Journal of Immunological Methods 475 (2019) 112631

Contents lists available at ScienceDirect

Journal of Immunological Methods

journal homepage: www.elsevier.com/locate/jim

Review

From big flow cytometry datasets to smart diagnostic strategies: The EuroFlow approach

C.E. Pedreira[a], E. Sobral da Costa[b], Q. Lecrevise[c], G. Grigore[d], R. Fluxa[d], J. Verde[d], J. Hernandez[d], J.J.M. van Dongen[e,**,1], A. Orfao[c,*,1], on behalf of EuroFlow

[a] Systems and Computing Department (PESC), COPPE, Federal University of Rio de Janeiro (UFRJ), Brazil
[b] School of Medicine, Federal University of Rio de Janeiro (UFRJ), Brazil
[c] Cancer Research Centre (IBMCC, USAL-CSIC), Department of Medicine and Cytometry Service (NUCLEUS), IBSAL and CIBERONC, University of Salamanca, Spain
[d] Cytognos SL, Salamanca, Spain
[e] Dept. of Immunohematology and Blood Transfusion (IHB), Leiden University Medical Center (LUMC), Leiden, the Netherlands

ABSTRACT

The rise in the analytical speed of mutiparameter flow cytometers made possible by the introduction of digital instruments, has brought up the possibility to manage progressively higher number of parameters simultaneously on significantly greater numbers of individual cells. This has led to an exponential increase in the complexity and volume of flow cytometry data generated about cells present in individual samples evaluated in a single measurement. This increase demands for new developments in flow cytometry data analysis, graphical representation, and visualization and interpretation tools to address the new big data challenges, i.e. processing data files of ≥10–25 parameters per cell in samples with > 5–10 million cells (= up to 250 million data points per cell sample) obtained in a few minutes.

Here, we present a comprehensive review of some of the tools developed by the EuroFlow consortium for processing flow cytometric big data files in diagnostic laboratories, particularly focused on automated EuroFlow approaches for: i) identification of all cell populations coexisting in a sample (automated gating); ii) smart classification of aberrant cell populations in routine diagnostics; iii) automated reporting; together with iv) new tools developed to visualize n-dimensional data in 2-dimensional plots to support expert-guided automated data analysis. The concept of using reference data bases implemented into software programs, in combination with multivariate statistical analysis pioneered by EuroFlow, provides an innovative, highly efficient and fast approach for diagnostic screening, classification and monitoring of patients with distinct hematological and immune disorders, as well as other diseases.

1. Introduction

In the last decades, parallel developments in multiple technologies devoted to high-throughput analysis of biological samples, (e.g. next-generation sequencing, gene expression profiling, proteomics, metabolomics and flow cytometry), together with the advances in computer hardware and computational tools, have placed biomedical research and laboratory diagnostics in the *big data* arena. Thereby, a lot of attention has been devoted to the potential utility of *big data* (analytical) approaches for fulfilling the increasingly high demand for innovative analytical solutions raised by the enormous complexity and volume of data generated in biology and medicine (Marx, 2013; Finak et al., 2014, Schultze, 2015). Among other sources of *big data*, flow cytometry is also included.

Early flow cytometry immunophenotypic studies were based on single marker stainings for evaluation of a few thousand cells. With the advent of ≥8-color flow cytometry devices in the past decade, an enormous increase in the complexity and amount of data generated about normal, reactive and malignant cells has occurred. Three main factors played a critical role in this rapid increase in data collection: i) the development of instrumentation progressively capable of higher number of parameters that can be simultaneously assessed for individual cells, ii) the availability of a greater number of fluorochromes compatible with these instruments and, iii) the higher speed of analysis brought up by digital (vs. analogic) instruments, which allows evaluation of several tens of thousands cells per second (Freer and Rindi, 2013; Martini et al., 2012). This exponential growth in the complexity of data generated about larger numbers of (i.e. tens of millions) individual cells (and their products), has pointed out the need to accelerate the incorporation of new developments in both data analysis and representation-visualization tools, and to re-evaluate and modify the conventional 1–2-dimensional-based flow cytometry data analysis procedures (Robinson et al., 2012; Pedreira et al., 2013; Orfao et al., 1999; Wood, 2016; Comans-Bitter et al., 1997; Chester and Maecker,

* Corresponding author at: Cancer Research Center, Paseo de la Universidad de Coimbra s/n, 37007 Salamanca, Spain.

** Corresponding author at: Leiden University Medical Center (LUMC), LUMC Main Building, Albinusdreef 2, 2333 ZA, Leiden, the Netherlands.

E-mail addresses: j.j.m.van_dongen@lumc.nl (J.J.M. van Dongen), orfao@usal.es (A. Orfao).

[1] Both authors should be considered as last author and both are correspondent authors.

https://doi.org/10.1016/j.jim.2019.07.003
Received 28 March 2019; Received in revised form 1 July 2019; Accepted 10 July 2019
Available online 12 July 2019

2015). Therefore, selection, validation and implementation (or even development) of novel multivariate analytical approaches and algorithms capable to mine large amounts of high-dimensional flow cytometry data using smart strategies and innovative tools, in a fast, robust, standardized and reproducible way, has become a major challenge (Robinson et al., 2012; Pedreira et al., 2013; Comans-Bitter et al., 1997; Chester and Maecker, 2015). Despite the great potential of such multivariate flow cytometry data analysis approaches vs individual human expertise, intuitive visualization of data in a user-friendly way, as well as efficient data storage and retrieval, are also key to allow access of experts to high-dimensional spaces where single data points (e.g. cells, cell populations, patients) fall, and facilitate interactive expert-based control (i.e. expert guidance and intervention) of the analytical process (Pedreira et al., 2013).

Due to its intrinsic analytical possibilities, multiparameter flow cytometry is currently considered an invaluable tool for both research and routine diagnostic purposes. Thus, ≥8-color flow cytometers are currently available in virtually every diagnostic and research laboratory around the world, particularly for evaluation of normal and/or malignant hematopoietic and lymphoid cells (Orfao et al., 1999; Wood, 2016; Comans-Bitter et al., 1997). Actual expansion of multiparameter flow cytometry to the field of immunology and hematology has been strongly facilitated by the fact that blood cells, including immune cells, typically circulate as single cells in our body, in the blood stream (and other body fluids). This makes it possible to analyze millions of single blood/immune cells via minimally-invasive diagnostic procedures (Robinson et al., 2012; Chester and Maecker, 2015; Mair et al., 2016).

Based on the identified data processing and analysis needs of clinical flow cytometry, the EuroFlow consortium has developed, validated and implemented, novel analytical tools in combination with smart diagnostic strategies, for processing *clinical flow cytometry big datasets* in routine diagnostics in the fields of leukemia and lymphoma and primary immune deficiencies of the lymphoid system, as reviewed below (Kalina et al., 2012).

2. Toward new ways to analyze and interpret flow cytometry data

For decades, conventional multiparameter flow cytometry data analysis procedures have been used to identify unique cell populations within a sample and characterize them, by defining gates in single parameter and bi-dimensional (parameter X vs parameter Y) plots, as already proposed in the early 1960's (Orfao et al., 1999; Bonner et al., 1972; Hulett et al., 1969). To solve problems related to the greater dimensionality of data when ≥10 parameters are simultaneously measured for single cells, sequential "*Boolean*" gating strategies (mainly) based on 2-D plots, were universally adopted for the identification of one or multiple cell populations coexisting in a sample, typically followed by visualization of the immunophenotypic profiles and relative distribution in the sample of the gated events in e.g. single parameter plots with descriptive statistics about them (Hunter et al., 1994). Whenever the number of markers in a panel required to stain a given sample was higher than the upper limit of the multicolor capabilities of the available flow cytometer instrument, the whole set of markers to be analyzed had to be investigated in two distinct aliquots of the same sample. In such case, data analysis is based in separate evaluation of the different data files generated about each aliquot of the sample, and the relationship between the phenotype of cells for markers evaluated in different aliquots of the same sample, fully relies on the interpretation of an expert.

