Alternative Approaches for Analysis of Complex Data Sets in Flow Cytometry

Carmen Gondhalekar

Abstract The field of flow cytometry is rapidly evolving as a result of advancements in sample preparation technology. Automated machinery, multi-channel PMTs, multiplexed assay formats, and mass cytometry introduce new opportunities in fields utilizing flow cytometry by creating very rich, albeit complex, data sets. Advances in cytometric data collection have always been accompanied by advances in analytical techniques. As high-content multi-parametric data sets are generated through new flow cytometry techniques and technologies, it is important to simultaneously develop computers and software with the ability to execute six major functions: manipulate an unlimited number of parameters, compare all data parameters in many different combinations, create and customize a data analysis strategy, see real-time changes in all parameters, offer advanced statistical tools, and rapidly report figures and graphs with endpoint conclusions. This chapter investigates methods of achieving each of these six functions by addressing data-analysis challenges in drug-screen assays, multiplexed cytokine assays, hyperspectral cytometry, and mass cytometry.

Keywords Data analysis · Software · Mass cytometry · Hyperspectral cytometry · Drug screens · Multiplexed assays · PlateAnalyzer · Cytospec

1 Introduction

For three decades, flow cytometry data analysis was focused on one to three parameters, then advanced to five to ten and subsequently up to twenty parameters. As the technology improved, data processing also increased in complexity to the point that we have shifted into a new age of research where many parameters can be

C. Gondhalekar (✉)
Weldon School of Biomedical Engineering and College of Veterinary Medicine, Room G221, Lynn Hall, 625 Harrison Street, West Lafayette, IN 47907, USA
e-mail: cgondhal@purdue.edu

© Springer Nature Singapore Pte Ltd. 2017

J.P. Robinson and A. Cossarizza (eds.), *Single Cell Analysis*, Series in BioEngineering, DOI 10.1007/978-981-10-4499-1_9

measured on millions of cells in a very short time period. The computational cost of such complexity increases exponentially when mass cytometry is brought into the picture; the total number of parameters can be as high as a hundred for several million cells. Traditional flow-cytometry analytical techniques were not originally designed for very high-content and high-throughput data, prompting us to consider what the next generation of flow-cytometry data analysis should look like.

If we examine the kind of software that is commercially available today, we can identify a number of popular commercial packages, including FCS express, Flow Jo, FACSDiva, Kalooza, and Winlist. While the degree varies among packages, each hosts a more-or-less friendly interface to allow users to navigate quickly between options for loading data, saving protocols, exporting statistics, and visualizing scatter plots and histograms. For file sizes ranging from a couple of thousand events to a couple of million, data loading time can take from a few seconds to thirty seconds. In fact, one of the fundamental changes made in version 2.0 of the FCS file format [1] was the capacity for the collection and analysis of multiple data sets within a listmode file; this was followed by FCS3.0, which accommodated instruments that could collect more than 100 million events [2]—something previously not attainable. This set the scene for more complex and much larger file sizes, which had previously been restricted by both the file standard and the capacity of disc drives to accommodate such large files.

Manipulating data after they have loaded is also dependent on the number of events and parameters, but very large files can cause sluggish execution of the tasks being demanded of conventional software. The way in which the software reads and stores data files largely determines computational speed. In conventional flow cytometry software, a comprehensive read of every FCS file was practical because users typically needed to analyze one FCS file at a time. However, with large quantities of multiparametric data, and the demand to analyze multiple FCS files simultaneously, new approaches to data management are needed. For example, demands in the area of high-throughput cytometry and the emergence of mass cytometry have created new burdens on data processing. Common data-analysis challenges faced by researchers in these fields will be described in the upcoming sections along with a number of software solutions.

Aside from faster computing speed, what else does conventional data analysis need in order to adapt to advancements and diversification in flow cytometry? We identified six aspects of cytometry analysis approaches that we considered critical to the next generation of cellular analysis.

1. Ability to manipulate an unlimited number of parameters
2. Ability to compare all data parameters in many different combinations
3. Ability to create and customize a data analysis strategy
4. Ability to see real-time changes in all parameters
5. Ability to integrate advanced statistical tools
6. Rapid reporting toolsets (figures and graphs) with endpoint conclusions

Why are the above aspects important in cytometry? The answer lies in adaptation to the changing experimental approaches driven by interdisciplinary science. As technology advances impact the field of cytometry, it is necessary to modify the analytical toolsets. For example, with increased parameter space driven by mass cytometry, performing simultaneous analysis of 50–100 parameters is an absolute need [3]. Another example involves the introduction of hyperspectral flow cytometry [4, 5], where there emerged a need to change the way fluorescence signals are analyzed: instead of the signal intensity within a narrow wavelength range being the primary feature as in traditional flow cytometry, the spectrum itself became a parameter. However, full-spectrum analysis could not be handled with traditional software. Third, a need for change in traditional software also emerged hand-in-hand with automated data collection. Along with robotically driven flow cytometry, the ability to collect thousands of samples in an hour changed both file-structure [6] and file-management principles [7]. Fourth, issues are encountered with increasingly complex assays such as multiplex-bead assays, where separate merchant-specific software is required for data analysis. The following discussion addresses the issues of data management, data display, and computing capability in each of the four cases above through two alternative, free data-analysis packages: *CytoSpec* and PlateAnalyzer. Other practical tools include methods of re-naming parameters in FCS files and extracting raw data from FCS files. These two points are addressed through two additional free software packages: FCS Rename and LData.

2 PlateAnalyzer: Drug-Screen Assays

Automation of sample prep and flow-cytometry instrumentation is one of the main drivers for new techniques in data analysis because data collected in processes such as automated drug screens are stored in ever larger and more complex FCS files [8–11]. One example is provided in a 2012 study conducted by Robinson et al. to examine the effects of a panel of drugs on cell toxicity [12]. A 384-well plate was used to test the toxicity of 32 drugs in a 10-step dilution series using three indicators (dyes) of HL-60 functional integrity. This assay design had to take into consideration 960 samples ($32 \times 10 \times 3$; drug, dosage, dye), with each sample representing 5000 cells, a total of 4.8 million events! This is one of many similar studies regularly used in the field of drug development and can be challenging to analyze using traditional analytical software because of the file size and complexity.

Before the use of automated sample-preparation machinery, flow cytometry data analysis generally focused on single-file analysis—samples were prepped and run one at a time in test tubes. Even when using microtiter plates, analytical techniques focused on reproducing the same traditional analysis model of single-sample analysis. The fundamental problem observed with current commercial analysis software is the degree of difficulty in reading 384 listmode files, bringing all the

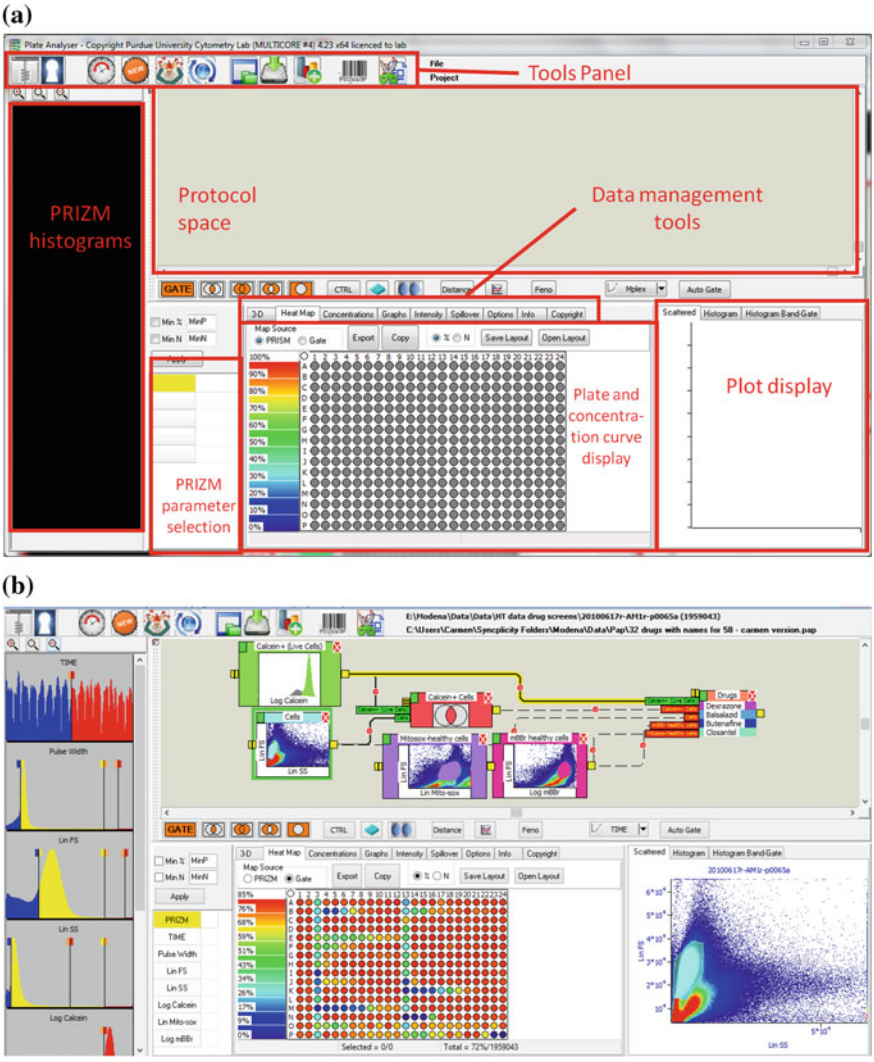

Fig. 1 To accommodate the analysis of complex data sets, PA utilizes methods of opening and closing files similar to those of most analysis software. However, it uploads all FCS files instead of just one. The upload of these files can be visualized in the plate and concentration-curve display. The microtiter-plate map is interactive, allowing the user to select the wells that will be plotted on the plot display. The workspace (**a**) allows the user to design a protocol that applies the gates to FCS files selected through the microtiter-plate display (**b**)—the key to rapid analysis of large data sets. Data provided by Robinson et al. [12]

