University of Zurich[UZH]

SIB Swiss Institute of Bioinformatics

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

IMLS Scientific Retreat 2017

Institute of Molecular Life Sciences
University of Zurich

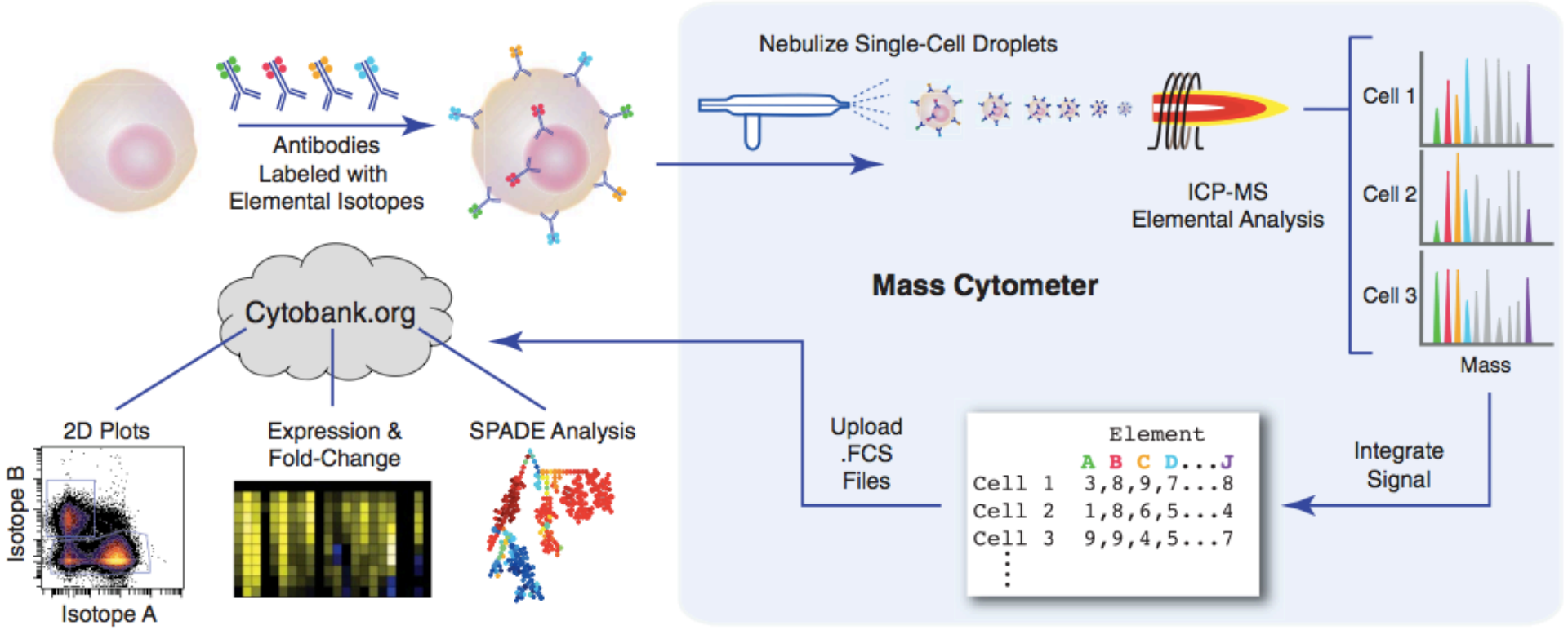
Lukas M. Weber

20 January 2017

Overview

1. High-dimensional cytometry (CyTOF and flow/FACS)
2. Data analysis and clustering
3. Results: comparison of clustering methods
4. Discussion

1. Mass cytometry (CyTOF) and high-dimensional flow cytometry / FACS



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

Mass cytometry (CyTOF)

- Measure protein expression levels in individual cells
  - surface and intracellular proteins

