Projects

Reminders:

- project “plan” by 4.12. (a few bullet points)
- due **12.01.2024** 18.00

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

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

Exercise notes

```r
pvalue <- lrt$table$PValue
pvalue_full <- lrt_full_knockout$table$PValue


threshold <- 0.1
knockout <- as.numeric(pvalue < threshold)
knockout_full <- as.numeric(pvalue_full < threshold)
```

```r
17
18  de_unadj <- rownames(subset(lrt_unadj_grp, PValue <= 0.05))
19  de_adj <- rownames(subset(lrt_adj_grp, PValue <= 0.05))
20
```

```r
1  # Extract all differentially expressed genes with p-value < 0.05
2  de_gene_limma <- de_gene_limma[de_gene_limma$P.Value < p_val, ]
```

Exercise notes

P-values!

If you learn only 1 thing in STA 426: DO NOT MAKE CUTOFFs on raw P-values (unless Bonferroni)

[decideTests(.., p.value = 0.05)]

```
17
18  de_unadj <- rownames(subset(lrt_unadj_grp, PValue <= 0.05))
19  de_adj <- rownames(subset(lrt_adj_grp, PValue <= 0.05))
20
```

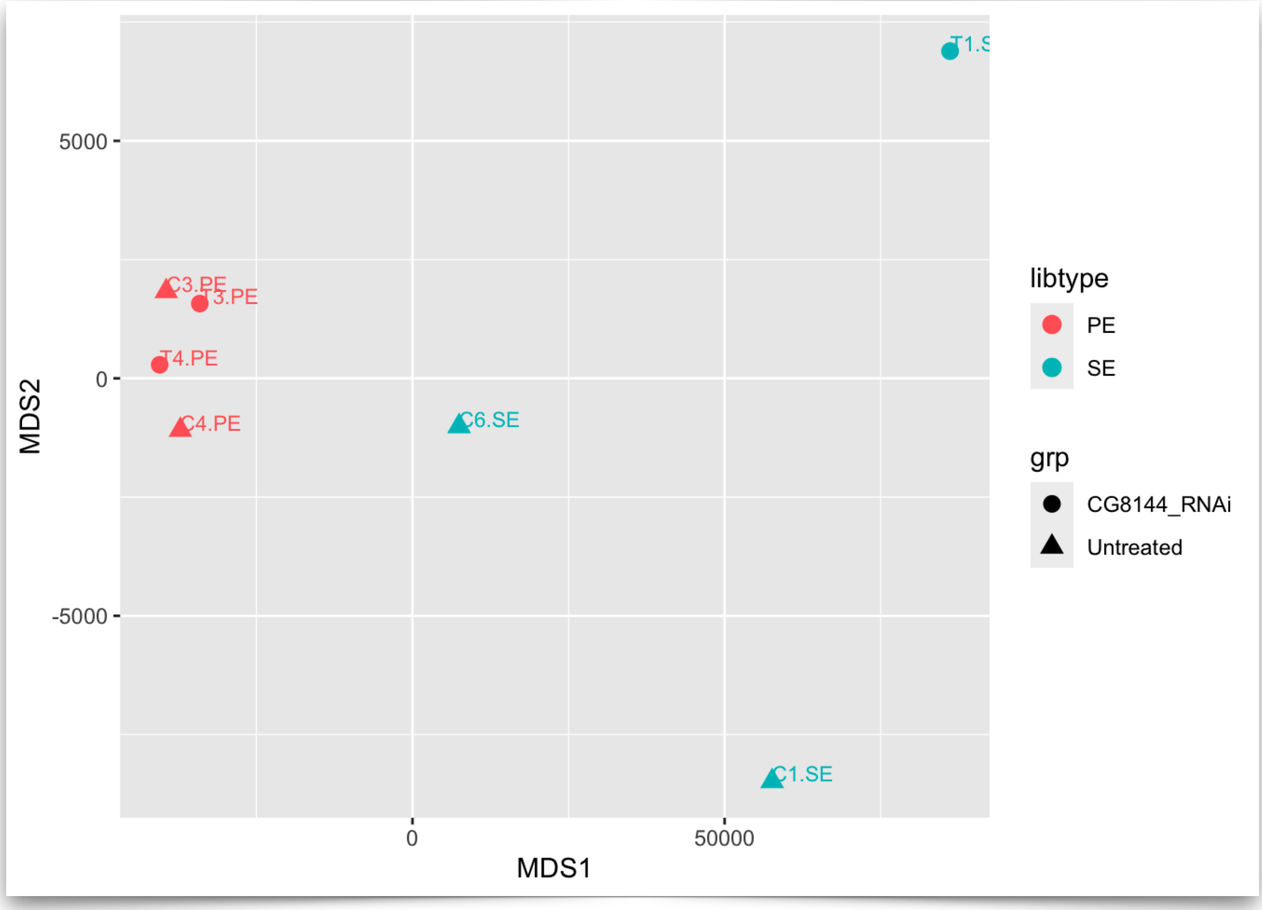
```
pvalue <- lrt$table$PValue
pvalue_full <- lrt_full_knockout$table$PValue


threshold <- 0.1
knockout <- as.numeric(pvalue < threshold)
knockout_full <- as.numeric(pvalue_full < threshold)
```

```
1  # Extract all differentially expressed genes with p-value < 0.05
2  de_gene_limma <- de_gene_limma[de_gene_limma$P.Value < p_val, ]
3  # Sort de_gene_limma by P-values
4  de_gene_limma <- de_gene_limma[order(de_gene_limma$P.Value), ]
5
6  head(de_gene_limma)
```

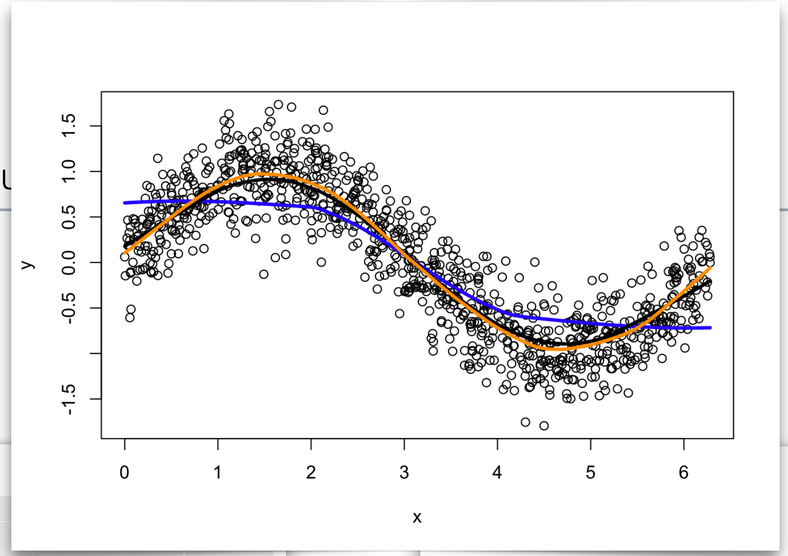
Exercise notes

?plotMDS —> Plot samples on a two-dimensional scatterplot so that distances on the plot approximate the typical log2 fold changes between the samples.