Between 1985 and 2005, most available clinical flow cytometers could only evaluate (simultaneously) 5–7 parameters per cell (Pedreira et al., 2013; Orfao et al., 1999) for a few hundred thousand cells. During this period, the number of markers required to compose a panel progressively increased to around 10–20 antibodies. Because of this, interpretation of progressively more complex, but objective, data derived from such quantifiable flow cytometry measurements, has gradually developed into a relatively subjective expert-based interpretation of "FCM images or pictures" (i.e. histograms and bivariate dot plots), in a similar way to what is done in conventional cytology and histopathology (e.g. interpretation of microscopic "images or pictures" of single cell smears, cytospins or histological sections). These expert-based data analysis and interpretation strategies necessarily demanded for highly-experienced and well-trained experts to select for the (right) cell population(s) of interest, at the expense of slower and/or less reproducible analyses; consequently, in many centers this has hampered the quality of the results obtained, while frequently limiting also the amount of data analyzed to (only) a small fraction of the total data generated, e.g. a few cell subsets from all cell populations in the sample (Pedreira et al., 2013).

Further parallel increase in multiparameter capabilities of digital flow cytometers, together with the number of distinct markers evaluated per antibody panel, emphasized the urgent need for novel data analyses strategies capable of i) extracting data from the new n-dimensional (n-D) spaces, ii) representing such multidimensional data in 1–2-dimensional graphics for expert intuitive data visualization, and iii) developing user-friendly interpretation-guided tools of both raw and processed big datasets (Robinson et al., 2012; Pedreira et al., 2013). Early introduction of multivariate flow cytometry data analysis algorithms, and their corresponding 2-D graphical representations, has rapidly highlighted the relevance of standardization of multiparameter flow cytometry procedures, to be able to take advantage of all potential benefits of automated pattern-guided data analysis approaches (Pedreira et al., 2013; Kalina et al., 2012; Costa et al., 2006; Costa et al., 2010). Since the earliest contributions to automated multiparameter clinical flow cytometry data analysis (Costa et al., 2006), an increased number of computational methods have been reported that might potentially overcome the limitations of conventional expert-based (manual) flow cytometry data analysis procedures (Costa et al., 2010; Quinn et al., 2007; Aghaeepour et al., 2011; Lo et al., 2008; Zare et al., 2012; Roederer et al., 2011; Qiu et al., 2011; Finak et al., 2016), as summarized elsewhere (Aghaeepour et al., 2013). Therefore, development, implementation and availability of new, up-to-date software tools for improved multivariate analysis and interpretation of complex multiparameter flow cytometry data, has become a priority. Many of the novel software tools and computational algorithms that are devoted to analysis and processing of multiparameter flow cytometry big data, have been implemented in open source software packages for easy access (Kalina et al., 2012; Qiu et al., 2011; Malek et al., 2015; van der Maaten and Hinton, 2008). Nevertheless, in many of them, user-friendly graphical interfaces are not provided, while several require the user to have some (basic) knowledge about software programming languages such as R, Python, Java or Matlab (Aghaeepour et al., 2011), in the absence of structured support to the user; in addition, none of these open source flow cytometry software tools has been (or is being) cleared for in vitro diagnostics (IVD). Altogether, these limitations become actual barriers for extended adoption and use of these software tools in most diagnostic laboratories around the world.

In parallel to the above initiatives, the EuroFlow consortium has pioneered the development of new software tools and strategies, as well as novel graphical representations for analysis and visualization of flow cytometry big data about single cells, beads and other types of events, which can be of great utility in both clinical and research flow cytometry laboratories. Such tools have been progressively implemented in the Infinicyt software (Cytognos SL, Salamanca, Spain). Altogether, this has paved the way for the introduction of (standardized) big data analytical tools to clinical flow cytometry in routine diagnostics (Flores-Montero et al., 2017; van der Burg et al., 2019; Lhermitte et al., 2018).

3. Automatic identification of cell populations

Identification of cell populations based on gating of flow cytometry data remains a basic and critical step in multiparameter flow cytometry data analysis in both diagnostic and research laboratories. Initially, gating was based on the establishment of a rectangular-shape region in a bi-dimensional dot plot. Subsequently, so called "Boolean" gating strategies were introduced, aiming at identifying a cell population based on a combination (e.g. the intersection) of multiple sequential gates established in a series of distinct single parameter histograms and/or, particularly 2-D dot plots (Hunter et al., 1994). Based on these gating approaches, a considerable large number of gates (or regions) is required for optimal identification of a given cell population in a sample that consists of multiple heterogeneous cell types and their subsets (Pedreira et al., 2013; Mair et al., 2016), as apart from selecting the cell population of interest, cell doublets and debris have to be discarded.

Early advances in software tools allowed to establish gates of different shapes, adapted to the distribution of the cell population being analyzed, in the most informative combination of (e.g. 2-D) dot plots. This has led to the introduction of elliptical, round-shaped and polygonal (i.e. very flexible) gates and gate combinations, in most available proprietary and open source flow cytometry software packages. However, this flexibility has further increased significantly the complexity of gating procedures, with a negative impact in reproducible gating (Sutherland et al., 1996). Such complexity has further extended significantly, when ≥8–color digital flow cytometers became available. This was due to the: i) greater number of dimensions (parameters) per cellular event; ii) the increased number of cells evaluated per single measurement, and iii) the higher number of cell populations identifiable in a single stain. Consequently, an exponential increase in time required for gating during flow cytometry data analysis, occurred. Of note, the specific gates and gating strategies used in this period still relied (mostly) on expert-based decisions, being thereby, associated with a significant component of individual subjectivity (Pedreira et al., 2013).

Once this problem had been recognized, the EuroFlow consortium started to search for innovative data analysis solutions, for faster and standardized (simultaneous) identification of all cell populations coexisting in a sample stained with an informative antibody combination. Since then, interactive automated gating approaches (Fig. 1) have been developed by EuroFlow, which rely on comparison of a given data file against flow cytometry reference big data bases composed of large sets of normal, reactive and/or patient data files, from matched samples stained with the same overlapping antibody combination and sample preparation and data acquisition SOPs. Overall, the new automated gating approaches developed by EuroFlow take advantage of a combination of previously developed clustering algorithms (Fig. 1, panels B1 and B2), to which a subsequent cluster classification step is added, to group the clusters of events into cell populations with a biological and/or clinical meaning (Fig. 1, panels C1 and C2). This classification step is reached via comparison of the immunophenotypic features of individual clusters of events against pre-existing, well-defined and standardized big data bases of matched normal and pathologic samples. Briefly, in a first step, all individual events contained in a given flow cytometry data file (Fig. 1 panels A1 and A2) are clustered (i.e. classified using unsupervised clustering algorithms). This is done by applying an agglomerative clustering approach starting from the less to the more dense regions in the multidimensional space. Thus, one generates, hundreds to thousands of different groups of similar events (clusters of events) which might have (or not have) biological and/or clinical significance, without (human) subjectivity (Fig. 1, panels B1 and B2). This unsupervised approach might eventually reveal populations of cells that do not appear in the training datasets. Although possible, a supervised scheme would limit identification of such cell populations present in the dataset. In a second step, each of the clusters (groups) of events is compared against the reference data base (Fig. 1, panels C1 and C2). Such data base comparisons open up three possibilities for each specific cluster of events: (i) the features of the cluster of events fully match (coincide) with those of given cell population in the data base, (ii) they might be similar to it, or (iii) they do not match at all the features of any of the cell populations represented in the reference data base. Whenever the interrogated cluster of events fully matches the features of a cell population contained in the data base, it might be classified as corresponding to that specific cell population; in turn, if despite being very similar to a given cell population, it does not fully match its features, then it is classified as potentially belonging to that specific cell population with the need for a further expert re-evaluation and decision about the cell population that cluster of events should be finally classified into. Finally, in case a cluster of events does

