University of Glasgow

Advanced methods in single cell

Thomas D. Otto

Montevideo 2025

First things first

- Any questions about yesterday?

Learning aims

- Introduce more technical concepts
- Explore how to process scRNA-Seq data
- Learn new methods, integration and DE in scRNA-Seq
- Critical evaluation of scRNA-Seq
- Overview of developments in scRNA-Seq

Today: Understand some of the computational concepts

1. Differential expression
2. Clustering (hierarchical / K-means)
3. Integration
4. Dimension reduction
    1. PCA
    2. T-SNE
5. Normalisation

1. Differential expression

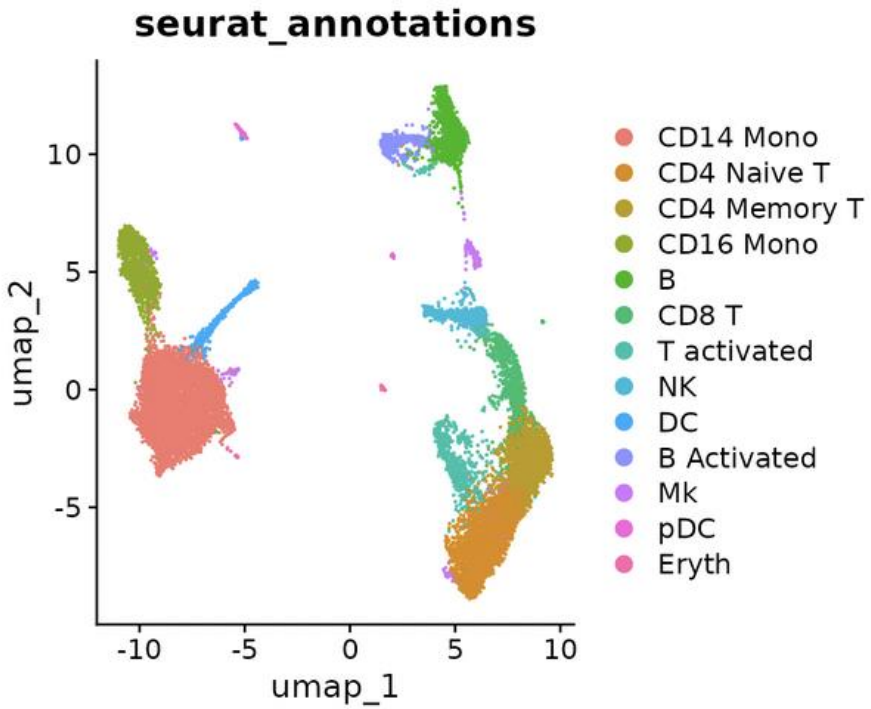
- How was it done for bulk RNA-Seq with DESeq2?
- What test do they use?
- What are the assumptions?

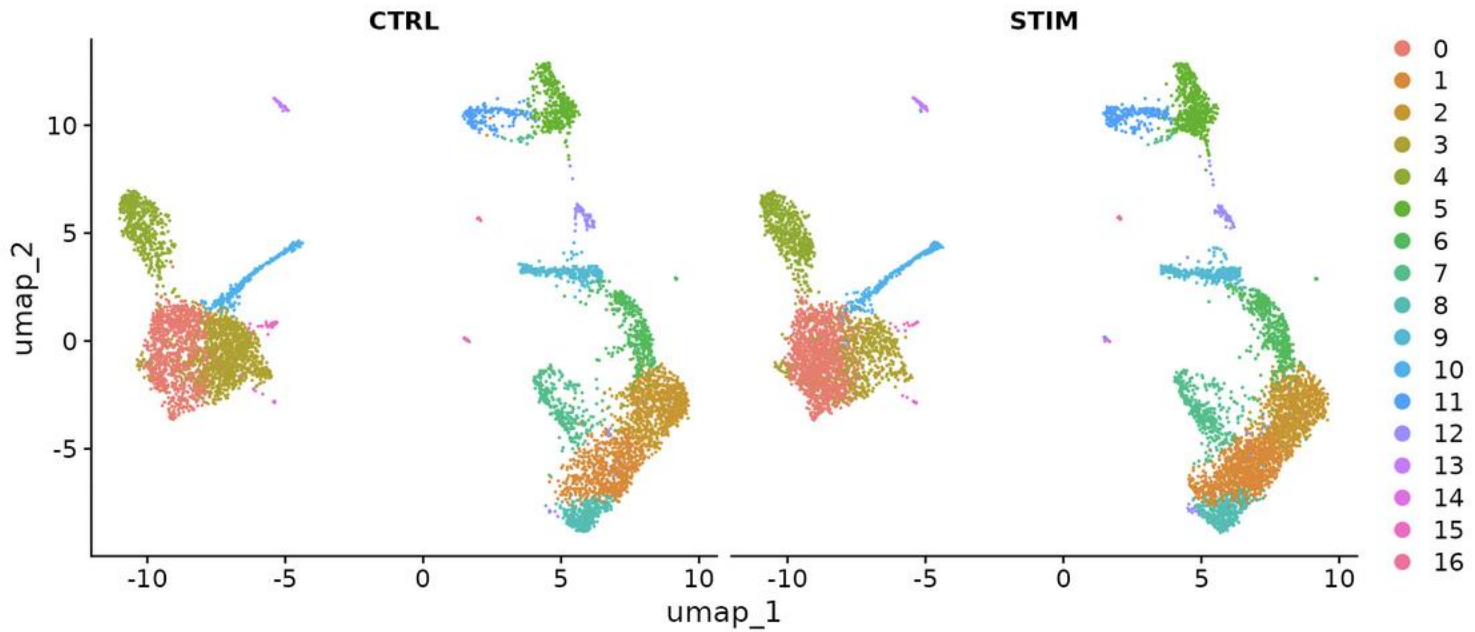
1. Differential expression

- Open question in scRNA-Seq!
- When are two genes differentially expressed – can see a difference in expression due to perturbation vs normal variation
- T-test?

Pseudo bulk

- You should know… ☺





Pseudo Bulk

- cells from each biological replicate are NOT independent (so different cells are NOT replicates)
- Merges the cell of each replicate and generate bulk
  - use DESeq2 or EDGER
  - Muscat
- Good FDR, but not enough power!
- Master project from last year…

https://www.nature.com/articles/s41467-022-35519-4

Pseudo Bulk

- cells from each biolc
  are NOT replicates)
- Merges the cell of ea
  - use DESeq2 or EDGE
  - Muscat
- Good FDR, but not enough power!
- Master project from

https://www.nature.com/articles/

Matters Arising | Open Access | Published: 22 December 2022

A balanced measure shows superior performance of pseudobulk methods in single-cell RNA-sequencing analysis

Alan E. Murphy & Nathan G. Skene

Nature Communications **13**, Article number: 7851 (2022) | Cite this article

2502 Accesses | **2** Citations | **15** Altmetric | Metrics

ARISING FROM Zimmerman, K. D., Espeland, M. A. & Langefeld, C. D. *Nature Communications* https://doi.org/10.1038/s41467-021-21038-1 (2021)

Recently, Zimmerman et al.[1], highlighted the importance of accounting for the dependence between cells from the same individual when conducting differential expression analysis on single-cell RNA-sequencing data. Their work proved the inadequacy of pseudoreplication approaches for such analysis—this was an important step forward that was conclusively proven by them. However, there appear to be limitations in both their benchmarking and simulation approaches. Here, we corrected these issues, reran the author's analysis and found that pseudobulk methods outperformed mixed models. Based on these findings, we recommend the use of pseudobulk approaches for differential expression in single-cell RNA-sequencing analyses.

Pseudo Bulk

- cells from each biological replicate are NOT independent (so different cells are NOT replicates)
- Merges the cell of each replicate and generate bulk
  - use DESeq2 or EDGER
  - Muscat
- Good FDR, but not enough power!
- Master project from last year…

https://www.nature.com/articles/s41467-022-35519-4

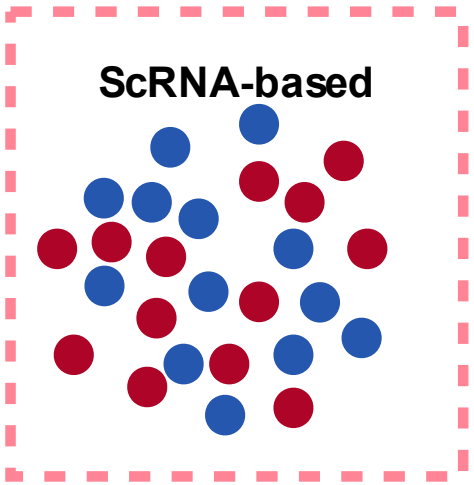
How to do pseudo bulk in Seurat?

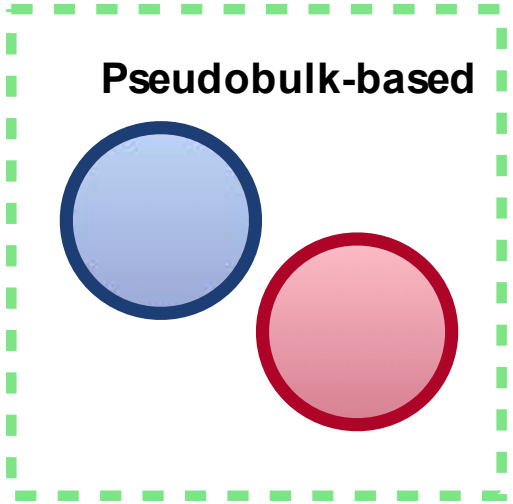
University of Glasgow

```
# pseudobulk the counts based on donor-condition-celltype

pseudo_ifnb <- AggregateExpression(ifnb, assays = "RNA", return.seurat =
T, group.by = c("stim", "donor_id", "seurat_annotations"))

pseudo_ifnb$celltype.stim <- paste(pseudo_ifnb$seurat_annotations, pseud
o_ifnb$stim, sep = "_")

Idents(pseudo_ifnb) <- "celltype.stim"


bulk.mono.de <- FindMarkers(object = pseudo_ifnb,
                            ident.1 = "CD14 Mono_STIM",
                            ident.2 = "CD14 Mono_CTRL",
                            test.use = "DESeq2")
```

Two approaches

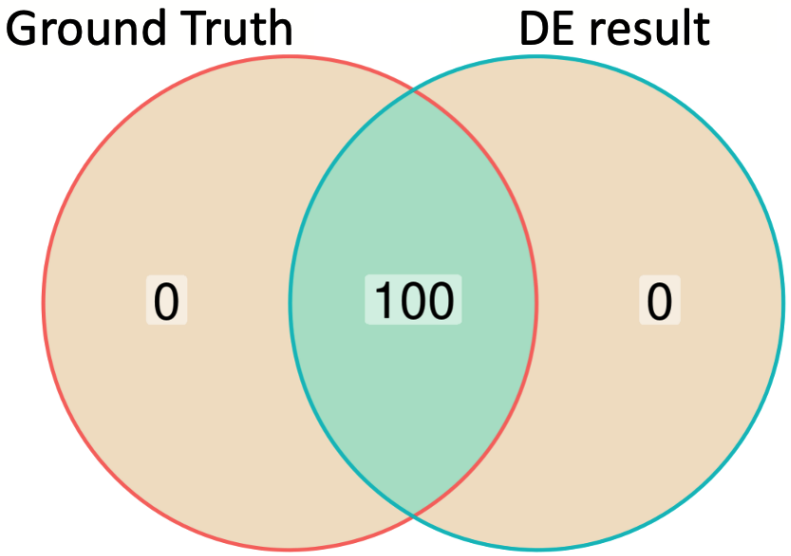
- Assuming all cells from one cluster!

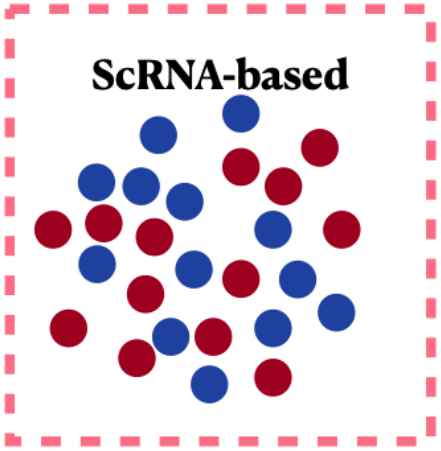


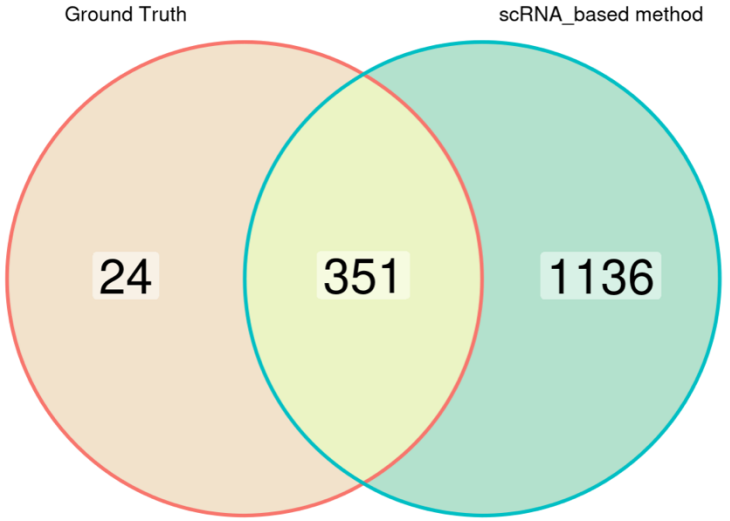


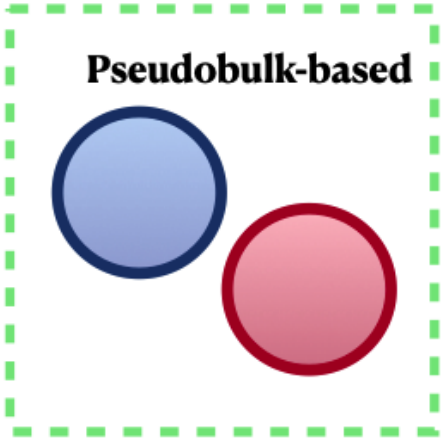
Slides from Yiyi Cheng (PhD student)

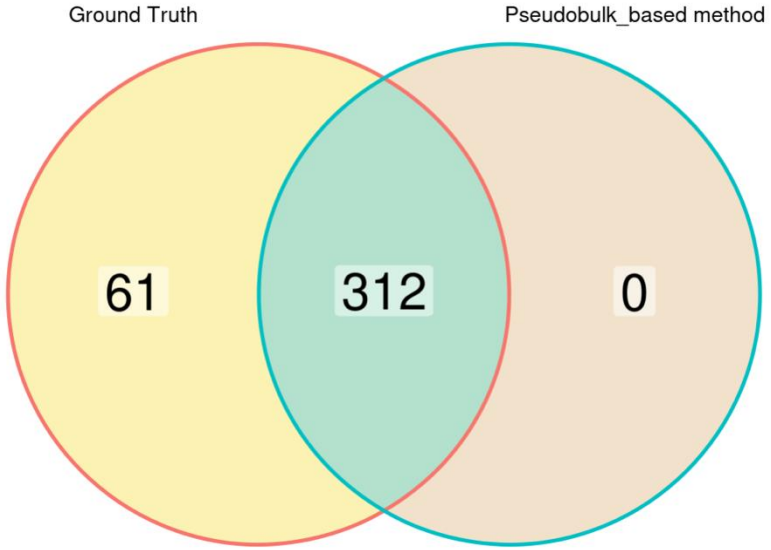
Testing the tools with perfect datasets







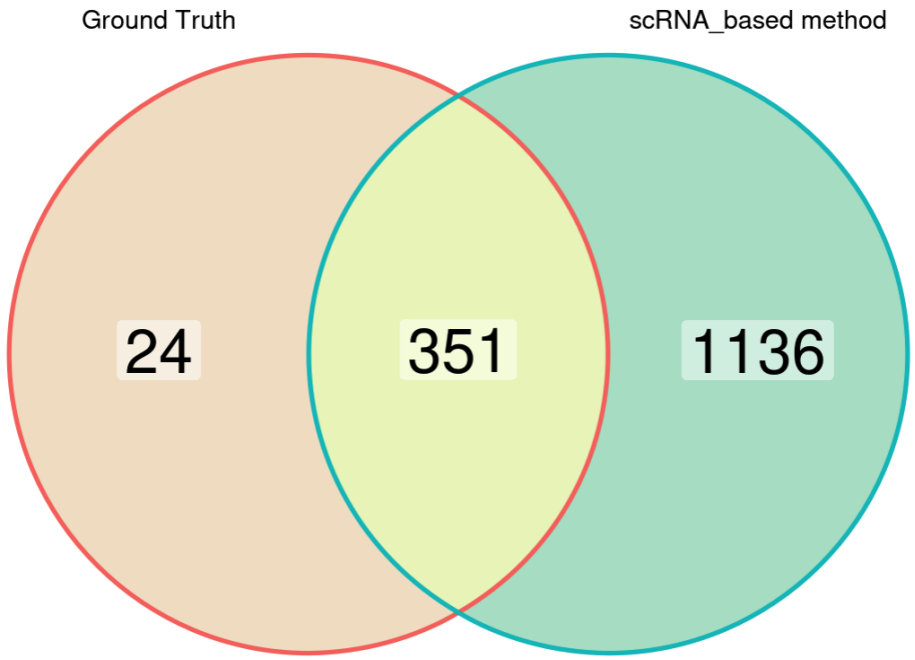




Concept of false positive / true positive

- We need a framework to classify how good the algorithms are

False negatives

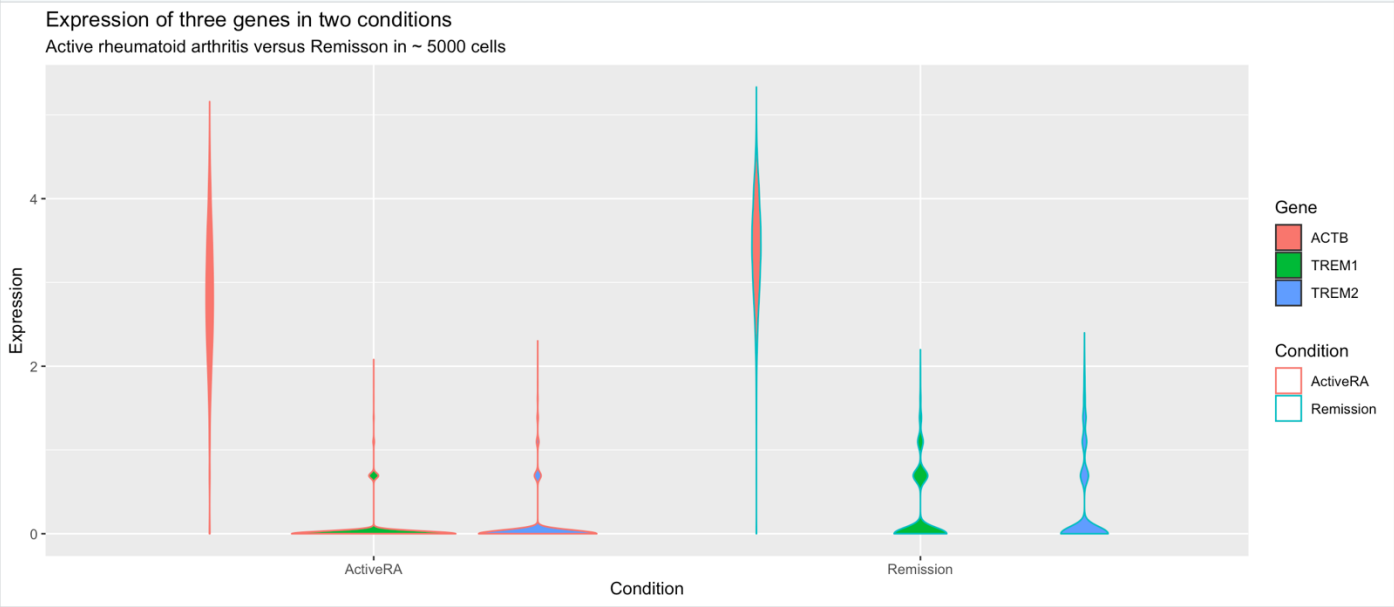


False positives

True positives

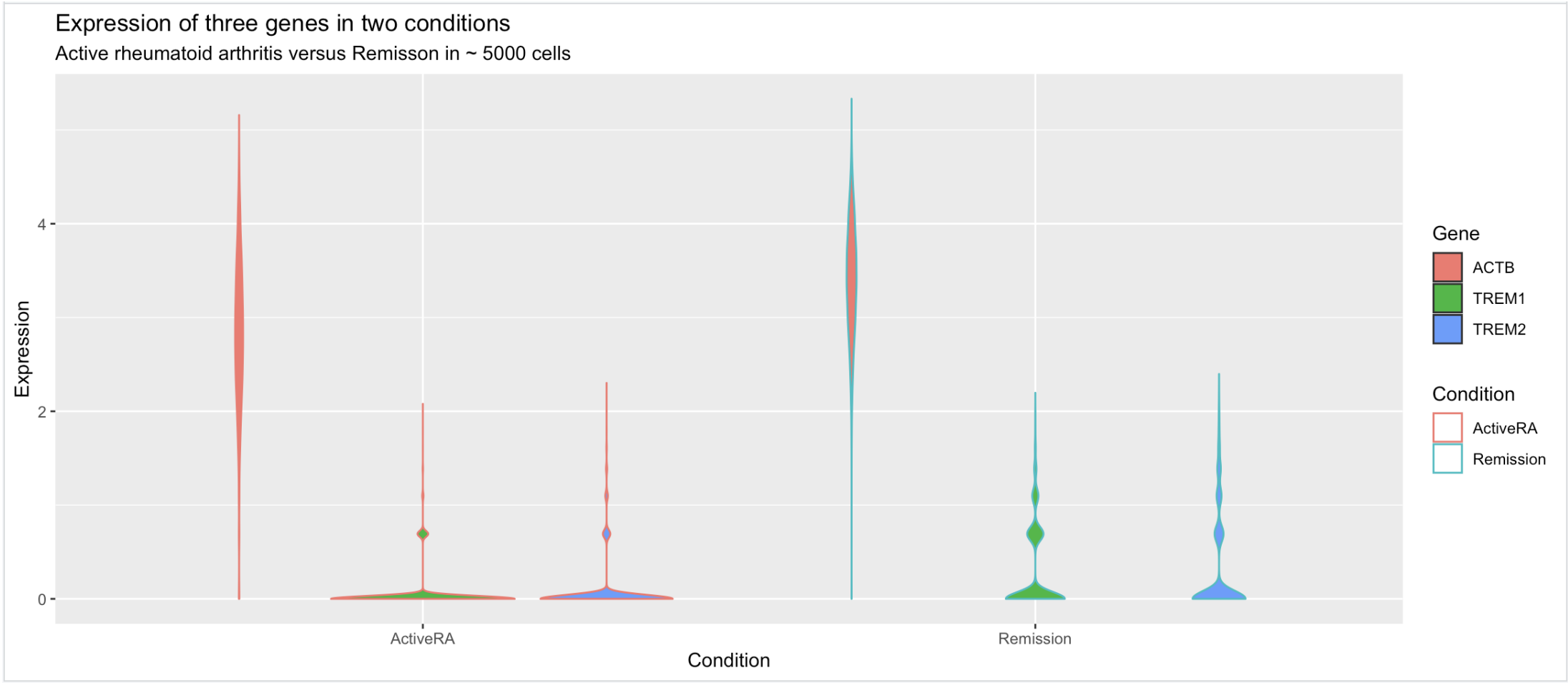
Challenges in DEG in scRNA-Seq

- **Dropout events**: Many genes have zero counts in individual cells, even if they are expressed at low levels.
- **Bimodality**: Gene expression in single cells often follows a bimodal distribution, with cells either expressing a gene (on) or not (off).
- **Heterogeneity**: Single-cell datasets are highly heterogeneous, with variability arising from both biological and technical sources.



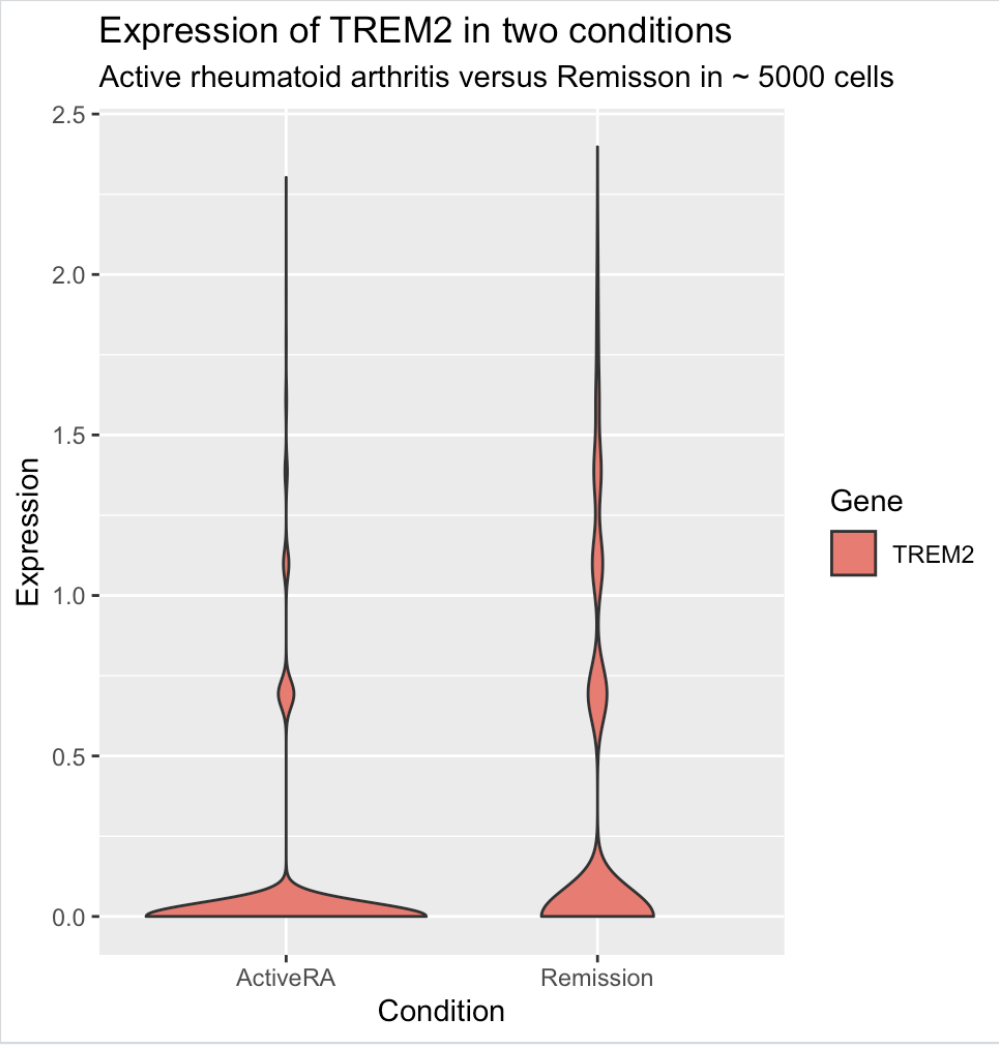
What are the challenges of scRNA-Seq?

- How well are data integrated?
- How well normalized?



t-test

- The ***t*-test** is any statistical hypothesis test in which the test statistic follows a Student's *t*-distribution under the null hypothesis.




```
> t.test(ac$Expression,re$Expression)

        Welch Two Sample t-test

data:  ac$Expression and re$Expression
t = -26.756, df = 7195, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2395823 -0.2068721
sample estimates:
 mean of x  mean of y
0.07437191 0.29759911
```

How are data normalised?
How many replicates?
Can I take cell as replicates?

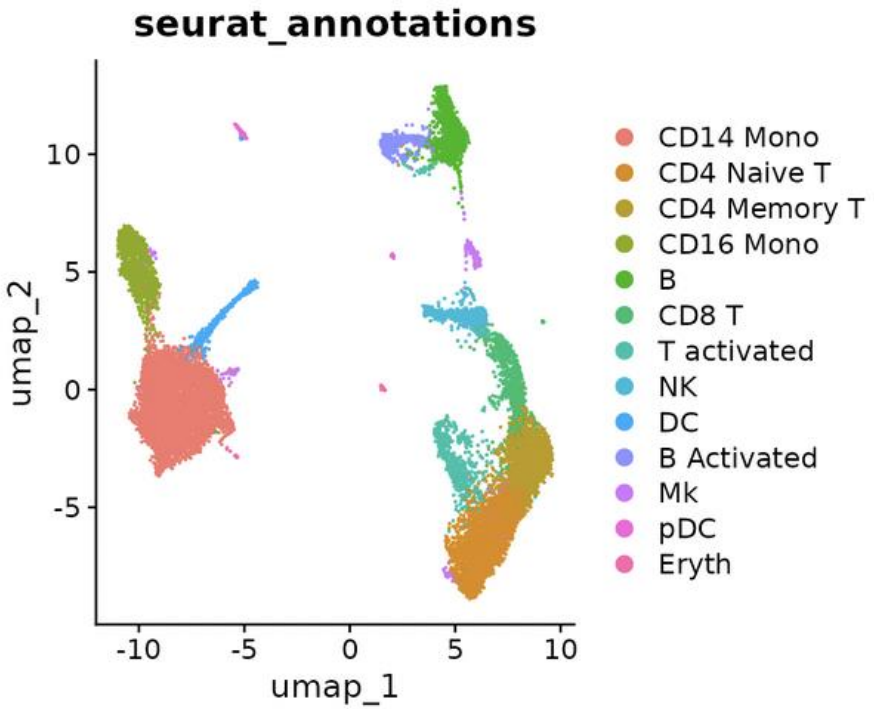
Different methods

- ?FindMarkers

- "wilcox" : Identifies differentially expressed genes between two groups of cells using a Wilcoxon Rank Sum test (default)
- "bimod" : Likelihood-ratio test for single cell gene expression, (McDavid et al., Bioinformatics, 2013)
- "roc" : Identifies 'markers' of gene expression using ROC analysis. For each gene, evaluates (using AUC) a classifier built on that gene alone, to classify between two groups of cells. An AUC value of 1 means that expression values for this gene alone can perfectly classify the two groupings (i.e. Each of the cells in cells.1 exhibit a higher level than each of the cells in cells.2). An AUC value of 0 also means there is perfect classification, but in the other direction. A value of 0.5 implies that the gene has no predictive power to classify the two groups. Returns a 'predictive power' (abs(AUC-0.5) * 2) ranked matrix of putative differentially expressed genes.
- "t" : Identify differentially expressed genes between two groups of cells using the Student's t-test.
- "negbinom" : Identifies differentially expressed genes between two groups of cells using a negative binomial generalized linear model. Use only for UMI-based datasets
- "poisson" : Identifies differentially expressed genes between two groups of cells using a poisson generalized linear model. Use only for UMI-based datasets
- "LR" : Uses a logistic regression framework to determine differentially expressed genes. Constructs a logistic regression model predicting group membership based on each feature individually and compares this to a null model with a likelihood ratio test.
- "MAST" : Identifies differentially expressed genes between two groups of cells using a hurdle model tailored to scRNA-seq data. Utilizes the MAST package to run the DE testing.
- "DESeq2" : Identifies differentially expressed genes between two groups of cells based on a model using DESeq2 which uses a negative binomial distribution (Love et al, Genome Biology, 2014).This test does not support pre-filtering of genes based on average difference (or percent detection rate) between cell groups. However, genes may be pre-filtered based on their minimum detection rate (min.pct) across both cell groups. To use this method, please install DESeq2, using the instructions at https://bioconductor.org/packages/release/bioc/html/DESeq2.html

The code Seurat

```
ifnb$celltype.stim <- paste(ifnb$seurat_annotations, ifnb$stim, sep = "_")
Idents(ifnb) <- "celltype.stim"
b.interferon.response <- FindMarkers(ifnb, ident.1 = "B_STIM", ident.2 = "B_CTRL", verbose = FALSE)*
head(b.interferon.response, n = 15)
```



- "wilcox" : Identifies differentially expressed genes between two groups of cells using a Wilcoxon Rank Sum test (default)

Wilcoxon Rank-Sum Test in Seurat

- **Non-Parametric**: It does not assume that the data follows a specific distribution (e.g., normal distribution), making it suitable for scRNA-seq data, which is often sparse and noisy.
- **Robust to Outliers**: It is less sensitive to extreme values compared to parametric tests like the t-test.
- **Handles Dropouts**: It works well with sparse data, where many genes have zero counts in individual cells.

Wilcoxon rank-sum test compares the distributions of gene expression

University of Glasgow

Step 1: Rank the Expression Values

- For each gene, the expression values across all cells are ranked from lowest to highest, regardless of group membership.
- If there are ties (e.g., multiple cells with the same expression value), they are assigned the average rank.

Step 2: Calculate the Test Statistic

- The ranks of the expression values are summed for each group.
- The test statistic U is calculated as:

$$U = R_1 - \frac{n_1(n_1+1)}{2}$$

Where:

- R_1 is the sum of ranks for the first group.
- n_1 is the number of cells in the first group.

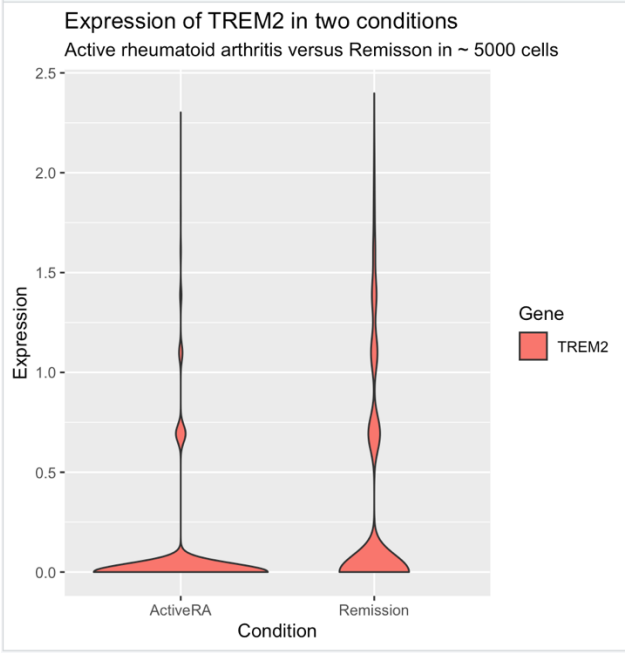
Step 3: Compute the p-value

- The p-value is calculated based on the distribution of the test statistic UU.
- A small p-value indicates that the gene expression distributions between the two groups are significantly different.

https://www.sciencedirect.com/science/article/pii/S0092867421005833?via%3Dihub; DeepSeek

MAST - Model-based Analysis of Single-cell Transcriptomics

- scRNA-Seq suffers from stochastic dropout and characteristic bimodal expression distributions in which expression is either strongly non-zero or non-detectable.
- Technical assay variability and extrinsic biological factors can significantly influence expression level measurements – modelled through CDR (Cell detection rate) co-variate
- MAST uses a two-part, generalized linear model (hurdle model) for such bimodal data that parameterizes both of these features



DOI 10.1186/s13059-015-0844-5

Cellular detection rate (CDR)

- Control of noise

The CDR for cell i is:

$$CDR_i = 1/N \sum_{g=1}^{N} z_{ig} \quad (1)$$

where z_{ig} is an indicator if gene g in cell i was expressed above *background*. We consider the implications of set-

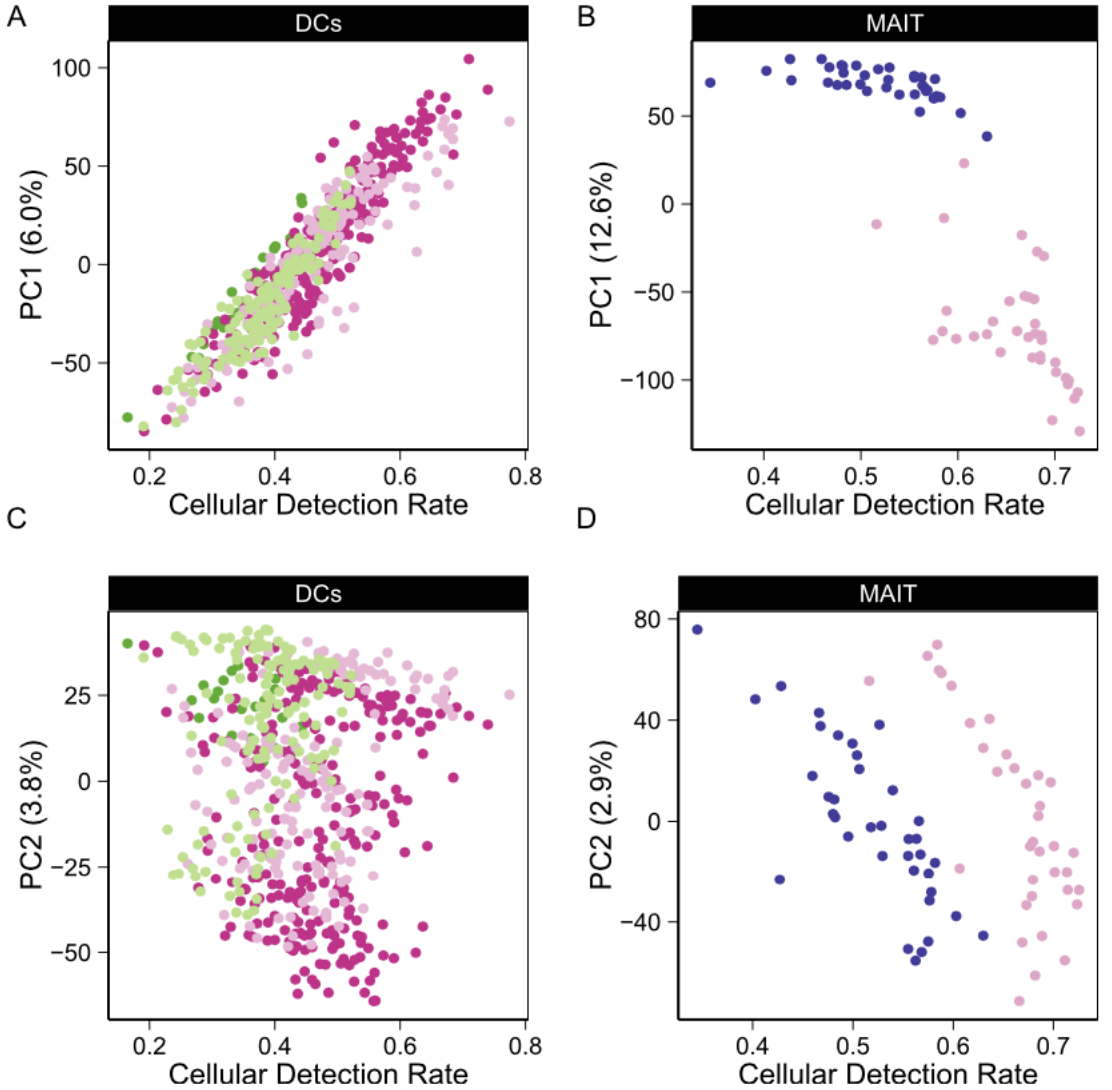


Fig. 1 Cellular detection rate correlates with the first two principal components of variation. The fraction of genes expressed, or cellular detection rate (CDR) correlates mostly with the **a**,**c**) first principal component (PC) of variation in the myeloid dendritic cells (DC) data set and mostly with the second PC in the **b**,**d**) mucosal-associated invariant T (MAIT) data set

MAST uses a **two-part model** to analyze gene expression:

University of Glasgow

Part 1: Hurdle Model for Detection (Binary Model)

- This part models the **probability that a gene is expressed** (i.e., not a dropout) in a given cell.
- It uses a **logistic regression framework** to predict whether a gene is "on" or "off" based on covariates (e.g., cell type, condition, or batch, modelled through CDR).
- The output is the **detection rate** (probability of expression) for each gene in each cell.

