Universität Zürich[UZH]

URPP Evolution in Action

SIB Swiss Institute of Bioinformatics

Statistical methods for flexible differential analysis of cross-sample 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

Quick talking points

- *Are statistical methods developed for bulk RNAseq data appropriate for single cell datasets?* —> Most definitely yes, but "it depends"
- A plea: when you talk about DE in scRNA-seq, specify what you mean, because I see at least 3 separate but related DE analyses:
    - marker gene DE (associate DE to trajectory)
    - differential abundance analysis
    - differential state analysis
    - ..

Quick advertisements

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

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

Postdoc position available: develop data science tools for (single-cell) -omics data

https://www.travelanddestinations.com/most-beautiful-places-to-visit-switzerland/

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

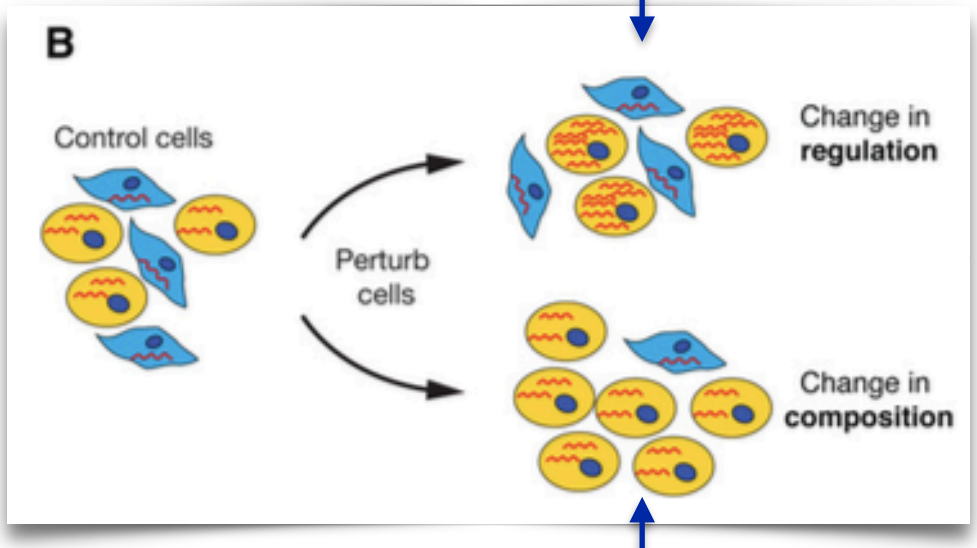
Perspective

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,*]

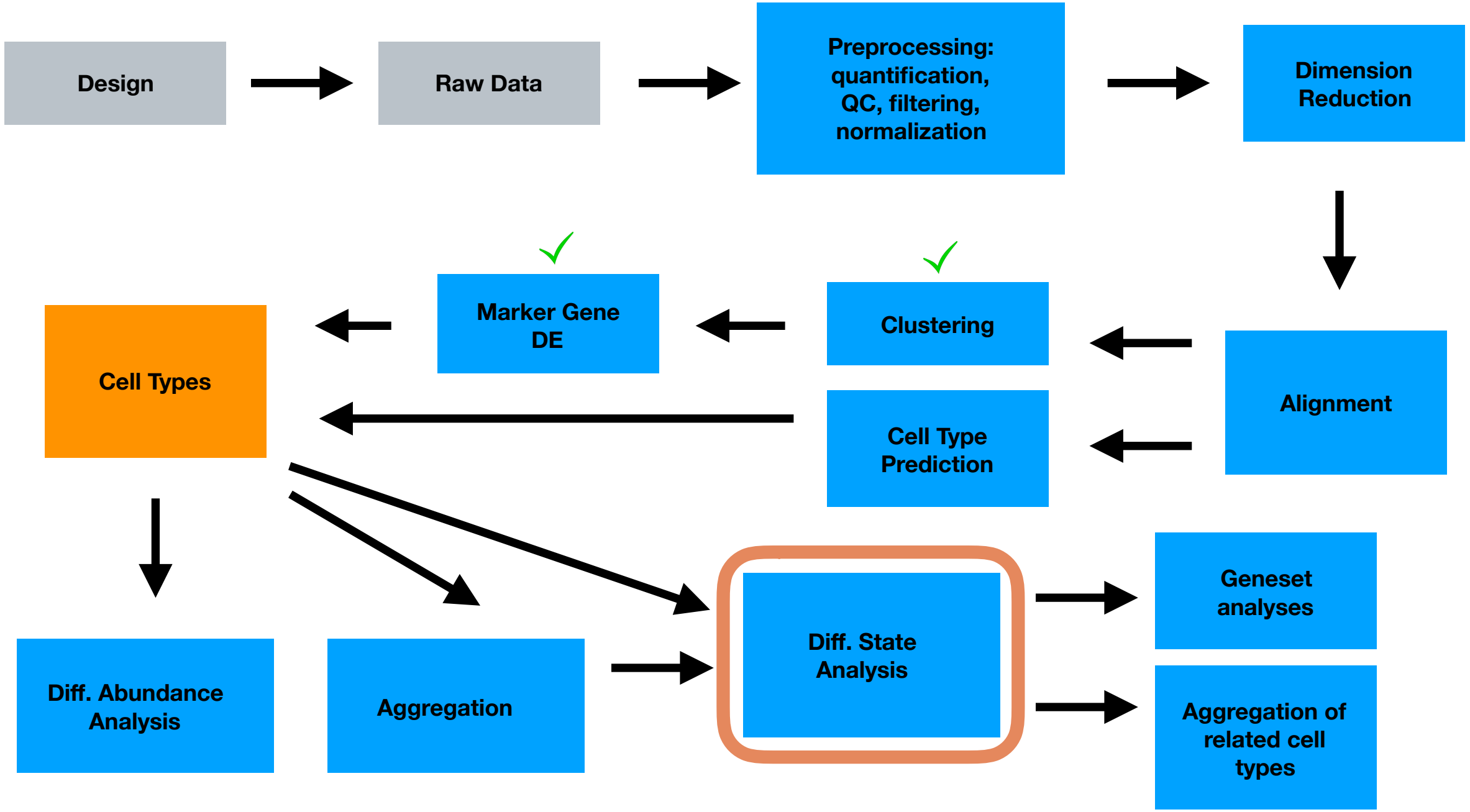
"We view a cell state as a secondary module operating in addition to the general cell type regulatory program."

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?


Design
Raw Data
Preprocessing: quantification, QC, filtering, normalization
Dimension Reduction
Alignment
Clustering
Marker Gene DE
Cell Types
Cell Type Prediction
Diff. Abundance Analysis
Aggregation
Diff. State Analysis
Geneset analyses
Aggregation of related cell types

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

Plea: test your new method on these (and other) benchmark datasets

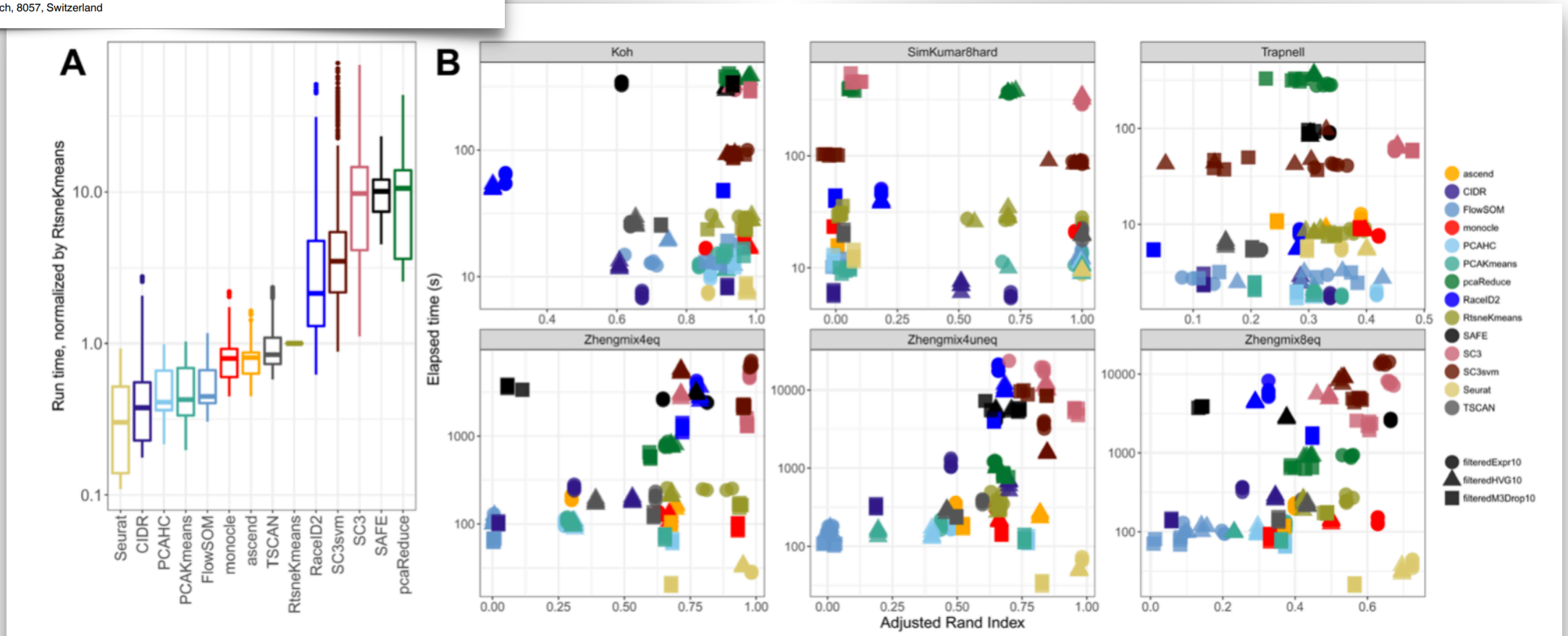


Figure 2. (**A**) Normalized run times, using `RtsneKmeans` as the reference method, across all data set instances and number of clusters. (**B**) Run time versus performance (ARI) for a subset of data sets and filterings, for the true number of clusters.

