



Projects

Reminders:

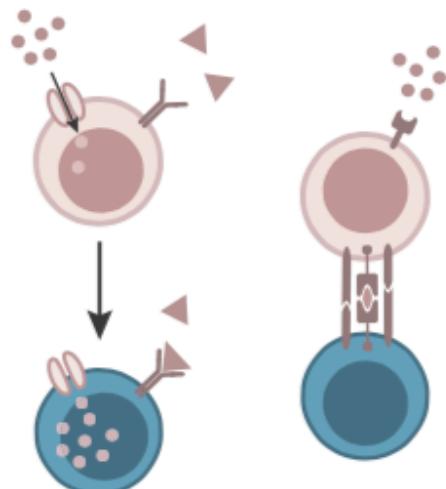
- project “plan” by **1.12.2025** (a few bullet points)
- due **9.01.2026** 18.00

—> When you have topic ready, organize a private Slack channel (group + Mark, Hubert, optionally collaborators).

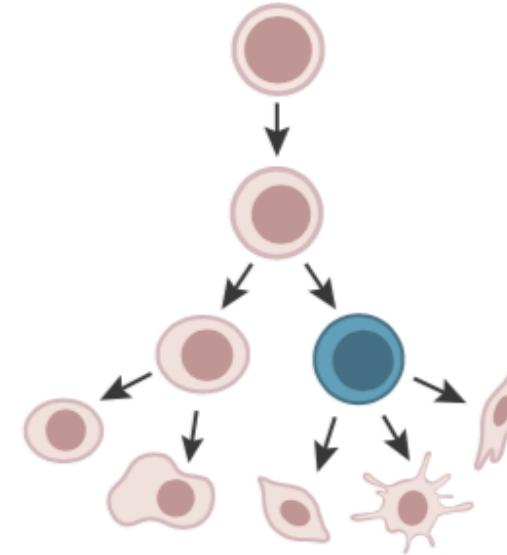
—> Private GitHub repo (of same name) will be made (Mark has a script); submit projects to this repo (as with exercises)

a

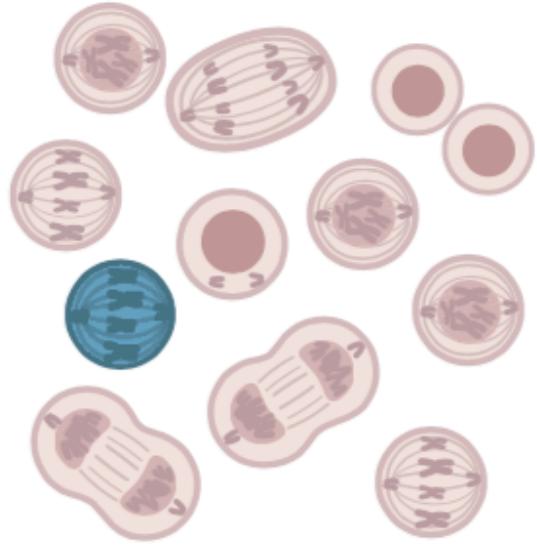
Environmental stimuli



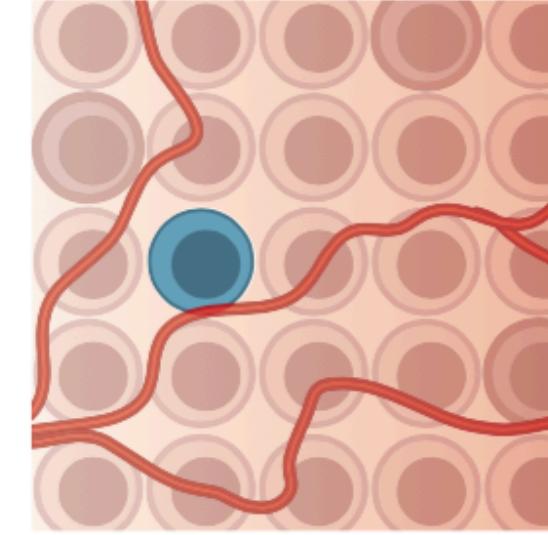
Cell development



Cell cycle



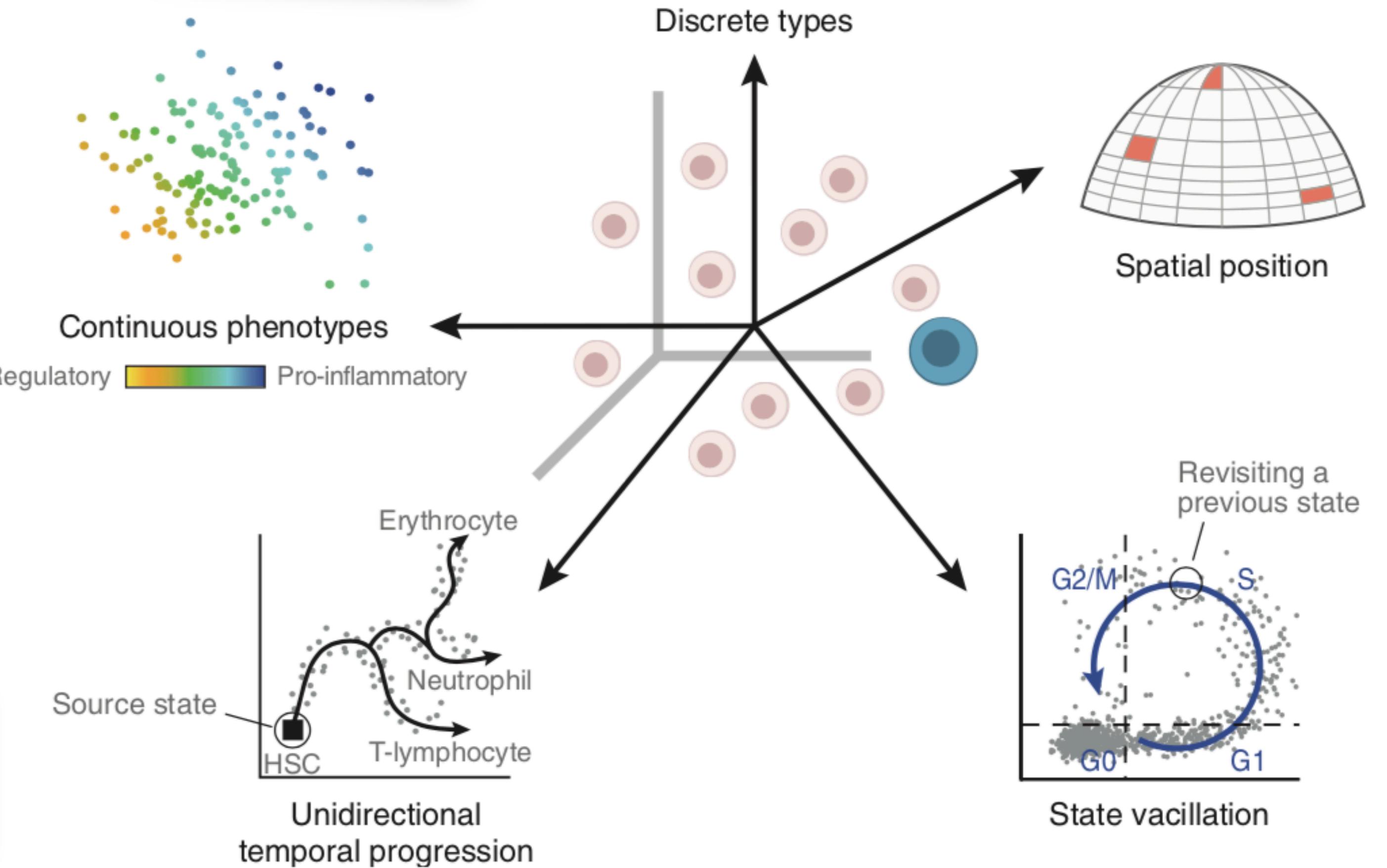
Spatial context



Applications

Revealing the vectors of cellular identity with single-cell genomics

Allon Wagner¹, Aviv Regev^{2,3,5} & Nir Yosef^{1,4,5}



Revealing the vectors of cellular identity with single-cell genomics

Allon Wagner¹, Aviv Regev^{2,3,5} & Nir Yosef^{1,4,5}

Perspective

Defining cell types and states with single-cell genomics

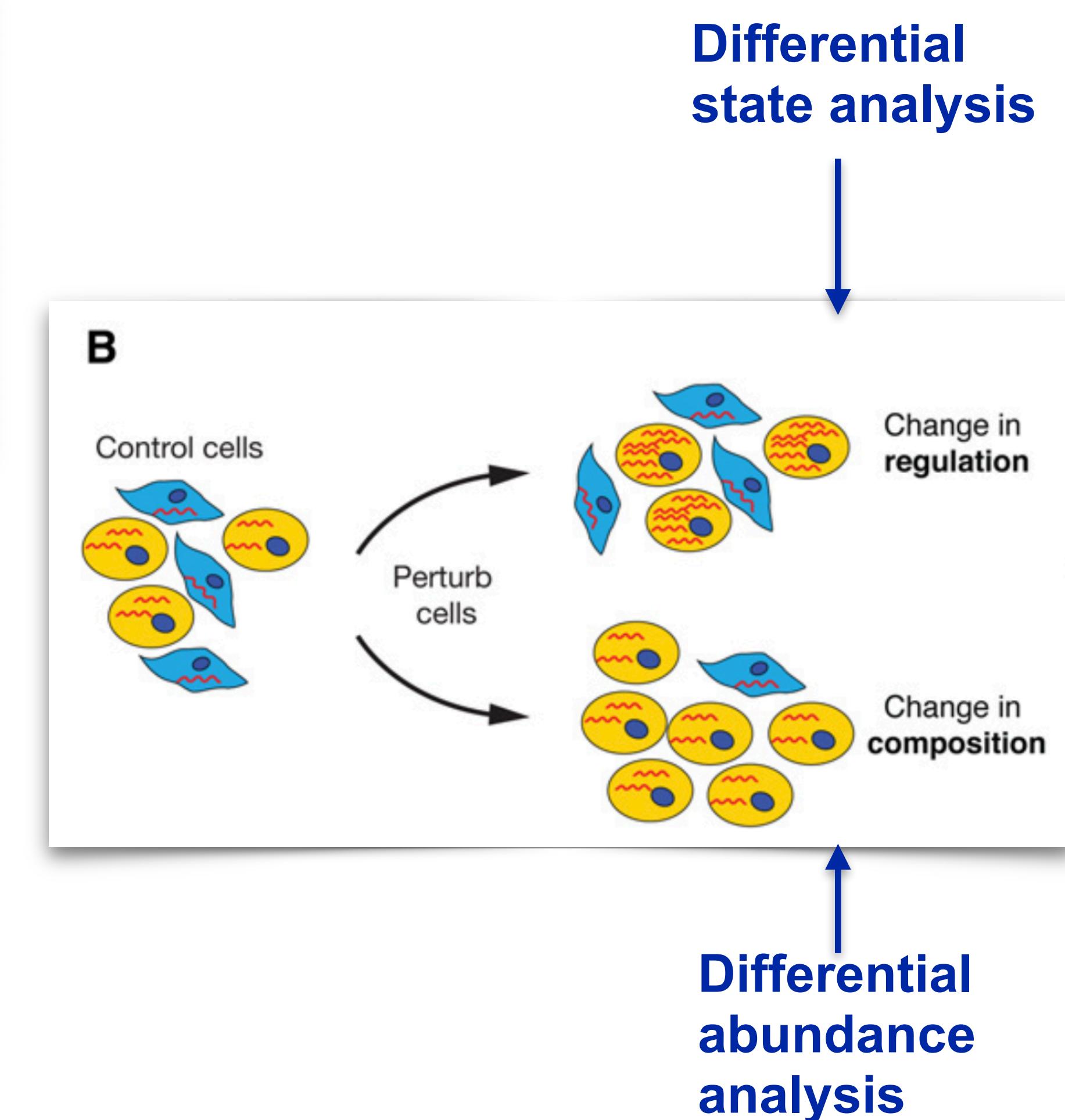
Cole Trapnell

Department of Genome Sciences, University of Washington, Seattle, Washington 98105, USA

Box 1 The many facets of a cell's identity

We define a cell's identity as the outcome of the instantaneous intersection of all factors that affect it. We refer to the more permanent aspects in a cell's identity as its type (e.g., a hepatocyte typically cannot turn into a neuron) and to the more transient elements as its state. Cell types are often organized in a hierarchical

Type: more permanent
State: more transient



Flow cytometry

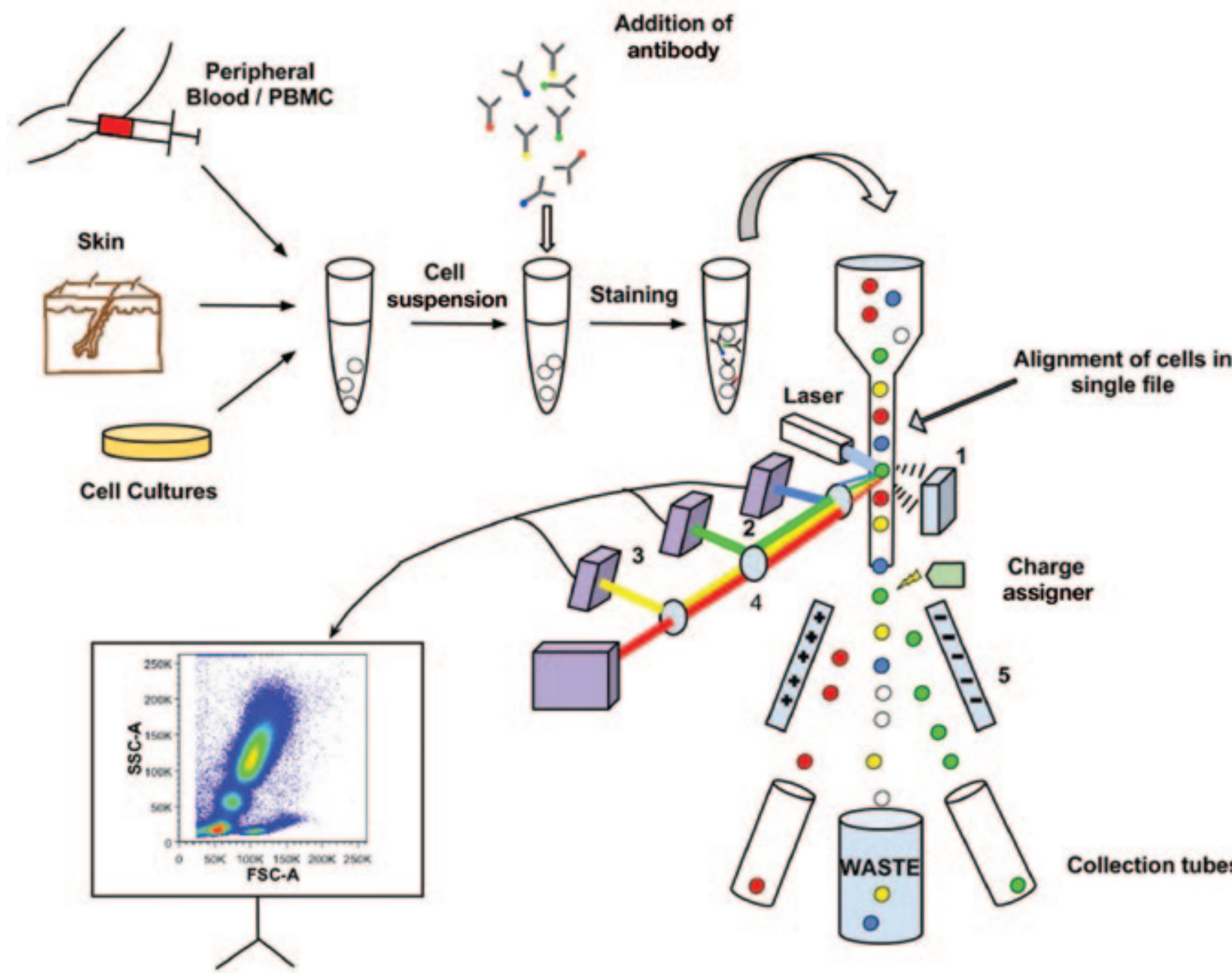
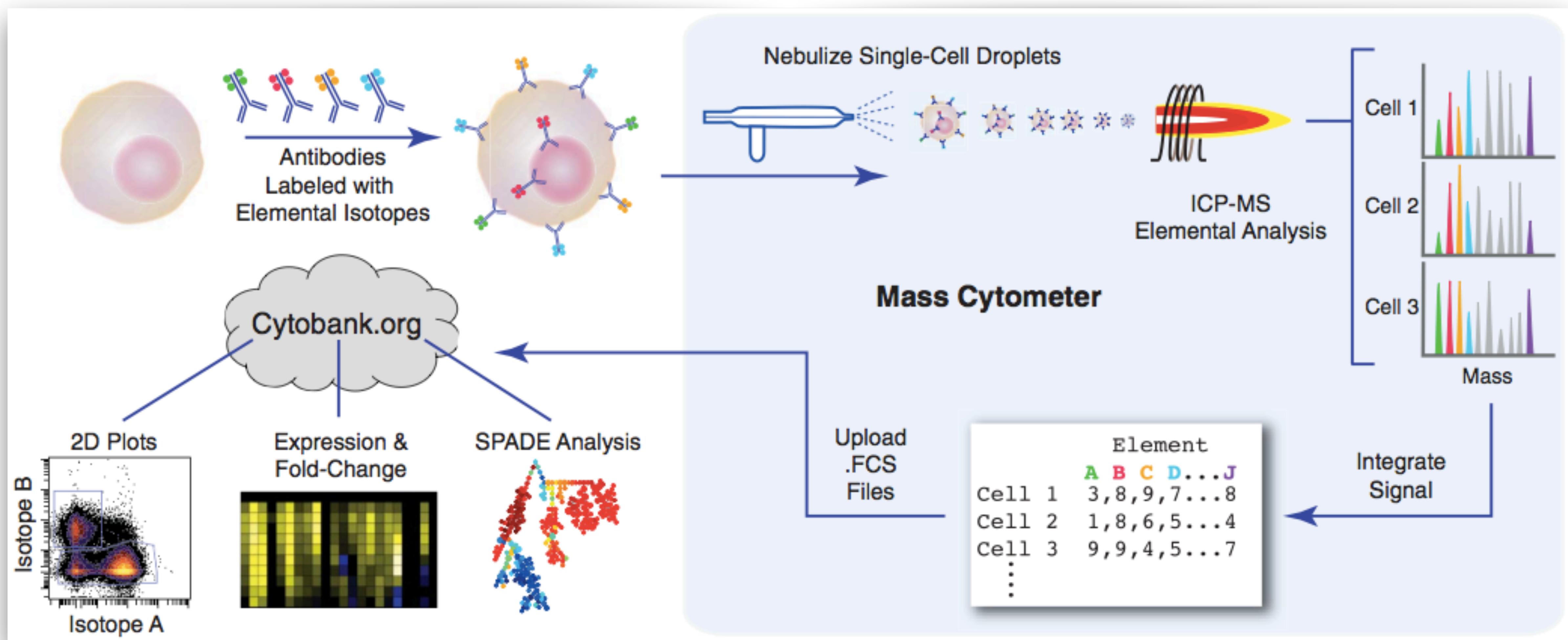
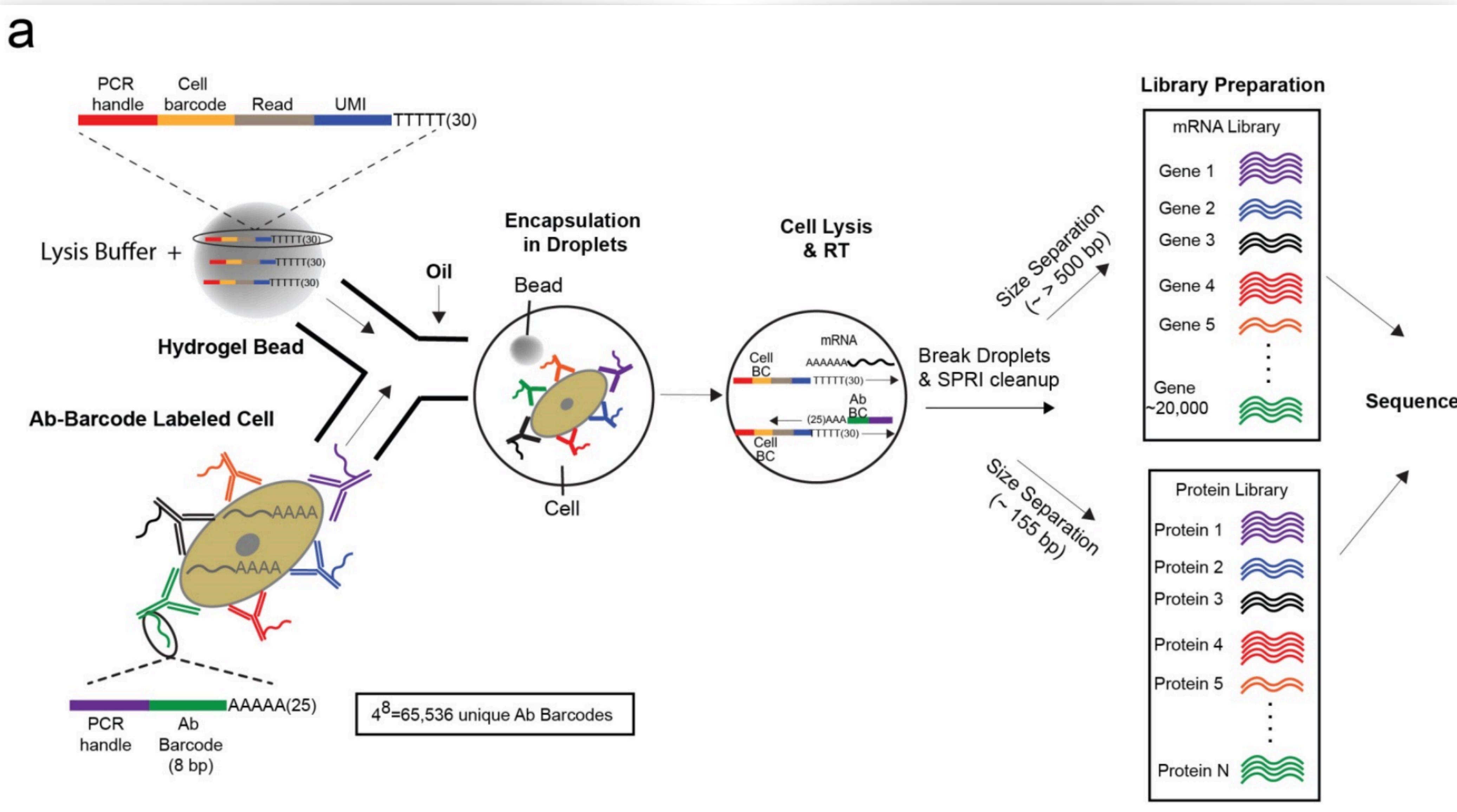


