



Single cell analysis

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



Why single cell?

“Bulk” versus single-cell

Discover and quantify abundance
of (new) cell types

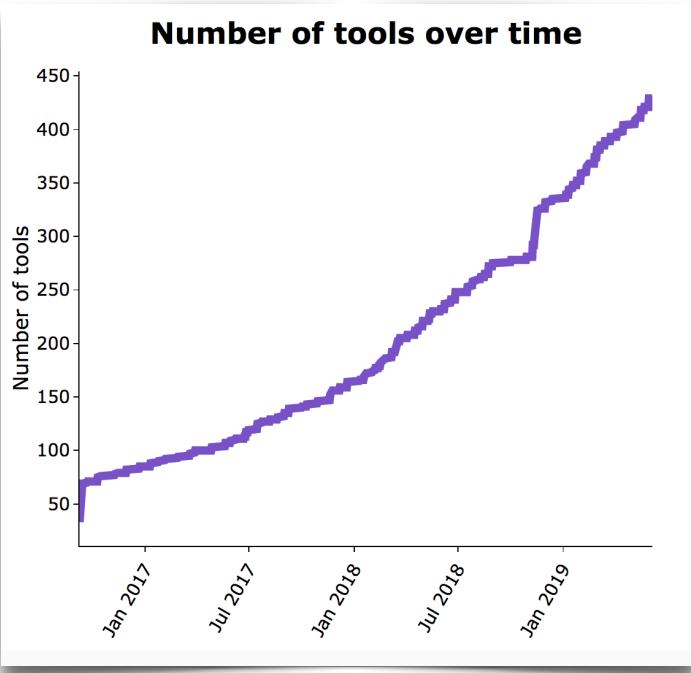
Study heterogeneity of gene
expression

Computational and analytical challenges in single-cell transcriptomics

Oliver Stegle¹, Sarah A. Teichmann^{1,2} and John C. Marioni^{1,2}

However, there are also important biological questions for which bulk measures of gene expression are insufficient¹⁴. 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 experimentally¹⁸. 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}.

Computational Tools

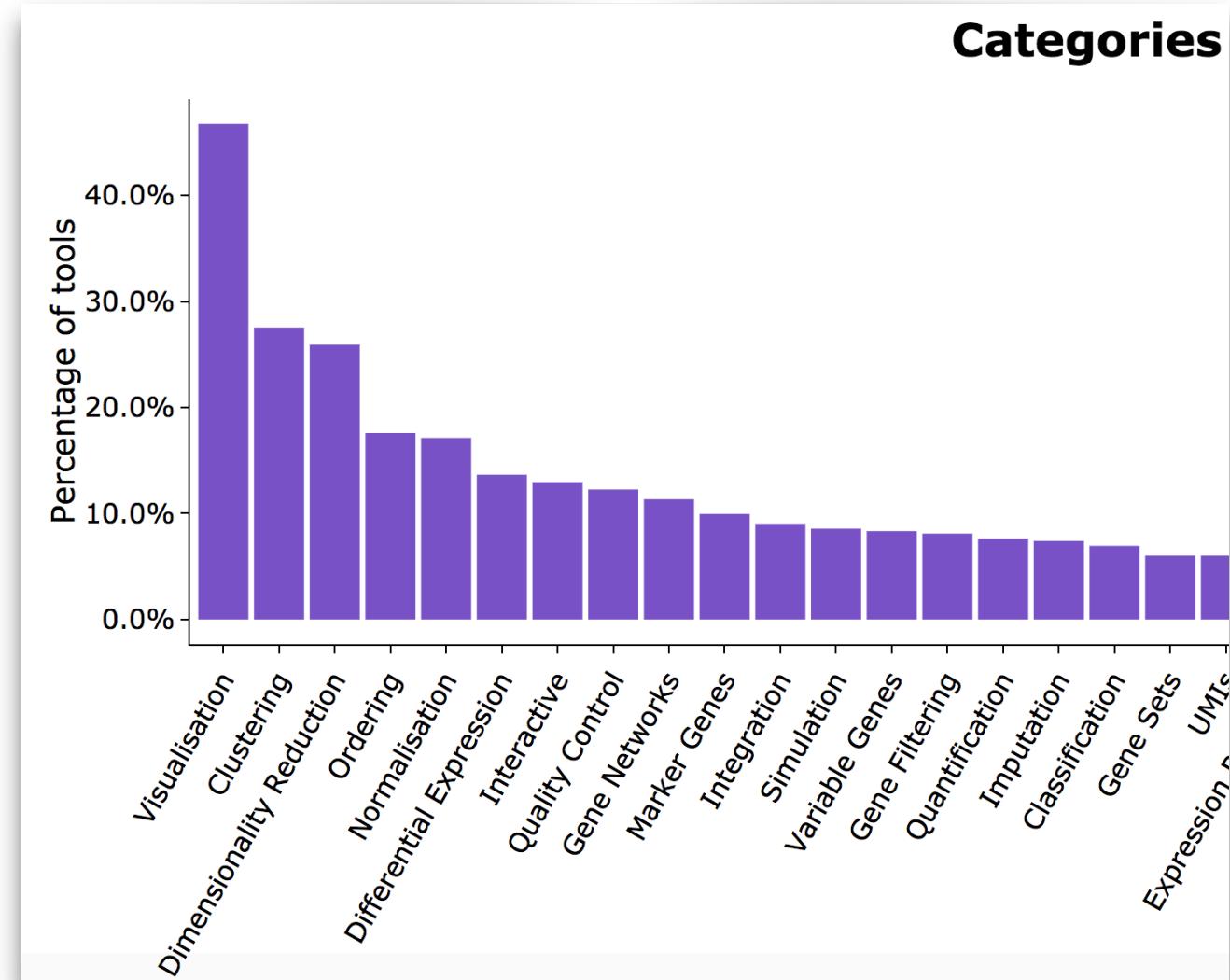


RESEARCH ARTICLE

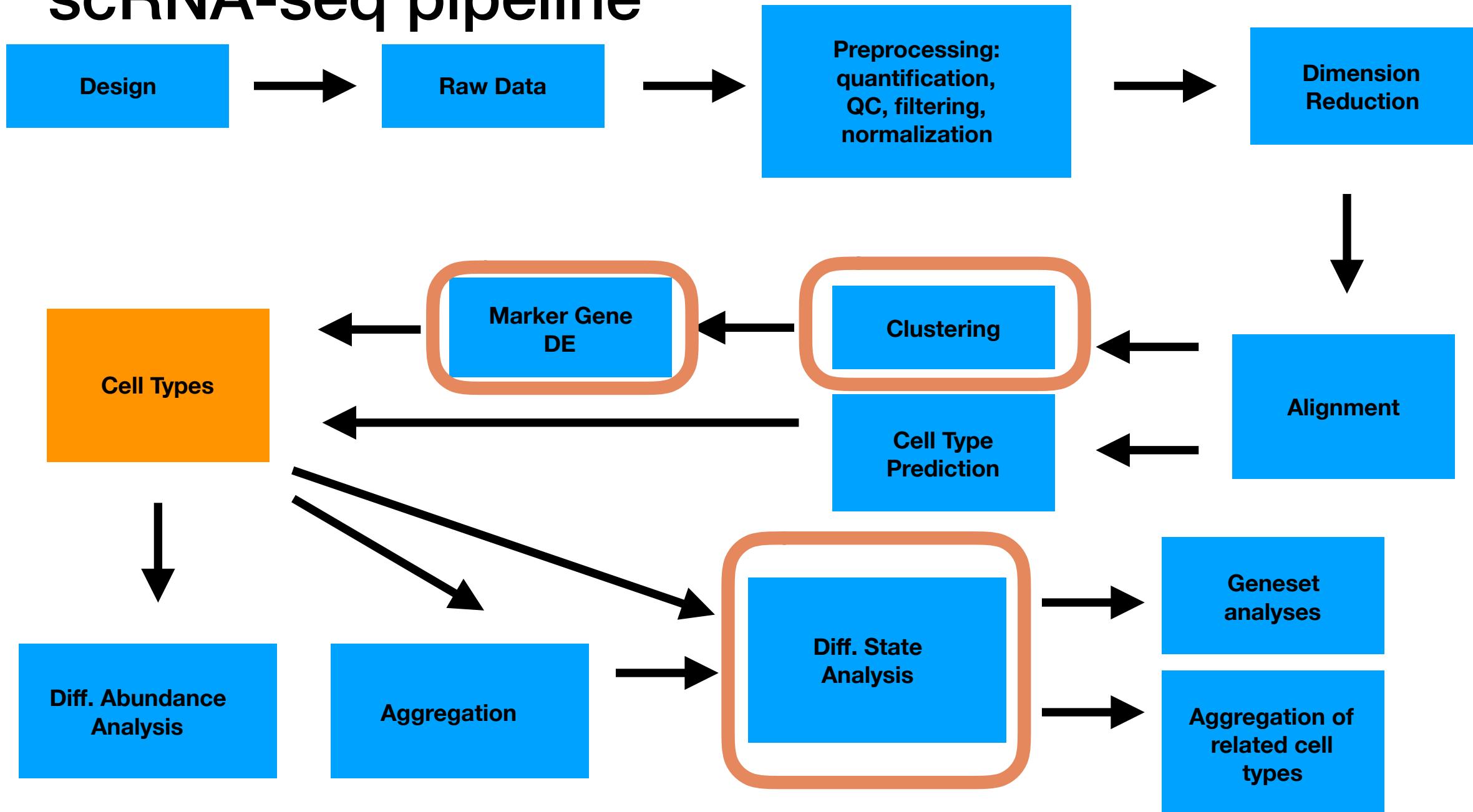
Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database

Luke Zappia^{1,2}, Belinda Phipson¹, 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



scRNA-seq pipeline





An application (motivation)

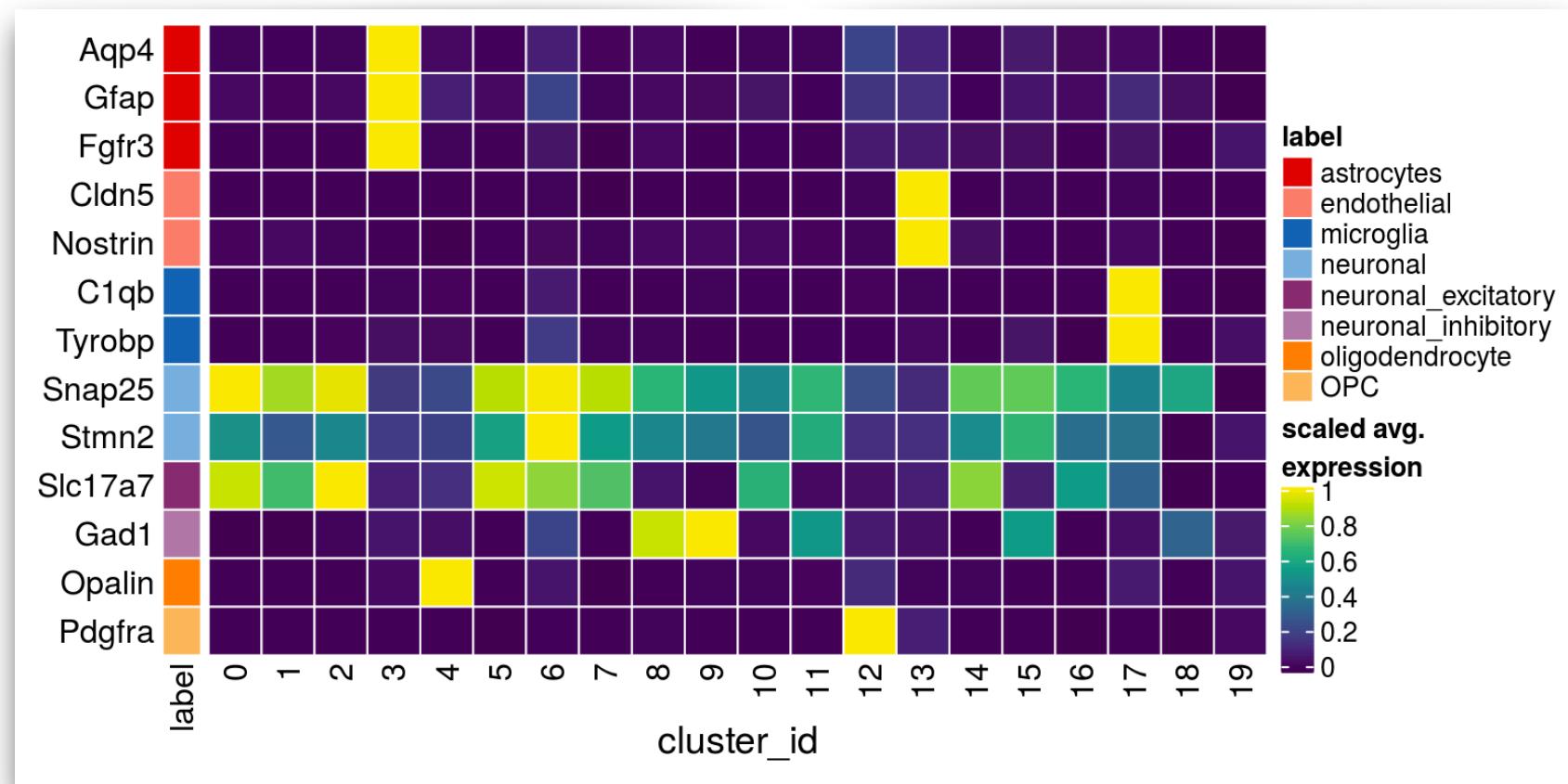
Application to LPS dataset: clustering + annotation of subpopulations

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

frontal cortex

single nuclei RNA-seq (10x)

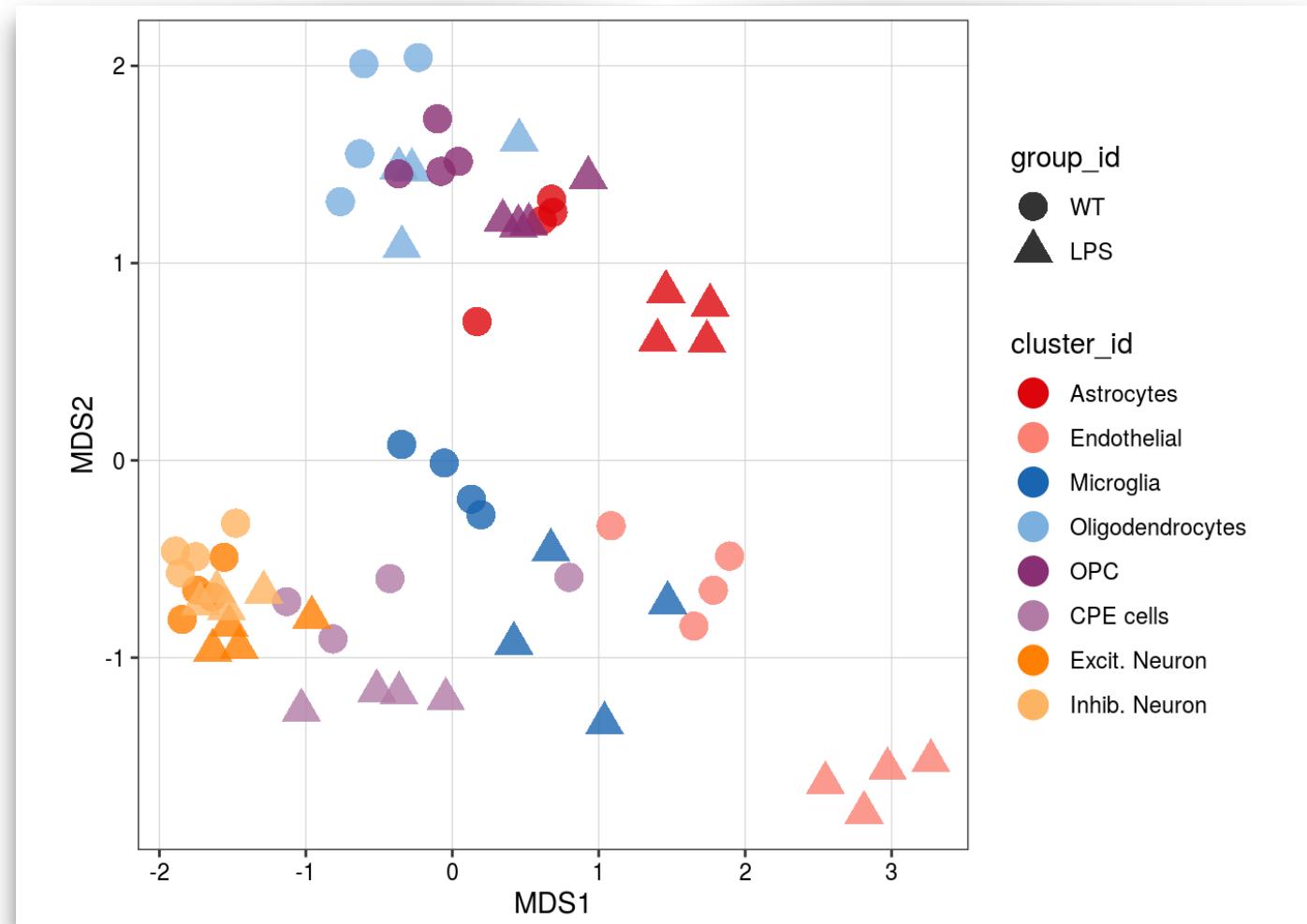
usual preprocessing:
filtering, doublet removal,
Seurat integration,
clustering