We know a bit about marker gene DE

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

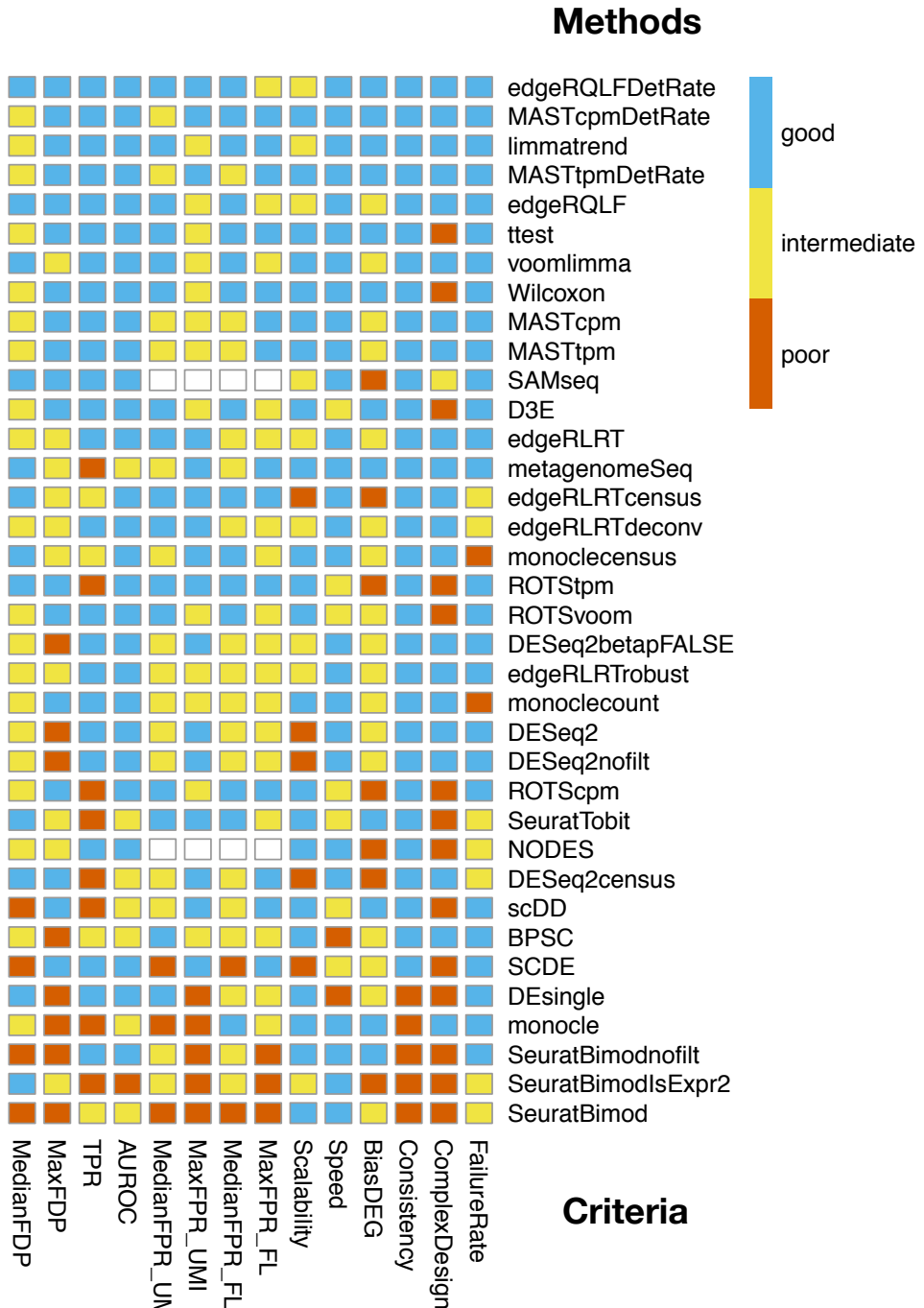
t-test and Wilcoxon perform surprisingly well

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

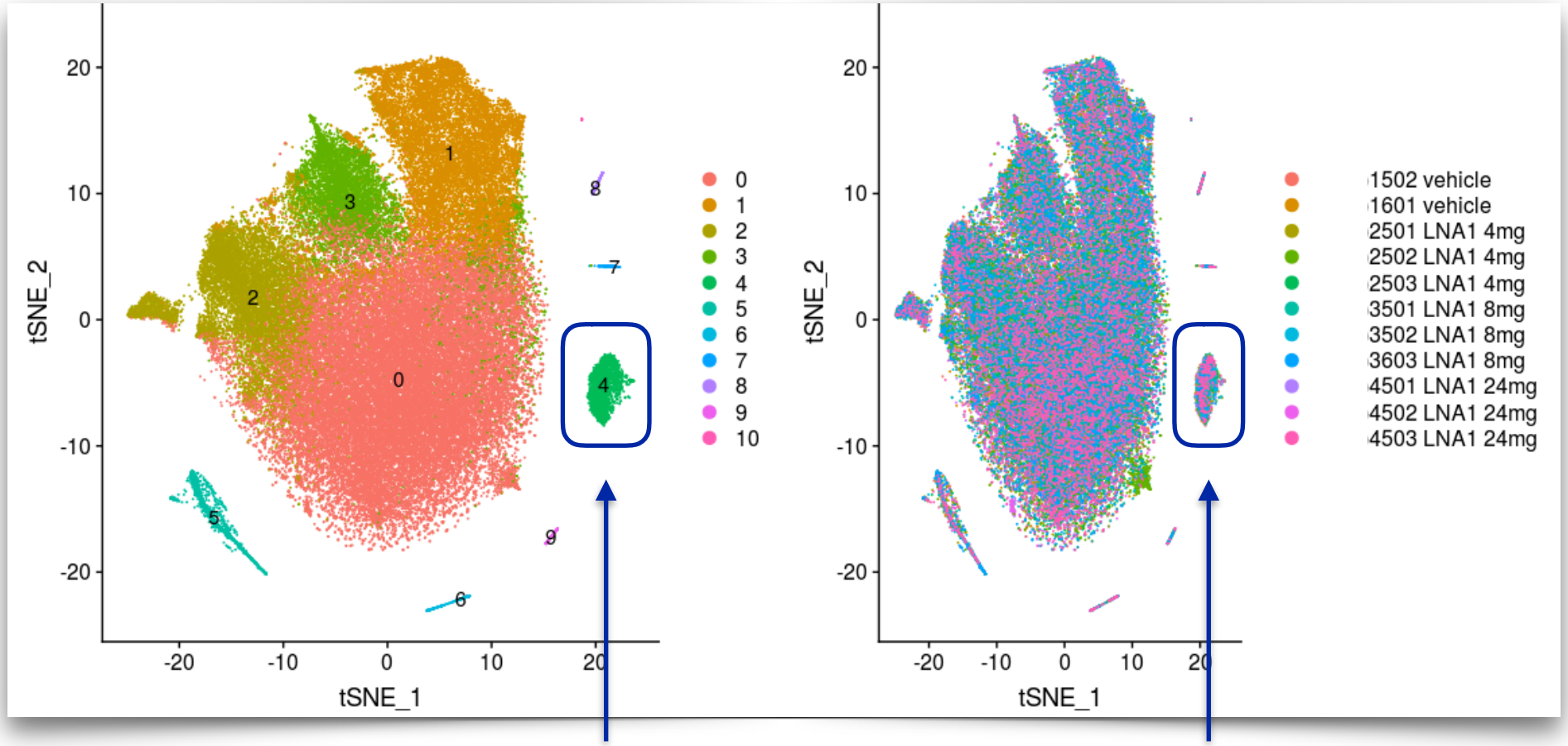
(Plea: test your new method on these (and other) benchmark datasets)

