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.

Image: Digital reproduction of *A guidance through time* by Casey Coolwell and Kyra Mancktelow

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 5 Cellular Transcriptomics

Room 304, Building 69

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

Slides and Practical notes:

https://cnsgenomics.com/data/teaching/GNGWS22/

Day 1 (June 23rd Thursday): Single cell analysis		
Lecture (Morning; single cell data and theory for common analyses)		
9-9:20am	Introduction to participants and instructors	All
9:30-9:40am	Introduction scRNA and spatial transcriptomics data	Quan Nguyen
9:40-10:00am	Data exploratory analysis and preprocessing	Quan Nguyen
10-10:20am	Data normalisation	Guiyan Ni
10:20-10:40am	Dimensionality reduction & Clustering	Quan Nguyen
10:40-11am	Break	
11:00-11:20am	Differential expression analysis	Guiyan Ni
11:20-11:30	Cell type analysis	Sally Mortlock
11:30-11:45	eQTL single cell/tissue/bulk	Sally Mortlock
11:50am-12:00pm	Questions and discussions and future perspectives	

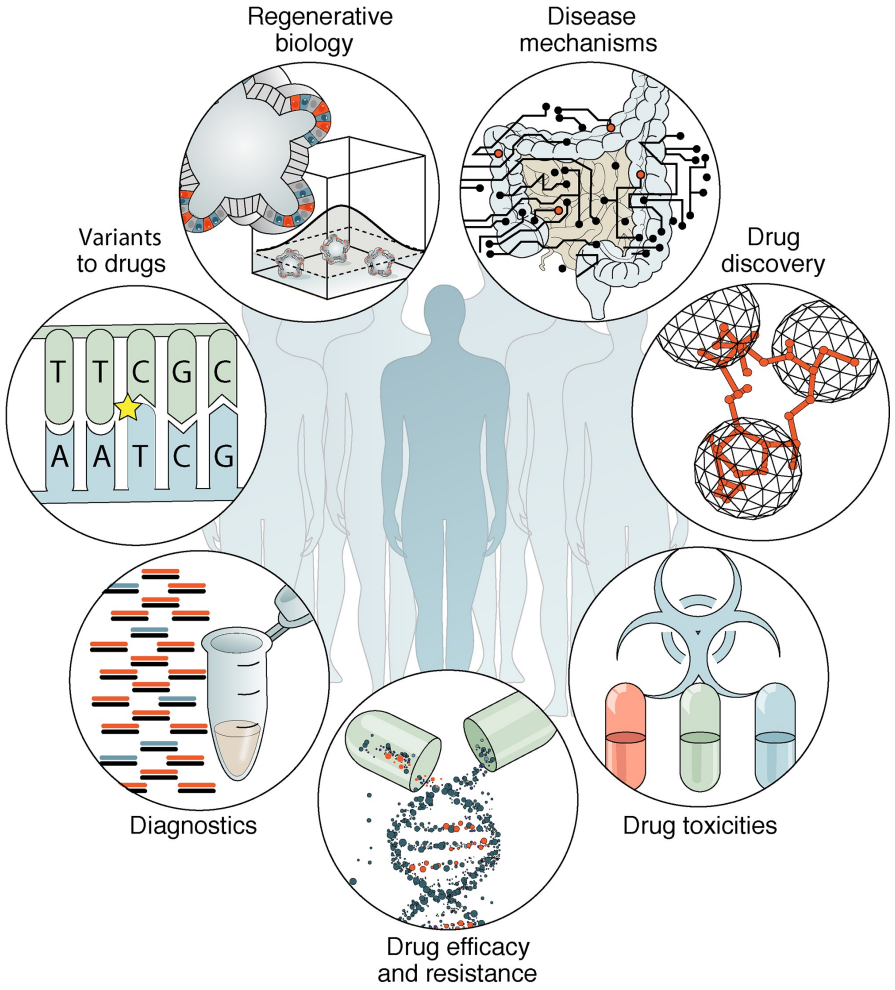
Single cell informatics

Scale

Resolution

INFORMATICS

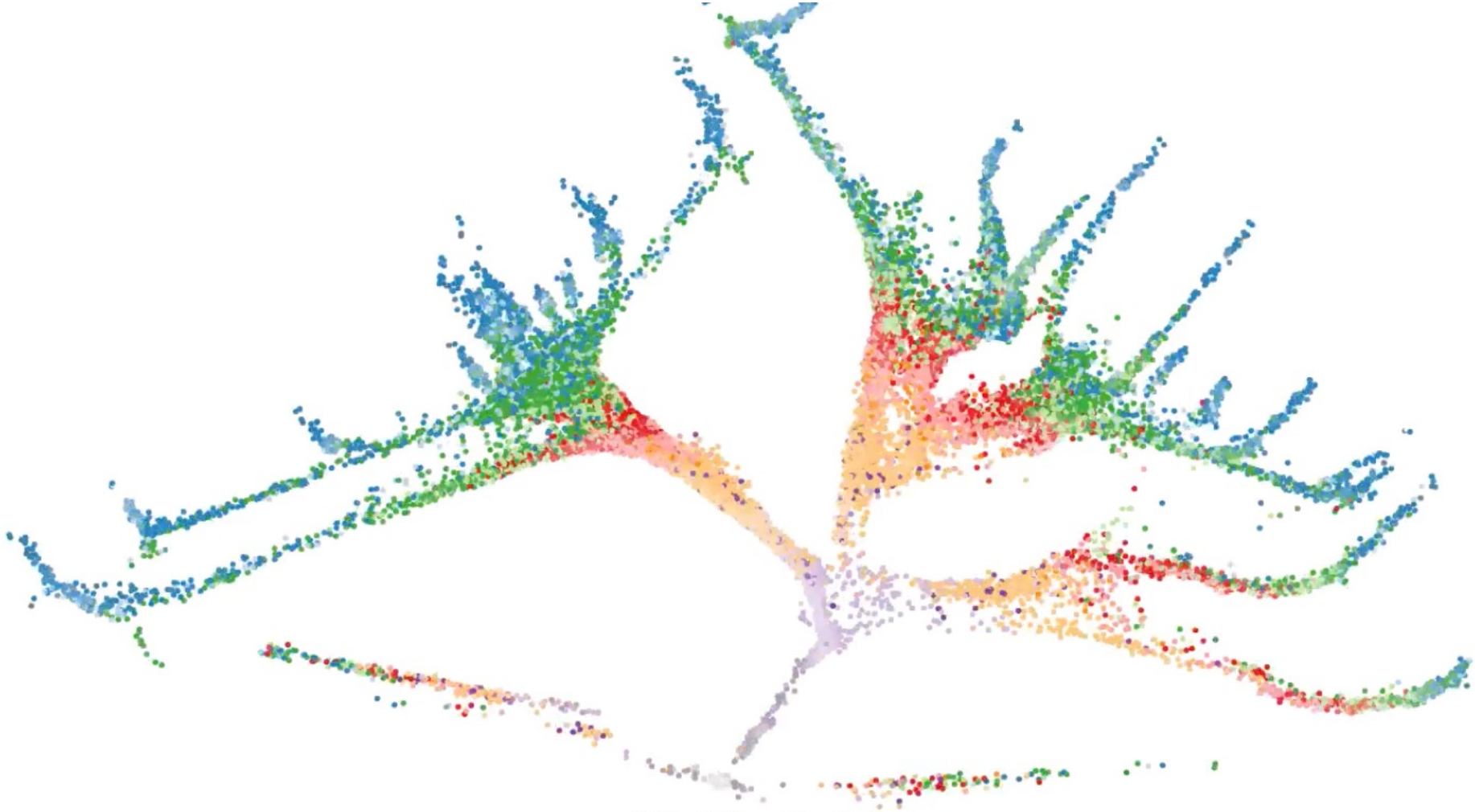
Precision Genomics Medicine



The G&G Cellomics Team

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

General introduction single cell and spatial transcriptomics



Science
AAAS

2018
BREAKTHROUGH
of the YEAR

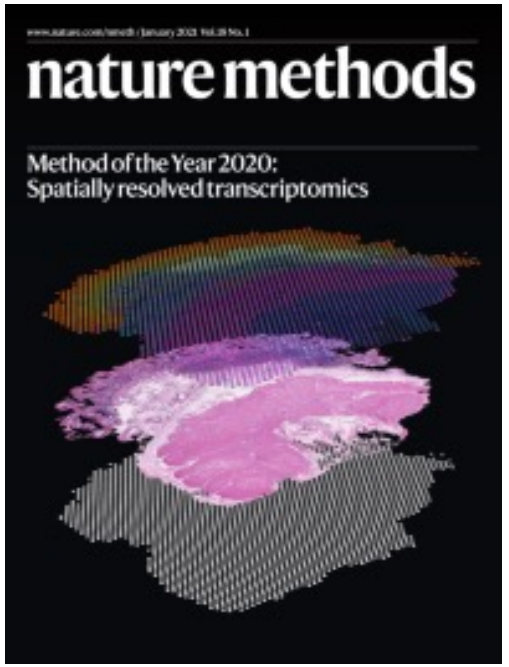
The single-cell revolution is just starting

Advanced genomics technologies

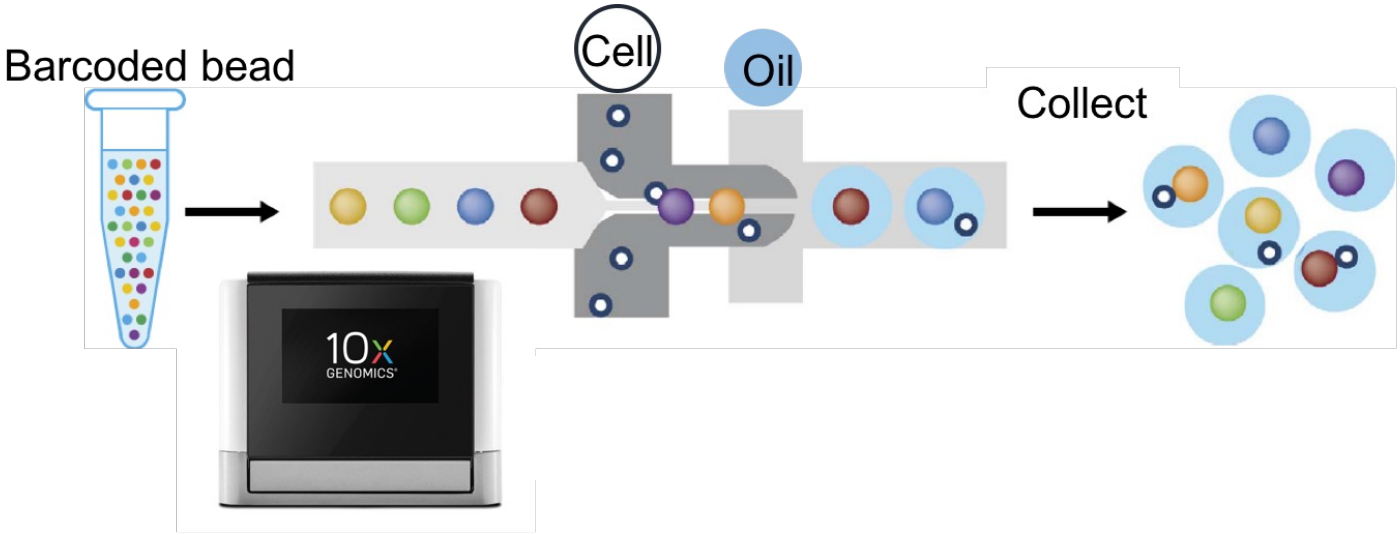
2018: Single Cell Transcriptomics

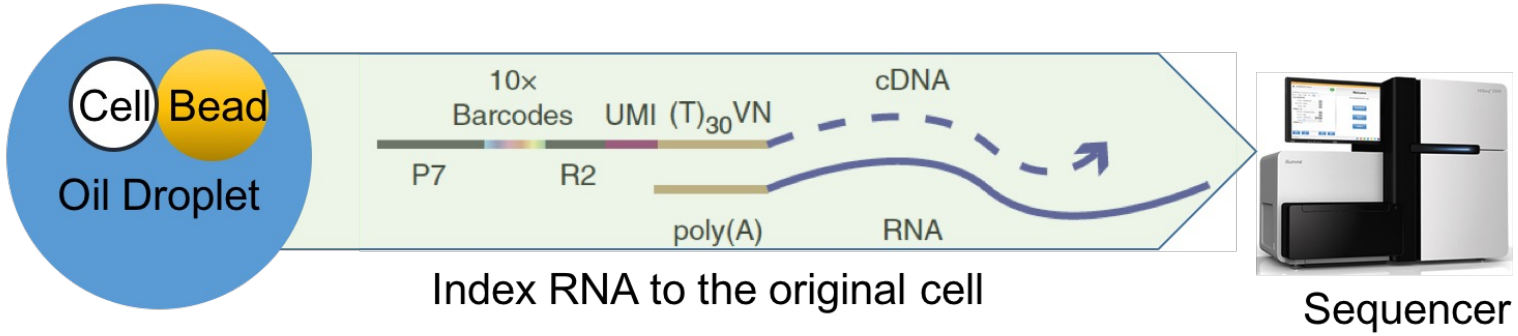
2019: Single Cell Multiomics

2020: Spatial Transcriptomics



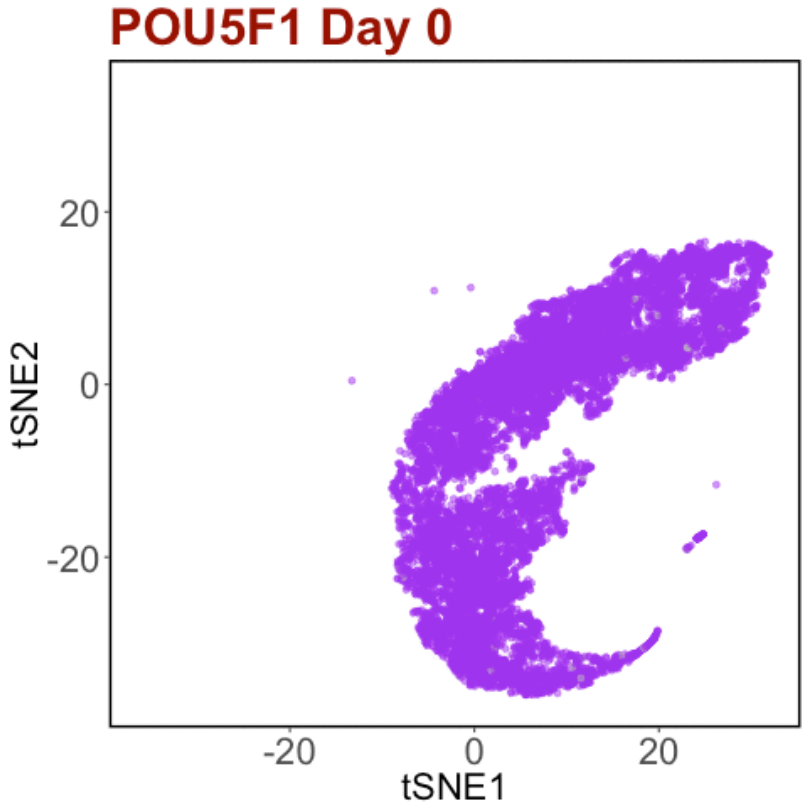
Single cell RNA sequencing





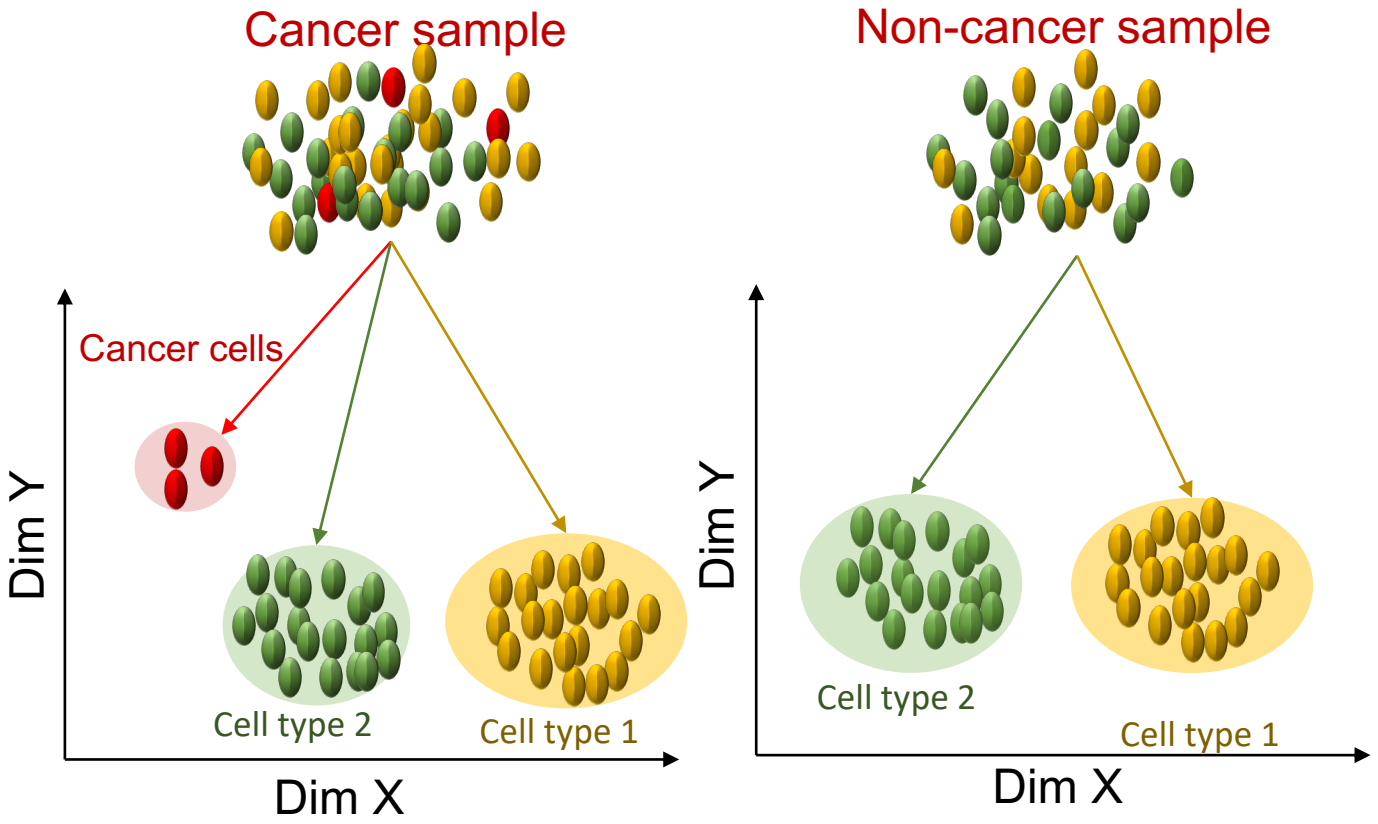
Index RNA to the original cell

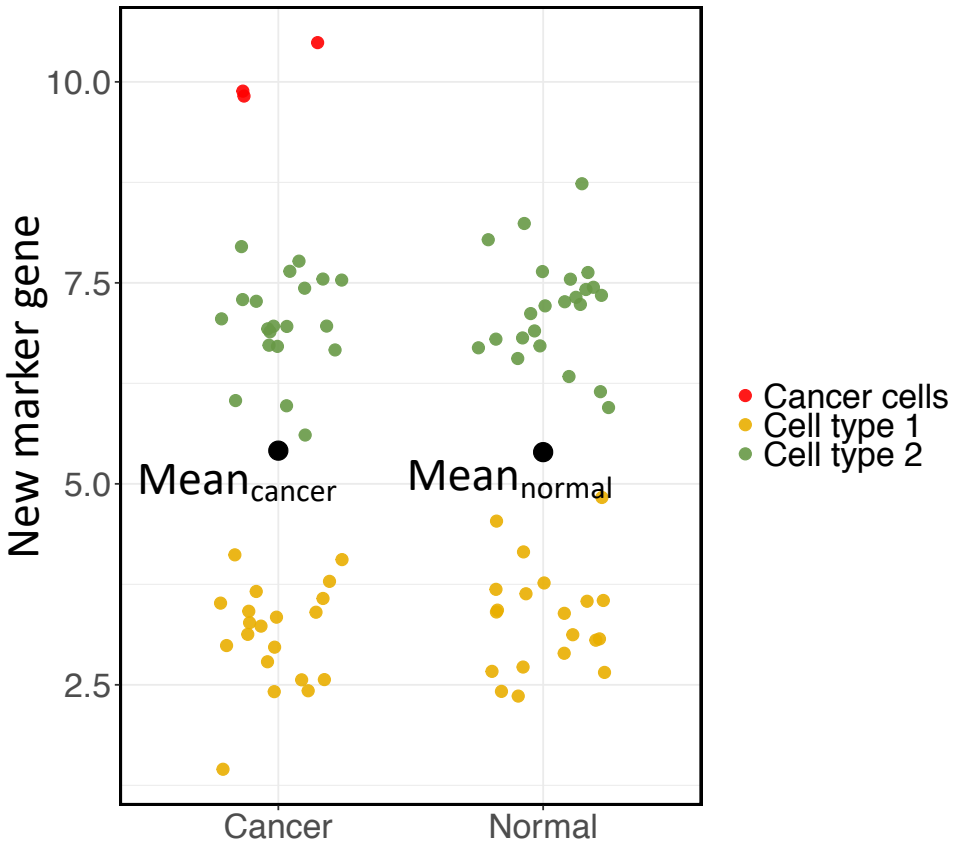
Sequencer



- 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

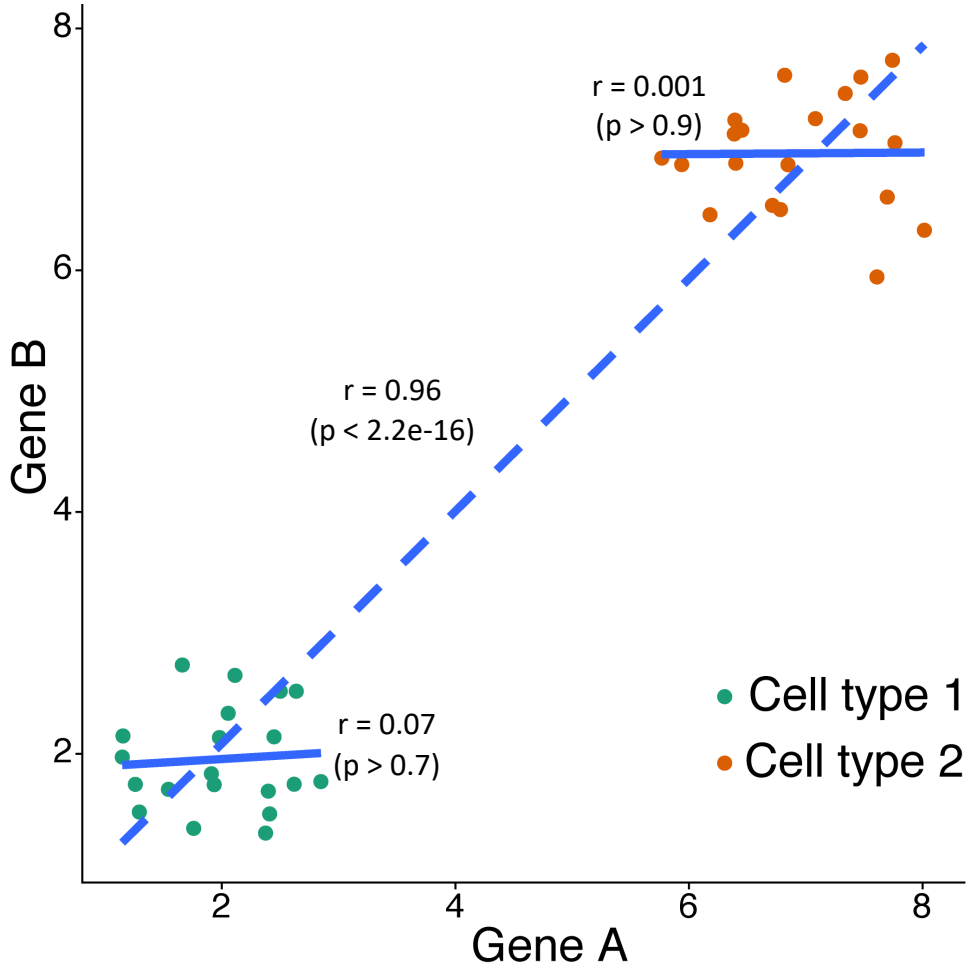
Disease at single-cell resolution

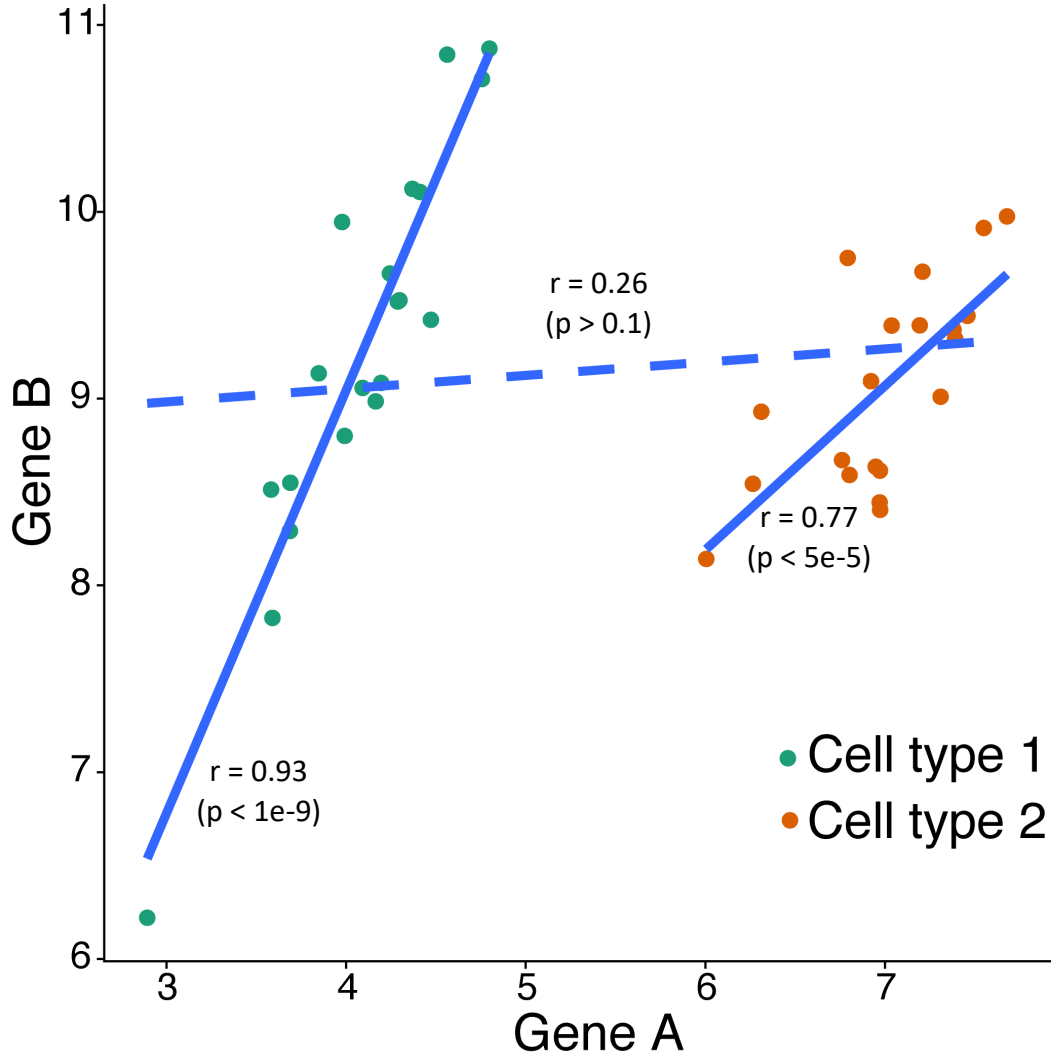




- Bulk RNA sequencing: no difference in mean expression
- Single-cell sequencing: can detect higher expression in cancer cells

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)

