vaccines

MDPI

Review

From Bivariate to Multivariate Analysis of Cytometric Data: Overview of Computational Methods and Their Application in Vaccination Studies

Simone Lucchesi [1], Simone Furini [2], Donata Medaglini [1] and Annalisa Ciabattini [1,*]

[1] Laboratory of Molecular Microbiology and Biotechnology (LA.M.M.B.), Department of Medical Biotechnologies, University of Siena, 53100 Siena, Italy; lucchesi5@student.unisi.it (S.L.); donata.medaglini@unisi.it (D.M.)

[2] Department of Medical Biotechnologies, University of Siena, 53100 Siena, Italy; simone.furini@unisi.it

* Correspondence: annalisa.ciabattini@unisi.it

Received: 27 February 2020; Accepted: 18 March 2020; Published: 20 March 2020

Abstract: Flow and mass cytometry are used to quantify the expression of multiple extracellular or intracellular molecules on single cells, allowing the phenotypic and functional characterization of complex cell populations. Multiparametric flow cytometry is particularly suitable for deep analysis of immune responses after vaccination, as it allows to measure the frequency, the phenotype, and the functional features of antigen-specific cells. When many parameters are investigated simultaneously, it is not feasible to analyze all the possible bi-dimensional combinations of marker expression with classical manual analysis and the adoption of advanced automated tools to process and analyze high-dimensional data sets becomes necessary. In recent years, the development of many tools for the automated analysis of multiparametric cytometry data has been reported, with an increasing record of publications starting from 2014. However, the use of these tools has been preferentially restricted to bioinformaticians, while few of them are routinely employed by the biomedical community. Filling the gap between algorithms developers and final users is fundamental for exploiting the advantages of computational tools in the analysis of cytometry data. The potentialities of automated analyses range from the improvement of the data quality in the pre-processing steps up to the unbiased, data-driven examination of complex datasets using a variety of algorithms based on different approaches. In this review, an overview of the automated analysis pipeline is provided, spanning from the pre-processing phase to the automated population analysis. Analysis based on computational tools might overcame both the subjectivity of manual gating and the operator-biased exploration of expected populations. Examples of applications of automated tools that have successfully improved the characterization of different cell populations in vaccination studies are also presented.

Keywords: vaccination; multiparametric flow cytometry; mass cytometry; computational data analysis; automated analysis; machine learning

1. Introduction

Flow cytometry allows to simultaneously quantify expression of extracellular and intracellular molecules targeted by dyes or monoclonal antibodies, as well as to measure multiple characteristics of a single cell such as size and granularity. This technology emerged as a powerful tool for detailed analysis of complex populations and several other factors have contributed to the success and widespread use of flow cytometry. These include the speed at which cells are analyzed, the high accuracy and resolution of the technology, and the low operating costs per sample. The more recent mass cytometry, or cytometry by time-of-flight mass spectrometry (CyTOF), is another technique for measuring the expression of more than 40 parameters on large number of cells. In mass cytometry, antibodies specific to markers of

interest are conjugated to heavy-metal isotopes and used to stain a population of cells. Compared to mass cytometry, conventional flow cytometry is a non-destructive techniques which can be used to sort cells for further analyses and offers the highest throughput with tens of thousands of cells measured per second [1]. Considering the similarity of their outputs (files in Flow Cytometry Standard format), flow cytometry and mass cytometry share many analysis tools. While both techniques allow to interrogate the immune system at a previously unprecedented level, scientific progress depends on our ability to interpret these results. Classical analysis is performed by the operator, which manually explores the cells, and identifies cellular subsets by specific gates, that can be further analyzed for the expression of different markers combinations, thus providing a hierarchical analysis strategy. Manual analysis of cytometry data is a simple and intuitive one. However, it constitutes a big source of variability and it is time consuming when a large number of samples and markers are analyzed [2–4]. Moreover, experts are typically looking for specific and expected cell types, excluding other cells from the analysis [5]. Operator subjectivity occurs at the level of choosing the hierarchy in which parameter combinations have to be considered, as well as in the shape and boundary of each gate specified in the analysis. To overcome these limitations, novel computational techniques have been developed in recent years, and computational flow cytometry has become a novel discipline useful for providing a set of tools to analyze, visualize, and interpret large amounts of cell data in a more automated and unbiased way.

Comparative studies between traditional manual gating versus automated analysis have demonstrated that many of the available tools can efficiently achieve the same results produced by manual analysis [6,7] with the advantage of being operator-independent and able to identify also unexpected cell populations, that would be hardly identified with traditional methods. Supported by these positive results, in the last years automated analyses have been applied to the identification of different cell populations in pre-clinical and clinical studies in different fields such as immunology, vaccinology, cell biology, oncology, and hematology, contributing to a deeper understanding of biological processes. Since flow cytometry is a powerful technology for studying multiple immune function in response to vaccination, ranging from the phenotypic and functional characterization of cellular immune responses to antibody detection and functional assessment, the use of computational tools represents a powerful strategy for the interpretation of large datasets that can be instrumental to profile the vaccine immune response. For these reasons, immunologists should be aware of the potentiality of automated tools, which should not remain exclusive to computer-science experts

In this review, a pipeline for the automated analysis of multiparametric cytometry data is described, focusing on both the pre-processing and analysis phases. The advantages/disadvantages, potentialities, and possible applications of the most commonly used algorithms currently available are described in order to make immunologists/vaccinologists aware of the added value of the computational analysis approach. Moreover, an overview of the impact of automated analysis in the knowledge of biological processes, especially in the vaccine field, is presented.

2. Automated Cytometry Data Analysis Workflow

The workflow for the analysis of cytometric data includes pre-processing, automated analysis with data-visualization, and result interpretation (Figure 1). Computational tools exist for each of these steps both as standalone software, FlowJo plugins, web services, and libraries for some of the most common programming languages. A high number of tools are available as packages for R, a highly popular language for statistical analyses, and stored on the Bioconductor repository, an open-source free software project for the analysis of high-throughput biomedical data [8] (Table 1).

The term computational cytometry is commonly referred to this vast arsenal of tools proposed for supporting the analysis of cytometry data. Different tools are characterized by a different degree of automation. At one side of the spectrum, there are tools that provide a complete unsupervised clustering of cell populations based on Machine Learning models, while at the other side, there are supervised tools that are specifically designed to assist in the manual analysis. In this review, the terms

automated or computational analysis and automated tools will be used for indicating the computational analysis workflow and the different algorithms, respectively. The entire workflow of the computational analysis will be described step by step, starting from the pre-processing tools, through the different analysis phases, up to the result interpretation (Figure 1).

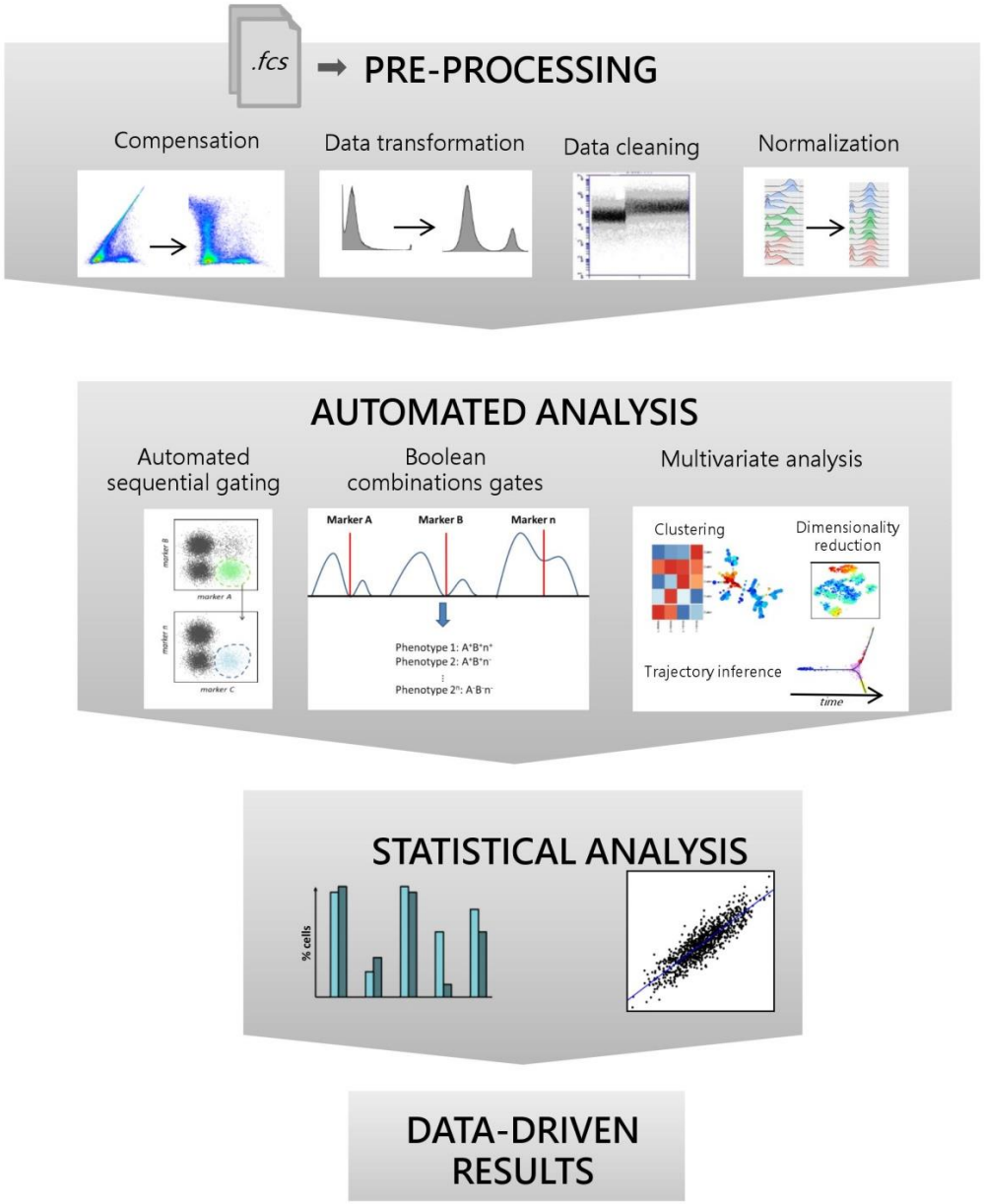

Figure 1. Automated cytometry data analysis workflow. Pipeline of the automated analysis steps of cytometric data. Starting from Flow Cytometry Standard (FCS) files, data are pre-processed to ensure reproducible and reliable results. Pre-processing phases include compensation (spectral overlap correction), data transformation (improvement of cell population visualization and automated cell types identification), data cleaning (removal of dead cells, debris, doublets, etc.), and normalizations (removal of batch effect between samples or balancing the contribution of each marker to the analysis). Pre-processed samples are analyzed with automated tools here classified as "Automated sequential gates", "Boolean combinations gates", and "Multivariate analysis" which include "Clustering algorithms", "Dimensionality reduction methods", and "Trajectory inference" techniques. Finally, statistical tests, correlation analysis and supervised machine learning techniques, such as regression and classification, can be applied to detect differences between experimental groups or to discover biomarkers.

