Automated Analysis of Flow Cytometry Data: A Systematic Review of Recent Methods

Taher Ahmed Ghaleb, Mawal Ali Mohammed and Emad Ramadan
Information and Computer Science Department
King Fahd University of Petroleum and Minerals
Dhahran 31261, Saudi Arabia
Emails: {g201106210,g201102570,eramadan}@kfupm.edu.sa

***Abstract*—Flow cytometry (FCM) is a very well-known method that is broadly used in clinical and research laboratories. Both clinical and research laboratories have been the target domains of FCM applications. The key research question in this particular field is "how to effectively automate FCM data analysis?". To answer this question, this paper systematically reviews current advances in the automation of FCM data analysis. All recent techniques have been studied in a way readers can recognize current trends, challenges, limitations and future directions. For future research, we have identified three main venues. First, the identification of the number of clusters prior to starting cell population identification is still a challenging process. Second, automating the process of cluster labeling still requires more improvement to be fully automated. Last, benchmark datasets are essential in order for researchers to be able to comparatively evaluate different techniques of FCM data analysis under fixed conditions. We end up this paper with a discussion about how flow cytometry data analysis techniques and datasets are correlated with open source technology.**

Keywords*—*Automated Gating; Clustering; Data analysis; Flow Cytometry (FCM); Multidimensional data; Open Source Software

I. Introduction

Flow cytometry (FCM) is a very well-known method that is largely applied in research and clinical laboratories to identify cell populations and their fluorescence, density, and morphological characteristics. Its usage has not been consistent as it has a variety of applications. For clinical purposes, it is used in transplantation, genetics, sex pre-selection, and diagnosis of lymphoma or leukemia patients. In research, it has extensively been used for detecting DNA damage and studying the structure and function of cells. Due to the rapid growth of FCM data volume and its complexity, the demand for reliable, fast, efficient, and accurate statistical methods has significantly increased.

The analysis of flow cytometric data has varied from one approach to another, from various perspectives. The major distinction between existing approaches is generally concerning the manipulation, analysis, visualization, or interpretation of flow cytometric data. From a different aspect, several methods concentrated their work on the automated gating of FCM data, while others attempted to enhance the obtained results by, for example, detecting and removing outliers, supporting multidimensional data, etc.

978-1-5090-4580-8/16/$31.00

Based on a survey published in 2009 by Bashashati *et al.* [1], we noticed that the literature is full of techniques that endeavored to improve the automation of FCM analysis. Another observation was related to the same study, which was not fully comprehensive as some critical state-of-the-art techniques were skipped. This was due to the use of one source of research indexing, which is Google Scholar. Therefore, this motivated us to go through the most recent advances in this context in order to 1) systematically review current approaches, 2) identify their limitations, and 3) recommend a set of future directions concerning the enhancement and automation of the entire process.

We organize the rest of this paper as follows. In Section II, we discuss the problem and limitation of related work and how it motivated us to come up with this work. Section III presents the methodology used to collect, review, and analyze all work related to the automated analysis of flow cytometry. Section IV discusses and analyzes the evaluation results of the reviewed techniques, which were based on well-defined criteria. We demonstrate in Section V the correlation between our study and open source technology. Finally, the paper concludes in Section VI suggests the possible future work.

II. Motivation

A previous survey [1] was conducted to encapsulate the analysis methods of the flow cytometry data. That survey provided an overview about some of the works that are conducted in this area. However, that work suffers serious issues that make us question the obtained results of that survey. In this section, we go over the following issues related to that survey:

- The criteria for selecting relevant research papers were somewhat limited and need to be extended. In addition, it adapted a framework that consists of 8 stages. However, you can rarely find a technique that applies all of such stages. Therefore, focusing on one or two core tasks that are central to the FCM analysis such as automated gating is preferable by readers.
- Some of the selected techniques in that survey were a bit irrelevant to the central theme of the survey. Examples of such are the ones in [2] and [3]. This urges to could actually urge readers to question the overall representation of the provided results.

- It is common for survey papers to provide an overview of the current state-of-the-art, the level of maturity, and the research gaps in the target topic of research. That survey lacks such information, which led to defacing the final conclusions.

