Single Cell School, Leysin, 17 October 2019

Universität Zürich[UZH]

URPP Evolution in Action

SIB Swiss Institute of Bioinformatics

Differential abundance (DA) and differential state (DS) analysis of single cell cytometry data

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

@markrobinsonca

Many many slides from:

Helena

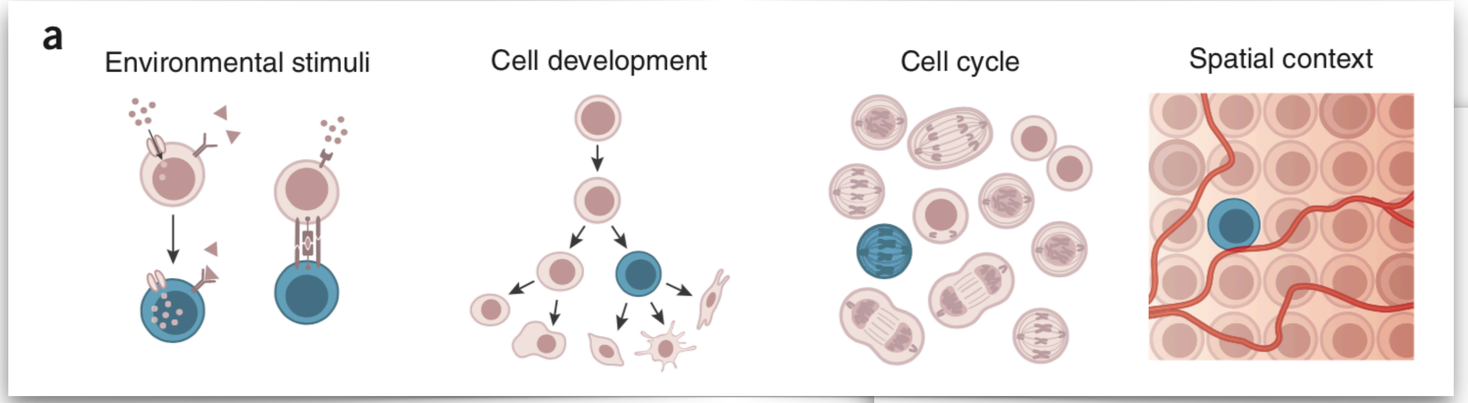
Lukas

Gosia

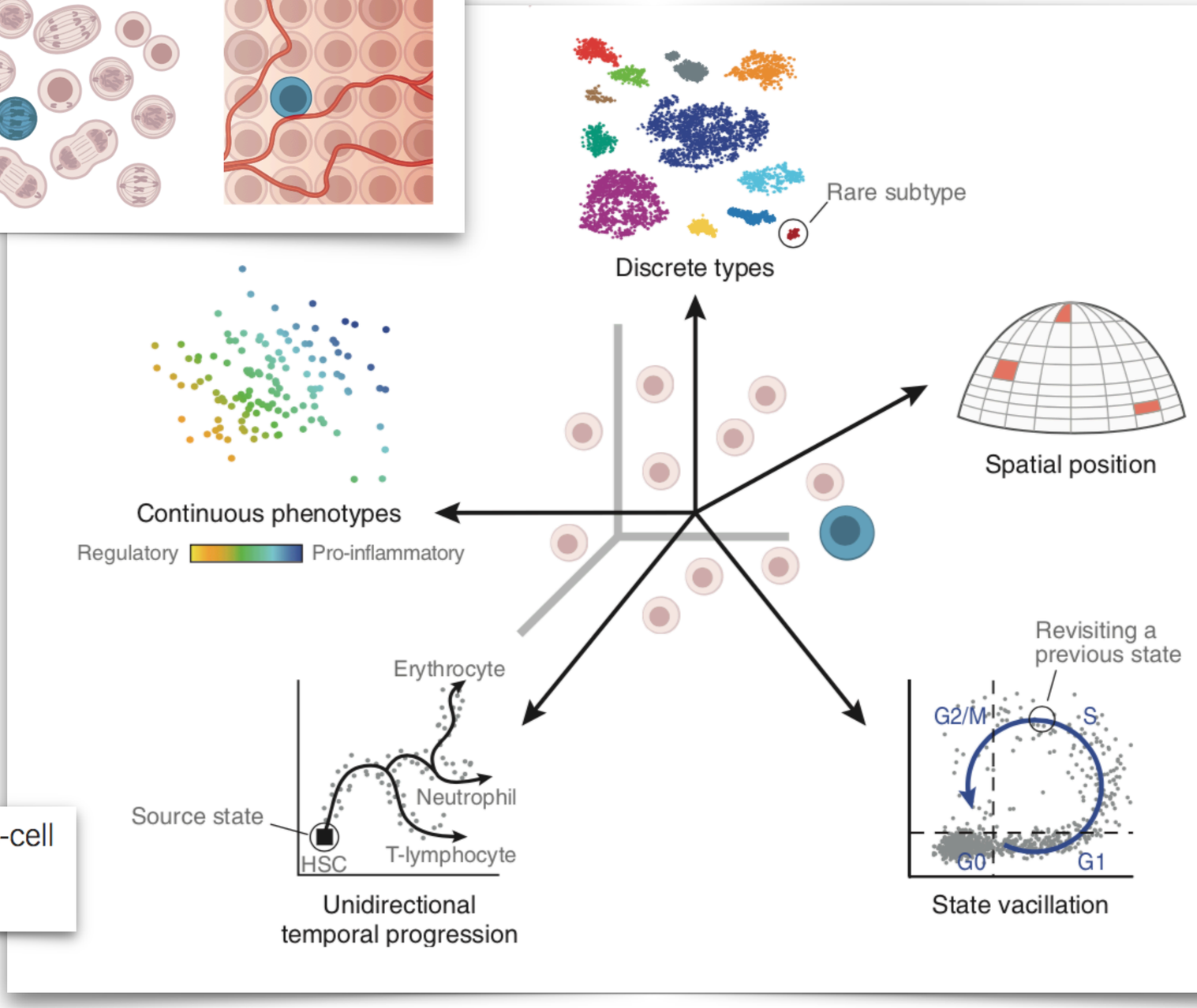
Charlotte

Outline

- Preliminaries
- **compensation** for CyTOF using single stains
- **clustering** “high” dimensional CyTOF
- **Differential analyses (diffcyt)**: abundance of populations, state transitions
- Parallels to single cell RNA-seq
- Working with tree representations of cells



Applications

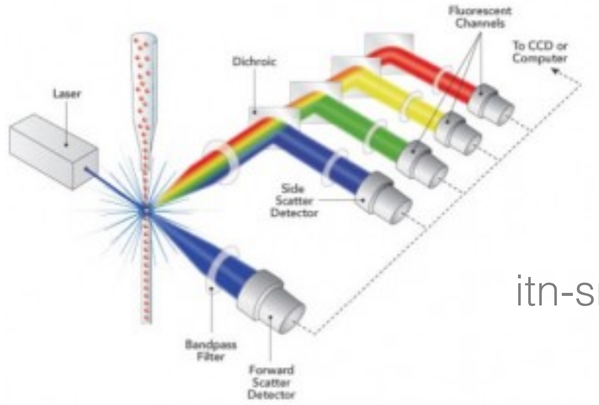


Revealing the vectors of cellular identity with single-cell genomics

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

High-dimensional cytometry

Measure **targeted protein expression levels** in single cells using **antibodies**



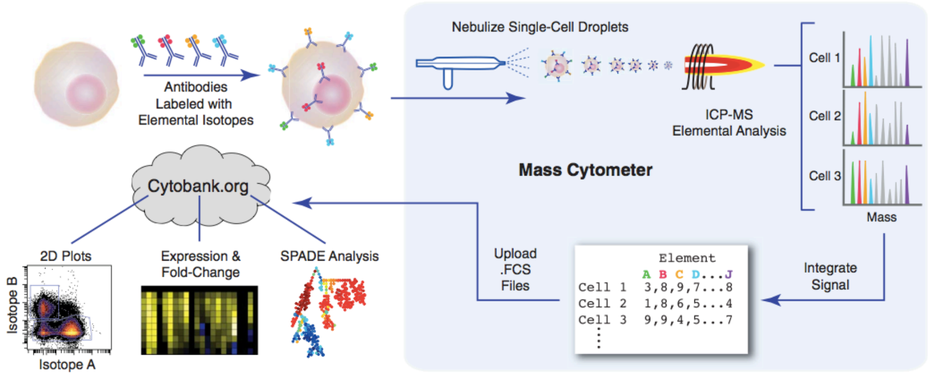
itn-snal.net

(A) Fluorescent flow cytometry / FACS

>20 proteins/cell; 1000s cells/sec

(B) Mass cytometry / CyTOF

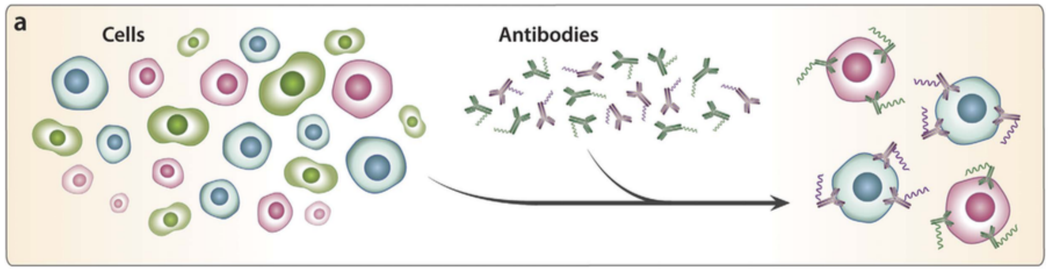
>40 proteins/cell; 100s cells/sec



Bendall et al. (2011), Fig. 1A

(C) sequence-based cytometry

>100 proteins/cell



Shahi et al. (2017), Fig. 1A

Revealing the vectors of cellular identity with single-cell genomics

Allon Wagner[1], 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

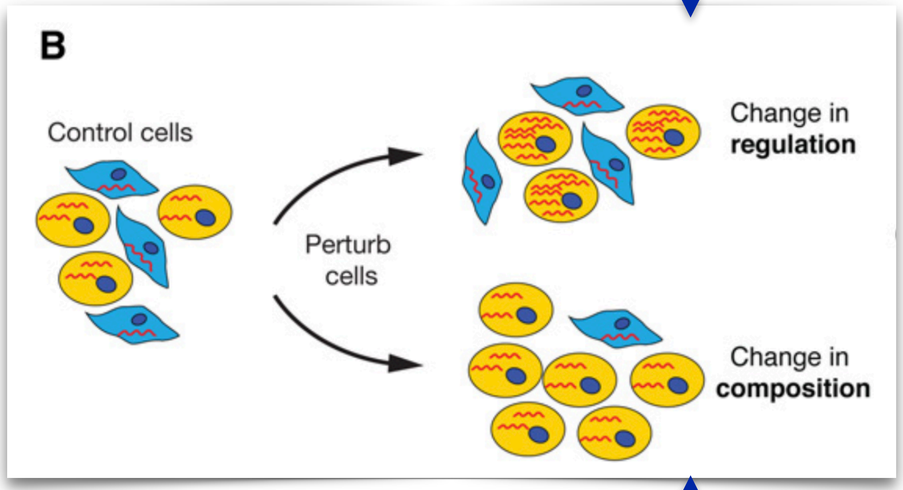
Box 1 The many facets of a cell's identity

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

Differential state analysis



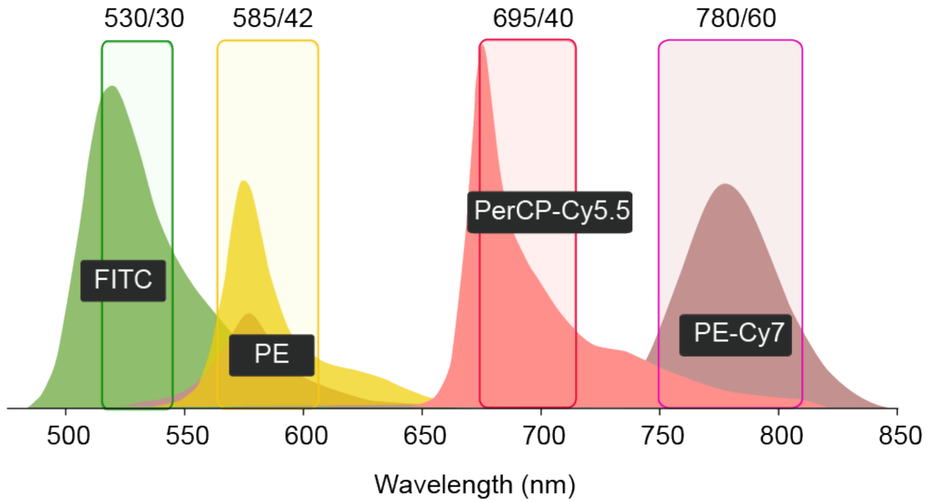
Differential abundance analysis

Spectral overlap vs. spillover

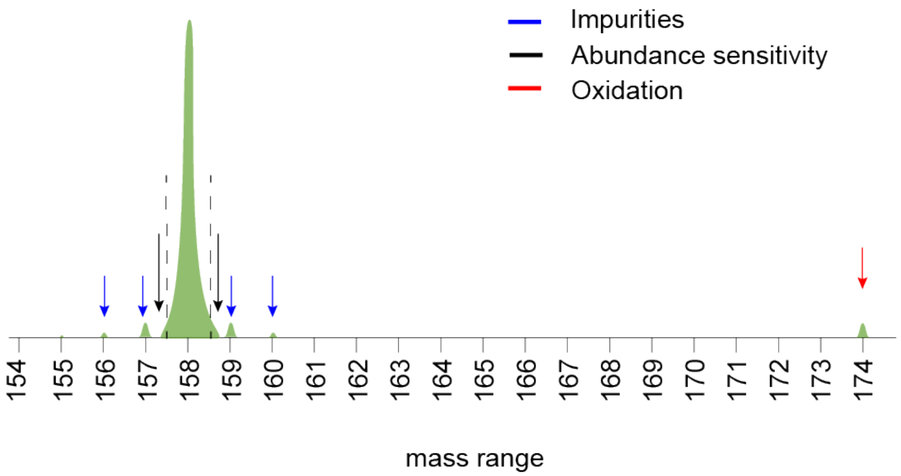
- CyTOF = increase in the number of parameters + massive decrease in spectral overlap
- but, still <u>three sources of signal overlap</u>:

1. **abundance sensitivity**:= (M±1) / M
2. **oxide formation**: +16M
3. **isotopic impurities**: up to ±6M

FACS



Mass cytometry



Do we need to compensate CyTOF data?

The ability to multiplex up to 40 cellular subset markers in mass cytometry, without a requirement for compensation for overlap in fluorescence signals as needed in conventional flow cytometry, makes mass cytometry an ideal technology to deeply phenotype cells in complex cell populations. This feature was elegantly demonstrated by

Atkuri et al. 2015 Drug Metabolism and Disposition

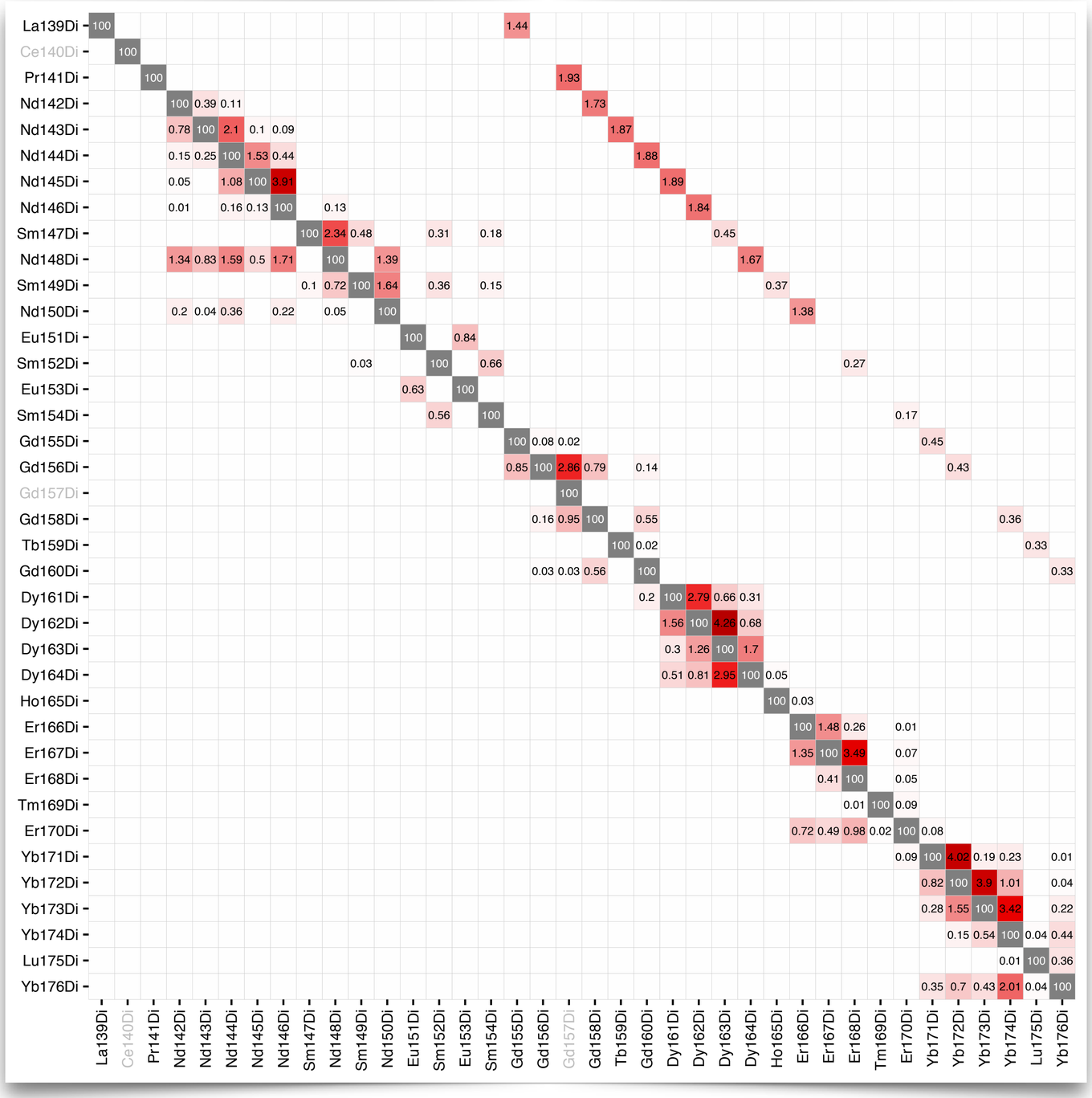
The metals that are sold as part of antibody labeling kits are of very high purity (98% and higher in most cases). As a practical matter, this means that "compensation" analogous to fluorescent antibodies is not needed, as most of the signal will be of the specified mass, with little to no signal at "M+1" or another contaminating mass. However, metal salts from other commercial sources may be of lesser purity. For example, the

Leipold al. 2015 Immunosenescence: Methods and Protocols, Methods in Molecular Biology

It should be made clear, though, that "minimal spillover" does not equal "no spillover."