Table 1. Automated tools for the analysis of cytometric data. For each tool is reported the function, the statistical platform in which they are available, and a brief description of the main function.

Function	Software	Availability	Description	Reference
Pre-processing	FlowCore	R, Bioconductor	Import, compensate and transform FCS files in R environment	[9]
	FlowStats	R, Bioconductor	Collection of algorithms to analyze flow cytometry data, including correction of batch effect	[10]
	FlowClean	R, Bioconductor FlowJo plugin	Quality control of data set based on compositional analysis	[11]
	FlowAI	R, Bioconductor FlowJo plugin	Quality control of data set based on flow rate, signal acquisition and dynamic range	[12]
	CATALYST	R, Bioconductor	Collection of algorithms to pre-process cytometric data and to perform data analysis (with FlowSOM clustering and dimensionality reduction)	[13]
	CytoNorm	R	Normalized batch effect using control sample and clustering algorithm	[14]
Automated sequential gating	FlowDensity	R, Bioconductor	Provides tools for automated 1-D and 2-D sequential gating	[15]
	OpenCyto	R, Bioconductor	Facilitates automated 1-D and 2-D gating methods in sequential way to mimic the manual gating	[16]
	AutoGate	Standalone software	Performs 2-D sequential gating to obviate the need to draw arbitrary gates to define the subsets in a gating	[17]
	cytometree	R	The algorithm relies on the construction of a binary tree, the nodes of which represents cellular populations	[18]
	EPP	Standalone software	AutoGate extension. Algorithm that detects the best 2-D gating strategy to identify cellular populations	[19]
Boolean combination gates	flowType	R, Bioconductor	Phenotyping cytometric using multi-dimensional expansion of 1-D partitions	[20]
	FloReMi	R	Starting from flowType results identifies the populations that best correlates with an external outcome	[21]
	RchyOptimyx	R, Bioconductor	Starting from flowType results, constructs a hierarchy of cells selecting the most informative phenotypes for biomarker detection	[22]
Clustering	FlowMeans	R, Bioconductor FlowJo plugin	Automated gating tool based on K-means algorithm	[23]
	SPADE	R, Matlab, Cytobank, FlowJo plugin	Clustering method based combining density-based sampling with hierarchical clustering	[24]
	HDPGMM	Python	Clustering based on hierarchical modeling extensions to the Dirichlet Process Gaussian Mixture Model	[25]
	Citrus	Cytobank, R	Identifies cell populations with hierarchical clustering and make prediction with regression model	[26]
	FlowSOM	R, Bioconductor FlowJo plugin, Cytobank	Clustering method combining SOM and hierarchical clustering	[27]

Table 1. *Cont.*

Function	Software	Availability	Description	Reference
	X-shift	Standalone software, FlowJo plugin	Clustering based on kNN density estimation and cluster merging according Mahalanobis distances	[28]
	flowClust	R, Bioconductor	Model-based clustering using a t-mixture model	[29]
	immunoClust	R, Bioconductor	Model-based clustering on individual samples. Includes an additional step to map cluster between samples	[30]
	SWIFT	Matlab	Clustering method based on splitting and merging of Gaussian mixture models	[31]
	FLOCK	C, Immport	Automated method partitioning of each dimension into bins, followed by merging of dense regions, and density-based clustering	[32]
	flowPeaks	R, Bioconductor	Clustering method combining density-based clustering and K-means	[33]
	ClusterX	R	Fast clustering by automatic search and find of density peaks	[34]
	PhenoGraph	Matlab, Python	Cells are visualized in a graph structure and connected with weighted edge based on neighbor shared by cell. Graph is then partitioned in group of cells sharing similar phenotypes	[35]
Dimensionality reduction	t-SNE	FlowJo plugin	Performs t-SNE in FlowJo, allowing to manually gate region in dimensionality reduced space to compare cell frequency across samples	[36]
	ACCENSE	Standalone software	Performs dimensionality reduction with t-SNE algorithm, followed by clustering of dimensionality reduced events with K-means or DBSCAN algorithms	[37]
	Rtsne	R	Performs t-SNE dimensionality reduction in R environment	[36]
	viSNE	Cytobank, Matlab	Visualization tool based on implementation of t-SNE algorithm	[38]
	EmbedSOM	R, Bioconductor FlowJo plugin	Dimensionality reduction technique based on SOM	[39]
	UMAP	R, Python, FlowJo plugin	Dimensionality reduction technique based on Uniform Manifold Approximation and Projection (UMAP)	[40]
	Destiny	R, Bioconductor	Performs dimensionality reduction with diffusion map	[41]
	Fit-SNE	R, Matlab, Python, FlowJo plugin	Tool to perform dimensionality reduction using Fast Fourier Transform-accelerated Interpolation-based t-SNE	[42]
Trajectory inference	Wanderlust	Matlab	Trajectory inference method based on kNN graph: Developed to identify linear transitions	[43]
	Wishbone	Matlab, Python	Evolution of Wanderlust, it can identify bifurcation in the trajectories	[44]
	Monocle	R, Bioconductor	Identification of bifurcated trajectory based on MST	[45]
	PHATE	Matlab, Python	Identification of trajectory preserving continual progressions, branches and clusters	[46]

R, package or code working on R; Bioconductor, R package available on Bioconductor repository [47]; Python, code or library written in Python language; Matlab, code or software based on Matlab language; C, code based on C programming language; FlowJo plugin, downloadable tools to expand FlowJo functionality [48]; Cytobank, online platform for single-cell analysis [49]; ImmPort, immunology database and analysis portal [50].

3. Data Pre-Processing

Cytometric data are usually provided in the form of FCS files. Before proceeding with further analyses, the raw-data included in FCS files needs to be pre-processed. Data pre-processing can be subdivided in four principal steps: (i) compensation; (ii) transformation; (iii) cleaning; and (iv) normalization (Figure 1).

Compensation. This step is necessary for adjusting the overlap between adjacent emission spectra of different fluorochromes. Compensation algorithms are included in most of the acquisition and analysis software packages that automatically perform this calculation.

Data transformation. Data transformation is an important step in automated analysis, as well as in manual analysis, that facilitates cell population visualization and identification. Automated tools can be influenced by asymmetric cell populations, frequent outlier events, cell populations whose variance depend on their mean fluorescence intensity, and multiplicative errors in the fluorescence channels. Data transformation plays an important role in mitigating these effects [51]. Data from flow cytometry experiments usually range over 3-5 orders of magnitude. Thus, mapping, at least part of the data range, to a logarithmic scale is often required, both for efficient visualization and automatic analysis with Machine Learning algorithms. However, standard logarithmic transformations are rarely adopted in flow cytometry for two main reasons. Firstly, as a consequence of the compensation step, described in the previous section, input data might assume negative values, which obviously cannot be represented into a pure logarithmic scale. Moreover, a logarithmic scale would shrink the data range close to zero, which could hamper the identification of cell populations with low marker values. A common strategy to avoid these problems is to transform data using the inverse of the hyperbolic sine, as this function guarantees an almost linear transformation for values close to zero, while it approaches a logarithmic scale both for highly positive and highly negative values. A drawback of the hyperbolic sine transformation is that it uses a single parameter to control both the width of the linear regime and its slope. In order to remove this restraint between width and slope, which could prevent the identification of an optimal transforming function, it was proposed to use biexponential functions for data transformation (which can be considered as generalization of the hyperbolic sine), and in particular a special class of biexponential functions known as the *logicle* transformation. An advantage of using the *logicle* transformation is that the parameters of the transforming function can be more easily defined, and that they are linked to easily interpretable characteristics of the original data (Figure 2).

Data cleaning. Marginal events, debris, dead cells, and doublets should be removed, either manually or automatically, as well as outliers, in order to use high quality data as input in the analysis. Tools like FlowClean [11] and FlowAI [12] aim to automatically remove cells derived from anomalies in the acquisition. Removing low quality cells reduces noise in data sets and avoids false positive results or loss of rare populations [12].

Normalizations. The previous steps are required both for manual and automated analyses, while data normalization is explicitly required only for multivariate analyses. The first step of data normalization is to estimate batch effects, i.e. inter-sample variation. The flowStats package includes two functions (warpSet and gaussNorm) to normalize the data on the base of high-density region landmarks for individual flow channels [10], while in mass cytometry normalization of batch effect can be performed with the packages CATALYST (Cytometry dATa anALYSis Tools) or CytoNorm [14,52,53].

In the second step of data normalization, the expression values of separate markers are modified so that different makers have similar expression ranges. This is needed as many clustering and dimensionality reduction algorithms compute the distance between cells or identify dense cell areas in multidimensional space, and these analyses would be hampered by the presence of highly different ranges among the various markers. To balance their contribution, each marker in the data set is normalized (normalization between markers, also referred to as scaling), employing the z-score or min-max normalization methods [32].