data together, and reducing the results to dose-response curves (a standard tool for interpreting drug screens). In the example of Robinson's data, the starting process would require gating of light scatter (thus 384 plots), followed by an analytical gate for each of the three assays per well (1152 plots), followed by generation of dose-response curves for each drug in each of the three assays (another 96 plots). As seen in the example above, the bottleneck of the drug-screening process is sample analysis, assuming that one has a mechanism for fast sample preparation (robotics) and collection (also robotics). No commercial software is capable of processing cytometric data in addition to generating IC50 curves. Since the goal of Robinson's assay was to generate dose-response curves, it was necessary to develop a technology that would rapidly perform the required calculations.

For data such as that generated by Robinson et al. in 2012, hours if not days are required to analyze and compare each well of a 384-well plate. The PlateAnalyzer (PA) software was developed specifically to analyze large volumes of complex data quickly and efficiently and fits perfectly into the drug-screening paradigm. The software is not commercial, but is robust and designed to answer the demands of large data set–based flow cytometry.

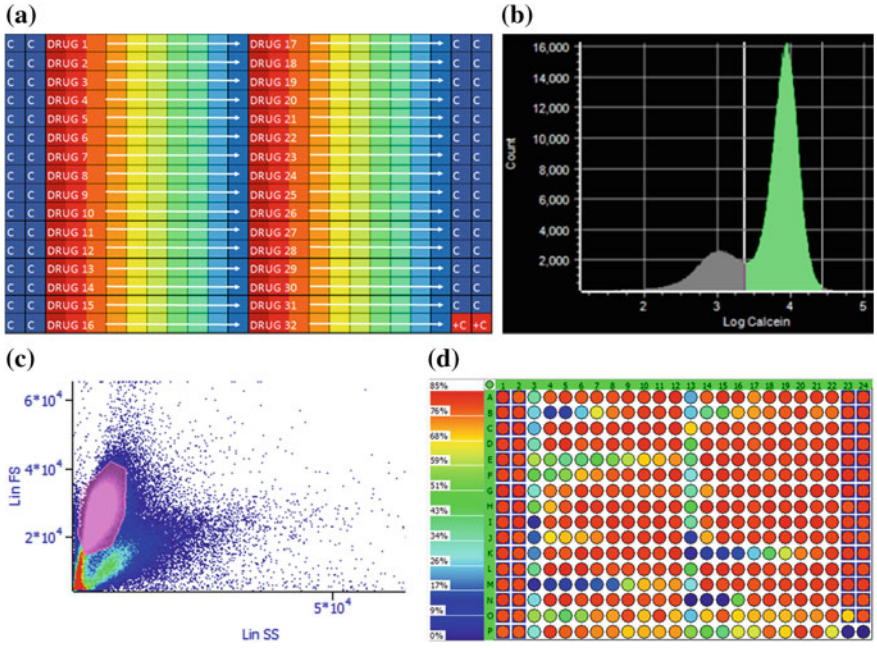

Fig. 2 **a** Drug-screen layout on a 384-well plate (C indicates control, C+ indicates positive control of cells treated with vancomycin); **b** gating live cells can be achieved by selecting cells positive for calcein (one of the three dyes used in the assay); **c** a gate (*pink*, indicating cells) is created on a forward-scatter (FS) versus side-scatter (SS) plot; **d** the number of events in the gate is displayed in the diagram of the 384-well plate as a heat map (*red* indicating a high number of cells). Data provided by Robinson et al. [12]

As previously mentioned, several criteria drive the development of solutions to analysis of high-throughput data sets, including the ability to read large amounts of data very quickly. While this appears to be a simple task, it is actually quite complex. PA utilizes several unique algorithms to increase the speed of data uploading and utilizes data layering to increase the speed of synthesizing large volumes of data. For a data set containing millions of data points, the applied algorithms and data layering shorten the computation time of uploading and sorting data during analysis. A typical set of 384 listmode files can be read in about 10 s, many times faster than pretty much any other software. Not only is the initial read speed faster, but the subsequent re-reading of the data set is almost instantaneous so that recalculations can be made in real time, even for very complex analyses. This feature allows the user to quickly test "what if" scenarios in PA before deciding on a plan for analysis—something often too difficult or time-consuming with traditional analytical approaches. Rapid sifting through data is a critical aspect of software design when dealing with very large data sets; one of the goals of PA is to reduce the total *time to result*.

The layout of PA is relatively fixed, as shown in Fig. 1. This layout shows the user all aspects of the analysis in a single view. Once data are uploaded to PA (Fig. 1b), they can be viewed through a number of interactive displays and graphs. PA has more gating options than most data analysis software, allowing users to make simple to very complex networks of gated populations that permit easy navigation of multi-parametric data. Like most analytical software, PA displays density scatter plots of the data. But unlike all current software packages, PA allows the display of only a single dot plot or histogram at a time (Fig. 2); these two plots are all that is needed, since all the FCS files of the microtiter plate are readily available to be analyzed simultaneously, individually, or in various combinations.

2.1 Alternative Data Displays

Each of the data display options can also be used for gating during the creation of a "logic map" (Fig. 3a). A logic map is a step-by-step pathway of analysis designed by the user. Whereas most analytical software limits the user to deduce the analytical pathway based on a simple display of scatter plots, histograms, and gates, PA clearly shows the analytical pathway as a network. The advantage of having a network is that there is a visible connection between gated plots. The network can be modified so the steps used to gate a certain population can be shuffled. This way, an analytical pathway is not only spatially re-organized, but the order in which populations are gated is changed (thereby changing the output). The ability to visually display the gating logic behind an output is particularly useful in multi-parameter data sets, where selecting for a unique population means selecting according to an organized series of scatter-plot gates. Therefore, logic maps range from being relatively simple (Fig. 3a), to very complex (Fig. 7e)—as is the case in many CyTOF experiments.

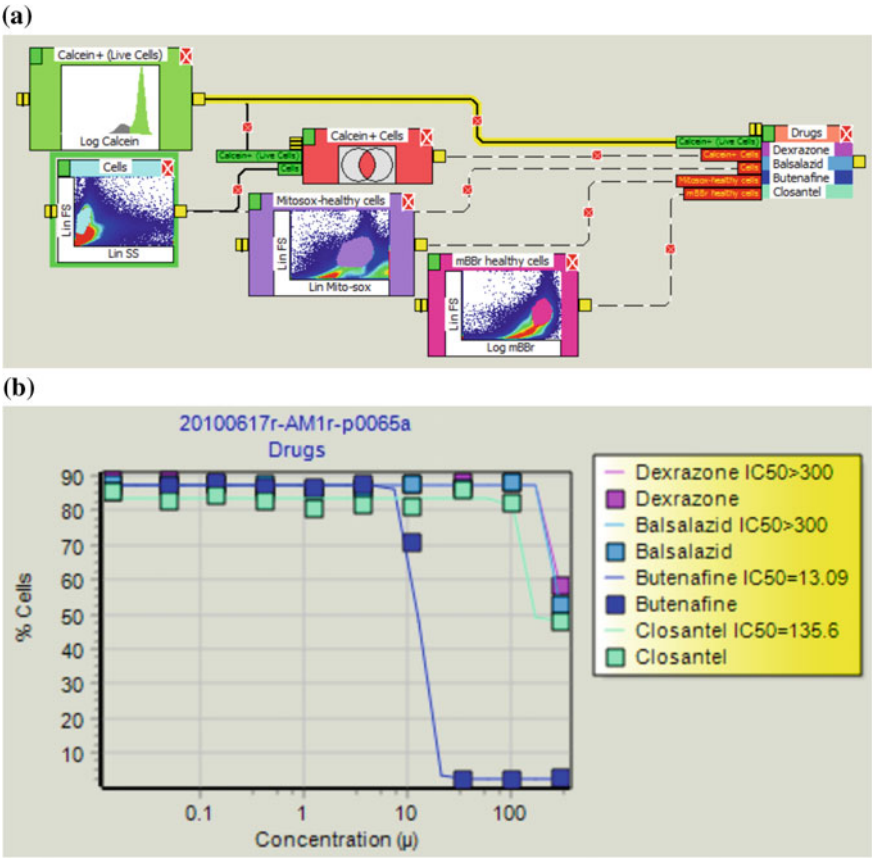

Fig. 3 **a** In the logic map, a number of different gates were made to investigate the effects of drugs on the dyes displayed in the table on HL-60 cells. One gate identifies calcein-positive cells. The calcein histogram gate can be networked with a forward-scatter (FS) versus side-scatter (SS) gate on HL-60 cells to have tighter criteria for selection. The process can be mimicked with gates for Mito-sox- and mBBr-positive populations. **b** A highlighted connection prompts PA to generate IC50 curves of the relationship between drug dosage and calcein dye intensity. Data provided by Robinson et al. [12]

