Algorithmic Clustering Of Single-Cell Cytometry Data-How Unsupervised Are These Analyses Really?

Pedersen, Christina Bligaard; Olsen, Lars Rønn

Published in:
Cytometry. Part A

Link to article, DOI:
10.1002/cyto.a.23917

Publication date:
2020

Document Version
Peer reviewed version

Algorithmic clustering of single-cell cytometry data - how unsupervised are these analyses really?

Christina Bligaard Pedersen[1,2] and Lars Rønn Olsen[1,2]

[1] Department of Health Technology, Technical University of Denmark
[2] Center for Genomic Medicine, Rigshospitalet - Copenhagen University Hospital

Contact
chrbl@dtu.dk, +45 50885052

Keywords
immunophenotyping, clustering, visualization, bioinformatics

Funding
This work was funded by the Independent Research Fund Denmark (grant 8048-00078B to LRO).

Running headline
Unsupervised clustering of cytometry data?

Conflict of interest
The authors declare no conflict of interest.

Across single-cell technologies, including flow and mass cytometry, as well as scRNA-seq, unsupervised clustering algorithms have become a staple of data analysis, and are often hailed as a replacement for manual gating with the promise of an unbiased interrogation of the data. There is no shortage of software for the purpose and many tools are produced with user friendly graphical interfaces for the less programming inclined part of the community. The algorithms boast a wide range of features: some excel at detecting rare cell populations, some provide suggestions for the number of different cell subsets in the data, some are fast, some are highly reproducible, etc. Common to almost all of them is that they are oversold on at least one aspect: they almost never provide an unsupervised, unbiased answer at the click of a button, but rather prompt a semi-supervised, iterative, inter-disciplinary process of computational analysis (e.g. by a bioinformatician) and domain expert interpretation (e.g. by an immunologist, haematologist, disease specialist, etc.) until a biologically meaningful clustering is achieved (1) (Figure 1). This is not to say that they are not useful - they certainly are - but the one-click, one size fits all analysis of single cell data remains elusive.

In the wake of heavy developments in algorithms and tools follow extensive testing and reviewing. In a key review of cytometry clustering tools, Robinson & Weber (2016) (2) highlighted a number of algorithms performing well on parameters such as the ability to detect rare or even novel cell populations, the ability to produce results mirroring those achieved by manual gating of the data, the reproducibility of the results from run to run, and the run times of the algorithms. The FlowSOM algorithm (3) came out on top in terms of speed, which combined with good clustering reproducibility has made it a go-to algorithm in studies involving both flow and mass cytometry. The big advantage of FlowSOM and similar

unsupervised clustering approaches over the traditional manual gating have been discussed extensively (1,2,4) with the key conclusion being that algorithmic clustering is not only more convenient than manual gating, but being unbiased by biological preconceptions, it also offers the potential to detect rare populations likely to be missed in manual approaches.

There are, however, a number of features of automated clustering that users need to be aware of. Firstly, mathematically optimal clustering is not the same as biologically meaningful clustering. The unsupervised algorithms remain ignorant of decades of biological research, as well as the technical uncertainty of the data as produced by the various technologies (5,6). We may know for a fact that two markers are never expressed simultaneously on the same cell lineage, but if the expression of all other markers happen to be similar, the algorithm will be none the wiser and likely combine the two cells in single cluster. This property can be argued both a feature and a bug at the same time - unbiased, naive data analysis is more likely to reveal rare or novel cell populations, but given the highly knowledge-based approach to constructing the phenotyping panels in these studies, how unbiased can we really expect the analysis results to be? Secondly, when evaluating the accuracy of clustering algorithms, we face the problem that we lack an objective benchmark - when attempting to expand the horizon of our current knowledge, the truth of course becomes a subjective matter, and even when simply attempting to replicate basic existing knowledge, the benchmark is usually a manually gated population, subject to the gating strategy applied. Lastly, no two algorithms produce the same results, and sometimes this is not even the case for two runs of the same algorithm on the same data. The reason for the latter is that most, if not all, of the widely used algorithms utilize a random start (meaning that unless the same seed is used, non-identical clusterings will result from each run). This is done to speed up these highly computationally demanding algorithms, and to speed up the analysis even further, users will often randomly downsample their data, which will of course also not produce exactly the same results each time a new sampling is done. The effects of these tricks range from basically undetectable to sometimes affecting the downstream biological interpretation of the results (2,7,8). The bottomline is this: the workflow in algorithmic clustering of cytometry data is rarely an unsupervised process, but more likely a semi-supervised iterative process of computational analysis and domain expert interpretation.