Leipold Cytometry Part A, 2015

Single stain beads highlight sparsity of spillover



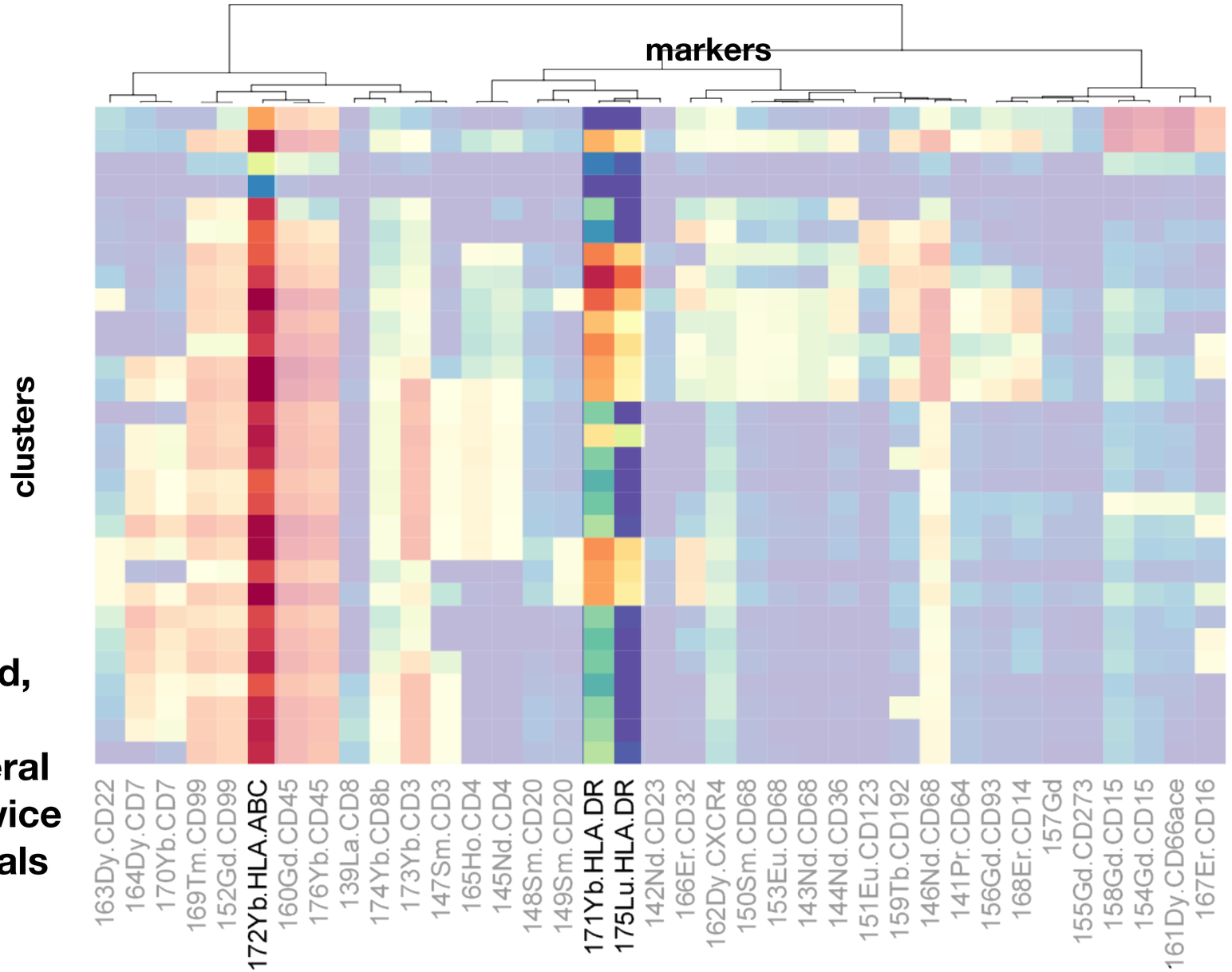
Compensation of Signal Spillover in Suspension and Imaging Mass Cytometry

Stéphane Chevrier,[1,4] Helena L. Crowell,[1,2,4] Vito R.T. Zanotelli,[1,3,4] Stefanie Engler,[1] Mark D. Robinson,[1,2,*] and Bernd Bodenmiller[1,5,*]

[1]Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland
[2]SIB Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland
[3]Systems Biology Ph.D. Program, Life Science Zürich Graduate School, ETH Zürich and University of Zürich, Zürich, Switzerland
[4]These authors contributed equally
[5]Lead Contact
*Correspondence: mark.robinson@imls.uzh.ch (M.D.R.), bernd.bodenmiller@imls.uzh.ch (B.B.)

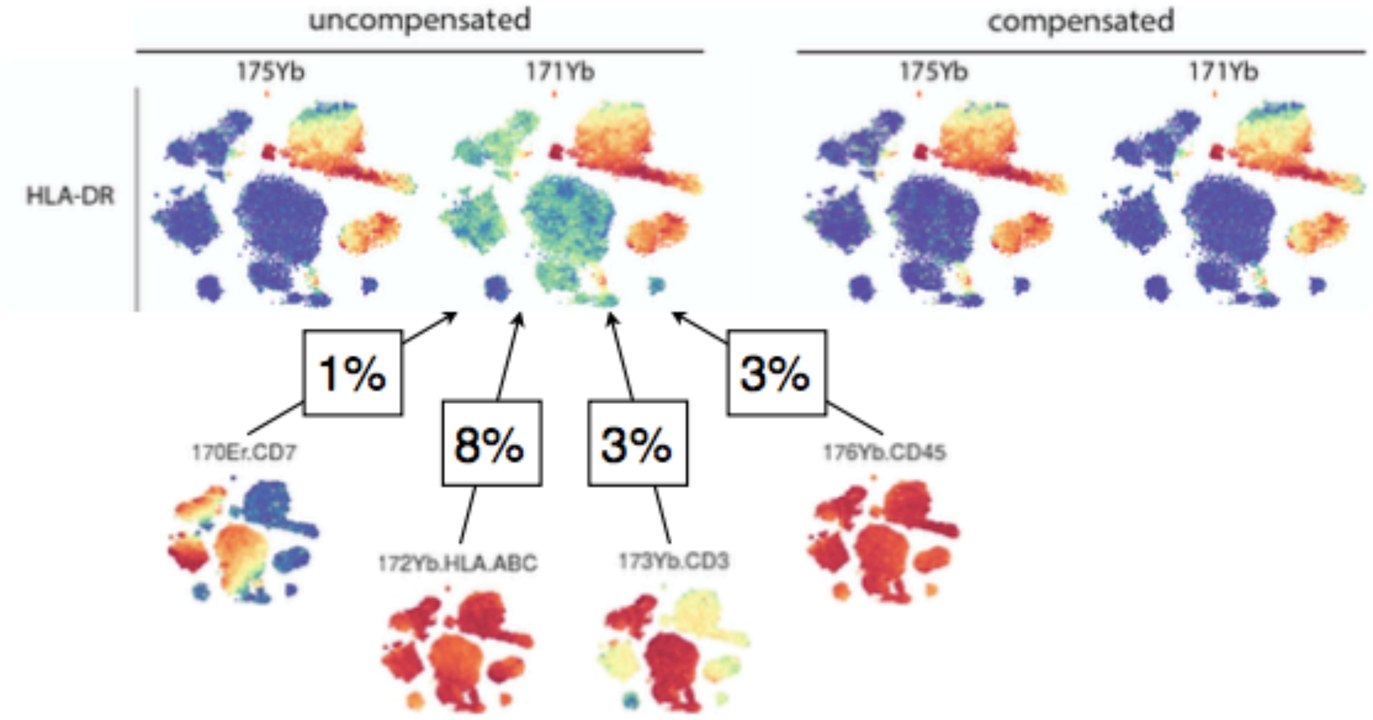
https://doi.org/10.1016/j.cels.2018.02.010

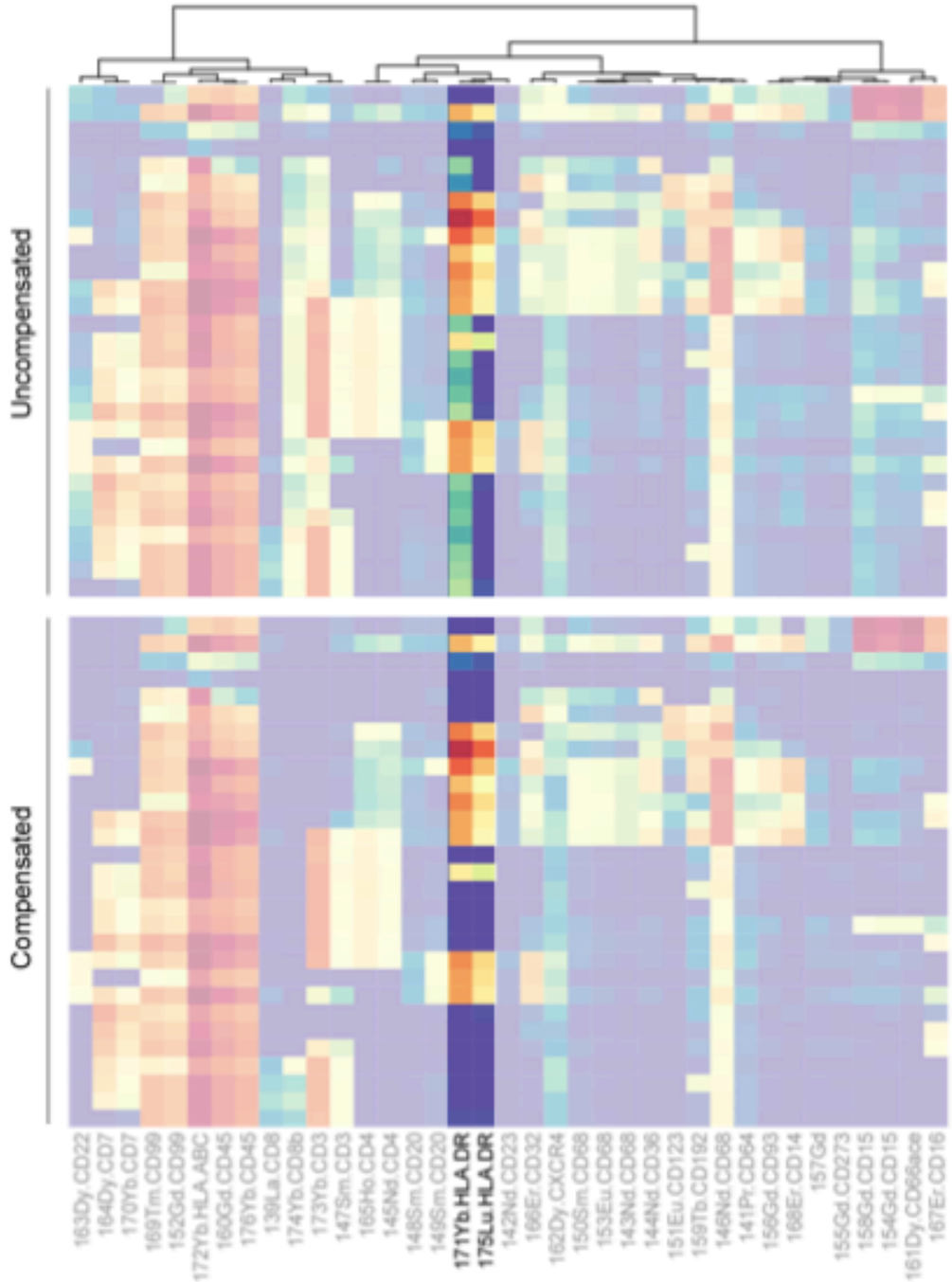
Short answer: Yes, even ~2-5% of a high signal can be significant



PBMCs measured, clustered with Phenograph; several antibodies used twice with different metals

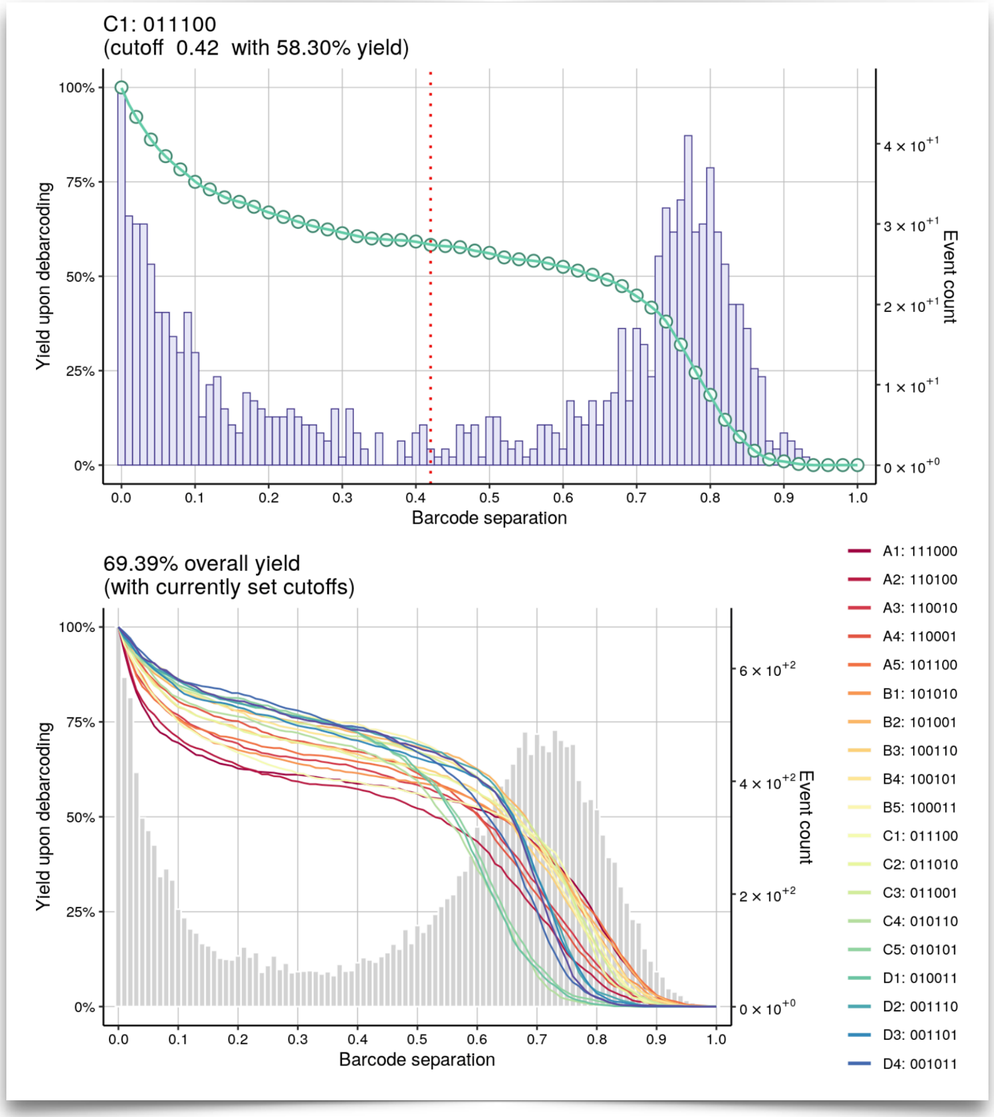
Correction of spillover artefacts on a 36ab panel





Spillover matrix estimated via single-stain beads:
non-negative least squares

R/Bioconductor CATALYST package: preprocessing (differential analysis comes later)



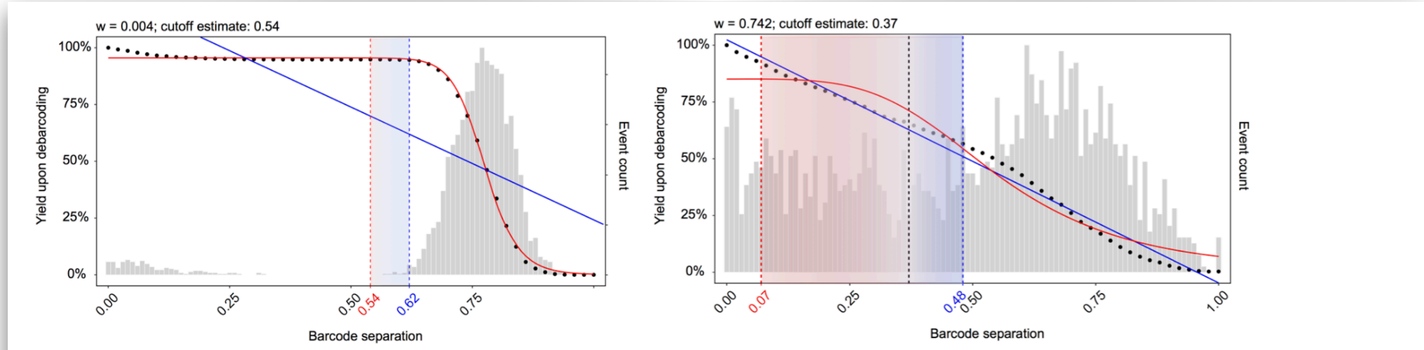
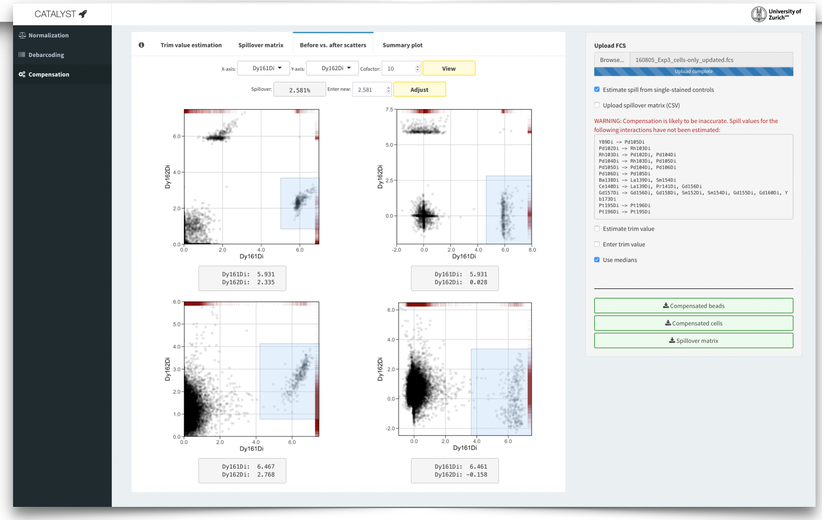


Figure 1: **Description of the automatic cutoff estimation for each individual population**
The bar graphs indicate the distribution of cells relative to the barcode distance and the dotted line corresponds to the yield upon debarcoding as a function of the applied separation cutoff. This curve is fitted with a linear regression (blue line) and a three parameter log-logistic function (red line). The cutoff estimate is defined as the mean of estimates derived from both fits, weighted with the goodness of the respective fit.

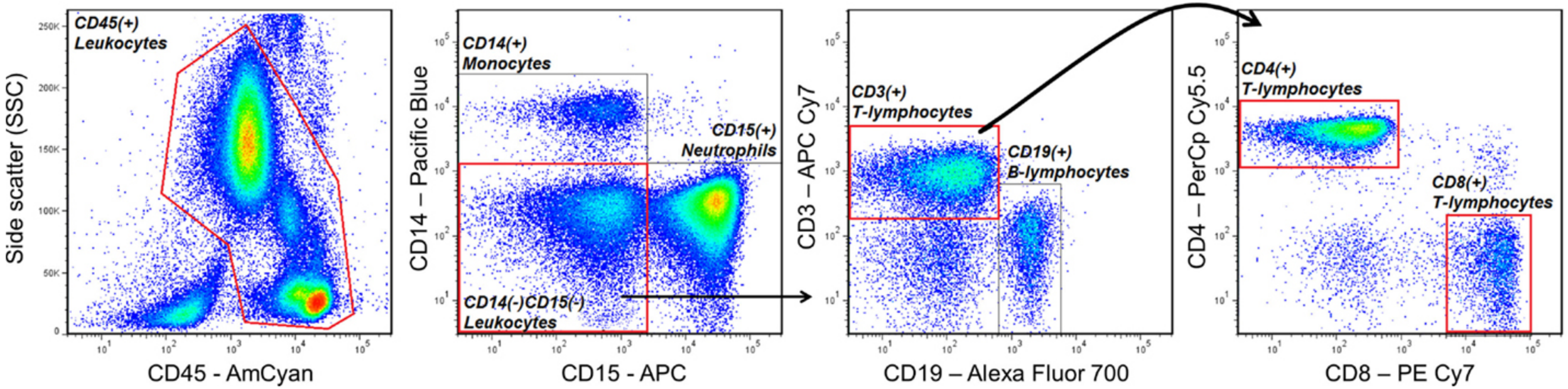


Helena

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

Manual gating

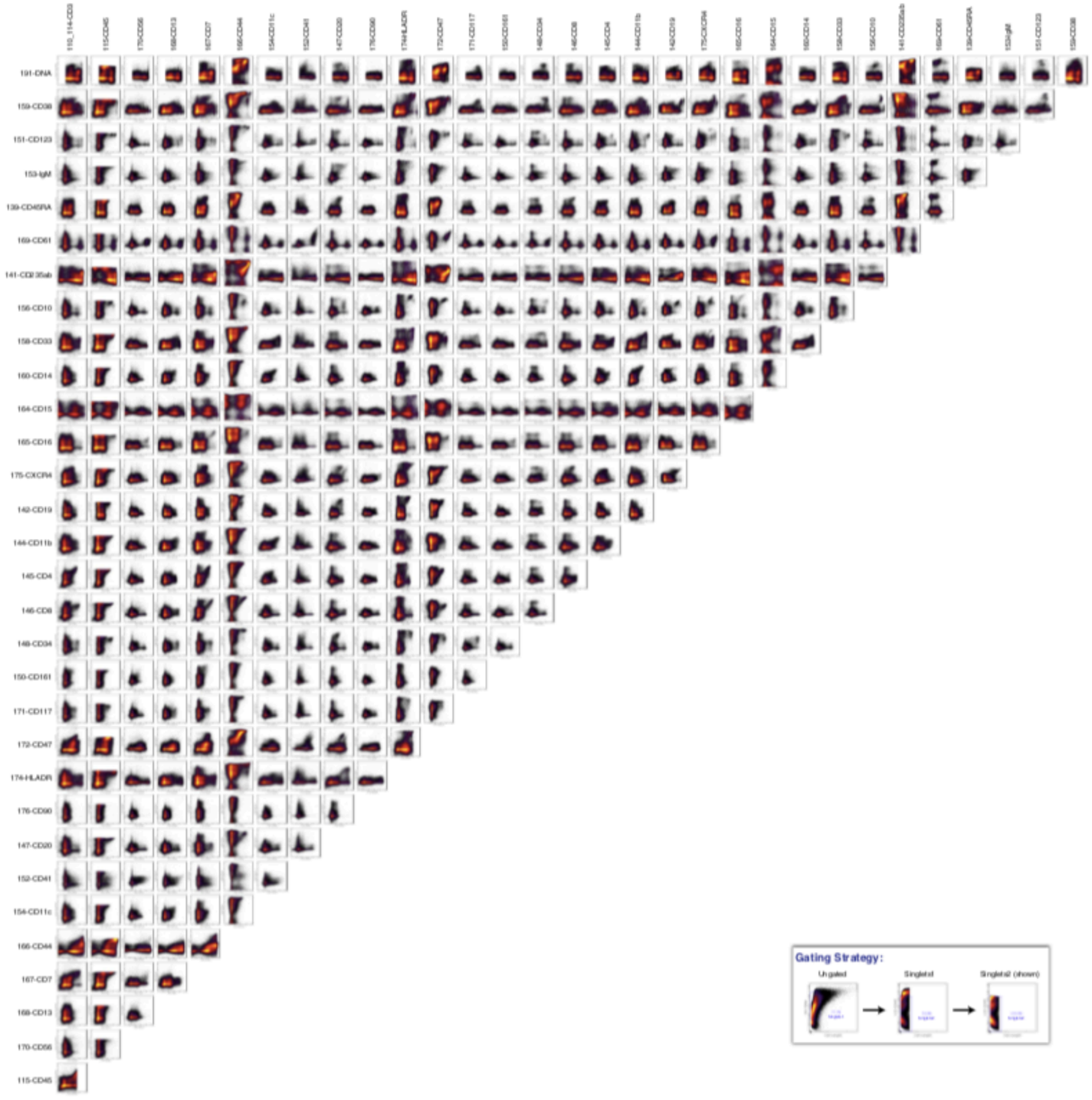
Identification of cell populations



Verschoor et al. (2015)

Manual gating

Not feasible in high-dimensional data



Bendall et al. (2011), Supp.