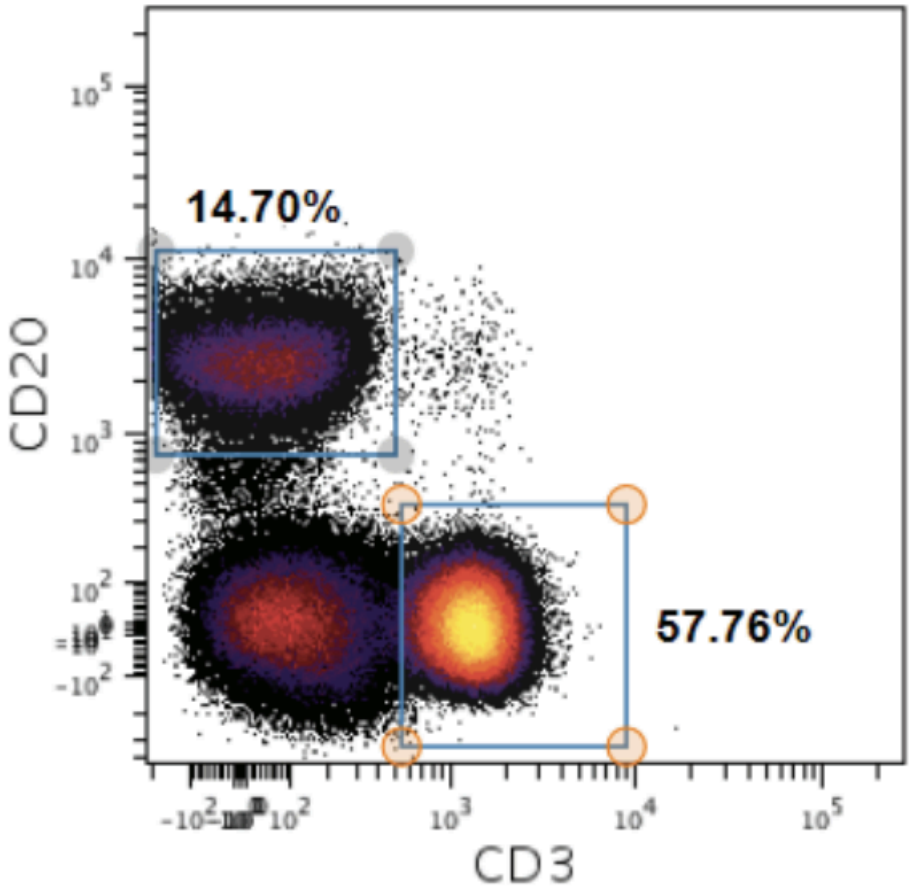
- CyTOF: 30-50 proteins per cell

- Large number of proteins per cell allows cell populations to be characterized in unprecedented detail

- (Latest flow cytometers can also reach 20-30 proteins per cell)

