Notes

- Projects
  - Many private channels are setup; will organise repos this week
  - office hours: get in touch for discussions (27 Dec - 5 Jan unlikely; before/after fine), otherwise ask questions via Slack.
- Exercises
  - Remember: top 9 exercises are counted (+3 for free to get mark out of 30)
- Journal clubs
  - 1/2 presentation; 1/2 feedback to others

Single cell analysis

- common threads of data analysis: dimension reduction, clustering, differential expression, differential abundance, (differential state, etc.)

SINGLE-CELL OMICS

Computational and analytical challenges in single-cell transcriptomics

Oliver Stegle[1], Sarah A. Teichmann[1,2] and John C. Marioni[1,2]

Why single cell?

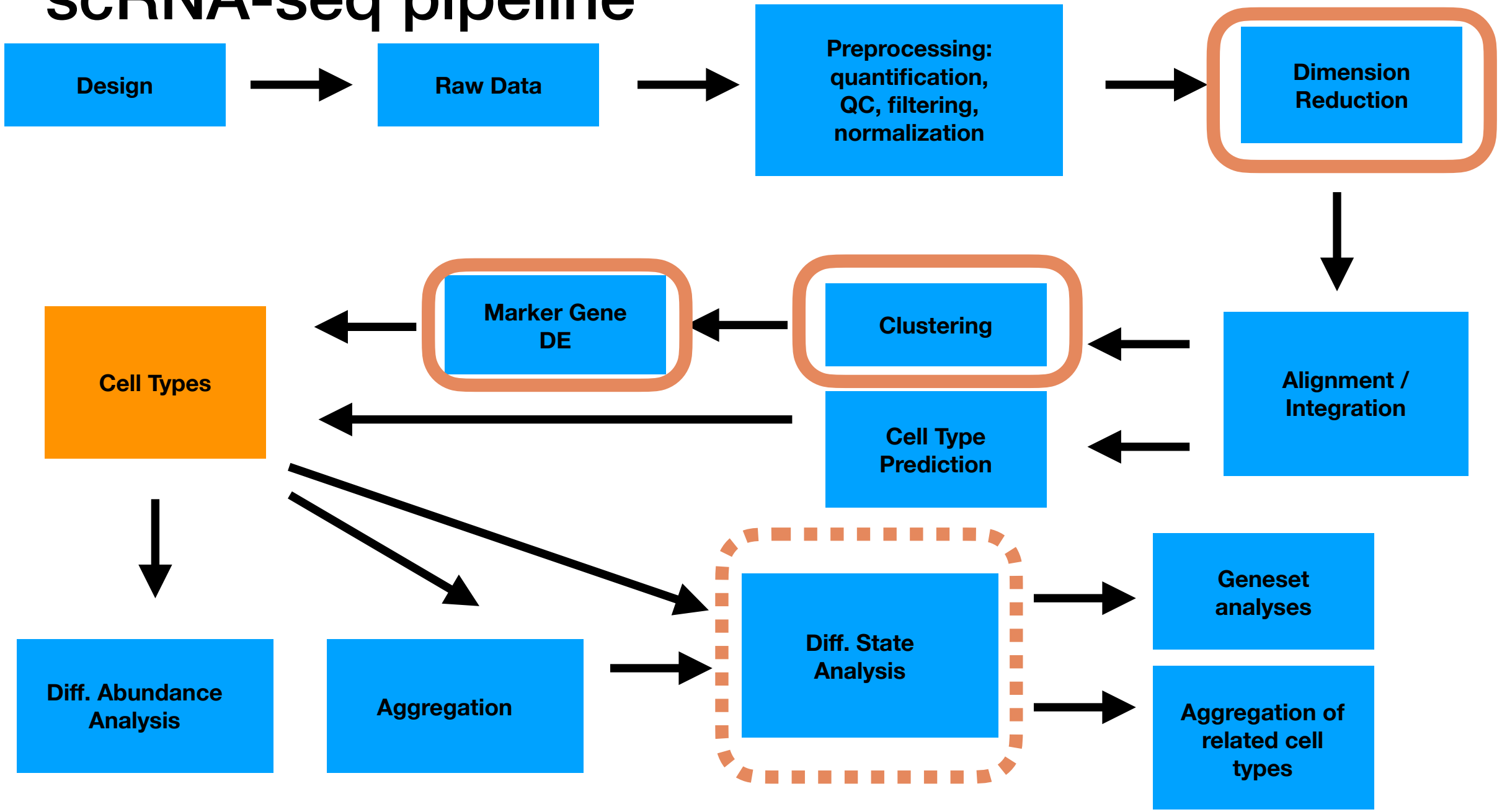
"Bulk" versus single-cell

Discover and quantify abundance of (new) cell types

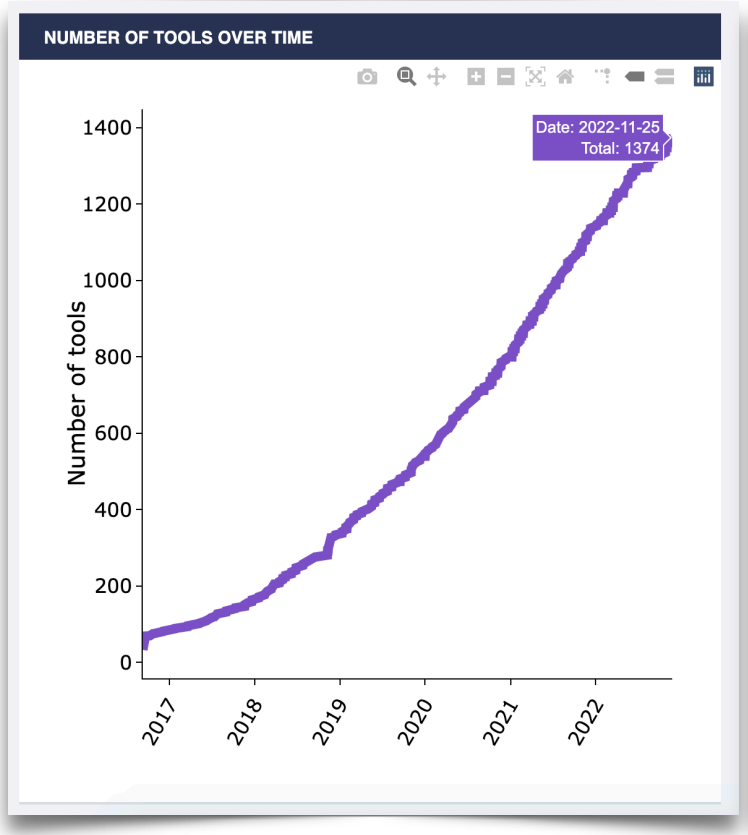
Study heterogeneity of gene expression

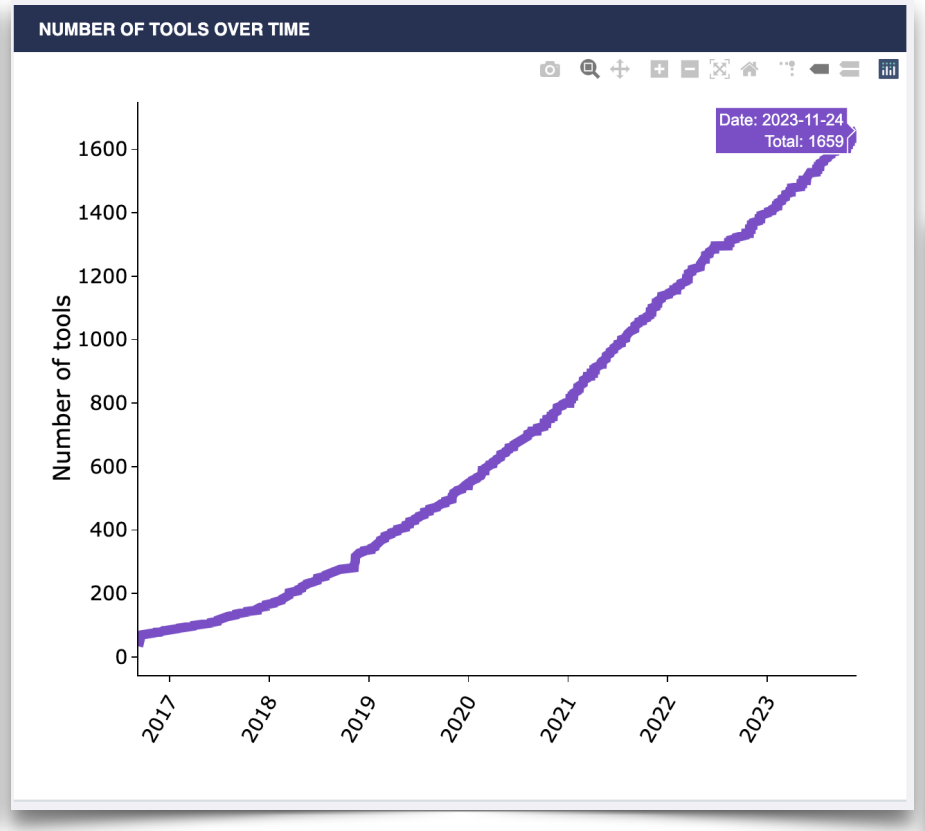
However, there are also important biological questions for which bulk measures of gene expression are insufficient[14]. For instance, during early development, there are only a small number of cells, each of which can have a distinct function and role[15–17]. Moreover, complex tissues, such as brain tissues, are composed of many distinct cell types that are typically difficult to dissect experimentally[18]. Consequently, bulk-based approaches may not provide insight into whether differences in expression between samples are driven by changes in cellular composition (that is, the abundance of different cell types) or by changes in the underlying phenotype. Finally, ensemble measures do not provide insights into the stochastic nature of gene expression[19,20].

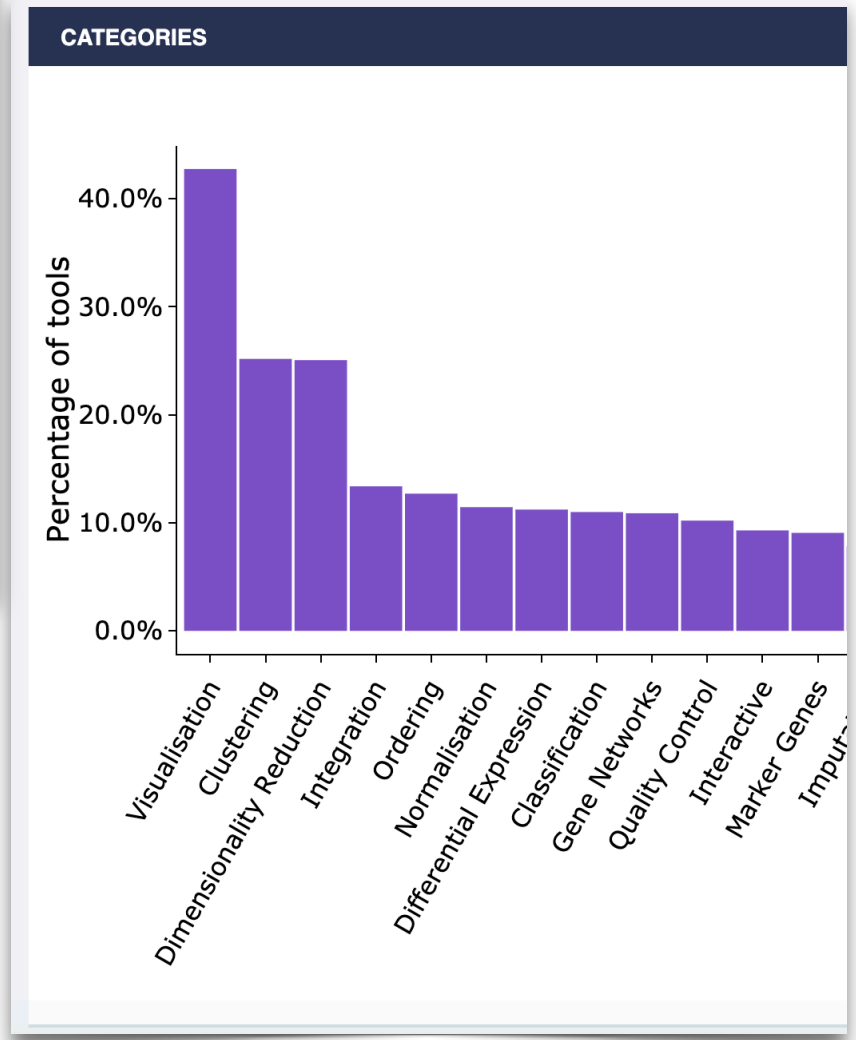
scRNA-seq pipeline



Method velocity







REVIEW Open Access

Over 1000 tools reveal trends in the single-cell RNA-seq analysis landscape

Luke Zappia[1,2] and Fabian J. Theis[1,2,3*]

RESEARCH ARTICLE

Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database

Luke Zappia[1,2], Belinda Phipson[1], Alicia Oshlack[1,2]*

1 Bioinformatics, Murdoch Children's Research Institute, Melbourne, Victoria, Australia, **2** School of Biosciences, Faculty of Science, University of Melbourne, Melbourne, Victoria, Australia

An application (motivation)