Part 2: Continuous Model for Expression Level

- This part models the **expression level** of a gene **only in cells where the gene is detected** (i.e., "on").
- It uses a **linear regression framework** (with a log-normal distribution) to predict the expression level based on covariates.
- The output is the **expected expression level** for each gene in each cell where it is expressed.

Combining the Two Parts

- **Part 1 (Hurdle Model)**:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1 X_{ij} + \beta_2 \text{CDR}_j + \epsilon_{ij}$$

Where:

- p_{ij} is the probability that gene i is expressed in cell j.
- X_{ij} represents other covariates (e.g., cell type or condition).
- CDR_j is the cellular detection rate for cell j.
- $\beta_0, \beta_1, \beta_2$ are coefficients to be estimated.

Part 2 (Continuous Model):

$$\log(y_{ij}) = \gamma_0 + \gamma_1 X_{ij} + \gamma_2 \text{CDR}_j + \eta_{ij}$$

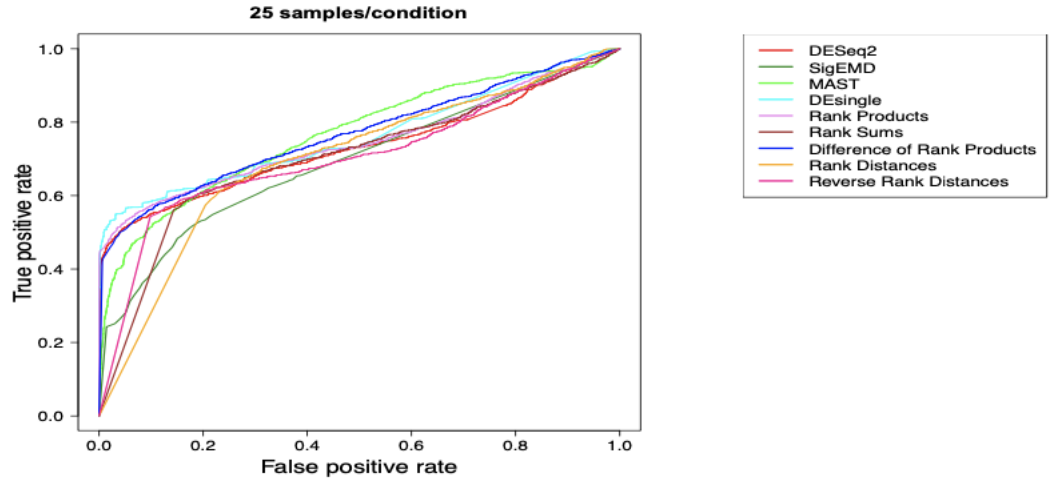
Where:

- y_{ij} is the expression level of gene i in cell j (for cells where the gene is expressed).
- X_{ij} represents other covariates.
- CDR_j is the cellular detection rate for cell j.
- $\gamma_0, \gamma_1, \gamma_2$ are coefficients to be estimated.

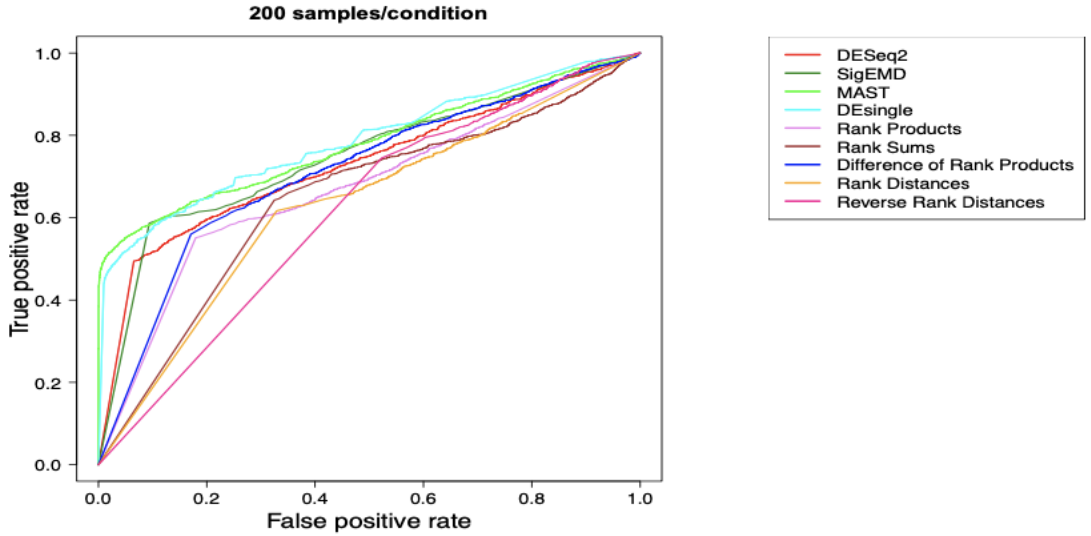
- MAST combines the results from the two parts to perform **differential expression analysis**.
- For each gene, it tests whether there are significant differences in:
  - The **probability of expression** (Part 1) between groups (e.g., cell types or conditions).
  - The **expression level** (Part 2) between groups.
- The final p-values from both parts are combined using a **meta-analysis approach** (e.g., Fisher's method) to determine overall significance

Power?

A)



D)



- Master thesis Olympia (2020)

50 cells per condition

Method	Number of Detected DE Genes	FDR $(\frac{FP}{FP+TP})$	TPR/Recall $(\frac{TP}{TP+FN})$	Precision $(\frac{TP}{TP+FP})$	F1 $(2\times(\frac{Precision \times Recall}{Precision+Recall}))$
MAST	134	0.015	0.092	0.985	0.168
SigEMD	905	0.281	0.454	0.719	0.557
DEsingle	1379	0.381	0.596	0.619	0.607
DESeq2	2266	0.583	0.652	0.417	0.509
Rank Products	2805	0.644	0.697	0.356	0.471
Rank Sum	4231	0.723	0.816	0.277	0.413
Rank Distance	4905	0.746	0.871	0.254	0.394
Reverse-Rank Distance	3934	0.719	0.773	0.281	0.413
Differential Rank Products	3424	0.664	0.802	0.336	0.473

Table 2: A summary statistics table showing the total number of detected differentially expressed genes by each package for 50 cells per condition. The calculated false discovery rate (FDR), true positive rate (TPR/Recall), precision and F1 score. Calculation formulas are included in the header where appropriate with number of false positives denoted as FP, the number of true positives as TP and the number of false negatives as FN.

Newer study – false positive

N_{ind}	N_{cells}	Two-part hurdle			Tweedie		GEE1	Pseudo-bulk		Tobit	Modified *t*
Replicates		Default	Corrected	RE	GLMM	GLM		Mean	Sum		
5	50	0.561	0.637	0.069	0.082	0.340	0.114	0.023	0.035	0.353	0.400
	100	0.677	0.719	0.064	0.084	0.463	0.110	0.022	0.032	0.471	0.510
	250	0.798	0.778	0.066	0.083	0.609	0.103	0.023	0.028	0.628	0.644
	500	0.862	0.803	0.065	0.081	0.705	0.104	0.023	0.026	0.725	0.718
40	50	0.561	0.602	0.051	0.054	0.345	0.055	0.025	0.013	0.340	0.393
	100	0.689	0.699	0.049	0.053	0.455	0.055	0.026	0.012	0.467	0.502
	250	0.820	0.803	0.044	0.053	0.607	0.053	0.022	0.010	0.622	0.639
	500	0.888	0.856	0.042	0.053	0.704	0.054	0.022	0.008	0.721	0.713

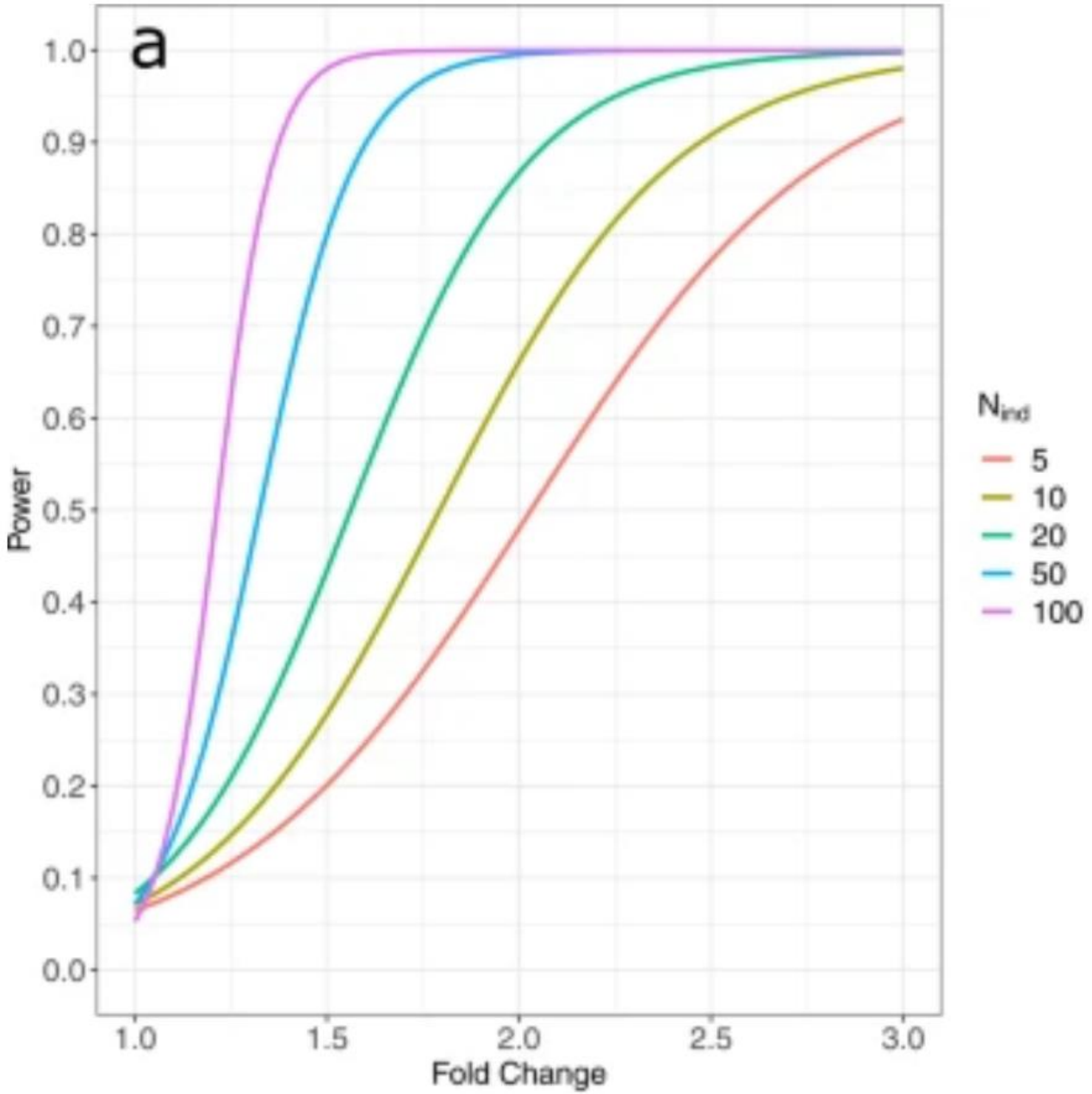
Default denotes MAST was implemented without random effects, RE denotes random effects, Corrected denotes data were batch-corrected for individual with ComBat prior to analysis without using individual as a random effect, GLM denotes generalized linear model, and GLMM denotes generalized linear mixed-effects model.

Two-part hurdle model as implemented in MAST, Tweedie distribution as implemented in "glmmTMB", GEE1 as implemented in "geepack", Pseudo-bulk averaged or summed across cells within an individual and was implemented in DESeq2, Modified *t* as implemented in ROTS, and Tobit as implemented in Monocle.

Zimmerman et al, 2020, Nature Communication

Power

Power curves for various, but likely, single-cell scenarios using MAST with a random effect for individual. Fold-change is simulated by multiplying the global mean gene expression values by the fold-change value for one group. All power is computed at $\alpha = 0.05$. **a** Differences in power when sample sizes range between 5 individuals per group to 100 when the number of cells per individual is held constant at 250



Replicates

- But how many replicates are feasible to do?

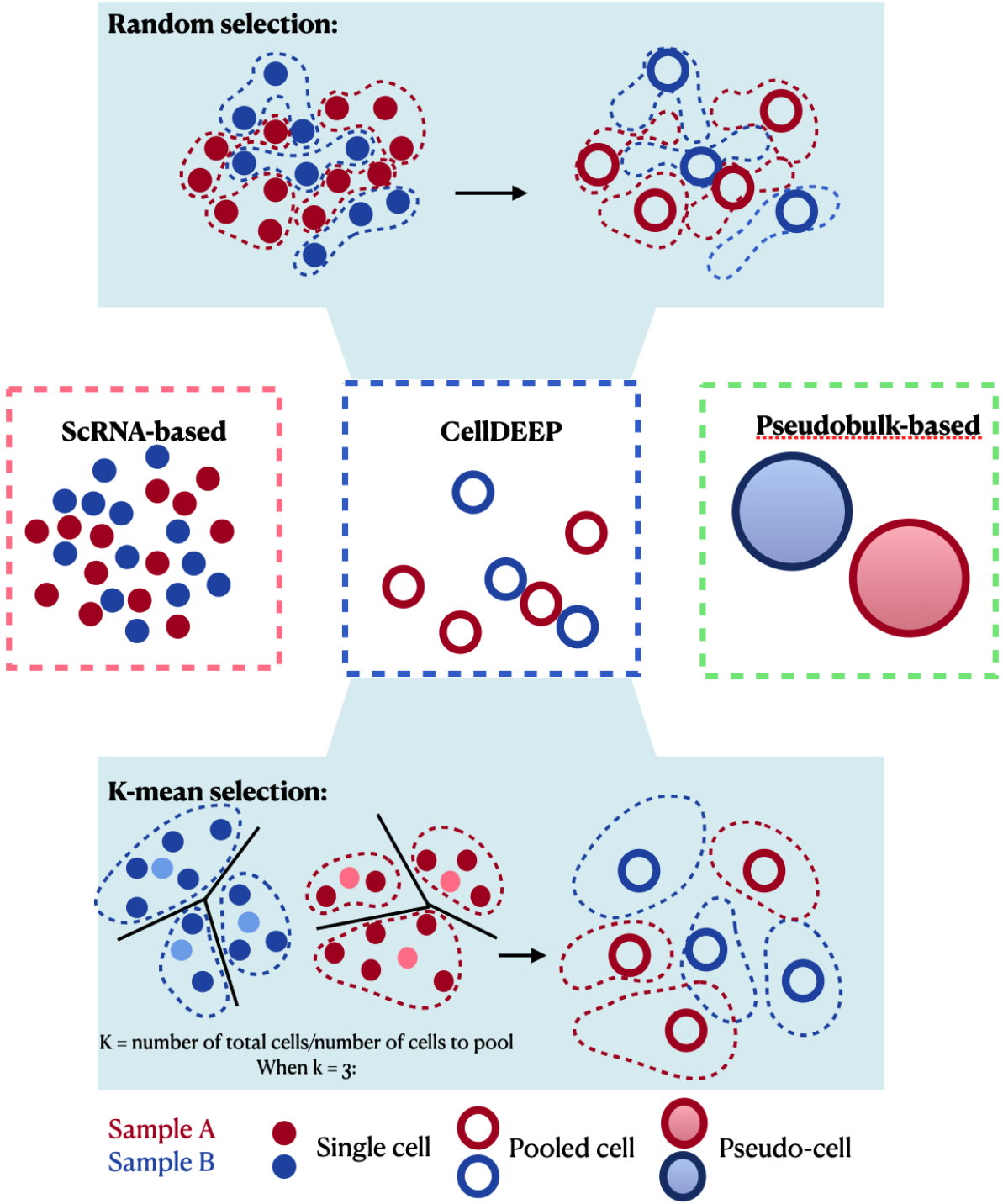
- Cells from one biological replicate are not independent

Pooling idea

- scRNA-Seq – MAST good but not enough signal per cell
- pseudo bulk good – but not enough power!

- Idea is to do something in the middle!!

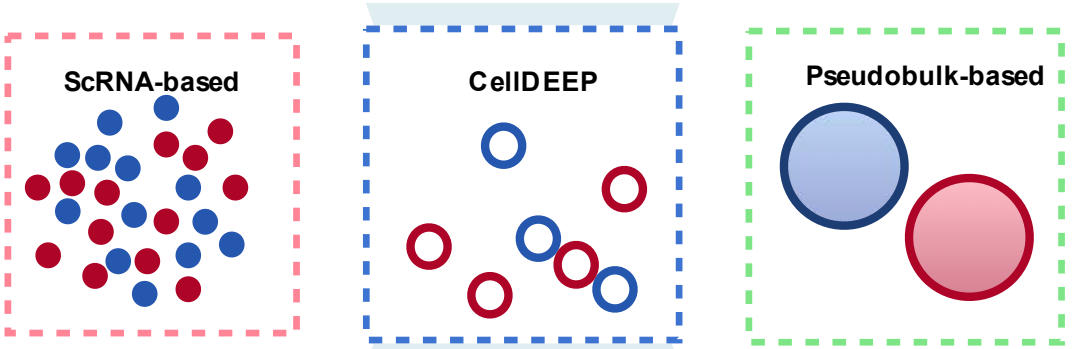
CellDeep

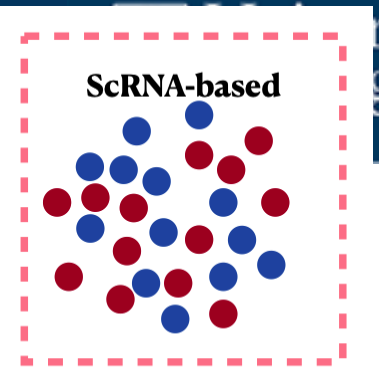


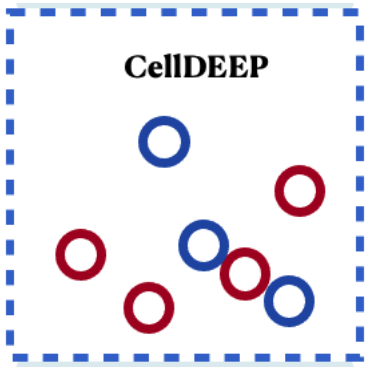
Yiyi Cheng 2021

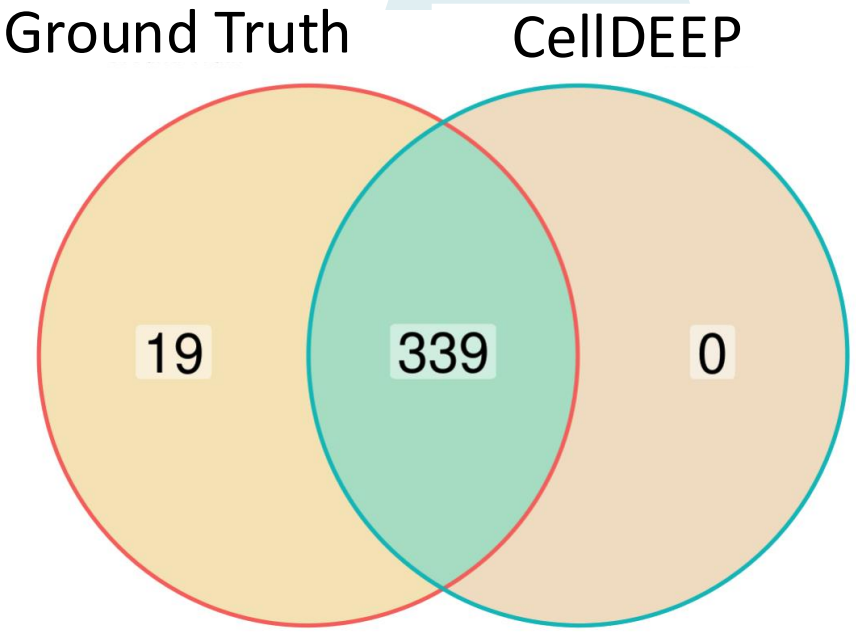
Yiyi has a solution

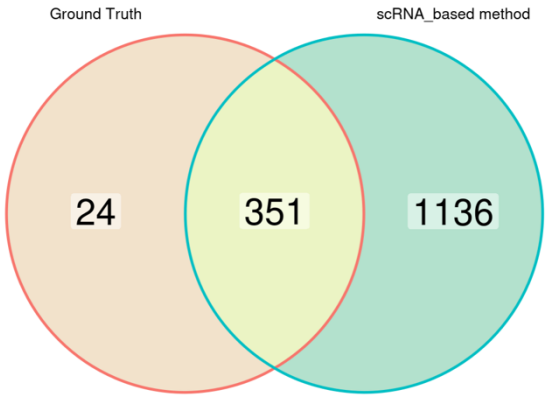
CellDEEP - Cell DiffErential Expression by Pooling

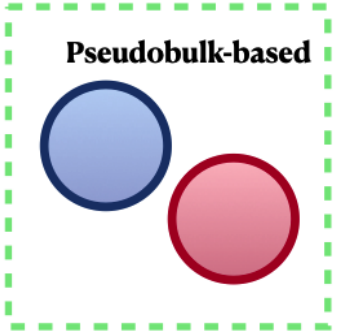


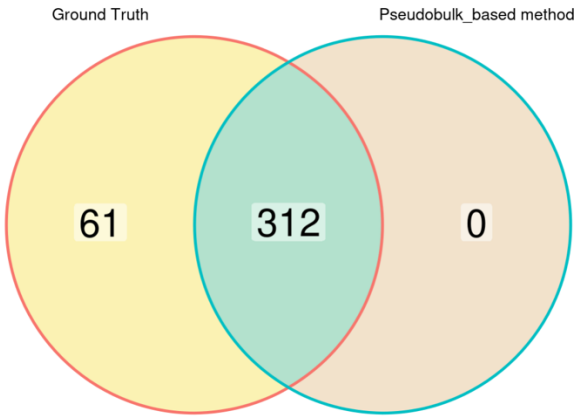




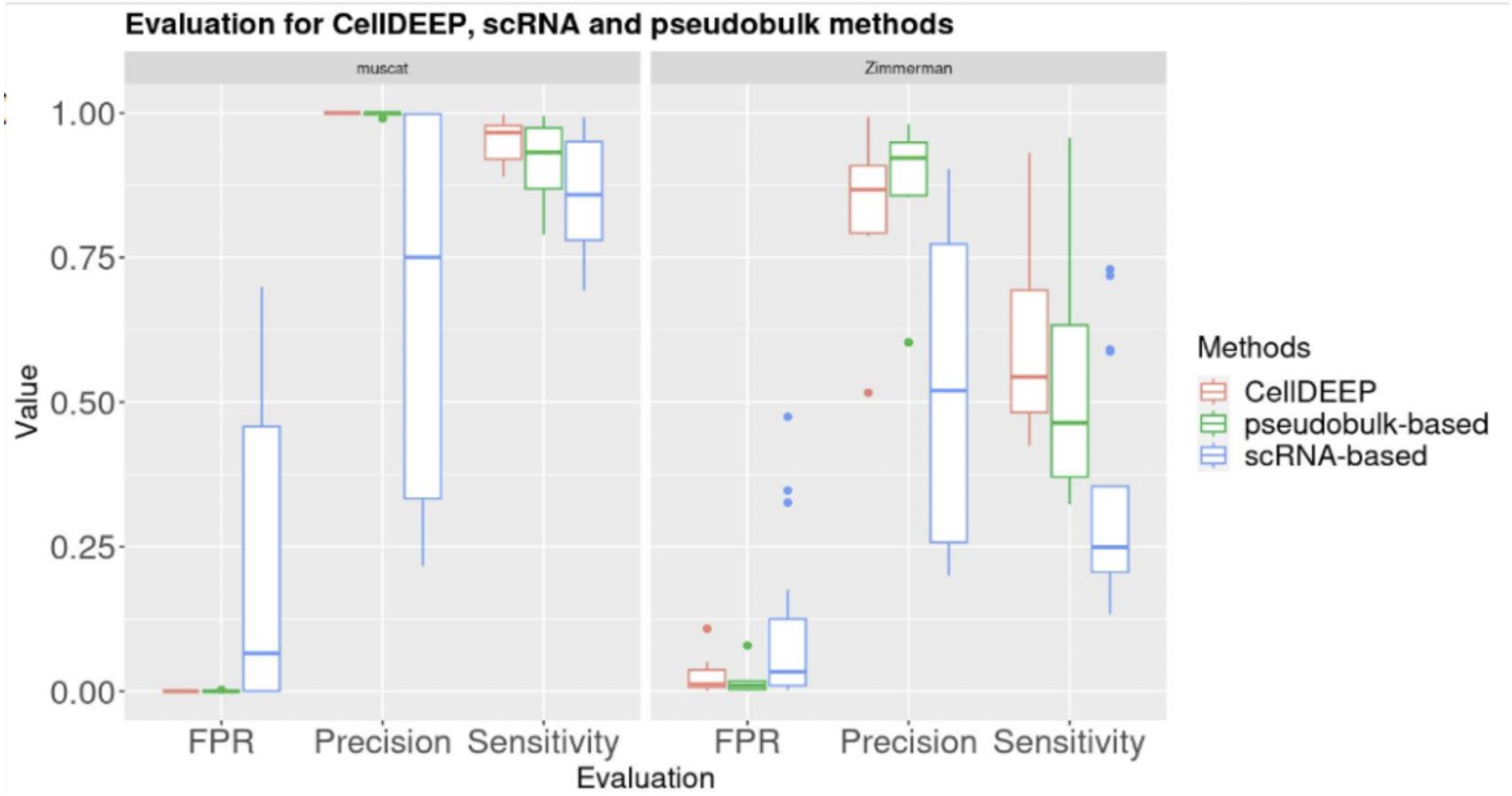






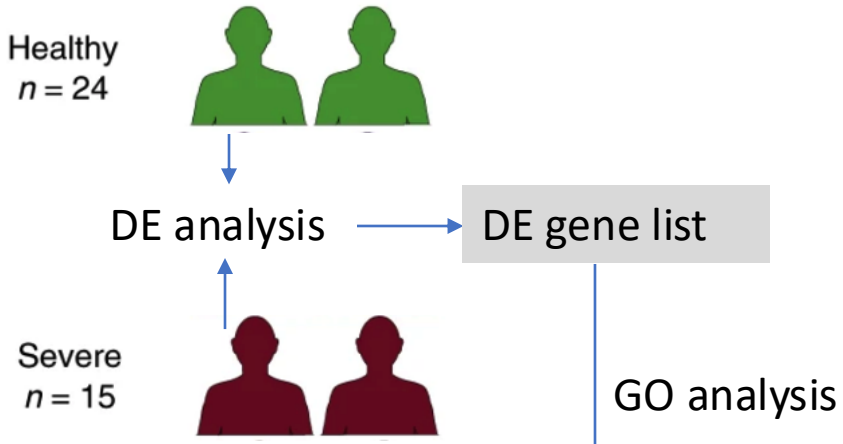


Simulation – CellDepp is pretty good!



Result 2: Covid dataset

Test 2: GO analysis

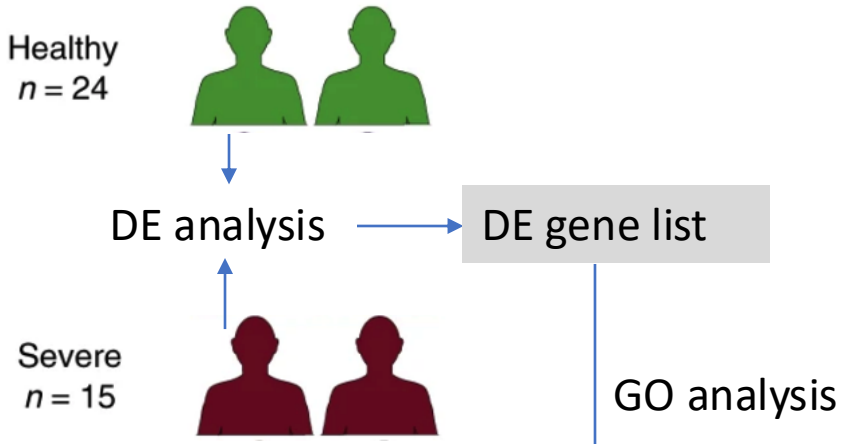


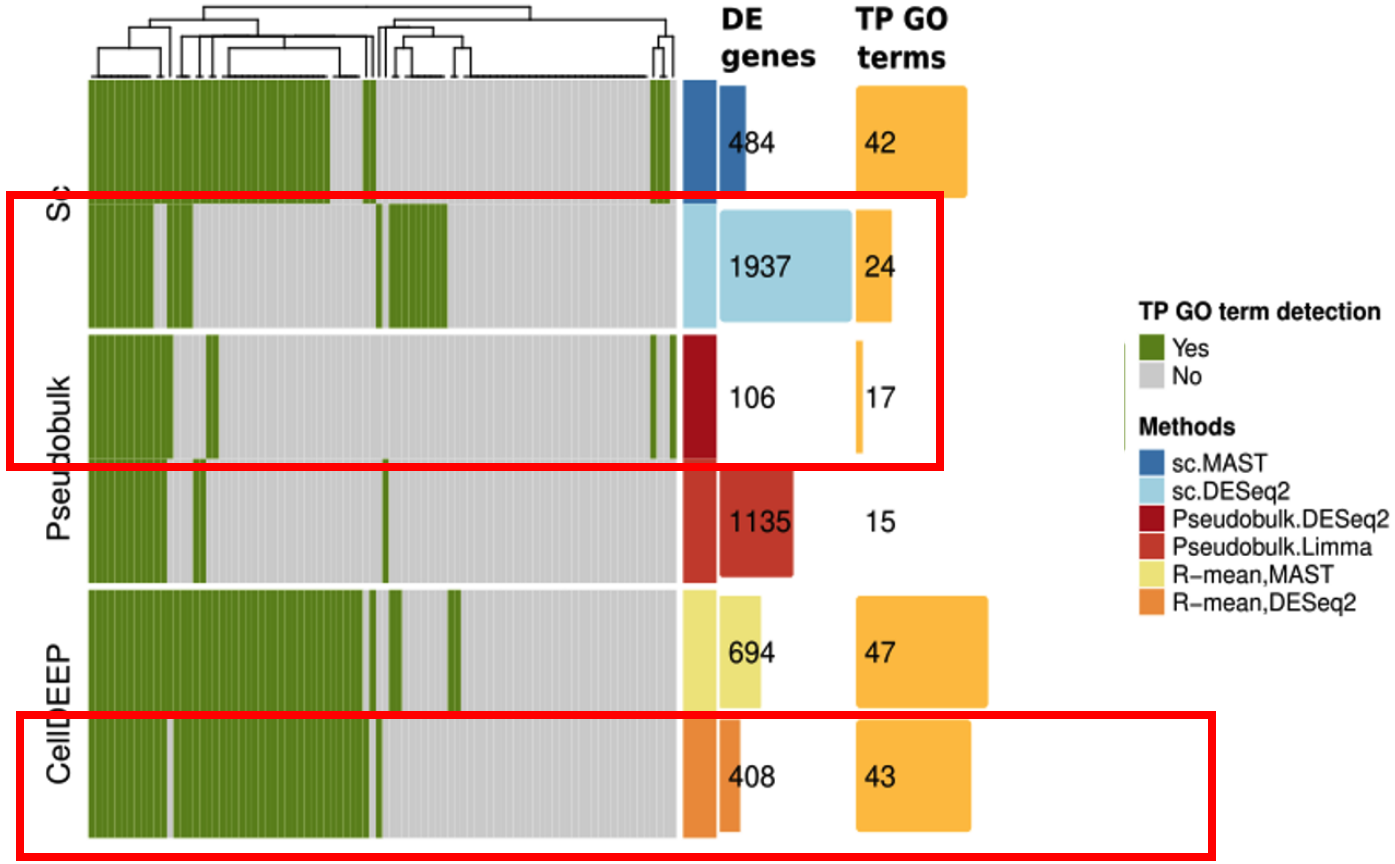
Check overlap!

90 pre-selected TP GO terms(@ Domenico Somma)

Result 2: Covid dataset

Test 2: GO analysis





Check overlap!

90 pre-selected TP GO terms

Result 2: Covid dataset

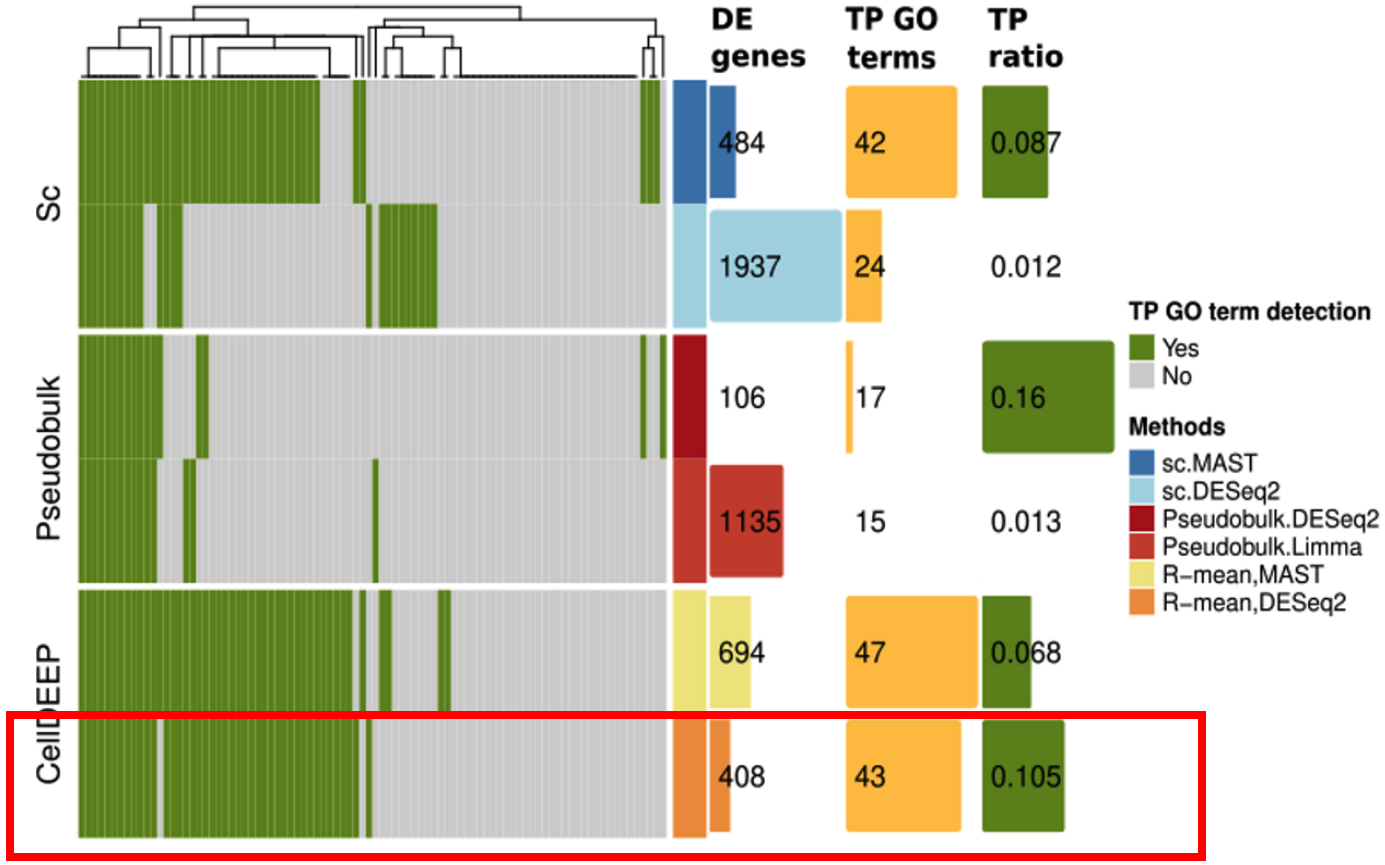
Test 2: GO analysis

TP: CellDEEP > scRNA > Pseudobulk.

