Transfer Learning in Single Cell Transcriptomics

ASA Non-Clinical Biostatistics Conference

June 17, 2019

Divyansh Agarwal

@divyansh_aga

Transfer Learning:

An approach for building (and estimating the parameters of) the model

What is the model for:

Denoising (Imputing) single cell RNA sequencing data

Single Cell Transcriptomics:

A relatively new technology which brings along new data challenges

What this talk is really about:

Getting you excited to think more about single cell data and our data denoising framework

Overview of my Talk

1. Single cell RNA sequencing (scRNA-seq)
    - Why is scRNA-seq data noisy? How can we address this problem?

450 ml

2% or whole?

+

10 g

Sliced or grated?

+

10 ml

+

2 tb sp

+

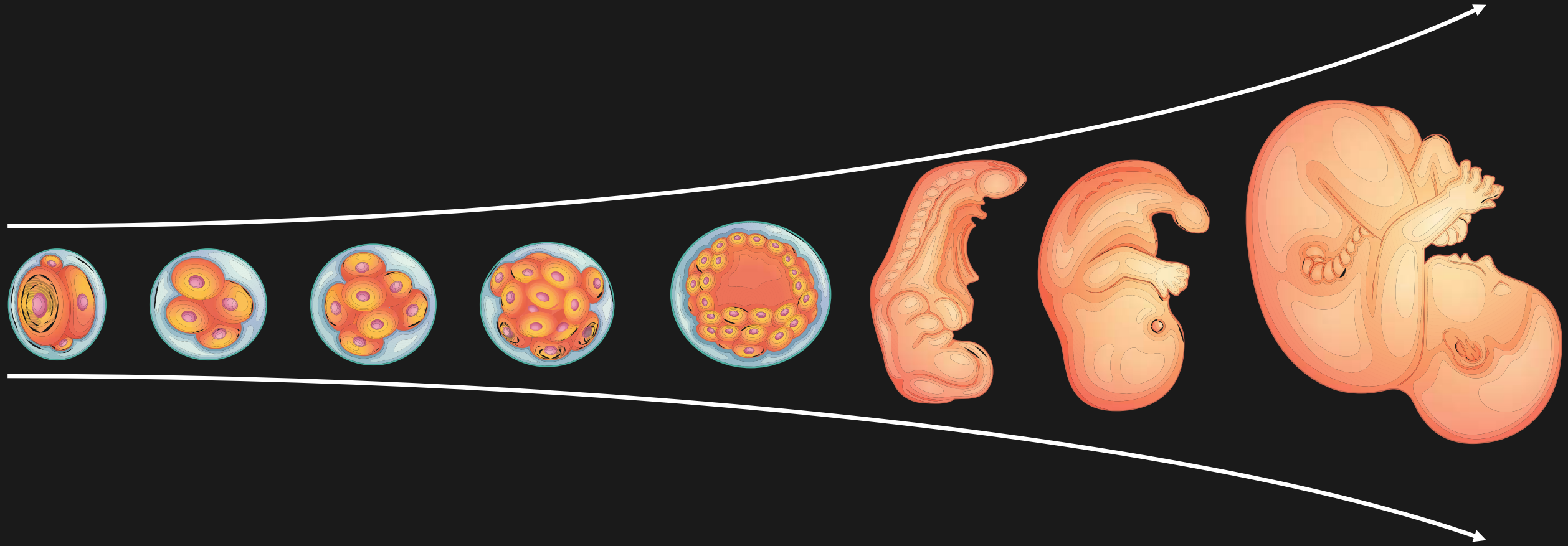
½ bowl

Vanilla/plain/fat-free?

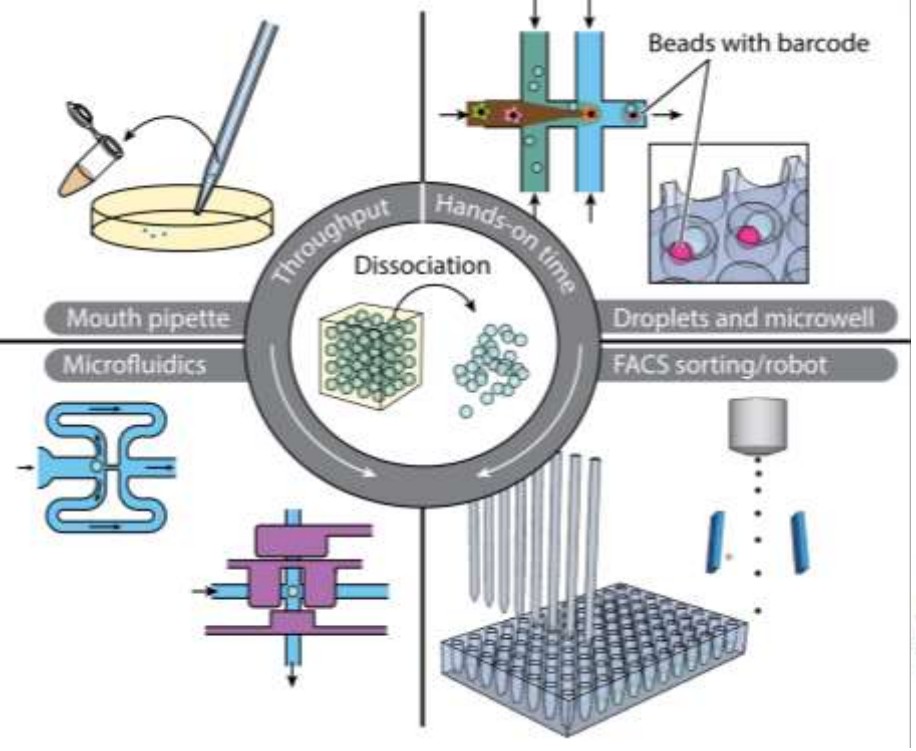
+

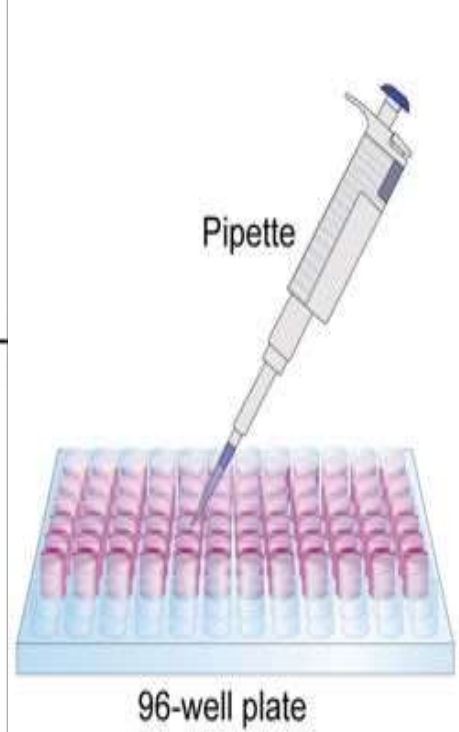
1 tb sp

At birth, we have well over 200 major cell types that constitute the human body

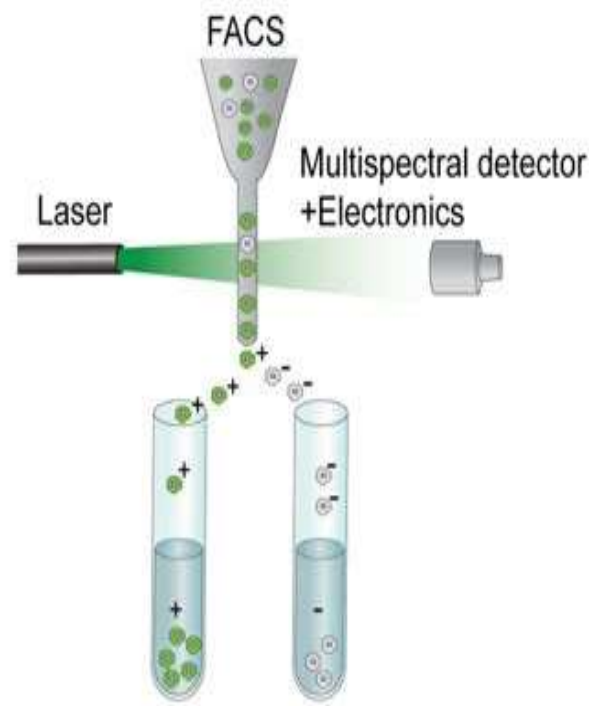


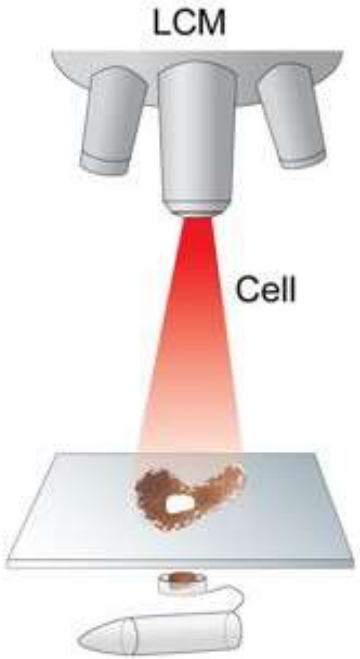
Just like appreciating a smoothie needs an understanding of its constituent ingredients, comprehending the complexity of life necessitates a grasp of its diverse cell type composition

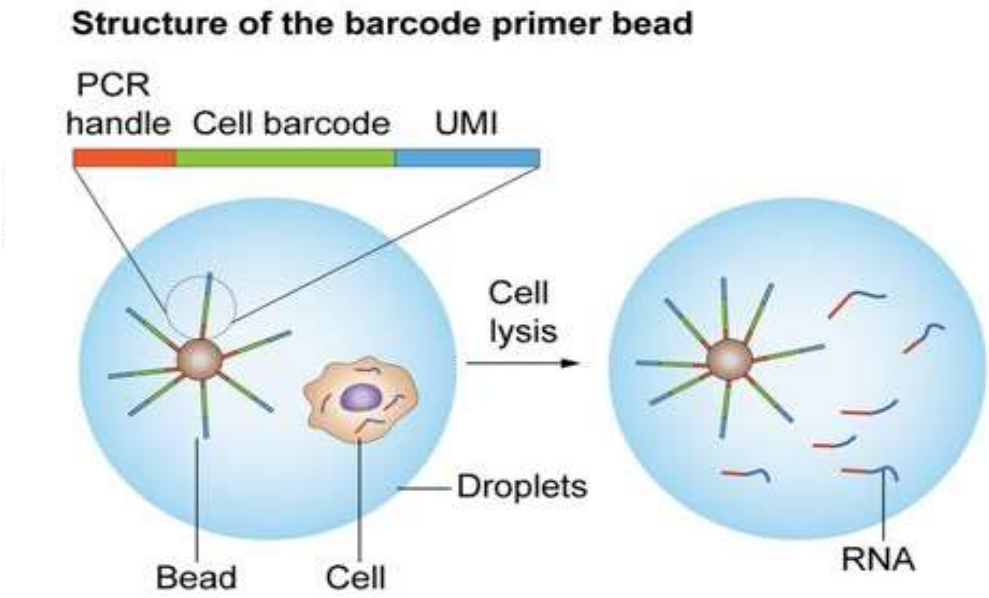

Beads with barcode
Throughput
Hands-on time
Dissociation
Mouth pipette
Microfluidics
Droplets and microwell
FACS sorting/robot

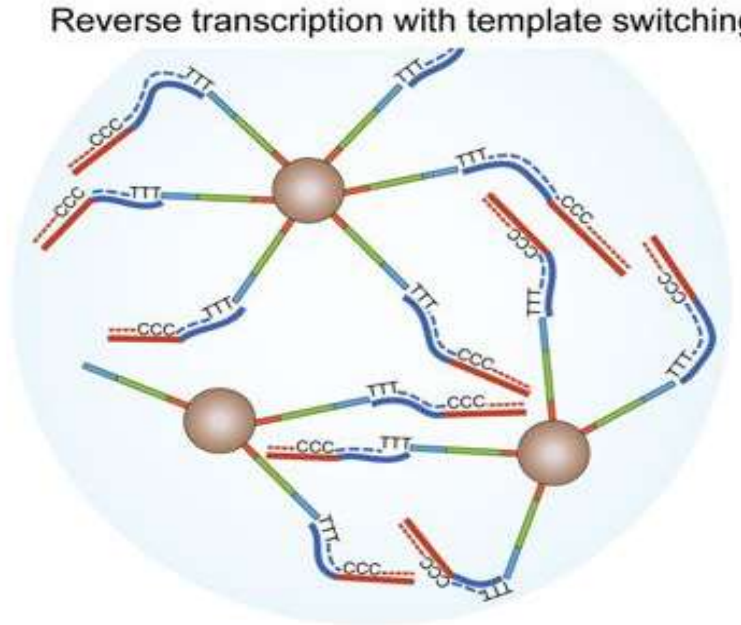

Pipette
96-well plate

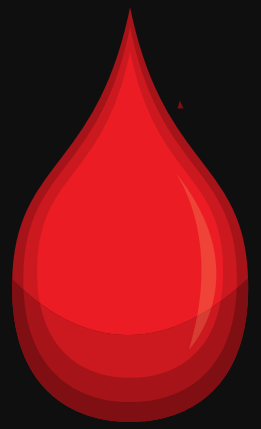

Microscope
Capillary pipette

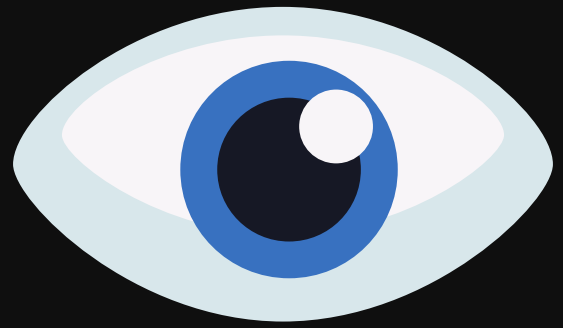

FACS
Laser
Multispectral detector
+Electronics

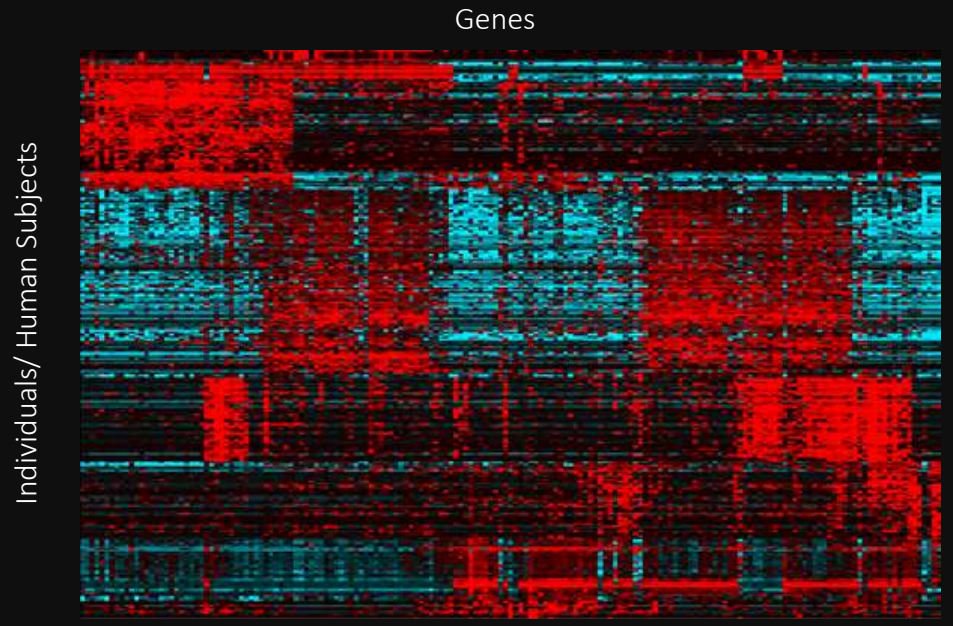

LCM
Cell

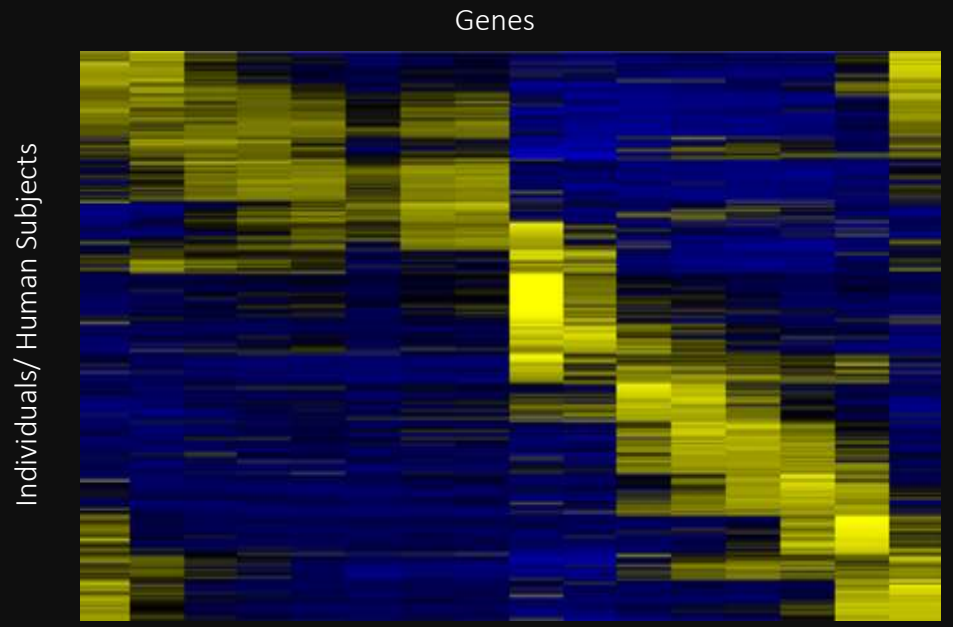

Structure of the barcode primer bead
PCR handle
Cell barcode
UMI
Cell lysis
Droplets
Bead
Cell
RNA
Break droplet


Reverse transcription with template switching



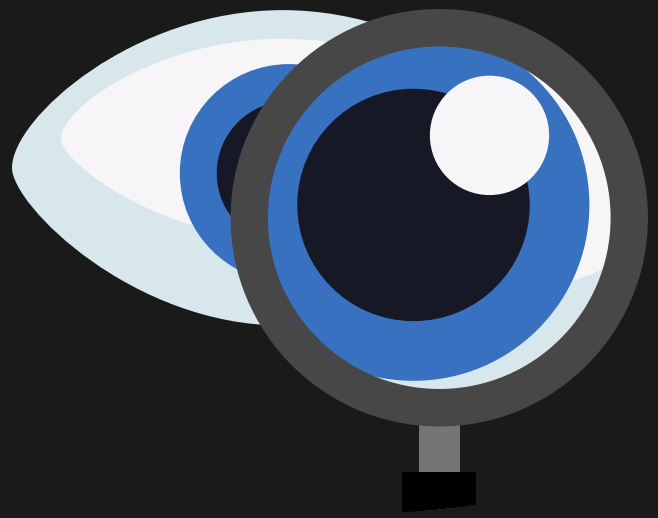


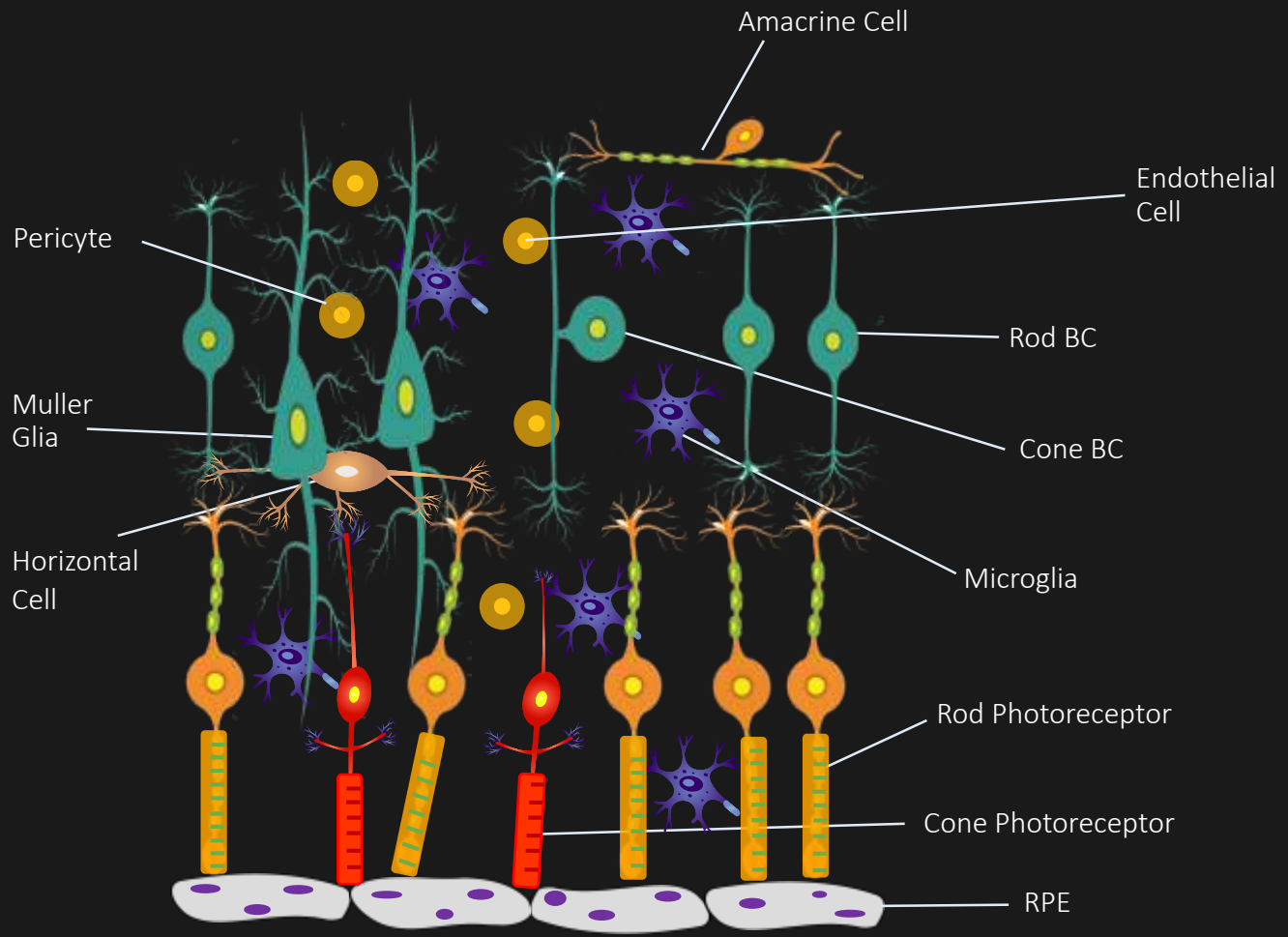




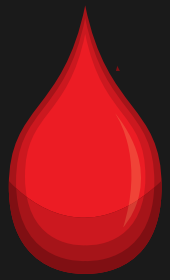
Cell types that make up the majority of a tissue dominate the gene expression patterns resulting from bulk, or while tissue, RNA sequencing

