

Acknowledgement of Country

The University of Queensland (UQ) acknowledges the Traditional Owners and their custodianship of the lands on which we meet.

We pay our respects to their Ancestors and their descendants, who continue cultural and spiritual connections to Country.

We recognise their valuable contributions to Australian and global society.



General Information:

- We are currently located in Building 69



Emergency evacuation point

- Food court and bathrooms are located in Building 63
- If you are experiencing cold/flu symptoms or have had COVID in the last 7 days please ensure you are wearing a mask for the duration of the module



Data Agreement

To maximize your learning experience, we will be working with genuine human genetic data, during this module.

Access to this data requires agreement to the following in to comply with human genetic data ethics regulations

Please email pctgadmin@imb.uq.edu.au with your name and the below statement to confirm that you agree with the following:

“I agree that access to data is provided for educational purposes only and that I will not make any copy of the data outside the provided computing accounts.”

Desktop Access

For non-UQ attendees, you are provided with a registration instruction for a guest account (A4 paper).

After you have completed the online registration, use the provided Username and the Password that you set to log into the desktop.

Cluster Access

- You have all been provided with login details to computing resources needed for the practical component
- An SSH terminal is needed to connect to the computing:
 - Windows: Install PuTTY
 - Hostname: as provided (203.101.228.xxx)
 - User: as provided
 - Check Connection > SSH > X11 > Enable X11 forwarding
 - Mac/Linux: Use the terminal
 - `ssh -X <user>@203.101.228.xxx`
- If interactive R plotting does not work on your machine, you can generate plot on the server and then download
 - Windows: use WinSCP -> enter login information
 - Or use Command Prompt -> `sftp <user>@203.101.228.xxx`
 - `get xxx.pdf` and the file will be in your user directory

Module 2 Cellular Transcriptomics

Room 304, Building 69

Quan Nguyen, Claire Cheng, Jacky Xie, Onkar Mulay

Slides and Practical notes:

https://cnsgenomics.com/data/teaching/GNGWS23/model*/

Day 1: June 19th 2023

Lecture (Morning; single cell data and theory for common analyses)

8-8:20am		Introduction to participants and instructors	All
8:20-8:40am		Introduction scRNA and spatial transcriptomics data	Quan Nguyen
8:40-9:00am		Data exploratory analysis and preprocessing	Claire
9-9:20am		Data normalisation	Quan Nguyen
9:20-9:40am		Dimensionality reduction & Clustering	Onkar
9:40-10am		Break	
10:00-12:00		Practical 1	Quan, Claire, Onkar, Jacky
1:00-1:20pm		Differential expression analysis	Onkar-Quan Nguyen
1:20-1:30		Cell type analysis	Claire
1:30-1:45		eQTL single cell/tissue/bulk	Quan Nguyen
1:50am-2:00pm		Questions and discussions and future perspectives	
2:00-4:00		Practical	
4:00-4:30		Introduce Spatial data	Quan
4:30-4:50		Data structure	Jacky

Day 2 : June 20th 2023

Lecture Spatial transcriptomics analysis

8:00-8:15am	Introduction to machine learning: machine learning vs statistical learning vs artificial intelligence in genomics and biological imaging	Quan Nguyen
8:15-8:30am	Introduction to machine learning: key concepts	Quan Nguyen
8:30-9:15am	Machine learning in single cell and spatial data	Jacky
9:15-9:30	Interpretability	Onkar
9:30-9:45	Uncertainty analysis and general discussion	Quan
9:45-10:00pm	Break	
10-12:00 pm	Practical	All

Single cell informatics

Scale

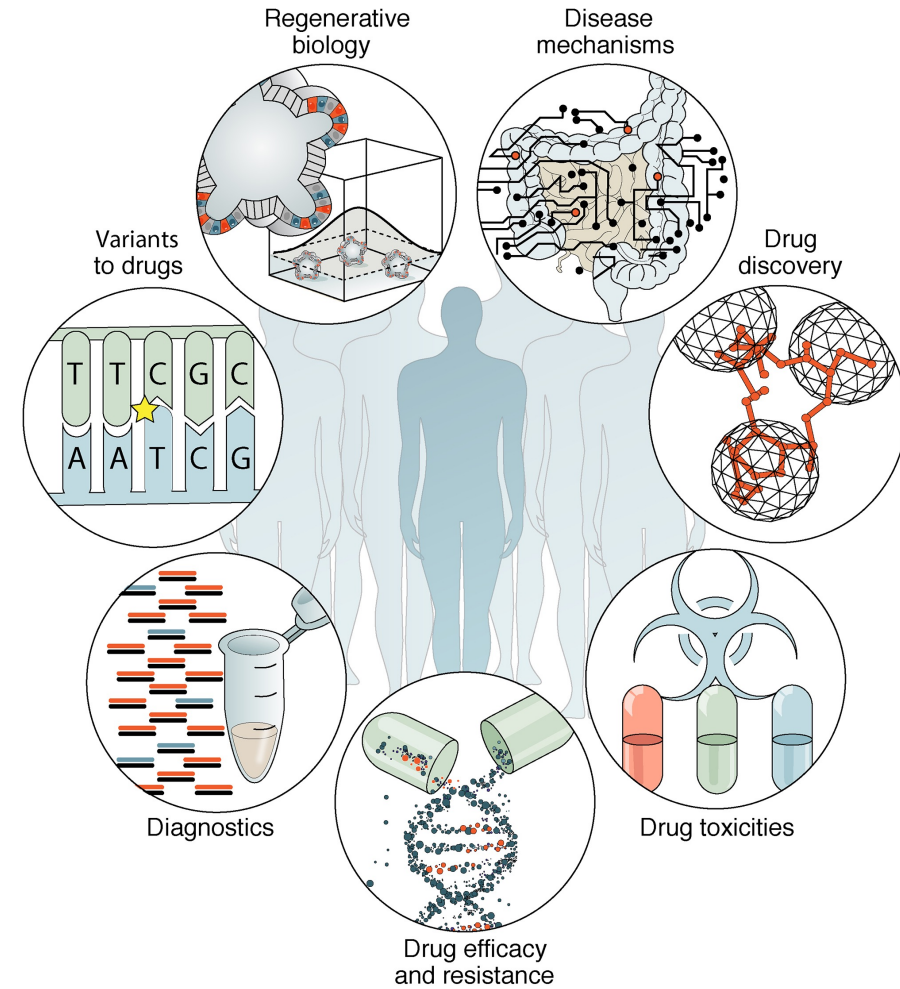


Resolution

INFORMATICS



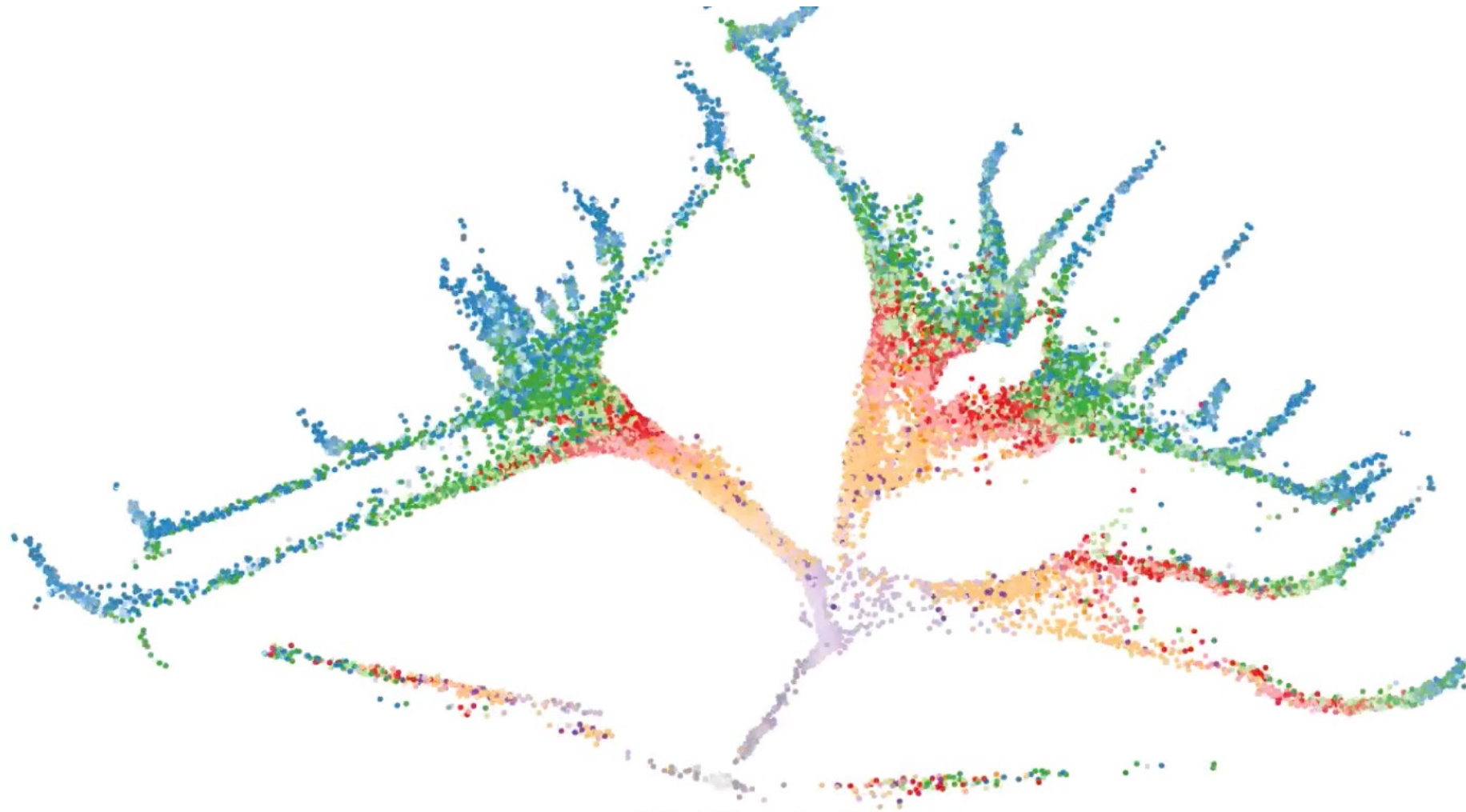
Precision Genomics Medicine



The G&G Cellomics Team

Quan Nguyen, Claire Cheng, Jacky X, Onkar Mulay

General introduction single cell and spatial transcriptomics

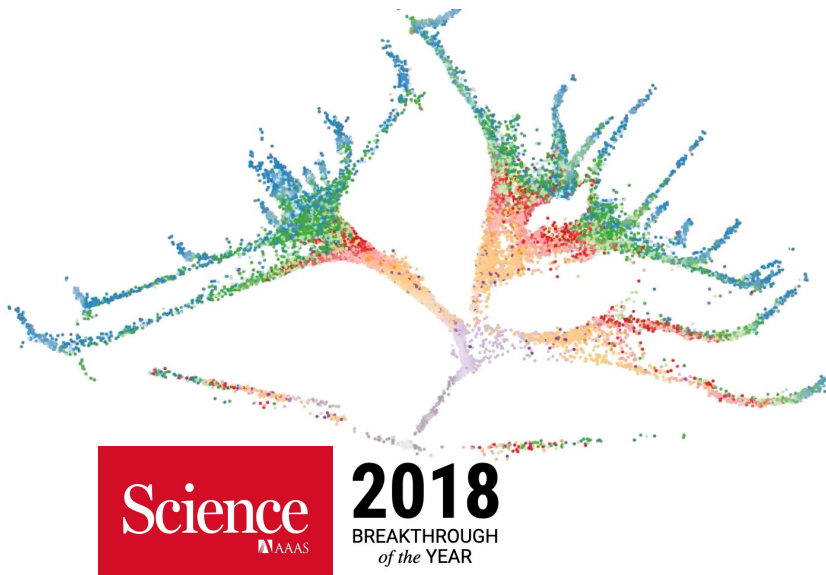


2018
BREAKTHROUGH
of the YEAR

The single-cell revolution is just starting

Advanced genomics technologies

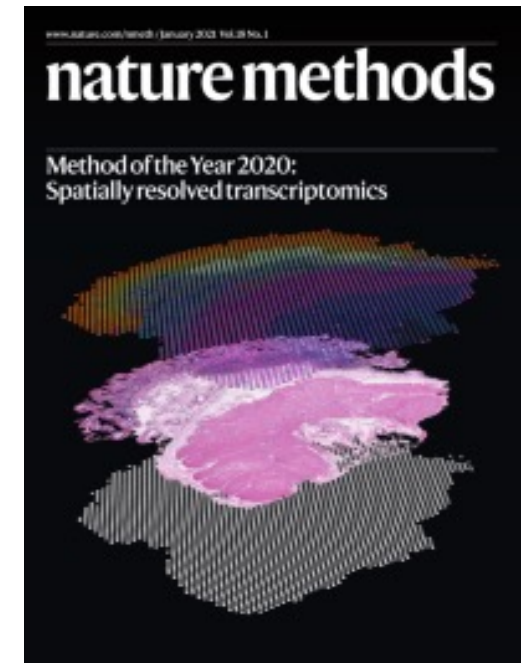
2018: Single Cell Transcriptomics



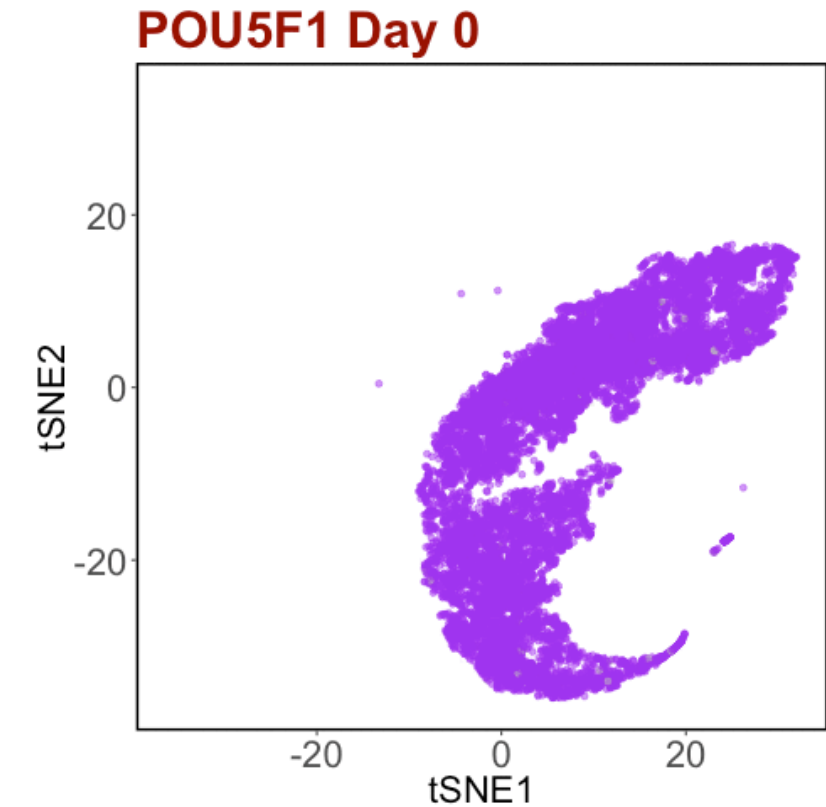
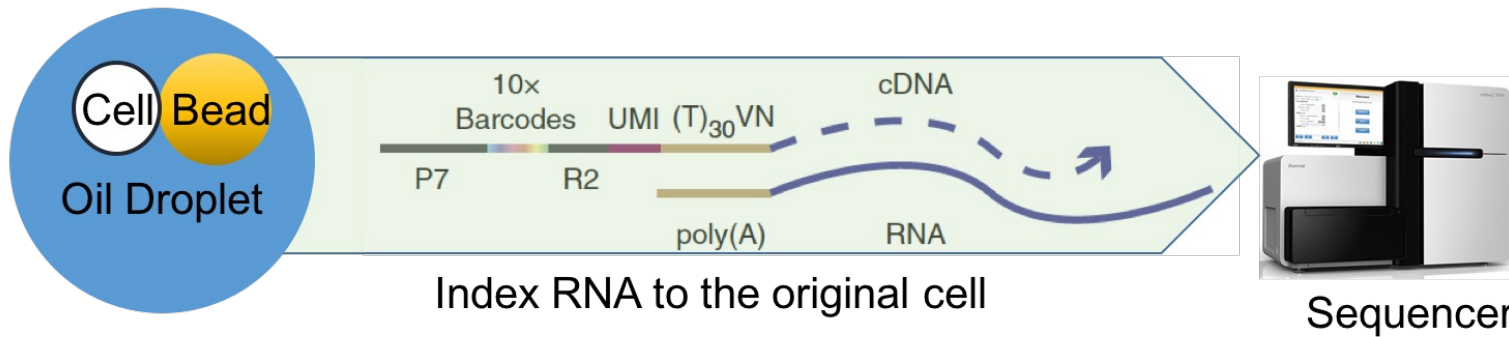
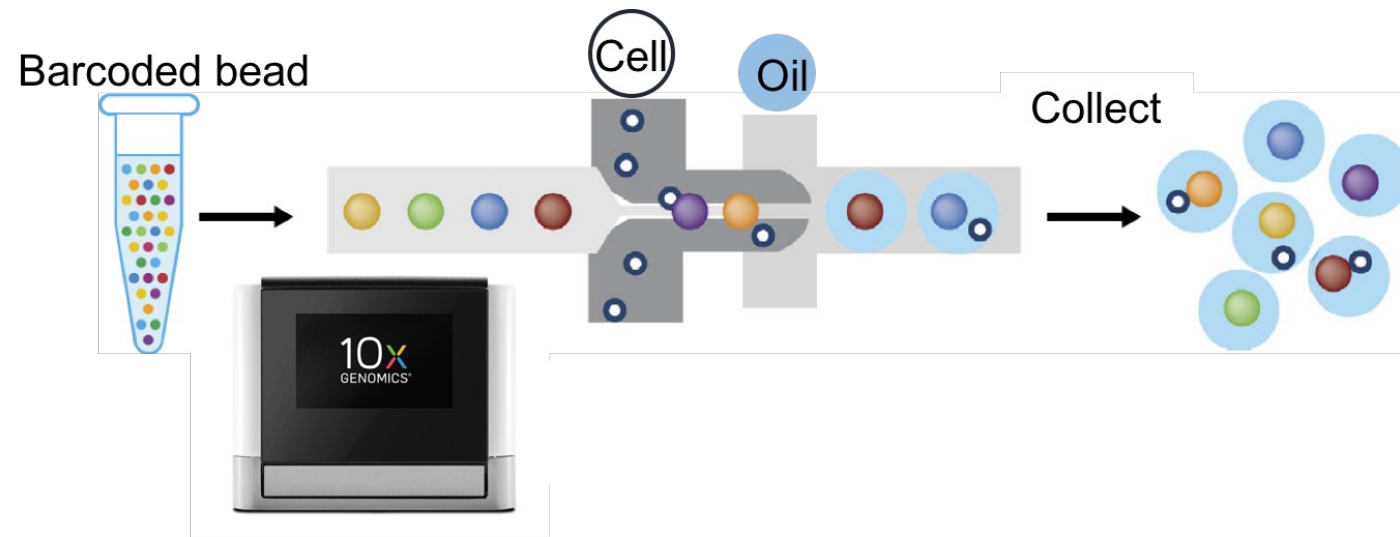
2019: Single Cell Multiomics



2020: Spatial Transcriptomics



Single cell RNA sequencing



- Single-cell RNA sequencing (scRNA-seq) measures thousands of genes in a separate cell
- How: 3 barcoding steps for sample, cell and RNA molecule
- Scale: bulk RNA-seq (5 samples) vs. scRNA-seq (45 K cells), a ~900 times bigger gene count matrix

Multiplexing and storage of single cell samples



Fix & permeabilize samples

Hybridize probes

Pool (optional)

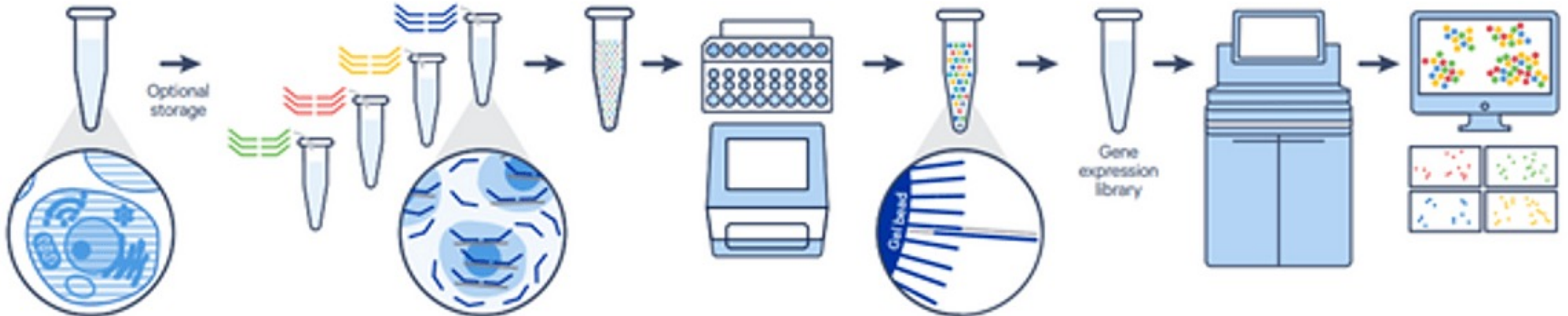
Partition in GEMs

Ligation & extension in GEMs

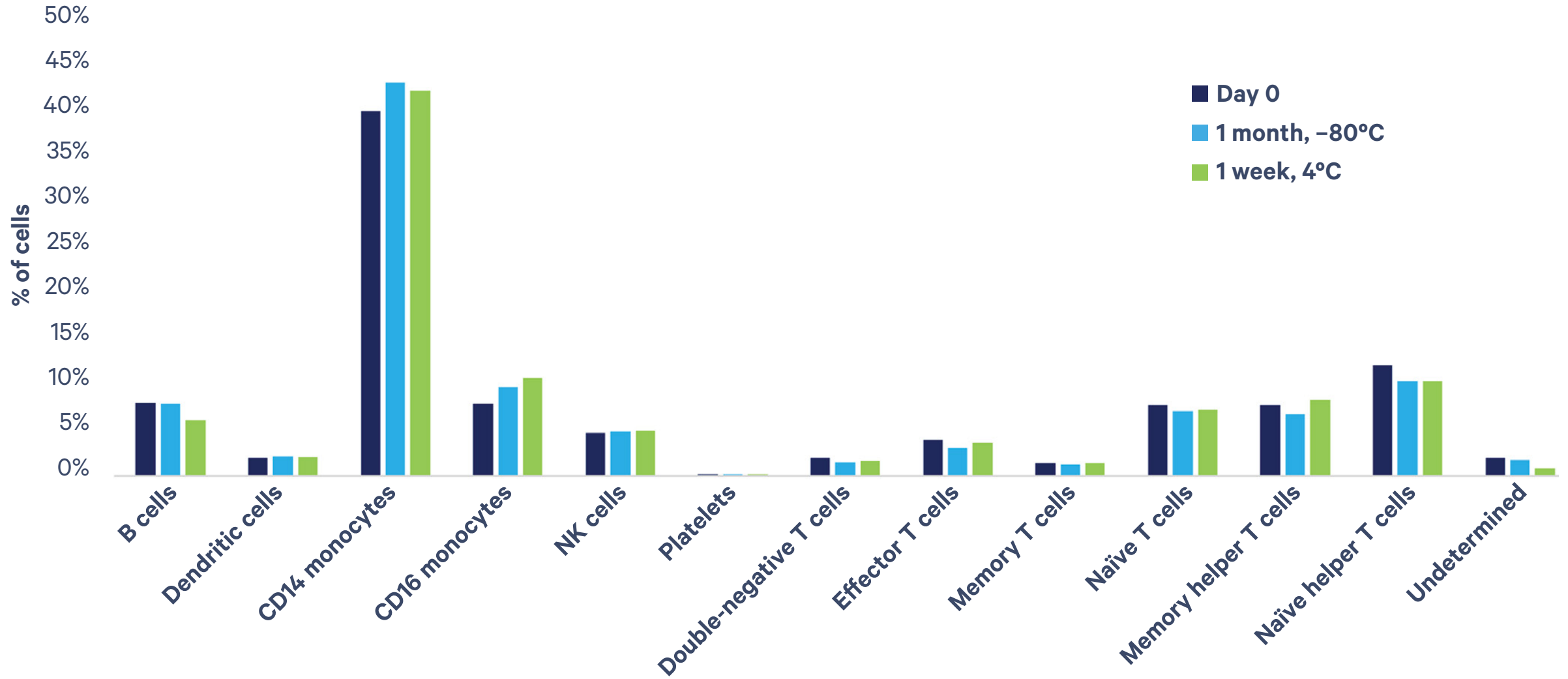
Library construction

Sequencing

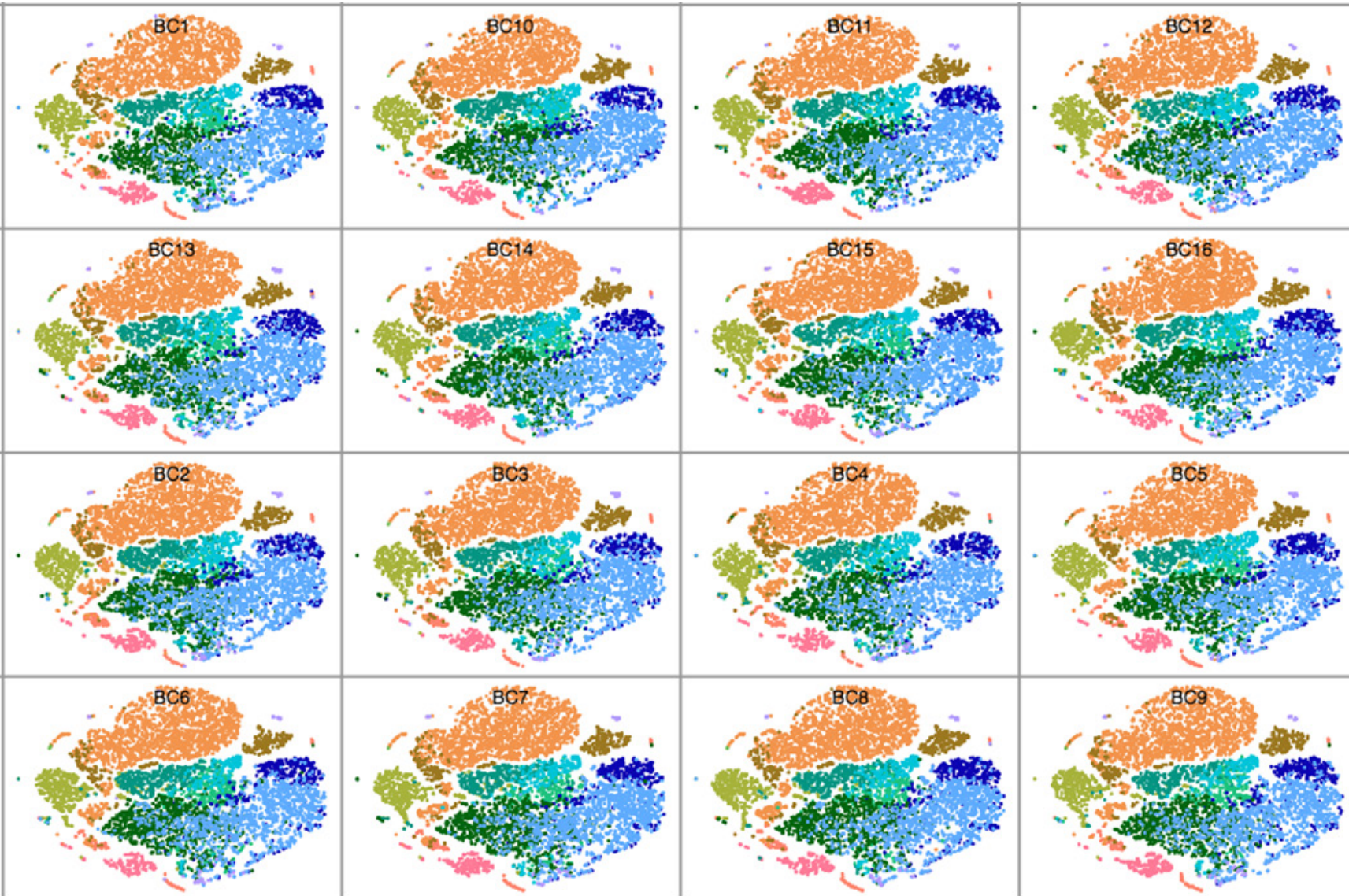
Data analysis



Multiplexing and storage of single cell samples



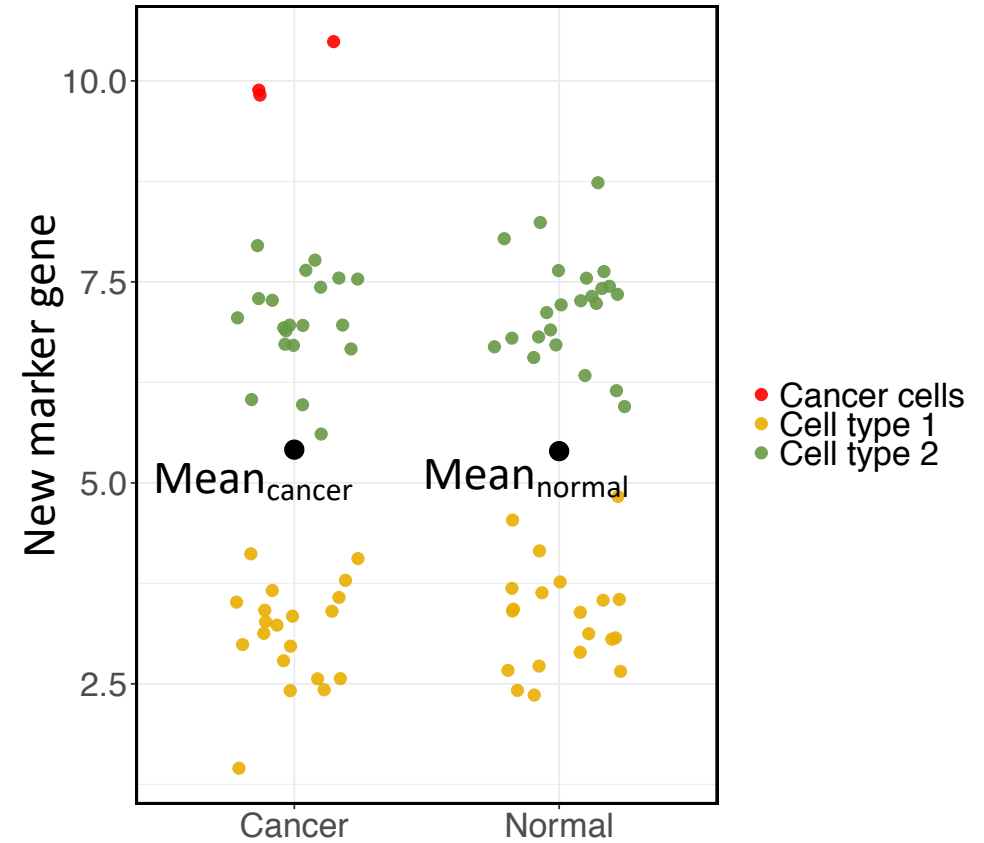
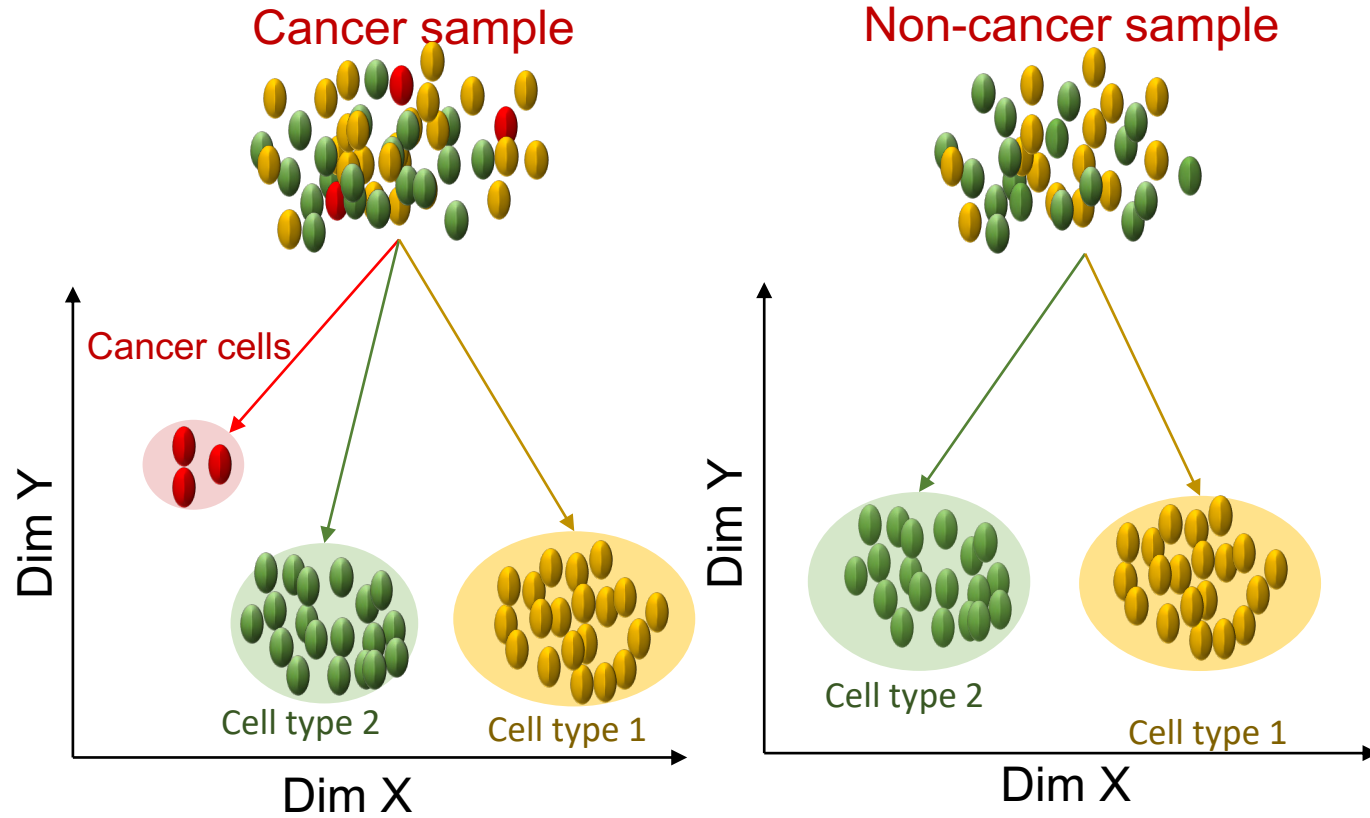
Multiplexing and storage of single cell samples



- Platelets
- Double-negative T cells
- Effector T cells
- Memory T cells
- Naïve T cells
- Memory helper T cells
- Naïve helper T cells
- Undetermined

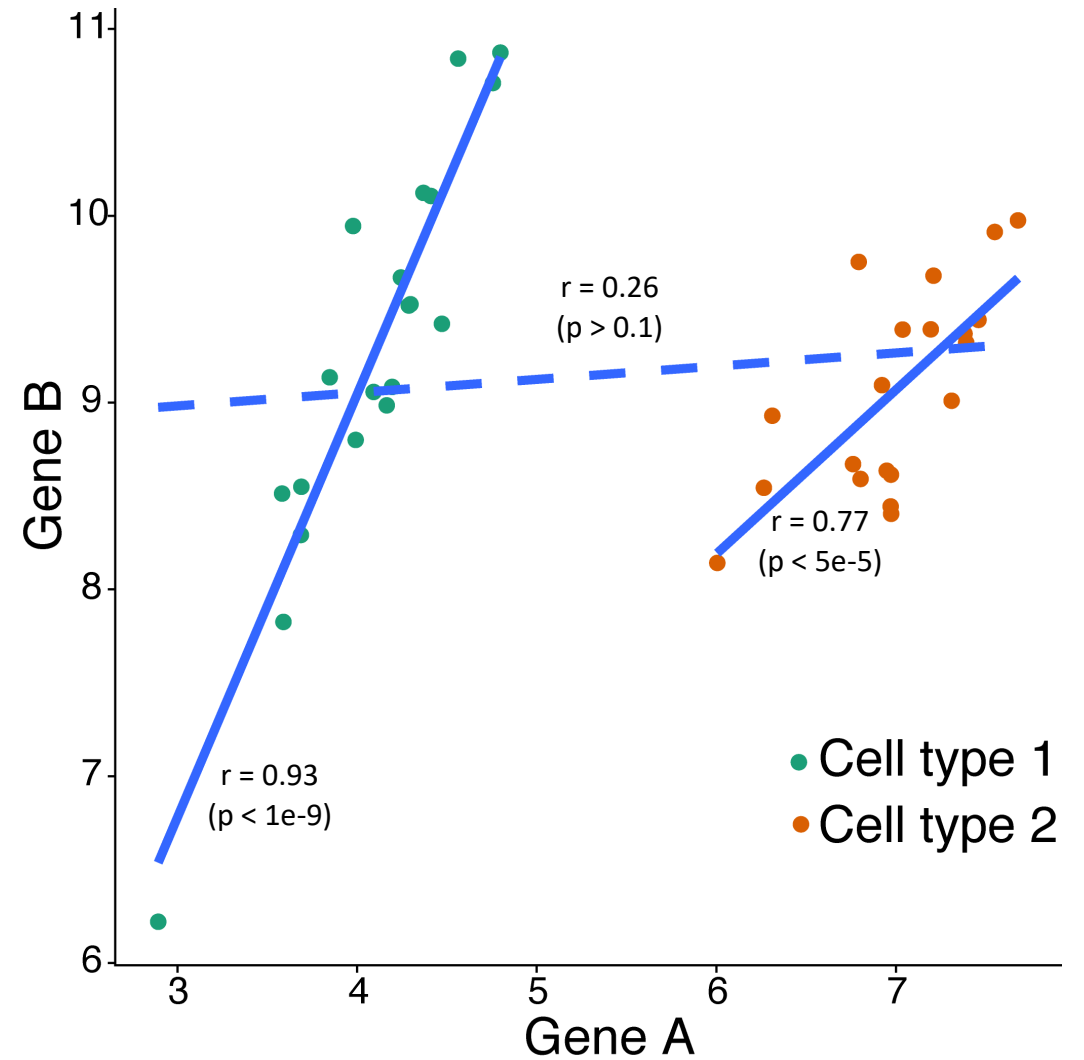
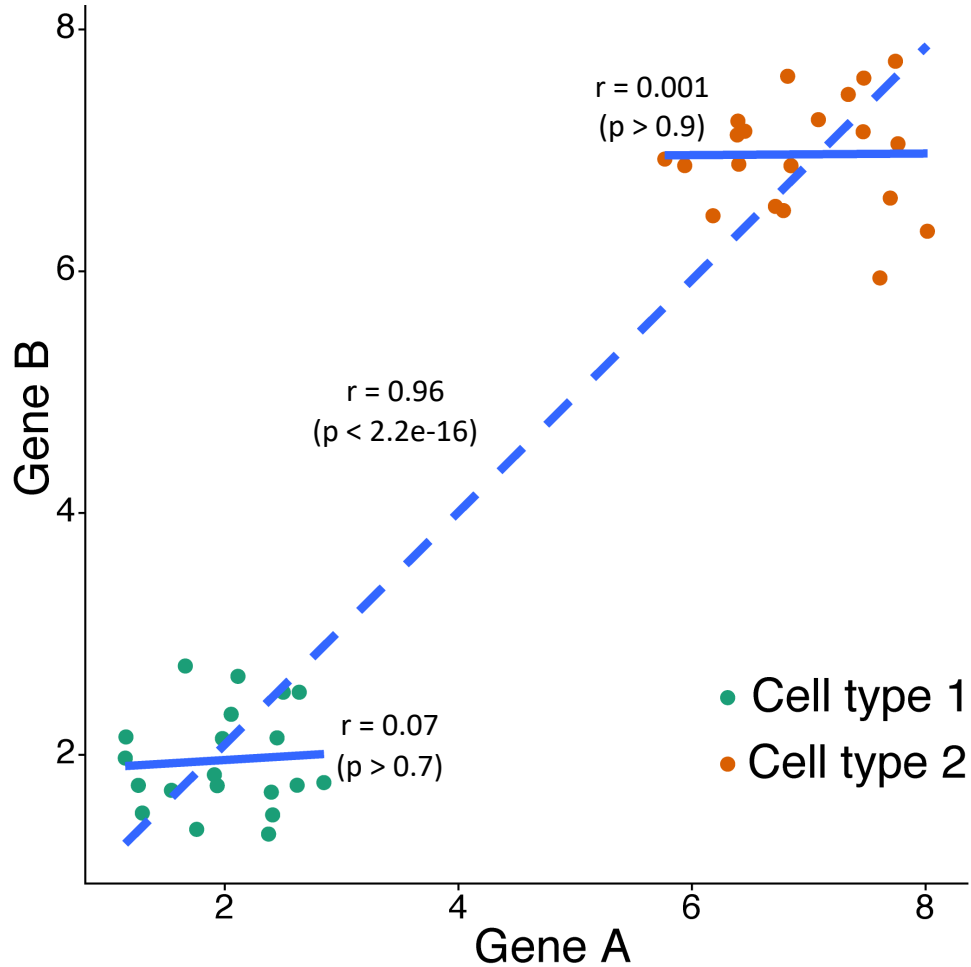


Disease at single-cell resolution



- Bulk RNA sequencing: no difference in mean expression
- Single-cell sequencing: can detect higher expression in cancer cells

Genes correlation detected at cell-type level



- Different results in gene expression patterns when looking at combined or separate cell types (cell-type specific signals need scRNAseq data)

Spatial transcriptomics approach

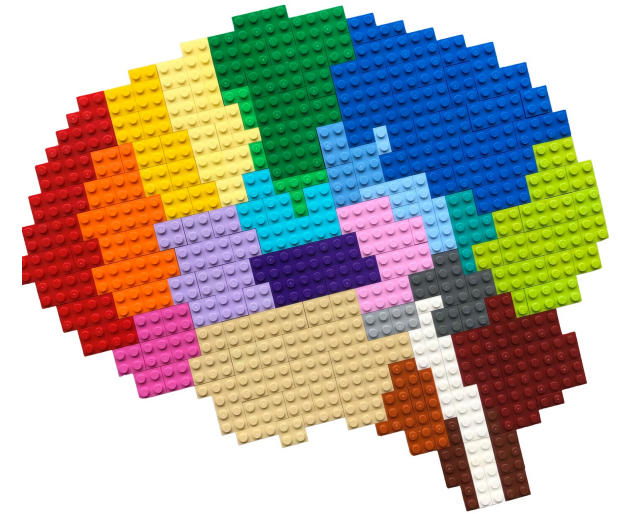
Bulk



Single cell



Spatial

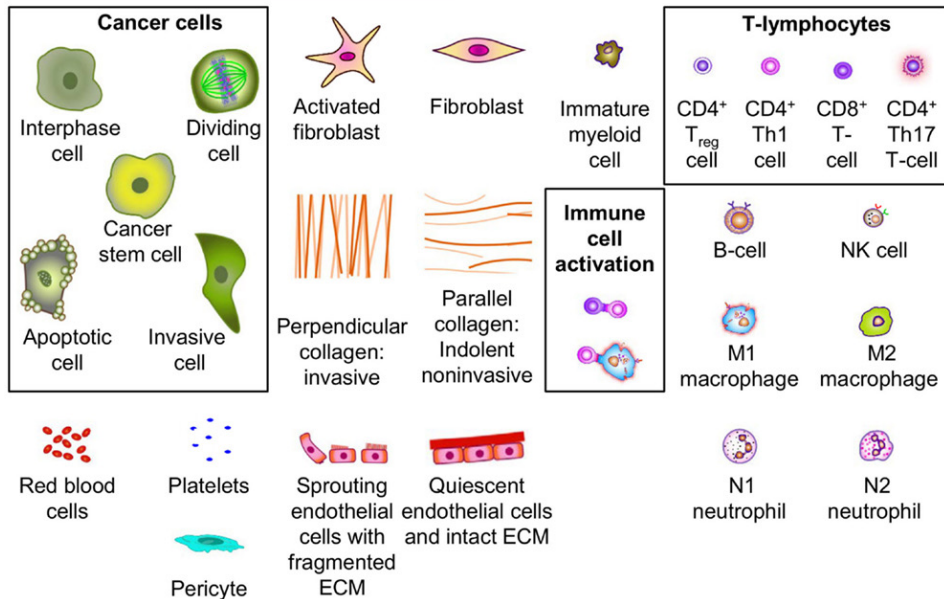
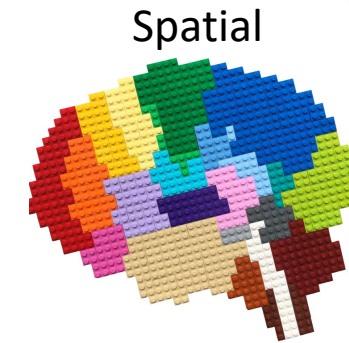
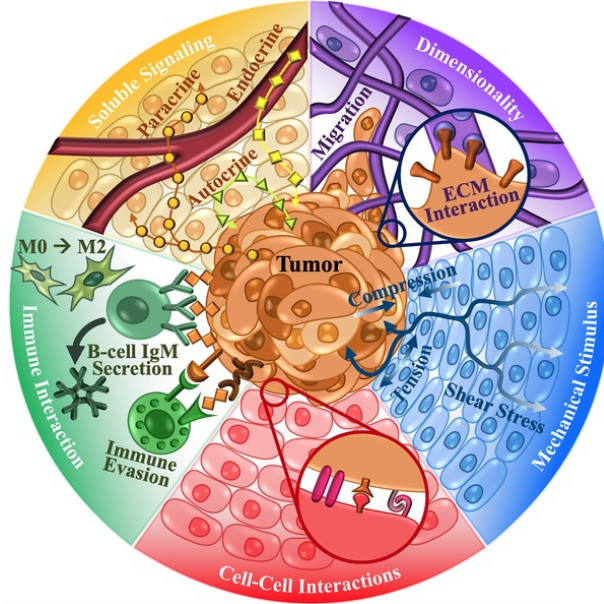
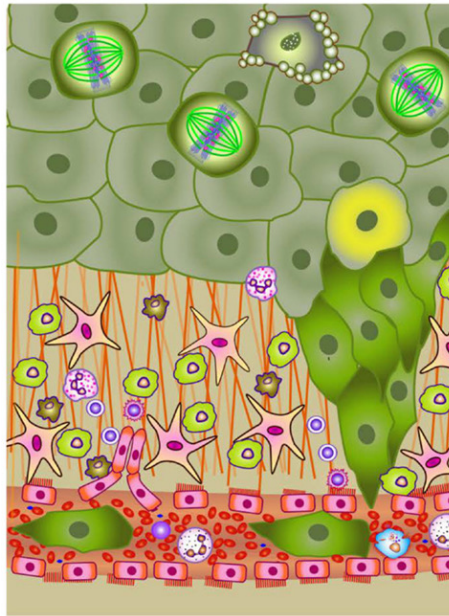


Lego:
(@boxia)

Fruit salad:
(@LGMartelotto)



Cellular ecosystem within a tissue



- Complex cellular ecosystem: cell-type composition, spatial organisation, cell-cell interaction, mechanical effect
- How to comprehensively investigate tissue ecosystem?

