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

Lecture 10: Deep Learning for single-cell genomics

Prof. Manolis Kellis
Guest lecture: Fabian Theis
Guest Lecture: Romain Lopez

Massachusetts Institute of Technology

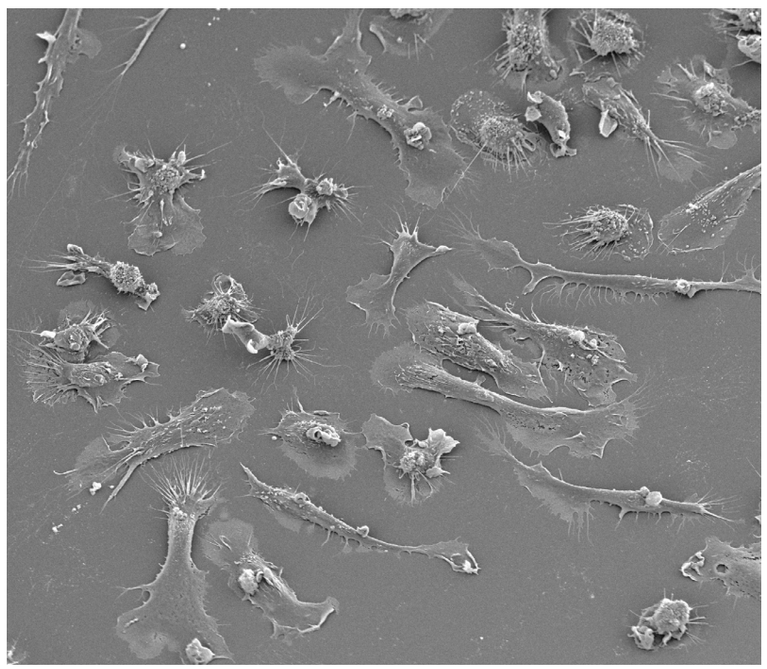
http://mit6874.github.io

Slides credit: Alex Shalek, Shahin Mohammadi, Fabian Theil, Romain Lopez

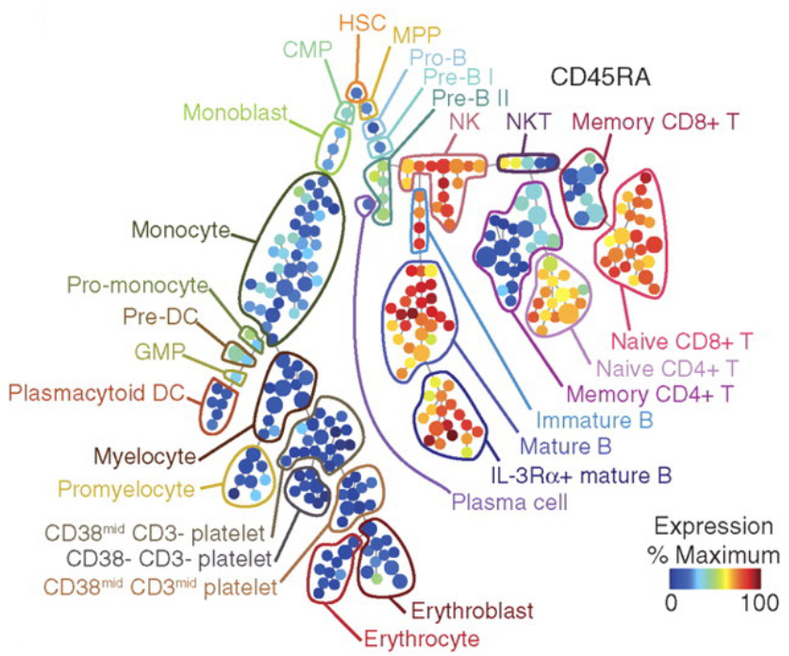
1. Foundations: Why single-cell profiling

Why single cells

Cellular heterogeneity

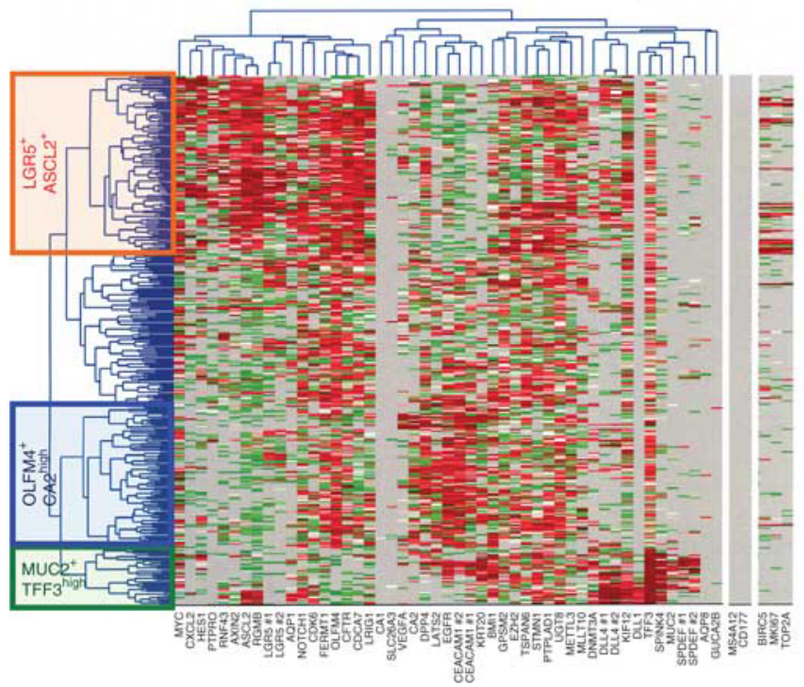


Differentiation trajectories



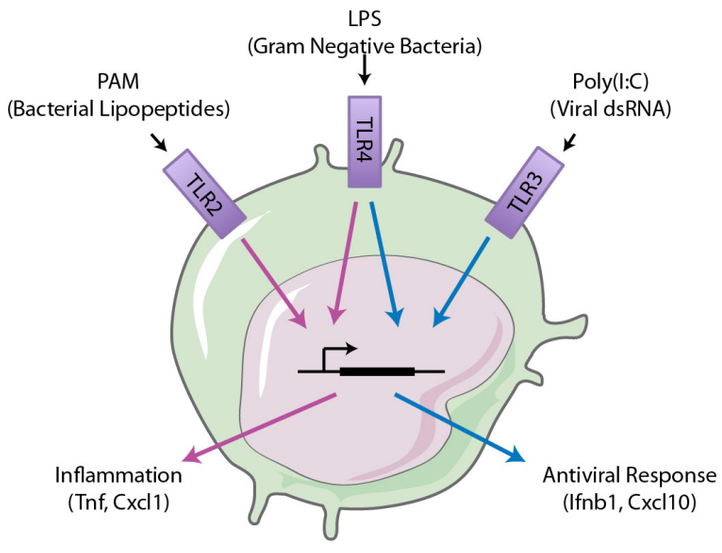
Bendall et al. (2011), Science

Within-cell-type differences

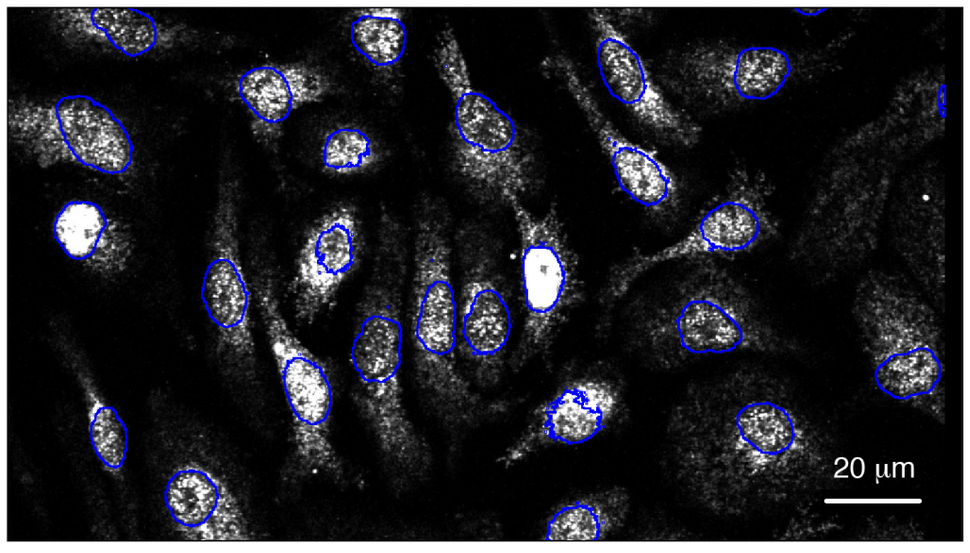


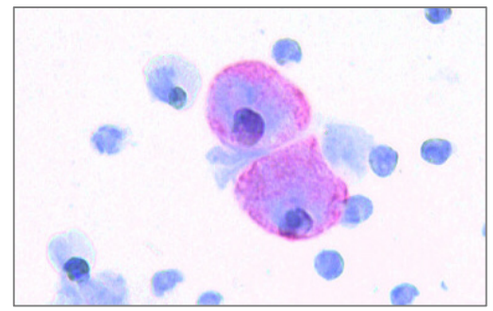
Dalerba et al. (2011), Nature Biotech

TLR Signaling



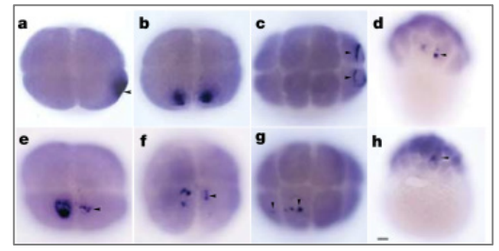
IRF3 Protein Levels - 4h LPS





fightcolorectalcancer.org/

Circulating Tumor Cells



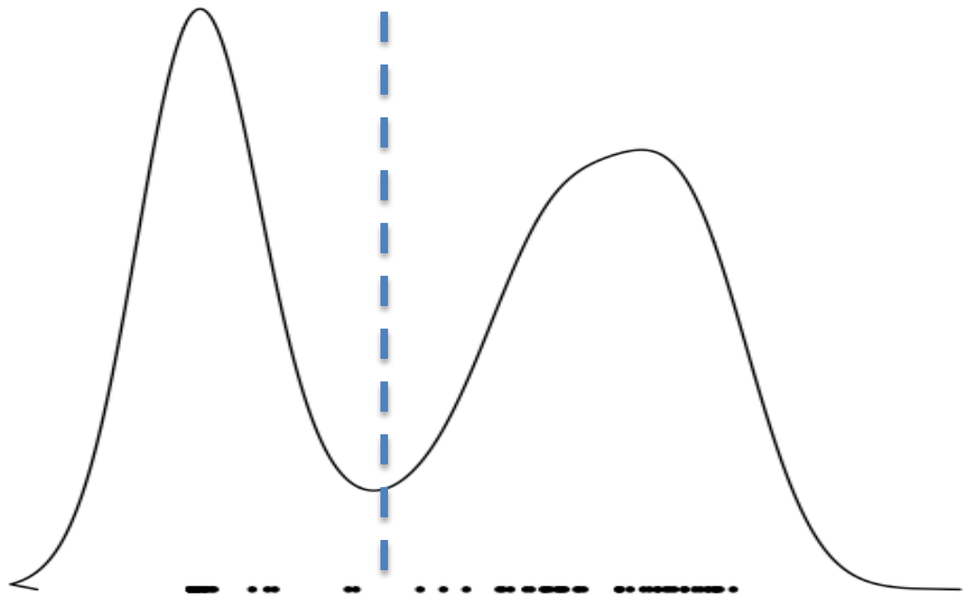
Gore et al. (2005), Nature

Zebrafish early embryo

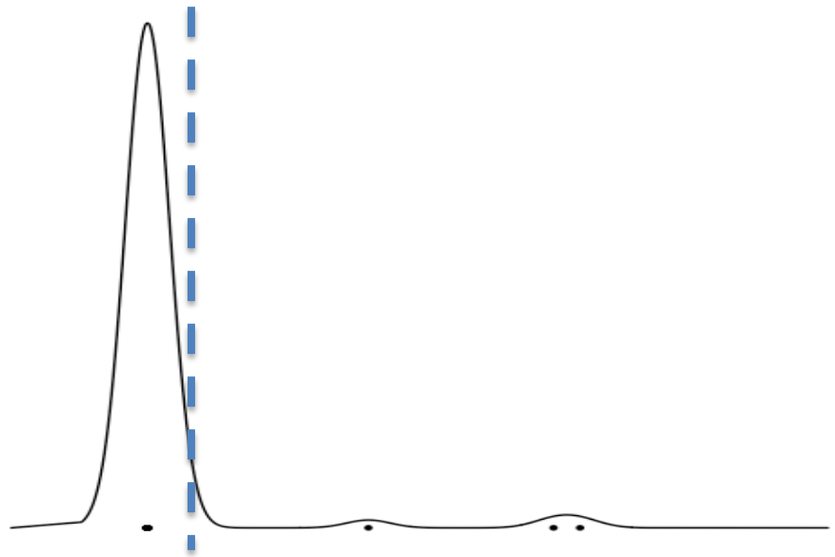
Cellular responses can vary substantially between “identical” cells.

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

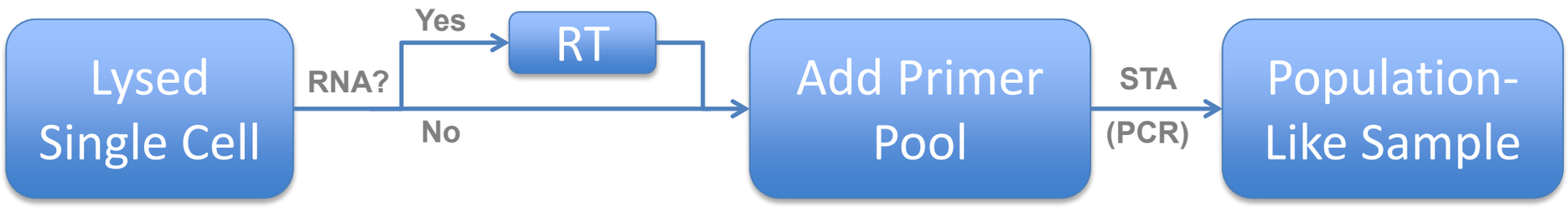
Method	Species?	Endogenous?	Real-Time?	# probes?	Amplification	Advantages	Disadvantages
FISH	DNA, RNA	Yes	No	<5	No	Spatial Info Lots of cells	Expensive
In-Situ Sequencing	DNA, RNA	Yes	No	Many	Yes	Spatial Info Lots of cells	Bias Slow
Single-cell (RT)-PCR	DNA, RNA	Yes	No	<500	Yes	Simple	Need to know targets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

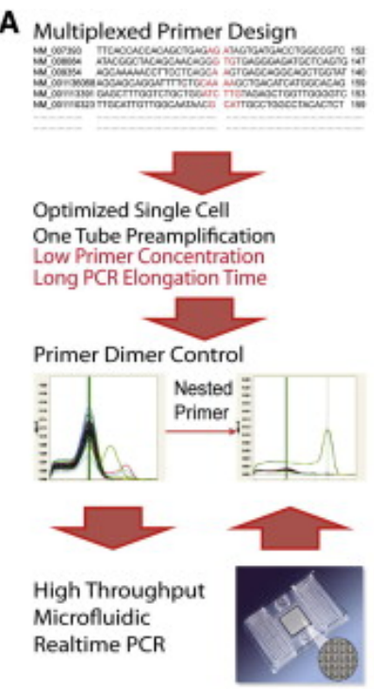
Pros: we can increase our multiplexing dramatically!

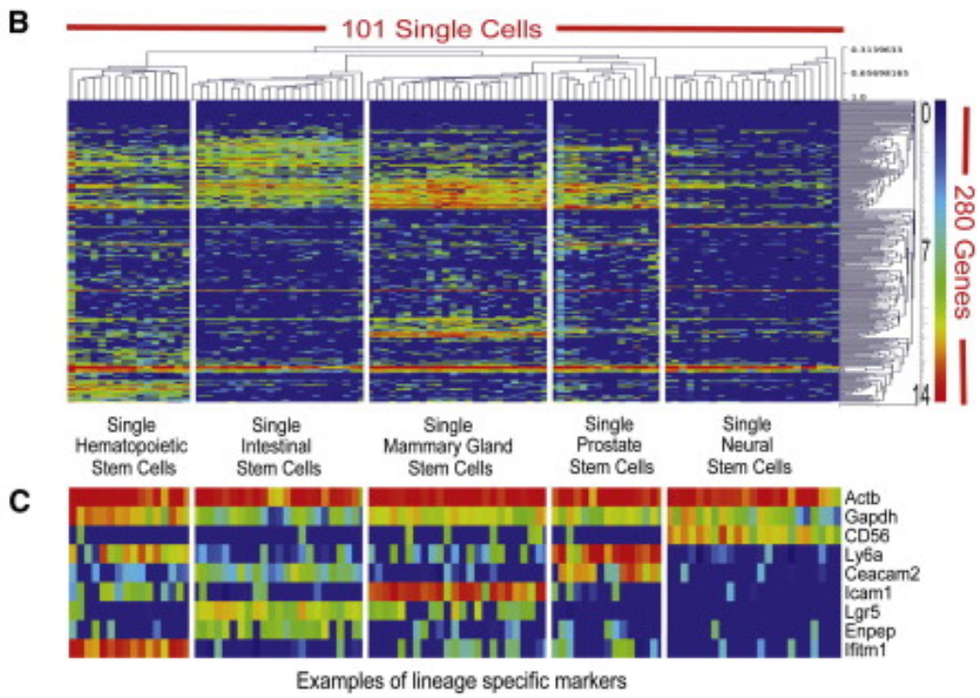
Cons: no dynamics!

Table adapted, in part, from: Raj & van Oudenaarden, *Annu. Rev. Biophys.* 2009.

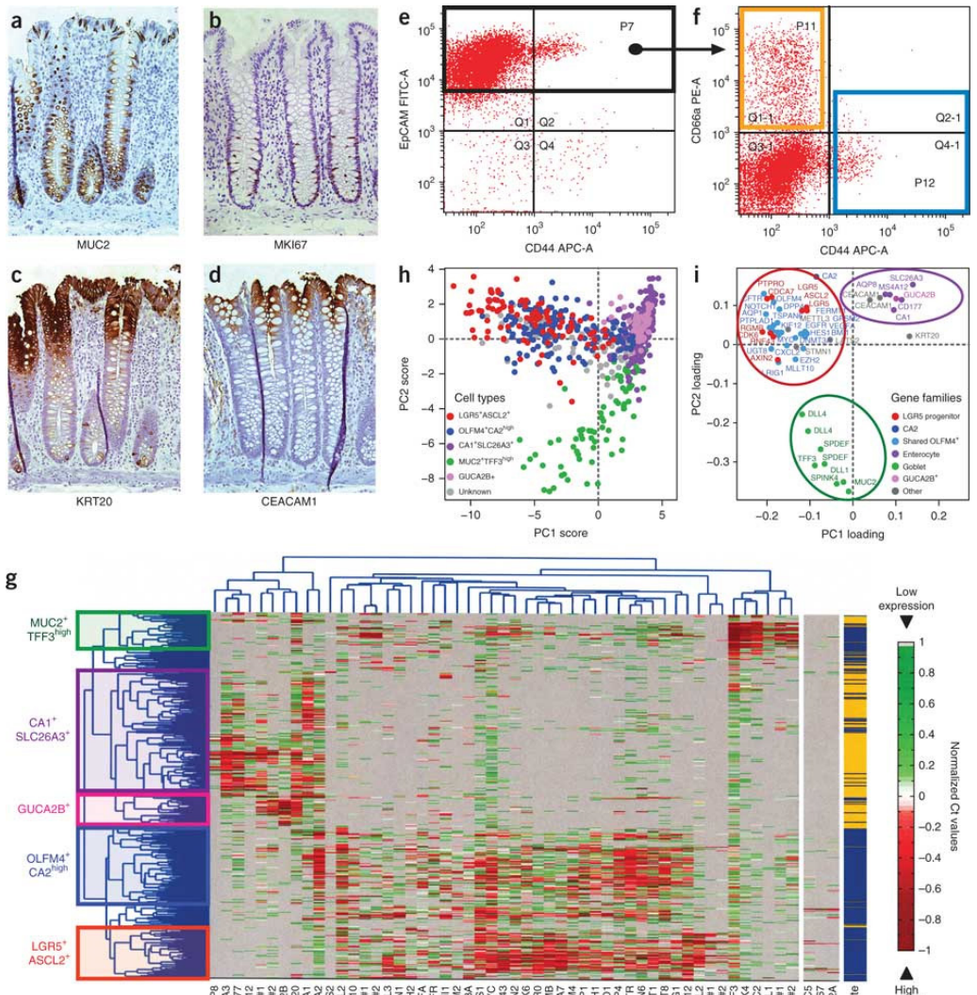
Foundational technology: (RT)-PCR





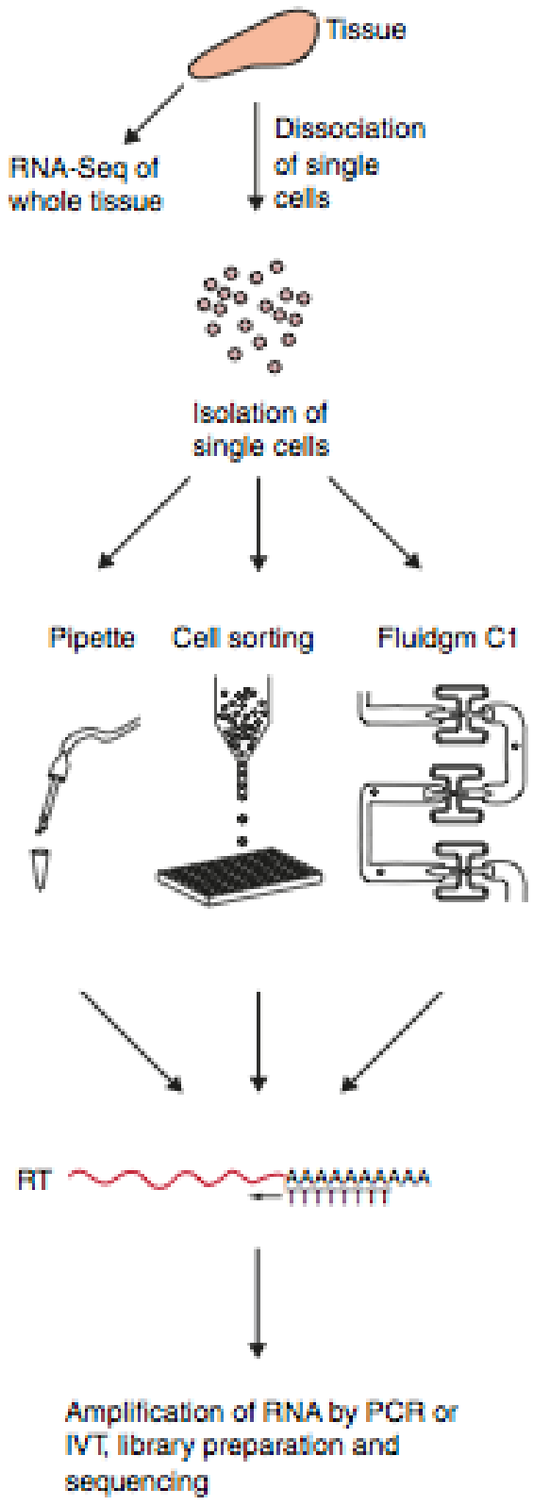


G Guo et al, Cell Stem Cell, 2013; 13: 492-505.

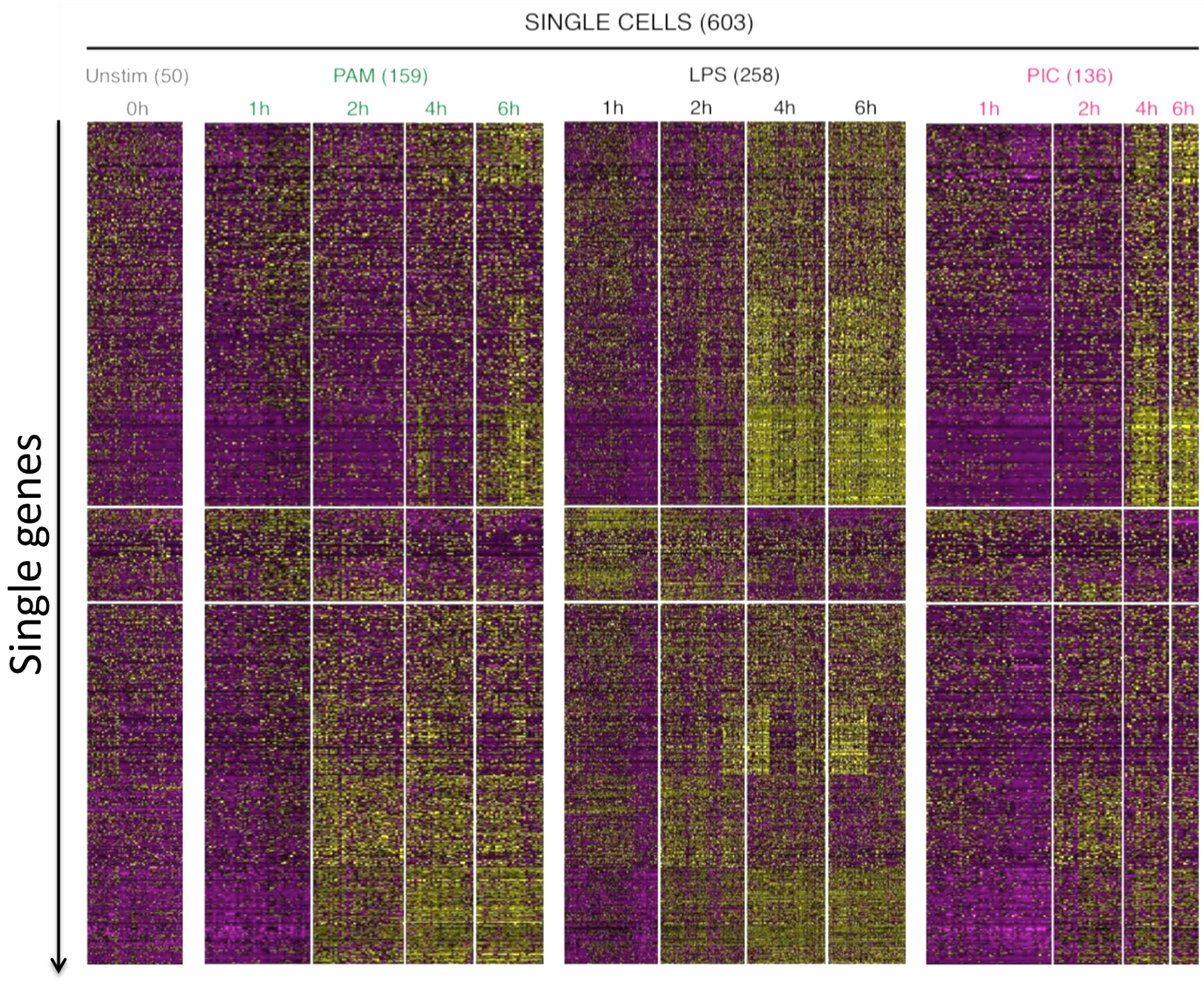


P Dalerba et al, Nature Biotech, 2011; 29: 1120-1127.

Scaling up: Single-cell RNA-Seq

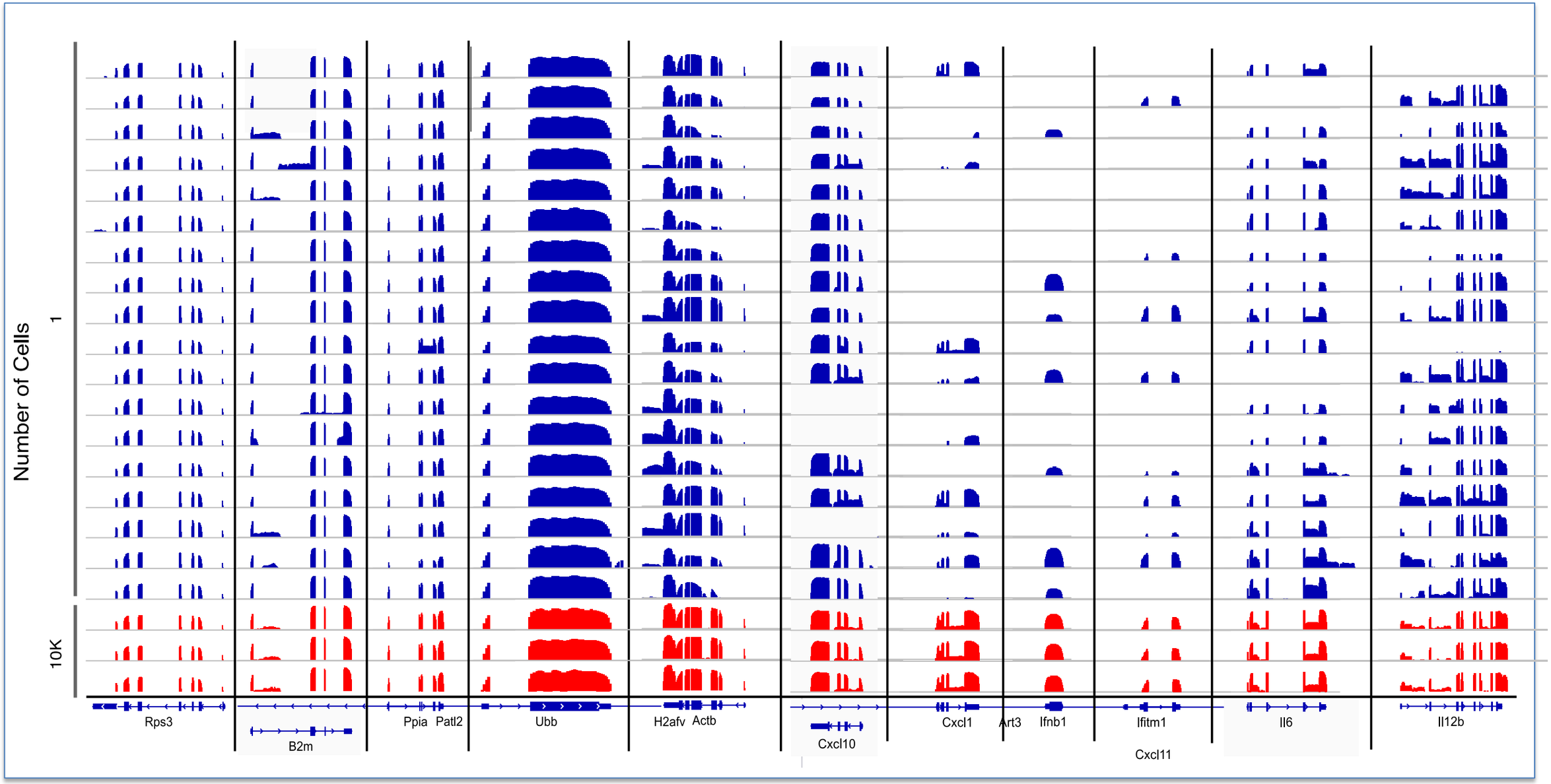


Avital et al., *Genome Biology* (2014).

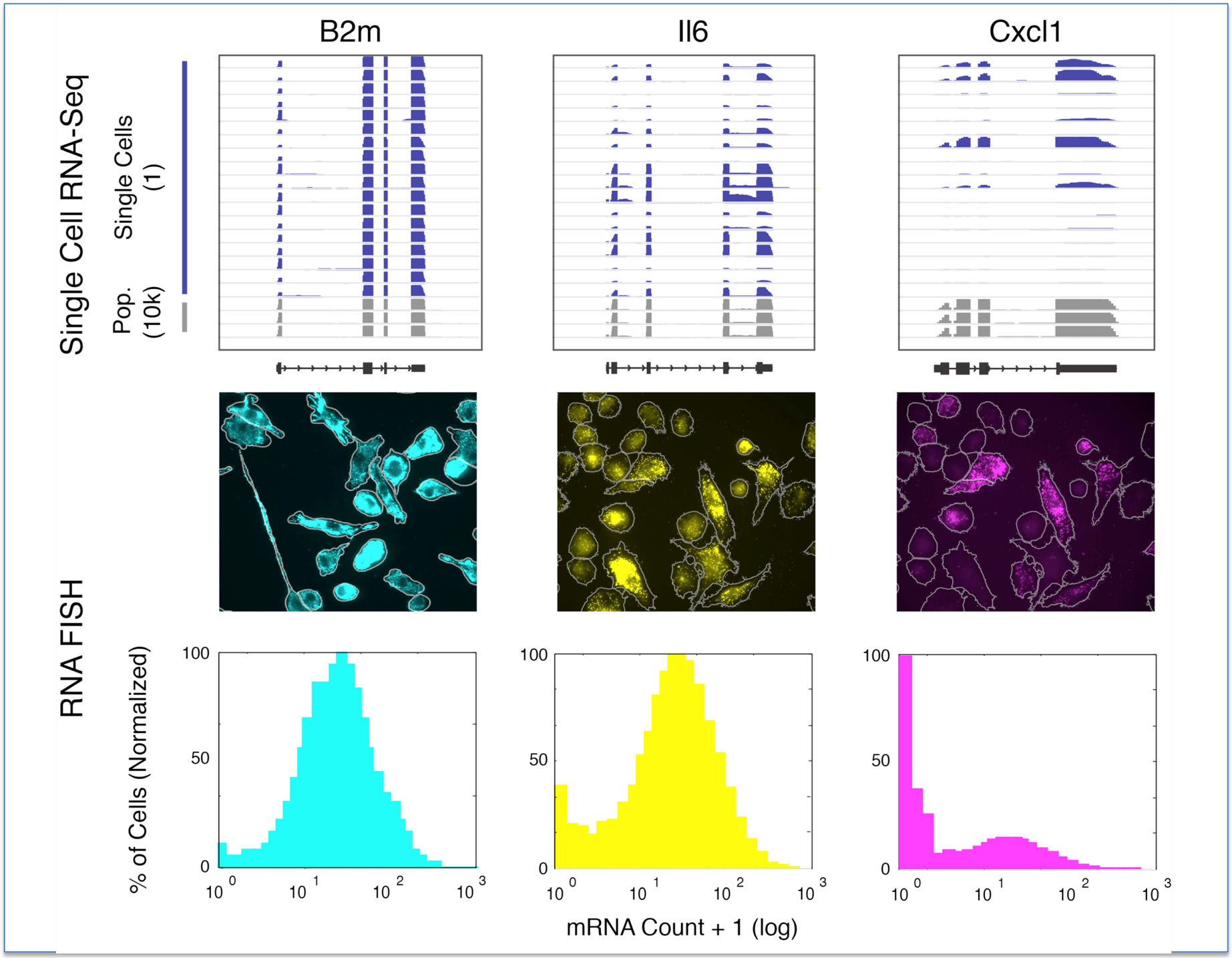


Shalek et al., *Nature* (2014).

scRNA data looks like RNA-seq



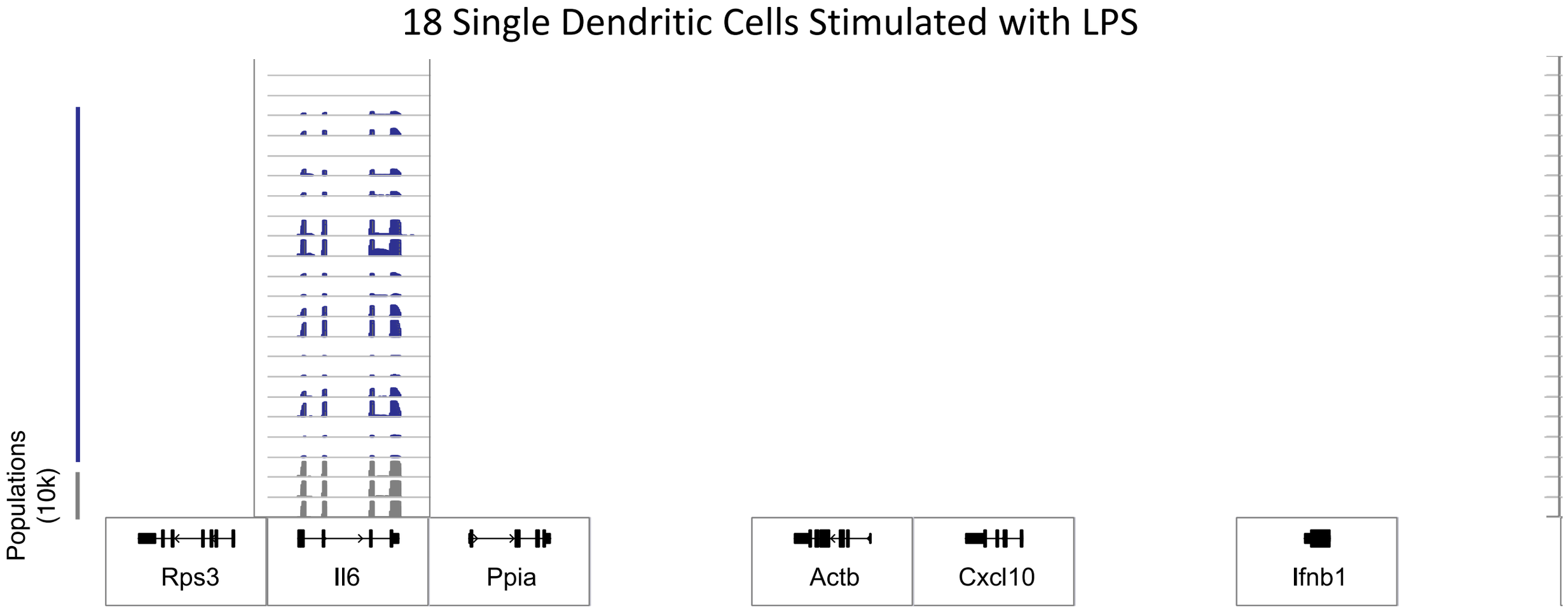
RNA FISH Validates Single-Cell RNA-Seq



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

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

Single-Cell RNA-Seq captures inter-cellular variability

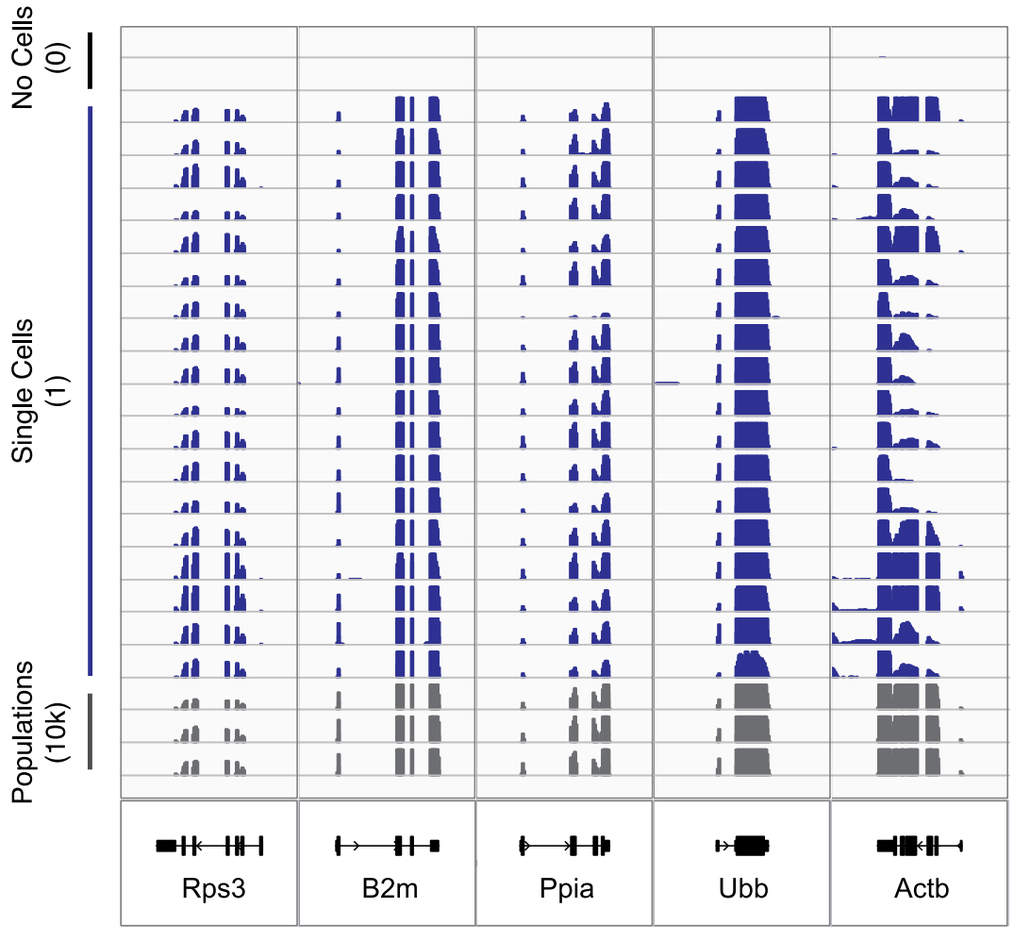


We can profile the full transcriptomes of cells with SMART-Seq.

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

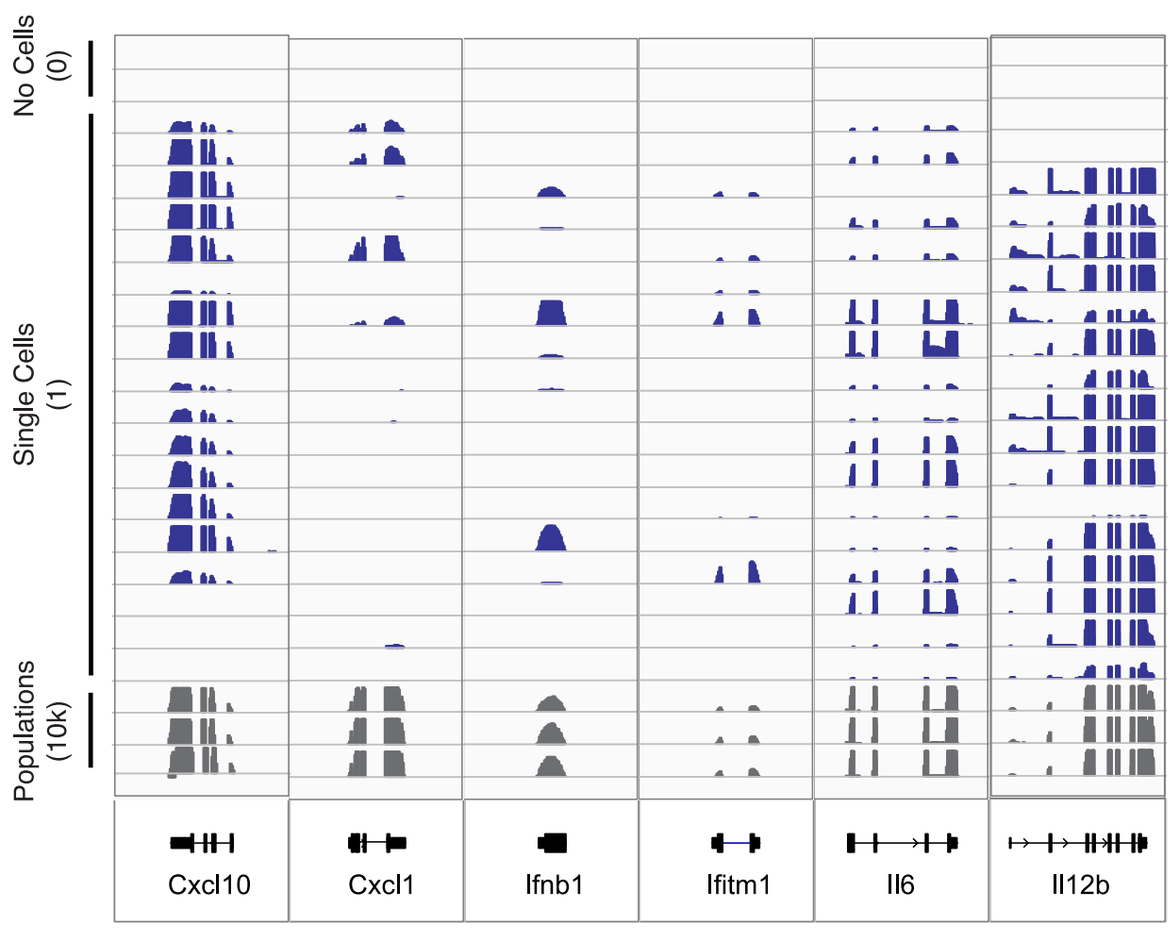
Housekeeping vs. variable genes

Least Variable Genes



Term	P-Value
Structural constituent of ribosome	1.66e-51
Structural molecule activity	1.16e-33

Most Variable Genes



Term	P-Value
Defense response	2.23e-29
Inflammatory response	7.99e-25

Screenshots are log scale in IGV

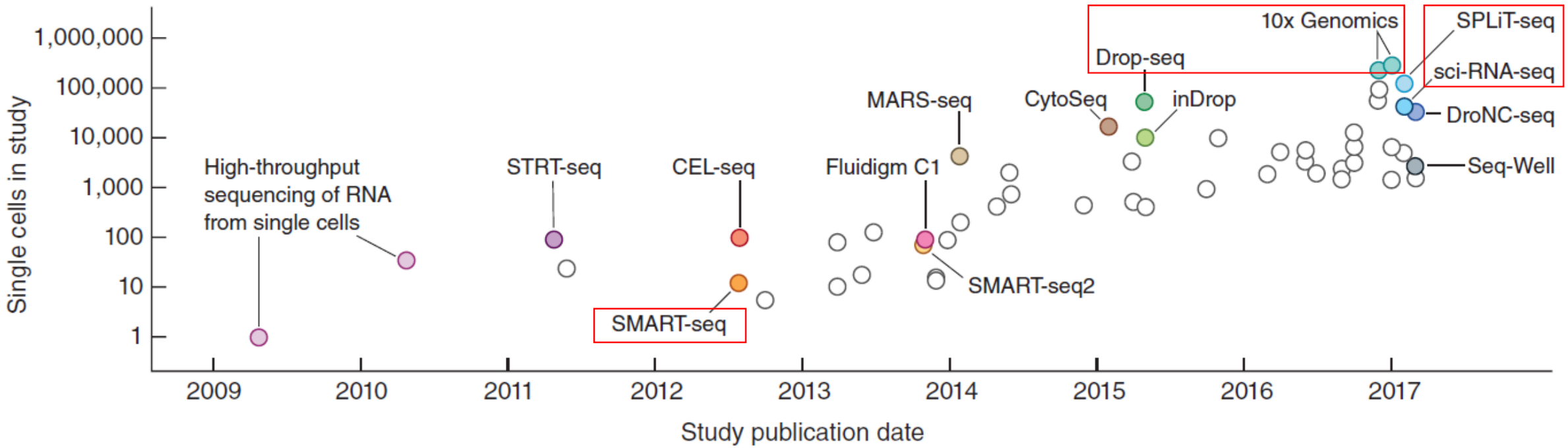
Housekeeping & ribosomal genes are among the least variable, while immune response elements are among the most variable.

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

2. Scaling up scRNA-seq technology

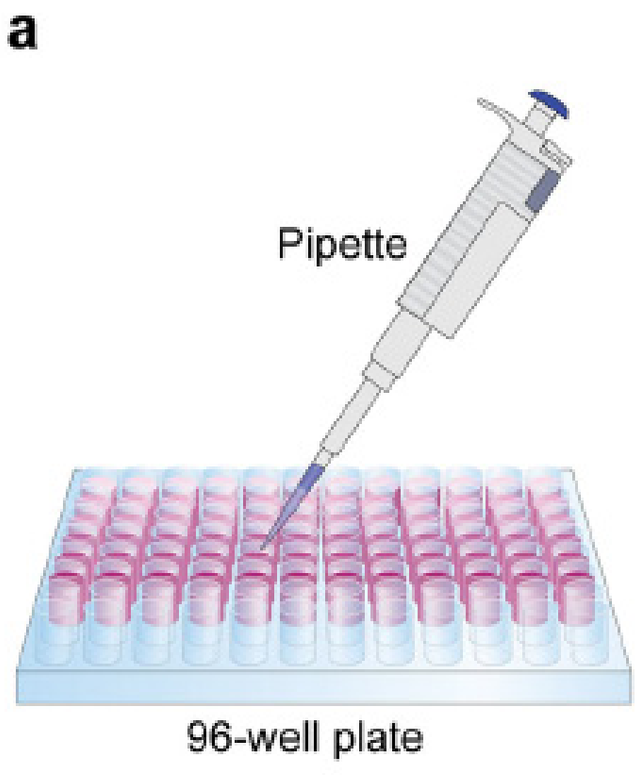
Exponential scaling of single-cell RNA-seq in the past decade

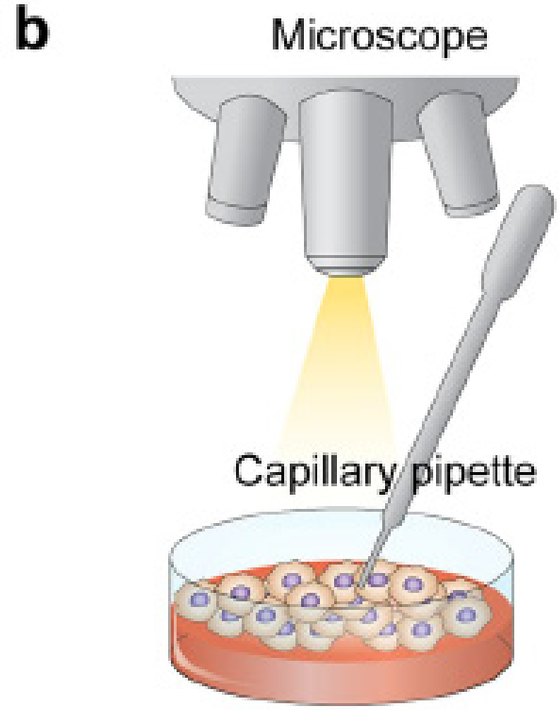
Cell numbers reported in representative publications by publication date. Key technologies are indicated.

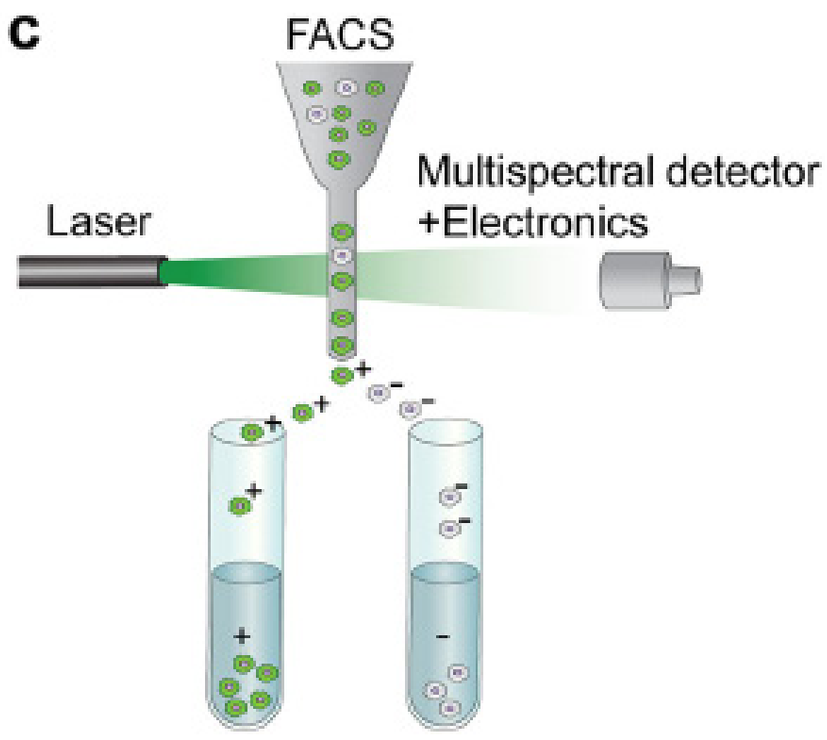


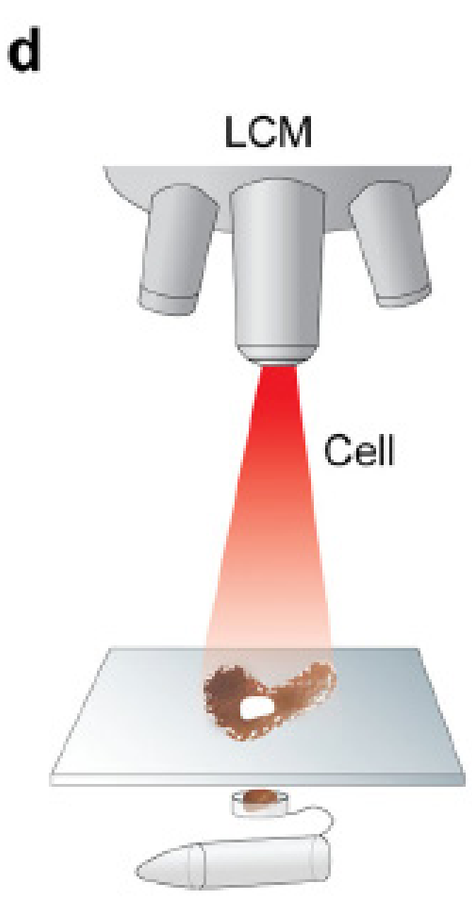
Svensson V, Vento-Tormo R, Teichmann SA. Nat Protoc. 2018 Apr;13(4):599-604.

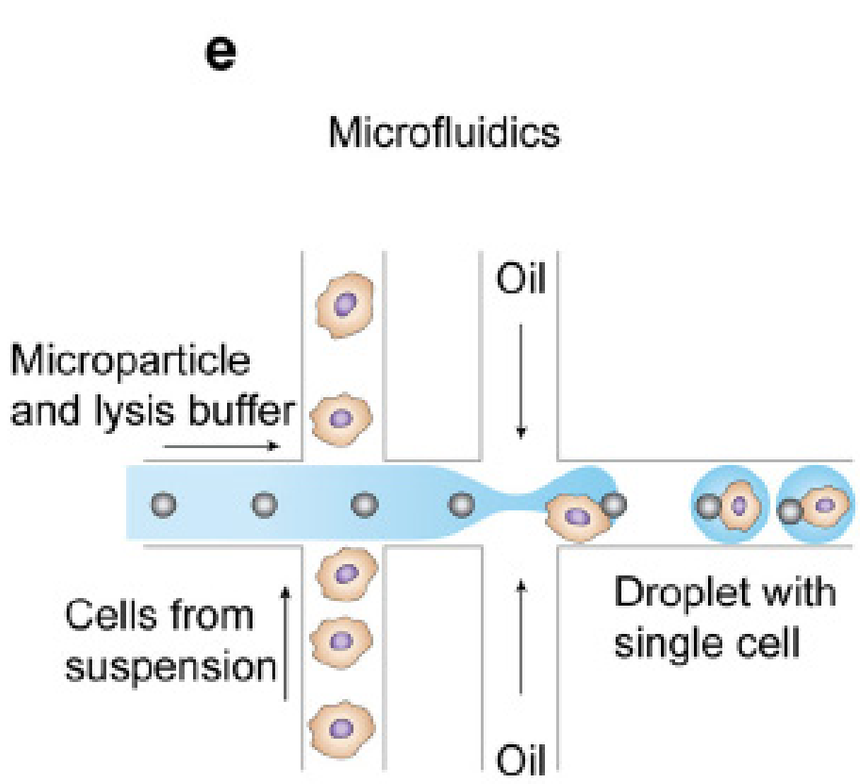
Evolution of methods for isolating single cells for profiling

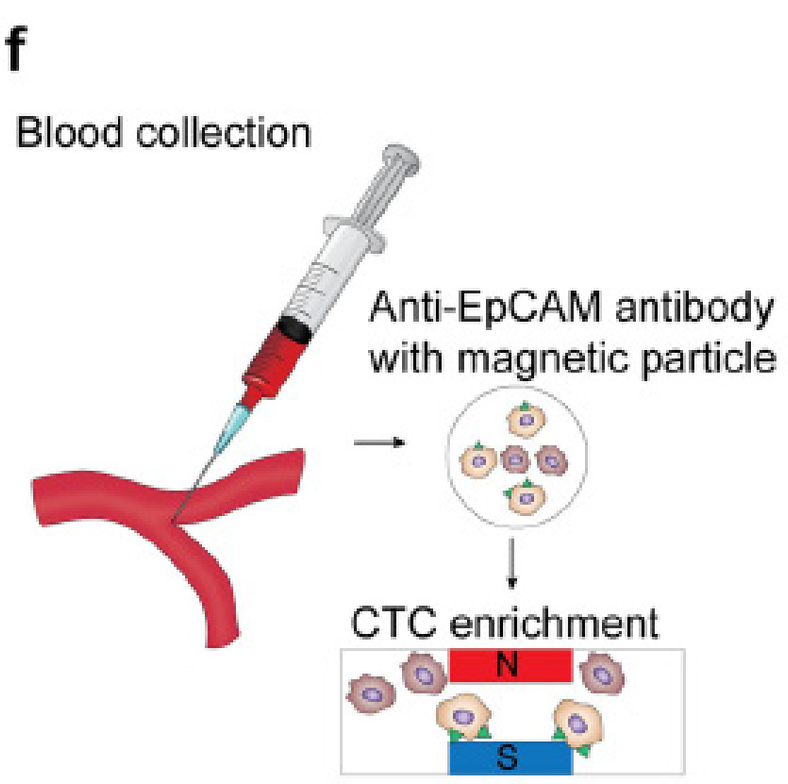




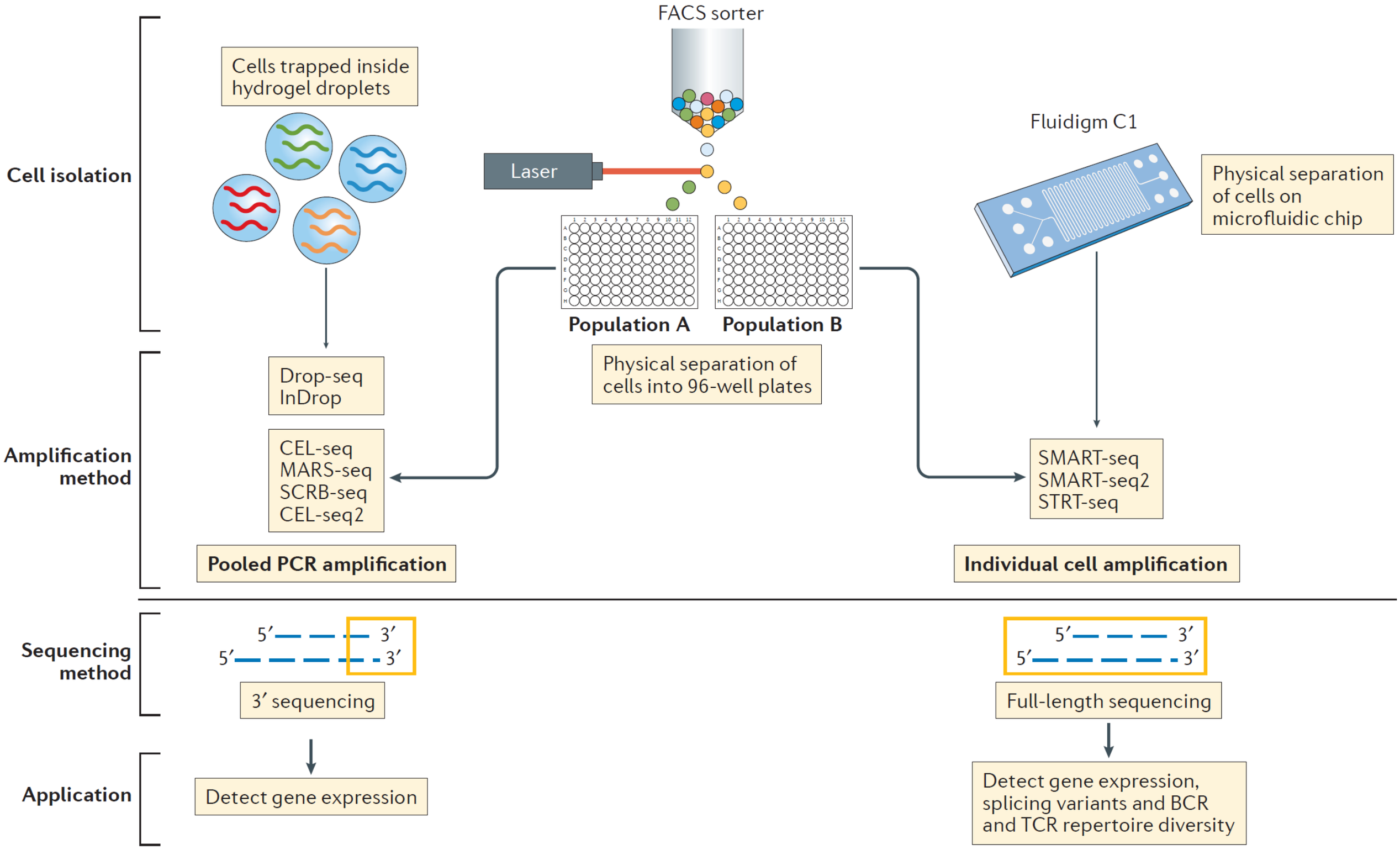








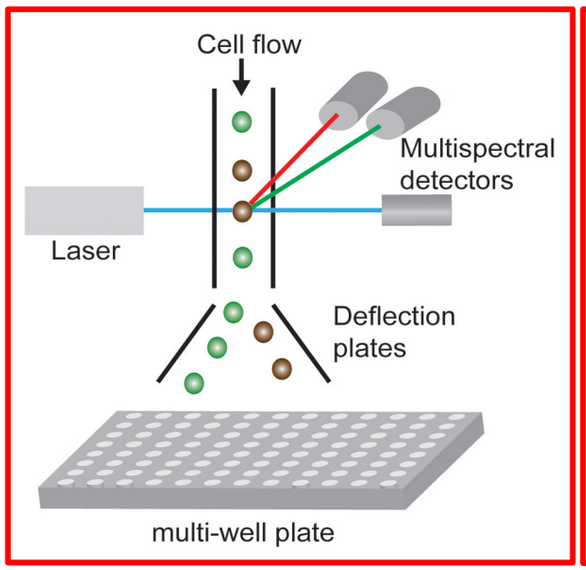
All methods seek to: separate cells, amplify RNA, sequence

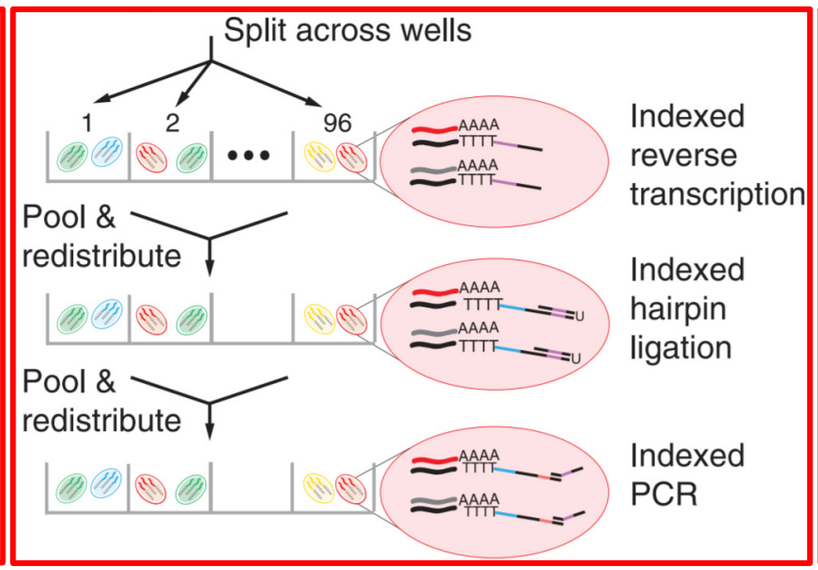


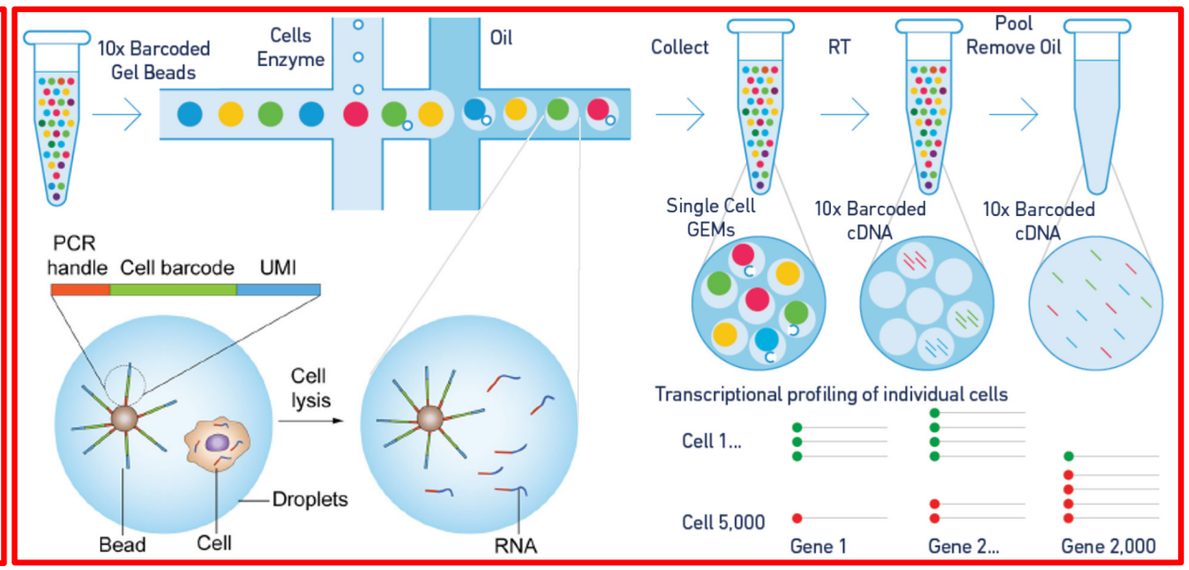
Paplexi et al. 2017

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	6,000–9,00 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

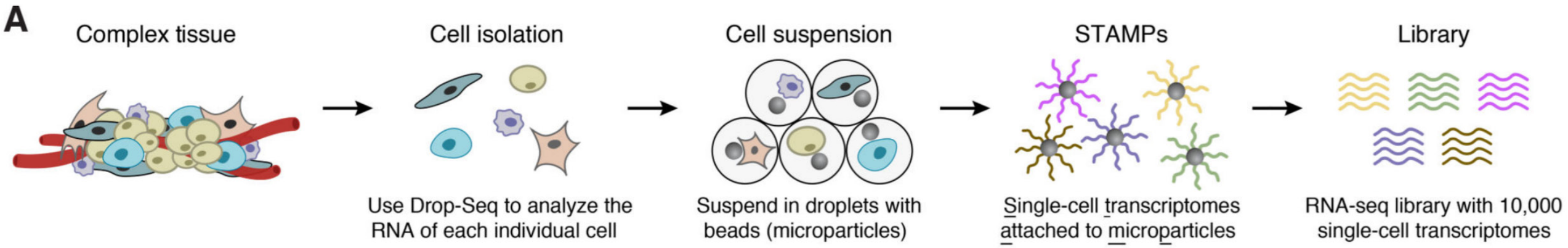


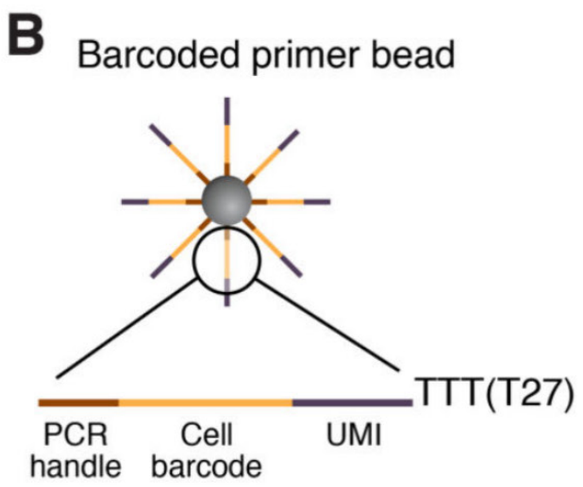


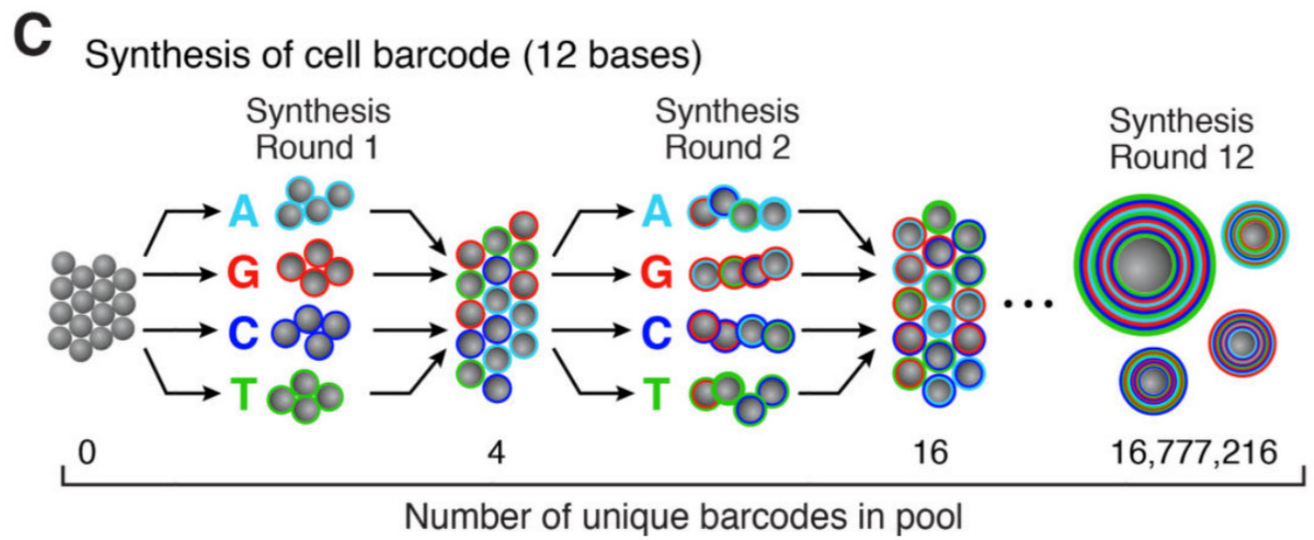


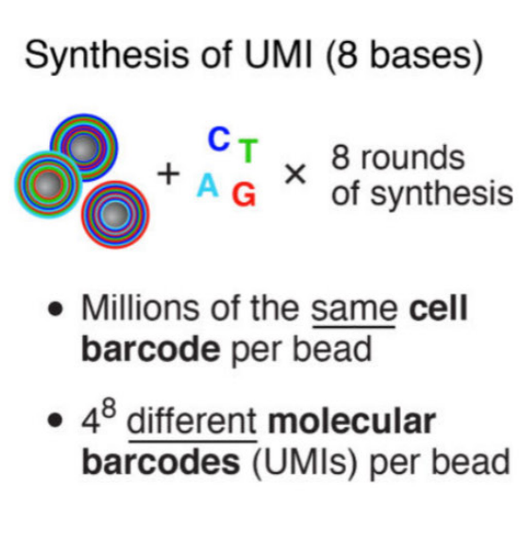
CEL-seq, cell expression by linear amplification and sequencing; CyTOF, cytometry by time of flight (mass cytometry); FACS, fluorescence-activated cell sorting; InDrop, indexing droplets sequencing; MARS-seq, massively parallel single-cell RNA sequencing; qPCR, quantitative PCR; SCRB-seq, single-cell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

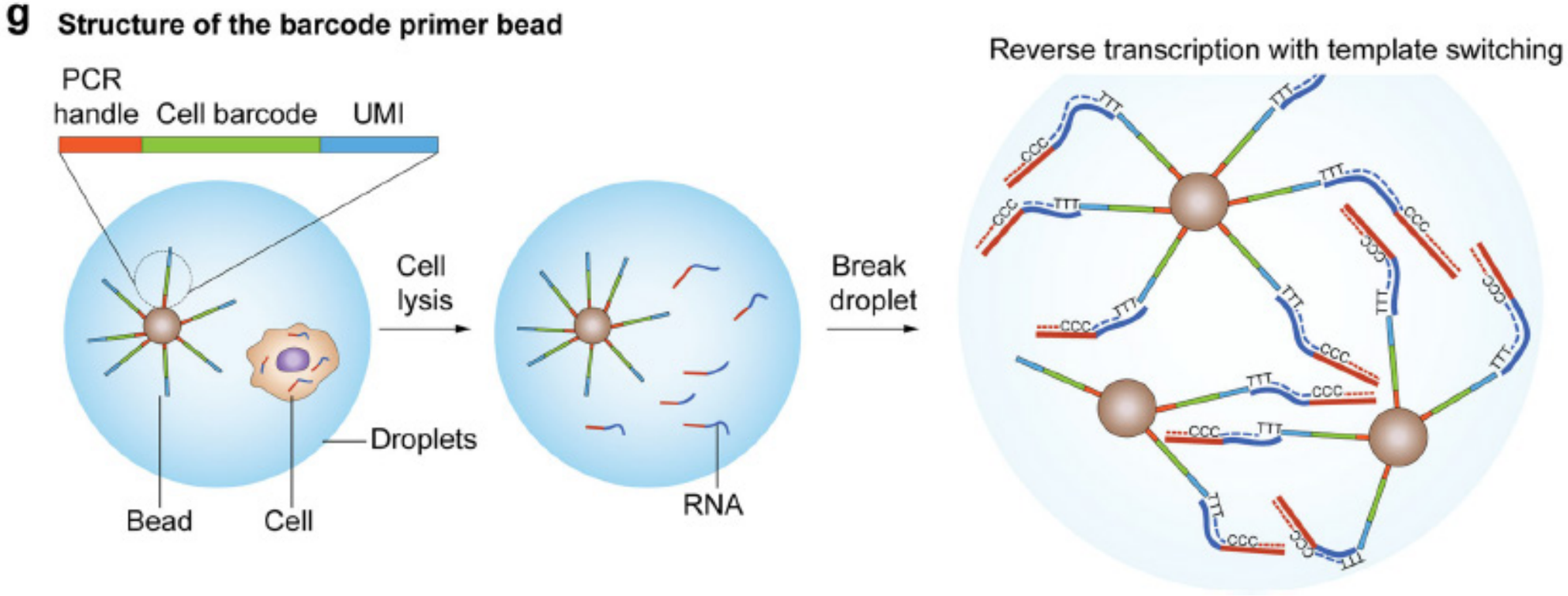
Drop-seq: Droplets as reaction chambers (10x)



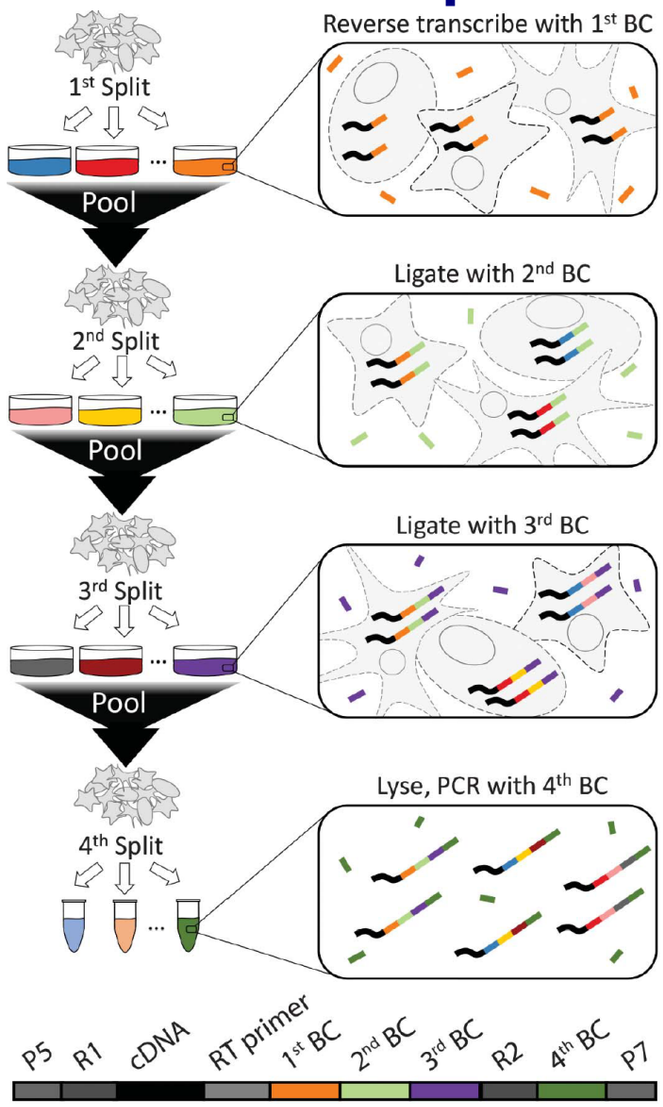








SPLIT-seq/sci-RNA-seq: Sequential combinatorial barcoding

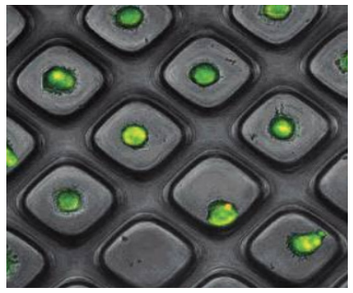


- Single cells never individually isolated
- Instead: fixed, and mRNA is manipulated in situ inside each cell
- Split cells into ~100 wells (e.g. 96 or 384-well plate) with unique barcodes in each well
- Labels all cells with a first barcode, for that well. Chance of same barcode: 1/100
- Pool cells, 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

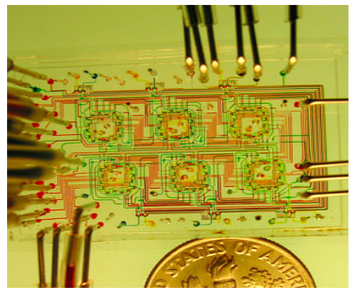
Single-cell Profiling technologies

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



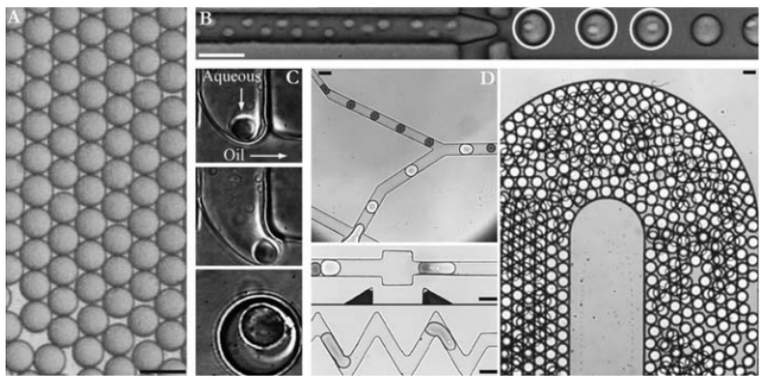
Passive wells



Active pumps and valves

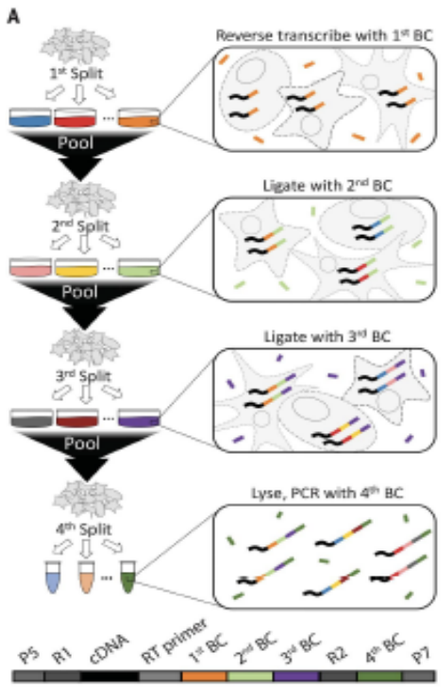
2. Droplets (Drop-seq, ddPCR)

- 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.