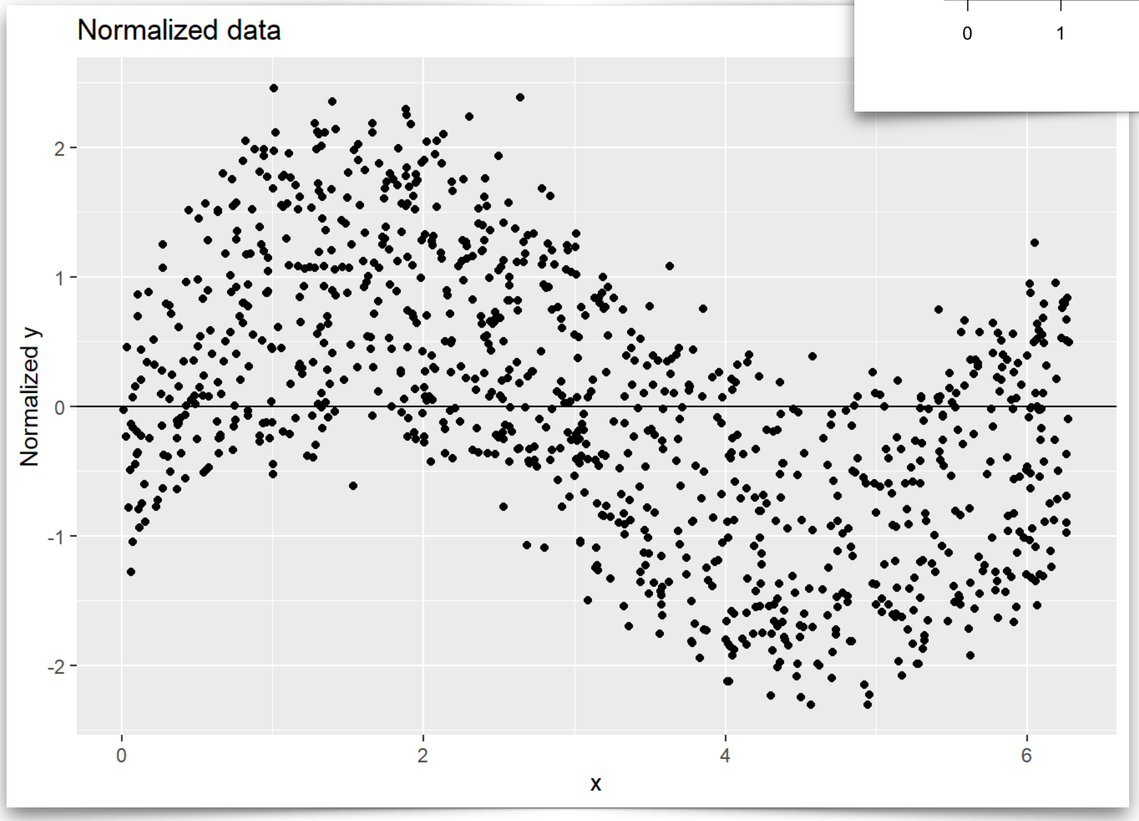
MDS plots

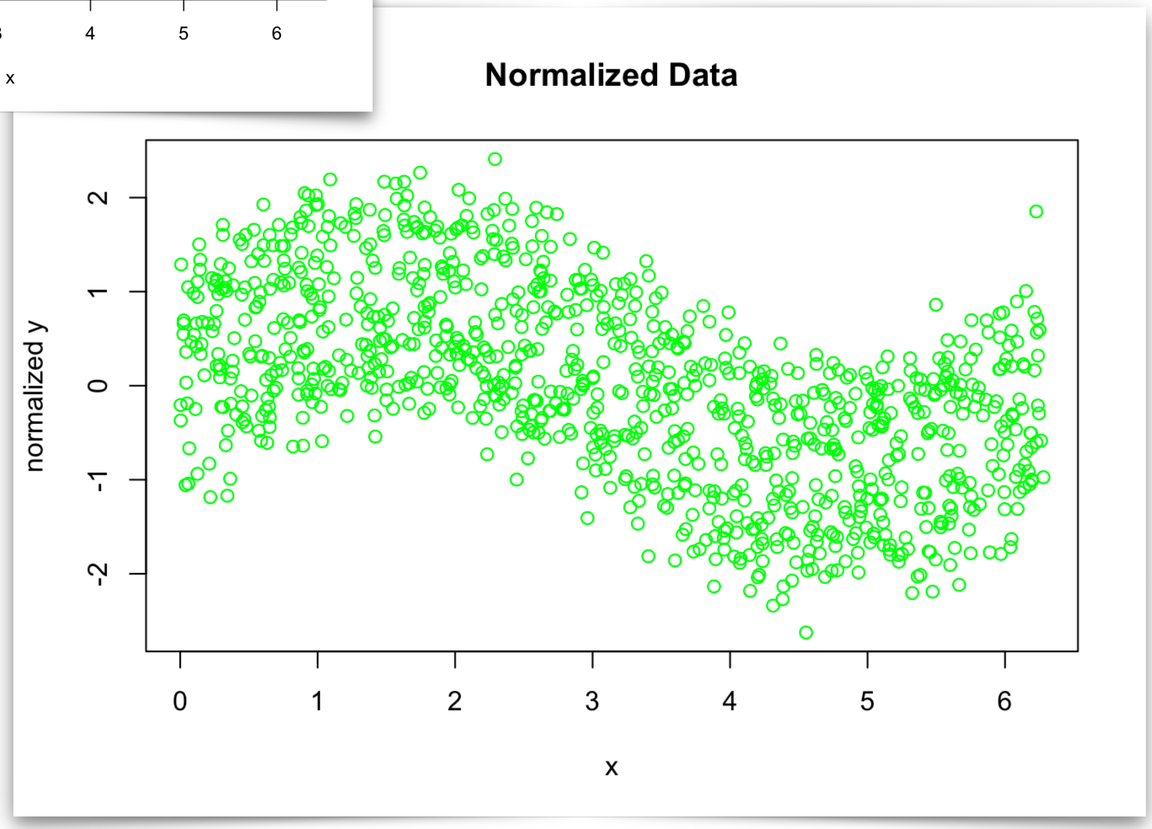


Exercise notes



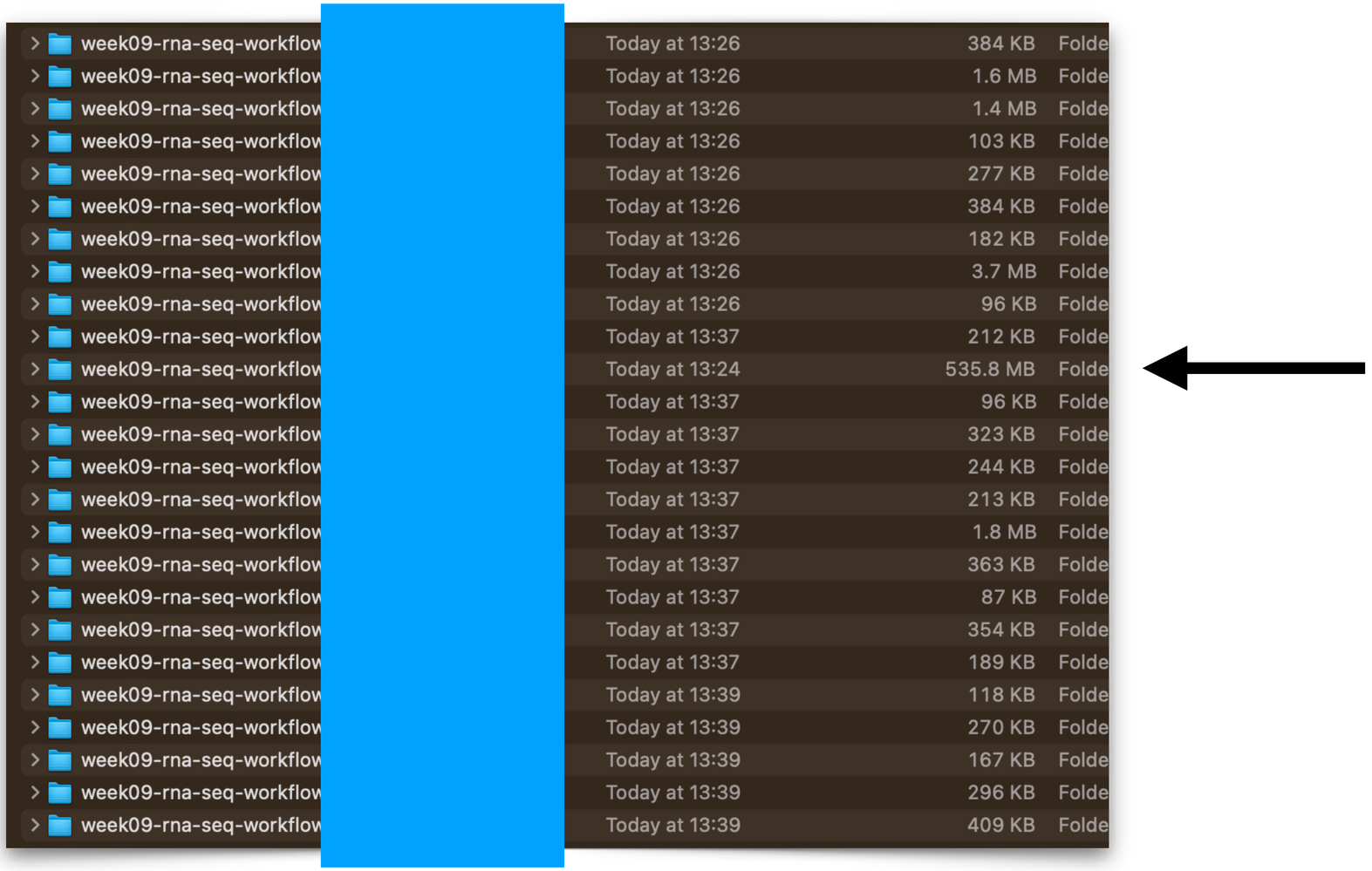
<— start with



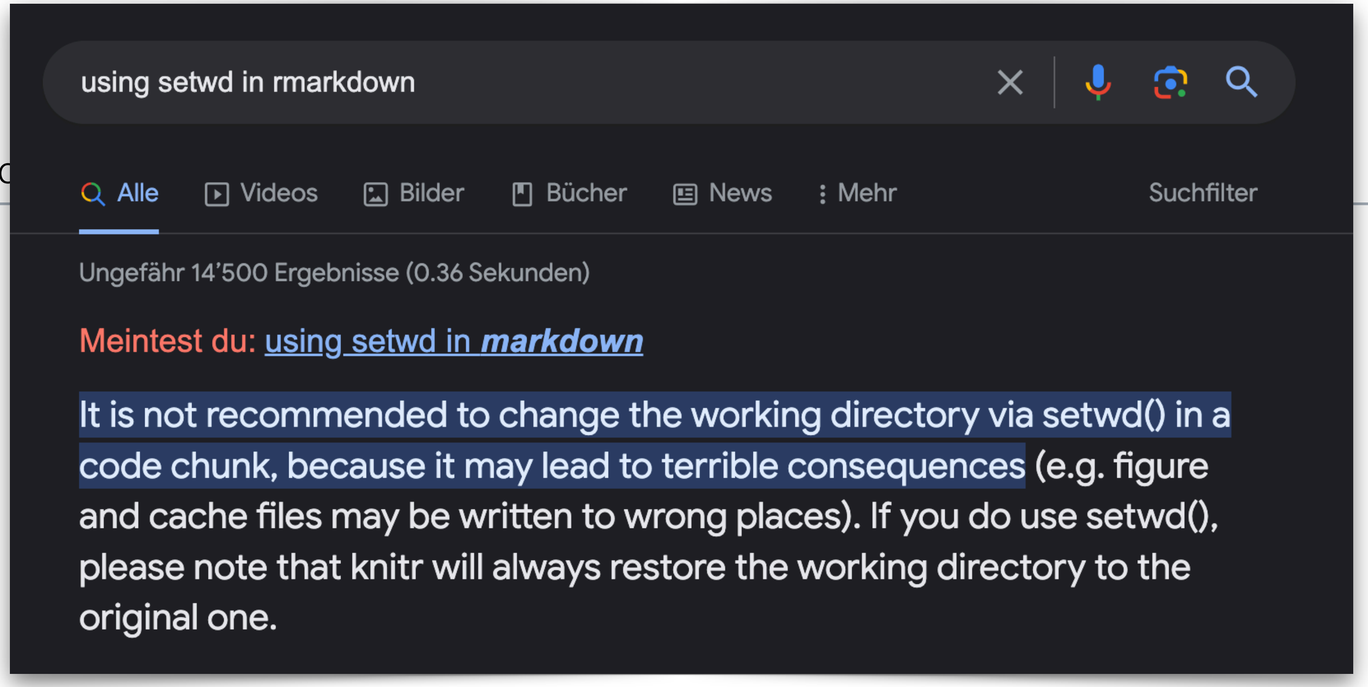


Exercise notes

git is meant for text files (and generally “small” ones); git ifs for larger files