Analysis tools for single cells and spatial data

Software programs

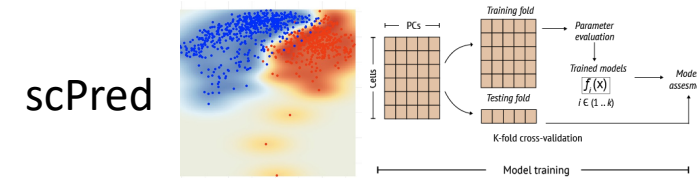
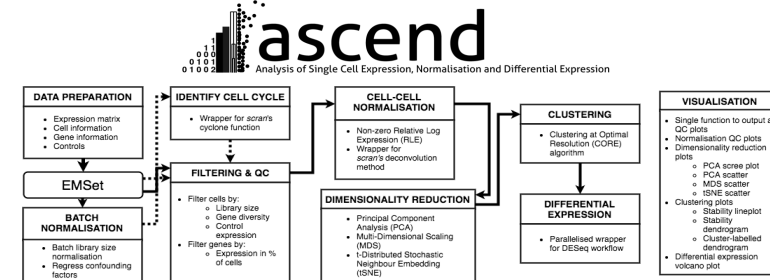
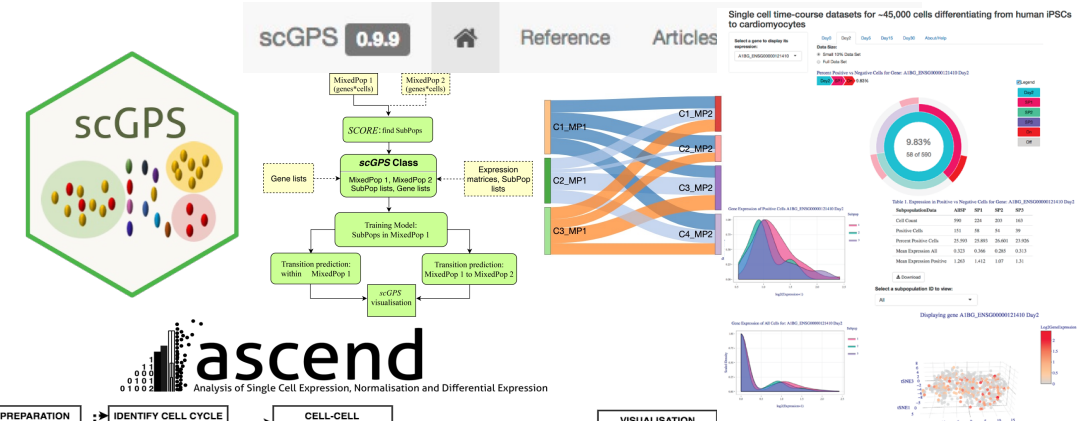
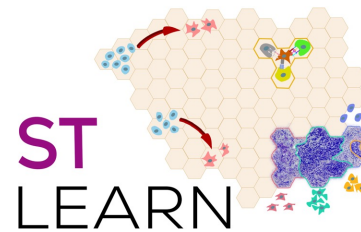
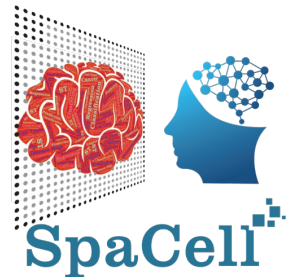
- scGPS: <https://github.com/BiomedicalMachineLearning/scGPS>
- ascend: <https://github.com/BiomedicalMachineLearning/ascend>
- scPred: <https://github.com/IMB-Computational-Genomics-Lab/scPred>
- CoreNET: <https://github.com/BiomedicalMachineLearning/CoreNET>
- HEMnet: <https://github.com/BiomedicalMachineLearning/HEMnet>
- scSplit: <https://github.com/ion-xu/scSplit>

scRNAseq visualisation

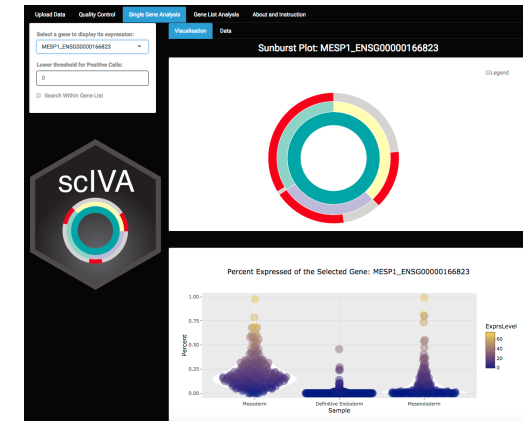
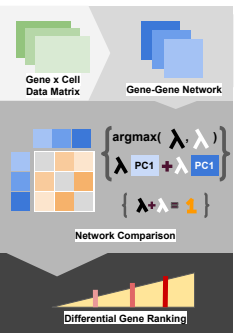
- HiPSC: <http://computationalgenomics.com.au/shiny/hipsc>
- Hipsc2cm: <http://computationalgenomics.com.au/shiny/hipsc2cm>
- scIVA: <http://computationalgenomics.com.au/shiny/scIVA/>

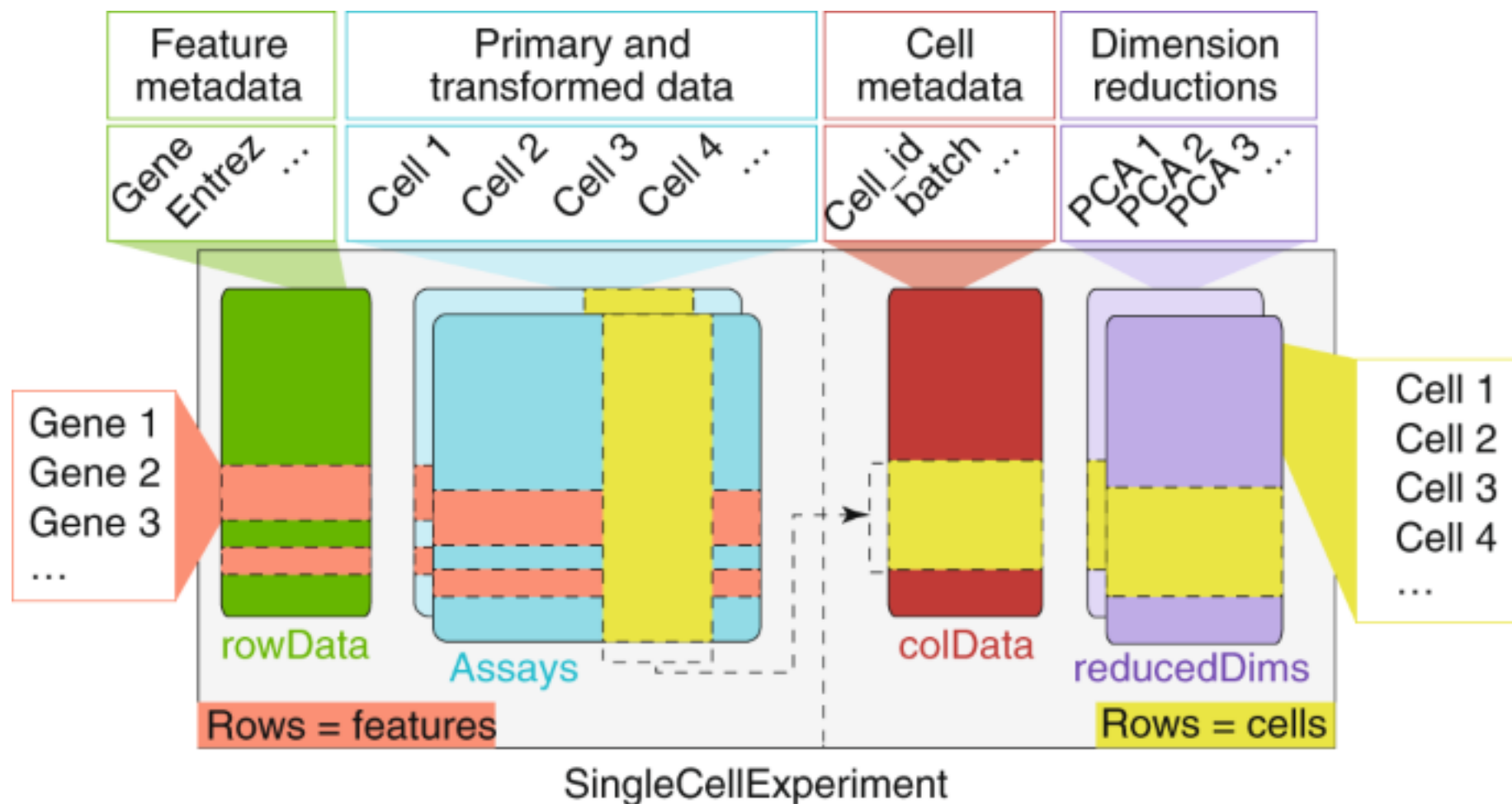
Spatial Transcriptomics

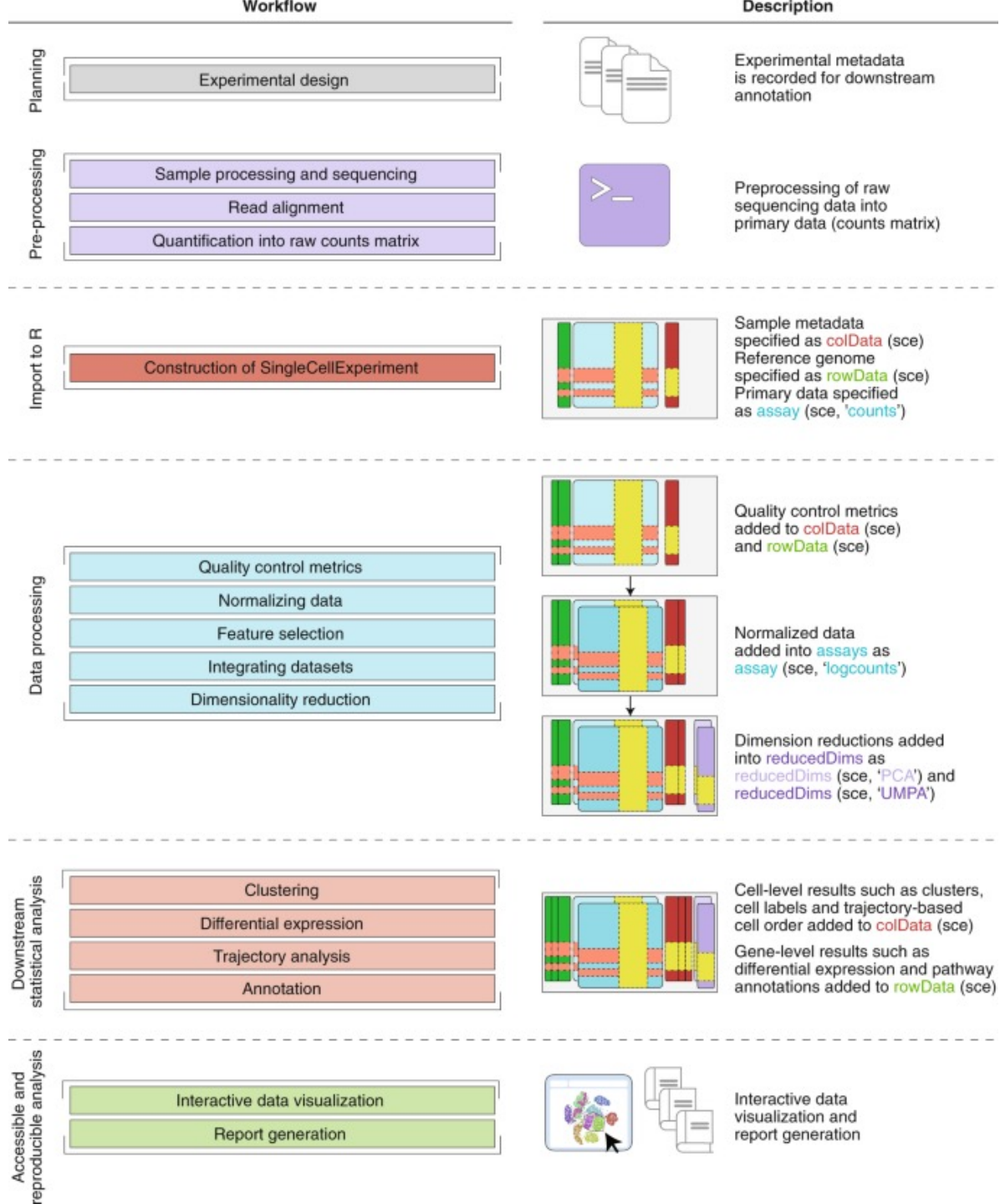
- SpaCell: <https://github.com/BiomedicalMachineLearning/Spacell>
- stLearn: <https://stlearn.readthedocs.io/en/latest/>



CoreNet







Data Preprocessing

Single cell data vs. bulk data

<https://github.com/IMB-Computational-Genomics-Lab/scIVA>

Upload Data
Quality Control
Single Gene Analysis
Gene List Analysis
About and Instruction

Upload Expression Matrix

Browse... expressionTestLarge.csv

Upload complete

Transpose Expression

Separator

Comma

Semicolon

Tab

Quote

None

Double Quote

Single Quote

Uploaded Expression Matrix

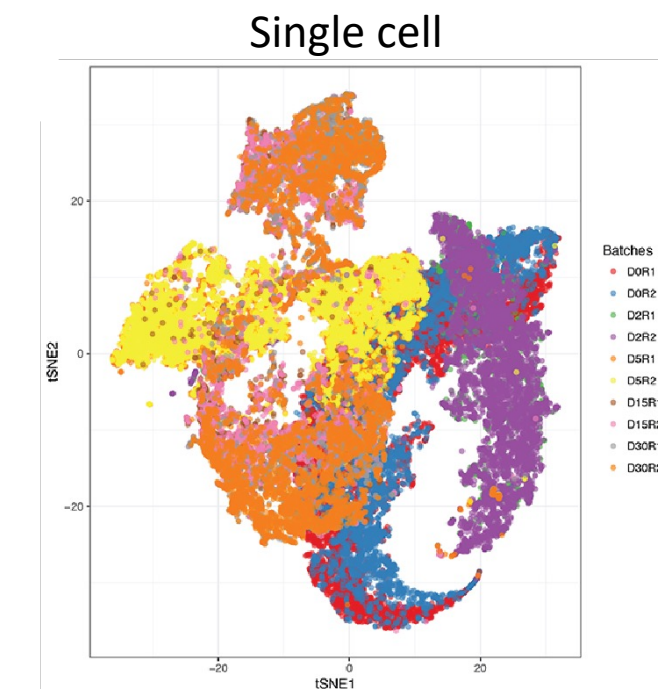
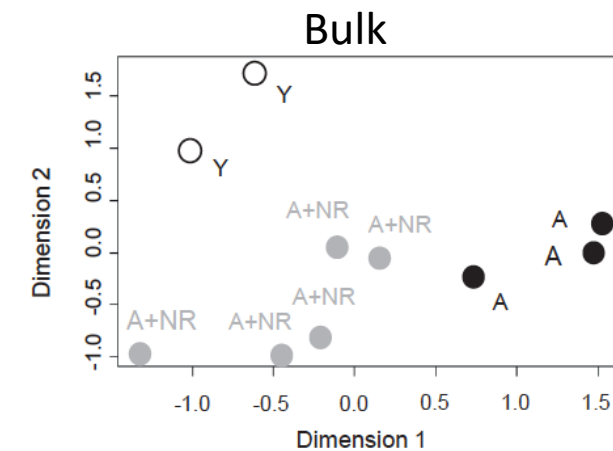
	1_AAACATACAGAATG-1	1_AAACATACCTTCTA-1	1_AAACATACGCAAGG-1	1_AAACATACGGGCAA-1	1_AAACATACGTCGAT-1
FO538757.1_ENSG00000279457	0.00	0.00	0.00	0.00	0.00
AP006222.2_ENSG00000228463	0.00	0.00	0.00	0.00	0.00
RP4-669L17.10_ENSG00000237094	0.00	0.00	0.00	0.00	0.00
RP11-206L10.9_ENSG00000237491	0.00	0.00	0.00	0.00	0.00
LINC00115_ENSG00000225880	0.00	0.00	0.00	0.00	0.00

No. of Genes

16561

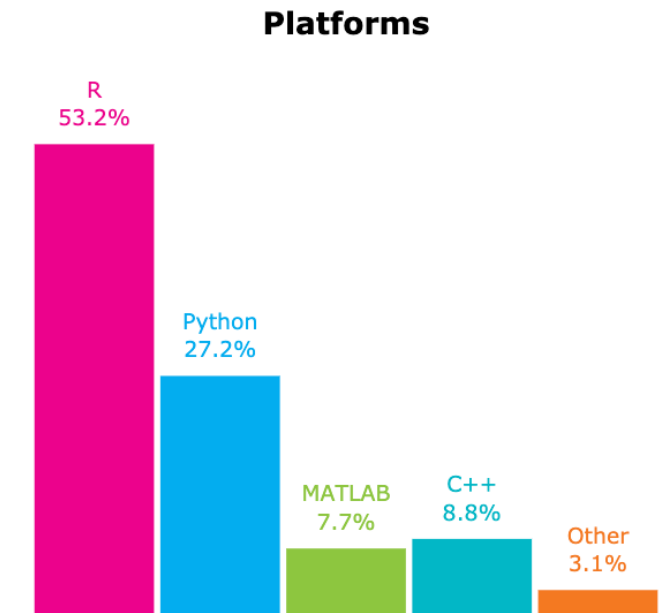
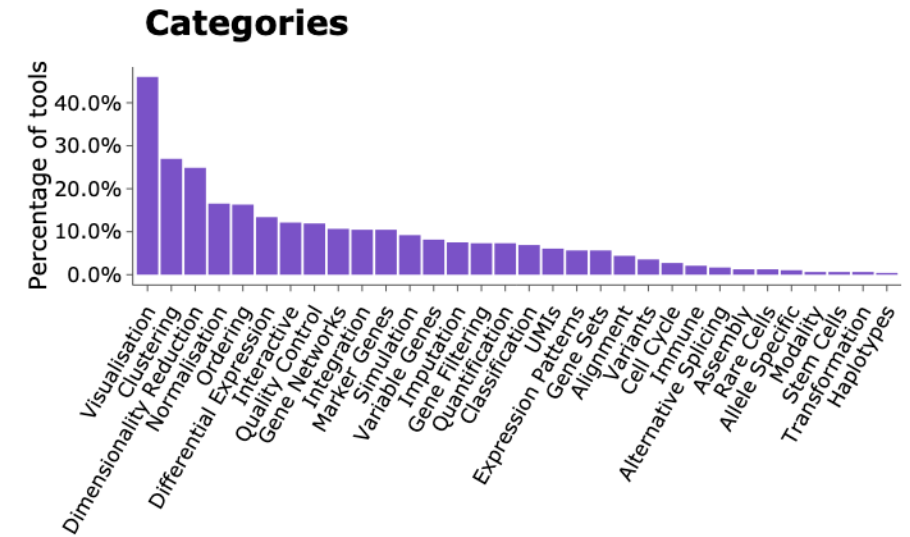
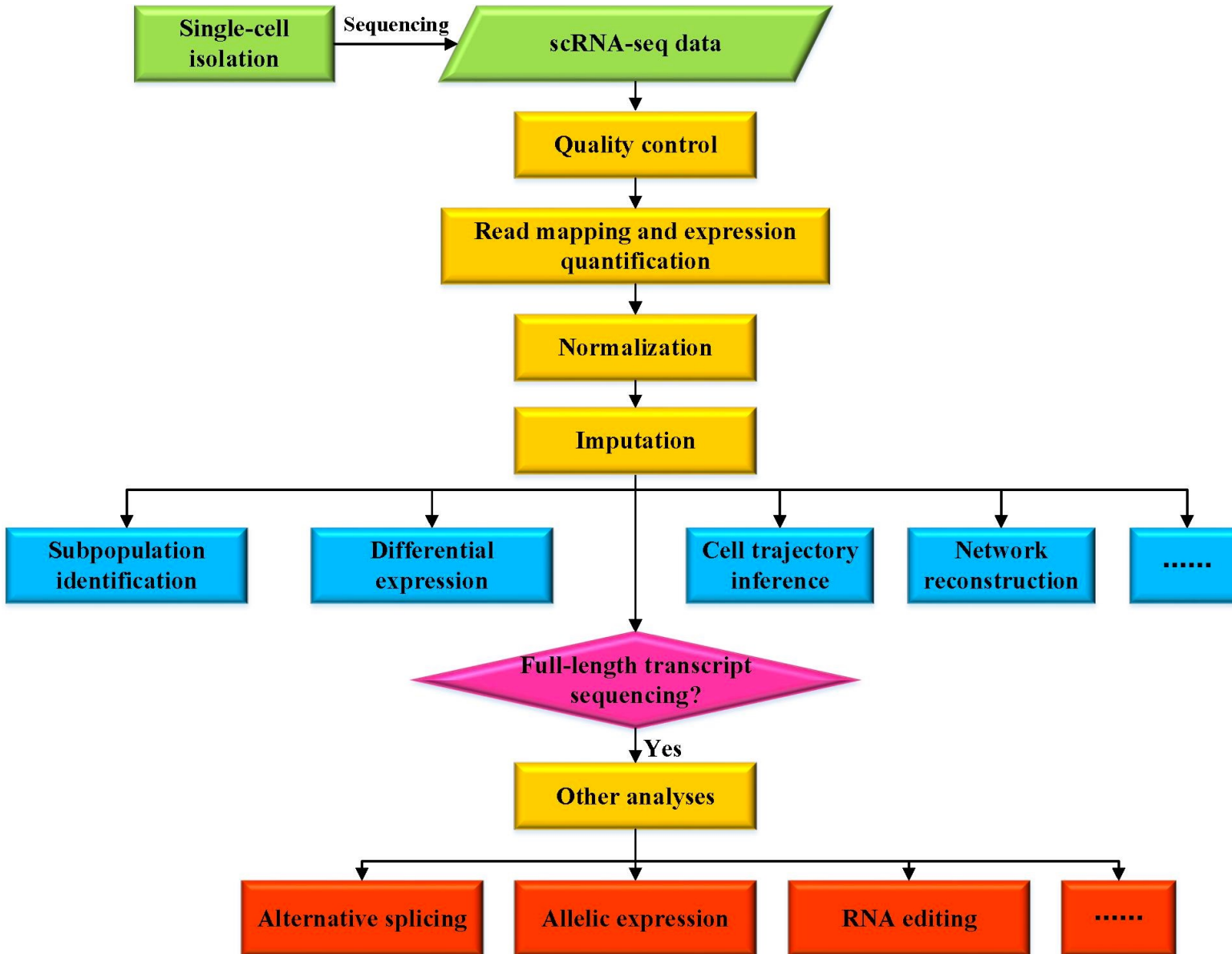
No. of Cells

13679

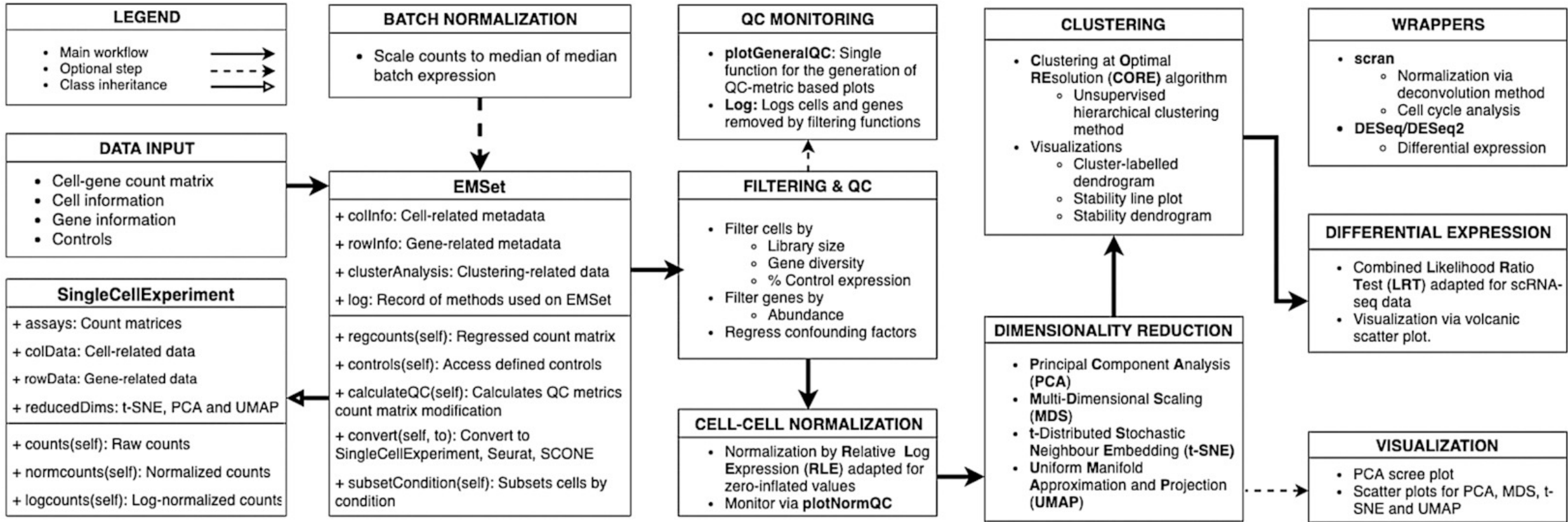


	Single cell	Bulk
Noisy data	Undetected genes (zero inflation)	Deep sequencing, most genes detected
Cell-cell variation	Measured	Not measured
Data size	Thousands of cells (1 cell ~ 1 bulk sample)	10-100 samples

Single cell data analysis



An analysis pipeline



Three main steps:

- 1) Data preprocessing and normalisation
- 2) Clustering to find subpopulations (a step applied in almost all cases)
- 3) Downstream analysis at cell-type specific level (genes, pathways, biological processes)

Analysis steps for the differentiation dataset

Sequenced 44,123 cells at 5 time points (10 samples)

↓ **QC and normalisation**

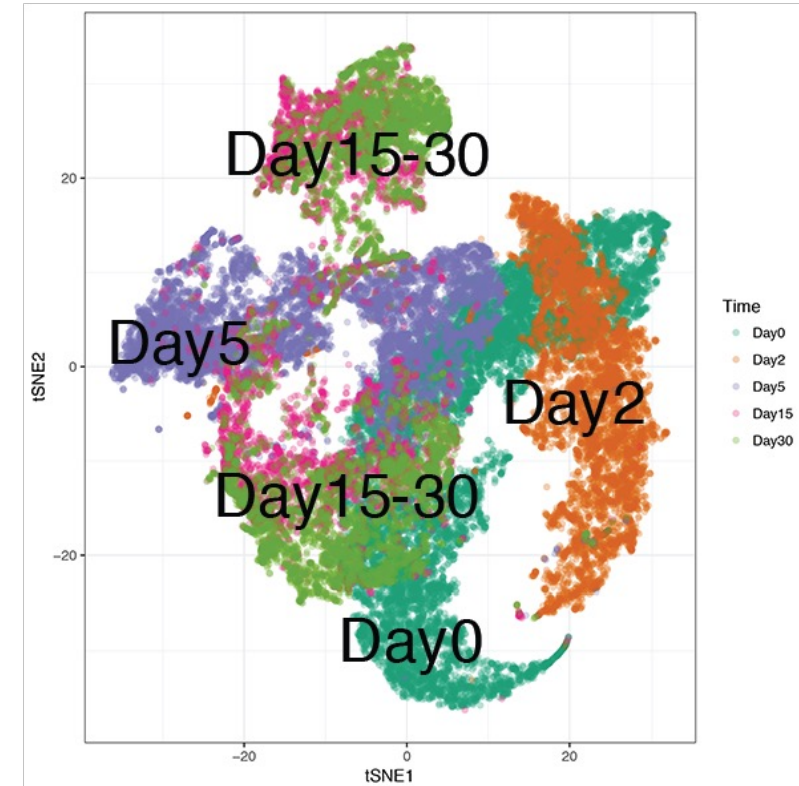
1. Data merging and normalising by batches (samples)
2. Data preprocessing (removing outlier cells and genes due to technical bias)
3. Cell-to-cell normalisation

↓ **Dimensionality reduction**

1. Dimensionality reduction (PCA, t-SNE, MDS, CIDR) and visualisation
2. Functionally evaluated scRNA data based on expression of known pluripotency and differentiation markers

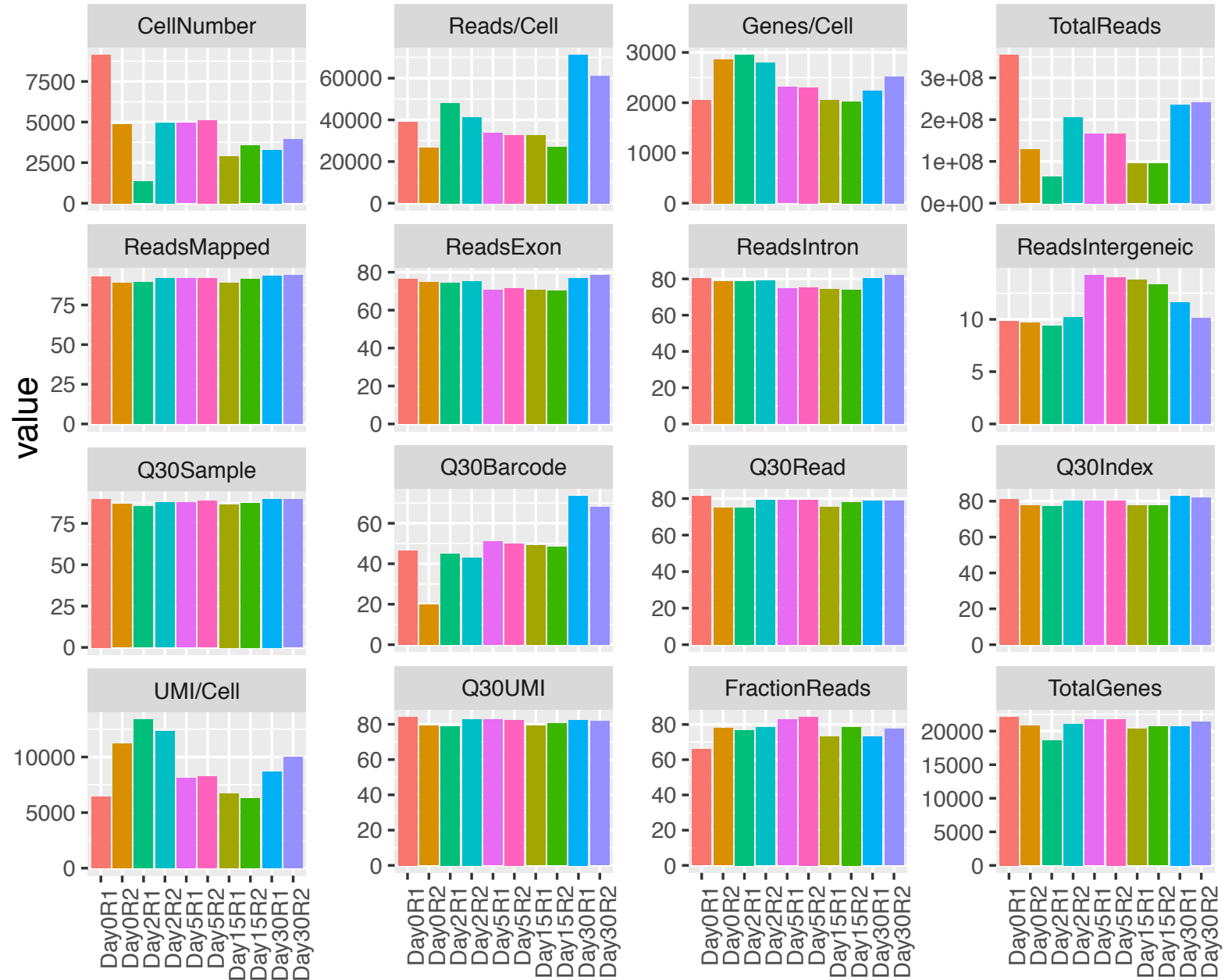
↓ **Clustering**

1. Developed a novel clustering method (CORE - Clustering at Optimal REsolution)
2. Implemented CORE to identify subpopulations within each time point
3. Validated CORE results by comparing with other methods and by functional analysis of each subpopulation

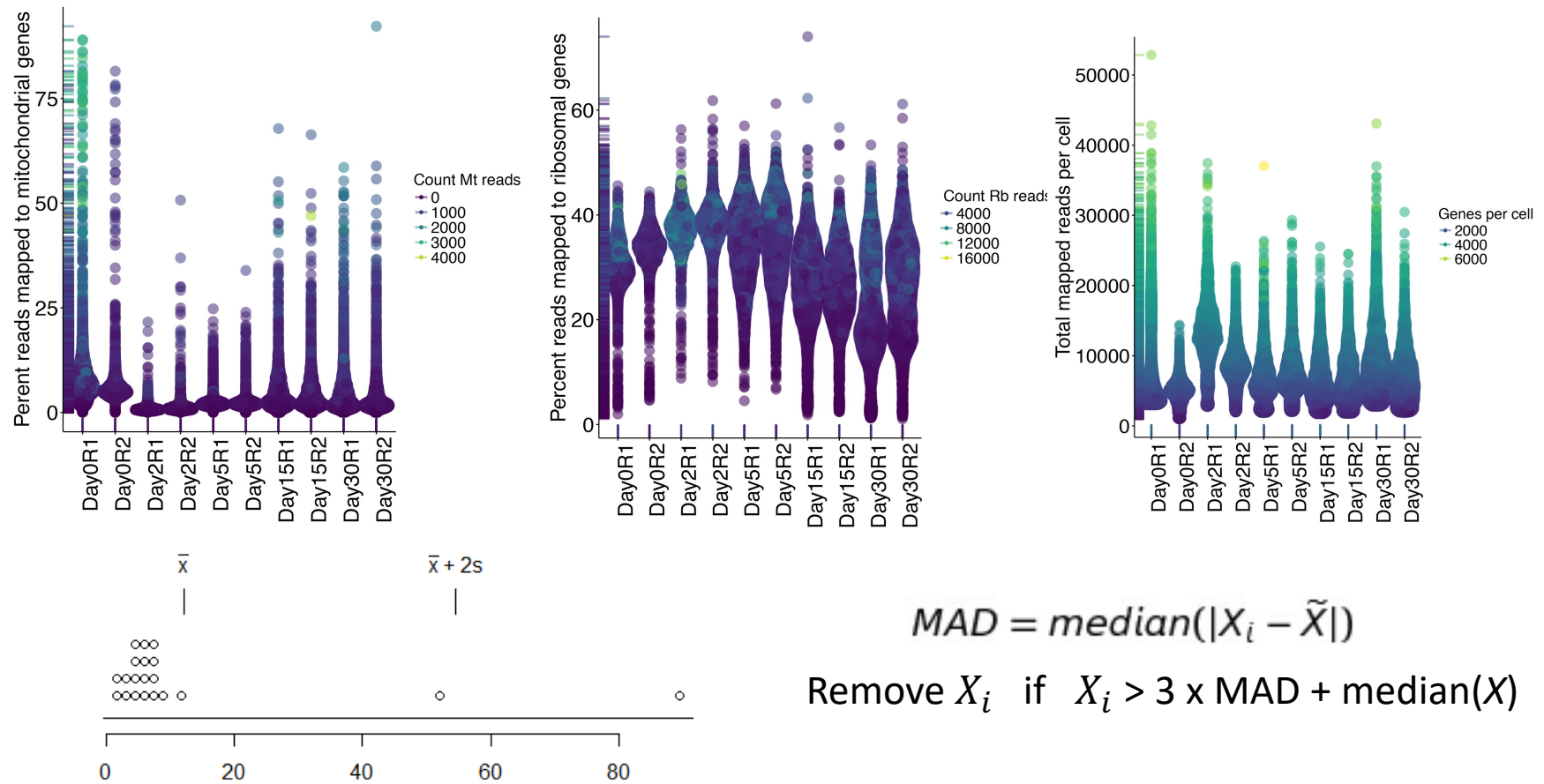


Data quality control: a range of QC measures

- 16 QC measures
- 10 scRNA-seq libraries

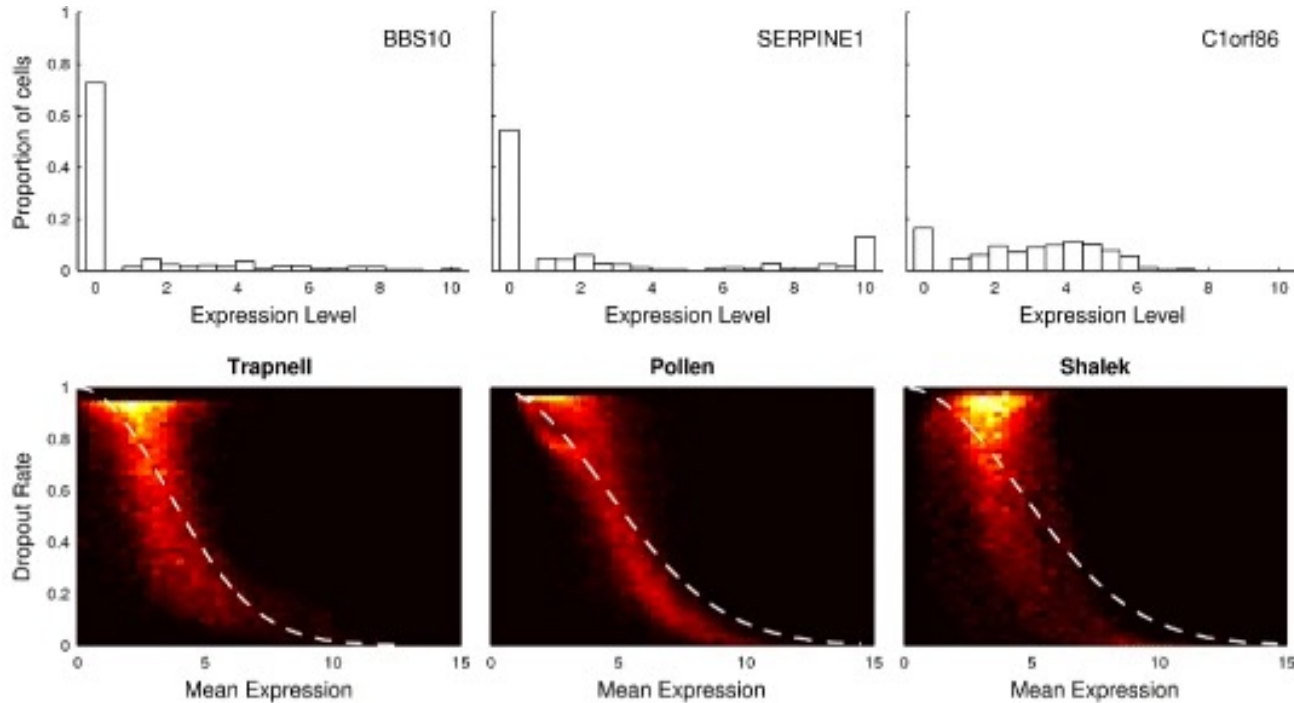


Data preprocessing: quality control and filtering genes and cells



- Median absolute deviation (MAD) is a simple measure of data dispersion that is more robust to cell outliers compared to other measures such as standard deviation
- Using MAD to remove cell outliers: 1) percent mapped reads to mitochondrial/ribosomal genes, 2) number of genes detected per cell, 3) total mapped reads per cell

Single cell data: zero inflation



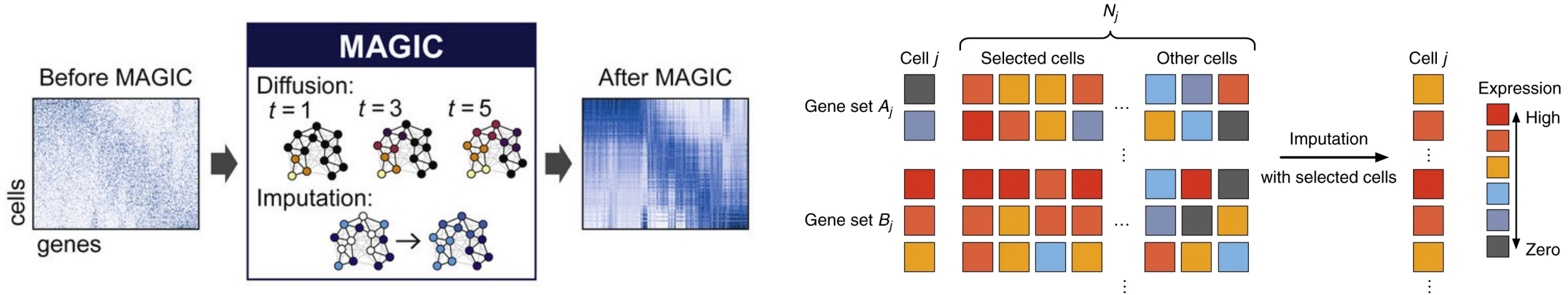
(Pierson and Yau, 2015)

Noise in scRNA-seq data derives from technological limitations:

- Sequencing library amplification bias
- Sequencing depth between cells and samples
- Low RNA capture rate (genes not detected even though they are expressed)
- Variable cell capture rate

$p_0 = \exp(-\lambda \mu^2)$, where λ is a fitted parameter, μ non-zero mean expression, p_0 gene dropout rate

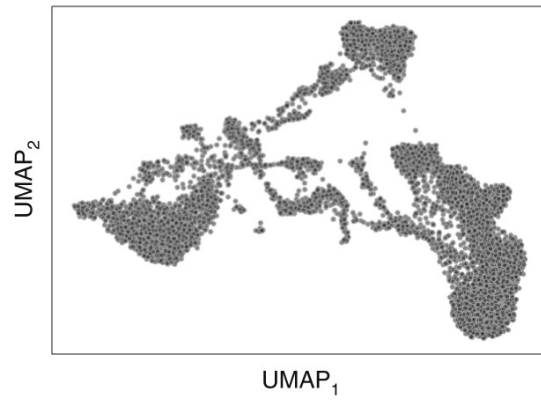
Single cell data: impute zero expression values



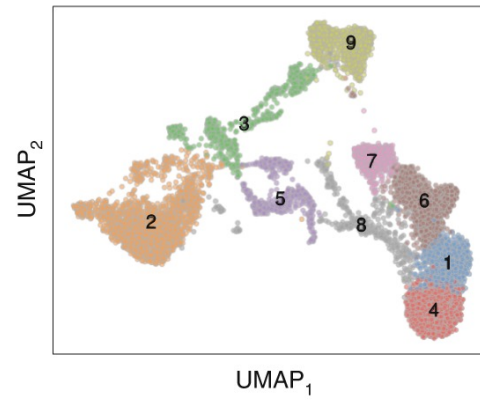
MAGIC: Markov Affinity-based Graph Imputation of Cells weights cells by Markov transition matrix (van Dijk et al., 2018)

scImpute: fits a mixture model to learn gene's dropout probability and borrows information of the same gene in other similar cells based on gene set B_j (Li & Li, 2018)

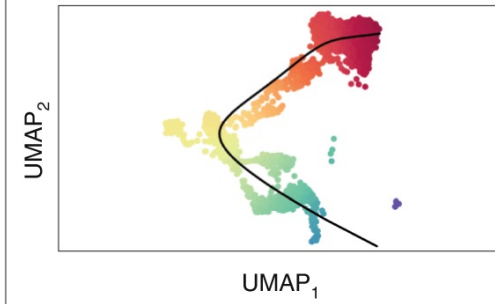
Dimensionality reduction



Clustering

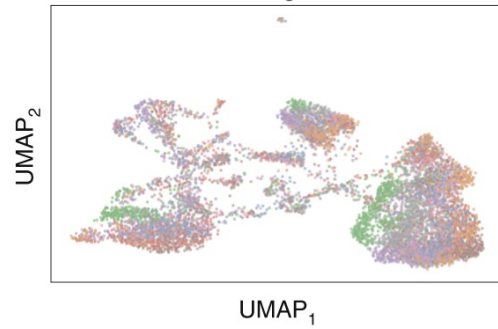


Trajectory analysis



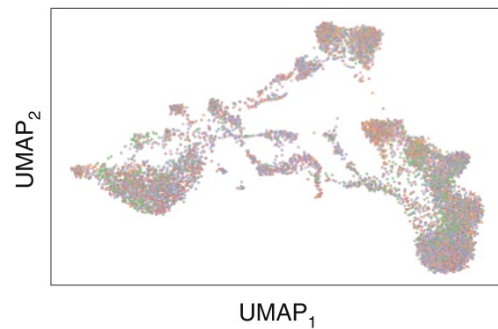
Integrating datasets

Pre-integration



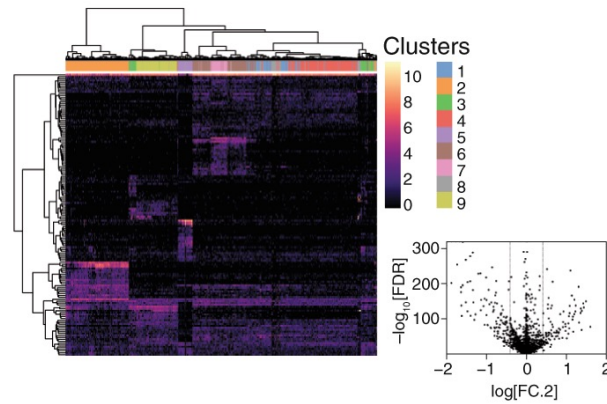
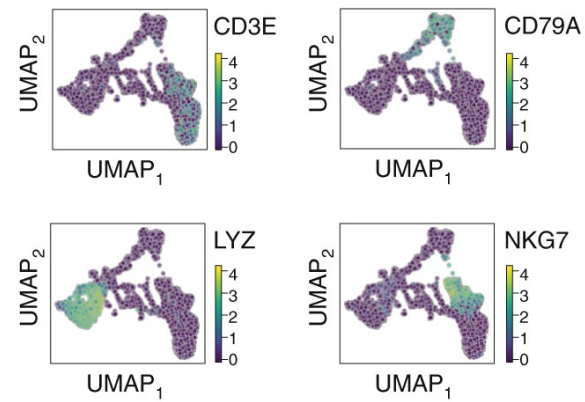
- Donor
- BM1
 - BM2
 - BM3
 - BM4
 - BM5
 - BM6
 - BM7
 - BM8

Post-integration

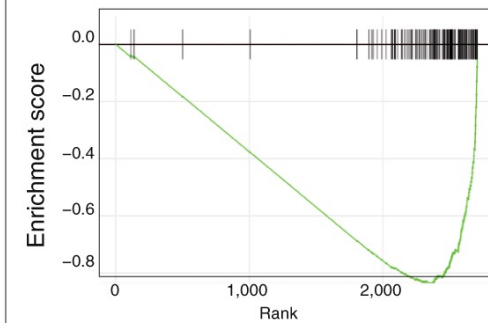


- Donor
- BM1
 - BM2
 - BM3
 - BM4
 - BM5
 - BM6
 - BM7
 - BM8

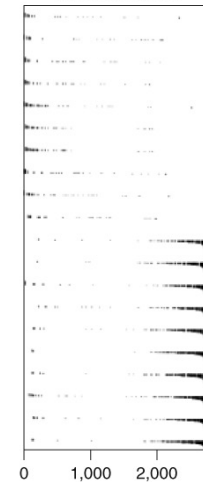
Differential expression



Annotation



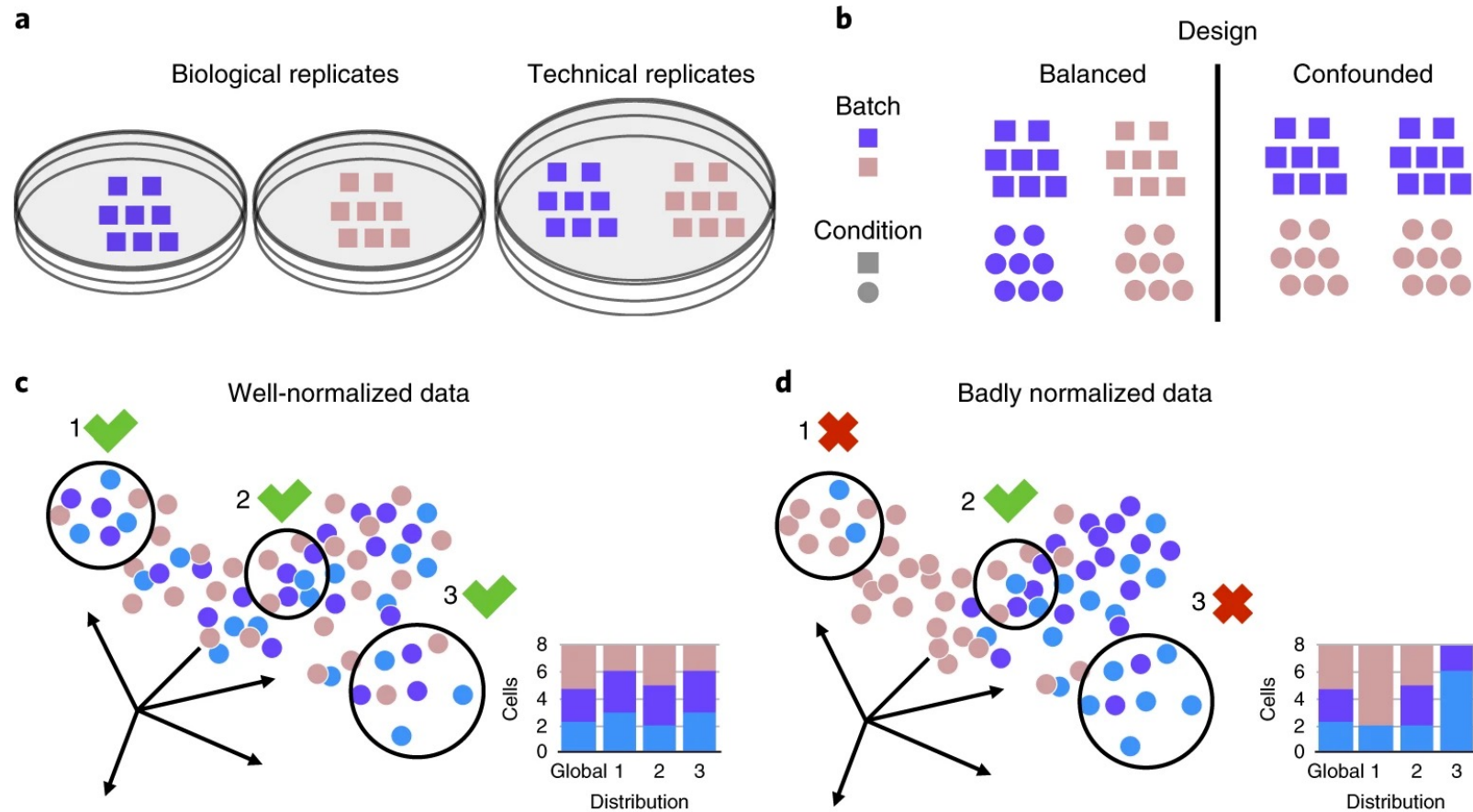
Gene ranks



Data Normalisation

Normalization - Motivation

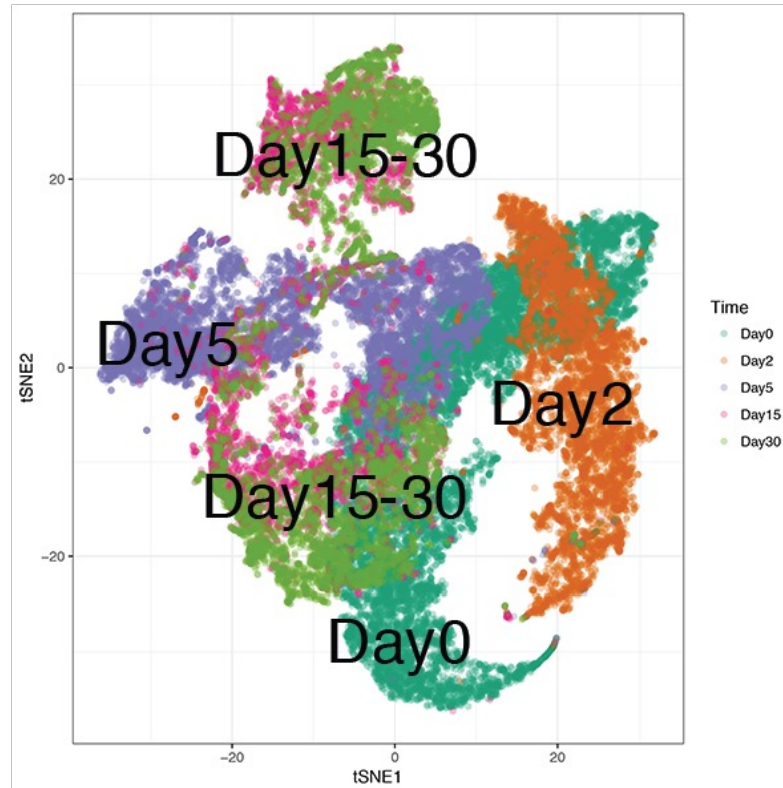
- Batch effects: technical differences induced by the operator or other experimental artifacts
- Often observe systematic differences in sequencing coverage between libraries (or cells)
- Normalization aims to remove these differences
- Such that they do not interfere with comparisons of the expression profiles between cells
- Ensure heterogeneity or differential expression within the cell population are driven by biology and not technical biases.