III. Survey Methodology

It has been known for literature surveys to employ automatic strategies to search for keywords concerned with its main domain (e.g., [4]). However, Brereton *et al.* [5] indicated that some digital libraries, in software engineering for example, do not well support the selection of primary studies and relevant research identification. Therefore, we decided to conduct a manual querying and filtering for selecting articles from venues that are highly relevant to Flow Cytometry.

We started our investigation by applying our search query at a very popular bibliographic database, Scopus. In order to achieve our goal, we have constructed a query that can retrieve the most relevant research articles to our selected topic. This general structure and semantic of our query are shown as follows.

"Look at the title, abstract, and keyword sections of the articles published after or during 2009 for 'Flow Cytometry' term in conjunction with one of the following phrases: 'Automated Gating', 'Automated Analysis', and 'Automated Clustering'"

Although this query has various forms to be implemented in the existing databases, its behavior and projected results remain the same. We demonstrate below and example form of our query that is compatible with Scopus search engine.

```
TITLE-ABS-KEY ( "Flow Cytometry"
             AND
               (  "Automated Gating"
               OR "Automated Analysis"
               OR "Automated Clustering"
               )
              )
              AND PUBYEAR > 2008
```

The query above has resulted in 56 publications displayed at Scopus search results. We went article-by-article through the titles, abstracts, and keywords of the resulting articles and we could, after filtering out the unrelated ones, end up with 25 relevant articles.

After that, we prepared a list of journals/conferences that publish articles related to flow cytometry. Then, we applied the same search query in each journal/conference to get a set of articles that has not already been found in Scopus, as shown in Table I. What urged us to proceed with this strategy is that we could find some relevant work that never appeared in Scopus, even with the change of query parameters. One of the

The next article selection includes a further reading of the papers that have been downloaded from the different sources. We carefully went through the abstract, introduction, and conclusion sections to get an outlook about the content. Once we notice that the paper is relevant, we dig deep into the other sections (i.e., methodology, results, discussion, etc.) to investigate the main characteristics and limitations of each article.

TABLE I: Result of our query at the different flow cytometry-related sources

Journal / Conference	result	filtered in
BMC bioinformatics	8	4
Cytometry Part A	5	1
Cytometry Part B	0	0
BIBE	1	1
Microbial Cell Factories	0	0
Bioinformatics	6	5
Chemometrics and Intelligent Laboratory Systems	8	1

IV. Evaluation

A. Evaluation Criteria

We have identified four major attributes to be used in the survey as evaluation criteria. These criteria are as follow:

1) **Outlier removal:** The outlier is an anomaly from the standard. In the case of cytometry data, it is the data points that are distant from the original cluster. The identification of the outliers and removing them is of great importance. This importance comes from the effect of these noise on the automated gating. These anomalies can cause deviation to the clustering results from the norm. However, the outlier removal process is a general process and it is not special to the FCM data analysis (i.e. there is no specifics that make the FCM data is special so it is treated differently from the other types of data).
2) **Automated gating:** Automated gating is the process of cell population identification [29]. The cell population identification can be considered as the process of the identification of the homogeneous cell groups that participate in a particular function.
   a) *The method of automated gating:* The methods of automated gating are mainly clustering techniques. Clustering is the process of grouping similar objects into the same group. Clustering analysis is not just an algorithm, it is a task to be solved. There are many algorithms that have been developed to solve this task. Each one of these algorithms has its own characteristics and each algorithm is more suitable for certain types of problems.
   b) *Type of techniques:* There are many classification schemes for the clustering techniques. One of these schemes is the classification of the clustering techniques into supervised vs. unsupervised techniques. The major difference between the supervised vs. unsupervised techniques is the labeling. In supervised techniques, the data is labeled while in the unsupervised technique the data is not labeled.
   c) *Multi-dimensionality:* Clustering space is one factor that affects the quality and the performance of clustering. For non-linear problems, the conversion of the