A granular understanding of gene expression required the ability to sequence individual cells



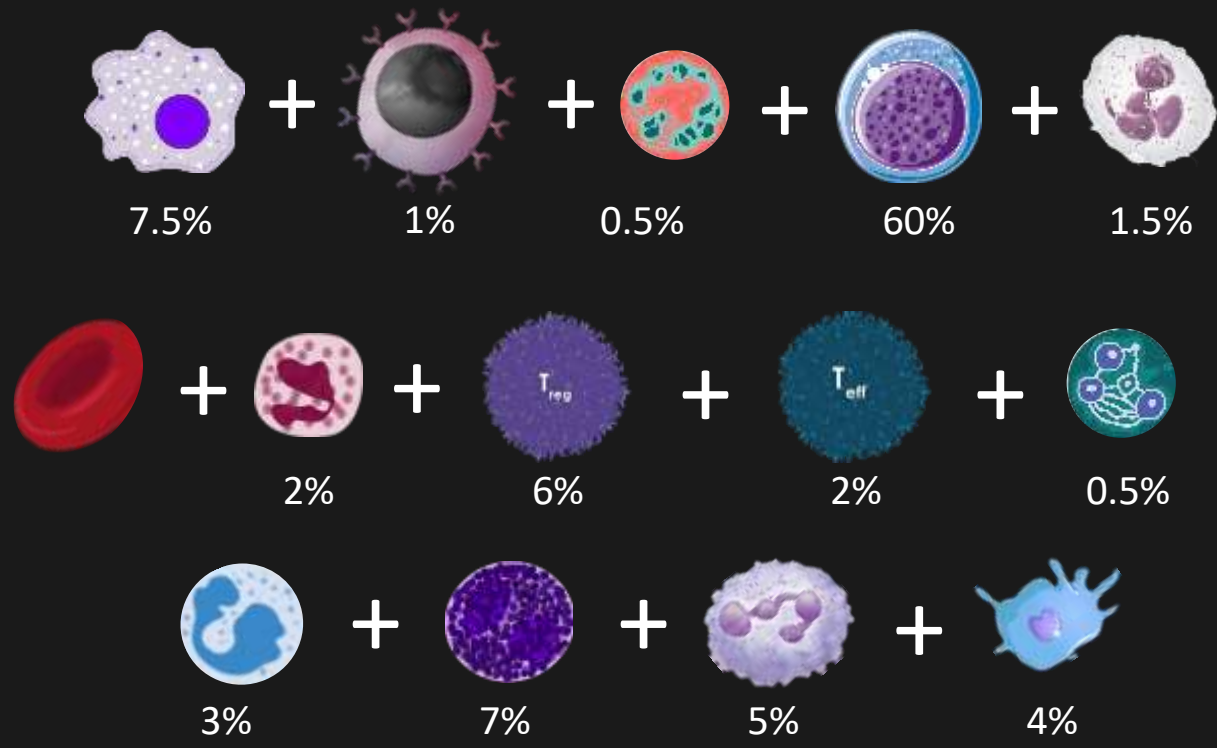


A New (Molecular) Microscope



450 ml + 10 g + ½ bowl + 2 tb sp + 1 tb sp + 10 ml

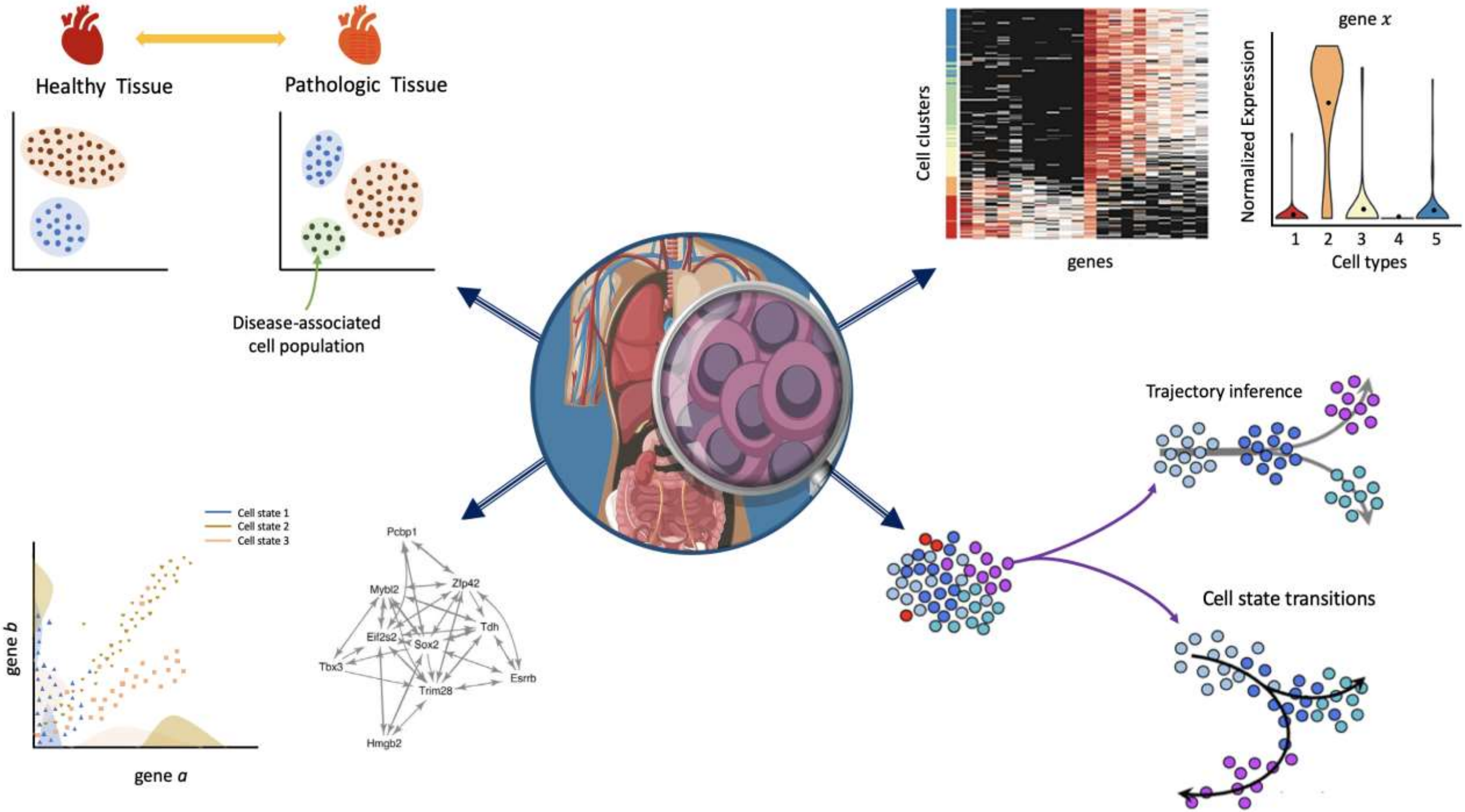
Single-cell transcriptomics



7.5% + 1% + 0.5% + 60% + 1.5%

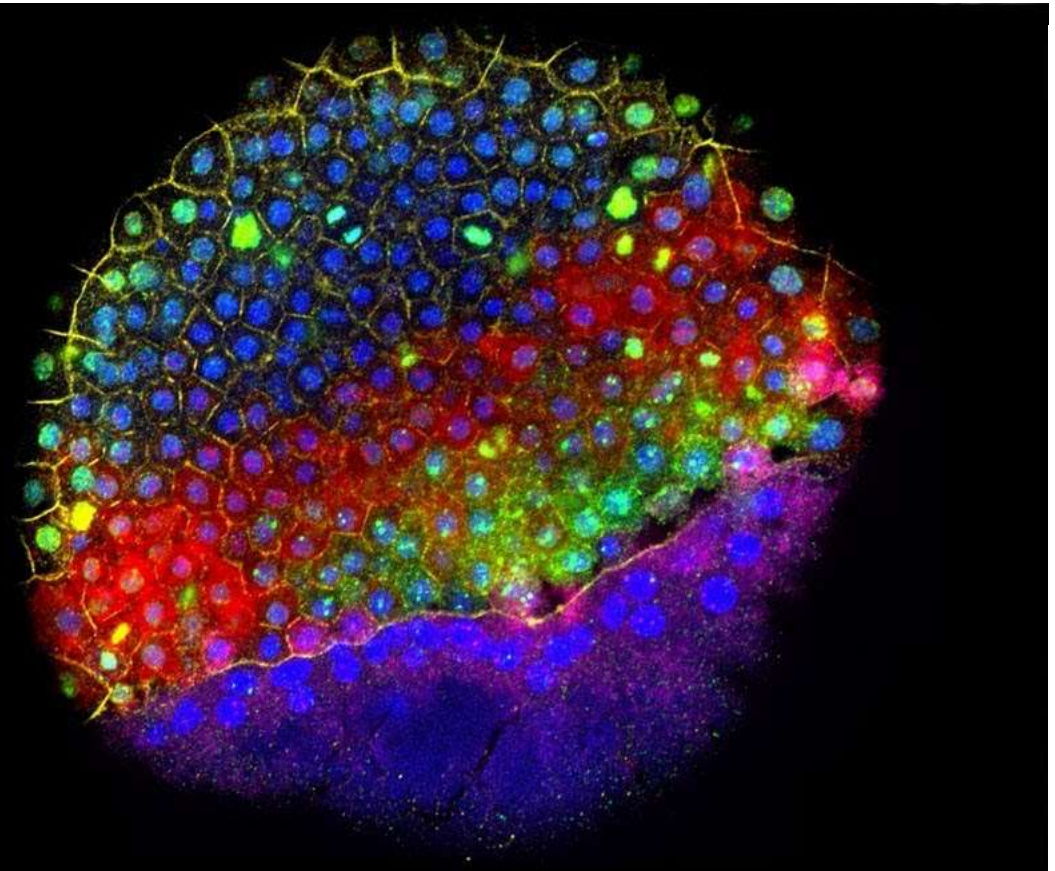
2% + 6% + 2% + 0.5%

3% + 7% + 5% + 4%


Healthy Tissue
Pathologic Tissue
Disease-associated cell population
Cell clusters
genes
gene x
Normalized Expression
1
2
3
4
5
Cell types
Trajectory inference
Cell state transitions
Cell state 1
Cell state 2
Cell state 3
gene b
gene a
Pcbp1
Mybl2
Zfp42
Tdh
Eif2s2
Sox2
Tbx3
Esrrb
Trim28
Hmgb2

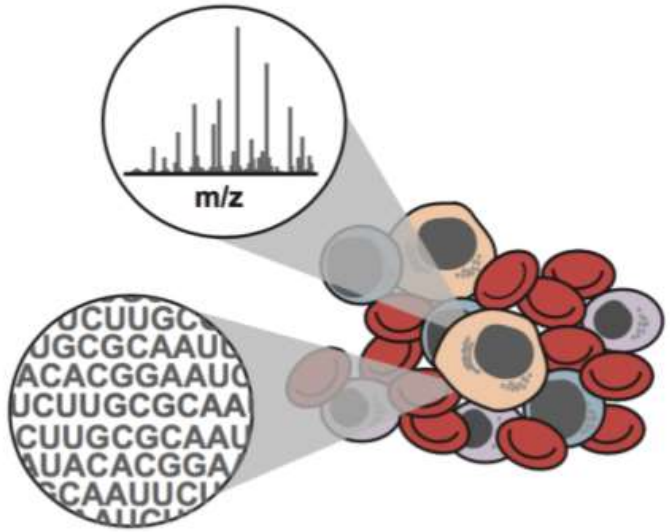
Single-cell profiling technologies → Single-cell data → Groundbreaking discoveries

Science AAAS

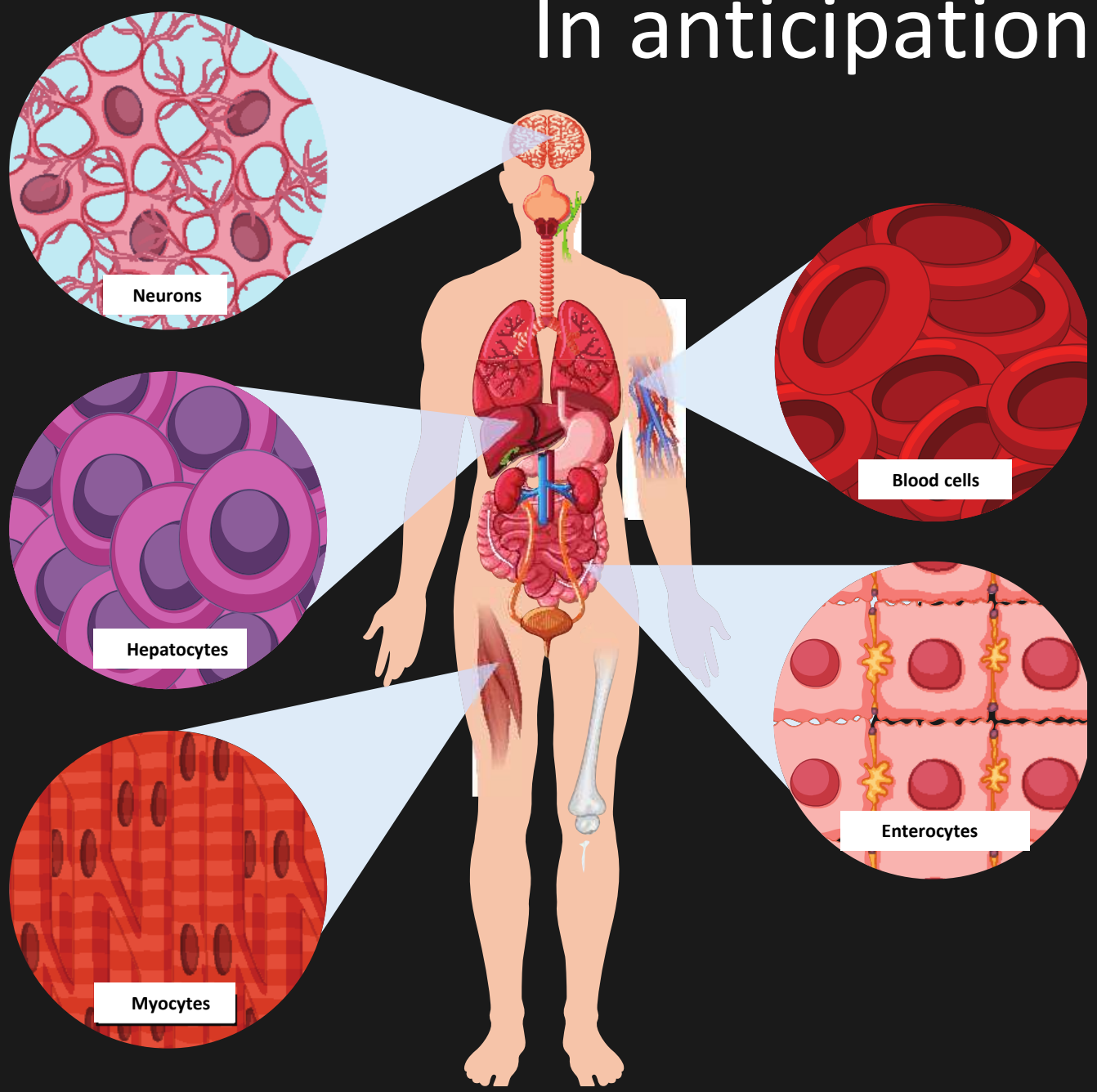


2018 BREAKTHROUGH *of the* YEAR

Biomedical expertise
Artificial intelligence
Complex data interpretation

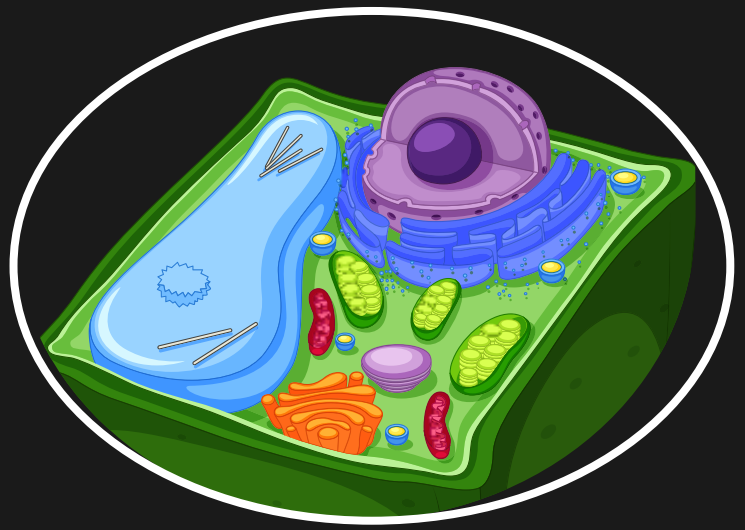


In anticipation of the Human Cell Atlas



Global efforts are underway to catalogue *all* the cell types and their transcriptomic patterns in the healthy human body

Single Cell experiments harbor multiple sources of *noise*



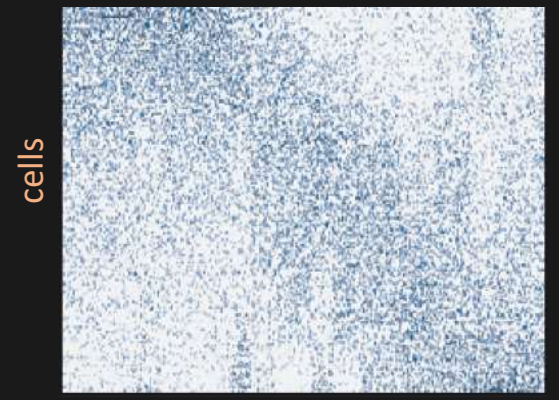
Some transcripts are lost during cell lysis

Some transcripts may not be converted to cDNA.

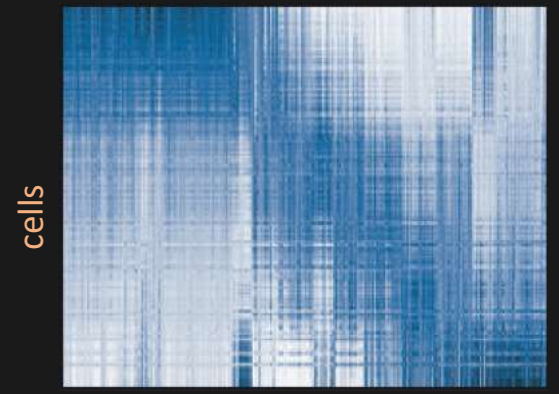
PCR amplification step introduces nonlinear biases

Some transcripts in library aren't sequenced.

What you observe:

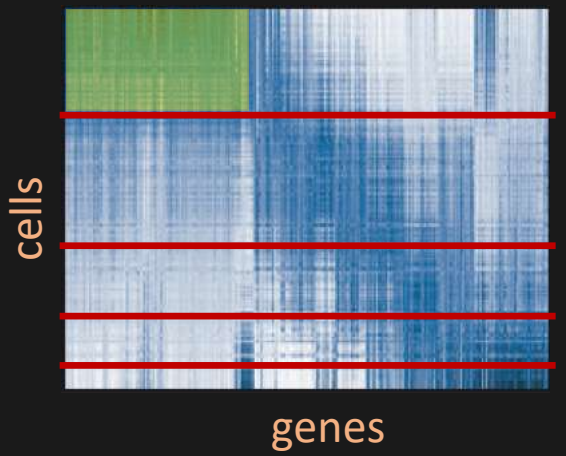


What the truth might be:



Why bother with denoising?

Because if this is the truth



We'd want to identify novel, rare cell types that associate with specific disease conditions and might be absent in an otherwise healthy individual

We'd like to characterize all the genes that mark a given cell population to be able to study them further, target them specifically, or isolate them using other experimental setups

Be careful what you denoise for

- Don't want false signals or oversmooth patterns
- Don't want to lose *true* biological information in the process

Overview of my Talk

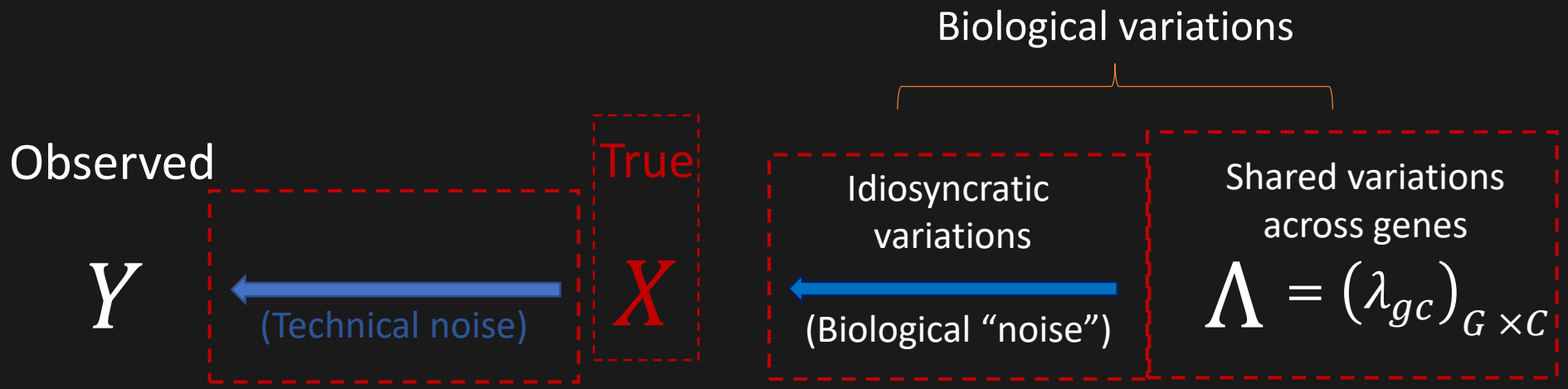
1. Single cell RNA sequencing (scRNA-seq)
   - Why is raw scRNA-seq data noisy? How can we address this problem?

