Published in final edited form as:
Cytometry A. 2010 July ; 77(7): 705–713. doi:10.1002/cyto.a.20901.

Data Analysis in Flow Cytometry: The Future Just Started

Enrico Lugli[1], **Mario Roederer**[1], and **Andrea Cossarizza**[2]

[1] ImmunoTechnology Section, Vaccine Research Center, National Institute of Allergy and Infectious Diseases, NIH, 40 Convent Drive, 20892, Bethesda, MD, USA

[2] Department of Biomedical Sciences, University of Modena and Reggio Emilia, via Campi 287, 41125, Modena, Italy

Abstract

In the last 10 years, a tremendous progress characterized flow cytometry in its different aspects. In particular, major advances have been conducted regarding the hardware/instrumentation and reagent development, thus allowing fine cell analysis up to 20 parameters. As a result, this technology generates very complex data sets that demand for the development of optimal tools of analysis. Recently, many independent research groups approached the problem by using both supervised and unsupervised methods. In this paper, we will review the new developments concerning the use of bioinformatics for polychromatic flow cytometry and propose what should be done in order to unravel the enormous heterogeneity of the cells we interrogate each day.

Key terms

polychromatic flow cytometry; data analysis; lymphocytes; T cells; immune system

Introduction

Differently to any other tissue in the body, cells being part of the immune system display a huge diversity and hundreds of subsets can been identified even within the same lineage, such as in CD4+ and CD8+ T cells or in dendritic cells. Identification of the heterogeneity of the immune system components can be only achieved through flow cytometry that allows the analysis of multiple surface and intracellular markers at the level of single cell. Major contributions in the last 10-15 years on the field of instrumentation, reagent development, and software analysis tools advanced the field of cell research forward – leading to the identification of specific subsets of cells with unique biological functions in normal and pathological conditions. The Herzenberg laboratory at Stanford University developed the first instrument able to detect 11 antigens in the same cell (1); this technology was later extended in the ImmunoTechnology Section at the Vaccine Research Center (VRC) at the NIH by utilizing quantum dots conjugated to monoclonal antibodies, upgraded instrumentation and allowed measurements up to 18 colors (2).

Meanwhile, development of new flow cytometric assays, *in primis* the capability to measure the release of cytokines by stimulated immune cells (3) and the analysis of the phosphorylation state of proteins involved in signal transduction (4), moved attention from basic phenotyping to more complex cell functions. All these aspects together undoubtedly

Correspondence to: Enrico Lugli, ImmunoTechnology Section, Vaccine Research Center, National Institute of Allergy and Infectious Diseases, NIH, 40, Convent Drive, 20892, Bethesda, MD, USA, Ph: +1 301 594 8602, Fax:+1 301 480 2788, luglie@mail.nih.gov or Andrea Cossarizza, MD, PhD., Chair of Immunology, Dept. of Biomedical Sciences, University of Modena and Reggio Emilia, Via Campi 287, 41125, Modena, Italy, Ph: +39 059 205 5415, Fax:+39 059 205 5426, andrea.cossarizza@unimore.it.

revealed multiple aspects of immune cell biology but, as a consequence, generate large and complex data sets. These data are best analyzed with the use of bioinformatic tools. Theoretically, millions of possible subpopulations can be identified in a single sample stained with 18 reagents; the number of variables measured can be increased by the different markers used in the analysis, by the experimental conditions (*e.g.*, stimulation time, concentration of the stimulus) or by the time points in an *in vitro* experiment or in a clinical trial. For example, more detailed analysis of generated data sets through the use of Bayesian networks revealed the existence of intracellular pathways not previously identified through classical biochemical approaches (5). Such large data sets can be certainly analyzed by the classical approach of sequential gating and then determining the representation of a particular cell subset expressing marker(s) of interest. However, in many cases this approach is both too time consuming and too subjective. The same problem arises when thousands of samples have to be analyzed in high-throughput screening experiments, for example in the context of data quality assessment, for which models have been recently proposed (6). More importantly, much information can be lost as our eyes and, as a consequence, our mind can only integrate information from two or perhaps three dimensions at a time in classical flow cytometric plots. Thus, the need to simplify analysis is absolutely necessary. Multiple attempts to do so have been proposed that can be roughly divided in two main categories, *i.e.*, supervised and unsupervised.

Most but not all of these new approaches are mainly explorative and not quantitative. Thus they cannot be used to identify statistically significant differences among groups of samples or cohorts of patients. Their main use is to identify trends in the data that otherwise would be missed by the classical sequential gating strategies. These trends must then be experimentally tested to see whether cell subsets or cells from healthy donors vs. patients behave differently or differentially respond to external stimuli.

In this paper, we review some of these types of analyses proposed over the last few years and discuss their pro and cons (summarized in Table 1); moreover, we will suggest what should be done and developed in the near future to obtain as much information from our data as possible.

Sequential gating: a universal method for analysis of flow cytometric data

Univariate histograms are widely used for the display of flow cytometric data, if only one parameter has to be visualized (7). They are useful when the same antigen is to be analyzed through multiple experimental samples (*e.g.*, the sample treated with different drugs or with different amounts of the same drug): in this case, the pattern of expression of a certain marker can be discerned by the graphs and quantitatively expressed as percentage of cells above a certain threshold, or alternatively, the mean (median) of the fluorescence intensity for the antigen. Most software packages provide tools for the overlay of histograms from multiple samples, allowing a rapid comparison of the parameter(s) of interest.