TABLE II: Summary of the comparison results

Ref.	Outlier removal	Method	Automated gating			Cluster Labelling	Interpretation (classification/ comparison of samples)
			Supervised/ Unsupervised	Multi-dimensional	Automated # of clusters		
[6]	—	A cluster analysis algorithm proposed by [7], which uses k-means	U	Y	M	—	[It enables easy comparative visualization and analysis of multiple datasets.]
[8]	—	Quasi-supervised learning algorithm	S	Y	—	—	[FlowCap-I Challenge with DLBCL and HSCT datasts]
[8]	—	Expectation-Maximization algorithm	U	Y	—	—	—
[8]	—	Divisive Binary Clustering	S	Y	Y	—	—
[9]	—	Second Order Polynomial Histogram Estimators (SOPHE)	U	Y	Y	—	[Using FlowJo over the human PBMC dataset and subsets of Lymphocyte Populations in Peripheral Blood.]
[10]	—	GemStone, a probability state model (PSM)	S	—	—	—	[Compared with a traditional gating analysis of GPI-deficient leukocytes performed by a trained expert.]
[11]	—	Statistical clustering	U	—	—	—	—
[12]	—	An automated data normalization algorithm by a combination of k-means clustering and Procrustes analysis, along with a density-based clustering method.	U	Y	M	—	[Applied to semen data samples collected from Dec.2010 to Mar.2011. Then, F-measure statistic was calculated for comparing the technique with SWIFT and OPTICS.]
[13]	—	GemStone probability state model (PSM), an automated data analysis, and QuantiCALC method.	U	2-D	—	—	[GemStone PSM and QuantiCALC were compared over 457 blood samples using PassingBablok regressions and Pearsons coefficient. The results were plotted using BlandAltman.]
[14]	—	X-Cyt by multivariating mixture modeling for partitioning cytometric data via an expectation-maximization (EM) algorithm	U	Y	M - adjustable	M	[Results were compared to proportions defined by an independent expert cytometry analyst with manual gating in FlowJo over a cohort of 236 healthy donor.]
[15]	[FSC-H and FSC-A scattered light parameters were used to filter outliers away]	Principle Component Analysis (PCA)	U	Y	—	E	[Tested on data set of peripheral blood mononuclear cells collected from 23 HIV-infected individuals, which were stimulated with overlapping HIV Gag-p55 and CMV-pp65 peptides or medium alone. Comparison results were produced as a Z-score matrix and a heat map.]
[15]	—	k-means cluster analysis	U	—	M	—	—
[15]	—	Z-score method for comparing proportions of cells	—	—	—	—	—
[16]	[Remove excessive boundary points and debris. Then, reduce skewness if it is heavy]	curvHDR: for automatic and semi-automatic gating by combining the notions of significant high negative curvature regions and highest-density regions	U	Y, but here it is limited to 1D-3D	M	—	[Tested over bivariate and trivariate samples. Compared with the flowClust method of Lo *et al.* [17]]
[18]	[Through MCL]	The SamSPECTRAL method which is composed of faithful sampling, computing a modified similarity matrix, and spectral clustering.	U	Y	Y	In steps: 'unregistered' –>'registering' –>'community p' =>'registered'.	[Compared with FlowMerge and FLAME over the GvHD dataset.]

