VIB, Ghent, 24 October 2019

Universität Zürich[UZH]

URPP Evolution in Action

SIB Swiss Institute of Bioinformatics

On the analysis of multi-sample multi-condition replicated single-cell RNA-seq datasets

Mark D. Robinson
Statistical Bioinformatics Group, IMLS@UZH+SIB

@markrobinsonca
https://robinsonlabuzh.github.io/


Helena Crowell


Charlotte Soneson


Pierre-Luc Germain


Fiona Huang


Simone Tiberi

Overview

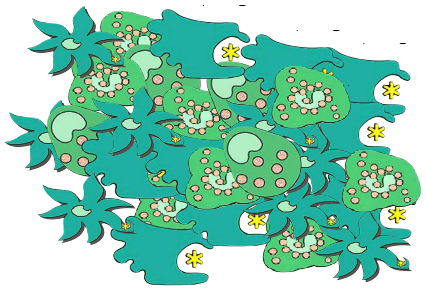
- Single cell RNA-seq data (scRNA-seq)
- A typical scRNA-seq pipeline
- Some comparisons of methods: clustering, “simple” differential expression
- Definitions: Cell type and cell state
- Multi-sample differential expression: **differential state analysis**
- Some things not in the preprint

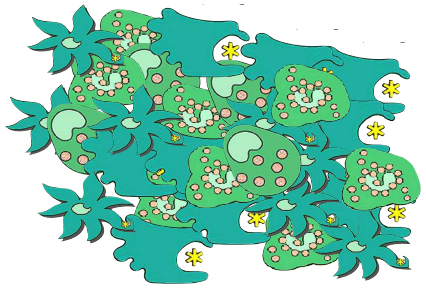
Bulk vs single-cell RNA-sequencing

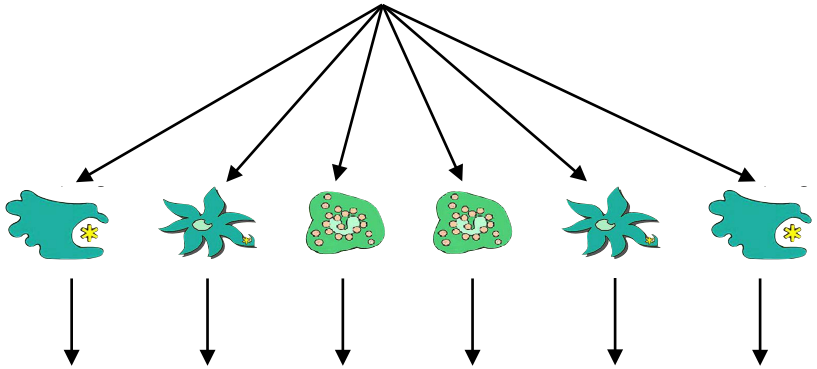
Cell sorting, tissue dissociation

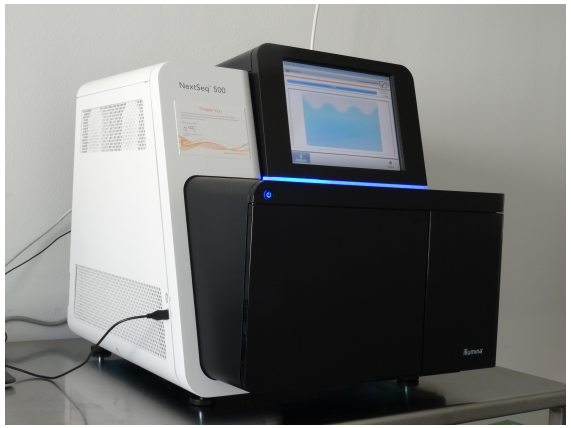
RNA extraction, preparation of cDNA, cell barcoding, UMIs (scRNA-seq only)

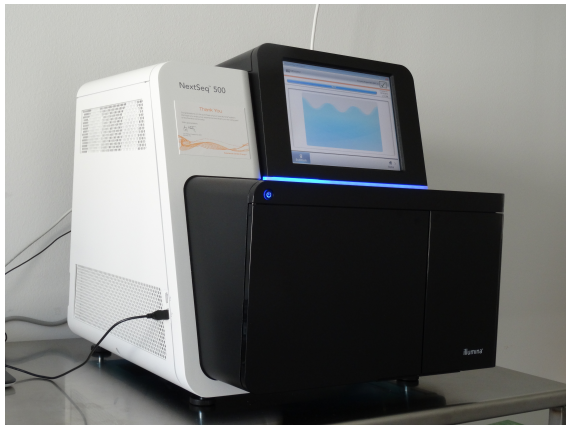
sequencing











Images modified from https://www.flickr.com/photos/konradfoerstner/21264667663 and https://commons.wikimedia.org/wiki/File:Innate_Immune_cells.svg

Introduction to Single-Cell RNA Sequencing

Thale Kristin Olsen[1] and Ninib Baryawno[1]

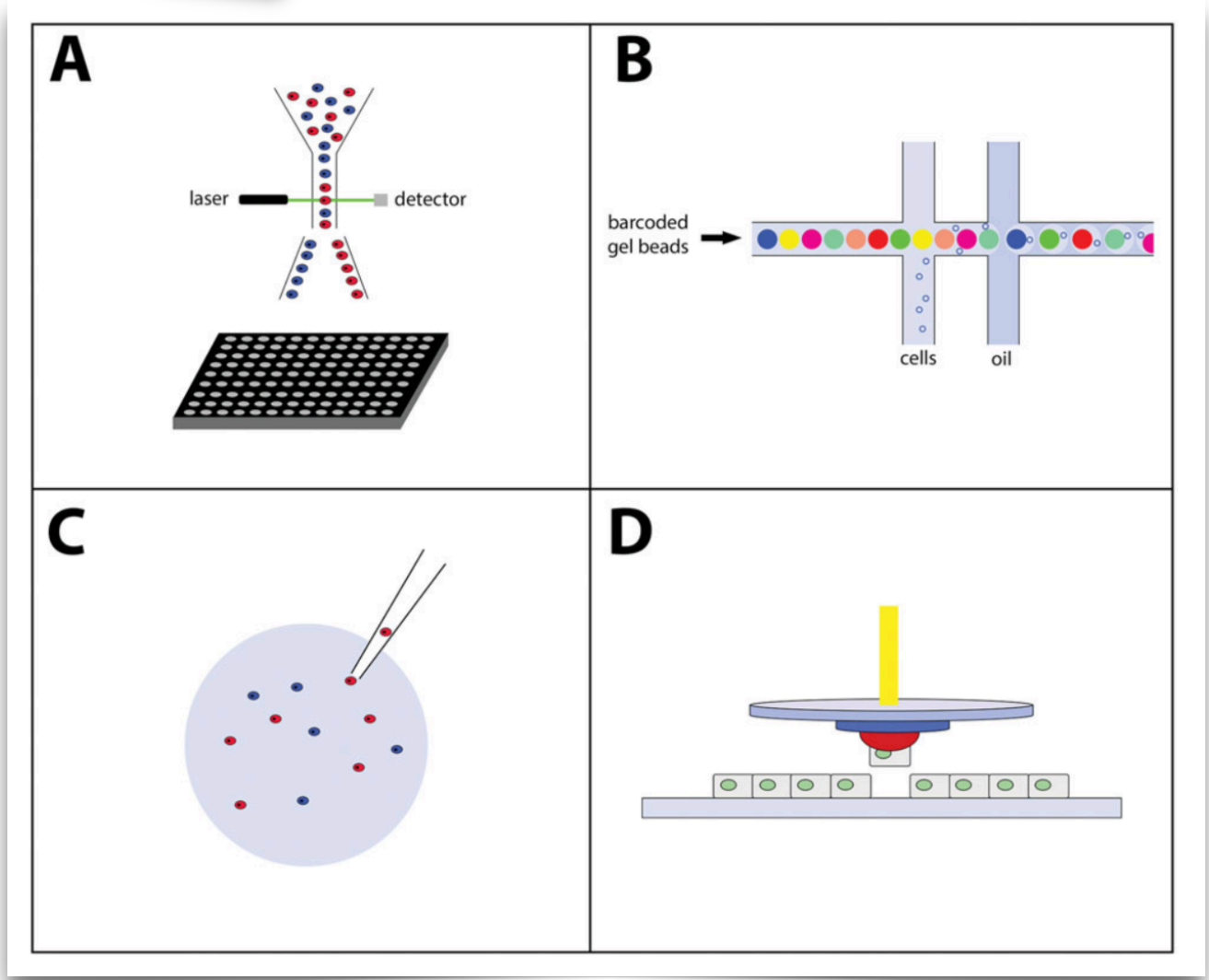
[1]Childhood Cancer Research Unit, Department of Women's and Children's Health, Karolinska Institutet, Stockholm, Sweden