Figure 1. Schematic representation of a flow cytometer. For details please see text. (1) Forward-scatter detector, (2) side-scatter detector, (3) fluorescence detector, (4) filters and mirrors, and (5) charged deflection plates.

Mass cytometry



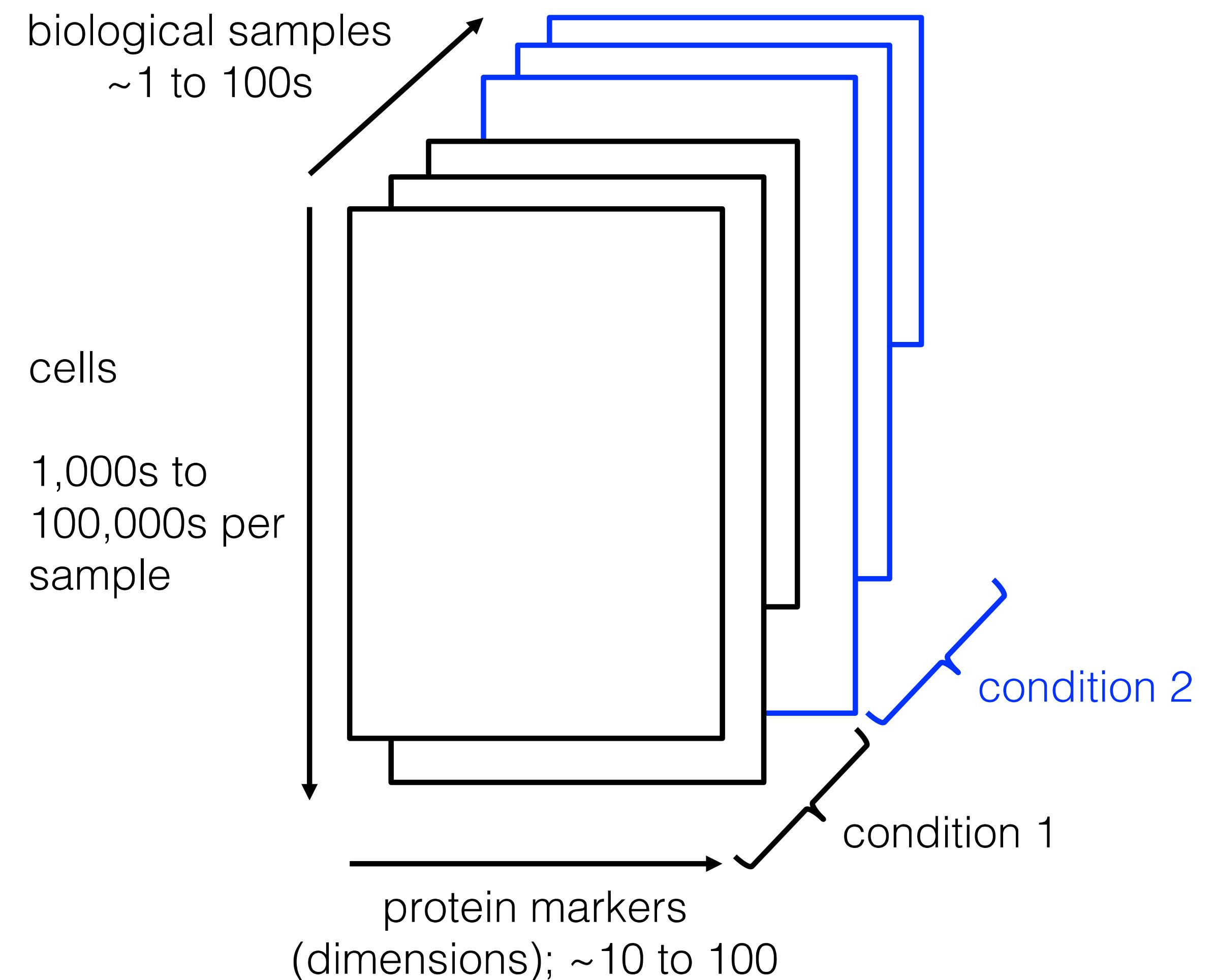
REAP-seq / CITE-seq



Data structure and differential analysis

Two types of differential analysis

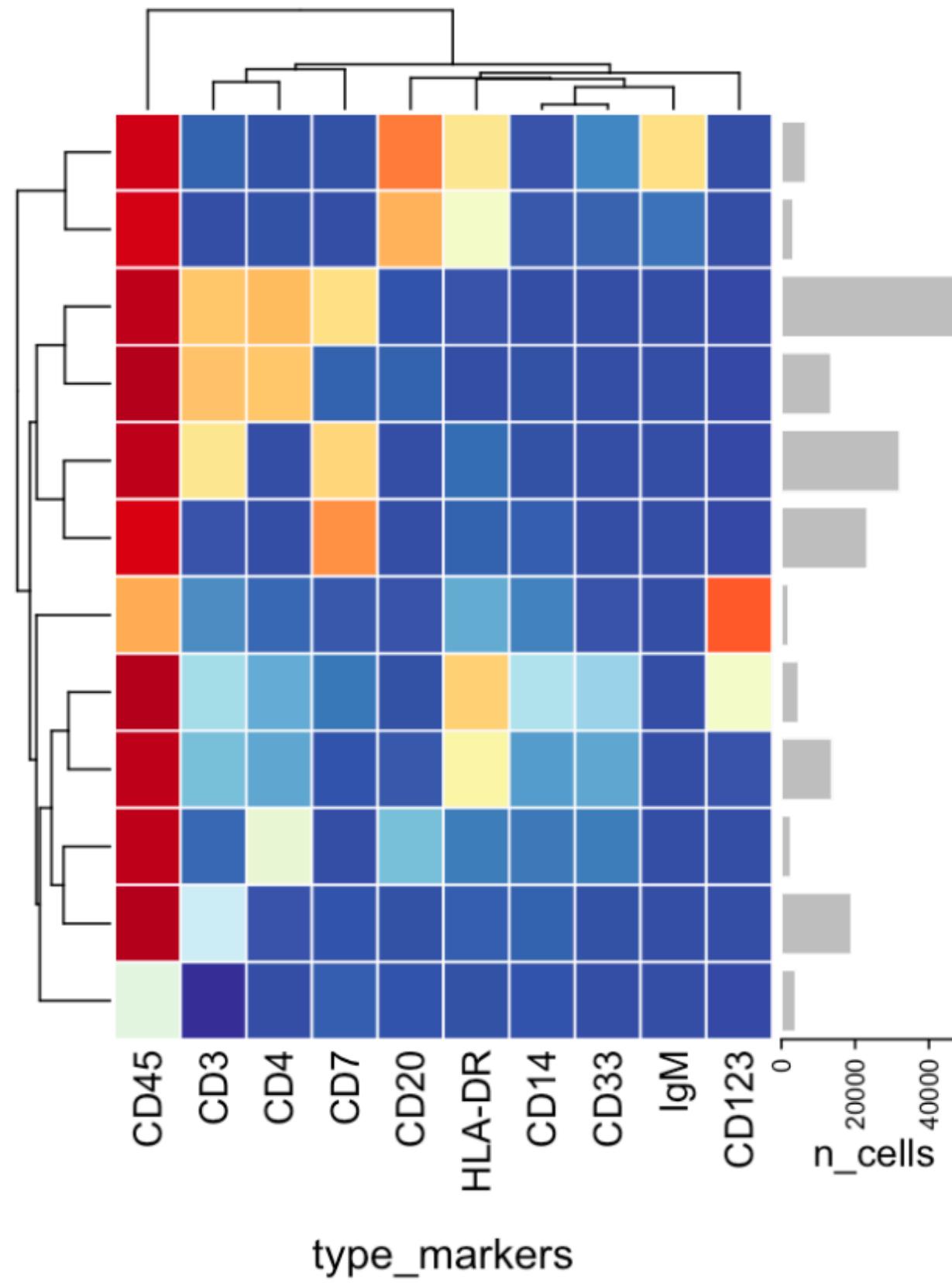
- **differential abundance** (DA) of cell populations
- **differential states**
 - e.g., differential expression of functional proteins (e.g., signaling) within cell populations



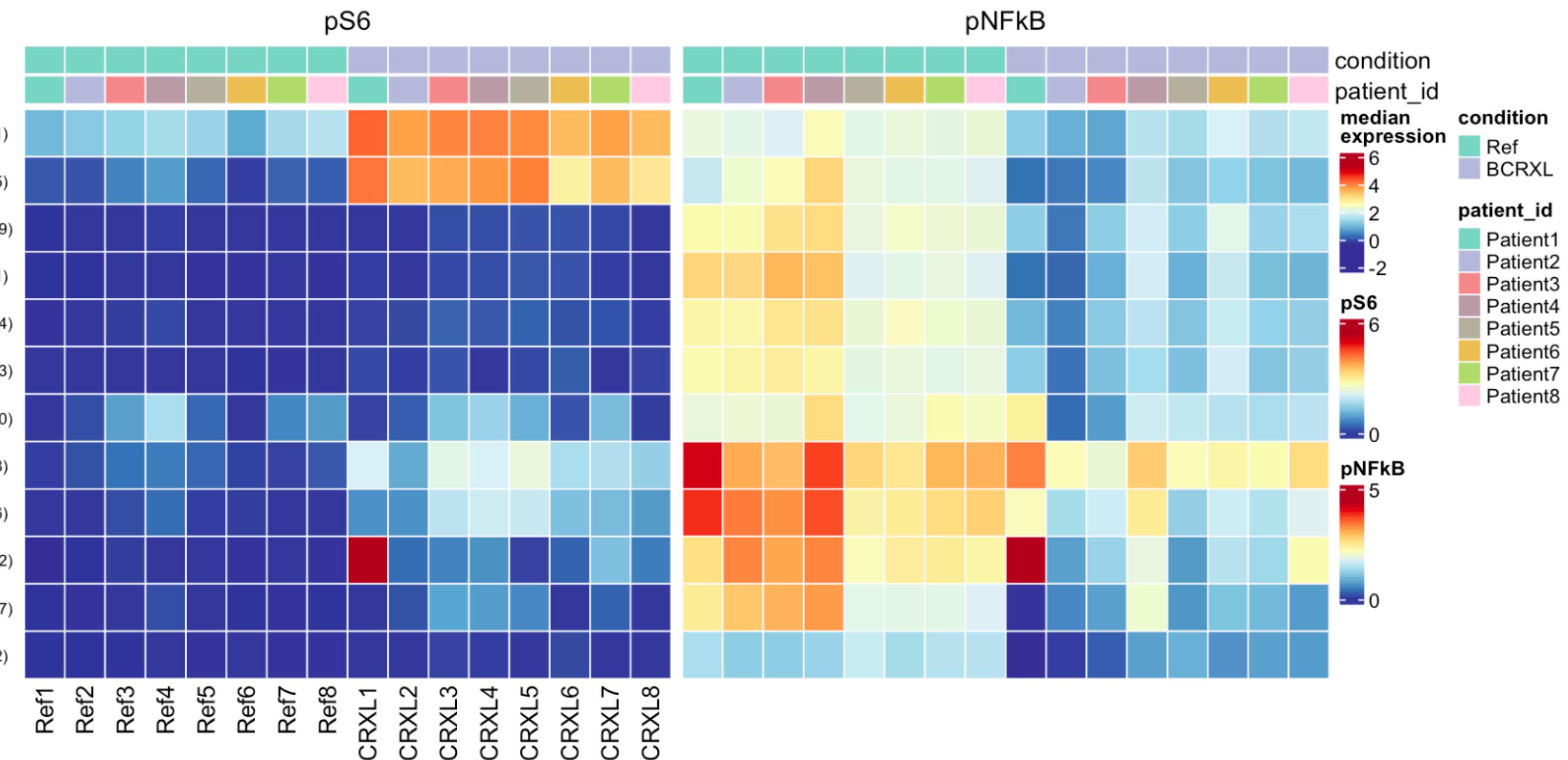
Canonical example: differential state analysis

Explicit split between “type” and “state” markers

10 type markers

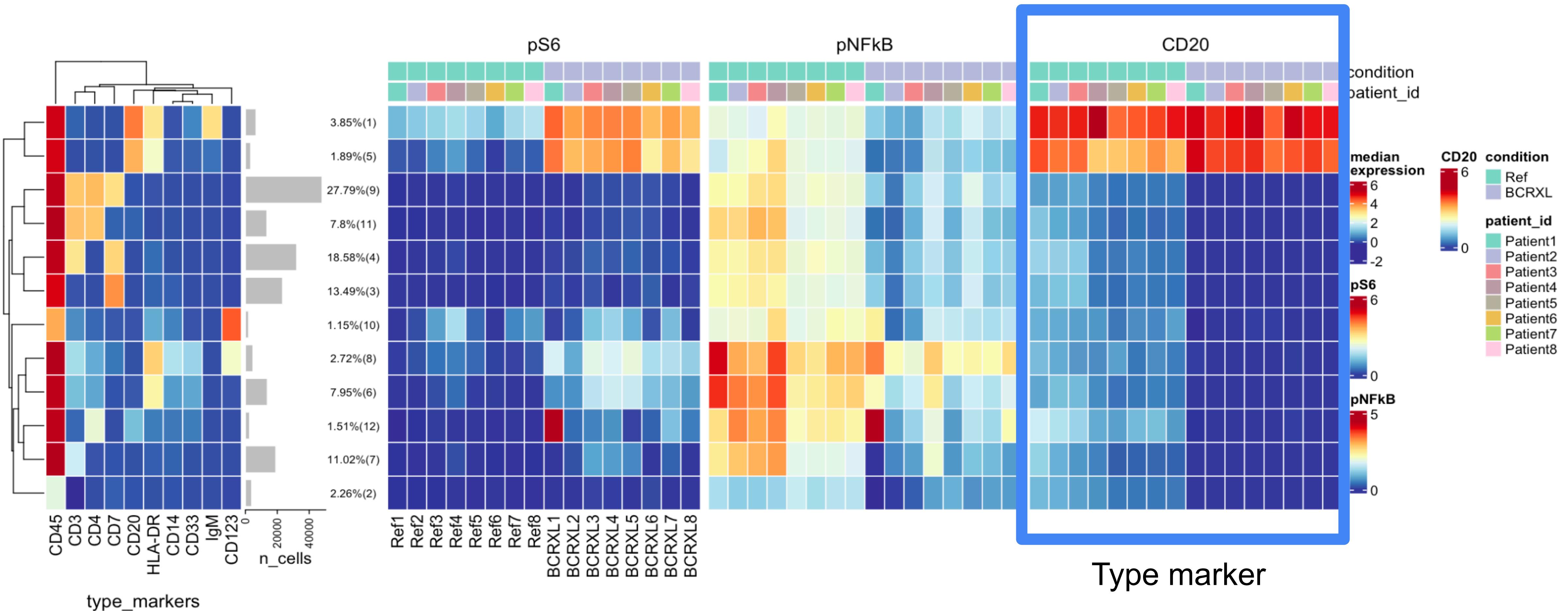


14 state markers (2 shown)



Data from Bodenmiller 2012 (8 patients; PBMCs +/- treatment with BCRXL)

With the expectation that type markers are constant



Data from Bodenmiller 2012 (8 patients; PBMCs +/- treatment with BCRXL)

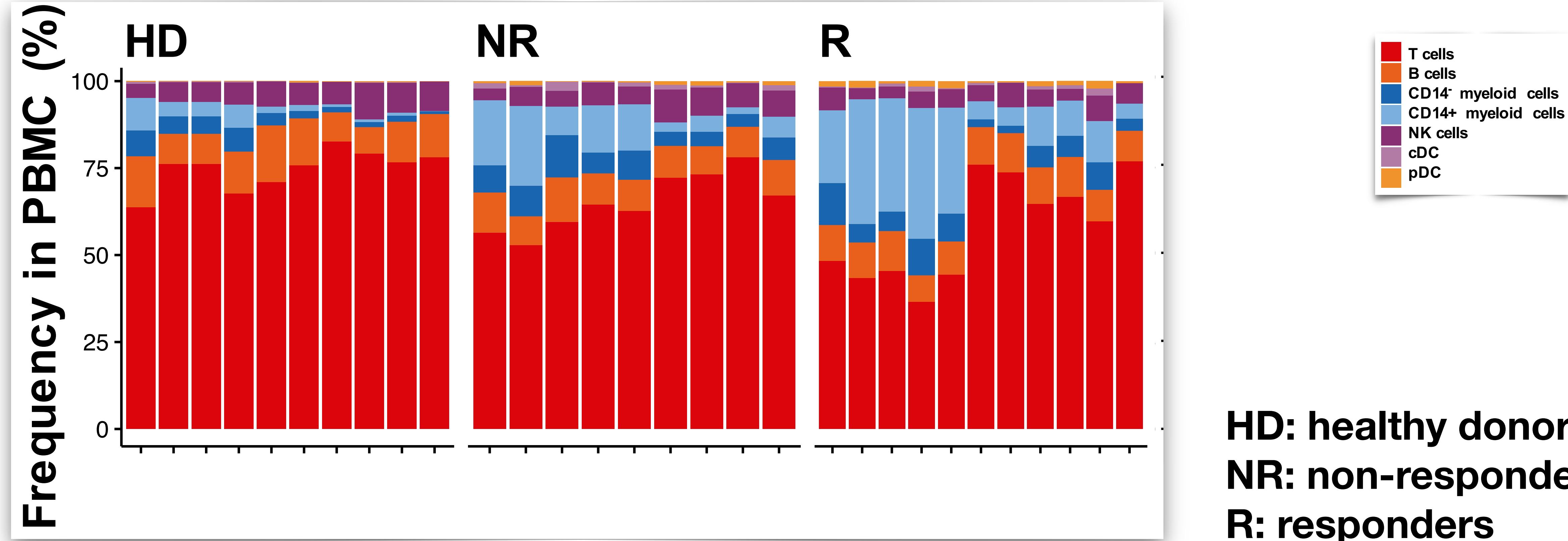
Good / Bad news: large batch effect, but nice experimental design (all conditions in every batch) so can be separated in statistical models.