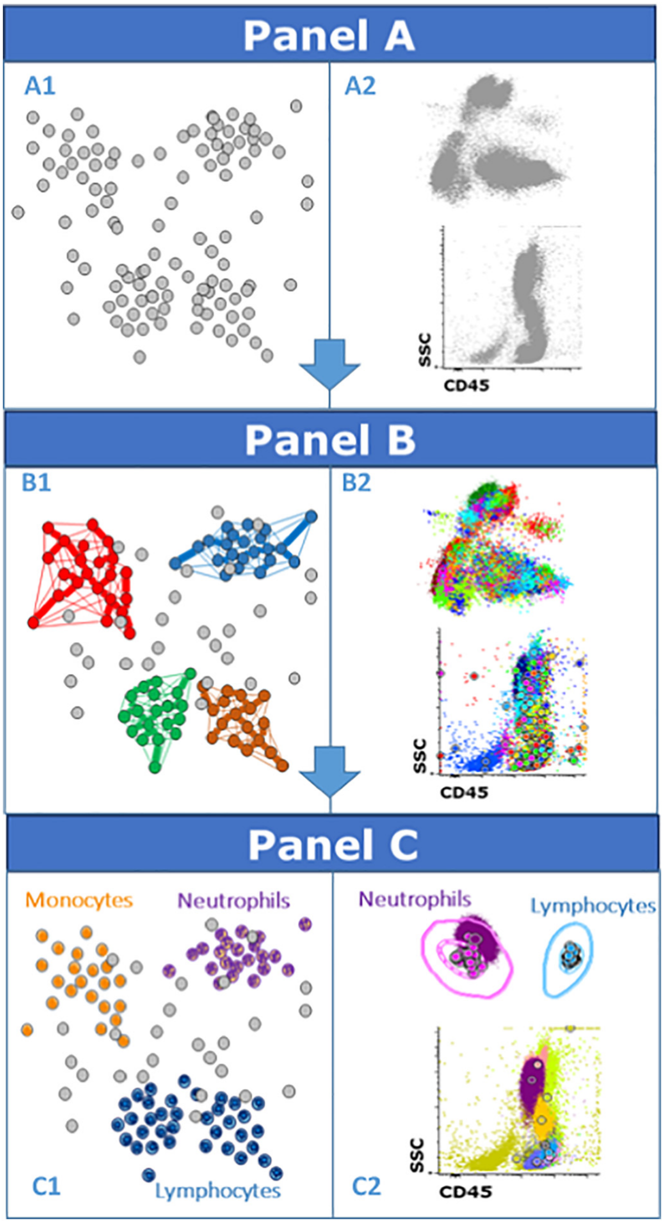

Fig. 1. Schematic representation of the automated gating and cell identification approach proposed by EuroFlow for analysis of flow cytometry data files. Unsupervised clustering analysis performed on raw flow cytometry data (panel A), leads to the identification of multiple clusters of events (panel B); through the comparison of each individual cluster of events (supervised) against a predefined data base, further classifies each of the clusters of events from panel B, into a specific cell population that fully matches the clusters' phenotypic features (panel C). In the left column (A1, B1 and C1), a schematic representation is shown, in which each point represents a cluster. Actual Flow cytometry data shown in the right column (A2, B2 and C2) in which each small point represents a cell event.

not match (i.e. significantly differs from) any of the cell populations represented in the database, it remains as an unclassified group of events (e.g. for further expert evaluation). Therefore, all groups of events obtained in a first step are classified in this second (data base comparison) step into: i) a well-defined cell population, debris or doublets, ii) a population of cells, debris or doublets that needs to be checked by an expert, or iii) an unclassifiable cluster of events for further expert evaluation. The results of the above automated gating steps are schematically illustrated in Fig. 1 (panels C1 and C2).

Based on the automated gating process described above, it should be emphasized here that, reference data bases must consist of representative sets of flow cytometry data files from healthy controls and/or patients, matched for the same type of sample, stained with exactly the same (or fully equivalent) multicolor antibody combination. To be considered as -fully equivalent-, a combination of antibodies should consist of antibody reagents that provide identical staining patterns to those obtained with the reference combination, meaning they recognize the same epitopes, they are conjugated with fluorochromes that can be measured in the same wavelengths and that they provide similar stain indices (e.g. < 30% difference in fluorescence intensity profiles) for both positive and negative cell populations coexisting in normal samples. In addition, they should include sufficient numbers of data files derived from distinct samples that mimic the acceptable levels of (technical and biological) variability that might be expected, for instance, for measurements performed at different days, in different instruments and distinct centers, by multiple technicians. Since these data bases are used as reference staining patterns against which stainings performed in a specific sample (or group of samples) from individual patients are directly compared, via the above described approaches and innovative software tools, different data bases per antibody combination, sample type, and even age, are required for optimal performance of the automated gating approach. Of note, in step 2, comparison of each of the clusters of events from a data file, against each of the different cell populations in the data base, requires algorithms (e.g. distinct multivariate analysis algorithms) and probability-based scoring criteria for the definition of the match vs. similar vs. unmatched results. Through this machine learning (Abu-Mostafa et al., 2012) approach, experts' knowledge introduced in the pre-defined cell populations (identified/gated) in the data base, is brought into a highly objective computer-based reference standard, which might be uniformly used across different laboratories, independently of the local expertise (e.g. knowledge and experience) (Kalina et al., 2012).

Validation of the EuroFlow automated gating approach vs expert-based manual gating for several different EuroFlow panels (van der Burg et al., 2019; van Dongen et al., 2012; Flores-Montero et al., 2019) and distinct types of samples (Flores-Montero et al., 2017), has shown that it is significantly faster, at the same time it is associated with high accuracy and reproducibility, with typically < 2% events in normal samples being required to be checked by the expert, based on comparison with all cell populations present in a data base matched for type of sample, antibody panel, sample preparation protocol and instrument set-up and calibration conditions (Flores-Montero et al., 2017; van Dongen et al., 2012; Flores-Montero et al., 2019; Kalina et al., 2018). For example, automated gating alarmed for the presence of abnormal cells in virtually all (> 450) patients diagnosed with acute leukemia, B cell chronic lymphoproliferative disorders (BCPD) and multiple myeloma (MM), after they had been stained with the EuroFlow ALOT (Acute Leukemia Orientation Tube), LST (Lymphocyte Screening Tube) and MM-MRD (Multiple Myeloma-Minimal Residual Disease) panel, respectively. Alarms were based on the presence of altered phenotypes and/or abnormally increased numbers of hematopoietic precursors, clonal/aberrant mature B-cells and plasma cells, respectively, as illustrated in Fig. 2 for a patient with a B-CLPD (Flores-Montero et al., 2017; van Dongen et al., 2012; Flores-Montero et al., 2019). In addition, automated gating also alarmed for specific technical problems in some data files (e.g. lack of reagent, too much debris, and inclusion of large tumor cells in the doublet gate).