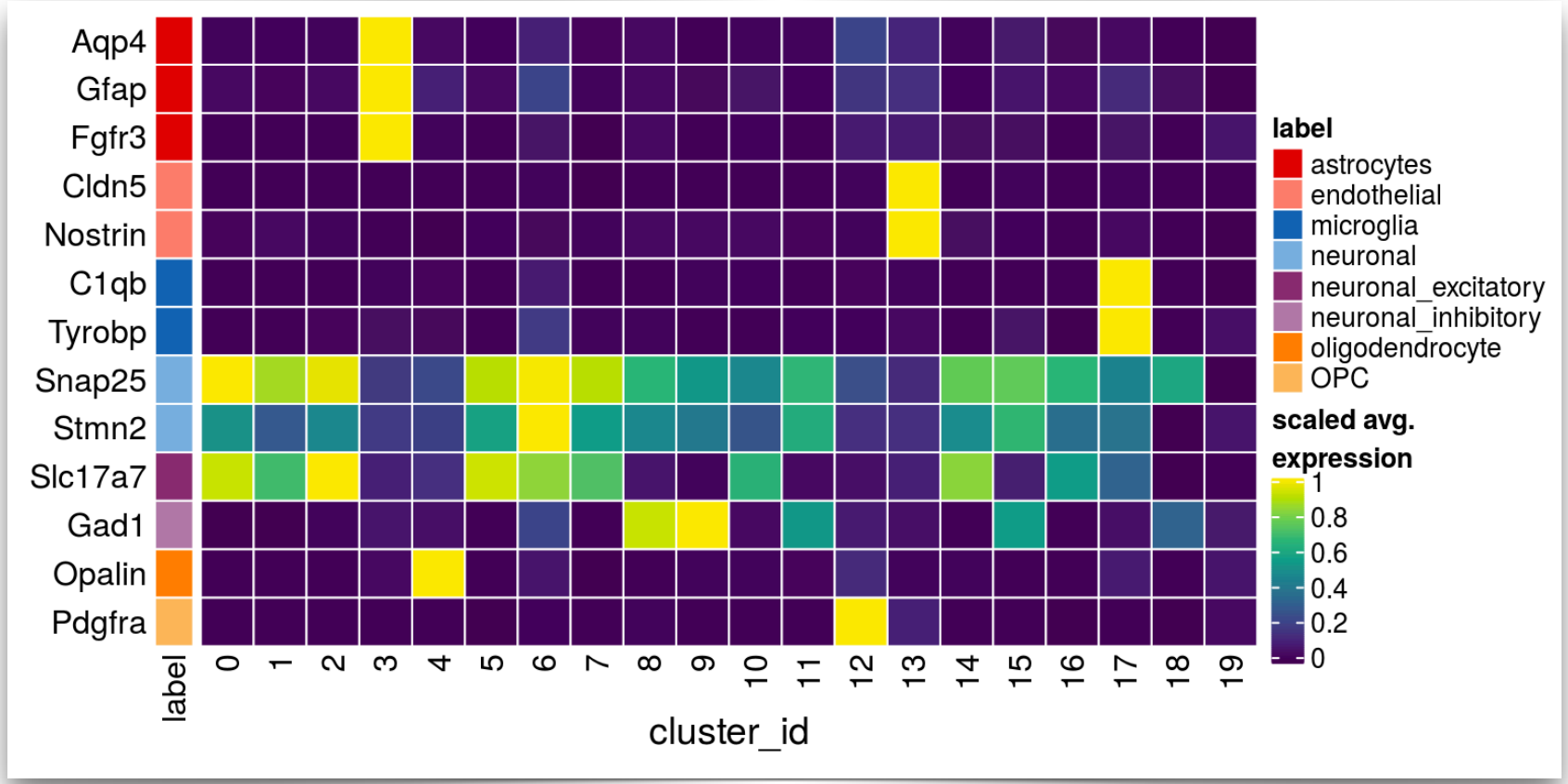
Application to LPS dataset: clustering + annotation of subpopulations

Data from:
4 mice treated with vehicle
4 mice treated with LPS

frontal cortex

single nuclei RNA-seq (10x)

usual preprocessing:
filtering, doublet removal,
Seurat integration,
clustering

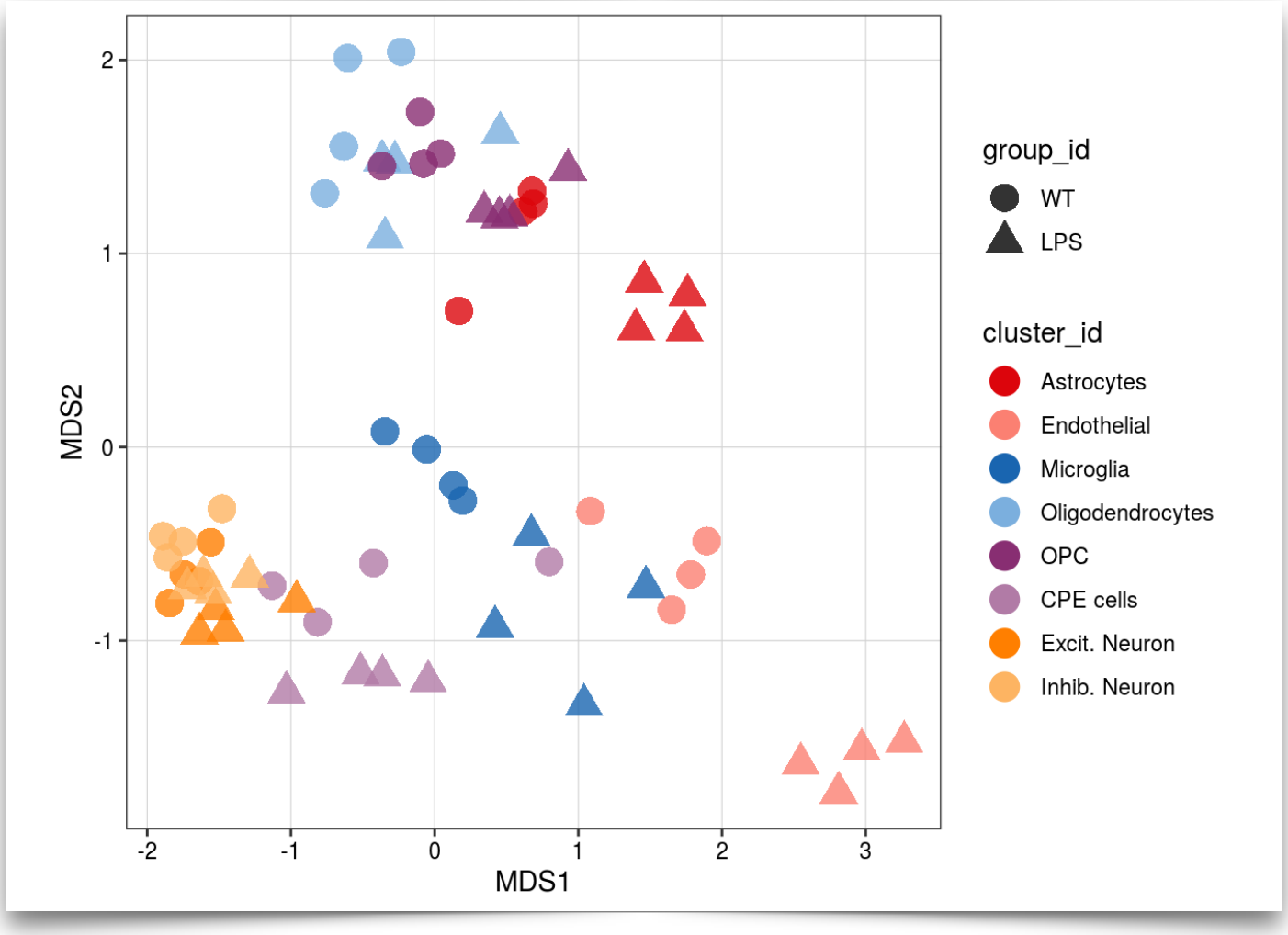


LPS dataset: interplay of cell type and cell state

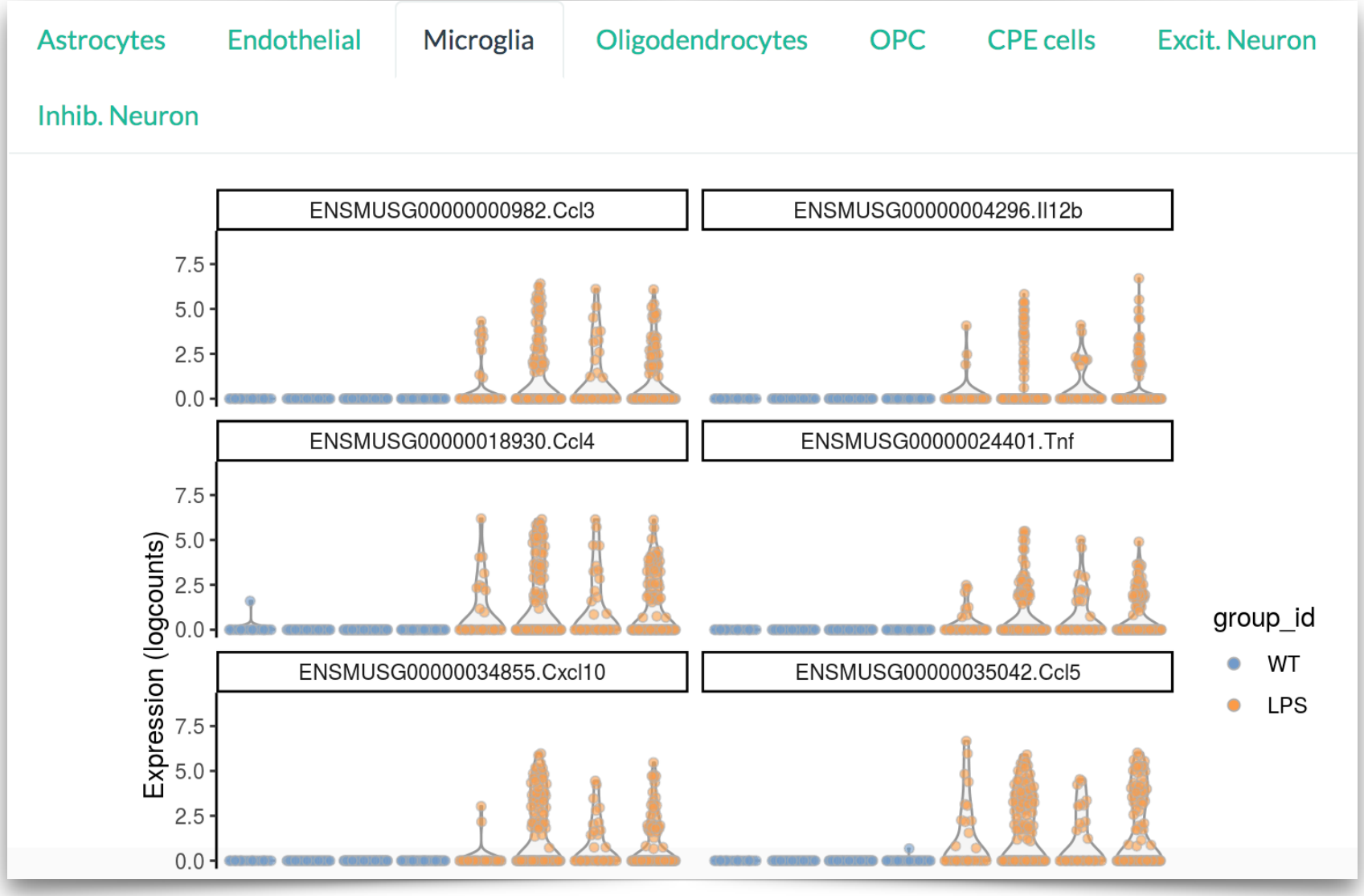
Application to LPS dataset: subpopulation-level visualization

Data from:
4 mice treated with vehicle
4 mice treated with LPS

Each dot is one subpopulation/ sample combination

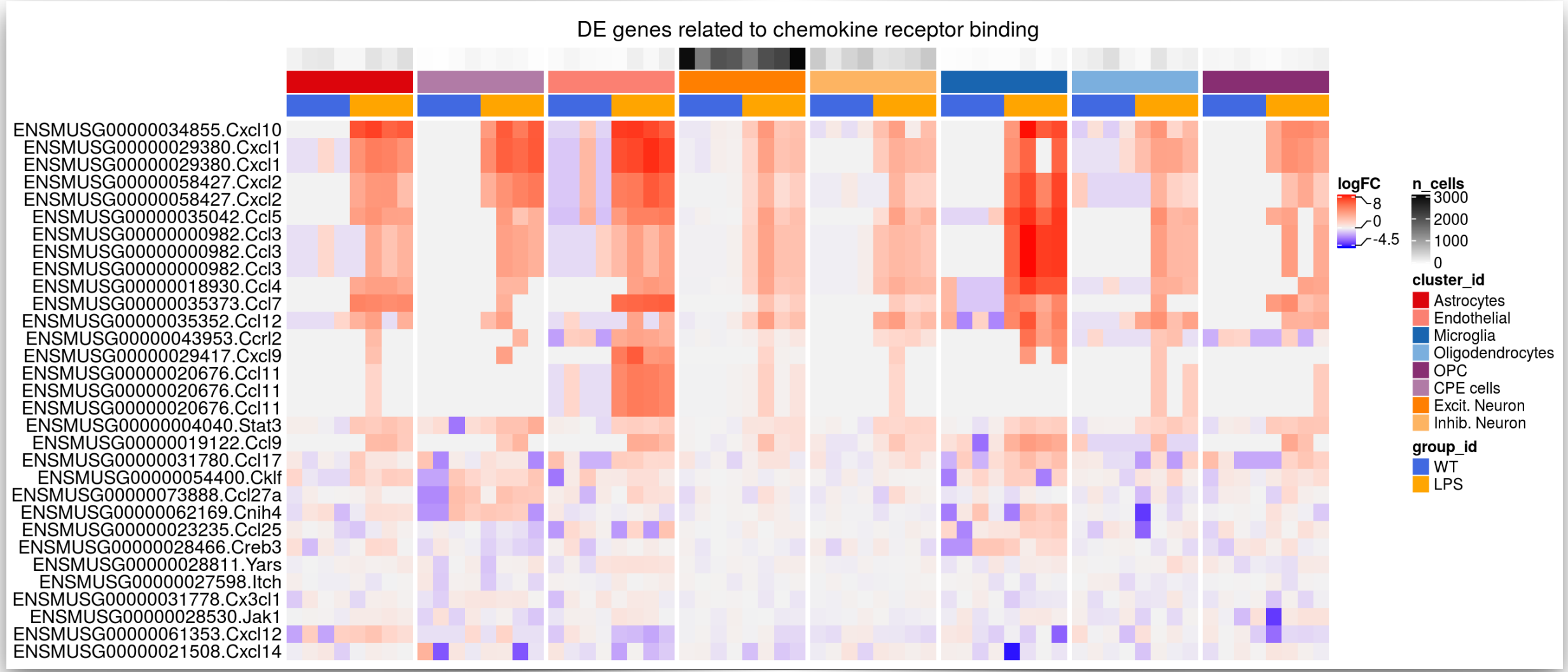


Application to LPS dataset: go back to cell-level response (discovery based on pseudobulk)