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

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



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



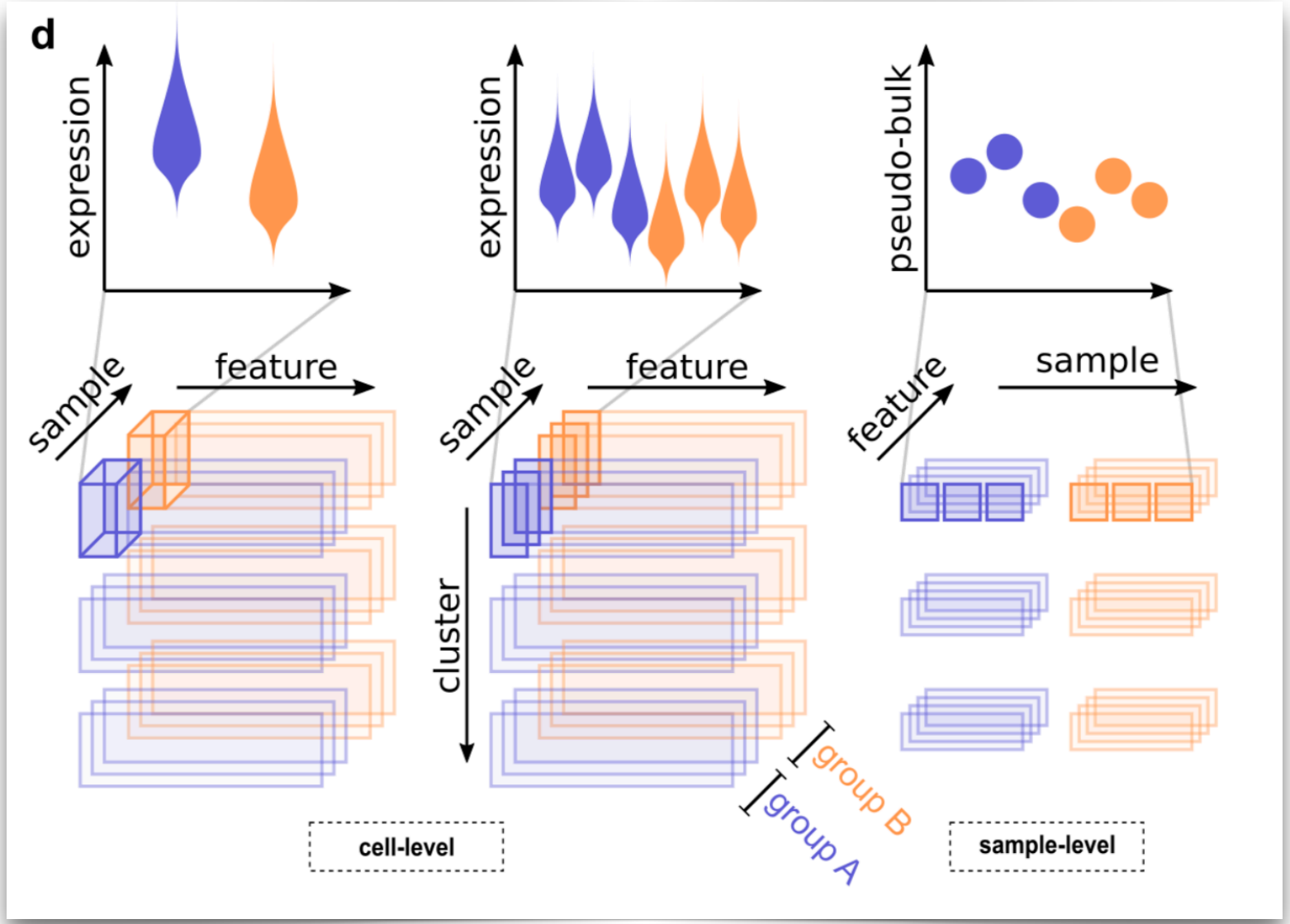
repeat for each population ..

Focus: Marker gene DE

Focus: cross-sample DE

Various ways to frame the inference

Multi-sample
Multi-condition
Multi-population



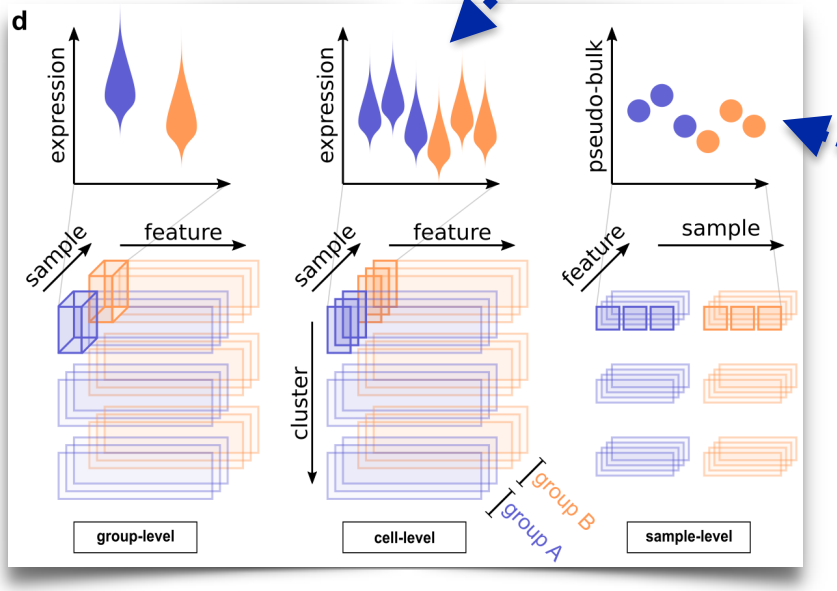
Some precedent, but different contexts

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



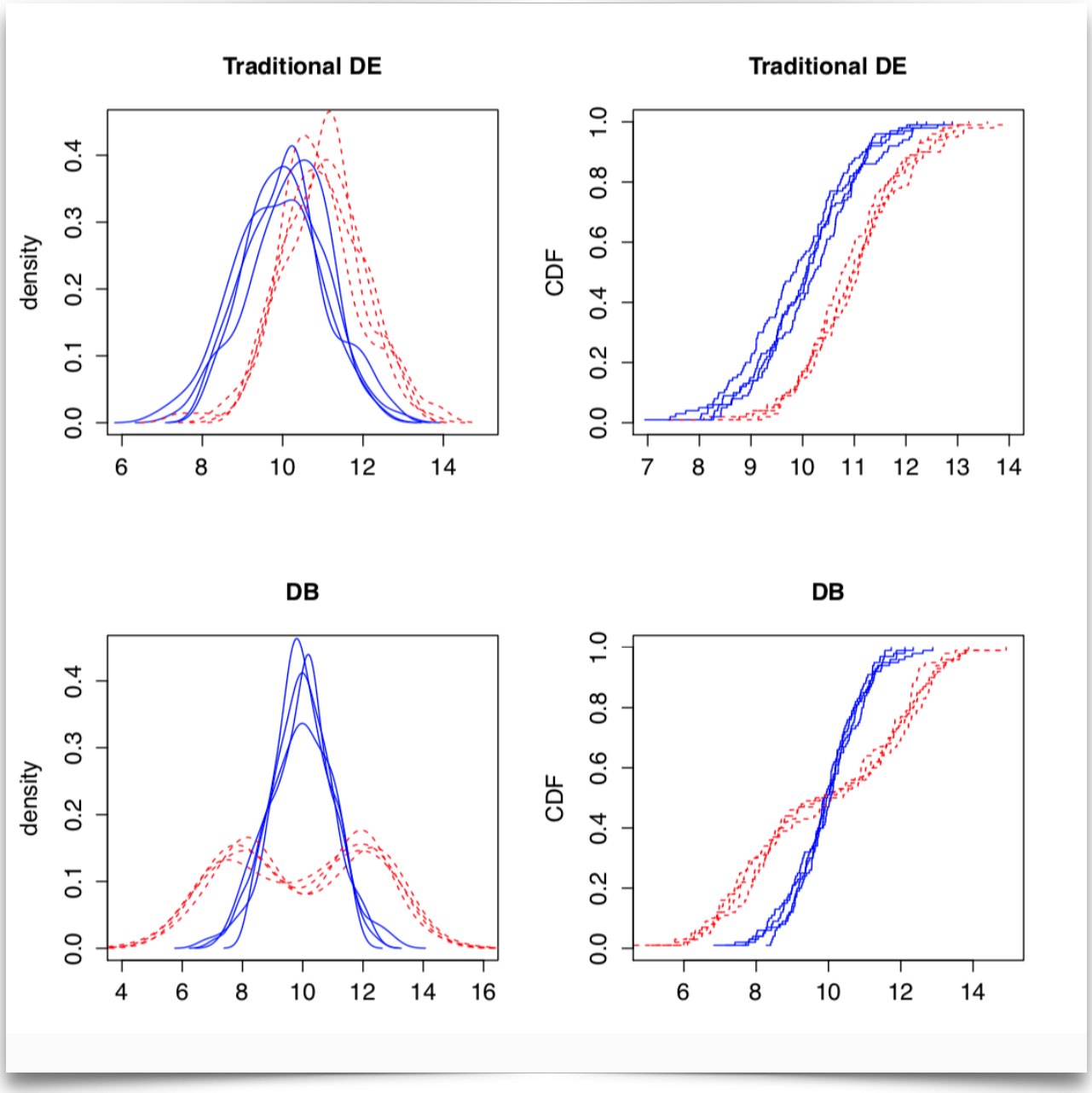
Overcoming confounding plate effects in differential expression analyses of single-cell RNA-seq data

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."

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

- k-sample Anderson-Darling test (Scholz and Stephens, 1987). Null distribution: all distributions are the same.
- functional data analysis?
- aggregate over cells?
- differential variability?



