University of Zurich[UZH]

RADIANT

Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data

C1omics 2015 conference

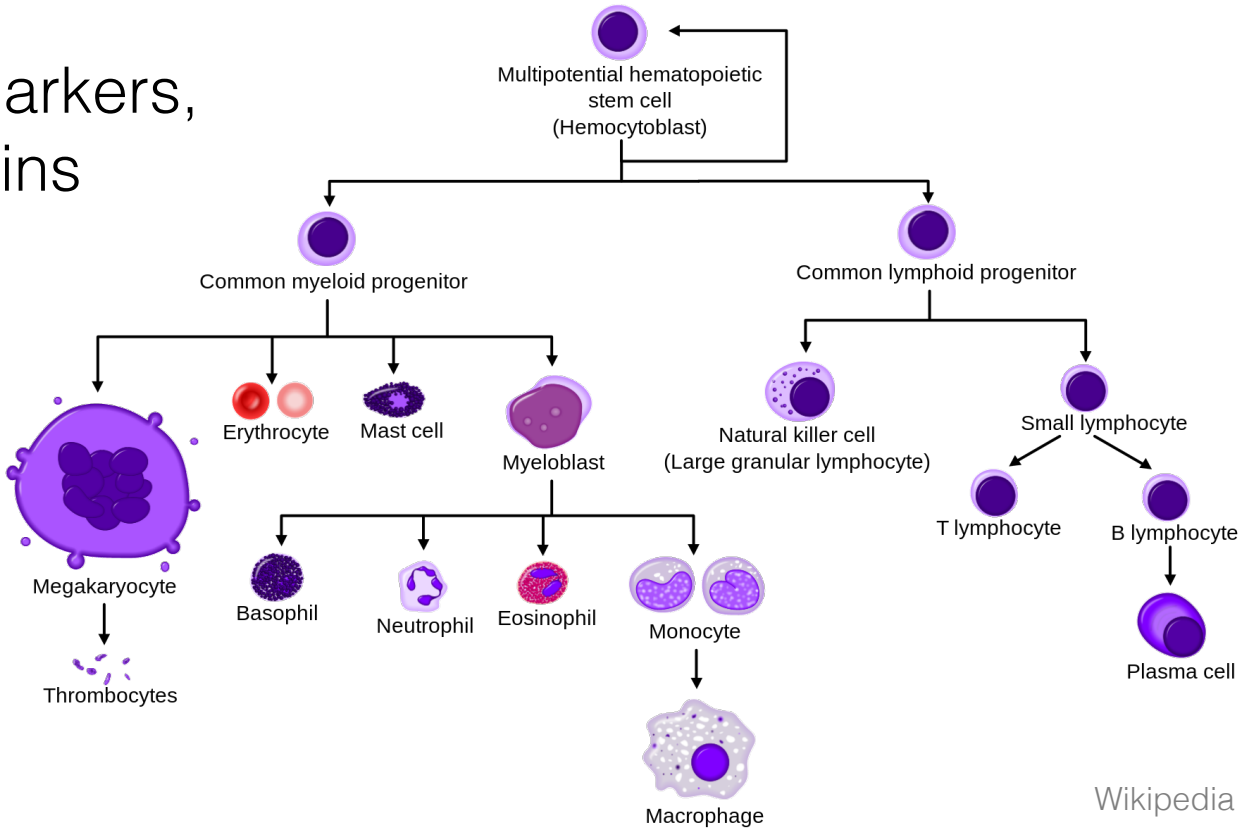
Single-cell omics methods and applications

Manchester, UK

Lukas Weber, 25 November 2015

Background

- single cell analysis: cytometry
- protein expression levels to characterize cell populations
- surface protein markers, intracellular proteins
- flow cytometry, mass cytometry



Wikipedia

Applications

- Analysis of immune cell populations in biological samples
  - disease diagnostics, e.g. leukemia
  - vaccine research

- Differential analysis: comparisons between groups of samples

- Focus may be on major populations or rare cell types
  - e.g. analysis of proportions of major populations in a sample
  - or detection of a single rare cell type

Technologies

- Flow cytometry: 1000s cells/sec, 5-15 proteins/cell
- Mass cytometry: 100s cells/sec, 40 proteins/cell

- FCS file format: n cells, p protein expression levels

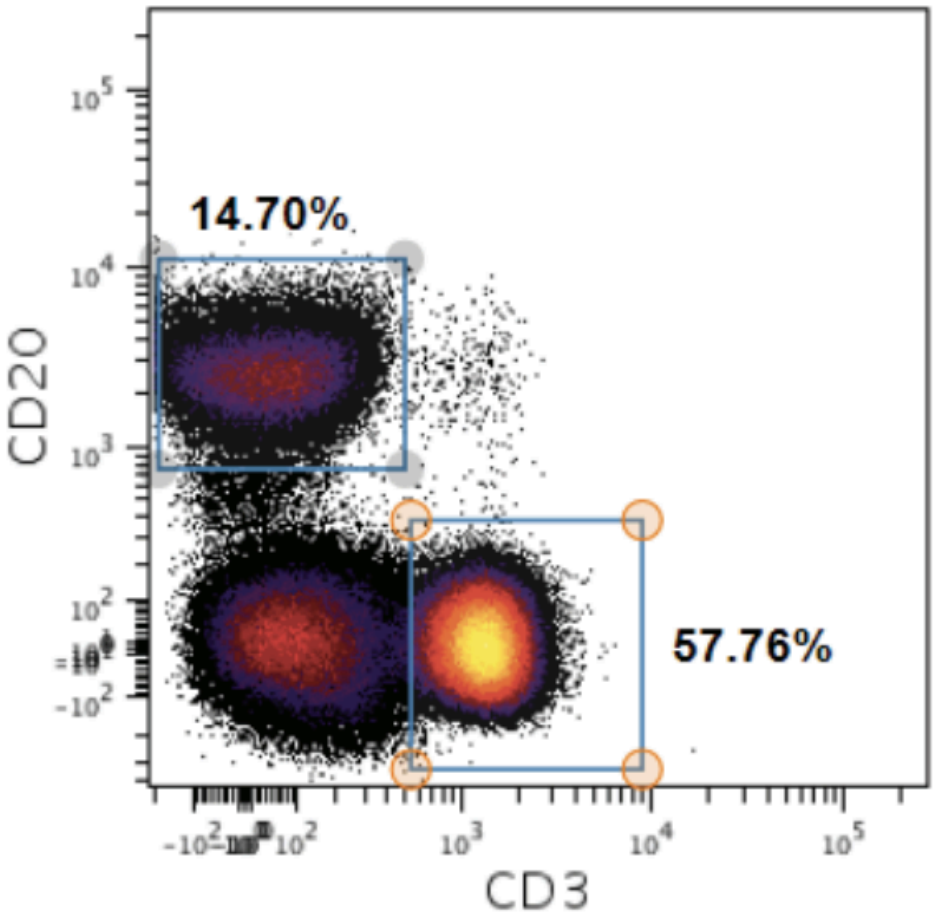
p proteins (dimensions)

n cells

```
> head(data)
     CD45RA(La139)Di CD133(Pr141)Di CD19(Nd142)Di CD22(Nd143)Di CD11b(Nd144)Di
[1,]      0.81704921     -0.1479468  -0.033481941     0.3321835   -0.045922440
[2,]      3.80138493     -0.1914464  -0.083273850     0.3723878    4.494378567
[3,]      3.20443869     -0.1611056   0.369612783    -0.2149521   -0.009404267
[4,]      2.23738217     -0.1380714  -0.088311136    -0.2204303    4.006597996
[5,]     -0.04404699     -0.1515095   0.402548134     2.5817690    6.742060184
[6,]      1.15033615     -0.1475202  -0.001792617    -0.1497730    1.529571056
```

Data analysis

- "gating"
  - visual inspection of series of 2D scatter plots
- problems
  - too many dimensions
  - multidimensional structure
  - unknown populations
  - subjective
- automated methods



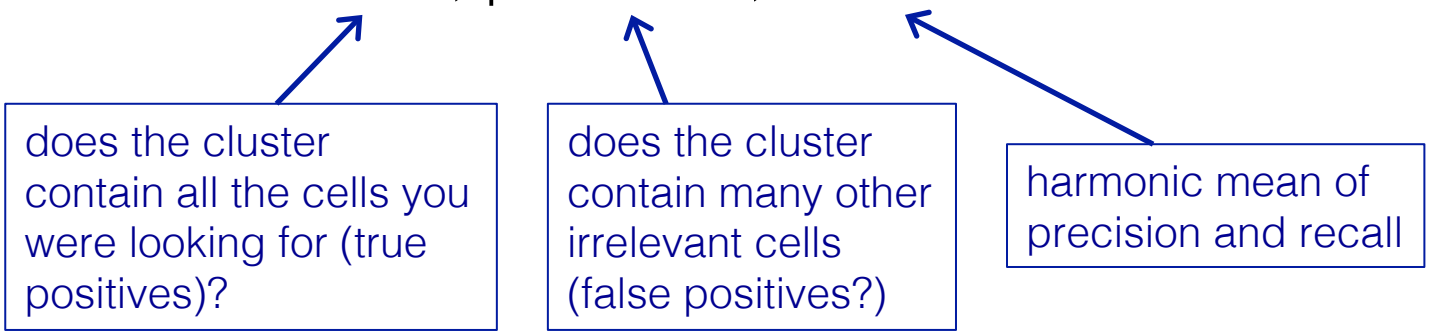
gating example:
http://www.cytobank.org

Our project

- Compare performance of clustering methods for automated detection of cell populations
  - specialized clustering methods for high-dimensional flow/mass cytometry data
  - active area of research, many new methods
  - ACCENSE, DensVM, FLOCK, flowMeans, FlowSOM, immunoClust, PhenoGraph, Rclusterpp, SamSPECTRAL, SWIFT, …

Our project

- 11 (13) methods
- 2 publicly available data sets, with manually gated cell populations as “truth”
  - Data set 1: 14 major populations
  - Data set 2: a single rare population of interest
- Metrics: recall, precision, F1 score



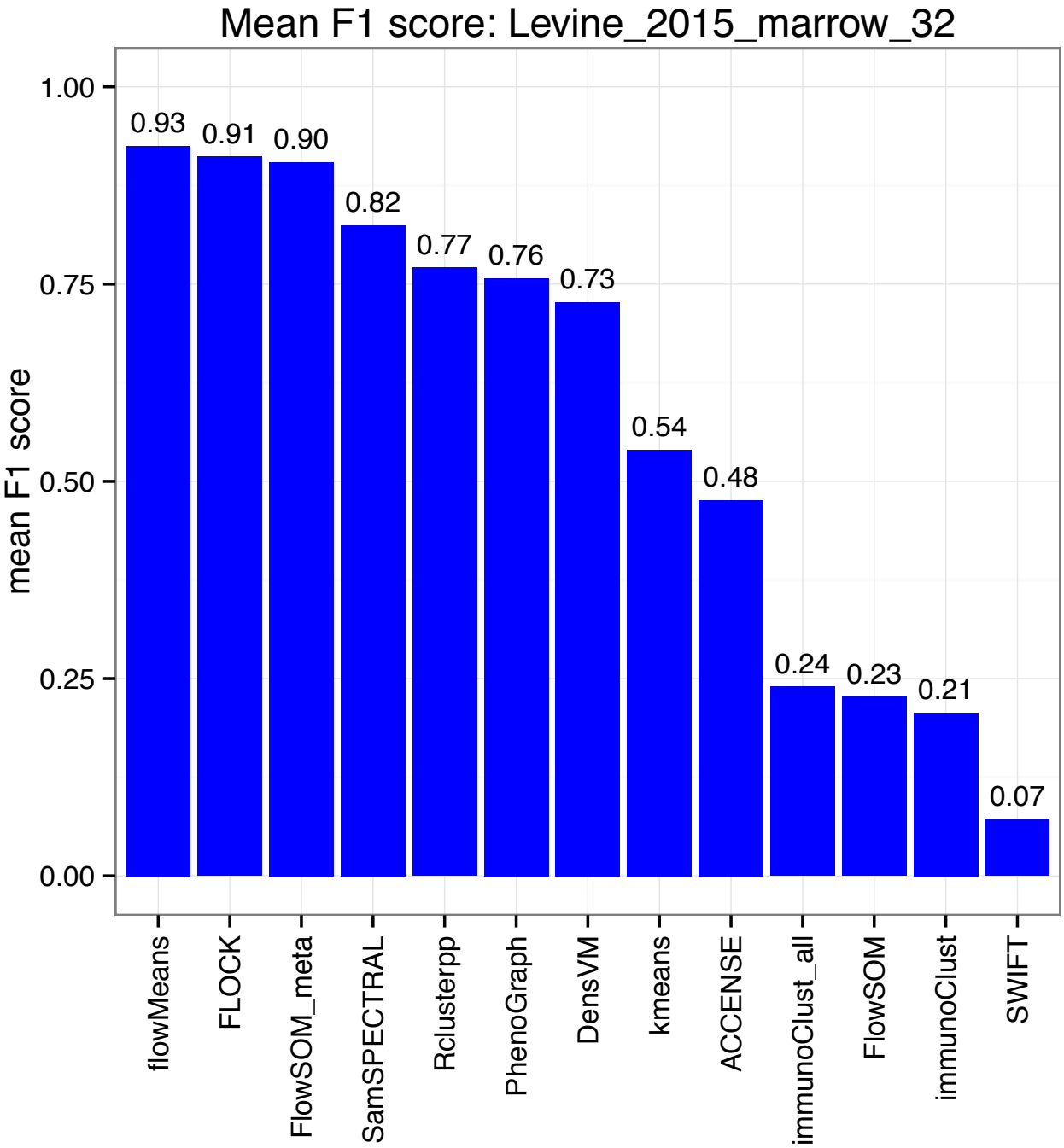
Results: Data set 1

- Task: detection of 14 major immune cell populations
- Data set source:
  - Levine et al. (2015), *Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis*, Cell.
- 32-dimensional mass cytometry data set
  - healthy human bone marrow mononuclear cells (BMMCs) from 2 individuals
  - 104,184 cells (size *n*)
  - expression levels of 32 surface marker proteins (dimensionality *p*)
  - 14 manually gated major cell populations (truth)

