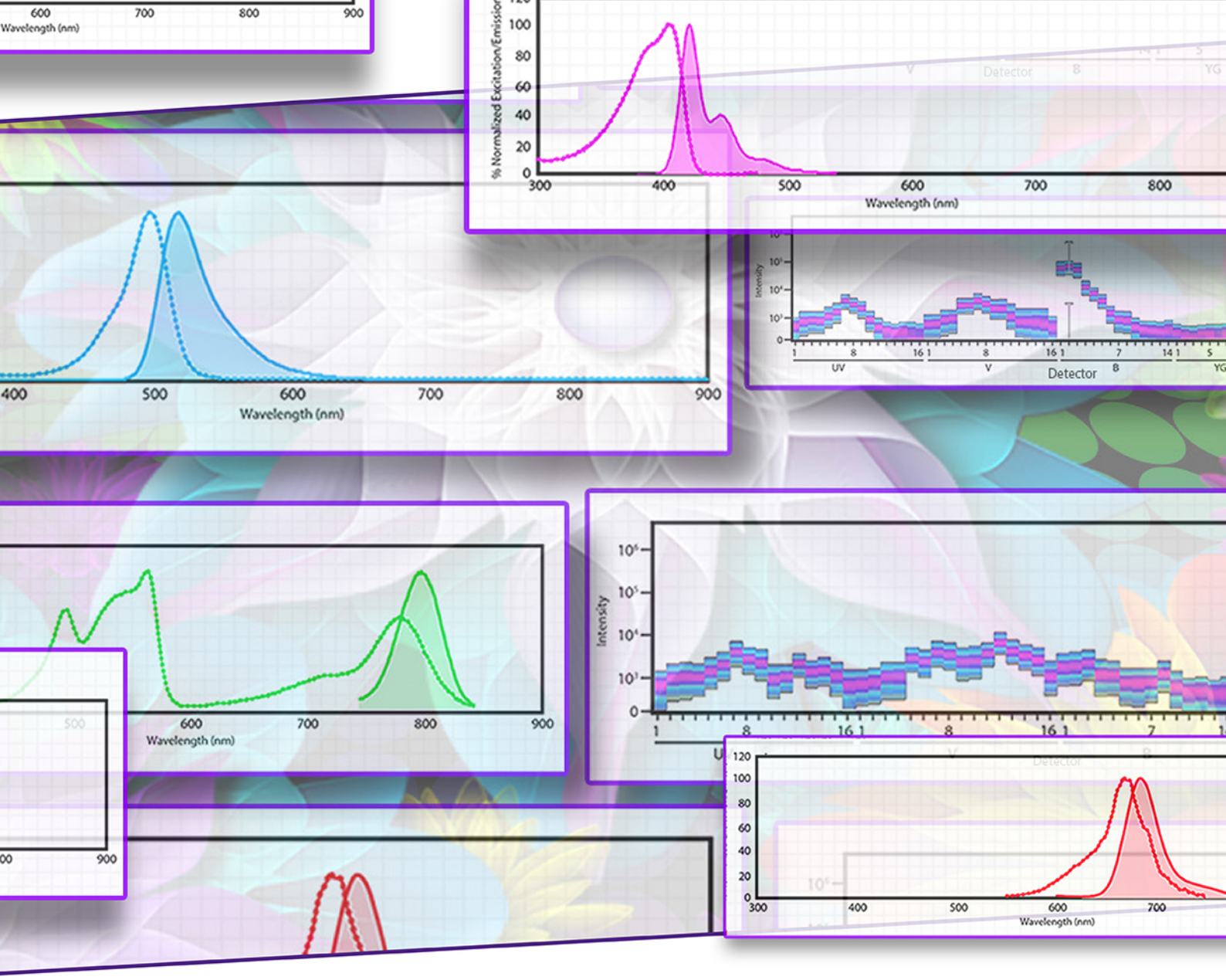
Current trends in flow cytometry automated data analysis software



Flow Cytometry Fluorophores Poster

As flow cytometers have advanced, new fluorophores have been developed to outfit them. Our new, updated Fluorophores for Flow Cytometry Poster provides details on options for both conventional and spectral unmixing cytometers. The poster provides full excitation and emission spectra as well as a brightness index for fluorophores offered by BioLegend.

Learn about different fluorophore families, including:

Spark Dyes: a family of small, synthetic fluorophores that fill spectral spaces between existing fluorophores.

Fire Dyes: tandem fluorophores that expand into spectral spaces previously unused in conventional cytometry.

Brilliant Violet™ Dyes: intensely bright polymers that maximize the utility of the violet laser.

Request the poster: biolegend.com/en-us/fluorophore-poster

World-Class Quality | Superior Customer Support | Outstanding Value

BioLegend products are manufactured in an ISO 13485:2016-certified facility to ensure the highest quality standards.

BioLegend®
biolegend.com

Received: 26 November 2020 | Revised: 21 January 2021 | Accepted: 28 January 2021

DOI: 10.1002/cyto.a.24320

REVIEW ARTICLE

ISAC CYTOMETRY Journal of Quantitative Cell Science PART A

Current trends in flow cytometry automated data analysis software

Melissa Cheung[1] | Jonathan J. Campbell[2] | Liam Whitby[3] | Robert J. Thomas[1] | Julian Braybrook[2] | Jon Petzing[1]

[1]Centre for Biological Engineering, Loughborough University, Loughborough, Leicestershire, United Kingdom

[2]National Measurement Laboratory, LGC, Teddington, United Kingdom

[3]UK NEQAS for Leucocyte Immunophenotyping, Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, United Kingdom

Correspondence
Melissa Cheung, Centre for Biological Engineering, Loughborough University, Loughborough, Leicestershire, UK.
Email: m.cheung@lboro.ac.uk

Funding information
EPSRC/MRC Doctoral Training Centre for Regenerative Medicine at Loughborough University, Grant/Award Number: EP/L105072/1; LGC; UK NEQAS

Abstract

Automated flow cytometry (FC) data analysis tools for cell population identification and characterization are increasingly being used in academic, biotechnology, pharmaceutical, and clinical laboratories. The development of these computational methods is designed to overcome reproducibility and process bottleneck issues in manual gating, however, the take-up of these tools remains (anecdotally) low. Here, we performed a comprehensive literature survey of state-of-the-art computational tools typically published by research, clinical, and biomanufacturing laboratories for automated FC data analysis and identified popular tools based on literature citation counts. Dimensionality reduction methods ranked highly, such as generic t-distributed stochastic neighbor embedding (t-SNE) and its initial Matlab-based implementation for cytometry data viSNE. Software with graphical user interfaces also ranked highly, including PhenoGraph, SPADE1, FlowSOM, and Citrus, with unsupervised learning methods outnumbering supervised learning methods, and algorithm type popularity spread across K-Means, hierarchical, density-based, model-based, and other classes of clustering algorithms. Additionally, to illustrate the actual use typically within clinical spaces alongside frequent citations, a survey issued by UK NEQAS Leucocyte Immunophenotyping to identify software usage trends among clinical laboratories was completed. The survey revealed 53% of laboratories have not yet taken up automated cell population identification methods, though among those that have, Infinicyt software is the most frequently identified. Survey respondents considered data output quality to be the most important factor when using automated FC data analysis software, followed by software speed and level of technical support. This review found differences in software usage between biomedical institutions, with tools for discovery, data exploration, and visualization more popular in academia, whereas automated tools for specialized targeted analysis that apply supervised learning methods were more used in clinical settings.

KEYWORDS
automation, cell therapy, data analysis, flow cytometry, gating, software

1 | INTRODUCTION

Flow cytometry (FC) is an important analytical technique for single-cell population identification and characterization. It is widely used within biotechnology, pharmaceutical and clinical laboratories, and biomanufacturing spaces. Reproducibility and rigor in results are very important, driven by the needs of regulators around the world, however, a major source of variation in FC lies within data analysis [1]. Conventional FC data analysis involves sequential manual selection (gating) of regions of interest typically in two-dimensional scatter or contour plots, viewing different combinations of parameters as axes. The analysis is straightforward with three- to four-color immunofluorescence data but becomes significantly more complex when examining an increasing number of cellular markers, leading to increasing human operator variation and issues of reproducibility [2, 3]. Current state-of-the-art flow cytometers are capable of measuring over 40 parameters, generating challenging complex, and time-consuming multidimensional data sets for manual analysis [4–6].

The past decade has seen a growth in the field of computational FC as researchers become increasingly motivated to solve the process bottlenecks, and reproducibility issues in manual gating, and improve standardization in immunophenotyping [7]. New automated data analysis software packages have emerged, making use of a range of different machine learning and clustering algorithms to replicate or aid manual data analysis tasks such as; data preprocessing, cell population identification and enumeration, feature extraction, and sample classification [8]. Visualization of data processed through algorithmic analyses is an essential aspect of analysis workflows, and is often embedded in the automated analysis itself, therefore making the distinction between pure analysis tools versus visualization tools somewhat blurred. Graphical outputs aid quality control checking and enables understanding and interpretation of the data. Examples of FC visualizations can be: (a) cell populations color-coded according to clustering results displayed on classic biaxial dot plots, (b) grouped populations in nodes arranged in the form of spanning trees, and (c) mapping of high-dimensional data to two-dimensional scatter plots representing data similarities, with color-coded cell clusters.

These data-driven automated algorithms have been demonstrated to improve the quality of flow cytometry data compared with centralized manual analysis, with potential benefits in lower technical variability in certain cell populations, reduced bias, and better efficiency [9]. Given the proliferation of such algorithms, verification methods to ensure correct choice would be recommended. It would be sensible for all users to contextually develop their own robust testing measures for automated analysis. However, this raises subjectivity issues if testing was based on users' own biological knowledge, compounded by the fact that there are no common toolsets to achieve this apart from real-world data sets which do not necessarily have an absolute cell count, and are inflexible compared with the potential of synthetic data.

Typical workflows in computational cytometry can be divided based on tools used for discovery versus targeted analysis, that is, the detection of unknown, novel cell populations compared with known well-defined ones. In both contexts, automation can help to reduce variability in the data analysis process. In discovery mode, automated tools can help uncover cell populations overlooked in sequential manual gating strategies, such as cells gated out in earlier steps. The value of automated tools in discovery mode is especially notable in facilitating the interpretation of high dimensional (>30) data, as the data can be reduced and visualized in two dimensions. These tools assist with the data exploration process, help to give an overview of the structure of the data, identify relationships between variables and offer novel insights. For comparison, in targeted analysis mode, the cell populations of interest are well characterized, the data analysis process follows a standard protocol that is likely to be validated and approved, for example, in clinical flow laboratories carrying out high throughput screening; measurement of clinical trial endpoints for hematological malignancies. The benefit of automated tools here may be in reducing the workload on users by automating the classification of healthy or disease cases, only flagging up uncertain cases for manual interpretation, thereby speeding up the data review process.