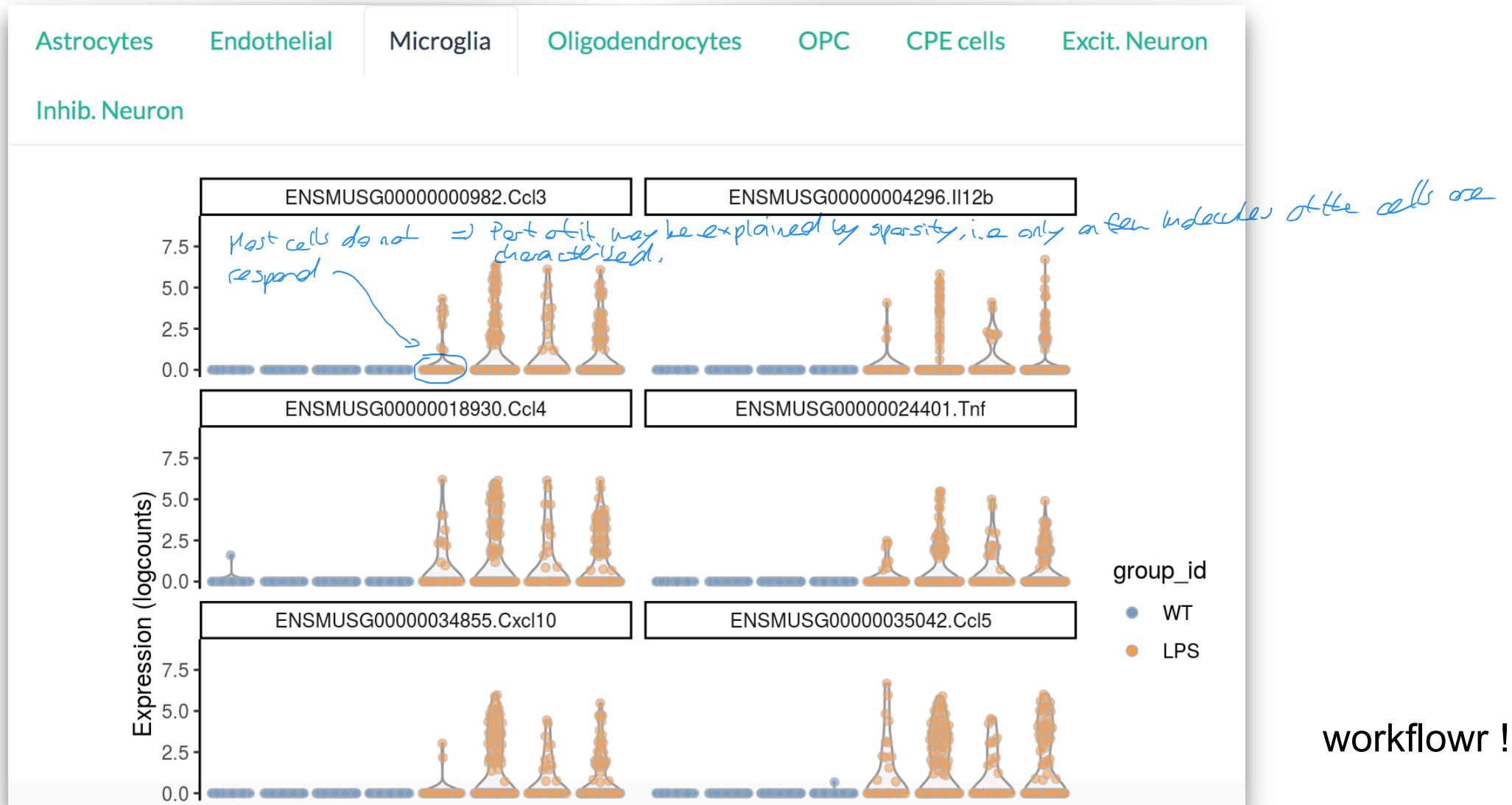
Application to LPS dataset: subpopulation-level visualization

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

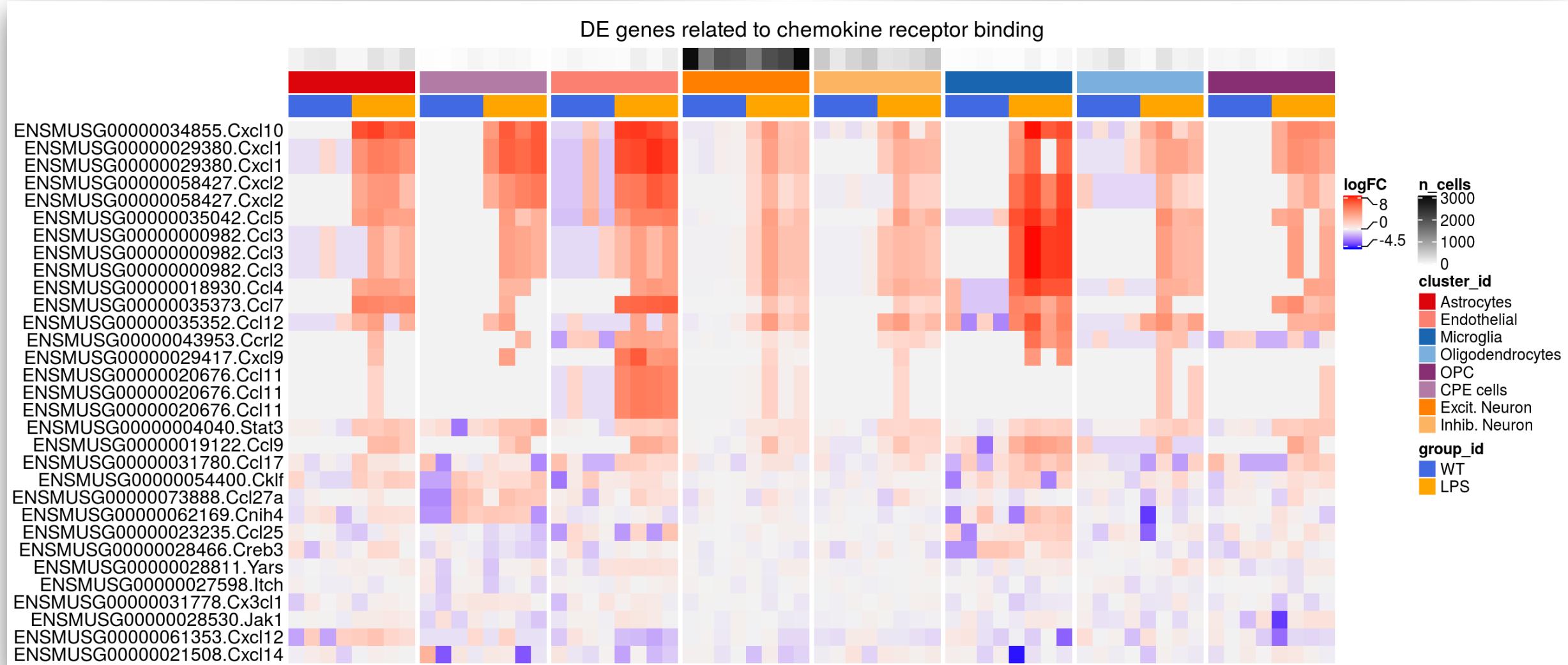
Each dot is one subpopulation/
sample combination



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



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



LPS dataset: interplay of cell type and cell state



cluster_id

● Astrocytes
● Endothelial
● Microglia

● Oligodendrocytes
● OPC
● CPE cells

● Excit. Neuron
● Inhib. Neuron

group_id

● WT
● LPS



Dimension reduction

Dimensionality reduction (generally)

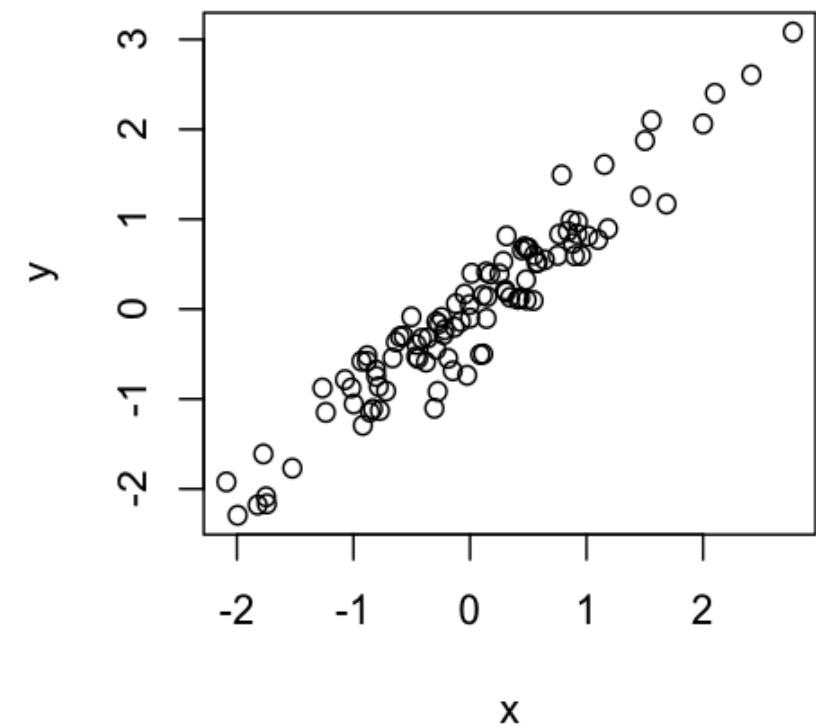
Techniques exist to **project** high-dimensional data (typical situation: 5k-10k gene expression measurements for each of N cells/samples) into a small number of dimensions (2 or 3)

Many techniques: **linear PCA**, multidimensional scaling (MDS), t-distributed stochastic neighbor embedding (**tSNE**), UMAP, ICA, diffusion maps,

Linear PCA: uses a linear combination of original variables such that the components decrease in variability (highest variance first) and are orthogonal to previous dimensions. Often, first 2 or 3 are used.

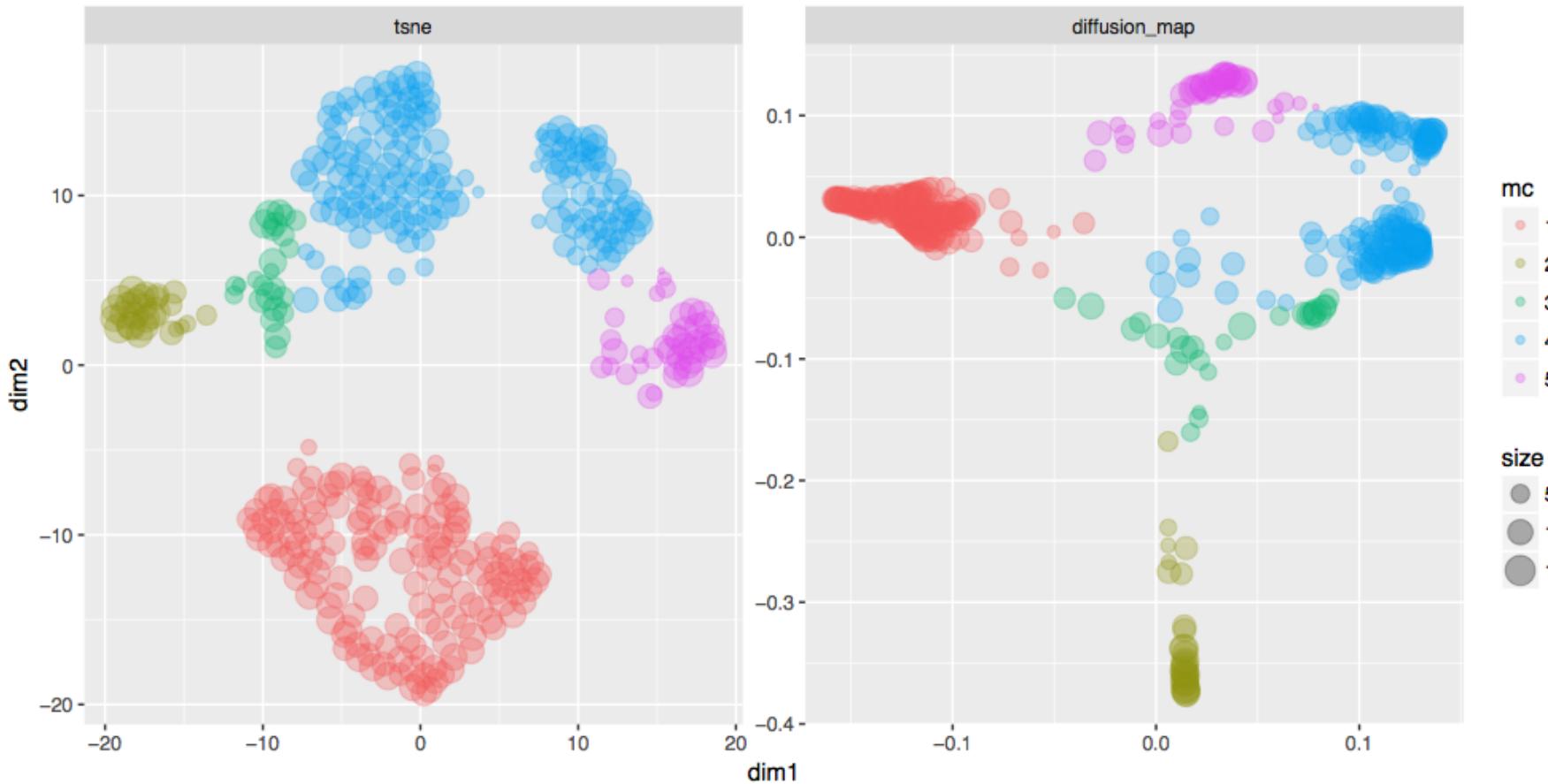
Visual explanation:

<http://setosa.io/ev/principal-component-analysis/>





tSNE (t-dist'd stochastic neighbour embedding) + diffusion maps

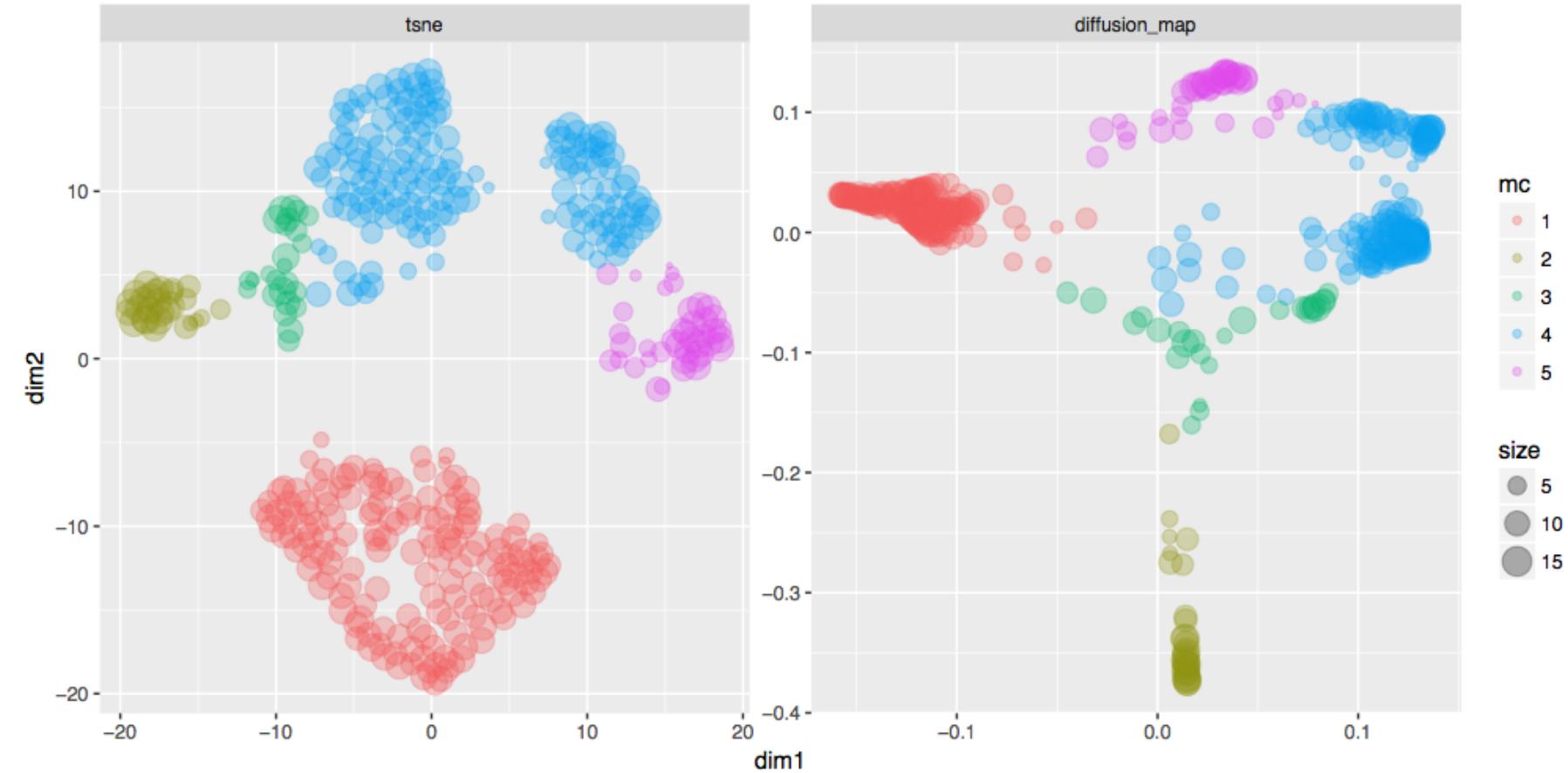


"if you haven't encountered t-SNE before, here's what you need to know about the math behind it. The goal is to take a set of points in a high-dimensional space and find a faithful representation of those points in a lower-dimensional space, typically the 2D plane. The algorithm is non-linear and adapts to the underlying data, performing different transformations on different regions. Those differences can be a major source of confusion."



tSNE (t-dist'd stochastic neighbour embedding) + diffusion maps

“Given data in a high-dimensional space .. find parameters that describe the lower-dimensional structures of which it is comprised. Unlike other popular methods such as PCA and MDS, diffusion maps are non-linear and focus on discovering the underlying manifold (lower-dimensional constrained “surface” upon which the data is embedded). By integrating local similarities at different scales, a global description of the data-set is obtained.





What dimension reduction should I use?

“Briefly, for cell clustering analysis, PCA, ICA, FA, NMF, and ZINB-WaVE are recommended for small data where computation is not a concern. PCA, ICA, FA, NMF are also recommended for large data where computation is a concern.”

Generally, non-linear methods are used for visualization; linear methods are used for input to downstream tasks (e.g., clustering, trajectory, etc.)

bioRxiv preprint first posted online May. 17, 2019; doi: <http://dx.doi.org/10.1101/641142>. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Accuracy, Robustness and Scalability of Dimensionality Reduction Methods for Single Cell RNAseq Analysis

Shiquan Sun^{1, 2}, Jiaqiang Zhu², Ying Ma² and Xiang Zhou^{2, 3, #}

1. School of Computer Science, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, P.R. China

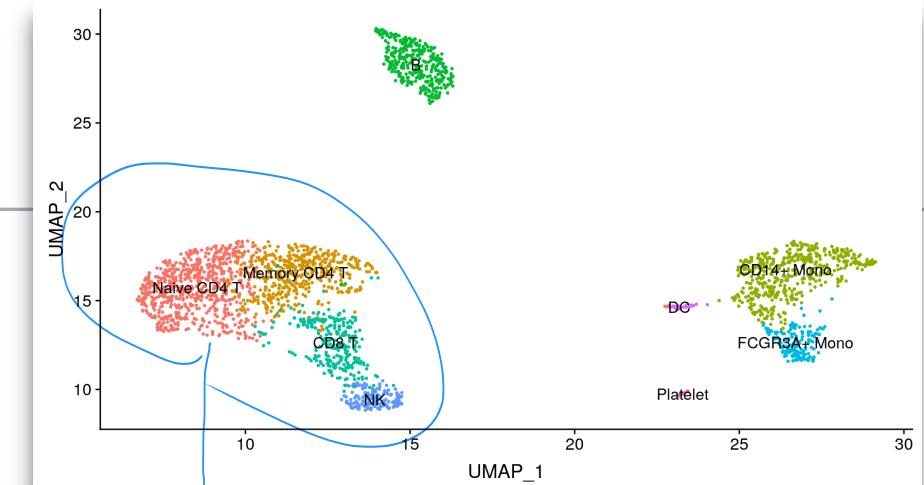
2. Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA

3. Center for Statistical Genetics, University of Michigan, Ann Arbor, MI 48109, USA

#: correspondence to XZ (xzhousph@umich.edu)

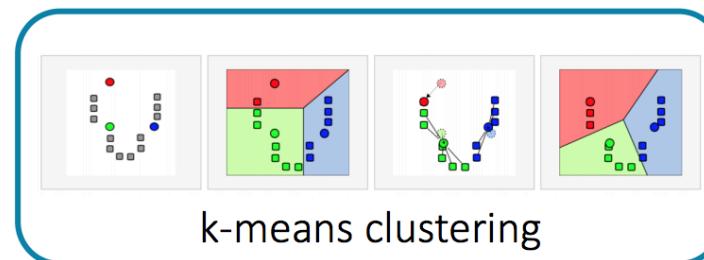
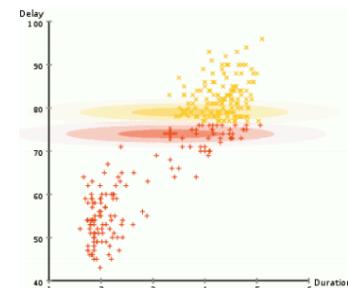
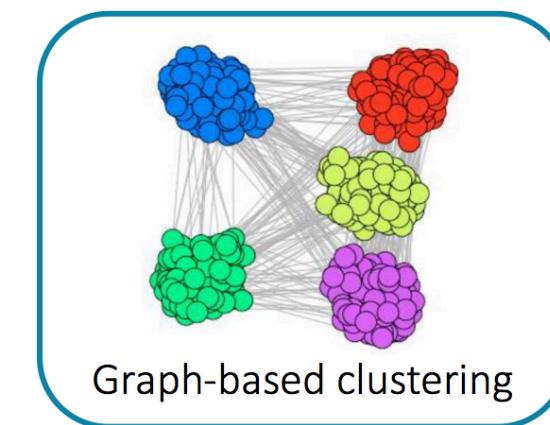
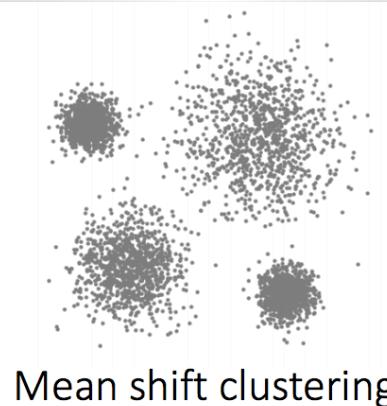
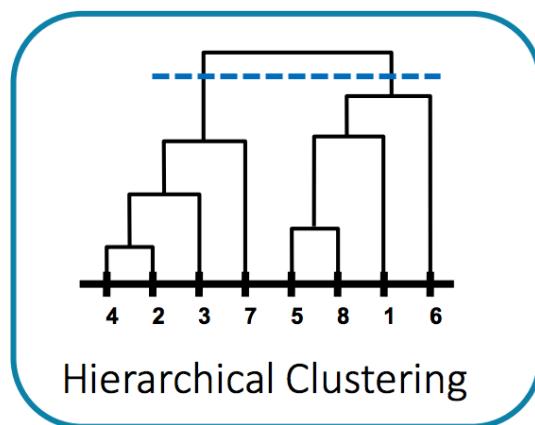


Clustering: an ill-posed problem?



clustering is very dependent on the variability of the features & thus the selected features.

If we simply focus on the high variability components of the group we may get a better separation.



12



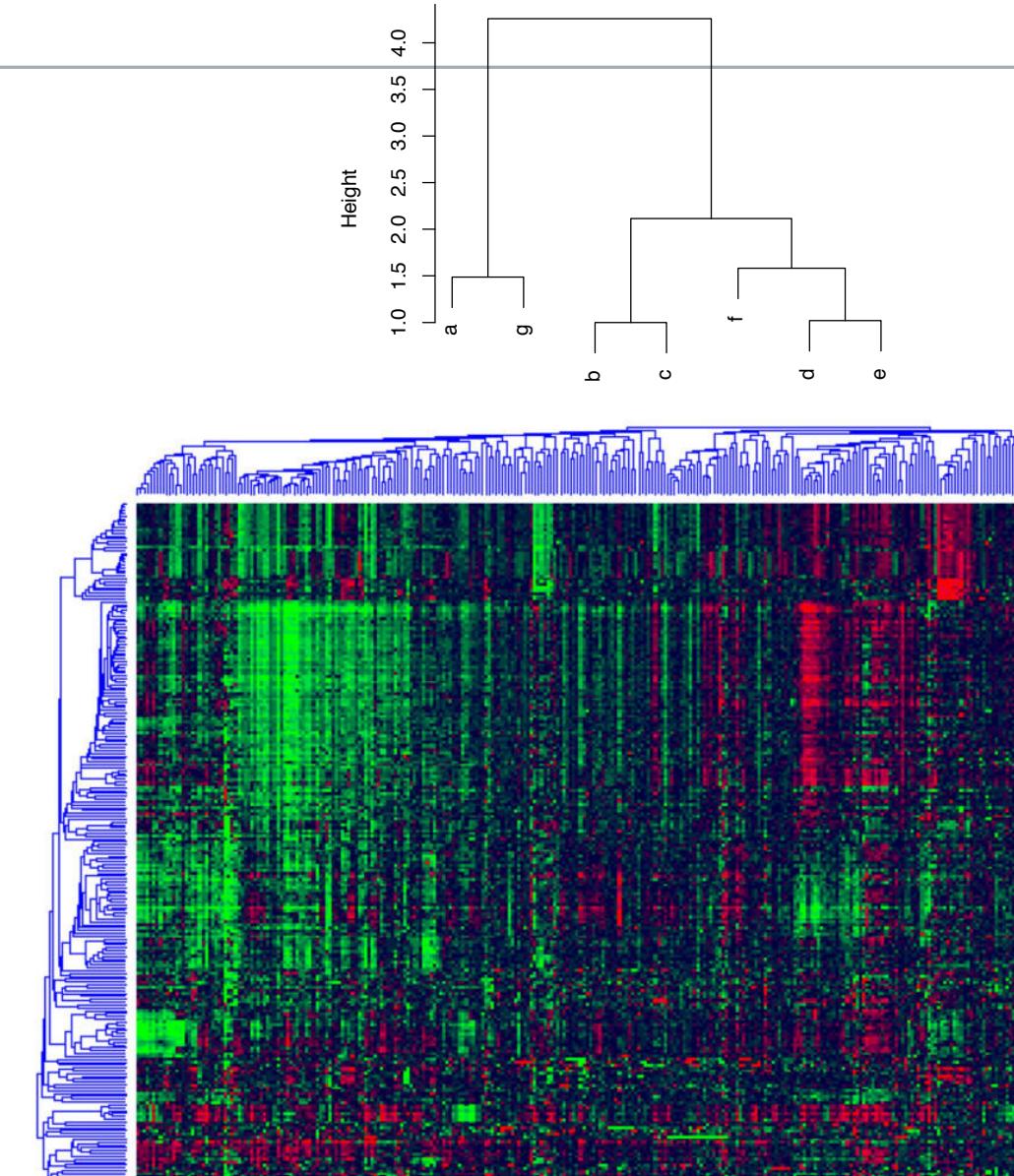
Hierarchical (Agglomerative) Clustering

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.