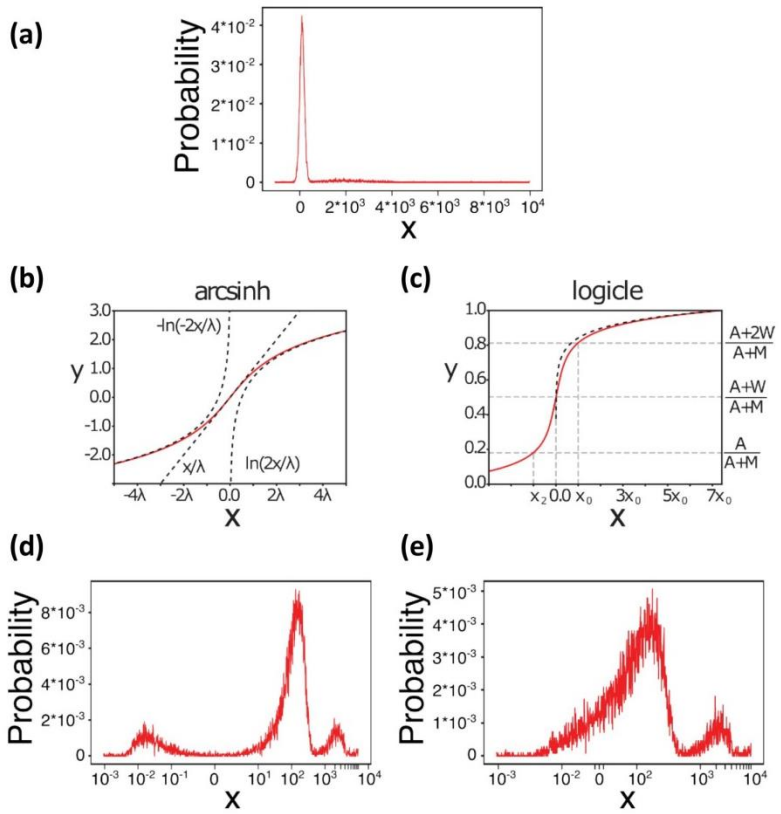

Figure 2. Data transformations. (**a**) A random data set was generated using two gaussian distributions, centered at 10^2 (10000 cells) and $2*10^3$ (1000 cells), and with standard deviation equal to 10^2 and 10^3, respectively. The probability histogram is shown on a linear scale. (**b**) Arcsinh transformation. The parameter λ defines both the width of the linear region, and its slope. The transforming function is approximately linear for λ close to zero, while it approaches logarithmic transformations when $x >> \lambda$ or when $x << -\lambda$. (**c**) Logicle transformation. The shape of the transforming function is defined by the parameters M, A, and W, which can be intuitively interpreted respectively as the number of decades, the number of negative decades, and the width of the linear region. (**d**) Probability histogram of the data in a transformed with the arcsinh function. (**e**) Probability histogram of the data in **a** transformed with the logicle function.

4. Automated Data Analysis

After pre-processing, the next phase in the automated analysis of cytometric data is the discovery and quantification of different cell populations. The main advantages and limits of different strategies are described below.

4.1. Automated Sequential Gating

Automated tools for sequential gating automatically compute gates around cell populations in bi-dimensional plots, overcoming the operator subjectivity due to manual drawing. Instead, the sequence of the markers analyzed is still defined by the operator. OpenCyto [16] and FlowDensity [15] R packages, as well as the standalone executable software AutoGate [17], assist the operator in the definition of mono- and bi-dimensional gates, by using methods for boundary definition based on density estimation techniques. The main advantage of the automated sequential gating approach is represented by the automated identification of the cell population in the bi-dimensional scatter plots, overcoming the limitations linked to the manual drawing of gate boundaries, thus improving reproducibility. An evolution of these algorithms is represented by tools that automatically identify the gating strategy. Cytometree, implemented as R package, aims to construct a binary tree in which the nodes represent gates and the binary tree represents the

best mono-dimensional sequential gating strategy used to identify the cellular sub-populations [18]. The AutoGate software has been recently implemented with the Exhaustive Projection Pursuit (EPP) clustering approach which automatically detects the best two-dimensional gating strategy to identify the cellular sub-populations [19]. Automated detection of the gating strategy allows to analyze all cells in the dataset, without discarding cells from the analysis, as occurs when the gating strategy is user dependent.

4.2. Boolean Combination Gates

Boolean combination gates analyze all the possible combination of marker expressions, overcoming the issue of selecting pairs of parameters and the hierarchy that characterize manual analysis. This approach is fast, capable of considering all cells in a dataset, and it allows to compare different phenotypes across samples. The main limitations of Boolean gates regard the visualization of the results when the number of parameters increase and the difficulty of separating populations in the mono-dimensional gating step (Supplementary Figure S1).

A tool for performing Boolean combination gating is flowType, which is an R package available on Bioconductor repository, that automatically bisects cells in positive and negative for each analyzed marker [20], and it provides as output all possible phenotype combinations, including parent populations. Although the high number of possible phenotypes hampers the visualization of the results, this approach is particularly suitable for biomarker identification, since it explores the dataset considering all the possible marker combinations. To this aim, FloReMi and RchyOptimyx packages might be used to better interpret flowType results [21,22]. In addition to R packages, Boolean combination gates are also available as a FlowJo tool.

4.3. Multivariate Approach

Algorithms based on clustering, dimensionality reduction, and trajectory inference fully switch from the univariate/bivariate analysis to a multivariate approach. These tools consider the distribution of all markers simultaneously in the whole dataset, overcoming many of the manual gating limitations.

4.3.1. Clustering

Clustering based approaches identify and separate cells with similar marker profiles into cell clusters. This is the only multivariate approach which allows to quantify cell subsets in different samples and to perform comparative analysis between different experimental groups (e.g. stimulated samples versus control). The clustering tools can be classified on the basis of the kind of algorithm used for diving the cells into separate populations. SPADE (spanning tree progression of density normalized events) [54] and Citrus [26] are based on hierarchical clustering algorithms. The popular K-means, in which events are iteratively assigned to k clusters, is the algorithm on which flowMeans is based [23]. In PhenoGraph, cells are connected by weighted edges, where sets of highly interconnected cells represent phenotypically similar cell (or "communities") that can be partitioned in clusters using similar community-detection algorithms used for the analysis of social networks [35]. Algorithms such as flowClust [29], immunoClust [30], SWIFT [31], and HDPGMM [25] are model-based techniques and assume each cell type can be modelled as a multivariate statistical distribution. Other tools are built upon density-based algorithm, such as FLOCK (FLOw Clustering without K) [32], X-shift [28], flowPeaks [33], and ClusterX [34], in which more dense regions are identified and used as cluster centers.

An evaluation of the performance of automated gating techniques can be found in Weber and Robinson [55], where different algorithms are compared with manual gating. In the Weber and Robinson benchmark, FlowSOM [27] has emerged as one of the algorithms with highest performance for the automated identification of cell populations (measures as F-score with respect to manual gating results), being at the same time one of the fastest ones. FlowSOM has been recently included in FlowJo as a plugin.

Visualization of clustering results is an important step to appropriately interpret the results and many tools include visualization implementation. Histograms and dot plots are used to display marker distribution in a cluster comparing with a parental population. Other approaches aim to visualize inter-cluster relationship, showing clusters centroids, cluster median values or frequency of positive cells with minimum spanning three (MST), heatmaps, scaffold map, and dimensionality reduction techniques [27,54–57].

4.3.2. Dimensionality Reduction

Dimensionality reduction techniques aim to map high-dimensional data into a lower-dimensional space by losing as little information as possible. In the field of cytometry, dimensionality reduction is usually adopted to easily visualize the data, generally in two- or three-dimensional plots. These low-dimensional plots provide a straightforward visualization of the structure of multidimensional data, maintaining the information of data at a single-cell level, which instead is lost in clustering analyses. Principal component analysis (PCA) is a widely used method for reducing the dimensionality of multivariate data by linearly mapping the original variables into a low number of principal components (PCs). The resulting PCs represent a new set of variables oriented along the direction of maximum variance in the original dataset. Since PCA performs linear transformations to reduce dimensionality, it might be not optimal for reducing the number of dimensions in biological systems, where nonlinear relationships are common. This shortcoming might produce artefacts in low-dimensional space, with two points close in the low-dimensional space but not in the original multidimensional space. In cytometry, one of the most commonly used dimensionally reduction technique, that overcomes the limitation of linear transformations inherent in PCA, is the t-distributed stochastic neighbor embedding (t-SNE) algorithm [36], a tool available in R and in FlowJo. This method aims to map points from the high-dimensional space to the low-dimensional map by minimizing the difference in all pairwise similarities. Two of the most used tools for analyzing cytometric data, based on the t-SNE algorithm, are viSNE [38] and Automatic Classification of Cellular Expression by Nonlinear Stochastic Embedding (ACCENSE) [37].

The main drawback of t-SNE is the high computational cost of the algorithm, with the consequence that usually the low-dimensional maps are built using a limited number of cells obtained with a down-sampling of the original data. Moreover, it is important to remark that the algorithm includes a series of stochastic steps and consequently different analyses will give slightly different results. Recently, new dimensionality reduction tools such as EmbedSOM [39], diffusion maps [58,59], Fit-SNE [42], and UMAP (Uniform Manifold Approximation and Projection) [40] have been developed and applied to single-cell data to overcame t-SNE limitations. Dimensionality reduction is purely a visualization tool and does not allow the exact quantification of the identified population that requires a subsequent step. In flowJo, a manual gating analysis can be performed on dimensional-reduced t-SNE map, while other tools such as ACCENSE [37] perform an automated gating on t-SNE map.

4.3.3. Trajectory Inference

The last and most recently developed approach for analyzing single-cell dynamic processes are the trajectory inference methods. This approach aims to model the cell development and the transitions between different cell states by following marker expression gradients in the multi-dimensional data set.