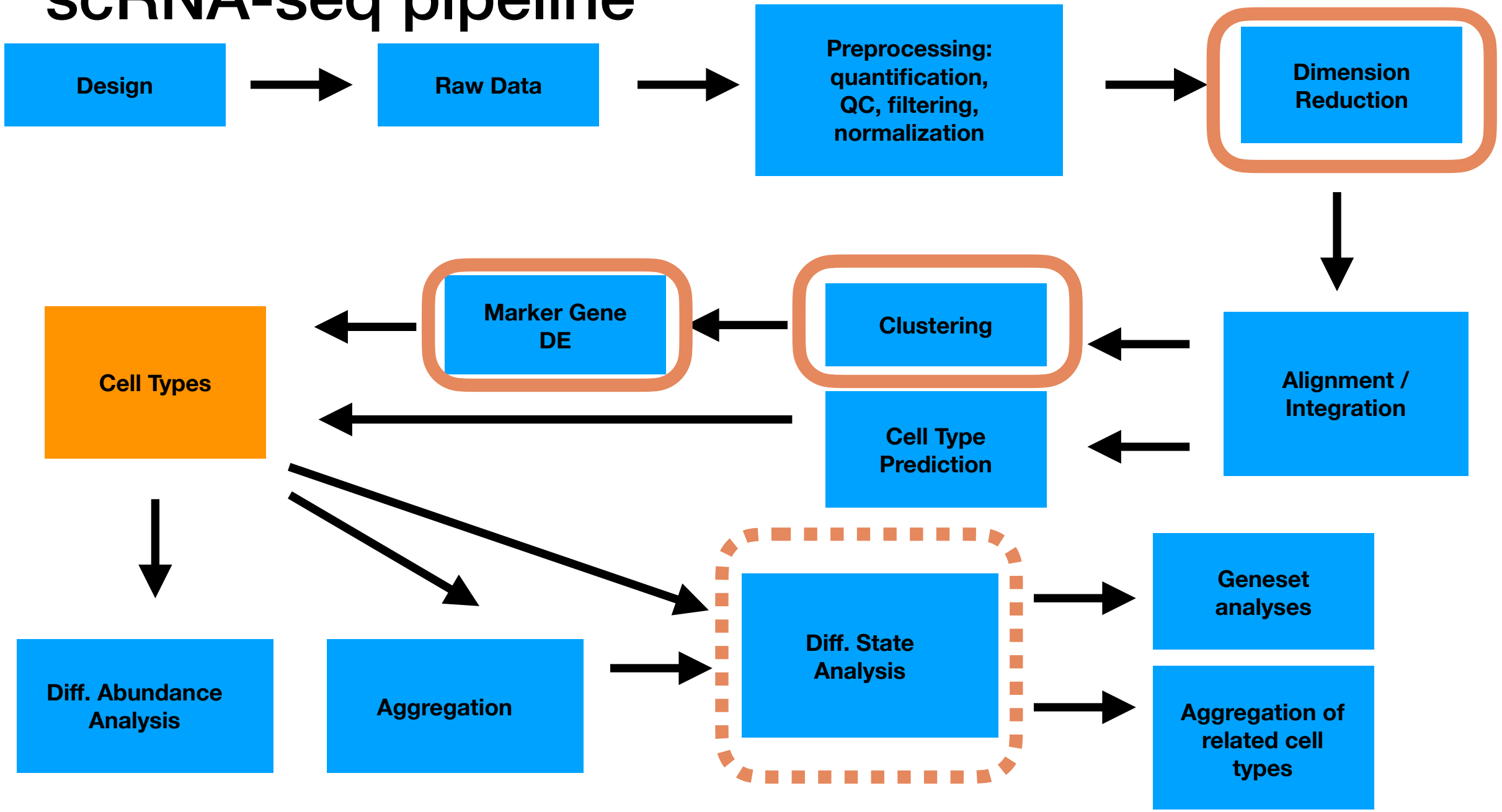


workflowr !

Application to LPS dataset: look at genes (genesets) changing {within specific, common across} subpopulations



scRNA-seq pipeline



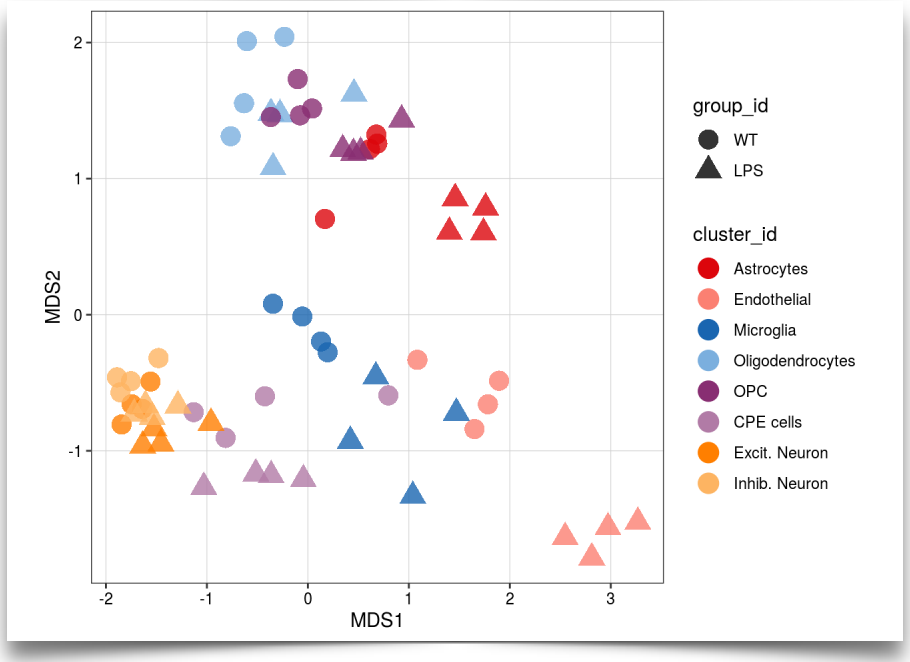
Dimension reduction

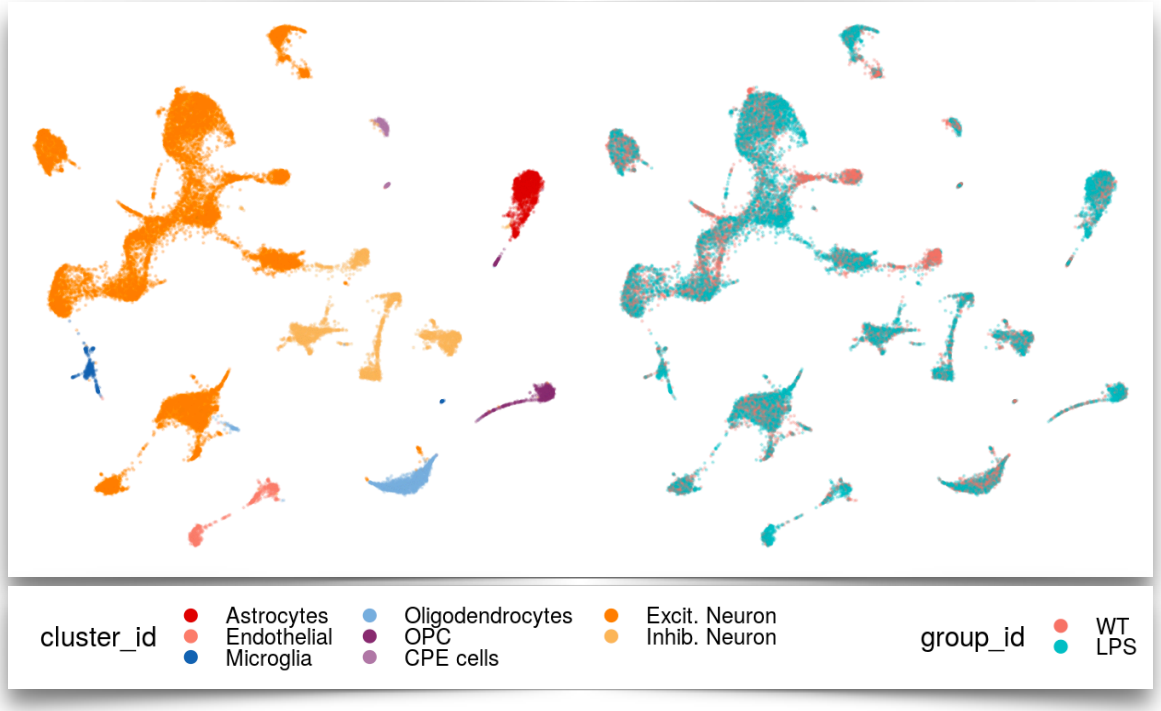
Dimension reduction: general introduction

- Single cell data comes as a matrix of N cells x G features
- Each cell is a point in G-dimensional space
- Goal: represent the data in 2-3 dimensions, but preserve **structure** as best as possible (i.e., points that are **close** in G dimensions should be close in 2/3 dimensions)

Dimension reduction: interpretation (single cell)

- Distances (in low-dimensional space) represent transcriptional changes: the larger the distance, the larger the transcriptional differences
- Hides: what are the transcriptional differences (e.g., few feature with large differences or many features with subtle differences)
- Local structure is typically better preserved than global structure





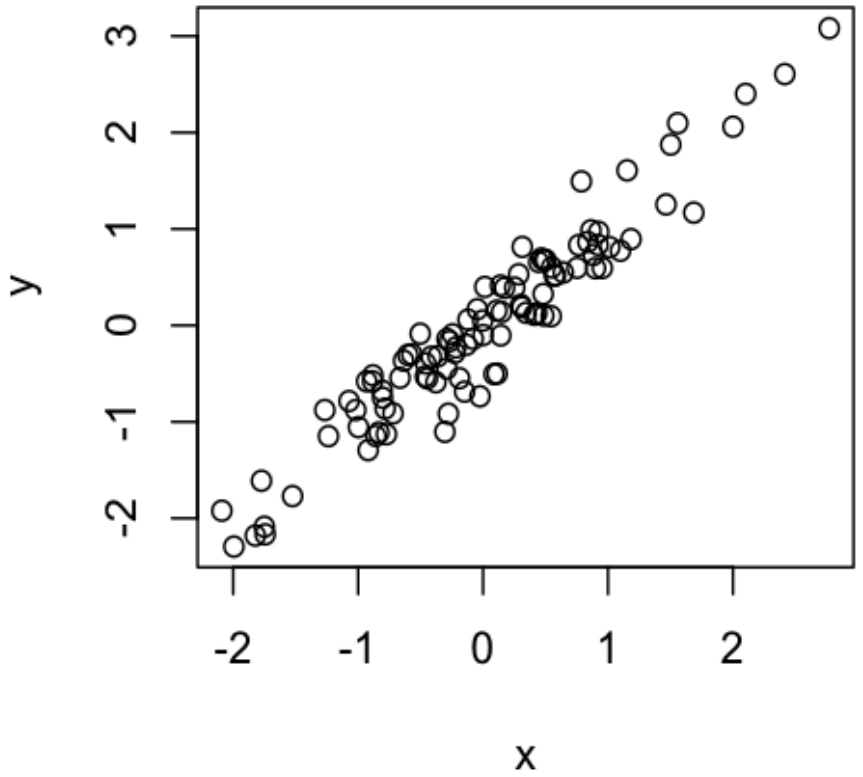
Introduction to dimension reduction: PCA (principal components analysis)

- Form successive *linear* combinations of the features that are: orthogonal, ordered by variance

$$Y = XA$$

$$Y_{rk} = a_{1k}x_{r1} + a_{2k}x_{r2} + \cdots + a_{pk}x_{rp}$$

- A is the loadings matrix
- Typically, first 2-3 columns ('principal components') of Y are retained for visualisation; often top P PCs are retained for other analyses (e.g., clustering)



https://strata.uga.edu/software/pdf/pcaTutorial.pdf

Introduction to dimension reduction: MDS (multi-dimensional scaling)

- Given a matrix of **distances** between each pair of cells, MDS places each cell into N-dimensional space such that the between-object distances are preserved as well as possible.

$$D := \begin{pmatrix} d_{1,1} & d_{1,2} & \cdots & d_{1,M} \\ d_{2,1} & d_{2,2} & \cdots & d_{2,M} \\ \vdots & \vdots & & \vdots \\ d_{M,1} & d_{M,2} & \cdots & d_{M,M} \end{pmatrix}.$$

The goal of MDS is, given D, to find M vectors $x_1, \ldots, x_M \in \mathbb{R}^N$ such that

$\|x_i - x_j\| \approx d_{i,j}$ for all $i, j \in 1, \ldots, M$,

$$\min_{x_1,\ldots,x_M} \sum_{i<j} (\|x_i - x_j\| - d_{i,j})^2.$$

https://en.wikipedia.org/wiki/Multidimensional_scaling

Introduction to dimension reduction: practical considerations

- initial subset of the number of features
- Is the data "linear"?
- distances in high dimensional space: be wary of the "curse of dimensionality"
- stochastic optimisation (seeds versus multiple runs)
- local structure versus global structure
- time complexity
- can new data be embedded?