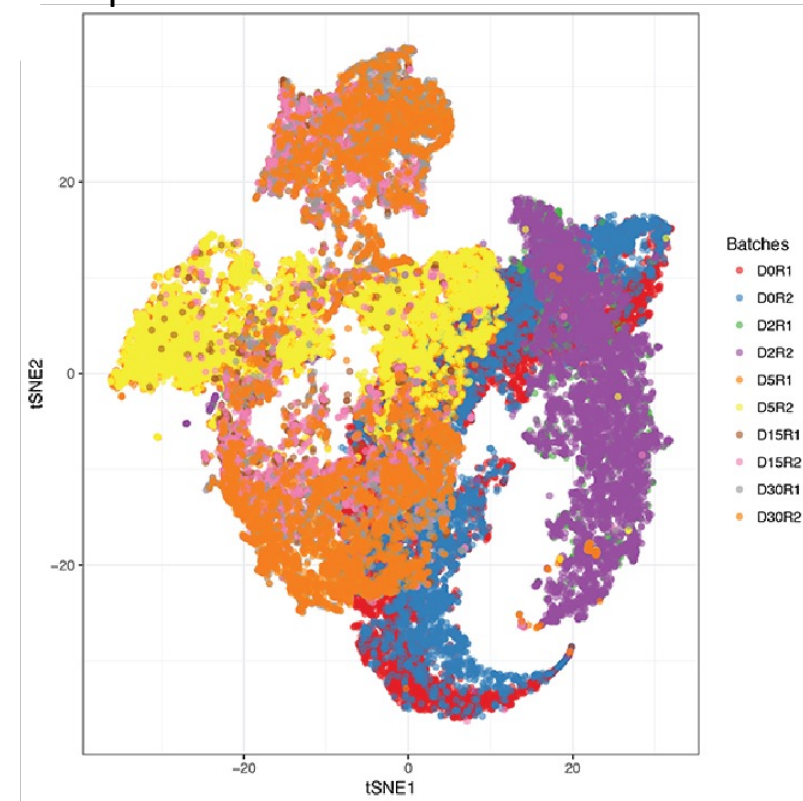
(Buttner et al, 2019)

Representation of biological and technical variation

Representation of biological variation



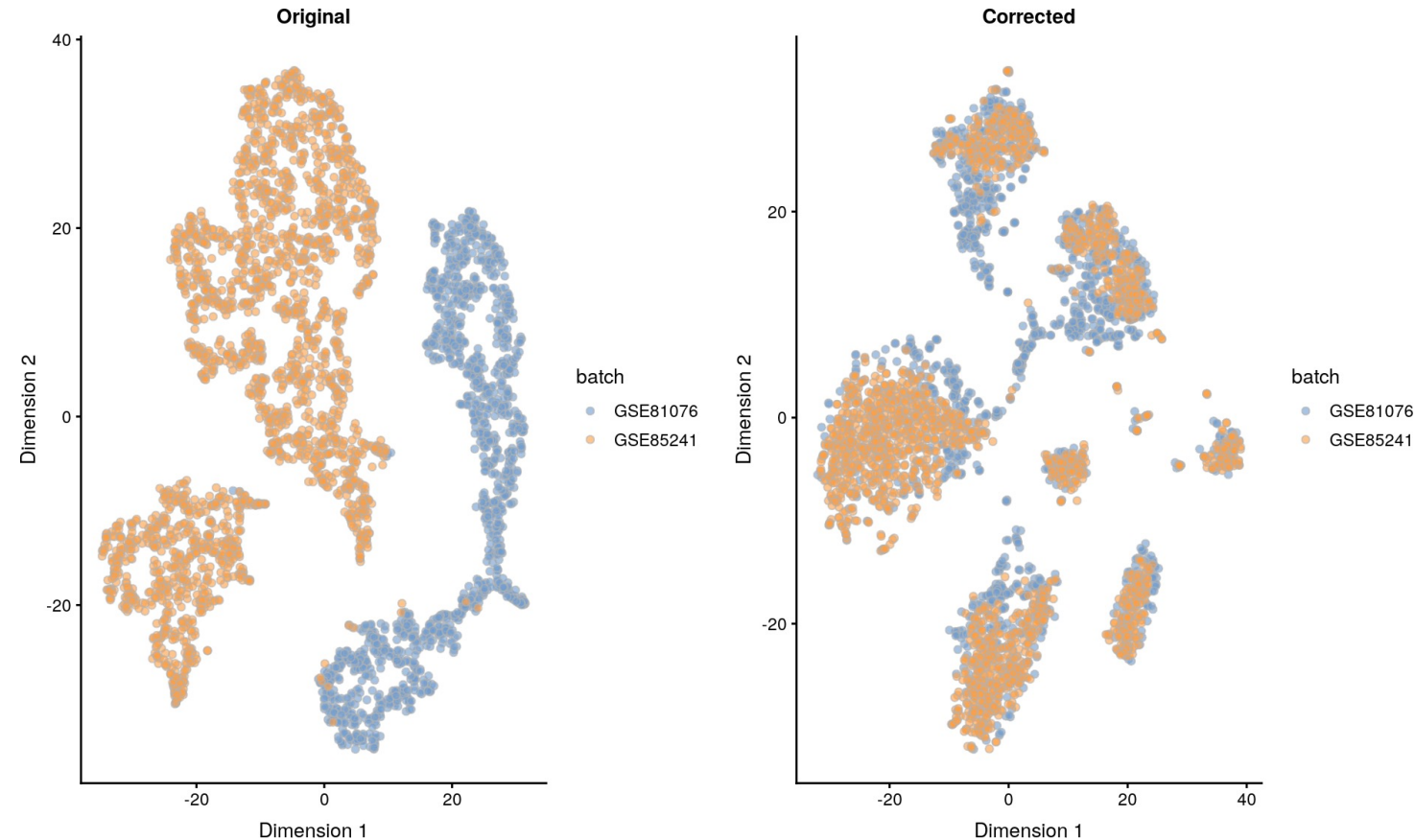
Representation of technical variation



Three levels of single cell data normalization

Three levels of technical variation in scRNA-seq data:

- Gene-specific effects within a cell: GC content, gene length
- Cell specific effects within a sample: each cell is amplified separately, causing amplification bias among cells
- Batch effects within a study: sample preparation or technology-specific effects



Cell to cell normalization: Library size normalization

	Cell1	Cell2	Cell3	Cell4	Cell5
gene1	0	0	0	0	0
gene2	0	0	0	0	0
gene3	3	0	1	0	1
gene4	0	1	3	3	0
gene5	1	4	2	1	2
colsum / library size	4	5	6	4	3
factor	0.91	1.14	1.36	0.91	0.68
Normalized library size	4.40	4.39	4.41	4.40	4.41

Total library size = 22

Nrcells = 5

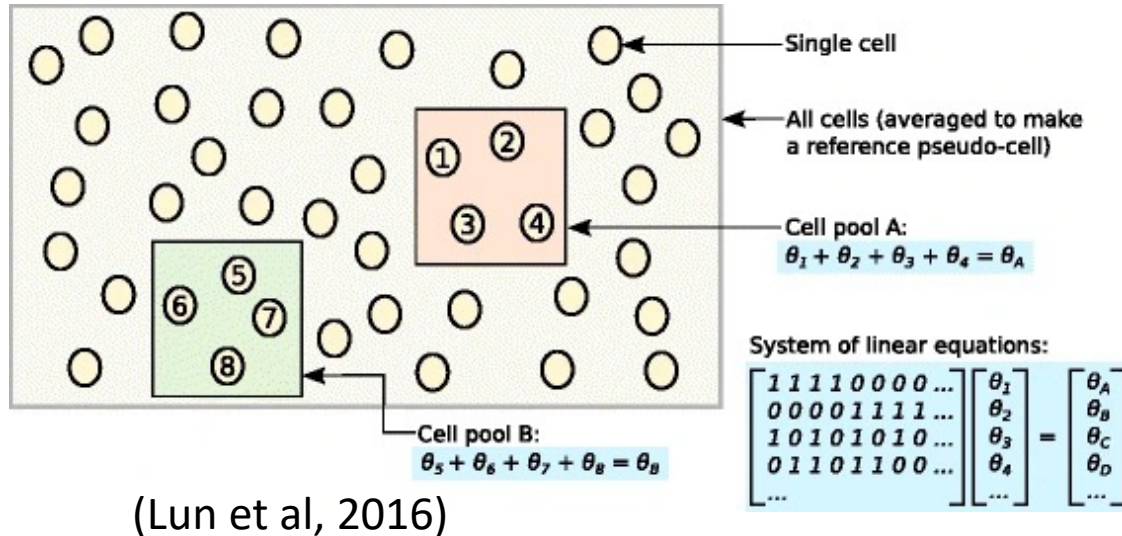
$$\text{Size factor} = \frac{\text{library size} * \text{nrCells}}{\text{Total library size}}$$

The mean size factor across all cells is equal to 1

Normalized expression values are on the same scale as the original counts,

Useful for interpretation especially when dealing with transformed data

Cell to cell normalisation: a pooling strategy to solve zero inflation



	Pool A		Pool B		Sum(poolA)	Sum(PoolB)	average	Sum(poolA)/average	Sum(poolB)/average
	Cell1	Cell 2	Cell 3	Cell 4					
g1	0	0	0	0	0	0	0	0	0
g2	0	0	0	0	0	0	0	0	0
g3	3	0	1	0	3	1	4/4	3/4/4	1/4/4
g4	0	1	3	3	1	6	7/4	1/7/4	6/7/4
g5	1	4	2	1	5	3	8/4	5/8/4	3/8/4

→ A demo

$$\begin{aligned}
 & \uparrow \qquad \qquad \qquad \uparrow \\
 & \theta_A \qquad \qquad \qquad \theta_B \\
 & = \theta_{cell1} + \theta_{cell2} \\
 & \uparrow \\
 & \text{Scaling factor of cell 1}
 \end{aligned}$$

$$E(V_{ik}) = \lambda_{i0} \sum_{j \in S_k} \theta_j \times t_j^{-1}$$

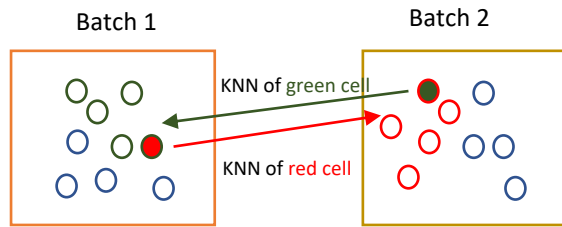
V_{ik} is the sum of adjusted expression value across all cells in pool V_k for gene i
 λ_{i0} is the expected transcript count and θ_j is the cell specific bias
 S_k is a pool of cell; $\theta_j \times t_j^{-1}$ is size factor for cell j

- Each cell is considered as a sequencing library, so the total reads per cell need to be normalised
- Pool cells to reduce the number of zeros
- Estimate the size factors for the pool
- Repeat many time and use deconvolution to estimate each cell size factor θ_j

Batch normalisation: Mutual nearest neighbour (MNN)

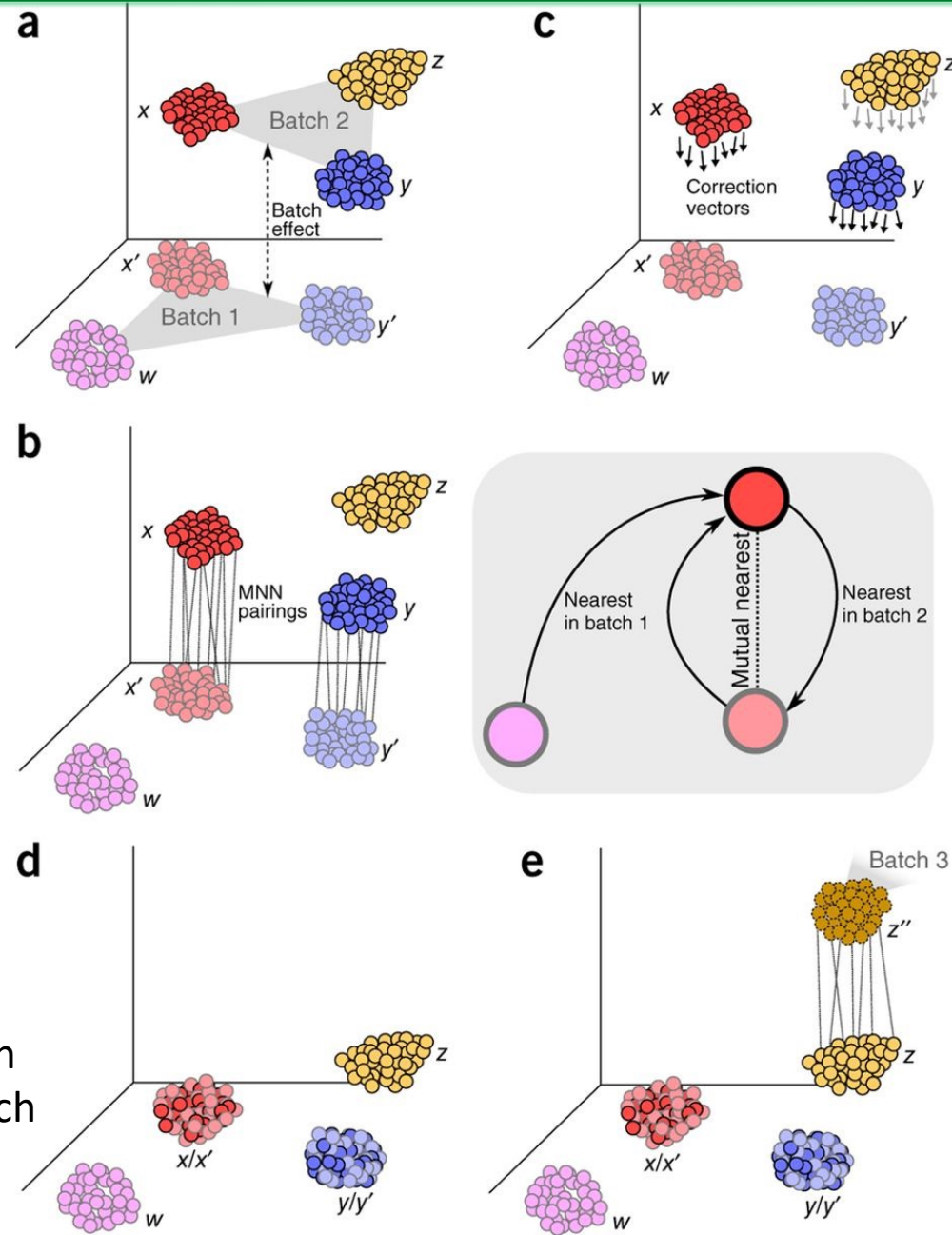
Three assumptions in MNN normalisation:

- (i) there is at least one cell population that is present in both batches,
- (ii) the batch effect is almost orthogonal to the biological subspace, and
- (iii) the batch-effect variation is much smaller than the biological-effect variation between different cell types



Red and green are MNN

Find KNN in another Batch



assume batch effects are mostly orthogonal to the biological manifold:

- ← batch effect: vertical
- ← biological manifold: horizontal

the cosine normalization

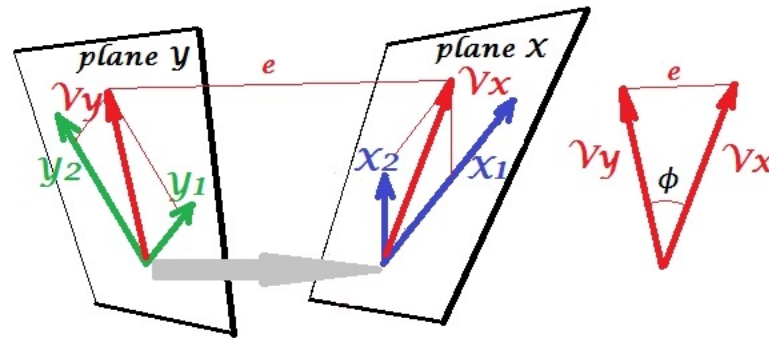
$$Y_x \leftarrow \frac{Y_x}{|Y_x|}$$

Batch normalisation: Canonical correlation analysis (CCA)

- CCA finds projection vectors u and v such that the correlation between the two datasets $u^T X$ and $v^T Y$ is maximized
- CCA vectors capture sources of variance that are shared between data sets.
- CC vectors are correlated, but not necessarily aligned between data sets
- Alignment finds cell in the other dataset with the most similar metagene expression while maintaining the relative ordering of cells within each data set

How to understand CCA

- Simple linear regression: $Y = bX + e$
- Multiple linear regression: $Y = b_1 X_1 + b_2 X_2 + \dots + e$
multiple Xs
objective function: $\sum (Y - \hat{Y})^2$
- CCA: $b'_1 Y_1 + b'_1 Y_1 + \dots + e' = b_1 X_1 + b_2 X_2 + \dots + e$
multiple Xs and multiple Ys
objective function: maximize the correlations between canonical variate pair



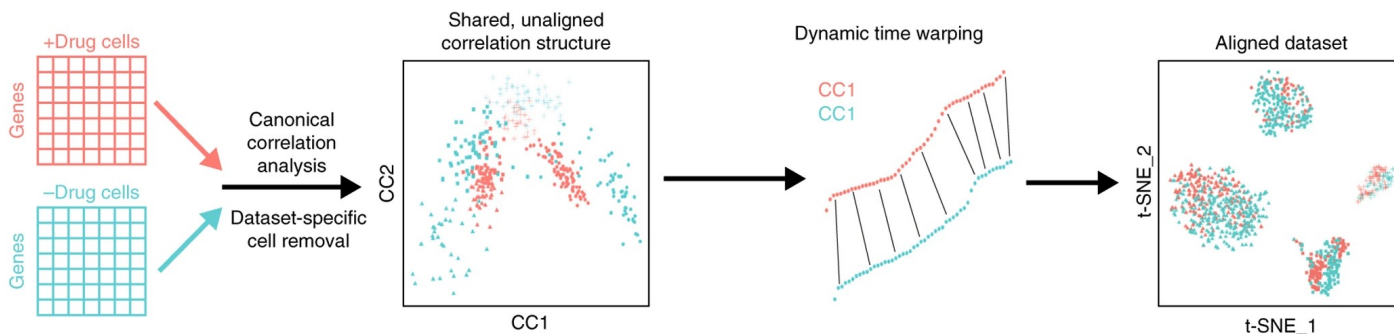
$$\begin{aligned}
 U_1 &= a_{11}X_1 + a_{12}X_2 + \dots + a_{1p}X_p \\
 U_2 &= a_{21}X_1 + a_{22}X_2 + \dots + a_{2p}X_p \\
 &\vdots \\
 U_p &= a_{p1}X_1 + a_{p2}X_2 + \dots + a_{pp}X_p \\
 \\
 V_1 &= b_{11}Y_1 + b_{12}Y_2 + \dots + b_{1q}Y_q \\
 V_2 &= b_{21}Y_1 + b_{22}Y_2 + \dots + b_{2q}Y_q \\
 &\vdots \\
 V_p &= b_{p1}Y_1 + b_{p2}Y_2 + \dots + b_{pq}Y_q
 \end{aligned}$$

Thus define

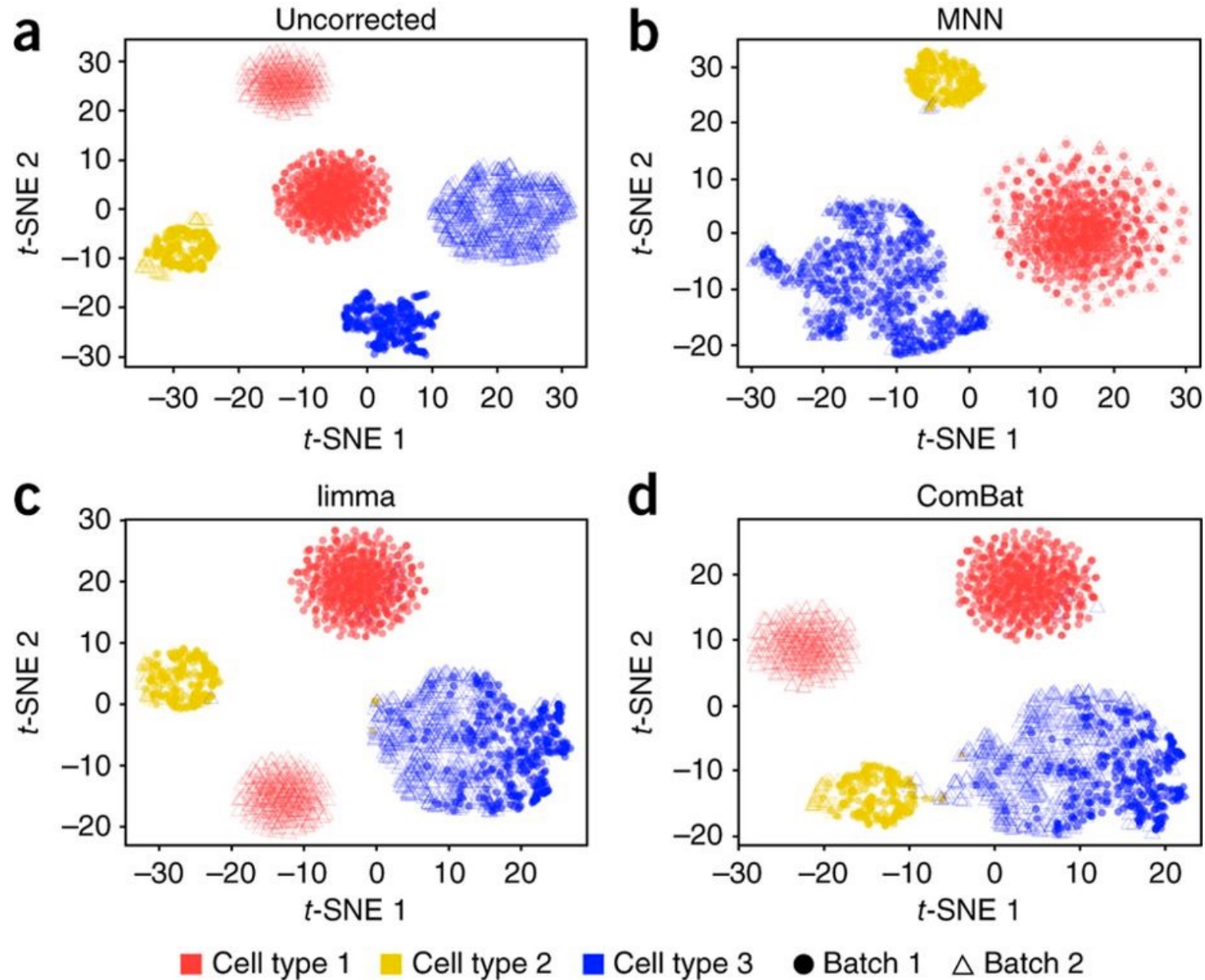
$$(U_i, V_i)$$

as the i^{th} canonical variate pair. (U_1, V_1) is the first canonical variate pair, similarly (U_2, V_2) would be the second canonical variate pair and so on. With $p \leq q$ there are p canonical covariate pairs.

We hope to find linear combinations that maximize the correlations between the members of each canonical variate pair.



Batch normalisation: Mutual Nearest Neighbour (MNN)



Log-transforms the normalized values

- The log-transformation is useful:
 - 1) differences in the log-values represent log-fold changes in expression.
 - 2) Gene X is more interesting:

raw count	cell A	Cell B
Gene X	50	10
Gene Y	1100	1000

Logcount(2)	cell A	Cell B
Gene X	5.64	3.32
Gene Y	10.1	9.97

Dimensionality Reduction

Dimensionality reduction: linear techniques

Why dimensionality reduction:

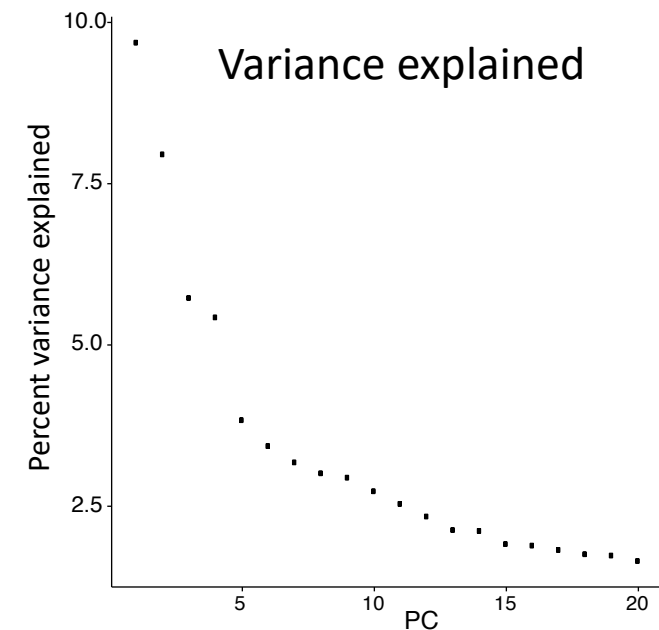
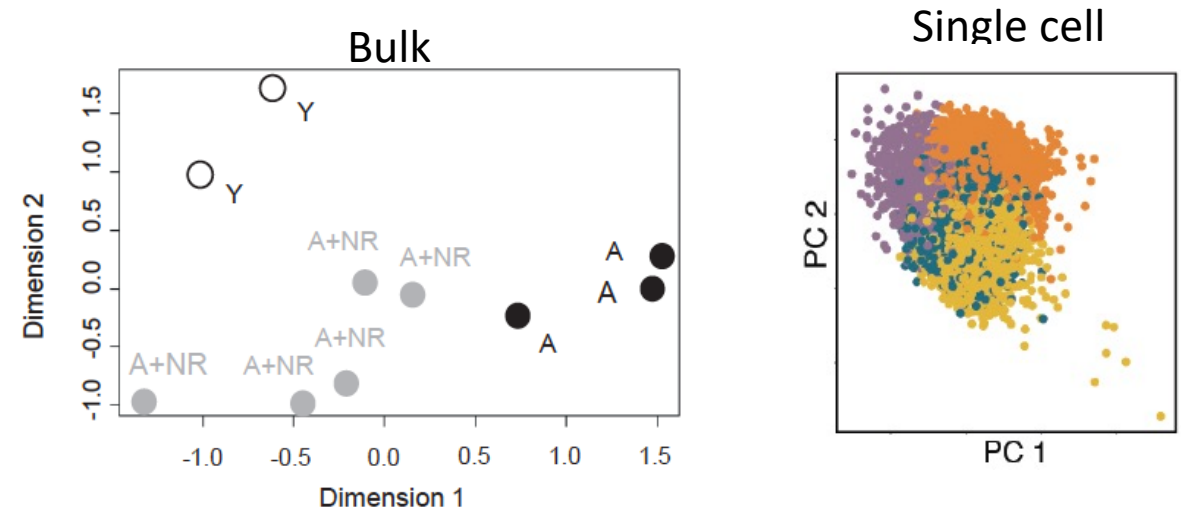
- Filters out noise
- Minimises curse of dimensionality
- Allows visualization with more separation of points
- Reduces computational load

Linear approaches:

- PCA (Principal Component Analysis)
- ICA (Independent Component Analysis)
- NMF (Non-negative Matrix Factorization)

Linear approaches:

- Capture the dimensions with higher variance
- Quantitative way to assess the amount of retained dimensions
- Preserve both long-range and short-range distance (i.e. cells that are very different or very similar)
- Different to bulk RNAseq data, the first few dimensions are not enough to capture scRNAseq data structure well

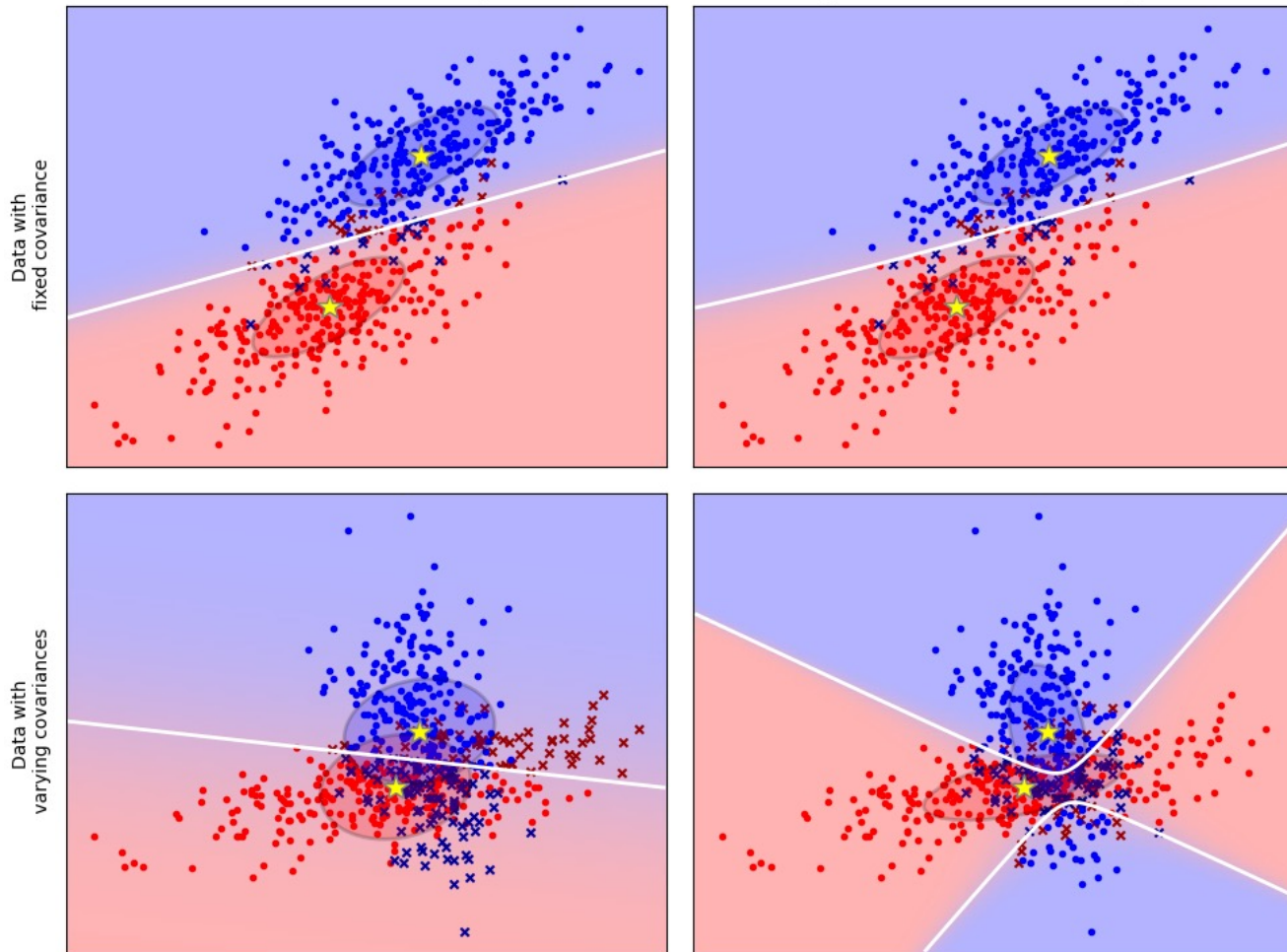


Supervised Dimensionality reduction techniques

Linear Discriminant Analysis vs Quadratic Discriminant Analysis

Linear Discriminant Analysis

Quadratic Discriminant Analysis



https://scikit-learn.org/stable/modules/lda_qda.html

- Probabilistic models which model the class conditional distribution of the data $P(\mathbf{X}|\mathbf{y}=\mathbf{k})$ for each class \mathbf{k} .

$$P(y = k|x) = \frac{P(x|y = k)P(y = k)}{P(x)} = \frac{P(x|y = k)P(y = k)}{\sum_l P(x|y = l) \cdot P(y = l)} \quad (1)$$

- For LDA and QDA, $P(\mathbf{x}|\mathbf{y})$ is modeled as a multivariate Gaussian distribution with density:

$$P(x|y = k) = \frac{1}{(2\pi)^{d/2} |\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_k)^t \Sigma_k^{-1} (x - \mu_k)\right) \quad (2)$$

LDA

$$\log P(y = k|x) = -\frac{1}{2}(x - \mu_k)^t \Sigma^{-1} (x - \mu_k) + \log P(y = k) + Cst$$

QDA

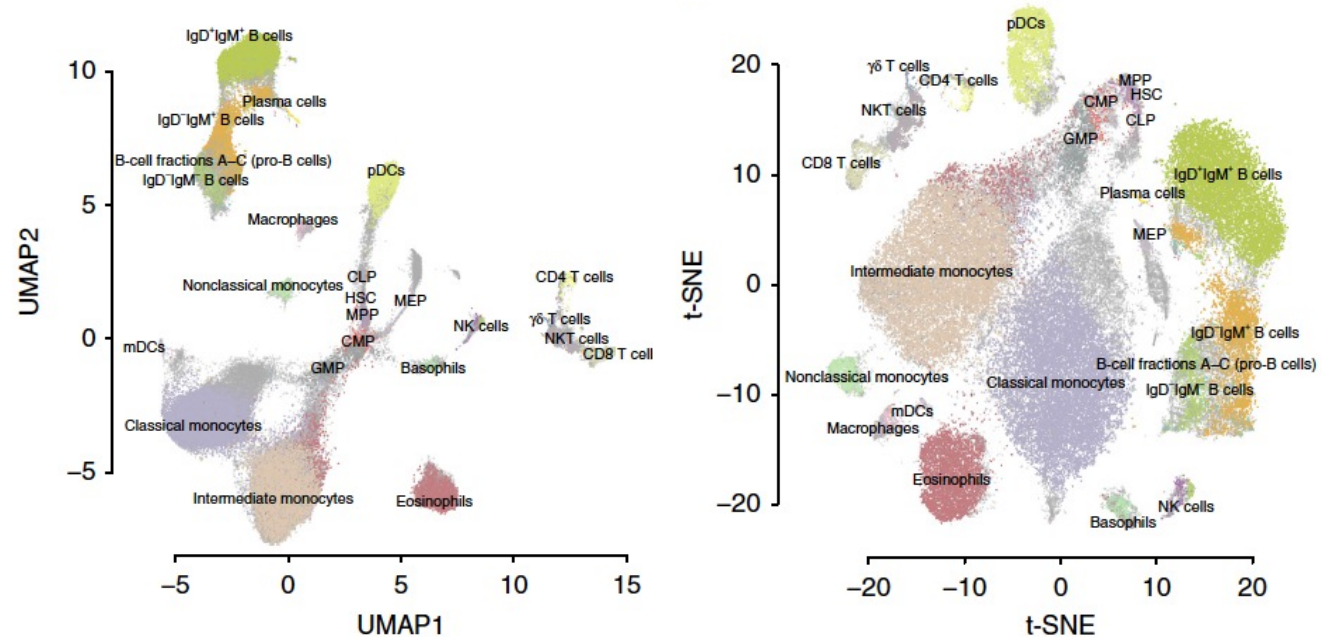
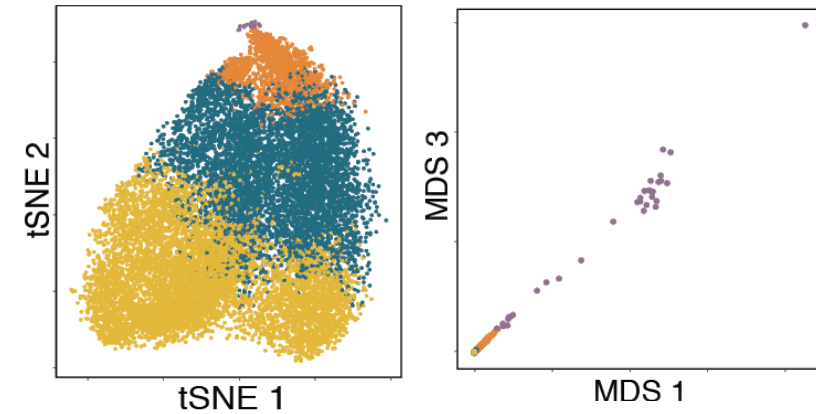
$$\begin{aligned} \log P(y = k|x) &= \log P(x|y = k) + \log P(y = k) + Cst \\ &= -\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x - \mu_k)^t \Sigma_k^{-1} (x - \mu_k) + \log P(y = k) + Cst \end{aligned}$$

- Perform supervised dimensionality reduction
- Projecting the input data to a linear subspace consisting of the directions which maximize the separation between classes

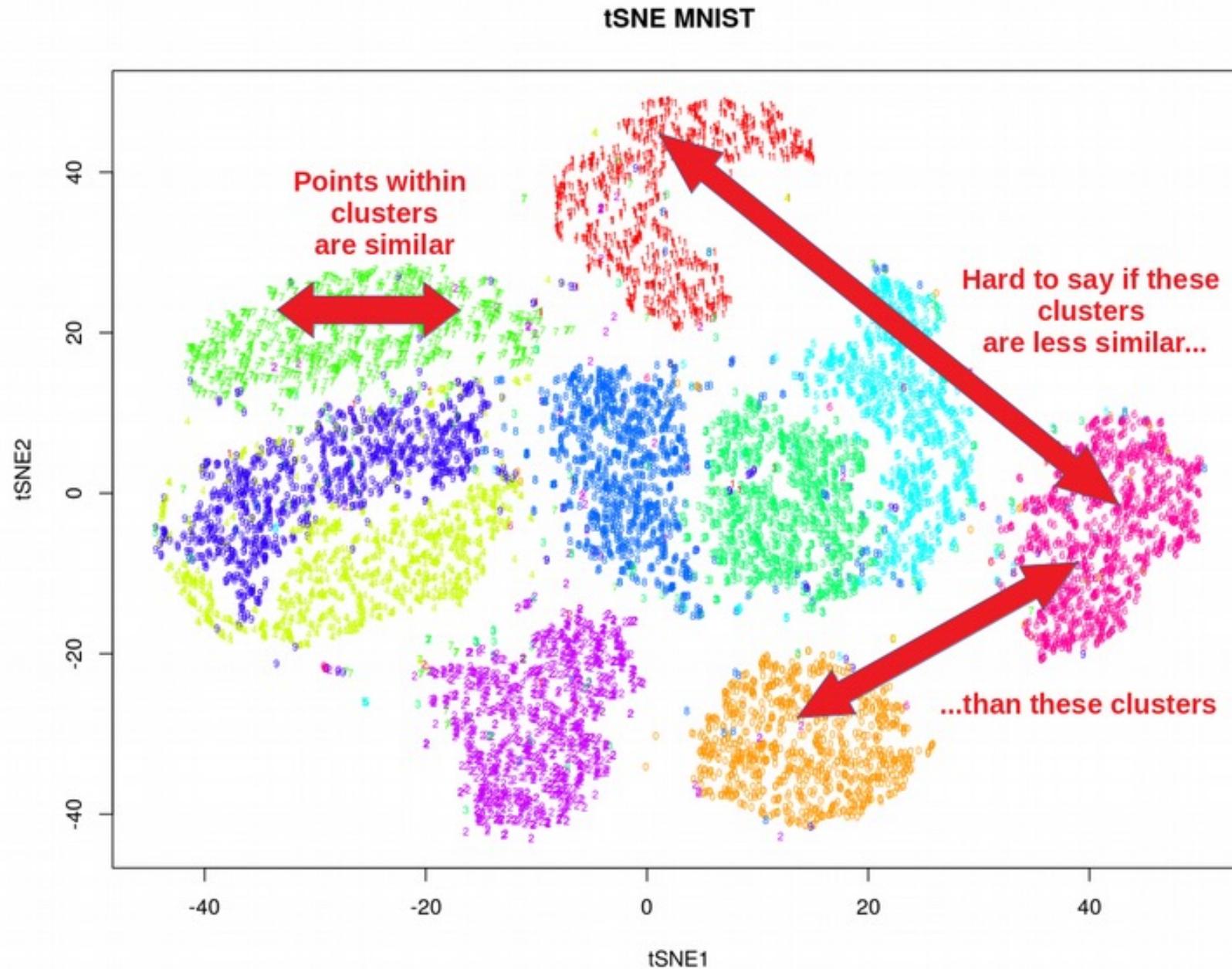
Dimensionality reduction: nonlinear techniques

- MDS (Multidimensional Scaling)
- Uniform manifold approximation and projection (UMAP)
- t-distributed Stochastic Neighbour Embedding (t-SNE)
- UMAP and tSNE: nonlinear embedding (mapping) of data points from high dimensional space to low dimensional space, so that the probability distance between these two space (KL divergence or cross entropy) is minimised
- Both methods: class of k-neighbour based graph learning algorithms, strong influence of hyperparameters, non-deterministic (stochastic)
- Nonlinear techniques solve the overcrowding representation, which is often seen in linear approaches for large scRNA-seq data
- UMAP preserves local & more of the global data structure than t-SNE

Overcrowding



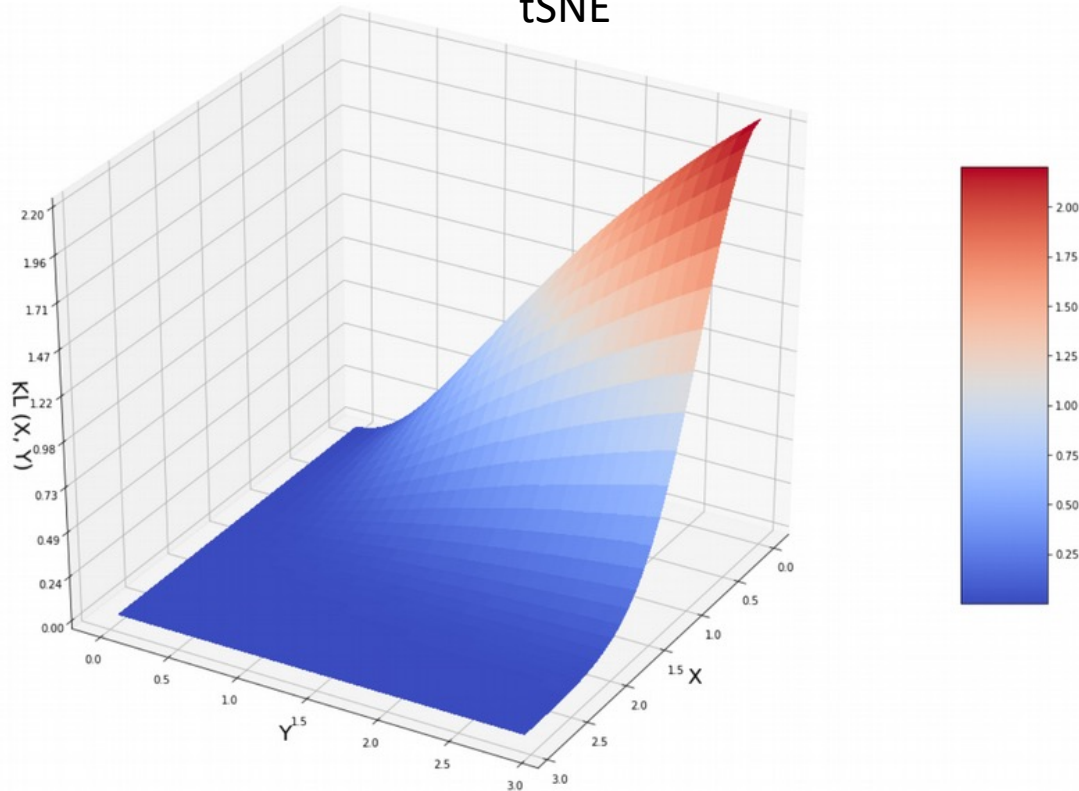
Global vs local distance in low dimensional space



tSNE does not preserve long distance - KL divergence

(Oskolkov N, 2019)

tSNE



- The embedding minimizes the Kullback-Leiber divergence of the distribution from Q to P calculated as: $KL(X, Y) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \approx e^{-X^2} \log(1 + Y^2)$