Continue of Table II: Summary of the comparison results

Ref.	Outlier removal	Method	Automated gating			Cluster Labelling	Interpretation (classification/ comparison of samples)
			Supervised/ Unsupervised	Multi-dimensional	Automated # of clusters		
[19]	[A rule of 95% quantile was used to identify outliers]	flowClust:a model-based clustering approach based on multivariate t mixture models with the Box-Cox transformation.	U	Y	They used the Bayesian Information Criterion (BIC) when the number of clusters is unknown,	—	[Compared to over GvHD data.]
[20]	[By monitoring cell population statistics from gated or ungated flow data conditioned on experiment-level metadata]	flowWorkspace and QUAliFiER. The flowWorkspace package is used to import the gating template from the FlowJo workspace. QUAliFiER involves importing the data, extracting cell population statistics, defining QA tasks, performing outlier calling, and then generating an quality assessment report	—	—	—	—	[A comparison of the Quality Assurance features of flowQ,FlowJo, and QUAliFiER was conducted over a dataset of 3000 FCS files from the Immune Tolerance Network.]
[21]	—	An ontology labeller	—	—	—	E	—
[22]	—	flowBin: based on K-means clustering and probability binning	U	Y	—	E	[Compared with NN merging over AML dataset (Flow Repository:FR-FCM-ZZYA) used in FlowCAP]
[23]	[Outliers were removed according to the gold standard]	flowPeaks: a hybrid approach of K-means followed by a finite mixture model and histogram spatial exploration	U	Y	Y	E	[Compared with Misty Mountain, FLOCK, flowMeans, flowMerge and FLAME over Barcode data, Simulated concave data, GvHD dataset, and Rituximab data.]
[24]	—	A computational approach that automatically reveals all possible cell subsets.	—	Y	M	—	[Applied to clinical data on HIV-infected military personnel since 1985.]
[25]	—	The misty mountain clustering algorithm	U	Y	Y	E	[The two fluorophores, APC-Cy7-A and Pacific Blue-A, in 853,674 U937 cells is used. This technique is compared to Manually gated 2D barcoding, Simulated 5D Gaussians, Simulated 2D non-convex, 3D rituximab, 4D GvHD and Manually gated 4D OP9.]
[26]	[Examining forward scatter characteristics (FSC)-H vs FSC-A dot plots. User involvement is required]	k- means clustering with 7-Attributes	U	Y	M	M	[Compared to the result done at the Methodist Hospital in Houston.]
[27]	[box-cox transformation]	t-Mixture Models	U	Y	Y (BIC method)	Two classification techniques: SVM and RFs	[438 lymphoma patients data.]
[19]	[box-cox transformation]	multi-variate t-Mixture Models	U	Y	Y	—	—
[28]	—	Gaussian-Mixture Model and Weighted Iterative Flow-clustering	U	Y	M	—	[A pair of datasets for which ground truth labels can be applied: one consisted of human peripheral blood cells, and the other consisted of mouse splenocytes.]

data into a higher dimensional space makes it easier to be solved. However, this comes with performance degradation.

d) *Automated number of clusters:* One of the major issues of clustering is the identification of the number of clusters. Some clustering techniques provide automated number of clusters identification while other techniques do not.

3) **Cluster labeling:** Cluster labeling is the process of identification of the similar cell populations among the different FCM data samples. The cluster labeling can be a stand-alone component or as an embedded component within the automated gating process, taking into account that it is not needed in supervised methods. We refer to manual labeling by 'M', while 'E' refers to labeling methods that are embedded in the automated gating.

4) **Performance evaluation:** The performance evaluation of the identified techniques is surveyed. The performance evaluation gives an indication of the quality of the developed techniques. This is a very important factor that differentiates the different techniques.

B. Analysis and Discussion

The application of the evaluation criteria to the different state-of-the-art techniques is summarized in Table II. Here, we analyze the obtained results and shed the light on the main research gaps in the area of FCM data analysis. We divide our discussion on the basis of the main points that need to be investigated and improved in the future for the sake of fully automating the process of FCM.

It can be noticed from the given table that most of the approaches do not employ cleaning and noise removal functionality, which may be due to the fact that the noise removal techniques are general to all types of datasets. On the other hand, outlier removal approaches applied by the techniques are distinct, meaning that no two techniques applied the same approach. This may actually imply that there is no outlier removal approach that can be applied in any technique, as this can depend on various factors, such as the used classifier, dataset, etc. Notice that some papers, such as [8] and [15], have applied different methods with different characteristics, and thus each method is listed separately in the table. flowAI [30] is a recent technique composed of two methods for cleaning FCM data from unwanted behaviors.