Same as simulation!

Not same as simulation:

Pseudobulk also find FP.



Conclusion DEG

- An open problem
- Be aware of limits and compare different methods

2. Clustering

- A dimension reduction methods, does not perform a grouping (clustering) of cells, but just projects points from a high dimensional space into 2d
- So how can we cluster our data, if PCA and t-SNE are just visualization tools?

https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/

Hierarchical clustering

Connectivity models: As the name suggests, these models are based on the notion that the data points closer in data space exhibit more similarity to each other than the data points lying farther away. These models can follow two approaches. In the first approach, they start with classifying all data points into separate clusters and then aggregating them as the distance decreases. In the second approach, all data points are classified as a single cluster and then partitioned as the distance increases. Also, the choice of distance function is subjective. These models are very easy to interpret but lacks scalability for handling big datasets. Examples of these models are **hierarchical clustering** algorithm and its variants.

Distances between 2 vectors

- Euclidean distance: $||a-b||_2 = \sqrt{(\Sigma(a_i-b_i))}$
- Squared Euclidean distance: $||a-b||_2^2 = \Sigma((a_i-b_i)^2)$
- Manhattan distance: $||a-b||_1 = \Sigma|a_i-b_i|$
- Maximum distance: $||a-b||_{INFINITY} = max_i|a_i-b_i|$
- Mahalanobis distance: $\sqrt{((a-b)^T S^{-1}(-b))}$ {where, s : covariance matrix}

```
euclidean <- function(a, b) sqrt(sum((a - b)^2))
```

h	i
1	2
2	2
2	3
2	2

i	j
2	9
2	8
3	7
2	8

What is the difference between squared and simple Euclidean distance?

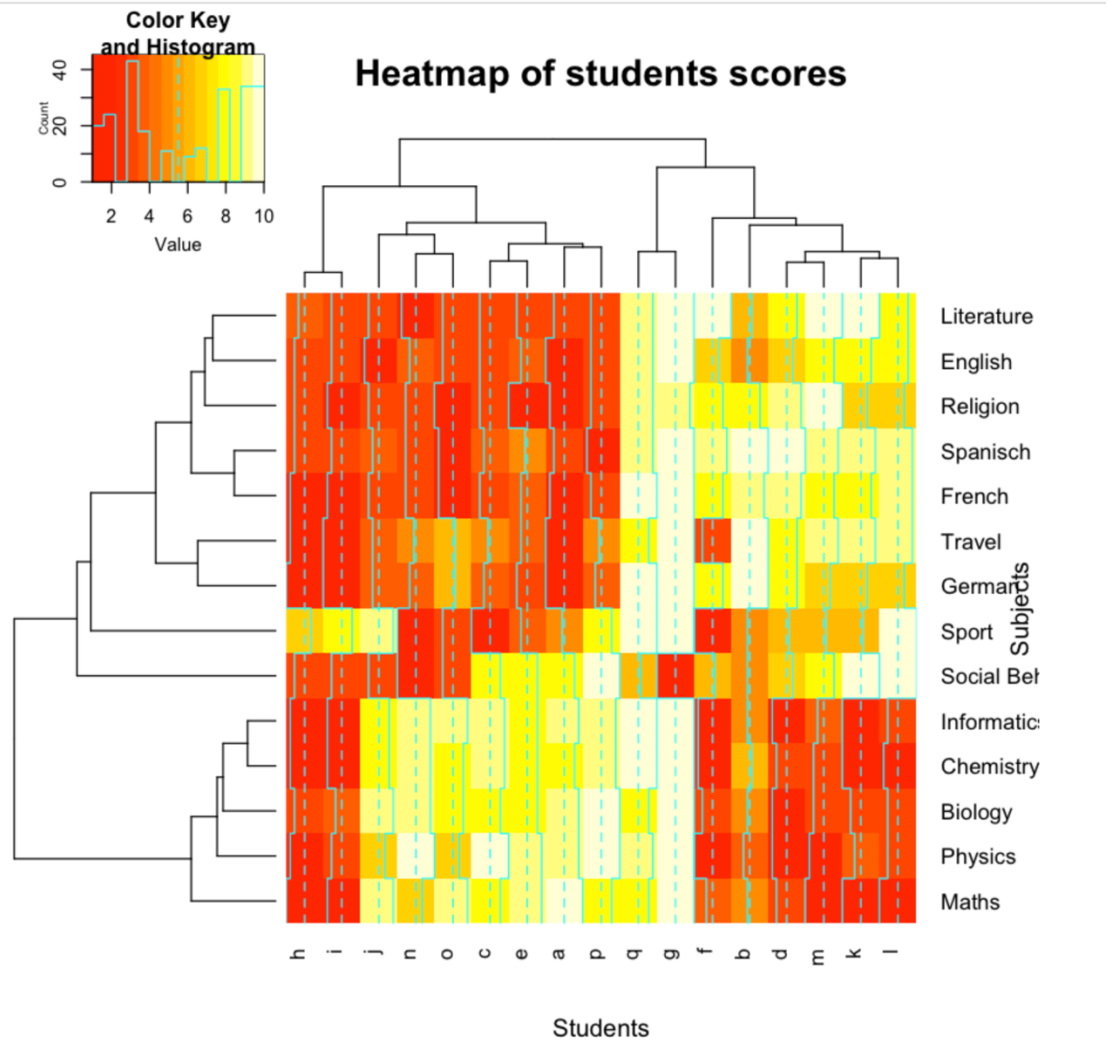
How to visualize this dataset?

Student	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q
Maths	10	5	8	3	9	4	10	1	2	9	1	1	1	7	9	8	8
Informatics	9	5	9	2	8	2	10	2	2	8	1	3	4	9	9	9	10
Physics	9	4	10	2	9	1	10	2	3	7	4	3	2	10	7	10	9
Chemistry	8	6	9	3	8	2	10	2	2	8	1	2	3	9	8	9	10
Biology	9	5	8	2	8	3	10	3	4	9	3	3	3	9	8	10	8
German	1	10	4	8	3	8	10	1	1	4	7	7	7	4	6	4	10
English	1	5	3	7	4	7	10	3	3	2	8	8	8	4	3	3	9
Sport	5	5	1	6	4	1	10	7	8	9	6	10	6	1	3	8	10
Travel	1	10	5	8	4	3	10	2	1	4	9	9	9	5	6	5	8
Social Behav	8	5	8	7	8	6	1	3	3	3	10	10	8	1	3	10	6
Literature	3	6	3	8	3	10	10	4	3	3	10	8	10	2	3	3	9
French	2	9	3	9	4	8	10	2	2	3	8	9	8	3	2	3	10
Spanisch	3	10	4	10	5	9	10	3	3	4	9	9	9	3	2	2	9
Religion	2	8	3	9	1	8	9	3	2	3	7	7	10	3	1	3	9

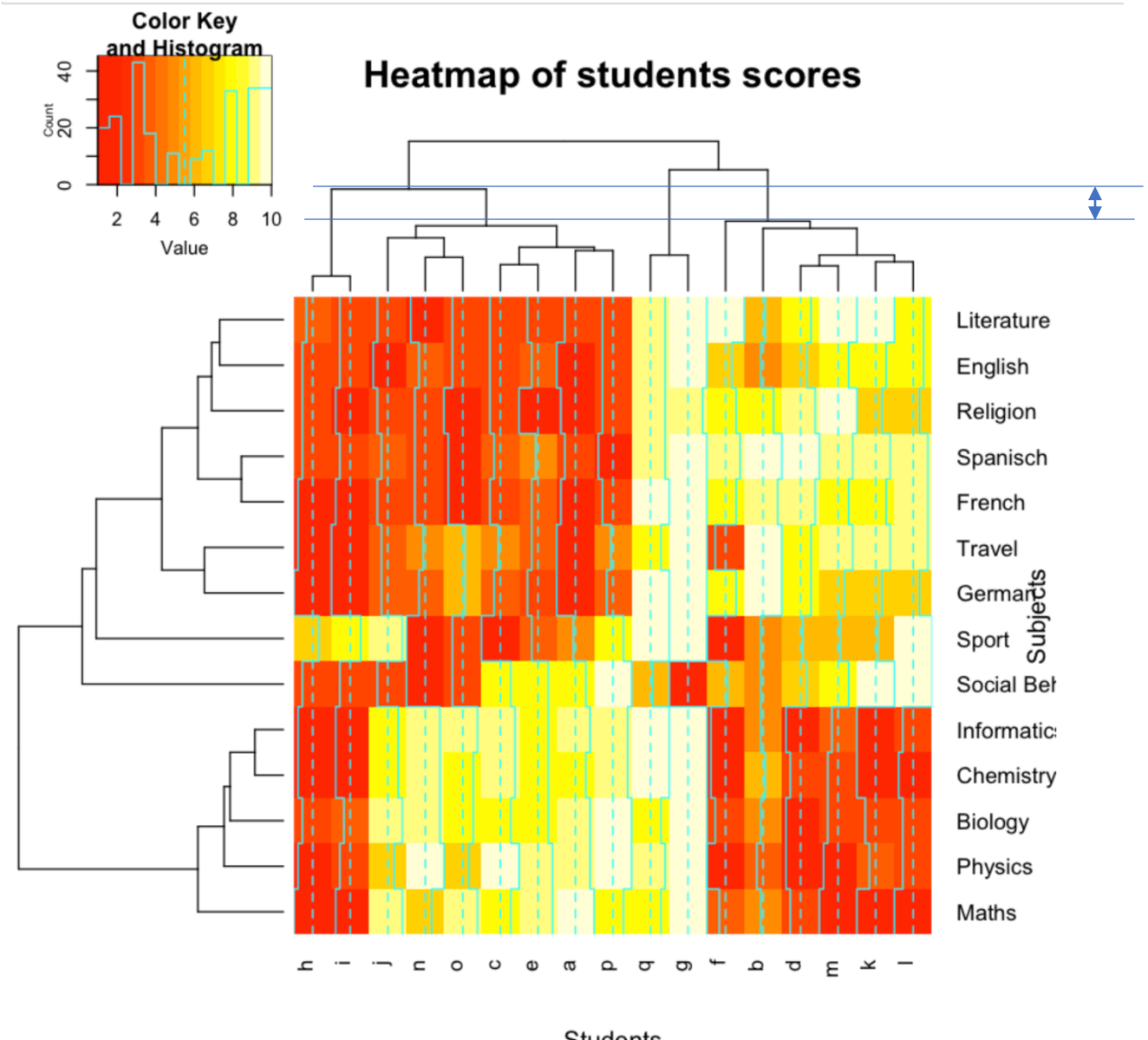
Hierarchical Clustering

- Think of the heatmap
- Each column is its own cluster
- group to closed cluster, based distance, and form new group
- This generates a dendrogram

- Heatmap in R uses this clustering (this afternoon)



How many clusters?

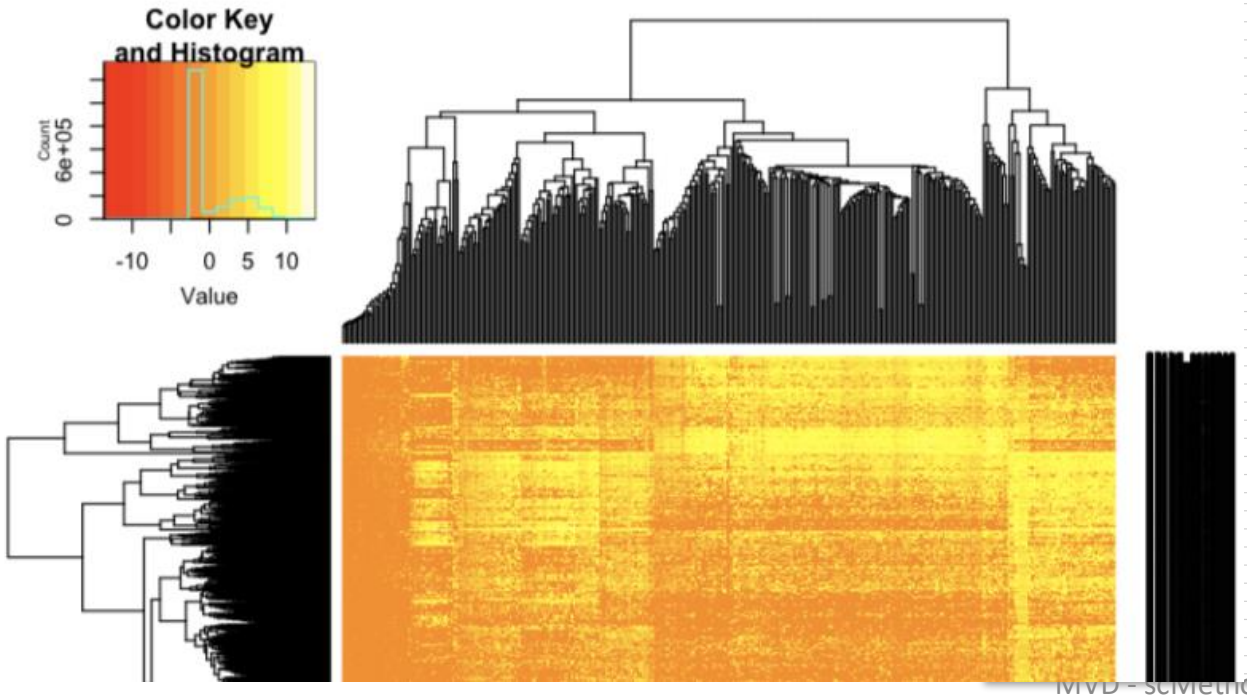


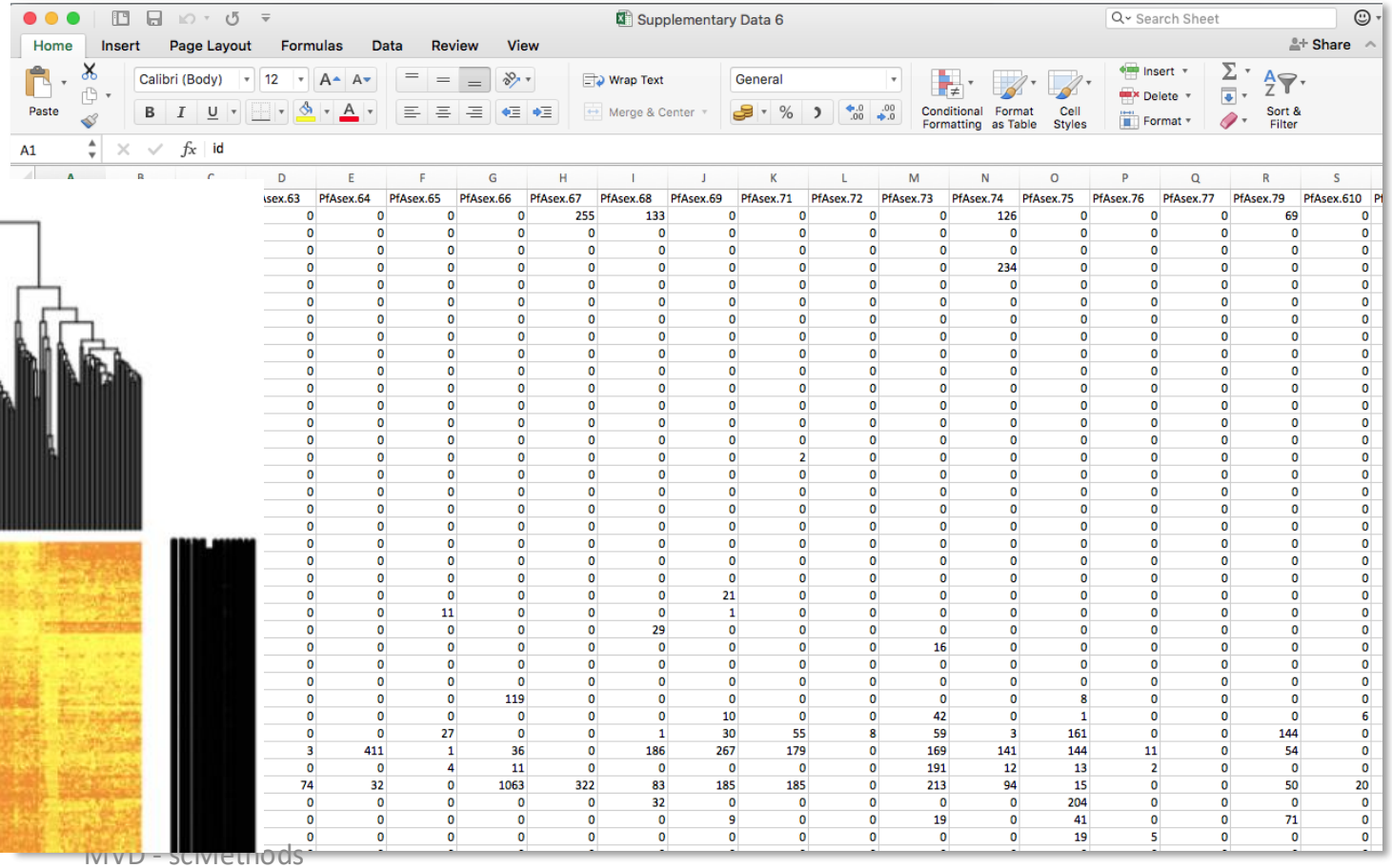
Maximal distance, seems here, so 4 groups

Does it make sense?

Output of single cell data is “HUGE”

- A table of ~20 thousands rows and ~10 thousands cells per run
- Each cell (column) is described by ~2,500 genes (row)
- Difficult to visualise this

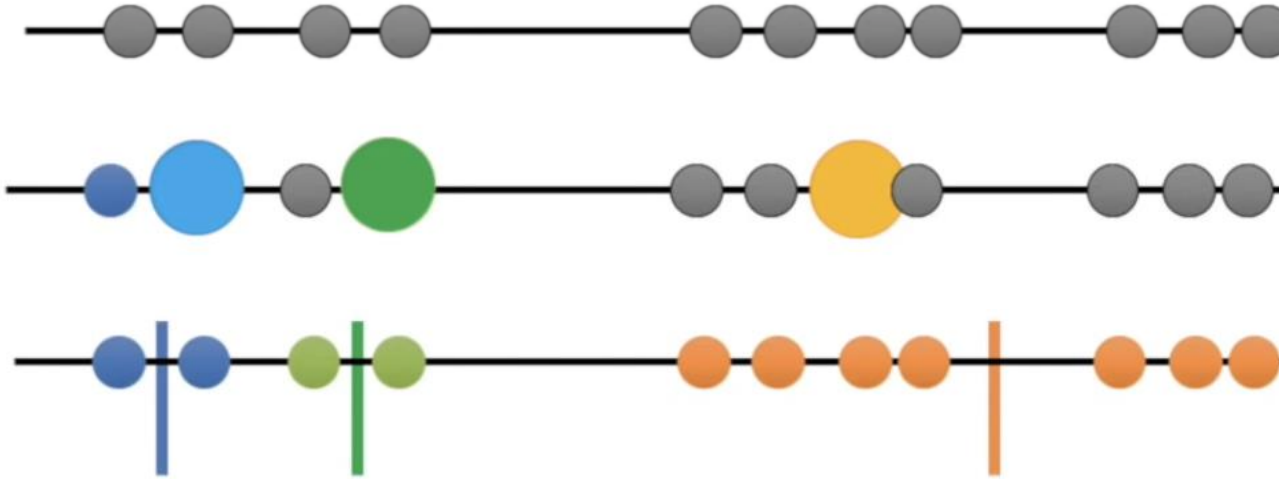


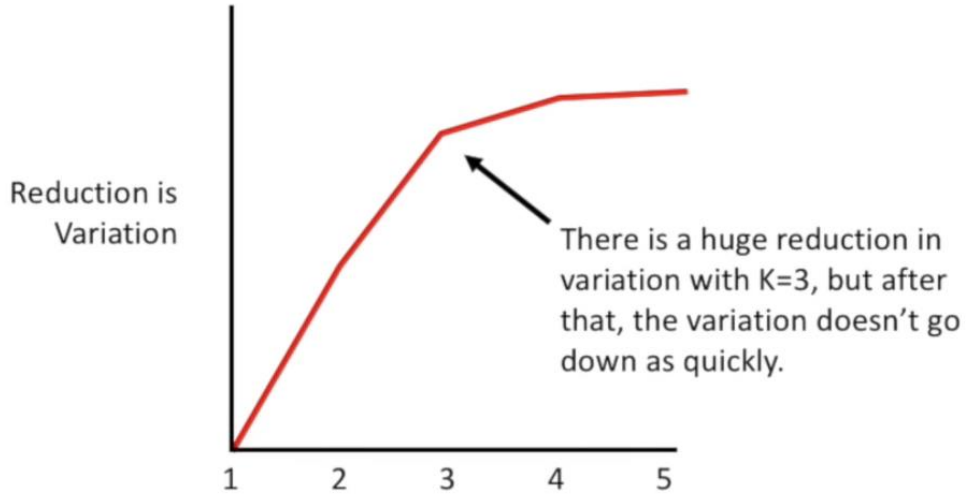


Centroid models: These are iterative clustering algorithms in which the notion of similarity is derived by the closeness of a data point to the centroid of the clusters**. K-Means clustering algorithm** is a popular algorithm that falls into this category. In these models, the number of clusters required at the end have to be mentioned beforehand, which makes it important to have prior knowledge of the dataset. These models run iteratively to find the local optima.

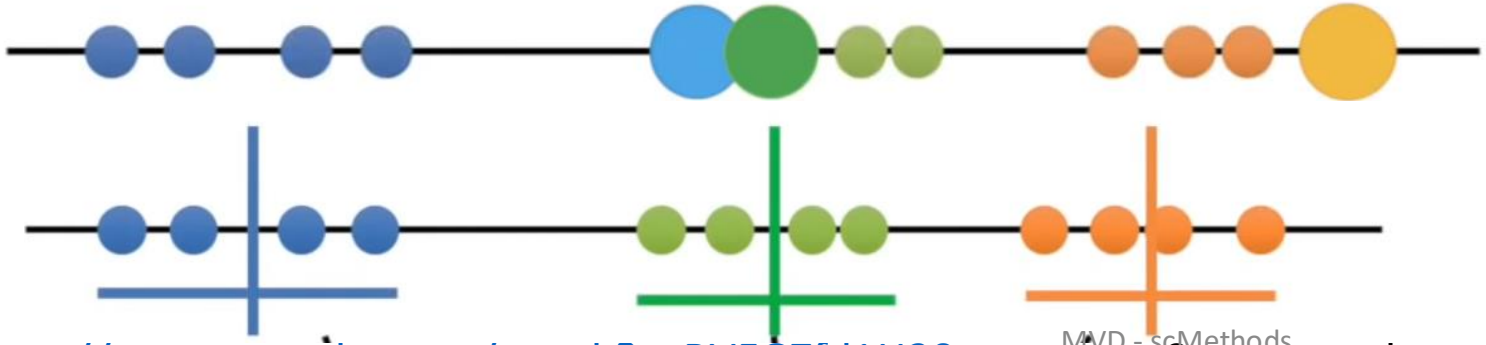
K-means

- Iterative approach; Easy to use; Random, difficult to get K





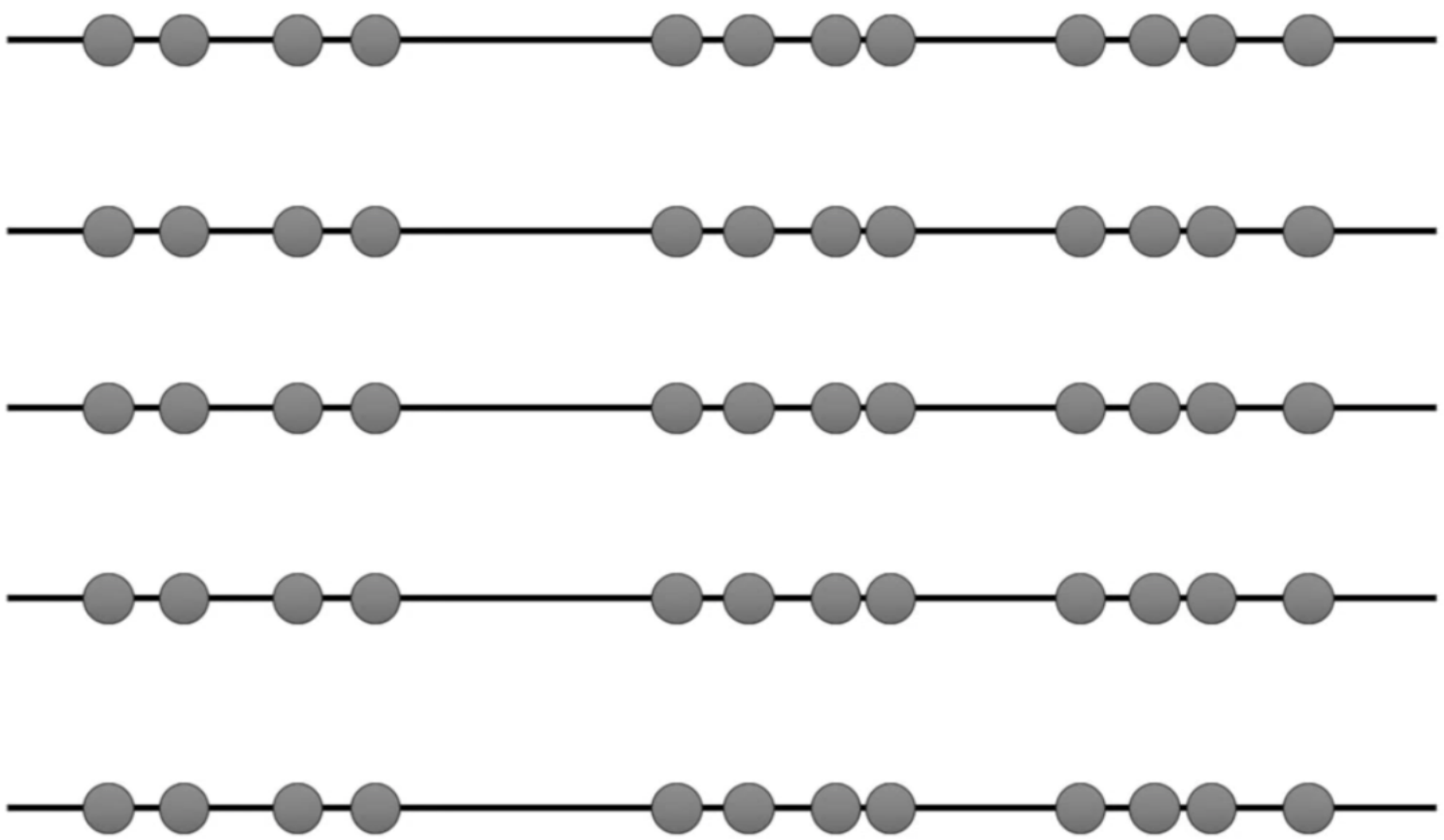
Different initialisation



https://www.youtube.com/watch?v=BVFG7fd1H30 more on Stat quest: https://www.youtube.com/watch?v=4b5d3muPQmA

K-means

- Iterative approach; Easy to use; Random, difficult to get K

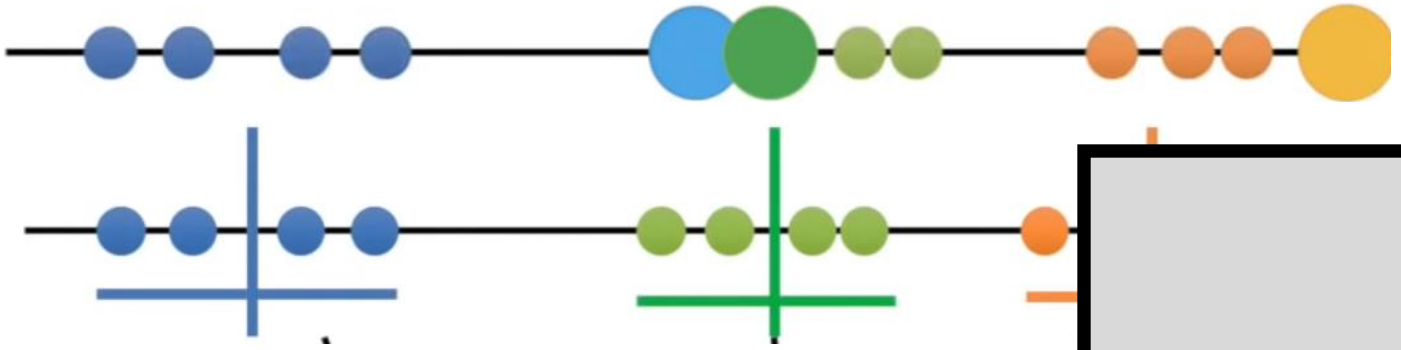


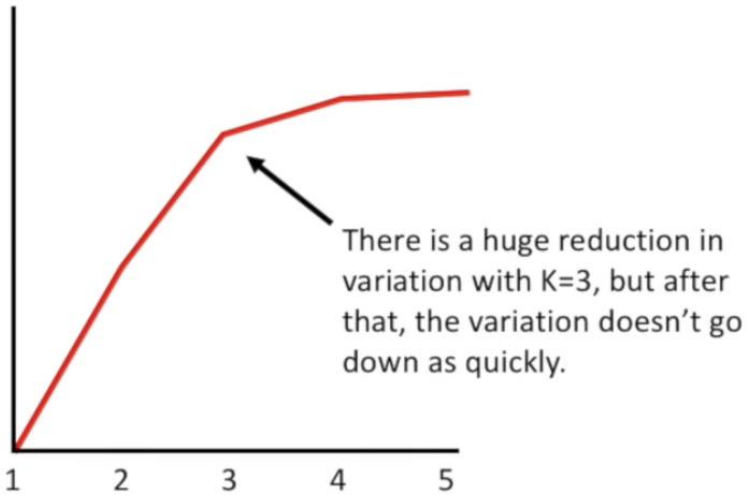
https://www.youtube.com/watch?v=BVFG7fd1H30 more on Stat quest: https://www.youtube.com/watch?v=4b5d3muPQmA

K-means

- Iterative approach; Easy to use; Random, difficult to get K

Iteration random initialisation to not just find local minima.



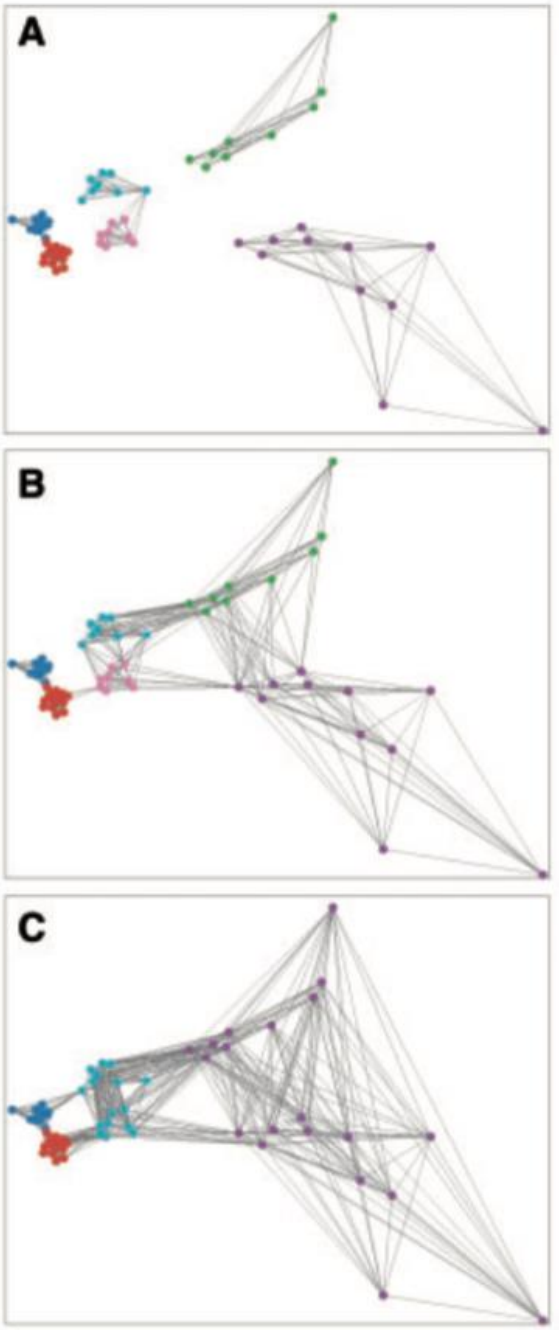


BTW, this is considered Machine Learning

https://www.youtube.com/watch?v=BVFG7fd1H30 more on Stat quest: https://www.youtube.com/watch?v=4b5d3muPQmA

Which clustering in scRNA-Seq?

- First k-means
- But k-means is not good for noisy high dimensional data
- shared nearest neighbor (SNN) -> graph theory-based algorithms
- Euclidean norm between two cells (expression difference of genes)
- Build graph
- Find cliques (connections between nodes)
- FindClusters(pbmc, **resolution = 0.5**)



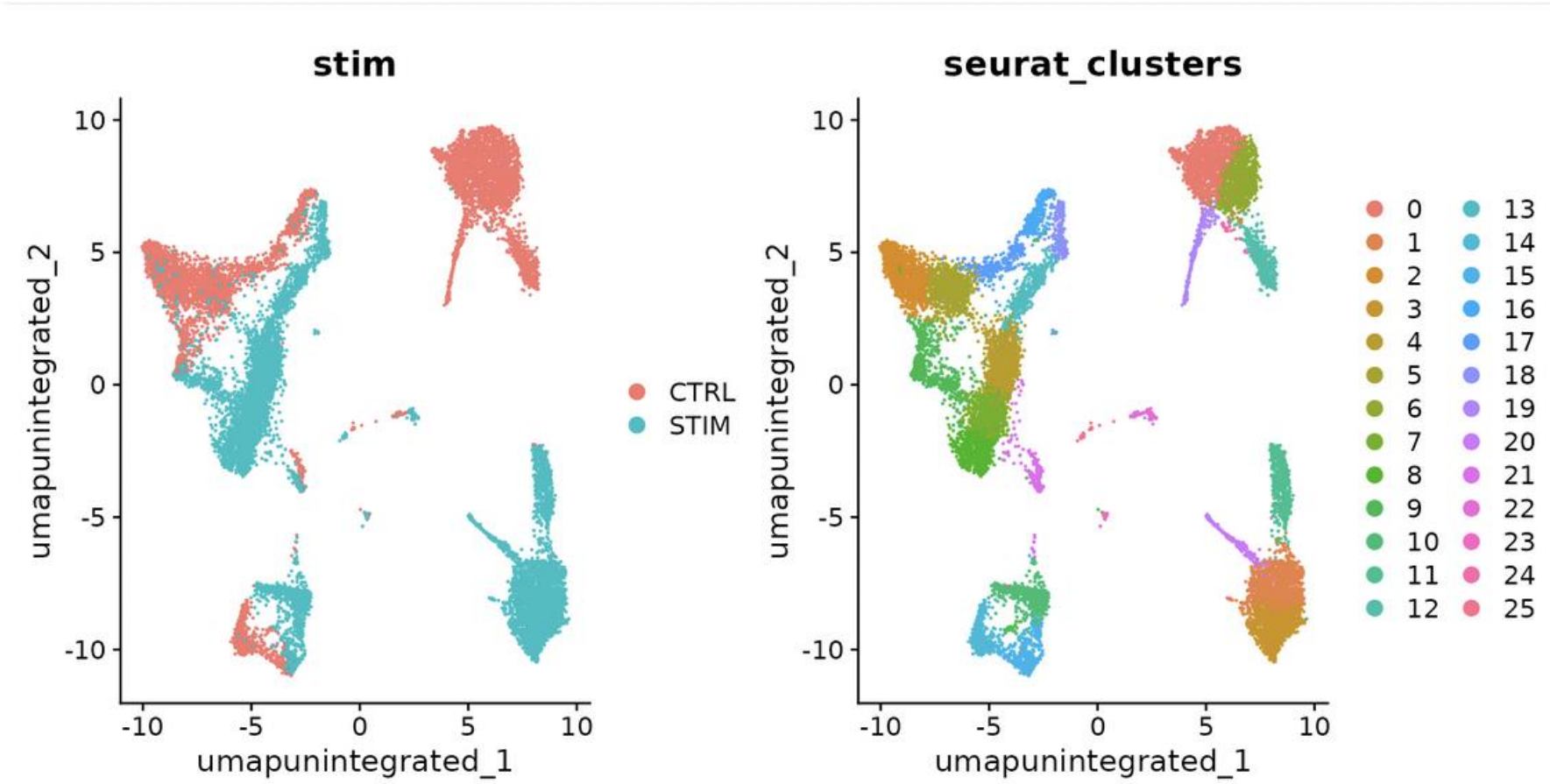
https://academic.oup.com/bioinformatics/article/31/12/1974/214505

Conclusion clustering

- Actually not so difficult!