Multiple choices are available when two parameters have to be visualized at the same time. Historically, dot plots are the most popular graphs used for this purpose. In this case, single events are depicted as dots on the screen. Monochromatic dot plots, however, have many limitations: for example, different cells displaying the same amount of fluorescence are overlayed in the plot and their further distinction becomes thus difficult if not impossible. This is particularly challenging when rare populations have to be investigated and a huge number of events needs to be acquired. Density information overcome this problem and can be provided graphically by either the use of contour plots or colored dot plots (7,8). Contours identify regions with equal cell density and the representation obtained is irrespective of the number of acquired events. Pseudocolor dot plots, as described in FlowJo

software, can be also used: different colors are used depending on the cell density and thus give the idea of the proportions of the different cell populations. The specific algorithms used to generate the colors or contour lines are important; the use of “probability” algorithms is often advantageous since the resulting graphics are highly similar irrespective of the total number of events collected and are thereby less prone to misinterpretation.

Pro and cons of sequential gating strategy

Histograms and dot plots are a very simple and intuitive way of analyzing flow cytometric data and allow gating of specific populations of interest that can be isolated for further analysis, if multiple markers are used in the same panel. As indicated above, quantitative results in terms of frequency (*i.e.*, percentage) or grade of positivity (*i.e.*, fluorescence intensity) can be easily obtained. However, they do not offer the visualization of the flow cytometric output as a whole. In fig. 1 we show that analyzing 6 antigens to define T cell differentiation state can lead to the identification of dozens of subsets of T cells that could follow specific and different dynamics under physiological or pathological conditions (1). Thus, careful analysis of these complex phenotypes by this approach can be time-consuming and may lead to the loss of important information. In this case, more sophisticated analysis are needed to better understand the dynamics of the data under consideration.

SPICE, a useful tool for the analysis of the antigen-specific immune response

Antigen-specific T cell response is a hot topic in T cell immunology since many years. Historically, to investigate the functionality and specificity of cytotoxic cells, researchers were using assays based on the release of radioactive elements (typically, ^{51}Cr) by target cells that were previously loaded with such tracer; the specificity of lymphocytes was investigated by limiting dilution assays, or by the analysis of cell proliferation measured by the incorporation of radioactive DNA precursors (typically, ^{3}H-thymidine) in growing cells. A revolution occurred in 1996 when Altman and colleagues described the use of MHC class I tetramers bearing a peptide for the analysis of antigen-specific CD8+T cells (9). Despite its great utility, this method does not directly address the function of these cells. Roughly at the same time, Louis Picker developed a method for the direct ex-vivo analysis of antigen-specific T cells by the intracellular analysis of cytokine production after stimulation with antigen or antigenic peptides (10). In this report, several cytokines were analysed in the same samples but not combined in the same assay, thus not giving any information about the heterogeneity of the immune response. The advent of the so-called polychromatic flow cytometry, that is advanced instrumentation and reagent development, allowed the simultaneous determination of 5 functions (IFN-γ, TNF-α, IL-2, IL-4 and MIP-1β) and demonstrated for the first time the complex heterogeneity of the antigen-specific T cell response (11). The software SPICE was developed at the VRC specifically to analyze the considerable heterogeneity of cell populations (n=31) that arise from this kind of analysis. In particular, SPICE is ideal for exploring (and quantitatively comparing) the functional and phenotypic profiles of subsets within a complex mixture.

Major guidelines for using SPICE

Originally, when intracellular cytokine staining (ICS) technique was first introduced, researchers measured only one function, *i.e.*, IFN-γ, as a surrogate marker for any T cell response. IFN-γ was thought to be the dominant immune function of the effector T cell response and many candidate vaccines moved to phase I-II because of their capability to induce a strong production of this cytokine. However, simultaneous staining of two cytokines at the same time, *e.g.*, IFN-γ and IL-2, revealed that T cell are heterogeneous, as

single-cytokine producing cells as well as double producers are present (12). The authors coined the word “polyfunctional” as referred to cells that were able to produce both cytokines at the same time. The first assessment of 5 cytokines simultaneously revealed that both CD4+ and CD8+ T cells were even more heterogeneous and multiple subsets of T cells with different functionality could be identified according to the cytokine-producing potential (11). From here the idea to brake down the immune response by identifying all combinations of cytokine-producing T cells. As we include more and more parameters, the number of possible combinations (2^n, where n is the number of parameters) grows exponentially. Thus, computer assistance is needed to simplify the evaluation of these data sets; it is this need that SPICE fills.

SPICE can also be used for multiple types of measurements that are not necessarily flow cytometry-derived. However, in this section, we will describe the most common and popular use, that is the analysis of the functionality of the immune response. In this specific case, the immune response, as measured by the simultaneous analysis of cytokine expression by antigen-stimulated cells, can be broken down in multiple populations, where each population is the result of the combination of positive or negative expression of single cytokines. A Boolean gating approach, such as that contained in FlowJo software (Tree Star Inc., Ashland, OR), can be used to automatically generate these combinations. Generally, data refer to a parent population such as cytokine-producing CD4+ and CD8+ T cells. If differentiation markers are included in the analysis, cytokine-producing cells can also be interrogated for their differentiation phenotype (central vs. effector vs. terminally differentiated memory cells, as an example). Data are exported and easily formatted by the support dedicated software Pestle (VRC, NIH). Pestle also allows the automatic subtraction of background cytokine response (the cytokine response in the unstimulated sample) from the antigen-induced response (the response in the stimulated sample). Within SPICE, variables are organized as *categorical* or as *value*. A *category* is used to organize the immune response data (time after vaccination, type of therapy, vaccination strategy, etc.), while *value* is the variable (*e.g.*, combination of cytokine expression) whose trend has to be analyzed. The power of SPICE comes in its organization of categories to allow for rapid isolation of a specific subset or comparison of different subsets for one or more of the measurement variables.

A wonderful example of SPICE application to the analysis of the immune response can be found in Betts *et al.* (13). Betts and coworkers used polychromatic flow cytometry to analyze the simultaneous production of IFN-γ, TNF-α, IL-2, MIP-1β and the degranulation marker CD107a and showed that highly polyfunctional HIV-specific CD8+ T cells (*i.e.*, cells producing 4 or 5 functional markers simultaneously) were more frequent in HIV long-term non-progressors than in progressors.