Trajectory inference makes a step forward compared to clustering and dimensionality reduction algorithms, allowing at a single-cell level the unbiased study of cell processes such as the cell cycle, cell differentiation and cell activation. Starting from a mixture of different cells, these algorithms reconstruct the development stages that cells are following, ideally sorting the immature cells first, followed by the transitional stages, and finally the mature cells. With multivariate algorithms, and in particular trajectory inference methods, the multicolor panel design is crucial, to ensure that all relevant transitional states can be detected [60]. While different trajectory inference approaches have been developed for single-cell transcriptomic application [61–65], a limited number of methods have been

applied to cytometry analysis. Wanderlust detects linear transition, starting from a user-defined starting cell and subsequently ordering the rest of the cells [43], on the other hand, Wishbone, Monocle, and PHATE are able also to detect a bifurcation in the trajectory, enabling to characterize different cell lines that are difficult to identify within a linear model [44–46]. The interest in the trajectory inference field is growing rapidly, with a rising number of tools developed each month [63]. Their application in studying cell differentiation and development is currently limited to single-cell transcriptomics data, but it is likely that in the near future many of these tools could also be used to analyze data from flow and mass cytometry.

4.3.4. Multivariate Analysis Settings

Multivariate analysis bypasses the need to make choices that could influence the results, such as the definition of gates or the sequence of analyzed markers. However, the selection of optimal parameters by the operator still plays a key role, and sometimes the analysis has to be repeated multiple times in order to identify the best settings [66]. A crucial point in multivariate analysis is the number of target cell populations that is required as input parameter in many tools such as FlowMeans, FlowSOM, or SPADE. In the definition of the number of clusters, two conflicting requirements need to be taken into account. With a high number of clusters, the clusters are more homogenous, and it is more likely to identify rare populations, but visualization and interpretation of the results is highly complicated by over-fragmentation and noise. On the other hand, a low number of clusters makes visualization and interpretation easier, but it increases the likelihood of missing interesting cell populations. A common method for estimating the optimal number of clusters is the "elbow-method", in which the sum of the square distances of all the samples from the corresponding cluster centers (cost function) is plotted as a function of the number of clusters. Nevertheless, it is not always easy to identify the optimal number of clusters, therefore, it is advisable to set the number of clusters slightly higher than expected populations in order to ensure that the relevant cell types can be found [60].

Algorithms that do not directly require the number of clusters could still include parameters that affect the number of populations. For example, FLOCK, a density-based clustering algorithms, has two main parameters (the number of bins and the density threshold) that influence the estimated density and, indirectly, the number of resulting cell populations [32]. In t-SNE, perplexity is a parameter that influences the similarity measure. Roughly, with a low perplexity, the algorithm considers as similar only the nearest cell, resulting in an over-fragmentation of the populations; while, with a high perplexity, all cells are considered to have the same similarity, resulting in random distributed points on the map. Typical perplexity value are between 5 and 50 [36] and to get the most from t-SNE, it is recommended to analyze multiple plots with different perplexities. Another important choice is the selection of the parameters (markers) to include in the analysis. The so-called "curse of dimensionality" affects multivariate analyses when a high number of variables are considered in once. It was suggested that the curse of dimensionality could also affect multivariate analysis of cytometric data [67]. However, comparative studies by the FlowCAP project have shown that many of multivariate tools have reached a level of maturity that matches, or even surpasses, the results produced by human experts [6,55,68]. Nevertheless, it is recommendable to choose the more appropriate markers to include in the analysis in order to reduce dimensionality, complexity, and noise of datasets (e.g.: removing from the analysis markers that show only negative population).

5. Interpretation of the Results

The final phase in an automated analysis pipeline is the interpretation of the data-driven results. Generally, the cell populations identified have to be compared among different experimental groups. In manual analysis, statistical tests such as Mann–Whitney or Kruskal–Wallis are generally performed to identify populations with statistically relevant differences between experimental groups. When used in automated analysis, the use of multiple tests correction becomes necessary, such as Benjamini–Hochberg

or Bonferroni, since many statistical comparisons are performed, increasing the probability that a type I error (false positive error) occurs.

Correlation test or supervised Machine Learning methods, such as multivariate regression and classification, can also be used to identify a signatures that correlates with an external variable [6]. Multivariate regression is used to model an association with a continuous outcome variable, while classification methods can be used to identify links with a categorical clinical outcome of interest, such as a pathology. Once trained, these machine learning methods can be used to make predictions about new samples, where the output variables, being continuous or categorical, is unknown. Some packages, such as Cytrus and FlowSOM, includes possible statistical tests to be applied down-stream to the clustering analysis.

6. Impact of Automated Analysis in the Knowledge of Biological Processes

Increasing numbers of automated analyses of multidimensional cytometry data have been published in the last years, as reported in Figure 3a. The analysis has been performed using Web of Science, starting from the articles describing the automated tools reported in Table 1, then selecting all the citing articles (7613), and refining the search for "cytometry" (1018 articles). The time course analysis shows a rising trend of publication starting from 2008, the year of publication of t-SNE [36] and flowClust [29], up to date, with a stronger increase in the 2014–2019 period. In these years, three special issues of Cytometry Part A, the journal specialized in quantitative single-cell analysis by cytometry techniques, have been entirely dedicated to the computational analysis of flow cytometry data. Two of them were built around the "Flow Cytometry: Critical Assessment of Population Identification Methods (FlowCAP)" project [69], aimed at advancing the development of computational methods for the identification of cell populations of interest in flow cytometry data, under the direction of an open consortium of immunologists, bioinformaticians, statisticians, and clinical scientists [70,71]. The third one, "Machine learning for single cell data", is a special issue focused on the development and comparative analysis of machine learning methods and their application to single cell data, planned to be published in February 2020.

Articles reported in Table 1 are technical reports on software/methods development, where test datasets have been employed for evaluating their power or comparing the performance of available tools. The analysis shows that vi-SNE and ACCENSE, implementations of t-SNE, are the tools most cited for dimensionality reduction analysis, while Phenograph, SPADE, Citrus, FlowSOM, and X-shift for clustering approaches (Figure 3b). For the pre-processing step, FlowCore has the highest number of citations, since it offers essential function such as compensation and transformation of data, while new tools, such as FlowClean and FlowAI, recently published (2016) are aimed at data refinement, such as elimination of the outliers and anomalies during acquisition. Most of these tools are available in the R platform, and their use has been partially limited to bioinformaticians and researcher with programming expertise, even though they are available as open-access software. Indeed, analyzing the category of the journals selected for publication (Figure 3c), the majority (about 70%) are specialized in cytometry, methods and computer science, while only about 20% are in multidisciplinary and less than 10% immunology/life science journals.

An effort made to simplify the use of some automated tools has been the development of software with user-friendly interfaces an plug-ins capable of extending the functionality of the FlowJo software. This strategy has the advantages of combining the use of one of the most popular software for flow cytometry with automated analysis, thus helping the researchers to approach to the computational analysis of multiparametric data.

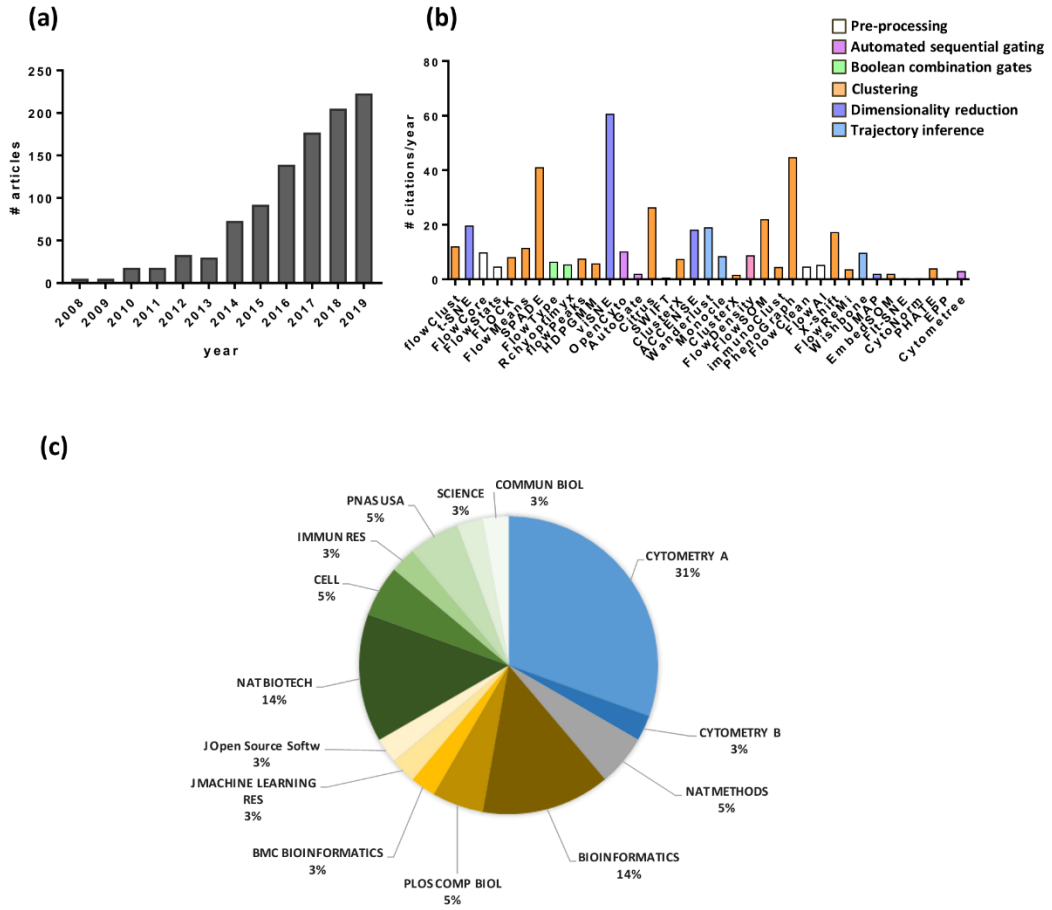

Figure 3. Analysis of the impact of automated tools on the scientific literature. (**a**) Numbers of publications per year, from 2008 to December 2019, on automated analyses of cytometry data. The query was performed with Web of Science starting from the articles reported in Table 1, selecting all the citing articles (7613), and refining the search for "cytometry" (1018 articles). (**b**) Numbers of citations each year of the reference articles of each automated tool reported in Table 1 (x axis); bars were colored according to the tool function. (**c**) Pie chart representation of the percentage of journals chosen for the publication of the reference article of each automated tool reported in Table 1.