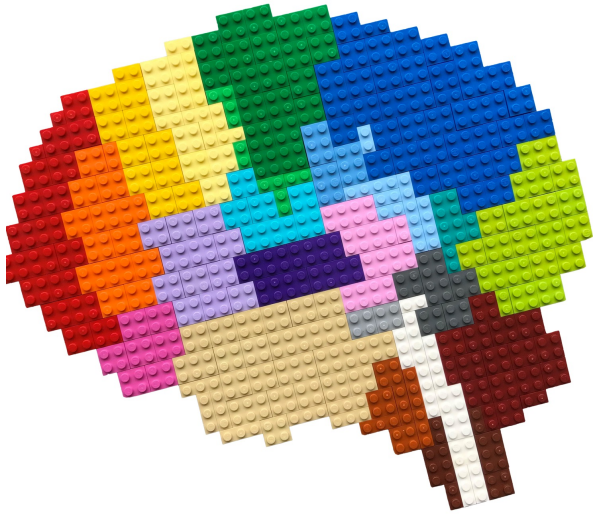
Spatial transcriptomics approach

Bulk

Single cell

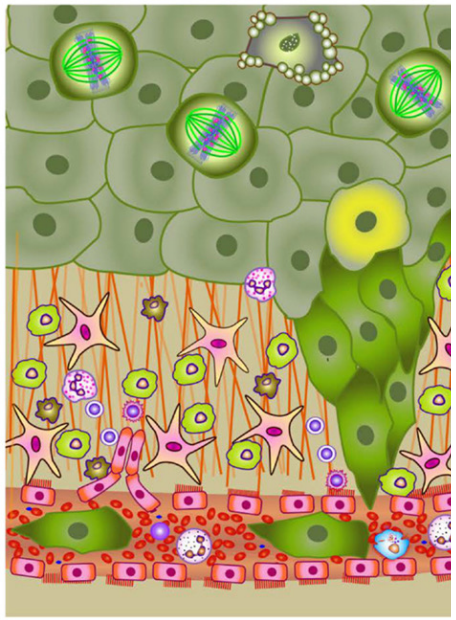
Spatial

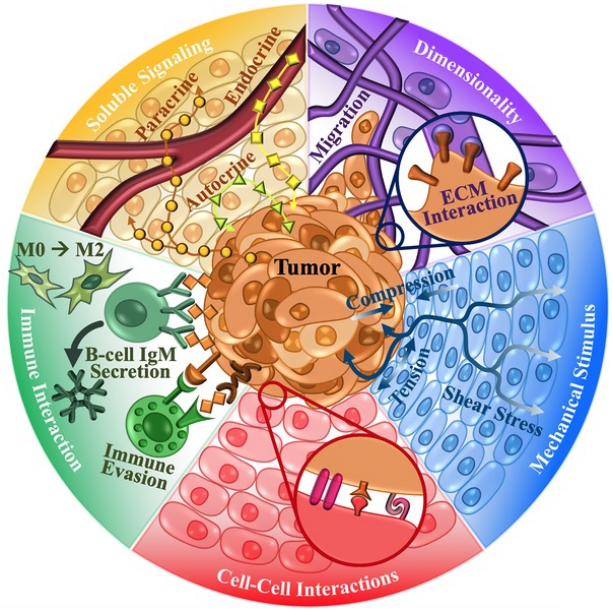
Lego:
(@boxia)

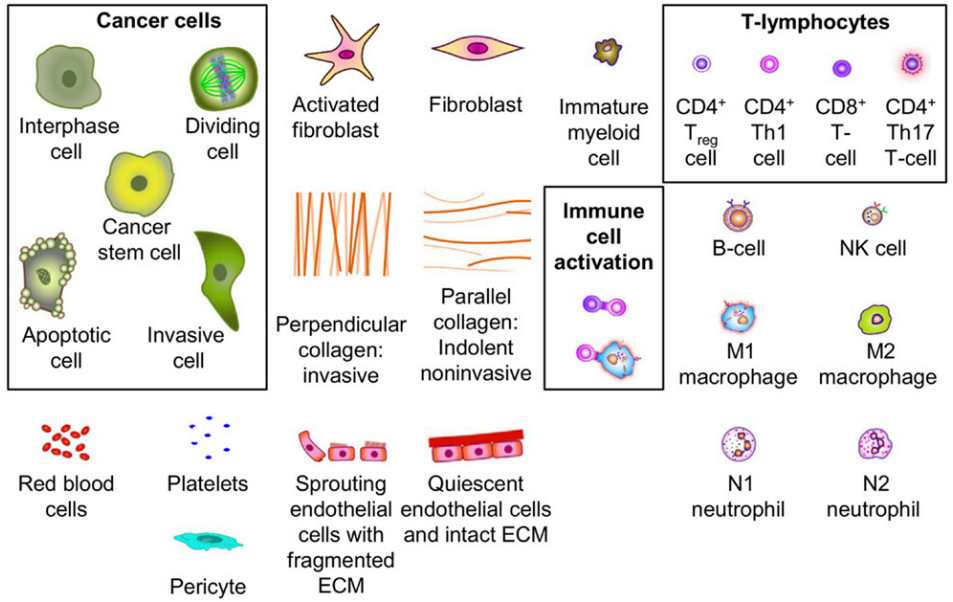


Fruit salad:
(@LGMartelotto)

Cellular ecosystem within a tissue







(Gerdes et al. 2014; Bregenzer et al. 2019)

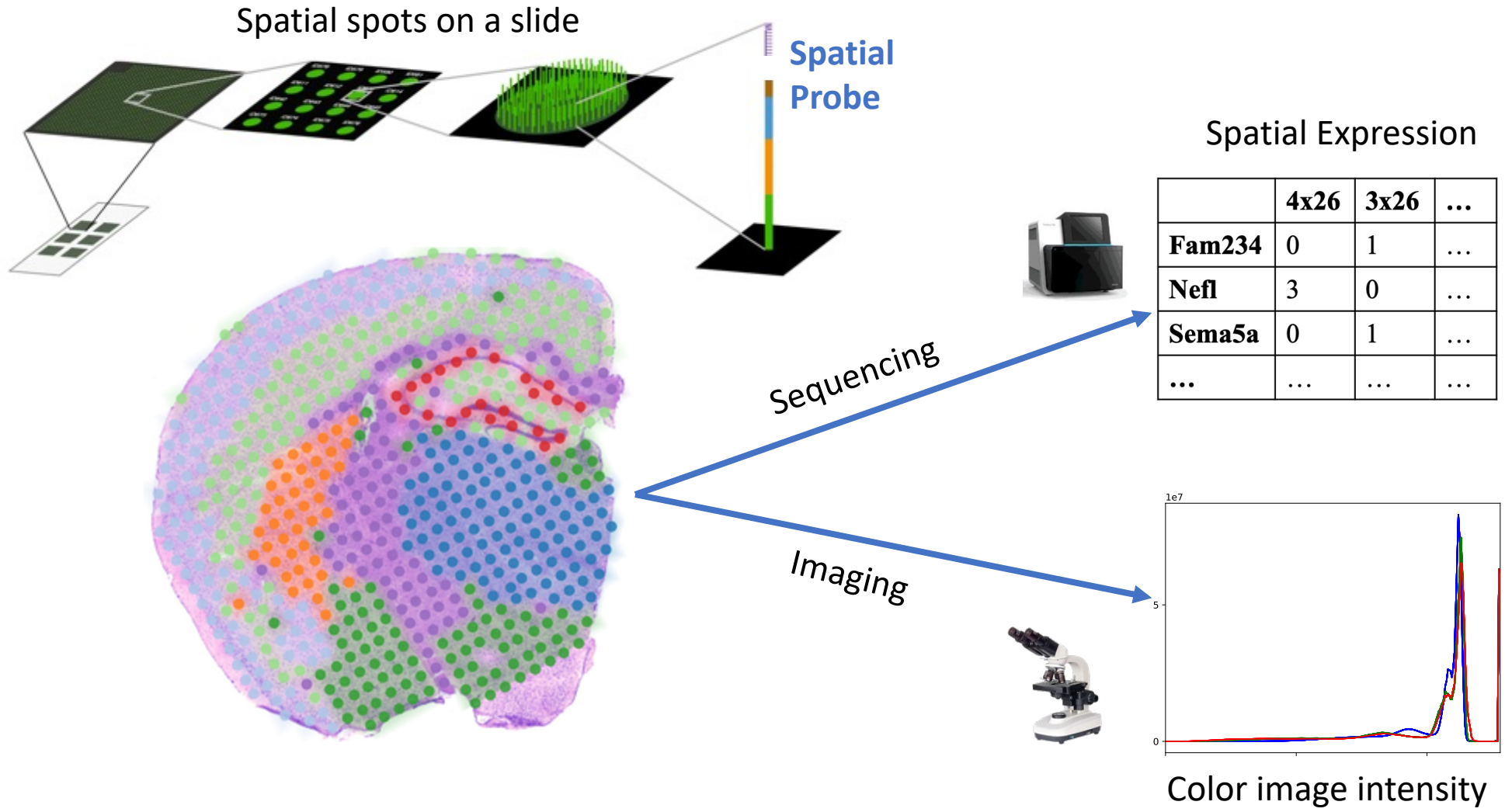
Bulk

Single cell

Spatial

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

Spatial transcriptomics adds spatial dimension and tissue morphology



	4x26	3x26	…
Fam234	0	1	…
Nefl	3	0	…
Sema5a	0	1	…
…	…	…	…

- 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.

Analysis tools for single cells and spatial data

Software programs

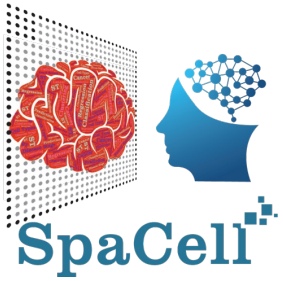
- scGPS: https://github.com/BiomedicalMachineLearning/scGPS
- ascend: https://github.com/BiomedicalMachineLearning/ascend
- scPred: https://github.com/IMB-Computational-Genomics-Lab/scPred
- CoreNET: https://github.com/BiomedicalMachineLearning/CoreNET
- HEMnet: https://github.com/BiomedicalMachineLearning/HEMnet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

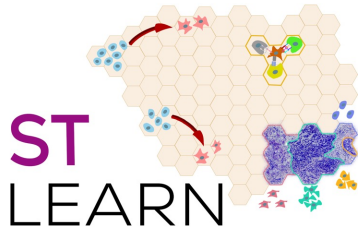
scRNAseq visualisation

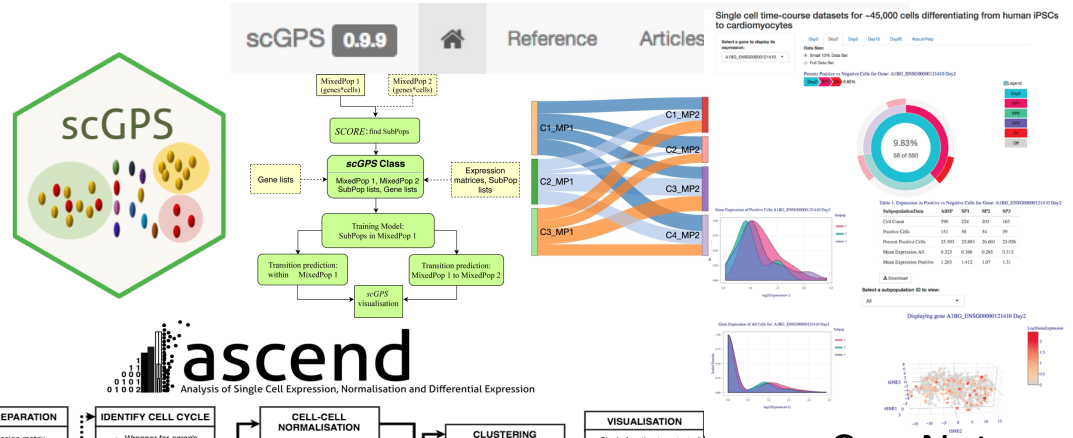
- HiPSC: http://computationalgenomics.com.au/shiny/hipsc
- Hipsc2cm: http://computationalgenomics.com.au/shiny/hipsc2cm
- scIVA: http://computationalgenomics.com.au/shiny/scIVA/

Spatial Transcriptomics

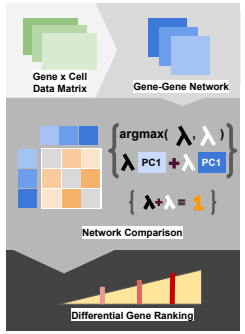
- SpaCell: https://github.com/BiomedicalMachineLearning/Spacell
- stLearn: https://stlearn.readthedocs.io/en/latest/



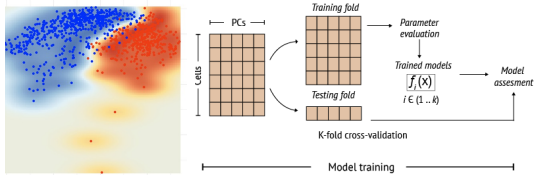


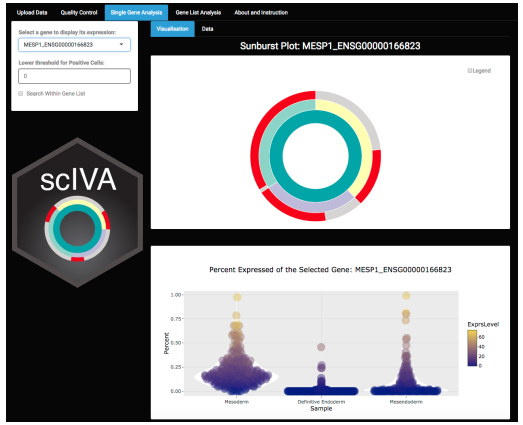


CoreNet



scPred

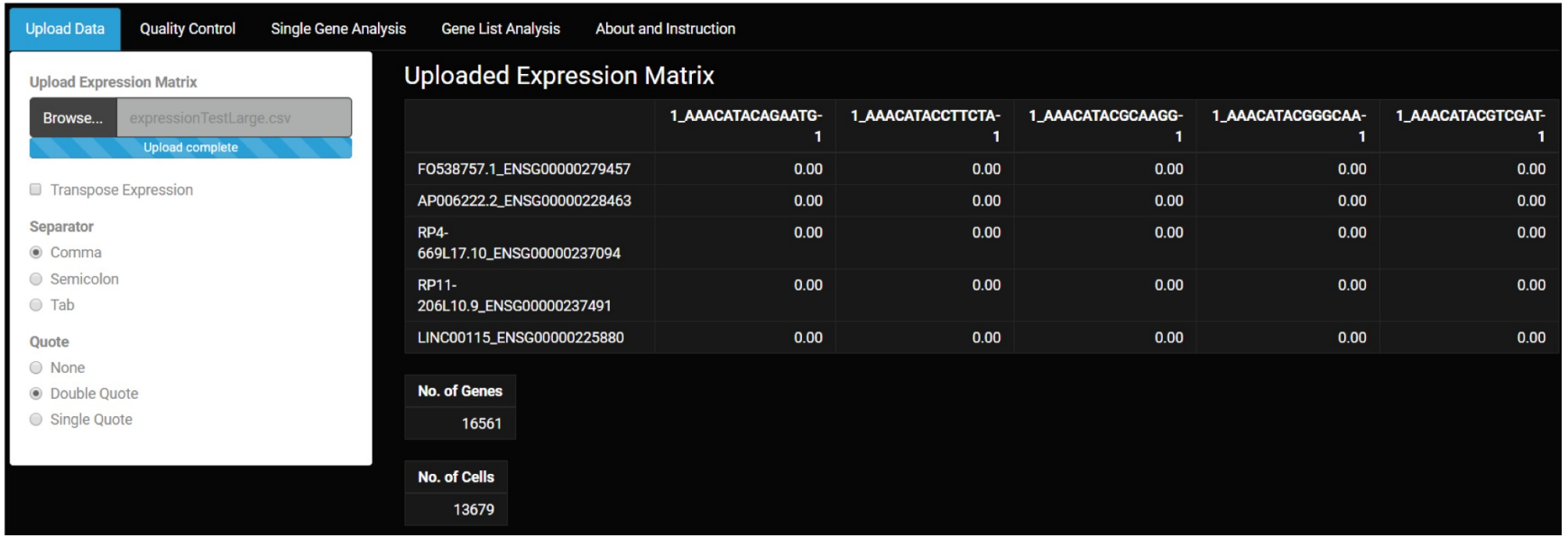




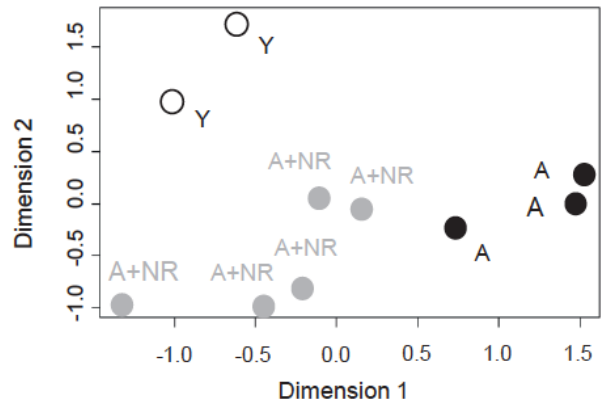
Data Preprocessing

Single cell data vs. bulk data

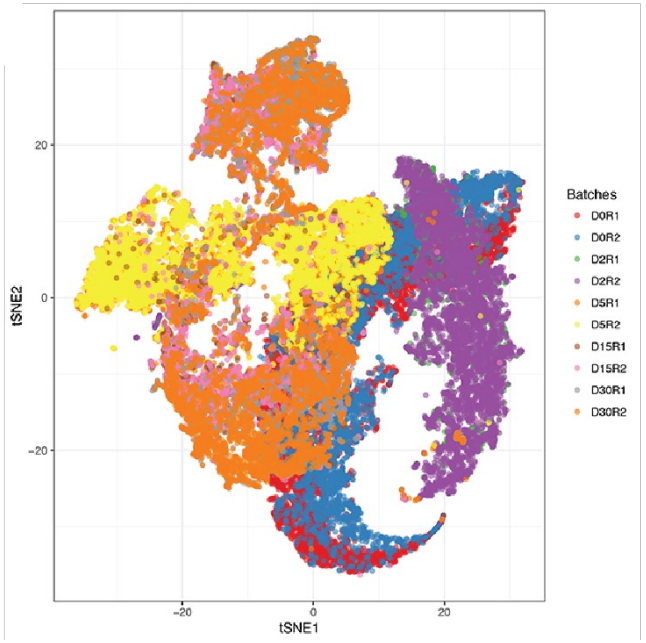
https://github.com/IMB-Computational-Genomics-Lab/scIVA



Bulk

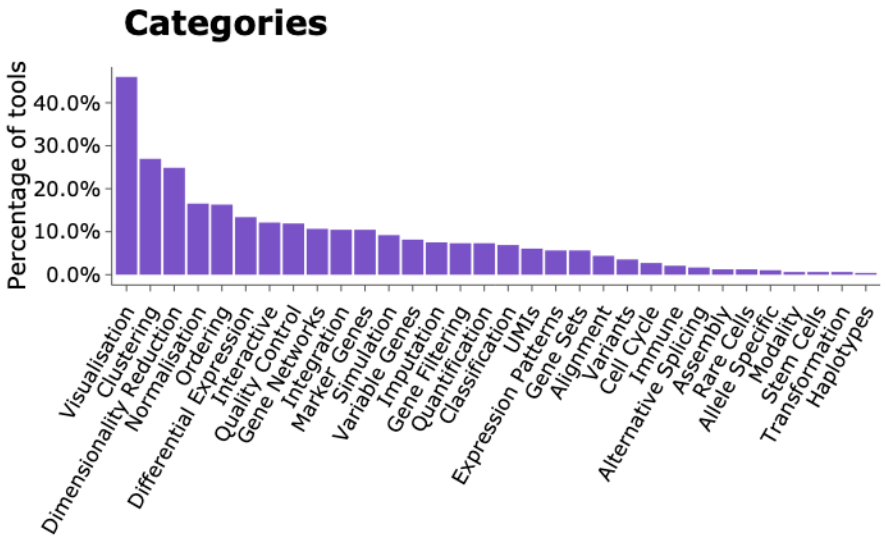


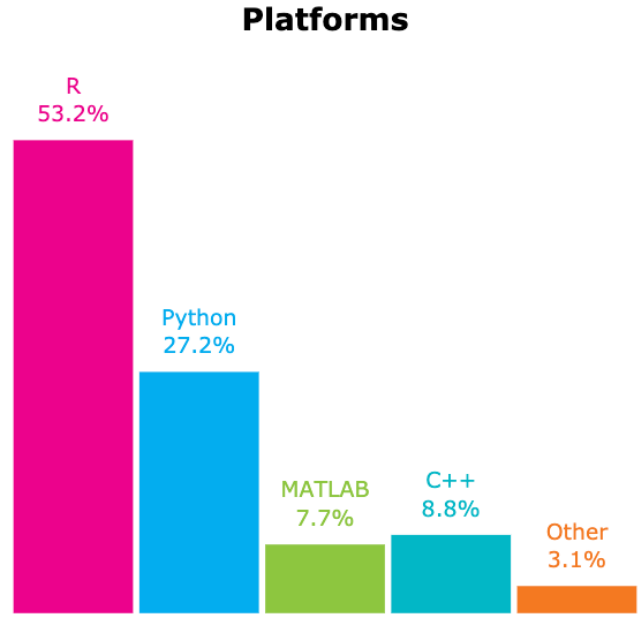
Single cell



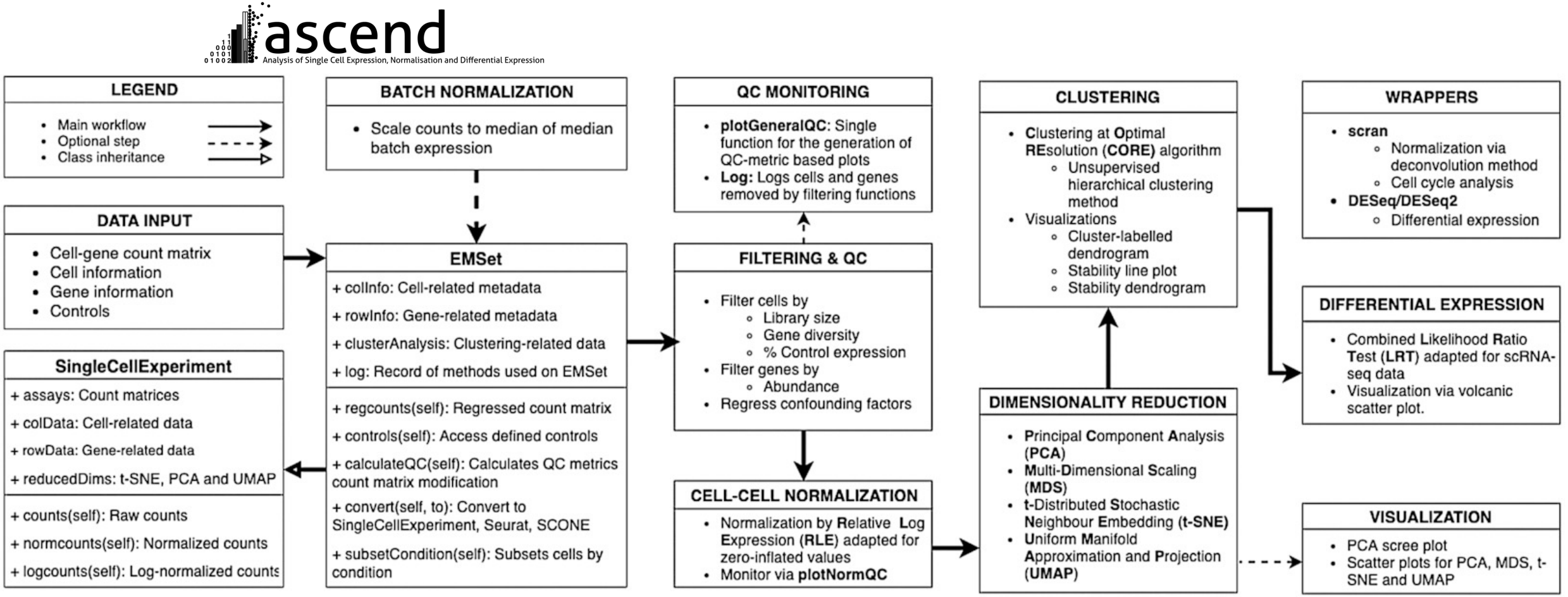
	Single cell	Bulk
Noisy data	Undetected genes (zero inflation)	Deep sequencing, most genes detected
Cell-cell variation	Measured	Not measured
Data size	Thousands of cells (1 cell ~ 1 bulk sample)	10-100 samples

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**

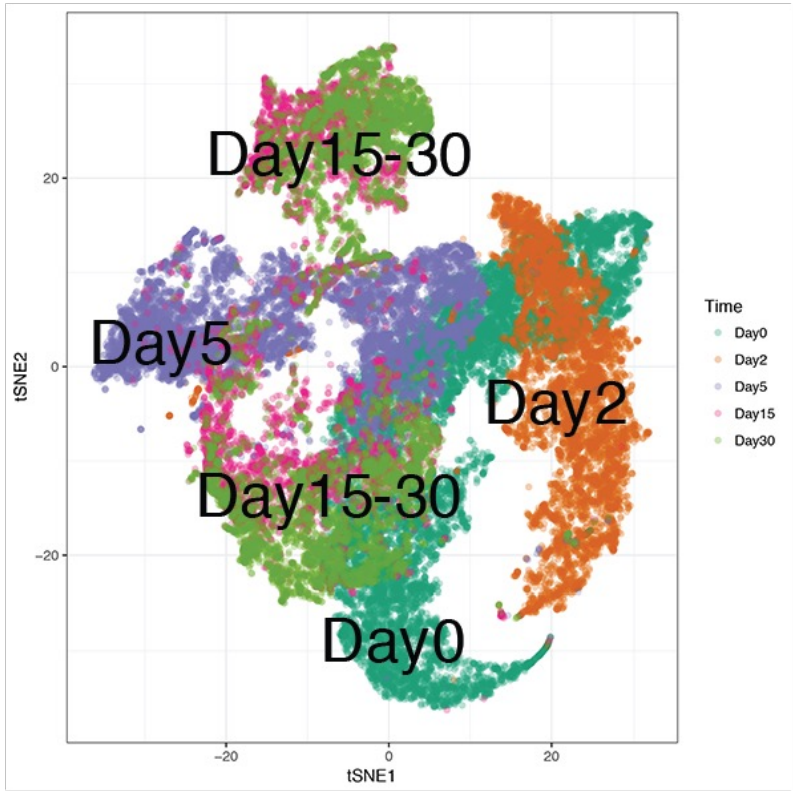
1. Data merging and normalising by batches (samples)
2. Data preprocessing (removing outlier cells and genes due to technical bias)
3. Cell-to-cell normalisation