Results

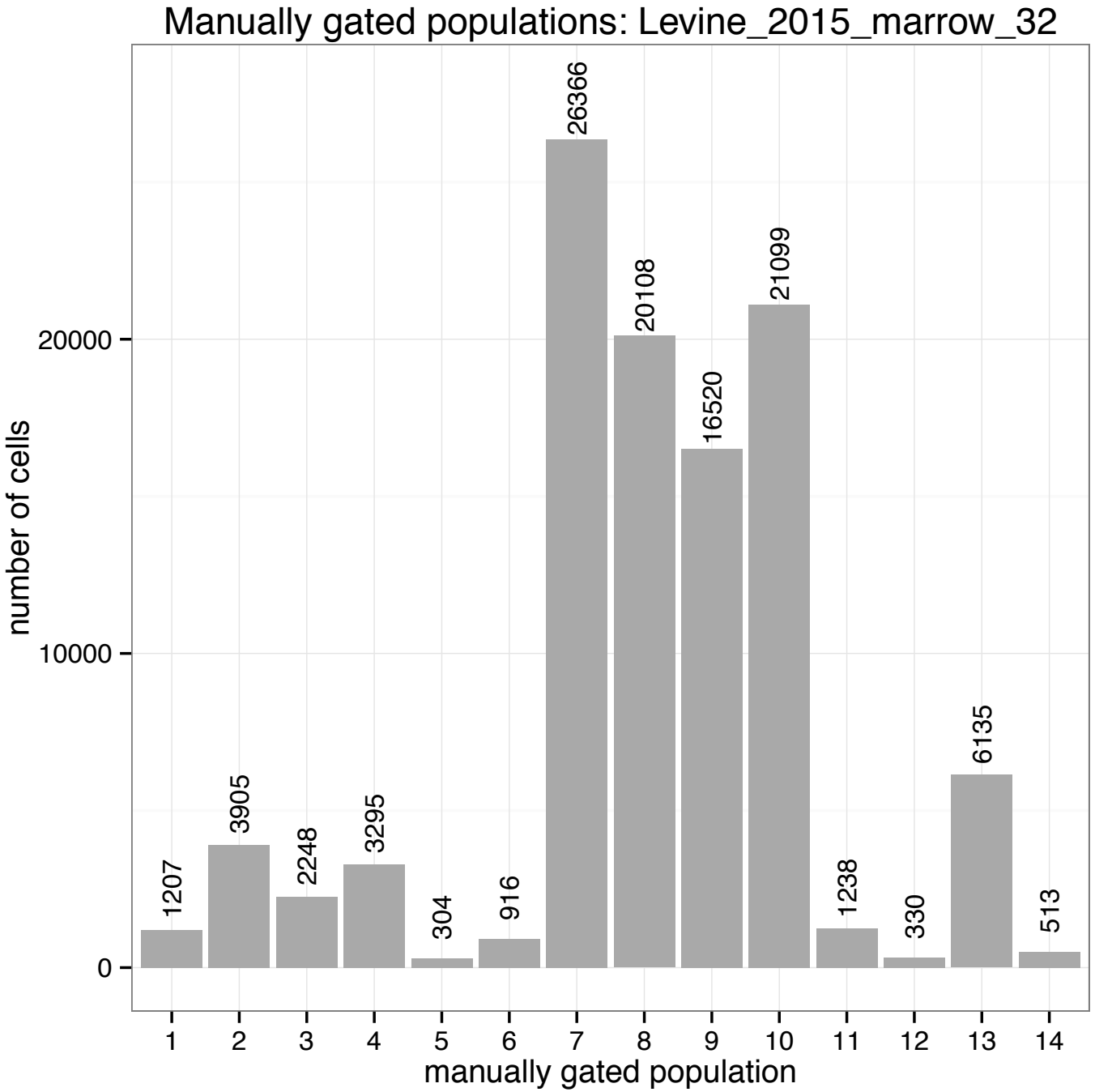
Mean F1 score

averaged over
14 true clusters,
weighted by
number of cells



Results

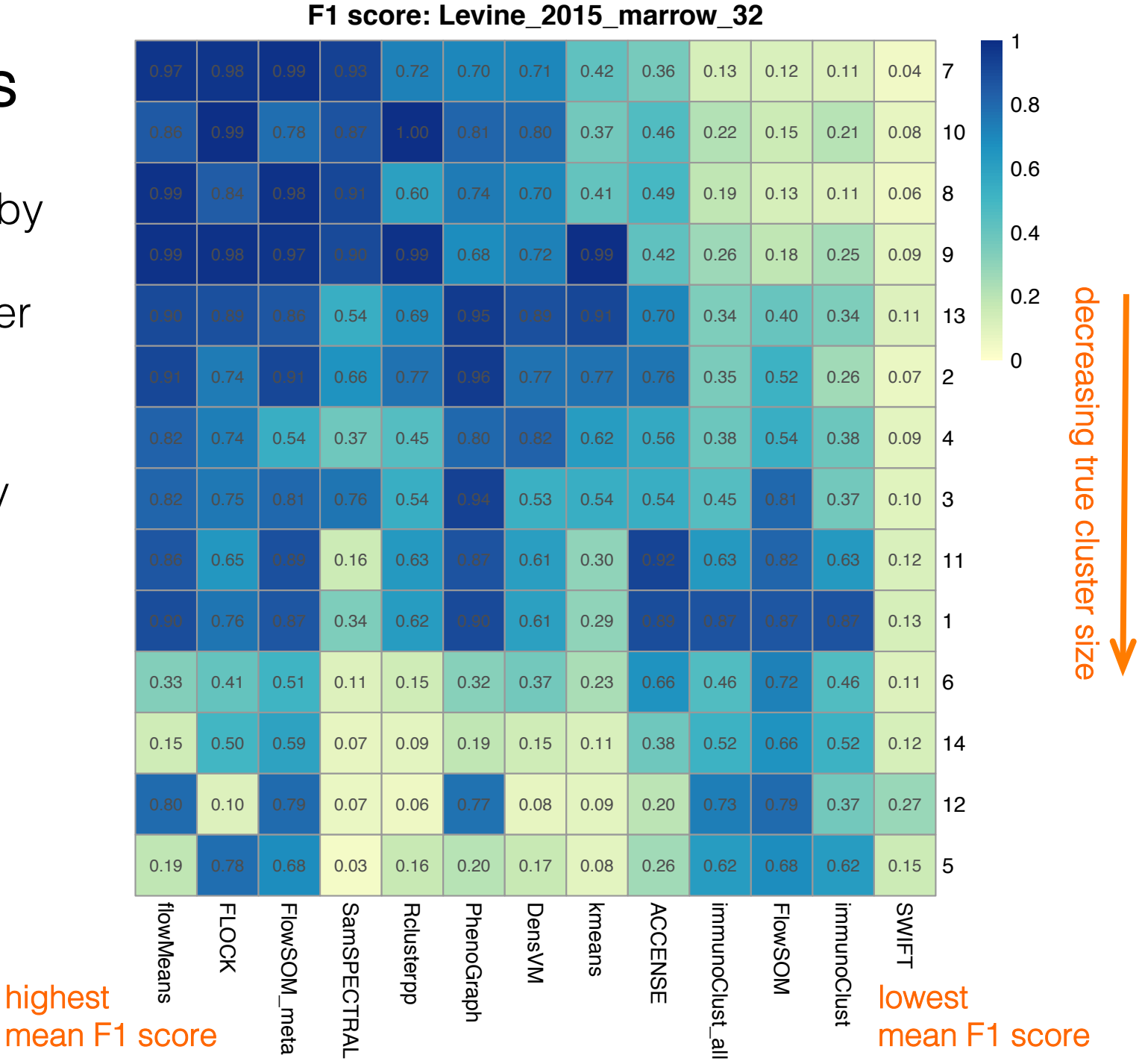
number of cells per true (manually gated) cluster