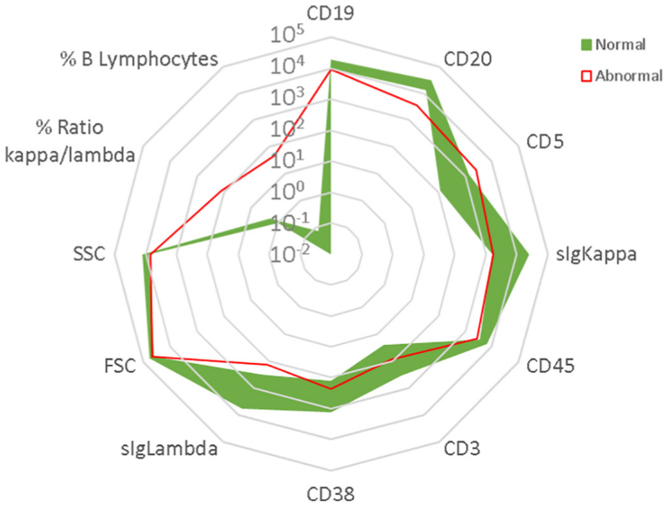

Fig. 2. Spider diagram illustrating the percent and immunophenotypic differences (software alarms) between a population of leukemia cells vs its normal counterpart. The data shown in the spider diagram derives from the comparison of a peripheral blood tumor chronic lymphocytic leukemia B-cell population (red line) against normal reference peripheral blood B-lymphocytes (green shaded areas) in a reference normal peripheral blood data base. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Despite all the above advances, automated identification and quantification of rare cells still remains a challenge. In 2008, we evaluated an approach (Pedreira et al., 2008) based on Principal Component Analysis (PCA) (Jolliffe and Cadima, 2016) and a probabilistic Bayesian model, which provided a tool for MRD detection in B-CLPD associated with relatively high sensitivity ($\leq 10^{-5}$). More recently (Qiu, 2015), rare cell identification was also approached through two techniques associated with very high computational cost: divergence (to measure pseudo-distances between two probability distributions) and an ensemble of Support Vector Machine (SVM). The experiment was done with training samples derived from the original samples, which roughly contained between 0.02% and 0.04% tumor cells, and produced an F-measure prediction value of only 0.69. EuroFlow has also designed and evaluated an approach based on the contribution of Canonical Correlation Analysis (Peltier et al., 2015), to which other multivariate analysis and clustering algorithms were added for comparison with cell populations in a data base; this approach has been subsequently validated and adopted for automated identification of rare cell populations (e.g. MRD detection) (Flores-Montero et al., 2017). Of note, while expert-based identification of MRD could reach a sensitivity (limit of detection) of 2×10^{-6}, parallel automatic gating alarmed for the presence of MRD at a limit of detection of 10^{-5}, which is still above the limit of sensitivity (within quantitative range) recommended by the International Myeloma working Group for MRD monitoring in MM (Kumar et al., 2016).

4. Data bases for classification of cell populations into distinct disease categories

The primary goal of clinical flow cytometry for diagnosis, classification and monitoring of leukemia and lymphoma and primary immune deficiencies, is to identify the presence of one or more populations of altered cells in a flow cytometry data set corresponding to a patient sample; in case such abnormal cell population(s) is identified, a second goal is pursued: to accurately link the altered cell population to the underlying disease condition and/or patient outcome (Pedreira et al., 2013).

As mentioned above the two distinct types of alterations most frequently identified in biological samples investigated by flow cytometry, for the presence of hematopoietic tumor cells or an underlying immunodeficiency, include: i) the presence of so-called aberrant phenotypes which are not seen in normal or reactive conditions, and ii) abnormally increased or decreased (absolute and/or reactive) cell counts. Distinct disease conditions are usually associated with uniquely altered cell distribution patterns and/or cell phenotypic profiles. Similarly to what has been designed for automated gating of flow cytometry data, EuroFlow has built a set of different data bases by merging flow cytometry data files of samples derived from multiple diseases and that contained the typically altered cell populations (e.g. tumor cell populations) from patients with e.g. distinct types of acute leukemia, B-CLPD, T-CLPD, myelodysplastic syndromes, paroxysmal nocturnal hemoglobinuria, primary immunodeficiencies, and other groups of diseases (Arber et al., 2016; Seidel et al., 2019). Once an altered cell population is identified in a flow cytometry data file from a given patient via comparison with a data base containing all normal cell populations present in that particular type of sample, the features (number and phenotype) of the altered (i.e. alarmed) cell population, can be further compared with those of cells from a data base that contains sets of representative cases of multiple diagnostic entities. Based on this latter comparison, the interrogated (altered) cell population(s) can be further classified into one (or more) or none of such diagnostic entities.

In this regard, EuroFlow has also constructed data bases containing sets of flow cytometry data files corresponding to distinct disease categories and specific for distinct EuroFlow screening tubes (e.g. ALOT) and antibody panels (e.g. BCLPD panel). Such data bases contain hundreds of patient samples classified according to the distinct World Health Organization (WHO) diagnostic categories (Arber et al., 2016). In turn, each patient sample consists of n-dimensional data about each individual cells in the data base (and/or the corresponding median fluorescence intensity values for each individual phenotypic parameter evaluated per patient sample); such data are then, either directly used for analysis, or employed to generate 2-D views of the original n-D space data, where optimal separation between groups of patients from distinct diagnostic categories in a given data base, is obtained through distinct (e.g. multivariate analysis) algorithms. Thus, reference cases within the same diagnostic category in the data base should form a cluster of patients with mean values and standard deviation values per diagnostic category, in the either supervised or unsupervised 2-D spaces generated for the comparison of newly interrogated altered cell populations vs. altered cell populations of cases from ≥1 diagnostic category in the data base. Thus, once a new case –not belonging to the reference data base- is compared against the reference data base values, it is also mapped in the same 2-D space and its location is confronted with that of the clusters of diseases contained within the reference cases. The decision to assign this cell population newly interrogated against the data base, to a given diagnostic category, relies on i) determining the probability it belongs to each of the distinct diagnostic categories in the data base, and ii) deciding to which one it belongs to because of being associated with the highest probability (Fig. 3). Of note, comparison of a new case against the reference cases in the data base can be done directly in the n-D space built on the basis of all parameters evaluated in common and visualized in 2-D graphics, or it can be directly compared in the 2-D space.

5. Smart classification of altered cell populations

Cells that belong to the same population are expected to show very similar immunophenotypic features and consequently, single events may be modeled as elements of an n-dimensional (n-D) space -formed by the n evaluated markers- where they cluster together and may be assigned (labeled) to specific classes that correspond to distinct cell populations. This opens the door to approach the classification of altered cell populations in a statistical pattern classification framework.

The strategies proposed by EuroFlow for smart classification of altered cell populations, are typically based on use of reference data files that contain reference cell populations that have been previously labeled by experts. In brief, the different EuroFlow strategies proposed for the classification of altered cell populations, such as those described below, are generally based on distinct algorithms that are applied to the reference data sets (i.e. the actual data in the reference data bases used), followed by the confrontation of an interrogated cell population (i.e. test case) against the transformed reference dataset (Pedreira et al., 2013; Costa et al., 2010; Lhermitte et al., 2018; Pedreira et al., 2008). In some of the strategies selected, the initial transformations aim at mapping the n-dimensional data into 2-D spaces, to allow the use of e.g. standard deviation (SD) based classification criteria. For other strategies, all transformations are directly done in the n-dimensional markers' space, without any previous reduction of data dimensionality, which avoids loss of part of the overall information, but prevents the use of standard deviation classification criteria. Of note, the choice of the panel of markers to be included in a data base is critical, because it determines the n-dimensional space in which the events are embedded and whether the distinct groups of events contained in the data base are susceptible of being clearly separated between them in such n-dimensional space. Thus, independently of the procedure used to classify a given cell population, the classification can only be fully successful when the distinct groups of reference cell populations are optimally separated in the n-dimensional space formed by the distinct scatter and phenotypic markers used to stain in common the cell populations in the data base; thus, the success of smart classification of cell populations intrinsically depends on appropriate design of the panel of markers used to achieve the required separation between the distinct relevant cell populations in the data base. Because of this, EuroFlow has carefully developed first, the appropriate antibody combinations and panels of markers that provide the required separation among the cell populations of interest in the n-dimensional markers' space (van der Burg et al., 2019; Theunissen et al., 2017; Blanco et al., 2018; Blanco et al., 2019; Damasceno et al., 2019), and subsequently, the analytical tools (Pedreira et al., 2013).