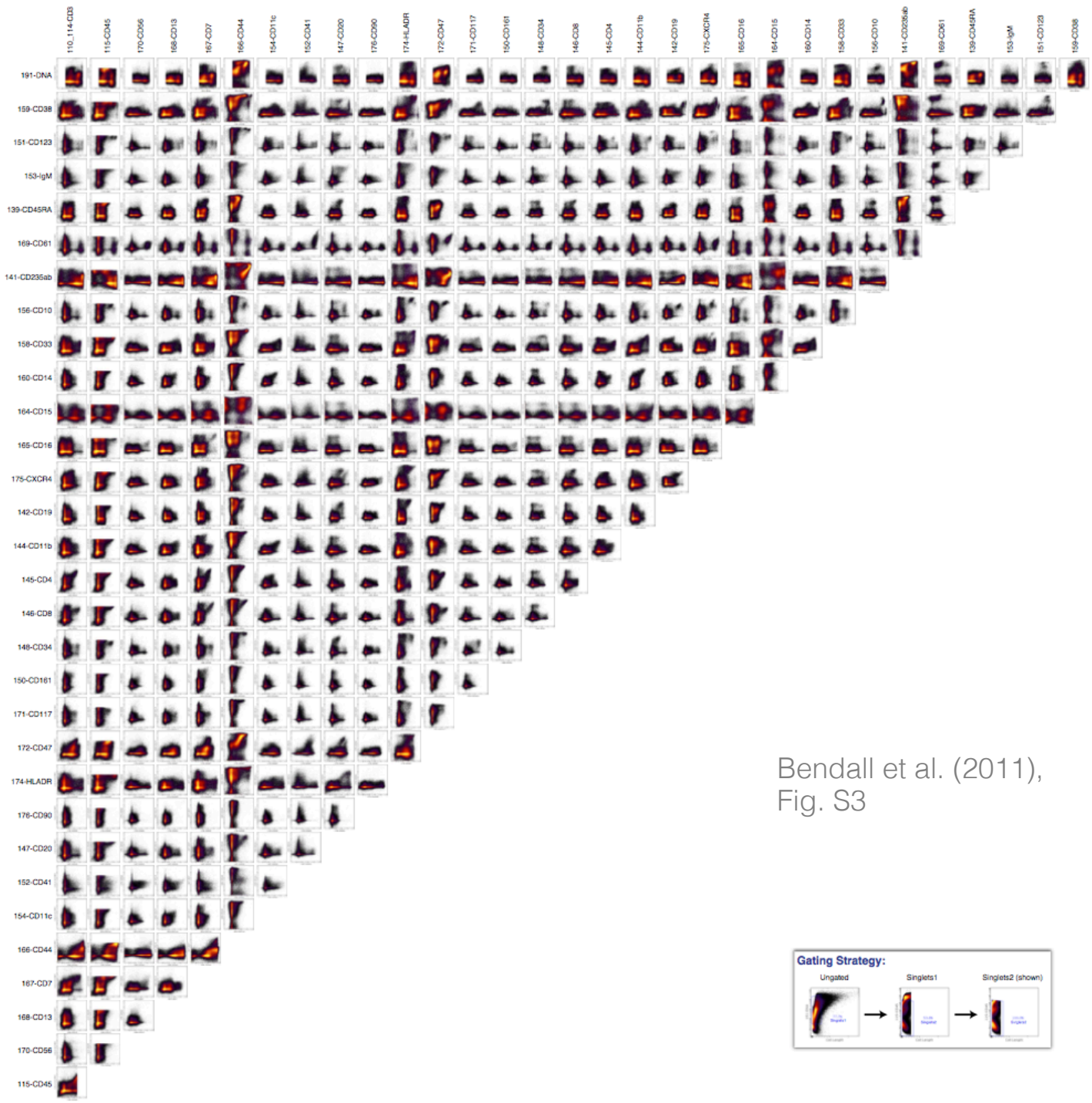
2. Data analysis

- Manual gating



gating example from Cytobank:
https://www.cytobank.org/

Manual gating

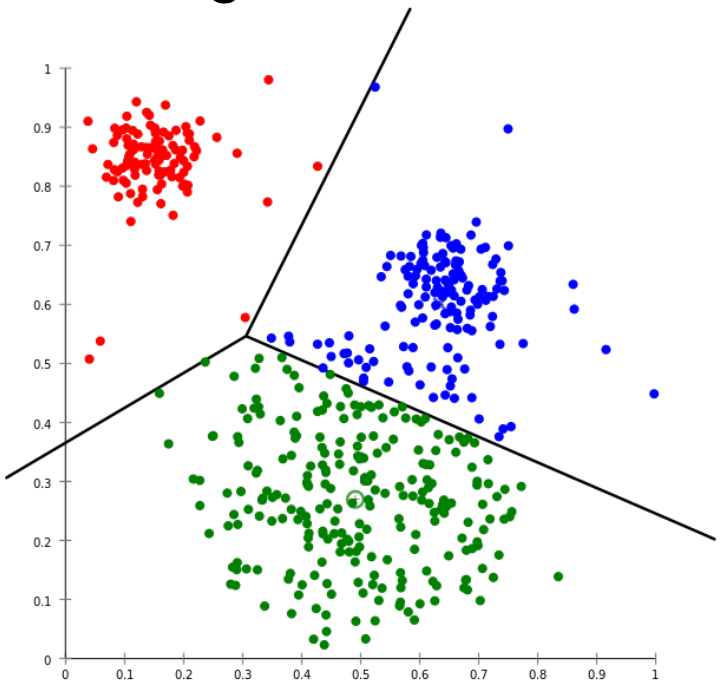


Bendall et al. (2011), Fig. S3

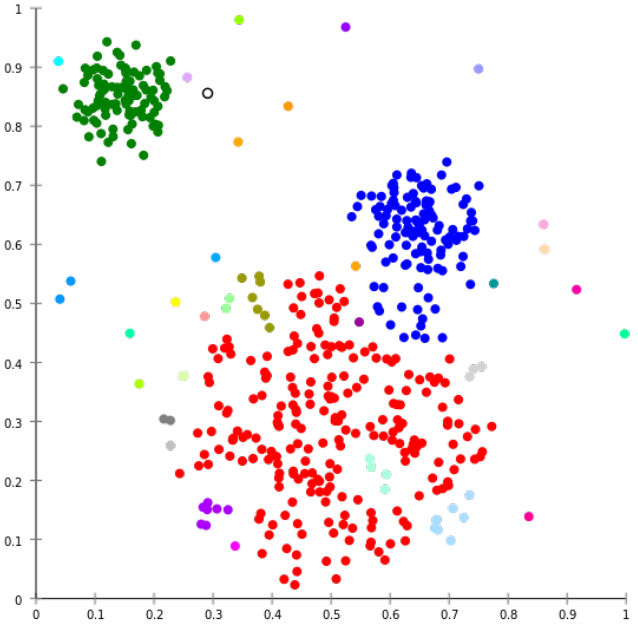
Data analysis

- Automated analysis methods

- Clustering



k-means



hierarchical clustering

Wikipedia

Clustering

- *many different mathematical/statistical/computational approaches*

- *specialized methods for high-dimensional data as well as specific application areas e.g. cytometry*

- FlowCAP-I: systematic comparison of clustering methods for (low-dimensional) flow cytometry data [Aghaeepour et al. 2013]

- no previous comprehensive comparison of clustering methods for high-dimensional cytometry data

3. Our project

- *Comprehensive comparison of clustering methods for high-dimensional flow and mass cytometry (CyTOF) data*
  - 18 clustering methods
  - 6 publicly available data sets from experiments in immunology
  - all major immune cell populations (4 data sets) and specific rare cell populations (2 data sets)
  - evaluation methodology adapted from FlowCAP-I (F1 scores)

