Workshop

CyTOF workflow: differential discovery in high-throughput high-dimensional cytometry datasets

Malgorzata Nowicka

University of Zurich

BioC 2017: Where Software and Biology Connect

Boston, 28 July 2017

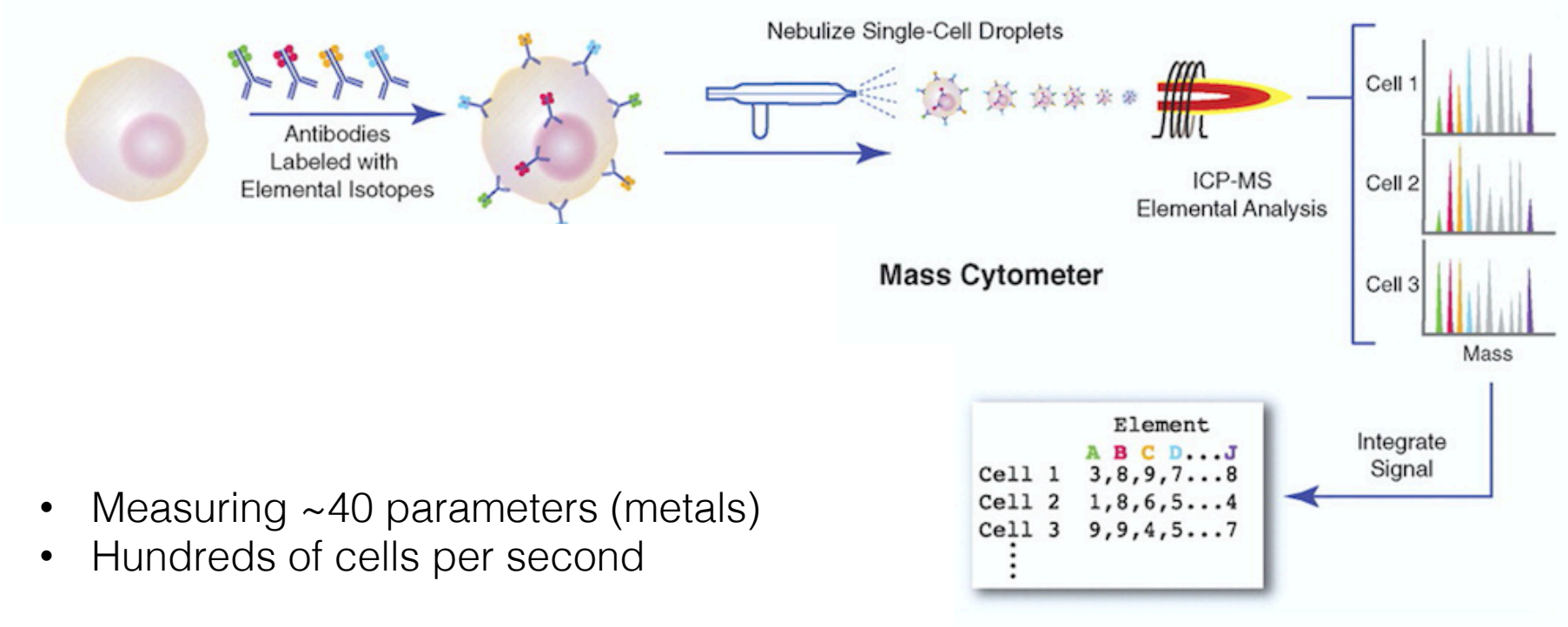
Have you analyzed CyTOF or flow cytometry data before?

My approach to this workshop

- This is just a shorter version of the original workflow
- The text is basically the same but maybe try to not focus on it so much, as you can carefully read it later
- Ideally, I would do the life coding but because of the time constrains, I will:
  - explain all the code in the vignette (html file)
  - copy-paste to R, to follow with you and have the opportunity to modify some things if needed

Introduction

CyTOF (mass cytometry) experiment



- Measuring ~40 parameters (metals)
- Hundreds of cells per second

Figure adapted from Bendall et al. Science 2011

Introduction

- High-dimensional cytometry because it is higher than before (from ~12 to ~40)
- We refer to this differential approach as “classic”
  - Common strategy:
    - identify cell populations of interest by manual gating or automated clustering
    - determine which of the cell subpopulations or protein markers are associated with a phenotype of interest using statistical tests
- Other approaches:
  - Citrus (Bruggner et al. 2014)
  - CellCnn (Arvaniti and Claassen 2016)
- Hybrid approaches (thanks to the modularity)
- This workflow can serve as a template

Overview

1. Data description
2. Data pre-processing (CATALYST pckg)
3. Data import – flowCore pckg
4. Data transformation - arcsinh with cofactor 5
5. Spot checks
   - MDS plots (limma pckg)
6. Cell clustering with FlowSOM and ClusterConsensusPlus pckgs
   - overclustering + manual merging
7. Visualization of the clusters with
   - t-SNE - Rtsne pckg
   - heatmaps - pheatmap pckg
   - ggplot pckg
8. Differential analysis with linear mixed models (lme4 and multcomp pckgs)
   - diff. cell population abundance
   - diff. marker expression

Data description

- Subset of data from the Bodenmiller et al. 2012 study; used mass cytometry to measure PBMC's
  - from 8 different patients
  - after 11 different stimulation conditions as well as an unstimulated "Reference" state
- Here, we use the samples that correspond to the BCR-crosslinking and Reference
- For each sample, 10 cell surface markers and 14 signaling markers were measured