High-dimensional single-cell analysis predicts response to anti-PD-1 immunotherapy

Carsten Krieg^{1,6} , Małgorzata Nowicka^{2,3}, Silvia Guglietta⁴, Sabrina Schindler⁵, Felix J Hartmann¹ , Lukas M Weber^{2,3} , Reinhard Dummer⁵, Mark D Robinson^{2,3} , Mitchell P Levesque^{5,7}  & Burkhard Becher^{1,7} 

Part 1:

Differential abundance of cell populations



After clustering (and manual merging), *generalized linear mixed model* is applied to cell count table to find differential abundance (n.b.: similar to RNA-seq differential expression).

Models for differential abundance similar to those for RNA-seq, but lower dimension



Manual merging of cell populations based on phenotypes

Generalized linear mixed models (differential abundance)

$$E(Y_{ij} | \beta_0, \beta_1, \gamma_i, \xi_{ij}) = \text{logit}^{-1}(\beta_0 + \beta_1 x_{ij} + \gamma_i + \xi_{ij}),$$

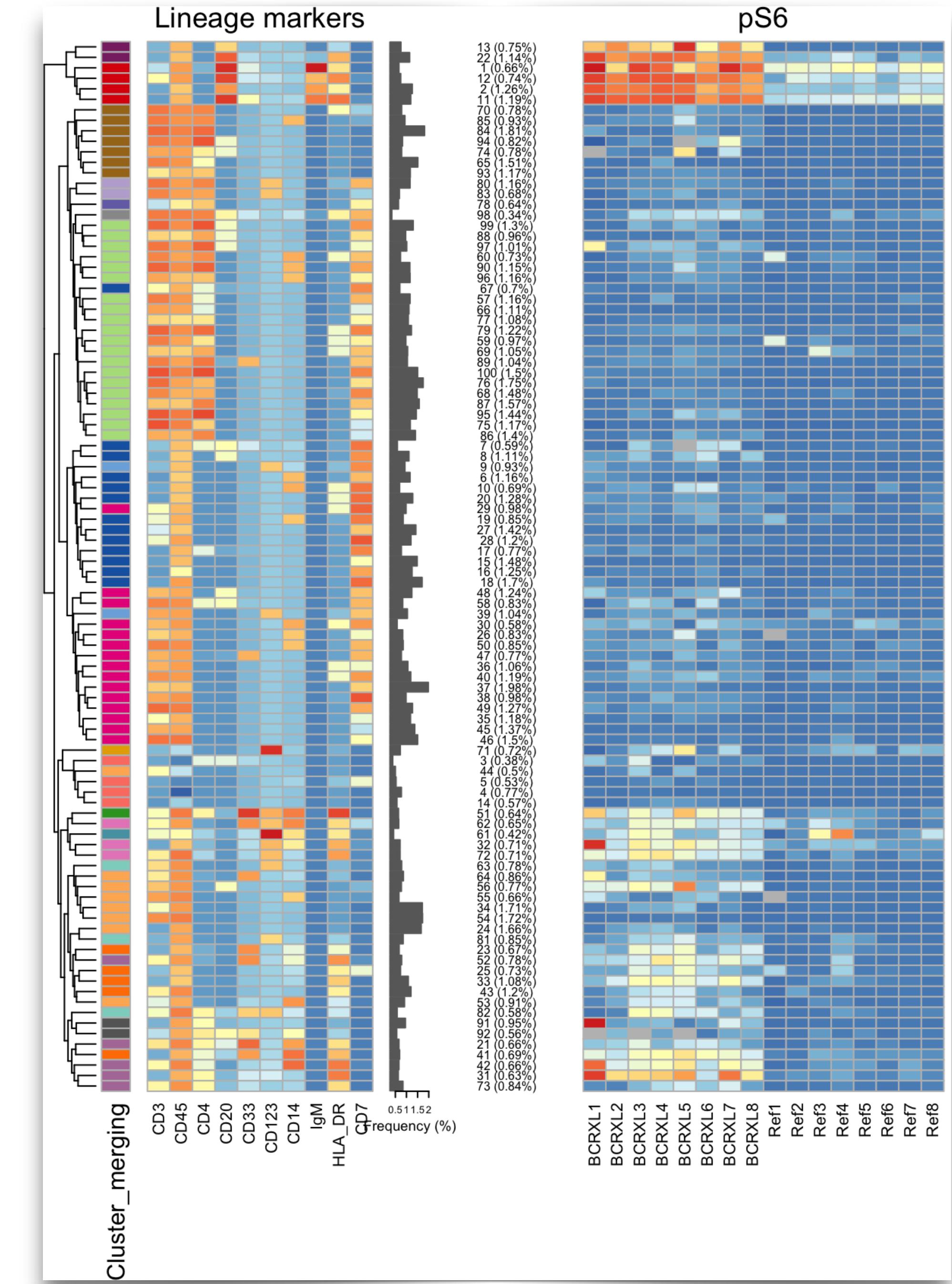
Linear mixed models (differential expression within populations)

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma_i + \epsilon_{ij},$$

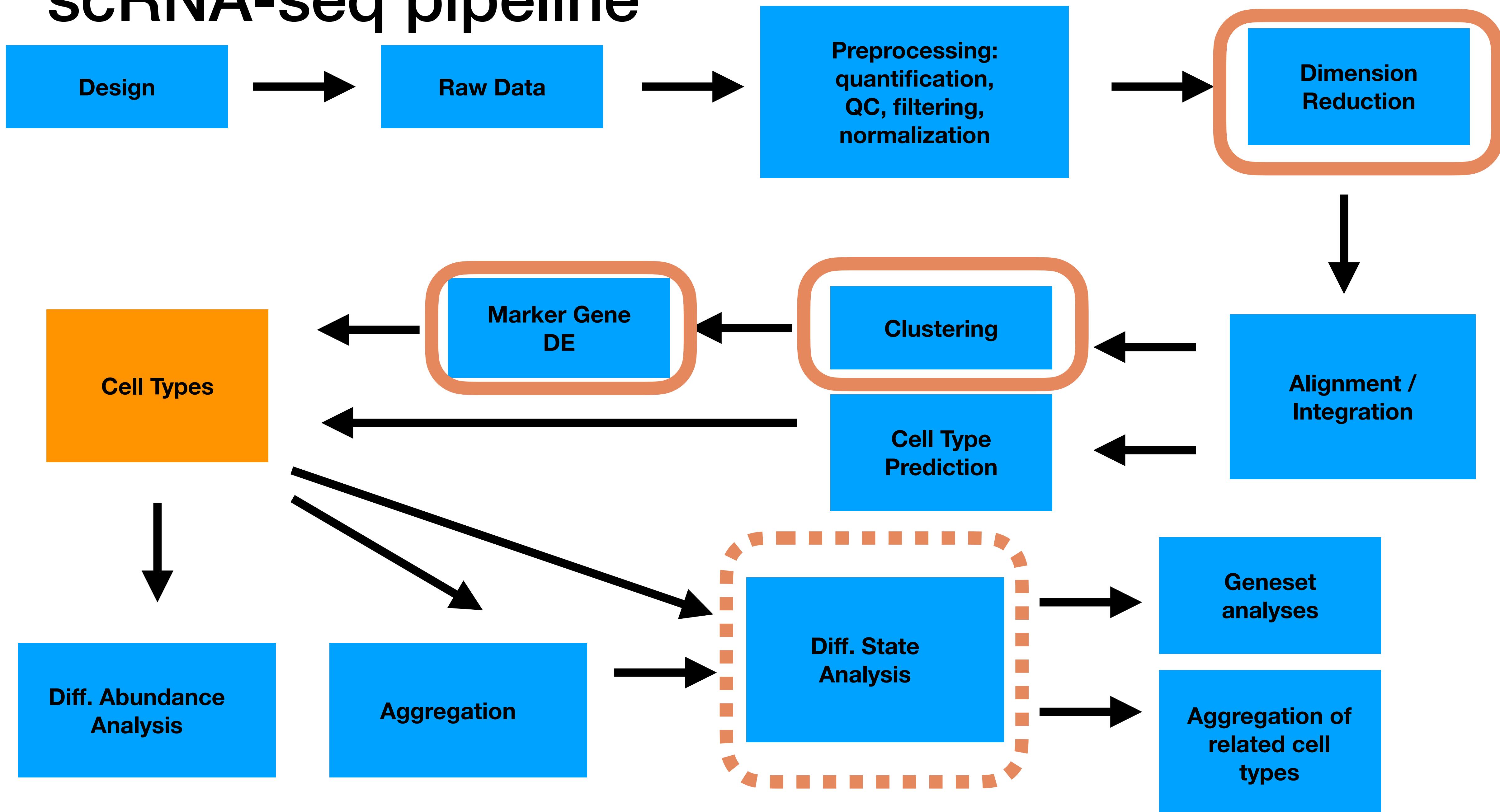
Part 2: subpopulation-specific differential analyses

Cluster to some number of groups based on lineage/type markers; look across samples in functional marker

→ median lineage marker signal by cluster



scRNA-seq pipeline



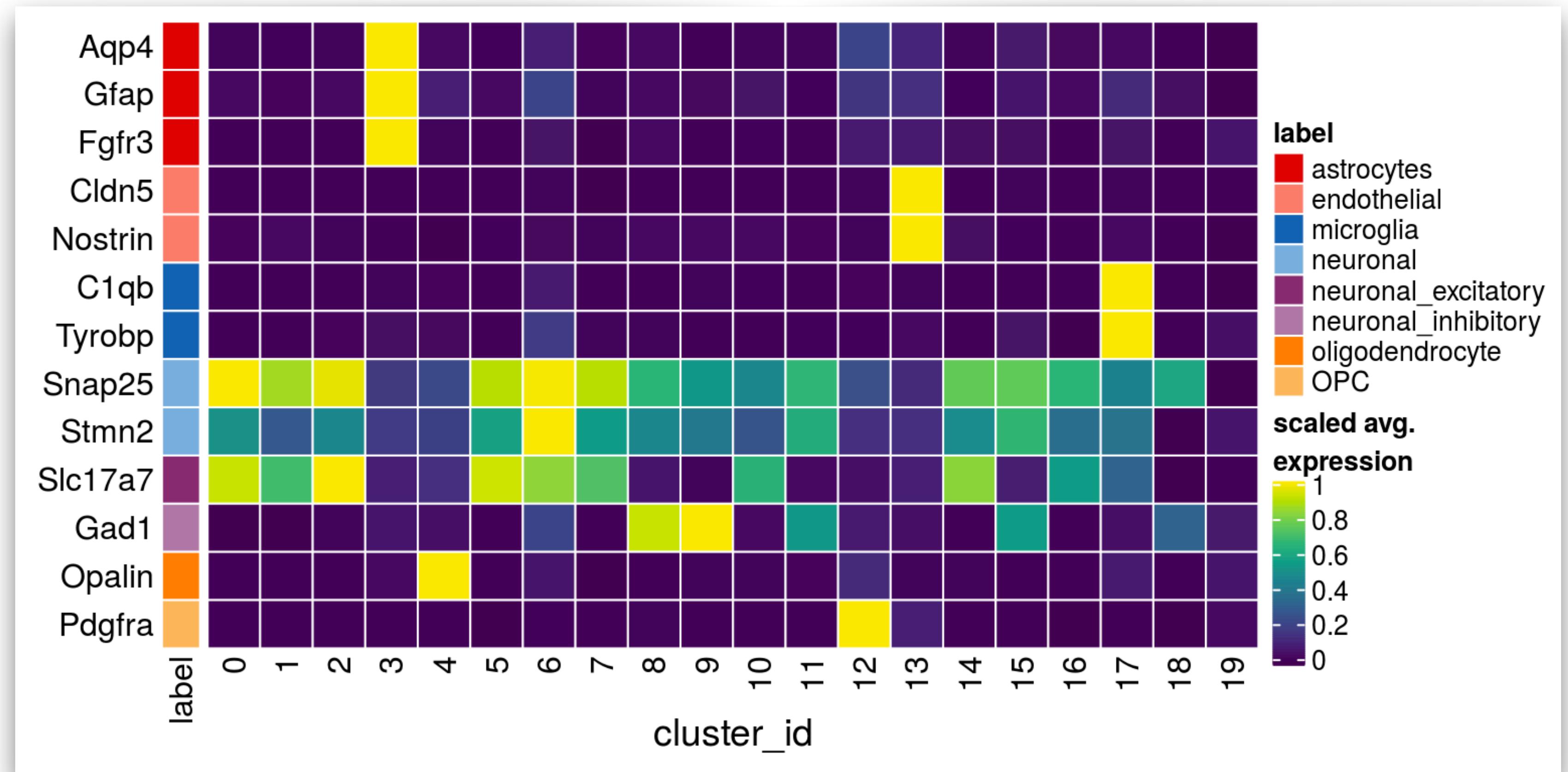
Application to LPS dataset: clustering + annotation of subpopulations

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

frontal cortex

single nuclei RNA-seq (10x)

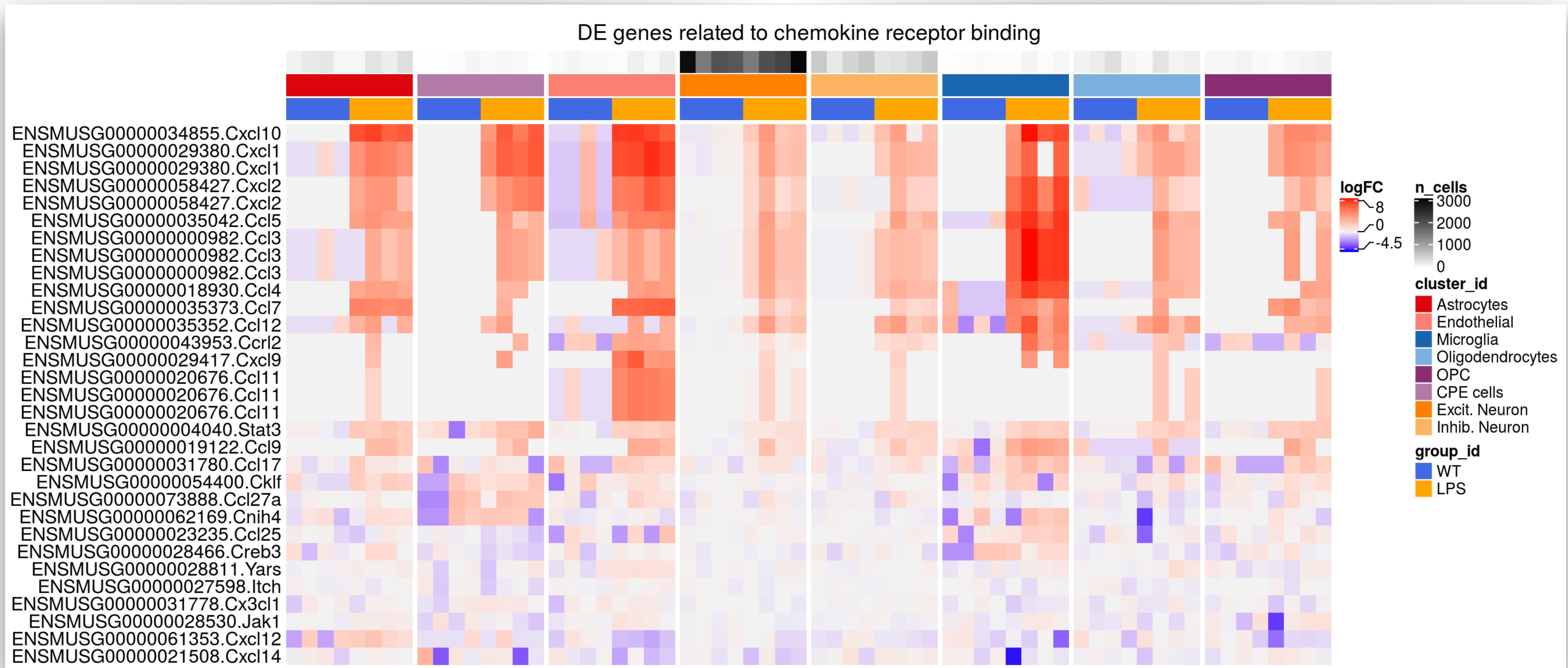
usual preprocessing:
filtering, doublet removal,
Seurat integration,
clustering