Exercise notes



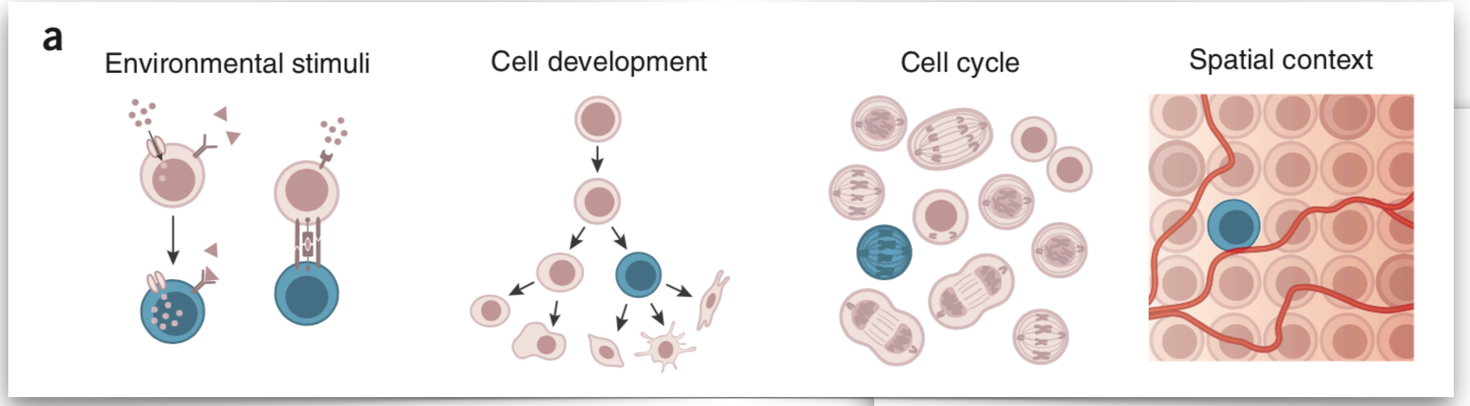

```
(samples <- read.table("/Use          Desktop/STA_E8/data/samples.txt", header=TRUE,
                       row.names=5, stringsAsFactors=FALSE))
```

Probably ok, but not great

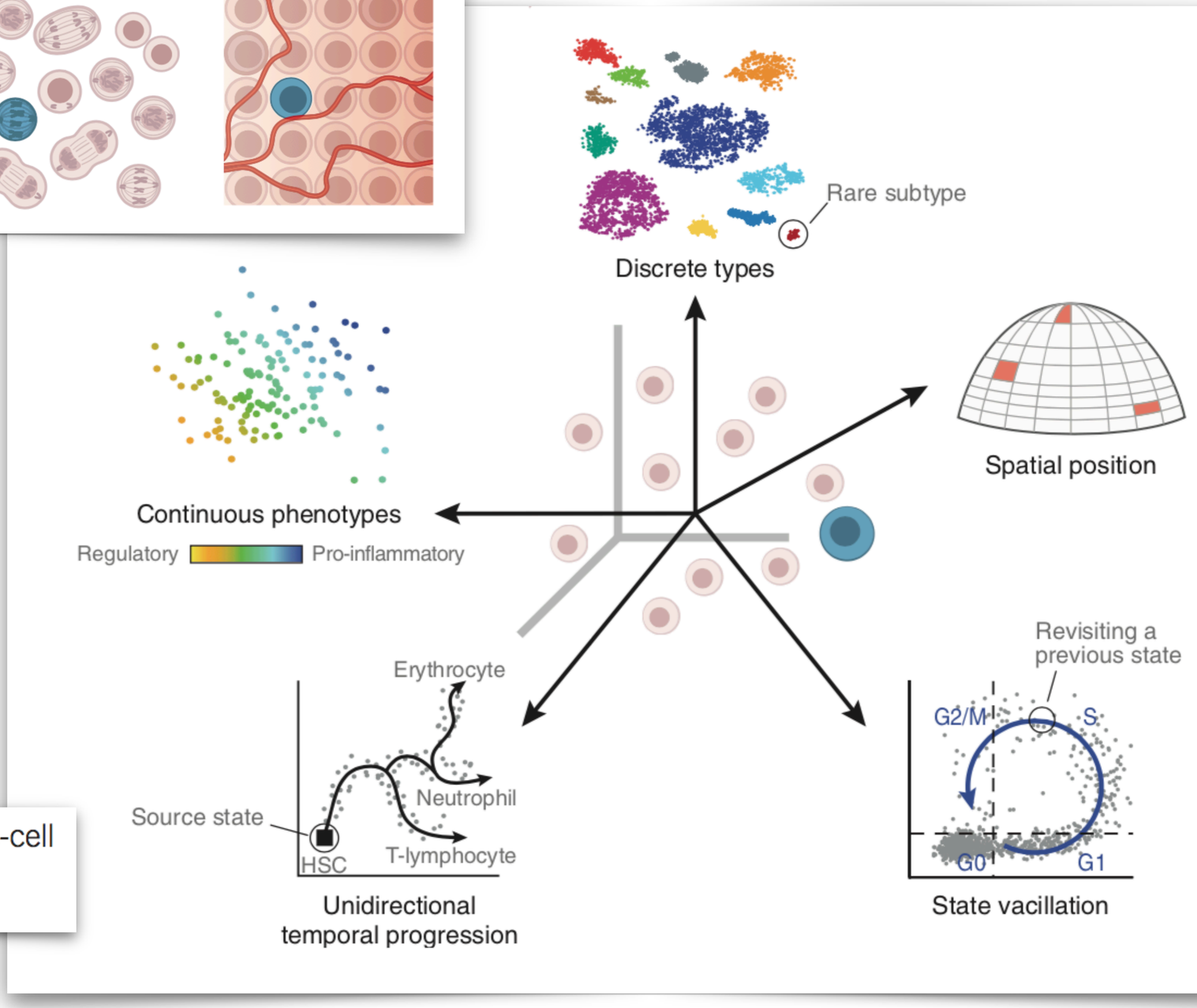
```
setwd('/Users        n/Desktop/STA_E8/data/')
counts <- readDGE(samples$countfile)$counts
```

Probably not ok

My suggestion: use R projects and put data in a /data/ directory, use "relative" paths



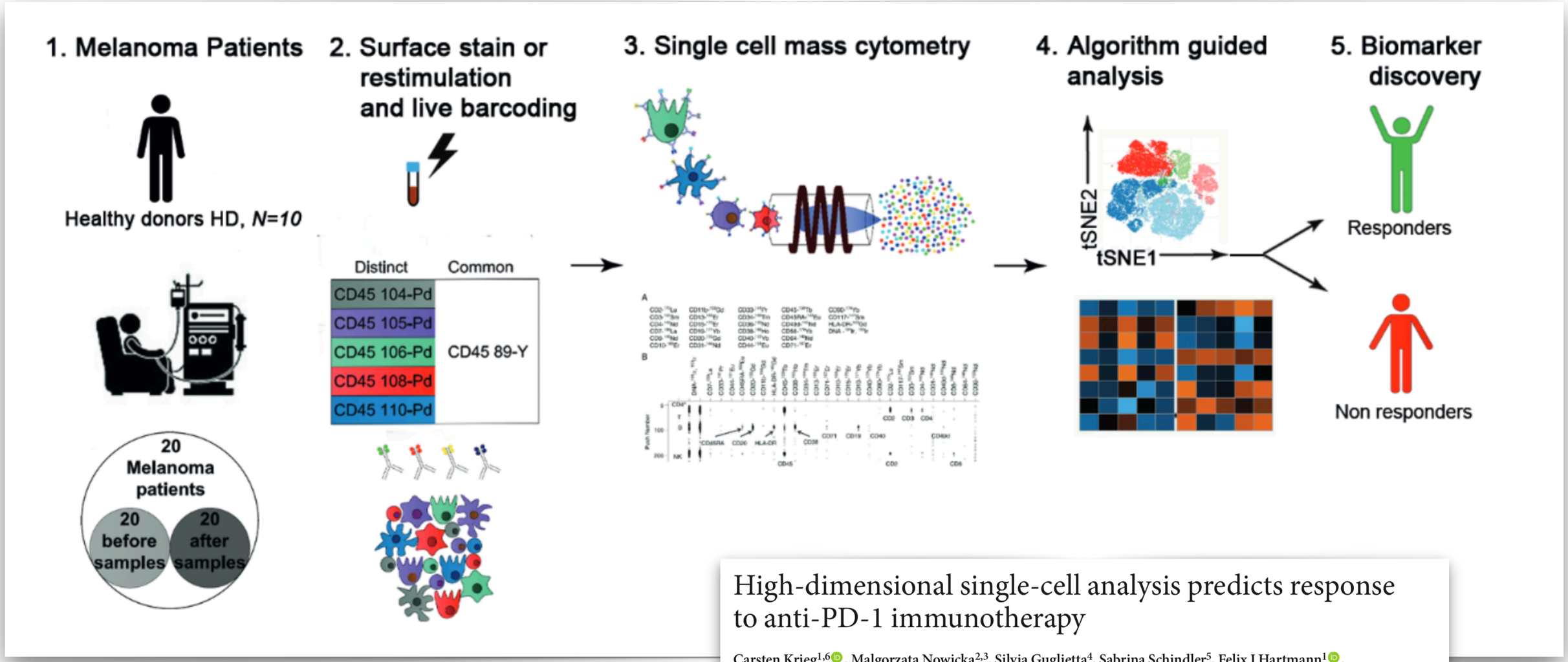
Applications



Revealing the vectors of cellular identity with single-cell genomics

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

Motivation for differential analysis of single cell (cytometry) data: *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]