Single cell isolation

FACS

Tissue dissociation

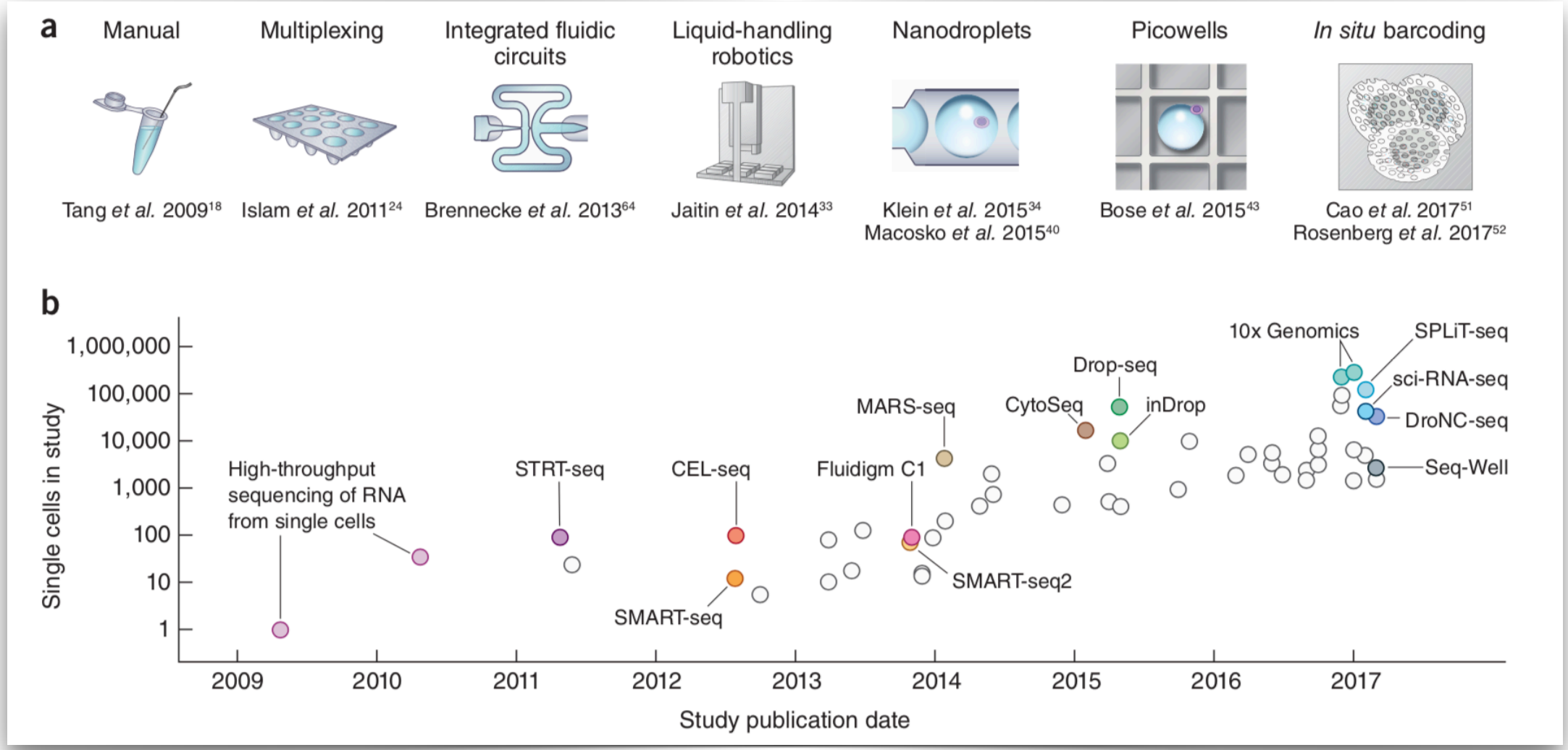
Micro-pipetting



droplet

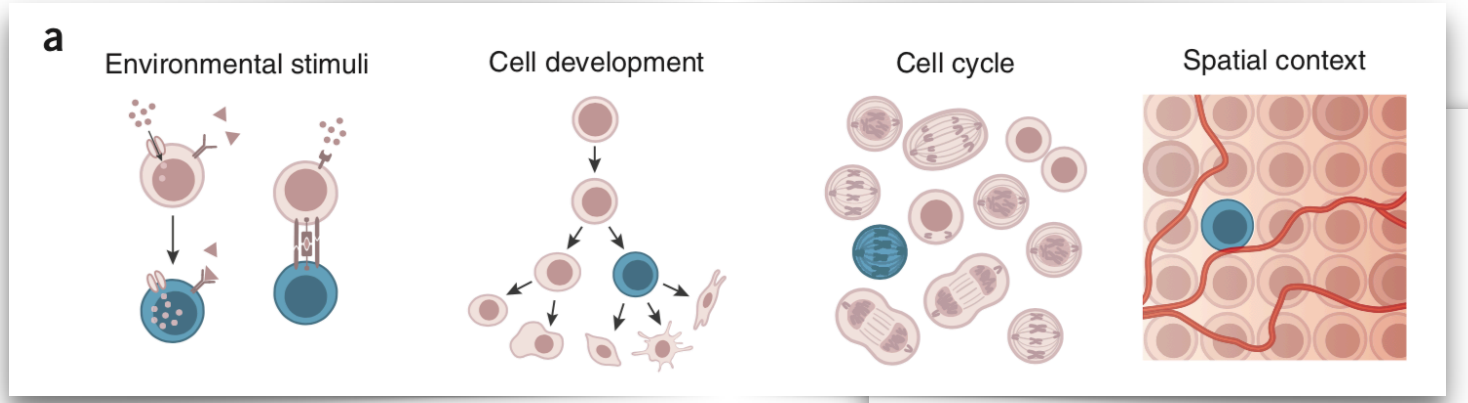
Laser Capture Microdissection

Platforms

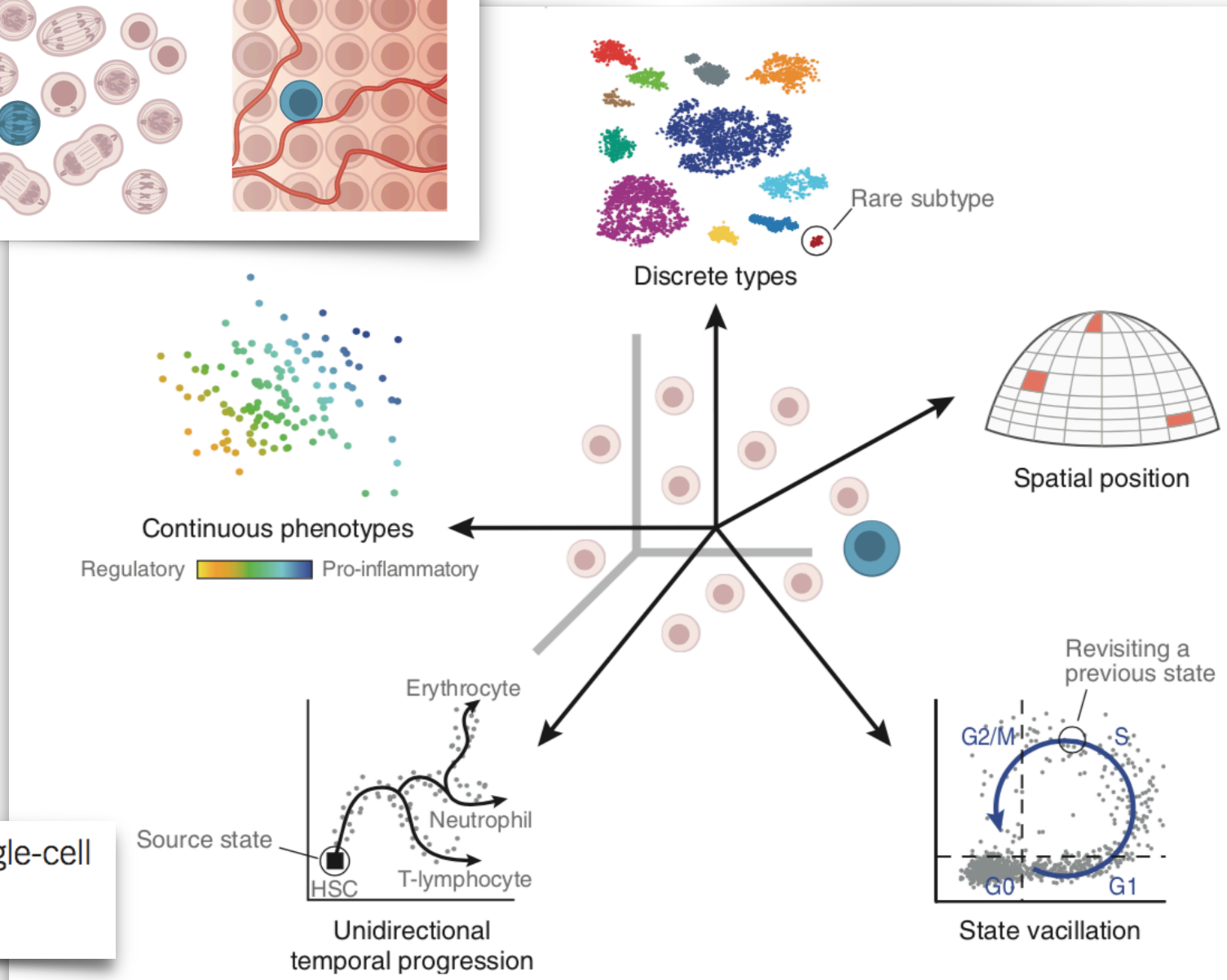


Note: tradeoff between number of cells and **depth per cell**

Svensson et al. 2018



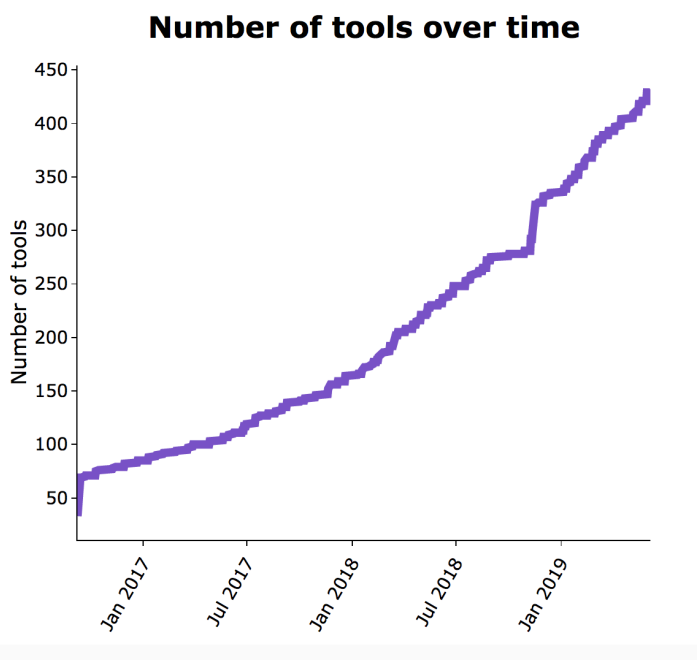
Applications



Revealing the vectors of cellular identity with single-cell genomics

Allon Wagner[1], Aviv Regev[2,3,5] & Nir Yosef[1,4,5]

Computational Tools

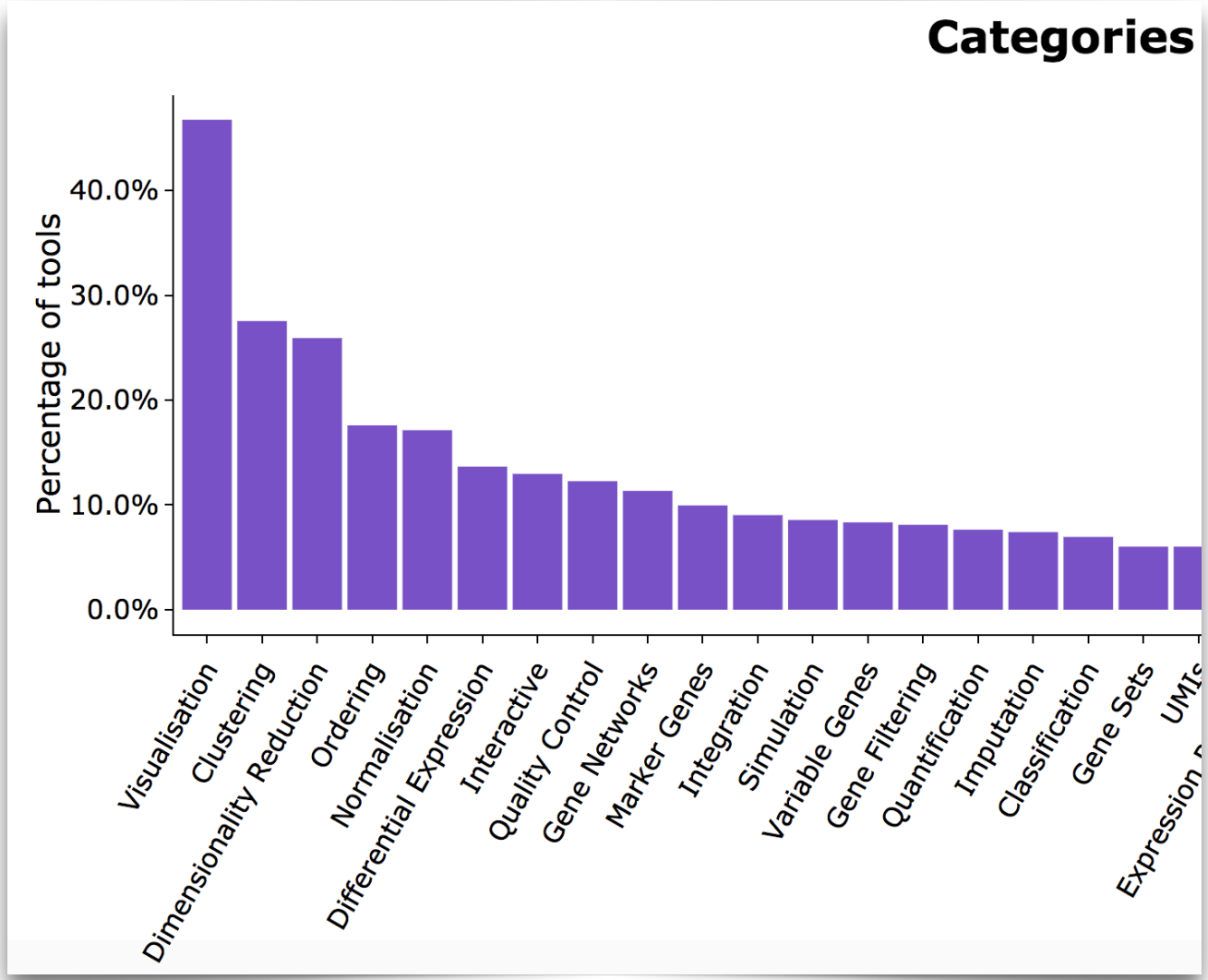


RESEARCH ARTICLE

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

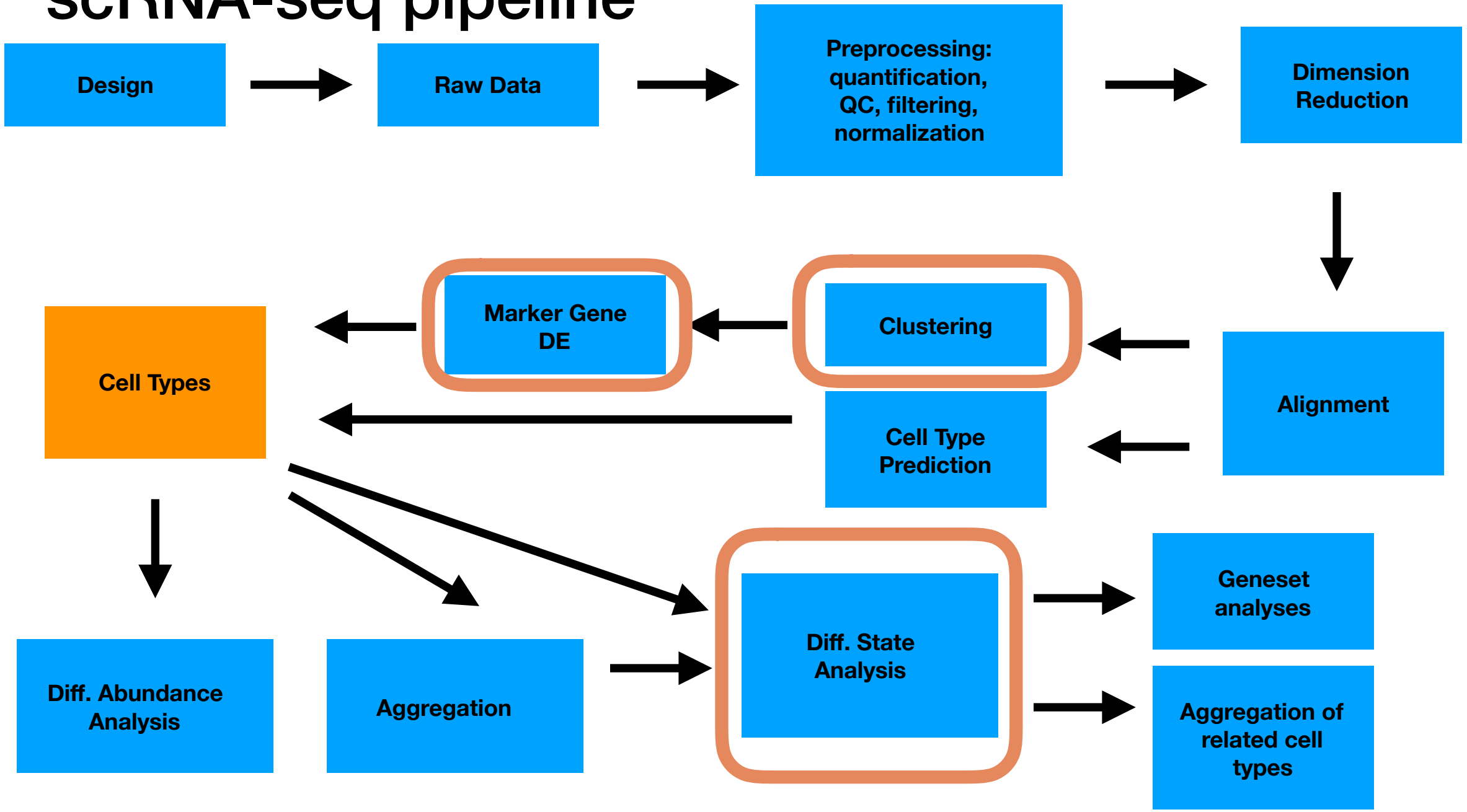
Luke Zappia[1,2], Belinda Phipson[1], Alicia Oshlack[1,2]*

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



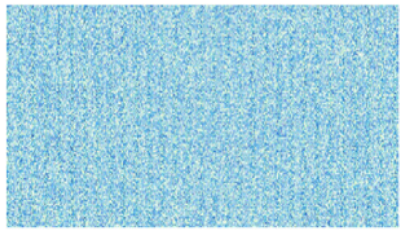
https://www.scrna-tools.org

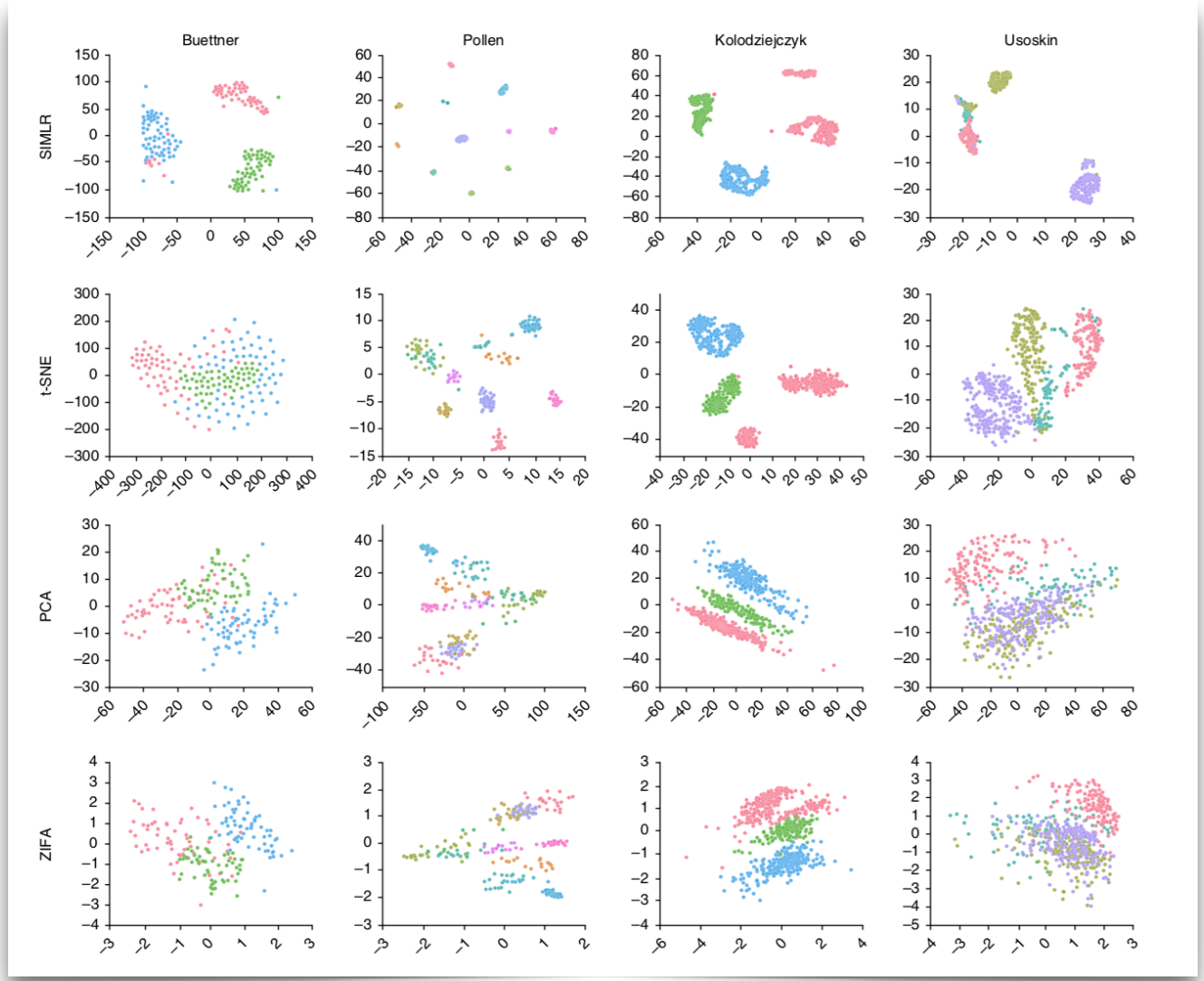
scRNA-seq pipeline



A couple early goals: find clusters, identify genes that distinguish clusters ("cell type") .. perhaps via *dimension reduction*

Gene expression matrix





Wang et al. 2017, SIMLR

Clustering methods

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 Son[1,2]

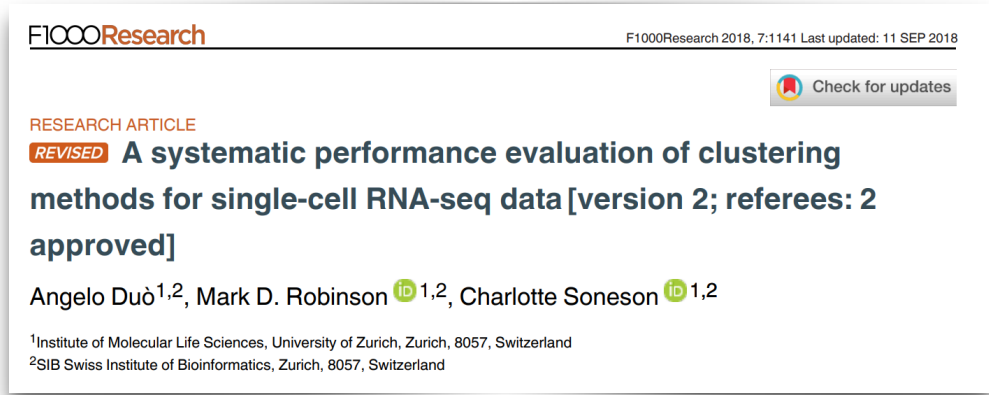
[1]Institute of Molecular Life Sciences, University of Zurich, Zurich, 8057, Switzerland
[2]SIB Swiss Institute of Bioinformatics, Zurich, 8057, Switzerland

Across various datasets (platforms), what methods work well?