Comparison of clustering algorithms

Algorithms for low-dimensional data: **FlowCAP consortium**

Critical assessment of automated flow cytometry data analysis techniques

Nima Aghaeepour[1], Greg Finak[2], The FlowCAP Consortium[3], The DREAM Consortium[3], Holger Hoos[4], Tim R Mosmann[5], Ryan Brinkman[1,7], Raphael Gottardo[2,7] & Richard H Scheuermann[6,7]

Clustering high-dimensional flow and mass cytometry

Motivation: Many new computational methods, explosion in the number of dimensions (both FACS and CyTOF) — what works “best”?

Lukas

EDITOR'S CHOICE

Cytometry PART A
Journal of the International Society for Advancement of Cytometry

Comparison of Clustering Methods for High-Dimensional Single-Cell Flow and Mass Cytometry Data

Lukas M. Weber,[1,2] Mark D. Robinson[1,2]*

Laundry list of methods

Table 1. Overview of clustering methods compared in this study

METHOD	ENVIRONMENT AND AVAILABILITY	SHORT DESCRIPTION	REF.
ACCENSE	Standalone application with graphical interface	Nonlinear dimensionality reduction (t-SNE) followed by density-based peak-finding and clustering in two-dimensional projected space.	22
ClusterX	R package (cytofkit) from Bioconductor	Density-based clustering on t-SNE projection map; faster than DensVM.	23
DensVM	R package (cytofkit) from Bioconductor	Density-based clustering on t-SNE projection map; similar to ACCENSE, with additional support vector machine to classify uncertain points.	24
FLOCK	C source code (also available in ImmPort online platform)	Partitioning of each dimension into bins, followed by merging of dense regions, and density-based clustering.	25
flowClust	R package from Bioconductor	Model-based clustering based on multivariate *t* mixture models with Box-Cox transformation.	26
flowMeans	R package from Bioconductor	Based on k-means, with merging of clusters to allow non-spherical clusters.	27
flowMerge	R package from Bioconductor	Extension of flowClust; merges cluster mixture components from flowClust.	28
flowPeaks	R package from Bioconductor	Peak-finding on smoothed density function generated by k-means; using finite mixture model.	29
FlowSOM	R package from Bioconductor	Self-organizing maps, followed by hierarchical consensus meta-clustering to merge clusters.	30
FlowSOM_pre	R package from Bioconductor	Same as FlowSOM, but without the final consensus meta-clustering step.	30
immunoClust	R package from Bioconductor	Iterative clustering based on finite mixture models, using expectation maximization and integrated classification likelihood.	31
k-means	R base packages (stats)	Standard k-means clustering.	
PhenoGraph	Graphical interface (cyt) launched from MATLAB (Python implementation also available)	Construction of nearest-neighbor graph, followed by partitioning of the graph into sets of highly interconnected points ("communities").	18
Rclusterpp	R package from GitHub (older version on CRAN)	Large-scale implementation of standard hierarchical clustering, with improved memory requirements.	32
SamSPECTRAL	R package from Bioconductor	Spectral clustering, with modifications for improved memory requirements.	33
SPADE	R package from GitHub (older version on Bioconductor; also available in Cytobank online platform)	"Spanning-tree progression analysis of density-normalized events"; organizes clusters into a branching hierarchy of related phenotypes.	34
SWIFT	Graphical interface launched from MATLAB	Iterative fitting of Gaussian mixture models by expectation maximization, followed by splitting and merging of clusters using a unimodality criterion.	35
X-shift	Standalone application (VorteX) with graphical interface (command-line version also available)	Weighted k-nearest-neighbor density estimation, detection of local density maxima, connection of points via graph, and cluster merging.	17

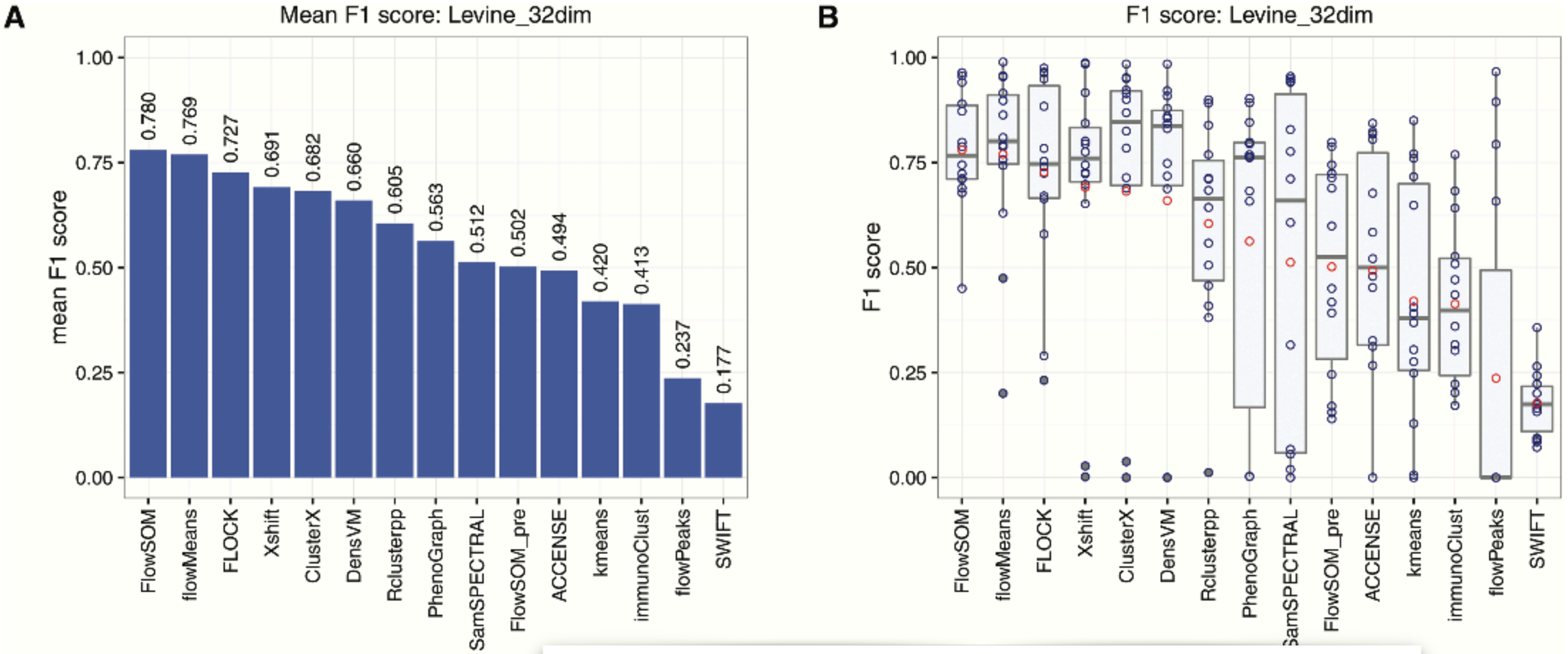
Manually gated populations

Manual gates = "truth"

Table 2. Summary of data sets used to evaluate clustering methods

DATA SET	CYTOF OR FLOW CYTOMETRY	CLUSTERING TASK	NO. OF CELLS	NO. OF DIMENSIONS	NO. OF MANUALLY GATED POPULATIONS OF INTEREST	NO. OF MANUALLY GATED CELLS	ORGANISM	NO. OF INDIVIDUALS (PATIENTS, MICE)	SAMPLE DESCRIPTION	REF.
Levine_32dim	CyTOF	Multiple populations	265,627	32 (surface markers)	14	104,184 (39%)	Human	2	Bone marrow cells from healthy donors	(18)
Levine_13dim	CyTOF	Multiple populations	167,044	13 (surface markers)	24	81,747 (49%)	Human	1	Bone marrow cells from healthy donor	(18)
Samusik_01	CyTOF	Multiple populations	86,864	39 (surface markers)	24	53,173 (61%)	Mouse	1	Replicate bone marrow samples from C57BL/6J mice (sample 01 only)	(17)
Samusik_all	CyTOF	Multiple populations	841,644	39 (surface markers)	24	514,386 (61%)	Mouse	10	Replicate bone marrow samples from C57BL/6J mice (all samples)	(17)
Nilsson_rare	Flow cytometry	Rare population	44,140	13 (surface markers)	1 (hematopoietic stem cells)	358 (0.8%)	Human	1	Bone marrow cells from healthy donor	(36)
Mosmann_rare	Flow cytometry	Rare population	396,460	14 (surface and intracellular)	1 (activated memory CD4 T cells)	109 (0.03%)	Human	1	Peripheral blood cells from healthy donor, stimulated with influenza antigens	(35)

Comparison of clustering methods



Hungarian algorithm to match clusters to populations

F1 score

From Wikipedia, the free encyclopedia

"F score" redirects here. For the significance test, see F-test.

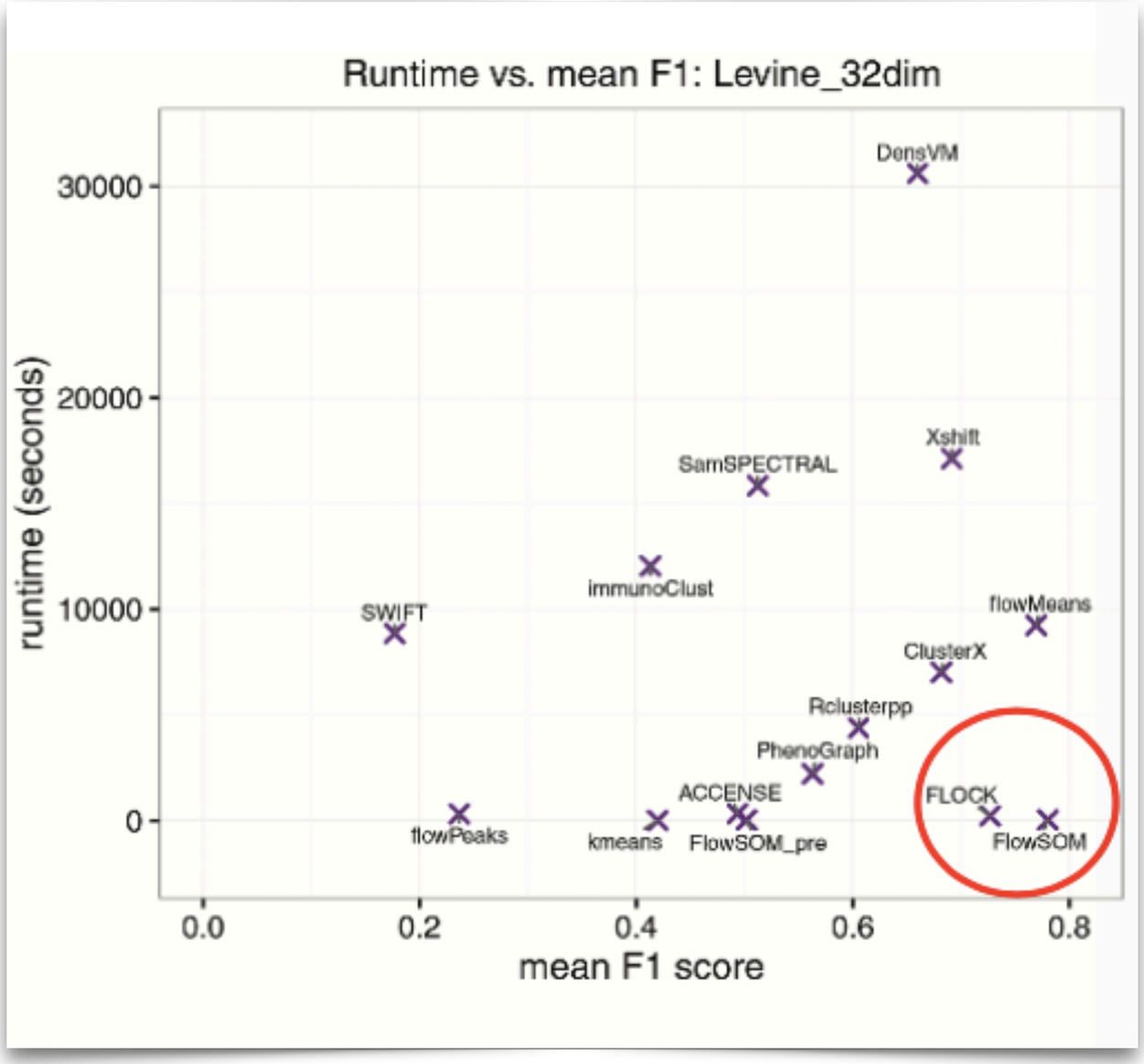
In statistical analysis of binary classification, the **F_1 score** (also **F-score** or **F-measure**) is a measure of a test's accuracy. It considers both the precision *p* and the recall *r* of the test to compute the score: *p* is the number of correct positive results divided by the number of all positive results, and *r* is the number of correct positive results divided by the number of positive results that should have been returned. The F_1 score can be interpreted as a weighted average of the precision and recall, where an F_1 score reaches its best value at 1 and worst at 0.

The traditional F-measure or balanced F-score (**F_1 score**) is the harmonic mean of precision and recall — multiplying the constant of 2 scales the score to 1 when both recall and precision are 1:

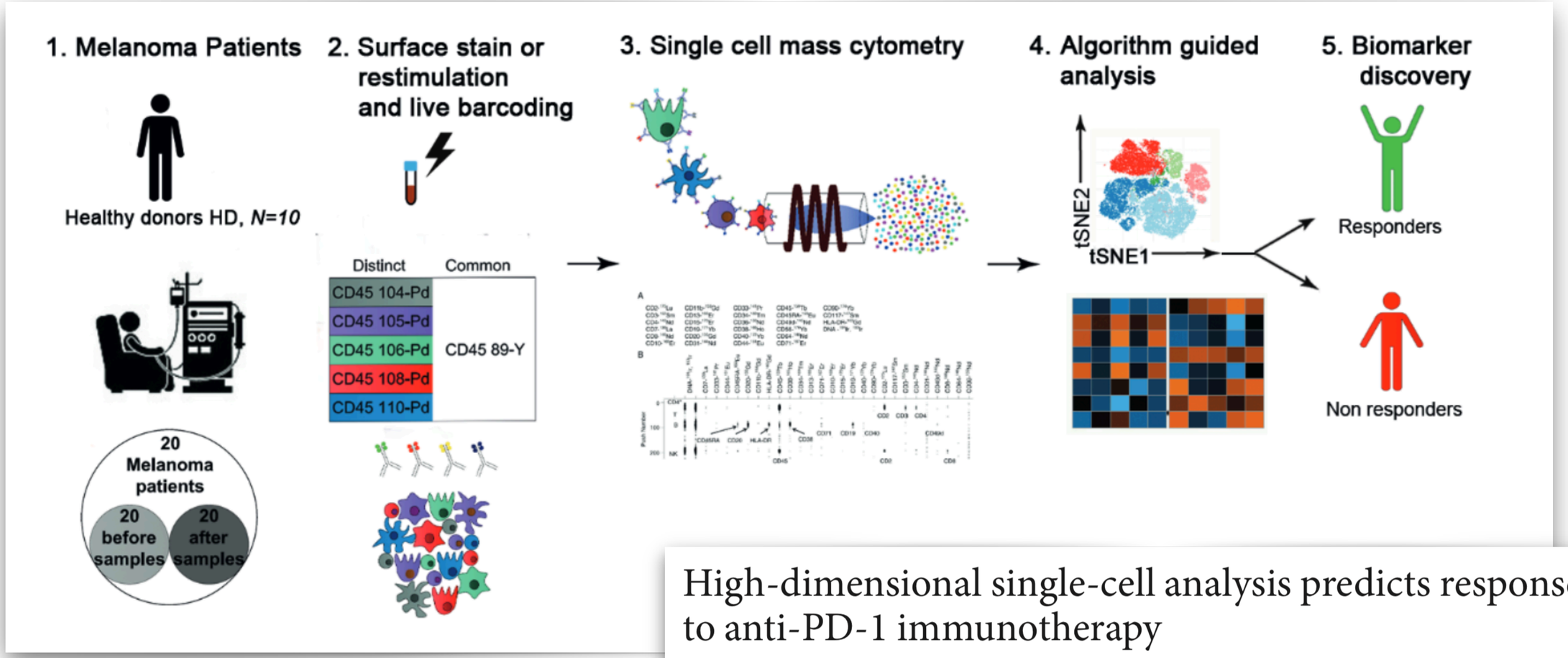
$$F_1 = 2 \cdot \frac{1}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

Comparison of clustering methods

- several methods performed well: *FlowSOM*, *X-shift*, *PhenoGraph*, *Rclusterpp*, *flowMeans*
- **FlowSOM** gave best performance (for several data sets) and was fast
- **X-shift** gave best performance for rare cell populations
- several methods sensitive to random starts (rare populations)
- code, data freely available