- The probability distance between two neighbouring cells is the joint probabilities $p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$

- Conditional probability of cell C_j given cell C_i is calculated as:

$$p_{j|i} = \frac{\exp\left(\frac{-d(C_i, C_j)^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} \exp\left(\frac{-d(C_i, C_k)^2}{2\sigma_i^2}\right)}$$

- For large distances X in high dimensions, the exponential term approaching 0, **so Y can be basically any value from 0 to ∞ and KL remains small**

- For small X, to minimise KL (cost/penalty), Y is small

- Pairwise similarity in t-SNE space: $q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq m} (1 + \|y_k - y_m\|^2)^{-1}}$, y_i and y_j are corresponding mapped points of cells C_i and C_j to t-SNE space, and **q_{ij} follows t distribution to avoid crowding**

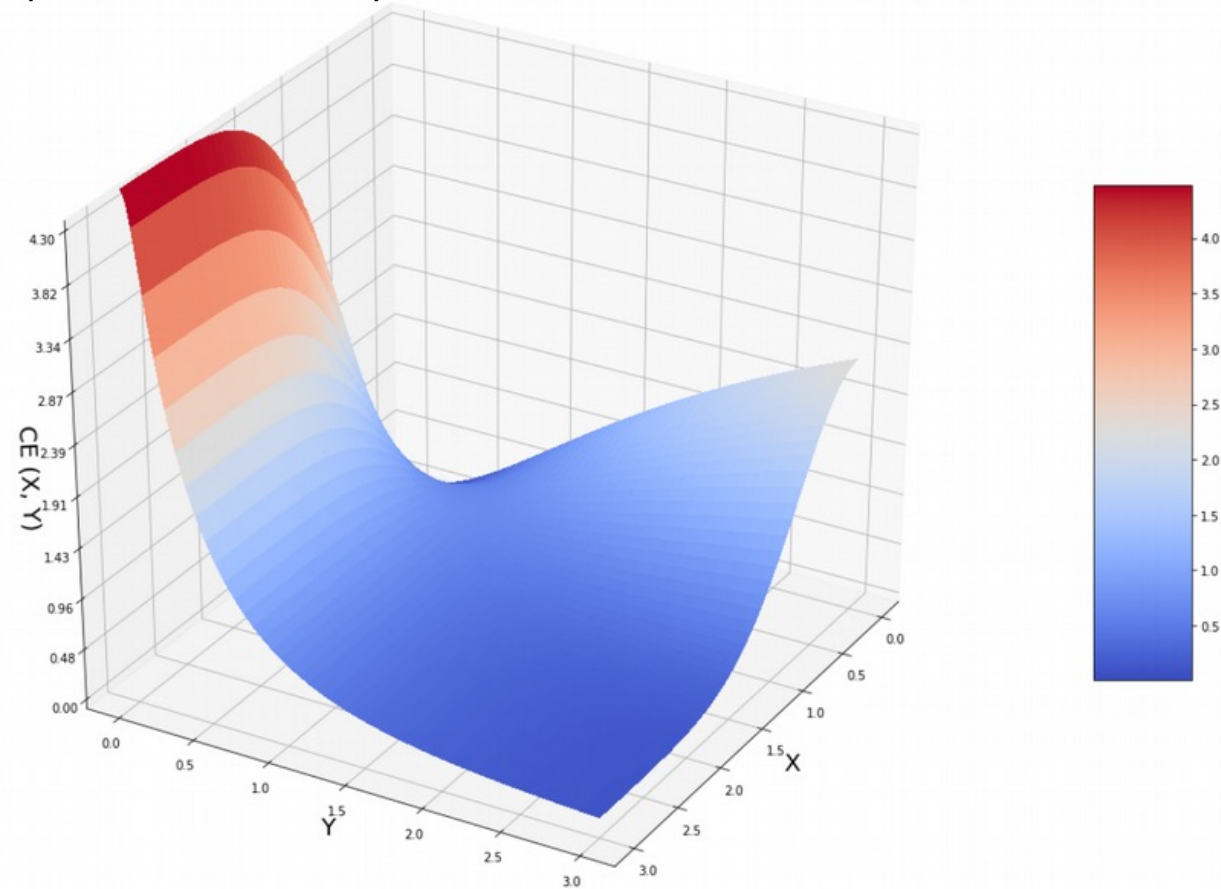
tSNE minimises Kullback-Leiber divergence $KL(X, Y)$

$$KL(X, Y) \approx -P(X) \log Q(Y) = e^{-X^2} \log(1 + Y^2)$$

UMAP preserves long distance - cross entropy

(Oskolkov N, 2019)

UMAP



$$X \rightarrow 0 : CE(X, Y) \approx \log(1 + Y^2)$$

When X small, Y is also approaching 0 to minimize CE

$$X \rightarrow \infty : CE(X, Y) \approx \log\left(\frac{1 + Y^2}{Y^2}\right)$$

When X large, Y is also large to minimize CE

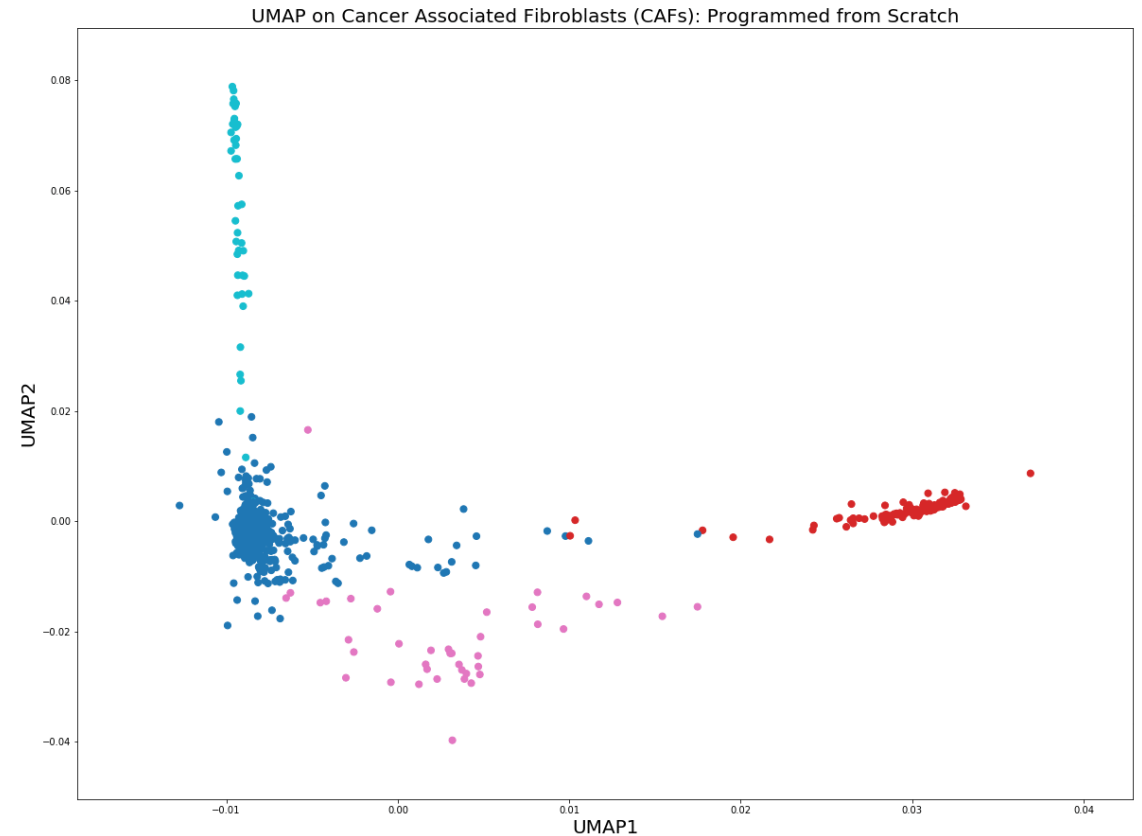
UMAP minimises cross entropy $CE(X, Y)$

$$CE(X, Y) = P(X) \log\left(\frac{P(X)}{Q(Y)}\right) + (1 - P(X)) \log\left(\frac{1 - P(X)}{1 - Q(Y)}\right)$$
$$\approx e^{-X^2} \log(1 + Y^2) + (1 - e^{-X^2}) \log\left(\frac{1 + Y^2}{Y^2}\right)$$

$$\text{tSNE: } KL(X, Y) \approx -P(X) \log Q(Y) = e^{-X^2} \log(1 + Y^2)$$

More about UMAP vs tSNE

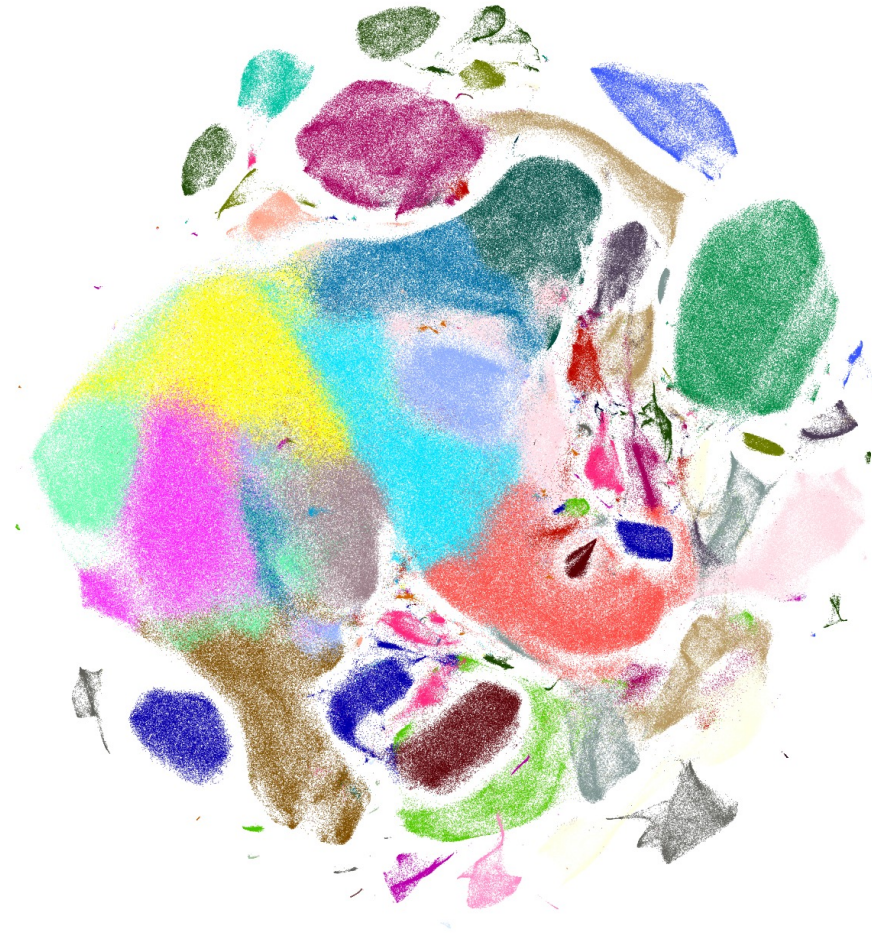
- To learn low-dimensional embeddings, UMAP assigns initial low-dimensional coordinates using **Graph Laplacian** (force directed graph layout algorithm) in contrast to **random normal initialization** used by tSNE. Therefore, UMAP is less dependent on random state (not changing from run to run)
- UMAP proceeds by iteratively applying attractive (among edges) and repulsive forces (among vertices) at each edge or vertex. Convergence is guaranteed by slowly decreasing the attractive and repulsive forces of the neighbour graph.
- UMAP has no computational restrictions on embedding dimension, making it viable as a general-purpose dimension reduction technique for machine learning (tSNE can only embed to 2-3 dimensions)



(Oskolkov N, 2019)

scRNAseq Data Clustering

Single Cell Clustering Analysis



Clustering in scRNAseq is a data-driven way to find cell (sub)types at single-cell resolution

Clustering to assess subpopulation heterogeneity

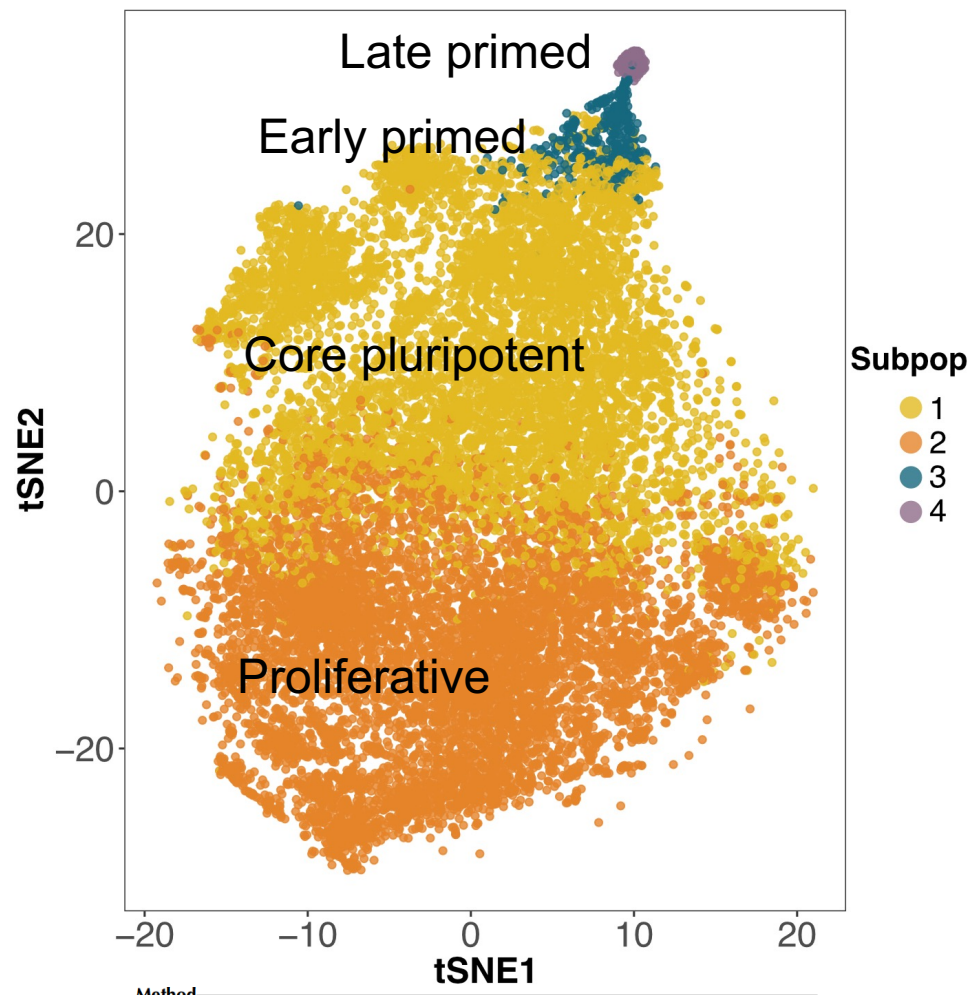
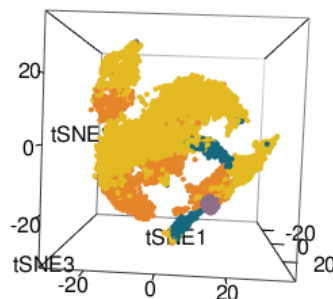
An example iPSC scRNA dataset:

- Sequenced > 18,000 cells (10x Genomics)
- Detected > 16,000 genes
- We proved that a seemingly homogeneous hiPSC population contains 4 subpopulations

Why study heterogeneity in development and diseases?

- More heterogeneous than expected
- Specific biological processes masked by mixed population-averaging effect
- Early disease diagnosis, specific markers
- Targeted drug discovery, treatment, and monitoring
- Personalised medicine

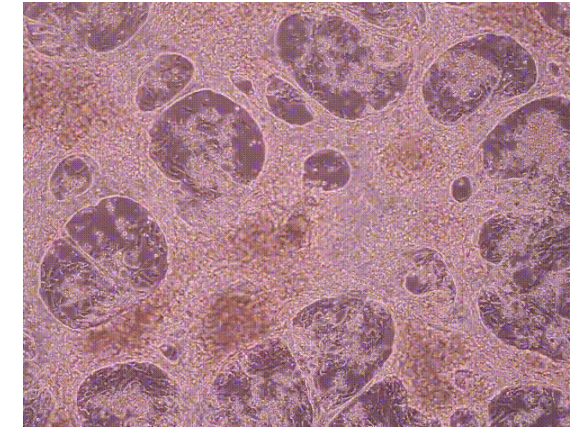
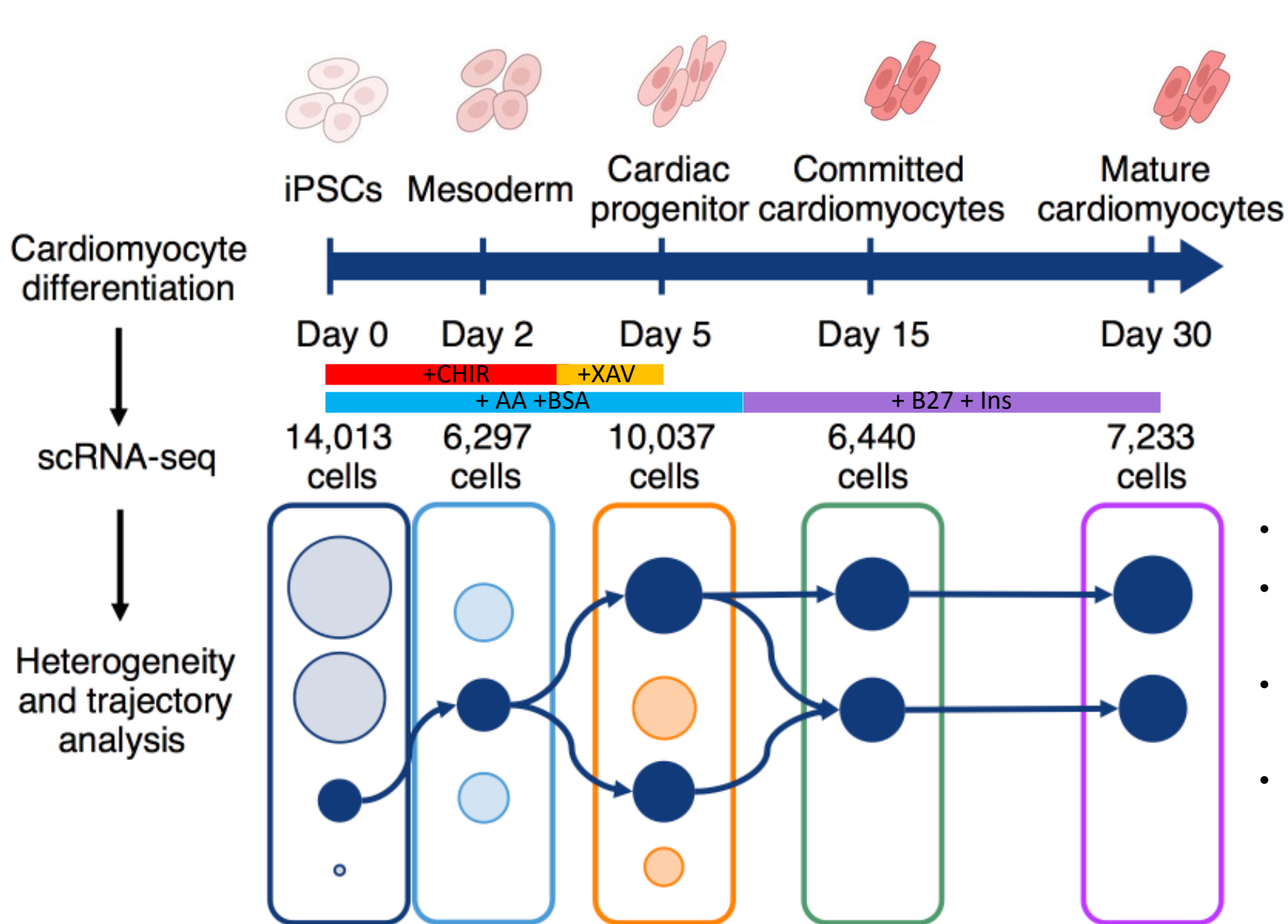
.....



Method _____
Single-cell RNA-seq of human induced pluripotent stem cells reveals cellular heterogeneity and cell state transitions between subpopulations

Quan H. Nguyen,^{1,4} Samuel W. Lukowski,^{1,4} Han Sheng Chiu,¹ Anne Senabouth,¹ Timothy J.C. Bruxner,¹ Angelika N. Christ,¹ Nathan J. Palpant,^{1,4} and Joseph E. Powell^{1,2,3,4}

Clustering to assess cell-type specific responses



(Fei Pei et al., 2017)

Question: differential responses at the subpopulation levels?

- 5 time points: days 0, 2, 5, 15 and 30
- Sequenced > 43,000 single-cell transcriptomes (10x Genomics)
- Detected > 17,000 genes at each time point
- Aim: to identify gene regulation changes at single-cell and subpopulation levels within and between time points

Cluster cells in expression space - Distance measures

1-Pearson's correlation coefficient (x_{ig} is the expression)	$d_{ij} = 1 - \frac{\sum_{g=1}^G (x_{ig} - \bar{x}_i) (x_{jg} - \bar{x}_j)}{\sqrt{\sum_{g=1}^G (x_{ig} - \bar{x}_i)^2} \sqrt{\sum_{g=1}^G (x_{jg} - \bar{x}_j)^2}}$
1-Spearman's correlation coefficient (r_{ig} expression rank)	$d_{ij} = 1 - \frac{\sum_{g=1}^G (r_{ig} - \bar{r}_i) (r_{jg} - \bar{r}_j)}{\sqrt{\sum_{g=1}^G (r_{ig} - \bar{r}_i)^2} \sqrt{\sum_{g=1}^G (r_{jg} - \bar{r}_j)^2}}$
Cosine distance	$d_{ij} = 1 - \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\ \mathbf{x}_i\ \cdot \ \mathbf{x}_j\ }$

Correlation-based and cosine distance metrics are **scale invariant**: they consider relative differences in values, making them more robust to library or cell size differences.

Classical clustering techniques

- Two examples of simple cases for K-mean and Hierarchical clustering techniques
- K-mean clustering:
 - Initialisation: given an initial set of K random centres and a distance matrix, finds the closest cluster centres for each of all cells, then updates the centres (average of all cells in a cluster).
 - Repeat the EM procedure till no more change in the centroids
 - K-mean requires a prior decision on the number of cell types
- Hierarchical clustering (Agglomerative/bottom-up approach):
 - Initialisation: HC begins with n clusters of size one
 - Merging (Ward's variance): the two clusters with the minimal increase in the distance $d_{AB} = SSE_{AB} - (SSE_A + SSE_B)$ are merged. The next decision to merge a subsequent cluster (C) to a {A, B} branch requires C to satisfy that the distance between C and {A, B} is minimised

$$SSE_A = \sum_{i=1}^{n_A} (a_i - \bar{a})'(a_i - \bar{a}), \text{ where } \bar{a} \text{ is the centroid cell of the cluster A}$$

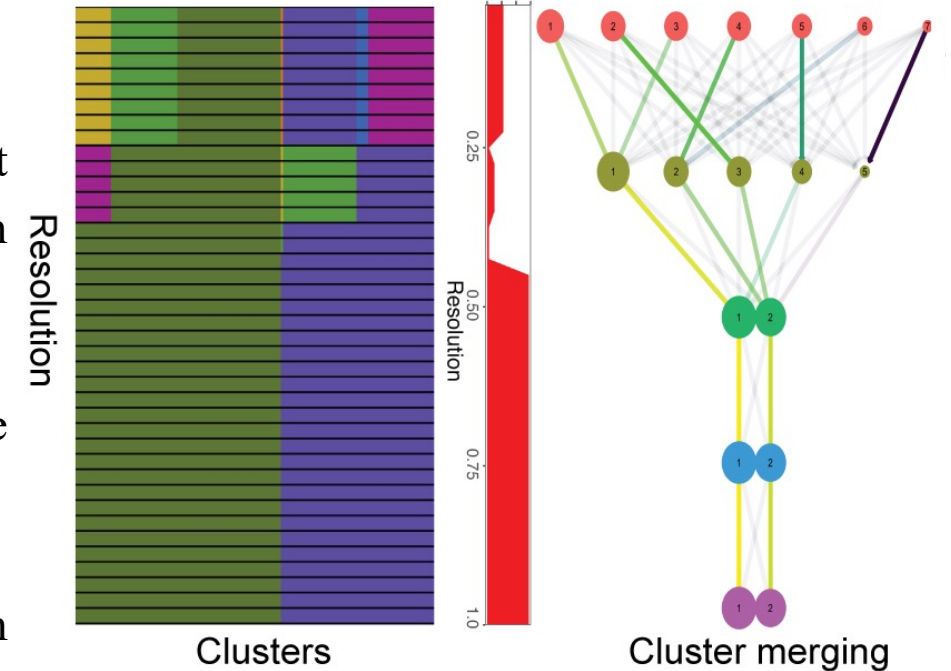
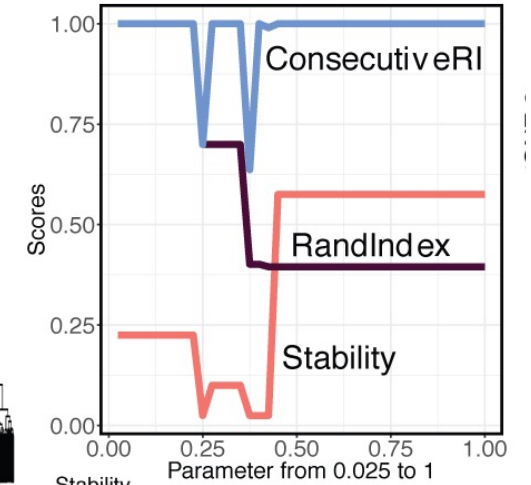
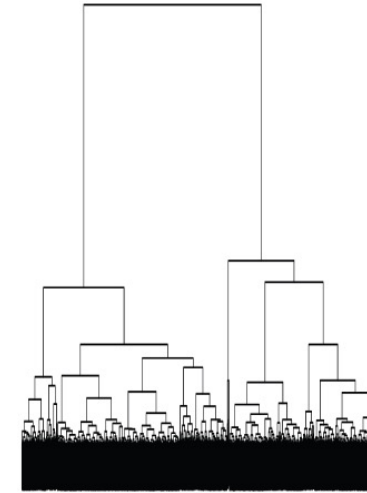
$$d_{C(AB)} = \frac{(n_A + n_C)}{(n_A + n_B + n_C)} d_{CA} + \frac{(n_B + n_C)}{(n_A + n_B + n_C)} d_{CB} - \frac{(n_C)}{(n_A + n_B + n_C)} d_{AB}$$

- A dendrogram tree is formed after the merging

SCORE (Stable Clustering at Optimal Resolution):

We improved HC clustering by first selecting for an optimal cluster resolution by implementing the following algorithm:

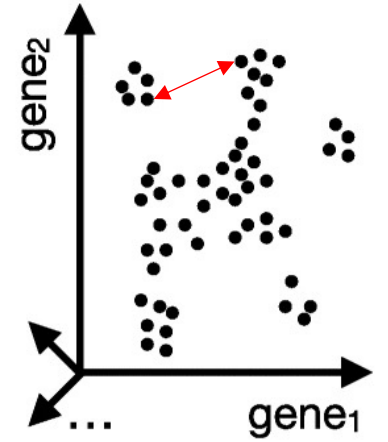
1. Apply cutreeDynamic 40 times to merge branches in 40 different height windows (defined the dendrogram area to be merged) from bottom ($W_1 = [0.025, 1]$) to the top ($W_1 = [1, 1]$).
2. Compute pairwise adjusted Rand index (AR_i) for every 2 consecutive windows (W_i and W_{i+1} for integers $i \in [1, 39]$)
3. Compute stability index S spanning the 40 iterations. S is the set of count values C_s for unique Rand index values AR_i that remain the same between consecutive W_i .
4. Determine the most stable clustering result C_s , where s is selected by the following criteria:
 - $AR_s = \max(S)$ and $\max(S)$ is different to AR_1 or AR_{40}
 - $s = 1$ or 40 if AR_1 or $AR_{40} = \max(S)$ and $C_s/40 > 0.5$ (i.e. stable in more than 50% of all iterations)



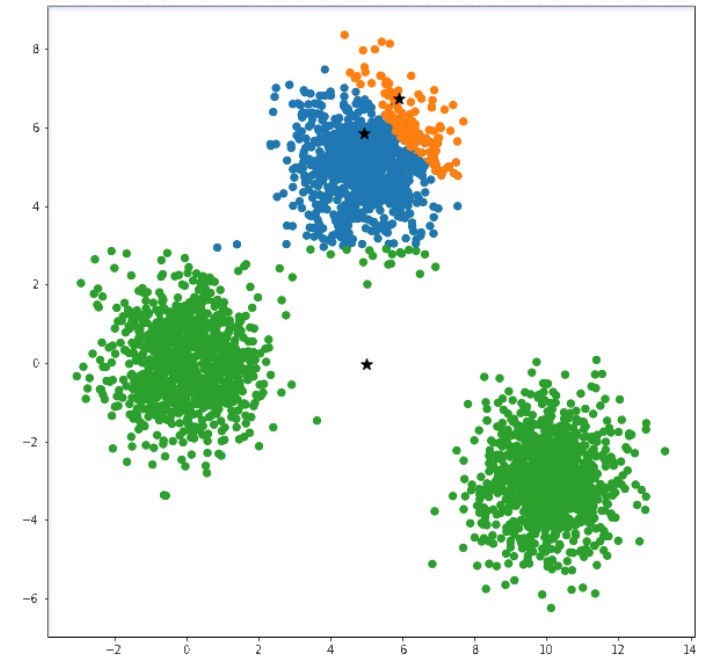
Cluster cells in expression space - Distance measures

- Clustering starts with computing a distance matrix between cells
- Distance between two cells i and j , x_{ig} is the expression of the gene g in the cell C_i

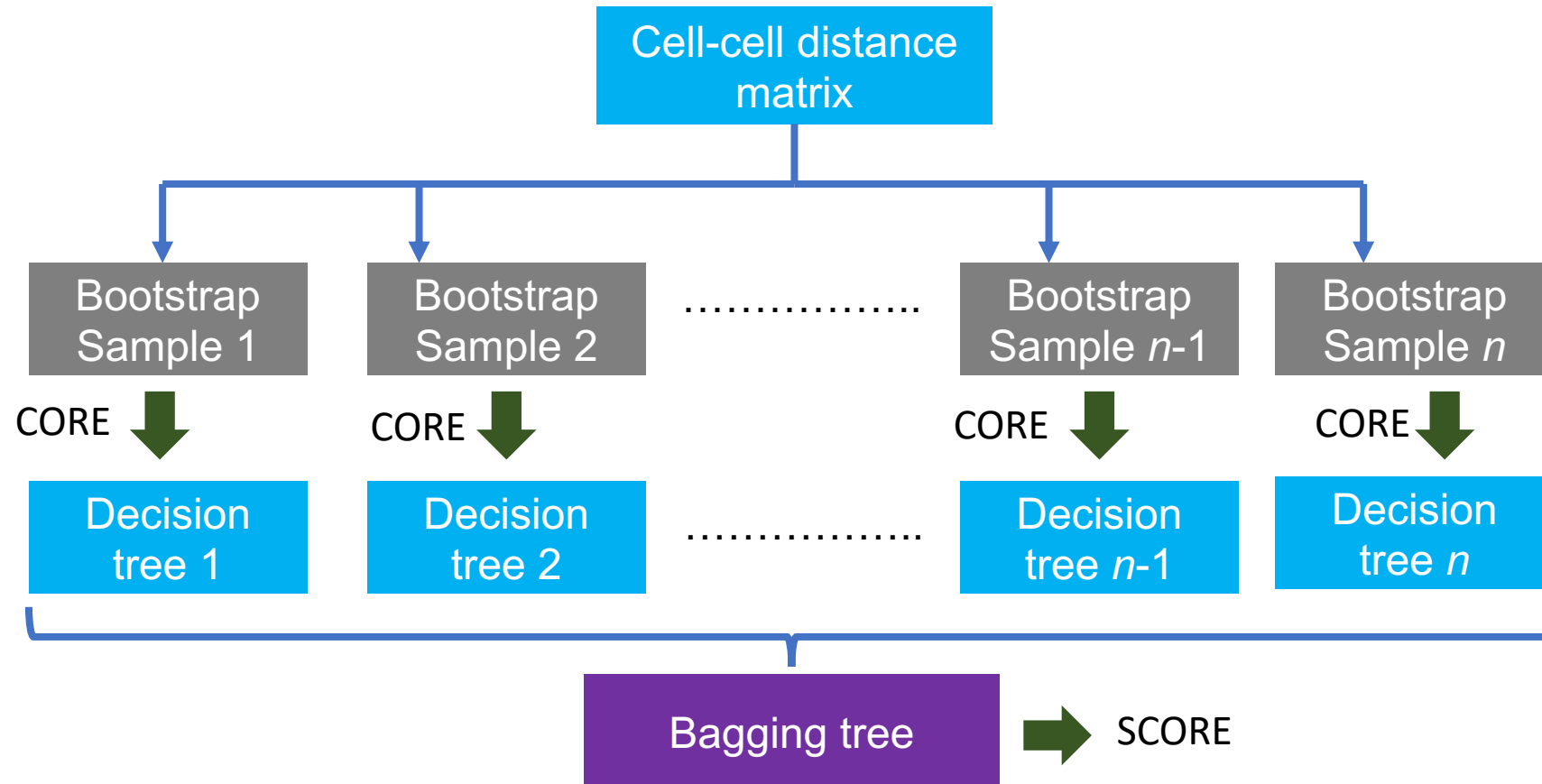
Euclidean distance	$d_{ij} = \sqrt{\sum_{g=1}^G (x_{ig} - x_{jg})^2}$
Manhattan distance	$d_{ij} = \sum_{g=1}^G x_{ig} - x_{jg} $
Maximum distance	$d_{ij} = \max_g x_{ig} - x_{jg} $



cells in gene expression space



Bootstrap and bagging strategies to select stable clusters



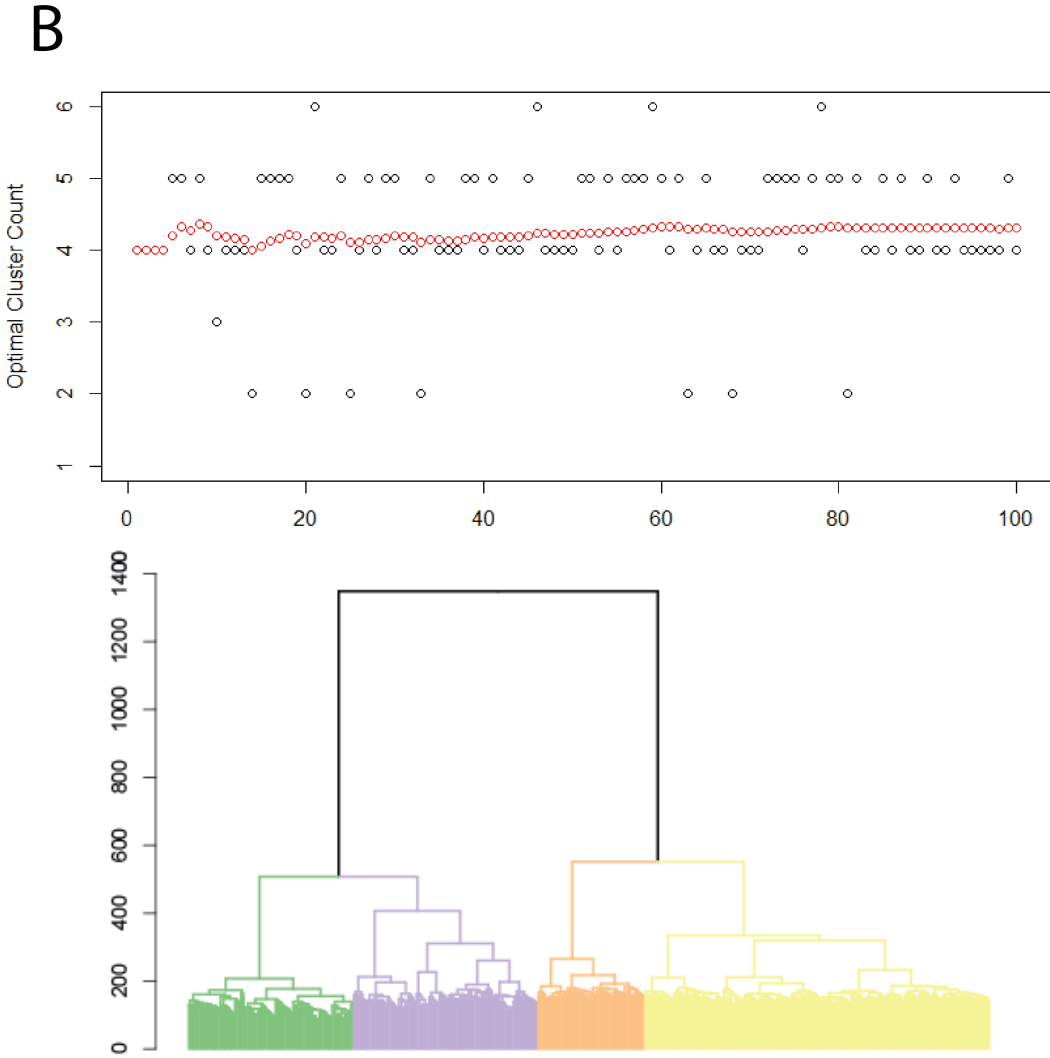
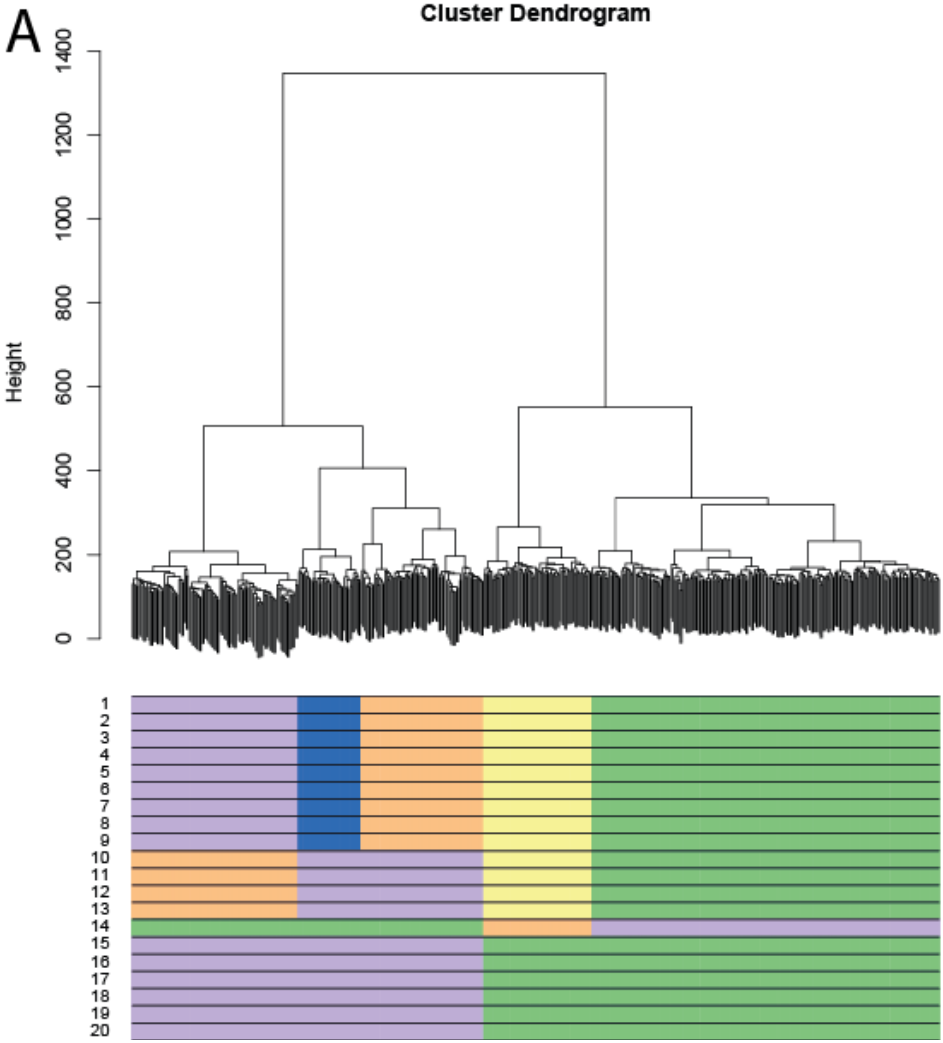
Clustering stability results from:

- Iterative grouping of cells in different search space of the clustering tree
- Bootstrap aggregating (bagging) ensemble algorithm

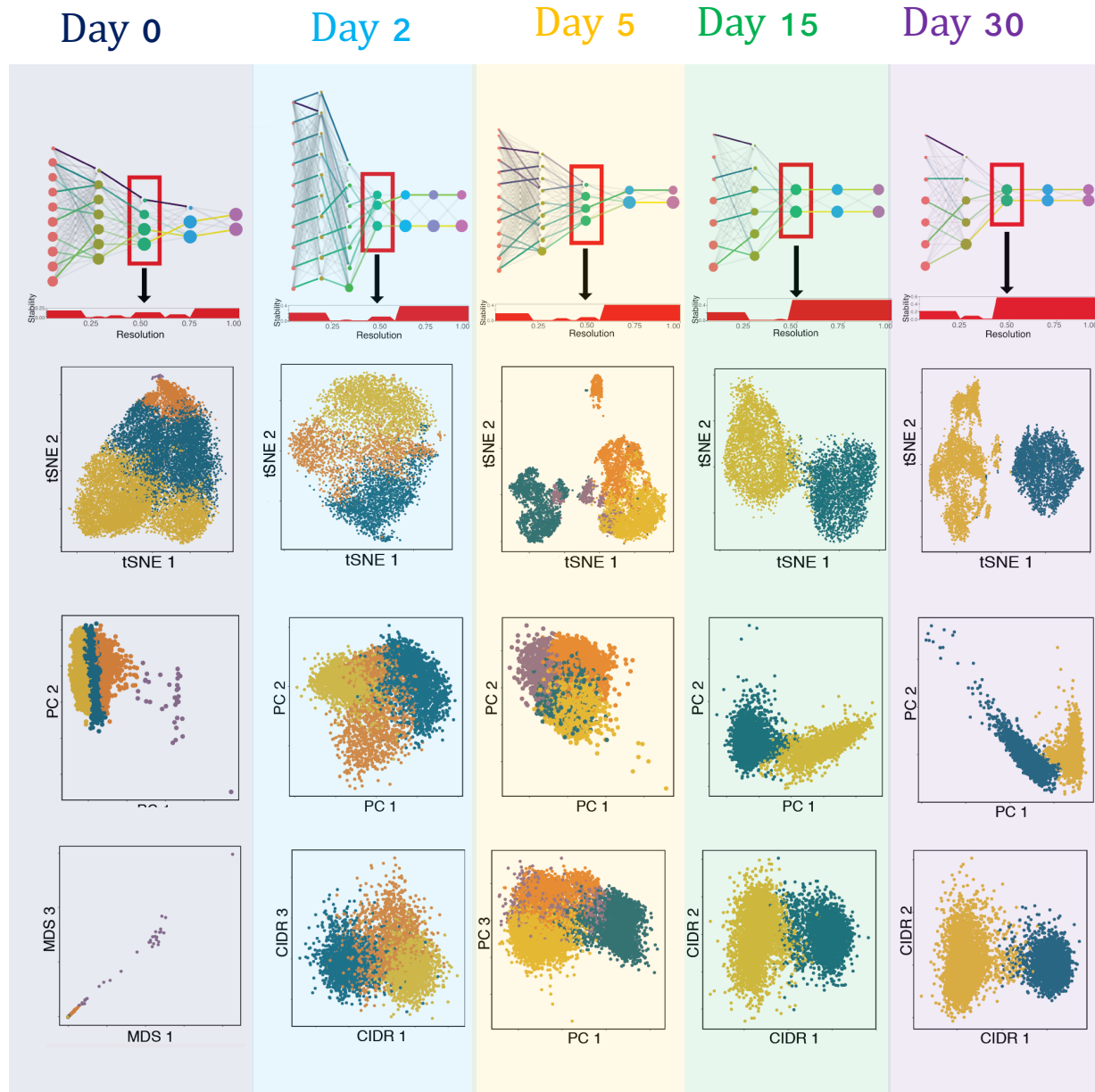
Bootstrap and bagging strategy to select stable clusters

1. Bagging strategies are used for re-clustering random sub-sets of cells from the population to generate additional dendrogram trees.
2. For each bagging run, choose a vector \mathbf{b}_k ($k= 1,2,\dots,m$) of length $p*\dim(C)$ ($p\leq 1$) containing a sample, with replacement, from set C and create a new matrix N_k , of Euclidean distances for the cells in \mathbf{b}_k .
3. For each N_k , a new dendrogram tree is generated and clustered, then an optimal stability is computed.
4. The most stable clustering result is then chosen from the original tree. By default the most commonly occurring stability from the bagging results and use it as the cluster count for the original dendrogram.