Introduction to dimension reduction: tSNE (t-distributed stochastic neighbor embedding)

- *If you do not know what tSNE is, .., you probably do not need to know because tSNE is basically dead by now* (https://towardsdatascience.com/how-exactly-umap-works-13e3040e1668)
- Governed by two laws:
  - 1. all points are repelled from each other
  - 2. points are attracted to their nearest neighbours
- $p_{j|i}$ - directional similarity of point j to point i
- perplexity related to sigma
- Minimize KL distance between p_{ij} and q_{ij}

$$p_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2/2\sigma_i^2)}$$

t-SNE aims to learn a d-dimensional map $\mathbf{y}_1, \ldots, \mathbf{y}_N$ (with $\mathbf{y}_i \in \mathbb{R}^d$) that reflects the similarities p_{ij} as well as possible. To this end, it measures similarities q_{ij} between two points in the map $\mathbf{y}_i$ and $\mathbf{y}_j$, using a very similar approach. Specifically, for $i \neq j$, define q_{ij} as

$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_k \sum_{l\neq k}(1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}$$

https://en.wikipedia.org/wiki/T-distributed_stochastic_neighbor_embedding

Introduction to dimension reduction: UMAP (Uniform Manifold Approximation and Projection)

- Very similar to tSNE, but notable differences:
  - locally adaptive exponential kernel (effectively, different metric for every point)
  - family of curves instead of t distributions
  - binary cross-entropy instead of KL
  - different initialization

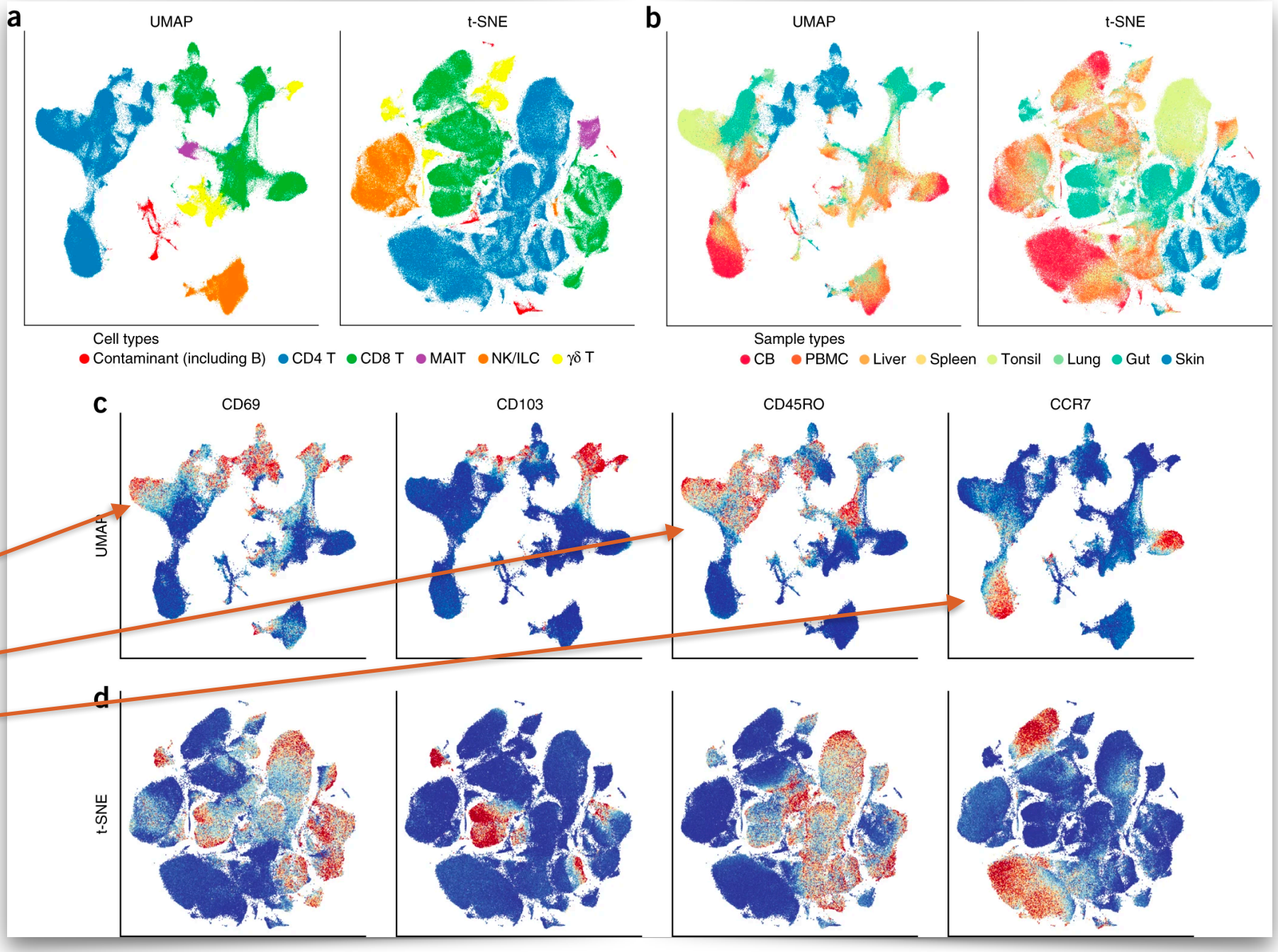
https://towardsdatascience.com/how-exactly-umap-works-13e3040e1668

How do we know a dimension reduction method is working well?

resident memory T cells

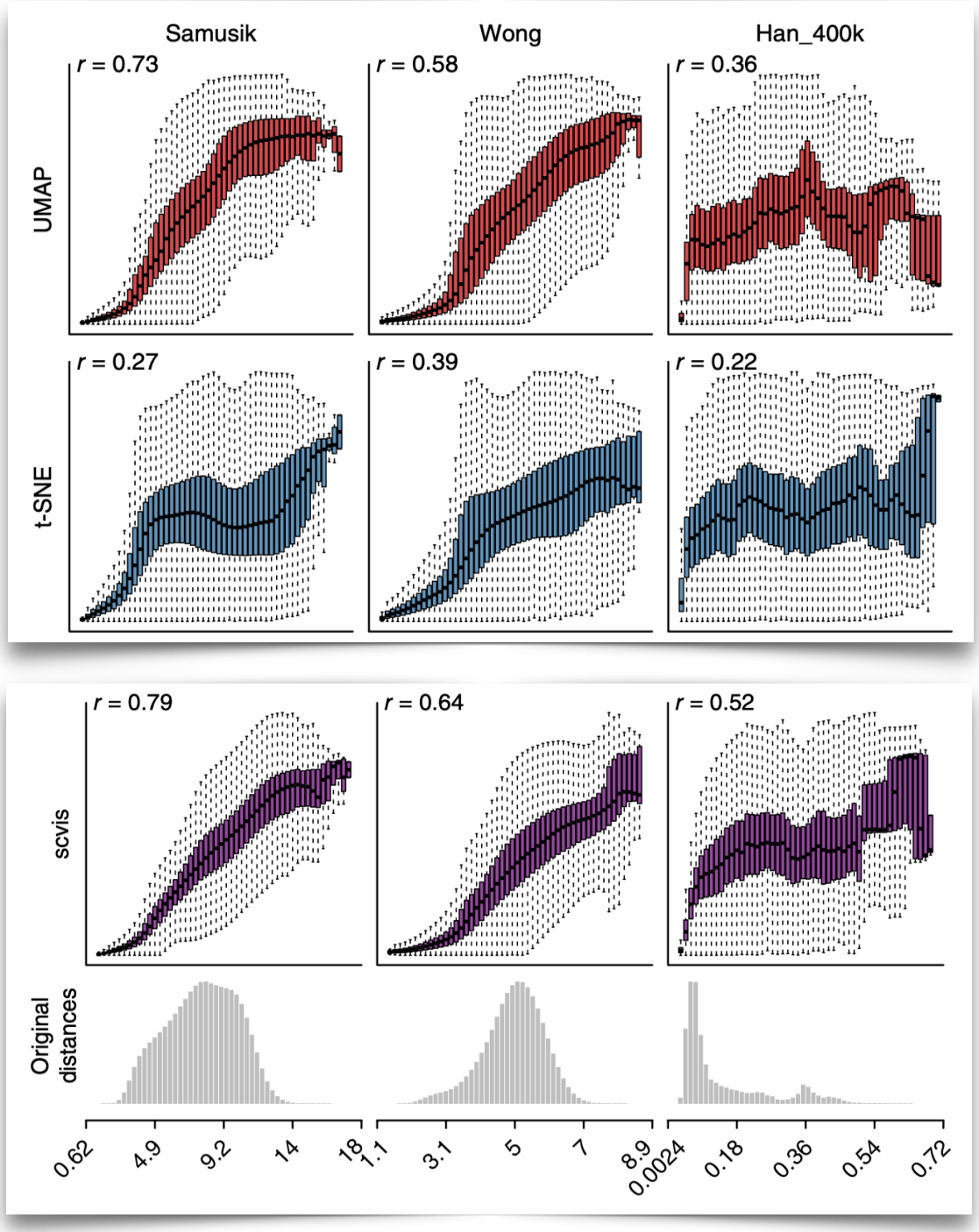
memory T cells

naive T cells



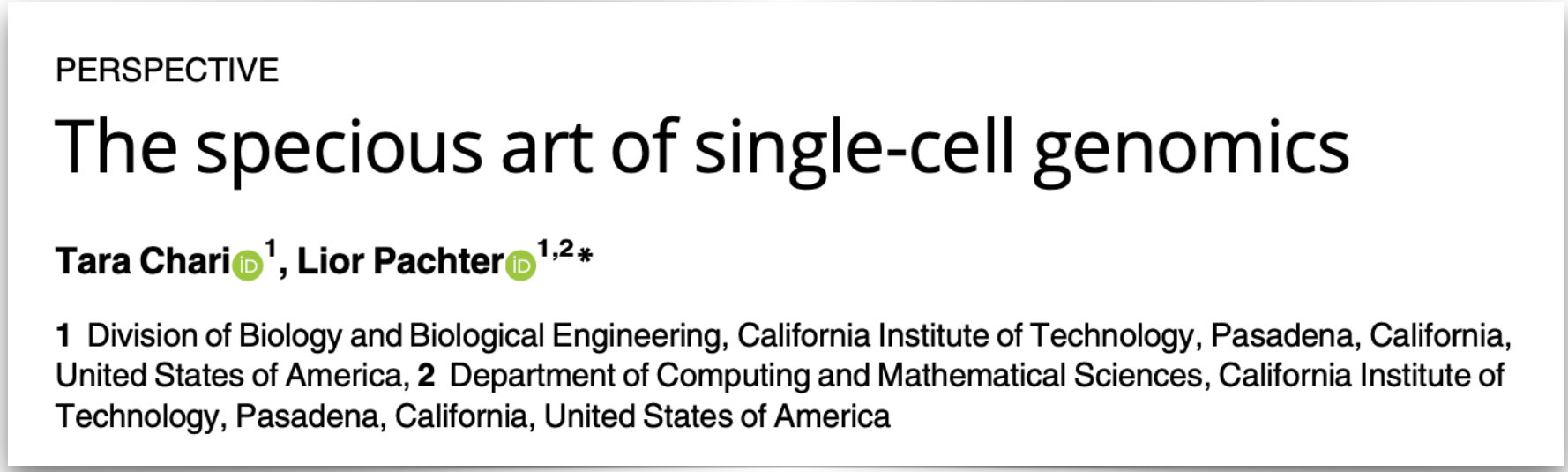
Becht et al. 2019

How do we know a dimension reduction method is working well?



Becht et al. 2019

Careful: dimension reduction can induce distortions



PERSPECTIVE

The specious art of single-cell genomics

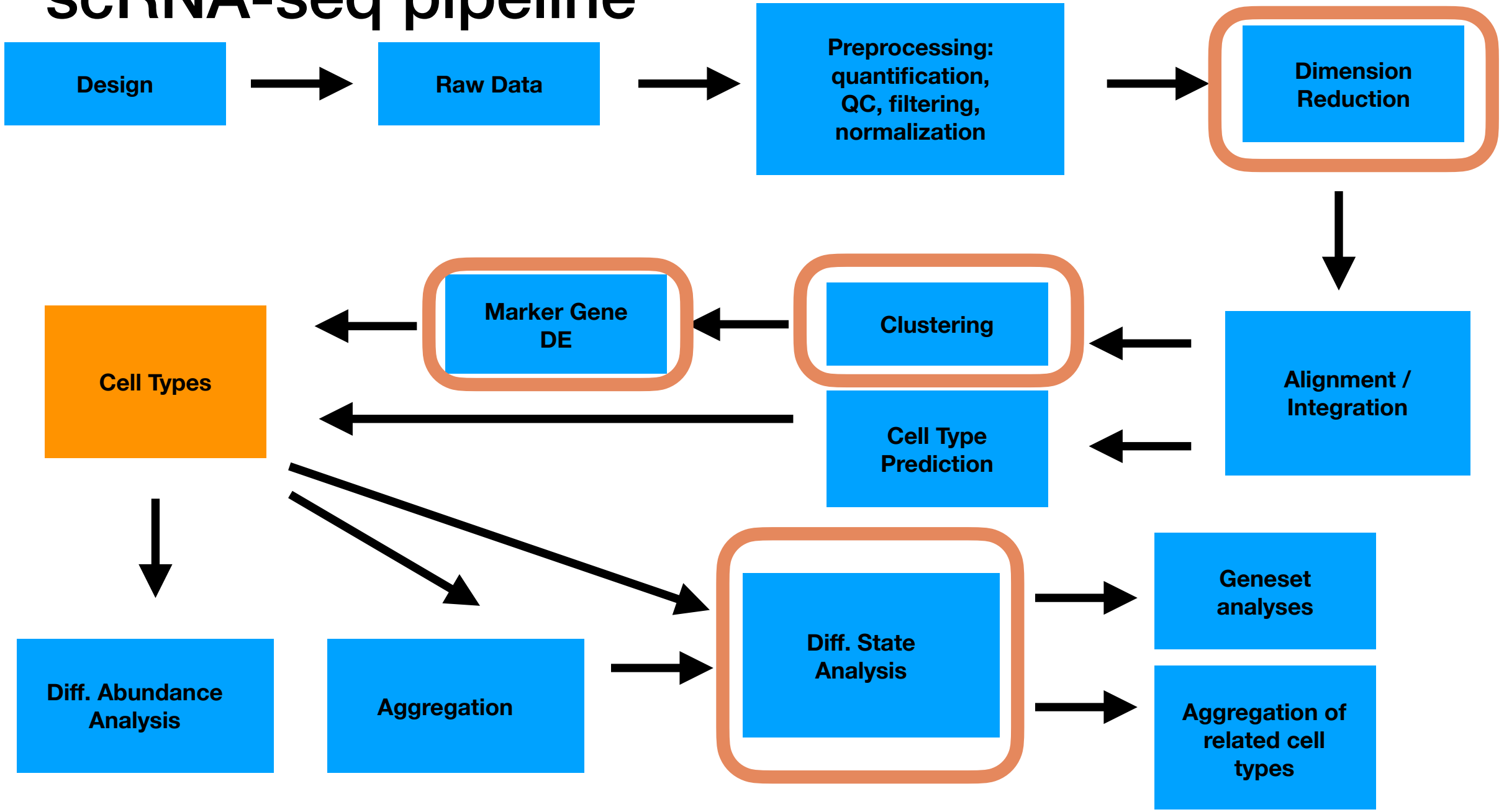
Tara Chari[1], **Lior Pachter**[1,2]*

1 Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California, United States of America, **2** Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, California, United States of America