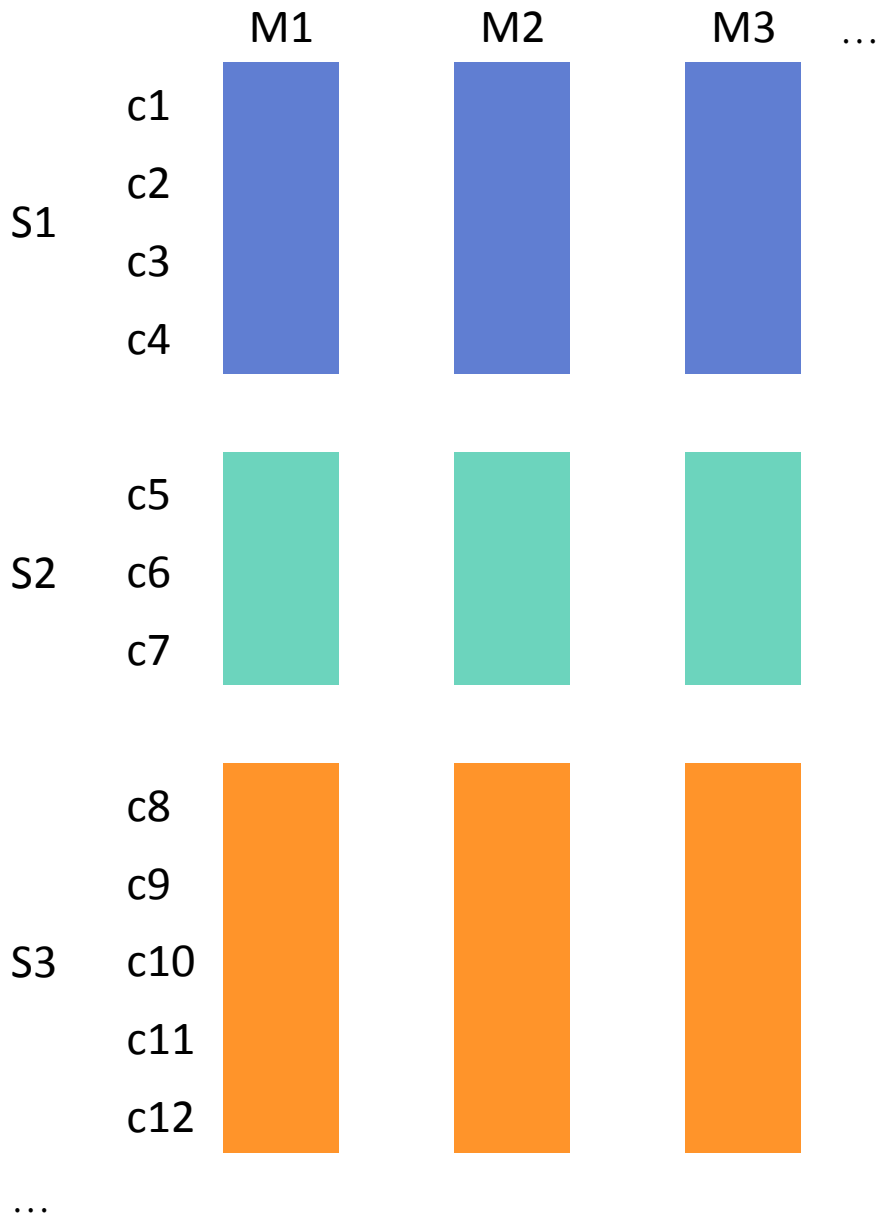
Data pre-processing

- The pre-processing steps in the Bodenmiller data, as we can download it, included removal of debris and de-barcoding
- In general, pre-processing steps may involve:
  - normalization using bead standards
  - de-barcoding
  - compensation
- CATALYST pckg

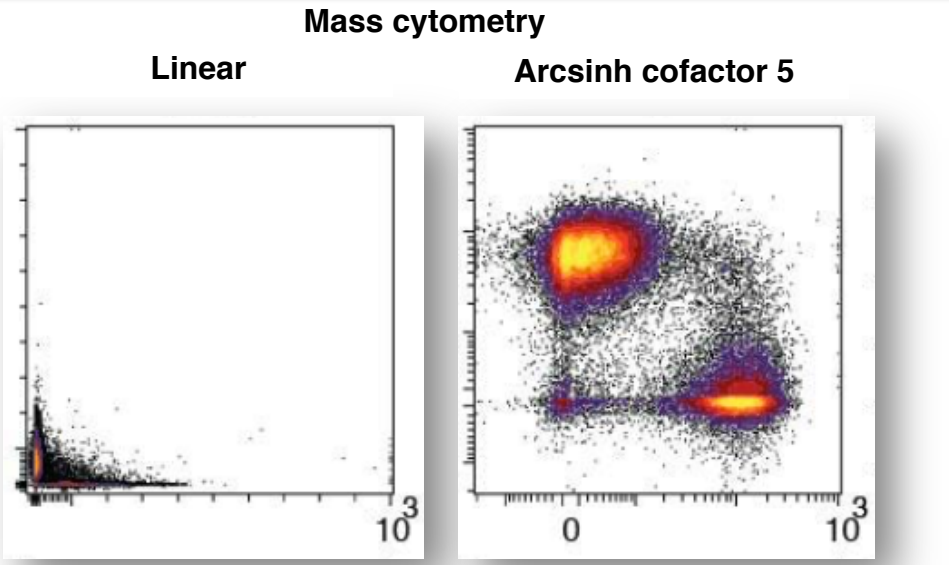
Data import

- All the data is available at the Robinson Lab server http://imlspenticton.uzh.ch/robinson_lab/cytofWorkflow/
- Downloading using the download.file()
  - Metadata
  - FCS files
  - Panel file
  - Files with cluster merging