LPS dataset: interplay of cell type and cell state



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





**University of
Zurich**^{UZH}

Statistical Bioinformatics // Institute of Molecular Life Sciences

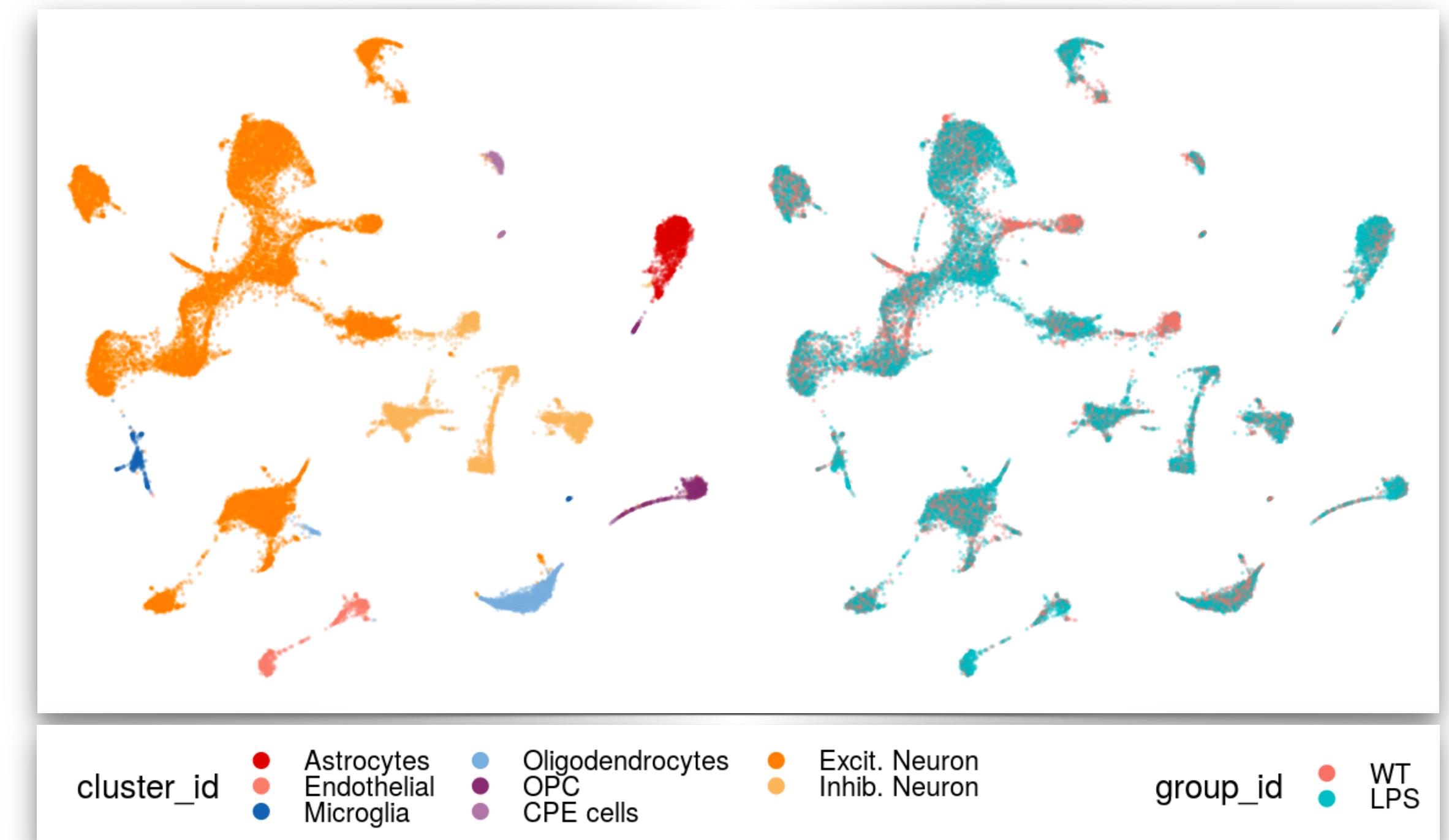
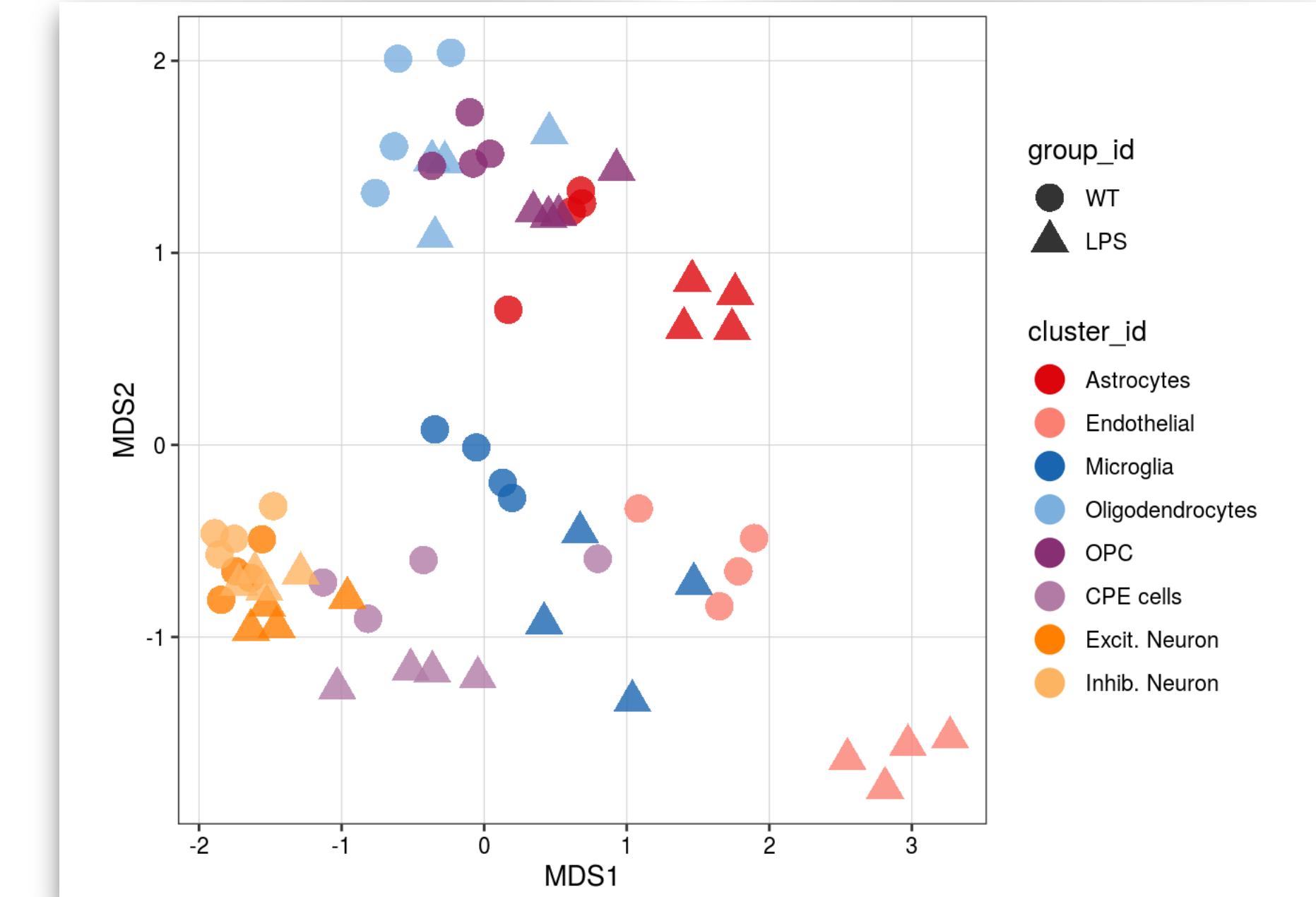
Dimension reduction

Dimension reduction: general introduction

- Single cell data comes as a matrix of N cells $\times G$ 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



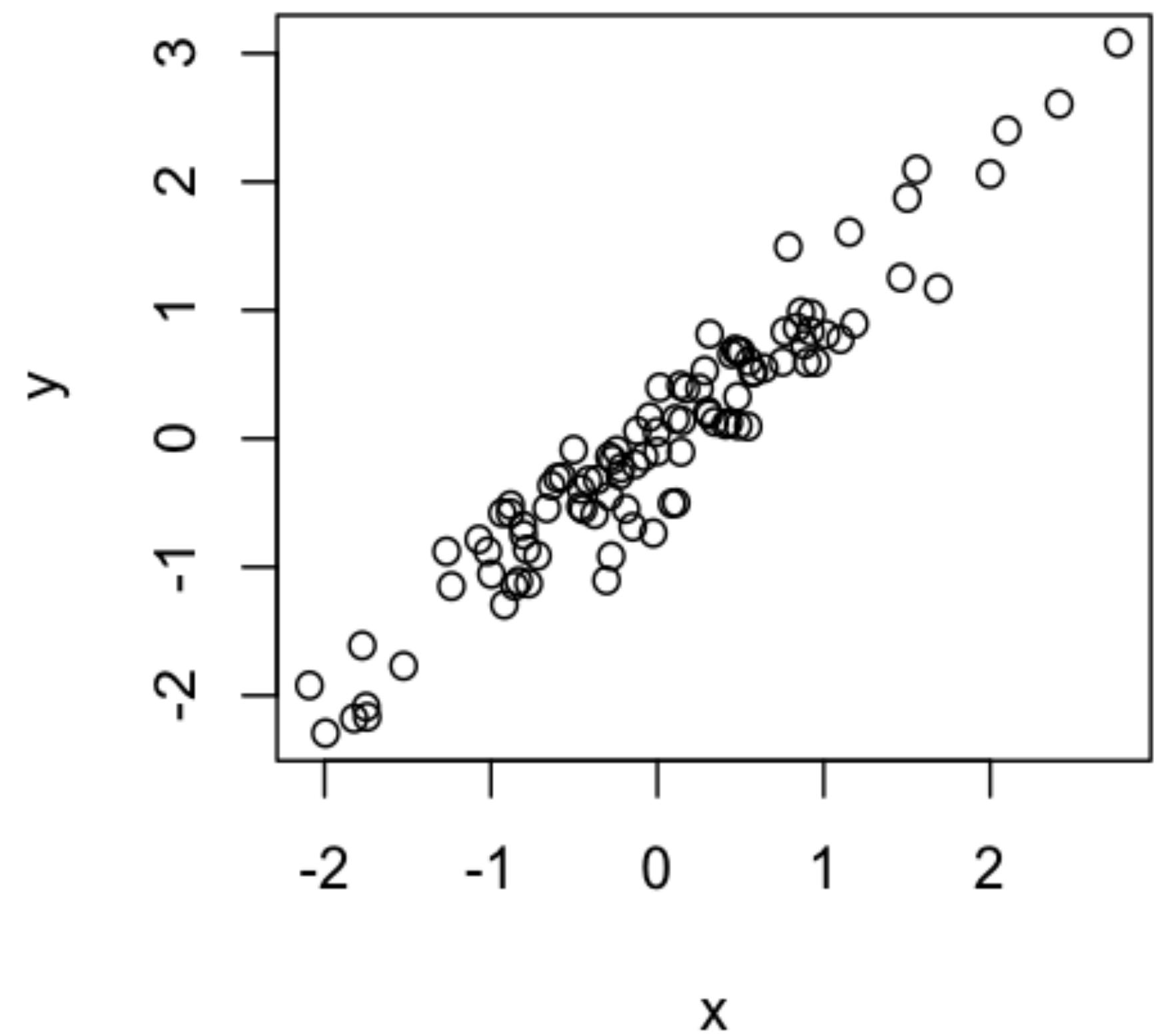
Introduction to dimension reduction: PCA (principal components analysis)

- 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)



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 $x_1, \dots, x_M \in \mathbb{R}^N$ such that

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

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

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}

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

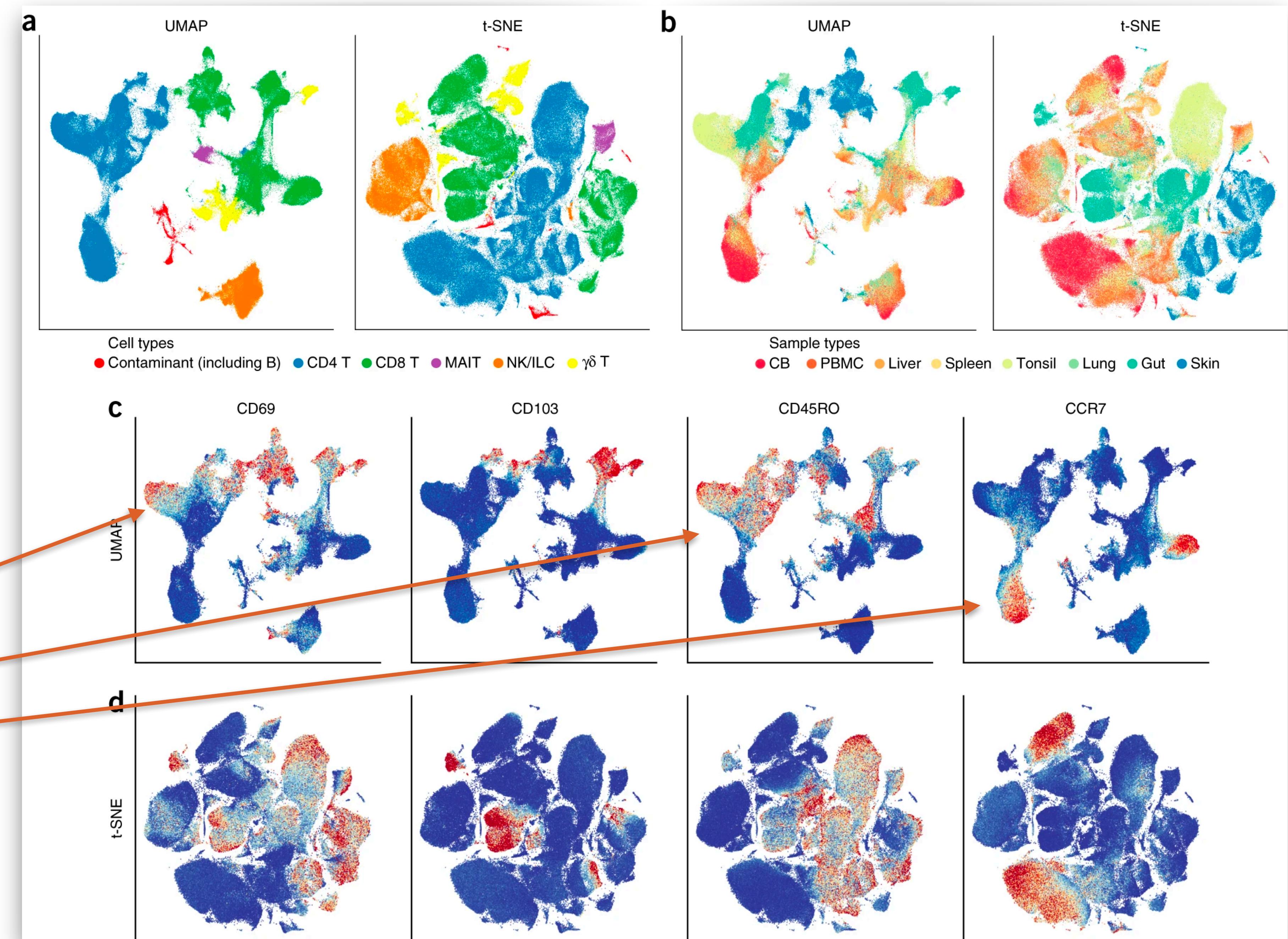
t-SNE aims to learn a d -dimensional map $\mathbf{y}_1, \dots, \mathbf{y}_N$ (with $\mathbf{y}_i \in \mathbb{R}^d$) that reflects the similarities p_{ij} as well as possible. To this end, it measures similarities q_{ij} between two points in the map \mathbf{y}_i and \mathbf{y}_j , using a very similar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

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

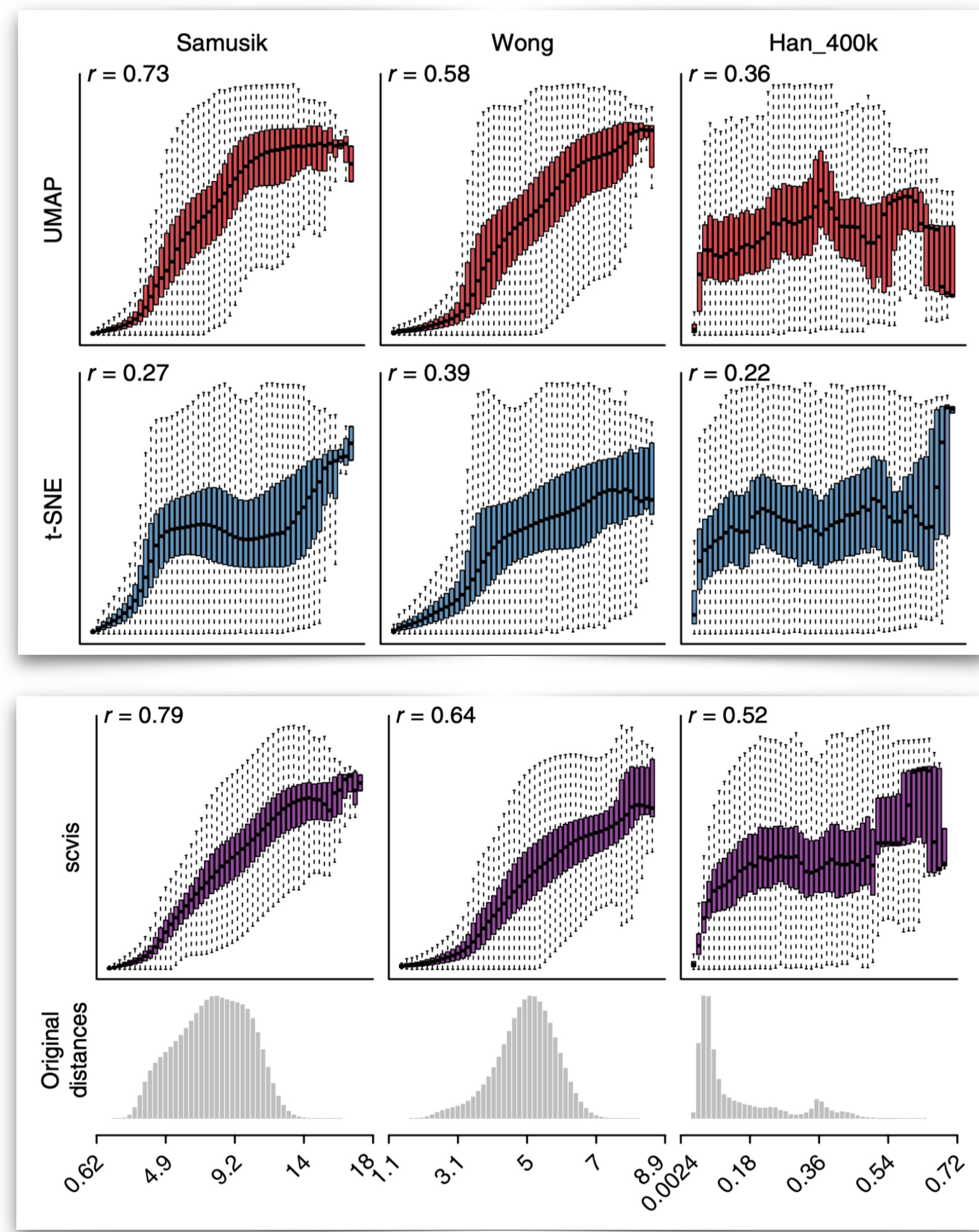
Introduction to dimension reduction: UMAP (Uniform Manifold Approximation and Projection)

- 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

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



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



Careful:
dimension
reduction
can induce
distortions

PERSPECTIVE

The specious art of single-cell genomics

Tara Chari¹, 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.*

Careful:
but it's
not all
that bad

FORMAL COMMENT

The art of seeing the elephant in the room: 2D embeddings of single-cell data do make sense

Jan Lause^{1,2}, Philipp Berens^{1,2}, Dmitry Kobak^{1,2,3*}

1 Hertie Institute for AI in Brain Health, University of Tübingen, Tübingen, Germany, **2** Tübingen AI Center, University of Tübingen, Tübingen, Germany, **3** IWR, Heidelberg University, Heidelberg, Germany

* dmitry.kobak@uni-tuebingen.de

A recent paper (Chari and Pachter, 2023) claimed that t-SNE and UMAP embeddings of single-cell datasets fail to capture true biological structure .. are as arbitrary and as misleading as forcing the data into an elephant shape. Here we show that this conclusion was based on inadequate and limited metrics of embedding quality. More appropriate metrics quantifying neighborhood and class preservation reveal the elephant in the room: while t-SNE and UMAP embeddings of single-cell data do not preserve high-dimensional distances, they can nevertheless provide biologically relevant information.