Results

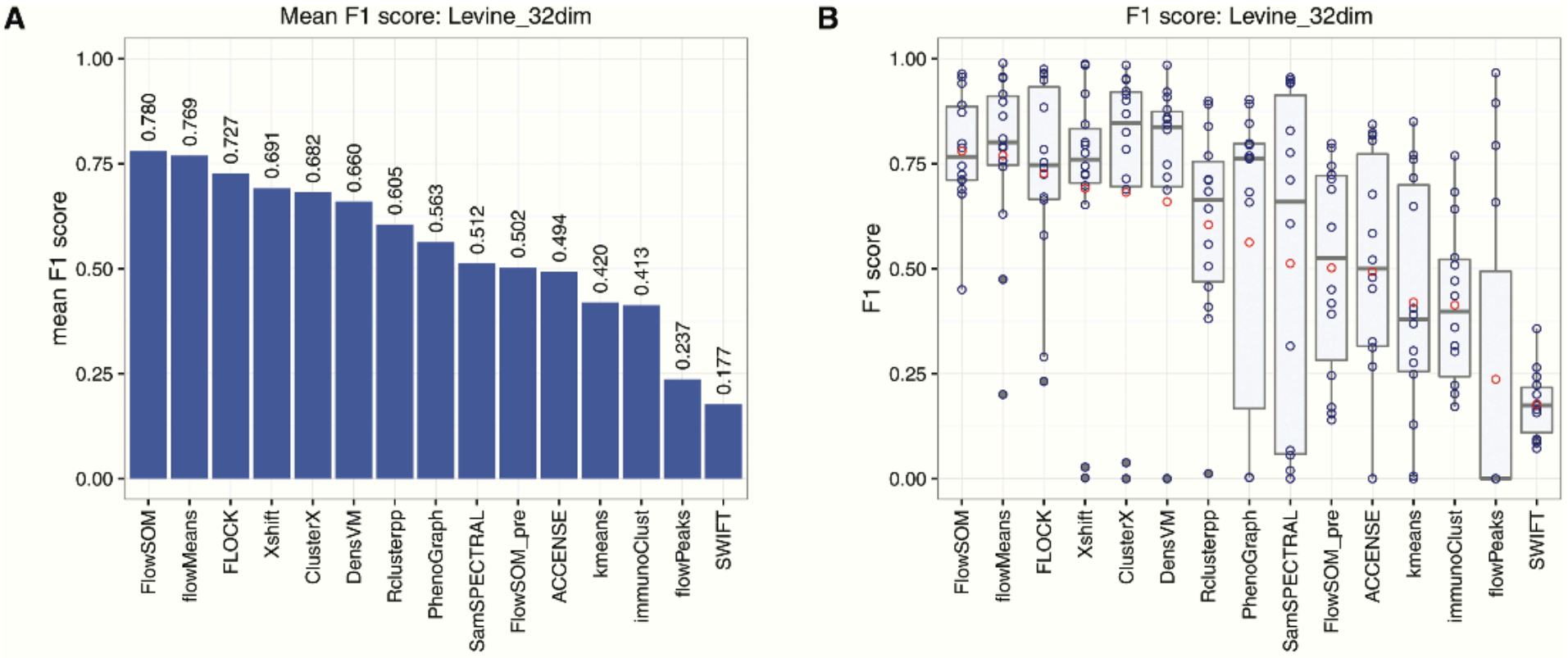
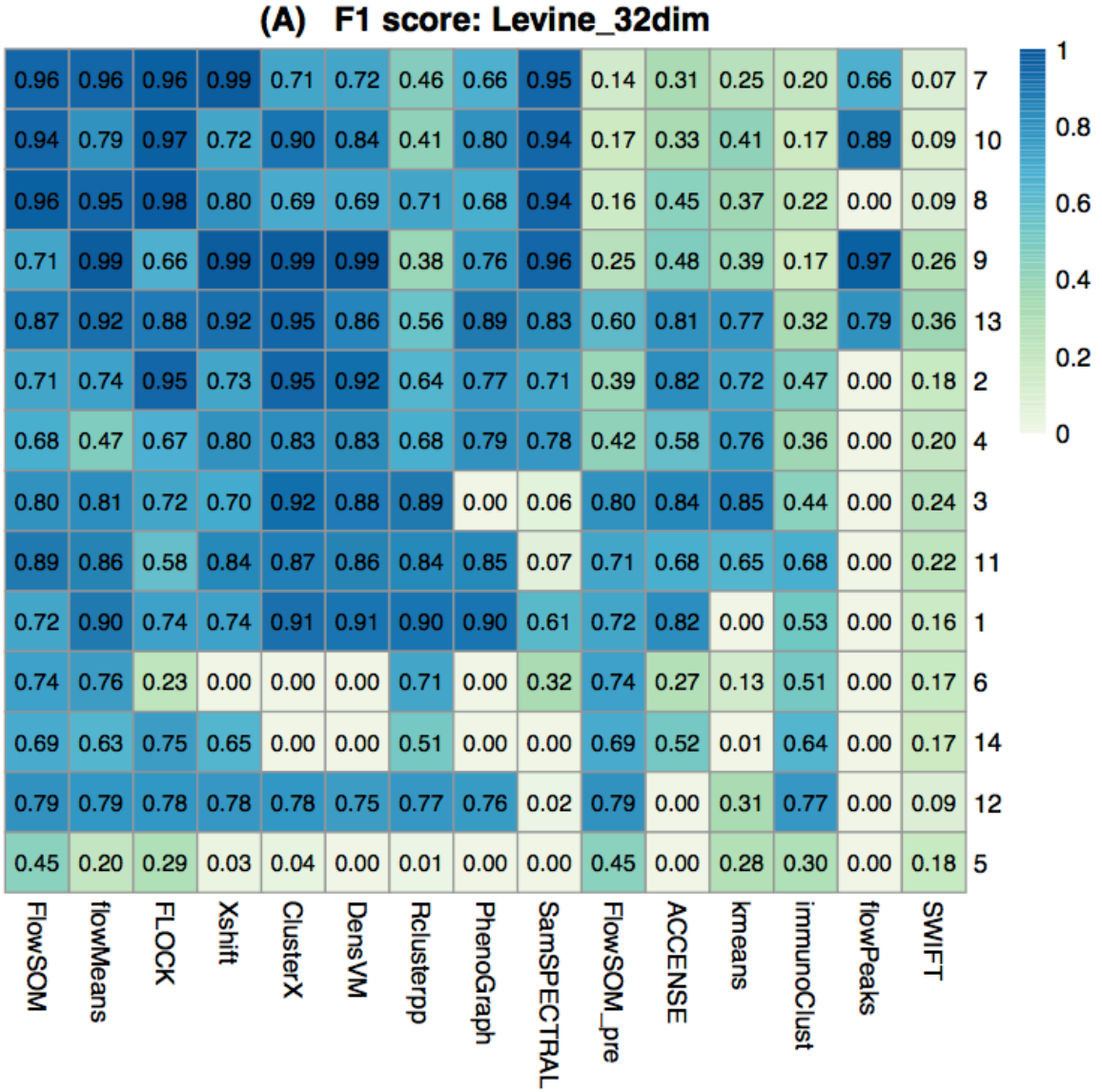