In this issue of Cytometry Part A (page XXX), Lacombe et al. describe an approach for analysis of flow cytometry data using FlowSOM (3) for clustering and the commercially licensed software Kaluza for interpreting the results, thereby providing a framework facilitating efficient application of the semi-supervised approach to clustering. By using Kaluza for post-processing, it is not only made easy for non-programmers to interrogate the resulting clusters using mean fluorescence intensity, cell numbers and percentages, and 2D histograms, but the labeling strategy can also be saved as a protocol for future use. One additional highlight in this work, is the use of two approaches for assigning additional “case” cells to existing “control” clusters: one based solely on healthy reference samples, and one including samples from both healthy individuals and leukemia patients. By first clustering and labeling healthy samples alone and subsequently ‘mapping’ the disease samples of interest to the predefined clusters of the healthy cells, it is possible to gain more control of the output as the healthy hematopoietic lineages are more easily defined than malignant ones. A similar approach has previously been suggested for mass cytometry data (9), but, as discussed by Lacombe et al. (in this issue page XXX), while the projection of diseased cells -

in their case leukemic cells - into the predefined clusters of healthy cells can be beneficial for the stable subsets, it also limits the opportunity for novel discoveries of populations that are unique to patient samples. As a result, the proposed method also includes a clustering scheme in which leukemic and healthy samples are clustered together, as means to detect minimal residual disease cells that are solely present in the leukemic samples. The work by Lacombe et al. very nicely exemplifies semi-supervised iterative analysis of cytometry data, and how this can be applied to answer a real-life research question. Additionally, their framework requires only little programming skills and is consequently accessible to most of the community, enabling researchers without programming skills to conduct the whole analysis themselves.

Other groups have suggested similar procedures for analysis of both flow and mass cytometry data (1,4,8,10), and both commercial and academic solutions (e.g. Cytosplore, Astrolabe, and Cytobank) to facilitate the process to various degrees do exist. However, for most of the free-to-use academic bioinformatics tools, smooth iterations of analysis and interpretation are limited by the user friendliness (many of the most popular algorithms are only command line executable) and run times of the analysis tools. FlowSOM is, as mentioned, a very fast clustering algorithm, which allows for the clustering of a million cells with 15 channels in ~1 minute, when using a high-performance computing dedicated CPU with 128 GB memory. However, other popular clustering algorithms including Phenograph (11), and X-shift (12) have much longer run times and higher memory requirements, with Phenograph taking ~50 minutes for a million cells with 15 channels and X-shift taking ~4 hours for just 250,000 cells. As can be seen in Figure 2A, the run time for X-shift scales very poorly prohibiting implementation in the iterative approaches. Both X-shift and Phenograph also face memory issues with higher cell counts, making it infeasible to run the algorithms on large datasets on a personal computer. Overall, the majority of cytometry clustering algorithms (with FlowSOM being the notable exception) are computationally demanding and time-consuming to run.

When considering run times of computational analyses in semi-supervised frameworks, it is important to also consider the additional frequently used analysis tools such as those for visualization of the data, for example, dimensionality reduction, density plots, heatmaps, etc. Dimensionality reduction is commonly used to examine the global structure of the data. In this category of algorithms, PCA is extremely fast even for millions of cells, but because of the complex structures of cytometry data, t-SNE and UMAP are much more commonly used to visualize clusters (13). A drawback of these methods is their run times, with UMAP requiring 30 minutes and t-SNE requiring more than 3 hours to process a million cells with 15 channels (Figure 2B). Generally, most visualizations of the data can take time and, due to the large size of cytometry datasets, also be quite computationally demanding and in most cases require at least some programming skills.

