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

University of Zurich^UZH

Swiss Institute of Bioinformatics

Lukas M. Weber and Mark D. Robinson

Institute of Molecular Life Sciences, University of Zurich, Switzerland

Introduction

Recent advances in mass cytometry (CyTOF) and multicolor flow cytometry enable measurement of 10-50 protein expression levels in thousands of cells per second.

These large, high-dimensional data sets can be used to study cell populations in unprecedented detail; e.g. detection, characterization, and comparison between different biological samples.

Clustering single-cell data

The use of clustering techniques to define cell populations is a key step in many automated analysis pipelines.

A number of new, specialized clustering algorithms for high-dimensional cytometry data have been published in the last 2-3 years.

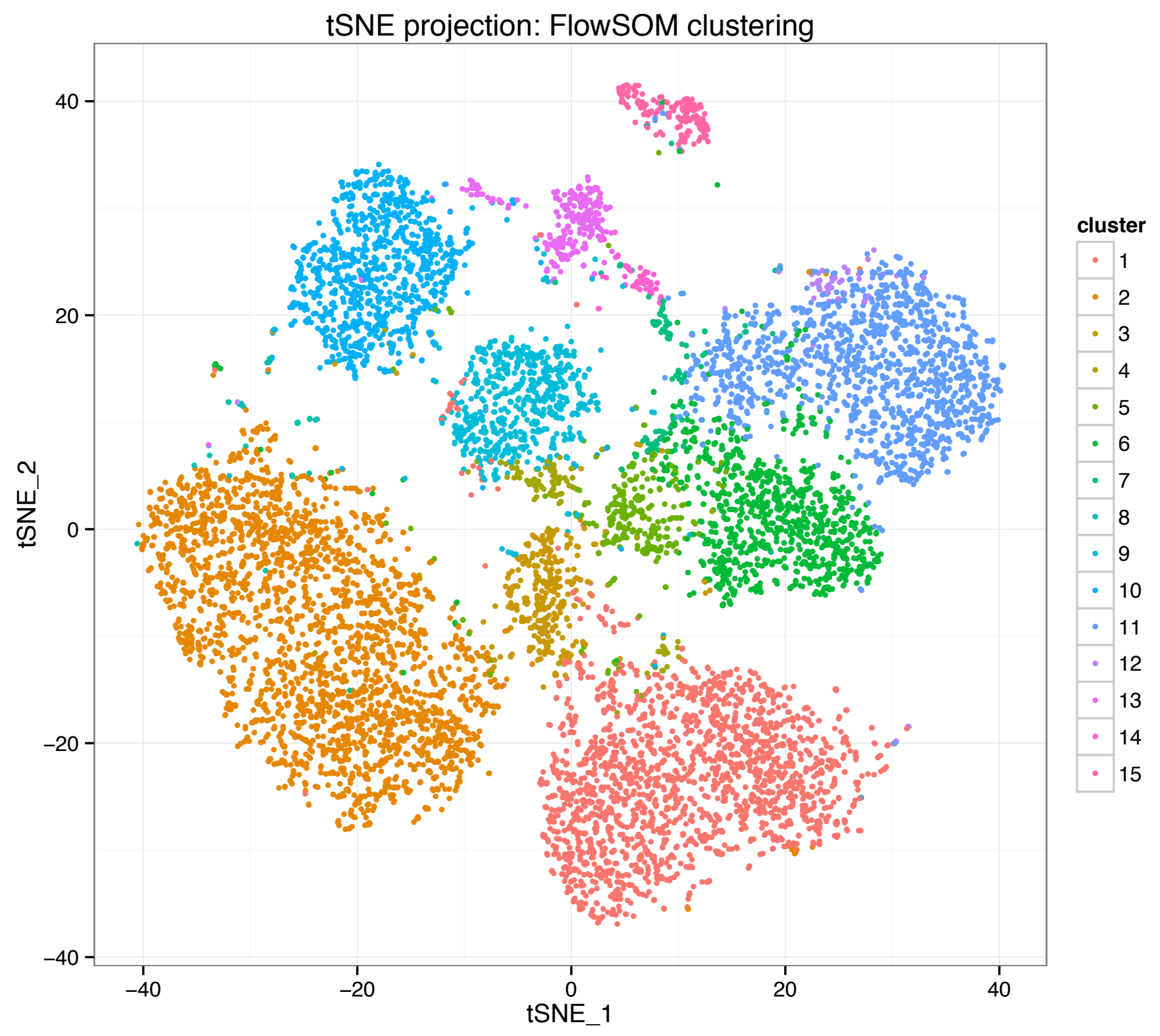


Figure 1. Example of clustering CyTOF data. (Data: Amir et al. 2013)