2.2 *Fast Gating of Multiple Parameters*

There are essentially no restrictions in gating in PA, as this can be achieved in five ways. The first two are standard for cytometry software—gating on the histogram or the density scatter plot. The next three are less conventional, but highly practical from the standpoint of analyzing large quantities of data distributed across a multi-well plate. One option is to create gates using an automated gating algorithm that identifies dense clusters of data in a scatter plot and automatically applies a gate. This is particularly useful when many populations exist, as will be seen in an

example of multiplexed bead analysis for cytokines discussed later in the chapter. Another method is to gate selected wells on the microtiter plate. For example, if a section of the plate is dedicated to controls, the user can create a gate that includes or excludes only controls. The fifth gating option applies a concept originally termed PRISM by Dr. James Wood when it was invented at the beginning of 1985 [13].

PRISM takes information from multiple parameters simultaneously and arranges them in histograms. Each histogram is associated with a linear gate separating negative events (below the threshold) and positive events (above the threshold). Analysis of each parameter involves selecting for the spectral "fingerprint" of a population, meaning the population's specific response to the group of antibodies being used as probes. Therefore, the PRISM function can be used as a gating criterion for selecting cells with different "fingerprints." The selection process is achieved in a single window by arranging the linear gates in automatically generated histograms, and identifying whether the desired population falls above or below the threshold for each histogram. This avoids the complex gating schemes commonly used to accomplish the same objective [13].

Gating populations is the first step in creating a logic map. The next step is taking each gated population displayed in the work-space and connecting the gates to refine the definition of the population of interest. Using available Boolean logic, gates can participate in multiple different analytical networks whose logic path is visually easy to follow. This offers a large advantage over traditional software where a user can easily get lost in the gating logic if multiple gates are needed to select for certain populations among many different samples. In PA, complex gating strategies can be applied and changed for all samples and viewed in real time for all populations.

For Robinson's drug screen, logic maps for three different assays are created for entire 384-well plates in just a few minutes. In conventional software, the final product would be a scatter plot or histogram and a couple of simple statistics regarding the percentage or intensity values of cells that fall inside or outside the gate. However, for the purpose of rapid drug-screening PA takes analysis one step further by making IC50 curves, and providing IC50 values for each drug identified (Fig. 3b). IC50 curves are a baseline evaluation tool for drug development, making its incorporation into flow cytometry analytical software particularly helpful. Within the PA software it is also possible to select any of four algorithms for IC50 generation, and it is also possible to switch off IC50 calculations to reduce demand on computing capacity during the process of designing the analytical logic map. In addition to generating IC50 curves, PA also exports into an excel sheet population data from scatter plots, histograms, and individual microtiter-plate wells.

Using Robinson et al.'s data as an example, we highlight the areas of modern flow cytometry techniques that are not compatible with the traditional analytical software, and resolve a number of these issues through a custom software specifically designed for rapid analysis of high-content, high-throughput data.

3 PlateAnalyzer: Multiplexed Cytokine Assays

Multiplexed bead assays are an example of an efficient way to analyze multiple analytes using very small volumes of sample. These types of assays have been performed for many years and were originally commercialized under the Luminex™ technology in the mid-1990s [14, 15]. Many companies offer kits to perform multiplexed bead assays ranging from single analytes to more than 30. The concept behind these assays is to provide a powerful tool for researchers to study multiple co-occurring factors in a single assay. This clever technique bypasses the obstacle faced for years when multiplexing was limited by fluorescence overlap. It does so by using sets of differentially sized and stained beads. The differences in size and staining of each set of beads act as an ID tag (Fig. 4). Each bead set is conjugated to a different anti-target antibody. The sets are combined to form a heterogeneous mixture that along with a detection antibody against each target is

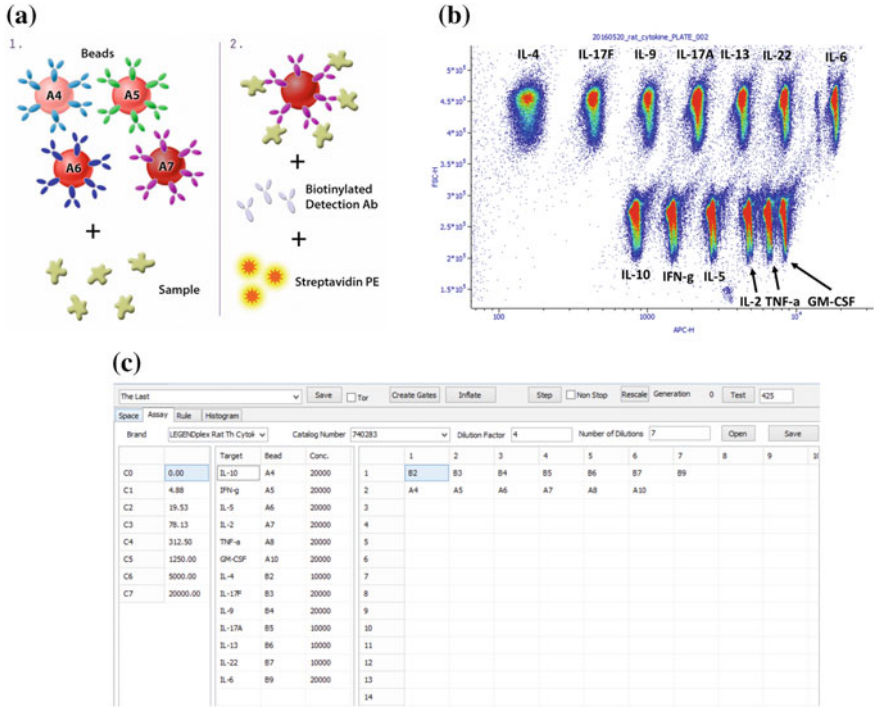

Fig. 4 **a** Biolegend kit (Catalog Number 740402) supplies differentially stained APC beads of two sizes conjugated to thirteen anti-cytokine antibodies. Once beads are introduced to the sample, an antibody conjugated to PE is used to detect the presence of the antigen on the bead. Part of the assay kit constitutes a standard dilution series in which the cytokine concentration is known [16]. **b** The differences between bead sets are observed in a FS versus APC scatter plot, where each bead set is a clear population distributed along the axes. **c** The information on the standard curve can be uploaded to PA through an interactive table

added to a sample containing unknown concentrations of the targets. During analysis, determination of the concentration of each target can be achieved by analyzing the correct bead, identified by its ID tag [16]. The samples can be in either a single test tube or a multi-well plate. The advantages of using a multi-well plate for multi-sample assays were emphasized in the previous section—being more efficient for experiments testing multiple different conditions.

An example of a multiplexed bead assay is demonstrated in an ongoing study where we describe a BioLegend 13-plex bead assay used to determine the cytokine response of plasma from rats experiencing vagus-nerve stimulation and/or treatments of saline or lipopolysaccharide (LPS) injection. In addition to the four treatments, samples are collected from each rat at six time points. A 96-well plate is used: columns 2–11 are loaded with rat plasma samples, while columns 1 and 12 are used to host a known 8-point dilution series of cytokines (these can be placed anywhere on the plate, but are described as being in columns 1 and 12 only for example), beads, and detection antibodies that will later be used to generate standard curves (Fig. 5).