Results

F1 score by individual true cluster

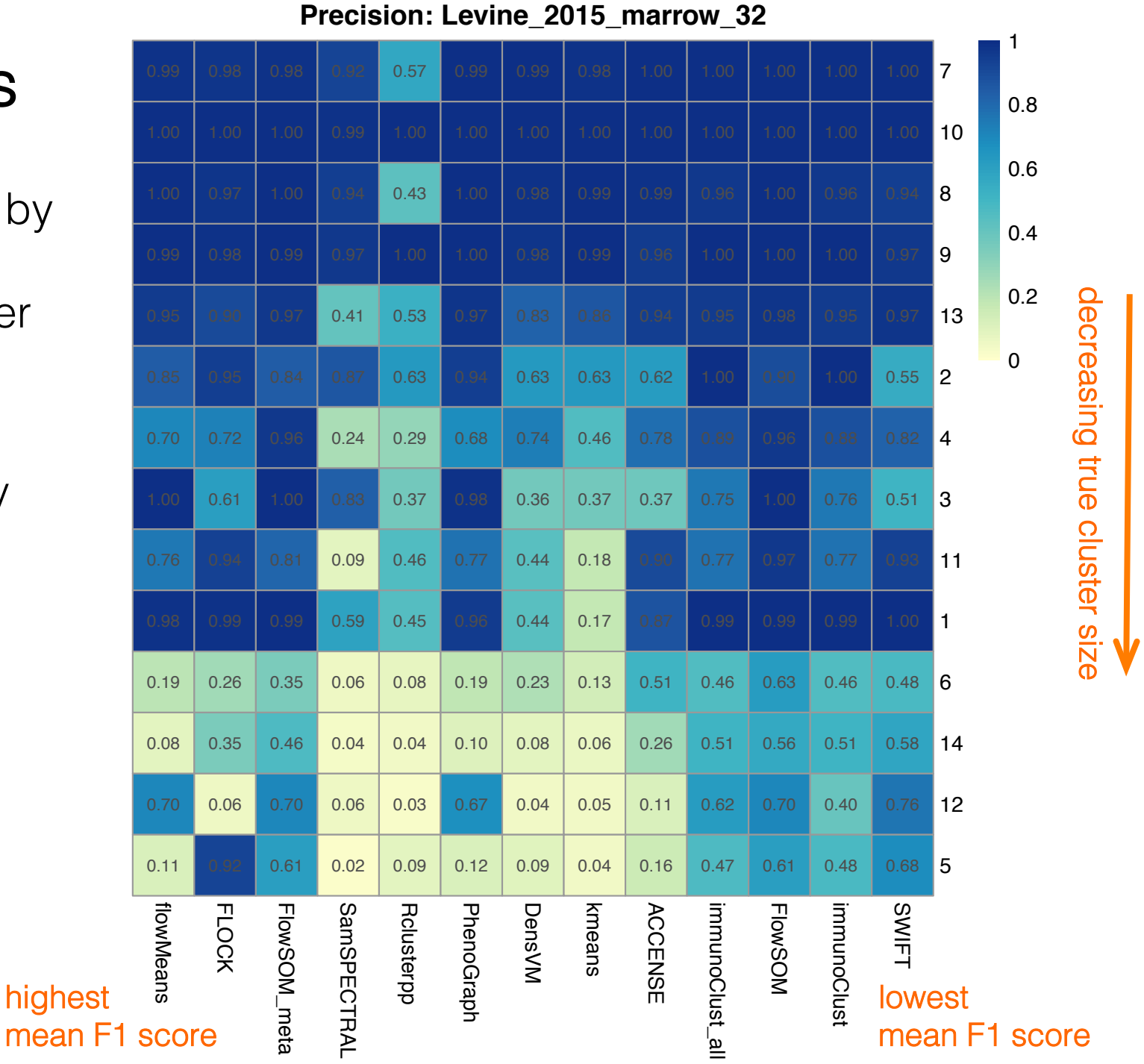
clusters ranked by size



Results

precision by individual true cluster

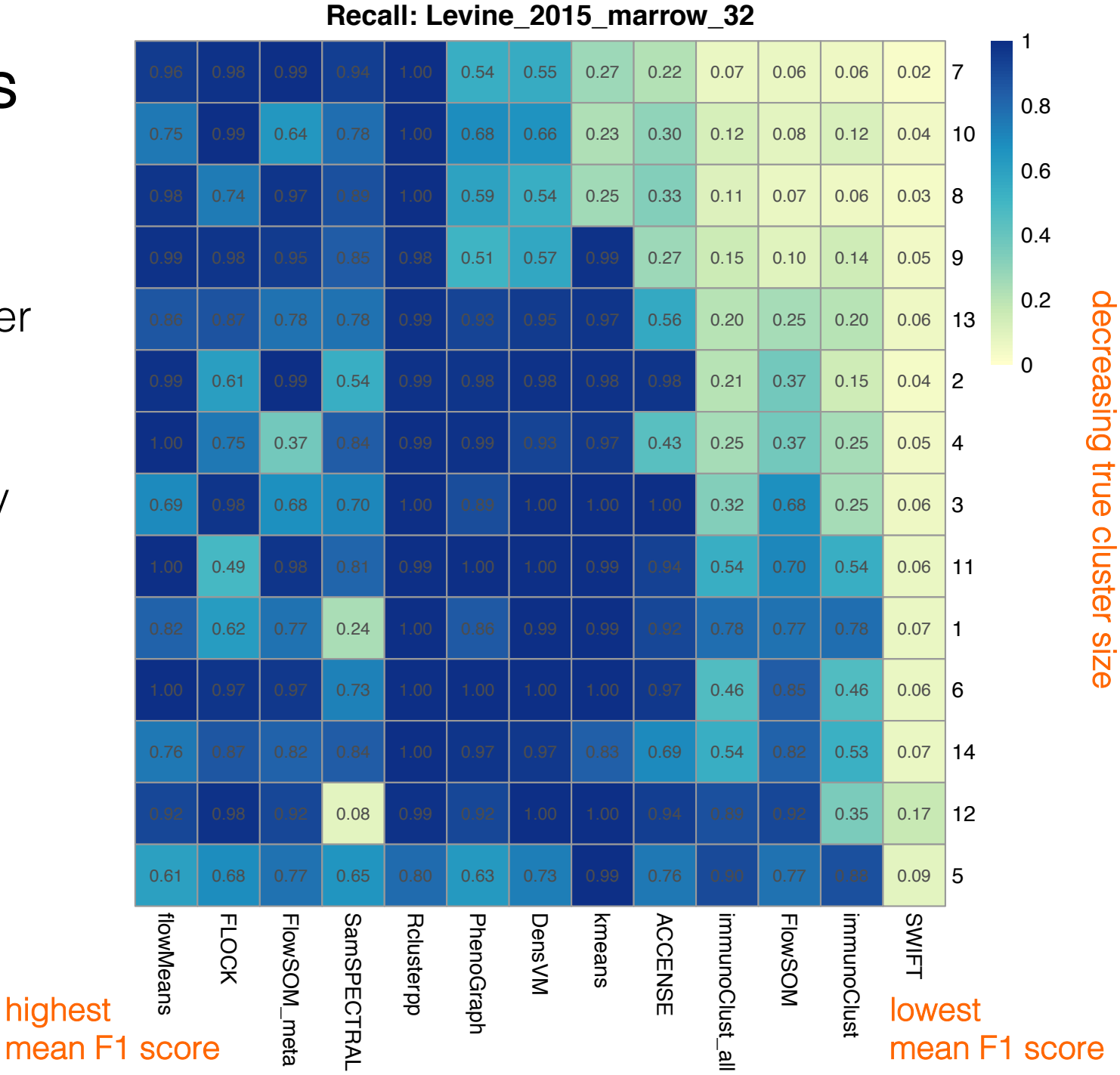
clusters ranked by size



Results

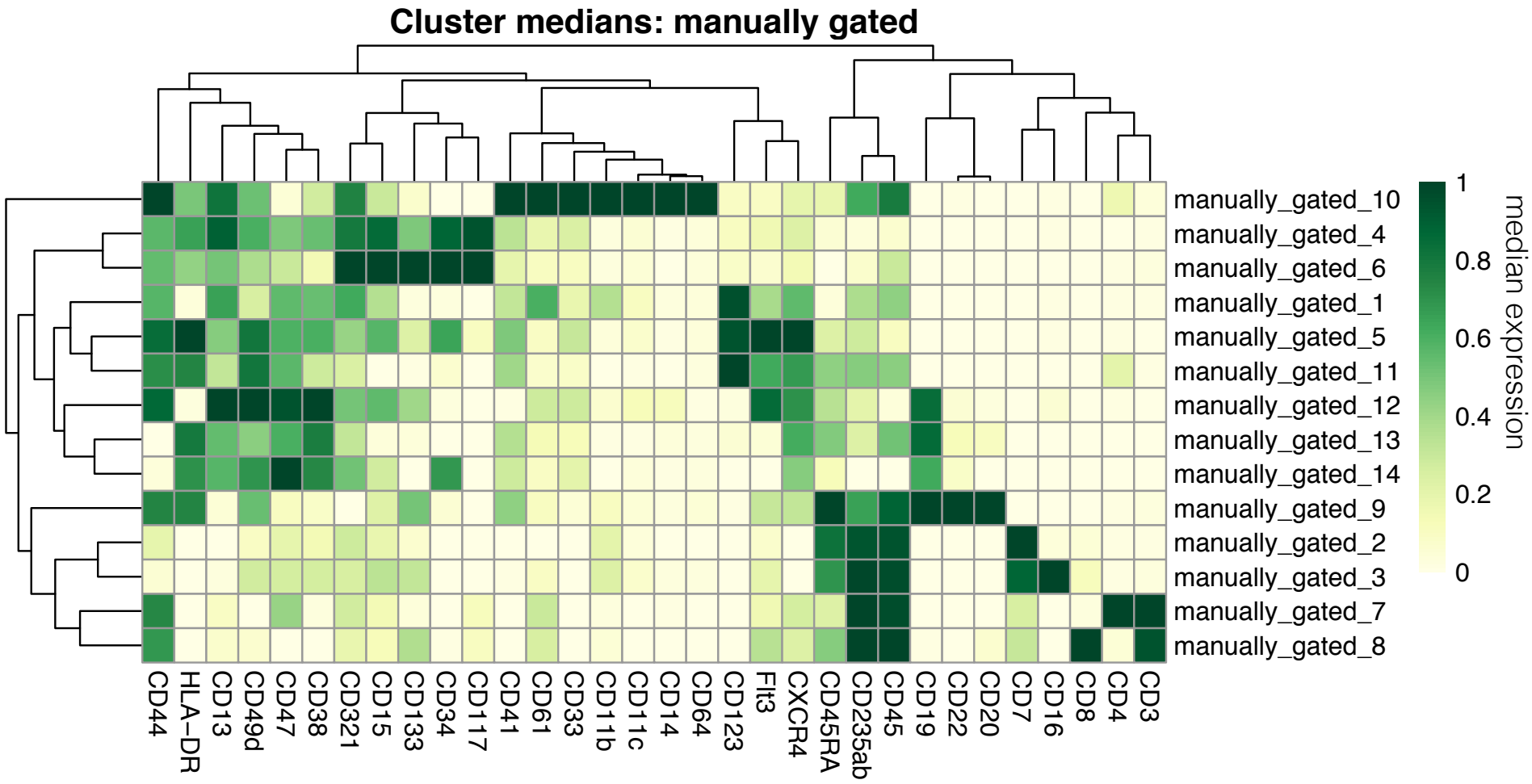
recall by individual true cluster

clusters ranked by size

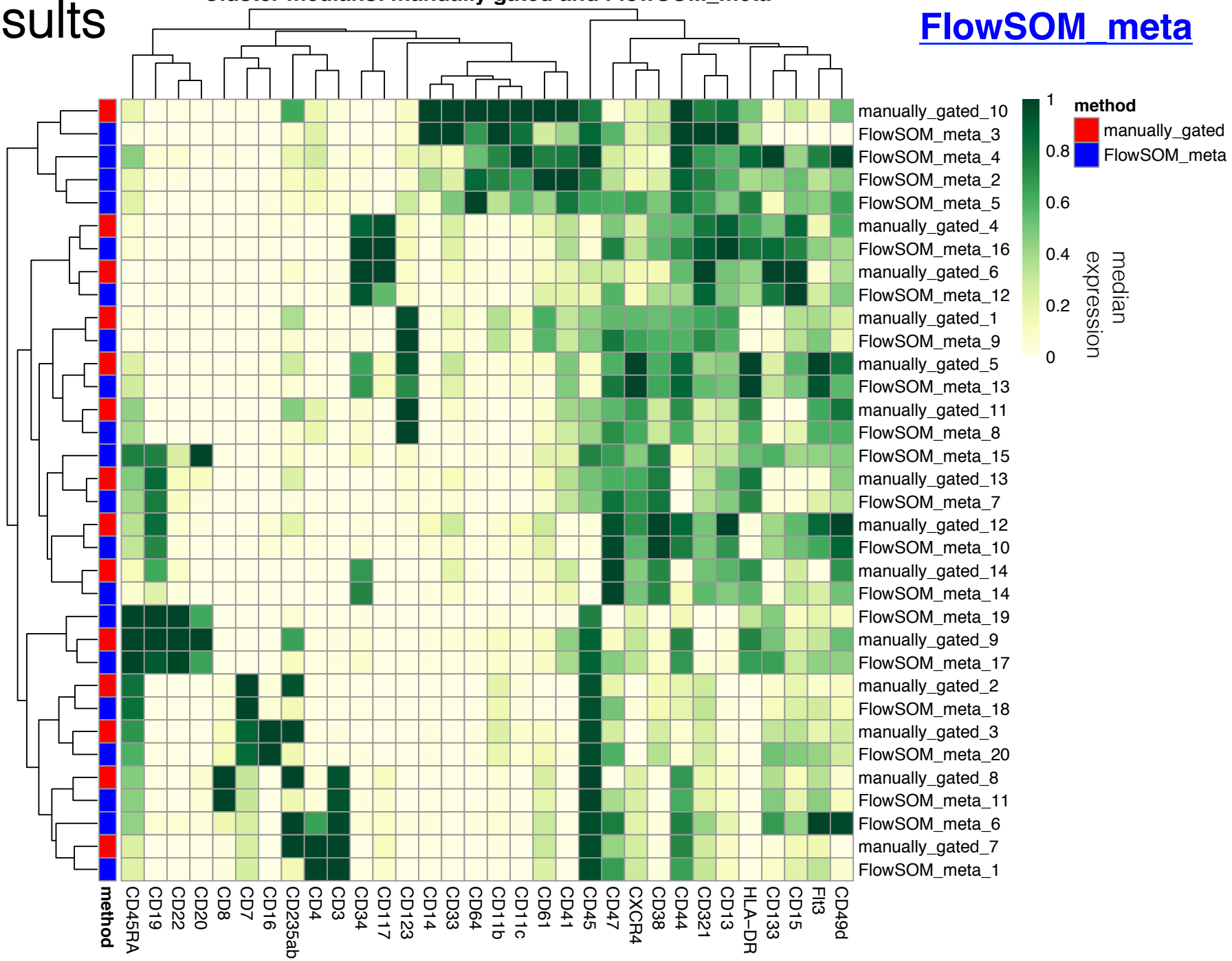


Results: cluster medians

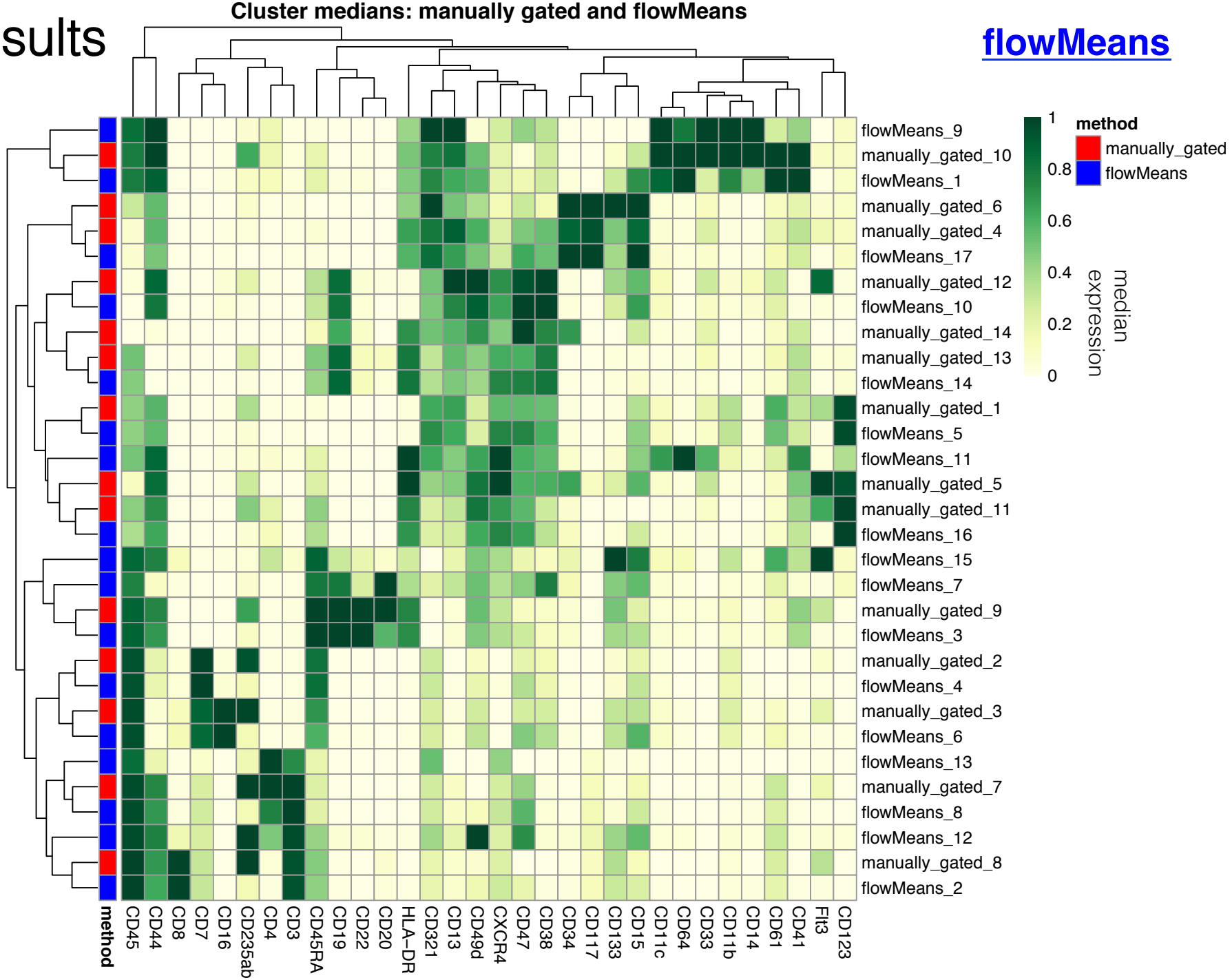
manually gated (truth)



Results

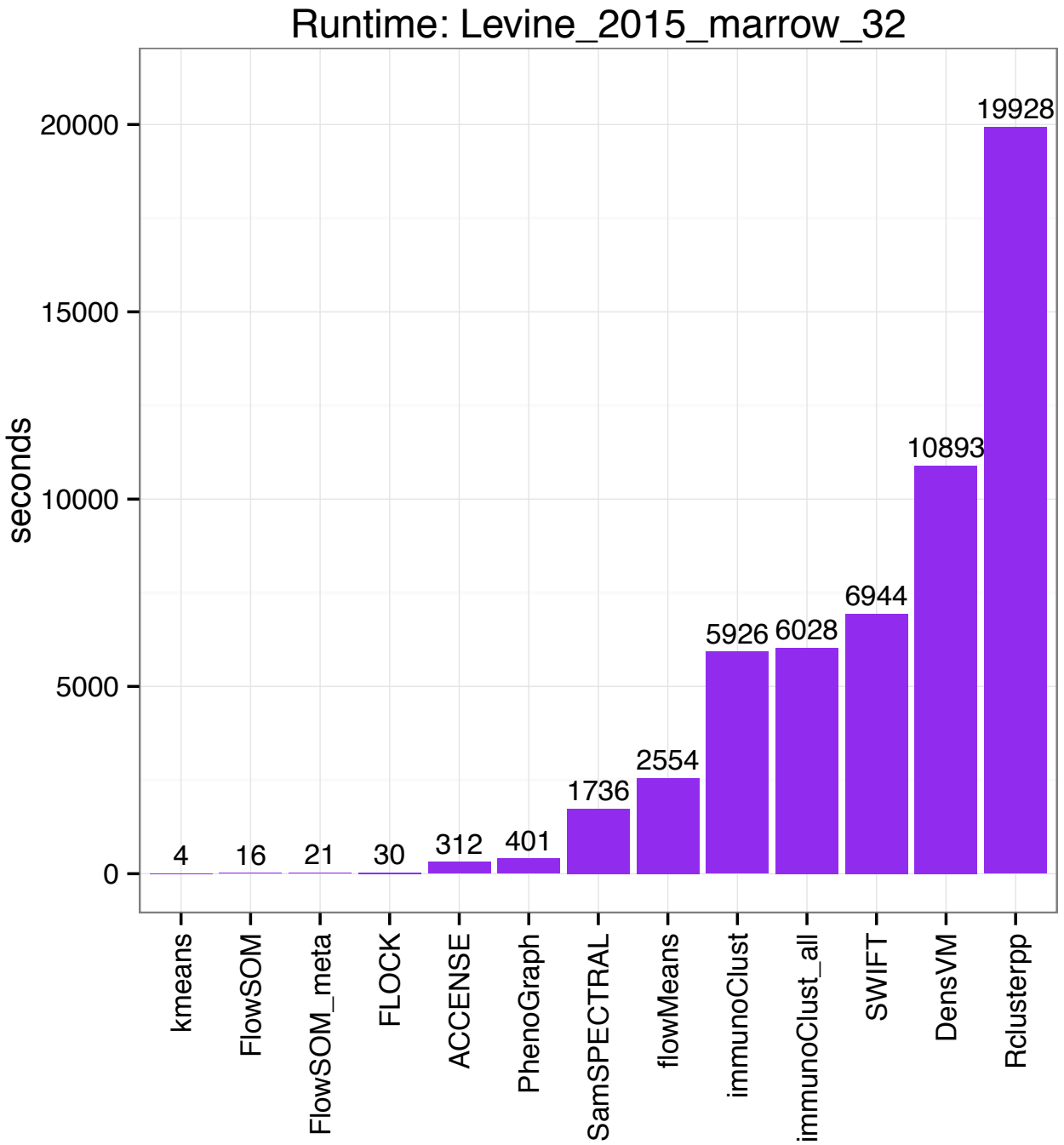

Cluster medians: manually gated and FlowSOM_meta
FlowSOM_meta
method
manually_gated
FlowSOM_meta
median expression
1
0.8
0.6
0.4
0.2
0
manually_gated_10
FlowSOM_meta_3
FlowSOM_meta_4
FlowSOM_meta_2
FlowSOM_meta_5
manually_gated_4
FlowSOM_meta_16
manually_gated_6
FlowSOM_meta_12
manually_gated_1
FlowSOM_meta_9
manually_gated_5
FlowSOM_meta_13
manually_gated_11
FlowSOM_meta_8
FlowSOM_meta_15
manually_gated_13
FlowSOM_meta_7
manually_gated_12
FlowSOM_meta_10
manually_gated_14
FlowSOM_meta_14
FlowSOM_meta_19
manually_gated_9
FlowSOM_meta_17
manually_gated_2