As the number of automated software increases, comparison studies have become important to provide guidance for users to determine which software to use for their analysis, and to evaluate the performance of the software. The flow cytometry: critical assessment of population identification methods (FlowCAP) consortium initiated a series of open challenges to objectively evaluate these new computational methods [10, 11]. FlowCAP provided benchmarking data sets to critically assess performance in population identification and sample classification tasks and used the F-measure (the harmonic mean of precision and recall) to rank the algorithms. These rankings helped inform potential users on the quality of automated methods based on different tasks. FlowCAP demonstrated certain automated methods were able to reliably replicate manual gating.

Several other recent comparison studies have evaluated selected unsupervised clustering methods in their abilities to reproduce manual gating, detect rare cell populations and their runtimes. Among those, one study [12] identified FlowSOM [13] as the best performing clustering method along with the fastest runtimes. X-shift [14], PhenoGraph [15], Rclusterpp, and FlowMeans [16] were also mentioned to perform well across six high dimensional data sets. A separate study [17] assessed FLOCK, SWIFT, and ReFlow on their ability to detect low-frequency populations compared with central manual gating. SWIFT was found to outperform the others in terms of the identification of populations <0.1%. This study noted the difficulties in implementing a fully automated workflow without human intervention. In addition, one study [18] evaluated the reproducibility and robustness of results based on the cluster stability using the Jaccard coefficient as the performance metric. PhenoGraph was observed to generate the highest proportion of stable clusters compared with SPADE1 and FLOCK.

Despite these recent benchmarking studies, uptake of automated analysis among academic, biotechnology, pharmaceutical, clinical laboratories, and contract research organizational researchers has been slow and manual gating remains the default method and standard. Manual analysis can be performed on instrument-packaged software

(e.g., Becton Dickinson FACS Diva, BD FACS Canto, Beckman Coulter Navios) or stand-alone FC analysis software (e.g., FlowJo, FCS Express, Kaluza, VenturiOne). The primary reasons for clinical centers not employing automated analysis were recently cited as being a lack of trust/understanding and lack of resources [19]. In this regard, this novel analysis of automated software provision and use presented here is intended for researchers and process operators familiar with FC who do not necessarily have a computational background, who are interested in implementing automated methods into their data analysis workflow and require a better understanding of the opportunities for automated software package selection.

This analysis begins with a comprehensive literature survey to identify the most frequently used tools in the past 10 years in FC automated data analysis software. Popular software are identified based on literature citations, then their common features are outlined to allow the determination of the toolset most relevant to individual need. In addition, automated data analysis software adoption trends from front line clinical laboratories are identified through a survey, and insights are provided on the reasons uptake of certain software is higher than others.

2 | SEARCH STRATEGY

The goal of this research was to understand current trends in automated data analysis software, the characteristics of these software, and identify which software were the most popular (although this is not a measure of most effective software). Software mentioned in recent reviews [10, 12, 17, 18, 20] were included. In addition, the Web of Science (WoS) database was searched using the following keywords; flow cytometry, automated, analysis. Using this search strategy, 89 software were identified from recent reviews and 108 publications were returned from the WoS database, typically output from research, clinical and biomanufacturing facilities. The WoS search strategy was designed to be as comprehensive as possible, although some tools may have been missed due to the fragmented nature of the field, such as FLOW-MAP force-directed graphs [21] and scaffold maps [22]. Use of additional keywords such as "computational" may have highlighted more software, however, in practice, the records retrieved from the database were either too restrictive with the AND Boolean search operator, or excessively broad with the OR search operator. After removing duplicates, the software identified in the search were refined based on the following specifications.

Inclusion criteria:

- Software is detailed in a publication from a peer-reviewed journal.
- Publication type: article.
- Software for flow cytometry or mass cytometry.
- Software for automated cell population identification (gating).
- Software intended for identification of human or mammalian cells.
- Software source code is available, or the program is made accessible by authors.

Exclusion criteria:

- Software lacking publication from a peer-reviewed journal.
- Publication type: conference proceedings, reviews, editorial material, book chapters. This exclusion criteria were applied in order to capture work that applied the data analysis software rather than just citing their use.
- Software unrelated to flow cytometry or mass cytometry technique.
- Software solely for automated data preprocessing, compensation, transformation, or other quality control feature.
- Software unrelated to the identification of human cells (e.g., beads, phytoplankton, bacterial identification) to focus the scope on cell therapy and medical applications.
- Software source code not provided, or program inaccessible.

Certain proprietary software that fell into the exclusion criteria include automated cell identification features in FACS Diva (Becton, Dickinson & Company [BD]), Kaluza (Beckman Coulter), FlowJo (BD), FCS Express (De Novo Software), Gemstone (Verity Software House), and VenturiOne (Applied Cytometry).

The number of software matching the criteria was refined to 51. Once shortlisted, software popularity was ranked according to the number of article citations. The sum total of the number of citations across all 51 software was 2027. Citing articles were refined to those matching "cytometry" as a keyword, included articles, and excluded conference proceedings, reviews, editorial material, and book chapters.

Additional software would have been identified if the search strategy were broadened to include automated single-cell analysis approaches from other technologies (e.g., RNA-sequencing analysis software in genomics, single-cell imaging, single-cell proteomics), and indeed many tools are transferable between different omics domains, however, this was beyond the scope of this work.

3 | GENERAL FINDINGS AND TRENDS

As of the end of 2019, this search strategy has been completed several times on an annual basis and has currently identified 51 automated flow cytometry software (Table S1). The earliest software was released in 2008 and subsequent years saw the number of different software released ranging from 1 to 6 per year, except for 2014 when a peak of 11 software were published (Figure 1A). When considering the country of origin, the USA has led the development with 29 software, followed by Canada with six software. Outside of North America, some European studies have come from The Netherlands, Belgium, France, and Germany (4, 2, 2, and 2 software, respectively). Australia and Singapore have also produced two apiece (Figure 1B).

The environment in which users interact with the software range from basic command line inputs to full graphical user interfaces (GUI). This survey found 41% of software could be accessed with a GUI, compared with 59% without GUIs (Figure 1C). A caveat here is that although most likely to have GUIs, as identified in section 2.0 proprietary computational tools lacking peer-reviewed publications and with

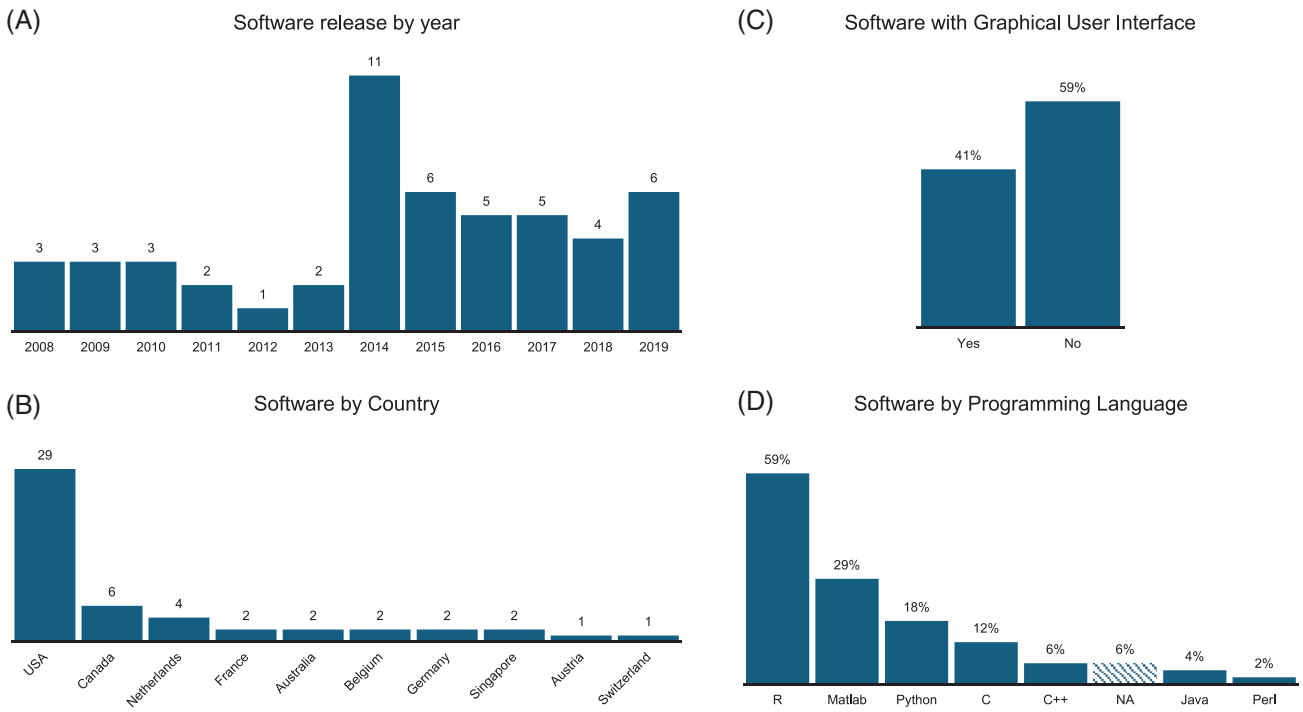


FIGURE 1 General trends in automated data analysis software. (A) Number of software released by year gradually increasing. (B) USA leads the development of software to analyze flow cytometry data. Counts based on first author affiliation in publications. (C) The majority of software are released without graphical user interfaces (GUI). (D) Technical trends. Software are released in multiple programming languages and implementations. R, Matlab and python are the languages most software are available [Color figure can be viewed at wileyonlinelibrary.com]

unavailable source code were excluded from our survey. Many of the tools were available in multiple programming languages, offering FC analysts a choice of integrated development environments. This survey found 59% of the software were available in R, 29% in Matlab and 18% in Python (Figure 1D).

3.1 | Most used software

The findings from the literature survey revealed the top five most cited automated data analysis software based on the search criteria and exclusion criteria were: viSNE, SPADE1, t-SNE, PhenoGraph and FLAME (Table S1). To balance out the effect of earlier software releases accumulating more citations over time, the number of citations were averaged over the number years in publication leading to an adjustment of the highest citation rates; viSNE, PhenoGraph, SPADE1, FlowSOM, and t-SNE (Figure 2A). Changes in individual software citations over time showed viSNE has been the top-cited software for the past three consecutive years (Figure 2B), and a recent rapid increase in FlowSOM citations, moving it from 23rd most cited software in 2017 to 7th highest in 2019. viSNE has a higher citation rate than its origin dimensionality reduction method t-SNE, suggesting many authors consider these separate tools and neglect to cite the original van der Maaten publication [23].

