. <https://doi.org/10.1038/s41592-018-0033-z>. Accessed 27 Mar 2019.
51. Li WW, Li JJ. An accurate and robust imputation method scimpute for single-cell RNA-seq data. *Nat Commun*. 2018;9(1):997.
52. Miao Z, Li J, Zhang X. scRecover: discriminating true and false zeros in single-cell RNA-seq data for imputation. *bioRxiv*. 2019665323. <https://doi.org/10.1101/665323>. Accessed 15 Oct 2019.
53. Chen M, Zhou X. VIPER: variability-preserving imputation for accurate gene expression recovery in single-cell RNA sequencing studies. *Genome Biol*. 2018;19(1):196.
54. Gong W, Kwak I-Y, Pota P, Koyano-Nakagawa N, Garry DJ. DrImpute: imputing dropout events in single cell RNA sequencing data. *BMC Bioinformatics*. 2018;19(1):220. <https://doi.org/10.1186/s12859-018-2226-y>. Accessed 27 Mar 2019.
55. Wagner F, Yan Y, Yanai I. K-nearest neighbor smoothing for high-throughput single-cell RNA-Seq data. *bioRxiv*. 2018217737. <https://doi.org/10.1101/217737>. Accessed 15 Oct 2019.
56. Moussa M, Mândoiu IL. Locality sensitive imputation for single cell RNA-Seq data. *J Comput Biol*. 2019. <https://doi.org/10.1089/cmb.2018.0236>. Accessed 27 July 2019.
57. Dijk DV, Sharma R, Nainys J, Yim K, Kathail P, Carr AJ, Burdziak C, Moon KR, Chaffer CL, Pattabiraman D, Bieri B, Mazutis L, Wolf G, Krishnaswamy S, Pe'er D. Recovering gene interactions from single-cell data using data diffusion. *Cell*. 2018;174(3):716–72927. <https://doi.org/10.1016/j.cell.2018.05.061>. Accessed 27 Mar 2019.
58. Jonathan Ronen AA. netsmooth: network-smoothing based imputation for single cell RNA-seq. *F1000Res*. 2018;7: <https://github.com/BIMSBbioinfo/netSmooth>.
59. Linderman GC, Zhao J, Kluger Y. Zero-preserving imputation of scRNA-seq data using low-rank approximation. *bioRxiv*. 2018. <https://www.biorxiv.org/content/10.1101/397588v1.abstract>.
60. Wagner F, Barkley D, Yanai I. Accurate denoising of single-cell RNA-Seq data using unbiased principal component analysis. *bioRxiv*. 2019655365. URL <https://doi.org/10.1101/655365>. Accessed 15 Nov 2019.
61. Chen C, Wu C, Wu L, Wang Y, Deng M, Xi R. scRMD: imputation for single cell RNA-seq data via robust matrix decomposition. *bioRxiv*. 2018459404. <https://doi.org/10.1101/459404>. Accessed 15 Oct 2019.
62. Kotliar D, Veres A, Nagy MA, Tabrizi S, Hodis E, Melton DA, Sabeti PC. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *Elife*. 2019;8:43803.
63. Buettner F, Pratanwanich N, McCarthy DJ, Marioni JC, Stegle O. f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq. *Genome Biol*. 2017;18(1):212. <https://doi.org/10.1186/s13059-017-1334-8>.
64. Verma A, Engelhardt BE. A robust nonlinear low-dimensional manifold for single cell RNA-seq data. *bioRxiv*. 2018443044. <https://doi.org/10.1101/443044>. Accessed 15 Nov 2019.
65. Durif G, Modolo L, Mold JE, Lambert-Lacroix S, Picard F. Probabilistic count matrix factorization for single cell expression data analysis. *Bioinformatics*. 2019. <https://doi.org/10.1093/bioinformatics/btz177>.
66. Stein-O'Brien GL, Clark BS, Sherman T, Zibetti C, Hu Q, Sealfon R, Liu S, Qian J, Colantuoni C, Blackshaw S, Goff LA, Fertig EJ. Decomposing cell identity for transfer learning across cellular measurements, platforms, tissues, and species. *Cell Syst*. 2019;8(5):395–4118. <https://doi.org/10.1016/j.cels.2019.04.004>.
67. Jung M, Wells D, Rusch J, Ahmad S, Marchini J, Myers SR, Conrad DF. Unified single-cell analysis of testis gene regulation and pathology in five mouse strains. *eLife*. 2019;8. URL <https://doi.org/10.7554/eLife.43966>.
68. Pierson E, Yau C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol*. 2015;16(1):241. <https://doi.org/10.1186/s13059-015-0805-z>. Accessed 27 Mar 2019.
69. Risso D, Perraudeau F, Gribkova S, Dudoit S, Vert J-P. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Commun*. 2018;9(1):284. <https://doi.org/10.1038/s41467-017-02554-5>.
70. Talwar D, Mongia A, Sengupta D, Majumdar A. AutoImpute: autoencoder based imputation of single-cell RNA-seq data. *Sci Rep*. 2018;8(1):16329. <https://doi.org/10.1038/s41598-018-34688-x>.
71. Wang T, Johnson TS, Shao W, Lu Z, Helm BR, Zhang J, Huang K. BERMUDA: a novel deep transfer learning method for single-cell RNA sequencing batch correction reveals hidden high-resolution cellular subtypes. *Genome Biol*. 2019;20(1):165. <https://doi.org/10.1186/s13059-019-1764-6>. Accessed 15 Nov 2019.
72. Arisdakessian C, Poirion O, Yunis B, Zhu X, Garmire L. DeepImpute: an accurate, fast and scalable deep neural network method to impute single-cell RNA-Seq data. *bioRxiv*. 2018. <https://www.biorxiv.org/content/10.1101/353607v1.abstract>.
73. Eraslan G, Simon LM, Mircea M, Mueller NS, Theis FJ. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat Commun*. 2019;10(1):390. <https://doi.org/10.1038/s41467-018-07931-2>. Accessed 27 Mar 2019.
74. Srinivasan S, Johnson NT, Korkin D. A hybrid deep clustering approach for robust cell type profiling using single-cell RNA-seq data. *bioRxiv*. 2019. <https://www.biorxiv.org/content/10.1101/511626v1.abstract>.
75. Zhang X-F, Ou-Yang L, Yang S, Zhao X-M, Hu X, Yan H. EnImpute: imputing dropout events in single cell RNA sequencing data via ensemble learning. *Bioinformatics*. 2019. <https://doi.org/10.1093/bioinformatics/btz435>.
76. Kinalis S, Nielsen FC, Winther O, Bagger FO. Deconvolution of autoencoders to learn biological regulatory modules from single cell mRNA sequencing data. *BMC Bioinformatics*. 2019;20(1):379. <https://doi.org/10.1186/s12859-019-2952-9>.
77. Badsha MB, Li R, Liu B, Li YI, Xian M, Banovich NE, Fu AQ. Imputation of single-cell gene expression with an autoencoder neural network. *bioRxiv*. 2018504977. <https://doi.org/10.1101/504977>. Accessed 15 Oct 2019.
78. Lin C, Jain S, Kim H, Bar-Joseph Z. Using neural networks for reducing the dimensions of single-cell RNA-Seq data. *Nucleic Acids Res*. 2017;45(17):156.
79. Amodio M, Dijk DV, Srinivasan K, Chen WS, Mohsen H, Moon KR, Campbell A, Zhao Y, Wang X, Venkataswamy M, Desai A, Ravi V, Kumar P, Montgomery R, Wolf G, Krishnaswamy S. Exploring single-cell data with deep multitasking neural networks. *bioRxiv*. 2019237065. <https://doi.org/10.1101/237065>. Accessed 15 Oct 2019.
80. Deng Y, Bao F, Dai Q, Wu LF, Altschuler SJ. Scalable analysis of cell-type composition from single-cell transcriptomics using deep recurrent learning. *Nat Methods*. 2019. <https://doi.org/10.1038/s41592-019-0353-7>.
81. Grønbech CH, Vording MF, Timshel P, Sønderby CK, Pers TH, Winther O. scVAE: Variational auto-encoders for single-cell gene expression data. *bioRxiv*. 2019318295. <https://doi.org/10.1101/318295>. Accessed 15 Oct 2019.
82. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. *Nat Methods*. 2018;15(12):1053–8. <https://doi.org/10.1038/s41592-018-0229-2>.
83. Ding J, Condon A, Shah SP. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nat Commun*. 2018;9(1):2002.
84. Wang D, Gu J. VASC: Dimension reduction and visualization of single-cell RNA-seq data by deep variational autoencoder. *Genomics Proteomics Bioinforma*. 2018;16(5):320–31.
85. Zhang C. Single-cell data analysis using mmd variational autoencoder for a more informative latent representation. *bioRxiv*. 2019613414. <https://doi.org/10.1101/613414>. Accessed 15 Oct 2019.
86. Leote AC, Wu X, Beyer A. Network-based imputation of dropouts in single-cell RNA sequencing data. *bioRxiv*. 2019611517. URL <https://doi.org/10.1101/611517>. Accessed 23 Apr 2019.
87. Wang J, Agarwal D, Huang M, Hu G, Zhou Z, Ye C, Zhang NR. Data denoising with transfer learning in single-cell transcriptomics. *Nat Methods*. 2019;16(9):875–8. <https://doi.org/10.1038/s41592-019-0537-1>. Accessed 15 Oct 2019.
88. Peng T, Zhu Q, Yin P, Tan K. SCRABBLE: single-cell RNA-seq imputation constrained by bulk RNA-seq data. *Genome Biol*. 2019;20(1):88. <https://doi.org/10.1186/s13059-019-1681-8>.
89. Zhu L, Lei J, Devlin B, Roeder K. A unified statistical framework for single cell and bulk RNA sequencing data. *Ann Appl Stat*. 2018;12(1):609–32. <https://doi.org/10.1214/17-AOAS1110>. Accessed 15 Nov 2019.

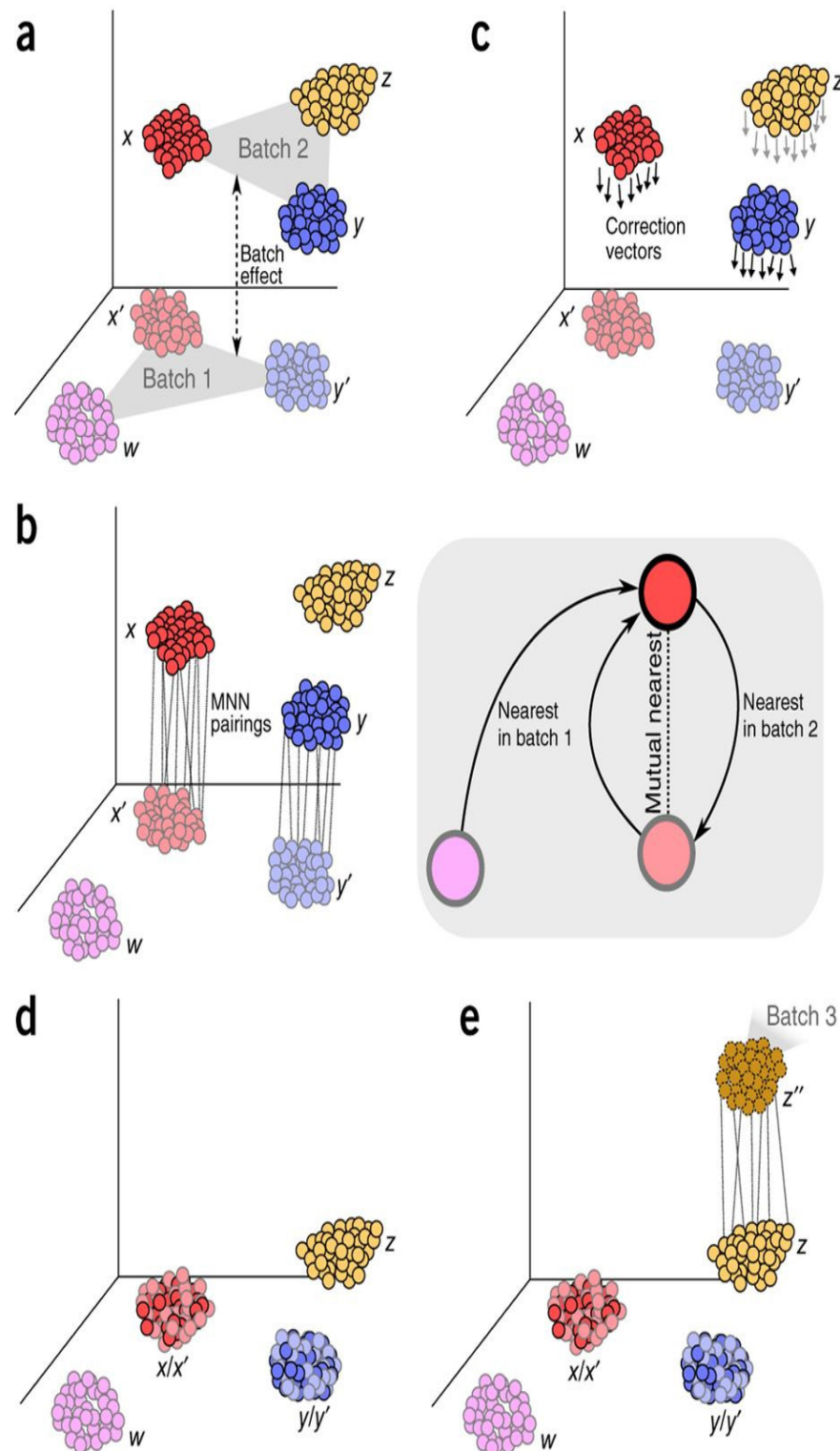
Integrating multiple single-cell datasets

Canonical Correlation Analysis (CCA)