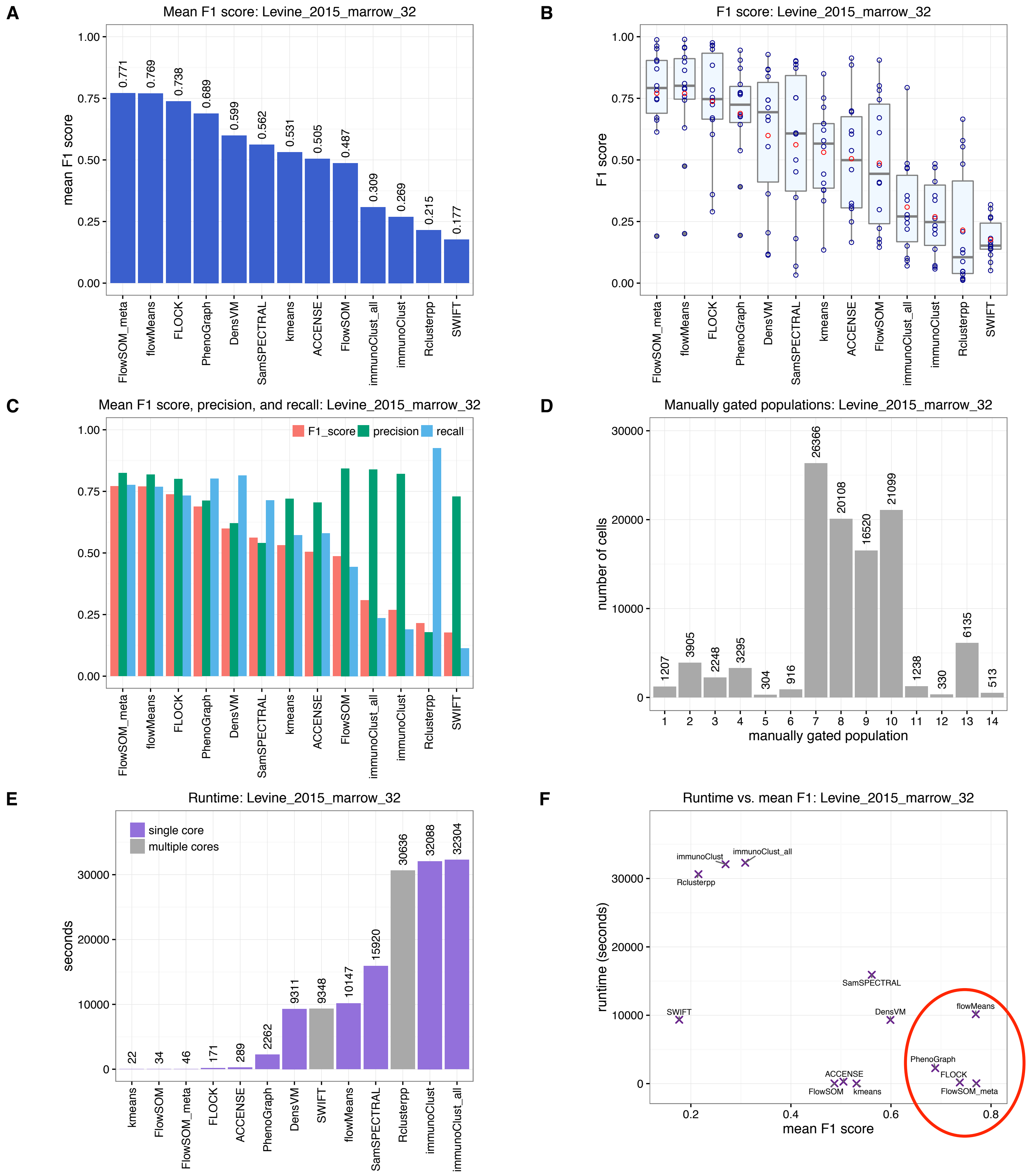


Figure 2. Results for 32-dimensional CyTOF data set from Levine et al. (2015). Data set contains 265,627 cells, 32 surface marker proteins, 14 manually gated populations. Healthy human bone marrow mononuclear cells (BMMCs) from 2 individuals.

(A) Mean F1 scores across populations. (B) Distributions of F1 scores across populations. (C) Mean precision, mean recall, and mean F1 scores. (D) Number of cells. (E) Runtimes. (F) Runtime vs. mean F1 score.

Results for all four data sets available in preprint and supplementary material.

Methods

We compared the performance of leading clustering methods for high-dimensional cytometry data using several publicly available data sets as benchmarks.

Clustering methods: 11 freely available methods (+2 variations): ACCENSE, DensVM, FLOCK, flowMeans, FlowSOM, FlowSOM_meta, immunoClust, immunoClust_all, k-means, PhenoGraph, Rclusterpp, SamSPECTRAL, SWIFT.

Data sets: 4 publicly available benchmark data sets from immunological experiments:

- 2x CyTOF data sets containing multiple cell populations
- 2x flow cytometry data sets with a single rare cell population of interest (0.8% and 0.03% of total cells)

Evaluation criteria: F1 score, precision, recall; using manually gated cell population labels as gold standard. Interpretation of protein expression profiles of detected clusters using hierarchical clustering. Runtimes; stability across random starts.