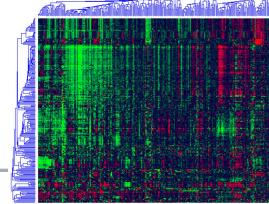
Linkage: determines how clusters are merged into a tree

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





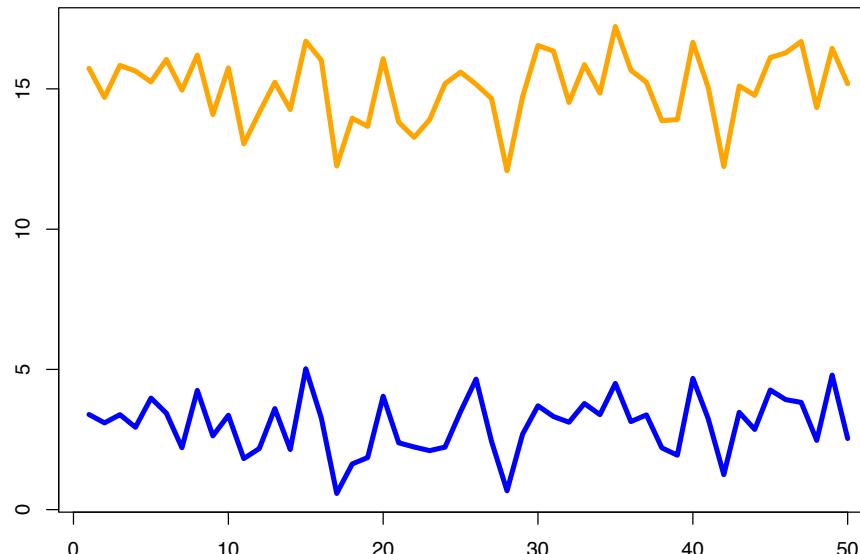
$$d(\mathbf{p}, \mathbf{q}) = d(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^n (q_i - p_i)^2}$$



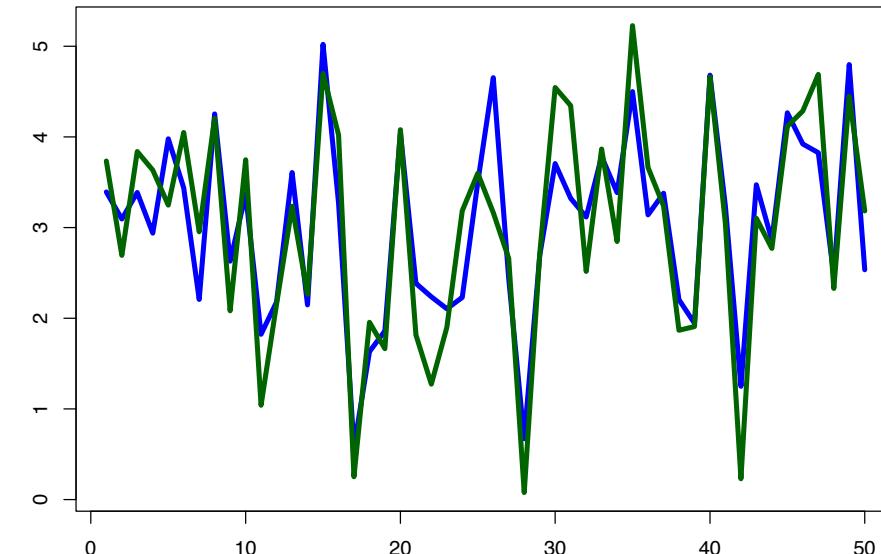
Are these “vectors” similar ?

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

It depends how you define similar.



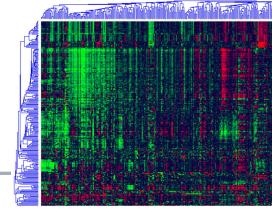
Euclidean distance: 84.84



3.92



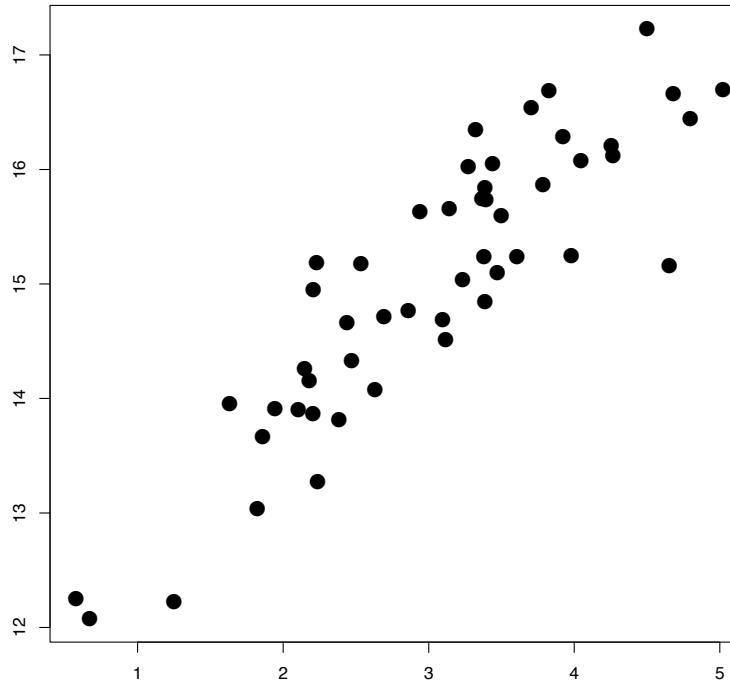
$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}},$$



Are these “vectors” similar ?

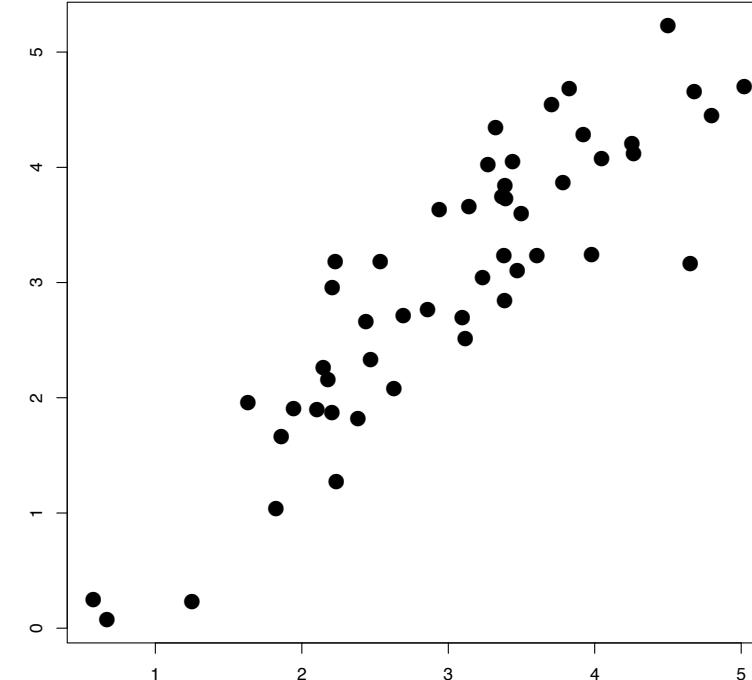
```
> cor(x,y)
[1] 0.8901139
> cor(x,y-12)
[1] 0.8901139
```

It depends how you define similar.

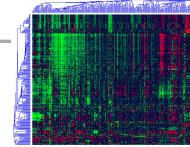


Correlation:

0.89



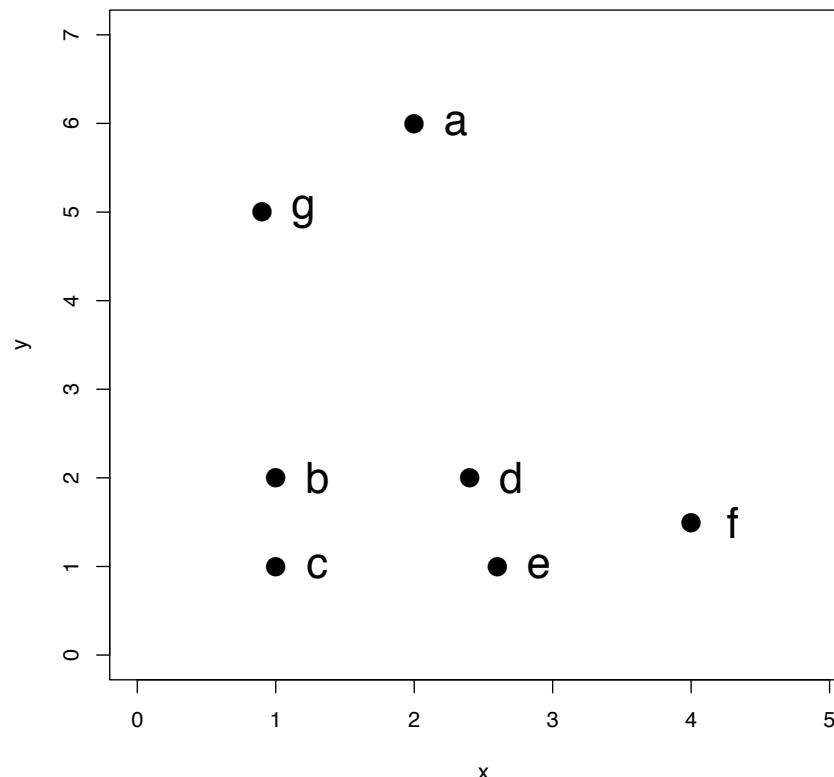
0.89



Hierarchical (Agglomerative) Clustering

Start with distances.

Linkage: determines how clusters are merged into a tree.

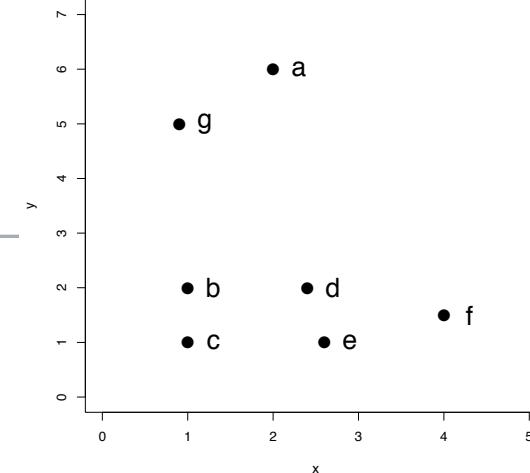
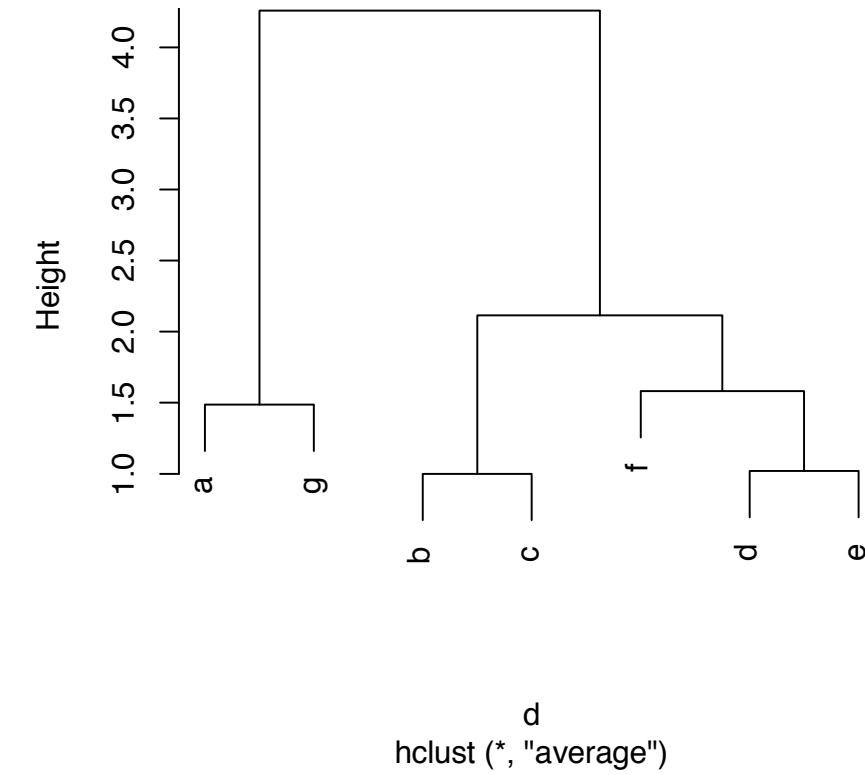
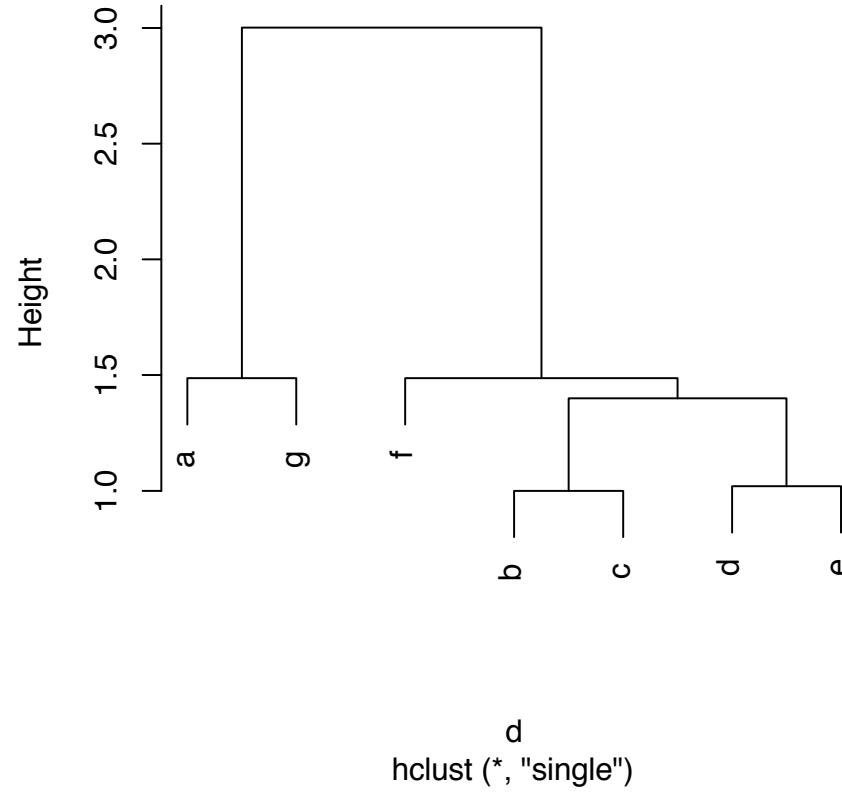


From eyeballing, here is a likely set of merges:

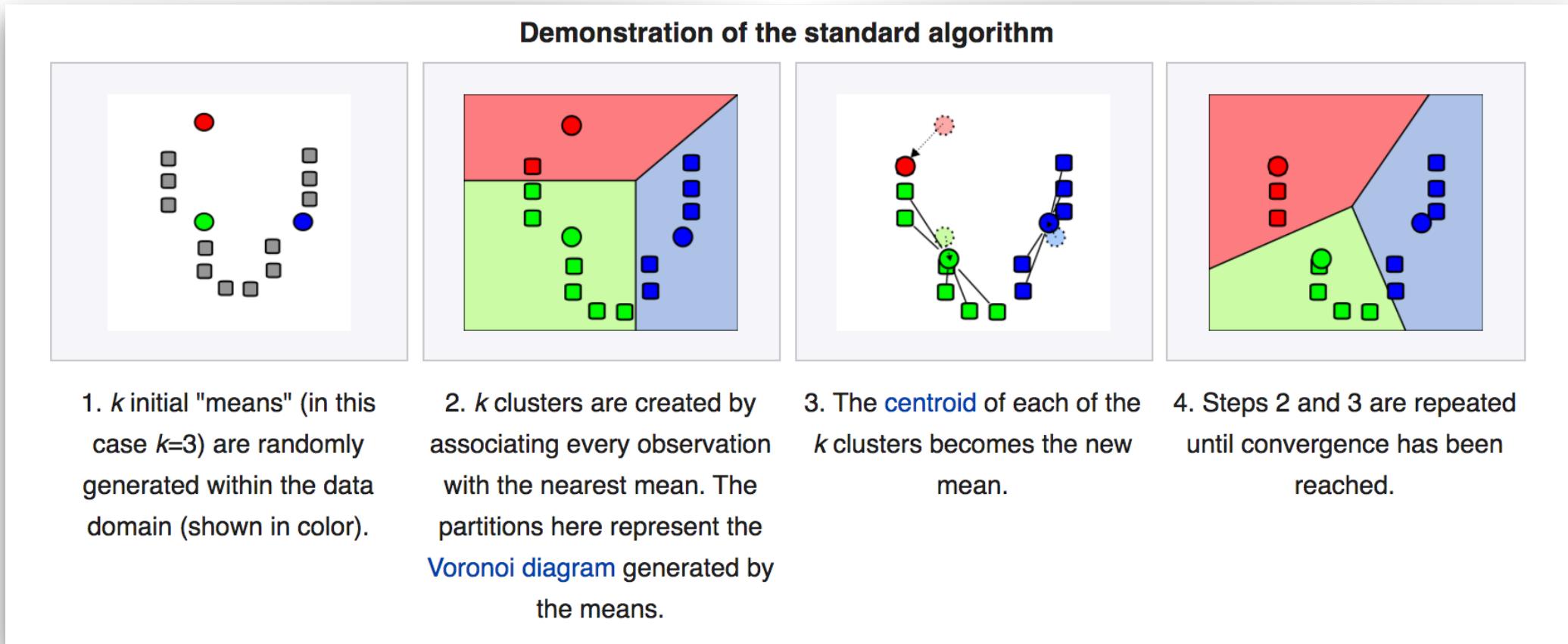
b,c
d,e
a,g,
(d,e),f
(b,c),((d,e),f)
ALL