Flow cytometry

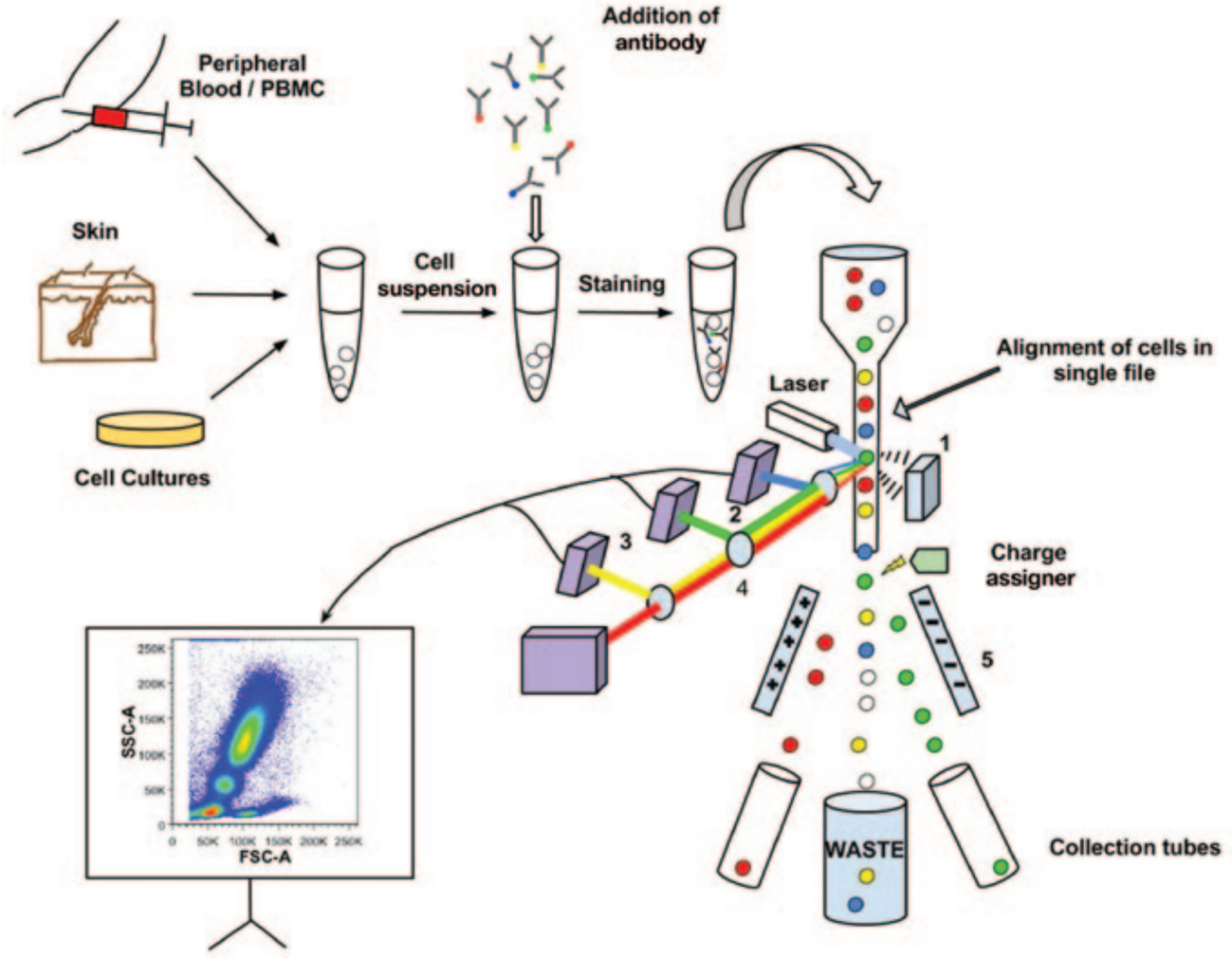
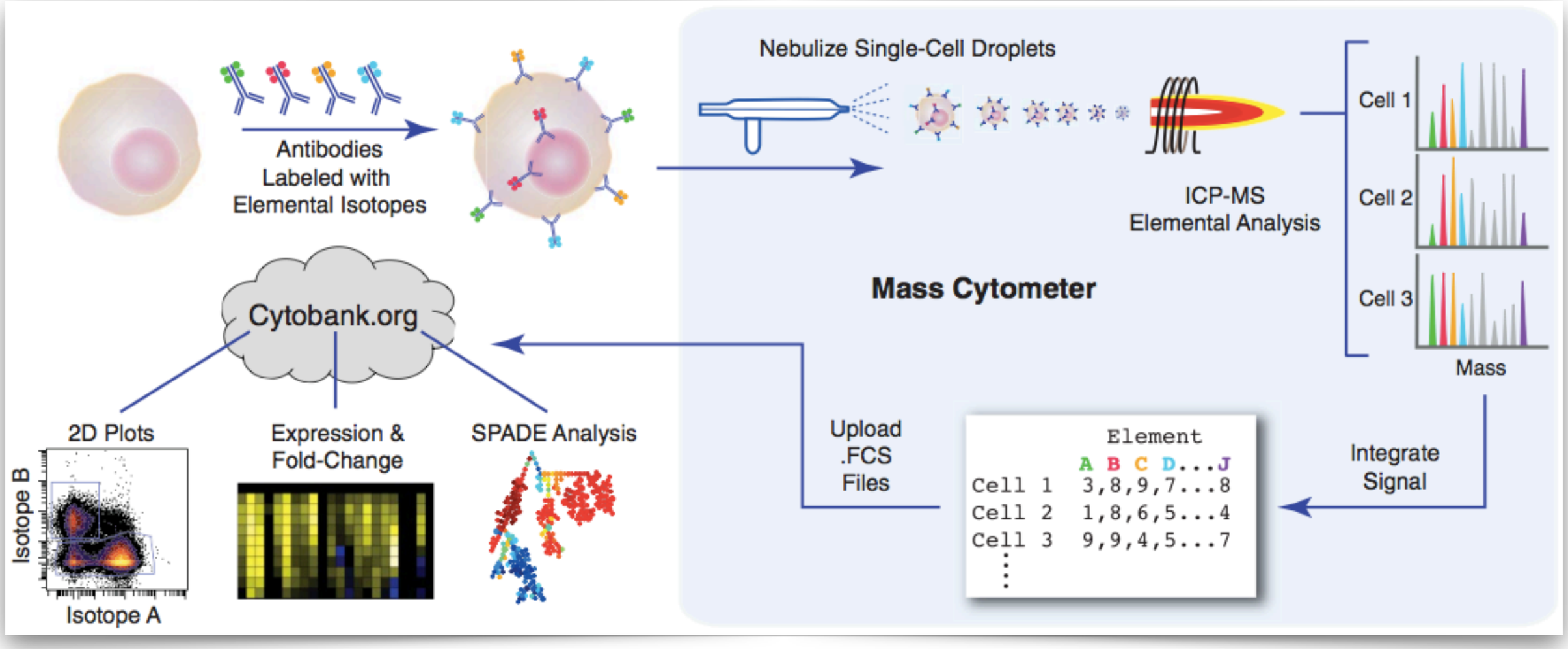
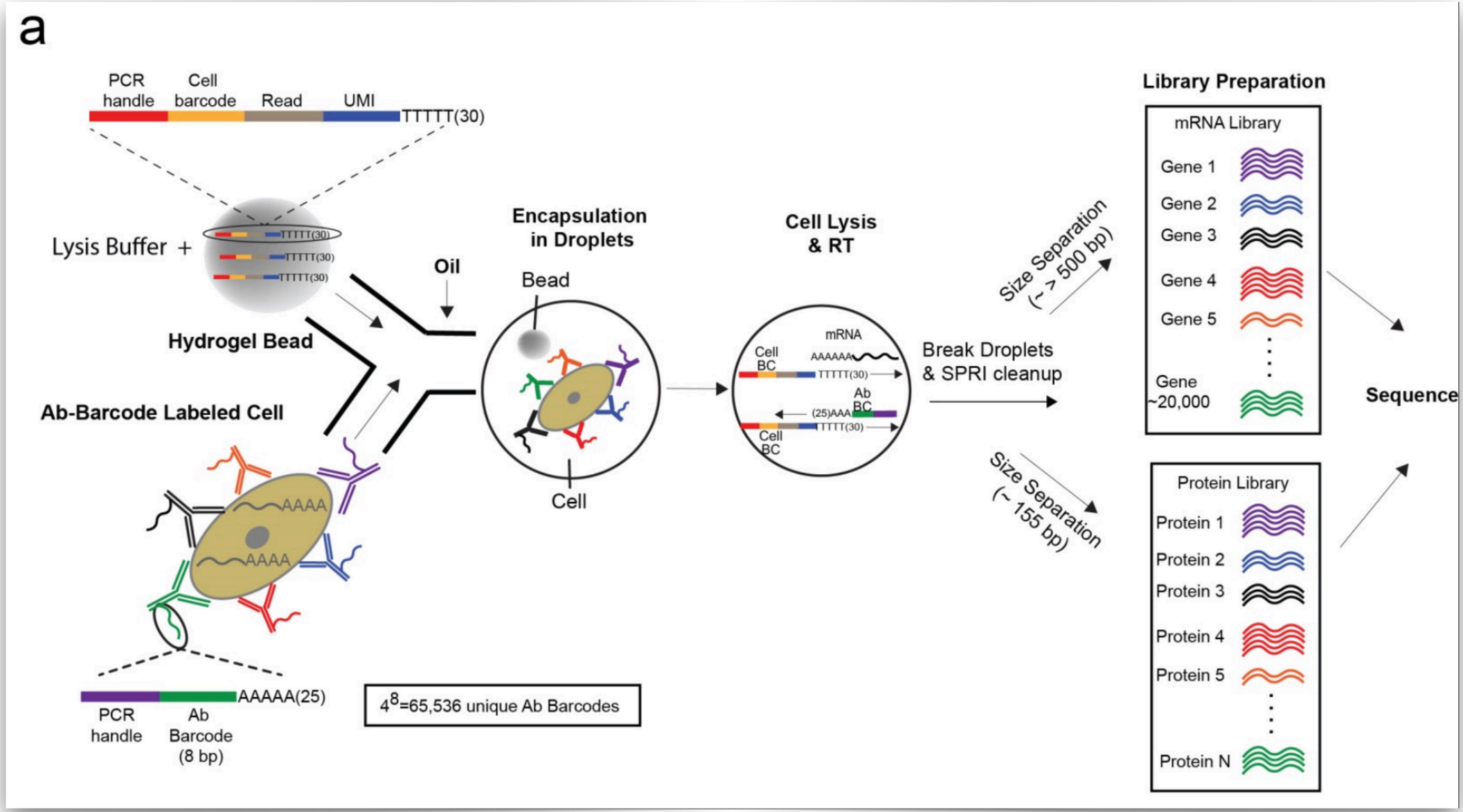


Figure 1. Schematic representation of a flow cytometer. For details please see text. (1) Forward-scatter detector, (2) side-scatter detector, (3) fluorescence detector, (4) filters and mirrors, and (5) charged deflection plates.