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

Equivalent **E**xpression

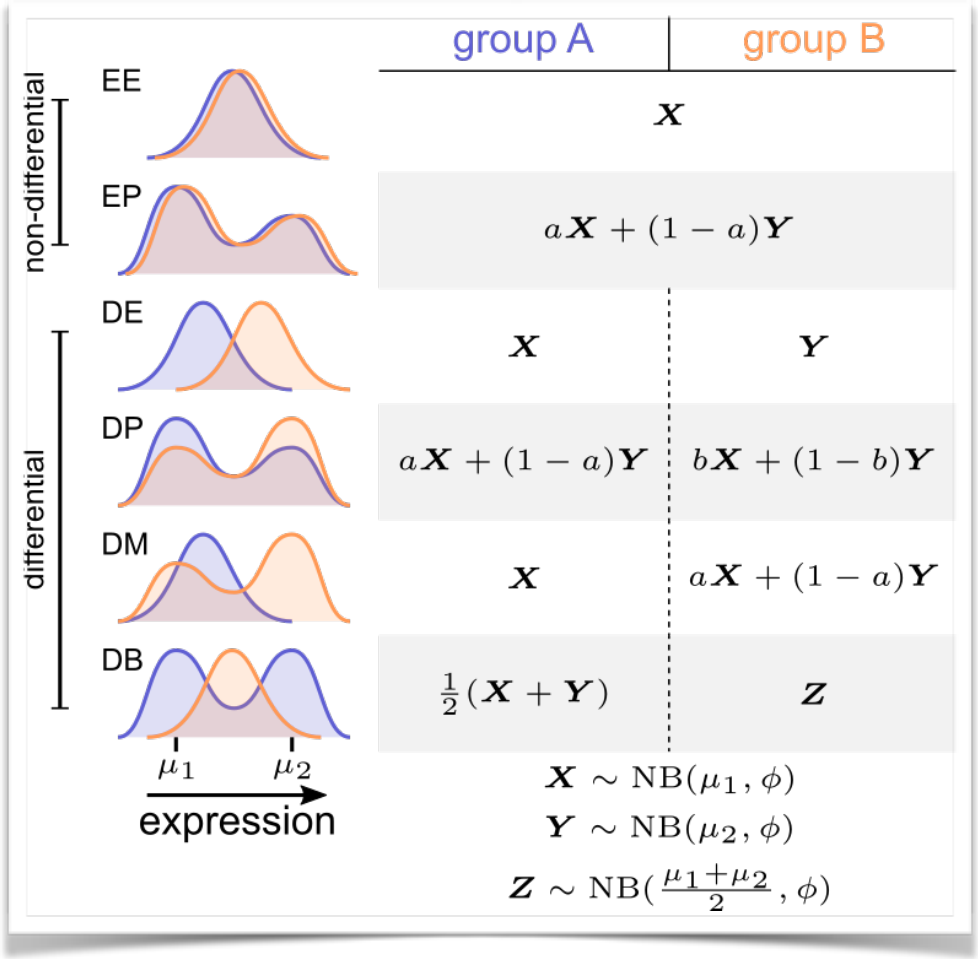
Equivalent **P**roportions

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

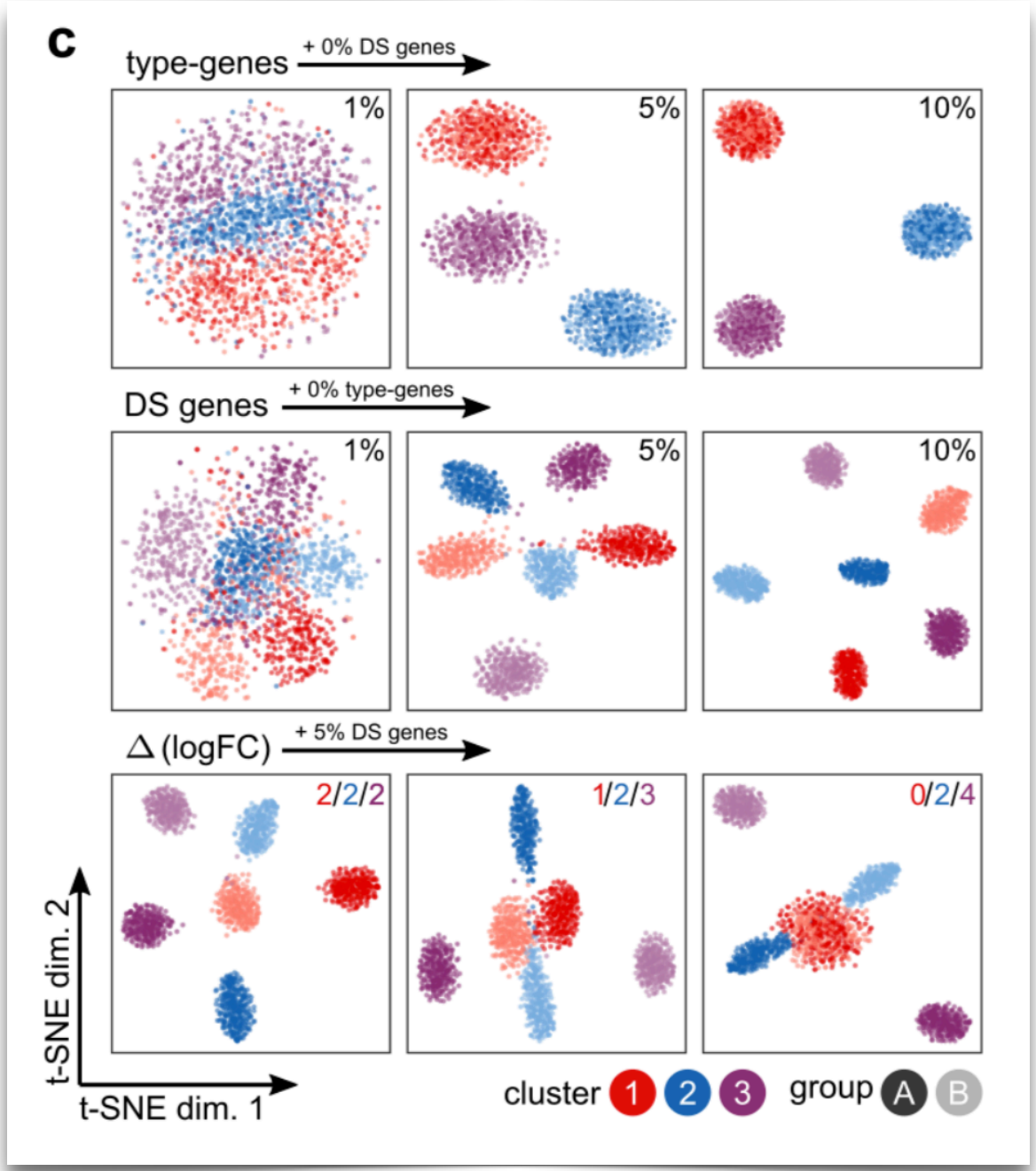
METHOD Open Access

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*]