Different linkages



k-means clustering (it has many variations)

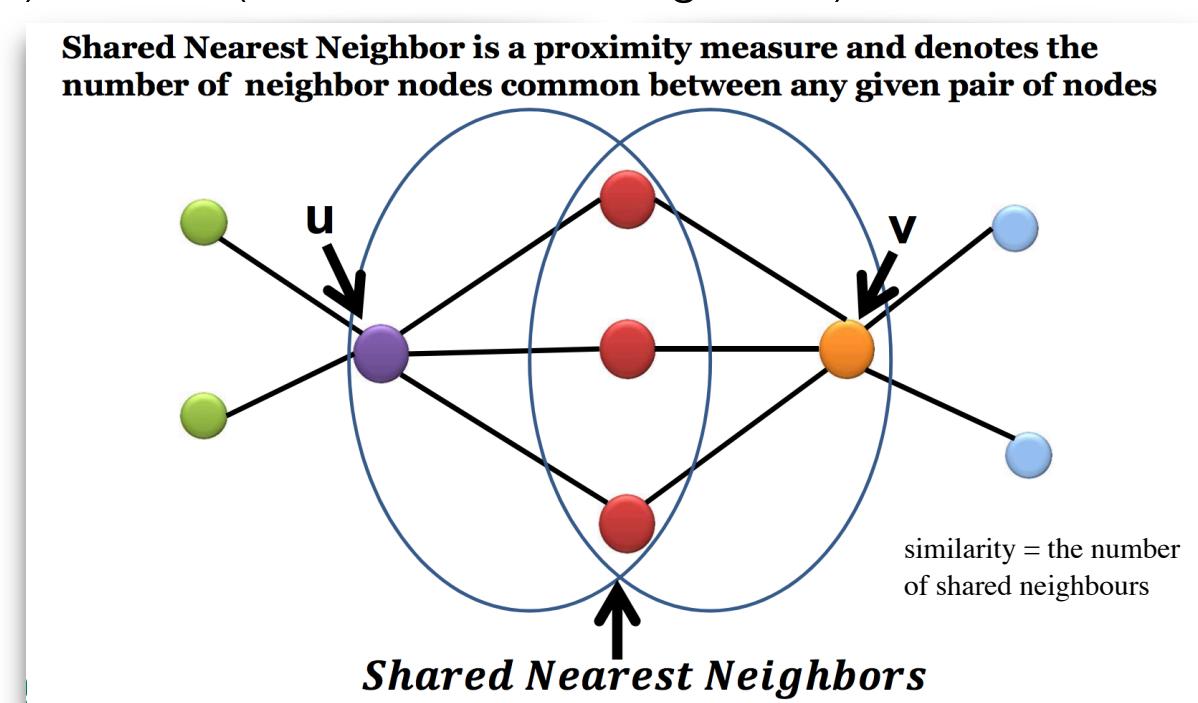


Disadvantages: hard to determine k , assumes spherical cluster shapes (depends on distance), different initial selection —> different clustering (typically run multiple times)

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.



From Louvain to Leiden: guaranteeing well-connected communities

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

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

⇒ Find modules with more connections inside the module rather than between.

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

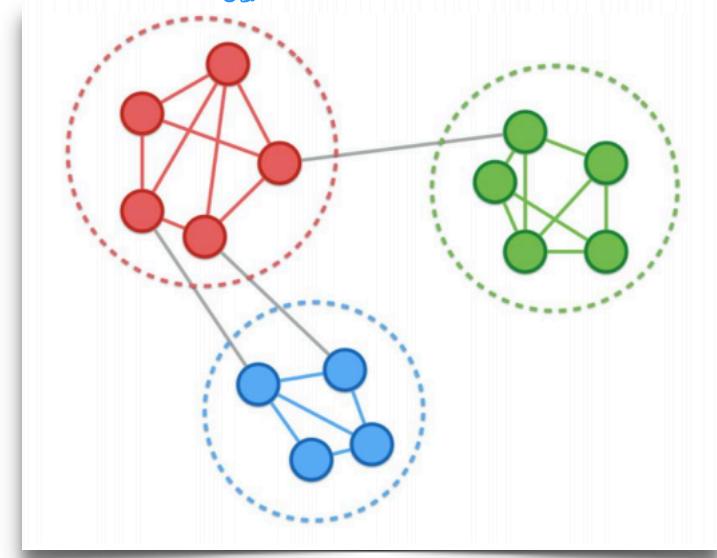
A_{ij} = edge weight between node i and j

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 $c_i = c_j$)



<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

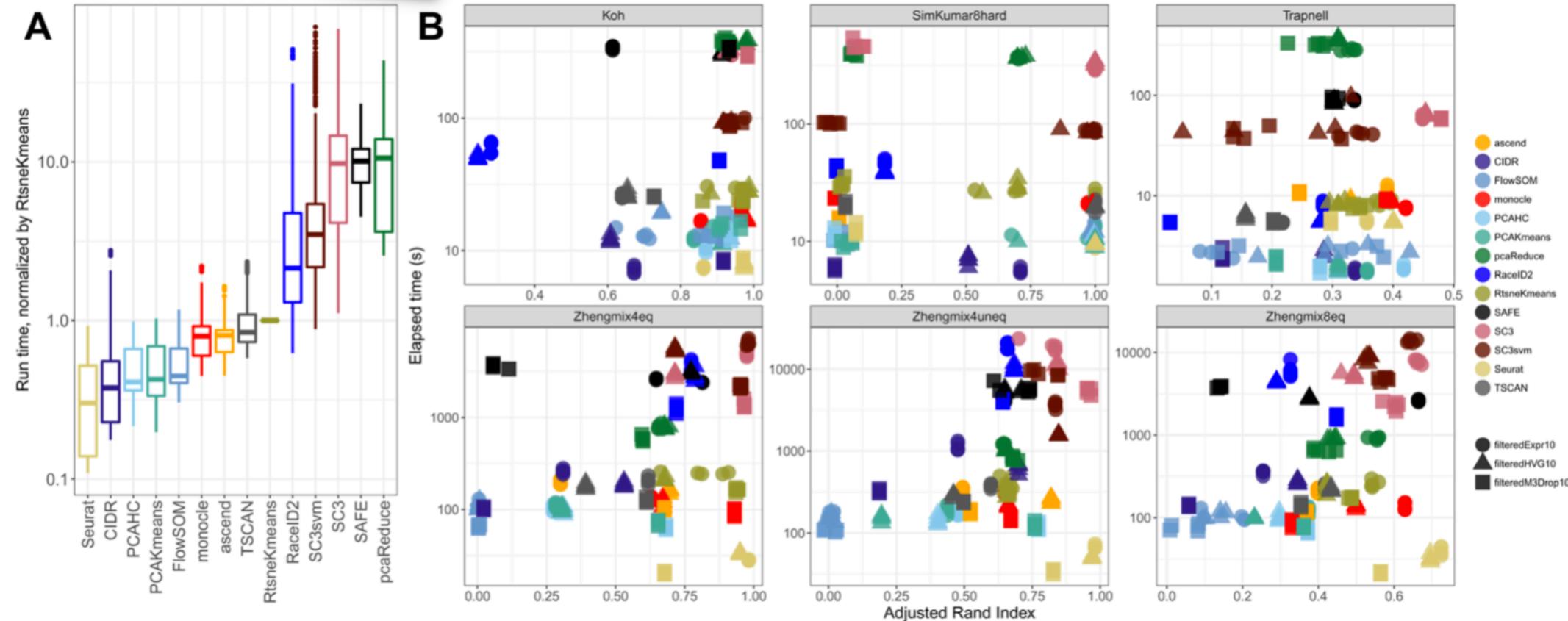


Figure 2. (A) Normalized run times, using RtsneKmeans as the reference method, across all data set instances and number of clusters. (B) Run time versus performance (ARI) for a subset of data sets and filterings, for the true number of clusters.



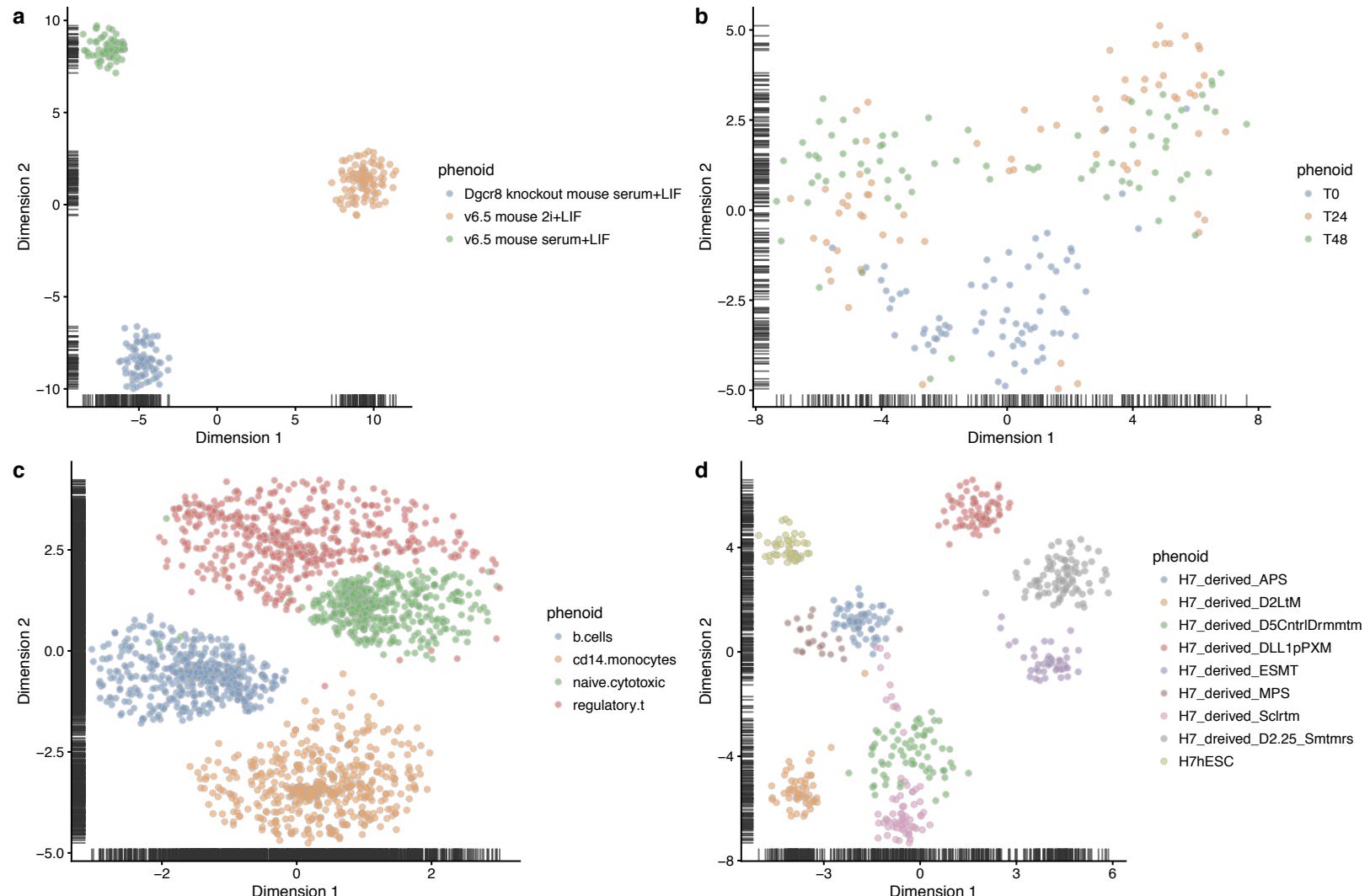
Differential expression of single cell RNA-seq data

How to cluster scRNA-seq data and/or how to find cell type markers ?

- Our strategy: datasets from conquer (<http://imlspenticton.uzh.ch:3838/conquer/>) with **predefined groups**: range of difficulty

→ By clustering & calculating p-values of DE genes we decrease the p-values (⇒ invalid p-values).

⇒ If we generate random noise & cluster the noise we would get more significant p-values by clustering.

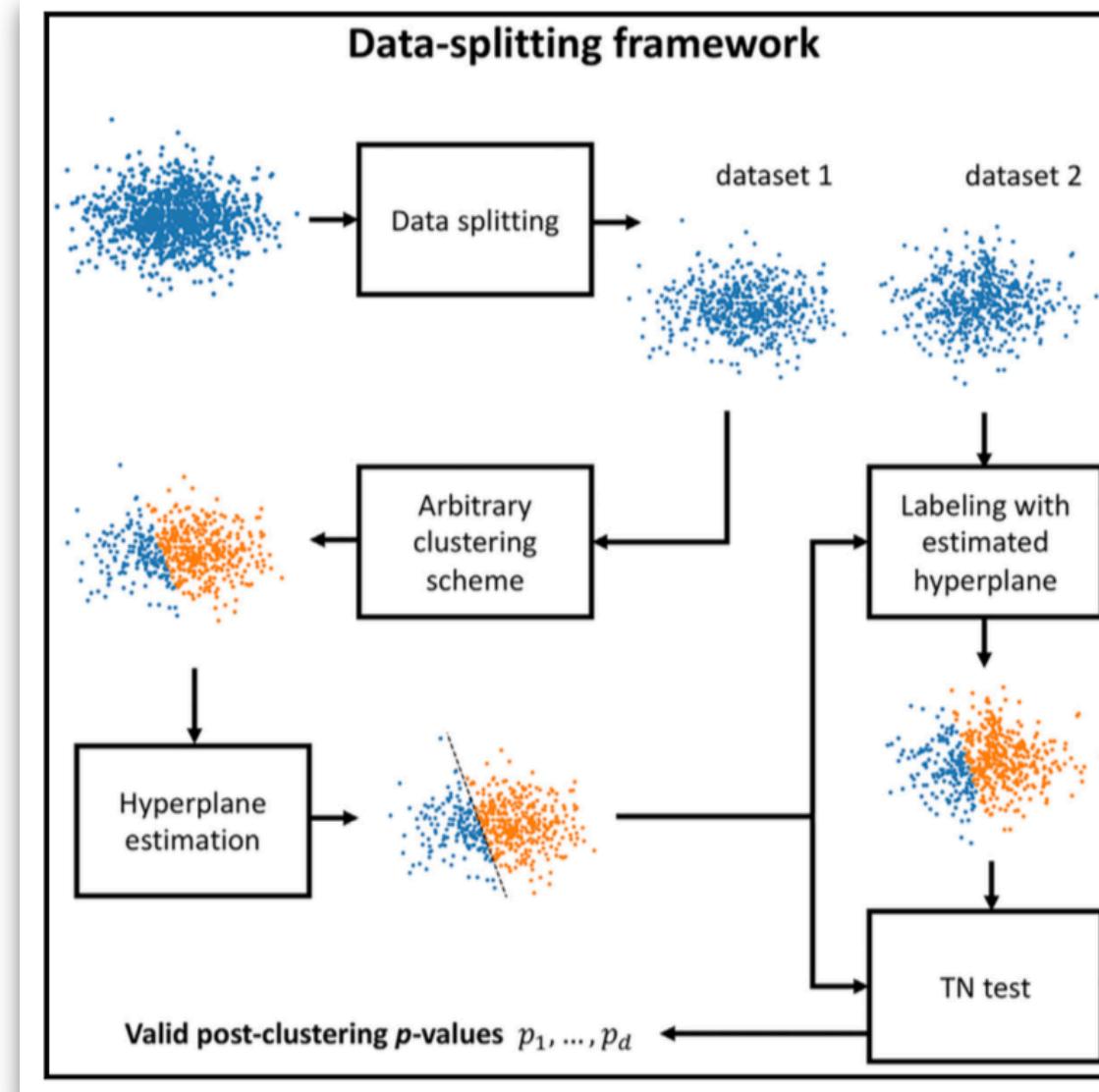


Statistical issue
lurking here:
clustering and
then testing
differences
between
clusters leads
to invalid P-
values

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

Authors

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



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:

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

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

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.

