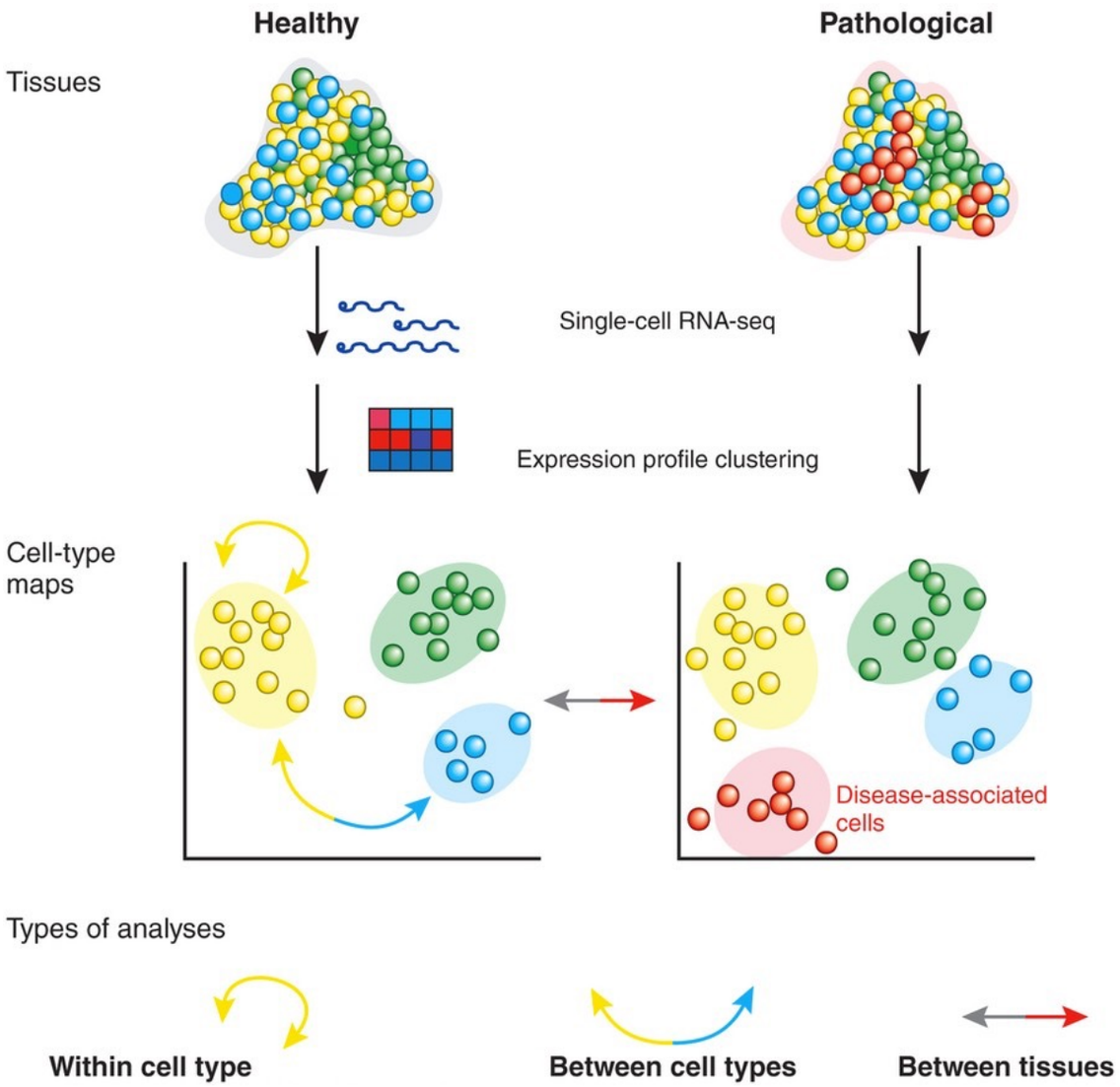
Ecole thématique sincellTE 2022

Downstream analysis

Akira Cortal Ph.D.

Roscoff, January 11th 2022

The context: Single-cell RNA-seq to uncover cell heterogeneity associated to distinct cellular phenotypes



Sandberg R. Nature Methods 11, 22–24 (2014) doi:10.1038/nmeth.2764

Heterogeneity between samples arising from:

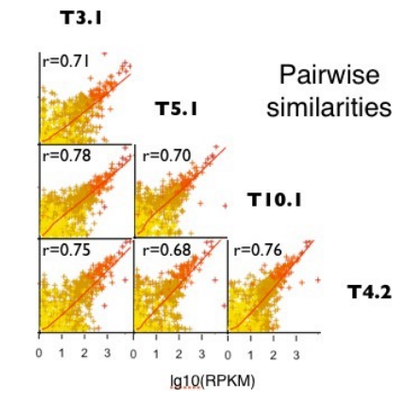
- Genetic factors (Donor)
- Environmental factors
- Treatments / Times of activation
- History of cells (e.g. clonal selection/expansion)
- Natural aging

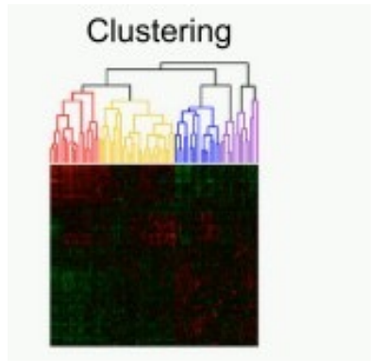
Heterogeneity within samples arising from:

- **Cell fate (permanent):**
  - Different lineages of differentiation
  - Different compositions of cell types
- **Cell state (transient):**
  - Stochasticity of gene expression
  - Pulsation / Circadian-like
  - Associated to cell cycle
  - **Different stages of activation**
- **Technical noise**

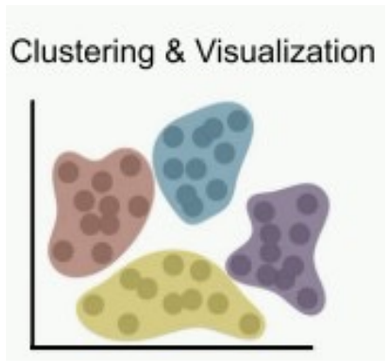
The questions

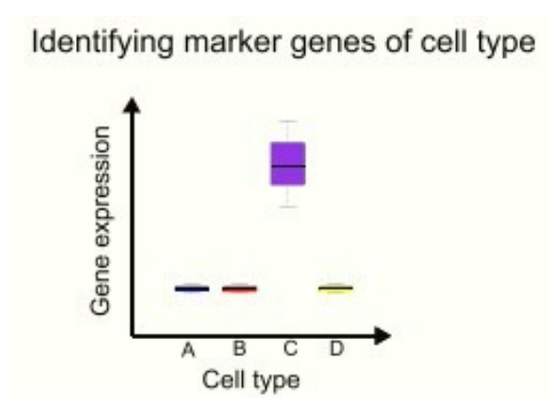
1- Is there **functionally relevant** cell heterogeneity in my data?



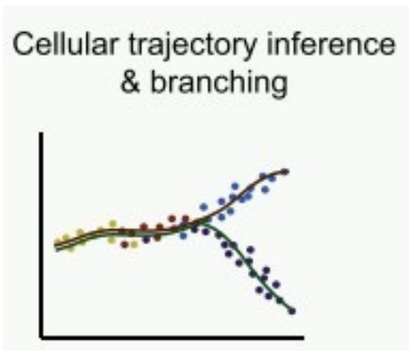


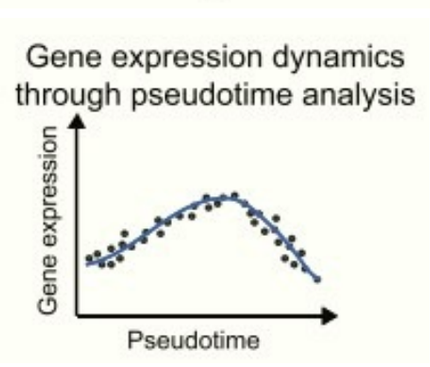
2- Are there distinct subpopulations of cells?





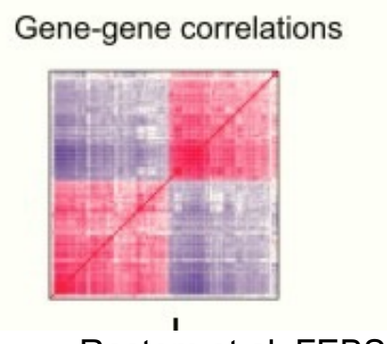
3- Are there continuums of differentiation / activation cell states?

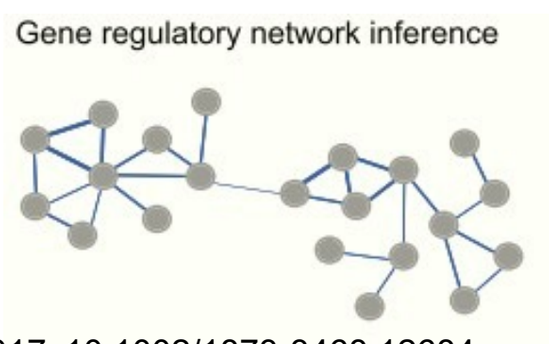




4- Which are the genes driving such heterogeneity?

5- May we learn something about the cellular / molecular mechanisms involved?: e.g. cell differentiation, biological process, pathways, regulatory modules, etc?





Rostom et al, FEBS 2017. 10.1002/1873-3468.12684

The computational challenges

1. Lower coverage/depth than bulk RNA-seq

2. Technical & biological noise

3. High dimensionality

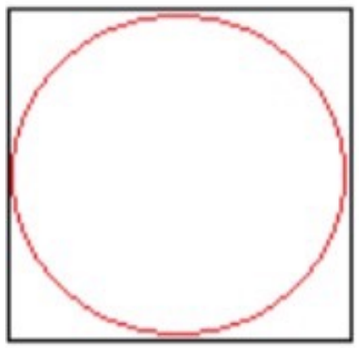
4. High variability

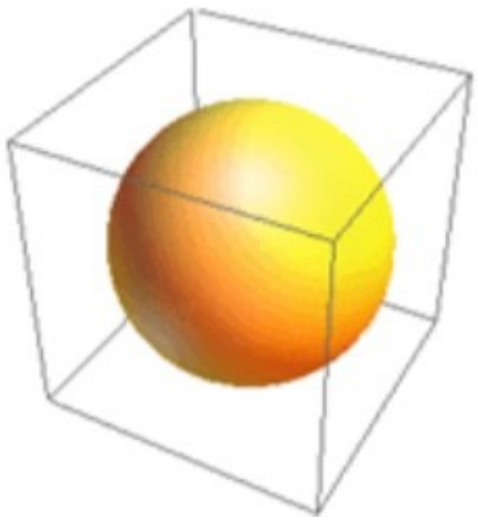
5. Dropouts => Zero-inflated data

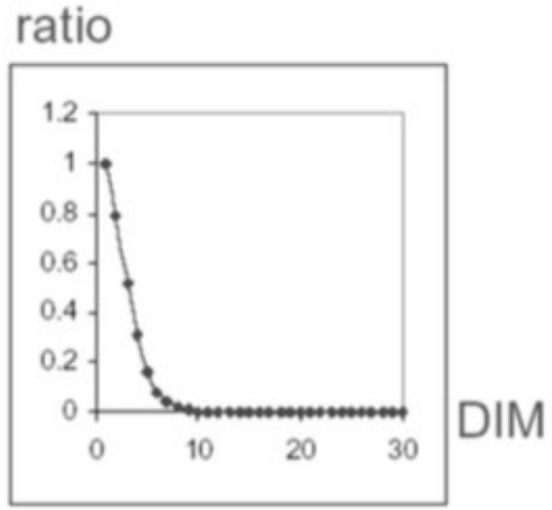
6. Multimodality

The challenges - High dimensionality

The curse of dimensionality: When dimensionality increases, data becomes increasingly sparse in the space that it occupies



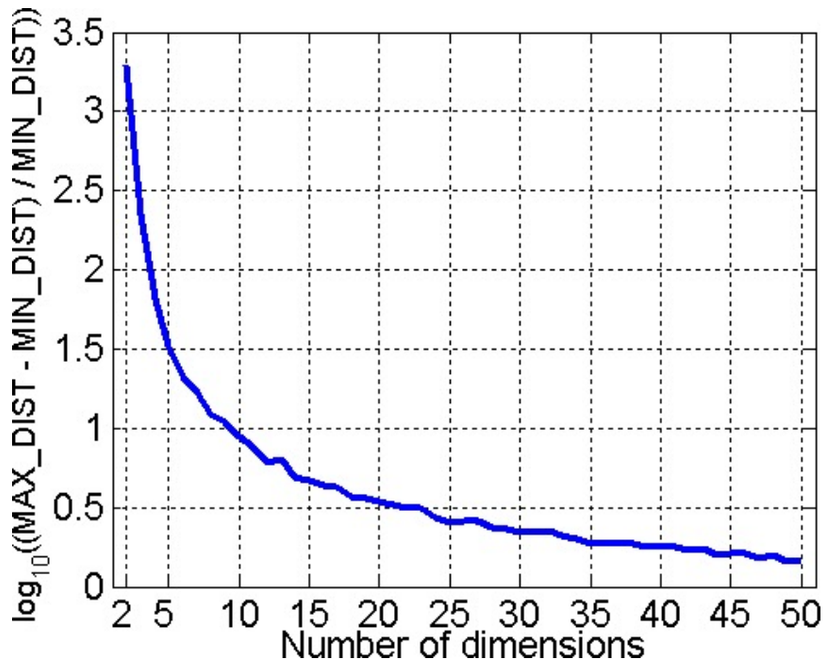




Definitions of density and distance between points, which is critical for clustering and outlier detection, become less meaningful

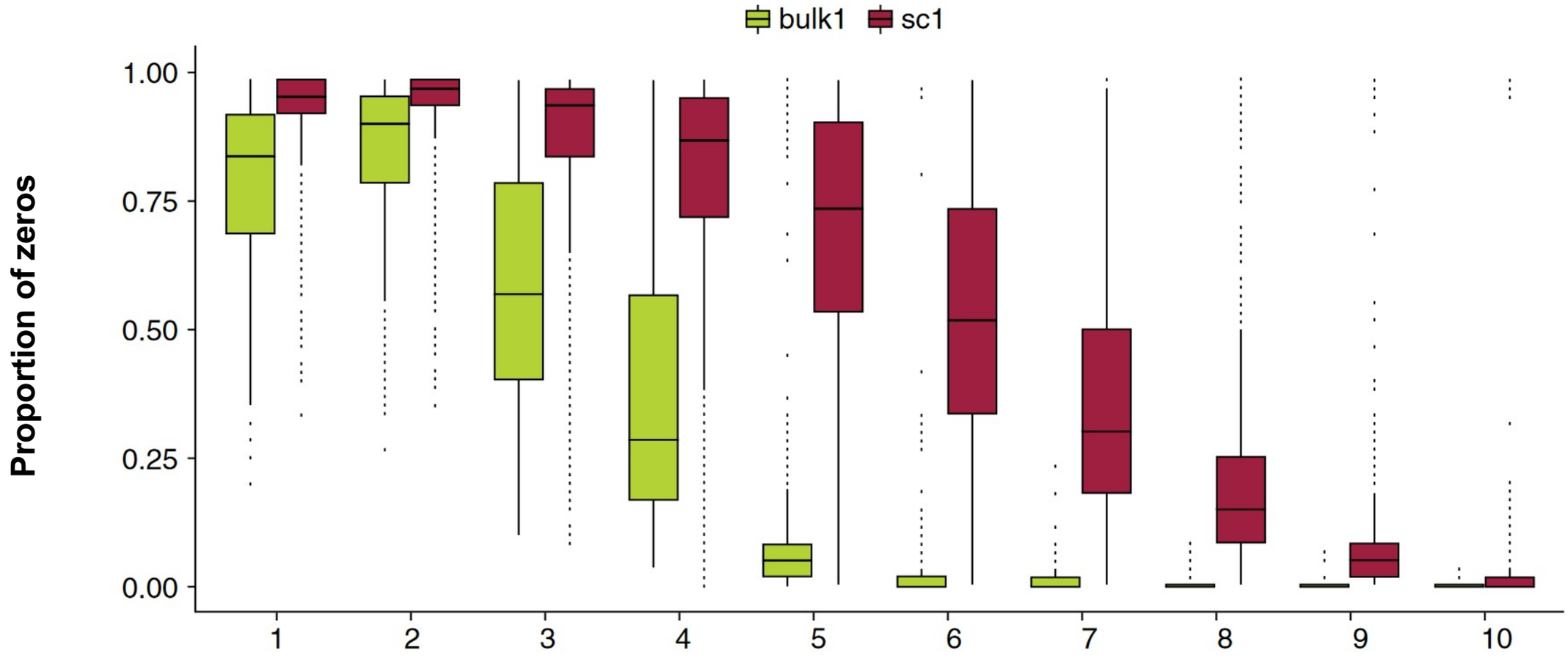
Randomly generate 500 points

Compute difference between max and min distance between any pair of points



Taken from Tan, Steinbach & Kumar, Introduction to Data Mining course
http://slideplayer.com/slide/6194466

The computational challenges - Zero Inflated data

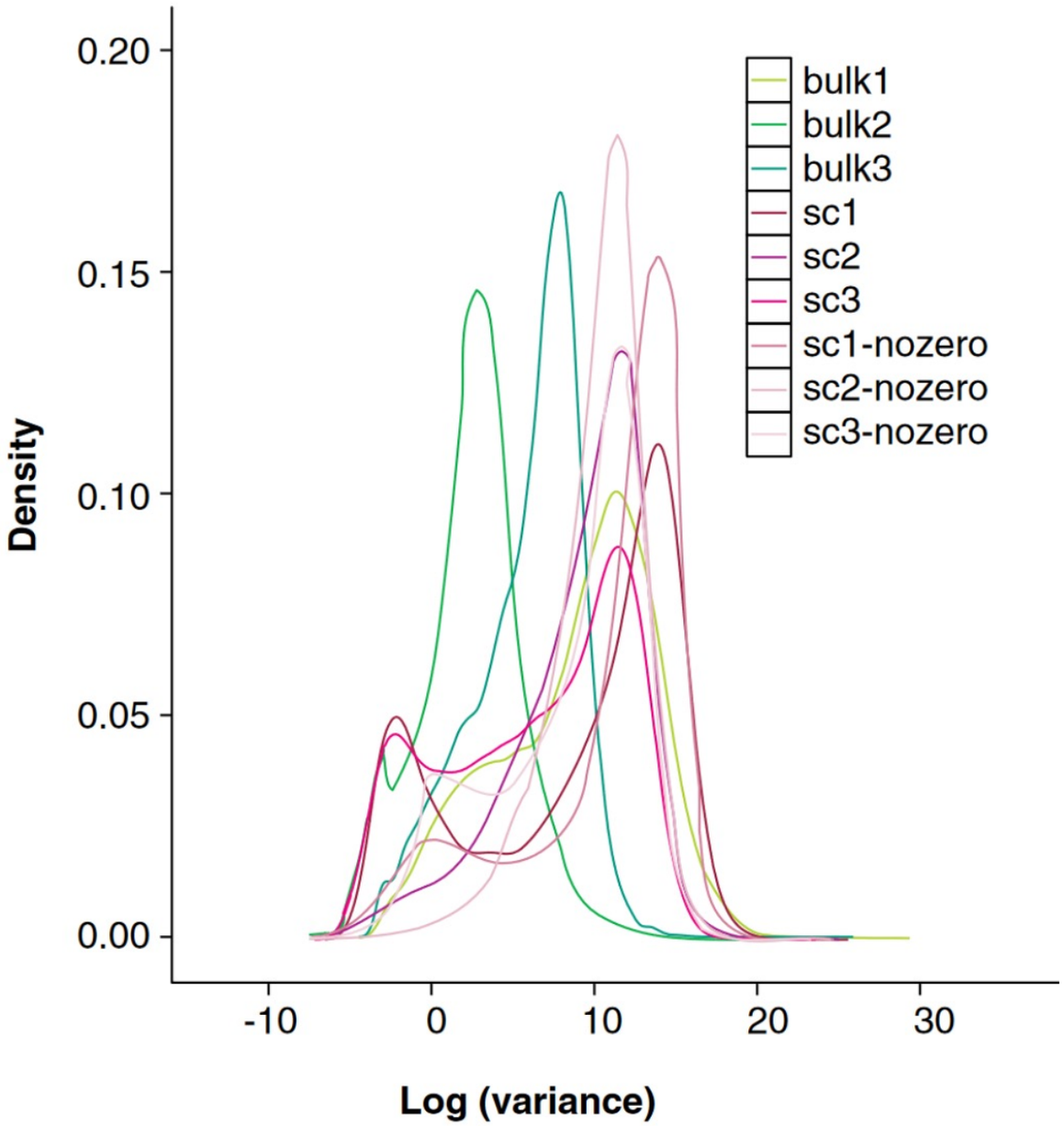


Bulk and SC sets with comparable depths

Bulk 1: 60 female bulk RNA-seq samples of individual Drosophila flies
SC1: 60 individual Mus musculus embryonic cells at various developmental time points

The computational challenges - High variability (overdispersion)

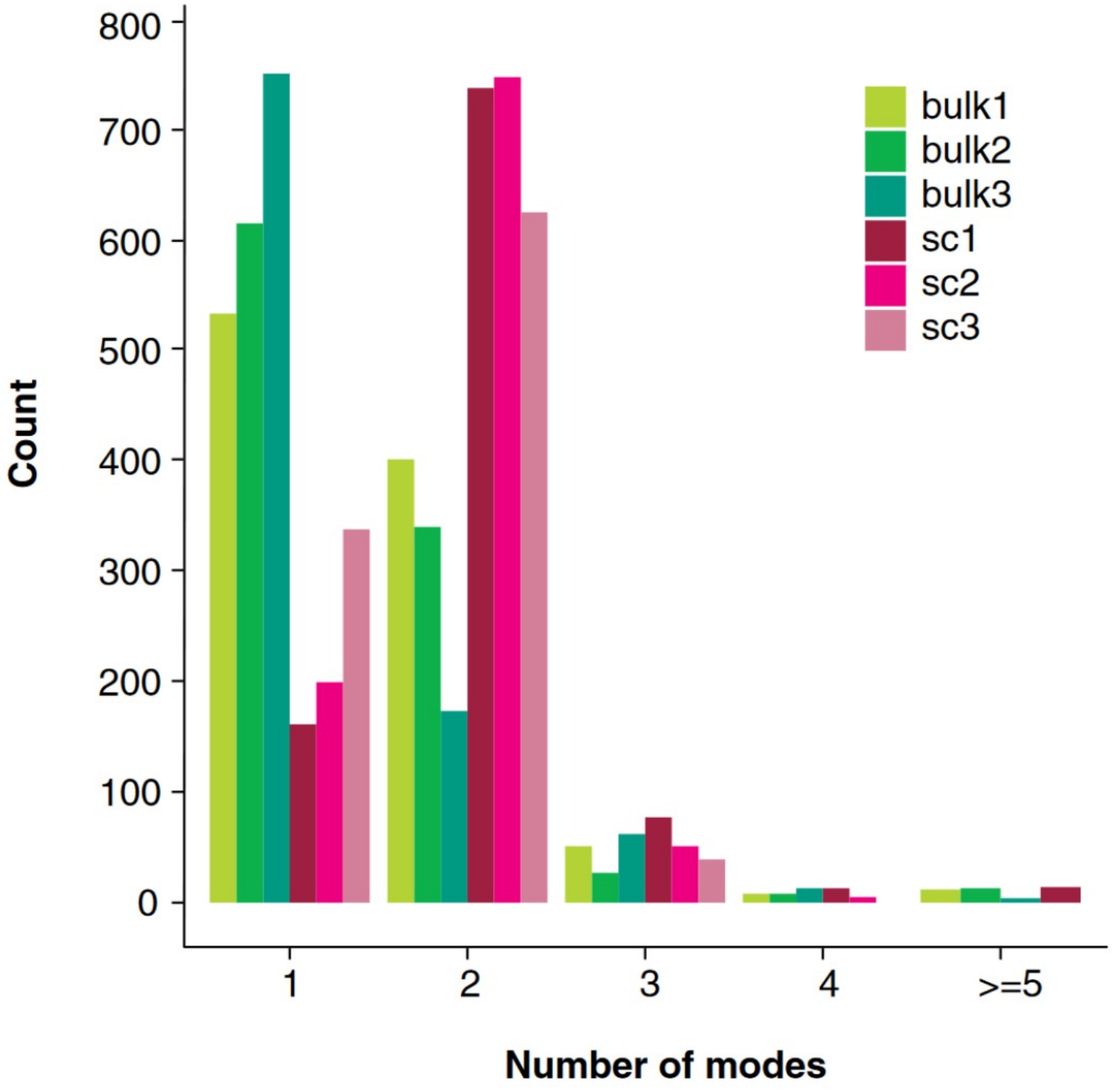
Densities of gene-specific log variance for all genes in three bulk and three single-cell RNA-seq dataset



- Much more variability of gene expression in single cell data than bulk.
- This is also true when taking in account the technical zero counts.

Bacher, Rhonda, and Christina Kendziorski. 'Design and Computational Analysis of Single-Cell RNA-Sequencing Experiments'. *Genome Biology* 17, no. 1 (avril 2016): 63. https://doi.org/10.1186/s13059-016-0927-y.