While looking at the different analysis techniques, we can see a wide variety of clustering methods has been adopted in the literature. These methods range from statistical to machine learning techniques. The core objective of such methods is the identification of cell population. The majority of such methods are shown to be unsupervised and can support multi-dimensionality. Notice that in some papers, we could not identify whether the classifier is supervised or not due to the lack of technical information about this regards in that papers. The identification of the number of classes, on the other hand, doest seem to be mature enough and still requires more improvements, as it was commonly carried out manually

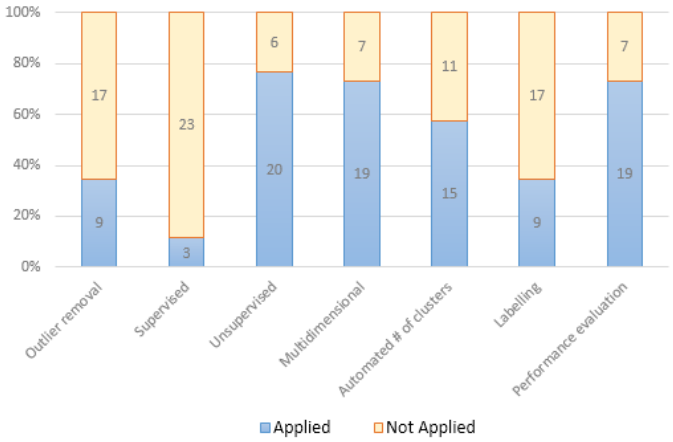

Fig. 1: Ratios of studies applying different data analysis attributes

in most of the techniques. There is also a room for future work in this perspective.

Some of the studies reported the need for user involvement in cluster labeling, which means that the entire process is not fully automated, while embedded labeling has been implemented in some other methods. May studies did not report what kind of cluster labeling approach they applied, which led us to consider its unavailability.

Almost all the studies conducted a performance evaluation, but not in a unified manner. On the other hand, FlowCap, introduced in [31], is a benchmark for implementing various analysis techniques of FCM data against different datasets. It enables future techniques to be fairly evaluated with other techniques without the need to re-implement them individually or convert data from one format to another.

To summarize the results of our study, we have plotted the bar chart in Fig. 1 that demonstrates the ratio of the studies applying each data analysis attribute presented in this paper. Note that we count here the number of methods, not the studies. In other words, techniques that were introduced in a single study are counted here separately. It is obvious that most of the techniques applied unsupervised algorithms with the support of multi-dimensional data and assessed their results in comparison with other methods. On the other hand, we can observe the inapplicability of supervised algorithms in the automated analysis of flow cytometry data due to its restriction in dealing with such kind of data. More than half the techniques employed an approach for determining the number of clusters, while less than 40% performed an outlier removal.

V. Correlation with Open Source Technology

What makes this work related to open source technology is the fact that the techniques and datasets used for flow cytometry analysis are available to researchers in various ways. This section presents the correlation between our study in this paper with open source technology to show how important this relationship is. We discuss the different aspects of deployment of techniques and datasets concerning flow cytometry analysis.

Open source software has recently concurred with the development of techniques in the area of flow cytometry analysis, which strongly contributed to the rapid growth of novel methods in this context (e.g., OpenCyto [32]). Aghaeepour *et al.* [31] aimed to build a framework that owns the capability of running most of the techniques in this manner under an integrated environment and over a unified benchmark of datasets.

A. Techniques Availability

Techniques used for the automated analysis of flow cytometry data are mainly based on classifiers, filters, and clustering techniques that are also applied in other fields of research, such as big data, machine learning, etc [33], [34], [35], [36]. This means that the backend implementation of such techniques is algorithmic methods, which implies that they could be written and implemented in any programming language as required. The description and documentation of such algorithms were published in various scientific research articles, which makes them reproducible by others.

On the other hand, most of these techniques are deployed as freely available packages and libraries in different programming environments, such as Java, R, Python, MATLAB, etc. Some of such libraries are open source under the GNU General Public License, while other are closed source but can be customized based on researchers' needs. Being open source is indeed helpful as researchers can extend or alter the functionality of the techniques to achieve their goals. In [31], a comprehensive list of analysis techniques of flow cytometry data is presented along with brief descriptions of their functionality and availability. A recent trend involved having online web-based analysis techniques [37].