Method	Description	Reference
`ascend` (v0.5.0)	PCA dimension reduction (dim=30) and iterative hierarchical clustering	36
`CIDR` (v0.1.5)	PCA dimension reduction based on zero-imputed similarities, followed by hierarchical clustering	37
`FlowSOM` (v1.12.0)	PCA dimension reduction (dim=30) followed by self-organizing maps (5x5, 8x8 or 15x15 grid, depending on the number of cells in the data set) and hierarchical consensus meta-clustering to merge clusters	38
`monocle` (v2.8.0)	t-SNE dimension reduction (initial PCA dim=50, t-SNE dim=3) followed by density-based clustering	25,39
`PCAHC`	PCA dimension reduction (dim=30) and hierarchical clustering with Ward.D2 linkage	33,40
`PCAKmeans`	PCA dimension reduction (dim=30) and K-means clustering with 25 random starts	33,41
`pcaReduce` (v1.0)	PCA dimension reduction (dim=30) and k-means clustering through an iterative process. Stepwise merging of clusters by joint probabilities and reducing the number of dimensions by PC with lowest variance. Repeated 100 times followed consensus clustering using the clue package	42
`RaceID2 (March 3, 2017 version)`	K-medoids clustering based on Pearson correlation dissimilarities	43
`RtsneKmeans`	t-SNE dimension reduction (initial PCA dim=50, t-SNE dim=3, perplexity=30) and K-means clustering with 25 random starts	34,41,44
`SAFE` (v2.1.0)	Ensemble clustering using SC3, CIDR, Seurat and t-SNE + Kmeans	45
`SC3` (v1.8.0)	PCA dimension reduction or Laplacian graph. K-means clustering on different dimensions. Hierarchical clustering on consensus matrix obtained by K-means	46
`SC3svm` (v1.8.0)	Using SC3 to derive the clusters for half of the cells, then using a support vector machine (SVM) to classify the rest	46,47
`Seurat` (v2.3.1)	Dimension reduction by PCA (dim=30) followed by nearest neighbor graph clustering	17
`TSCAN` (v1.18.0)	PCA dimension reduction followed by model-based clustering	48

Reference datasets

Data set	Sequencing protocol	# cells	# features	Median total counts per cell	Median # features per cell	# subpopulations	Description	Ref.
Koh	SMARTer	531	48,981	1,390,268	14,277	9	FACS purified H7 human embryonic stem cells in different differention stages	24
KohTCC	SMARTer	531	811,938	1,391,012	66,086	9	FACS purified H7 human embryonic stem cells in different differention stages	24
Kumar	SMARTer	246	45,159	1,687,810	26,146	3	Mouse embryonic stem cells, cultured with different inhibition factors	23
KumarTCC	SMARTer	263	803,405	717,438	63,566	3	Mouse embryonic stem cells, cultured with different inhibition factors	23
SimKumar4easy	-	500	43,606	1,769,155	29,979	4	Simulation using different proportions of differentially expressed genes	29
SimKumar4hard	-	499	43,638	1,766,843	30,094	4	Simulation using different proportions of differentially expressed genes	29
SimKumar8hard	-	499	43,601	1,769,174	30,068	8	Simulation using different proportions of differentially expressed genes	29
TrapnellI	SMARTer	222	41,111	1,925,259	13,809	3	Human skeletal muscle myoblast cells, differention induced by low-serum medium	25
TrapnellITCC	SMARTer	227	684,953	1,819,294	66,864	3	Human skeletal muscle myoblast cells, differention induced by low-serum medium	25
Zhengmix4eq	10xGenomics GemCode	3,994	15,568	1,215	487	4	Mixtures of FACS purified peripheral blood mononuclear cells	5
Zhengmix4uneq	10xGenomics GemCode	6,498	16,443	1,145	485	4	Mixtures of FACS purified peripheral blood mononuclear cells	5
Zhengmix8eq	10xGenomics GemCode	3,994	15,716	1,298	523	8	Mixtures of FACS purified peripheral blood mononuclear cells	5



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

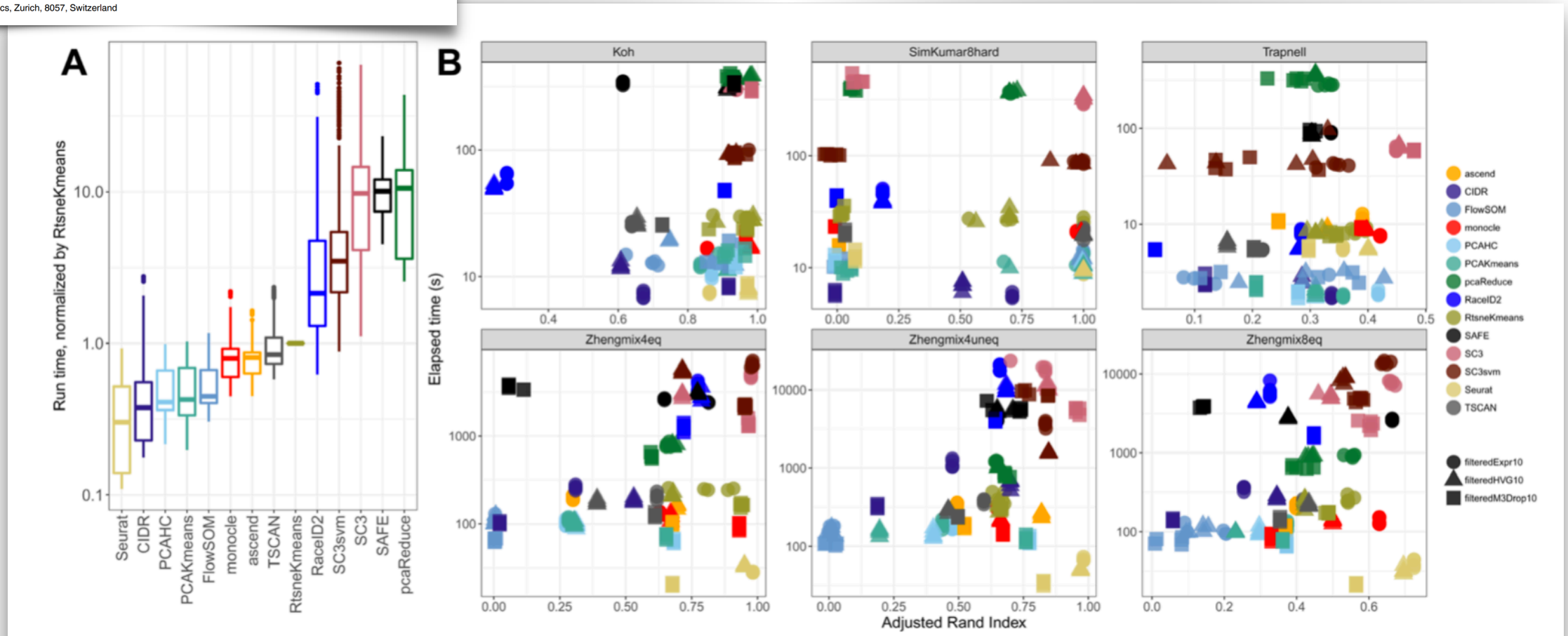
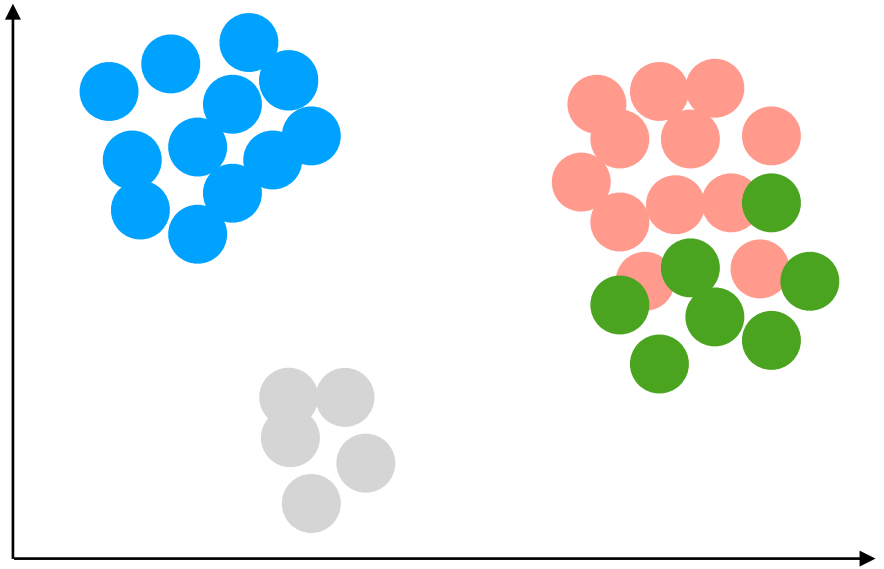


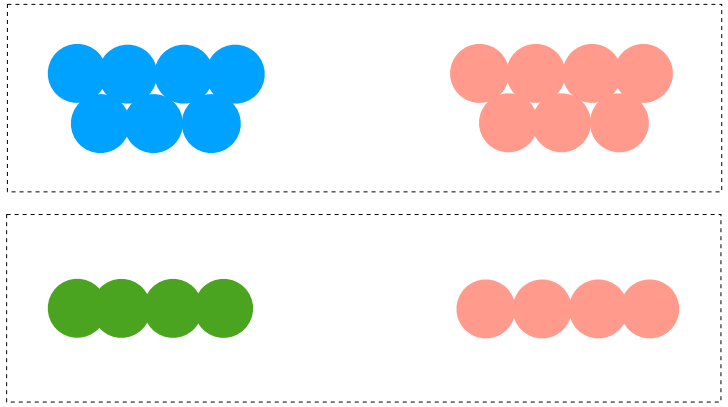
Figure 2. (**A**) Normalized run times, using `RtsneKmeans` as the reference method, across all data set instances and number of clusters. (**B**) Run time versus performance (ARI) for a subset of data sets and filterings, for the true number of clusters.

Similarly, how well do differential expression methods work?

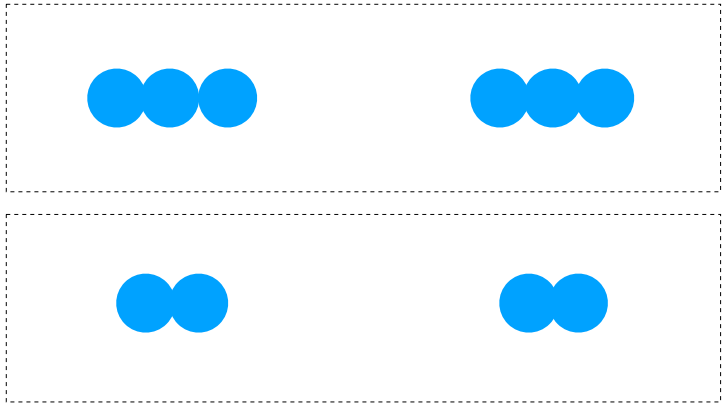
Using labels from independent truth (time point, fluoresence)



“Signal” data sets



“Mock” or null data sets



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

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

RECEIVED 6 JUNE 2017; ACCEPTED 16 JANUARY 2018; PUBLISHED ONLINE 26 FEBRUARY 2018;

conquer - reprocessed data analysis read public scRNA-seq datasets

- http://imlspenticton.uzh.ch:3838/conquer/
- Currently 40 data sets, both full-length and UMI-based protocols

conquer

About | scRNA-seq data sets | Changelog | Excluded samples | Tutorial

Show 10 entries Search:

	Data set	ID	organism	ncells	MultiAssayExperiment	MultiQC report	scater report	salmon archive
1	EMTAB2805 (PMID 25599176)	Buettner2015	Mus musculus	288	Download .rds (2017-07-23)	Download .html (2017-07-23)	Download .html (2017-07-23)	Download .tar.gz (2016-07-23)
2	EMTAB3929 (PMID 27062923)	Petropoulos2016	Homo sapiens	1529	Download .rds (2017-07-22)	Download .html (2017-07-22)	Download .html (2017-07-22)	Download .tar.gz (2017-07-20)
3	GSE41265 (PMID 23685454)	Shalek2013	Mus musculus	18	Download .rds (2017-07-23)	Download .html (2017-07-23)	Download .html (2017-07-23)	Download .tar.gz (2016-09-09)
4	GSE44183-GPL11154 (PMID 23892778)	Xue2013	Homo sapiens	29	Download .rds (2017-07-23)	Download .html (2017-07-23)	Download .html (2017-07-23)	Download .tar.gz (2016-07-24)
5	GSE44183-GPL13112 (PMID 23892778)	Xue2013	Mus musculus	17	Download .rds (2017-07-23)	Download .html (2017-07-23)	Download .html (2017-07-23)	Download .tar.gz (2016-07-24)
6	GSE44183-GPL13112-trimmed (PMID 23892778)	Xue2013	Mus musculus	17	Download .rds (2017-07-23)	Download .html (2017-07-23)	Download .html (2017-07-23)	Download .tar.gz (2016-07-28)

Charlotte Soneson

scRNA-seq “simple” differential expression: Evaluated methods

- 36 approaches to differential gene expression (19 different packages/ tests, multiple settings)

BPSC
D3E
MAST
monocle
NODES
scDD
SCDE
Seurat
DEsingle

single-cell DE methods

DESeq2
edgeR-LRT
edgeR-QLF
limma-trend
voom-limma
SAMseq

“bulk” DE methods

metagenomeSeq
ROTS
t-test
Wilcoxon test

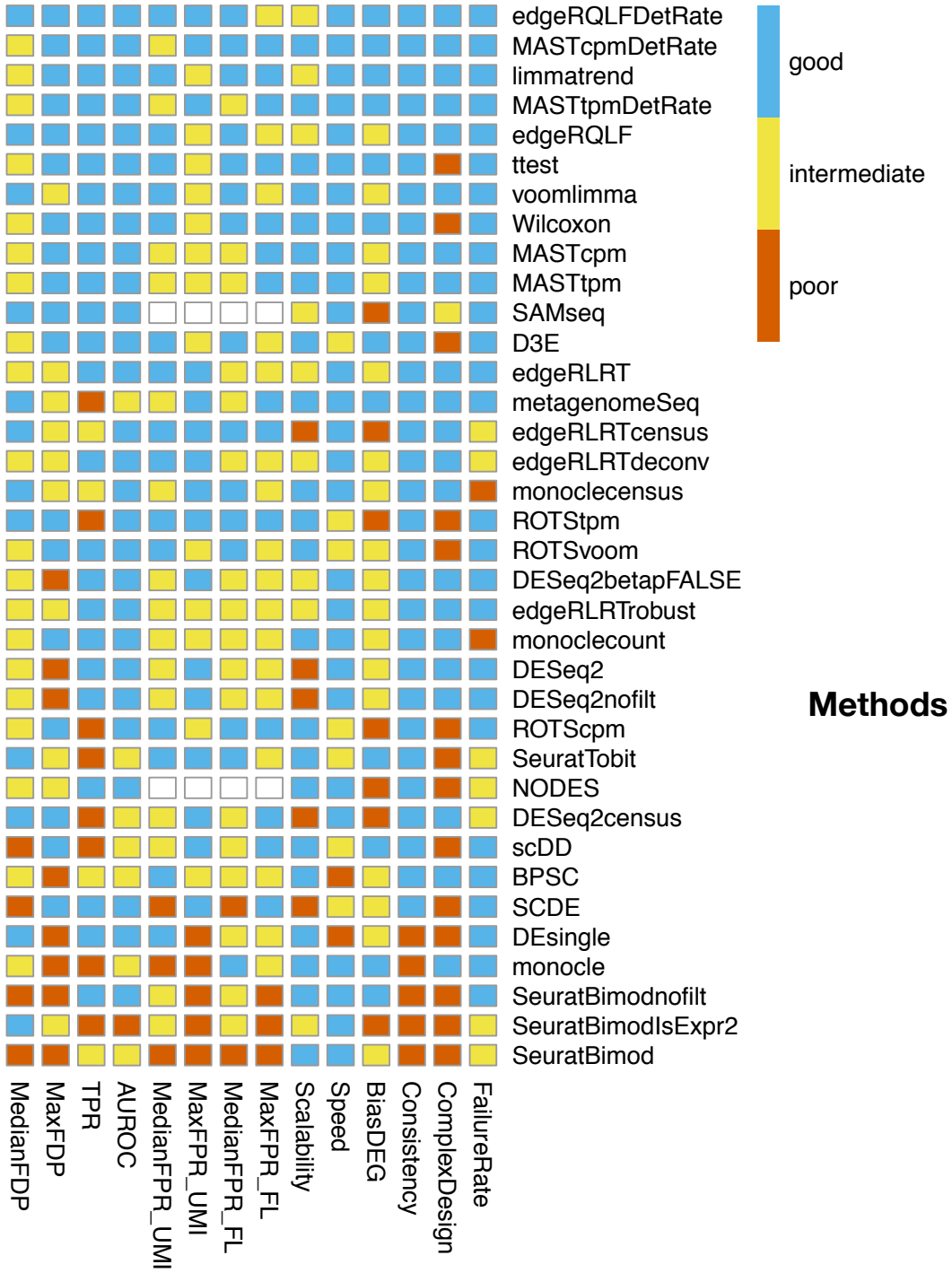
Code and data are available on GitHub:
https://github.com/csoneson/conquer_comparison

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”