Single-Cell Expression Profiling Pipeline

1. Cell Harvest

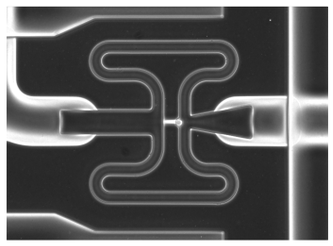
- Harvest cells in media
- Pre-enrich (FACS)

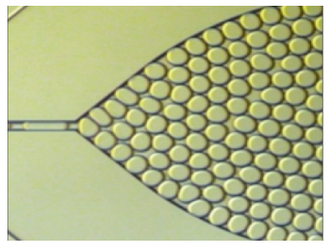
Cell work

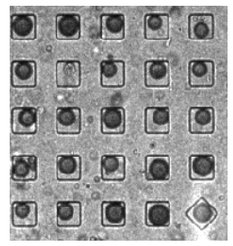
2. Single Cell Preparation

(a) C_1: Cells → Whole Transcriptome Amplification
(b) Multiwell Plates
(c) Next Gen Technologies

Single cell processing

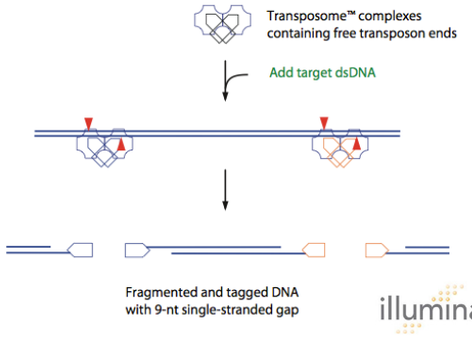






3. Expression Profiling

Library construction, validation, HiSeq

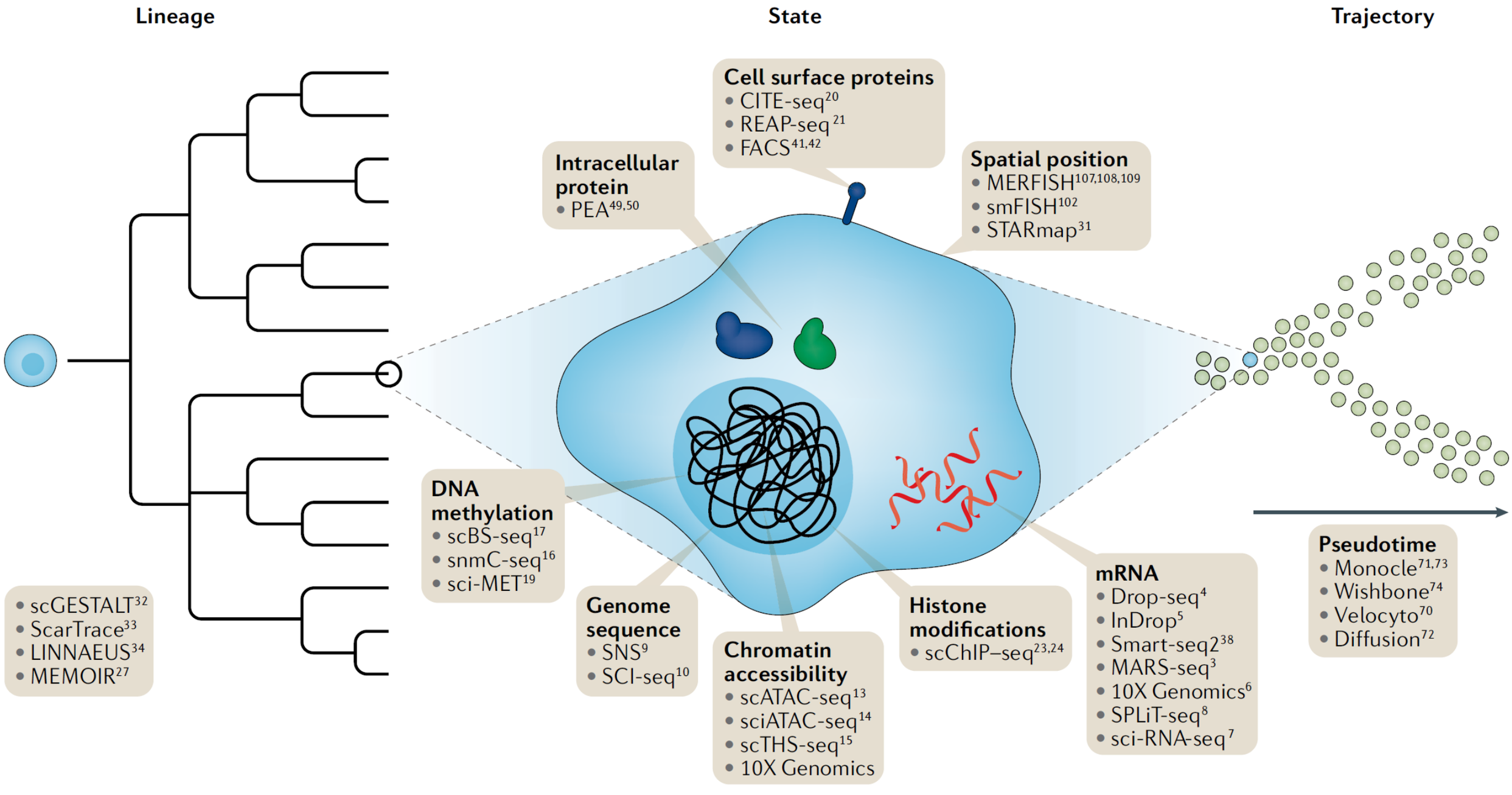


We can go from single cells to aligned reads in less than a day

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

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

Diverse technologies for sc profiling



SINGLE-CELL OMICS

Integrative single-cell analysis

Tim Stuart[1] *and Rahul Satija*[1,2]*

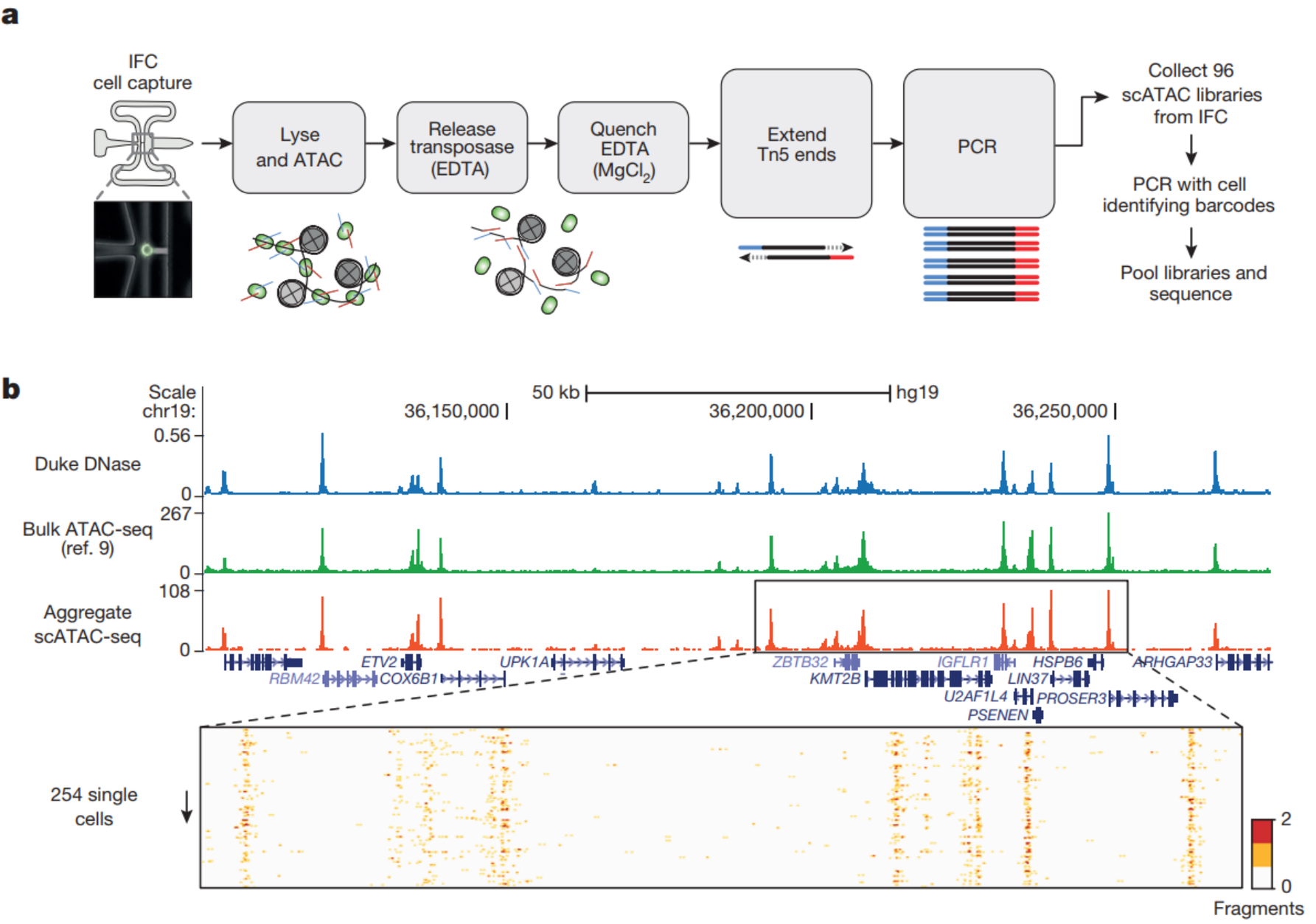
Data types	Method name	Feature throughput	Cell throughput	Refs
Unimodal				
mRNA	Drop-seq	Whole transcriptome	1,000–10,000	4
	InDrop	Whole transcriptome	1,000–10,000	5
	10X Genomics	Whole transcriptome	1,000–10,000	6
	Smart-seq2	Whole transcriptome	100–300	38
	MARS-seq	Whole transcriptome	100–300	3
	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
Multimodal				
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

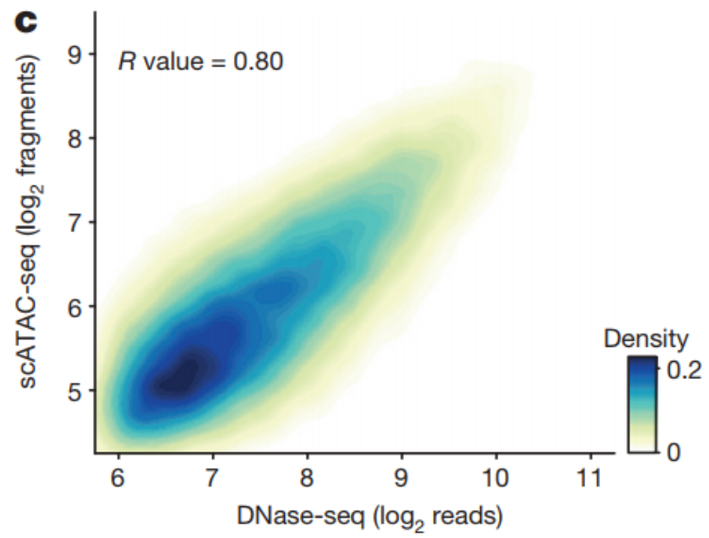
SINGLE-CELL OMICS

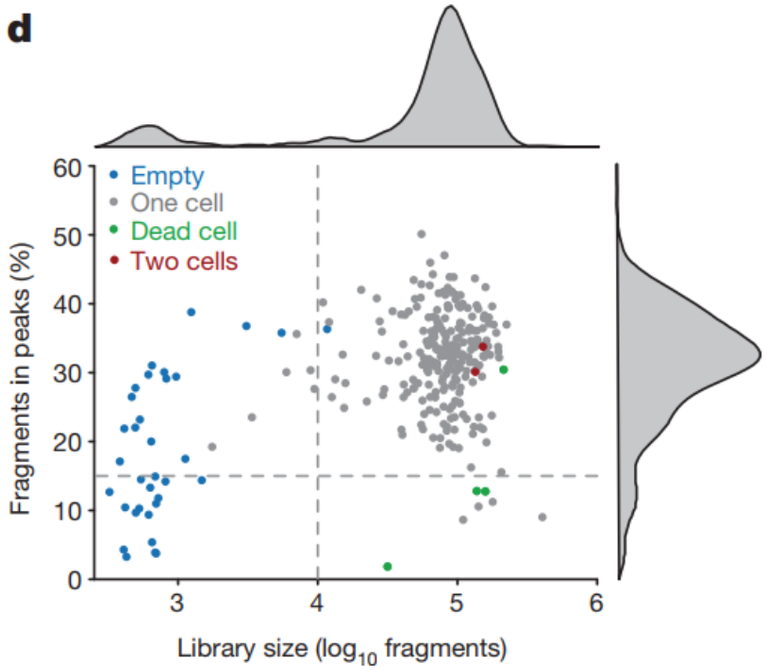
Integrative single-cell analysis

Tim Stuart[1] *and Rahul Satija*[1,2]*

Single-cell Epigenomics (scATAC-Seq)

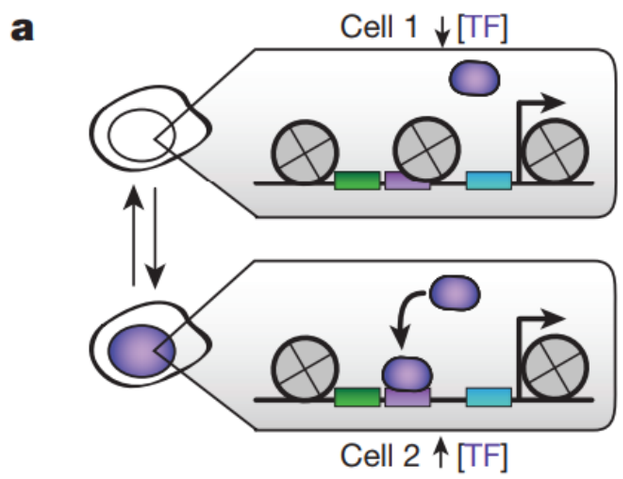


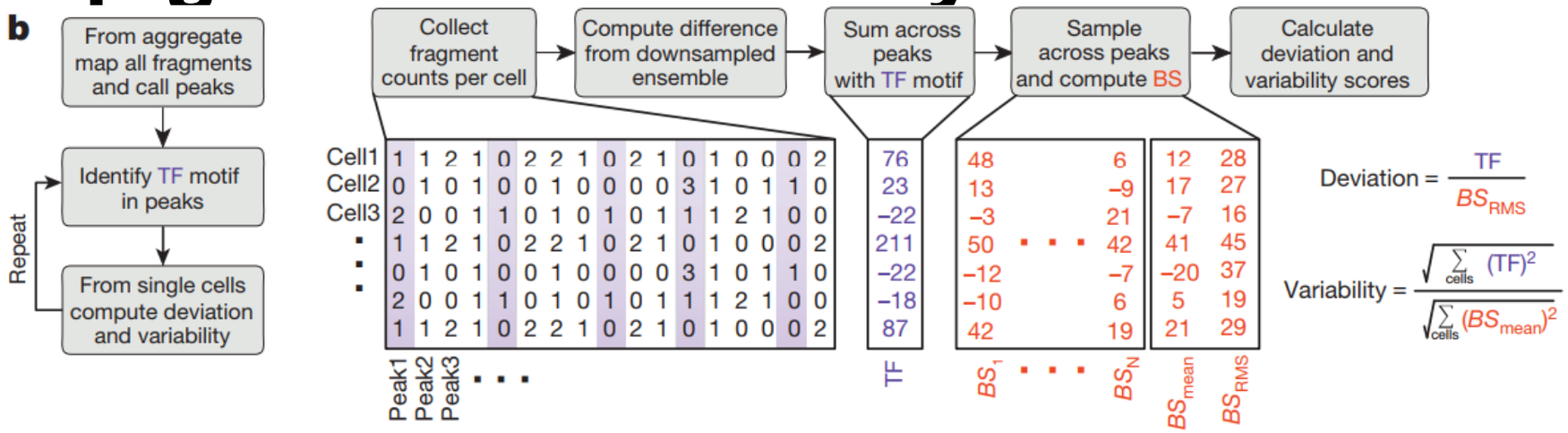


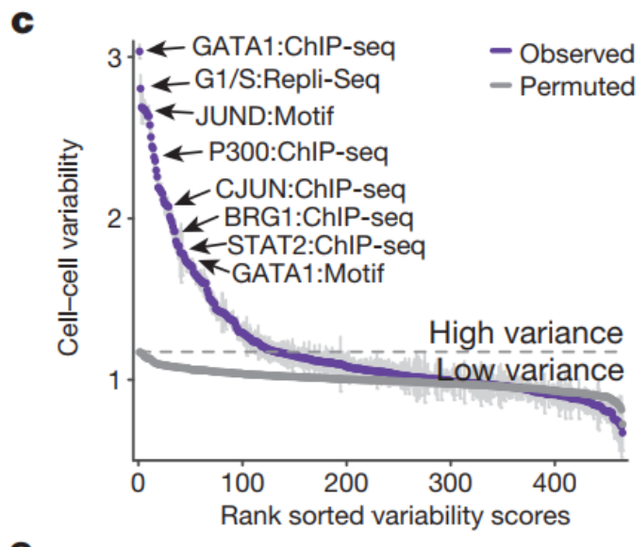


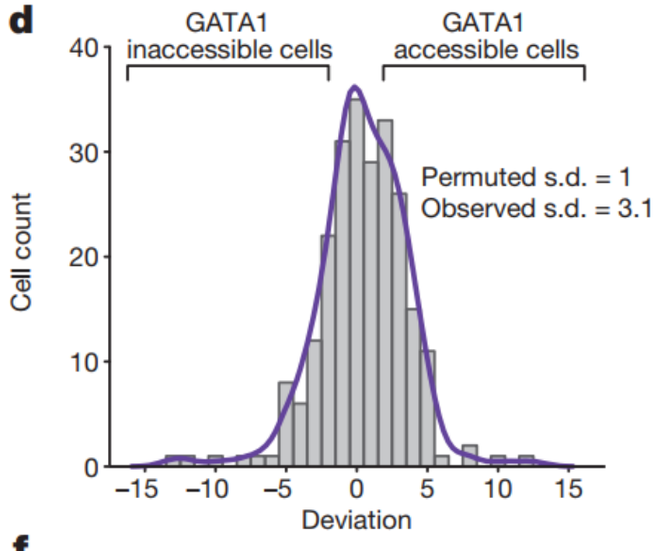
Buenrostro *et al., 2015*

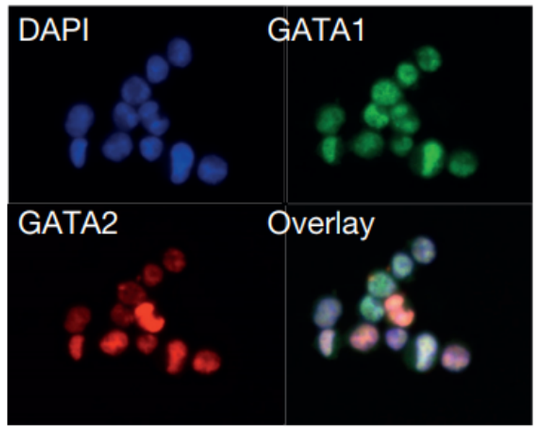
Trans-factors are associated with single-cell epigenomic variability

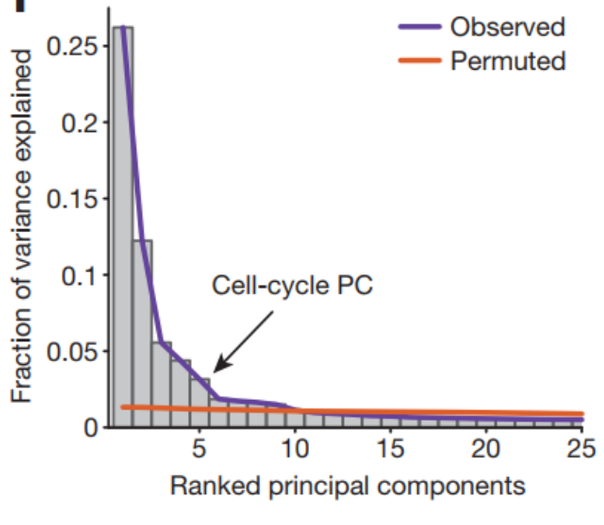


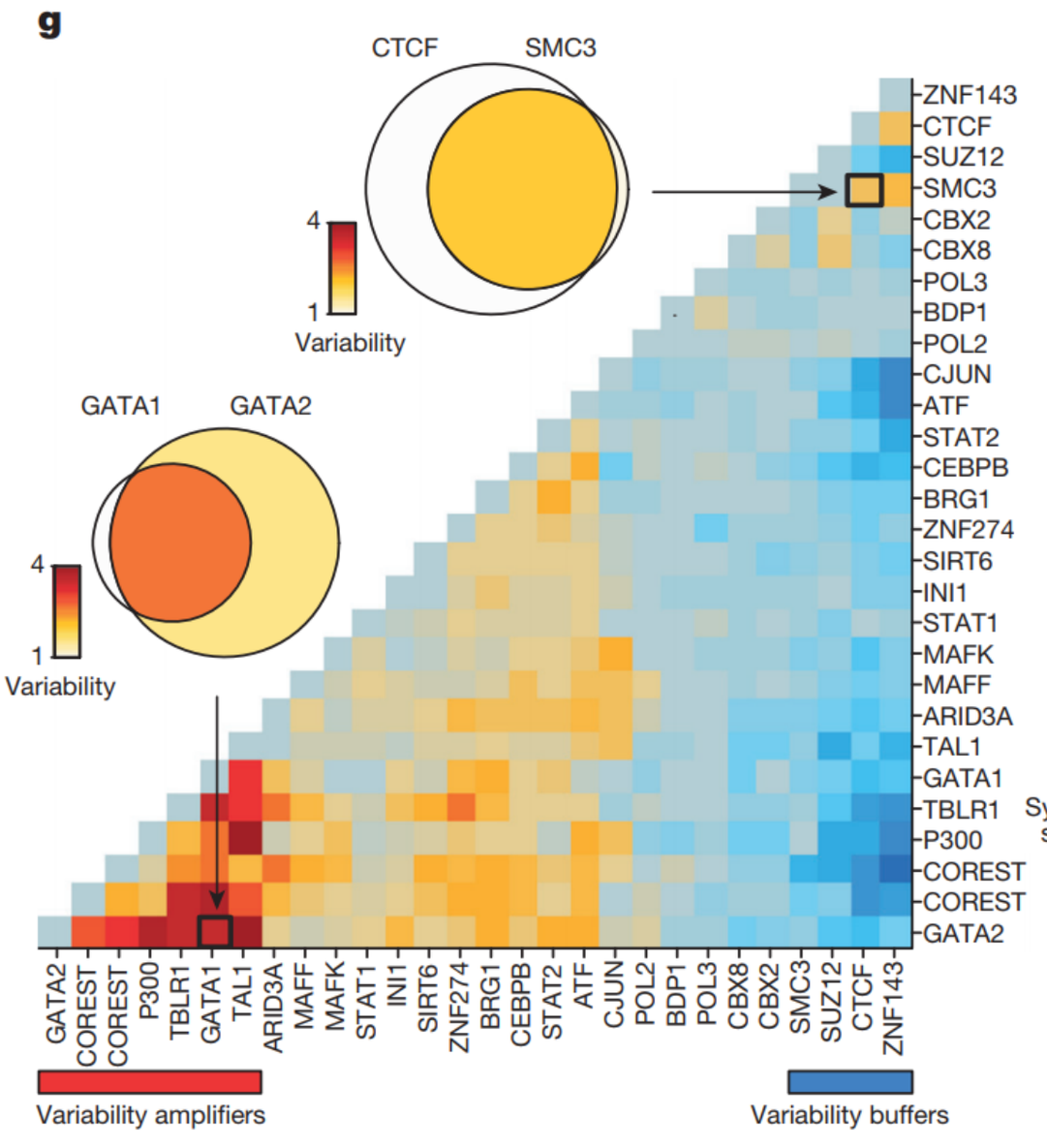




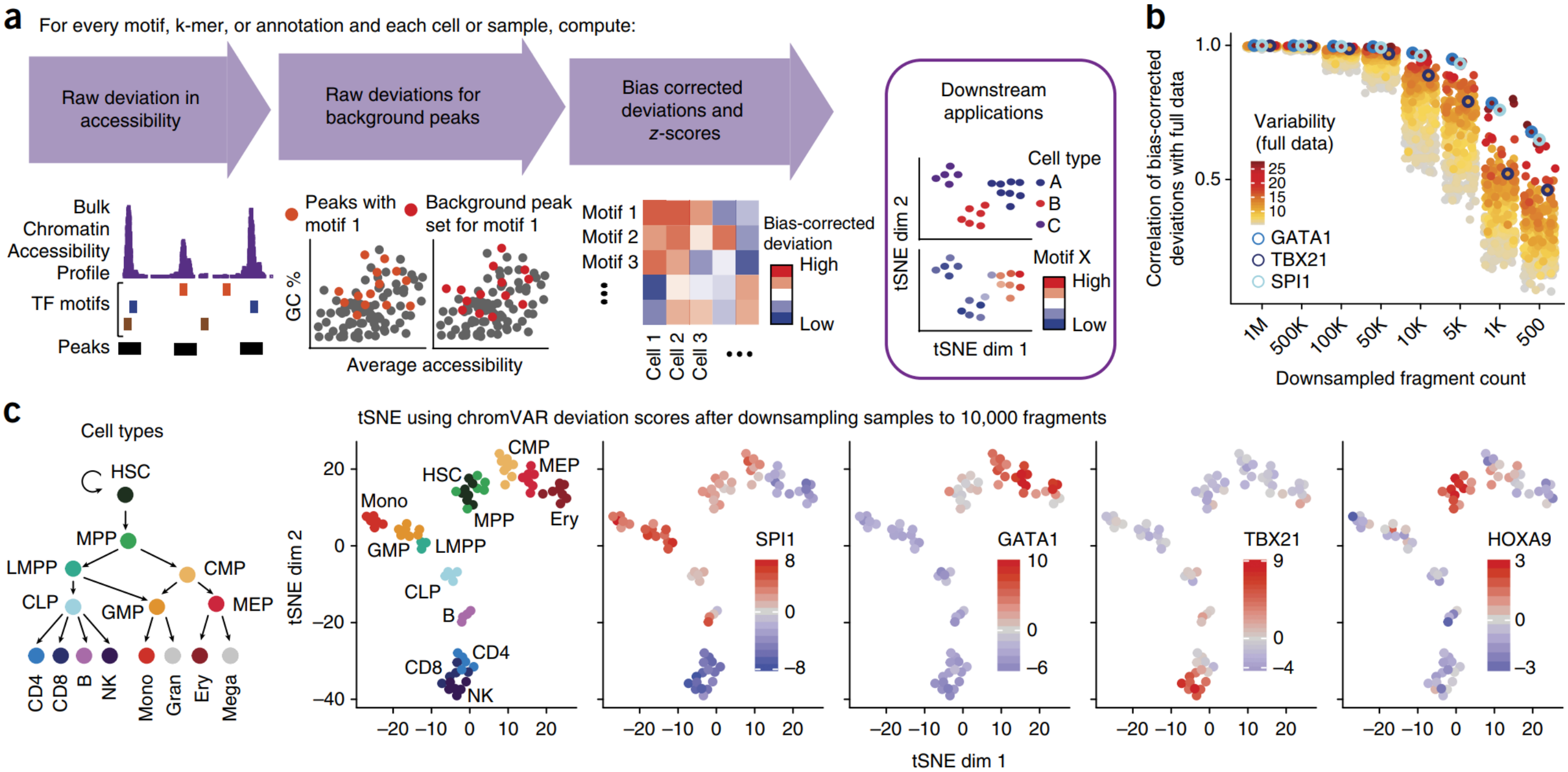




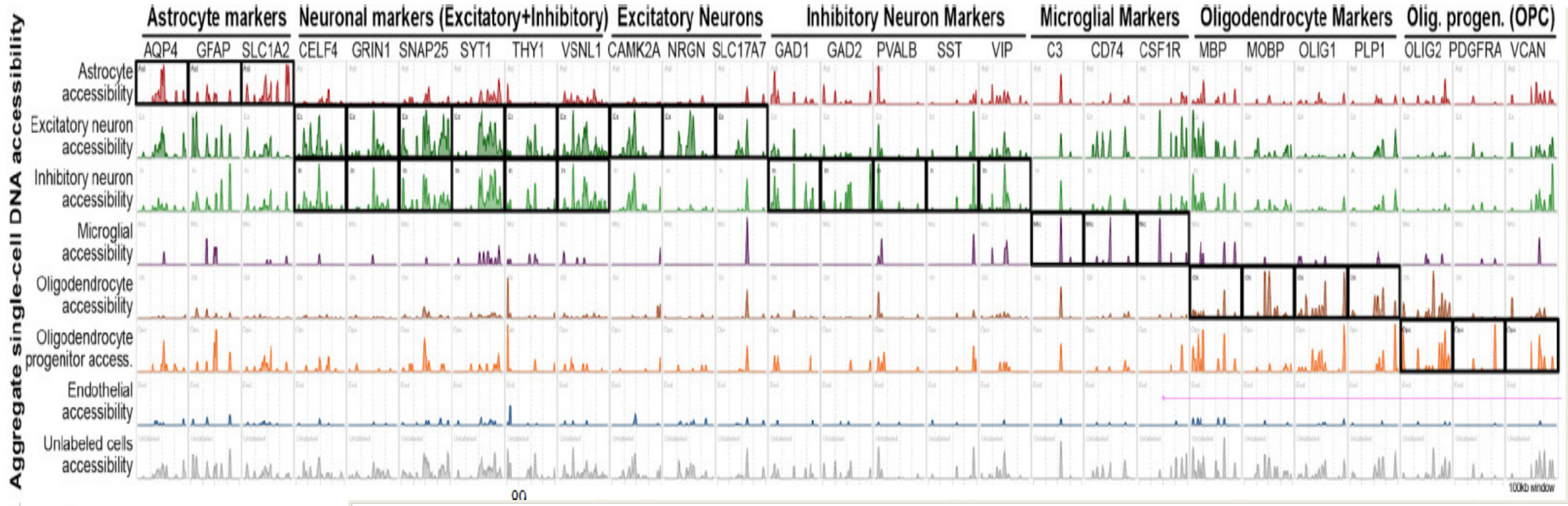


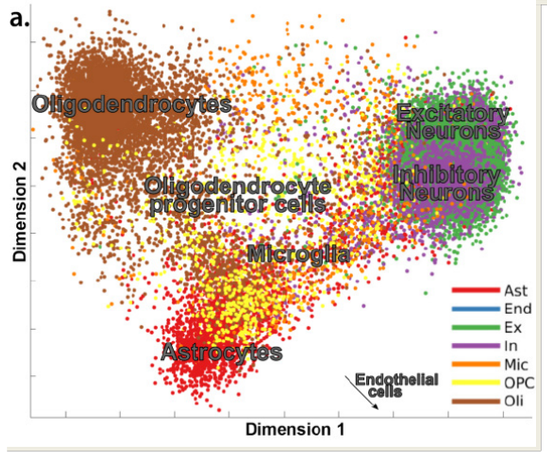


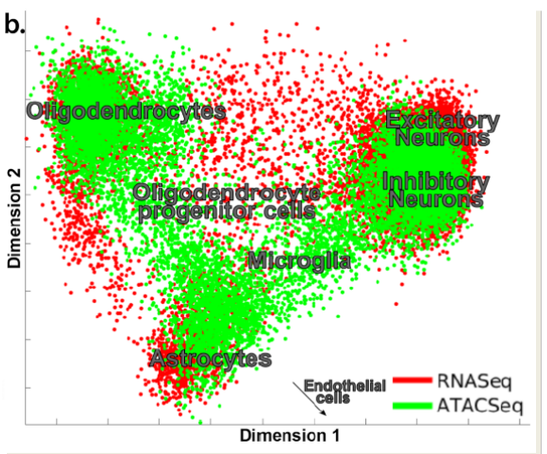
Integrate scATAC + scRNA using ChromVAR

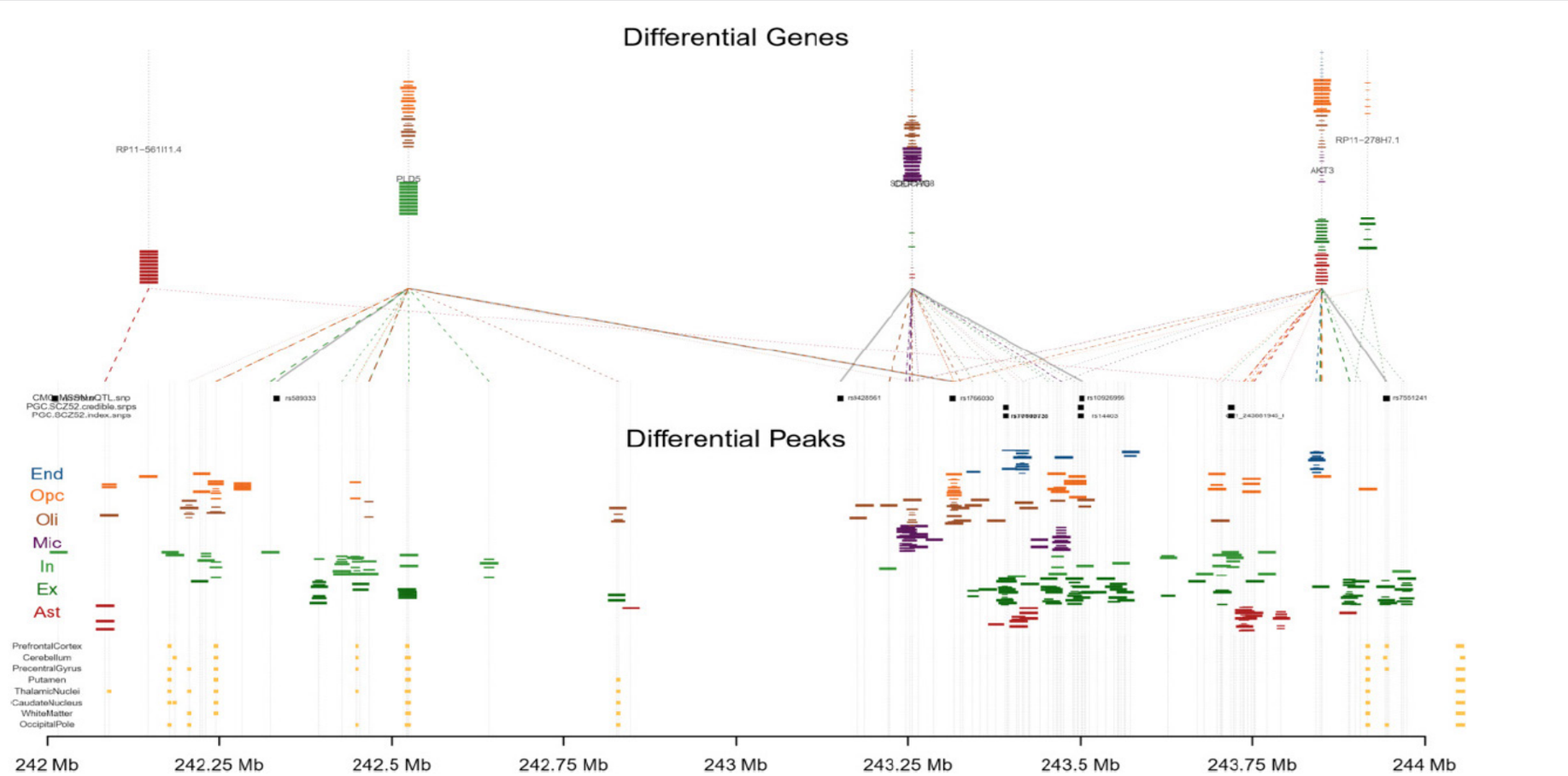


Link single-cell epigenomics and single-cell transcriptomics

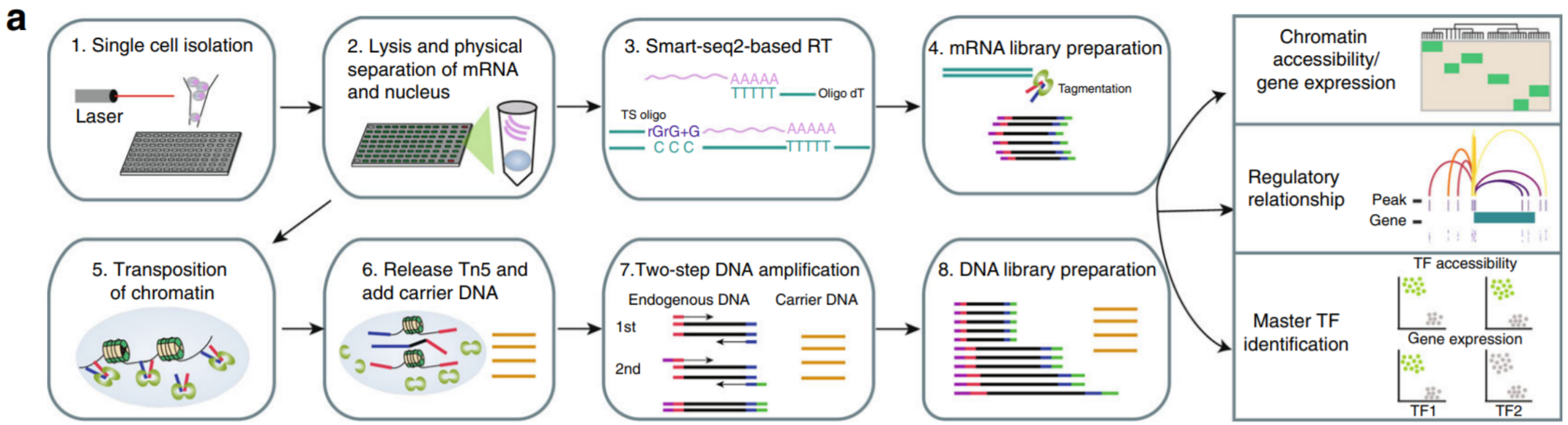


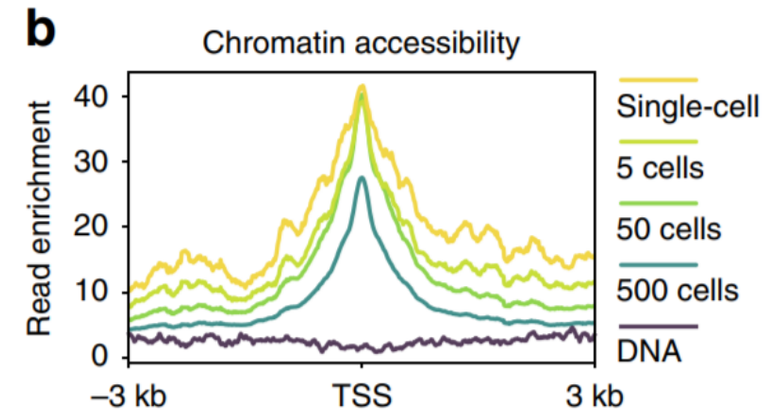


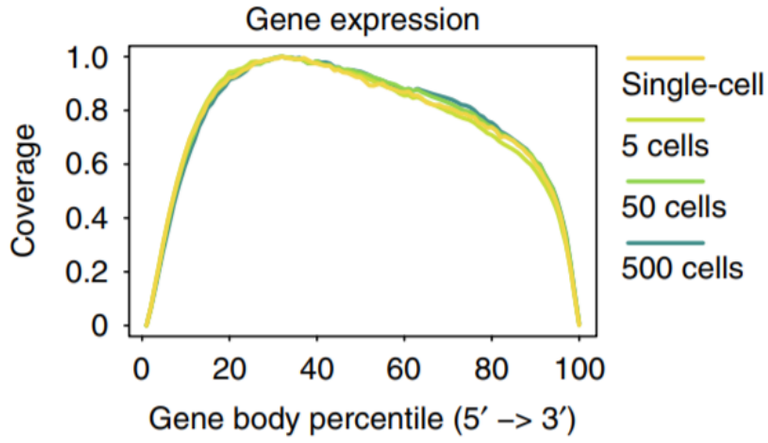


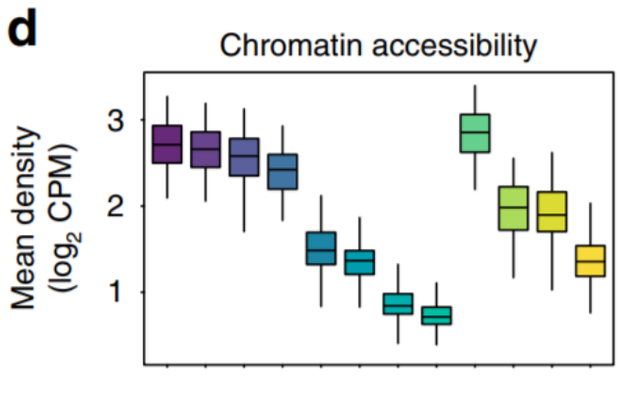


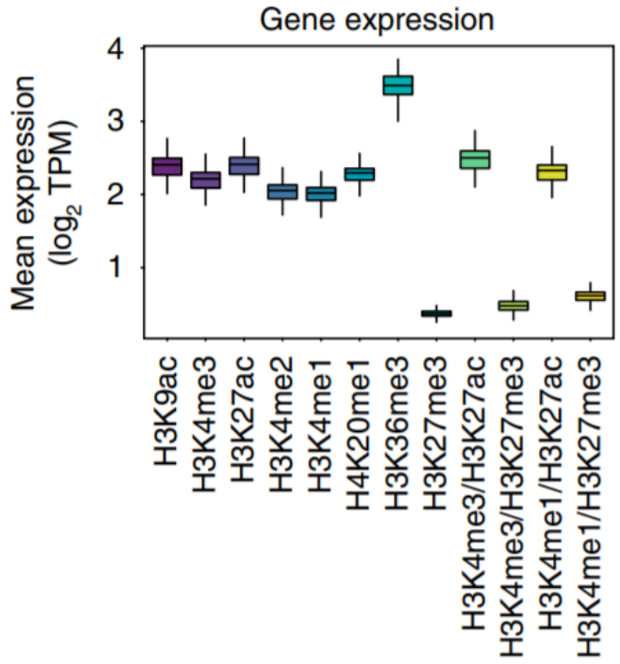
scMulti-Omics: Multiple profiling of the same cell

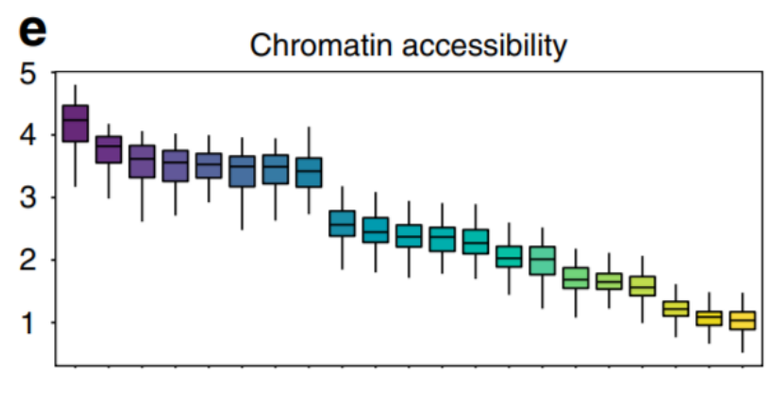


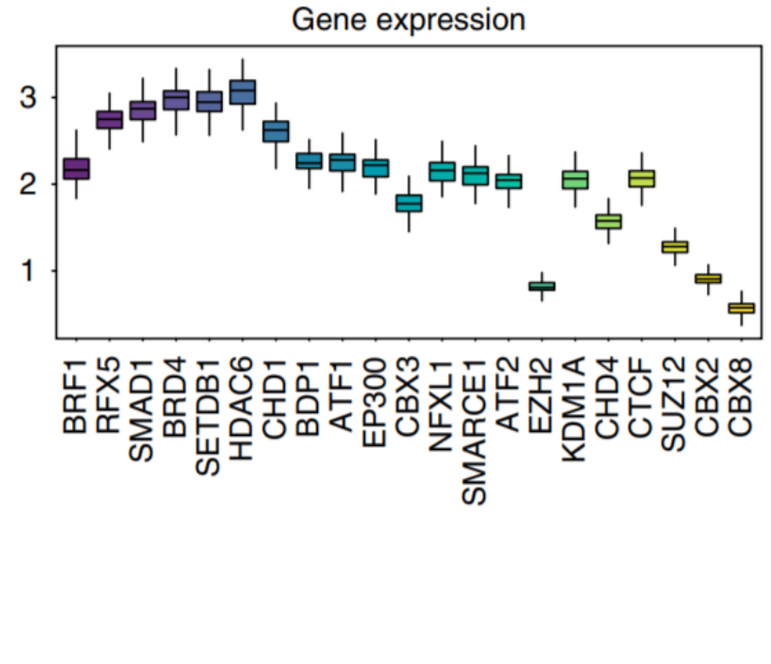




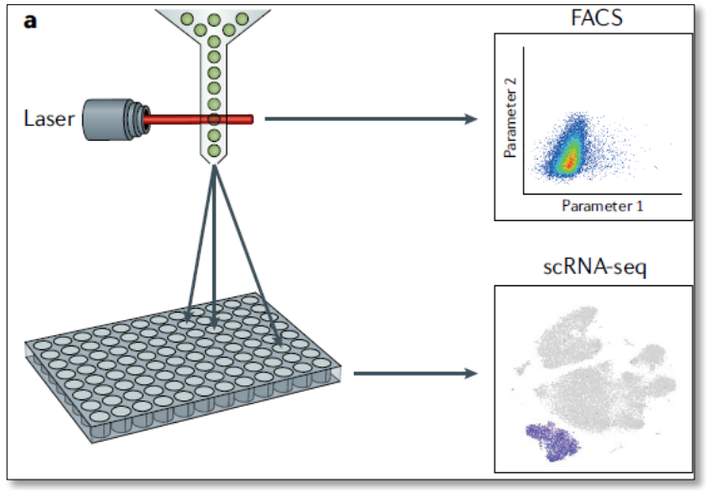


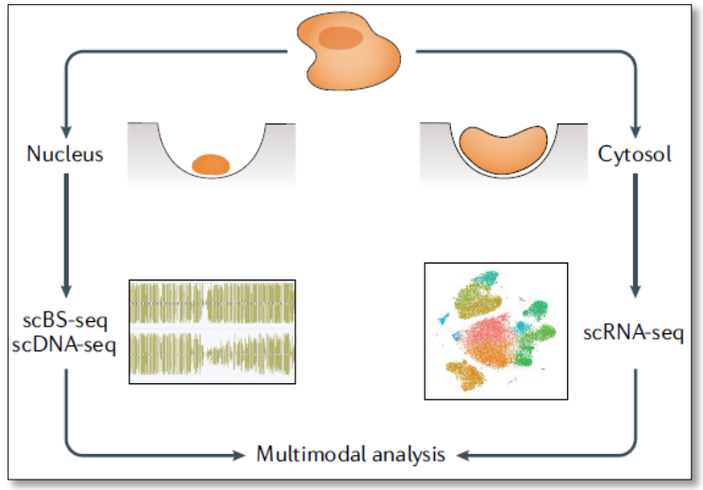


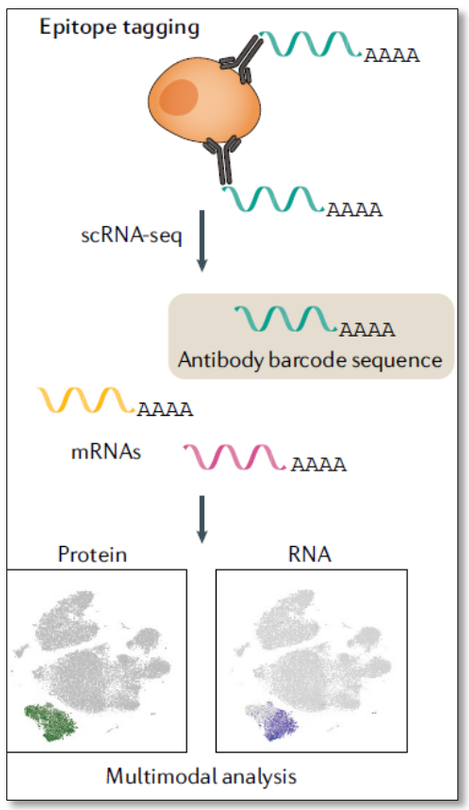


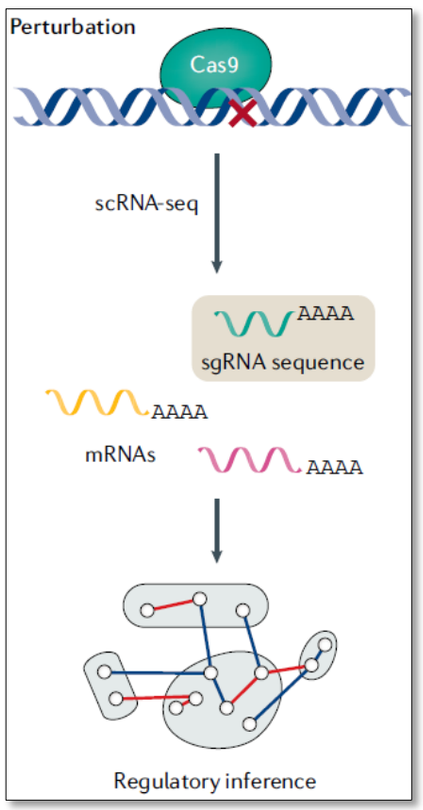


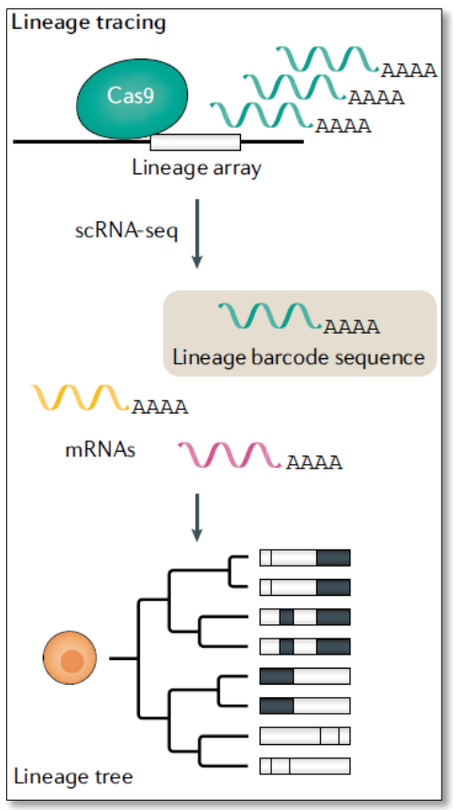
Diverse approaches for sc multi-omics

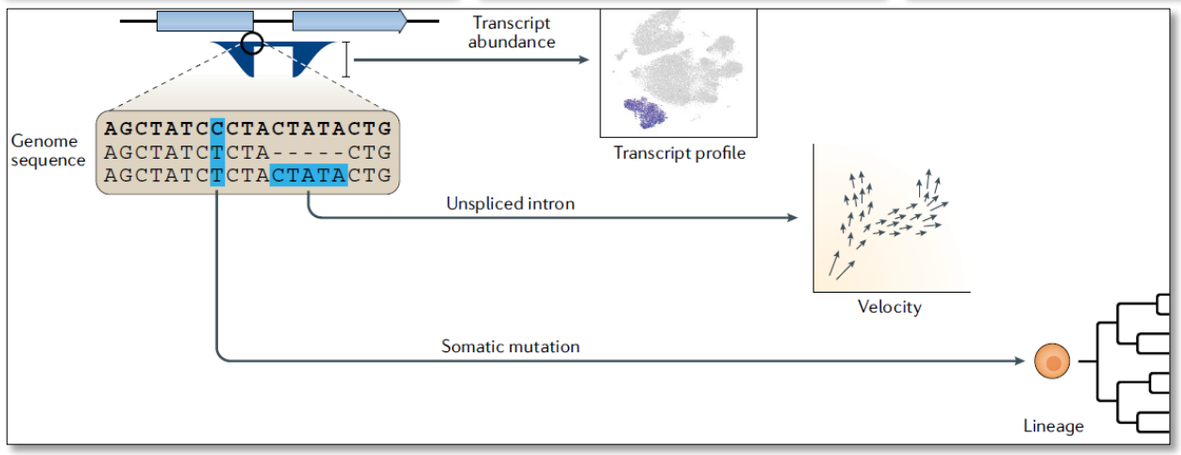












- a | Gathering cytometric single- cell measurements using multiparameter fluorescenceactivated cell sorting (FACS) before single- cell RNA sequencing (scRNA- seq) can allow fluorescence- based measurements of protein levels to be later linked to cellular transcriptomes; hence, RNA and protein levels 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 proteins20,21 (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.

SINGLE-CELL OMICS

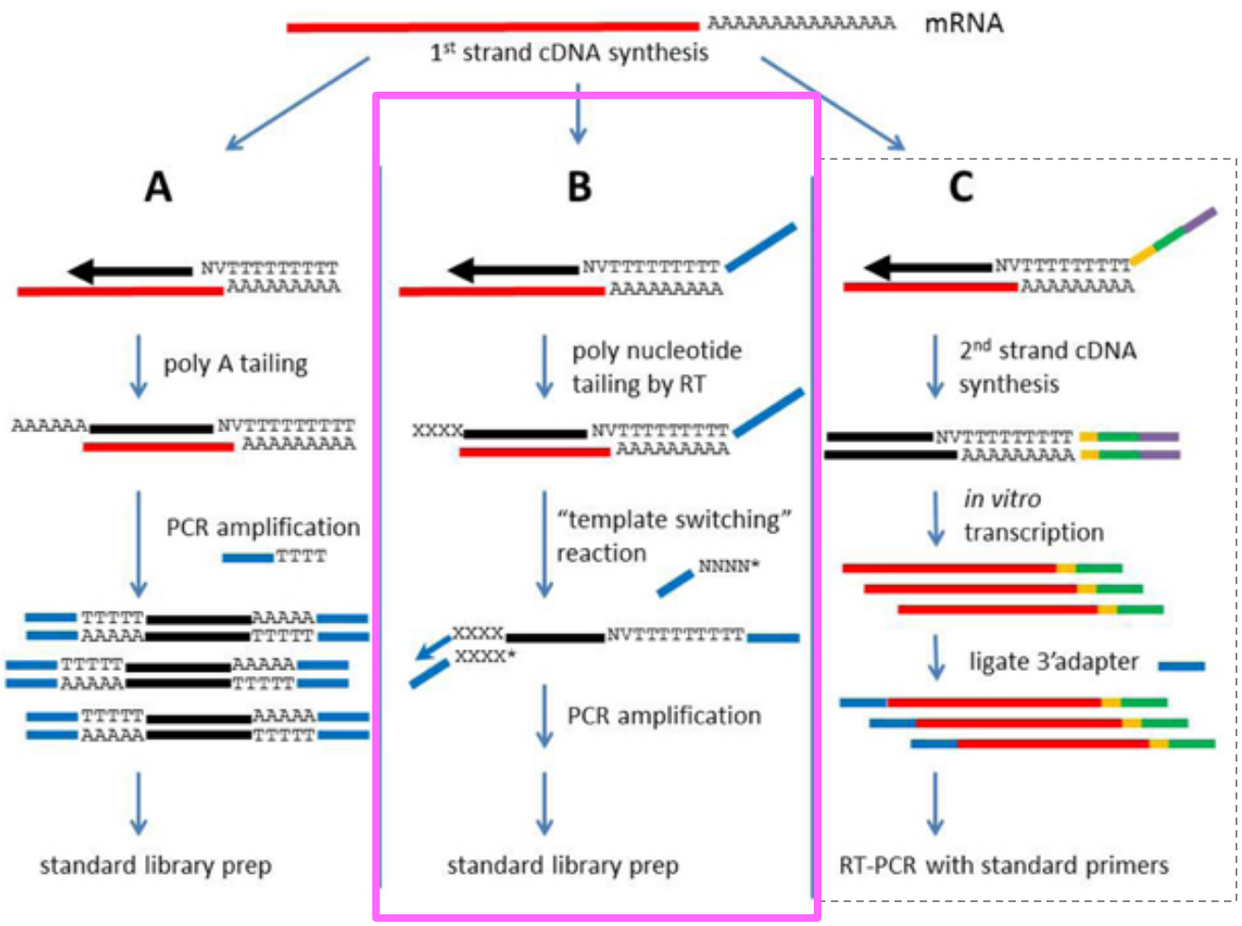
Integrative single-cell analysis

Tim Stuart[1] *and Rahul Satija*[1,2]*

4. Dealing with noise and doublets in single-cell data

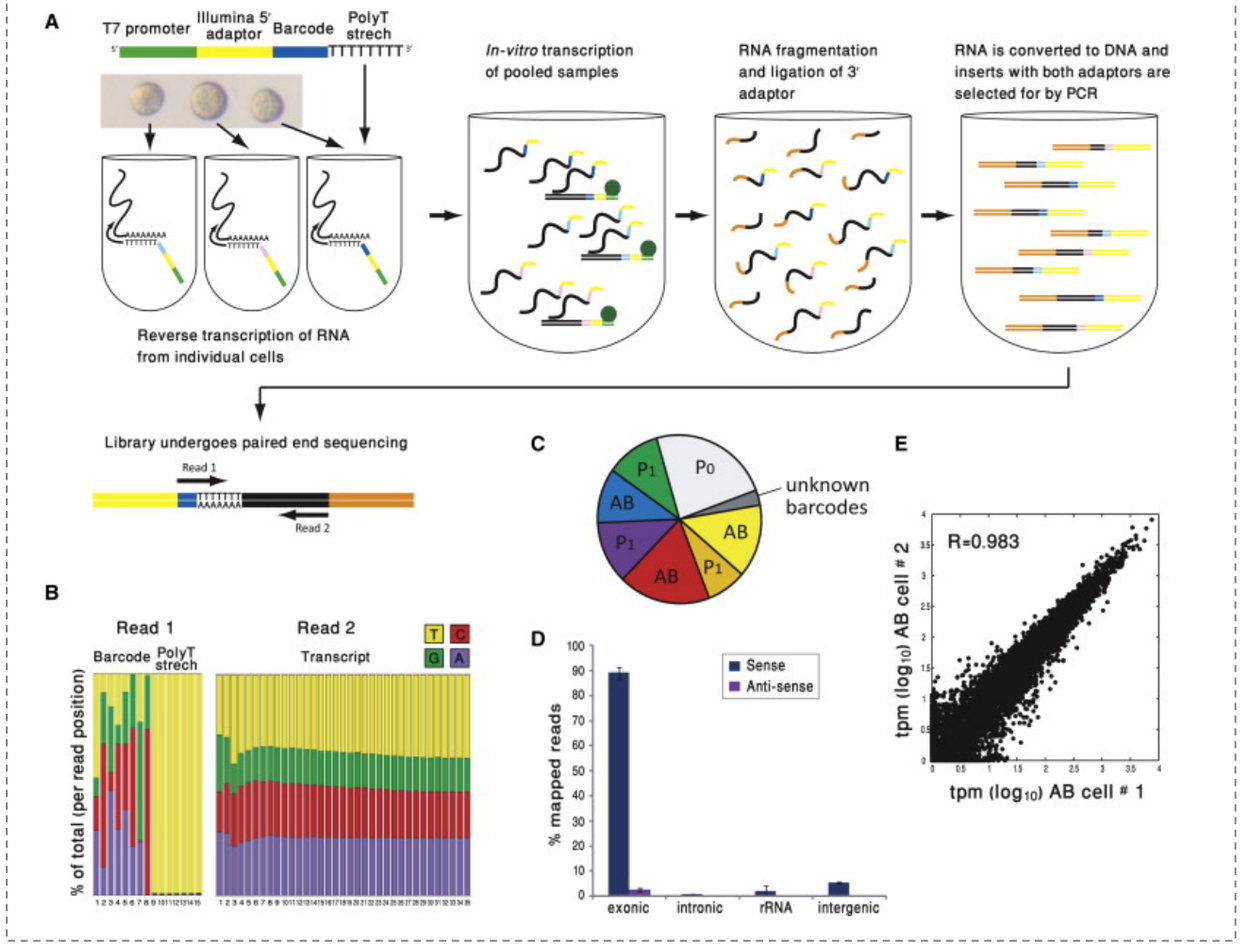
Dealing with rRNA contamination

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



From www.biotechniques.com

CEL-Seq



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

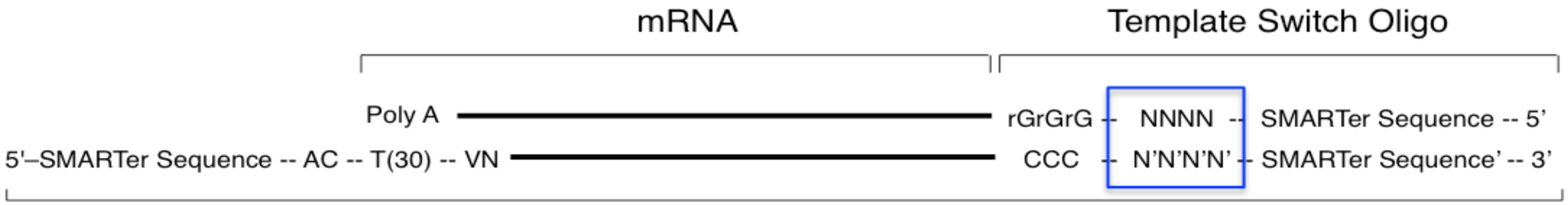
Ribosomal RNA contamination

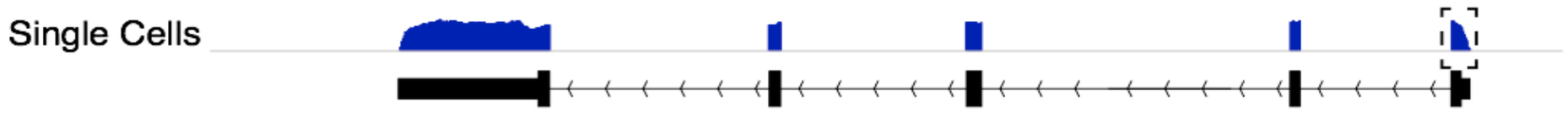
- rRNA overwhelms mRNA (~98%)
- polyA-selection is too inefficient

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. Beyond scRNA-seq: scATAC-seq, multi-omics
4. Dealing with noise, doublets, and other sc issues
5. Computational challenges in single-cell data analysis
6. Deep learning methods for single-cell data analysis
7. Guest lecture: Fabian Theis
8. Guest lecture: Romain Lopez

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



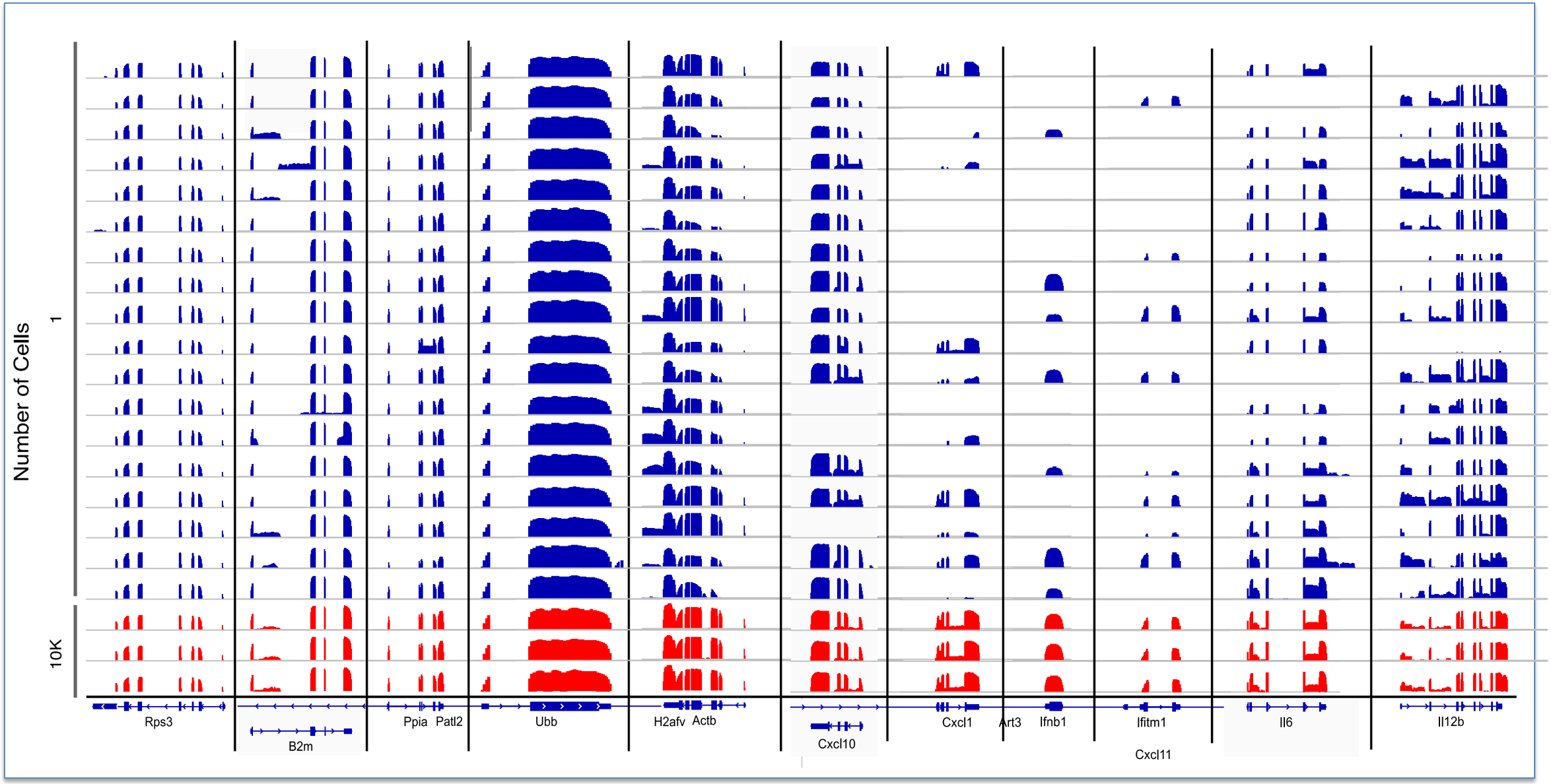


SMARTer Sequence	NNNN	GGG	Read
AAGCAGTGGTATCAACGCAGAGT	GCGG	GGG	GAAAAGGGGCTGTTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCC
AAGCAGTGGTATCAACGCAGAGT	AGAG	GGG	GAGGGCTGTTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCCTGTG
AAGCAGTGGTATCAACGCAGAGT	ACGA	GGG	GGGCTGTTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCGGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCCTGTGAA
AAGCAGTGGTATCAACGCAGAGT	CGGG	GGG	GAAAAGGGGCTGTTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCC
AAGCAGTGGTATCAACGCAGAGT	AGCG	GGG	AGGGCTGTTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCCTGTGA
AAGCAGTGGTATCAACGCAGAGT	CCGG	GGG	GCAAAGGGCTGTTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCCT
AAGCAGTGGTATCAACGCAGAGT	CCCG	GGG	GGCTGTTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCCTGTGAAG
AAGCAGTGGTATCAACGCAGAGT	AGCG	GGG	GAAAAGGGGCTGTTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCC
AAGCAGTGGTATCAACGCAGAGT	AGAG	GGG	AGGGCTGCTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCCTGTGA
AAGCAGTGGTATCAACGCAGAGT	AGCG	GGG	GAGGGCTGTTTCAGGGCTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCCTGTG
AAGCAGTGGTATCAACGCAGAGT	CCCG	GGG	AGGGCTGTTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCCTGTGA
AAGCAGTGGTATCAACGCAGAGT	GTAT	GGG	GAAAAGGGGCTGTTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCC
AAGCAGTGGTATCAACGCAGAGT	AAGG	GGG	CAACGCAGAGTAATCGGGGAGGGCTGTTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTG
AAGCAGTGGTATCAACGCAGAGT	AAGG	GGG	GAGGGCTGTTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCCTGTG
AAGCAGTGGTATCAACGCAGAGT	GTCT	GGG	GAGGGCTGTTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCCTG
AAGCAGTGGTATCAACGCAGAGT	GTAT	GGG	GGGAGGGCTGTTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCCTG
AAGCAGTGGTATCAACGCAGAGT	AGCG	GGG	CAACGCAGAGTAATCGGGGAGGGCTGTTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTC
AAGCAGTGGTATCAACGCAGAGT	CCCG	GGG	GAGGGCTGTTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCCTGTG
AAGCAGTGGTATCAACGCAGAGT	AAGG	GGG	GAAAAGGGGCTGTTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCC
AAGCAGTGGTATCAACGCAGAGT	AGCG	GGG	GAGGGCTGTTTCAGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCCTGTG
AAGCAGTGGTATCAACGCAGAGT	AGCG	GGG	AGGGCTGTTTTTTGGGTTTGGGTTAGTGAGCCTCATCCTGGCAGTTATTTTATAGTAAAGAACATTCAAGTGCTCTGCCTACCTAGGGCCCTGTGA

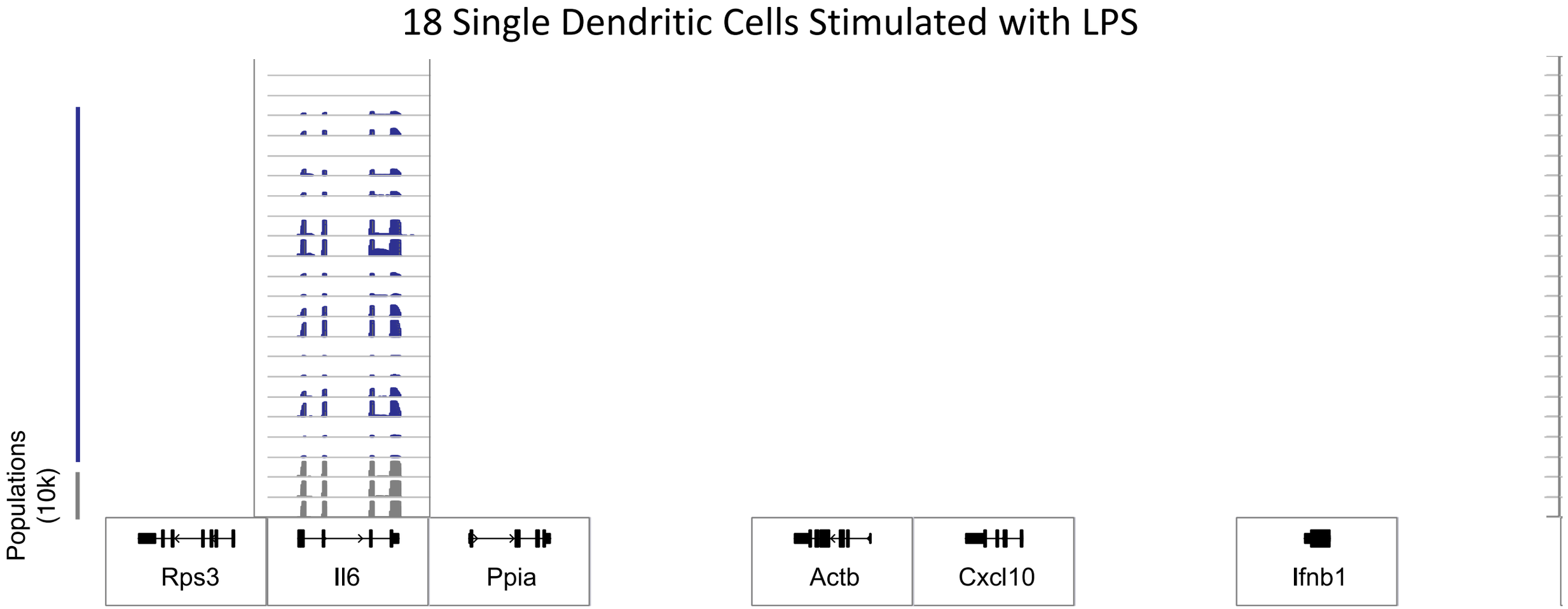
Ultra-low input RNA-seq is problematic

- 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

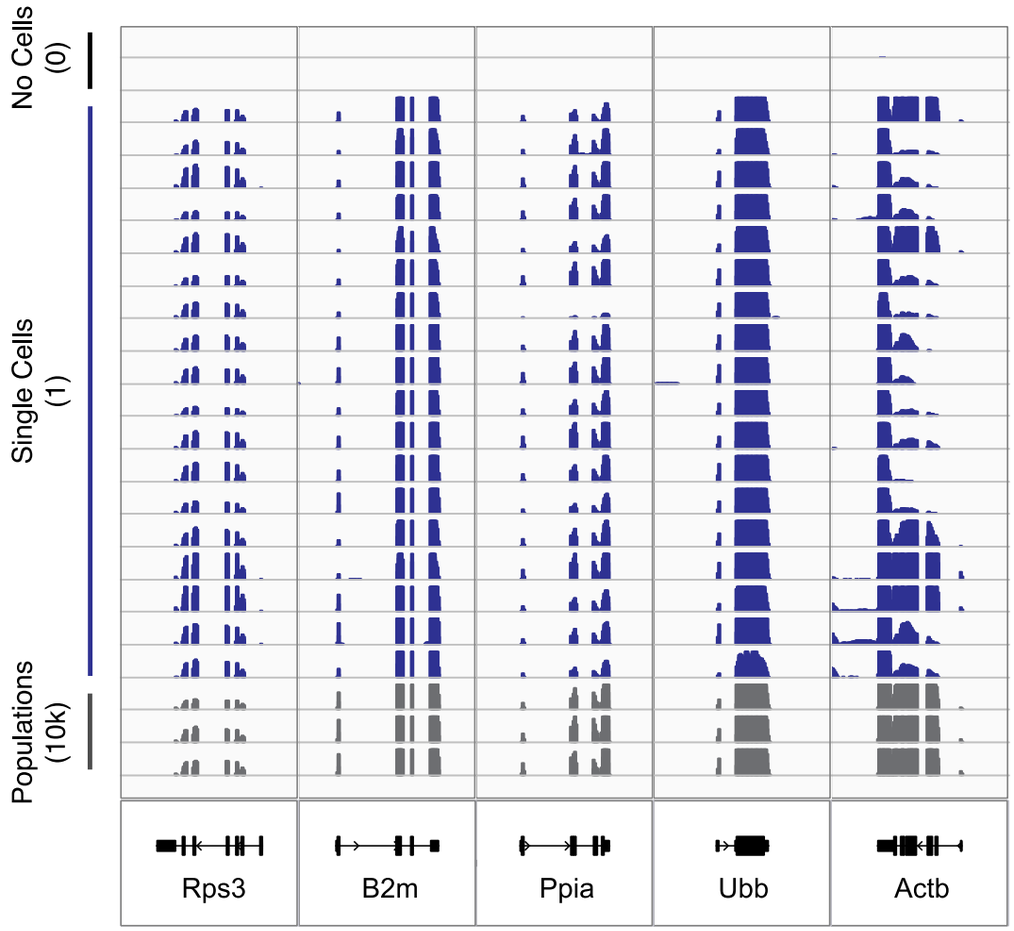


We can profile the full transcriptomes of cells with SMART-Seq.