Software that provided a GUI were considerably more cited than those without—command line-based software (Figure 3A). The combined total number of citations for software with GUIs was 1459 compared with 613 for those without GUIs. Command line-based software require computer programming knowledge, which acts as a barrier to many biomedical researchers. Another factor that influences software selection is cost and availability. There are three broad levels of cost in accessing automated flow cytometry data analysis software: free open source software on a free platform, free open source software on a platform requiring a license fee or subscription, and, commercial software on a standalone or paid platform. Currently, access to software are mostly free and open-source, however, some platforms require a paid subscription. Software are available as packages built within the Matlab or R statistical software environments, plugins as part of specialist FC manual data analysis software (such as FlowJo, FCS Express), and applications on web-based platforms such as Cytobank [24]. The same software can be implemented and be available on more than one platform. Cost does not appear to be a deciding factor for users, because the most cited software were accessed through paid platforms (Figure 3B). The levels of usability and software support provided typically increase in line with cost.

4 | SOFTWARE ALGORITHM TYPES

For further insight, the software were separated based on the algorithm type. The algorithms broadly fall into two categories: supervised and unsupervised learning. Thirty-four of 51 software in our survey employed unsupervised learning algorithms, and 17 used supervised learning algorithms (Figure 4) [25–41].

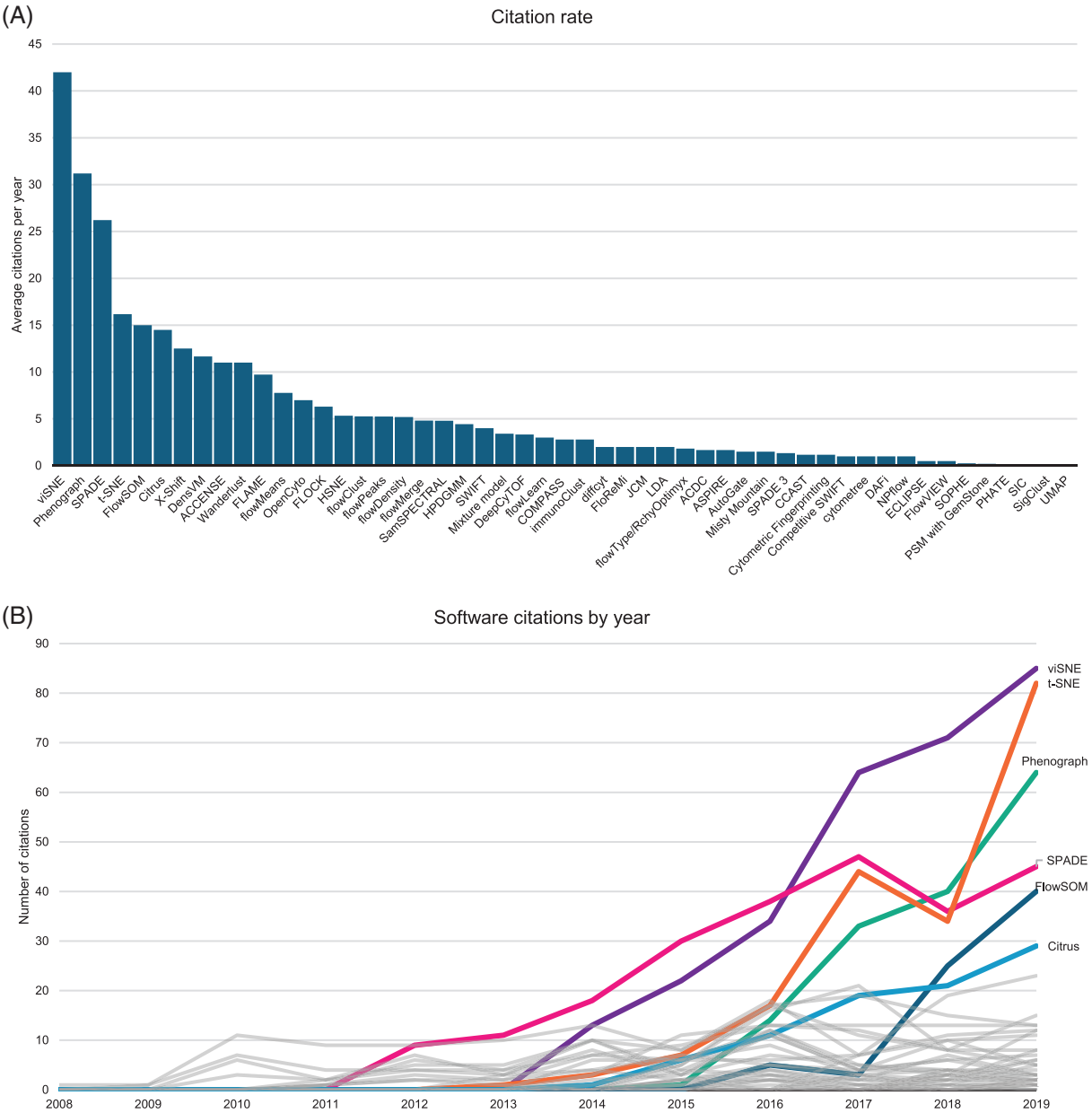


FIGURE 2 (A) viSNE, Phenograph and SPADE are the highest ranked software based on average number of citations per year. (B) Software citation trends by year. viSNE has been the top cited software for the past three consecutive years. t-SNE, Phenograph and SPADE are also highly cited. FlowSOM citations have risen steeply since 2017 compared with other software [Color figure can be viewed at wileyonlinelibrary.com]

4.1 | Supervised learning methods

Supervised learning methods aim to solve classification and regression problems. These algorithms require training data with known outcomes to learn from, in order to build a model to classify new inputs. In practical FC applications, manually annotated cell populations associated with healthy or diseased patients could be used as training data. Cell marker expression features that correlate with the two outcomes would be extracted from the data and then a model built to classify the disease status of new samples.

The limitation of these methods is that the algorithm is only as good as the training data sets available for it to learn from, and it is also possible to overstrain a learning algorithm. Furthermore, there are insufficient publicly available training data sets for all possible scenarios in clinical settings, especially those focused on rare cell identification. The FlowCAP-II sample classification challenge used three real-world patient data sets, half of each data set (training set) was labeled with patient clinical outcomes and the challenge was to correctly label the other half (test set). The comparison study found many algorithms achieved perfect classification accuracy on two data sets (acute myeloid leukemia detection and HIV vaccination antigen stimulation groups), but all performed poorly on a third (HIV exposure on African infants) [10]. Because the current number of supervised learning software in FC data

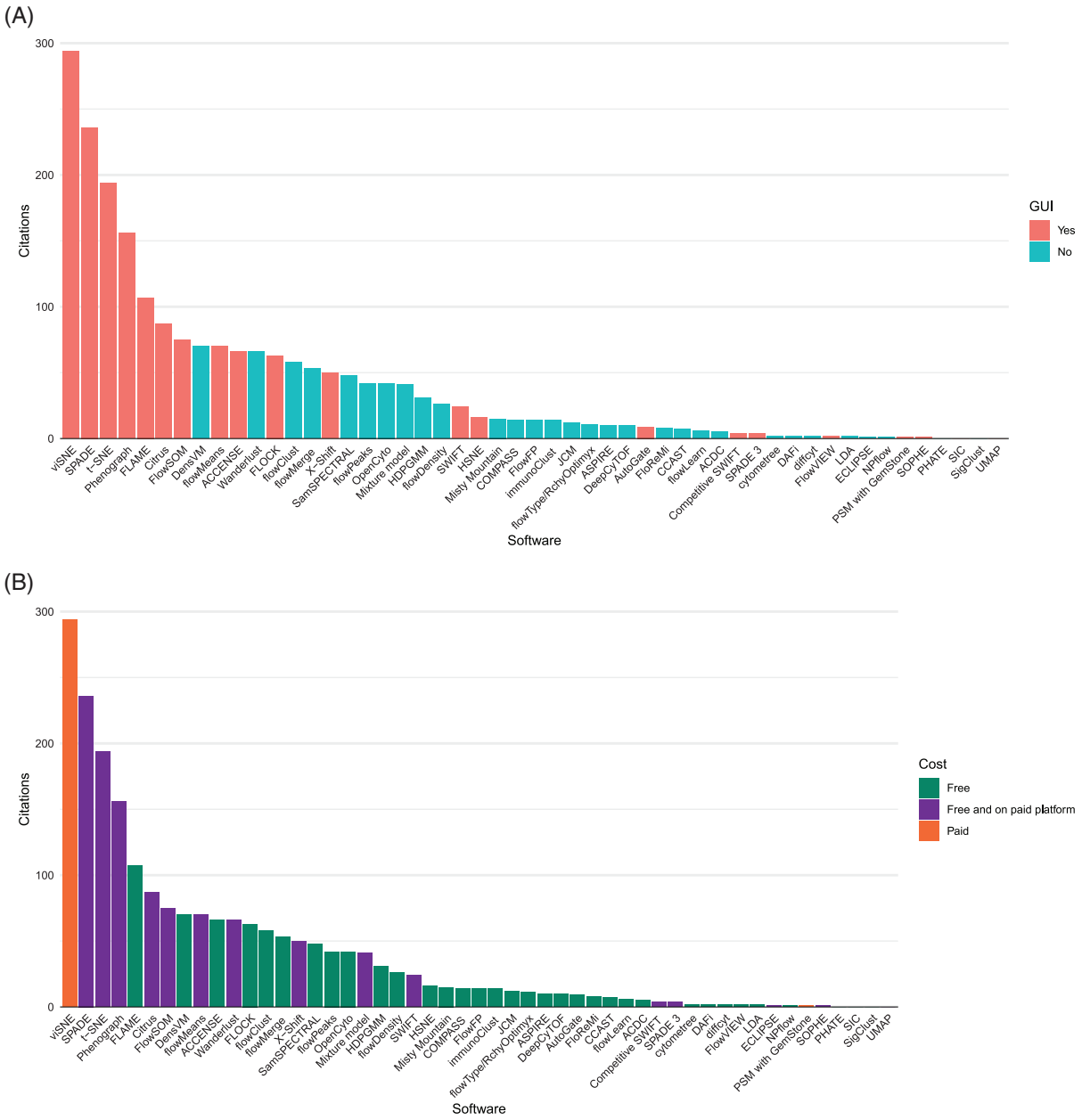


FIGURE 3 Factors for software frequency of citation. (A) Software with graphical user interfaces (GUIs) are more highly cited. (B) Cost is not an important factor for users, with the most cited software requiring fees or access to a paid platform [Color figure can be viewed at wileyonlinelibrary.com]

analysis is low, and there is limited availability of large training data sets, the majority of this analysis concentrates on the significant number of unsupervised methods.