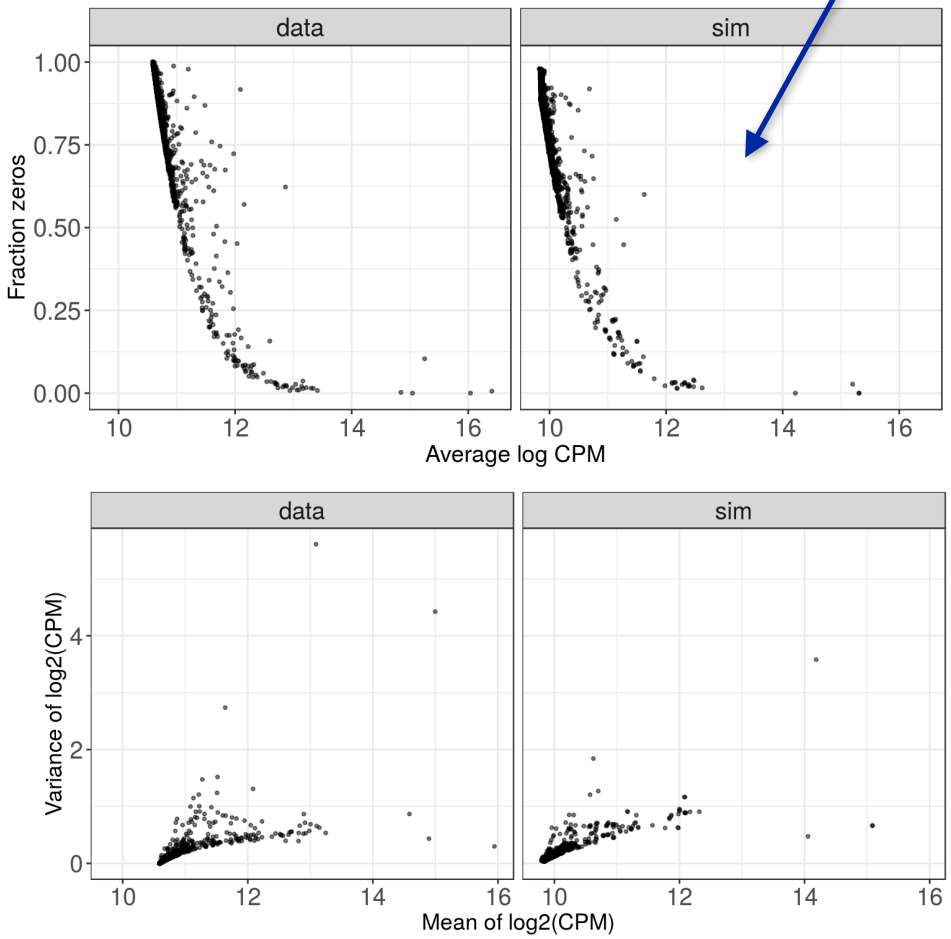
Flexibility of simulation

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

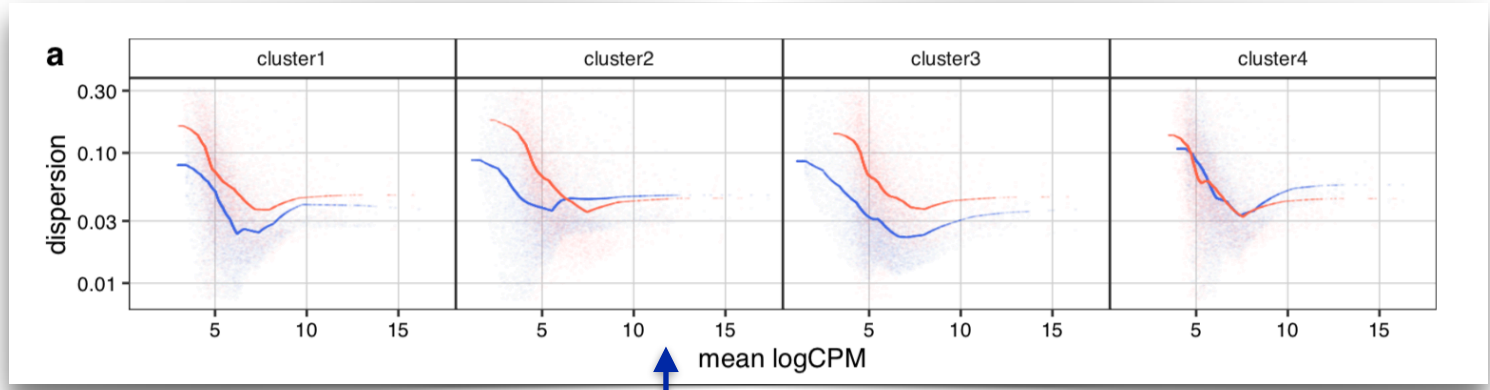


countsimQC: comparing simulated and real data

cell-level properties



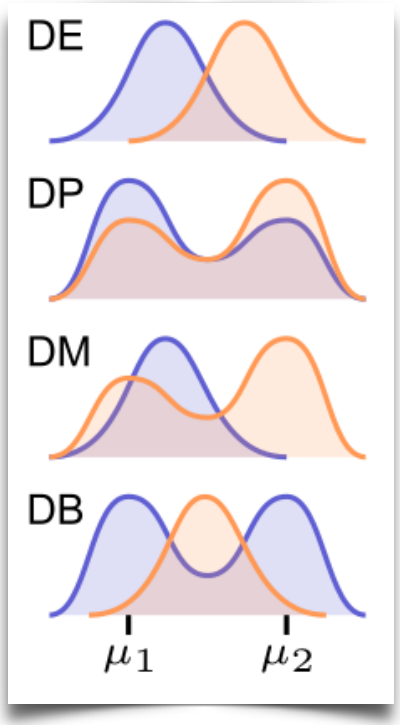
pseudobulk-level dispersion-mean relationships