FlowSOM_meta_18
manually_gated_3
FlowSOM_meta_20
manually_gated_8
FlowSOM_meta_11
FlowSOM_meta_6
manually_gated_7
FlowSOM_meta_1
method
CD45RA
CD19
CD22
CD20
CD8
CD7
CD16
CD235ab
CD4
CD3
CD34
CD117
CD123
CD14
CD33
CD64
CD11b
CD11c
CD61
CD41
CD45
CD47
CXCR4
CD38
CD44
CD321
CD13
HLA-DR
CD133
CD15
Flt3
CD49d

Results


Cluster medians: manually gated and flowMeans
flowMeans
method
manually_gated
flowMeans
median expression
1
0.8
0.6
0.4
0.2
0
flowMeans_9
manually_gated_10
flowMeans_1
manually_gated_6
manually_gated_4
flowMeans_17
manually_gated_12
flowMeans_10
manually_gated_14
manually_gated_13
flowMeans_14
manually_gated_1
flowMeans_5
flowMeans_11
manually_gated_5
manually_gated_11
flowMeans_16
flowMeans_15
flowMeans_7
manually_gated_9
flowMeans_3
manually_gated_2
flowMeans_4
manually_gated_3
flowMeans_6
flowMeans_13
manually_gated_7
flowMeans_8
flowMeans_12
manually_gated_8
flowMeans_2
method
CD45
CD44
CD8
CD7
CD16
CD235ab
CD4
CD3
CD45RA
CD19
CD22
CD20
HLA-DR
CD321
CD13
CD49d
CXCR4
CD47
CD38
CD34
CD117
CD133
CD15
CD11c
CD64
CD33
CD11b
CD14
CD61
CD41
Flt3
CD123