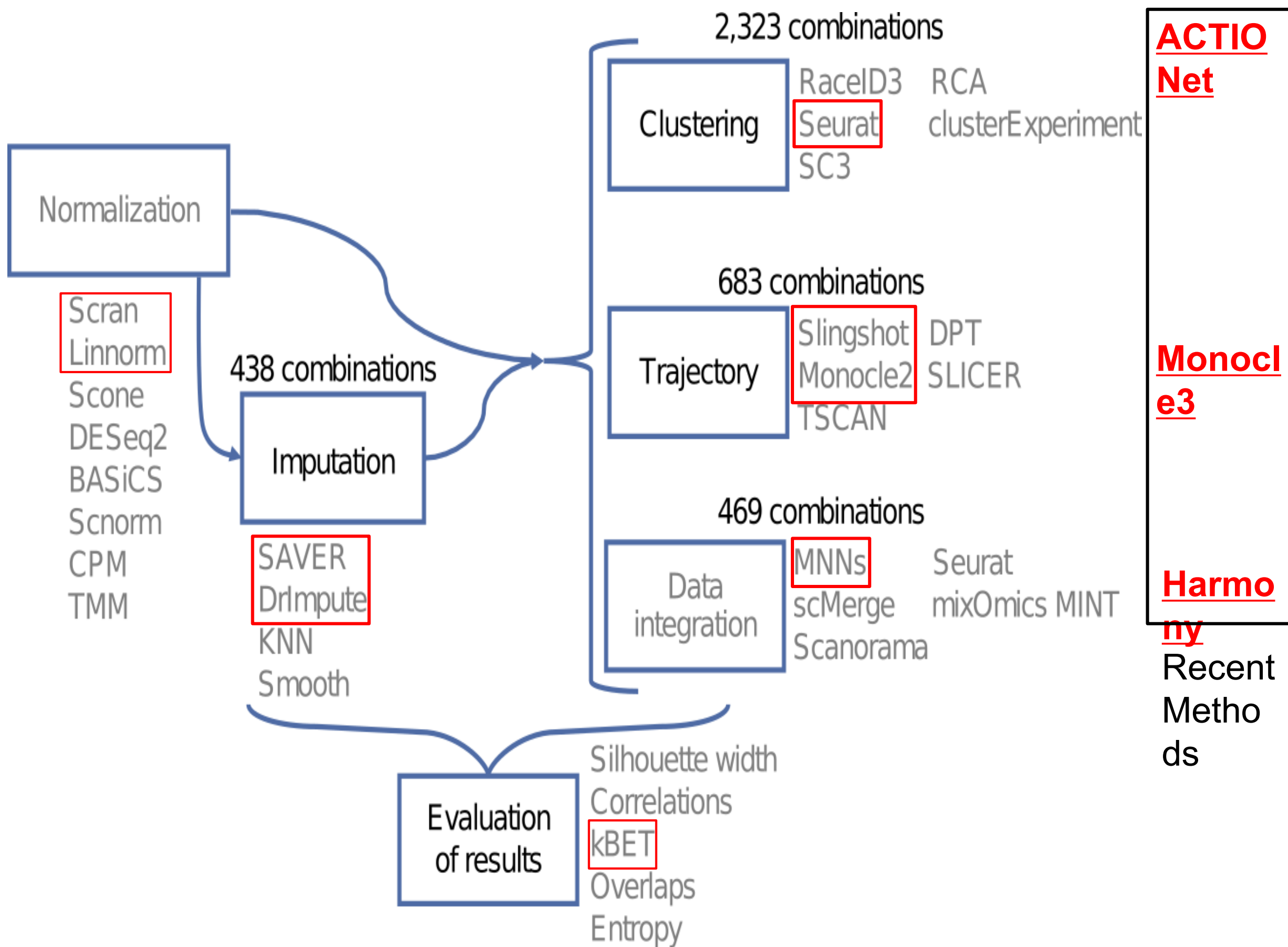
From: Butlet *et al.*, 2018

Mutual nearest neighbors (MNN) correction



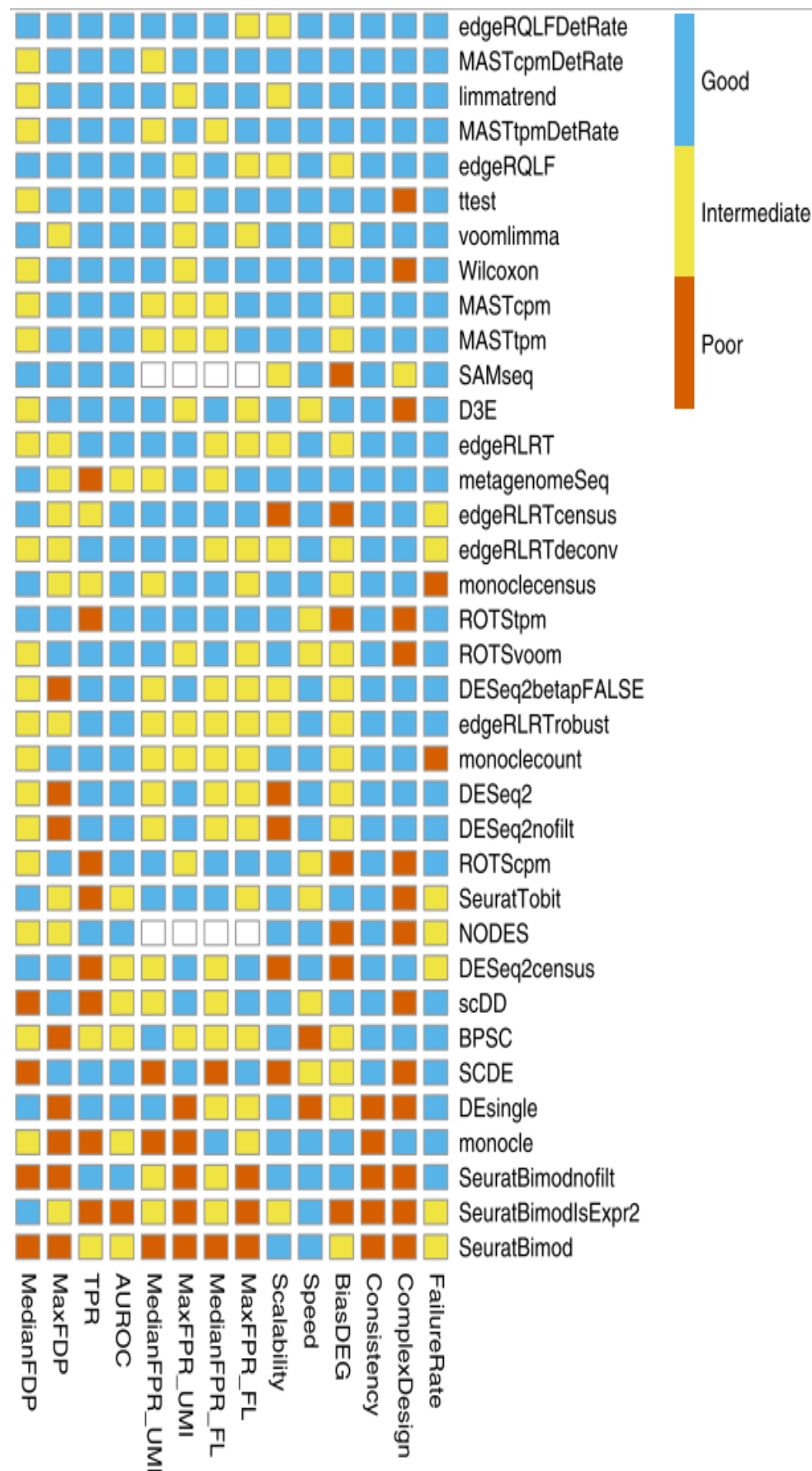
From: Haghighverdi *et al.*, 2018

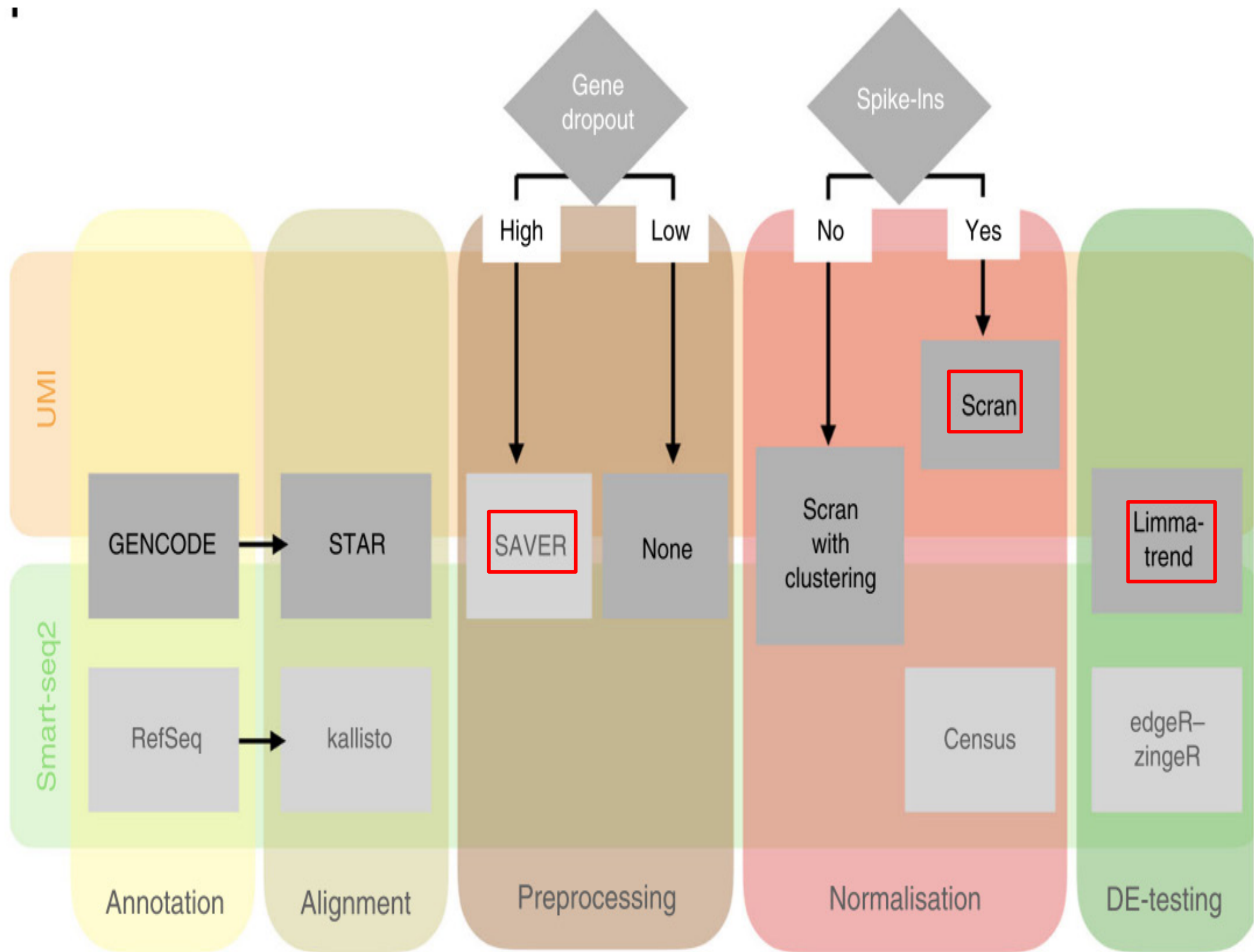
Summary and Method comparison



Comparison of differential expression methods

Found Limma-trend, MAST, edgeR, also t-test and Wilcoxon to perform well





Summary

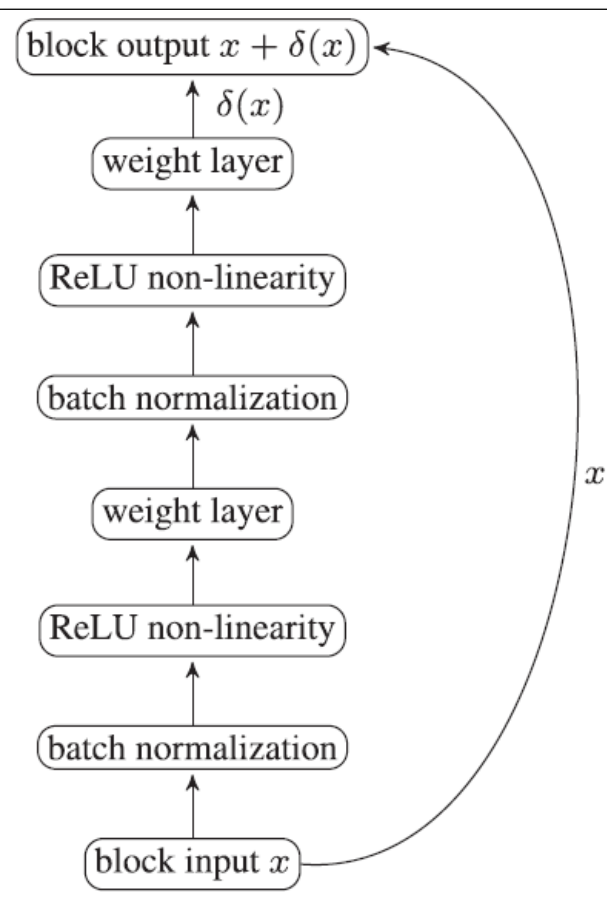
- Normalization
 - Scrان and Linnorm
- Imputation
 - SAVER
- Batch-correction
 - [fast]MNN and Harmony
- Clustering
 - ACTIONet and Seurat
- Trajectory detection
 - Monocle3 and Slingshot
- Differential expression
 - Limma-trend

Today: Deep Learning for Single-cell Genomics

1. Why single cells, traditional approaches, scRNA-seq
2. Scaling up single-cell technologies: evolution of scRNA-seq
3. Dealing with noise, doublets, and other scRNA-seq issues
4. Computational challenges in scRNA-seq data analysis
5. Deep learning methods for scRNA-seq data analysis
6. Beyond scRNA-seq: scATAC-seq, multi-omics

5. Deep Learning methods for scRNA-seq

MMD-ResNet: Autoencoder for batch correction



MMD (Gretton *et al.*, 2006, 2012) is a measure for distance between two probability distributions p, q . It is defined with respect to a function class $\mathcal{F}$ by

$$\text{MMD}(\mathcal{F}, p, q) \equiv \sup_{f \in \mathcal{F}} (\mathbb{E}_{x \sim p} f(x) - \mathbb{E}_{x \sim q} f(x)).$$

When $\mathcal{F}$ is a reproducing kernel Hilbert space with kernel k , the MMD can be written as the distance between the mean embeddings of p and q

$$\text{MMD}^2(\mathcal{F}, p, q) = \|\mu_p - \mu_q\|_{\mathcal{F}}^2, \quad (1)$$

where $\mu_p(t) = \mathbb{E}_{x \sim p} k(x, t)$. Equation (1) can be written as

$$\text{MMD}^2(\mathcal{F}, p, q) = \mathbb{E}_{x, x' \sim p} k(x, x') - 2\mathbb{E}_{x \sim p, y \sim q} k(x, y) + \mathbb{E}_{y, y' \sim q} k(y, y'), \quad (2)$$

where x and x' are independent, and so are y and y' . Importantly, if k is a universal kernel, then $\text{MMD}(\mathcal{F}, p, q) = 0$ iff $p = q$. In practice, the distributions p, q are unknown, and instead we are given observations $X = \{x_1, \dots, x_n\}, Y = \{y_1, \dots, y_m\}$, so that the (biased) sample version of (2) becomes

$$\begin{aligned} \text{MMD}^2(\mathcal{F}, X, Y) = & \frac{1}{n^2} \sum_{x_i, x_j \in X} k(x_i, x_j) \\ & - \frac{2}{nm} \sum_{x_i \in X, y_j \in Y} k(x_i, y_j) \\ & + \frac{1}{m^2} \sum_{y_i, y_j \in Y} k(y_i, y_j). \end{aligned}$$

Maximum Mean Discrepancy (MMD) loss function

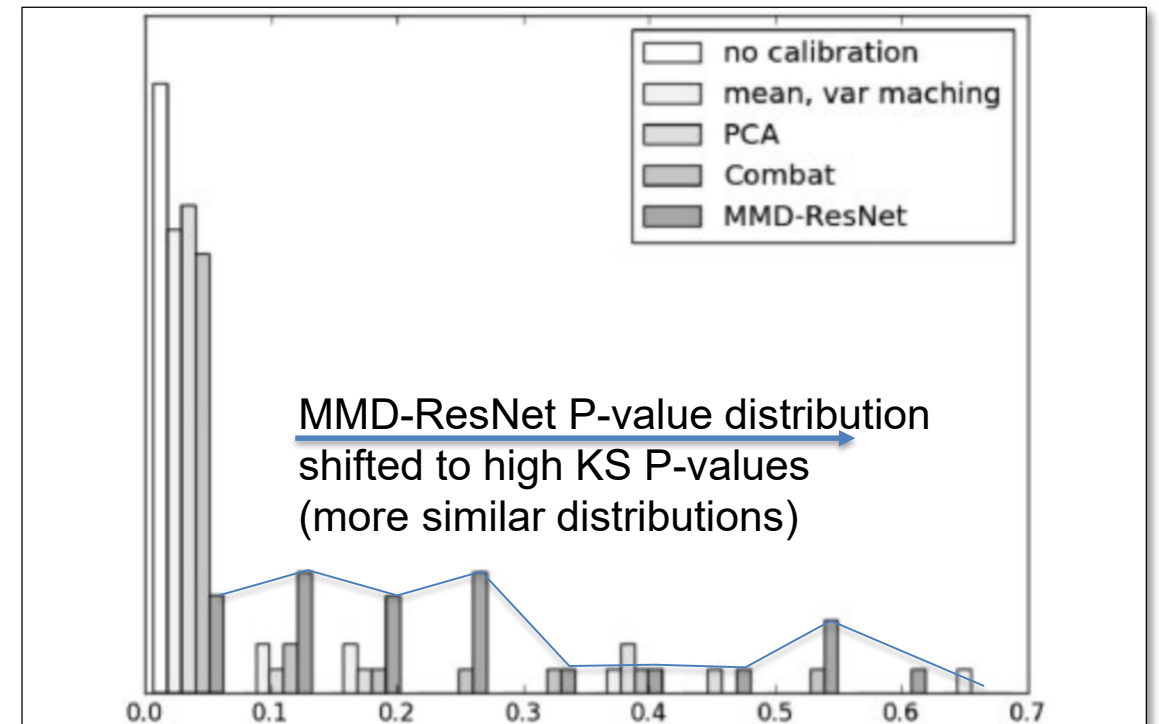


Fig. 5. Histograms of the 25 P -values of Kolmogorov-Smirnov tests, comparing the distributions of the calibrated data with the target distribution of each of the 25 markers

