Bioinformatics, 32(16), 2016, 2473–2480
doi: 10.1093/bioinformatics/btw191
Advance Access Publication Date: 10 April 2016
Original Paper

OXFORD

Data and text mining

flowAI: automatic and interactive anomaly discerning tools for flow cytometry data

Gianni Monaco[1,2,*], Hao Chen[1], Michael Poidinger[1], Jinmiao Chen[1], João Pedro de Magalhães[2] and Anis Larbi[1,*]

[1]Singapore Immunology Network (SIgN), Agency for Science Technology and Research (A*STAR), Singapore 138648, Singapore and [2]Integrative Genomics of Ageing Group, Institute of Integrative Biology, University of Liverpool, Liverpool L69 7ZB, UK

*To whom correspondence should be addressed.

Associate Editor: Robert Murphy

Received on December 17, 2015; revised on April 3, 2016; accepted on April 4, 2016

Abstract

Motivation: Flow cytometry (FCM) is widely used in both clinical and basic research to characterize cell phenotypes and functions. The latest FCM instruments analyze up to 20 markers of individual cells, producing high-dimensional data. This requires the use of the latest clustering and dimensionality reduction techniques to automatically segregate cell sub-populations in an unbiased manner. However, automated analyses may lead to false discoveries due to inter-sample differences in quality and properties.

Results: We present an R package, flowAI, containing two methods to clean FCM files from unwanted events: (i) an automatic method that adopts algorithms for the detection of anomalies and (ii) an interactive method with a graphical user interface implemented into an R shiny application. The general approach behind the two methods consists of three key steps to check and remove suspected anomalies that derive from (i) abrupt changes in the flow rate, (ii) instability of signal acquisition and (iii) outliers in the lower limit and margin events in the upper limit of the dynamic range. For each file analyzed our software generates a summary of the quality assessment from the aforementioned steps. The software presented is an intuitive solution seeking to improve the results not only of manual but also and in particular of automatic analysis on FCM data.

Availability and implementation: R source code available through Bioconductor: http://bioconductor.org/packages/flowAI/

Contacts: mongianni1@gmail.com or Anis_Larbi@immunol.a-star.edu.sg

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Flow cytometry (FCM) is a laser-based methodology designed to capture the physical and biochemical characteristics of a cell or a particle in a stream of fluid. Fluorescence-conjugated antibodies are used to target antigens expressed inside or at the surface of the cells of interest. As cells pass through the laser (excitation), the fluorochrome will change its state of energy and emit a light (emission) that is captured by a series of detectors. FCM applications have been developed mainly for both research and clinical settings in medicine but also for other non-biomedical domains such as marine and plant biology. The most common application is the immune-phenotyping of blood samples and thus the quantification of the number and frequency of various immune cell populations. In hematology, FCM is the technology of choice, as, for example, it requires only few drops of blood to diagnose leukemia through the detection of the perturbation of normal cell frequencies (Brown and Wittwer, 2000). Moreover, FCM helped increase our understanding of cellular functions of the immune system and is widely used in cell cycle analysis, pre-transplant crossmatching, cell sorting, apoptosis, vaccine development and other applications that scrutinize cellular properties

(Jaye *et al.*, 2012; Mulley and Kanellis, 2011; Pozarowski and Darzynkiewicz, 2004; Vermes *et al.*, 2000).

The data are stored in Flow Cytometry Standard (FCS) files that include the fluorescence and scattered light levels for each cell that passed through the laser beams. Nowadays it is possible to analyze up to 20 markers at a time in a single staining panel by using an equal number of different fluorochromes detected in separate channels. The common approach used to analyze the data produced by FCM is to visually select cells of interest through 1 or 2 markers known to be highly specific. However, to delineate the high heterogeneity of immune cell populations, it is necessary to look simultaneously at the whole staining panel. Principal component analysis has been used to detect the complexity in CD8 T cell populations characterized by intermediate phenotypes that show a continuum of expression of different combinations of cytokines and surface markers (Newell *et al.*, 2012). Another dimensionality reduction technique called t-Distributed Stochastic Neighbor Embedding (t-SNE) (Becher *et al.*, 2014; Maaten and Hinton, 2008; Shekhar *et al.*, 2014) was successfully applied to identify ambiguous cell populations, including monocyte–macrophage intermediates and granulocyte variants in a mass cytometry experiment based on a 38-antibody panel (Becher *et al.*, 2014).