Data import: expr



Data transformation



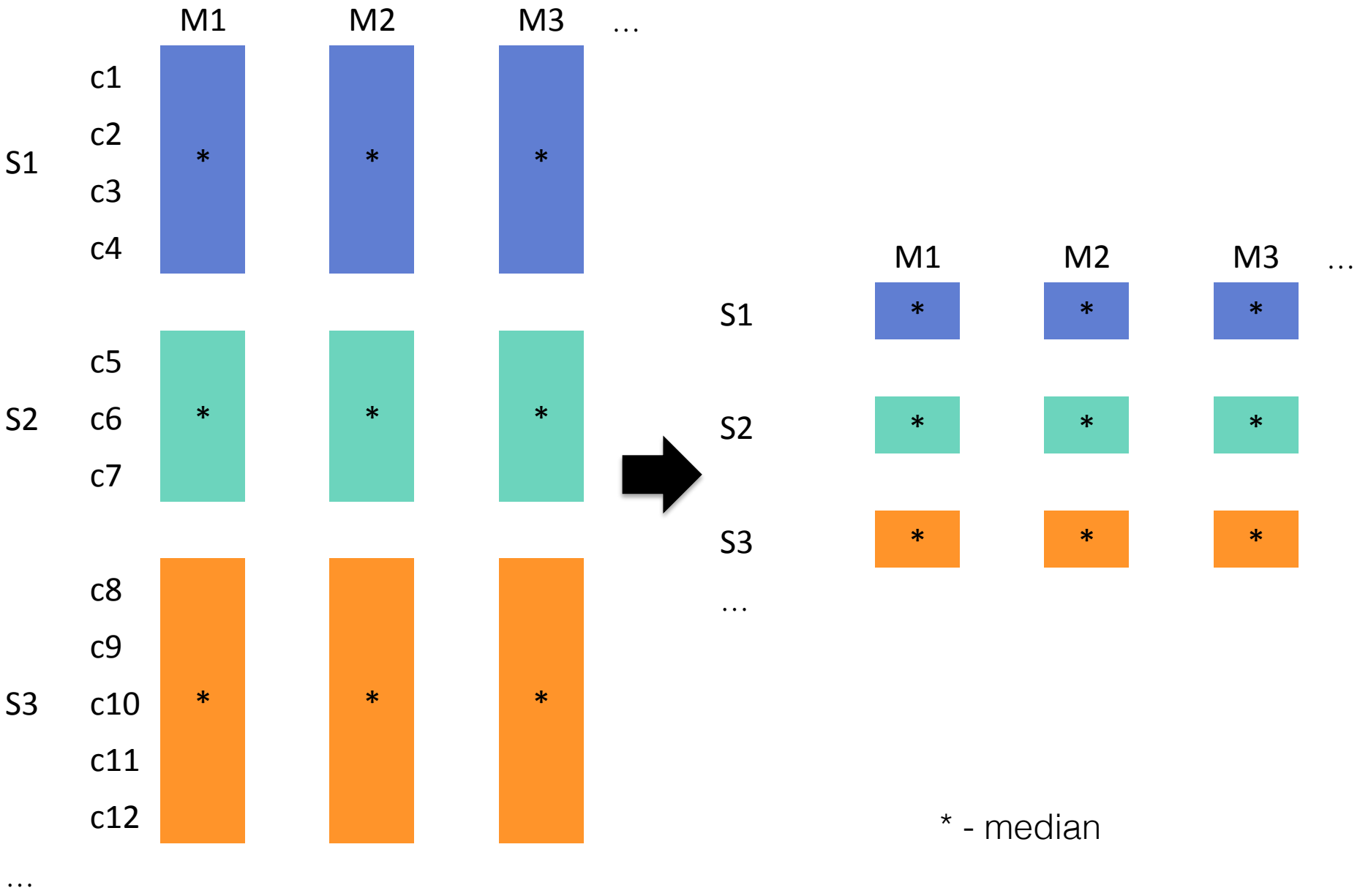
- Skewed distributions
  - Hard to distinguish between positive and negative populations
- Magic transformation - arcsinh (hyperbolic inverse sine) with cofactor 5
  - linear for the lower expression,
  - logarithmic for the higher expression,
  - works for the negative and zero values (they do not have to be excluded from the analysis)

Figure adapted from Bendall et al. Science 2011

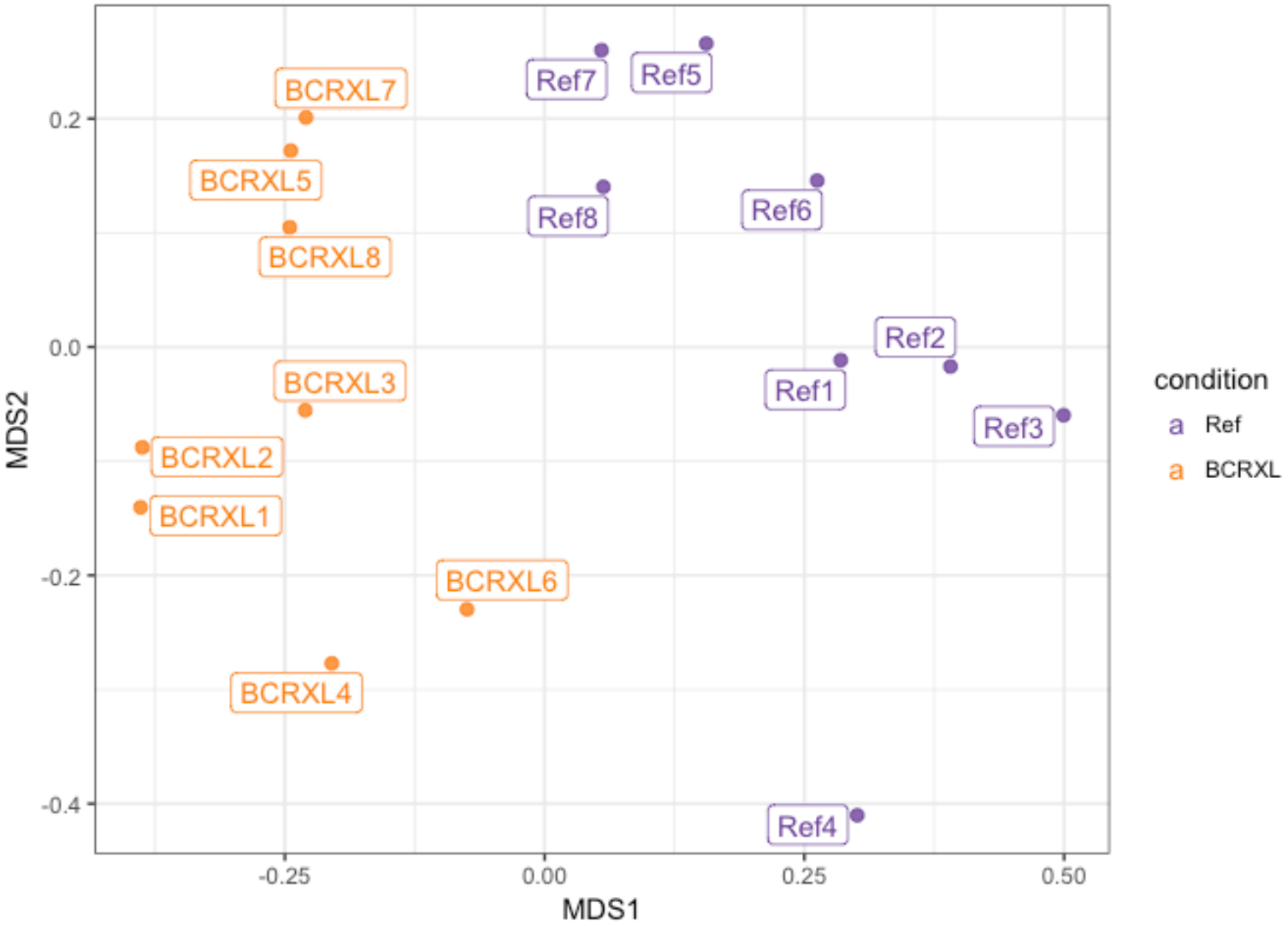
Diagnostic plots – quick look on

- Plot with per-sample marker expression distributions, colored by condition
  - Identify problematic samples or markers
- Cell counts in samples
- MDS plot (or PCA plot)
  - Standard usage in the RNA-seq analysis
  - Generated with the plotMDS() function from the limma pckg
  - As we want to plot samples we need to summarize the information about cells to the sample level