4.2 | Unsupervised learning methods

With unsupervised learning, no training data set is needed, and the goal is to correctly identify and quantify cell populations in FC data. Automated gating of cell subtypes is viewed as a clustering problem. The unsupervised learning software in this survey apply different clustering methods such as hierarchical clustering, partition clustering, model-based clustering, density-based clustering (Figure S1). Dimensionality reduction is also used to simplify multiparameter data sets. Below is a brief overview of the most frequently used clustering algorithms. For a comprehensive survey of clustering algorithms, see Reference [42].

4.2.1 | Hierarchical clustering

Hierarchical clustering has two strategies to group similar datapoints together, agglomerative, and divisive [43]. The agglomerative method follows a bottom-up approach, where neighboring datapoints are merged to form sequentially larger clusters, until only one cluster remains. The divisive method follows a top-down approach, starting with the

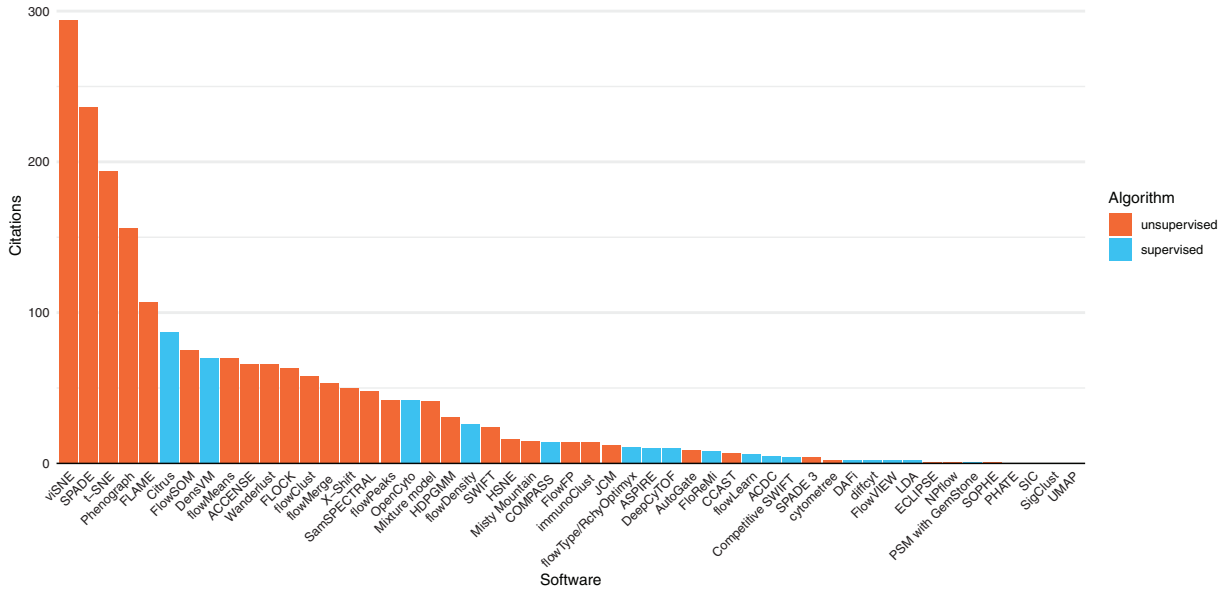


FIGURE 4 Software citations by computational method. Number of citations by machine learning method. Unsupervised learning methods include clustering, dimensionality reduction and do not require training data. Supervised learning methods include classification and regression such as support vector machines, artificial neural networks. They require manually labeled training data to build a model and preform predictions. The most frequently cited flow cytometry software algorithms apply unsupervised learning approaches [Color figure can be viewed at wileyonlinelibrary.com]

whole data set as one cluster and partitioning it to form smaller clusters down to the level of individual datapoints. The target number of clusters is determined by the user. The resulting clustered data can be visualized as a hierarchical tree structure (dendrogram) which resembles phylogenetic trees. Thus, hierarchical clustering appears well suited to classifying data sets with evolutionary observations and may have natural uses for analyzing cell development, maturation, and differentiation data from time course experiments.

The second most frequently cited software in this survey, SPADE1, applies agglomerative hierarchical clustering in its algorithm [44]. A prior density-based down-sampling step is performed to equalize low density populations with high density ones. Down-sampling reduces the time complexity of the hierarchical clustering step, and also increases the prevalence of rare cell types and noise events. The SPADE1 algorithm overcomes the problem of selecting the number of clusters by over-clustering the data set (e.g., instead of three nodes, set 100 nodes). The algorithm builds a minimum spanning tree (MST) from the clustered data, and then relies on expert operator manual analysis to partition the MST to determine correct number of cell populations. An improvement on the SPADE1 algorithm, SPADE3, has been released to remove the stochastic nature of the original agglomerative algorithm by implementing a deterministic K-means clustering algorithm, and to introduce a semiautomated interpretation of the MST [45], thus creating a new software (albeit with the same name) with different mathematical definitions and characteristics, and potentially different data analysis outcomes. In addition to these algorithmic differences between the versions, SPADE3 is primarily implemented in Matlab although stand-alone executable code does exist, SPADE1 and its updated version SPADE2 (better GUI and runtimes) are implemented in R and are available on Cytobank and as a plugin on FlowJo.

4.2.2 | K-means clustering

The K-means clustering method was first published in 1955 and is one of the most popular clustering algorithms used in pattern recognition [46]. K denotes the number of clusters, which is user defined. The K-means algorithm begins with K seed points randomly scattered in the data set acting as cluster centers. Neighboring datapoints are assigned to their nearest seed to form the initial clusters. The center of the clusters, the centroid, is calculated and repositioned. The algorithm repeats the assignment of datapoints to the updated centroid, and then updates the centroid, and so on. Further iterations to update the clustering are performed until cluster membership stabilizes. K-means is an efficient algorithm, with faster run times compared with hierarchical and model-based clustering. However, the drawbacks are its requirement for a predefined number of clusters, its limitation to spherically shaped data and sensitivity to outliers. These are key issues that need to be addressed for correct analysis of FC data, which are usually non-convex shaped and noisy.

The software flowMeans [16] and flowPeaks [47] are based on K-means clustering, and attempt to solve these limitations of K-means clustering on FC data by over-clustering the data then merging nearby clusters to obtain a single population. flowMeans applies a change point detection algorithm to detect the number of clusters, whereas flowPeaks fits a Gaussian finite mixture model to the initial K-means clustered data then generates a density function to search and merge peaks. The results successfully identify nonspherical cluster shapes, however, rare clusters remain difficult to uncover.

4.2.3 | K-medoids clustering

K-medoids clustering, also known as partition around medoids (PAM), is similar to the K-means method, intending to partition the data set

into K clusters, but instead of using centroids (the mean of the datapoints in a cluster) to assign nearby objects, K-medoids uses the representative object of a cluster with minimal average dissimilarity to its assigned objects [43]. K-medoids is less sensitive to outliers than K-means, however, its main disadvantage is the high computational cost for analyzing large data sets. Sampling of the data set is one strategy to reduce runtimes (CLARA) [43]. A modified version of PAM has been proposed for use in a clustering analysis pipeline to identify cell populations [48].

4.2.4 | Density-based clustering

Density-based clustering algorithms such as DBSCAN (density-based spatial clustering of applications with noise) [49] and OPTICS [50] views datapoints in high density regions as clusters, separated by regions of low density. Density-based clustering identifies core points belonging to a cluster as well as noise points. These algorithms are intended to discover clusters of arbitrary shape, such as geographical data. Key requirements are a threshold for the minimum number of points in a neighborhood and an arbitrary distance measure for the density-reachability of a point to a core point. Since the number of clusters is not a required input parameter, this method is useful for FC data analysis where the number of cell subtypes is unknown. Generically, density-based clustering algorithms appear to be a widespread strategy for software developers to identify cell populations, and are used by some software: ACCENSE [51], DensVM [25], Flock [52], flowDensity [26], Misty Mountain [53] and others [54–56], noting that mathematical implementations and algorithms may vary depending upon the data analysis approach.

4.2.5 | Model-based clustering

Model-based clustering assumes the data follows a statistical distribution and models this onto the data set. For example, Gaussian mixture modeling (GMM) views the data as consisting of several Gaussian (normal) distributions and merges the data to the predetermined number of clusters fitting the model. There are numerous mathematical models available, so basic problems arise in selecting an appropriate model and choosing the number of clusters for fitting the model. The optimal model neither underfits nor overfits data, and can be estimated using criteria such as the Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) [57]. This approach to fit each model to the data to find the best fit is computationally expensive.

Model based clustering methods are the most frequent in this survey and may be due to the plethora of statistical models to choose from. These include models based on mixtures of Gaussian distributions, Student's t-distributions, and skew t-distributions. The following examples of software use model-based clustering methods: FLAME [58], flowClust [59], flowMerge [60], SWIFT [61, 62] and others [63–70].

4.2.6 | Spectral clustering

Spectral clustering is based on graph theory where each datapoint represents a node, and the edges are weighted based on a similarity criterion. Clustering is achieved through graph partitioning [71]. Spectral clustering is used by the software SamSPECTRAL [72] which includes a subsampling step to reduce runtimes. Wanderlust also applies a graph-based representation of data in its algorithm [73].

4.2.7 | Self-organizing map

The self-organizing map (SOM) is based on a model of neural network learning [74]. The premise is to construct a grid and map random datapoints one at a time onto each node of the grid. The grid self-organizes so that neighboring nodes have greater similarity, and less similar nodes are moved further away. The next input datapoint is applied to the node that matches best with it. In the end, a large high dimensional data set is reduced to a low dimensional space while retaining the global structure of the original data [75]. The resulting SOM can be clustered further to group similar nodes, using traditional methods such as hierarchical agglomerative clustering and K-means clustering [76]. The FC data analysis software FlowSOM builds a minimal spanning tree from the SOM, followed by a consensus hierarchical clustering step to give the expected number of cell types [13].