MMD-ResNet outperforms PCA, Combat, and

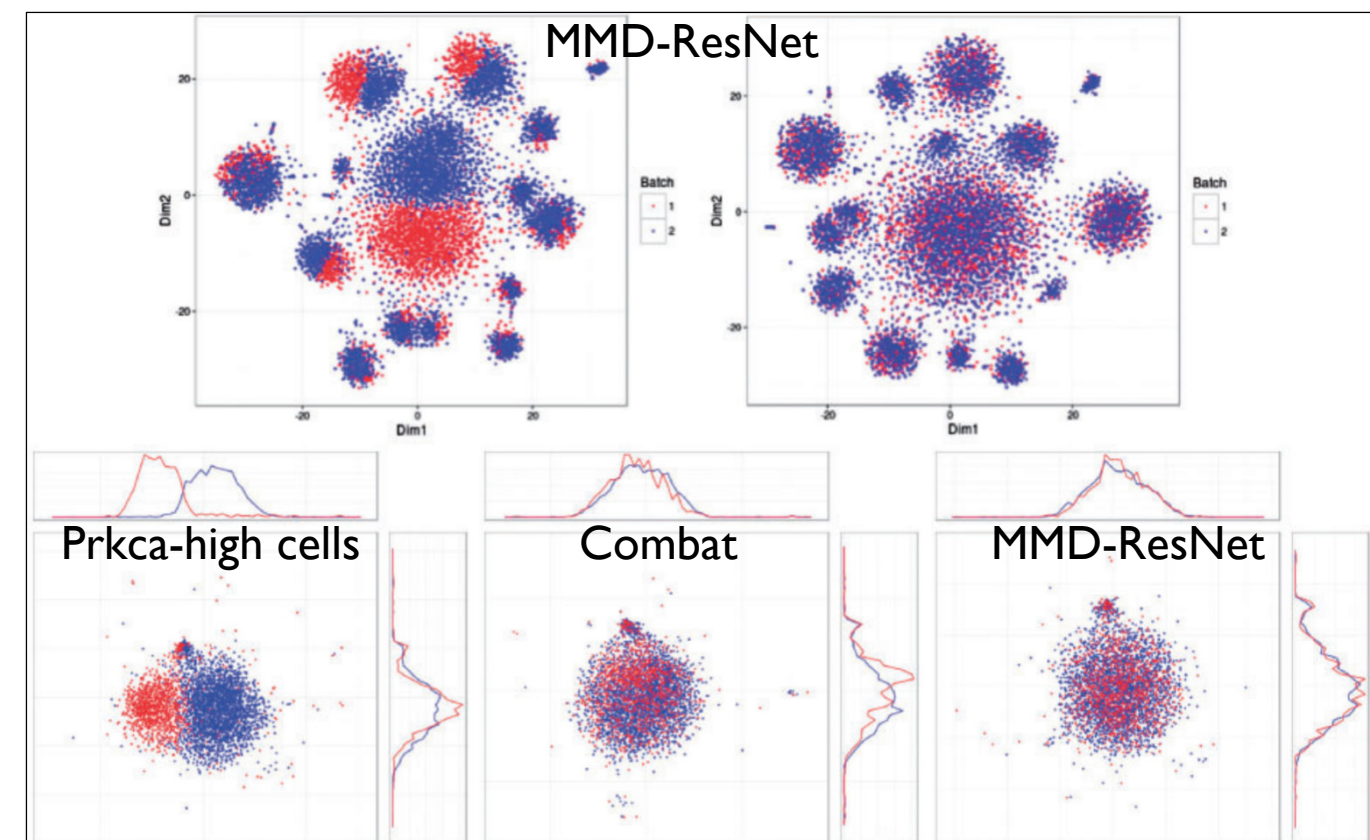
$$L(w) = \sqrt{\text{MMD}^2(\{\hat{\psi}(x_1), \dots, \hat{\psi}(x_n)\}, \{y_1, \dots, y_m\})},$$

Train ResNet with loss MMD score function

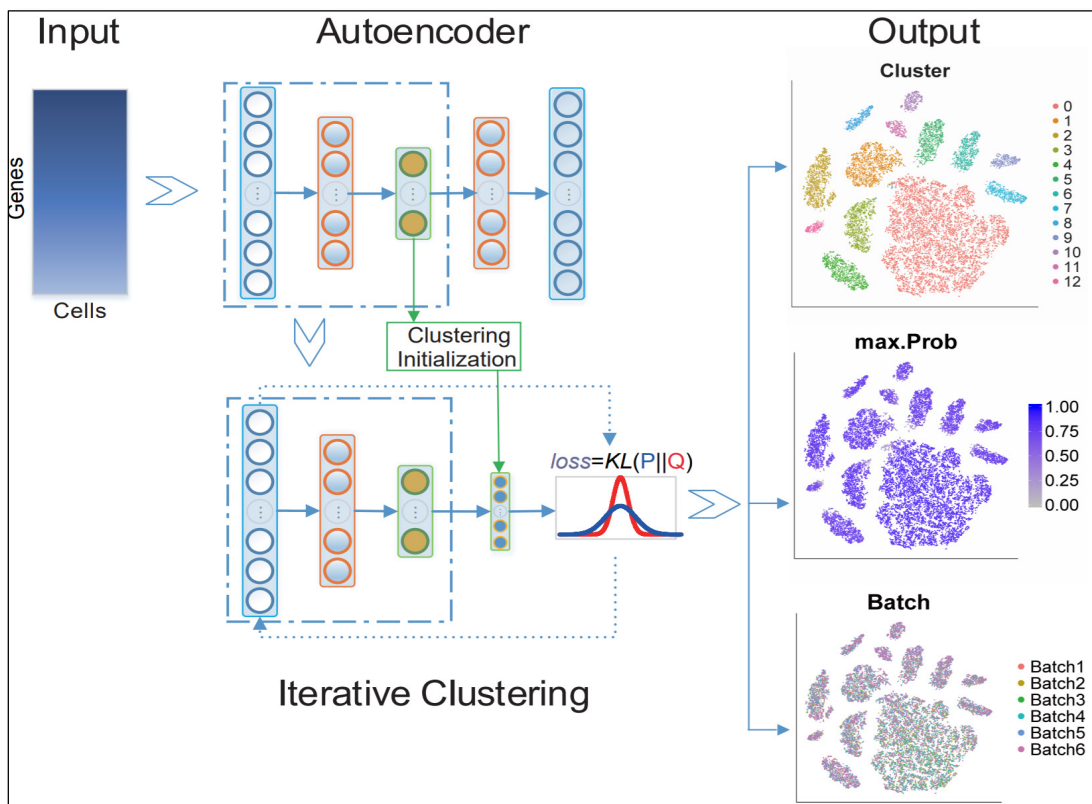
Removal of batch effects using distribution-matching residual networks

Uri Shaham^{1,†}, Kelly P. Stanton^{2,3,†}, Jun Zhao³, Huamin Li⁴, Khadir Raddassi⁵, Ruth Montgomery⁶ and Yuval Kluger^{2,3,4,*}

Shaham et al., Bioinf, 2017



DESC: Deep embedding for cell-type-specific batch correction

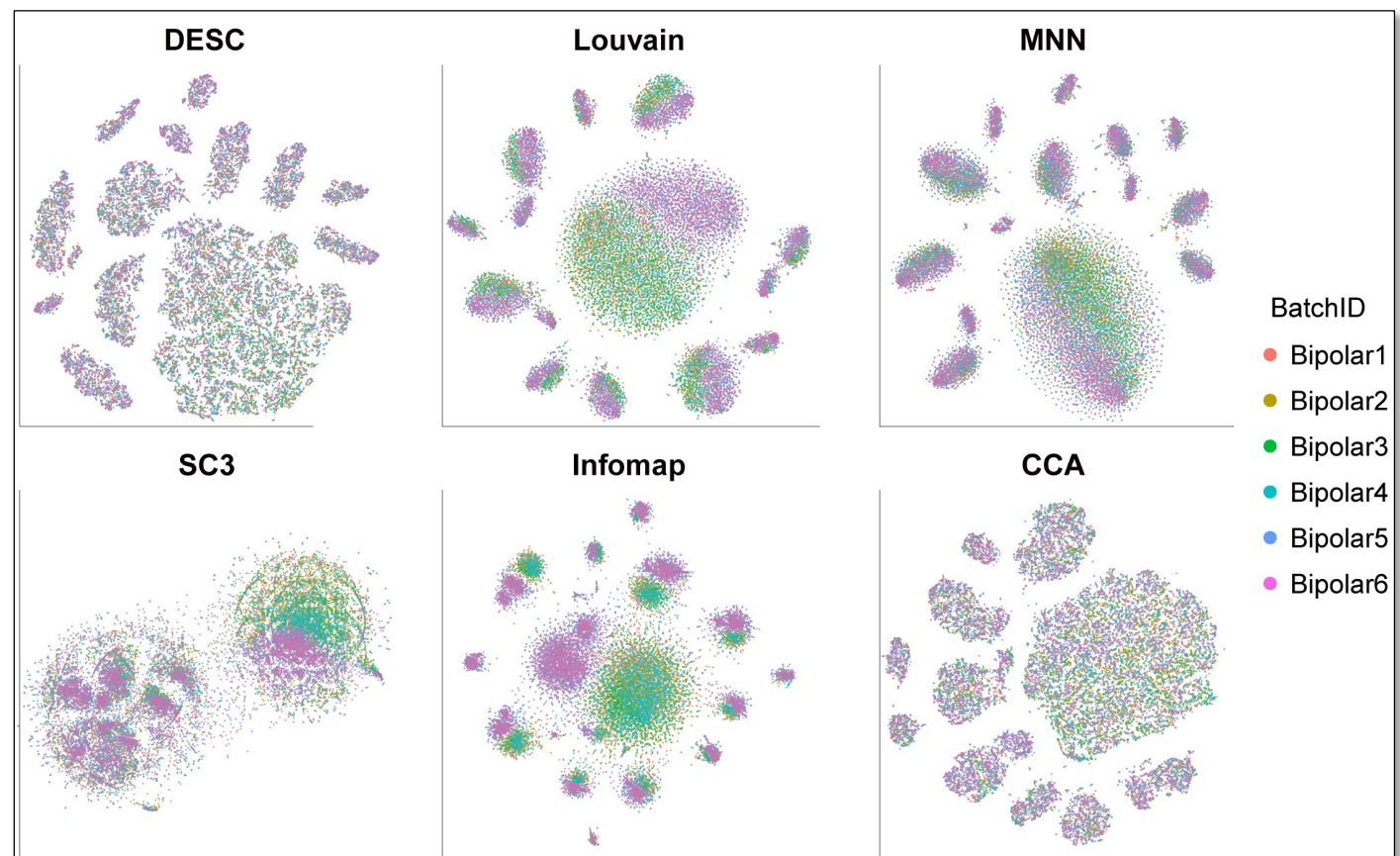


DESC (Deep Embedding for single-cell clustering):

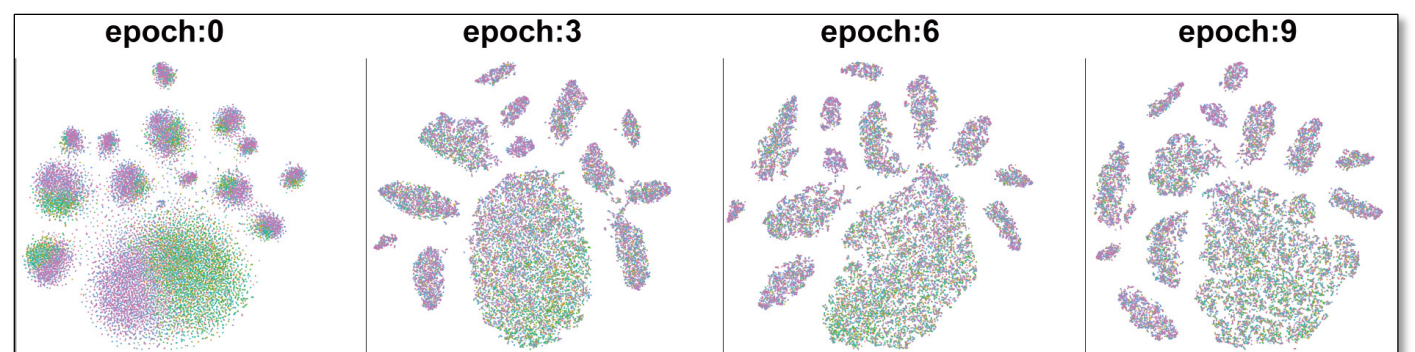
- Stacked auto-encoder learns cluster-specific gene expression representation and cluster assignments for scRNA-seq data clustering
- Initialize clustering obtained from autoencoder
- Learn non-linear mapping from original space to a low-dimensional space
- iteratively optimize clustering objective function
 - Move each cell to nearest cluster
 - balance biological and technical differences between clusters
 - reduce influence of batch effect
- Enables soft clustering by assigning cluster-specific probabilities to each cell
- Facilitates clustering of cells with high confidence

Deep learning enables accurate clustering and batch effect removal in single-cell RNA-seq analysis

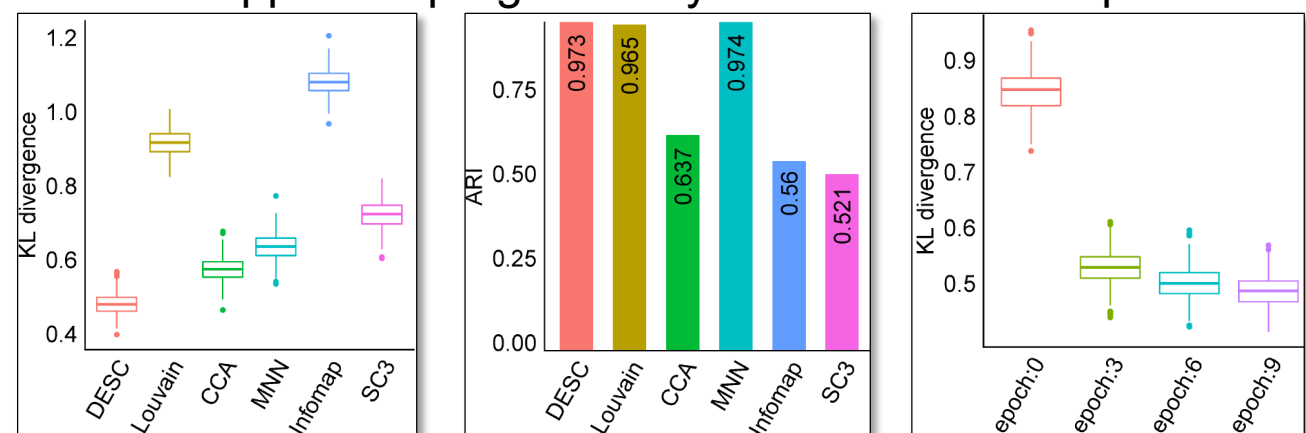
Xiangjie Li, Yafei Lyu, Jihwan Park, Jingxiao Zhang, Dwight Stambolian, Katalin Susztak, Gang Hu, Mingyao Li
doi: <https://doi.org/10.1101/530378>



DESC avoids cluster-specific batch effects found in other methods

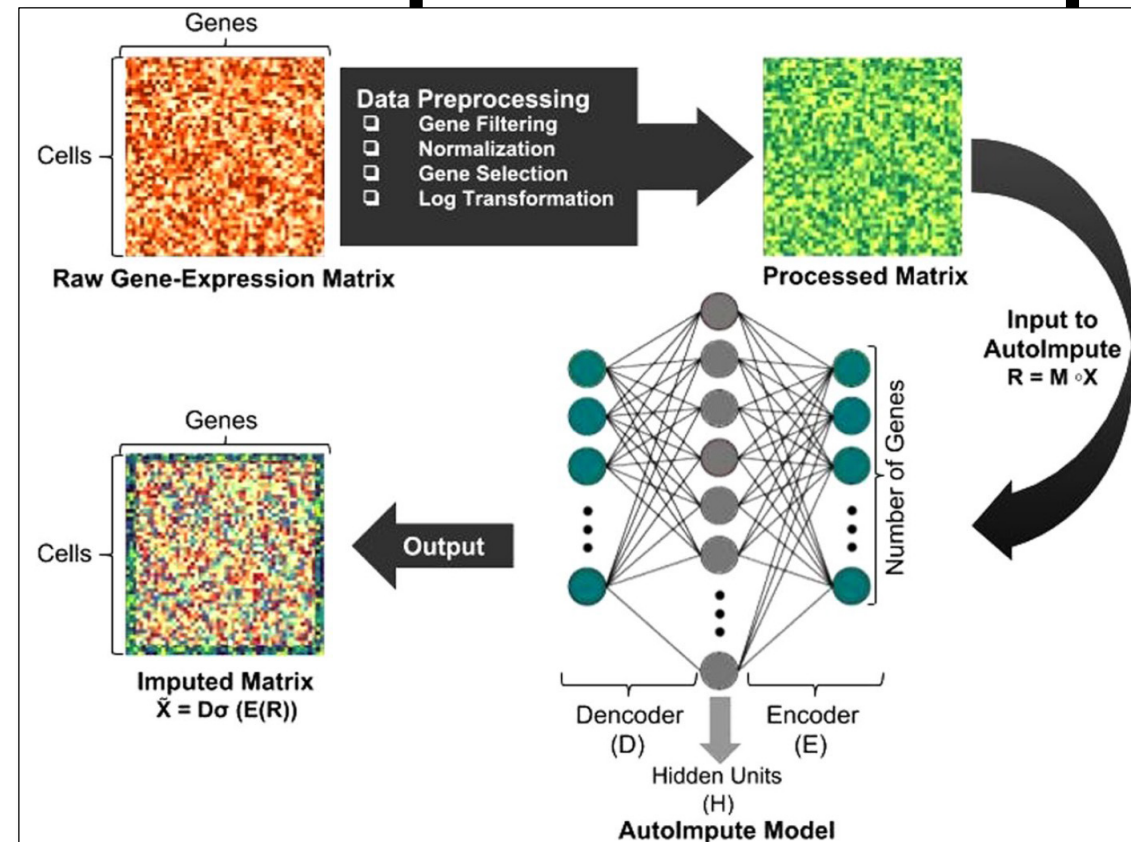


Iterative approach progressively removes cluster-specific batch effects



Rand Index (RI) = measure of the similarity between two data clusterings
ARI = Adjusted Rand Index, adjusted for the chance grouping of elements