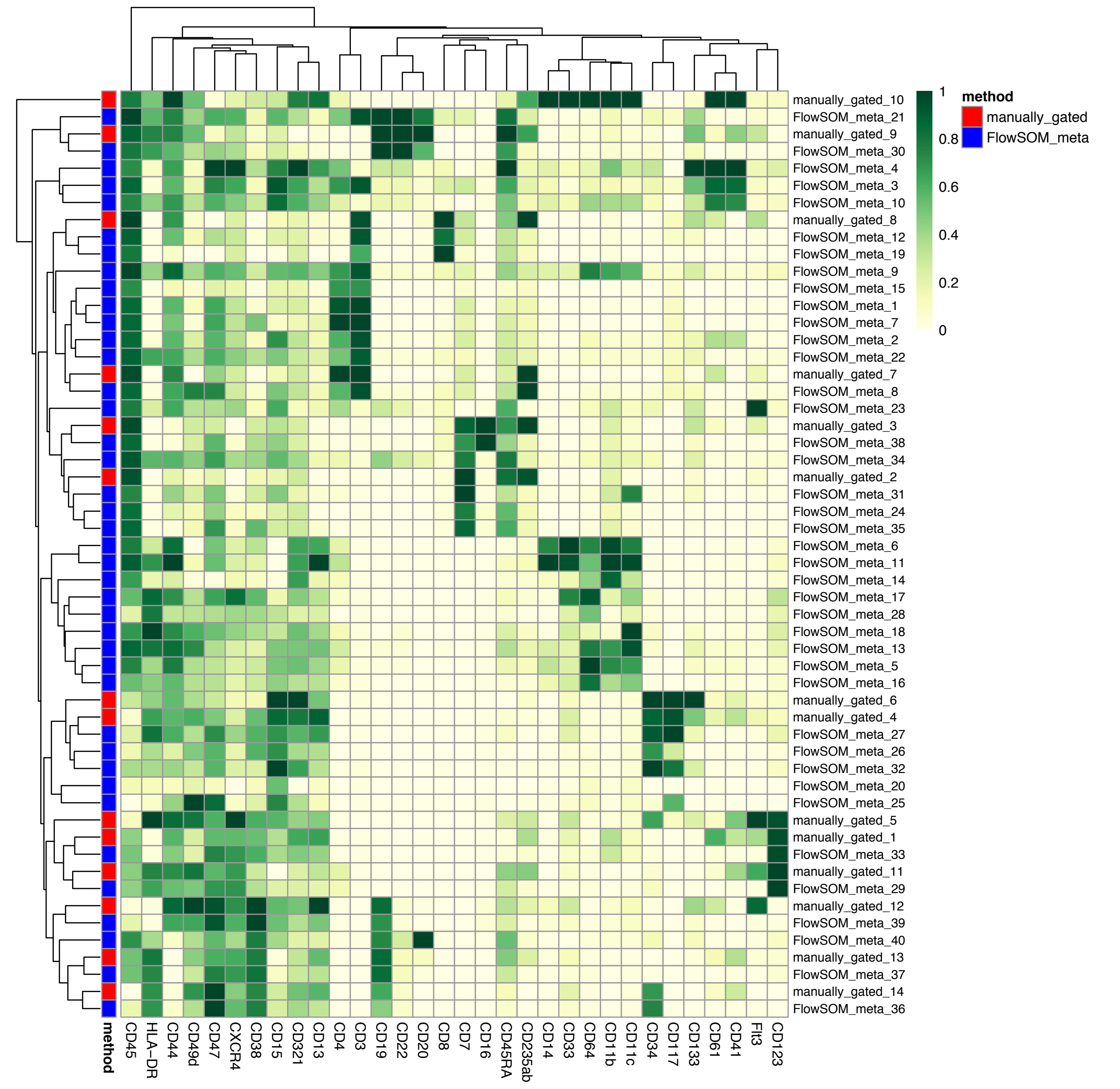