Mass cytometry



REAP-seq / CITE-seq

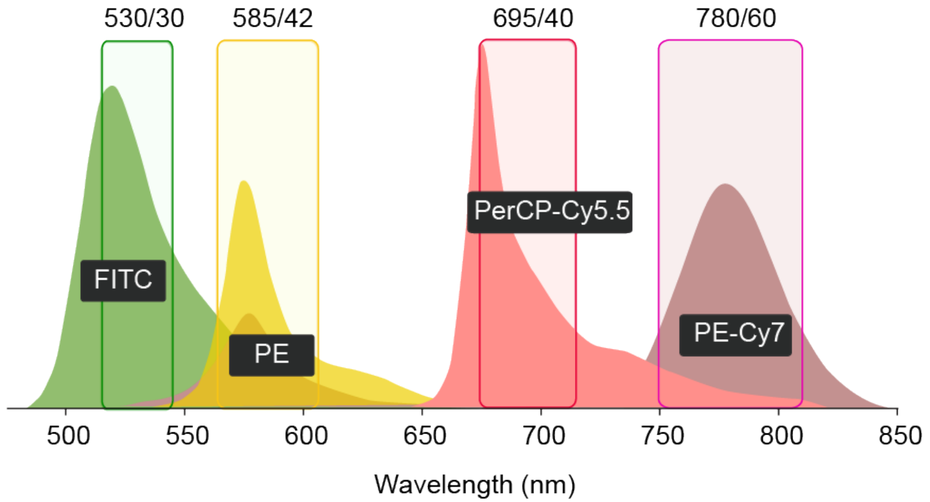


Spectral overlap vs. spillover

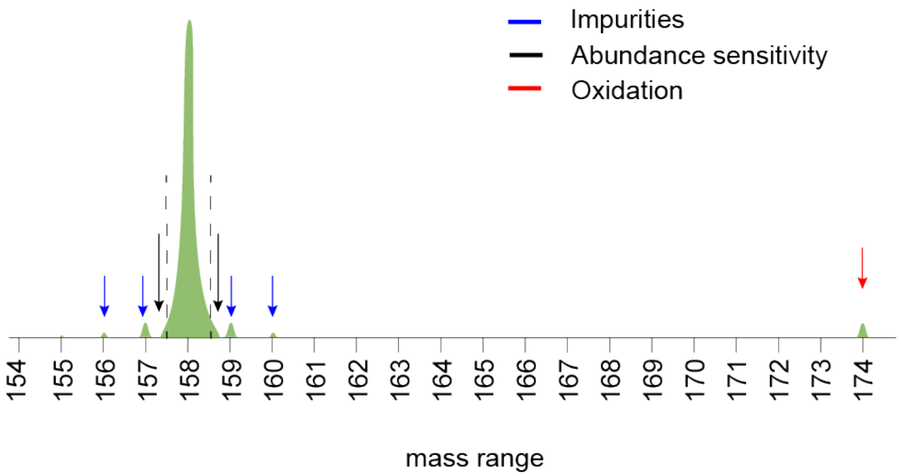
- CyTOF = increase in the number of parameters + massive decrease in spectral overlap
- but, still <u>three sources of signal overlap</u>:

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

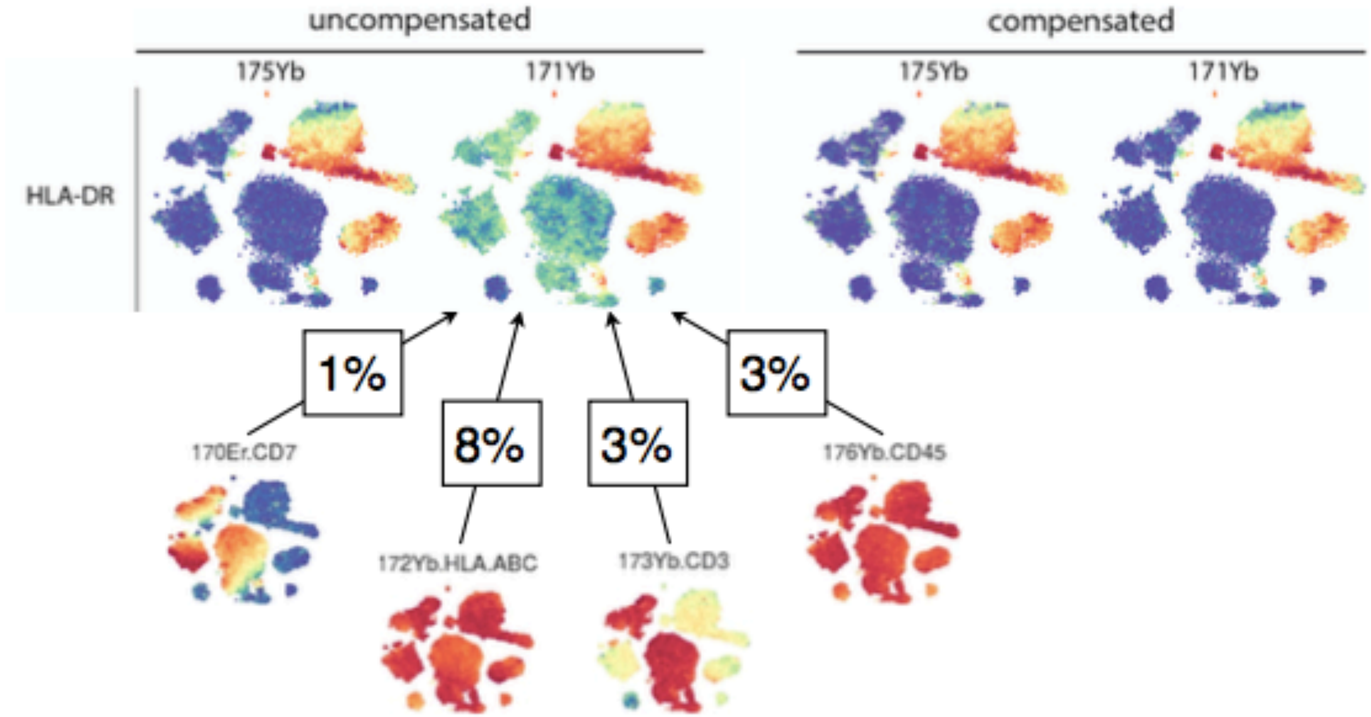
The metals that are sold as part of antibody labeling kits are of very high purity (98% and higher in most cases). As a practical matter, this means that "compensation" analogous to fluorescent antibodies is not needed, as most of the signal will be of the specified mass, with little to no signal at "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

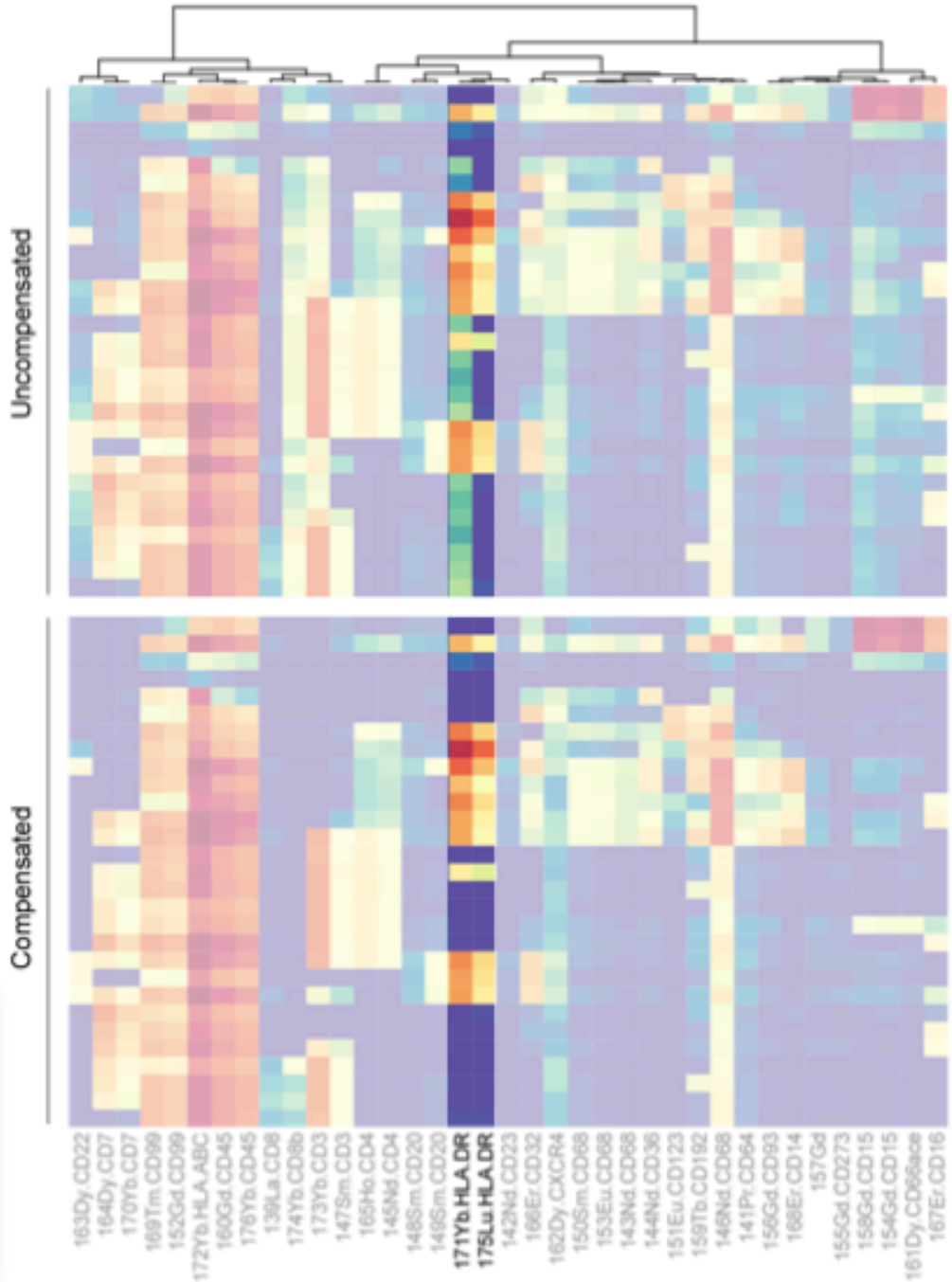
Correction of spillover artefacts on a 36ab panel



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

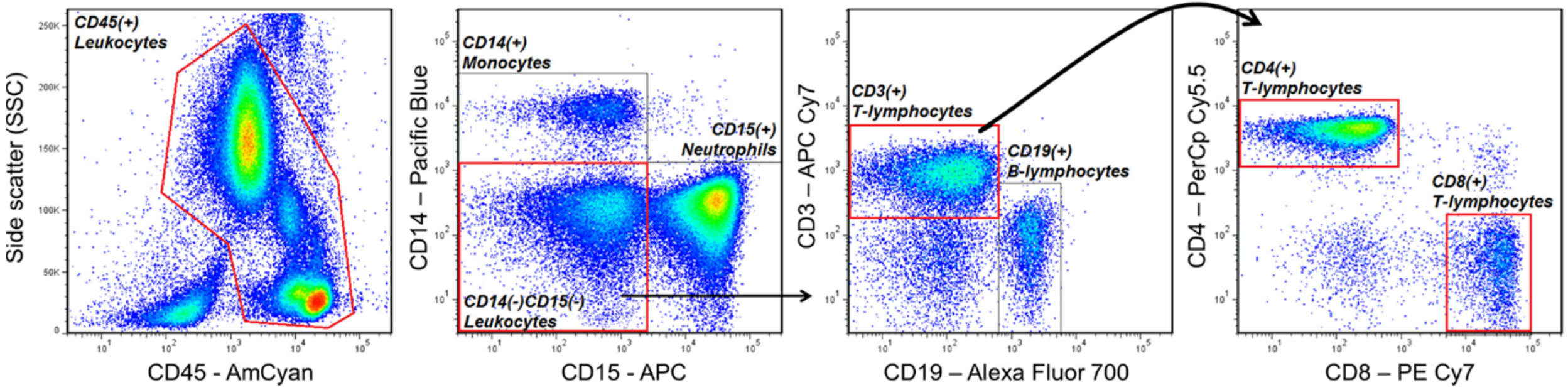
Compensation of Signal Spillover in Suspension and Imaging Mass Cytometry