Results

runtime in seconds



Results: Data set 2

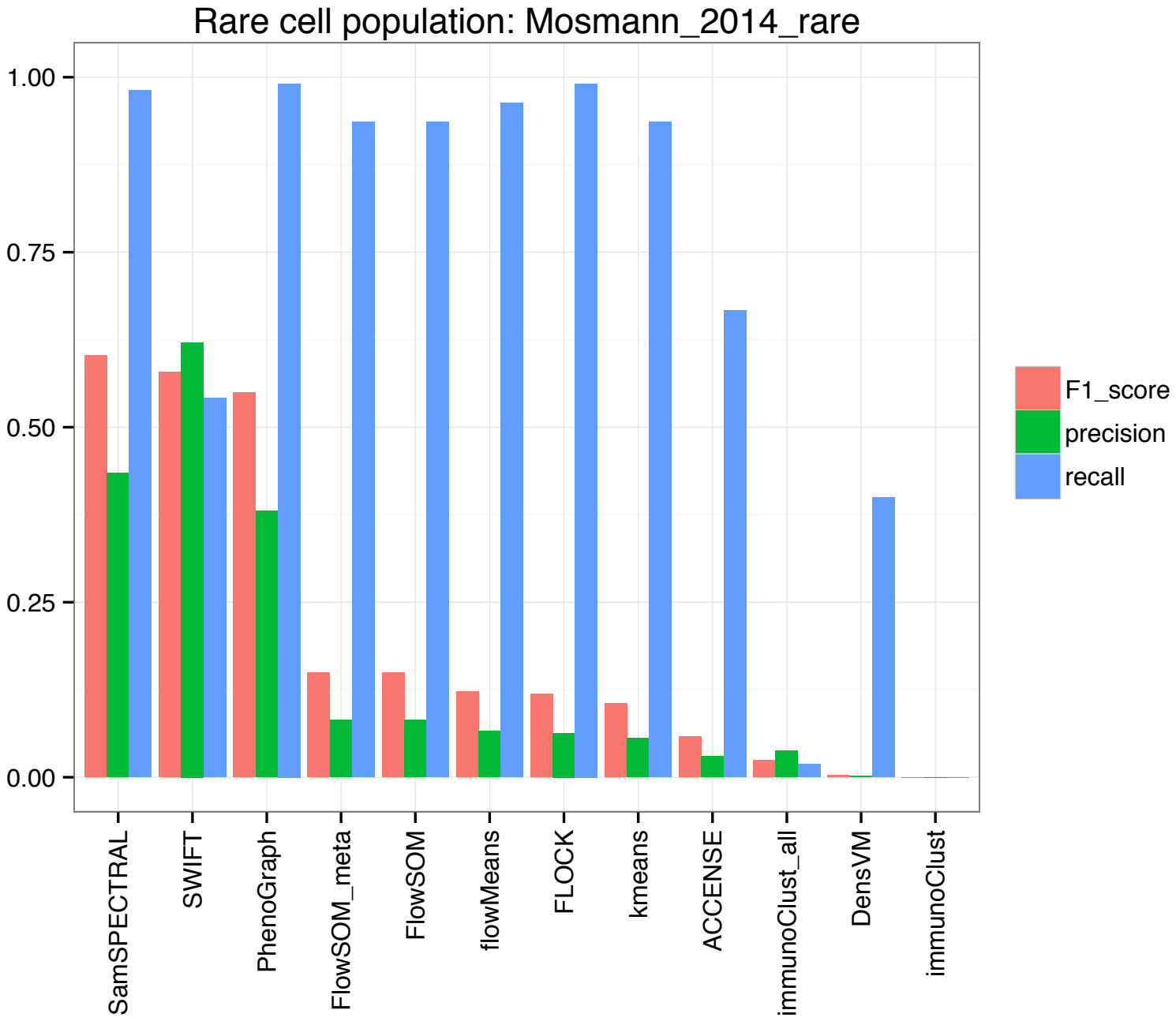
- Task: detection of a single rare cell population of interest
- Data set source:
  - Mosmann et al. (2014), *SWIFT – Scalable Clustering for Automated Identification of Rare Cell Populations in Large, High-Dimensional Flow Cytometry Datasets, Part 2: Biological Evaluation*, Cytometry Part A.
- 15-dimensional flow cytometry data set
  - healthy human peripheral blood mononuclear cells (PBMCs) exposed to influenza antigens
  - rare population of live, activated (cytokine-producing) memory CD4 T cells
  - 396,460 cells (size *n*); 109 cells from rare population
  - expression levels of 15 proteins (dimensionality *p*)

Results

detection of rare cell population

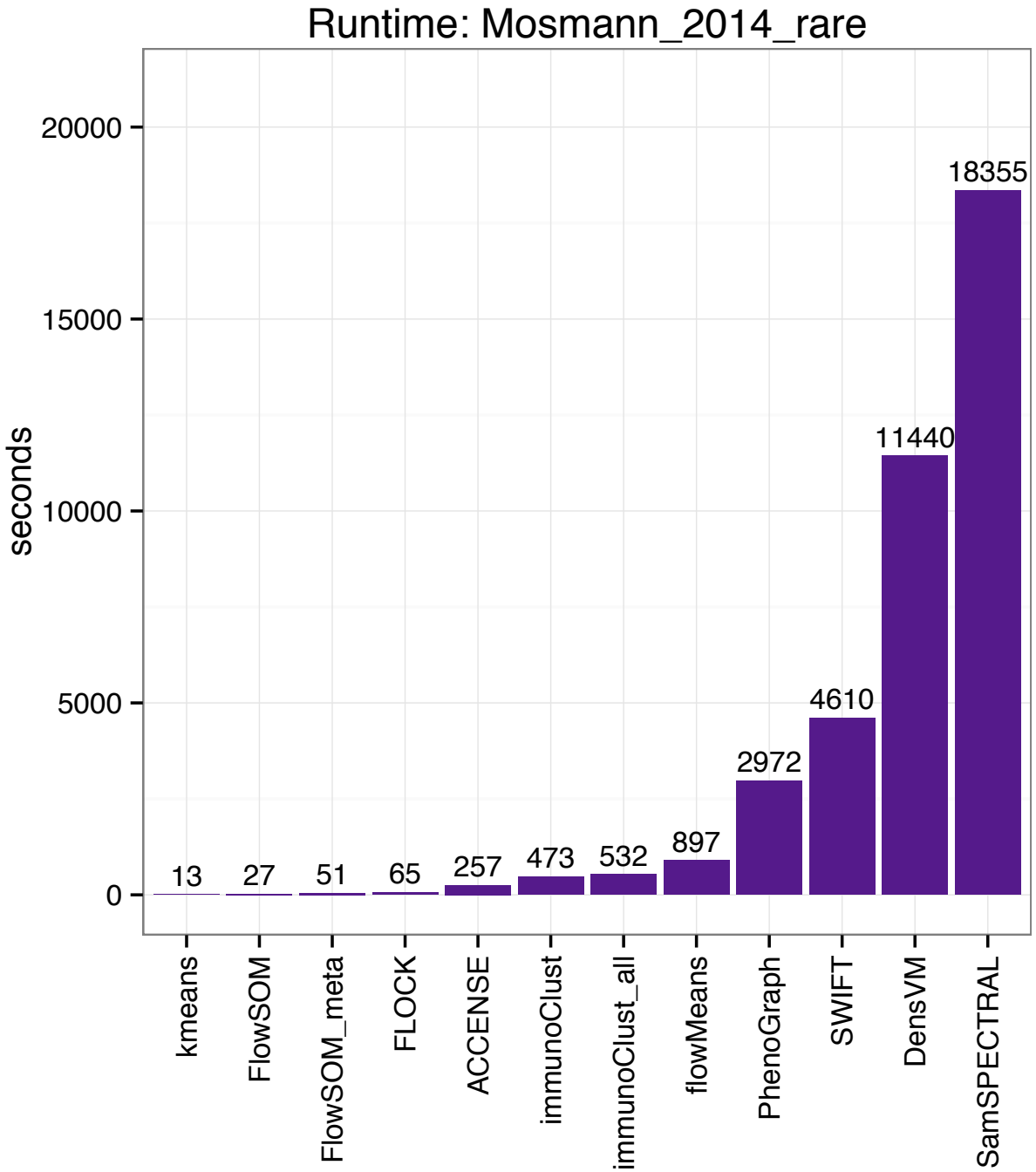
F1 score, precision, recall

109 cells (out of 396,460)



Results

runtime in seconds



Conclusions

- Data set 1: 14 major cell populations
  - FlowSOM_meta performs very well; fast runtime
- Data set 2: single rare population
  - PhenoGraph, SamSPECTRAL, SWIFT
- Runtime
- Next steps: more data sets; robustness

Acknowledgments

Mark Robinson

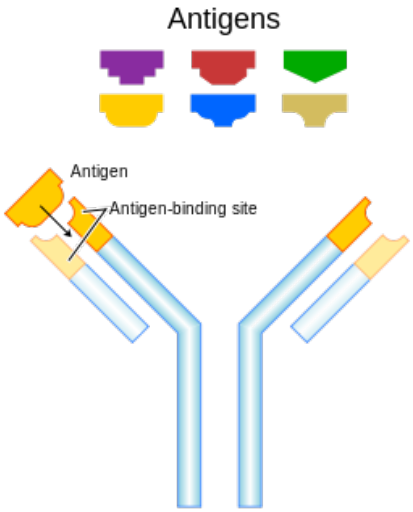
Robinson lab (UZH)

Alexander Roth (UZH)
Michael Baudis (UZH)

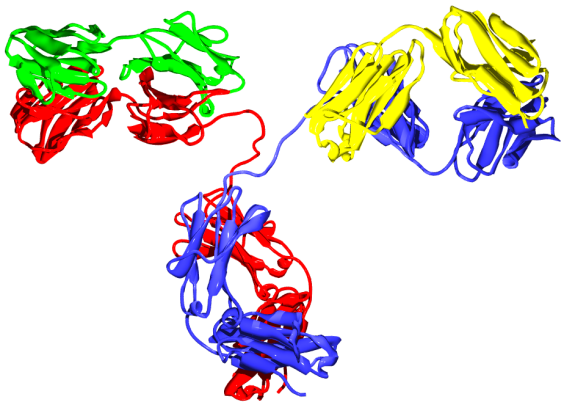
RADIANT

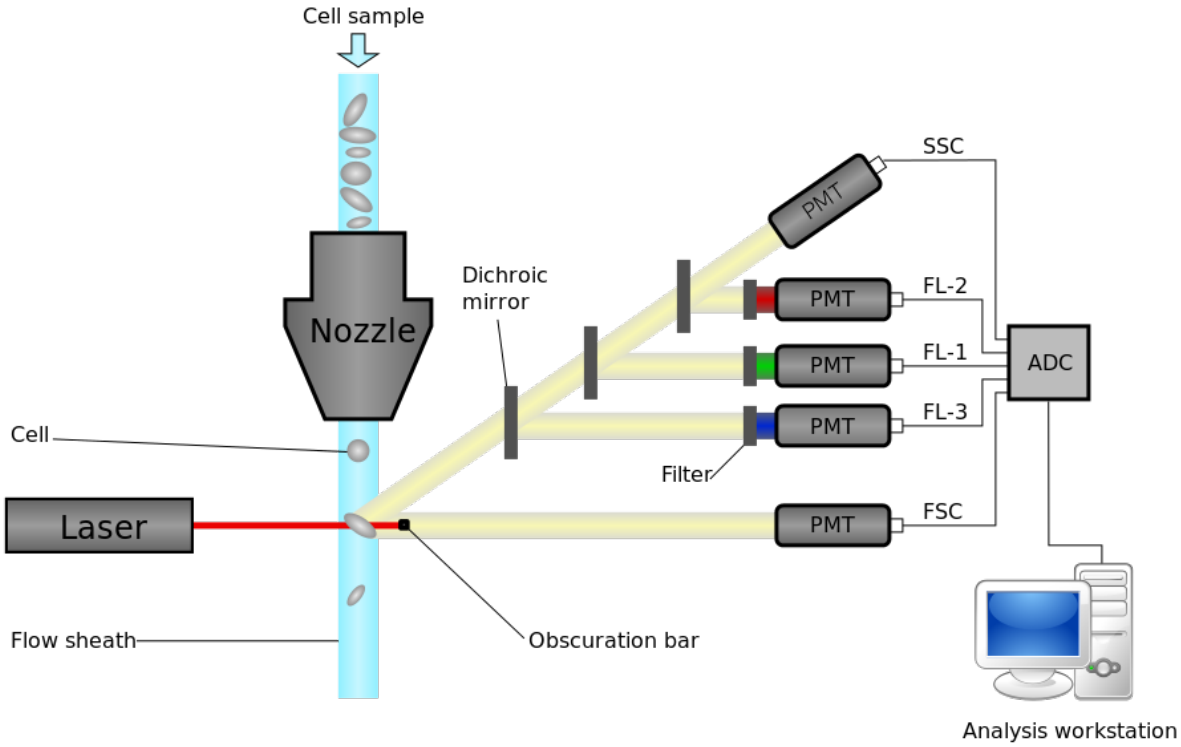
Additional slides

Flow cytometry (FACS)