The computational challenges - Multimodality



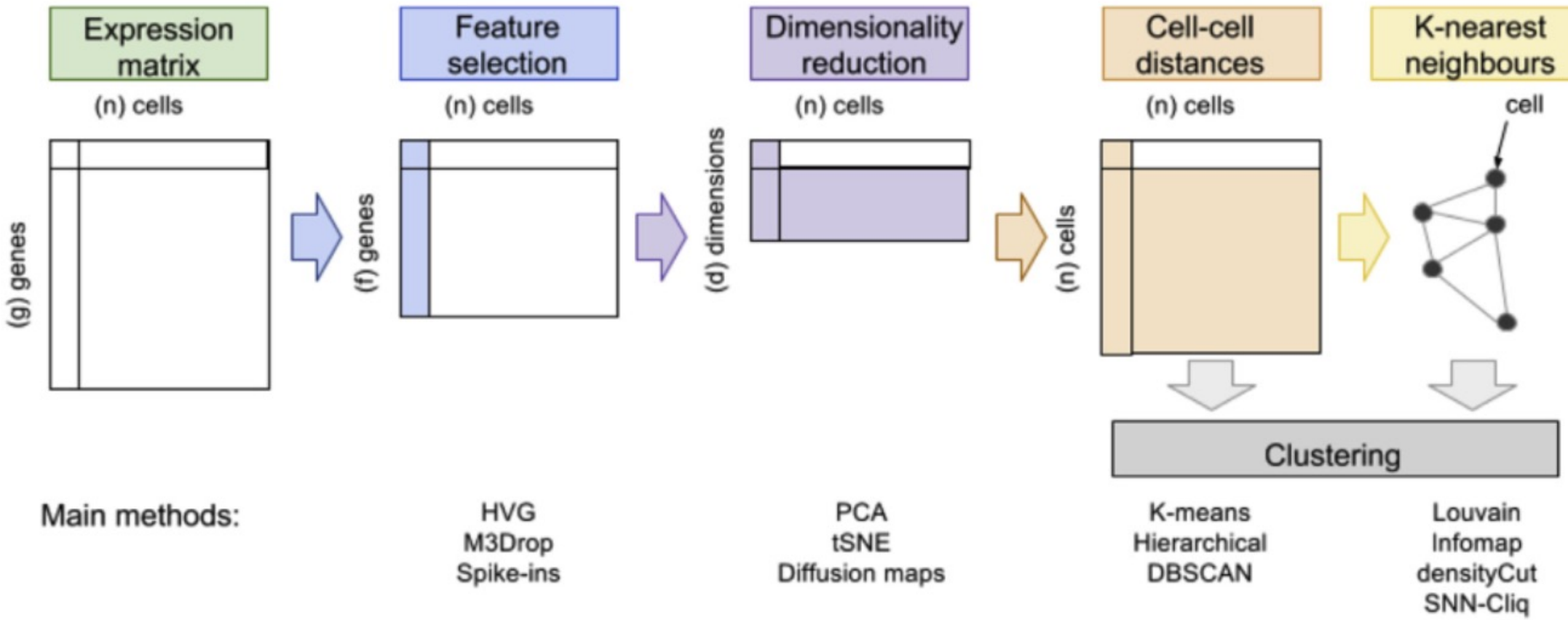
- Much more “modes” of expression in single cell data.
- 1 modes -> “expressed – non expressed”
- 2 modes -> “non expressed, lowly expressed, highly expressed”
- 3 modes -> low, medium, high expression

Bacher, Rhonda, and Christina Kendziorski. ‘Design and Computational Analysis of Single-Cell RNA-Sequencing Experiments’. *Genome Biology* 17, no. 1 (avril 2016): 63. https://doi.org/10.1186/s13059-016-0927-y.

The bioinformatics pipeline: main “modular” components

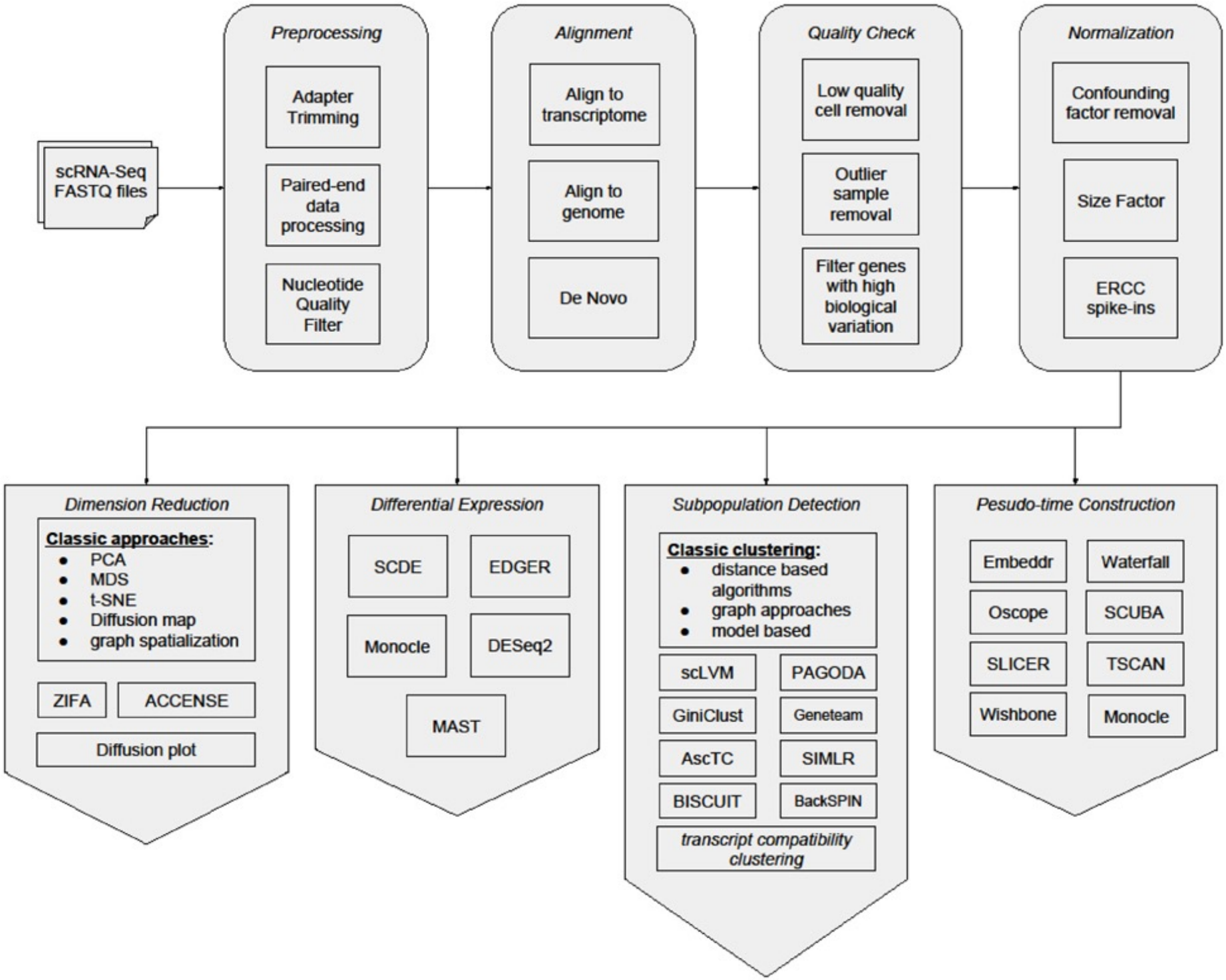
1. **Feature selection**
2. **Dimensionality Reduction**
3. **Clustering / Hiearchy**
4. **Differential Expression / Gene signature extraction**
5. **Functional interpretation**
6. **Batch effect correction**

The bioinformatics pipeline: Example 1



Andrews, Tallulah S., and Martin Hemberg. 'Identifying Cell Populations with ScRNASeq'. *Molecular Aspects of Medicine*, The emerging field of single-cell analysis, 59 (février 2018): 114–22. https://doi.org/10.1016/j.mam.2017.07.002.

The bioinformatics pipeline: Example 2



Poirion, Olivier B., Xun Zhu, Travers Ching, and Lana Garmire. 'Single-Cell Transcriptomics Bioinformatics and Computational Challenges'. *Frontiers in Genetics* 7 (2016): 163. https://doi.org/10.3389/fgene.2016.00163.

Online catalogue: scRNA-tools database

sc RNA-tools Table Tools Analysis Updates Submit FAQs

Tools Table

Sort, filter and download the database

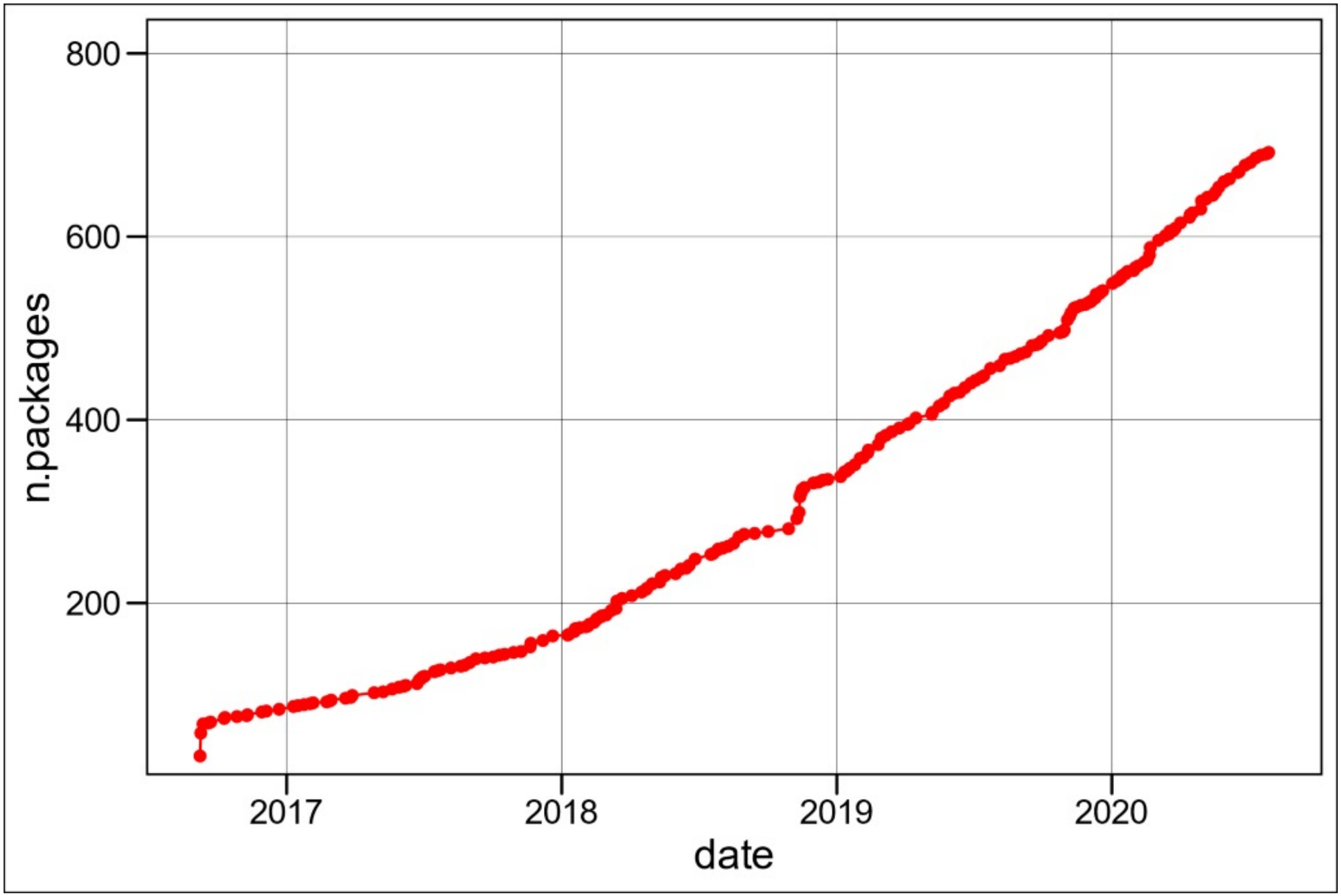
Search

NAME	PLATFORM	DOIS	CITATIONS	LICENSE	CATEGORIES
acorde	R	10.1101/2021.05.07.441841	0	GPL-3.0	Alternative Splicing, Differential Expression, Visualisation
ACTINN	Python	10.1093/bioinformatics/btz592, 10.1101/532093	18	GPL-3.0	Classification
ACTION	C++/R/MATLAB	10.1038/s41467-018-03933-2, 10.1101/081273	33	-	Clustering, Dimensionality Reduction, Gene Networks
ACTIONet	R/C++	10.1038/s41467-020-18416-6, 10.1101/746339	12	GPL-2.0-or-later	Classification, Clustering, Dimensionality Reduction, Gene Sets, Integration, Normalisation, Visualisation
ACTIVA	Python	10.1101/2021.01.28.428725	1	MIT	Simulation
ADC	Python	10.1371/journal.pcbi.1009548	0	-	Gene Networks
ADImpute	R	10.1101/611517	5	GPL-3.0	Imputation
adobo	Python	10.1093/bioinformatics/btaa269	20	GPL-3.0	Cell Cycle, Classification, Clustering, Differential Expression, Dimensionality Reduction, Gene Filtering, Imputation, Normalisation, Quality Control, Visualisation

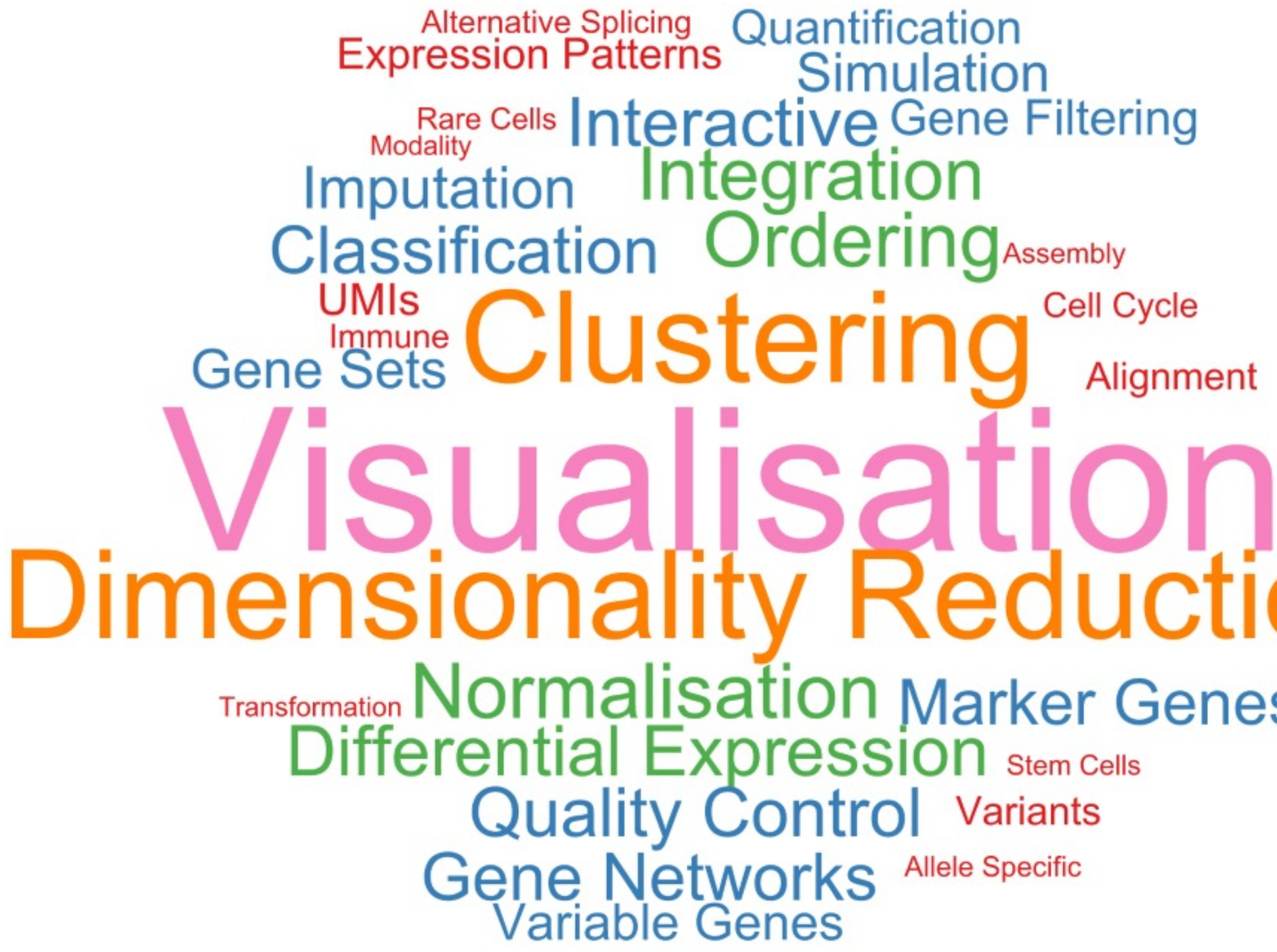
Zappia, L., Phipson, B., Oshlack, A., 2018. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. PLOS Computational Biology 14, e1006245. https://doi.org/10.1371/journal.pcbi.1006245

Online catalogue: scRNA-tools database

Number of single cell methods



Online catalogue: scRNA-tools database



Examples of *all-in-one* environments: SEURAT

Expression QC

Normalization

Highly variable genes

Dealing with confounders

Dimensional Reduction

Visualization

Marker genes

Cell Cycle Regression

Clustering cells

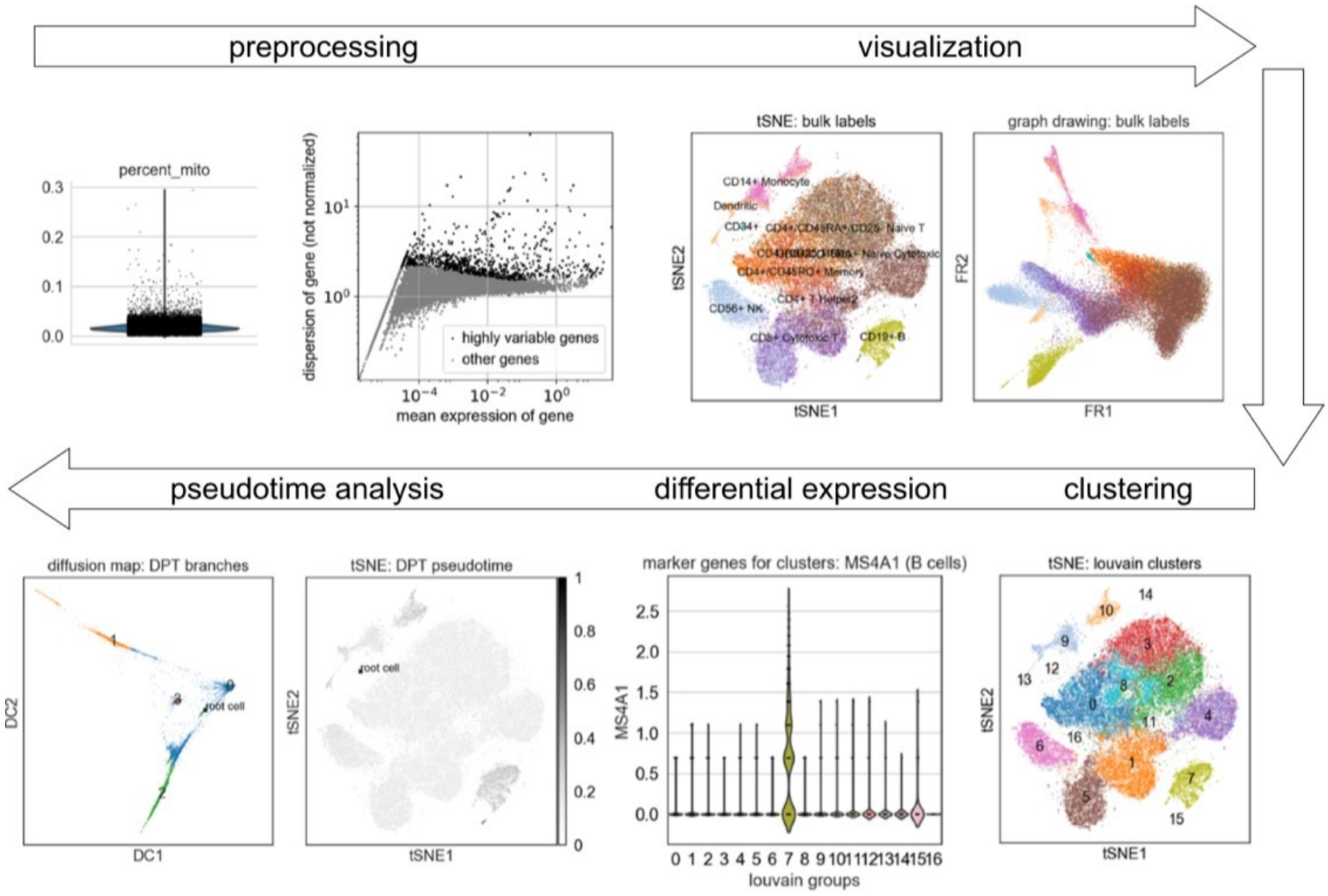
Differential expression

Multimodal Analysis

https://satijalab.org/seurat/

See tutorial at:
https://hemberg-lab.github.io/scRNA.seq.course/seurat-chapter.html

Examples of *all-in-one* environments: SCANPY



SCANPY: large-scale single-cell gene expression data analysis.
Wolf et al. Genome Biology 2018, 19:15 https://doi.org/10.1186/s13059-017-1382-0

A word about Hardware

Laptop
CPU: 2~8 cores
RAM:4~16 Gb

Analysis of one sample with 2000~10000 cells

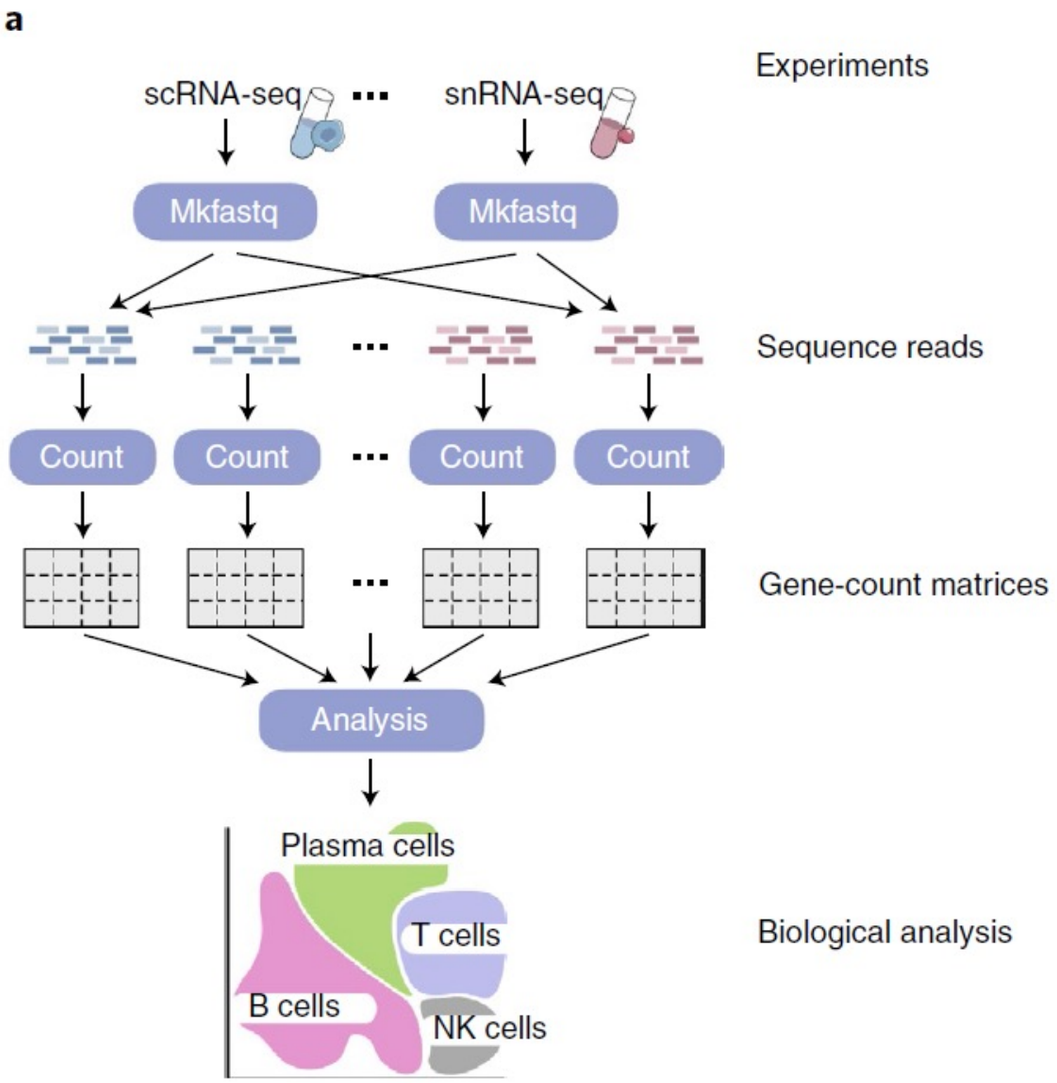
Workstation
CPU: 8~32 cores
RAM:16~128 Gb