Bootstrap and bagging strategy to select stable clusters



Subpopulations identified by CORE are distinguishable



Day 0	Day 2	Day 5	Day 15	Day 30
D0:S1 Core pluripotent	D2:S1 Definitive endoderm	D5:S1 CM precursor	D15:S1 Non-contractile	D30:S1 Non-contractile
D0:S2 Proliferative	D2:S2 Mesoderm	D5:S2 Definitive endoderm	D15:S2 Committed CM	D30:S2 Definitive CM
D0:S3 Early-primed	D2:S3 Mesendoderm	D5:S3 Cardiovascular progenitor		
D0:S4 Late-primed		D5:S4 Intermediate		

*CM = Cardiomyocyte

- From a mixed population at each time point, CORE identified 2 to 4 homogenous clusters
- The identified subpopulations were confirmed by independent methods: PCA, MDS, tSNE, CIDR
- The subpopulations are biologically distinct

Graph-based Clustering

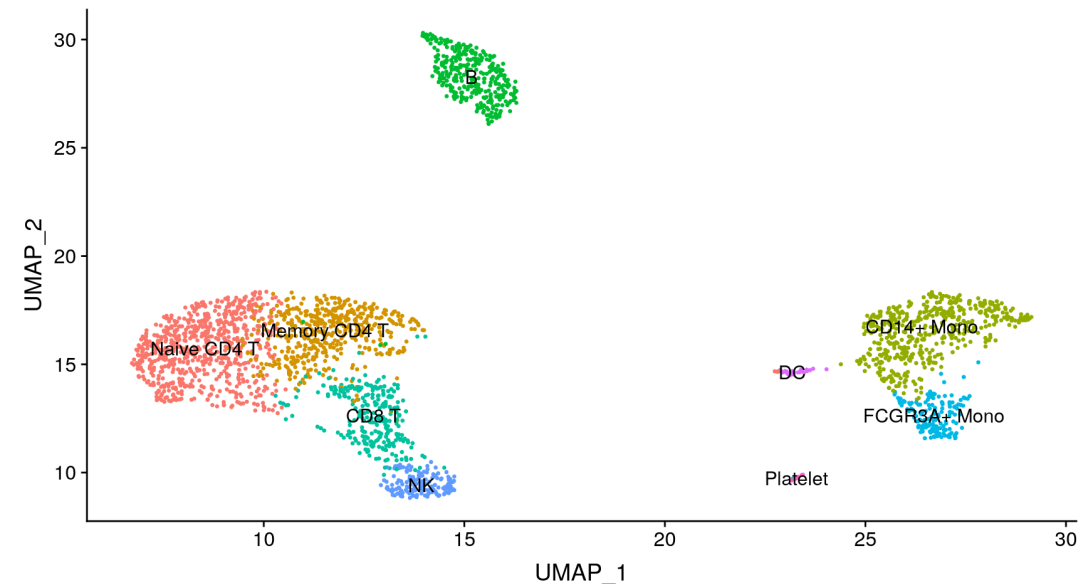
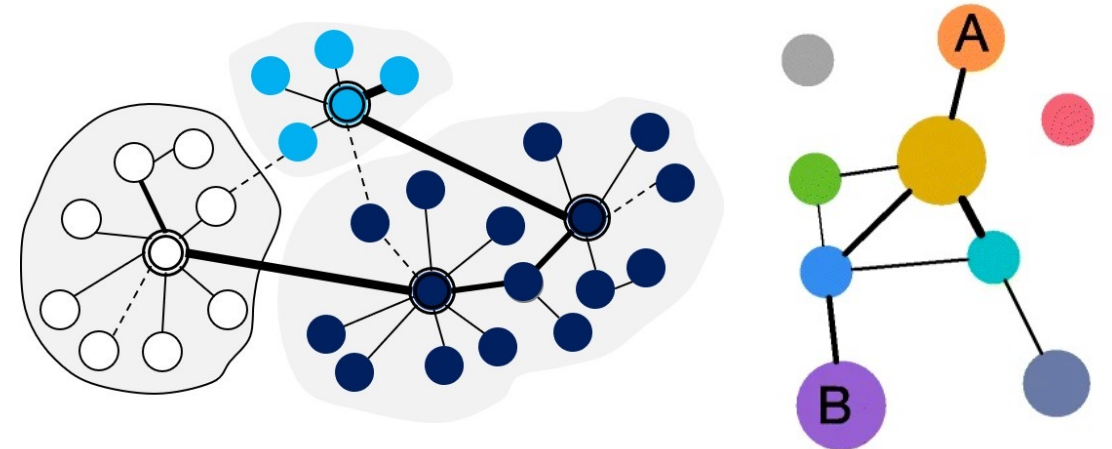
Two main steps:

1) Embed cells in a graph structure:

- K-nearest neighbour (KNN) graph (cells with similar expression patterns identified by Euclidean distance in PCA space)
- Edge weights between any two cells based on the shared overlap in their local neighbourhoods (Jaccard similarity)

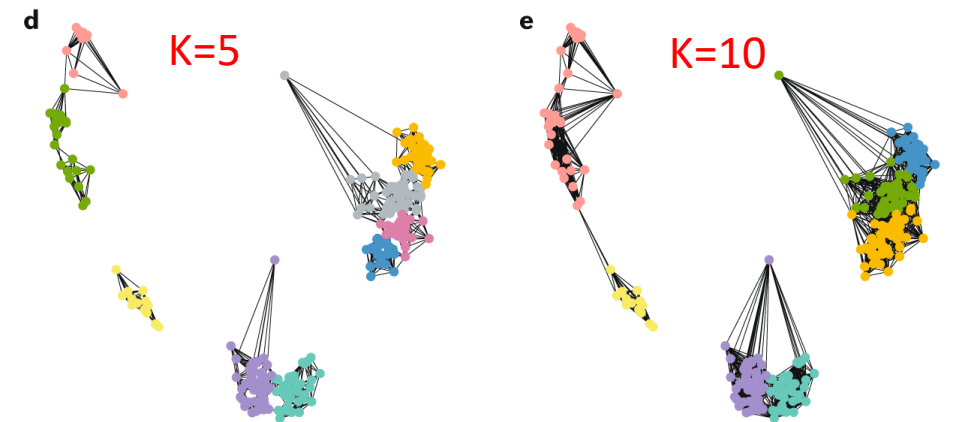
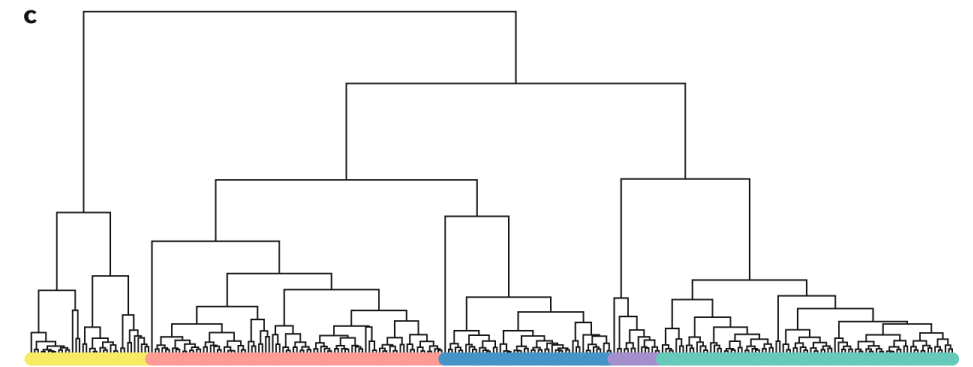
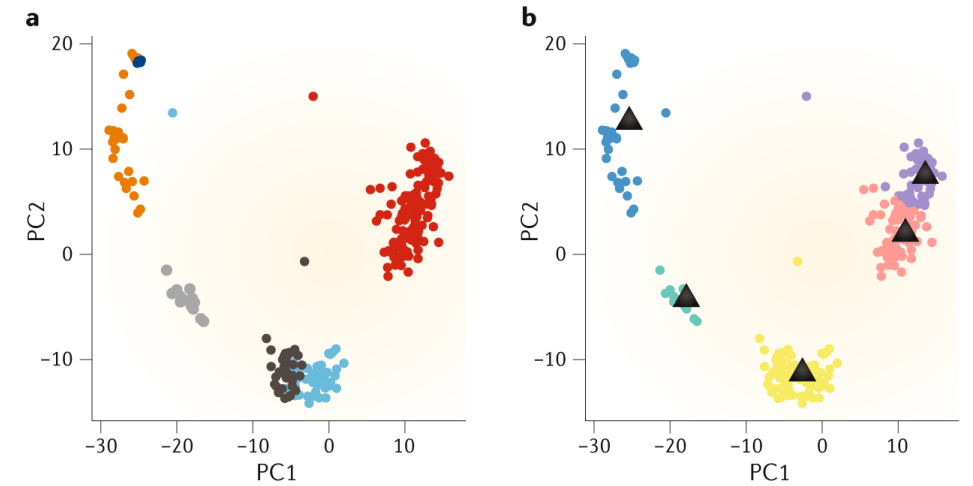
2) Community detection to partition cells in graph into groups of cells

- Modularity optimization techniques such as the Louvain algorithm
- Modularity: measures the density of edges inside communities to edges outside communities
- Louvain iteratively groups cells together, with the goal of optimizing the standard modularity function

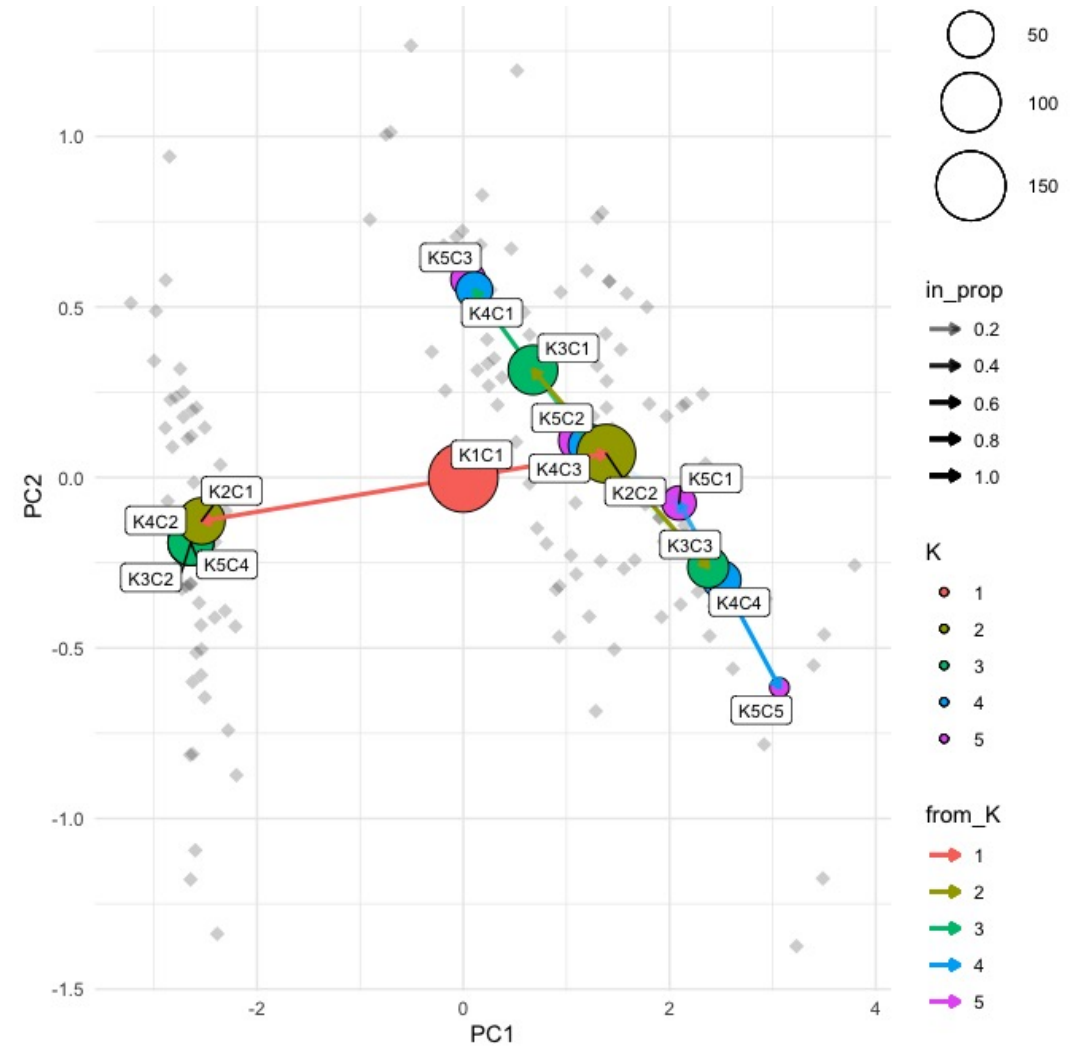
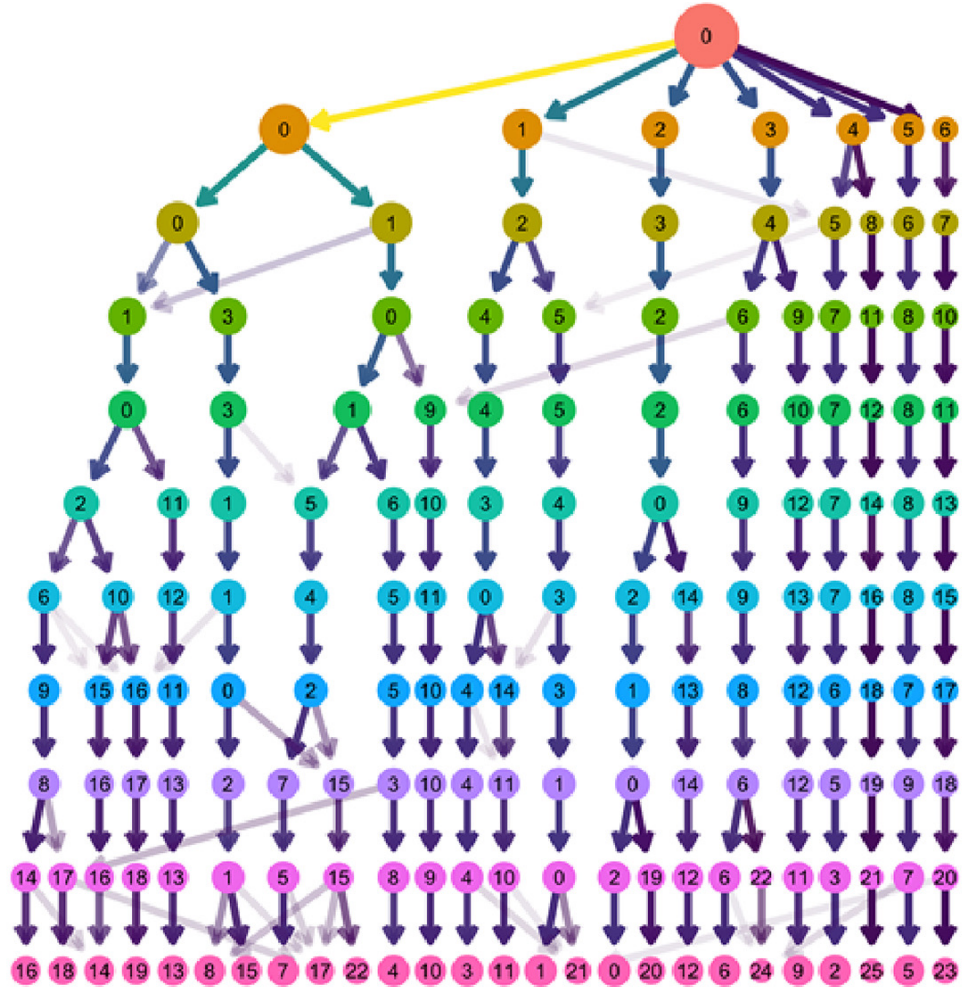


Graph-based Clustering

- Build shared-nearest-neighbour graph connecting the cells and finds tightly connected communities
- Increasing the number of neighbours when constructing the cell-cell graph indirectly decreases the resolution of graph-based clustering



Visualise clustering results



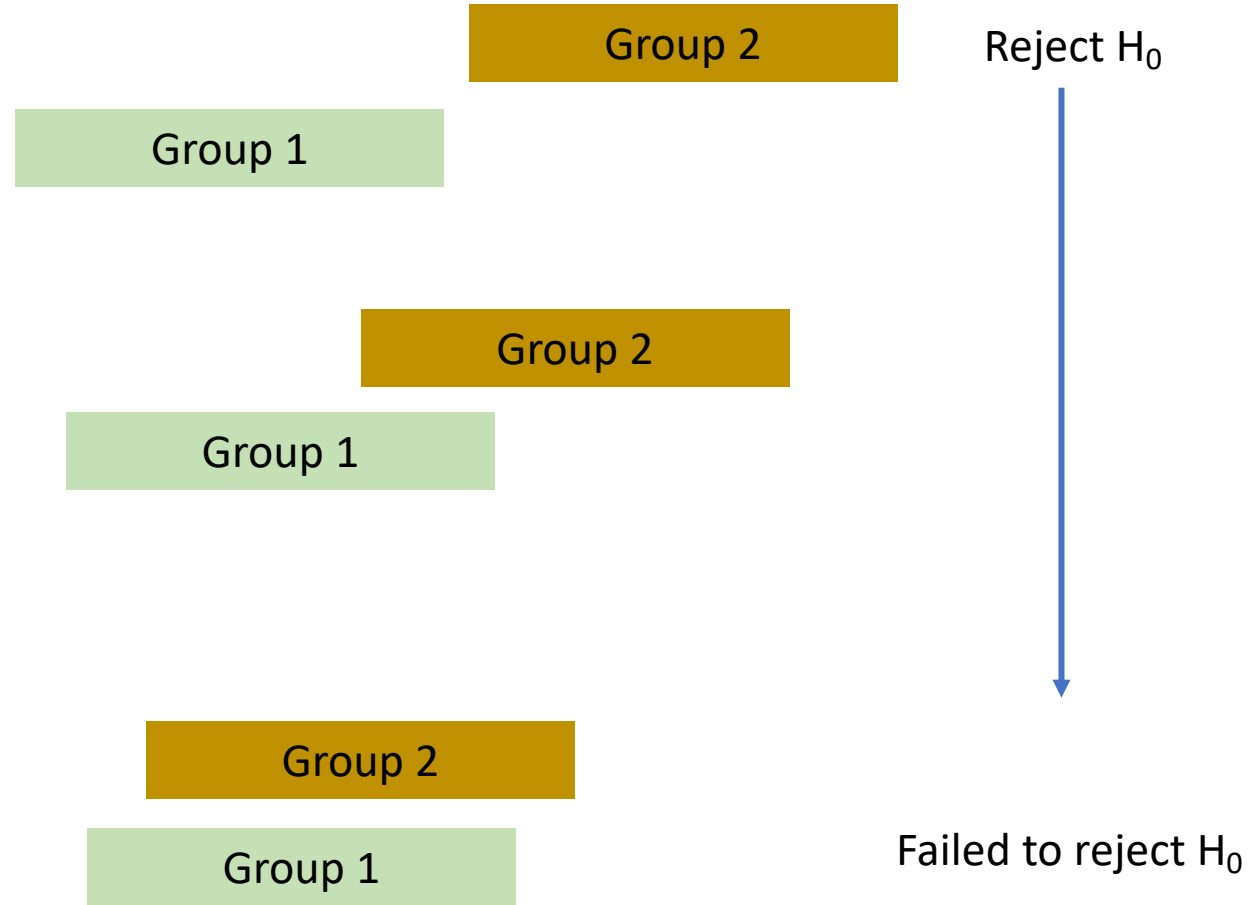
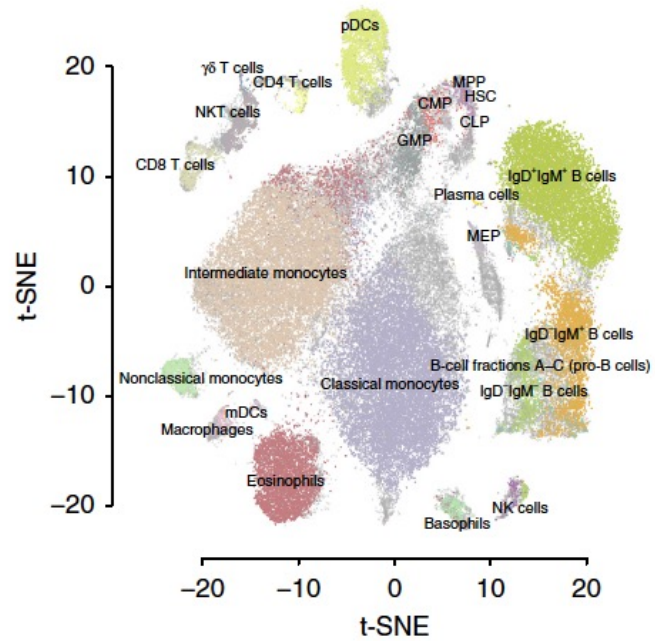
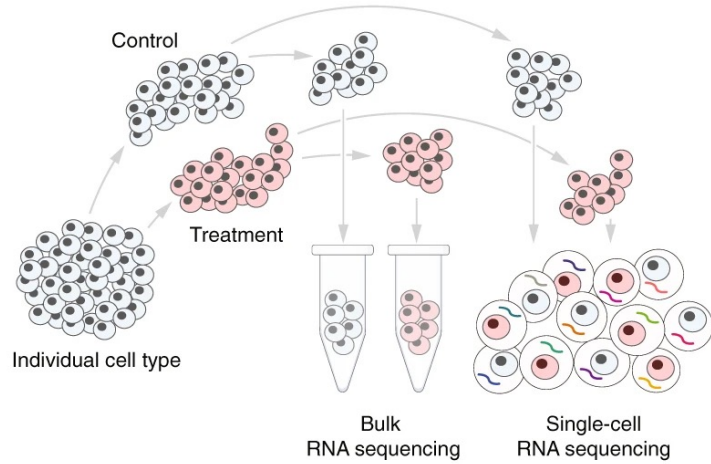
Statistical evaluation of clustering results

Adjusted Rand index (ARI)	$\text{ARI} = \frac{2(ad - bc)}{(a + b)(b + d) + (a + c)(c + d)}$
Jaccard index	$\text{Jaccard} = \frac{a}{a + b + c}$
Fowlkes–Mallows index (FM)	$\text{FM} = \sqrt{\left(\frac{a}{a + b}\right) \left(\frac{a}{a + c}\right)}$

- a: the number of **pairs** of cells **correctly** partitioned into the same cluster
 - b: the number of **pairs** of cells **wrongly** partitioned into the same cluster
 - c: the number of **pairs** of cells **wrongly** partitioned into different clusters
 - d: the number of **pairs** of cells **correctly** partitioned into different clusters
- > higher index scores (max = 1) mean more accurate clustering results

Differential expression analysis

Why DE



Four main categories

- Parametric tests
 - E.g. T-test
- Non-parametric tests
 - Wilcoxon rank-sum test, Kolmogorov–Smirnov (KS) test
 - Convert observed expression to ranks, then test whether the distribution of ranks for one group is significantly different from the other group
- Bulk RNA-seq based method
 - e.g edgeR DEseq2
- scRNA-seq specific methods
 - e.g MAST (Model-based Analysis of Single-cell Transcriptomics), SCDE
 - Large number of samples (ie. cells) → whole distribution of expression values in each group

Parametric tests

- T test
- Testing for the location of / comparing two means
- Assume: each population is normally distributed

$$\text{T distribution} = \frac{\text{standard normal}}{\text{sqrt}(\text{chi-squared} / \text{df})}$$

- X_1, X_2, \dots, X_m are iid $N(\mu_1, \sigma^2)$, Y_1, Y_2, \dots, Y_n are iid $N(\mu_2, \sigma^2)$

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t_{m+n-1}$$

$$S = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-1}$$

Non-parametric tests

Wilcoxon rank-sum test \approx Mann-Whitney test

- Nonparametric, alternative to the two- sample t-test
- Testing for the location of / comparing two medians
- Can be used with quantitative data or ordinal data
- The data do not have to be normally distributed but do have similar shapes
- Use the RANKS of observations

Non-parametric tests- step by step

Wilcoxon rank-sum test \approx Mann-Whitney test

Tied rank,
average



Group	Treatment	T	T	T	T	Control	C	C	C	C
Gene count	110	100	90	80	70	40	30	20	10	10
Rank	1	2	3	4	5	6	7	8	9.5	9.5

$$\sum(Treatment) = 1+2+3+4+5=15 \quad \sum(control) = 40$$

If the treatment does not have effect, we expect the sum of the rank to be near the middle: $(15+40)/2 = 27.5$

$$\mu_w = \frac{n_1(n_1 + n_2 + 1)}{2} \quad \sigma_w = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} \quad Z = \frac{w - \mu_w}{\sigma_w}$$

w is sum of the ranks for the sampler sample, if $n_1 \neq n_2$

Linear model for differential expression

LIMMA

- Generalized linear model
- $\log(y_{igk}) = \mu_j + \alpha_{ig} + error_{igk}$
 - Separate model for each gene g
 - k is a specific sample
 - μ_g is mean expression for gene g over all samples
 - α_{ig} is deviation of the mean of the i th condition from the overall mean
- $H_0: \alpha_{treat, gene g} = \alpha_{control, gene g}$ no difference in treatment and control group

Assumption using log as link function: $y_{igk} \sim \text{Poisson} \rightarrow \text{mean} = \text{variance}$

However, often observe mean < variance \rightarrow thus, Log-normal over correct data dispersion $\rightarrow y_{ijk} \sim \text{negative binomial distribution}$

edgeR

- Generalized linear model

Expression level of interest

$$y_{gi} \sim NB(\mu_{gi}, \varphi_g) = NB(M_{gi}\lambda_{gi}, \varphi_g)$$

Raw count for gene g, sample i

Normalization factor

Dispersion for gene g

$Var(y_{gi}) = \mu_{gi} + \varphi_g \mu_{gi}^2$ if $\varphi_g = 0 \rightarrow$ NB becomes Poisson

Gamma-Poisson mixture

Biological variance \sim Gamma

Measurement error \sim Poisson

$$H_0: \lambda_{gi} = \lambda_{gj}$$

MAST

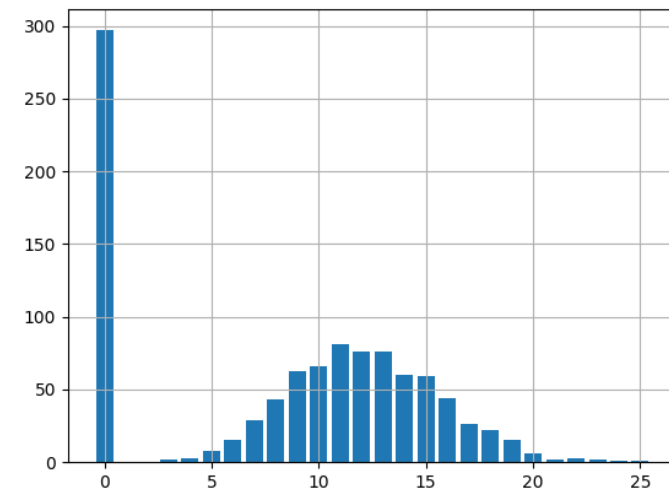
Hurdle model

- a two-part generalized linear model
 - models the rate of expression over the background of various transcripts
 - the positive expression mean.

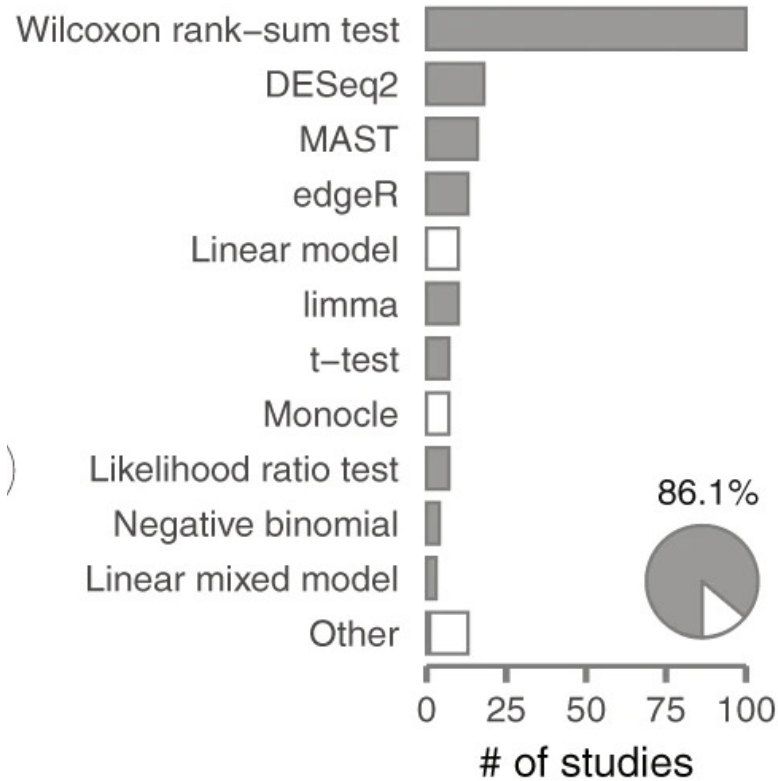
$$\text{logit}(\Pr(Z_{ig} = 1)) = X_i \beta_g^D$$

$$\Pr(Y_{ig} = y | Z_{ig} = 1) = N(X_i \beta_g^C, \sigma_g^2)$$

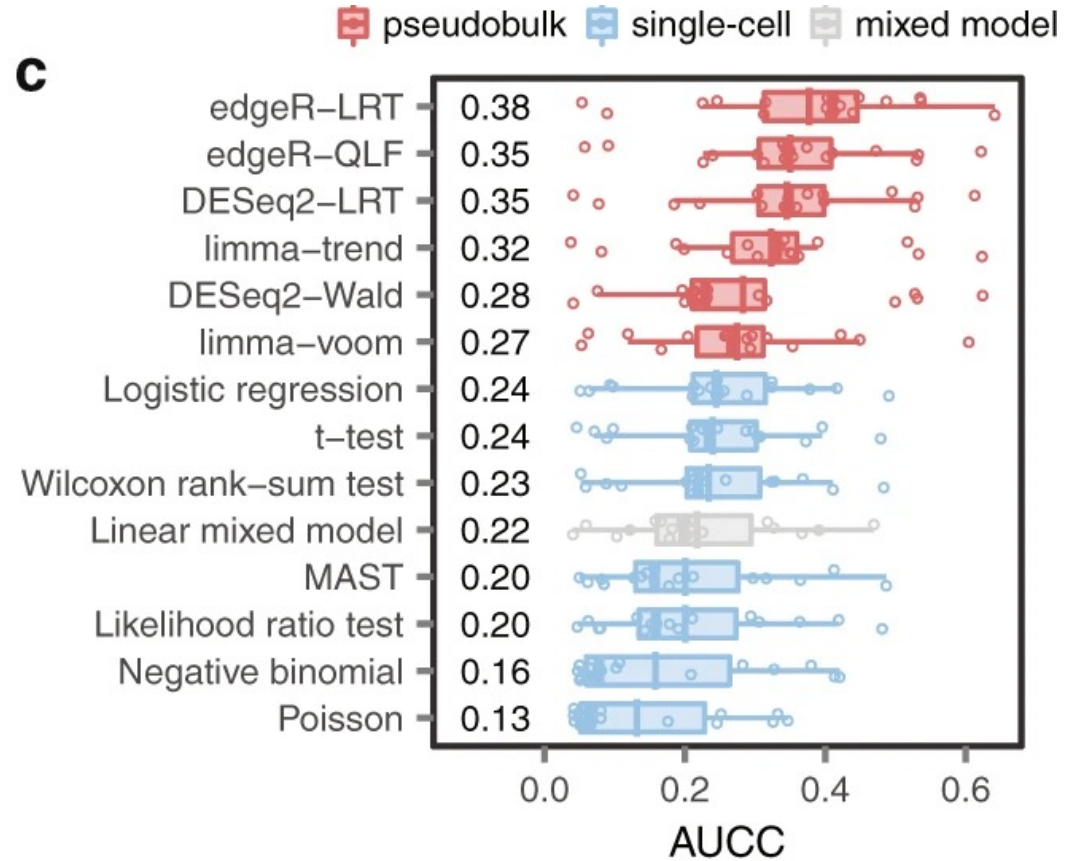
→ logistic regression to estimate the probability being zero
linear regression to estimate the mean of non-zero



Comparison between different methods



c

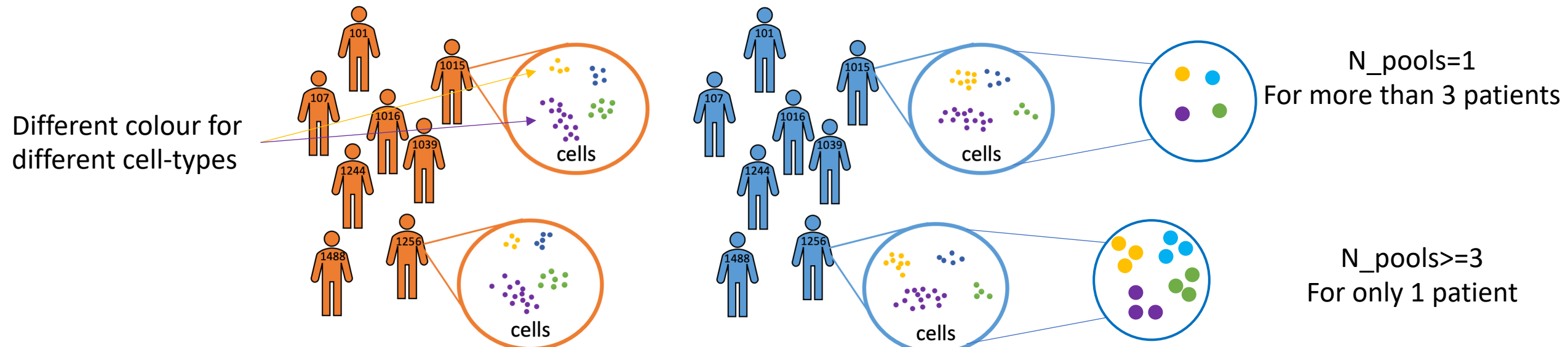


Pseudobulk differential expression analysis

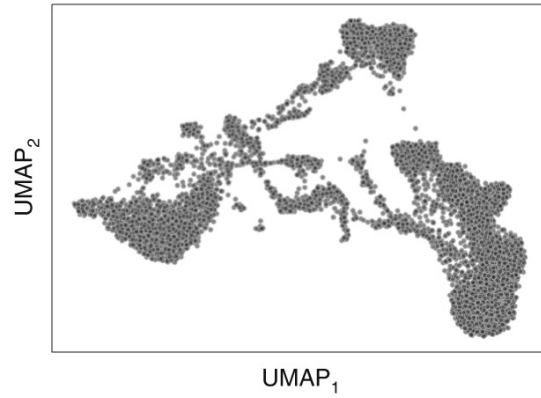
A pseudobulk sample is formed by aggregating the expression values from a group of cells (single-cell) from the same individual. The cells are typically grouped by cell type assignment, clustering or they are randomly sampled into multiple groups ($N \geq 3$).

Why?

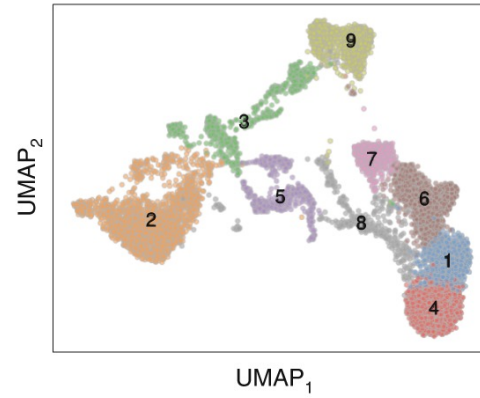
1. Forming pseudobulk samples is important to perform accurate differential expression analysis.
2. Cells from the same individual are more similar to each other than to other individuals' cells so, treating each cell as an independent sample leads to underestimation of the variance and misleadingly small p-values.
3. Working on the level of pseudobulks ensures reliable statistical tests because the samples correspond to the units of replication.



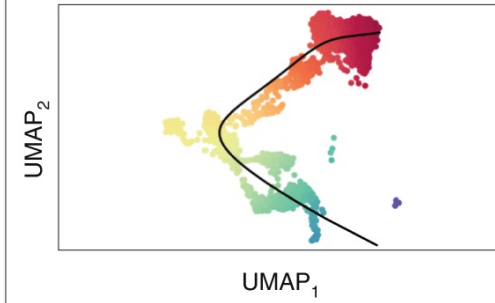
Dimensionality reduction



Clustering

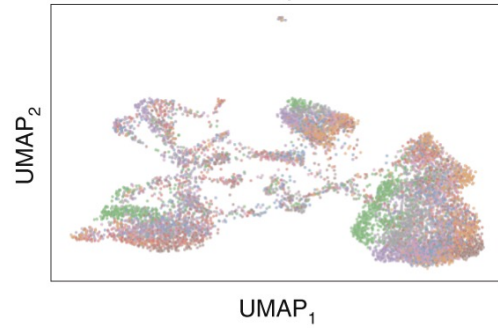


Trajectory analysis



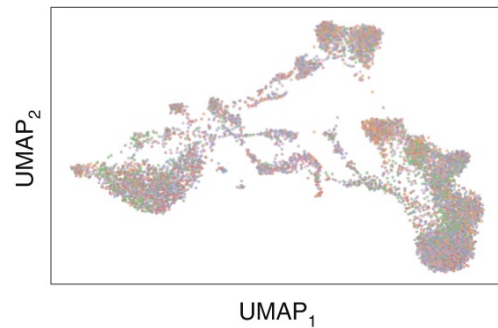
Integrating datasets

Pre-integration



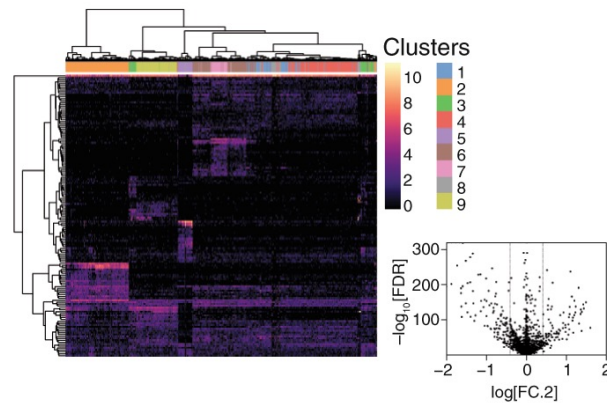
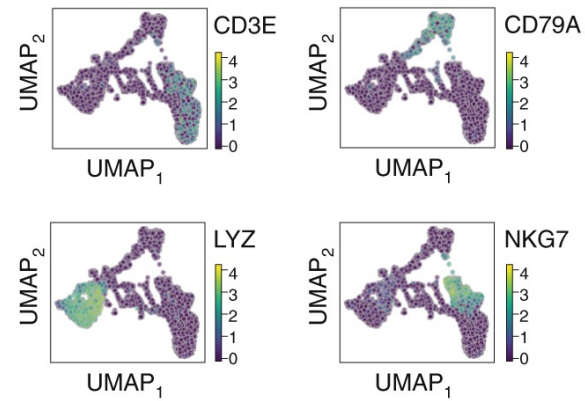
- Donor
- BM1
 - BM2
 - BM3
 - BM4
 - BM5
 - BM6
 - BM7
 - BM8

Post-integration

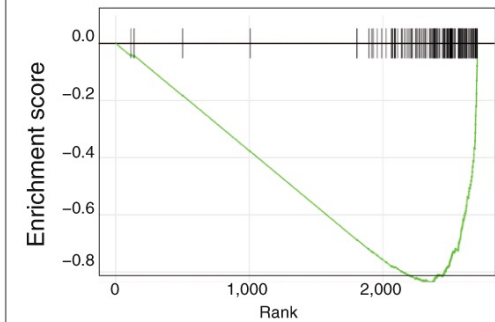


- Donor
- BM1
 - BM2
 - BM3
 - BM4
 - BM5
 - BM6
 - BM7
 - BM8

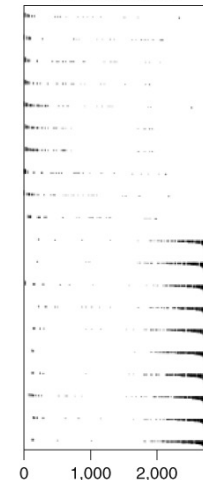
Differential expression



Annotation



Gene ranks

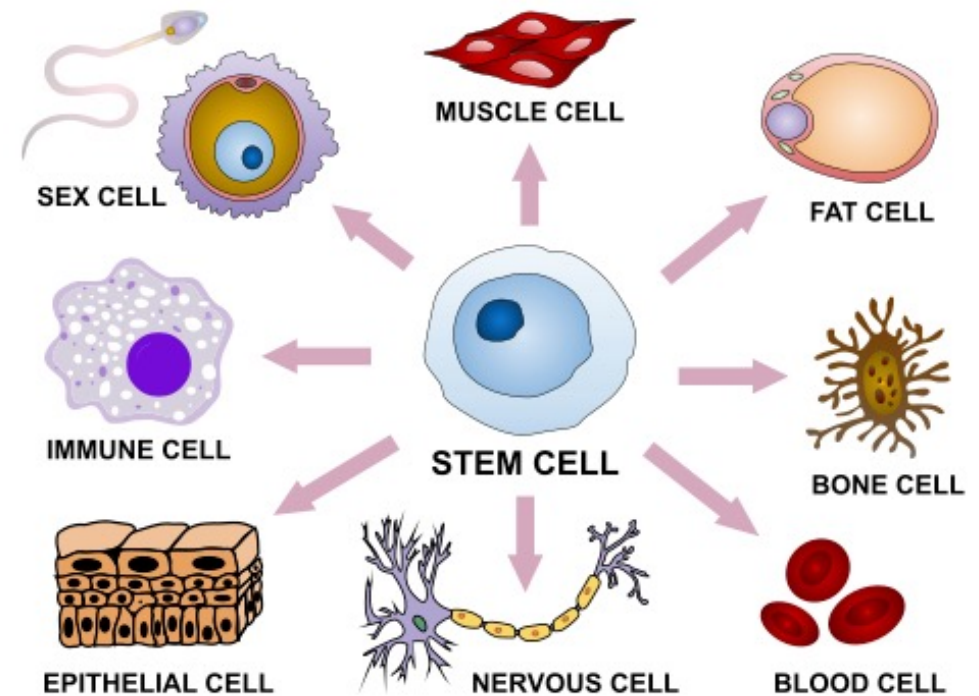


Cell Type Analysis

Cell Type Analysis

What is a cell type?

Cells can be organized into groups based on shared, quantifiable, features (lineage, location, morphology, activity, cell interactions, epigenetic state, cellular response, and molecular composition (mRNA and protein levels)).



Cell Type Analysis

scRNA-seq-based cell classification:

Partition cells into “clusters” based on expression signatures representing a “putative cell type”. This may not correspond to all features above and is also sensitive to cell state.

Cell Type Classification

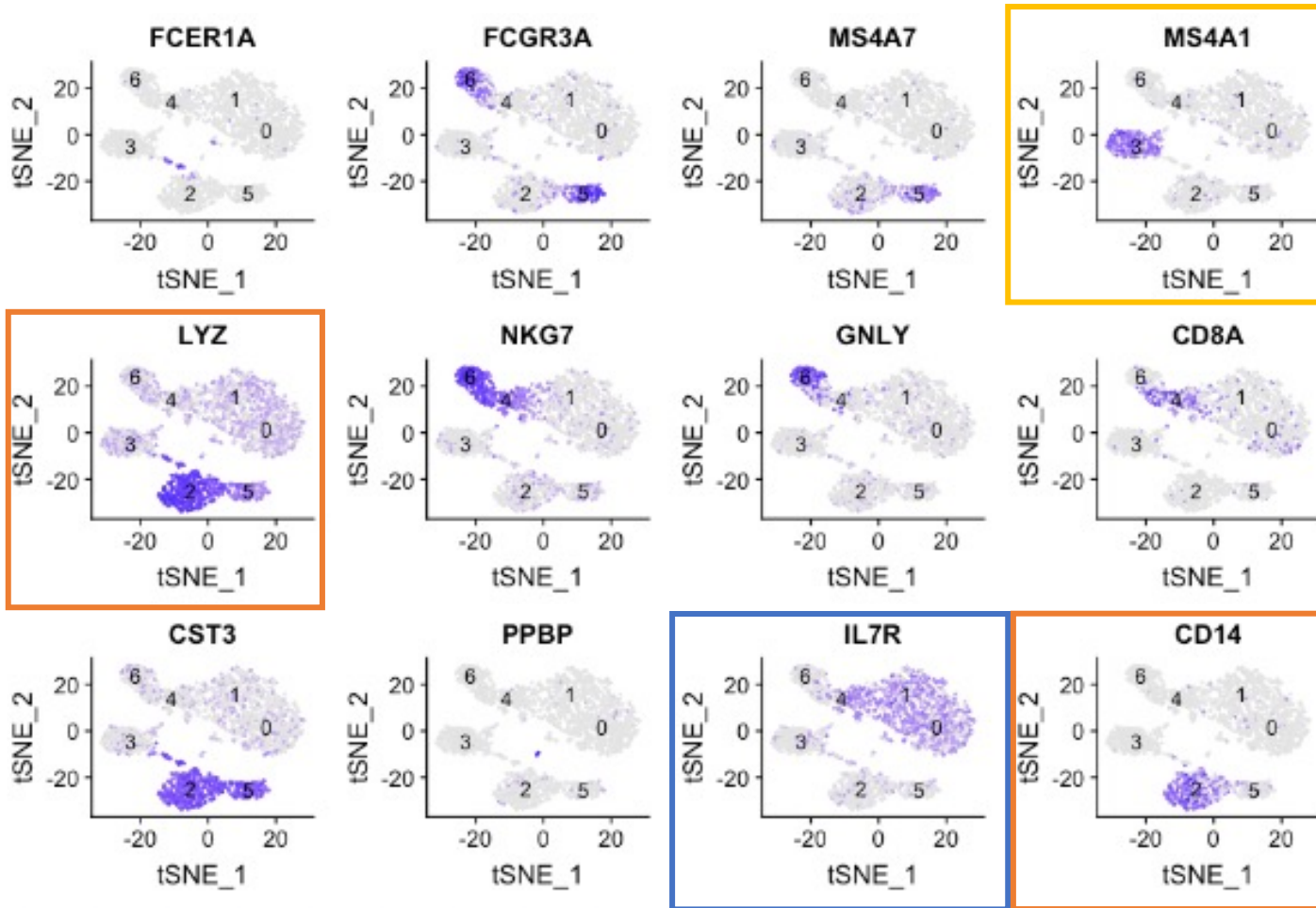
Unsupervised

- Clustering algorithms - cluster cells into groups based on the similarities of the gene expression profiles.
- Use known cell type marker gene lists.
- Cell type labels are assigned to each cluster by manual inspection of gene expression profile of a cluster or by computational tools.
- Can be challenging to specify biologically appropriate number of clusters.
- Relies on expert curated known marker gene lists.
- Seurat v3 clustering, raceID3, LIGER, SC3, Monocle3, TSCAN, pcaReduce and CIDR, SAME-clustering and SHARP.

Supervised

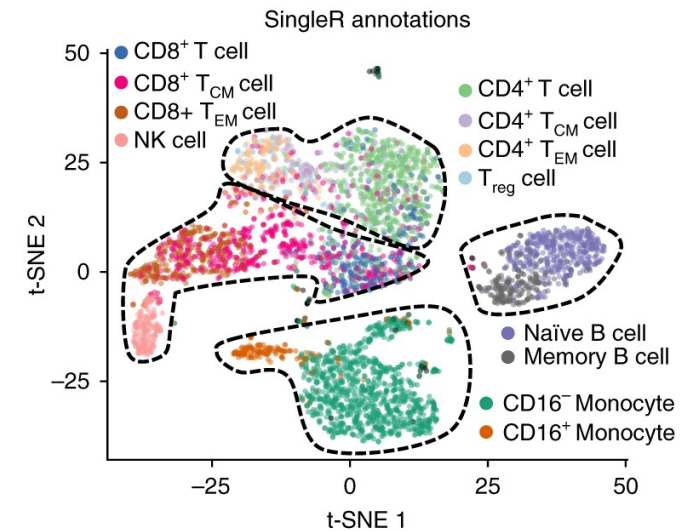
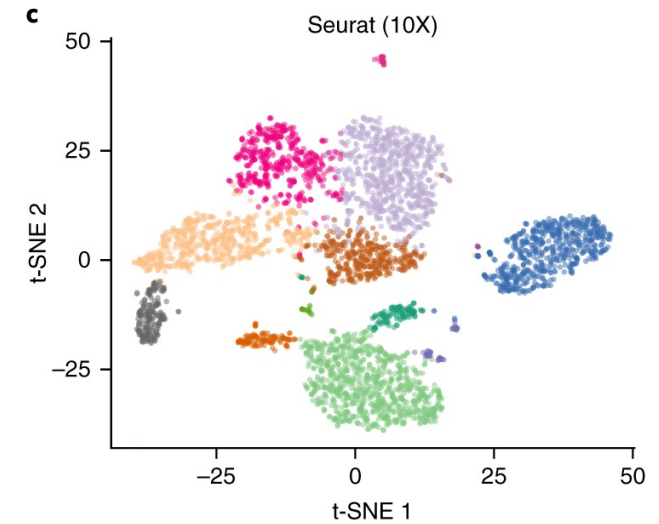
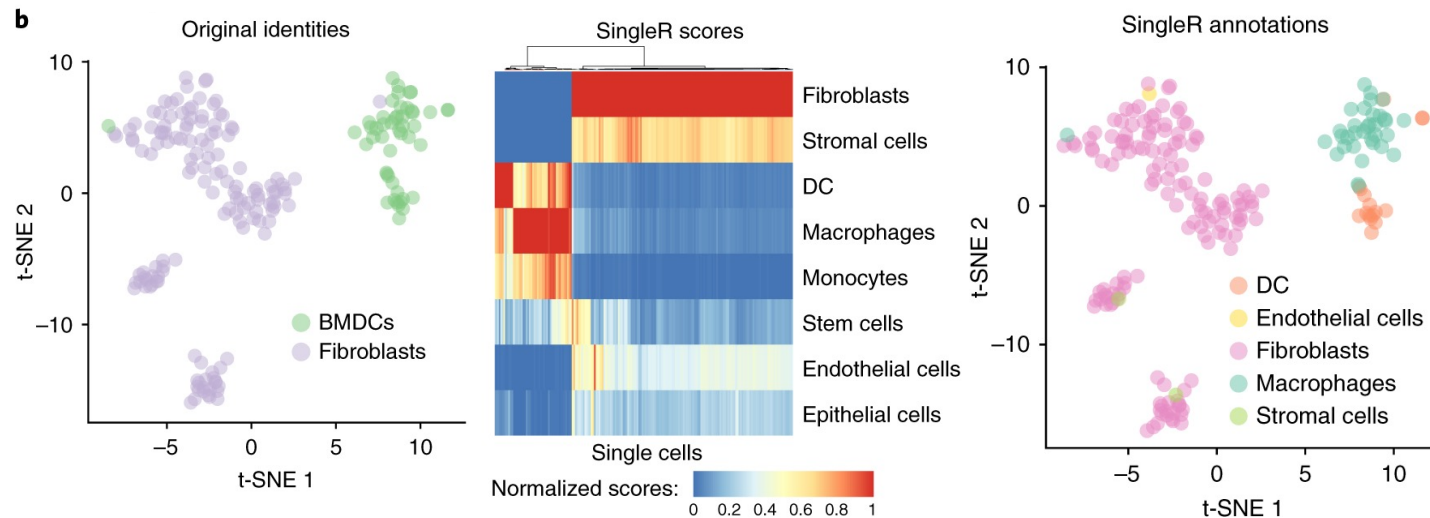
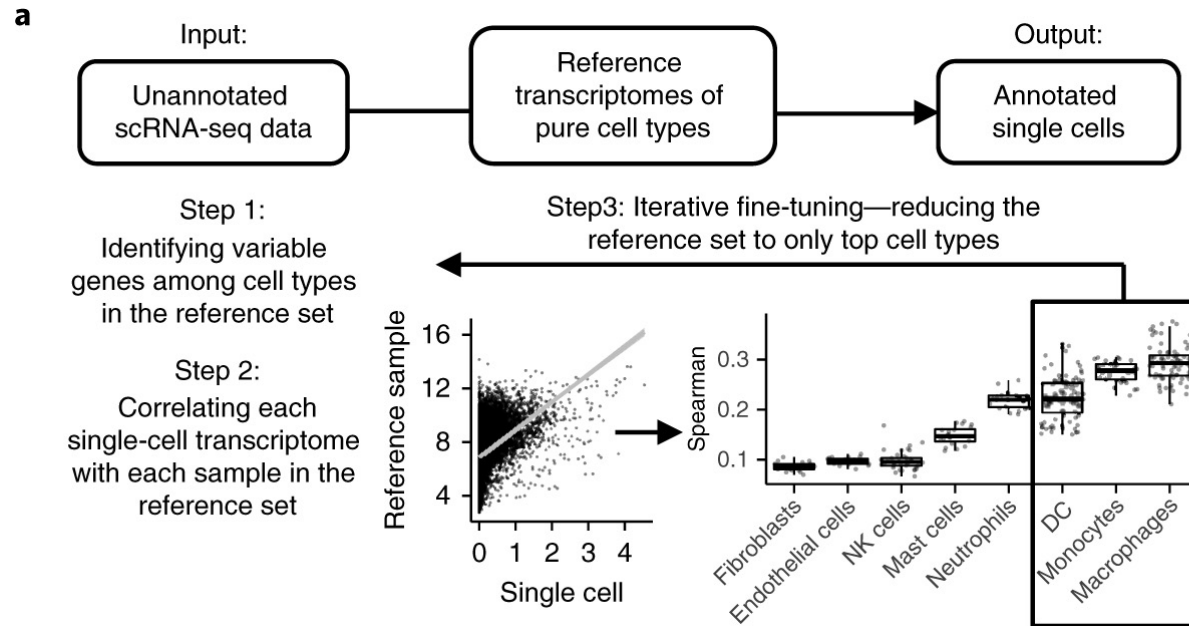
- Require a reference dataset with known cell type annotations.
- They train a classifying model on the reference data, and then apply the trained model to predict the cell types in an unannotated dataset.
- Restricted to the cell types included in the reference data.
- Can be challenging to obtain a suitable reference dataset, especially for novel tissue types.
- scPred, CellAssign, Seurat v3 mapping, scmap-cluster, scmap-cell, singleR, CHETAH, Garnett and SingleCellNet.