AutoImpute: Overcomplete autoencoder for filling in zeros



AutoImpute

Filter raw gene expression data for bad genes

(normalize by library size, prune by gene-selection, log transform)

Feed processed matrix to AutoImpute model

- learn expression data representation
- reconstruct imputed matrix

Use overcomplete autoencoders to capture distribution of sparse gene expression data, and regenerate complete version of it

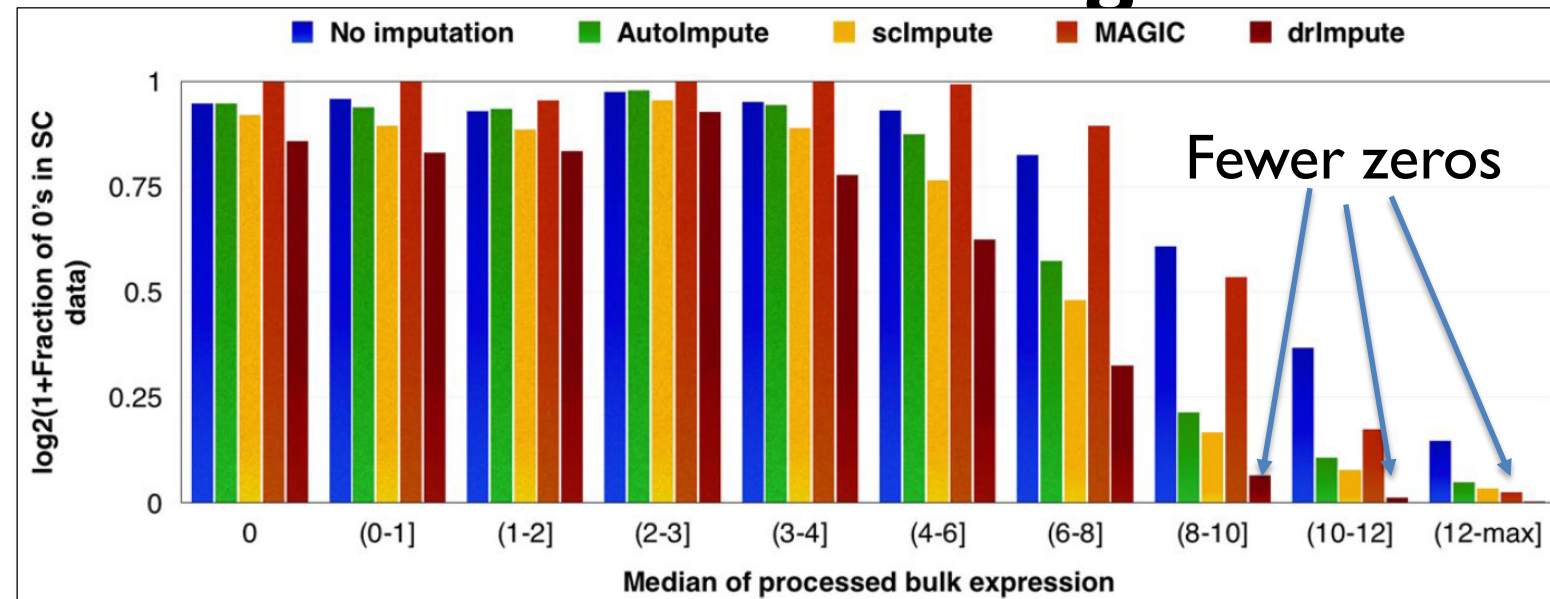
- Feeding sparse gene expression matrix as input to autoencoder
- train it to learn the encoder and decoder functions that best regenerate imputed expression with no dropouts
- back-propagating errors only for non-zero counts in sparse matrix

Training and Hyper-parameter Selection. The autoencoder network consists of a fully-connected multi-layer perceptron (MLP), with three layers: input, hidden and the output layer. It is trained using gradient descent with gradients computed by back-propagation to reach the minimum of the cost function (equation 8). RMSProp Optimizer was used to adjust the learning rate, such that, we avoid getting stuck at local minima and reach the minimum of the cost function faster. Both E - encoder matrix and D - decoder matrix were initialized from a random normal distribution.

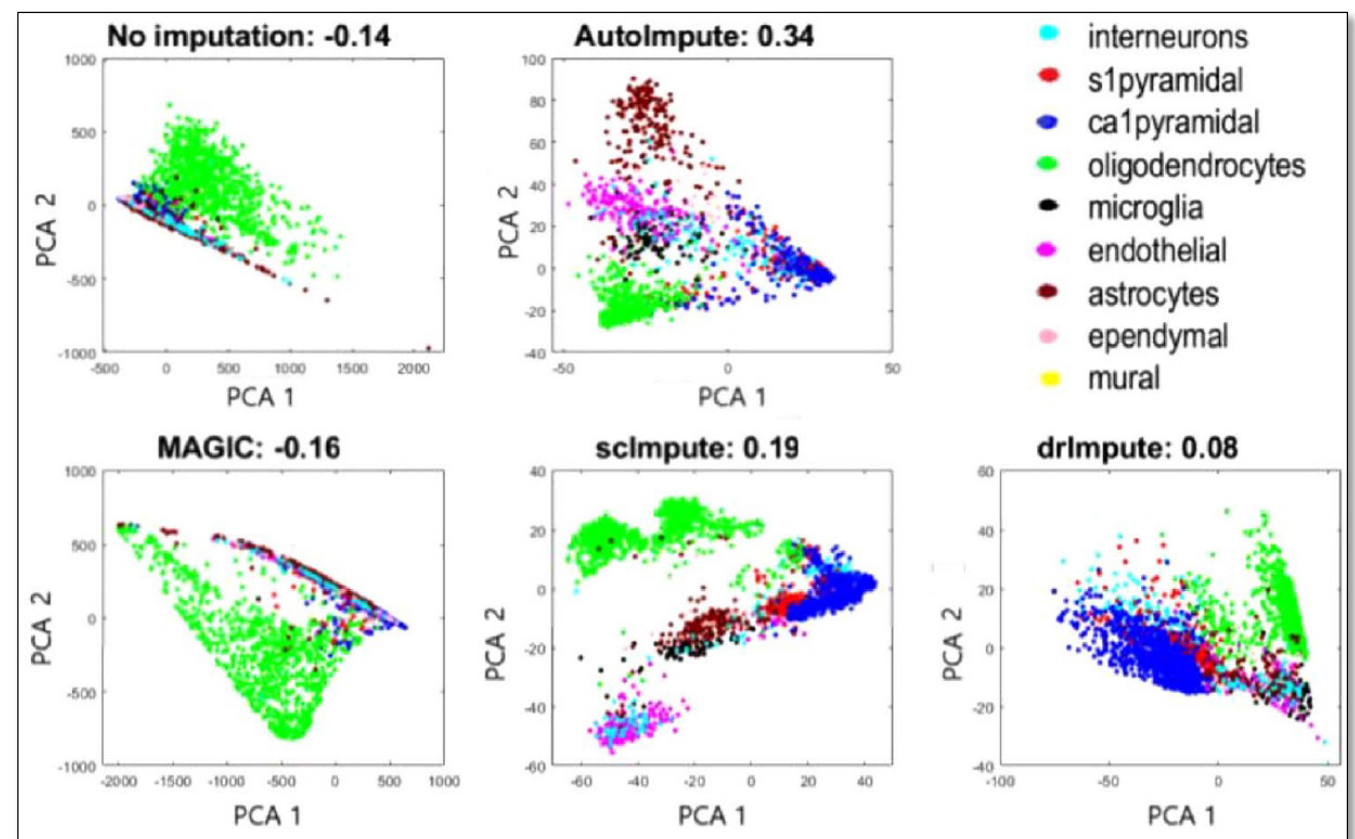
The hyper-parameter selection was done after doing an extensive grid search on the following hyper-parameters:

- λ - the regularization coefficient, to control the contribution of the regularization term in the loss or cost function.
- Size of the hidden layer or latent space dimensionality.
- Initial value of learning rate.
- Threshold value - We stop the gradient descent after the change in loss function value in consecutive iterations is less than the threshold value, implying convergence.

The best results were observed on the hyper-parameter choices shown in Table 2.



Autoimpute captures more non-zero values for highly-expressed genes

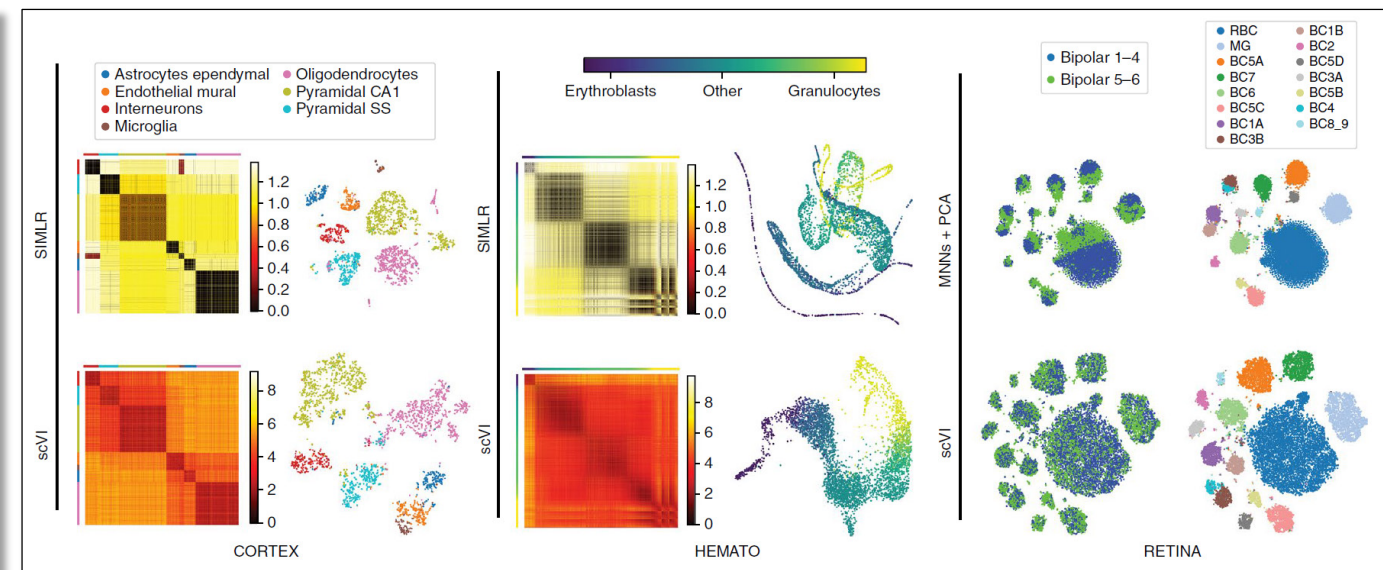
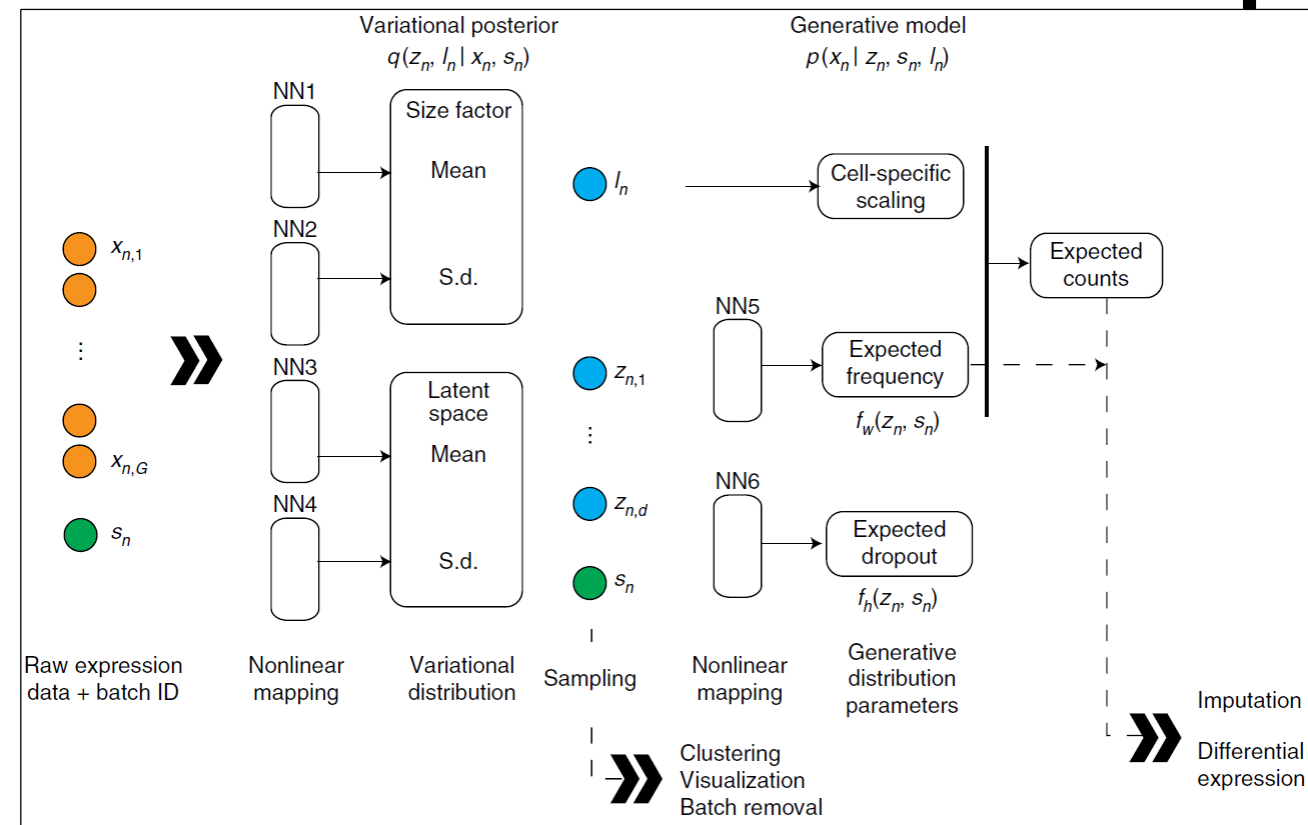