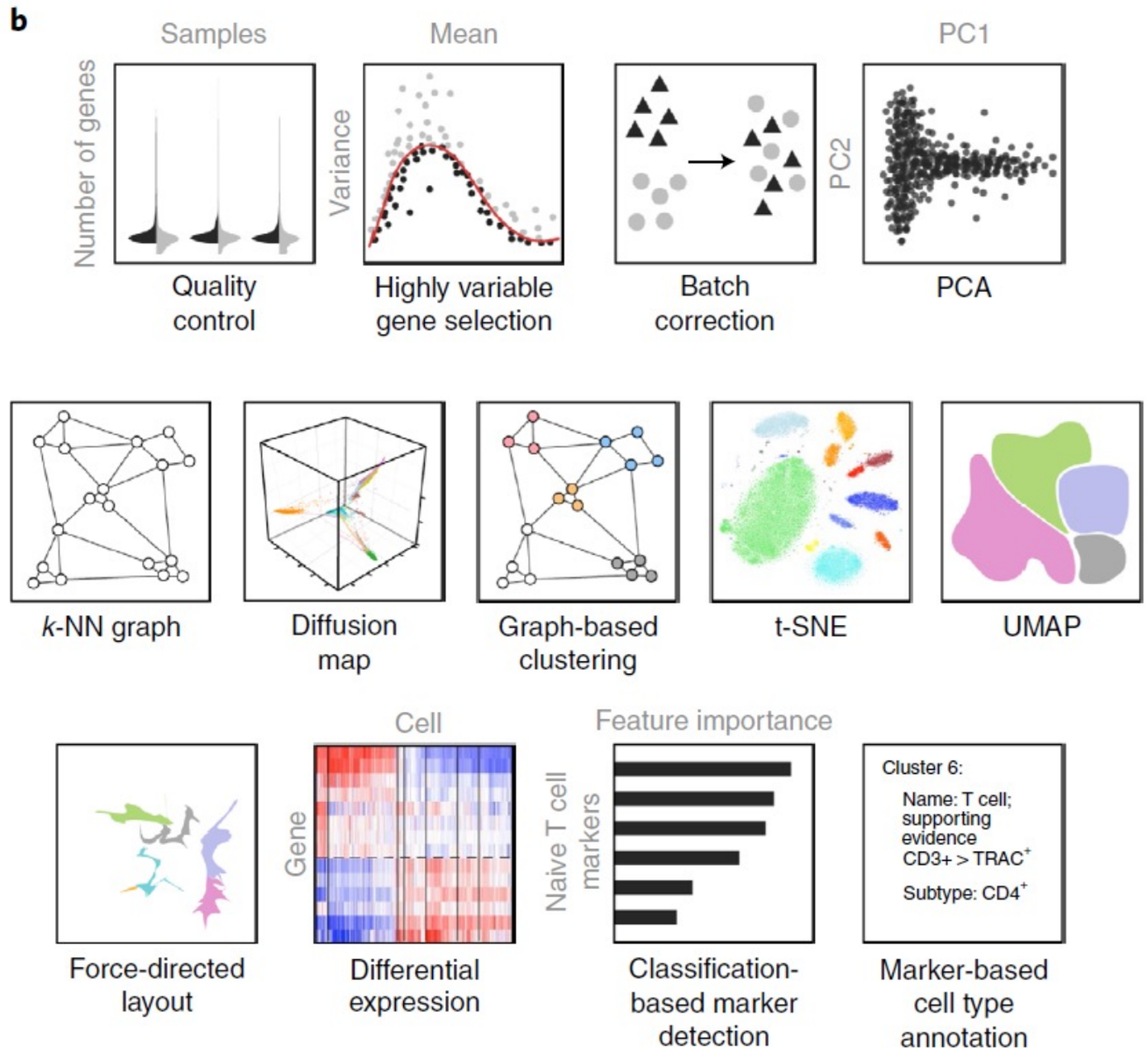
Analysis of medium scale experiment with 4 ~ 8 samples 50000~100000 cells

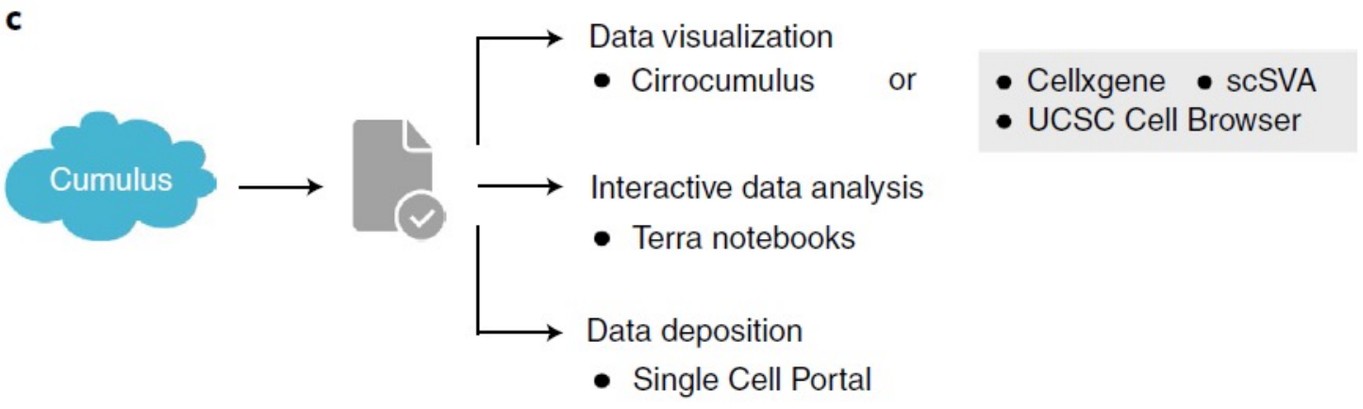
Server
CPU: 16~128 cores
RAM: 256Gb~4 Tb

Analysis of large scale experiment and atlases with more than 100000 cells in total

Cumulus cloud solution







Li, B. *et al.* (2020) 'Cumulus provides cloud-based data analysis for large-scale single-cell and single-nucleus RNA-seq', *Nature Methods*, 17(8), pp. 793–798. doi:10.1038/s41592-020-0905-x.

Cumulus cloud solution

Table 1 | Cumulus is computationally efficient and cost-effective

	Time[a]			Cost per sample[b]
	Cell Ranger + Seurat version 3	Cell Ranger + SCANPY	Cumulus	Cumulus
Total	10 d, 5 h, 38 min	9 d, 5 h, 35 min	15 h, 15 min	US$1.832
Mkfastq	13 h, 18 min	13 h, 18 min	7 h, 54 min	US$0.22
Count	8 d, 14 h, 12 min	8 d, 14 h, 12 min	6 h, 44 min	US$1.61
Analysis	26 h, 8 min	2 h, 5 min	37 min	US$0.002

[a]Total execution time on the bone marrow dataset of the Cumulus, Cell Ranger + Seurat version 3 or Cell Ranger + SCANPY pipeline, running on a 32-CPU-thread, 120-GB-memory Google Cloud virtual machine instance (Methods). [b]Average computational cost for running Cumulus per sample of ~4,000 cells (Methods).

Li, B. *et al.* (2020) 'Cumulus provides cloud-based data analysis for large-scale single-cell and single-nucleus RNA-seq', *Nature Methods*, 17(8), pp. 793–798. doi:10.1038/s41592-020-0905-x.

The bioinformatics pipeline: main “modular” components

1. **Feature selection**
2. **Dimensionality Reduction**
3. **Clustering / Hiearchy**
4. **Differential Expression / Gene signature extraction**
5. **Functional interpretation**
6. **Batch effect correction**

Feature extraction (I)

A. Simple filtering criteria, eg:

- Filtering of lowly expressed genes expressed in < x% of cells
- Genes with a mean average of expression < threshold
- Restrict to protein coding genes

B. M3Drop: Dropout-based feature selection for scRNASeq:

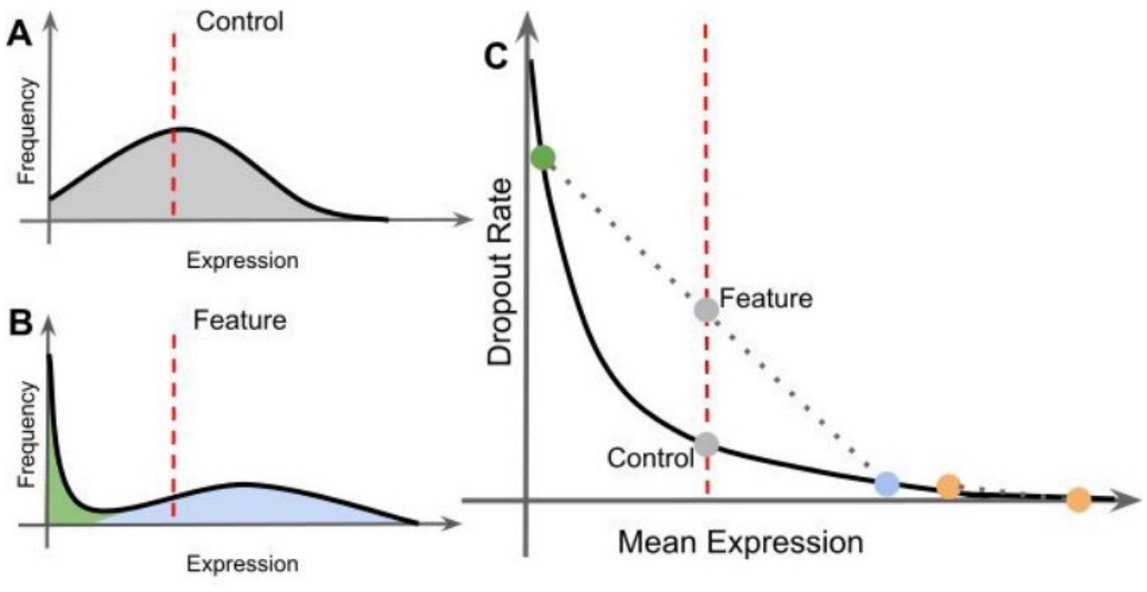


Figure 1: Differentially expressed genes exhibit bimodal expression which increases the dropout rate relative to the mean expression. (**A & B**) Genes with the same mean expression (dashed red line), but (A) is expressed evenly across cells, whereas (B) is highly expressed in some cells (blue) and lowly expressed in others (green). (**C**)This leads to a surplus of dropouts since mean and dropout rate average linearly (dotted line) whereas the expectation (black line) is non-linear. Orange points indicate a gene with very high expression where differential expression leads to only a small increase in dropout-rate.

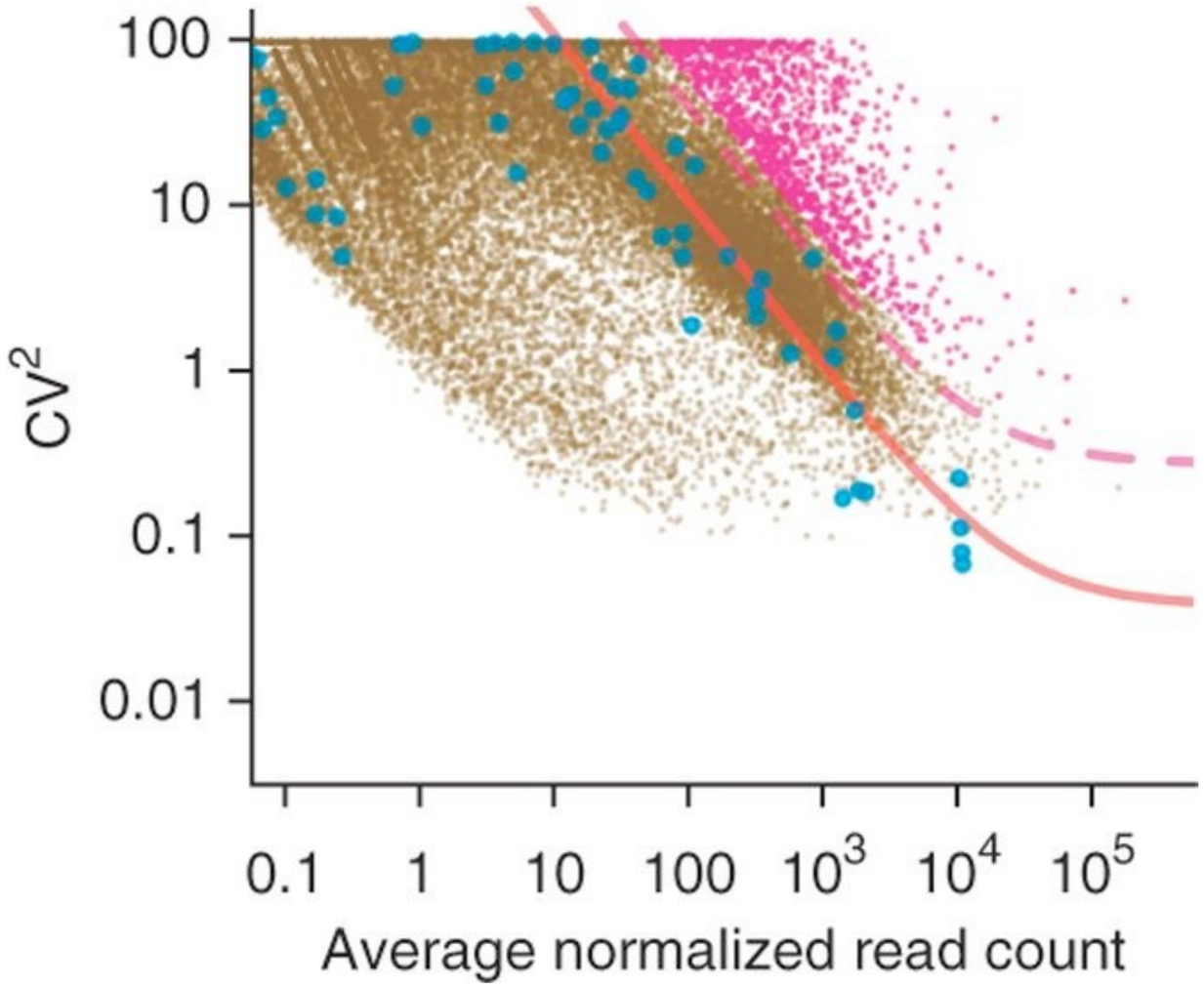
Michaelis-Menten function to the relationship between mean expression (S) and dropout-rate (M3Drop).

$$P_{dropout} = 1 - \frac{S}{K_M + S}$$

Since the Michaelis-Menten function has a single parameter (Km), we can test the hypothesis that the gene-specific Ki is equal to the Km that was fit for the whole transcriptome. This can be done by propagating errors on both observed dropout rate and observed mean expression to estimate the error of each Ki. The significance can then be evaluated using a t-test

Andrews and Hemberg. Bioinformatics (2018)
https://doi.org/10.1093/bioinformatics/bty1044

Feature extraction (II) - Highly Variable Genes (Brennecke et al)



Accounting for technical noise in single-cell RNA-seq experiments.
Brennecke et al. Nature Methods (2013) 10:1093–1095

The coefficient of variation (CV) is defined as the ratio of the standard deviation to the mean

$$c_{\mathrm{v}} = \frac{\sigma}{\mu}.$$

The bioinformatics pipeline: main “modular” components

1. **Feature selection**
2. **Dimensionality Reduction**
3. **Clustering / Hiearchy**
4. **Differential Expression / Gene signature extraction**
5. **Functional interpretation**
6. **Batch effect correction**

Dimensionality reduction - Why?

1. Need of an orthogonal space

2. Minimize curse of dimensionality

3. Filter out noise

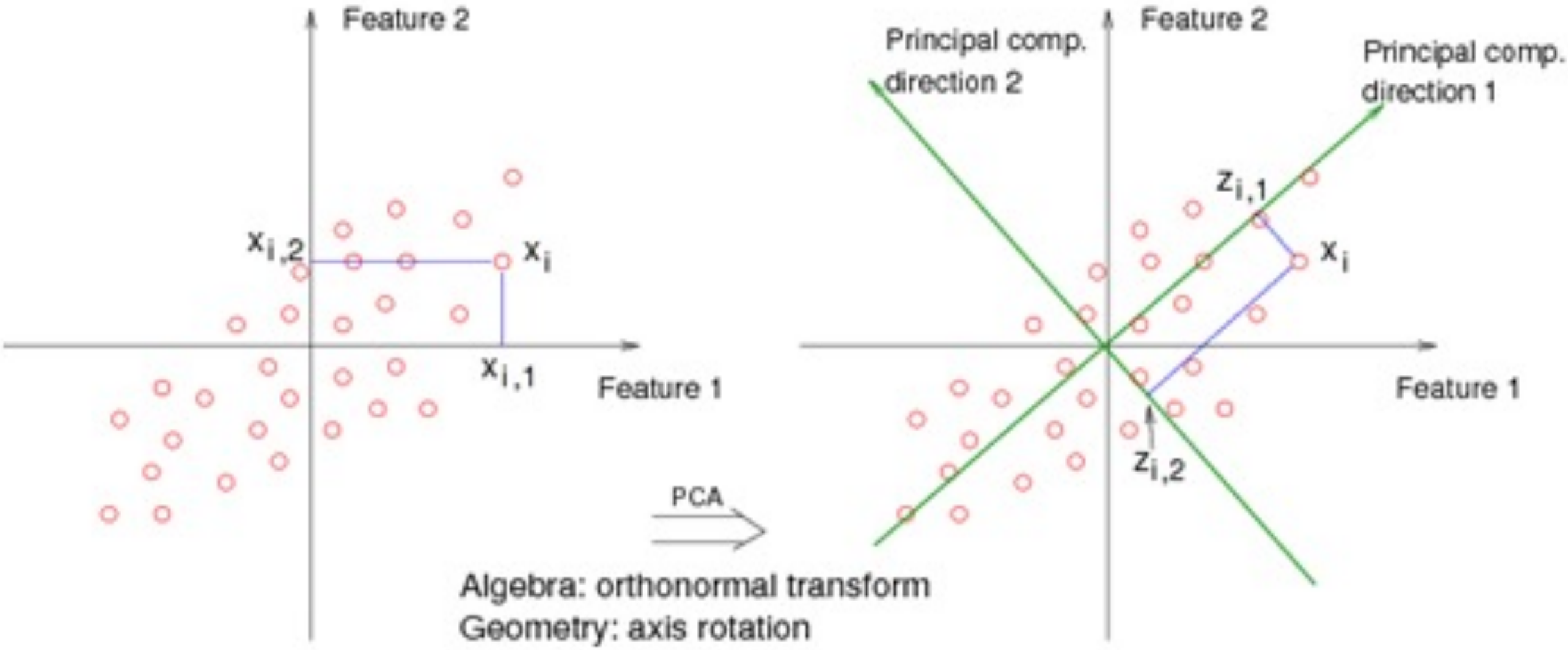
4. Allow visualization

5. Reduce computational load

Popular methods used for single- cell data analysis:

1. PCA
2. tSNE
3. UMAP
4. Others : Diffusion map, Isomap

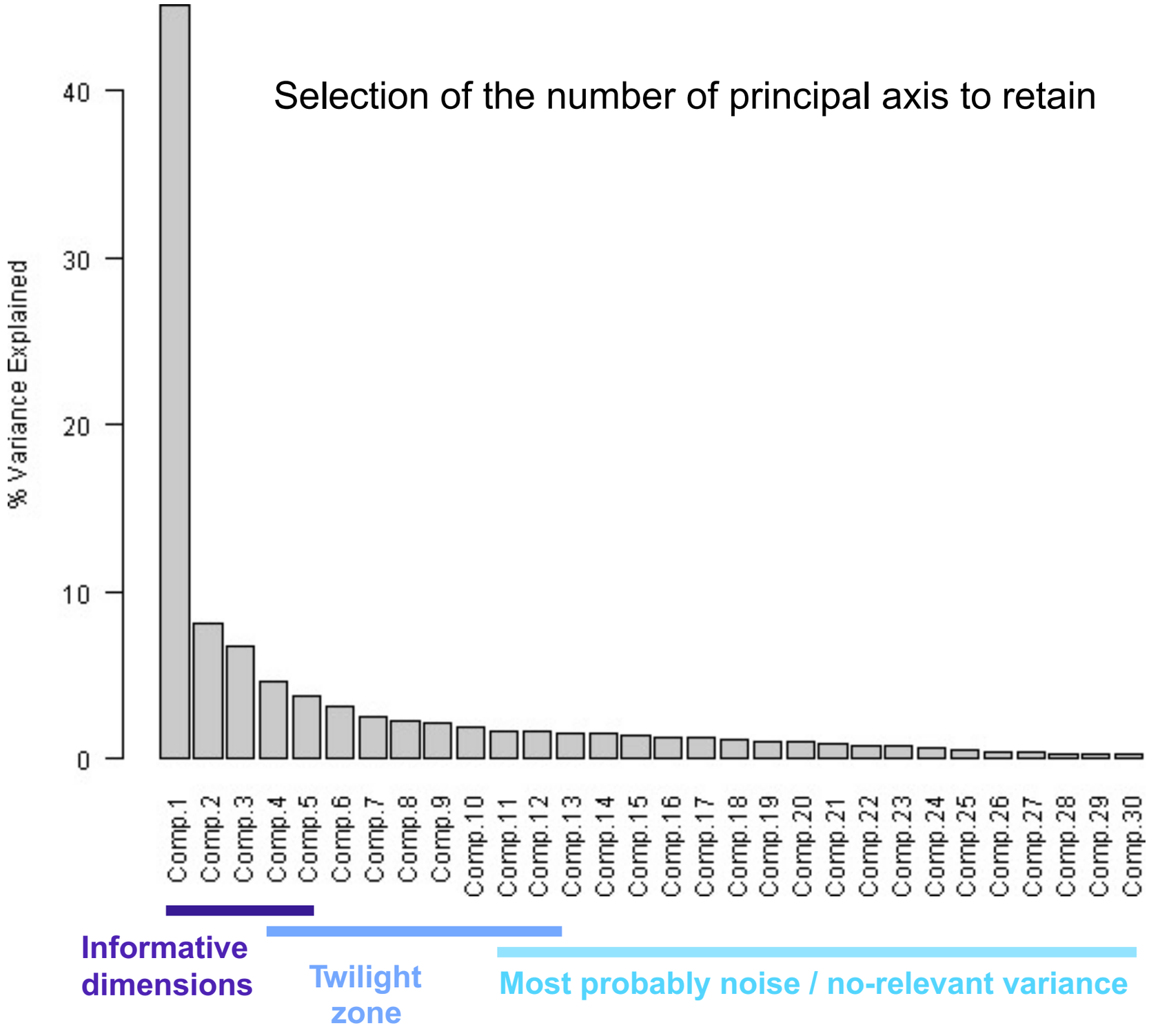
Dimensionality reduction (I) - Principal Component Analysis (PCA) (I)



Source URL: https://onlinecourses.science.psu.edu/stat857/node/35

Dimensionality reduction (I) - Principal Component Analysis (PCA) (II)

Selection of the number of principal axis to retain



https://www.r-bloggers.com/visualizing-principal-components/

Further reading: https://hemberg-lab.github.io/scRNA.seq.course/seurat-chapter.html#significant-pcs

Dimensionality reduction (I) - Principal Component Analysis (PCA) (III)

PCA Advantages and limitations

- Based on linear transformations
- Captures the dimensions with higher variance
- Objective control on the amount of retained dimensions
- Fast & scalable
- Preserves both long-range and short-range relationships

Extensions of the PCA approach

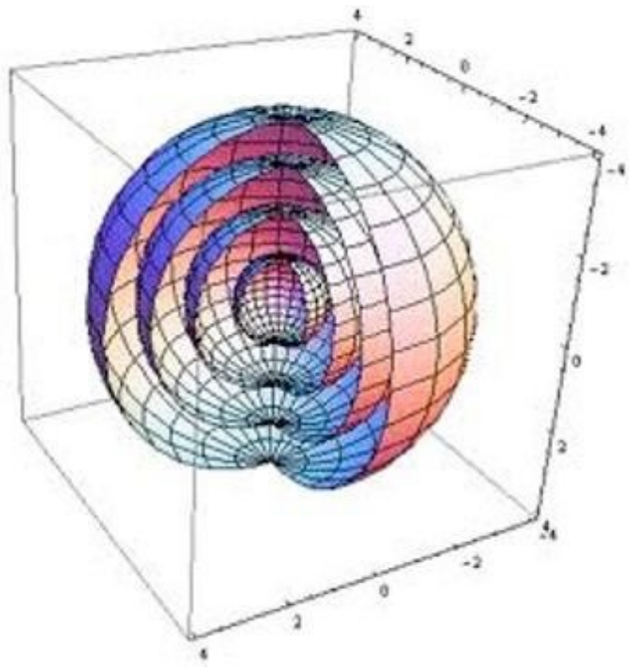
- A variation of PCA which explicitly deals with the large number of zero-values in scRNASeq data has been developed (ZIFA, Pierson and Yau, 2015) but the zero-inflation model employed may not fit all datasets (Andrews and Hemberg, 2016).

- Risso et al. (2017) proposed a method similar to PCA based on a zero-inflated negative binomial model instead of a Gaussian model.