Motivation for differential analysis: finding cancer biomarkers



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

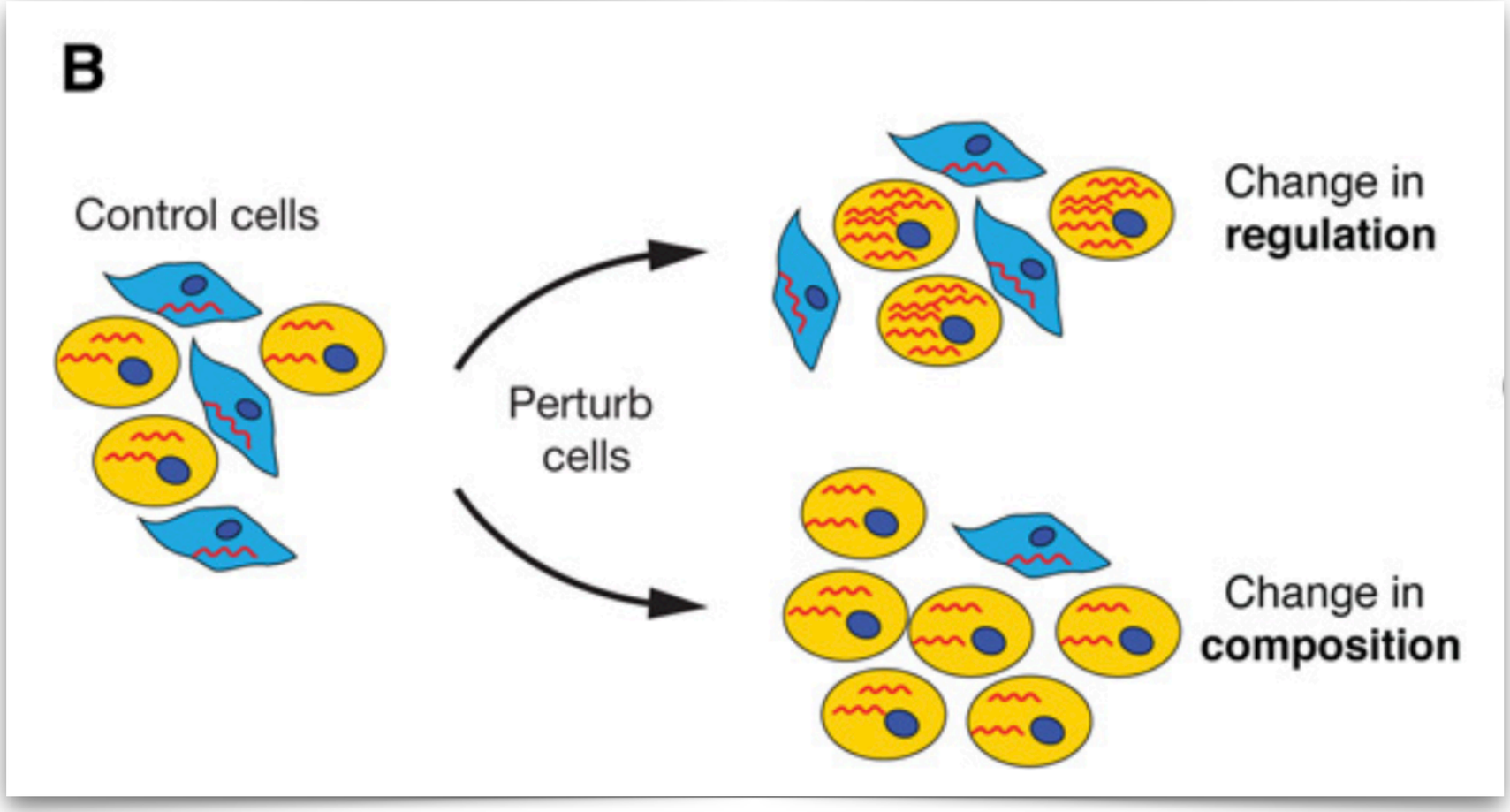
Carsten Krieg[1,6], Malgorzata Nowicka[2,3], Silvia Guglietta[4], Sabrina Schindler[5], Felix J Hartmann[1], Lukas M Weber[2,3], Reinhard Dummer[5], Mark D Robinson[2,3], Mitchell P Levesque[5,7] & Burkhard Becher[1,7]

Perspective

Defining cell types and states with single-cell genomics

Cole Trapnell

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



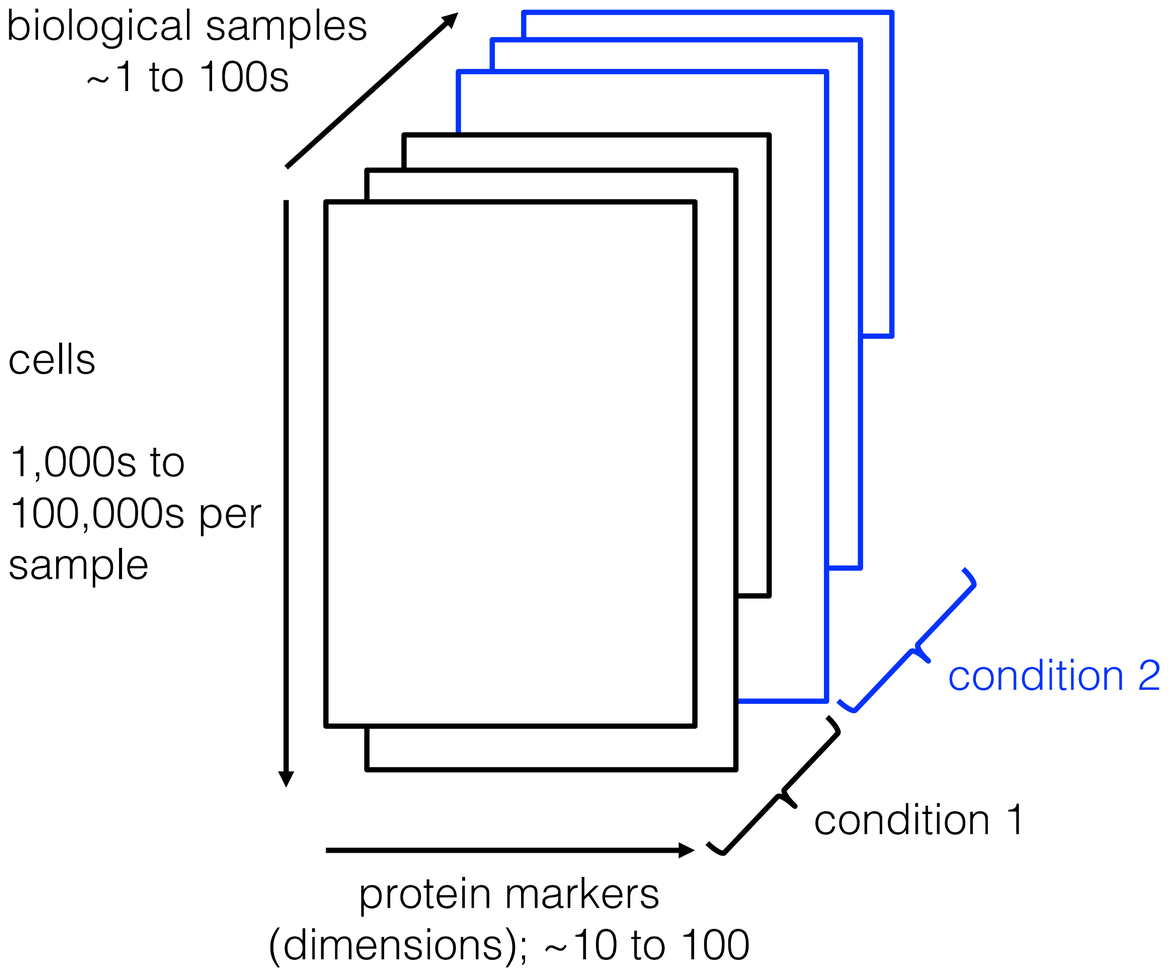
Differential state analysis

Differential abundance analysis

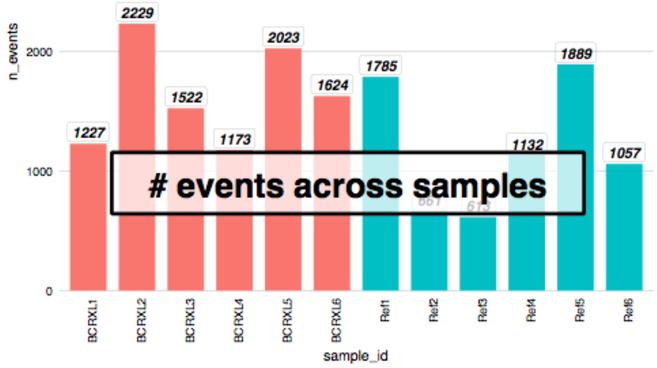
Data structure and differential analysis

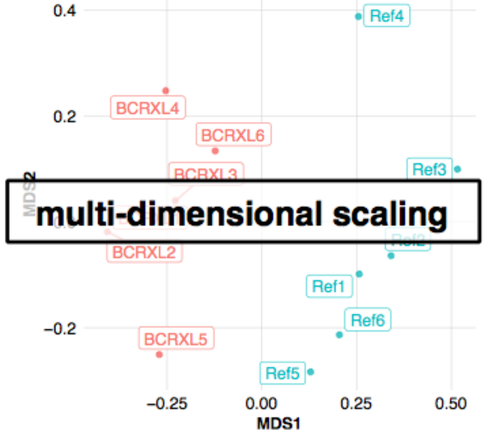
Two types of differential analysis

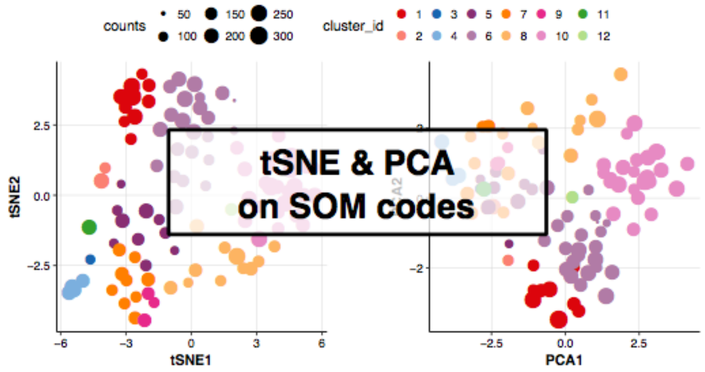
- **differential abundance** (DA) of cell populations
- **differential states**
  e.g., differential expression of functional proteins (e.g., signaling) within cell populations

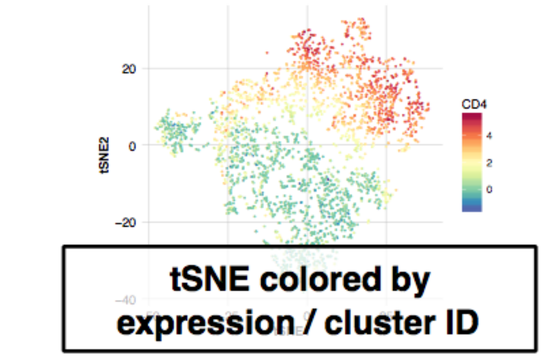


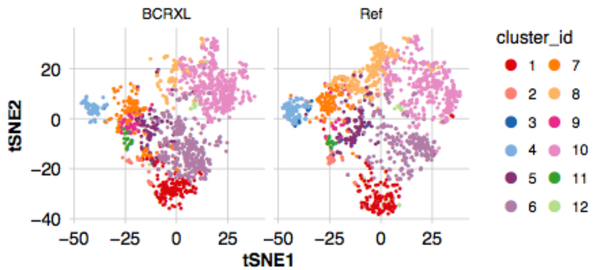
Analysis pipelines







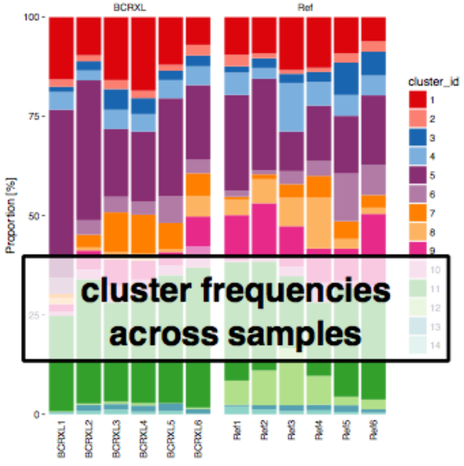


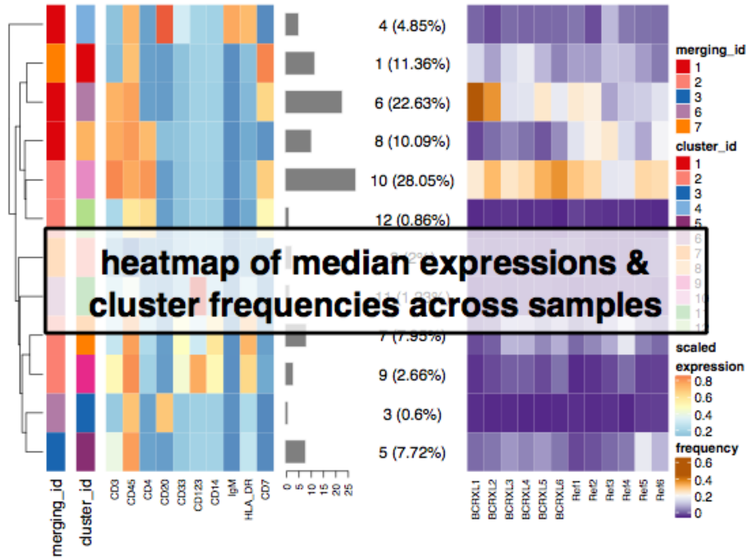


METHOD ARTICLE

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]





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?

Cytometry workflow: looking across multiple samples

Gosia

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]

preprocessing

↓

cluster all cells, all samples (merging or over-clustering)

↓

differential statistics

F1000 Bioconductor channel workflow published May 2017; updated May 2019 with drastically simplified code (functionality in CATALYST); will be updated again in Oct 2019 because of changes in BioC (again simplifications)

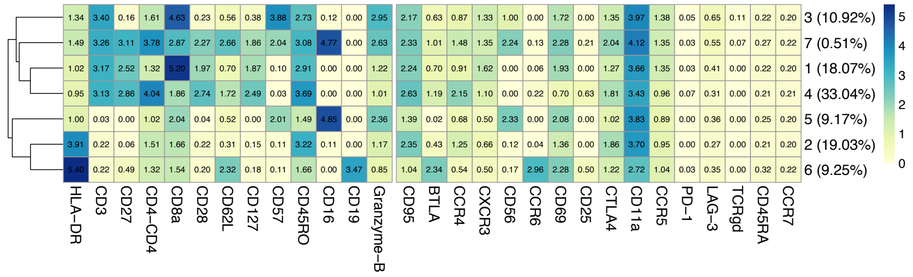
Helena

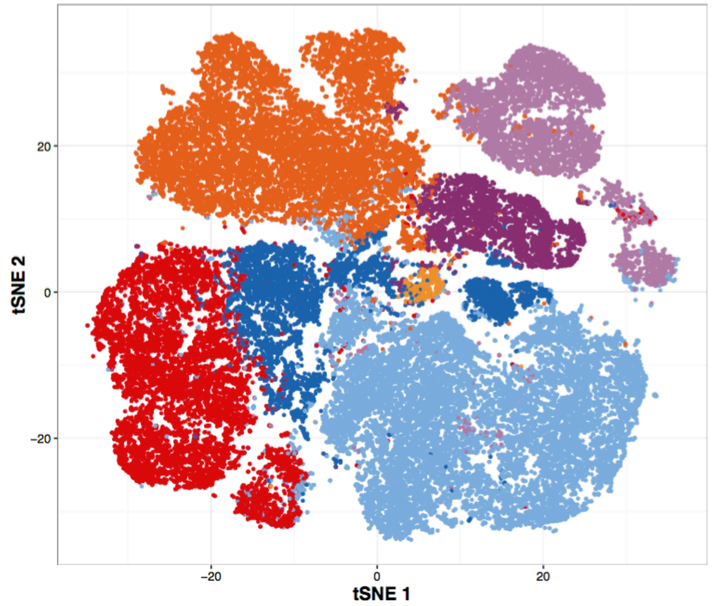
Key elements of CyTOF workflow

- Exploration of various data aspects at each step
- Separation of **type** and **state** markers
- Put all samples together and cluster (FlowSOM or other)
- Optional: manually merge clusters (via visualizations: heatmaps, low dimensional projects)
- Differential abundance analysis (count-based model, somewhat similar to RNA-seq)
- For **state** markers, differential state analysis (aggregate and use linear model)

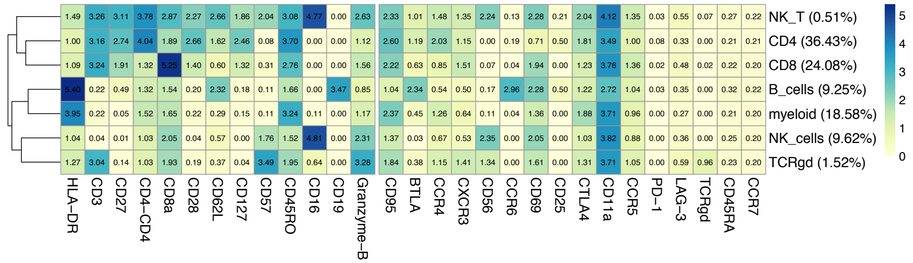
Merging clusters from 20 to 7

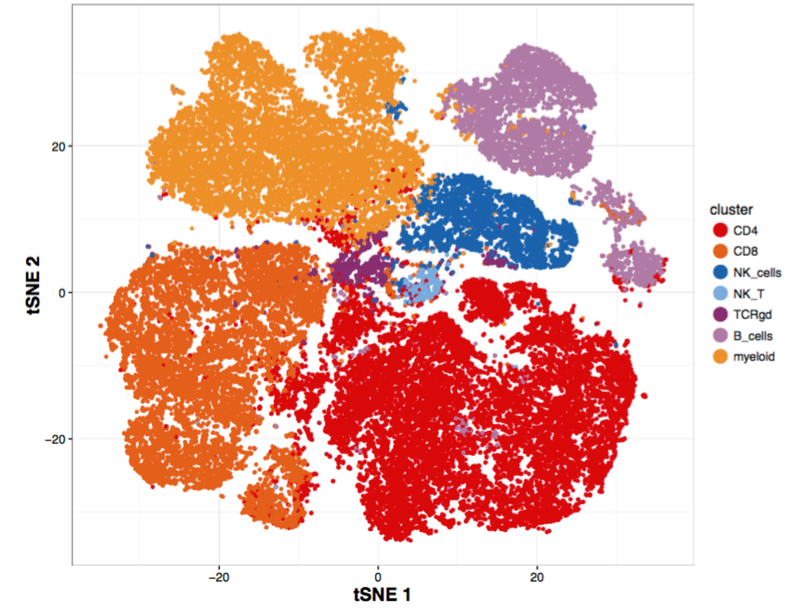
7 clusters





7 clusters by expert





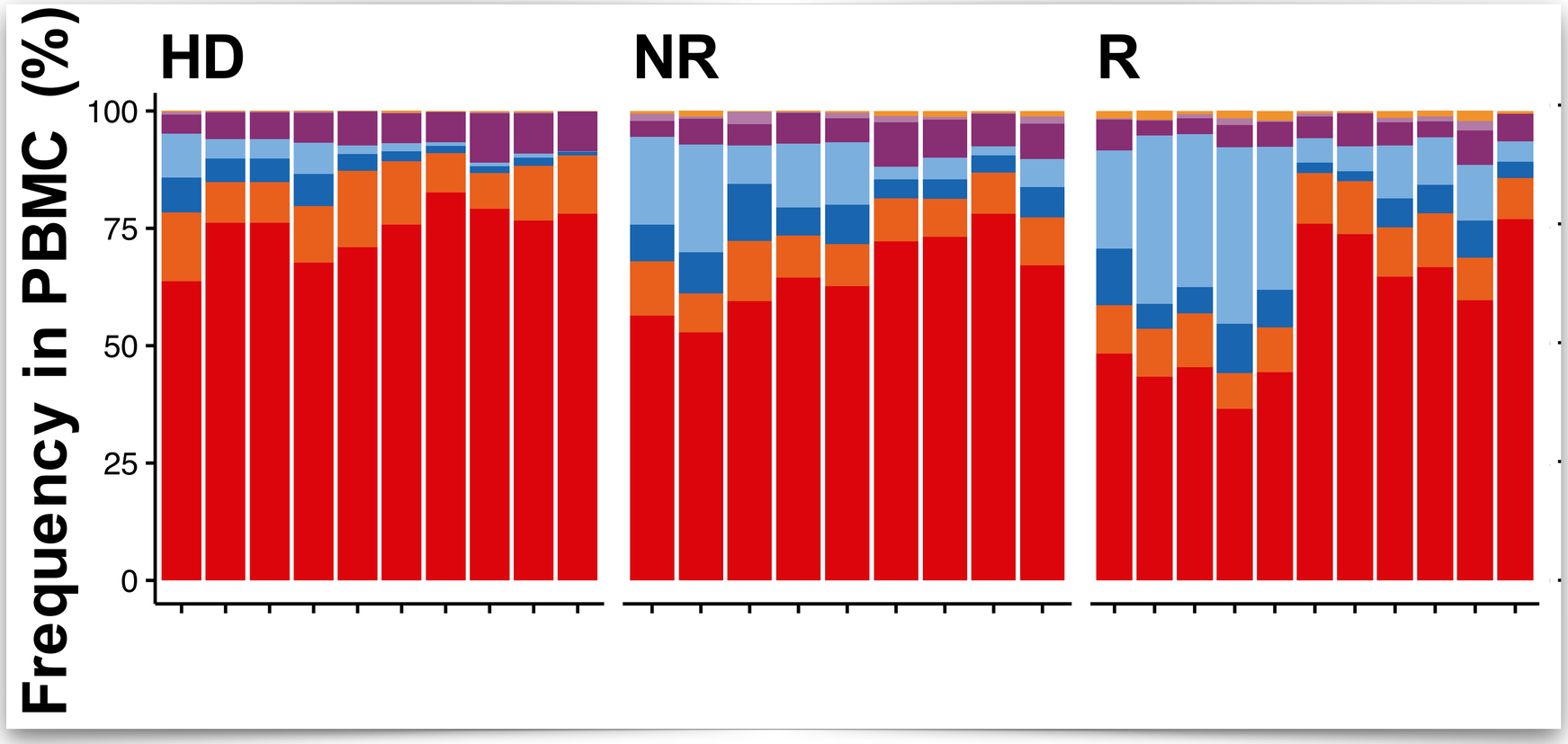
Carsten Krieg, now at MUSC
PBMCs from metastatic melanoma patients, comparing responders to non-responders)

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], Malgorzata Nowicka[2,3], Silvia Guglietta[4], Sabrina Schindler[5], Felix J Hartmann[1], Lukas M Weber[2,3], Reinhard Dummer[5], Mark D Robinson[2,3], Mitchell P Levesque[5,7] & Burkhard Becher[1,7]

Part 1: Differential abundance of cell populations



HD: healthy donors
NR: non-responders
R: responders

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).