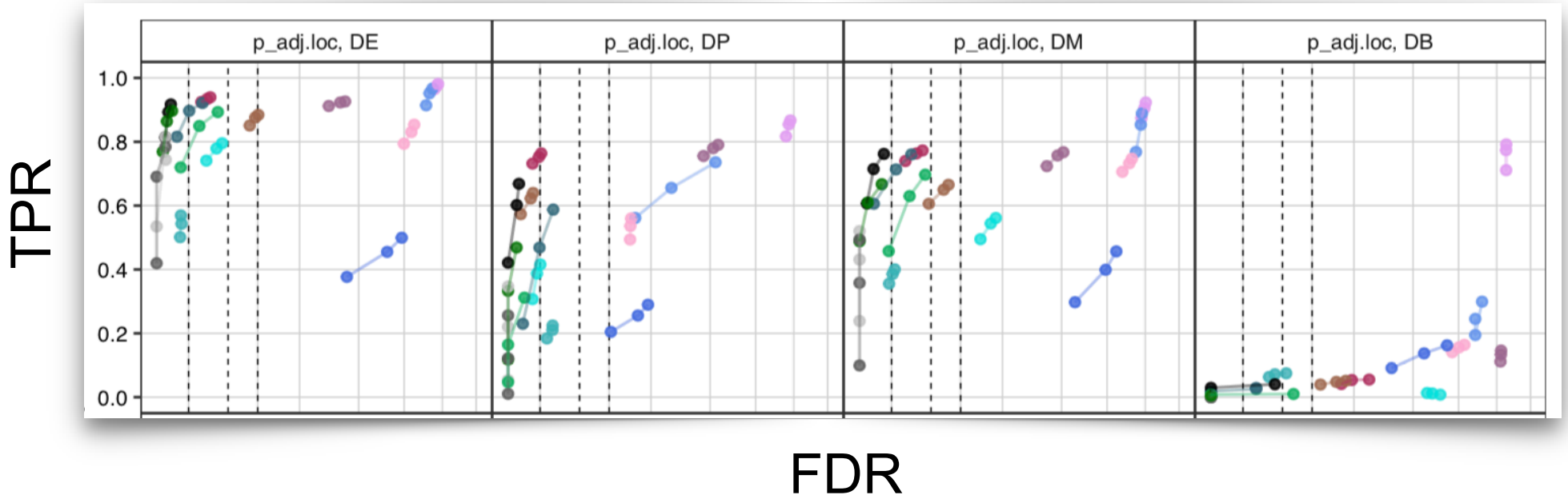


aggregate-level properties

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

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





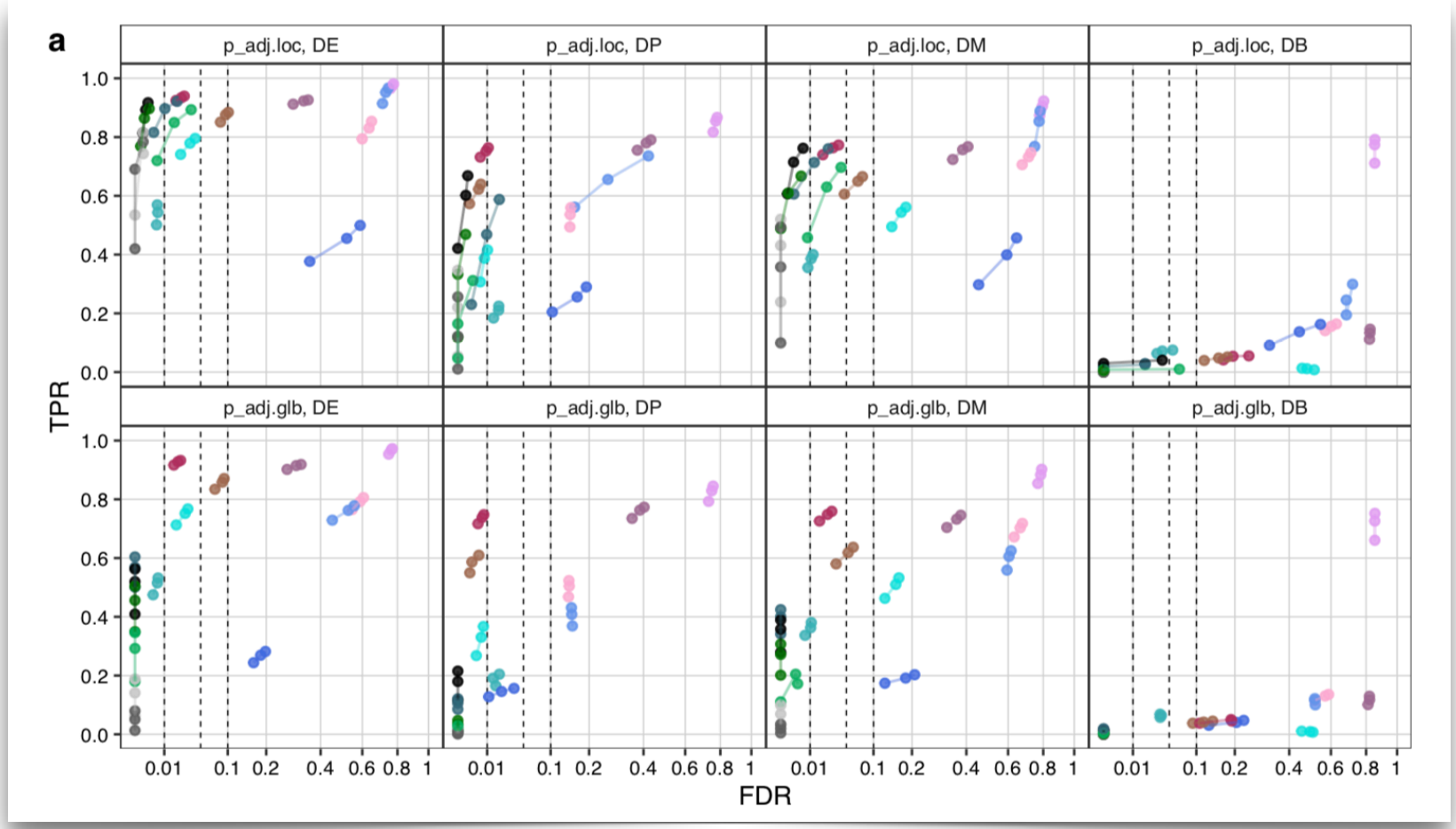
AD = Anderson-Darling
MM = mixed models

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

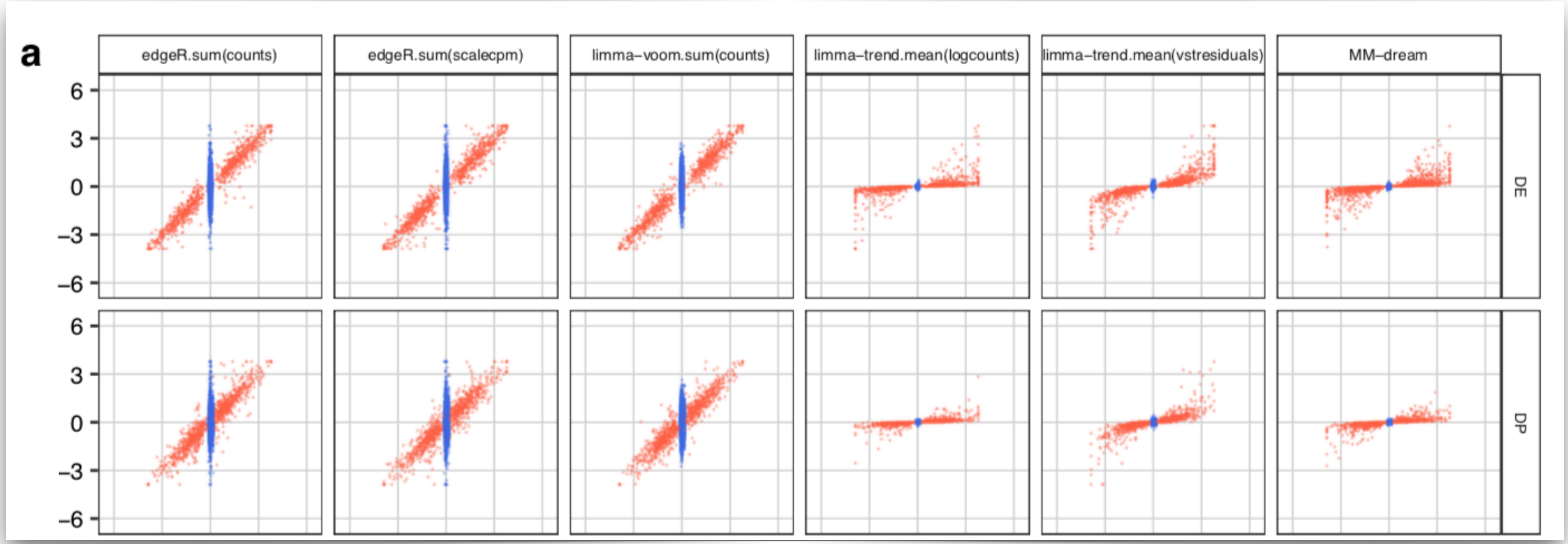
Multiple testing correction should be done locally (separate for each cluster/population)

local

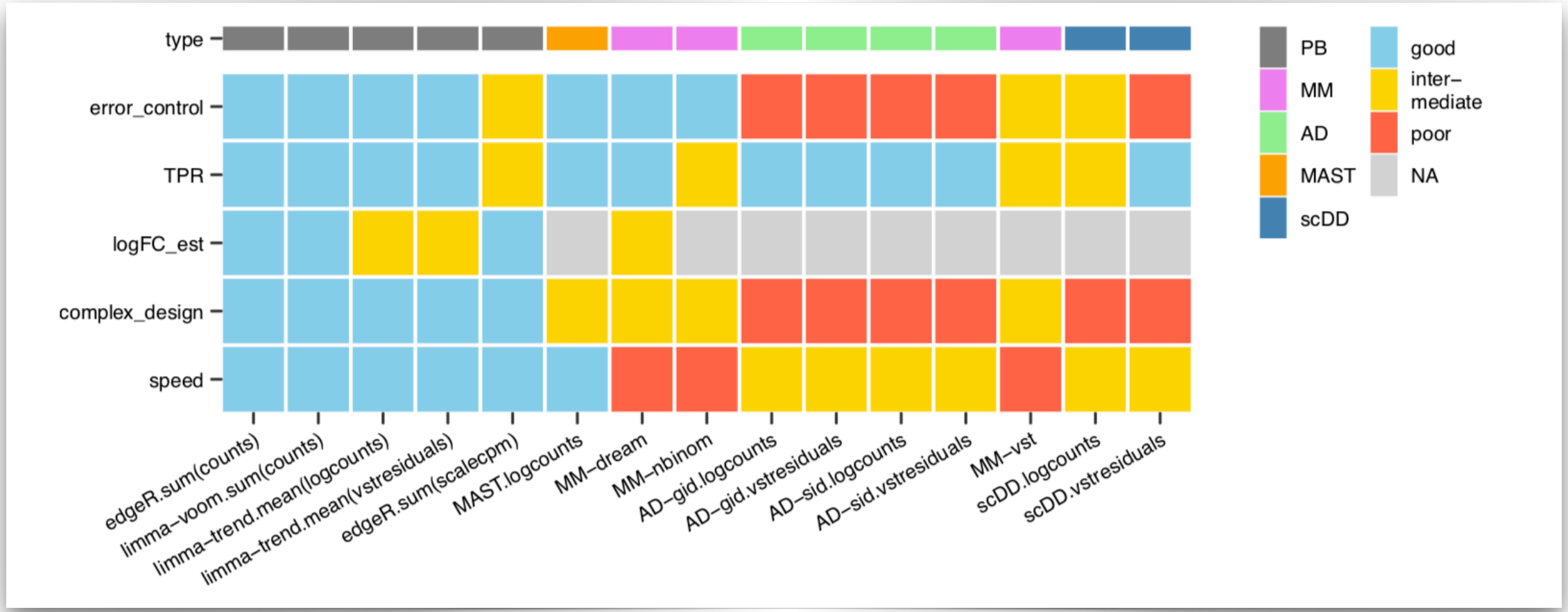
global



Pick your data to model wisely



Current rating



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

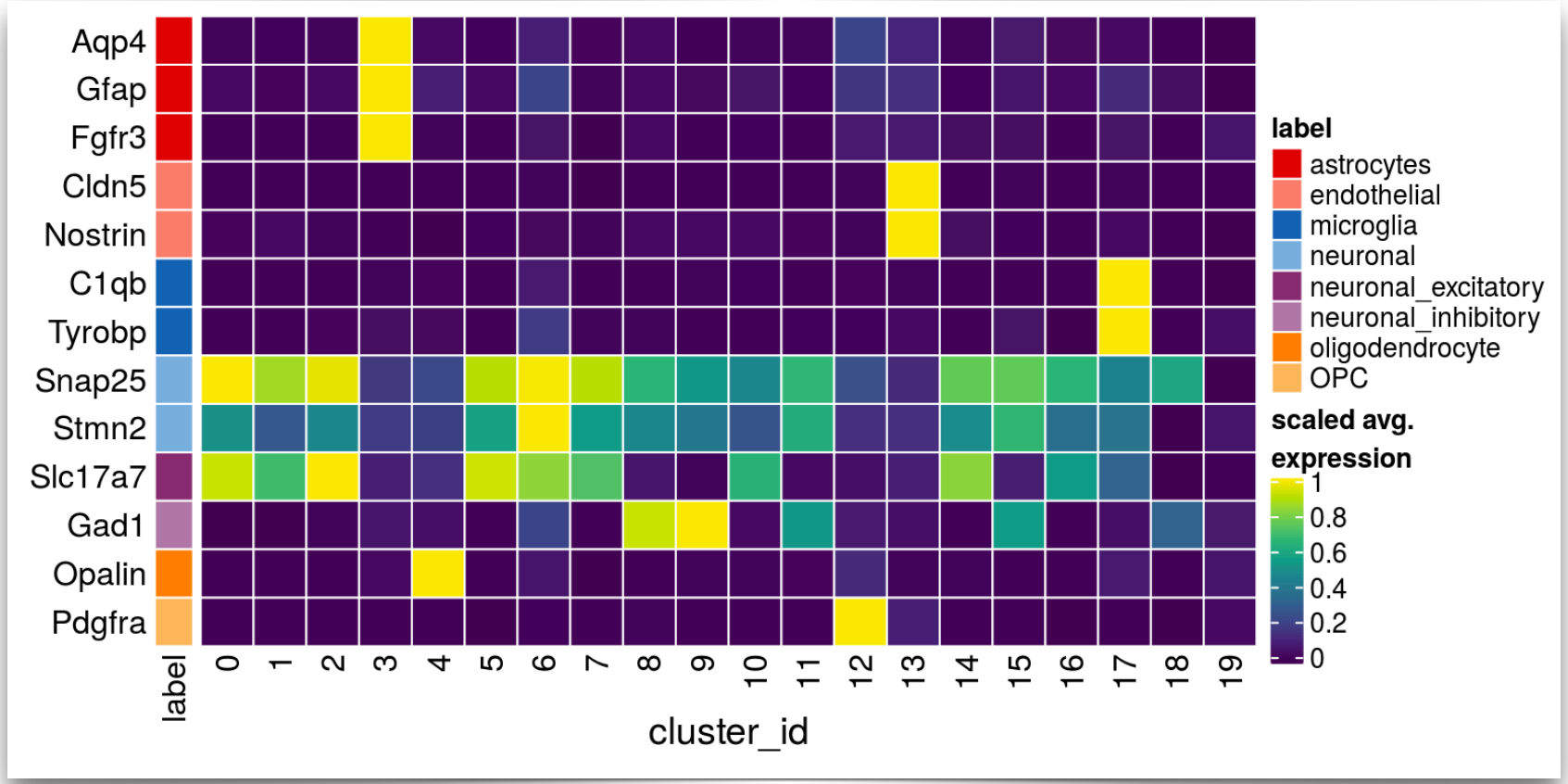
Application to LPS dataset: clustering + annotation subpopulations