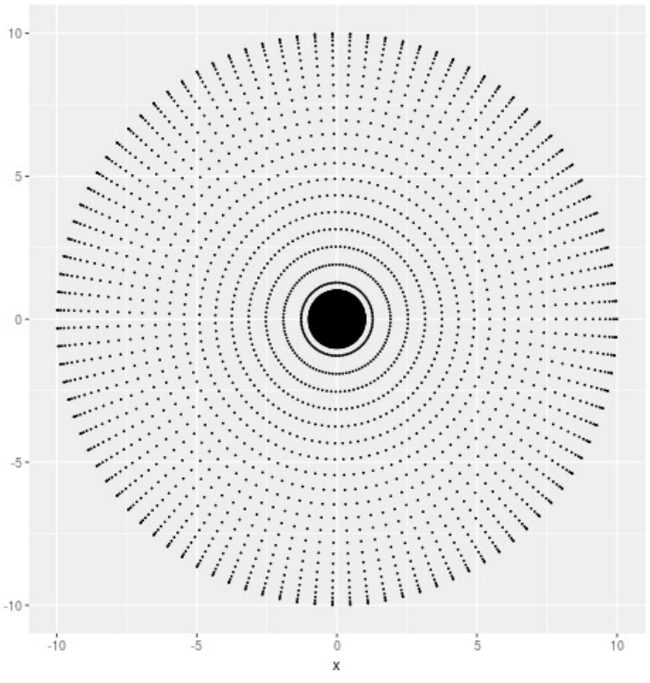
Dimensionality reduction (II) - tSNE
t-distributed stochastic neighbor embedding

tSNE is a non-linear dimension reduction technique able to show structures in the data that cannot be found simply by changing the direction in which you look

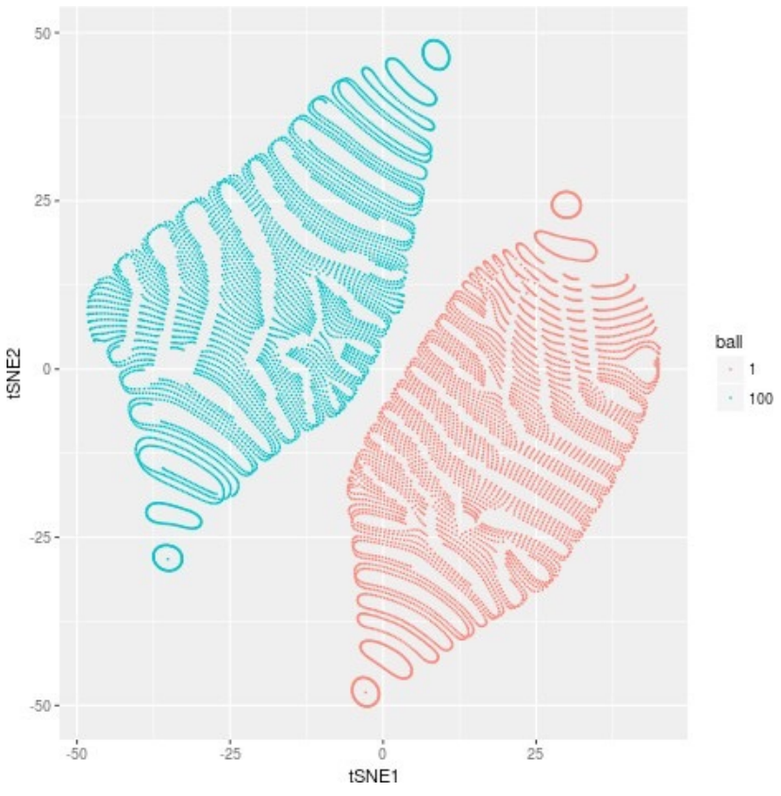
Original



PCA



tSNE



tSNE: What the hell is it?, by Matthew Young
https://constantamateur.github.io/2018-01-02-tSNE/

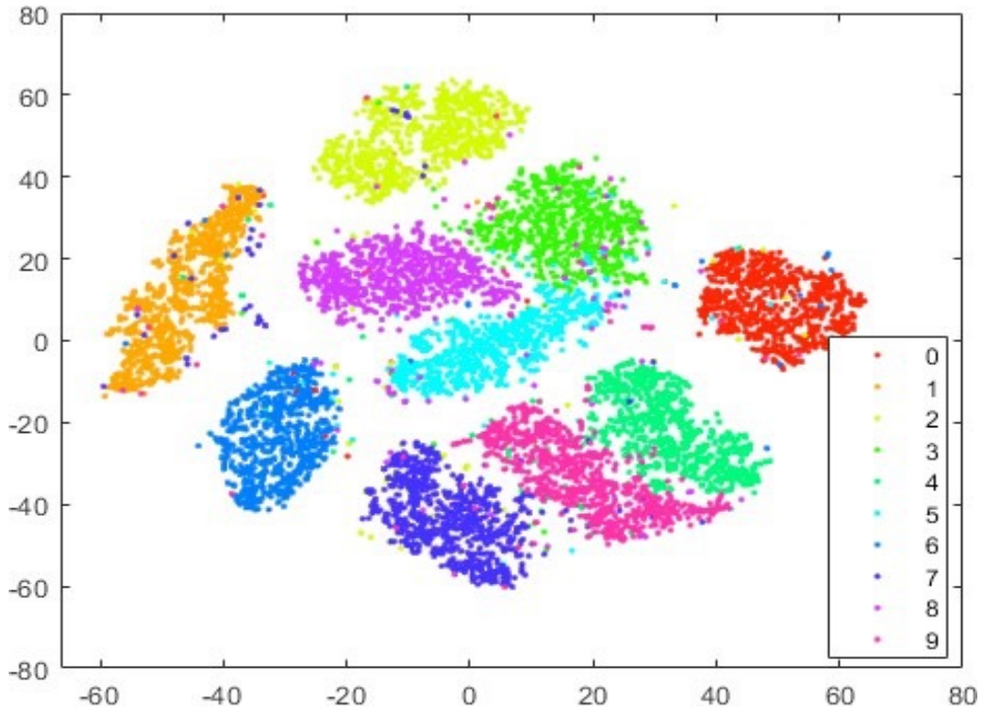
Dimensionality reduction (II) - tSNE

The Art of Using T-SNE for Single-Cell Transcriptomics
https://doi.org/10.1038/s41467-019-13056-x.

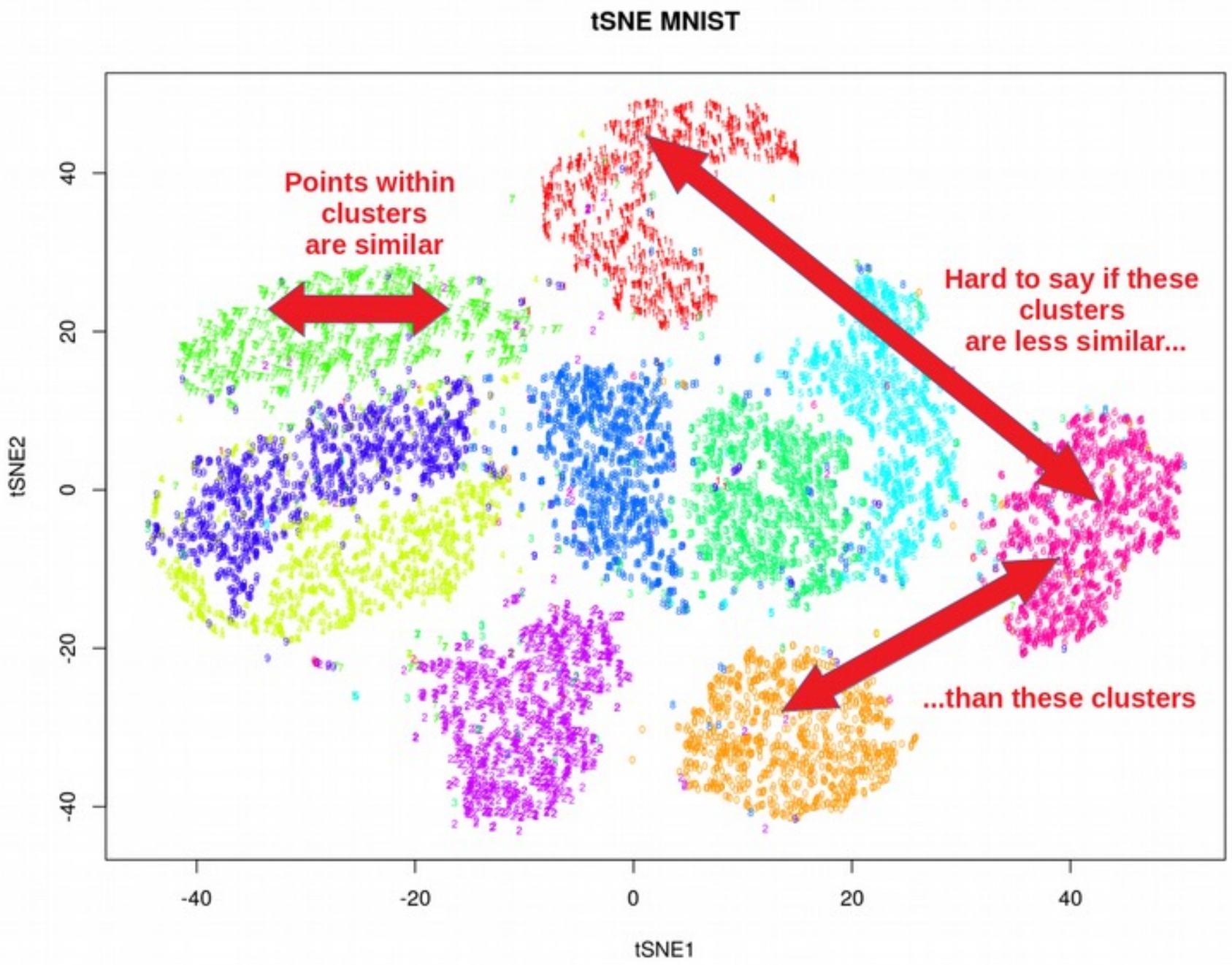
tSNE: What the hell is it?, by Matthew Young
https://constantamateur.github.io/2018-01-02-tSNE/

How to Use t-SNE Effectively, Wattenberg, et al. Distill, 2016
https://distill.pub/2016/misread-tsne/

t-SNE in wikipedia
https://en.wikipedia.org/wiki/T-distributed_stochastic_neighbor_embedding



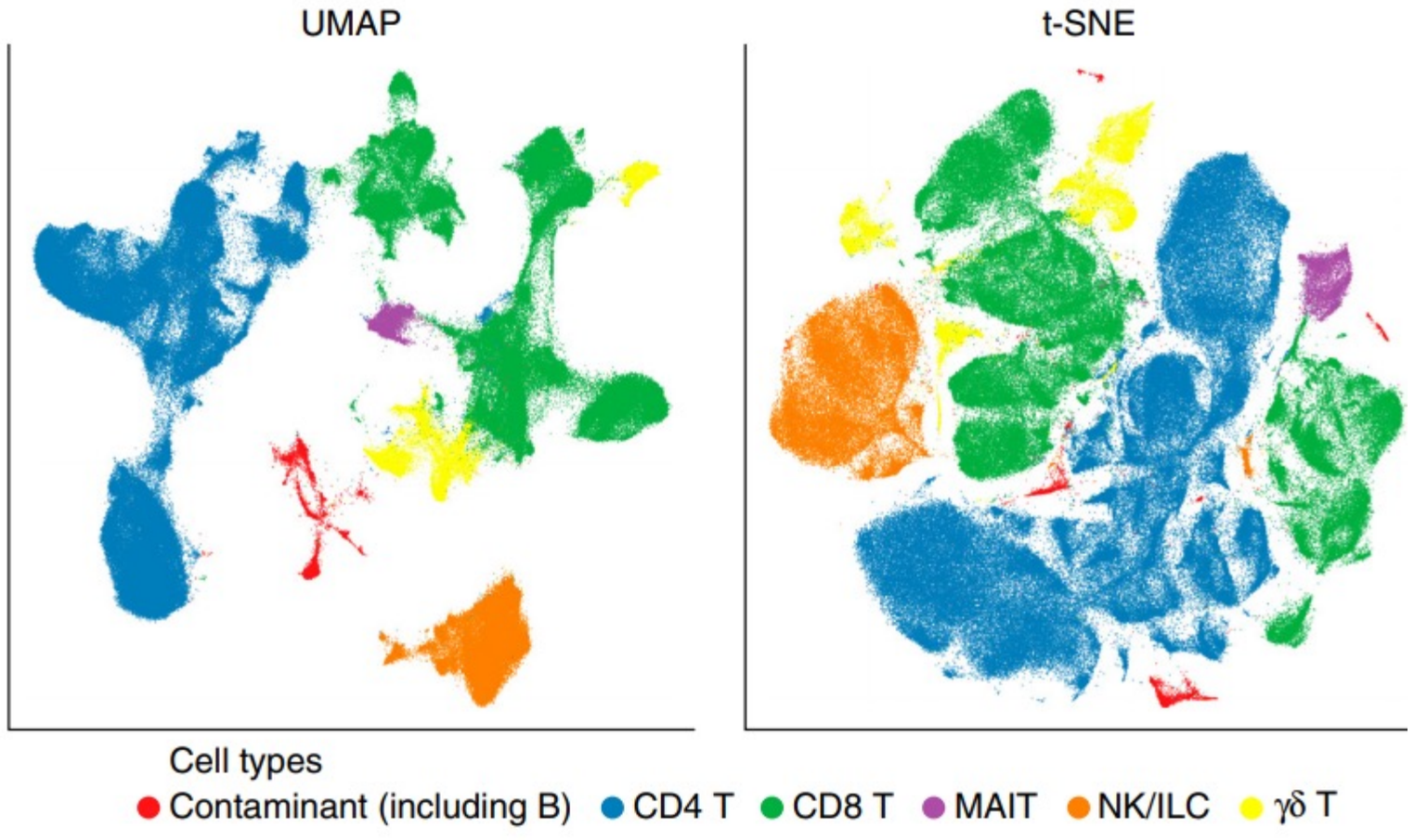
Dimensionality reduction (II) - tSNE



tSNE doesn't preserve global data structure…

Dimensionality reduction (III) - Uniform Manifold Approximation and Projection (UMAP)

Uniform manifold approximation and projection (UMAP):
- Claimed to preserve as much the local and global data structure than t-SNE
- Shorter run time.



Dimensionality reduction for visualizing single-cell data using UMAP
Becht et al. Nature Biotechnology 2018. http://www.nature.com/doifinder/10.1038/nbt.4314

Dimensionality reduction (III) - Uniform Manifold Approximation and Projection (UMAP)

Dmitry Kobak @hippopedoid · 13 févr. 2020

Becht et al.: UMAP preserves global structure better than t-SNE.

@GCLinderman & me: only because you used random init for t-SNE but spectral init for UMAP.

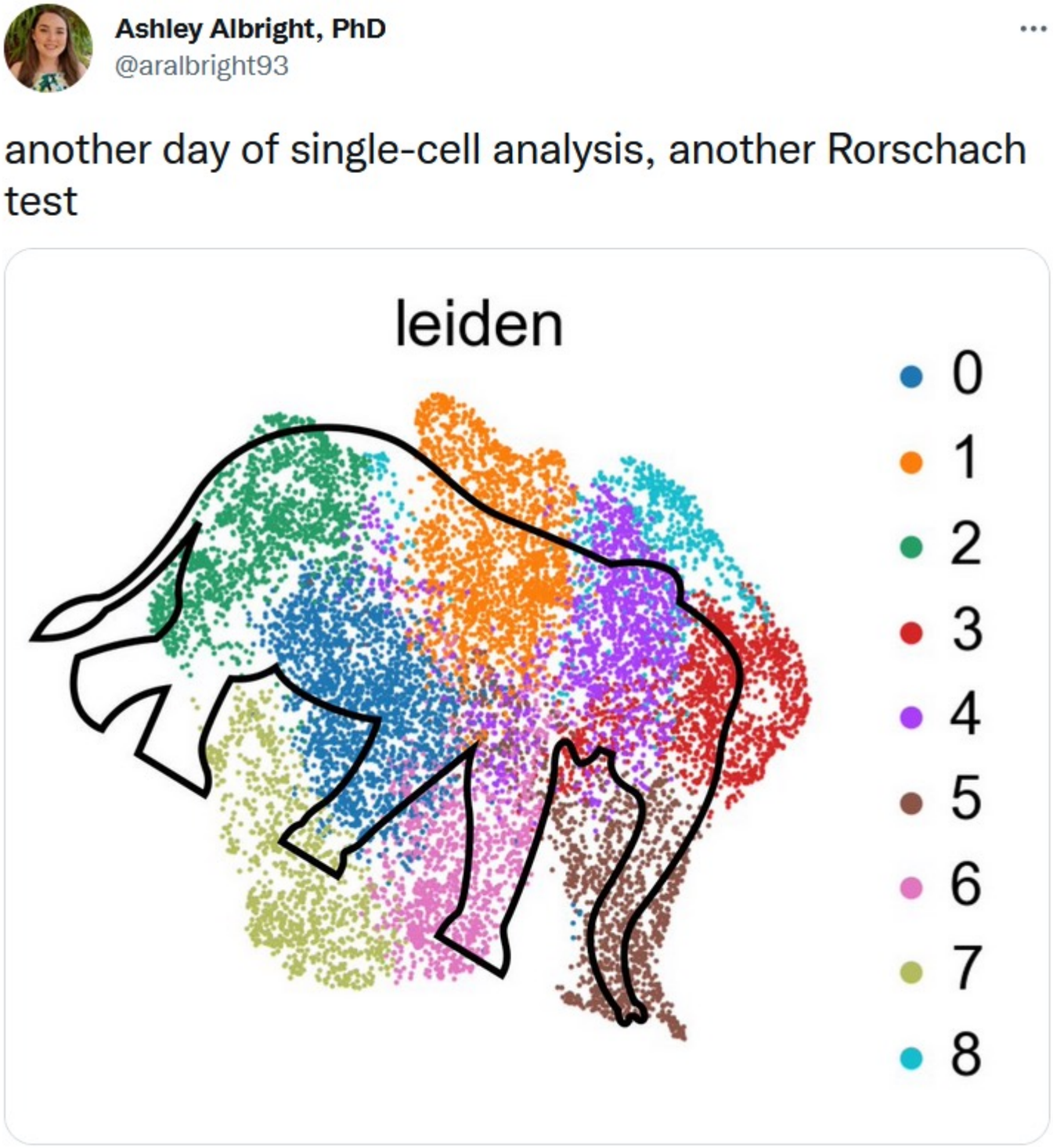
@NikolayOskolkov: that's wrong; init does not matter; the loss function does.

	Preservation of pairwise distances			Reproducibility of large-scale structures		
	Samusik	Wong	Han	Samusik	Wong	Han
UMAP, LE init.	**0.70**	0.58	**0.31**	0.94	**0.98**	0.49
UMAP, random init.	0.40	0.39	0.14	0.24	0.21	0.22
t-SNE, PCA init.	0.60	**0.66**	0.29	**0.95**	**0.98**	**0.92**
t-SNE, random init.	0.32	0.37	0.18	0.29	0.33	0.06

Table 1: Performance of t-SNE and UMAP with random and informative initializations, using data sets and evaluation metrics from Becht et al. For the reproducibility metric, the average over three random subsamples of size $n = 200\,000$ is reported.

Kobak, D. and Linderman, G.C. (2021) 'Initialization is critical for preserving global data structure in both t-SNE and UMAP', *Nature Biotechnology*, 39(2), pp. 156–157. doi:10.1038/s41587-020-00809-z.

Dimensionality reduction (III) - Uniform Manifold Approximation and Projection (UMAP)



- **Can you really blindly trust what you are seeing on these embeddings?**
- **Can it be used for downstream analysis such as clustering?**

Dimensionality reduction (III) - Uniform Manifold Approximation and Projection (UMAP)

Lior Pachter ✓
@lpachter

It's time to stop making t-SNE & UMAP plots. In a new preprint w/ Tara Chari we show that while they display some correlation with the underlying high-dimension data, they don't preserve local or global structure & are misleading. They're also arbitrary. 🧵 biorxiv.org /content/10.110...

Lior Pachter ✓ @lpachter · 28 août 2021

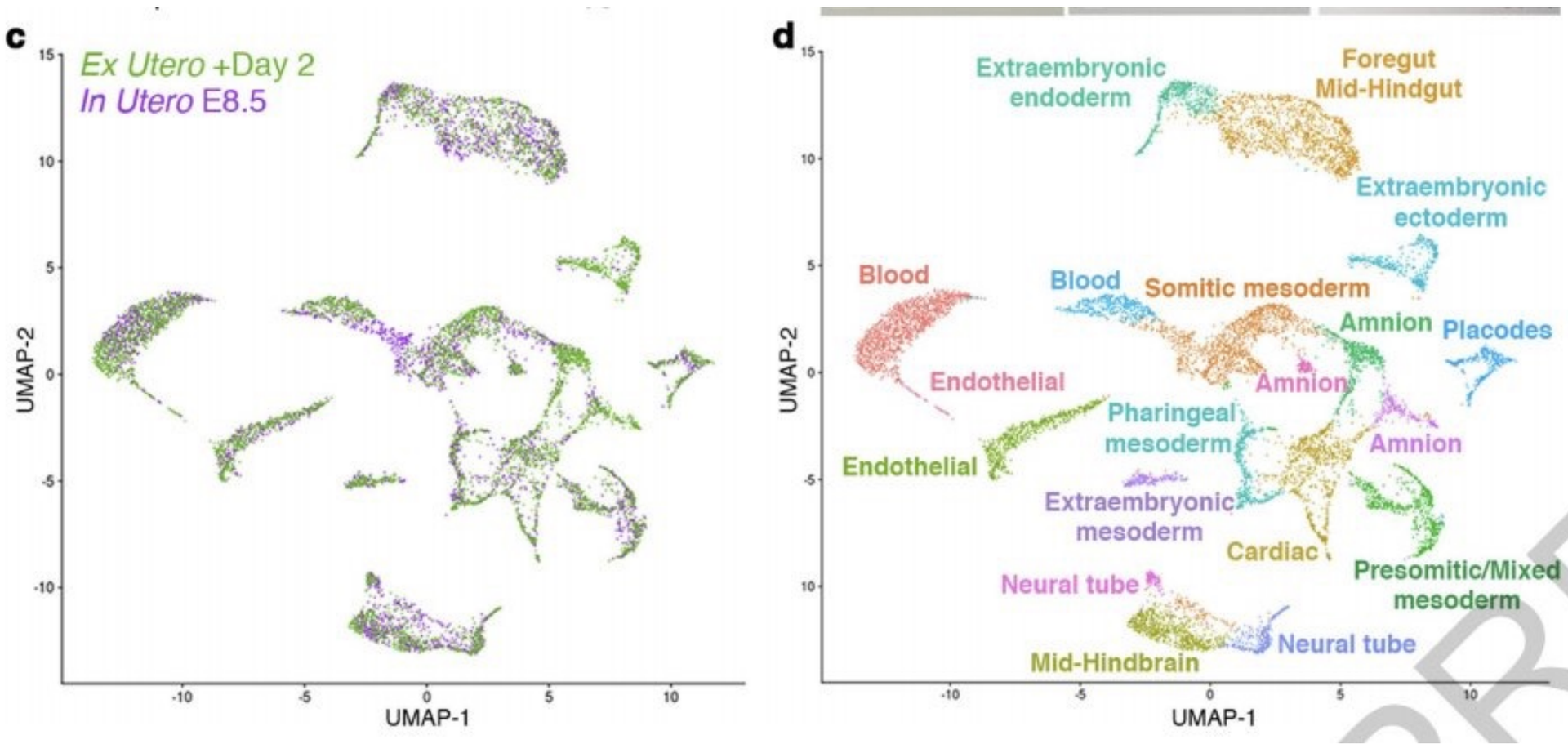
Perhaps it's time for everyone to say out loud what we've all known for some time, but have had difficulty admitting: t-SNE, UMAP and relatives are just specious art and we risk fooling ourselves when we start to believe in the mirages they present.

21 75 282

Dimensionality reduction (III) - Uniform Manifold Approximation and Projection (UMAP)

Rahul Satija
@satijalab

Remarkable overlap between scRNA-seq data of in utero and ex utero (!) mouse embryogenesis

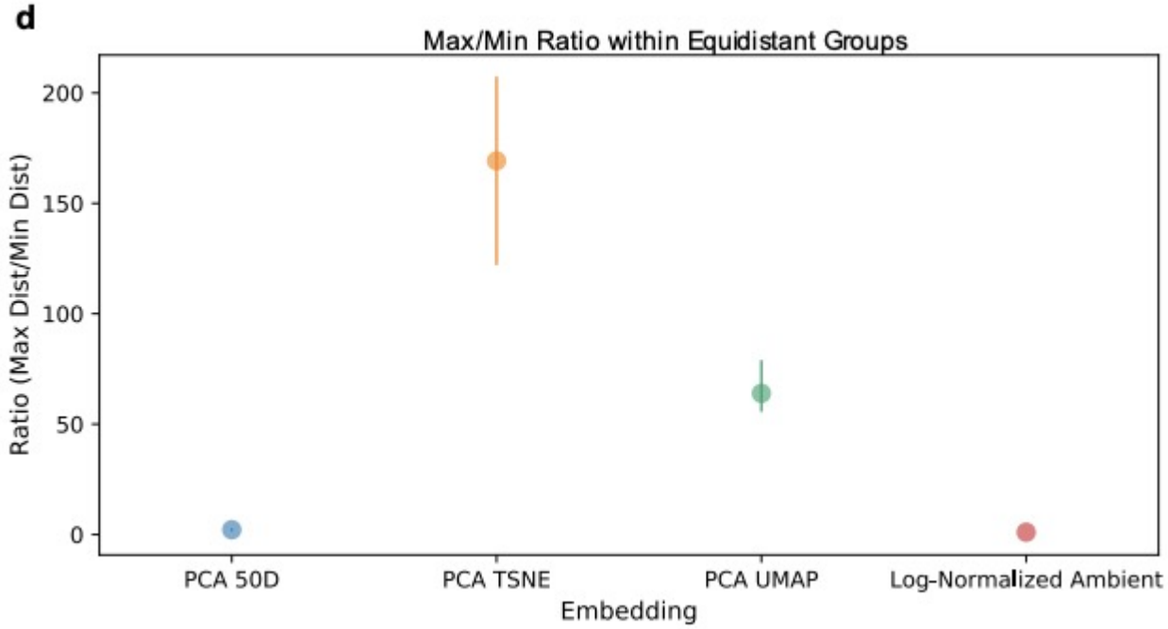


https://twitter.com/satijalab/status/1372243222223806468?s=20

Dimensionality reduction (III) - Uniform Manifold Approximation and Projection (UMAP)

Lior Pachter @lpachter · 28 août 2021

These distortions are inevitable. Cells or cell types that are equidistant in high dimension must exhibit increasing distortion as they increase in number. Actually, UMAP and t-SNE distortions are even worse (much worse!) than the lower bounds from theory.