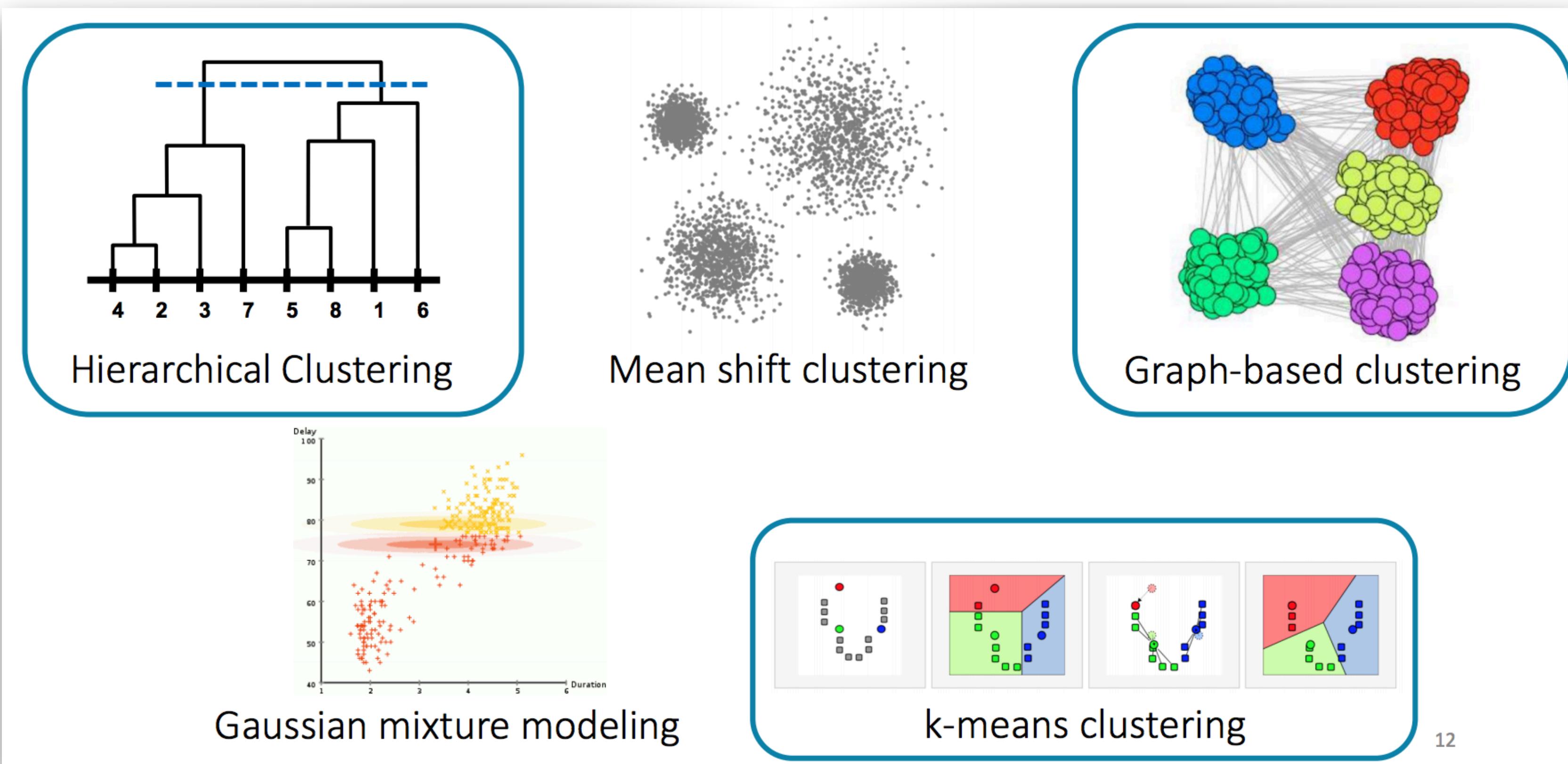
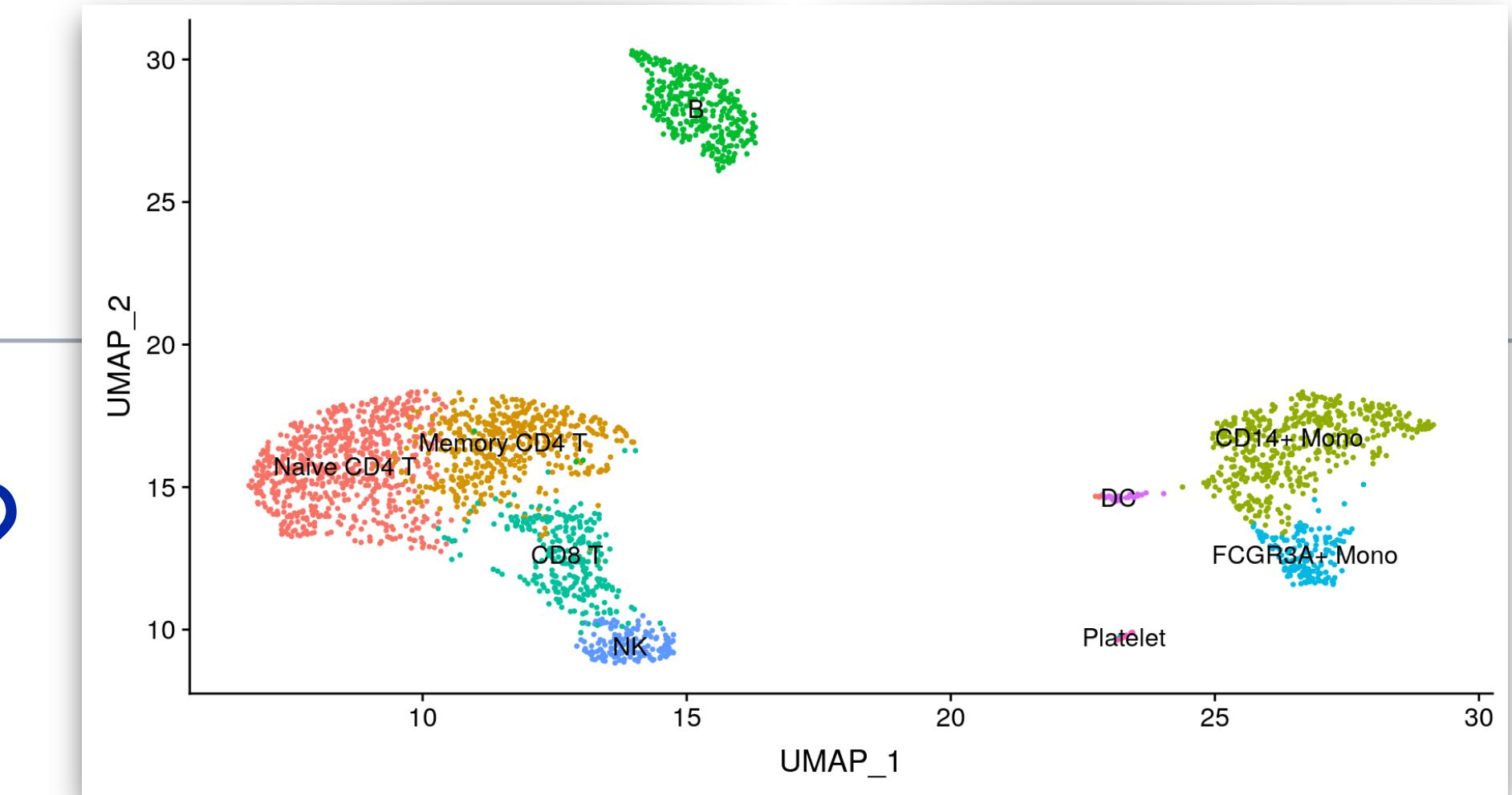
**University of
Zurich**^{UZH}

Statistical Bioinformatics // Institute of Molecular Life Sciences

Clustering



Clustering: an ill-posed problem?





Cluster Dendrogram

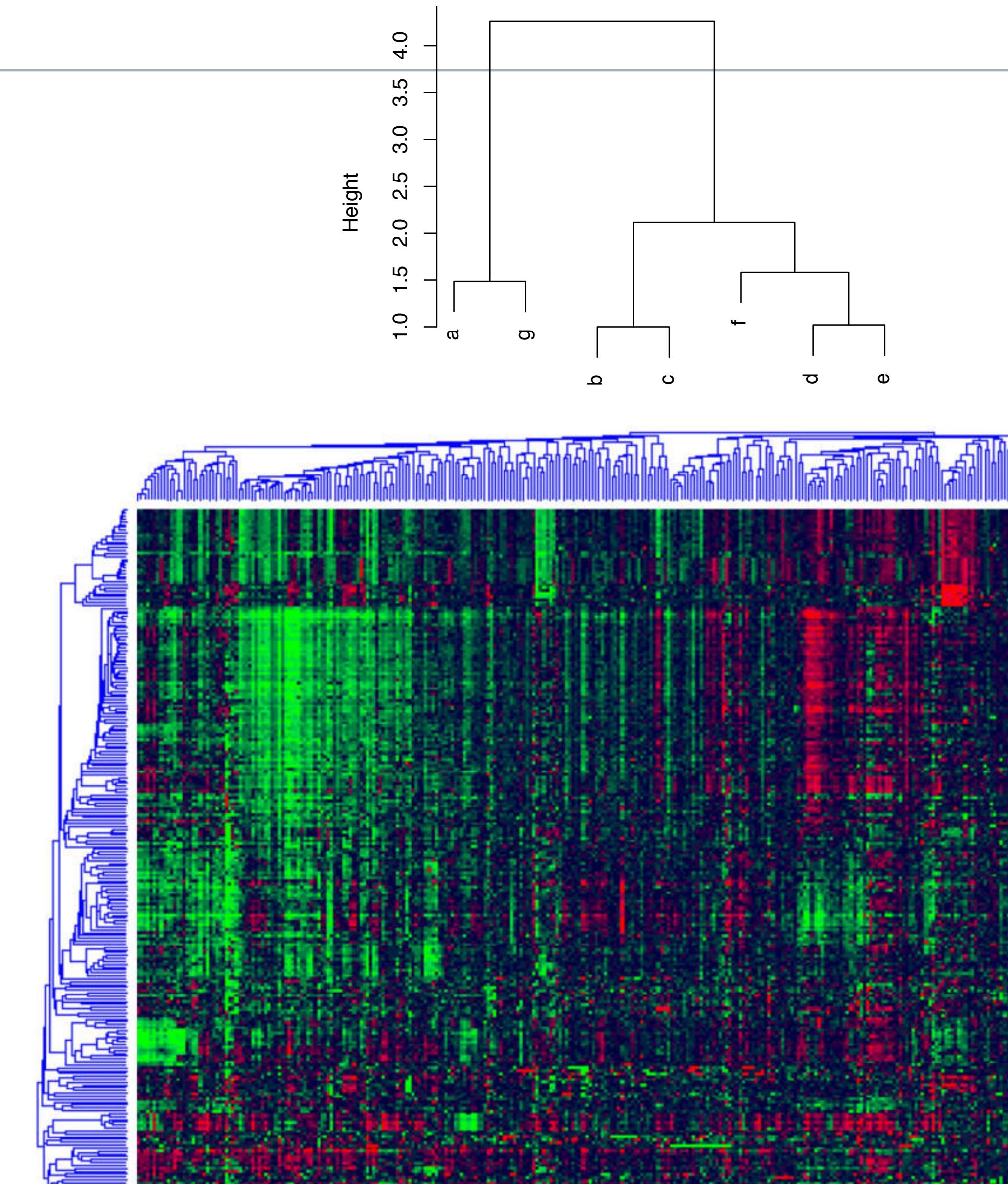
Hierarchical (Agglomerative) Clustering

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

Metric: to define how similar any two vectors are.

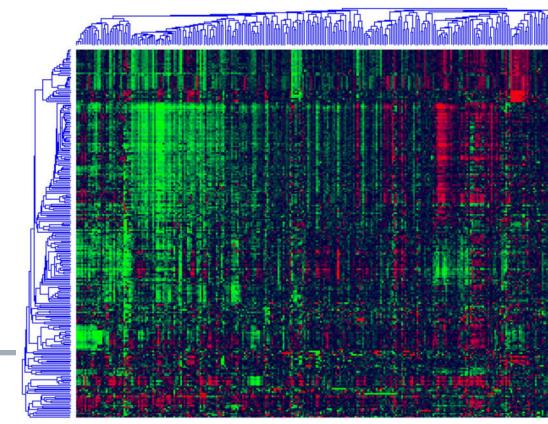
Linkage: determines how clusters are merged into a tree

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





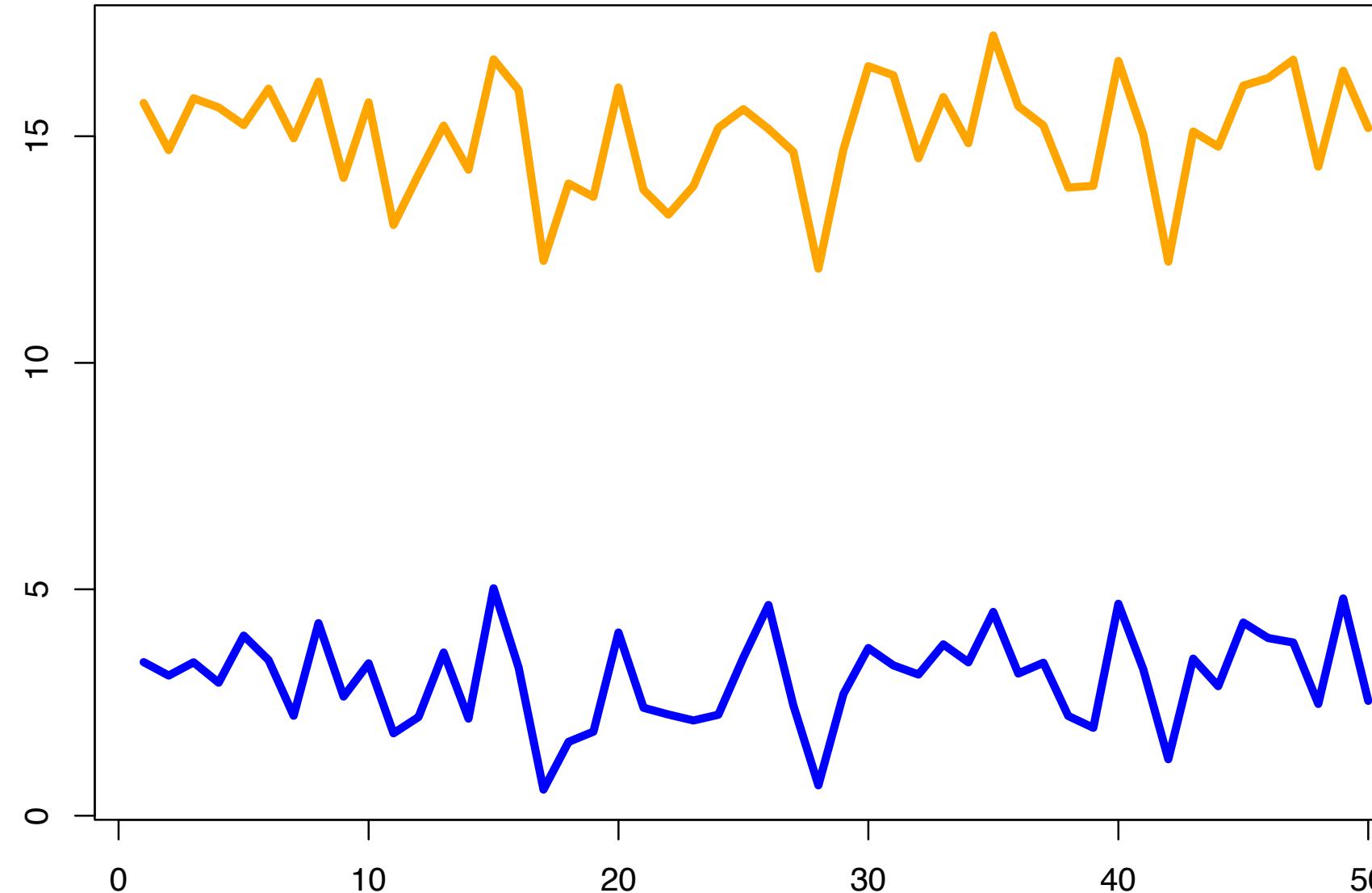
$$d(\mathbf{p}, \mathbf{q}) = d(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^n (q_i - p_i)^2}.$$



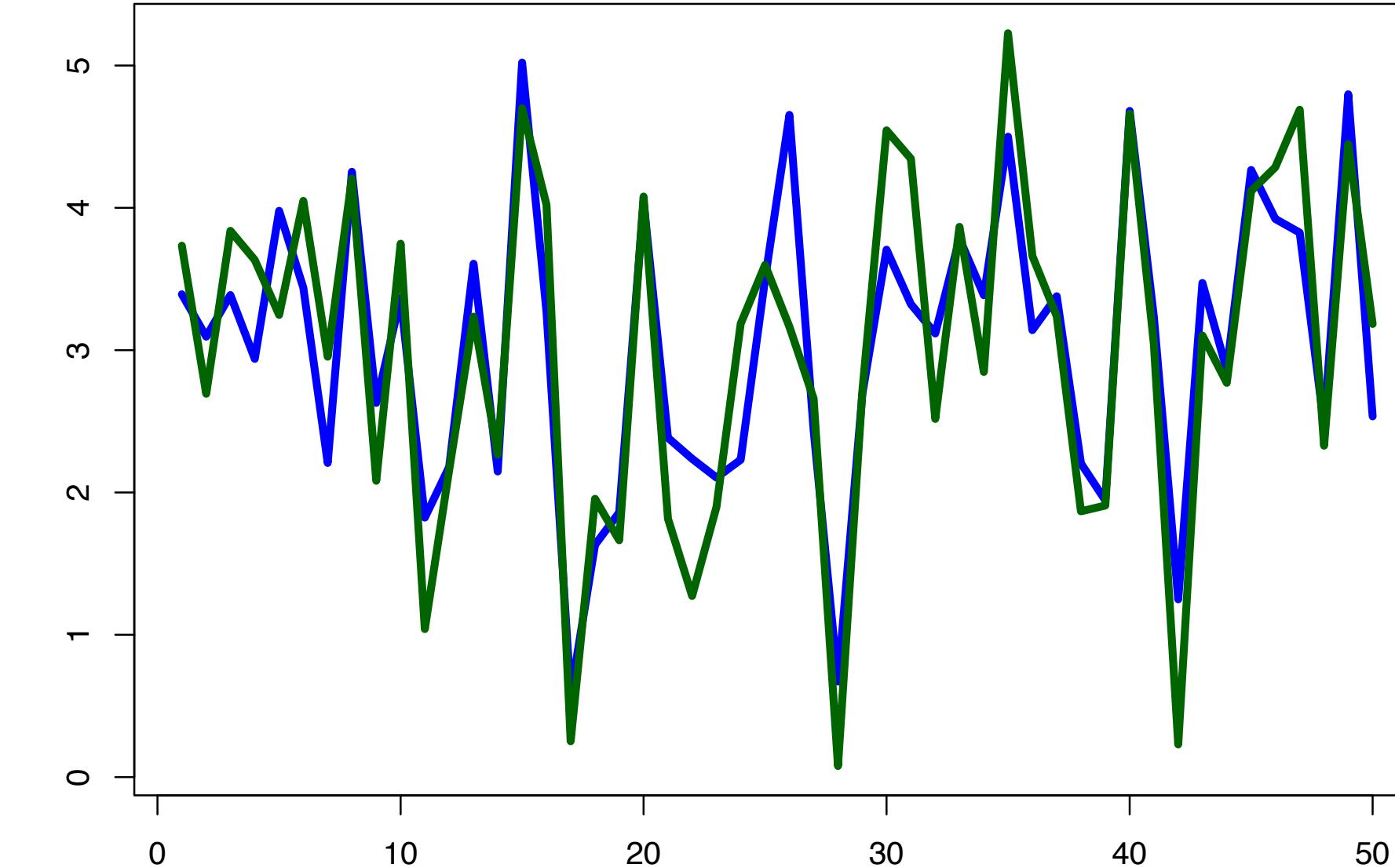
Are these “vectors” similar ?