Overall, the different classification strategies can be divided in two main groups. The first group comprises approaches that first map data from the n-dimensional space into a 2-D space (Fig. 3C), followed by application of a pre-defined classification algorithm to decide which specific (e.g. diagnostic) label should be assigned to each individual test case, based on the resulting 2-dimensional transformed data. This includes several approaches that differ among them in the way dimensionality is reduced, such as PCA (Costa et al., 2010), Canonical Correlation Analysis (CA) (Peltier et al., 2015), and Robust (curve) analysis (Daszykowski et al., 2007) illustrated in the Fig. 3, panels C2, C3 and C4 respectively. The second group of classification approaches embraces strategies such as the neighbor APS (Fig. 3, panel C5) and Support Vector Machine (SVM) based algorithms (Goldberger et al., 2004), in which the decision is directly generated using the n-dimensional data, with subsequent visualization in 2-D plots.

From the first group of approaches selected for reducing multidimensional data, the Automatic Population Separator (APS) plots implemented > 15 years ago in the Infinicyt software, use PCA (Costa et al., 2010) to bring data from n-D to 2-D spaces (Fig. 3C2). PCA is a transformation that produces n linear combinations of the markers (the so-called n principal components), in such a way that the first principal component accounts for as much of the variability in the data as possible, while the second, third and the following components account for the maximum variability not explained by the previous ones. Thus, through PCA, principal components maximize description of data variability both among, and within, the groups, the first few principal components being usually sufficient to appropriately represent the structure of data (Fig. 3C2). In turn, CA (Peltier et al., 2015) – Fig. 3C3, despite following the PCA principles, is a supervised approach that searches for components (directions) in such a way that, the Euclidean

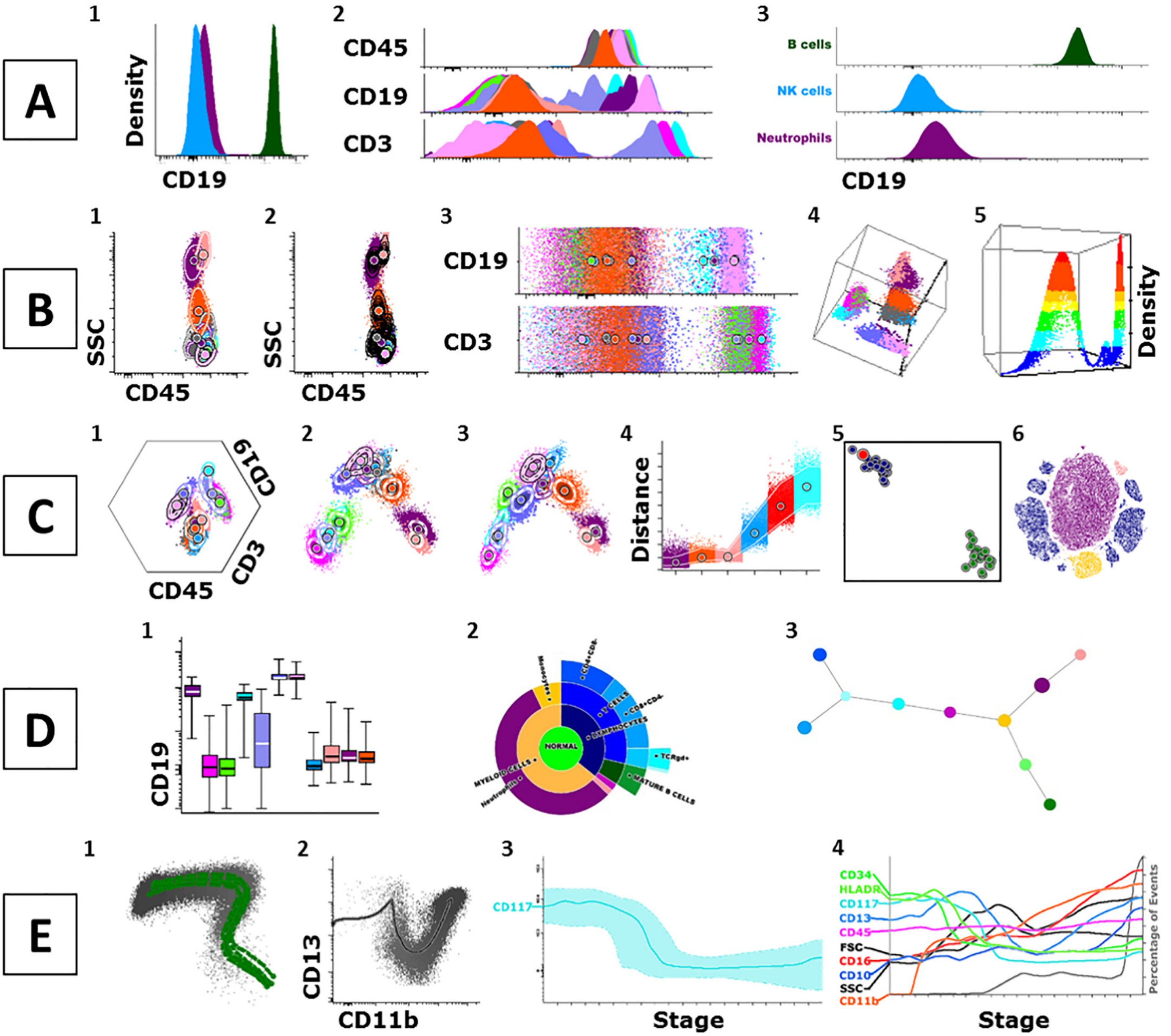

Fig. 3. Classical vs novel graphical tools for visualization of flow cytometry data. Classical single parameter (vs cell count) histograms (panel A1)) and histogram arrays for several parameters (panel A2) and multiple cell populations (panel A3) are shown in A. In B, conventional 2- (panels B1–B3) and 3-dimensional (panels B4 and B5) dot plots (panels B1, B3 and B4) and density plots (panels B2 and B5) are displayed. > 3-dimensional (panel C) flow cytometry data. In C, a multiview plot (panel C1) together of several 2-dimensional graphical plots of data corresponding to distinct multivariate analysis approaches are shown in panels C2–C6: principal component analysis (APS) in panel C2, canonical correlation analysis in panel C3, Robust curve analysis in panel C4, neighbor-APS (NAPS) in panel C5 and t-SNE in panel C6. In D, expression of the CD19 marker in different cell subsets coexisting in a sample is shown in a box plot graphical display (panel D1), together with two distinct graphical representations of population trees reflecting their size and the relationship among them (panels D2 and D3). In E, a principal component (PC) 1 vs PC2 APS plot (panel E1) as well as a conventional bivariate dot plot (panel E2) showing the maturation pathway of normal BM neutrophils is displayed, together with the pattern of expression of CD117 (panel E3) and multiple other markers (panel E4), along the different BM maturation stages (X-axis in panels E4 and E5).

distances in the transformed 2-dimensional space reflect Mahalanobis distances in the original n-dimensional space, resulting in maximum separation among the groups with minimal intragroup variation (Fig. 3C3). Calculation of Mahalanobis distances requires pre-established (supervised) definition of the distinct groups. Finally, the so-called Robust curve analysis, is based on robust distances (Daszykowski et al., 2007) between events and groups, calculated by reducing the influence of outliers in Mahalanobis distances, i.e. by replacing, in the calculation of Mahalanobis distances, the covariance matrix by its robust counterpart. The 2-D transformed dataset is set up by pairs containing the two robust distance of each event to each of two confronted groups (e.g. disease categories) (Fig. 3C4). For each of the above three strategies, transformed 2-D datasets are generated first; subsequently, the test-case is interrogated against all two-by-two group comparisons, using the (e.g. 2) standard deviation criterion. Based on this criterion, the test-case is associated with one of the two groups, e.g. if it is inside the 2 standard deviations contour of this group and clearly outside the 2 standard deviation contour of the other group. The final label is defined by a score, based on the results obtained from all two-by-two group comparisons. Thus, when a (test) tumor cell population is to be classified among three possible groups (disease categories) such as reference AML, BCP-ALL and T-ALL groups, and it falls inside the AML reference group in both the AML vs BCP-ALL and AML vs T-ALL comparisons, while it falls outside the BCP-ALL and T-ALL groups in the BCP-ALL vs

T-ALL comparison, this case would be classified as AML (Costa et al., 2010; Lhermitte et al., 2018).