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

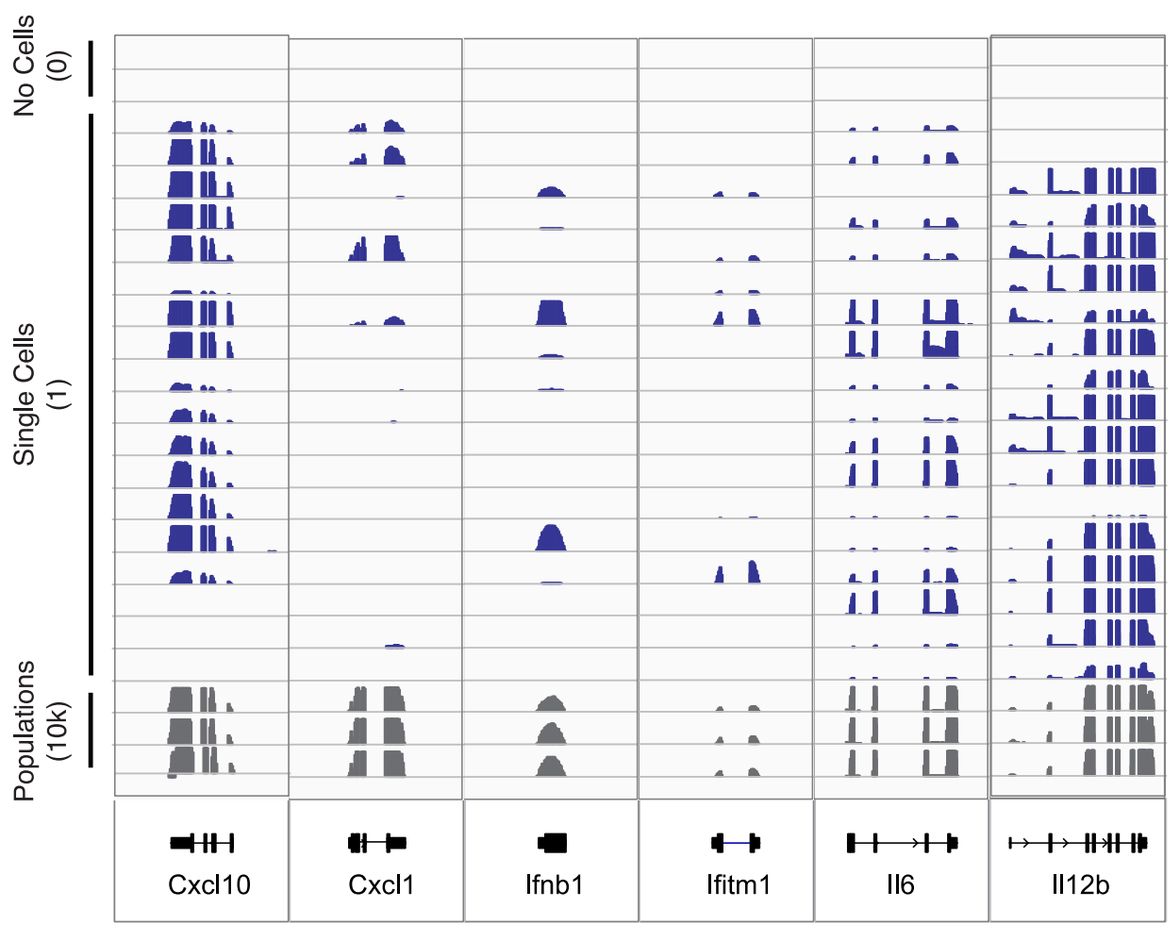
Housekeeping vs. variable genes

Least Variable Genes



Term	P-Value
Structural constituent of ribosome	1.66e-51
Structural molecule activity	1.16e-33

Most Variable Genes



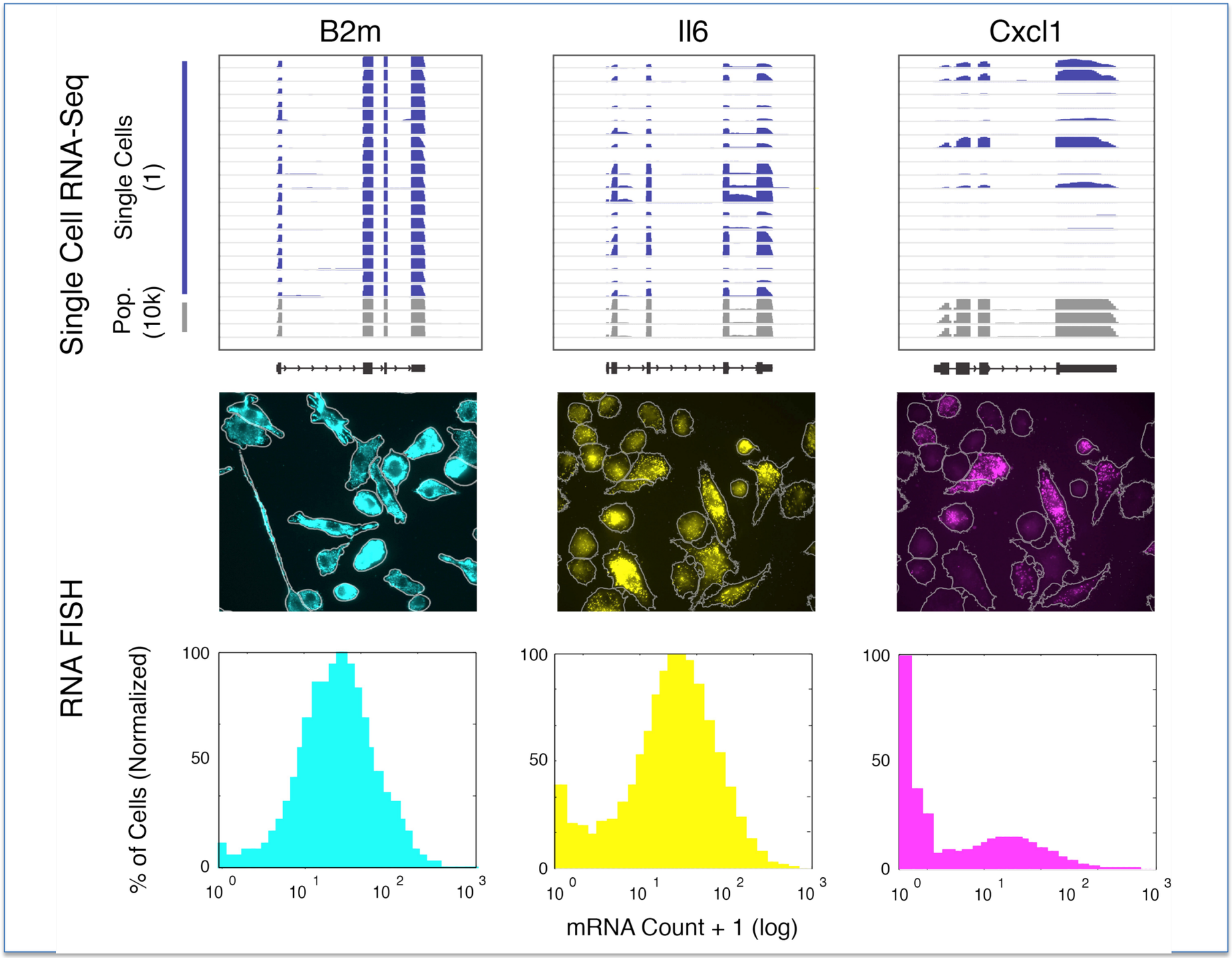
Term	P-Value
Defense response	2.23e-29
Inflammatory response	7.99e-25

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)

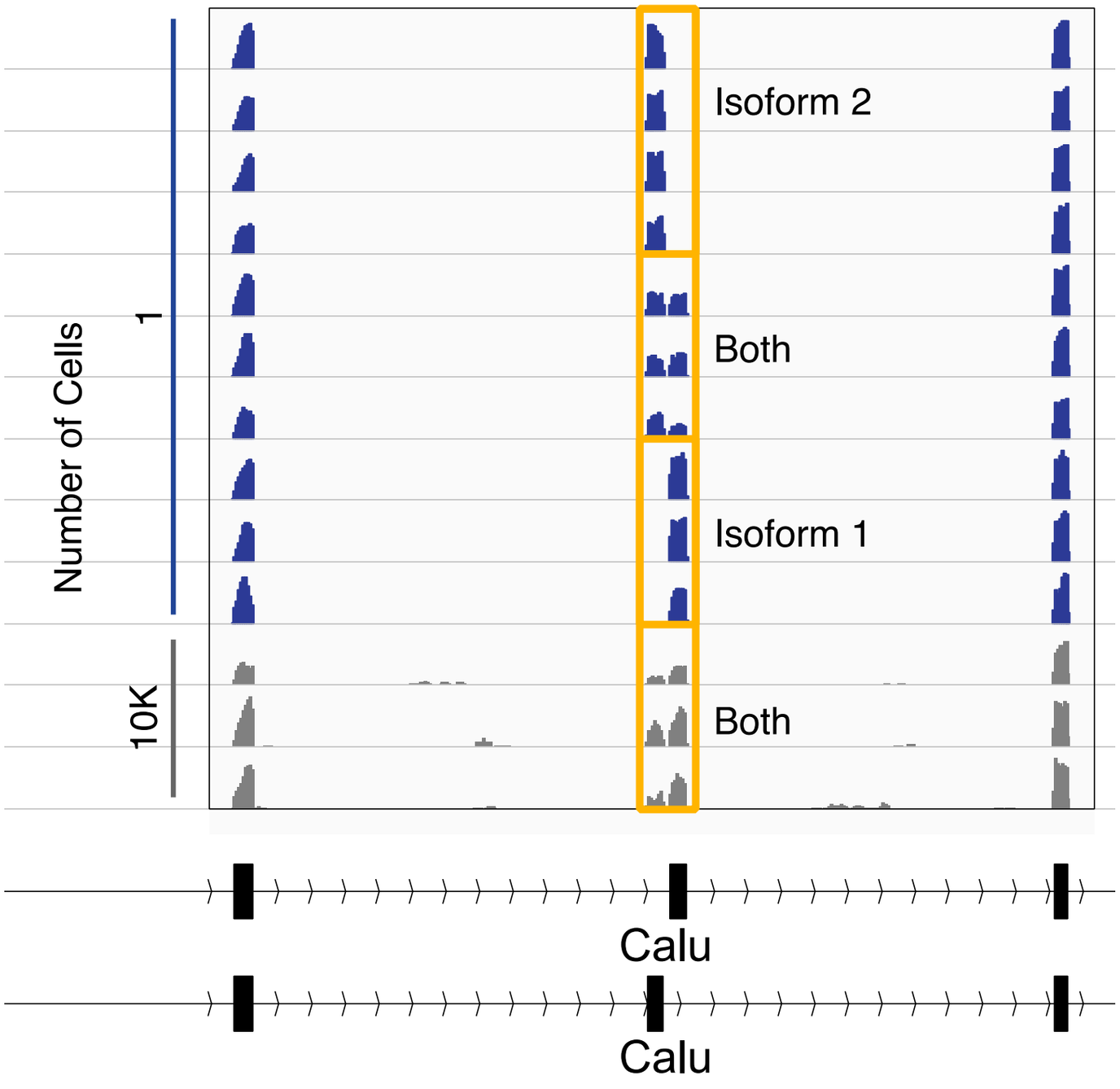
RNA FISH Validates Single-Cell RNA-Seq



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

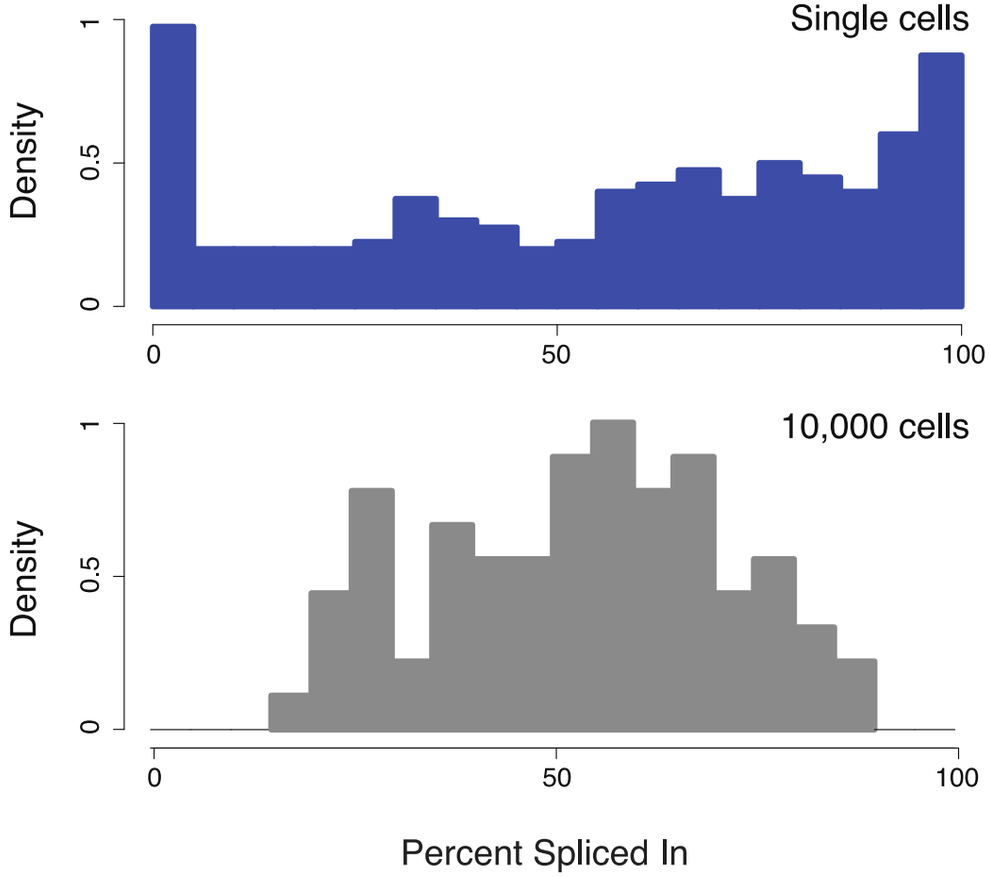
Shalek*, Satija* et al, *Nature*, **498**, (2013)

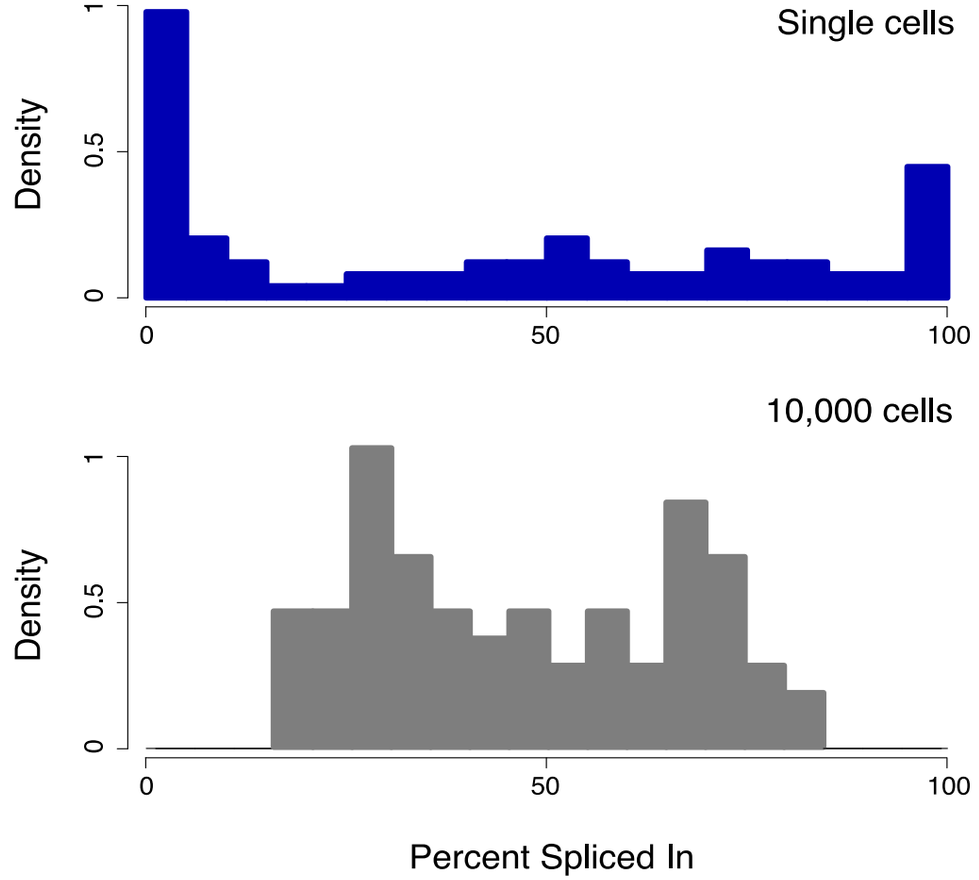
Two sources of noise in single cell data



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

Two sources of noise in single cell data

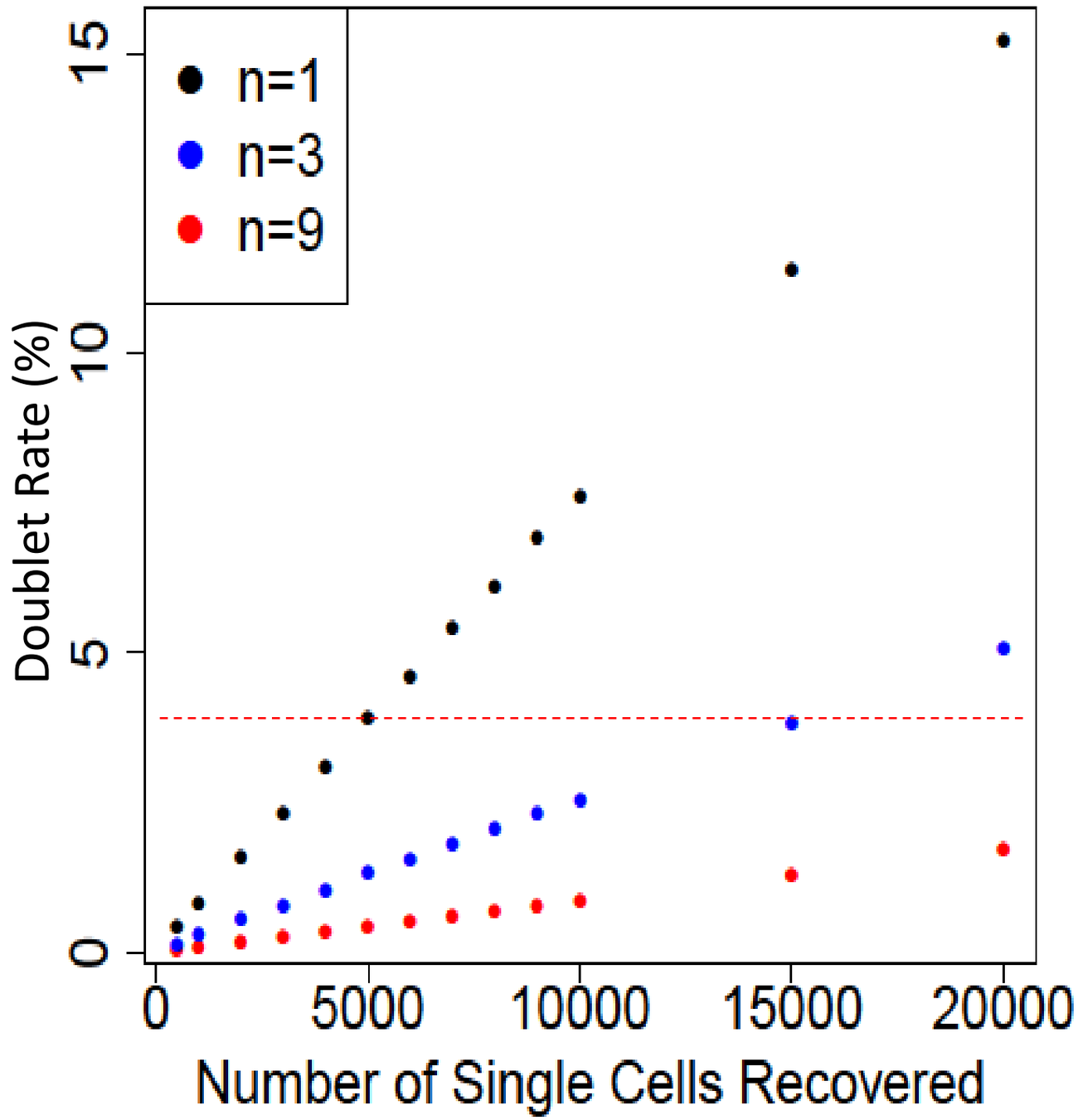




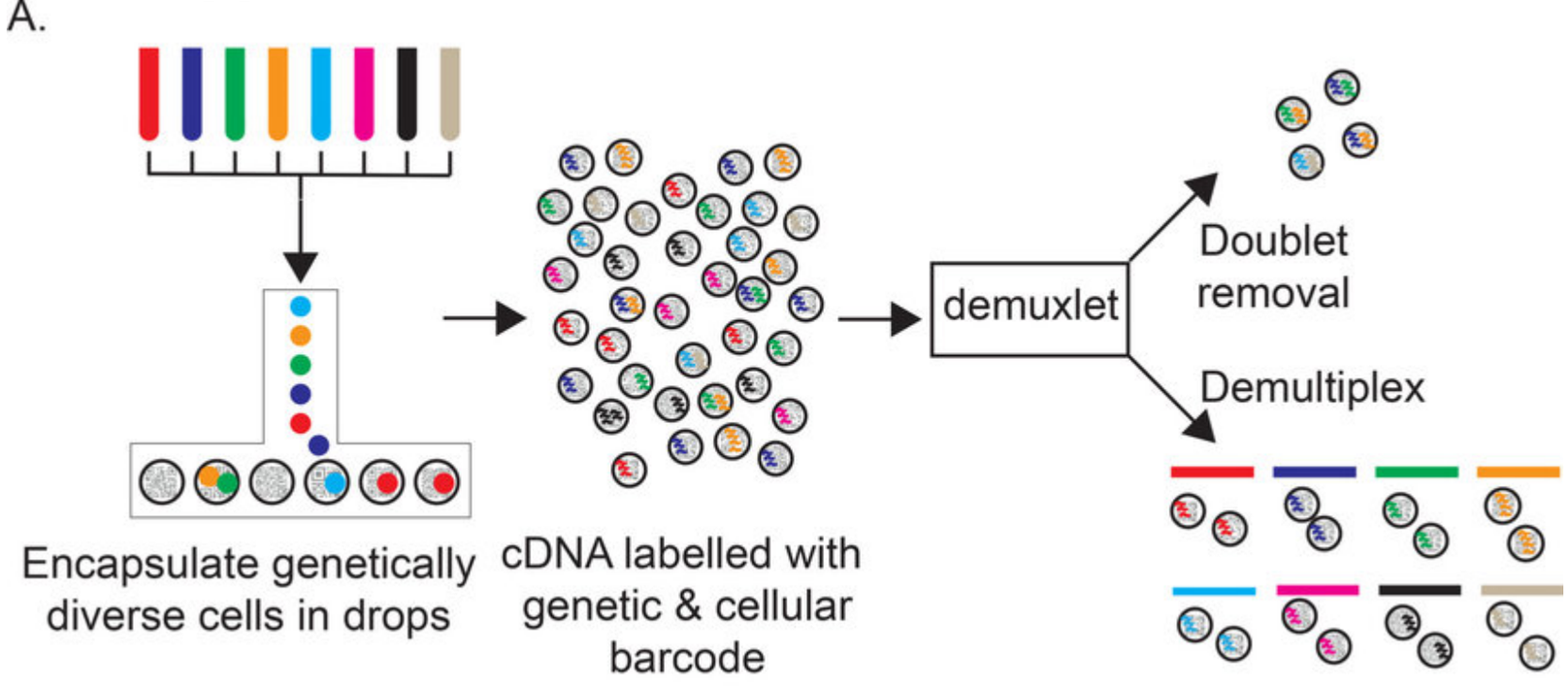
One way to overcome this – focus on highly expressed genes

Doublets increase with numbers of cells profiled

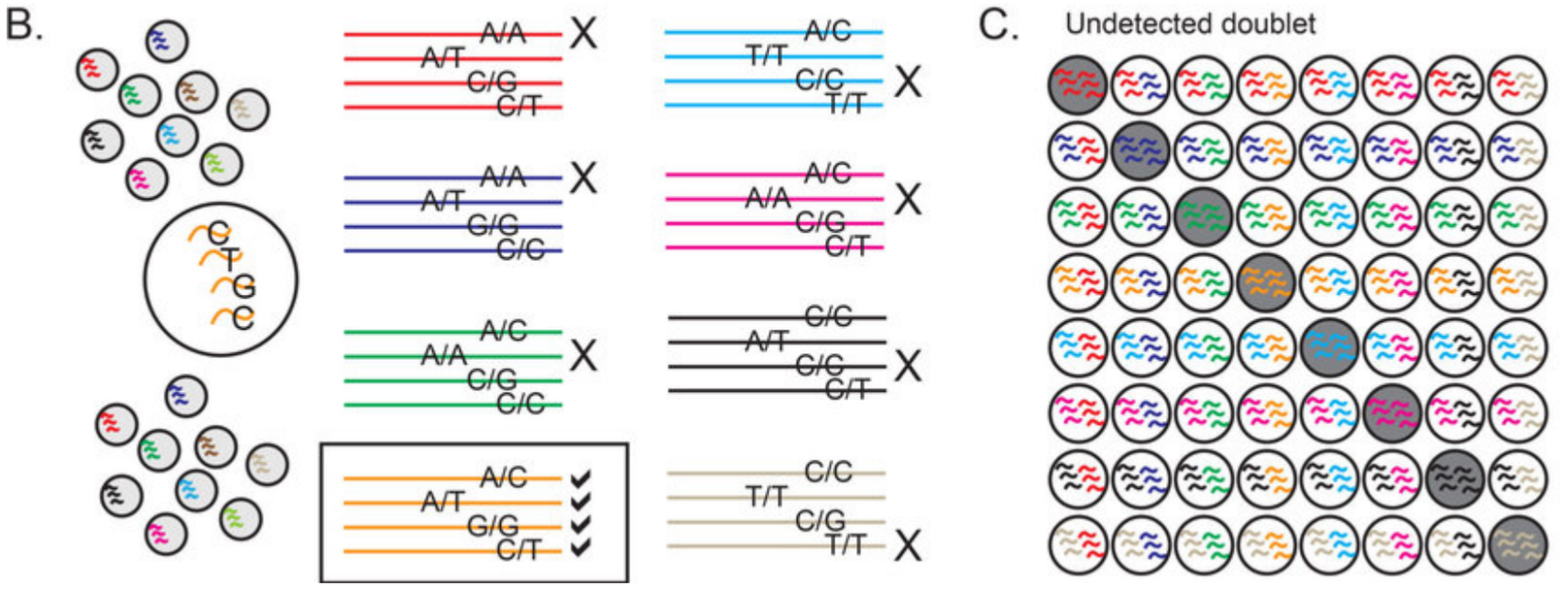
- Prohibitive cost (~$2,000 per sample)
- Prominent batch effects
- Limitations to the number of cells that can be captured
  - As more cells are captured more doublets appear in the data



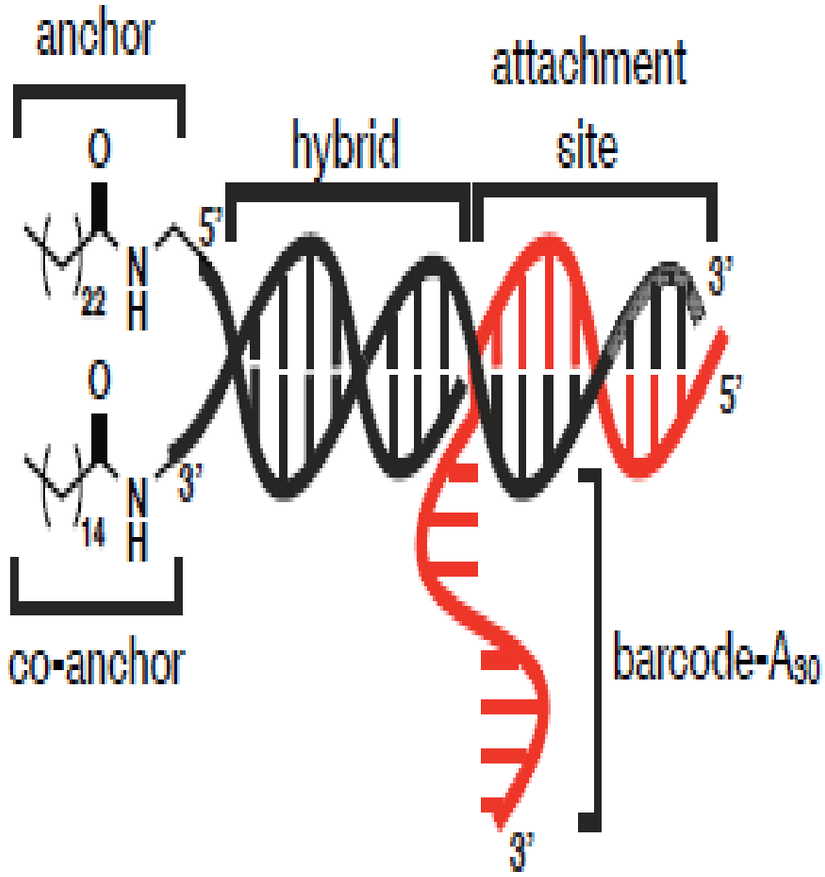
Recognizing doublets with cellular barcodes



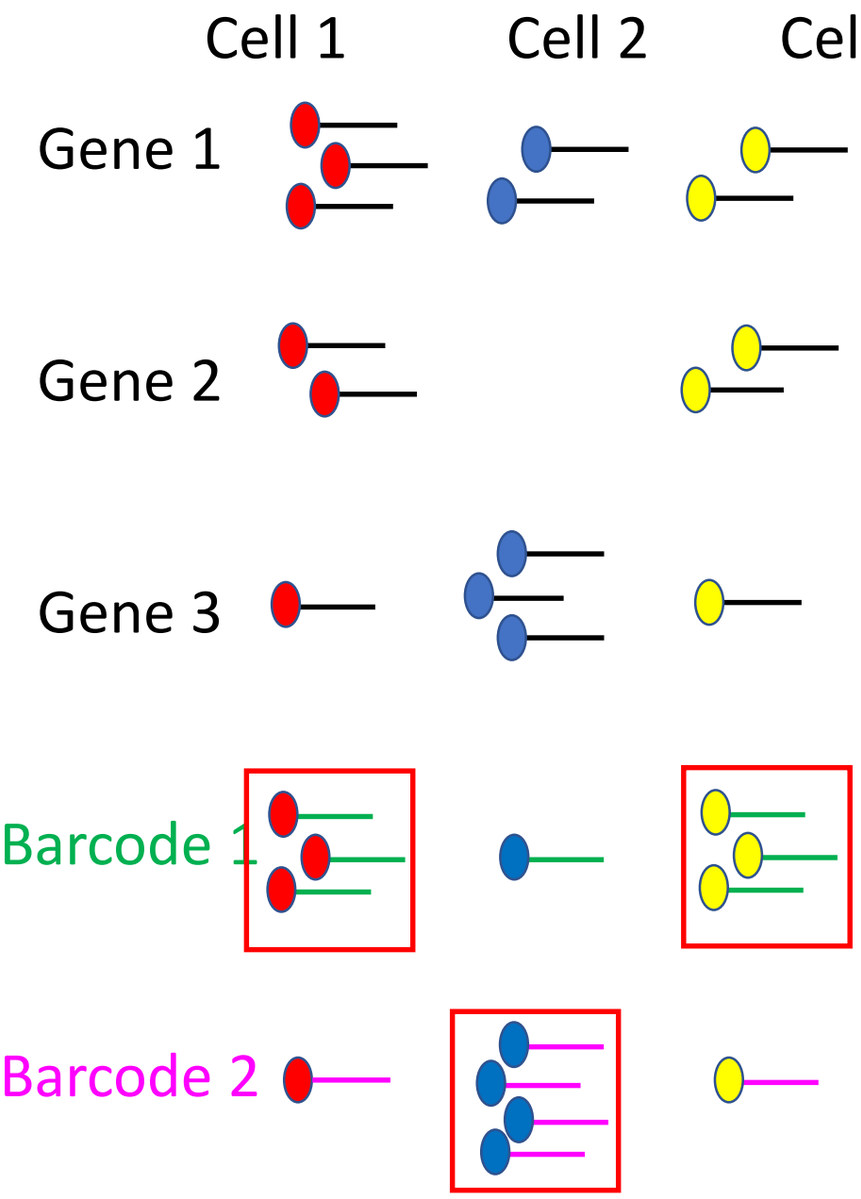
Recognizing doublets with genetic ‘barcodes’



Multiplexing Using Lipid-Tagged Indices
MULTI-Seq



McGinnis et al. Nature Methods 2019



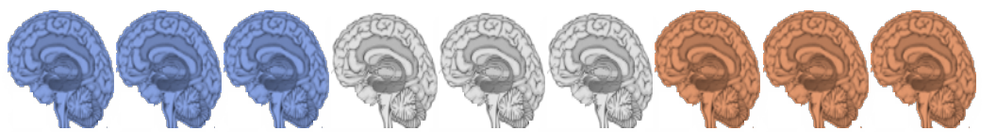
24 Sz + 24 BD + 24 Ctrl = Multiplexing (batches of 9)

24 SZ

24 BD

24 CON

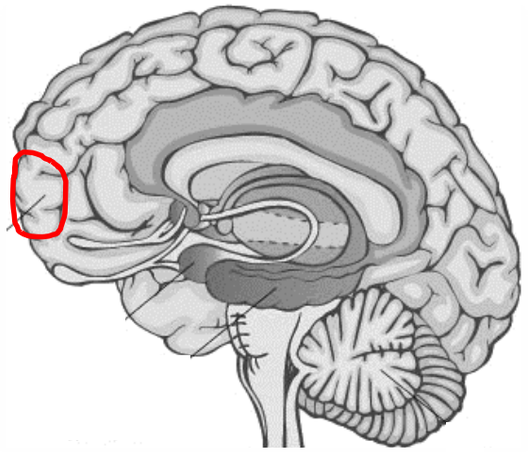
8 batches, 9 individuals each, mix phenotypes



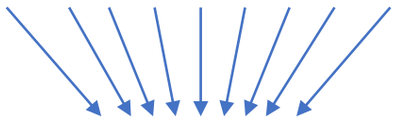
Schizo **Control** **Bipolar**

Isolate nuclei and label each sample with a unique oligo hashtag

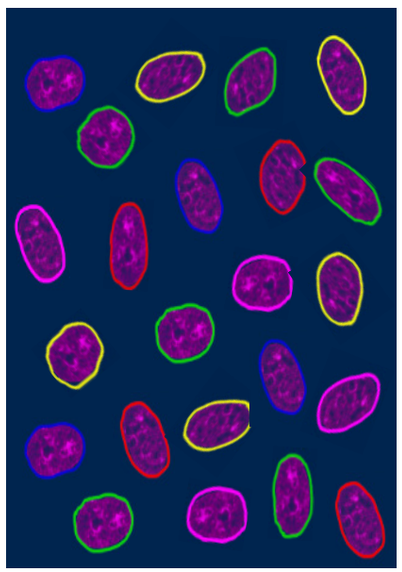
Pooled nuclear suspension spread across eight 10x channels



Frontopolar Cortex – BA10
72 individuals total

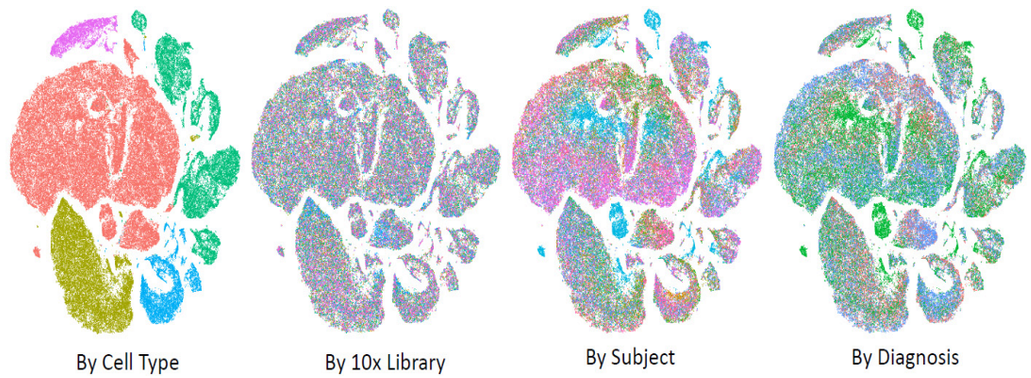


Pool all material into a single nuclear suspension



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

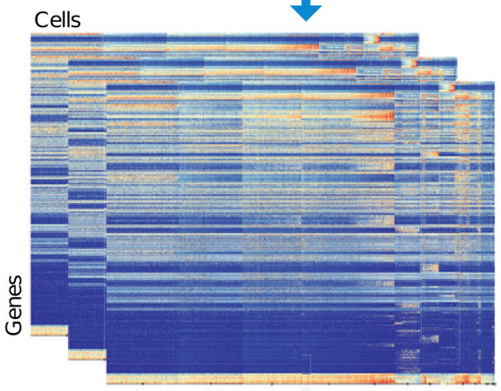
Brad Ruzicka

5. Computational challenges in single-cell data analysis

PRE-PROCESSING

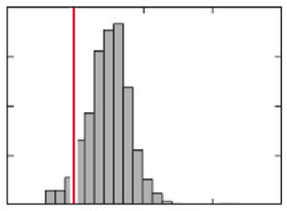
DOWNSTREAM ANALYSIS

Raw data processing

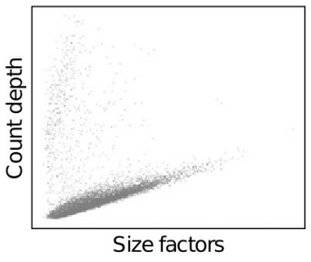


Count matrices

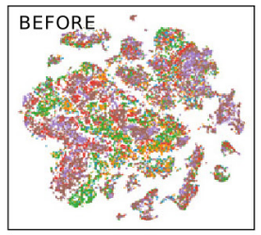
Quality control



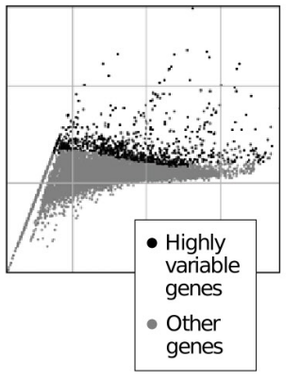
Normalization



Data correction (e.g. batch)

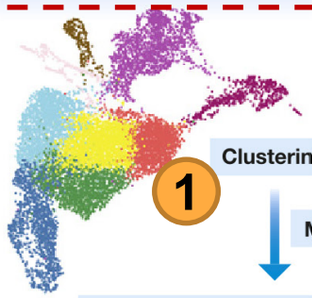


Feature selection



Visualization

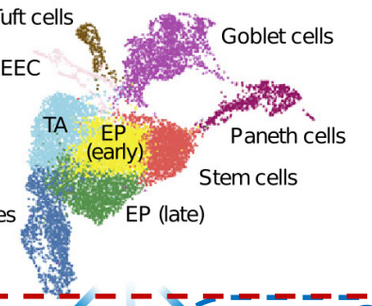
Cell-level analyses



Clustering

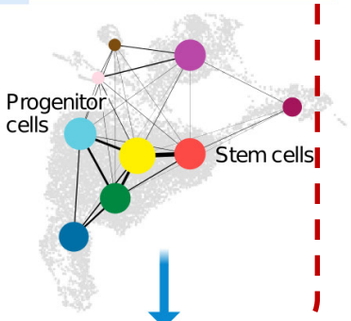
Marker identification

Cluster annotation

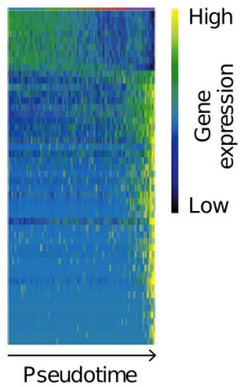


Multi-resolution analysis

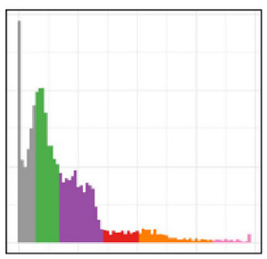
Trajectory inference



Gene dynamics

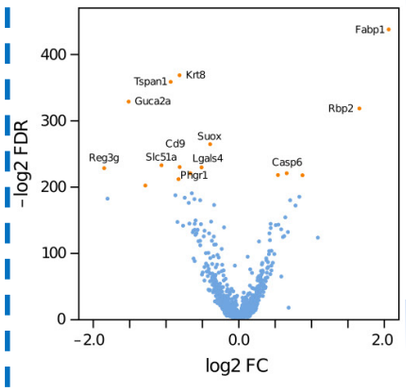


Metastable states

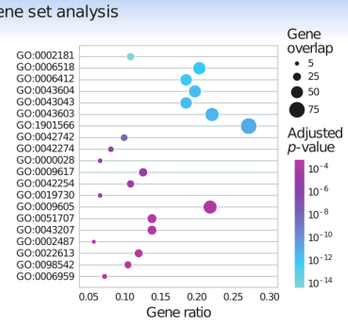


Gene-level analyses

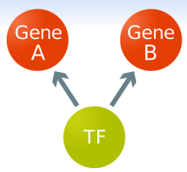
Differential expression



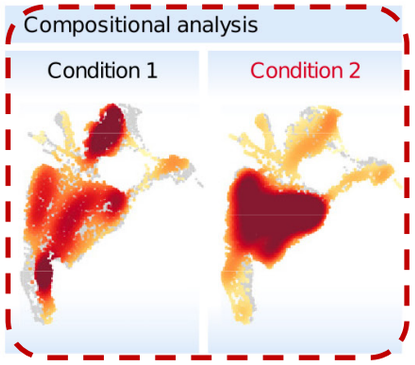
Gene set analysis



Gene regulatory networks



Compositional analysis



Multi-Omics analyses

Clustering similar genes

Dimensionality Reduction

Missing data imputation

Deconvolution of bulk RNA-seq

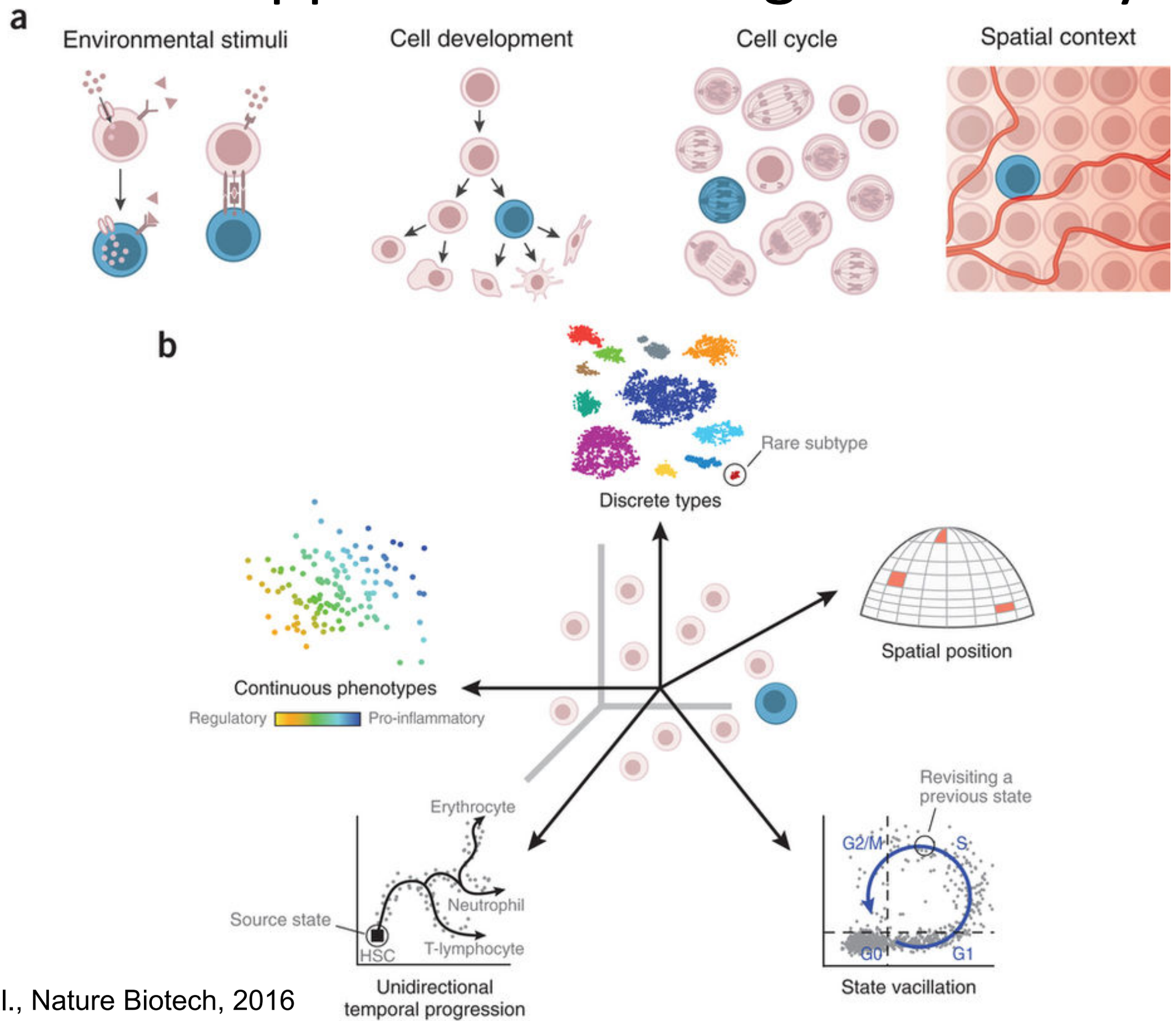
Luecken, M. D., & Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, *15*(6).

Extracting biological insights from scRNA-seq data

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

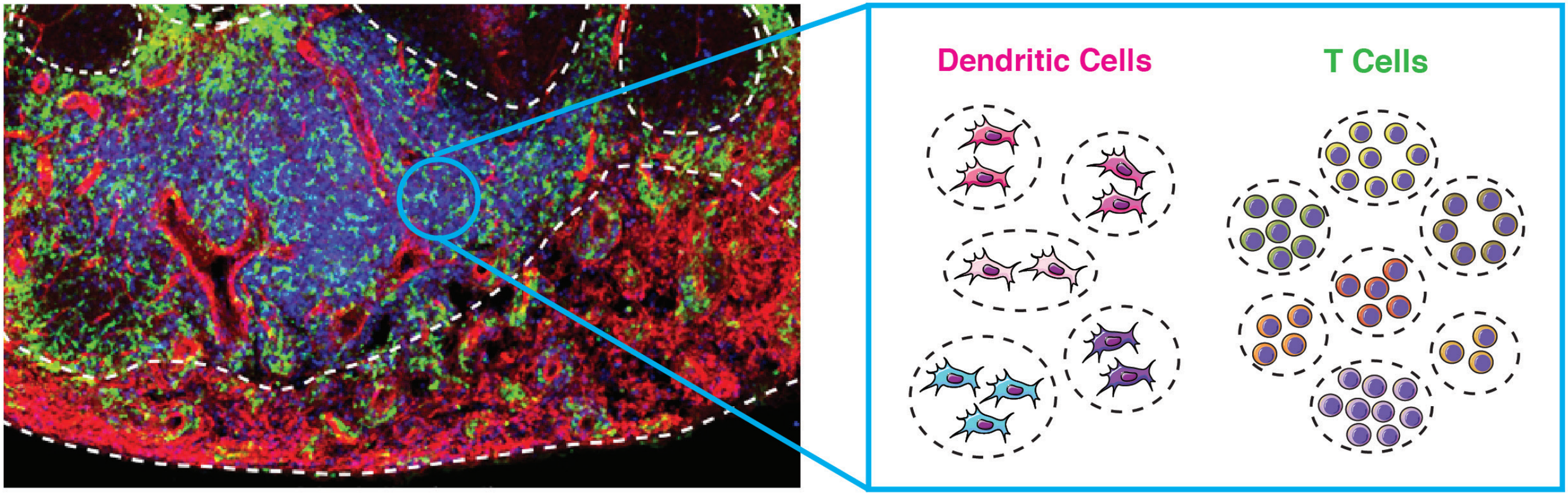
Clustering similar cells

Methods + applications of single-cell analysis



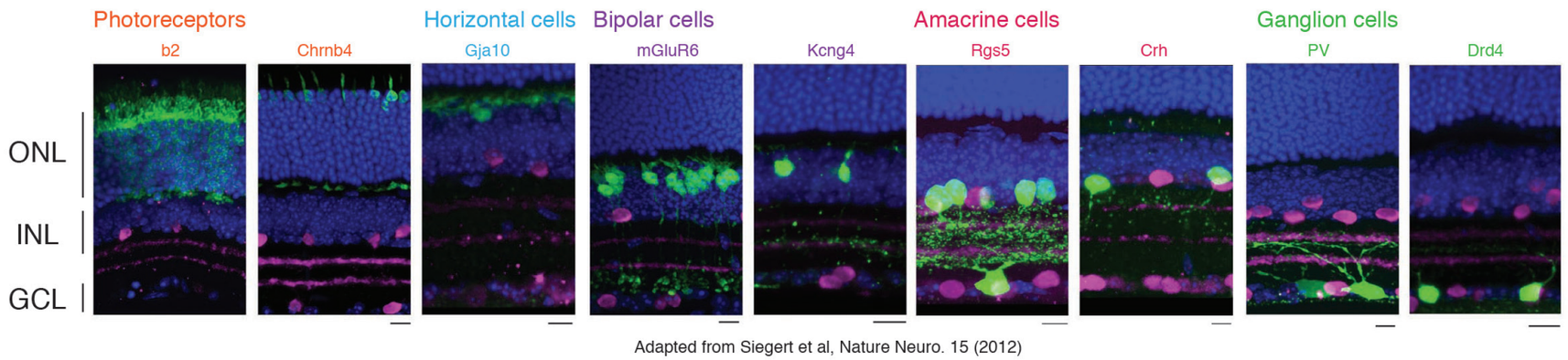
Wagner et al., Nature Biotech, 2016

From Complex Tissues to individual cell types

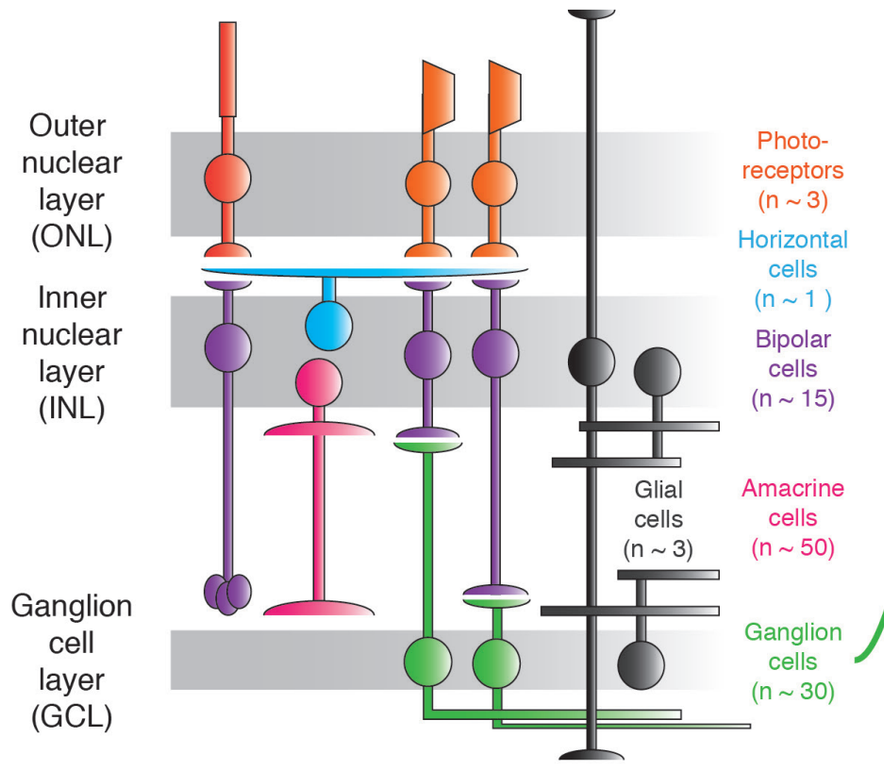


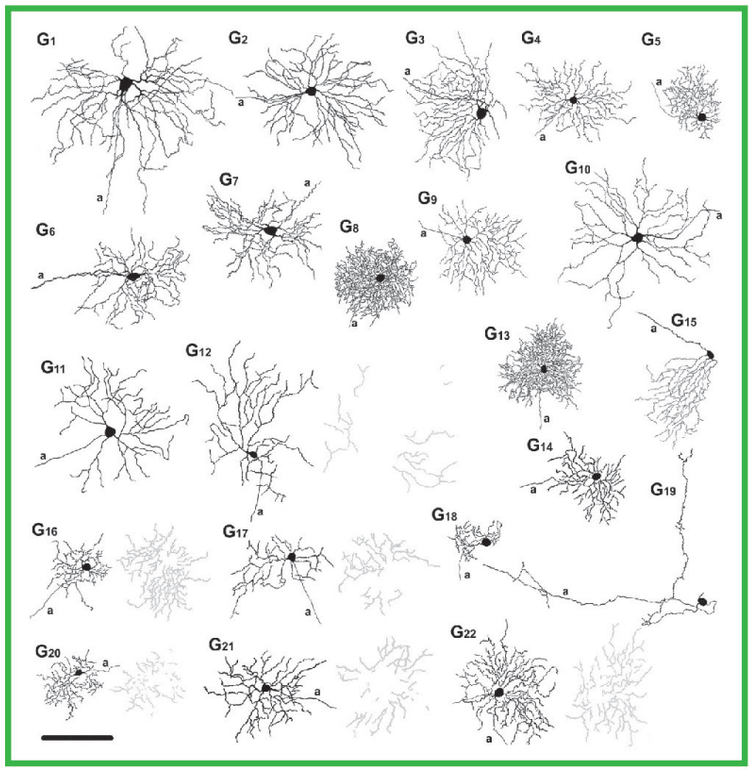
Can we identify the different cell types/states in a complex tissue?

Brain Case Study: The Mouse Retina



Adapted from Siegert et al, Nature Neuro. 15 (2012)





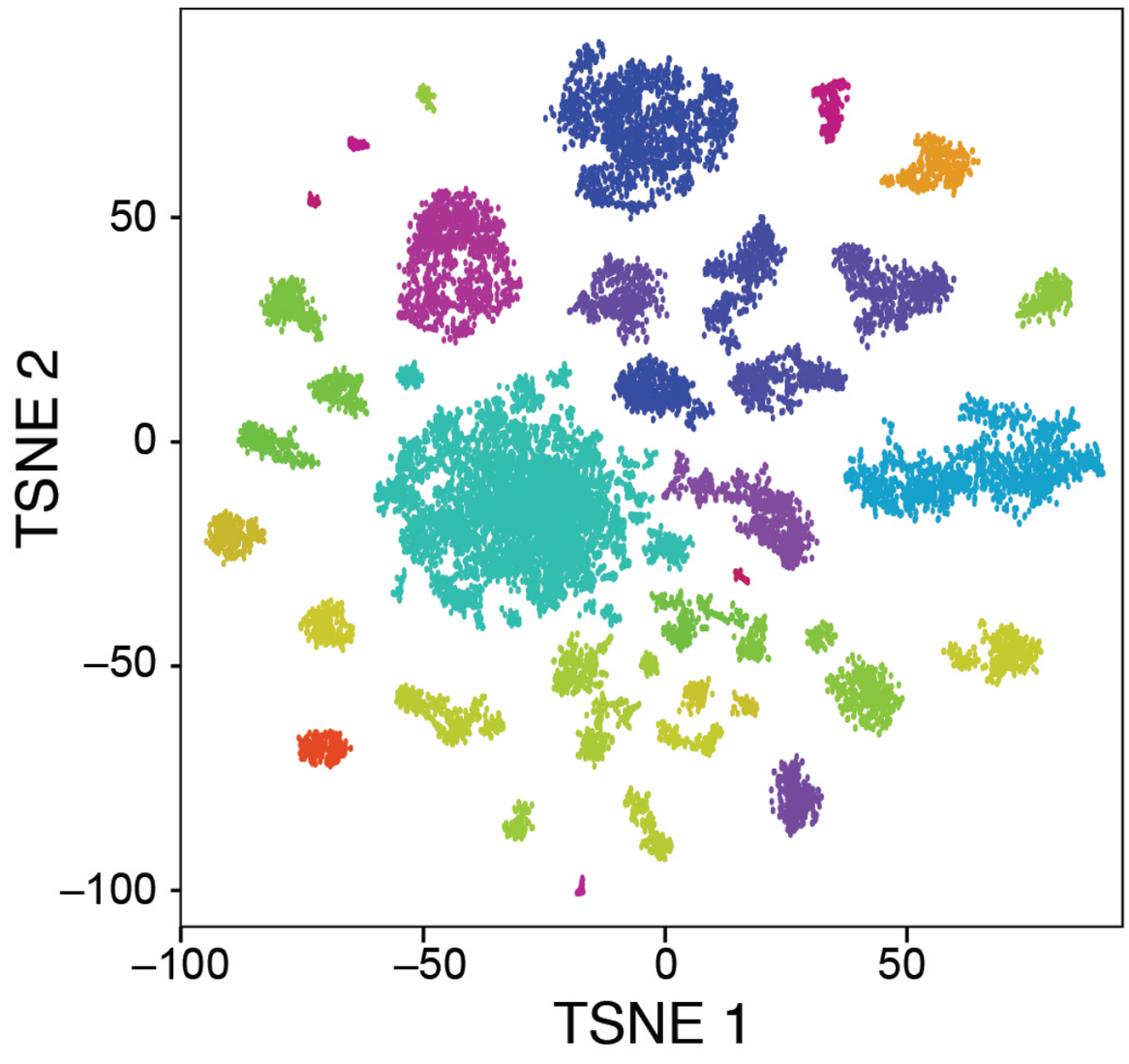
Adapted from Volgyi et al, *J. Comparative Neurology* **512** (2009)

~100 cell subtypes, only some with molecular markers

Macosko et al, *Cell*, **161**, 2015

49,300 Retina Cells Grouped Into 39 Clusters

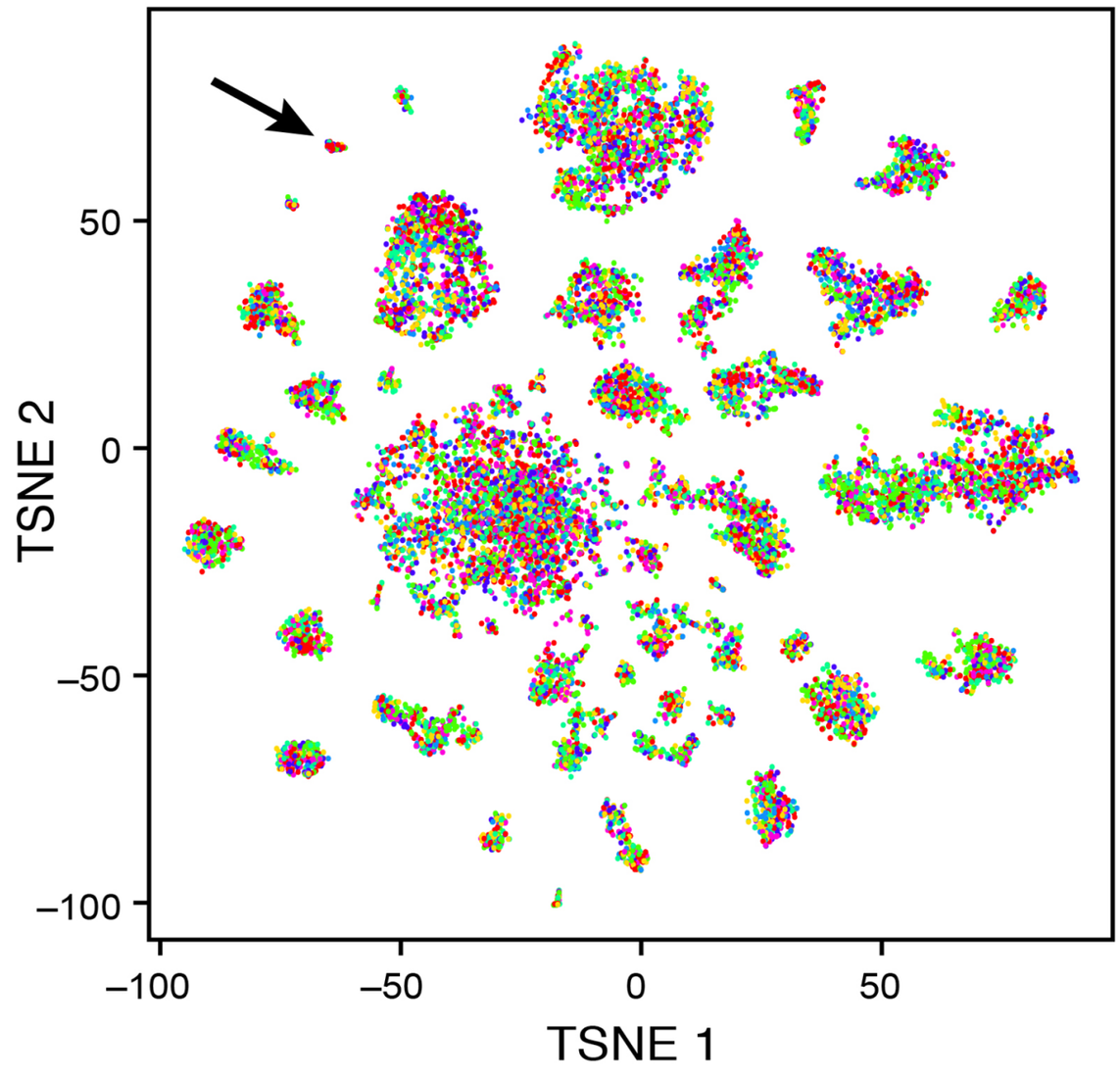
- **Drop-Seq:** 49,300 cells from dissociated mouse retina (P14) (~15k reads per cell)
- **Computational pipeline**: Select 25% best coverage cells, Dimensionality reduction (PCA+tSNE), Project remaining cells, Identify cell types (density clustering), Refine clusters (differential expression)



Macosko et al, *Cell*, **161**, 2015

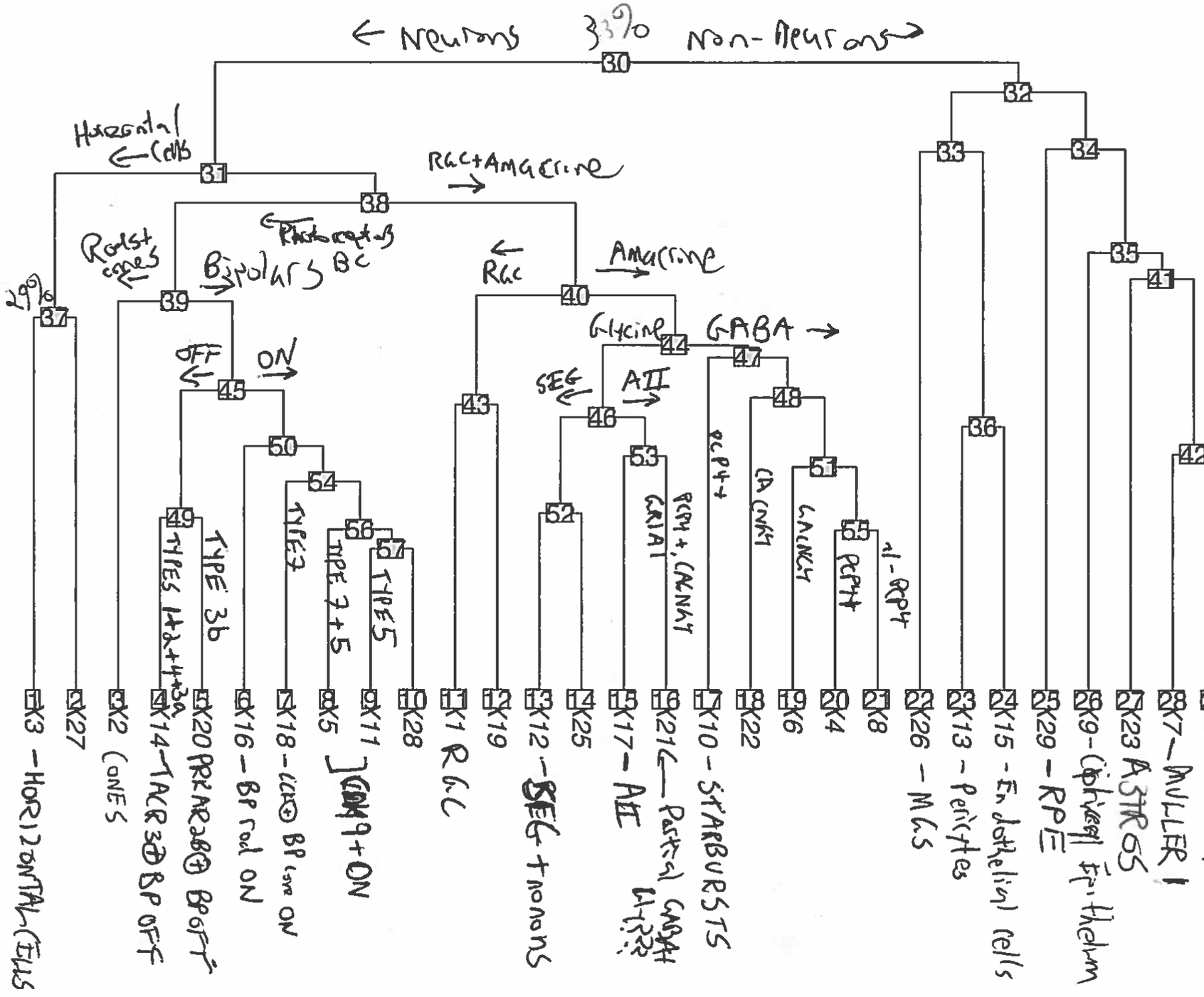
49,300 Retina Cells Grouped Into 39 Clusters

- **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)



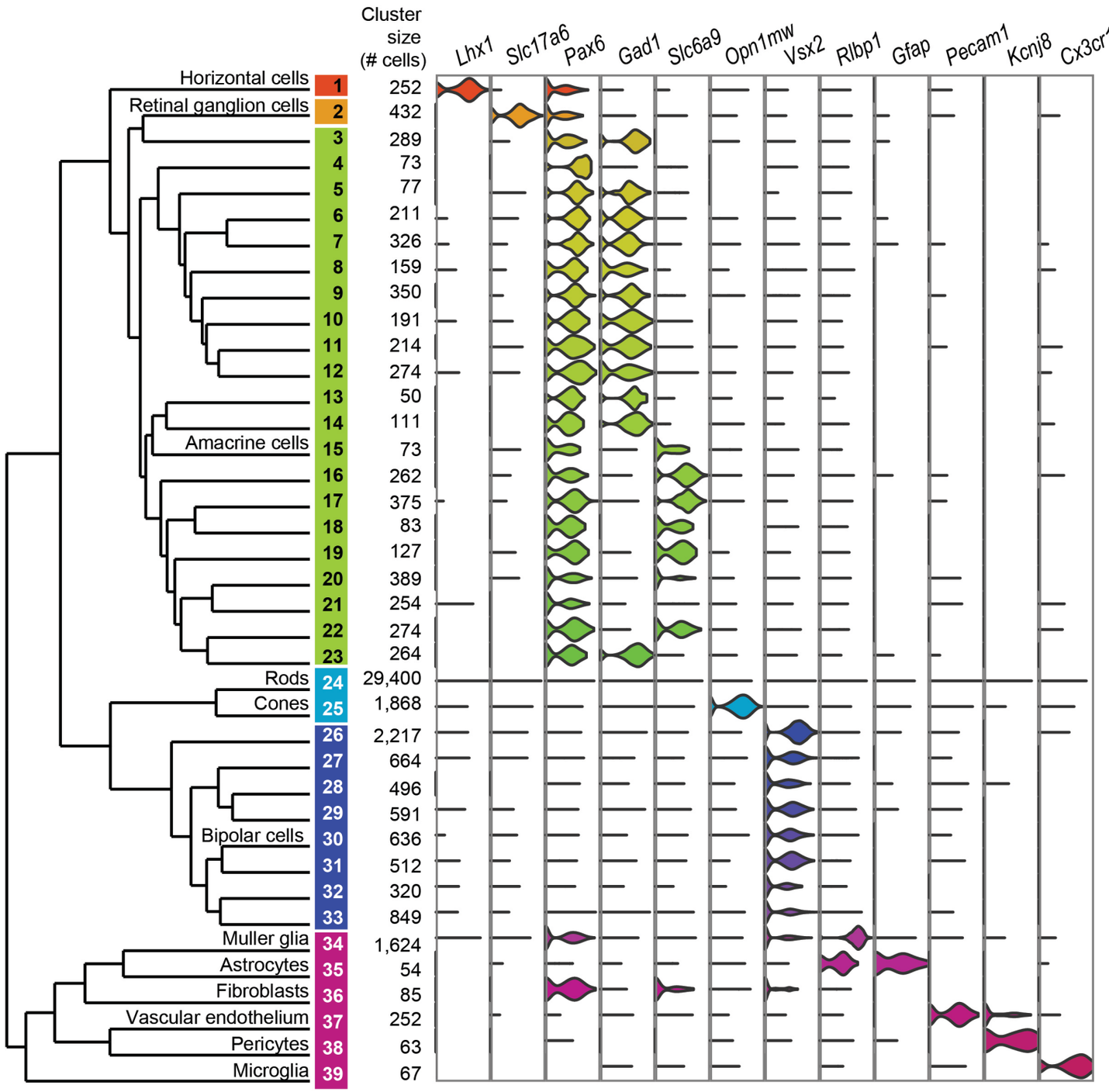
Macosko et al, *Cell*, **161**, 2015

Annotating A Cell Atlas



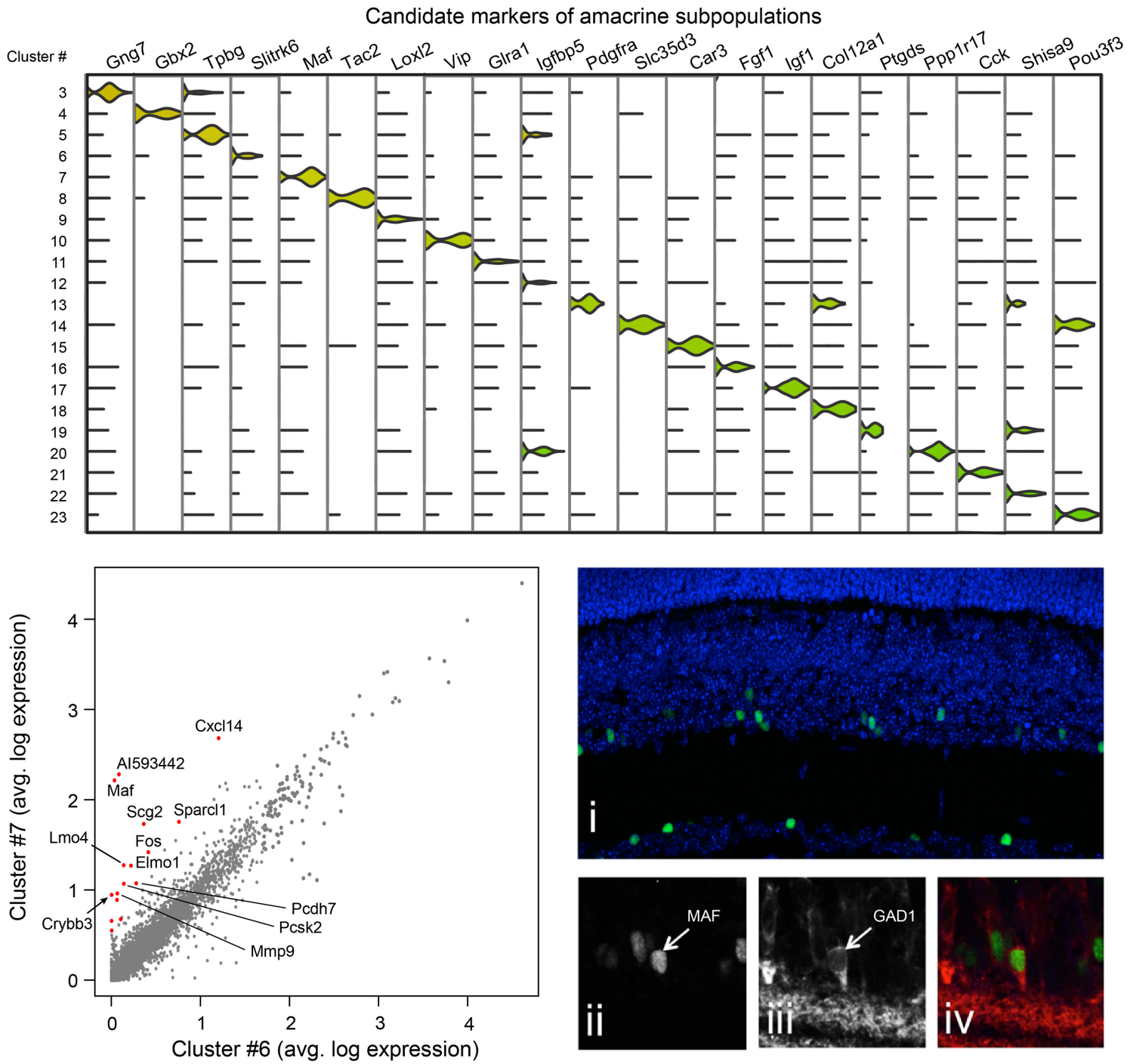
Macosko et al, *Cell*, **161**, 2015

39 Clusters: Known Cell Types & Relationships



Macosko et al, *Cell*, **161**, 2015

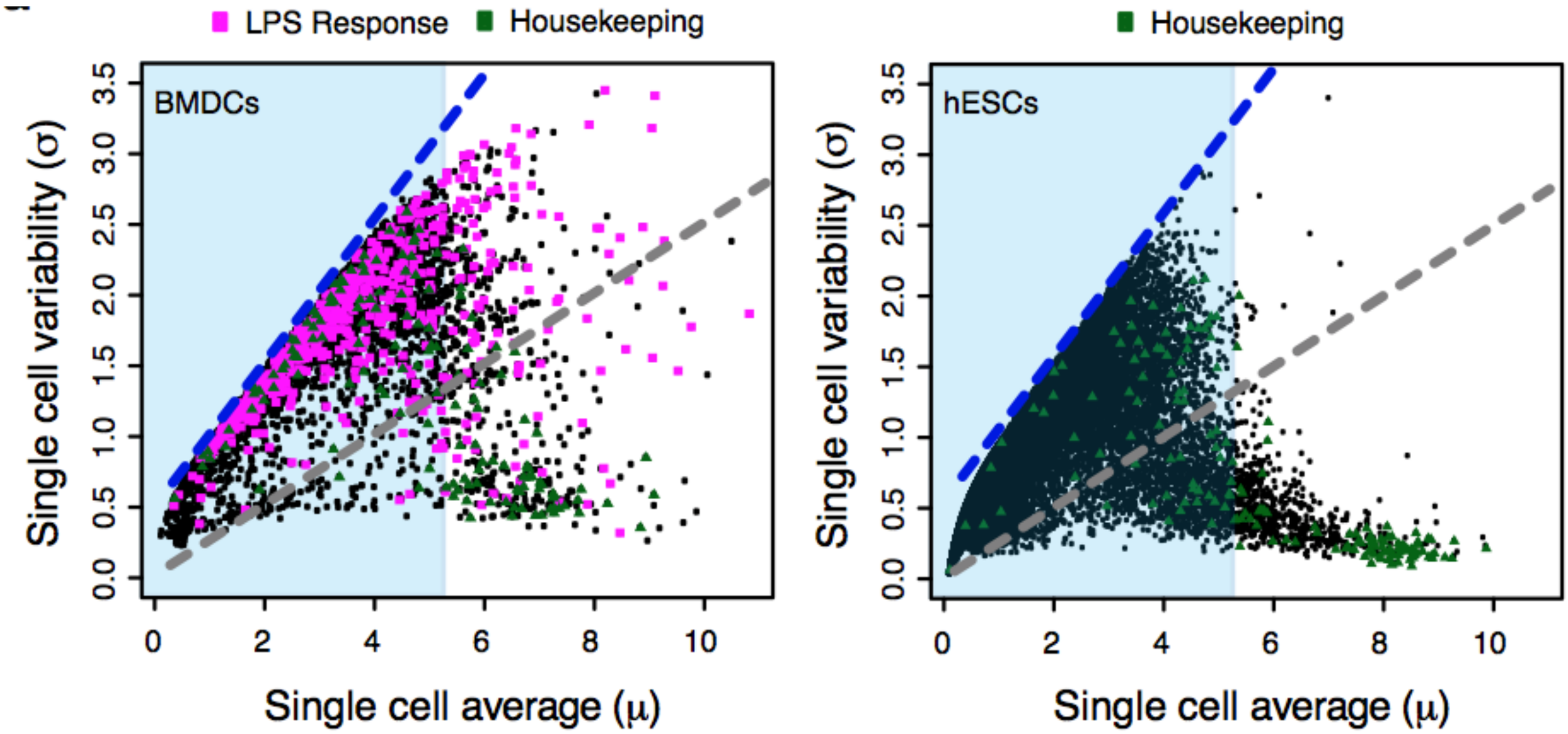
39 Clusters: New Markers That Can Be Validated!