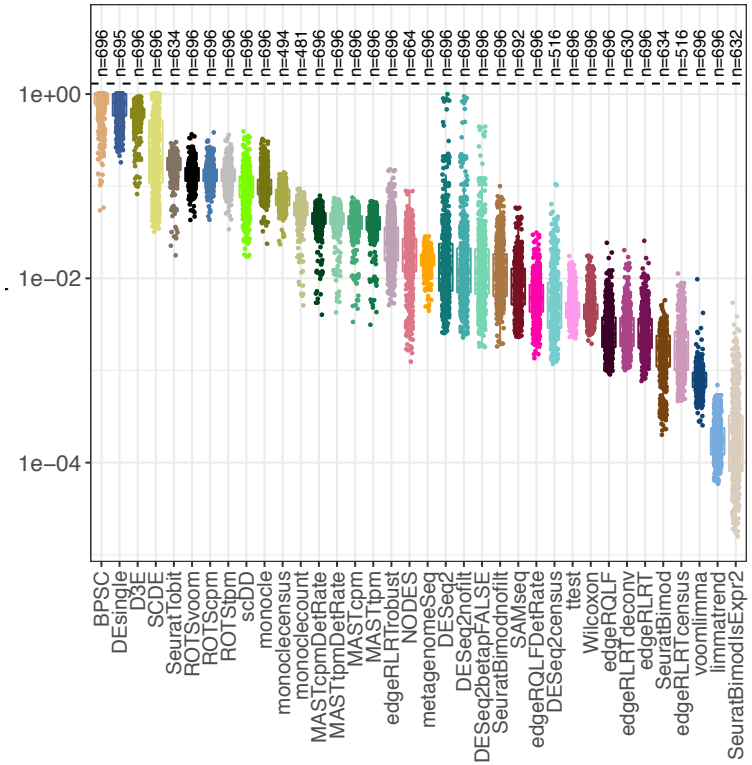


Beyond the punchline

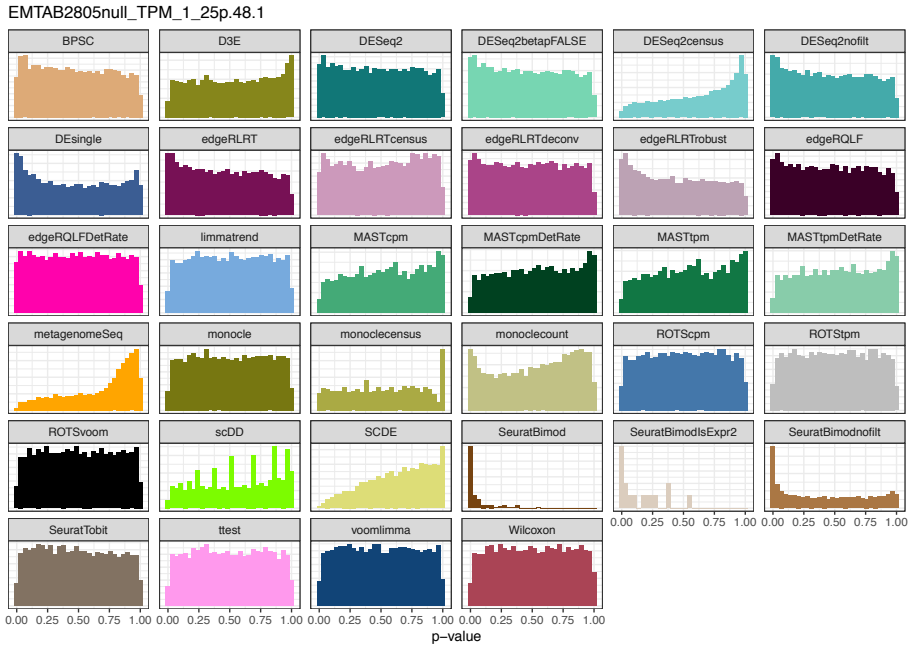
Many more details! >30 Supplementary Figures

http://imlspenticton.uzh.ch:3838/scrnaseq_de_evaluation/

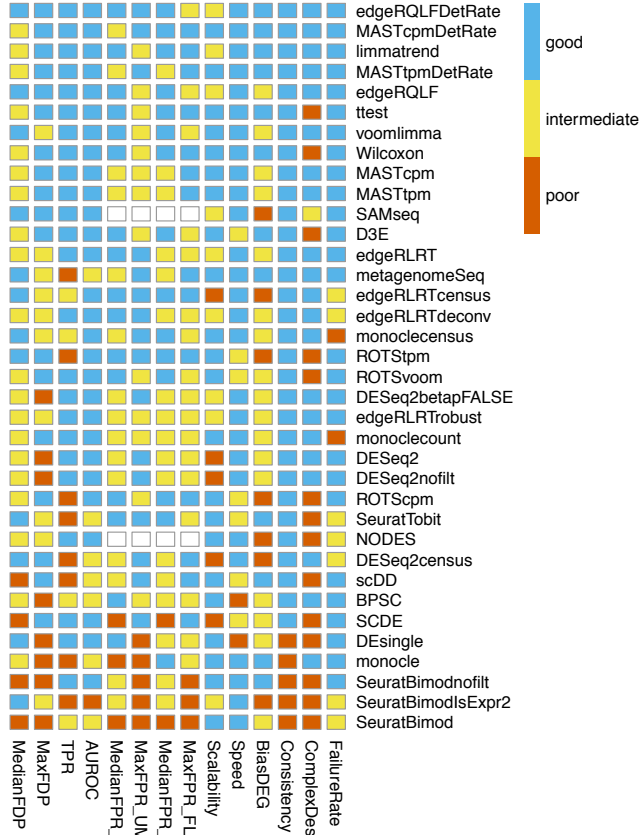
Relative run times



P-value distributions (under null)



Supplementary Figure 14: Representative p-value histograms for all methods returning nominal p-values applied to one of the scRNA-seq null dataset instances, after gene prefiltering retaining only genes with an estimated expression exceeding 1 TPM in more than 25% of the cells.



Theme: meta-research

Genome Biology ✓
@GenomeBiology

Following

We're very excited to be launching a special issue on benchmarking of bioinformatic tools, together with the guest editors @markrobinsonca and @olgavitek. For further information, see this link: biomedcentral.com/collections/be …

8:16 AM - 5 Mar 2019

REVIEW Open Access

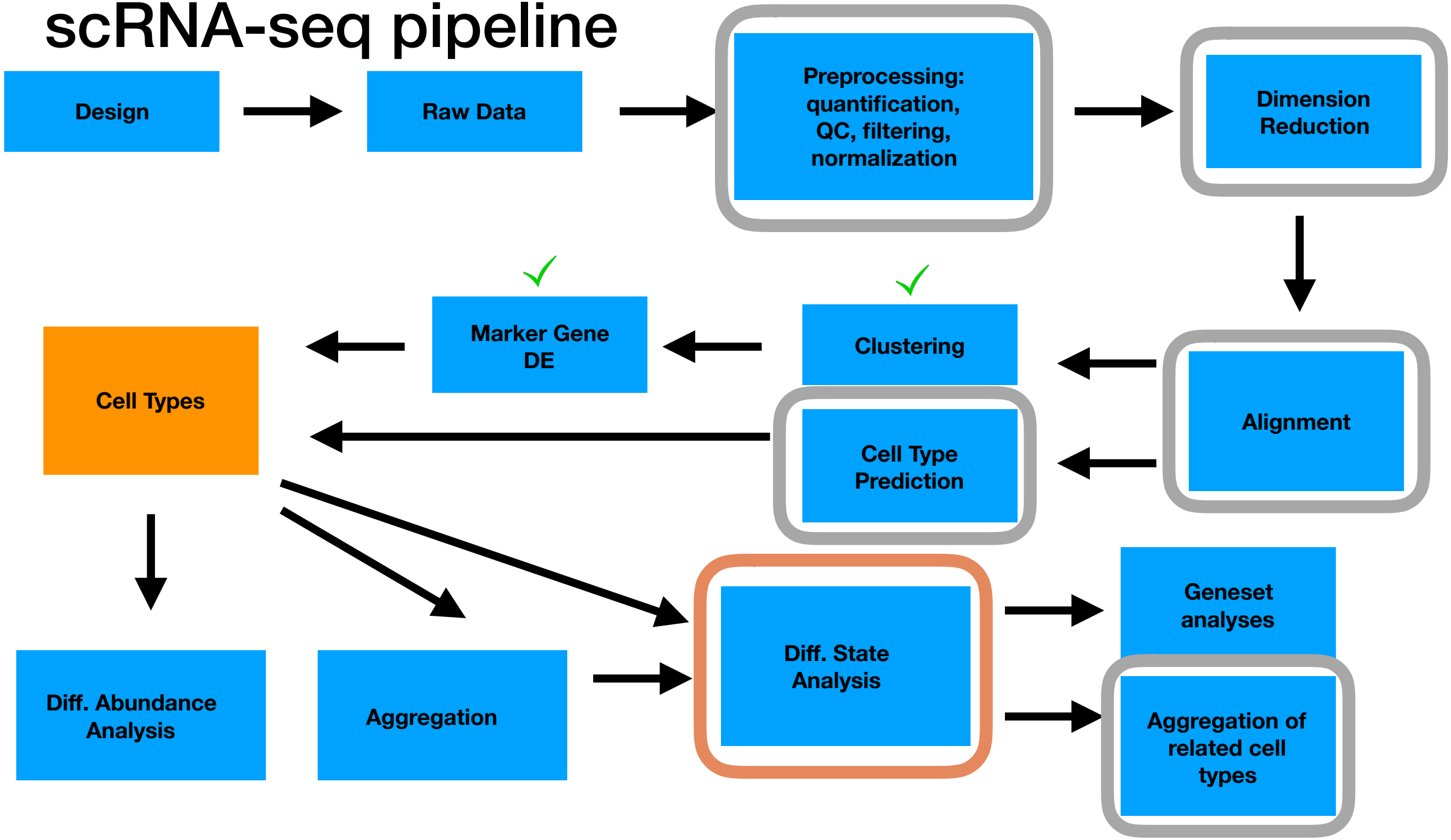
Essential guidelines for computational method benchmarking

Check for updates

Lukas M. Weber[1,2], Wouter Saelens[3,4], Robrecht Cannoodt[3,4], Charlotte Soneson[1,2,8], Alexander Hapfelmeier[5], Paul P. Gardner[6], Anne-Laure Boulesteix[7], Yvan Saeys[3,4*] and Mark D. Robinson[1,2*]

One critical aspect: making reference datasets available so that new methods can be tested.

scRNA-seq pipeline



Revealing the vectors of cellular identity with single-cell genomics

Allon Wagner[1], 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

Type: more permanent

State: more transient

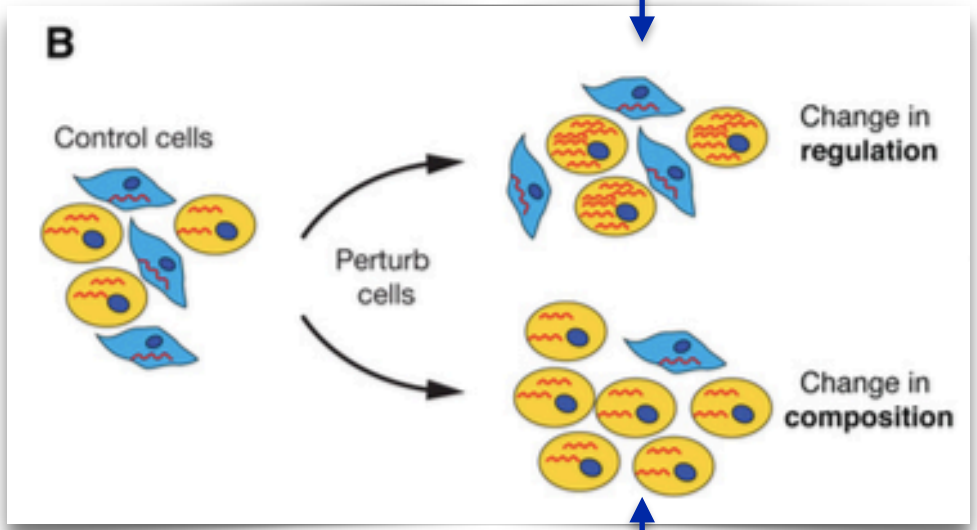
Perspective

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 Xia[1] and Itai Yanai[1,2,*]

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

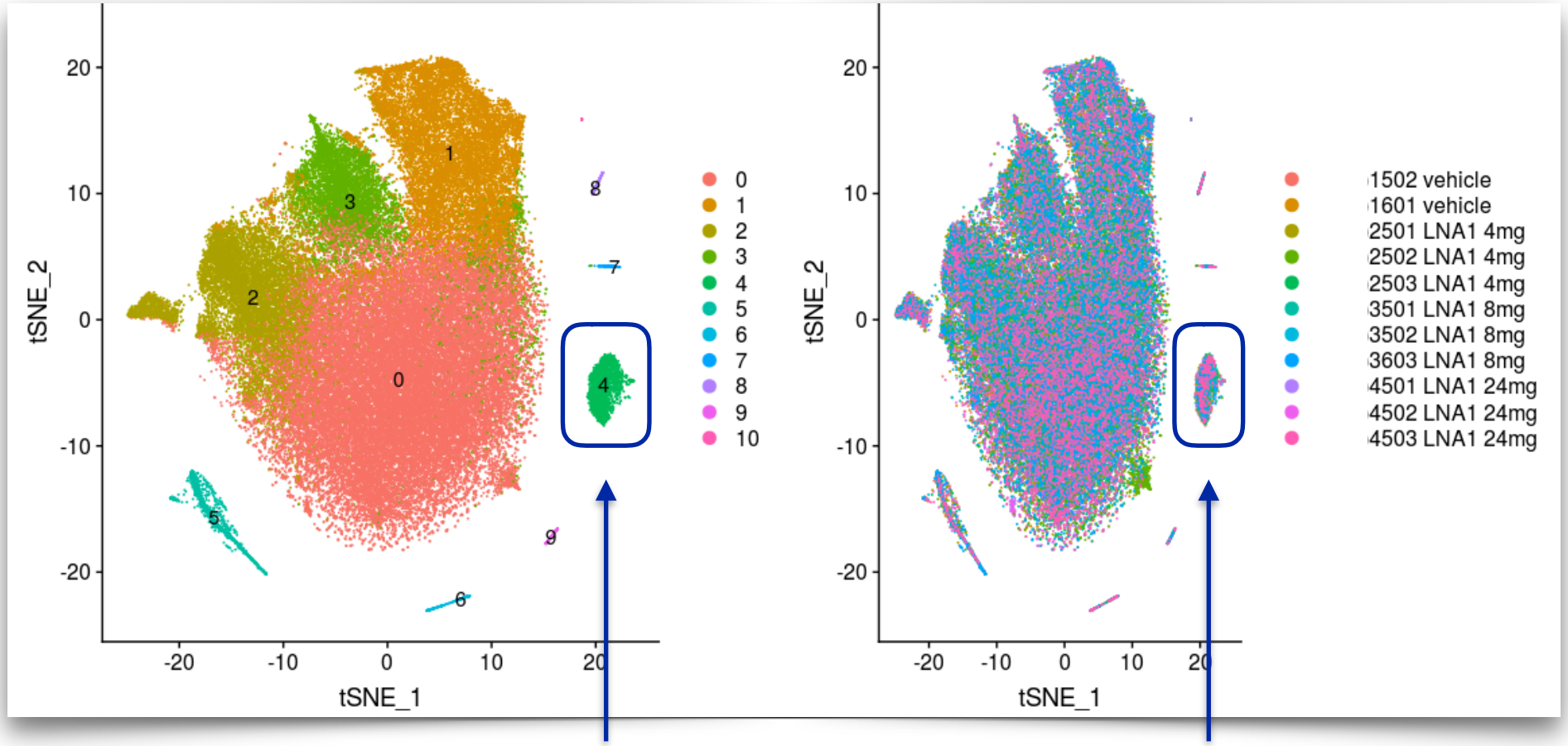
SPOTLIGHT

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



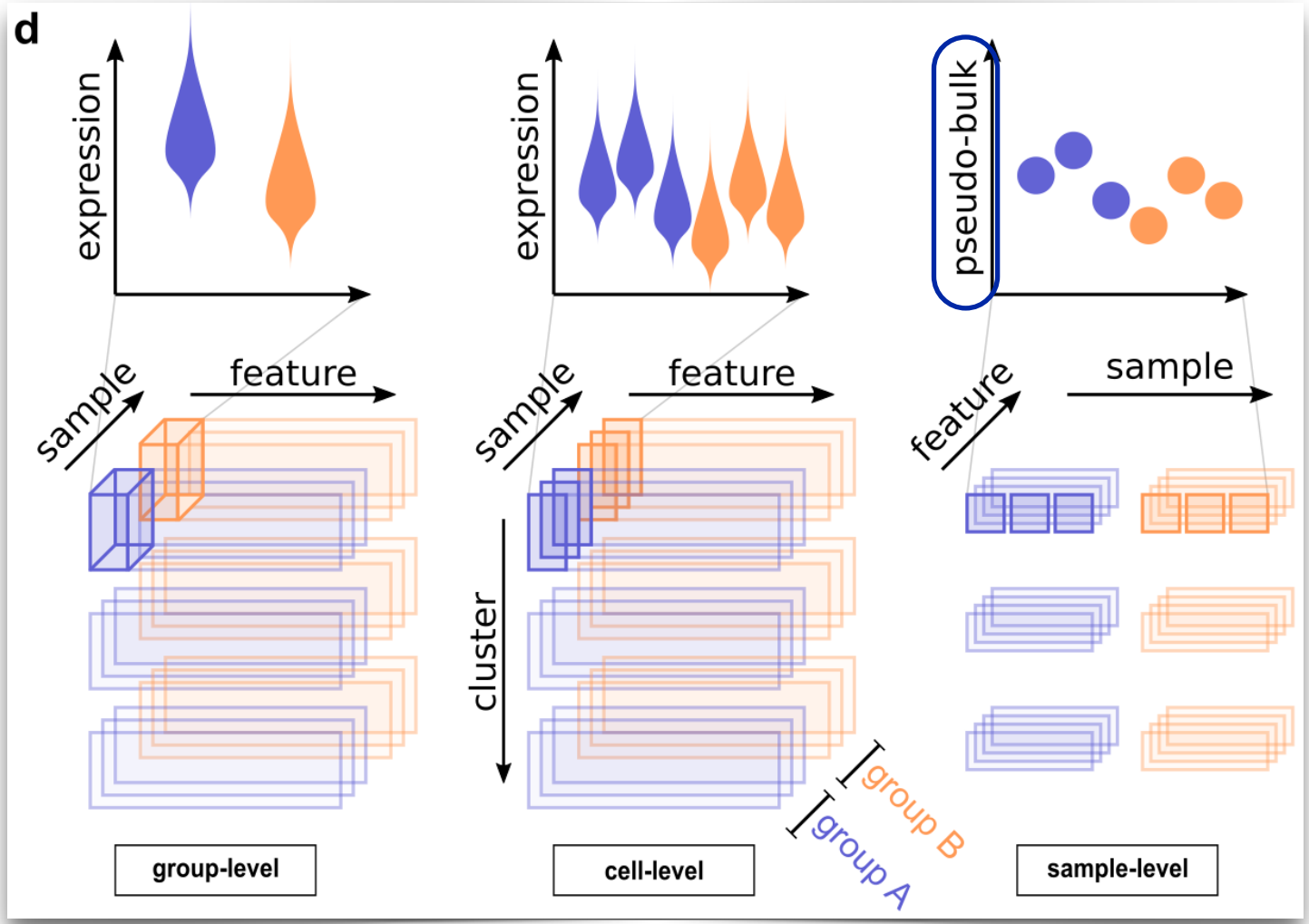
repeat for each population ..