MAST

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

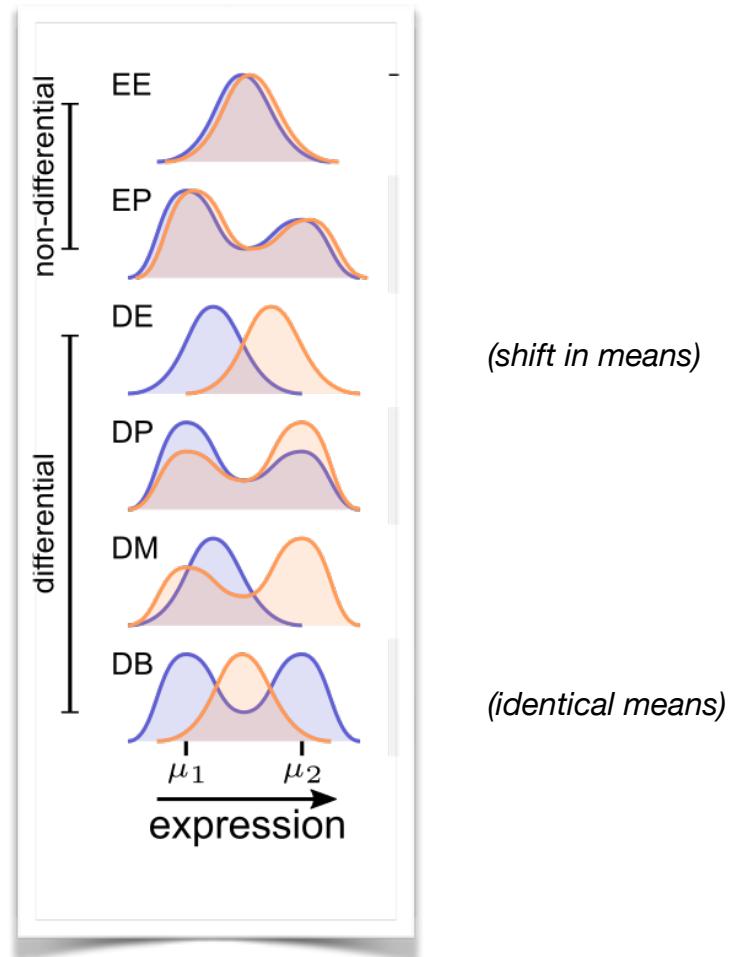
where p_d is the probability of observing a dropout event in cell c for a gene expressed at an average level x in S , $p_{\text{Poisson}}(x)$ and $p_{\text{NB}}(x \mid r_c)$ are the probabilities of observing expression magnitude of r_c in case of a dropout (Poisson) or successful amplification (NB) 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*}

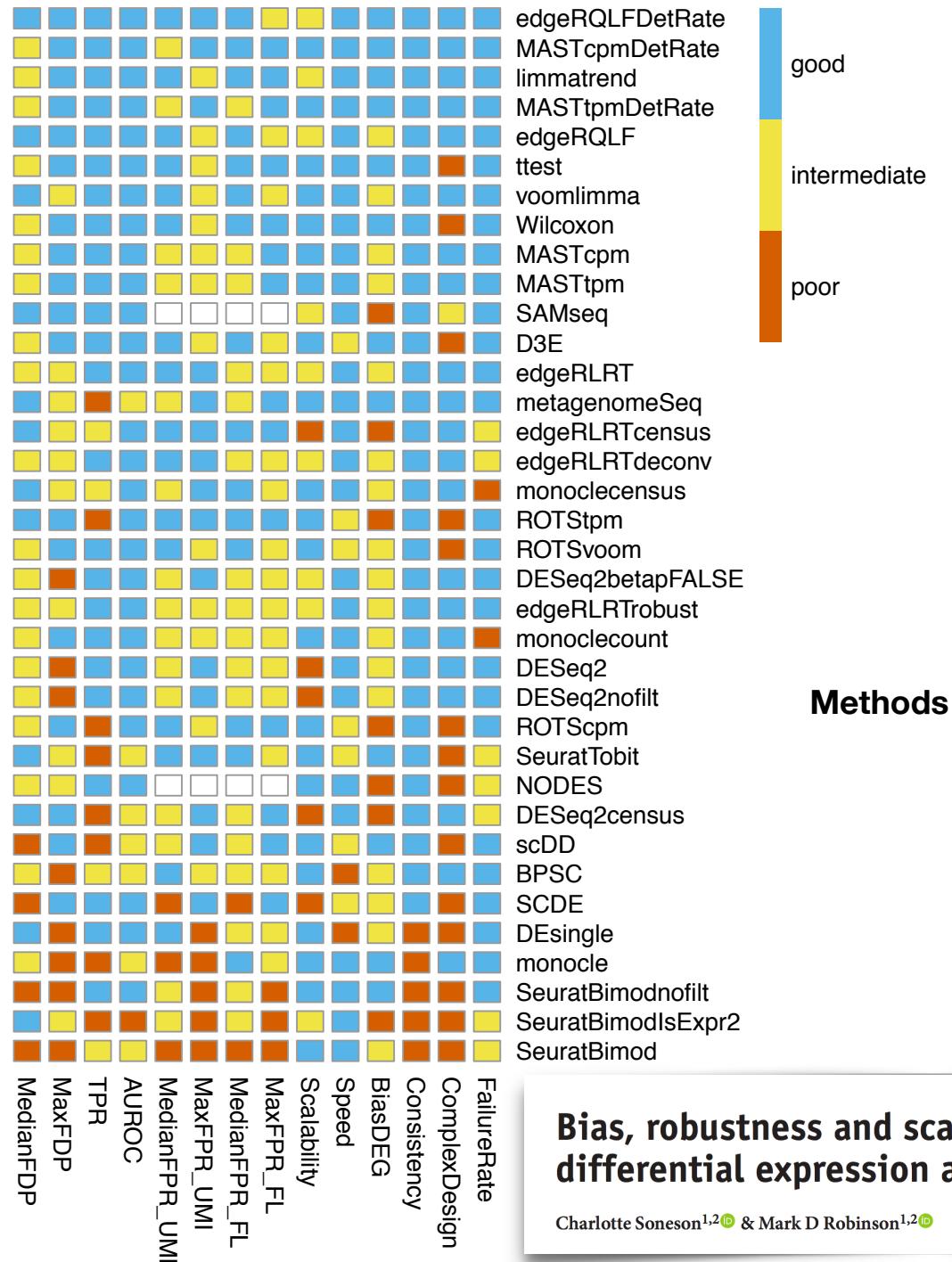
Punchline

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



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

Charlotte Soneson^{1,2} & Mark D Robinson^{1,2}



Differential states: A different *differential expression* problem (in single cell RNA-seq data)

Revealing the vectors of cellular identity with single-cell genomics

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

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

→ only changes during cell differentiation

Type: more permanent

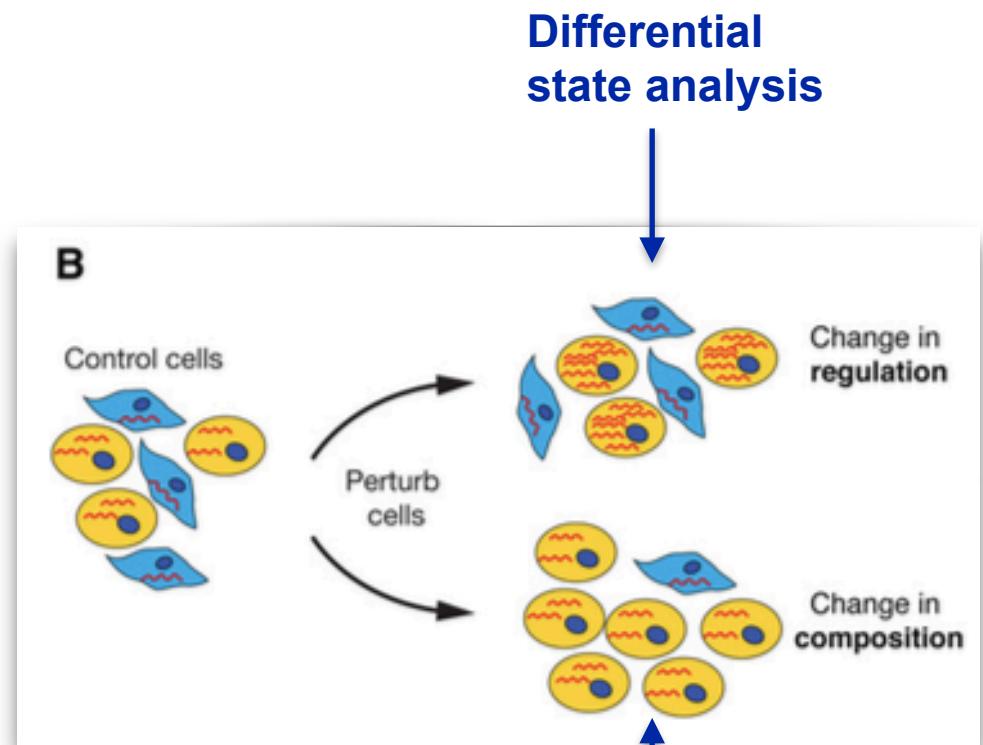
State: more transient

Perspective

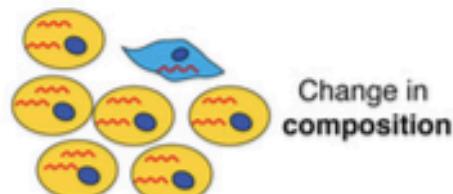
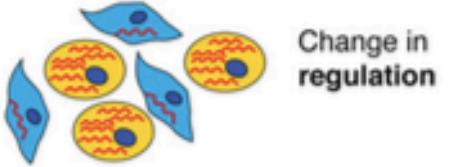
Defining cell types and states with single-cell genomics

Cole Trapnell

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



Differential state analysis



Differential abundance analysis

HYPOTHESIS

A periodic table of cell types

Bo Xia¹ and Itai Yanai^{1,2,*}

→ poses question if the cell state signal
can be detected compared to the
cell type signal

"We view a cell state as a secondary
module operating in addition to the
general cell type regulatory program."

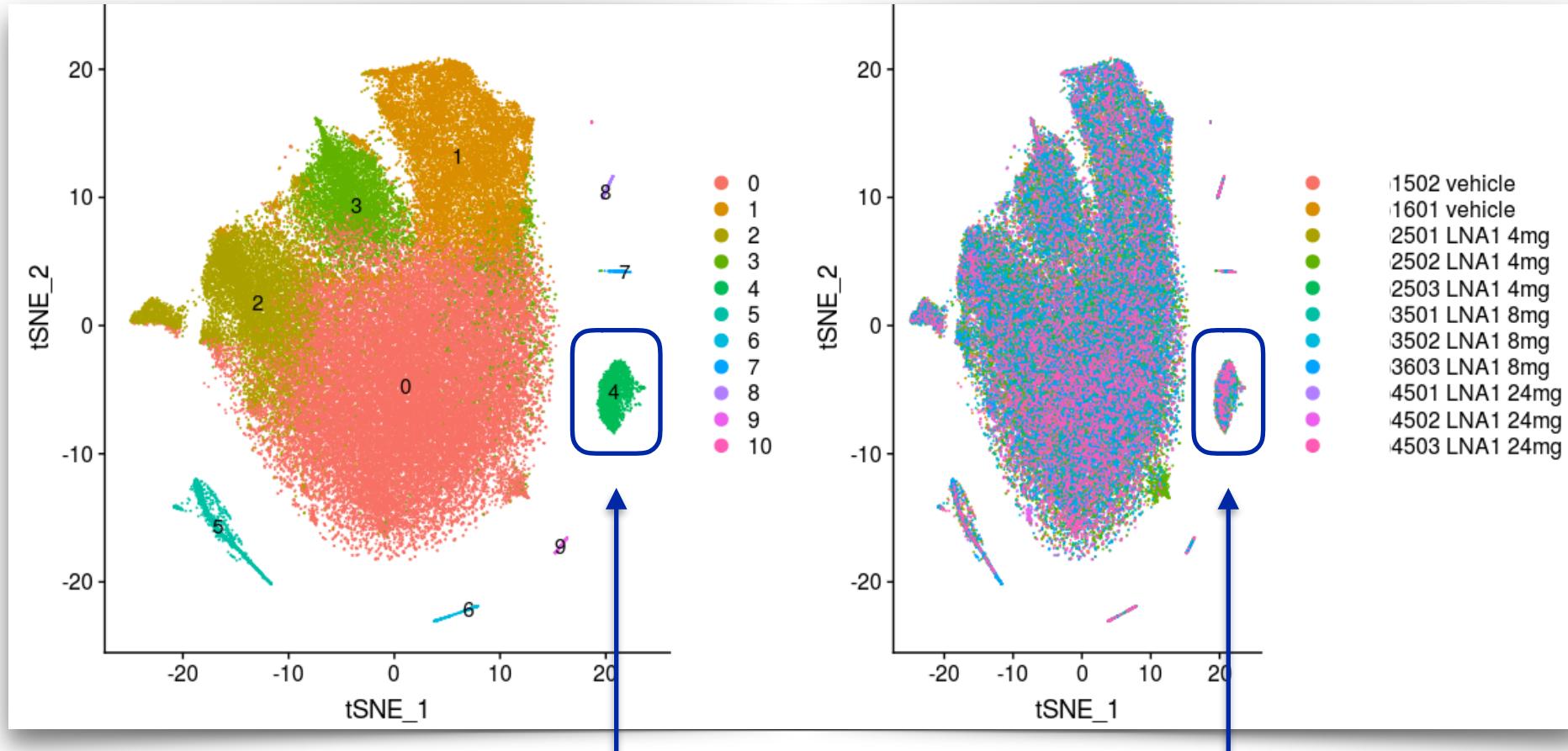
SPOTLIGHT

The evolving concept of cell identity in the single cell era

Samantha A. Morris^{1,2,3,*}

"how can we be confident that a novel transcriptional
signature represents a new cell type rather than a known cell
type in an unrecognized state?