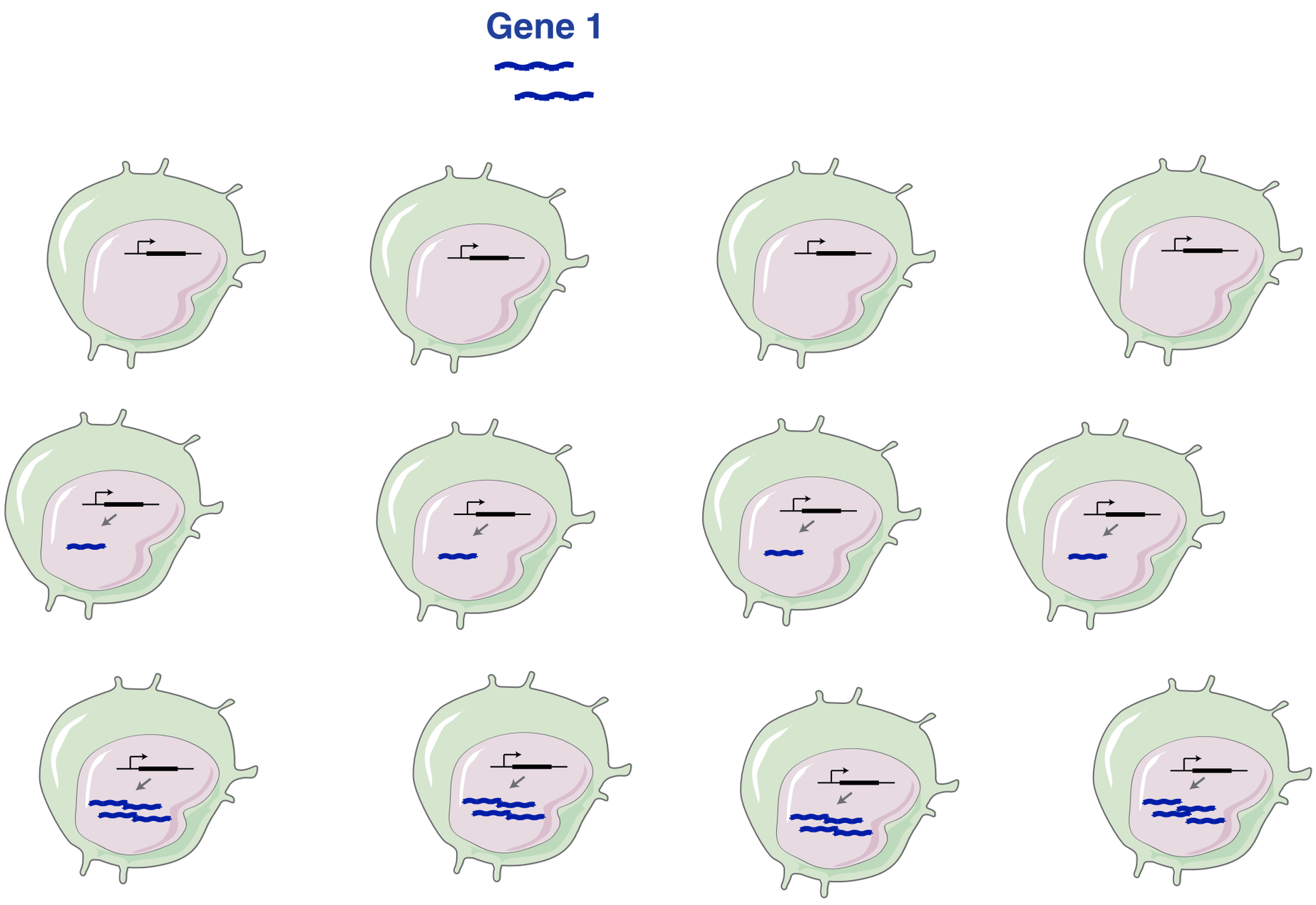
Macosko et al, *Cell*, **161**, 2015

Clustering similar genes

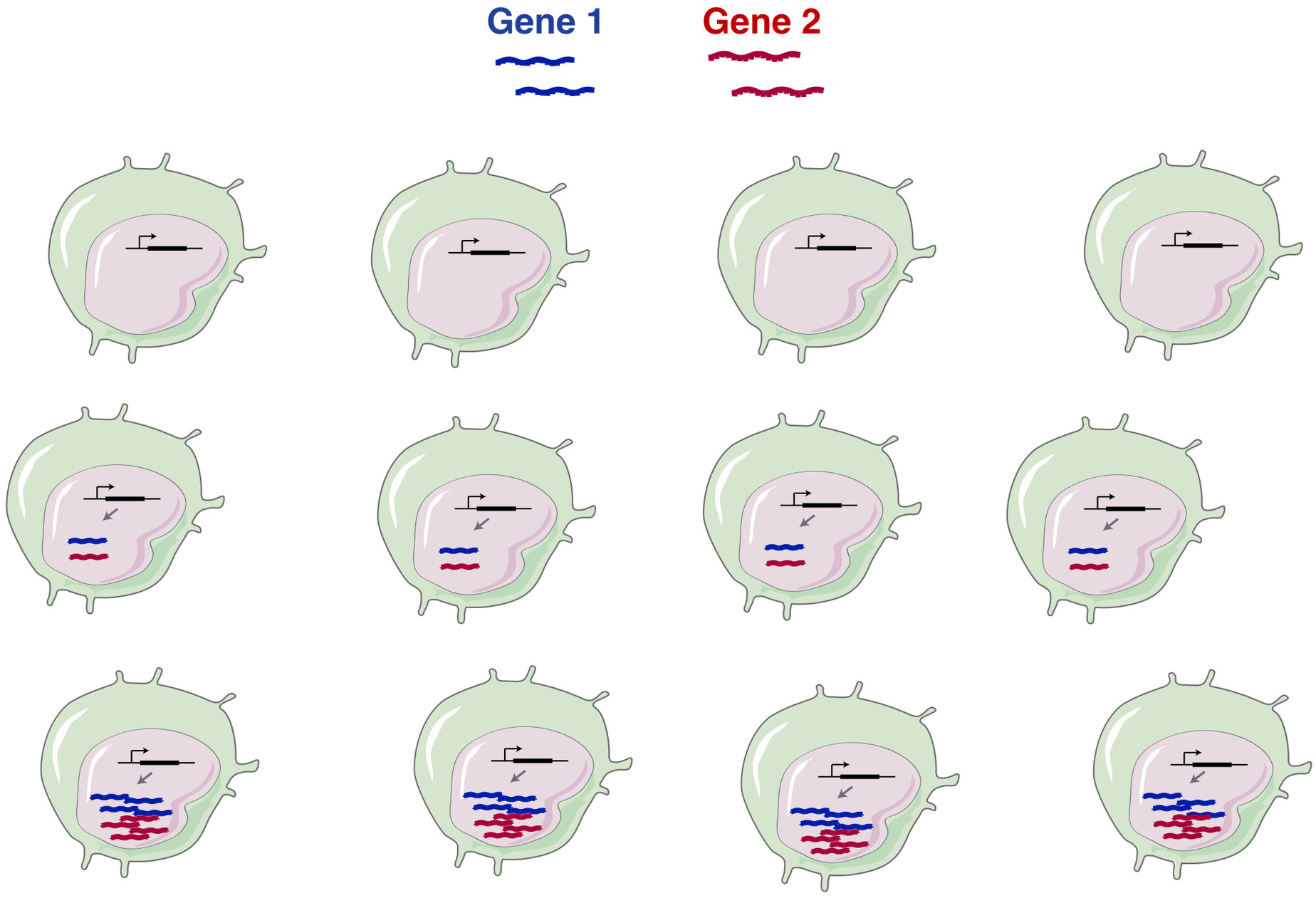
Identifying ‘variable’ genes



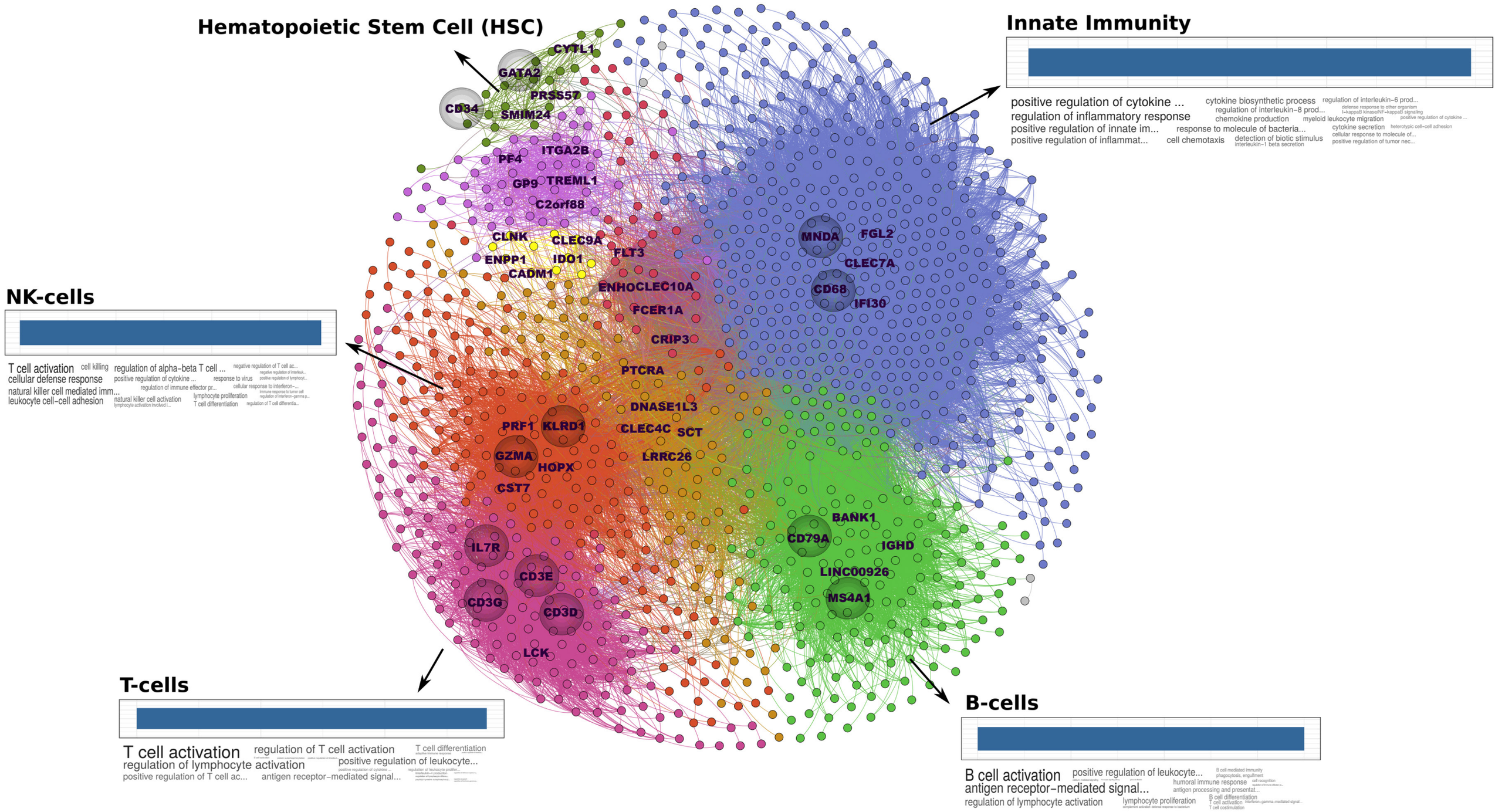
Variation is interesting



Co-variation implies co-regulation

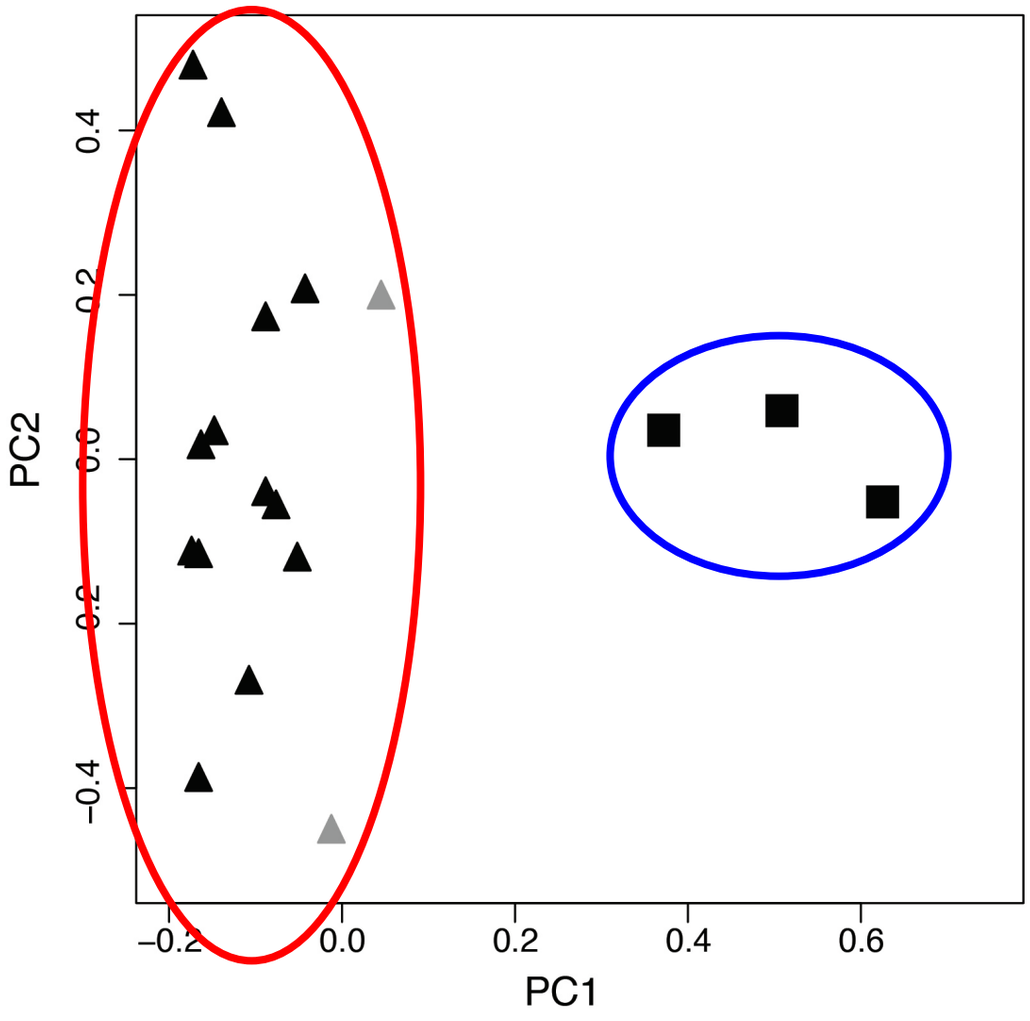


Sparsity-based gene network inference



CO-Dependency network of genes

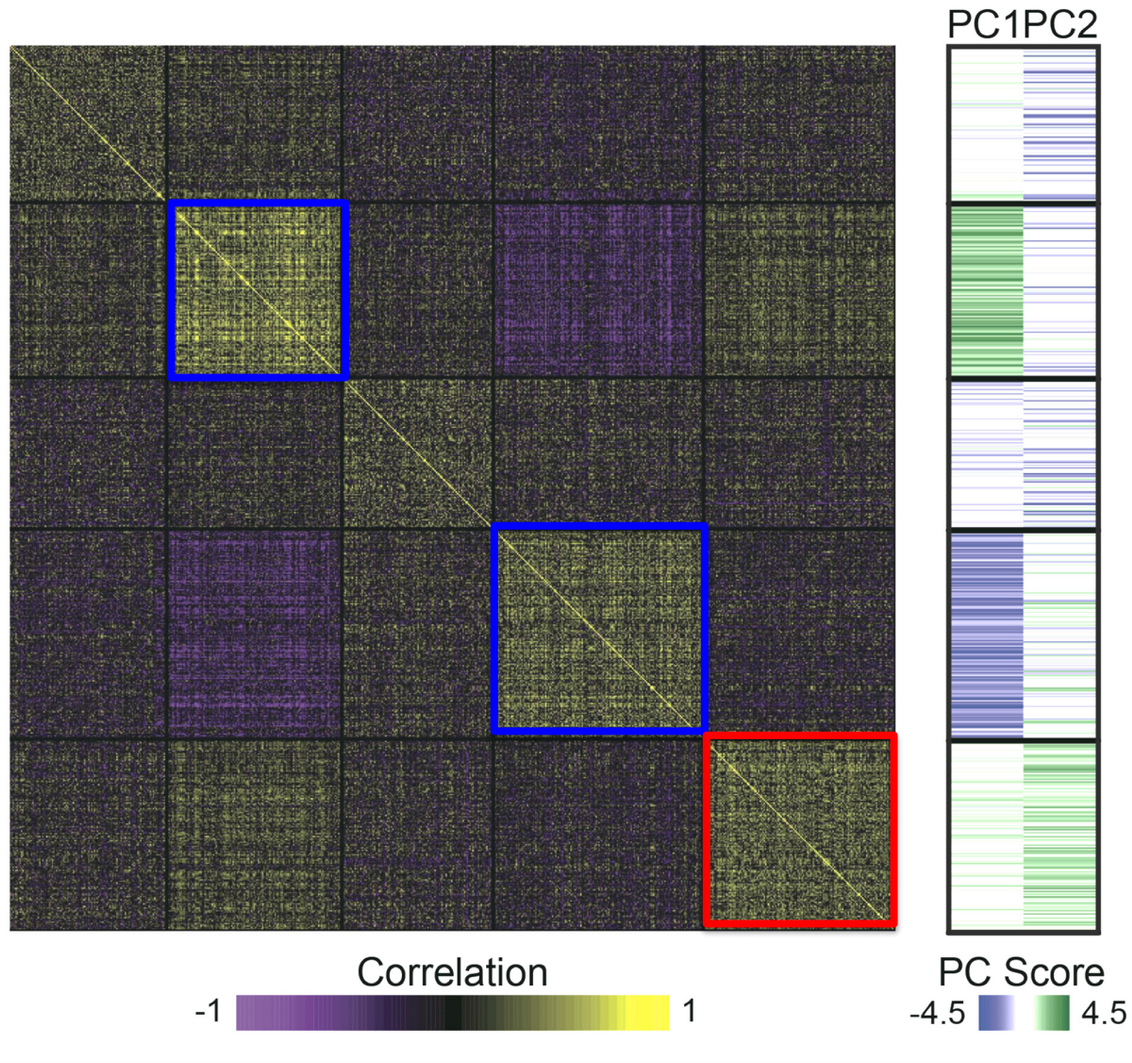
Genes Co-Vary Across Single Cells



Cluster-disrupted
Ccr7, Ccl22,
Serpinb9, Cd83

Maturing
Tnf, Il1a,
Cxcl10, Cd88

Antiviral
Ifit1, Ifih1,
Stat2, Irf7



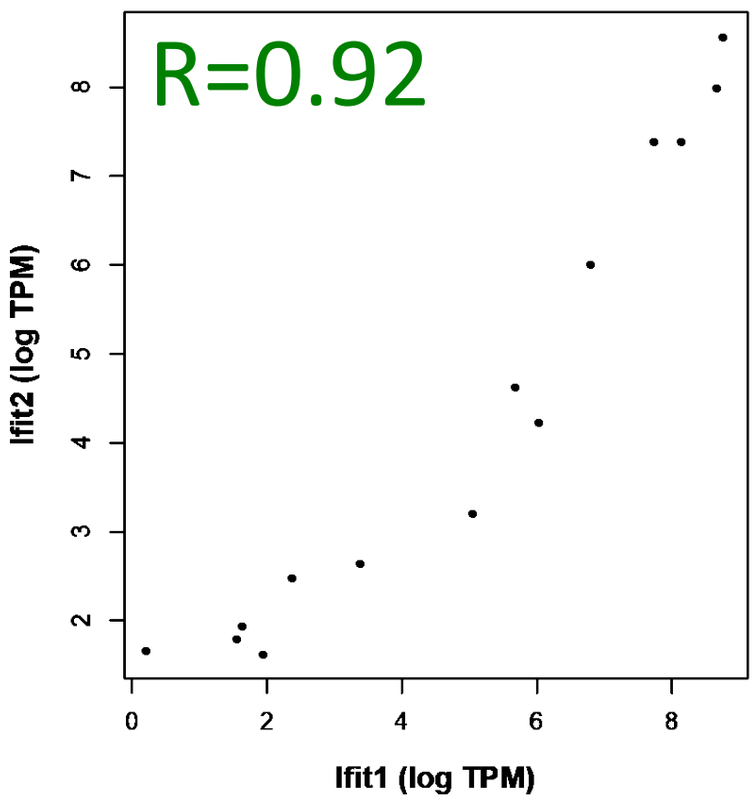
We can uncover cell states and circuits, as well as their markers and drivers, from structures in cell-to-cell variation

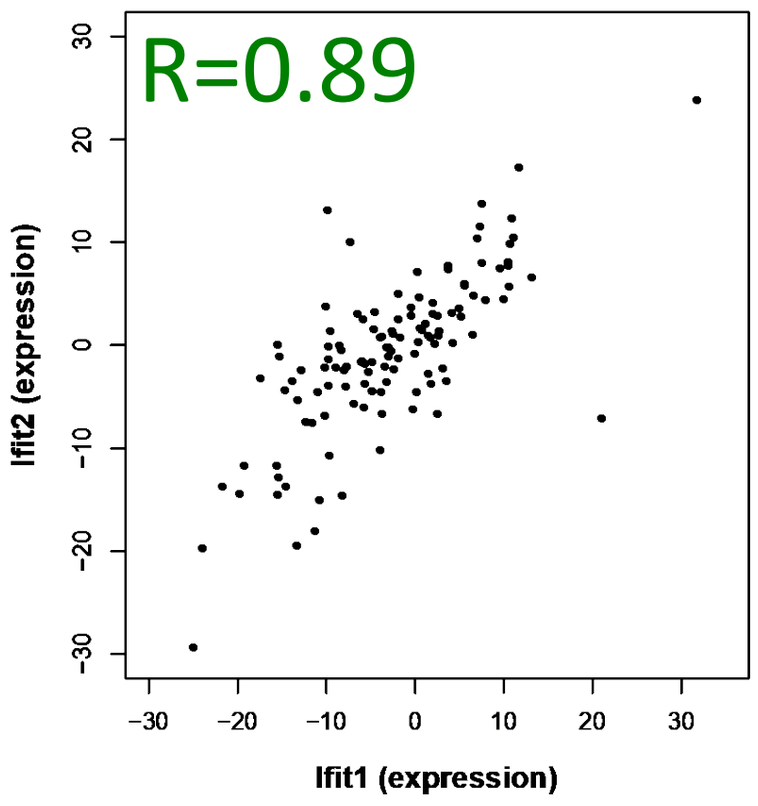
Shalek*, Satija* et al, *Nature*, **498**, (2013)

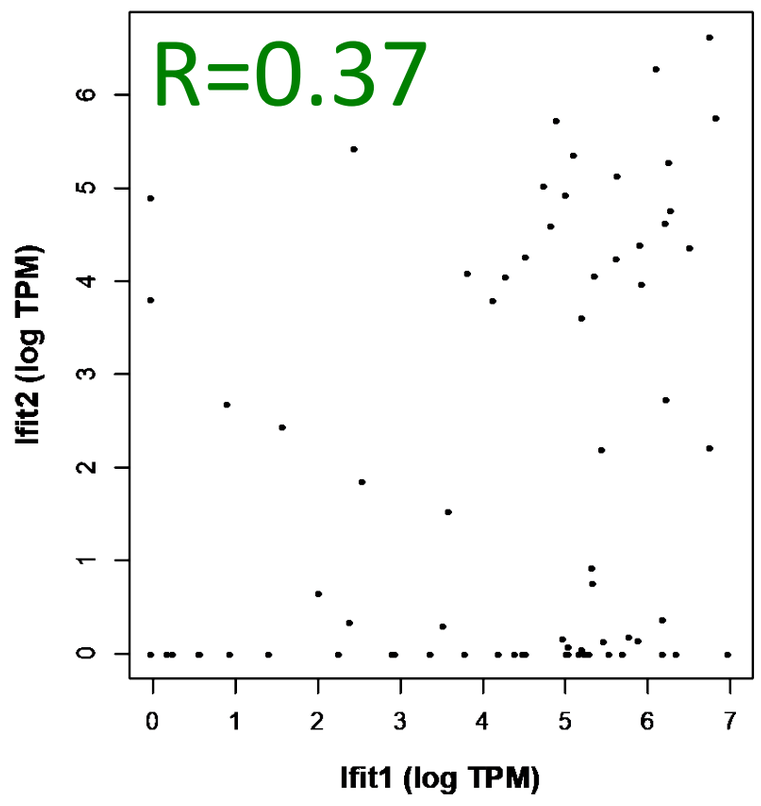
Correlation is not well-suited for single-cell analysis

POPULATIONS

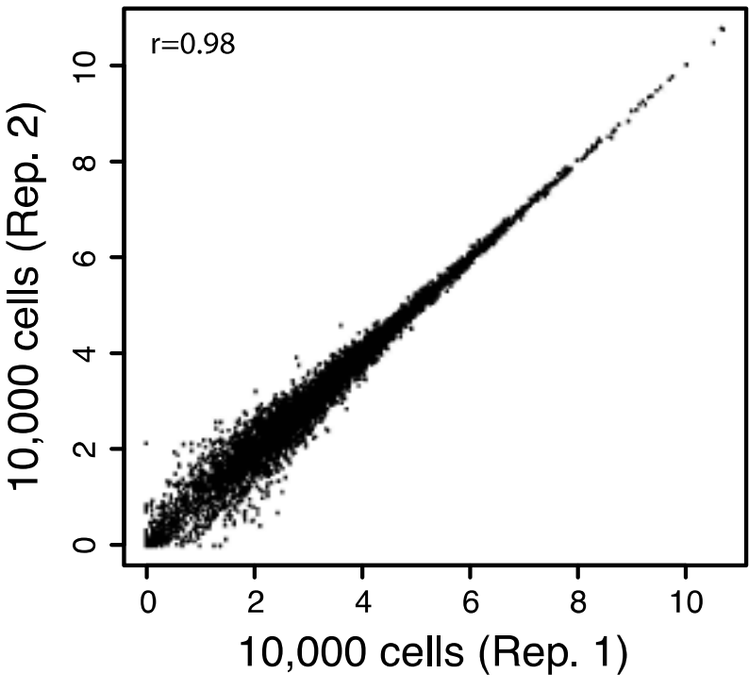
SINGLE CELLS

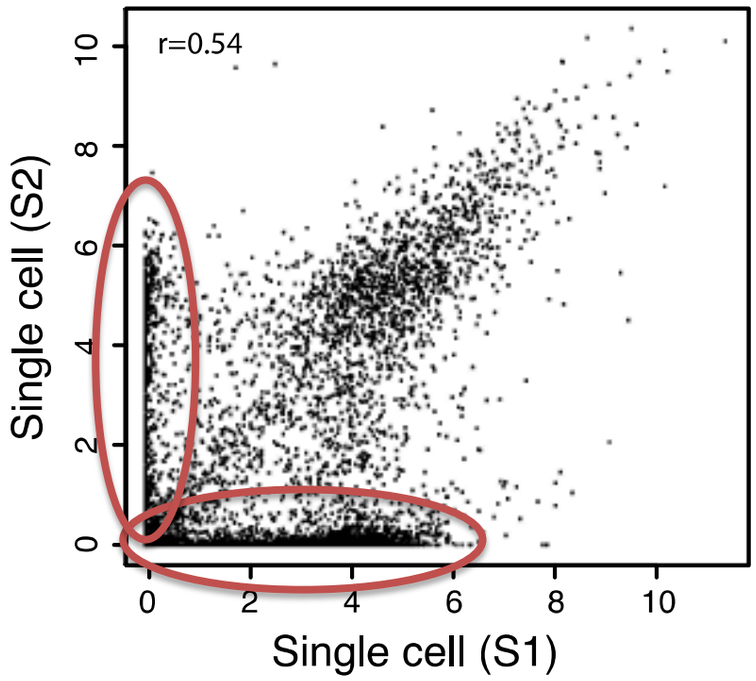


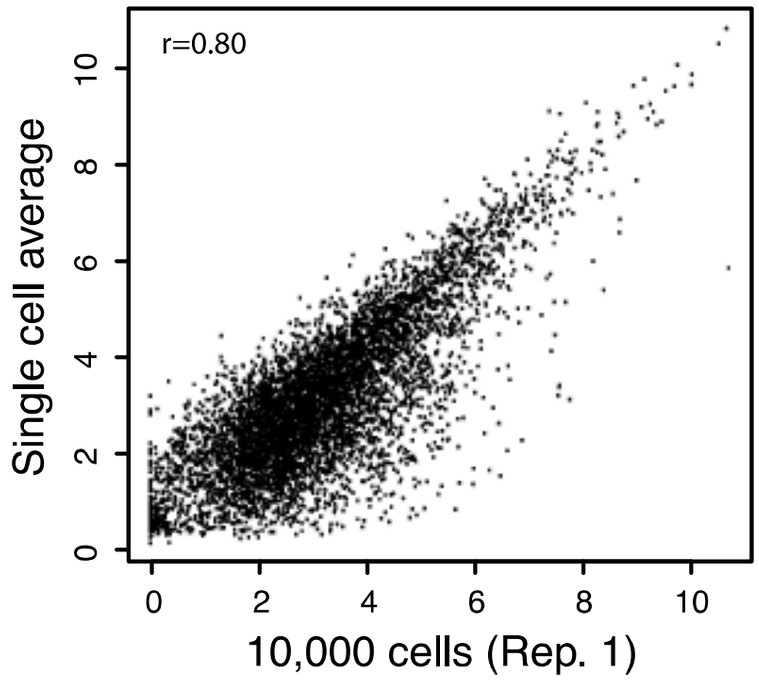




scRNA-Seq data has many many zeros





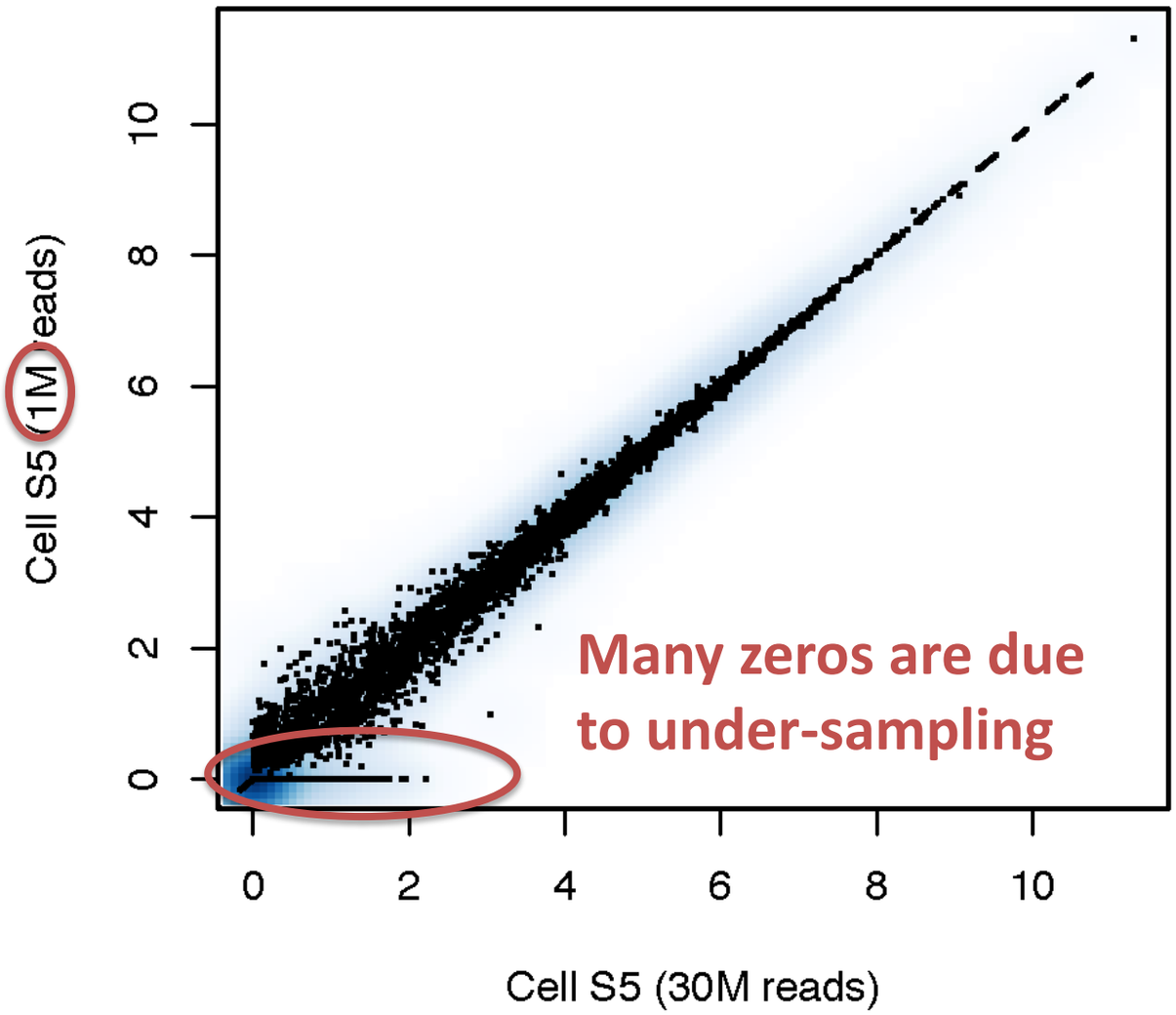


Transcriptome-wide, single cells are very different.

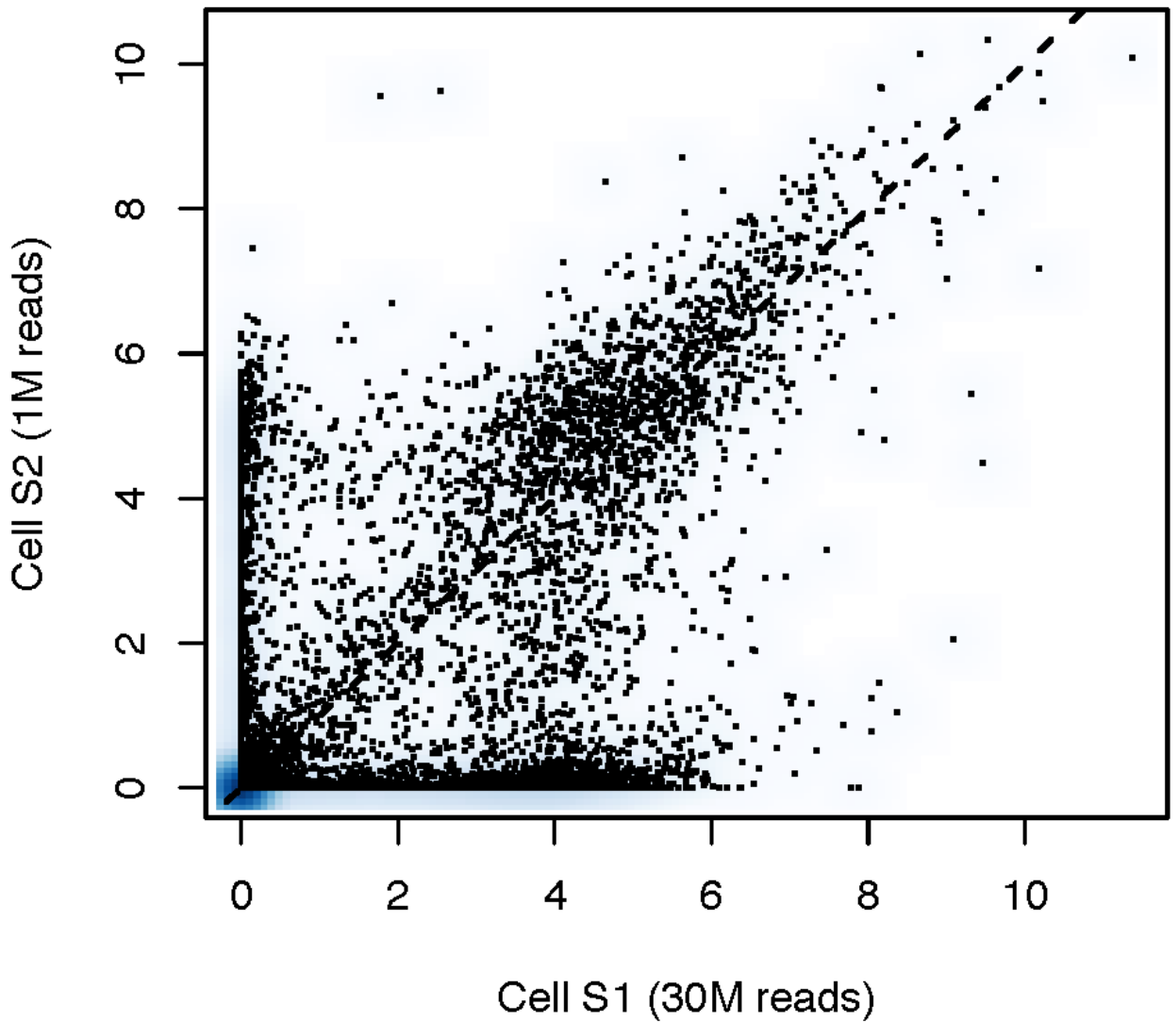
Shalek*, Satija* et al, *Nature*, **498**, (2013)

Variability due to sampling vs. biology

Sampling Bias



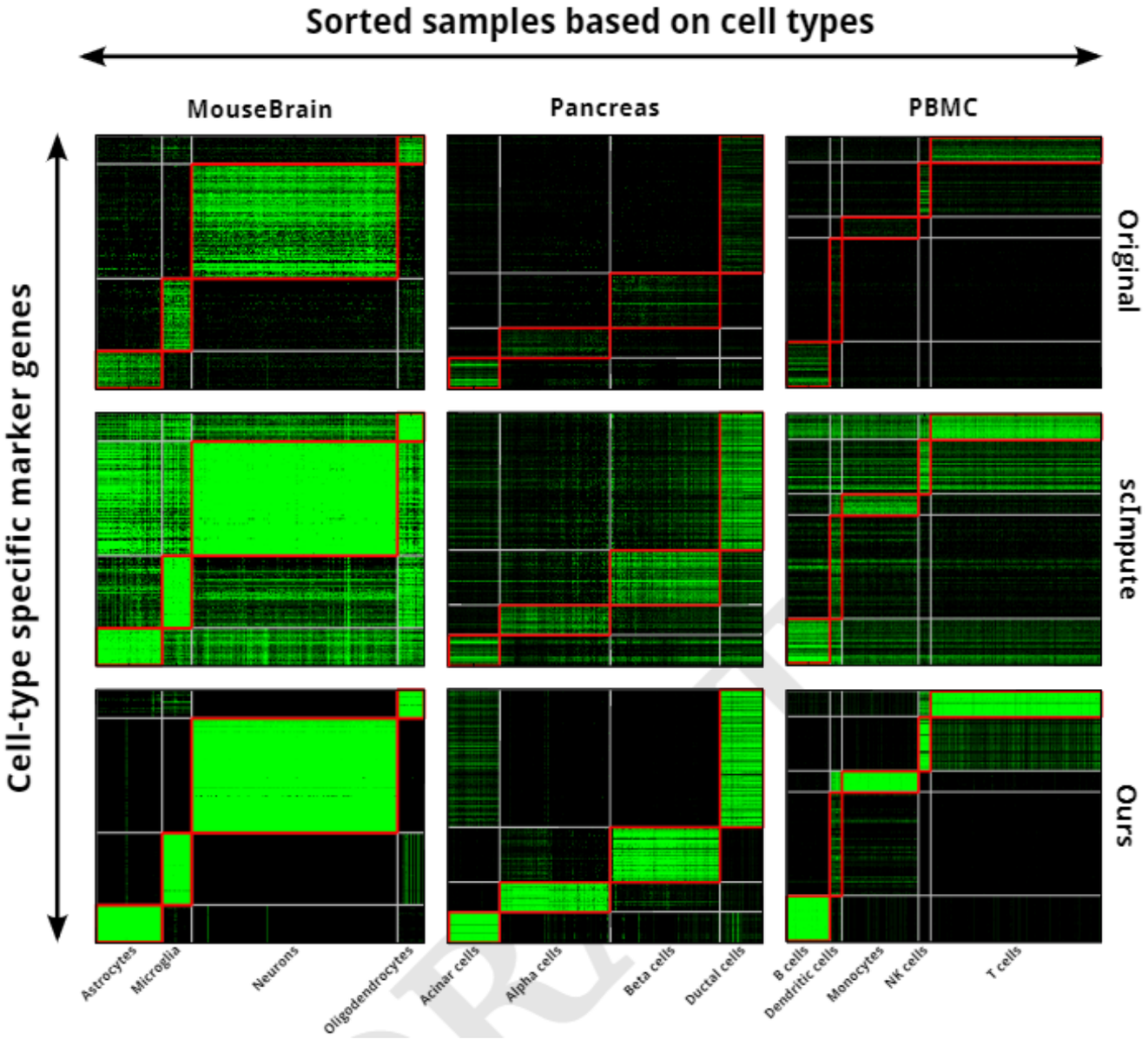
Cellular Variation



Shalek*, Satija* et al., Nature, 2013

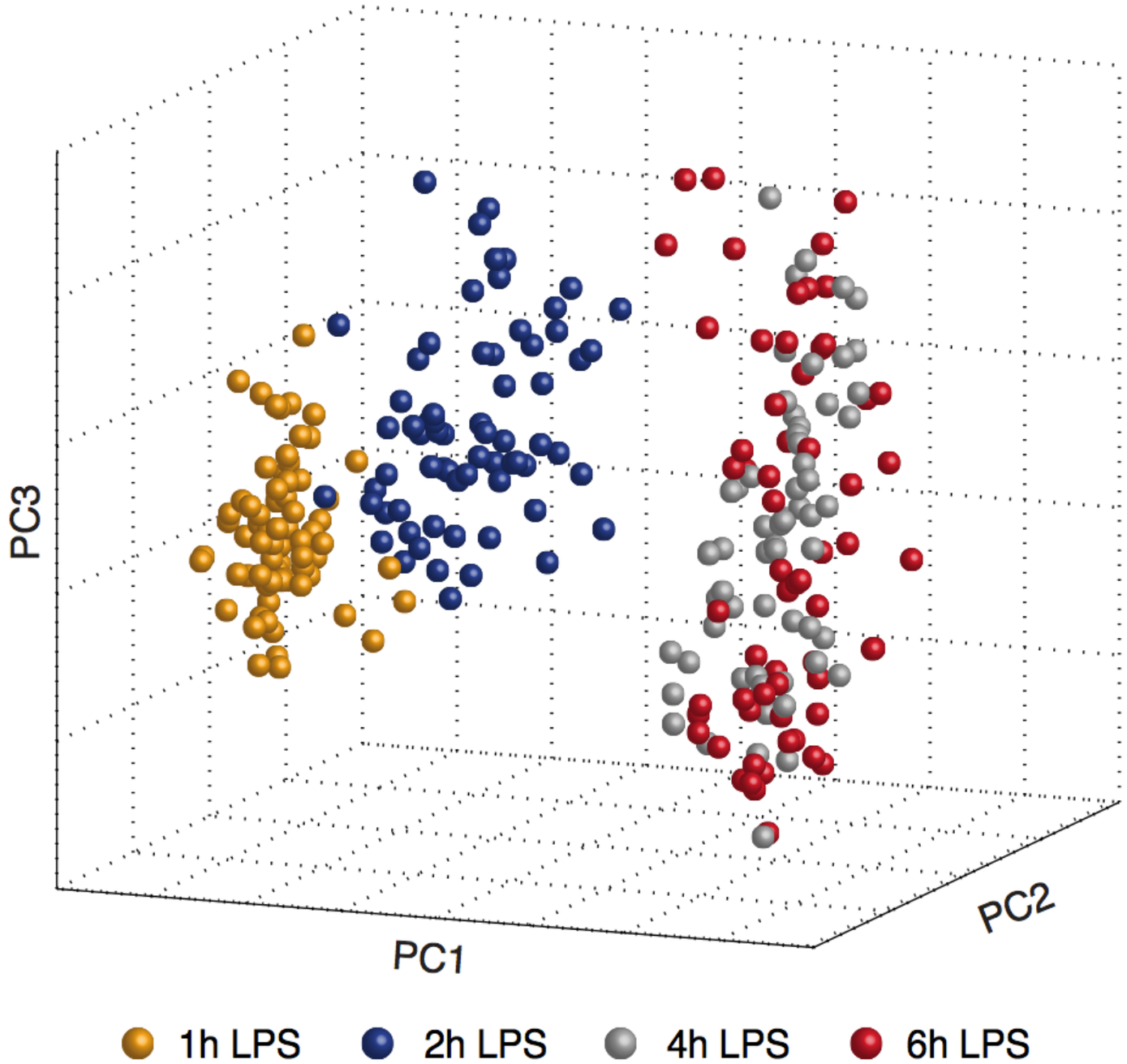
Dimensionality reduction

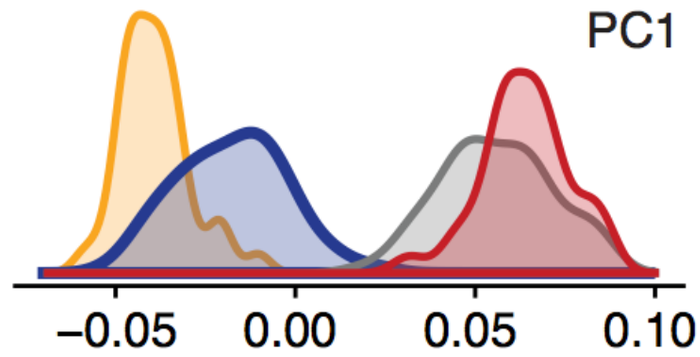
Dimensionality Reduction

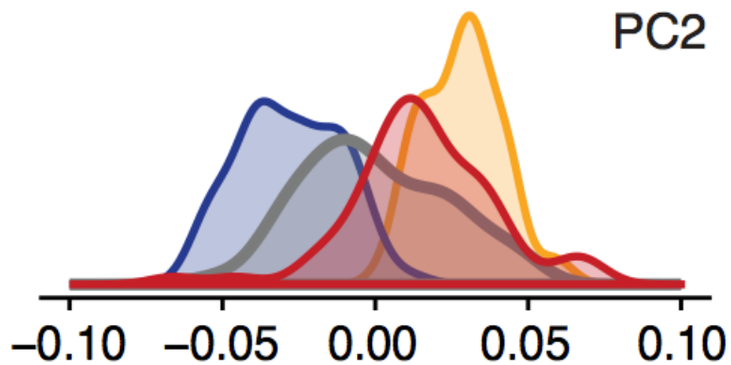


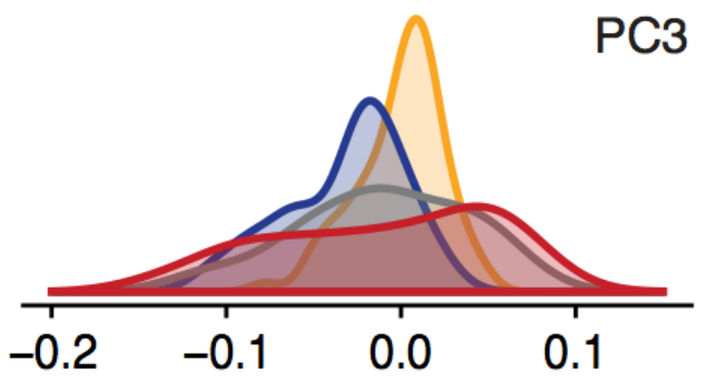
- Curse of dimensionality
- Easier to visualize/process
- Reduce noise
- Linear methods: *PCA*
  - Identifying batch/cell cycle effects effects
- Nonlinear methods: t-distributed stochastic neighbor embedding (*t-SNE*)
  - Exploratory data analysis

PCA – 300 cell dataset









Important consideration for PCA

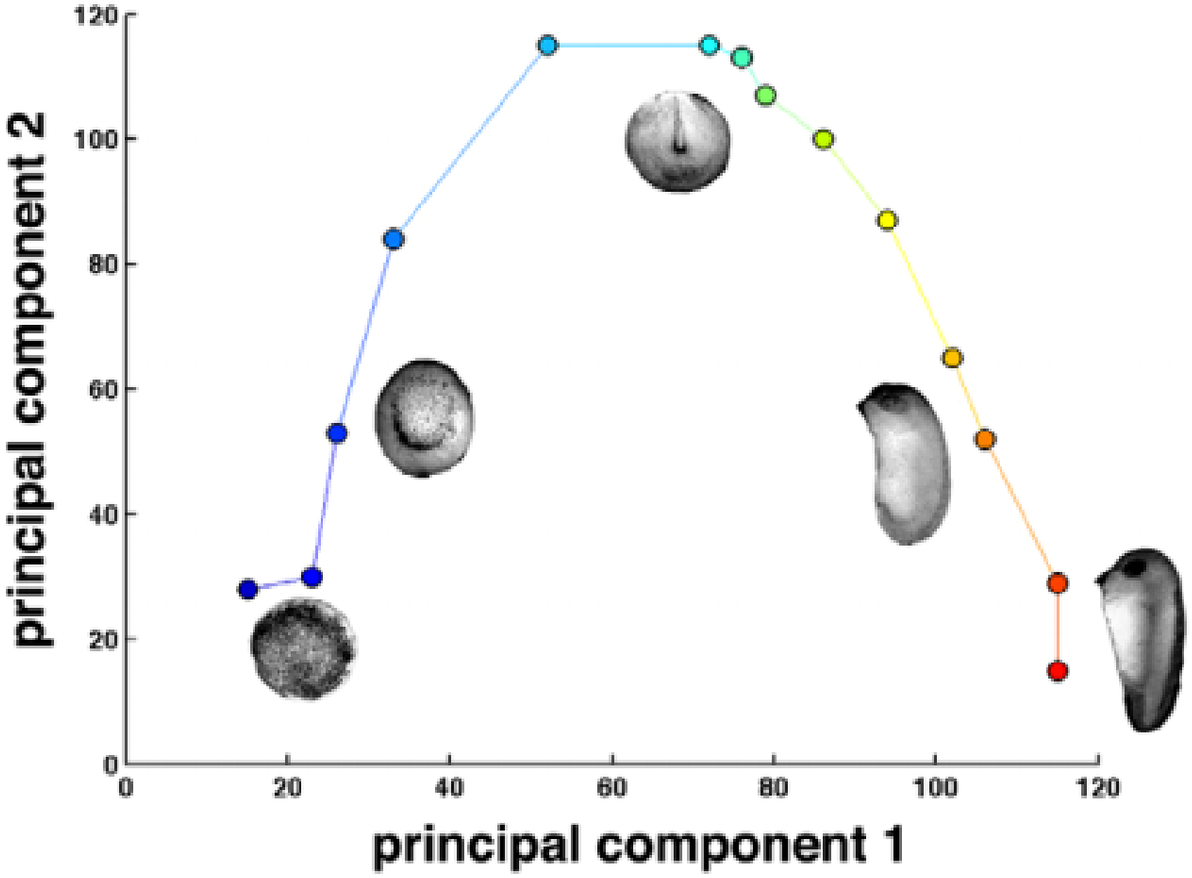
- **Input gene list**
  - Can dramatically alter output

- **Interpretation:**
  - ‘Assigning ‘biology’ or function requires prior knowledge
  - PCs often correlate with technical quality
  - Not all PCs are significant (Chung, Storey, arXiv.org)

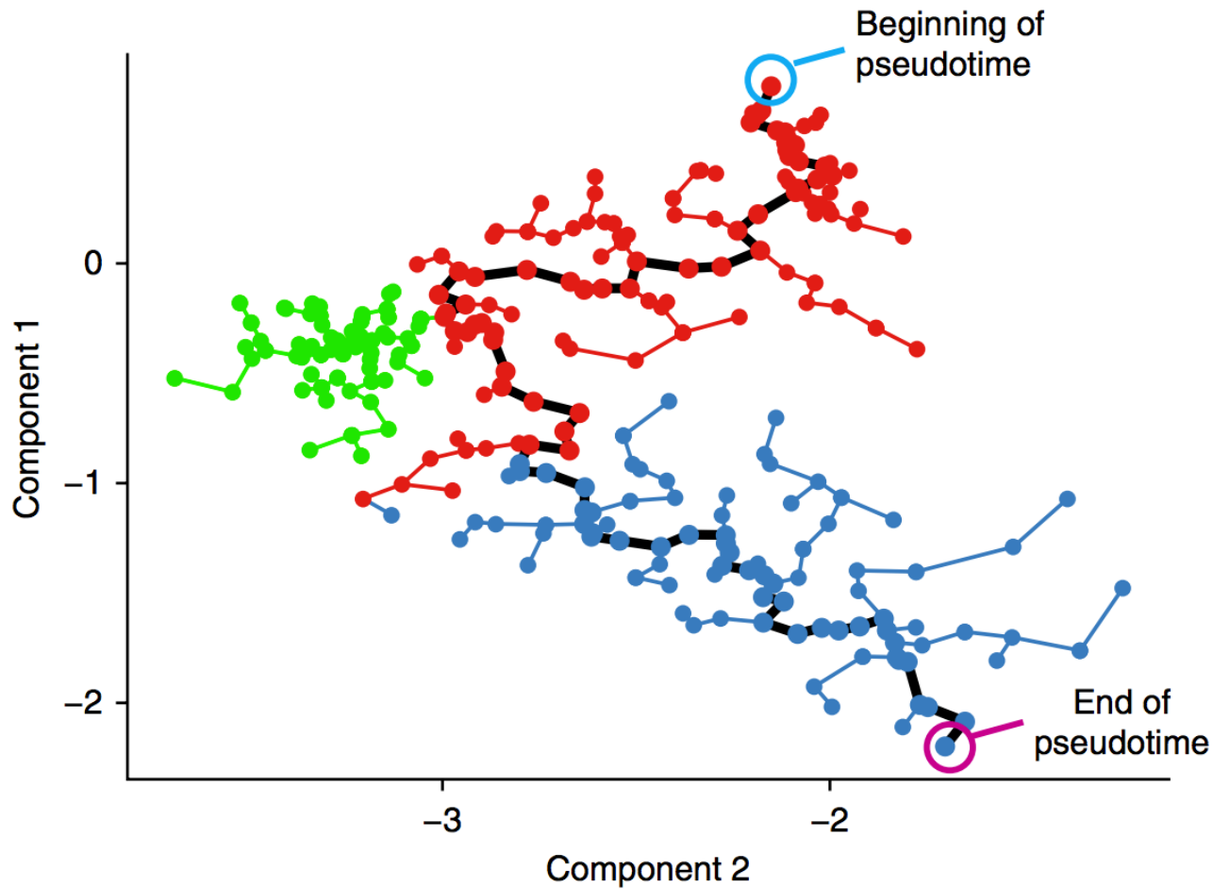
- **Limitations/extensions:**
  - PCs represent **linear** combination of individual features

Interpreting dimensionality reduction

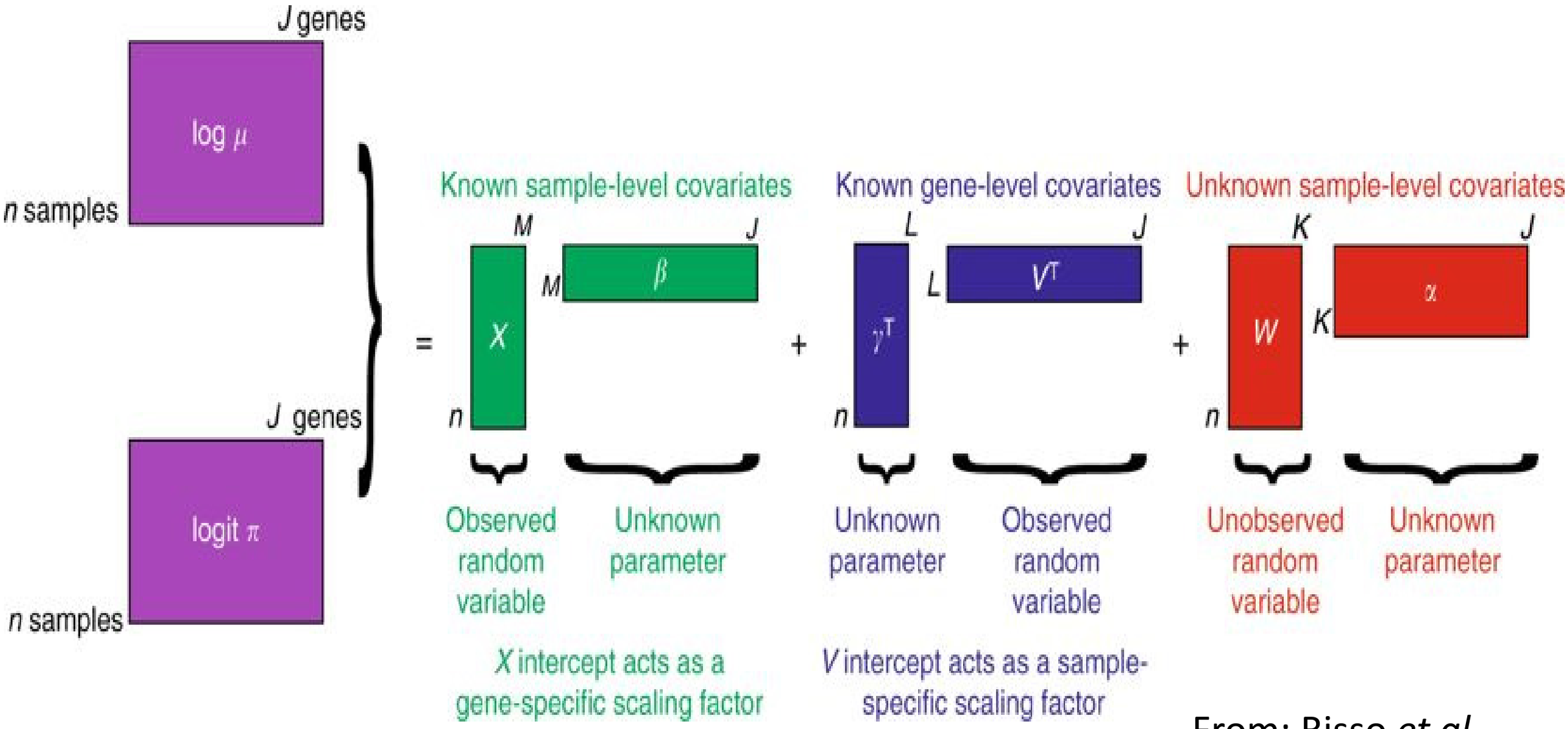
Anavy et al, Development, 2014



Trapnell et al, Nat. Biotech., 2014



Zero-inflated negative binomial model (ZINB-WaVE)



From: Risso *et al., 2018*

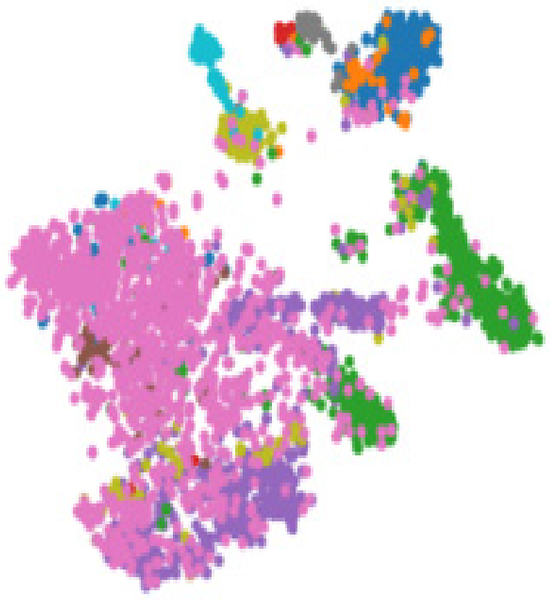
- A generalized linear factor analysis model

Dropping factor analysis in favor of deep autoencoders

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

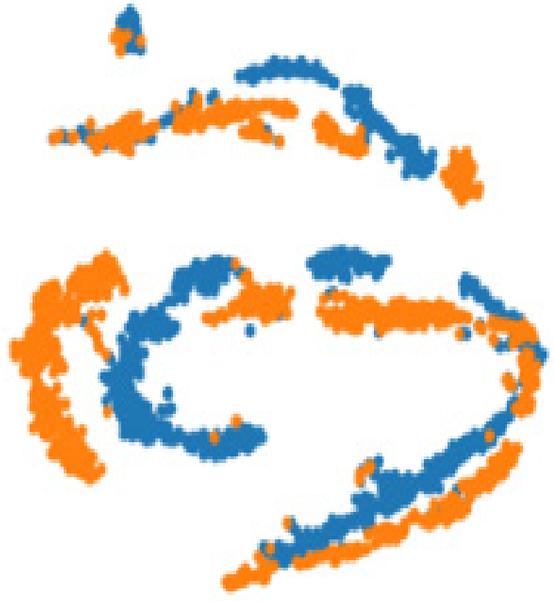
tSNE of low dimensional representation

DCA - Rosenberg 2018

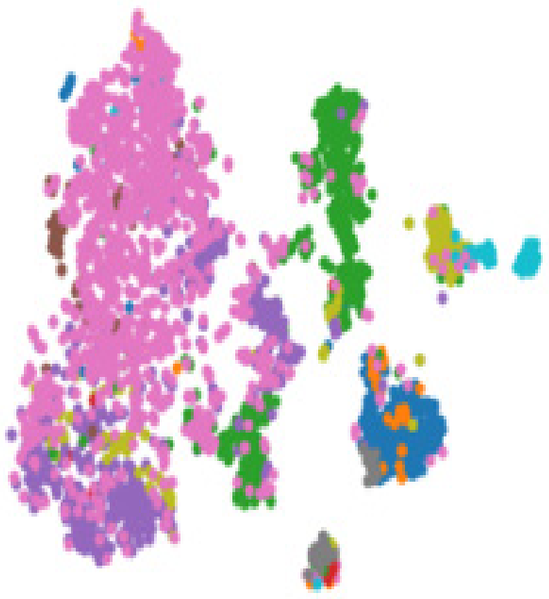


DCA - Zeisel 2018

DCA - Lukassen 2018

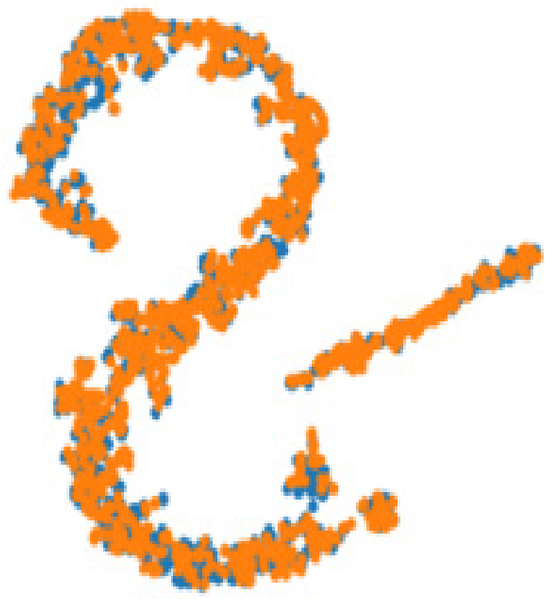


scVI - Rosenberg 2018



scVI - Zeisel 2018

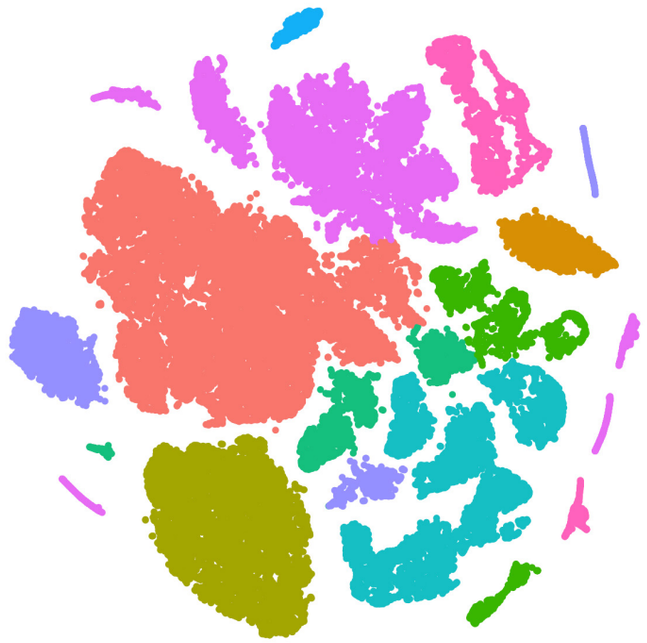
scVI - Lukassen 2018



From: V. SVENSSON, 2018

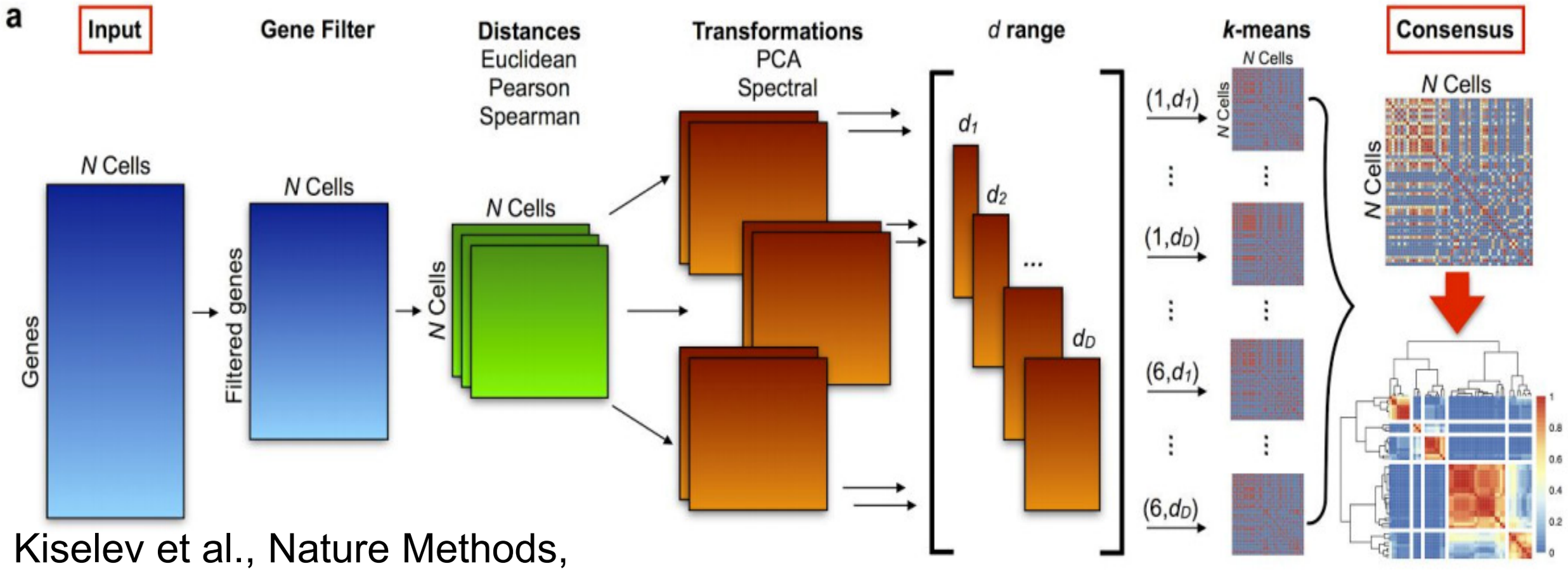
Dinstinguishing different cell types

Discrete cell type identification



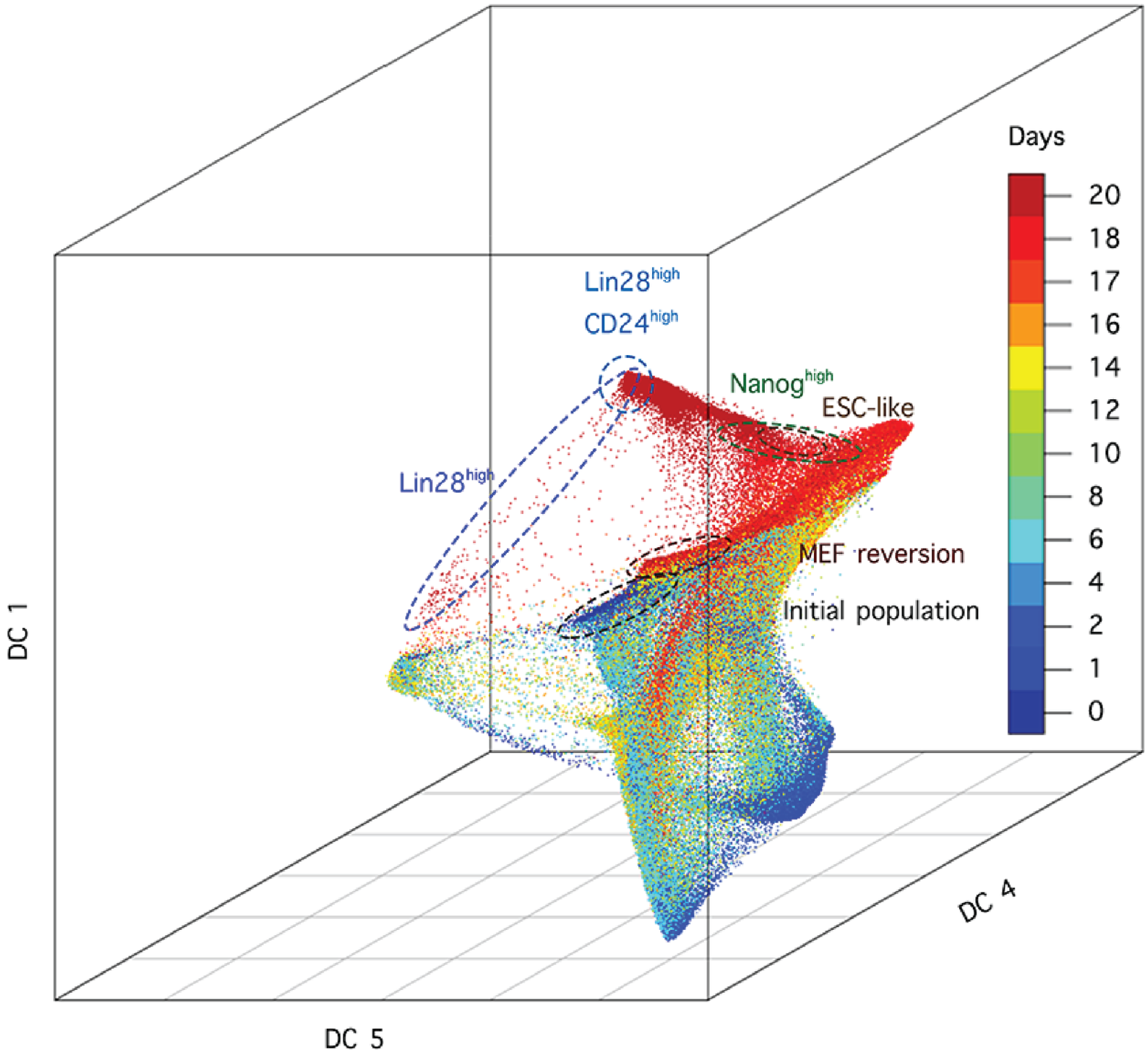
- **Based on traditional clustering approaches:**
  <u>k-means, hierarchical, and graph-based clustering techniques</u>
- **tSNE + k-means (traditional)**
- **SINCERA (Guo et al. 2015)**
  - Based on hierarchical clustering
  - Data is converted to z-scores before clustering
- **SNNCliq (C. Xu and Su 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 explaning 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)



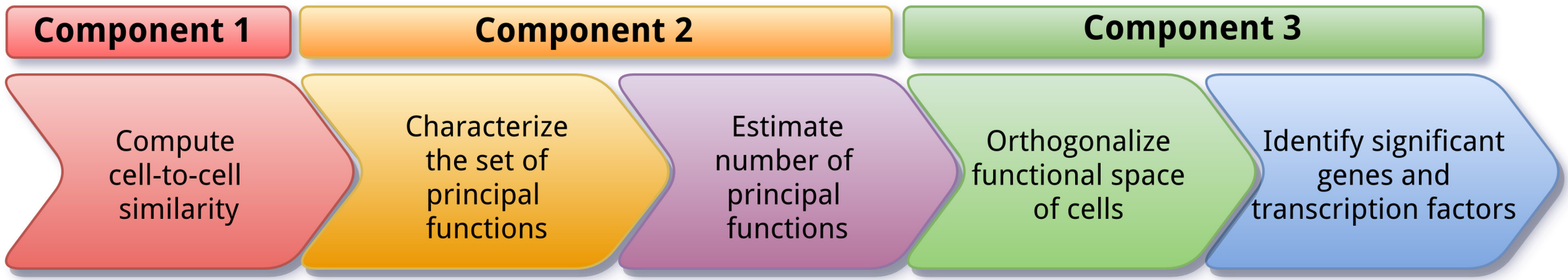
Kiselev et al., Nature Methods, 2017

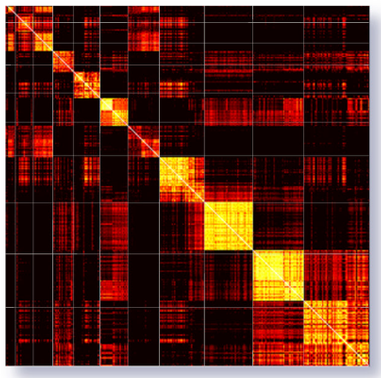
Continuous cell states: diffusion map

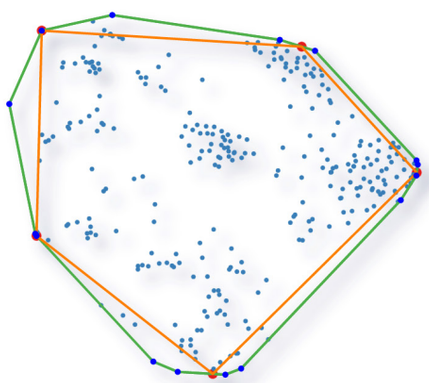


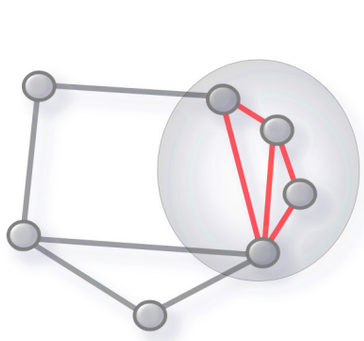
Angerer *et al*, Bioinformatics 2016

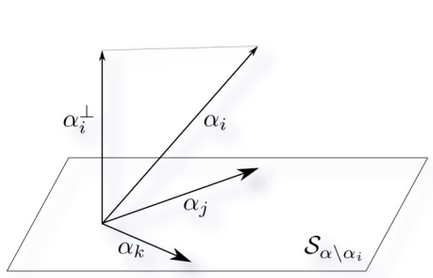
Archetypal-analysis for Cell type indentificaTION (ACTION)