Pro and cons of SPICE

As already explained above, SPICE is very useful to simplify the visualization of complex data sets. Many variables, such as those resulting from the combination of positive and negative expression of antigens, can be overlaid in a single graph and grouped depending on the categories utilized. A picture of the whole trend in the data can thus be obtained very quickly. An example of the SPICE output is shown in Fig. 2, where pies show the differences in the composition of the peripheral CD4+ T cell pool between a group of healthy donors and patients with HIV infection. However, SPICE, as obvious, is totally dependent on the gating strategy applied and, as a consequence, does not offer an unsupervised method of analysis of the flow cytometric data.

Probability Binning and Frequency Difference Gating

An important analysis element in flow cytometry is the identification of differences between samples – *e.g.*, treatment responses, genetic variability, etc. Historically, such differences have been quantified by changes the representation of a gated subset (or, perhaps, phenotypic differences identified by median fluorescence intensity changes). Direct comparison of the raw data, however, is desirable under many circumstances: it does not presume identification of specific subsets through subjective (manual) gating; it may identify subtle changes in subsets that are not apparent within grossly-defined gates; and it should be able to take into account all measurement parameters. Historically, the only well-characterized algorithms to do this comparison operated on univariate data, by essentially doing histogram subtraction. These include the Komogorovs-Smirnoff (K-S) statistic (14,15), Overton subraction (16), SED (17), and some limited parametric models (18,19), which compare the density of events at each position within an intensity histogram and report a statistical significance associated with the difference. The primary problem is that these algorithms are typically far too sensitive (identifying differences among samples that are not biologically different). Furthermore, extension of these algorithms to multiparameter data is not easily done.

These nonparametric algorithms estimate event density in the fluorescence distribution by "binning" the data into equal-sized bins. A typical univariate histogram is thereby reduced to 256 bins; the number of events in each bin is compared between a test and experimental sample. If this were extended to two dimensions, the number of bins would be 256 × 256, or more than 65,000; with 8 dimensions, the number of bins would be 1.8×10^{19}. Even if it were possible with today's computers to compare across this many bins, it should be noted that one could never collect enough events so as to populate them sufficiently to do a comparison across samples. Clearly, the number of bins can be reduce by dividing each measurement into fewer divisions – but even dividing each measurement into 4 would result in over 65,000 bins. Each bin would cover one-fourth of the measurement space for each parameter, making it insensitive to subtle changes.

This problem was solved by the use of "probability binning" (PB) to dissect multidimensional space (20,21). PB creates small bins where there are many events (and therefore much more information about event density), and big bins where there are few events. Indeed, in the final binning, each bin contains the *same* number of events. Therefore, when comparing frequency distributions between samples, this algorithm minimizes the maximum expected variance, with a result that the statistical power is as robust as possible. A significant advantage of the PB algorithm is that computation time does not significantly increase with the number of parameters.

PB provides a statistic to compare different multivariate distributions. In addition, it can be used to rank the "distance" between different multivariate distributions. In a test example, PB was used on 4-color immunophenotyping of B cells from identically aged, genetically disparate strains of mice. The difference between littermates was negligible; the difference between each strain and the F1 cross was far less than the difference between the parental strains, as expected. This type of analysis can be used to identify genetic elements controlling leukocyte homeostasis – without *a priori* knowledge of the subsets present.

A useful outcome of the PB algorithm is the identification of which of the multidimensional bins are the ones with the biggest differences between the two samples. By grouping these "most-different" bins together, a gate can be constructed that selectively identifies the region of multidimensional space which has the biggest differences between two samples. This gate can then be applied to a large set of samples – the frequency of events within this gate will

presumably be related to how closely related a sample is to either the control or experimental sample. This process is termed “Frequency Difference Gating” (FDG) – *i.e.*, the definition of a gate based on the maximal difference in the frequency of events between two samples (22). The utility of FDG was shown by comparing B cell distributions in 6 and 8-week old mice. PB identifies subtle changes between these distributions. FDG creates a gate that identifies the regions with maximal variance; this gate can be applied to many samples to both quantify the differences between samples but also to further characterize the populations that are different.

Pros and cons of PB and FDG

A major advantage of these algorithms is that they scale to highly multidimensional data without significant impact on computation time. They are objective methods, and provide a robust statistic that can be used to compare distributions across samples. Finally, the ability to generate a gate representing the “most” different regions is a powerful tool to explore and quantify distributions. It is quite different from “cluster” analyses, in that it makes no assumption about the separation of those events from a main population – and thereby can identify shifts in the fluorescence distribution rather than wholesale changes in subsets.

A major disadvantage of these algorithms, as currently implemented, is that they are highly sensitive to variations in instrument settings and sensitivity. Apparent changes in fluorescence intensity can be biological or simply mildly different experimental staining conditions or instrument settings. While the latter can be controlled with very careful adherence to protocol, it may not always be possible. Future algorithms will hopefully incorporate fluorescence normalization techniques that attempt to remove experimental and instrumental variability in order to best reveal biological variability. As such, the most robust comparisons are derived from samples that are processed and stained in parallel, and analyzed sequentially on the same instrument.

Cluster analysis

Heat maps are a relatively simple and intuitive way to simultaneously visualize the trend of multiple variables following experimental perturbation. They are known since the late 1950's and became very popular in biology after the introduction of microarray technology. A color code is used to indicate the relative expression of a variable compared to a control sample: generally, black for unchanged, green for downregulated and red for upregulated, as proposed by Eisen and colleagues (23). Heat maps can also be coupled to clustering, whose algorithms are able to group variables on the basis of similarities and differences. In this way, patterns in the data or in patient cohorts can be discerned by an unsupervised bioinformatic approach. For a nice overview about clustering algorithms applied to flow cytometry see Bashashaty and Brinkman (24).