MDS plot: expr_median_sample_tbl



MDS plot


BCRXL7
BCRXL5
BCRXL8
BCRXL3
BCRXL2
BCRXL1
BCRXL6
BCRXL4
Ref7
Ref5
Ref8
Ref6
Ref2
Ref1
Ref3
Ref4
0.2
0.0
-0.2
-0.4
MDS2
-0.25
0.00
0.25
0.50
MDS1
condition
a Ref
a BCRXL

Cell population identification

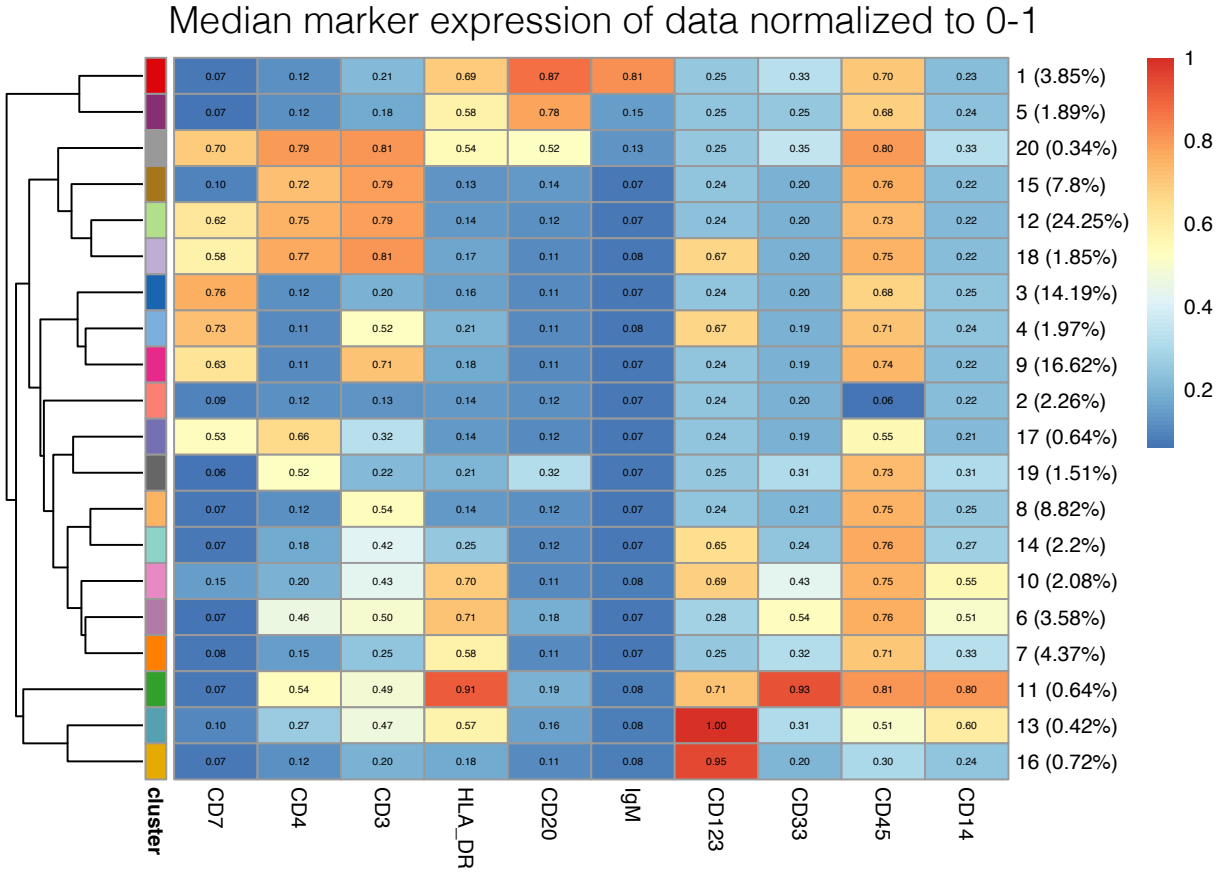
- Manual gating
- Comparison of clustering algorithms by Lukas Weber et al., Cytometry Part A, 2016
- FlowSOM (+ ClusterConsensusPlus) one of the best performing methods and super fast
- 3 steps:
  1. Building of the self-organizing map (SOM) with the BuildSOM function - cells are assigned according to their similarities to 100 grid points (or, so-called codes) of the SOM
  2. Building of a minimal spanning tree, which is mainly used for graphical representation of the clusters
  3. Metaclustering of the SOM codes performed directly with the ConsensusClusterPlus function

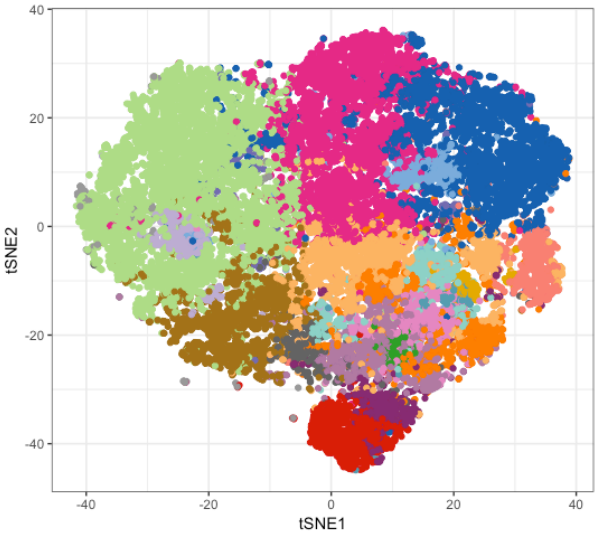
Over-clustering

- Some level of over-clustering is necessary, in order to detect somewhat rare populations
- In addition, merging can always follow an over-clustering step, but splitting of existing clusters is generally not feasible