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

It depends how you define similar.



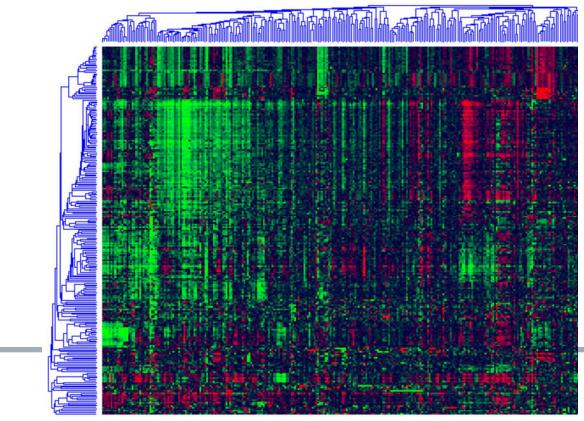
Euclidean distance: 84.84



3.92



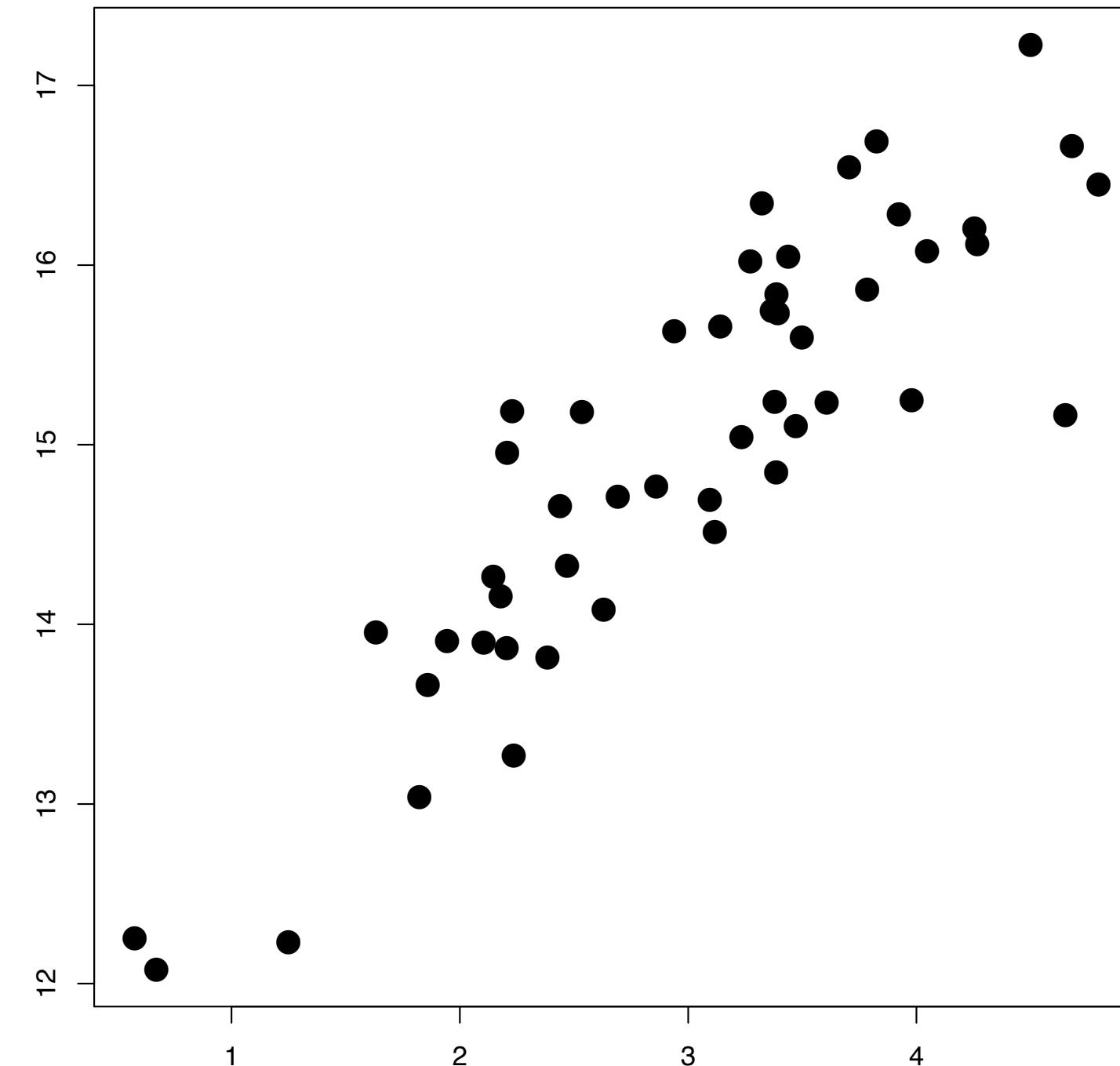
$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}},$$



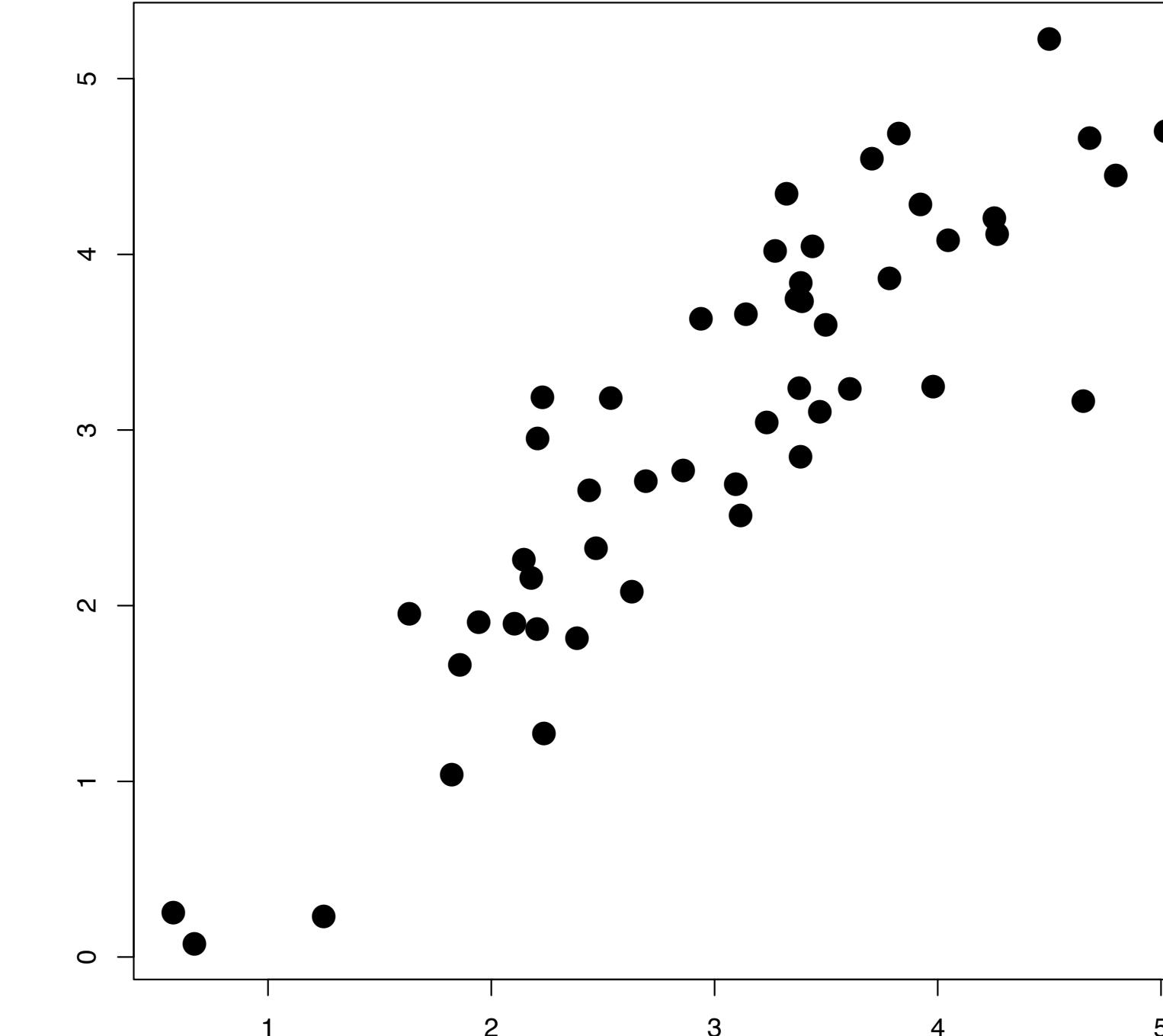
Are these “vectors” similar ?

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

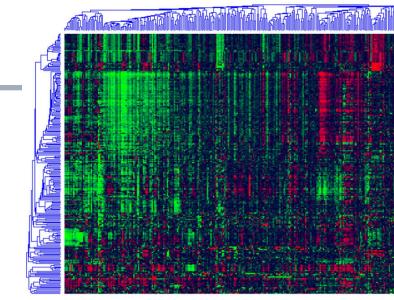
It depends how you define similar.



Correlation: 0.89



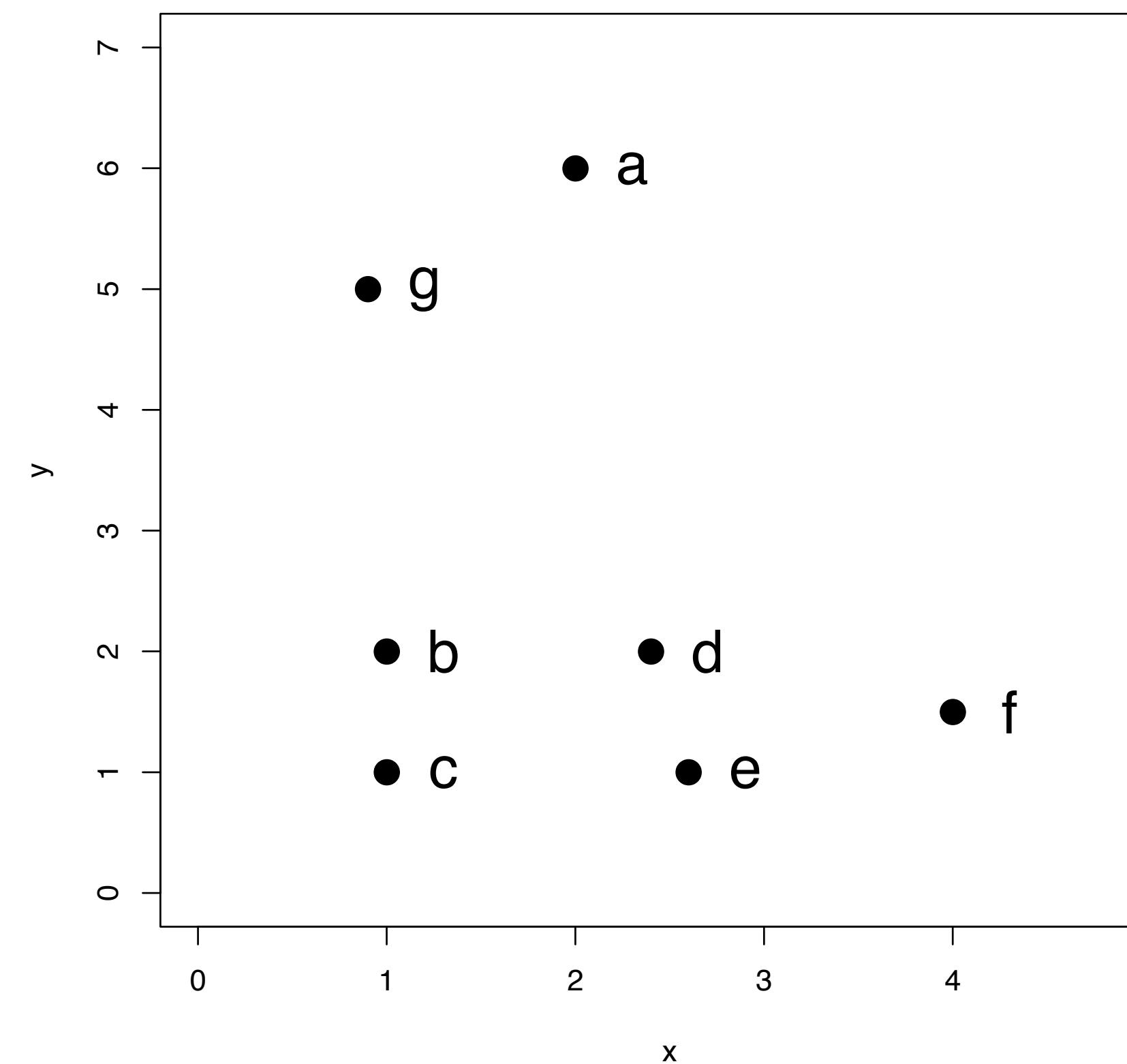
0.89



Hierarchical (Agglomerative) Clustering

Start with distances.

Linkage: determines how clusters are merged into a tree.