2. The ideas underlying our proposed solution
   - Exploring the power (and the limits) of transfer learning

Single-cell **A**nalysis **V**ia **E**xpression **R**ecovery by harnessing e**X**ternal data:

SAVER-X

Our model setup is intuitive and comprehensive



$Y_{gc}|X_{gc} \overset{\text{ind}}{\sim} F_{gc}(X_{gc})$

$X_{gc}|\Lambda \overset{\text{ind}}{\sim} H_g(\lambda_{gc})$ $\Lambda \sim$ some structure

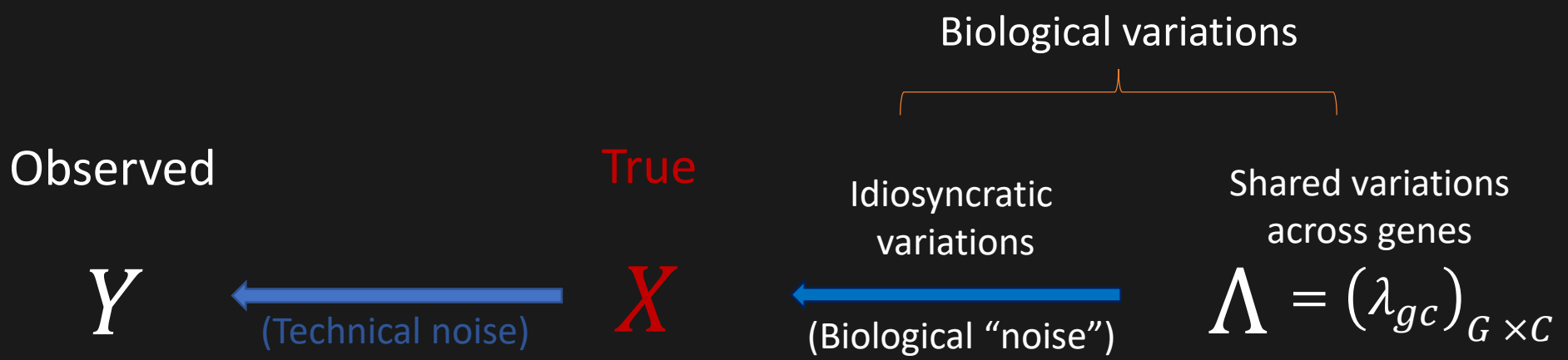
Biological variations:

- Shared variations across genes
- Purely random, unpredictable variations
  - Stochastic gene expression and its consequences [*Cell, 2008*]
  - Functional roles for noise in genetic circuits [*Nature, 2010*]

Biological noise is not just NOISE

Learning across similar genes

Our model setup is intuitive and comprehensive



$Y_{gc}|X_{gc} \overset{\text{ind}}{\sim} F_{gc}(X_{gc})$ $\quad X_{gc}|\Lambda \overset{\text{ind}}{\sim} H_g(\lambda_{gc})$ $\quad \Lambda \sim$ some structure

$$Y_{gc}|X_{gc} \overset{\text{ind}}{\sim} \text{Poisson}(l_c X_{gc})$$

Cell-specific size factor (known)

Poisson-alpha is well-suited to model technical noise

Validating the noise model

$$Y_{gc}|X_{gc} \overset{\text{ind}}{\sim} \text{Poisson}(l_c X_{gc})$$

l_c: Cell-specific efficiency constant (known)

Find a situation where true distribution is ***known:***

$$X_{gc} \overset{i.i.d}{\sim} H_g$$

?

Technical noise

Observed RNA counts*:*

$$Y_{gc}$$

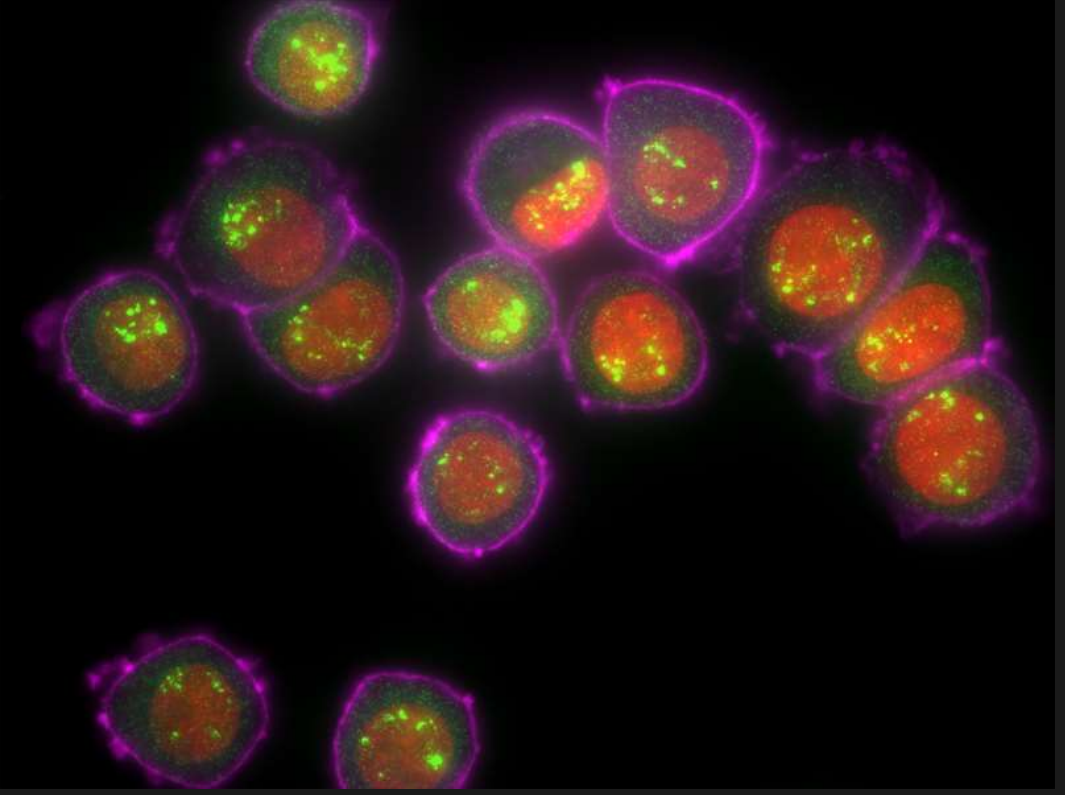
?

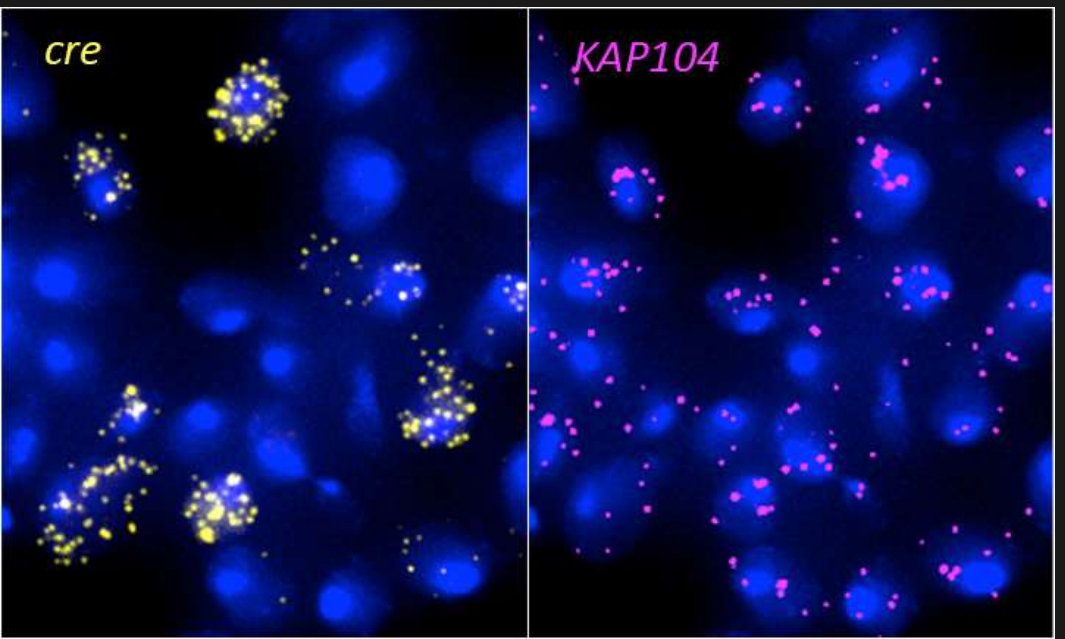
$$\widehat{H}_g$$

Distribution Deconvolution assuming a given noise model:

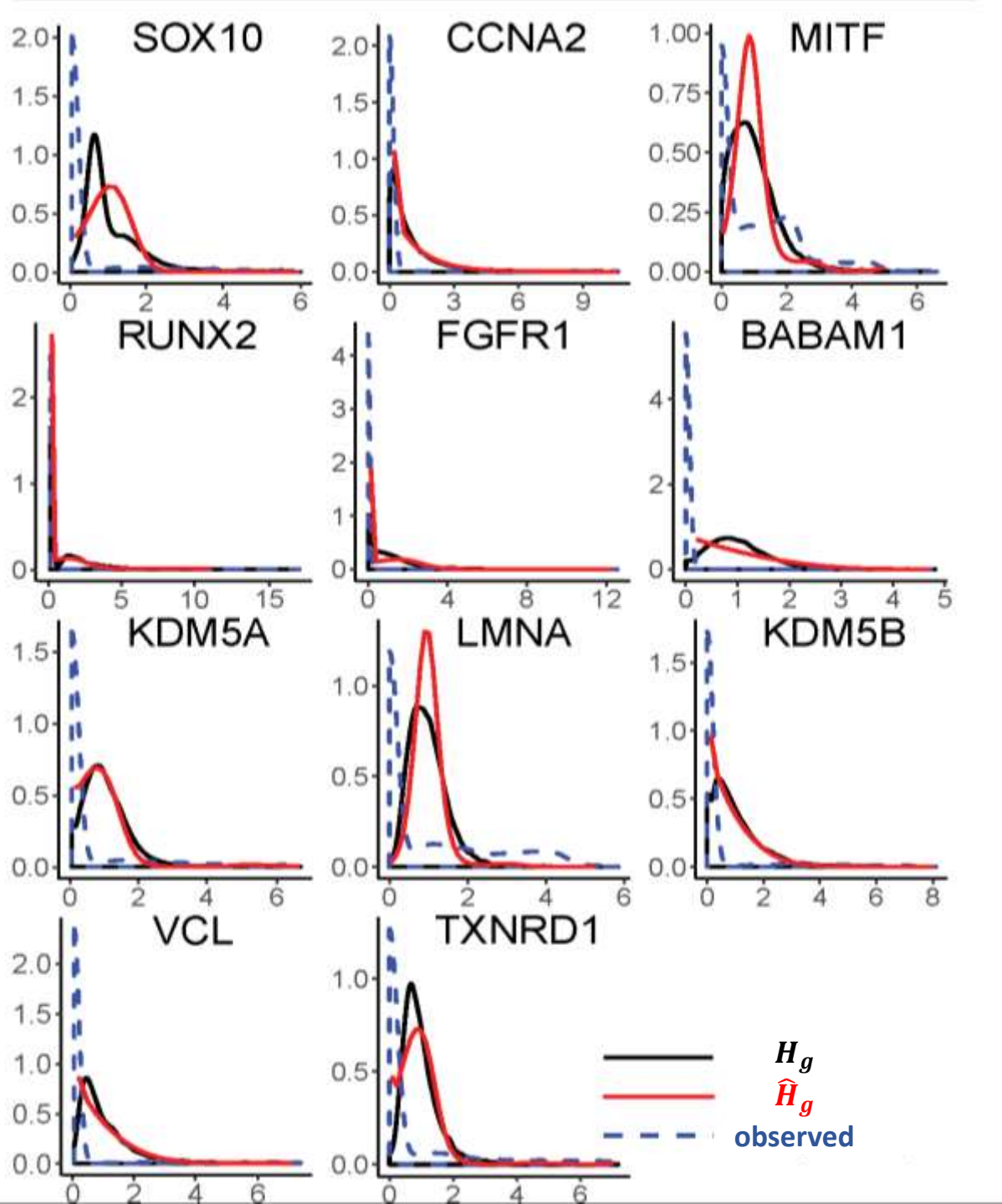
$$Y_{gc}|X_{gc} \overset{\text{ind}}{\sim} \text{Poisson}(l_c X_{gc})$$

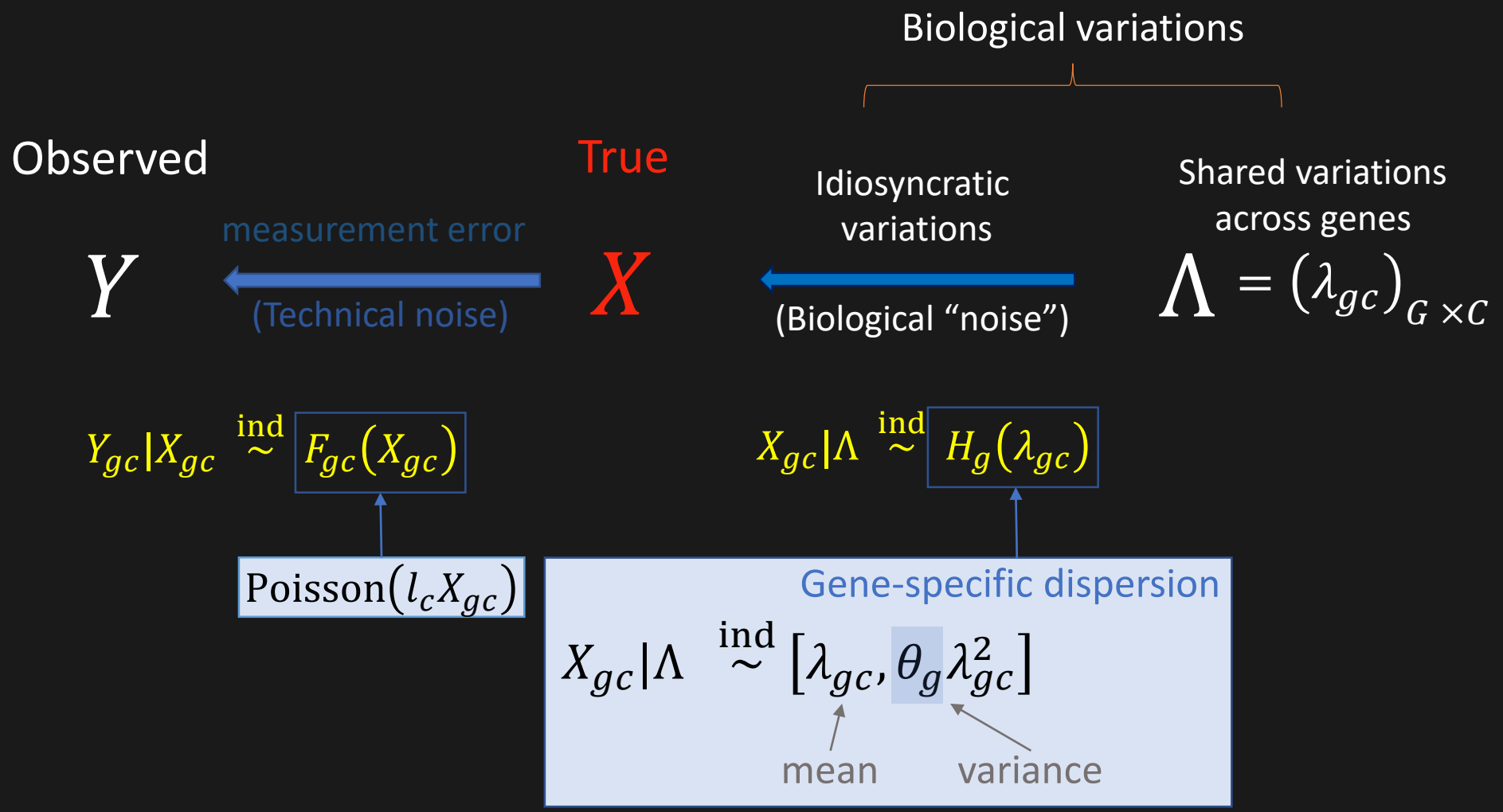
Gene expression distribution deconvolution in single-cell RNA sequencing, PNAS 2018.



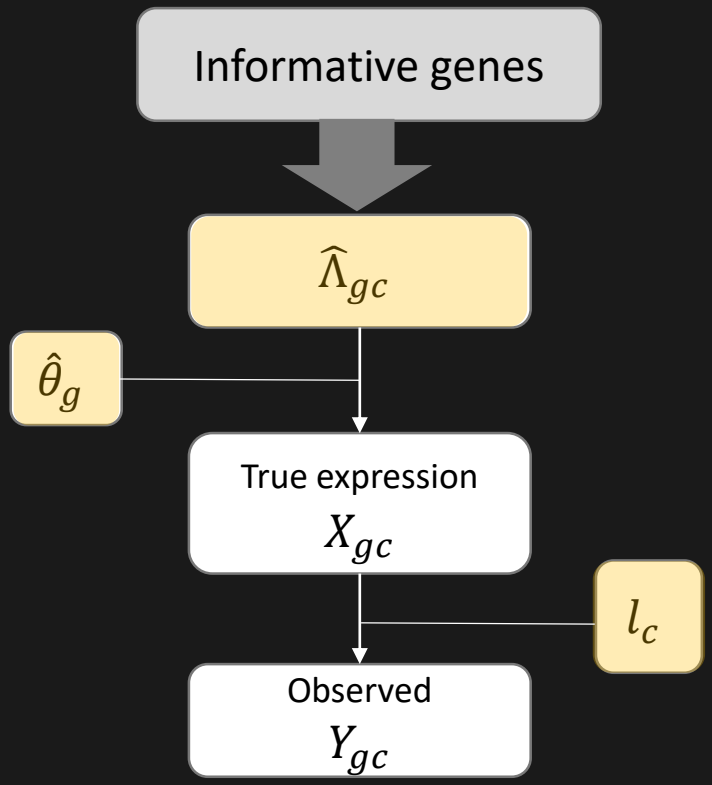


$\widehat{H}_g$ V.S. H_g

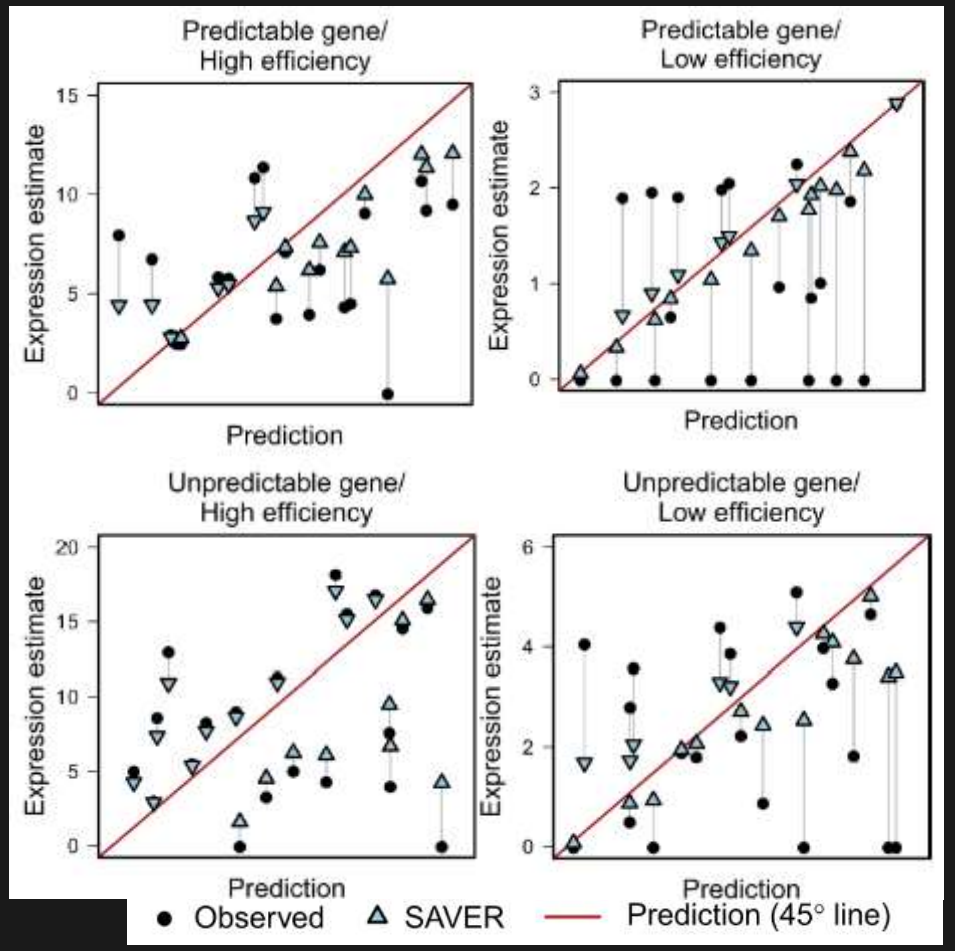



Biological variations
Observed
True
Idiosyncratic variations
Shared variations across genes
measurement error
(Technical noise)
(Biological "noise")
Y
X
$\Lambda = (\lambda_{gc})_{G \times C}$
$Y_{gc}|X_{gc} \overset{\text{ind}}{\sim} F_{gc}(X_{gc})$
$X_{gc}|\Lambda \overset{\text{ind}}{\sim} H_g(\lambda_{gc})$
$\text{Poisson}(l_c X_{gc})$
Gene-specific dispersion
$X_{gc}|\Lambda \overset{\text{ind}}{\sim} [\lambda_{gc}, \theta_g \lambda_{gc}^2]$
mean
variance