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



Historically, manual gating was (is) used for cytometry data

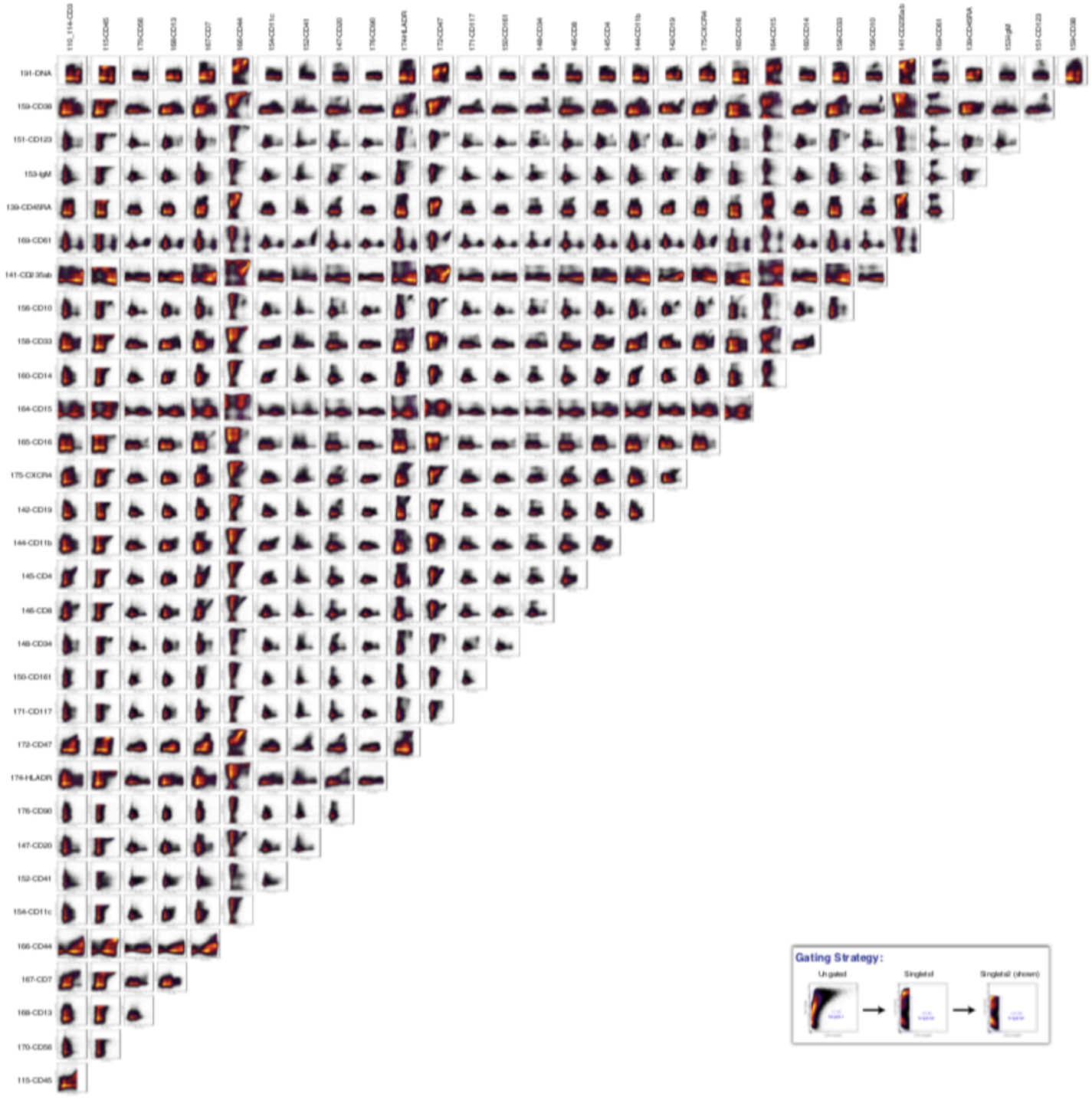
Identification of cell populations



Verschoor et al. (2015)

Manual gating

Not feasible in high-dimensional data



Bendall et al. (2011), Supp.

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

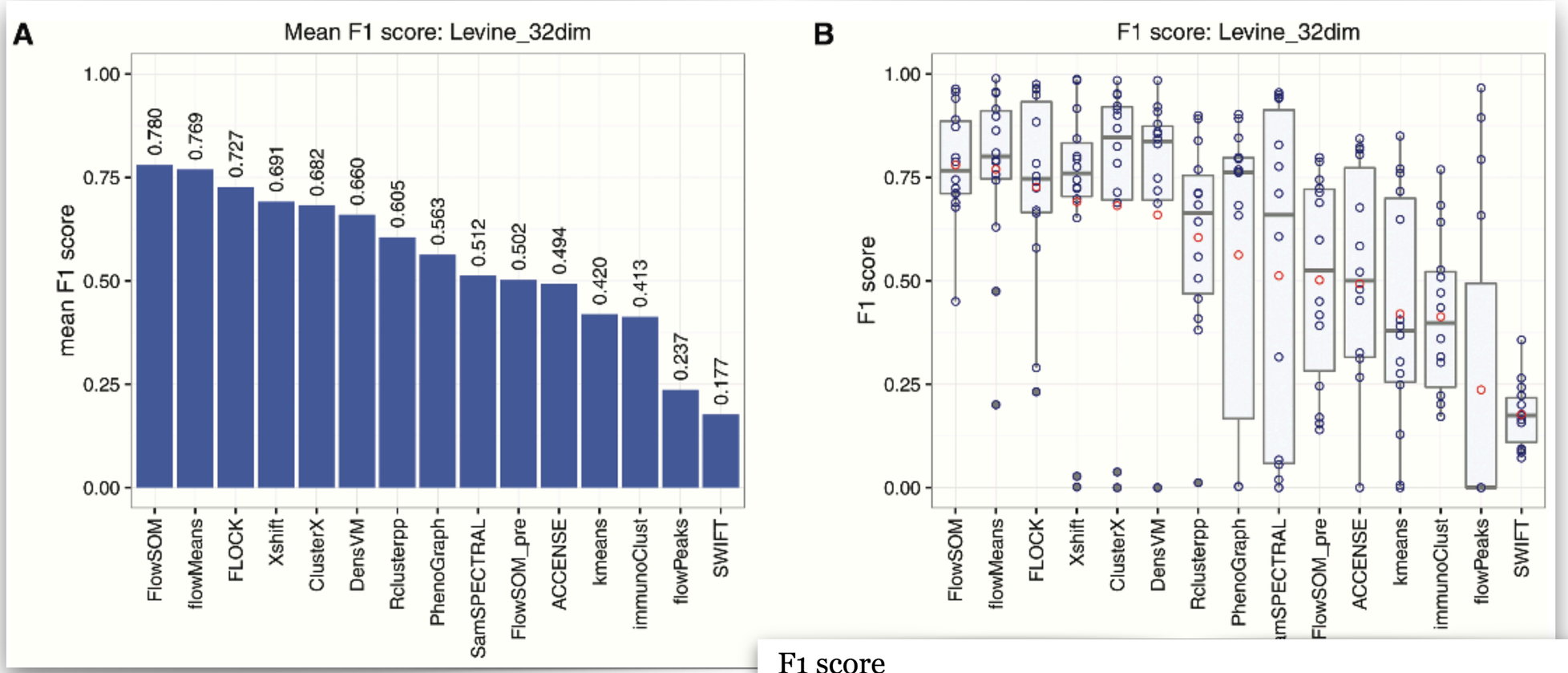
EDITOR'S CHOICE

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]*

Comparison of clustering methods



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]*

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:

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

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

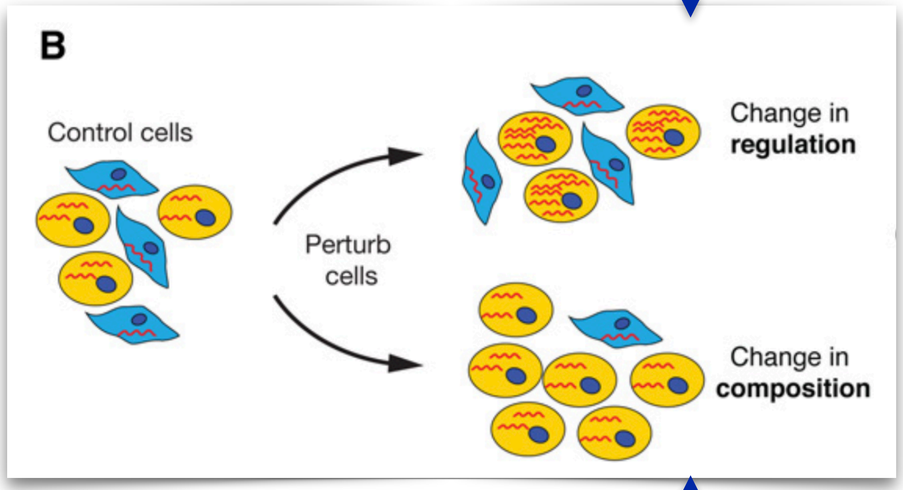
Box 1 The many facets of a cell's identity