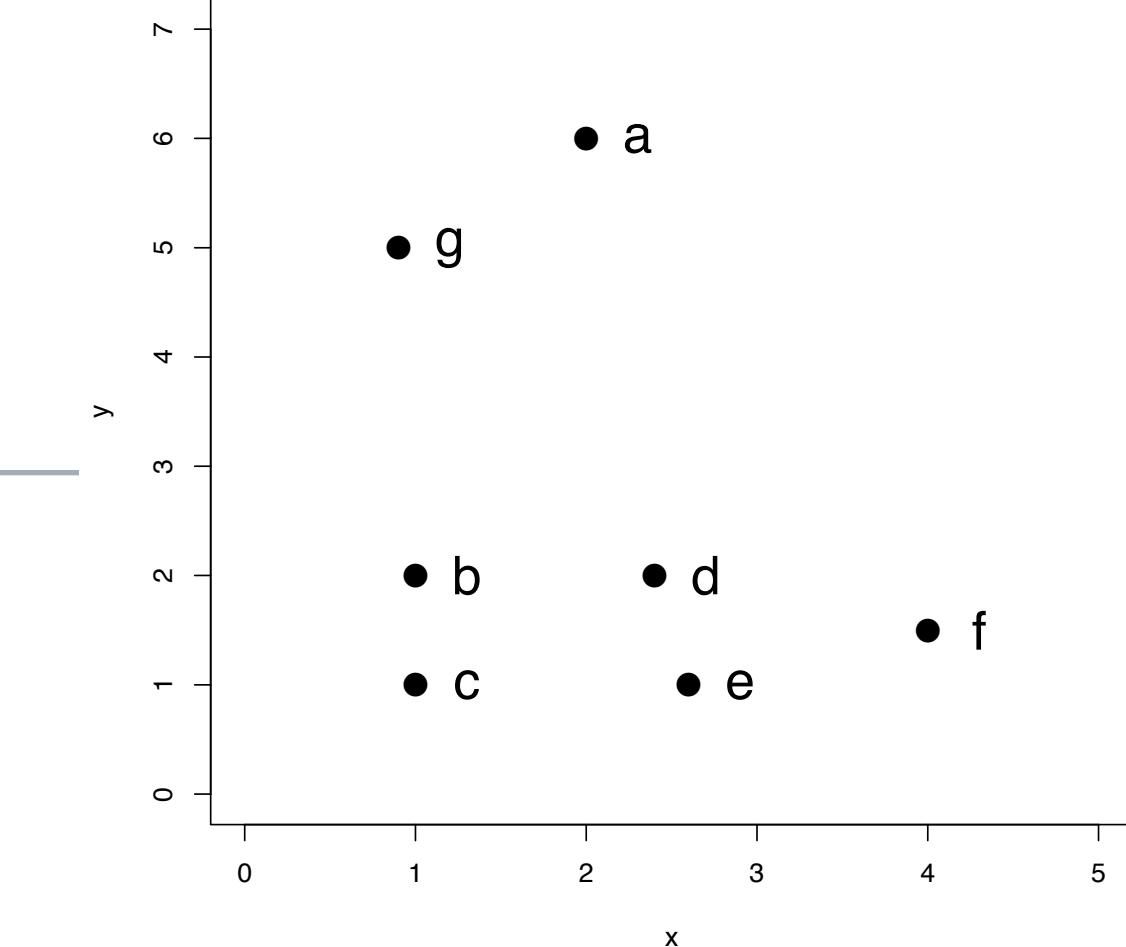
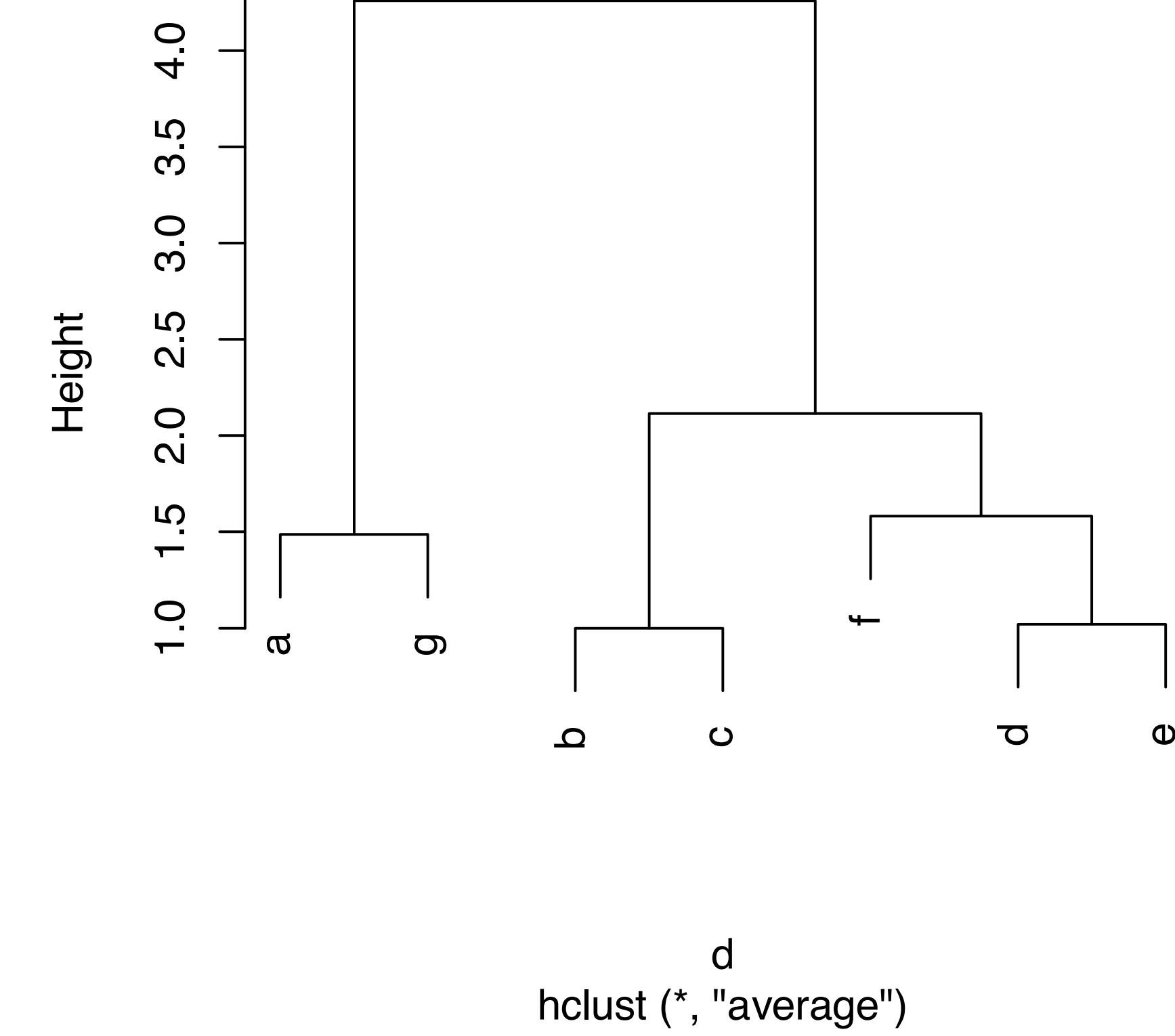
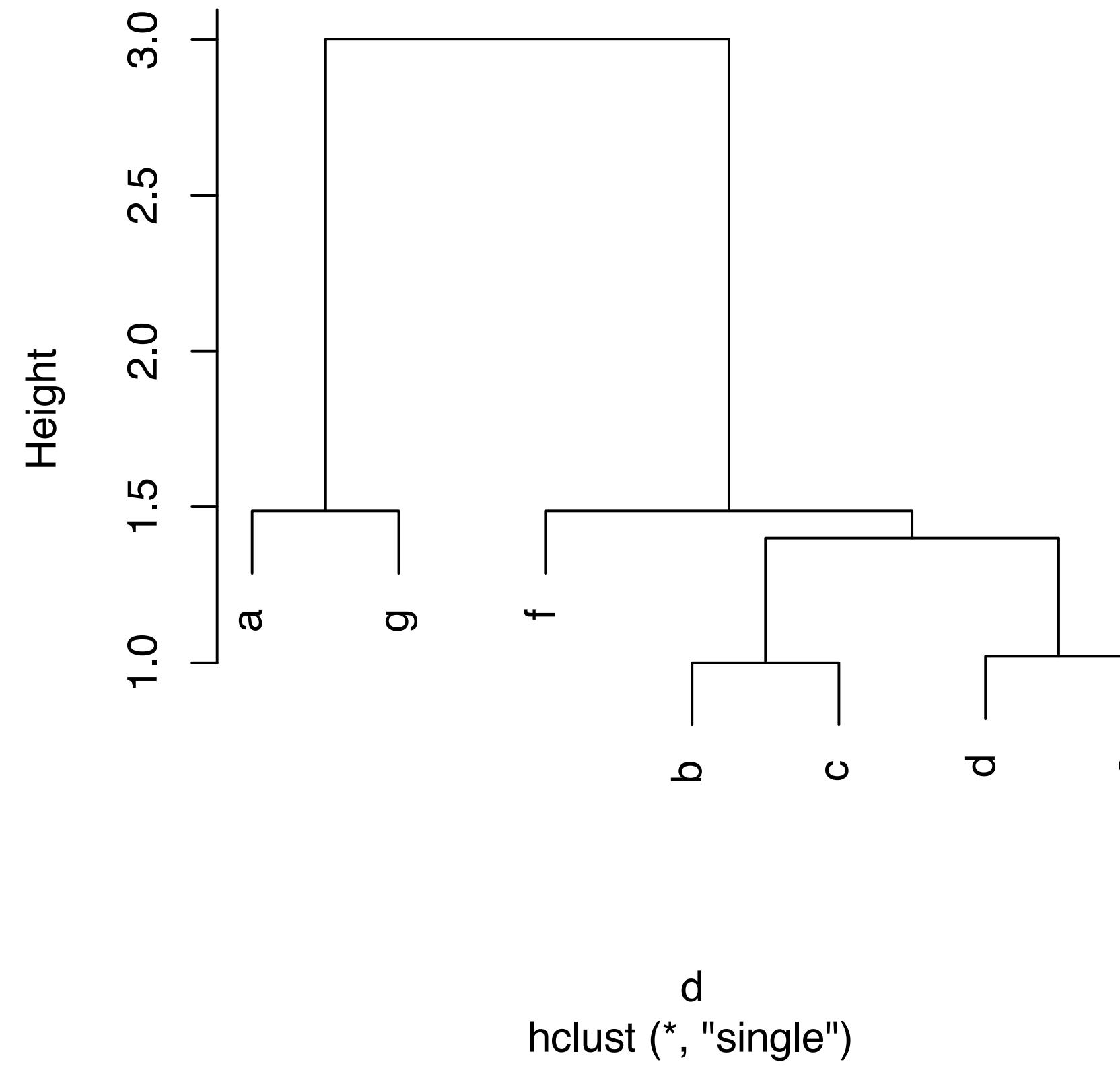


From eyeballing, here is a likely set of merges:

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



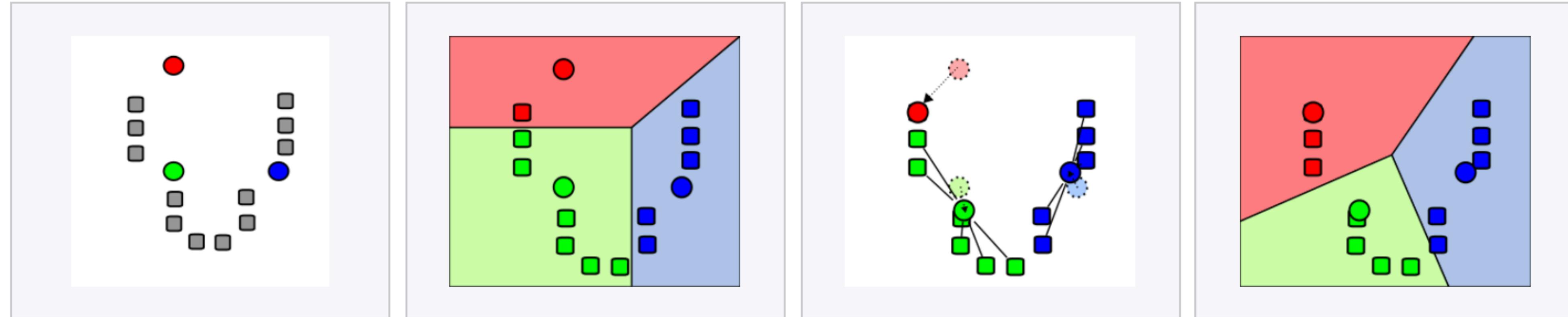
Different linkages





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.

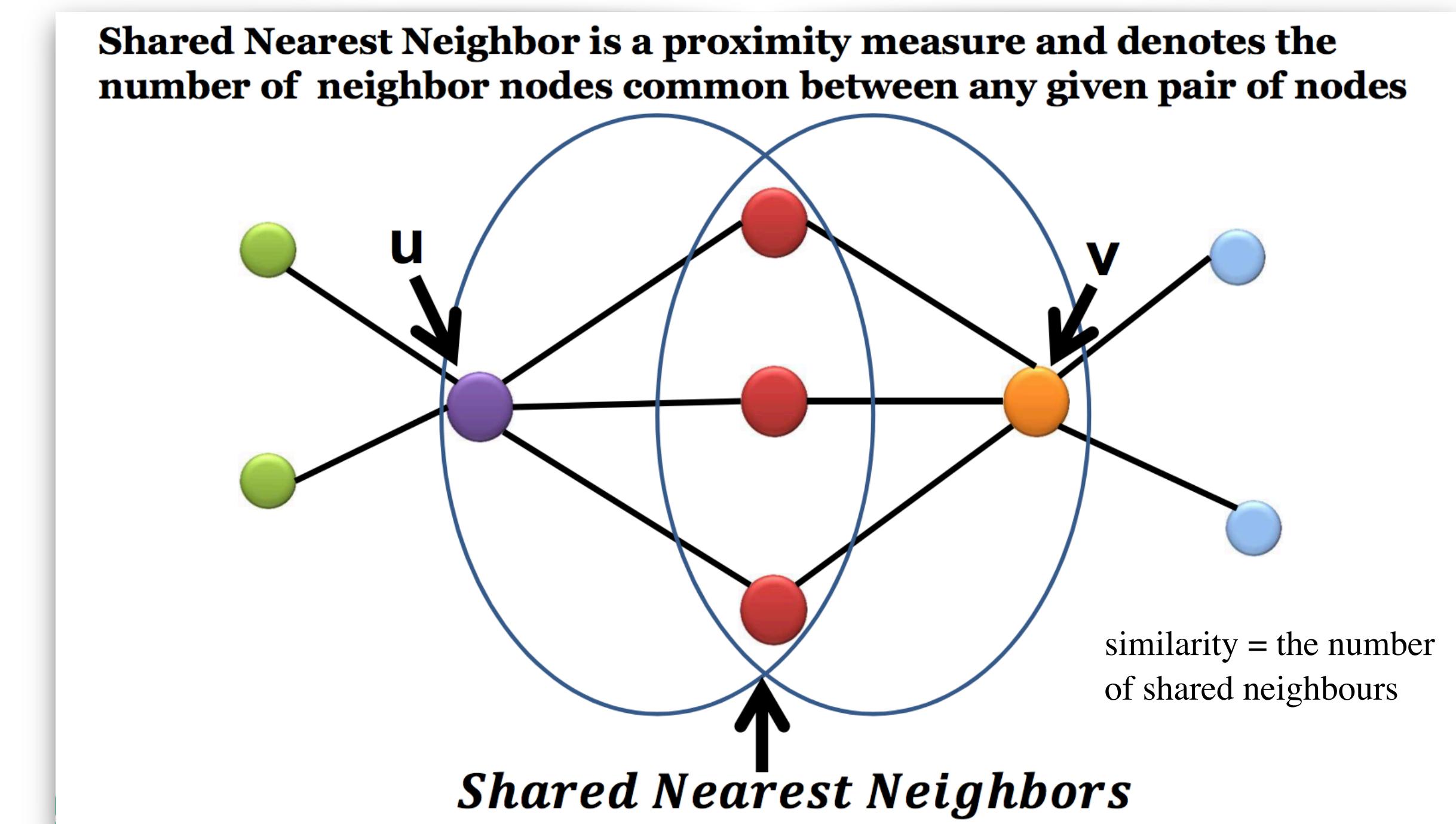
Disadvantages: hard to determine k , assumes spherical cluster shapes (depends on distance), different initial selection —> different clustering (typically run multiple times)



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.





Graph-based clustering

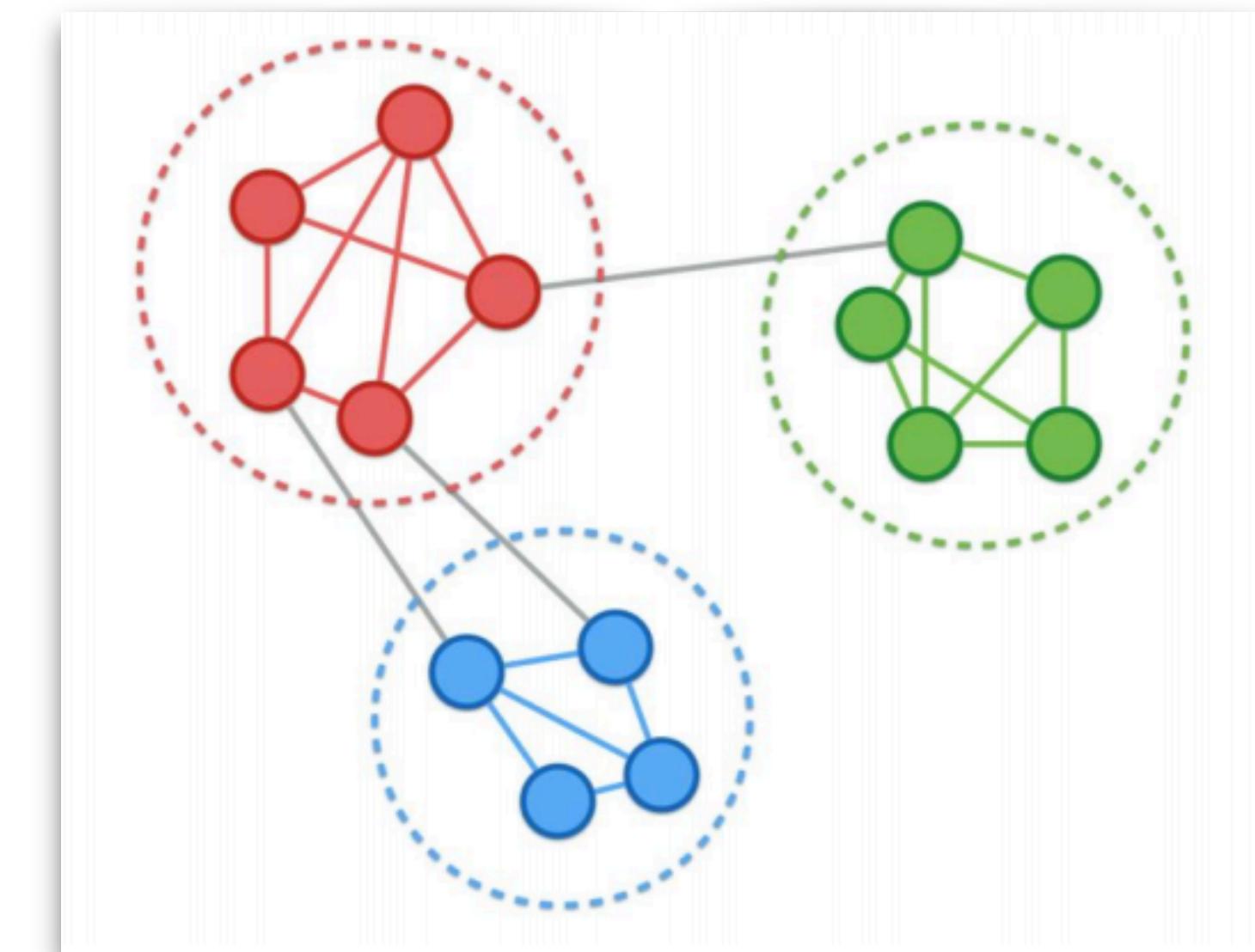
- From graph (originally depends on k), “community detection” to break into clusters
- Want to optimize “modularity” —> i.e. find more edges *inside* the groups than edges linking nodes with rest of the graph.
- Louvain (2008) is a well known heuristic method to optimize modularity

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

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
 δ = kronecker delta (1 if $c_i = c_j$)

From Louvain to Leiden: guaranteeing well-connected communities

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



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



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}¹Institute of Molecular Life Sciences, University of Zurich, Zurich, 8057, Switzerland²SIB Swiss Institute of Bioinformatics, Zurich, 8057, Switzerland

scRNA-seq: No “best” clustering algorithm

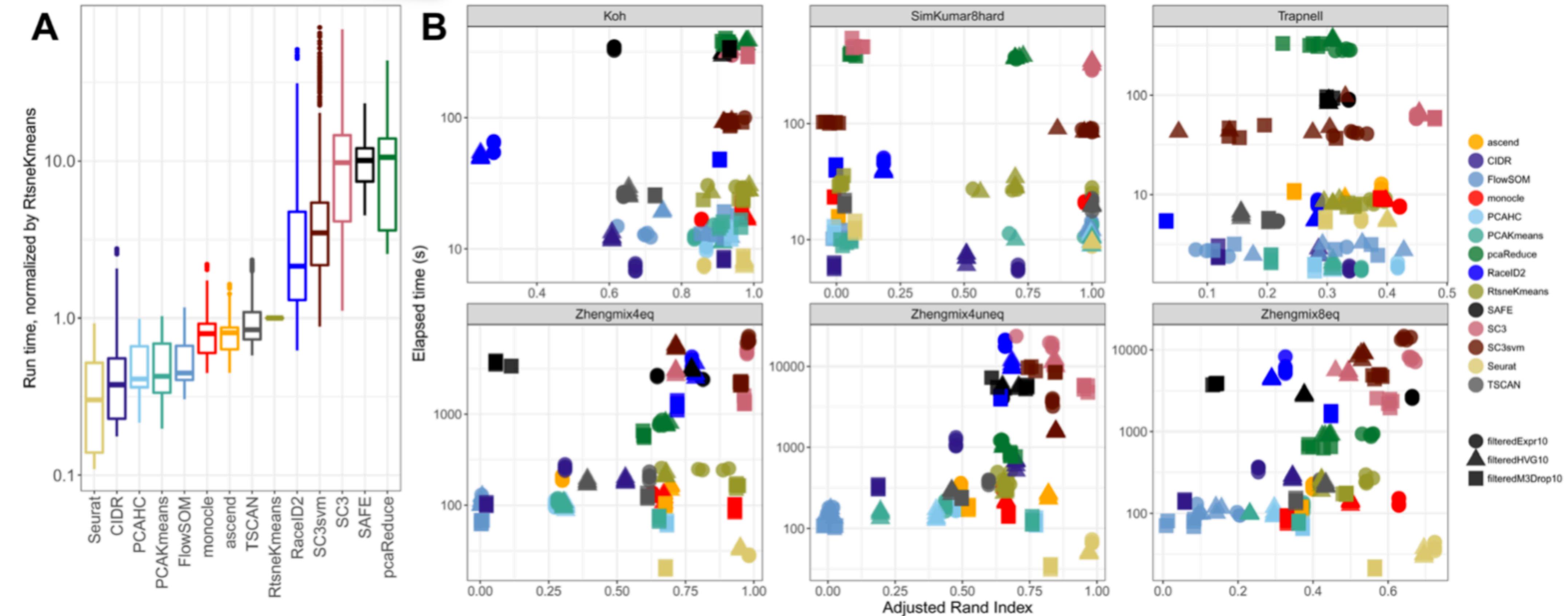
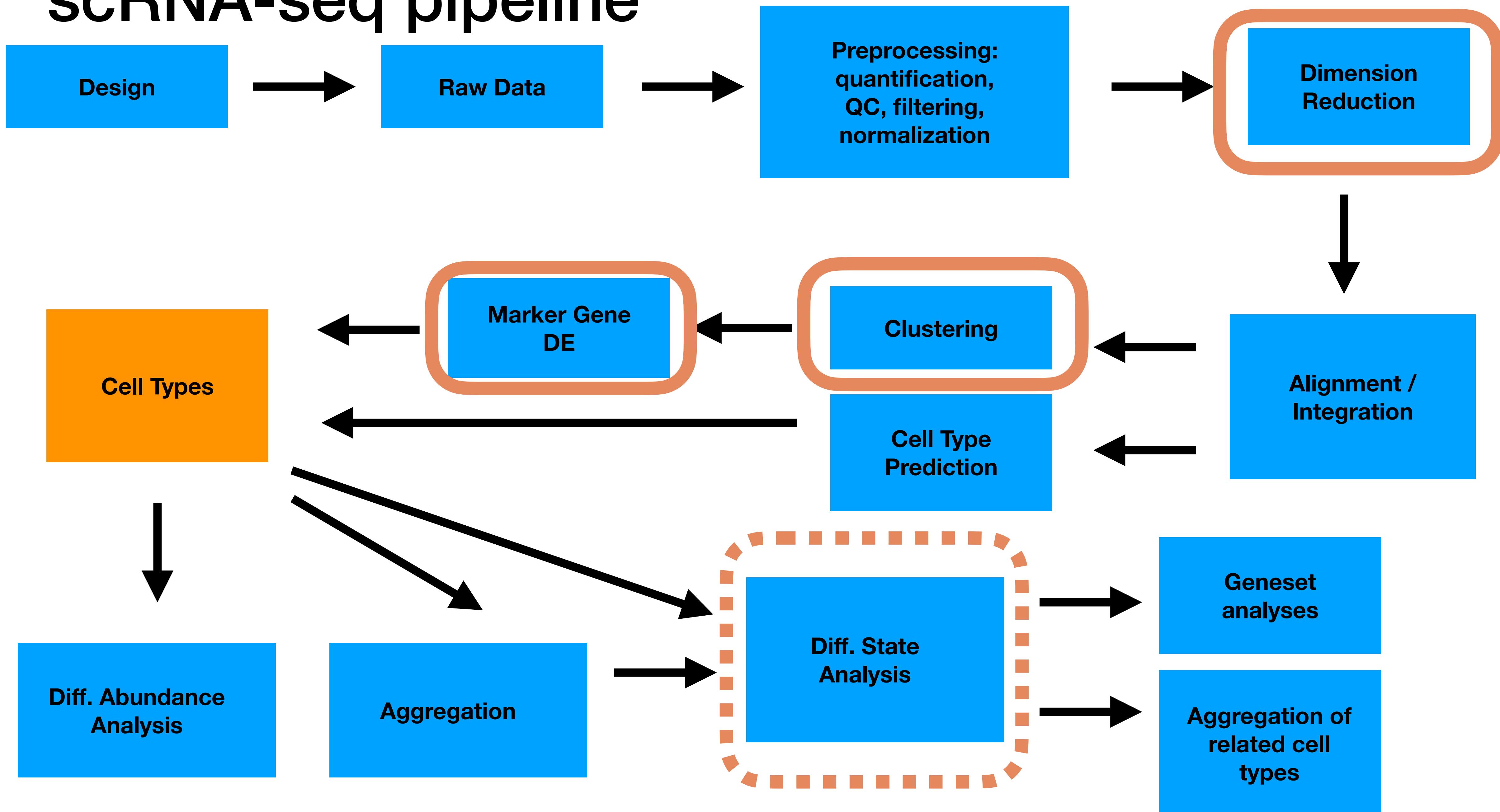