Since comparative studies between manual versus automated analysis have demonstrated that many of the available tools can efficiently achieve the same results produced by manual gating [6,7], automated analysis have been applied, in recent years, for the identification of different cell populations in biomedical research and clinical diagnostic analysis [72]. One of the first computational approaches applied to clinical data analysis was conducted for identifying immunological correlations of HIV protection. Automated analysis was applied to a dataset derived from a large retrospective study of individuals at the early stage of HIV infection, and allowed to identify three T-cell subsets whose frequency during early infection had a statistically significant relationship with clinical progression to AIDS [20]. In the field of hematologic malignancies, successful application of computational methodologies have also been reported for acute lymphoblastic leukemia (AML) aimed at improving the discrimination between leukemic and normal cells [73,74], identifying B cell precursor as predictors of disease relapse [75], monitoring the minimal residual disease [76,77], evaluating the disease progression [78], or characterizing the immune alterations in AML patients [79]. Moreover, computational methods have been used on existing clinical flow cytometric data to improve diagnostic accuracy to distinguish mantle cell lymphoma from small lymphocytic lymphoma [80], or discriminate various subpopulations of blood cells in the context of B-chronic lymphocytic leukemia [81]. Complex data sets generated by multi-parametric flow cytometry have been analyzed with automatic tools

for characterizing myeloid and lymphoid cells in steady state [82–84], during the differentiation process [85,86], and in pathological conditions [87–93].

The automated analysis approach is therefore a powerful tool for unambiguous and unbiased characterization of cells, their subpopulations, functions, and roles in physiological and pathological conditions, applicable both in biomedical research and clinical diagnostic analysis [72,94].

7. Flow Cytometry in Vaccine Studies and the Advantages of Computational Analysis

Flow cytometry is a powerful technology for the characterization of multiple immune functions in response to vaccination, and both humoral and cellular components can be measured and characterized by flow cytometry-based assays. Multiparametric flow cytometry can be particularly suitable for the deep characterization of cellular immune responses, allowing to measure the phenotype and the functional features of rare cells, such as antigen-specific cells. The study of the CD4+ T cell activation and their effector function is fundamental in the characterization of immune responses to vaccination [95]. T helper cells are indeed closely related with long-term humoral immunity and modulate the functions of macrophages and CD8+ cytotoxic T cells through cytokines secretion, thus playing a central role in mediating vaccine immune responses [96]. Through flow cytometry, it is possible to directly and specifically identify antigen-specific T cells, using the major histocompatibility complex (MHC) tetramer staining technology [97,98], a procedure that has been used for characterizing antigen-specific T cell responses both in pre-clinical and clinical studies [99–101]. Furthermore, the combination of tetramer-staining with intracellular cytokine detection allows to assess, at single-cell level, the polyfunctional activity of antigen-specific T cells [102,103]. These procedures can be applied to better understand the complex functional profile of CD8+ and CD4+ T cell responses upon vaccination or infection.

Multiparametric flow cytometry can be particularly suitable also for characterizing polyclonal antibody responses elicited by vaccines, through a set of antibody-detection or cell-based functionality assays that can allow to identify humoral features that correlate with protection [104]. Antibody Fc-mediated mechanisms, such as cellular cytotoxicity, phagocytosis, direct pathogen killing, and modulation/stimulation of innate and adaptive immunity, can contribute, beyond neutralization, to confer protection against many pathogens. These mechanisms can be measured trough a range of different flow cytometry-based functional assays, that integrated with biophysical assays through Machine Learning methods, can contribute to profile the polyclonal antibody response and to identify immunological correlates and mechanism of humoral protection [104,105]. A complementary approach to the antibody response characterization is the study of the B cell response to vaccination, in which the production of plasma cells and memory B cells can be deeply analyzed by flow cytometry and its development and dissemination can be tracked between lymphoid organs and blood.

Automated analysis of cytometry data represents a powerful tool for the interpretation of large datasets in an unbiased way, that can be instrumental to profile the vaccine immune response. This analysis might unmask the detection of specific phenotypes/effector cells, that could be hardly detected with the manual analysis, and identify particular cell types (biomarkers) that can be specifically induced by tested vaccine formulations. Automated analysis has become particularly necessary as the size of marker panels has increased and consequently the number of cell populations identified by the combination of different markers has exponentially raised. Thanks to the computational approach, it is possible not only to identify cell populations according to the expression of two or three well-known specific surface markers, but also to distinguish different subsets within a population, based on combination of the other surface molecules expression. These subsets can be cells at intermediate stages of differentiation, or novel unexpected phenotypes. By applying the FlowSOM clustering approach different clusters of B cells elicited by immunization with a tuberculosis vaccine antigen combined with the liposome-based adjuvant CAF01 have been characterized [57]. Employing a computational approach, it was possible to identify many plasmablast subsets and different germinal center B cell subtypes. The clustering approach, followed by a statistical analysis between groups immunized with

or without the adjuvant component, has allowed us to identify a group of plasma cells as a specific biomarker of immunization with the adjuvanted-vaccine formulation [57]. Another clustering tool, FLOCK, was used for characterizing seventeen different B-cell subsets in human blood and to identify and quantify novel plasmablast subsets responding transiently to tetanus and other vaccinations (diphtheria toxoid, trivalent influenza vaccine 2009, H1N1 monovalent influenza vaccine, Hepatitis A, and Hepatitis B) [32].

Automated analysis can be particularly efficient also for identifying polyfunctional antigen-specific T cells elicited by vaccine administration or natural infection. Different computational tools, ranging from the Boolean combination gates, FlowSOM, or integrated approach combining targeted feature extraction (OpenCyto) with dimension reduction (t-SNE) have indeed been used to profile the polyfunctional activity of tuberculosis antigen-specific T cells and visualize treatment-specific differences between different vaccine formulations [106–108]. These studies demonstrate the importance of automated approaches to identify and visualize changes in very rare, multifunctional, antigen-specific T cells across different conditions, in flow cytometry datasets.

8. Conclusions

Automated analysis of cytometric data has widely been demonstrated to efficiently achieve reproducible results compared to manual analysis, with the important advantages of eliminating the bias toward expected populations, the subjectivity in manual drawing of gates and in marker selection, and most importantly the possibility to identify unexpected cell populations. The potentiality of the automated analysis of cytometry data ranges from the improvement of the data quality in the pre-processing steps up to the unbiased, data-driven examination of complex dataset using a variety of algorithms based on different approaches. Automated tools such as clustering algorithms or dimensionality reduction techniques fully switch from the bi-variate to a multi-variate analysis, overcoming most of the drawbacks that affect classical manual analysis, which are still partially present in automated sequential gating and Boolean combination gates. Moreover, combined approaches using more than one algorithm can further improve the automated analysis [109]. The development of automated tools addresses many needs associated with high-dimensional datasets, and the awareness of their potential is now expanding from computer scientists to immunologists/biologists, as demonstrated by the rising numbers of scientific publication in fields such as oncology and immunology, reported in recent years. Nevertheless, this process is still at the beginning, and efforts aimed at encouraging interdisciplinary cooperation, simplifying the graphical user interface of the computational tools, and training the next generation of flow cytometry experts are necessary to further increase the application of automated analysis to complex cytometry data. The use of automated tools can significantly contribute to the interpretation of cytometric data in a more reliable and efficient way, and to improve the knowledge of cellular populations, their function and roles in physiological and pathological conditions. Cellular profiles obtained with automated analysis of complex flow cytometry datasets can be integrated through a systems biology approach with the molecular profile achieved with the *omic* technologies, such as genomics, transcriptomics, proteomics, and metabolomics, together with clinical readouts, for better understanding the behavior of the immune system in response to antigenic challenges, such as vaccination or infection.

Supplementary Materials: The following are available online at http://www.mdpi.com/2076-393X/8/1/138/s1, Figure S1: Comparison between Boolean combination of 1D gates and 2D gates.