Iterative approach progressively removes cluster-specific batch effects

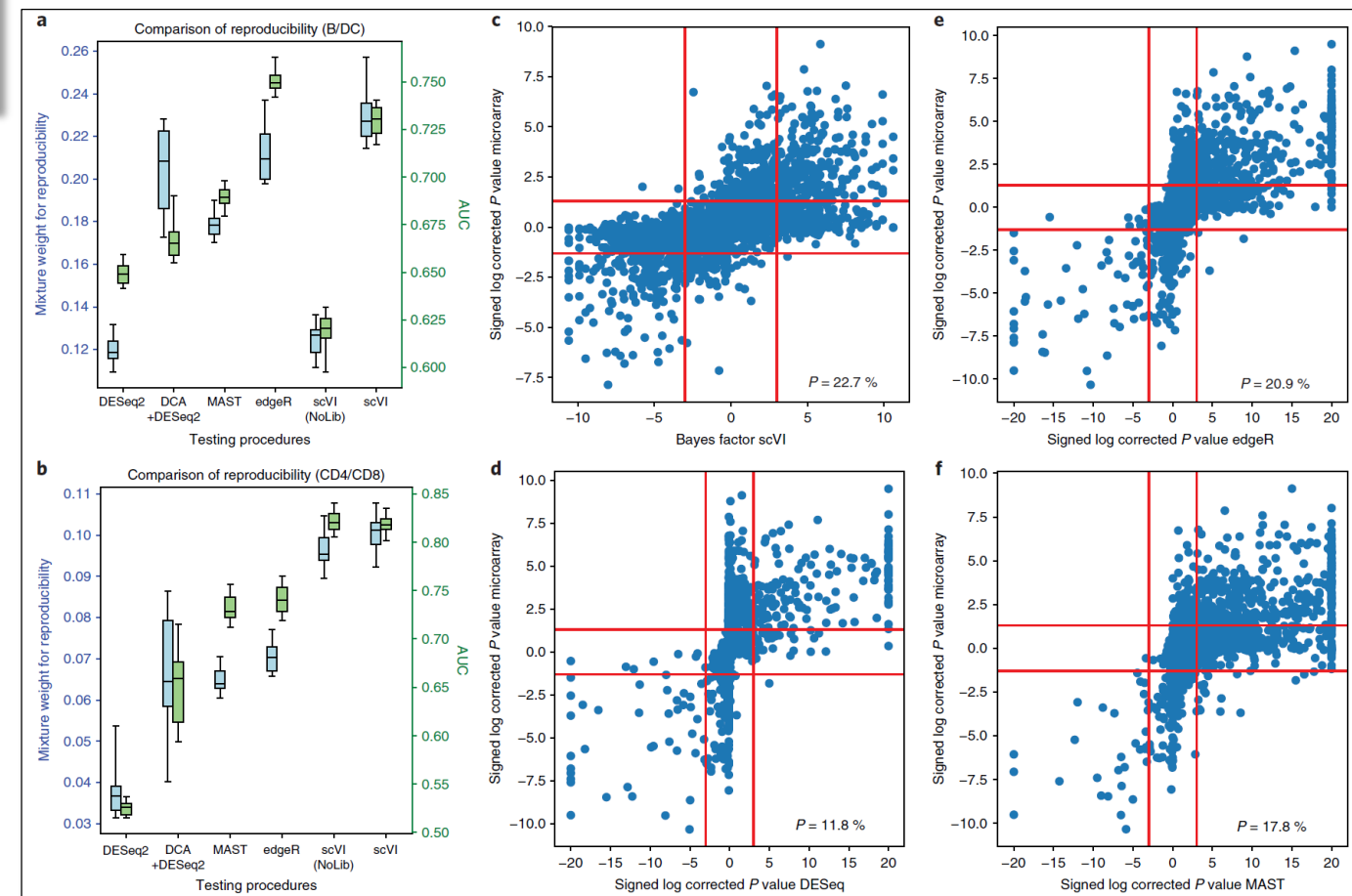
AutoImpute: Autoencoder based imputation of single-cell RNA-seq data

Divyanshu Talwar¹, Aanchal Mongia¹, Debarka Sengupta^{1,3} & Angshul Majumdar²

scVI: Use NN to estimate params in variational inference



scVI retains biological signal in diverse datasets



scVI enables differential expression analysis

scVI: Learn non-linear embedding of cells for multiple analysis tasks
 NN=Neural networks used to compute embedding and expr. distribution
 f_w, f_h : functional representations NN5,6 to capture parameters of Gaussians

Modeled observed expression x_{ng} (gene g, cell n) as sample
 Drawn from zero-inflated negative binomial (ZINB) distribution
 Conditioned on the batch annotation s_n of each cell (if available)

And on two additional, unobserved random variables:

- ρ_g^n **nuisance variation**, 1-D Gaussian, model differences in capture efficiency & sequencing depth, cell-specific scaling factor
- z_n **remaining variation**, 10-D Gaussian, model biological differences between cells.

Represent each cell as point in low-dimensional latent space (for visualization and clustering).

Neural network maps the latent variables to ZINB distribution parameters (Fig. 1a, neural networks 5 and 6).

This mapping goes through intermediate variables:

- batch-corrected, normalized estimate of the percentage of transcripts in each cell n that originate from each gene g

Use these estimates for differential expression analysis

Use scaled version (multiplying by estimated library size) for imputation.

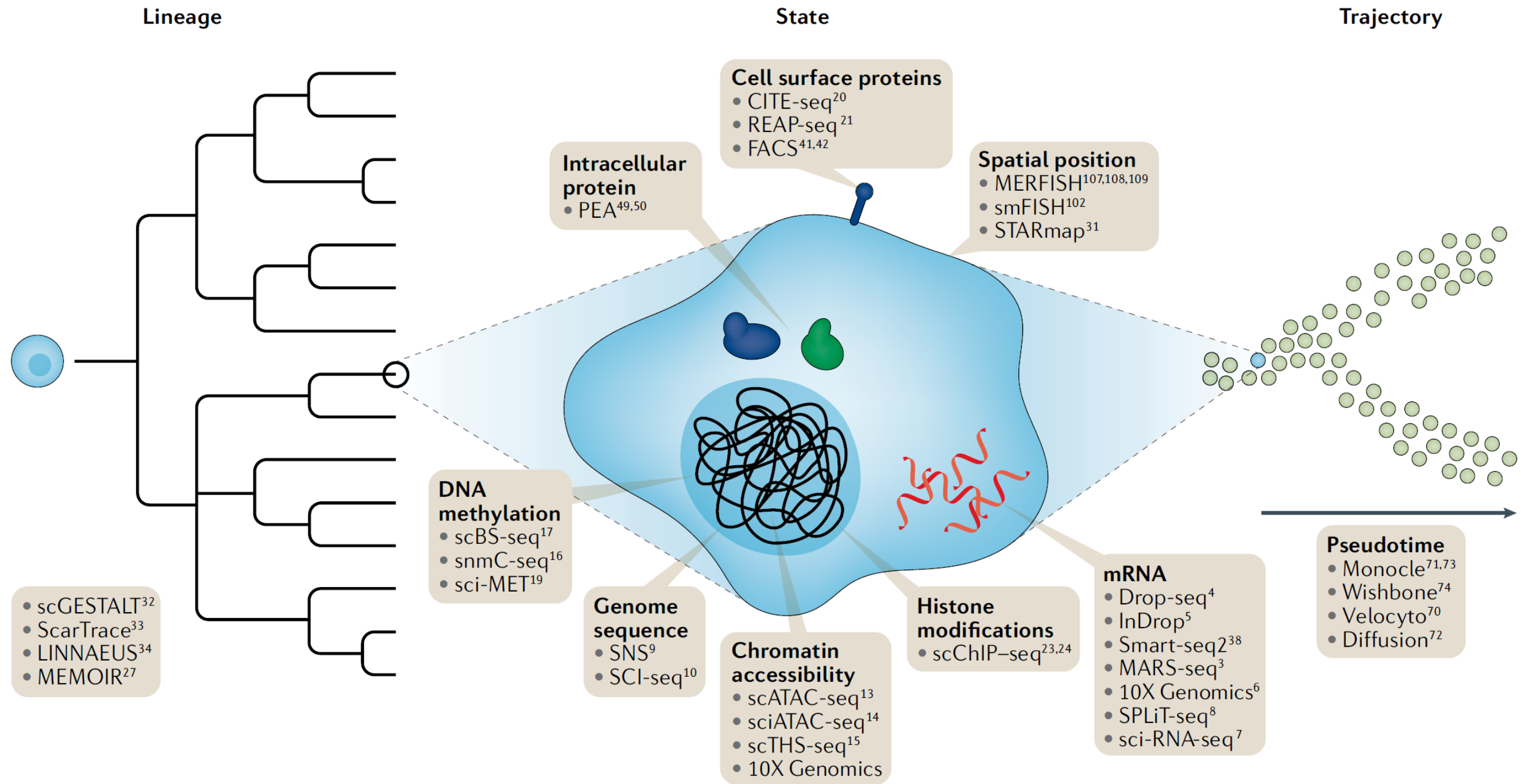
Derived approximation for posterior distribution of latent variables q by training another neural network using variational inference and a scalable stochastic optimization procedure (NN1-NN4).

Today: Deep Learning for Single-cell Genomics

1. Why single cells, traditional approaches, scRNA-seq
2. Scaling up single-cell technologies: evolution of scRNA-seq
3. Dealing with noise, doublets, and other scRNA-seq issues
4. Computational challenges in scRNA-seq data analysis
5. Deep learning methods for scRNA-seq data analysis
6. Beyond scRNA-seq: scATAC-seq, multi-omics

6. Beyond RNA: scATAC-seq, Multi-Omics

Diverse technologies for sc profiling



Data types	Method name	Feature throughput	Cell throughput	Refs
<i>Unimodal</i>				
mRNA	Drop-seq	Whole transcriptome	1,000–10,000	4
	InDrop	Whole transcriptome	1,000–10,000	5
	10X Genomics	Whole transcriptome	1,000–10,000	6
	Smart-seq2	Whole transcriptome	100–300	38
	MARS-seq	Whole transcriptome	100–300	3
	CEL-seq	Whole transcriptome	100–300	1
	SPLiT-seq	Whole transcriptome	≥ 50,000	8
	sci-RNA-seq	Whole transcriptome	≥ 50,000	7
Genome sequence	SNS	Whole genome	10–100	9
	SCI-seq	Whole genome	10,000–20,000	10
Chromatin accessibility	scATAC-seq	Whole genome	1,000–2,000	13
	sciATAC-seq	Whole genome	10,000–20,000	14
	scTHS-seq	Whole genome	10,000–20,000	15
DNA methylation	scBS-seq	Whole genome	5–20	17
	snmC-seq	Whole genome	1,000–5,000	16
	sci-MET	Whole genome	1,000–5,000	19
	scRRBS	Reduced representation genome	1–10	18
Histone modifications	scChIP-seq	Whole genome + single modification	1,000–10,000	24
Chromosome conformation	scHi-C-seq	Whole genome	1–10	26
<i>Multimodal</i>				
Histone modifications + spatial	NA	Single locus + single modification	10–100	23
mRNA + lineage	scGESTALT	Whole transcriptome	1,000–10,000	32
	ScarTrace	Whole transcriptome	1,000–10,000	33
	LINNAEUS	Whole transcriptome	1,000–10,000	34
Lineage + spatial	MEMOIR	NA	10–100	27
mRNA + spatial	osmFISH	10–50 RNAs	1,000–5,000	35
	STARmap	20–1,000 RNAs	100–30,000	31
	MERFISH	100–1,000 RNAs	100–40,000	108
	seqFish	125–250 RNAs	100–20,000	29
mRNA + cell surface protein	CITE-seq	Whole transcriptome + proteins	1,000–10,000	20
	REAP-seq	Whole transcriptome + proteins	1,000–10,000	21
mRNA + chromatin accessibility	sci-CAR	Whole transcriptome + whole genome	1,000–20,000	48
mRNA + DNA methylation	scM&T-seq	Whole genome	50–100	46
mRNA + genomic DNA	G&T-seq	Whole genome + whole transcriptome	50–200	44
mRNA + intracellular protein	NA	96 mRNAs + 38 proteins	50–100	50
		82 mRNAs + 75 proteins	50–200	49
DNA methylation + chromatin accessibility	scNOMe-seq	Whole genome	10–20	11

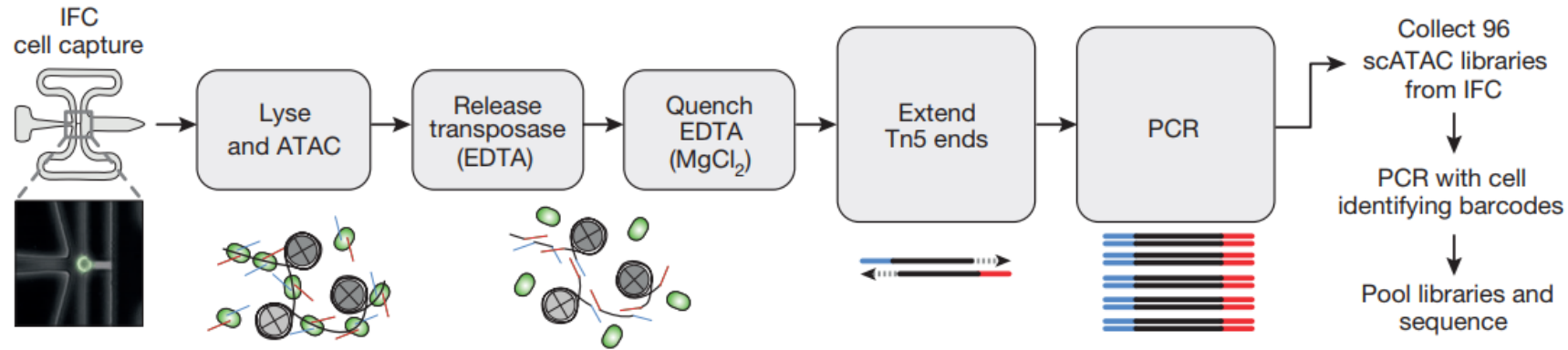

SINGLE-CELL OMICS

Integrative single-cell analysis

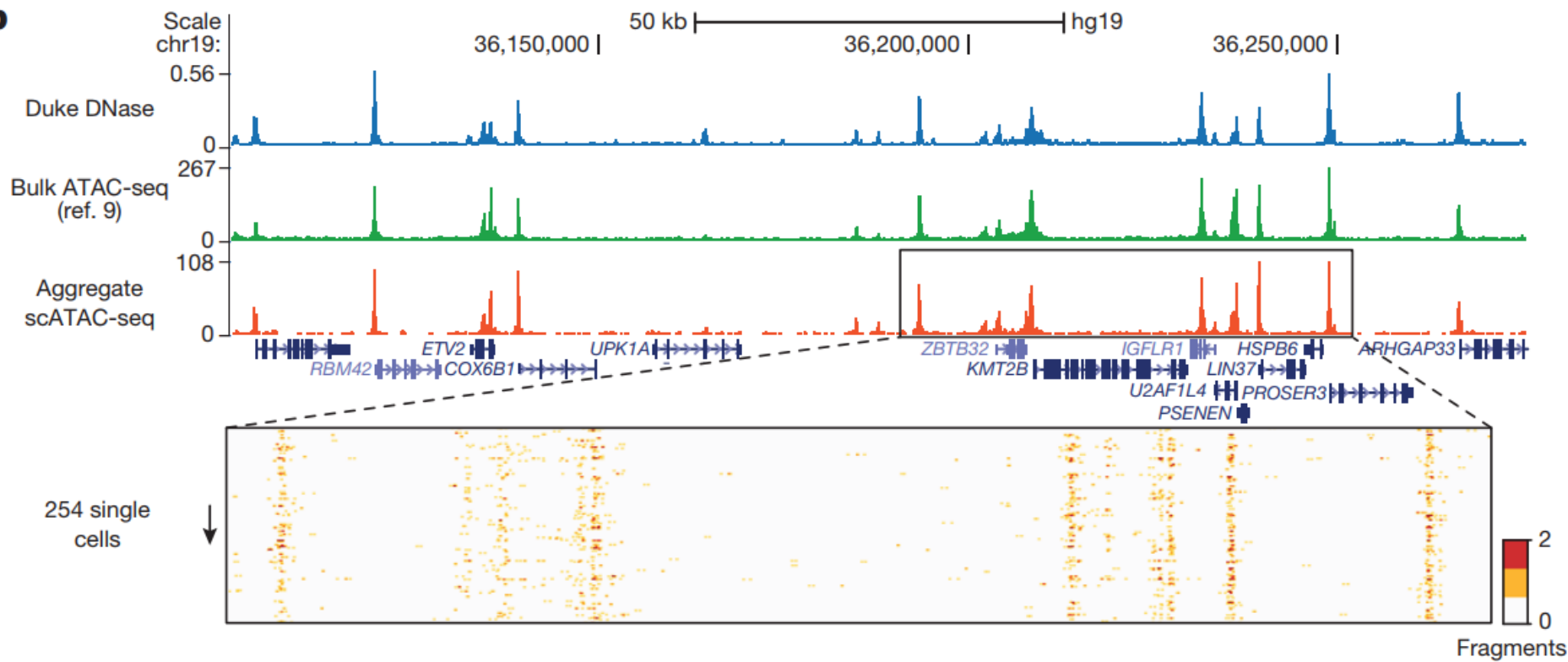
Tim Stuart¹ and Rahul Satija^{1,2*}

Single-cell Epigenomics (scATAC-Seq)