- **UMAP and tSNE distort the similarity/distance of the cells drastically compared to PCA.**
- **One should not use these embeddings for downstream analysis.**

https://twitter.com/lpachter/status/1431326001414299650

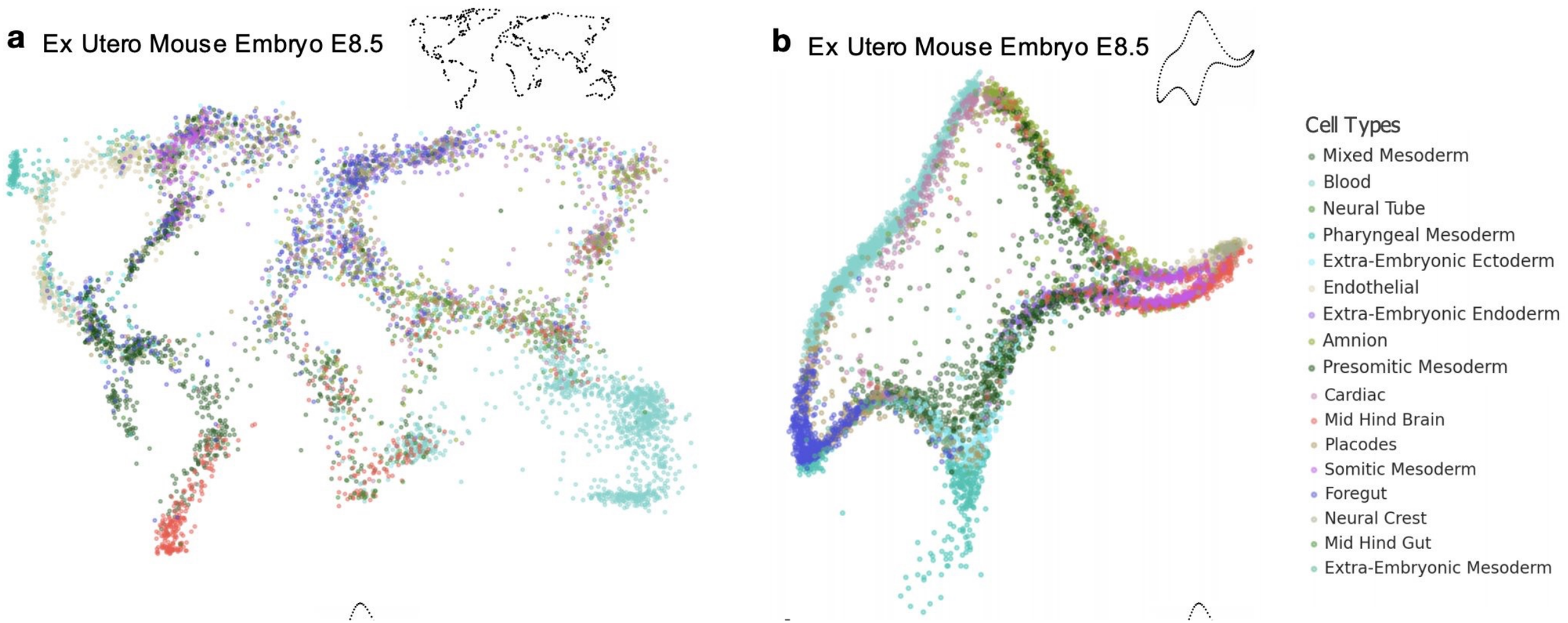
Dimensionality reduction (III) - Uniform Manifold Approximation and Projection (UMAP)

Lior Pachter @lpachter · 28 août 2021

Ok.. but.. maybe t-SNE & UMAP (or your favorite 2D viz) aren't perfect, but they are "canonical" and not arbitrary. Nope. They're just art. We developed Picasso for embedding your data into any shape, with less distortion than t-SNE & UMAP (see elephant at the start of the🧵)

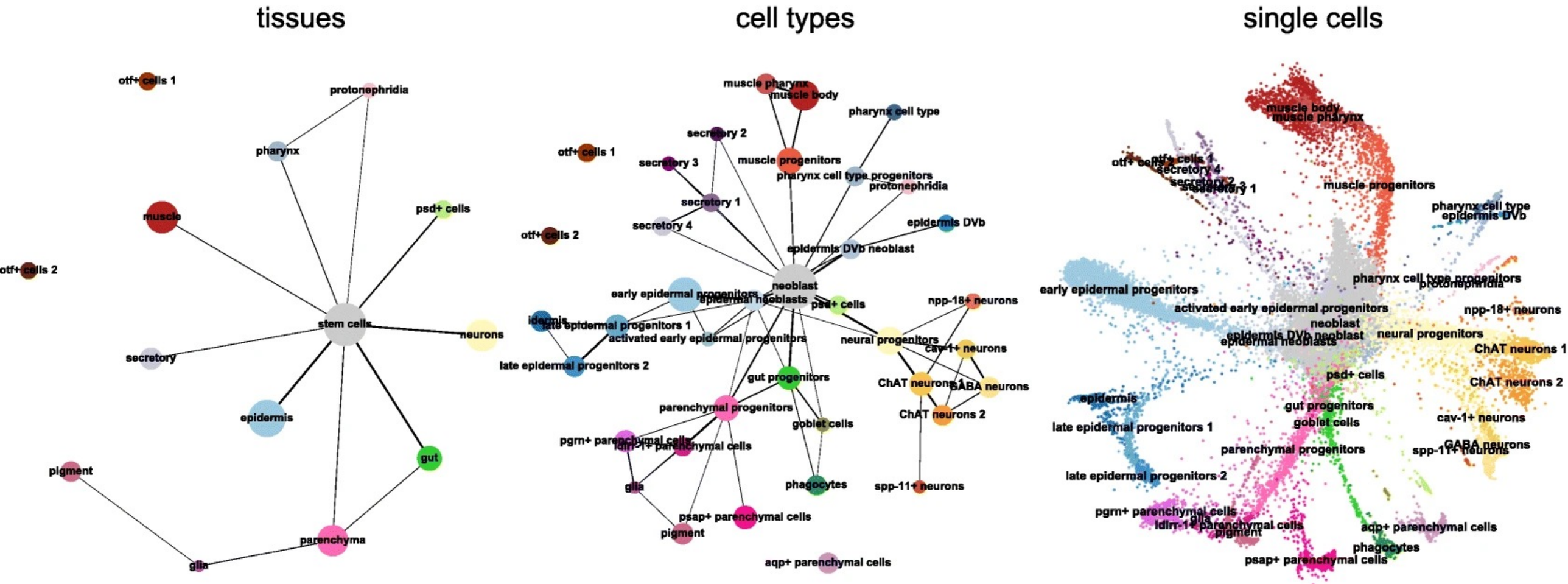
Lior Pachter @lpachter · 28 août 2021

Picasso is available via @GoogleColab so you can experiment with turning any dataset you like into any shape you want...while producing a better representation of the data than w/ t-SNE or UMAP. No more Rorschach testing necessary! Make your own elephant!



Dimensionality reduction (IV) - PAGA

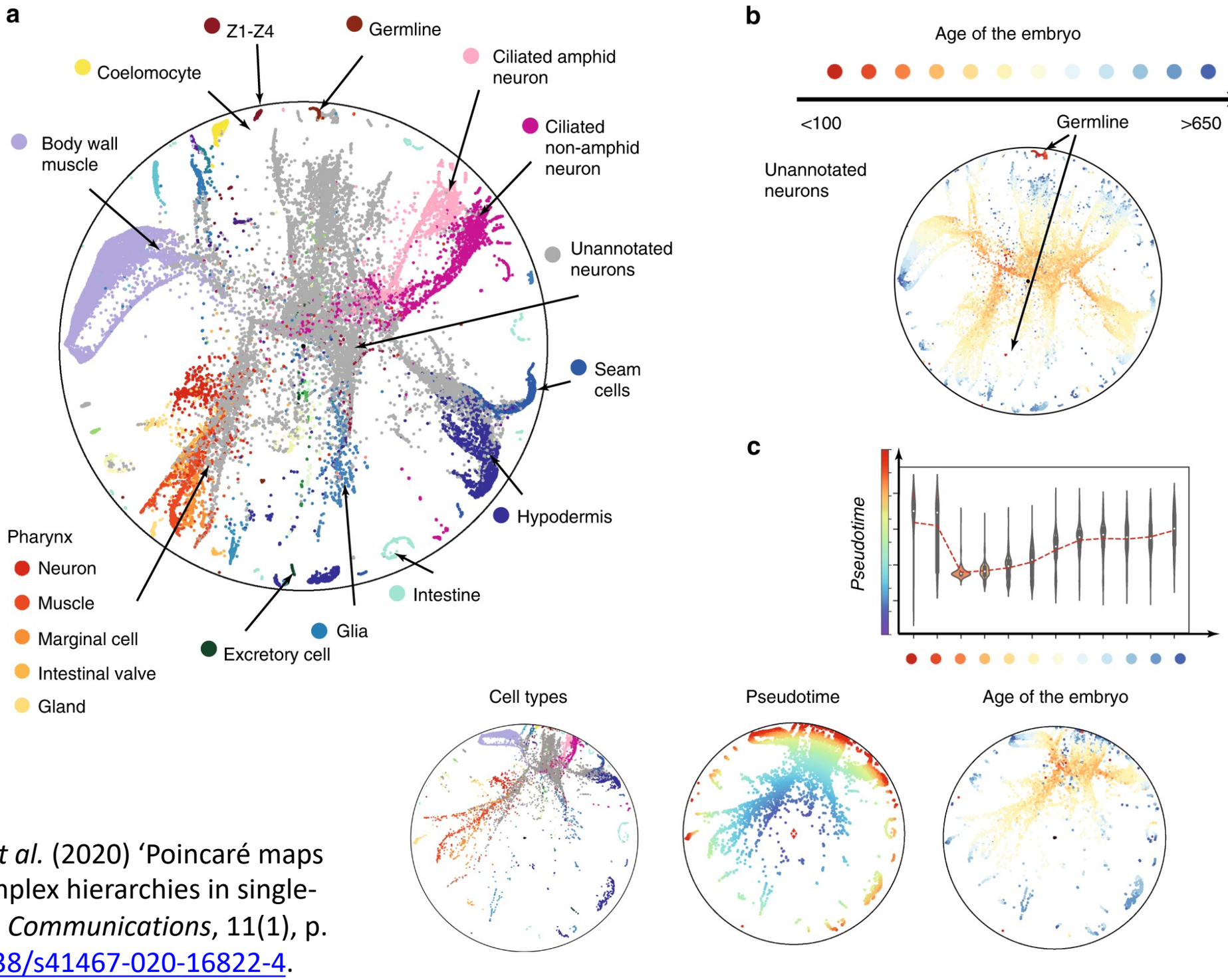
PAGA graphs for data for the flatworm *Schmidtea mediterranea*



Wolf, F. A. *et al.* (2019) 'PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells', *Genome Biology*, 20(1), p. 59. doi: 10.1186/s13059-019-1663-x.

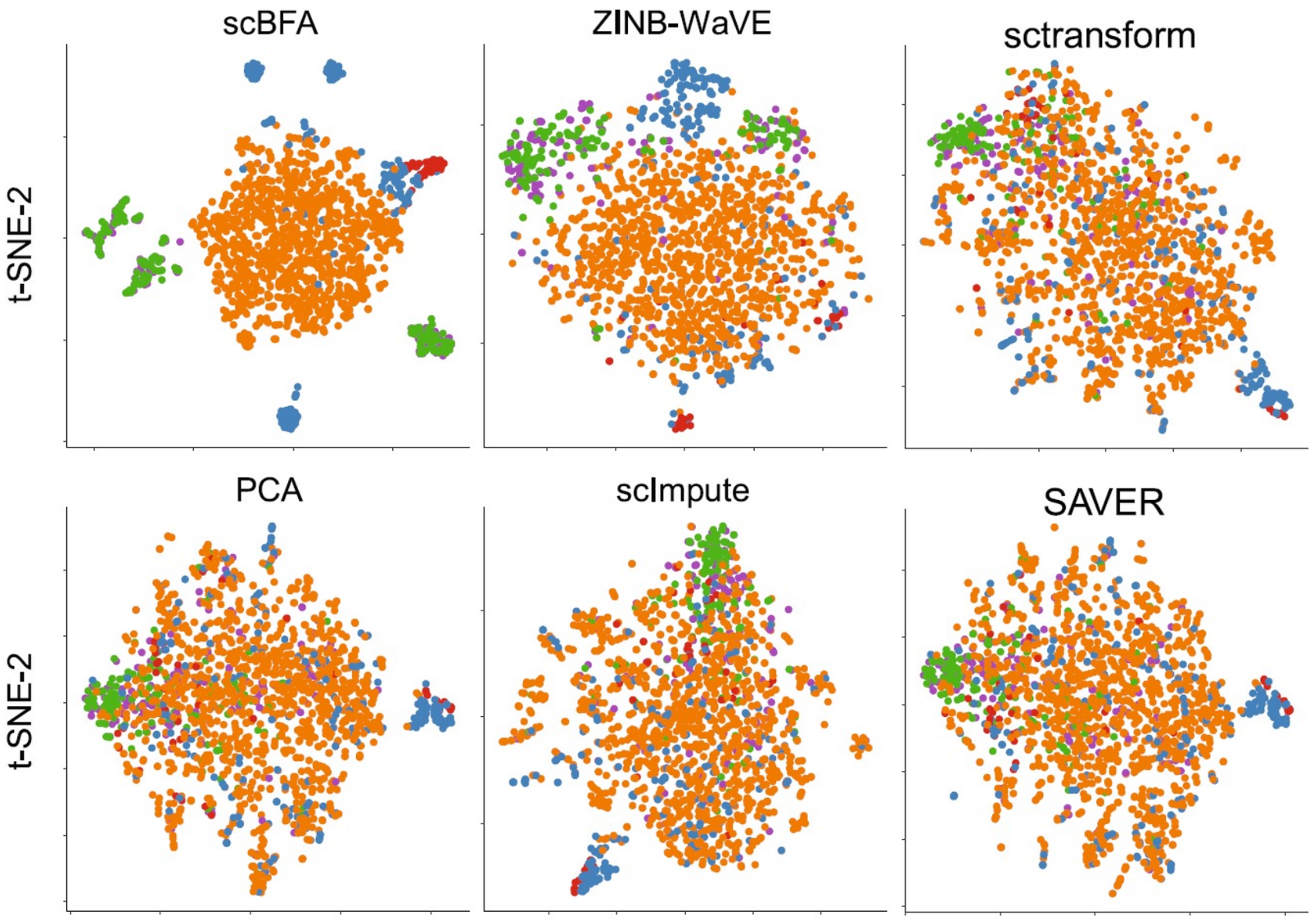
Dimensionality reduction (V) - Poincarémap

Analysis of *C. elegans* cell atlas with poincaré map



Klimovskaia, A. *et al.* (2020) 'Poincaré maps for analyzing complex hierarchies in single-cell data', *Nature Communications*, 11(1), p. 2966. doi: 10.1038/s41467-020-16822-4.

Dimensionality reduction (V) scBFA

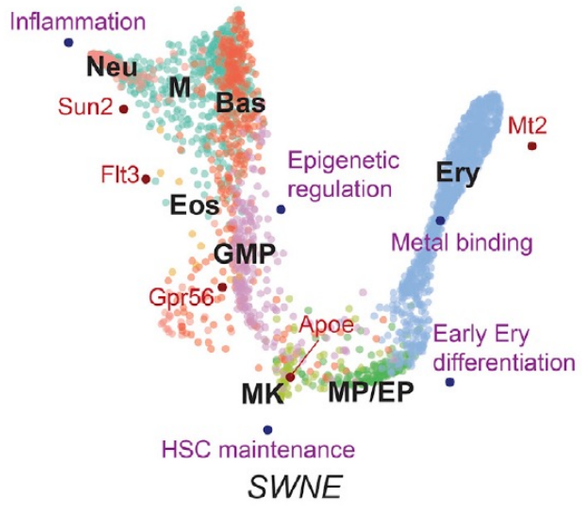


- **Based on gene detection pattern and not count.**
- **Especially useful for highly noisy data with low amount of UMI.**

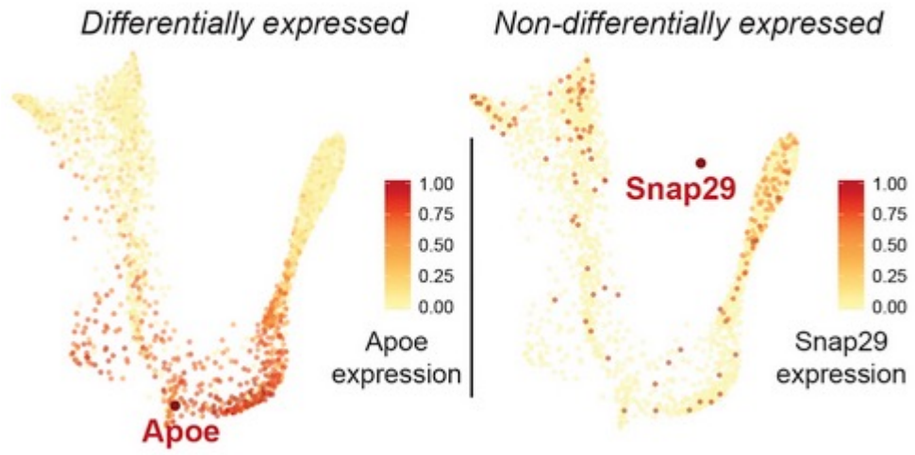
Li, R. and Quon, G. (2019) 'scBFA: modeling detection patterns to mitigate technical noise in large-scale single-cell genomics data', *Genome Biology*, 20(1), p. 193. doi:10.1186/s13059-019-1806-0.

Dimensionality reduction (VI) - SWNE/SIMBA embedding of both cells and genes

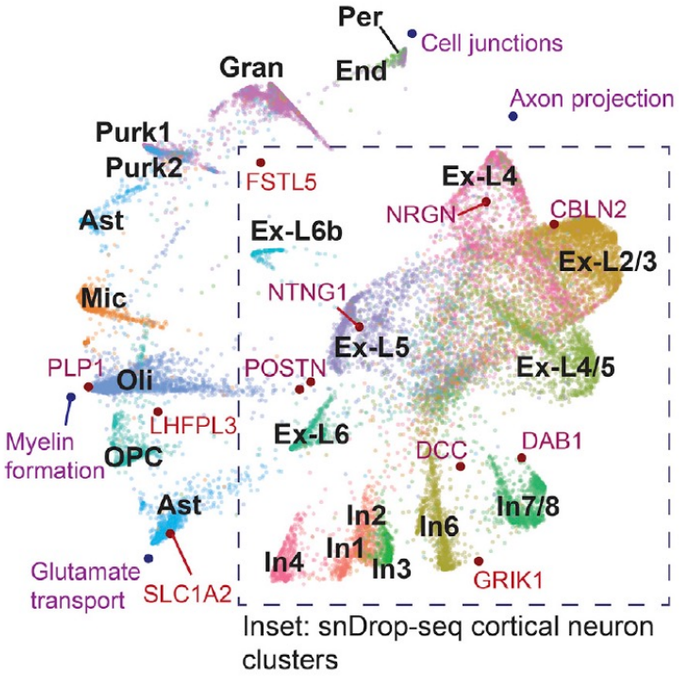
Hematopoiesis cell type visualizations



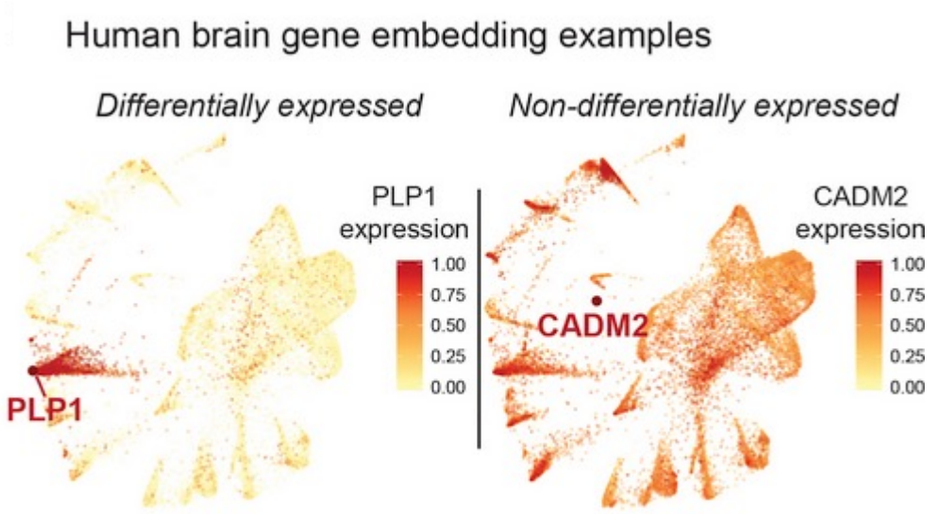
Hematopoiesis gene embedding examples



Human brain cell type visualizations



Human brain gene embedding examples



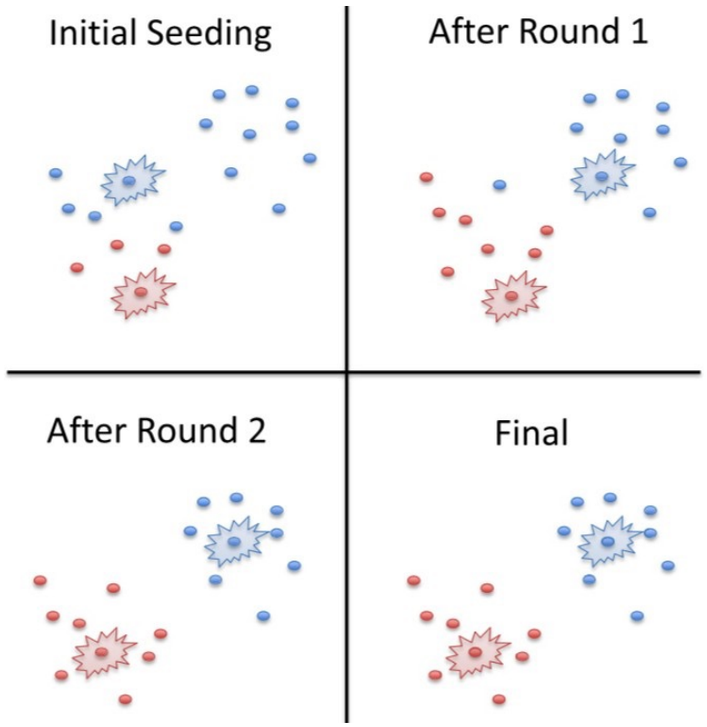
Wu, Y., Tamayo, P. and Zhang, K. (2018) 'Visualizing and Interpreting Single-Cell Gene Expression Datasets with Similarity Weighted Nonnegative Embedding', *Cell Systems*, 7(6), pp. 656-666.e4. doi:10.1016/j.cels.2018.10.015.