Several computational tools that aim to automatically characterize cell populations without losing multi-dimensional information are constantly developed and periodically benchmarked by the FlowCAP consortium (Aghaeepour *et al.*, 2013). Undoubtedly, the widest range of tools has been distributed by the BioConductor platform based on the R programming language. The root package for FCM data is flowCore, since it defines the container class and it enables to perform essential manipulations such as compensation and transformation (Hahne *et al.*, 2009). In addition, a series of complementary packages has been developed for further operations, such as visualization, quality assessment, statistical analysis and automated gating (Finak *et al.*, 2014; Hahne *et al.*; Sarkar *et al.*, 2008).

To accompany and support the large development of automatic methods to define populations, it is crucial to use high quality FCM data as input in order to optimize the robustness of the results. This is especially true since research is looking deeper into the complexity of cell distribution. For instance, target cell sub-populations may represent as low as 0.05% of the total cell population suggesting that minute variation in the quality of the data may lead to false positive results or loss of signal. Standardization, calibration and quality control guidelines using beads have been defined to ensure that the signal acquired is the most accurate and with the least variation (Oldaker, 2007; Perfetto *et al.*, 2006). Nonetheless, these procedures are not always carefully monitored and even having the FCM instrument at optimal conditions before sample processing does not exclude electronic drifts or fluidic instability issues at the time of data recording. An R package, flowQ (Bashashati and Brinkman, 2009; Gentleman *et al.*, 2006), creates concise reports of quality checks on single and multi-panel experiments to highlight issues that can be encountered in data acquisition. The reports indicate the number of cells, percentage of boundary events and anomalies on the fluidics and signal acquisition over time. Another package, flowClean (Fletez-Brant, 2014), determines and marks low quality cells using compositional data analysis. In brief, it splits the time in equally sized bins and flags the events that are within time frames containing unusual ratios of cell populations. However, flowQ does not actively detect and remove the anomalies and flowClean is poorly intuitive and thus it does not allow to infer the source of the anomalies.

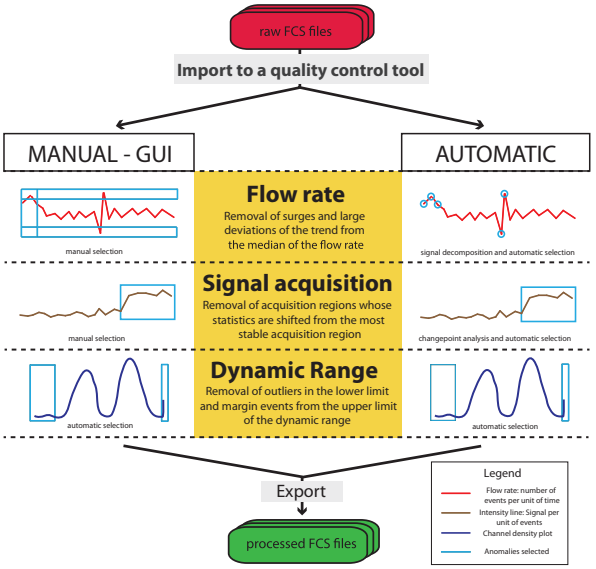

Fig. 1. Workflow of the quality control of FCM data using the flowAI package. Data can be processed manually with a Shiny application or automatically with the call of an R function. The steps are complementary in both cases. On the one hand, the manual method allows the user to interactively choose appropriate thresholds on plots portraying flow rate and signal acquisition through visual inspection. On the other hand, the automatic method performs this selection through anomaly detection algorithms. Both the interactive and automatic methods eliminate events recorded outside an acceptable dynamic range using standard thresholds

We present our package called flowAI that provides two solutions, one automatic and one interactive, to discard cells from FCM data that do not reach appropriate quality standards. Our workflow adapts and expands previous ideas with methods never implemented before to provide a more objective, efficient and intuitive solution for the quality control of FCM data.

2 Implementation and methods

2.1 The software

Both the automatic and interactive methods have been implemented in the R package flowAI and distributed by the Bioconductor platform (http://bioconductor.org/packages/flowAI/). Our tools incorporate functionalities from several other R packages. For example, the automatic method integrates functions from the mFilter (Balcilar, 2007) and changepoint (Killick and Eckley, 2014) packages in the algorithms aiming to automatically detect the anomalies while the interactive method leverages on the R shiny package (Chang *et al.*, 2105) to build the web graphical interface.

2.2 Workflow

The entire quality control analysis of flowAI contains three main steps to detect and remove anomalies from FCM data complementary for both the automatic and the manual methods (Fig. 1).

2.2.1 Flow rate check

The first step evaluates the steadiness of the flow rate of the analysis. The flow rate is reconstructed by reporting the number of cells acquired per unit of time. This is only possible for FCS files of version equal or greater than 3.0 which implement the keyword $TIMESTEP to allow for kinetic analysis (Seamer *et al.*, 1997).