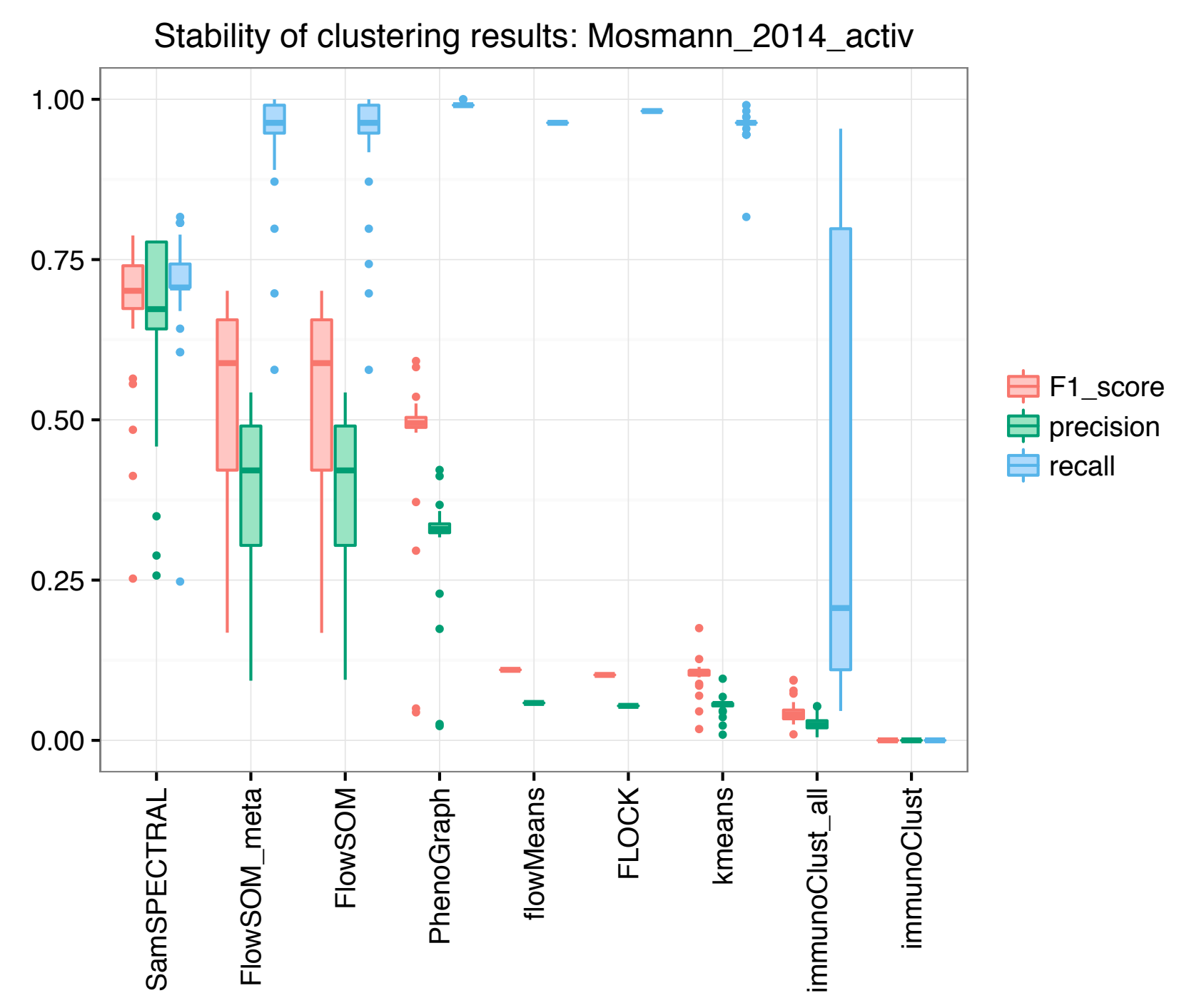