The bioinformatics pipeline: main “modular” components

1. **Feature selection**
2. **Dimensionality Reduction**
3. **Clustering / Hiearchy**
4. **Differential Expression / Gene signature extraction**
5. **Functional interpretation**
6. **Batch effect correction**

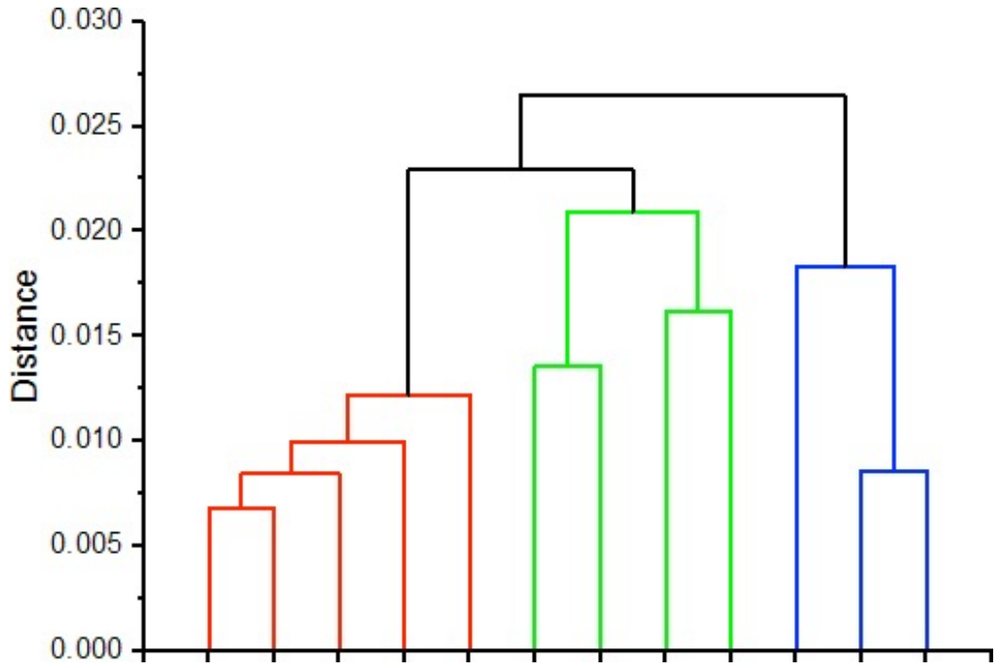
Unsupervised clustering: broad method catergories borrowed for scRNA-seq data analysis

1) K-means based

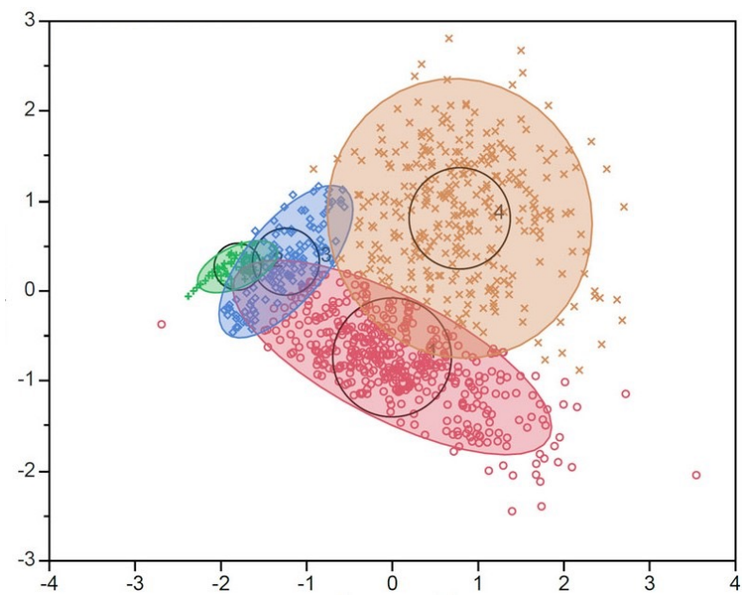


Page et al. BMC Research Notes 2014, 7:829
http://www.biomedcentral.com/1756-0500/7/829

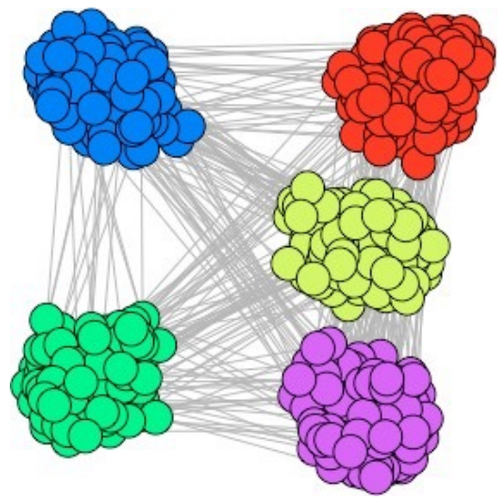
2) Hierarchical clustering



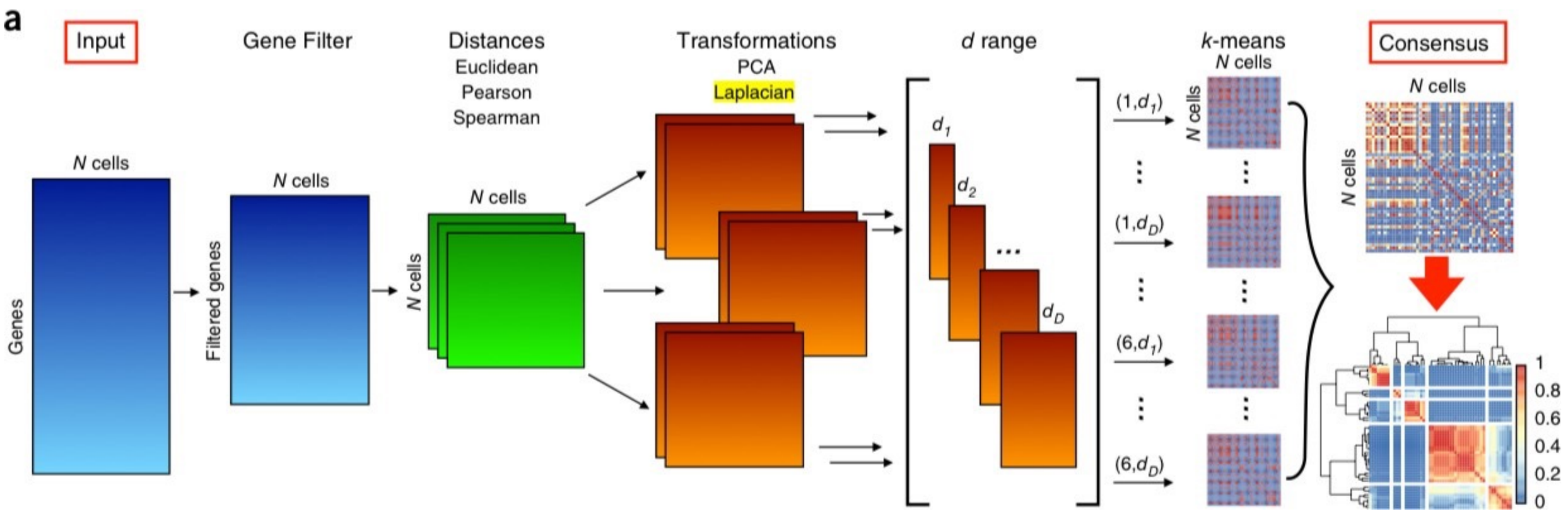
3) Model-based clustering (Mclust)



4) Graph-based clustering (iGraph)



Unsupervised clustering. Examples of dedicated methods for scRNA-seq (I): SC3



SC3: consensus clustering of single-cell RNA-seq data. Kiselev et al. Nature Methods 2017, 14:483–486

Unsupervised clustering. Other dedicated methods for scRNA-seq (III)

Method	Version	Class of clustering technique	Publication
ascend	v0.9.0	Hierarchical	Senabouth et al. (2019)
CIDR	v0.1.5	Hierarchical	Lin et al. (2017)
DIMMSC	v0.2.1	Model-based	Sun et al. (2018)
Linnorm	v2.6.1	Partitioning	Yip et al. (2017)
monocle3	v2.99.2	Multiple choices	Qiu et al. (2017)
pcaReduce	v1.0	Hierarchical	Zurauskiene and Yau (2016)
RaceID3	v0.1.3	Multiple choices	Herman and Grün (2018)
SC3	v1.10.1	Partitioning	Kiselev et al. (2017)
Seurat	v2.3.4	Graph-based	Macosko et al. (2015)
SIMLR	v1.8.1	Partitioning	Wang et al. (2017)
sincell	v1.14.1	Multiple choices	Julia et al. (2015)
sscClust	v0.1.0	Multiple choices	Ren et al. (2019)
TSCAN	v1.20.0	Model-based	Ji and Ji (2016)

Krzak, M. *et al.* (2019) 'Benchmark and Parameter Sensitivity Analysis of Single-Cell RNA Sequencing Clustering Methods', *Frontiers in Genetics*, 10, p. 1253. doi:10.3389/fgene.2019.01253.

Benchmark of (some) clustering methods (I)

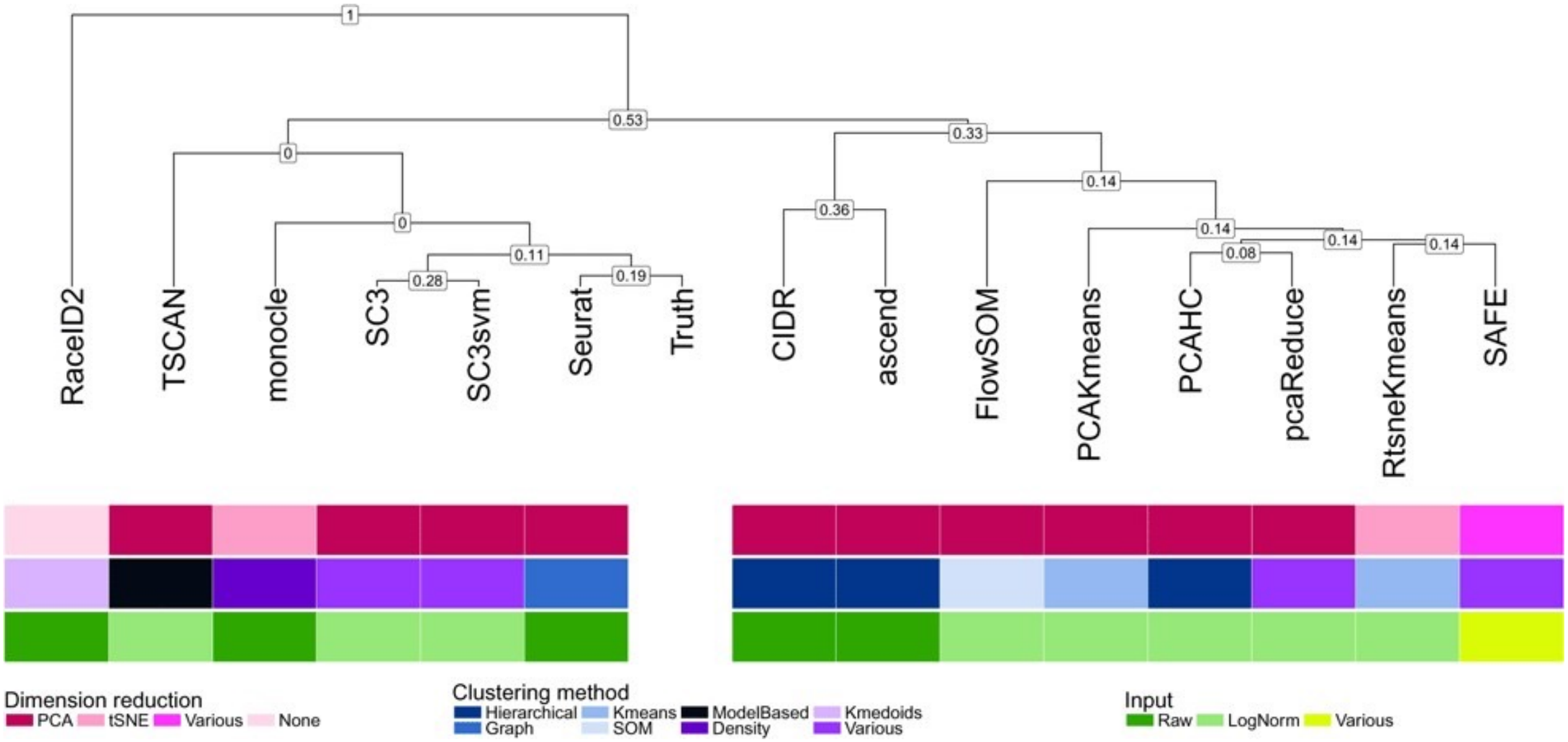


Figure 4. Clustering of the methods based on the average similarity of their partitions across data sets

Duò A, Robinson MD and Soneson C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data [version 2]. F1000Research 2018, 7:1141 (doi: 10.12688/f1000research.15666.2)

Benchmark of (some) clustering methods (II)

... when forcing the methods to cluster with the right number of groups as truth...

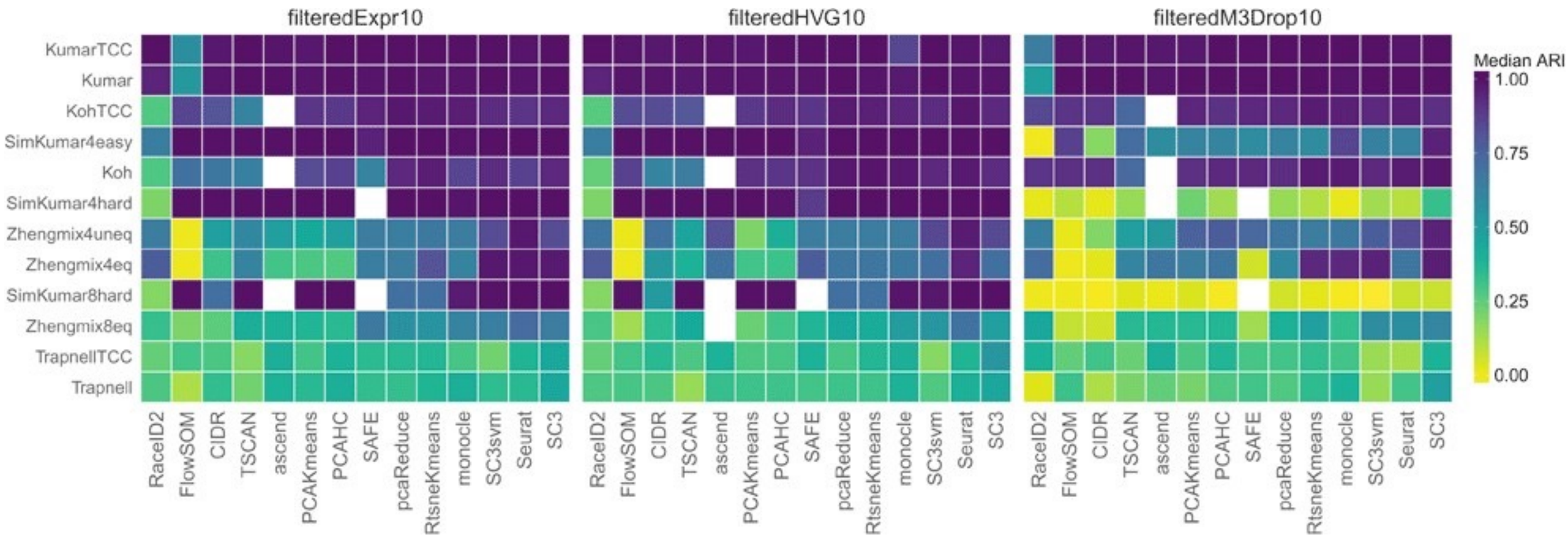
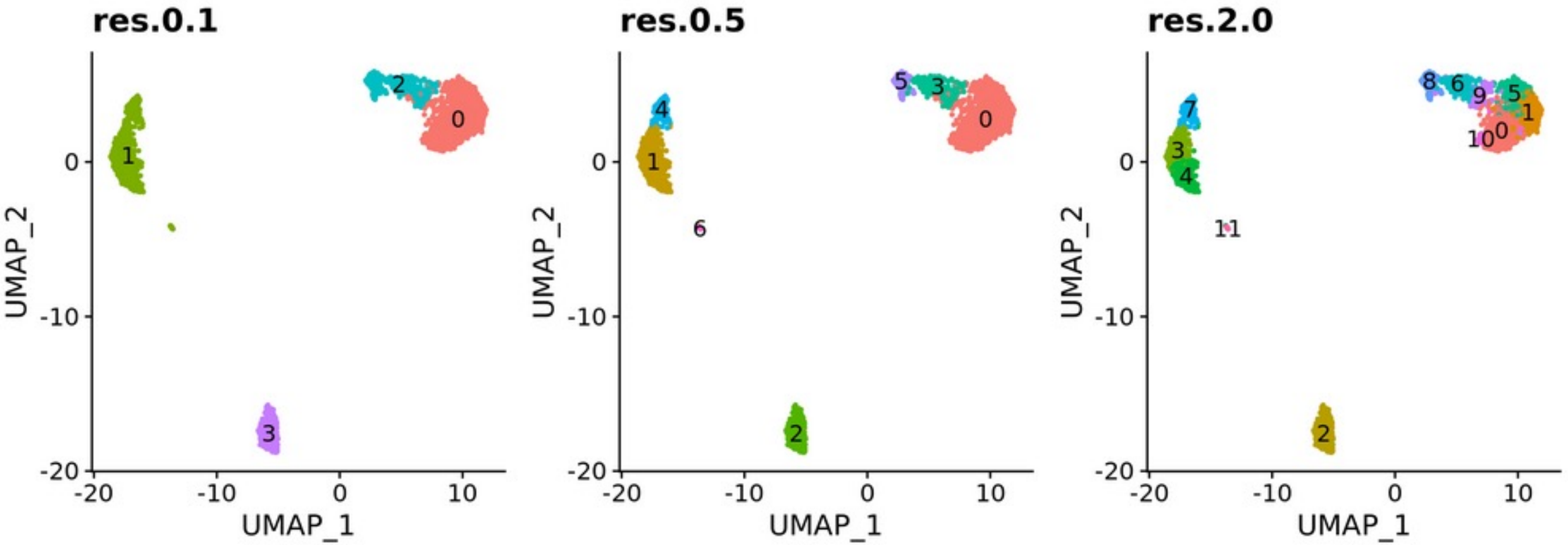


Figure 1. Median ARI scores, representing the agreement between the true partition and the one obtained by each method, when the number of clusters is fixed to the true number.

Each row corresponds to a different data set, each panel to a different gene filtering method, and each column to a different clustering method. The methods and the data sets are ordered by their mean ARI across the filterings and data sets. Some methods failed to return a clustering with the correct number of clusters for certain data sets (indicated by white squares).

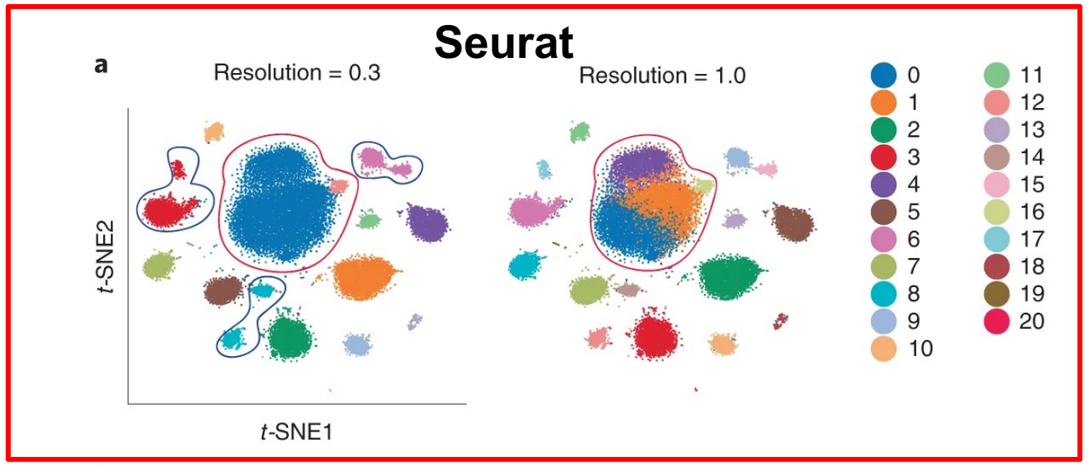
Duò A, Robinson MD and Soneson C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data [version 2]. F1000Research 2018, 7:1141 (doi: 10.12688/f1000research.15666.2)

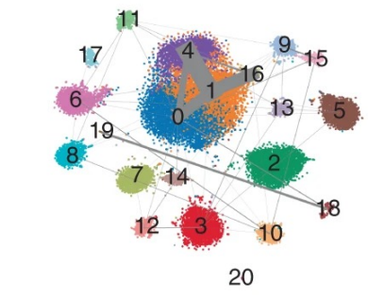
How many clusters?

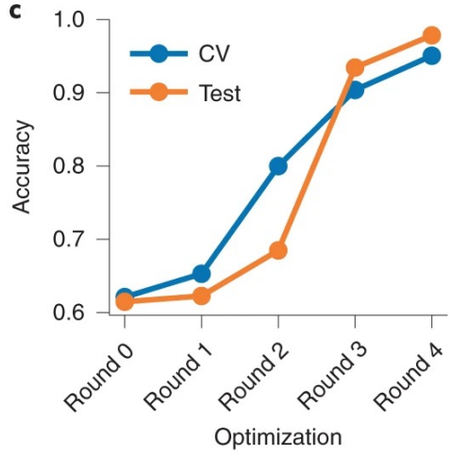


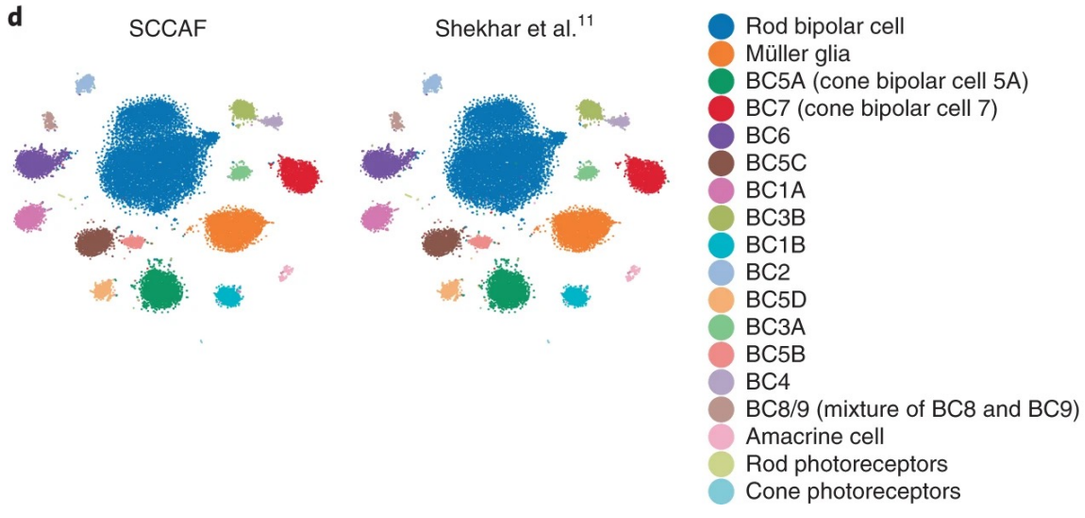
- **The number of cluster depends on the biological hypothesis you want to answer.**

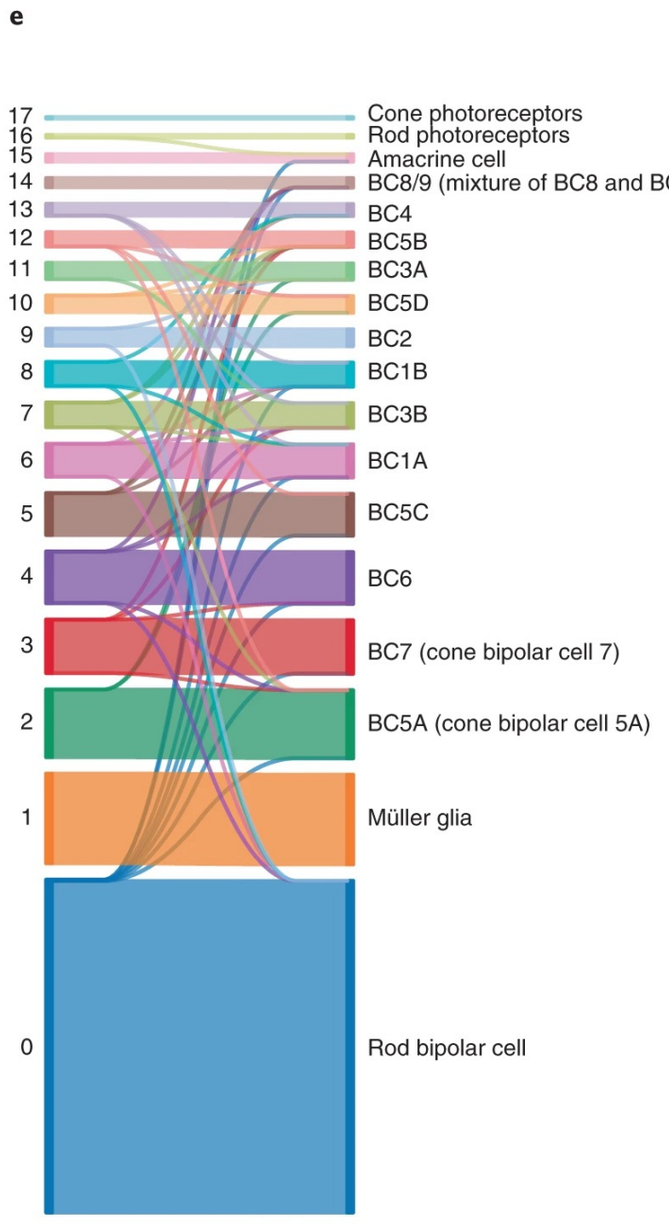
Alternative – SCAAF







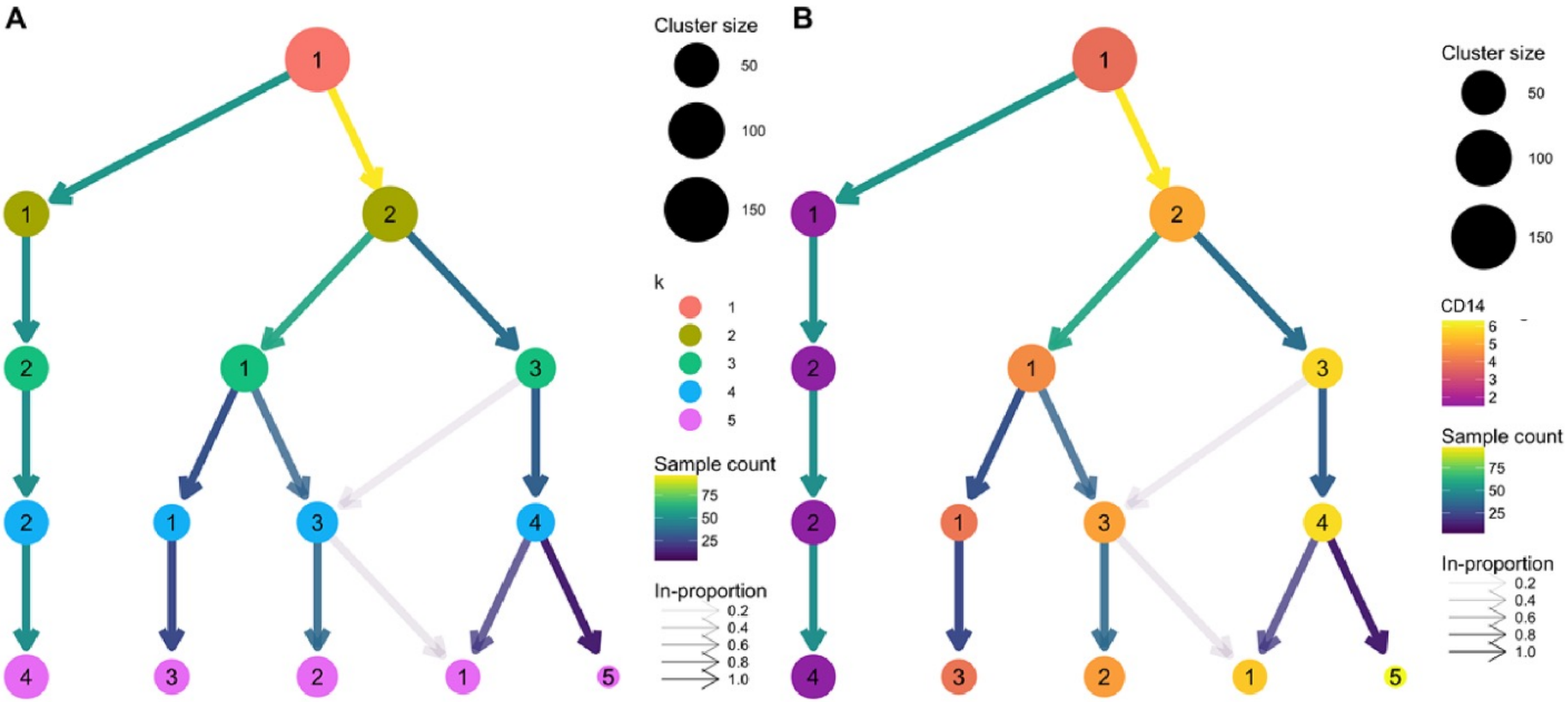




- **Through a series of clever optimisation SCAAF is able to find the « ground truth » cell types without specifying the number of cluster.**
- **… But again what is really ground truth? Cell types? Cell states?**

Miao, Z. *et al.* (2020) 'Putative cell type discovery from single-cell gene expression data', *Nature Methods*, 17(6), pp. 621–628. doi:10.1038/s41592-020-0825-9.