The keyword stores a value corresponding to the resolution of the 'Time' channel in terms of seconds or fractions of a second. Ideally, the detection of anomalies in the flow rate check should be performed at the maximum time resolution allowed by the FCM instrument. However, the setting of a larger time step for the analysis greatly decreases the running time and memory usage.

A stable flow rate of FCM instruments can be pictured by a line with non-periodic fluctuations but with a constant variation. The anomalies in the flow rate that mostly affect the quality of signal acquisition are abrupt surges and significant changes in the speed of the fluid, generally caused by factors such as debris and air intrusion in the fluidic system. To discard anomalies through the interactive method, users can adjust two horizontal sliding bars to eliminate flow rate surges and two vertical sliding bars to discard regions at the beginning and the end of the flow rate where the instabilities mostly occur. Instead, for the automatic version we designed an anomaly detection algorithm built upon the generalized extreme studentized deviate (ESD) test (Rosner, 1983) and optimized to work on time series data.

As stated in a review of outlier detection methods, the anomalies are contextual to the nature of the data (Chandola *et al.*, 2009) and hence it is preferable to develop techniques customized for the domain of interest. The patterns depicted by the flow rate of FCM data are generally similar to the ones treated by economists, engineers and social scientists in time series analyses, whose basic idea is to extract additional information from time series data by splitting into its components.

As a first step for our automatic method, we implemented the Christiano–Fitzgerald band pass filter (Balcilar, 2007; Christiano and Fitzgerald, 2003) to split the value (y_t), corresponding to the number of events recorded at the time point t, into the trend (τ_t) and cyclical (c_t) components:

$$y_t = \tau_t + c_t \tag{1}$$

The trend component will be a smooth line that indicates long-term increase or decrease in the flow rate, while the cyclical component will contain the non-periodic fluctuations and abrupt surges from the trend line.

Second, the flow rate values are penalized by adding or subtracting the corresponding absolute values of the cyclical component according to their direction from their median:

$$y.pen_t = \begin{cases} y_t + |c_t|, & y_t \geq \text{median}(y) \\ y_t - |c_t|, & y_t < \text{median}(y) \end{cases} \tag{2}$$

Lastly, the generalized ESD test is applied on the penalized flow rate to detect the anomalies. This method, with an iterative process, searches for a number outliers not exceeding a predefined threshold k, $r_{1:k} = \{r_1, r_2, \ldots, r_i, \ldots, r_k\}$, in a dataset of sample size n. At each iteration, an observation r_i is tested as a potential outlier and it is removed from the data before the next iteration. Practically, an exemplary iteration has the following steps:

1. Extraction of the observation that largely deviates from the central tendency indicator (mean or median) scaled by the measure of dispersion (standard deviation or median absolute deviation):

$$r_i = \frac{\max |y.pen_i - \text{median}(y)|}{MAD(y)} \tag{3}$$

2. Computation of the critical value lambda λ_i from the t distribution using a defined level of significance α. The observation is flagged as an outlier if its value is higher than lambda: $r_i > \lambda_i$.
3. The observation r_i is removed from the data that is now reduced to the sample size $n - i$.

Our procedure uses the median and the median absolute deviation (MAD) because, particularly in presence of outliers, they are a more robust alternative to the mean and standard deviation (Leys *et al.*, 2013).

2.2.2 Signal acquisition check

The second step verifies the stability of the signal acquired over time. A common practice to verify the quality of signal acquisition is to use Levy–Jennings-type graphs, where fluorescence is plotted against time (Barnett and Reilly, 2007). A stable signal acquisition should produce intensity values whose distribution is consistent throughout the course of the entire experiment. This is the expected behavior if we assume that cells from a heterogeneous sample are randomly aspirated by the FCM tube over time. Therefore, changes in the signal intensities are not due to biological variation but rather to technical issues such as defective laser-detection system, voltage instability or poor quality of sample preparation, for example, inadequate vortexing.

For each channel, flowAI creates Levy–Jennings-type graphs by splitting the intensity values of a marker in equally sized bins and plotting their median against time. This method is already implemented by the flowQ package, where the user can infer the quality of an FCS file from the visualization of time line plots. However, in addition to that, flowAI allows the removal of the regions with an unstable signal. As for the flow rate, this operation can be performed manually through visual inspection or automatically. The latter method implements a step detection algorithm to identify shifts in the mean and variance of the intensity values. The algorithm used, binary segmentation, is implemented in the changepoint package (Killick and Eckley, 2014). Its basic concept has been firstly described by the genetists Edwards and Cavalli-Sforza as a new clustering method based on the analysis of variance (Edwards and Cavalli-Sforza, 1965). This method is computationally fast and most frequently used among the changepoint detection methods.