Two types of differential expression: marker gene DE, differential state analysis



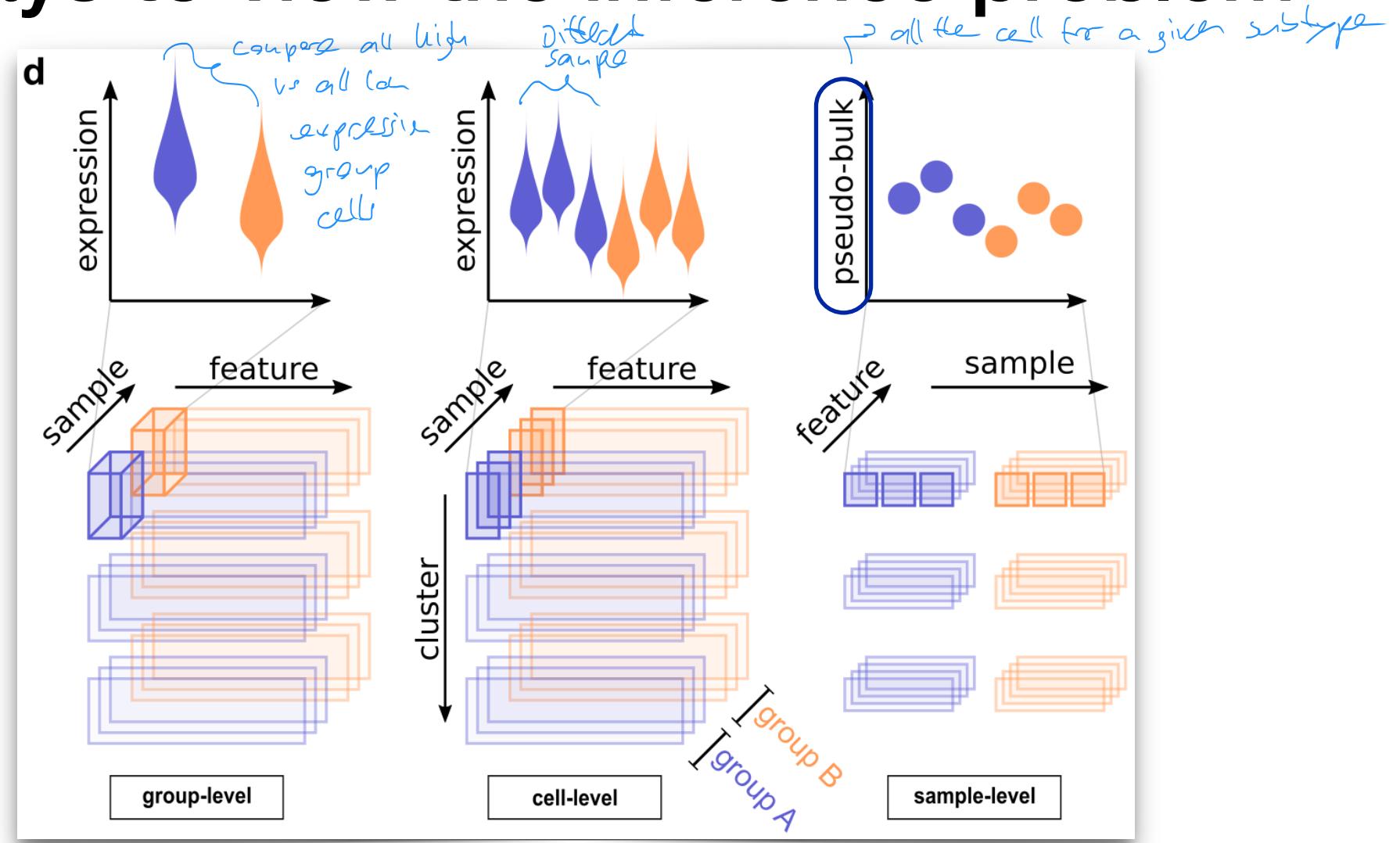
repeat for each population ..

Focus: Marker gene DE

Focus: cross-sample DE

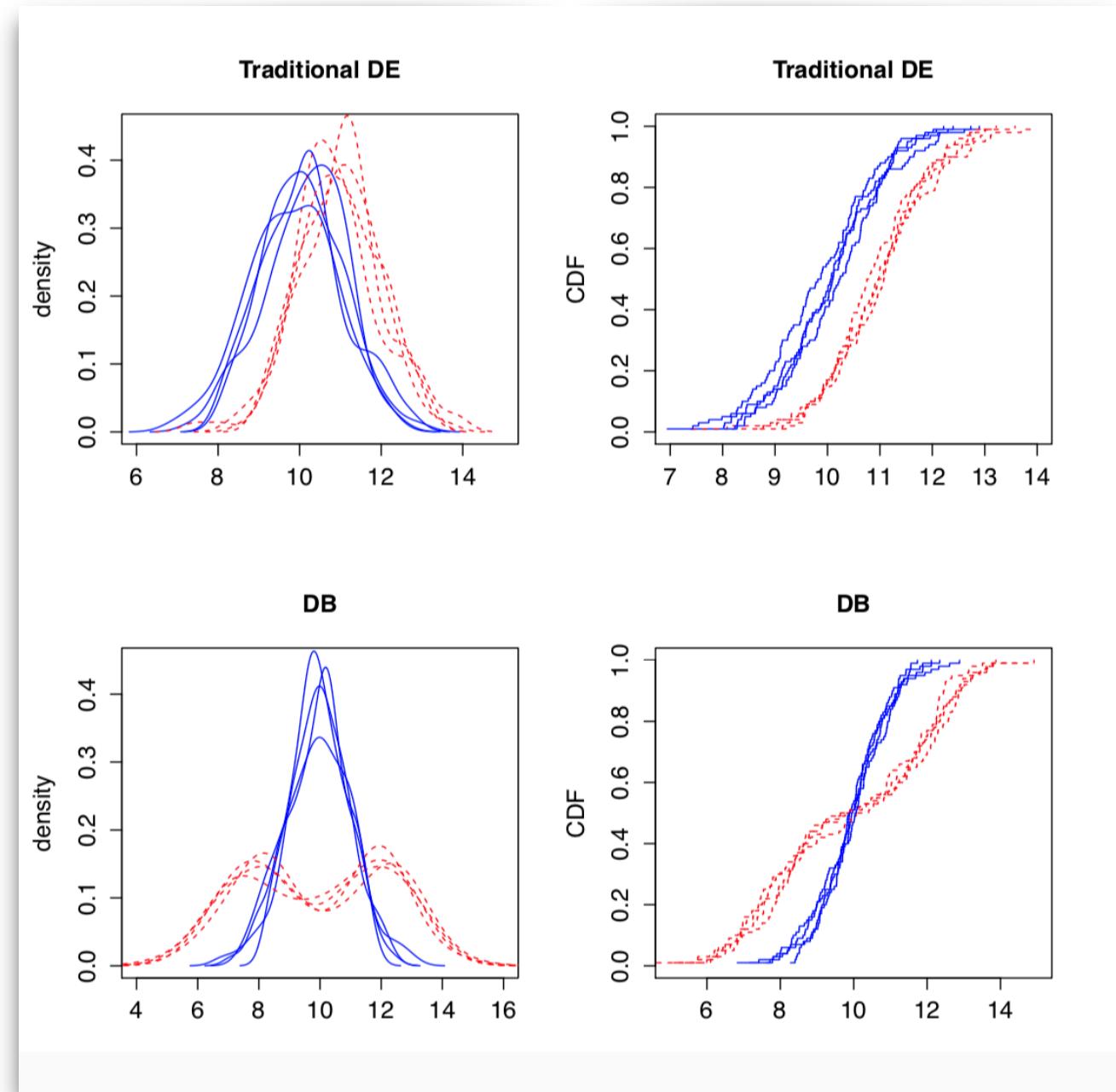
After “Cell Type Prediction” / “Clustering”, various ways to view the inference problem

Multi-sample
Multi-condition
Multi-population



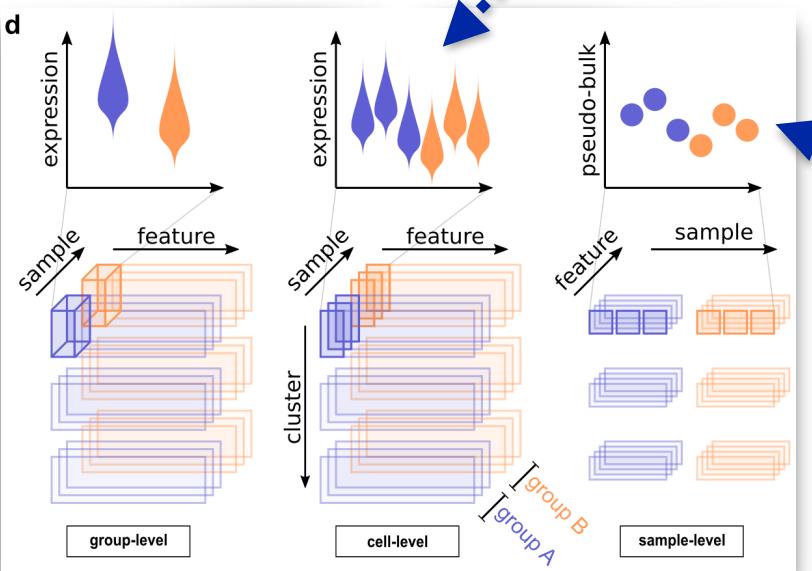
Limited “off-the-shelf” options for comparison of distributions (with replication)

- What is the null distribution? —> all distributions are the same.
- k-sample Anderson-Darling test (Scholz and Stephens, 1987)
- functional data analysis?



Some precedent, but different contexts

Multi-sample
Multi-condition
Multi-population



Batch effects and the effective design of single-cell gene expression studies

Po-Yuan Tung^{1,*}, John D. Blischak^{1,2,*}, Chiaowen Joyce Hsiao^{1,*}, David A. Knowles^{3,4}, Jonathan E. Burnett¹, Jonathan K. Pritchard^{3,5,6} & Yoav Gilad^{1,7}

mixed models

Overcoming confounding plate effects in differential expression analyses of single-cell RNA-seq data

AARON T. L. LUN*

Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK
aaron.lun@cruk.cam.ac.uk

JOHN C. MARONI

Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK
EMBL European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK and Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK
maroni@ebi.ac.uk

“A solution is proposed whereby counts are summed from all cells in each plate and the count sums for all plates are used in the DE analysis.”