3. Integration

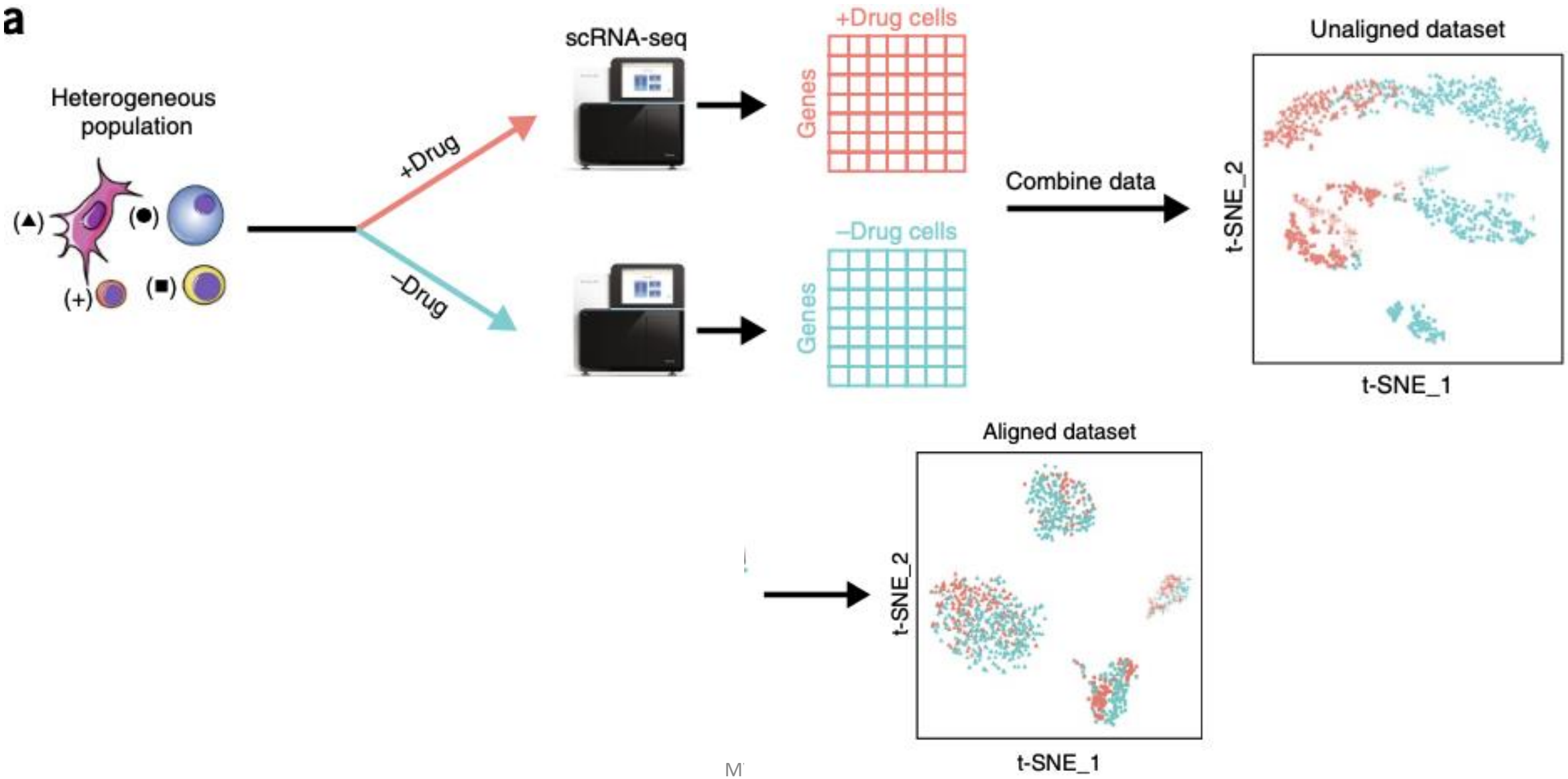
- Why do we need to integrate our data?



Integration

- **Preserve biological variation**: Ensure that meaningful biological differences (e.g., cell types or states) are retained.
- **Remove batch effects**:
  - Different protocols (amount UMI, different machines)
  - Different days
  - Different person
  - scRNA-Seq is just different
- Seurat CCA methods
- BBKNN
- Harmony
- STACAS

Seurat CCA – “default”



Canonical Correlation Analysis (CCA)

- CCA is a statistical method that identifies **linear relationships** between two datasets by finding pairs of vectors (called **canonical vectors**) that maximise the correlation between the datasets.
- For two datasets Y_1 and Y_2, CCA finds vectors u_1 and u_2 such that the correlation between Y_1 and Y_2 is maximised.
- In Seurat, CCA is applied to the normalised expression matrices of the two datasets, using the shared HVGs.

$$\mathrm{CCA}: \max_{\boldsymbol{u}_1,\boldsymbol{u}_2} \rho = \mathrm{corr}\left(\boldsymbol{Y}_1\boldsymbol{u}_1, \boldsymbol{Y}_2\boldsymbol{u}_2\right) = \frac{\boldsymbol{u}_1^T \mathbf{0\text{-}k} \sum_{12} \boldsymbol{u}_2}{\sqrt{\boldsymbol{u}_1^T \sum_{11} \boldsymbol{u}_1} \sqrt{\boldsymbol{u}_2^T \sum_{22} \boldsymbol{u}_2}}.$$

 https://pmc.ncbi.nlm.nih.gov/articles/PMC7416047/

How is roughly works

i. learns a shared gene correlation structure that is conserved between the data sets using canonical correlation analysis
ii. identifies individual cells that cannot be well described by this shared structure
iii. aligns the data sets into a conserved low-dimensional space (latent space / PCA), using nonlinear ‘warping’ algorithms to normalise for differences in feature scale in a manner that is robust to shifts in population density
iv. proceeds with an integrated downstream analysis, for example, identifying discrete subpopulations through clustering, or reconstructing continuous developmental processes
v. It performs comparative analysis on aligned subpopulations

How is roughly works

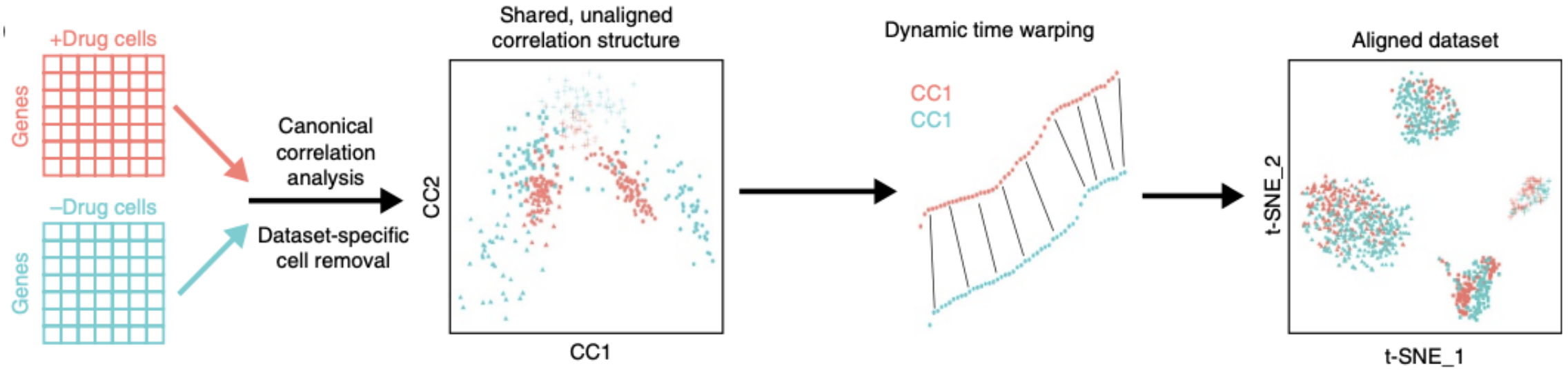
i. learns a shared gene correlation structure that is conserved between the data sets using canonical correlation analysis

ii. identifies individual cells that cannot be well described by this shared structure

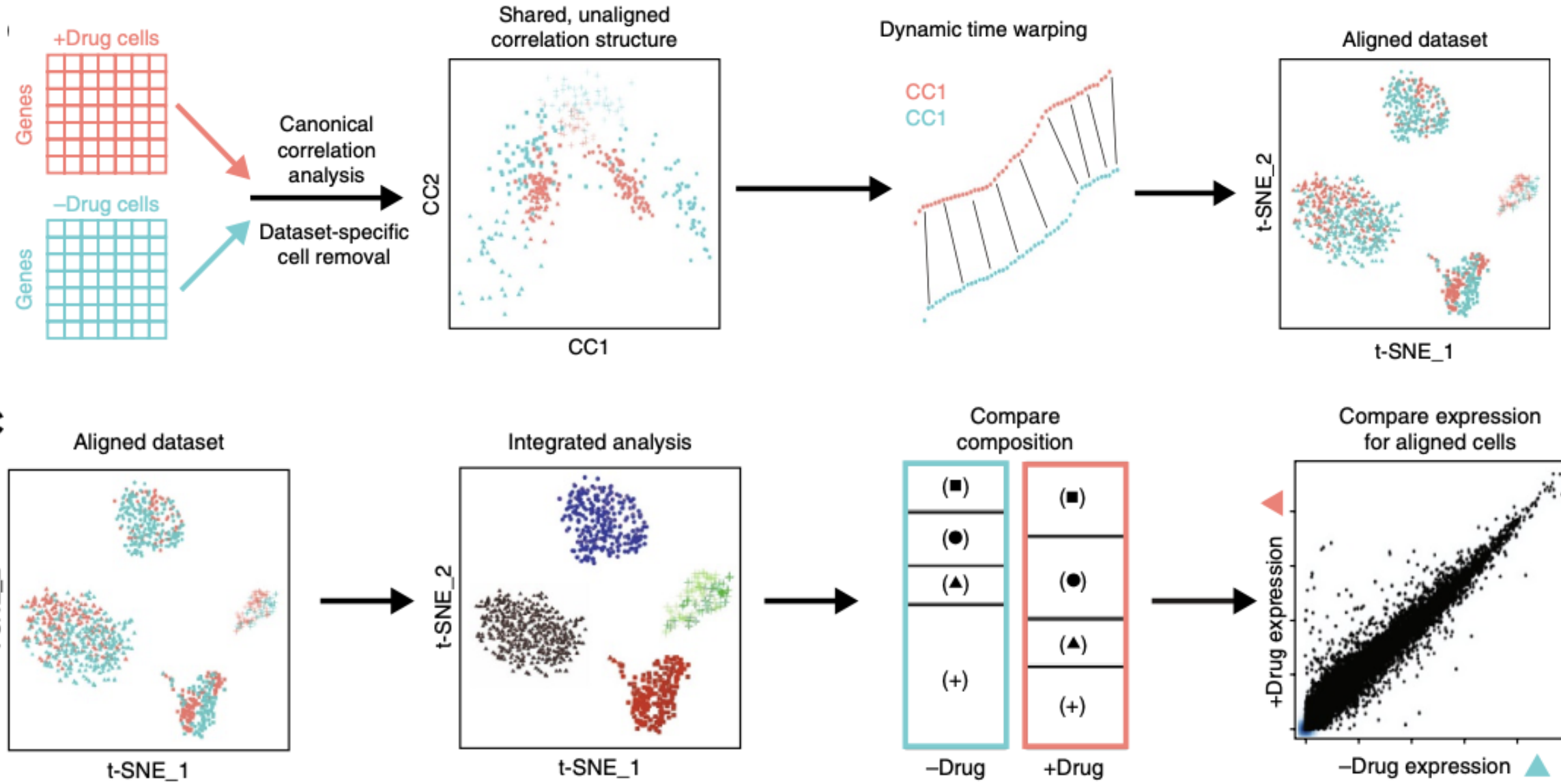
iii. aligns the data sets into a conserved low-dimensional space, using nonlinear ‘warping’ algorithms to normalize for differences in feature scale, in a manner that is robust to shifts in populations density

iv. proceeds with an integrated downstream analysis, for example, identifying discrete subpopulations through clustering, or reconstructing continuous developmental processes

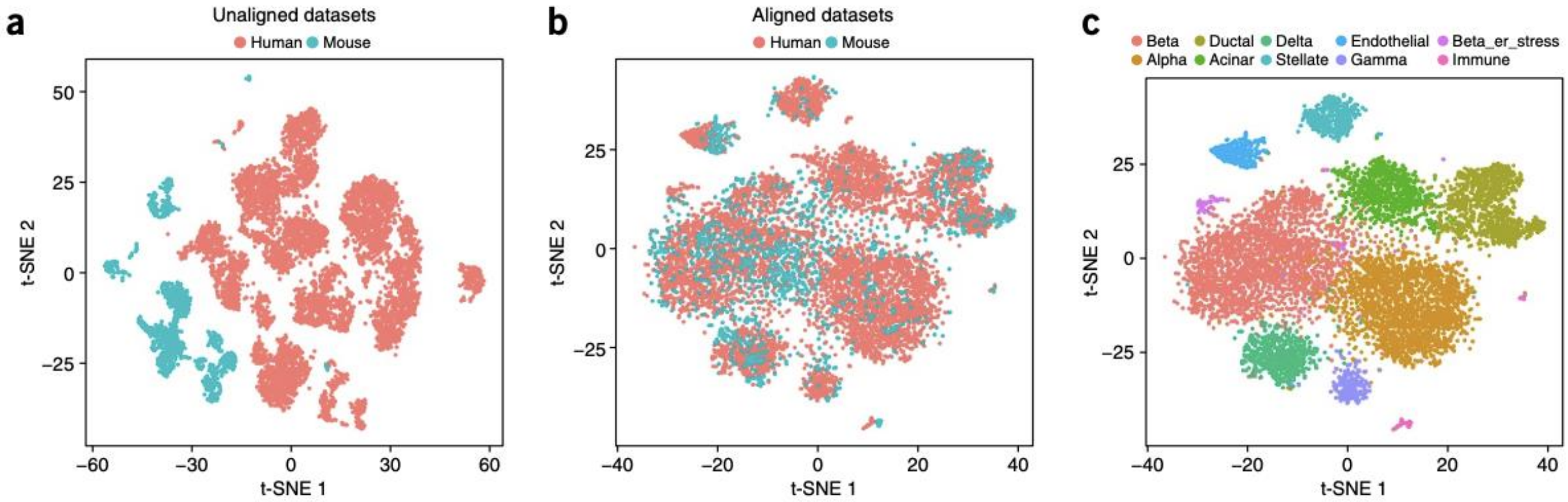
v. It performs comparative analysis on aligned subpopulations



How is roughly works



Works nice on datasets of paper

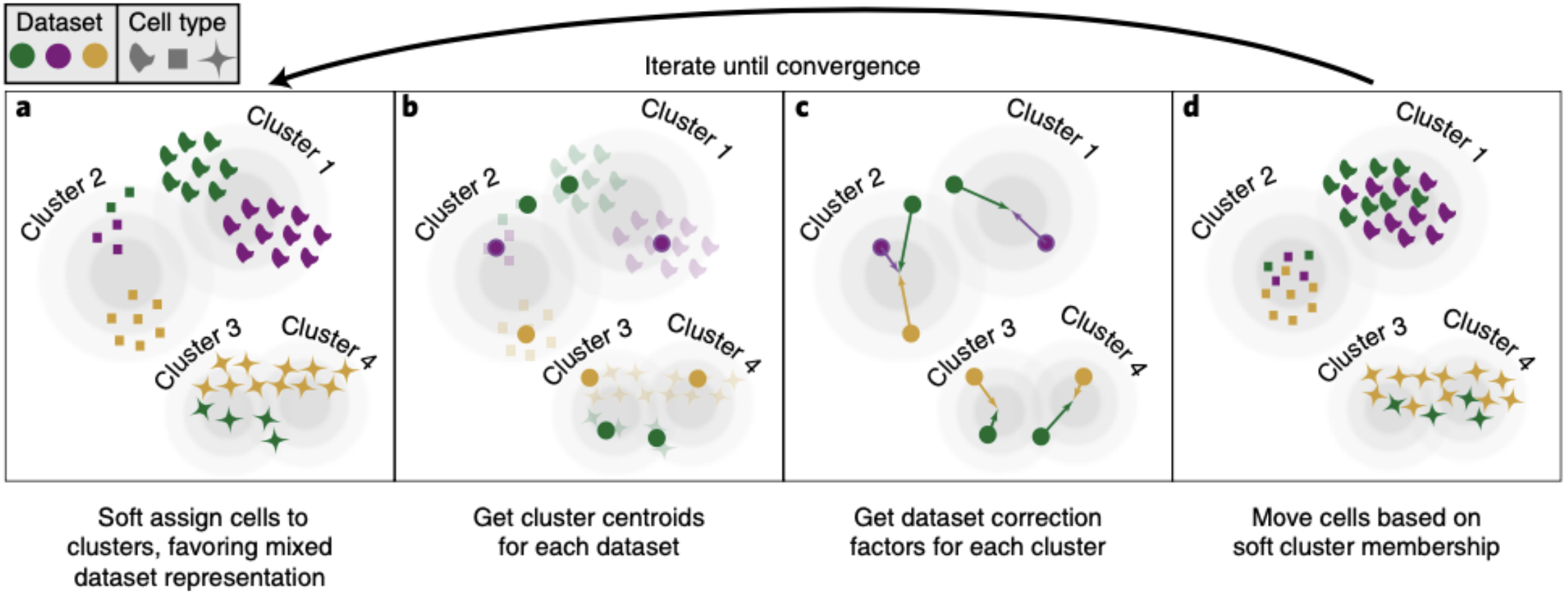


Joint identification of cell types across human and mouse islet scRNA-seq atlases.

doi:10.1038/nbt.4096

Harmony

- Projects cells into a shared embedding in which cells group by cell type rather than dataset specific conditions



https://www.nature.com/articles/s41592-019-0619-0

Comparison - time

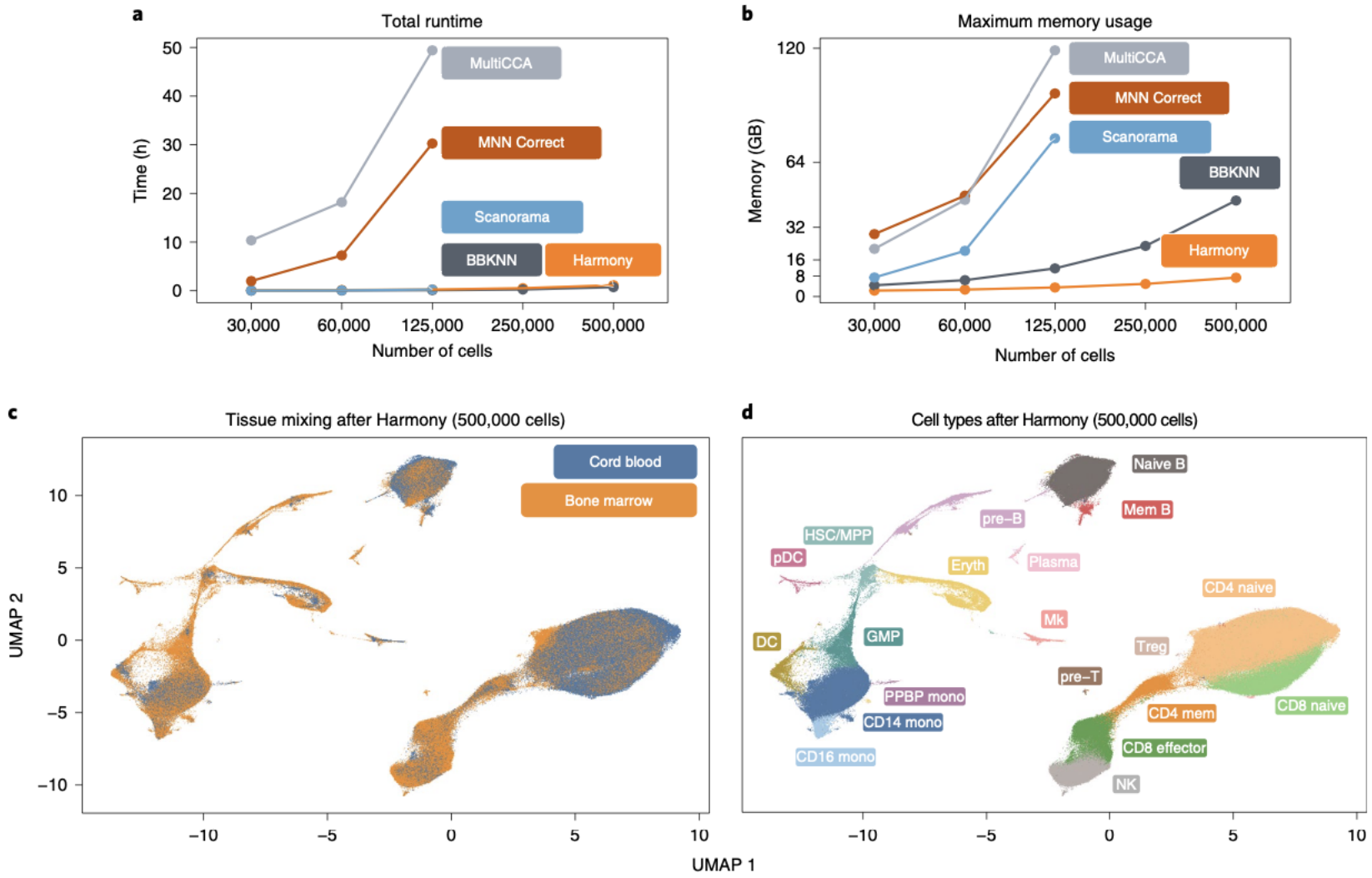


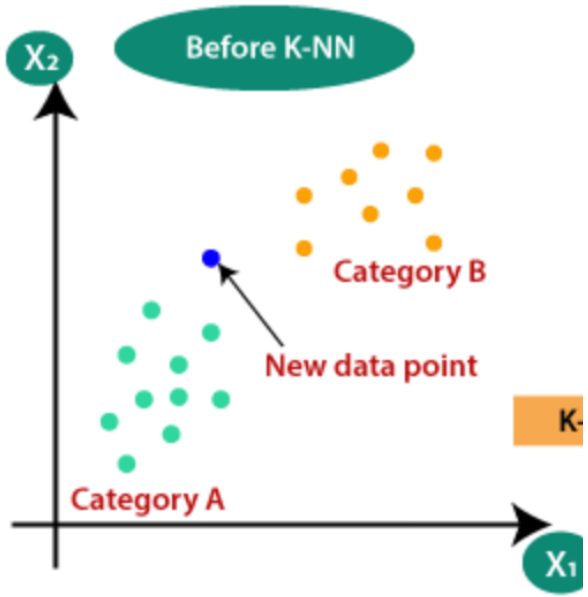
Fig. 3 | Computational efficiency benchmarks. BBKNN, Scanorama, MNN Correct and MultiCCA are compared on five downsampled HCA datasets of increasing sizes. a,**b**, Total runtime (**a**) and maximum memory (**b**) required to analyze each dataset are shown. Scanorama, MultiCCA and MNN Correct were terminated for excessive memory requests on the 250,000 and 500,000 cell datasets. **c**, The mixing between tissues in the Harmony embedding of

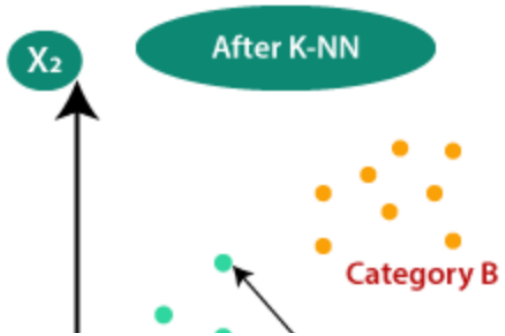
BBKNN – batch balanced k nearest neighbors

University of Glasgow

- Implemented in Python
- **Main assumptions:**
  - **at least some cells of the same type exist across batches**
  - **that the differences between the same cell types across batch caused by batch effects are less than the differences between cells of different types within a batch**
- Says, it is the best ☺

K-Nearest Neighbor(KNN) Algorithm for Machine Learning





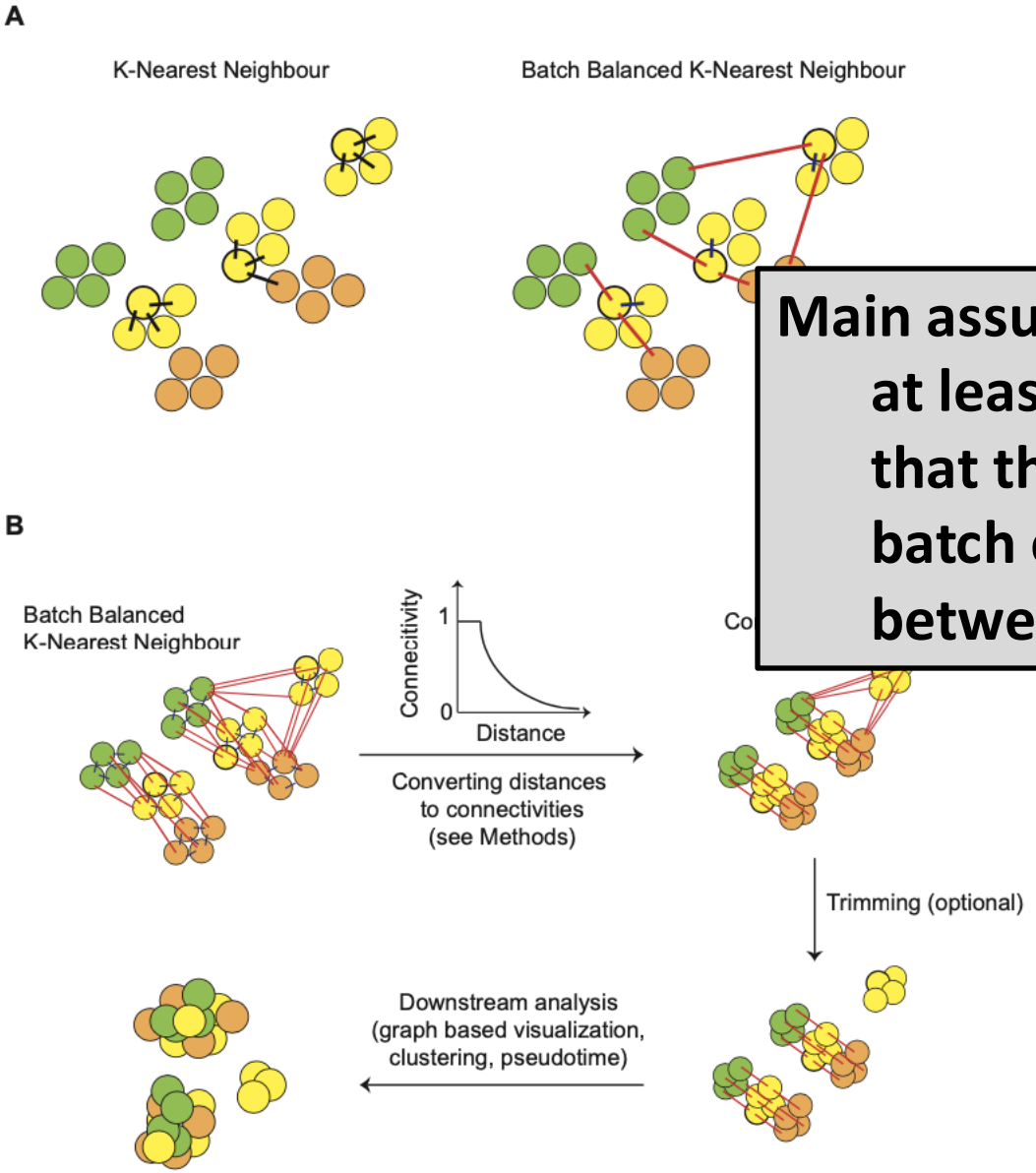
How does K-NN work?

The K-NN working can be explained on the basis of the below algorithm:

- **Step-1:** Select the number K of the neighbors
- **Step-2:** Calculate the Euclidean distance of **K number of neighbors**
- **Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance.
- **Step-4:** Among these k neighbors, count the number of the data points in each category.
- **Step-5:** Assign the new data points to that category for which the number of the neighbor is maximum.
- **Step-6:** Our model is ready.

https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

How is that different?



Supplementary Figure 1: A conceptual schematic of BBKNN's operation. Identifying a cell's k nearest neighbours for the purpose of constructing a KNN graph compared to the batch balanced counterpart in BBKNN (A). The neighbour distance collection is then converted to exponentially related connectivities, which BBKNN trims to weed out any erroneous connections between

Main assumptions:

- **at least some cells of the same type exist across batches**
- **that the differences between the same cell types across batch caused by batch effects are less than the differences between cells of different types within a batch**

$$\sum_{a \in S_c} \alpha_{ca} = \lambda \log_2(|S_c|) \qquad (5)$$

where λ is a bandwidth parameter (set to 1 by default) that controls how quickly the connectivity values decay to 0 with distance.

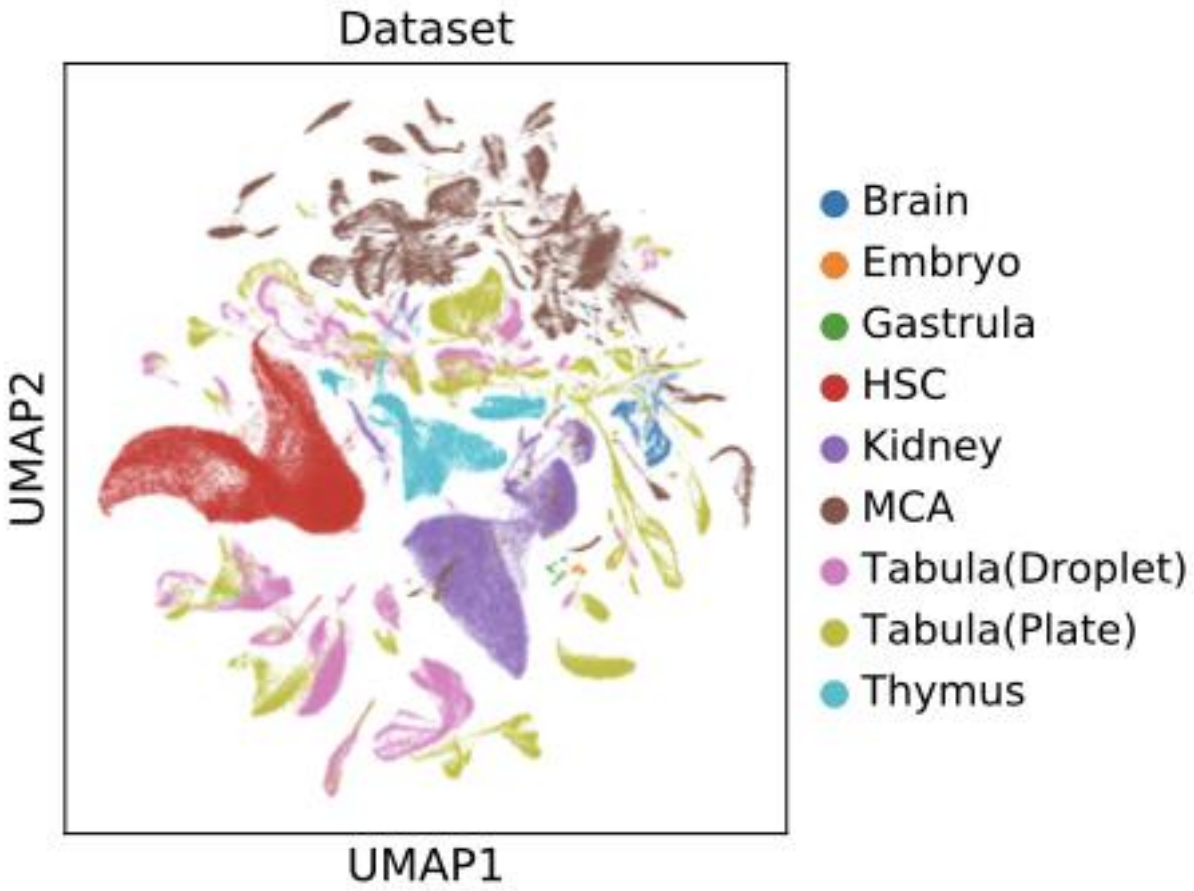
The connectivity score α_{ca} is then made symmetric to give,

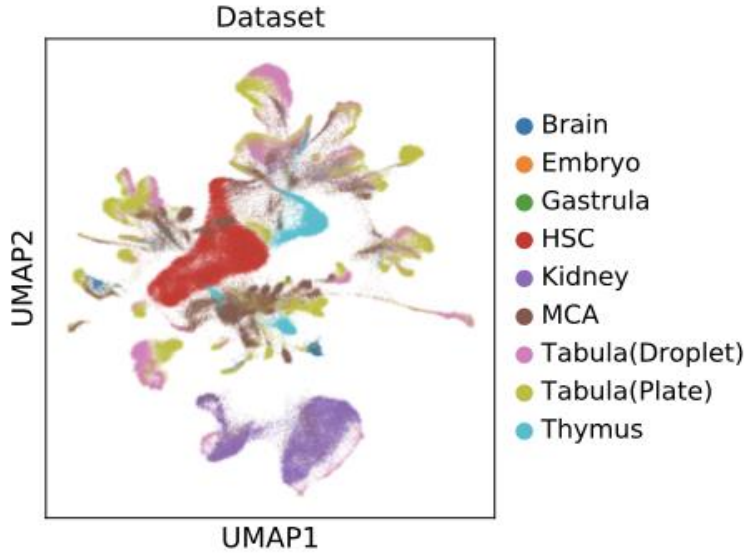
$$w_{ca} = w_{ac} = \alpha_{ac} + \alpha_{ca} - \alpha_{ac}\alpha_{ca} \qquad (6)$$

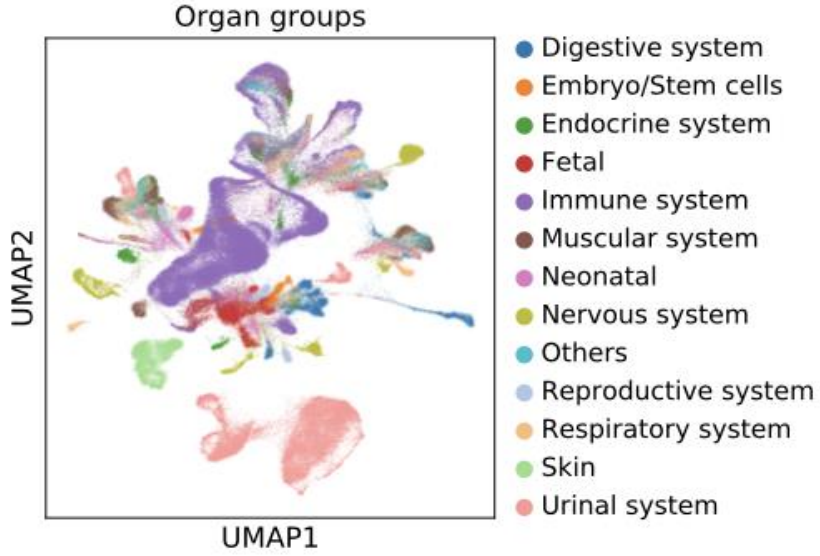
Each connection between cells is given the corresponding weight, w_{ca}, producing a weighted network representation of the data.