This approach iteratively splits the data in two groups at a time simply applying the method of least squares. In our case, given an ordered set of n fluorescence values $m_{1:n} = (m_1, m_2, \ldots, m_i, \ldots, m_n)$ corresponding to the medians of all bins, the total sum of squares (SST) from their mean is calculated as a measure of dispersion:

$$SST = \sum_{i=1}^{n} (m_i - \bar{m})^2 \tag{4}$$

A changepoint m_i that splits the data in two segments, $s_1 = (m_1, \ldots, m_i)$ and $s_2 = (m_{i+1}, \ldots, m_n)$, is detected when the cost function, represented by the within-groups sum of squares (SSW), is minimized:

$$\arg\min_i \sum_{s_1=1}^{i} (m_{s_1} - \bar{m}_{s_1})^2 + \sum_{s_2=i+1}^{n} (m_{s_2} - \bar{m}_{s_2})^2 \tag{5}$$

The minimization of the cost function (5) is equivalent to the maximization of the between-group sum of squares (SSB), and the sum between SSW and SSB results in the SST.

Each new segment created is in turn split in two segments by the repetition of the same procedure. The search of new changepoints terminates if either the minimized cost function is higher than a defined threshold or if a pre-established maximum number of

changepoints has been detected. In flowAI we used a variant of this method provided by the changepoint package that not only searches for shifts in the mean but also in the variance.

The binary segmentation algorithm is performed independently on each fluorescence channel and lastly the longest region that does not contain changepoints in any of the channels is chosen as high quality signal.

2.2.3 Dynamic range check

A third quality step is performed on the lower and upper limit of the dynamic range. Signals recorded by FCM instruments can only fall within a determined dynamic range. The last generation of FCM has reached a dynamic range of 2^{24} channels (Novo and Wood, 2008), but most of the instruments nowadays used in laboratories and clinics have a range of 2^{18}. Due to this limitation, all measurements with a real value higher than the upper limit will be recorded in the last channel of the dynamic range causing an accumulation of signals that is not directly comparable with the rest of the data. These values are commonly called margin events. Our package allows the removal of events where at least one of the parameters has an intensity value on the upper limit of the dynamic range.

The values of the lower limit are treated in a different way. For the signal of the light scatter channels (reflecting the morphology of the cells) any value less than zero is removed. Instead, for the immunofluorescence channel, small fluctuations in the range of negative values are usually acceptable since they are the byproduct of standard operations such as correction of background noise, auto fluorescence and spectral overlap. Nonetheless, technical issues, such as flow rate surges or voltage instability, can exacerbate the magnitude of a negative value to an unacceptable range that would also interfere with the downstream signal processing, such as logicle transformation or automatic gating.

The flowAI package uses an outlier detection method to remove the outliers among the negative values. Every value that is inferior to a certain threshold is labelled as outlier and consequently removed. For each channel, a threshold referred to as Z-score is computed with a method recommended by Iglewicz and Hoaglin (1993). The formula is given in (6), where the threshold is obtained for a set of n negative values $x_{1:n} = (x_1, \ldots, x_n)$:

$$Z = \frac{-3.5 \ \mathrm{MAD}(x_{1:n})}{0.6745} + \mathrm{median}(x_{1:n}) \qquad (6)$$

Alternatively to the removal of negative outliers, the lower limit of the dynamic range can be truncated to the cut-off suggested by the FCS file. This method was previously adopted as preprocessing step for the cleaning of FCM data from erroneous measurement (Qian *et al.*, 2012; Van Gassen *et al.*, 2016).

2.2.4 Results evaluation

At the completion of the analysis with the automatic method, a report is generated indicating the percentage of cells that did not pass the quality checks and a series of graphs showing where the anomalies in terms of time and parameters were detected. Our suggestion is first to run the automatic method with default settings on a small sample of FCM data, second to customize the settings if necessary, third to perform the quality control automatically on the entire dataset and lastly to intervene manually only for those files whose automatic control is not able to meet the accuracy required.

3 Results and discussion

Here, we provide analysis results obtained using the automatic method in flowAI on several FCM data. We studied the nature of the abnormalities detected in each quality control step and then we evaluated the overall improvement of computational analysis with the cleaned data.