Bioconductor workflow

Manual merging of cell populations based on phenotypes

Generalized linear mixed models (differential abundance)

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

Linear mixed models (differential expression within populations)

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

Limitations of existing methods

Table 1 Overview of existing methods and limitations

Method	Short description	Limitations	Ref.
Citrus	Uses hierarchical clustering and regularized regression or classification models to select predictive features, such as cluster abundances or median expression of functional markers, that are associated with an outcome of interest	• Detected features cannot be ranked by importance • Lasso-regularized models cannot easily detect multiple correlated features • Rare cell populations cannot easily be detected, due to minimum cluster size requirement and computational limitations • Response variable is the clinical outcome variable, which makes it difficult to account for complex experimental designs (including batch effects, paired designs, and continuous covariates)	9
CellCnn	Applies convolutional neural networks in a representation learning framework to detect rare cell populations associated with an outcome of interest; designed specifically for detecting rare cell populations	• Ranking of detected cells cannot be interpreted in terms of statistical significance • Interpretation of detected populations (referred to as filters) can be difficult, since they may be composed of multiple distinct cell populations • Response variable is the clinical outcome variable, which makes it difficult to account for complex experimental designs (including batch effects, paired designs, and continuous covariates) • All protein markers are treated identically; there is no conceptual split between cell type and cell state (or functional) markers	10
cydar	Assigns cells to overlapping hyperspheres in the high-dimensional space; tests for differential abundance between hyperspheres using moderated tests from *edgeR*[15,16], while controlling the spatial false discovery rate among overlapping hyperspheres	• Rare cell populations cannot easily be detected, due to their relatively small volume in the high-dimensional space • All protein markers are treated identically; there is no conceptual split between cell type and cell state (or functional) markers	11
classic regression-based approach	Automated clustering using *FlowSOM*[14], followed by manual merging and annotation to define cell populations; differential testing of features such as population abundances or median expression of functional markers using generalized linear mixed models, linear mixed models, or linear models	• Manual merging and annotation step requires expert biological knowledge, and can be time-consuming and subjective • When testing large numbers of clusters, e.g. to detect rare cell populations: loss of statistical power due to multiple testing penalty; no sharing of information across clusters	12

Overview of recently developed methods for performing differential analyses in high-dimensional cytometry data. For each method, a short description of the methodology and a summary of limitations are provided

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

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

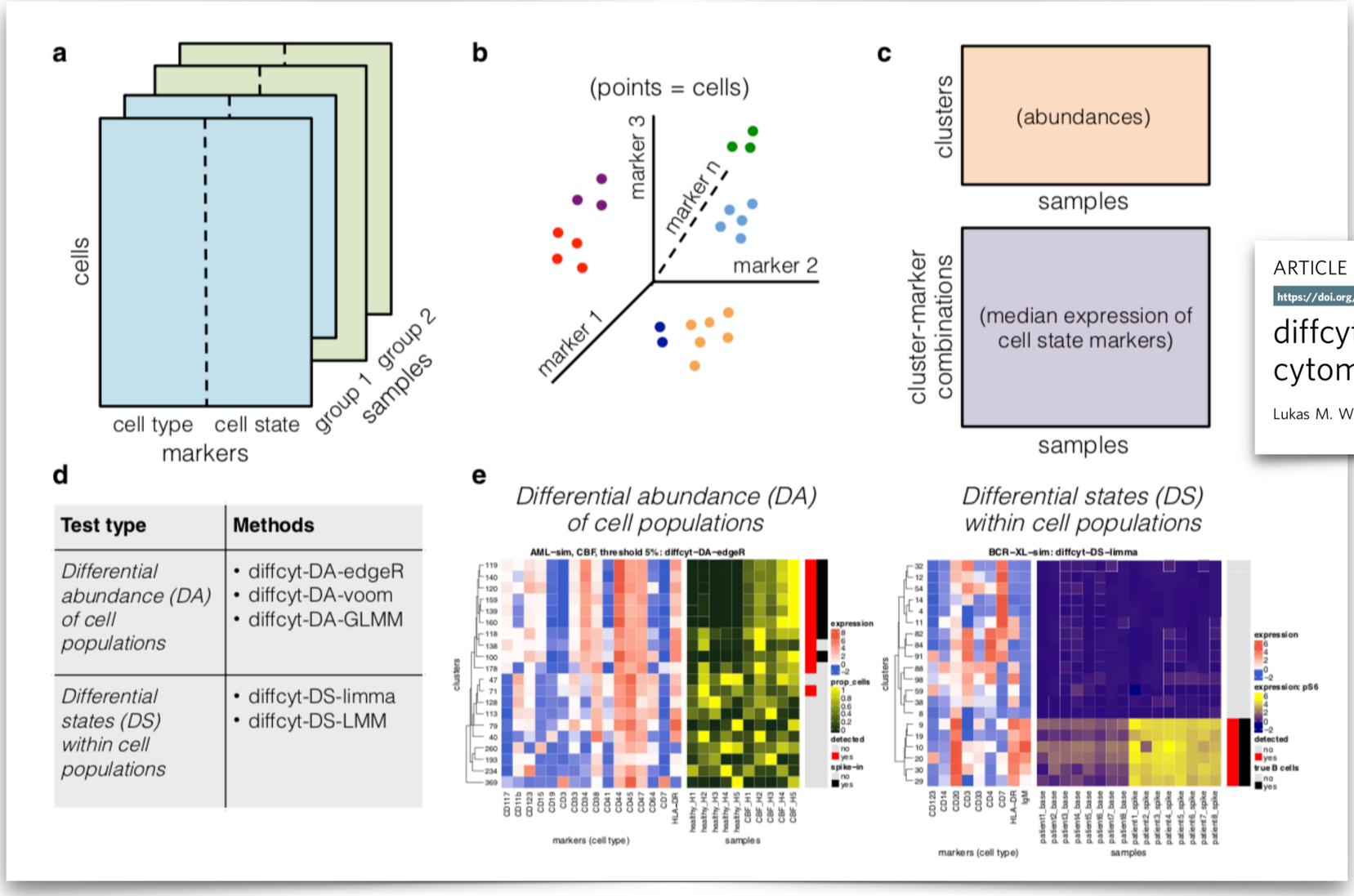
Lukas M. Weber[1,2], Malgorzata Nowicka[1,2,3], Charlotte Soneson[1,2,4] & Mark D. Robinson[1,2]

n.b. Citrus/CellCnn models are reversed to ours (response variable: patient/experimental condition; explanatory variables: CyTOF measurements)

cydar doesn't distinguish type and state

diffcyt: differential tests more formalised

Lukas



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

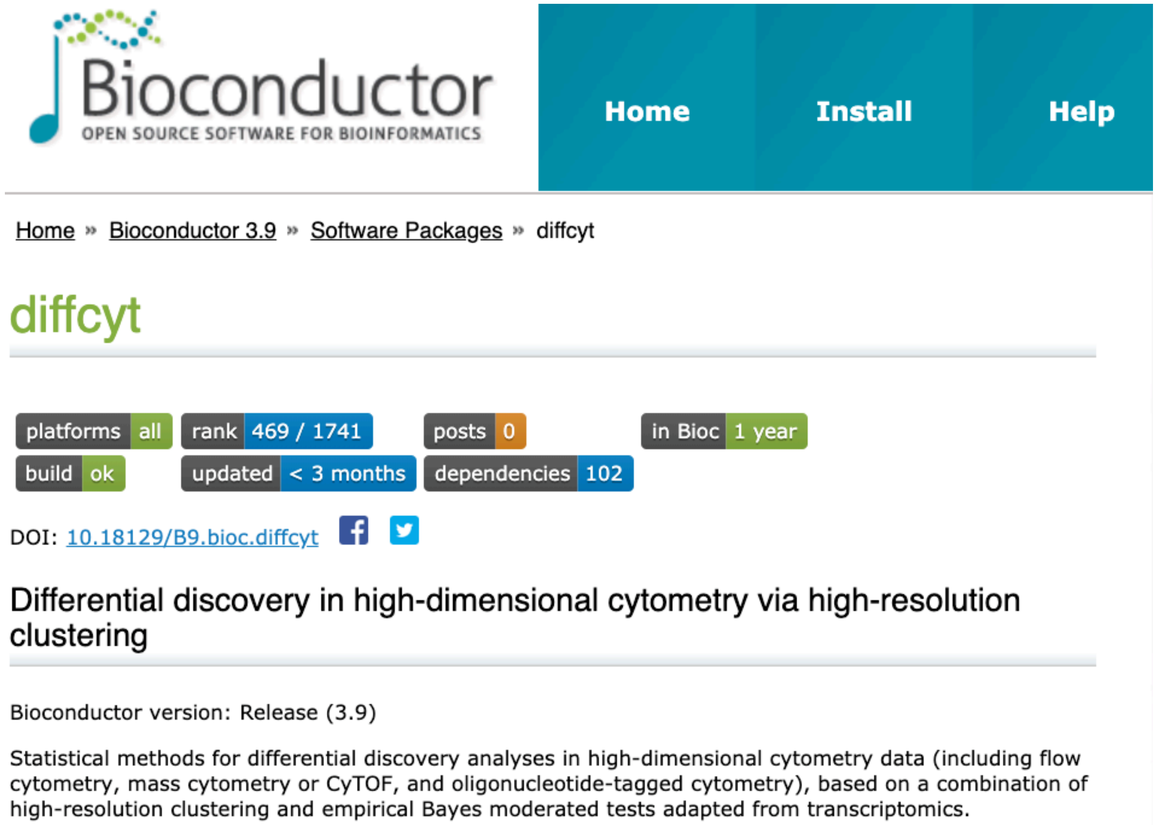
https://doi.org/10.1038/s42003-019-0415-5 OPEN

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]

Note: for differential state analysis, aggregates are always taken. We are testing this now with scRNA-seq data

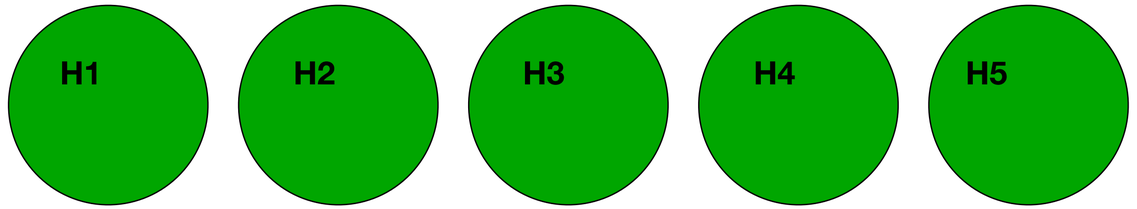
diffcyt: Bioconductor package



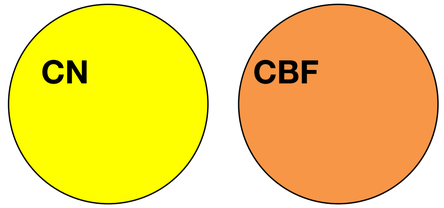
Interoperable with CATALYST for pipelines

Creation of a benchmark: AML-sim data generation strategy

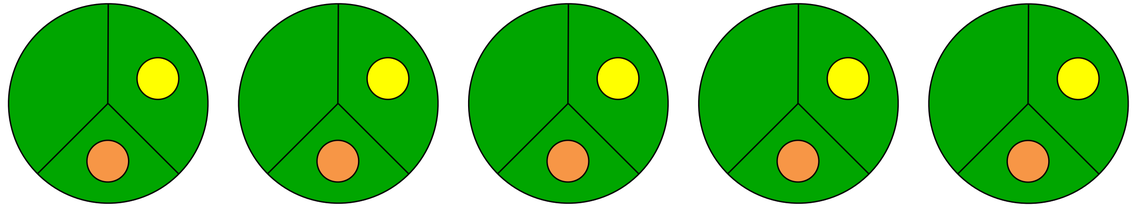
5x healthy samples



AML: 1x CN, 1x CBF



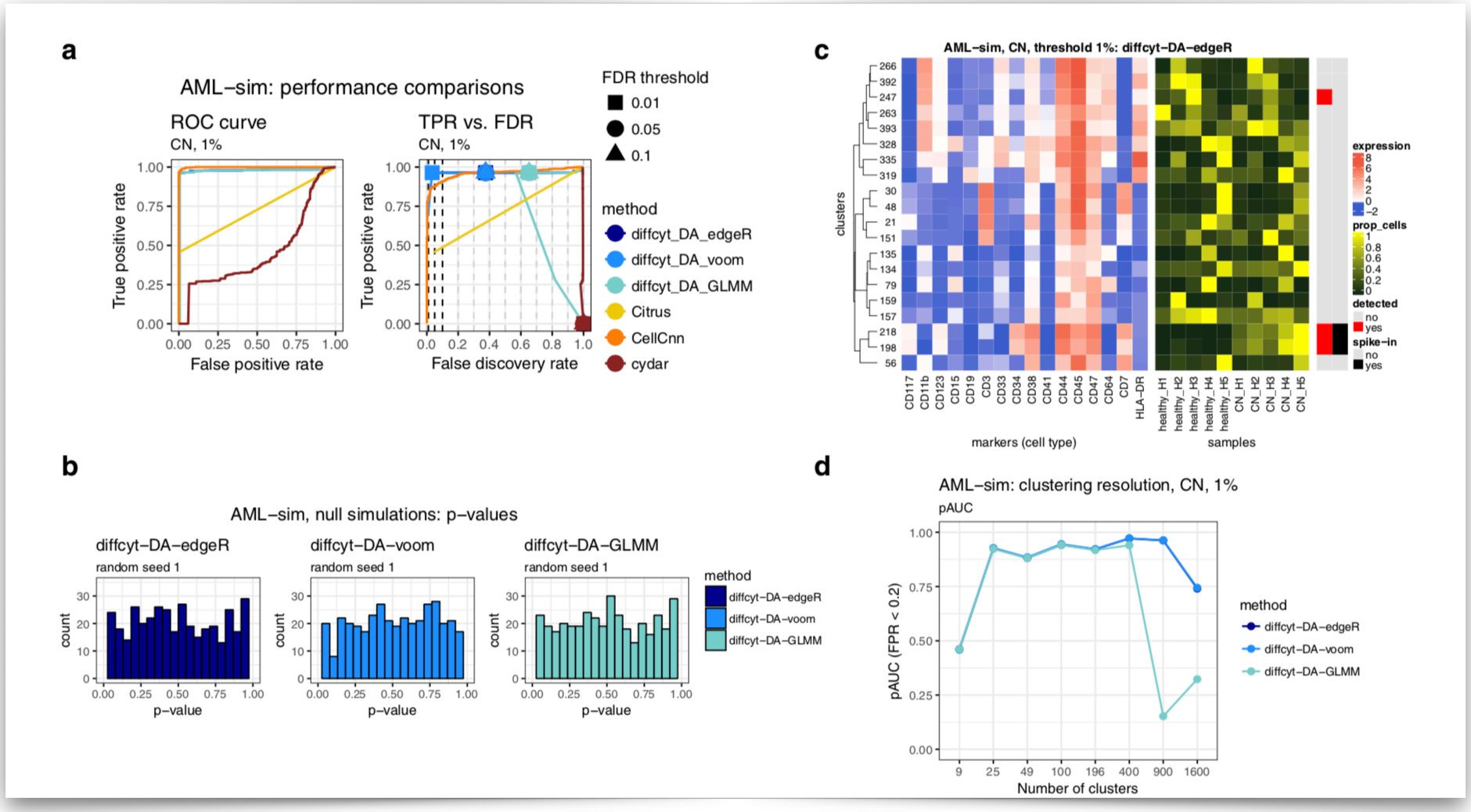
Split each healthy sample into 3 equal parts; computationally "spike in" CN and CBF cells



Repeat for different thresholds: 5%, 1%, 0.1%, 0.01%

(strategy adapted from Arvaniti et al., 2017)

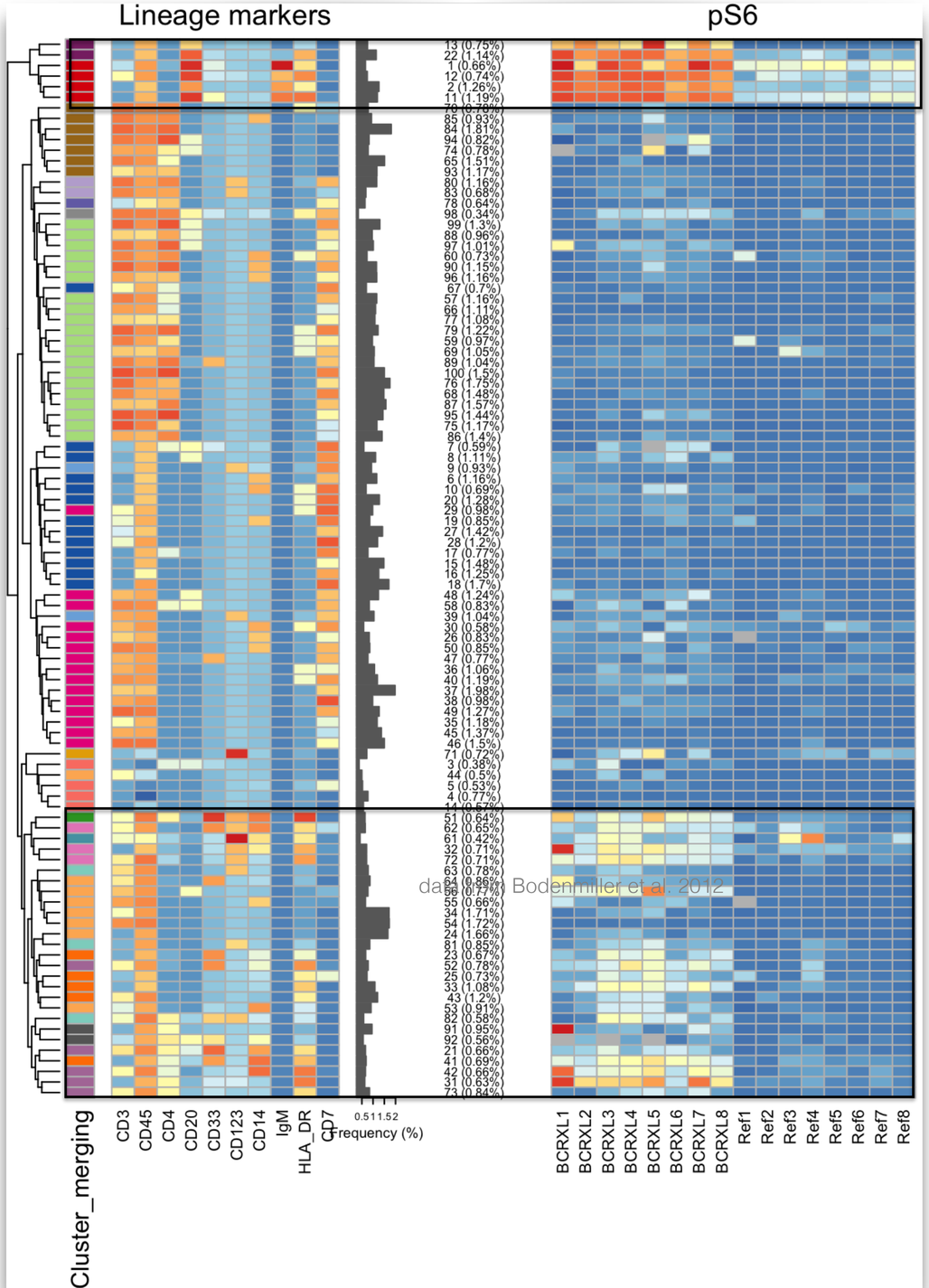
Differential abundance detection performance across methods



Part 2: subpopulation-specific differential analyses (“state”)