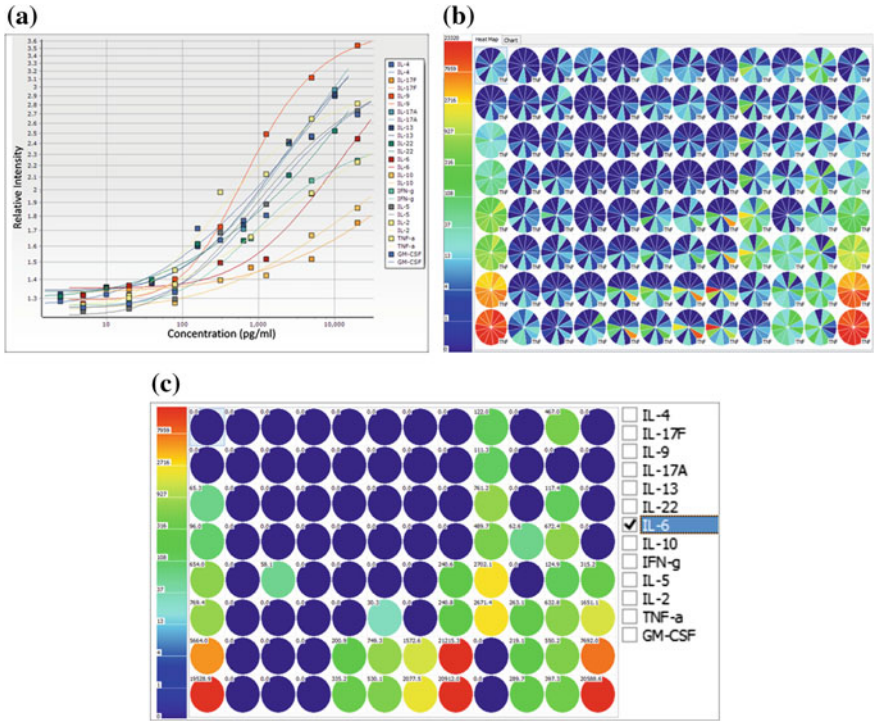

Fig. 5 Once information regarding the standard dilution series is input to the database, calibration curves are automatically generated by PA (**a**) and calculations are performed to determine the concentration (pg/ml) of each cytokine in each well. The calculated results can be qualitatively determined using a colorimetric chart where each pie slice represents a cytokine (**b**), or qualitatively determined by selecting the cytokine of interest (**c**). Data provided by Dr. Joseph Paul Robinson's lab group (unpublished)

3.1 Conventional Versus Alternative Approaches

How would conventional software be used to determine the concentration of each analyte for each well in a 96-well plate? Considering that each well is associated with one FCS file, the user would first upload 96 FCS files. The next step would be to make a protocol, gating each of 13 bead sets, and applying the protocol to all 96 files. The data on gated populations would then be exported (one well at a time) to another program so analyte concentration values could be obtained from a calibration curve that the user must also build. This would equate to 1248 calculations (96×13) that would need to be performed using a secondary software package after the relevant data were extracted from the listmode files of each well.

It is evident that multiplexed bead assays on microtiter plates are very challenging to analyze using standard software. To address this obstacle, manufacturers of multiplexed bead assay kits also sell data-analysis software specifically for those assays. However, the software sold by one company is not necessarily compatible with the multiplexed bead assay kit sold by another company, or with an assay designed by the user. Though differences between kits make them incompatible with mismatched software, bead kits have a number of common descriptors. One regards characteristics of the kit: the number of bead sets, how beads are identified, and the concentration and dilution steps of the standard curves. Another characteristic is directly related to the user: arrangement of experimental and standard samples on a microtiter plate and alterations of the kit's intended protocol. Either information regarding each characterizing feature of the kit must be embedded into the software, or the software must allow the user to easily input this information. To explore methods of resolving the latter of the two tasks, PA software uses an interactive assay-plate display (Fig. 5). The user can tell the program which wells were used for the calibration standards by highlighting them and selecting an option that opens an empty table dedicated to the input of assay information (Fig. 4c). The table is simplistic, prompting the user to enter the number of dilution steps, dilution factor, and initial/highest concentration. In the center of the table, the user can indicate the spatial arrangement and ID of the beads according to their location on a scatter plot. For the BioLegend 13-plex bead assay kit indicated previously, this entire process of detailing the kit characteristics and saving it to the PA kit database takes at maximum 5 min. The information input to the table can be saved, thereby embedding information about the kit in the software. The process of loading a saved kit and applying it to another experiment takes seconds. Using these tools, a user can not only build a database of commercial assay kits, but also develop and save his own custom assay protocols.

PA features other characteristics to facilitate multiplexed bead analysis, compared to both standard flow-cytometry software and software specific for multiplexing. One of these is the method of gating. A time-consuming aspect of multiplexed bead assays is gating each of the multiple bead sets. The number of

bead sets in a single assay is increasing as more sophisticated ways to multiplex with flow cytometry emerge. Consequently, as more bead sets are added to a single kit, it will take more and more time to gate each population. In addition, between experiments users must gate each bead set based on the same criteria for gate size—though this process is up to the discretion of the user and susceptible to human error. Instead of requiring hand-made gates, PA can automatically identify dense clusters of data points and assign size-adjustable gates to each cluster—a process called auto-gating. This approach not only defines gate size according to a single algorithm applied to every experiment, eliminating human variation in gate creation, but also bypasses the need for manual gating—though this is still an option for an event where a combination of auto and manual gating is required.

3.2 Data Output Possibilities

Once data are appropriately gated, it is convenient to expedite the process of exporting the data regarding each gate. This is particularly important in a multiplexed bead assay, where there may be up to 30 gates. The next generation of flow-cytometry software should not only be capable of exporting data, but also make graphs and charts that are appropriate for the type of study being conducted. In the case of a multiplexed bead assay, such graphs and charts include calibration curves (Fig. 5a), automatic calculation of analyte concentration, and automatic application of those calculations to an easily understood visual display (Fig. 5b, c). Based on this need, PA has an integrated function that reads data from the wells that the user has identified as containing the calibration standards and automatically performs all the concentration calculations. The numerical data can be not only produced, but also displayed in an easily interpretable concentration heat map of all analytes. PA achieves this by arranging data in colorimetric pie charts (each slice indicating a different analyte) arranged in the same format as the microtiter plate.

Multiplexed bead assay kits are a major step forward in the field of flow cytometry because they permit simultaneous analysis of a multitude of variables. This feature is particularly powerful when conducting experiments in which variables of interest form a complex interconnected network. Software to examine such assays are available, but a single software compatible with kits from any manufacturer or even custom-built assays is not available. We have identified a number of unique features important to make this software of the future an effective tool: ability to generate calibration curves, rapid concentration calculations, automatic gating, easily interpretable graphics, and simultaneous analysis and visualization of all wells in the plate. These features as well as those common to most flow cytometry software were incorporated into PA, significantly simplifying calculations and abbreviating the time it takes to analyze a 13-plex bead assay to no more than a minute for an entire plate.

4 PlateAnalyzer: Mass Cytometry

Thus far in the chapter, we have noted two major advancements in the field of flow cytometry—automated high-throughput flow cytometry on microtiter plates and multiplexed bead assays. In both cases, traditional flow cytometry approaches are not designed, or are unable to offer the tools needed, for rapid analysis of very large and complex data sets directly from listmode files. The upcoming generation of software must feature methods by which to expedite sample processing by rapidly loading data containing multiple parameters that can be visualized and manipulated simultaneously. This feature is particularly important in the rapidly growing field of time-of-flight mass cytometry (CyTOF). As mentioned in the introduction, mass cytometry can involve the analysis of a few million events, each described by up to 100 parameters. One mechanism that allows for the use of so many parameters is "bar-coding" cells [17] in each treatment with unique combinations of metals/metal isotopes. If a user is conducting the assay in a microtiter plate where each well is experiencing a different experimental condition, bar-coding allows the user to identify which cell originates from which well if all the wells are combined into a single test tube before being processed by CyTOF. The strategy of condensing all experimental treatments to a single tube reduces the time it takes to run a microtiter plate and avoids washing and clearing the instrument between samples, a somewhat time-consuming process.

To keep pace with CyTOF's emerging barcoding technology [18], providers of barcoding kits such as Fluidigm also offer de-barcoding software as part of the instrument's collection and analysis routines. However, it is quite easy to perform

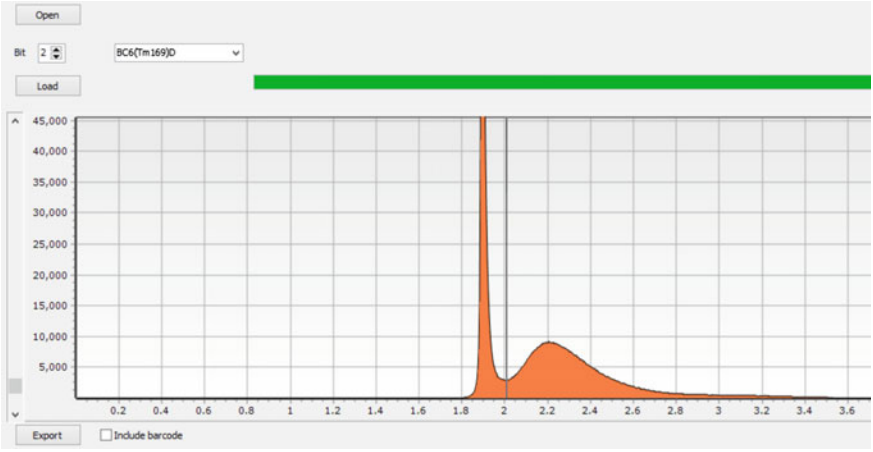

Fig. 6 A dropdown menu lists all the parameters analyzed. BC 1, 2, 3 4, 5, 6, and 7 were the barcodes used in the study by Bodenmiller et al. [3]; each barcode is selected and assigned a bit number. Histograms of each barcode are displayed upon their selection so that a linear gate can be used to select events to the right of the gate and eliminate the noise to the left of the gate