Recently, some groups used hierarchical clustering to distinguish different types of hematological malignancies or to determine prognosis in B-CLL on the basis of the relative expression (percentage or MFI) of surface antigens commonly used for cancer cell differentiation/classification (25,26). Despite their great usefulness in translational medicine, the outputs obtained with these approaches were based on the expression of individual markers and no information was derived from the mutual expression patterns of different surface proteins or on the complexity of cancer cell phenotypes.

Several reports describe the use of cluster analysis (CA) applied to raw flow cytometric data as an approach to group cells with similar fluorescence patterns (27,28). This is done to overcome the subjectivity of manual gating, as well as to identify all possible cell populations contained in a given dataset. K-means clustering was proposed two decades ago

as an unsupervised method to identify populations (27,28). In the K-means approach, the number of clusters in which the data will be catalogued has to be predetermined, thus rendering difficult the clustering of non-naturally partitioned data. Moreover, each data point (event, in this case) can belong to only one cluster. A modified version of the K-means, termed fuzzy K-means, allows each point to belong to multiple clusters with different grades of association (29). Clustering of the raw data could also be done hierarchically, as it is done for gene array studies. However, the considerable number of events that are generally acquired at the flow cytometer renders this approach untenable in terms of processing time; thus previous partitioning (gating) of the data should be considered. This is especially true when multiple parameters are analyzed at the same time, as it is in polychromatic flow cytometry.

A specific approach of data processing before CA has been proposed by a number of groups, *i.e.*, Boolean combination of gates identifying positive and negative expression of antigens. In 2006, two reports described this approach prior to CA and identified the most common lymphocyte populations in different organs of the mice or the differentiation pattern of T cells stimulated with a combination of cytokines (30,31). In the same way, Kitsos and colleagues determined the cell state in response to stimuli (32). These authors stimulated HL-60 cells with different combinations of external stimuli at various concentrations and measured cell differentiation (through the analysis of lineage markers by flow cytometry) or apoptosis and described unusual patterns of differentiation that otherwise could have been missed by classical analyses.

Shortly thereafter, we used polychromatic (8-color) flow cytometry and hierarchical clustering to classify people of different ages, *i.e.*, young (20 year old), middle age (60 year old) adults, and centenarians, by considering the T cell flow cytometric profile as a whole (33). Indeed, by combining the expression of CD45RA, CCR7, CD95, CD127 and CD38, we described 48 subpopulations on both CD4+ and CD8+ T cells as obtained by Boolean combination of gates. Besides the well-known observation that memory T cells accumulate with age concomitant with the depletion of naïve T cells, the CA identified those subpopulations that were expanded in the centenarian cohort compared to other cohorts (and vice versa), thereby revealing dynamics of T cell phenotypes that occur with ageing. Subject classification was, in this case, driven by all 48 phenotypes identified by the Boolean combination of gates. Interestingly, we found the above cohorts clustered much better on the basis of the CD8+ rather than the CD4+ T cell phenotype, thus indicating the former as a more homogeneous population than the latter.

Another report used CA to analyze polychromatic flow cytometric data obtained from profiling of healthy donors and ankylosing spondylitis (AS) patients (34). Four different panels were used to describe multiple leukocytes populations including granulocytes, CD4+ and CD8+ T cells, NK cells, B cells and monocytes and up to 80 variables were included in the data set and used for the subjects' classification. Healthy donors and AS patients nicely segregated in two different clusters on the basis of the flow cytometric profile and the authors identified those populations that were specifically expanded or contracted in the two cohorts.

Principal component analysis

Because of bivariate plots, flow cytometric data are generally displayed two dimensions at a time. However, as discussed before, they can be imagined in a multi-dimensional space where each dimension corresponds to single fluorescent parameters or scatters. If we think that flow cytometry allows analysis at the level of single cells, a huge amount of data can thus be generated by a single file. Similarly, a single flow cytometric profile can generate

multiple variables as a result of the Boolean combination of gating, as described above. Thus, each subject, patient or sample can be projected onto an n-dimensional space where each dimension is represented by a combination of antigens. Since it is not possible to display such data in a multi-dimensional space, we adopted the use of Principal Component Analysis (PCA) to reduce the multidimensionality of the data set (33). PCA is an unsupervised dimension-reduction method that generates a new set of decorrelated variables (called principal components) as linear combinations of the original variables (in our case, represented by T cell subsets-combination of phenotypes) (35). The majority of the variation of flow cytometric datasets (subjects) can be captured by the most dominant principal components that become the new axes in a two or three dimensional representation. The loss of information occurring with this transformation is minimal and allows the classification of experimental samples by considering the flow cytometric output as a whole. The use of PCA in flow cytometry was not new as it was proposed for the first time in 1984 and again in 1987 (36,37). Both reports aimed to reduce the multidimensionality of the raw flow cytometric data. In particular, Kosugi *et al.* (37) were able to identify four different clusters of cells in a bivariate plot resulting from the PCA of cells identified by four different parameters, that is two fluorescent parameters (cells positive for CD3 and CD8 surface molecules) and forward (FSC) and side (SSC) scatters. Two of these populations were distinguishable as expressing the CD3 and CD8 antigens while the remaining two were negative for CD3 and CD8 and were distinguishable by the different scatter properties (cells with low FSC, presumably dead cells, and monocytes, which have higher FSC compared to any other cell population after Ficoll isolation).

Differently, we applied the PCA to the data set described in the previous section. In our case, we simply assumed that not all phenotypes (combination of antigens) contributed in the same way to the definition of the data set as some of them do not change with immunological ageing. Similarly to CA, PCA grouped subjects of different age and identified the variables that were contributing the most to the definition of the cohort. Moreover, PCA was able to identify the specific phenotype(s) that was (were) best describing a certain subject. In this way, “immunological ageing” of an individual could be determined if phenotyped for the same markers and then inserted into the data set as a test, independent sample. Our data thus demonstrated that the flow cytometric profile considered in its entirety could be a useful tool to classify subjects on the basis of phenotypic characteristics. This approach not only has implications for research purposes as it can identify the major phenotypes in a certain cohort of individuals but could also help diagnosis as it can identify similarities and differences among unknown phenotypes.