With all of this being said, clustering-based analysis of cytometry data has offered many new insights into the mechanisms of health and disease in the past decade, and efforts in developing better and more efficient computational analysis tools, as well as frameworks for interpretation, continue to enhance the knowledge output from immunophenotyping data. While a fast, unsupervised, unbiased approach to analyzing cytometry data has yet to see the light of day, one thing is certain: these are not only challenging, but exciting times for cytometry.

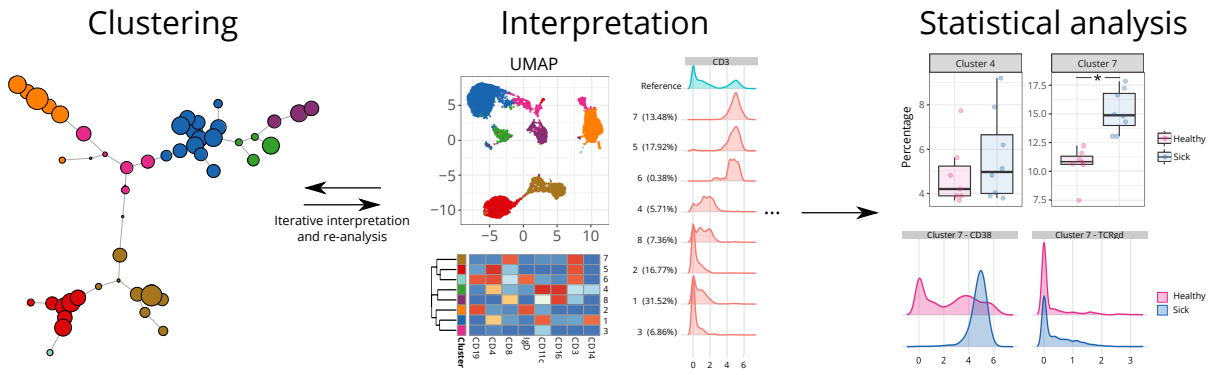

Figure 1. Example of a semisupervised workflow for the analysis of single-cell cytometry data. First, a clustering of the cells is carried out using a selected set of markers. This is followed by in-depth analysis and interpretation of the clustering, typically using dimensionality reduction plots, heatmaps, and density plots. These visualizations facilitate the labeling of each cluster but may also prompt the use of alternative clustering approaches. Once the clustering and labeling are accepted, it is possible to carry out statistical analysis of the results, for example, comparing different sample groups or time points.