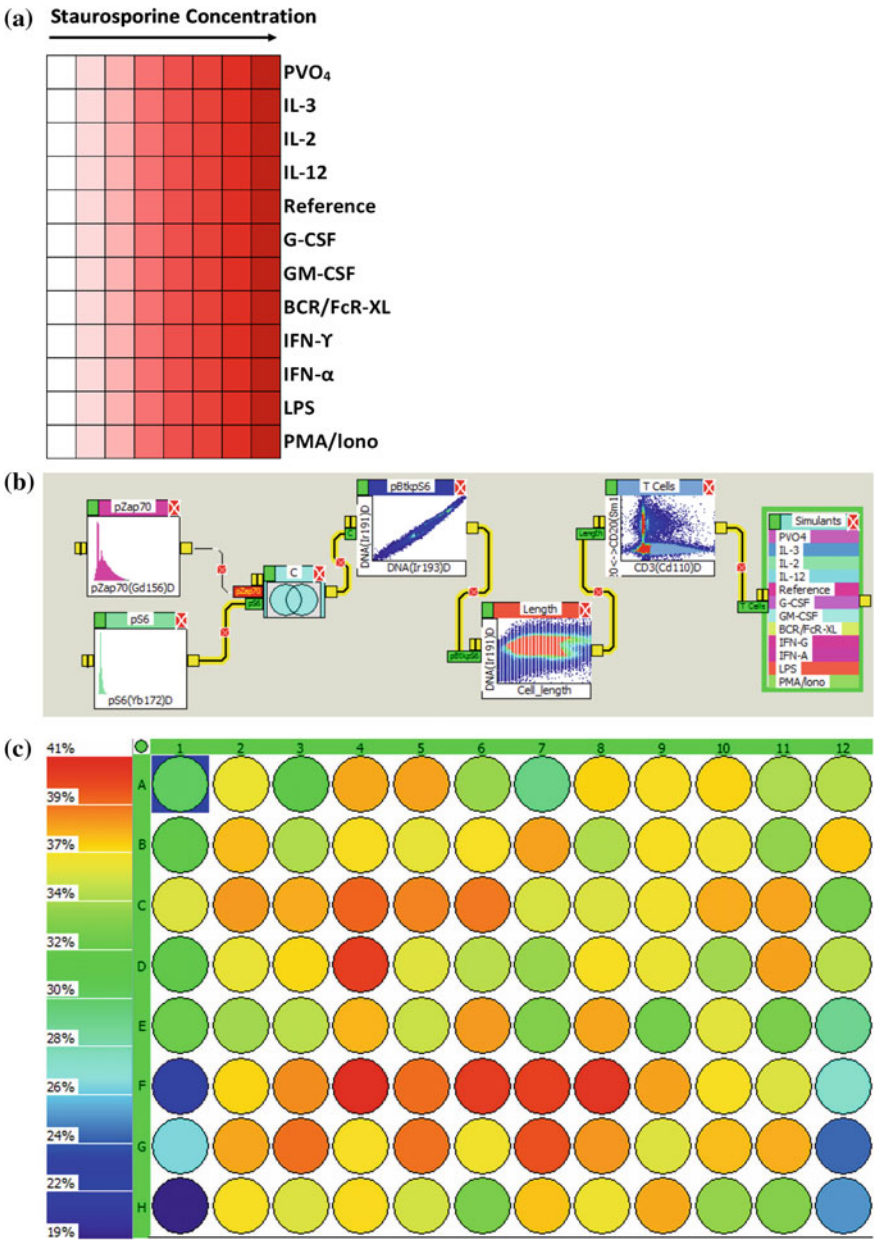

Fig. 7 (continued)

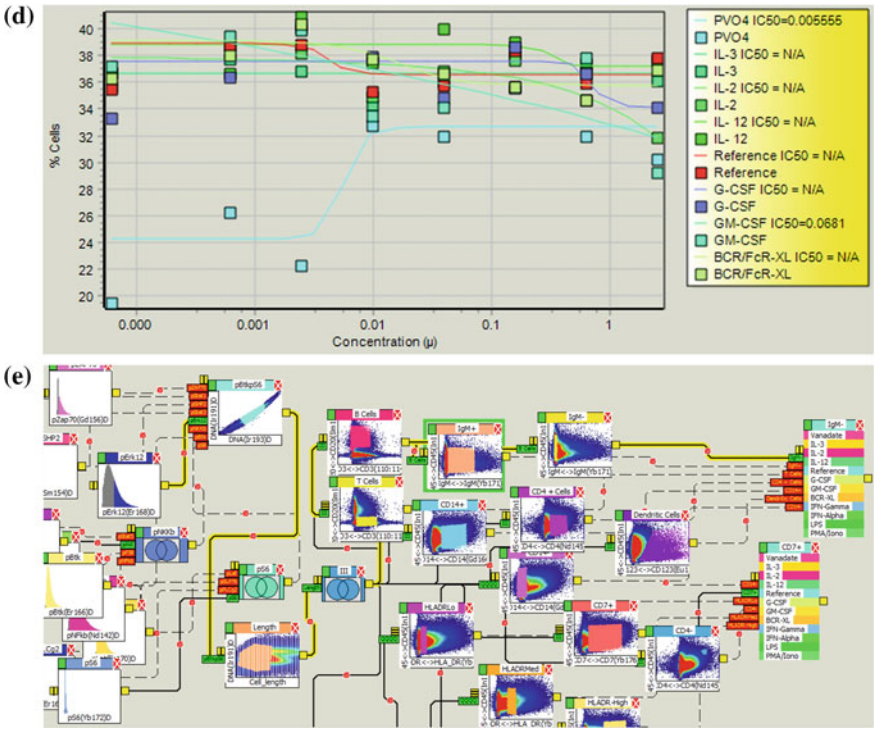

Fig. 7 **a** Display of the assay layout, indicating the stimulants (*rows*) and the 8-point staurosporine dilution (reproduced from reference [3]) series (*columns*). **b** A simple logic map is constructed to determine phosphorylation of pZap70 and pS6 among live T cells in the presence of certain stimulants and staurosporine. **c** A visual effect of the number of events that fall within the gates determined in the protocol. **d** IC50 curves and values of a 10-point dilution of staurosporine for the gated populations. **e** Significantly more complex logic maps can be created. The screen shot of the logic map shown displays the effect of staurosporine on each cell in the presence of every stimulant. The highlighted path tells the user that data on the population indicated through the highlighted path are being displayed on the IC50 graph and the heat map (both not shown). Data provided by Dr. Bernd Bodenmiller [3]

the "de-barcoding" independently, giving the user more control of the output. PA has a "de-barcoding" feature integrated in the software (Fig. 6), allowing the user to manipulate the extracted CyTOF data independently of the instrument software.

An additional feature of the PA de-barcoding algorithm is that it allows the user to keep or remove the barcode parameters within the de-barcoded data file. If you remove these parameters, the file size is significantly smaller, but you are unable to cross check the de-barcoding process. Once the file is de-barcoded PA appropriately separates the data into the correct well according to the original microtiter-plate assay layout and stores the data from each well as an individual FCS file. Once the data are sorted, the user encounters a similar obstacle as with drug screens and cytokine assays: high-content and multi-parametric data are difficult to analyze on

traditional software because such software was originally designed to evaluate a limited number of parameters or events per test tube. An example of the complexity of a data set produced by CyTOF is demonstrated by a study conducted by Dr. Bernd Bodenmiller [3]. The study aimed to examine the effect of different doses of staurosporine on 14 phosphorylation sites in 10 types of peripheral blood mononuclear cells in the presence of 12 stimuli (Fig. 7a). In sum, this would be 2352 8-point dose-response titrations (14 cell types $\times$ 14 phosphorylation sites $\times$ 12 stimuli). Comparison of each well would be time-consuming using conventional software, making PA's ability to view all the data at once while simultaneously making protocols very useful. Figure 7 shows one method by which to analyze two phosphorylation sites in T cells. However, the logic map can be far more extensive than that displayed in this figure, such as the one in Fig. 7e. At the end of the gating process, the desired results are IC50 curves to evaluate dose response. As discussed in the drug-screen example above, PA can generate IC50 curves without needing to export data into a different software (Fig. 7d). Another important feature of the approach we have developed is that one can visualize changes in any parameter displayed on the logic map in real time. For example, if one wished to evaluate the impact of a drug only on cells that expressed very high levels of an antigen (e.g., only the brightest CD3 cells) one could move the gate of the CD3 cell population and in real time observe dose-response curves for any other set of parameters.