4.2.8 | Dimensionality reduction

Dimensionality reduction is not strictly a clustering method. The idea is to take data containing multiple parameters and reduce it to (usually) two dimensions which can be easily interpreted. Principle component analysis (PCA) is an established dimensionality reduction method, however, newer algorithms such as t-stochastic neighborhood embedding (t-SNE) are a significant improvement that preserves (to a limited extent) both the local and global structure of the high-dimensional data, and generates a visual map of the data where similar points are clustered together [23]. Albeit very large data sets (>10^6 events) can cause crowding in the layouts that limit meaningful interpretation of the data, and runtimes are slow [77]. The t-SNE algorithm and its implementation in viSNE successfully visualizes a variety of large real-world data sets and appear well suited to analysis of large multidimensional FC data [78]. This is reflected in their overwhelming popularity in this survey with viSNE and t-SNE ranking first and third respectively in the software citation analysis, and their numbers combined make up 24% (488 out of 2072) of all citations. Dimensionality reduction is increasingly being used as the first step of a data analysis pipeline to extract initial clusters, followed by a clustering step to identify cell populations [79].

The benefits of data visualization and interpretation following dimensionality reduction have encouraged further development of similar data analysis tools that improve scalability, runtimes and are better able to handle large (>10^6) data sets and represent the global

structure. These tools include hierarchical stochastic neighbor embedding (HSNE) [80], PHATE [81] and uniform manifold approximation projection (UMAP) [77].

4.3 | Preprocessing tools

Although excluded from this study, automated preprocessing tools play an important role in FC data analysis because they enable high-quality input data for all the analysis approaches mentioned above. Preprocessing tools used to clean raw data include quality control tools to remove fluorescence anomalies (flowClean, flowAI), perform transformation (flowCore) and normalization (flowStats) [82–85]. Manual gates that exclude doublets, debris and dead cells can be imported from FlowJo into R using flowWorkspace [86], and these manual gates can also be automatically replicated using flowDensity [26].

In summary, the popularity of FC automated data analysis software may depend on the convenience of having a GUI. Currently, unsupervised learning methods receive more citations than supervised methods. Among unsupervised methods, dimensionality reduction algorithms are more popular than other clustering algorithms, because it seems users value the automatic visual output of high-dimensional data presented in an intuitive way that retains local and global structure. Among the other unsupervised methods there was no specific class of algorithm that was more popular than others, although analysis methods that provide novel data visualizations (e.g., SPADE1, Phenograph, FlowSOM) received more citations than algorithms in the same class. A caveat in focusing on the popularity of a tool is that it does not necessarily provide information on its fitness for purpose, in this regard further investigations on a correlation between popularity and performance is warranted.

This analysis of journal publications is a historical viewpoint over 10 years, but it does not necessarily provide a real-time perspective in this highly dynamic environment with new toolsets appearing on a yearly basis. Therefore, it was important to gather new information in the communities that perform flow cytometry data analysis on a regular basis. The external quality assessment (EQA) space with a large range of clinical participants was an ideal platform to investigate this issue.

5 | CLINICAL LABORATORY USERS SURVEY

To obtain a full picture of the popularity of automated flow cytometry data analysis software, it was important to gain insight on their actual use within clinical centers, not apparent from literature citations. An invitation to participate in a survey was distributed to laboratories worldwide registered with the EQA/proficiency testing programme from UK NEQAS for Leucocyte Immunophenotyping. The survey aimed for a broad overview and was not intended to extract actual participant use of specific functions of software. Survey distribution occurred in January 2020 and responses were gathered over 1 month. The online survey of eight questions (Table S2) was developed to expand on the literature review to understand the potential use of automated software in clinical laboratories.

5.1 | Survey results

The survey received 49 responses out of 310 potential respondents, a response rate of 16% which is consistent with typical response rates of 15%–20% from email invitations to participate in online, non-incentivized surveys [87]. The quality of respondents is high because of the targeted nature of the survey to subscribers of an EQA programme. Although conclusions from 49 responders should be carefully considered, the survey is valuable in providing strong suggestions of behavior on the current use of automated FC tools in clinical laboratories. The survey found more than half of respondents (26 out of 49, 53%) never use automated FC software and only use manual gating to identify cell populations (Figure 5A). Thirteen of 49 (27%) mainly use manual gating but occasionally use automated software, and 9 of 49 (18%) split their analysis between manual and automated methods. One respondent mainly uses automated software but occasionally use manual gating. The results suggest (on this basis) that most clinical laboratories rely on manual gating to identify cell populations and the use of automated methods have yet to be firmly established. The observed pattern of adoption is expected given the emerging nature of the software.

The survey asked participants to identify which automated data analysis software they used (Figure 5B). Nine software platforms were identified among the 16 respondents who used automated software, the most frequently identified of which was Infinicyt (63%). Other software selected included AutoGate (31%) [88] and FACSCanto (19%).

The survey also asked participants to identify software they were aware of but do not currently use (Figure 5C). A total of 37 automated data analysis tools were identified by respondents, an increase of 28 from the number of software respondents actually used. Once again, Infinicyt software was the most popular response (60%), followed by FlowSOM (44%), t-SNE (28%), viSNE (24%), and COMPASS (24%). For further insight, responses were grouped according to manual-only users (never use automated software) and automated users. This grouping revealed the automated-user base of respondents were aware of a wider range of software available compared with manual-only respondents: 36 software were identified by (13 of 23) automated users compared with eight identified by (11 out of 26) manual users.

The results gathered from this question suggest many laboratories were aware of what software was available but perhaps have not had the time or resources available to validate and implement changes to a manual gating protocol to incorporate automated analysis. It is also possible that laboratories first consider the many software packages available before committing to purchase only one software package, such as Infinicyt. Furthermore, software selection may be partly

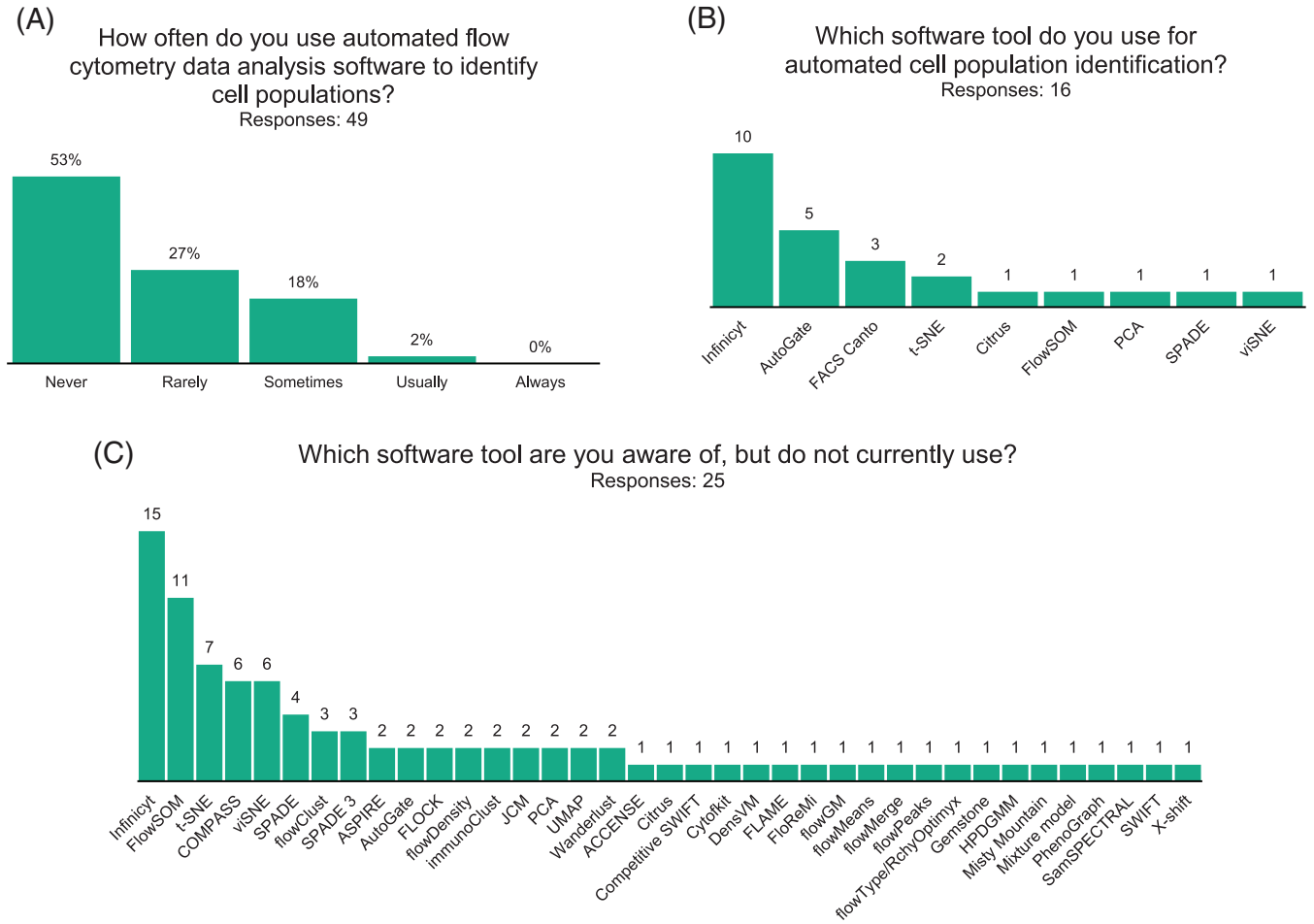


FIGURE 5 Results of a survey of clinical laboratories on the use of automated flow cytometry software [Color figure can be viewed at wileyonlinelibrary.com]

influenced by common consortium recommendations or EQA schemes.

To understand the factors which users consider important when using automated software, survey participants were asked to grade the importance of factors along a 5-point scale from "not important at all" to "extremely important" (Figure S2A,B). Results from this question revealed the most important factors for users was the software data output quality, followed by software speed, and the level of technical support. Of lesser importance, scored in decreasing order, were factors such as compatibility with other software, cost, software reputation, software availability, and, seen in the literature. The appearance of software was the lowest ranked importance factor in the survey.

To further understand the user interaction with automated software and the potential impact this has on software selection, development and data quality, the survey participants were asked in Question 8 to assess the software they were most familiar with by responding to 10 usability statements on a 1 to 5 score scale from "strongly disagree" to "strongly agree." The statements are based on the System Usability Scale (SUS) and are designed to provoke extreme disagreement or agreement among all respondents [89]. Statements that commonly lead to strong disagreement alternate with those that lead to strong agreement, to prevent response biases. This arrangement allows calculation of the SUS score, where (a) the score of each odd-numbered statement minus 1, and (b) the score of each even-numbered statement taken away from five, are summed then multiplied by 2.5 to obtain a score out of 100, with higher scores indicating better usability. Scores for individual statements are not meaningful on their own and need to be taken together to give a measure of the overall software usability.