Achieving a balance between the predicted and the observed

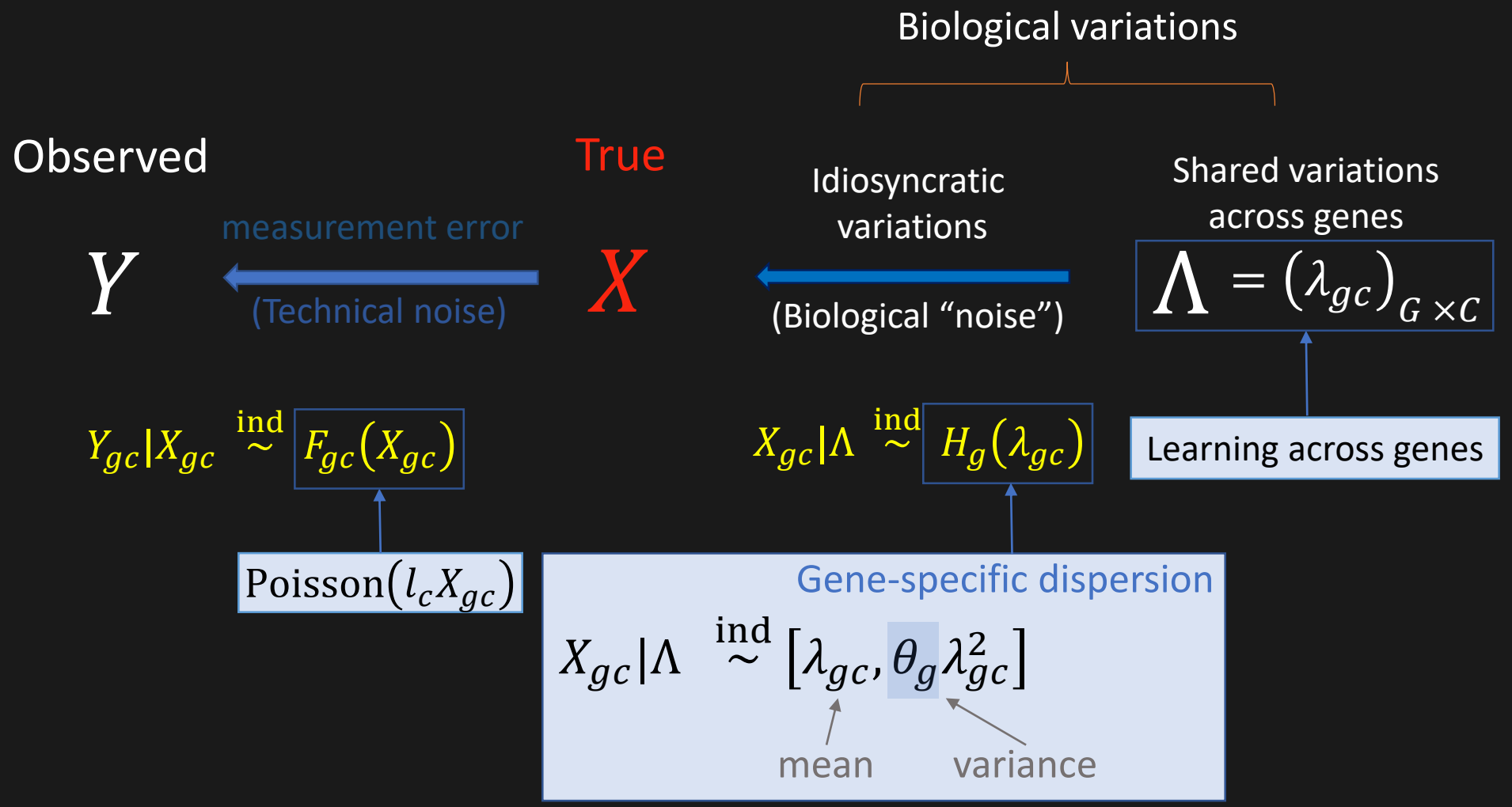


$$\hat{X}_{cg} = w_{cg}\frac{Y_{cg}}{s_c} + (1 - w_{cg})\hat{\Lambda}_{cg}$$



Gene expression recovery for single-cell RNA sequencing. Nature Methods 2018.

Decomposing the variation in three components



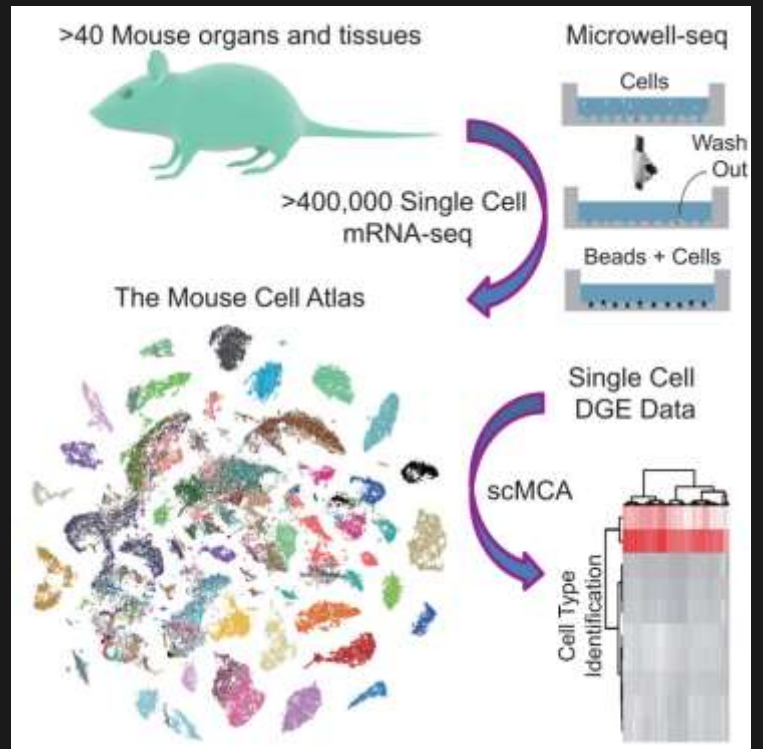
Can we use existing data in the public domain to denoise new scRNAseq datasets being generated?

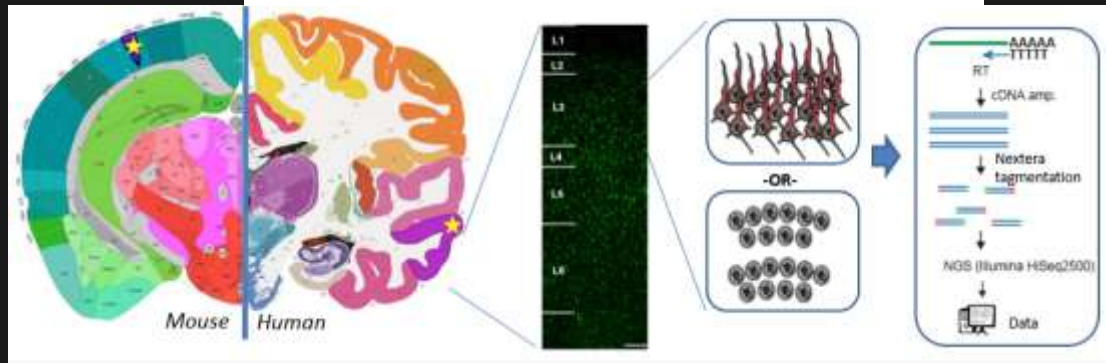
If the original study is of relatively low quality

or

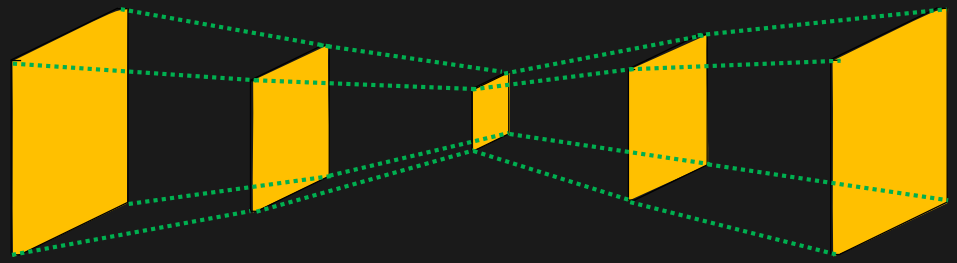
It hasn't profiled enough cells of a particular type that one might be interested in

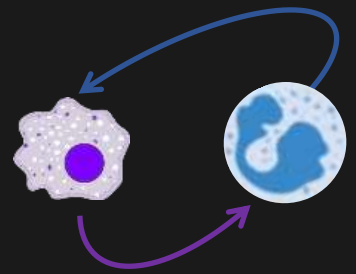
Data Model Result





At the backend, we pretrain an autoencoder using publicly available data





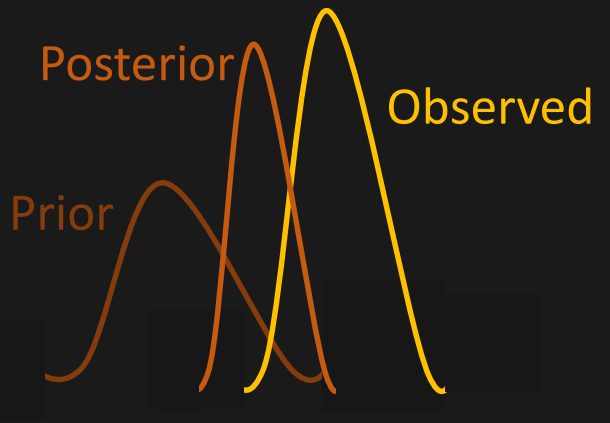
	Cell 1	Cell 2	…
Gene 1	0	1	3
Gene 2	1	0	1
…	0	2	2

Initialize the weights of the autoencoder by pretraining on cells extracted from public repositories. The weights are then updated to fit the target data.

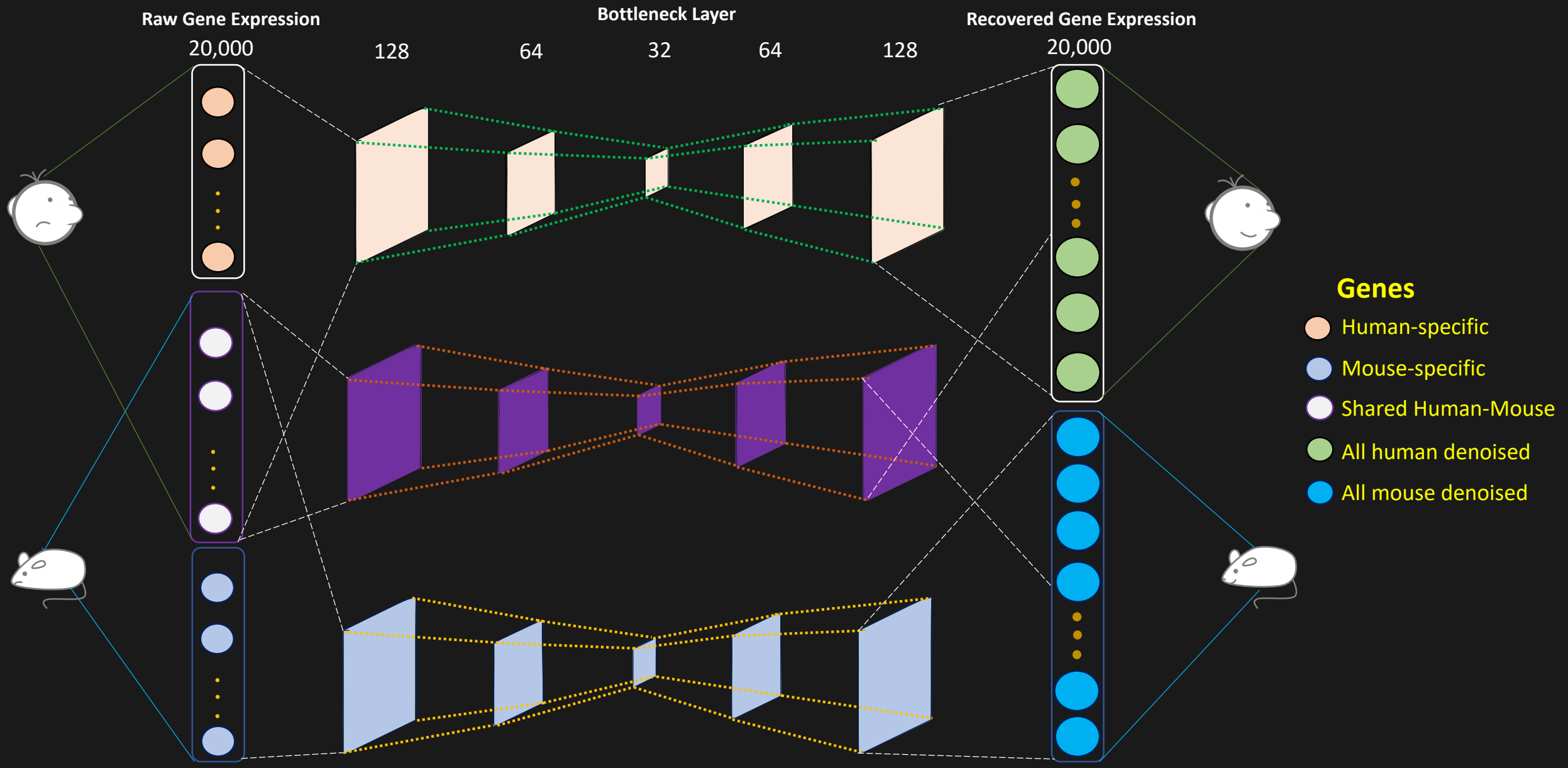
	Cell 1	Cell 2	…
Gene 1	0.23	1.5	3.3
Gene 2	1.3	0.4	1.7
…	0.1	2.6	2.05

Filter Genes

Bayesian shrinkage computes a weighted average of the predicted values and the observed data.



https://singlecell.wharton.upenn.edu/saver-x/



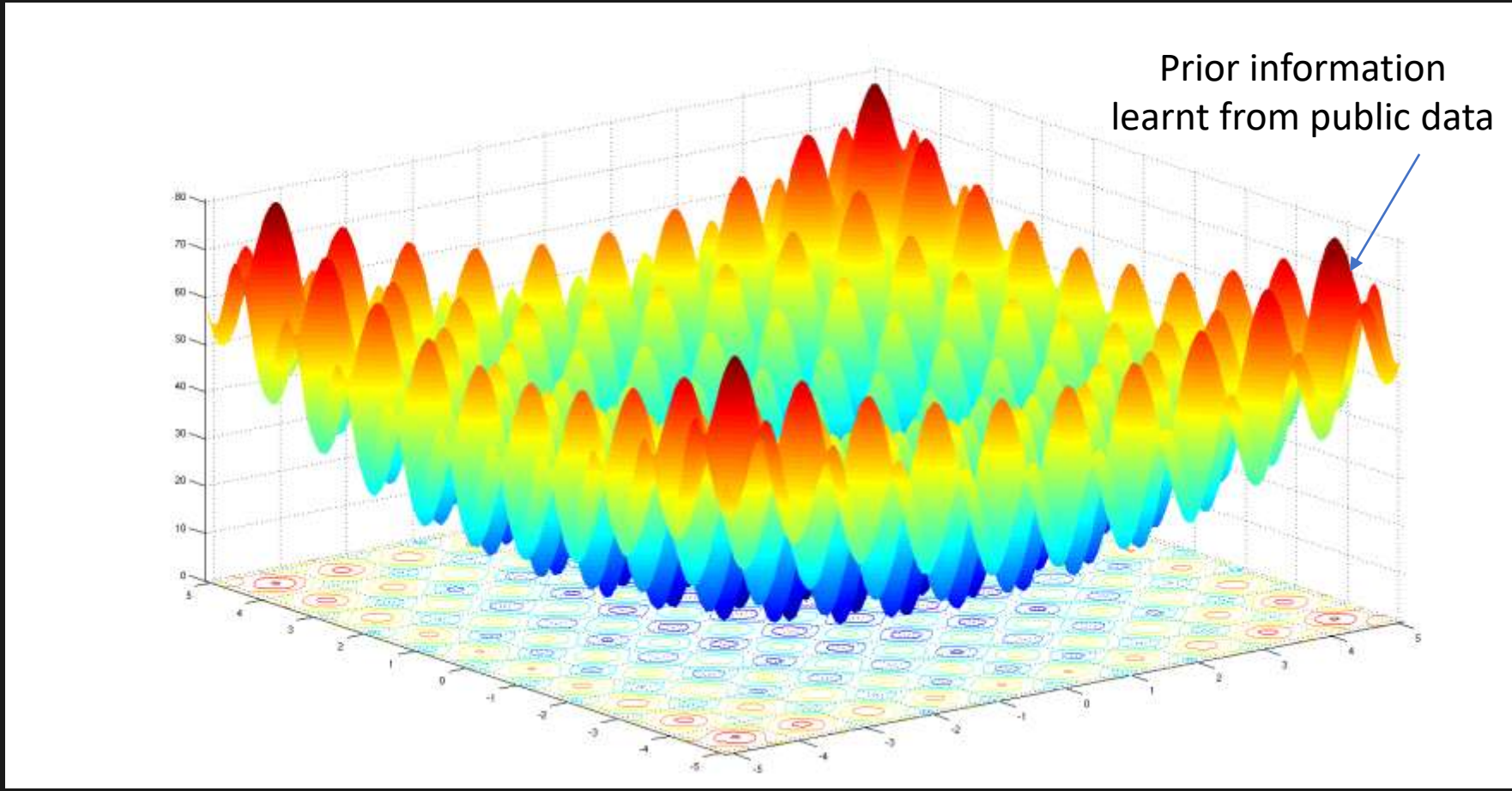
Loss function: maximize (quasi-) log-likelihood

$L(\Lambda, \vec{\alpha}; Y)$

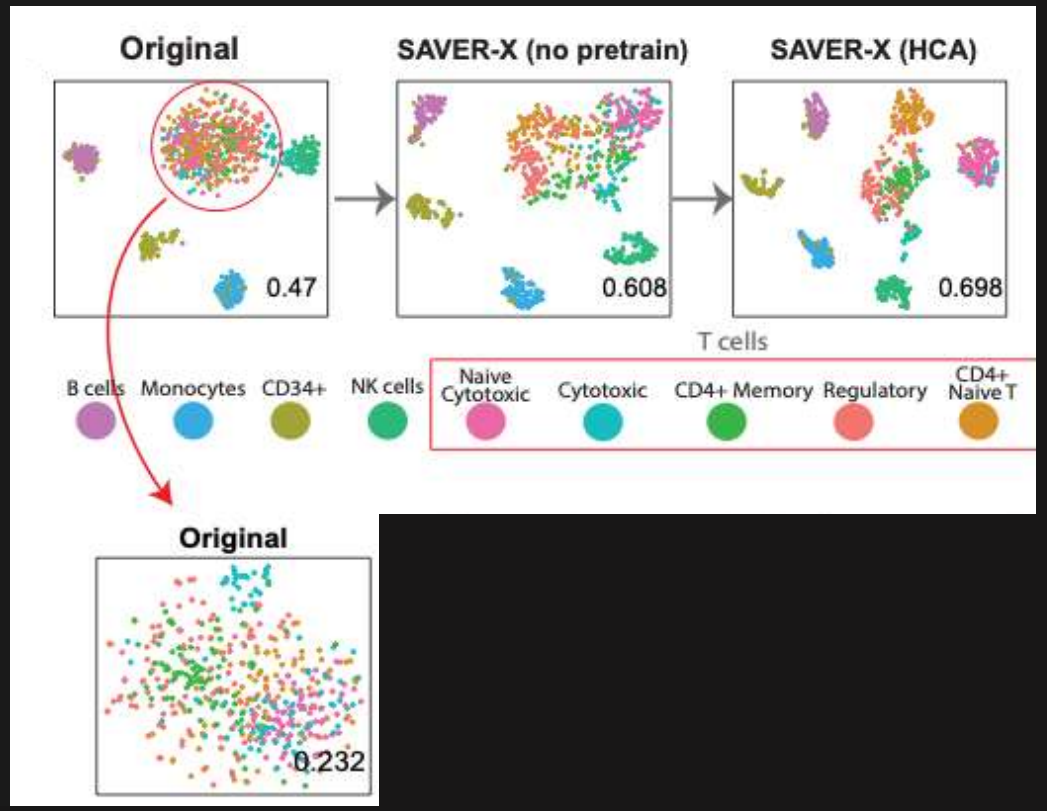
$\sim$ Negative binomial$[Y; \mu(\xi(Y)), \vec{\alpha}]$

What is transfer learning doing?