Clustered here to 100 groups; for each, look across samples in functional marker

→ median **lineage** marker signal by cluster

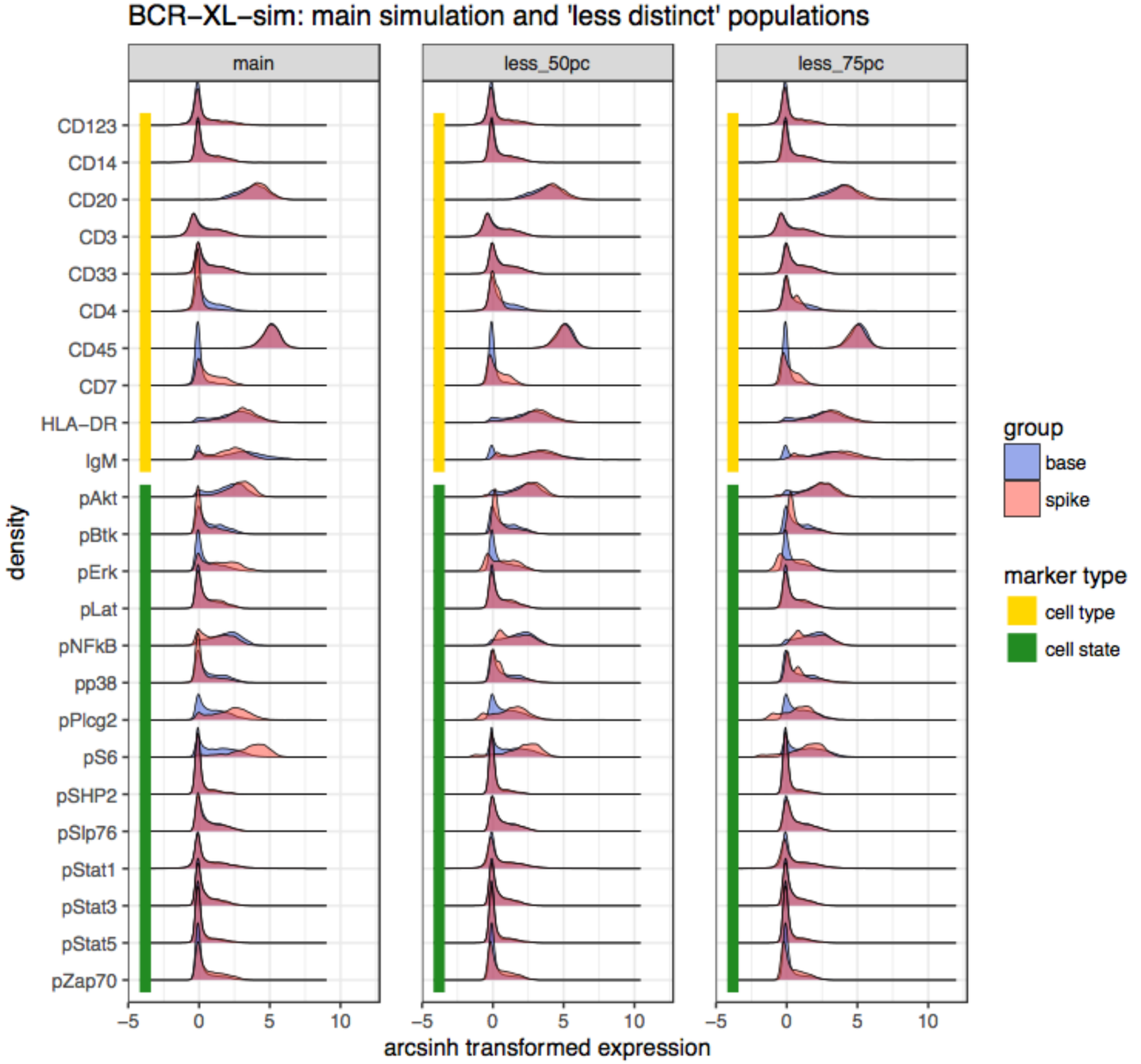


data from Bodenmiller et al. 2012

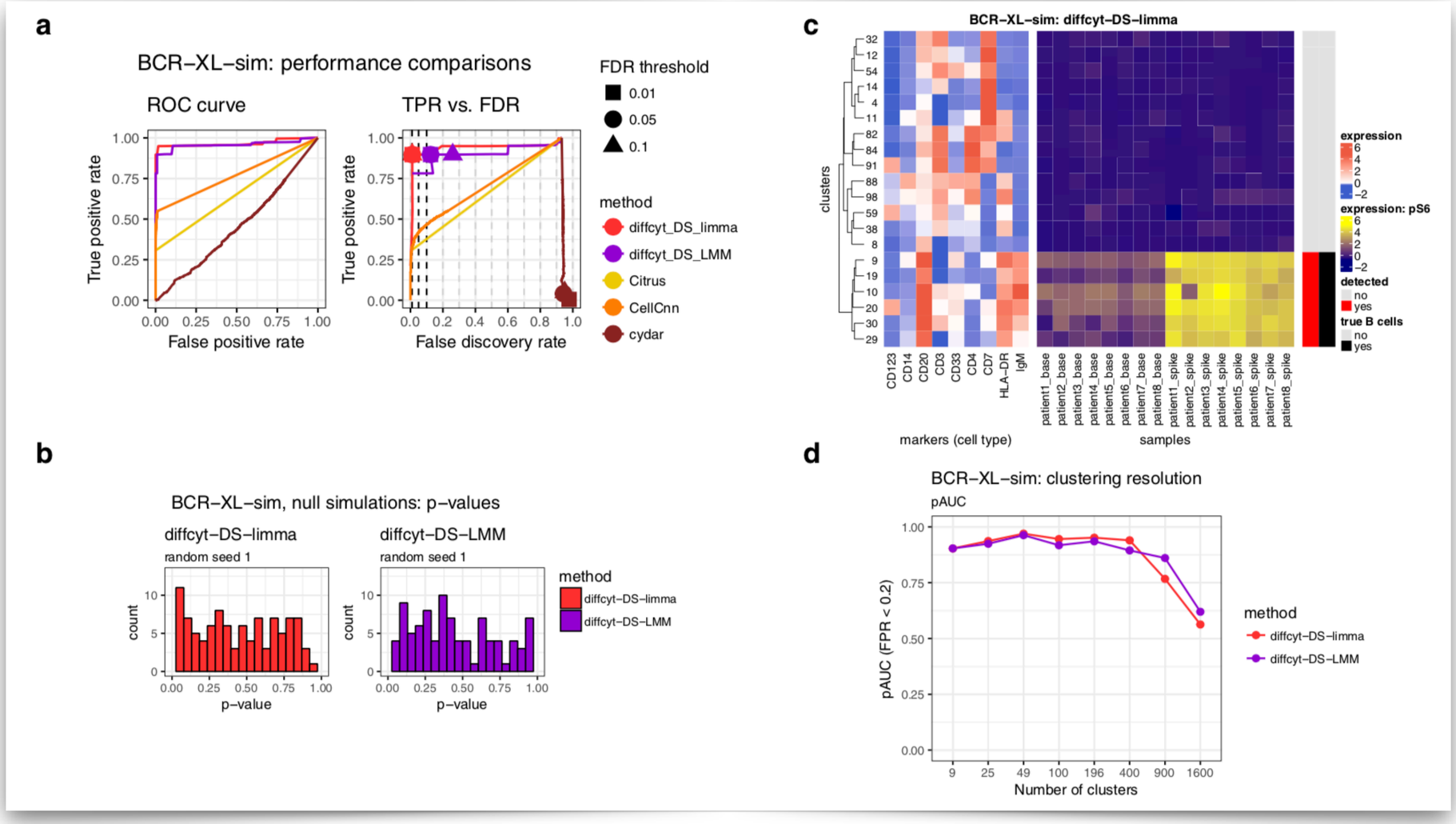
← *median* **functional** marker signal by sample

diffcyt: benchmarking

Sensitivity



Differential state detection performance across methods



Some notes on CyTOF differential discovery

- Flexibility in definition of type and state
- For rare populations, ability to detect changes in abundance is driven jointly by “distinctness” (clustering) and “rarity” (more abundant —> easier); also related to depth of sampling
- Fairly wide range in the “sweet spot” of clustering
- Cell type assignment as an alternative to clustering is easily accommodated

HDCytoData package

Collection of benchmark datasets in Bioconductor formats

HDCytoData

platforms all | rank 134 / 371 | posts 0 | build ok
updated before release | dependencies 93

DOI: 10.18129/B9.bioc.HDCytoData

Collection of high-dimensional cytometry benchmark datasets in Bioconductor object formats

Bioconductor version: Release (3.9)

Data package containing a collection of high-dimensional cytometry benchmark datasets saved in SummarizedExperiment and flowSet Bioconductor object formats, including row and column metadata describing samples, cell populations (clusters), and protein markers.

Author: Lukas M. Weber [aut, cre], Charlotte Soneson [aut]

Maintainer: Lukas M. Weber <lukmweber at gmail.com>

Citation (from within R, enter `citation("HDCytoData")`):

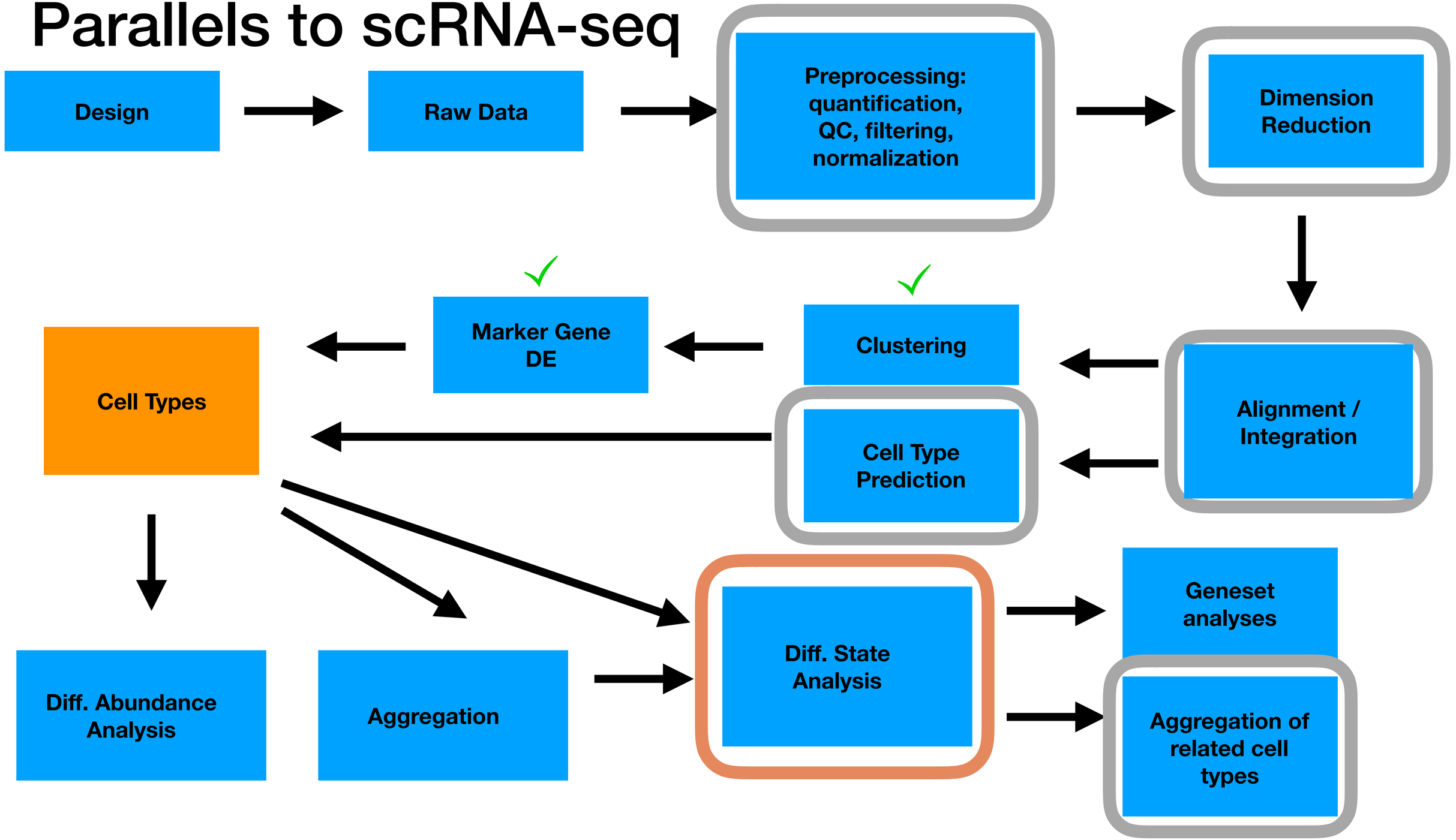
Weber L, Soneson C (2019). *HDCytoData: Collection of high-dimensional cytometry benchmark datasets in Bioconductor object formats*. R package version 1.4.0, https://github.com/lmweber/HDCytoData.

HDCytoData package

Bioconductor ExperimentHub

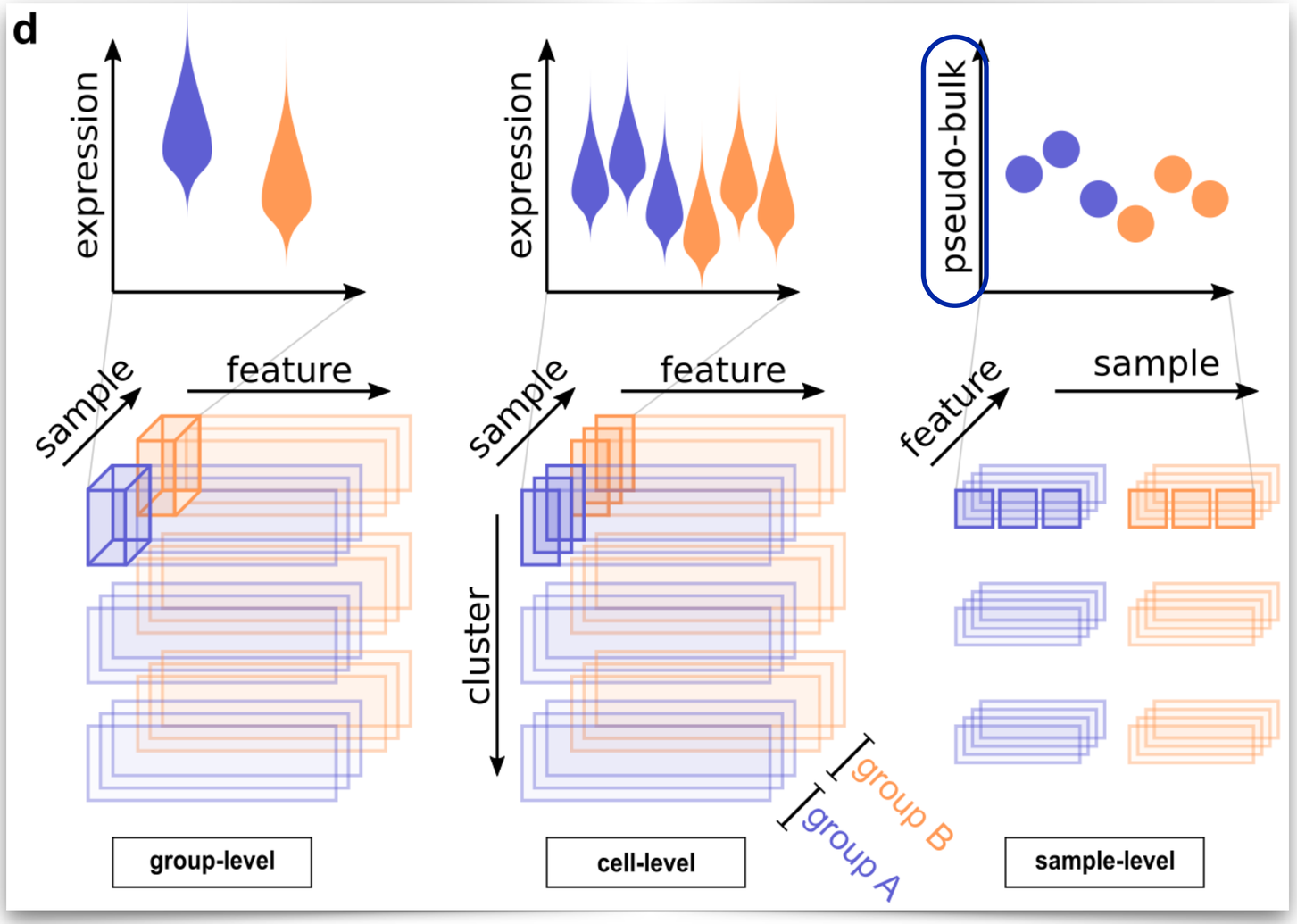
```
> library(HDCytoData)

> data_SE <- Levine_32dim_SE()
> data_flowSet <- Levine_32dim_flowSet()

> ehub <- ExperimentHub()
> query(ehub, "HDCytoData")
> data_SE <- ehub[["EH1119"]]
> data_flowSet <- ehub[["EH1120"]]
```

Parallels to scRNA-seq



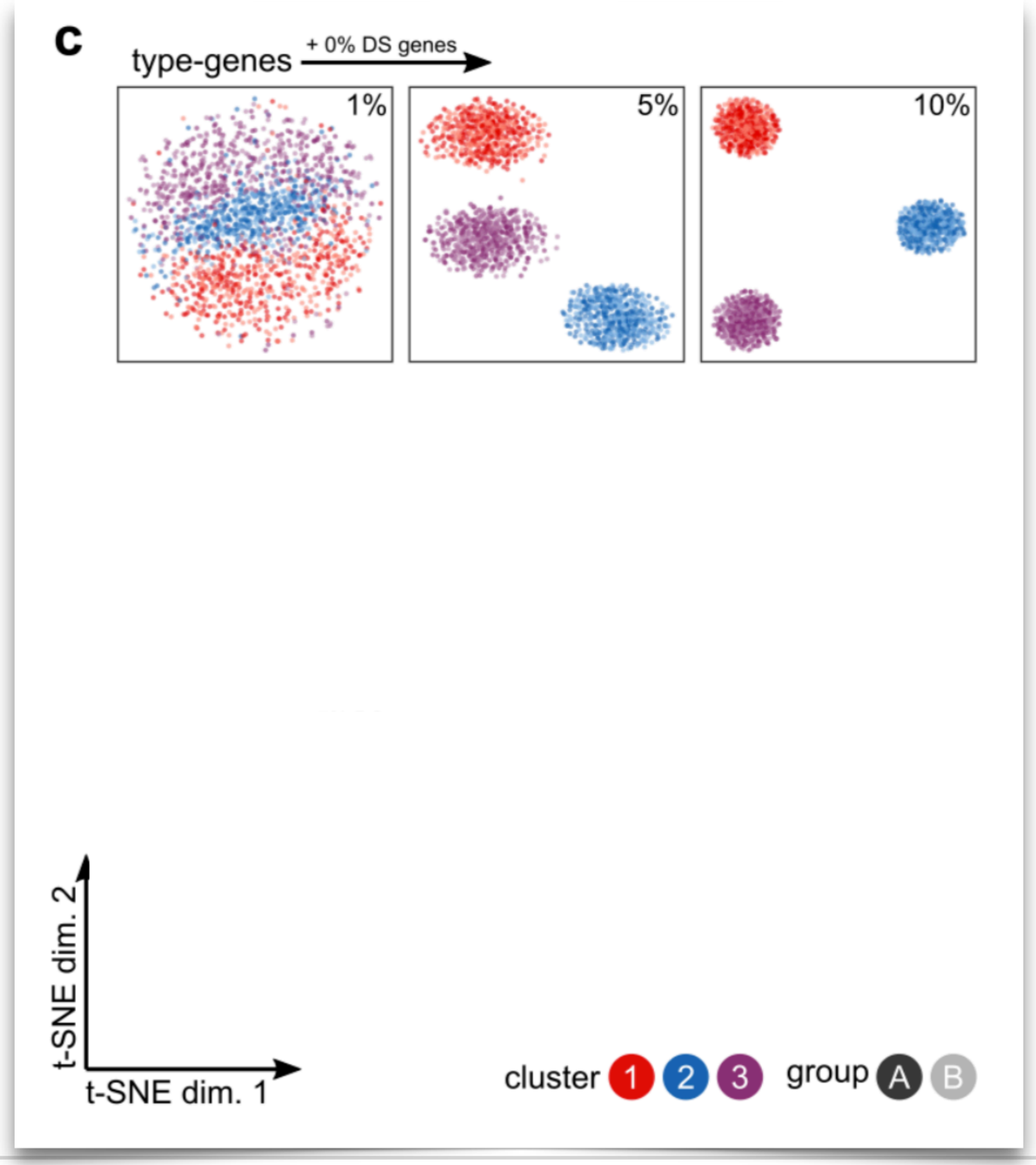
After “Cell Type Prediction” / “Clustering”, various ways to view the inference problem