↓ **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

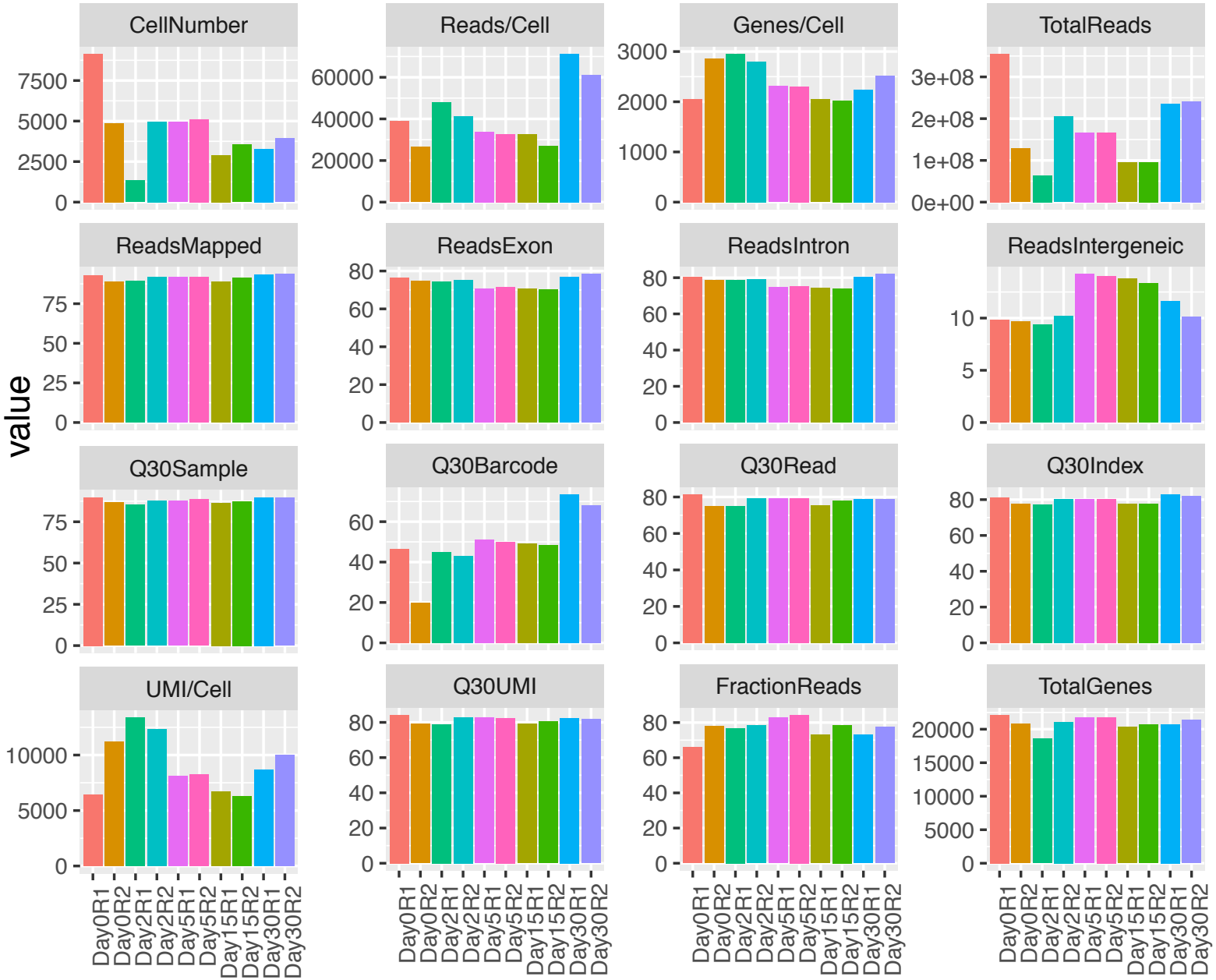
↓ **Clustering**

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

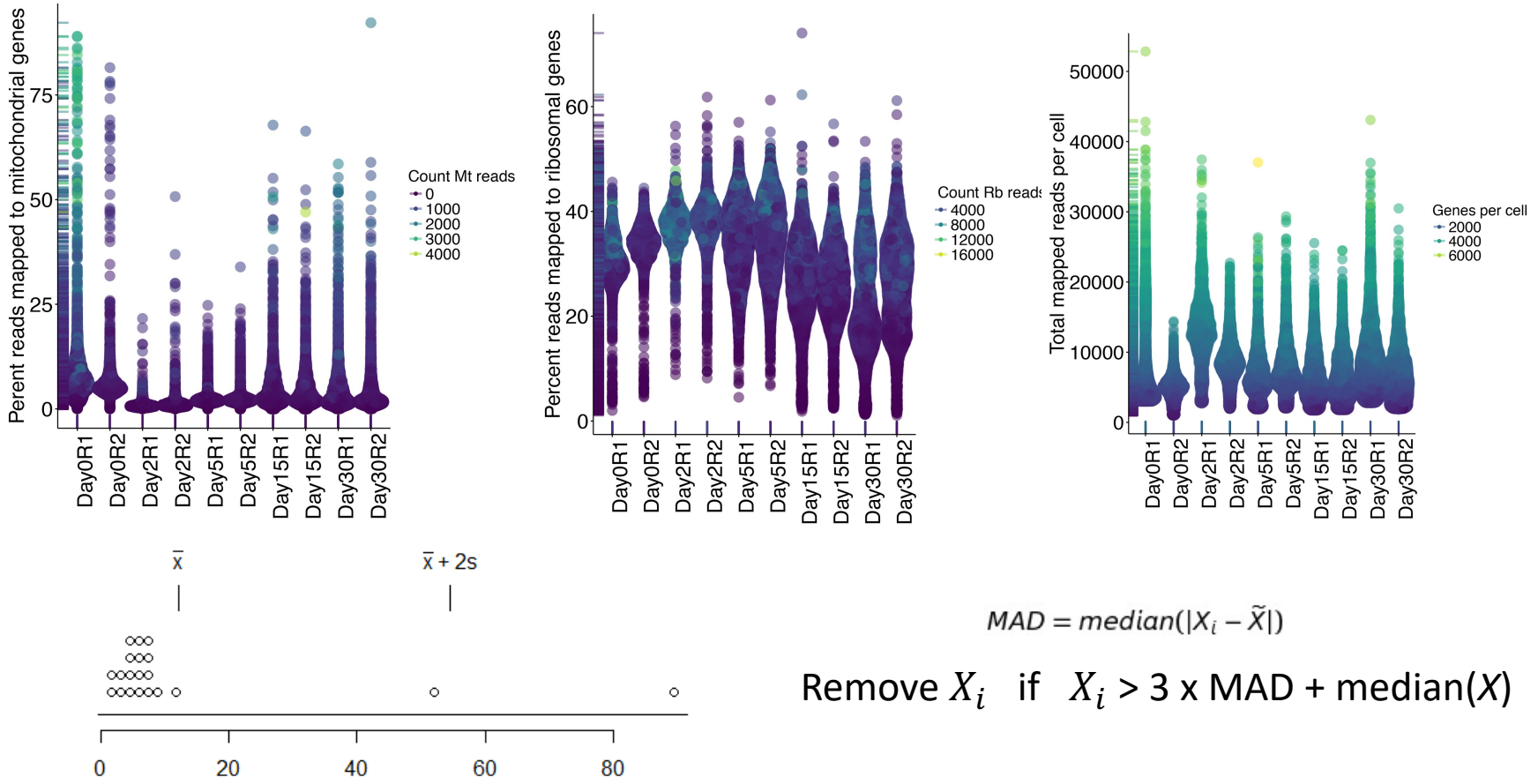


Data quality control: a range of QC measures

- 16 QC measures
- 10 scRNA-seq libraries



Data preprocessing: quality control and filtering genes and cells

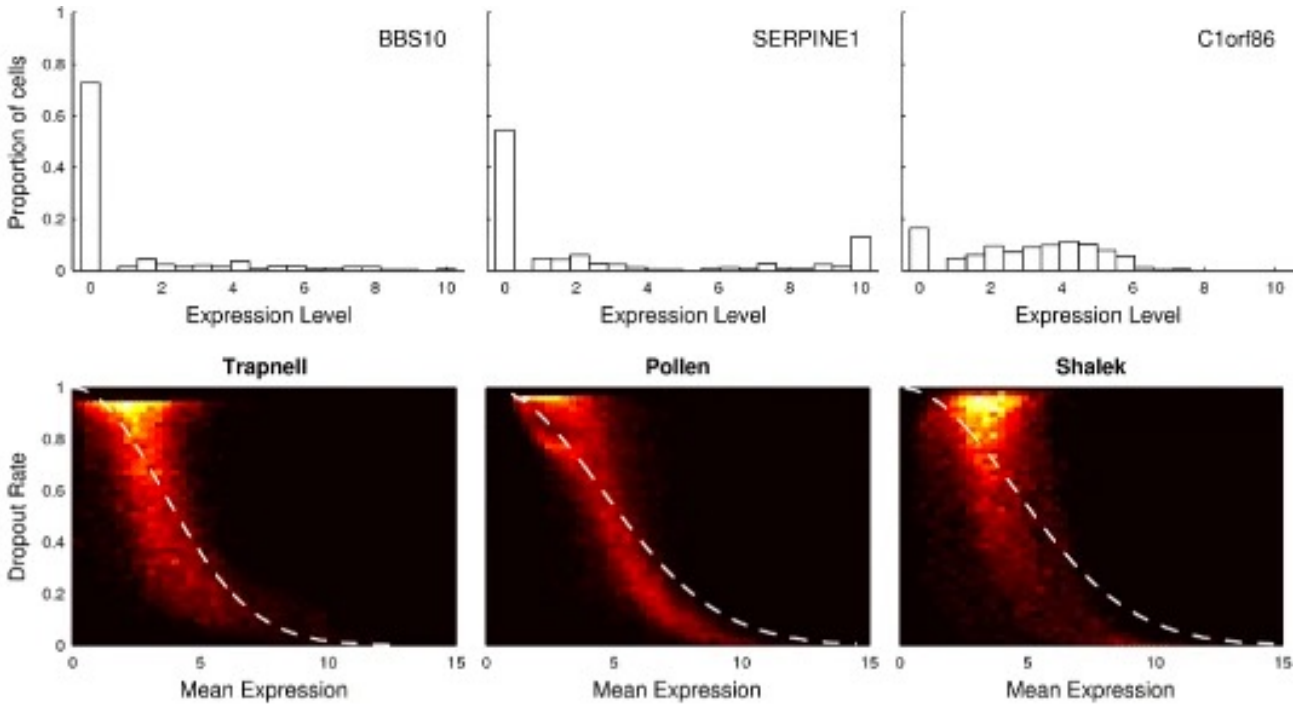


$$MAD = median(|X_i - \tilde{X}|)$$

Remove X_i if X_i > 3 x MAD + median(X)

- 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



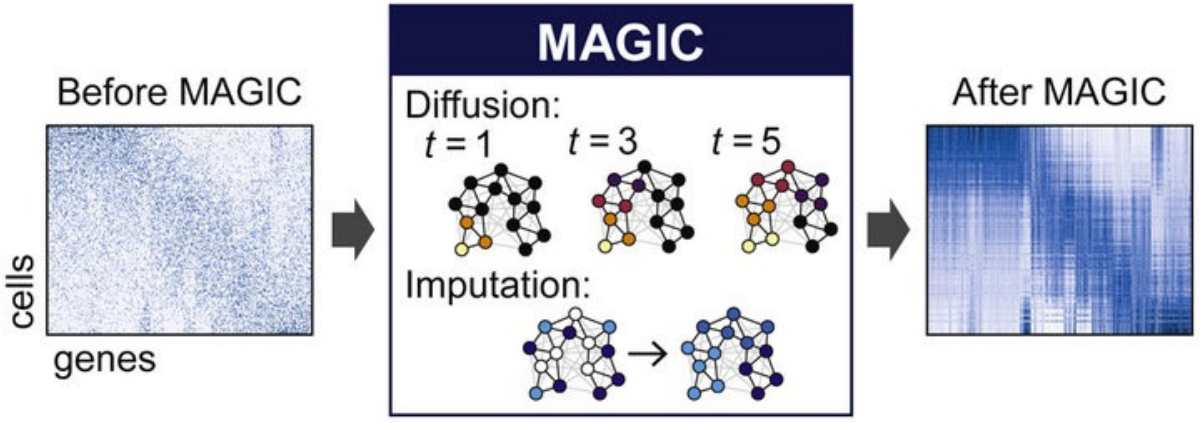
(Pierson and Yau, 2015)

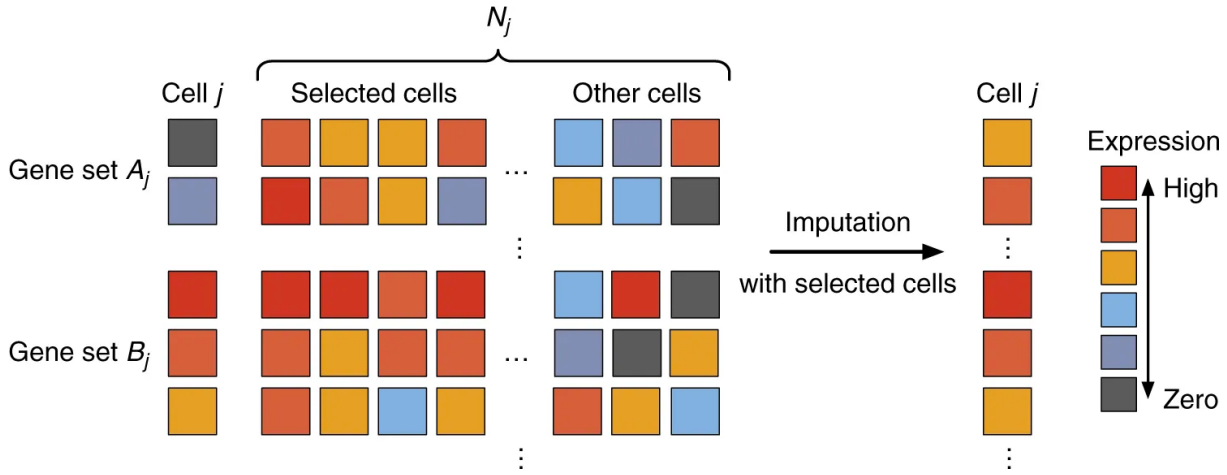
$p_0 = \exp(-\lambda \mu^2)$, where λ is a fitted parameter, μ non-zero mean expression, p_0 gene dropout rate

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

Single cell data: impute zero expression values





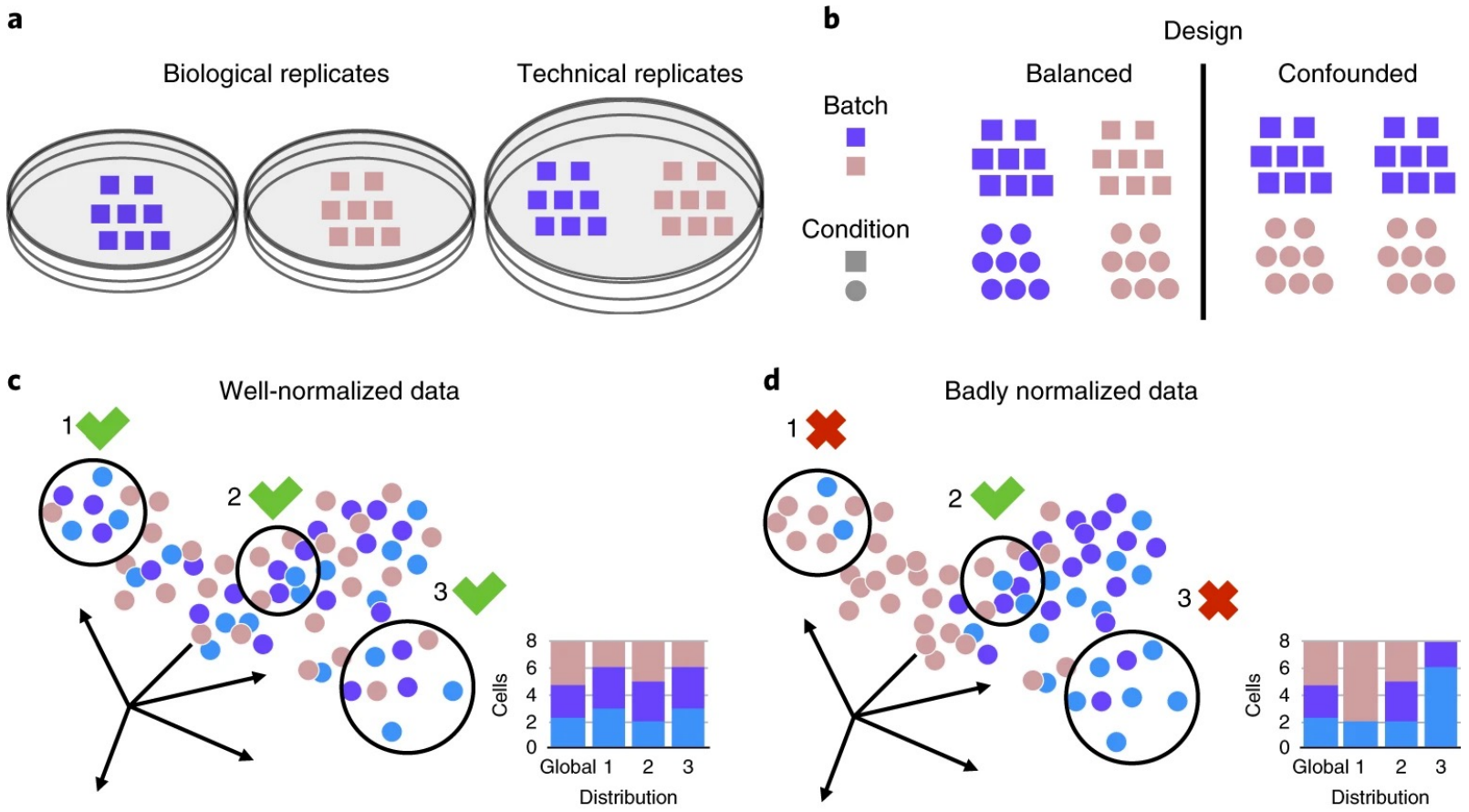
MAGIC: Markov Affinity-based Graph Imputation of Cells weights cells by Markov transition matrix (van Dijk et al., 2018)

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)

Data Normalisation

Batch effects

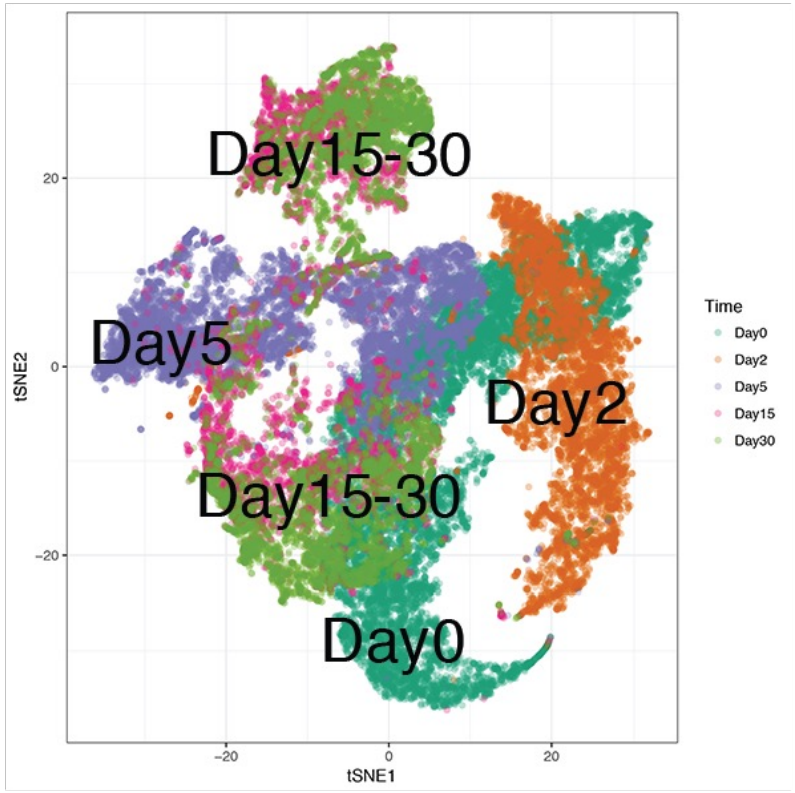
- Batch effects: technical differences induced by the operator or other experimental artifacts
- A balanced experimental design allocates samples evenly between batches, so that the effect can be easily regressed out in a linear model by setting appropriate covariates
- Assumption of orthogonality between the batch effect and the biological subspace



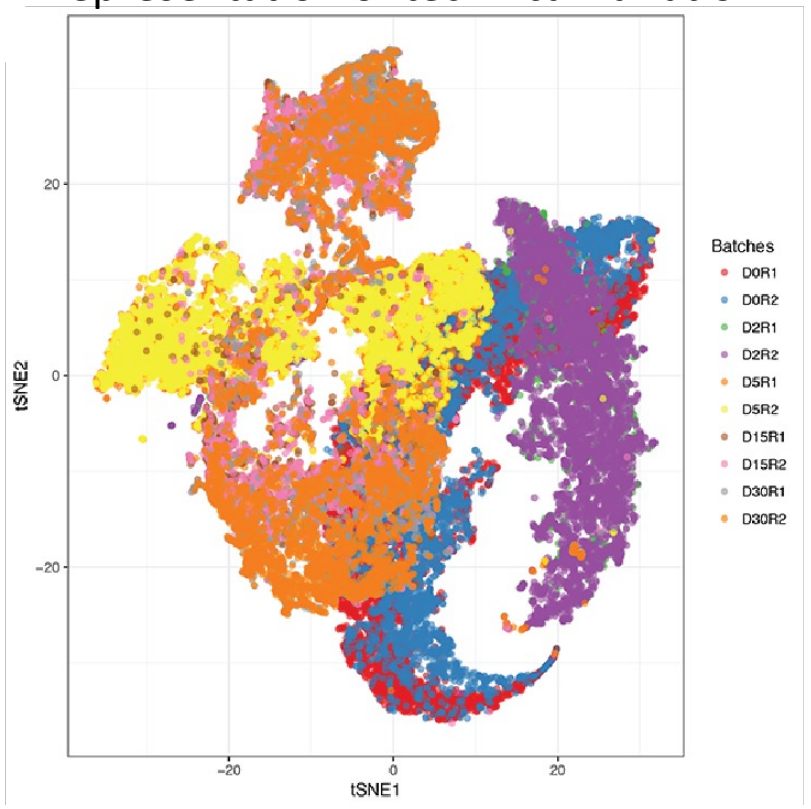
(Buttner et al, 2019)

Representation of biological and technical variation

Representation of biological variation



Representation of technical variation



$$Rate_i = min(MMR_j) \times N_i \times \frac{RF_i}{TMR_i}$$

Batch normalisation by sequencing depth:

$Rate_i$ is the binomial rate parameter for sampling reads in sample *i*

MMR_j ratio sequencing depth to mean depth

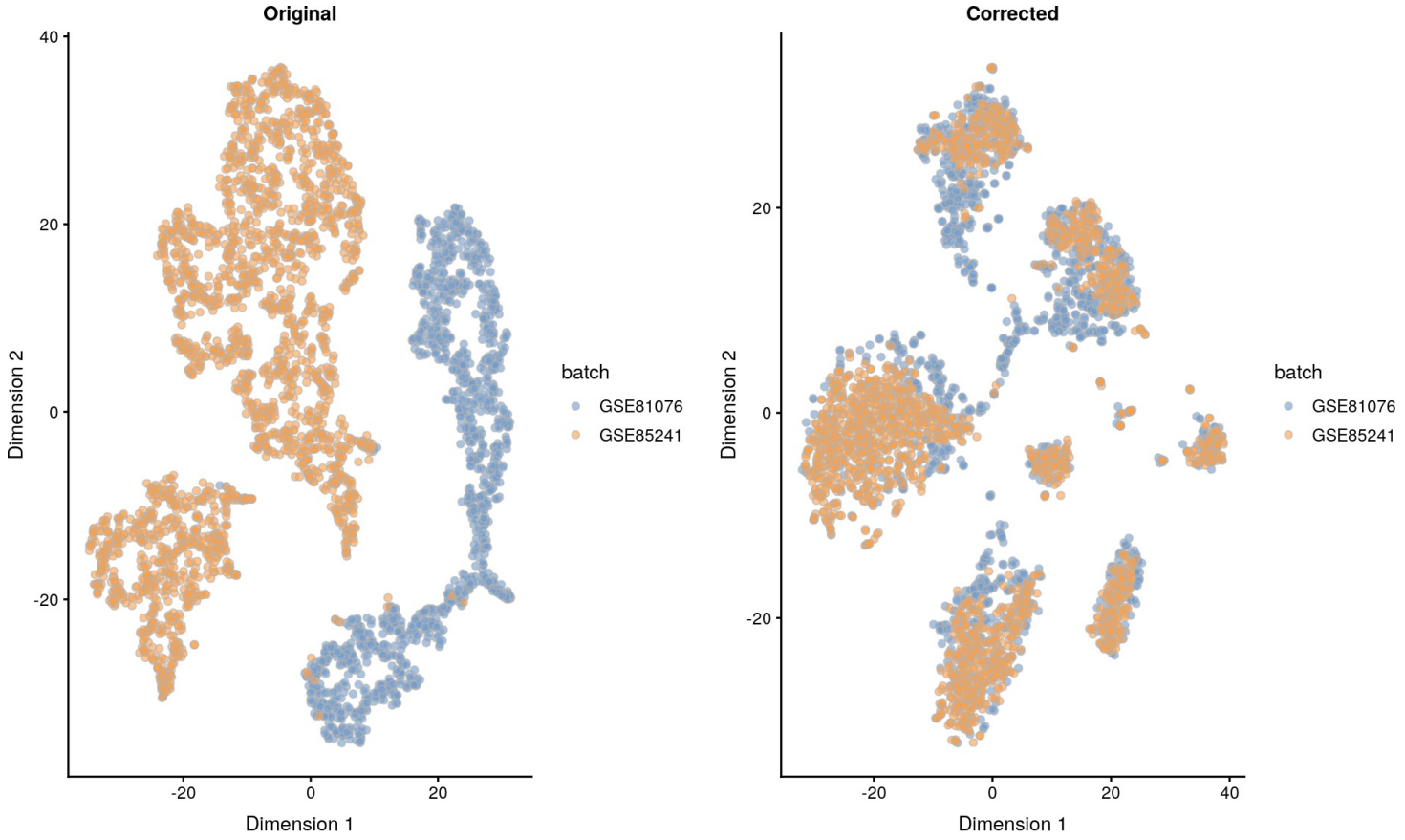
N_i is the number of cells in sample *i*

RF_i is the fraction of mapped reads; TMR_i is the total mapped reads

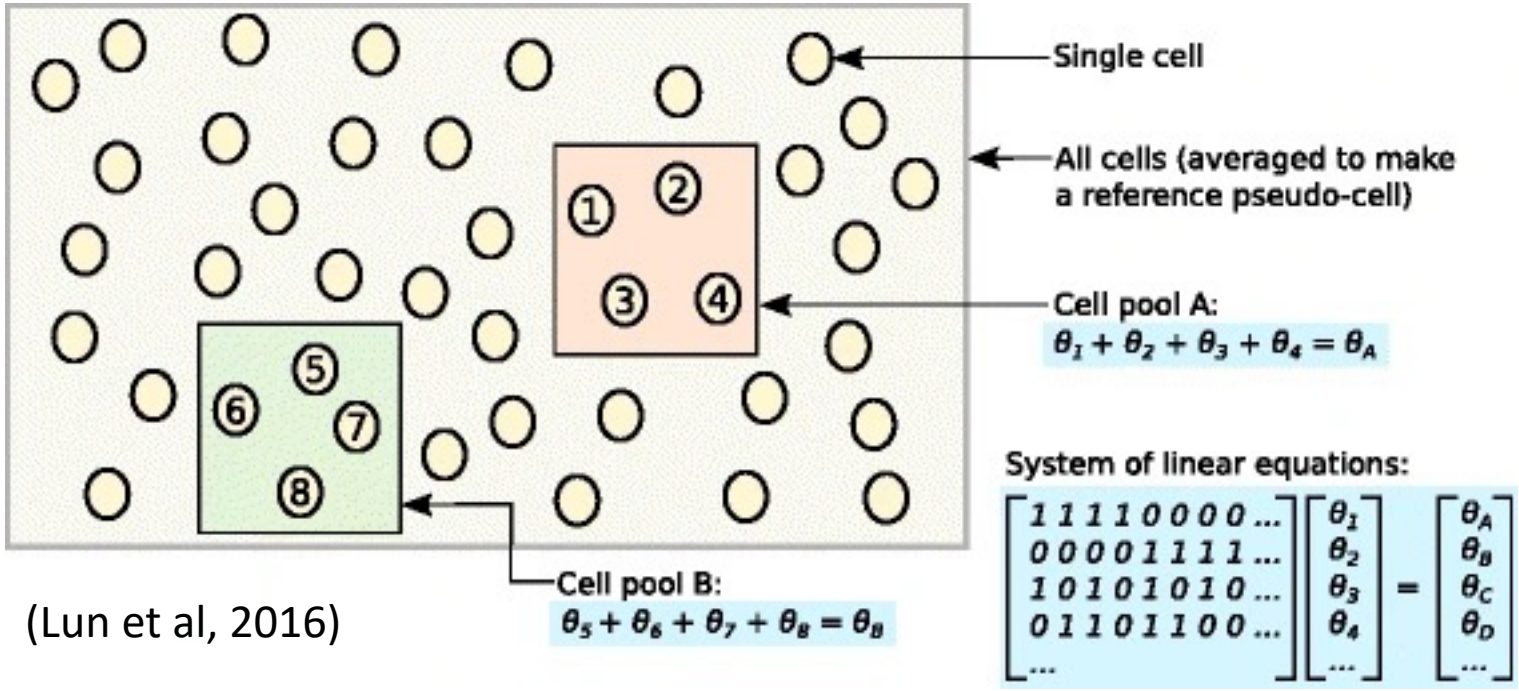
Three levels of single cell data normalization

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 normalisation: a pooling strategy to solve zero inflation



(Lun et al, 2016)