Figure 2. (A) Normalized run times, using RtsneKmeans as the reference method, across all data set instances and number of clusters. (B) Run time versus performance (ARI) for a subset of data sets and filterings, for the true number of clusters.



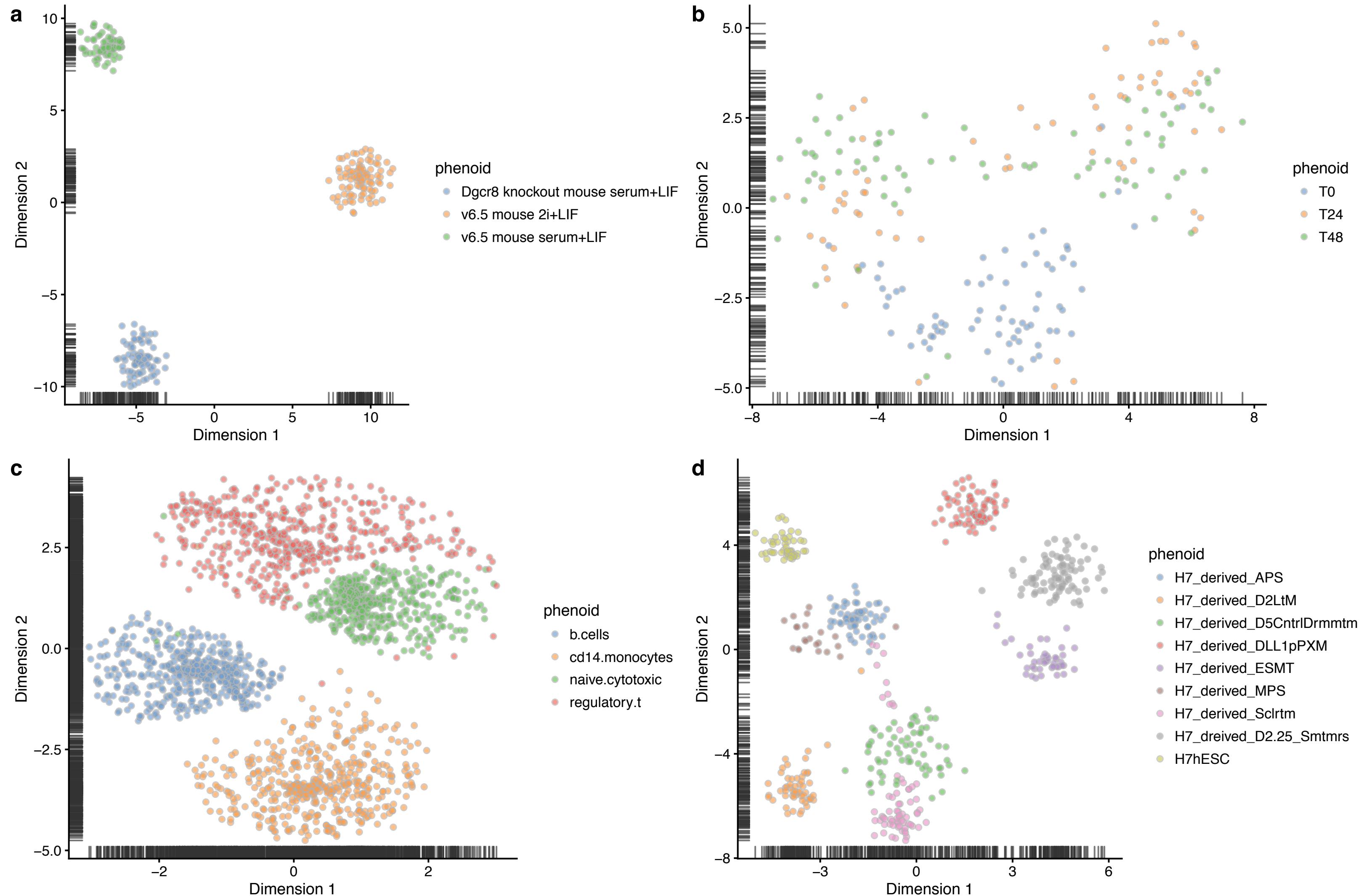
(Marker gene) Differential expression of single cell RNA-seq data

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.

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:

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

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

MAST

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.

mixture model

hurdle model



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

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

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

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

SCDE

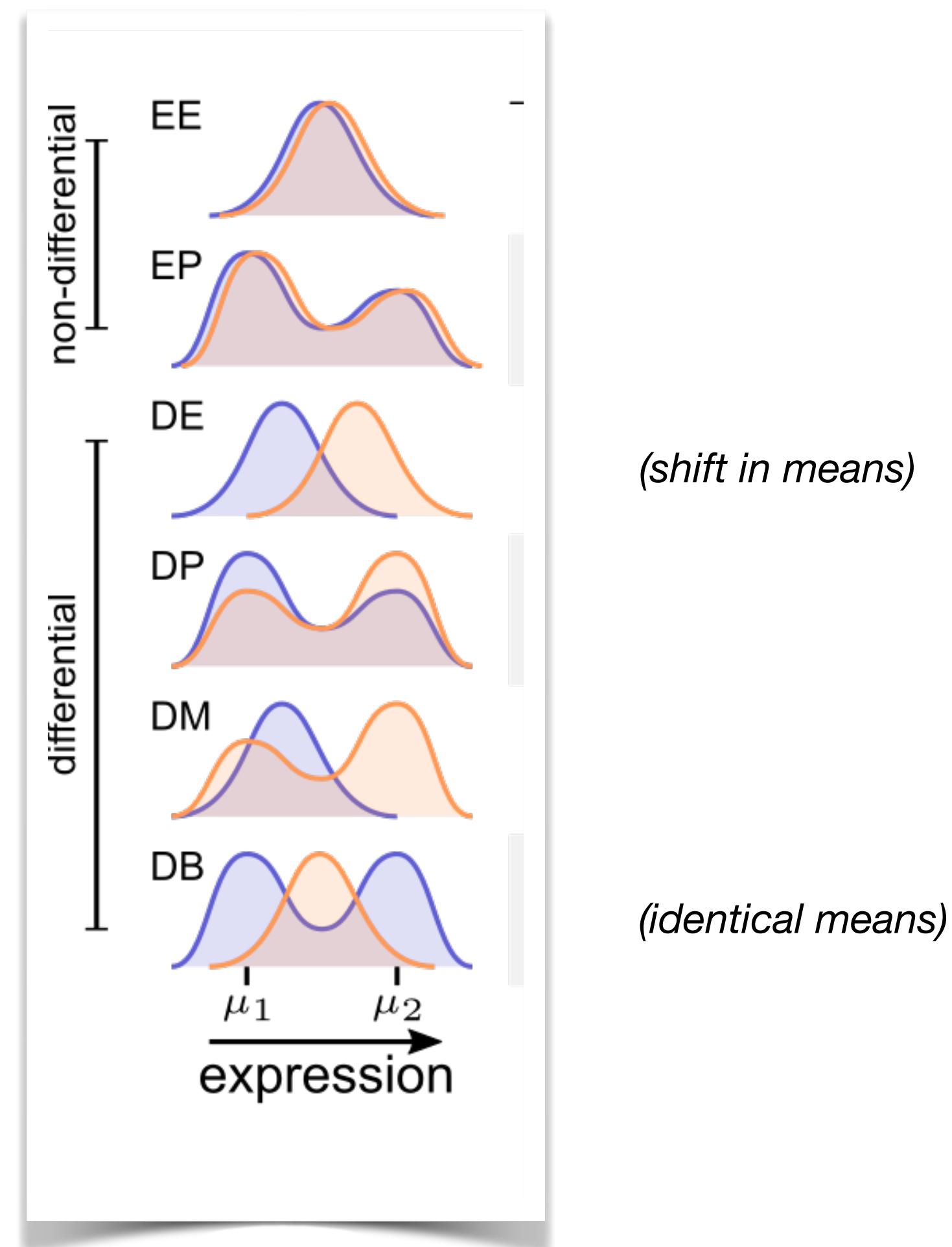
where p_d is the probability of observing a dropout event in cell c for a gene expressed at an average level x in S , $p_{\text{Poisson}}(x)$ and $p_{\text{NB}}(x \mid r_c)$ are the probabilities of observing expression magnitude of r_c in case of a dropout (Poisson) or successful amplification (NB) 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 Expression
- Equivalent Proportions
- Differential Expression
- Differential Proportions
- Differential Modality
- Both, Differential modality & component means



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

Genome Biology

Open Access



CrossMark

METHOD

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

Keegan D. Korthauer^{1,2}, Li-Fang Chu³, Michael A. Newton^{4,5}, Yuan Li⁵, James Thomson^{3,6,7}, Ron Stewart³ and Christina Kendziora^{4,5*}

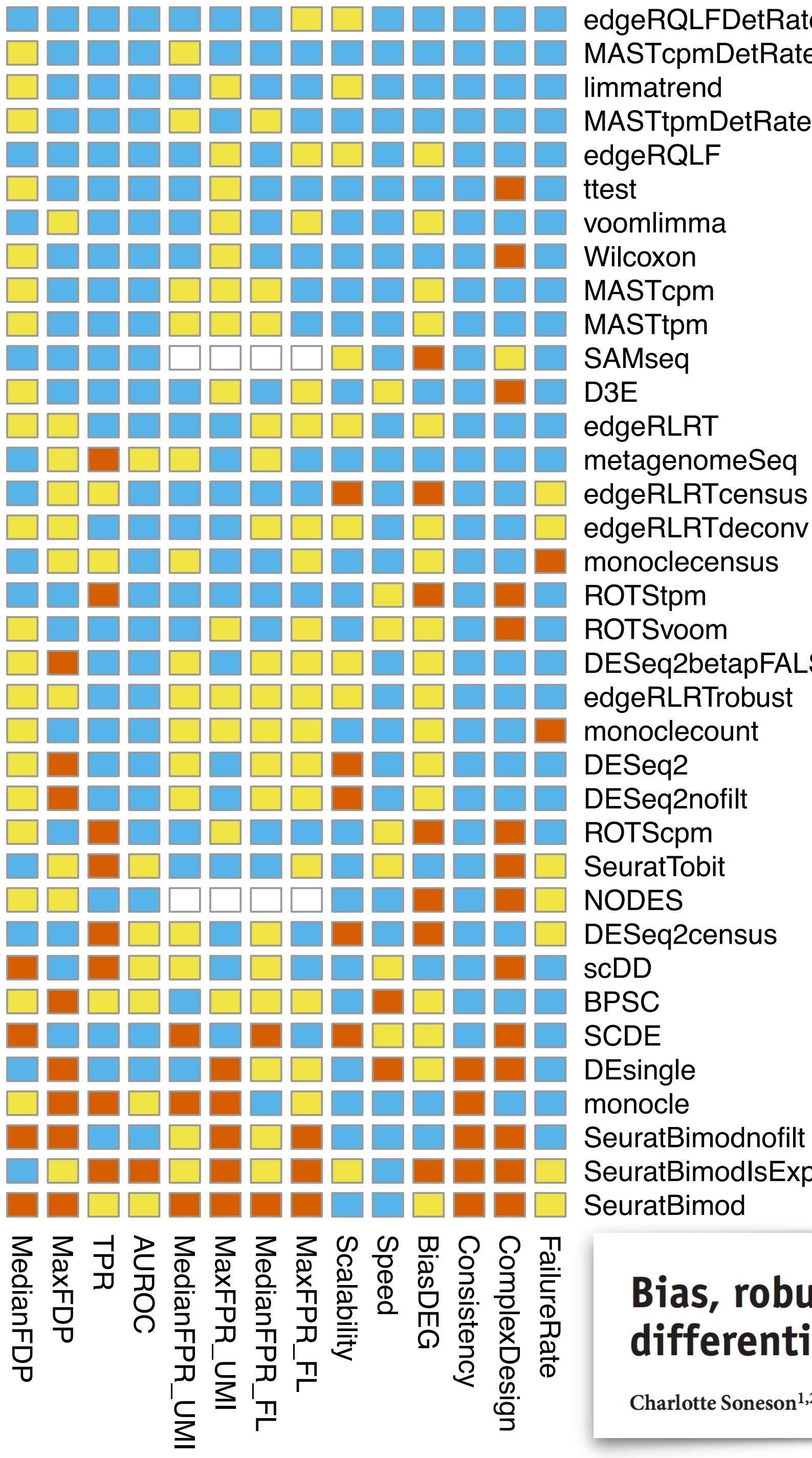
Punchline

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”

Criteria



Methods

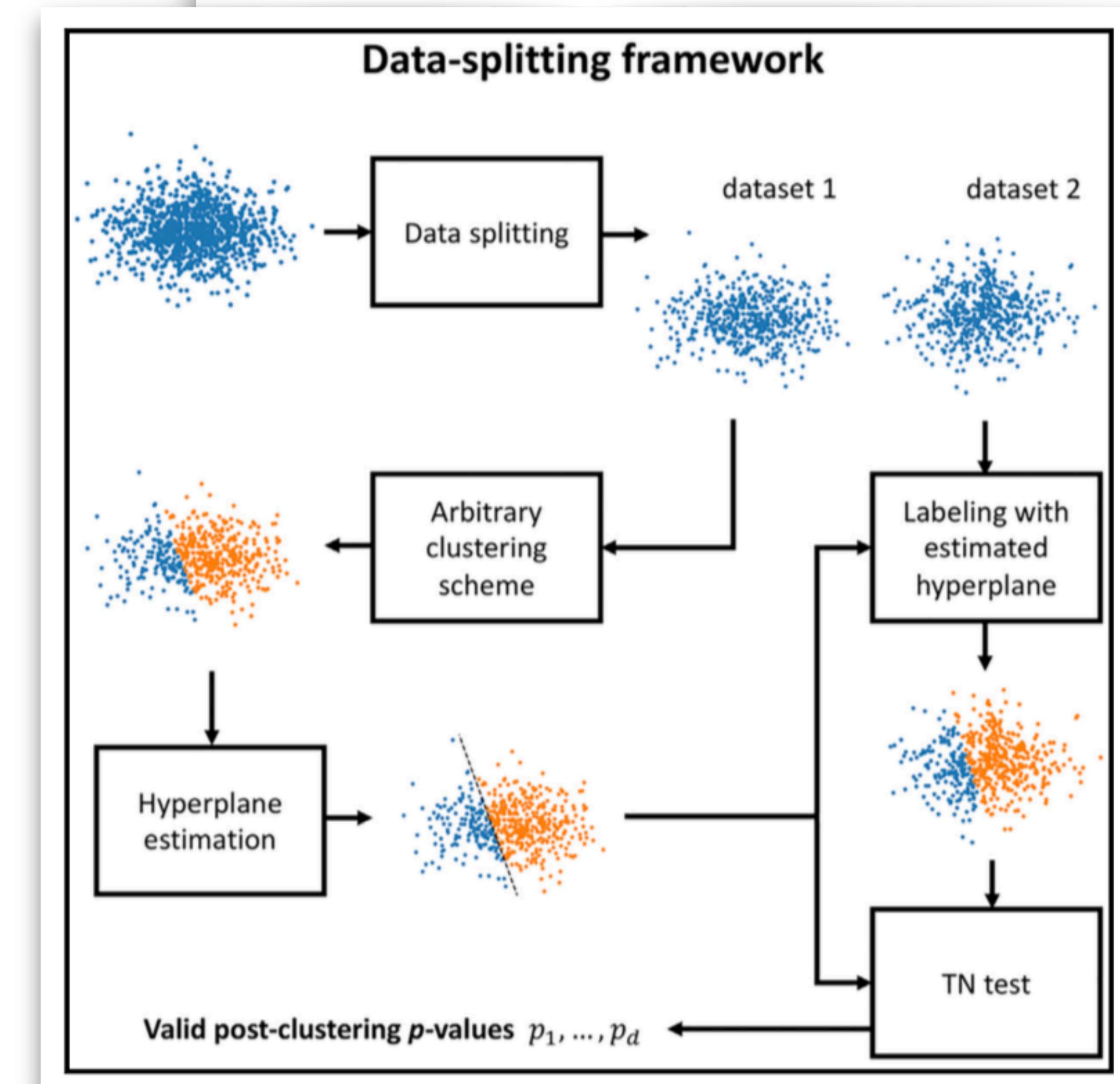
Bias, robustness and scalability in single-cell differential expression analysis

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 (Shannon) entropy? Which has low entropy?