Multi-sample
Multi-condition
Multi-population



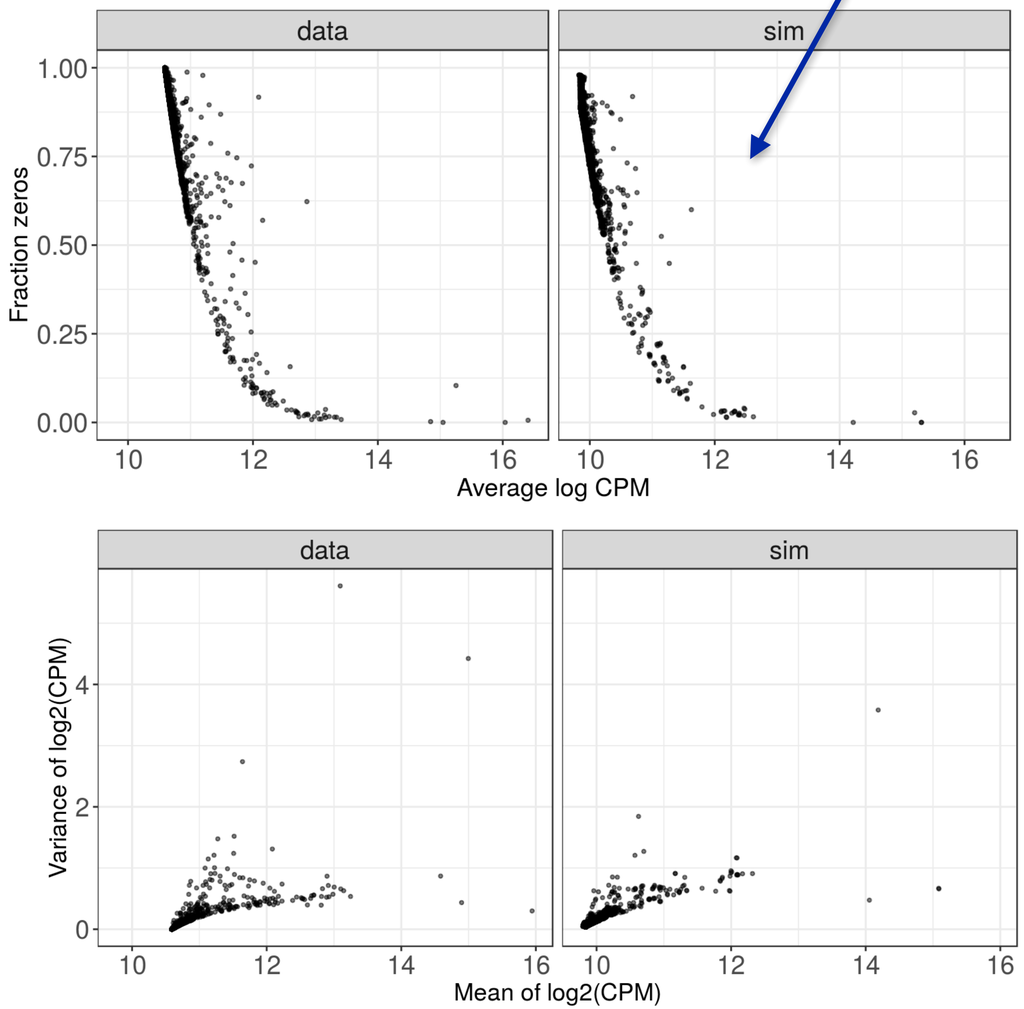
Flexible simulation

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

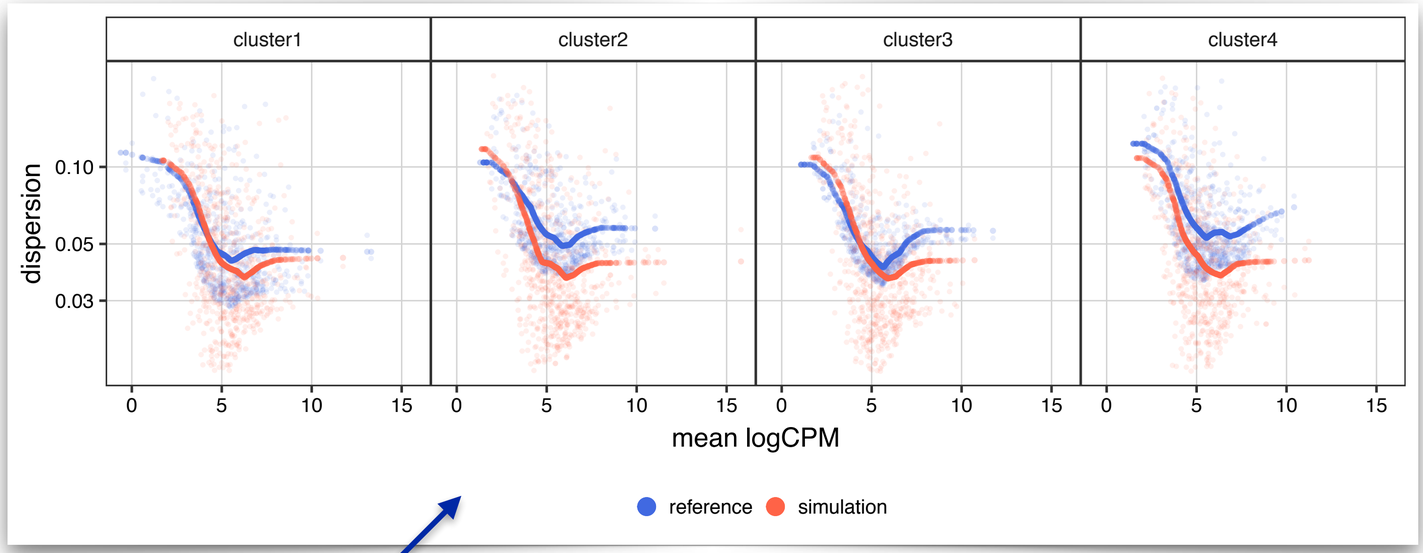


countsimQC: comparing simulated data to real data

cell-level properties



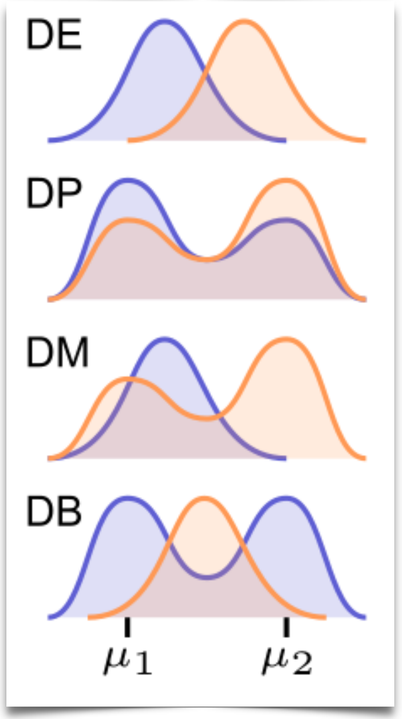
pseudobulk-level dispersion-mean relationships

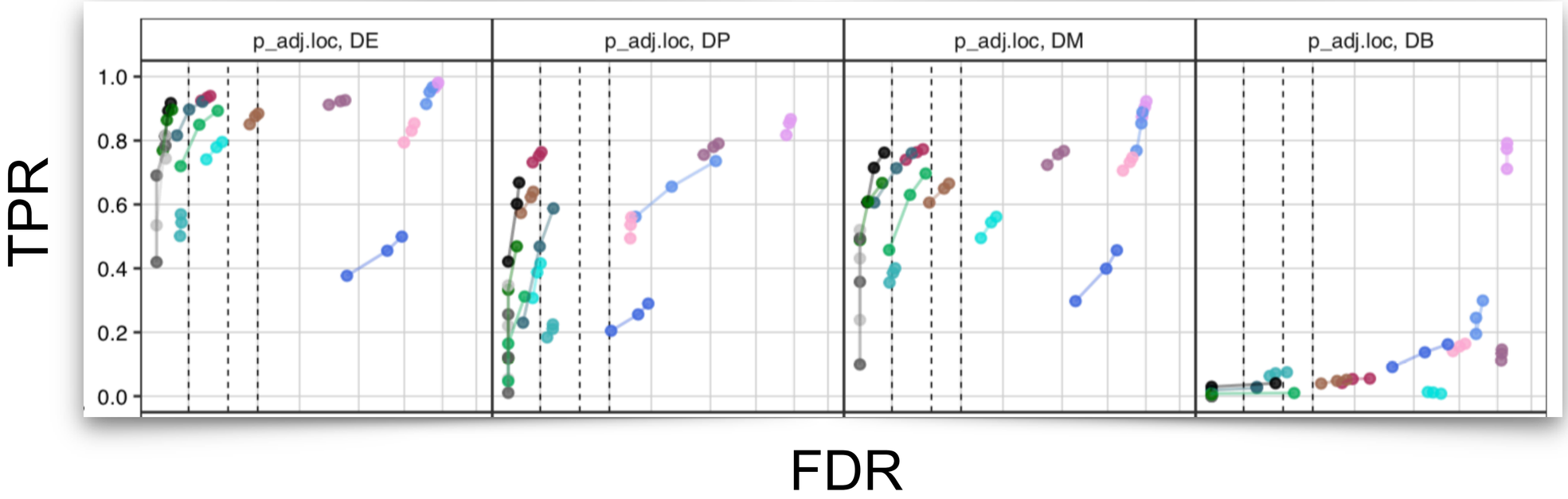


aggregate-level properties

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

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

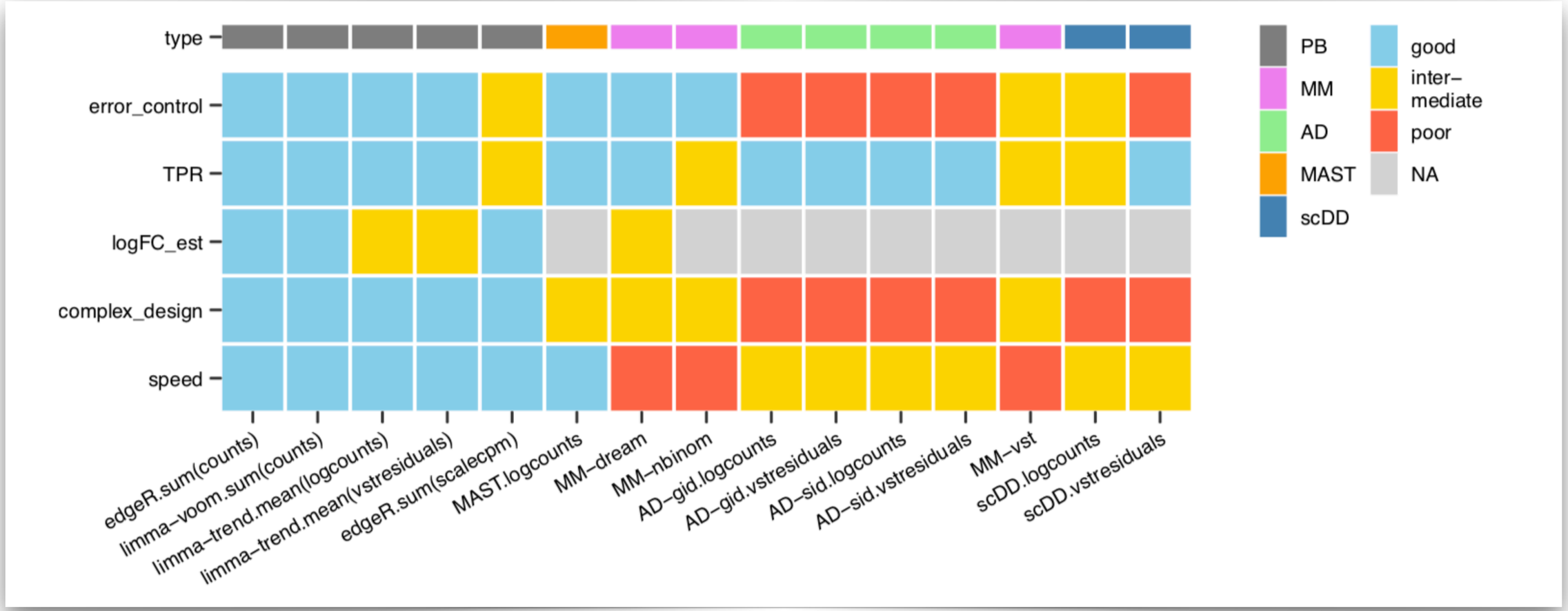




AD = Anderson-Darling
MM = mixed models

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

Current rating



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

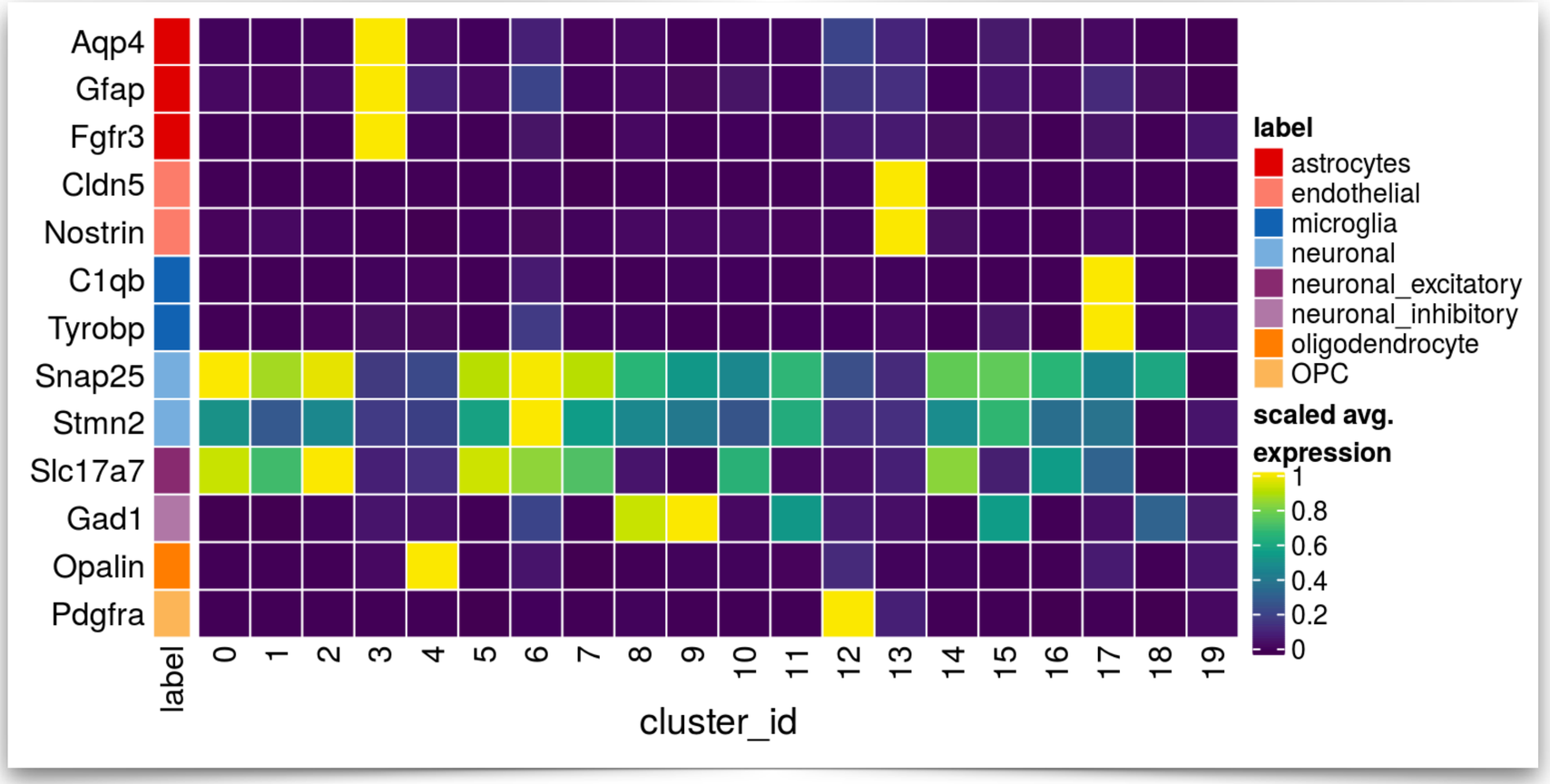
Application to LPS dataset: clustering + annotation subpopulations

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

frontal cortex

single nuclei RNA-seq (10x)

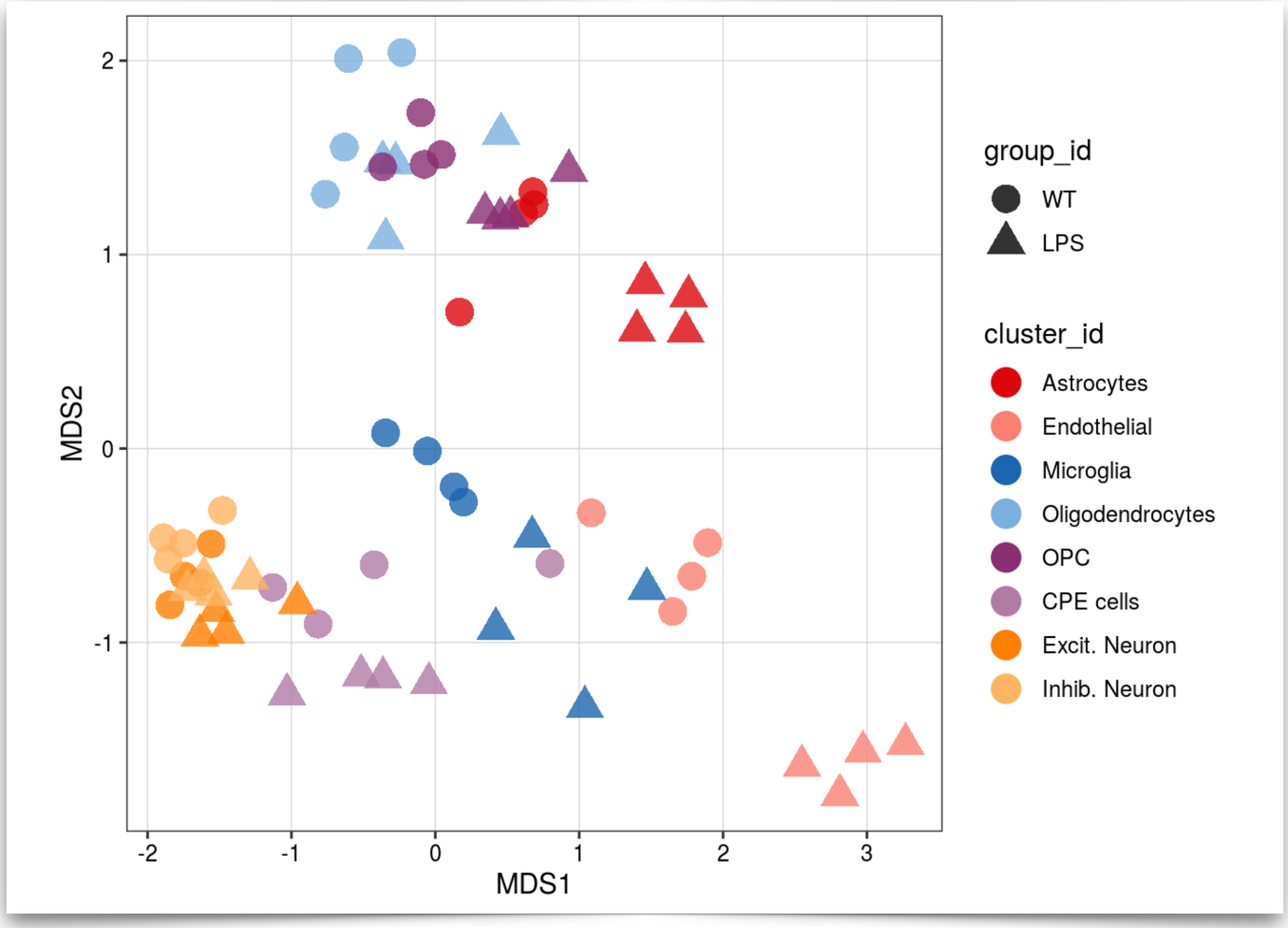
usual preprocessing:
filtering, doublet removal,
Seurat integration,
clustering