Author Contributions: S.L. and A.C. writing—original draft preparation; S.L., S.F., D.M. and A.C. writing—review and editing; D.M. funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by European Commission, grant number TRANSVAC2-730964.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Adan, A.; Alizada, G.; Kiraz, Y.; Baran, Y.; Nalbant, A. Flow cytometry: Basic principles and applications. *Crit. Rev. Biotechnol.* **2017**, *37*, 163–176. [CrossRef]
2. Nomura, L.; Maino, V.C.; Maecker, H.T. Standardization and optimization of multiparameter intracellular cytokine staining. *Cytom. Part A* **2008**, *73*, 984–991. [CrossRef] [PubMed]
3. Maecker, H.T.; McCoy, J.P.; Nussenblatt, R. Standardizing immunophenotyping for the Human Immunology Project. *Nat. Rev. Immunol.* **2012**, *12*, 191–200.
4. Gouttefangeas, C.; Chan, C.; Attig, S.; Køllgaard, T.T.; Rammensee, H.-G.; Stevanović, S.; Wernet, D.; thor Straten, P.; Welters, M.J.P.; Ottensmeier, C.; et al. Data analysis as a source of variability of the HLA-peptide multimer assay: From manual gating to automated recognition of cell clusters. *Cancer Immunol. Immunother.* **2015**, *64*, 585–598. [CrossRef] [PubMed]
5. Irish, J.M. Beyond the age of cellular discovery. *Nat. Immunol.* **2014**, *15*, 1095–1097. [CrossRef]
6. Aghaeepour, N.; Finak, G.; FlowCAP Consortium; DREAM Consortium; Hoos, H.; Mosmann, T.R.; Brinkman, R.; Gottardo, R.; Scheuermann, R.H. Critical assessment of automated flow cytometry data analysis techniques. *Nat. Methods* **2013**, *10*, 228–238. [CrossRef]
7. Conrad, V.K.; Dubay, C.J.; Malek, M.; Brinkman, R.R.; Koguchi, Y.; Redmond, W.L. Implementation and Validation of an Automated Flow Cytometry Analysis Pipeline for Human Immune Profiling. *Cytom. Part A* **2019**, *95*, 183–191. [CrossRef]
8. Huber, W.; Carey, V.J.; Gentleman, R.; Anders, S.; Carlson, M.; Carvalho, B.S.; Bravo, H.C.; Davis, S.; Gatto, L.; Girke, T.; et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods* **2015**, *12*, 115. [CrossRef]
9. Hahne, F.; LeMeur, N.; Brinkman, R.R.; Ellis, B.; Haaland, P.; Sarkar, D.; Spidlen, J.; Strain, E.; Gentleman, R. flowCore: A Bioconductor package for high throughput flow cytometry. *BMC Bioinform.* **2009**, *10*, 106. [CrossRef]
10. Hahne, F.; Khodabakhshi, A.H.; Bashashati, A.; Wong, C.-J.; Gascoyne, R.D.; Weng, A.P.; Seyfert-Margolis, V.; Bourcier, K.; Asare, A.; Lumley, T.; et al. Per-channel basis normalization methods for flow cytometry data. *Cytom. Part A* **2010**, *77*, 121–131. [CrossRef]
11. Fletez-Brant, K.; Špidlen, J.; Brinkman, R.R.; Roederer, M.; Chattopadhyay, P.K. flowClean: Automated identification and removal of fluorescence anomalies in flow cytometry data. *Cytom. Part A* **2016**, *89*, 461–471. [CrossRef] [PubMed]
12. Monaco, G.; Chen, H.; Poidinger, M.; Chen, J.; de Magalhães, J.P.; Larbi, A. flowAI: Automatic and interactive anomaly discerning tools for flow cytometry data. *Bioinformatics* **2016**, *32*, 2473–2480. [CrossRef] [PubMed]
13. Crowell, H.L.; Zanotelli, V.R.T.; Chevrier, S.; Robinson, M.D.; Bodenmiller, B. CATALYST: Cytometry dATa anALYSis Tools. Available online: https://bioconductor.org/packages/release/bioc/html/CATALYST.html (accessed on 19 March 2020).
14. Van Gassen, S.; Gaudilliere, B.; Angst, M.S.; Saeys, Y.; Aghaeepour, N. CytoNorm: A Normalization Algorithm for Cytometry Data. *Cytom. Part A* **2019**, *97*, 268–278. [CrossRef] [PubMed]
15. Malek, M.; Taghiyar, M.J.; Chong, L.; Finak, G.; Gottardo, R.; Brinkman, R.R. flowDensity: Reproducing manual gating of flow cytometry data by automated density-based cell population identification. *Bioinformatics* **2015**, *31*, 606–607. [CrossRef]
16. Finak, G.; Frelinger, J.; Jiang, W.; Newell, E.W.; Ramey, J.; Davis, M.M.; Kalams, S.A.; De Rosa, S.C.; Gottardo, R. OpenCyto: An open source infrastructure for scalable, robust, reproducible, and automated, end-to-end flow cytometry data analysis. *PLoS Comput. Biol.* **2014**, *10*, e1003806. [CrossRef]
17. Meehan, S.; Walther, G.; Moore, W.; Orlova, D.; Meehan, C.; Parks, D.; Ghosn, E.; Philips, M.; Mitsunaga, E.; Waters, J.; et al. AutoGate: Automating analysis of flow cytometry data. *Immunol. Res.* **2014**, *58*, 218–223. [CrossRef]
18. 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*, 1132–1140. [CrossRef]
19. Meehan, S.; Kolyagin, G.A.; Parks, D.; Youngyunpipatkul, J.; Herzenberg, L.A.; Walther, G.; Ghosn, E.E.B.; Orlova, D.Y. Automated subset identification and characterization pipeline for multidimensional flow and mass cytometry data clustering and visualization. *Commun. Biol.* **2019**, *2*, 229. [CrossRef]

20. Aghaeepour, N.; Chattopadhyay, P.K.; Ganesan, A.; O'Neill, K.; Zare, H.; Jalali, A.; Hoos, H.H.; Roederer, M.; Brinkman, R.R. Early immunologic correlates of HIV protection can be identified from computational analysis of complex multivariate T-cell flow cytometry assays. *Bioinformatics* **2012**, *28*, 1009–1016. [CrossRef]
21. Van Gassen, S.; Vens, C.; Dhaene, T.; Lambrecht, B.N.; Saeys, Y. FloReMi: Flow density survival regression using minimal feature redundancy. *Cytom. Part A* **2016**, *89*, 22–29. [CrossRef]
22. Aghaeepour, N.; Jalali, A.; O'Neill, K.; Chattopadhyay, P.K.; Roederer, M.; Hoos, H.H.; Brinkman, R.R. RchyOptimyx: Cellular hierarchy optimization for flow cytometry. *Cytom. Part A* **2012**, *81*, 1022–1030. [CrossRef] [PubMed]
23. Aghaeepour, N.; Nikolic, R.; Hoos, H.H.; Brinkman, R.R. Rapid cell population identification in flow cytometry data. *Cytom. Part A* **2011**, *79*, 6–13. [CrossRef] [PubMed]
24. Qiu, P.; Simonds, E.F.; Bendall, S.C.; Gibbs, K.D.; Bruggner, R.V.; Linderman, M.D.; Sachs, K.; Nolan, G.P.; Plevritis, S.K. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat. Biotechnol.* **2011**, *29*, 886–891. [CrossRef] [PubMed]
25. Cron, A.; Gouttefangeas, C.; Frelinger, J.; Lin, L.; Singh, S.K.; Britten, C.M.; Welters, M.J.P.; van der Burg, S.H.; West, M.; Chan, C. Hierarchical Modeling for Rare Event Detection and Cell Subset Alignment across Flow Cytometry Samples. *PLOS Comput. Biol.* **2013**, *9*, e1003130. [CrossRef] [PubMed]
26. Bruggner, R.V.; Bodenmiller, B.; Dill, D.L.; Tibshirani, R.J.; Nolan, G.P. Automated identification of stratifying signatures in cellular subpopulations. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, E2770–E2777. [CrossRef] [PubMed]
27. Van Gassen, S.; Callebaut, B.; Van Helden, M.J.; Lambrecht, B.N.; Demeester, P.; Dhaene, T.; Saeys, Y. FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytom. Part A* **2015**, *87*, 636–645. [CrossRef]
28. Samusik, N.; Good, Z.; Spitzer, M.H.; Davis, K.L.; Nolan, G.P. Automated mapping of phenotype space with single-cell data. *Nat. Methods* **2016**, *13*, 493–496. [CrossRef]
29. Lo, K.; Hahne, F.; Brinkman, R.R.; Gottardo, R. flowClust: A Bioconductor package for automated gating of flow cytometry data. *BMC Bioinform.* **2009**, *10*, 145. [CrossRef]
30. 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*, 603–615. [CrossRef]
31. Mosmann, T.R.; Naim, I.; Rebhahn, J.; Datta, S.; Cavenaugh, J.S.; Weaver, J.M.; Sharma, G. 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*, 422–433. [CrossRef]
32. Qian, Y.; Wei, C.; Eun-Hyung Lee, F.; Campbell, J.; Halliley, J.; Lee, J.A.; Cai, J.; Kong, Y.M.; Sadat, E.; Thomson, E.; 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. B Clin. Cytom.* **2010**, *78* (Suppl. 1), S69–S82. [CrossRef] [PubMed]
33. Ge, Y.; Sealfon, S.C. flowPeaks: A fast unsupervised clustering for flow cytometry data via K-means and density peak finding. *Bioinformatics* **2012**, *28*, 2052–2058. [CrossRef] [PubMed]
34. Rodriguez, A.; Laio, A. Clustering by fast search and find of density peaks. *Science* **2014**, *344*, 1492–1496. [CrossRef] [PubMed]
35. Levine, J.H.; Simonds, E.F.; Bendall, S.C.; Davis, K.L.; Amir, E.D.; Tadmor, M.D.; Litvin, O.; Fienberg, H.G.; Jager, A.; Zunder, E.R.; et al. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* **2015**, *162*, 184–197. [CrossRef]
36. van der Maaten, L.; Hinton, G. Viualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
37. Shekhar, K.; Brodin, P.; Davis, M.M.; Chakraborty, A.K. Automatic Classification of Cellular Expression by Nonlinear Stochastic Embedding (ACCENSE). *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 202–207. [CrossRef]
38. Amir, E.D.; Davis, K.L.; Tadmor, M.D.; Simonds, E.F.; Levine, J.H.; Bendall, S.C.; Shenfeld, D.K.; Krishnaswamy, S.; Nolan, G.P.; Pe'er, D. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.* **2013**, *31*, 545–552. [CrossRef]
39. Kratochvíl, M.; Koladiya, A.; Balounova, J.; Novosadova, V.; Fišer, K.; Sedlacek, R.; Vondrášek, J.; Drbal, K. Rapid single-cell cytometry data visualization with EmbedSOM. *bioRxiv* **2018**, 496869. [CrossRef]