Example

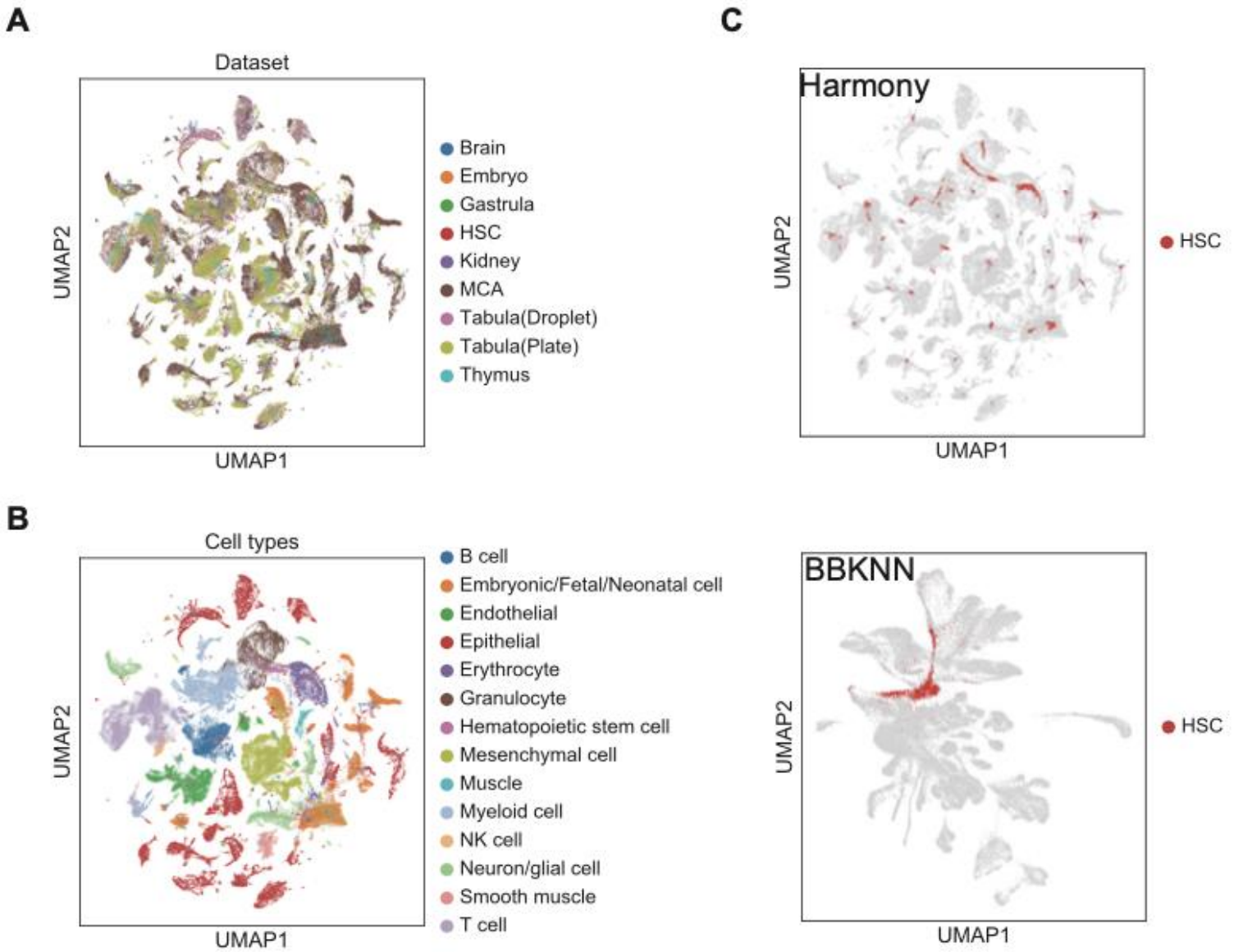
Mouse atlas: Analysing the complete 267,690 cell murine atlas collection. Merging all of the data sources leads to a clear divide based on the study of origin (A), which is successfully amended by BBKNN (B,C).







Better than harmony?



Harmony batch correction of the murine atlases. The datasets are well mixed (A), and the cell types are successfully reconnected (B) in most cases. However, the resulting manifold is considerably more fragmented than the one proposed by BBKNN, with the purified hematopoetic stem cell population from the HSC dataset split across the whole space instead of forming a centralised hub (C).

STACAS

- We are using STACAS
- “works” best
- How do we know that an integration worked?

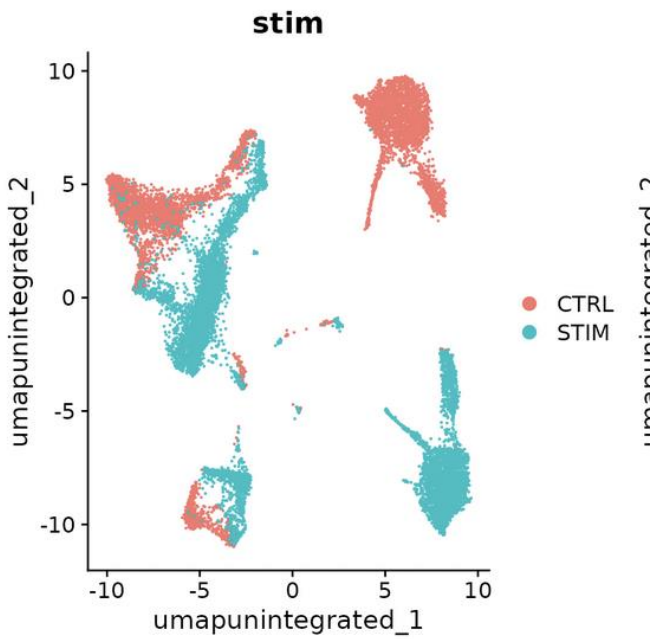
JOURNAL ARTICLE

STACAS: Sub-Type Anchor Correction for Alignment in Seurat to integrate single-cell RNA-seq data

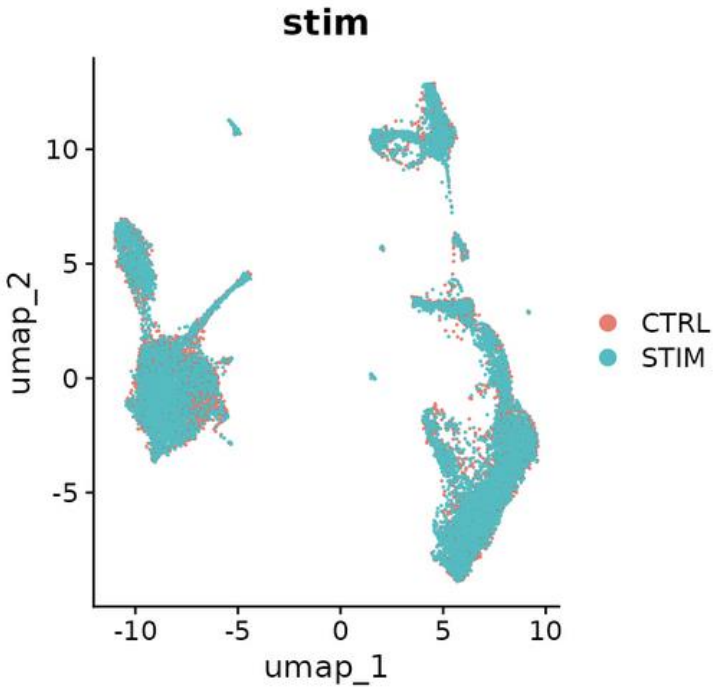
Massimo Andreatta, Santiago J Carmona

Bioinformatics, Volume 37, Issue 6, March 2021, Pages 882–884, https://doi.org/10.1093/bioinformatics/btaa755

Published: 26 August 2020 **Article history** ▾



Not integrated



Integrated

Which is the best tool?

- Independent comparison needed.

Spoiler alert, there is no winner!!!!

Bad for us, as we will need to test

But that will keep us our jobs ;-)

But what about multiple testing?

nature methods

Explore content ∨ About the journal ∨ Publish with us ∨

nature › nature methods › analyses › article

Analysis | Open access | Published: 23 December 2021

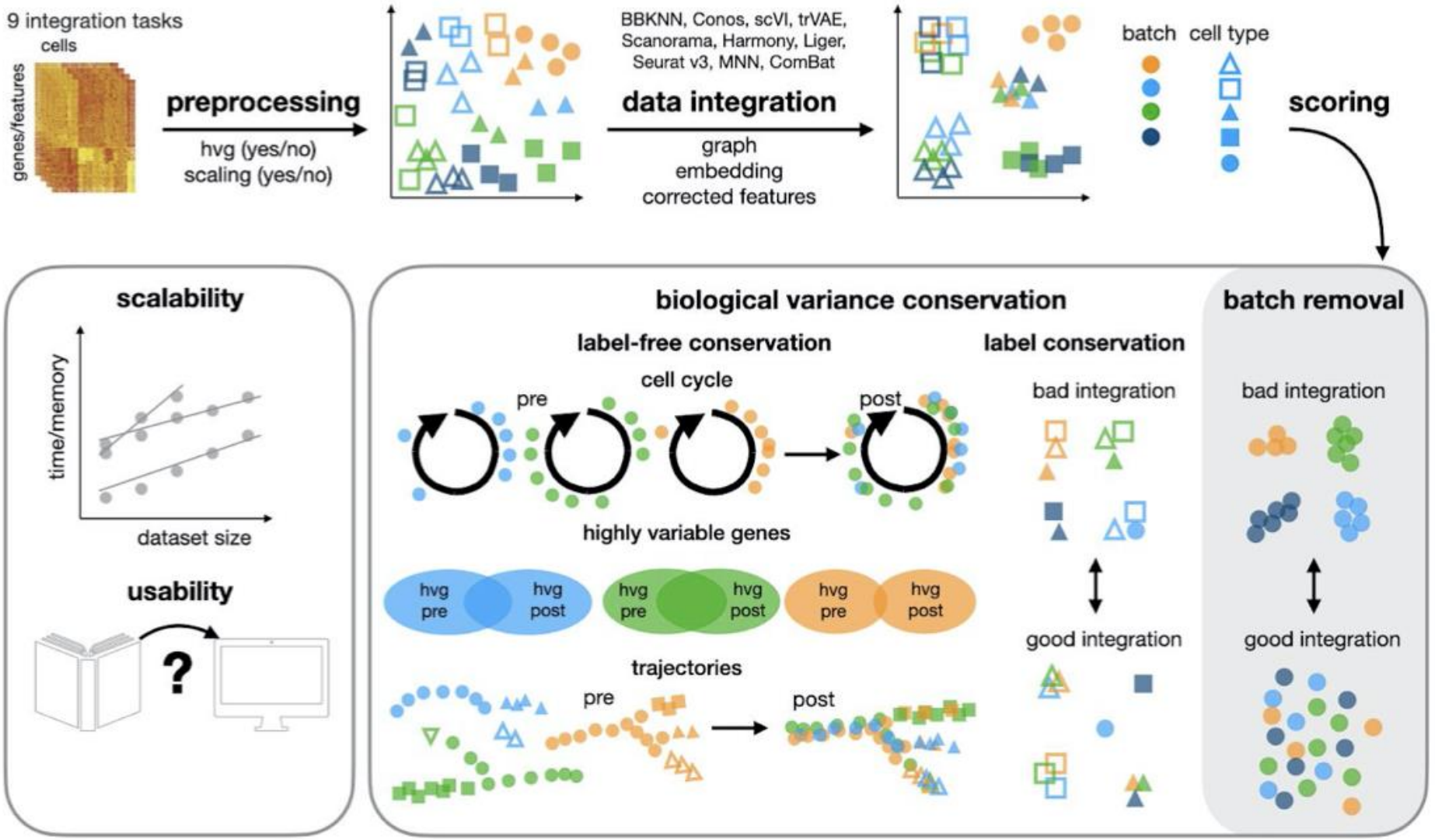
Benchmarking atlas-level data integration in single-cell genomics

Malte D. Luecken, M. Büttner, K. Chaichoompu, A. Danese, M. Interlandi, M. F. Mueller, D. C. Strobl, L. Zappia, M. Dugas, M. Colomé-Tatché & Fabian J. Theis

Nature Methods **19**, 41–50 (2022) | Cite this article

108k Accesses | **225** Citations | **358** Altmetric | Metrics

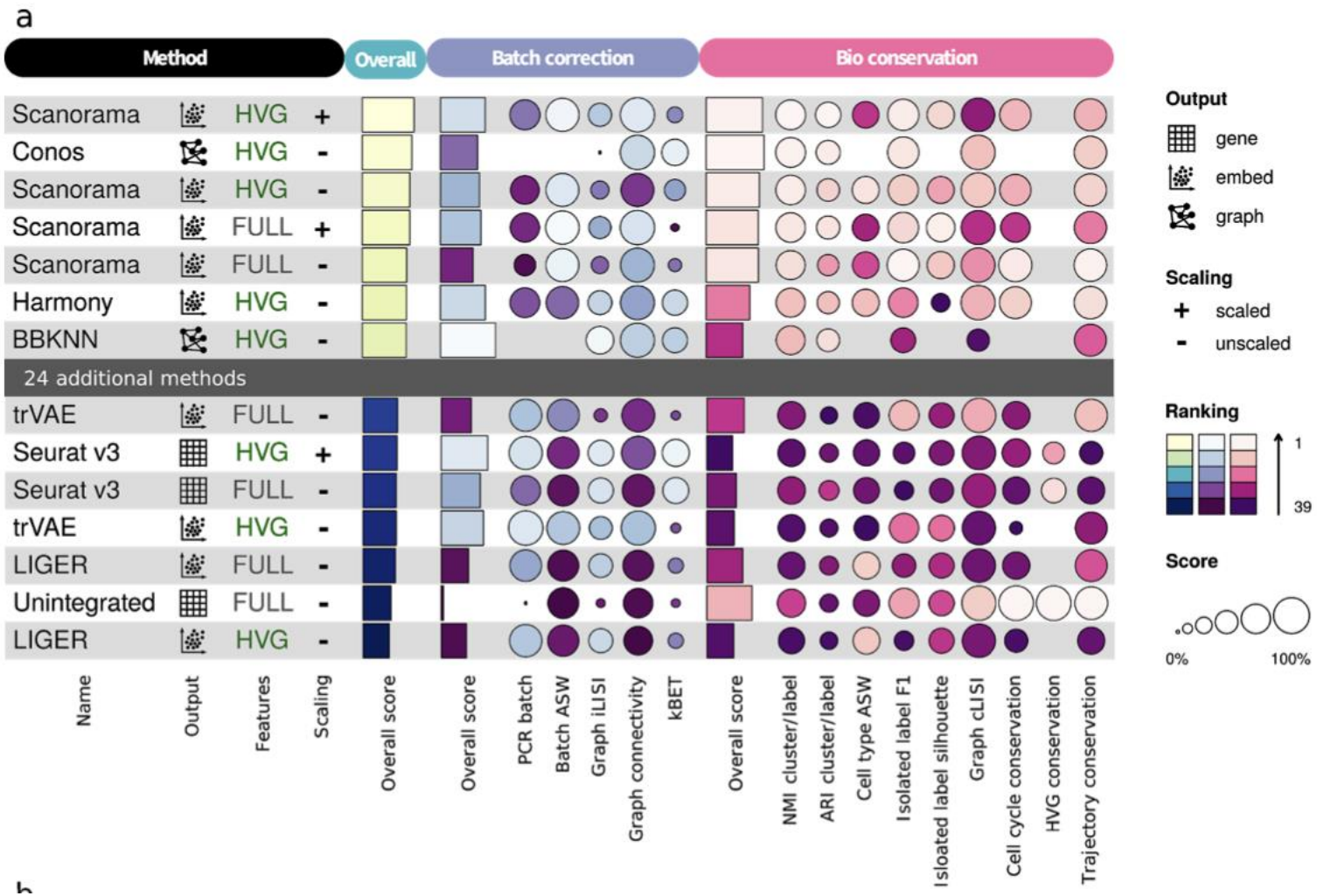
How to compare tools?



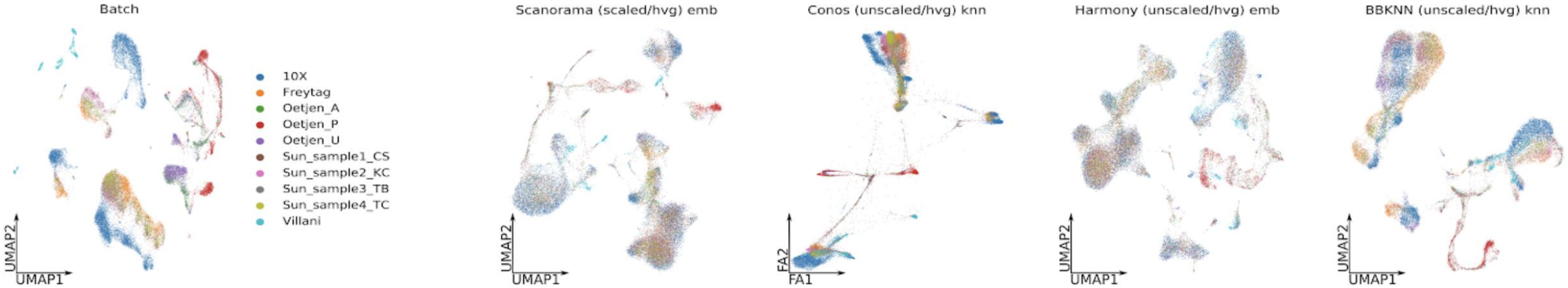
The methods

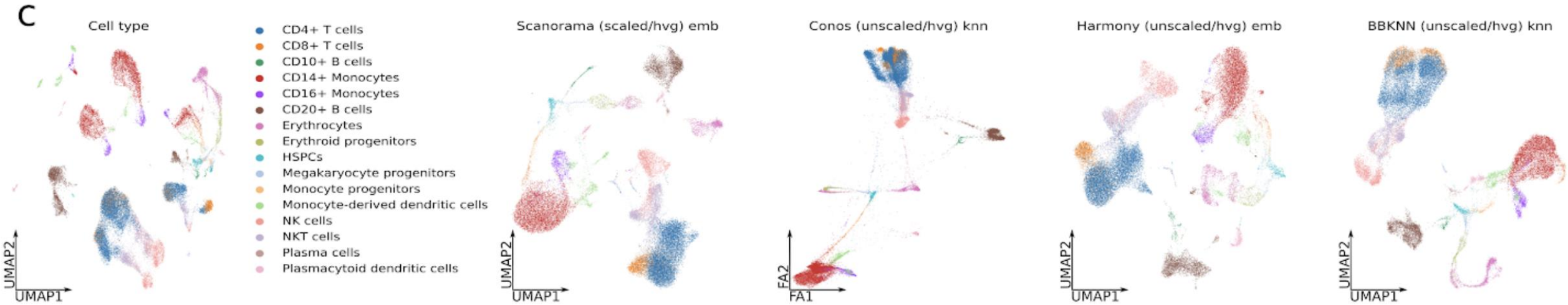
- benchmark 38 method and preprocessing combinations on 77 batches of gene expression, chromatin accessibility, and simulation data from 23 publications, altogether representing >1.2 million cells distributed in nine atlas-level integration tasks
- freely available reproducible python module can be used to identify optimal data integration methods for new data, benchmark new methods, and improve method development.
- BBKNN, Scanorama, and scVI perform well, particularly on complex integration tasks;
- Seurat v3 performs well on simpler tasks with distinct biological signals
- Where is harmony?
- Also do scATAQ-Seq

Result 1:human immune cell task



Result 1:human immune cell task

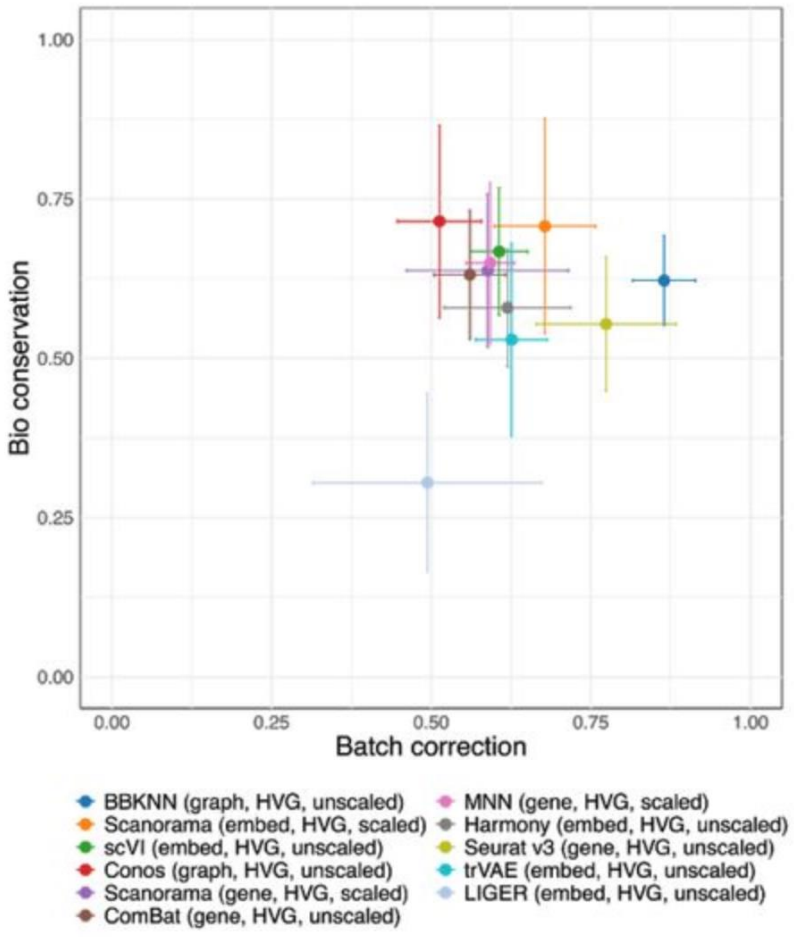


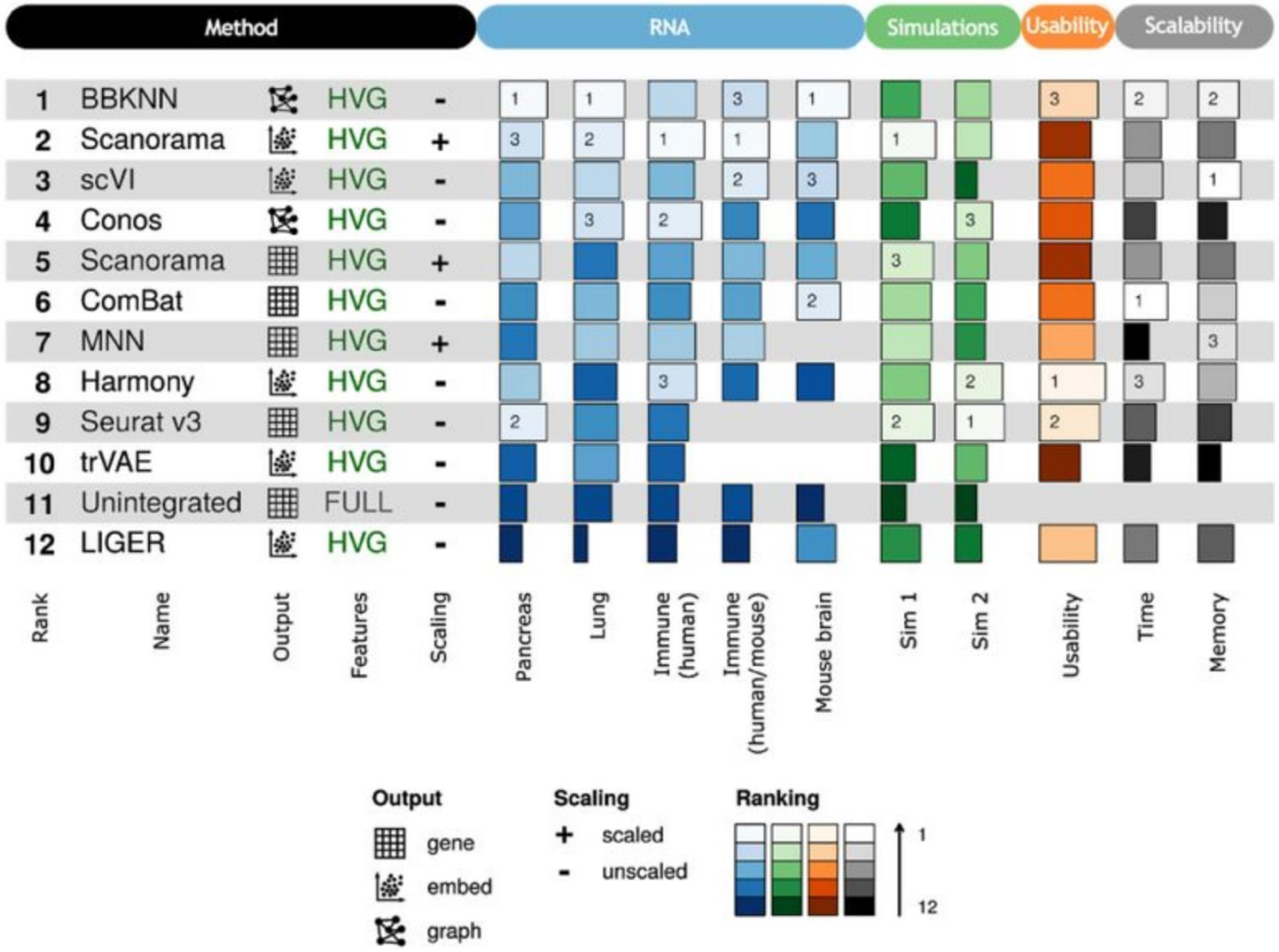


Critical thinking!

- Did they do a fair test?
- How is it better than the other papers (comparing their own tool).

Overview



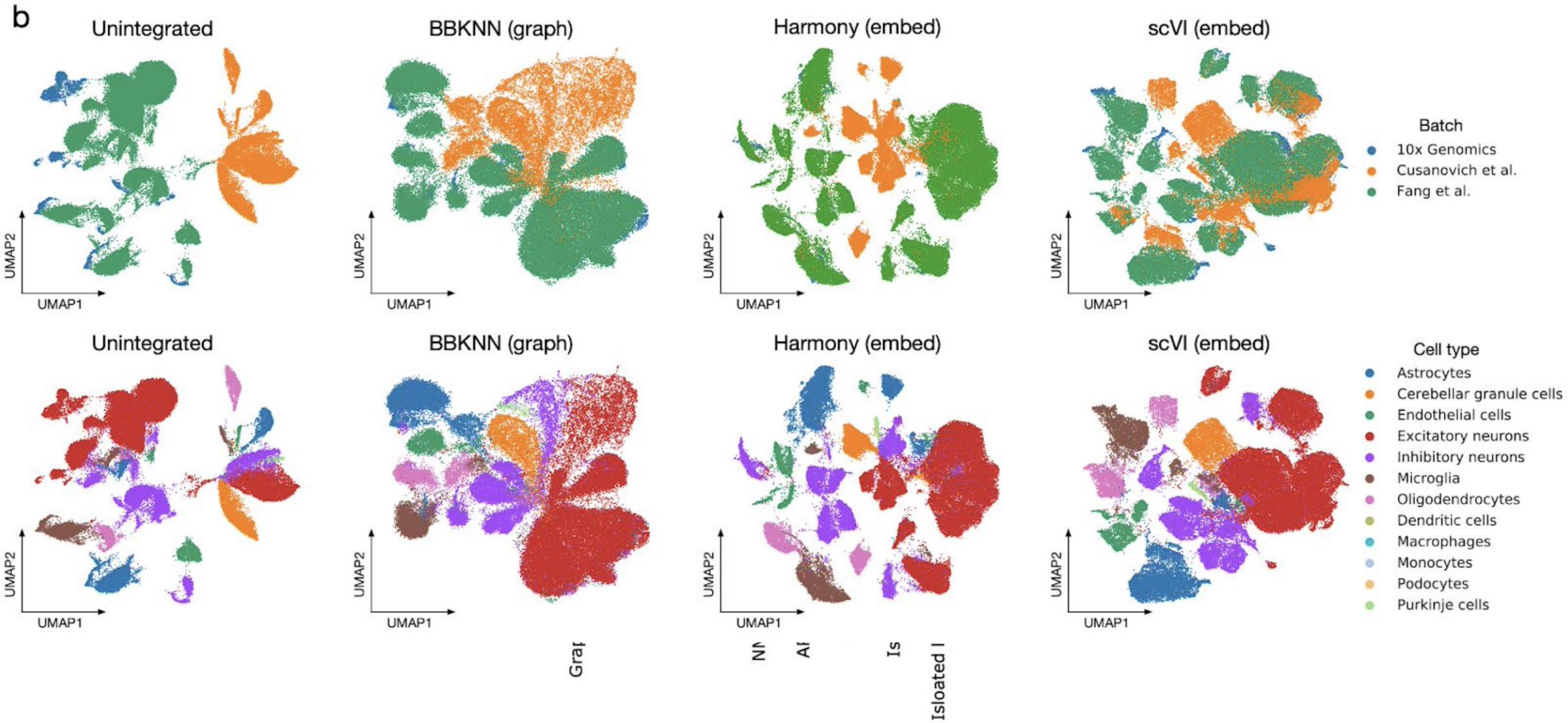


3 of 4

MVD - scMethods

Overview of benchmarking results on all RNA integration tasks and simulations, including usability and scalability results.

Result 3: large mouse brain ATAC task



Conclusion

- No really a winner!
- Basically, for complex datasets you need a pipeline that can integrate your datasets of choice and then you decide the best one.
  - BBKNN, Scanorama, and scVI perform well, particularly on complex integration tasks;
  - Seurat v3 performs well on simpler tasks with distinct biological signals

- For our work larger tasks, we mostly use Harmony
- Pairwise, ok with Seurat -> part of the exercise

And more reviews to reads…

nature reviews genetics

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › nature reviews genetics › expert recommendation › article

Expert Recommendation | Published: 31 March 2023

Best practices for single-cell analysis across modalities

Lukas Heumos, Anna C. Schaar, Christopher Lance, Anastasia Litinetskaya, Felix Drost, Luke Zappia, Malte D. Lücken, Daniel C. Strobl, Juan Henao, Fabiola Curion, Single-cell Best Practices Consortium, Herbert B. Schiller & Fabian J. Theis ✉

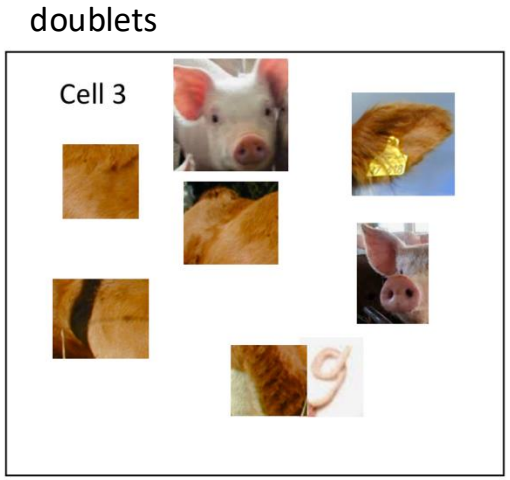
Nature Reviews Genetics **24**, 550–572 (2023) | Cite this article

Conclusion Integration

- Sorry, there is not one solution!
- Important, just the embedding get transformed, not the raw reads
- Don’t trust the comparison of the tool developers but of others

4. Dimension reduction

- Our data are high-dimensional
- A lot of drop outs
- Noisy data…



Our data are high-dimensional!

- Every gene is one component, with 65000 possible categories
- The expression of each cell (in bulk sample) is a combination of the 65000 genes
- We speak of an N-dimensional space, as we cannot image more than 3 dimension
- Most values might be close to zero, and some **values are more relevant**
- Some low dimensional pattern are batch effects, need to get rid
- So we need to reduce the dimensionality….

Visualisation

Get rid of things we don't need

Many methods to reduce dimensionality

1. PCA (linear)
2. t-SNE (non-parametric/ nonlinear)
3. Sammon mapping (nonlinear)
4. Isomap (nonlinear)
5. LLE (nonlinear)
6. CCA (nonlinear)
7. SNE (nonlinear)
8. MVU (nonlinear)
9. Laplacian Eigenmaps (nonlinear)
10. UMAP (nonlinear)
11. Diffusion Maps (nonlinear)
12. Phate (nonlinear)

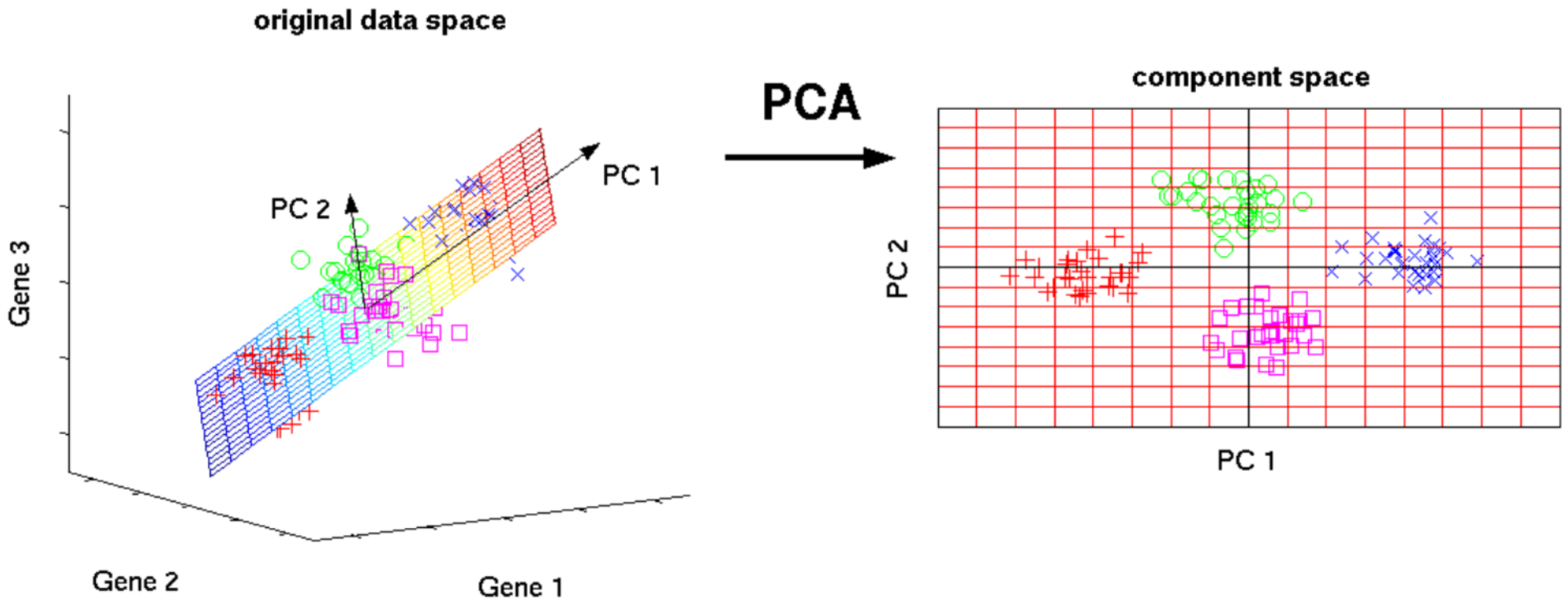
- The good news is that you need to use only two of the algorithms mentioned above to effectively visualize data in lower dimensions – PCA and UMAP.