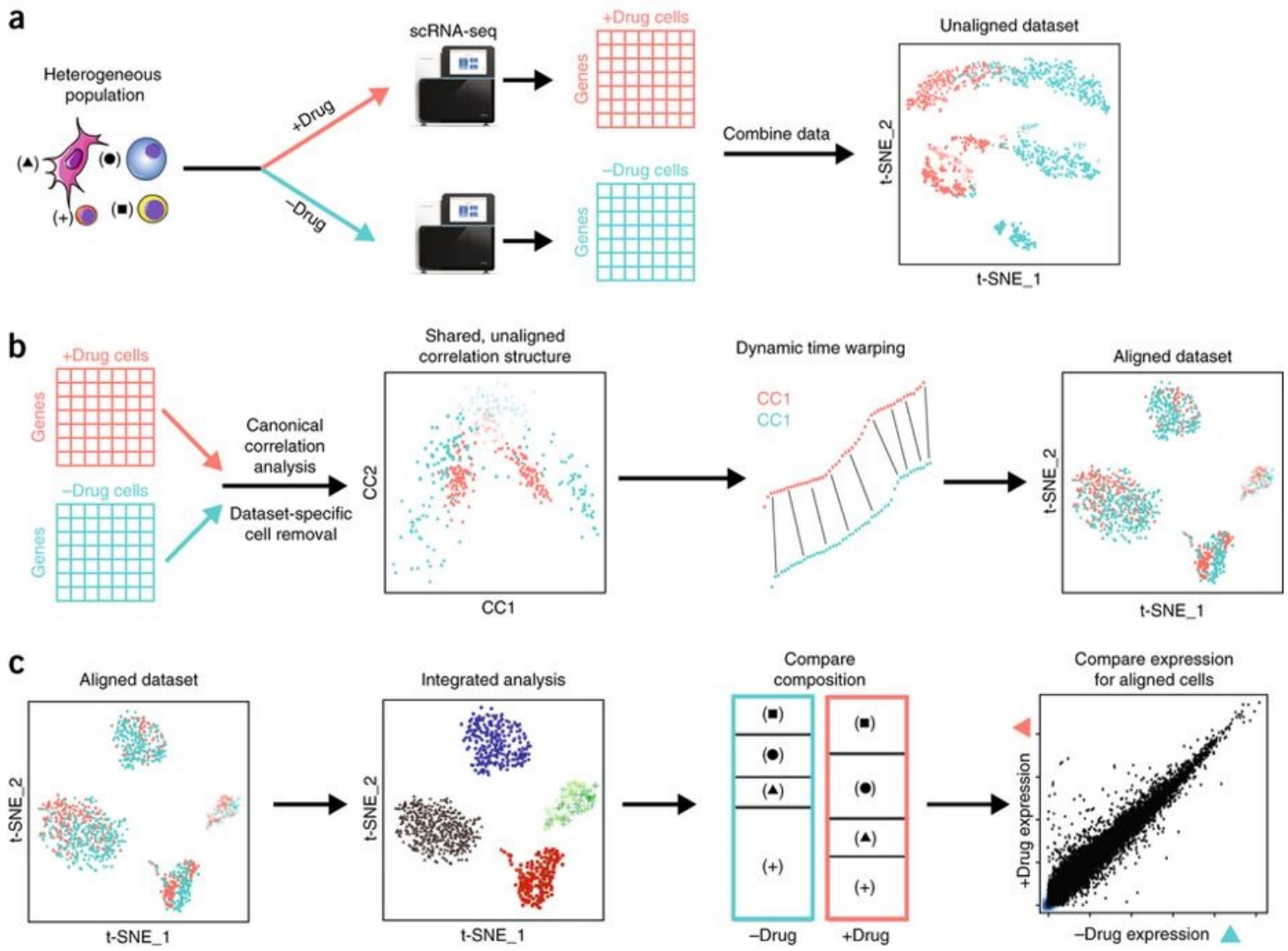
$$E(V_{ik}) = \lambda_{i0} \sum_{\mathrm{j} \epsilon \mathrm{S_k}} \theta_j \times t_j^{-1}$$

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: Canonical correlation analysis (CCA)

- CCA finds projection vectors u and v such that the correlation between the two datasets u^TX and v^TY 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

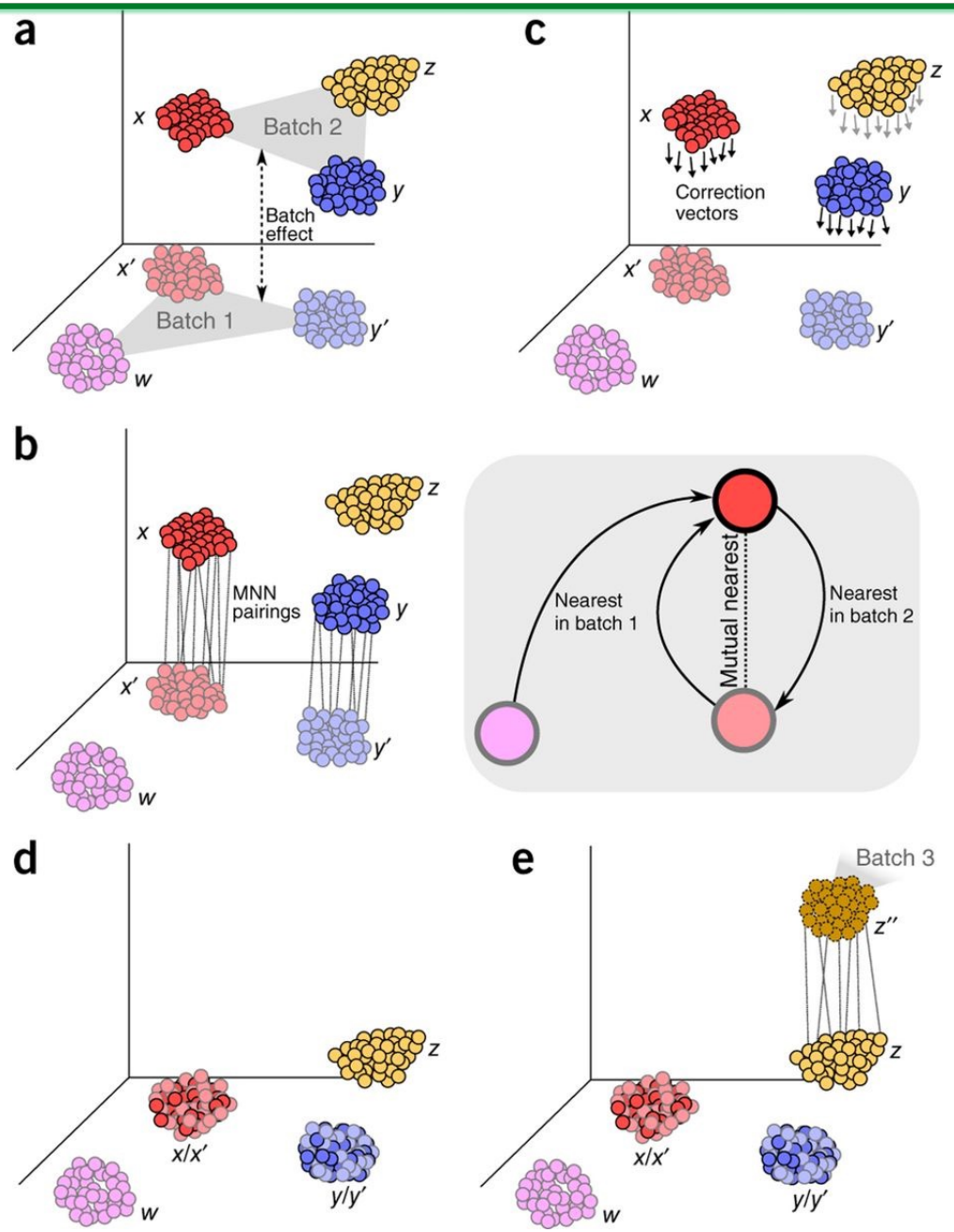


(Butler et al, 2018)

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

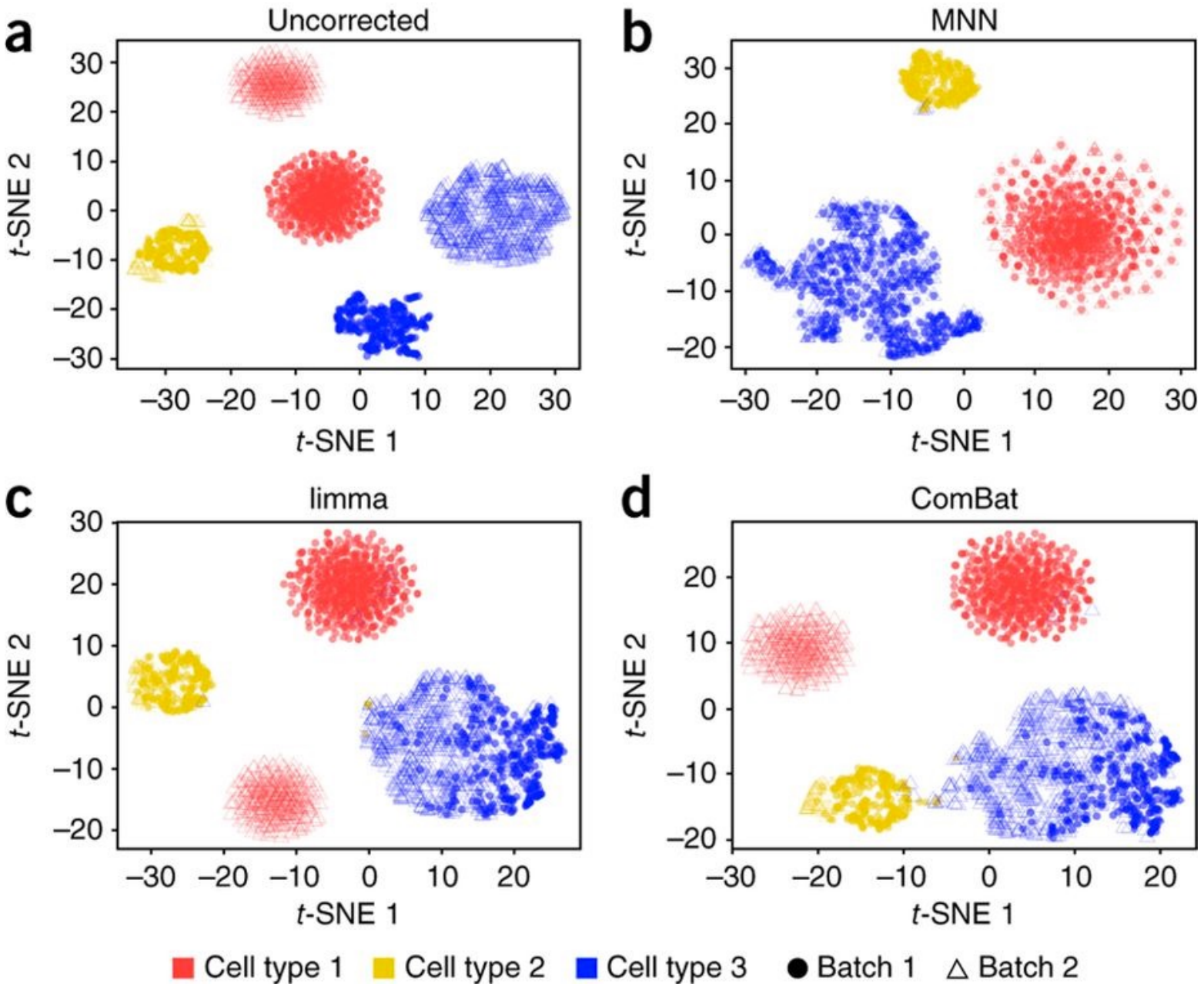


the cosine normalization

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

(Haghverdi et al, 2018)

Batch normalisation: Mutual Nearest Neighbour (MNN)



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

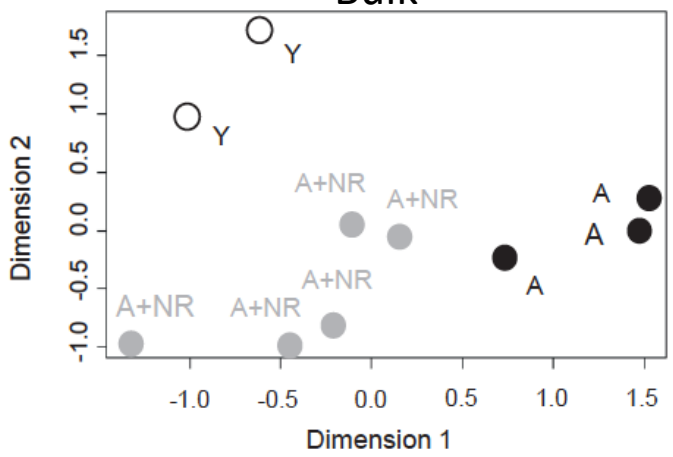
Linear approaches:

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

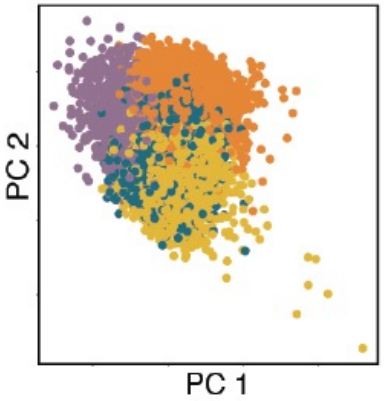
Linear approaches:

- 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

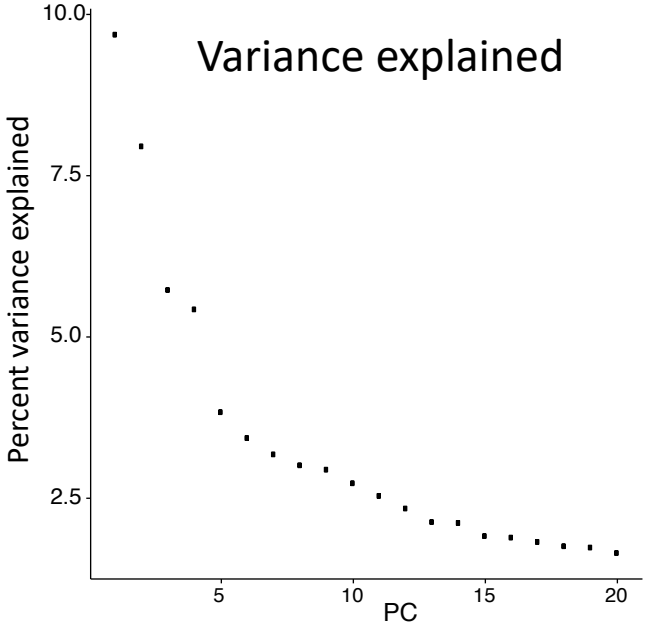
Bulk



Single cell



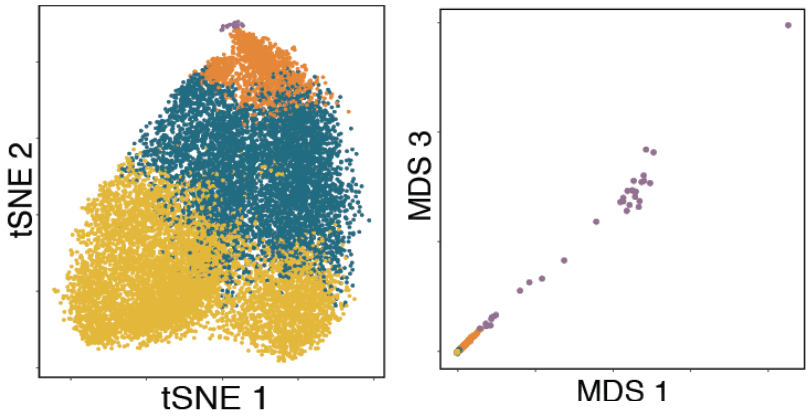
Variance explained

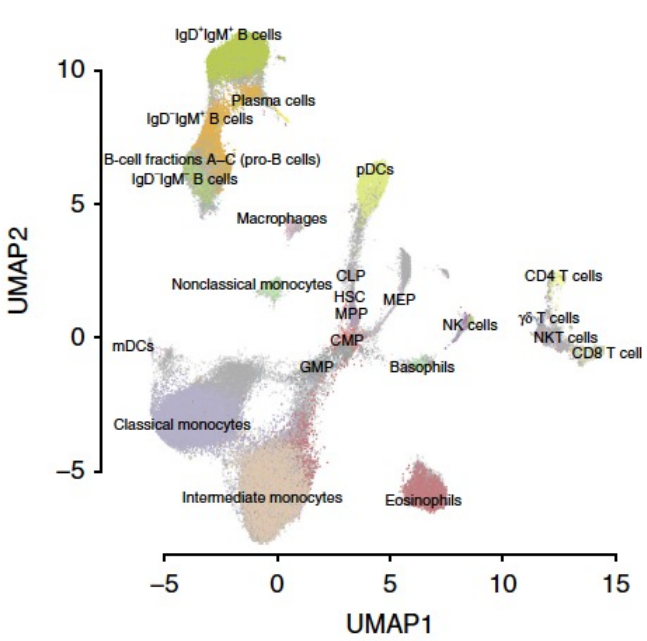


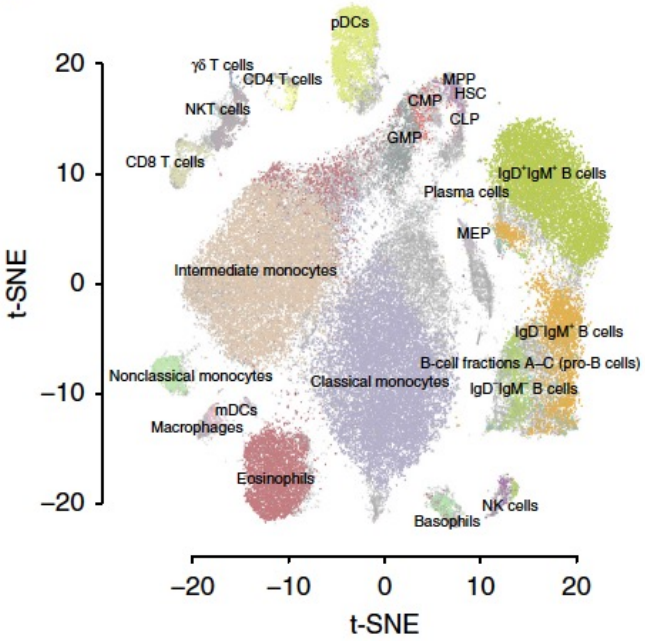
Dimensionality reduction: nonlinear techniques

- 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

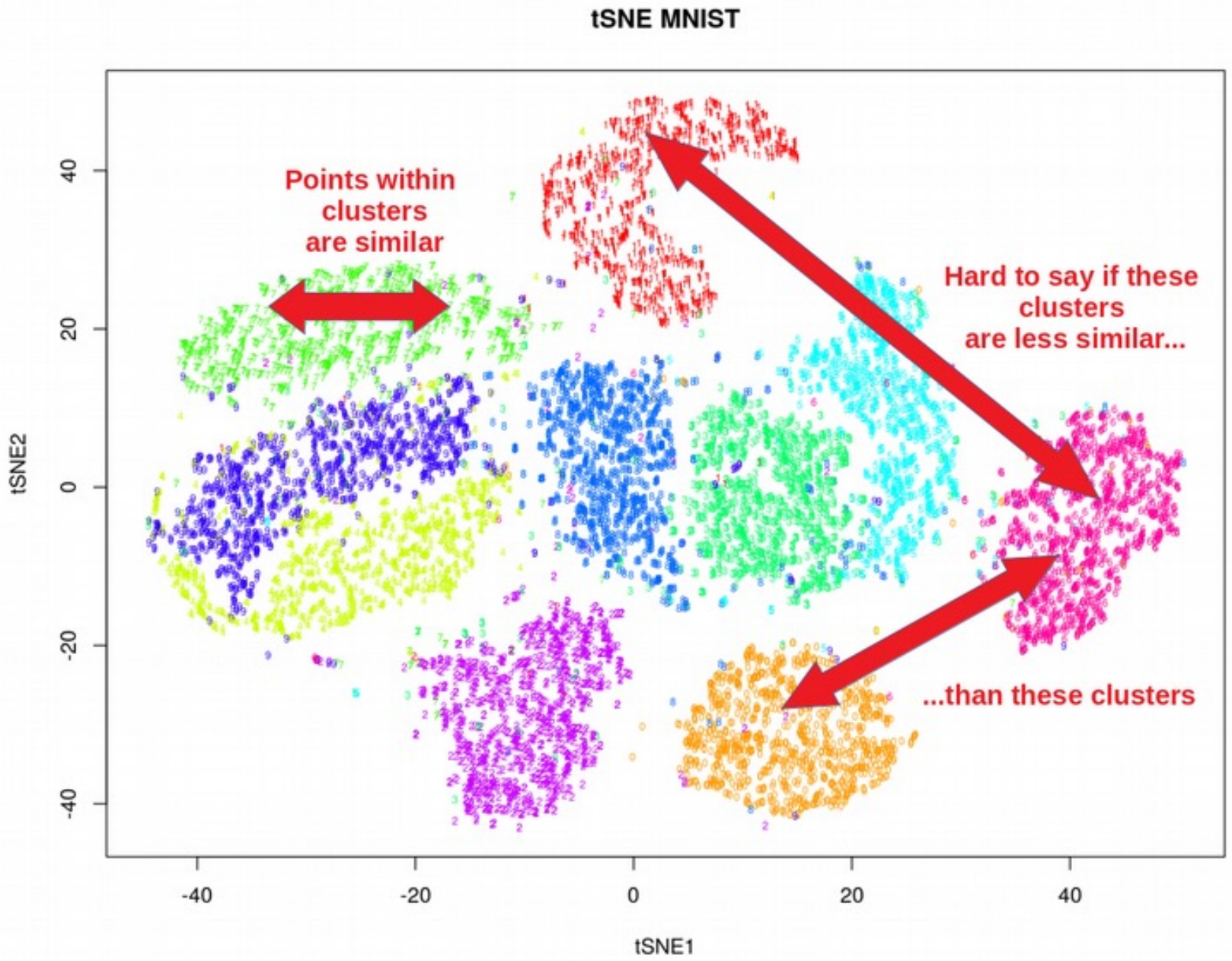
Overcrowding







Global vs local distance in low dimensional space

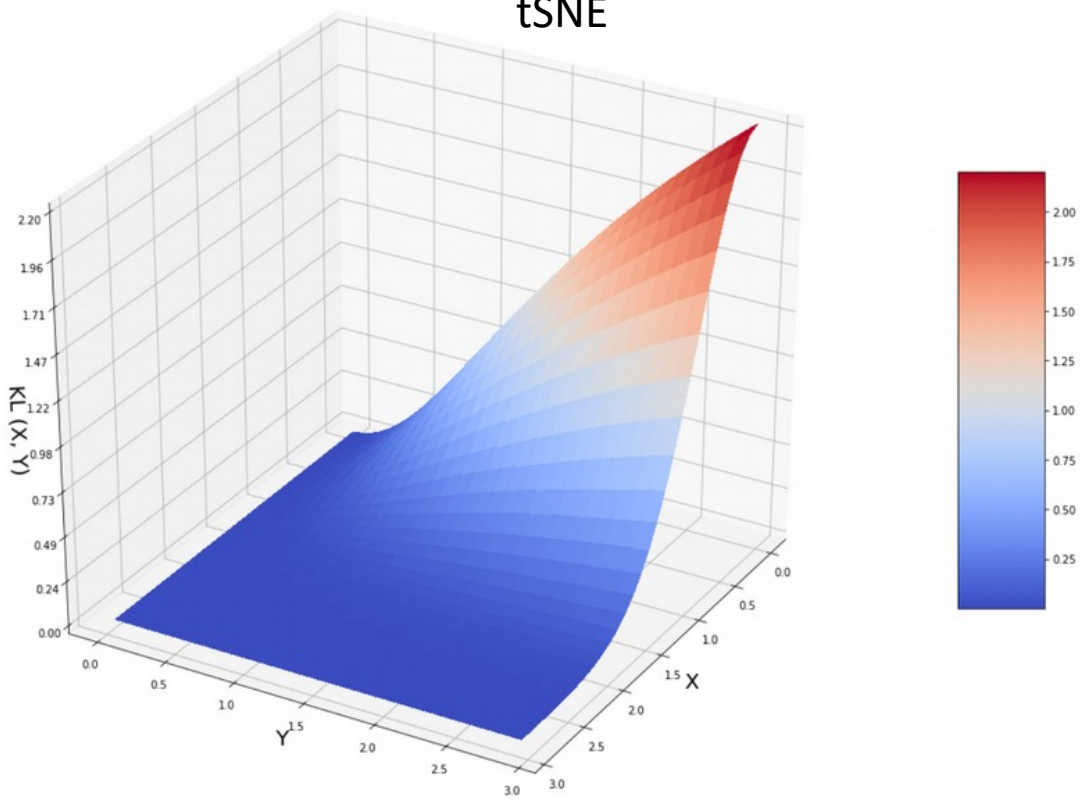


(Oskolkov N, 2019)

tSNE does not preserve long distance - KL divergence

(Oskolkov N, 2019)

tSNE



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

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

- 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:

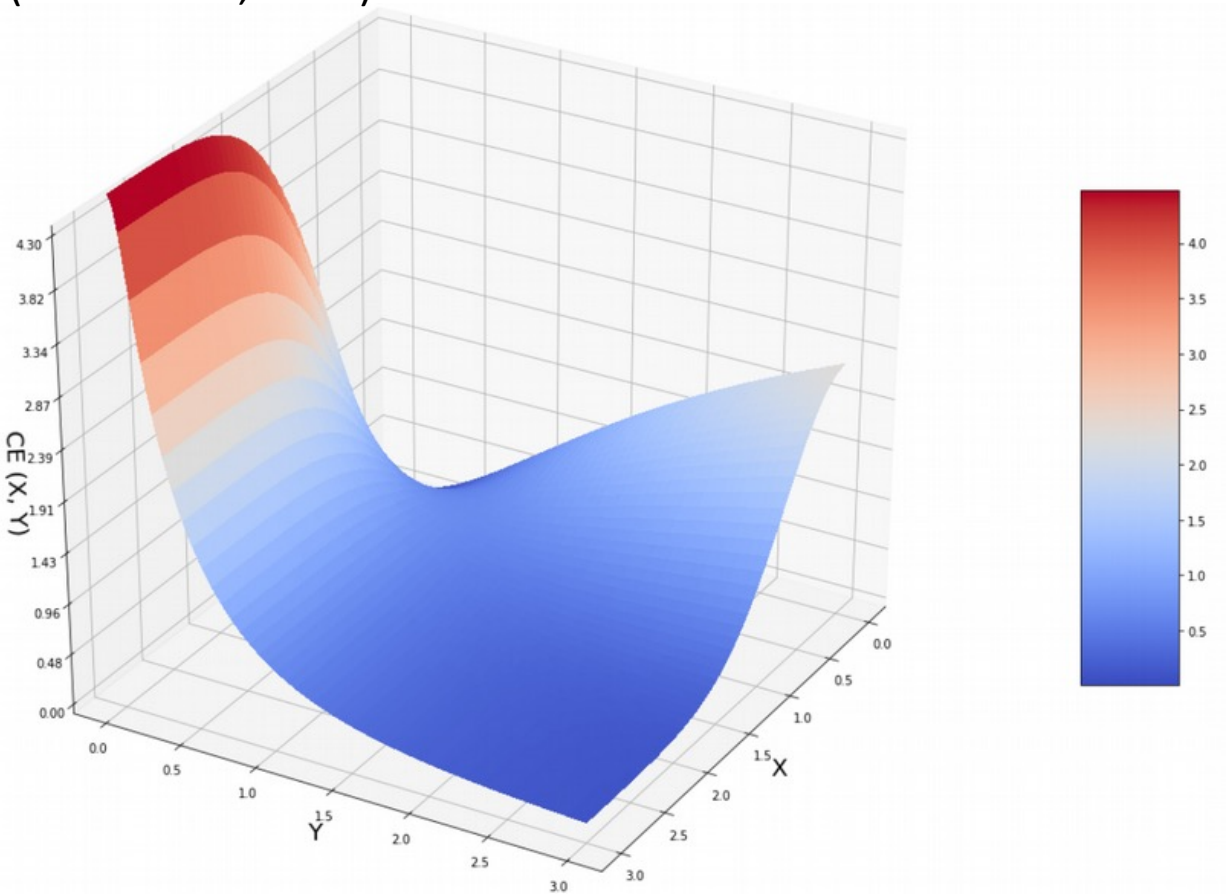
$$p_{j|i} = \frac{exp\left(\frac{-d\left(C_i,C_j\right)^2}{2\sigma_i^2}\right)}{\Sigma_{k\neq i}\ exp\left(\frac{-d(C_i,C_k)^2}{2\sigma_i^2}\right)}$$

- 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: $q_{ij} = \frac{\left(1+\|y_i - y_j\|^2\right)^{-1}}{\sum_{k\neq m}(1+\|y_k - y_m\|^2)^{-1}}$, 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**

UMAP preserves long distance - cross entropy

(Oskolkov N, 2019)

UMAP



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

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

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

When X large, Y is also large to minimize CE

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

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

$$\approx e^{-X^2}\log\left(1+Y^2\right) + \left(1-e^{-X^2}\right)\log\left(\frac{1+Y^2}{Y^2}\right)$$

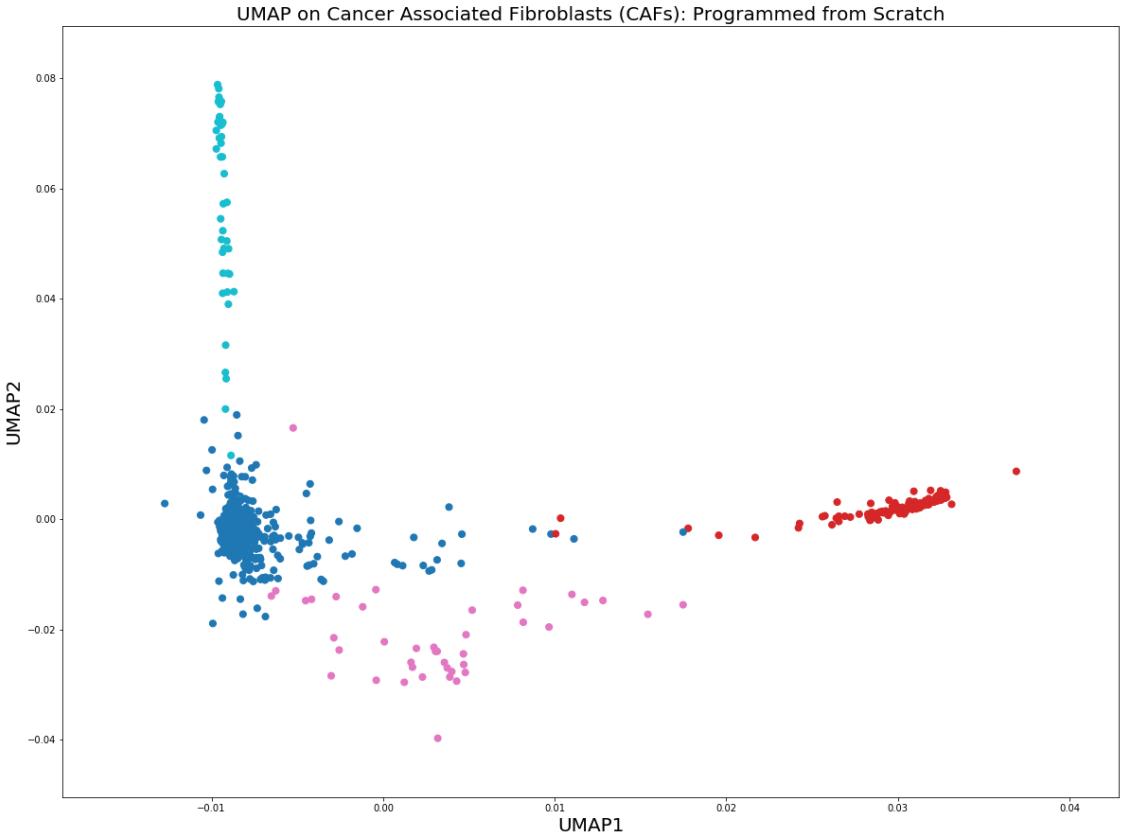
tSNE: $KL(X,Y) \approx -P(X)\log Q(Y) = e^{-X^2}\log\left(1+Y^2\right)$

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.