Figure 3. Interpretability of clusters detected by FlowSOM_meta. Heatmap shows median expression intensities of each protein marker (columns), for each detected cluster and manually gated population (rows). Data set: Levine_2015_marrow_32.

Figure 4. Stability of clustering results across random starts. Figure shows variation across 30 random starts, for data set Mosmann_2014_activ (contains a rare population of CD4 T cells; 0.03% of total cells).



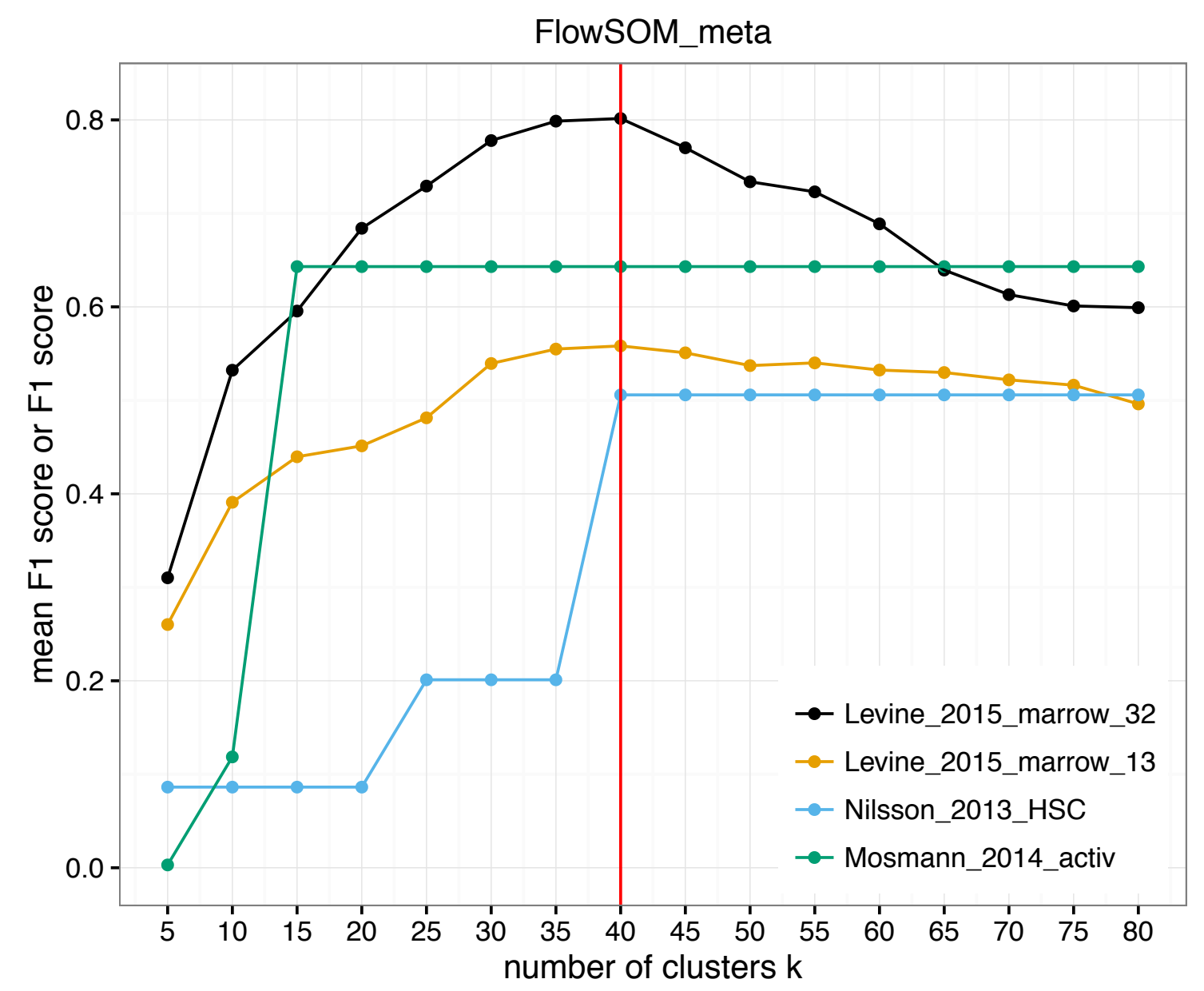


Figure 5. Optimal number of clusters for FlowSOM_meta. The selection of an optimal number of clusters was a key consideration throughout the comparisons; automatic options often did not perform well.

Results and discussion

FlowSOM (with optional meta-clustering but without automatic selection of the number of clusters) performed well across all data sets, and had among the fastest runtimes.

- Runtime in seconds to minutes enables interactive, exploratory analyses on a standard laptop.
- Several other methods also performed well, including *PhenoGraph*, *flowMeans*, and *FLOCK*.

Automatic selection of number of clusters performed poorly for several methods. A simple, direct parameter input to manually select number of clusters was often more practical; but was not available for many methods.

Several clustering methods were sensitive to random starts when detecting rare cell populations.

This study extends previous comparisons (e.g. FlowCAP) by including new clustering methods and focusing on high-dimensional data. Our results provide a practical guide for researchers deciding between clustering methods for analyzing data from CyTOF or high-dimensional flow cytometry experiments.

Availability

bioRxiv (preprint):
http://biorxiv.org/content/early/2016/04/07/047613

GitHub (code scripts):
https://github.com/lmweber/cytometry-clustering-comparison

FlowRepository (data):
https://flowrepository.org/id/FR-FCM-ZZPH

Acknowledgments

University of Zurich: Charlotte Soneson, Stéphane Chevrier, Vito Zanotelli, Bernd Bodenmiller, Felix Hartmann; and members of the Robinson, Baudis, and von Mering labs.

Data sets: We thank members of the following labs for creating the data sets used in this study, and in particular for making these publicly available:

- Pe'er lab (Columbia University)
- Nolan lab (Stanford University)
- Pronk lab (Lund University, Sweden)
- Mosmann lab (University of Rochester)
- Sharma lab (University of Rochester)

Email: lukas.weber@imls.uzh.ch; mark.robinson@imls.uzh.ch