B. Datasets Availability

FCM data is normally considered to be complex and of large sizes containing gene arrays of human or animal cell populations [31] (e.g., GvHD, DLBCL, HSCT, WNV, and ND), but having heterogeneous versions of them makes it possible to apply different techniques over them easily. Each technique requires data to be of a certain format in order to accept it as valid input. After the collection of FCM data, it is stored in files of different formats depending on the collection method applied. After that, flow cytometry datasets are created and deployed publicly using open-source and customizable representations. Such representations allow users to freely and thoroughly access, parse, filter, diagnose, and process flow cytometry data. Some techniques may be interested in specific features of the data while ignoring other unimportant ones. In addition, it is sometimes essential to normalize data to improve the cohesion and integrity of data and to reduce redundancy. All such operations cannot be accomplished if data was provided in fixed or closed-source structure, which expresses the usefulness of open source philosophy.

Flow Cytometry Standard (FCS) is the standard representation of data generated in laboratories, but there exist many other additional formats. Therefore, there exist different techniques that can convert data from one format to another. Usually, such techniques are developed to be capable of dealing with different standard and well-known formats. In special cases, some analysis techniques implicitly employ such a conversion mechanism, while some others require users to do it manually. For example, FCSTrans [38] is an open source tool for converting FCM data files and transforming data from/to different formats. To conclude, having FCM data stored in open representations makes it more usable. However, having a unified representation (or at least conversion mechanism) would increase that usability in addition to improved efficiency where conversion operations could be left out.

VI. Conclusion

We conducted this survey to evaluate the state-of-the-art in the area of flow cytometry data analysis and to identify the research gaps in this area as well. This work is considered as a completion of a previous study that surveyed the state-of-the-art techniques till the mid of 2009. In other words, we conducted our survey to review papers from 2009 (excluding papers that were previously included in the aforementioned survey) until present. Different sources containing peer-reviewed articles have been used in order perfectly accomplish this work. All results obtained have been summarized in a comprehensive table to allow better recognition of the similarities and differences between techniques. We then discussed the correlation of flow cytometry data analysis techniques and datasets with open source technology.

Although there has been much work in this area, there is still a room for further improvement. Expressively, in this survey, we have identified three gaps to be addressed in future work: (1) In the gating process, specifying the number of clusters prior to clustering the cytometry data is still a challenging task, which requires being enhanced by automating the process of identifying their number. (2) We found that the process of labeling clusters is also still carried out manually in various studies, especially the ones employ unsupervised techniques for the automated gating. Again, automating such a process is the desired enhancement option we recommend for future work. (3) Having a unified benchmark dataset in a popular format is very important so that state of the art approaches can be fairly evaluated and compared in terms of performance, accuracy, and effectiveness.

References

[1] A. Bashashati and R. R. Brinkman, "A survey of flow cytometry data analysis methods," *Advances in bioinformatics*, vol. 2009, 2009.

[2] M. A. Suni, H. S. Dunn, P. L. Orr, R. De Laat, E. Sinclair, S. A. Ghanekar, B. M. Bredt, J. F. Dunne, V. C. Maino, and H. T. Maecker, "Performance of plate-based cytokine flow cytometry with automated data analysis," *BMC immunology*, vol. 4, no. 1, p. 1, 2003.

[3] M. Roederer, A. Treister, W. Moore, and L. A. Herzenberg, "Probability binning comparison: a metric for quantitating univariate distribution differences," *Cytometry*, vol. 45, no. 1, pp. 37–46, 2001.

[4] S. Beecham, N. Baddoo, T. Hall, H. Robinson, and H. Sharp, "Motivation in software engineering: A systematic literature review," *Information and software technology*, vol. 50, no. 9, pp. 860–878, 2008.

[5] P. Brereton, B. A. Kitchenham, D. Budgen, M. Turner, and M. Khalil, "Lessons from applying the systematic literature review process within the software engineering domain," *Journal of systems and software*, vol. 80, no. 4, pp. 571–583, 2007.

[6] S. M. Dabdoub, W. C. Ray, and S. S. Justice, "Find: A new software tool and development platform for enhanced multicolor flow analysis," *BMC bioinformatics*, vol. 12, no. 1, p. 145, 2011.

[7] T. C. Bakker Schut, B. G. Grooth, and J. Greve, "Cluster analysis of flow cytometric list mode data on a personal computer," *Cytometry Part A*, vol. 14, no. 6, pp. 649–659, 1993.