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

frontal cortex

single nuclei RNA-seq (10x)

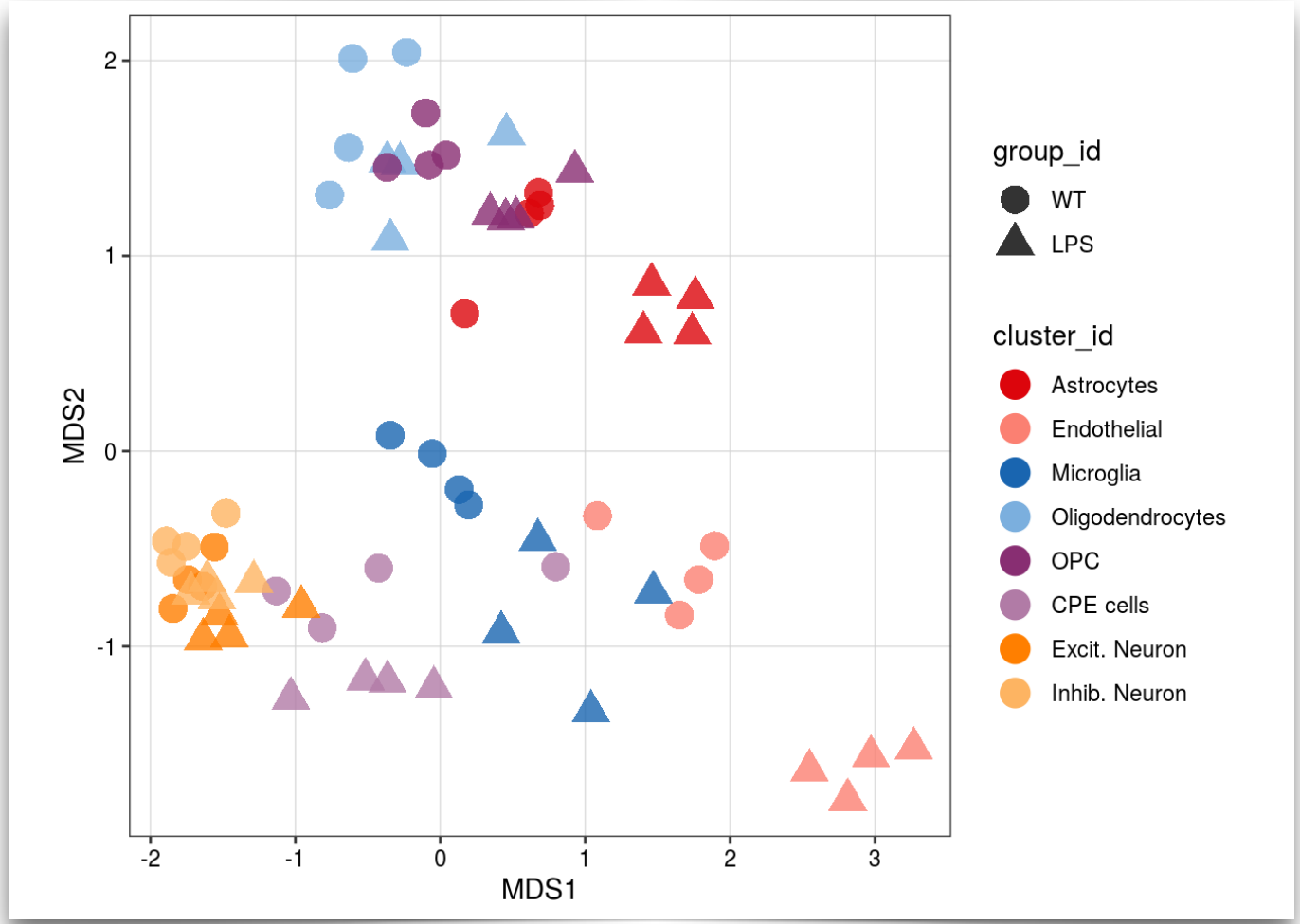
usual preprocessing:
filtering, doublet removal,
Seurat integration,
clustering



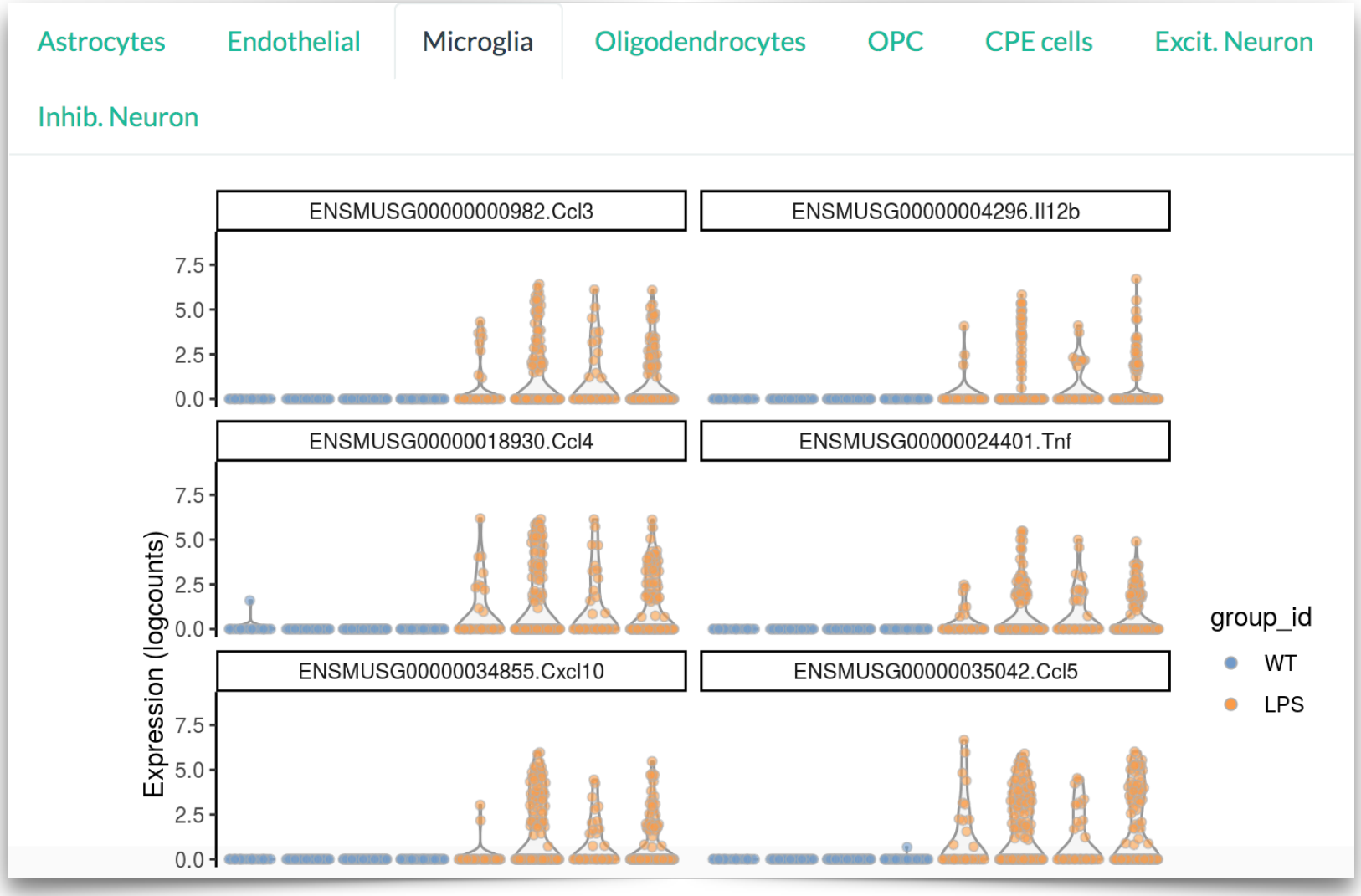
Application to LPS dataset: subpopulation-level visualization