Unsupervised example



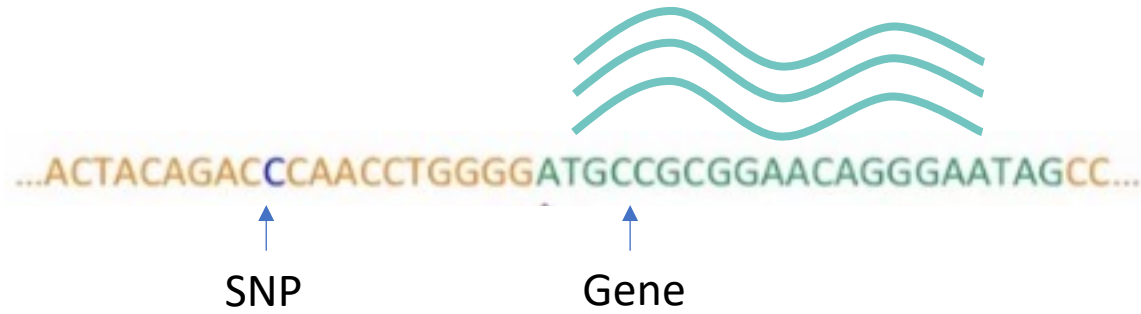
Cluster	Marker	Cell Type
0-1	IL7R	CD4 T cells
2	CD14, LYZ	CD14+ Monocytes
3	MS4A1	B cells
4	CD8A	CD8 T cells
5	FCGR3A, MS4A7	FCGR3A+ Monocytes
6	GNLY, NKG7	NK cells
Unidentified	FCER1A, CST3	Dendritic Cells
Unidentified	PPBP	Megakaryocytes

Supervised example - SingleR

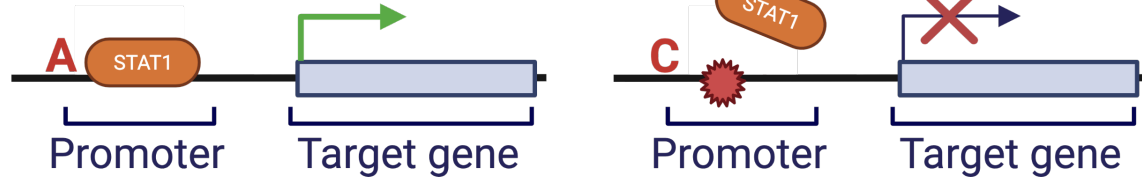


Single-cell eQTL

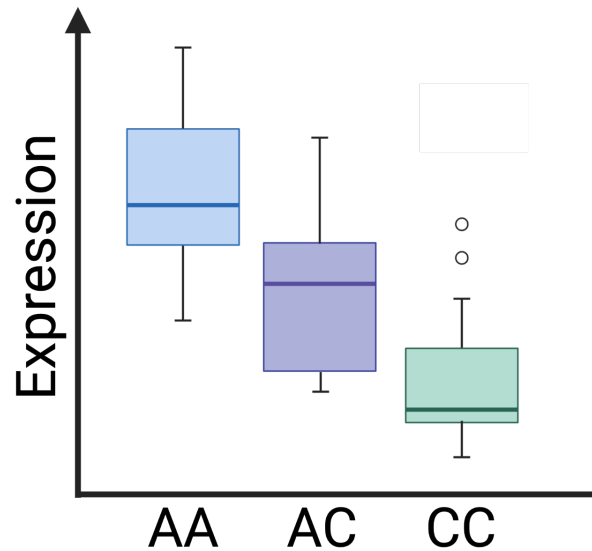
Integration with genomics



Transcription factor



Expression Quantitative Trait Loci (eQTL)



eQTL model: linear regression

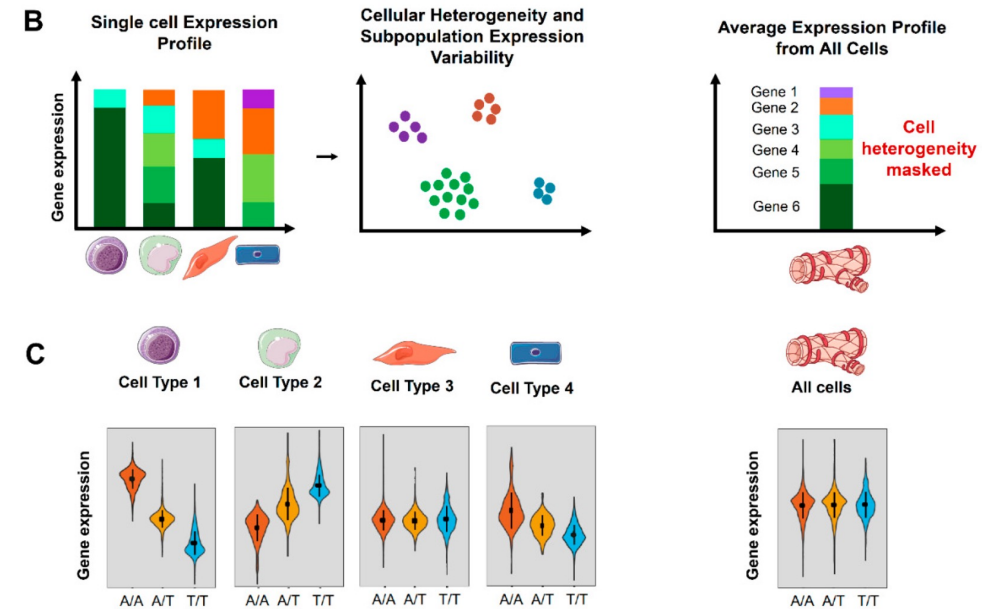
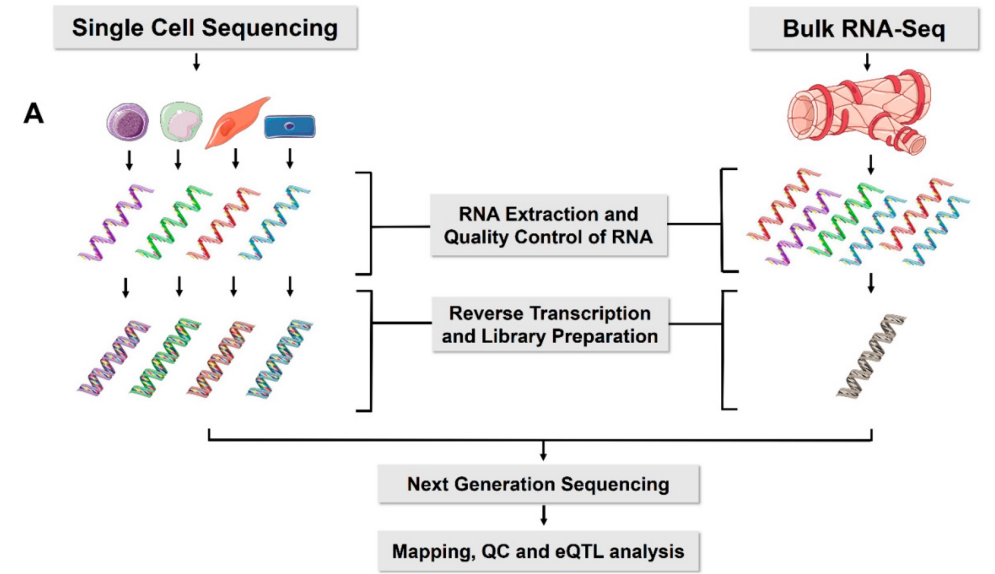
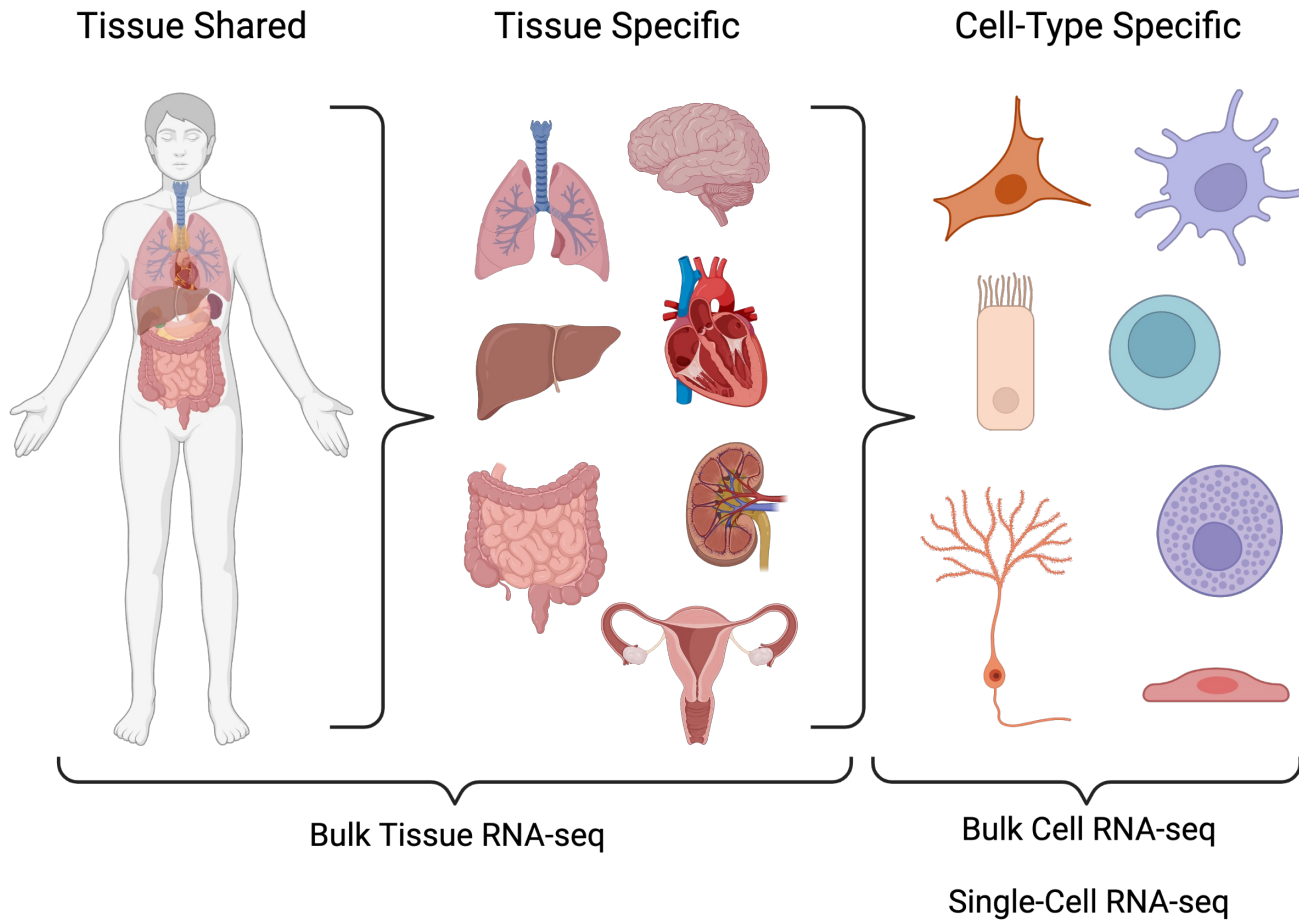
$$\gamma = x\beta + C\alpha + \epsilon$$

Phenotype Effect size residual error

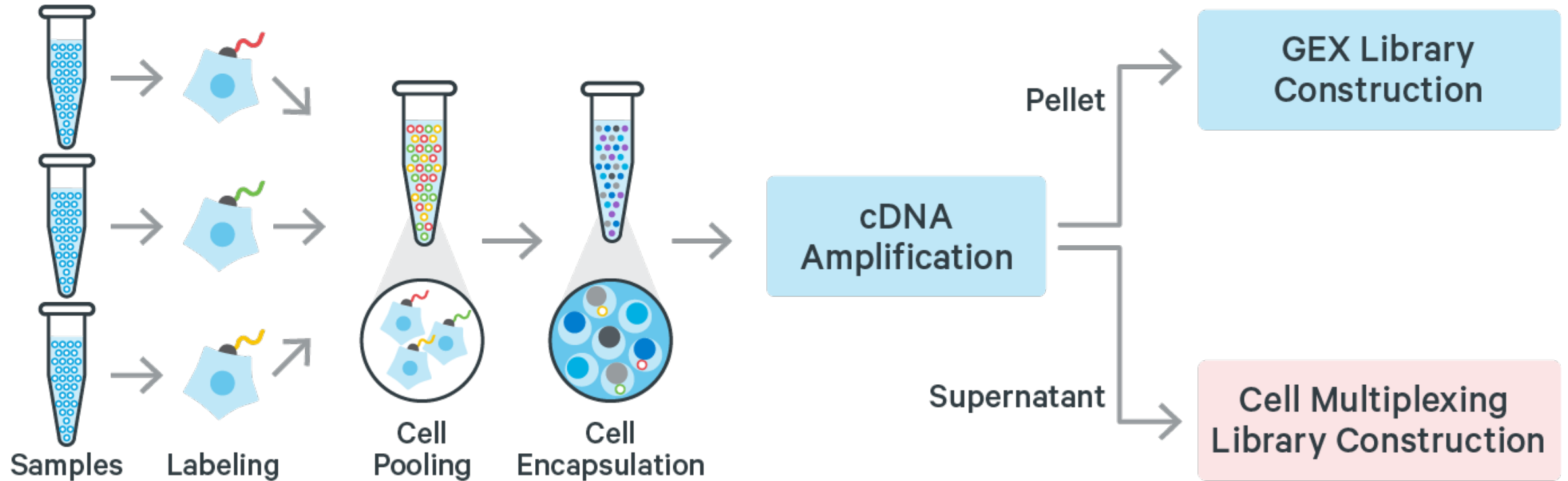
Genotype Covariates: PCs, batches, age, sex

The equation represents the linear regression model for eQTL. The phenotype (γ) is determined by the genotype (xβ), covariates (Cα), and residual error (ε).

Cell-type specific eQTLs



Multiplexing - labeling



Multiplexing and storage of single cell samples



Fix & permeabilize samples

Hybridize probes

Pool (optional)

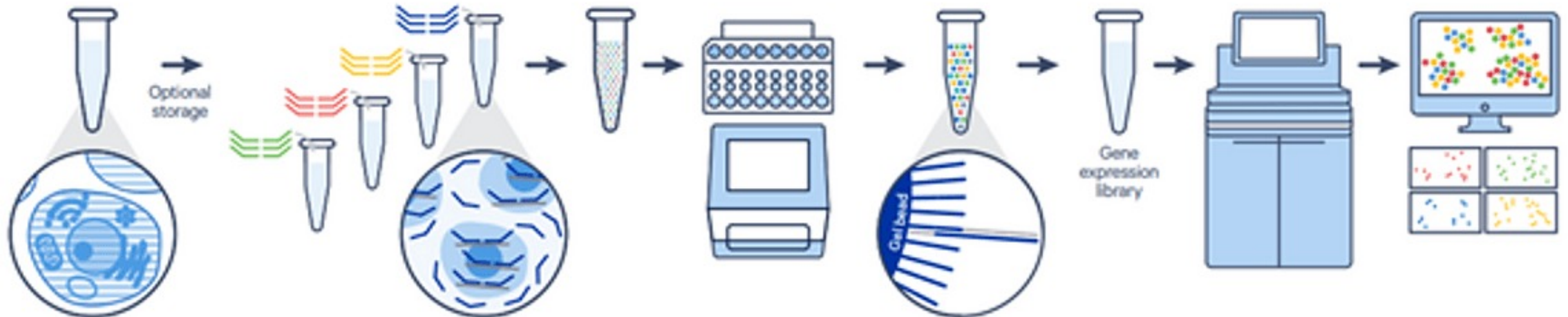
Partition in GEMs

Ligation & extension in GEMs

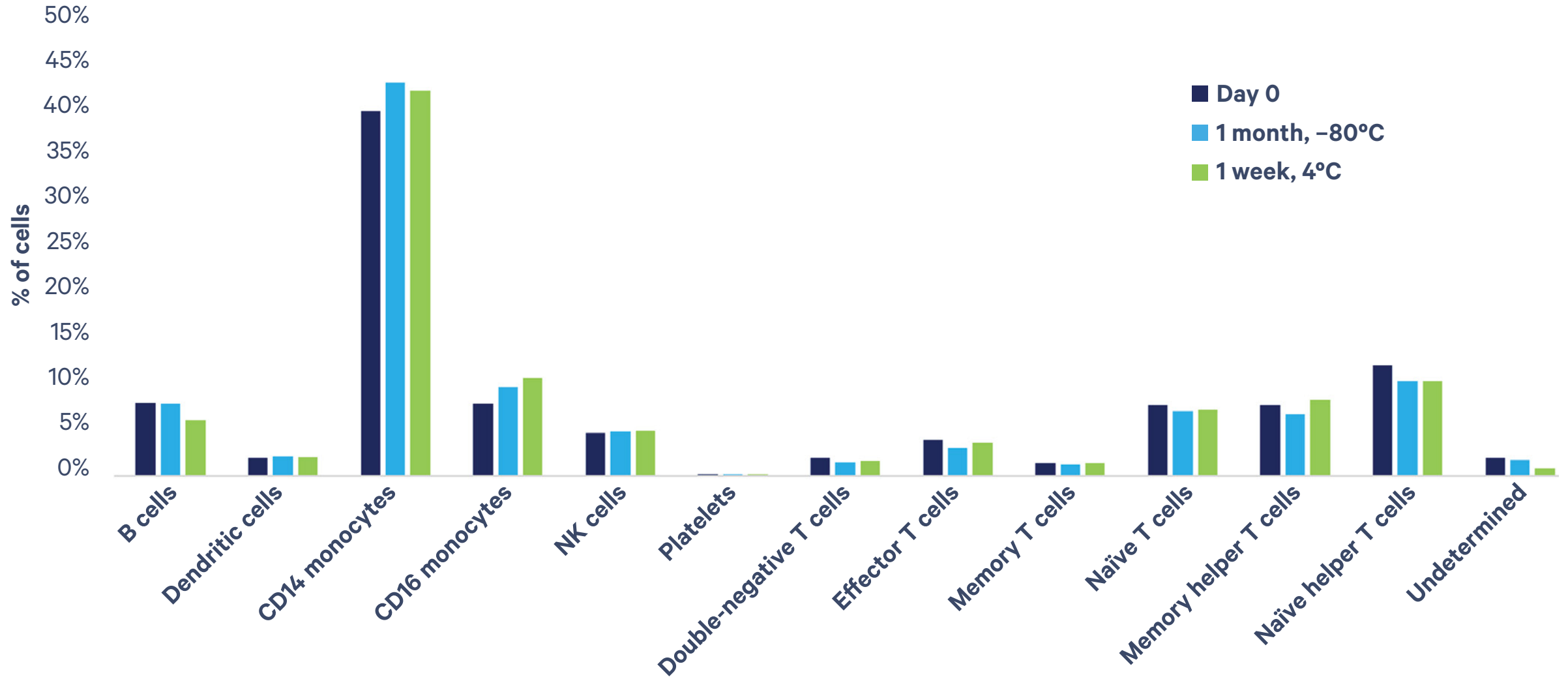
Library construction

Sequencing

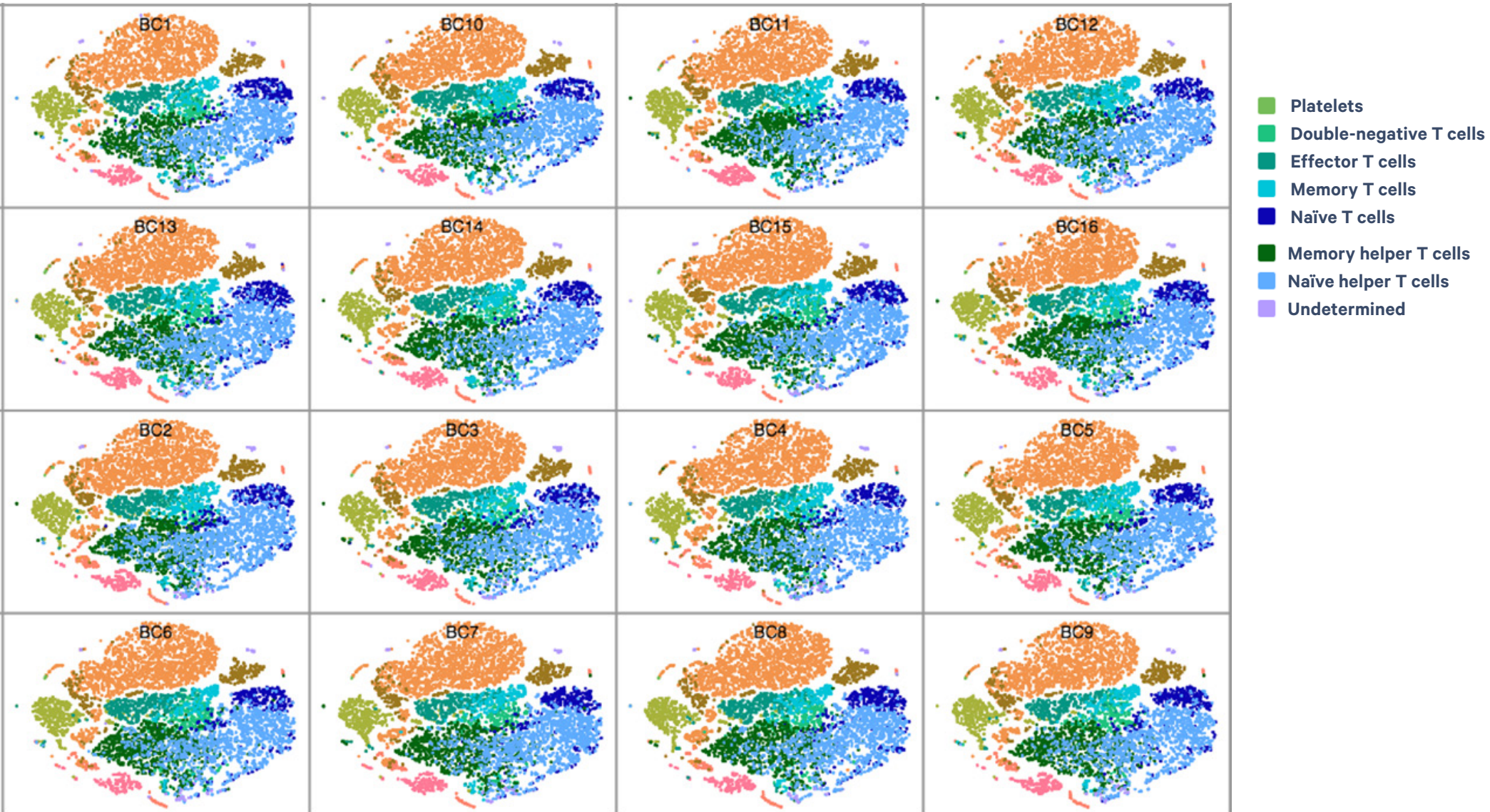
Data analysis



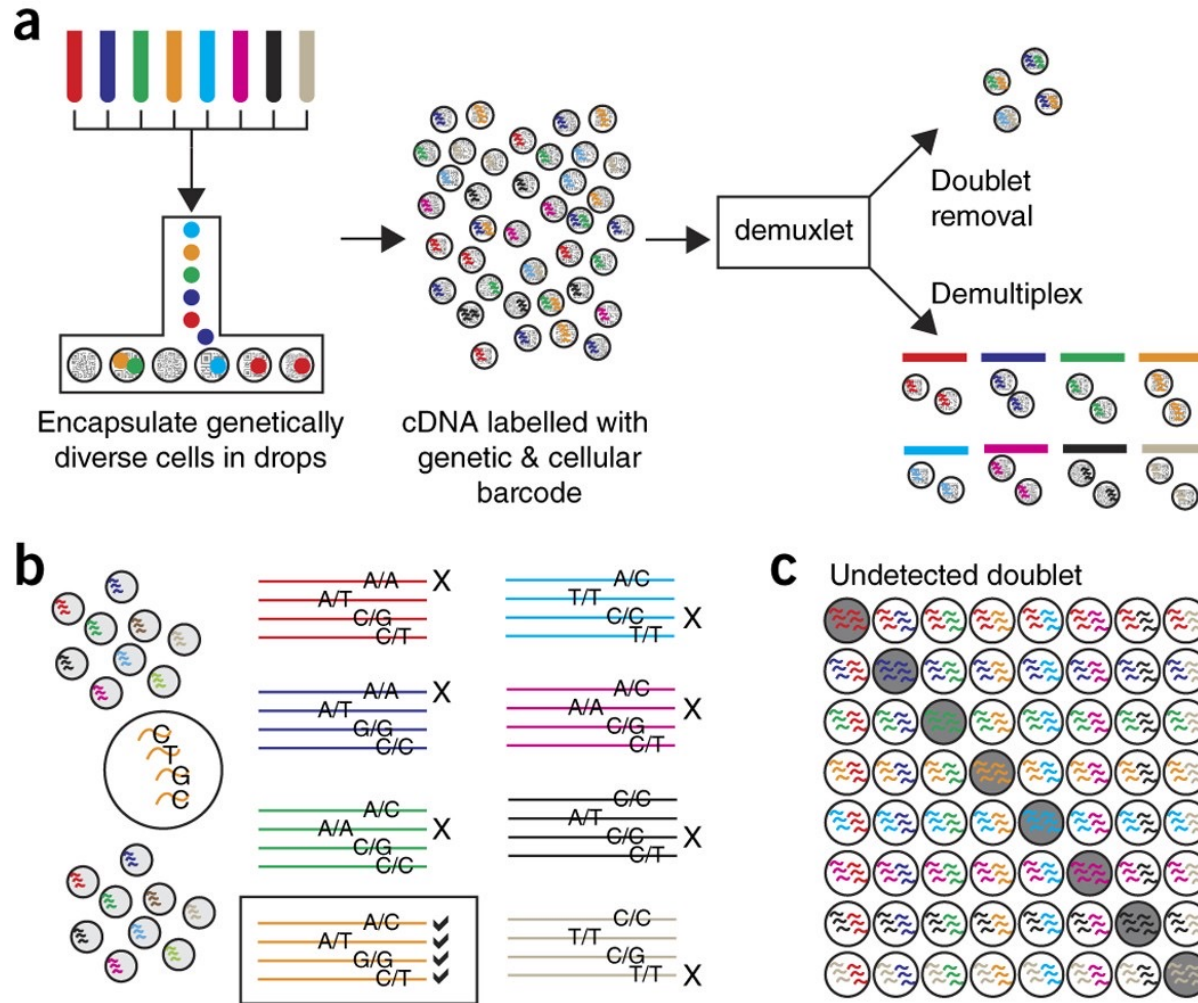
Multiplexing and storage of single cell samples



Multiplexing and storage of single cell samples



Multiplexing - genetic



Xu et al. *Genome Biology* (2019) 20:290
<https://doi.org/10.1186/s13059-019-1852-7>

Genome Biology

METHOD

Open Access

Genotype-free demultiplexing of pooled single-cell RNA-seq



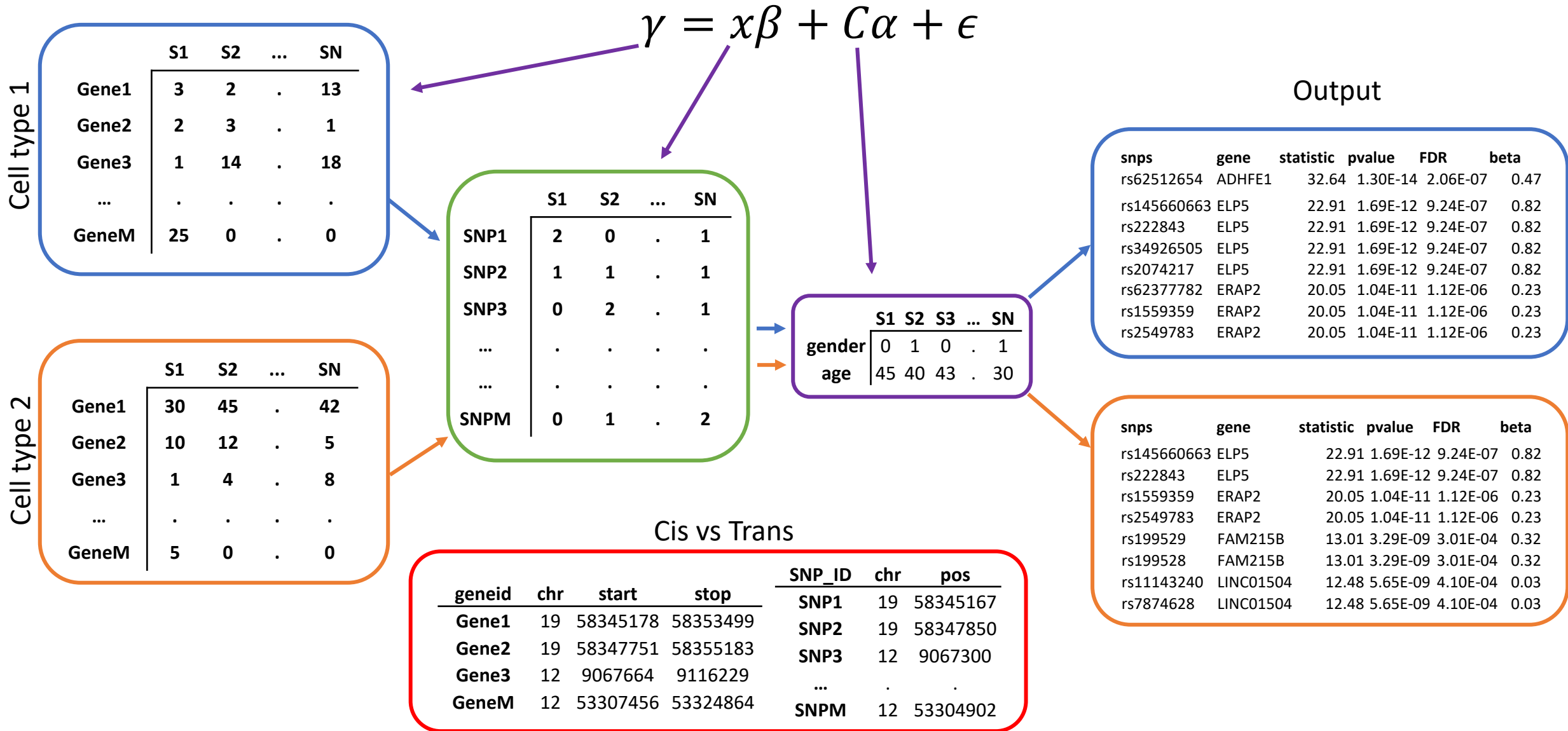
Jun Xu¹, Caitlin Falconer², Quan Nguyen², Joanna Crawford², Brett D. McKinnon^{2,5}, Sally Mortlock², Anne Senabouth⁴, Stacey Andersen^{1,2}, Han Sheng Chiu², Longda Jiang², Nathan J. Palpant^{1,2}, Jian Yang^{2,10}, Michael D. Mueller⁵, Alex W. Hewitt^{7,8,9}, Alice Pébay^{6,7,8}, Grant W. Montgomery^{1,2}, Joseph E. Powell^{3,4} and Lachlan J.M. Coin^{1,2,11,12,13*}

Abstract

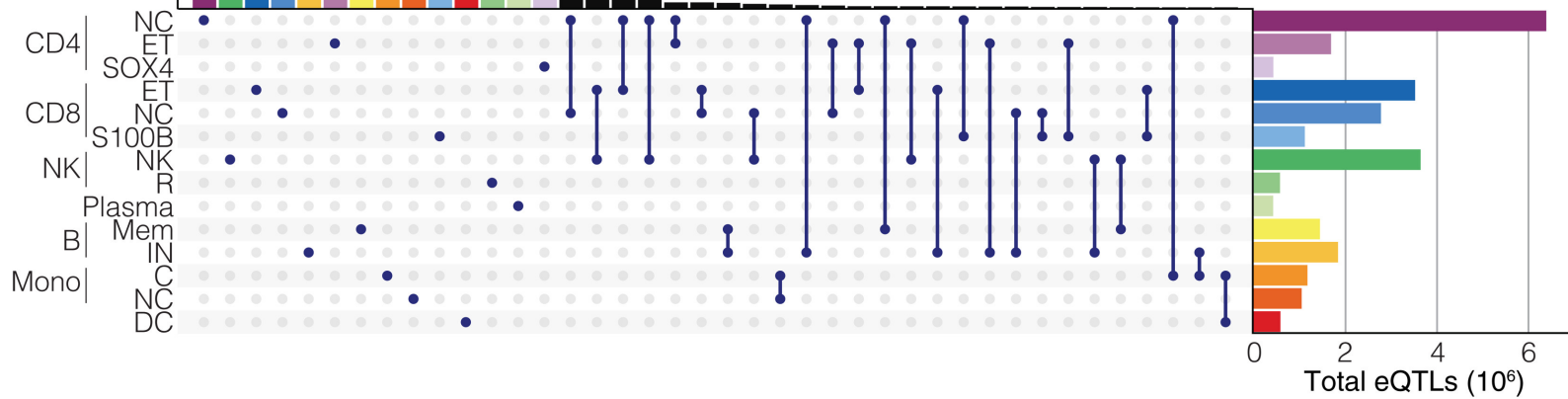
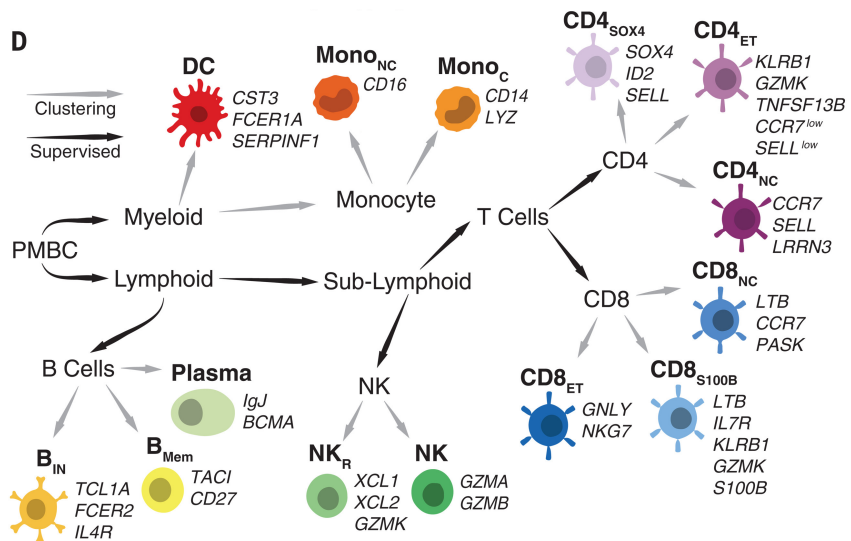
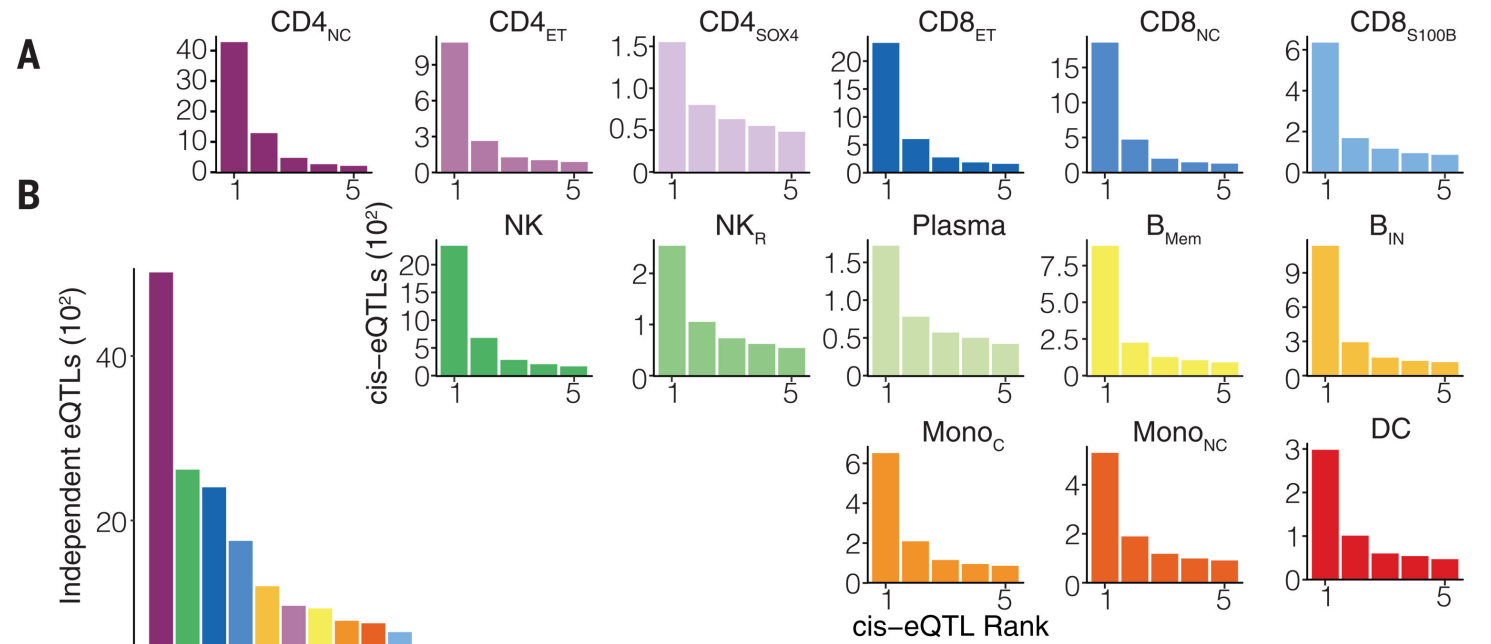
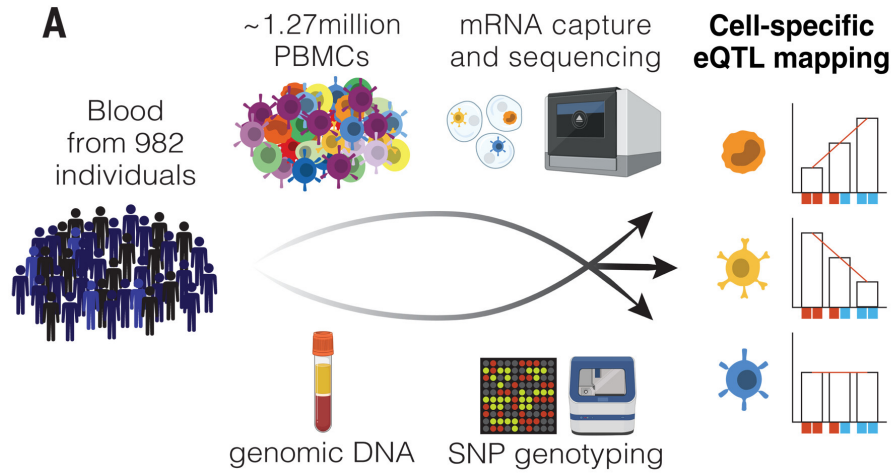
A variety of methods have been developed to demultiplex pooled samples in a single cell RNA sequencing (scRNA-seq) experiment which either require hashtag barcodes or sample genotypes prior to pooling. We introduce scSplit which utilizes genetic differences inferred from scRNA-seq data alone to demultiplex pooled samples. scSplit also enables mapping clusters to original samples. Using simulated, merged, and pooled multi-individual datasets, we show that scSplit prediction is highly concordant with demuxlet predictions and is highly consistent with the known truth in cell-hashing dataset. scSplit is ideally suited to samples without external genotype information and is available at: <https://github.com/jon-xu/scSplit>

Keywords: scSplit, scRNA-seq, Demultiplexing, Machine learning, Unsupervised, Hidden Markov Model, Expectation-maximization, Genotype-free, Allele fraction, Doublets

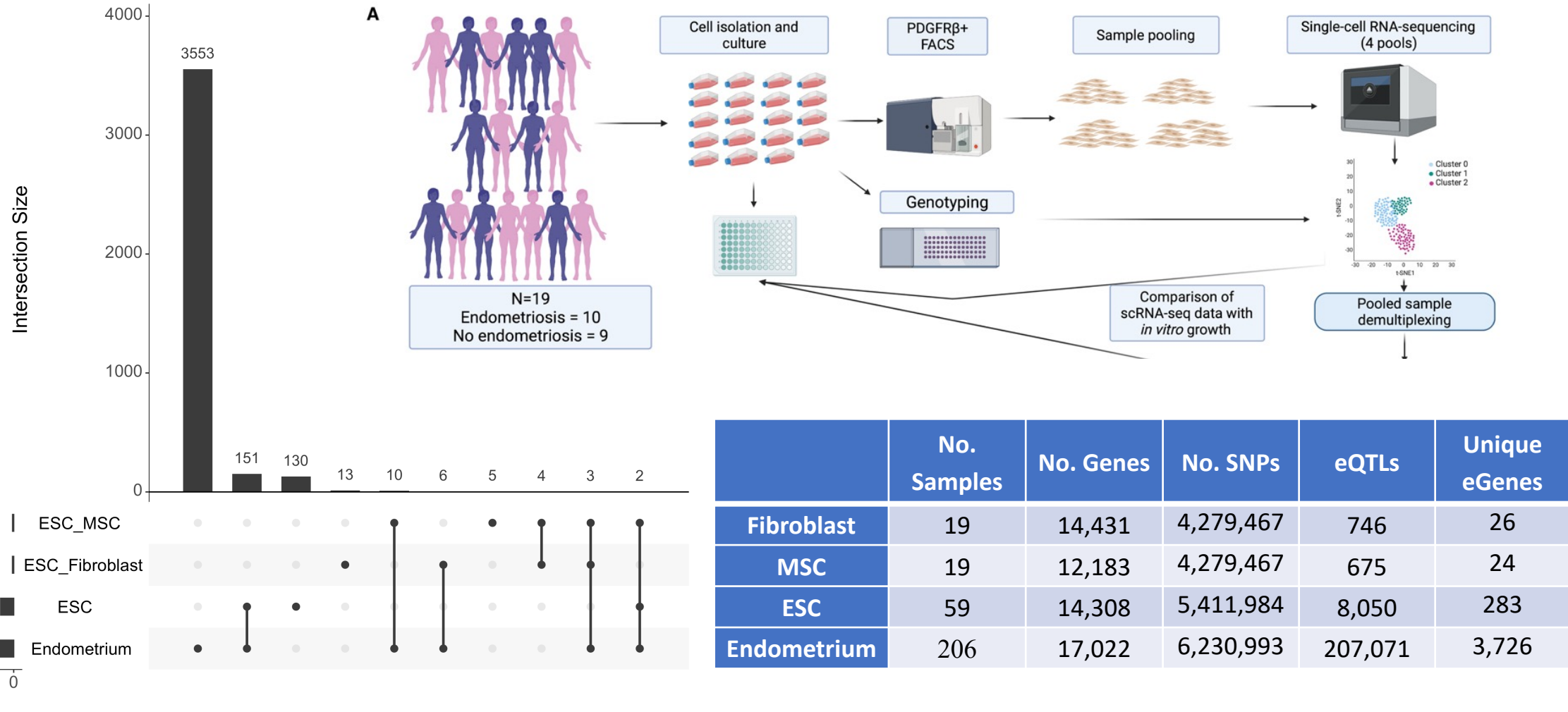
eQTL Analysis



Example – OneK1K



Example - Single-Cell Endometrial eQTLs



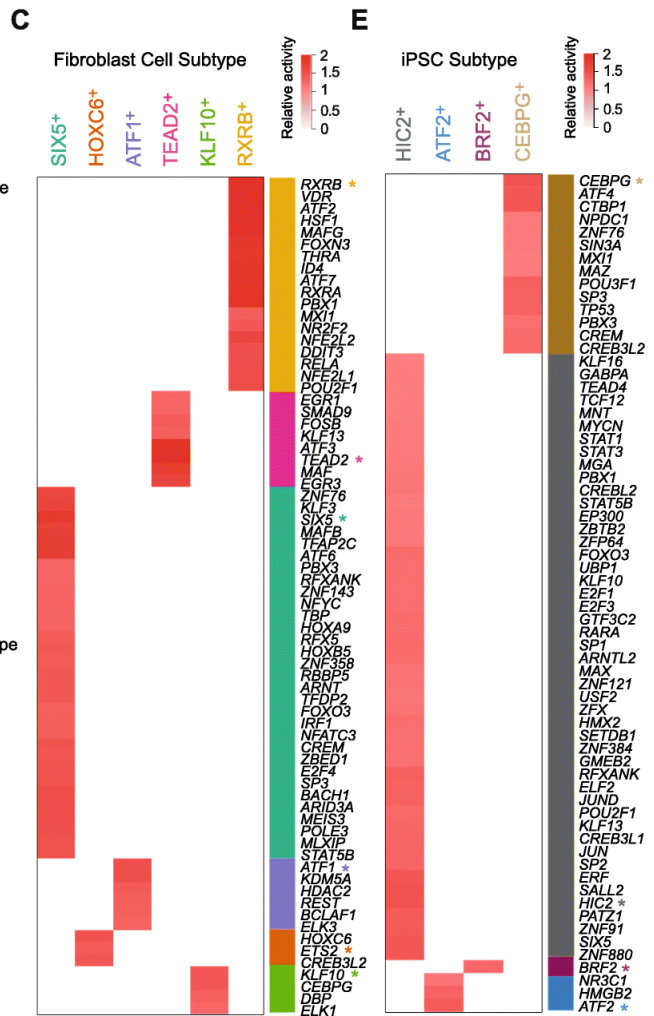
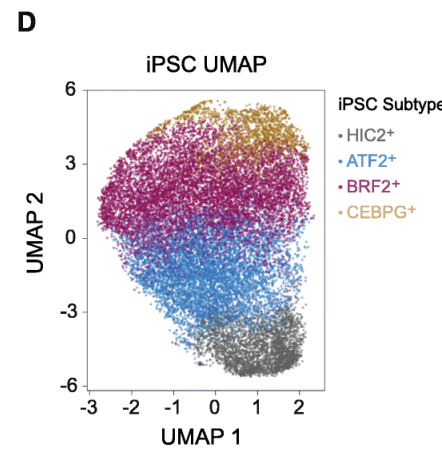
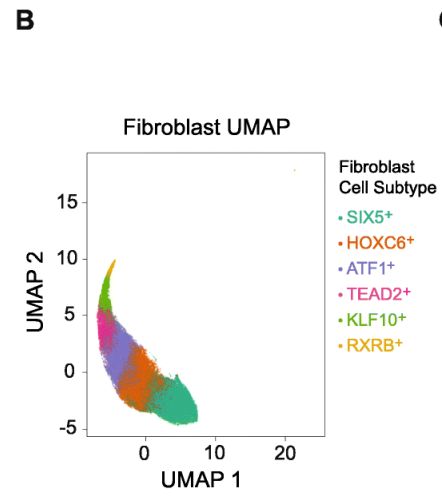
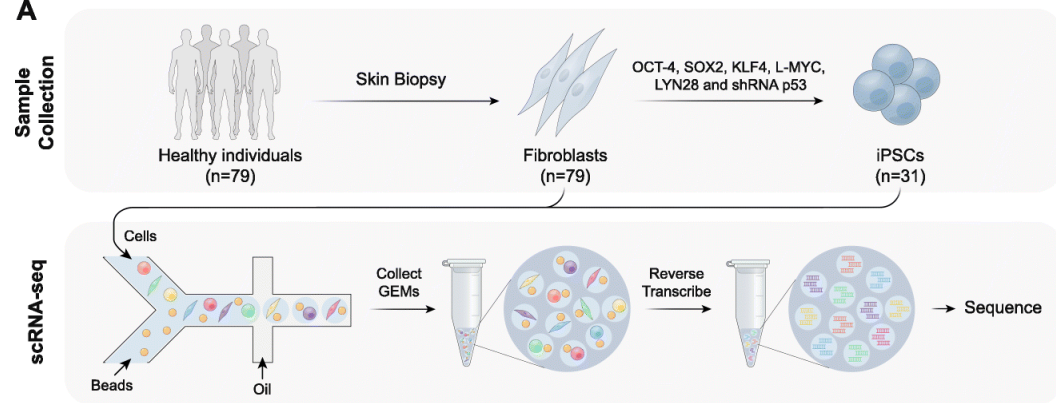
RESEARCH