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

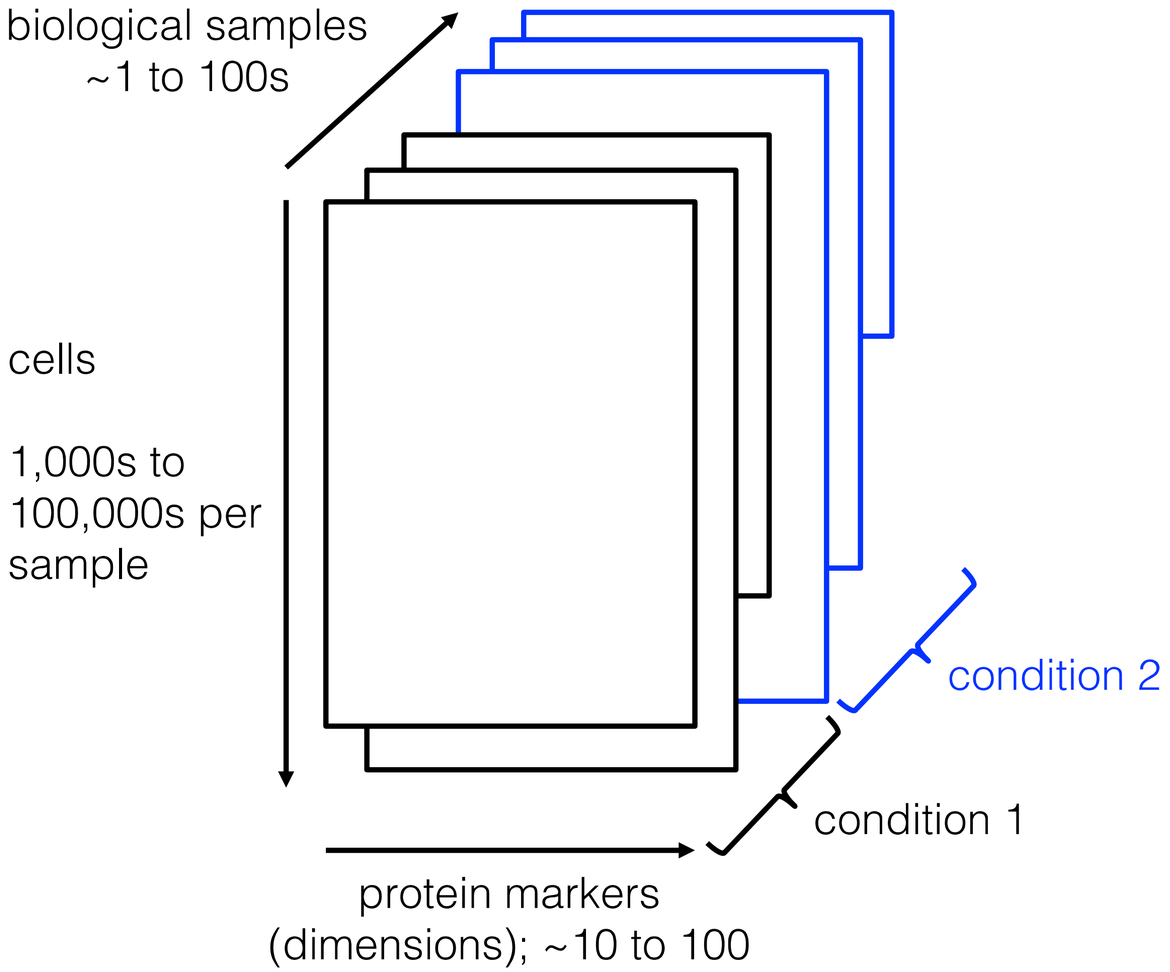


Differential abundance analysis

Data structure and differential analysis

Two types of differential analysis

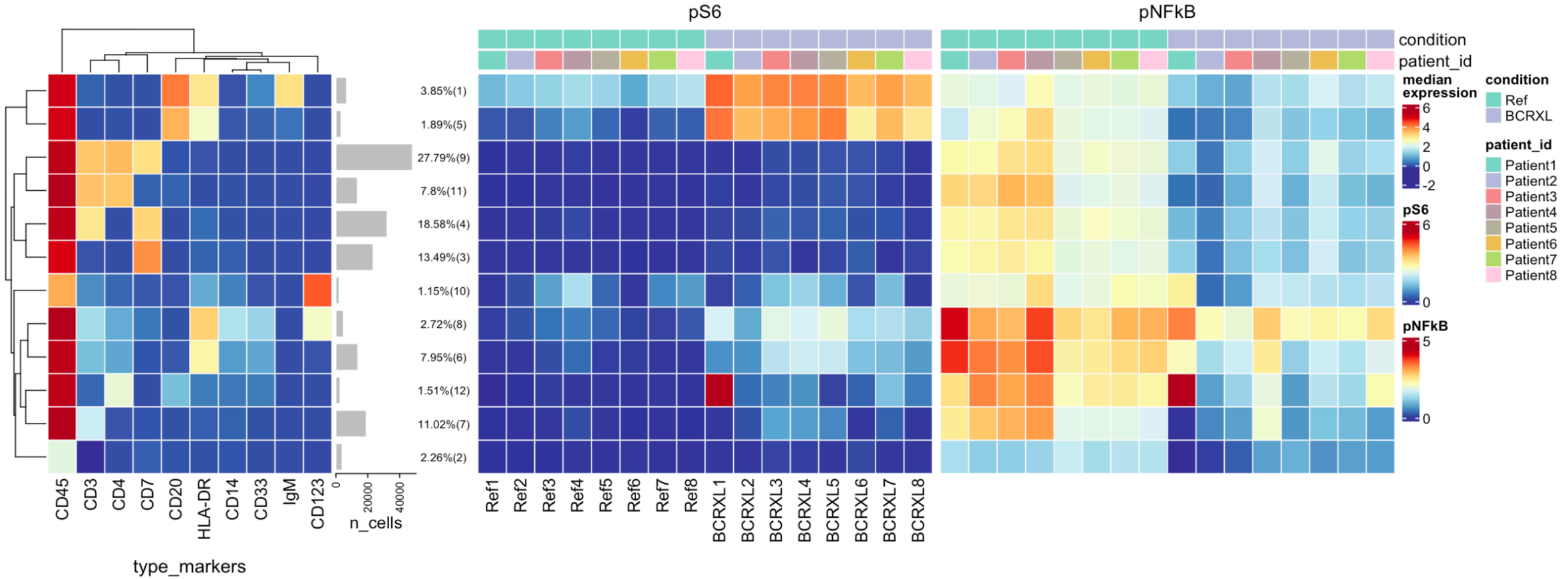
- **differential abundance** (DA) of cell populations
- **differential states**
  e.g., differential expression of functional proteins (e.g., signaling) within cell populations



Canonical example: differential state analysis
Explicit split between “type” and “state” markers

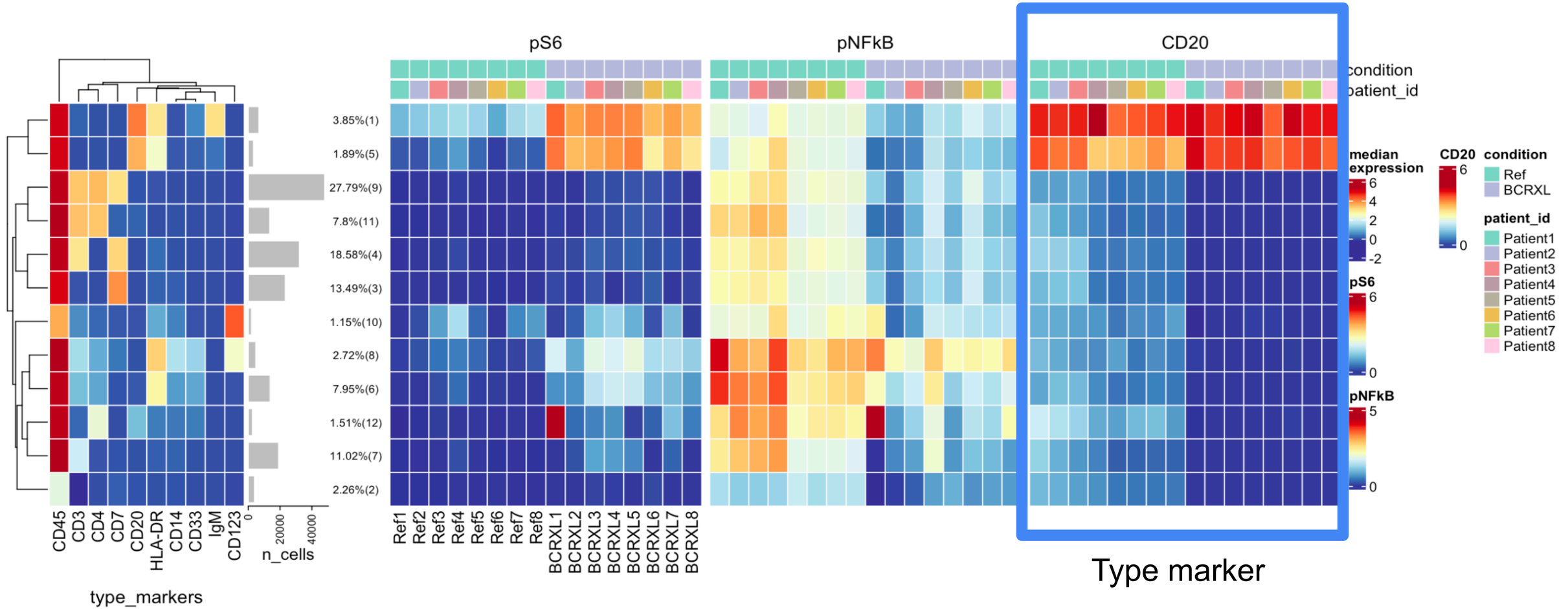
10 type markers

14 state markers (2 shown)