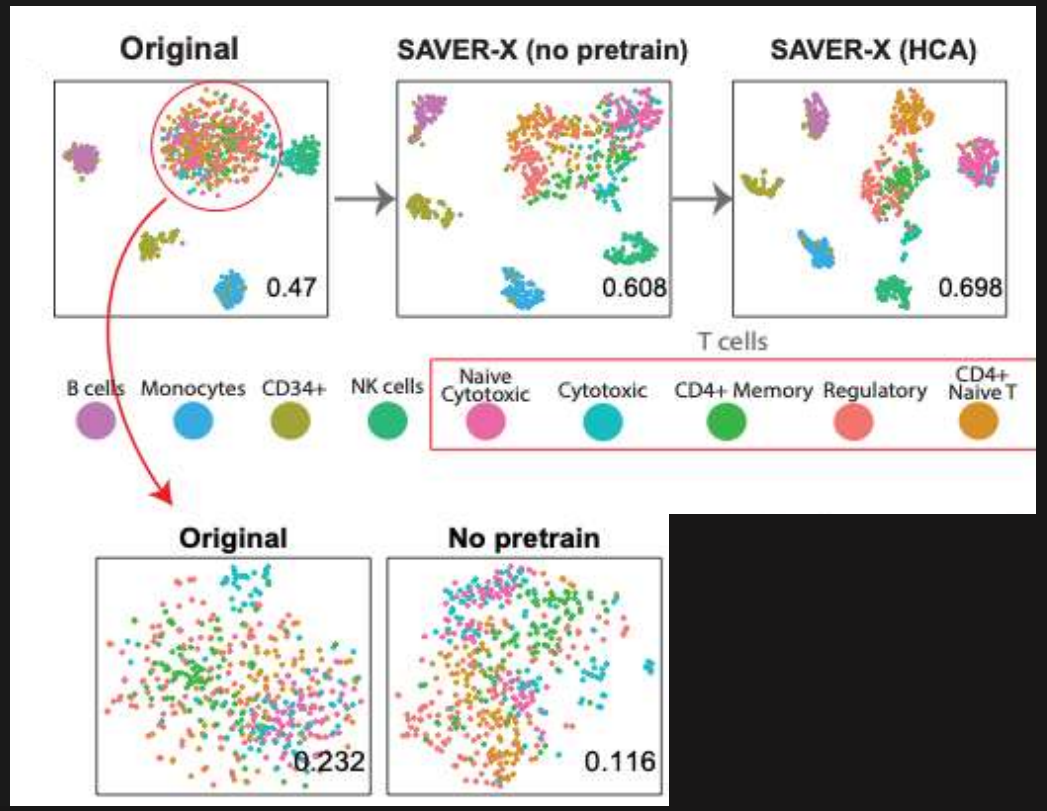
Estimating the parameters in our model (autoencoder) for better initialization



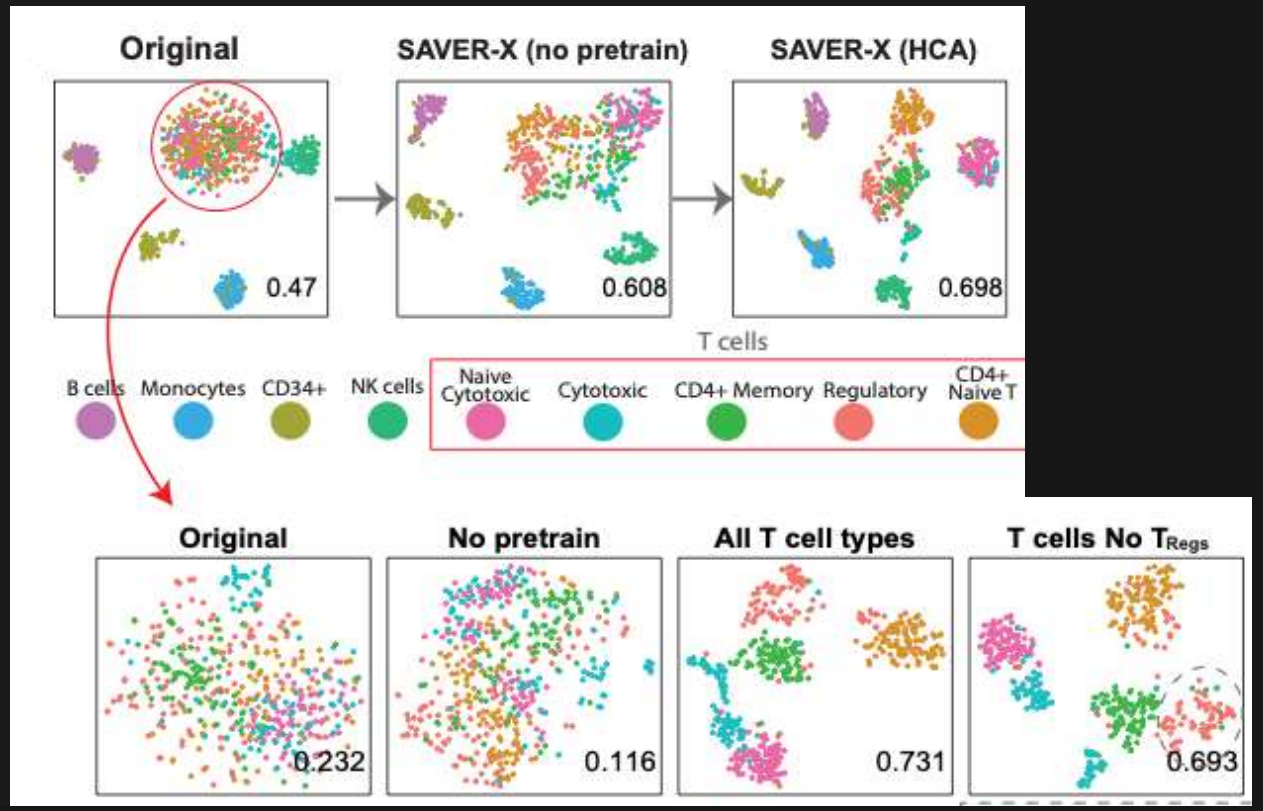
Let’s look at how SAVER-X does on real data



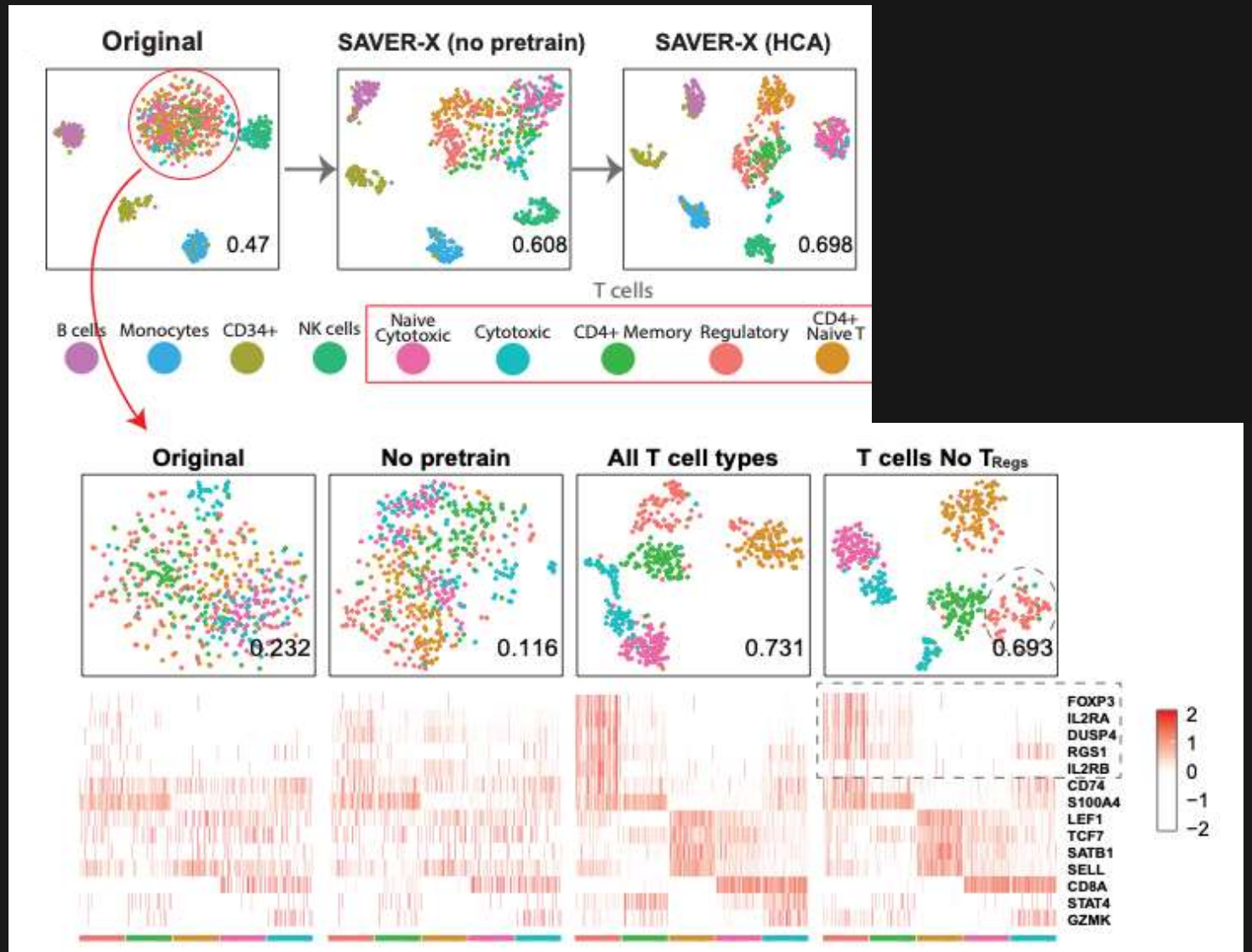
Let’s look at how SAVER-X does on real data



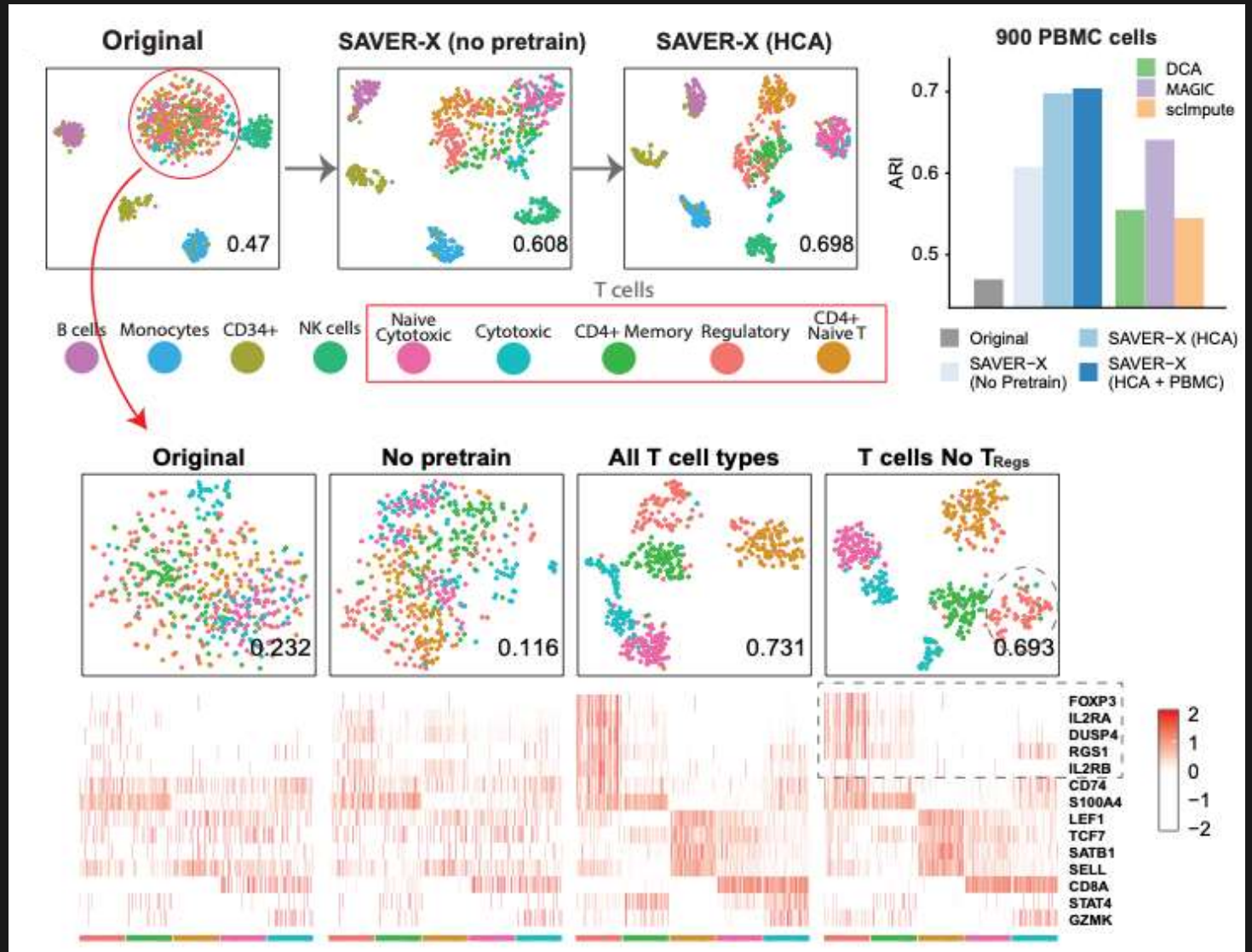
Let’s look at how SAVER-X does on real data



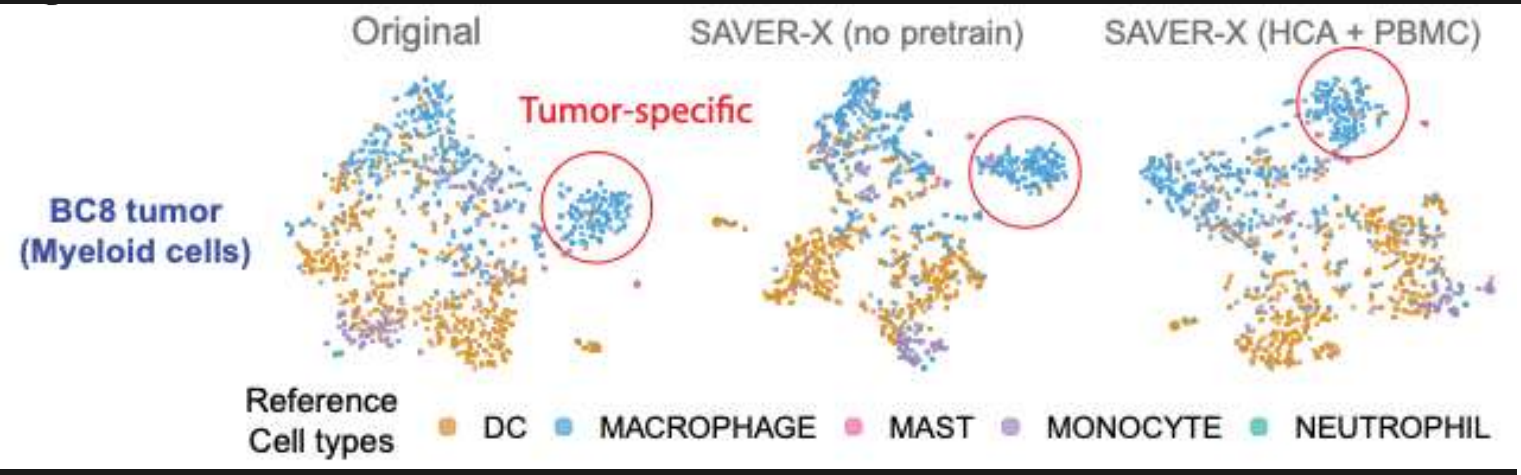
Let’s look at how SAVER-X does on real data

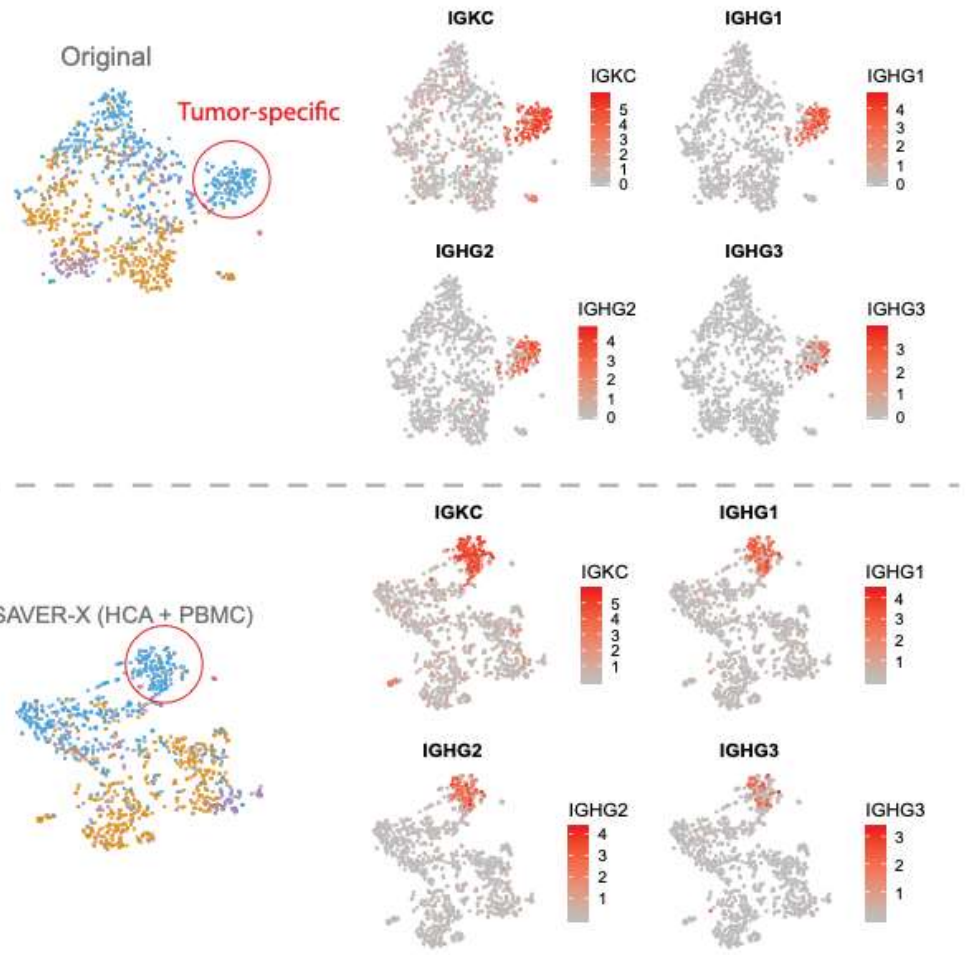


Let’s look at how SAVER-X does on real data



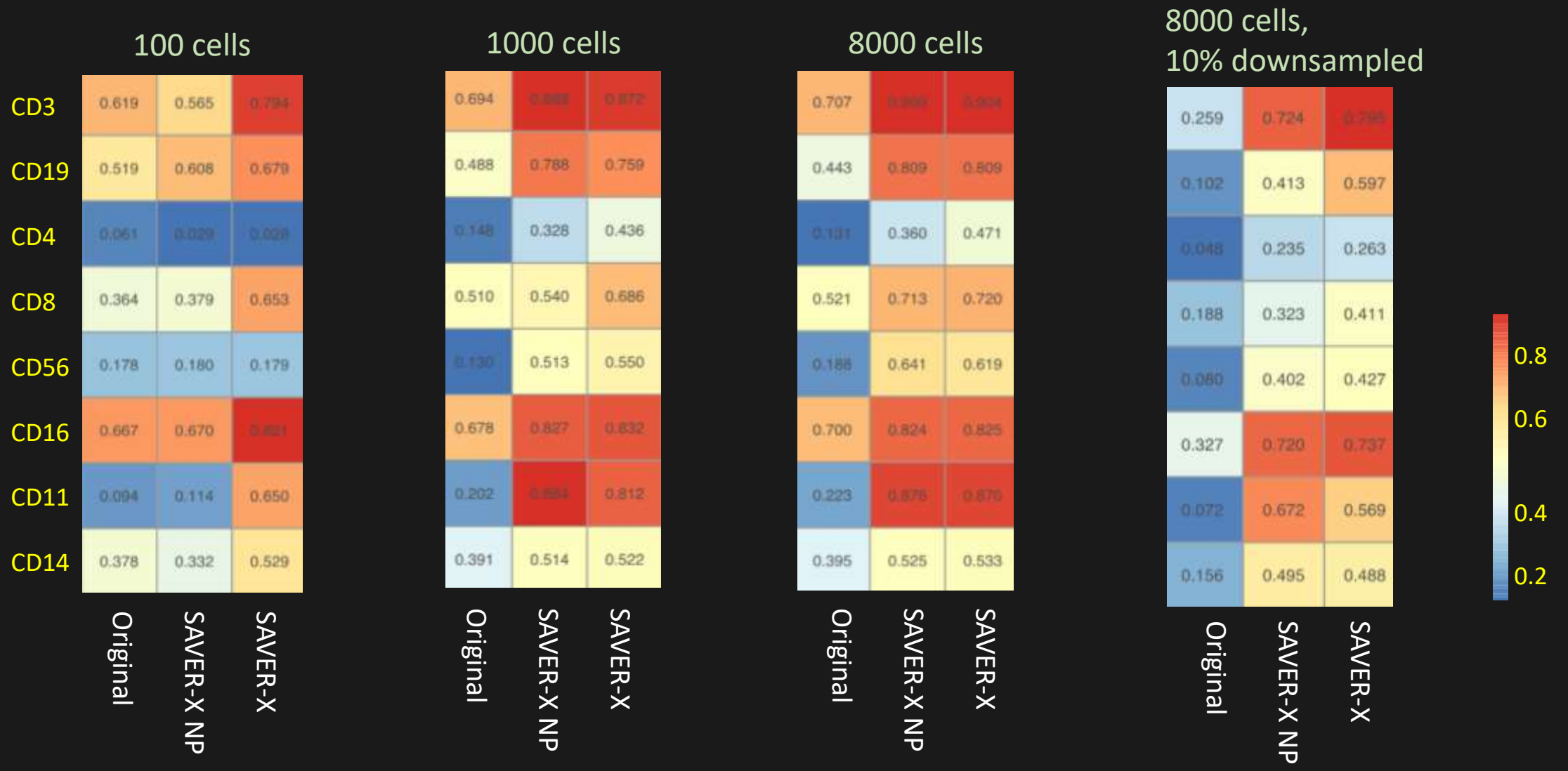
Denoising datasets in disease settings by borrowing information from related datasets in the healthy domain



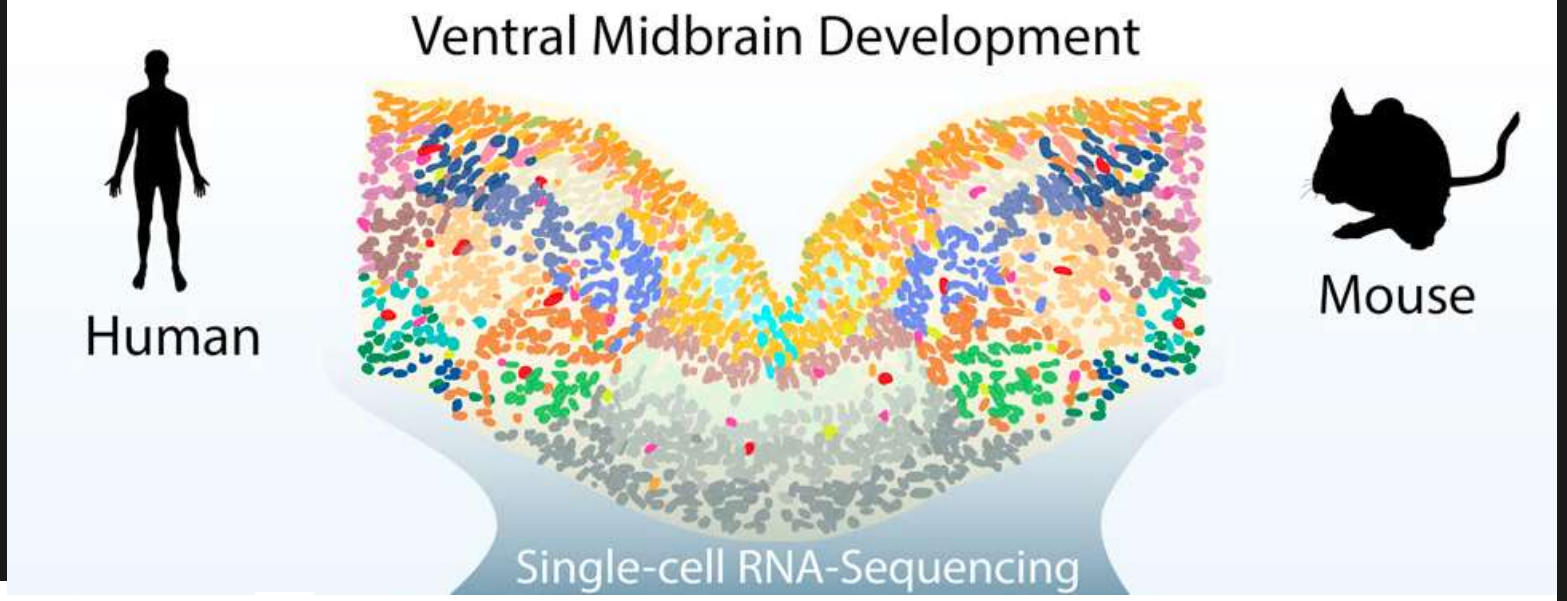


While preserving disease-specific cell types and their signatures

SAVER-X improves correlations between cell surface proteins and their corresponding genes



Can mouse data help denoise human data?