PCA

- Principal Component Analysis (PCA) is a dimension-reduction tool that can be used to reduce a large set of variables to a small set that still contains most of the information in the large set.
- Principal components are new variables that are constructed as linear combinations of the initial variables.
- Transform correlated variables into uncorrelated variables

- PCA tries to detect the highest variability

- For single cell use it to find real signal versus confounders. Generally 15 - 30



Source: nlpca

PCA – at different levels

- An expression of a cell is described by the expression of its genes
- Taken n genes ($g_1..g_n$), part of the same metabolic pathways (eg fatty acid biosynthesis). One is upregulated, then the others as well, as when the process starts, all genes are expressed higher - as they are correlated – like weight and height is generally correlated
- Simplest explanation: PCA reduces features, as one variable might be enough to describe that our pathway is up
- PCA build new features (principal components) that are linear combination of old features ($g_1+4*g_2-0.5*g_3$ etc)
- This new feature should contain most of the variance – so describes best if the genes of the pathway are changing **or** if you would reconstruct the expression of the genes ($g_1..g_n$) the error is minimal

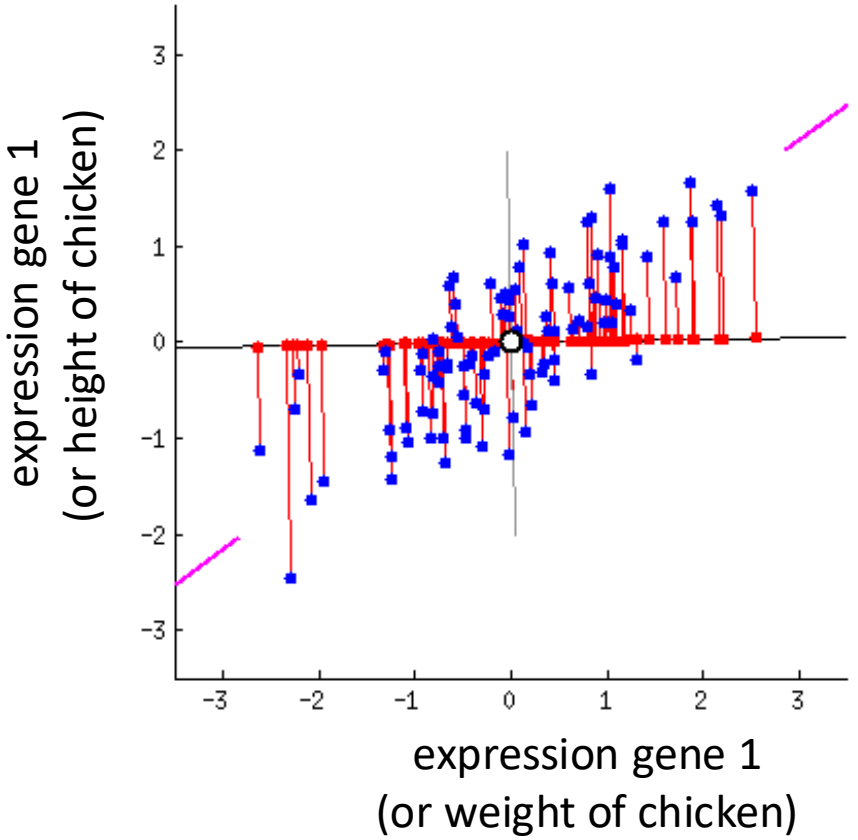
https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues

PCA – at different levels

- new property is a line w_1x+w_2y
- spread on project (red dot on black line) highest – average squared distance from the center of the gene cloud to each red dot; *variance*

$$\sigma_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

- **or** the total reconstruction error is measured as the average squared length of the corresponding red lines
- so: the higher the variance caputed by PC1 the lower the error



https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues

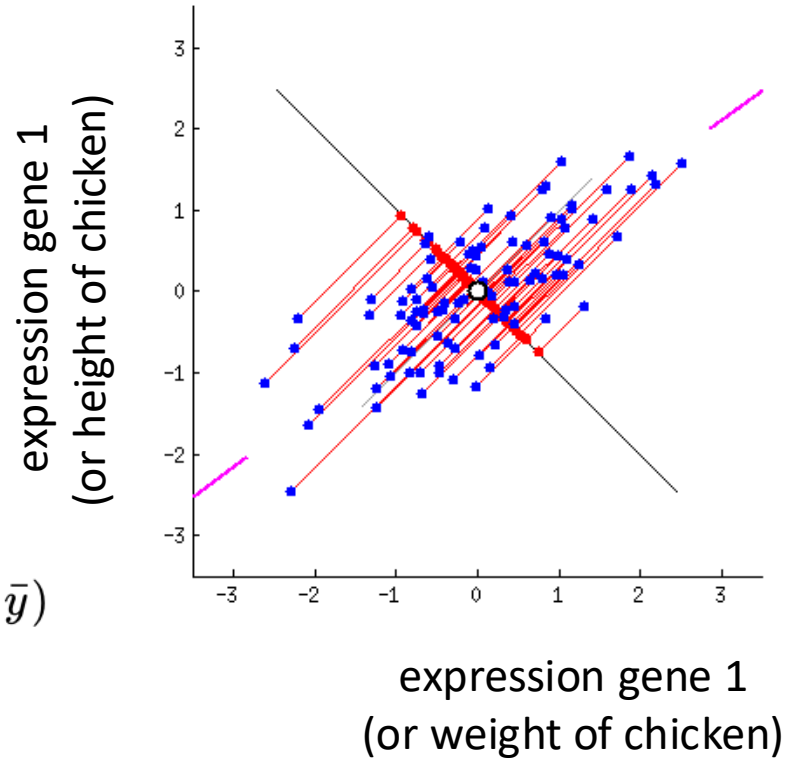
PCA – eigenvector eigenvalues

- Variance measures the variation of a single random variable
- covariance is a measure of how much two random variables vary together
- *covariance matrix:*

$$C = \begin{pmatrix} \sigma(x,x) & \sigma(x,y) \\ \sigma(y,x) & \sigma(y,y) \end{pmatrix}$$

co-variance

$$\sigma(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$



Co-variance of these data

$$\begin{pmatrix} 1.07 & 0.63 \\ 0.63 & 0.64 \end{pmatrix}$$

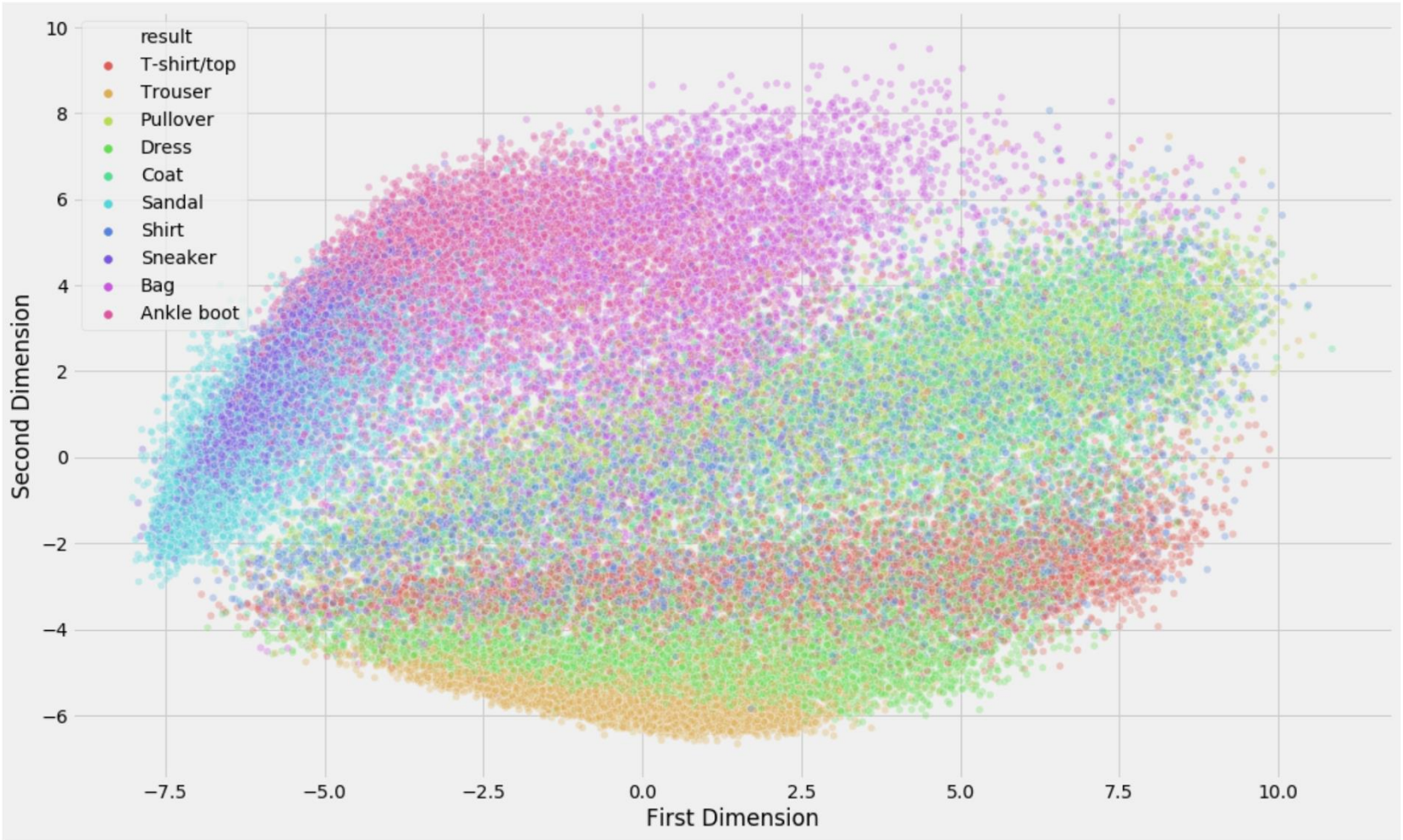
Spectral theorem (Eigenvectors)

$$\begin{pmatrix} 1.52 & 0 \\ 0 & 0.19 \end{pmatrix}$$

Eigenvalues

https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues

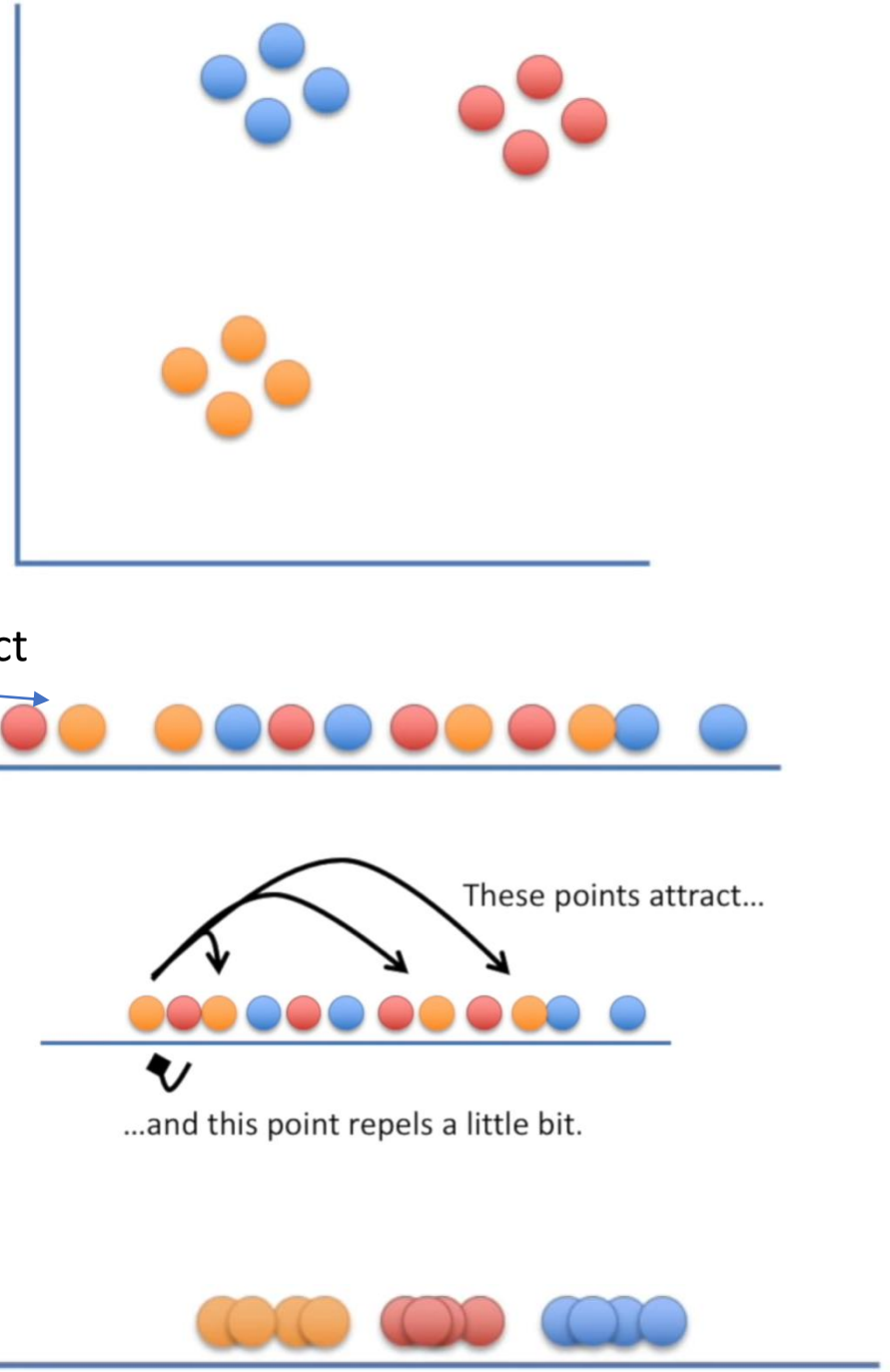
MNIST fashion



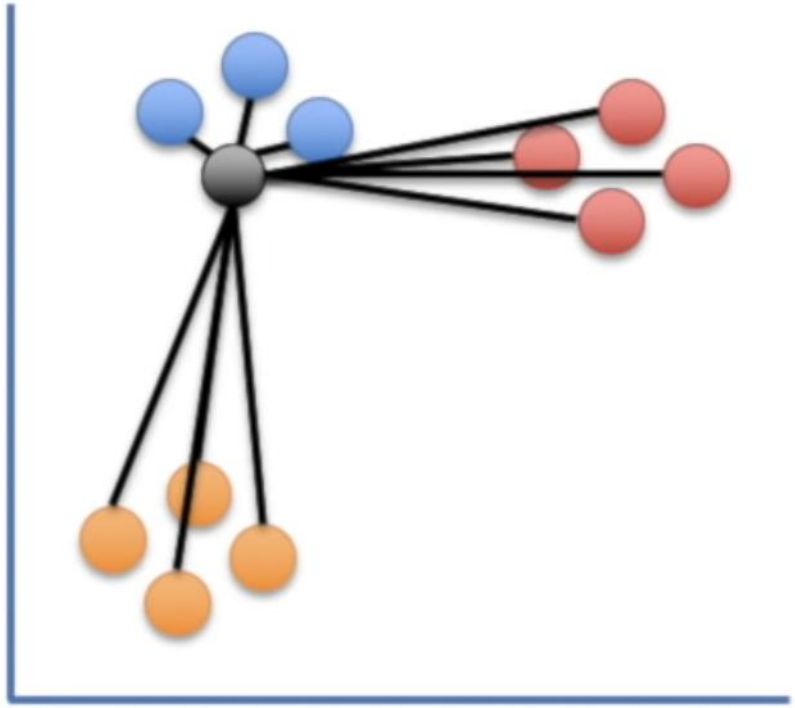
https://dax-cdn.cdn.appdomain.cloud/dax-fashion-mnist/1.0.2/data-preview/Part%201%20-%20Data%20Exploration.html

t-Distributed Stochastic Neighbor Embedding (t-SNE)

- non-linear technique for dimensionality reduction
- extensively applied in image processing, NLP, genomic data and speech processing
- Clustering data that preserves distances varying scales
- Ignores intermediate and long distances



https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf



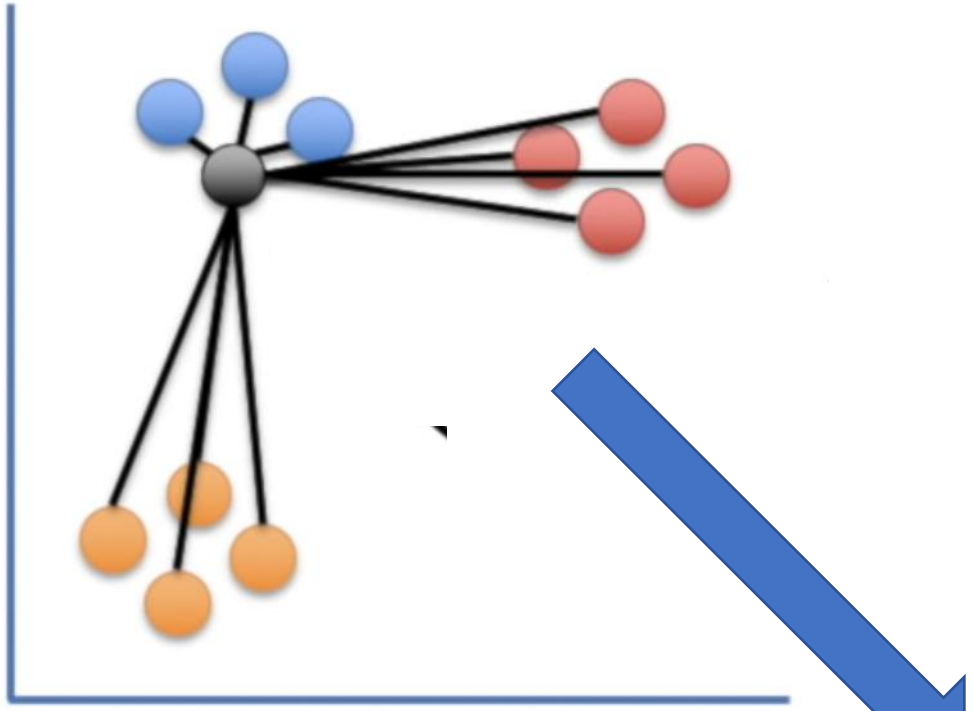
Standard deviation σ_i (**perplexity parameter)**

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)}$$

The probability of point x_i to have x_j as it's neighbor

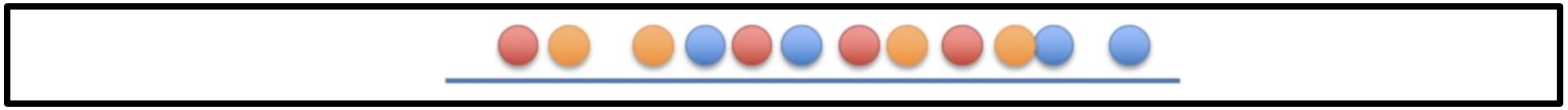
Average distance between all points

Calculating a joint probability distribution that represents the similarities between the data points



$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)}$$

The probability of point x_i to have x_j as it's neighbor

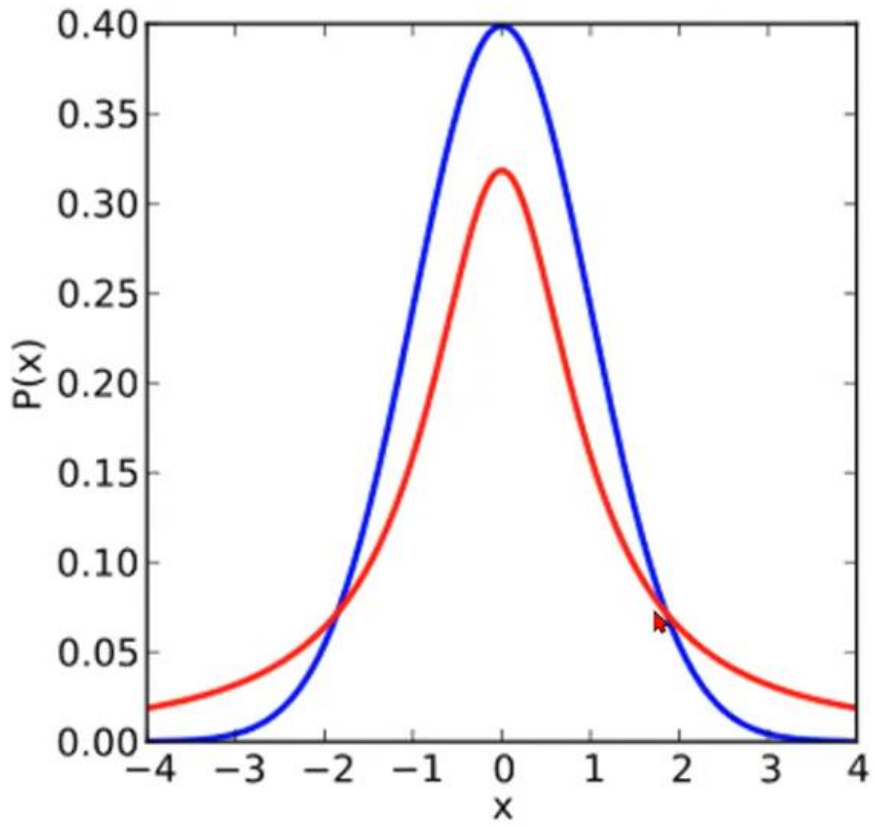


Creating a dataset of points in the target dimension and then calculating the joint probability distribution for them as well.

$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \|y_k - y_l\|^2\right)^{-1}}$$

(t-distributed student test)

Low dimensional embedding using a Student t-distribution to avoid overcrowding



Red – Student t-distribution (1 degree of freedom)

Blue - Gaussian

Also explains the t- in the name of t-SNE

Need to compare

Kullback-Leibler divergence between P and Q.

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

P and Q defined on the same probability space, $\mathcal{X}$, the relative entropy from Q to P

Use the Kullback-Leibler equation as a cost function that we want to optimise.

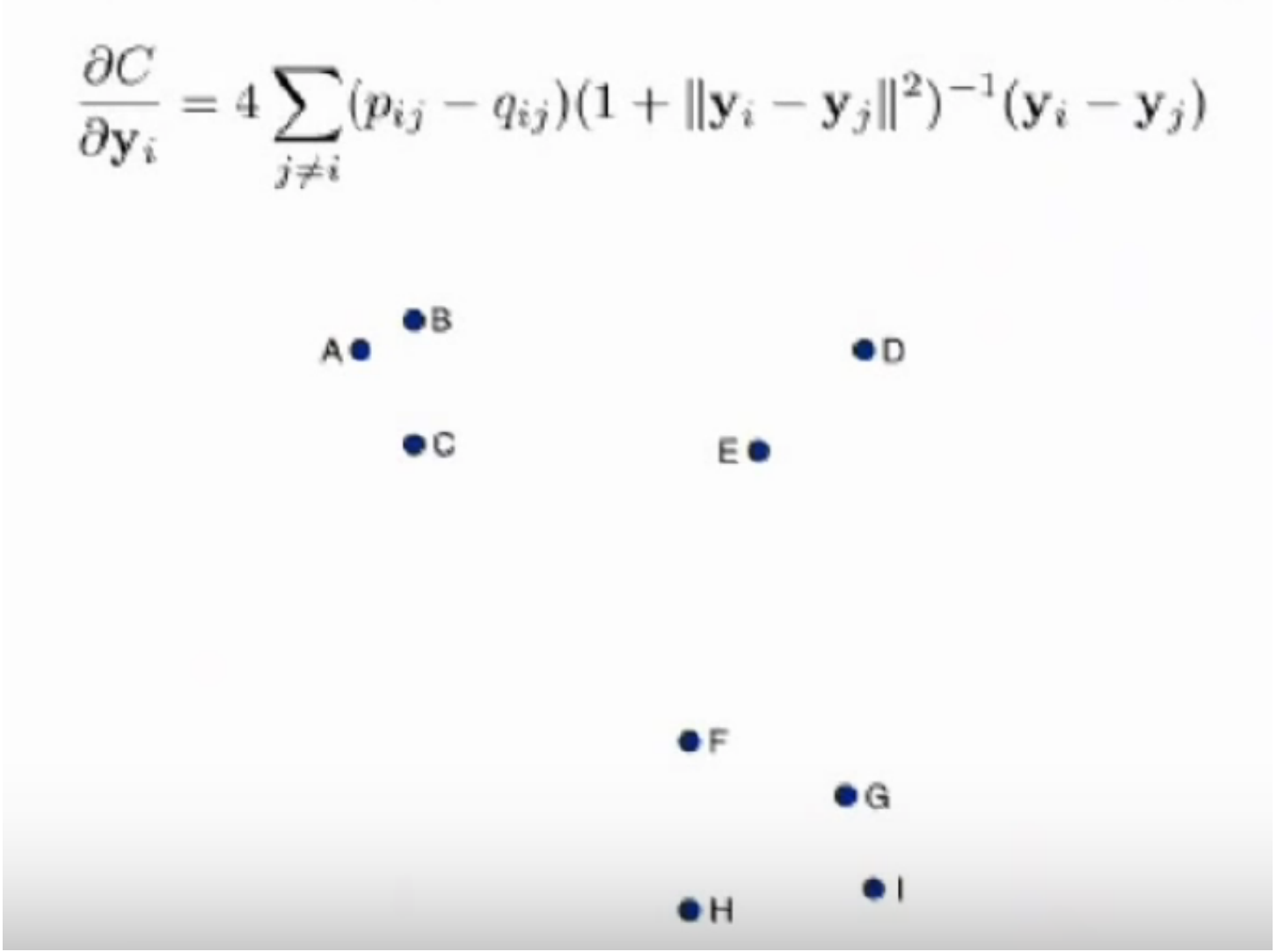
$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$$

- if *p* = *q* then log(1) = 0

- Penalize when *p* != *q*
  - Large ***p*** modeled by small ***q***: Big penalty
  - Small ***p*** modeled by large ***q***: Small penalty

Gradient decent optimization

$$\frac{\partial C}{\partial \mathbf{y}_i} = 4 \sum_{j \neq i} (p_{ij} - q_{ij})(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1} (\mathbf{y}_i - \mathbf{y}_j)$$

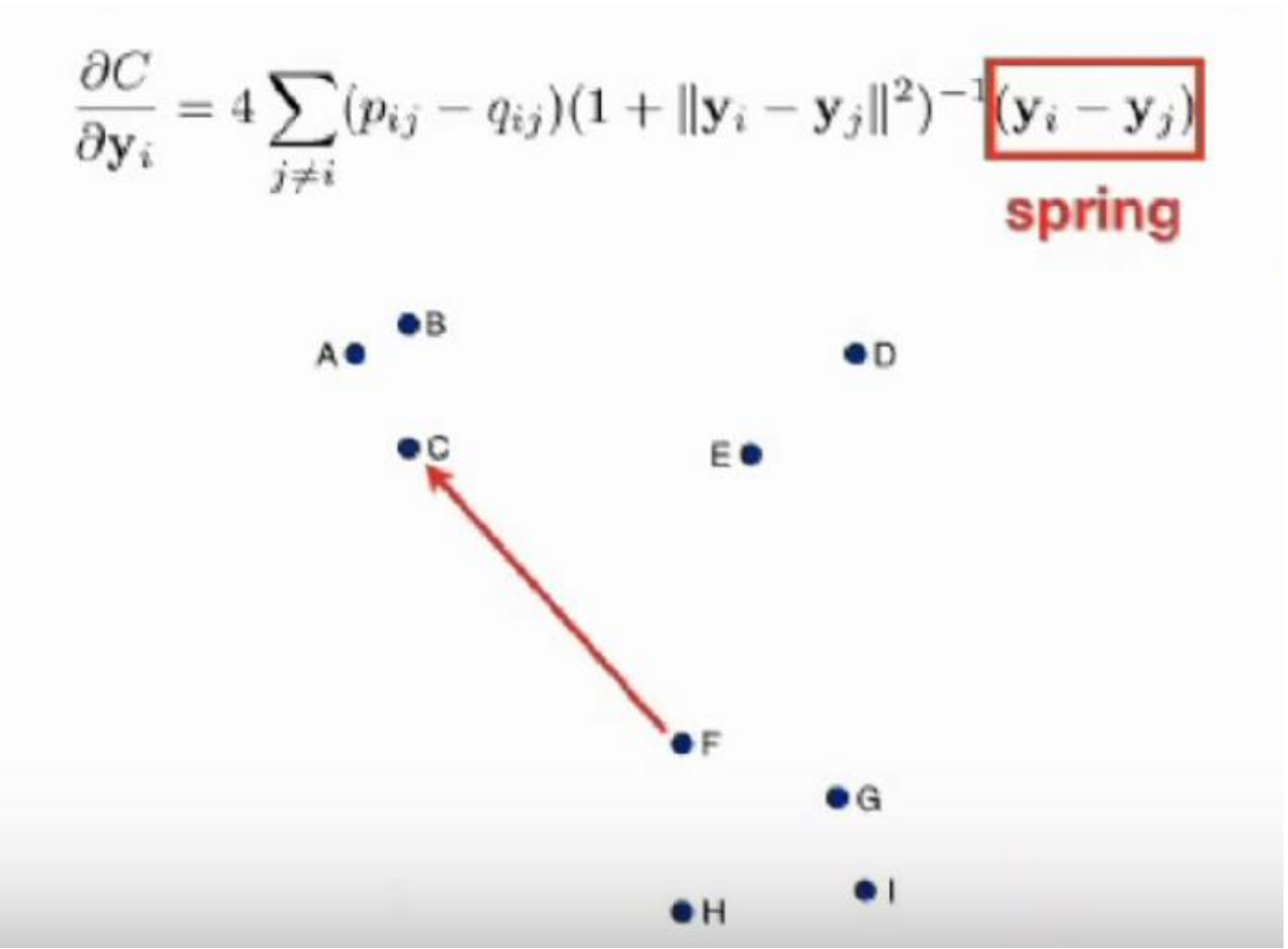


https://www.youtube.com/watch?v=RJVL80Gg3lA

Gradient decent optimization

$$\frac{\partial C}{\partial \mathbf{y}_i} = 4 \sum_{j \neq i} (p_{ij} - q_{ij})(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1} (\mathbf{y}_i - \mathbf{y}_j)$$

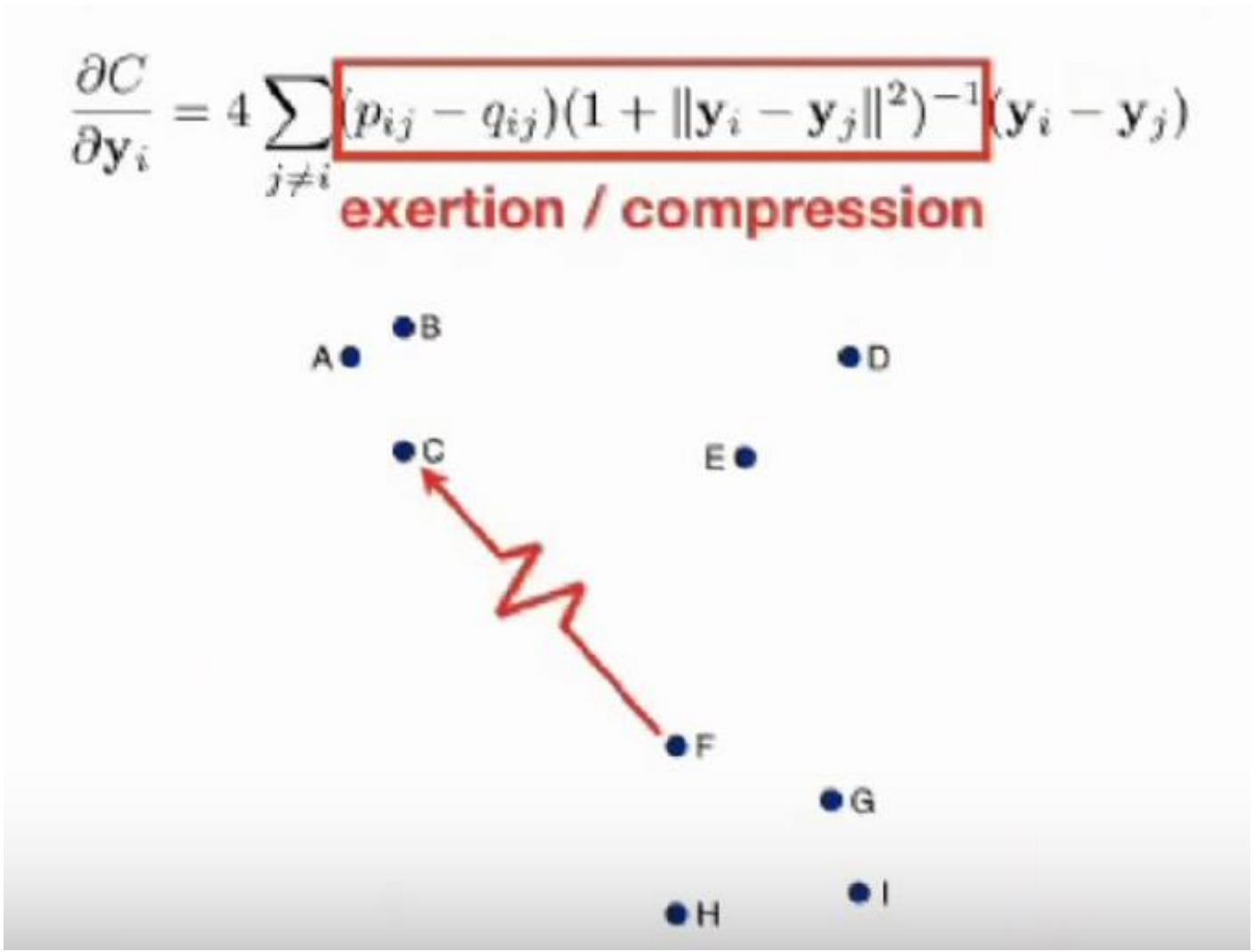
spring



Gradient decent optimization

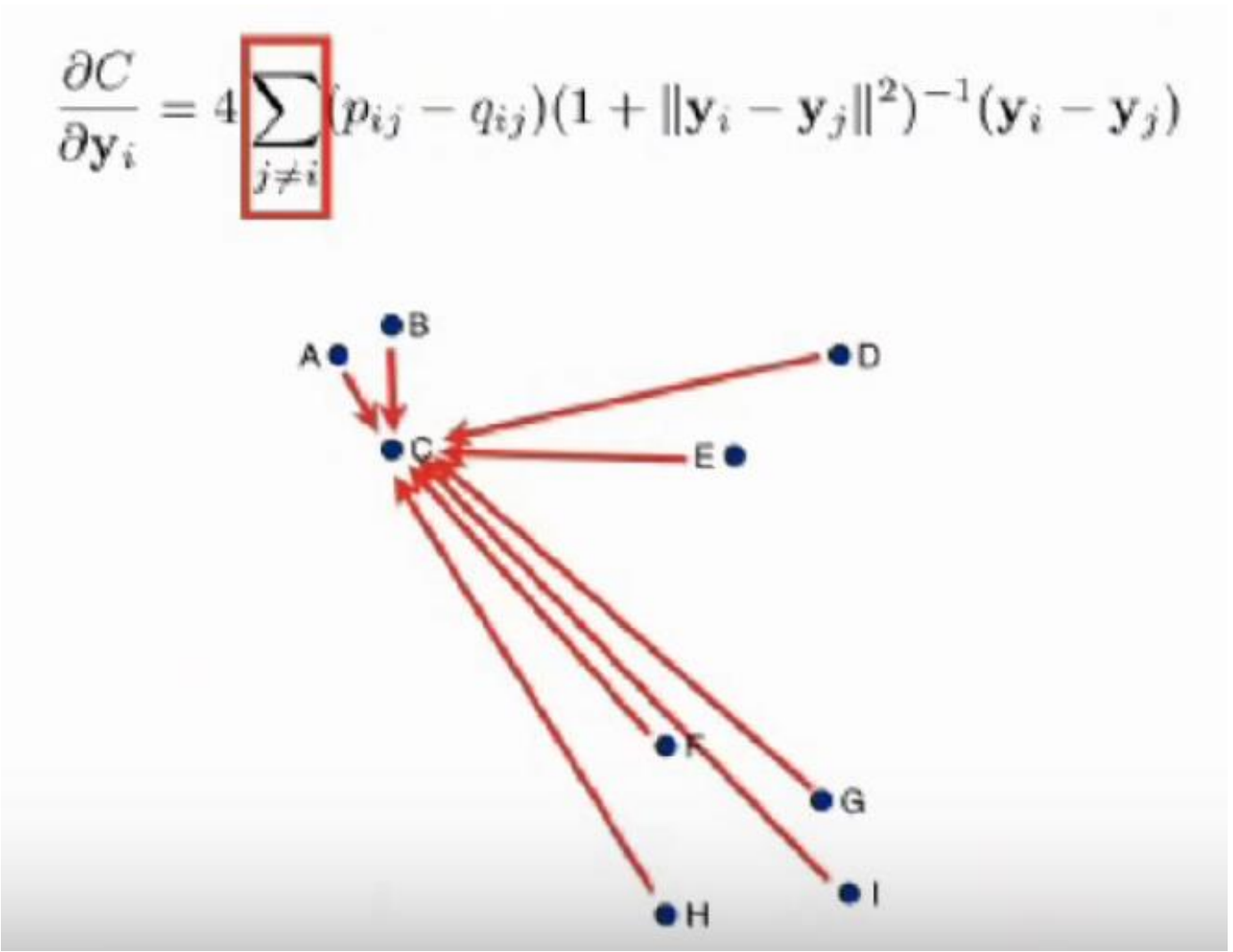
$$\frac{\partial C}{\partial \mathbf{y}_i} = 4 \sum_{j \neq i} (p_{ij} - q_{ij})(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1} (\mathbf{y}_i - \mathbf{y}_j)$$

exertion / compression

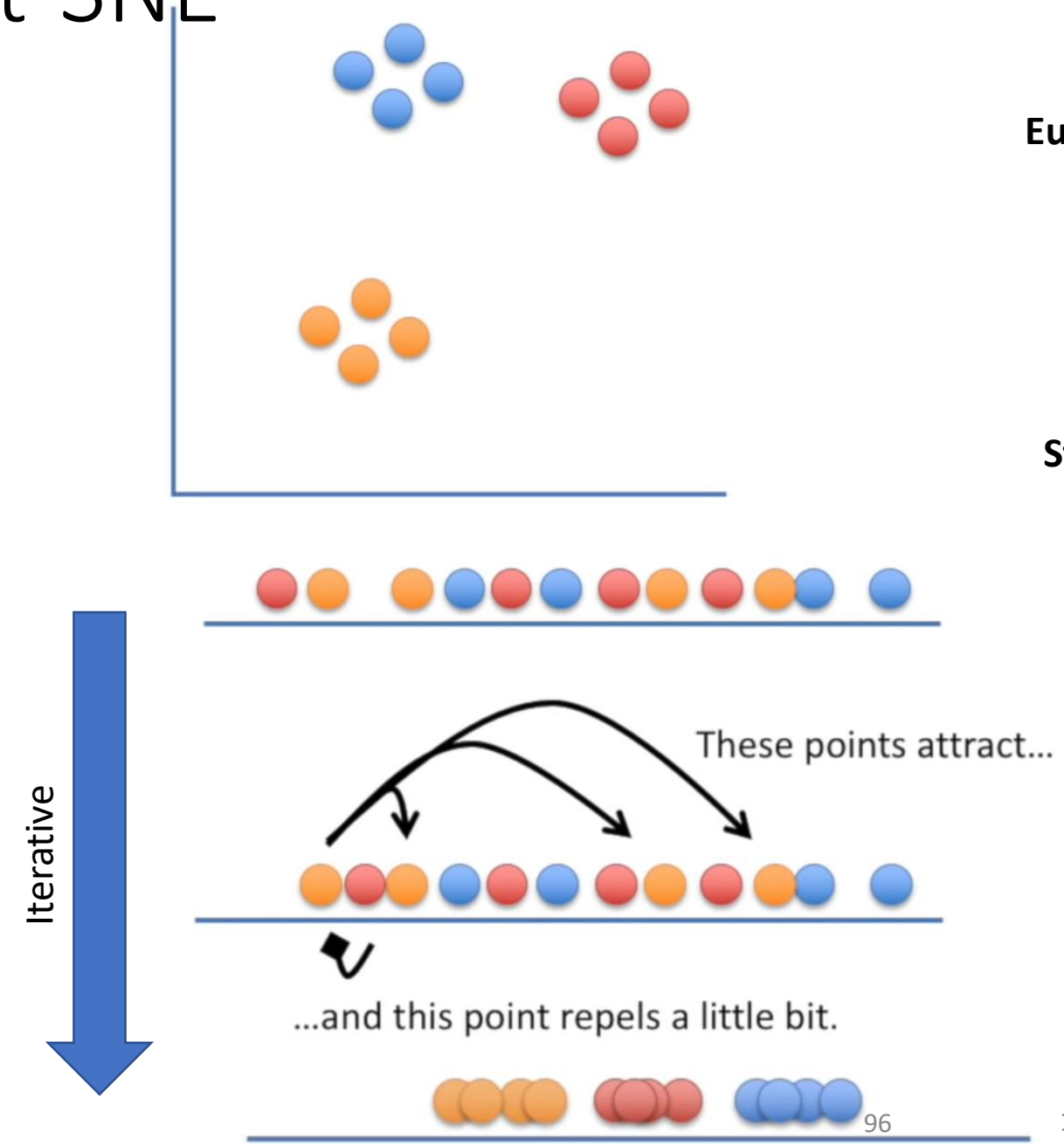


Gradient decent optimization

$$\frac{\partial C}{\partial \mathbf{y}_i} = 4 \sum_{j \neq i} (p_{ij} - q_{ij})(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}(\mathbf{y}_i - \mathbf{y}_j)$$



t-SNE



Euclidean distances

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)}$$

The probability of point x_i to have x_j as it's neighbor

Student t-distribution

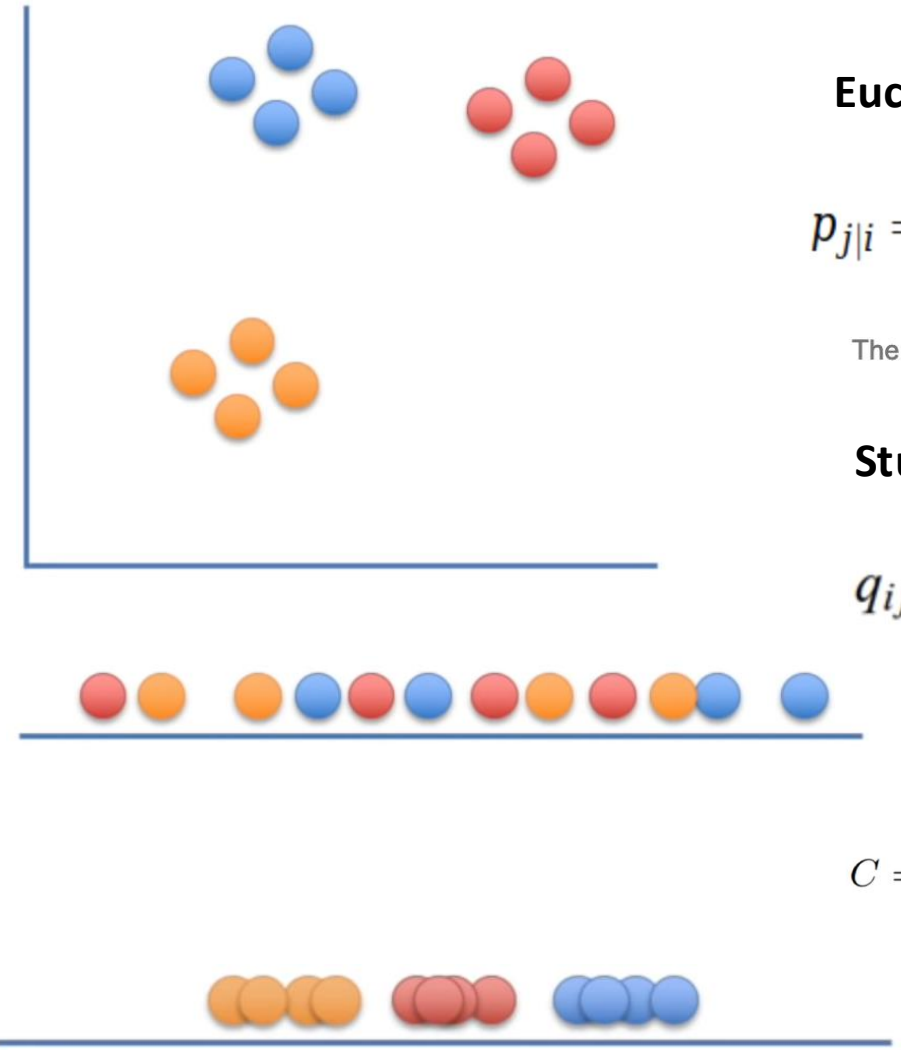
$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \|y_k - y_l\|^2\right)^{-1}}$$