3.1 Overview of the datasets

A total of 4469 FCM files from 11 different datasets, precisely 2 in-house and 9 from the online database FlowRepository (Spidlen *et al.*, 2012), were used for our evaluation. The two in-house datasets contain 84 samples each, and are part of a larger project called the Singapore Longitudinal Aging Study (SLAS). Ethical approval was obtained from the National University of Singapore Institutional Review Board for SLAS blood collection and experiments. A different panel was used for the two datasets. Panel 1 consisted of 16 antibodies targeting markers for the overall white blood cell populations: CD16, CD4, CD38, CD62L, CD19, CD66b, CD45, CD27, CD56, CD3, CD8, CD14, CD123, HLA-DR. Panel 2 consisted of 14 antibodies targeting the B lymphocyte populations: CD19, CD20, CD21, CD23, CD24, CD27, CD38, IgG, IgM, IgD, HLA-DR. Regarding the 9 datasets retrieved online, we selected the ones used for the flowCAP contests. Data and details are available on flowrepository.org under the IDs with the prefix FR-FCM- and followed by: ZZYA, ZZZU, ZZY2, ZZY3, ZZYY, ZZY6, ZZYZ, ZZZV, ZZ99.

3.2 Examination of anomalies in FCM data from different perspectives

In this section, the anomalies detected in each quality control step is analyzed separately. The main consideration is that even though our workflow schematizes the quality control in three different steps, they are usually strictly related. For example, a surge in the flow rate often corresponds to an unstable signal acquisition that in turn would potentially result in a value in the upper margin or in the negative outlier space of the dynamic range. Nonetheless, given the high variability of anomalies that can occur in a FCM experiment, the division of the quality control in the three steps defined in our work is necessary to assure the detection of those anomalies that are not visible from a single perspective.

In this manuscript, we focus on the file 220662.fcs from the ZZZV dataset to show how a complete quality control with flowAI works on an FCS file. In addition, numerous other examples are reported in the supplementary material.

3.2.1 Surges and trend shifts in the flow rate

The flow rate was recreated dividing the time channel of an FCS file in equal intervals with a time step of 1/10 of a second. Fluidics' stability in the sample is a good indicator for the absence of anomalies such as clogging and air bubbles in the flow cell and other disturbances in the flow stream. Our algorithm has been designed to acknowledge cyclical patterns to detect local anomalies, i.e. surges, as well as to remove global anomalies, i.e. large deviations of the trend from the median flow rate (Fig. 2a). From all the FCS files analyzed, we verified that the beginning and the end of the flow rate are the regions where irregularities occur the most. FCM experts recognize these patterns as being frequent and mainly due to air bubbles, debris or clogged cells (Supplementary Fig. S1a, e, f). In Figure 2a, the flow rate takes about 10 s to stabilize but usually strong fluctuations vanish more quickly (Supplementary Figs S1a, S2a and S3a).

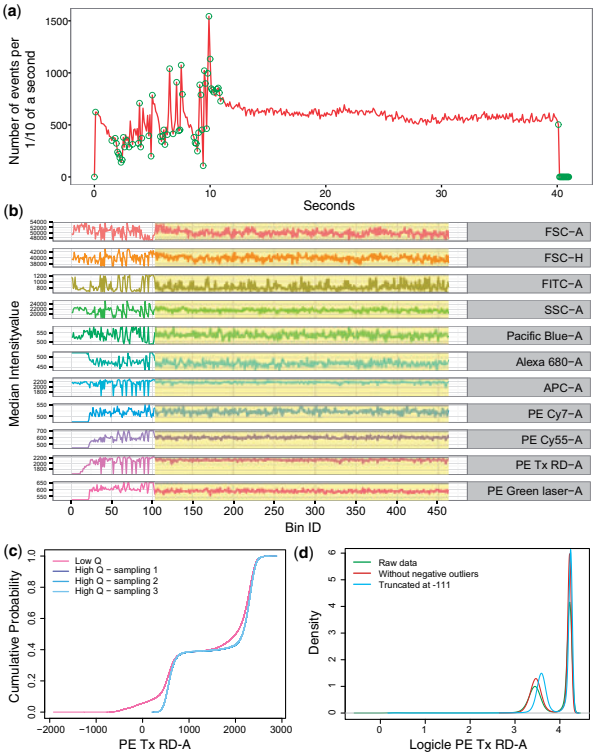

Fig. 2. Quality control results of the file 220662.fcs from the ZZZV dataset. The plots (**a**) and (**b**) were extracted from the report generated by the automatic method of the flowAI package using default settings. (a) Strong fluctuations are detected in the flow rate at the beginning of the experiment. The anomalies detected are indicated with green circles. (b) Changepoint detection in signal intensity over time, represented as median of equally sized bins. The region discarded is complementary to the one detected as instable in the flow rate check. The yellow region is selected as being steady and therefore categorized as high quality. (**c**) ECDF curves of raw intensity values of the low (in red) and high (shades of blue) quality events of the PE Tx RD-A channel. The sample size of the three high quality samplings equals the number of low quality events detected. (**d**) Density plots of the logicle transformed data of the PE Tx RD-A channel using the logicle parameters estimated from raw data (green line), from data with negative values truncated at -111 (blue line), and from data without negative outliers (red line). The density curves vary among the three sets of data, indicating the repercussions on the estimation of the logicle parameters according to the dynamic range used for the data