Regarding those strategies that are associated with classification decisions directly in the n-dimensional data space that have been evaluated by EuroFlow for classification of cell populations, based on flow cytometry reference data bases, SVM and NAPS (Neighborhood Automatic Population Separator), resulted particularly attractive. SVM, is based on the classical Support Vector Machine algorithm (Goldberger et al., 2004) and aims at finding an optimal separation hyperplane directly in the high dimensional space; such hyperplane maximizes the separation between different groups of cell populations in the training data set (e.g. data base), in such a way that all cell populations from one group are located as far as possible from those belonging to another group. SVM is by construction, a binary classifier, i.e. it only separates two groups at a time. Thus, the normalized distance to the limits of the two groups are used to estimate the probability that the test-cell population should be classified as belonging to one or the other group. In turn, the NAPS approach is based on the Neighborhood Components Analysis algorithm (NCA) – Fig. 3C5 (Cortes and Vapnik, 1995). The key concept behind NCA is to learn a distance metric, in the n-dimensional markers' space, that maximizes the probability that new (test) cell populations are correctly classified. After submitting the n-dimensional data to a linear transformation, provided by the learned metric, the K-Nearest Neighborhood algorithm may be applied in a much more efficient way to classify new cases. Furthermore, a soft neighborhood assignment is used through which a point selects another point as its neighbor with a certain probability. This step is what makes feasible a numerical solution to the optimization problem –i.e. to maximize the probability that new cases are correctly classified- (Cortes and Vapnik, 1995).

Besides the above described supervised classification methods, other clustering (unsupervised) algorithms might also be used to gather groups of cell populations, without the need to use a pre-labeled training set. Two main pathways are possible to achieve this goal. The first consists on reducing dimensionality to 2-D through PCA, and then apply the classical K-means. K-Means could be directly applied in the n-dimension space, but results are often not satisfactory, mainly due to the sparsity of the space. Alternatively, the t-Distributed Stochastic Neighbor Embedding (t-SNE) (van der Maaten and Hinton, 2008) might be applied. t-SNE aims at visualizing high-dimension data by projecting it into a 2-D or 3-D space(Fig. 3C6). It belongs to the same family of methods as NCA (Goldberger et al., 2004) and tries to place individual events from an n-dimensional space in a 2-D or 3-D space in which their neighborhood identity in n-D, is optimally preserved (van der Maaten and Hinton, 2008) (Fig. 3C6).

6. Graphical visualization of n-dimension data

Dimensionality reduction from the n-D space generated by multicolor flow cytometry measurements, to the visually accessible 2-D and 3-D spaces, allows flow cytometry experts to visually infer the relative position of two events (or groups of events) in the original high-dimensional space (Fig. 3A–B). This approach has been used in clinical flow cytometry to support expert decisions about the similarities between groups of events and/or cell populations and their classification into normal vs pathological cell populations (Pedreira et al., 2013). Thus, plots that represent multi-dimensional sets of flow cytometry data have been traditionally used to select (e.g. gate) groups-of-events, and consequently, to identify cell populations, using direct 1-D to 3-D representations of data (Fig. 3A–B), in an expert-based subjective way; i.e., the expert selects from all possible 1-D to 3-D data plots, which ones she/he will focus for the identification of the cell populations of interest present in a flow cytometry data file. Through this approach, graphical representations of multidimensional data, provide the opportunity to the decision-maker, to supervise the result provided by the system since she/he can directly interact with the elements plotted in these 2-D graphics to select events, groups of events and visualize the corresponding labels and statistic data. However, any dimensionality reduction entails loss of a fraction of the original information present in the n-dimensional dataset. Thus, whenever possible, 1-D to 3-D graphics used to visualize n-D flow cytometry data should be used as support tool for data visualization, particularly when automated data analysis algorithms are applied to identify and classify cell populations in a sample; in contrast, these graphics should not be "the only tool" to rely on when analyzing n-D flow cytometry data, since it has become a time consuming, expert-based, subjective approach, whose applicability and feasibility decrease as the number of parameters and cell population in a flow cytometry data set increases (Pedreira et al., 2013).

Currently, there are several ways to reduce data dimensionality in the (computing) literature (van der Maaten and Hinton, 2008; Rogers and Holyst, 2009; Flores-Montero et al., 2017; Yan and Xu, 2007; Peres et al., 2013; Mead, 1992). Thus, in addition to the classical graphical tools currently available in virtually every flow cytometry software such as single-parameter histograms, bivariate and three-D dot plots (Fig. 3A–B), 2-D graphical representations of n-D data based on distinct multivariate data analysis approaches, such as PCA (Costa et al., 2010), CA, Robust (curve) analysis (Daszykowski et al., 2007), NAPS/NCA (Cortes and Vapnik, 1995) and t-SNE (van der Maaten and Hinton, 2008) (Fig. 3, panels C2–C6), as well as other types of plots illustrated in Fig. 3D–E, need also to be considered as standard graphical plots for visualization of multicolor flow cytometry data.

7. Concluding remarks

The increased complexity and volume of flow cytometry data generated in diagnostic laboratories has fostered the design, implementation and validation of novel data analysis tools and strategies for: i) automated flow cytometry gating and identification of multiple populations of cells coexisting in a biological sample like blood, bone marrow and lymph nodes, and ii) classification of altered tumor and immune cells. In the last decade, EuroFlow has developed and validated such strategies and tools via combined: i) clustering algorithms, ii) multivariate data analysis approaches to classify samples with altered cell populations into specific disease categories, and iii) the corresponding 2-D visualization plots for expert-guided visualization and control of the whole process. Such tools have taken advantage of supervised approaches for labeling of relevant individual (biological or clinical) cell populations, based on well-defined and pre-classified data bases of normal and/or pathologic cells (Flores-Montero et al., 2019).

The use of these new tools and strategies will contribute to standardization of clinical flow cytometry, through faster and more objective analysis and interpretation of flow cytometric big data both in research and clinical settings. At the same time the new EuroFlow tools and strategies provide the basis for ongoing analysis of flow cytometry big data sets derived from > 20-color experiments in which hundreds of distinct cell populations are simultaneously identified (Liechti and Roederer, 2019; Nettey et al., 2018).

Acknowledgements

E.S.d.C. and C.E.P. acknowledge Research Foundation of the State of Rio de Janeiro, Rio de Janeiro, Brazil (FAPERJ - E26/010.001659/2016, E26/110.105/2014, E26/102.191/2013, E26/202.838/2017-CNE), CAPES - Coordination of Improvement of Higher Education Personnel (PROEX - 1201036), Programa de Oncobiologia and Conselho Nacional de Desenvolvimento Cientifico e Tecnologico–CNPQ of Brazil (409440/2016-7, 304243/2015, PVE4001942014-7).

References

Abu-Mostafa, Y., Malik Magdon, I., Lin, T., 2012. Learning from Data. AMLBooks.

Aghaeepour, N., Nikolic, R., Hoos, H.H., Brinkman, R.R., 2011. Rapid cell population identification in flow cytometry data. Cytometry A 79, 6–13. https://doi.org/10.1002/cyto.a.21007.

Aghaeepour, N., Finak, G., FlowCAP Consortium, DREAM Consortium, Hoos, H., Mosmann, T.R., Brinkman, R., Gottardo, R., Scheuermann, R.H., 2013. Critical assessment of automated flow cytometry data analysis techniques. Nat. Methods 10 (3), 228–238. https://doi.org/10.1038/nmeth.2365.

Arber, D.A., Orazi, A., Hasserjian, R., Thiele, J., Borowitz, M.J., et al., 2016. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. Blood 127, 2391–2405. https://doi.org/10.1182/blood-2016-03-643544.

Blanco, E., Pérez-Andrés, M., Arriba-Méndez, S., et al., 2018. Age-associated distribution of normal B-cell and plasma cell subsets in peripheral blood. J. Allergy Clin. Immunol. 141 (6), 2208–2219. .e16. https://doi.org/10.1016/j.jaci.2018.02.017.