La Manno (2016)
Developing
Ventral Midbrain

Mouse Cell Atlas
(2018)
Developing Brain Cells

Human

Mouse

random split

1907 cells

7187 cells

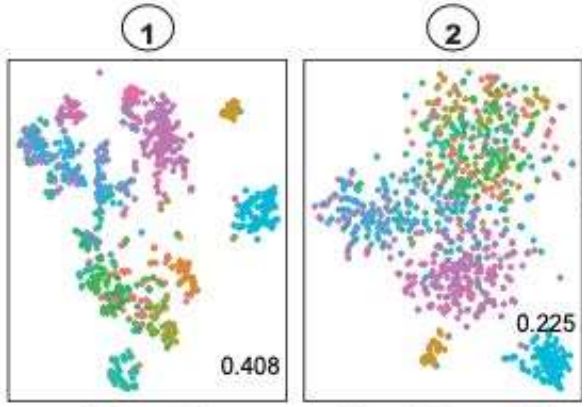
1 3 4 5

1000 cells 977 cells (Avg. UMI/cell: 8123) (Avg. UMI/cell: 675)

Down-sampling
with average efficiency **10%**

2 **(low quality)**

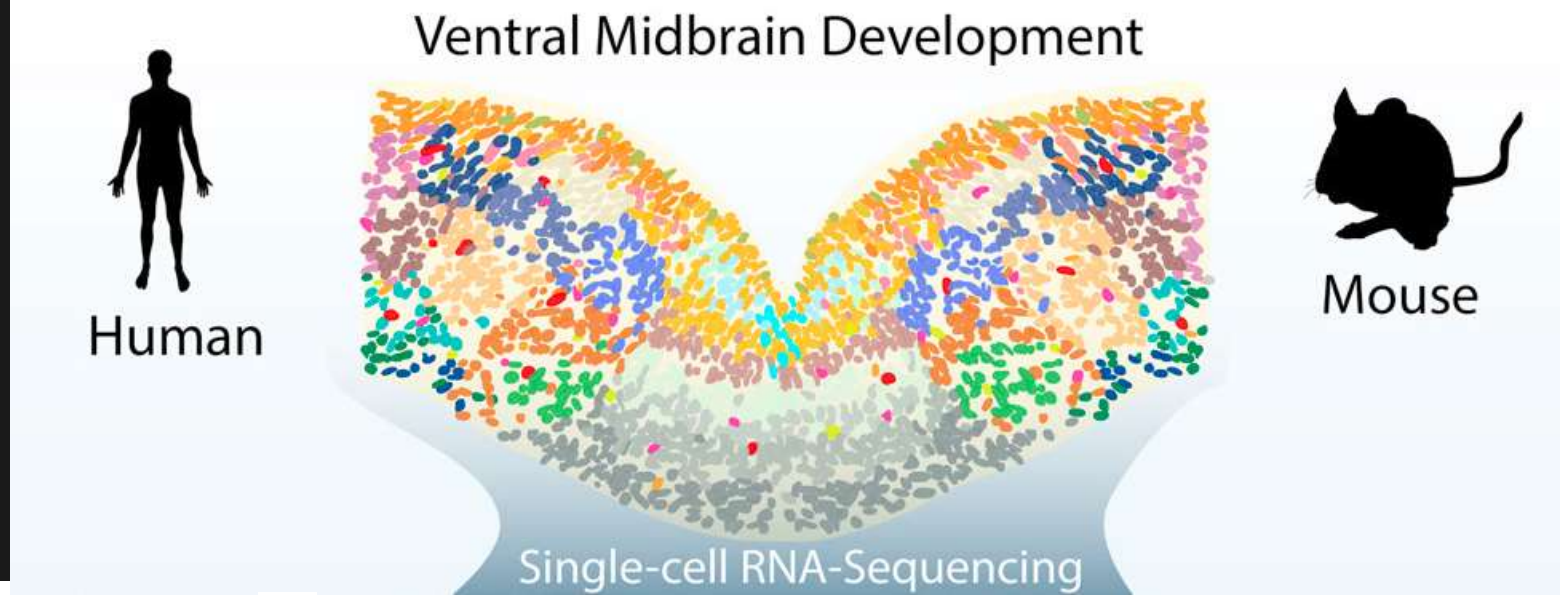
(Avg. UMI/cell: 452)



Major cell types

- hEndo
- hGaba
- hNbM
- hNbML1
- hNProg
- hOMTN
- hPeric
- hProgBP
- hProgFPL
- hProgFPM
- hProgM
- hRgl2a
- hRgl2b

Can mouse data help denoise human data?



La Manno (2016)
Developing
Ventral Midbrain

Mouse Cell Atlas
(2018)
Developing Brain Cells

Human

Mouse

random split

1907 cells

7187 cells

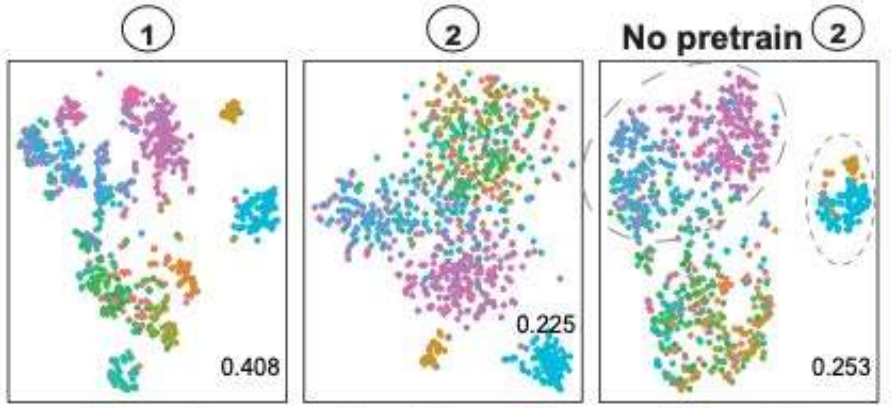
1 3 4 5

1000 cells 977 cells (Avg. UMI/cell: 8123) (Avg. UMI/cell: 675)

Down-sampling
with average efficiency **10%**

2 **(low quality)**

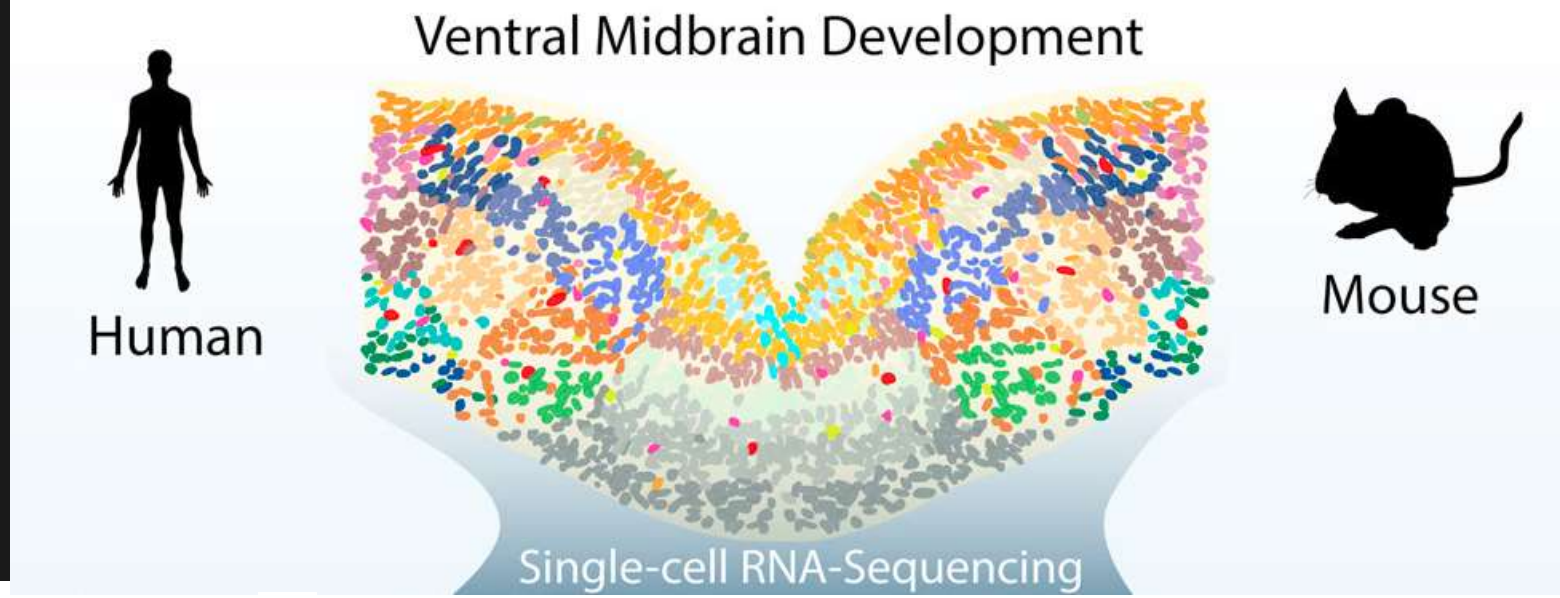
(Avg. UMI/cell: 452)



Major cell types

- hEndo
- hGaba
- hNbM
- hNbML1
- hNProg
- hOMTN
- hPeric
- hProgBP
- hProgFPL
- hProgFPM
- hProgM
- hRgl2a
- hRgl2b

Can mouse data help denoise human data?



La Manno (2016)
Developing
Ventral Midbrain

Mouse Cell Atlas
(2018)
Developing Brain Cells

Human

Mouse

random split

1907 cells

7187 cells

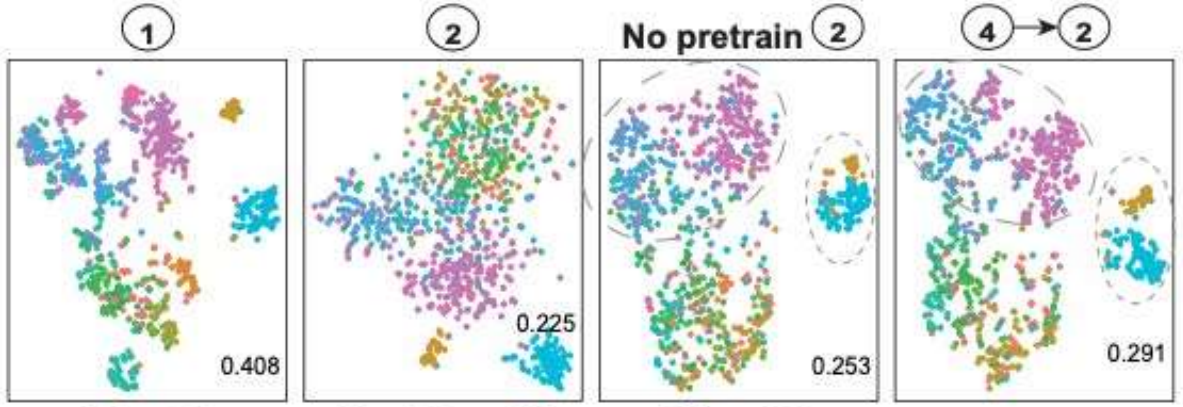
1 3 4 5

1000 cells 977 cells (Avg. UMI/cell: 8123) (Avg. UMI/cell: 675)

Down-sampling
with average efficiency **10%**

2 **(low quality)**

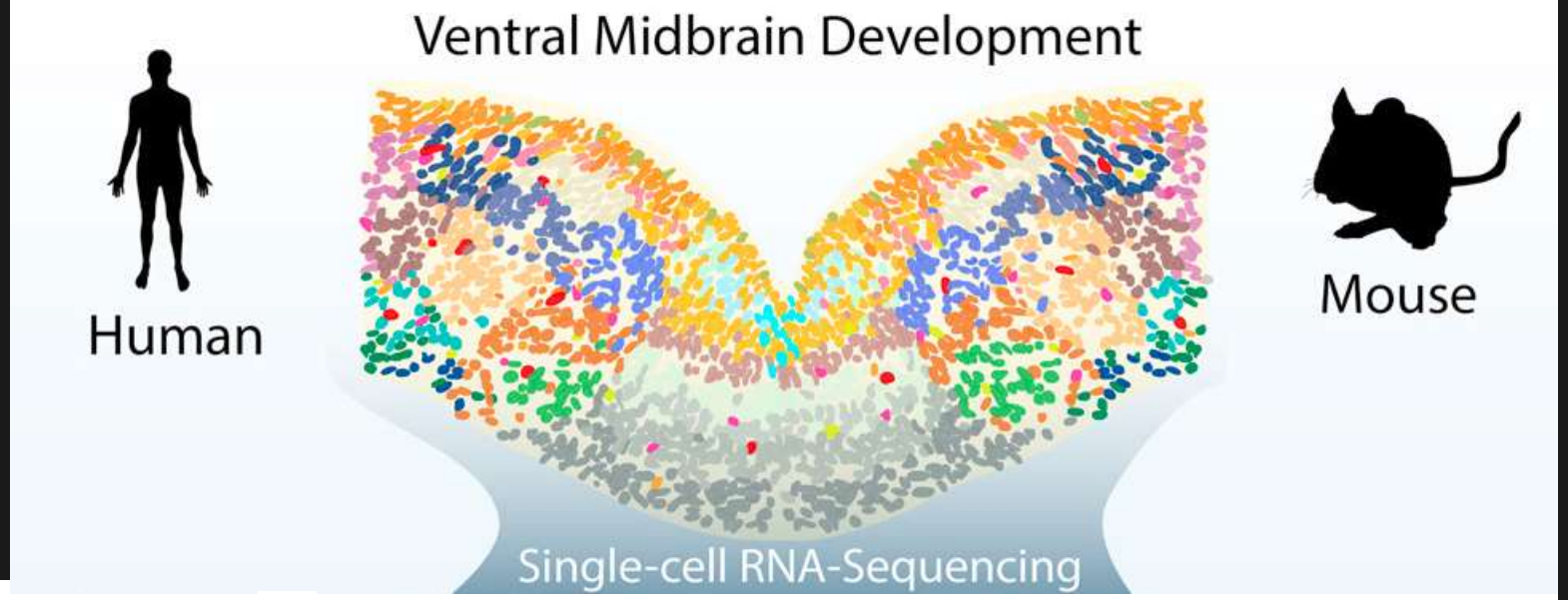
(Avg. UMI/cell: 452)



Major cell types

- hEndo
- hGaba
- hNbM
- hNbML1
- hNProg
- hOMTN
- hPeric
- hProgBP
- hProgFPL
- hProgFPM
- hProgM
- hRgl2a
- hRgl2b

Can mouse data help denoise human data?



La Manno (2016)
Developing
Ventral Midbrain

Mouse Cell Atlas
(2018)
Developing Brain Cells

Human

Mouse

random split

1907 cells

7187 cells

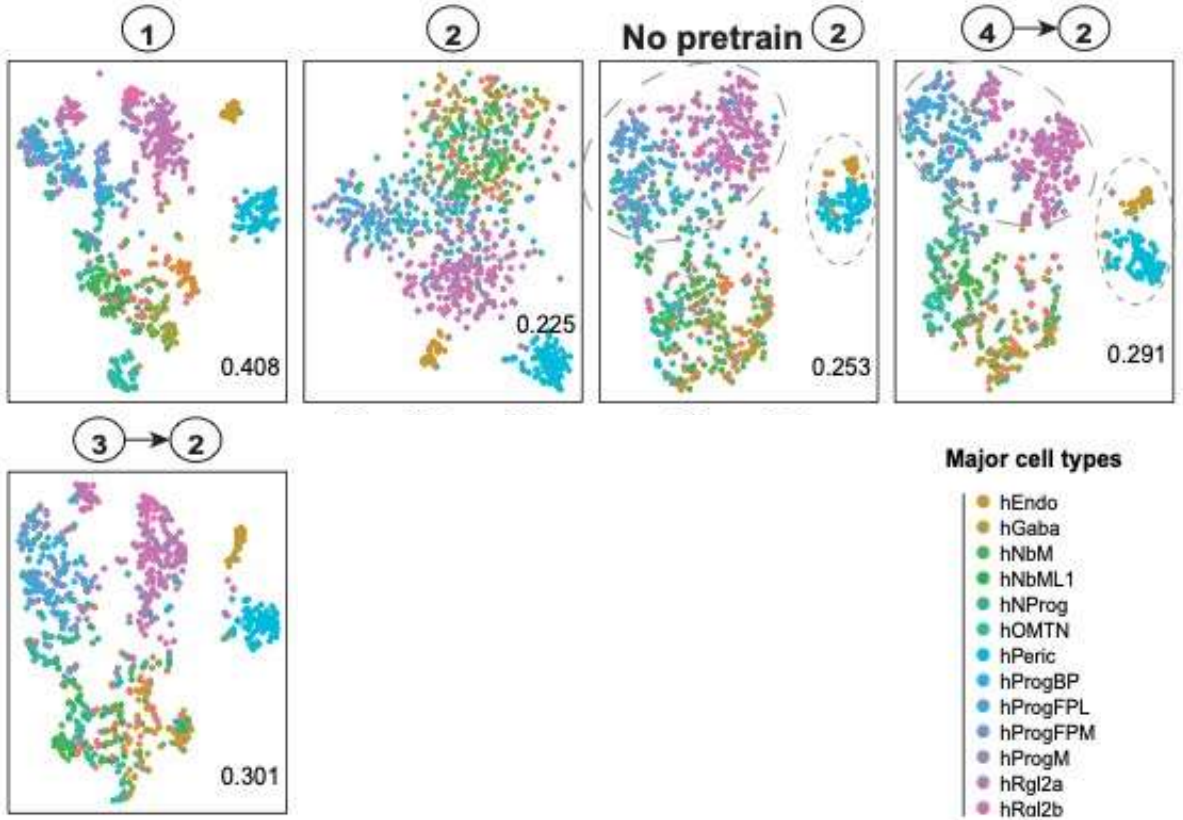
1 3 4 5

1000 cells 977 cells (Avg. UMI/cell: 8123) (Avg. UMI/cell: 675)

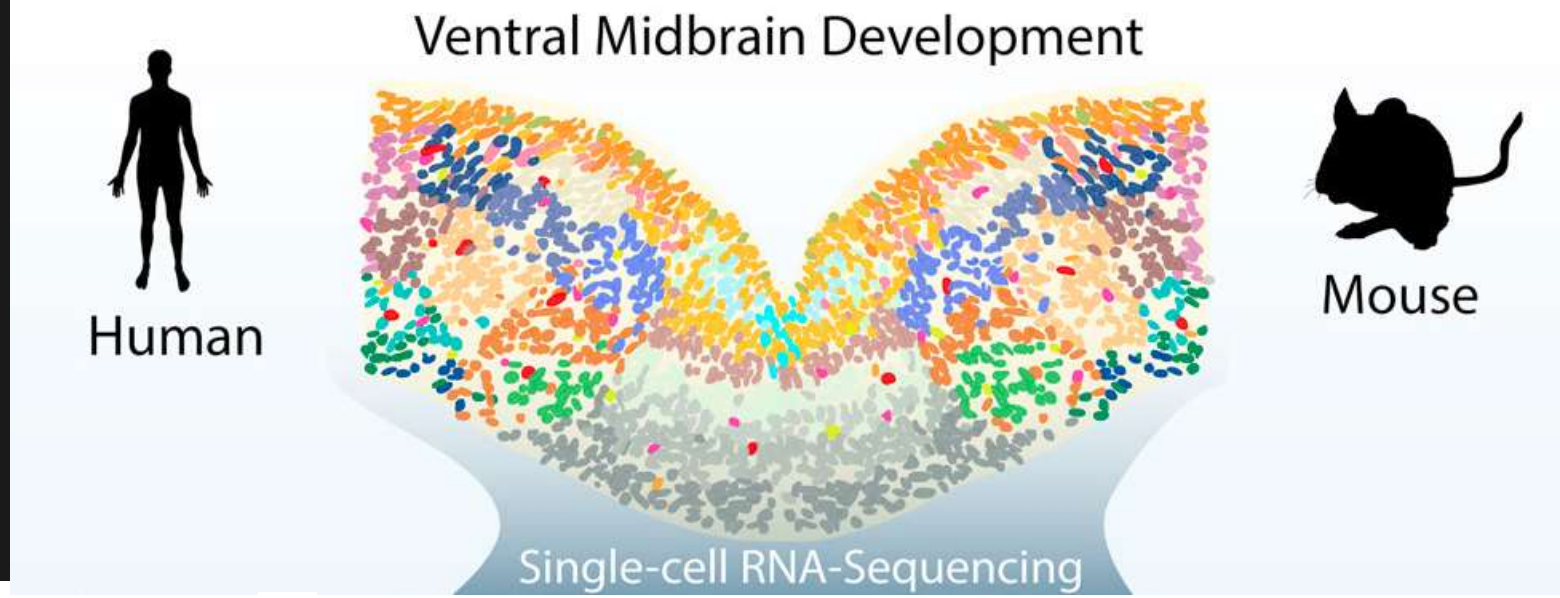
Down-sampling
with average efficiency **10%**

2 **(low quality)**

(Avg. UMI/cell: 452)



Can mouse data help denoise human data?