5 Renaming FCS Files

For any analysis software, an important aspect to consider, one which is often overlooked, is the proper naming of parameters. Before data are collected on a flow cytometer, parameter names are created and assigned to one of the cytometer PMTs. These parameter names are then reflected on the software displaying data as they are collected on the instrument, and are integrated into the resulting FCS file. However, if parameters are not named *before* data collection they often remain titled FL1 or FL2, etc. Names such as FL1 or FL2 are occasionally included on papers and presentations without appropriate explanation as to the actual parameter identity—something that is unacceptable in either publications or laboratory records. Another significant disadvantage and source of confusion when parameters remain unnamed is the lack of accurate metadata when data sets are sent to collaborators for analysis. On some occasions familiar to many, digital data files are shelved for some time and notebooks containing details on the assay and parameter names are lost. In such cases, the parameters associated with FL1, FL2, etc. may be forgotten or inaccurately remembered, making analysis of otherwise high-quality

data impossible. Currently, there is no software capable of changing parameters names in an FCS file because of the complex relationship between the binary data and the file description information at the beginning of the FCS file format. In addition, this task would be computationally expensive if the data were to be analyzed by the majority of analytical software packages because every FCS file would need to have its parameters re-named during analysis. This is especially problematic when dealing with microtiter-plate assays, which can have hundreds of FCS files per plate. In the FCS file specification, the starting block of data in the binary file is determined when the file is created. Changing any parameter information in the American Standard Code for Information Interchange (ASCII) data file will change that binary starting block and render the file unreadable by any flow cytometry software. Therefore, changing parameter names is not a simple task. The program FCS Rename (Fig. 8) was constructed to rename an FCS file so that when data are visualized on analysis software the correct parameter is listed even if it was not collected when the sample was run on the flow cytometer. With standard flow software, every FCS file of a microtiter plate would need to be renamed prior to analysis because the software gathers parameter information from each FCS file as it is being read. If parameter information were collected from only one FCS file, and applied to the entire plate of data, then one would not need to rename every FCS file, but only the first. This is in fact how PA reads data files, making the application of FCS Rename very practical for data sets consisting of hundreds of FCS files. In the next section, we will explore a program that can allow access to all the details of the FCS file, making certain previously inaccessible information accessible.

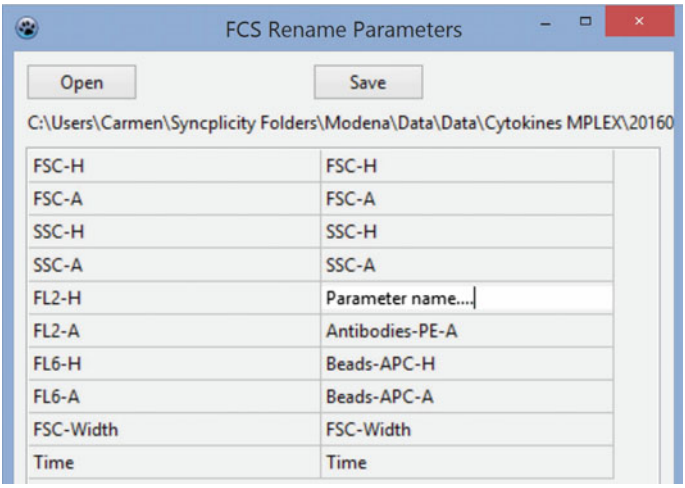

Fig. 8 Parameters from the multiplexed cytokine example described previously are being re-named from the titles indicated on the left-hand side of the column to the titles listed on the right-hand side. The re-named parameters will then be displayed on the axes of graphs and charts made in PA

6 LData: Extracting Raw Data from an FCS File

As seen in the previous section, FCS Rename allows the user to access the parameter names of an FCS file in a very specific manner. On occasion, a deeper level of access to the FCS file is needed to extract raw data.

The program LData (Fig. 9) was developed to transform an FCS file into an ASCII or a.txt file. In this format, raw data from the FCS file can be input to an excel file, parameter names can be manually changed, and the flow cytometer data-collection settings can be viewed. In a clinical laboratory, some information contained in the FCS file must remain confidential as required by the Institutional Review Board (IRB) [19] because it regards a patient's personally identifiable information (PII). For this reason, LData was designed with the ability to erase PII prior to data export, a function called "Incognify."

Using a software such as LData to access information in an FCS file in a form that can easily be changed, exported, and modified is a very powerful tool for flow cytometrists. It can be used in developing software that manipulates data in new and interesting ways, such as PA, FCS Rename, and CytoSpec—software designed for a new generation of flow cytometers discussed in the next section.

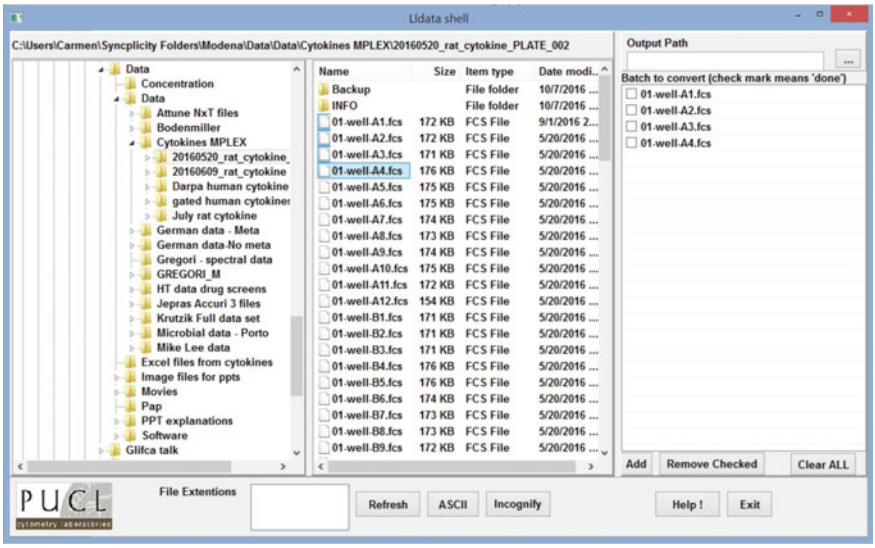

Fig. 9 The LData program interface allows the user to select the file/files to convert to ASCII or comma-separated (.txt) format. Once the data are selected and added into the table on the right, the ASCII option is selected to convert files. The "Incognify" option is specific for FCS files containing personally identifiable information (PII), such as FCS files on patient samples. "Incognify" obscures PII to facilitate data sharing

7 Cytospec: Analyzing Hyperspectral Data

Flow cytometers typically use PMTs to amplify and detect signals with a specific wavelength, each PMT having a single channel. Whereas the early days of flow cytometry used two to three PMTs for two- to three-color detection, recent commercial instruments can detect 10–12 colors using more single-channel PMTs [20, 21]. Though increasing the number of PMTs also increases the degree of multiplexing, there is a limit to the number of colors one can detect simultaneously owing to spectral overlap between fluorophores, the number of lasers available, the fluorophores, and the number of channels. Spectral overlap can to a large degree be resolved through a process of compensation [22]. Increasing the number of PMTs has additional limitations: sensitivity is compromised owing to loss of signal as it passes through more filters, and the design of small instruments becomes more challenging [23].

To address the need of multi-color detection in flow cytometry while overcoming the limitations of traditional systems, a study was conducted in 2004 to construct a prototype hyperspectral flow cytometer in which only one 32-channel PMT was used to detect fluorophores simultaneously per event [4], the equivalent of having a flow cytometer with 32 PMTs. Hyperspectral cytometry places less emphasis on signal intensity at specific wavelengths, and more emphasis on detecting a broad range of wavelengths. This allows for emission profiles of fluorophores to be recorded, not just their intensity at a specific wavelength. In fact, spectral cytometry uses the entire spectrum as a parameter. To maximize the number of fluorophores that could be detected in a single event, an additional study was conducted in 2012 that utilized an approach mathematically similar to compensation: spectral un-mixing [23]. Traditional compensation methods that are still applied in today's flow cytometry software involve mathematically eliminating the fluorescence signal that spills into the wrong channel [24, 25]. However, this process can eliminate potentially useful information. To use the information from spectral spillover in the 2012 study [23], a new and increasingly accepted method based on techniques in image analysis mathematically considers the spilled-over signal as part of the signal of interest. Doing so places more emphasis on measuring a fluorophore spectrum instead of a single emission wavelength and more accurately describes the number of fluorophores present in the analyte (Fig. 10). Approaching spectral un-mixing from the angle of image analysis offers other advantages. The formulas used for spectral un-mixing in image analysis do not assume that the number of detectors equals the number of fluorophores [22]. However, the concept applied to conventional flow cytometry is that the number of PMTs present or "on" is equal to the number of fluorophores being detected. The traditional spectral-compensation approach poses complications for analyzing data produced by new cytometers (where the PMTs are all "on" when the instrument is operating) and hyperspectral cytometry (a single 32-channel PMT). Since hyperspectral cytometry data cannot be analyzed by traditional software because of the constraints needed for compensation, the software *CytoSpec* was created. *CytoSpec*