Recently, Kalina *et al.* (38) phenotyped B cells from healthy donors and common variable immunodeficiency patients by 6-colour flow cytometry and subsequently applied the probability binning algorithm. The coordinated of “overfed” or “underfed” bins' means were then subjected to PCA to identify the portions of the six-parameter space containing the “overfed” or “underfed” bins. Interestingly, they found that a specific cell population was present in the PCA plot from patients but absent in the one obtained from in the healthy donors. After gating, they showed that this population represented CD27- CD24bright CD38 bright CD19+ transitional B cells which were specifically expanded in the immunodeficient patients.

Pro and cons of CA and PCA

The aforementioned approaches can be very useful to simplify data analysis in polychromatic flow cytometry experiments. In particular they can: i) identify trends in the data that otherwise could be missed by classical approaches simply because the amount of generated variables is very high (CA); ii) group subjects or patients on the basis of the whole

flow cytometric profile (CA and PCA); iii) identify the differentially represented phenotypes among cohorts of donors or patients (CA and PCA).

CA and PCA are explorative, not quantitative types of analysis. Thus, they should be used as means to identify similarities and differences among groups. Moreover, multiple algorithms for CA are available but not all of them are suitable for flow cytometric data. For example, the previously cited K-means or fuzzy K-means only cluster data with spherical distribution, which usually does not occur in flow cytometry. Thus, recognition of discrete populations can be hardly possible with certain types of data. Clustering of the raw data can also be hierarchical but the number of events acquired for each sample can limit this performance in terms of processing time.

Identifying trends in the data is the most attractive capability of these approaches. However, trends need to be experimentally tested to prove a real biological effect. Nevertheless, huge data sets can be used to build models and to predict cellular behaviour, as proposed by Janes *et al.* (39). These authors measured multiple pathways of intracellular signaling in response to different combinations and concentrations of external stimuli including TNF-α, insulin and EGF and determined cellular apoptotic phenotype by measuring multiple apoptotic parameters. A model based on principal component analysis and partial least square regression combined stimulation and intracellular signaling and was able to predict the cellular response, *i.e.*, apoptosis, with high accuracy. The same model also revealed connections among previously unrelated external stimuli in determining the survival or the death of the target cell.

Conclusions and future directions

We have reviewed some of the new developments in the field of data analysis in polychromatic flow cytometry. The huge data sets generated by this technology can be overwhelming (*e.g.*, consider that a 10 colour staining allows the recognition of 1,024 different cell populations, not to mention possible differences in FSC and SSC) and some useful aspects can be lost or ignored if the proper analysis approach is not adopted. Certainly, the future in this field is the use of automatic tools of data analysis that can identify cell populations and possibly underline their relative importance. We believe that manual gating strategies will be largely reconsidered and modified over the next few years as robust and reliable methods of automated gating are being proposed (40). In particular, unsupervised approaches applied to raw data, such as CA and PCA (Table 1), will become more and more popular, as they consider the fluorescence values of each single cells and do not presume the previous partition of the data, *e.g.*, distinction between positive and negative expression. The main goal of these automatic tools of analysis will become the identification of similarities and differences among samples. Certainly, an algorithm cannot substitute the expertise of the operator in data analysis, especially when specific populations of interest need to be identified. However, different applications such as high-throughput and high-content screening assays for compounds nowadays require the fast extrapolation of information (41). Especially in these disciplines, it is necessary to obtain data that go beyond a mere positive and negative result.

Analysis of multiple parameters at the same time can be only beneficial for unraveling the mechanism of action of specific compounds. For instance, the analysis of multiple reactive oxygen species in the same cell revealed that quercetin, a flavonoid known as antioxidant, can exert pro-oxidant functions in certain cellular systems by generating the superoxide anion O_2-(42). The same approach was used to describe the loss of glutathione and the increase in oxygen free radicals in adipocytes treated with antiretroviral drugs, such as stavudine (43). We expect that multiparameter flow cytometric assays for compound

screening will be largely utilized in the future. As a consequence, rapid and simple methods of analysis are absolutely required.

The same scenario can be envisaged for the role of flow cytometry in the diagnosis of hematological malignancies. More and more markers are now combined to better characterize malignant cells; automated analytical tools applied to multiparameter data set will allow us to better define the type of disease, its stage and the progression rate. Computers and algorithms combined with advanced operator expertise and instrument standardization will certainly improve our knowledge in basic and translational research.