Antibodies

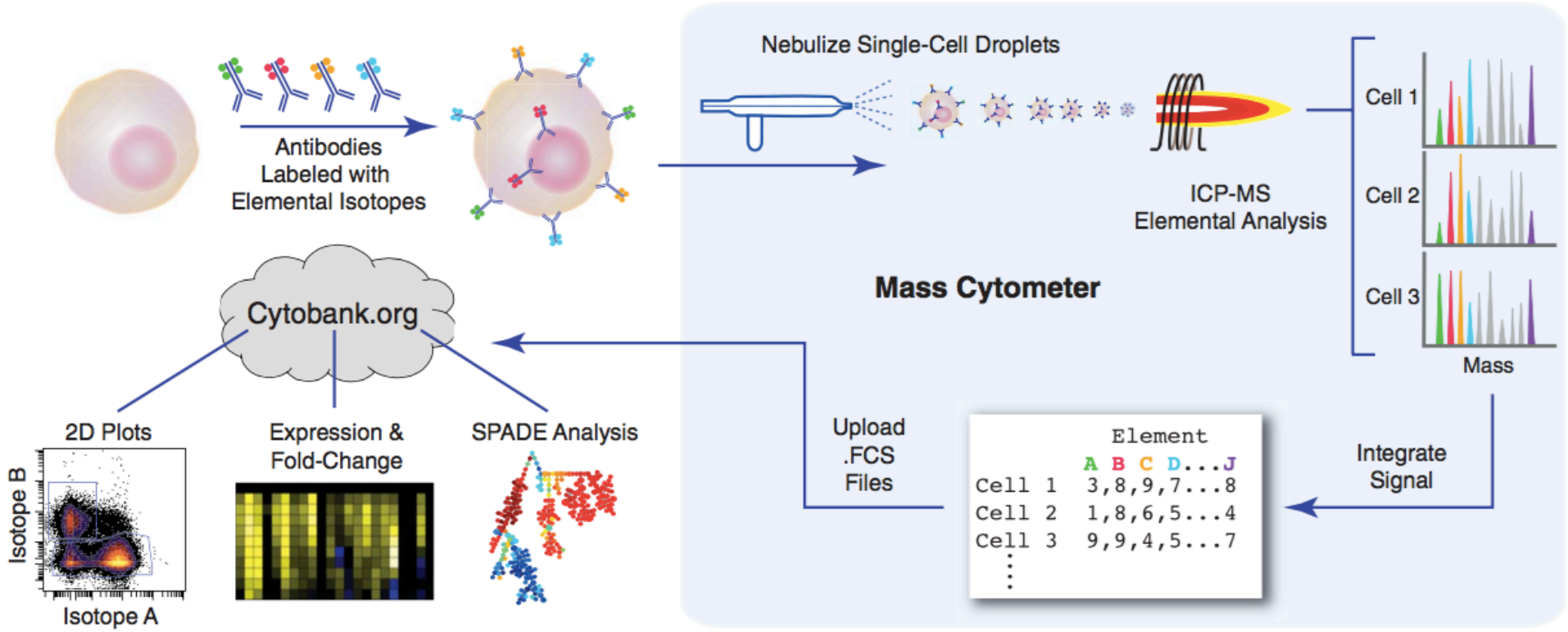




Flow cytometry schematic

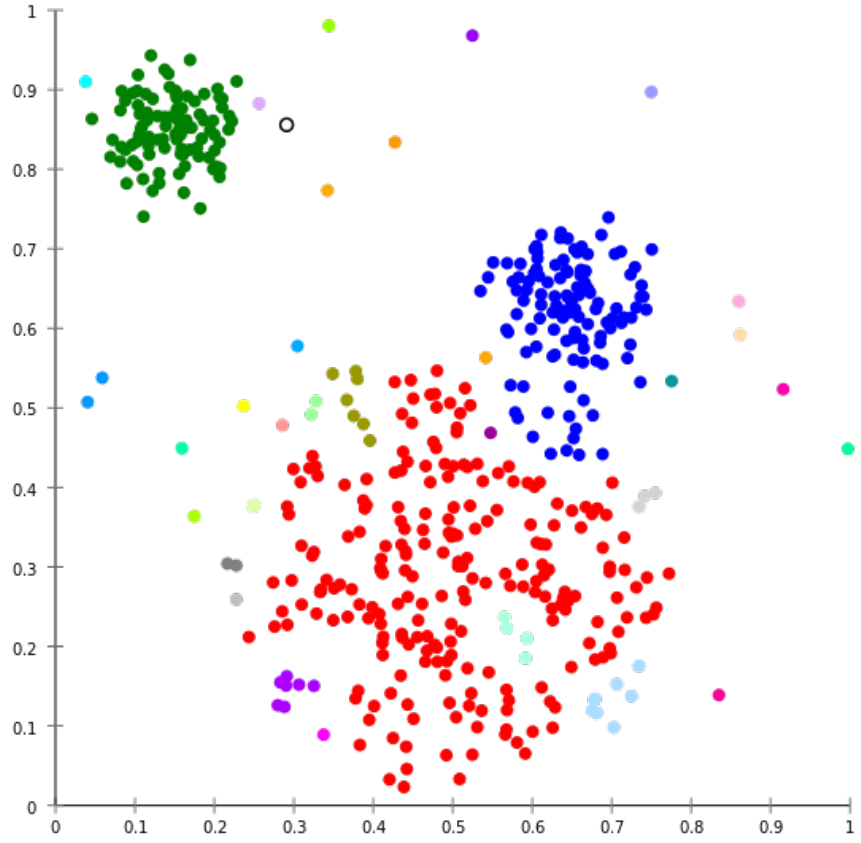
Wikipedia

Mass cytometry

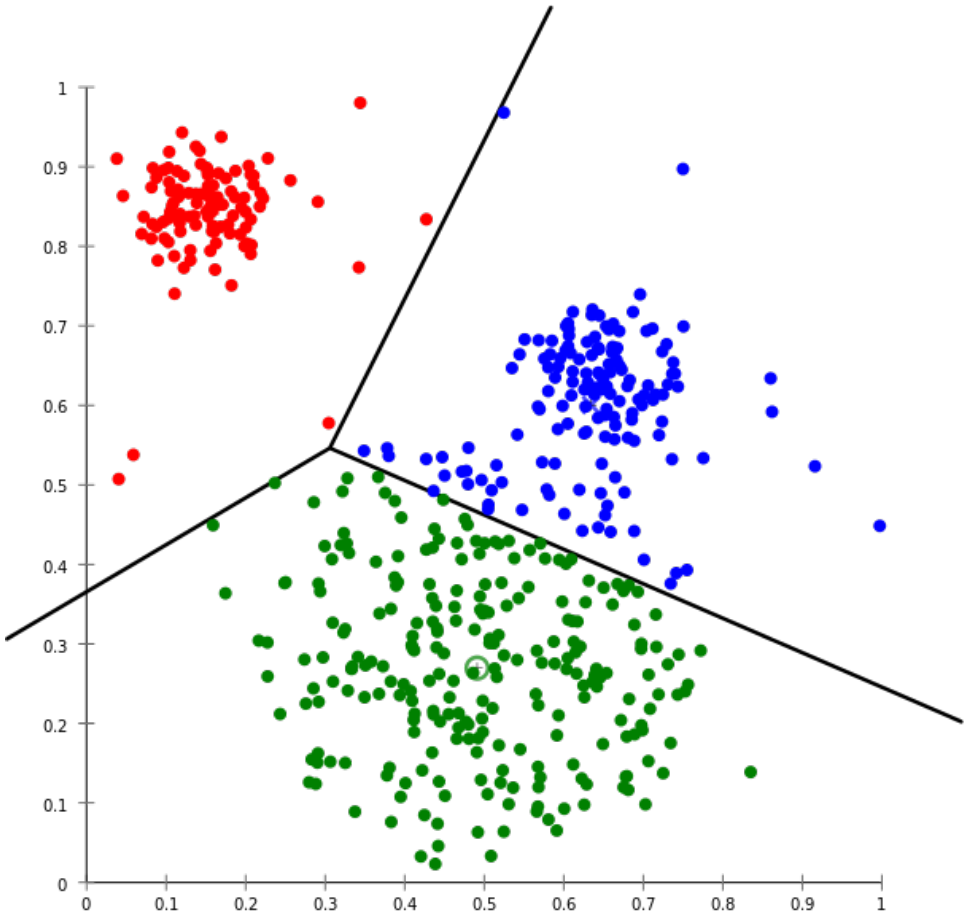


Bendall et al. (2011), Fig. 1A

Cluster analysis



hierarchical clustering



k-means

Wikipedia

F1 score

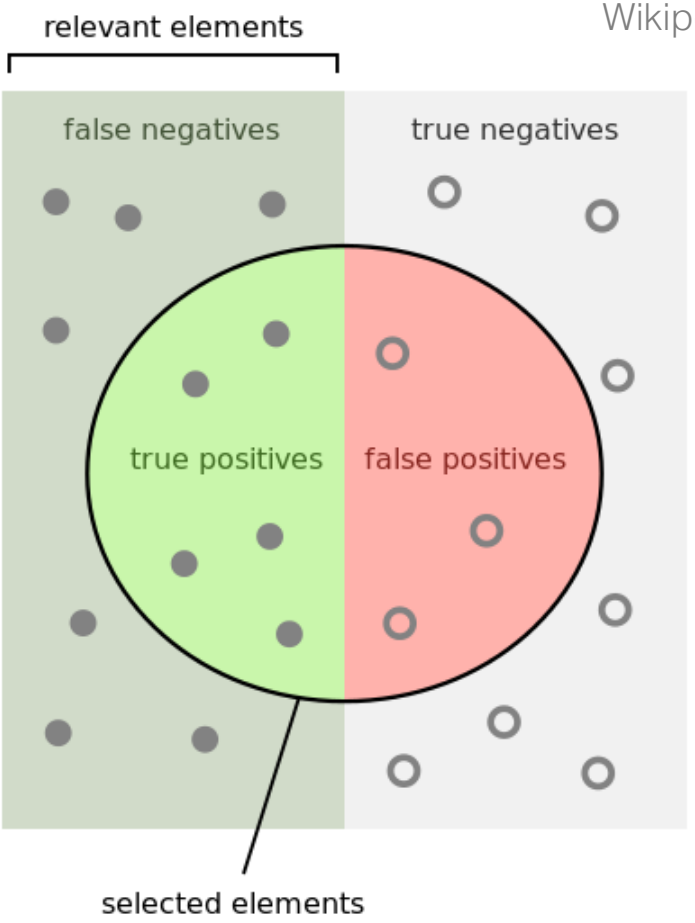
harmonic mean of precision and recall

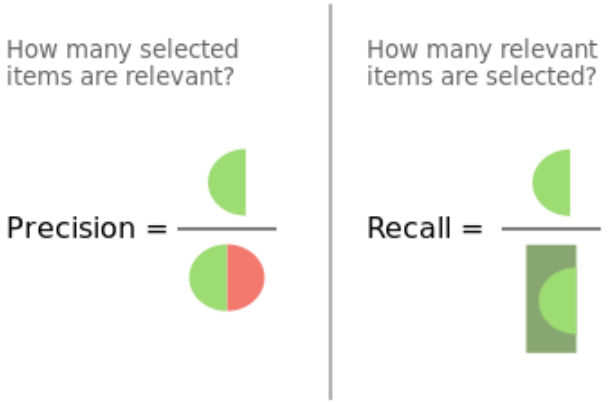
$$F_1 = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

precision = positive predictive value

recall = sensitivity = true positive rate (TPR)