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

Each dot is one subpopulation/ sample combination

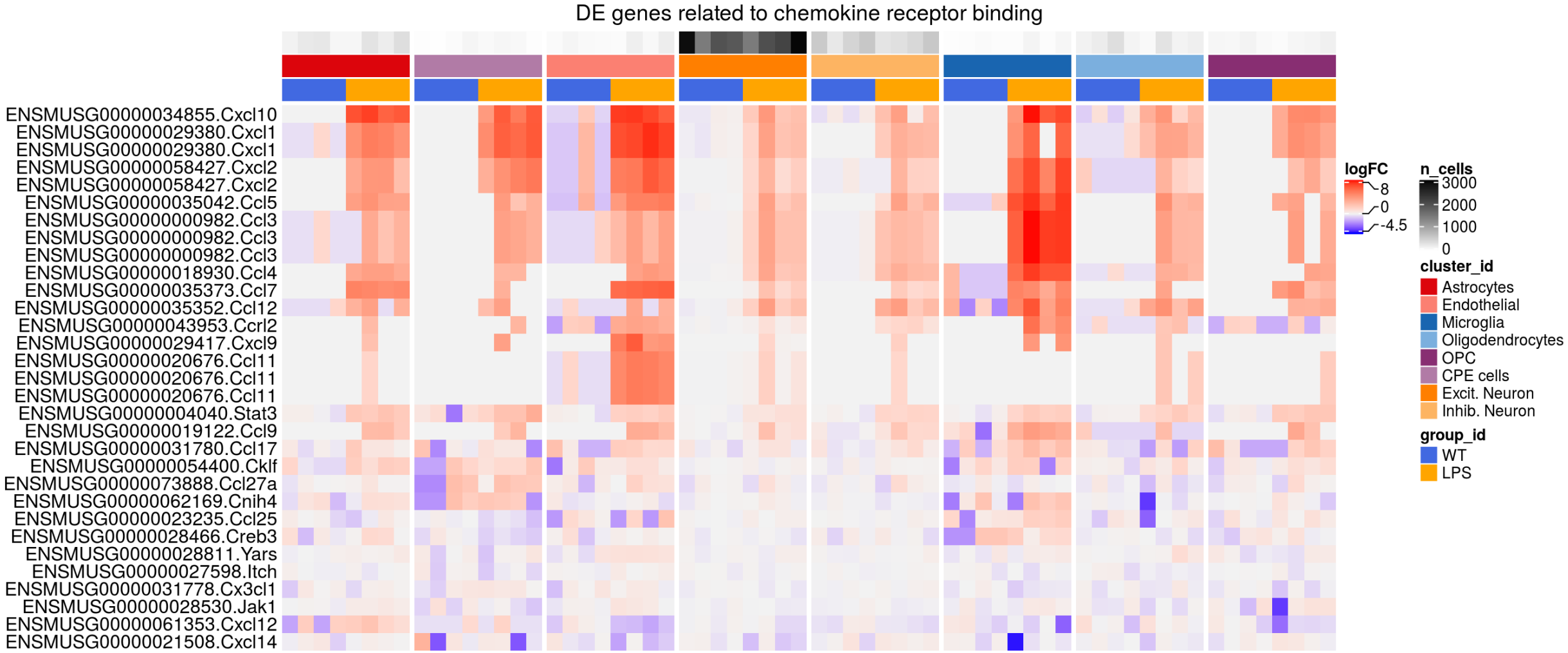


Application to LPS dataset: go back to cell-level response (discovery based on pseudobulk)



workflowr !

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



LPS dataset: interplay of cell type and cell state

There is still a lot to say

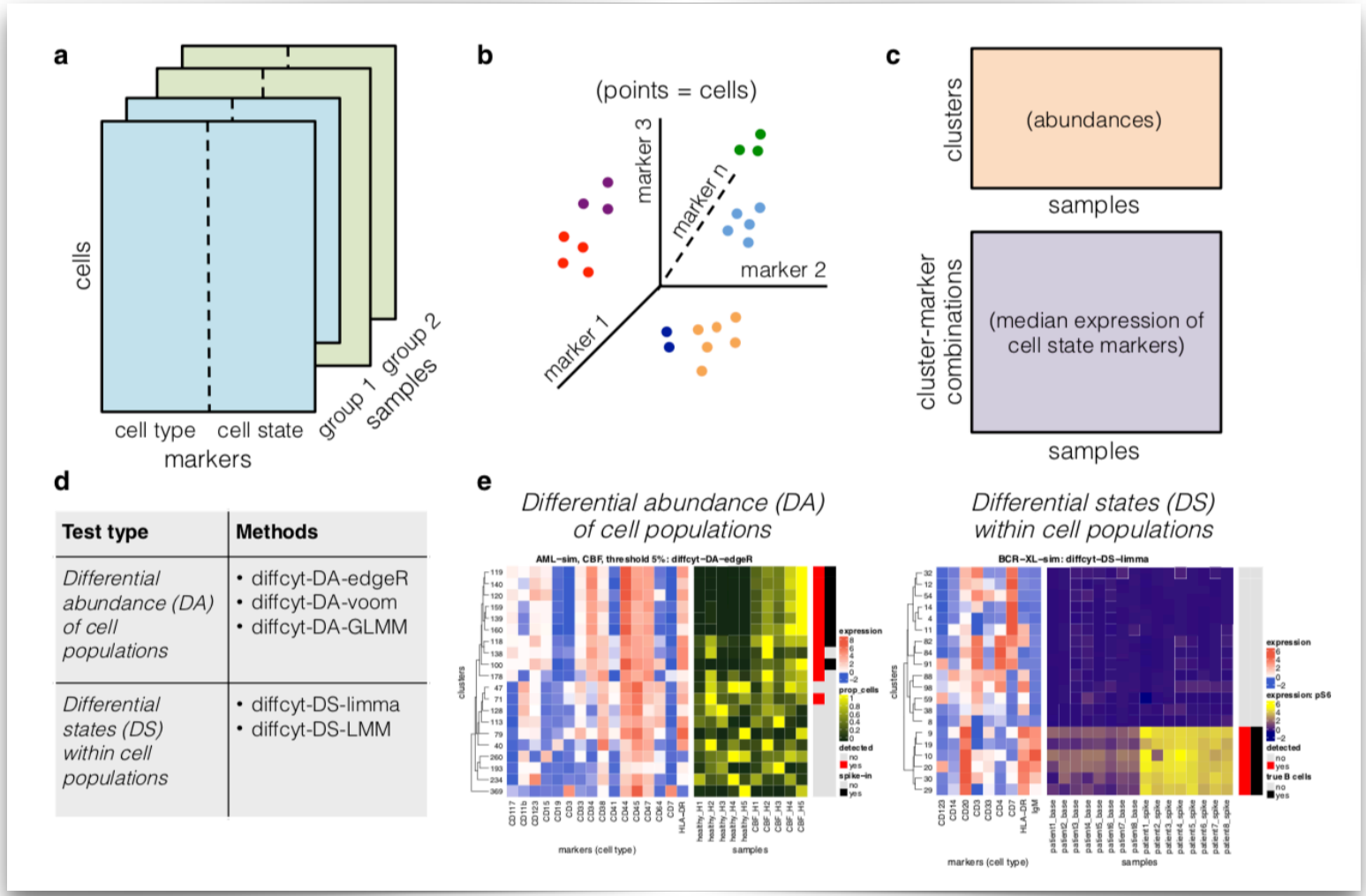
- multi-sample multi-condition multi-subpopulation datasets —> *in silico* sorting + **differential state analysis**
- Aggregation (e.g., pseudobulk counts) works well, is fast, flexible and modular
- software: https://github.com/HelenaLC/muscat

- Are we getting deep enough (per cell, per subpopulation)? —> Power differs by cell type
- Interplay between definition of type and state
- Can we get everything from aggregates?
- 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?

Many parallels to CyTOF data analysis



Gosia Lukas Helena



Test type	Methods
Differential abundance (DA) of cell populations	• diffcyt-DA-edgeR • diffcyt-DA-voom • diffcyt-DA-GLMM
Differential states (DS) within cell populations	• diffcyt-DS-limma • diffcyt-DS-LMM

ARTICLE

https://doi.org/10.1038/s42003-019-0415-5 OPEN

diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering

Lukas M. Weber[1,2], Malgorzata Nowicka[1,2,3], Charlotte Soneson[1,2,4] & Mark D. Robinson[1,2]

F1000Research F1000Research 2019, 6:748 Last updated: 24 MAY 2019

Check for updates

METHOD ARTICLE

REVISED CyTOF workflow: differential discovery in high-throughput high-dimensional cytometry datasets [version 3; peer review: 2 approved]

Malgorzata Nowicka[1,2], Carsten Krieg[3], Helena L. Crowell[1,2], Lukas M. Weber[1,2], Felix J. Hartmann[3], Silvia Guglietta[4], Burkhard Becher[3], Mitchell P. Levesque[5], Mark D. Robinson[1,2]

Statistical Bioinformatics Group, IMLS, UZH

Charlotte

Roche

Dheeraj Maholtra
Daniela Calini

Helena

Angelo Duo
Simone Tiberi

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