Open Access

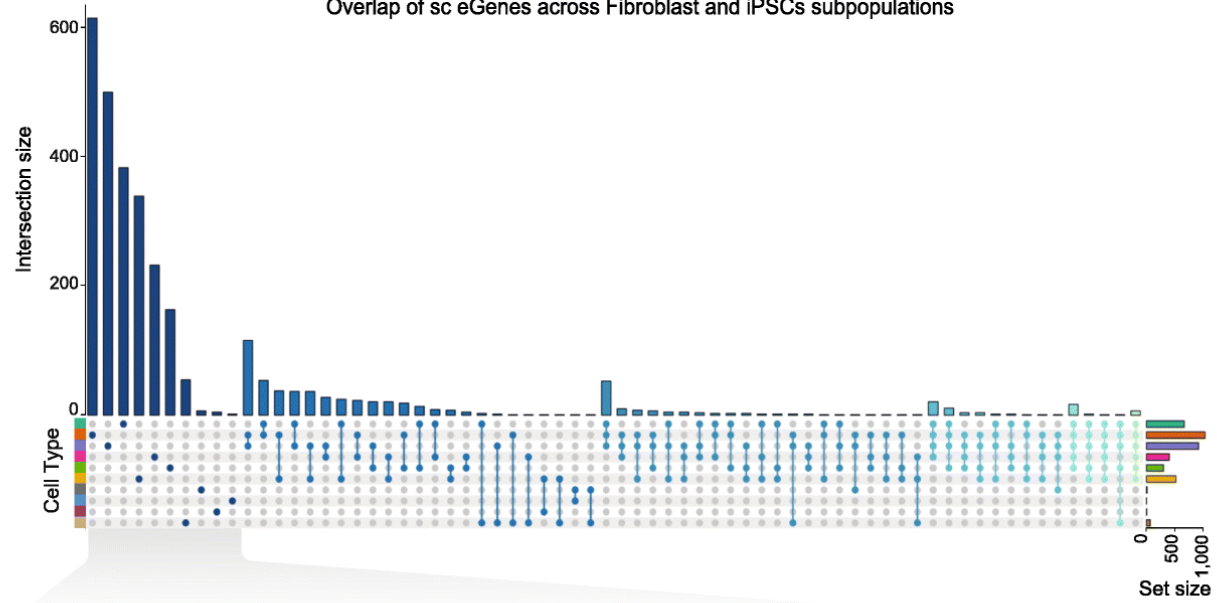
Single cell eQTL analysis identifies cell type-specific genetic control of gene expression in fibroblasts and reprogrammed induced pluripotent stem cells



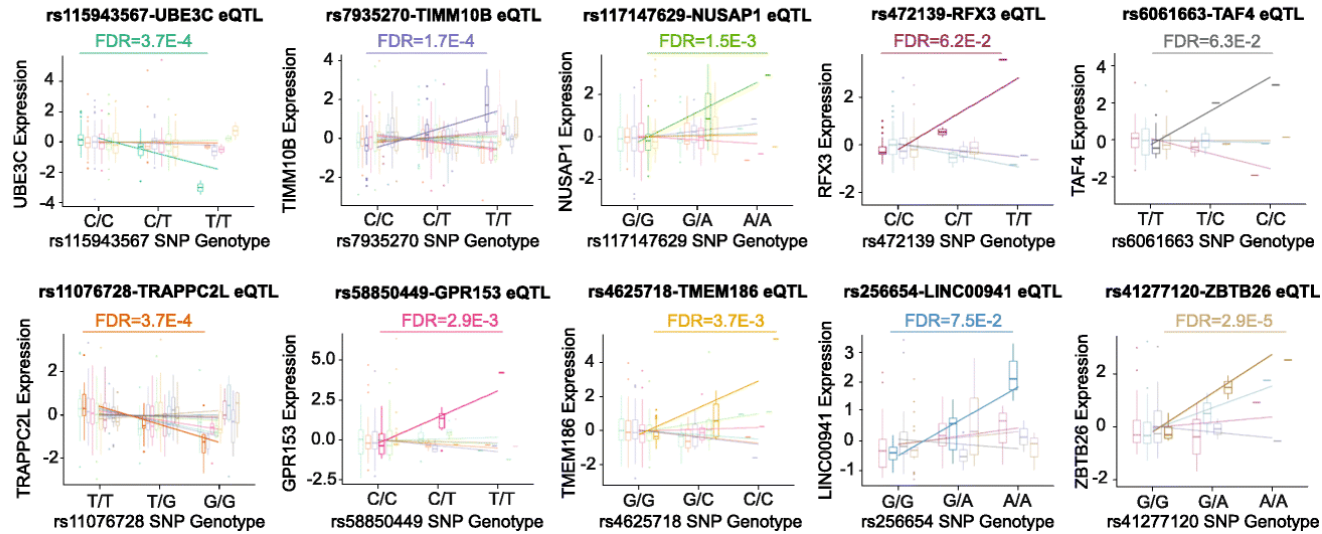
Drew Neavin^{1†}, Quan Nguyen^{2†}, Maciej S. Daniszewski^{3,4,5}, Helena H. Liang^{3,4}, Han Sheng Chiu², Yong Kiat Wee¹, Anne Senabouth¹, Samuel W. Lukowski², Duncan E. Crombie^{3,4}, Grace E. Lidgerwood^{3,4,5}, Damián Hernández^{3,4,5}, James C. Vickers⁶, Anthony L. Cook⁶, Nathan J. Palpant^{2†}, Alice Pébay^{3,4,5†}, Alex W. Hewitt^{3,4,7†} and Joseph E. Powell^{1,8*†}



Overlap of sc eGenes across Fibroblast and iPSCs subpopulations



E



Cell Subtype Key

Fibroblast
iPSC

SIX5⁺
HIC2⁺

HOXC6⁺
ATF2

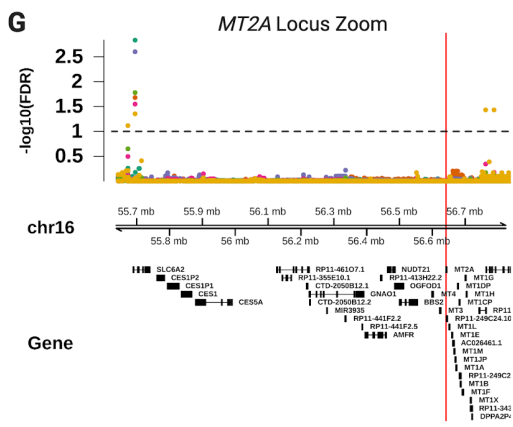
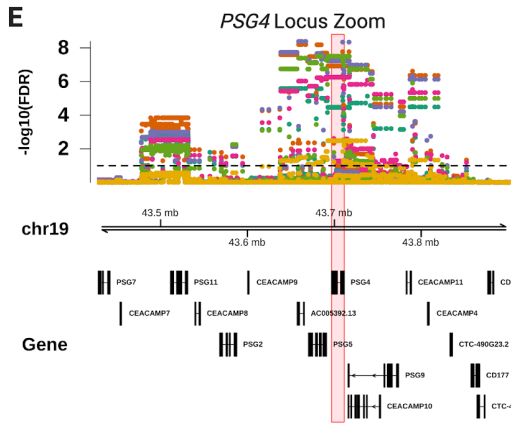
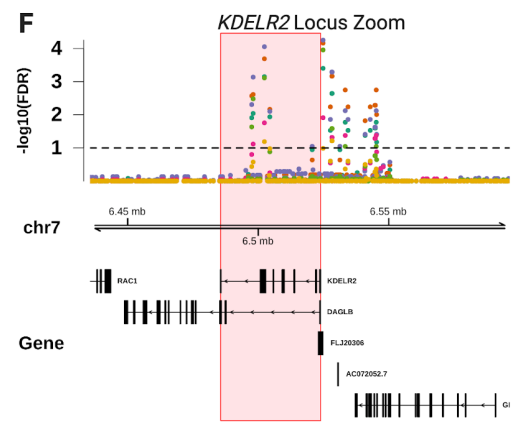
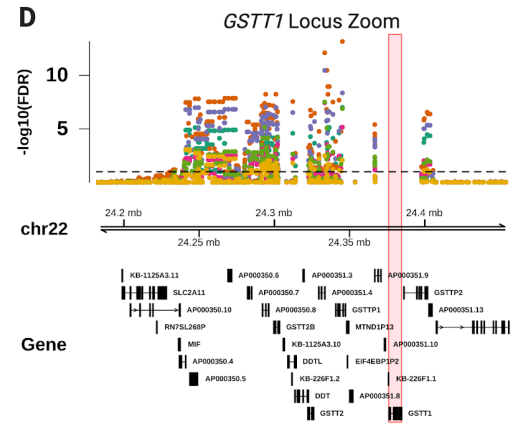
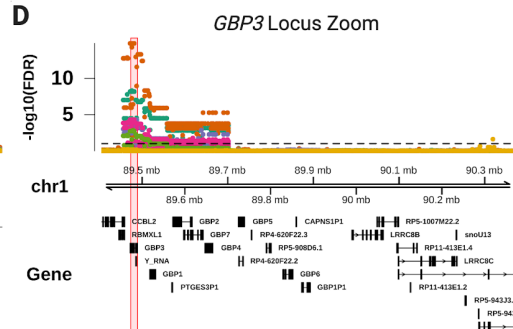
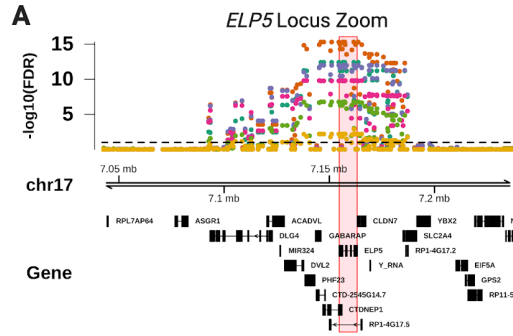
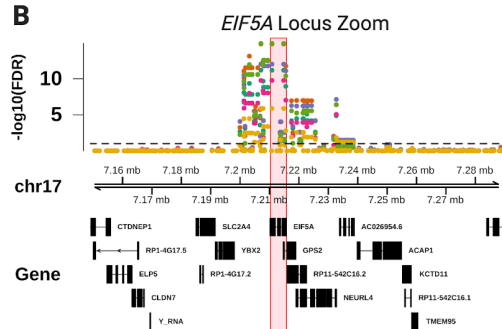
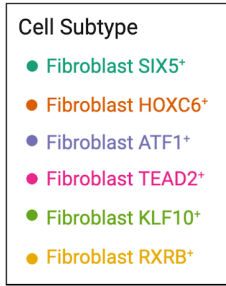
ATF1⁺
BRF2

TEAD2⁺
CEBPG⁺

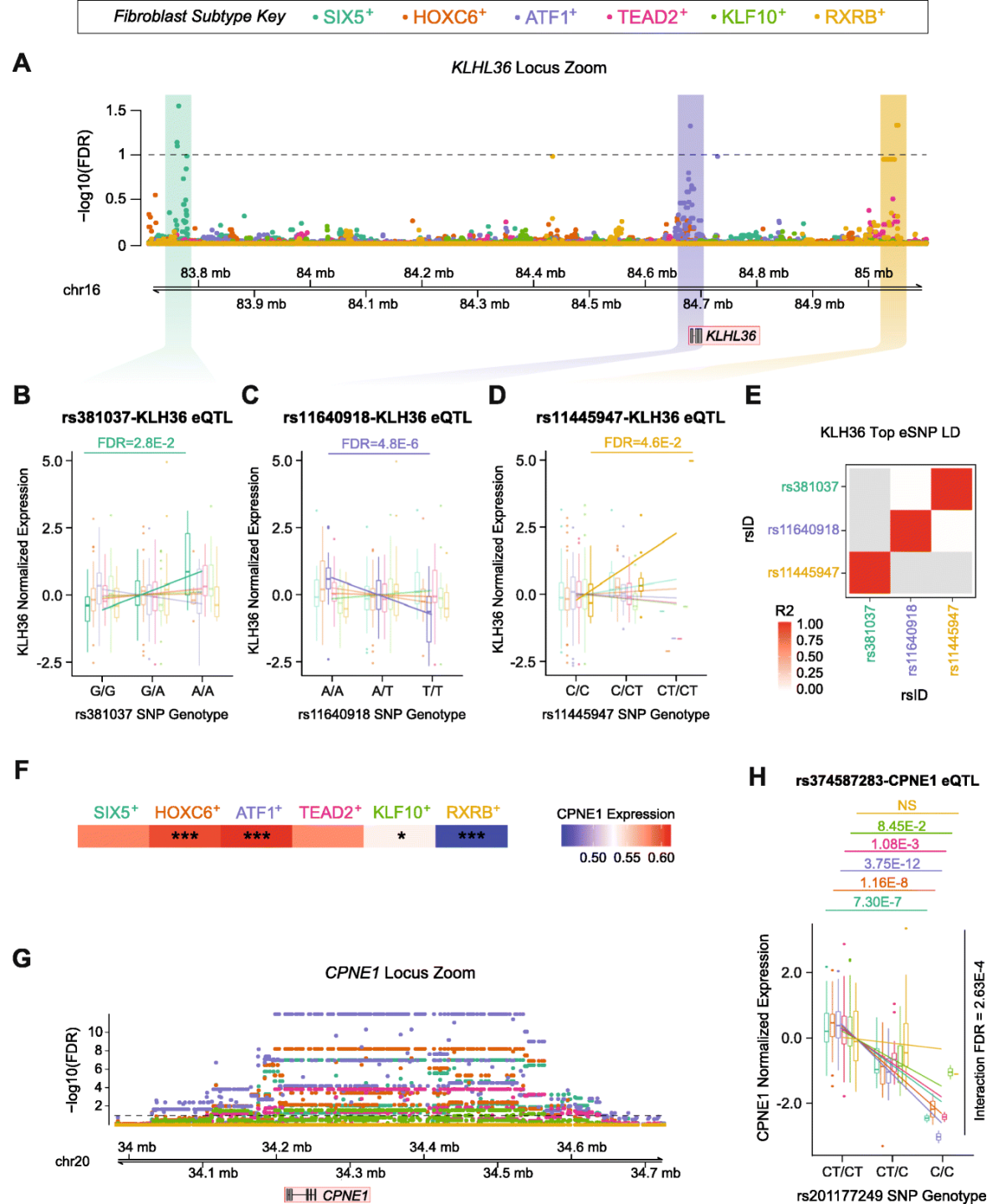
KLF10⁺

RXRβ⁺

eGenes that were significant in all six fibroblast subtypes

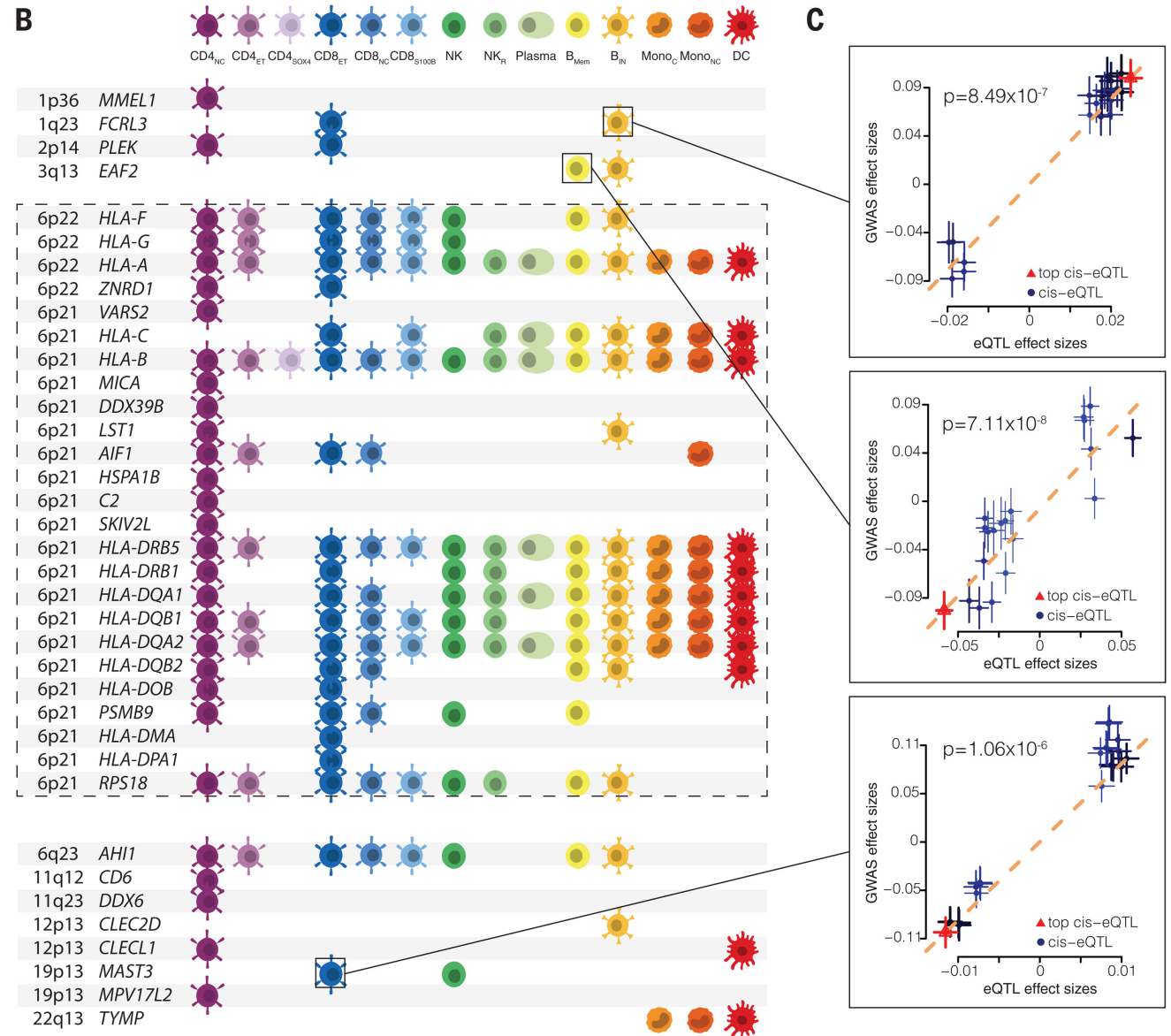


Single cell eQTL analysis identifies cell type-specific genetic control of gene expression in fibroblasts and reprogrammed induced pluripotent stem cells

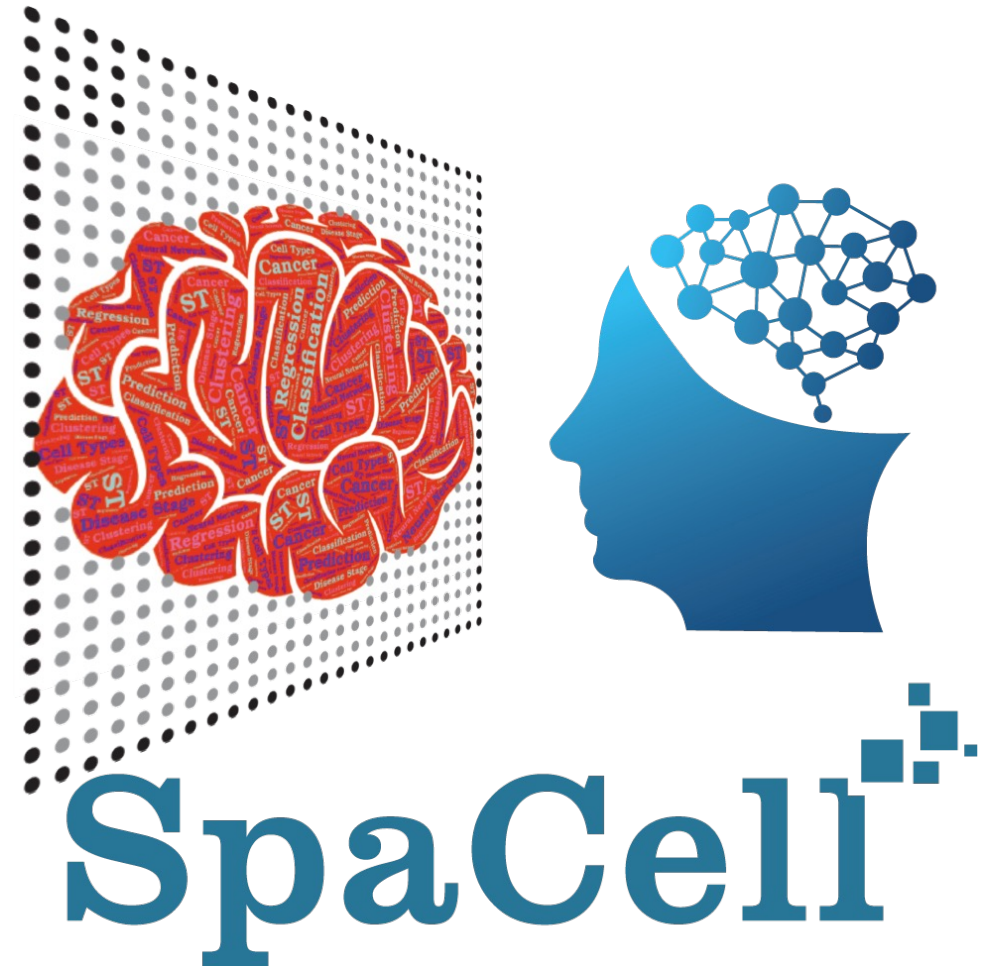
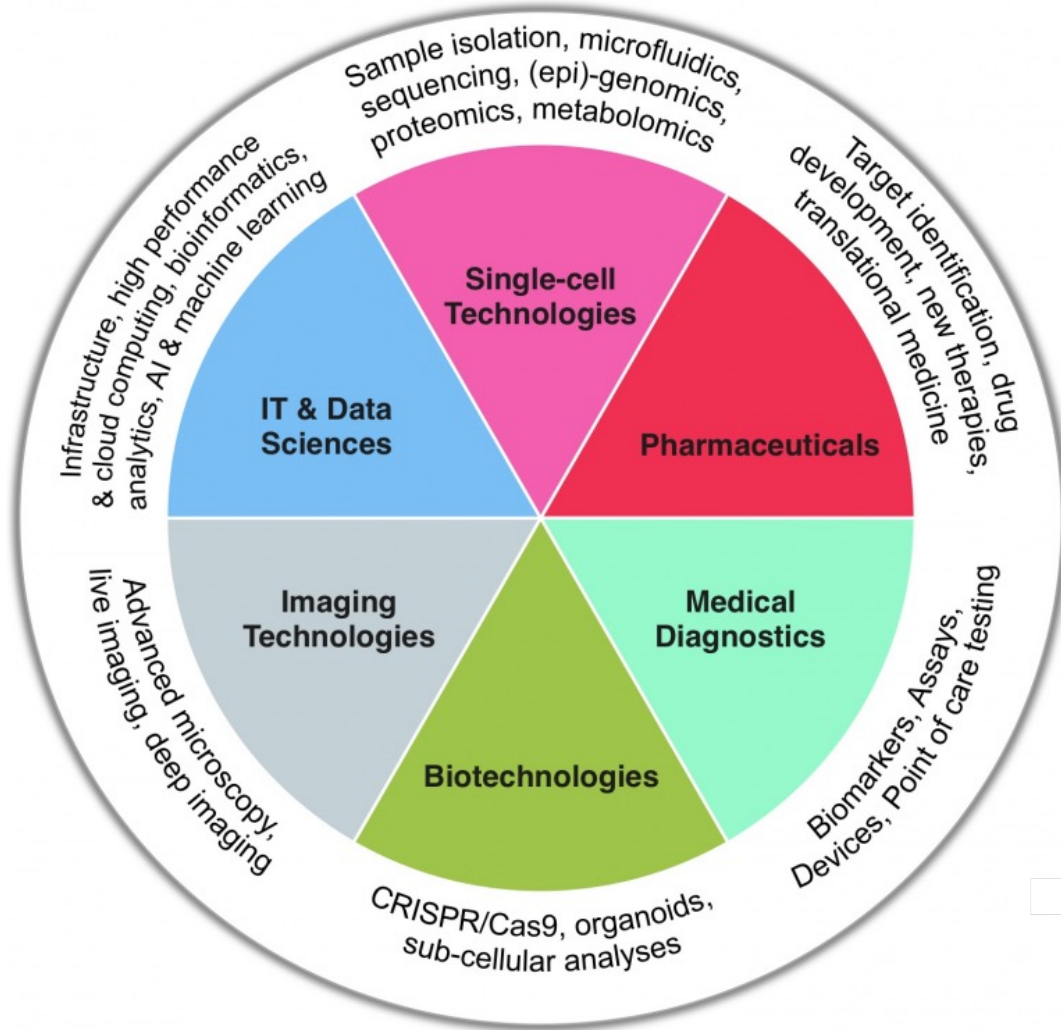


Multiple sclerosis example

- Identified overlapping cis-eQTL for 108 risk genes using coloc.
- Of the 108 genes, 69 show eQTL overlap in just a single cell type.
- 39 genes identified using SMR.



Spatial transcriptomics and Machine learning

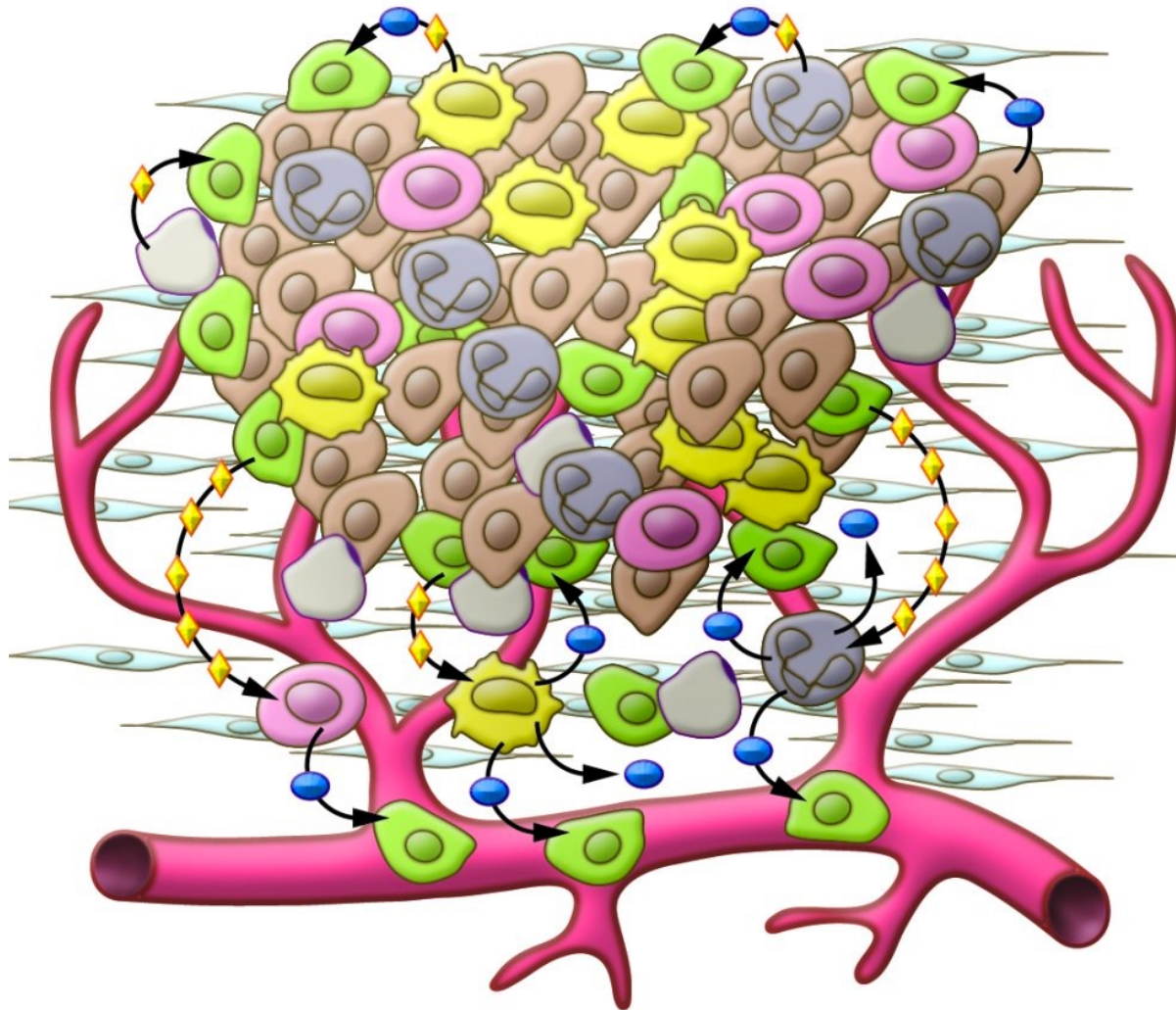


The G&G Cellomics Team

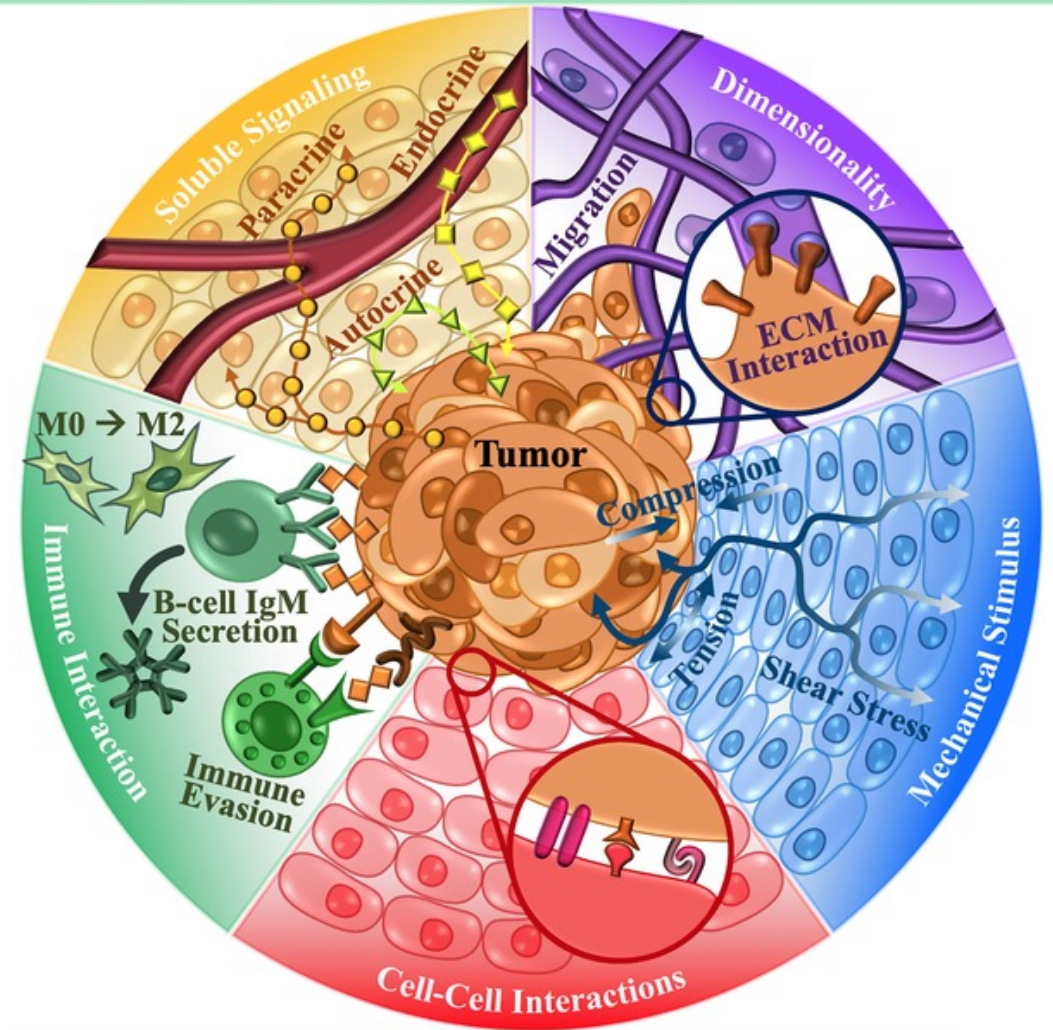
Quan Nguyen, Guiyan Ni, Sally Mortlock, Duy Pham, Xiao Tan

Introduction to spatial transcriptomics

Cancer in a native tissue



(Korkaya et al, 2011)



(Bregenzler et al, 2019)

- Cell-type composition and organisation and cell-cell interactions are important
- Complex in vivo processes have direct effects on or are the consequences of transcriptional regulation

Spatial transcriptomics approach

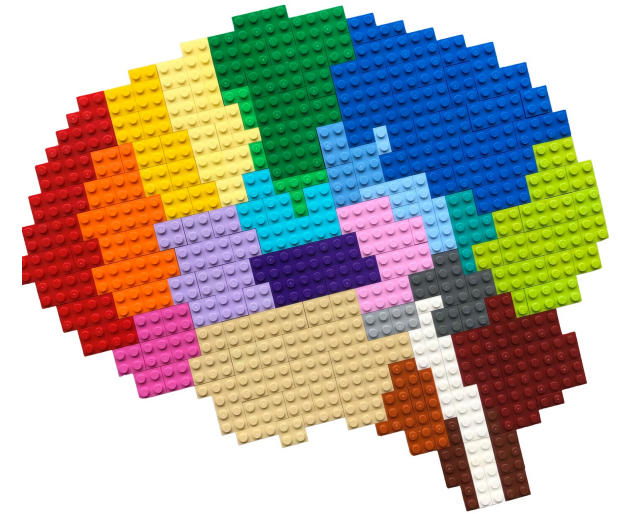
Bulk



Single cell

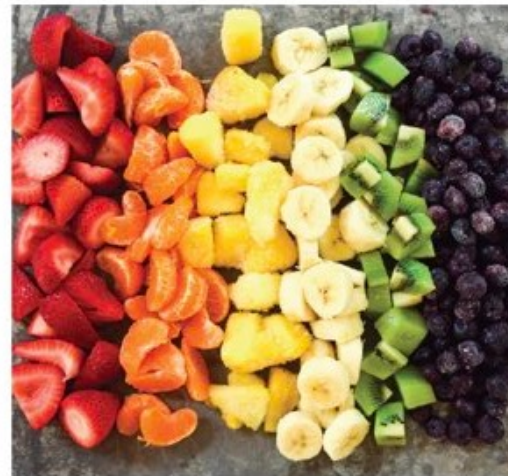


Spatial

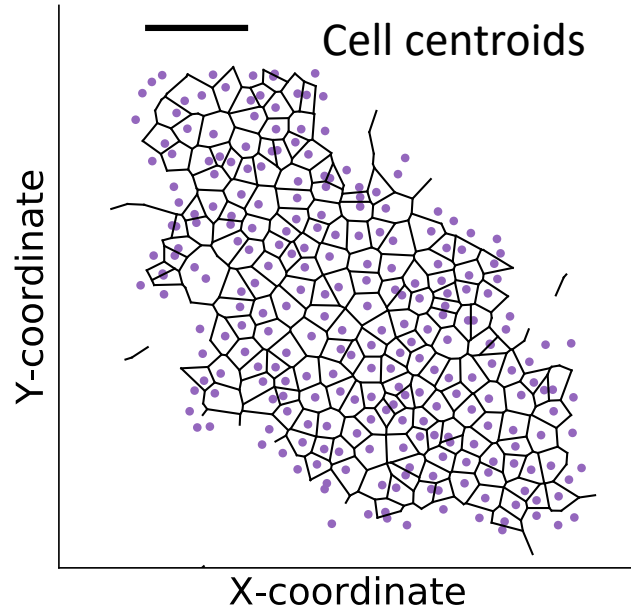


Lego:
(@boxia)

Fruit salad:
(@LGMartelotto)



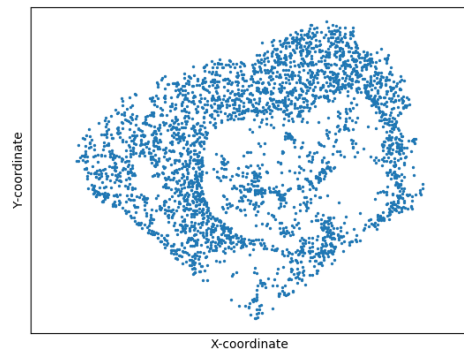
Spatial Transcriptomics Data (seqFISH): expression + location



(2050 cells and ~10,000 genes)

Field of View	Cell ID	X	Y	Aanat	Aasdh	Aatf	Abat	Abca16	Abca17	...
0	0	1	1766.40	283.42	0	0	2	0	0	0 ...
1	0	2	1891.40	348.38	0	0	0	0	2	0 ...
2	0	3	1548.70	351.11	0	0	0	0	0	0 ...
3	0	4	1657.60	357.37	0	0	0	2	0	0 ...
4	0	5	1767.40	392.22	0	0	0	0	0	0 ...

Fluorescence single molecule counts

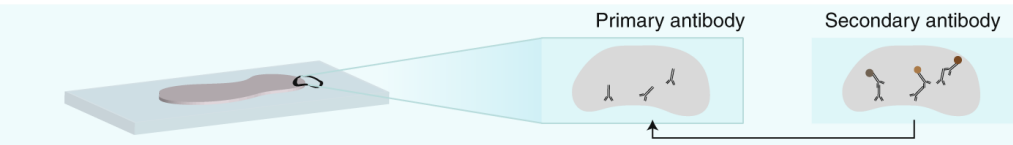
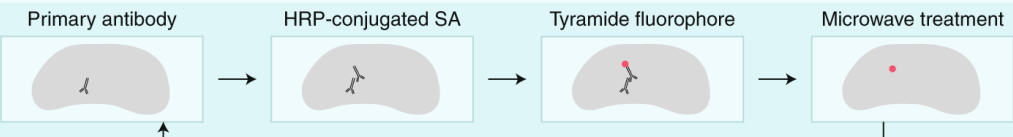
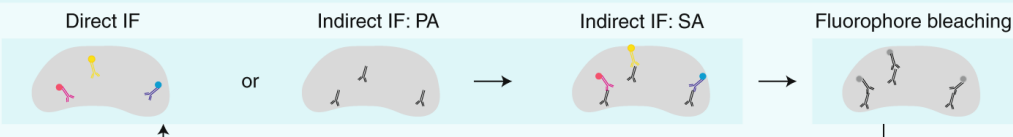
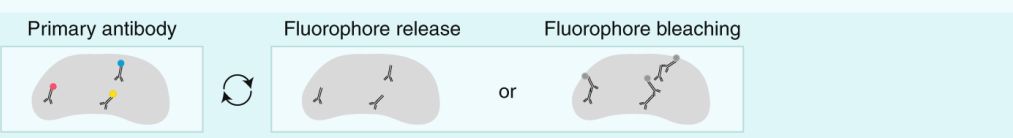
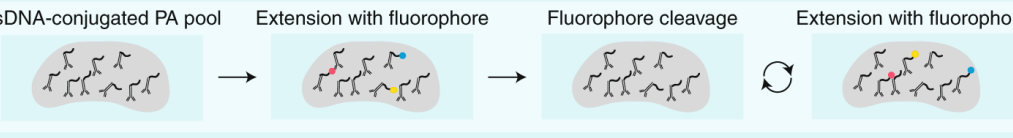
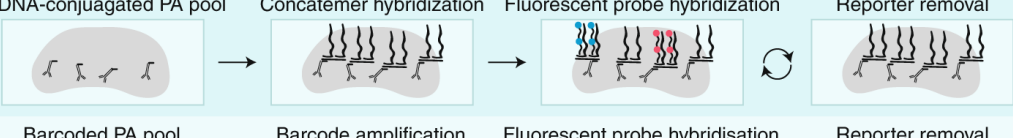






Example of seqFISH RNA in a cell: 3247 genes

Gene ID	1	19	23	44	53	57	63	70	71	72	...
0	653.00	675.24	687.21	733.85	615.16	663.99	611.06	669.65	638.03	601.10	...
1	434.34	428.89	479.06	472.43	469.95	464.81	443.74	417.42	430.46	472.07	...