Over-clustering

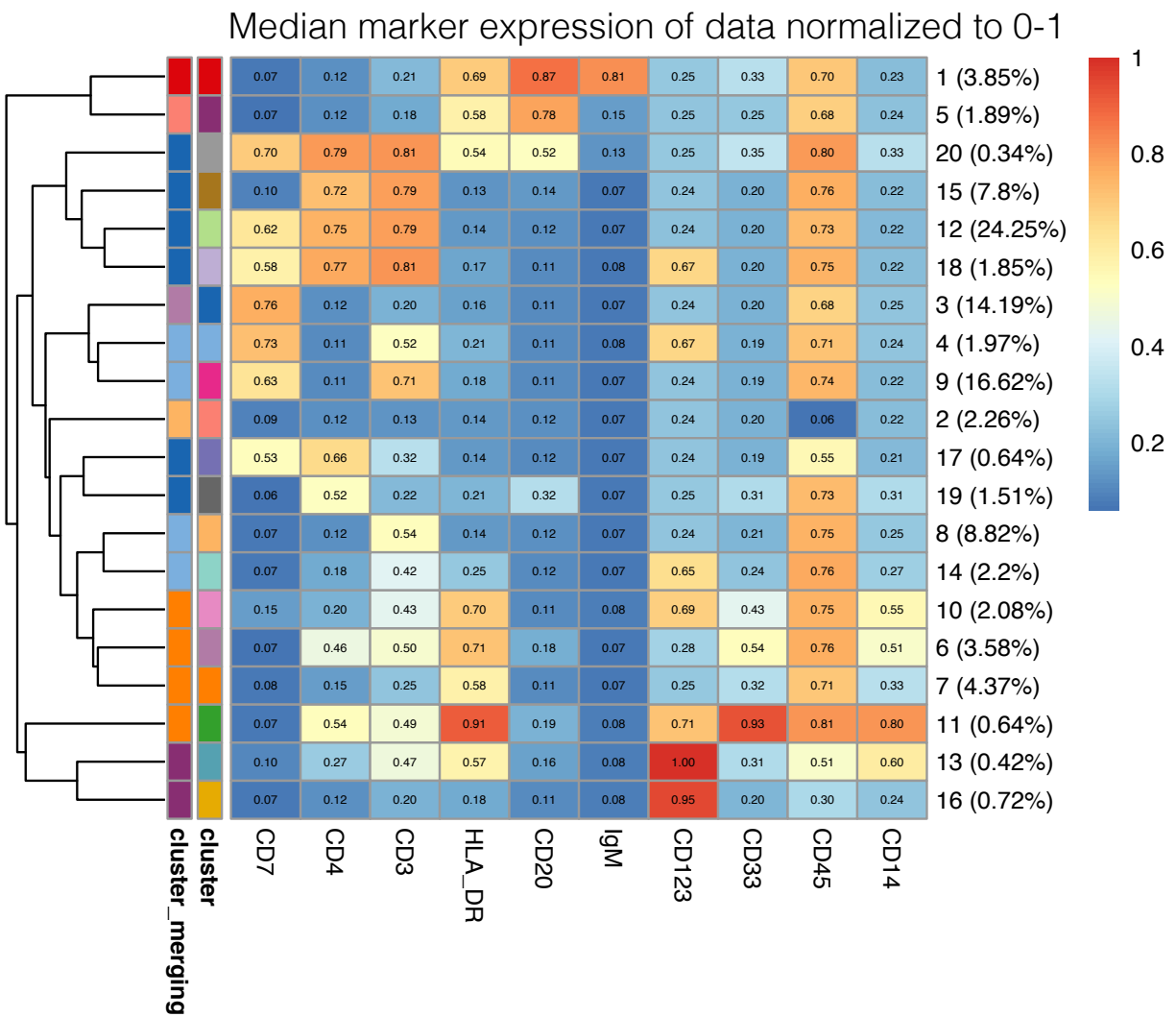
Over-clustering into 20 groups

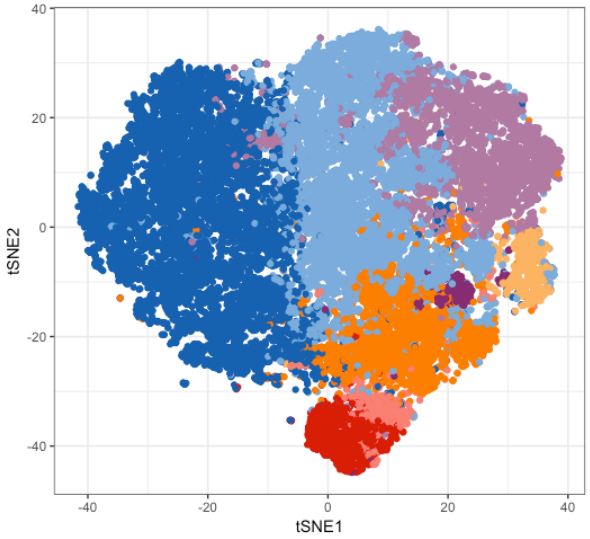




Figures adapted from Nowicka et al. F1000Research 2017

Merging by an expert