Nevertheless, there are cases of flow rate surges interspersed over the entire course of the experiment (Supplementary Figs S4a and S5a) possibly caused by clusters of debris suddenly aspirated by the FCM tube (Supplementary Fig. S5a–c). However, even though it was not always possible to associate flow rate surges with debris or clogged cells, surges removal is still necessary because of their association with signal intensity variation.

Lastly, in an FCS file we observed a steady change of the flow rate, and hence the signal, in the last part of the analysis. The resulting low quality cells have a distribution uniformly shifted compared to the one of the high quality cells. This is probably due to the manipulation of the speed settings by the instrument operator during the running of the experiment (Supplementary Fig. S6).

3.2.2 Mean and variance deviation from stable acquisition regions

For each channel, the signal acquisition over time is reconstructed first dividing the total number of cells in equally sized bins and second calculating the median value of each bin. The output is graphically shown with line plots (Fig. 2b). Mean and variance shifts in the signal acquisition are detected using the binary segmentation method from the changepoint package (see Section 2.2.2).

In most of the analyzed cases, signal instability is strongly related to flow rate fluctuations (Fig. 2, Supplementary Figs S1–S3 and S6). However, anomalies caused by laser-detection systems can eventually occur independently of the speed variations of the flow rate. In Supplementary Figure S4, for example, the numerous flow rate surges are hardly detectable in the signal plots and the channels storing the signal elicited by the green laser (G780-A, G710-A, G660-A and G610-A) show a delay in the reaching of stability that warrants a careful monitoring of the functionality of that specific laser-detection system.

In Supplementary Figure S5, even though the flow rate surges are associated clearly with the signal plots, the signal acquisition gradually weakens at different rates in different channels after a first region of steadiness (FSC-A, FSC-H and APC-A), while in other channels it remains constant for a longer period. In this rare case, other technical issues should be investigated. Some of the factors that might cause less common anomalies, but have to be kept in mind, are laser power instability, detection system irregularities, poor quality of the sheath fluid and accumulation of dirt in the flow cell.

3.2.3 Refining the dynamic range: removal of negative outliers and margin events

Because of the quantum nature of light, both the scatter and fluorescence channel values cannot theoretically fall in the negative range of values. However, because of the background and noise correction of the optical detection system of FCM instruments, negative values are recorded for both light scatter and immunofluorescence channels. This problem is also exacerbated by instable signal acquisition, for instance during flow rate surges (Supplementary Figs S2a–c and S4a–c), or by compensation, where a value proportionate to the spectra overlap of other channels is subtracted from each channel. Negative estimates are considered part of a negative population of cells with a low mean and a large coefficient of variation. Therefore, with the logarithmic transformation not being able to handle negative values, new transformation methods have been developed. Probably the most popular one is the logicle transformation, also called 'bi-exponential' (Parks *et al.*, 2006). With this method, values with an absolute small magnitude are scaled linearly, while large values are scaled in a log-like fashion. The transition from the linear to the logarithmic scaling is defined by the ω parameter of the formula. It determines the width of the linearized data and its value is directly estimated from the fifth percentile of the values below zero. We noticed that this estimation method lacks accuracy when the outliers in the negative range are more than 5% of negative values and precision when the negative values acquired are low and with sparse values. To overcome the arbitrary estimation of the ω parameter, a cut-off at the value -111 has been suggested (Qian *et al.*, 2012). Nevertheless, this procedure does not have any theoretical explanation either and, as the authors of the logicle transformation method also implied, the truncation of the values would deform the distribution of the negative population and result in an improper estimation of its statistics (Parks *et al.*, 2006). Our idea is to use an outlier detection method to remove only the negative values that stray from the ones that compactly aggregate around zero. In Figure 2d, we depicted the differences among the distributions of the logicle transformed data for a channel of the 220662.fcs file where the ω parameter was estimated on the raw data, after removing the

negative outliers and after truncating the data at −111. Generally speaking, a better estimation of the parameters of a negative cell population is expected, since the data are neither affected by outliers nor by a truncation to an arbitrary threshold. Overall, although this procedure might not give any substantial advantage for downstream manual analysis, it should improve the quality of the results for any kind of automatic analysis, from simple statistics calculations to gating.