40. McInnes, L.; Healy, J.; Saul, N.; Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *JOSS* **2018**, *3*, 861. [CrossRef]
41. Angerer, P.; Haghverdi, L.; Büttner, M.; Theis, F.J.; Marr, C.; Buettner, F. destiny: Diffusion maps for large-scale single-cell data in R. *Bioinformatics* **2016**, *32*, 1241–1243. [CrossRef]
42. Linderman, G.C.; Rachh, M.; Hoskins, J.G.; Steinerberger, S.; Kluger, Y. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nat. Methods* **2019**, *16*, 243–245. [CrossRef] [PubMed]
43. Bendall, S.C.; Davis, K.L.; Amir, E.D.; Tadmor, M.D.; Simonds, E.F.; Chen, T.J.; Shenfeld, D.K.; Nolan, G.P.; Pe'er, D. Single-Cell Trajectory Detection Uncovers Progression and Regulatory Coordination in Human B Cell Development. *Cell* **2014**, *157*, 714–725. [CrossRef] [PubMed]
44. Setty, M.; Tadmor, M.D.; Reich-Zeliger, S.; Angel, O.; Salame, T.M.; Kathail, P.; Choi, K.; Bendall, S.; Friedman, N.; Pe'er, D. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat. Biotechnol.* **2016**, *34*, 637–645. [CrossRef] [PubMed]
45. Trapnell, C.; Cacchiarelli, D.; Grimsby, J.; Pokharel, P.; Li, S.; Morse, M.; Lennon, N.J.; Livak, K.J.; Mikkelsen, T.S.; Rinn, J.L. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **2014**, *32*, 381–386. [CrossRef]
46. Moon, K.R.; van Dijk, D.; Wang, Z.; Gigante, S.; Burkhardt, D.B.; Chen, W.S.; Yim, K.; van den Elzen, A.; Hirn, M.J.; Coifman, R.R.; et al. Visualizing structure and transitions in high-dimensional biological data. *Nat. Biotechnol.* **2019**, *37*, 1482–1492. [CrossRef]
47. Bioconductor-Home. Available online: https://bioconductor.org/ (accessed on 19 March 2020).
48. FlowJo Exchange. Available online: https://www.flowjo.com/exchange/#/ (accessed on 19 March 2020).
49. Cytobank. Available online: https://www.cytobank.org/ (accessed on 19 March 2020).
50. ImmPort Shared Data. Available online: https://www.immport.org/shared/home (accessed on 19 March 2020).
51. Finak, G.; Perez, J.-M.; Weng, A.; Gottardo, R. Optimizing transformations for automated, high throughput analysis of flow cytometry data. *BMC Bioinform.* **2010**, *11*, 546. [CrossRef]
52. Finck, R.; Simonds, E.F.; Jager, A.; Krishnaswamy, S.; Sachs, K.; Fantl, W.; Pe'er, D.; Nolan, G.P.; Bendall, S.C. Normalization of mass cytometry data with bead standards. *Cytom. Part A* **2013**, *83*, 483–494. [CrossRef]
53. Zunder, E.R.; Finck, R.; Behbehani, G.K.; Amir, E.-A.D.; Krishnaswamy, S.; Gonzalez, V.D.; Lorang, C.G.; Bjornson, Z.; Spitzer, M.H.; Bodenmiller, B.; et al. Palladium-based mass tag cell barcoding with a doublet-filtering scheme and single-cell deconvolution algorithm. *Nat. Protoc.* **2015**, *10*, 316–333. [CrossRef]
54. Anchang, B.; Hart, T.D.P.; Bendall, S.C.; Qiu, P.; Bjornson, Z.; Linderman, M.; Nolan, G.P.; Plevritis, S.K. Visualization and cellular hierarchy inference of single-cell data using SPADE. *Nat. Protoc.* **2016**, *11*, 1264–1279. [CrossRef]
55. Weber, L.M.; Robinson, M.D. Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. *Cytom. Part A* **2016**, *89*, 1084–1096. [CrossRef]
56. Hsiao, C.; Liu, M.; Stanton, R.; McGee, M.; Qian, Y.; Scheuermann, R.H. Mapping cell populations in flow cytometry data for cross-sample comparison using the Friedman-Rafsky test statistic as a distance measure. *Cytom. Part A* **2016**, *89*, 71–88. [CrossRef] [PubMed]
57. Lucchesi, S.; Nolfi, E.; Pettini, E.; Pastore, G.; Fiorino, F.; Pozzi, G.; Medaglini, D.; Ciabattini, A. Computational Analysis of Multiparametric Flow Cytometric Data to Dissect B Cell Subsets in Vaccine Studies. *Cytom. Part A* **2019**, *97*, 259–267. [CrossRef] [PubMed]
58. Coifman, R.R.; Lafon, S.; Lee, A.B.; Maggioni, M.; Nadler, B.; Warner, F.; Zucker, S.W. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *PNAS* **2005**, *102*, 7426–7431. [CrossRef] [PubMed]
59. Haghverdi, L.; Buettner, F.; Theis, F.J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **2015**, *31*, 2989–2998. [CrossRef]
60. Saeys, Y.; Van Gassen, S.; Lambrecht, B.N. Computational flow cytometry: Helping to make sense of high-dimensional immunology data. *Nat. Rev. Immunol.* **2016**, *16*, 449–462. [CrossRef]
61. Cannoodt, R.; Saelens, W.; Saeys, Y. Computational methods for trajectory inference from single-cell transcriptomics. *Eur. J. Immunol.* **2016**, *46*, 2496–2506. [CrossRef]
62. Moon, K.R.; Stanley, J.S.; Burkhardt, D.; van Dijk, D.; Wolf, G.; Krishnaswamy, S. Manifold learning-based methods for analyzing single-cell RNA-sequencing data. *Curr. Opin. Syst. Opin.* **2018**, *7*, 36–46. [CrossRef]
63. Saelens, W.; Cannoodt, R.; Todorov, H.; Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **2019**, *37*, 547–554. [CrossRef]

64. Jin, S.; MacLean, A.L.; Peng, T.; Nie, Q. scEpath: Energy landscape-based inference of transition probabilities and cellular trajectories from single-cell transcriptomic data. *Bioinformatics* **2018**, *34*, 2077–2086. [CrossRef]
65. Shi, J.; Teschendorff, A.E.; Chen, W.; Chen, L.; Li, T. Quantifying Waddington's epigenetic landscape: A comparison of single-cell potency measures. *Brief. Bioinform.* **2018**, *21*, 248–261. [CrossRef]
66. Pedersen, C.B.; Olsen, L.R. Algorithmic Clustering Of Single-Cell Cytometry Data-How Unsupervised Are These Analyses Really? *Cytom. Part A* **2019**, *97*, 219–221. [CrossRef] [PubMed]
67. Orlova, D.Y.; Herzenberg, L.A.; Walther, G. Science not art: Statistically sound methods for identifying subsets in multi-dimensional flow and mass cytometry data sets. *Nat. Rev. Immunol.* **2018**, *18*, 77. [CrossRef] [PubMed]
68. Saeys, Y.; Van Gassen, S.; Lambrecht, B. Response to Orlova et al. "Science not art: Statistically sound methods for identifying subsets in multi-dimensional flow and mass cytometry data sets". *Nat. Rev. Immunol.* **2018**, *18*, 78. [CrossRef] [PubMed]
69. FlowCAP—Flow Cytometry: Critical Assessment of Population Identification Methods. Available online: http://flowcap.flowsite.org/ (accessed on 19 March 2020).
70. Brinkman, R.R.; Aghaeepour, N.; Finak, G.; Gottardo, R.; Mosmann, T.; Scheuermann, R.H. Automated analysis of flow cytometry data comes of age. *Cytom. Part A* **2016**, *89*, 13–15. [CrossRef]
71. Brinkman, R.R.; Aghaeepour, N.; Finak, G.; Gottardo, R.; Mosmann, T.; Scheuermann, R.H. State-of-the-Art in the Computational Analysis of Cytometry Data. *Cytom. Part A* **2015**, *87*, 591–593. [CrossRef]
72. Mittag, A.; Tarnok, A. Recent advances in cytometry applications: Preclinical, clinical, and cell biology. *Methods Cell Biol.* **2011**, *103*, 1–20.
73. Coustan-Smith, E.; Song, G.; Shurtleff, S.; Yeoh, A.E.-J.; Chng, W.J.; Chen, S.P.; Rubnitz, J.E.; Pui, C.-H.; Downing, J.R.; Campana, D. Universal monitoring of minimal residual disease in acute myeloid leukemia. *JCI Insight* **2018**, *3*, 98561. [CrossRef]
74. DiGiuseppe, J.A.; Tadmor, M.D.; Pe'er, D. Detection of minimal residual disease in B lymphoblastic leukemia using viSNE. *Cytom. B Clin. Cytom.* **2015**, *88*, 294–304. [CrossRef]
75. Good, Z.; Sarno, J.; Jager, A.; Samusik, N.; Aghaeepour, N.; Simonds, E.F.; White, L.; Lacayo, N.J.; Fantl, W.J.; Fazio, G.; et al. Single-cell developmental classification of B cell precursor acute lymphoblastic leukemia at diagnosis reveals predictors of relapse. *Nat. Med.* **2018**, *24*, 474–483. [CrossRef]
76. Reiter, M.; Diem, M.; Schumich, A.; Maurer-Granofszky, M.; Karawajew, L.; Rossi, J.G.; Ratei, R.; Groeneveld-Krentz, S.; Sajaroff, E.O.; Suhendra, S.; et al. Automated Flow Cytometric MRD Assessment in Childhood Acute B-Lymphoblastic Leukemia Using Supervised Machine Learning. *Cytom. Part A* **2019**, *95*, 966–975. [CrossRef]
77. Ko, B.-S.; Wang, Y.-F.; Li, J.-L.; Li, C.-C.; Weng, P.-F.; Hsu, S.-C.; Hou, H.-A.; Huang, H.-H.; Yao, M.; Lin, C.-T.; et al. Clinically validated machine learning algorithm for detecting residual diseases with multicolor flow cytometry analysis in acute myeloid leukemia and myelodysplastic syndrome. *EBioMedicine* **2018**, *37*, 91–100. [CrossRef] [PubMed]
78. Rajwa, B.; Wallace, P.K.; Griffiths, E.A.; Dundar, M. Automated Assessment of Disease Progression in Acute Myeloid Leukemia by Probabilistic Analysis of Flow Cytometry Data. *IEEE Trans. Biomed. Eng.* **2017**, *64*, 1089–1098. [CrossRef] [PubMed]
79. Chretien, A.-S.; Granjeaud, S.; Gondois-Rey, F.; Harbi, S.; Orlanducci, F.; Blaise, D.; Vey, N.; Arnoulet, C.; Fauriat, C.; Olive, D. Increased NK Cell Maturation in Patients with Acute Myeloid Leukemia. *Front. Immunol.* **2015**, *6*, 564. [CrossRef]
80. Zare, H.; Bashashati, A.; Kridel, R.; Aghaeepour, N.; Haffari, G.; Connors, J.M.; Gascoyne, R.D.; Gupta, A.; Brinkman, R.R.; Weng, A.P. Automated analysis of multidimensional flow cytometry data improves diagnostic accuracy between mantle cell lymphoma and small lymphocytic lymphoma. *Am. J. Clin. Pathol.* **2012**, *137*, 75–85. [CrossRef] [PubMed]
81. Lakoumentas, J.; Drakos, J.; Karakantza, M.; Nikiforidis, G.C.; Sakellaropoulos, G.C. Bayesian clustering of flow cytometry data for the diagnosis of B-chronic lymphocytic leukemia. *J. Biomed. Inform.* **2009**, *42*, 251–261. [CrossRef]
82. Becher, B.; Schlitzer, A.; Chen, J.; Mair, F.; Sumatoh, H.R.; Teng, K.W.W.; Low, D.; Ruedl, C.; Riccardi-Castagnoli, P.; Poidinger, M.; et al. High-dimensional analysis of the murine myeloid cell system. *Nat. Immunol.* **2014**, *15*, 1181–1189. [CrossRef]