[8] B. E. Kokturk and B. Karacali, "Model-free expectation maximization for divisive hierarchical clustering of multicolor flow cytometry data," in *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*. IEEE, 2014, pp. 267–272.

[9] J. Zaunders, J. Jing, M. Leipold, H. Maecker, A. D. Kelleher, and I. Koch, "Computationally efficient multidimensional analysis of complex flow cytometry data using second order polynomial histograms," *Cytometry Part A*, 2015.

[10] D. T. Miller, B. C. Hunsberger, and C. B. Bagwell, "Automated analysis of gpi-deficient leukocyte flow cytometric data using gemstone," *Cytometry Part B: Clinical Cytometry*, vol. 82, no. 5, pp. 319–324, 2012.

[11] F. Ribalet, D. M. Schruth, and E. V. Armbrust, "flowphyto: enabling automated analysis of microscopic algae from continuous flow cytometric data," *Bioinformatics*, vol. 27, no. 5, pp. 732–733, 2011.

[12] H. Babamoradi, J. M. Amigo, F. van den Berg, M. R. Petersen, N. Satake, and G. Boe-Hansen, "Quality assessment of boar semen by multivariate analysis of flow cytometric data," *Chemometrics and Intelligent Laboratory Systems*, vol. 142, pp. 219–230, 2015.

[13] L. Wong, B. L. Hill, B. C. Hunsberger, C. B. Bagwell, A. D. Curtis, and B. H. Davis, "Automated analysis of flow cytometric data for measuring neutrophil cd64 expression using a multi-instrument compatible probability state model," *Cytometry Part B: Clinical Cytometry*, 2015.

[14] X. Hu, H. Kim, P. J. Brennan, B. Han, C. M. Baecher-Allan, P. L. De Jager, M. B. Brenner, and S. Raychaudhuri, "Application of user-guided automated cytometric data analysis to large-scale immunoprofiling of invariant natural killer t cells," *Proceedings of the National Academy of Sciences*, vol. 110, no. 47, pp. 19 030–19 035, 2013.

[15] J. Frederiksen, M. Buggert, A. C. Karlsson, and O. Lund, "Netfcm: A semi-automated web-based method for flow cytometry data analysis," *Cytometry Part A*, vol. 85, no. 11, pp. 969–977, 2014.

[16] U. Naumann, G. Luta, and M. P. Wand, "The curvhdr method for gating flow cytometry samples," *BMC bioinformatics*, vol. 11, no. 1, p. 44, 2010.

[17] K. Lo, R. R. Brinkman, and R. Gottardo, "Automated gating of flow cytometry data via robust model-based clustering," *Cytometry Part A*, vol. 73, no. 4, pp. 321–332, 2008.

[18] H. Zare, P. Shooshtari, A. Gupta, and R. R. Brinkman, "Data reduction for spectral clustering to analyze high throughput flow cytometry data," *BMC bioinformatics*, vol. 11, no. 1, p. 403, 2010.

[19] K. Lo, F. Hahne, R. R. Brinkman, and R. Gottardo, "flowclust: a bioconductor package for automated gating of flow cytometry data," *Bmc Bioinformatics*, vol. 10, no. 1, p. 145, 2009.

[20] G. Finak, W. Jiang, J. Pardo, A. Asare, and R. Gottardo, "Qualifier: An automated pipeline for quality assessment of gated flow cytometry data," *BMC bioinformatics*, vol. 13, no. 1, p. 252, 2012.

[21] M. Courtot, J. Meskas, A. D. Diehl, R. Droumeva, R. Gottardo, A. Jalali, M. J. Taghiyar, H. T. Maecker, J. P. McCoy, A. Ruttenberg *et al.*, "flowcl: ontology-based cell population labelling in flow cytometry," *Bioinformatics*, vol. 31, no. 8, pp. 1337–1339, 2015.

[22] K. ONeill, N. Aghaeepour, J. Parker, D. Hogge, A. Karsan, B. Dalal, and R. R. Brinkman, "Deep profiling of multitube flow cytometry data," *Bioinformatics*, p. btv008, 2015.