References

1. De Rosa SC, Herzenberg LA, Roederer M. 11-color, 13-parameter flow cytometry: identification of human naive T cells by phenotype, function, and T-cell receptor diversity. Nat Med. 2001; 7(2): 245–8. [PubMed: 11175858]
2. Chattopadhyay PK, Price DA, Harper TF, Betts MR, Yu J, Gostick E, Perfetto SP, Goepfert P, Koup RA, De Rosa SC, et al. Quantum dot semiconductor nanocrystals for immunophenotyping by polychromatic flow cytometry. Nat Med. 2006; 12(8):972–7. [PubMed: 16862156]
3. Suni MA, Picker LJ, Maino VC. Detection of antigen-specific T cell cytokine expression in whole blood by flow cytometry. J Immunol Methods. 1998; 212(1):89–98. [PubMed: 9671156]
4. Perez OD, Nolan GP. Simultaneous measurement of multiple active kinase states using polychromatic flow cytometry. Nat Biotechnol. 2002; 20(2):155–62. [PubMed: 11821861]
5. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP. Causal protein-signaling networks derived from multiparameter single-cell data. Science. 2005; 308(5721):523–9. [PubMed: 15845847]
6. Le Meur N, Rossini A, Gasparetto M, Smith C, Brinkman RR, Gentleman R. Data quality assessment of ungated flow cytometry data in high throughput experiments. Cytometry A. 2007; 71(6):393–403. [PubMed: 17366638]
7. Lugli E, Troiano L, Cossarizza A. Investigating T cells by polychromatic flow cytometry. Methods Mol Biol. 2009; 514:47–63. [PubMed: 19048213]
8. Herzenberg LA, Tung J, Moore WA, Parks DR. Interpreting flow cytometry data: a guide for the perplexed. Nat Immunol. 2006; 7(7):681–5. [PubMed: 16785881]
9. Altman JD, Moss PA, Goulder PJ, Barouch DH, McHeyzer-Williams MG, Bell JI, McMichael AJ, Davis MM. Phenotypic analysis of antigen-specific T lymphocytes. Science. 1996; 274(5284):94–6. [PubMed: 8810254]
10. Picker LJ, Singh MK, Zdraveski Z, Treer JR, Waldrop SL, Bergstresser PR, Maino VC. Direct demonstration of cytokine synthesis heterogeneity among human memory/effector T cells by flow cytometry. Blood. 1995; 86(4):1408–19. [PubMed: 7632949]
11. De Rosa SC, Lu FX, Yu J, Perfetto SP, Falloon J, Moser S, Evans TG, Koup R, Miller CJ, Roederer M. Vaccination in humans generates broad T cell cytokine responses. J Immunol. 2004; 173(9):5372–80. [PubMed: 15494483]
12. Harari A, Vallelian F, Meylan PR, Pantaleo G. Functional heterogeneity of memory CD4 T cell responses in different conditions of antigen exposure and persistence. J Immunol. 2005; 174(2): 1037–45. [PubMed: 15634928]
13. Betts MR, Nason MC, West SM, De Rosa SC, Migueles SA, Abraham J, Lederman MM, Benito JM, Goepfert PA, Connors M, et al. HIV nonprogressors preferentially maintain highly functional HIV-specific CD8+ T cells. Blood. 2006; 107(12):4781–9. [PubMed: 16467198]
14. Finch PD. Substantive difference and the analysis of histograms from very large samples. J Histochem Cytochem. 1979; 27(3):800. [PubMed: 479554]
15. Young IT. Proof without prejudice: use of the Kolmogorov-Smirnov test for the analysis of histograms from flow systems and other sources. J Histochem Cytochem. 1977; 25(7):935–41. [PubMed: 894009]
16. Overton WR. Modified histogram subtraction technique for analysis of flow cytometry data. Cytometry. 1988; 9(6):619–26. [PubMed: 3061754]

17. Bagwell C. A journey through flow cytometric immunofluorescence analyses. Clin Immunol Newsletter. 1996; (16):33–37.

18. Lampariello F. Evaluation of the number of positive cells from flow cytometric immunoassays by mathematical modeling of cellular autofluorescence. Cytometry. 1994; 15(4):294–301. [PubMed: 8026220]

19. Lampariello F, Aiello A. Complete mathematical modeling method for the analysis of immunofluorescence distributions composed of negative and weakly positive cells. Cytometry. 1998; 32(3):241–54. [PubMed: 9667514]

20. Roederer M, Moore W, Treister A, Hardy RR, Herzenberg LA. Probability binning comparison: a metric for quantitating multivariate distribution differences. Cytometry. 2001; 45(1):47–55. [PubMed: 11598946]

21. Roederer M, Treister A, Moore W, Herzenberg LA. Probability binning comparison: a metric for quantitating univariate distribution differences. Cytometry. 2001; 45(1):37–46. [PubMed: 11598945]

22. Roederer M, Hardy RR. Frequency difference gating: a multivariate method for identifying subsets that differ between samples. Cytometry. 2001; 45(1):56–64. [PubMed: 11598947]

23. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. 1998; 95(25):14863–8. [PubMed: 9843981]

24. Bashashaty A, Brinkman R. A Survey of Flow Cytometry Data Analysis Methods. Advances in Bioinformatics. 2009

25. Maynadie M, Picard F, Husson B, Chatelain B, Cornet Y, Le Roux G, Campos L, Dromelet A, Lepelley P, Jouault H, et al. Immunophenotypic clustering of myelodysplastic syndromes. Blood. 2002; 100(7):2349–56. [PubMed: 12239142]

26. Zucchetto A, Sonego P, Degan M, Bomben R, Dal Bo M, Russo S, Attadia V, Rupolo M, Buccisano F, Del Principe MI, et al. Signature of B-CLL with different prognosis by Shrunken centroids of surface antigen expression profiling. J Cell Physiol. 2005; 204(1):113–23. [PubMed: 15605425]

27. Murphy RF. Automated identification of subpopulations in flow cytometric list mode data using cluster analysis. Cytometry. 1985; 6(4):302–9. [PubMed: 4017796]

28. Mann RC. On multiparameter data analysis in flow cytometry. Cytometry. 1987; 8(2):184–9. [PubMed: 3582064]

29. Wilkins MF, Hardy SA, Boddy L, Morris CW. Comparison of five clustering algorithms to classify phytoplankton from flow cytometry data. Cytometry. 2001; 44(3):210–7. [PubMed: 11429771]

30. Hofmann M, Zerwes HG. Identification of organ-specific T cell populations by analysis of multiparameter flow cytometry data using DNA-chip analysis software. Cytometry A. 2006; 69A(6):533–40. [PubMed: 16646049]

31. Petrausch U, Haley D, Miller W, Floyd K, Urba WJ, Walker E. Polychromatic flow cytometry: a rapid method for the reduction and analysis of complex multiparameter data. Cytometry A. 2006; 69A(12):1162–73. [PubMed: 17089357]

32. Kitsos CM, Bhamidipati P, Melnikova I, Cash EP, McNulty C, Furman J, Cima MJ, Levinson D. Combination of automated high throughput platforms, flow cytometry, and hierarchical clustering to detect cell state. Cytometry A. 2007; 71A(1):16–27. [PubMed: 17211881]