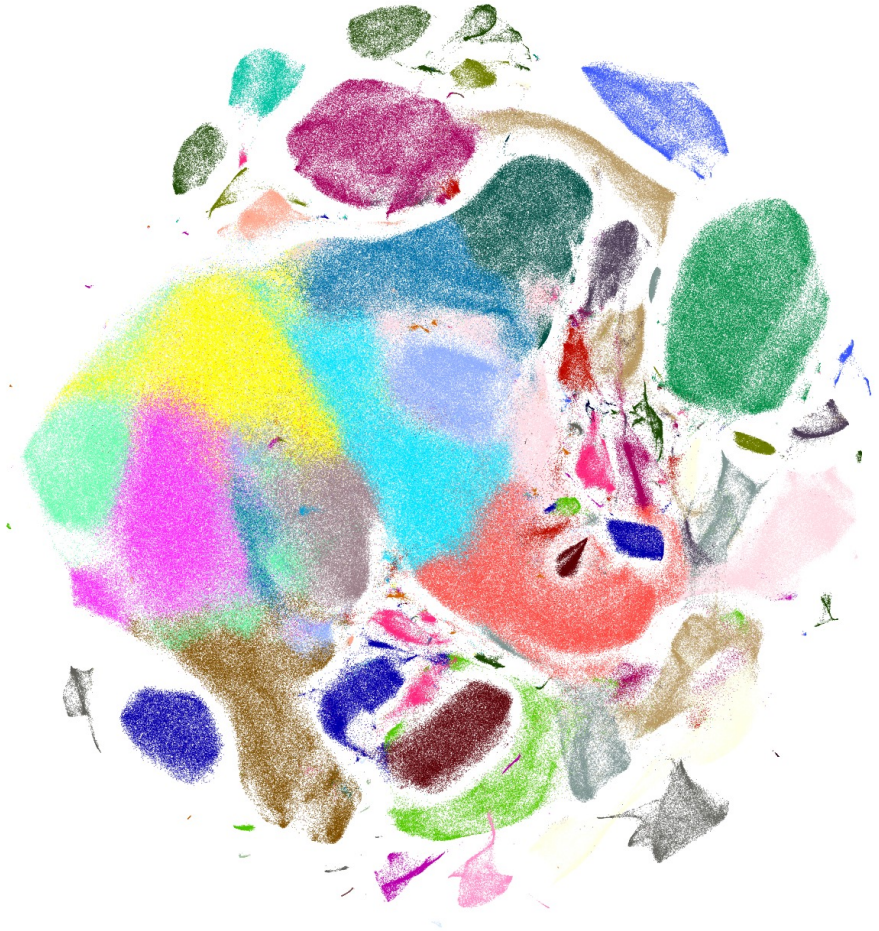
- UMAP has no computational restrictions on embedding dimension, making it viable as a general-purpose dimension reduction technique for machine learning (tSNE can only embed to 2-3 dimensions)



(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

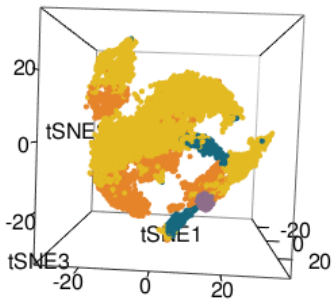
An example iPSC scRNA dataset:

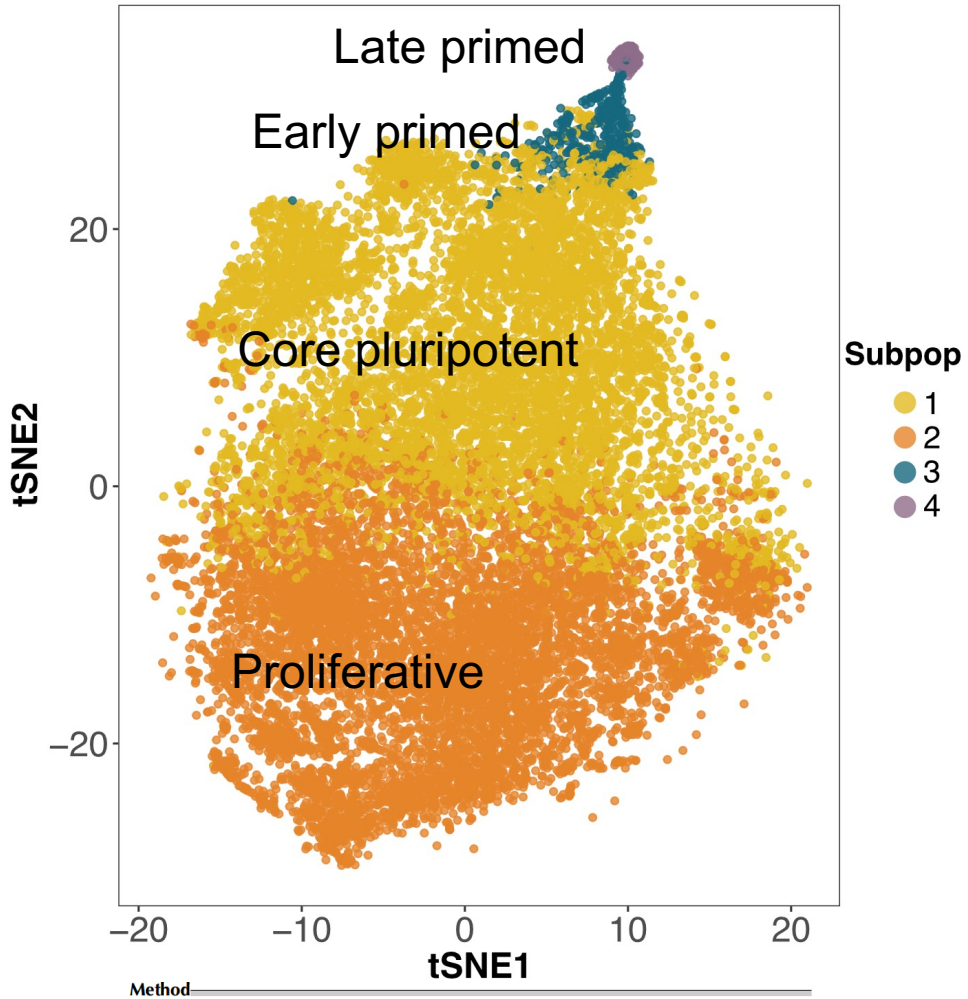
- Sequenced > 18,000 cells (10x Genomics)
- Detected > 16,000 genes
- We proved that a seemingly homogeneous hiPSC population contains 4 subpopulations

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

.....





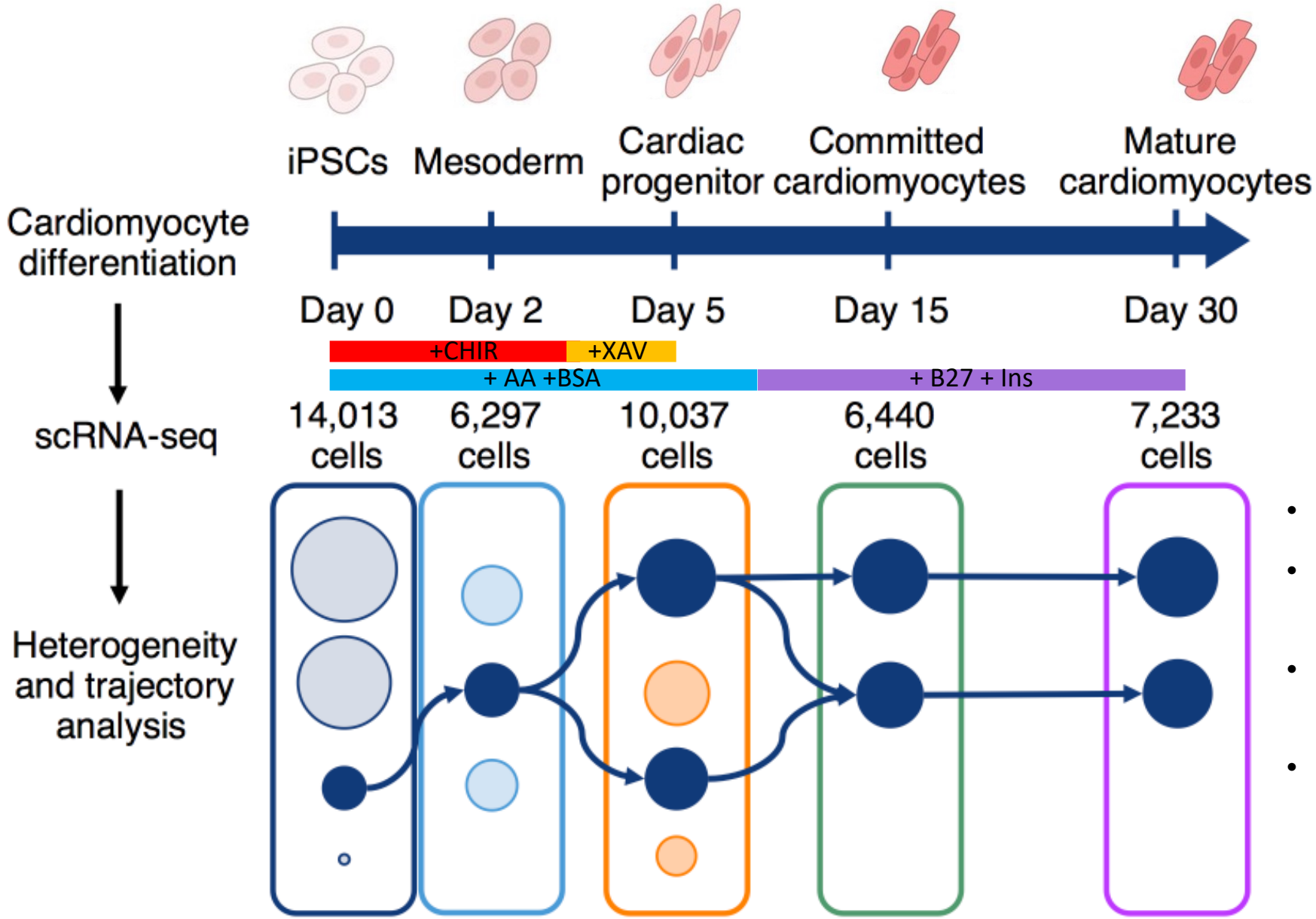
Method

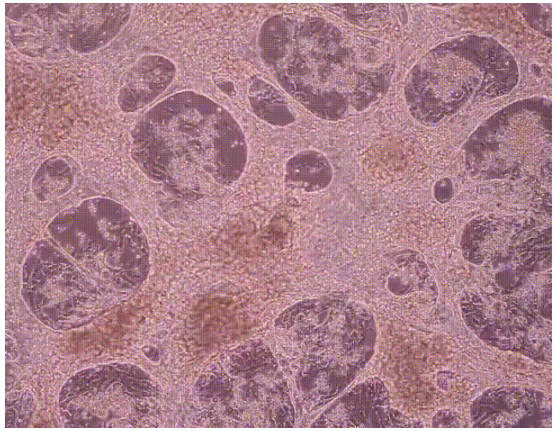
Single-cell RNA-seq of human induced pluripotent stem cells reveals cellular heterogeneity and cell state transitions between subpopulations

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]

GENOME RESEARCH

Clustering to assess cell-type specific responses





(Fei Pei et al., 2017)

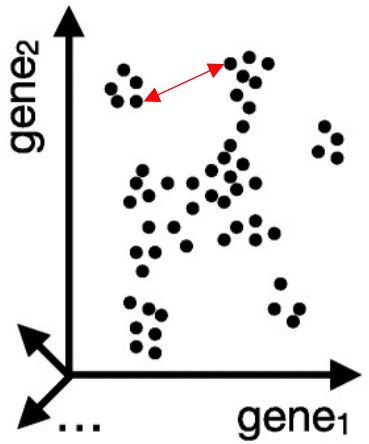
Question: differential responses at the subpopulation levels?

- 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

- 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	$d_{ij} = \sum_{g=1}^{G} \mid x_{ig} - x_{jg} \mid$
Maximum distance	$d_{ij} = \max_{g} \mid x_{ig} - x_{jg} \mid$



cells in gene expression space

Cluster cells in expression space - Distance measures

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

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 (**S**table **C**lustering at **O**ptimal **RE**solution):

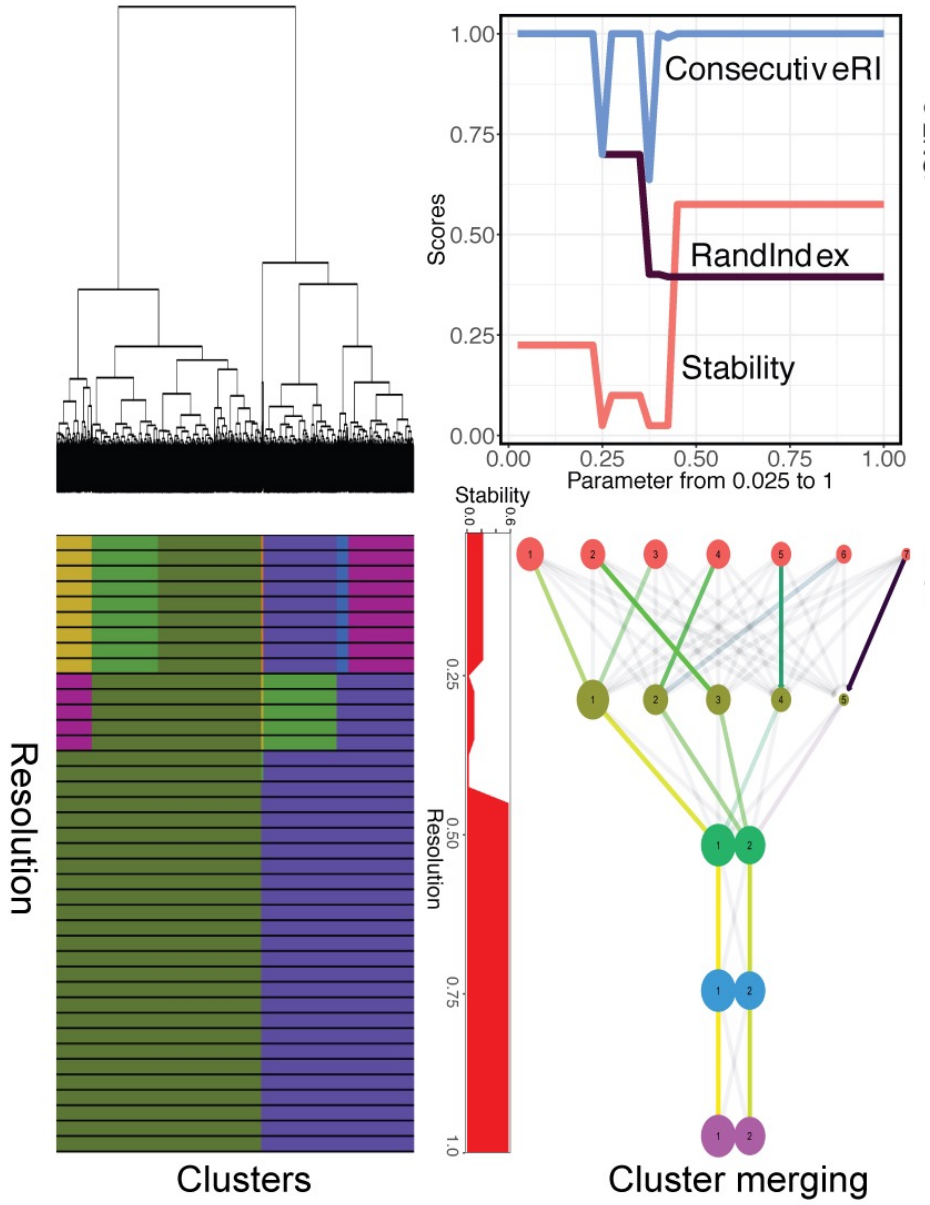
We improved HC clustering by first selecting for an optimal cluster resolution by implementing the following algorithm:

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 ($W_1 = [1, 1]$).

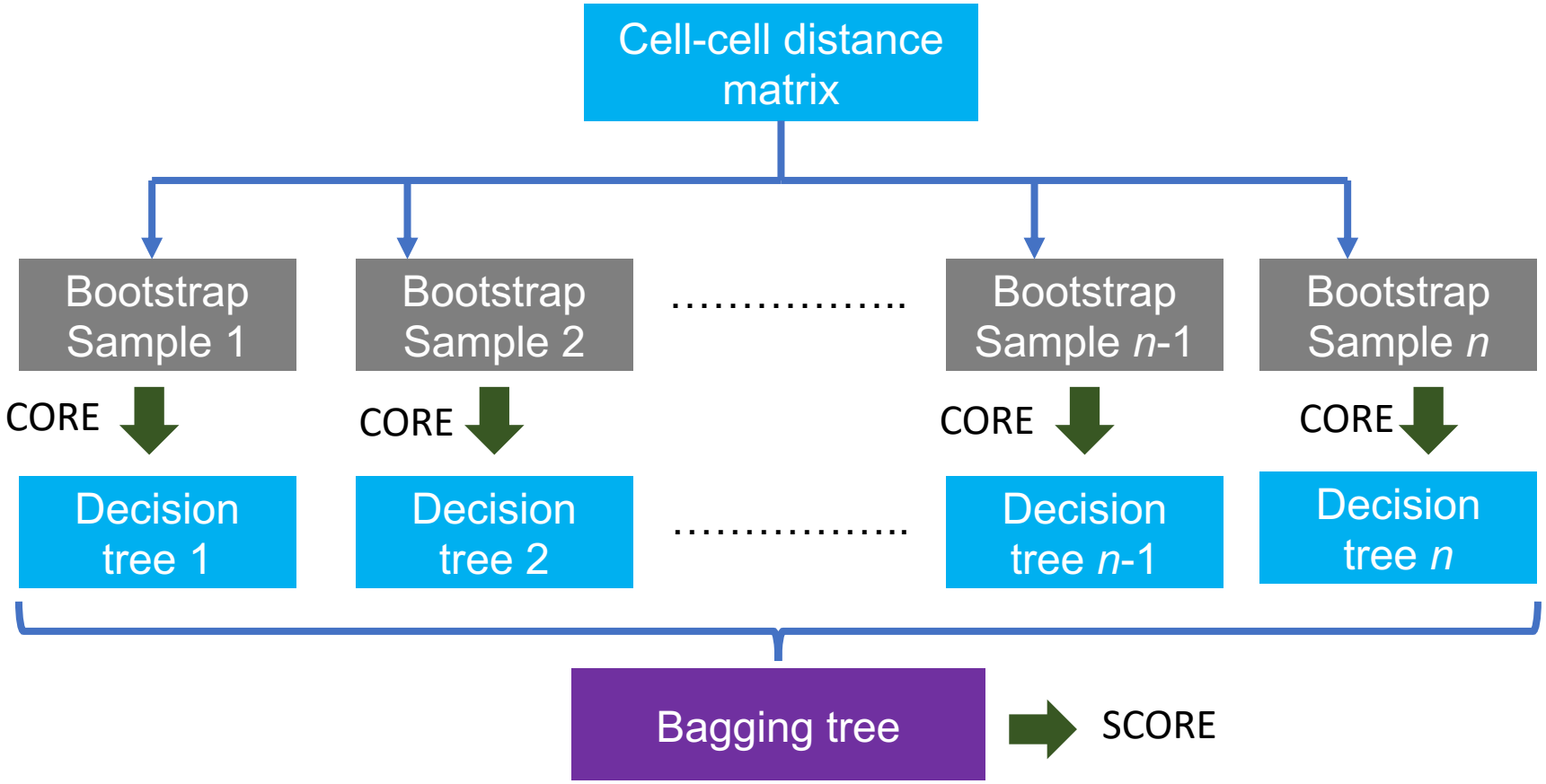
2. Compute pairwise adjusted Rand index (AR_i) for every 2 consecutive windows (W_i and W_{i+1} for integers $i \epsilon [1, 39]$)

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)



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

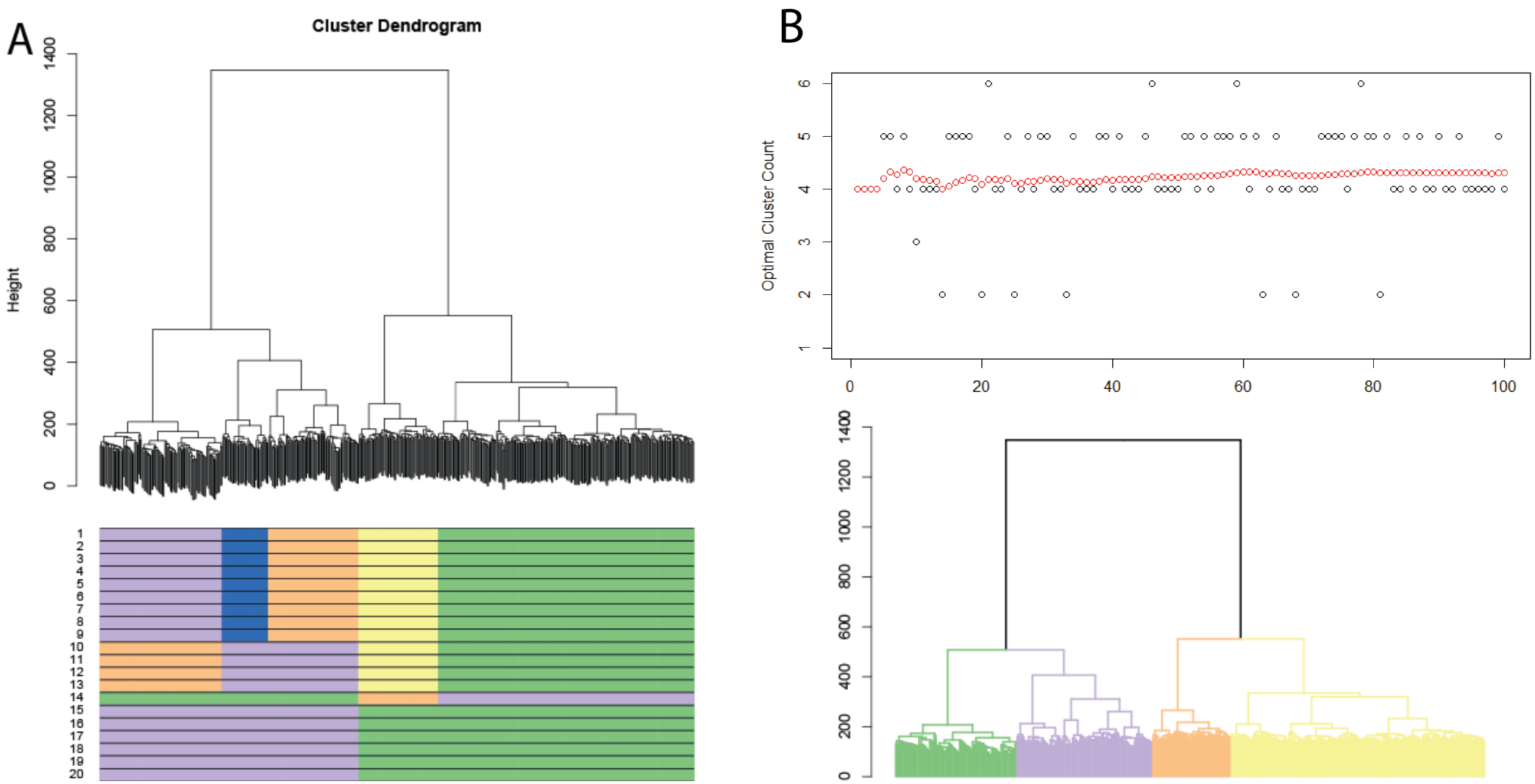
1. Bagging strategies are used for re-clusteing random sub-sets of cells from the population to generate additional dendrogram trees.

2. For each bagging run, choose a vector $\mathbf{b}_k$(k= 1,2,...,m) of length p∗dim(C) (p≤1) containing a sample, with replacement, from set C and create a new matrix N_k, of Euclidean distances for the cells in $\mathbf{b_k}$.

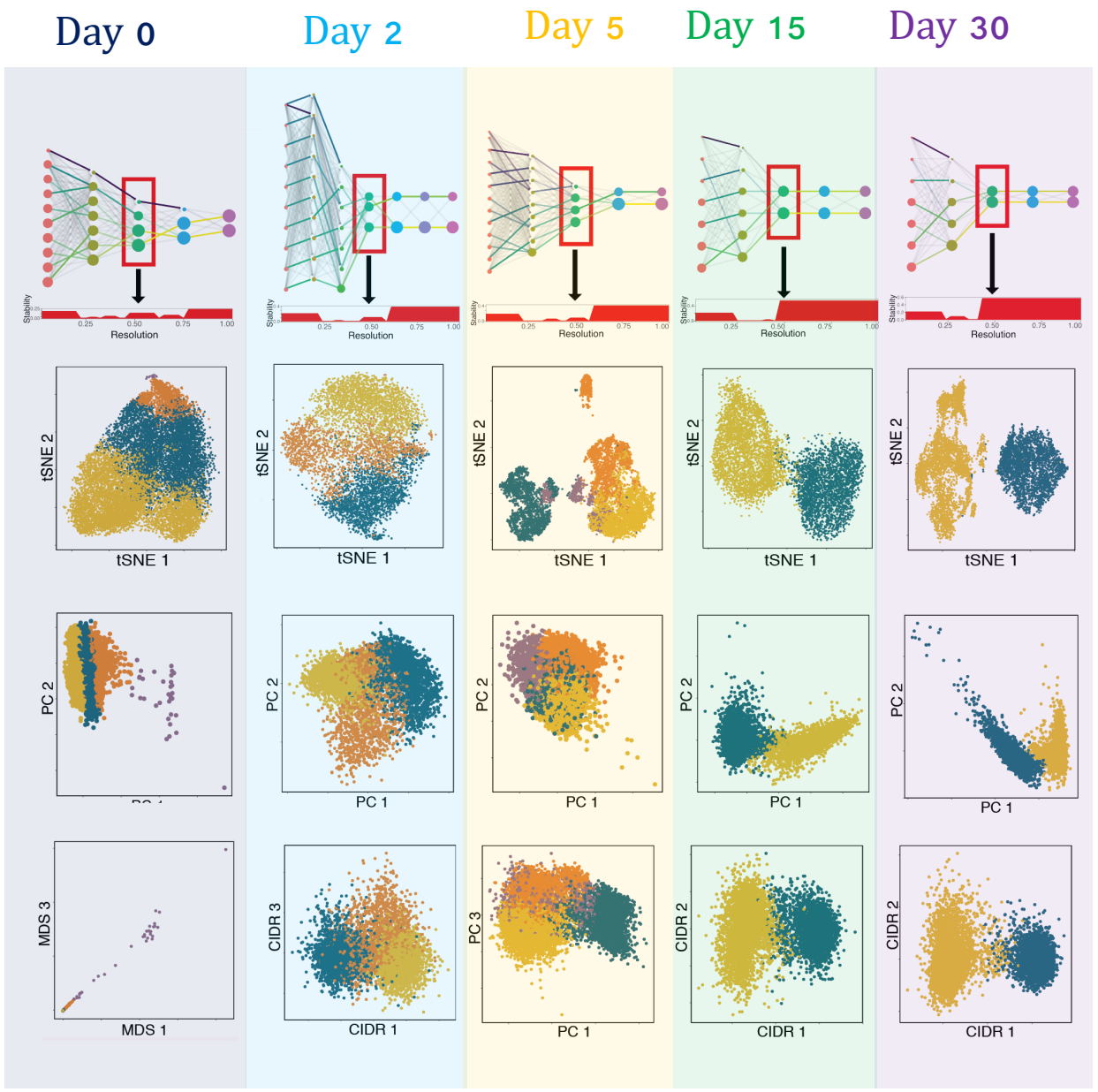
3. For each N_k, a new dendrogram tree is generated and clustered, then an optimal stability is computed.