Blanco, E., Pérez-Andrés, M., Arriba-Méndez, S., et al., 2019. Defects in memory B-cell and plasma cell subsets expressing different immunoglobulin-subclasses in CVID and Ig-subclass deficiencies. J. Allergy Clin. Immunol. https://doi.org/10.1016/j.jaci.2019.02.017.

Bonner, W.A., Hulett, H.R., Sweet, R.G., Herzenberg, L.A., 1972. Fluorescence activated cell sorting. Rev. Sci. Instrum. 43, 404–409. https://doi.org/10.1063/1.1685647.

Chester, C., Maecker, H.T., 2015. Algorithmic tools for mining high-dimensional cytometry data. J. Immunol. 195, 773–779. https://doi.org/10.4049/jimmunol.1500633.

Comans-Bitter, W.M., de Groot, R., van den Beemd, R., Neijens, H.J., Hop, W.C., Groeneveld, K., Hooijkaas, H., van Dongen, J.J., 1997. Immunophenotyping of blood lymphocytes in childhood. Reference values for lymphocyte subpopulations. J. Pediatr. 130, 388–393. https://doi.org/10.1016/s0022-3476(97)70200-2.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297.

Costa, E.S., Arroyo, M.E., Pedreira, C.E., García-Marcos, M.A., Tabernero, M.D., Almeida, J., Orfao, A., 2006. A new automated flow cytometry data analysis approach for the diagnostic screening of neoplastic B-cell disorders in peripheral blood samples with absolute lymphocytosis. Leukemia 20, 1221–1230. https://doi.org/10.1038/sj.leu.2404241.

Costa, E.S., Pedreira, C.E., Barrena, S., Lecrevisse, Q., Flores, J., Quijano, S., Almeida, J., del Carmen García-Macias, M., Bottcher, S., Van Dongen, J.J., Orfao, A., 2010. 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 24, 1927–1933. https://doi.org/10.1038/leu.2010.160.

Damasceno, D., Teodosio, C., van den Bossche, W.B., et al., 2019. Distribution of subsets of blood monocytic cells throughout life. J. Allergy Clin. Immunol. https://doi.org/10.1016/j.jaci.2019.02.030.

Daszykowski, M., Kaczmarek, K., Vander Heyden, Y., Walczak, B., 2007. Robust statistics in data analysis – a review basic concepts. Chemom. Intell. Lab. Syst. 85, 203–219.

Finak, G., Frelinger, J., Jiang, W., Newell, E.W., Ramey, J., Davis, M.M., Kalams, S.A., De Rosa, S.C., Gottardo, R., 2014. OpenCyto: an open source infrastructure for scalable, robust, reproducible, and automated end-to-end flow cytometry data analysis. PLoS Comput. Biol. 10 (8), e1003696. https://doi.org/10.1371/journal.pcbi.1003806.

Finak, G., Langweiler, M., et al., 2016. Standardizing flow cytometry immunophenotyping analysis from the Human ImmunoPhenotyping Consortium. Sci. Rep. 6, 20686. https://doi.org/10.1038/srep20686.

Flores-Montero, J., Flores, L.S., Paiva, B., Puig, N., García-Sánchez, O., Böttcher, S., van der Velden, V.H., Pérez-Morán, J.J., Vidriales, M.B., García-Sanz, R., Jimenez, C., González, M., Martinez-López, J., Mateos, A.C., Grigore, G.E., Fluxá, R., Pontes, R., Caetano, J., Sedek, L., Del Cañizo, M.C., Bladé, J., Lahuerta, J.J., Aguilar, C., Bárez, A., García-Mateo, A., Labrador, J., Leoz, P., Aguilera-Sanz, C., San-Miguel, J., Mateos, M.V., Durie, B., van Dongen, J.J., Orfao, A., 2017. Next generation flow (NGF) for highly sensitive and standardized detection of minimal residual disease in multiple myeloma. Leukemia. https://doi.org/10.1038/leu.2017.29.

Flores-Montero, J., Grigore, G., Fluxá, R., et al., 2019. Data bases of normal samples as internal reference for automated gating and identification of abnormal cells. J. Immunol. Methds (in this number).

Freer, G., Rindi, L., 2013. Intracellular cytokine detection by fluorescence-activated flow cytometry: basic principles and recent advances. Methods. https://doi.org/10.1016/j.ymeth.2013.03.035.

Goldberger, J., Roweis, S., Hinton, G., Salakhutdinov, R., 2004. Neighbourhood components analysis. In: Proceedings of the 17th International Conference on Neural Information Processing Systems, pp. 513–520.

Hulett, H.R., Bonner, W.A., Barrett, J., Herzenberg, L.A., 1969. Cell sorting: automated separation of mammalian cells as a function of intracellular fluorescence. Science 166, 747–749. https://doi.org/10.1126/science.166.3906.747.

Hunter, S.D., Peters, L.E., Wotherspoon, J.S., Crowe, S.M., 1994. Lymphocyte subset analysis by Boolean algebra: a phenotypic approach using a cocktail of 5 antibodies and 3 color immunofluorescence. Cytometry 15, 258–266. https://doi.org/10.1002/cyto.990150311.

Jolliffe, I.T., Cadima, J., 2016. Principal component analysis: a review and recent developments. Philos. Trans. A Math. Phys. Eng. Sci. 374 (2065), 20150202.

Kalina, T., Flores-Montero, J., van der Velden, V.H., Martin-Ayuso, M., Böttcher, S., Ritgen, M., Almeida, J., Lhermitte, L., Asnafi, V., Mendonça, A., de Tute, R., Cullen, M., Sedek, L., Vidriales, M.B., Pérez, J.J., te Marvelde, J.G., Mejstrikova, E., Hrusak, O., Szczepański, T., van Dongen, J.J., Orfao, A., EuroFlow Consortium (EU-FP6, LSHB-CT-2006-018708), 2012. EuroFlow standardization of flow cytometer instrument settings and immunophenotyping protocols. Leukemia 26, 1986–2010.

Kalina, T., Brdickova, N., Glier, H., Fernandez, P., Bitter, M., Flores-Montero, J., van Dongen, J.J.M., Orfao, A., 2018. Frequent issues and lessons learned from EuroFlow QA. J. Immunol. Methods. https://doi.org/10.1016/j.jim.2018.09.008.

Kumar, S., Paiva, B., Anderson, K.C., et al., 2016. International Myeloma Working Group consensus criteria for response and minimal residual disease assessment in multiple myeloma. Lancet Oncol. 17, e328–e346. https://doi.org/10.1016/S1470-2045.

Lhermitte, L., Mejstrikova, E., van der Sluijs-Gelling, A.J., Grigore, G.E., Sedek, L., Bras, A.E., Gaipa, G., Sobral da Costa, E., Novakova, M., Sonneveld, E., Buracchi, C., de Sá Bacelar, T., Te Marvelde, J.G., Trinquand, A., Asnafi, V., Szczepanski, T., Matarraz, S., Lopez, A., Vidriales, B., Bulsa, J., Hrusak, O., Kalina, T., Lecrevisse, Q., Martin Ayuso, M., Brüggemann, M., Verde, J., Fernandez, P., Burgos, L., Paiva, B., Pedreira, C.E., van Dongen, J.J.M., Orfao, A., van der Velden, V.H.J., 2018. Automated database-guided expert-supervised orientation for immunophenotypic diagnosis and classification of acute leukemia. Leukemia 32, 874–881. https://doi.org/10.1038/leu.2017.313.

Liechti, T., Roederer, M., 2019. OMIP-051 – 28-color flow cytometry panel to characterize B cells and myeloid cells. Cytometry A 95, 150–155. https://doi.org/10.1002/cyto.a.23689.

Lo, K., Brinkman, R.R., Gottardo, R., 2008. Automated gating of flow cytometry data via robust model-based clustering. Cytometry A 73, 321–332. https://doi.org/10.1002/cyto.a.20531.