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

With the expectation that type markers are constant



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

Key elements of CyTOF workflow

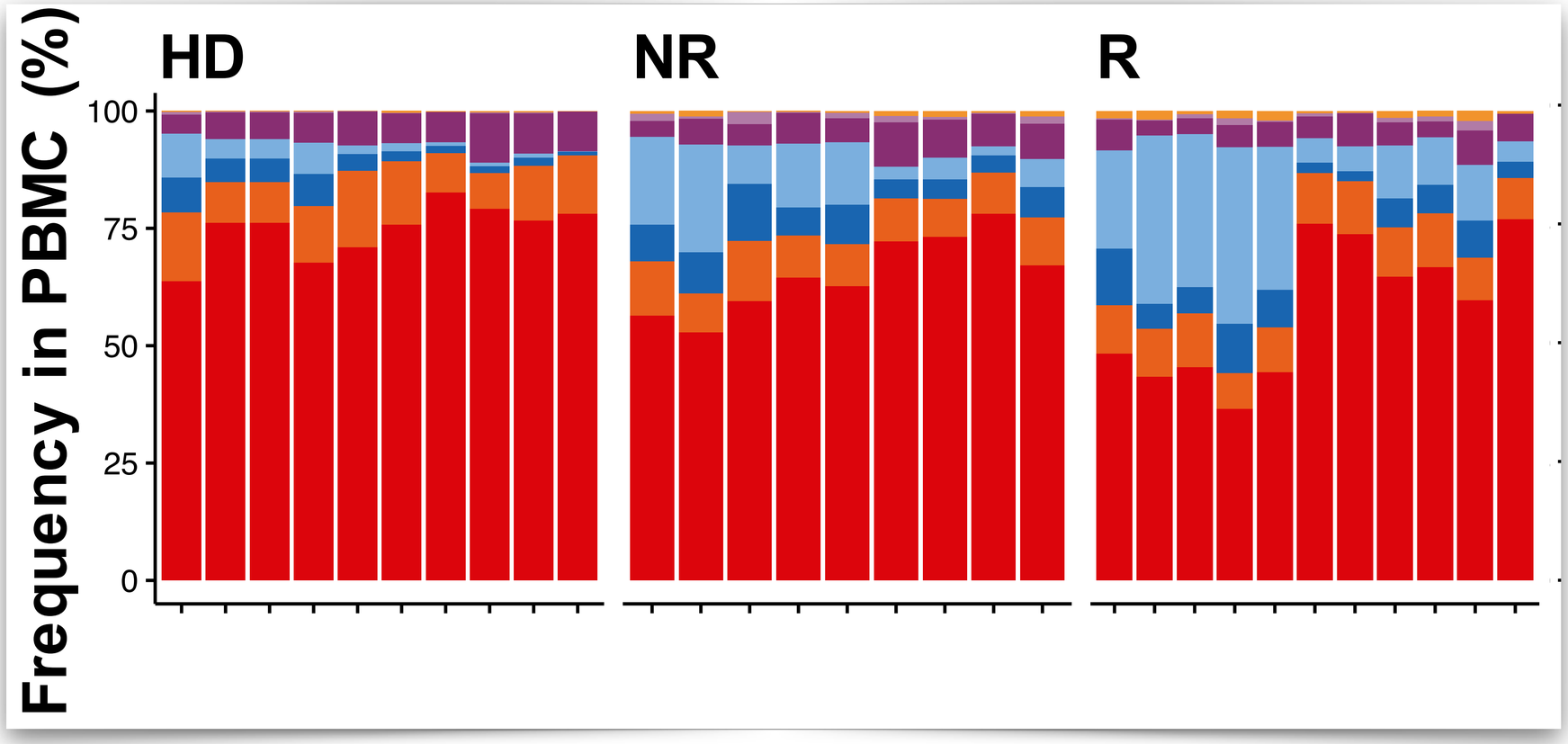
- Exploration of various data aspects at each step
- Separation of **type** and **state** markers
- Put all samples together and cluster (FlowSOM or other)
- Optional: manually merge clusters (via visualizations: heatmaps, low dimensional projections)
- Differential abundance analysis (count-based model,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

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).

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

Manual merging of cell populations based on phenotypes

Generalized linear mixed models (differential abundance)

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

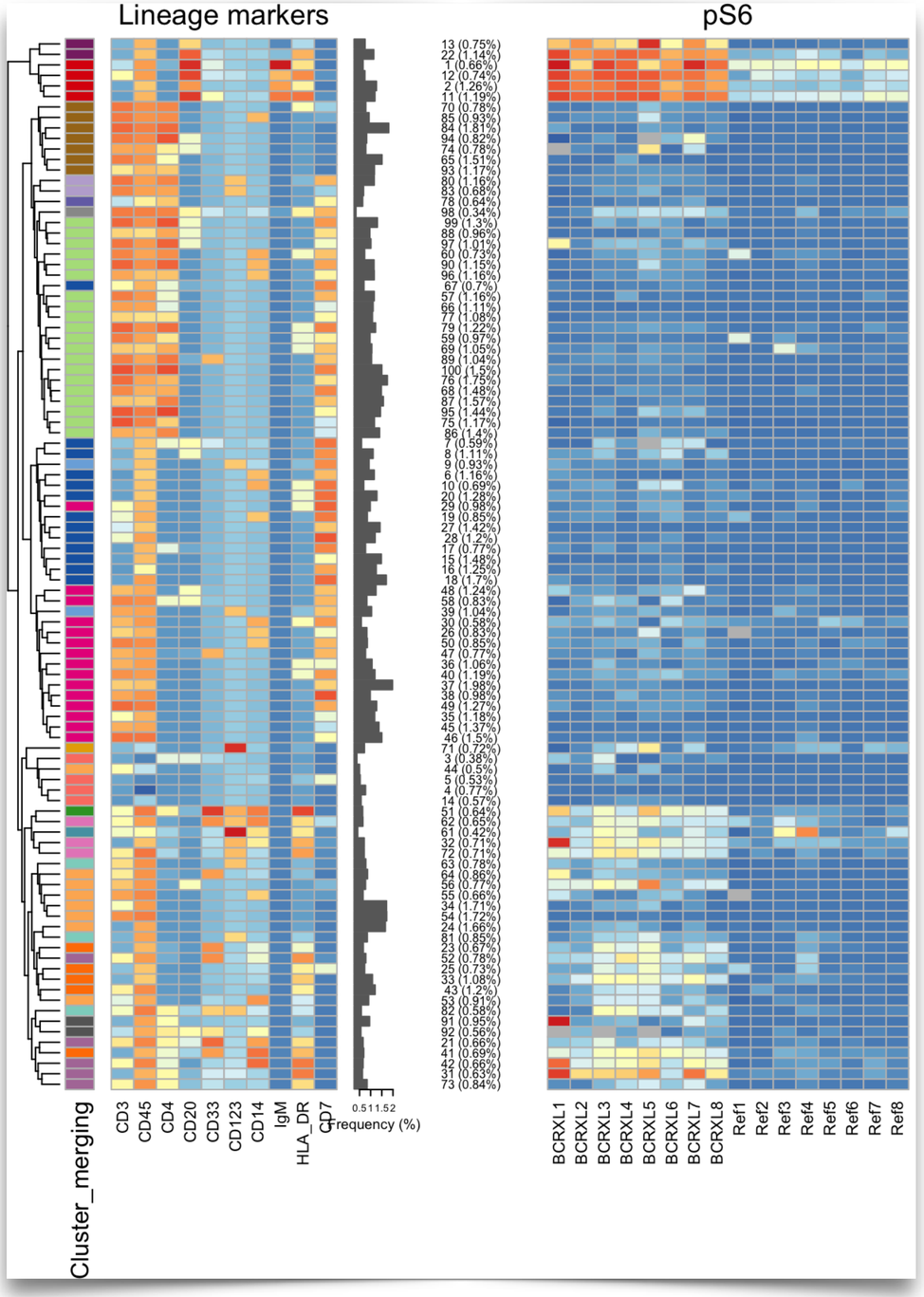
Linear mixed models (differential expression within populations)

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

Part 2: subpopulation-specific differential analyses

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

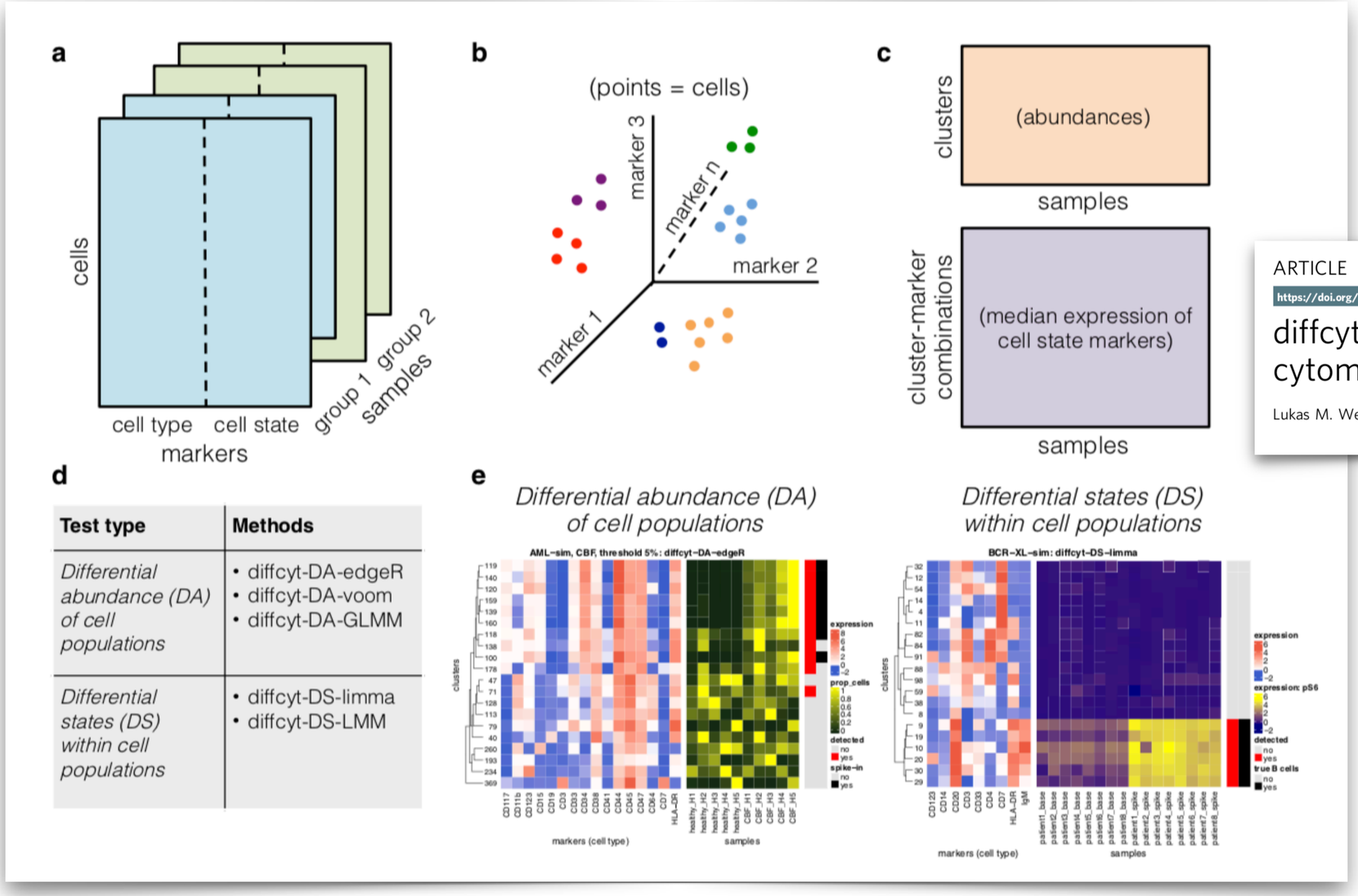
median lineage marker signal by cluster



median functional marker signal by sample

data from Bodenmiller et al. 2012

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