A last issue to consider when analyzing FCM data is the signal which value exceeds the limitations of the machine, thus generating the so called margin events. In fact, the signal can only be recorded up to the upper value of a dynamic range pre-set by the manufacturer of a FCM instrument. Therefore, it is impractical to discern subpopulations of cells whose values are stored in the upper margin of the dynamic range. This is already a common practice especially among computational biologists that require clean data to improve the quality of the analysis which is why we implemented it in our pipeline.

3.3 Overall improvement using computational methods

In the previous sections, we described each step of our pipeline separately in order to examine the anomalies from different perspectives. Instead, in this section, we look at the final results using approaches to analyze the multi-dimensionality data in its entire complexity.

3.3.1 Disappearance of undefined populations in high quality data

We used SPADE to identify and visualize populations from high dimensional FCM data (Qiu *et al.*, 2011). In brief, SPADE first prunes high density regions, second identifies clusters and third links them together with a minimum spanning tree.

The SPADE results before and after quality control of the file 220662.fcs are reported in Figure 3. The FCS file was part of an experiment where the functionality of CD4 and CD8 T cells in response to an HIV vaccination was assessed through intracellular cytokine staining. Looking at the SPADE results through the markers CD3, CD4 and CD8 it is possible to identify CD4 T cells at the bottom-right branch and CD8 T cells at the top-right branch (Fig. 3a).

The analysis was made with default settings and, from the 200 populations identified by SPADE in the original file (Fig. 3a), 43 disappeared in the high quality data (Fig. 3b and Supplementary Fig. S7). To explain the nature of the faulty populations, we examined the graphs reporting the coefficient of variation and a high variability was found for the markers CD3 and CD8 in the discarded populations. One may also suspect that those are new undefined populations that solicit further investigation. However, plotting the CD3 channel against FSC-A with the flowJo software, it was possible to identify the faulty populations only in the files with high instability in the flow rate (Supplementary Fig. S7).

3.3.2 Erratic populations revealed using dimensionality reduction

Another approach consisted in applying a dimensionality reduction method, t-SNE (Maaten and Hinton, 2008), to capture non-linear relationships in the high dimensional space with the intensity values of high and low quality events. For the analysis we used the R package cytofkit that includes also an algorithm based on support vector machine to identify the clusters from the new components defined by t-SNE (Supplementary Fig. S8a and b).

Using 2D plots of the first two components, we noticed that in most of the files a fraction of low quality cells was still superimposing to the populations of high quality cells while a remaining

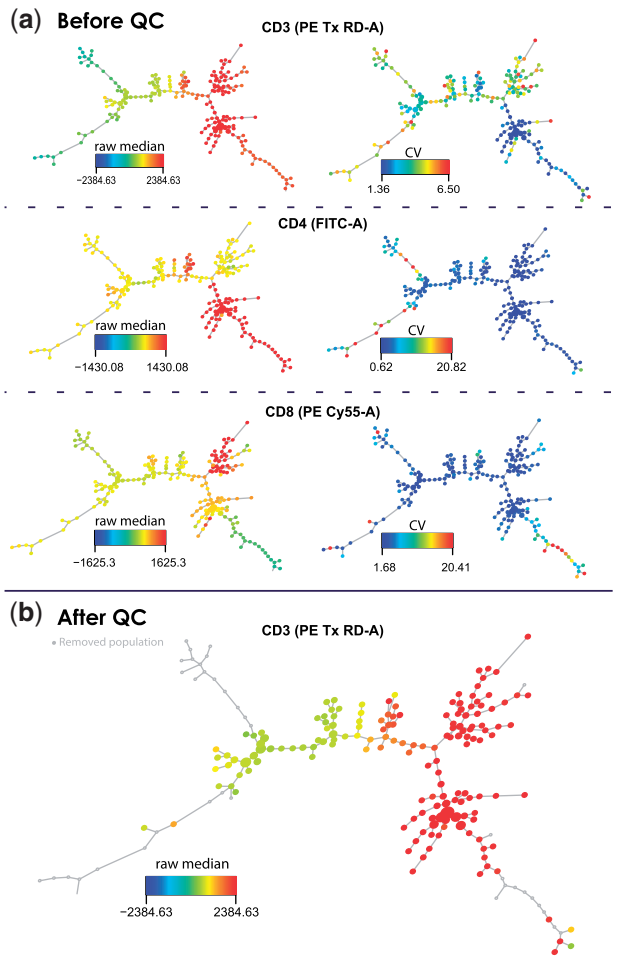

Fig. 3. SPADE analysis of the file 220662.fcs from the ZZZV dataset (**a**) before and (**b**) after quality control. The raw intensity median values and the coefficient of variation of the CD3, CD4 and CD8 channels are used as color-code for the population identified by SPADE. In (a) it is possible to localize the populations of CD4 and CD8 T cells and nodes with high coefficient of variation. In (b) the grey nodes are the population removed by the quality control