KL-divergence

$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$$

t-SNE

Another Approach:



Euclidean distances

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)}$$

The probability of point x_i to have x_j as it's neighbor

exponential probability distribution (not normalisations)

$$p_{i|j} = e^{-\frac{d(x_i, x_j) - \rho_i}{\sigma_i}}$$

Student t-distribution

$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \|y_k - y_l\|^2\right)^{-1}}$$

$$q_{ij} = \left(1 + a(y_i - y_j)^{2b}\right)^{-1}$$

KL-divergence

$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$$

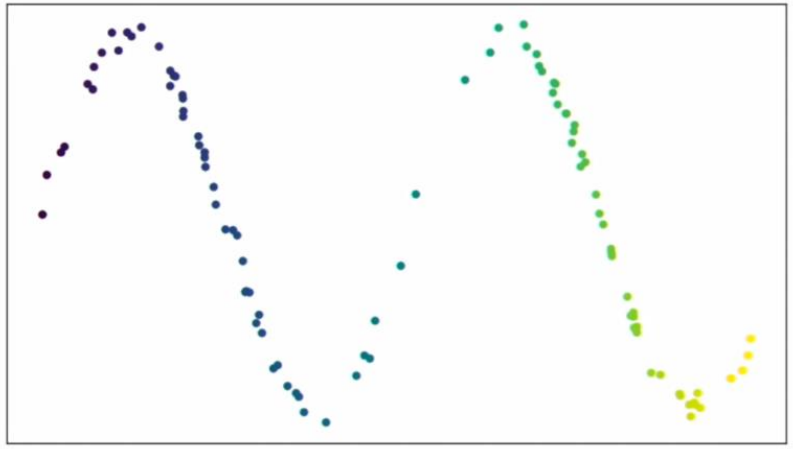
binary cross-entropy

$$CE(X,Y) = \sum_i \sum_j \left[p_{ij}(X) \log\left(\frac{p_{ij}(X)}{q_{ij}(Y)}\right) + (1 - p_{ij}(X)) \log\left(\frac{1 - p_{ij}(X)}{1 - q_{ij}(Y)}\right) \right]$$

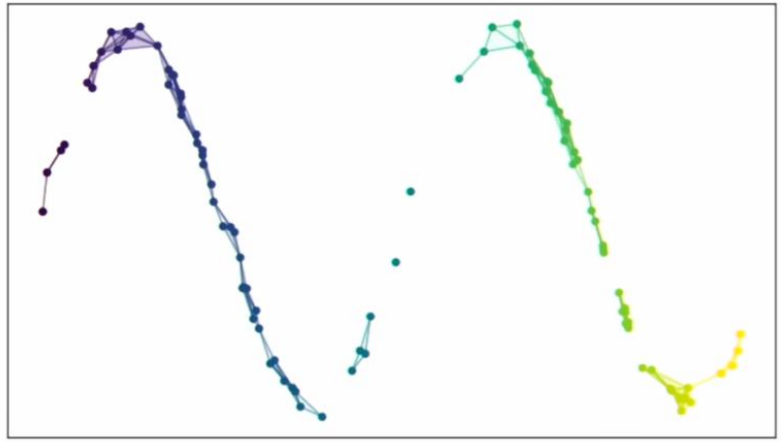
This method is called UMAP

UMAP: Uniform Manifold Approximation and Projection

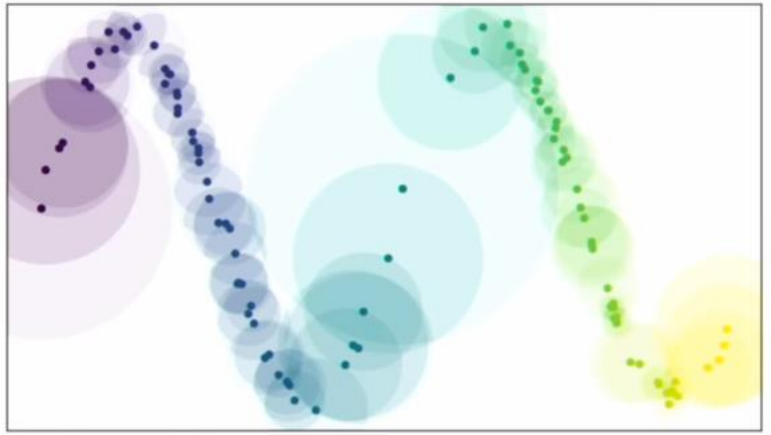
- Use topology to estimate the binary cross entropy
- Use algebraic topology
- UMAP essentially constructs a weighted graph from the high dimensional data – simplex
- edge strength representing how “close” a given point is to another
- projects this graph down to a lower dimensionality (force-directed graph layout algorithm)



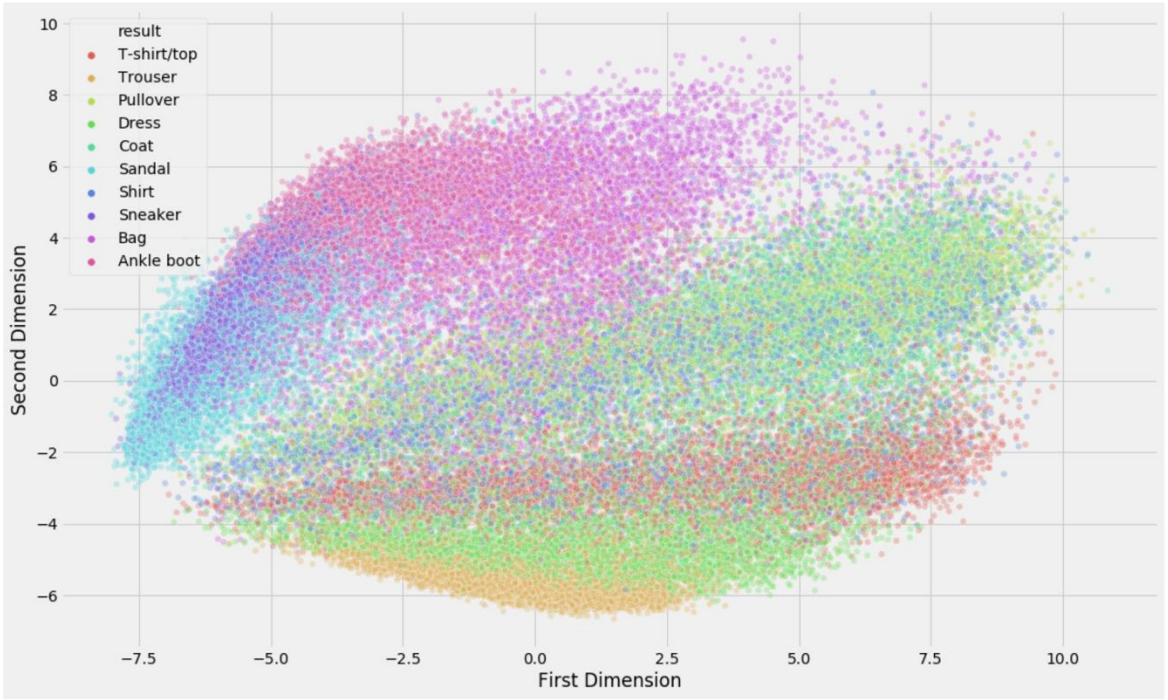
Simplex



fuzzy simplicial complex

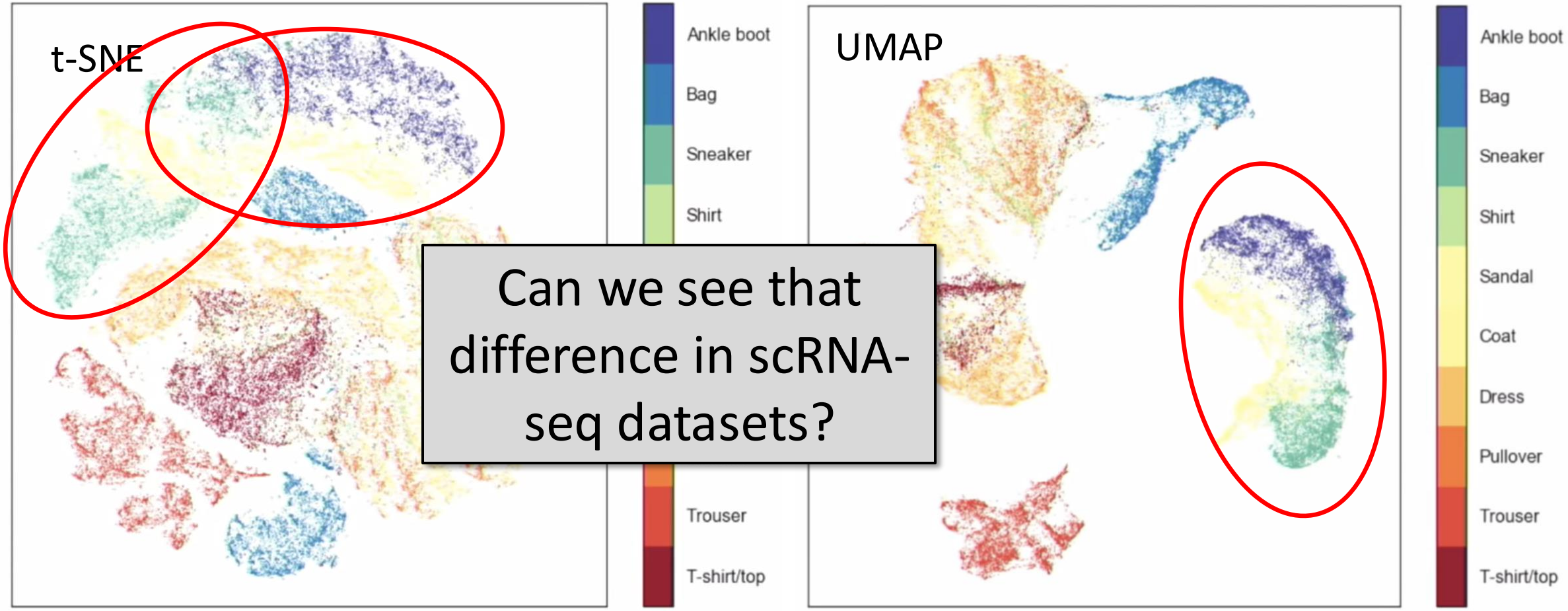


Standard example: Fashion MNIST (784-dimensional)



https://github.com/zalandoresearch/fashion-mnist

Standard example: Fashion MNIST (784-dimensional)



Can we see that difference in scRNA-seq datasets?

+ different classes pick out nicely

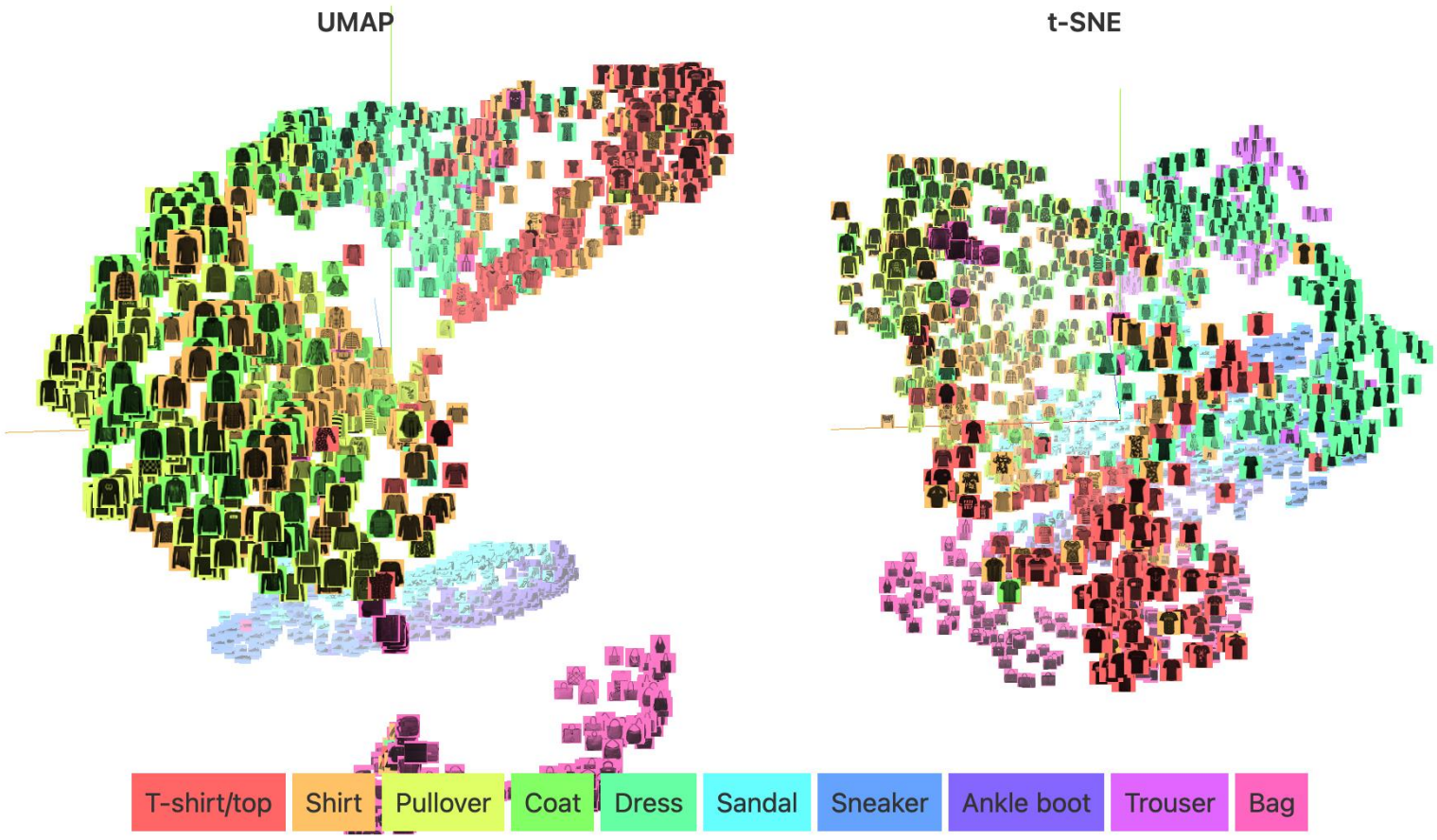
15 minutes on a desktop

+ More understanding between the classes-
→ Global conservation of distance

78 seconds~ 11x faster!

want to play a little bit?

https://pair-code.github.io/understanding-umap/



Conclusions?

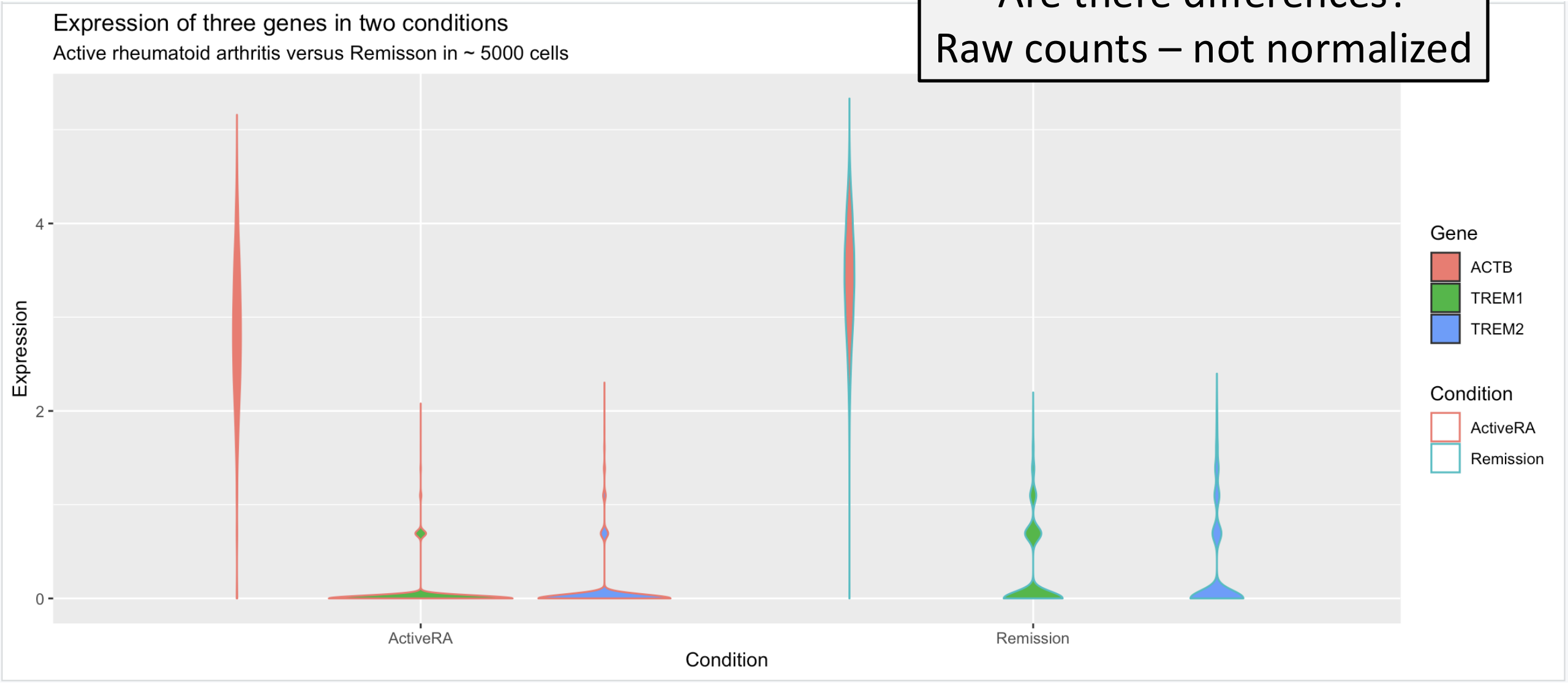
- What do you think – T-SNE or UMAP?

Last but not least - normalise

What is our problem?

- cell-specific biases

Are there differences?
Raw counts – not normalized



Normalization

- default Seurat – log scale - zero mean, variance one
- scTransform
- Scran

But why do we need to normalize data?
What happens, if we don't?

Default Seurat method

- Employ a global-scaling normalization method “LogNormalize” that normalizes the feature expression measurements for each cell by the total expression, multiplies this by a scale factor (10,000 by default), and log-transforms the result.
- Normalized values are stored in pbmc[["RNA"]]@data.

Scaling the data

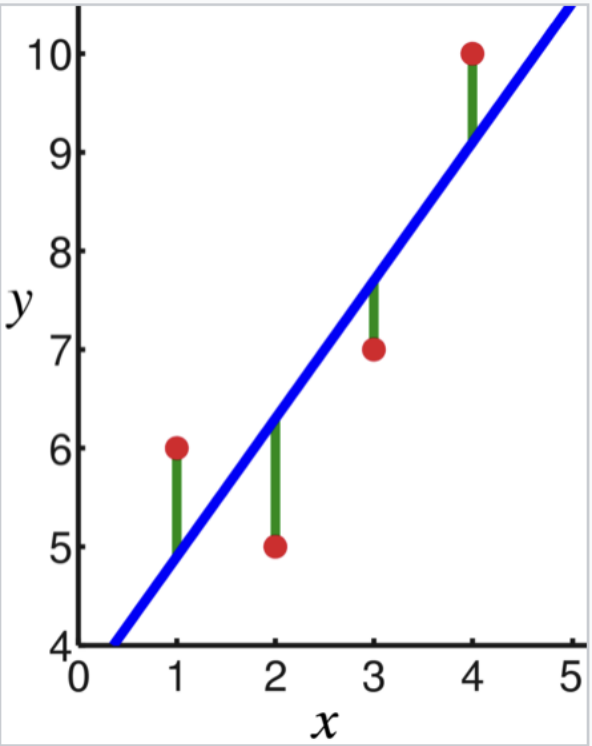
- Shifts the expression of each gene, so that the mean expression across cells is 0
- Scales the expression of each gene, so that the variance across cells is 1
  - This step gives equal weight in downstream analyses, so that highly-expressed genes do not dominate
- The results of this are stored in pbmc[["RNA"]]@scale.data

Linear regression – heart about it?

- Linear regression is a statistical technique used to model the relationship between a dependent variable (response variable) and one or more independent variables (predictor variables)

- Goal: Find the best-fit line that describes the linear relationship between the dependent variable and independent variables.

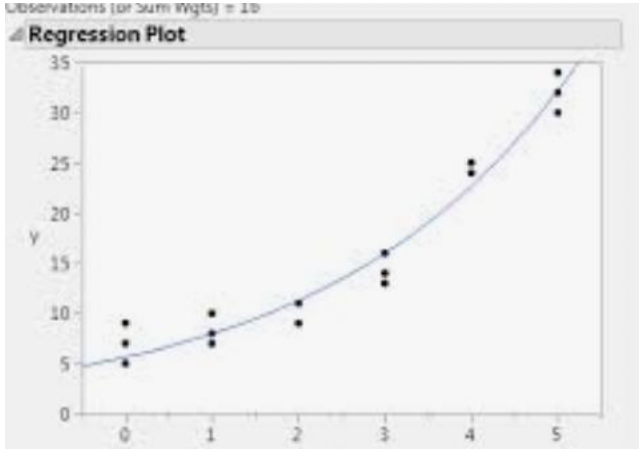
- y = b0 + b1*x1 + b2*x2 + ... + bn*xn



In linear regression, the observations (**red**) are assumed to be the result of random deviations (**green**) from an underlying relationship (**blue**) between a dependent variable (*y*) and an independent variable (*x*).

Generalized linear model (GLM)

- flexible generalization of ordinary linear regression
- GLM generalizes linear regression by allowing the linear model to be related to the response variable via a *link function* and by allowing the magnitude of the variance of each measurement to be a function of its predicted value
- Basically it can adapt through a function



- in R: ?glm()

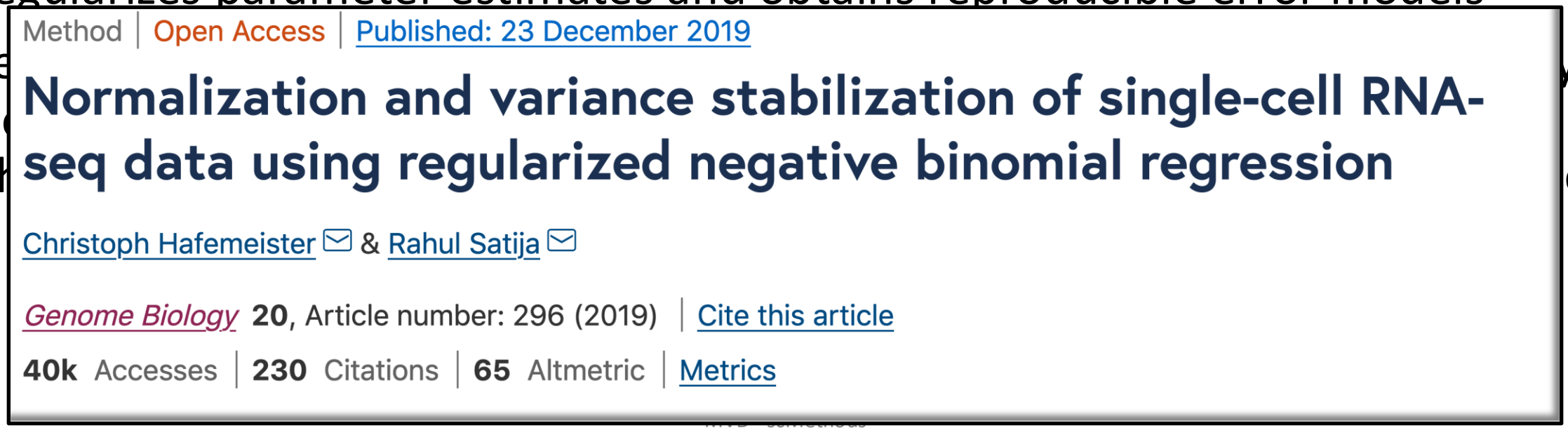
Wiki

scTransform

- SCTransform(pbmc, vars.to.regress = "percent.mt", verbose = FALSE)
- scRNA-Seq is confounded by technical factors including sequencing depth
- Use a modelling framework for the normalization and variance stabilization of molecular count data
- Omits the need for heuristic steps including pseudo counts addition or log-transformation
- Improves common downstream analytical tasks such as variable gene selection, dimensional reduction, and differential expression

scTransform (2/2)

- Show that different groups of genes cannot be normalized by the same constant factor
- Construct a generalized linear model (GLM) for each gene with UMI counts as the response and UMI counts as the explanatory variable
- pooling information across genes with similar abundances, scTransform regularizes parameter estimates and obtains reproducible error models
- re
  n
  ch



Scran

- “size factors” like in DESeq2
- normalization is performed on pooled counts for multiple cells, where the incidence of problematic zeroes is reduced by summing across cells
- pooled size factors are then deconvolved to infer the size factors for the individual cell

METHOD | Open Access | Published: 27 April 2016

Pooling across cells to normalize single-cell RNA sequencing data with many zero counts

Aaron T. L. Lun, Karsten Bach & John C. Marioni

Genome Biology **17**, Article number: 75 (2016) | Cite this article

Conclusion

- Try different methods
- I would suggest to use scTransform
- For parasites dataset we have good results with SCRAN as it deals better with uneven RNA content
- Ross likes the log p + 1 method

Single cell underlying methods

- A lot of math & stats
- Similar concepts are used again and again
- Be critical in usage – tools will always do something- they don’t say when it does not make sense

- Integration
- Differentially expression

- Let’s introduce pseudo time tomorrow

Conclusions

single cell RNA-Seq is noisy

Need to relate the finding to “biology”

Many methods exists

Still open problems: How good is integration? DE!

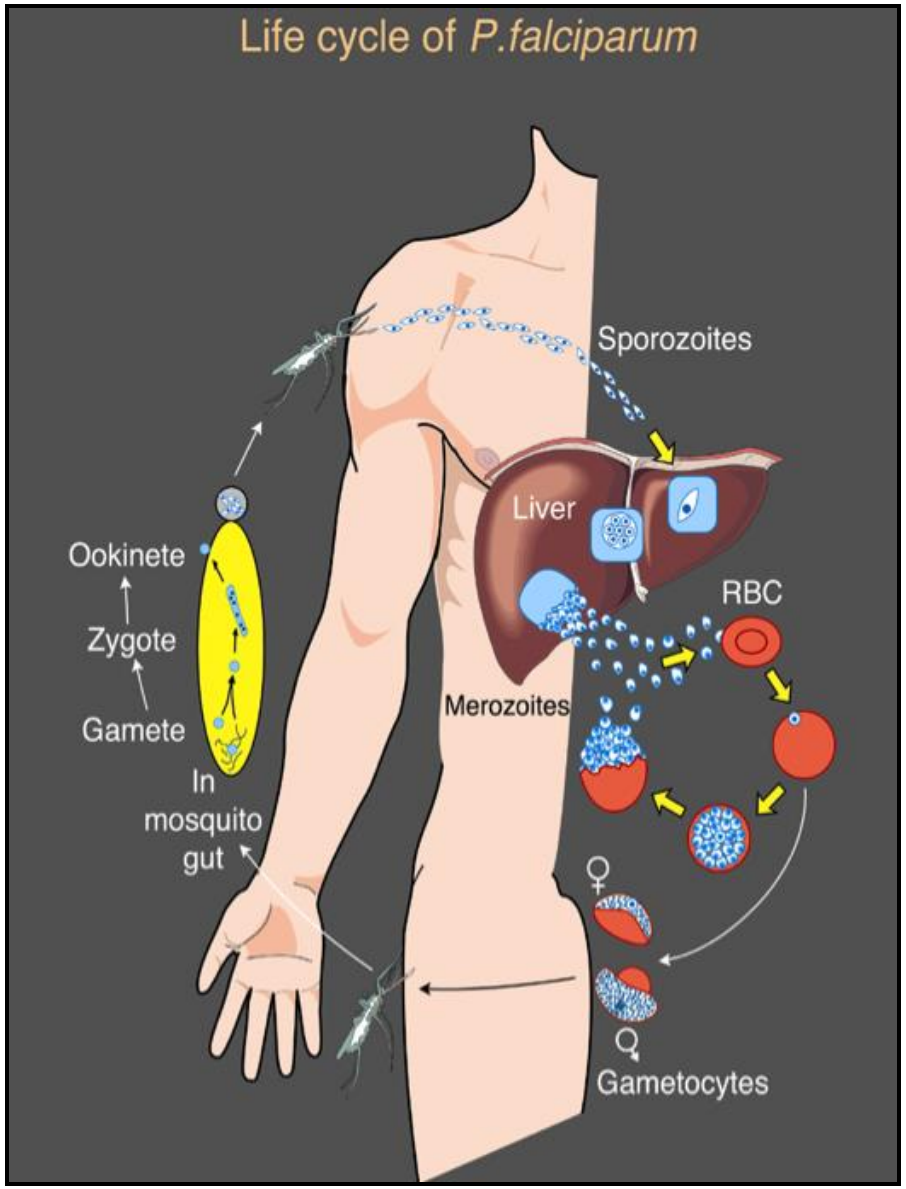
Vehicle to understand your biological question

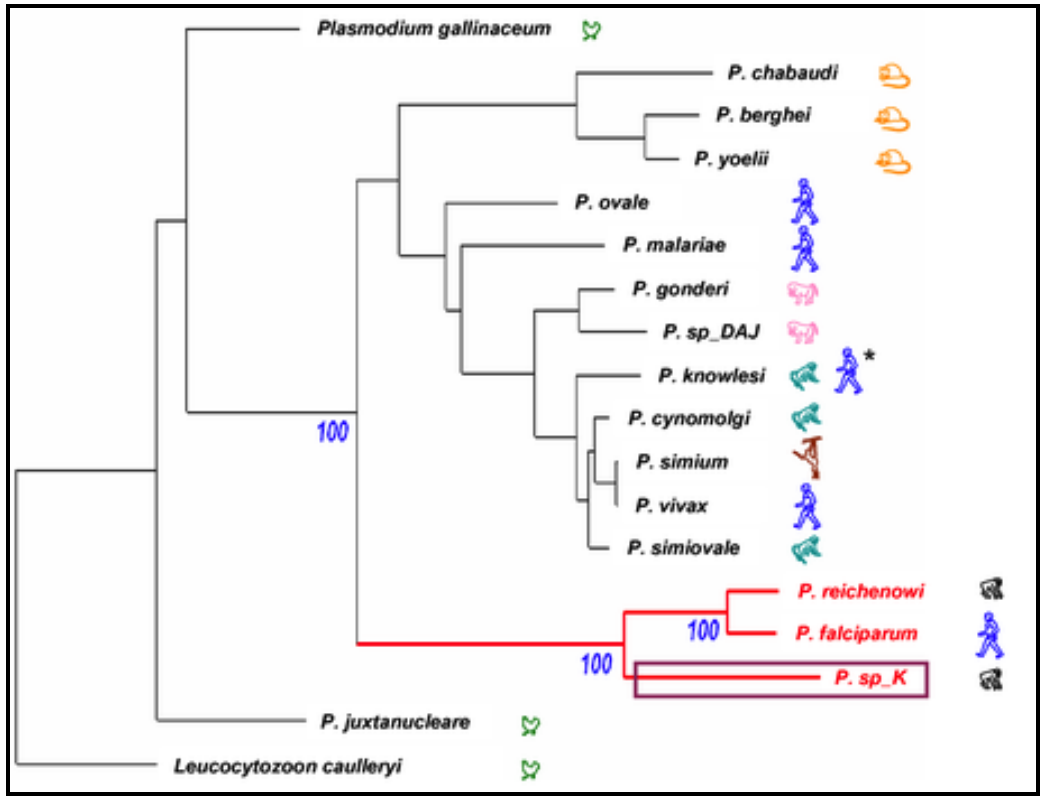
Pseudo time is powerful to find dynamics

Several exiting new tools

Another examples?