Wikipedia

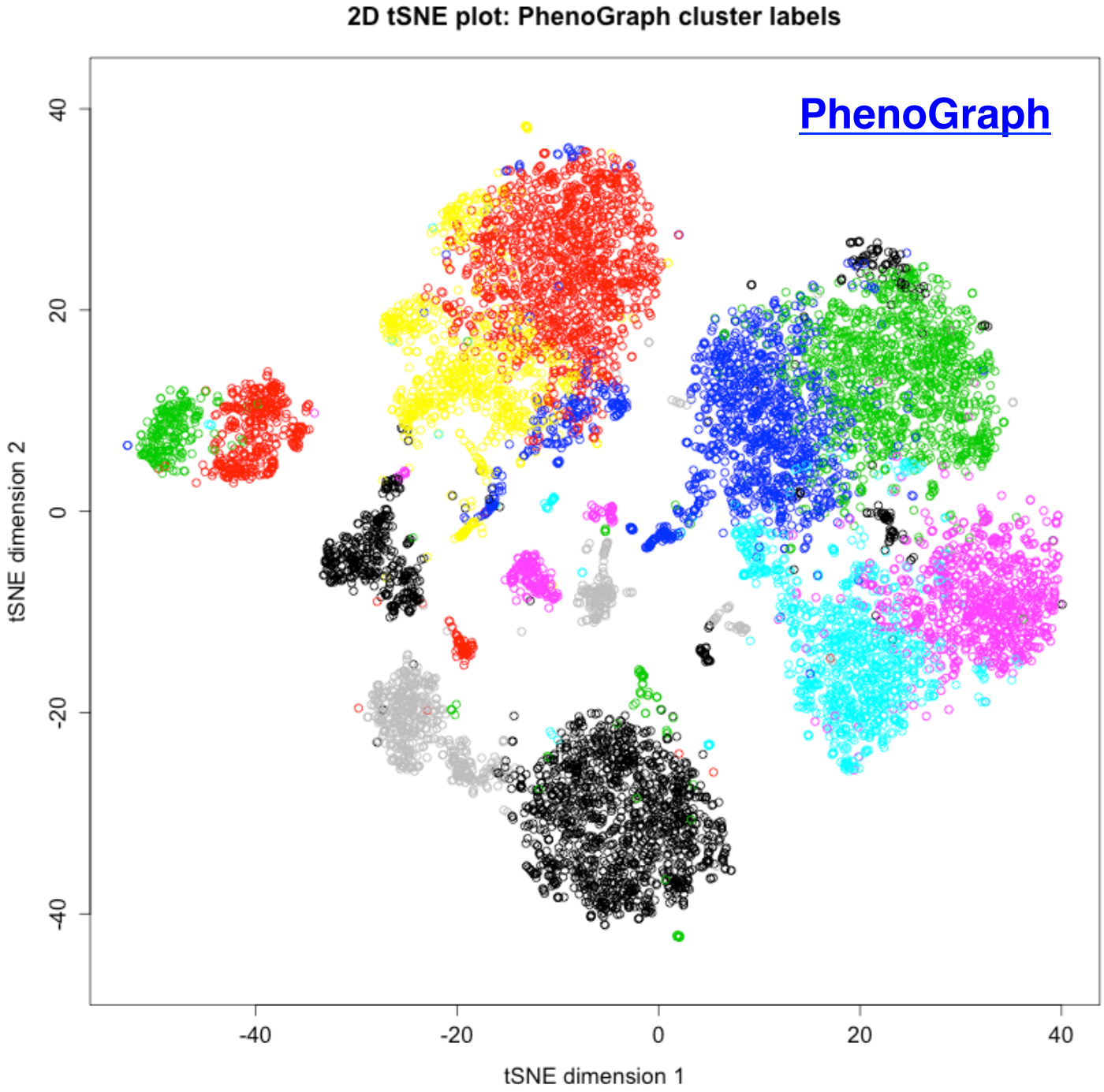




Results

visualization with tSNE / viSNE

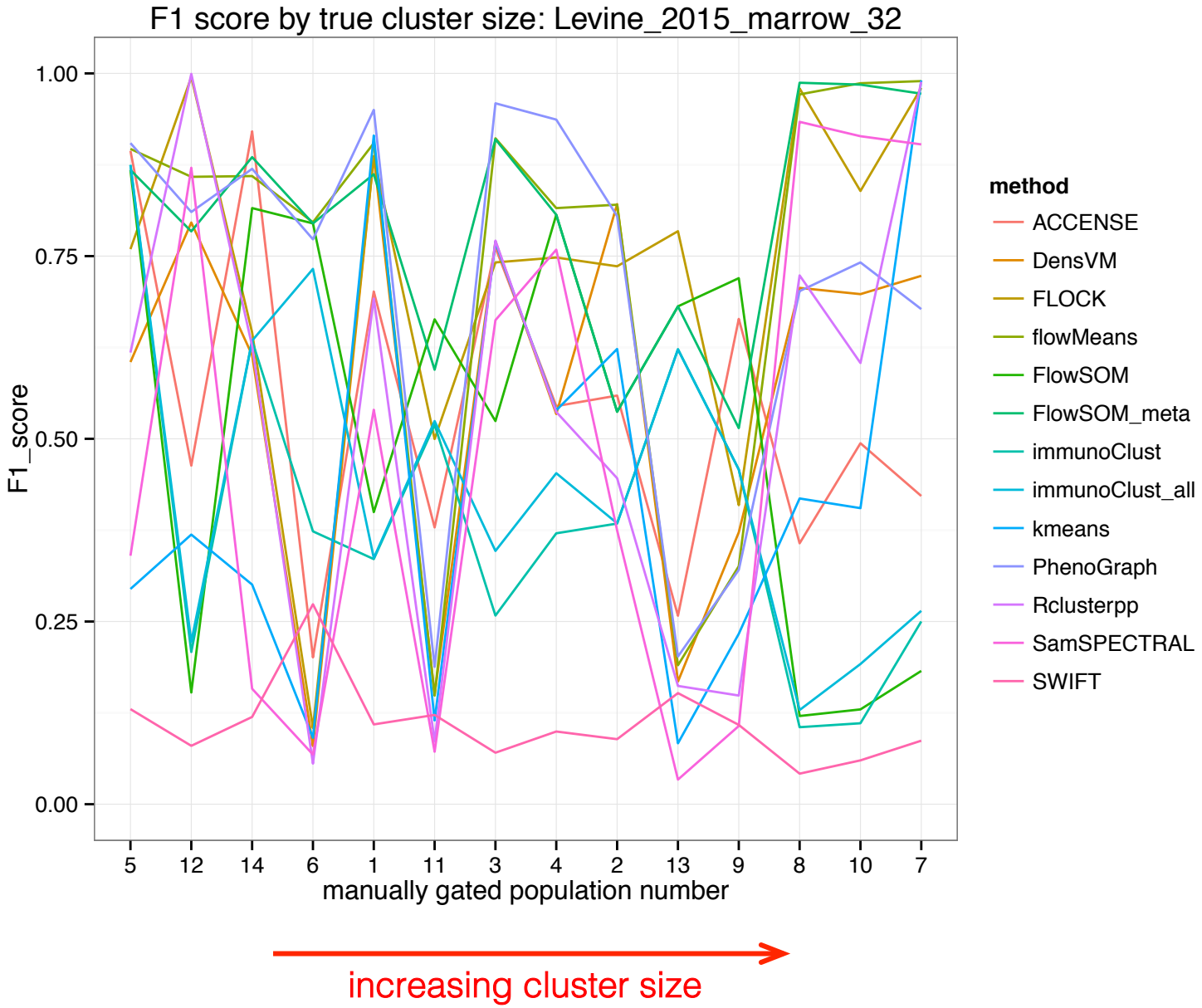
nonlinear 2D projection



Results

F1 score by individual true cluster

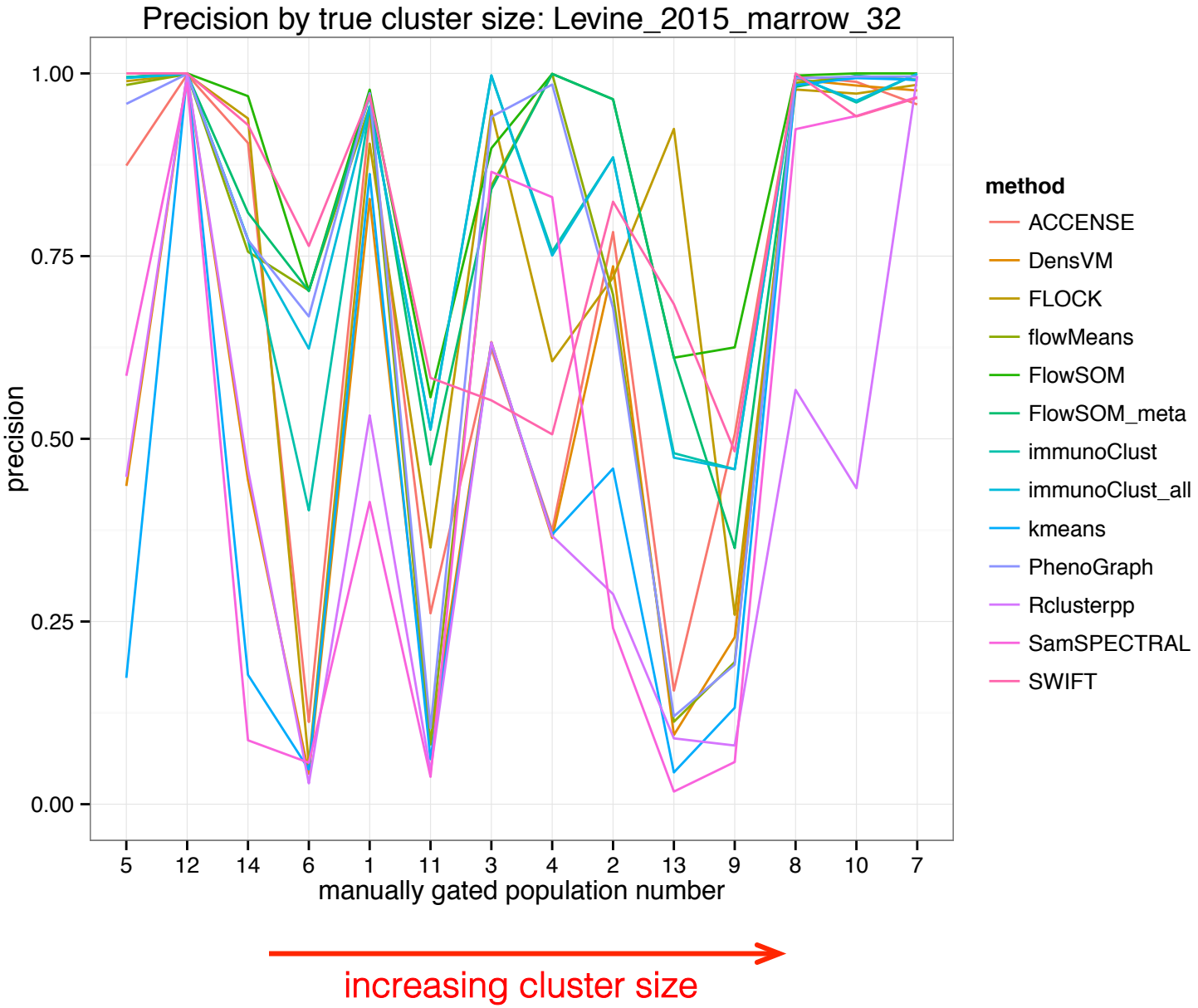
clusters ranked by size



Results

precision by individual true cluster

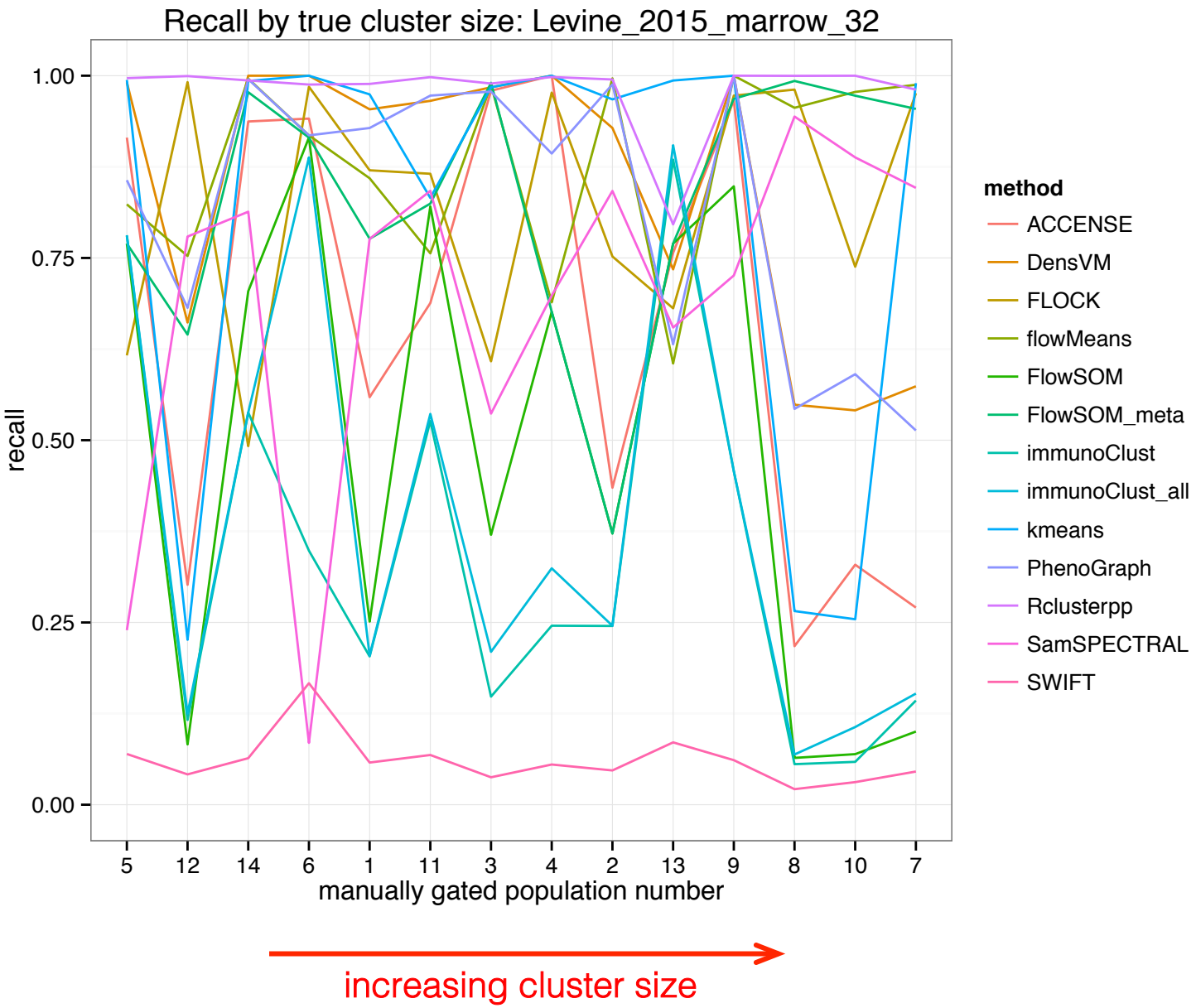