La Manno (2016)
Developing
Ventral Midbrain

Mouse Cell Atlas
(2018)
Developing Brain Cells

Human

Mouse

Mouse

random split

1907 cells

7187 cells

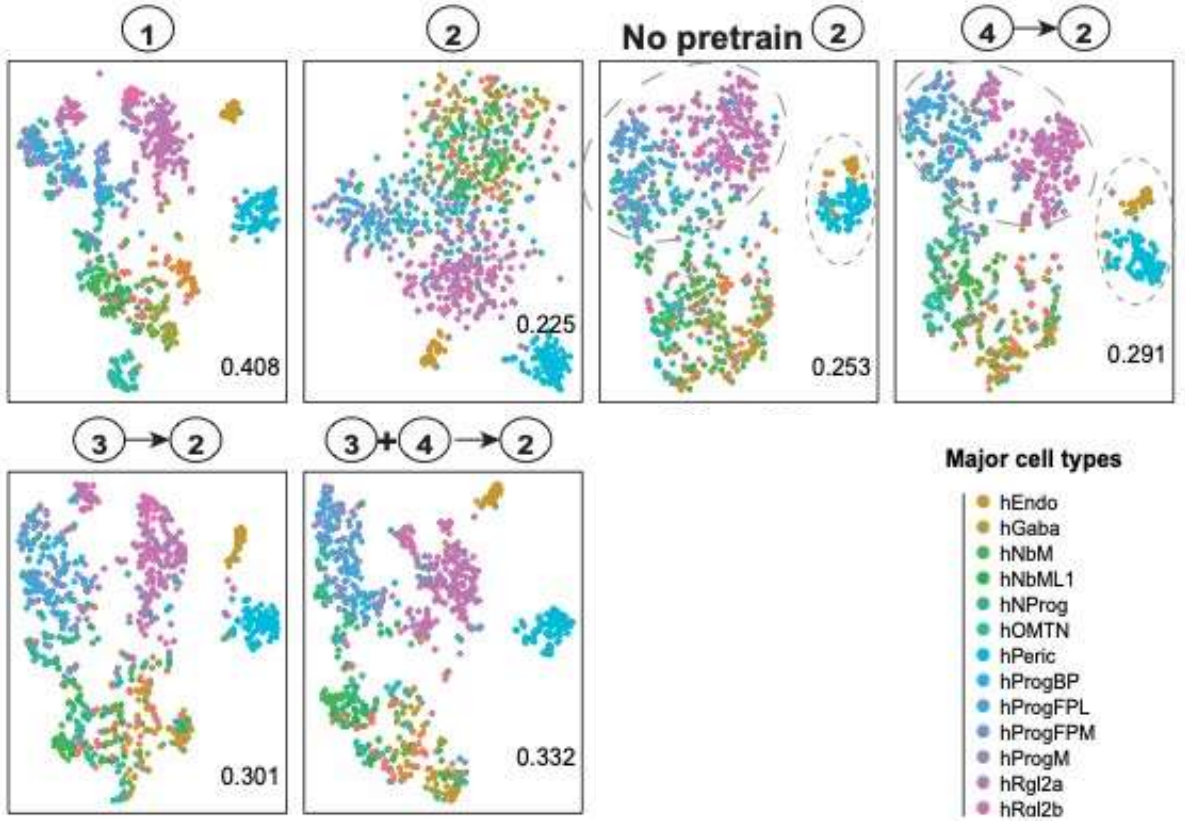
1 3 4 5

1000 cells 977 cells (Avg. UMI/cell: 8123) (Avg. UMI/cell: 675)

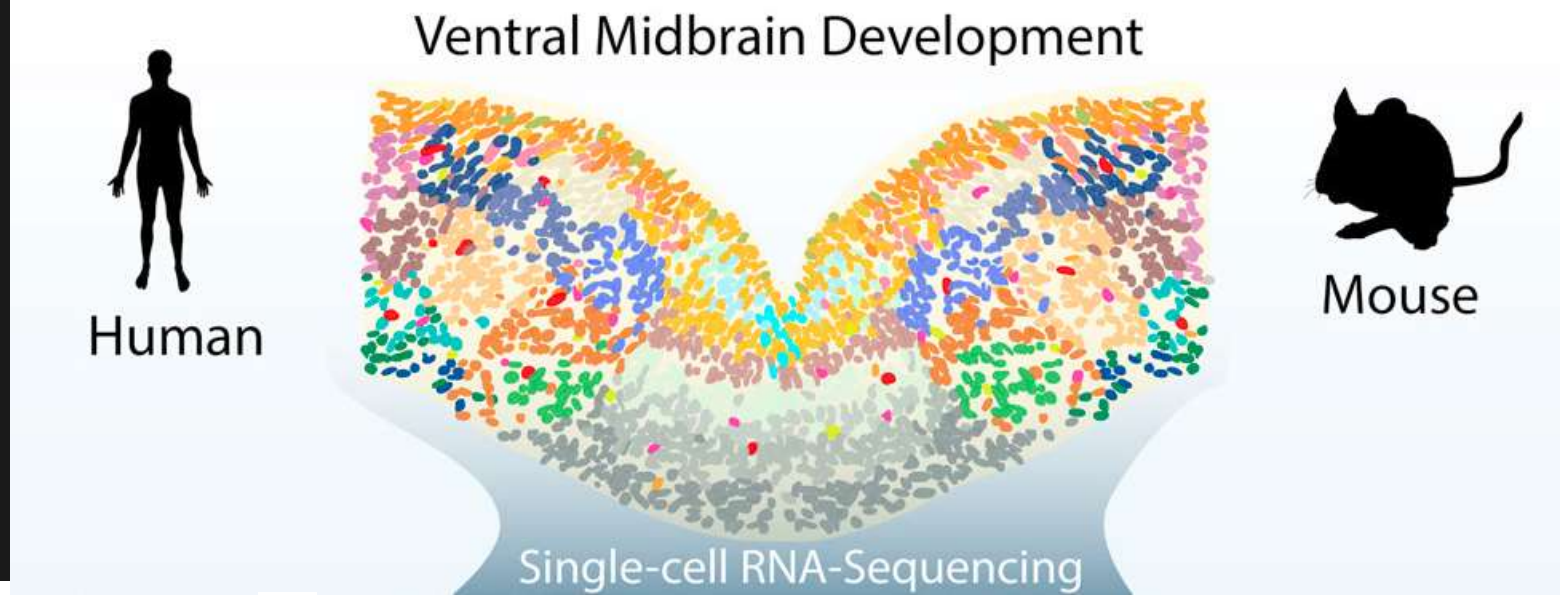
Down-sampling
with average efficiency **10%**

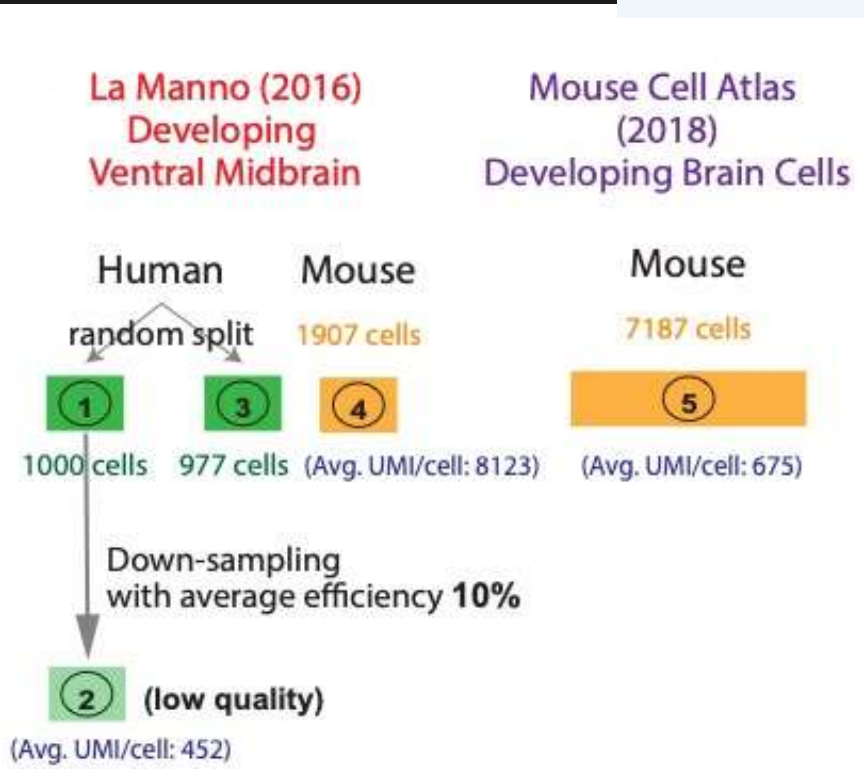
2 **(low quality)**

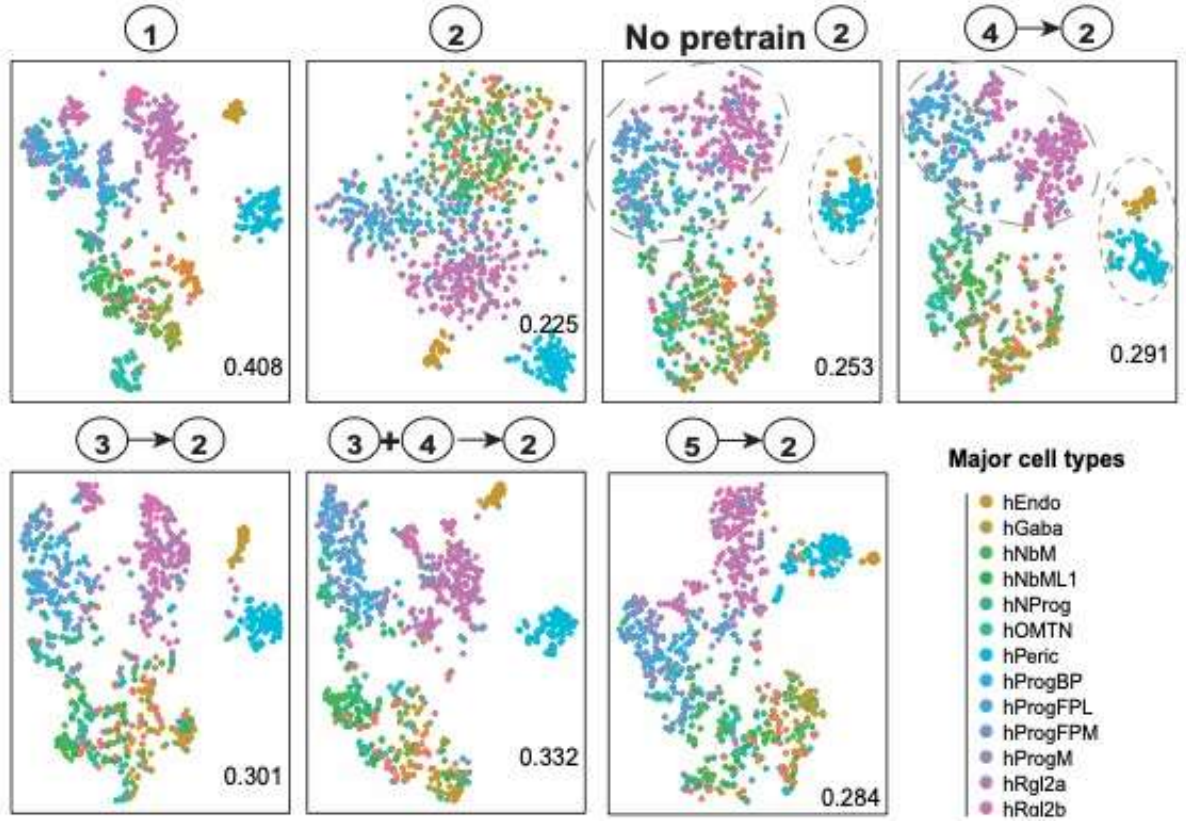
(Avg. UMI/cell: 452)



Yes, mouse data help denoise human data!

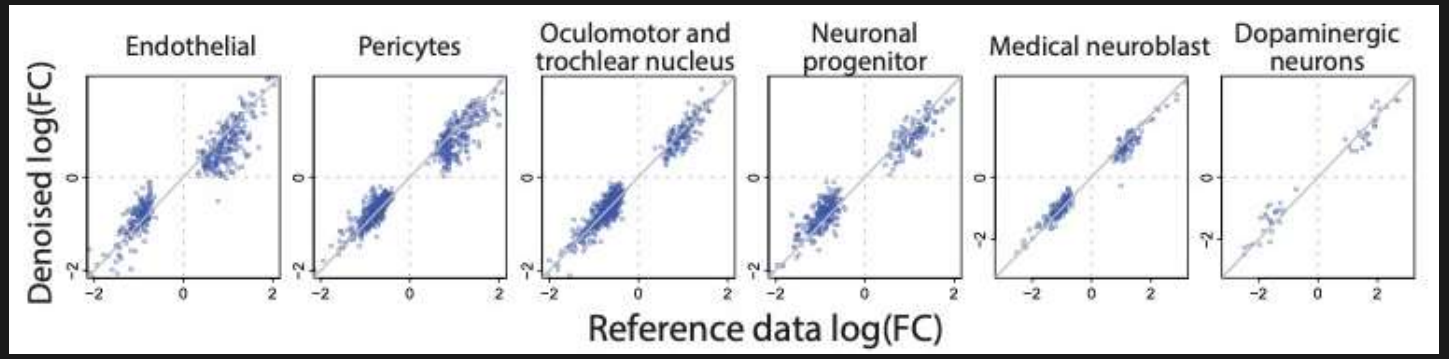






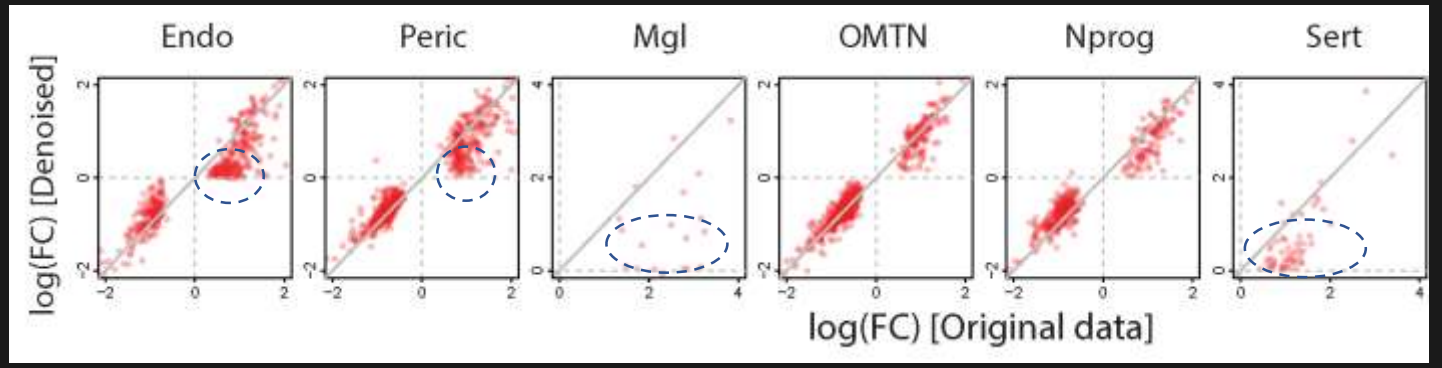
SAVER-X does not bias towards external data

SAVER-X preserves gene expression patterns that are unique to human



- Training after initialization
- Cross-validation to only transfer for "predictive" genes
- Empirical Bayes shrinkage for estimating X

Without cross-validation / EB shrinkage



Overview of my Talk

1. Single cell RNA sequencing (scRNA-seq)
    - Why is raw scRNA-seq data noisy? How can we address this problem?

2. The ideas underlying our proposed solution
    - Exploring the power (and the limits) of transfer learning

3. Statistical inference on the denoised values: why you should care

Propagating uncertainty in downstream analyses

Gene-level analyses:

Determining gene-gene correlations

Inference of regulatory networks

Cell-level analyses:

Computing cell-to-cell distance for clustering or visualization

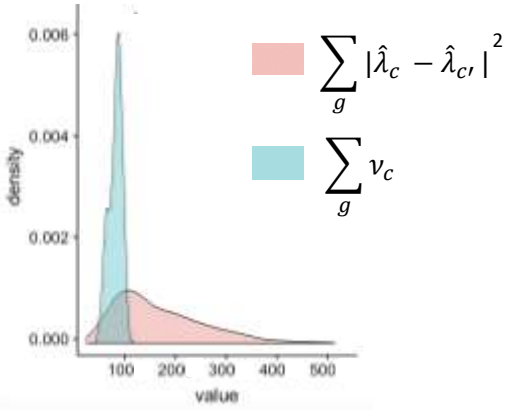
Uncertainty-adjusted Euclidean distance between cells

$$E[\,\|X_c - X_{c'}\|^2\,|Y] = \left\|\hat{X}_c - \hat{X}_{c'}\right\|^2 + \sum_g \hat{v}_{gc} + \sum_g \hat{v}_{gc'}$$

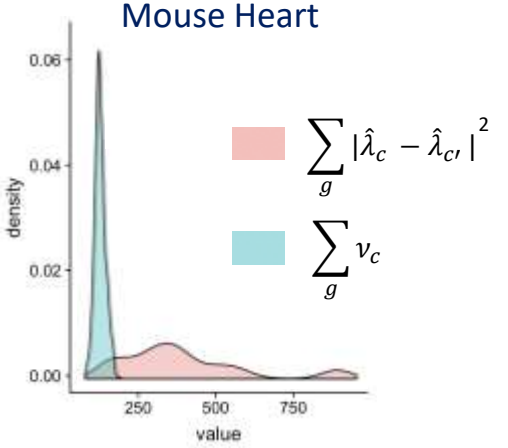
versus

Sampling of $\hat{X}_{gc}$ from its posterior distribution

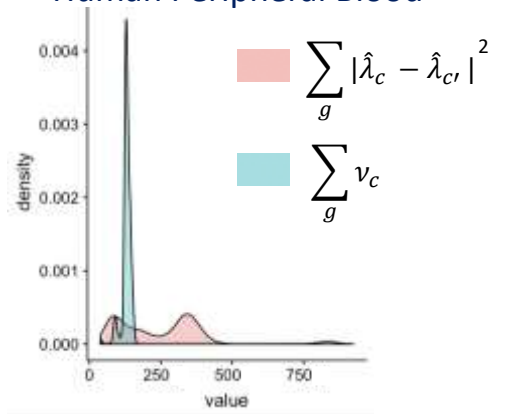
Mouse Kidney



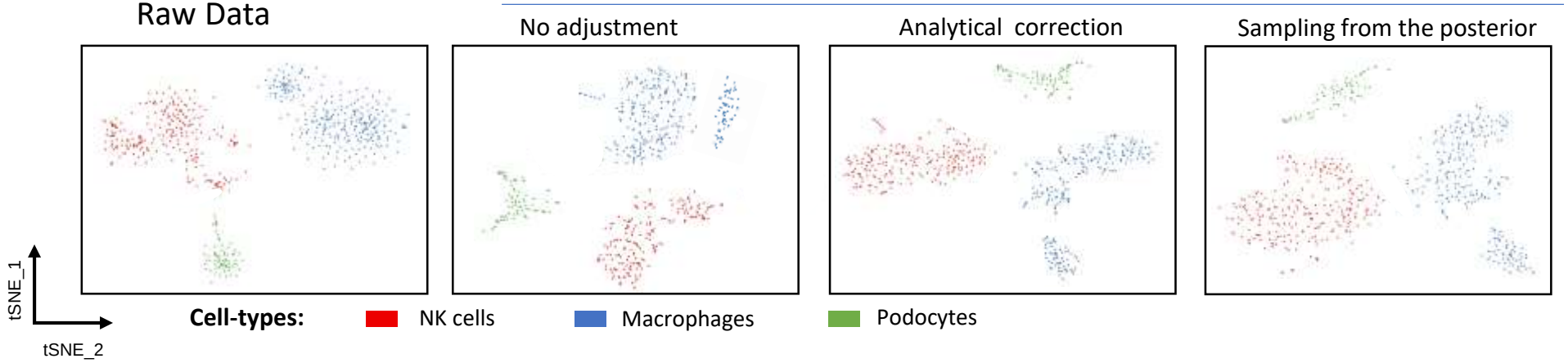
Mouse Heart

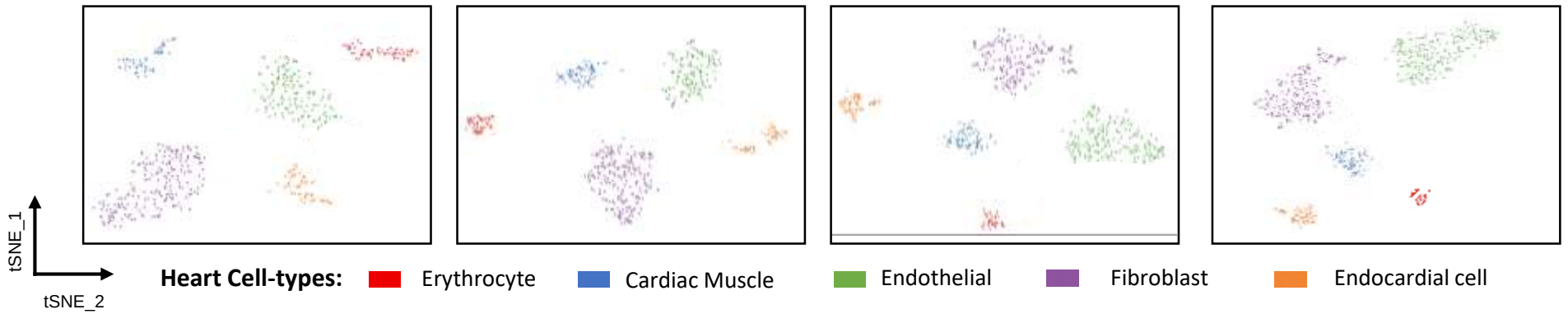


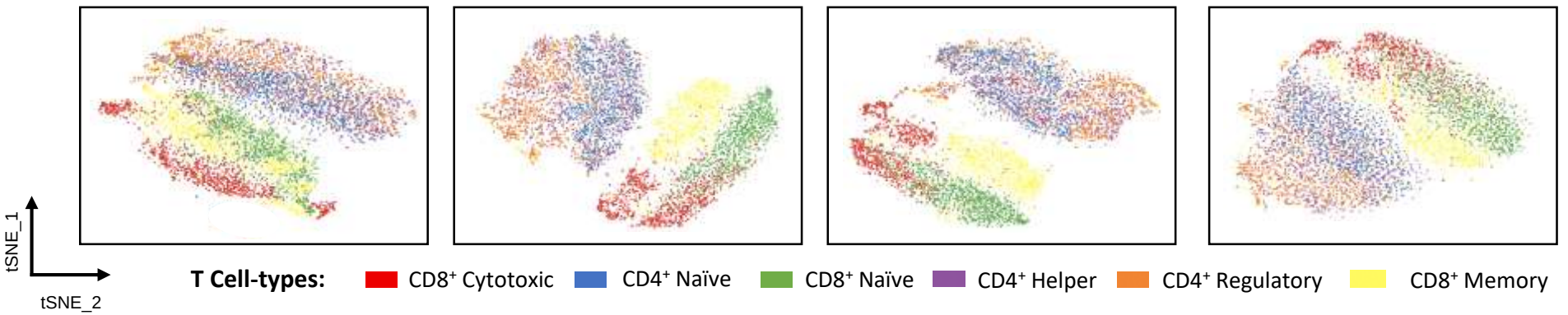
Human Peripheral Blood



Working with SAVER-denoised Values







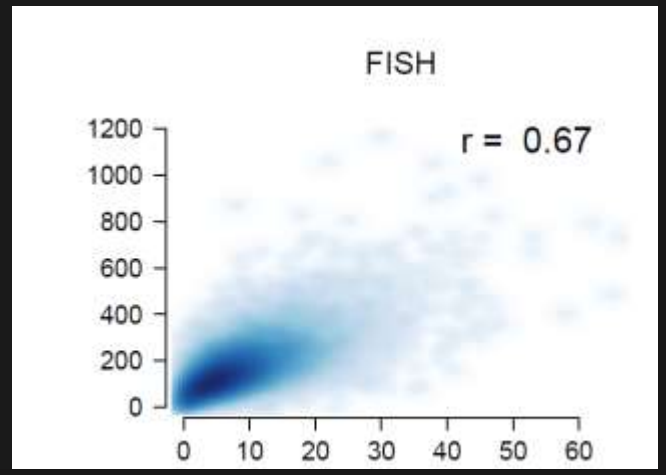
Recovering gene–gene relationships

$$\lambda_{cg} | \, Y_{cg}, \mu_{cg}, \hat{\sigma}_{cg} \sim \Gamma(\hat{\lambda}_{cg}, v_{cg})$$

$$Cor(\lambda_{cg}, \lambda_{cg'})$$

$$= Cor(\hat{\lambda}_{cg}, \hat{\lambda}_{cg'}) \times f_g \times f_{g'}$$

f_g has simple analytical formula.