33. Lugli E, Pinti M, Nasi M, Troiano L, Ferraresi R, Mussi C, Salvioli G, Patsekin V, Robinson JP, Durante C, et al. Subject classification obtained by cluster analysis and principal component analysis applied to flow cytometric data. Cytometry A. 2007; 71A(5):334–44. [PubMed: 17352421]

34. Steinbrich-Zollner M, Grun JR, Kaiser T, Biesen R, Raba K, Wu P, Thiel A, Rudwaleit M, Sieper J, Burmester GR, et al. From transcriptome to cytome: integrating cytometric profiling, multivariate cluster, and prediction analyses for a phenotypical classification of inflammatory diseases. Cytometry A. 2008; 73A(4):333–40. [PubMed: 18307258]

35. Wold S, Esbensen K, Geladi P. Principal component analysis. Chemom Intell Lab Syst. 1987; 2:37–52.

36. Mann RC, Popp DM, Hand RE Jr. The use of projections for dimensionality reduction of flow cytometric data. Cytometry. 1984; 5(3):304–7. [PubMed: 6734355]

37. Kosugi Y, Sato R, Genka S, Shitara N, Takakura K. An interactive multivariate analysis of FCM data. Cytometry. 1988; 9(4):405–8. [PubMed: 3261233]

38. Kalina T, Stuchly J, Janda A, Hrusak O, Ruzickova S, Sediva A, Litzman J, Vlkova M. Profiling of polychromatic flow cytometry data on B-cells reveals patients' clusters in common variable immunodeficiency. Cytometry A. 2009; 75A(11):902–9. [PubMed: 19802875]

39. Janes KA, Albeck JG, Gaudet S, Sorger PK, Lauffenburger DA, Yaffe MB. A systems model of signaling identifies a molecular basis set for cytokine-induced apoptosis. Science. 2005; 310(5754):1646–53. [PubMed: 16339439]

40. Lo K, Brinkman RR, Gottardo R. Automated gating of flow cytometry data via robust model-based clustering. Cytometry A. 2008; 73A(4):321–32. [PubMed: 18307272]

41. Naumann U, Wand MP. Automation in high-content flow cytometry screening. Cytometry A. 2009; 75A(9):789–97. [PubMed: 19548208]

42. Ferraresi R, Troiano L, Roat E, Lugli E, Nemes E, Nasi M, Pinti M, Fernandez MI, Cooper EL, Cossarizza A. Essential requirement of reduced glutathione (GSH) for the anti-oxidant effect of the flavonoid quercetin. Free Radic Res. 2005; 39(11):1249–58. [PubMed: 16298752]

43. Cossarizza A, Ferraresi R, Troiano L, Roat E, Gibellini L, Bertoncelli L, Nasi M, Pinti M. Simultaneous analysis of reactive oxygen species and reduced glutathione content in living cells by polychromatic flow cytometry. Nat Protoc. 2009; 4(12):1790–7. [PubMed: 20010930]

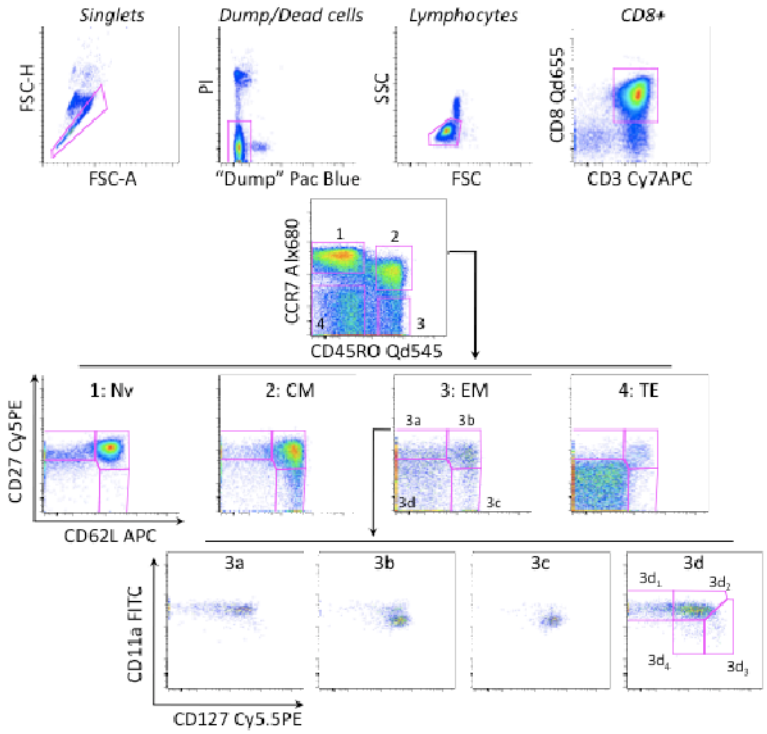

Fig. 1. A simple 6 differentiation-antigen staining can identify dozens of subsets of CD8+ T cells

Magnetically-enriched CD8+ peripheral blood mononuclear cells from a healthy donor were stained with multiple fluorescently-conjugated monoclonal antibodies directed to CD45RO, CCR7, CD62L, CD27, CD127 and CD11a to determine T cell differentiation state. Cells were acquired with a modified FACSAria (BD, San Josè, CA) able to detect 20 parameters (Details on the machine configuration can be found at: http://www3.niaid.nih.gov/labs/aboutlabs/VRC/flowCytometryCoreLaboratory/). Singlets were selected on the basis of FSC-A and FSC-H. Monocytes, B cells and dead cells were excluded by gating on Dump- (CD14-/CD19-) and PI- cells. CD8+ T cells were further selected for CD3 and CD8 positivity. Naïve (Nv; 1), Central Memory (CM; 2), Effector Memory (EM; 3) and Terminal Effectors (TE; 4) were defined on the basis of CD45RO and CCR7 expression. Within these populations, four different subsets (a, b, c, d) can be further defined by the expression of CD62L and CD27. As an example, subsequent analysis of CD127 and CD11a expression in the CD62L-,CD27- EM population can identify four more subsets. The same type of sequential analysis can be applied to all populations in a hierarchical way, thus leading to the hypothetic identification of 64 subsets. Data were compensated and analyzed with FlowJo version 9 (Treestar, Ashland, OR, USA). PI: Propidium Iodide; Pac Blue: Pacific Blue; Qd: Quantum Dot; Alx: Alexa.