Coordinates









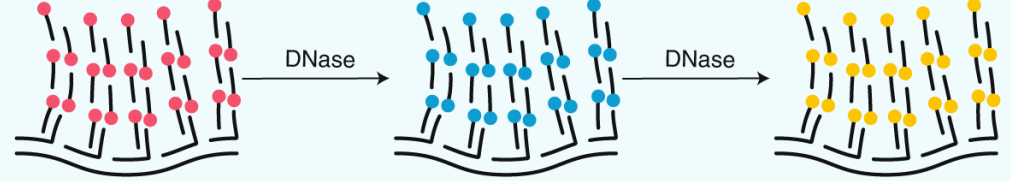



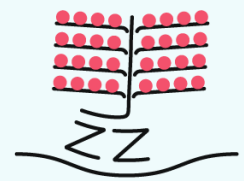
Spatial proteomics

				No. of targets	Tissue prep.			
Iterative	mIHC		Primary antibody	Secondary antibody	30	FFPE		
	OPAL		Primary antibody	HRP-conjugated SA	Tyramide fluorophore	Microwave treatment	10	FFPE
Iterative (fluorescence)	CyclIF		Direct IF	or Indirect IF: PA	Indirect IF: SA	Fluorophore bleaching	60	FFPE
	REAdye_release and REAfinity		Primary antibody	Fluorophore release	or Fluorophore bleaching		100 (400)	FFPE
Iterative (fluorescence)	CODEX		dsDNA-conjugated PA pool	Extension with fluorophore	Fluorophore cleavage	Extension with fluorophore	60	FF* FFPE
	Immuno-SABER		ssDNA-conjugated PA pool	Concatemer hybridization	Fluorescent probe hybridization	Reporter removal	10 (50)	Whole-mount FF* FFPE
	InSituPlex		Barcoded PA pool	Barcode amplification	Fluorescent probe hybridisation	Reporter removal	10	FFPE
TOF-mass pectrometry	IMC		Metal-conjugated PA pool	UV laser ablation	TOF mass spectrometry		40 (100)	FF FFPE
	MIBI		Metal-conjugated PA pool	Ion beam gun	TOF mass spectrometry		40 (100)	FF FFPE
Sequencing	DSP		Stain + oligonucleotide-conjugated PA pool	Oligonucleotide cleavage	Quantitative analysis		44 (100)	FF* FFPE

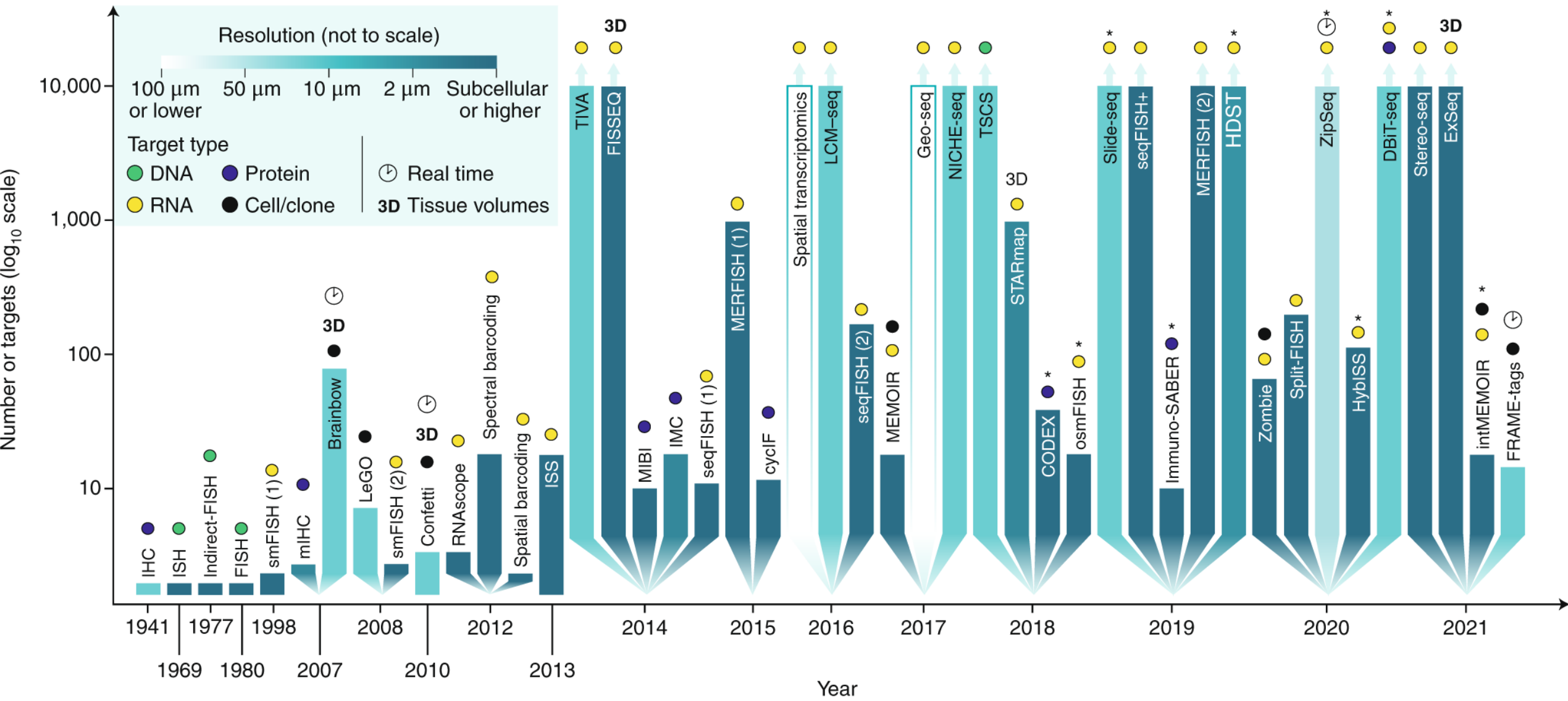
Spatial transcriptomics (sequencing)

				No. of targets	Tissue prep.	
LCM based (e.g., LCM-seq)	Image, laser capture	Tissue digestion, mRNA collection, cDNA synthesis	Sequence cDNA	10,000+	FF	
mRNA capture (e.g., spatial transcriptomics)	Stain, image	Permeabilize tissue, mRNA capture, in situ cDNA synthesis	Sequence cDNA	10,000+	FF FFPE	
Microfluidics based (e.g. DBIT-seq)	Permeabilize tissue, microfluidic barcoding, image	In situ cDNA synthesis	Sequence cDNA	10,000+	FF FFPE	
	Round 1	Round 2	Round <i>n</i>	Barcode		
ISS					31 (256)	FF FFPE
FISSEQ					10,000+	FF FFPE

Spatial transcriptomics (FISH)

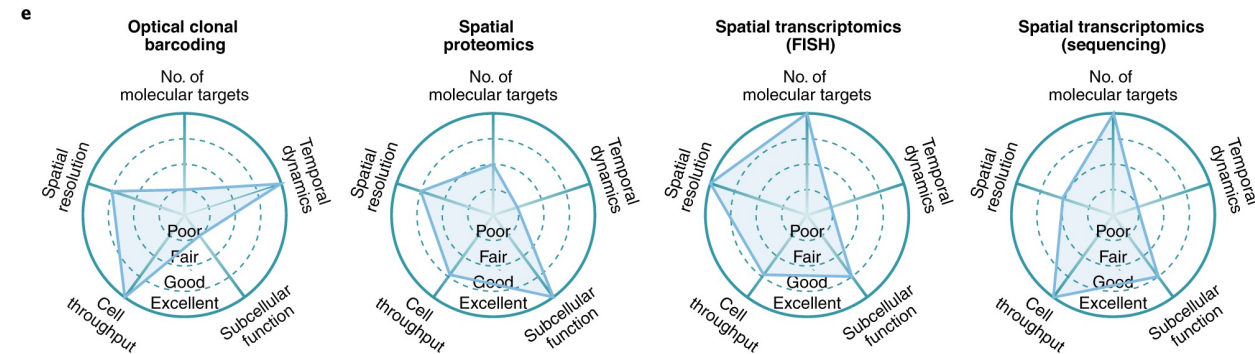
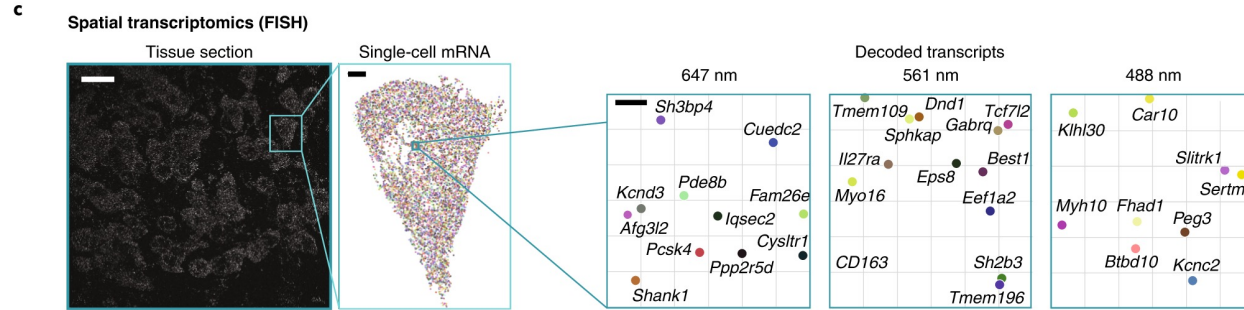
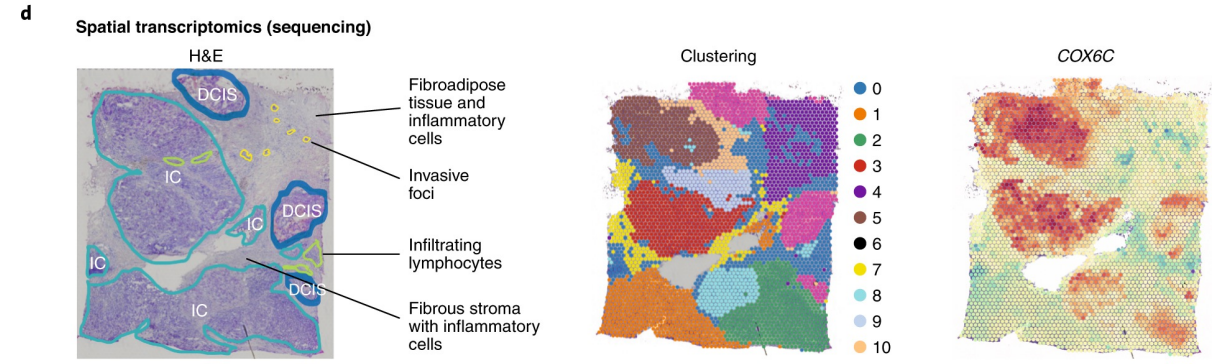
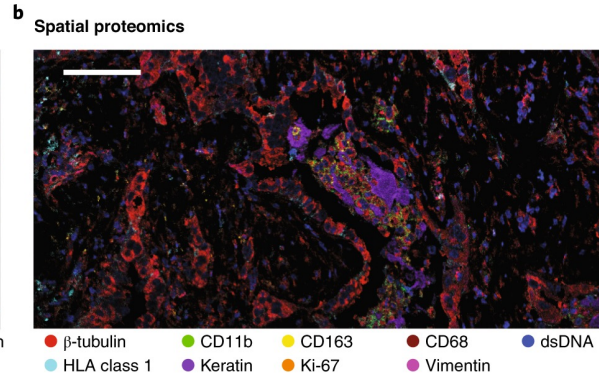
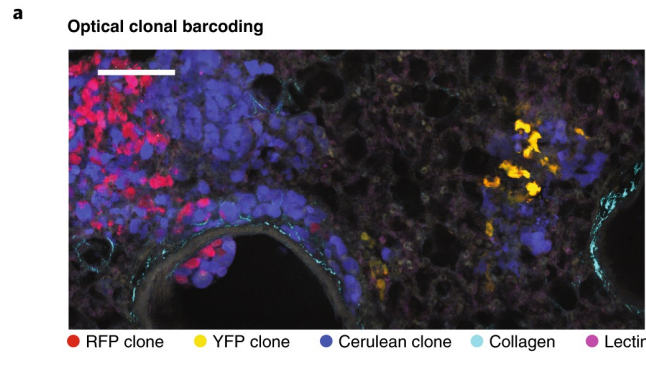
		Barcode	No. of targets	Tissue prep.
smFISH		NA	<10	FF FFPE
Spectral barcoding			32 (792)	NA
Spatial barcoding			<10	NA
	Round 1 Round 2 Round <i>n</i>			
osmFISH		NA	33	FF
MERFISH			10,000	FF
seqFISH			249	FF
seqFISH+			10,000	FF
RNAscope		NA	12	FF FFPE

Rapid technology development – Opportunities and challenges

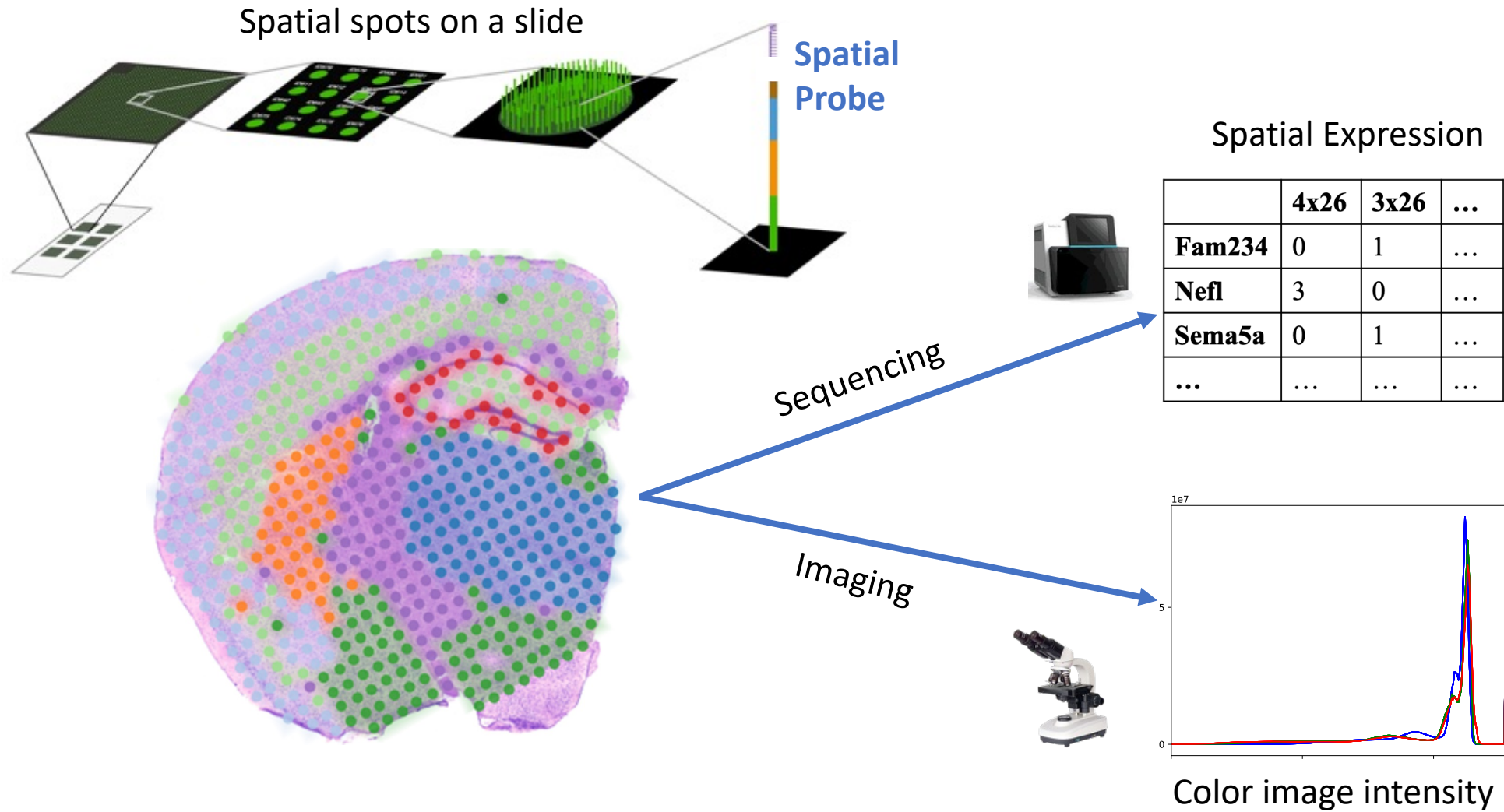


(Lewis et al., 2021, Nat Methods)

Rapid technology development – Opportunities and challenges

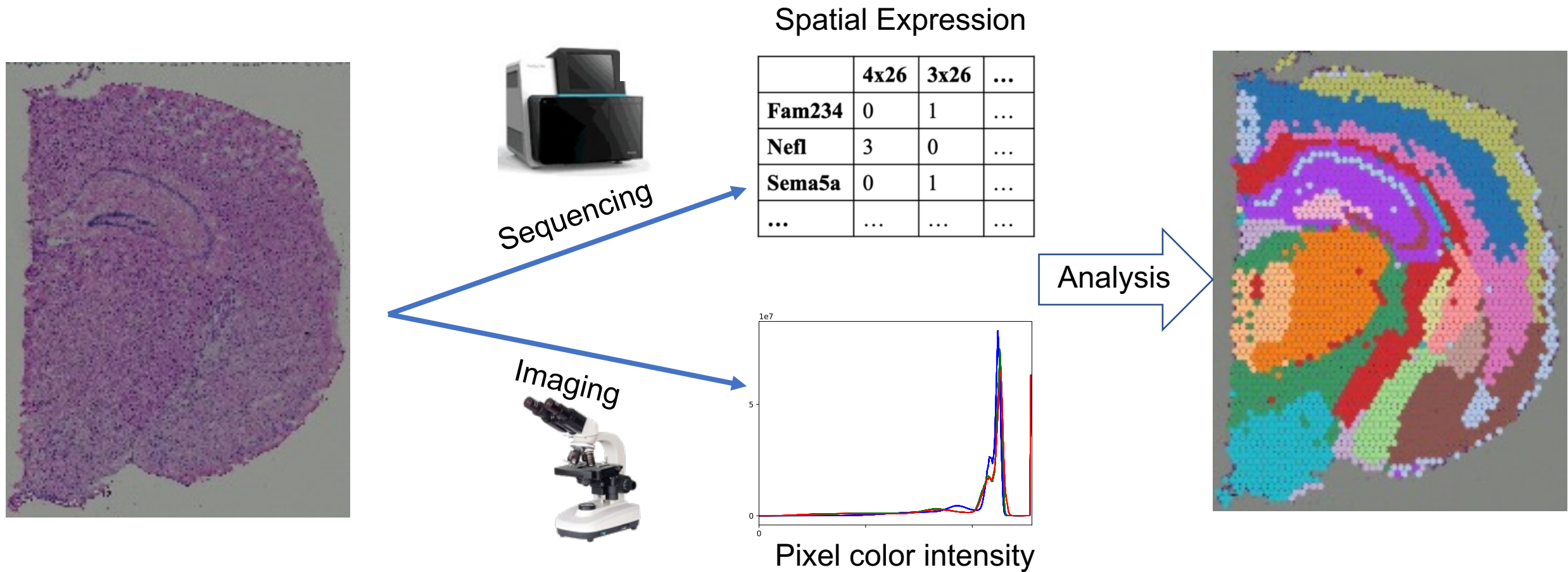


Spatial transcriptomics adds spatial dimension and tissue morphology



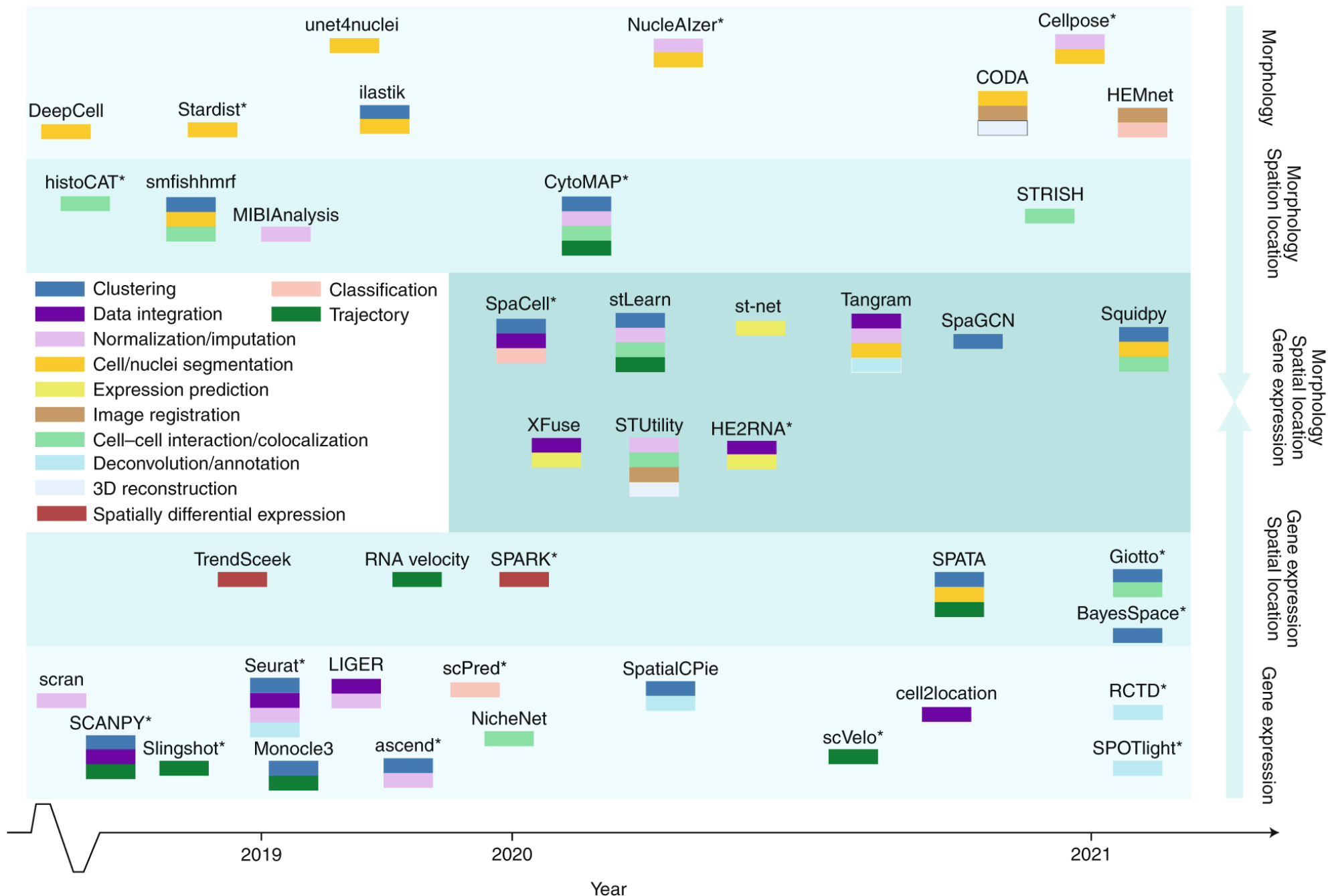
- On-tissue expression profiling (>20,000 genes); each spot contains ~1-9 cells; tissue < 6.5 mm x 6.5 mm
- Other spatial technologies are different (complementary) in resolution, throughput, scale, sensitivity ect.

Spatial transcriptomics captures tissue morphology and spatially-resolved transcriptome

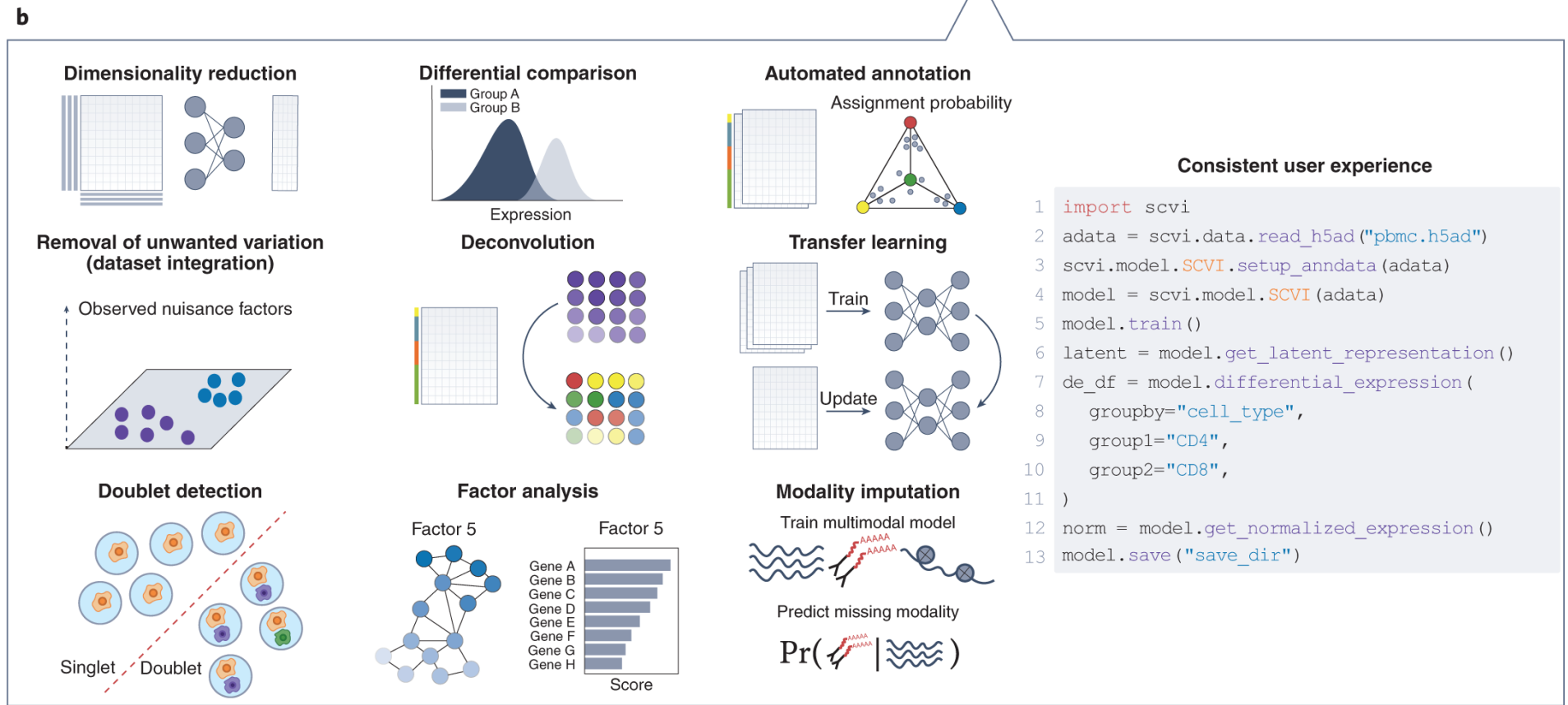
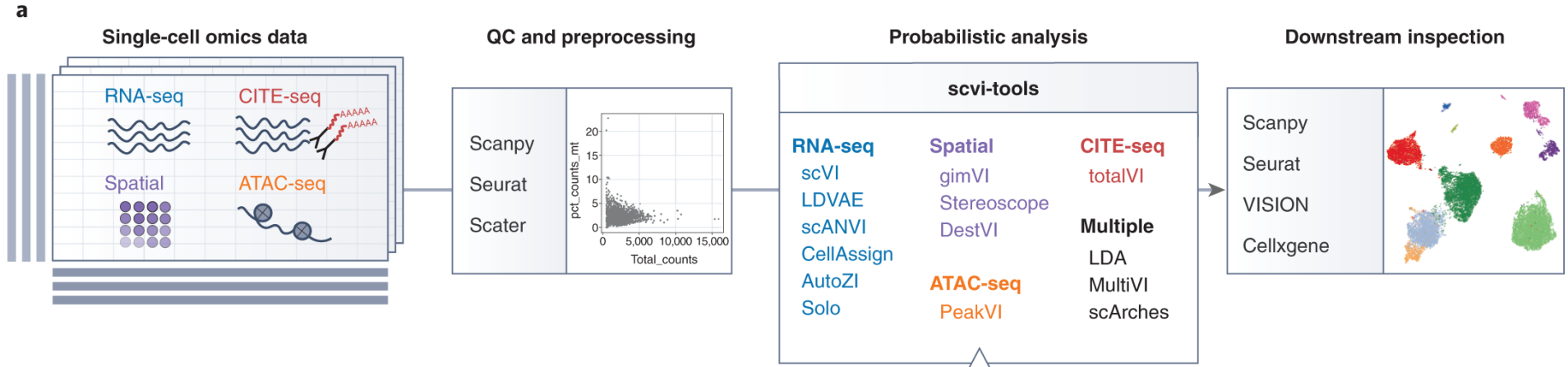


- On-tissue expression profiling (>20,000 genes); each spot contains ~1-9 cells; tissue < 6.5 mm x 6.5 mm
- Other spatial technologies are different (complementary) in resolution, throughput, scale, sensitivity ect.

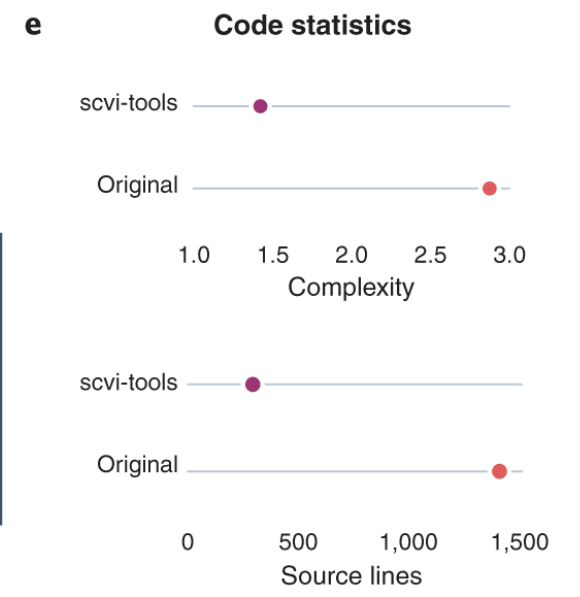
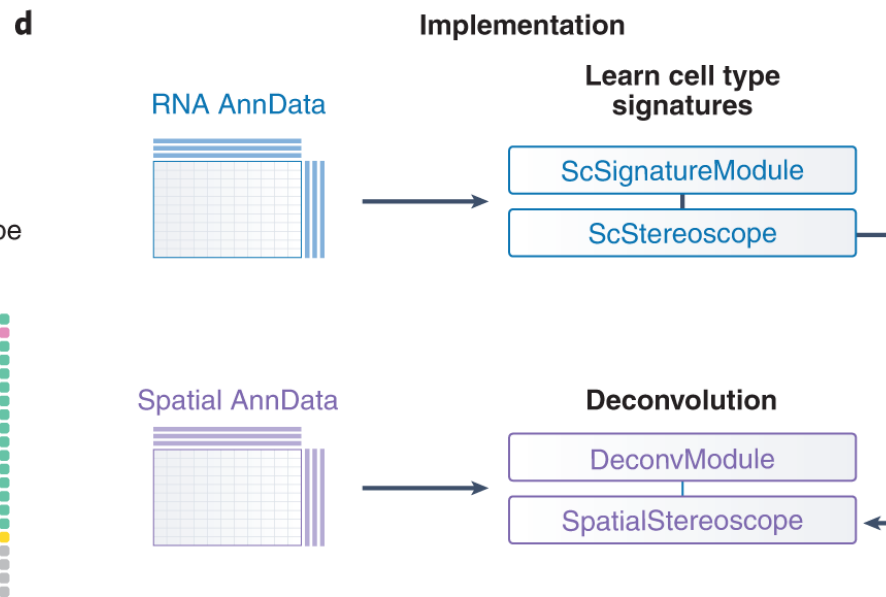
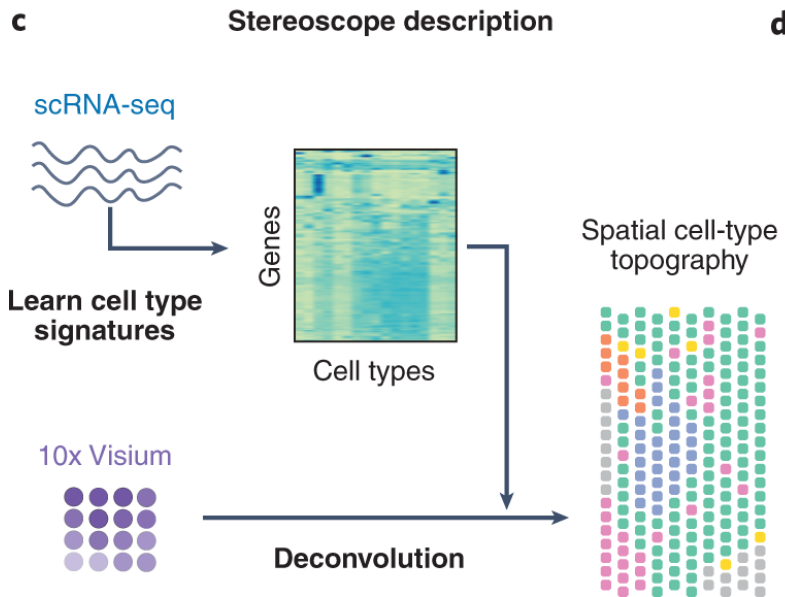
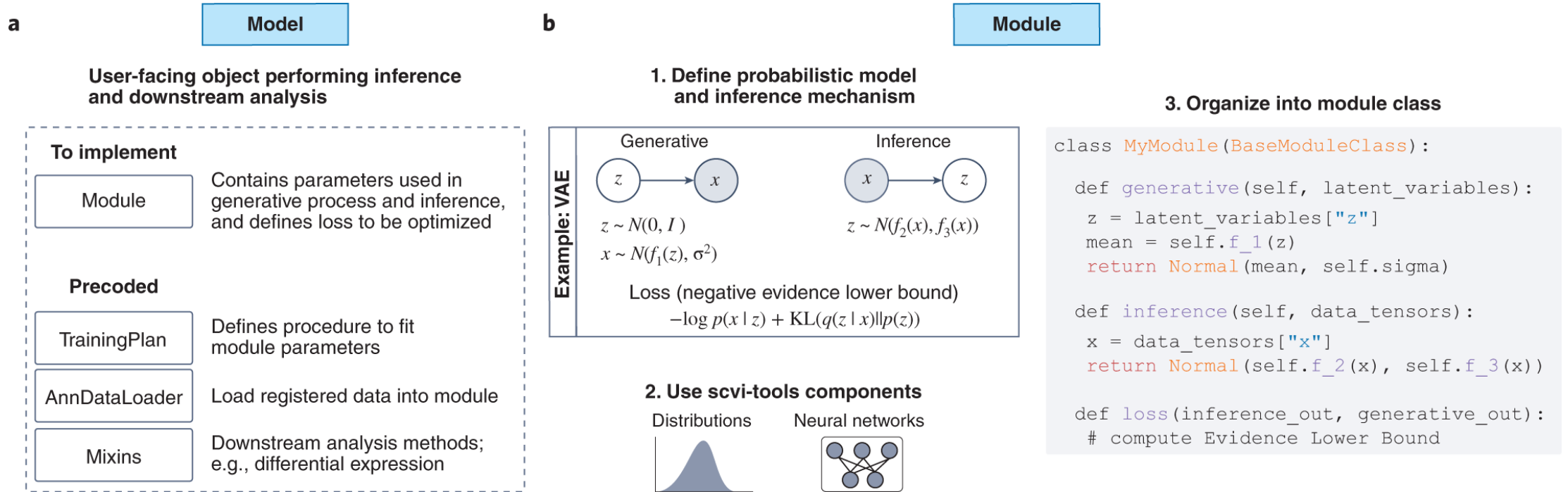
Computational Analysis



Probabilistic Approach



Probabilistic Approach



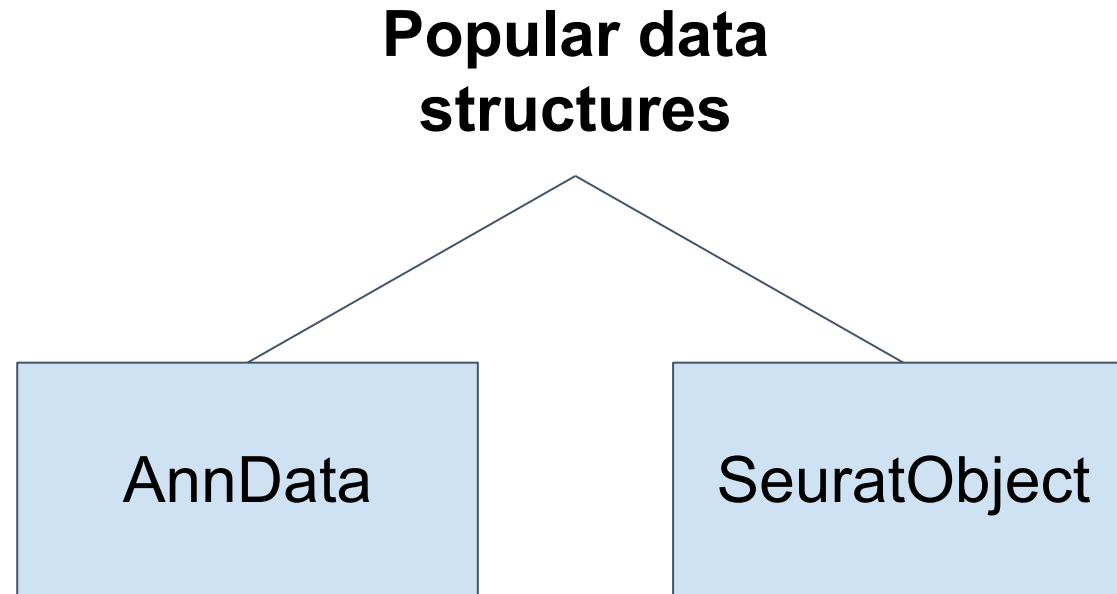
Data structure of scRNAseq and Spatial transcriptomics

Definition

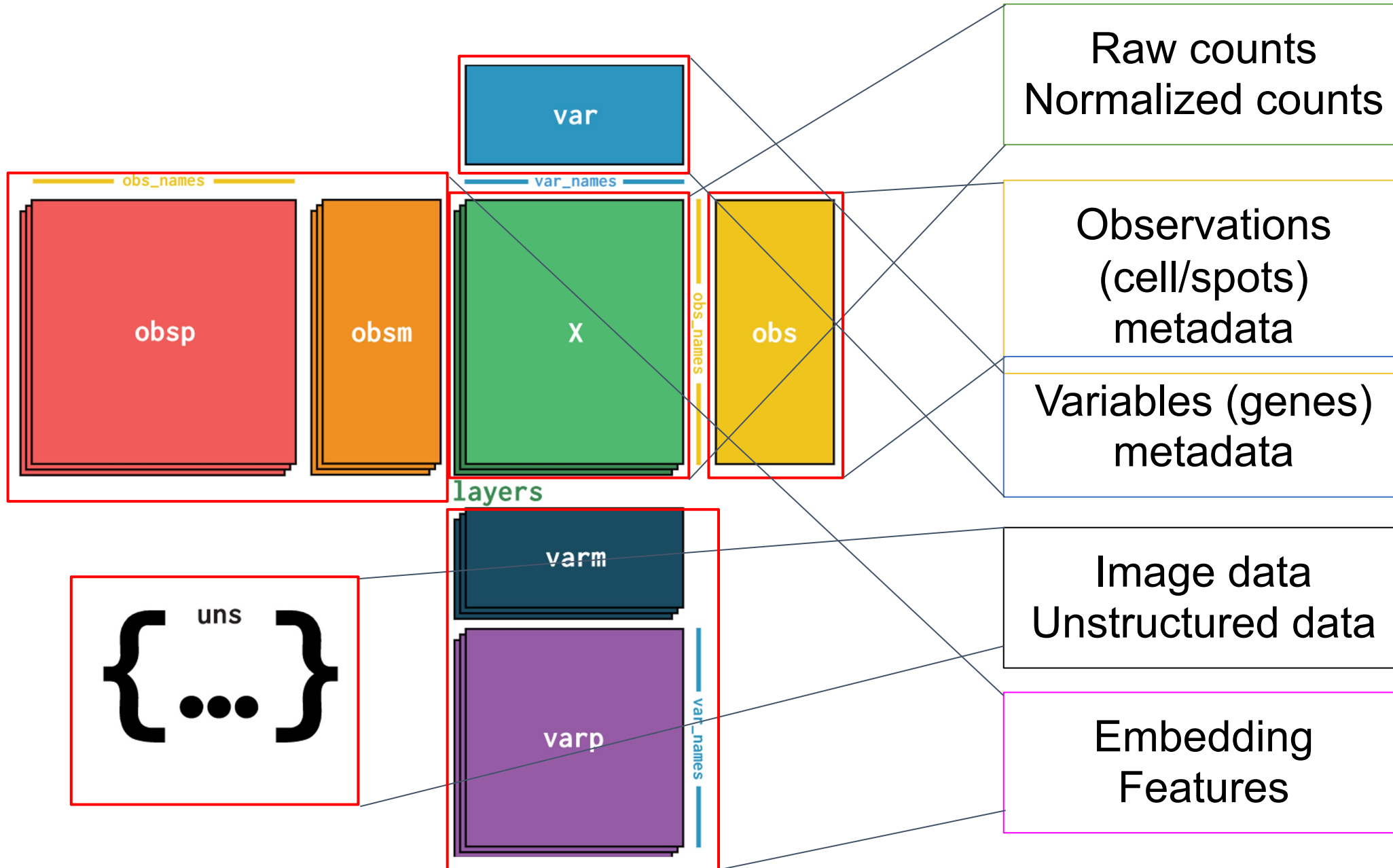


- **Data:** Collection of raw facts (numeric, categorical, etc.)
- **Data structure:** specialized format for *organizing* and *storing* data in memory that contains not only the *elements* stored but also *their relationship* to each other

Popular data structures



AnnData (Annotated data) - Python



SeuratObject - R

Seurat Object

Assays

Raw counts
Normalised Quantitation

Metadata

Experimental Conditions
QC Metrics
Clusters

Embeddings

Nearest Neighbours
Dimension Reductions

Variable Features

Variable Gene List

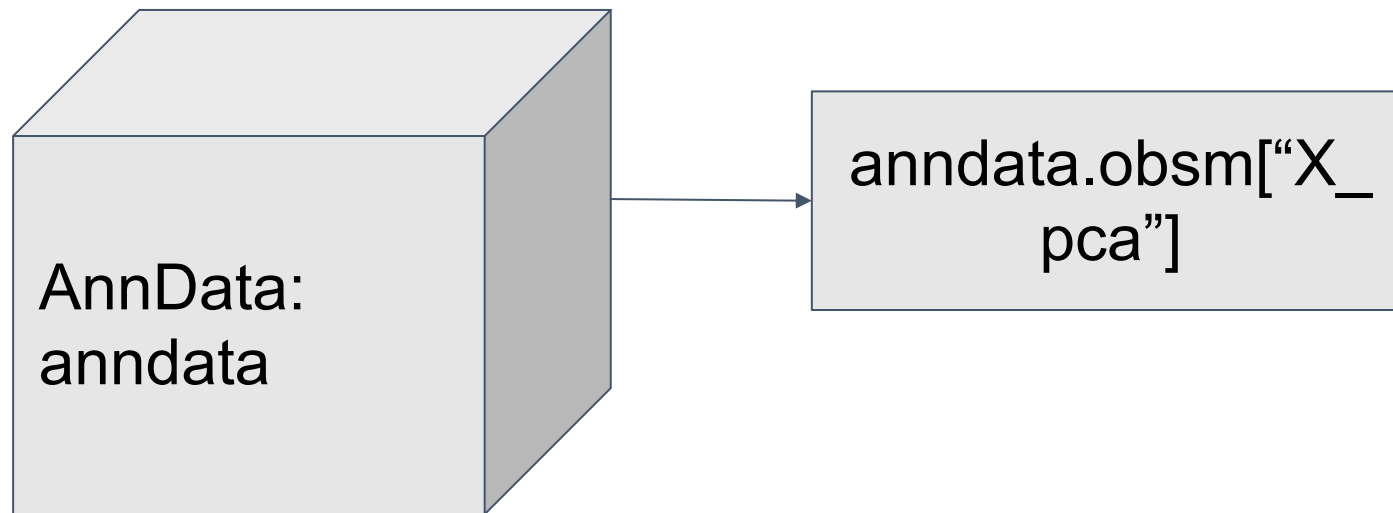
Use case:

Perform K-means clustering and store to AnnData

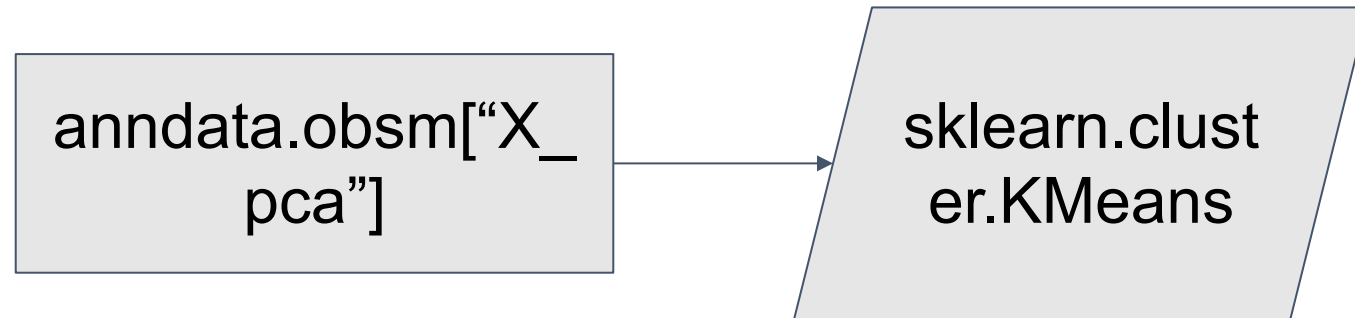
How?

1. Extract the PCs components from AnnData for every cells/spots
2. Using external scikit-learn package for K-means clustering
3. Get the K-means clustering results
4. Add results to observation annotation of AnnData object

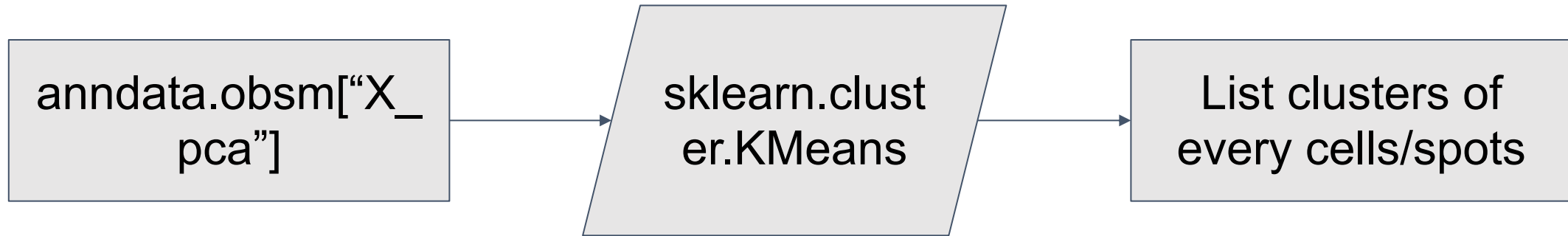
1. Extract the PCs components from AnnData for every cells/spots



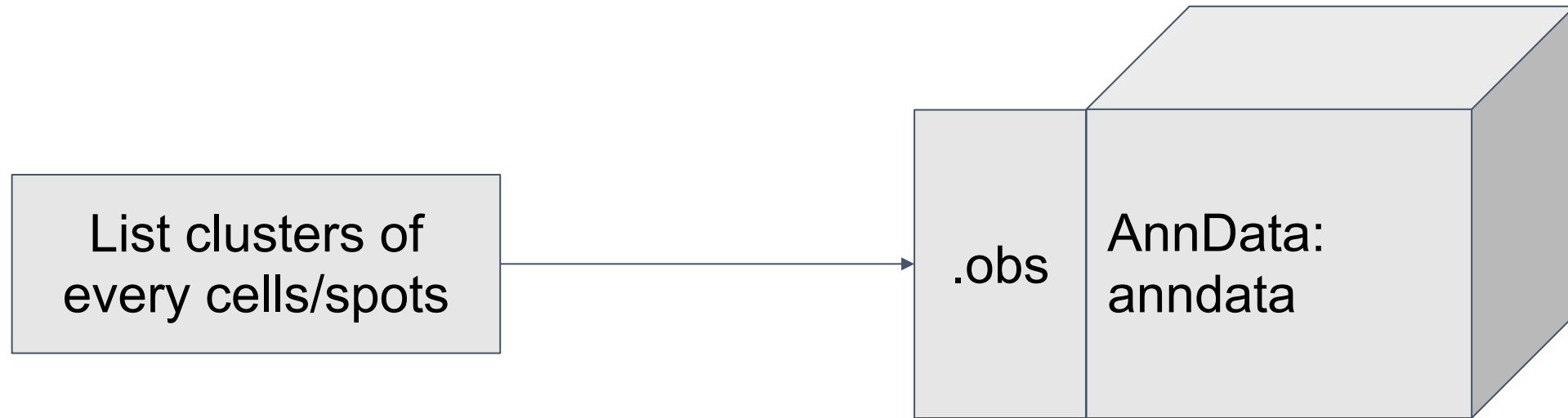
2. Using external scikit-learn package for K-means clustering



3. Get the K-means clustering results

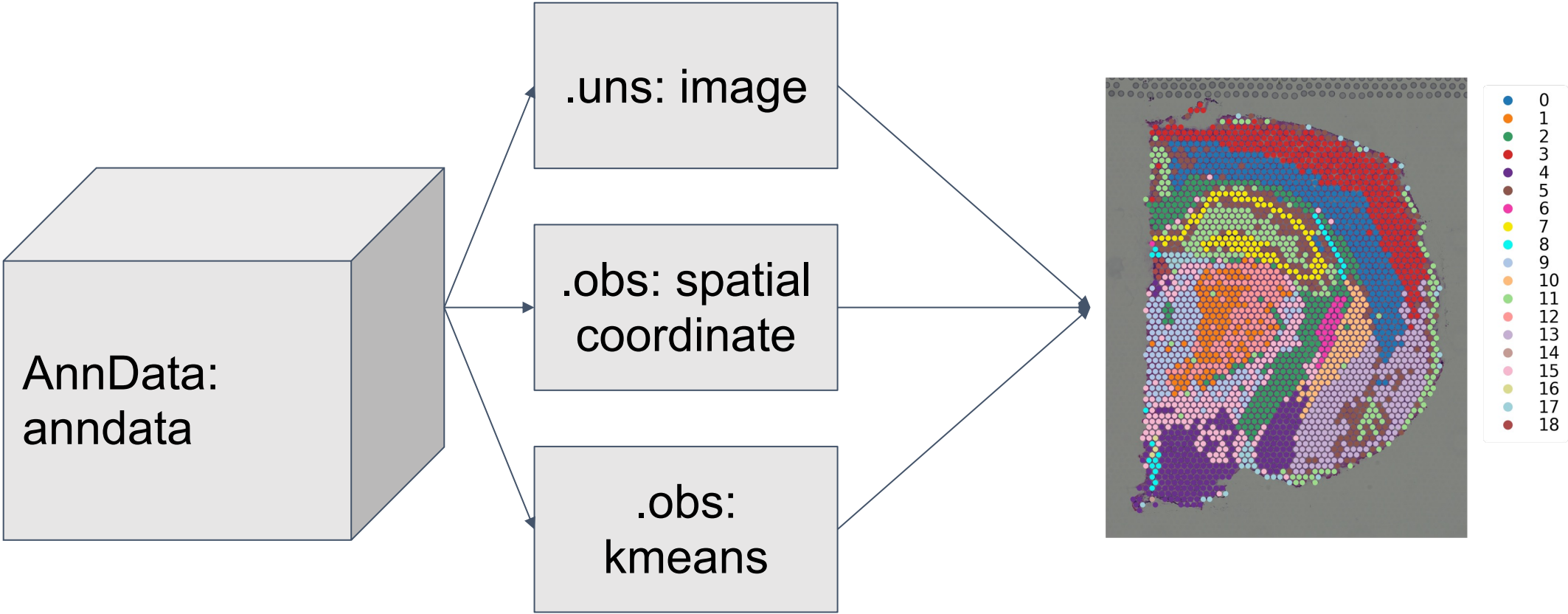


4. Add results to observation annotation of AnnData object



Use case:

Plotting Kmeans results for spatial transcriptomics



Discussion and Future perspectives