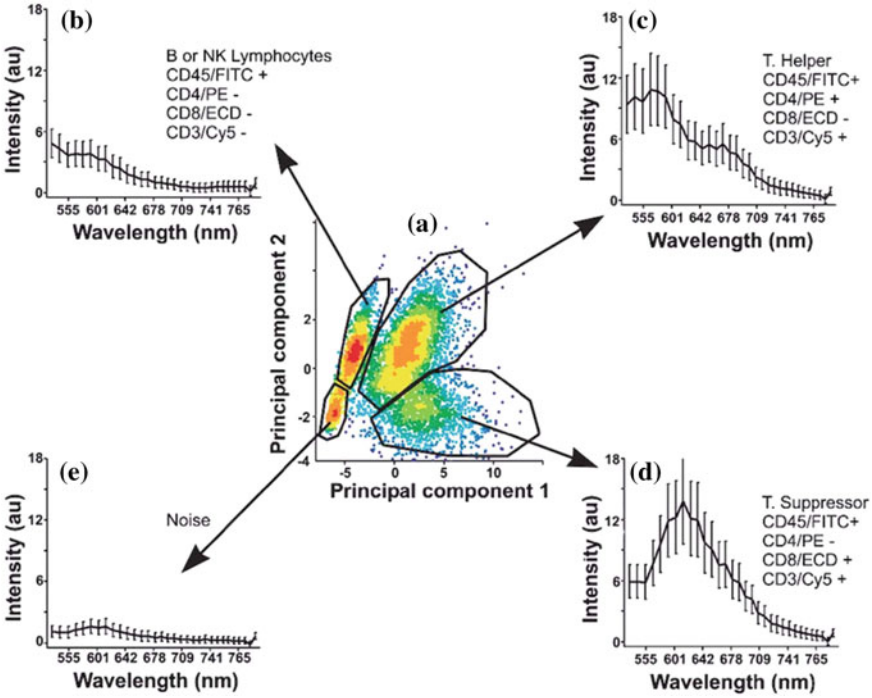

Fig. 10 Image reproduced from Grégori et al. [23]. A spectral analysis of the principal components can be conducted to identify subsets of cells based on the spectra and spatial arrangement characterizing each population

applies spectral un-mixing techniques from image analysis to accommodate the mismatch between the individual PMTs and fluorophores, and utilizes spectral spillover to detect and identify the emission profile of all fluorophores in the analyte.

7.1 Principal Component Analysis

Like most cytometry software, *Cytospec* was designed for not only unmixing spectral data, but analyzing it as well. In multi-parametric studies such as those performed in hyperspectral cytometry, the relationship between parameters is often of interest. To tease apart these relationships, standard flow cytometry data-analysis techniques call for scatter plots comparing each variable or combinations of variables. Given a 10-color data set, 45 scatter plots would be needed to examine different combinations of 2 parameters. This number significantly increases as combinations of parameters are compared, and data analysis becomes challenging for those conducting the study. *Cytospec* provides the user with the option of

performing principal component analysis (PCA), a function that analyzes all data points to identify and group observations according to patterns in the data. The output is a list of principal components, each representing an axis of a coordinate system designed to visually identify patterns in the data (Fig. 11). Until recently, software has not incorporated PCA in the toolset for data analysis because flow cytometry software was originally designed for analyzing much simpler 2- or 3-color data sets [26]. However, to accommodate highly multiplexed studies conducted in hyperspectral cytometry and multi-color flow cytometry, PCA was incorporated into *CytoSpec*. Through *CytoSpec*, PCA can be performed on any number of parameters collected by the flow cytometer and generates a list of principal components that can then be compared using scatter plots. This facilitates

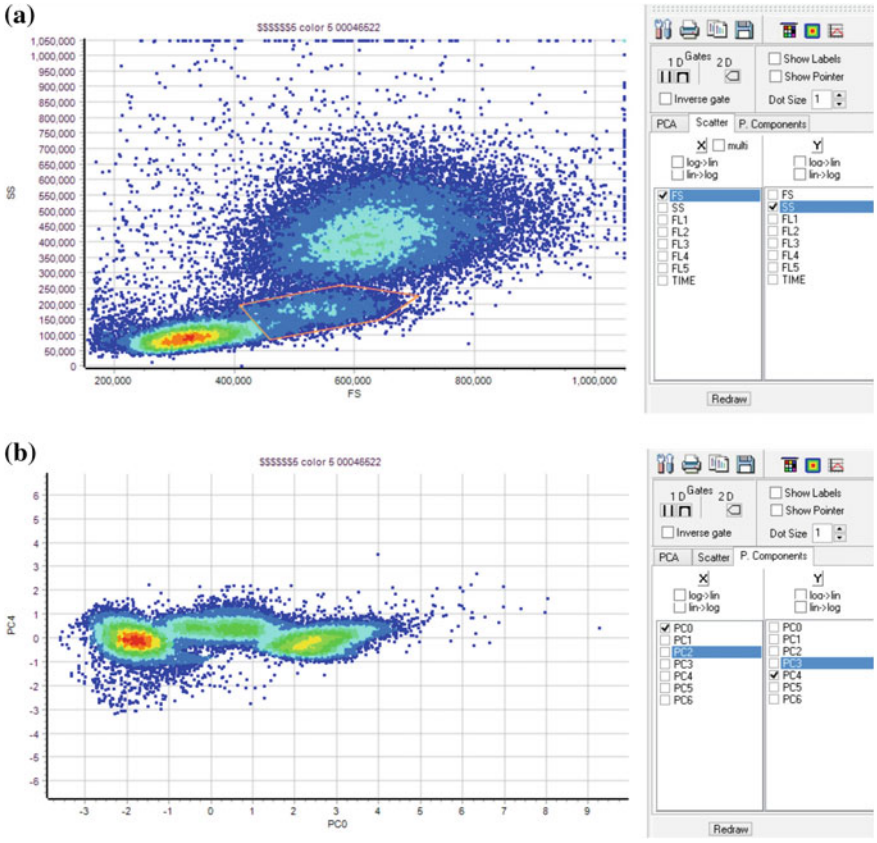

Fig. 11 Human plasma was collected and run on a flow cytometer to observe populations of lymphocytes, monocytes, and granulocytes. **a** CytoSpec was used to analyze the data and create scatterplot displays. **b–d** Cytospec can perform principal component analysis on a selected number of parameters. Different combinations of principal components (PCs) were compared to identify unique groups in the data. Data provided by Dr. J. Paul Robinson's lab group (unpublished)

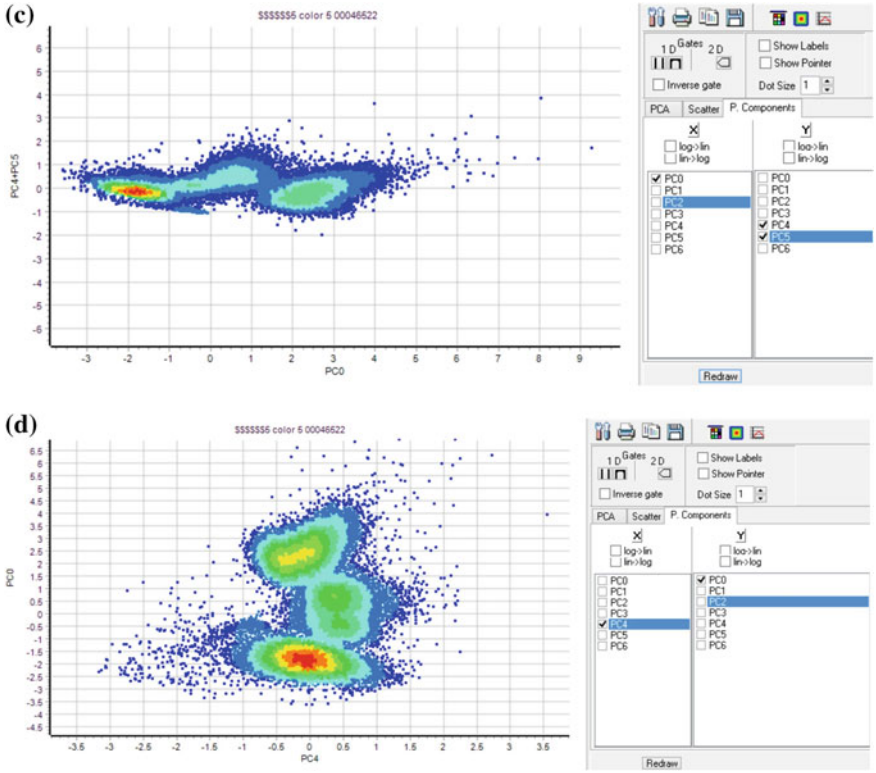

Fig. 11 (continued)

elimination of noise and identification of unique populations that would otherwise have remained invisible.