This question received six responses ranking 5 software (Figure S2C). From the individual surveys, AutoGate, FACS Canto and FlowMerge received SUS scores above 70, therefore were judged to have "acceptable" usability based on the benchmark provided by Bangor et al. [90]. Compass received a SUS score below 70, indicating "marginally acceptable" usability. Infinicyt received a SUS score below 50, falling into the "unacceptable" region. To our knowledge, this is the first application of the SUS to quantify the usability performance of flow cytometry automated software. While the number of responses to this question were too low to draw conclusions from, it was interesting to note that the most identified software among the survey was also the least user-friendly, and as we anticipate the field of computational cytometry to mature and for user uptake to increase, these initial SUS scores

we have calculated will provide a critical baseline for future benchmarking studies to compare against.

The clinical survey results showed that 16 respondents identified nine software they make use of, and 25 respondents identified 37 software they were aware of but do not use. Excluding duplicates, 38 unique software were identified. A cross comparison with the 51 software identified from the literature review reveal the majority (36 of the 38) were included in both surveys. Two software do not appear in the literature review, which happened to be commercially available ones. This shows good correlation between the two information streams.

The analysis of literature captured the software used at a point in time (over 10 years) that precedes current usages, whereas the clinical survey revealed the most up to date patterns of use. Because of this contrast in timepoints, the clinical survey captured only two more software that slipped through the literature review.

The most frequently identified software by the online survey, Infinicyt, was not identified as a function of the original literature search strategy, and not mentioned in previous reviews on automated analysis tools. Infinicyt is proprietary software for analysis of multidimensional flow cytometry data, developed with support from the EuroFlow Consortium for standardization of immunophenotyping protocols [91]. The main feature of Infinicyt is the supervised learning algorithm for automatic identification and classification of cell populations based on reference databases built from merged multicenter patient files [92, 93]. Application of these Infinicyt tools are optimized to samples acquired following fully standardized EuroFlow standard operating procedures, reagents, instrument settings, and eight-color antibody panels for hematological malignancies [94]. The database-guided tool has been shown to successfully classify acute leukemia cases using a database constructed from 656 patients [93]. The software is also designed to be integrated with a laboratory information system (LIS) for secure handling of patient data. The highly specialized purpose of Infinicyt for clinical diagnostics explains its common use in clinical laboratories survey, and perhaps its underrepresentation in research areas.

Another popular software among the clinical laboratories, FACSCanto, was not featured in the original literature search because of the lack of peer-reviewed published work on its automated cell population identification function, but the clinical survey has identified it. The survey participants used FACSCanto software for analysis of CE-in vitro diagnostic (IVD)-marked assays such as CD4 and CD34 absolute count analysis. The software provides automated analysis of workflows and, similar to Infinicyt, is designed for clinical cytometry with LIS enabled connectivity. FACSCanto software popularity is possibly influenced by its bundled distribution with BD cytometer equipment and is therefore used by default by operators.

Common software highly ranked in both the literature citation analysis and the online survey were: FlowSOM, t-SNE, viSNE, and SPADE1. Overall, although uptake of automated software is growing, manual gating remains the standard practice. For clinical laboratory users, the most important component of automated software is the data output quality. This factor was not obvious from the findings from the literature citations, and it will be interesting to investigate whether software popularity translates to data quality. For automated analysis techniques to overtake manual gating, not only do the cell population identification results have to replicate expert manual analysis, but the results obtained from algorithms must also be robust with cell population numbers that can be reported with confidence.

6 | CONCLUSIONS

Flow cytometry has evolved to a stage where data analysis can be approached with unsupervised and supervised learning methods that automatically cluster cell populations and classify samples corresponding to clinical outcomes. Automated techniques allow FC analysis without manual variability, subjectivity, and bias of gating, and thus many new methods have been developed in the field in the past decade. However, it should be recognized that many of the automated techniques require moderate to significant operator control of software variables (beyond the default settings) and hence human subjectivity within the data processing chain may still be apparent.

In this literature survey, the current state-of-the-art software have been identified and their popularity ranked based on literature citations. Although citation counts do not necessarily reflect the use of software in labs, they give a good indication. The purpose of this study was to define the prevalence and perceived volume of use of automated software, not specifics of use in a laboratory or manufacturing company. Highly ranked software included: viSNE, t-SNE, SPADE1, PhenoGraph, FlowSOM, and Citrus. A common attribute of these software packages is the availability of a GUI that increase ease of use and appearance. This highlights the importance of usability as a factor for uptake of automated software in the community. Moreover, these software are implemented in multiple platforms (Bioconductor, FlowJo, Cytobank), and provide novel visualization outputs to aid interpretation of the data. Trends between software frequency of citation and factors such as cost or the underlying algorithm type were not apparent.

In addition to the literature survey, an online questionnaire of clinical laboratories on the use of automated FC software was completed via the external quality assessment (EQA)/proficiency testing programme from UK NEQAS for Leucocyte Immunophenotyping. This survey collected actual real-world usage data and opinions about automated FC data analysis software from a global targeted audience which could not be obtained from the literature search. Noting that this analysis was based on 49 respondents out of a possible 310 participants, a strength of this survey lies in its distribution through the EQA network rather than a public medium, which ensured genuine trustworthy responses. Very few surveys of this nature have been published in the literature. The online questionnaire did not capture users in similarly highly regulated spaces such as biotechnology, pharmaceutical and contract research organizations and so the rollout of a similar survey is planned to better understand automated software usage trends within these groups. However, distribution of a survey to those parties will be more difficult because they do not necessarily

subscribe to a comparable EQA network, so networks from the International Society for Advancement of Cytometry (ISAC) and the International Clinical Cytometry Society (ICCS) could potentially be explored in the future.

Most frequently identified automated software for clinical cytometrists were Infinicyt and FACSCanto, noting that 53% of participants stated that they never used these automated tools. Infinicyt in particular makes use of large reference patient databases to classify new patient samples using a supervised learning algorithm. These software have highly specialized workflows for analysis of regulated clinical assays to automated immunophenotyping, along with an important feature to connect with a hospital laboratory information system (LIS) to securely manage patient data.

The contrast in software popularity between the two complimentary surveys reflect the different needs and behaviors of the two communities. Clinical users are more likely to run routine, well-defined assays with standardized processes to enable confident diagnostics of patient samples. For example, the highly standardized ISHAGE protocol for enumeration of hematopoietic stem cells in peripheral blood recommends the use of specific antibody conjugates and prescribes manual gating strategies to identify target cell populations [95]. In this respect, clinical users lean toward tools that replicate expert manual gating and can automate targeted analysis of well-defined populations.

This is different in academia, where research is performed on well-defined cell subsets alongside unknown target cell populations, and hence users make more use of automated tools that support discovery and exploratory research.

The standardized data sets produced across clinical settings with the same experimental parameters, and crucially linked with specific patient outcomes, can be grouped to build a large database collection that allows for their use as training data sets for the development of supervised learning algorithms. In comparison, the academic space is less likely to have a large and diverse resource of labeled data to use for training purposes, and therefore is dominated by use of unsupervised learning methods. Overall, there is no "best" method. The most suitable automated analysis tools to use will be context dependent, on factors such as cell type, the data structure, and the purpose of the analysis. The best case is to provide users with complete details of how tools work, for them to make a well-informed decision. This may call for additional benchmarking methods/results from a wider selection of data sets.

More than half of the respondents from the clinical survey never use automated analysis tools and only use manual gating protocols, suggesting barriers to adoption of software are widespread.

The questionnaire gave an insight into the clinical users' software preferences when incorporating automated workflows into their data analysis. High value was given to the data output quality, speed of software and level of technical support. The low take-up in automated software may be down to shortcomings in all three factors in the current software available. The most critical factor, quality of the data, is a major driver for the use of automated software. Tools that aid rigor and reproducibility are expected to be welcomed, so it is intriguing that adoption rates are low, but it may be down to human sentiment and trust in manual methods.

With respect to the speed of software, because results need to reported in a timely (or possibly urgent) manner for clinicians to make decisions on patient treatment strategies, the analysis time needs to be in the order of seconds and minutes rather than hours and days. Current automated software may not offer significantly faster gains in analysis times over manual analysis that would incentivize uptake. Finally, better documentation in the form of detailed user manuals, video tutorials, and troubleshooting guides would increase the level of technical support, and make automated analysis more widely used.

Regulatory requirements are a possible factor for the low uptake of automated methods in the clinical laboratory. Implementation of new diagnostic methods is driven by international guidelines (e.g., World Health Organization (WHO), International Council for Standardization in Hematology (ICSH), International Clinical Cytometry Society (ICCS)). Consensus guidelines regarding use of automated methods have yet to be established. Even once guidelines are published, implementing new protocols at the laboratory level requires documenting process change controls, validations, and verifications in line with quality management system ISO 15189:2012 [96]. The increased regulatory requirements in clinical spaces compared with academia may be a barrier to uptake. Diagnostic methods are typically developed on an individual disease or biomarker basis, so are narrow in scope by nature. This means the pace of automated adoption occurs one test at a time, rather than all the tests involving flow cytometry changing to automated analysis at once.

An interesting factor to investigate further may be whether the number of colors in a staining panel correlates with uptake or use of an algorithm. As the burden of manual analysis increases with the number of parameters in a panel, perhaps clinical laboratories with more complex panels will be keener adopters of automated software that offer more efficient, scalable and unbiased analyses.

The awareness of new tools can be more dated among the clinical workforce because day-to-day sample processing demands reduces the time available to keep up to date with the latest literature.

There are now trends for academic users to acquire programming skills in R, Python and Matlab to keep up with data analysis requirements. This is a less likely scenario in clinical laboratories and may be the reason for the lower uptake of tools that are executed in those programming environments.

To a certain degree, usage of these tools relies on the efforts of commonly used stand-alone software packages (e.g., FlowJo, FCSExpress) to implement automated tools as plugins integrated into their GUIs. The skills shortage presents a risk to employers, whether to train up staff to be knowledgeable in coding but lose that tacit knowledge when they leave the company, or to buy in a ready-made software with full GUI that does not require specialist training and is easy to learn for new users. Indeed, this study has shown a user preference for tools with GUI. The implication could be for high performing software without a GUI losing ground to lower quality but easier to use software.

In this paper, we have investigated the current usage trends and popularity of automated flow cytometry data analysis software. However, it is worth emphasizing that the popularity of a tool does not indicate whether it is the correct or best approach of analyzing data, and therefore a key question that has emerged from this study is whether popularity translates to quality. It is clear that challenges in the data output quality from automated software remain a hurdle to the widespread uptake of software in flow cytometry. This is an opportunity for further work to assess the actual performance of different algorithm types through a range of benchmarking real-world experimental and simulated data sets with controlled cell characteristics.