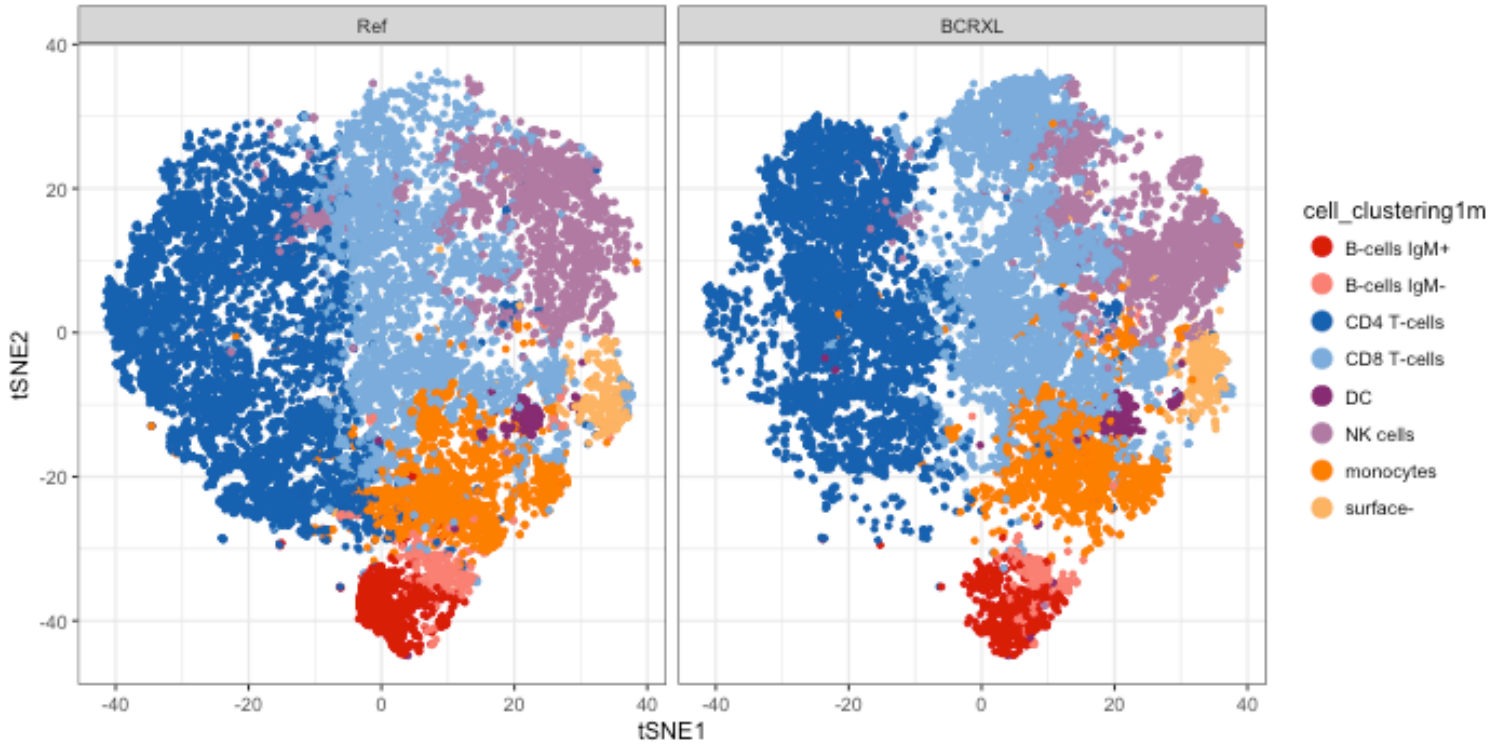


cluster_merging

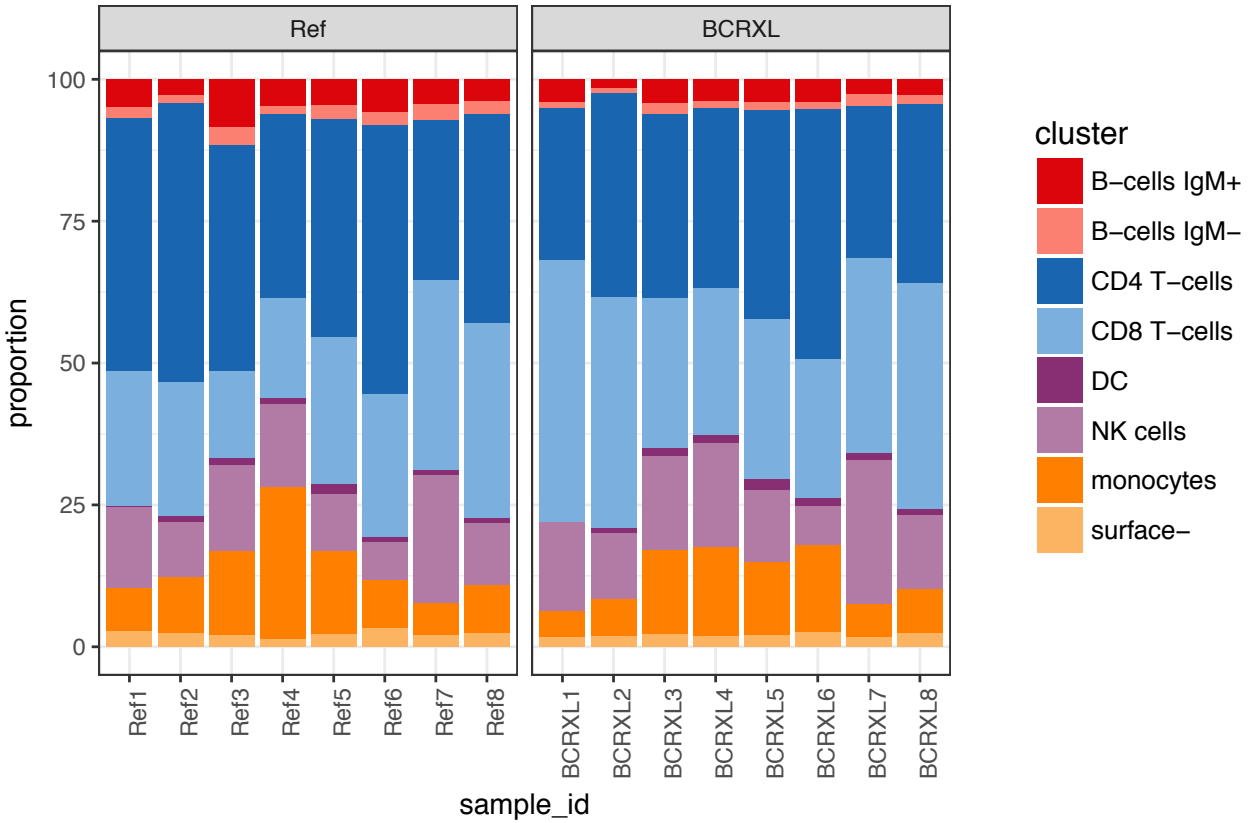
- B–cells IgM+
- B–cells IgM–
- CD4 T–cells
- CD8 T–cells
- DC
- NK cells
- monocytes
- surface–