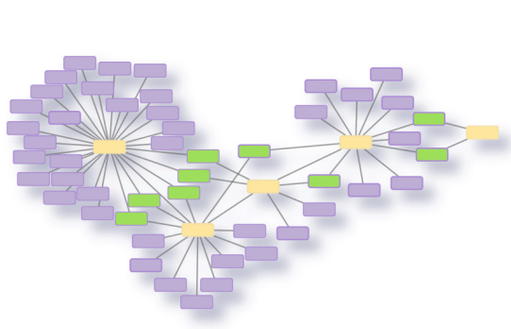






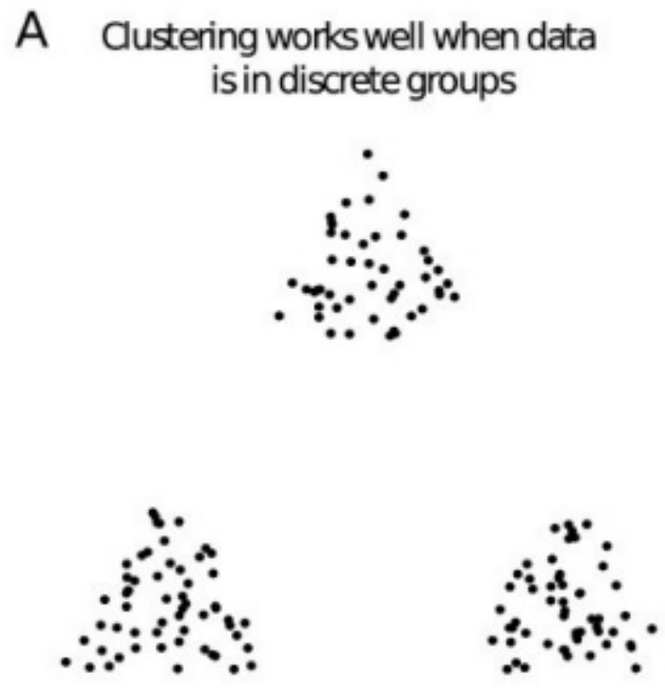


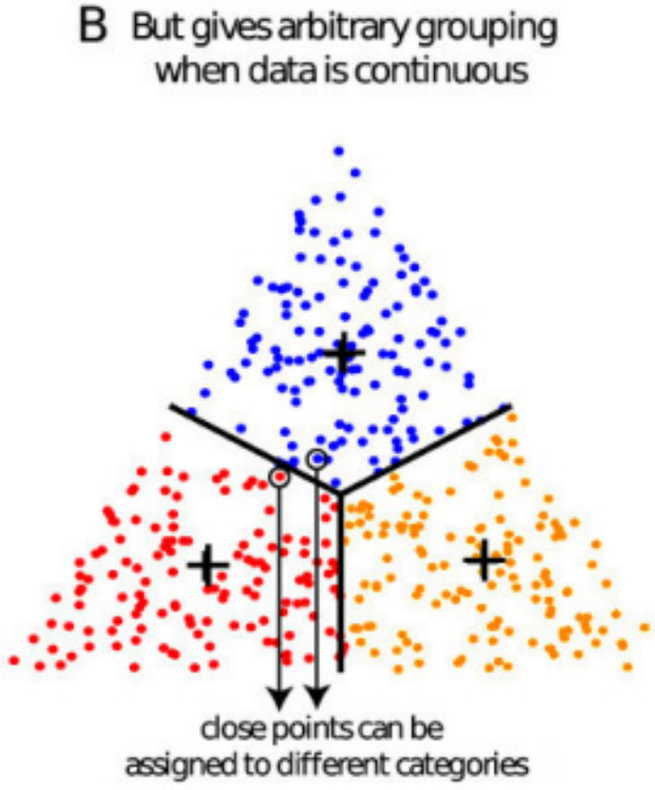


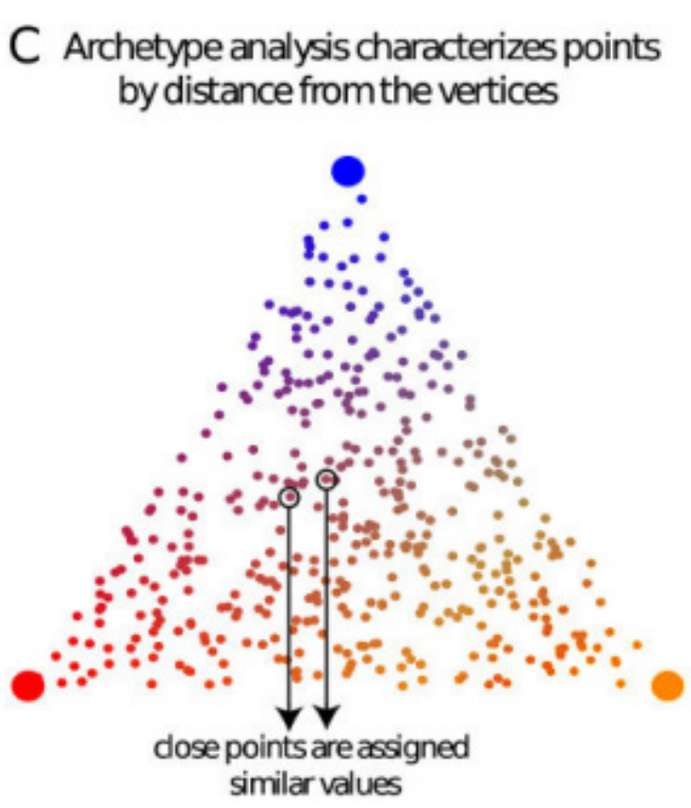


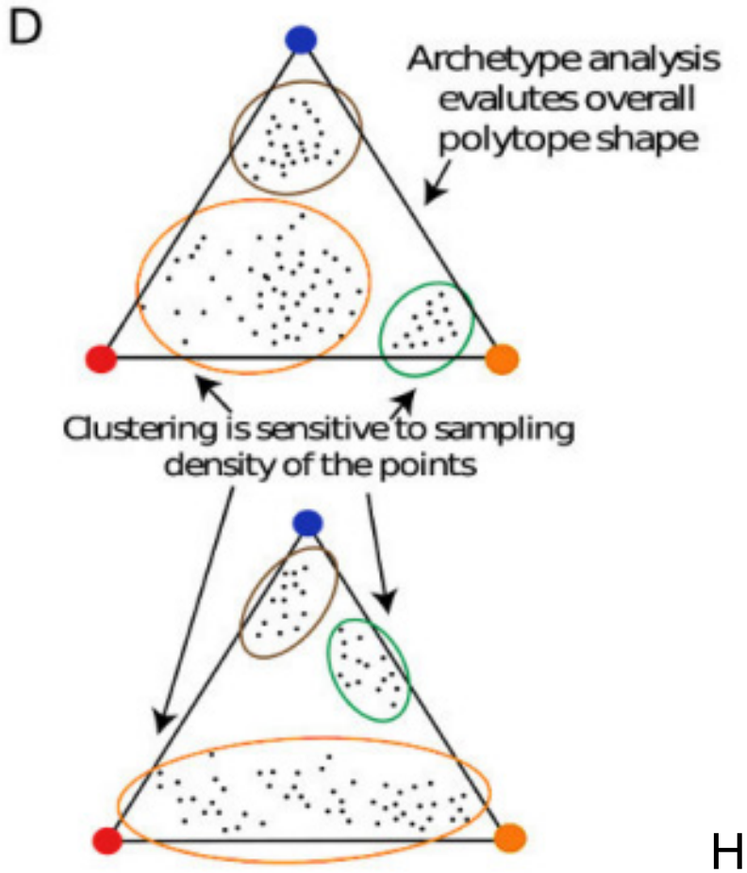
Mohammadi *et al.*, BioRxiv 2016, Nature Communications, under review

Combine discrete + continuous: archetype analysis



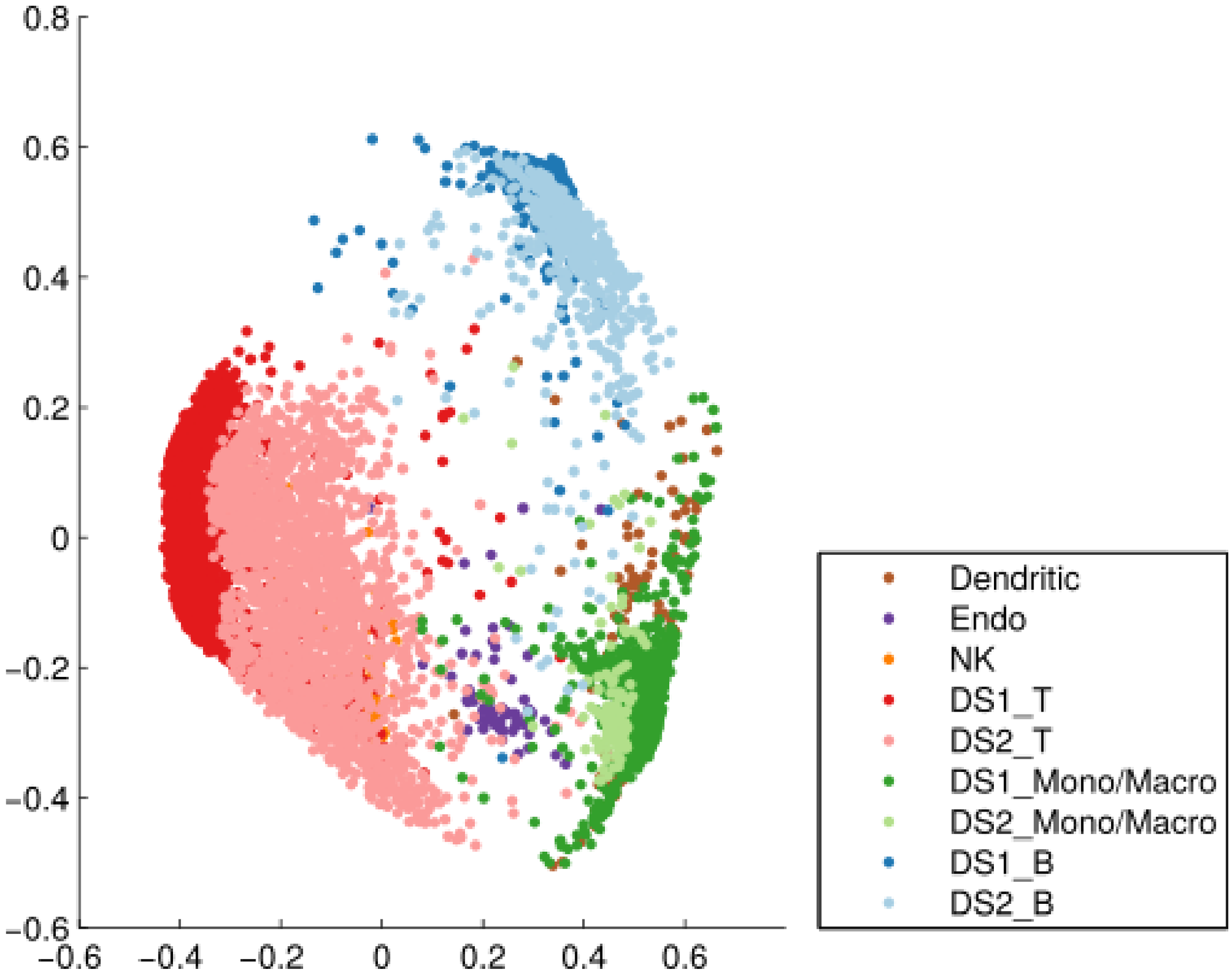






Hart *et al*, Nature Methods 2015

Matching cell types across datasets



Alignment of PBMC vs. Tumor scRNA

Multi-resolution analysis

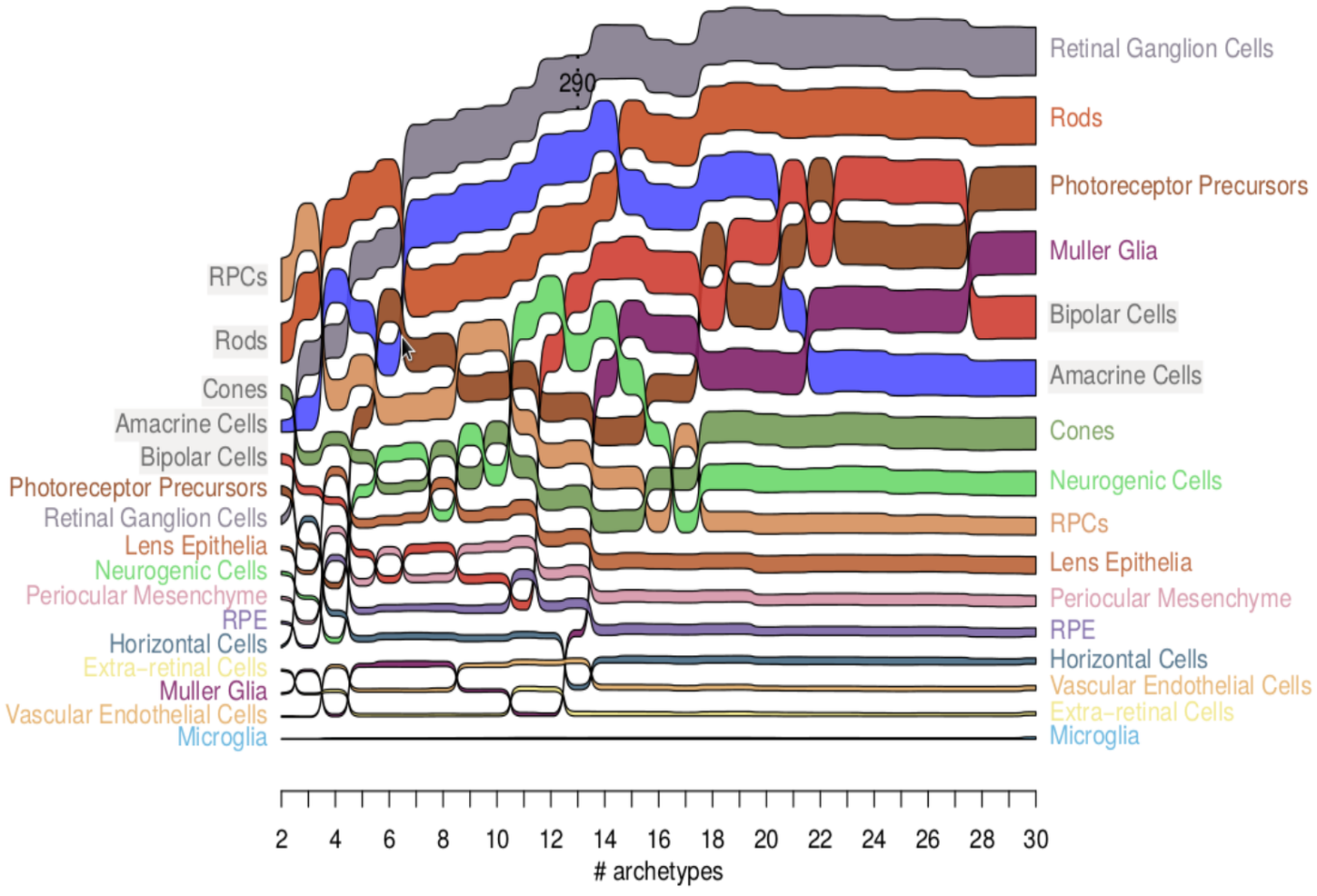
ACTIONet

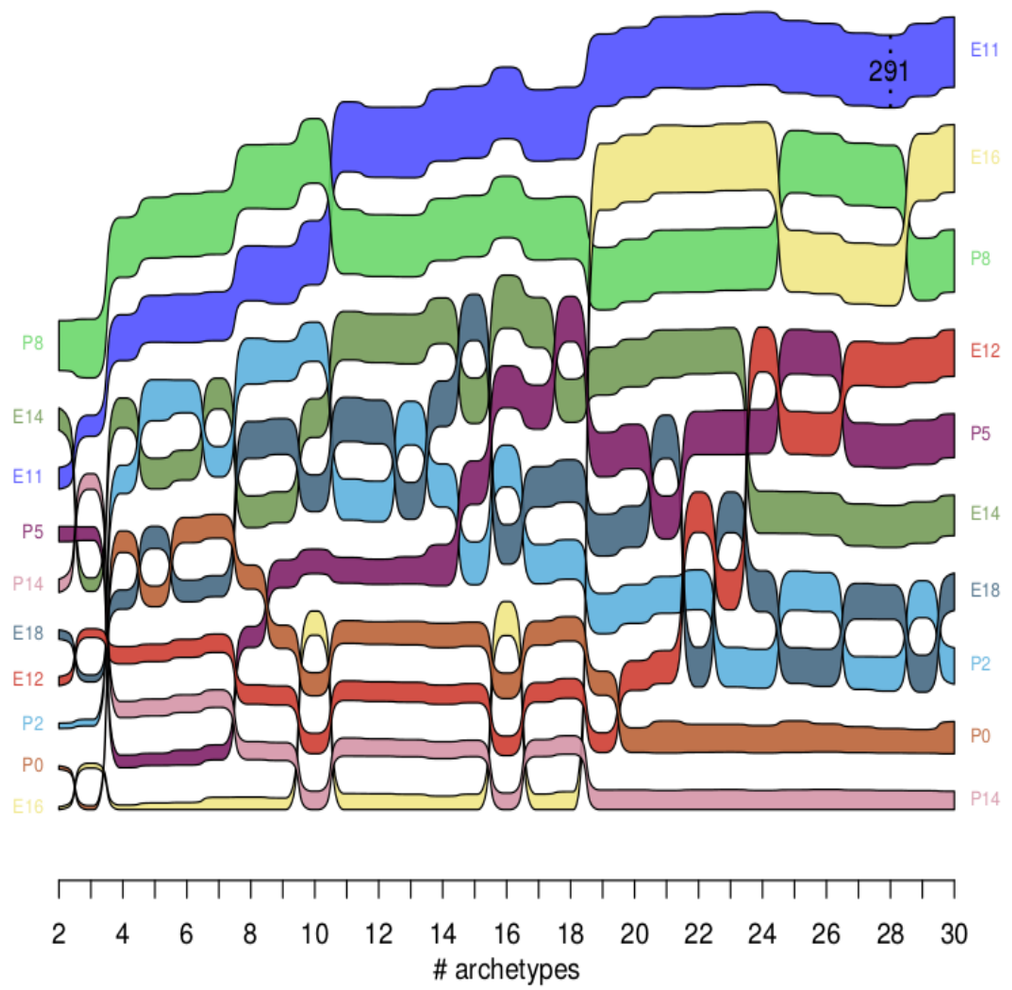
Main issues with parametric methods

How many archetypes?
How many factors?
How many clusters?

Optimal number of factors differs by celltype/age

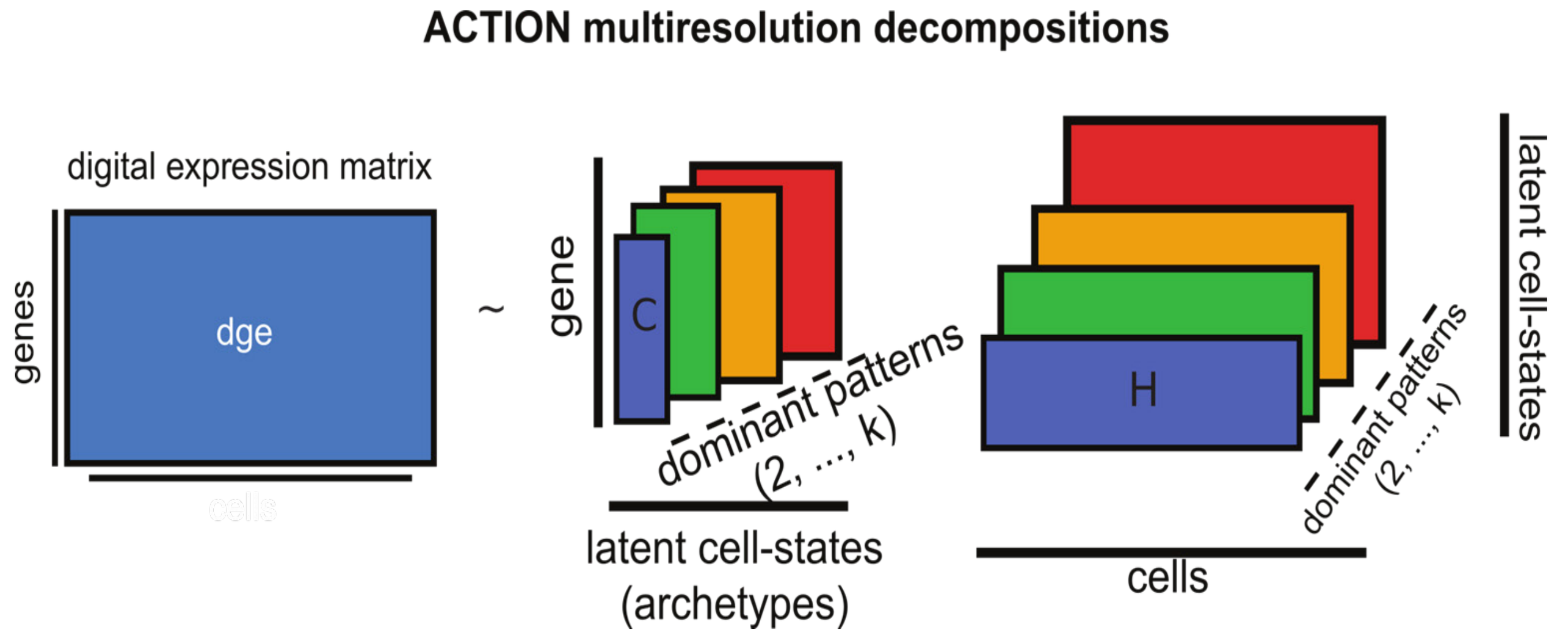
(Ex.: Mouse retina development -- similar results with other species/tissues)





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

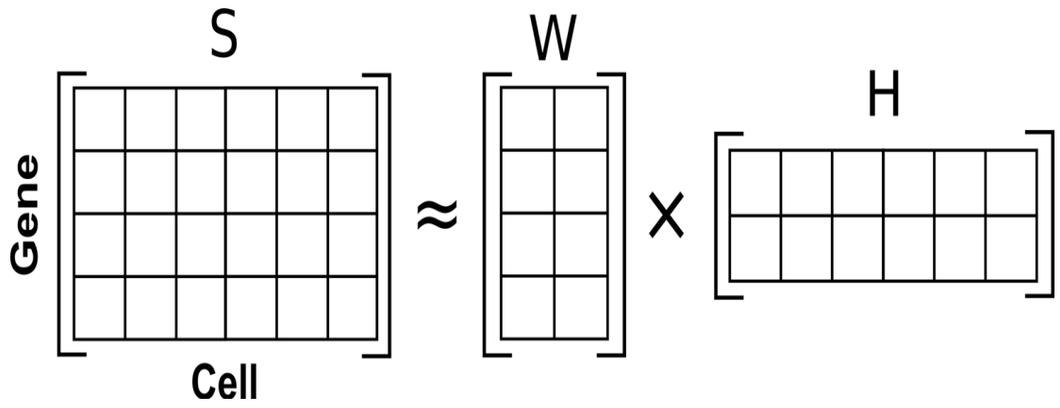
Recently developed method: ACTIONet



Complementary approaches

Factor analysis

ACTION



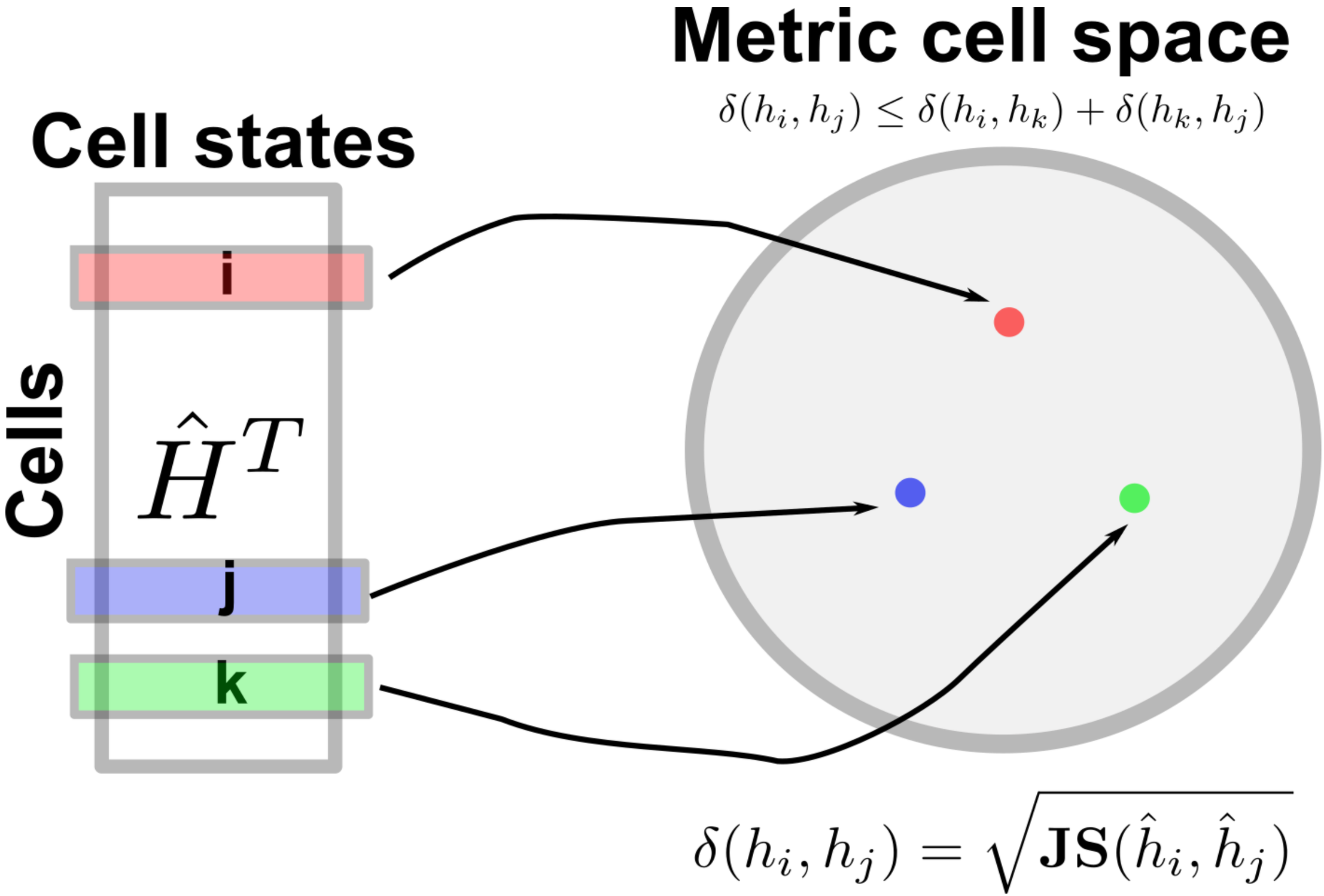
- Identifies hidden cell states/gene programs
  - Biological
  - Technical
- Deconvolves complex biological processes
- Uncovers discriminating/marker genes

Network analysis

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

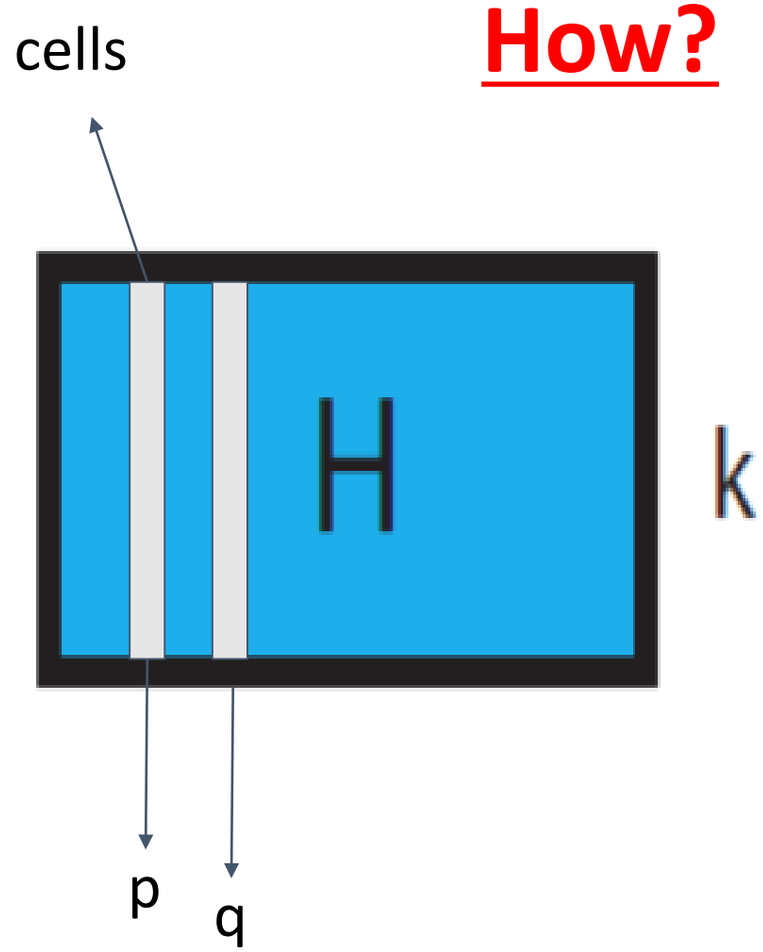
ACTIONet

Step 1: Define a metric cell space



Step 1: Define a metric cell space

How?



Kullback–Leibler divergence

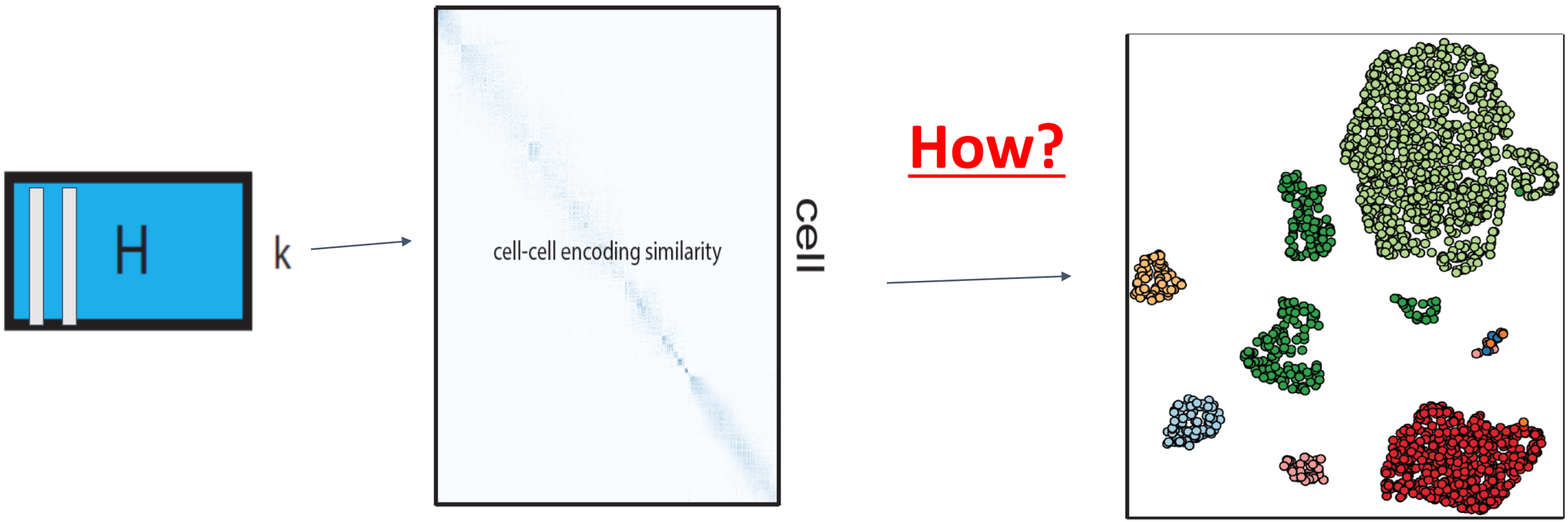
$$\mathbf{KLD}(p \parallel q) = \sum x_x p(x) \mathbf{log}\left(\frac{q(x)}{p(x)}\right)$$

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)

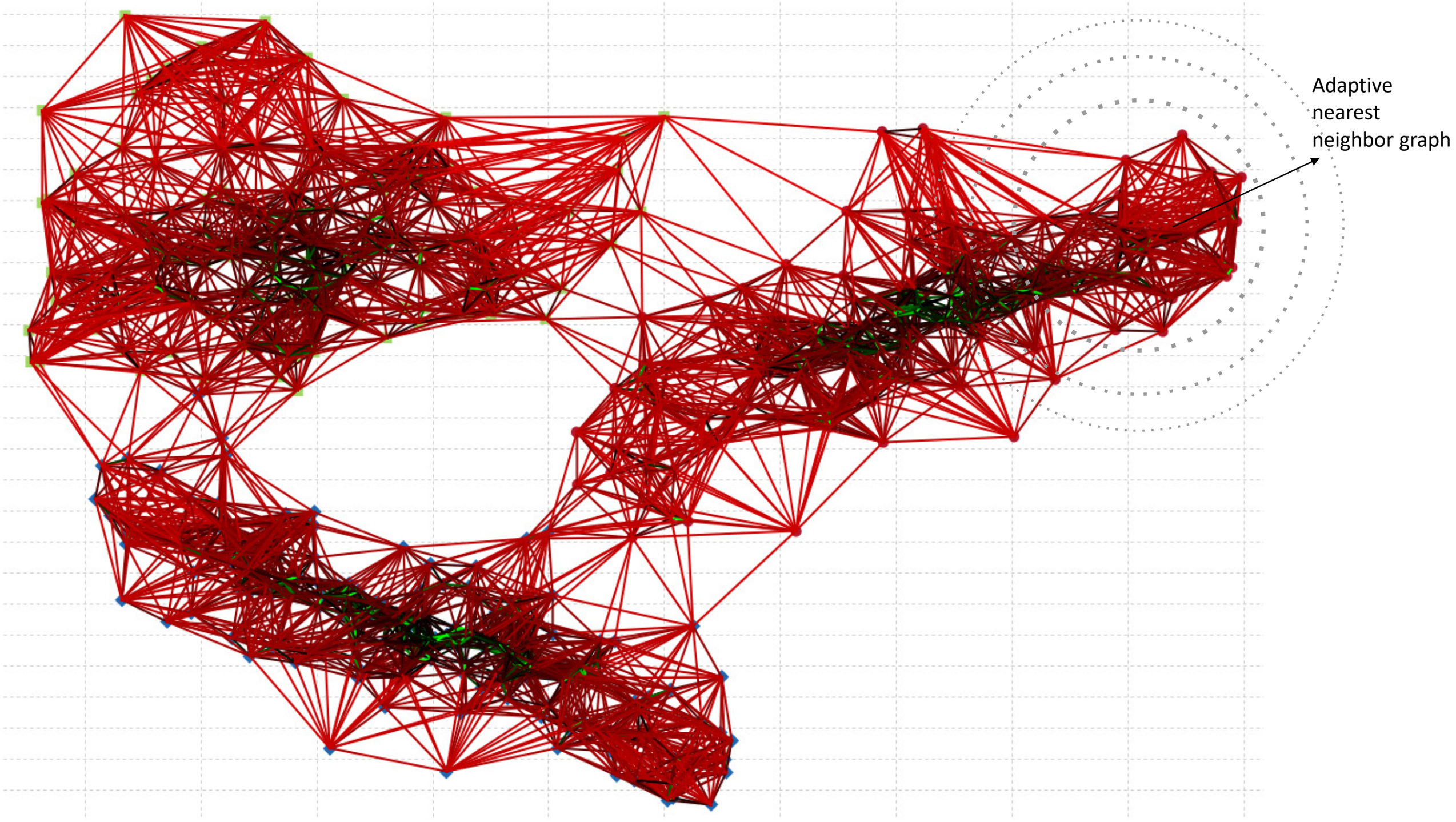
Step 2: Construct a network representation of the cell space



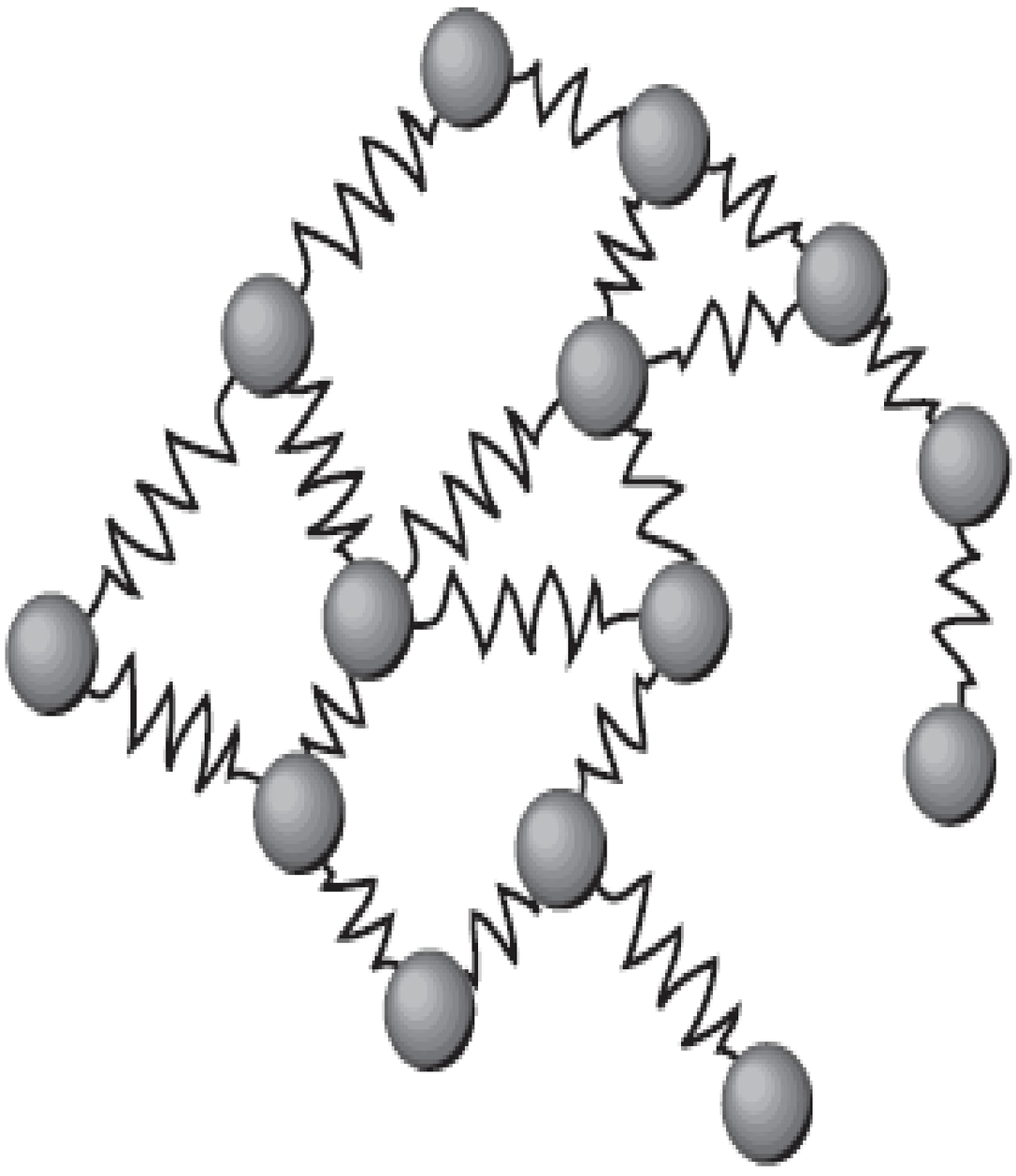
Density-dependent adaptive nearest neighbor graph

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

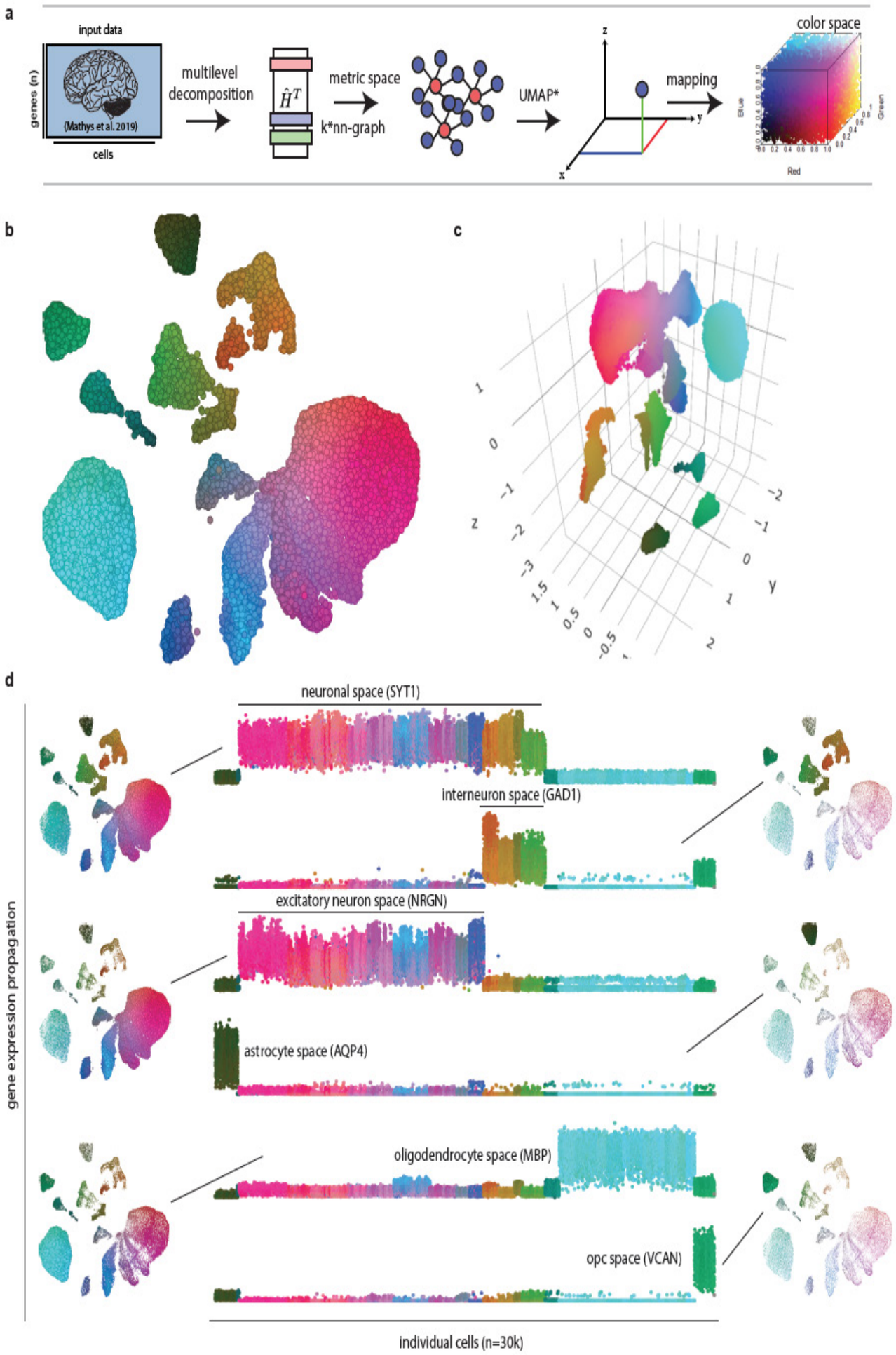
Step 2: Construct a network representation of the cell space



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

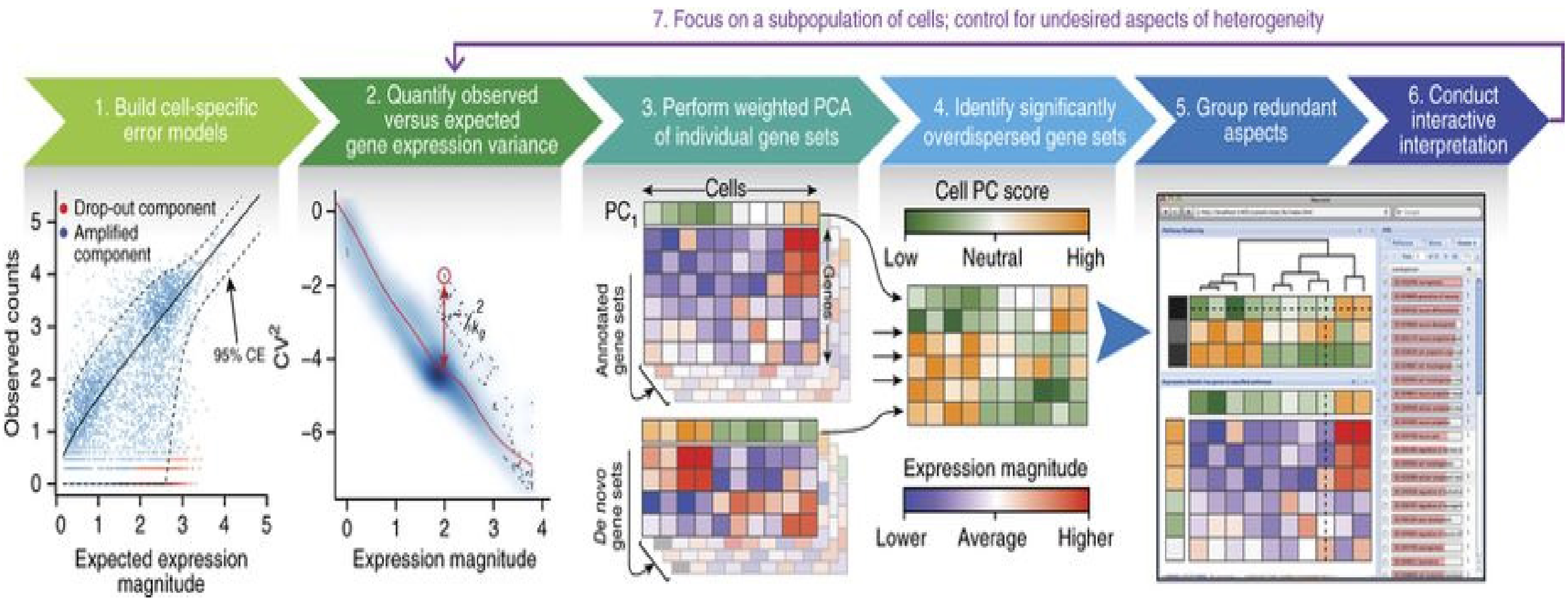


Step 4: Color-coding cells

- 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*

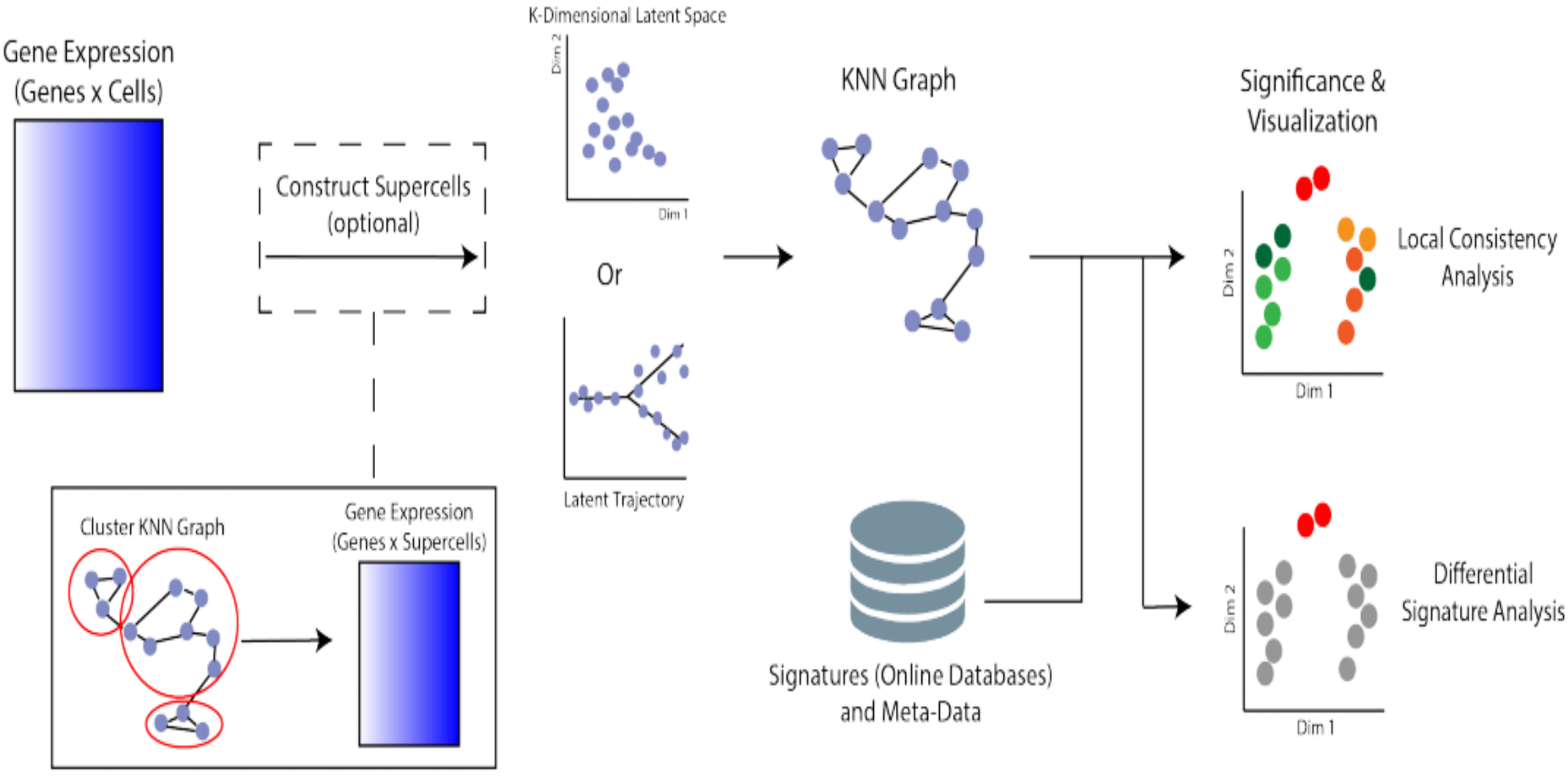
Interpreting cell-to-cell variabilities using known genesets/pathways

Pathway and gene set overdispersion analysis (PAGODA)



From: Fan *et al.,* 2016

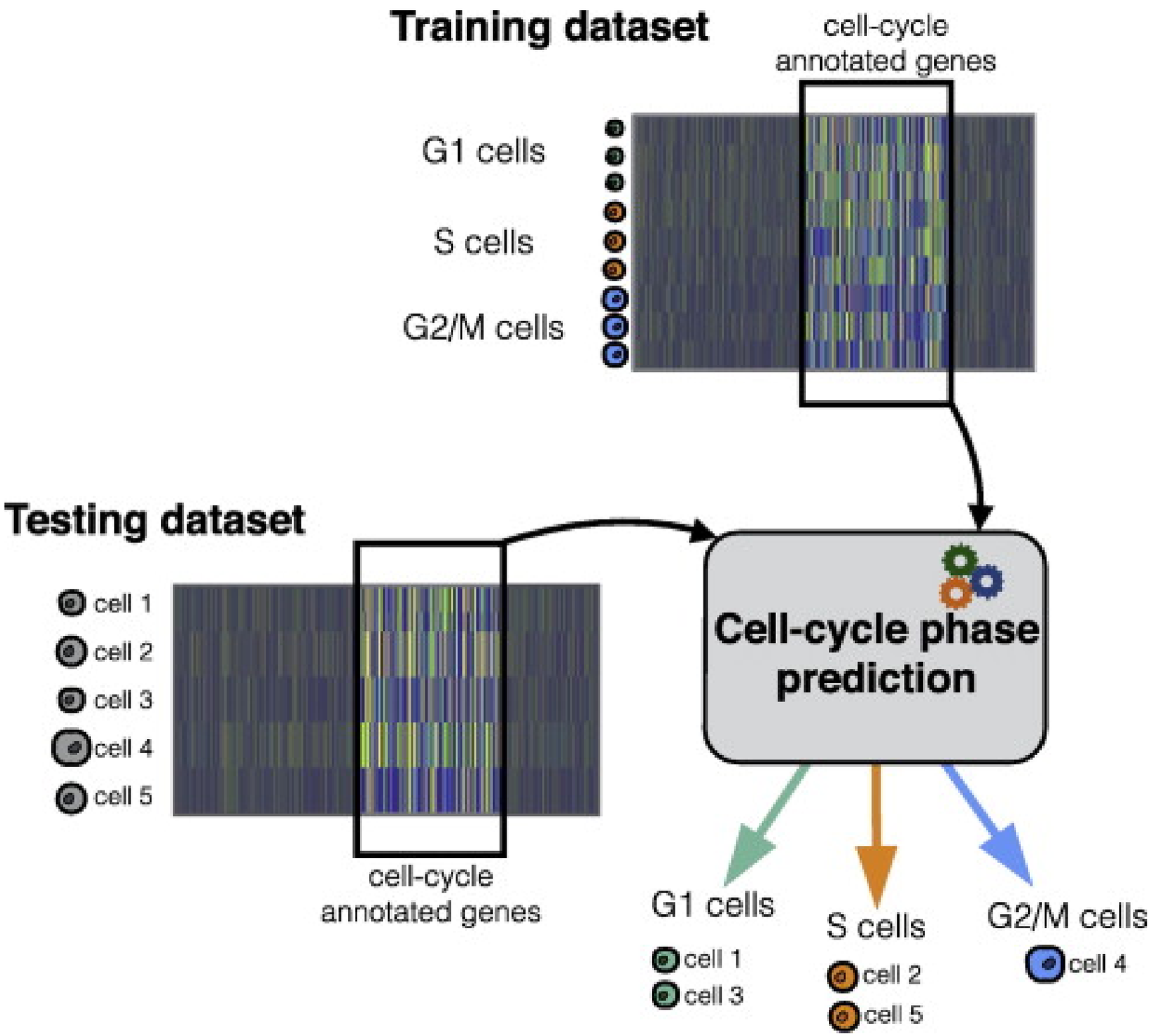
VISION method



From: DeTomaso *et al.,* 2018

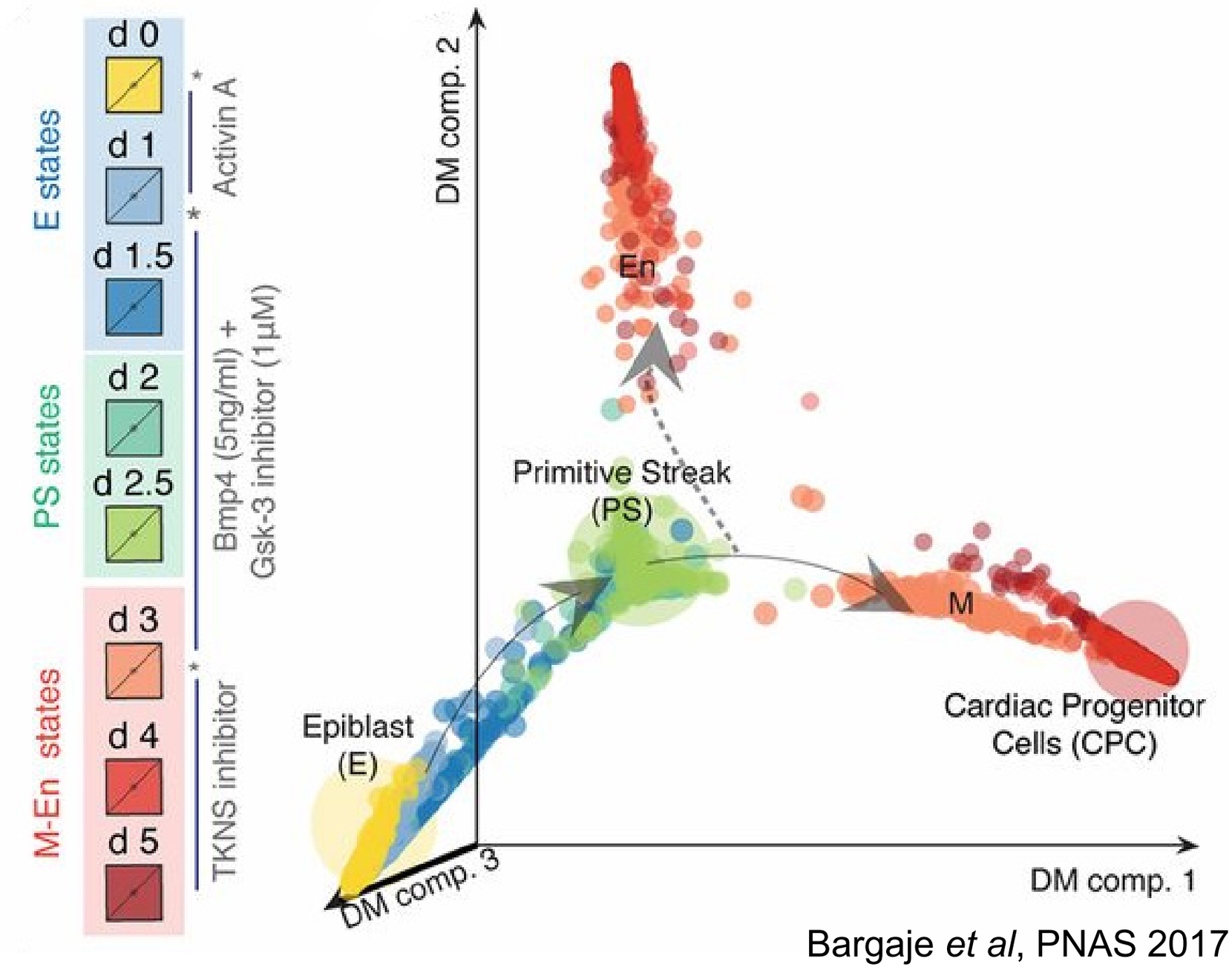
Trajectories through cell space

Cell-cycle phase prediction



Scialdone *et al.*, 2015

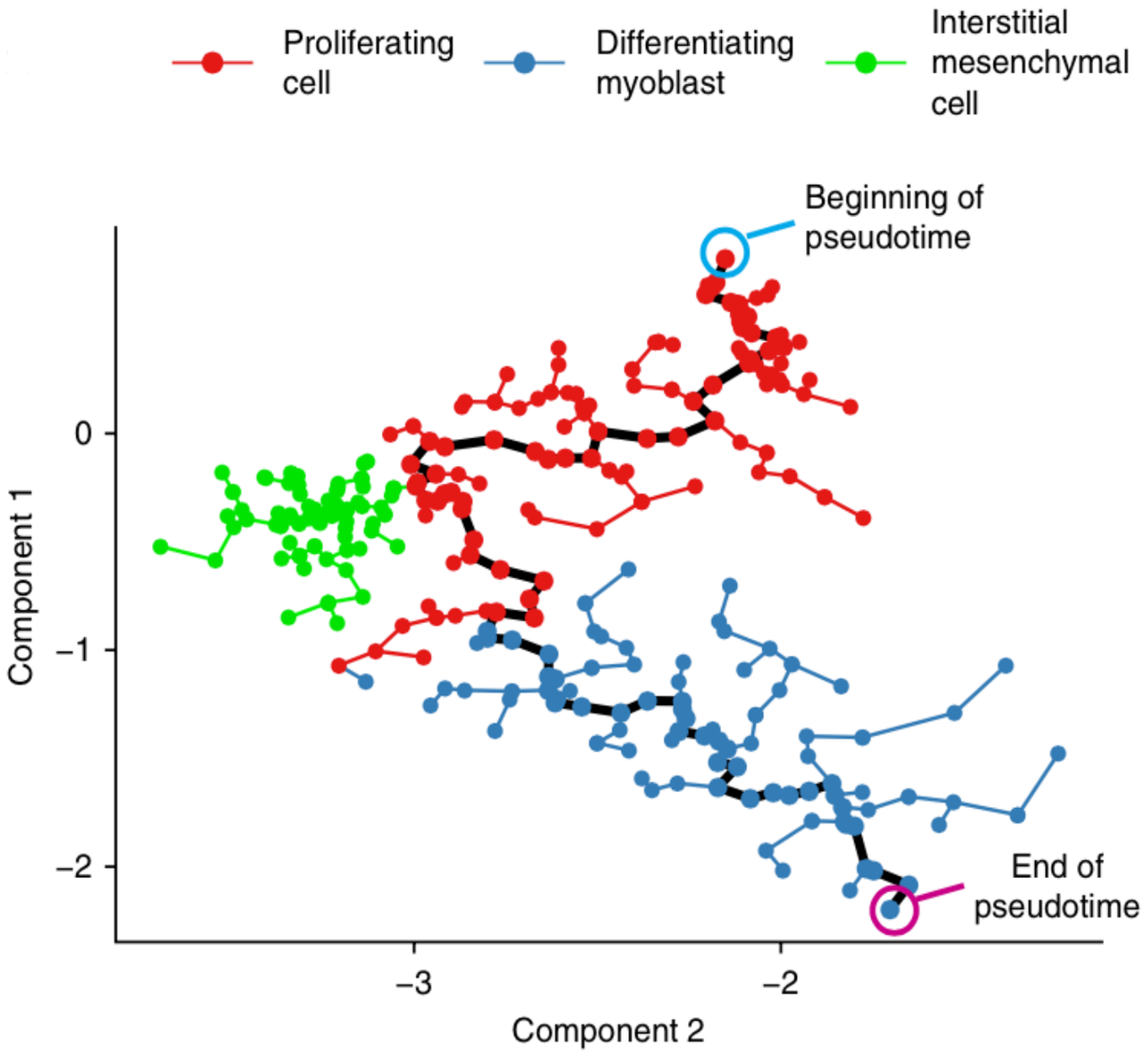
Trajectory inference



Bargaje *et al*, PNAS 2017

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

Trajectory inference methods

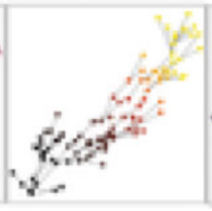
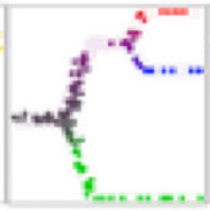
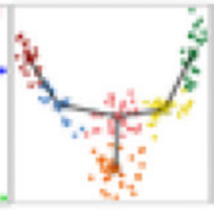
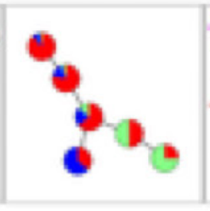
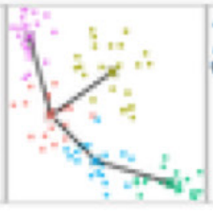
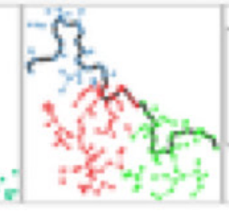
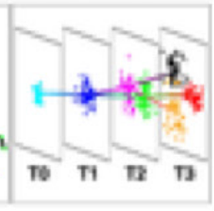


TSCAN pseudotime reconstruction with monocle

- 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

Ji *et al*, NAR 2016

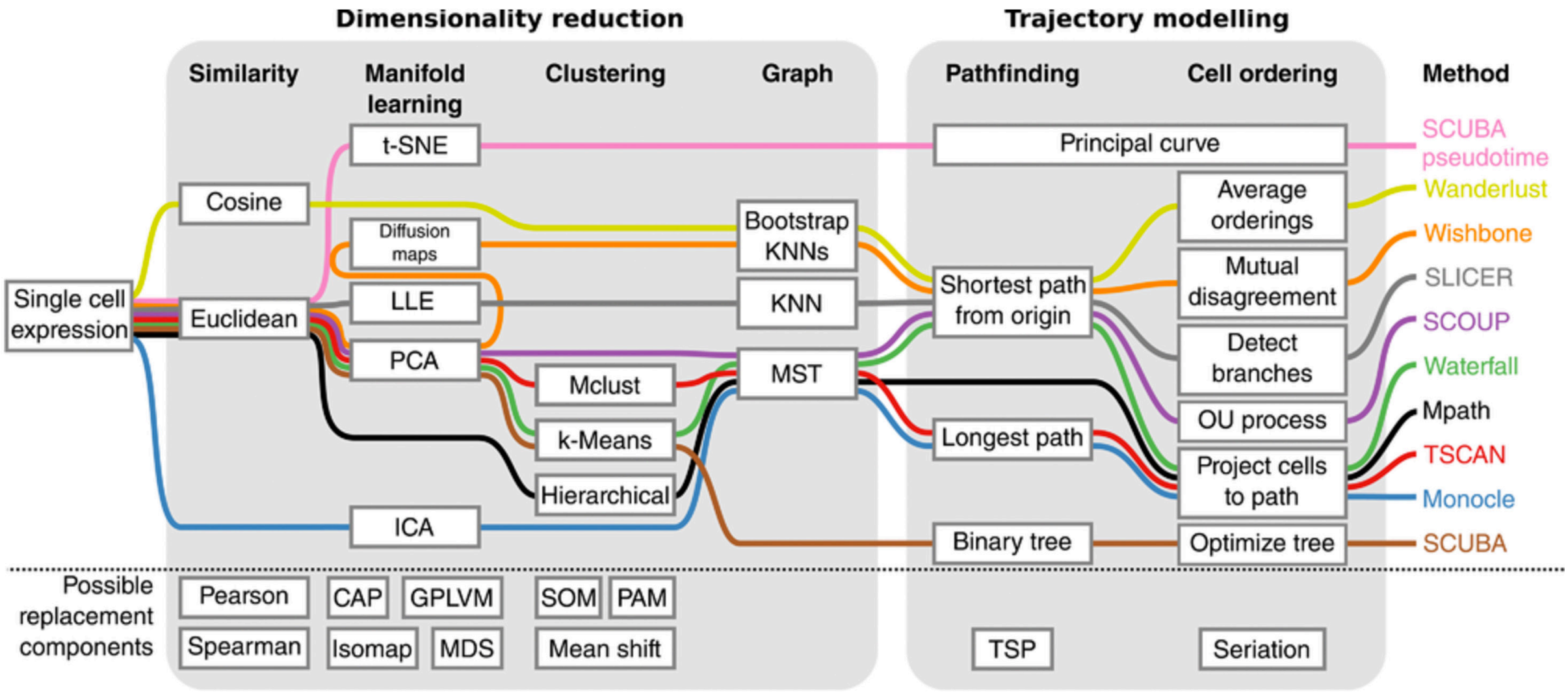
Overview of trajectory identification methods

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

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

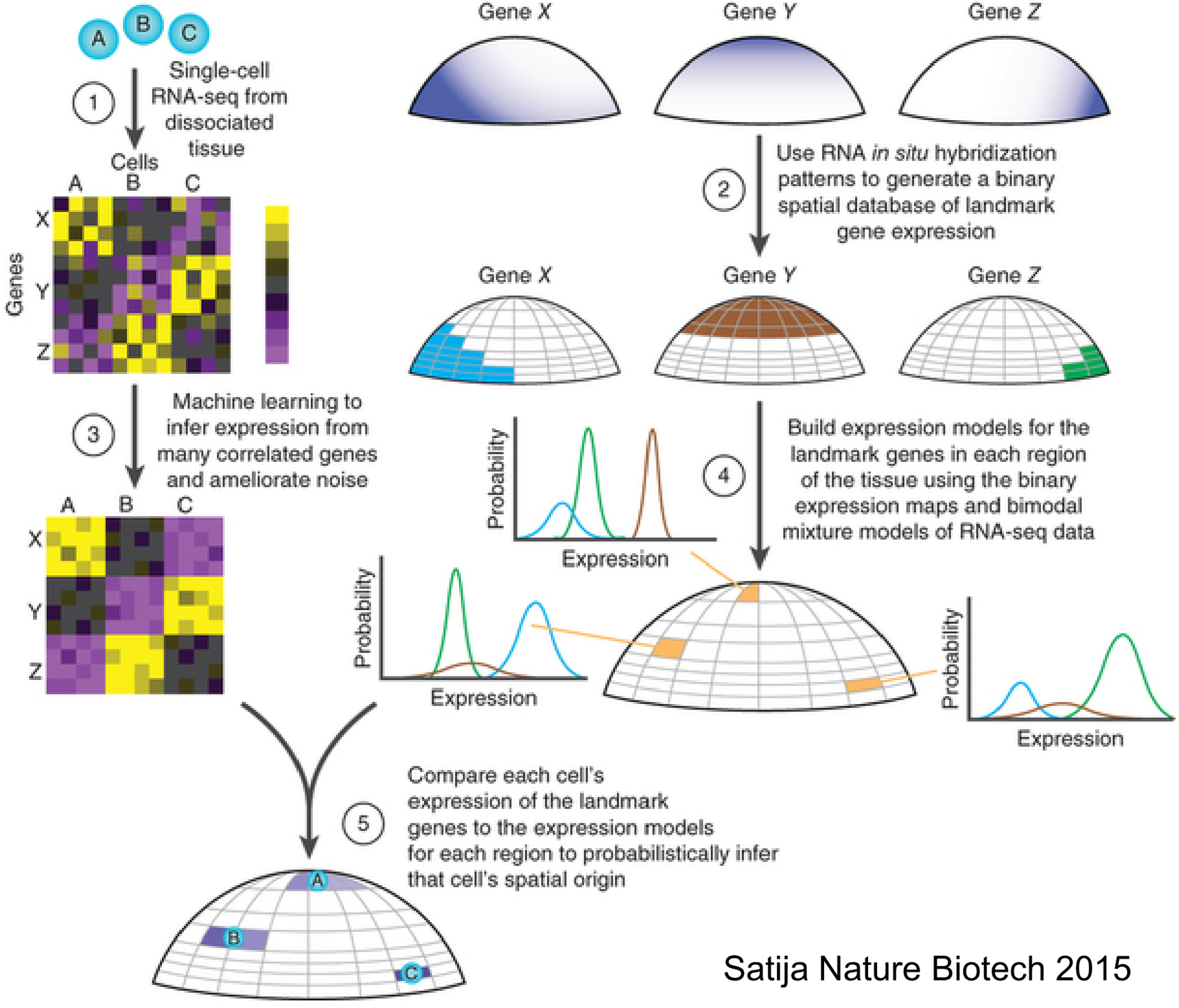
Trajectory identification: meta-method view



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

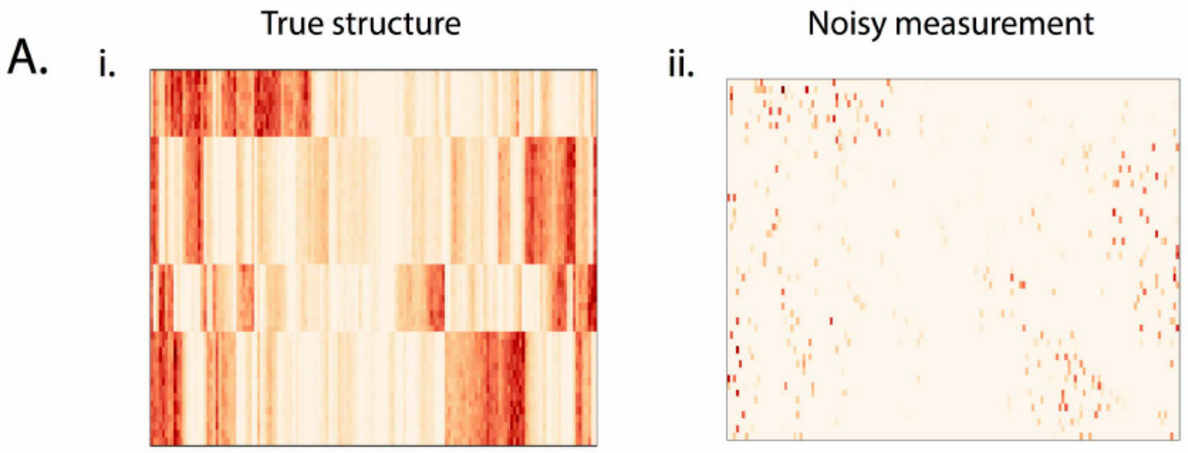
Dataset completion & missing data imputation

Spatial reconstruction of single-cell gene expression



Satija Nature Biotech 2015

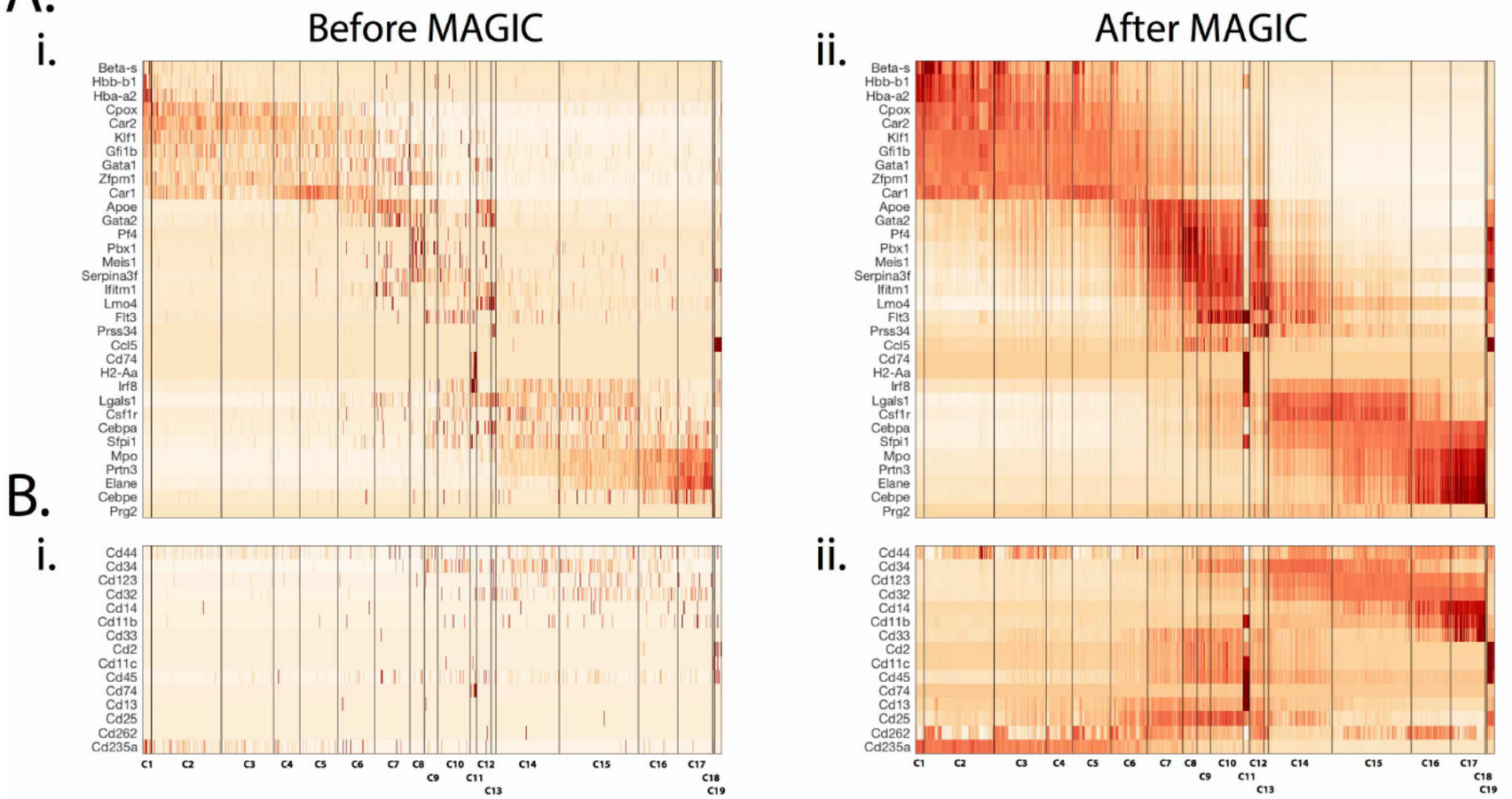
Missing value imputation with MAGIC (Markov Affinity-based Graph Imputation of Cells)

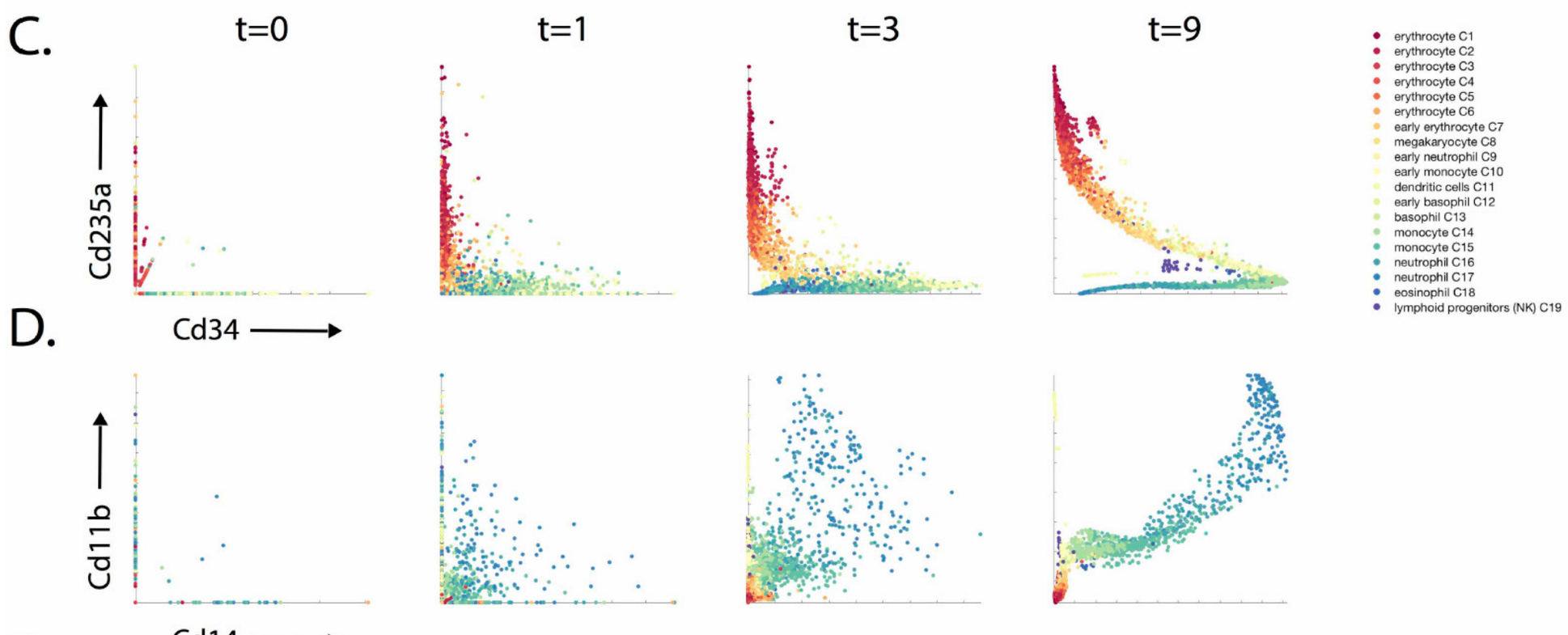


Random walk on cell-cell similarity graph

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

Imputation reveals gene-gene correlation patterns





A: 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]	*k*-means clustering of PCs of correlation matrix
knn-smooth	[55]	*k*-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 factorizatio