Focus: Marker gene DE

Focus: cross-sample DE

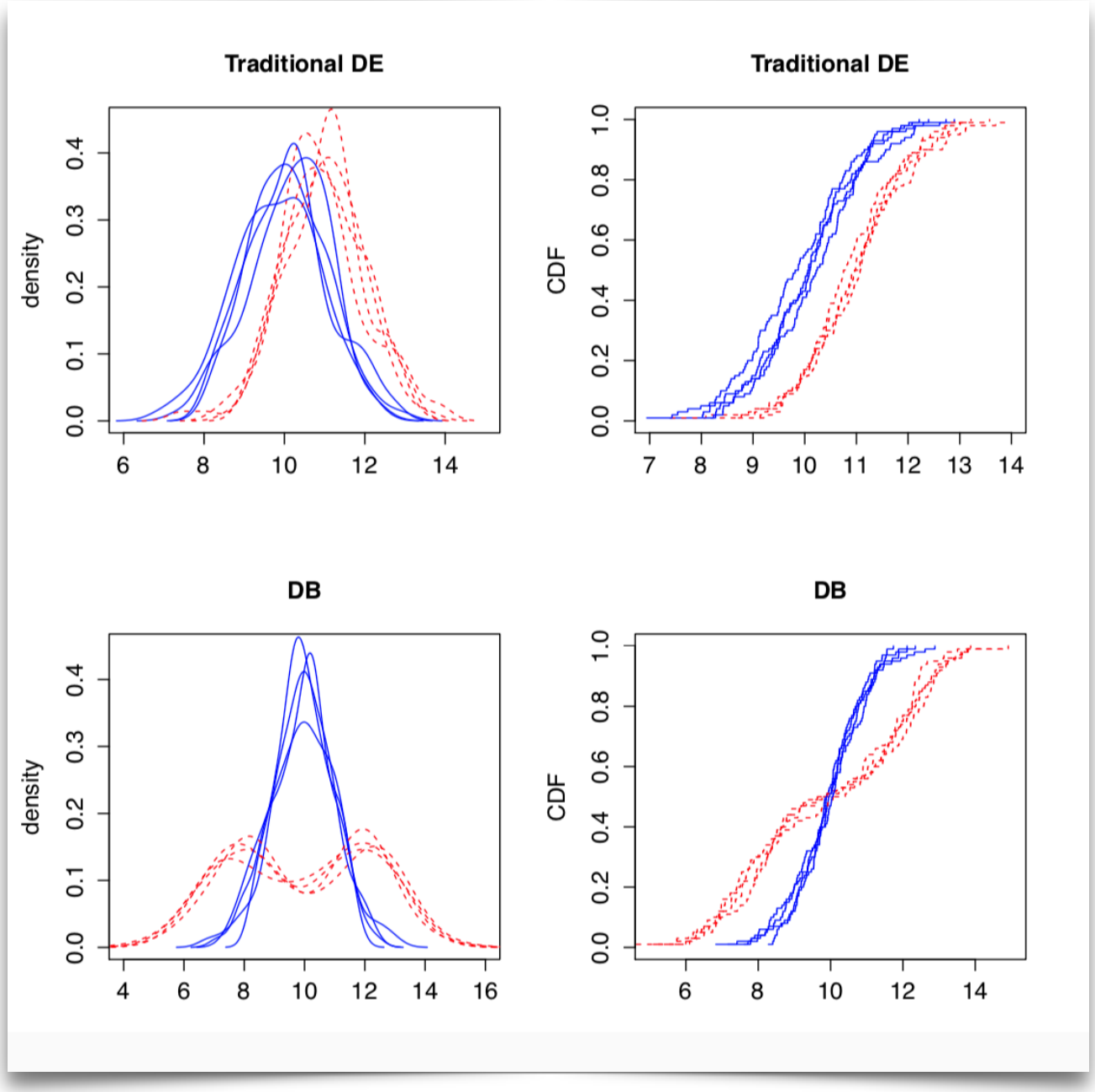
After “Cell Type Prediction” / “Clustering”, various ways to view the inference problem

Multi-sample
Multi-condition
Multi-population



Limited “off-the-shelf” options for comparison of distributions

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



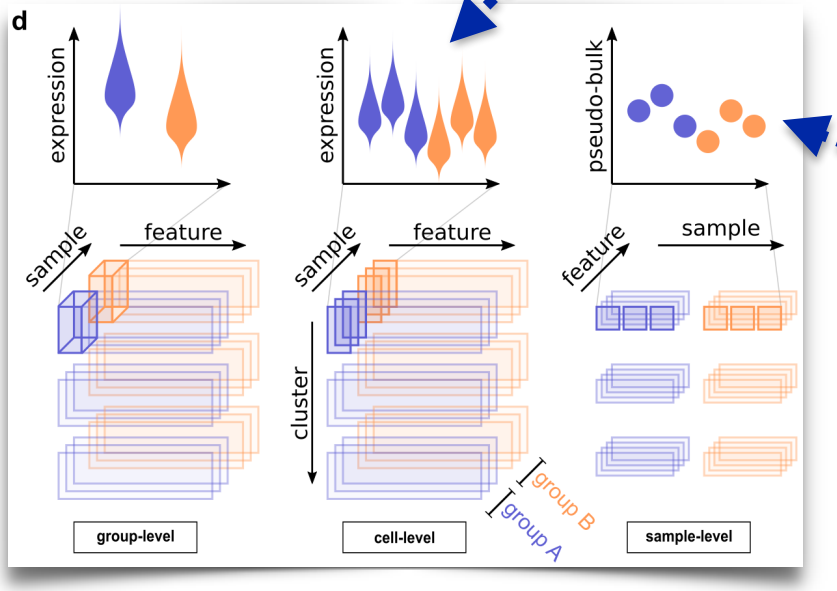
Some precedent, but different contexts

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. Burnett[1], 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. MARIONI
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
marioni@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 **E**xpression

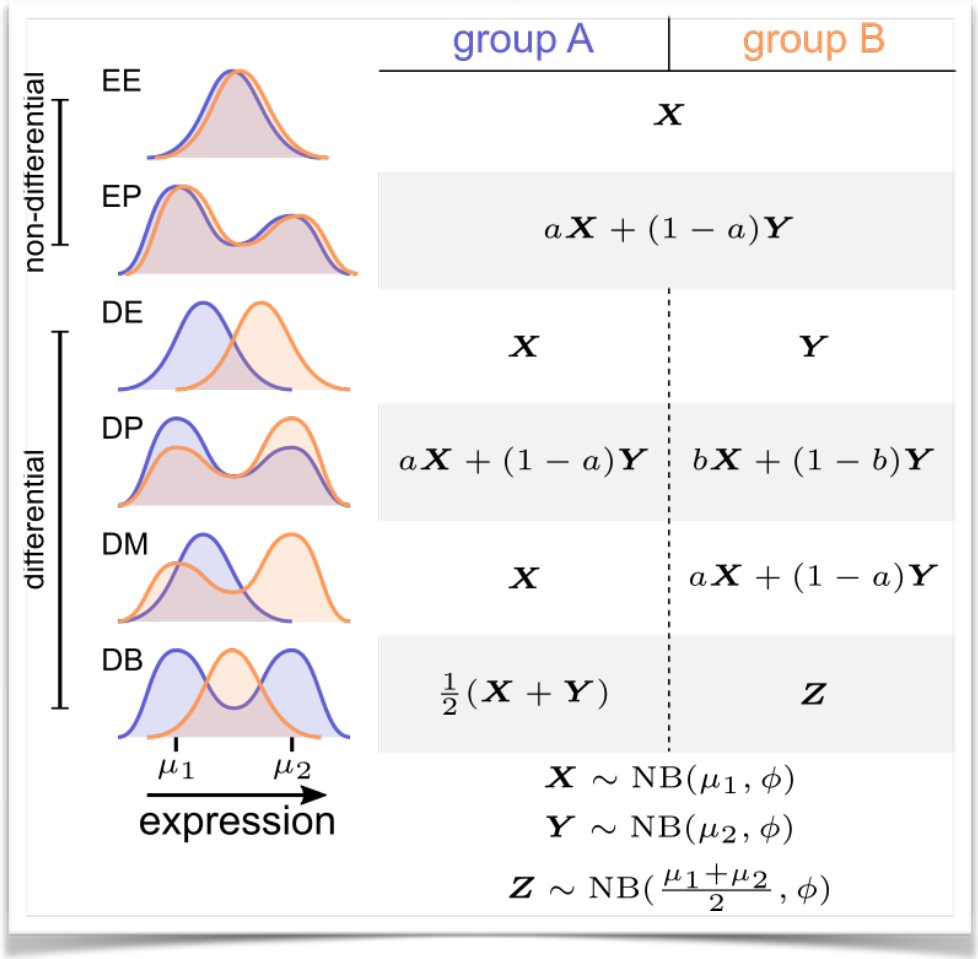
Equivalent **P**roportions

Differential **E**xpression

Differential **P**roportions

Differential **M**odality

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



(shift in means)

(identical means)