Figure 1

Results



Supplementary Figure S7

Results

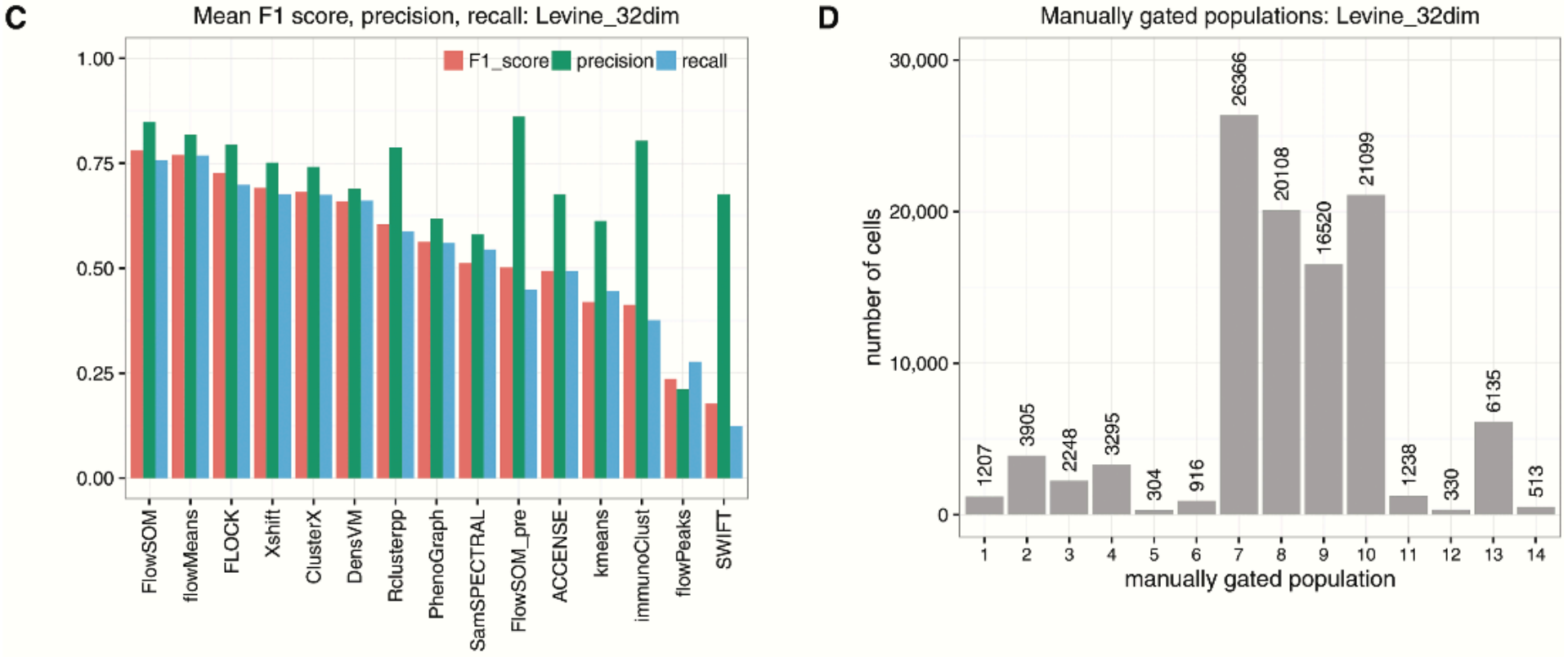


Figure 1

Results

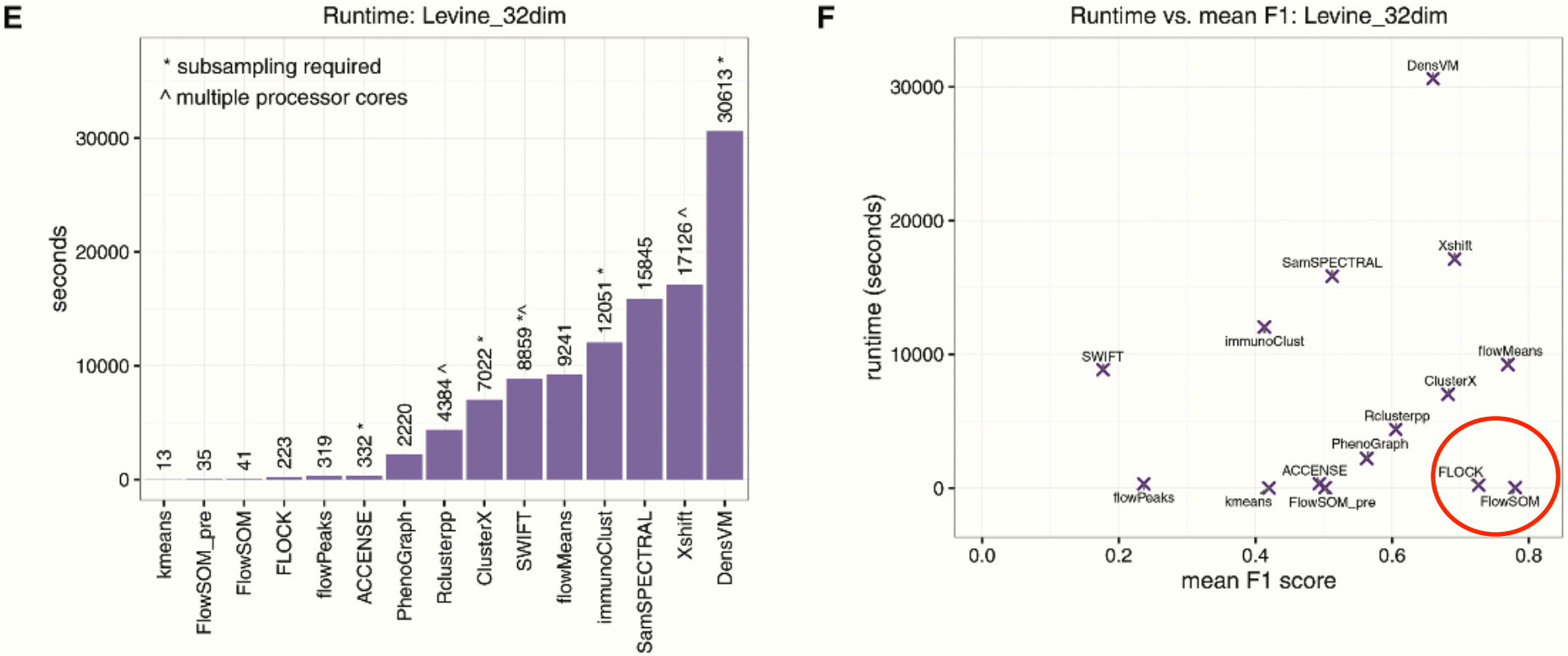


Figure 1

4. Discussion

- Overall results:
  - several methods performed well (i.e. accurately reproduced expert manual gating): *FlowSOM*, *X-shift*, *PhenoGraph*, *Rclusterpp*, *flowMeans*
  - *FlowSOM* gave best performance (for several data sets) and was also one of the fastest methods (this was unexpected!)
  - FlowSOM is fast enough to enable exploratory, interactive analysis on a standard laptop or desktop computer (seconds to minutes)
  - *X-shift* gave best performance for rare cell populations
  - several methods sensitive to random starts (for rare populations)

Discussion

- Now published (December 2016) in *Cytometry Part A:*
  - http://onlinelibrary.wiley.com/doi/10.1002/cyto.a.23030/full

- Open science / open access

  - all code scripts available on GitHub (to facilitate reproducibility and extensions/adaptations): https://github.com/lmweber/cytometry-clustering-comparison

  - data files available on FlowRepository: http://flowrepository.org/id/FR-FCM-ZZPH

  - bioRxiv preprint prior to formal publication (downloads, useful feedback): http://biorxiv.org/content/early/2016/11/10/047613

Thank you!

Acknowledgments

- Mark Robinson
- Charlotte Soneson
- Stéphane Chevrier
- Vito Zanotelli
- Bernd Bodenmiller
- Felix Hartmann
- Nikolay Samusik (Stanford)

- Robinson lab (UZH)
- Baudis lab (UZH)
- von Mering lab (UZH)

Additional slides

Why clustering?

- Clustering is used as an intermediate step within *analysis pipelines*

- Example CyTOF data analysis pipeline:
  - design antibody panel and run experiment
  - *pre-process data* (normalization, compensation, transformation)
  - use *clustering* to detect/define *cell populations*
  - *downstream statistical analysis* (e.g. differential analysis of cell population abundances/frequencies; or differential analysis of functional marker expression within cell populations; between two groups of biological samples e.g. diseased vs. healthy)

Why clustering?

- In general: useful in experiments where *cell populations* are of interest

- Example: Collaboration project with Felix Hartmann, UZH
  - Analyzed differences in median expression of functional markers within cell populations defined by clustering (using FlowSOM) between two groups of samples (narcolepsy patients vs. healthy controls)
  - Hartmann F.J. et al. (2016) *High-dimensional single-cell analysis reveals the immune signature of narcolepsy.* Journal of Experimental Medicine, 213(12), 2621–2633. Available at: https://www.ncbi.nlm.nih.gov/pubmed/27821550

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

Our paper (Weber and Robinson, 2016), Table 1

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)