[23] Y. Ge and S. C. Sealfon, "flowpeaks: a fast unsupervised clustering for flow cytometry data via k-means and density peak finding," *Bioinformatics*, vol. 28, no. 15, pp. 2052–2058, 2012.

[24] N. Aghaeepour, P. K. Chattopadhyay, A. Ganesan, K. O'Neill, H. Zare, A. Jalali, H. H. Hoos, M. Roederer, and R. R. Brinkman, "Early immunologic correlates of hiv protection can be identified from computational analysis of complex multivariate t-cell flow cytometry assays," *Bioinformatics*, vol. 28, no. 7, pp. 1009–1016, 2012.

[25] I. P. Sugár and S. C. Sealfon, "Misty mountain clustering: application to fast unsupervised flow cytometry gating," *BMC bioinformatics*, vol. 11, no. 1, p. 502, 2010.

[26] M.-C. Shih, S.-H. S. Huang, and C.-C. J. Chang, "A multidimensional flow cytometry data classification," in *Bioinformatics and BioEngineering, 2009. BIBE'09. Ninth IEEE International Conference on*. IEEE, 2009, pp. 356–359.

[27] A. Bashashati, K. Lo, R. Gottardo, R. D. Gascoyne, A. Weng, and R. Brinkman, "A pipeline for automated analysis of flow cytometry data: preliminary results on lymphoma sub-type diagnosis," in *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*. IEEE, 2009, pp. 4945–4948.

[28] I. Naim, S. Datta, J. Rebhahn, J. S. Cavenaugh, T. R. Mosmann, and G. Sharma, "Swiftscalable clustering for automated identification of rare cell populations in large, high-dimensional flow cytometry datasets, part 1: Algorithm design," *Cytometry Part A*, vol. 85, no. 5, pp. 408–421, 2014.

[29] F. Mair, F. J. Hartmann, D. Mrdjen, V. Tosevski, C. Krieg, and B. Becher, "The end of gating? an introduction to automated analysis of high dimensional cytometry data," *European journal of immunology*, vol. 46, no. 1, pp. 34–43, 2016.

[30] G. Monaco, H. Chen, M. Poidinger, J. Chen, J. P. de Magalhães, and A. Larbi, "flowai: automatic and interactive anomaly discerning tools for flow cytometry data," *Bioinformatics*, p. btw191, 2016.

[31] N. Aghaeepour, G. Finak, H. Hoos, T. R. Mosmann, R. Brinkman, R. Gottardo, R. H. Scheuermann, F. Consortium, D. Consortium *et al.*, "Critical assessment of automated flow cytometry data analysis techniques," *Nature methods*, vol. 10, no. 3, pp. 228–238, 2013.

[32] G. Finak, J. Frelinger, W. Jiang, E. W. Newell, J. Ramey, M. M. Davis, S. A. Kalams, S. C. De Rosa, and R. Gottardo, "Opencyto: an open source infrastructure for scalable, robust, reproducible, and automated, end-to-end flow cytometry data analysis," *PLoS Comput Biol*, vol. 10, no. 8, p. e1003806, 2014.

[33] P. Berkhin, "A survey of clustering data mining techniques," in *Grouping multidimensional data*. Springer, 2006, pp. 25–71.

[34] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Transactions on neural networks*, vol. 16, no. 3, pp. 645–678, 2005.

[35] F. Pereira, T. Mitchell, and M. Botvinick, "Machine learning classifiers and fmri: a tutorial overview," *Neuroimage*, vol. 45, no. 1, pp. S199–S209, 2009.

[36] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," 2007.

[37] Y. Qian, H. Kim, S. Purawat, J. Wang, R. Stanton, A. Lee, W. Xu, I. Altintas, R. Sinkovits, and R. H. Scheuermann, "Flowgate: towards extensible and scalable web-based flow cytometry data analysis," in *Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure*. ACM, 2015, p. 5.

[38] Y. Qian, Y. Liu, J. Campbell, E. Thomson, Y. M. Kong, and R. H. Scheuermann, "Fcstrans: An open source software system for fcs file conversion and data transformation," *Cytometry Part A*, vol. 81, no. 5, pp. 353–356, 2012.