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 Commited 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

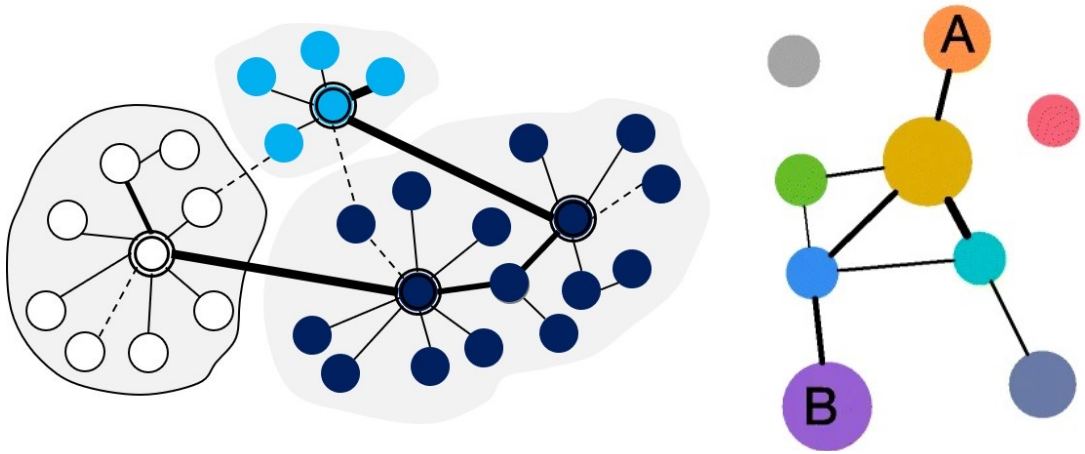
- From a mixed population at each time point, CORE identified 2 to 4 homogenous clusters
- The identified subpopulations were confirmed by independent methods: PCA, MDS, tSNE, CIDR
- The subpopulations are biologically distinct

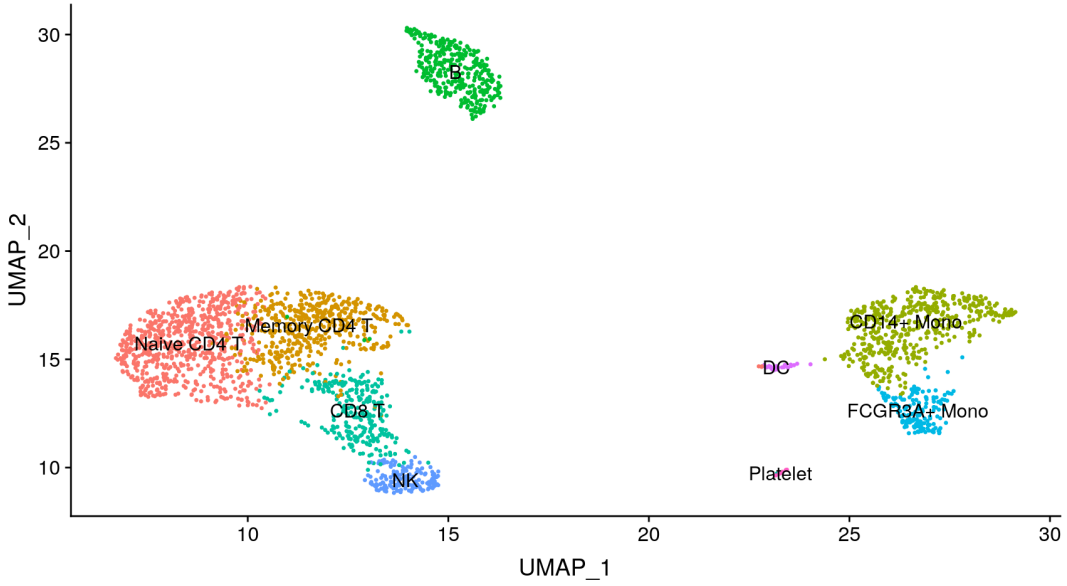
Graph-based Clustering

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)

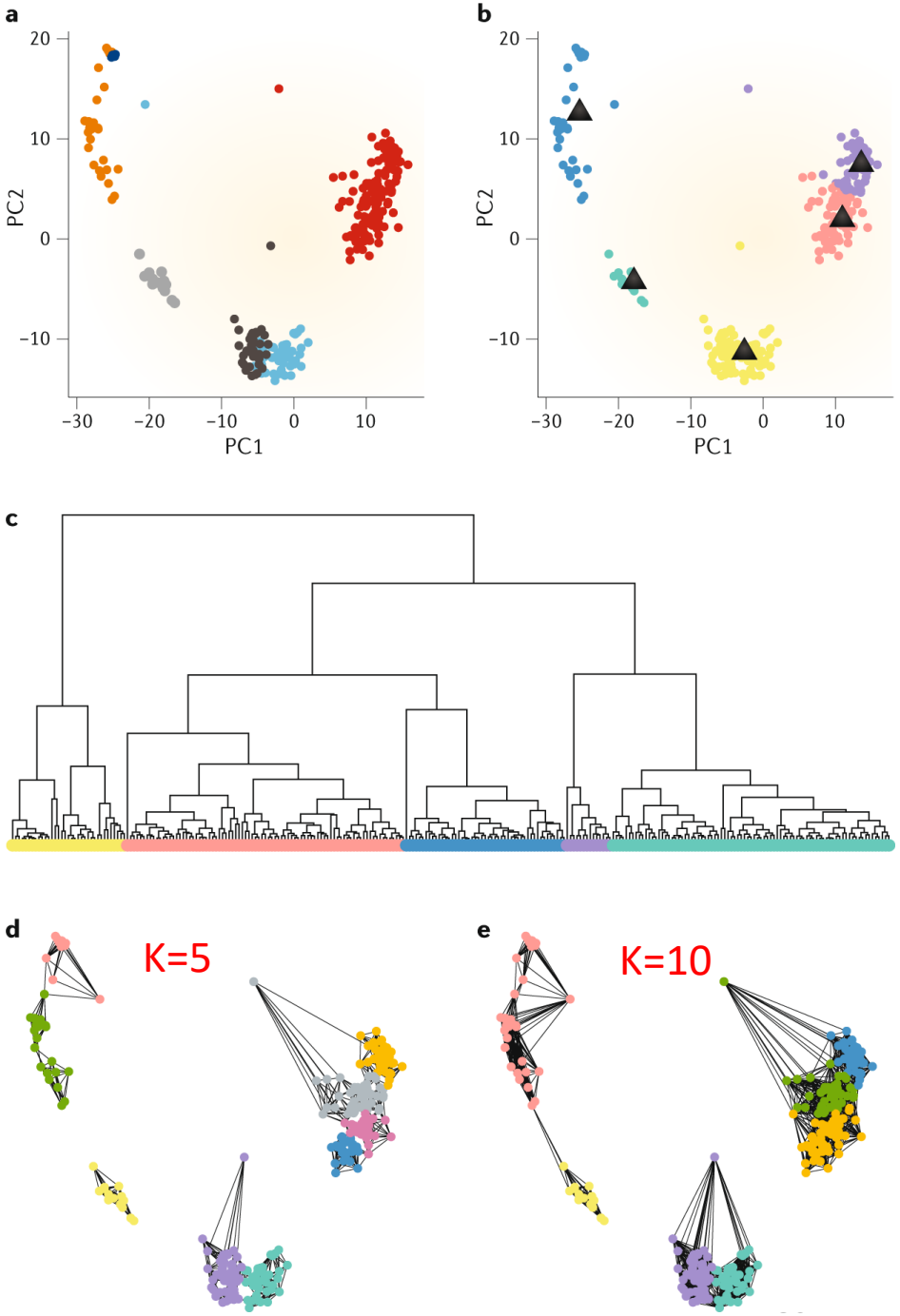
2) Community detection to partition cells in graph into groups of cells
   - Modularity optimization techniques such as the Louvain algorithm
   - Modularity: measures the density of edges inside communities to edges outside communities
   - Louvain iteratively groups cells together, with the goal of optimizing the standard modularity function





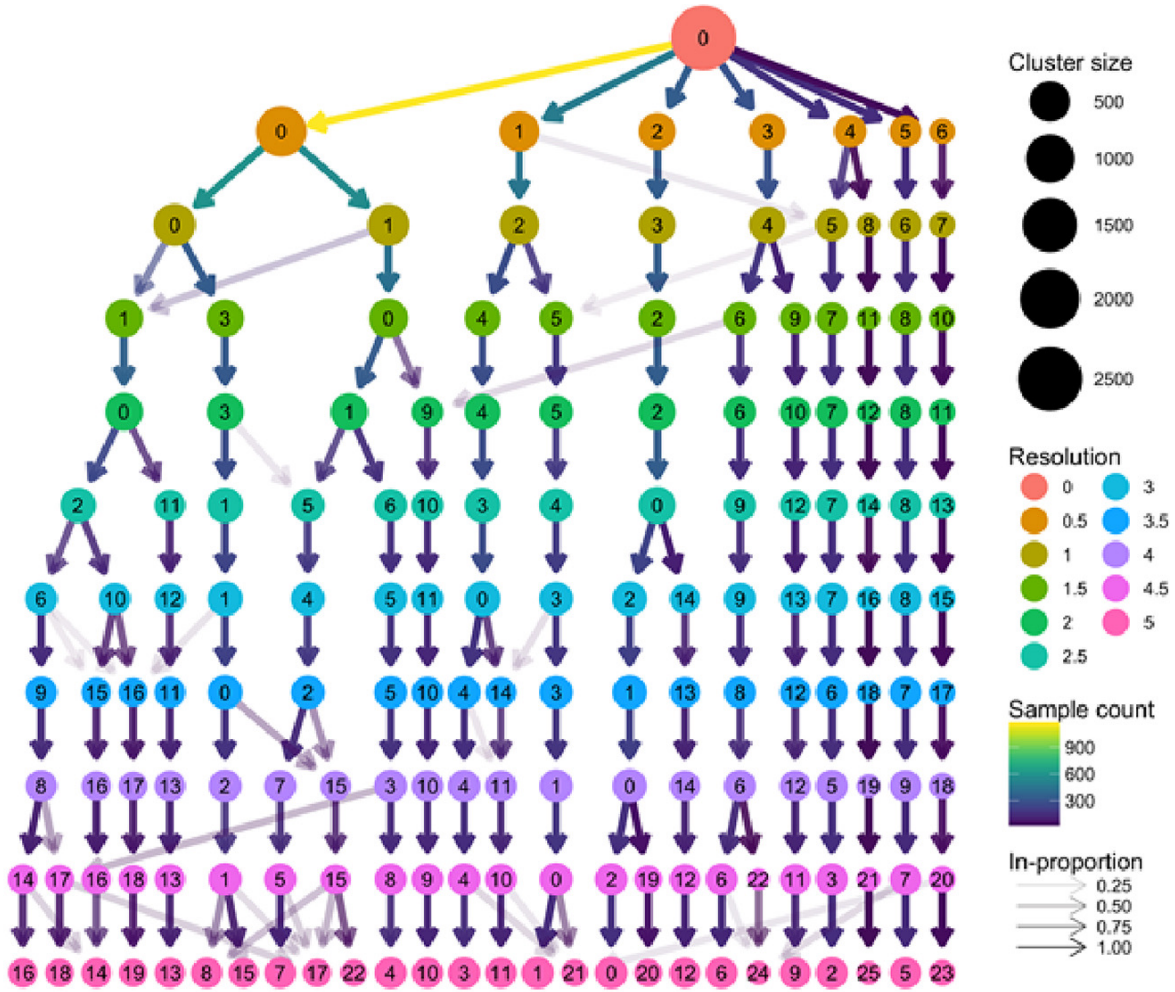
Graph-based Clustering

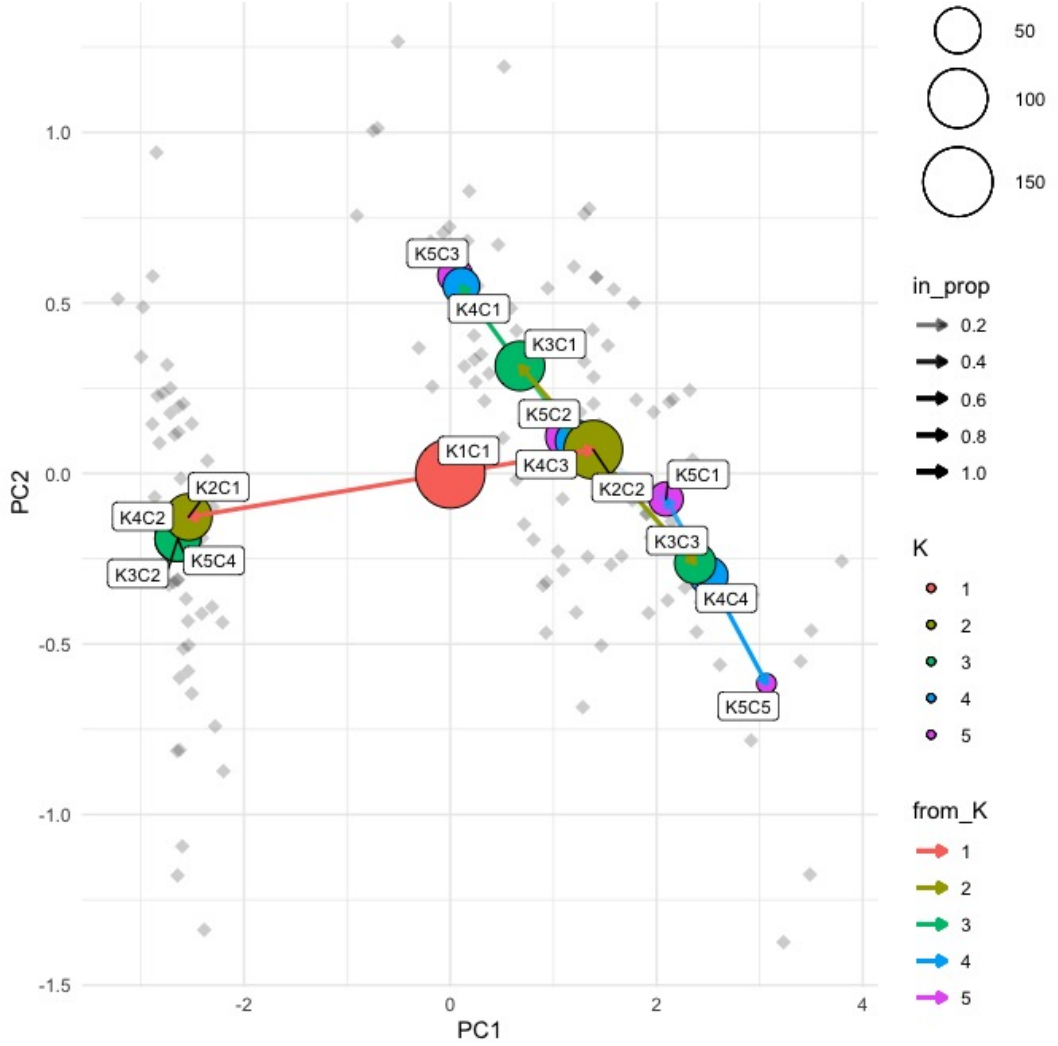
- Build shared-nearest-neighbour graph connecting the cells and finds tightly connected communities
- Increasing the number of neighbours when constructing the cell–cell graph indirectly decreases the resolution of graph-based clustering



Nature Reviews Genetics, 20, (2019)

Visualise clustering results





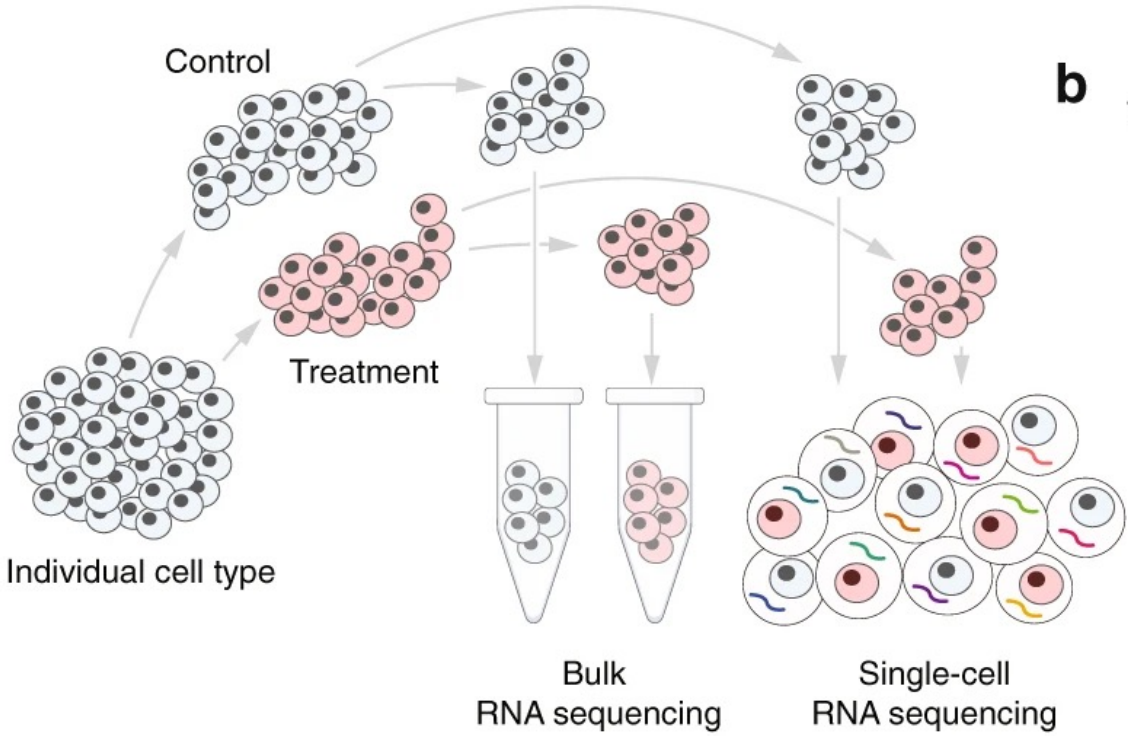
Statistical evaluation of clustering results

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

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

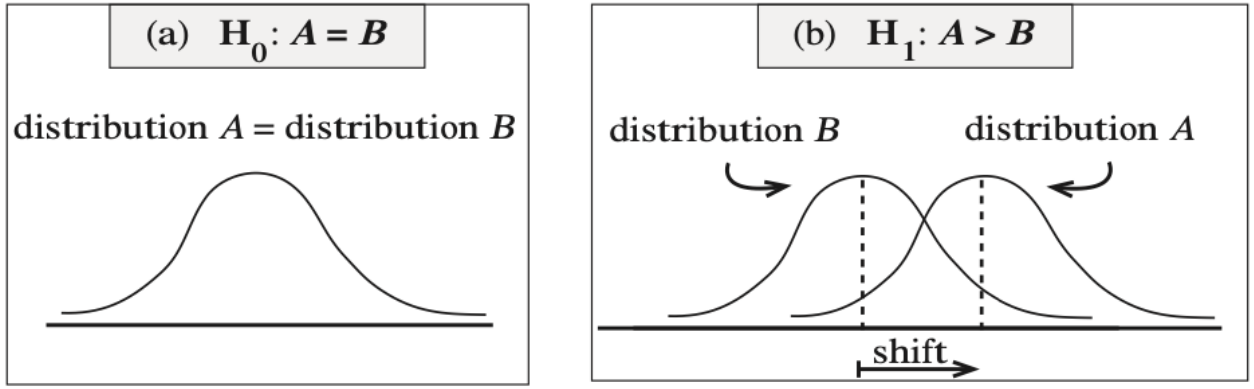

Control
b
Treatment
Individual cell type
Bulk
RNA sequencing
Single-cell
RNA sequencing

Three main categories

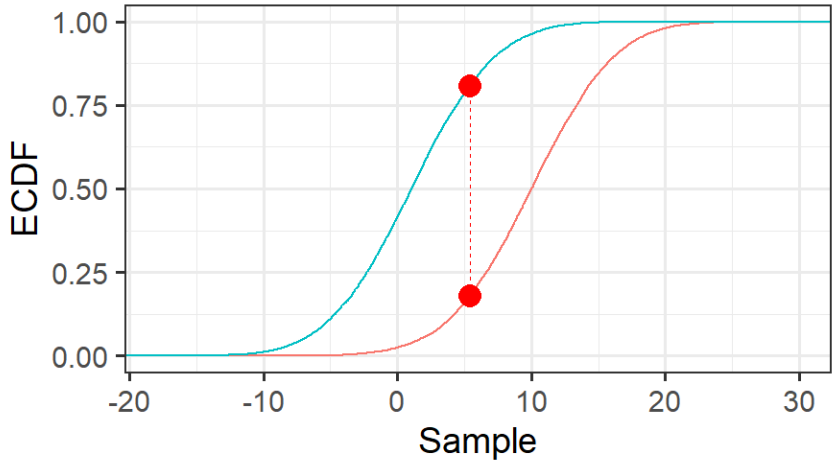
- Non-parametric tests
  - Wilcoxon rank-sum test, Kolmogorov–Smirnov (KS) test
  - Convert observed expression to ranks, then test whether the distribution of ranks for one group is significantly different from the other group
- Bulk RNA-seq based method
  - e.g edgeR DEseq2
- scRNA-seq specific methods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

Non-parametric tests

Wilcoxon rank-sum test



KS test



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 ith condition form the overall mean
- H_0: $\alpha_{treat,geneg} = \alpha_{control,geneg}$ no difference in treatment and control group

Assumption using log as link function: y_{igk}~ Poisson → mean= variance
However, often observe mean < variance → thus, Log-normal over correct data dispersion → y_{ijk}~ 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 φ_g=0 → NB becomes Poisson

Gamma-Poisson mixture

Biological variance ~ Gamma

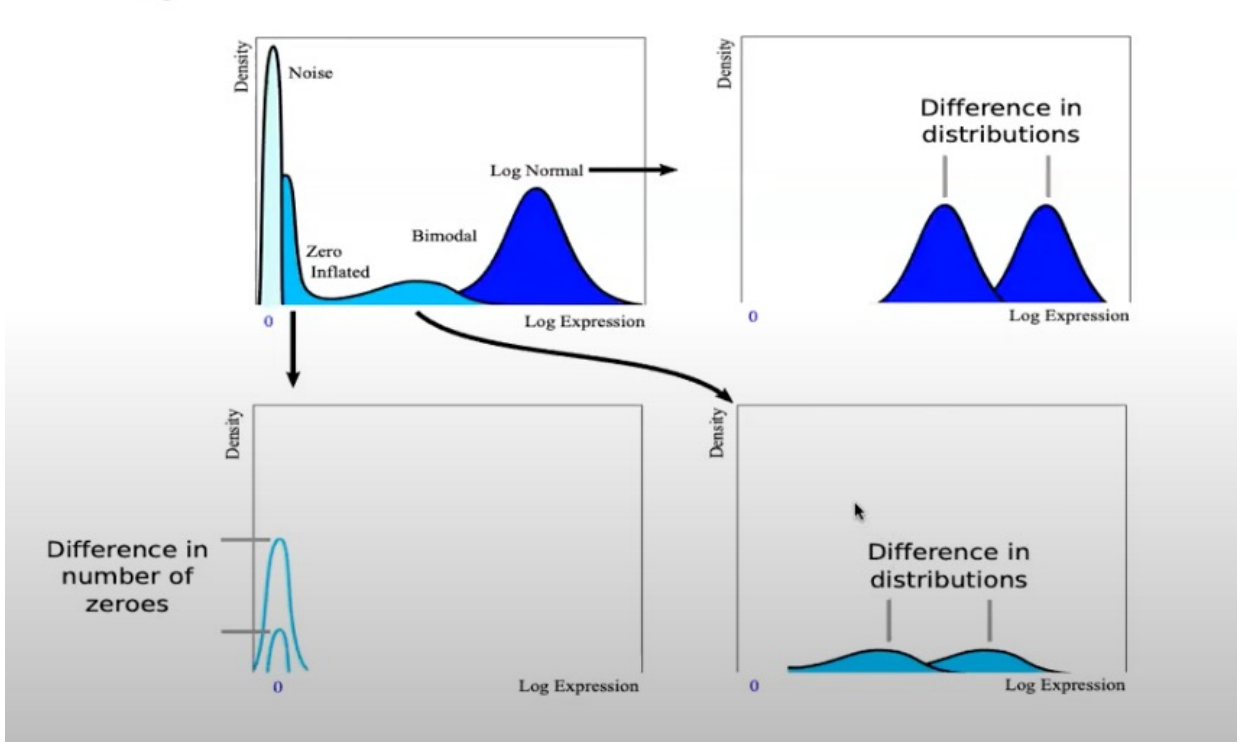
Measurement error ~ Poisson

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

MAST

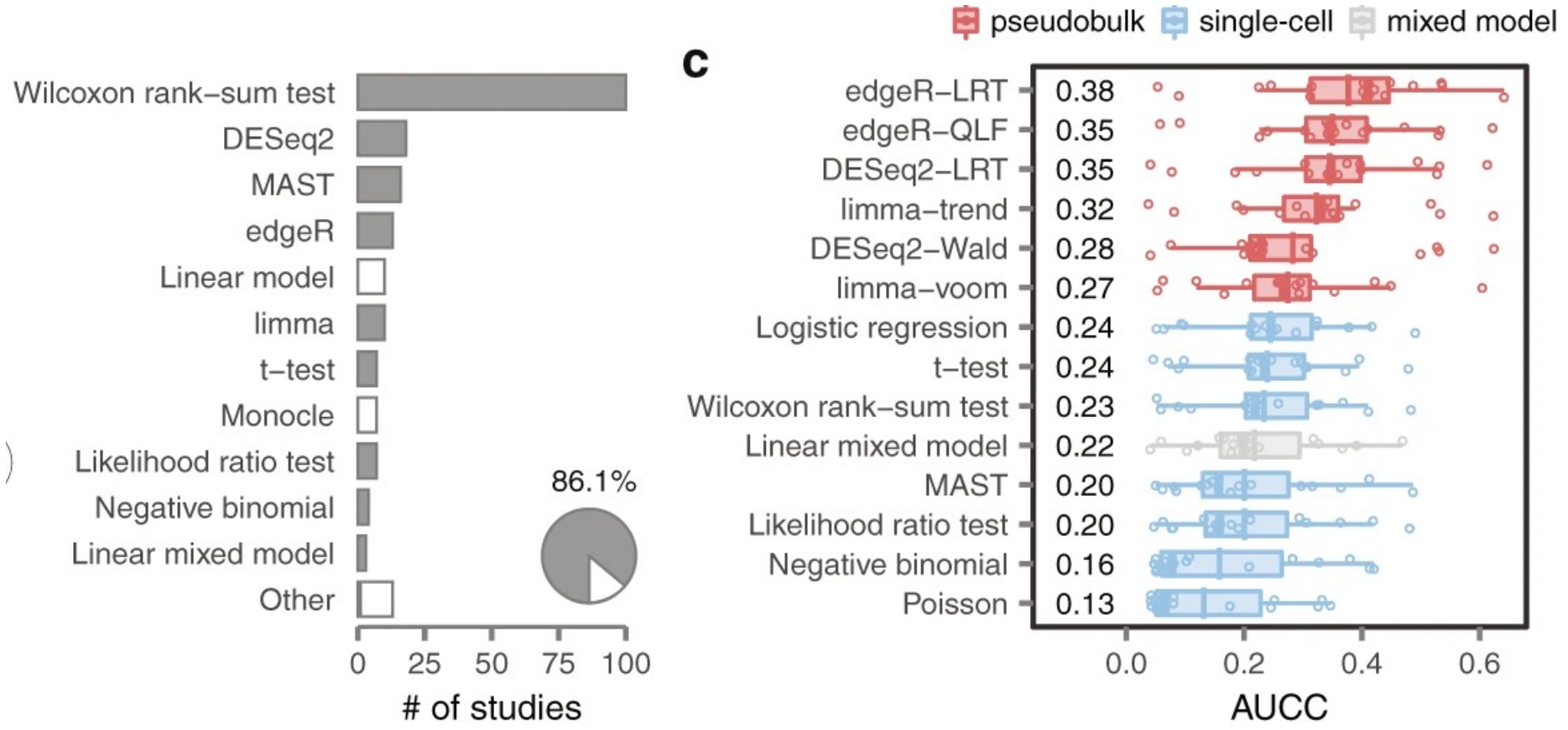
Hurdle model

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



From Xiaole Liu 2020 STAT115

Comparison between different methods



Squair et al, NC, 2021

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)).