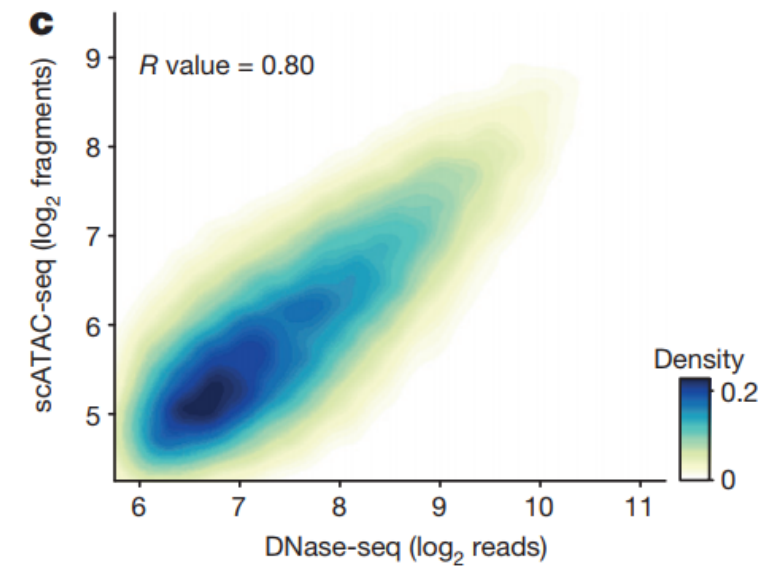
a



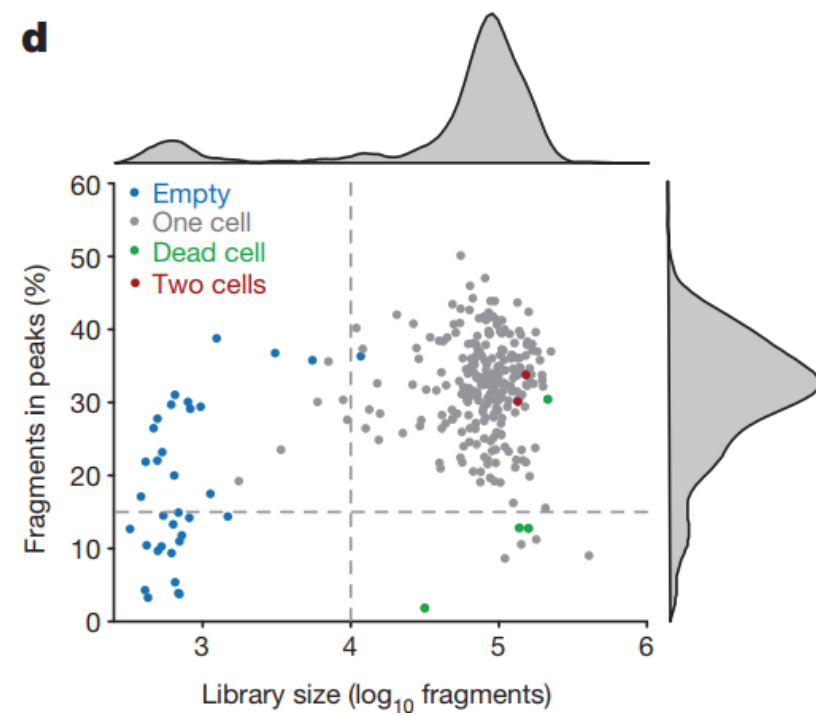
b



c



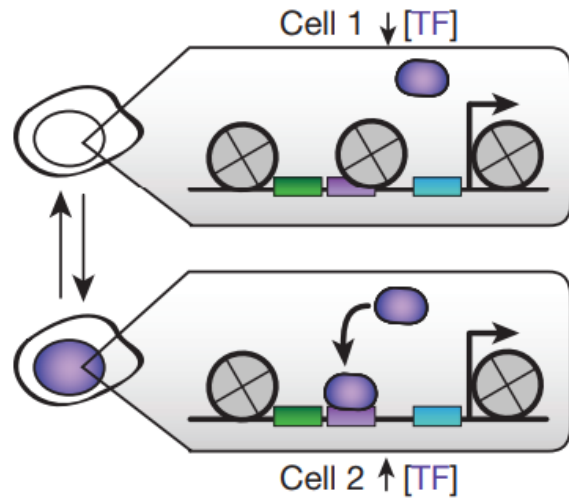
d



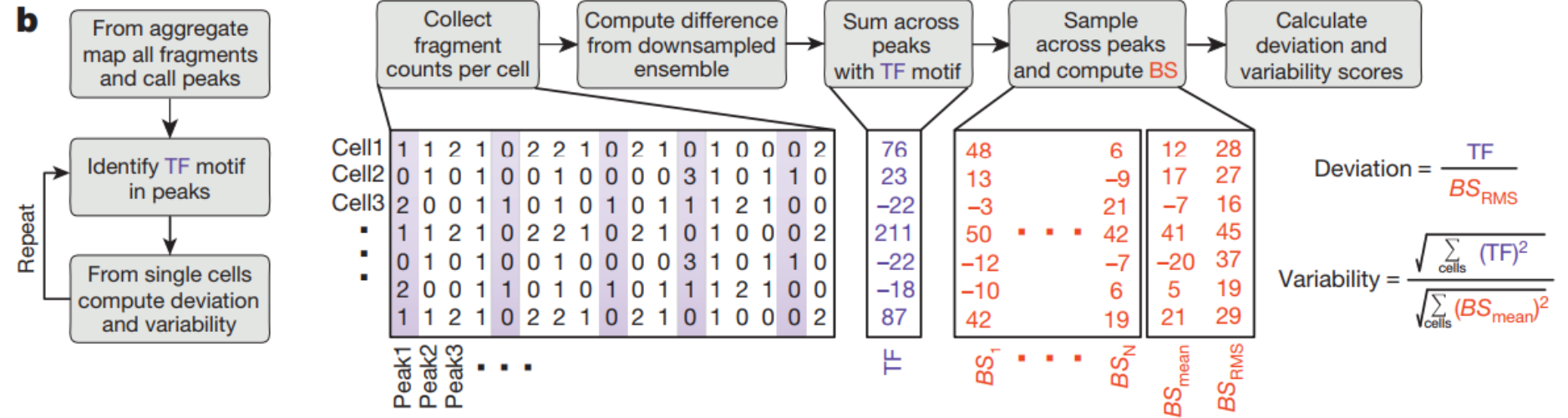
Buenrostro *et al.*, 2015

Trans-factors are associated with single-cell epigenomic variability

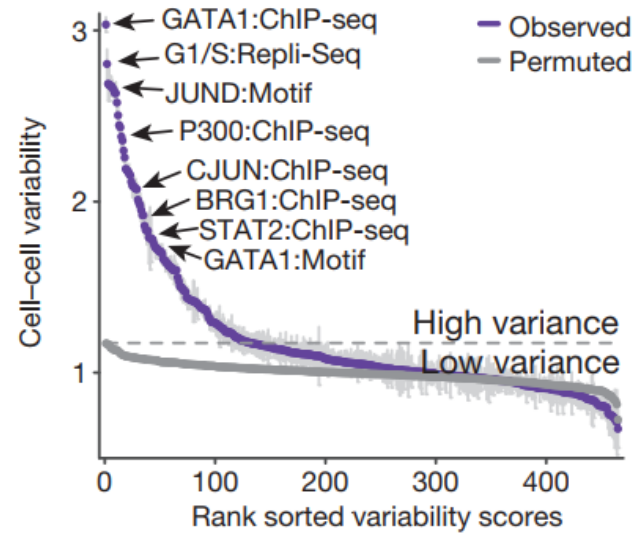
a



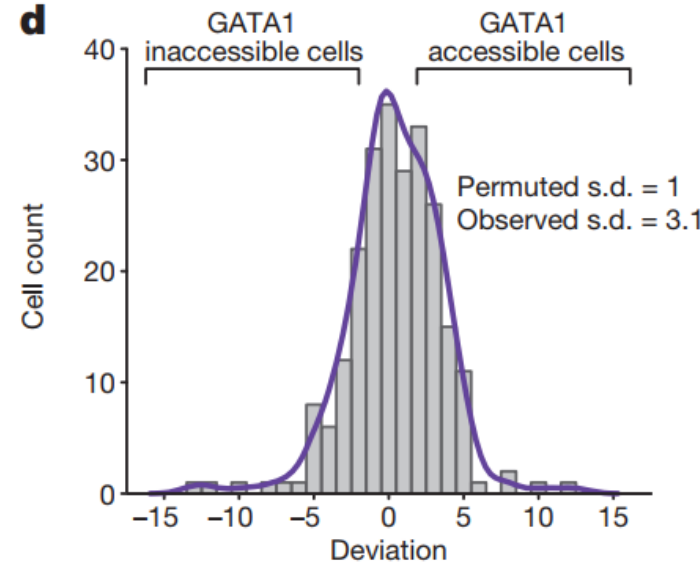
b



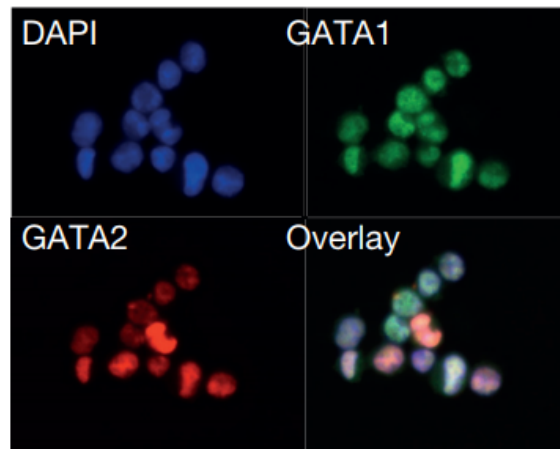
c



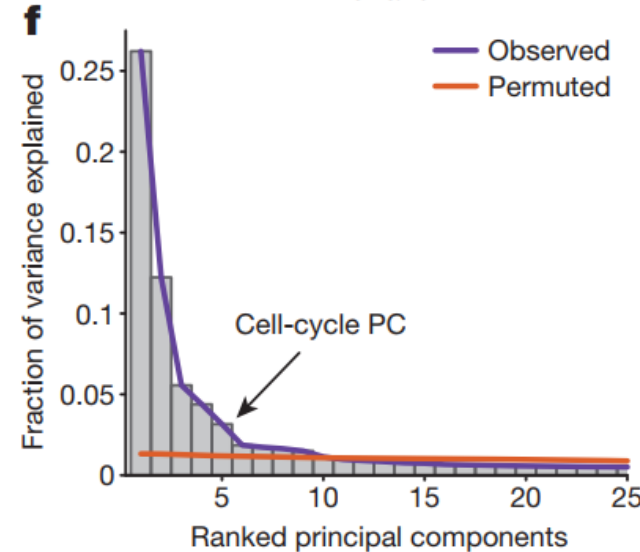
d