…idea adapted from Korthauer et al., 2016


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

Genome Biology

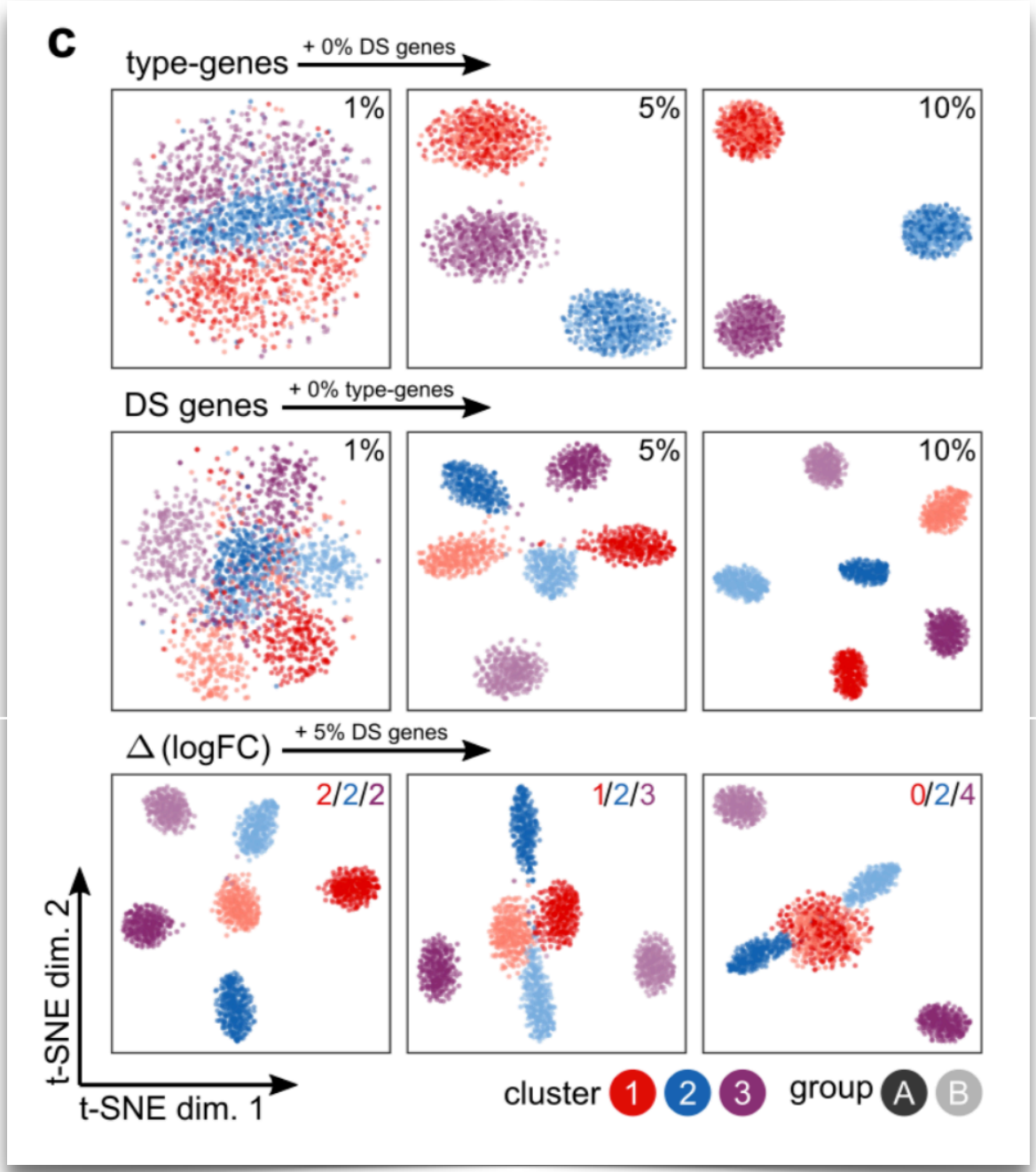
METHOD Open Access

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

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

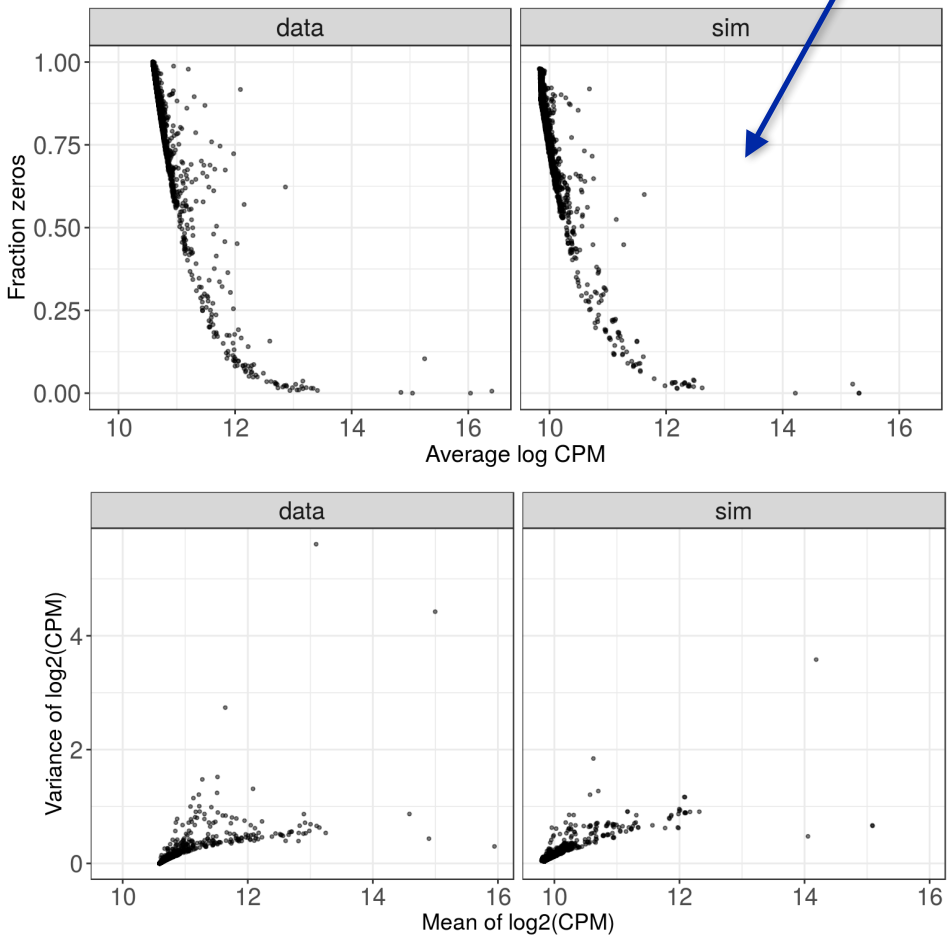
Flexibility of simulation

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

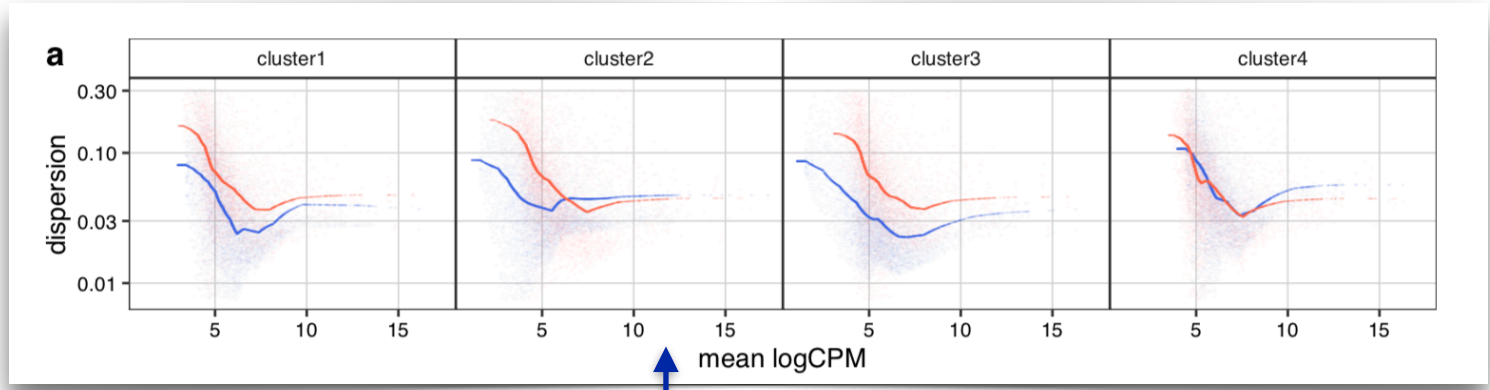


countsimQC: comparing simulated and real data

cell-level properties



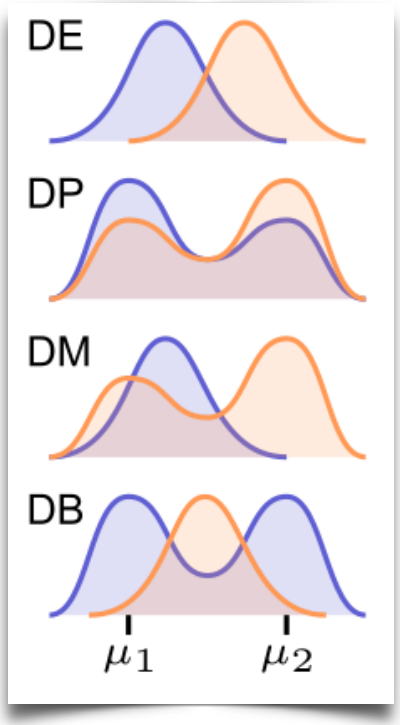
pseudobulk-level dispersion-mean relationships

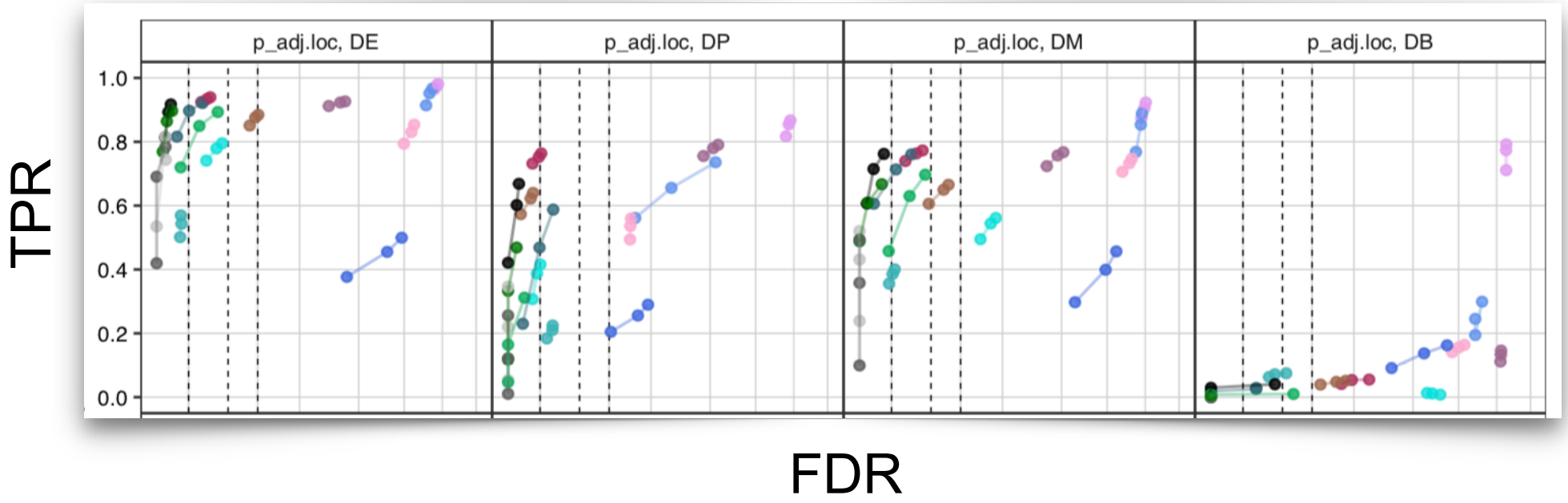


aggregate-level properties

https://bioconductor.org/packages/release/bioc/html/countsimQC.html

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

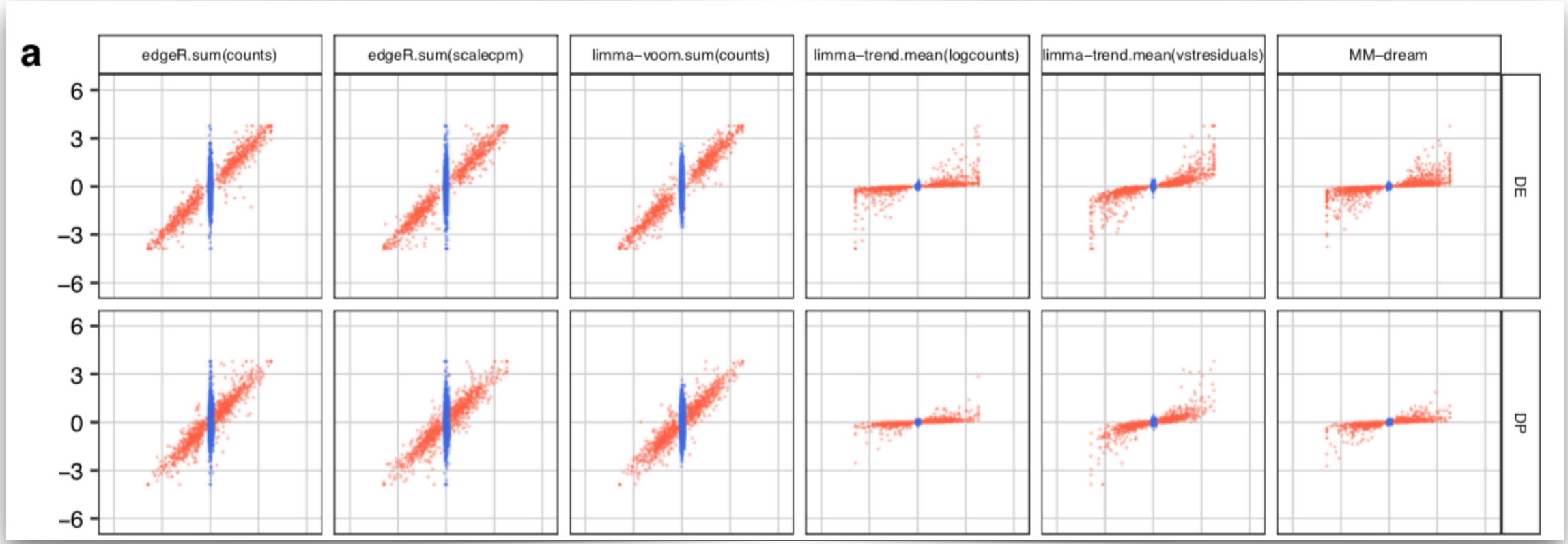




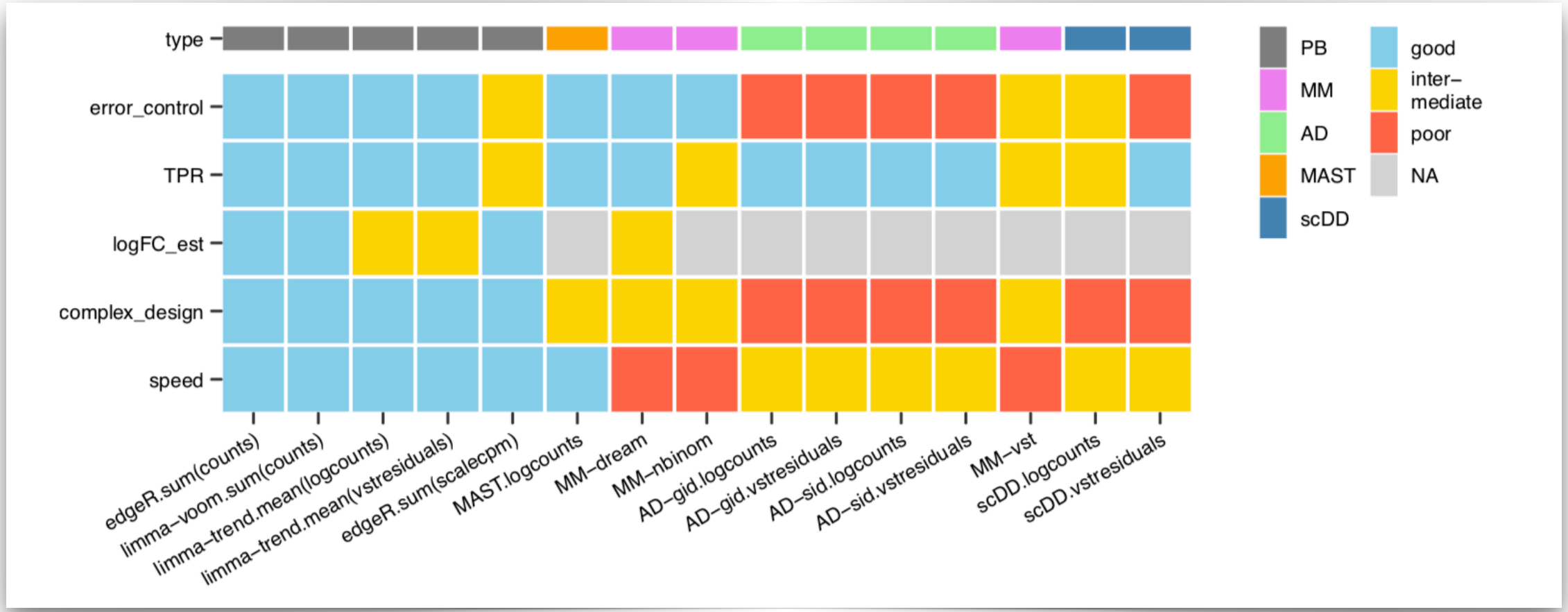
AD = Anderson-Darling
MM = mixed models

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

Pick your data to model wisely



Current rating



PB = pseudobulk
AD = Anderson-Darling
MM = mixed models

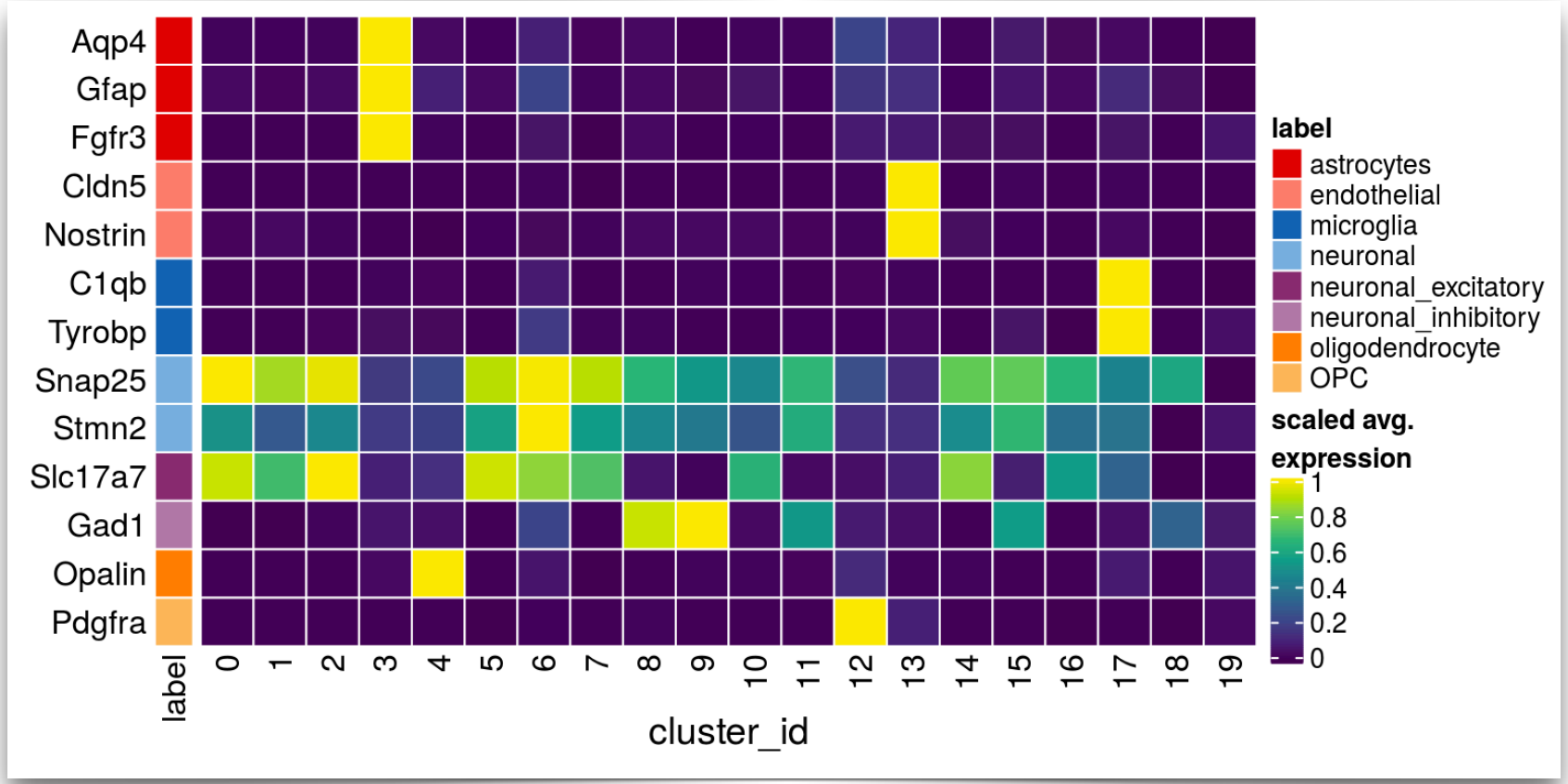
Application to LPS dataset: clustering + annotation subpopulations

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

frontal cortex

single nuclei RNA-seq (10x)