ACKNOWLEDGMENTS

The authors would like to acknowledge all participants that completed the clinical laboratory survey.

AUTHOR CONTRIBUTIONS

Melissa Cheung: Conceptualization; formal analysis; investigation; methodology; visualization; writing-original draft; writing-review & editing. **Jonathan Campbell:** Writing-review & editing. **Liam Whitby:** Writing-review & editing. **Robert Thomas:** Writing-review & editing. **Julian Braybrook:** Writing-review & editing. **Jon Petzing:** Conceptualization; funding acquisition; project administration; supervision; writing-review & editing.

CONFLICT OF INTEREST

The authors have no conflicts of interest.

ORCID

Liam Whitby https://orcid.org/0000-0002-5218-2593

REFERENCES

1. Maecker HT, Rinfret A, D'Souza P, Darden J, Roig E, Landry C, et al. Standardization of cytokine flow cytometry assays. BMC Immunol. 2005;6:1–18.
2. Grant R, Coopman K, Medcalf N, Silva-Gomes S, Campbell JJ, Kara B, et al. Understanding the contribution of operator measurement variability within flow cytometry data analysis for quality control of cell and gene therapy manufacturing. Measurement. 2020;150:106998.
3. Grant R, Coopman K, Medcalf N, Silva-Gomes S, Kara B, Campbell JJ, et al. Quantifying operator subjectivity within flow cytometry data analysis as a source of measurement uncertainty and the impact of experience on results. PDA J Pharm Sci Technol. 2020;pdajpst.2019.011213. https://doi.org/10.5731/pdajpst.2019.011213.
4. Bendall SC, Nolan GP, Roederer M, Chattopadhyay PK. A deep profiler's guide to cytometry. Trends Immunol. 2012;33(7):323–32.
5. Bendall SC, Simonds EF, Qiu P, El-ad DA, Krutzik PO, Finck R, et al. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. Science (80-). 2011;332(6030):687–96.
6. Mair F, Prlic M. OMIP-044: 28-color immunophenotyping of the human dendritic cell compartment. Cytom Part A. 2018;93(4):402–5.
7. O'Neill K, Aghaeepour N, Špidlen J, Brinkman R. Flow cytometry bioinformatics. PLoS Comput Biol. 2013;9(12).
8. Bashashati A, Brinkman RRA. Survey of flow cytometry data analysis methods. Adv Bioinforma. 2009;2009:1–19.
9. Finak G, Langweiler M, Jaimes M, Malek M, Taghiyar J, Korin Y, et al. Standardizing flow cytometry Immunophenotyping analysis from the human ImmunoPhenotyping consortium. Sci Rep. 2016;6:1–11.
10. Aghaeepour N, Finak G, Hoos H, Mosmann TR, Brinkman R, Gottardo R, et al. Critical assessment of automated flow cytometry data analysis techniques. Nat Methods. 2013;10(3):228–38.
11. Aghaeepour N, Chattopadhyay P, Chikina M, Dhaene T, Van Gassen S, Kursa M, et al. A benchmark for evaluation of algorithms for identification of cellular correlates of clinical outcomes. Cytom Part A. 2016;89(1):16–21.
12. Weber LM, Nowicka M, Soneson C, Robinson MD. diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering.. Commun Biol. 2019;2(183):1–11.
13. Van Gassen S, Callebaut B, Van Helden MJ, Lambrecht BN, Demeester P, Dhaene T, et al. FlowSOM: using self-organizing maps for visualization and interpretation of cytometry data. Cytom Part A. 2015;87(7):636–45.
14. Samusik N, Good Z, Spitzer MH, Davis KL, Nolan GP. Automated mapping of phenotype space with single-cell data. Nat Methods. 2016;13(6):493–6.
15. Levine JH, Simonds EF, Bendall SC, Davis KL, Amir EAD, Tadmor MD, et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. Cell. 2015;162(1):184–97.
16. Aghaeepour N, Nikolic R, Hoos HH, Brinkman RR. Rapid cell population identification in flow cytometry data. Cytom Part A. 2011;79 A (1):6–13.
17. Pedersen NW, Chandran PA, Qian Y, Rebhahn J, Petersen NV, Hoff MD, et al. Automated analysis of flow cytometry data to reduce inter-lab variation in the detection of major histocompatibility complex multimer-binding T cells. Front Immunol. 2017;8:1–12.
18. Melchiotti R, Gracio F, Kordasti S, Todd AK, de Rinaldis E. Cluster stability in the analysis of mass cytometry data. Cytom Part A. 2017;91(1):73–84.
19. Czechowska K, Lannigan J, Aghaeepour N, Back JB, Begum J, Behbehani G, et al. Cyt-Geist: current and future challenges in cytometry: reports of the CYTO 2019 conference workshops. Cytom Part A. 2019;95(12):1236–74.
20. Saeys Y, Van Gassen S, Lambrecht BN. Computational flow cytometry: helping to make sense of high-dimensional immunology data. Nat Rev Immunol. 2016;16(7):449–62.
21. Zunder ER, Lujan E, Goltsev Y, Wernig M, Nolan GP. A continuous molecular roadmap to iPSC reprogramming through progression analysis of single-cell mass cytometry. Cell Stem Cell. 2015;16(3):323–37.
22. Spitzer MH, Gherardini PF, Fragiadakis GK, Bhattacharya N, Yuan RT, Hotson AN, et al. An interactive reference framework for modeling a dynamic immune system. Science (80-). 2015;349(6244).
23. Maaten L, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res. 2008;9(Nov):2579–605.
24. Kotecha N, Krutzik PO, Irish JM. Web-based analysis and publication of flow cytometry experiments. Curr Protoc Cytom. 2010;53(1):10–17.
25. Becher B, Schlitzer A, Chen J, Mair F, Sumatoh HR, Teng KWW, et al. High-dimensional analysis of the murine myeloid cell system. Nat Immunol. 2014;15(12):1181–9.
26. Malek M, Taghiyar MJ, Chong L, Finak G, Gottardo R, Brinkman RR. FlowDensity: reproducing manual gating of flow cytometry data by automated density-based cell population identification. Bioinformatics. 2015;31(4):606–7.
27. Bruggner RV, Bodenmiller B, Dill DL, Tibshirani RJ, Nolan GP. Automated identification of stratifying signatures in cellular subpopulations. Proc Natl Acad Sci. 2014;111(26):E2770–7.
28. Finak G, Frelinger J, Jiang W, Newell EW, Ramey J, Davis MM, et al. OpenCyto: an open source infrastructure for scalable, robust,