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

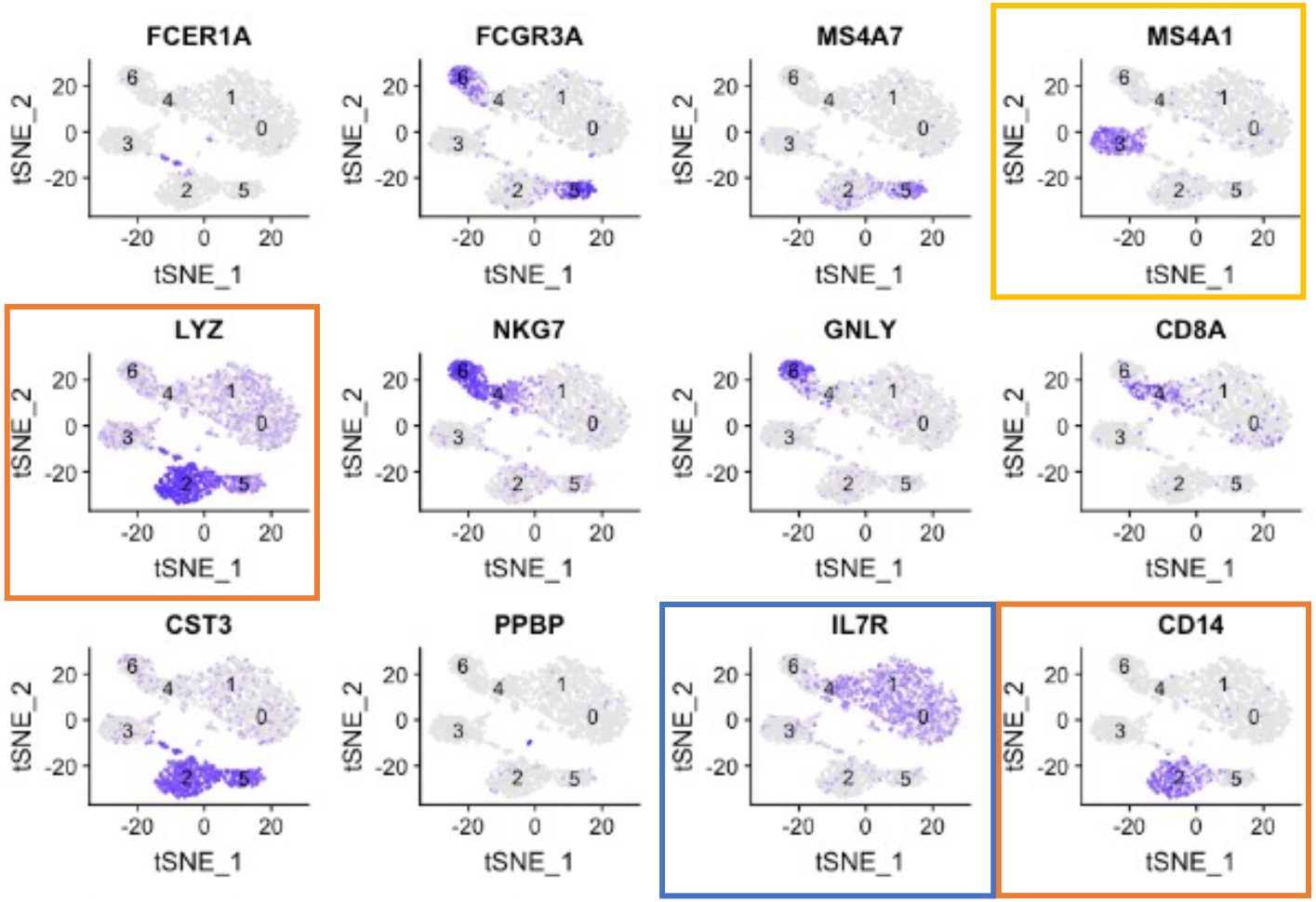
Unsupervised

- 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.

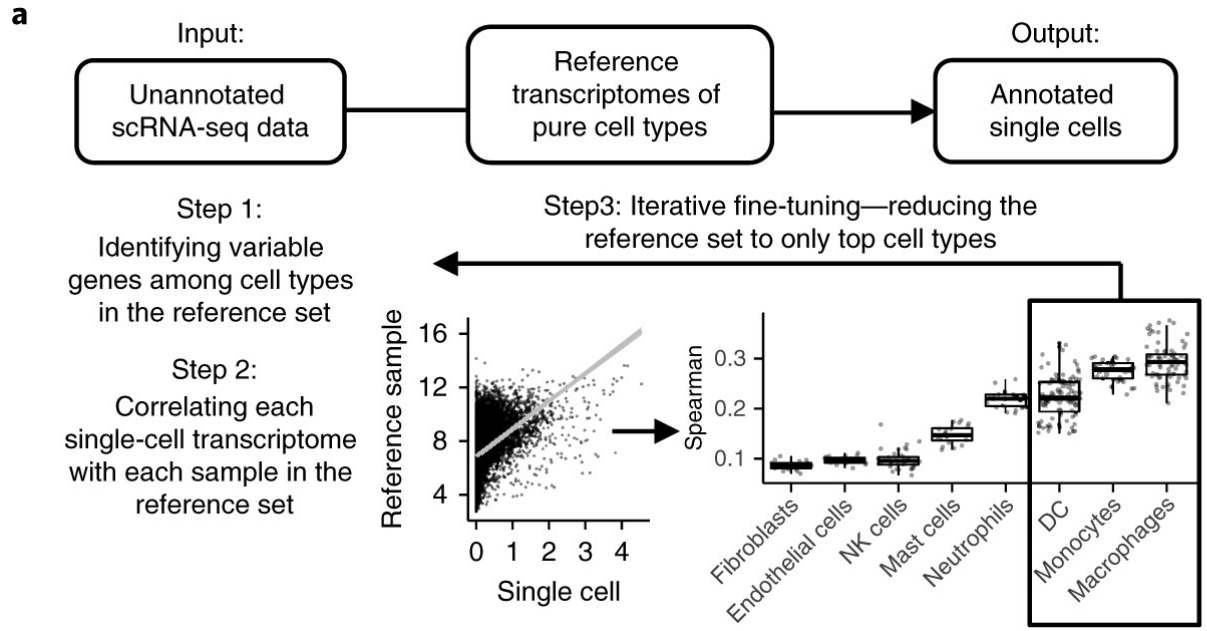
Unsupervised example

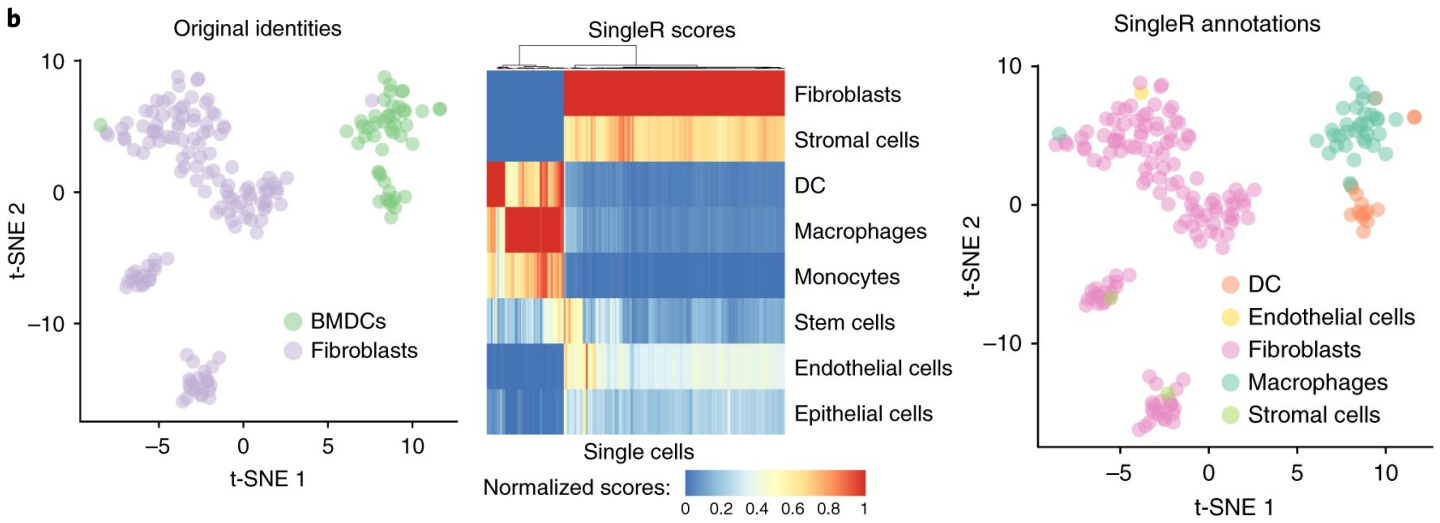


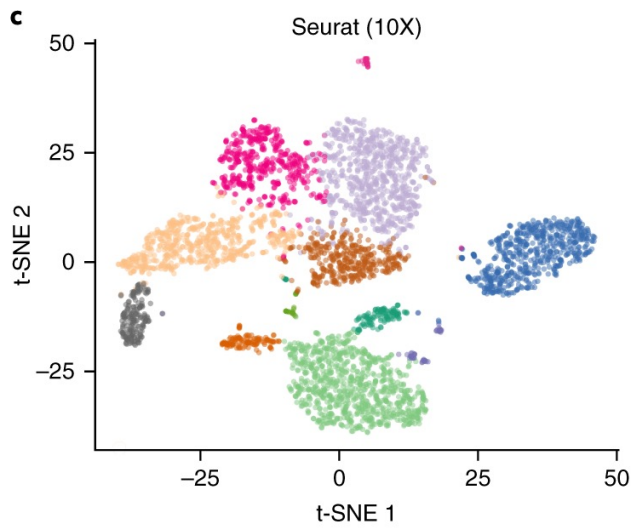
Cluster	Marker	Cell Type
0-1	IL7R	CD4 T cells
2	CD14, LYZ	CD14+ Monocytes
3	MS4A1	B cells
4	CD8A	CD8 T cells
5	FCGR3A, MS4A7	FCGR3A+ Monocytes
6	GNLY, NKG7	NK cells
Unidentified	FCER1A, CST3	Dendritic Cells
Unidentified	PPBP	Megakaryocytes

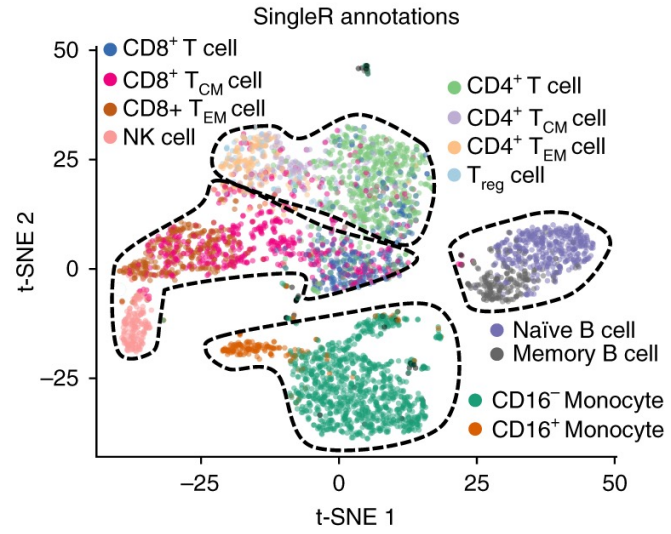
https://hbctraining.github.io/In-depth-NGS-Data-Analysis-Course/sessionIV/lessons/SC_clustering_analysis.html

Supervised example - SingleR





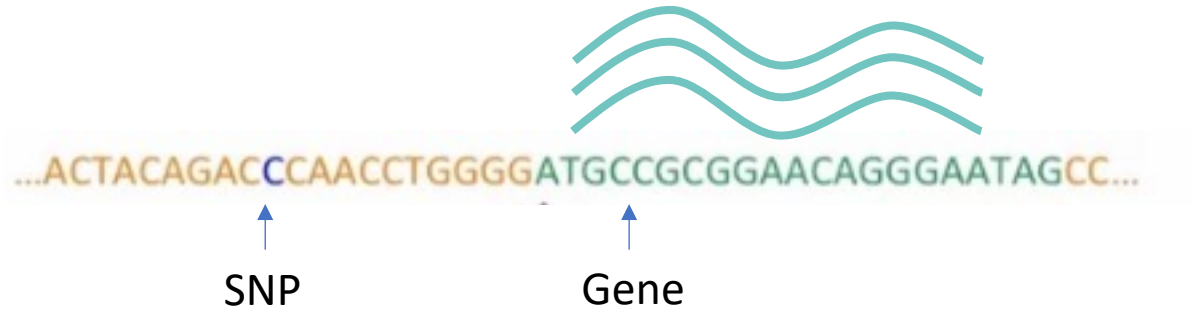


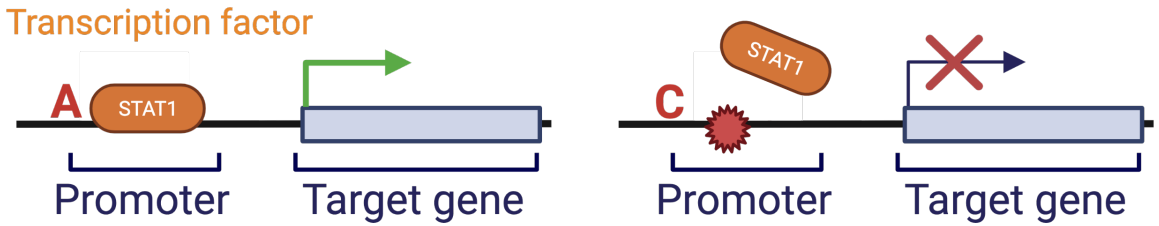


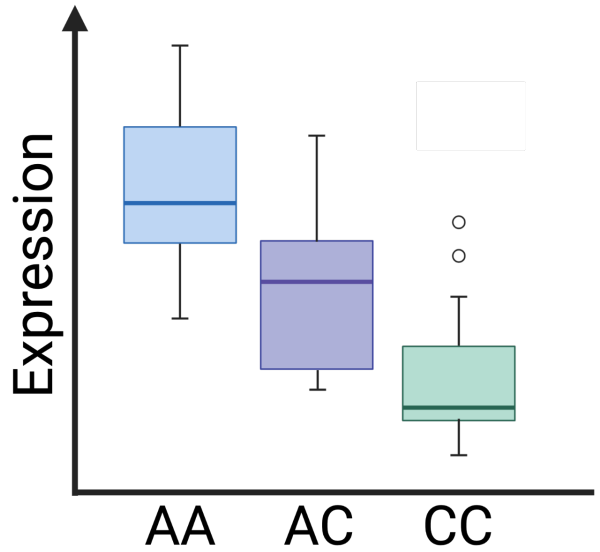
Aran, D., Looney, A.P., Liu, L. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol* **20,** 163–172 (2019). https://doi.org/10.1038/s41590-018-0276-y

Single-cell eQTL

Integration with genomics







Expression Quantitative Trait Loci (eQTL)

eQTL model: linear regression

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

Phenotype

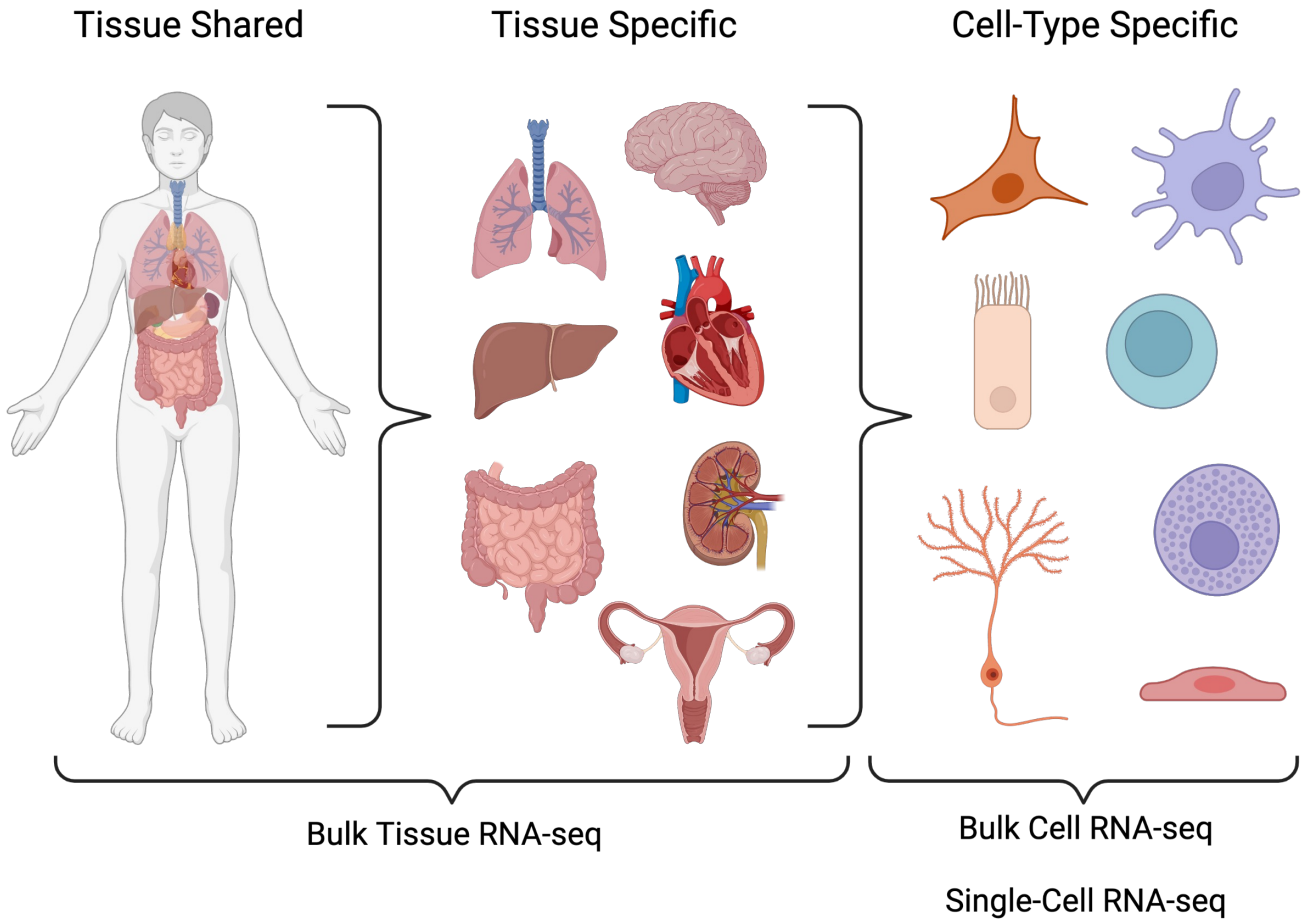
Genotype

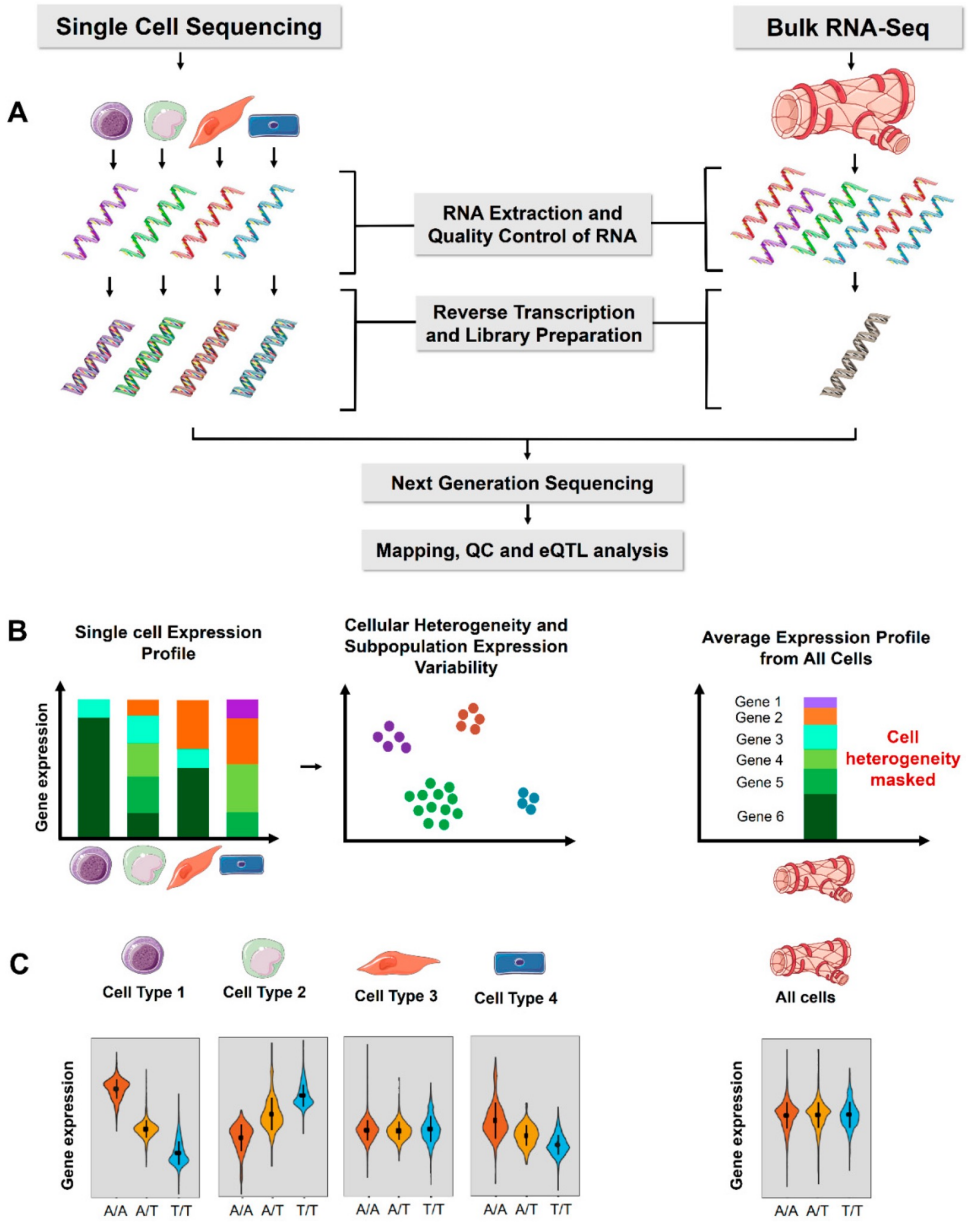
Effect size

Covariates: PCs, batches, age, sex

residual error

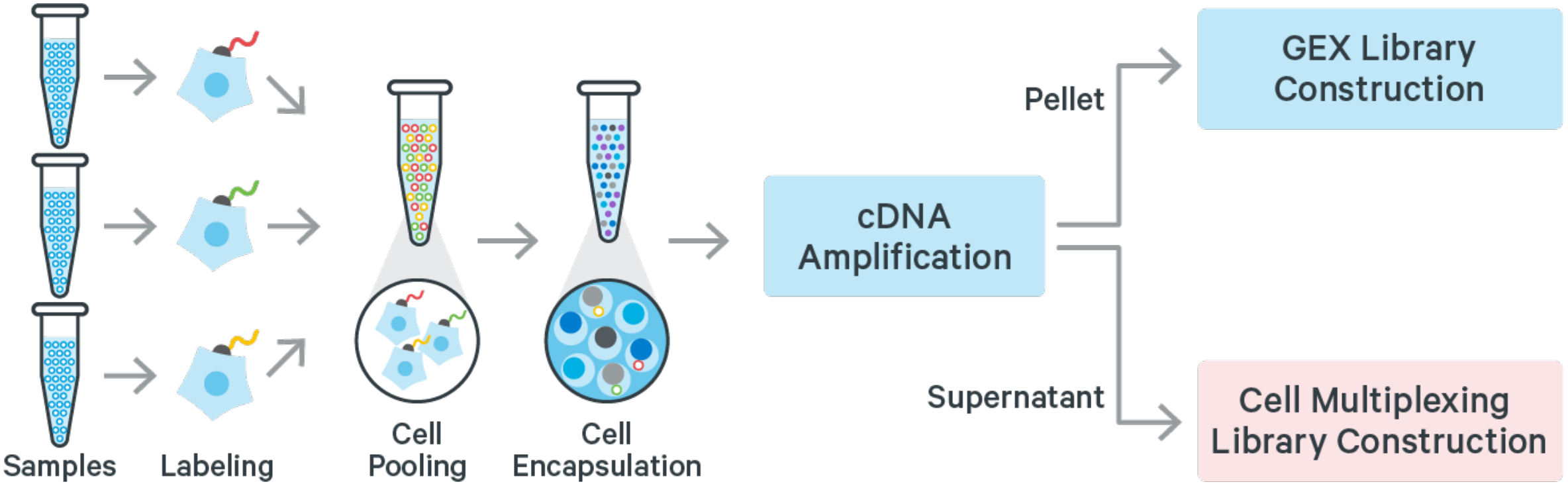
Cell-type specific eQTLs





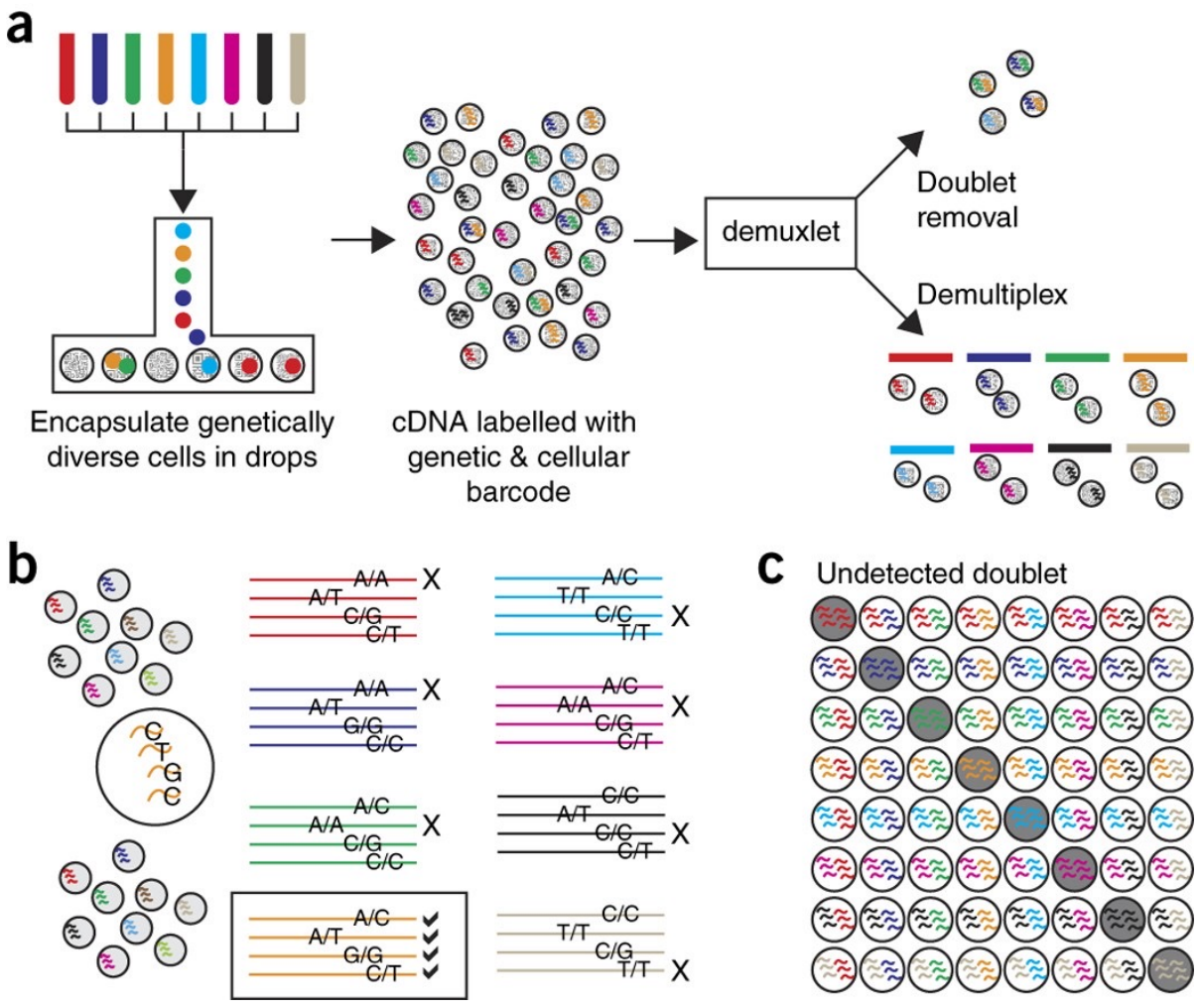
Maria et al. 2022

Multiplexing - labeling



10X Genomics

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 Xu[1], Caitlin Falconer[2], Quan Nguyen[2], Joanna Crawford[2], Brett D. McKinnon[2,5], Sally Mortlock[2], Anne Senabouth[4], Stacey Andersen[1,2], Han Sheng Chiu[2], Longda Jiang[2], Nathan J. Palpant[1,2], Jian Yang[2,10], Michael D. Mueller[5], 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