Simulation: multi-sample, multi-subpopulation, multi-condition

Equivalent Expression

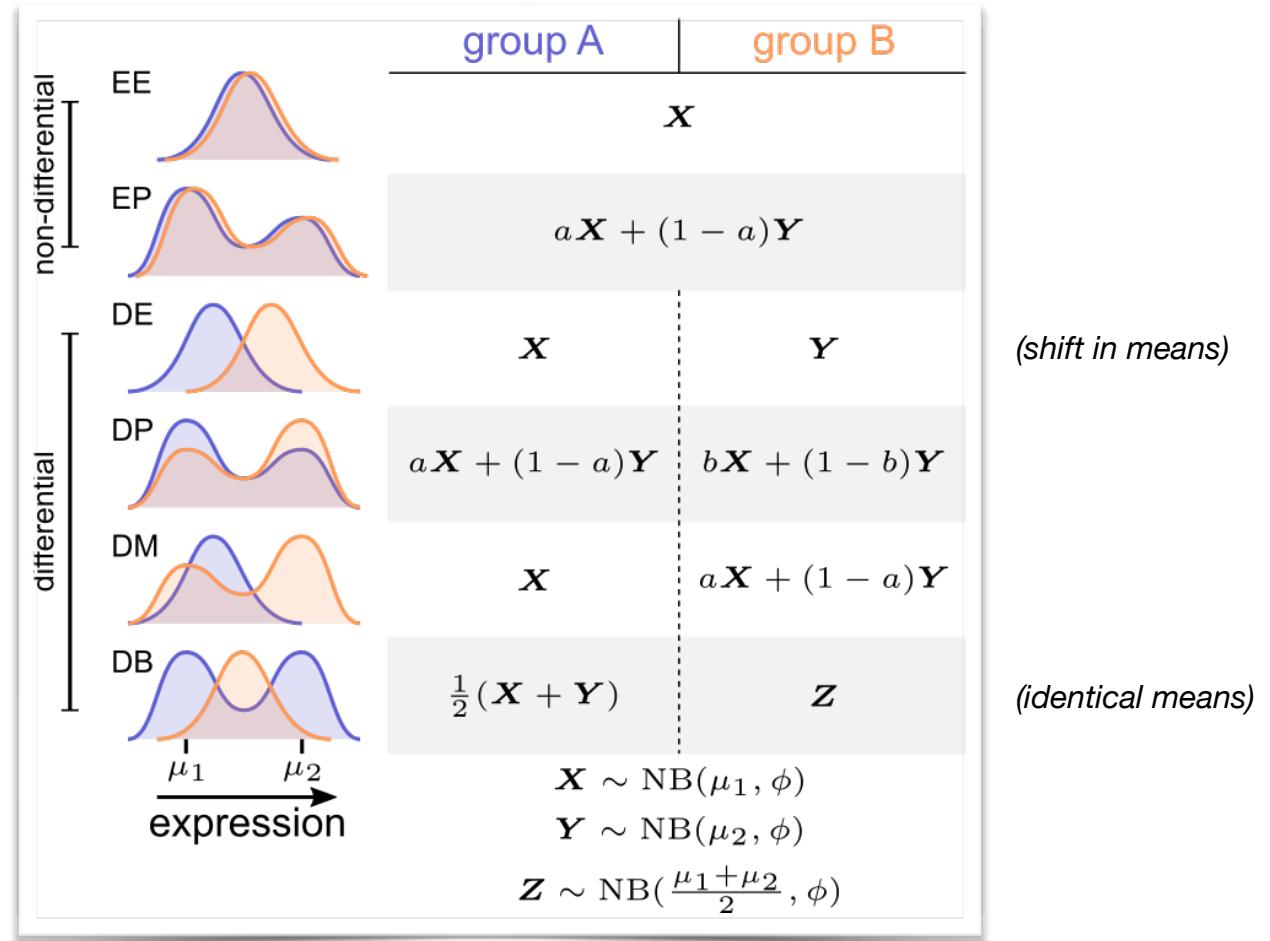
Equivalent Proportions

Differential Expression

Differential Proportions

Differential Modality

Both, Differential modality & component means



...idea adapted from Korthauer et al., 2016

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

Genome Biology

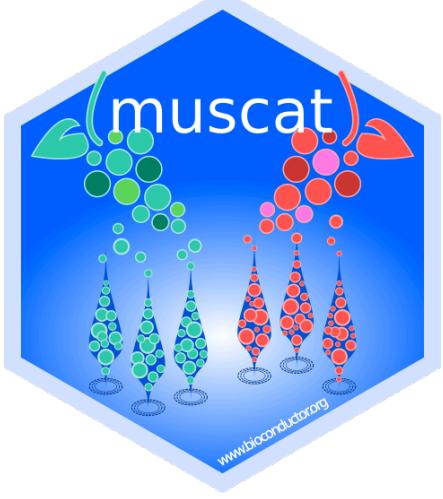
METHOD

Open Access



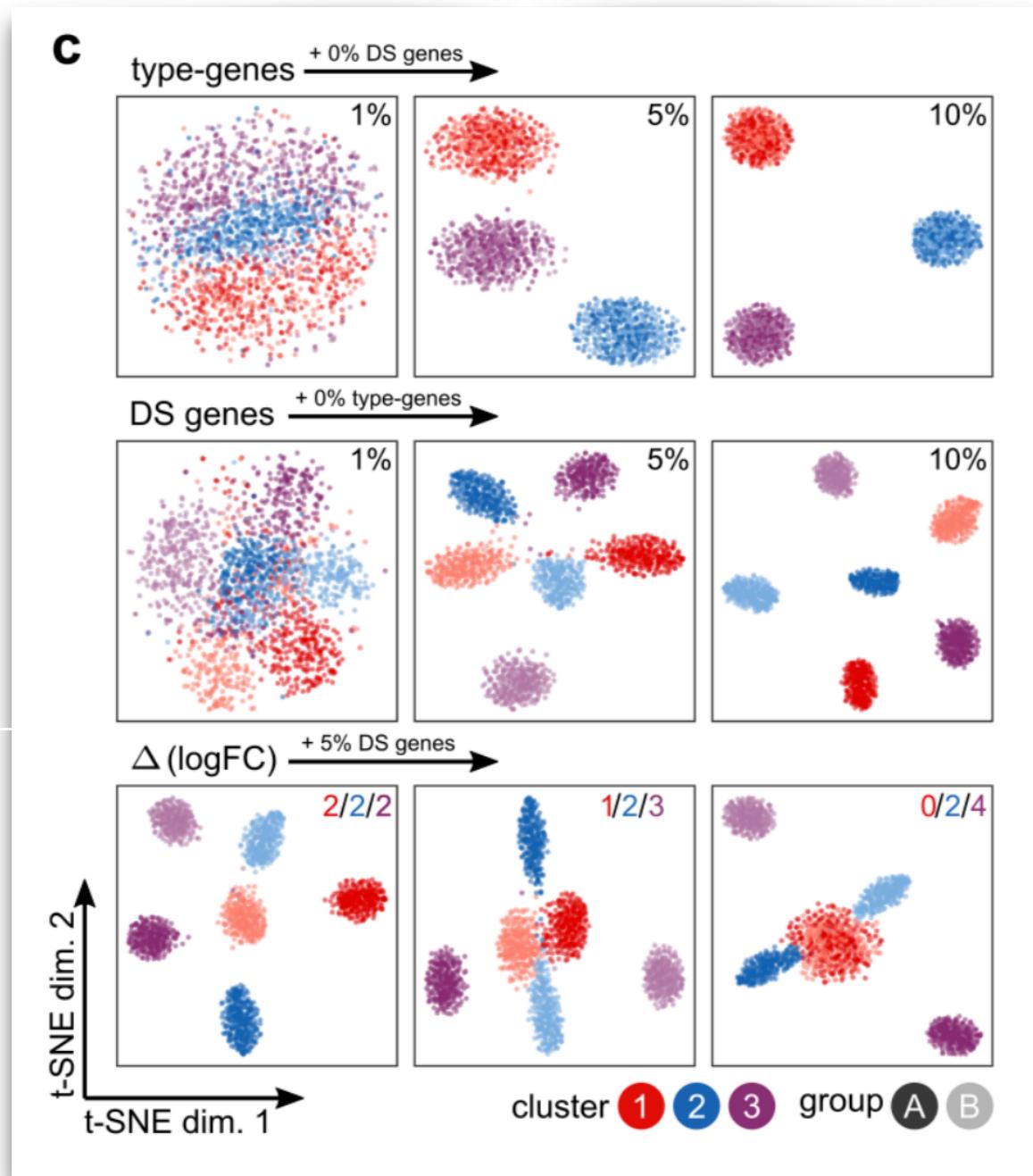
A statistical approach for identifying differential distributions in single-cell RNA-seq experiments

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

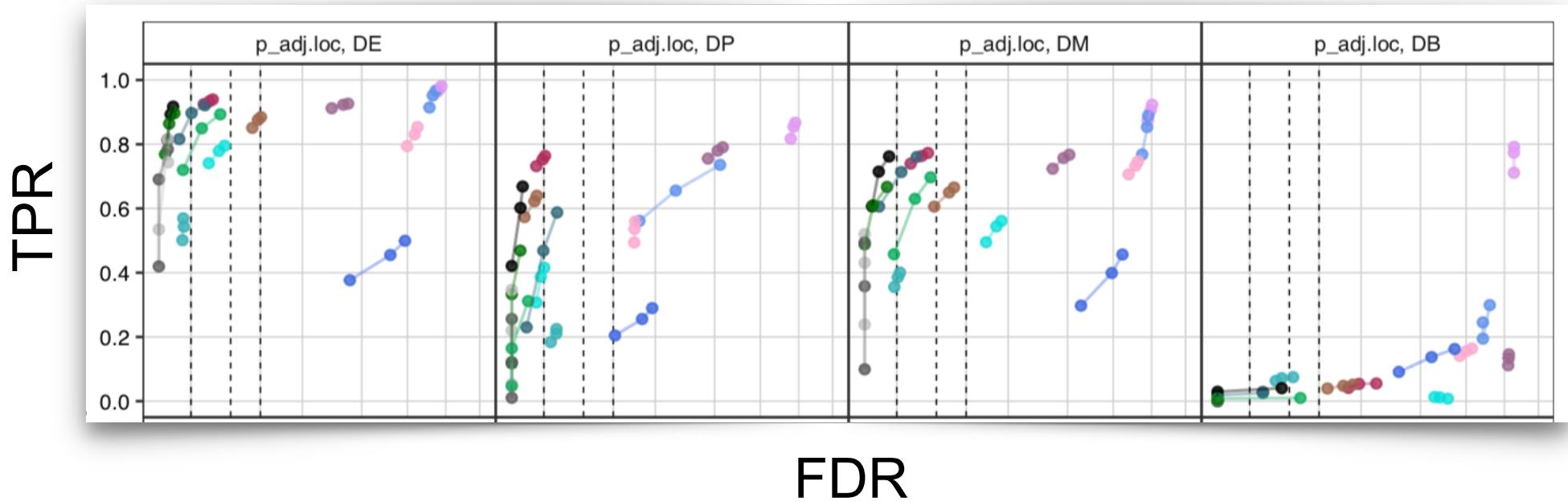
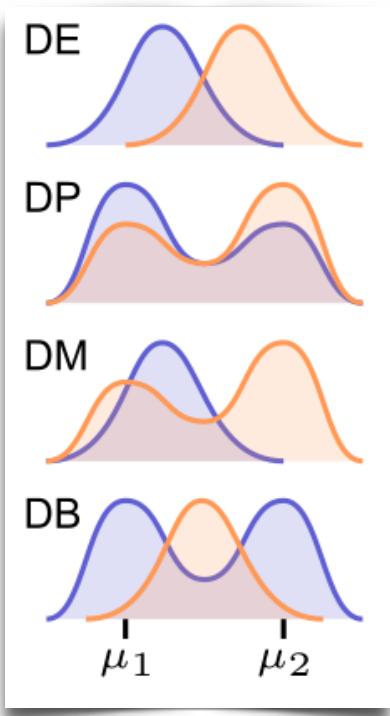


A flexible multi-sample simulation

- knobs for: sample size, # of cells, changes in abundance, subpopulation-specific state changes
- batch effects?



Differential state: Aggregation works well, mixed models work well. DB especially difficult to detect



AD = Anderson-Darling

MM = mixed models

- edgeR.sum(counts)
- edgeR.sum(scalecpm)
- limma-voom.sum(counts)
- limma-trend.mean(logcounts)
- limma-trend.mean(vstresiduals)
- MM-dream
- MM-nbinom
- MM-vst
- scDD.logcounts
- scDD.vstresiduals
- MAST.logcounts
- AD-gid.logcounts
- AD-gid.vstresiduals
- AD-sid.logcounts
- AD-sid.vstresiduals



Other

Intriguing possibility: deconvolute bulk RNA-seq into cell type specific expression profiles

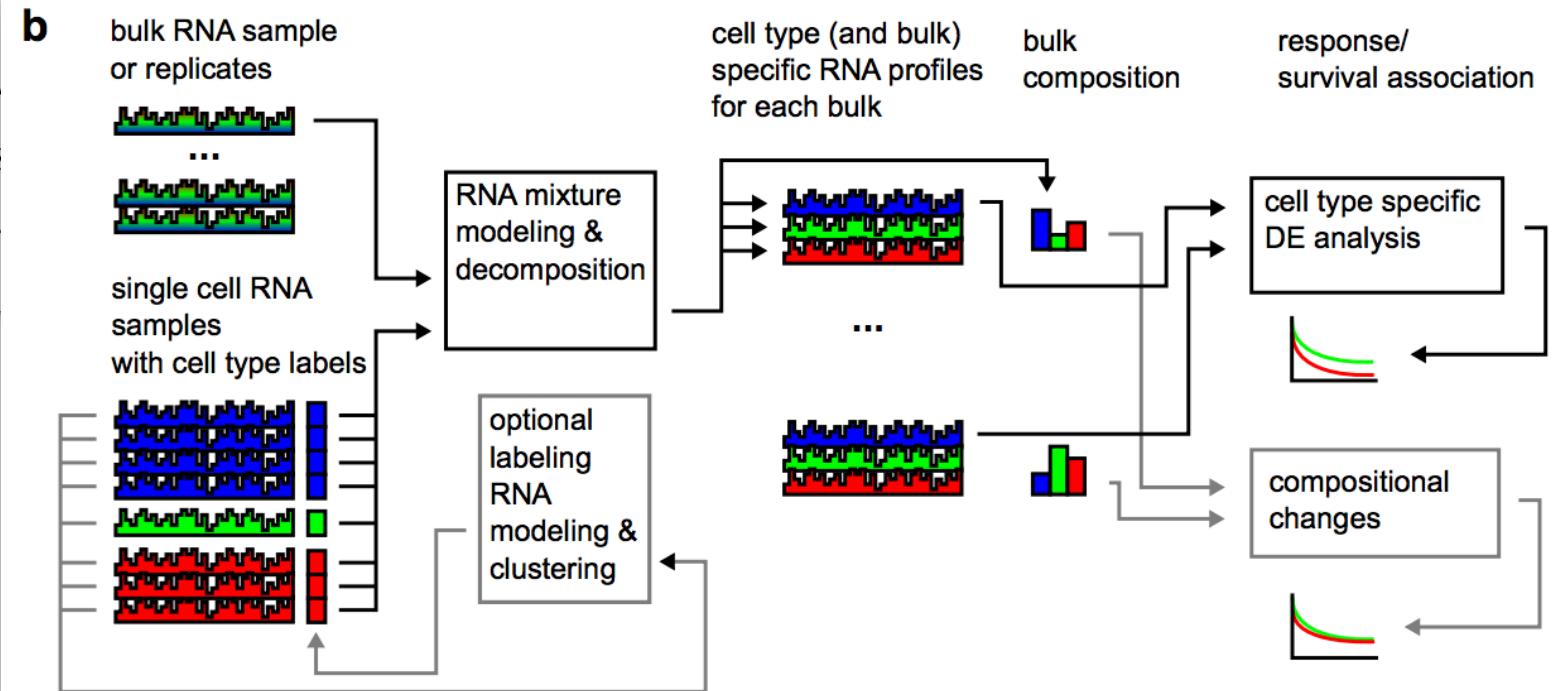
bioRxiv preprint first posted online Nov. 25, 2019; doi: <http://dx.doi.org/10.1101/854505>. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

PRISM: Recovering cell type specific expression profiles from composite RNA-seq data

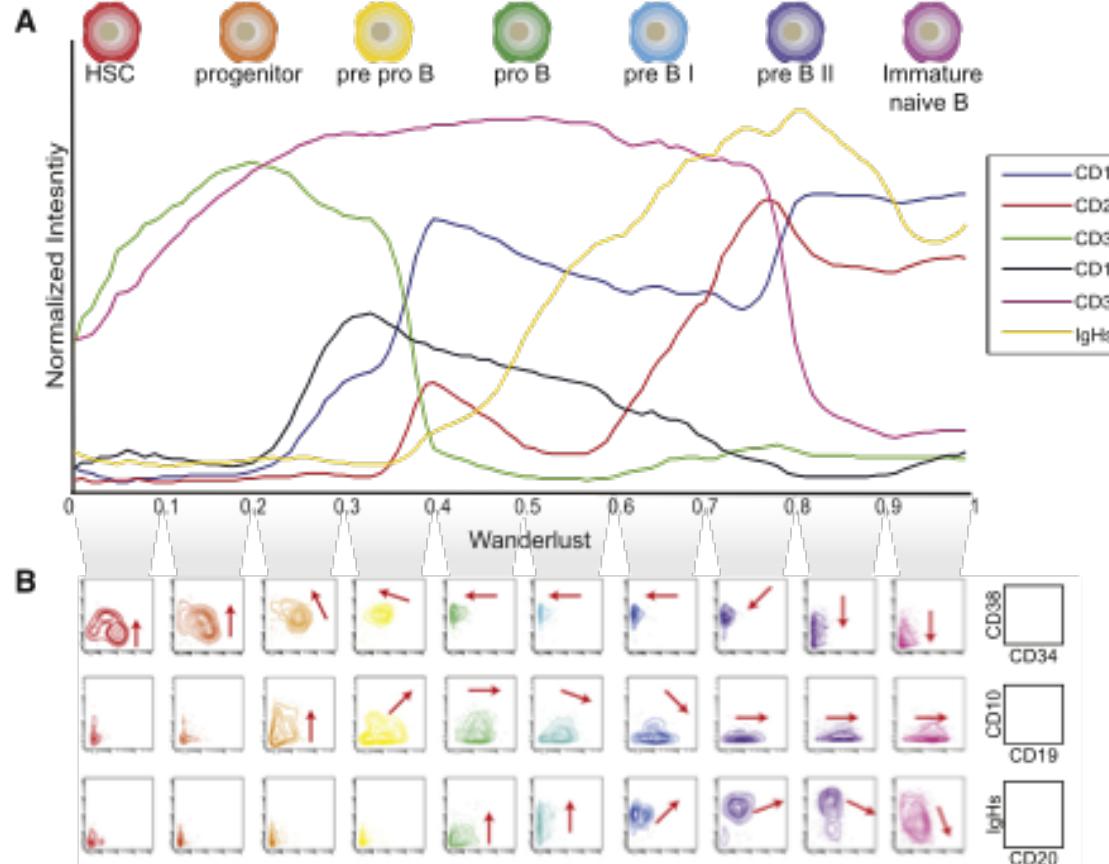
Antti Häkkinen¹, Kaiyang Zhang¹, Amjad Alkodsi^{1,2}, Noora A

Jun Dai¹, Katja Kaipio⁴, Tarja Lamminen⁴, Naziha Mansuri⁴, R

Olli Carpén^{1,3,4}, Johanna Hynninen⁵, Sakari Hietanen⁵, Rainer



Trajectory analysis



Amir et al., NBT, 2013

Mark D. Robinson, IMLS, UZH

Cell development

Revealing the vectors of cellular identity with single-cell genomics

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

A comparison of single-cell trajectory inference methods

Wouter Saelens^{1,2,6}, Robrecht Cannoodt^{1,3,4,6}, Helena Todorov^{1,2,5} and Yvan Saeyns^{1,2,*}

Trajectory inference approaches analyze genome-wide omics data from thousands of single cells and computationally infer the order of these cells along developmental trajectories. Although more than 70 trajectory inference tools have already been developed, it is challenging to compare their performance because the input they require and output models they produce vary substantially. Here, we benchmark 45 of these methods on 110 real and 229 synthetic datasets for cellular ordering, topology, scalability and usability. Our results highlight the complementarity of existing tools, and that the choice of method should depend mostly on the dataset dimensions and trajectory topology. Based on these results, we develop a set of guidelines to help users select the best method for their dataset. Our freely available data and evaluation pipeline (<https://benchmark.dynverse.org>) will aid in the development of improved tools designed to analyze increasingly large and complex single-cell datasets.