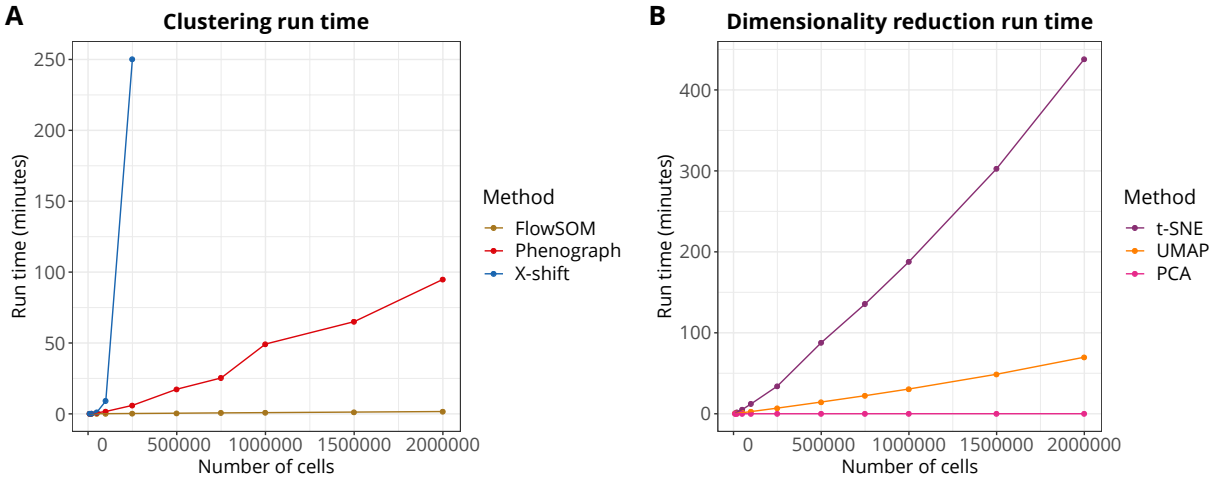

Figure 2. Run times for some of the most popular clustering (A) and dimensionality reduction (B) algorithms on a mass cytometry data set with 15 lineage markers used. All run time analyses were performed on a server with 28 dedicated cores and 128 GB memory—even when the tool in question did not support multithreading (only X-shift and the UMAP R package uwot supported this). FlowSOM (v. 1.16.0) was run using the FlowSOM R package from bioconductor including the meta-clustering step, Phenograph was run using the implementation in the cytofkit (v. 1.12.0) R package, and X-shift was run as a standalone from the 26-Apr-2018 VorteX release with a memory limit of 64 GB. PCA was run using the base R implementation, and t-SNE was run using the Rtsne (v. 0.15) package without an initial PCA step. UMAP was run using the uwot (v. 0.1.3) R package using 15 neighbors, an effective minimum distance of 0.2, and 28 threads. R v. 3.6.1 was used throughout the analysis and timing was measured out using R's built-in time. All tools were run three times and the means are reported.

LITERATURE CITED

1. Olsen LR, Pedersen CB, Leipold MD, Maecker HT. Getting the Most from Your High-Dimensional Cytometry Data. Immunity 2019;50:535–536.

2. Weber LM, Robinson MD. Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. Cytometry A 2016;89:1084–1096.

3. Van Gassen S, Callebaut B, Van Helden MJ, Lambrecht BN, Demeester P, Dhaene T, Saeys Y. FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. Cytometry A 2015;87:636–645.

4. Nowicka M, Krieg C, Crowell HL, Weber LM, Hartmann FJ, Guglietta S, Becher B, Levesque MP, Robinson MD. CyTOF workflow: differential discovery in high-throughput high-dimensional cytometry datasets. [version 3; peer review: 2 approved]. F1000Res. 2017;6:748.

5. Olsen LR, Leipold MD, Pedersen CB, Maecker HT. The anatomy of single cell mass cytometry data. Cytometry A 2019;95:156–172.

6. Seiler C, Kronstad LM, Simpson LJ, Gars ML, Vendrame E, Blish CA, Holmes S. Uncertainty Quantification in Multivariate Mixed Models for Mass Cytometry Data. BioRxiv 1903.07976 [Preprint] 2019.

7. Melchiotti R, Gracio F, Kordasti S, Todd AK, de Rinaldis E. Cluster stability in the analysis of mass cytometry data. Cytometry A 2017;91:73–84.

8. Qiu P. Toward deterministic and semiautomated SPADE analysis. Cytometry A 2017;91:281–289.

9. Good Z, Sarno J, Jager A, Samusik N, Aghaeepour N, Simonds EF, White L, Lacayo NJ, Fantl WJ, Fazio G, Gaipa G, Biondi A, Tibshirani R, Bendall SC, Nolan GP, Davis KL. Single-cell developmental classification of B cell precursor acute lymphoblastic leukemia at diagnosis reveals predictors of relapse. Nat. Med. 2018;24:474–483.

10. Lee AJ, Chang I, Burel JG, Lindestam Arlehamn CS, Mandava A, Weiskopf D, Peters B, Sette A, Scheuermann RH, Qian Y. DAFi: A directed recursive data filtering and clustering approach for improving and interpreting data clustering identification of cell populations from polychromatic flow cytometry data. Cytometry A 2018;93:597–610.

11. Levine JH, Simonds EF, Bendall SC, Davis KL, Amir ED, Tadmor MD, Litvin O, Fienberg HG, Jager A, Zunder ER, Finck R, Gedman AL, Radtke I, Downing JR, Pe'er D, Nolan GP. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. Cell 2015;162:184–197.

12. Samusik N, Good Z, Spitzer MH, Davis KL, Nolan GP. Automated mapping of phenotype space with single-cell data. Nat. Methods 2016;13:493–496.

13. Becht E, McInnes L, Healy J, Dutertre C-A, Kwok IWH, Ng LG, Ginhoux F, Newell EW. Dimensionality reduction for visualizing single-cell data using UMAP. Nat. Biotechnol. 2018.