reproducible, and automated, end-to-end flow cytometry data analysis. PLoS Comput Biol. 2014;10(8):1–12.
29. Lin L, Finak G, Ushey K, Seshadri C, Hawn TR, Frahm N, et al. COMPASS identifies T-cell subsets correlated with clinical outcomes. Nat Biotechnol. 2015;33(6):610–6.
30. O'Neill K, Jalali A, Aghaeepour N, Hoos H, Brinkman RR. Enhanced flowType/RchyOptimyx: a bioconductor pipeline for discovery in high-dimensional cytometry data. Bioinformatics. 2014;30(9): 1329–30.
31. Dundar M, Akova F, Yerebakan HZ, Rajwa B. A non-parametric Bayesian model for joint cell clustering and cluster matching: identification of anomalous sample phenotypes with random effects. BMC Bioinformat. 2014;15(1):1–15.
32. Li H, Shaham U, Stanton KP, Yao Y, Montgomery RR, Kluger Y. Gating mass cytometry data by deep learning. Bioinformatics. 2017;33(21): 3423–30.
33. Van Gassen S, Vens C, Dhaene T, Lambrecht BN, Saeys Y. FloReMi: flow density survival regression using minimal feature redundancy. Cytom Part A. 2016;89(1):22–9.
34. Lux M, Brinkman RR, Chauve C, Laing A, Lorenc A, Abeler-Dörner L, et al. FlowLearn: fast and precise identification and quality checking of cell populations in flow cytometry. Bioinformatics. 2018;34(13): 2245–53.
35. Lee HC, Kosoy R, Becker CE, Dudley JT, Kidd BA. Automated cell type discovery and classification through knowledge transfer. Bioinformatics. 2017;33(11):1689–95.
36. Rebhahn JA, Roumanes DR, Qi Y, Khan A, Thakar J, Rosenberg A, et al. Competitive SWIFT cluster templates enhance detection of aging changes. Cytom Part A. 2016;89(1):59–70.
37. Lee AJ, Chang I, Burel JG, Lindestam Arlehamn CS, Mandava A, Weiskopf D, et al. DAFi: a directed recursive data filtering and clustering approach for improving and interpreting data clustering identification of cell populations from polychromatic flow cytometry data. Cytom Part A. 2018;93(6):597–610.
38. Weber LM, Nowicka M, Soneson C, Robinson MD. diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering. Commun. Biol. 2019;2(183):1–11.
39. Reiter M, Rota P, Kleber F, Diem M, Groeneveld-Krentz S, Dworzak M. Clustering of cell populations in flow cytometry data using a combination of Gaussian mixtures. Pattern Recogn. 2016;60: 1029–40.
40. Abdelaal T, van Unen V, Höllt T, Koning F, Reinders MJT, Mahfouz A. Predicting cell populations in single cell mass cytometry data. Cytom Part A. 2019;95(7):769–81.
41. Wong L, Hill BL, Hunsberger BC, Bagwell CB, Curtis AD, Davis BH. Automated analysis of flow cytometric data for measuring neutrophil CD64 expression using a multi-instrument compatible probability state model. Cytom Part B - Clin Cytom. 2015;88(4):227–35.
42. Xu D, Tian Y. A comprehensive survey of clustering algorithms. Ann Data Sci. 2015;2(2):165–93.
43. Kaufman L, Rousseeuw PJ. In: Kaufman L, Rousseeuw PJ, editors. Finding groups in data: an introduction to cluster analysis. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 1990 (Wiley Series in Probability and Statistics).
44. Qiu P, Simonds EF, Bendall SC, Gibbs KD, Bruggner RV, Linderman MD, et al. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. Nat Biotechnol. 2011;29(10): 886–93.
45. Qiu P. Toward deterministic and semiautomated SPADE analysis. Cytom Part A. 2017;91(3):281–9.
46. Jain AK. Data clustering: 50 years beyond K-means. Pattern Recogn Lett. 2010 Jun 1;31(8):651–66.
47. Ge Y, Sealfon SC. Flowpeaks: a fast unsupervised clustering for flow cytometry data via K-means and density peak finding. Bioinformatics. 2012;28(15):2052–8.
48. Pouyan MB, Nourani M. Identifying cell populations in flow cytometry data using phenotypic signatures. IEEE/ACM Trans Comput Biol Bioinforma. 2017;14(4):880–91.
49. Ester M, Kriegel H-P, Sander J, Xu X. A Density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the 2nd international conference on knowledge discovery and data mining. Vol. 96, 34th ed. Menlo Park, California: AAAI Press; 1996. p. 226–31.
50. Ankerst M, Breunig MM, Kriegel HP, Sander J. OPTICS: ordering points to identify the clustering structure. SIGMOD Rec (ACM Spec Interes Gr Manag Data). 1999;28(2):49–60.
51. Shekhar K, Brodin P, Davis MM, Chakraborty AK. Automatic classification of cellular expression by nonlinear stochastic embedding (ACCENSE). Proc Natl Acad Sci. 2014;111(1):202–7.
52. Qian Y, Wei C, Eun-Hyung Lee F, Campbell J, Halliley J, Lee JA, et al. Elucidation of seventeen human peripheral blood B-cell subsets and quantification of the tetanus response using a density-based method for the automated identification of cell populations in multidimensional flow cytometry data. Cytom Part B Clin Cytom. 2010; 78B(S1):S69–82.
53. Sugar IP, Sealfon SC. Misty Mountain clustering: application to fast unsupervised flow cytometry gating. BMC Bioinformat. 2010;11 (1):502.
54. Folcarelli R, van Staveren S, Bouman R, Hilvering B, Tinnevelt GH, Postma G, et al. Automated flow cytometric identification of disease-specific cells by the ECLIPSE algorithm. Sci Rep. 2018;8(1):1–18.
55. Zaunders J, Jing J, Leipold M, Maecker H, Kelleher AD, Koch I. Computationally efficient multidimensional analysis of complex flow cytometry data using second order polynomial histograms. Cytom Part A. 2016;89(1):44–58.
56. Meehan S, Kolyagin GA, Parks D, Youngyunpipatkul J, Herzenberg LA, Walther G, et al. Automated subset identification and characterization pipeline for multidimensional flow and mass cytometry data clustering and visualization. Commun Biol. 2019;2 (1):1–12.
57. Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. J Am Stat Assoc. 2002;97(458):611–31.
58. Pyne S, Hu X, Wang K, Rossin E, Lin T-I, Maier LM, et al. Automated high-dimensional flow cytometric data analysis. Proc Natl Acad Sci. 2009;106(21):8519–24.
59. Lo K, Hahne F, Brinkman RR, Gottardo R. flowClust: a Bioconductor package for automated gating of flow cytometry data. BMC Bioinformat. 2009;10(1):145.
60. Finak G, Bashashati A, Brinkman R, Gottardo R. Merging mixture components for cell population identification in flow cytometry. Adv Bioinforma. 2009;2009:1–12.
61. Naim I, Datta S, Rebhahn J, Cavenaugh JS, Mosmann TR, Sharma G. SWIFT-scalable clustering for automated identification of rare cell populations in large, high-dimensional flow cytometry datasets, part 1: algorithm design. Cytom Part A. 2014;85(5):408–21.
62. Mosmann TR, Naim I, Rebhahn J, Datta S, Cavenaugh JS, Weaver JM, et al. SWIFT-scalable clustering for automated identification of rare cell populations in large, high-dimensional flow cytometry datasets, part 2: biological evaluation. Cytom Part A. 2014;85(5):422–33.
63. Boedigheimer MJ, Ferbas J. Mixture modeling approach to flow cytometry data. Cytom Part A. 2008;73(5):421–9.
64. Cron A, Gouttefangeas C, Frelinger J, Lin L, Singh SK, Britten CM, et al. Hierarchical Modeling for rare event detection and cell subset alignment across flow cytometry samples. Altan-bonnet G. PLoS Comput Biol. 2013 Jul 11;9(7):e1003130.
65. Rogers WT, Moser AR, Holyst HA, Bantly A, Mohler ER, Scangas G, et al. Cytometric fingerprinting: quantitative characterization of multivariate distributions. Cytom Part A. 2008;73(5):430–41.
66. Sörensen T, Baumgart S, Durek P, Grützkau A, Häupl T. immunoClust-an automated analysis pipeline for the identification of

immunophenotypic signatures in high-dimensional cytometric datasets. Cytom Part A. 2015;87(7):603–15.
67. Pyne S, Lee SX, Wang K, Irish J, Tamayo P, Nazaire MD, et al. Joint modeling and registration of cell populations in cohorts of high-dimensional flow cytometric data. PLoS One. 2014;9(7):e100334.
68. Anchang B, Do MT, Zhao X, Plevritis SK. CCAST: a model-based gating strategy to isolate homogeneous subpopulations in a heterogeneous population of single cells. PLoS Comput Biol. 2014;10(7):13–7.
69. Commenges D, Alkhassim C, Gottardo R, Hejblum B, Thiébaut R. Cytometree: a binary tree algorithm for automatic gating in cytometry analysis. Cytom Part A. 2018;93(11):1132–40.
70. Hejblum BP, Alkhassim C, Gottardo R, Caron F, Thiébaut R. Sequential Dirichlet process mixtures of multivariate skew t-distributions for model-based clustering of flow cytometry data. Ann Appl Stat. 2019;13(1):638–60.
71. Von Luxburg U. A tutorial on spectral clustering. Stat Comput. 2007 Dec 22;17(4):395–416.
72. Zare H, Shooshtari P, Gupta A, Brinkman RR. Data reduction for spectral clustering to analyze high throughput flow cytometry data. BMC Bioinformat. 2010;11:403.
73. Bendall SC, Davis KL, Amir EAD, Tadmor MD, Simonds EF, Chen TJ, et al. Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development. Cell. 2014;157(3):714–25.
74. Kohonen T. The self-organizing map. Neurocomputing. 1998;21(1–3):1–6.
75. Kohonen T. Essentials of the self-organizing map. Neural Netw. 2013 Jan 1;37:52–65.
76. Vesanto J, Alhoniemi E. Clustering of the self-organizing map. IEEE Trans Neural Netw. 2000;11:586–600.
77. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol. 2019;37(1):38–47.
78. Amir EAD, Davis KL, Tadmor MD, Simonds EF, Levine JH, Bendall SC, et al. ViSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. Nat Biotechnol. 2013;31(6):545–52.
79. Diggins KE, Brent Ferrell P, Irish JM. Methods for discovery and characterization of cell subsets in high dimensional mass cytometry data. Methods. 2015;82:55–63.
80. Van Unen V, Höllt T, Pezzotti N, Li N, Reinders MJT, Eisemann E, et al. Visual analysis of mass cytometry data by hierarchical stochastic neighbour embedding reveals rare cell types. Nat Commun. 2017;8(1):1–10.
81. Moon KR, van Dijk D, Wang Z, Gigante S, Burkhardt DB, Chen WS, et al. Visualizing structure and transitions in high-dimensional biological data. Nat Biotechnol. 2019;37(12):1482–92.
82. Fletez-Brant K, Špidlen J, Brinkman RR, Roederer M, Chattopadhyay PK. flowClean: automated identification and removal of fluorescence anomalies in flow cytometry data. Cytom Part A. 2016;89(5):461–71.
83. Monaco G, Chen H, Poidinger M, Chen J, de Magalhães JP, Larbi A. flowAI: automatic and interactive anomaly discerning tools for flow cytometry data. Bioinformatics. 2016;32(16):2473–80.
84. Hahne F, LeMeur N, Brinkman RR, Ellis B, Haaland P, Sarkar D, et al. flowCore: a Bioconductor package for high throughput flow cytometry. BMC Bioinformat. 2009;10(1):1–8.
85. Hahne F, Gopalakrishnan N, Khodabakhshi AH, Wong C, Lee K. flowStats: statistical methods for the analysis of flow cytometry data. R Package version. 2009;3(1).
86. Finak G, Jiang M. flowWorkspace: infrastructure for representing and interacting with the gated cytometry. R Package version. 2011;3(3).
87. Pedersen MJ, Nielsen CV. Improving survey response rates in online panels: effects of Low-cost incentives and cost-free text appeal interventions. Soc Sci Comput Rev. 2016;34(2):229–43.
88. Meehan S, Walther G, Moore W, Orlova D, Meehan C, Parks D, Ghosn E, Philips M, Mitsunaga E, Waters J, Kantor A, Okamura R, Owumi S, Yang Y, Herzenberg LA, Herzenberg LA. AutoGate: automating analysis of flow cytometry data. Immunol Res. 2014;58(2-3):218–223.
89. Brooke J. SUS-A quick and dirty usability scale. Usability Eval Ind. 1996;189(194):4–7.
90. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the system usability scale. Int J Hum Comput Interact. 2008;24(6):574–94.
91. Kalina T, Flores-Montero J, Van Der Velden VHJ, Martin-Ayuso M, Böttcher S, Ritgen M, et al. EuroFlow standardization of flow cytometer instrument settings and immunophenotyping protocols. Leukemia. 2012;26(9):1986–2010.
92. Costa ES, Pedreira CE, Barrena S, Lecrevisse Q, Flores J, Quijano S, et al. Automated pattern-guided principal component analysis vs expert-based immunophenotypic classification of B-cell chronic lymphoproliferative disorders: a step forward in the standardization of clinical immunophenotyping. Leukemia. 2010;24(11):1927–33.
93. Lhermitte L, Mejstrikova E, Van Der Sluijs-Gelling AJ, Grigore GE, Sedek L, Bras AE, et al. Automated database-guided expert-supervised orientation for immunophenotypic diagnosis and classification of acute leukemia. Leukemia. 2018;32(4):874–81.
94. Van Dongen JJM, Lhermitte L, Böttcher S, Almeida J, Van Der Velden VHJ, Flores-Montero J, et al. EuroFlow antibody panels for standardized n-dimensional flow cytometric immunophenotyping of normal, reactive and malignant leukocytes. Leukemia. 2012;26(9):1908–75.
95. Sutherland DR, Anderson L, Keeney M, Nayar R, Chin-Yee I. The ISHAGE guidelines for CD34+ cell determination by flow cytometry. J Hematother. 1996;5(3):213–26.
96. ISO 15189:2012 Medical laboratories — Requirements for quality and competence.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Cheung M, Campbell JJ, Whitby L, Thomas RJ, Braybrook J, Petzing J. Current trends in flow cytometry automated data analysis software. *Cytometry*. 2021;99:1007–1021. https://doi.org/10.1002/cyto.a.24320