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 *k*-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)

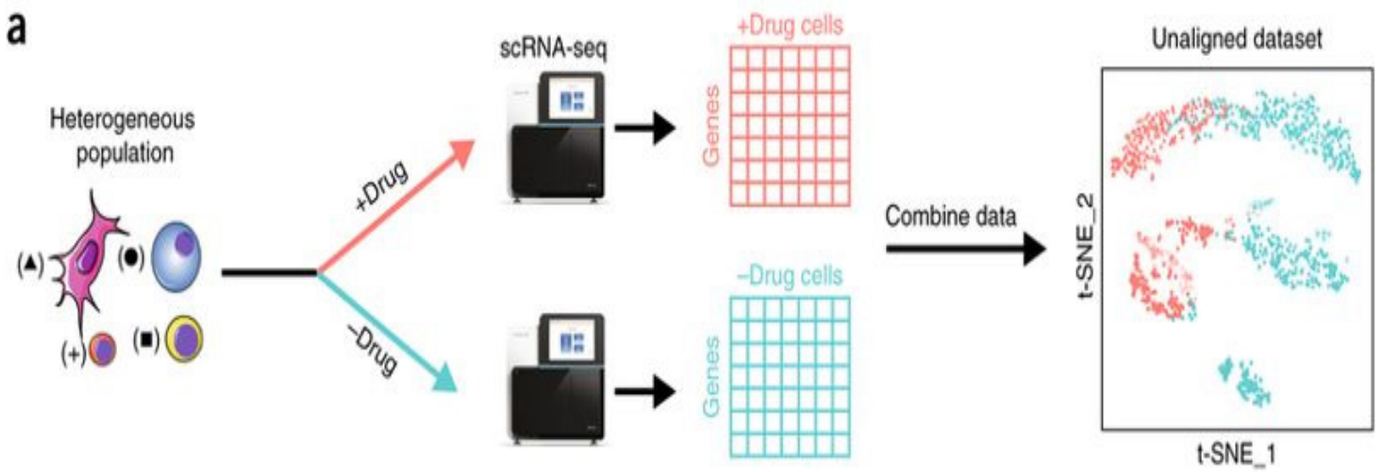
T: using external information

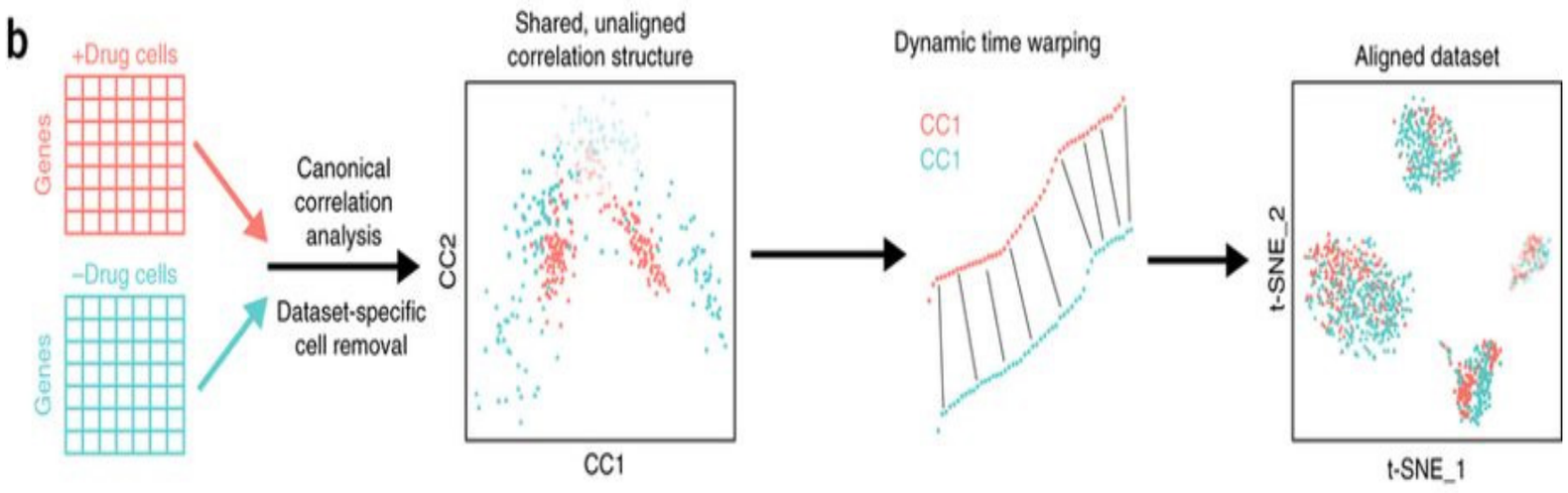
ADImpute	[86]	Gene regulatory network information
netSmooth	[58]	PPI network information
SAVER-X	[87]	Transfer learning with atlas-type resources
SCRABBLE	[88]	Matched bulk RNA-seq data
TRANSLATE	[77]	Transfer learning with atlas-type resources
URSM	[89]	Matched bulk RNA-seq data

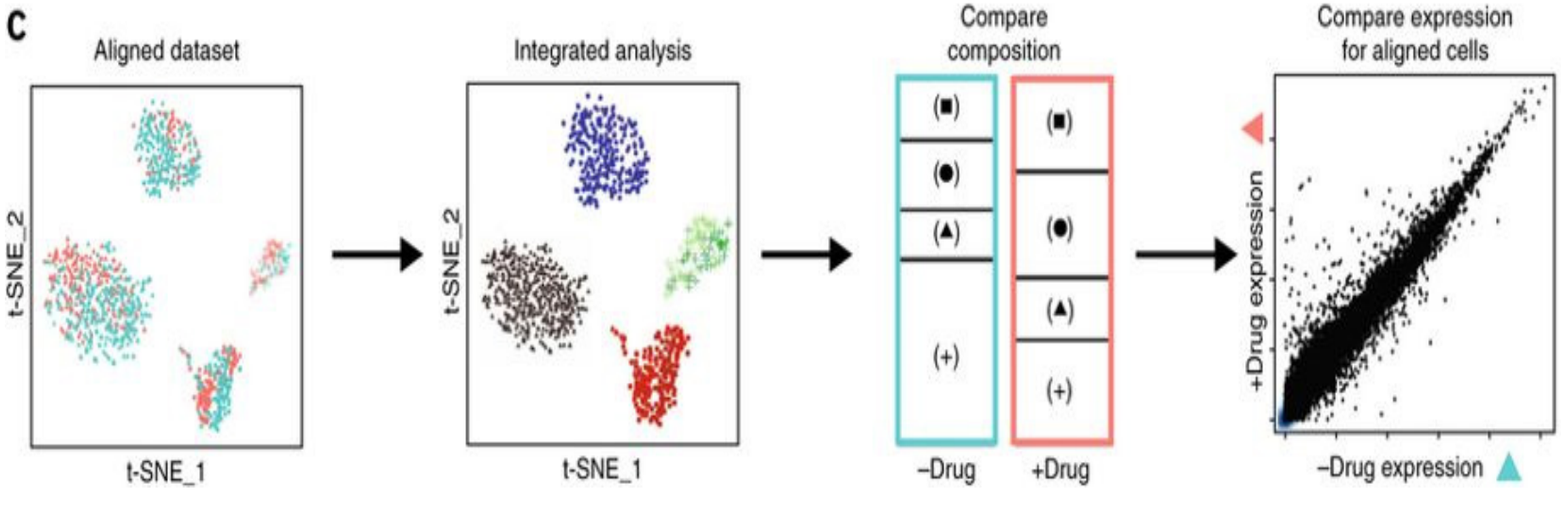
47. Tang W, Bertaux F, Thomas P, Stefanelli C, Saint M, et al. bayNorm: Bayesian gene expression recovery, imputation and normalisation for single cell RNA-sequencing data. bioRxiv. 2018. https://www.biorxiv.org/content/10.1101/384586v2.abstract.
48. Azizi E, Prabhakaran S, Carr A, Pe'er D. Bayesian 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 JI, 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 WV, 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, Măndoiu II. 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, Bierie 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, Yunits 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)

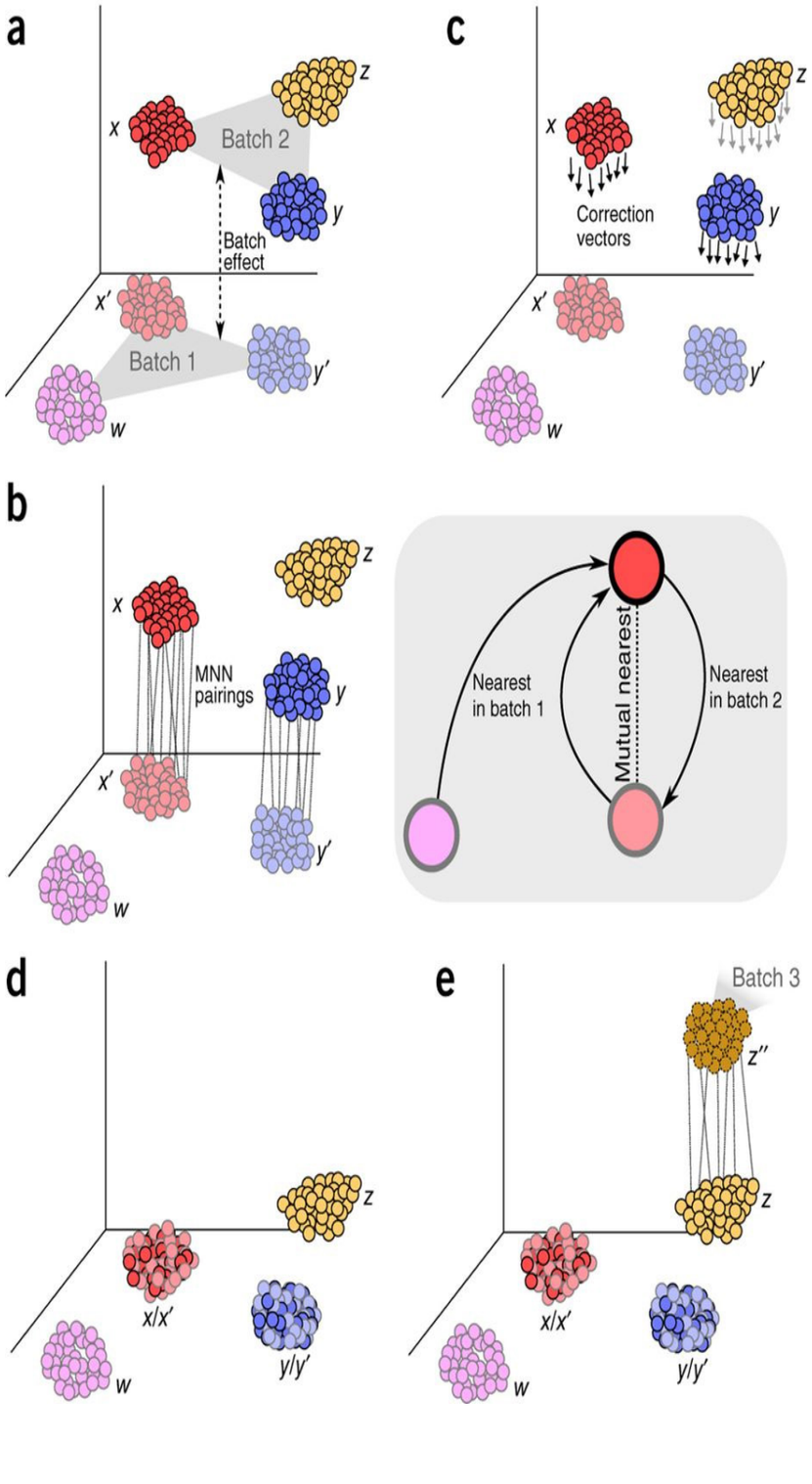






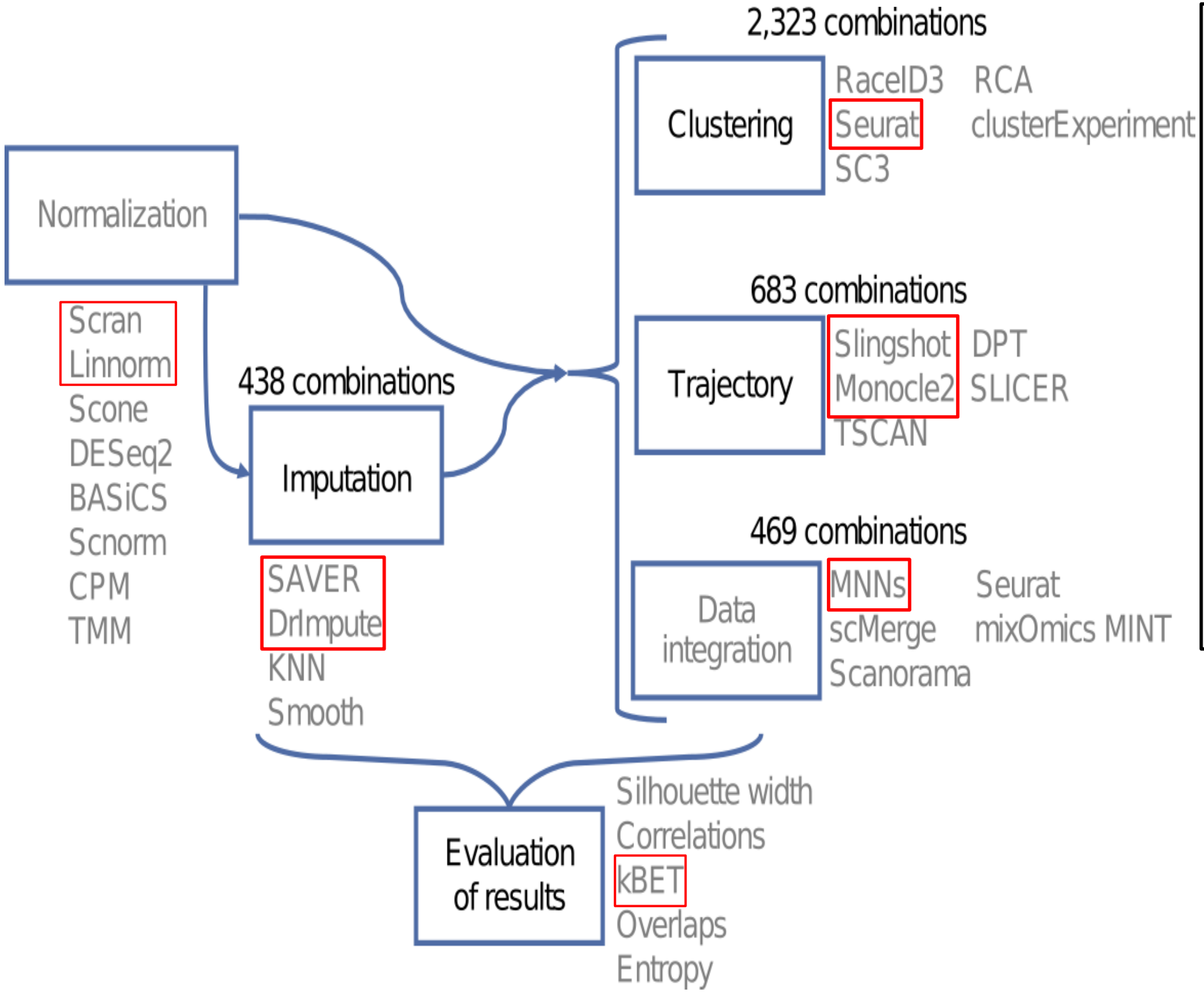
From: Butlet *et al.,* 2018

Mutual nearest neighbors (MNN) correction



From: Haghverdi *et al., 2018*

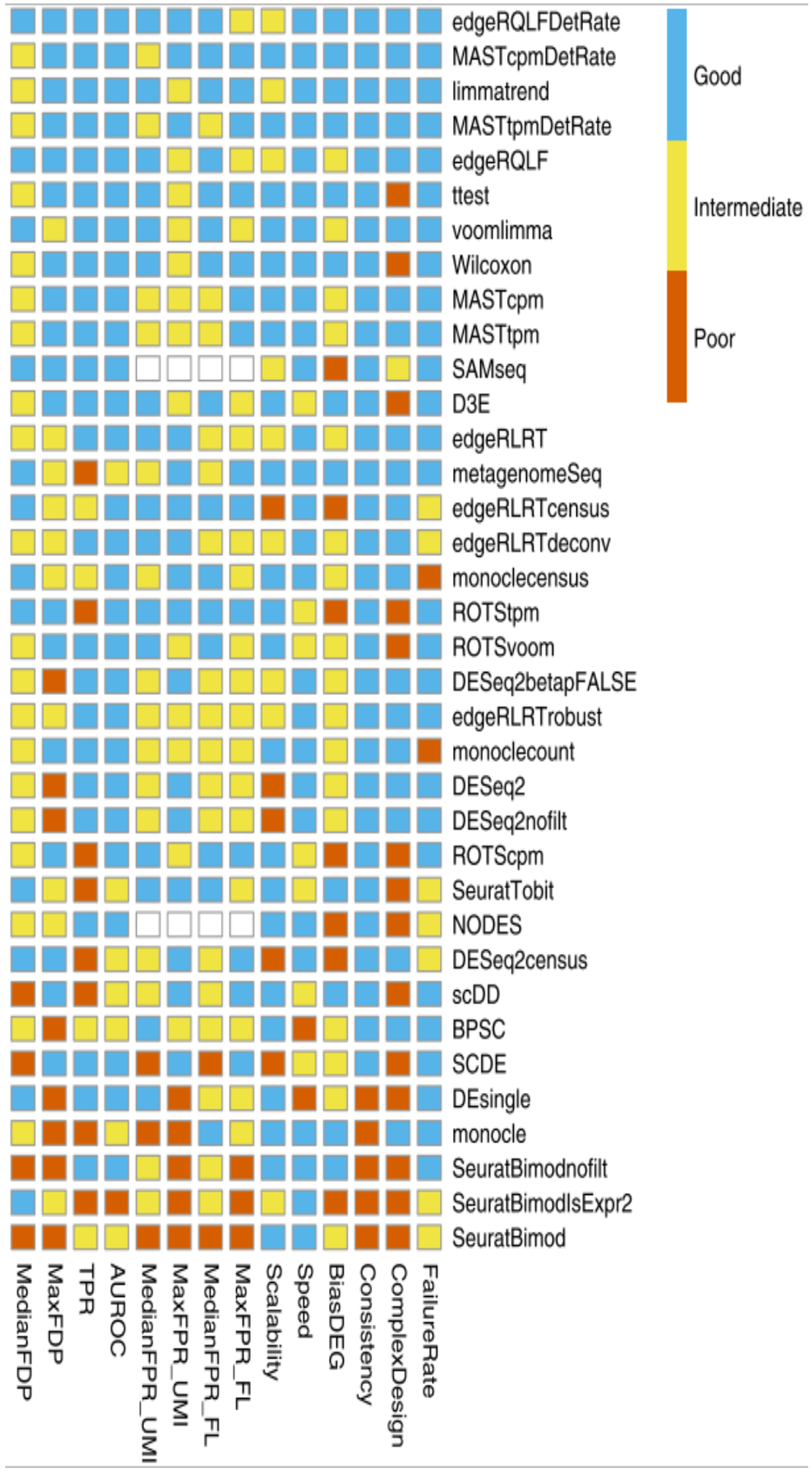
Summary and Method comparison



Recent Methods

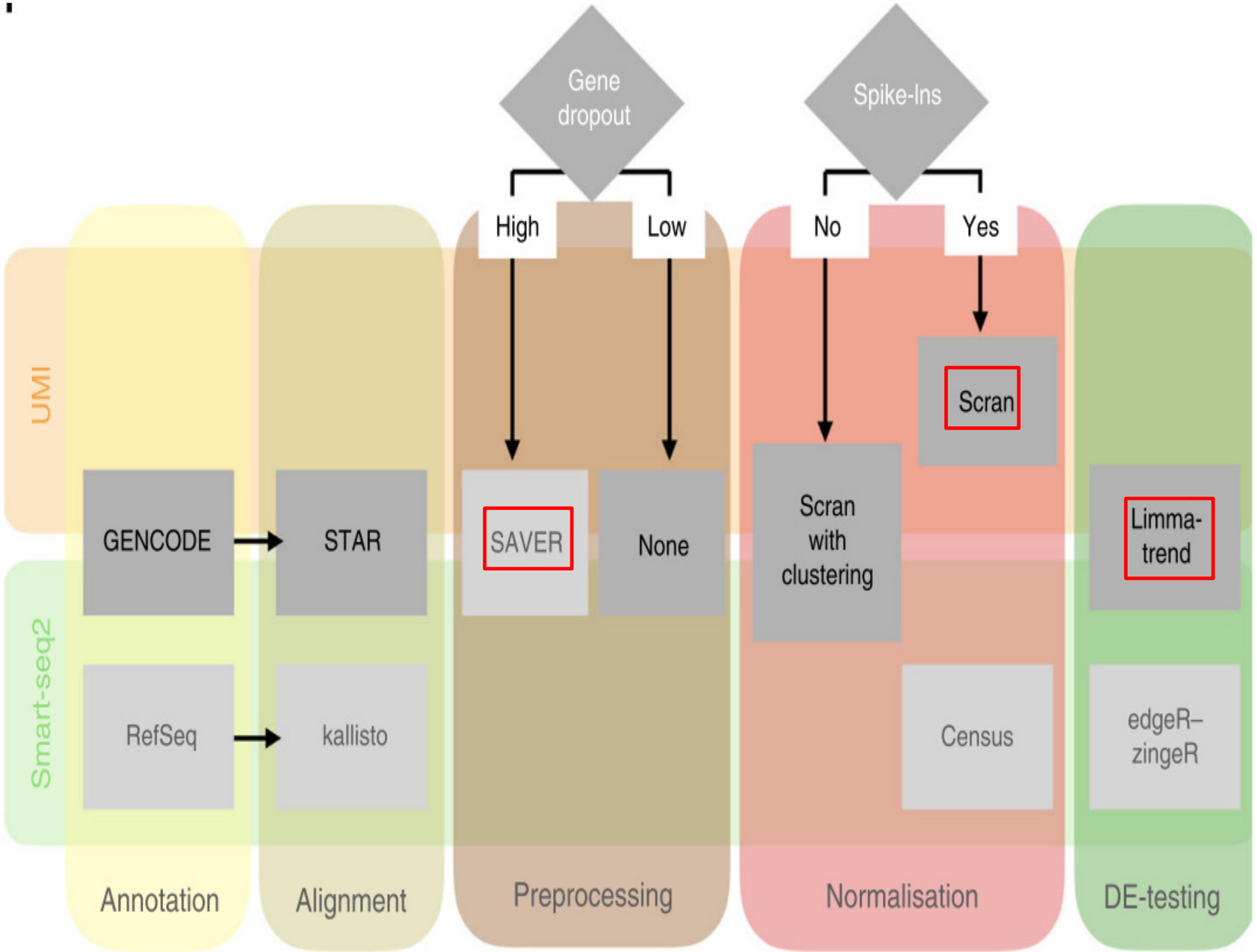
Tian, L., Dong, X., Freytag, S., Lê Cao, K.-A., Su, S., JalalAbadi, A., … Ritchie, M. E. (2019). Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. Nature Methods, 16(6), 479–487.

Comparison of differential expression methods



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

Soneson, C., & Robinson, M. D. (2018). Bias, robustness and scalability in single-cell differential expression analysis. Nature Methods, 15(4), 255–261.



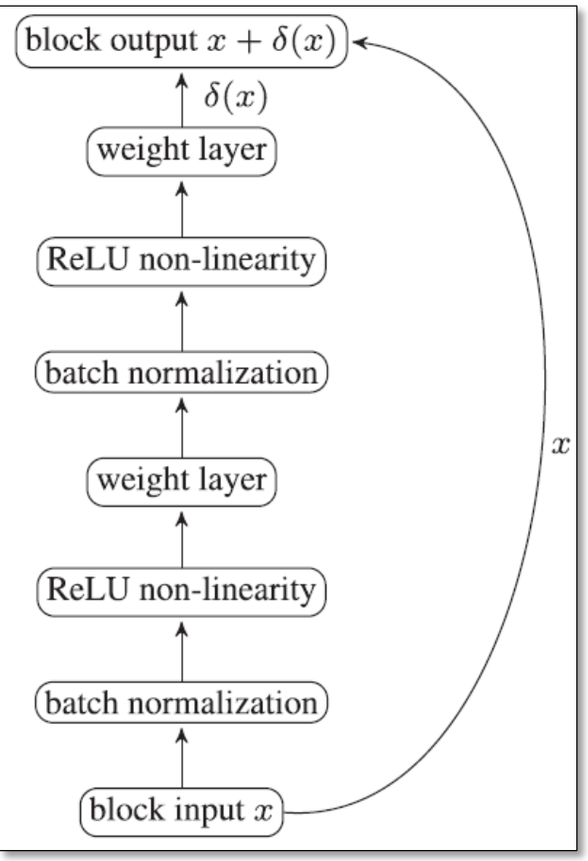
Vieth, B., Parekh, S., Ziegenhain, C., Enard, W., & Hellmann, I. (2019). A systematic evaluation of single cell RNA-seq analysis pipelines. Nature Communications, 10(1), 4667.

Summary

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

6. Deep Learning methods for scRNA-seq

MMD-ResNet: Autoencoder for batch correction



Autoencoder ResNet arch.

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

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

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

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

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

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

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

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

Maximum Mean Discrepancy (MMD) loss function

$$L(w) = \sqrt{\mathrm{MMD}^2(\{\widehat{\psi}(x_1), \ldots \widehat{\psi}(x_n)\}, \{y_1, \ldots y_m\})},$$

Train ResNet with loss MMD score function

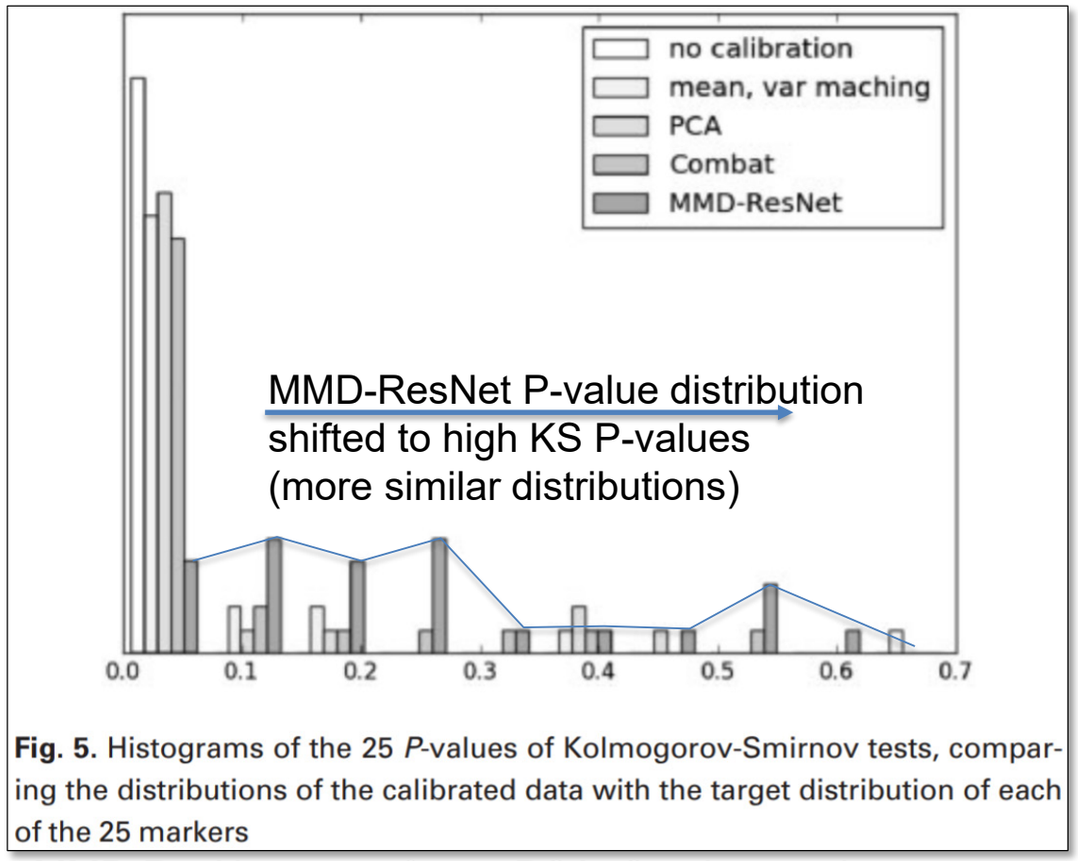


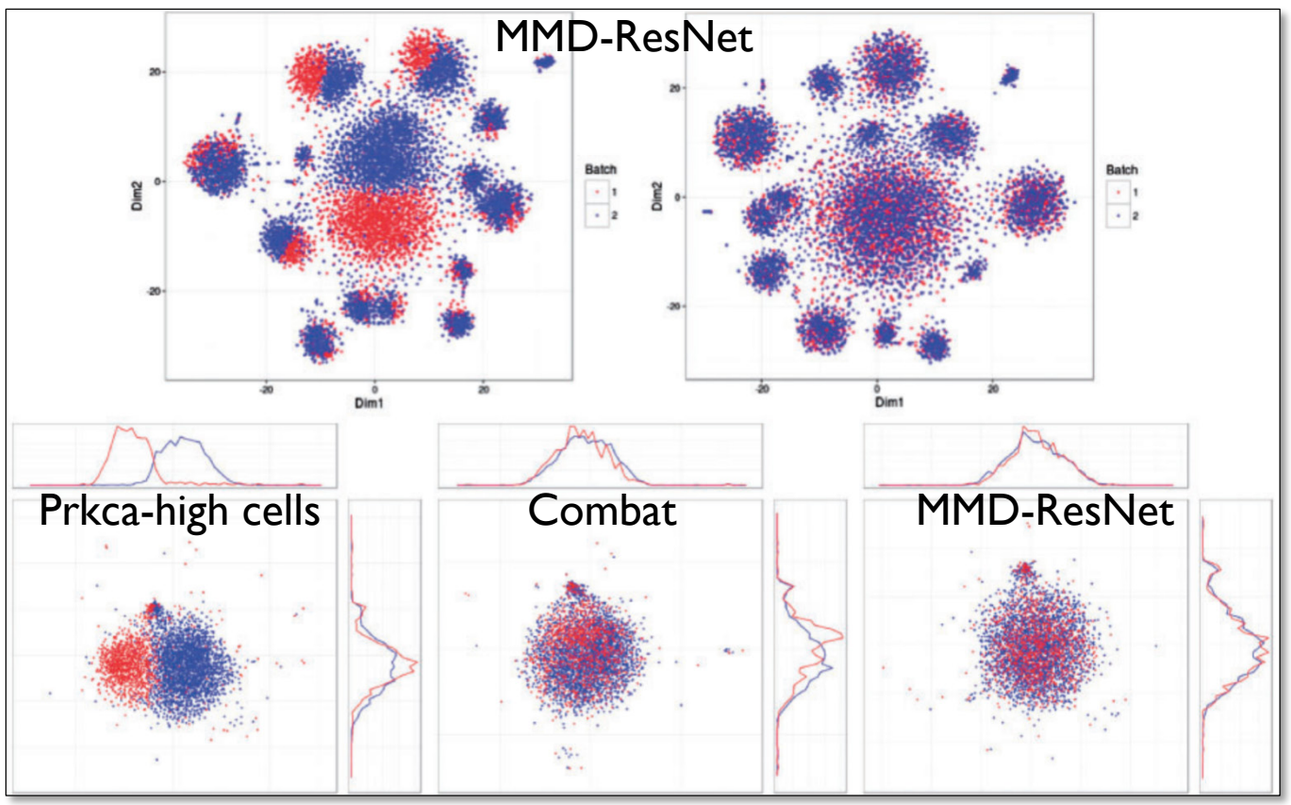
Fig. 5. Histograms of the 25 *P*-values of Kolmogorov-Smirnov tests, comparing the distributions of the calibrated data with the target distribution of each of the 25 markers

MMD-ResNet outperforms PCA, Combat, and

Removal of batch effects using distribution-matching residual networks

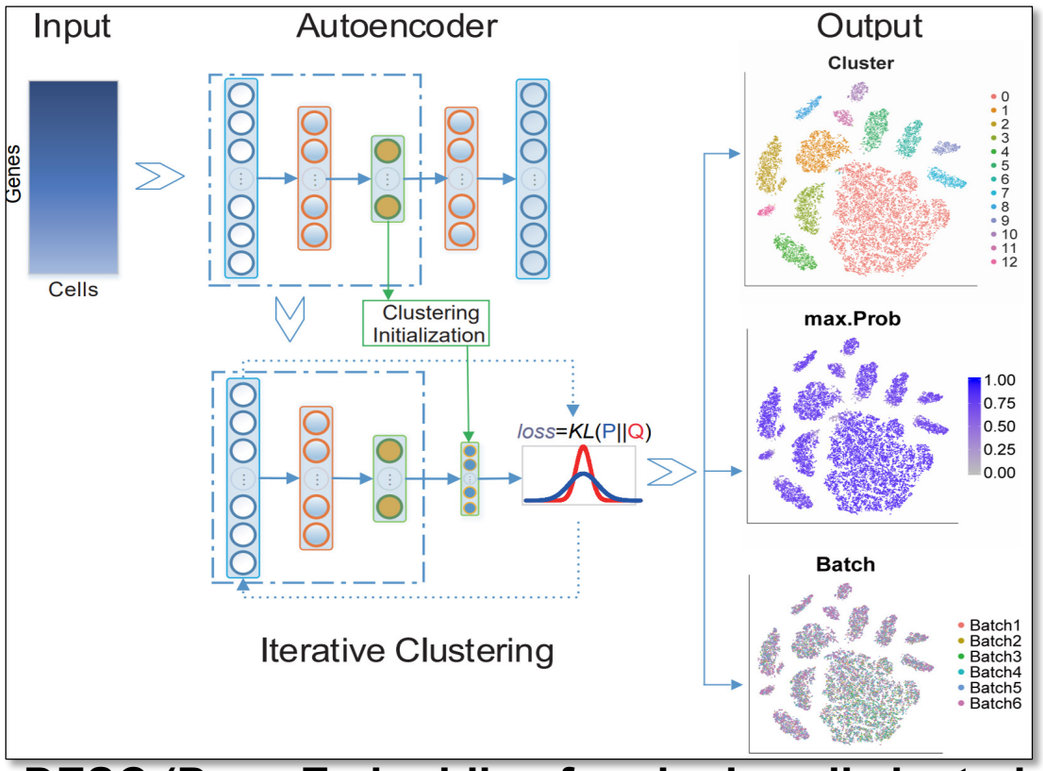
Uri Shaham[1,†], Kelly P. Stanton[2,3,†], Jun Zhao[3], Huamin Li[4], Khadir Raddassi[5], Ruth Montgomery[6] and Yuval Kluger[2,3,4,*]

Shaham et al., Bioinf, 2017



t-SNE plots before (left) and after (right) calibration

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

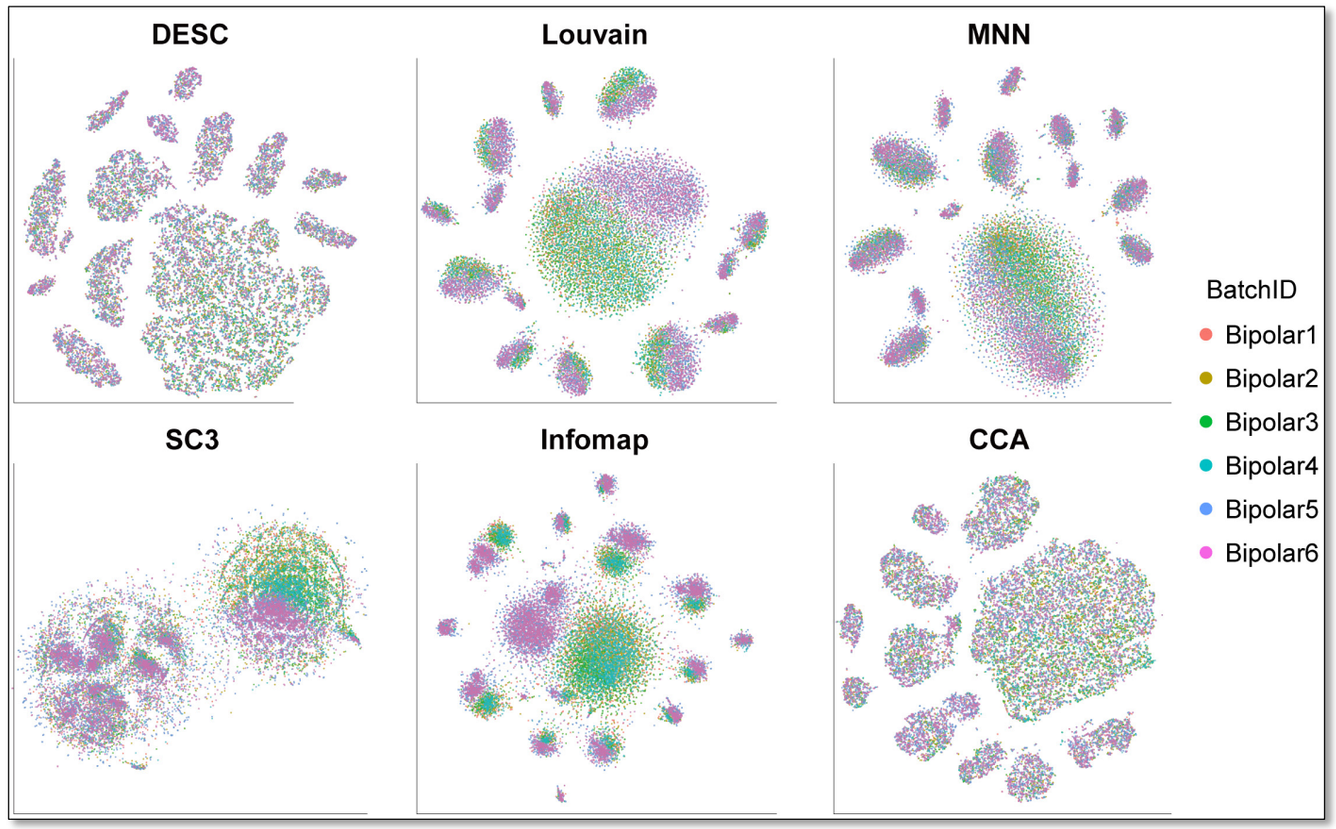


DESC (Deep Embedding for single-cell clustering):

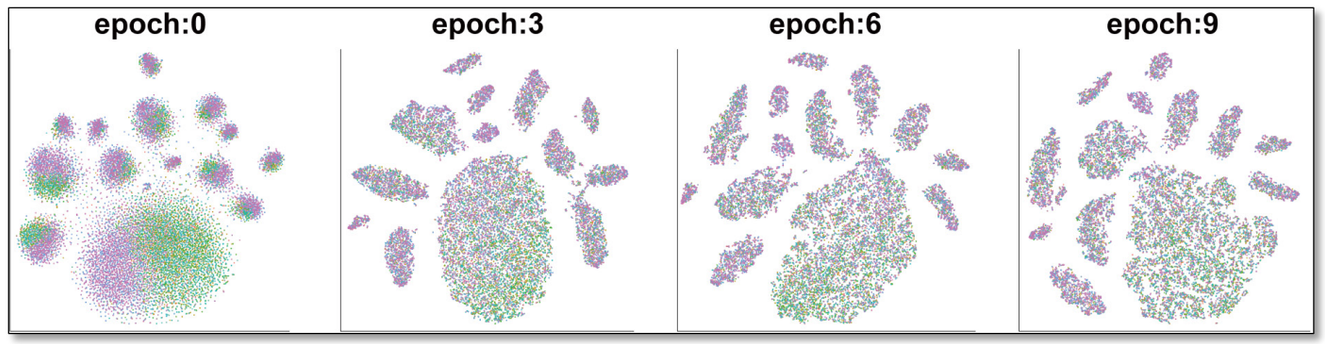
- Stacked auto-encoder learns cluster-specific gene expression representation and cluster assignments for scRNA-seq data clustering
- Initialize clustering obtained from autoencoder
- Learn non-linear mapping from original space to a low-dimensional space
- iteratively optimize clustering objective function
  - 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

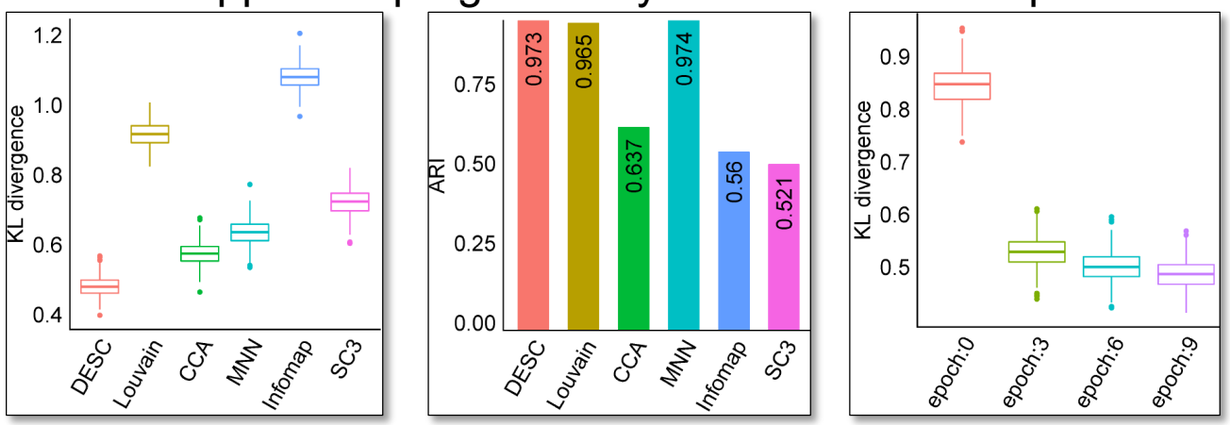
Xiangjie Li, Yafei Lyu, Jihwan Park, Jingxiao Zhang, Dwight Stambolian, Katalin Susztak, Gang Hu, Mingyao Li
doi: https://doi.org/10.1101/530378



DESC avoids cluster-specific batch effects found in other methods

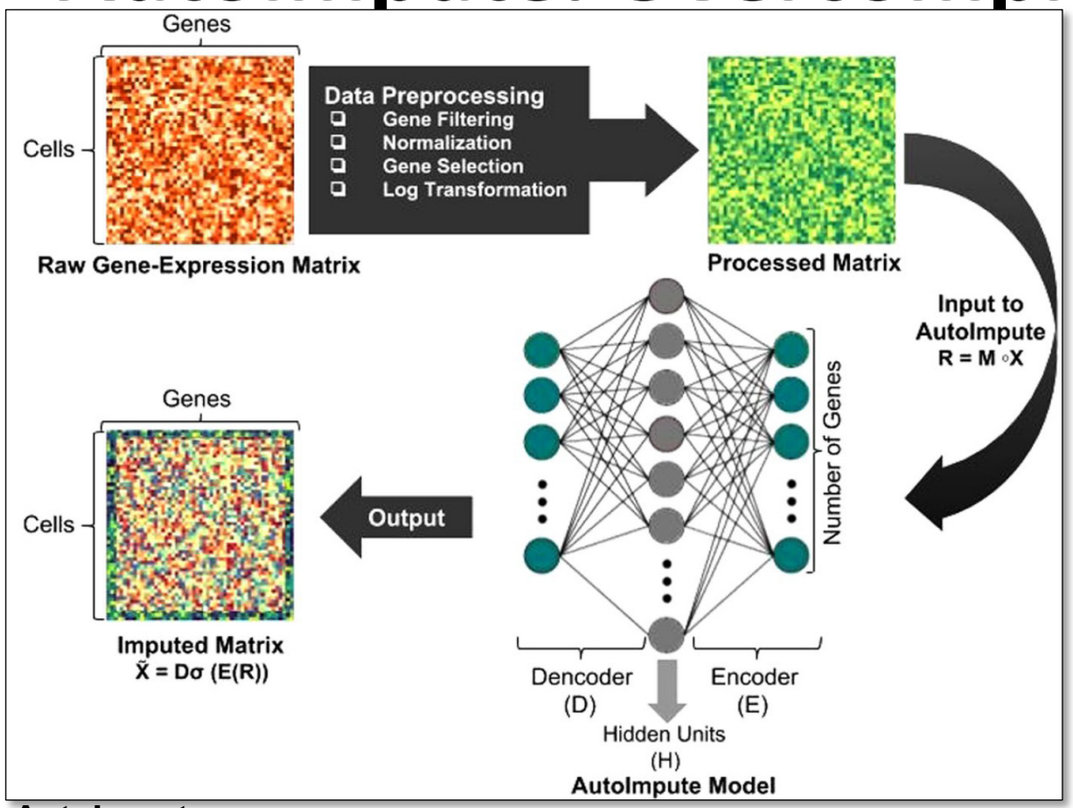


Iterative approach progressively removes cluster-specific batch effects



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

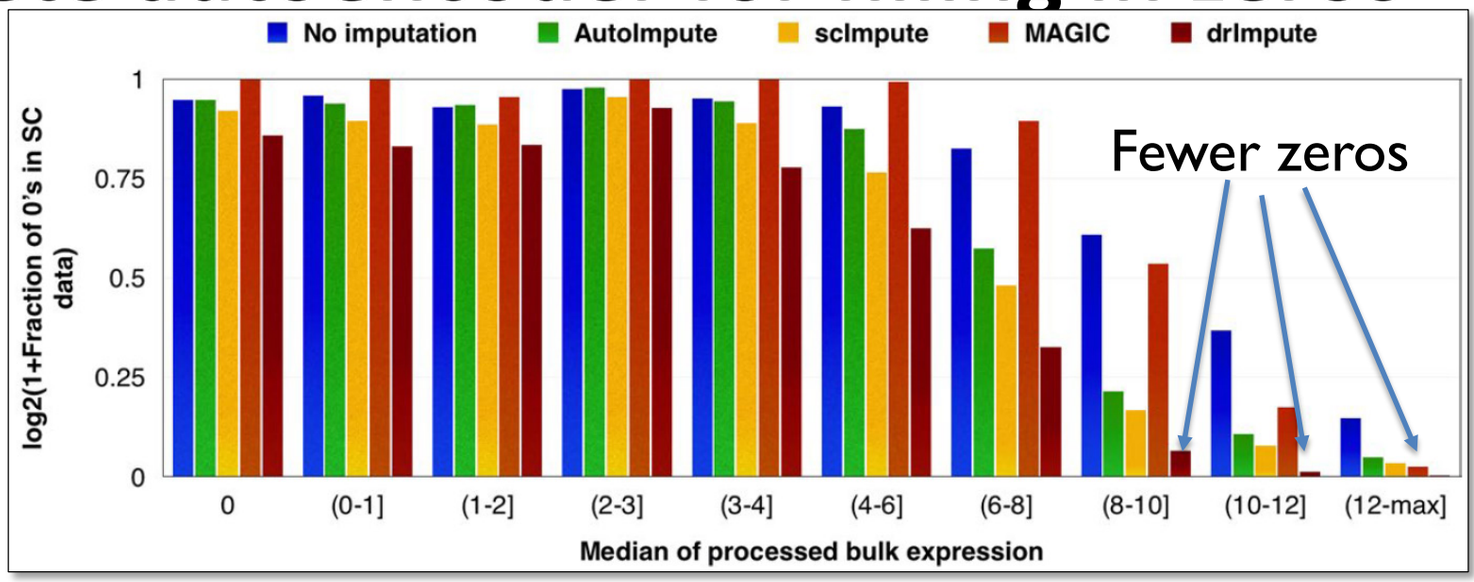
- Feeding sparse gene expression matrix as input to autoencoder
- train it to learn the encoder and decoder functions that best regenerate imputed expression with no dropouts
- back-propagating errors only for non-zero counts in sparse matrix

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.

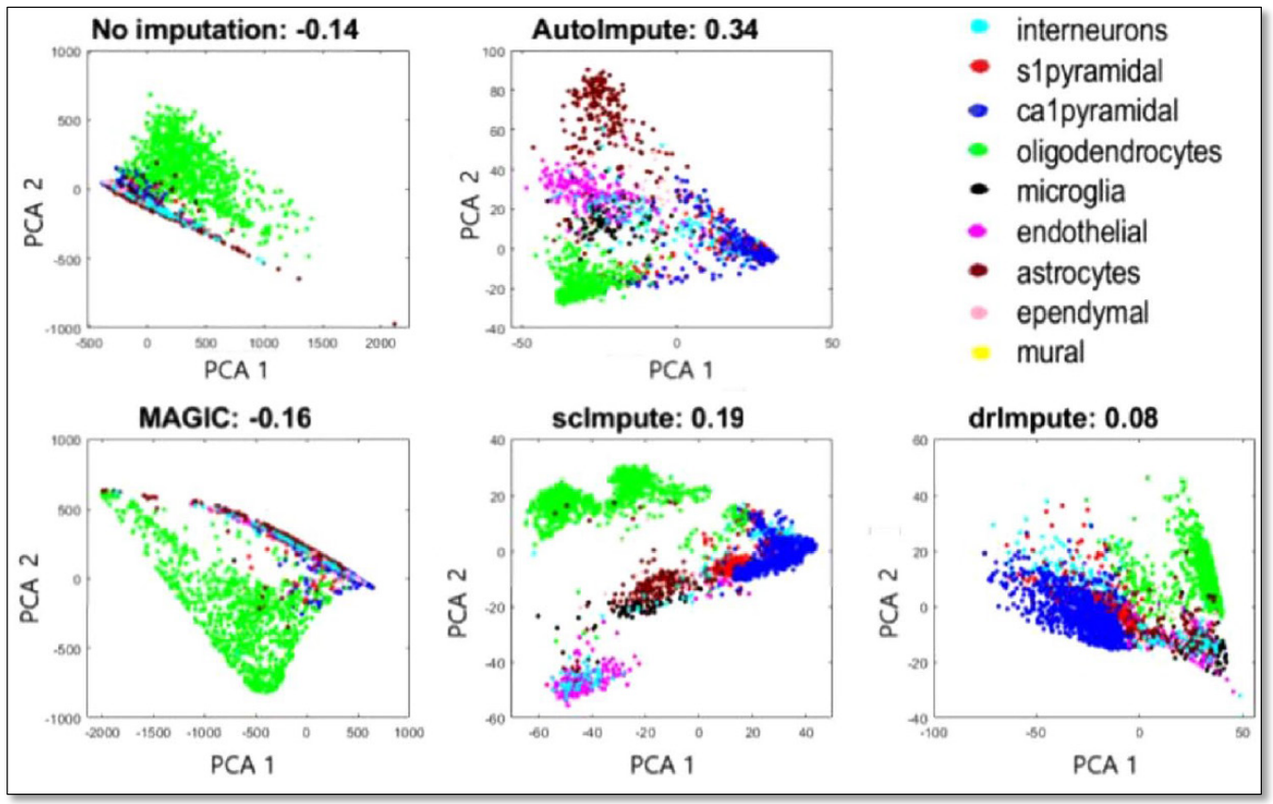
The hyper-parameter selection was done after doing an extensive grid search on the following hyper-parameters:

- 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

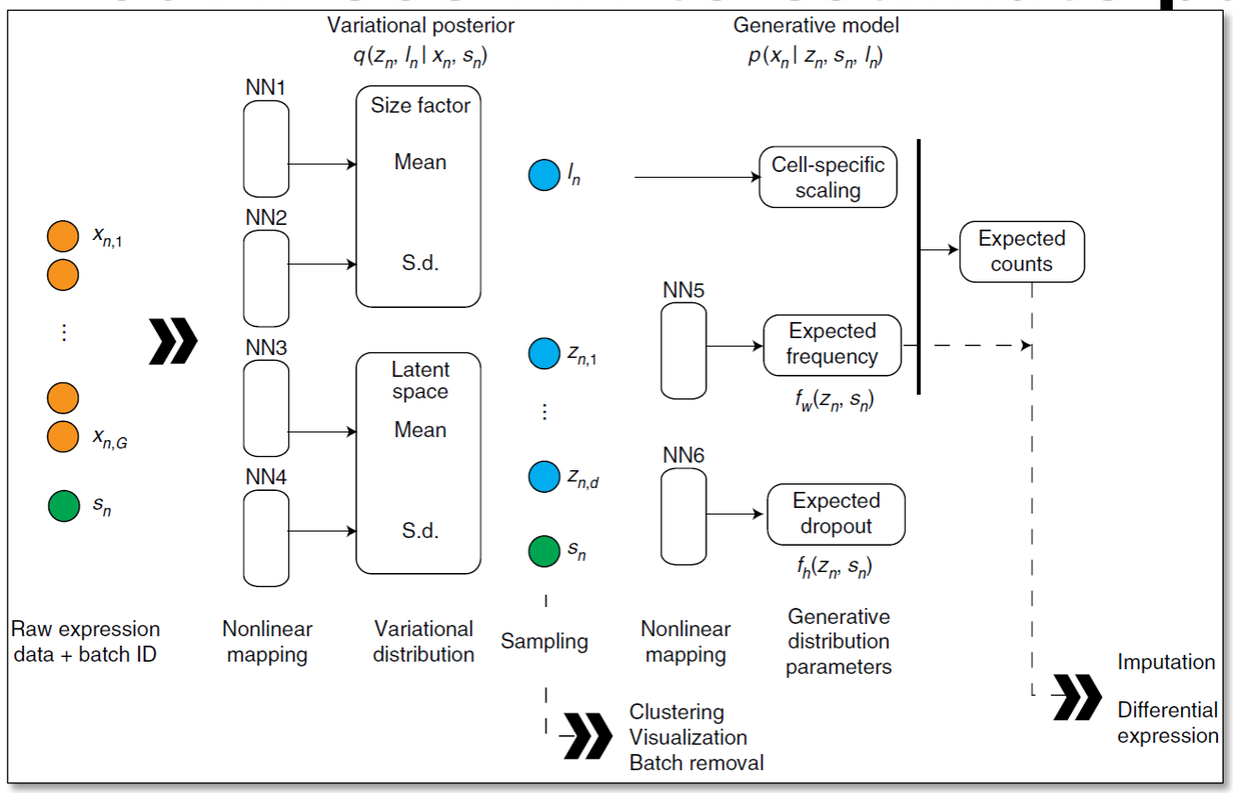


Iterative approach progressively removes cluster-specific batch effects

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

Divyanshu Talwar[1], Aanchal Mongia[1], Debarka Sengupta[1,3] & Angshul Majumdar[2]

scVI: Use NN to estimate params in variational inference



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

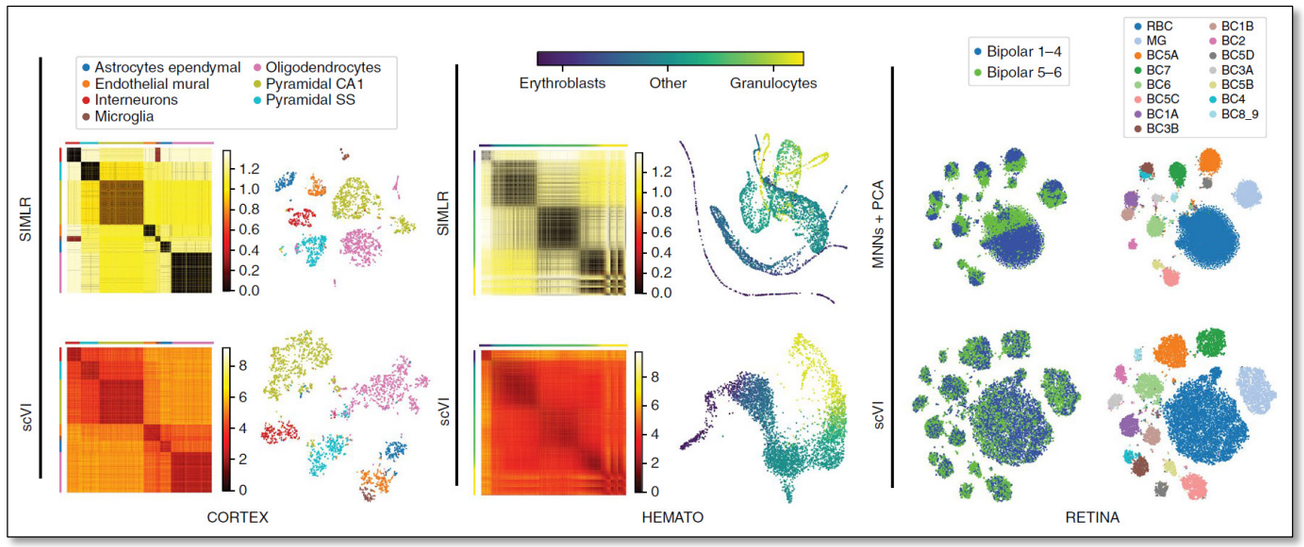
Modeled observed expression x_{ng} (gene g, cell n) as sample
Drawn from zero-inflated negative binomial (ZINB) distribution
Conditioned on the batch annotation s_n of each cell (if available)
And on two additional, unobserved random variables:

- 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.

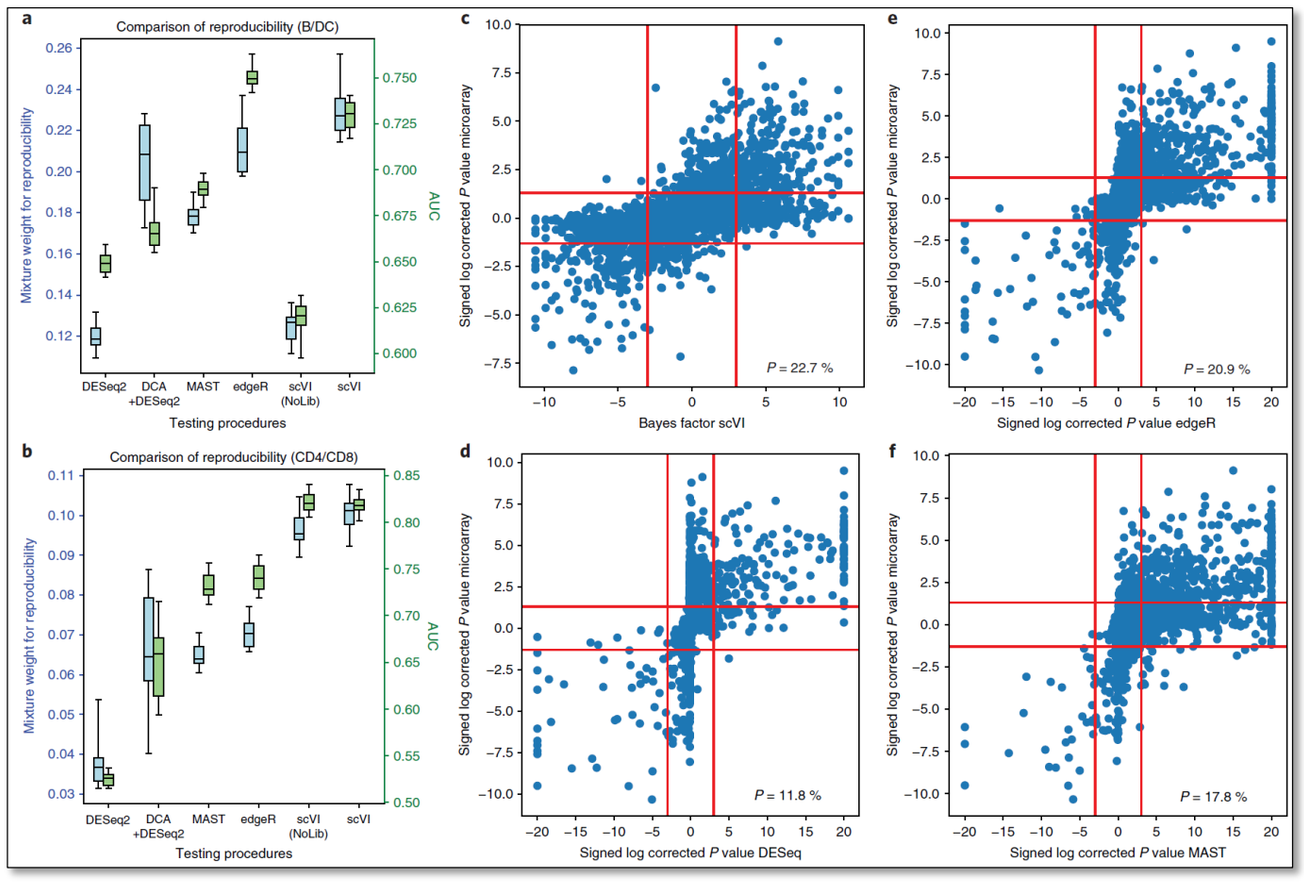
Represent each cell as point in low-dimensional latent space (for visualization and clustering).
Neural network maps the latent variables to ZINB distribution parameters (Fig. 1a, neural networks 5 and 6).
This mapping goes through intermediate variables:

- 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).



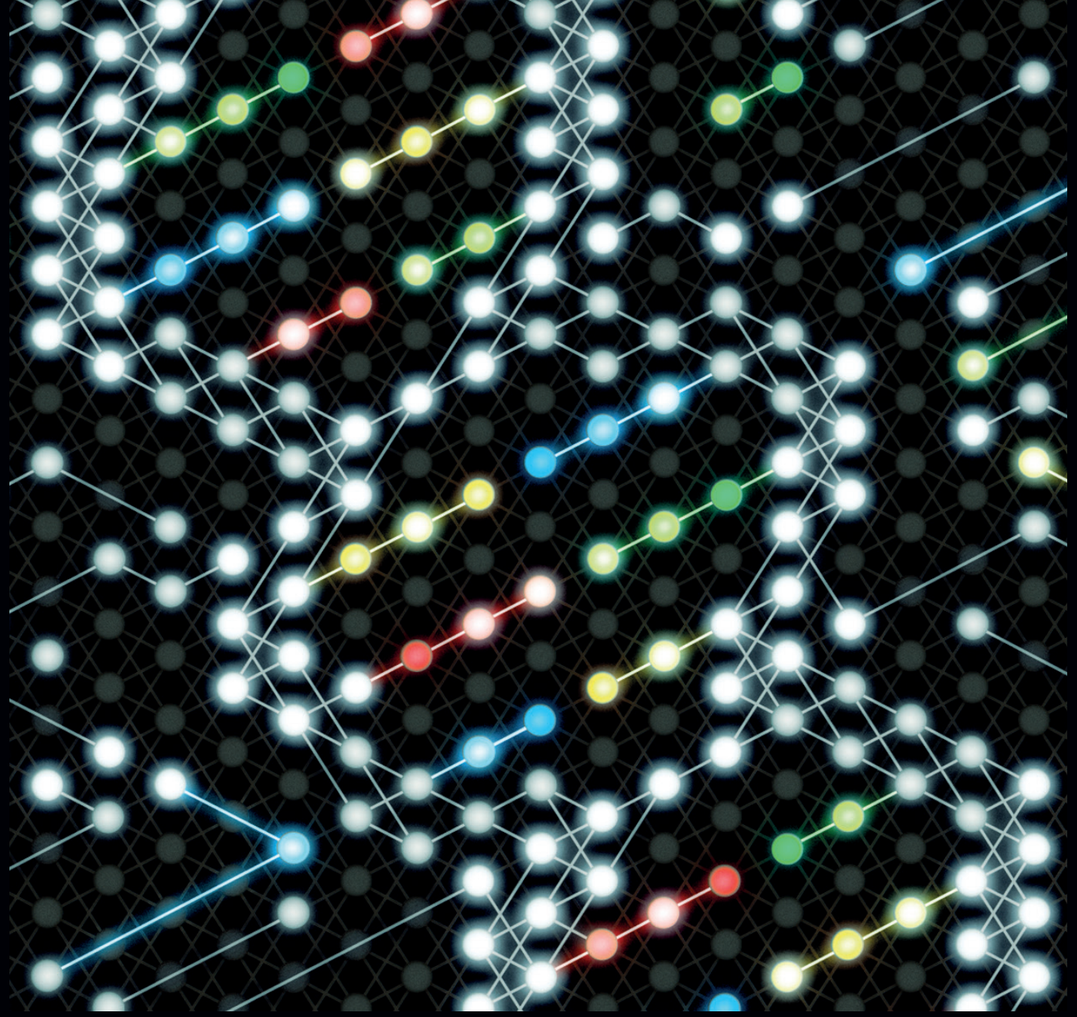
scVI retains biological signal in diverse datasets



scVI enables differential expression analysis

7. Guest Lecture: Fabian Theis

Deep representation learning in single cell genomics



Fabian J. Theis

Institute of Computational Biology, Helmholtz Munich &
Department of Mathematics, TU Munich & Wellcome Trust Sanger Institute

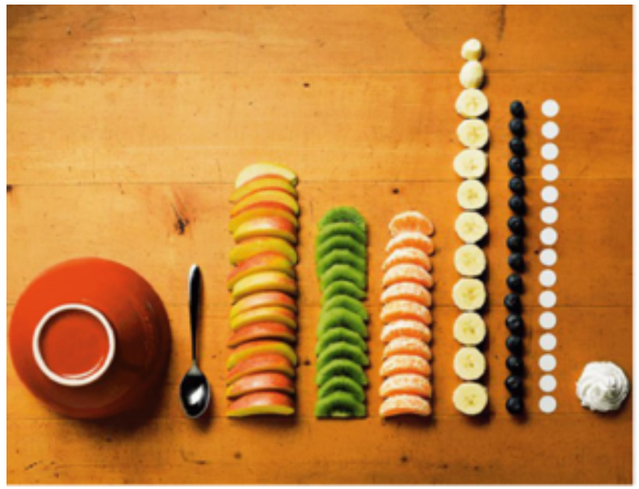
www.comp.bio @fabian_theis

HELMHOLTZAI

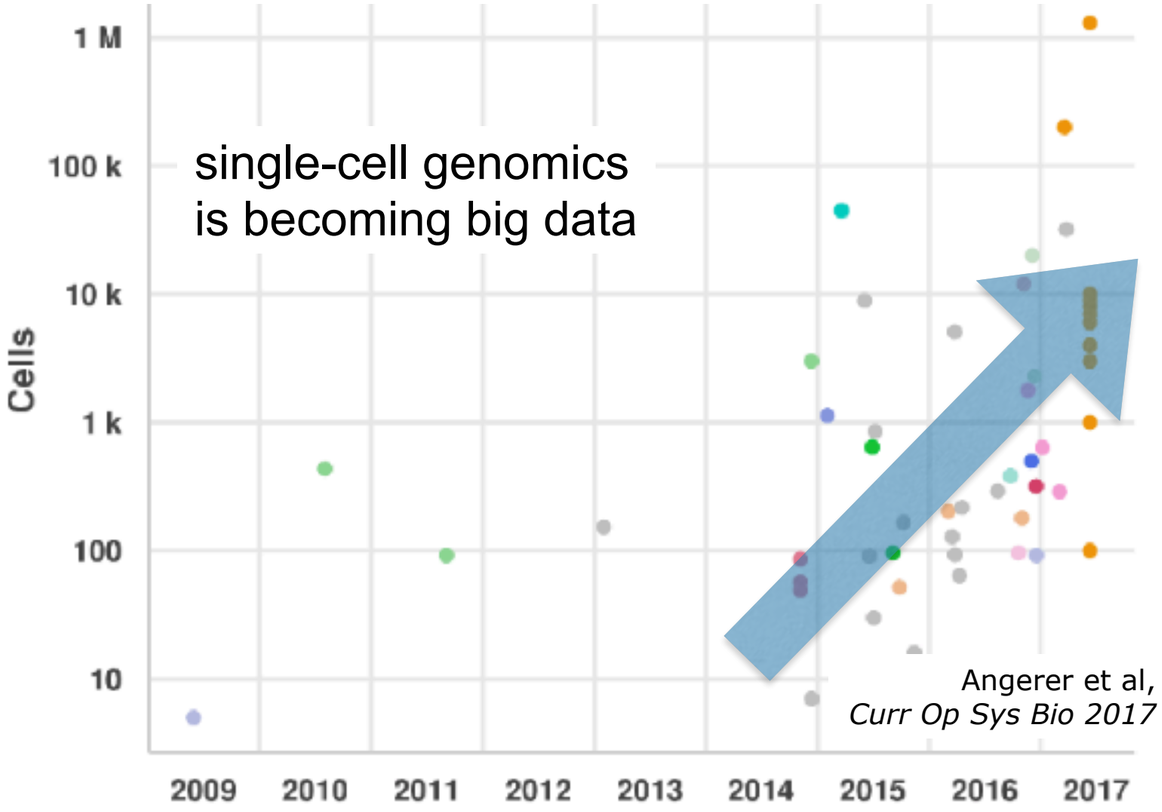
ellis unit munich

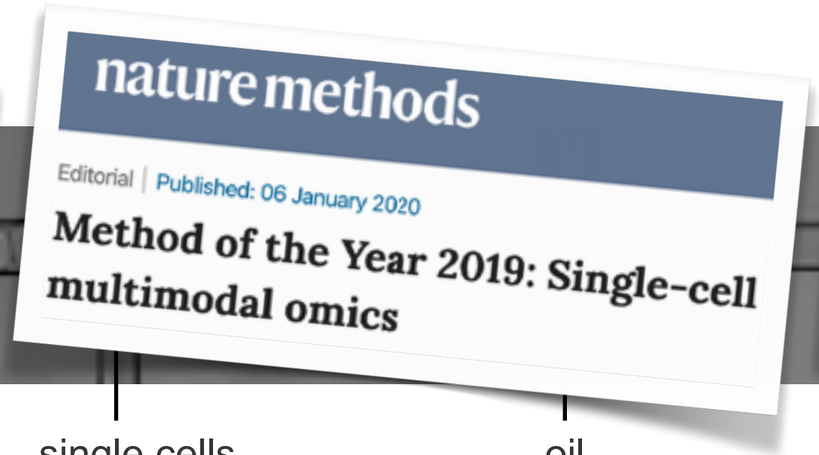
unbiased description of cellular state

bulk genomics



single-cell genomics





gel beads

single cells

oil

adapted from Shalek & Regev & G Zheng (10X Genomics)

single-cell transcriptome analysis

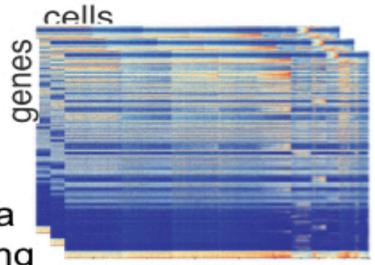
github.com/theislab/scanpy

pypi v0.1.8 | docs passing | build passing

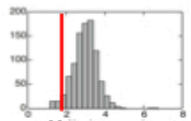
scanpy

Wolf et al, *Genome Biology* 2018

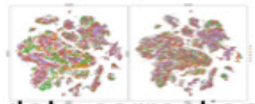
raw data processing



count matrices

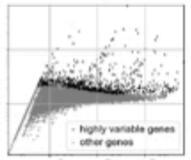


quality control



data correction (e.g. batch)

normalization

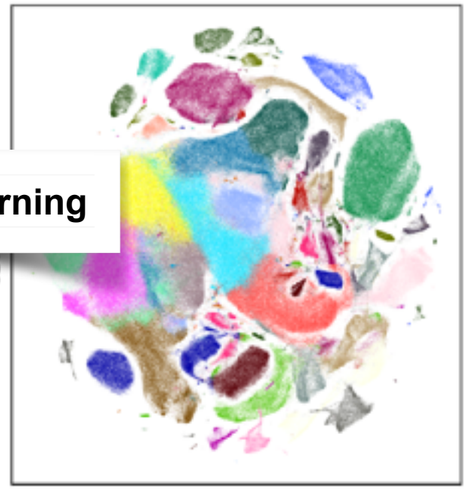


feature selection

visualization

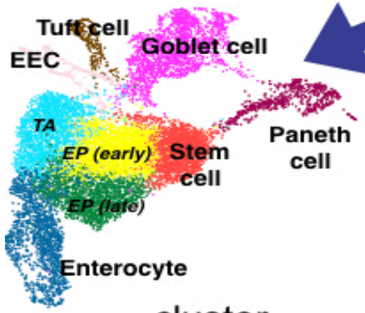
pre-processing

latent space learning



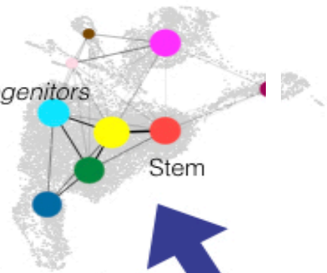
clustering

marker identification

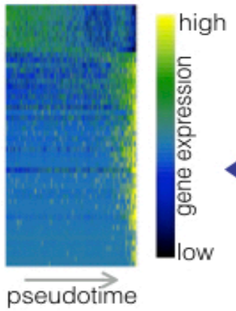


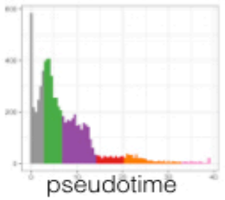
cluster annotation

annotation

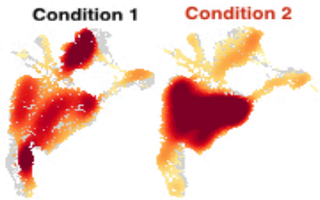


trajectory inference

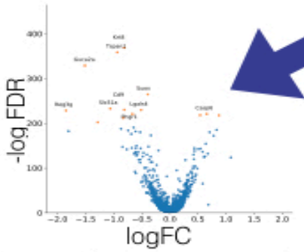




gene dynamics & metastable states



compositional analysis

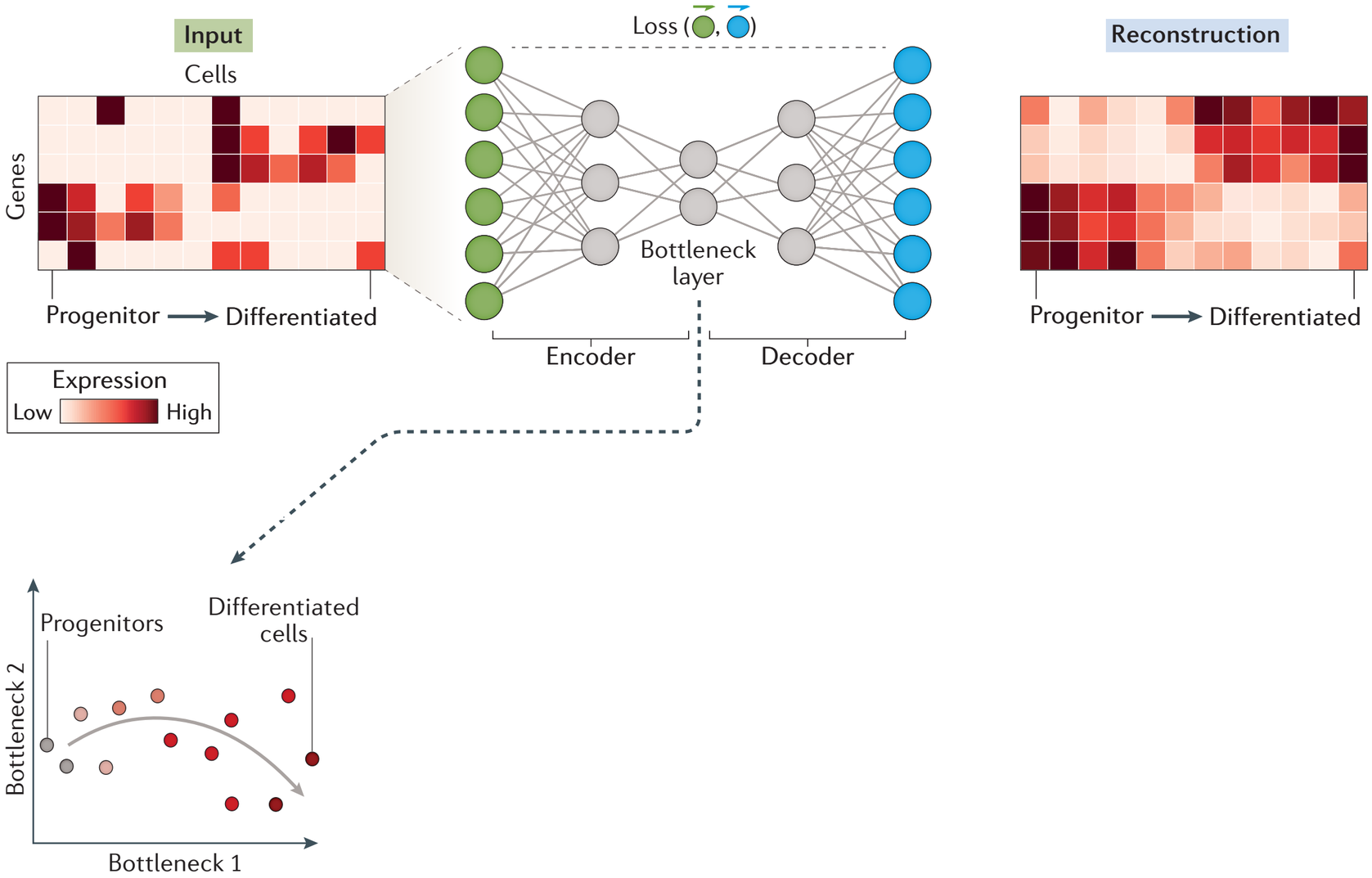


differential expression

downstream analysis

Luecken & T, *Mol Sys Bio* 2019

neural networks for robust latent space learning in scRNA-seq



Deep learning for genomics
Eraslan & Avsec et al, *Nat Rev Genetics* 2019

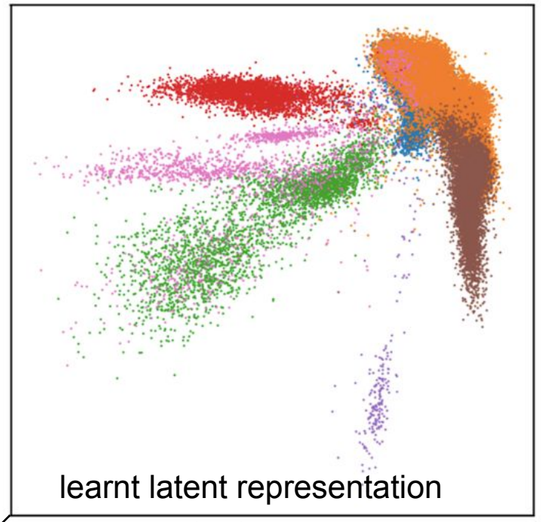
neural networks for robust latent space learning in scRNA-seq