Figures adapted from Nowicka et al. F1000Research 2017

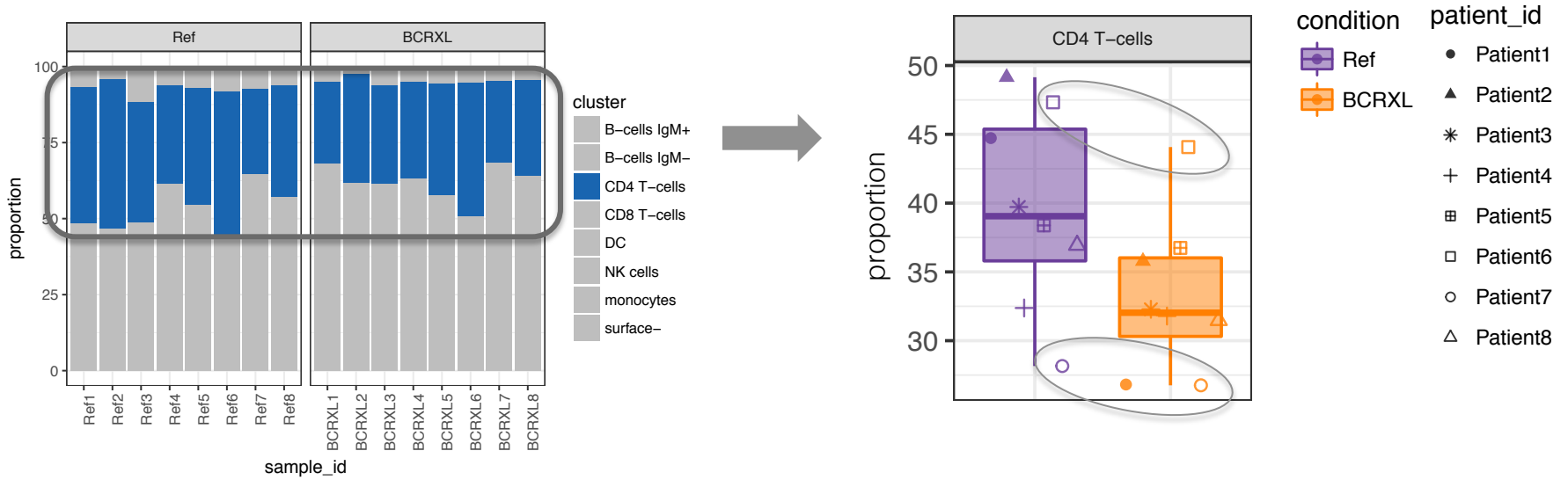
Statistical testing – differential abundance