83. Wong, M.T.; Chen, J.; Narayanan, S.; Lin, W.; Anicete, R.; Kiaang, H.T.K.; De Lafaille, M.A.C.; Poidinger, M.; Newell, E.W. Mapping the Diversity of Follicular Helper T Cells in Human Blood and Tonsils Using High-Dimensional Mass Cytometry Analysis. *Cell Rep.* **2015**, *11*, 1822–1833. [CrossRef]
84. Hu, X.; Kim, H.; Brennan, P.J.; Han, B.; Baecher-Allan, C.M.; De Jager, P.L.; Brenner, M.B.; Raychaudhuri, S. Application of user-guided automated cytometric data analysis to large-scale immunoprofiling of invariant natural killer T cells. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 19030–19035. [CrossRef]
85. Lin, J.; Kim, D.; Tse, H.T.; Tseng, P.; Peng, L.; Dhar, M.; Karumbayaram, S.; Di Carlo, D. High-throughput physical phenotyping of cell differentiation. *Microsyst. Nanoeng.* **2017**, *3*, 17013. [CrossRef]
86. Li, N.; van Unen, V.; Abdelaal, T.; Guo, N.; Kasatskaya, S.A.; Ladell, K.; McLaren, J.E.; Egorov, E.S.; Izraelson, M.; de Sousa Lopes, S.M.C.; et al. Memory CD4+ T cells are generated in the human fetal intestine. *Nat. Immunol.* **2019**, *20*, 301–312. [CrossRef]
87. Liu, M.; Barton, E.S.; Jennings, R.N.; Oldenburg, D.G.; Whirry, J.M.; White, D.W.; Grayson, J.M. Unsupervised learning techniques reveal heterogeneity in memory CD8+ T cell differentiation following acute, chronic and latent viral infections. *Virology* **2017**, *509*, 266–279. [CrossRef] [PubMed]
88. Barcenilla, H.; Åkerman, L.; Pihl, M.; Ludvigsson, J.; Casas, R. Mass Cytometry Identifies Distinct Subsets of Regulatory T Cells and Natural Killer Cells Associated With High Risk for Type 1 Diabetes. *Front. Immunol.* **2019**, *10*, 982. [CrossRef] [PubMed]
89. Emmaneel, A.; Bogaert, D.J.; Van Gassen, S.; Tavernier, S.J.; Dullaers, M.; Haerynck, F.; Saeys, Y. A Computational Pipeline for the Diagnosis of CVID Patients. *Front. Immunol.* **2019**, *10*, 2009. [CrossRef] [PubMed]
90. Mukherjee, R.; Barman, P.K.; Thatoi, P.K.; Tripathy, R.; Kumar Das, B.; Ravindran, B. Non-Classical monocytes display inflammatory features: Validation in Sepsis and Systemic Lupus Erythematous. *Sci. Rep.* **2015**, *5*, 13886. [CrossRef]
91. Lacombe, F.; Lechevalier, N.; Vial, J.P.; Béné, M.C. An R-Derived FlowSOM Process to Analyze Unsupervised Clustering of Normal and Malignant Human Bone Marrow Classical Flow Cytometry Data. *Cytom. Part A* **2019**, *95*, 1191–1197. [CrossRef]
92. Coindre, S.; Tchitchek, N.; Alaoui, L.; Vaslin, B.; Bourgeois, C.; Goujard, C.; Lecuroux, C.; Bruhns, P.; Le Grand, R.; Beignon, A.-S.; et al. Mass Cytometry Analysis Reveals Complex Cell-State Modifications of Blood Myeloid Cells During HIV Infection. *Front. Immunol.* **2019**, *10*, 2677. [CrossRef]
93. Leite Pereira, A.; Bitoun, S.; Paoletti, A.; Nocturne, G.; Marcos Lopez, E.; Cosma, A.; Le Grand, R.; Mariette, X.; Tchitchek, N. Characterization of Phenotypes and Functional Activities of Leukocytes From Rheumatoid Arthritis Patients by Mass Cytometry. *Front. Immunol.* **2019**, *10*, 2384. [CrossRef]
94. Duetz, C.; Bachas, C.; Westers, T.M.; van de Loosdrecht, A.A. Computational analysis of flow cytometry data in hematological malignancies: Future clinical practice? *Curr. Opin. Oncol.* **2020**, *32*, 162–169. [CrossRef]
95. Ciabattini, A.; Pettini, E.; Medaglini, D. CD4(+) T Cell Priming as Biomarker to Study Immune Response to Preventive Vaccines. *Front. Immunol.* **2013**, *4*, 421. [CrossRef]
96. Jelley-Gibbs, D.M.; Strutt, T.M.; McKinstry, K.K.; Swain, S.L. Influencing the fates of CD4 T cells on the path to memory: Lessons from influenza. *Immunol. Cell Biol.* **2008**, *86*, 343–352. [CrossRef]
97. Altman, J.D.; Moss, P.A.; Goulder, P.J.; Barouch, D.H.; McHeyzer-Williams, M.G.; Bell, J.I.; McMichael, A.J.; Davis, M.M. Phenotypic analysis of antigen-specific T lymphocytes. *Science* **1996**, *274*, 94–96. [CrossRef] [PubMed]
98. Moon, J.J.; Chu, H.H.; Pepper, M.; McSorley, S.J.; Jameson, S.C.; Kedl, R.M.; Jenkins, M.K. Naive CD4(+) T cell frequency varies for different epitopes and predicts repertoire diversity and response magnitude. *Immunity* **2007**, *27*, 203–213. [CrossRef] [PubMed]
99. Prota, G.; Christensen, D.; Andersen, P.; Medaglini, D.; Ciabattini, A. Peptide-specific T helper cells identified by MHC class II tetramers differentiate into several subtypes upon immunization with CAF01 adjuvanted H56 tuberculosis vaccine formulation. *Vaccine* **2015**, *33*, 6823–6830. [CrossRef] [PubMed]
100. Ciabattini, A.; Pettini, E.; Fiorino, F.; Pastore, G.; Andersen, P.; Pozzi, G.; Medaglini, D. Modulation of Primary Immune Response by Different Vaccine Adjuvants. *Front. Immunol.* **2016**, *7*, 427. [CrossRef]
101. Uchtenhagen, H.; Rims, C.; Blahnik, G.; Chow, I.-T.; Kwok, W.W.; Buckner, J.H.; James, E.A. Efficient ex vivo analysis of CD4+ T-cell responses using combinatorial HLA class II tetramer staining. *Nat. Commun.* **2016**, *7*, 1–12. [CrossRef]

102. Pastore, G.; Carraro, M.; Pettini, E.; Nolfi, E.; Medaglini, D.; Ciabattini, A. Optimized Protocol for the Detection of Multifunctional Epitope-Specific CD4+ T Cells Combining MHC-II Tetramer and Intracellular Cytokine Staining Technologies. *Front. Immunol.* **2019**, *10*, 2304. [CrossRef]
103. Tesfa, L.; Volk, H.D.; Kern, F. A protocol for combining proliferation, tetramer staining and intracellular cytokine detection for the flow-cytometric analysis of antigen specific T-cells. *J. Biol. Regul. Homeost. Agents* **2003**, *17*, 366–370.
104. Chung, A.W.; Kumar, M.P.; Arnold, K.B.; Yu, W.H.; Schoen, M.K.; Dunphy, L.J.; Suscovich, T.J.; Frahm, N.; Linde, C.; Mahan, A.E.; et al. Dissecting Polyclonal Vaccine-Induced Humoral Immunity against HIV Using Systems Serology. *Cell* **2015**, *163*, 988–998. [CrossRef]
105. Kimble, J.B.; Malherbe, D.C.; Meyer, M.; Gunn, B.M.; Karim, M.M.; Ilinykh, P.A.; Iampietro, M.; Mohamed, K.S.; Negi, S.; Gilchuk, P.; et al. Antibody-Mediated Protective Mechanisms Induced by a Trivalent Parainfluenza Virus-Vectored Ebolavirus Vaccine. *J. Virol.* **2019**, *93*, e01845-18. [CrossRef]
106. Lin, L.; Frelinger, J.; Jiang, W.; Finak, G.; Seshadri, C.; Bart, P.-A.; Pantaleo, G.; McElrath, J.; DeRosa, S.; Gottardo, R. Identification and visualization of multidimensional antigen-specific T-cell populations in polychromatic cytometry data. *Cytom. Part A* **2015**, *87*, 675–682. [CrossRef]
107. Ciabattini, A.; Pettini, E.; Fiorino, F.; Lucchesi, S.; Pastore, G.; Brunetti, J.; Santoro, F.; Andersen, P.; Bracci, L.; Pozzi, G.; et al. Heterologous Prime-Boost Combinations Highlight the Crucial Role of Adjuvant in Priming the Immune System. *Front. Immunol.* **2018**, *9*, 380. [CrossRef] [PubMed]
108. Billeskov, R.; Wang, Y.; Solaymani-Mohammadi, S.; Frey, B.; Kulkarni, S.; Andersen, P.; Agger, E.M.; Sui, Y.; Berzofsky, J.A. Low Antigen Dose in Adjuvant-Based Vaccination Selectively Induces CD4 T Cells with Enhanced Functional Avidity and Protective Efficacy. *J. Immunol.* **2017**, *198*, 3494–3506. [CrossRef] [PubMed]
109. Kvistborg, P.; Gouttefangeas, C.; Aghaeepour, N.; Cazaly, A.; Chattopadhyay, P.K.; Chan, C.; Eckl, J.; Finak, G.; Hadrup, S.R.; Maecker, H.T.; et al. Thinking outside the gate: Single-cell assessments in multiple dimensions. *Immunity* **2015**, *42*, 591–592. [CrossRef] [PubMed]