(published August 2023)

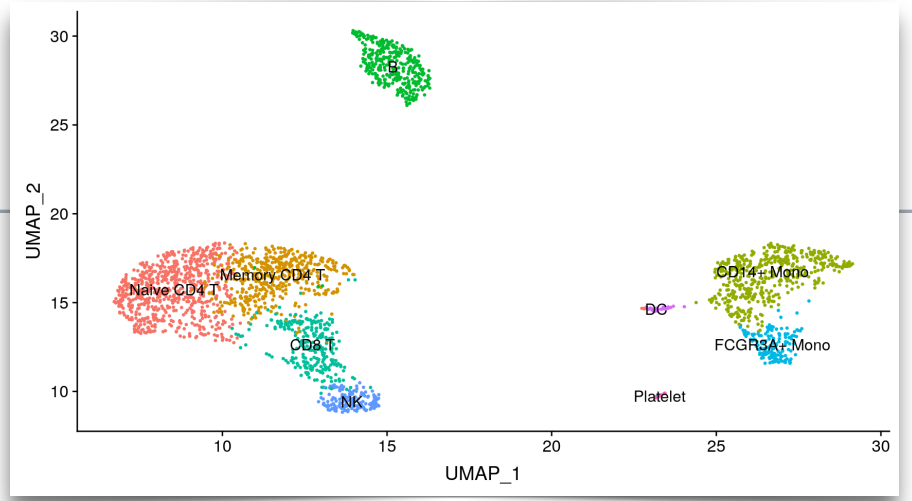
single-cell genomics studies typically begin with reduction to 2 or 3 dimensions to produce "all-in-one" visuals of the data .. these are subsequently used for qualitative and ~~quantitative~~ exploratory analysis .. there is little theoretical support for this practice, and we show that extreme dimension reduction .. inevitably induces significant distortion of high-dimensional datasets.

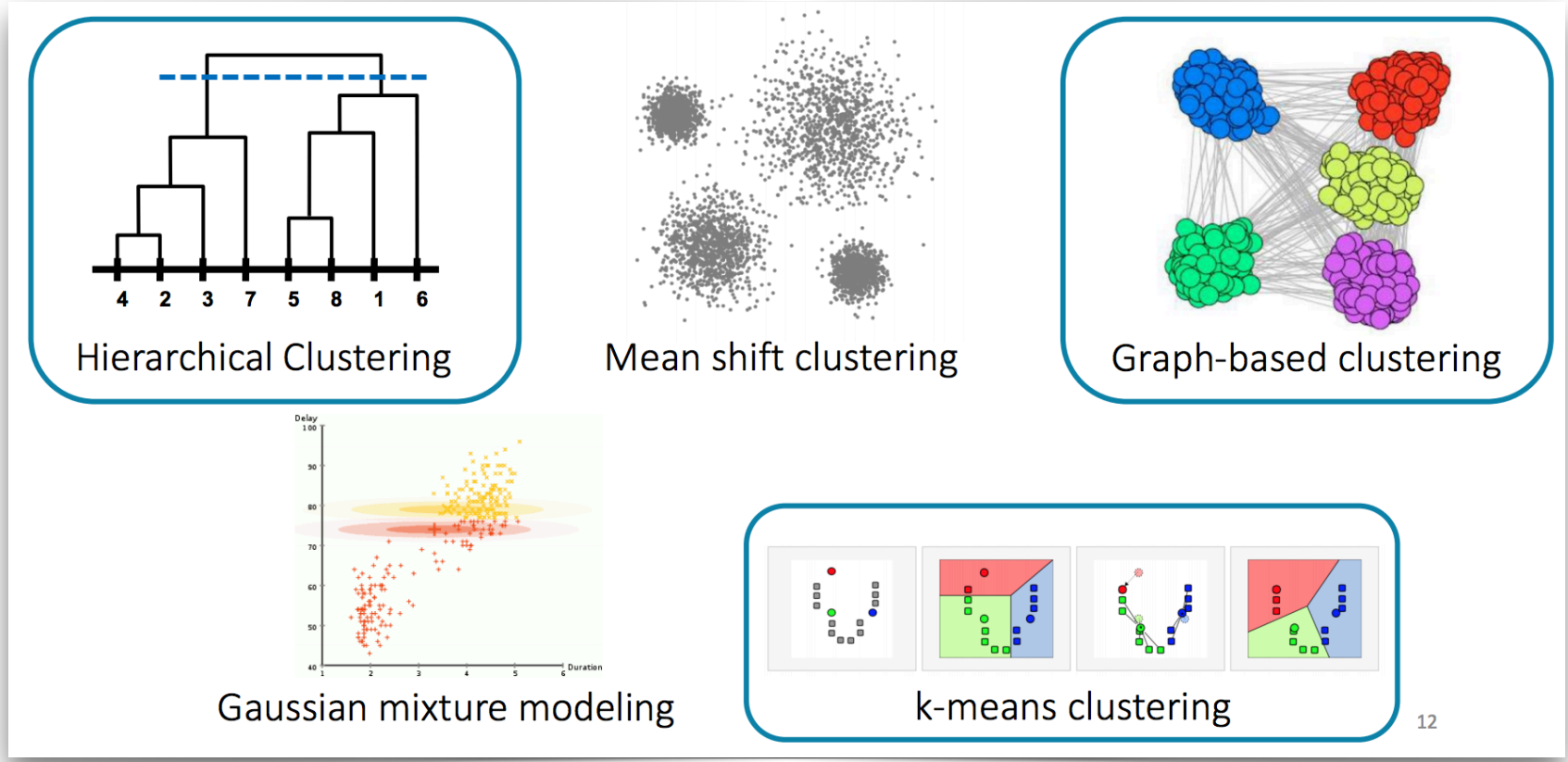
scRNA-seq pipeline



Clustering

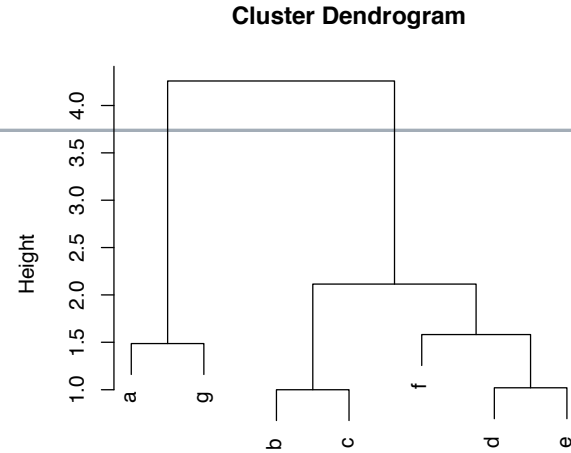
Clustering: an ill-posed problem?





http://en.wikipedia.org/wiki/Hierarchical_clustering

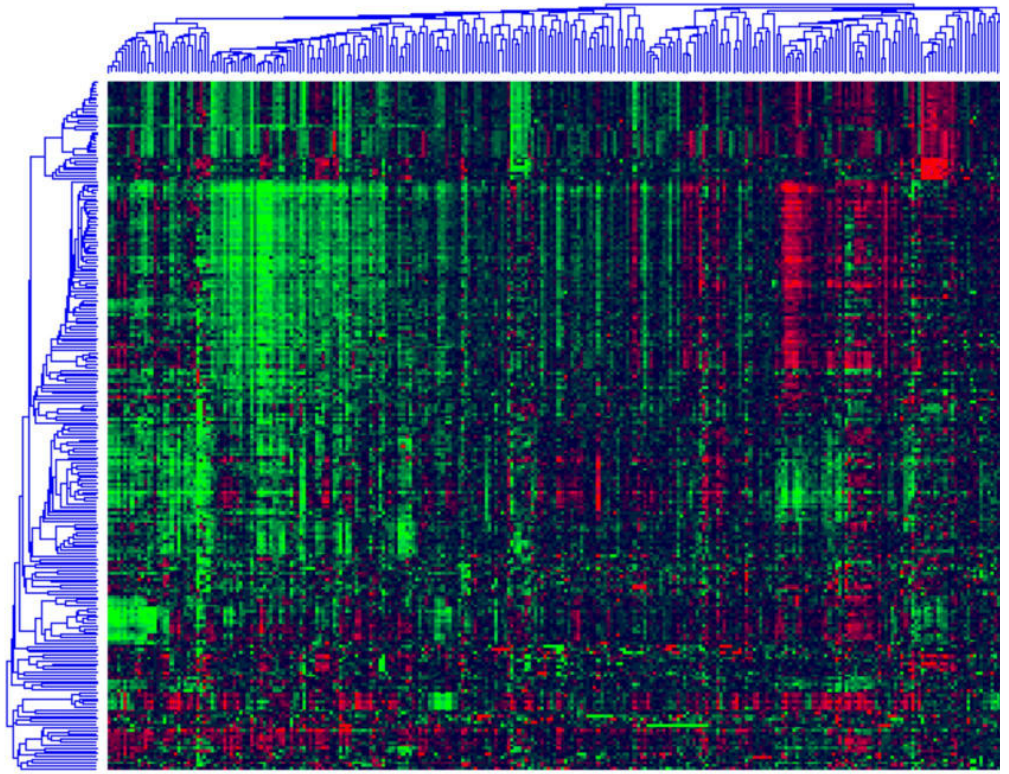
Hierarchical (Agglomerative) Clustering



Divisive (all features start as 1 cluster, then subsequently split) versus Agglomerative (every feature is it's own cluster, then subsequently merged)

Metric: to define how similar any two vectors are.

Linkage: determines how clusters are merged into a tree

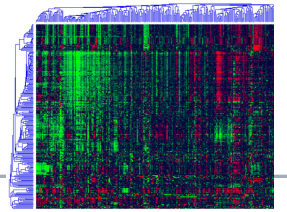


Disadvantages: doesn't scale to large datasets (all pairwise comparisons), may want to cut tree at different heights at different parts of tree

http://en.wikipedia.org/wiki/Euclidean_distance

Euclidean distance:

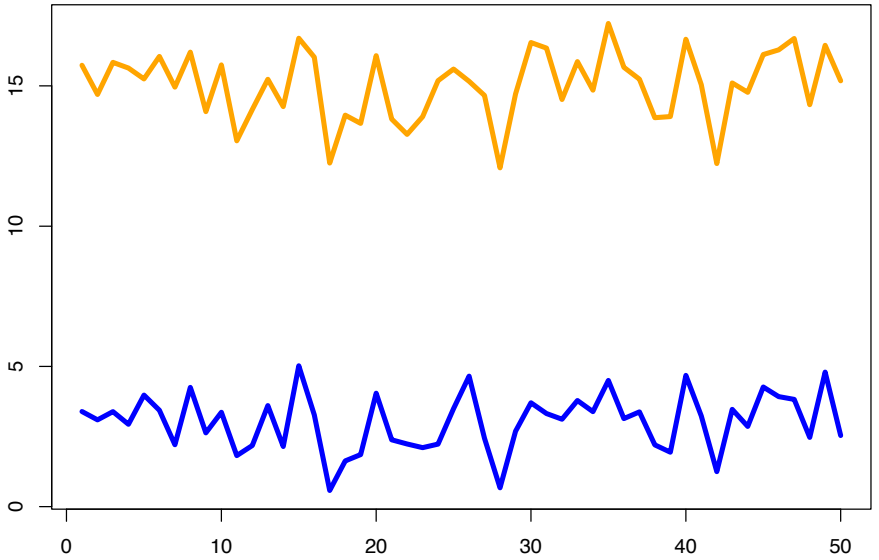
$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1-p_1)^2+(q_2-p_2)^2+\cdots+(q_n-p_n)^2} = \sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}.$$

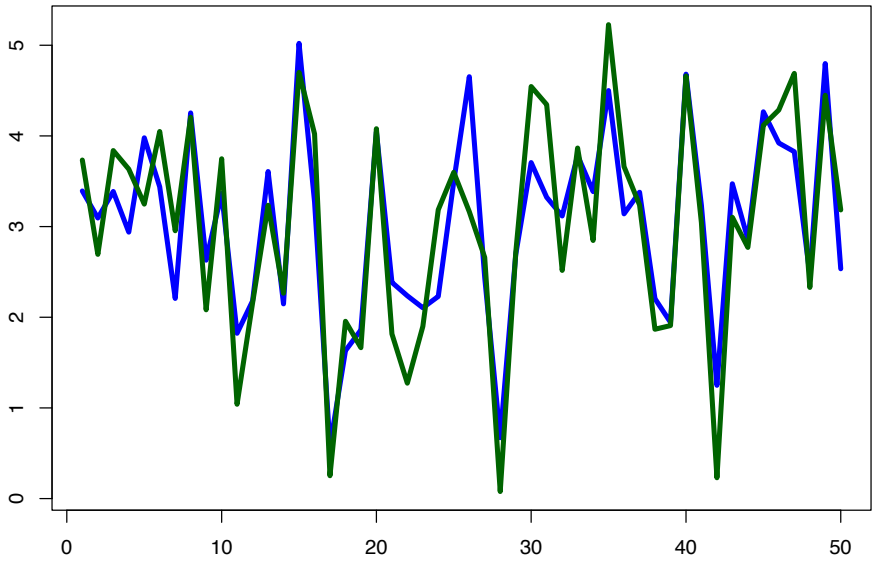


Are these “vectors” similar ?

```
> sqrt(sum((x-(y-12))^2))
[1] 3.926007
> sqrt(sum((x-y)^2))
[1] 84.84028
```

It depends how you define similar.