Statistical testing – differential abundance



Statistical testing – differential abundance



<u>The generalized linear mixed model (GLMM) for each cell population</u>

Y – number of CD4 cells
π – proportion of CD4 cells
m – total number of cells

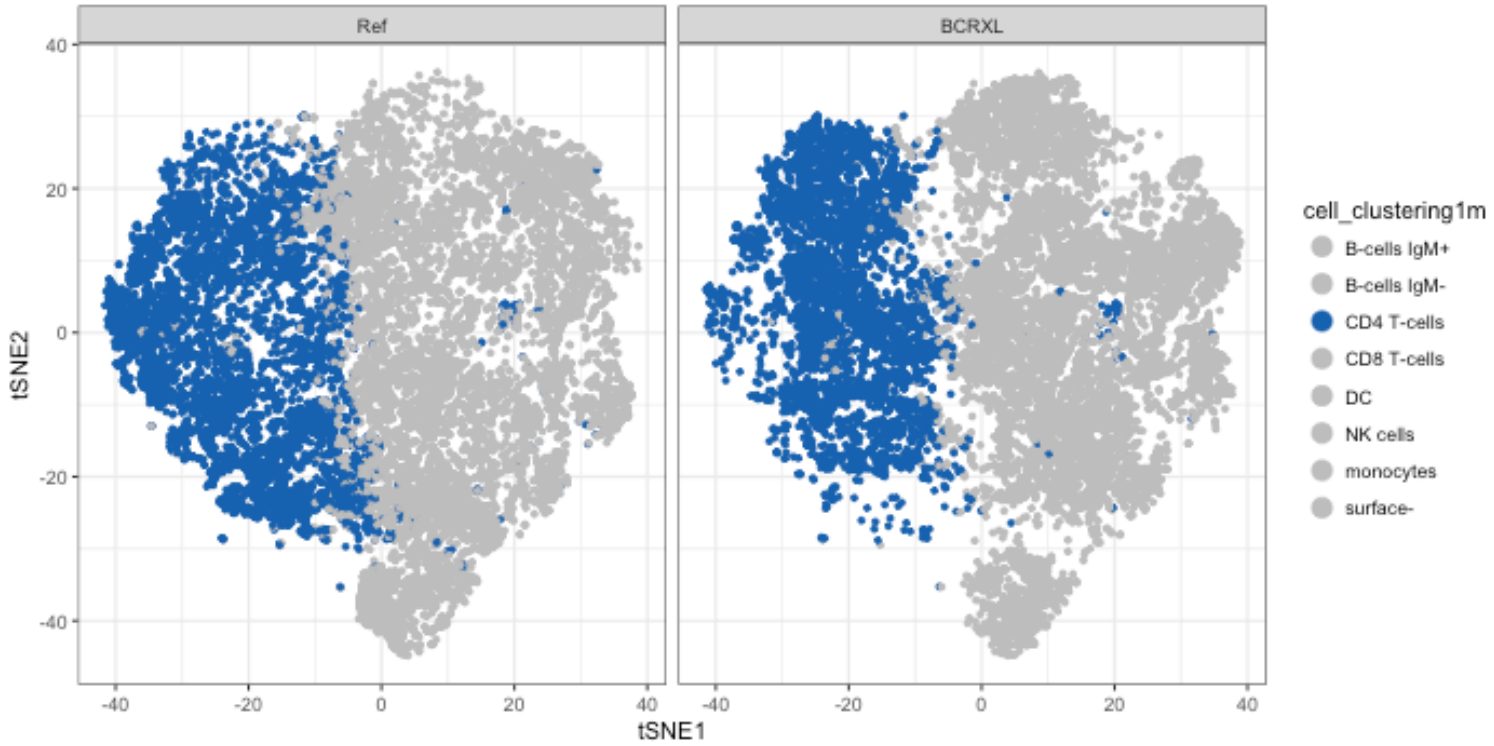
$$Y \sim Bin(m, \pi)$$

i = 1,…,8 – patient ID
j = 1, 2 – condition

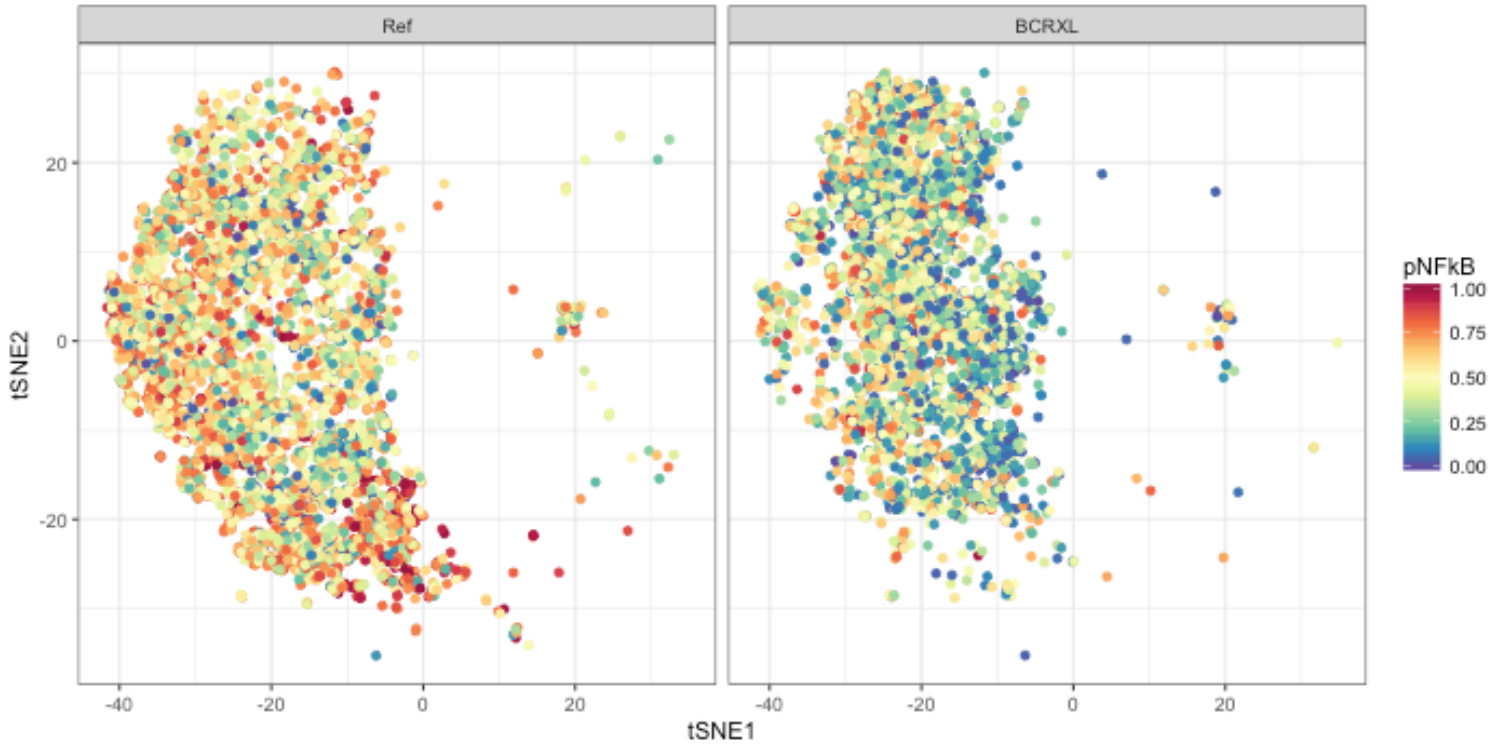
$$logit(\pi_{ij}) = \beta_0 + \beta_1 x_{ij} + \xi_{ij} + \gamma_i$$

- observation-level random effects $\xi_{ij} \sim N(0, \sigma_\xi^2)$
- random intercept for each patient $\gamma_i \sim N(0, \sigma_\gamma^2)$

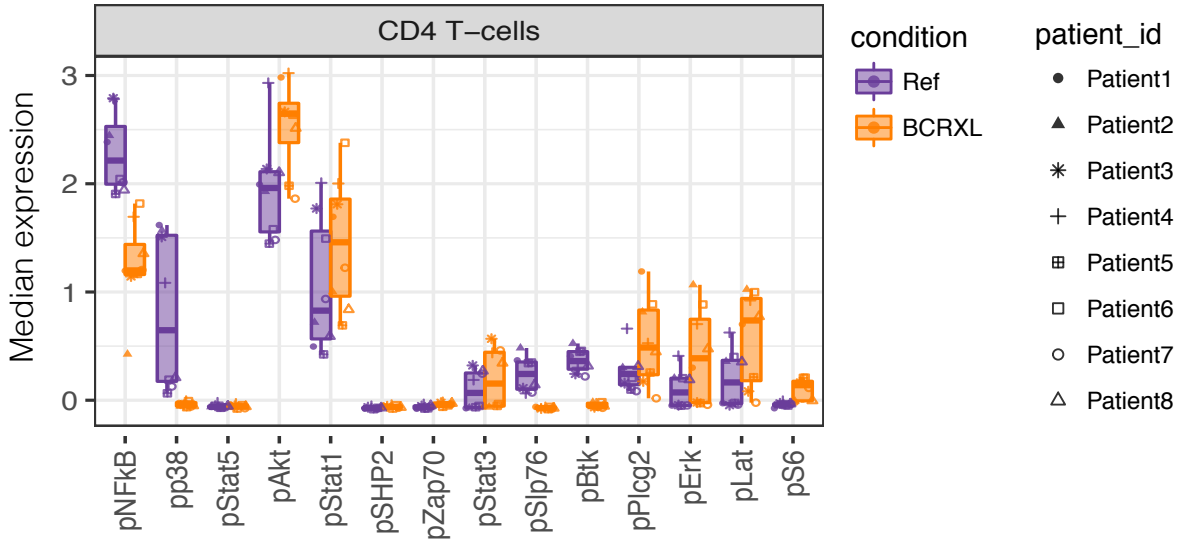
Statistical testing – differential expression of signaling markers



Statistical testing – differential expression of signaling markers



Statistical testing – differential expression of signaling markers



<u>The linear mixed model (LMM) for each cell population</u>

Y – median expression of marker in CD4 cells

$$Y \sim N(\mu, \sigma^2)$$

i = 1,…,8 – patient ID
j = 1, 2 – condition

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

$$\epsilon_{ij} \sim N(0, \sigma^2)$$

- random intercept for each patient

$$\gamma_i \sim N(0, \sigma_\gamma^2)$$