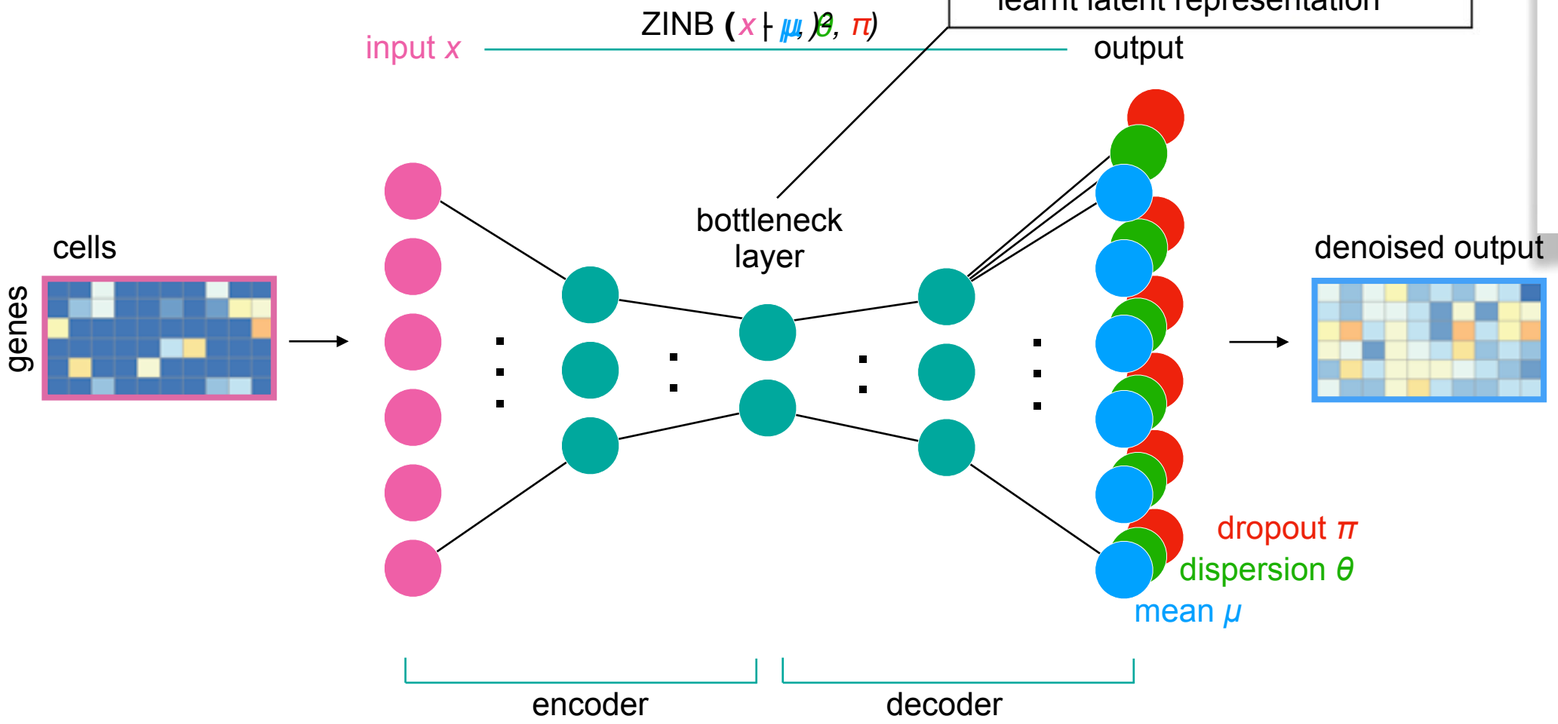
github.com/theislab/DCA

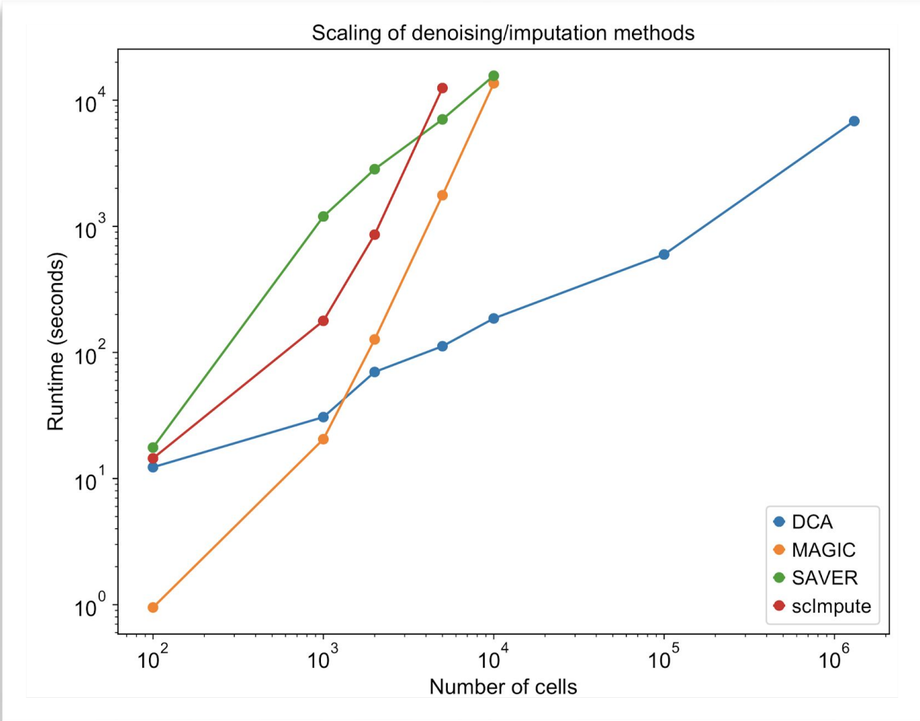
pypi v0.1.8 | docs passing | build passing

increasing interest: scVAE, scVI, VASC, SAUCIE, MAGAN, version of GANs (*bioRxiv, Nat Meth 2018*)

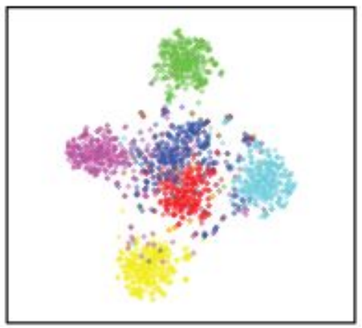
example: denoising by a **deep count autoencoder**:
- compress expression profiles to reduce noise
- replace MSE cost function by adapted ZINB loss



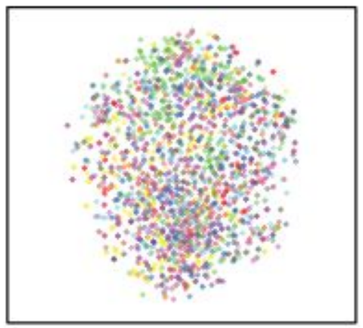




DCA denoised

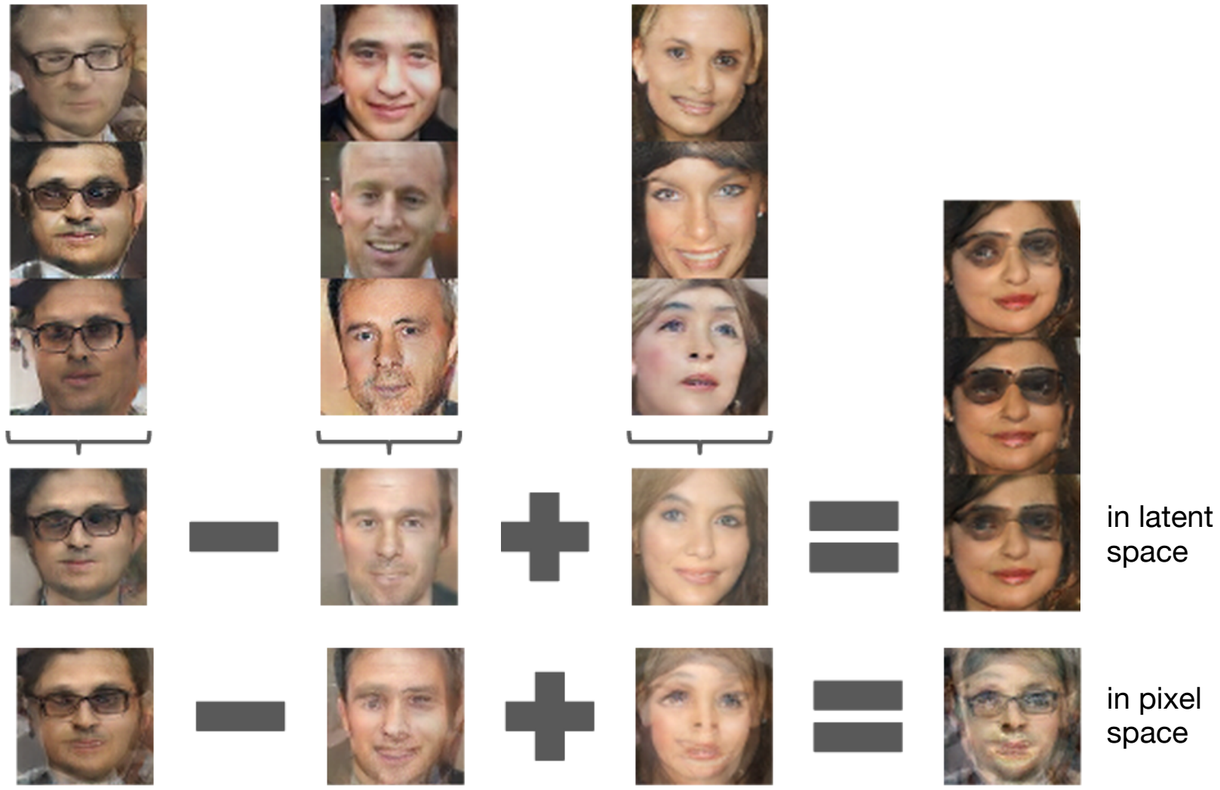


MSE denoised



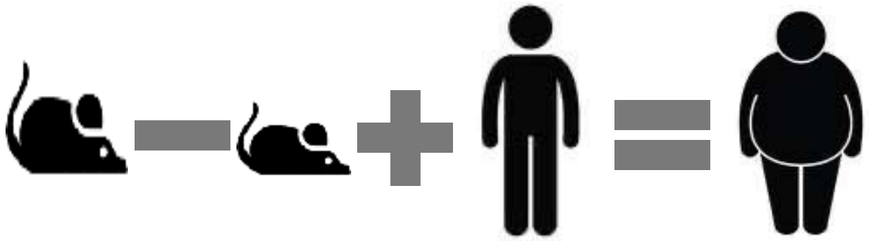
Eraslan & Simon et al. *Nat Comm* 2019

style transfer & domain adaptation by generative neural networks



deep convolutional generative adversarial networks, Radford et al, ICLR 2016

Question
Can we predict perturbation effect of a cell type given observed effects in other cell types?

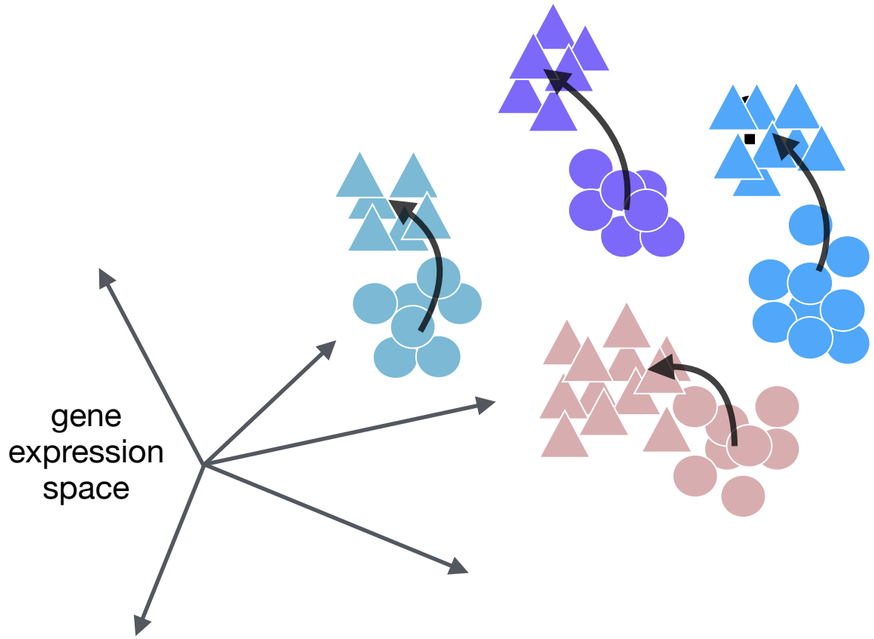


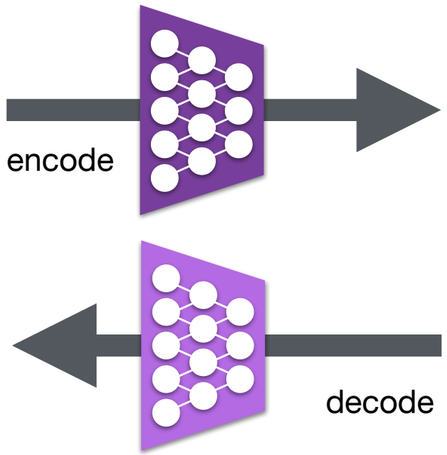
scGen: predicting single-cell perturbation effects using generative models

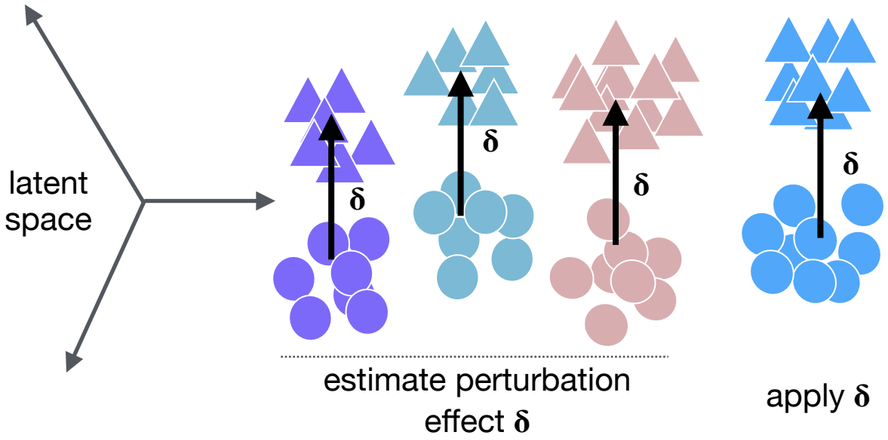
github.com/theislab/scGen

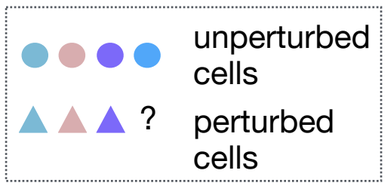
scGen = VAE + latent-space vector arithmetics

goal: out-of-sample prediction



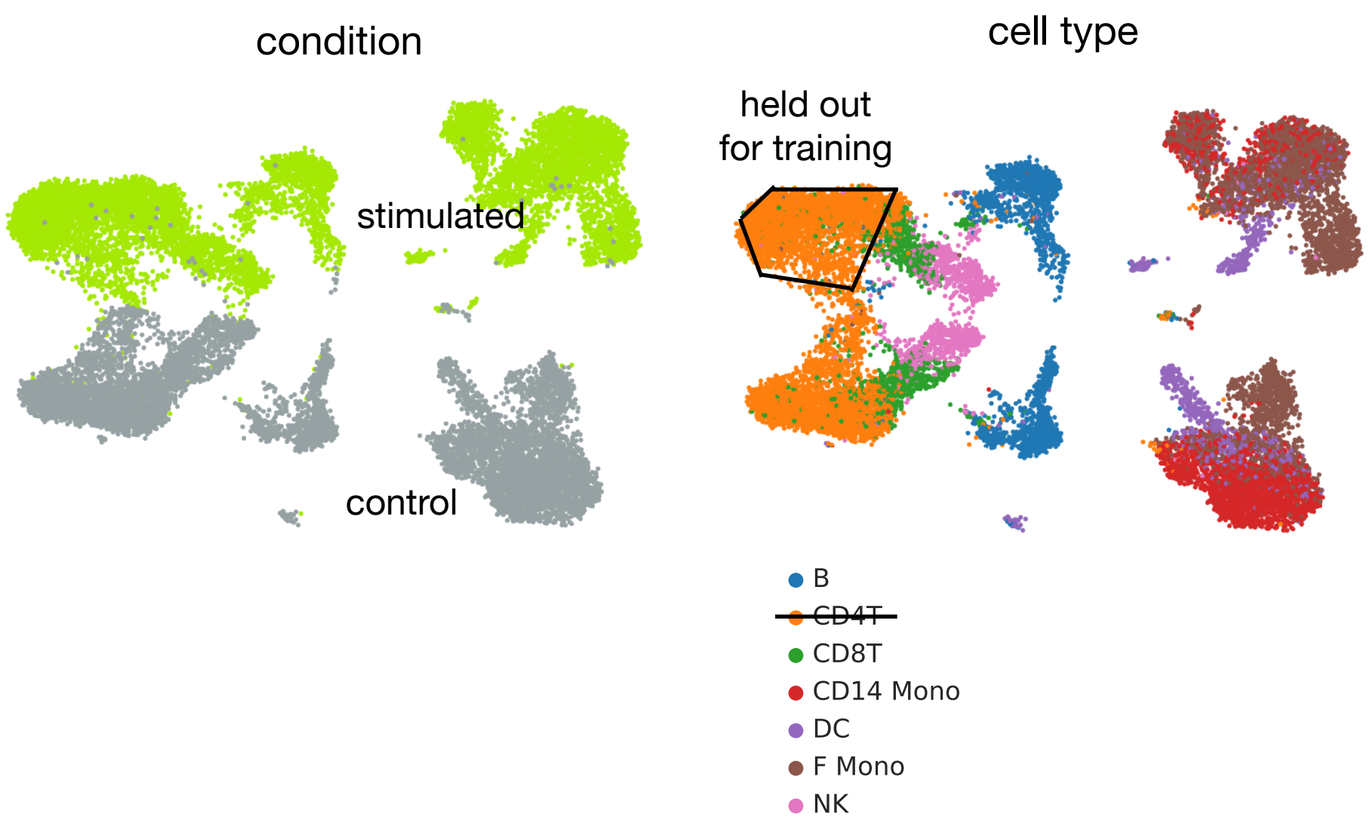


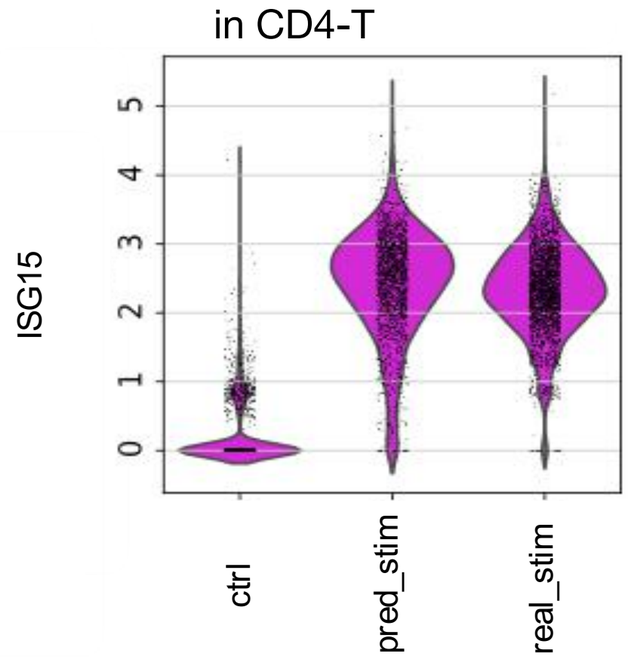


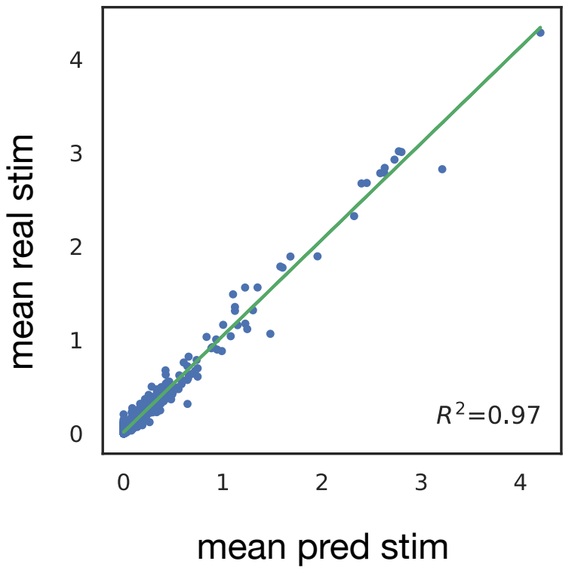


Lotfallahi et al. *Nat Meth 2019*

scGen predicts single-cell perturbation effects for unseen phenotypes

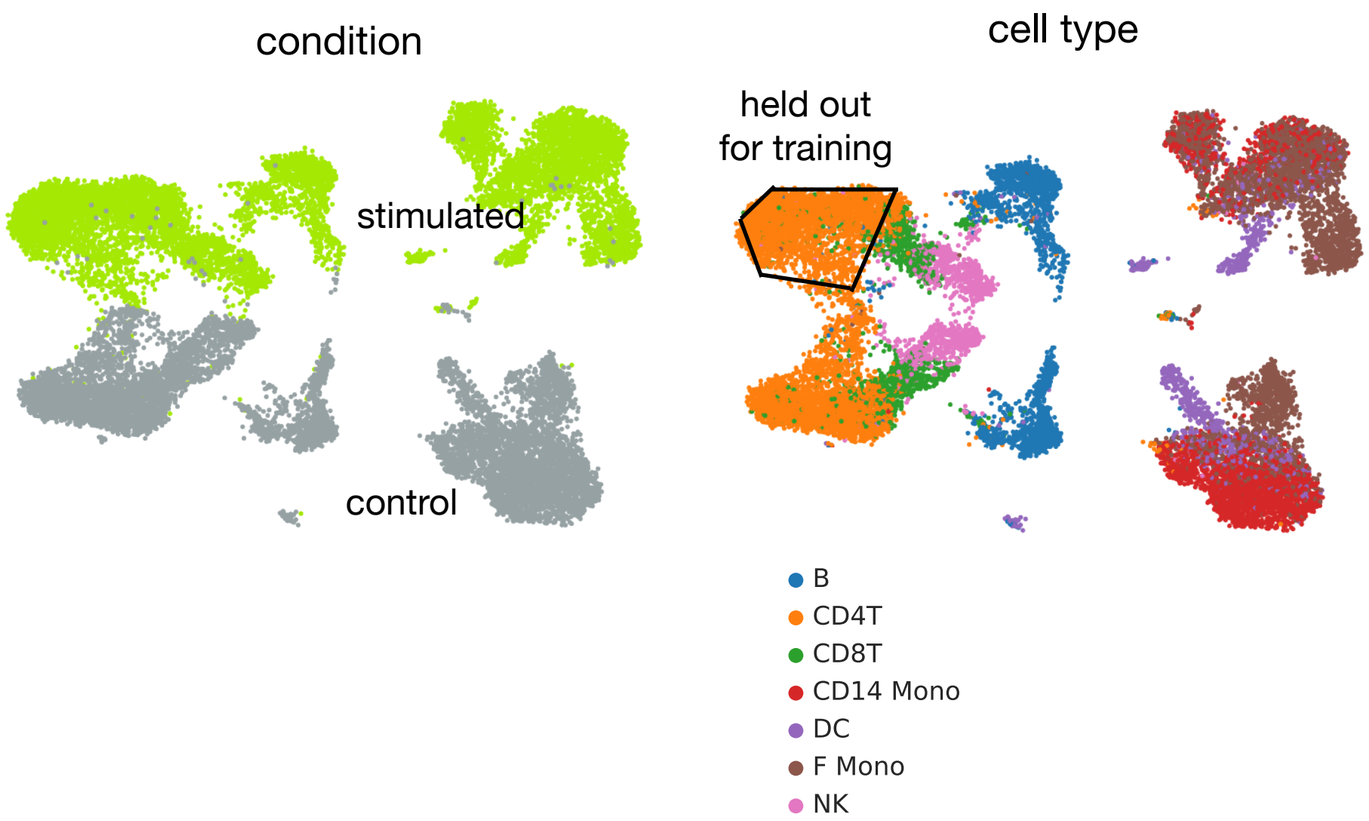


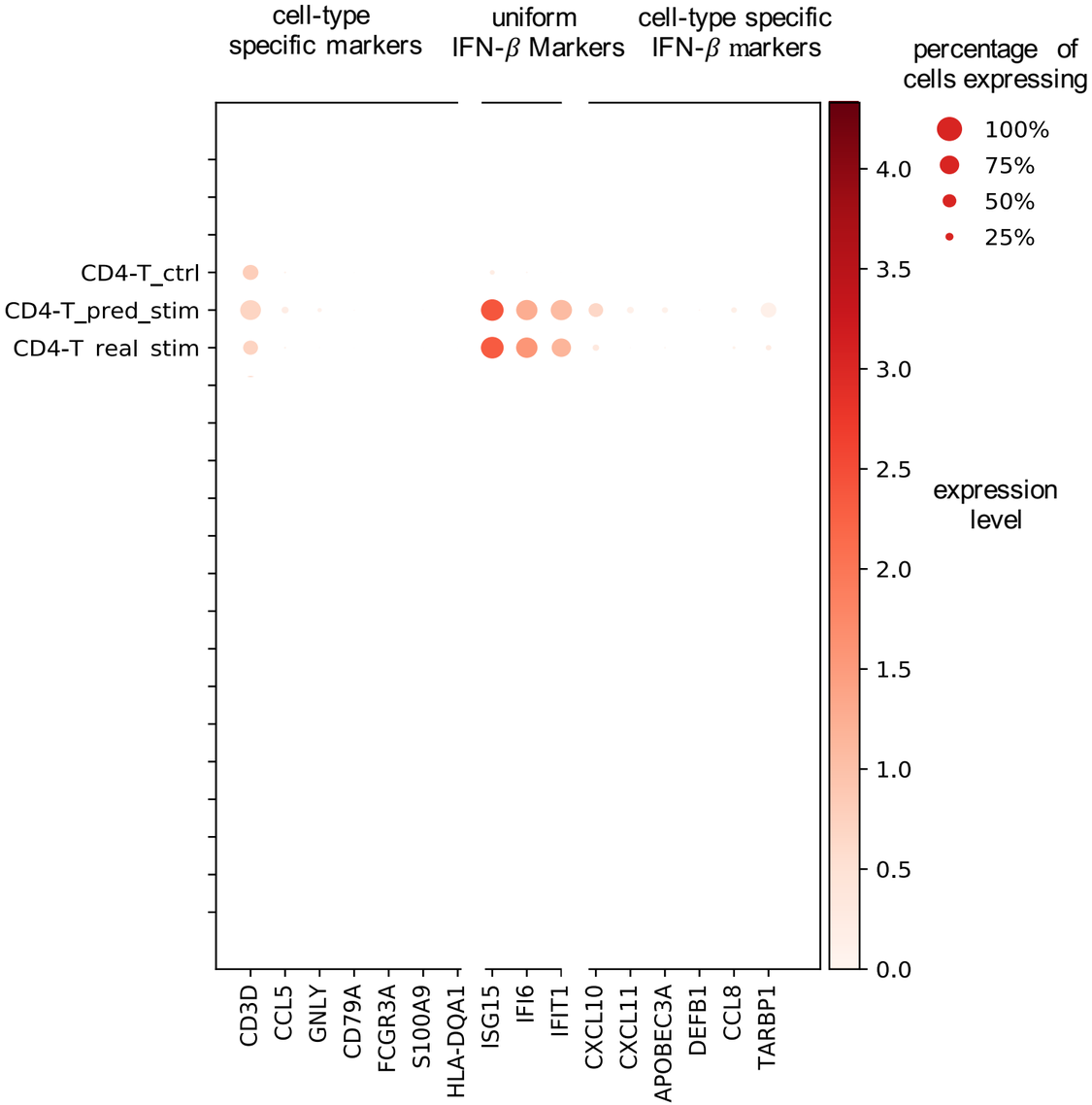




unstimulated + IFN-β-stimulated PBMCs (Kang et al. Nature Biotech, 2018)

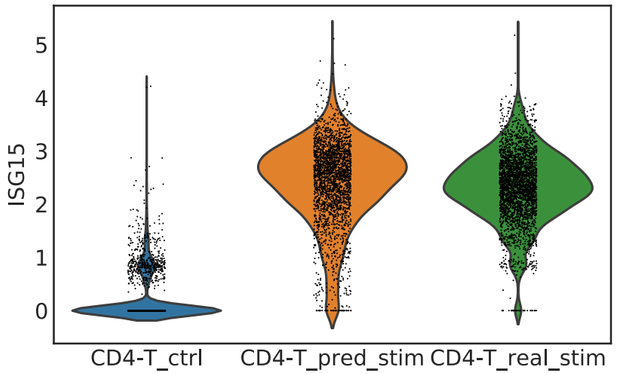
scGen predicts single-cell perturbation effects for unseen phenotypes

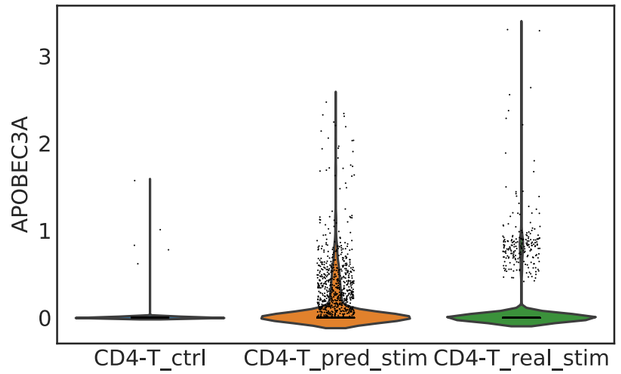


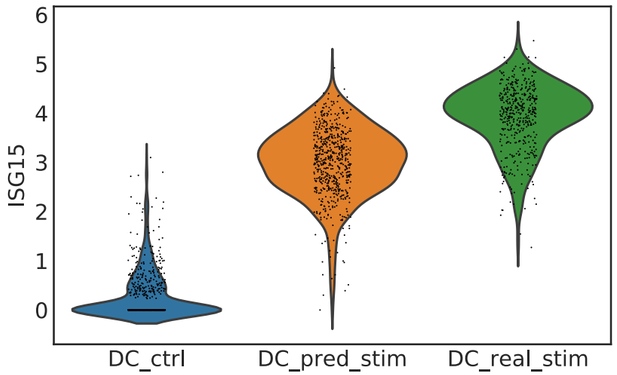


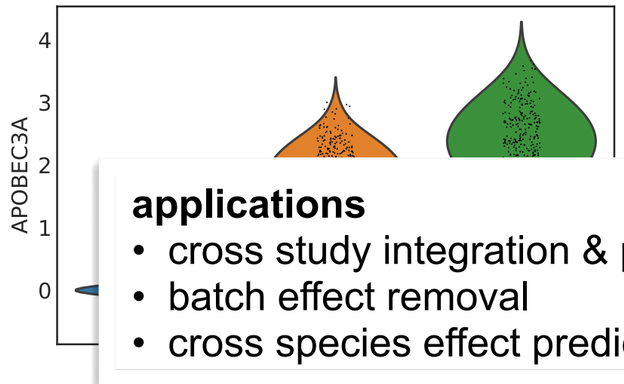
unstimulated + stimulated PBMCs (Kang et al. Nature Biotech, 2018)

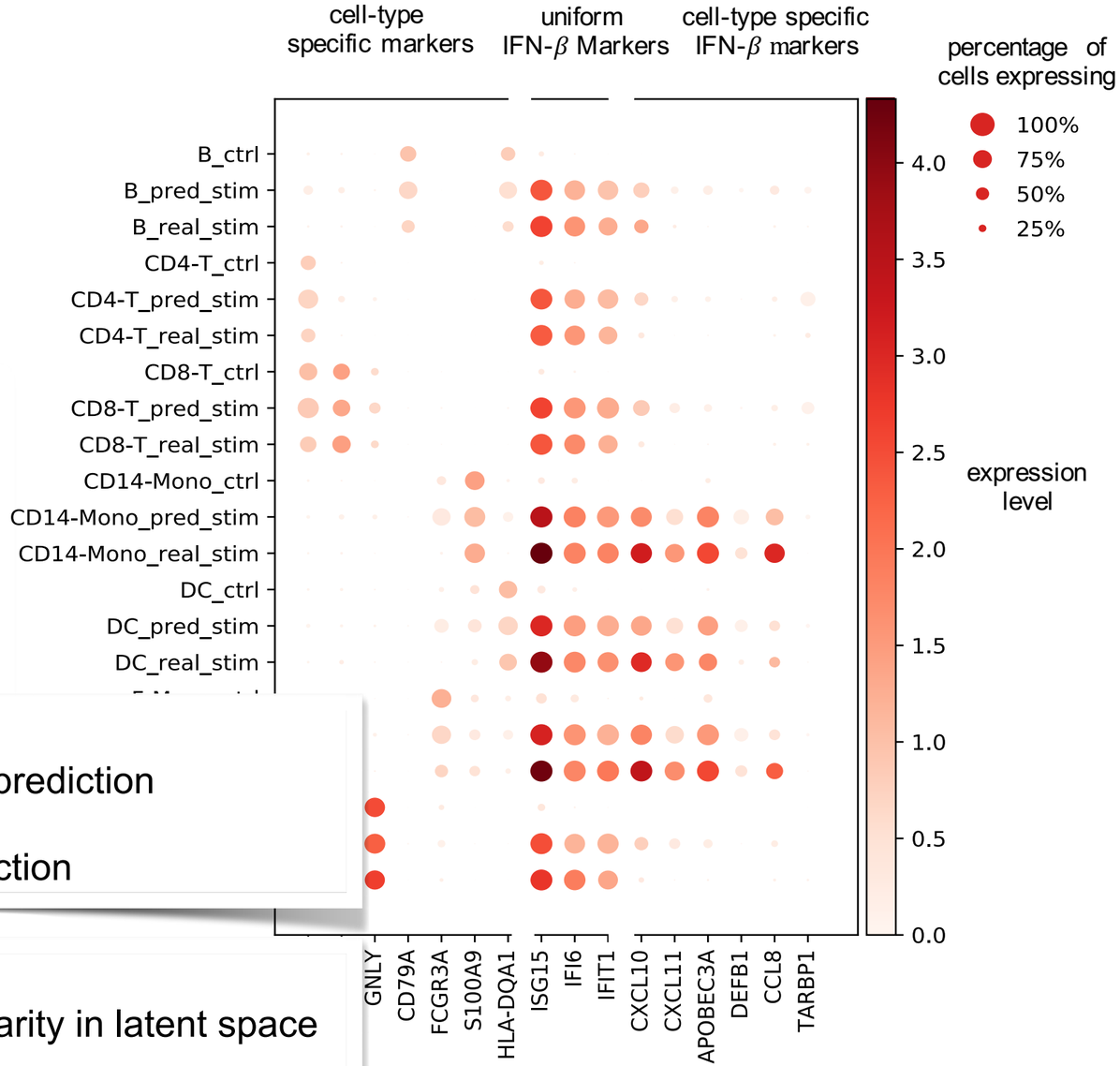
scGen predicts single-cell perturbation effects for unseen phenotypes











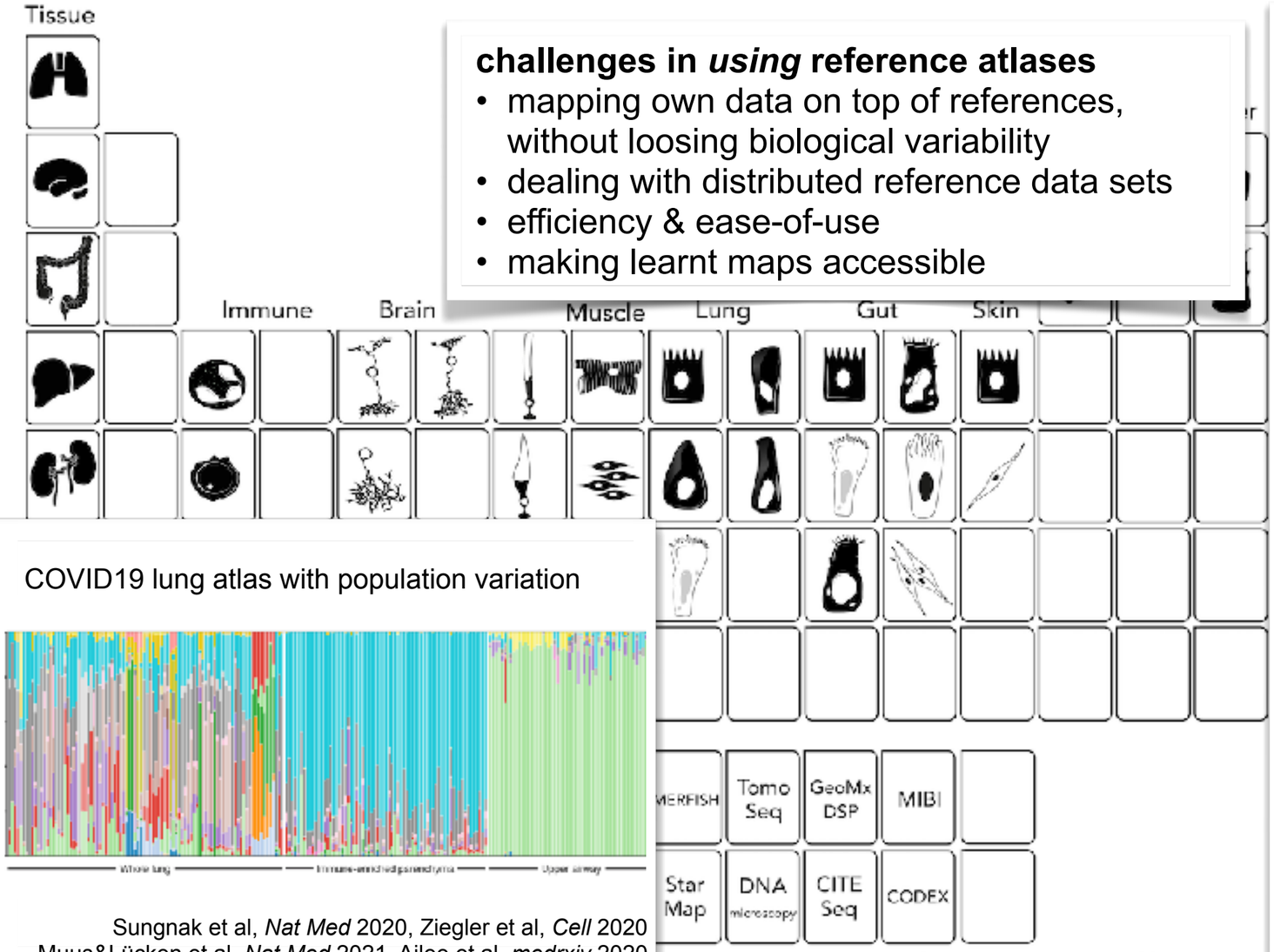
applications
- cross study integration & prediction
- batch effect removal
- cross species effect prediction

limitations
- rigid model, empirical linearity in latent space
- only single perturbation

unstimulated + stimulated PBMCs (Kang et al. Nature Biotech, 2018)

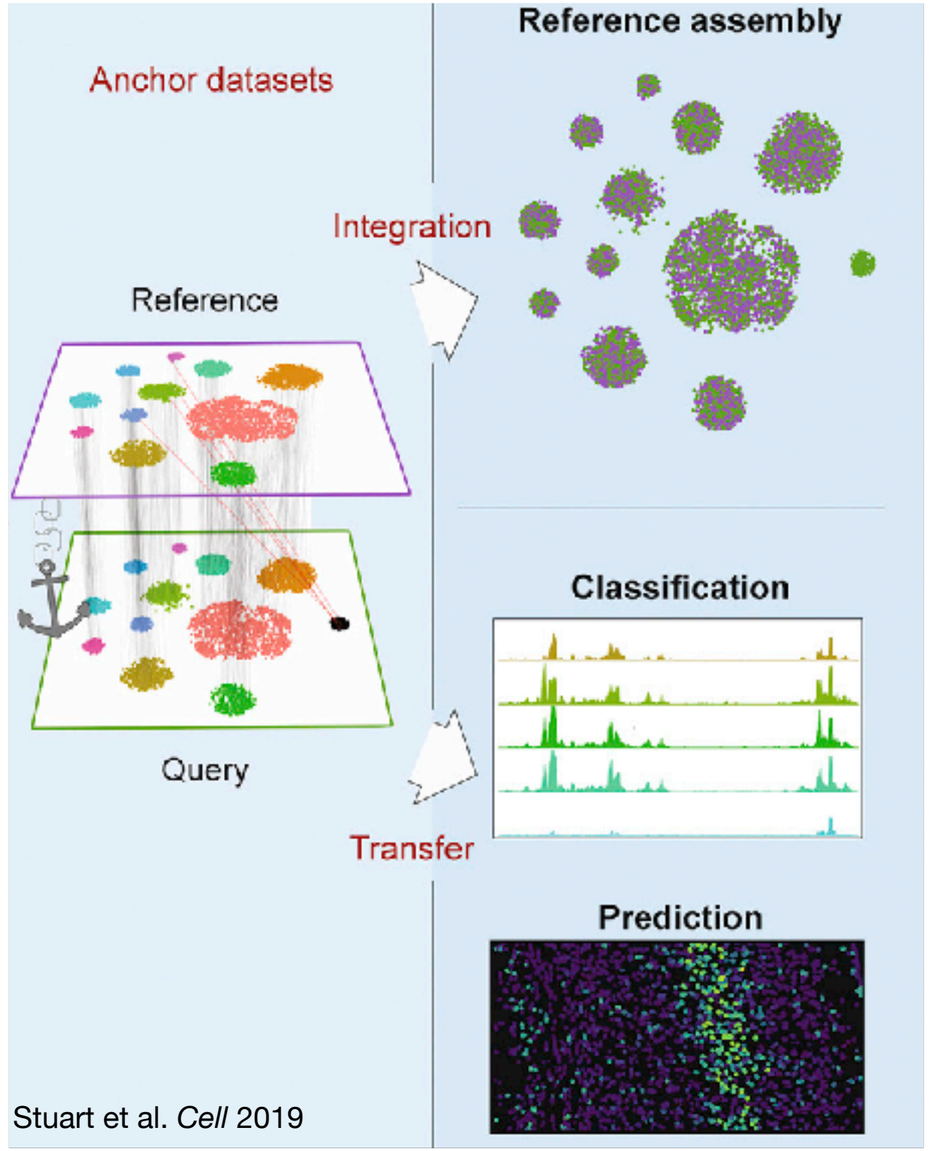
Using a reference cell atlases

challenges in *using* reference atlases
- mapping own data on top of references, without loosing biological variability
- dealing with distributed reference data sets
- efficiency & ease-of-use
- making learnt maps accessible



COVID19 lung atlas with population variation

Sungnak et al, *Nat Med* 2020, Ziegler et al, *Cell* 2020
Muus&Lücken et al, *Nat Med* 2021, Ailee et al, *medrxiv* 2020



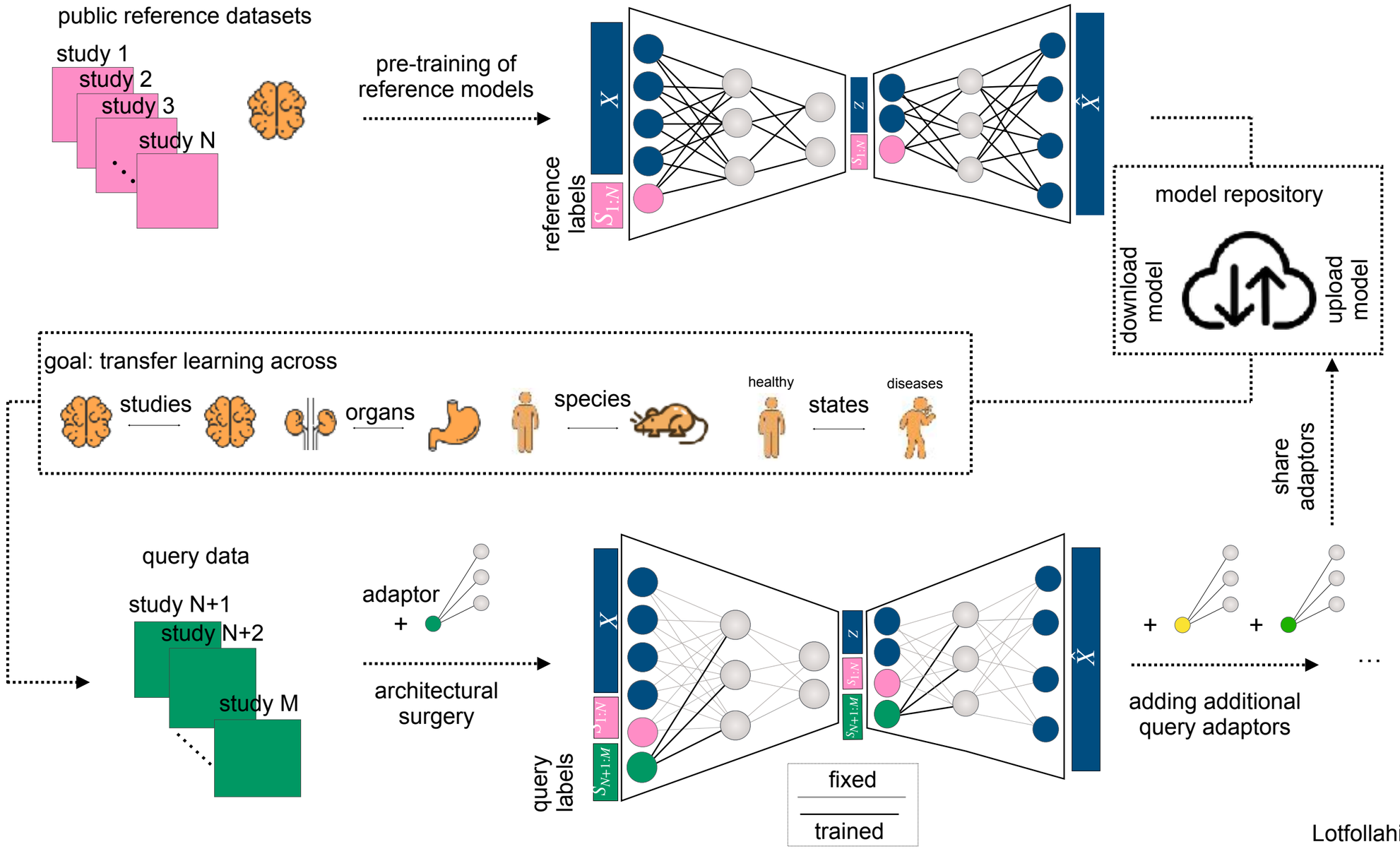
Stuart et al. *Cell* 2019

github.com/theislab/scArches

pypi v0.1.8 | docs passing | build passing

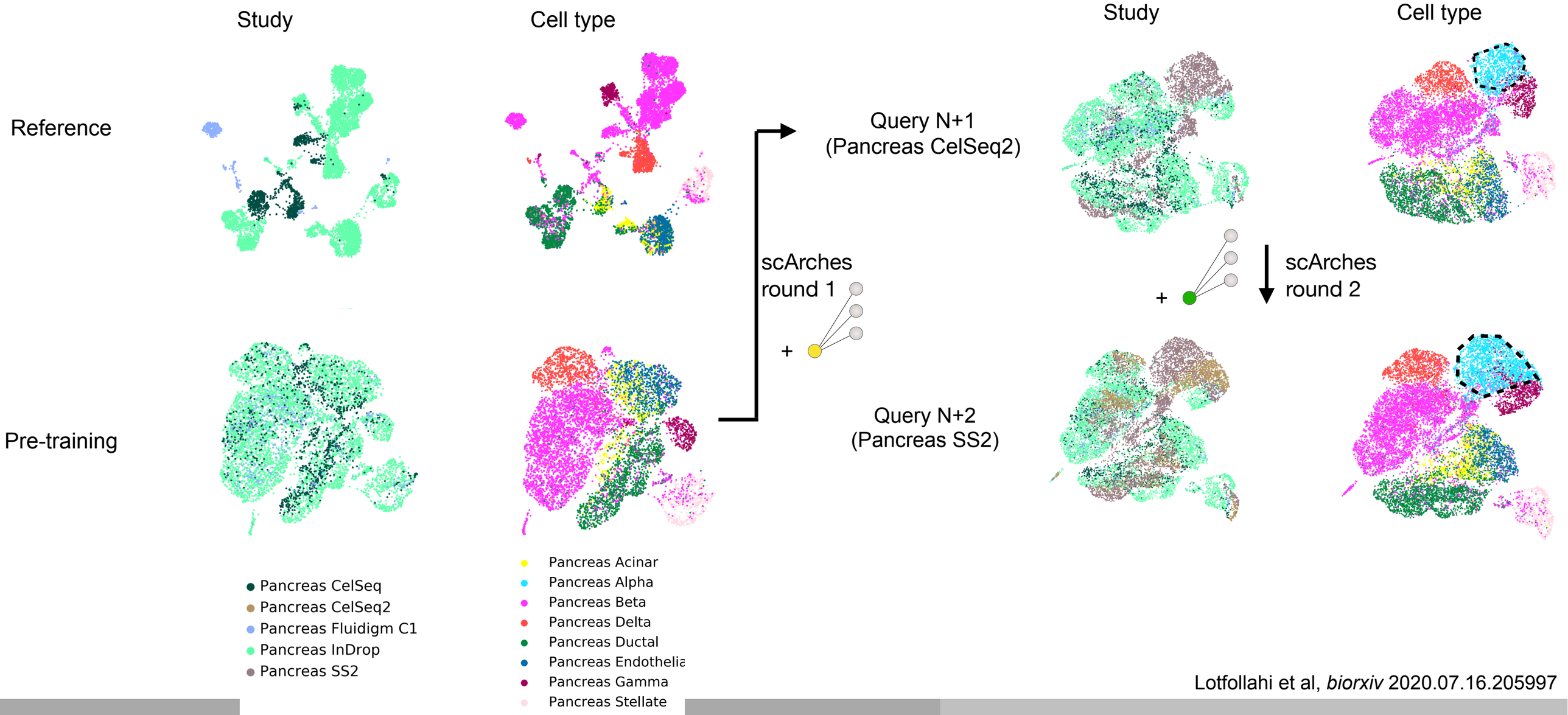
Query-to-reference data integration by transfer learning

M Lotfallahi



Lotfollahi et al, *biorxiv* 2020.07.16.205997

Query-to-reference data integration by transfer learning

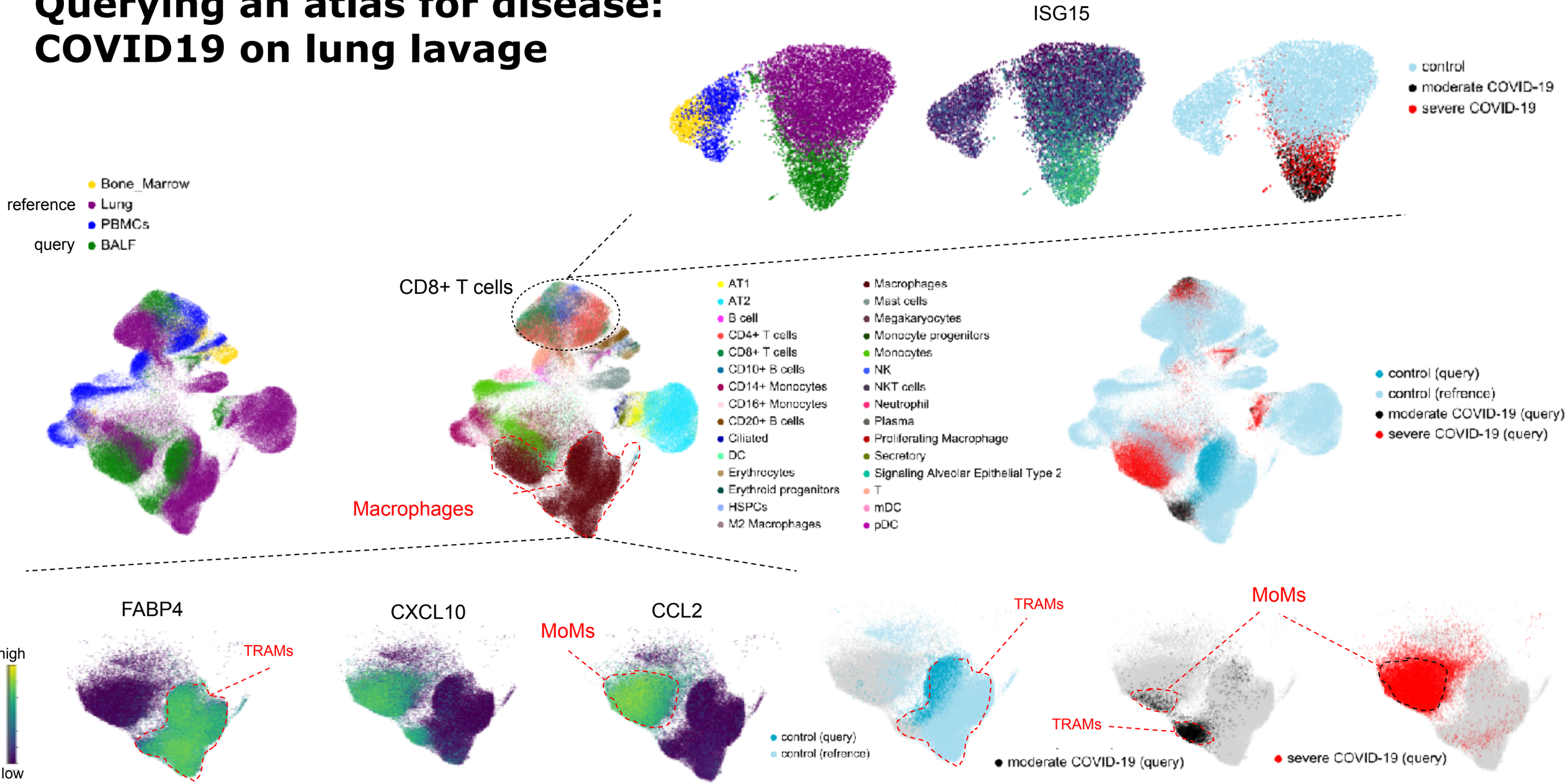


Lotfollahi et al, *biorxiv* 2020.07.16.205997

github.com/theislab/scArches

pypi v0.1.8 | docs passing | build passing

Querying an atlas for disease: COVID19 on lung lavage



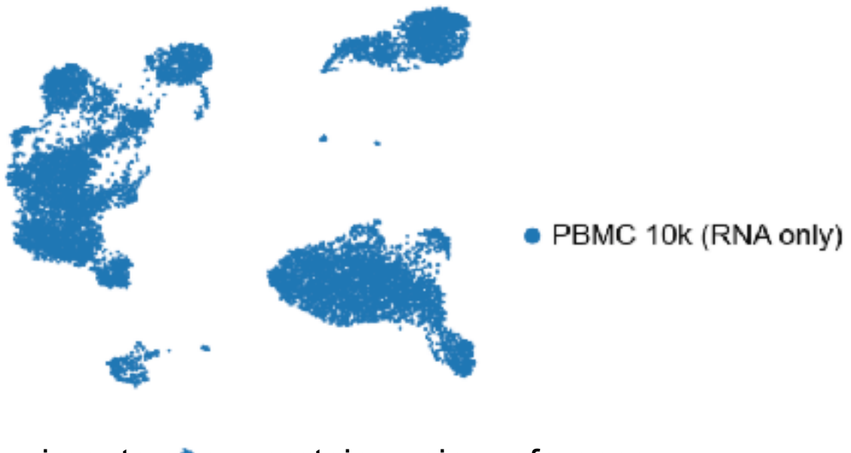
Lotfollahi et al, *biorxiv* 2020.07.16.205997

scArches allows construction of multi-modal reference atlases

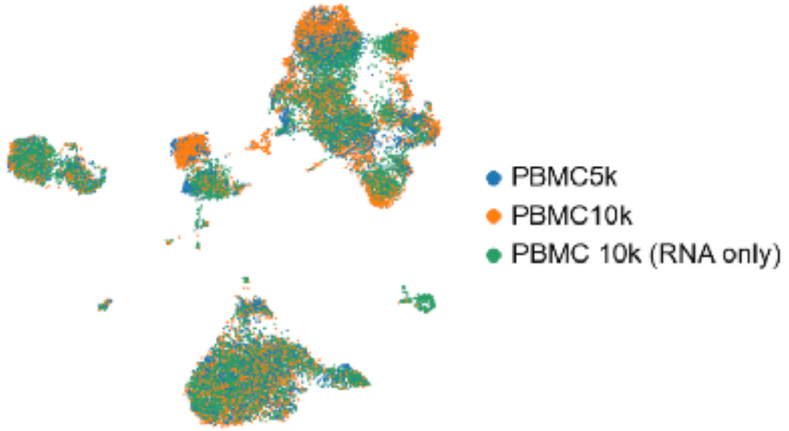
i**dea**: use multi-modal latent space model (totalVI from Yosef lab) on reference to reconstruct query proteins

reference (RNA+protein via CITE-seq)

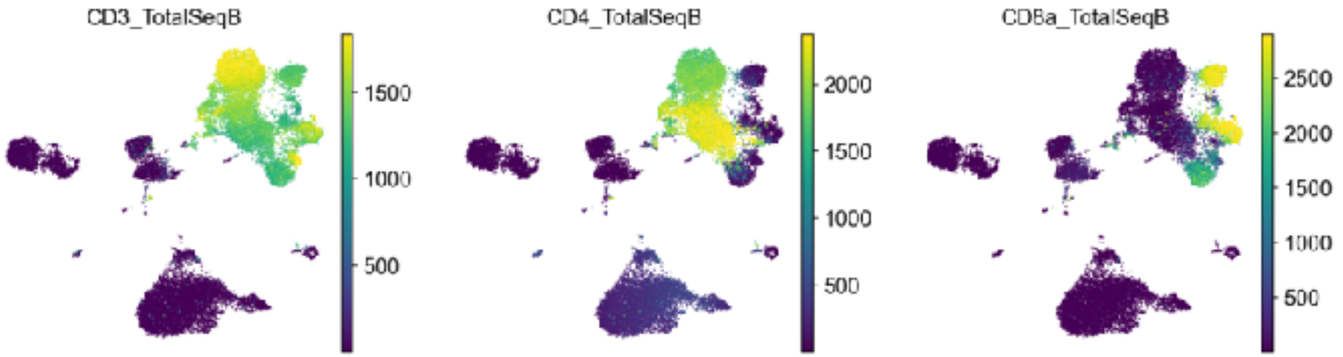
query (RNA only)



reference and query

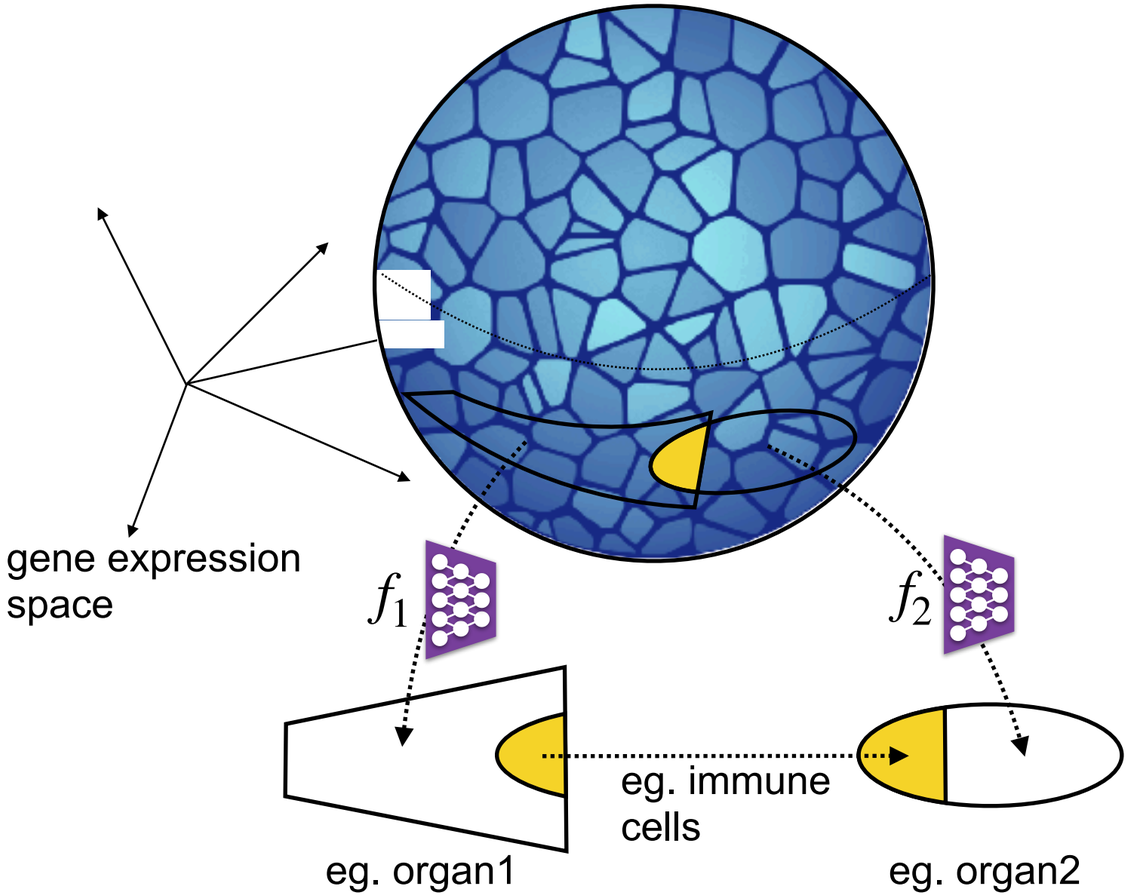


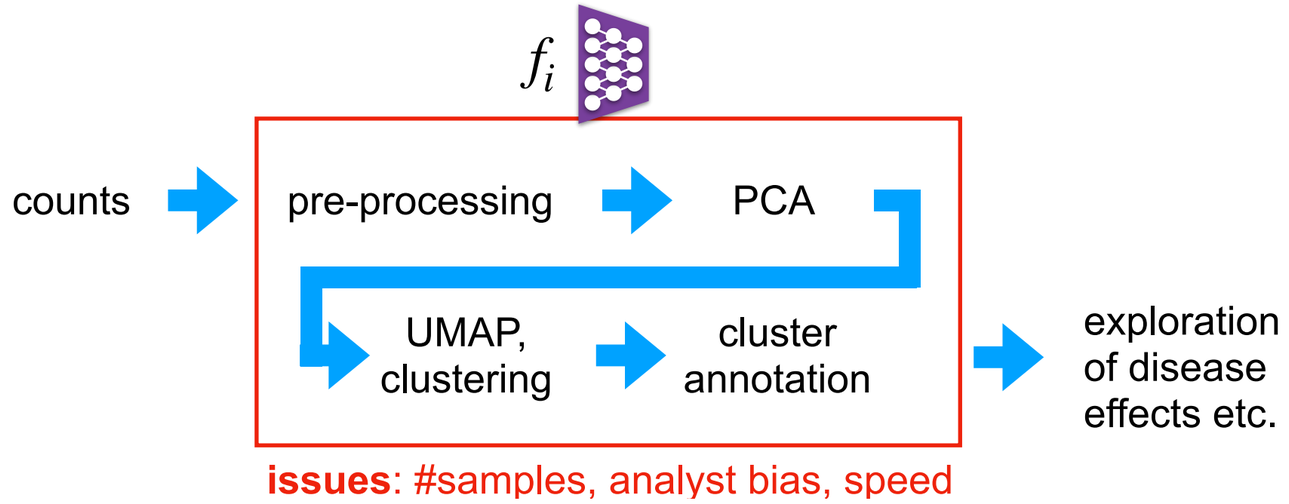
impute query proteins using reference

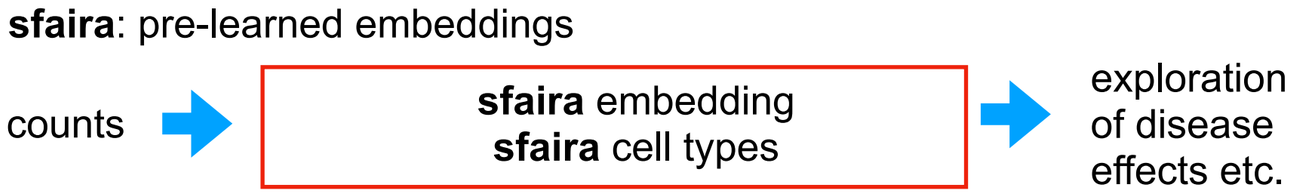


from Sergei Rybakov (T lab) & Adam Gayoso (Yosef lab); data sets from 10x Genomics; totalVI for analysis of CITE-seq data (Gayoso et al 2020)

how to easily use neural networks? → manifold & atlas idea







sfaira - single-cell model zoo

github.com/theislab/sfaira

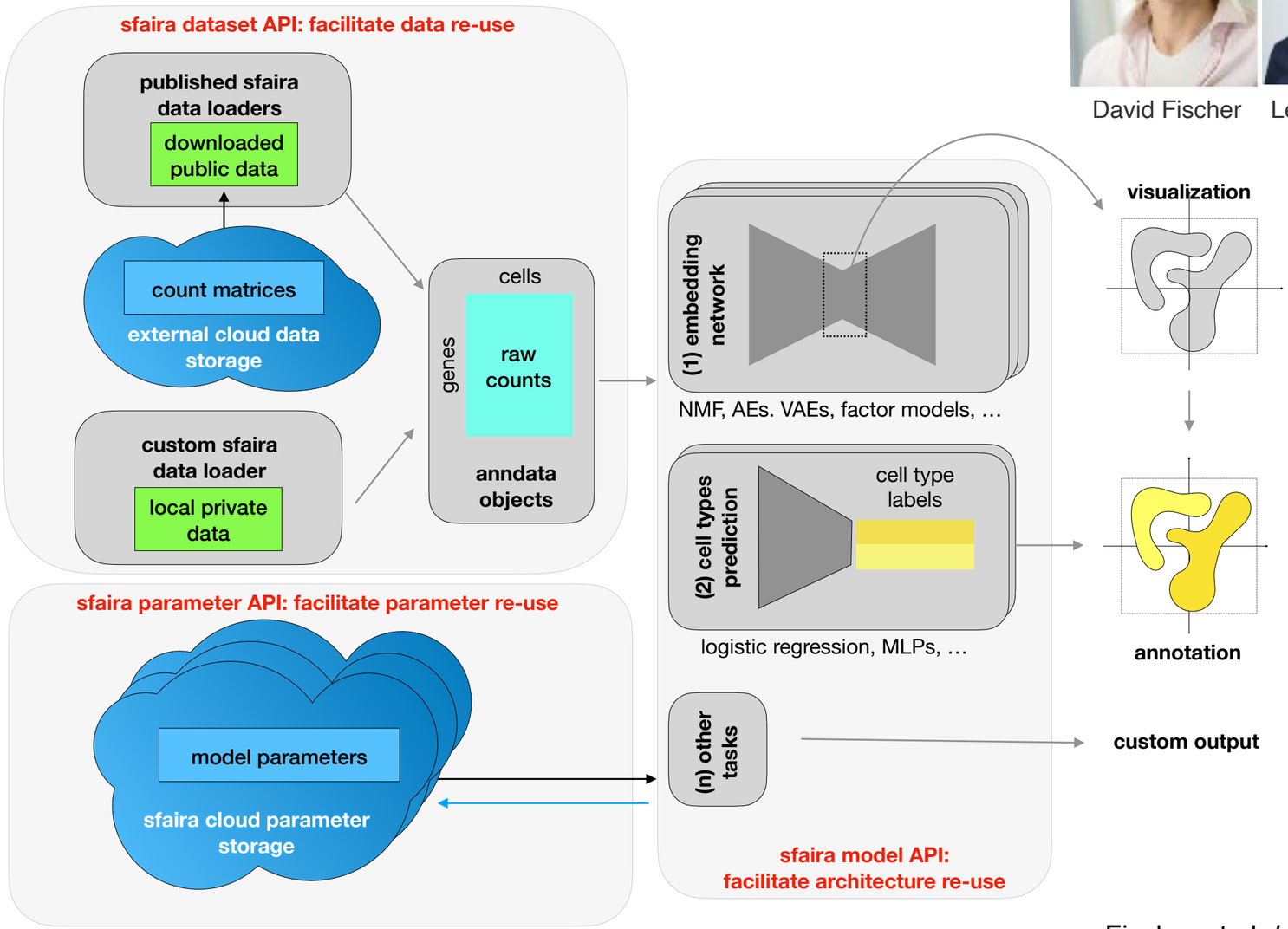
pypi v0.1.8 | docs passing | build passing

David Fischer

Leander Dony

aim: comparable, reproducible & easy access to annotated single-cell data sets and trained network models

sfaira = (dataset, annotation, model, parameters)

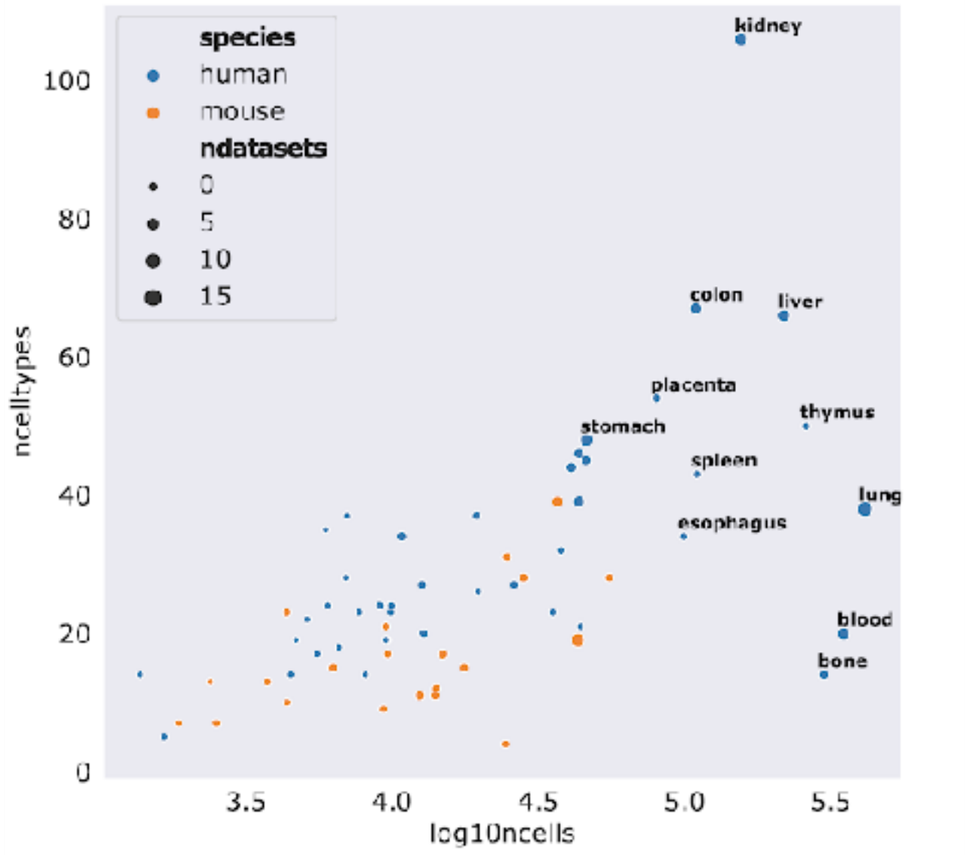


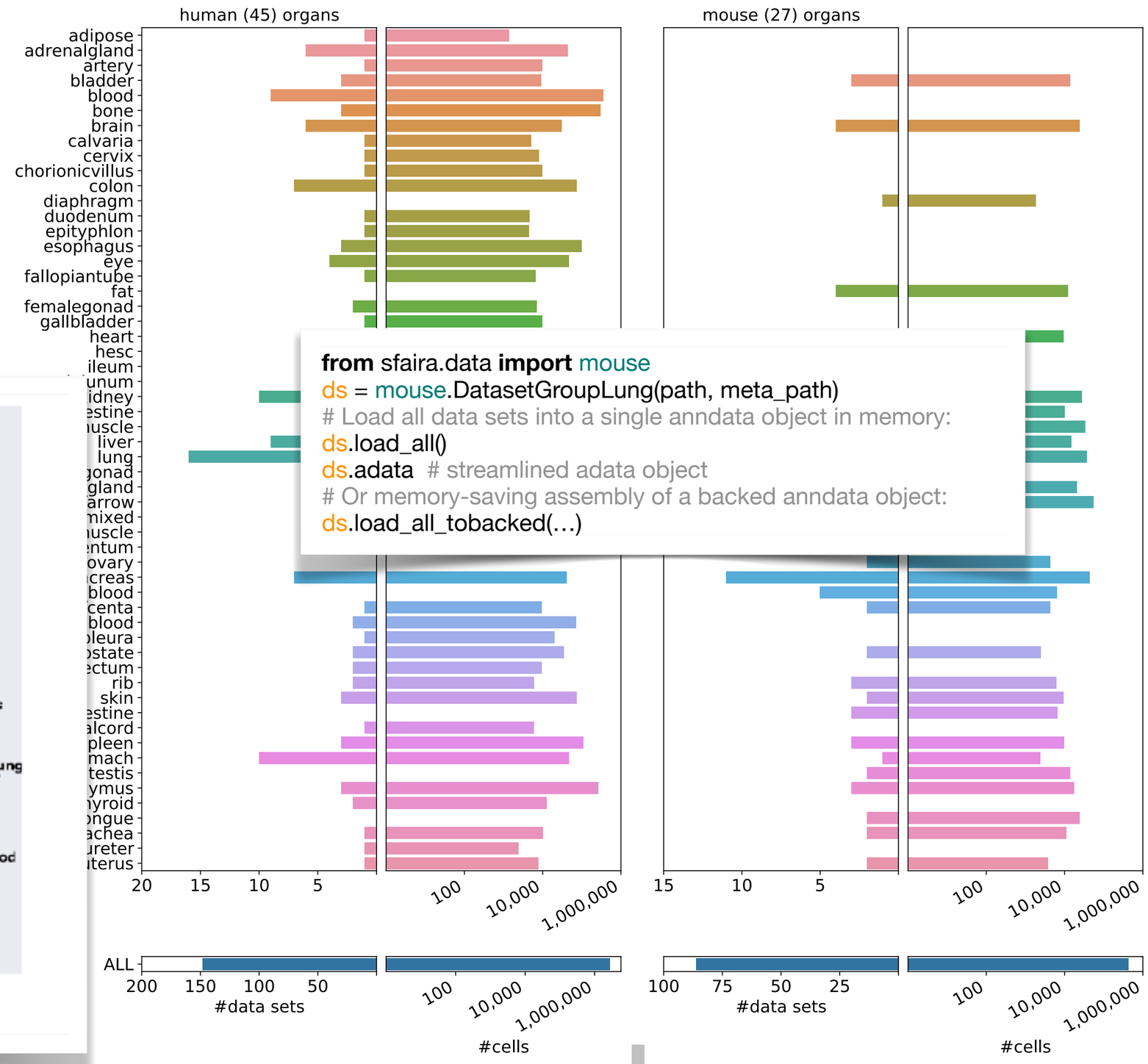
Fischer et al, *biorxiv* 2020

sfaira - data sets

240 data sets, 55 organs, 3.2M cells

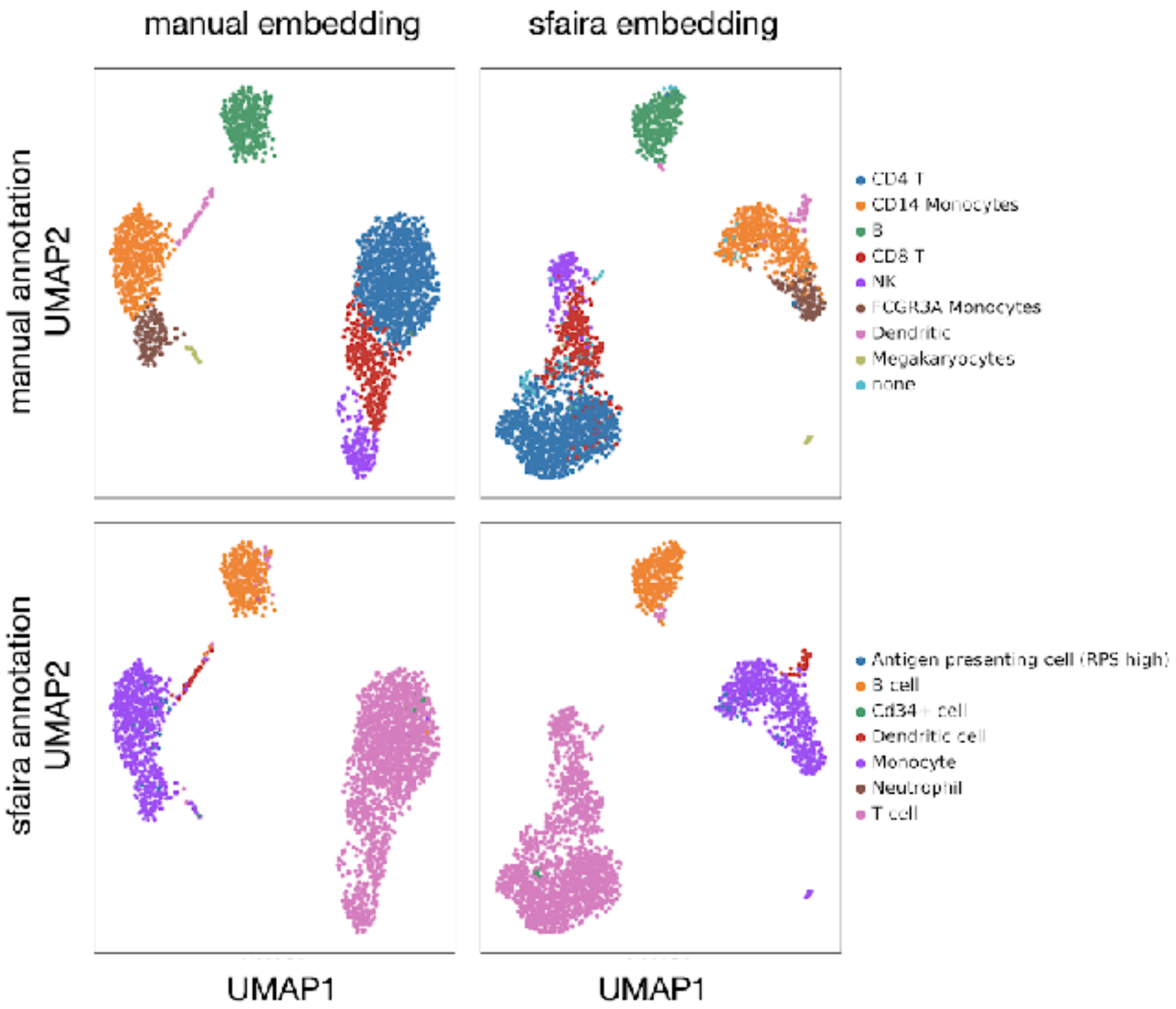
can easily stream eg. all human cells (2.7M, 20k genes, 200 GB h5ad file) using mini batching






```
from sfaira.data import mouse
ds = mouse.DatasetGroupLung(path, meta_path)
# Load all data sets into a single anndata object in memory:
ds.load_all()
ds.adata  # streamlined adata object
# Or memory-saving assembly of a backed anndata object:
ds.load_all_tobacked(...)
```

sfaira - easily use trained latent space embeddings to facilitate standard single-cell workflows