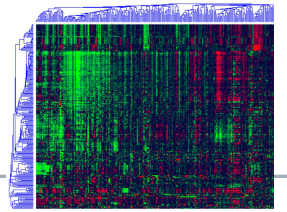




Euclidean distance: 84.84 3.92

http://en.wikipedia.org/wiki/Correlation_and_dependence

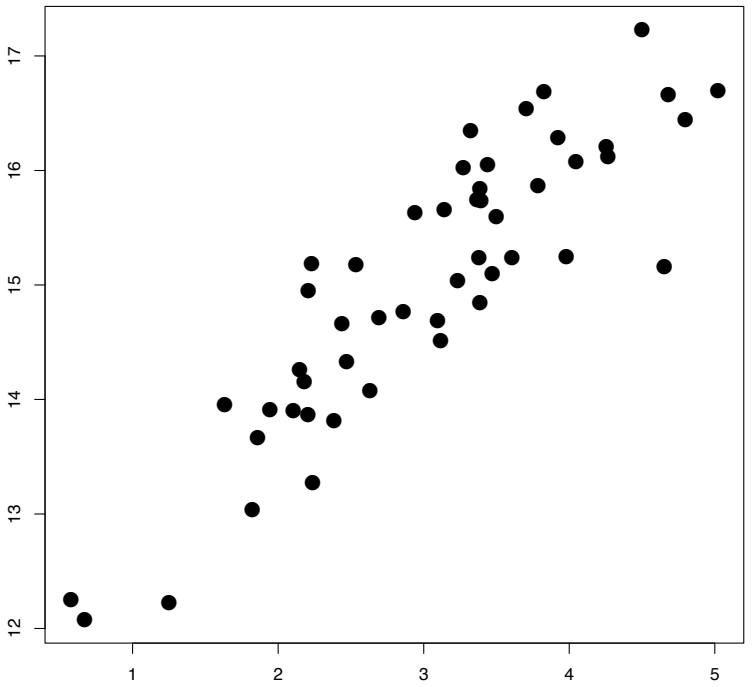
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

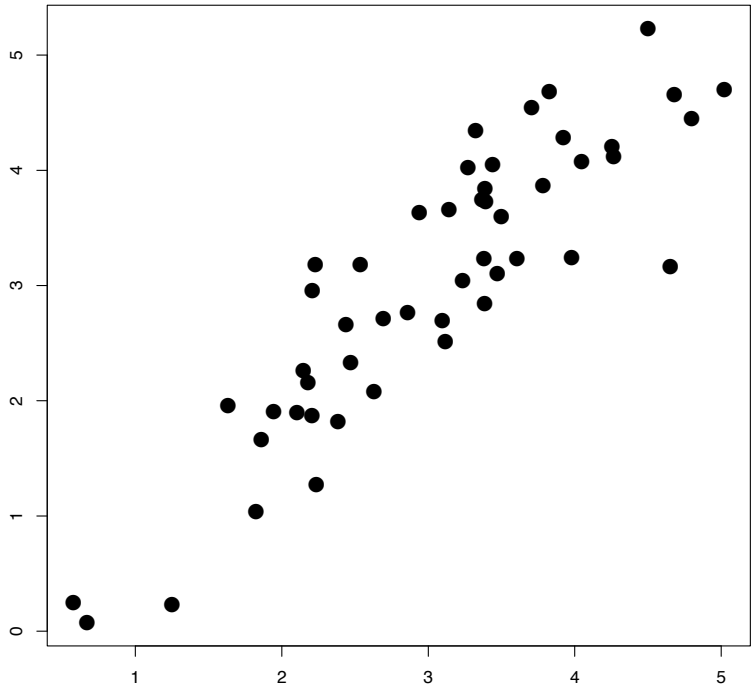


Are these “vectors” similar ?

```
> cor(x,y)
[1] 0.8901139
> cor(x,y-12)
[1] 0.8901139
```

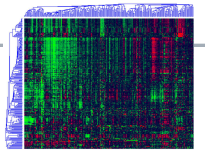
It depends how you define similar.





Correlation: 0.89 0.89

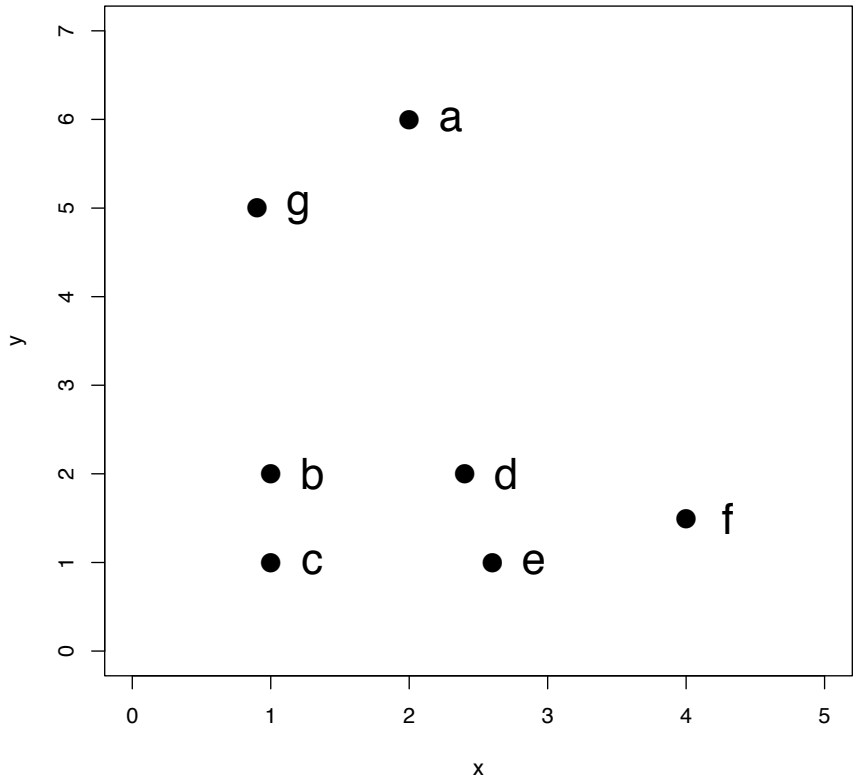
http://en.wikipedia.org/wiki/Hierarchical_clustering



Hierarchical (Agglomerative) Clustering

Start with distances.

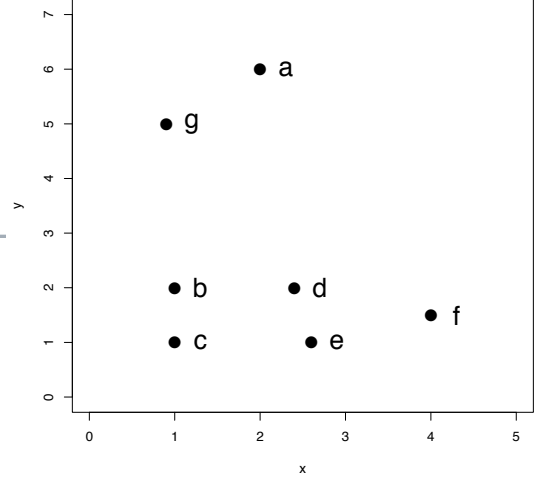
Linkage: determines how clusters are merged into a tree.

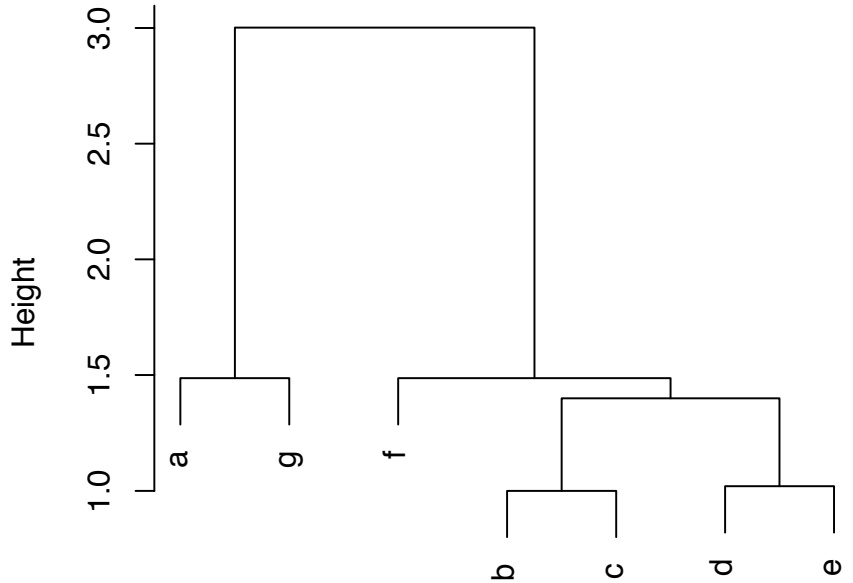


From eyeballing, here is a likely set of merges:

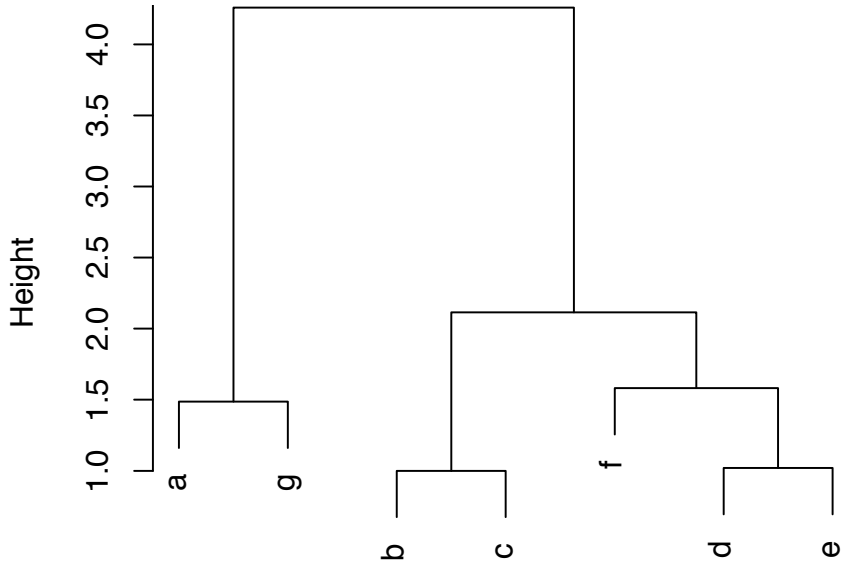
b,c
d,e
a,g,
(d,e),f
(b,c),((d,e),f)
ALL

Different linkages





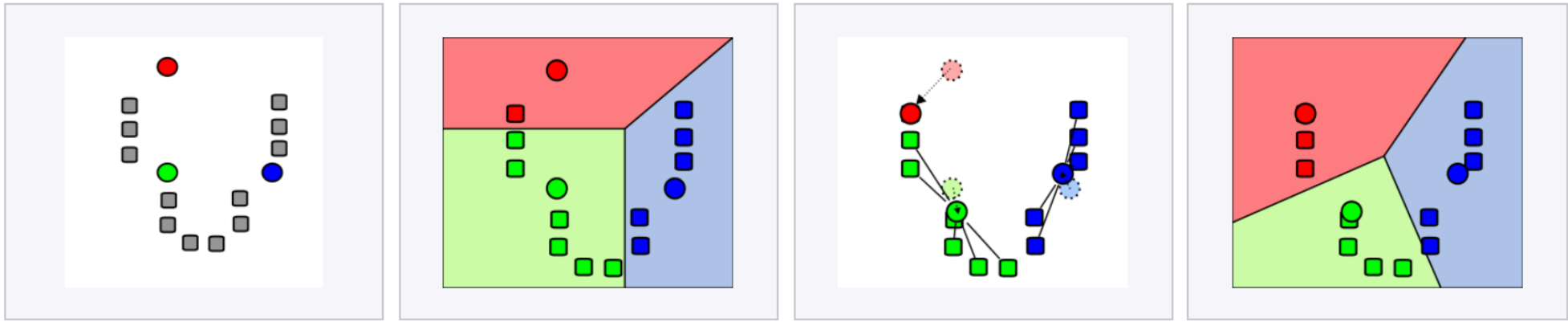
d
hclust (*, "single")



d
hclust (*, "average")

k-means clustering (it has many variations)

Demonstration of the standard algorithm



1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).

2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

3. The centroid of each of the *k* clusters becomes the new mean.

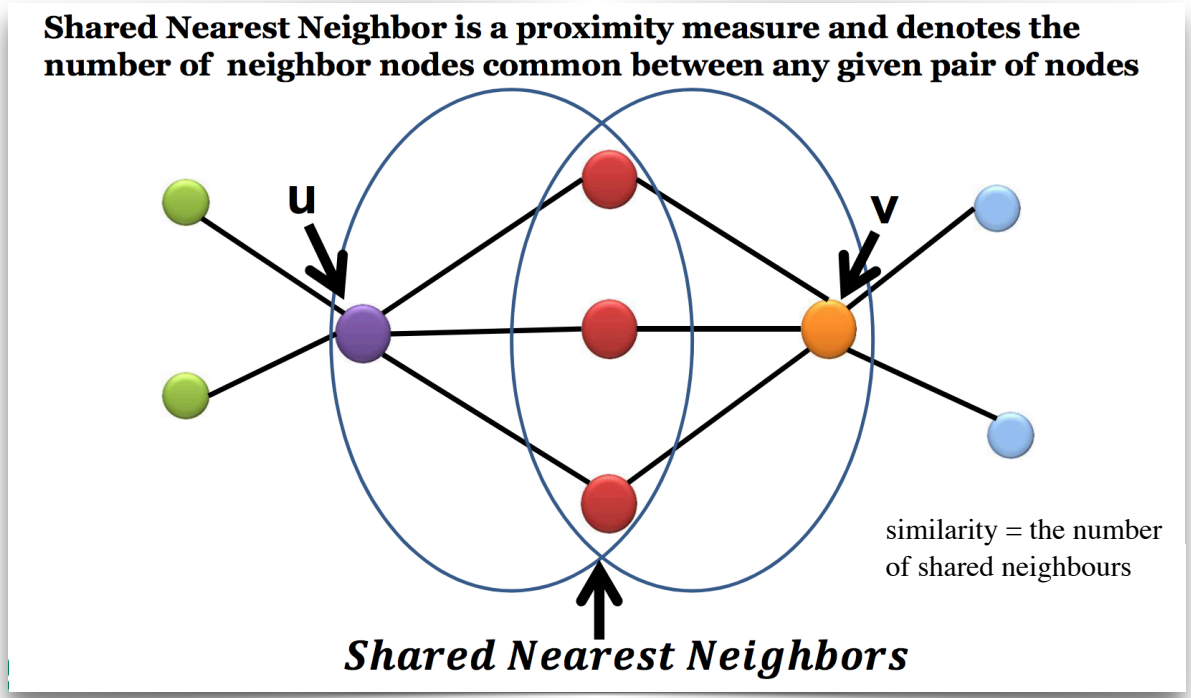
4. Steps 2 and 3 are repeated until convergence has been reached.