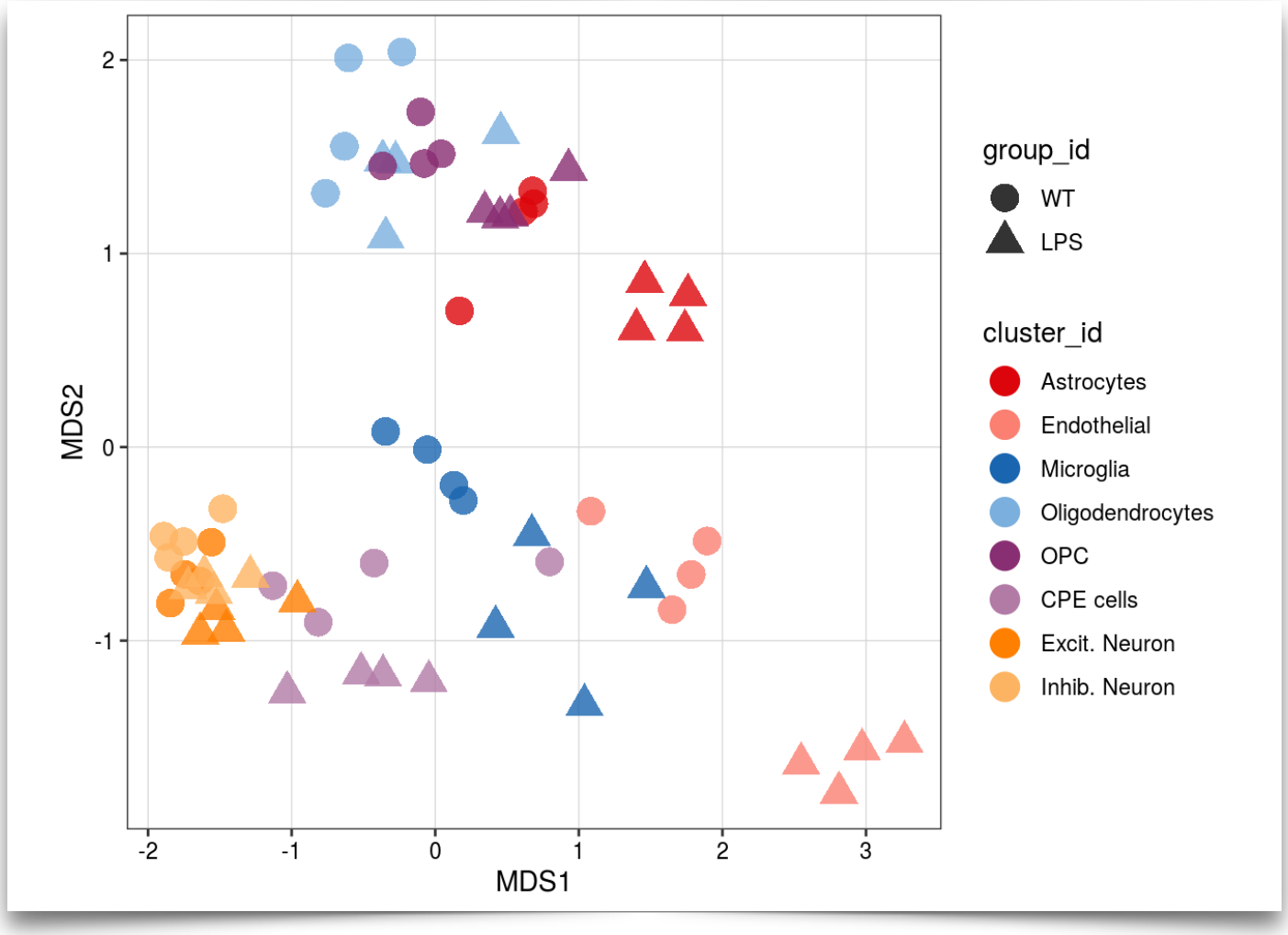
usual preprocessing:
filtering, doublet removal,
Seurat integration,
clustering



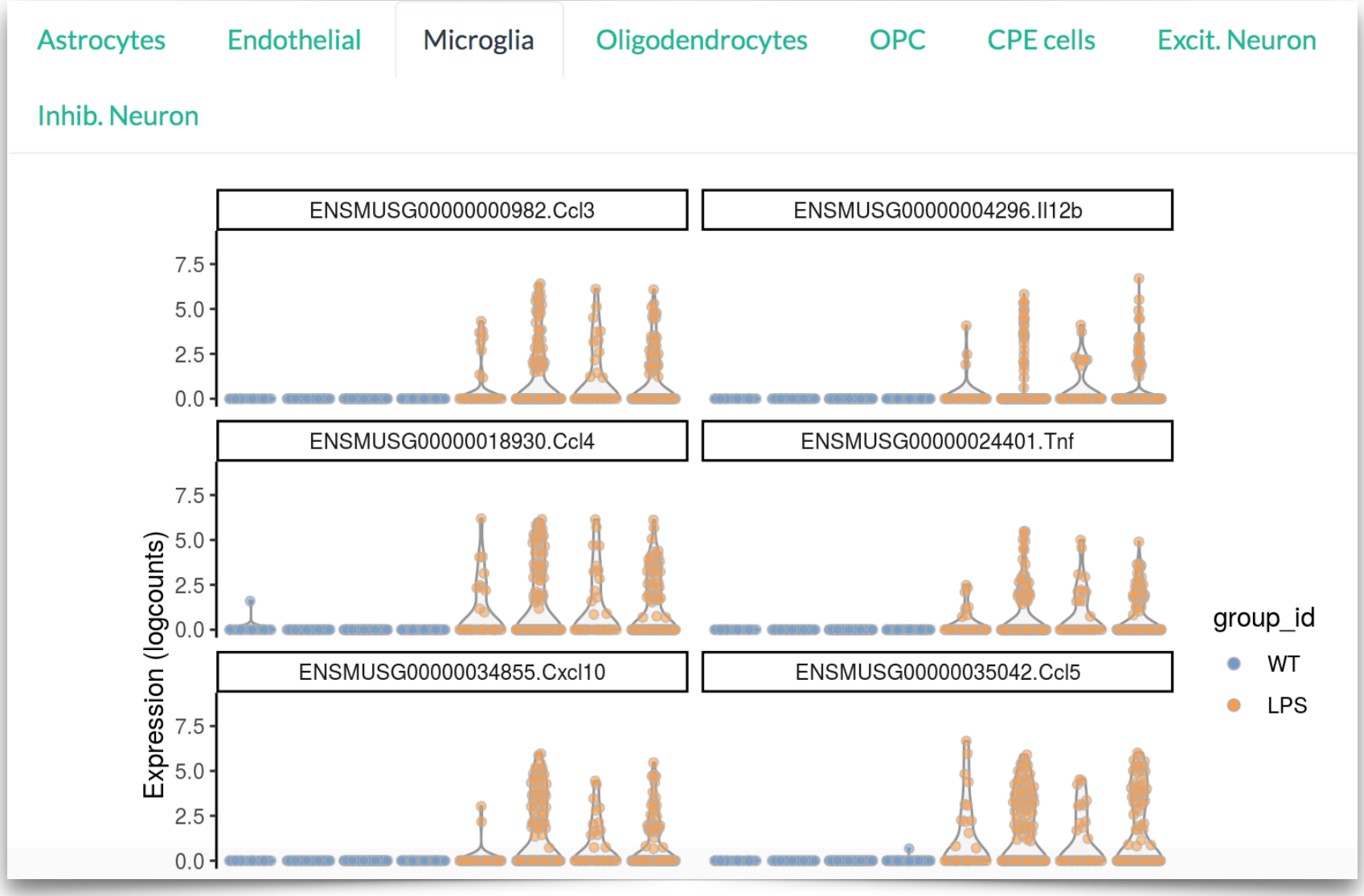
Application to LPS dataset: subpopulation-level visualization

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

Each dot is one subpopulation/ sample combination

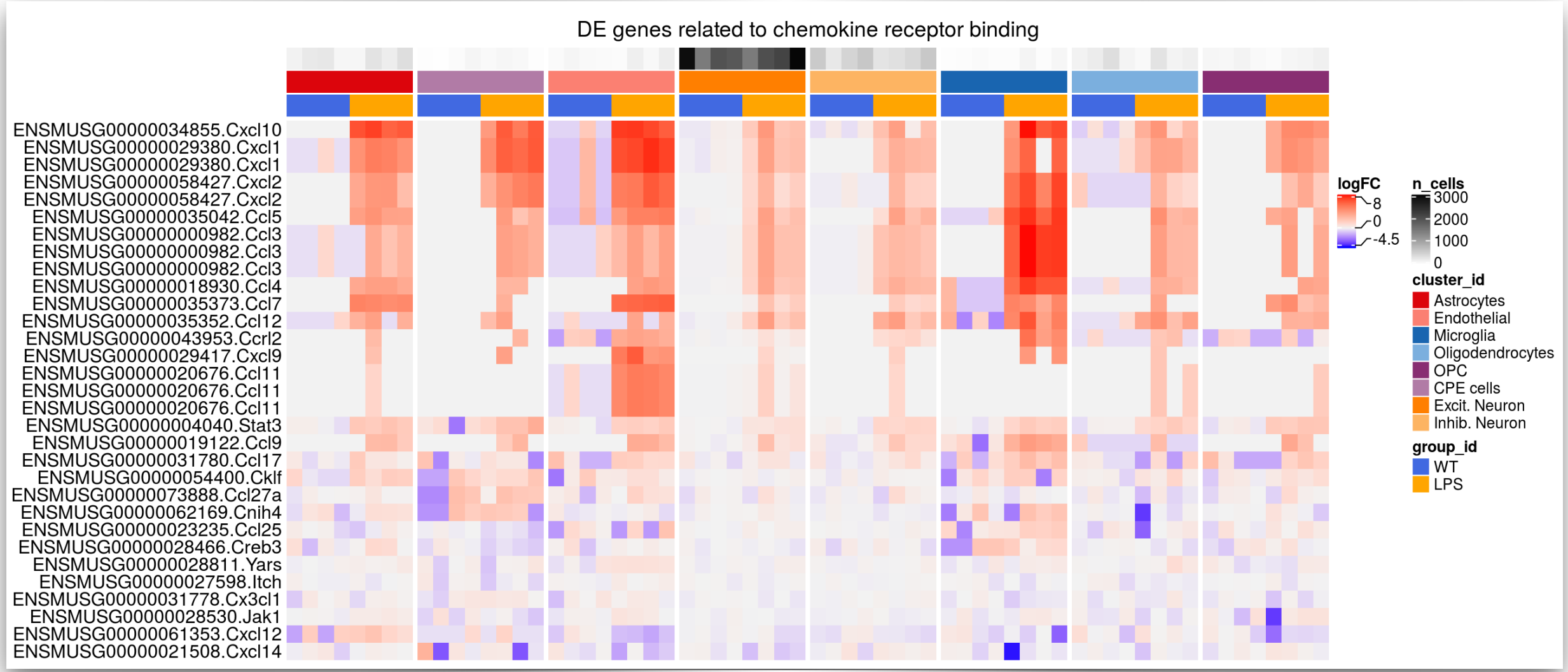


Application to LPS dataset: go back to cell-level response (discovery based on pseudobulk)



workflowr !

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

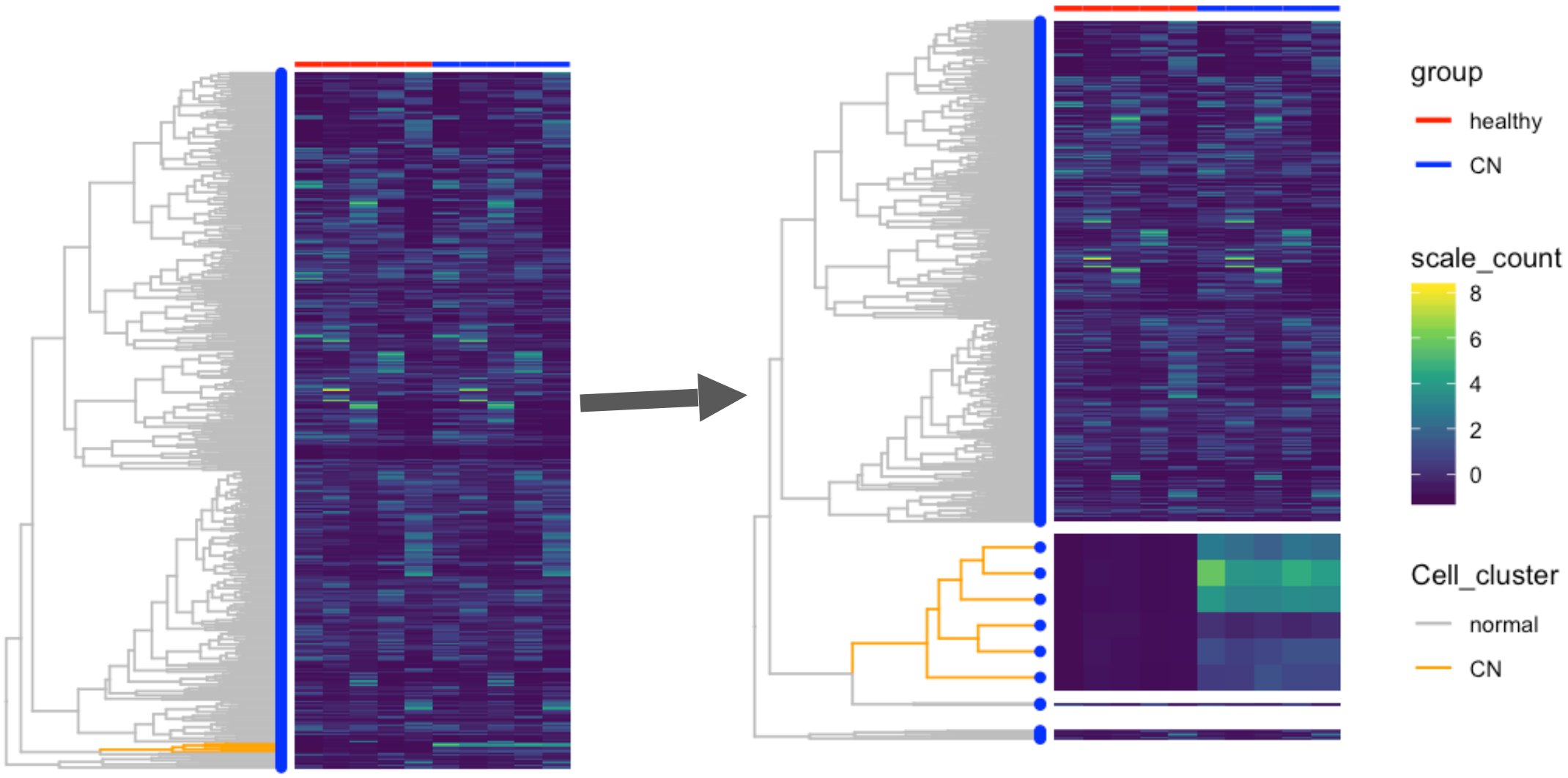


LPS dataset: interplay of cell type and cell state

A couple things you can’t read in the preprint

- Maybe relationships between cells are important —> use a tree of such relations to guide the inference of differential expression
- Can we do better by looking at full distributions instead of aggregating?

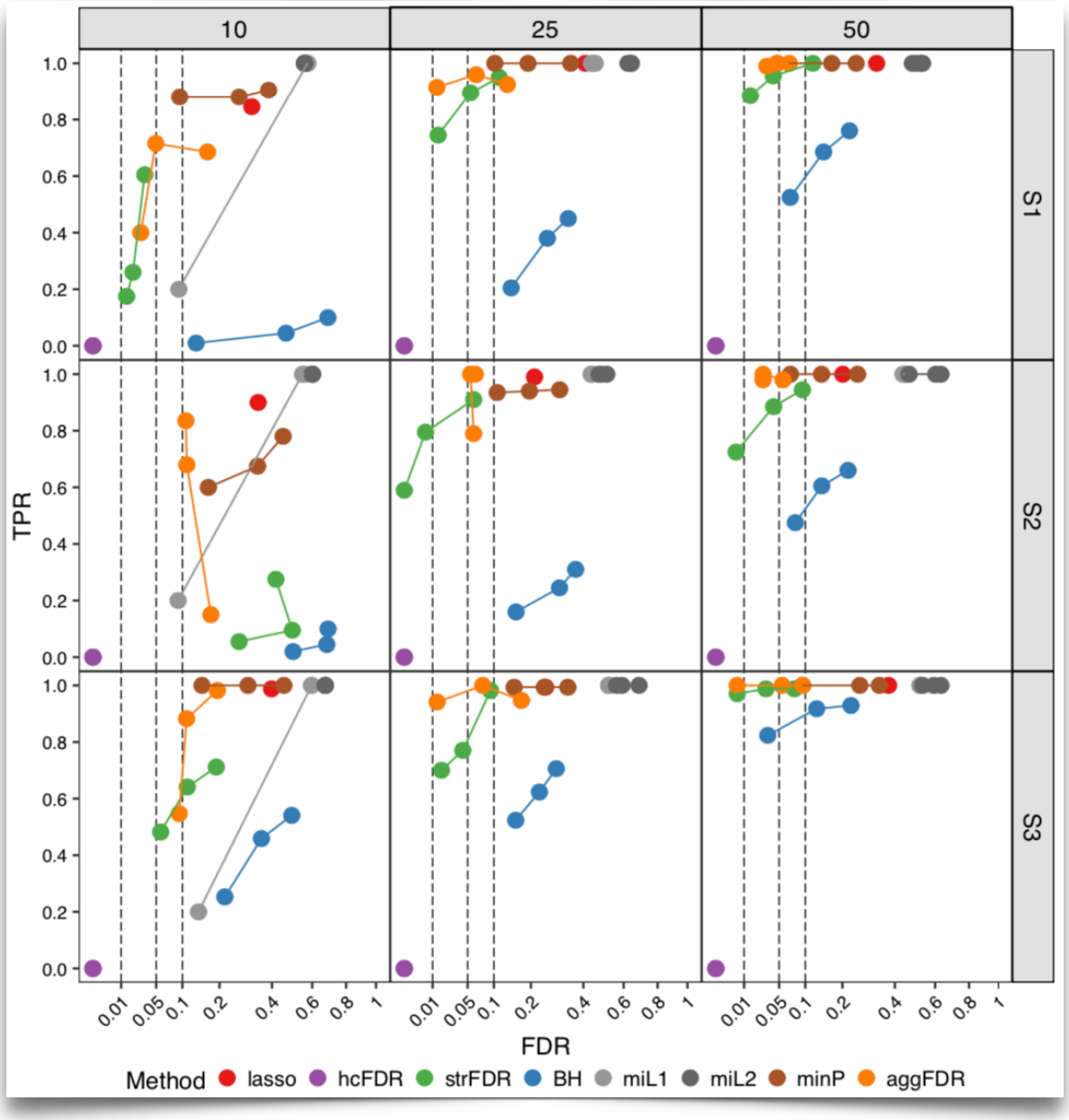
Motivation: can we use the tree information and perform differential inferences across resolutions?