BABAM1

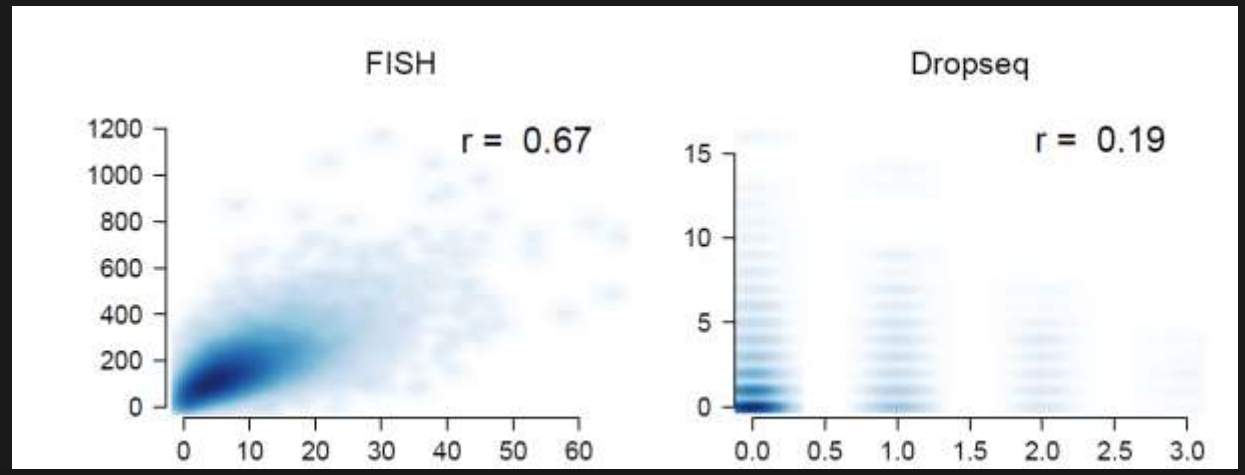
Recovering gene–gene relationships

$$\lambda_{cg} | \, Y_{cg}, \mu_{cg}, \hat{\sigma}_{cg} \sim \Gamma(\hat{\lambda}_{cg}, v_{cg})$$

$$Cor(\lambda_{cg}, \lambda_{cg'})$$

$$= Cor(\hat{\lambda}_{cg}, \hat{\lambda}_{cg'}) \times f_g \times f_{g'}$$

f_g has simple analytical formula.



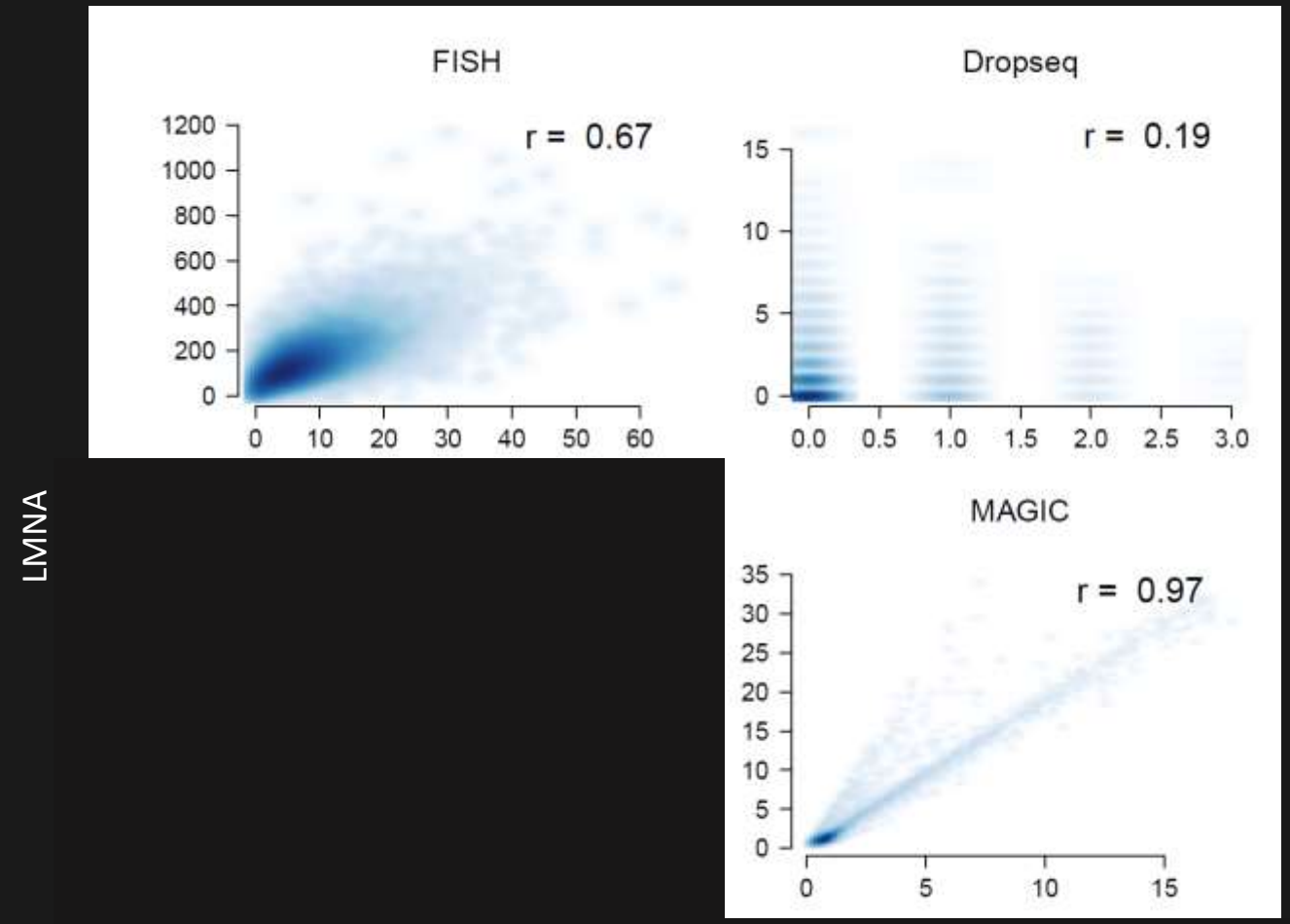
Recovering gene–gene relationships

$$\lambda_{cg} | \, Y_{cg}, \mu_{cg}, \hat{\sigma}_{cg} \sim \Gamma(\hat{\lambda}_{cg}, v_{cg})$$

$$Cor(\lambda_{cg}, \lambda_{cg'})$$

$$= Cor(\hat{\lambda}_{cg}, \hat{\lambda}_{cg'}) \times f_g \times f_{g'}$$

f_g has simple analytical formula.



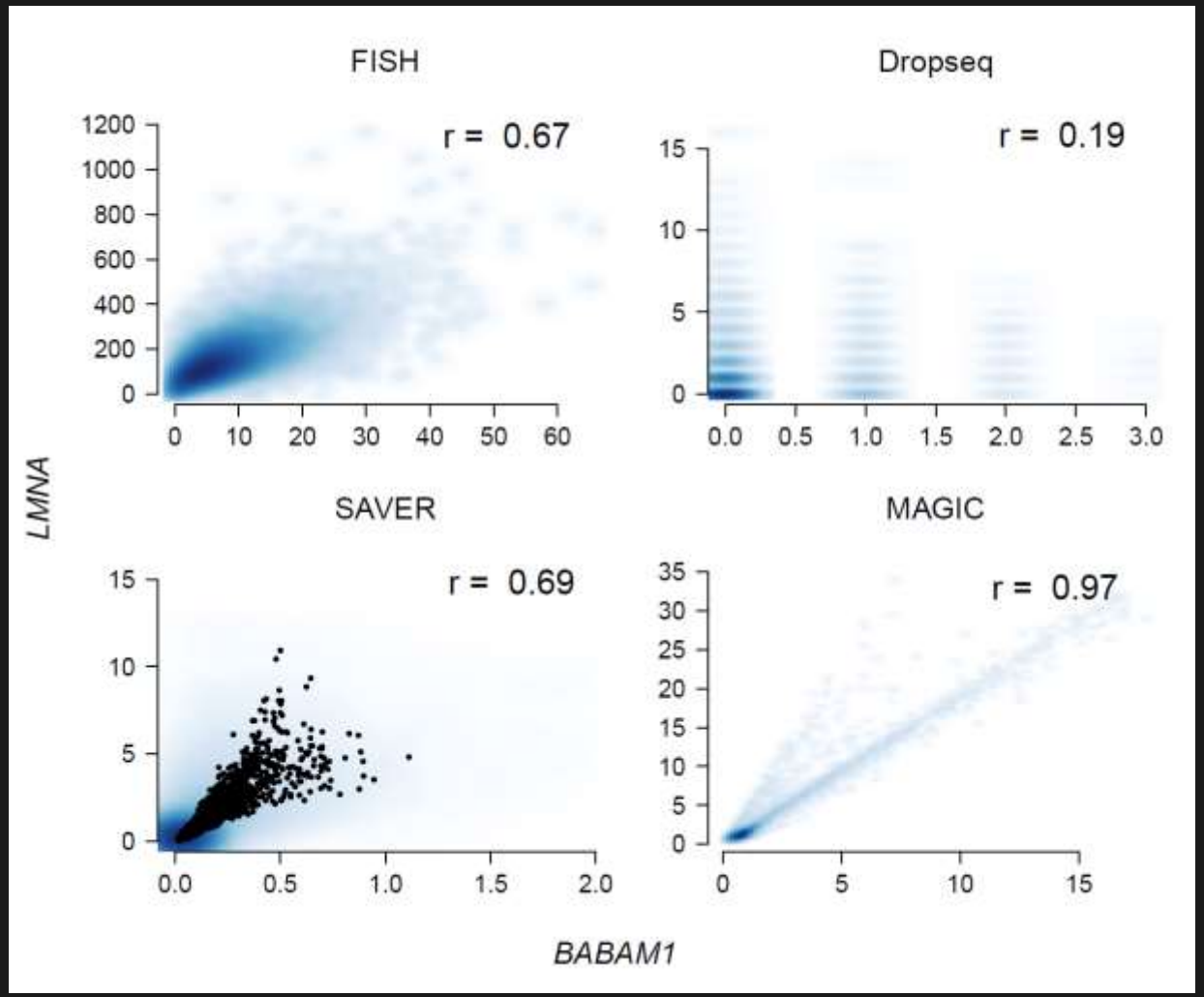
Recovering gene–gene relationships

$$\lambda_{cg} | \, Y_{cg}, \mu_{cg}, \hat{\sigma}_{cg} \sim \Gamma(\hat{\lambda}_{cg}, v_{cg})$$

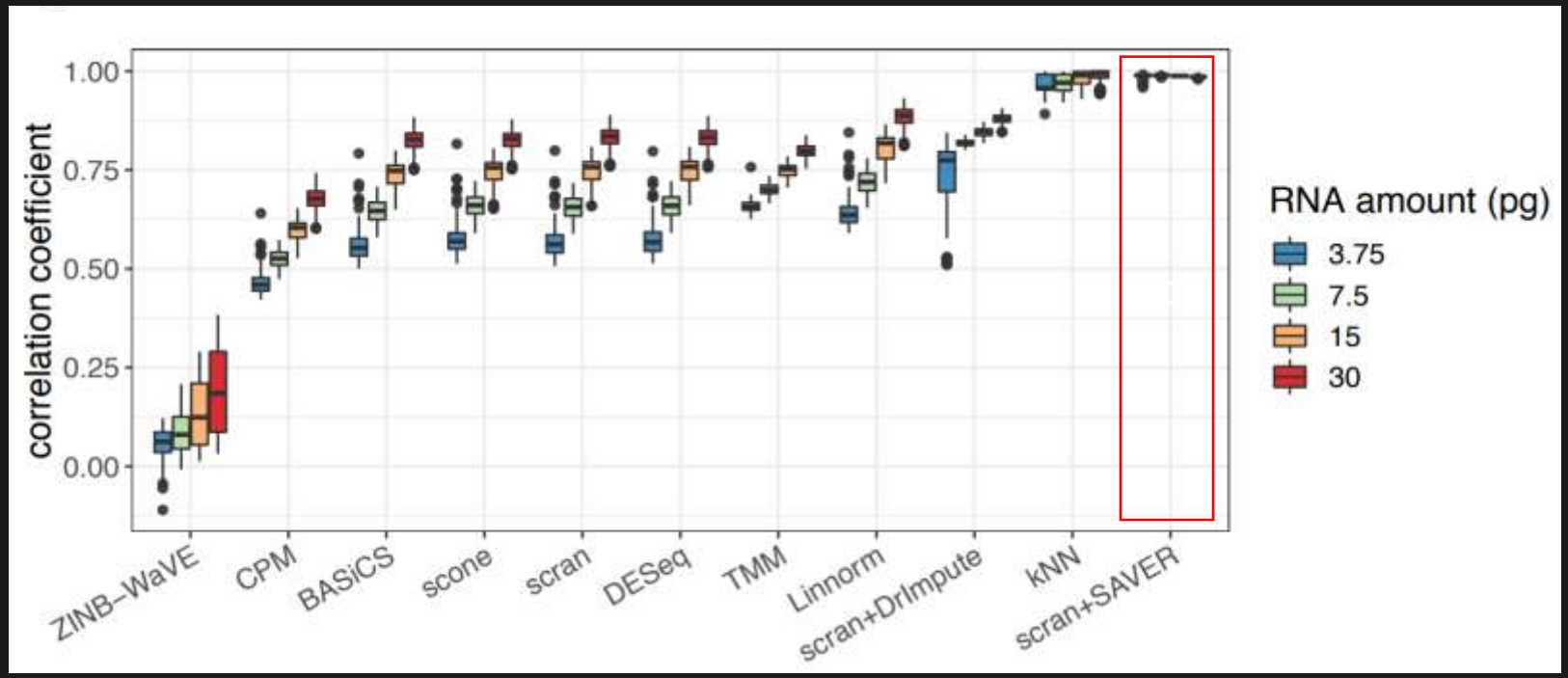
$$Cor(\lambda_{cg}, \, \lambda_{cg'})$$

$$= Cor(\hat{\lambda}_{cg}, \, \hat{\lambda}_{cg'}) \times f_g \times f_{g'}$$

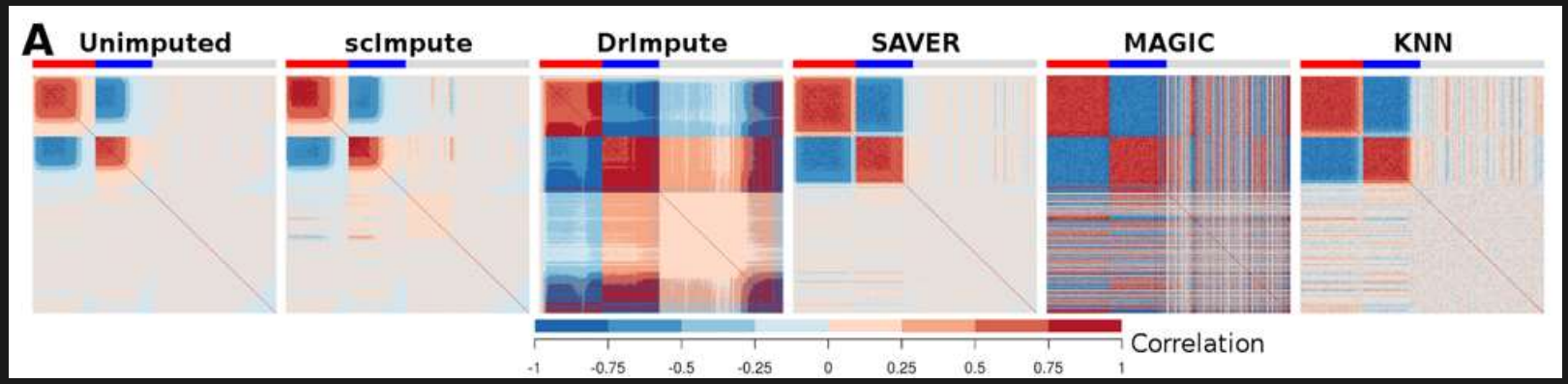
f_g has simple analytical formula.



The SAVER framework has been validated by third parties



Andrews and Hemberg. F1000 Research (2019)



Tian et al. scRNA-mixology (2018)

Take Home Messages

1. Transfer learning in single-cell transcriptomics improves data denoising and pattern discovery

   https://www.biorxiv.org/content/10.1101/457879v2

2. Denoise your single cell transcriptomics data using our gateway:

   https://singlecell.wharton.upenn.edu/saver-x/

3. Put your statistical hat on while using the denoised values for identifying new biomarkers (target cell types *and/or* genes)

This Work is the Brain Child of…

Nancy R. Zhang

Jingshu Wang

Mo Huang

NIH National Institutes of Health

NSF

Zilu Zhou, University of Pennsylvania

Chengzhong Ye, Tsinghua University

Gang Hu, Nankai University

Wharton Research Computing Staff

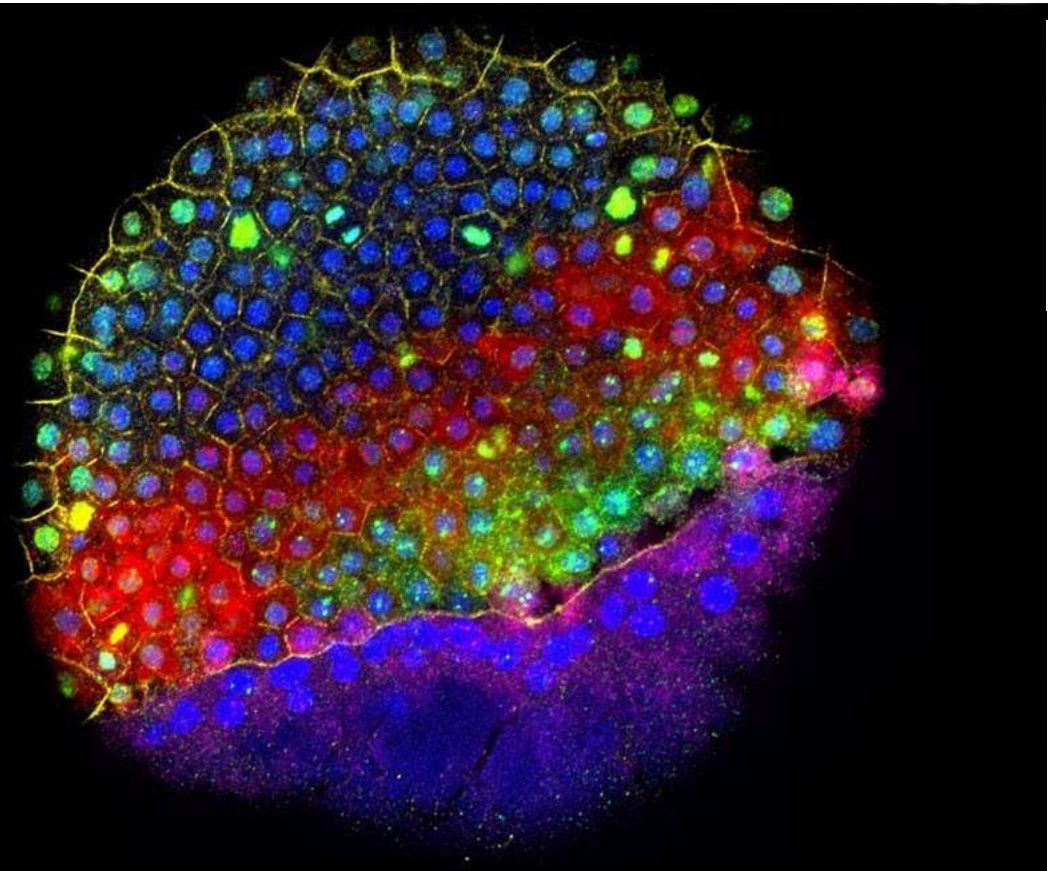
XSEDE Extreme Science and Engineering Discovery Environment

Blavatnik Family Foundation Fellowship

Vecteezy

Single-cell profiling technologies → Single-cell data → Groundbreaking discoveries

Science AAAS



2018 BREAKTHROUGH *of the* YEAR

Biomedical expertise
Artificial intelligence
Complex data interpretation