Our paper
(Weber and Robinson, 2016), Table 2

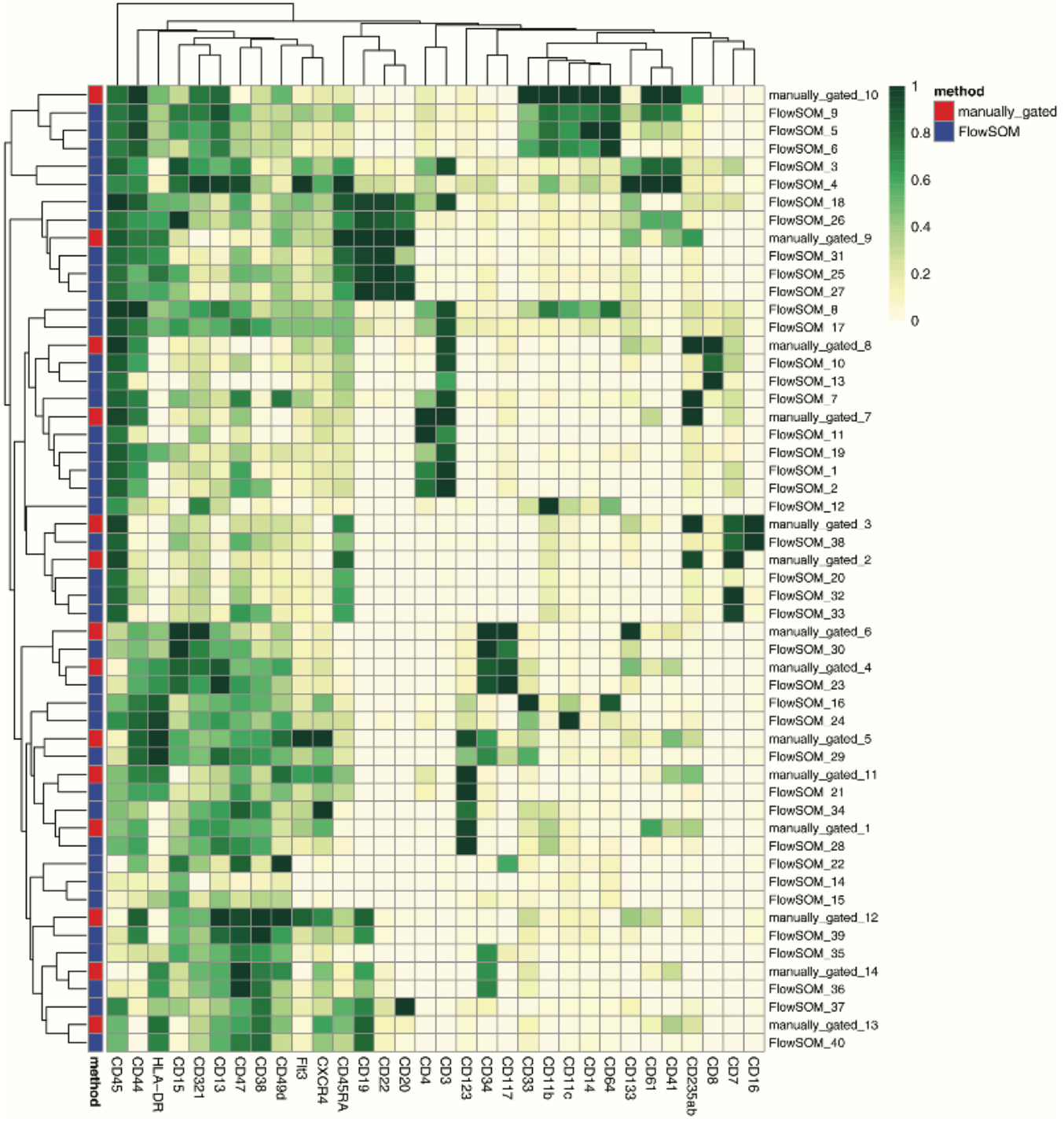
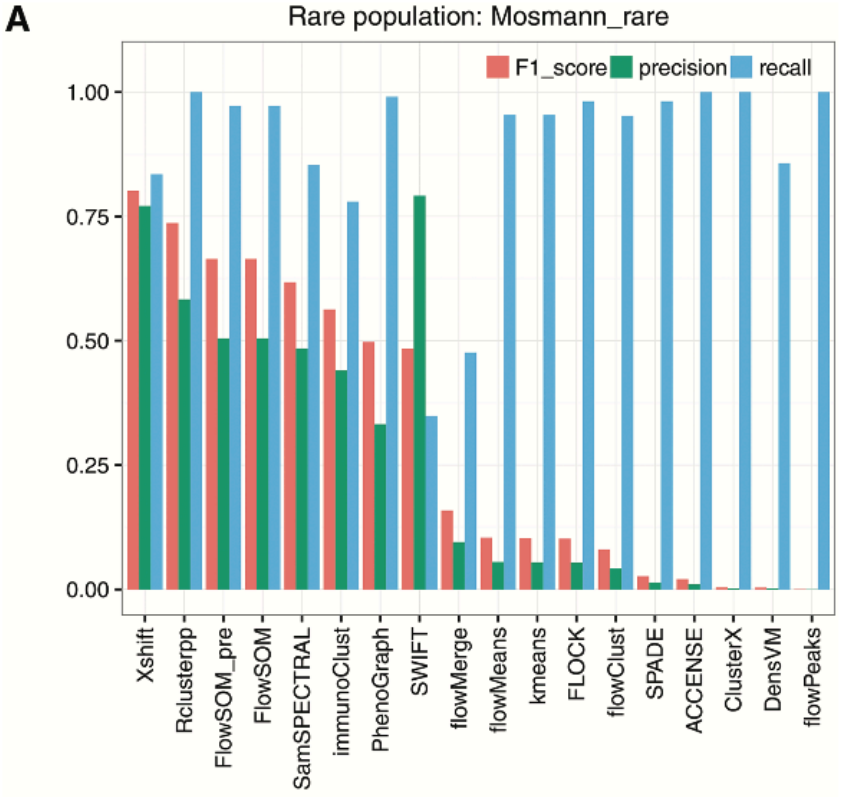
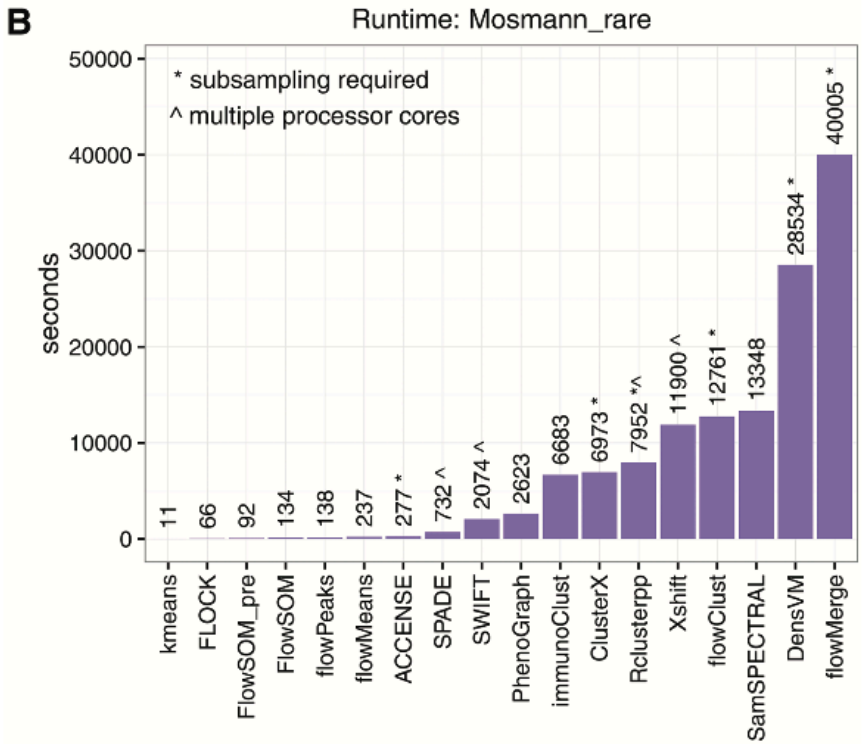