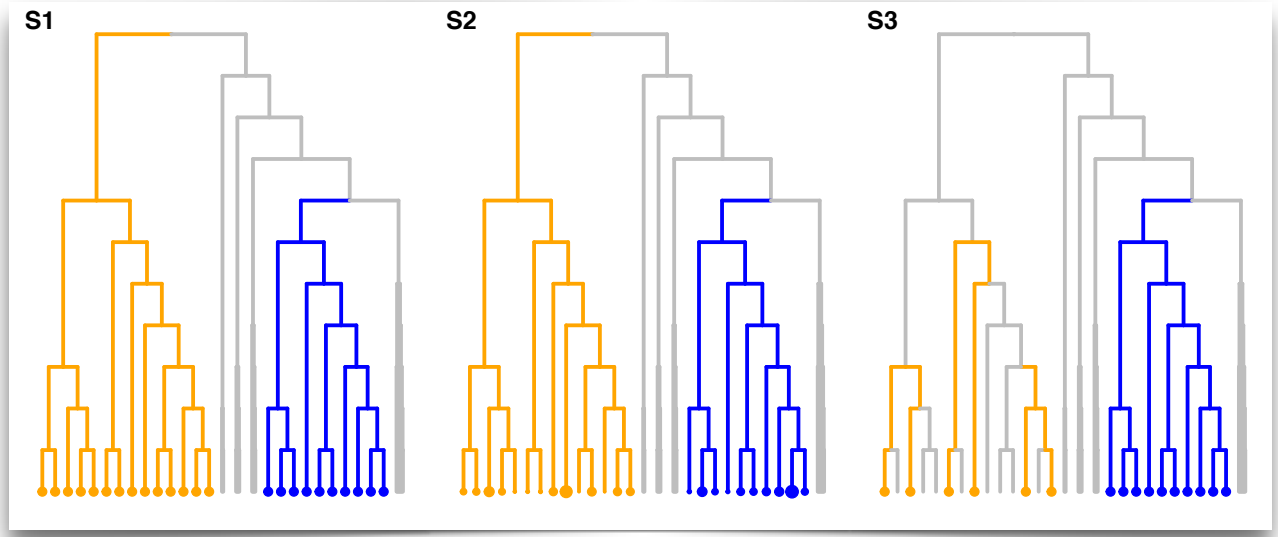


Give more space to orange branch; The visualization is on the leaf level (blue points)

Fiona

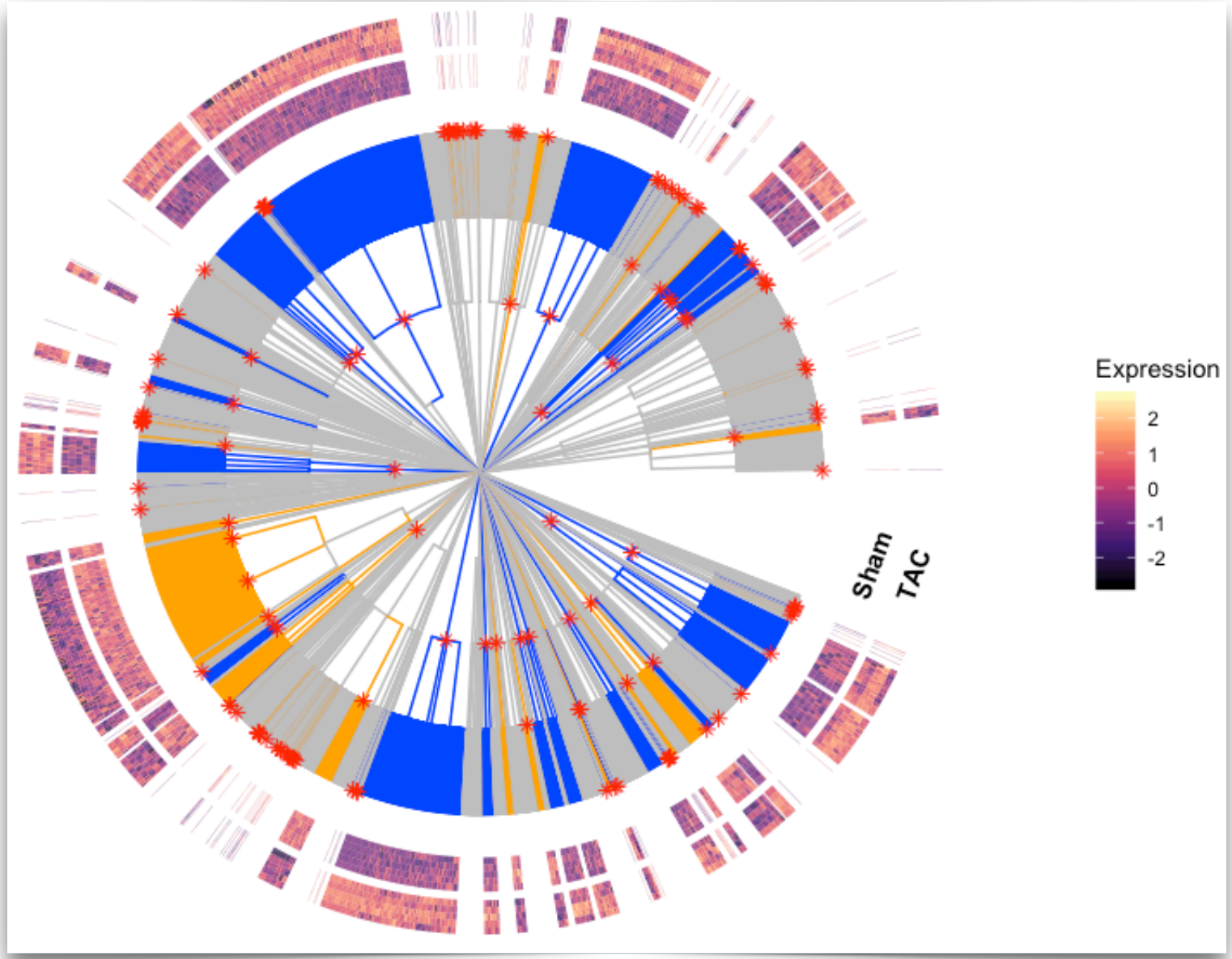
aggFDR compares favourably across scenarios and sample sizes





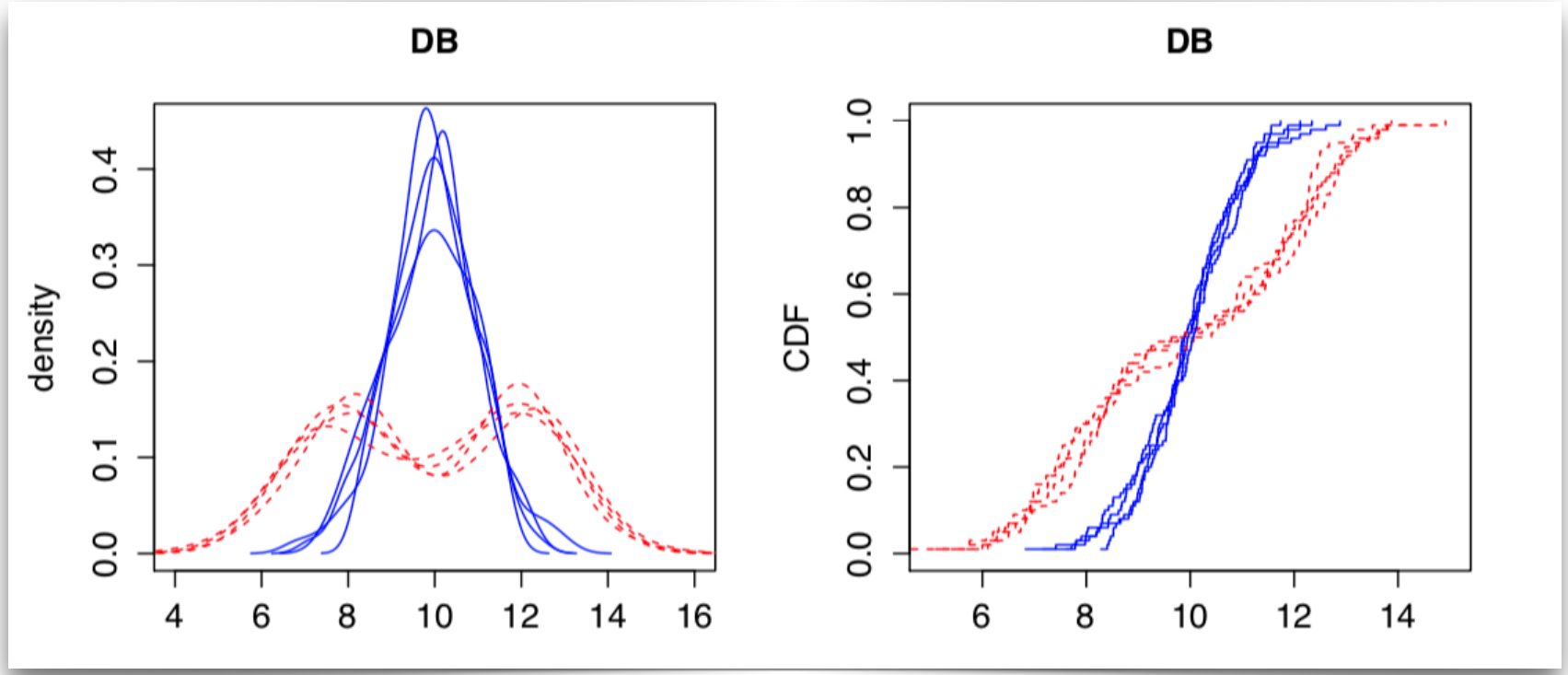
For single cell: generate tree from clustering of type genes, use **aggFDR** to report data-dependent resolution of differential states

Analyses across resolutions



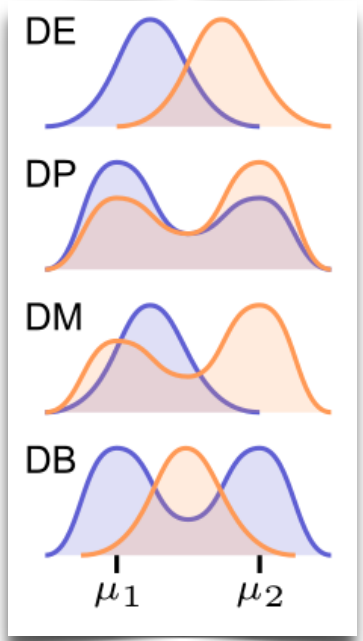
(here: microRNA relationships and corresponding gene expression, but could be cells)

Differential state from full distributions (no aggregation)

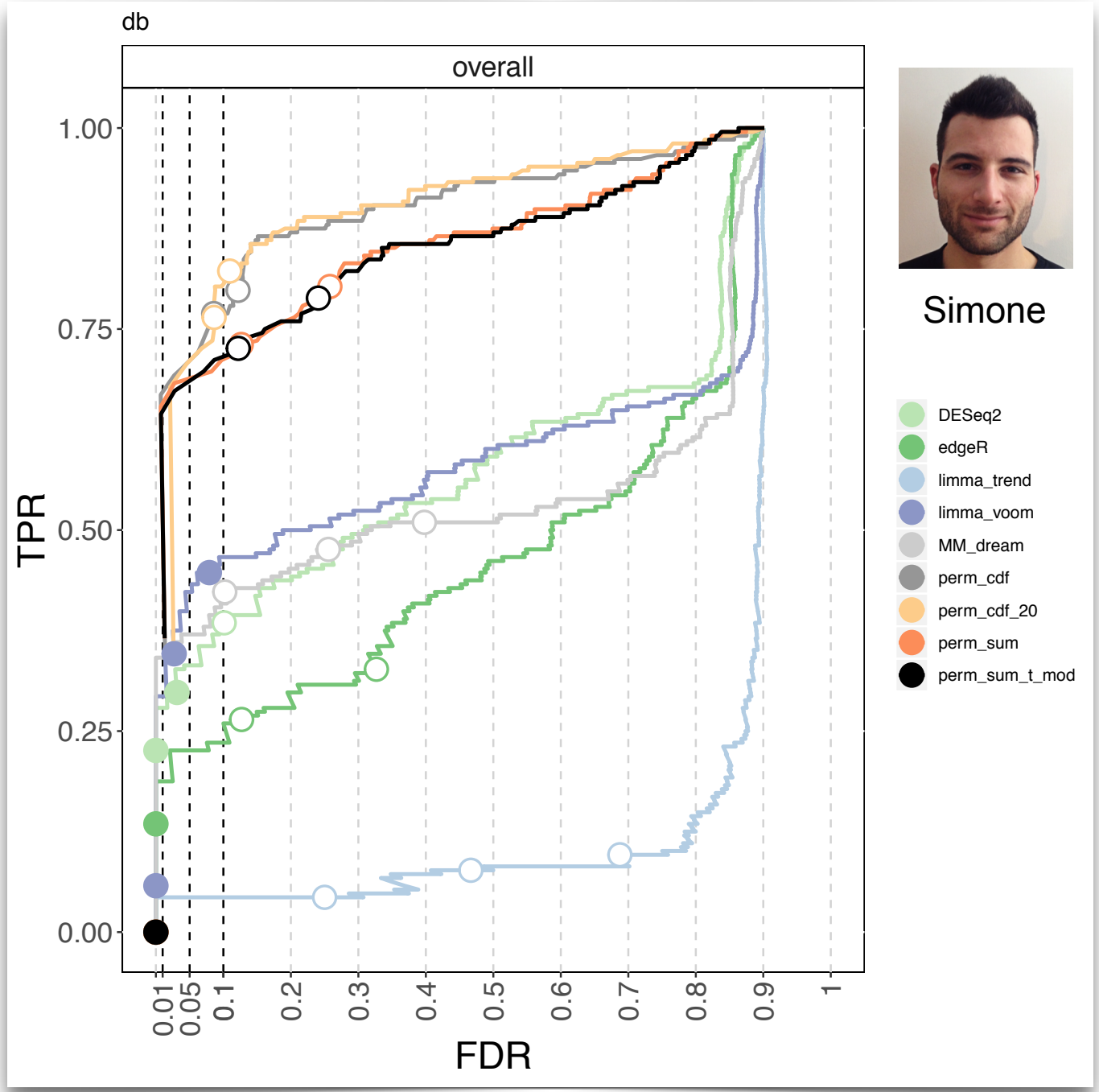


Idea: shuffle cells from multiple samples to generate a permutation test, compare some summary measure of the distribution

Differential state from full distributions (no aggregation)



Permutation testing shows big gains over aggregation for DB. Comparable performance for DE, DM, DP.



Simone

Comments

- multi-sample multi-condition multi-subpopulation datasets —> *in silico* sorting + **differential state analysis**
- Aggregation (e.g., **pseudobulk** counts) works well, is fast, flexible and modular
- software: https://github.com/HelenaLC/muscat, aggFDR, ..

- Are we getting deep enough (per cell, per subpopulation)? —> power differs by cell type
- Interplay between definition of type and state: discretization, but at what resolution? —> data-driven aggregation along tree
- Should we fit separate models for each subpopulation (what we do now) or one model over all subpopulations?
- How to best use batch correction methods, cell type assignment methods
- Extensions to trajectories?

Statistical Bioinformatics Group, IMLS, UZH

Roche

Dheeraj Maholtra
Daniela Calini

SIB Swiss Institute of Bioinformatics

Universität Zürich[UZH]

FNSNF
Fonds national suisse
Schweizerischer Nationalfonds
Fondo nazionale svizzero
Swiss National Science Foundation

University of Zurich[UZH]
URPP Evolution in Action: From Genomes to Ecosystems

Chan Zuckerberg Initiative

Preprint

http://bit.ly/2K4jKzK
or Google: “crowell biorxiv muscat”

On the discovery of population-specific state transitions from multi-sample multi-condition single-cell RNA sequencing data

Helena L. Crowell[1,2], Charlotte Soneson[1,2,3,*], Pierre-Luc Germain[1,4,*], Daniela Calini[5], Ludovic Collin[5], Catarina Raposo[5], Dheeraj Malhotra[5], and Mark D. Robinson[1,2]

[1] *Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland*
[2] *SIB Swiss Institute of Bioinformatics, Zurich, Switzerland*
[3] *Present address: Friedrich Miescher Institute for Biomedical Research and SIB Swiss Institute of Bioinformatics, Basel, Switzerland*
[4] *D-HEST Institute for Neuroscience, Swiss Federal Institute of Technology, Zurich, Switzerland*
[5] *F. Hoffmann-La Roche Ltd, Pharma Research and Early Development, Neuroscience, Ophthalmology and Rare Diseases, Roche Innovation Center Basel, Basel, Switzerland*
[*] *These authors contributed equally.*

July 26, 2019