Elimination of noise is an important part of flow cytometry. Scatter plots of parameters and principal component analyses are both methods by which to discriminate signal from noise. In cases where there is a large degree of noise, populations with low counts can be obscured, and statistics extracted from gated populations can be affected. It is a common feature among analytical software packages to gate out noise, but it may also be desirable to create a new data file on the population of interest that omits most noise or unwanted populations. CytoSpec enables users to generate new FCS files containing only information on a gated set of data points.

In summary, *CytoSpec* is a software tool capable of processing both standard and hyperspectral flow cytometry data using new analytical techniques such as principal component analysis, spectral fingerprint identification, and generation of new FCS files with reduced noise.

8 Conclusion

Instrumentation for fluorescence-based flow cytometry has advanced significantly since its origins in the 1950s. It is noted by Dr. Howard Shapiro that during this time, improvements in flow cytometry automation were out-pacing advancements in data analysis—primarily because computers were largely unavailable to flow cytometrists [27]. As computers became more common, the ability to process the large amount of data produced by flow cytometers caught up to the advancements in instrumentation. Now that the field of flow cytometry has made another great leap forward with automated sample prep, high-throughput sampling, and multi-parametric studies, the supporting analytical tools must again keep pace. This chapter highlights several broad areas of study in which flow cytometry is rapidly advancing: high-throughput drug screens, multi-parametric studies with CyTOF, multiplexed bead assays, and hyperspectral cytometry. Among these areas, six aspects were identified as being of primary importance for the next generation of analytical software:

1. Ability to manipulate an unlimited number of parameters
2. Ability to compare all data parameters in many different combinations
3. Ability to create and customize a data-analysis strategy
4. Ability to see real-time changes in all parameters
5. Ability to integrate advanced statistical tools
6. Rapid reporting toolsets (figures and graphs) with endpoint conclusions.

Based on these six aspects, PlateAnalyzer, *CytoSpec*, FCS Rename, and LData were created to explore a range of software solutions. For high-throughput drug screens, these solutions included simultaneous data visualization, creation and modification of simple-to-extensive logic maps, and generation of IC50 curves, all within the same software. A very similar strategy was taken for analyzing CyTOF data, with the added feature of de-barcoding samples and re-arranging them according to the original microtiter plate array. Multiplexed bead assay analysis was simplified and expedited by using auto-gating systems and ready-to-upload assay databases containing modifiable information on the assay layout. Hyperspectral cytometry posed a new challenge, given that the output, a spectrum of fluorophores, was significantly different from that of conventional cytometers. The software approach tested through *CytoSpec* involved spectral un-mixing and pattern-recognition techniques new to the flow cytometry field. As support for PA and *CytoSpec*, FCS Rename and LData were created to permit the user to change parameter names or extract raw data from FCS files. Information for free download of all four software packages is available on http://www.cyto.purdue.edu/Purdue_software. By creating these data-analysis software options, we hope to form a basis for the next generation of commercial flow cytometry software, which will be geared towards simplifying and expediting high-throughput, multiparametric studies.

References

1. Dean PNBC, Lindmo T, Murphy RF, Salzman GC (1990) Introduction to flow cytometry data file standard. Cytometry 11:321–322. doi:10.1002/cyto.990110302
2. Seamer LC, Bagwell CB, Barden L, Redelman D, Salzman GC, Wood JC, Murphy RF (1997) Proposed new data file standard for flow cytometry, version FCS 3.0. Cytometry 28:118–122
3. Bodenmiller B, Zunder ER, Finck R, Chen TJ, Savig ES, Bruggner RV, Simonds EF, Bendall SC, Sachs K, Krutzik PO, Nolan GP (2012) Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators. Nat Biotechnol 30:67–858. doi:10.1038/nbt.2317
4. Robinson JP (2004) Multispectral cytometry: the next generation. Biophotonics Intl 36–40. doi:10.1017/S1431927605510328
5. Grégori G, Patsekin V, Rajwa B, Jones J, Ragheb K, Holdman C, Robinson JP (2012) Hyperspectral cytometry at the single-cell level using a 32-channel photodetector. Cytometry Part A 81:35–44. doi:10.1002/cyto.a.21120
6. Robinson JP, Durack G, and Kelley S (1991) An innovation in flow cytometry data collection & analysis producing a correlated multiple sample analysis in a single file. Cytometry 12:82–90. doi:10.1002/cyto.990120112
7. Edwards BS, Kuckuck F, Sklar LA (1999) Plug flow cytometry: An automated coupling device for rapid sequential flow cytometric sample analysis. Cytometry 37:156–159. doi:10.1002/cyto.990120112
8. Bocsi J, Tárnok A (2008) Toward automation of flow data analysis. Cytometry Part A 73A:679–680. doi:10.1002/cyto.a.20617
9. Chen X, Hasan M, Libri V, Urrutia A, Beitz B, Rouilly V, Duffy D, Patin É, Chalmond B, Rogge L, Quintana-Murci L, Albert ML, Schwikowski B (2015) Automated flow cytometric analysis across large numbers of samples and cell types. J Clin Immunol 157:249–260. doi:10.1016/j.clim.2014.12.009
10. Finak G, Perez J-M, Weng A, Gottardo R (2010) Optimizing transformations for automated, high throughput analysis of flow cytometry data. BMC Bioinform 11:546. doi:10.1186/1471-2105-11-546
11. Black CB, Duensing TD, Trinkle LS, Dunlay RT (2011) Cell-based screening using high-throughput flow cytometry. Assay and Drug Dev Technol 9:13–20. doi:10.1089/adt.2010.0308
12. Robinson JP, Patsekin V, Holdman C, Ragheb K, Sturgis J, Fatig R, Avramova LV, Rajwa B, Davisson VJ, Lewis N, Narayanan P, Li N, Qualls CW Jr (2013) High-throughput secondary screening at the single-cell level. J Lab Autom 18:85–98. doi:10.1177/2211068212456978
13. Wood J, *Personal Communication on PRISM Processor*, J.P. Robinson, Editor. 2011
14. Elshal MF, McCoy JP (2006) Multiplex bead array assays: performance evaluation and comparison of sensitivity to ELISA. Methods 38:317–323. doi:10.1016/j.ymeth.2005.11.010
15. Horan P, Wheeless L (1977) Quantitative single cell analysis and sorting. Science 198:149–157. doi:10.1126/science.905822
16. BioLegend (2014) Principle of the Assay. Available from: http://www.biolegend.com/legendplex
17. Krutzik PO, Nolan GP (2006) Fluorescent cell barcoding in flow cytometry allows high-throughput drug screening and signaling profiling. NatMethods 3:361–368. doi:10.1038/nmeth872
18. Bendall SC, Simonds EF, Qiu P, Amir el AD, Krutzik PO, Finck R, Bruggner RV, Melamed R, Trejo A, Ornatsky OI, Balderas RS, Plevritis SK, Sachs K, Pe'er D, Tanner SD, Nolan GP (2011) Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. Science 332:96–687. doi:10.1126/science.1198704

19. US Department of Health and Human Services. Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (HIPAA) privacy rule. Available from: http://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html
20. Bendall SC, Nolan GP, Roederer M, Chattopadhyay PK (2012) A deep profiler's guide to cytometry. Trends Immunol 33:323–332. doi:10.1016/j.it.2012.02.010
21. BioLegend (2014) History of Flow Cytometry. Available from: http://www.biolegend.com/historyofflow
22. Novo D, Grégori G, Rajwa B (2013) Generalized unmixing model for multispectral flow cytometry utilizing nonsquare compensation matrices. Cytometry Part A: J Intl Soc Anal Cytol 83:508–520. doi:10.1002/cyto.a.22272
23. Grégori G, Patsekin V, Rajwa B, Jones J, Ragheb K, Holdman C, Robinson JP (2012) Hyperspectral cytometry at the single-cell level using a 32-channel photodetector. Cytometry Part A 81A:35–44. doi:10.1002/cyto.a.21120
24. Bagwell CB, Adams EG (1993) Fluorescence spectral overlap compensation for any number of flow cytometry parameters. Ann NY Acad Sci 677:84–167. doi:10.1111/j.1749-6632.1993.tb38775.x
25. Roederer M (2002) Compensation in flow cytometry. Curr Protoc Cytom 1–14. doi:10.1002/0471142956.cy0114s22
26. Loken MR, Lanier LL (1984) Three-color immunofluorescence analysis of Leu antigens on human peripheral blood using two lasers on a fluorescence-activated cell sorter. Cytometry 5:151–158. doi:10.1002/cyto.990050209
27. Shapiro HM (2005) History. In: Practical flow cytometry. Wiley, Hoboken, pp 73–100