Demuxlet - Kang et al. 2017

eQTL Analysis

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

Cell type 1

	S1	S2	...	SN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
...	.	.	.	.
GeneM	25	0	.	0

Cell type 2

	S1	S2	...	SN
Gene1	30	45	.	42
Gene2	10	12	.	5
Gene3	1	4	.	8
...	.	.	.	.
GeneM	5	0	.	0

	S1	S2	...	SN
SNP1	2	0	.	1
SNP2	1	1	.	1
SNP3	0	2	.	1
...	.	.	.	.
...	.	.	.	.
SNPM	0	1	.	2

	S1	S2	S3	...	SN
gender	0	1	0	.	1
age	45	40	43	.	30

Cis vs Trans

geneid	chr	start	stop
Gene1	19	58345178	58353499
Gene2	19	58347751	58355183
Gene3	12	9067664	9116229
GeneM	12	53307456	53324864

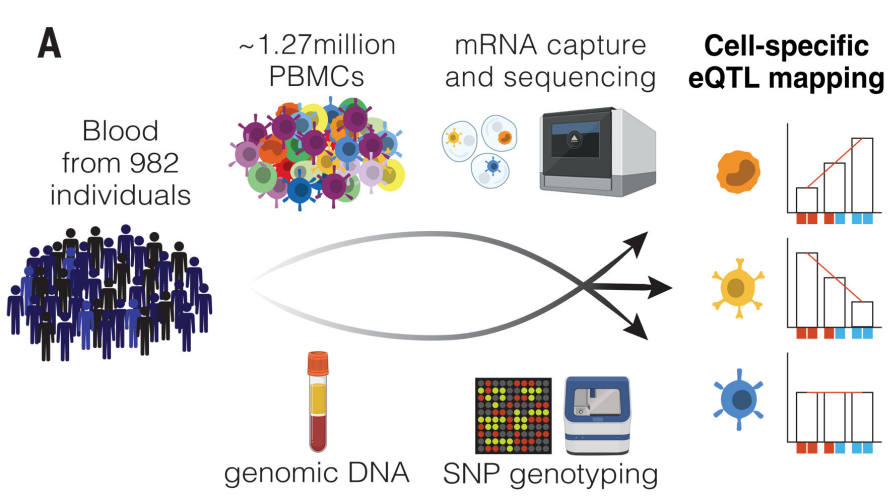
SNP_ID	chr	pos
SNP1	19	58345167
SNP2	19	58347850
SNP3	12	9067300
...	.	.
SNPM	12	53304902

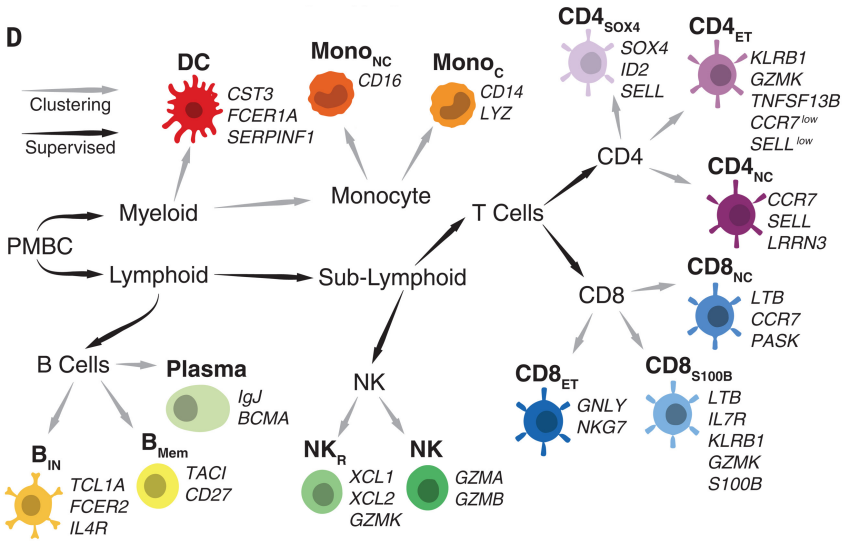
Output

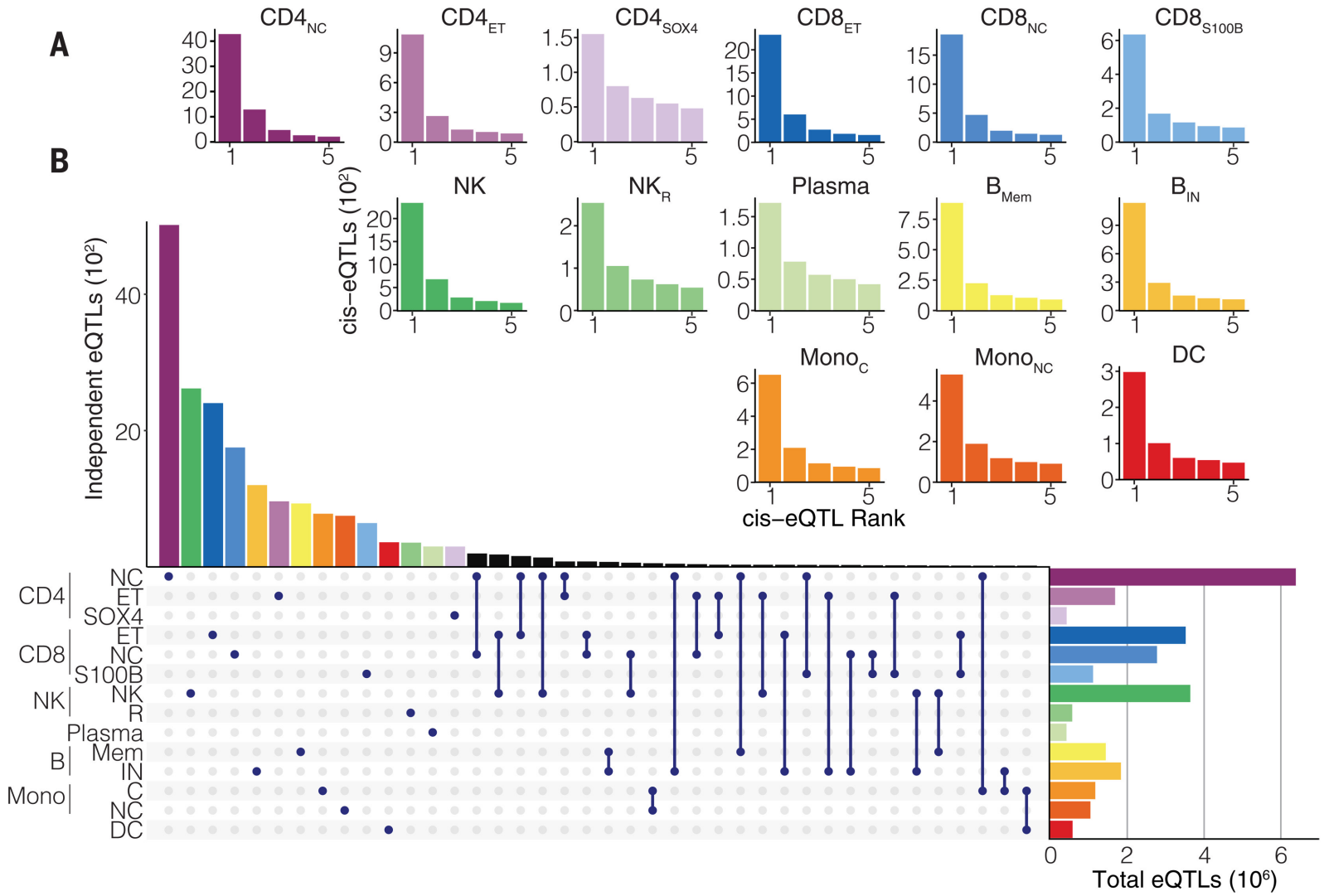
snps	gene	statistic	pvalue	FDR	beta
rs62512654	ADHFE1	32.64	1.30E-14	2.06E-07	0.47
rs145660663	ELP5	22.91	1.69E-12	9.24E-07	0.82
rs222843	ELP5	22.91	1.69E-12	9.24E-07	0.82
rs34926505	ELP5	22.91	1.69E-12	9.24E-07	0.82
rs2074217	ELP5	22.91	1.69E-12	9.24E-07	0.82
rs62377782	ERAP2	20.05	1.04E-11	1.12E-06	0.23
rs1559359	ERAP2	20.05	1.04E-11	1.12E-06	0.23
rs2549783	ERAP2	20.05	1.04E-11	1.12E-06	0.23

snps	gene	statistic	pvalue	FDR	beta
rs145660663	ELP5	22.91	1.69E-12	9.24E-07	0.82
rs222843	ELP5	22.91	1.69E-12	9.24E-07	0.82
rs1559359	ERAP2	20.05	1.04E-11	1.12E-06	0.23
rs2549783	ERAP2	20.05	1.04E-11	1.12E-06	0.23
rs199529	FAM215B	13.01	3.29E-09	3.01E-04	0.32
rs199528	FAM215B	13.01	3.29E-09	3.01E-04	0.32
rs11143240	LINC01504	12.48	5.65E-09	4.10E-04	0.03
rs7874628	LINC01504	12.48	5.65E-09	4.10E-04	0.03

Example – OneK1K

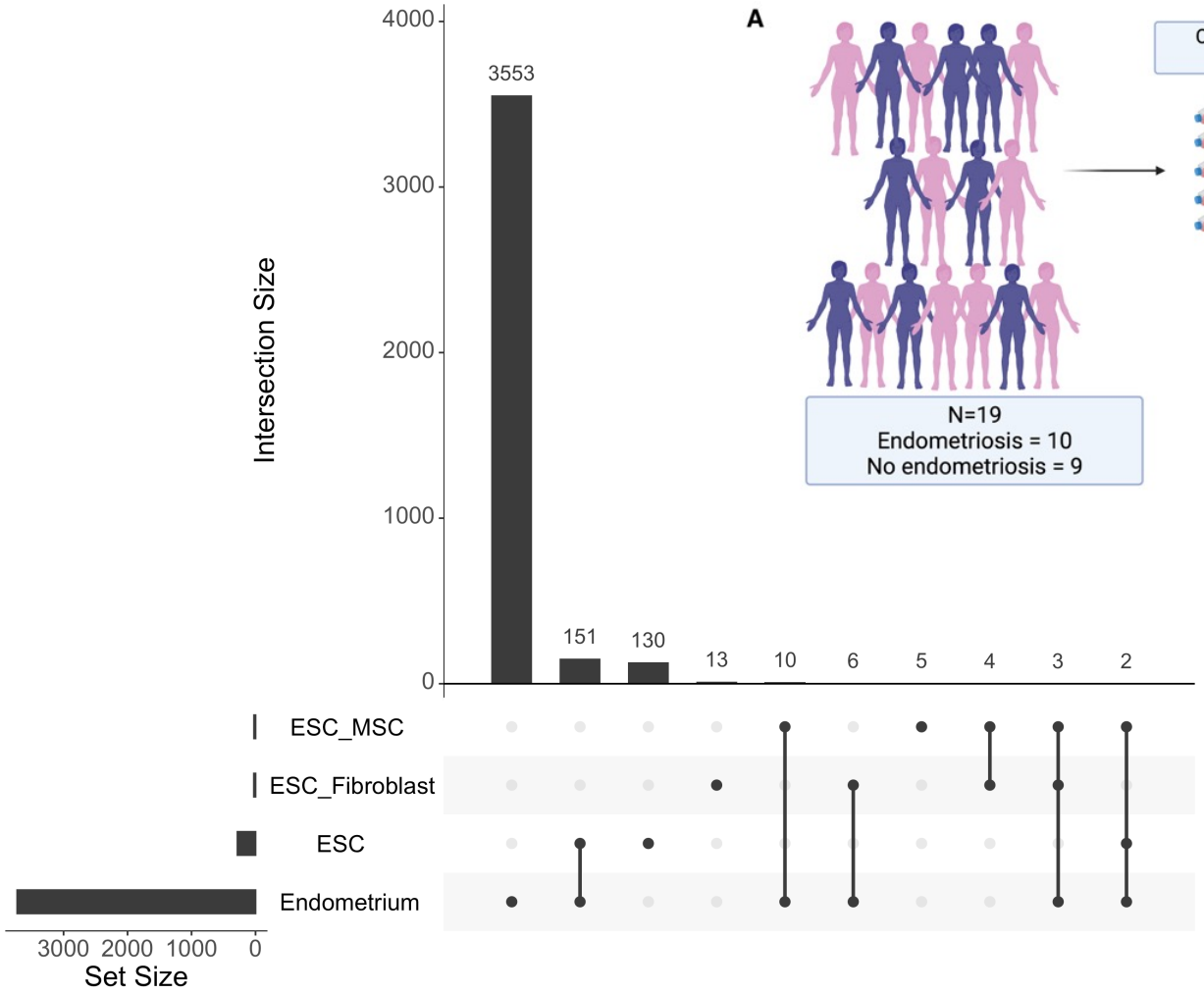


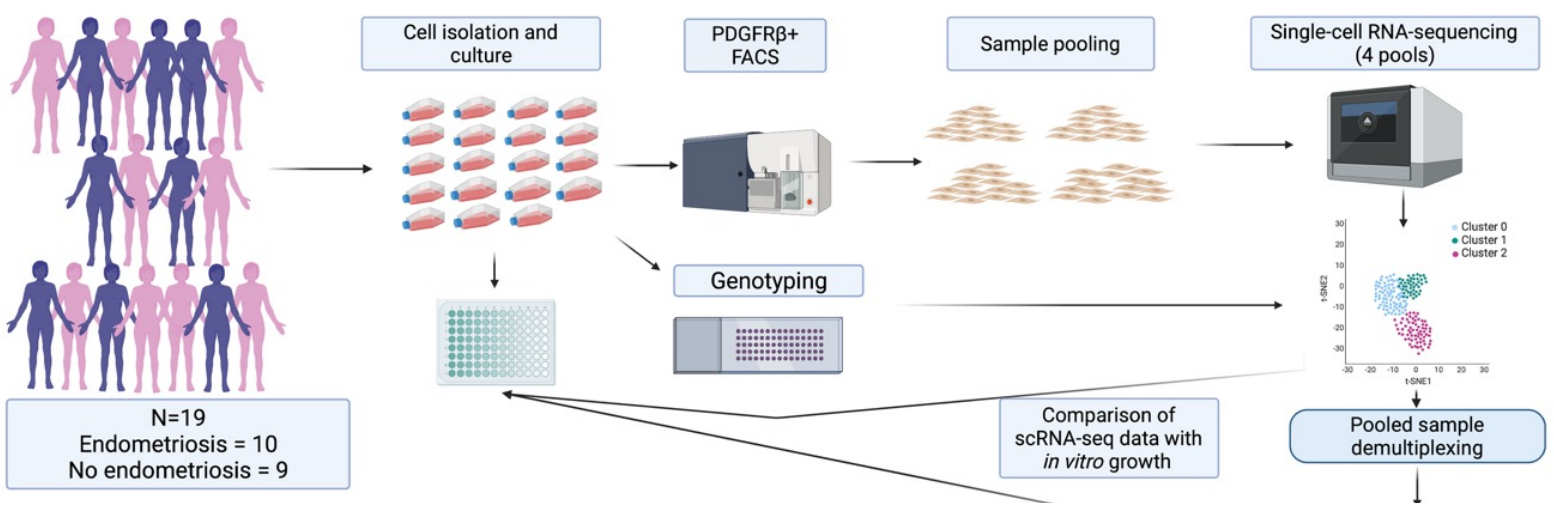




S. Yazar, et al. **Single-cell eQTL mapping identifies cell type–specific genetic control of autoimmune disease.** Science Vol. 376 Issue 6589 Pages eabf3041

Example - Single-Cell Endometrial eQTLs

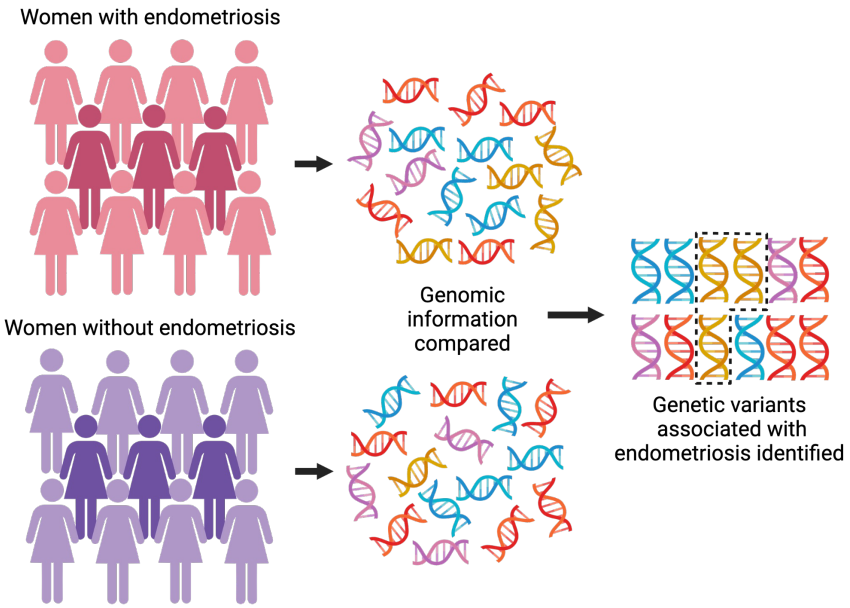




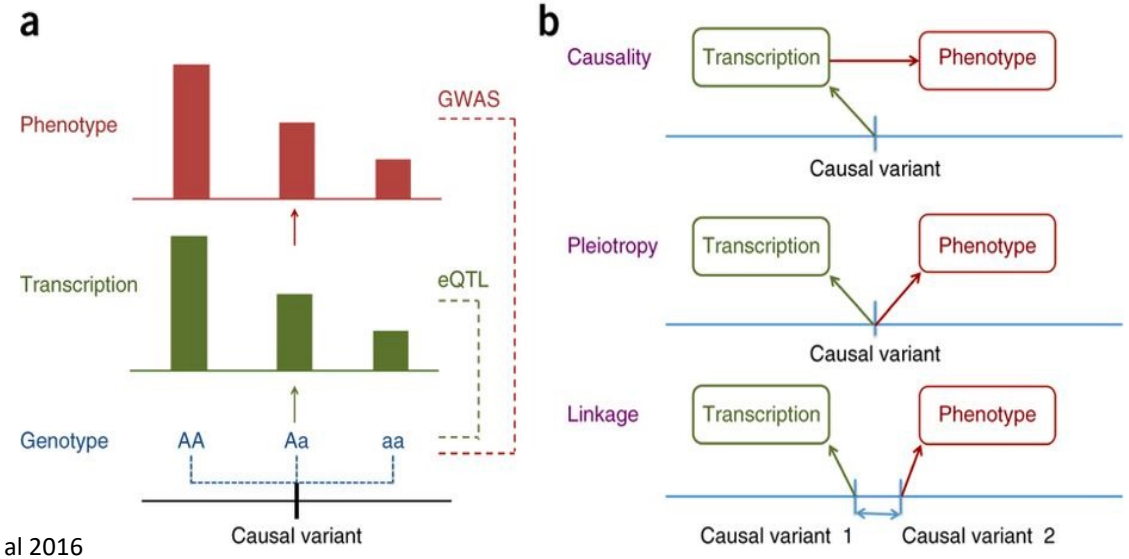
	No. Samples	No. Genes	No. SNPs	eQTLs	Unique eGenes
Fibroblast	19	14,431	4,279,467	746	26
MSC	19	12,183	4,279,467	675	24
ESC	59	14,308	5,411,984	8,050	283
Endometrium	206	17,022	6,230,993	207,071	3,726

McKinnon et al. 2022

Integrating GWAS Data

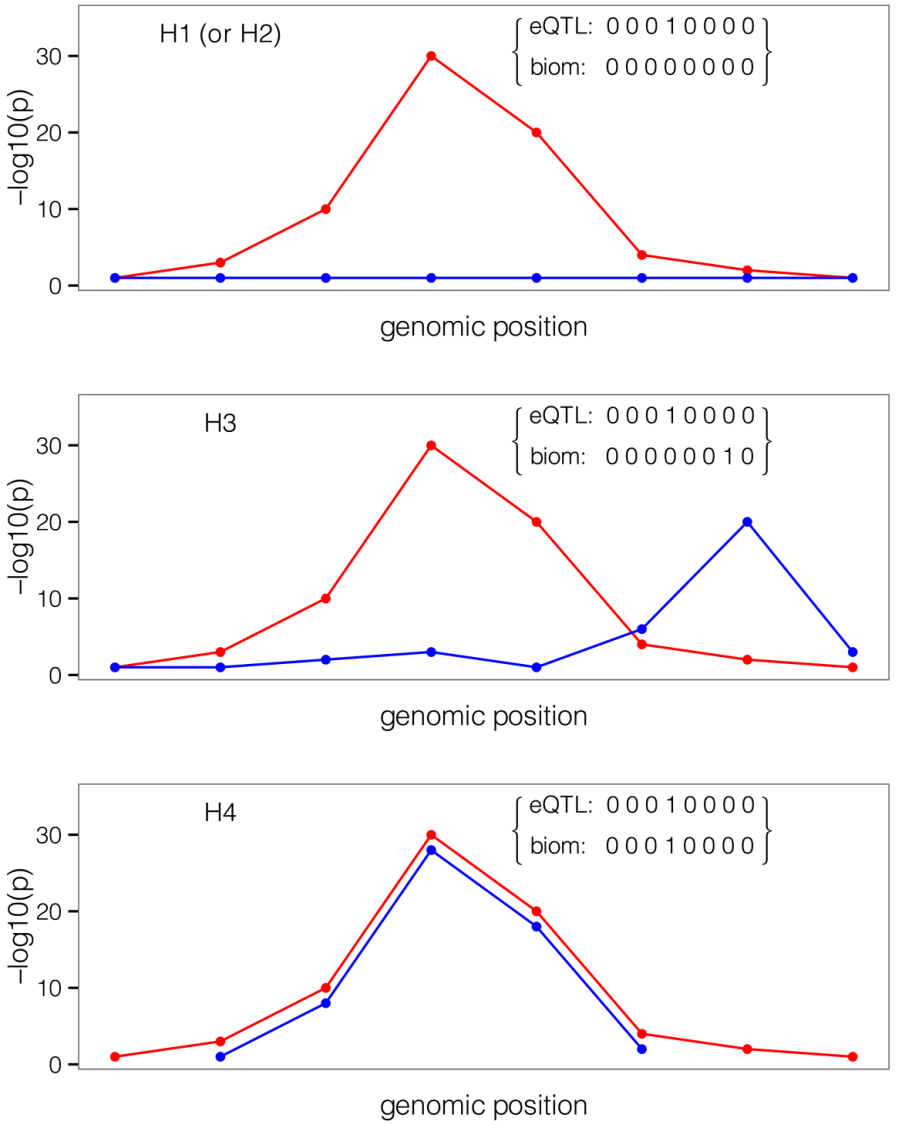


Summary-data-based Mendelian Randomisation (SMR)



Zhu et. al 2016

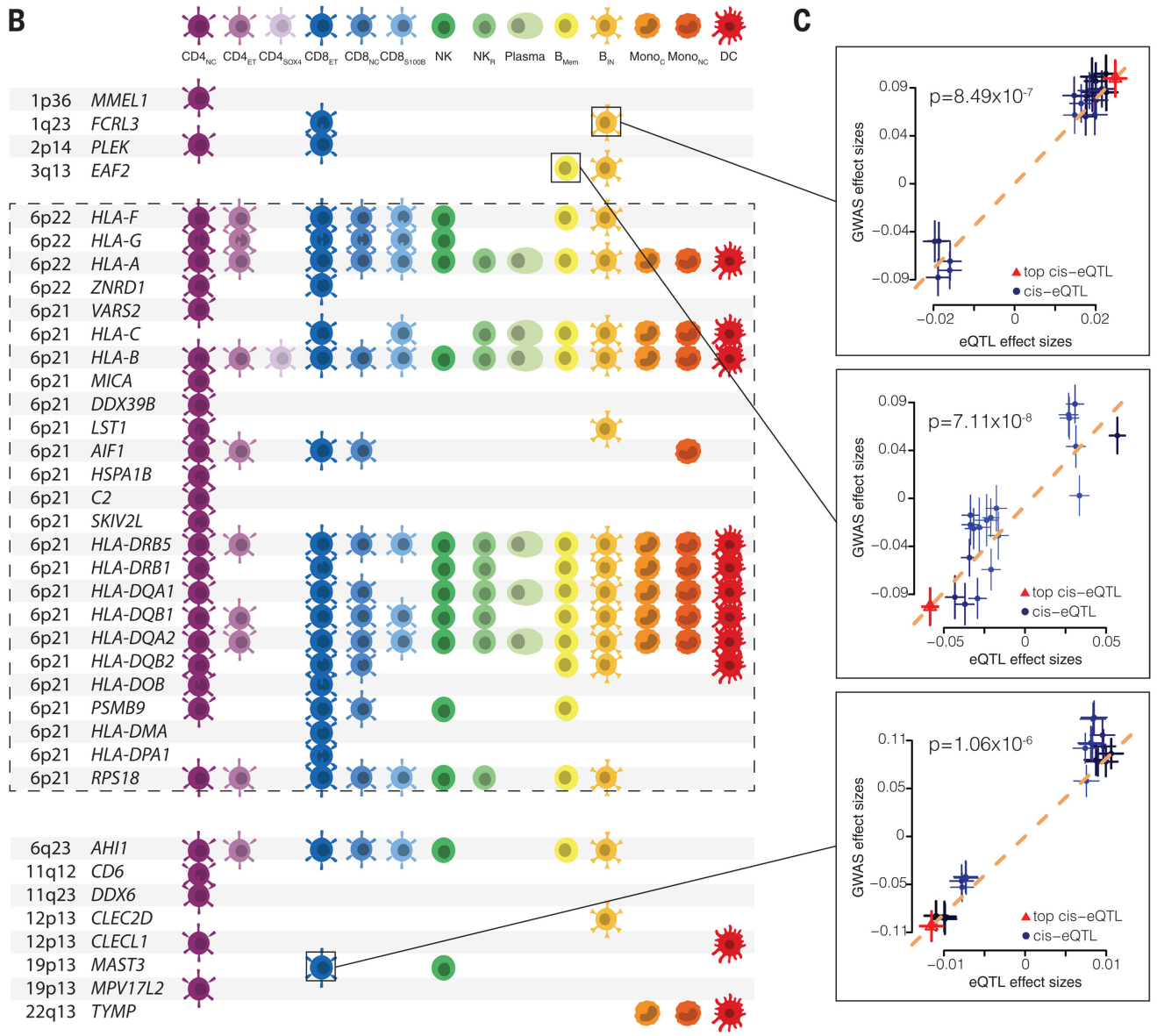
Bayesian colocalization method (coloc)



Giambartolomei et.al. 2014

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.



S. Yazar, et al. **Single-cell eQTL mapping identifies cell type–specific genetic control of autoimmune disease.** Science Vol. 376 Issue 6589 Pages eabf3041

Discussion and Future perspectives