Disadvantages: hard to determine k, assumes spherical cluster shapes (depends on distance), different initial selection —> different clustering (typically run multiple times)

Graph-based clustering

- (made popular for scRNA-seq by the Seurat package)
- Euclidean distance is affected by “curse of dimensionality”; typically top PCs are used to help this
- graph is built as kNN (k nearest neighbour) or SNN (shared nearest neighbour)

The k-nearest neighbor graph: two vertices u and v are connected by an edge, if the distance between u and v is among the k-th smallest distances from u to other objects.



https://www.csc2.ncsu.edu/faculty/nfsamato/practical-graph-mining-with-R/slides/pdf/Graph_Cluster_Analysis.pdf

Graph-based clustering

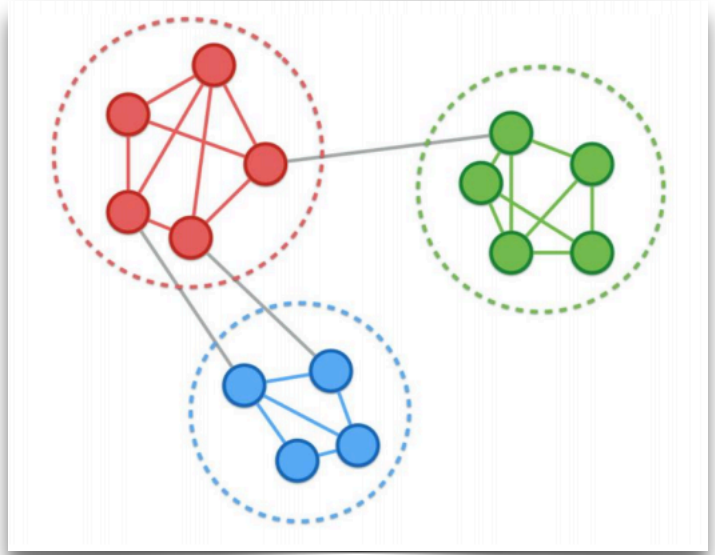
From Louvain to Leiden: guaranteeing well-connected communities

V. A. Traag, L. Waltman & N. J. van Eck

- From graph (originally depends on k), “community detection” to break into clusters
- Want to optimize “modularity” —> i.e. find more edges *inside* the groups than edges linking nodes with rest of the graph.
- Louvain (2008) is a well known heuristic method to optimize modularity

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

A_{ij} = edge weight between node *i* and *j*
k_i = sum of weights attached to
c_i = community of node *i*
m = sum of all edge weights in graph
delta = kronecker delta (1 if $c_i = c_j$)



https://nbisweden.github.io/excelerate-scRNAseq/session-clustering/Clustering.pdf

F1000Research F1000Research 2018, 7:1141 Last updated: 11 SEP 2018

Check for updates

RESEARCH ARTICLE

REVISED **A systematic performance evaluation of clustering methods for single-cell RNA-seq data [version 2; referees: 2 approved]**

Angelo Duò[1,2], Mark D. Robinson[1,2], Charlotte Soneson[1,2]

[1]Institute of Molecular Life Sciences, University of Zurich, Zurich, 8057, Switzerland
[2]SIB Swiss Institute of Bioinformatics, Zurich, 8057, Switzerland

scRNA-seq: No “best” clustering algorithm

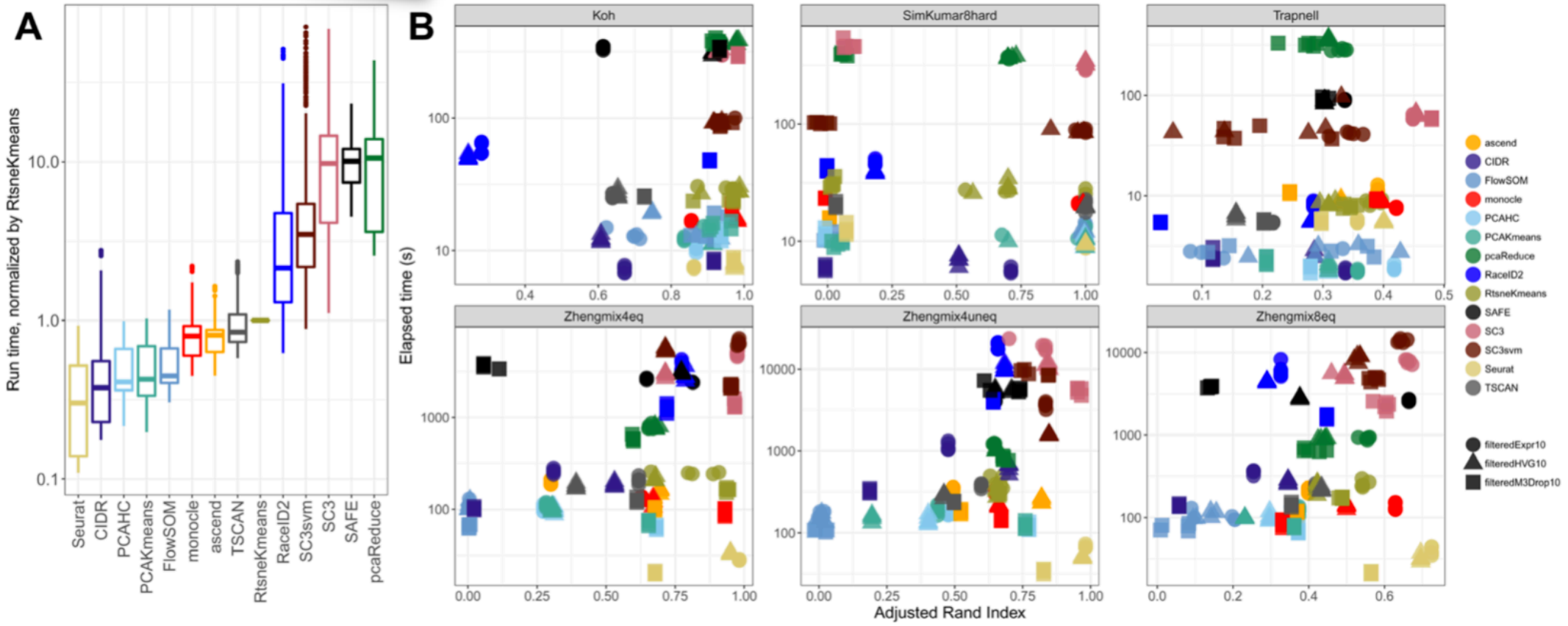
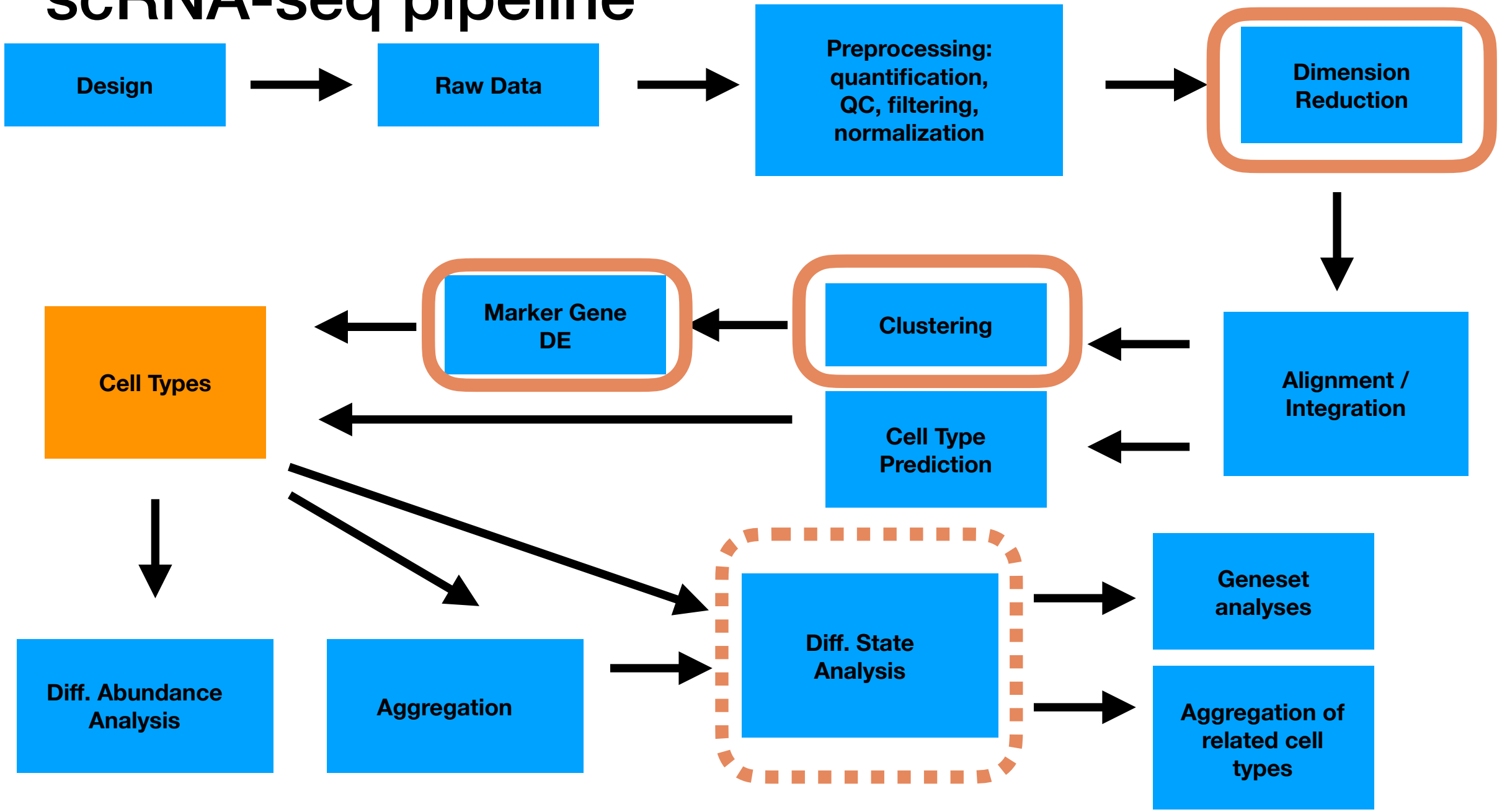


Figure 2. (**A**) Normalized run times, using `RtsneKmeans` as the reference method, across all data set instances and number of clusters. (**B**) Run time versus performance (ARI) for a subset of data sets and filterings, for the true number of clusters.

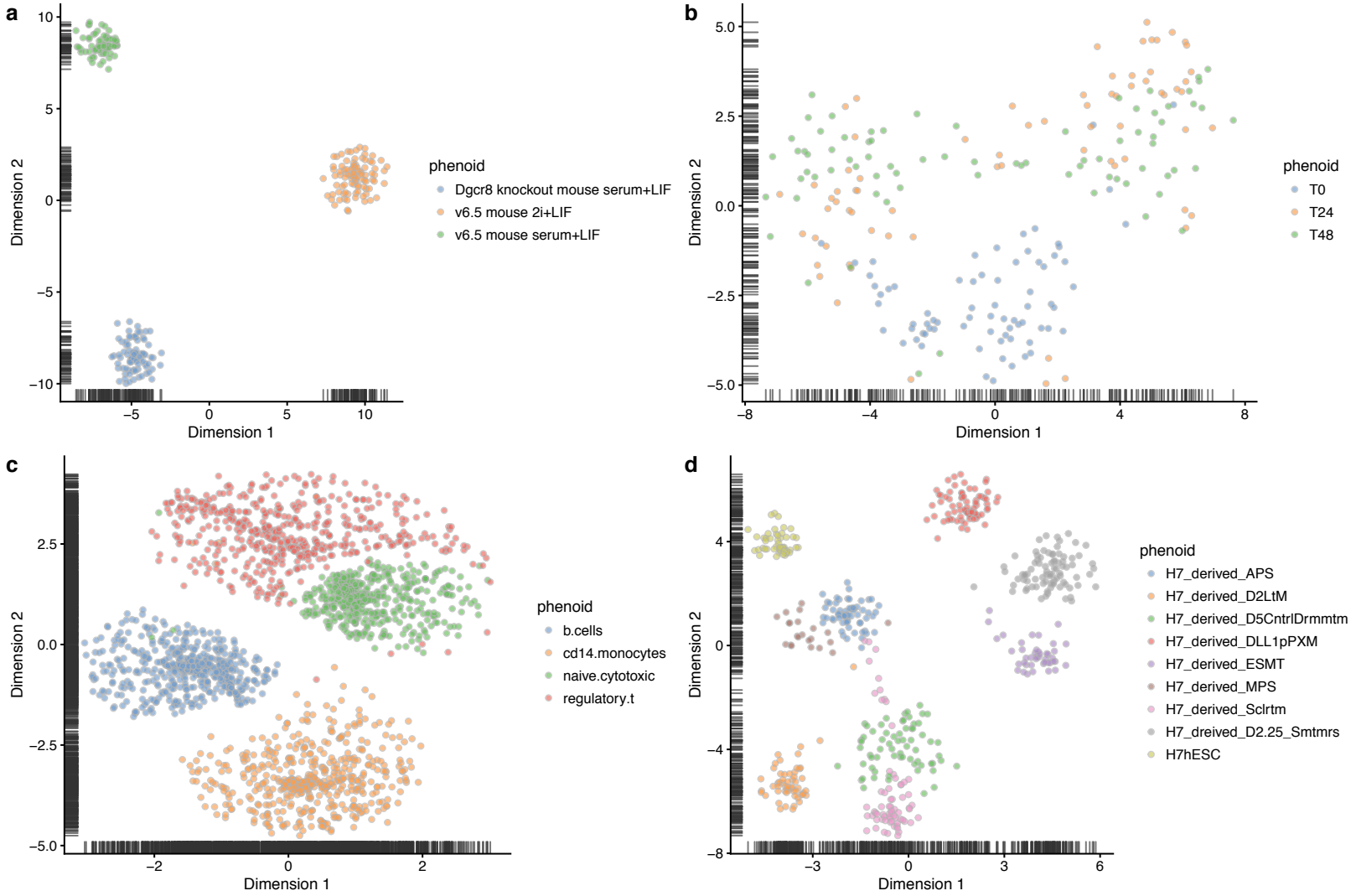
(Marker gene) Differential expression of single cell RNA-seq data

scRNA-seq pipeline



After clustering, how to find cell type markers ?