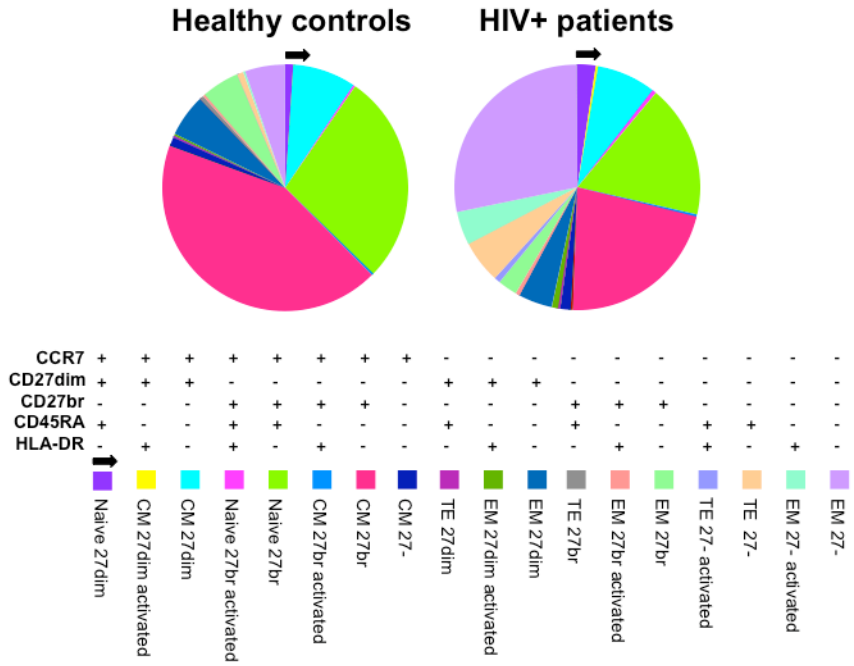

Figure 2. Use of SPICE for the analysis of differentiation and activation of peripheral CD4+ T cells

Another manner of representing data related to the expression of differentiation and activation markers is shown in this Figure. The expression of CCR7, CD45RA, CD27, and HLA-DR has been investigated in living CD3+,CD4+,CD8- T cells from healthy donors (whose mean CD4+ T cell count in peripheral blood was 45%, with 1,109 cells/μL) and patients with HIV infection who were out of treatment (mean CD4+ T cell count 21%, 480 cells/μL). Positive and negative expression of antigen [for CD27 relative expression was further distinguished beween dim (dim) and bright (br)] were combined by Boolean gating to generate all possible subsets. Each colour in the pie corresponds to a specific combination of antigens indicated in the bottom part of the figure. CD45RA and CCR7 were used to define Naïve (CD45RA+CCR7+), CM (CD45RA-CCR7+), EM (CD45RA-CCR7-) and TE (CD45RA+CCR7). Note the striking difference in the composition of CD4+ T cell population. The black arrows indicate the way of reading the colours (clockwise in the upper part, left to right in the lower part).

Table 1

Summary of the applications, pro and cons of the data analysis approaches discussed in this article

Analysis approach	Common Applications	Pro	Cons
Sequential Gating [7,8]	Standard analysis of flow data: identifies subset of cells at operator discretion	• Simple and intuitive • Used when a specific population has to be identified	• Allows the visualization of only two parameters at a time • Gates are dependent on the subjectivity of the operator
SPICE[11,13]	Simplified visualization of complex and big data sets. Suitable for the analysis of the: • antigen-specific immune response • antigen expression across multiple cellular subsets	• Rapid and simplified visualization of the trends in the data across multiple experimental samples and time points • Quantitative: allows comparison and statistical analysis	• Dependent on the sequential gating strategy • Supervised approach
Probability Binning [20, 21]	Applied to raw flow cytometric data: identifies differences in the distribution of events in a multidimensional space between a control and (a) test sample(s)	• Unsupervised • Multidimensional • Quantitative (provide robust statistics)	• Sensitive to variations in instrument settings and sensitivity
Frequency Difference Gating [22]	Allows multidimensional gating of the bins that contain the biggest differences, as identified by the PB algorithm	• Unsupervised • Multidimensional	• Sensitive to variations in instrument settings and sensitivity
Cluster Analysis (CA) [25-27,29-34,40]	Analysis of subpopulations generated by the combination of positive and negative expression of antigens	• The flow cytometric output set can be visualized as a whole • Identifies trends in the data • Semi-quantitative (population dynamics is depicted by a colour-coded heat map)	• Explorative (but semi-quantitative) approach • Results need to be biologically tested
Principal Component Analysis (PCA) [33,35-37]	Analysis of subpopulations generated by the combination of positive and negative expression of antigens	• 2D representation of multi-dimensional data sets • Identifies trends in the data (subset dynamics)	• Explorative approach • Results need to be biologically tested
Single cell CA/ PCA [38]	Analysis of raw flow cytometric data to identify: • clusters of cells with similar fluorescence patterns • trends in the raw data (if multiple samples are compared)	• Unsupervised approach • Without previous partitioning of the data, cells are grouped or distributed on the basis of their fluorescence pattern	• Explorative approach • Results need to be biologically tested • Not all cluster algorithms are suitable