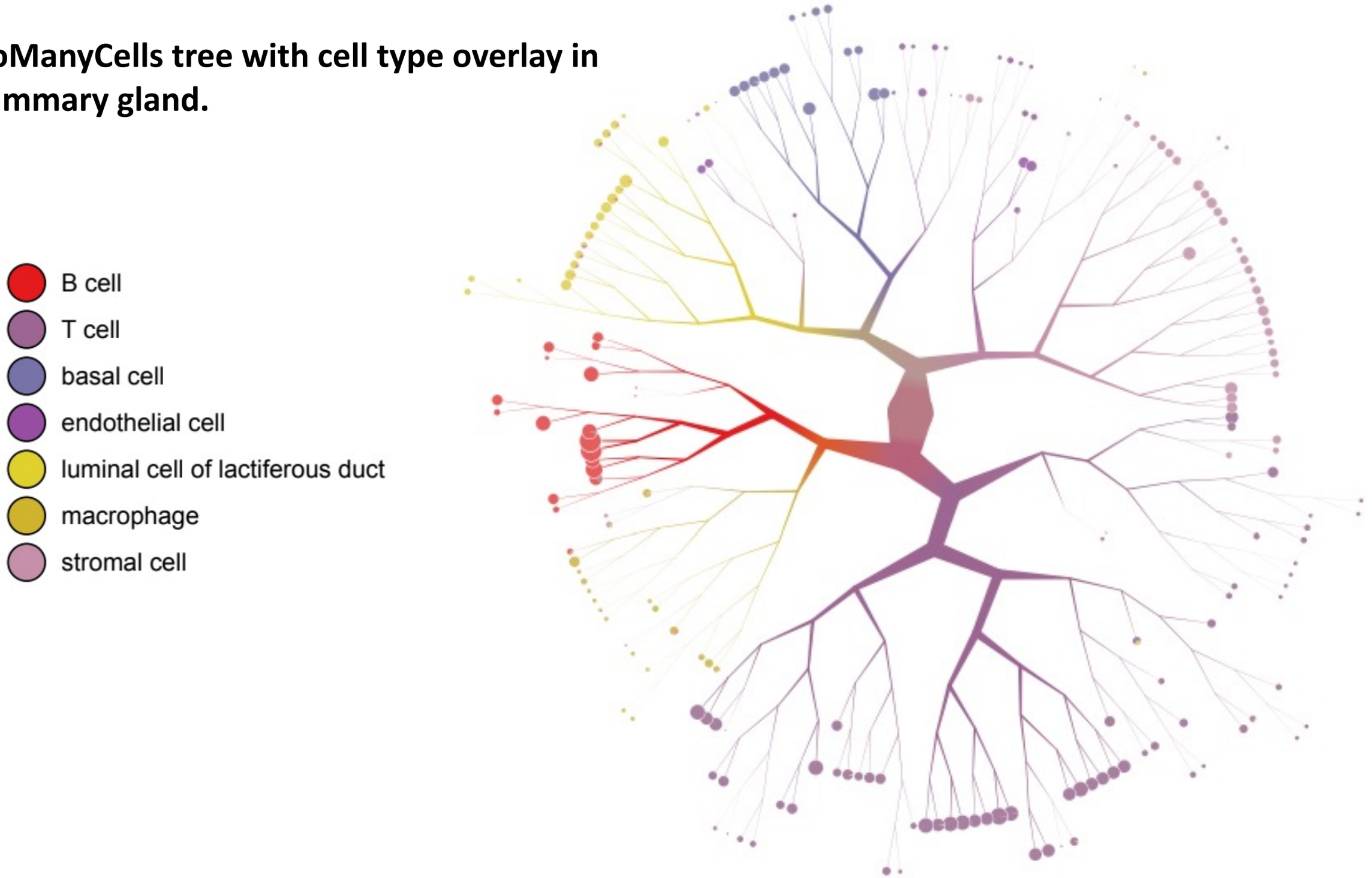
How many clusters: Clustree



Zappia, L. and Oshlack, A. (2018) 'Clustering trees: a visualization for evaluating clusterings at multiple resolutions', *GigaScience*, 7(7), p. giy083. doi:10.1093/gigascience/giy083.

Alternative – TooManyCells

TooManyCells tree with cell type overlay in mammary gland.

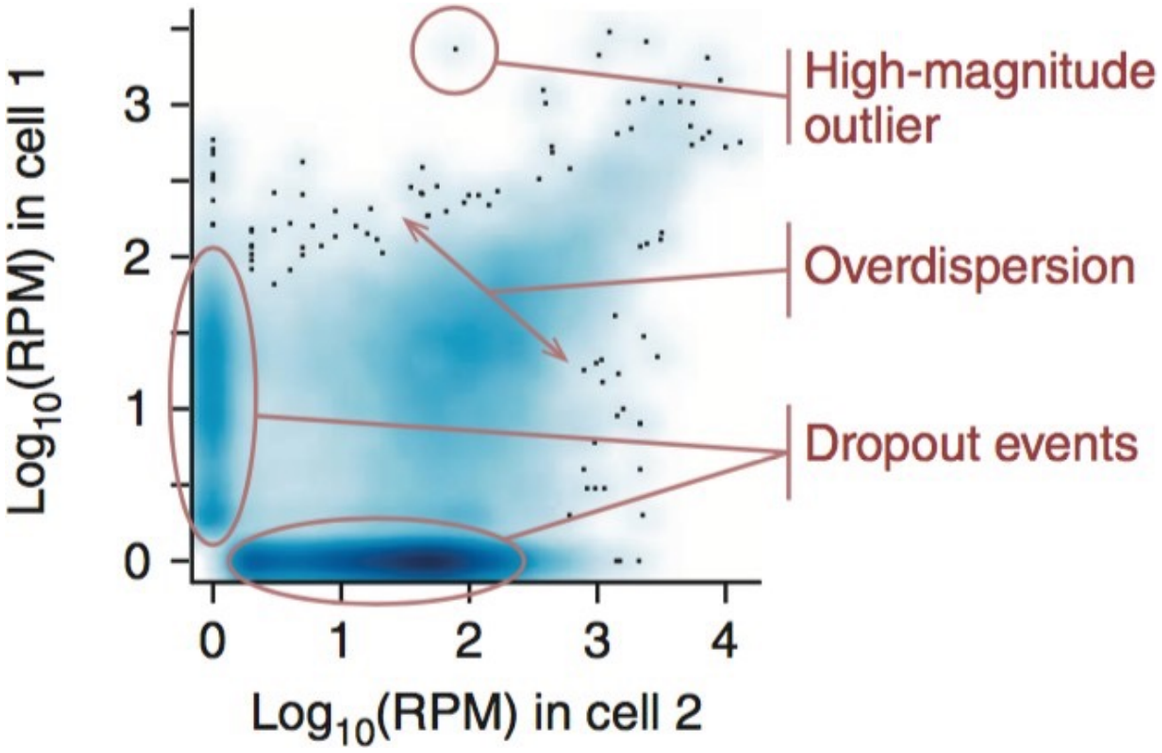


Schwartz, G. W. *et al.* (2020) 'TooManyCells identifies and visualizes relationships of single-cell clades', *Nature Methods*, 17(4), pp. 405–413. doi: 10.1038/s41592-020-0748-5.

The bioinformatics pipeline: main “modular” components

1. **Feature selection**
2. **Dimensionality Reduction**
3. **Clustering / Hiearchy**
4. **Differential Expression / Gene signature extraction**
5. **Functional interpretation**
6. **Batch effect correction**

Differential expression analysis of single-cell data: The statistical concepts



Kharchenko et al. Nature Methods (2014)

1st. Modeling the measurement of cells as a mixture of two probabilistic processes:

i.The transcript is amplified & detected at a level correlating with its abundance (count data)
Negative binomial distribution

ii.The transcript fails to amplify or is not detected for other reasons (to account for abundance of dropout events)
Poisson distribution
Zero-inflated negative binomial

2nd. Empirical Bayesian framework to regularize model parameters

- helps to improve inference for genes with sparse expression
- based on measurements of individual cells in order to estimate both the likelihood of a gene being expressed at any given average level in each subpopulation and the likelihood of expression fold change between them

SCDE Kharchenko et al. Nature Methods (2014) 11:740

3rd. Extend to Generalized linear modeling (GLM) in order to:

- Accommodate complex experimental designs
- Controlling for covariates (including technical factors) in both the discrete and continuous parts of the model.

MAST. Finak et al. Genome Biology 2015, 16:278

Differential expression analysis: The methods (I)

Tool	Prog. Language	Input format	Model	Year/ version	URL
SCDE	R	Read counts	Poisson and negative binomial model	2014/2.2.0	http://bioconductor.org/packages/release/bioc/html/scde.html
MAST	R	TPM/FPKM	Generalized linear model	2015/1.0.5	http://bioconductor.org/packages/release/bioc/html/MAST.html
scDD	R	TPM/FPKM	Conjugate Dirichlet process mixture	2016/0.99.0	http://bioconductor.org/packages/devel/bioc/html/scDD.html
EMDomics	R	TPM/FPKM	Non-parametric earth mover's distance	2016/2.4.0	https://www.bioconductor.org/packages/release/bioc/html/EMDomics.html
D3E	Python	Read counts	Cramér-von Mises test, Kolmogorov-Smirnov test, likelihood ratio test	2016/	https://github.com/hemberg-lab/D3E
Monocle2	R	TPM/FPKM	Generalized additive model	2014/2.2.0	http://bioconductor.org/packages/release/bioc/html/monocle.html
SINCERA	R	TPM/FPKM/Read counts	Welch's t-test and Wilcoxon rank sum test	2015/	https://research.cchmc.org/pbge/sincera.html
edgeR	R	Read counts	Negative binomial model, Exact test	2010/3.16.5	http://bioconductor.org/packages/release/bioc/html/edgeR.html
DESeq2	R	Read counts	Negative binomial model, Exact test	2014/1.14.1	http://bioconductor.org/packages/release/bioc/html/DESeq2.html
DEsingle	R	Read counts	Zero inflated negative binomial	2018/1.2.0	https://bioconductor.org/packages/release/bioc/html/DEsingle.html
SigEMD	R	TPM/FPKM	Non-parametric earth mover's distance	2018/0.21.1	https://github.com/NabaviLab/SigEMD

Wang, T. *et al.* (2019) 'Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data', *BMC Bioinformatics*, 20(1), pp. 1–16. doi:10.1186/s12859-019-2599-6.

Differential expression analysis: The methods (II)

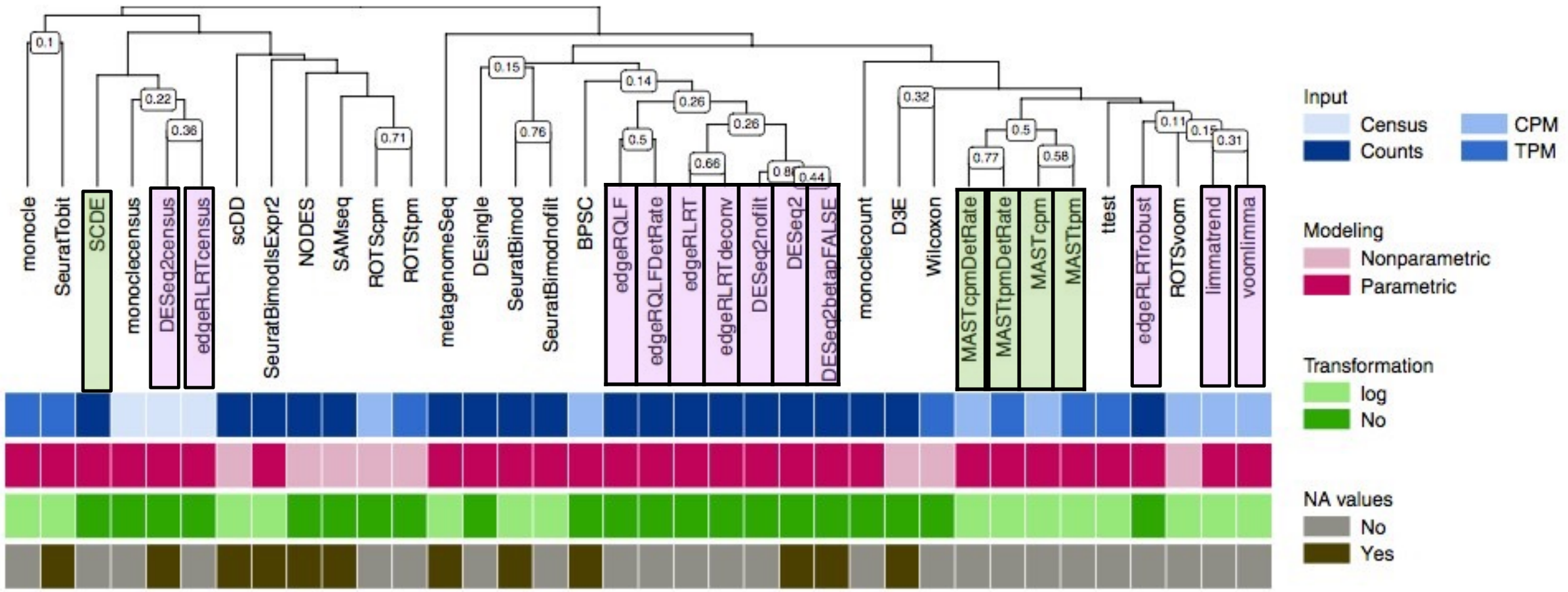
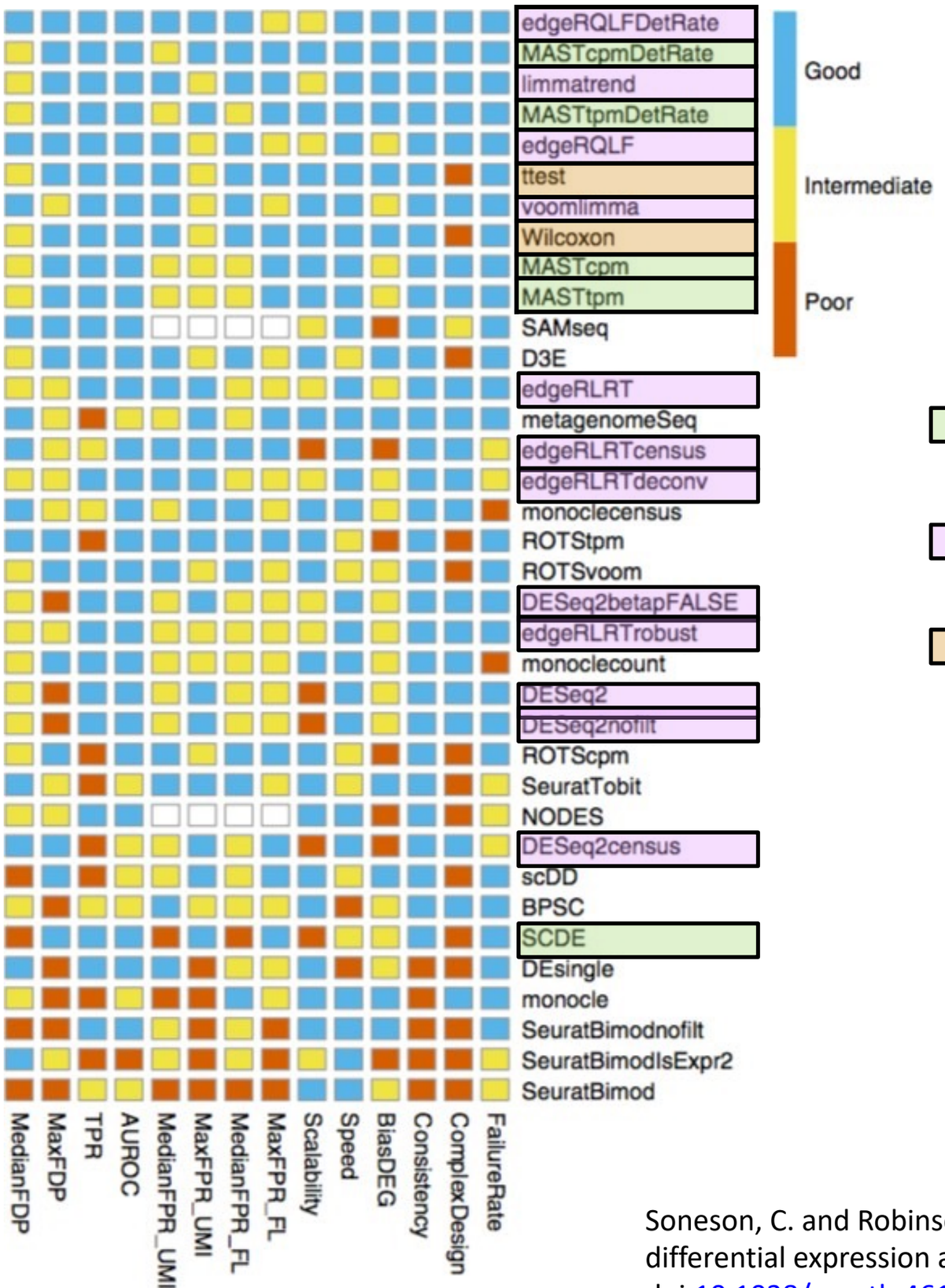


Figure 3 | Average similarities between gene rankings obtained by the evaluated DE methods. The dendrogram was obtained by complete-linkage hierarchical clustering based on the matrix of average AUCC values across all data sets. The labels of the internal nodes represent their stability across data sets (fraction of instances where they are observed). Only nodes with stability scores of at least 0.1 are labeled. Colored boxes represent method characteristics.

Soneson, C. and Robinson, M.D. (2018) 'Bias, robustness and scalability in single-cell differential expression analysis', *Nature Methods*, 15(4), pp. 255–261. doi:10.1038/nmeth.4612.

Differential expression analysis: The benchmark



Single cell specific methods

Methods borrowed from bulk-RNA-seq

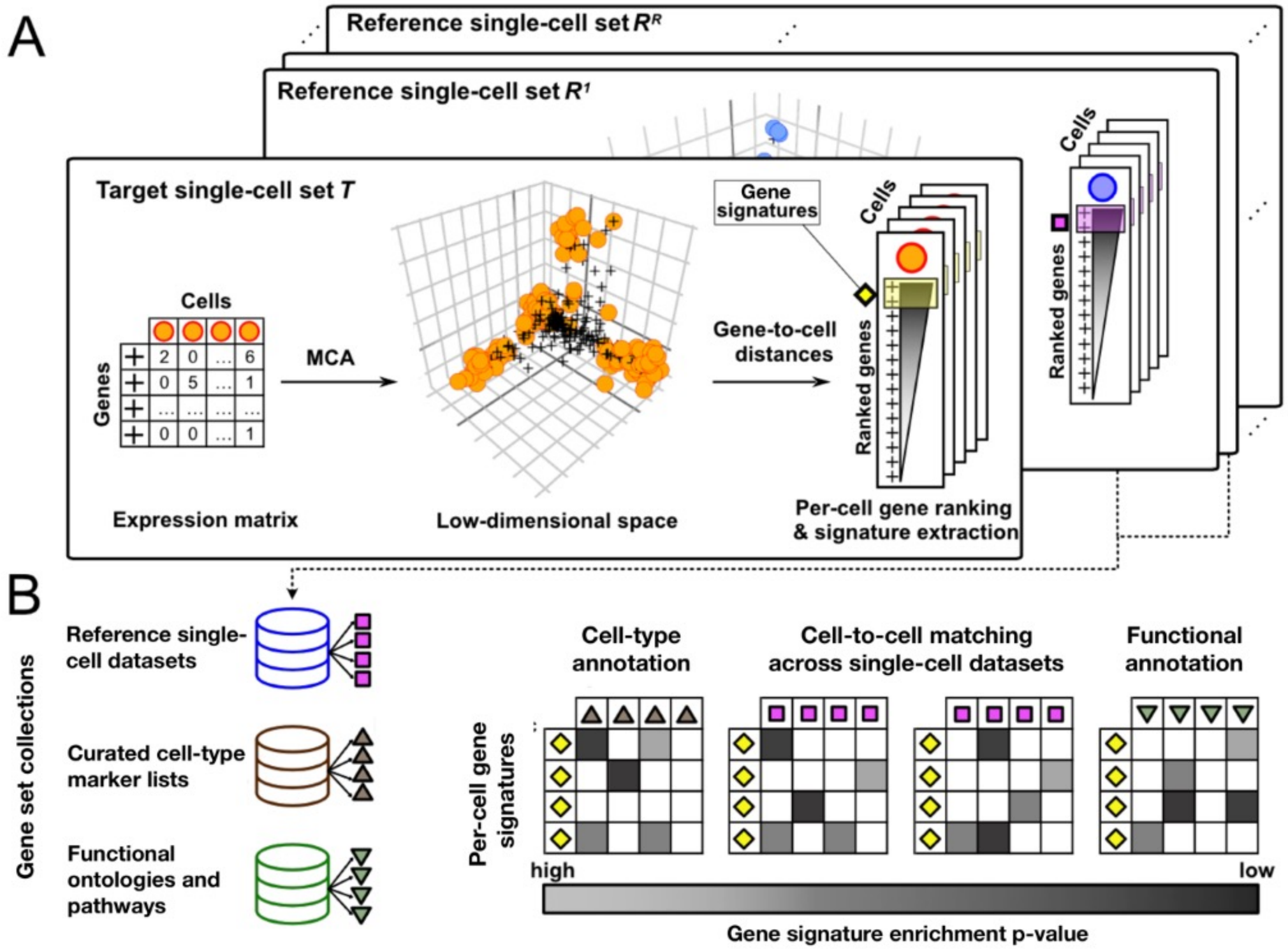
Naïve approaches

Soneson, C. and Robinson, M.D. (2018) 'Bias, robustness and scalability in single-cell differential expression analysis', *Nature Methods*, 15(4), pp. 255–261. doi:10.1038/nmeth.4612.