Mair, F., Felix, J., Hartmann, F.J., Mrdjen, D., Tosevski, V., Krieg, C., Becher, B., 2016. The end of gating? An introduction to automated analysis of high dimensional cytometry data. Eur. J. Immunol. 46, 34–43. https://doi.org/10.1002/eji.201545774.

Malek, M., Taghiyar, M.J., Chong, L., Finak, G., Gottardo, R., Brinkman, R.R., 2015. flowDensity: reproducing manual gating of flow cytometry data by automated density-based cell population identification. Bioinformatics 31, 606–607. https://doi.org/10.1093/bioinformatics/btu677.

Martini, J., Recht, M.I., Huck, M., Bern, M.W., Johnson, N.M., Kiesel, P., 2012. Time encoded multicolor fluorescence detection in a microfluidic flow cytometer. Lab Chip 12, 5057–5062. https://doi.org/10.1039/c2lc40515f.

Marx, V., 2013. Biology: the big challenges of big data. Nature 498, 255–260. https://doi.org/10.1038/498255a.

Mead, A., 1992. Review of the development of multidimensional scaling methods. J. R. Stat. Soc. D 41 (1), 27–39.

Nettey, L., Giles, A.J., Chattopadhyay, P.K., 2018. OMIP-050: a 28-color/30-parameter fluorescence flow cytometry panel to enumerate and characterize cells expressing a wide array of immune checkpoint molecules. Cytometry A 93, 1094–1096. https://doi.org/10.1002/cyto.a.23608.

Orfao, A., Schmitz, G., Brando, B., Ruiz-Arguelles, A., Basso, G., Braylan, R., Rothe, G., Lacombe, F., Lanza, F., Papa, S., Lucio, P., San Miguel, J.F., 1999. Useful information provided by the flow cytometric immunophenotyping of hematological malignancies: current status and future directions. Clin. Chem. 45, 1708–1717.

Pedreira, C.E., Costa, E.S., Almeida, J., Fernandez, C., Quijano, S., Flores, J., Barrena, S., Lecrevisse, Q., Van Dongen, J.J.M., Orfao, A., on behalf of EuroFlow Consortium, 2008. A probabilistic approach for the evaluation of minimal residual disease by multiparameter flow cytometry in leukemic B-cell chronic lymphoproliferative disorders. Cytometry A 12, 1141–1150. https://doi.org/10.1002/cyto.a.20638.

Pedreira, C.E., Costa, Elaine S., Lecrevisse, Q., van Dongen, J.J.M., Orfao, A., EuroFlow Consortium, 2013. Overview of clinical flow cytometry data analysis: recent advances and future challenges. Trends Biotechnol. 31 (7), 415–427. https://doi.org/10.1016/j.tibtech.2013.04.008.

Peltier, C., Visalli, M., Schlich, P., 2015. Comparison of canonical variate analysis and principal component analysis on 422 descriptive sensory studies. Food Qual. Prefer. 40, 326–333.

Peres, R.T., Aranha, C., Pedreira, C.E., 2013. Optimized bi-dimensional data projection for clustering visualization. Inf. Sci. 232, 104–115.

Qiu, P., 2015. Computational prediction of manually gated rare cells in flow cytometry data. Cytometry A 87A, 594–602. https://doi.org/10.1002/cyto.a.22654.

Qiu, P., Simonds, E.F., Bendall, S.C., Gibbs Jr., K.D., Bruggner, R.V., Linderman, M.D., Sachs, K., Nolan, G.P., Plevritis, S.K., 2011. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. Nat. Biotechnol. 29, 886–891. https://doi.org/10.1038/nbt.1991.

Quinn, J., Fisher, P.W., Capocasale, R.J., Achuthanandam, R., Kam, M., Bugelski, P.J., Hrebien, L., 2007. A statistical pattern recognition approach for determining cellular viability and lineage phenotype in cultured cells and murine bone marrow. Cytometry A 71, 612–624. https://doi.org/10.1002/cyto.a.20416.

Robinson, J.P., Rajwa, B., Patsekin, V., Davisson, V.J., 2012. Computational analysis of high-throughput flow cytometry data. Expert Opin. Drug Discovery 7, 679–693. https://doi.org/10.1517/17460441.2012.693475.

Roederer, M., Nozzi, J.L., Nason, M.C., 2011. SPICE: exploration and analysis of post-cytometric complex multivariate datasets. Cytometry A 79A, 167–174. https://doi.org/10.1002/cyto.a.21015.

Rogers, W.T., Holyst, H.A., 2009. A bioconductor package for fingerprinting flow cytometric data. AdvBioinform. https://doi.org/10.1155/2009/193947. 193947–11.

Schultze, J.L., 2015. Teaching 'big data' analysis to young immunologists. Nat. Immunol. 16, 902–905. https://doi.org/10.1038/ni.3250.

Seidel, M.G., Kindle, G., Gathmann, B., et al., 2019. Registry working definitions for the clinical diagnosis of inborn errors of immunity. J. Allergy Clin. Immunol. Pract. piihttps://doi.org/10.1016/j.jaip.2019.02.004. S2213-2198(19) 30168-0.

Sutherland, D.R., Anderson, L., Keeney, M., Nayar, R., Chin-Yee, I., 1996. The ISHAGE guidelines for CD34+ cell determination by flow cytometry. International Society of Hematotherapy and Graft Engineering. J. Hematother. 5 (3), 213–226. https://doi.org/10.1089/scd.1.1996.5.213.

Theunissen, P., Mejstrikova, E., Sedek, L., et al., 2017. Standardized flow cytometry for highly sensitive MRD measurements in B-cell acute lymphoblastic leukemia. Blood 129, 347–357. https://doi.org/10.1182/blood-2016-07-726307.

van der Burg, M., Kalina, T., Perez-Andres, M., et al., 2019. The euroflow pid orientation tube for flow cytometric diagnostic screening of primary immunodeficiencies of the lymphoid system. Front. Immunol. 4 (10), 246. https://doi.org/10.3389/fimmu.2019.00246.

van der Maaten, L.J.P., Hinton, G.E., 2008. Visualizing high-dimensional data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605.

van Dongen, J.J., Lhermitte, L., Böttcher, S., Almeida, J., van der Velden, V.H., Flores-Montero, J., Rawstron, A., Asnafi, V., Lécrevisse, Q., Lucio, P., Mejstrikova, E., Szczepański, T., Kalina, T., de Tute, R., Brüggemann, M., Sedek, L., Cullen, M., Langerak, A.W., Mendonça, A., Macintyre, E., Martin-Ayuso, M., Hrusak, O., Vidriales, M.B., Orfao, A., 2012. EuroFlow Consortium (EU-FP6, LSHB-CT-2006-018708). EuroFlow antibody panels for standardized n-dimensional flow cytometric immunophenotyping of normal, reactive and malignant leukocytes. Leukemia 26, 1908–1975. https://doi.org/10.1038/leu.2012.120.

Wood, B.L., 2016. Principles of minimal residual disease detection for hematopoietic neoplasms by flow cytometry. Cytometry B Clin. Cytom. 90, 47–53. https://doi.org/10.1002/cyto.b.21239.

Yan, S., Xu, D., 2007. Graph embedding and extensions: a general framework for dimensionality reduction. IEEE Trans. Pattern Anal. Mach. Intell. https://doi.org/10.1109/TPAMI.2007.12.

Zare, H., Bashashati, A., Kridel, R., Aghaeepour, N., Haffari, G., Connors, J.M., Gascoyne, R.D., Gupta, A., Brinkman, R.R., Weng, A.P., 2012. Automated analysis of multidimensional flow cytometry data improves diagnostic accuracy between mantle cell lymphoma and small lymphocytic lymphoma. Am. J. Clin. Pathol. 137, 75–85. https://doi.org/10.1309/AJCPMMLQ67YOMGEW.