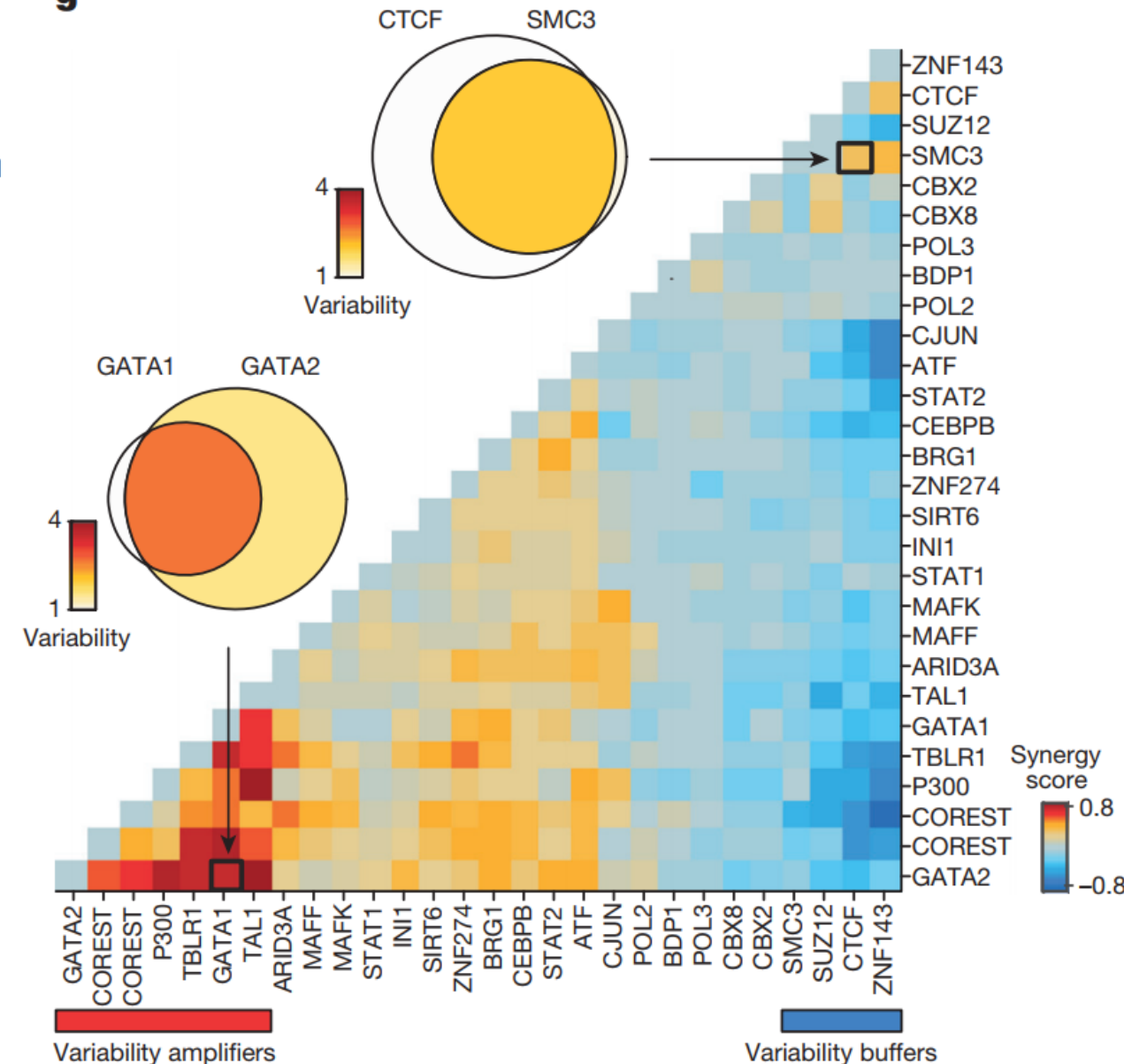
e



f

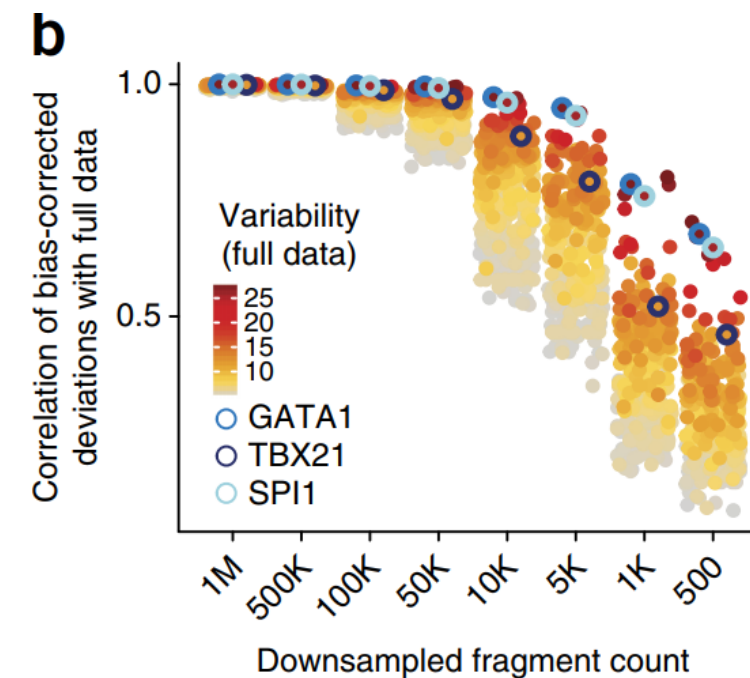
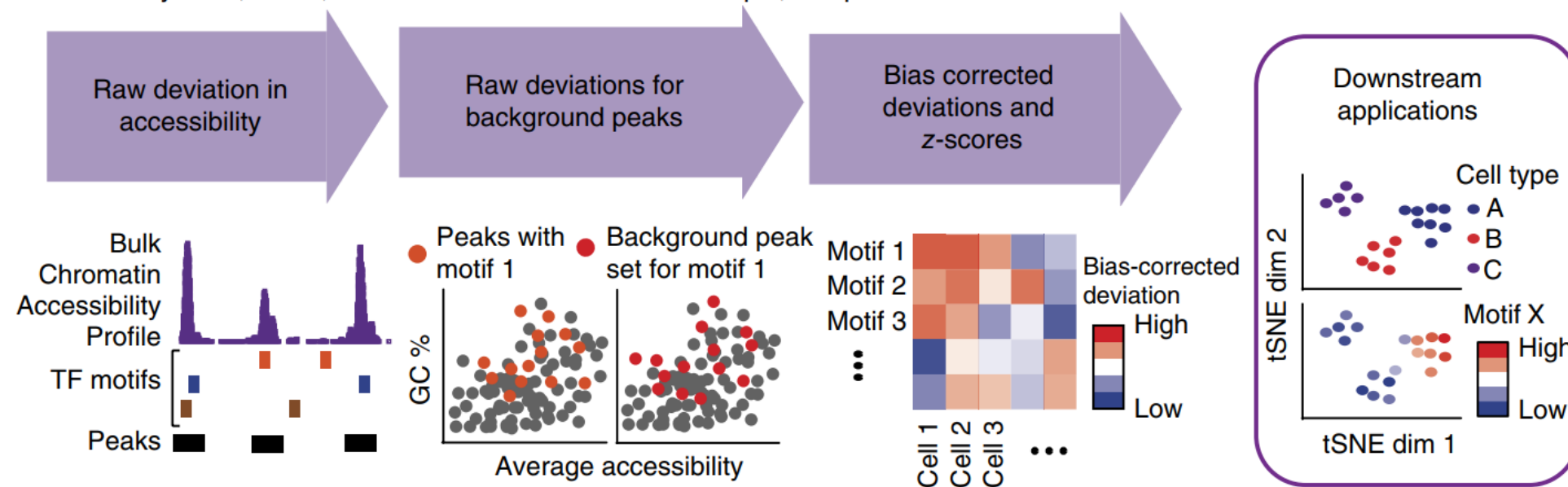


g

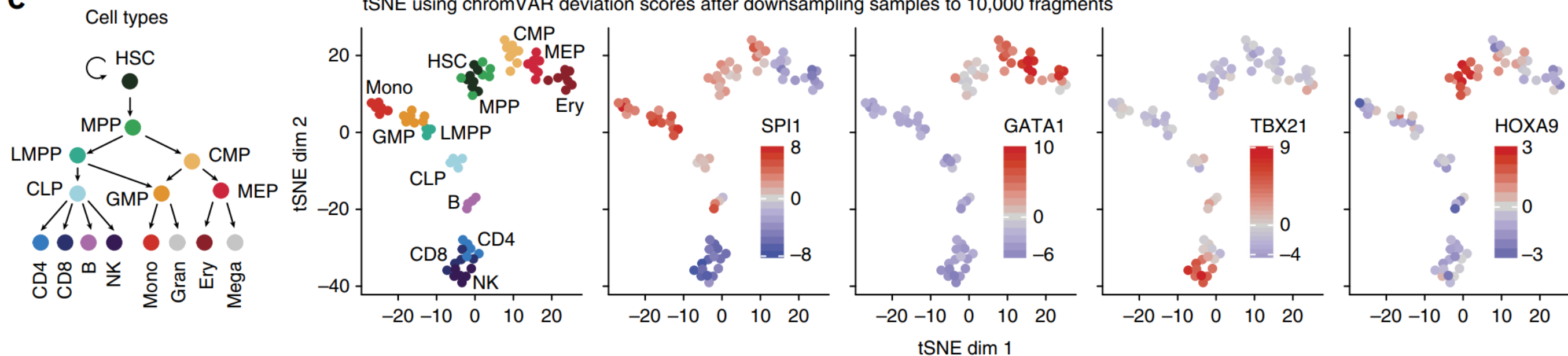


Integrate scATAC + scRNA using ChromVAR

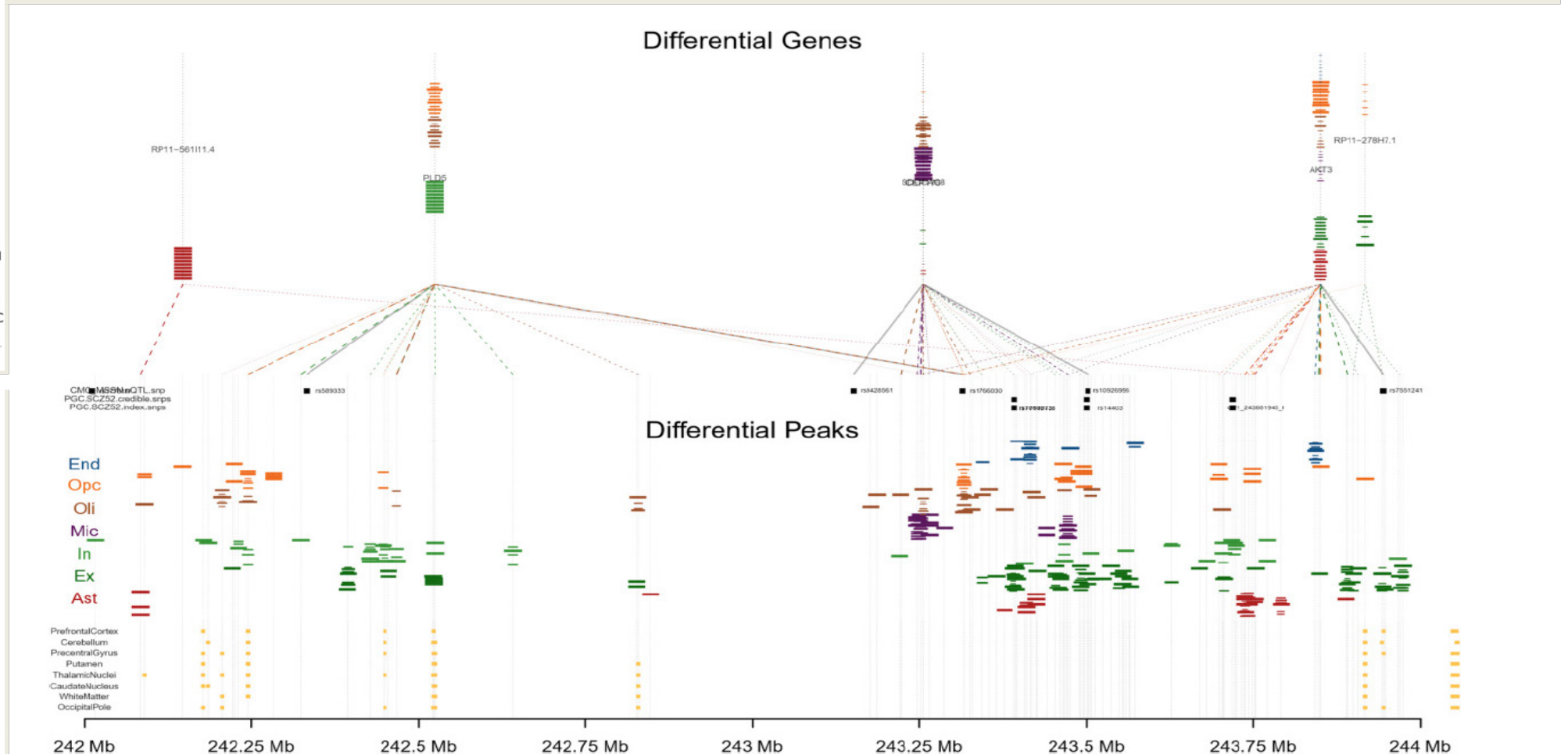
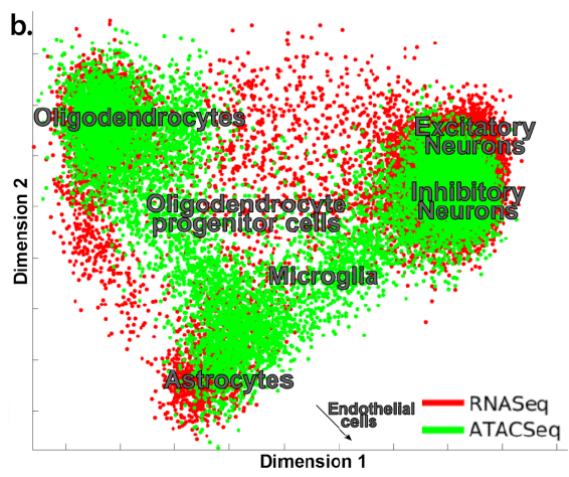
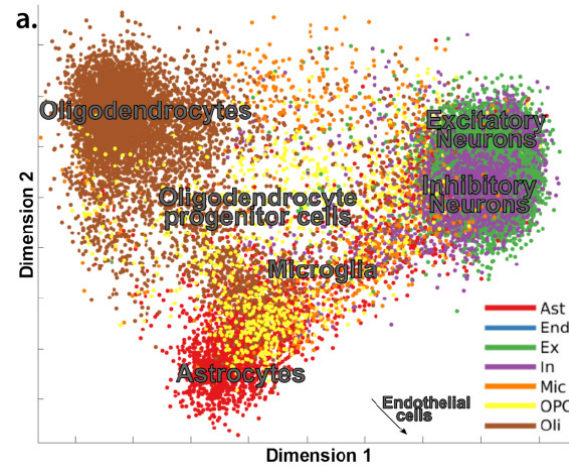
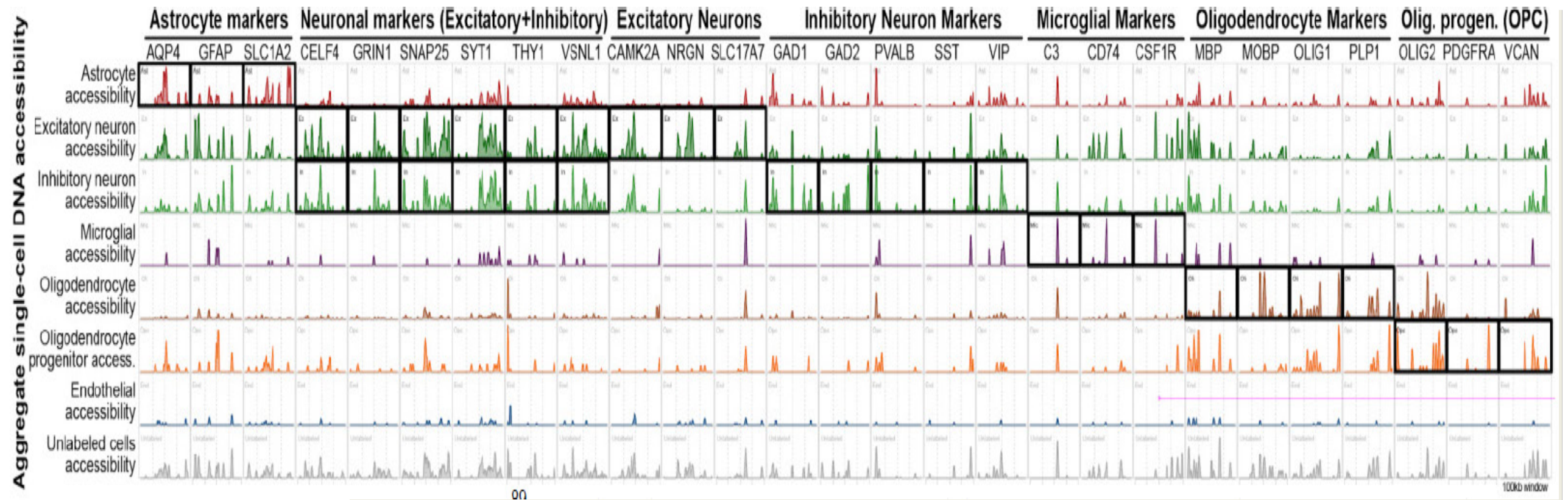
a For every motif, k-mer, or annotation and each cell or sample, compute:



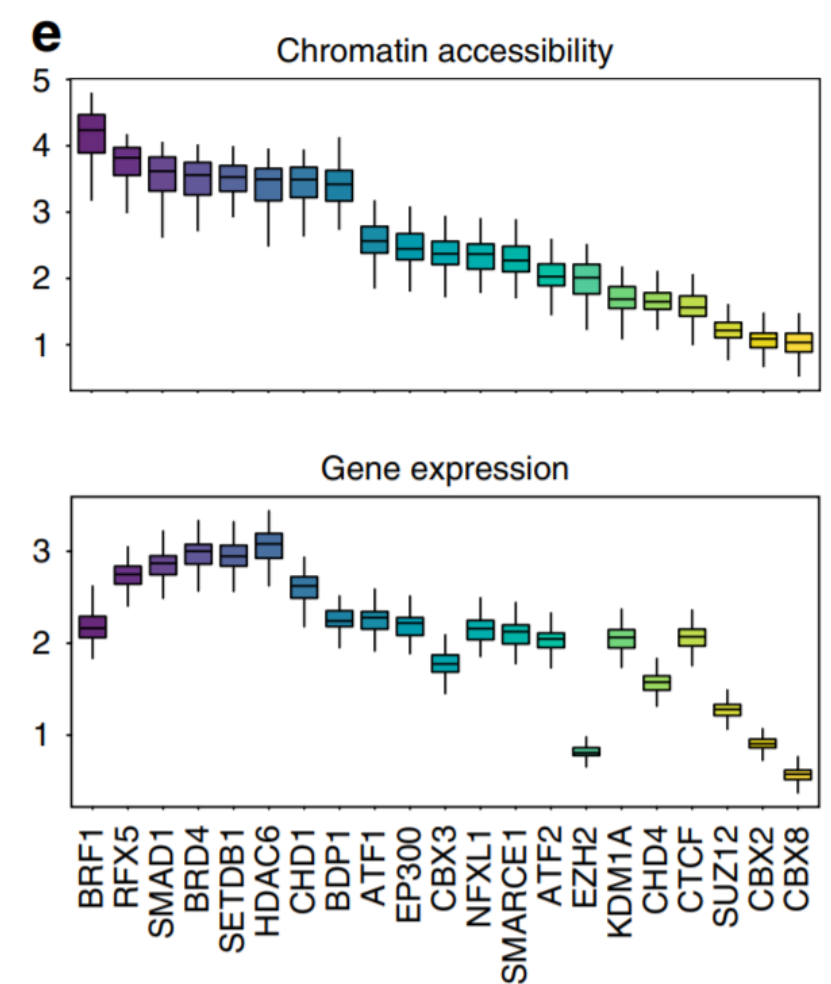
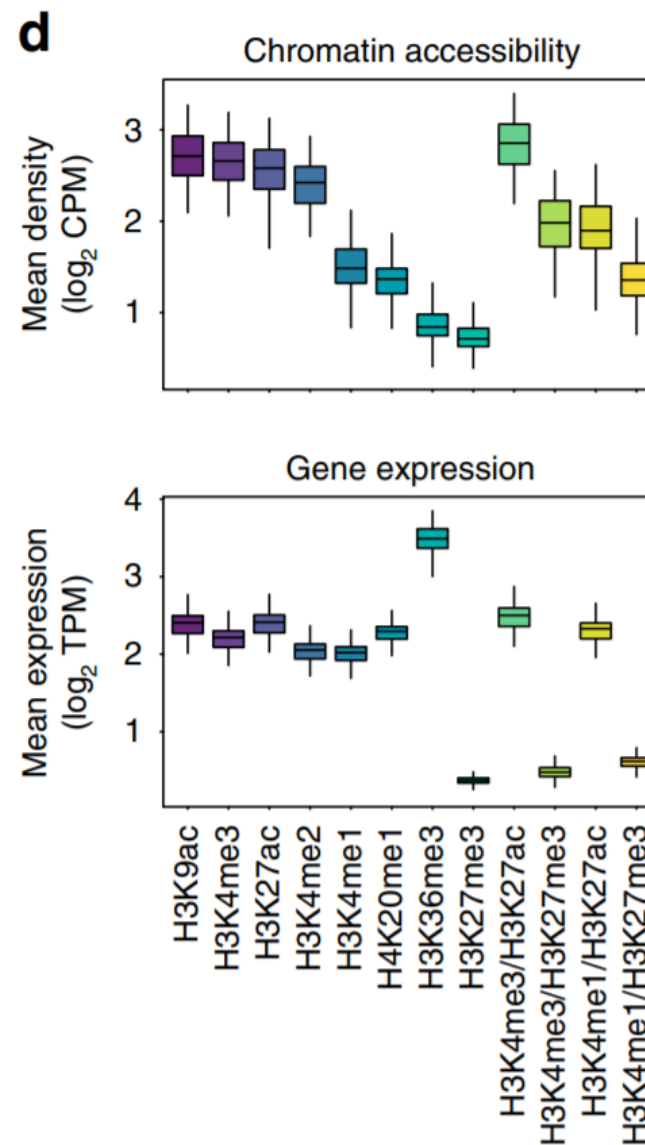
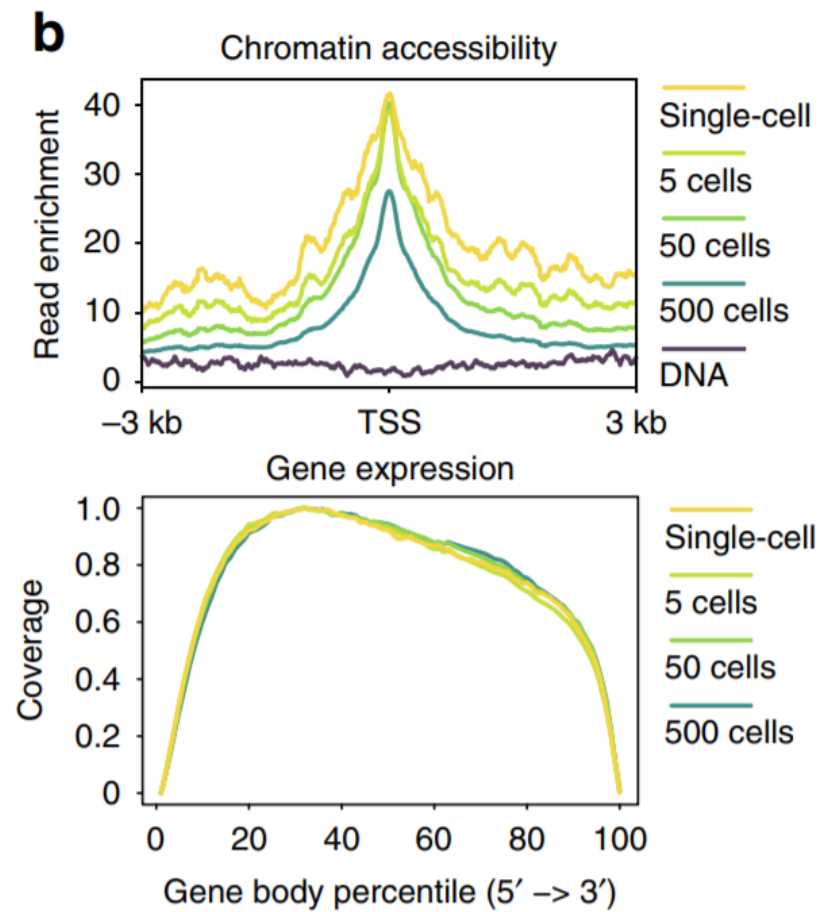
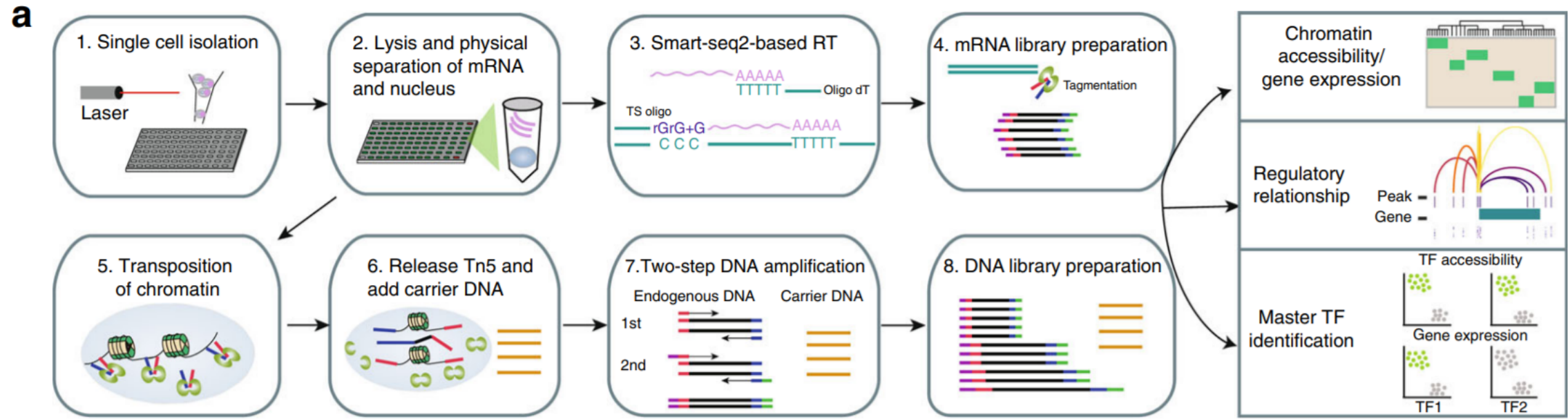
c tSNE using chromVAR deviation scores after downsampling samples to 10,000 fragments



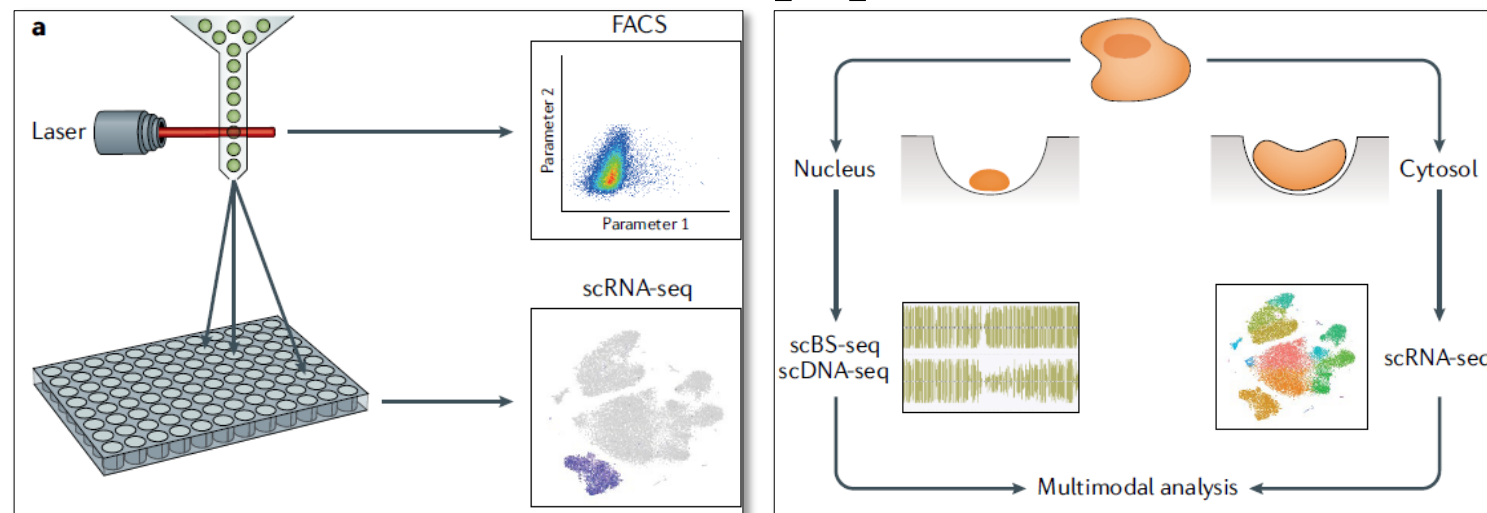
Link single-cell epigenomics and single-cell transcriptomics



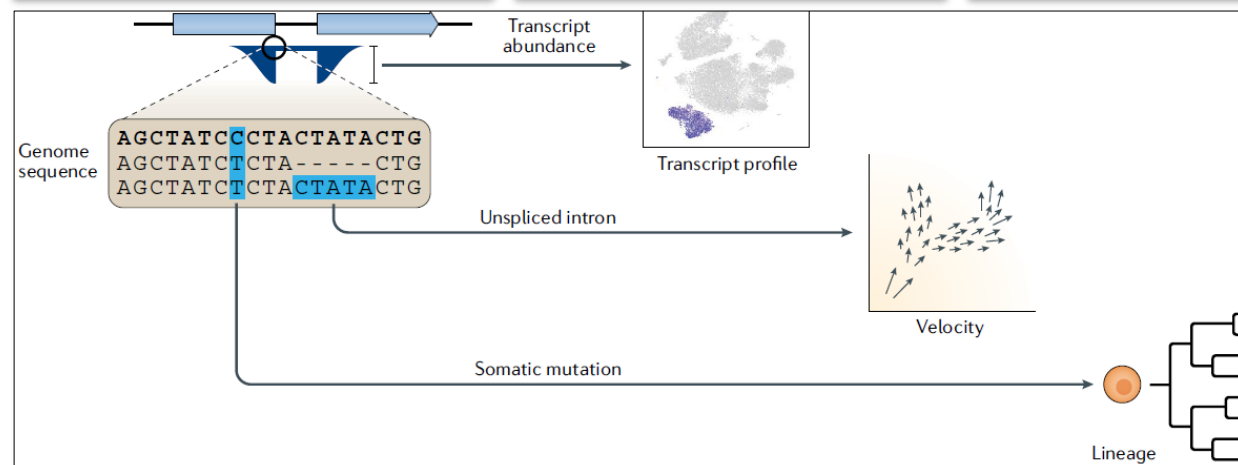
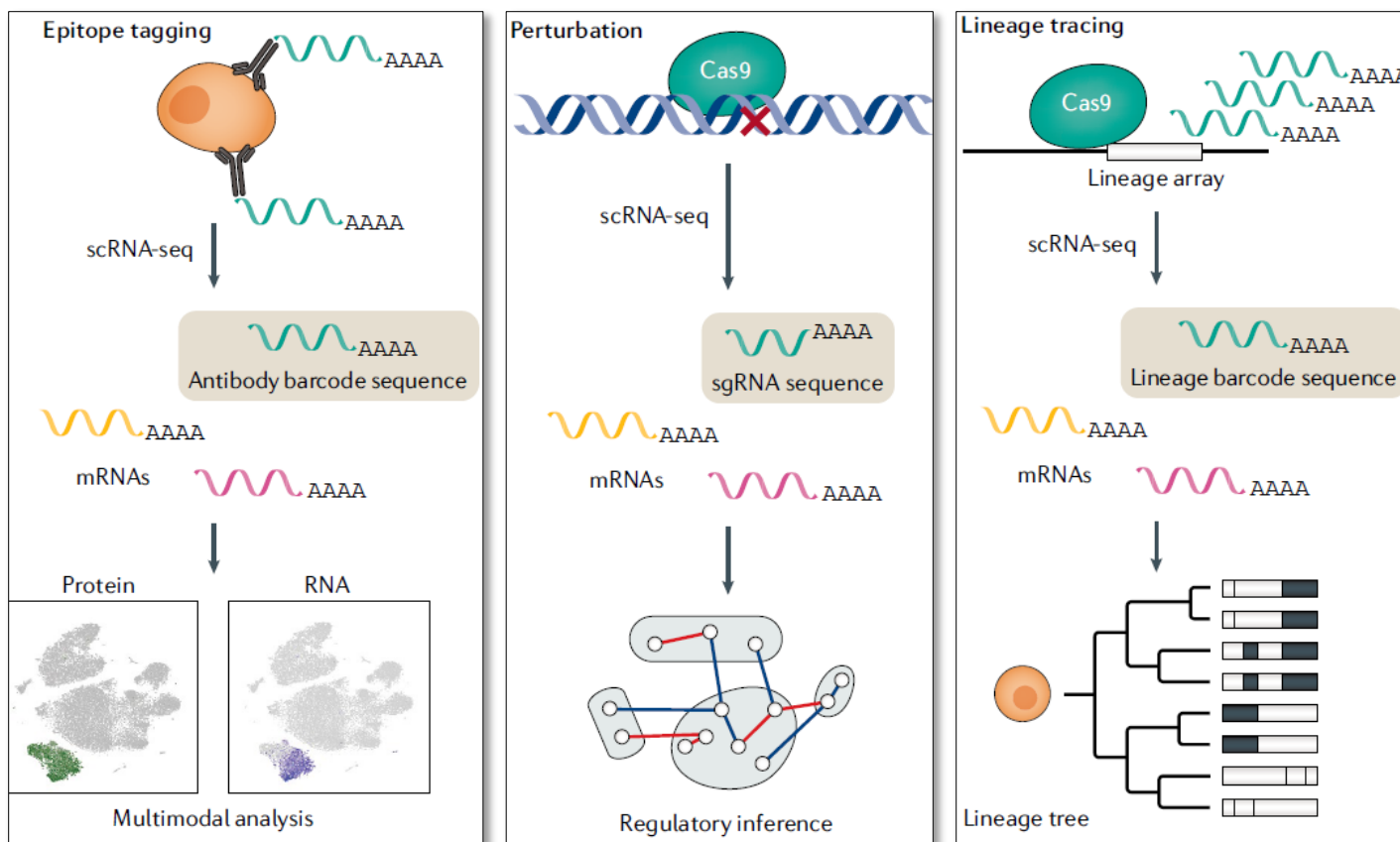
scMulti-Omics: Multiple profiling of the same cell



Diverse approaches for sc multi-omics



- a | Gathering cytometric single-cell measurements using multiparameter fluorescence-activated cell sorting (FACS) before single-cell RNA sequencing (scRNA-seq) can allow fluorescence-based measurements of protein levels to be later linked to cellular transcriptomes; hence, RNA and protein levels can be analysed jointly in the same cell.
- b | A lyse- and-split strategy can allow parallel workflows to be performed on different cellular fractions. For example, the cytosol can be physically separated from the nucleus to allow measurement of cytosolic mRNAs through scRNA-seq and measurements of the genomic DNA using whole-genome sequencing or bisulfite sequencing to gather complementary data on the cell genotype or methylome, respectively.
- c1 | Innovative barcoding strategies can enable standard scRNA-seq methods to capture important additional information to enhance the analysis of cell transcriptomes. Cell surface protein abundance can be captured using standard scRNA-seq methods by conjugating polyadenylated antibody barcodes to antibodies targeting cell surface proteins^{20,21} (left panel).
- c2 | These antibody barcode sequences can be captured alongside polyadenylated mRNAs and decoded to provide an estimate of protein levels for each cell. Allelic information can be encoded by the single-guide RNA (sgRNA) sequence used to guide Cas9 in pooled genetic screens, allowing gene knockout information to be associated with single-cell transcriptional profiles (middle panel).
- c3 | Cell lineage can also be encoded in a polyadenylated barcode sequence through the cumulative editing of a lineage array sequence by Cas9 (right panel). Over time, Cas9 will cut the lineage array, resulting in mutations at different points in the array. Cells sharing common mutations in the lineage array are likely to have originated from the same progenitor. By placing the lineage array sequence under the control of an RNA polymerase II promoter, these sequences can also be captured alongside endogenous mRNAs.
- d | Additional information can be extracted from scRNA-seq data beyond a typical analysis that provides only estimates of transcript counts in each cell. Somatic mutations can be identified from sequencing reads for each individual cell and can be used to reconstruct lineage relationships between cells. Retained introns can also be detected and can be used to give an estimate of the rate of change in transcript abundance (RNA velocity⁷⁰). scBS-seq, single-cell bisulfite sequencing; scDNA-seq, single-cell DNA sequencing.



Today: Deep Learning for Single-cell Genomics

1. Why single cells, traditional approaches, scRNA-seq
2. Scaling up single-cell technologies: evolution of scRNA-seq
3. Dealing with noise, doublets, and other scRNA-seq issues
4. Computational challenges in scRNA-seq data analysis
5. Deep learning methods for scRNA-seq data analysis
6. Beyond scRNA-seq: scATAC-seq, multi-omics