The bioinformatics pipeline: main “modular” components

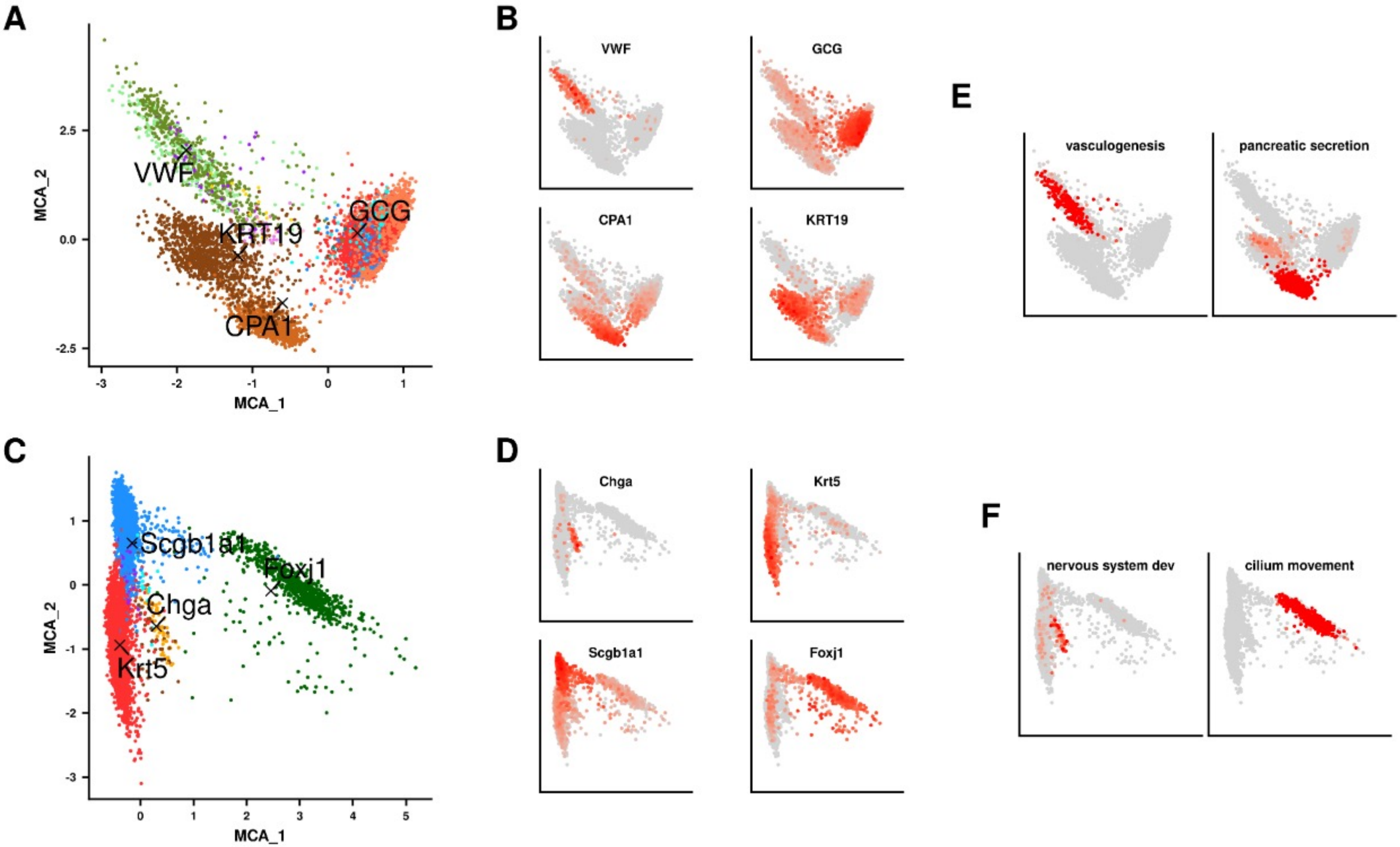
1. **Feature selection**
2. **Dimensionality Reduction**
3. **Clustering / Hierarchy**
4. **Differential Expression**
5. **Functional interpretation**
6. **Batch effect correction**

CellD: cell signatures extraction and enrichment analysis



Cortal, A. *et al.* (2021) 'Gene signature extraction and cell identity recognition at the single-cell level with Cell-ID', *Nature Biotechnology*, 39(9), pp. 1095–1102. doi:10.1038/s41587-021-00896-6.

CellD: cell signatures extraction and enrichment analysis

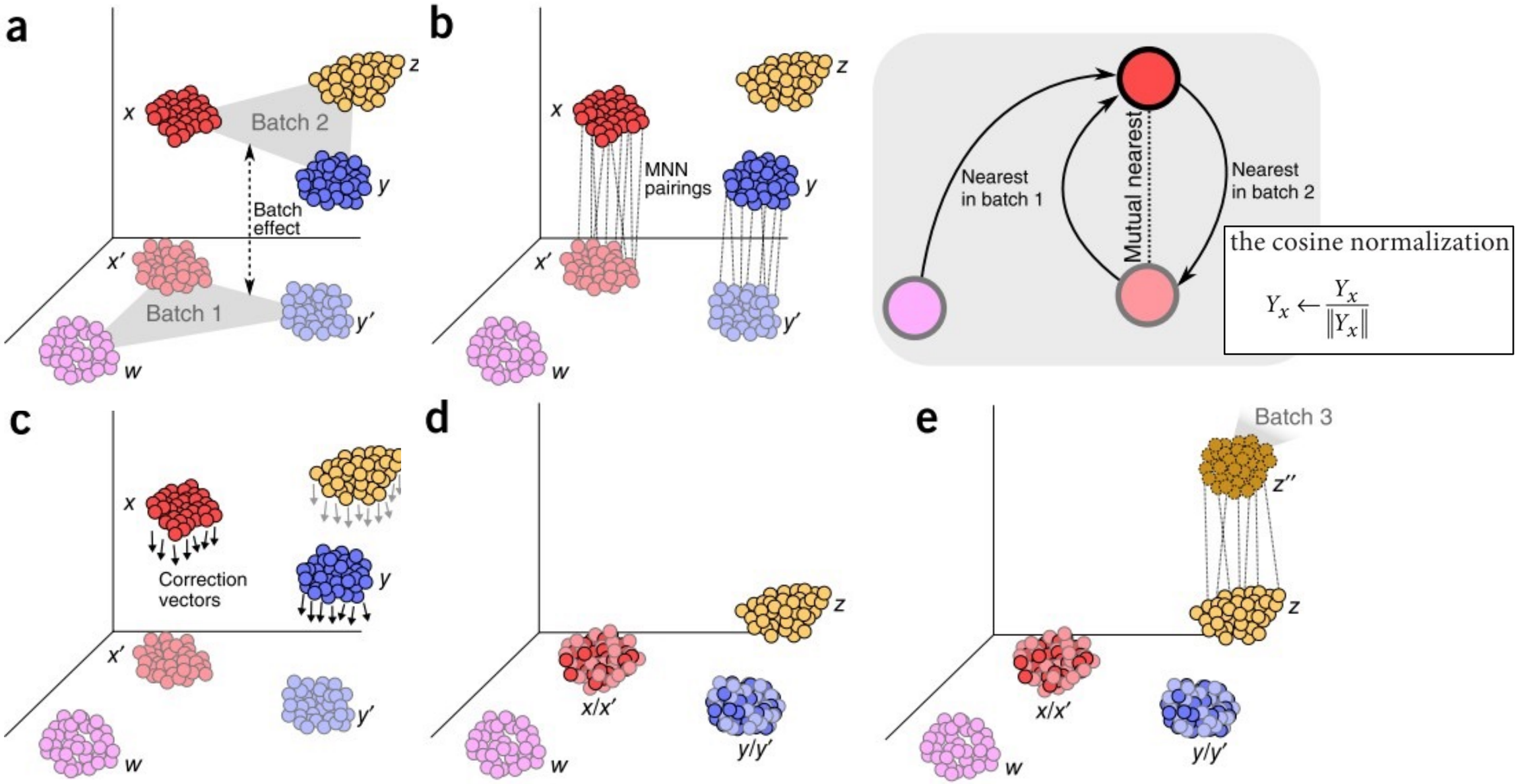


Cortal, A. *et al.* (2021) 'Gene signature extraction and cell identity recognition at the single-cell level with Cell-ID', *Nature Biotechnology*, 39(9), pp. 1095–1102. doi:10.1038/s41587-021-00896-6.

The bioinformatics pipeline: main “modular” components

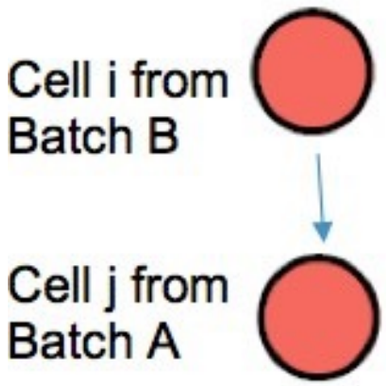
1. **Feature selection**
2. **Dimensionality Reduction**
3. **Clustering / Hiearchy**
4. **Differential Expression**
5. **Functional interpretation**
6. **Batch effect correction**

Mutual Nearest Neighbors (MNN): (I) Overview



Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Haghverdi et al. Nature Biotechnology 2018 doi:10.1038/nbt.4091

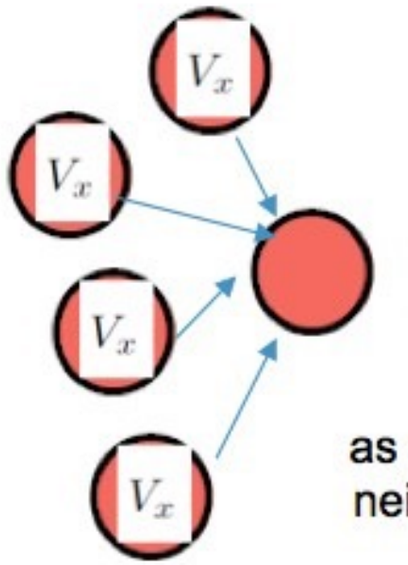
Mutual Nearest Neighbors (MNN): (II) Correction vectors



1) For each MNN pair, a pair-specific batch-correction vector is computed as the vector difference between the expression profiles of the paired cells.

$$V_x = \begin{pmatrix} gene1_a - gene1_b \\ gene2_a - gene2_b \\ gene3_a - gene3_b \\ \cdots \\ geneN_a - geneN_b \end{pmatrix}$$

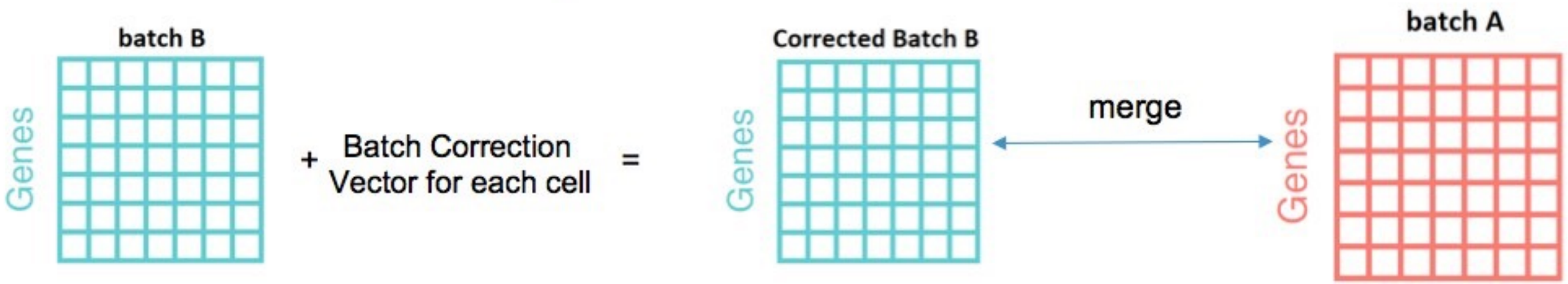
2) A cell-specific batch-correction vector is then calculated as a weighted average of these pair-specific vectors, as computed with a Gaussian kernel.



Gaussian Kernel Smoothing

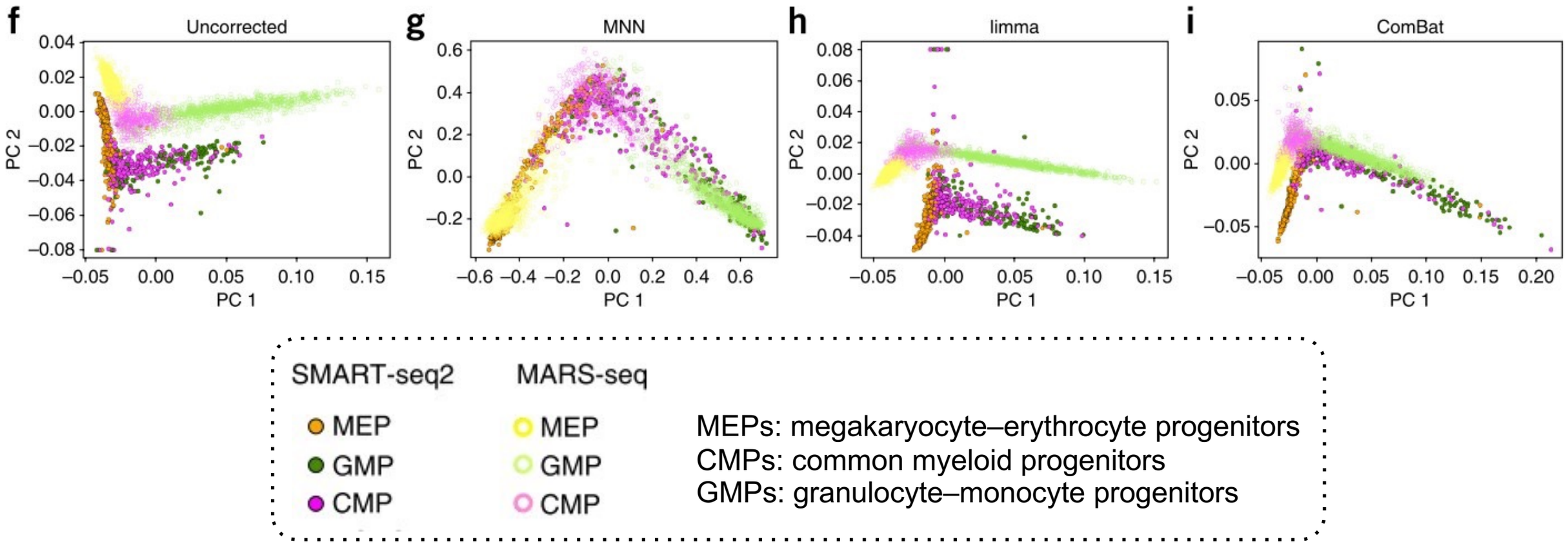
Real valued function
$f : R^p \rightarrow R$
as the weighted average of neighboring observed data

Batch Correction vector for each cell



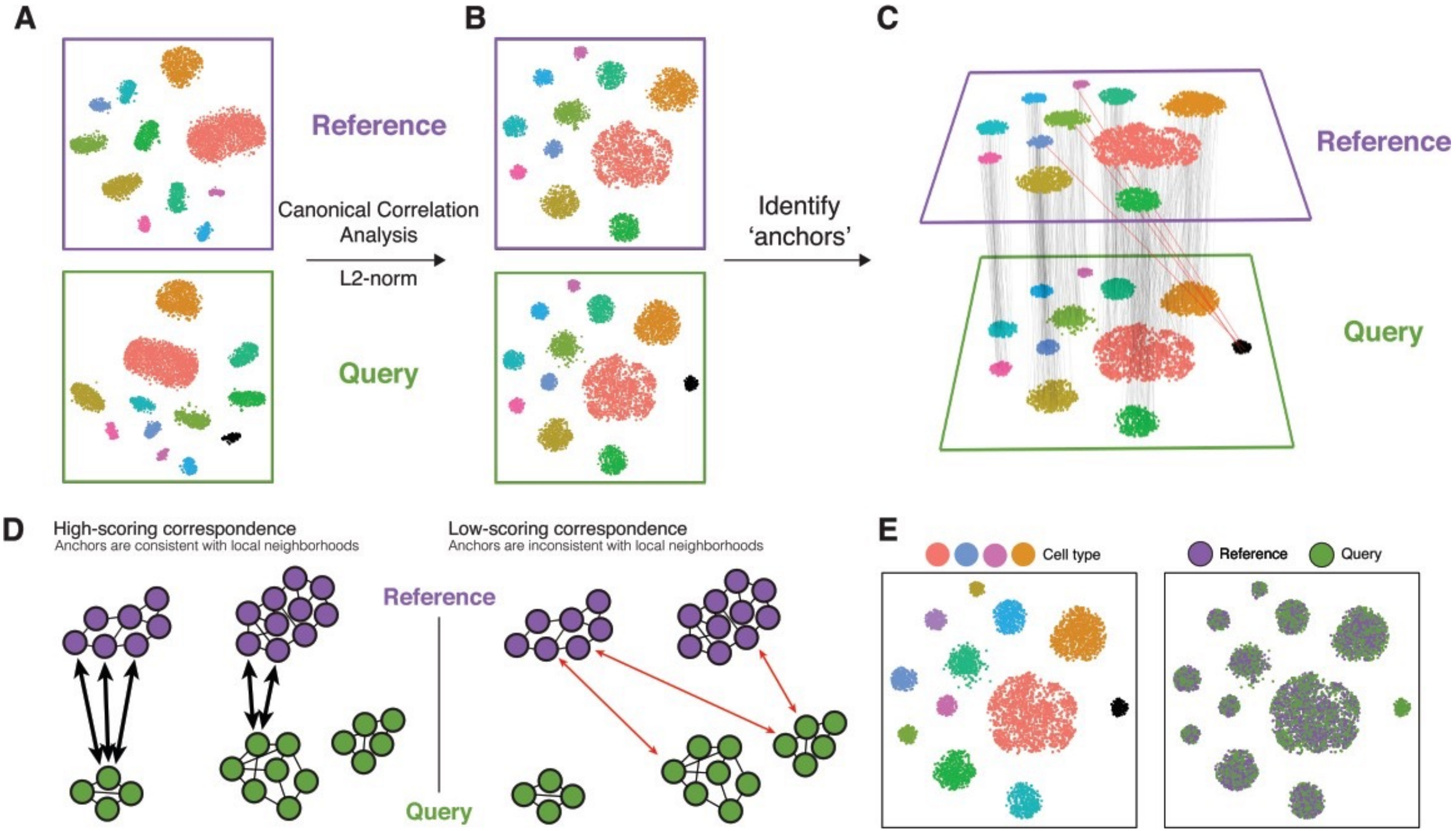
Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Haghverdi et al. Nature Biotechnology 2018 doi:10.1038/nbt.4091

Mutual Nearest Neighbors (MNN): (III) Example



Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Haghverdi et al. Nature Biotechnology 2018 doi:10.1038/nbt.4091

Seurat v3's integration: CCA + "anchors"



Comprehensive integration of single cell data Stuart et al. BioRxiv (2018)
https://www.biorxiv.org/content/10.1101/460147v1

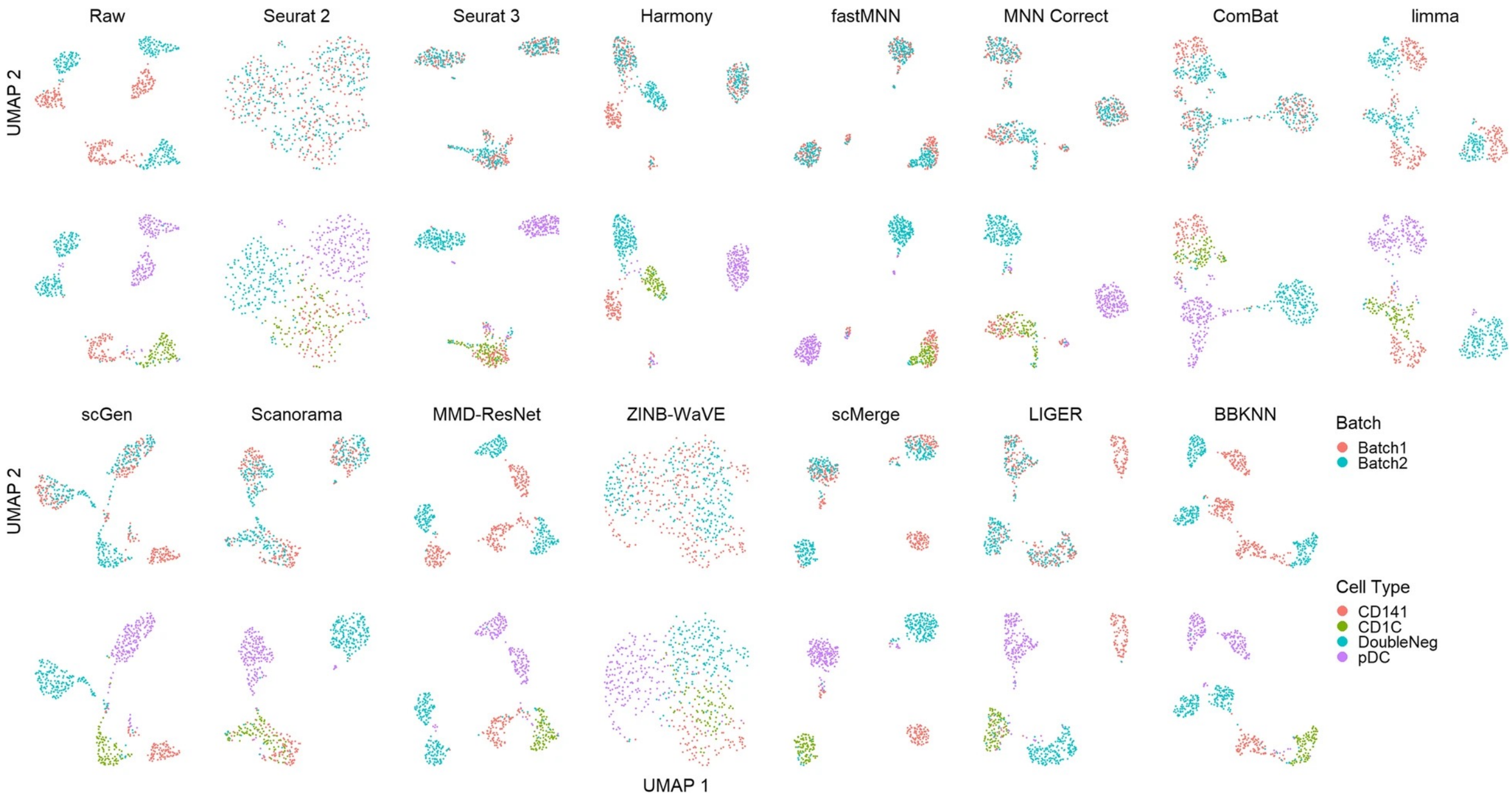
Summary of integrative methods

Tool	Lang.	Output	Correction principle	Installation	License	Ref
mnnCorrect	R	Counts matrix	Mutual nearest neighbour detection across batches.	Batchelor	GPL-3	(16)
Limma	R	Counts matrix	Fits linear model to remove batch effect components.	Limma	GPL (>=2)	(21)
ComBat	R	Counts matrix	Adjusts for known batches using an empirical Bayesian framework.	Sva	Artistic-2.0	(22)
Seurat	R	Counts matrix	Diagonalized CCA to reduce dimensionality and MNN detection in this space.	Seurat (CRAN)	GPL-3	(23)
Scanorama	Python	Counts matrix	SVM to reduce dimensionality and mutual nearest neighbor detection and panoramic stitching.	pip	MIT	(24)
Harmony	R	Embedding	Iterative soft k-means clustering algorithm in dimensionally reduced space.	Github	GPL-3	(25)
fastMNN	R	Embedding	Mutual nearest neighbor detection after multi-sample PCA.	Batchelor	GPL-3	(16)
BBKNN	Python	Graph	Mutual nearest neighbour pair selection across batches in PCA space.	pip3	MIT	(26)

R, Bioconductor, Conda, Python.

Chazarra-Gil, R. *et al.* (2021) 'Flexible comparison of batch correction methods for single-cell RNA-seq using BatchBench', *Nucleic Acids Research*, 49(7), p. e42. doi:10.1093/nar/gkab004.

Summary of integrative methods



Tran, H.T.N. *et al.* (2020) 'A benchmark of batch-effect correction methods for single-cell RNA sequencing data', *Genome Biology*, 21(1), p. 12. doi:10.1186/s13059-019-1850-9.

My Routine Pipeline

Framework: Seurat

Normalisation: scran deconvolution normalisation

Highly variable genes: Seurat FindVariableGenes

Dimensionality reduction: Seurat PCA

Two dimensional embeddings: Seurat UMAP, SWNE

Clustering: Seurat graph clustering

Differential expression: Wilcoxon-test or MAST

Functionnal Analysis: CellID, AUCell

Integration methods: Harmony

Cell type inference: sciBet, SCINA

Benchmarks evaluating each of the analytical steps:

Assessment of Single Cell RNA-Seq Normalization Methods.
http://www.g3journal.org/content/7/7/2039.long

Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data
https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bby011/4898116

A systematic performance evaluation of clustering methods for single-cell RNA-seq data
https://f1000research.com/articles/7-1141/v2

Bias, robustness and scalability in single-cell differential expression analysis.
https://www.nature.com/articles/nmeth.4612

A comparison of single-cell trajectory inference methods: towards more accurate and robust tools
https://www.nature.com/articles/s41587-019-0071-9

A test metric for assessing single-cell RNA-seq batch correction (KBet)
https://www.nature.com/articles/s41592-018-0254-1

Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments
https://www.nature.com/articles/s41592-019-0425-8

A benchmark of batch-effect correction methods for single-cell RNA sequencing data
https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1850-9

A comparison of automatic cell identification methods for single-cell RNA sequencing data
https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1795-z

Recommended online courses

Orchestrating Single-Cell Analysis with Bioconductor
https://osca.bioconductor.org/

Seurat Vignettes
https://satijalab.org/seurat/vignettes.html

Complete course on Single-cell RNA-seq data analysis from U. Cambridge
http://hemberg-lab.github.io/scRNA.seq.course/index.html

Bioinformatics Training channel on YouTube
https://www.youtube.com/channel/UCsc6r6UKxb2qRcDQPix2L5A

A step-by-step workflow for low-level analysis of single-cell RNA-seq data
https://f1000research.com/articles/5-2122/v1

Thanks for your attention!