manually_gated_10
FlowSOM_9
FlowSOM_5
FlowSOM_6
FlowSOM_3
FlowSOM_4
FlowSOM_18
FlowSOM_26
manually_gated_9
FlowSOM_31
FlowSOM_25
FlowSOM_27
FlowSOM_8
FlowSOM_17
manually_gated_8
FlowSOM_10
FlowSOM_13
FlowSOM_7
manually_gated_7
FlowSOM_11
FlowSOM_19
FlowSOM_1
FlowSOM_2
FlowSOM_12
manually_gated_3
FlowSOM_38
manually_gated_2
FlowSOM_20
FlowSOM_32
FlowSOM_33
manually_gated_6
FlowSOM_30
manually_gated_4
FlowSOM_23
FlowSOM_16
FlowSOM_24
manually_gated_5
FlowSOM_29
manually_gated_11
FlowSOM_21
FlowSOM_34
manually_gated_1
FlowSOM_28
FlowSOM_22
FlowSOM_14
FlowSOM_15
manually_gated_12
FlowSOM_39
FlowSOM_35
manually_gated_14
FlowSOM_36
FlowSOM_37
manually_gated_13
FlowSOM_40
method
CD45
CD44
HLA−DR
CD15
CD321
CD13
CD47
CD38
CD49d
Flt3
CXCR4
CD45RA
CD19
CD22
CD20
CD4
CD3
CD123
CD34
CD117
CD33
CD11b
CD11c
CD14
CD64
CD133
CD61
CD41
CD235ab
CD8
CD7
CD16
1
0.8
0.6
0.4
0.2
0
method
manually_gated
FlowSOM

Figure 2


A
Rare population: Mosmann_rare
F1_score
precision
recall
1.00
0.75
0.50
0.25
0.00
Xshift
Rclusterpp
FlowSOM_pre
FlowSOM
SamSPECTRAL
immunoClust
PhenoGraph
SWIFT
flowMerge
flowMeans
kmeans
FLOCK
flowClust
SPADE
ACCENSE
ClusterX
DensVM
flowPeaks


B
Runtime: Mosmann_rare
* subsampling required
^ multiple processor cores
seconds
50000
40000
30000
20000
10000
0
11
66
92
134
138
237
277 *
732 ^
2074 ^
2623
6683
6973 *
7952 *^
11900 ^
12761 *
13348
28534 *
40005 *
kmeans
FLOCK
FlowSOM_pre
FlowSOM
flowPeaks
flowMeans
ACCENSE
SPADE
SWIFT
PhenoGraph
immunoClust
ClusterX
Rclusterpp
Xshift
flowClust
SamSPECTRAL
DensVM
flowMerge

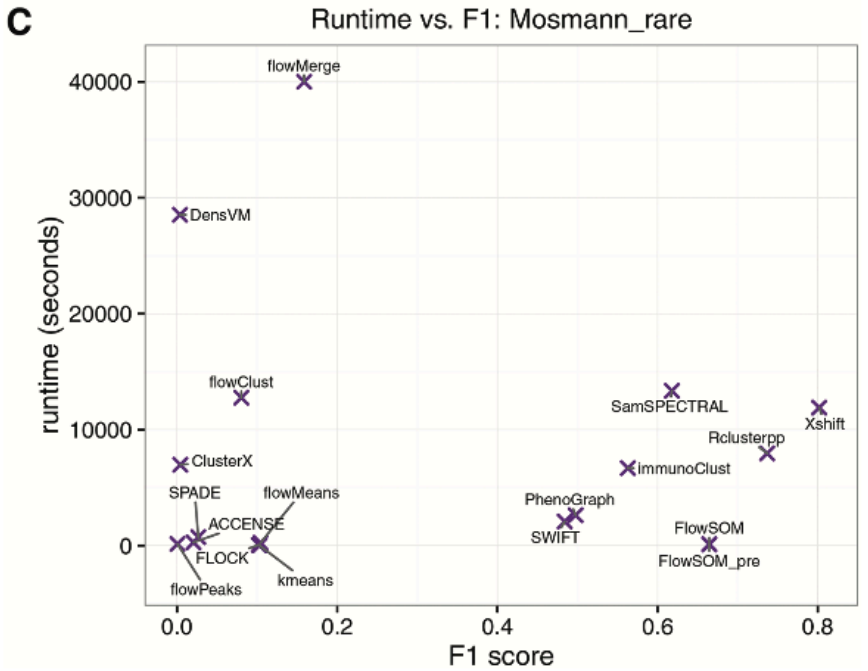

C
Runtime vs. F1: Mosmann_rare
runtime (seconds)
40000
30000
20000
10000
0
flowMerge
DensVM
flowClust
ClusterX
SPADE
flowMeans
ACCENSE
FLOCK
flowPeaks
kmeans
SamSPECTRAL
Xshift
Rclusterpp
immunoClust
PhenoGraph
SWIFT
FlowSOM
FlowSOM_pre
0.0
0.2
0.4
0.6
0.8
F1 score

Figure 3