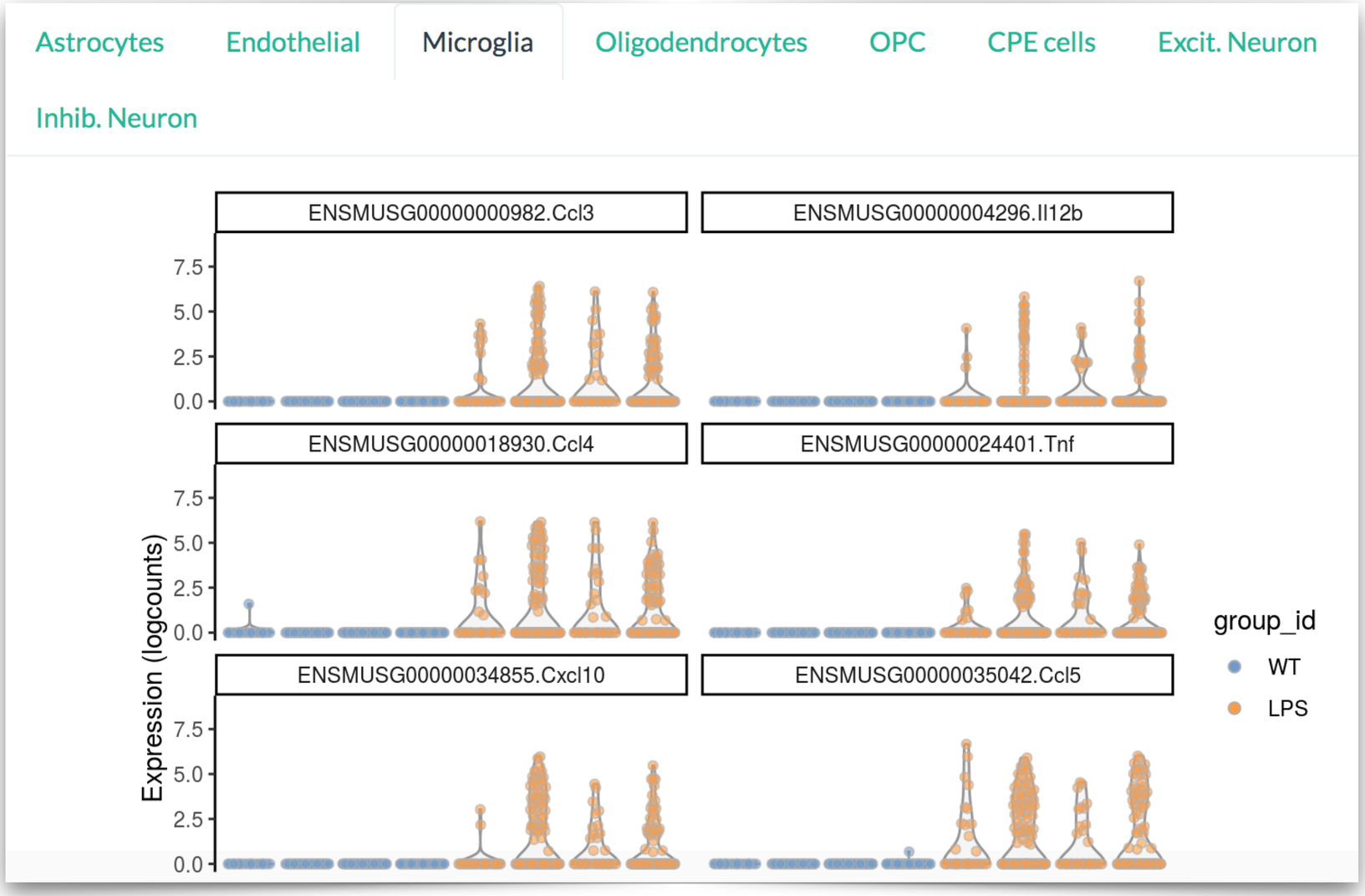
Application to LPS dataset: subpopulation-level visualization

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

Each dot is one subpopulation/ sample combination

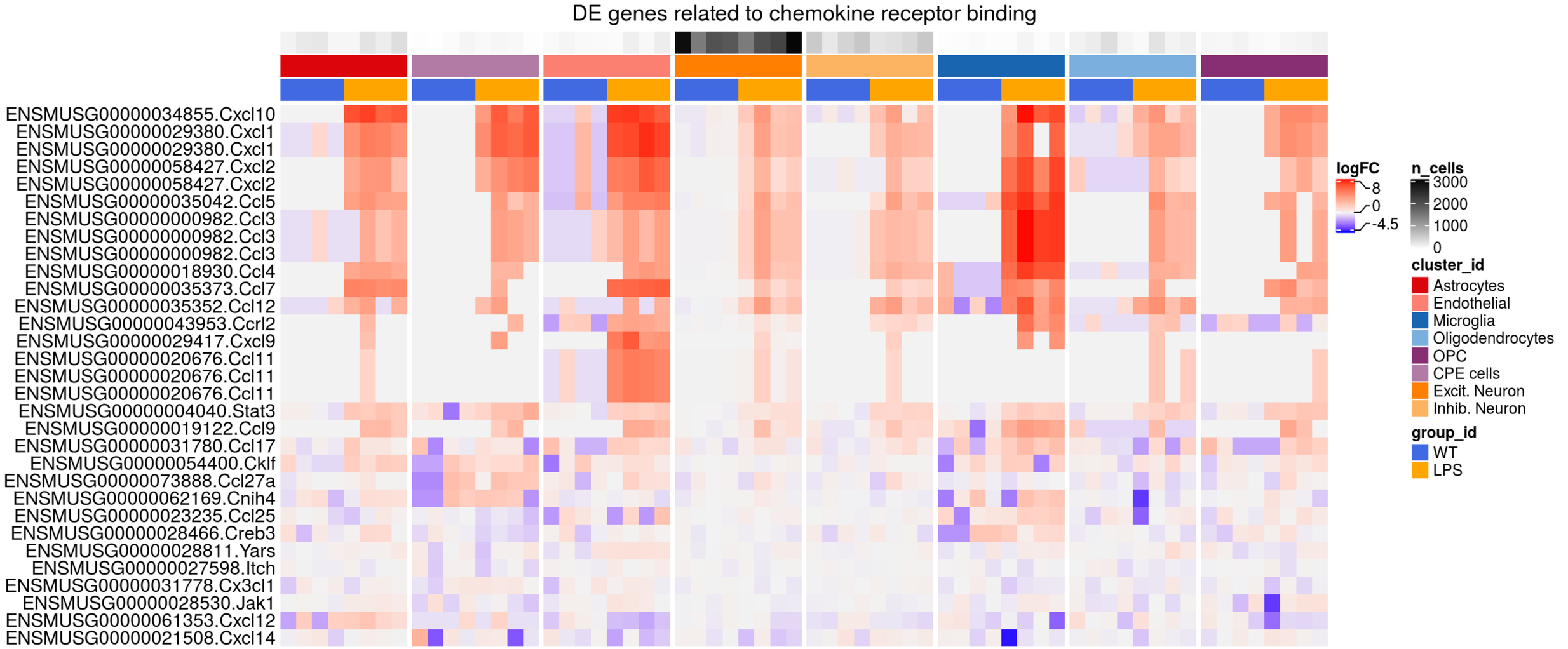


Application to LPS dataset: go back to cell-level response (discovery based on pseudobulk)



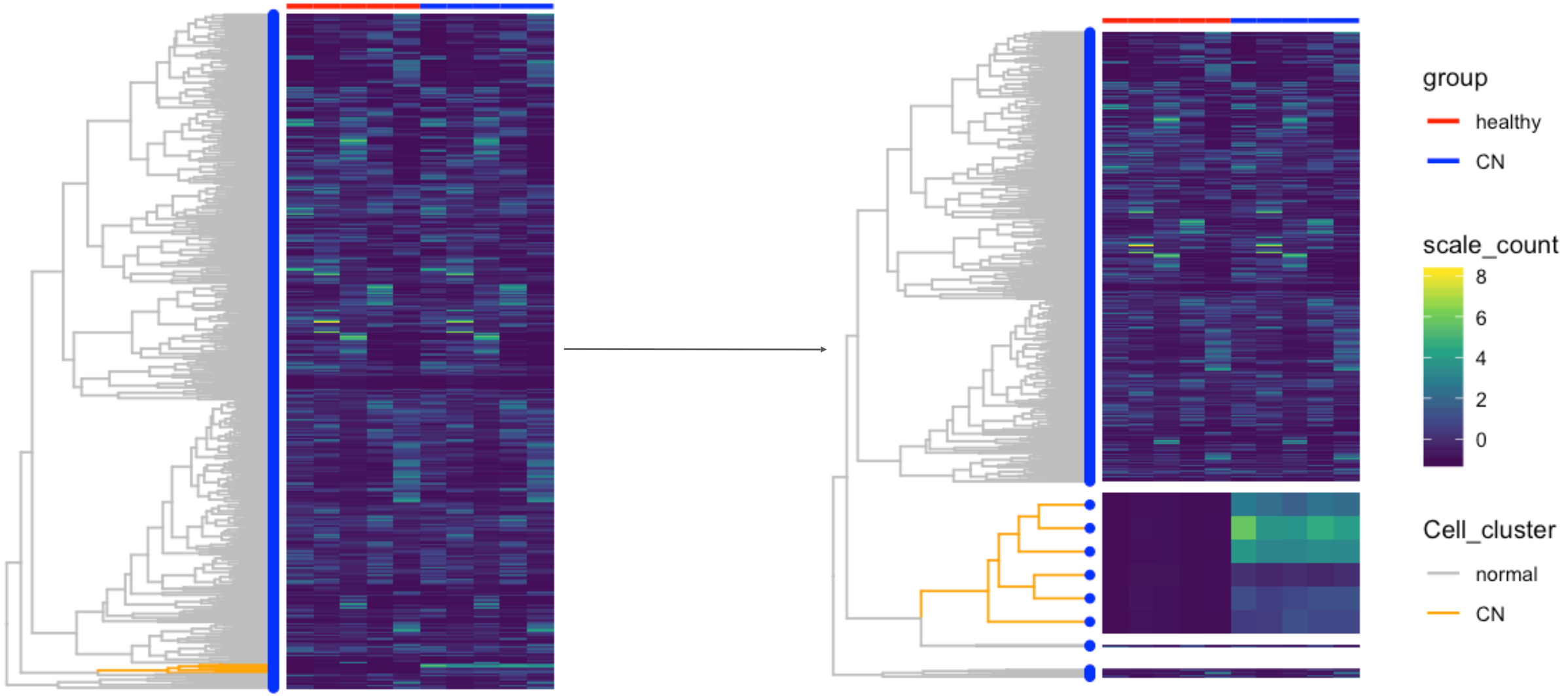
workflowr !

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



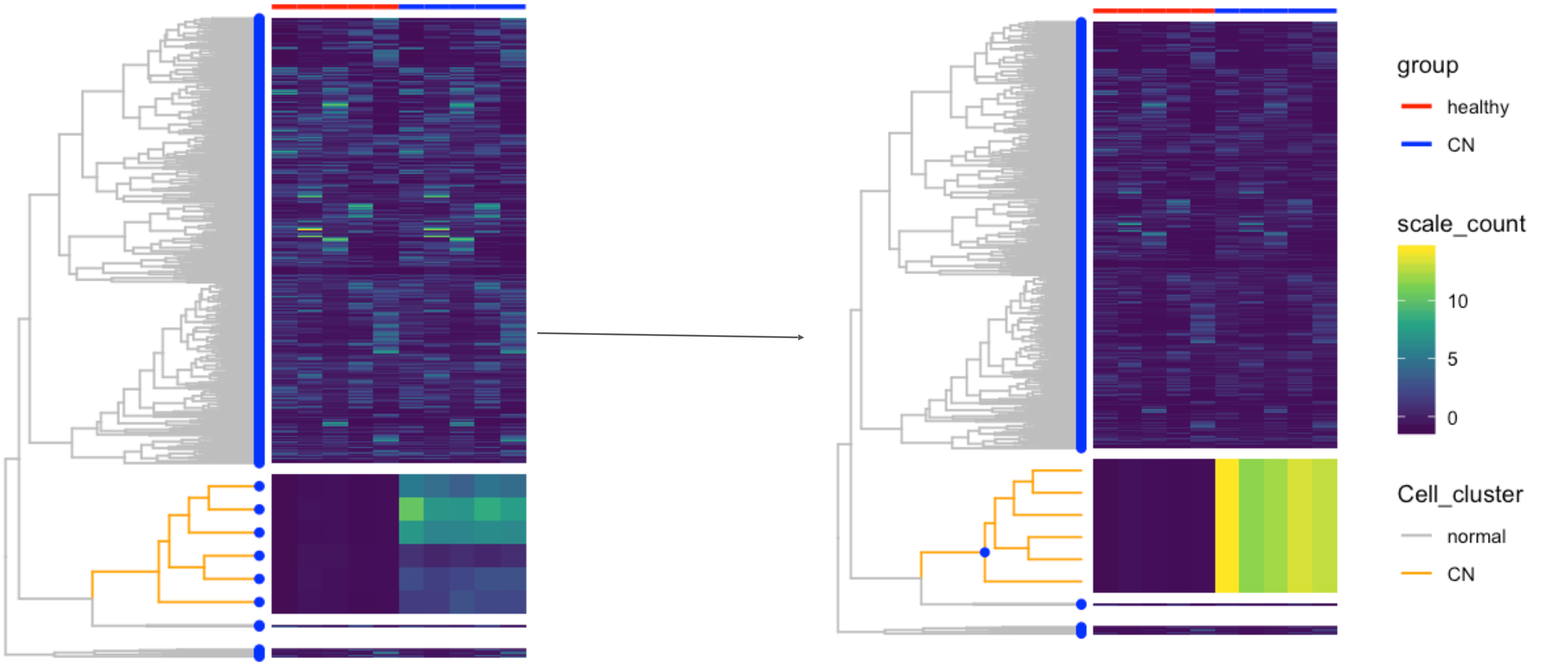
LPS dataset: interplay of cell type and cell state

Motivation: can we use the tree information in the differential inferences?



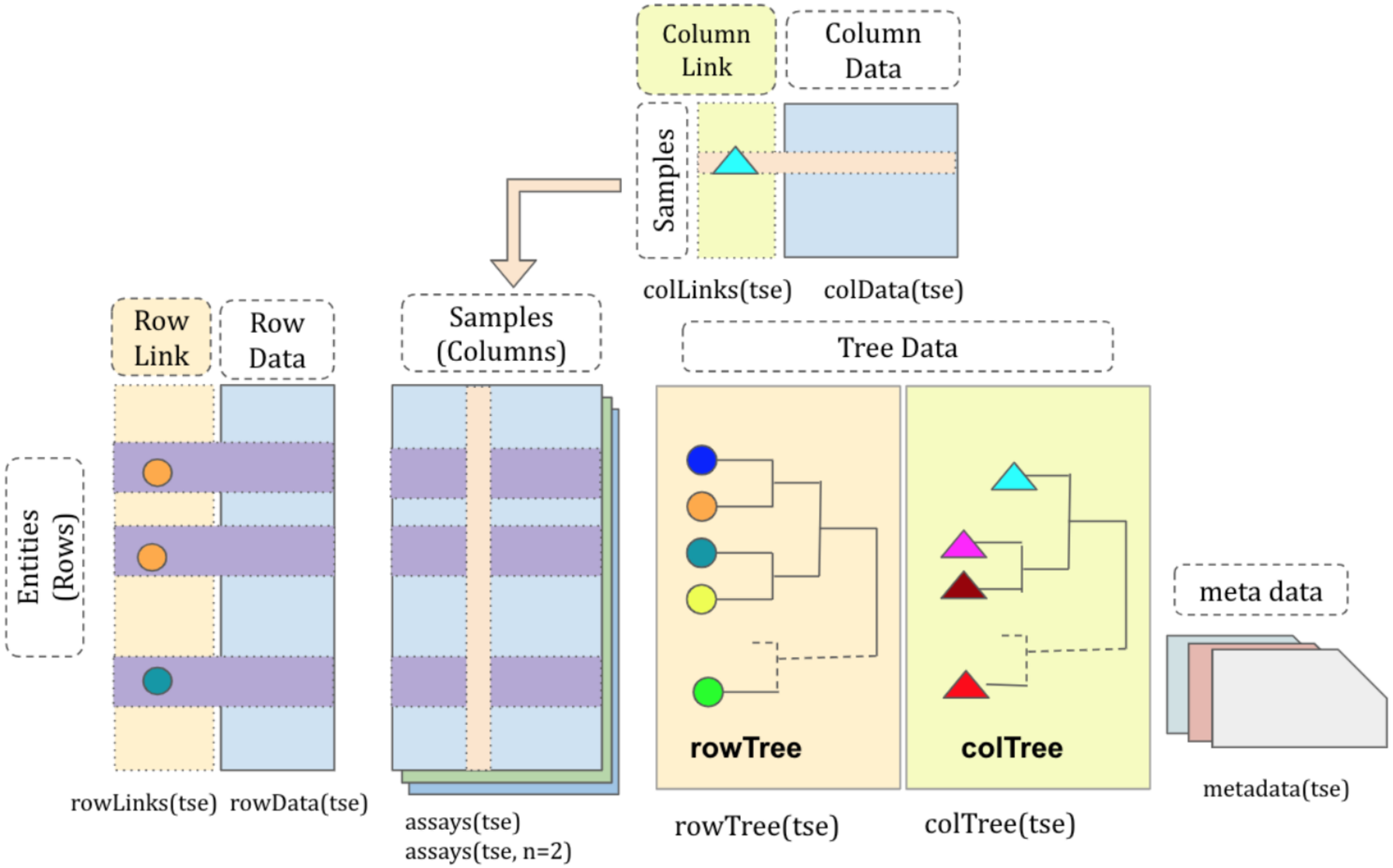
Give more space to orange branch; The visualization is on the leaf level (blue points)

Visualization: TreeHeatmap



Change visualization to a specified level

The structure



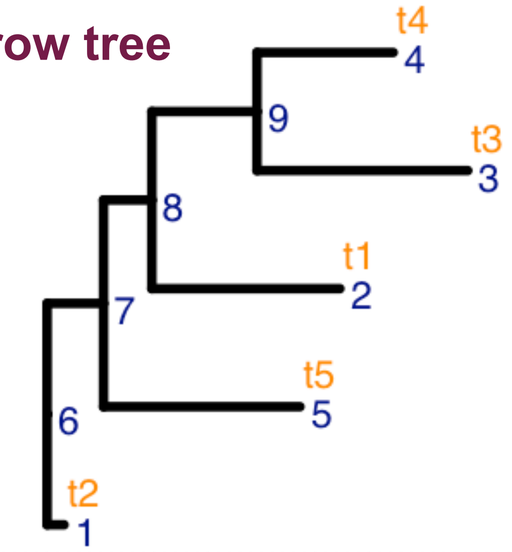
The components

```
> row_data
DataFrame with 5 rows and 2 columns
                var1      var2
         <character> <logical>
entity1            a      TRUE
entity2            b     FALSE
entity3            a      TRUE
entity4            b     FALSE
entity5            b      TRUE
```

```
> assay_data
         A_1 A_2 B_1 B_2
entity1    0   0   0   0
entity2    1   5   9  13
entity3    2   6  10  14
entity4    3   7  11  15
entity5    4   8  12  16
```

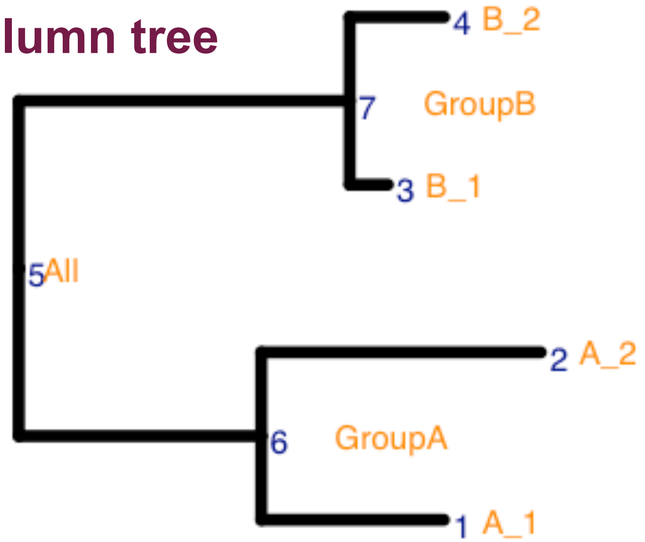
```
> col_data
DataFrame with 4 rows and 2 columns
           gg       group
    <numeric> <character>
A_1         1           A
A_2         2           A
B_1         3           B
B_2         3           B
```

The row tree



Note: The tip labels of the row tree are different to the row names of the row data

The column tree



Discussion points

- Framework that explicitly builds in **type** and **state**
- Differential (relative) abundance bears similarity to RNA-seq DE: cluster cell counts
- Differential state also bears similarity to DE, but on aggregates (microarrays for CyTOF, RNA-seq for scRNA-seq)
- Linear modelling approaches always worth their weight in gold .. flexibility for experimental designs .. fast, sensitive to find many types of changes

- Can we get everything from aggregates? (We are still finding out)
- How to best use batch correction methods, cell type assignment methods

- Use of trees (TreeSummarizedExperiment and friends)

- Code/data is available for basically everything we do

Statistical Bioinformatics Group, IMLS, UZH

Roche

scRNAseq:
Dheeraj Maholtra
Daniela Calini

CyTOF:
Carsten Krieg
Burkhard Becher
Mitch Levesque

CATALYST:
Helena Crowell
Vito Zanotelli
Stephane Chevrier
Bernd Bodenmiller

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