- here, **predefined groups**: range of difficulty



Differential expression: zero inflation / model dropout, mixture models, etc.

MAST

mixture model

hurdle model

SCDE

Single-cell RNA-seq hurdle model

We model the $\log_2(\text{TPM}+1)$ expression matrix as a two-part generalized regression model. The gene expression rate was modeled using logistic regression and, conditioning on a cell expressing the gene, the expression level was modeled as Gaussian.

Given normalized, possibly thresholded (see Additional file 1), scRNA-seq expression $Y=[y_{ig}]$, the rate of expression and the level of expression for the expressed cells are modeled conditionally independent for each gene g. Define the indicator $Z=[z_{ig}]$, indicating whether gene g is expressed in cell i (i.e., $z_{ig}=0$ if $y_{ig}=0$ and $z_{ig}=1$ if $y_{ig}>0$). We fit logistic regression models for the discrete variable Z and a Gaussian linear model for the continuous variable $(Y \mid Z=1)$ independently, as follows:

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

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

The regression coefficients of the discrete component are regularized using a Bayesian approach as implemented in the *bayesglm* function of the *arm* R package, which uses weakly informative priors [30] to provide sensible estimates under linear separation (See Additional file 1 for details). We also perform regularization of the continuous model variance parameter, as described below, which helps to increase the robustness of gene-level differential expression analysis when a gene is only expressed in a few cells.

Differential expression analysis. With a Bayesian approach, the posterior probability of a gene being expressed at an average level x in a subpopulation of cells S was determined as an expected value (E) according to

$$p_S(x) = E\left[\prod_{c\in B} p(x\,|\,r_c,\Omega_c)\right]$$

where B is a bootstrap sample of S, and $p(x|r_c,\Omega_c)$ is the posterior probability for a given cell c, according to

$$p(x\,|\,r_c,\Omega_c) = p_d(x)p_{Poisson}(x) + (1-p_d(x))p_{NB}(x\,|\,r_c)$$

where p_d is the probability of observing a dropout event in cell c for a gene expressed at an average level x in S, $p_{Poisson}(x)$ and $p_{NB}(x|r_c)$ are the probabilities of observing expression magnitude of r_c in case of a dropout (Poisson) or successful amplification (NB) of a gene expressed at level x in cell c, with the parameters of the distributions determined by the Ω_c fit. For the differential expression analysis, the posterior probability that the gene shows a fold expression difference of f between subpopulations S and G was evaluated as

$$p(f) = \sum_{x\in X} p_S(x)p_G(fx)$$

where x is the valid range of expression levels. The posterior distributions were renormalized to unity, and an empirical P value was determined to test for significance of expression difference.

Differential distributions

Equivalent **E**xpression

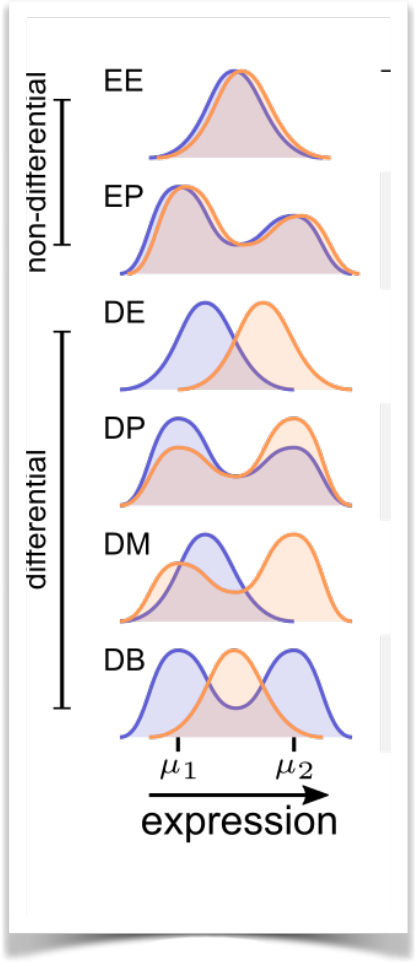
Equivalent **P**roportions

Differential **E**xpression

Differential **P**roportions

Differential **M**odality

Both, **D**ifferential modality & component means



(shift in means)

(identical means)

Korthauer *et al. Genome Biology* (2016) 17:222
DOI 10.1186/s13059-016-1077-y

Genome Biology

METHOD

Open Access

A statistical approach for identifying differential distributions in single-cell RNA-seq experiments

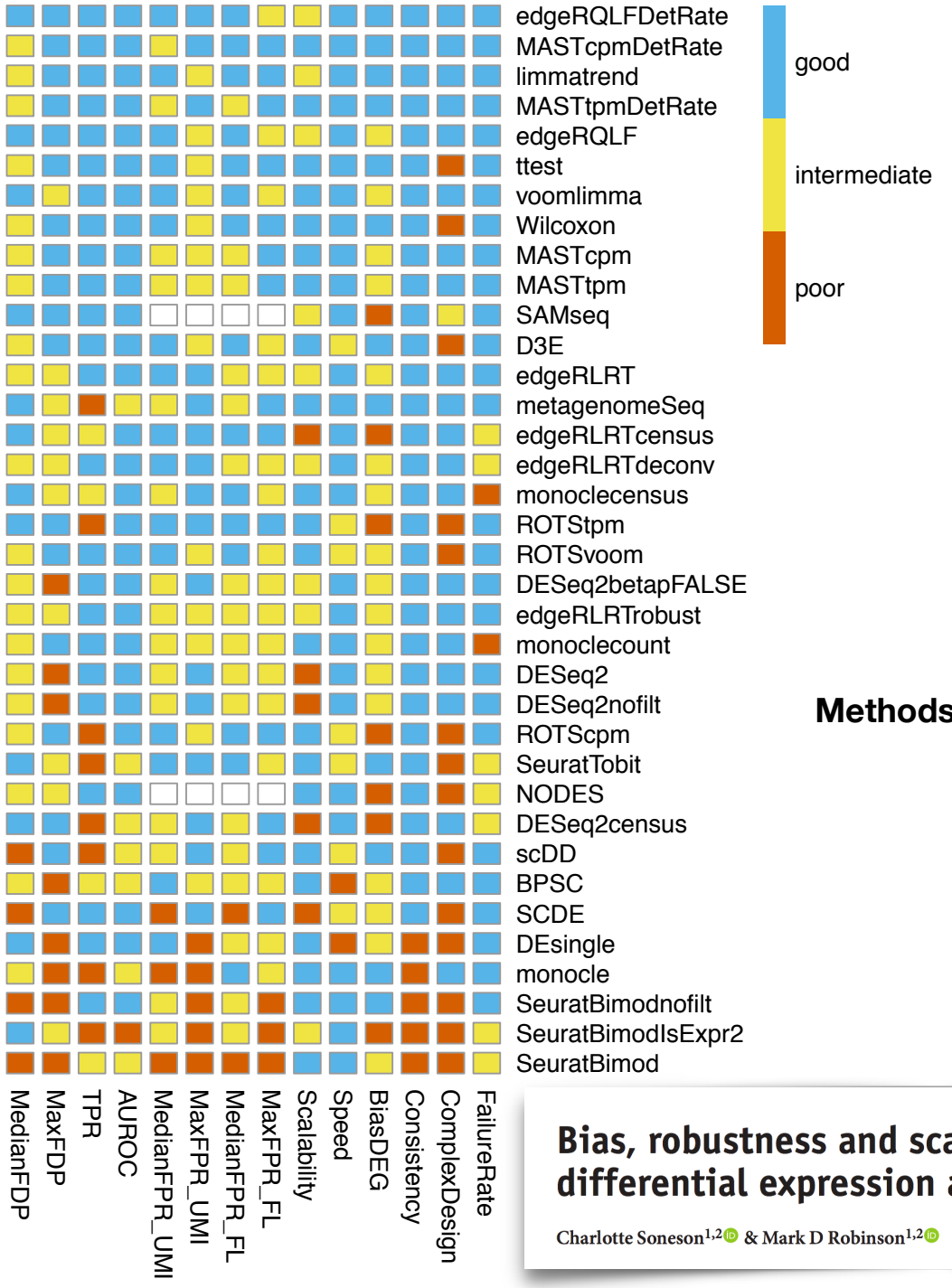
Keegan D. Korthauer[1,2], Li-Fang Chu[3], Michael A. Newton[4,5], Yuan Li[5], James Thomson[3,6,7], Ron Stewart[3] and Christina Kendziorski[4,5*]

Punchline

Several methods work well, including a mix of single-cell-specific and bulk methods

t-test and Wilcoxon perform surprisingly well

“we found that bulk RNA-seq analysis methods do not generally perform worse than those developed specifically for scRNA-seq”



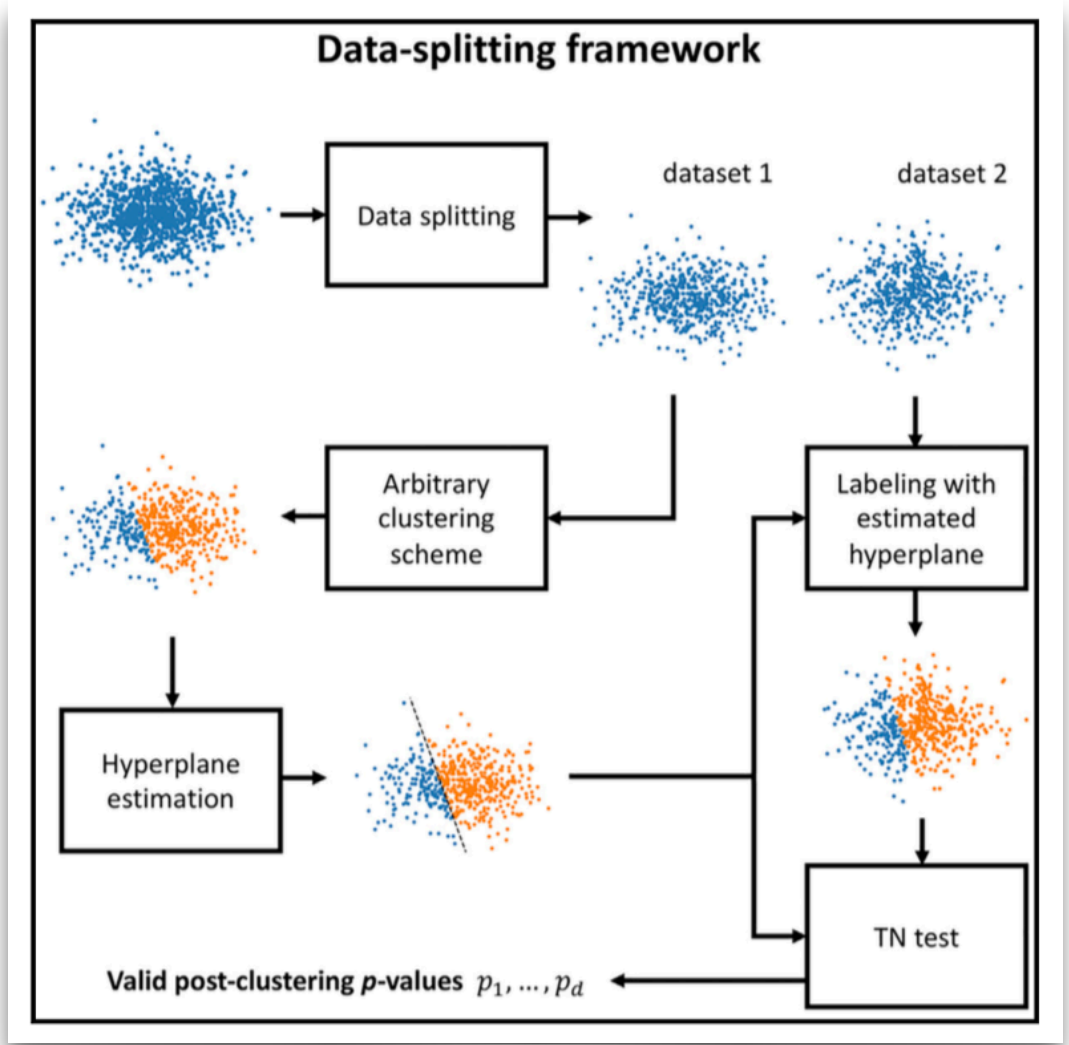
Bias, robustness and scalability in single-cell differential expression analysis

Charlotte Soneson[1,2] & Mark D Robinson[1,2]

Statistical issue lurking here: clustering and then testing differences between clusters leads to invalid P-values

Valid Post-clustering Differential Analysis for Single-Cell RNA-Seq

Graphical Abstract



Authors

Jesse M. Zhang, Govinda M. Kamath, David N. Tse

Often what we do: find DE between clusters; ignore the magnitude of the P-value

Two genes shown below have roughly same magnitude of DE (e.g., using ANOVA F-statistic on log-normalized-counts) —> Which is the better “marker gene”? Which has high entropy? Which has low entropy?