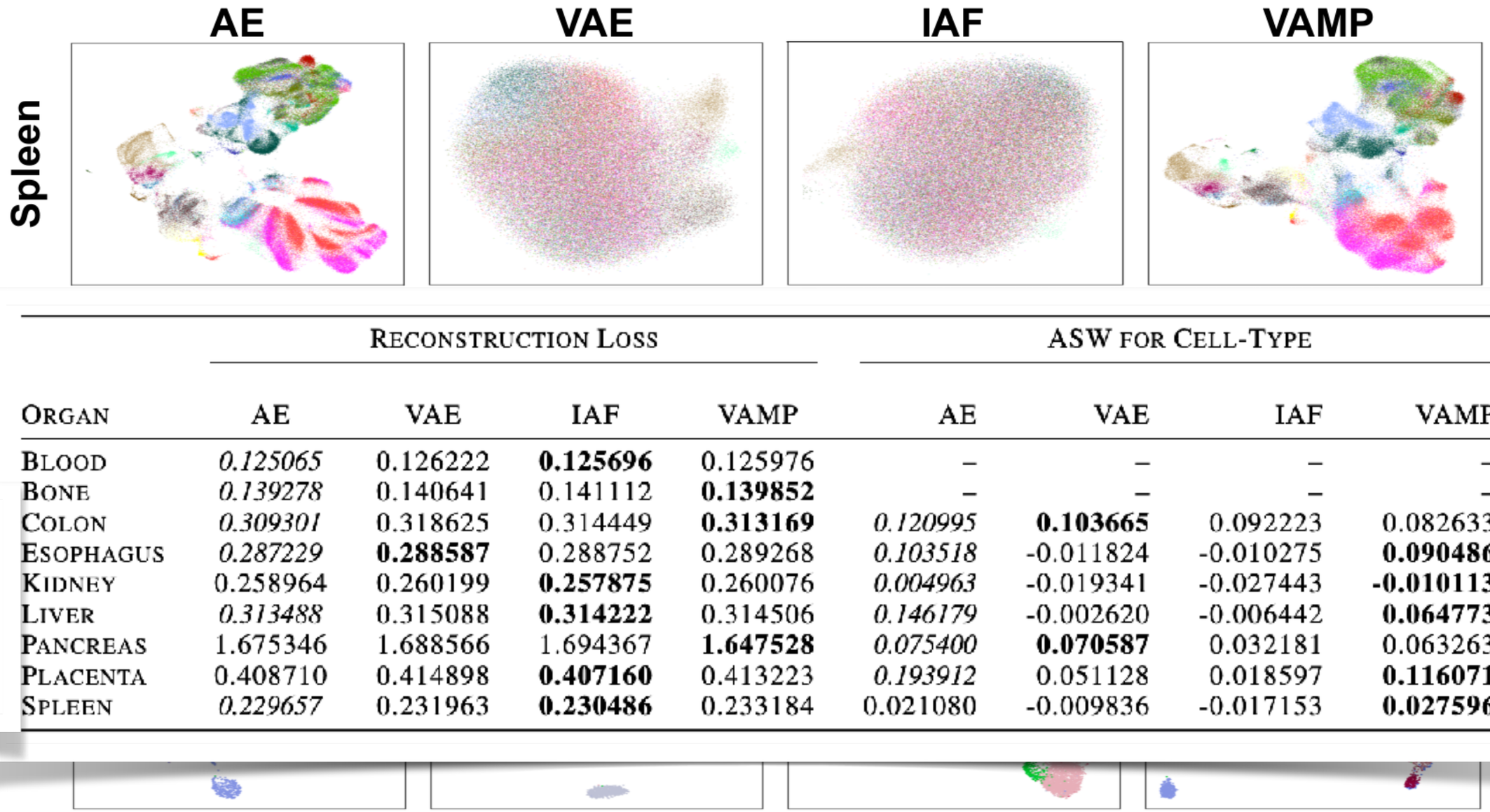


Fischer et al, *biorxiv* 2020

github.com/theislab/sfaira

pypi v0.1.8 | docs passing | build passing

example application: evaluate priors in single-cell VAEs



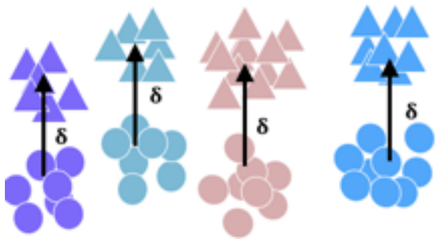
	Reconstruction Loss				ASW for Cell-Type			
Organ	AE	VAE	IAF	VAMP	AE	VAE	IAF	VAMP
Blood	*0.125065*	0.126222	**0.125696**	0.125976	–	–	–	–
Bone	*0.139278*	0.140641	0.141112	**0.139852**	–	–	–	–
Colon	*0.309301*	0.318625	0.314449	**0.313169**	*0.120995*	**0.103665**	0.092223	0.082633
Esophagus	*0.287229*	**0.288587**	0.288752	0.289268	*0.103518*	-0.011824	-0.010275	**0.090486**
Kidney	0.258964	0.260199	**0.257875**	0.260076	*0.004963*	-0.019341	-0.027443	**-0.010113**
Liver	*0.313488*	0.315088	**0.314222**	0.314506	*0.146179*	-0.002620	-0.006442	**0.064773**
Pancreas	1.675346	1.688566	1.694367	**1.647528**	*0.075400*	**0.070587**	0.032181	0.063263
Placenta	0.408710	0.414898	**0.407160**	0.413223	*0.193912*	0.051128	0.018597	**0.116071**
Spleen	*0.229657*	0.231963	**0.230486**	0.233184	0.021080	-0.009836	-0.017153	**0.027596**

sfaira enabled **quick evaluation on 16 public scRNA-seq data sets** across 9 tissues and 700k cells

VAMP prior - Tomczak & Welling, AISTATS 2018
Leander Dony et al, *ICML-WCB* 2020

github.com/theislab

pypi v0.1.8 | docs passing | build passing



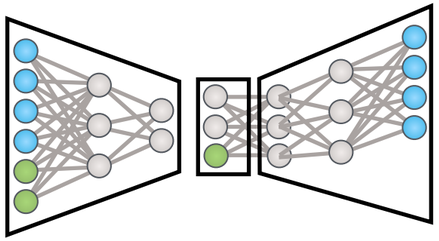
conclusion

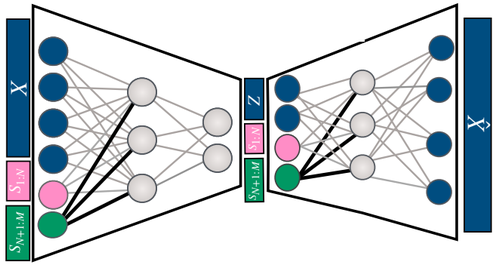
» *latent space learning* in single-cell genomics*:* using autoencoders

» *scgen*: model perturbations as linear shifts in latent space

» *scArches:* use single-cell atlases to query own data via architectural surgery

» *sfaira*: model zoo with pre-learned embeddings and easy data loaders

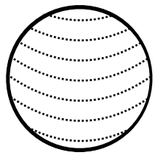


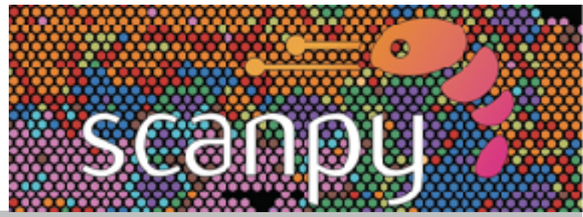


outlook

» extension towards spatial data & models (CNN + graph CNNs) - *squidpy*

» interpreting *latent spaces,* including dynamic information





Acknowledgements

theislab@
Institute of Computational Biology

@fabian_theis www.comp.bio

Bundesministerium für Bildung und Forschung

Deutsche Forschungsgemeinschaft DFG

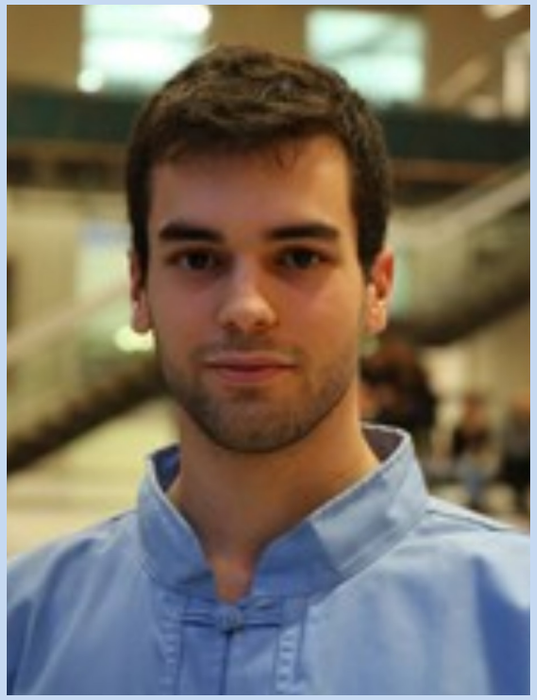
CHAN ZUCKERBERG INITIATIVE

single cell omics germany

MUDS | MUNICH SCHOOL FOR DATA SCIENCE HELMHOLTZ | TUM | LMU

HELMHOLTZ AI

ellis unit munich



8. Guest Lecture: Romain Lopez

Deep Generative Models for Single-cell Transcriptomics

Romain Lopez

University of California, Berkeley

Slide credits: Jeffrey, Nir & Romain

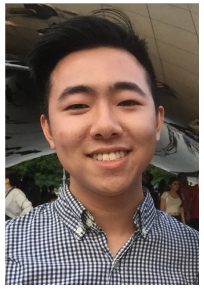
The scVI collaboration

Romain Lopez

Pierre Boyeau

Adam Gayoso

Chenling Xu



Galen Xing

Jeff Regier

Mike Jordan

Nir Yosef

& Maxime Langevin, Edouard Melhman, Jules Samaran, Achille Nazaret, Gabriel Misrachi, Oscar Clivio, Yining Liu

Outline

1. Background & Review
   - Single-cell Transcriptomics
   - Bayesian Modeling
   - Deep Generative Models
2. Single-cell Variational Inference (scVI)
3. Probabilistic Annotation (scANVI)
4. Information constraints on Auto-Encoding Variational Bayes (HCV)
5. Decision-making with Auto-Encoding Variational Bayes
6. Open-source scientific research: making VI more accessible

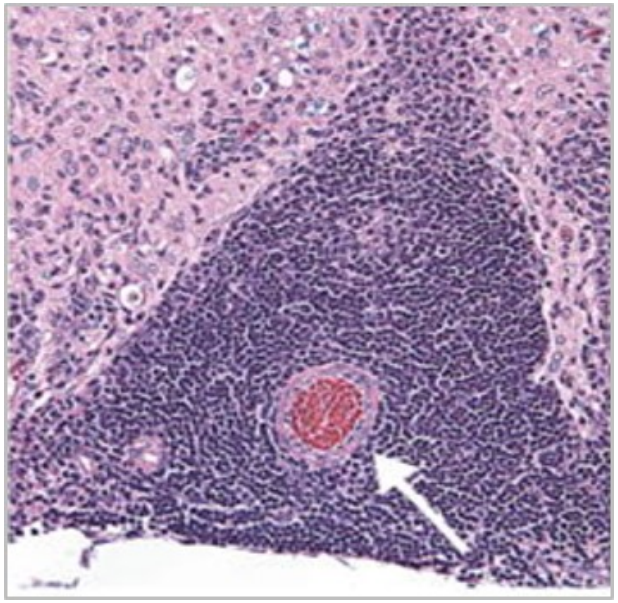
Outline

Background & Review

Single-cell Transcriptomics

Cells share the same DNA but have distinct functions

Complex Tissue



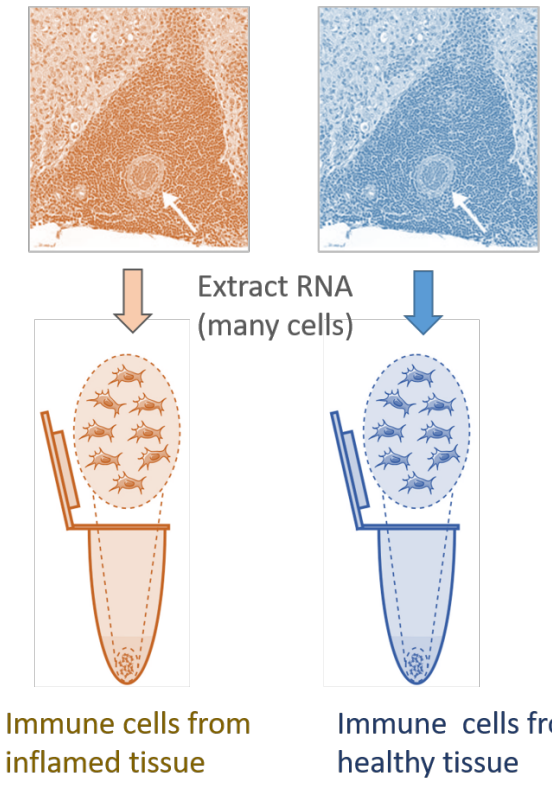
H&E stain of the spinal cord of mice with a multiple sclerosis model.

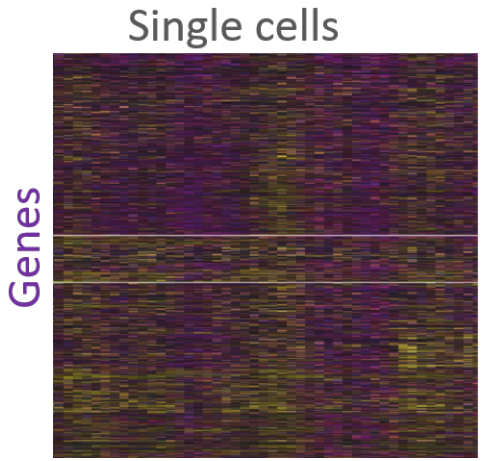
Peters et al, Immunity (2011)

Biological questions

- What type of cells are present in the tissue?
- Which functions do these cells carry?
- How are these functions different from the healthy tissue?

scRNA-seq measures gene expression at the cellular scale





scRNA-seq datasets
(one displayed here)

Several exciting technological advances:

- Sequencing of one million cells (10x Genomics, 2017)
- Multi-modal data: CITE-Seq (2017), Slide-seq (2019) and others

scRNA-seq yields quantitative answers to biological questions

Most biological questions can be casted into either **cell-level** or **gene-cell-level** algorithmic queries.

Comp. Bio. Tasks	Definition
Stratification via Embedding	Project the cells for identification (e.g., clustering/trajectory analysis)
Harmonization	Provide a batch-effect-free embedding to compare across conditions
Annotation	Transfer cell type labels from one dataset to another
Normalization/Imputation	Compute average expression levels while removing technical artifacts
Differential Expression	Find gene expression discrepancies between cell types

Overarching goal: probabilistic stratification and annotation of single-cell transcriptomes

Approach: learning cell-level and gene-cell-level similarity while correcting for technical biases

scRNA-seq data analysis remains challenging

1. scRNA-seq measurements are affected by technical noise
    - variable sequencing depth
    - batch-effects
2. Data is generated from a multivariate count distributions (non-Gaussian measurements)
3. Analysis requires scalable methods

scRNA-seq workflows combine many standard ML methods

1. **Normalize** the data adequately; there exists at least 30 possible combinations,
2. **Reduce the dimension** of the data (e.g., using PCA),
3. **Apply an ad-hoc algorithm** to correct for batch-effects,
4. **Cluster** the data to identify cell states,
5. Perform **differential expression** to match the clusters to known cell types (e.g., using DESeq2 on the raw counts),

How to find *unifying modeling assumptions* across the whole pipeline?

Starting point: improving the PCA

$$
\begin{aligned}
z &\sim \mathrm{Normal}(0, I) \\
x \mid z &\sim \mathrm{Normal}(Wz + v, \sigma^2 I)
\end{aligned}
\tag{1}
$$

Probabilistic intepretation of PCA suggests why it is inadequate for scRNA-seq data:

1. The expression levels are **not Gaussian**: data must be normalized.
2. There is **no basis for assuming linearity** between latent variables and gene expression levels.
3. PCA is for $\sigma^2 \to 0$. This is **not a fully probabilistic model** and cannot carry uncertainty of the measurements.

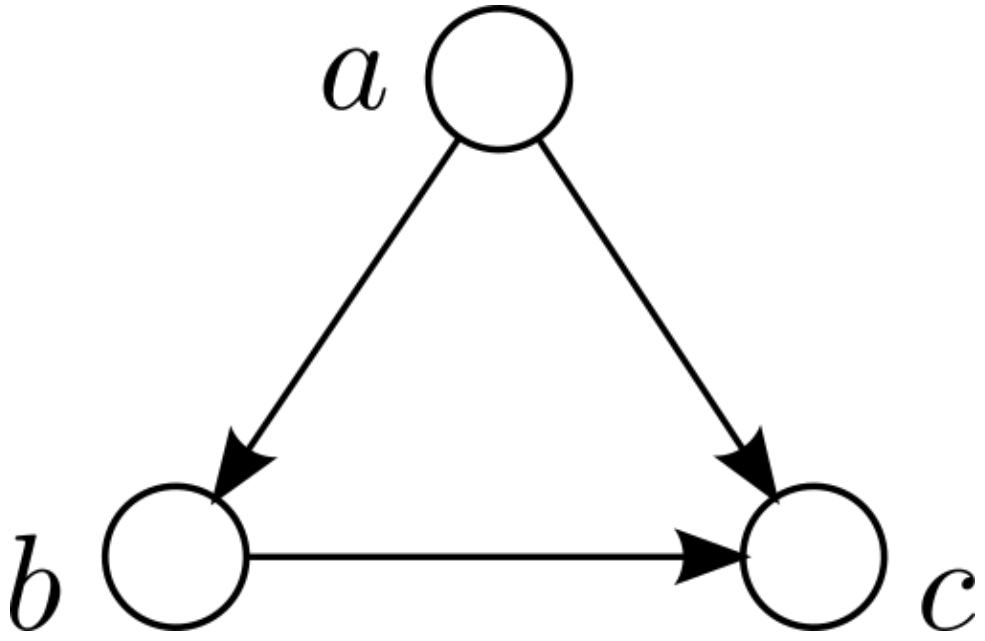
Room for improvement: a scalable and consistent framework for fully-probabilistic analysis of scRNA-seq data

scVI: a deep generative model that addresses all tasks and scales easily by leveraging stochastic optimization

Background & Review

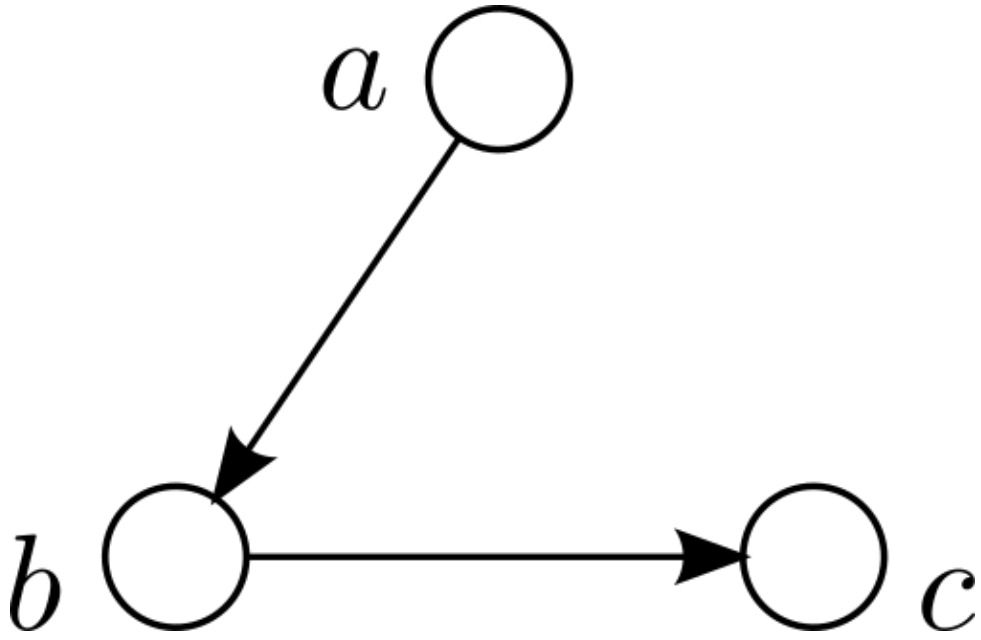
Bayesian Modeling

A graphical model shows a factorization of a joint distribution



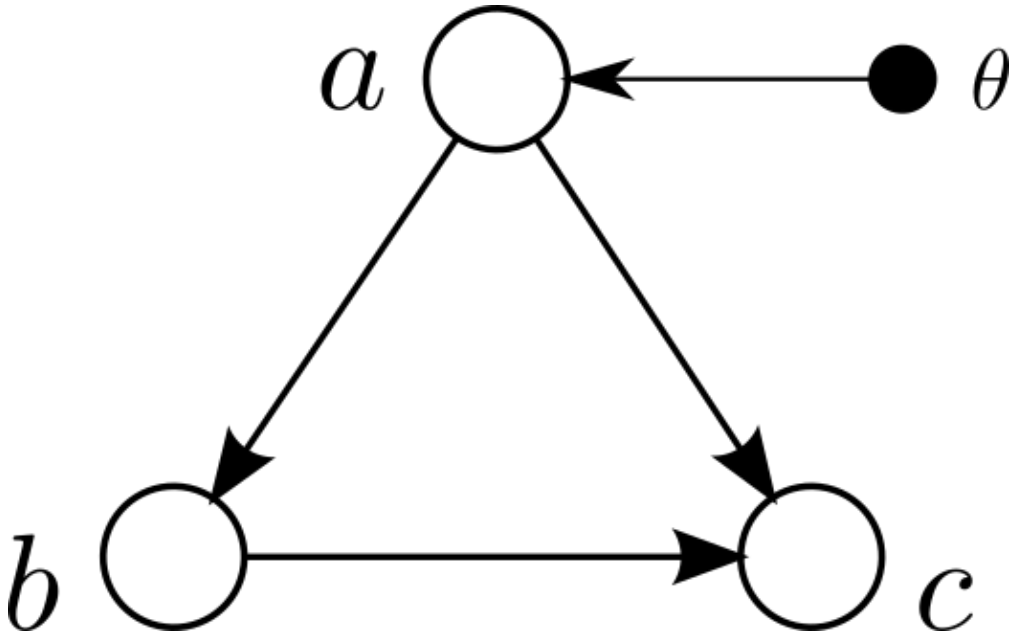
$$p(a, b, c) = p(c \mid a, b)p(b \mid a)p(a)$$

Omit edges to represent conditional independence



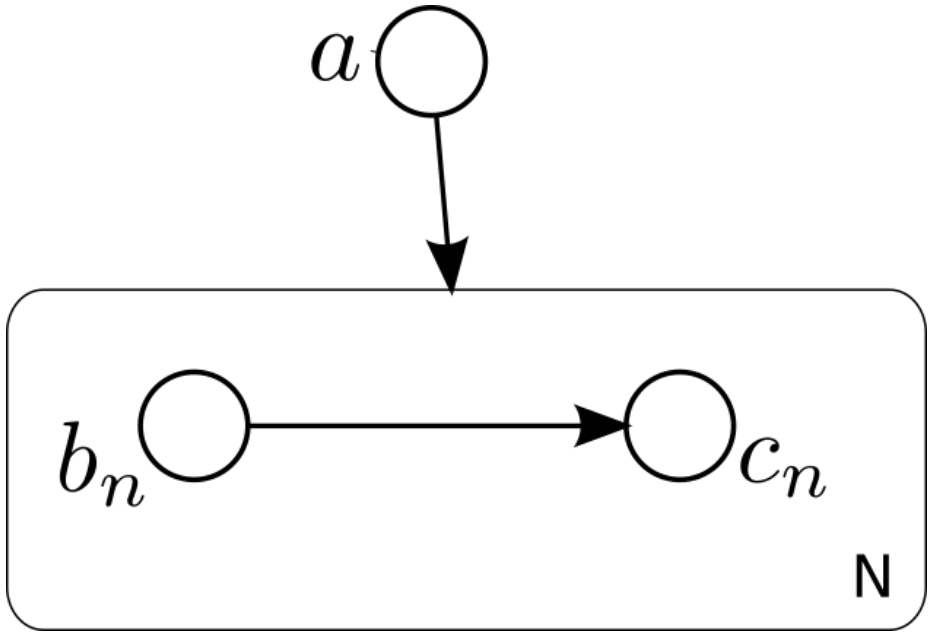
$$p(a, b, c) = p(c \mid b)p(b \mid a)p(a)$$

Solid dots represent unknown constants



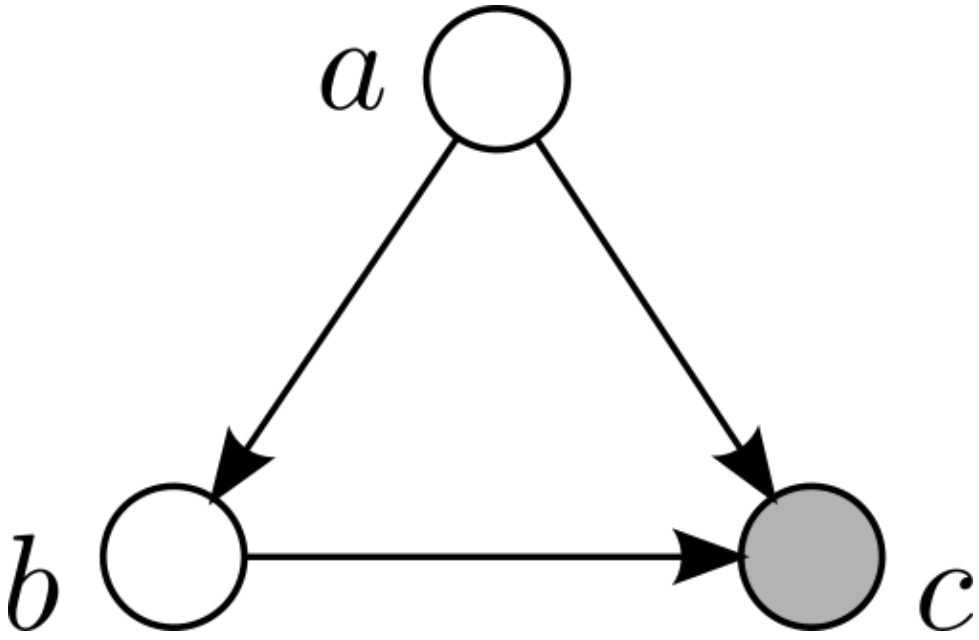
$$p(a, b, c) = p(c \mid a, b)p(b \mid a)p_\theta(a)$$

Rectangles denote independent replication



$$p(a, \mathbf{b}, \mathbf{c}) = \prod_{n=1}^{N} \left[p(c_n \mid a, b_n) p(b_n \mid a) \right] p(a)$$

Shaded nodes are observed.
Empty nodes are latent.



$$\underbrace{p(a, b \mid c)}_{\text{posterior}} \propto \underbrace{p(c \mid a, b)}_{\text{likelihood}} \underbrace{p(b \mid a)p(a)}_{\text{prior}}$$

Bayes rule usually cannot be applied

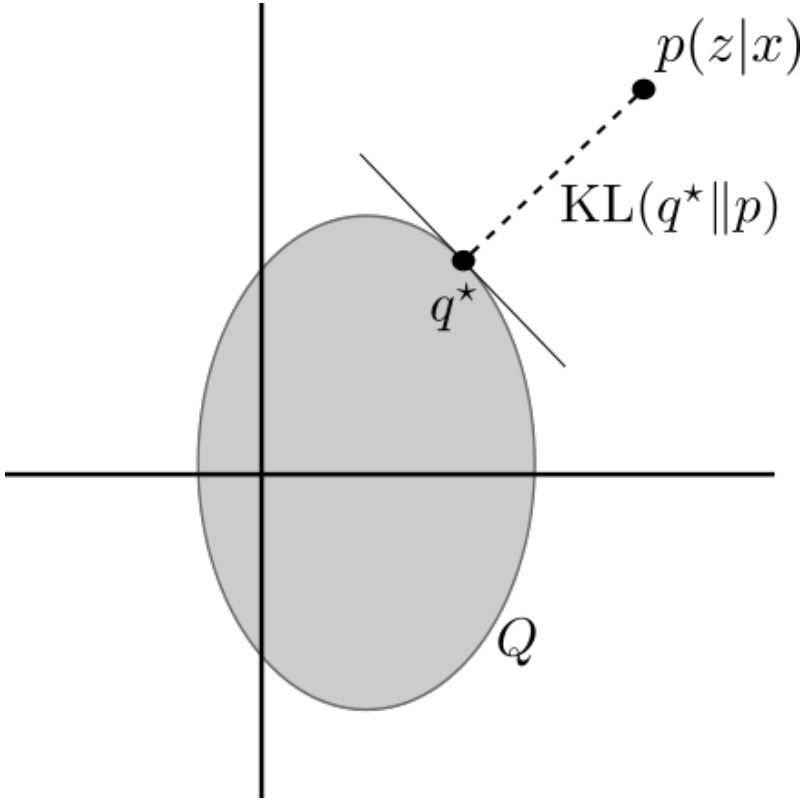
Let z denote the latent random variables. The posterior distribution of z typically is intractable:

$$p(z \mid x) = \frac{p(x \mid z)p(z)}{p(x)}$$

where

$$p(x) = \int p(x \mid z)p(z)\, dz.$$

Variational inference approximates the posterior



We cast the inference problem into an optimization one!

Variational inference is based on a mathematical trick

The optimization problem can be written without $p(x)$ or $p(z \mid x)$:

$$q^\star = \arg\min_{q \in Q} \mathrm{KL}(\, q(z) \,\|\, p(z \mid x)\,) \tag{2}$$

$$= \arg\min_{q \in Q} \mathbb{E}_q \left[\log q(z) - \log p(z \mid x)\right] \tag{3}$$

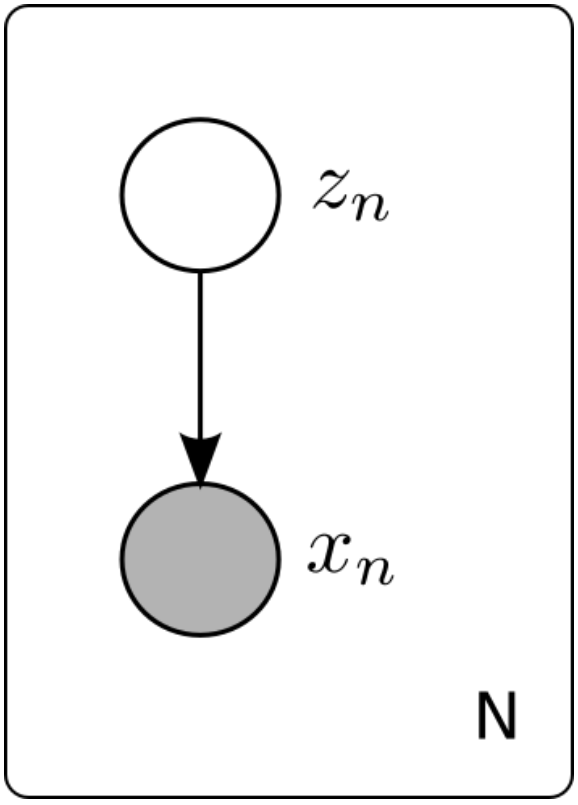
$$= \arg\min_{q \in Q} \mathbb{E}_q \left[\log q(z) - \log p(z, x)\right] + \log p(x) \tag{4}$$

$$= \arg\min_{q \in Q} \mathbb{E}_q \left[\log q(z) - \log p(z, x)\right]. \tag{5}$$

Background & Review

Deep Generative Models

Idea: Use neural networks to encode conditional probabilities!

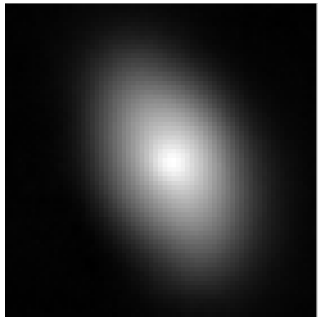


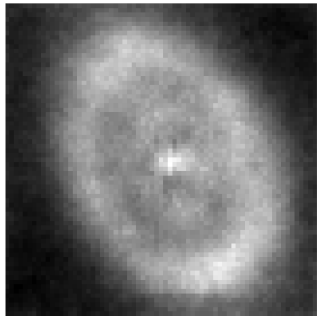
$$z_n \sim \mathcal{N}(0, \mathrm{I})$$

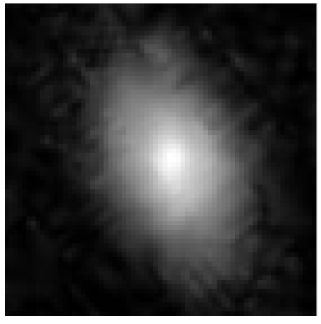
$$x_n \mid z_n \sim \mathcal{N}(\mu(z_n), \sigma(z_n))$$

Example

$$z_n = [0.1, -0.5, 0.2, 0.1]^\top$$

$\mu(z_n) =$ 

$\sigma(z_n) =$ 

$x_n =$ 

Variational inference maximizes a lower bound

Learning the model parameters

$$\log p(x) = \log \mathbb{E}_q \left[\frac{p(x,z)}{q(z \mid x)} \right] \geq \mathbb{E}_q \log \left[\frac{p(x,z)}{q(z \mid x)} \right].$$

VI maximizes this lower bound w.r.t. the parameters of $p(x,z)$ and $q(z \mid x)$.

Auto-encoding Variational Bayes With several observations, one must maintain a posterior approximation for each datapoint. Instead, we use neural networks to parameterize its *variational* approximation $q(z \mid x)$:

$$q(z \mid x) \sim \mathcal{N}(\hat{\mu}(x), \hat{\sigma}(x)).$$

This is usually referred to as "amortized VI" or AEVB.

Outline

scVI is a Bayesian model that separates biological signal from technical effects

The process that generated the gene expression count x_{ng} for a cell n, with batch identifier s_n and a gene g is

$$
\begin{aligned}
z_n &\sim \mathrm{Normal}(0, I) && \text{Cell embedding} \\
\ell_n &\sim \mathrm{LogNormal}(\ell_\mu, \ell_\sigma^2) && \text{Library size} \\
\rho_n &= f_w(z_n, s_n) && \text{Normalized expression} \\
x_{ng} &\sim \mathrm{NegativeBinomial}\left(\ell_n \rho_{ng}, \theta_g\right) && \text{Raw data}
\end{aligned}
$$

where f_w is a neural network with a softmax non-linearity in its last layer. z_n is made **invariant** to s_n as well as ℓ_n.

Lopez et al., *Nature Methods*, 2018

The same model can be used for several downstream analysis tasks

Our set of **modeling hypothesis** is now **common** to each task. This ensure reproducibility and avoid statistical artifacts.

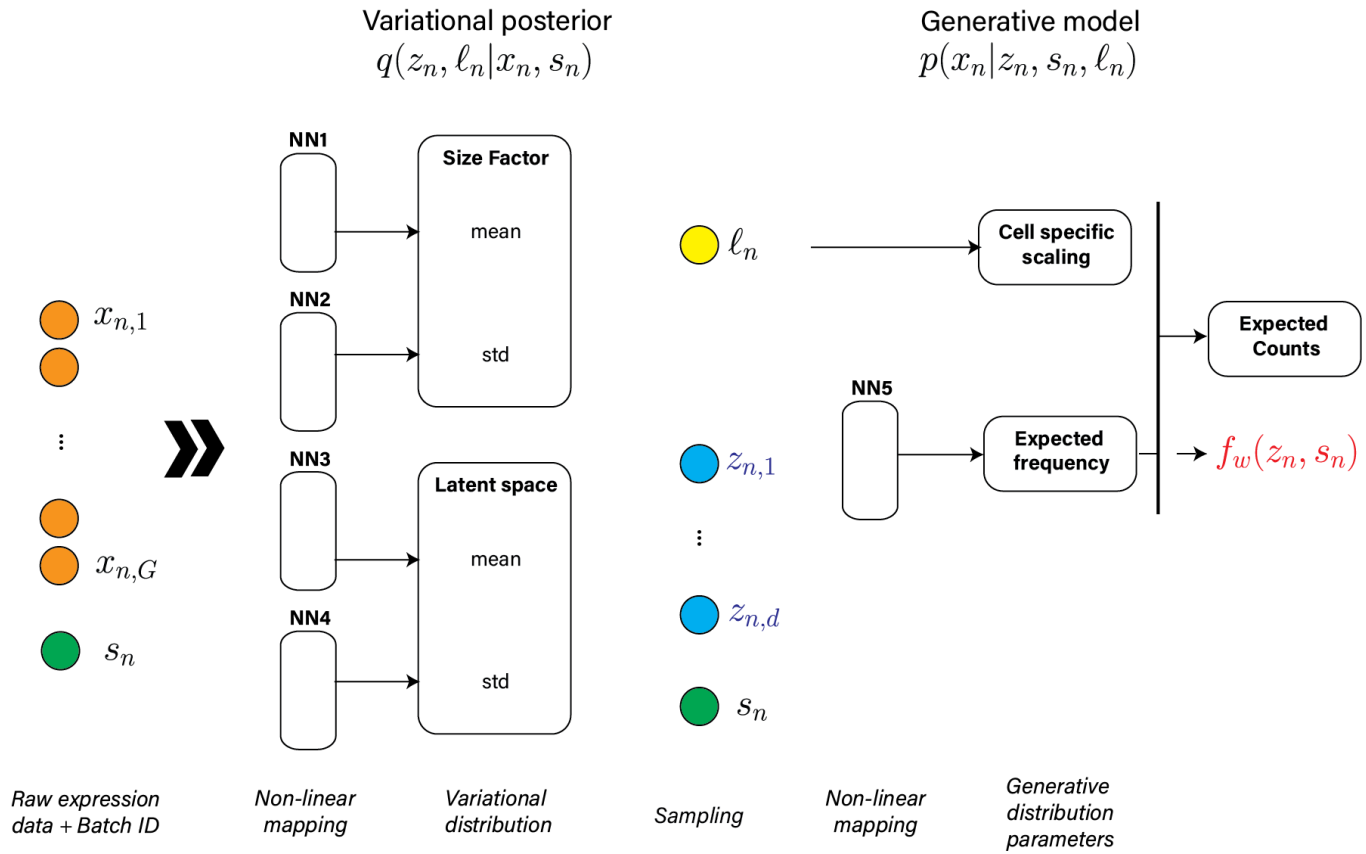
Comp Bio	**Bayesian Statistics**
Embedding	Posterior sampling for z_n
Harmonization	Conditioning on batch-information s_n
Normalization	Conditioning on hidden scalar ℓ_n
Imputation	Posterior sampling for ρ_{ng}
Differential Expression	Bayesian Hypothesis testing for ρ_{ng}

Lopez et al., *Nature Methods*, 2018

scVI is an algorithm

Inference can be done with Auto-encoding Variational Bayes!

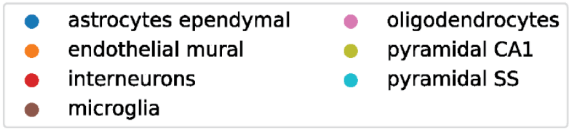
All CompBio tasks are well defined!



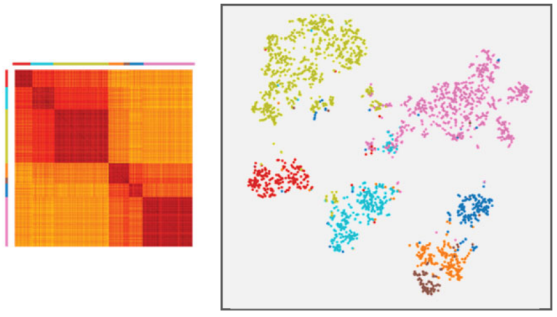
Lopez et al., *Nature Methods*, 2018

Latent variable z_n effectively recovers biological structure...

Hierarchical clusters:
~3k cortex cells
(Zeisel et al 2015)

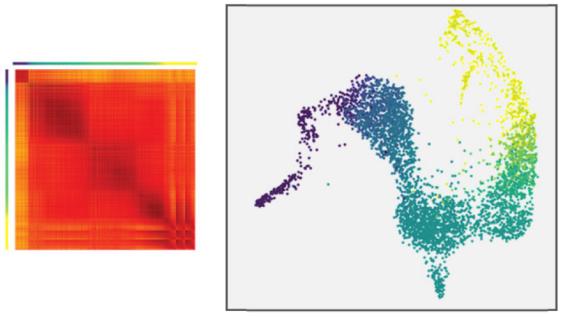


Cell- cell similarity (Matrix/ tSNE)



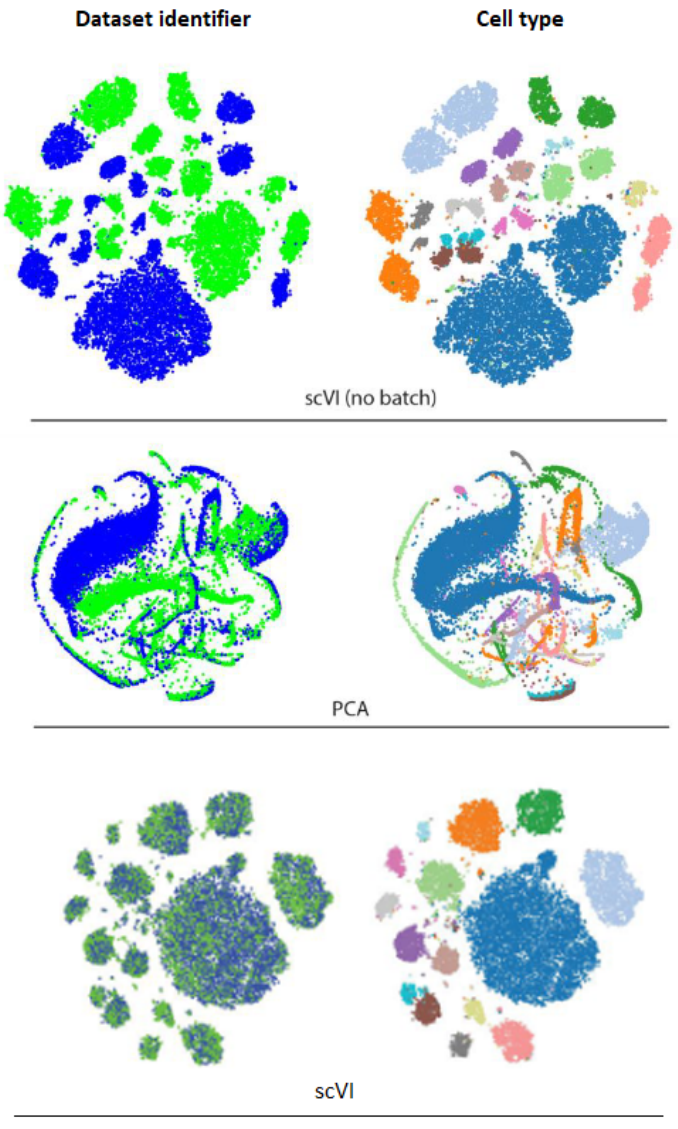
Developmental gradient:
hematopoiesis ~4k cells
(Tusi et al 2018)

Cell- cell similarity (Matrix/ tSNE)



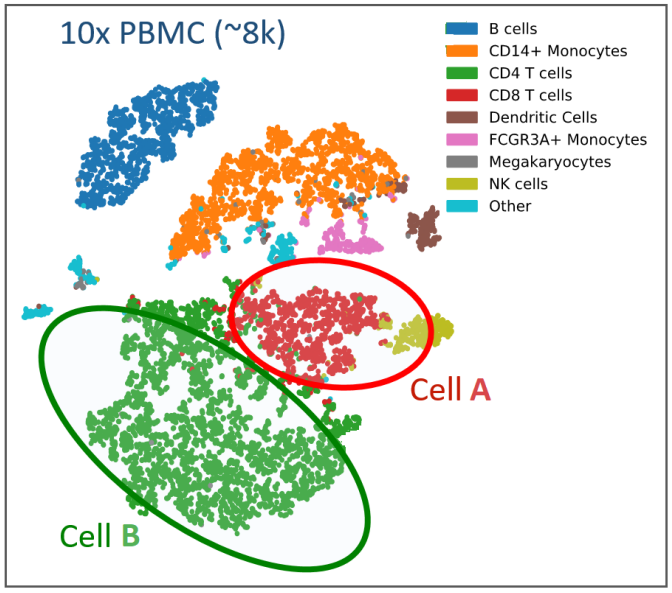
Versatile **stratification** thanks to the embedding $\mathbb{E}_{q(z_n|x_n)}[z_n]$.

... and corrects for batch-effects



Latent variable ρ_n permits differential expression assessment

Example with PBMCs



Bayesian estimates of fold-change

$$p\left(\left|\log \frac{\rho_{ag}}{\rho_{bg}}\right| > \delta \mid x_a, x_b\right)$$

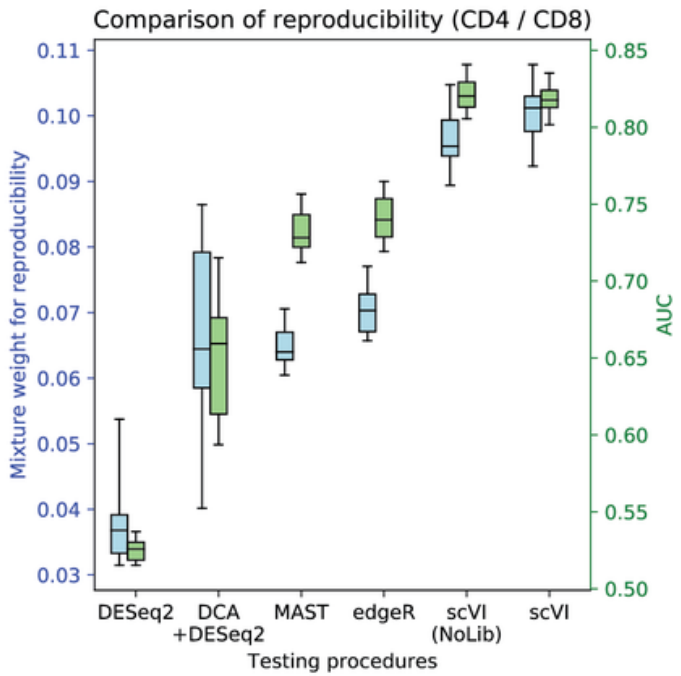
Approximated with $q(z \mid x)$

LFC estimation: Boyeau et al. (2019)

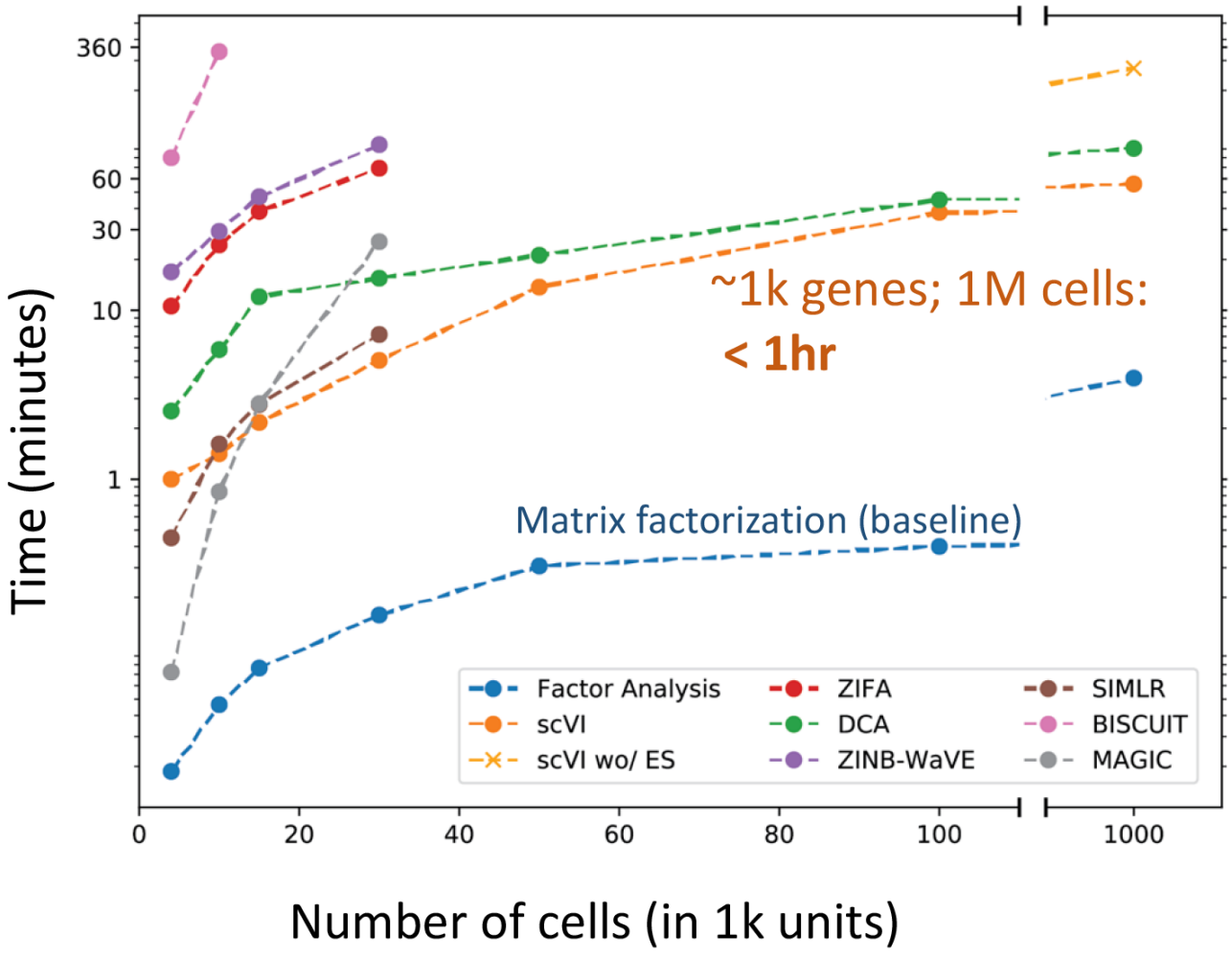
Evaluation

Reproducibility between scRNA-seq analysis and microarray reference

Results



scVI scales to large datasets



Dataset: 1.3M mouse cortex cells 10x Genomics

Outline

The problem of data annotation

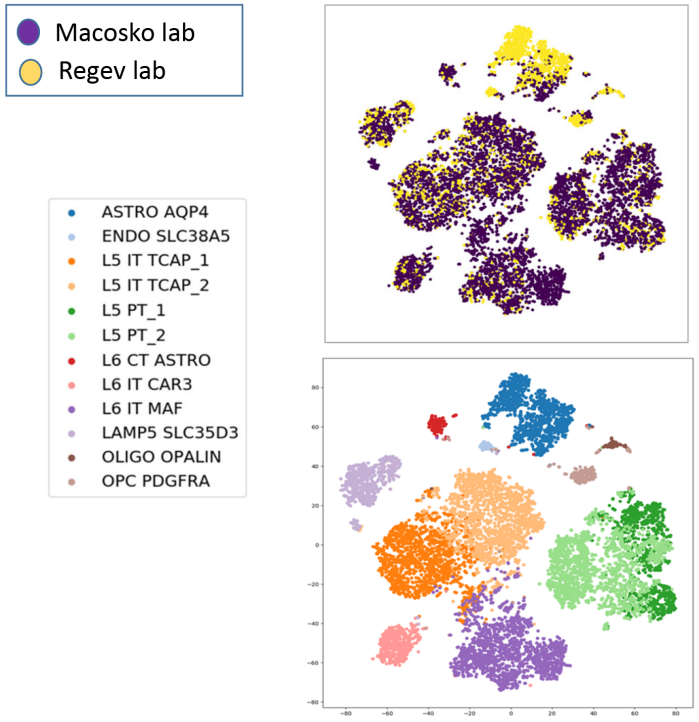
Data harmonization scenarios:

- Multiple samples
- Across labs
- Across technologies

Objective:

- Batch mixing
- Retain original structure
- *Use one data set to accurately annotate the other*

Formulation: a **domain adaptation** problem, with possible **semi-supervision**



Xu*, Lopez*, Melhman*, Regier, Jordan and Yosef, Molecular Systems Biology, 2021

Existing methods and limitations

Harmonization and Annotation are well studied in machine learning and computer vision, with applications to single-cell data

- Mutual Nearest Neighbors (Nature Biotechnology, 2018)
- Seurat Anchors (Cell, 2019)
- LIGER (Cell, 2019)

These methods make use of combinations of algorithms and heuristics (matching clusters or relying on PCA). These might require manual intervention to work well and cannot perform DE without querying the raw data.

We propose scANVI as an **end-to-end probabilistic method**.

scANVI annotates data sets using semi-supervised learning

Raw data

$$\begin{bmatrix} x_n \\ s_n \\ c_n^* \end{bmatrix}$$

Raw count matrix

Batch ID

Cell type ID (opt.)

$$\begin{bmatrix} 0 & 1 & \dots & 0 \\ 2 & 0 & \dots & 1 \\ \vdots & \vdots & \ddots & \\ 0 & 3 & \dots & 1 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \dots & 1 \end{bmatrix} \quad \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 2 \end{bmatrix} \quad \begin{bmatrix} 0 \\ \text{na} \\ \vdots \\ 5 \\ \vdots \\ \text{na} \end{bmatrix}$$

x_n s_n c_n^*

Annotation Scenarios

- One dataset partially annotated
- Transfer labels across datasets

Motivation De-novo annotation is tedious and prone to errors

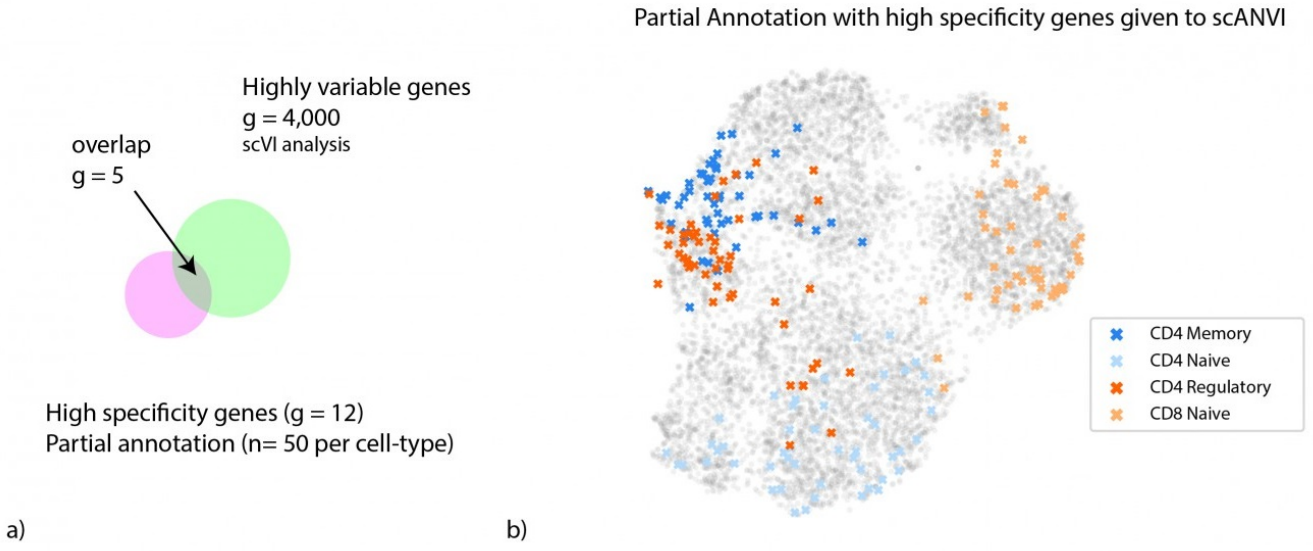
scANVI's approach

- Change the prior for z from isotropic normal to a mixture model
- Treat the semi-supervision as a missing value problem for mixture assignment (i.e., cell type) c

Xu*, Lopez*, Melhman*, Regier, Jordan and Yosef, Molecular Systems Biology, 2021

Application of scANVI to propagation of seed labels in T cells

Seed labels based on specific genes (0.5% annotated)

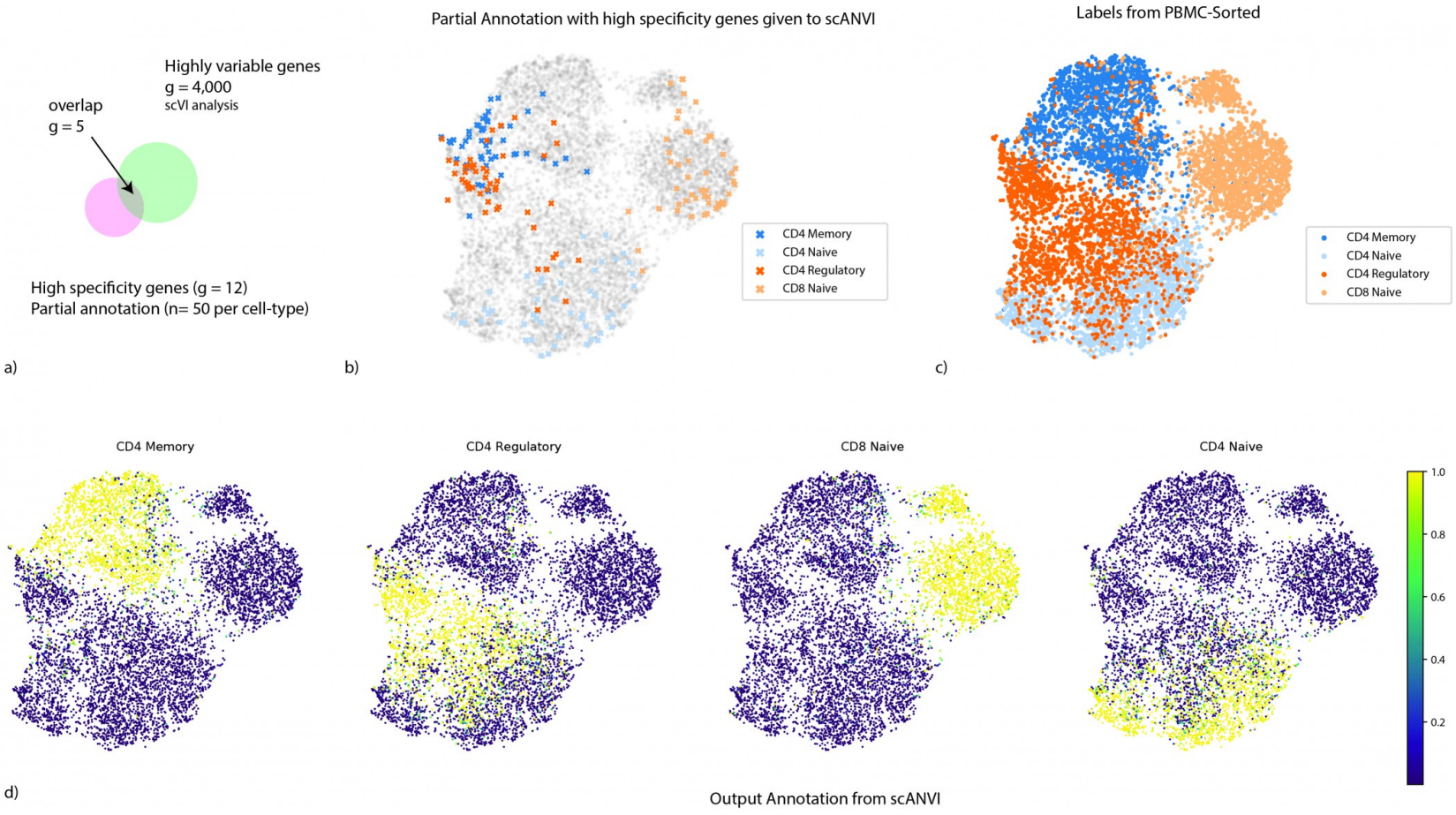


Possible approaches

- Supervised learning
- Clustering plus majority assignment
- Semi-supervision with scANVI

Xu*, Lopez*, Melhman*, Regier, Jordan and Yosef, Molecular Systems Biology, 2021

Application of scANVI to propagation of seed labels in T cells

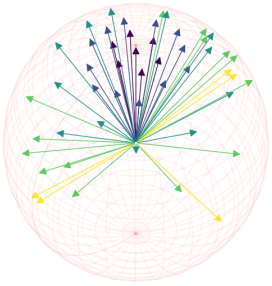


Xu*, Lopez*, Melhman*, Regier, Jordan and Yosef, Molecular Systems Biology, 2021

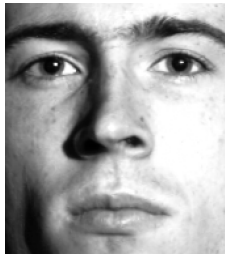
Outline

The problem of learning invariant representations

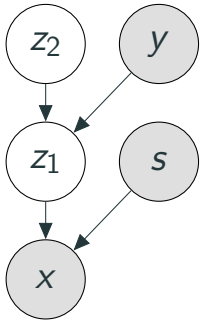
- How to obtain a representation $\mathbb{E}_{p_{\text{data}}(X)}\left[q(Z \mid X)\right]$ independent from an observed nuisance parameter S ?
- Statistician's answer: Condition on S and learn $p(Z \mid X, S)$!



s: angle between the camera and the light source



One image x for a given lighting condition s and person y



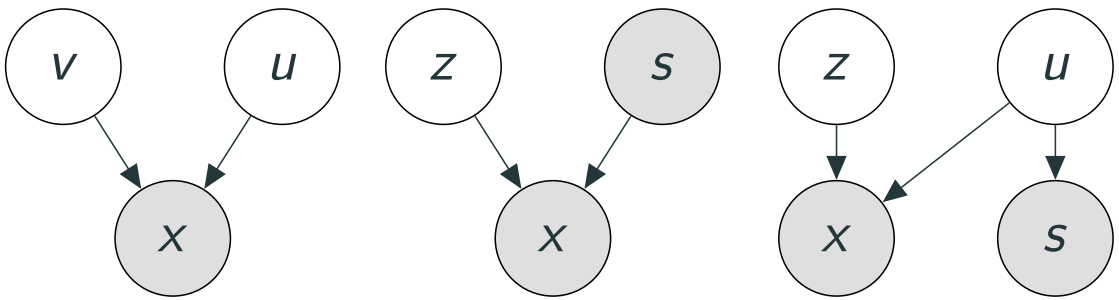
Complete graphical model

- In practice: $\mathbb{E}_{p_{\text{data}}(X,S)}\left[q(Z \mid X, S)\right]$ is still correlated with S

Lopez et al., Neural Information Processing Systems, (2018)

AEVB might not be satisfactory for non-trivial modeling purposes

- Sampling from the *aggregated posterior* $\hat{q}(Z) = \mathbb{E}_{p_{\text{data}}(X)}\left[q(Z \mid X)\right]$ is common to recover a representation used for downstream analysis
- Graphical model assumptions of conditional independence might **not be respected** in the aggregated posterior $\hat{q}(Z)$ due to over-flexibility of neural networks (Louizos et al. 2015).
- Modeling instances :



Learning Interpretable Representations | Learning Invariant Representations | Learning Denoised Representations

Lopez et al., Neural Information Processing Systems, (2018)

Information constraints on AEVB

- We restrain the search space for the variational distribution: in particular, we wish to enforce statements of the form $q(u) \perp\!\!\!\perp q(v)$.
- **Problem**: any measure of mutual information is intractable from the current graphical model and its variational approximation.
- **Solution**: we compute on each mini-batch a non-parametric measure of dependence from kernel embedding of joint distributions :

$$-\lambda \widehat{\mathrm{HSIC}}(q(u, v)),$$

where $\widehat{\mathrm{HSIC}}$ is the empirical estimate of the Hilbert-Schmidt norm of the cross-covariance operator $\mathcal{C}_{q(u,v)}$ that embeds the joint.

We call this modification **HSIC Constrained VAEs (HCV)**.

Lopez et al., Neural Information Processing Systems, (2018)

Outline

scVI performs differential expression via Bayesian model selection.

Setup: Let (a, b) be two cells, (x_a, x_b) their respective measurements and (ρ_a, ρ_b) the normalized gene expression levels. For every gene g, we have at disposal two models of the world:

$$\mathcal{M}_1^g : \left| \log \frac{\rho_a^g}{\rho_b^g} \right| > \delta \quad \text{and} \quad \mathcal{M}_0^g : \left| \log \frac{\rho_a^g}{\rho_b^g} \right| \leq \delta. \tag{6}$$

In scVI, we select the most likely model based on the Bayes factor:

$$\mathbf{BF}_g = \frac{p_\theta \left(\mathcal{M}_1^g \mid x_a, x_b \right)}{p_\theta \left(\mathcal{M}_0^g \mid x_a, x_b \right)}, \tag{7}$$

and we define a gene g to be differentially expressed if $\mathbf{BF}_g > 10$.

Limitations of this approach

There are two clear limitations to this formulation:

1. this approach is potentially biased, as we use the variational distribution to compute the posterior probability of differential expression:

$$p_g := p_\theta\left(\mathcal{M}_1^g \mid x_a, x_b\right) \approx \mathbb{E}_{q_\phi(z_a|x_a)}\mathbb{E}_{q_\phi(z_b|x_b)}\mathbb{1}\left\{\left|\log\frac{\rho_a^g}{\rho_b^g}\right| > \delta\right\}. \quad (8)$$

2. applicability is limited as Bayes factor are not intuitive for practitioners. Rather, we would like to control the *posterior expected False Discovery Rate*.

Estimating posterior expectations with VAEs

Either problem reduces to calculating posterior expectations accurately, of the form:

$$\mathcal{Q}(f, x) = \mathbb{E}_{p_\theta(z|x)} f(z).$$

We have access to samples $(z_i)_{1 \le i \le n}$ from the variational distribution $q_\phi(z \mid x)$. A naive but practical approach is to consider a plugin estimator:

$$\hat{\mathcal{Q}}^n_{\mathrm{P}}(f, x) = \frac{1}{n} \sum_{i=1}^{n} f(z_i). \tag{9}$$

or as a proposal for self-normalized importance sampling (SNIPS):

$$\hat{\mathcal{Q}}^n_{\mathrm{IS}}(f, x) = \frac{\sum_{i=1}^{n} w(x, z_i) f(z_i)}{\sum_{j=1}^{n} w(x, z_j)}. \tag{10}$$

Here the importance weights are $w(x, z) := {p_\theta(x,z)}/{q_\phi(z|x)}$.

Estimating posterior expectations with VAEs (continued)

These approaches may fail for two reasons:

1. the model fit by the VAE may not be equal to the underlying data distribution,
2. there may be strong discrepancies between the variational distribution and the posterior.

More specifically, variational approximations often underestimate the variance of the posterior (Turner et al., 2011). Consequently, they may yield poor proposals for importance sampling.

Background: variational bounds for Bayesian inference

There are many decomposition of the evidence used for variational inference:

$$\underbrace{\log p_\theta(x)}_{\text{evidence}} = \underbrace{\mathbb{E}_{q_\phi(z|x)} \log \frac{p_\theta(x,z)}{q_\phi(z \mid x)}}_{\text{ELBO}} + \underbrace{\Delta_{\mathrm{KL}}(q_\phi \parallel p_\theta)}_{\text{reverse KL VG}}, \tag{VI}$$

$$\underbrace{\log p_\theta(x)}_{\text{evidence}} = \underbrace{\log \mathbb{E}_{p_\theta(z|x)} \frac{p_\theta(x,z)}{q_\phi(z \mid x)}}_{\text{EUBO}} - \underbrace{\Delta_{\mathrm{KL}}(p_\theta \parallel q_\phi)}_{\text{forward KL VG}}, \tag{RWS}$$

$$\underbrace{\log p_\theta(x)}_{\text{evidence}} = \underbrace{\frac{1}{2} \log \mathbb{E}_{q_\phi(z|x)} \left(\frac{p_\theta(x,z)}{q_\phi(z \mid x)} \right)^2}_{\text{CUBO}} - \underbrace{\frac{1}{2} \log \left(1 + \Delta_{\chi^2}(p_\theta \parallel q_\phi)\right)}_{\chi^2 \text{ VG}}. \tag{CHIVI}$$

One may also apply importance sampling to tighten some of those bounds (i.e., IWAE).

Our three-step procedure

We propose a simple three-step procedure for Bayesian decision-making with VAEs:

(a) Fit multiple VAEs, each with a different variational distribution (e.g., IWAE, RWS, CHIVAE);

(b) Keep the best model based on a surrogate of the likelihood;

(c) Learn several variational approximations to the model posterior;

(d) Estimate the optimal decision via multiple importance sampling;

Lopez et al., Neural Information Processing Systems, (2020)

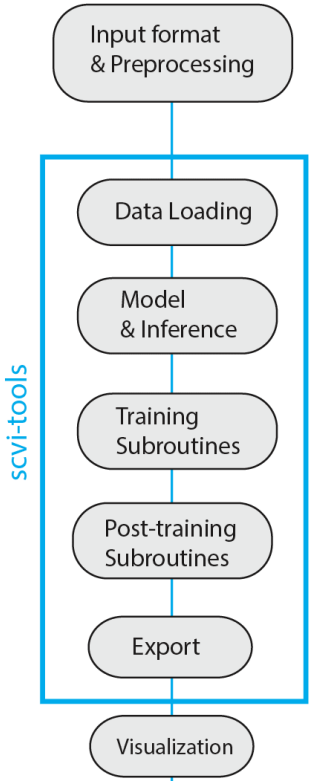
Other content of the paper

- A more general presentation of these ideas for Bayesian decision-making,
- Further theoretical developments on the pPCA model,
- Our novel formulation of the VAE learned via CHIVI,
- Supplementary experiments on pPCA and MNIST,
- A full-fledged application to differential expression.

Lopez et al., Neural Information Processing Systems, (2020)

Outline

We aim at making VI more accessible: scvi-tools is a public open-source repository



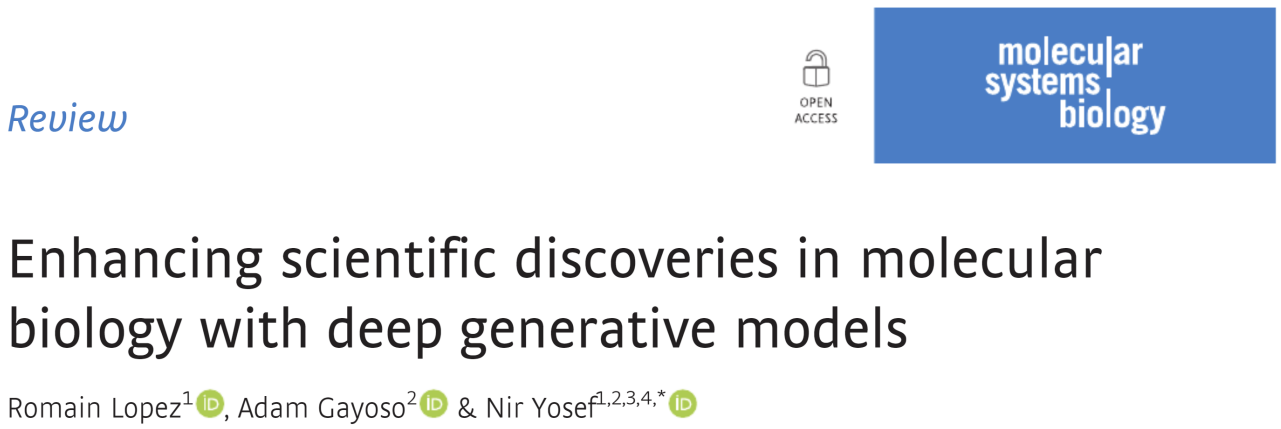
https://scvi-tools.org

- Our codebase contains multiple algorithms for single-cell omics analysis (scRNA-seq, CITE-seq, spatial transcriptomics, ATAC-seq) as well as tutorials;
- We conceived a high-level interface to probabilistic programming languages (Pyro, PyTorch). It is simple to prototype **new methods**;

Come contribute !

Gayoso*, Lopez*, Xing* et al. In Preparation, (2021)

We aim at making VI more accessible: review on DGMs for molecular biology



We recently published a review on applications of deep generative models in molecular biology.

Lopez et al. Mol Sys Bio, (2020)

References

Romain Lopez, Adam Gayoso and Nir Yosef
Enhancing Scientic Discoveries in Molecular Biology with Deep Generative Models.
Molecular Systems Biology, 2020

Chenling Xu*, Romain Lopez*, Edouard Mehlman*, Jeffrey Regier,
Michael I. Jordan and Nir Yosef
Probabilistic Harmonization and Annotation of Single-cell Transcriptomics data with Deep Generative Models.
Molecular Systems Biology, 2021

Romain Lopez, Jeffrey Regier, Michael Cole, Michael I. Jordan and Nir Yosef
Deep Generative Modeling for Single-cell Transcriptomics
Nature Methods, 2018

Romain Lopez, Pierre Boyeau, Nir Yosef, Michael I. Jordan and Jeffrey Regier
Decision-making with Auto-Encoding Variational Bayes.
Advances in Neural Information Processing Systems, 2020

Romain Lopez, Jeff Regier, Michael I. Jordan and Nir Yosef
Information Constraints on Auto-Encoding Variational Bayes.
Advances in Neural Information Processing Systems, 2018

For more, visit our codebase: `https://scvi-tools.org`

Email: romain_lopez@berkeley.edu

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. Beyond scRNA-seq: scATAC-seq, multi-omics
4. Dealing with noise, doublets, and other sc issues
5. Computational challenges in single-cell data analysis
6. Deep learning methods for single-cell data analysis
7. Guest lecture: Fabian Theis
8. Guest lecture: Romain Lopez