Malaria & Phylogeny



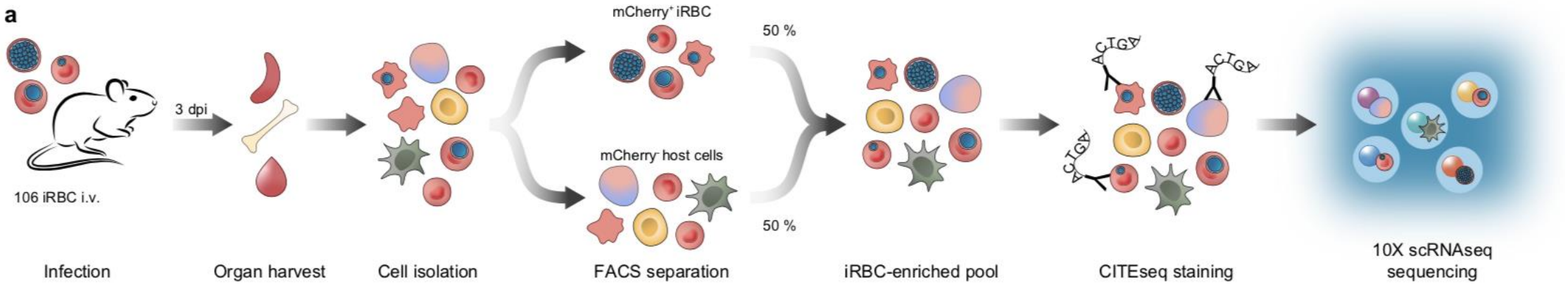


207 million clinical episodes, approx. 627,000 deaths (2012); 408,000 in 2018; 620,00 2020

No vaccines // Fast development of drug resistance

Parasites needs to balance in-host replication and between-host transmission

Host parasite interaction in Malaria



Franziska Hentzschel

Science Advances

Current Issue | First release papers | Archive | About | Submi

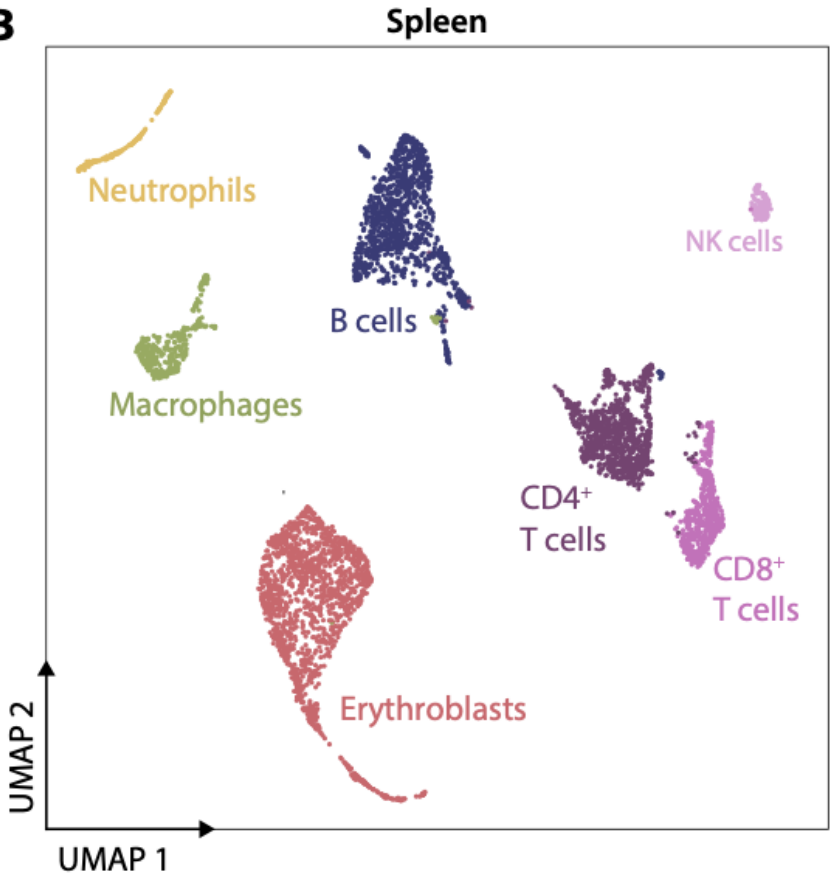
HOME > SCIENCE ADVANCES > VOL. 8, NO. 17 > HOST CELL MATURATION MODULATES PARASITE INVASION AND SEXUAL DIFFERENTIATION IN PLASMODIUM BERGHEI

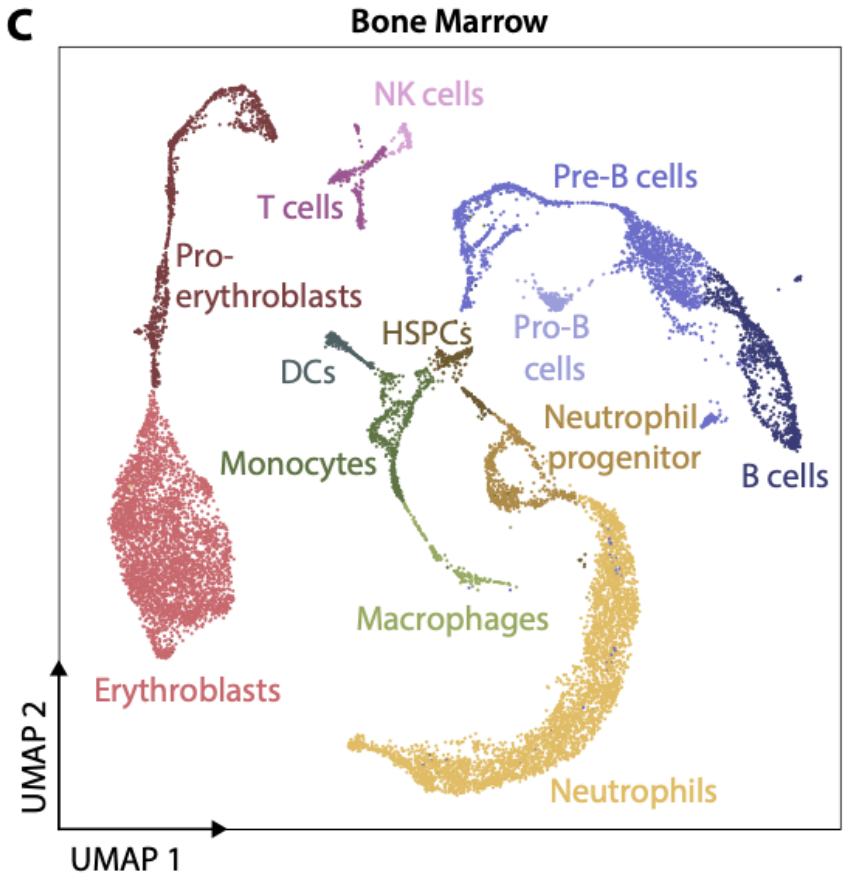
RESEARCH ARTICLE | MICROBIOLOGY

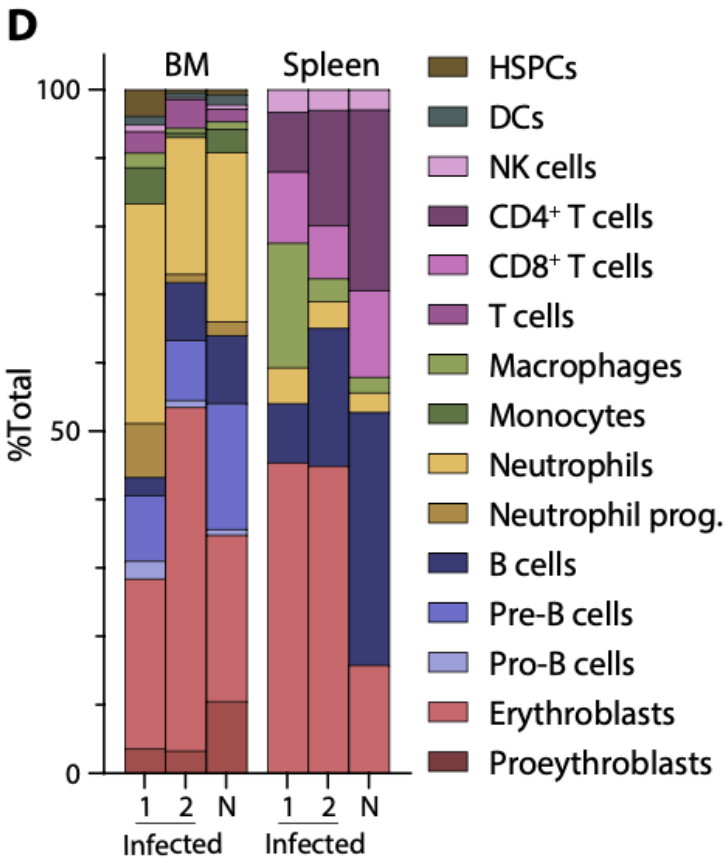
Host cell maturation modulates parasite invasion and sexual differentiation in *Plasmodium berghei*

FRANZISKA HENTZSCHEL, MATTHEW P. GIBBINS, CHARALAMPOS ATTIPA, DARIO BERALDI, CHRISTOPHER A. MOXON, THOMAS D. OTTO, AND MATTHIAS MARTI Authors Info & Affiliations

Samples in Spleen and bone marrow



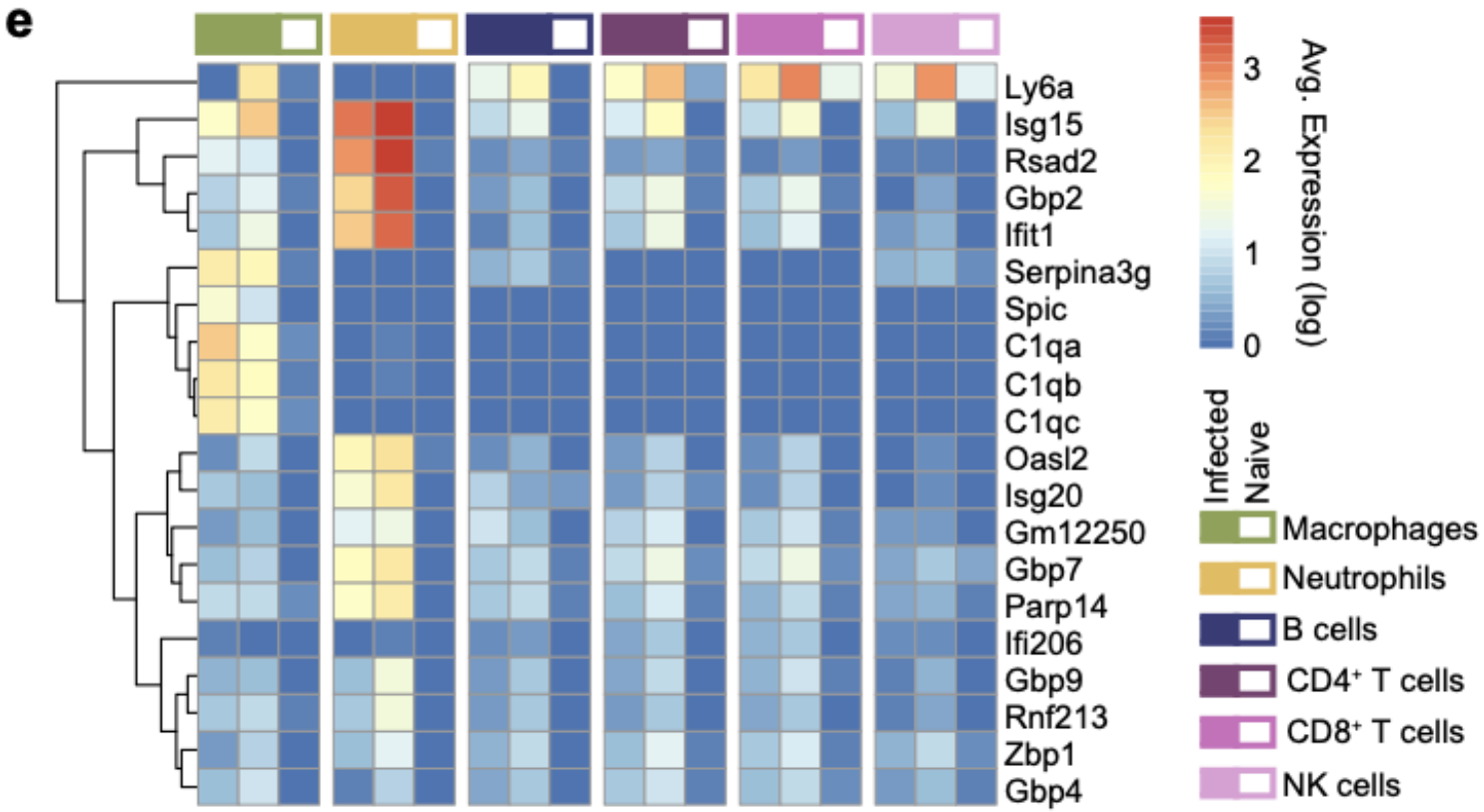




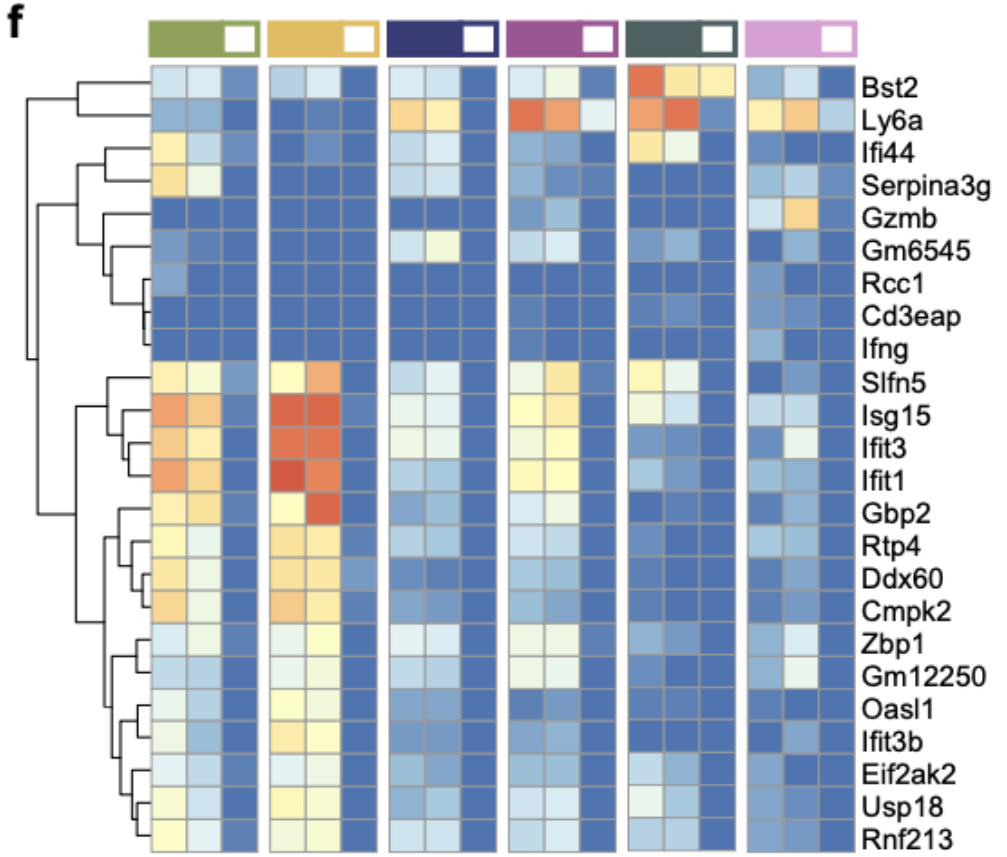
Relative abundance needs more replicates

Differentially expressed genes?

Spleen

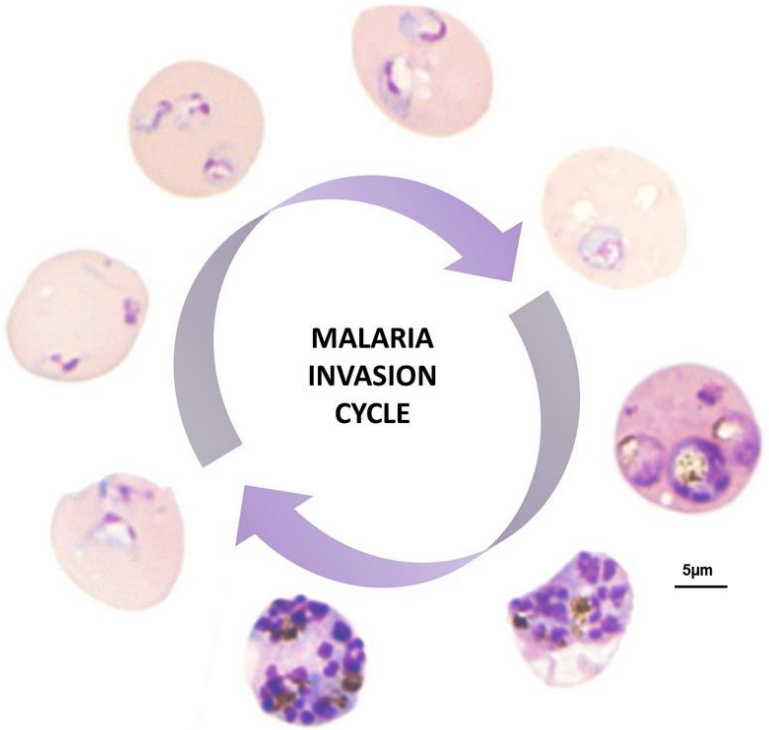


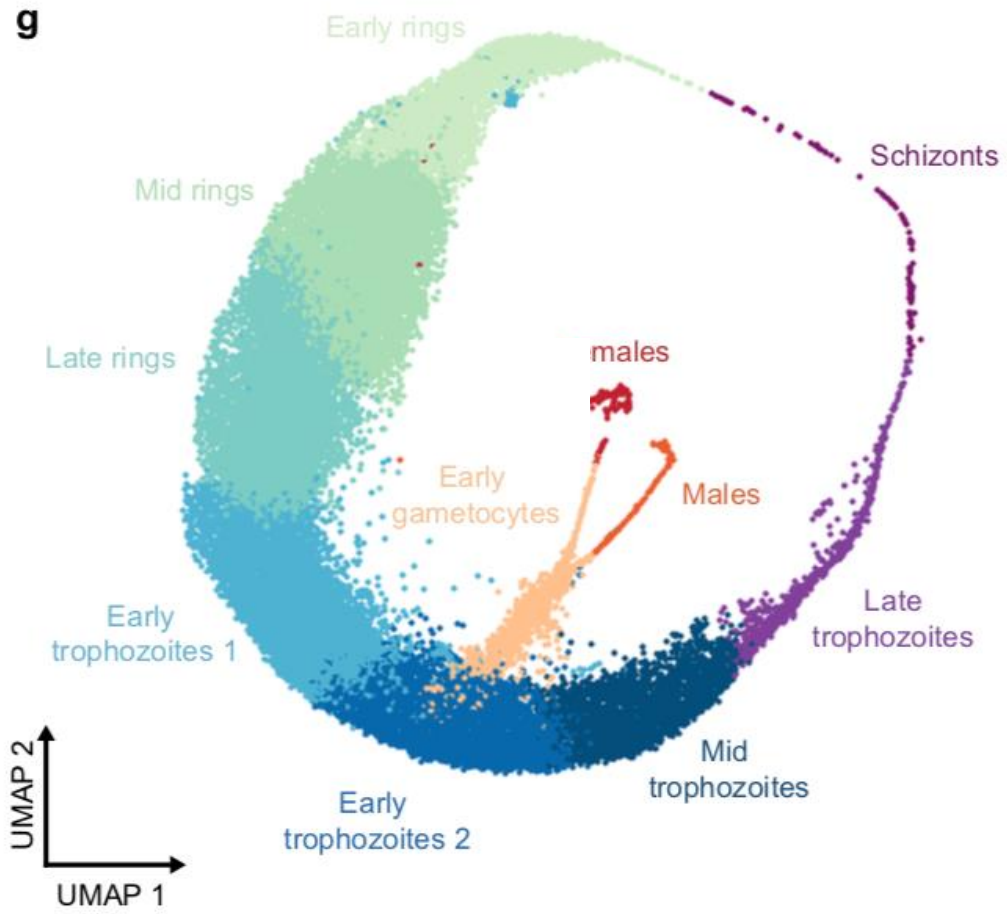
Bone Marrow



Generally more going on in MK / Monocytes – Ifits1, Lsg15

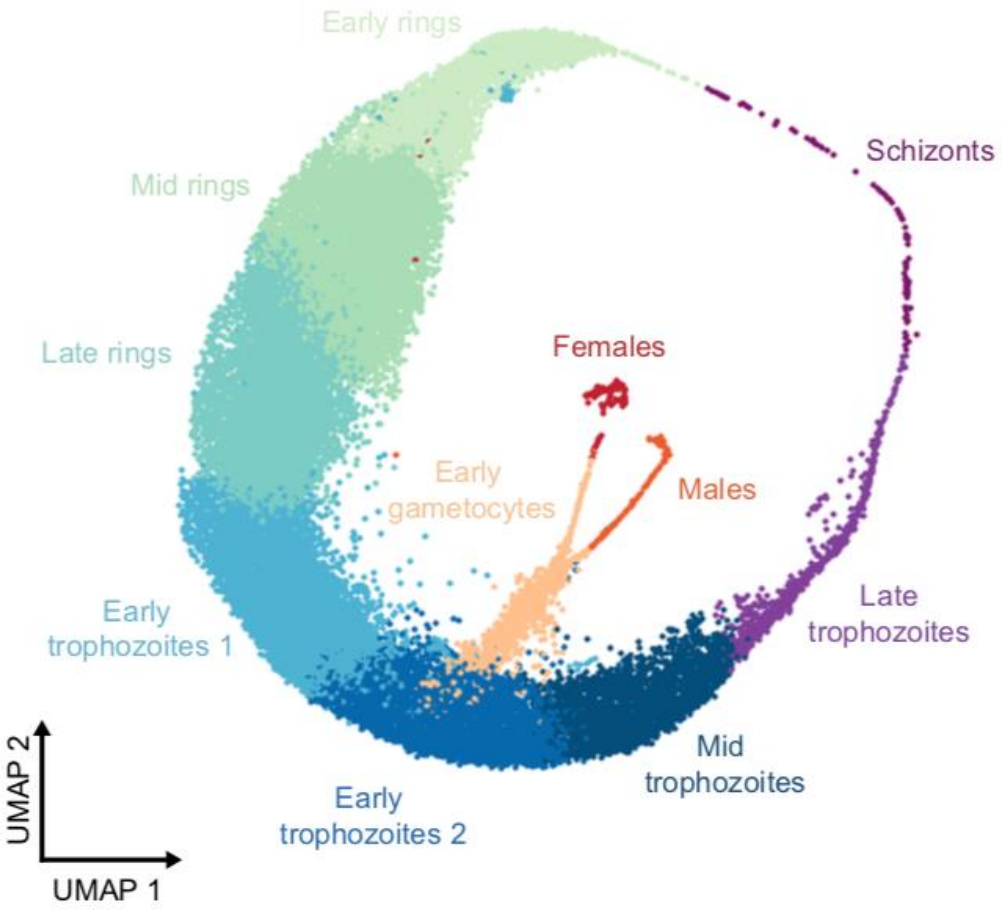
Invasion phenotype

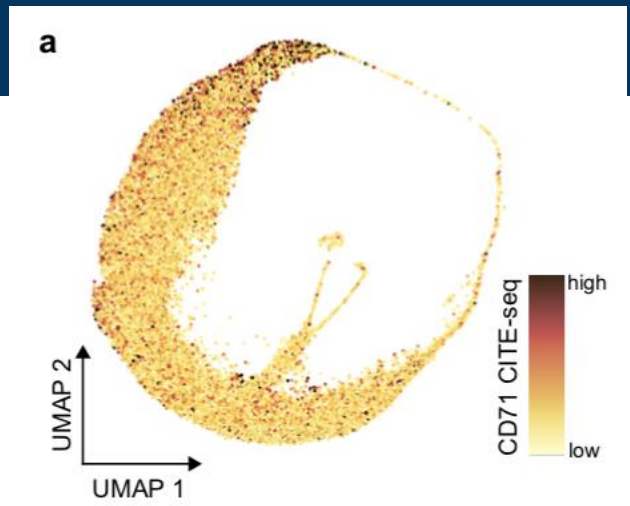


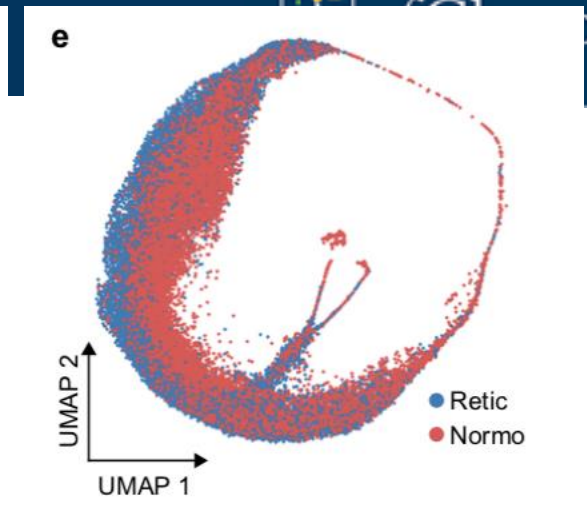


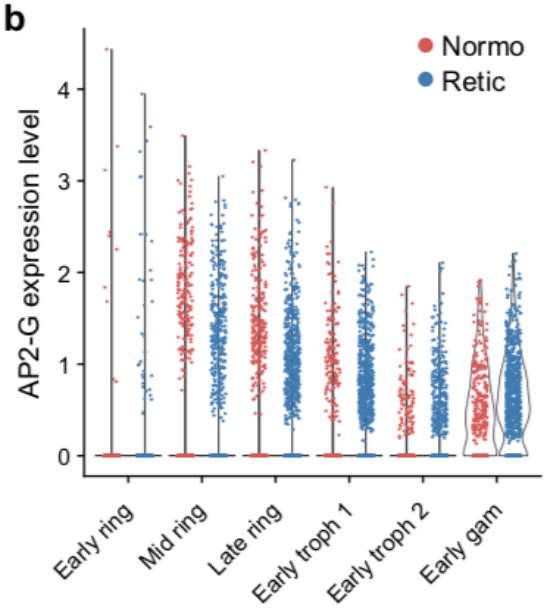
***P. berghei* sexual development occurs preferentially in reticulocytes**

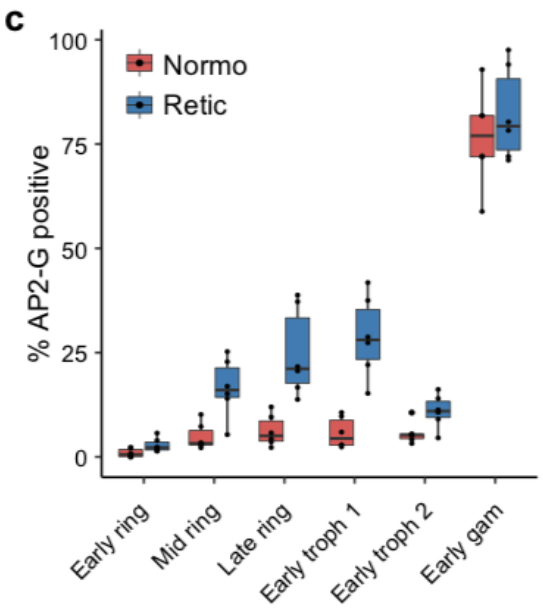
Invasion phenotype











***P. berghei* sexual development occurs preferentially in reticulocytes**

Hentzschel *et al*, Science advances 2022

Summary

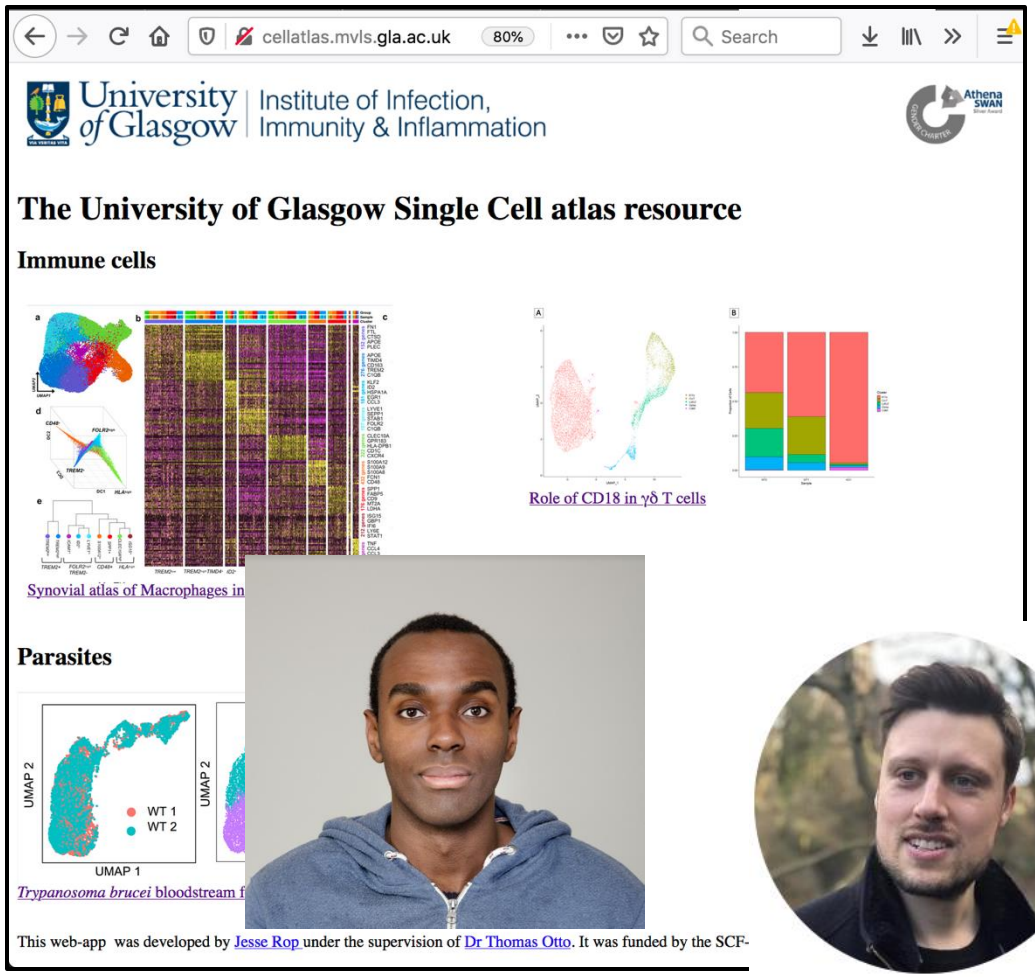
- dual scRNA-Seq is powerful…
- … but not correct setup

- But how could you look at these data?

Learning aims

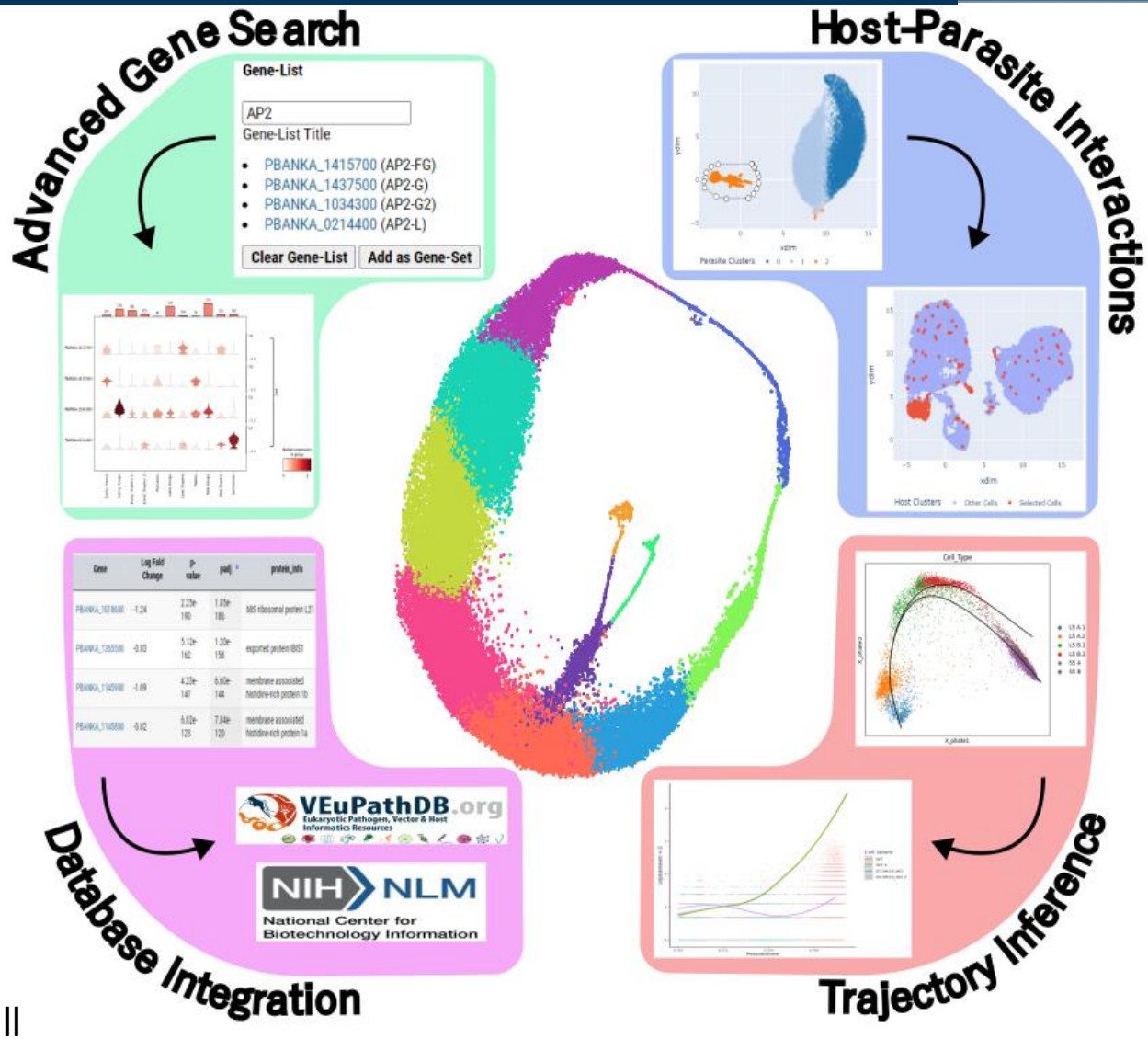
- Introduce more technical concepts
- Explore how to process scRNA-Seq data
- Learn new methods, normalization, integration and DE in scRNA-Seq
- Critical evaluation of scRNA-Seq
- Overview of developments in scRNA-Seq

Data accessibility - ParaCell



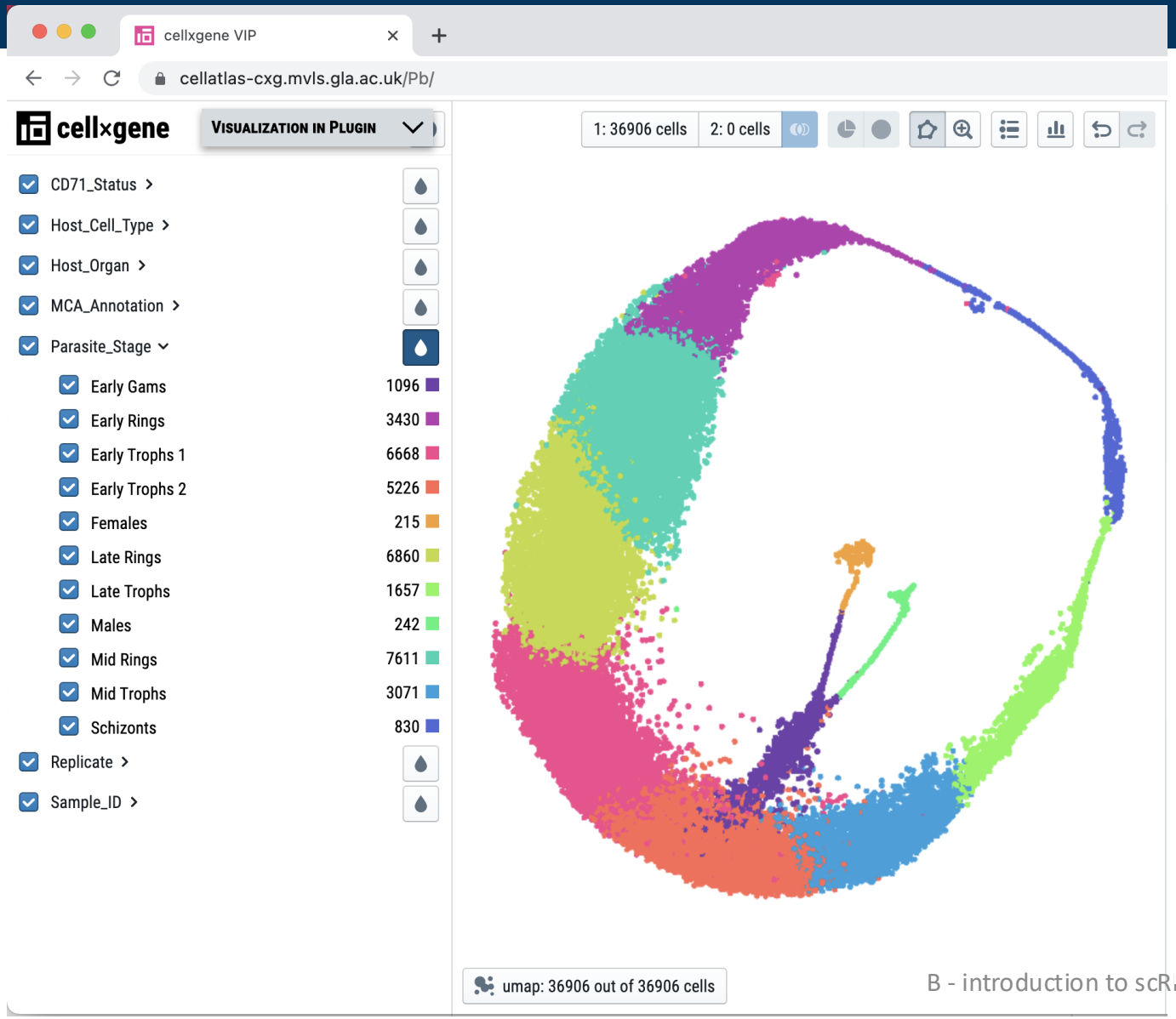
Edward Agboraw

William Haese-Hill



http://cellatlas.mvls.gla.ac.uk/

Cellxgene - paraCell



JOURNAL ARTICLE

paraCell: a novel software tool for the interactive analysis and visualization of standard and dual host–parasite single-cell RNA-seq data

Edward Agboraw, William Haese-Hill, Franziska Hentzschel, Emma Briggs, Dana Aghabi, Anna Heawood, Clare R Harding, Brian Shiels, Kathryn Crouch, Domenico Somma ...
Show more

Nucleic Acids Research, Volume 53, Issue 4, 28 February 2025, gkaf091, https://doi.org/10.1093/nar/gkaf091
Published: 20 February 2025 **Article history** ▾

Show example

- http://cellatlas.mvls.gla.ac.uk/


University of Glasgow
Questions?
#UofGWorldChangers
@UofGlasgow