clusters ranked by size



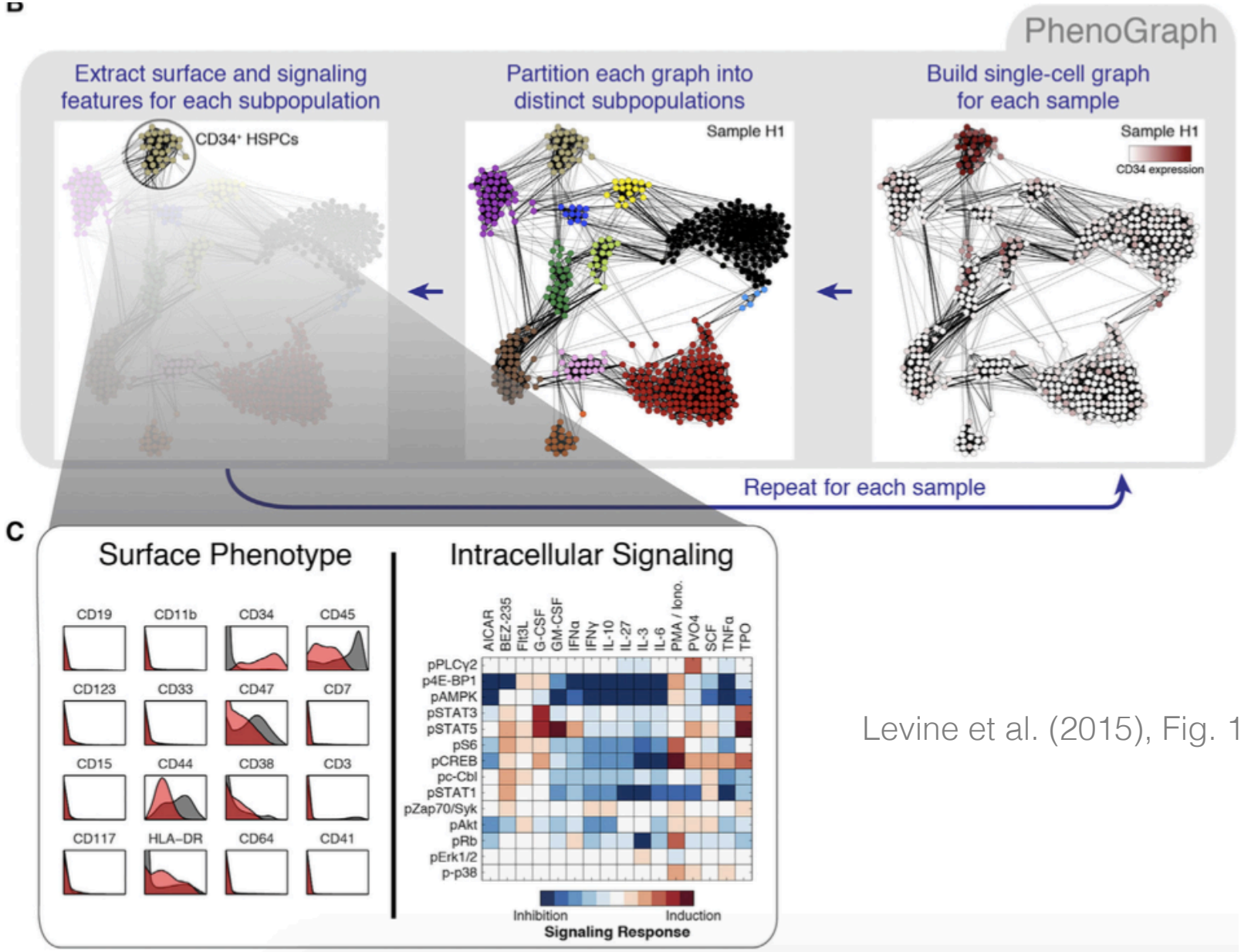
Results

recall by individual true cluster

clusters ranked by size

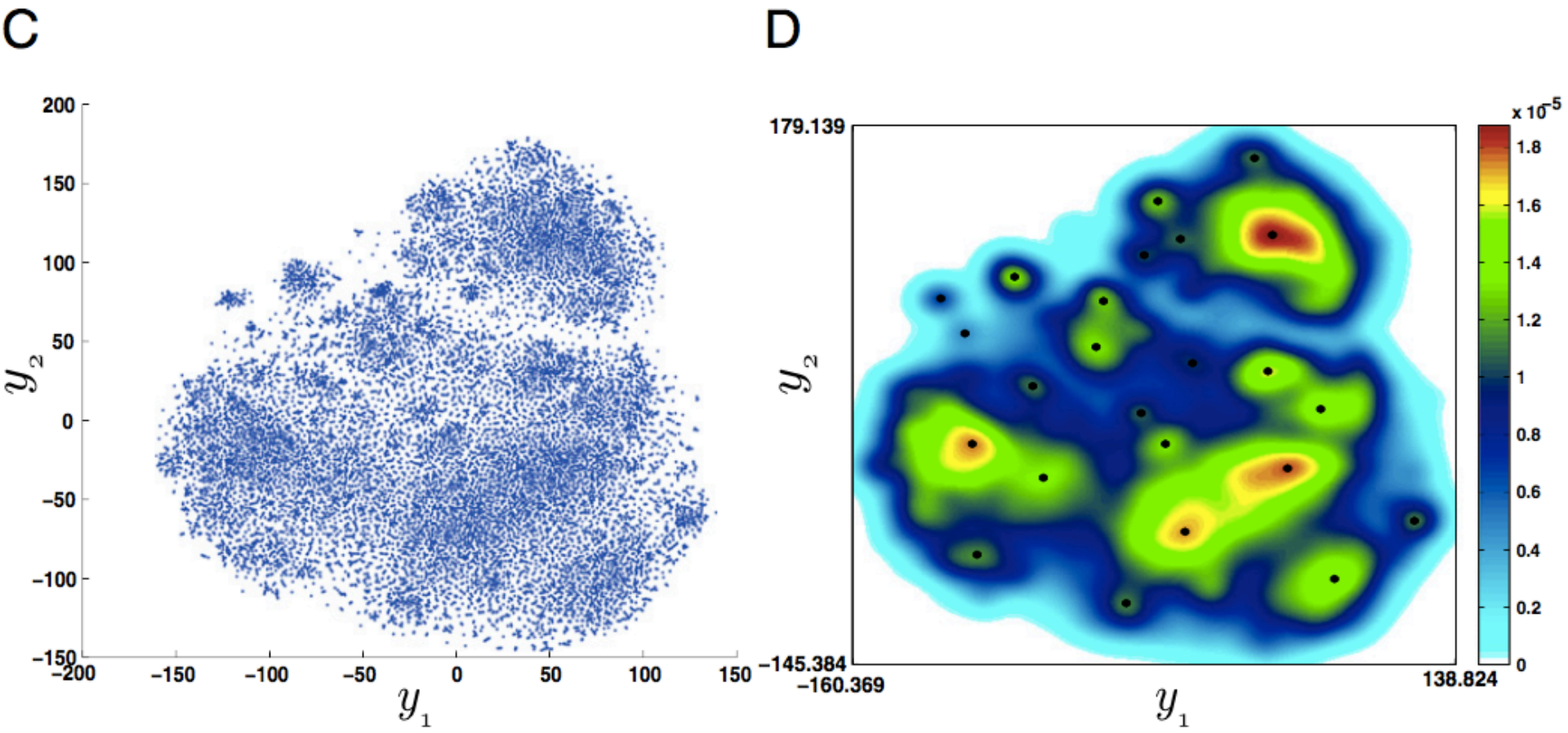


PhenoGraph



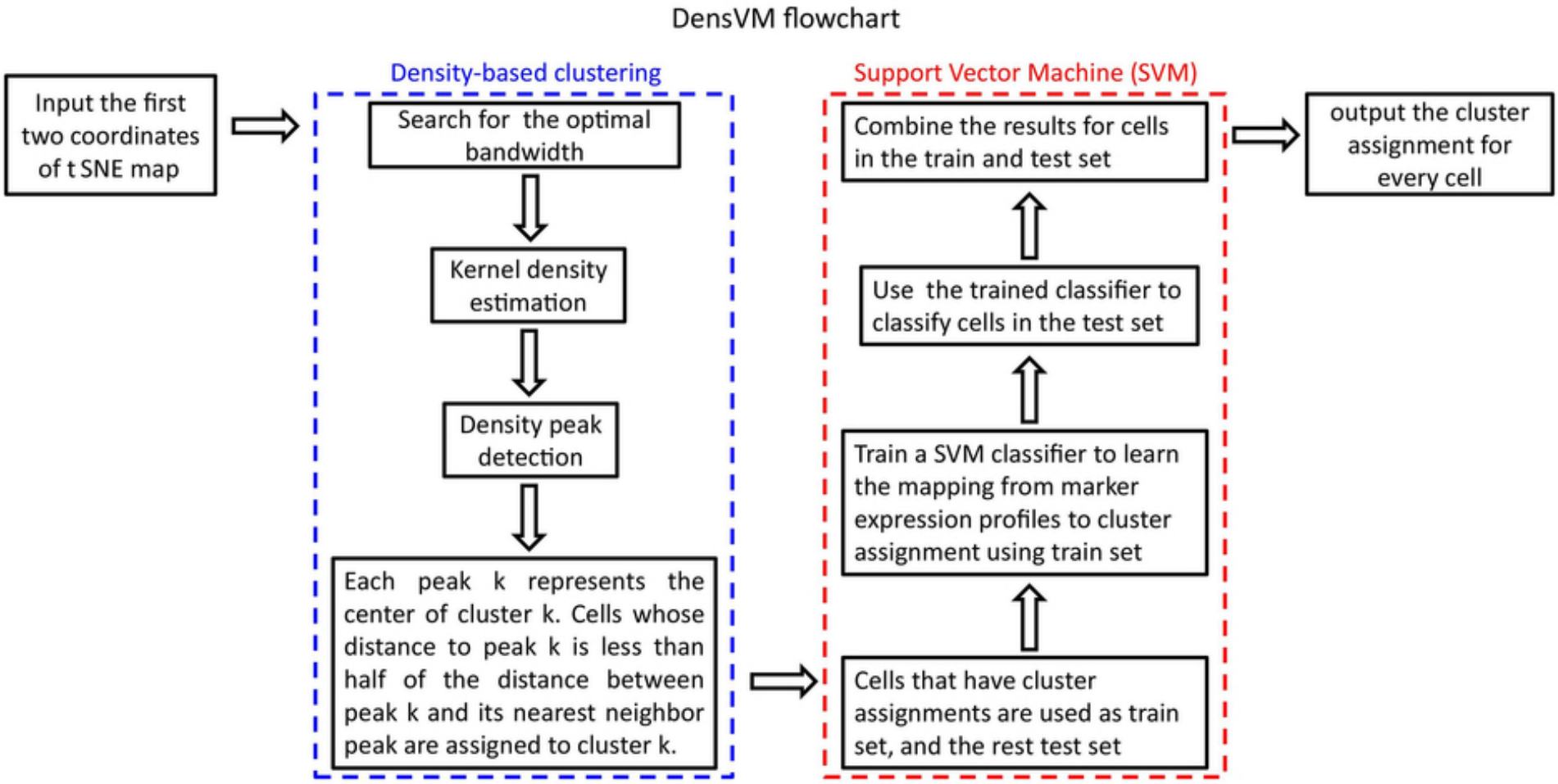
Levine et al. (2015), Fig. 1

ACCENSE



Shekhar et al. (2014), Fig. 1

DensVM



Becher et al. (2014), Supp. Fig. 2

FlowSOM

van Gassen et al. (2015), Fig. 1