fraction formed separate sub-populations of events. In an FCS file from the SLAS dataset (Panel 1), we ascertained that the new populations in the low quality data mainly derived from dead cells and margin events; the borders are jagged and the shape is irregular reflecting the erratic nature of the acquired signal (Supplementary Fig. S8b). In contrast, the populations of high quality cells have smooth borders and a regular round shape.

T-SNE was then computed on a B cell population preprocessed with flowJo, where debris, doublets and dead cells were removed (Supplementary Fig. S8c and d). In Supplementary Figure S8c, an irregular CD19 population is revealed that was not found in the analysis of the raw data (Supplementary Fig. S8b). Further analysis revealed that the expression values of the CD19 channel were recorded at the upper margin of the dynamic range. This demonstrates that anomalies in only one channel can be easily camouflaged as valid cell populations in a multi-dimensional analysis if a careful quality control has not been applied beforehand. Lastly, in Supplementary Figure S8d, a significant shift in the average acquisition signal is visible in the t-SNE analysis by the formation of adjacent complementary population.

In summary, we advocate the importance of making a comprehensive cleaning on the data from different perspectives. Once faulty

Table 1. Pairwise agreement scores among the quality control made manually with flowJo, and automatically with flowAI and flowClean

Dataset (*n*)*	Median kappa coefficients (*n*)**		
	flowJo – flowAI	flowJo – flowClean	flowAI – flowClean
ZZZV (240)	0.9 (177)	0.25 (88)	0.26 (86)
ZZZU (308)	0.33 (255)	0.33 (3)	0.26 (64)
ZZ99 (766)	0.81 (390)	0.7 (327)	0.82 (328)
SLAS panel I (84)	0.07 (73)	0.23 (4)	0.018 (3)
SLAS panel II (84)	0.57 (82)	0.1 (43)	0.07 (39)

*Total number of files per dataset.

**Total number of Cohen's kappa tests with P-value < 0.05 selected for the calculation of the median kappa coefficient.

signals are included in downstream analyses, it becomes hard to detect them and they would eventually lead to false discoveries.

3.4 Benchmarking and performance

The automatic method in flowAI was compared both with a manual quality control using flowJo and the method in R package flowClean. The flowQ package was excluded from the comparison because it does not actively detect anomalies.

3.4.1 Agreement assessment with other approaches

A fundamental element for the quality control of an FCS file to is the time channel. The datasets ZZYA, ZZY2, ZZY3, ZZYY, ZZY6 and ZZYZ seemed to be already pre-processed and did not have a proper time channel. Although flowAI is still able to check the signal and dynamic range of a FCS file without the time channel, in this case it is impossible for flowClean and impractical for flowJo to do the quality control. Therefore, only the remaining datasets with a proper time channel were used for the benchmarking.

The flowJo analysis was executed by removing the margin events from the FSC-A and SSC-A scatterplot and unstable acquisition regions from the channel with more visible anomalies plotted against time. Regarding flowClean and the automatic method in flowAI, they were both run with default settings. The Cohen's kappa test was used to measure the agreement of two quality control methods on each FCS file. For the kappa statistic, a minimum value of anomalies was required to reach the significance level. For each dataset, the median of the significant kappa coefficients has been reported in Table 1.

Overall, flowAI showed a stronger agreement with the manual quality control than with flowClean. Also, flowAI was the most stringent towards anomalies while flowClean was the most tolerant (Table 1 and Supplementary Fig. S7). Nonetheless, both flowAI and flowClean still require a fine tuning of the settings for certain datasets to perform optimally. For example, better agreements would have been reached for the SLAS panel I dataset if less stringent settings were used for flowAI. In this respect, a decisive advantage of flowAI is its intuitiveness. In fact, based on the flow rate and signal plots, it is relatively easy to establish if the settings have to be more or less stringent. On the contrary, we found the diagnostic plot produced by flowClean harder to interpret.

3.4.2 Running time

The running time of the automatic method in flowAI was measured on a laptop with a 2.7 GHz CPU and 16 GB of RAM. We used four batches of datasets to evaluate the time performance. Each batch consists of five datasets of increasing size (100, 500, 1000, 1500 and 2000 MB) formed using an increasing number of FCS files with same size, number of events and parameters (Fig. 4).

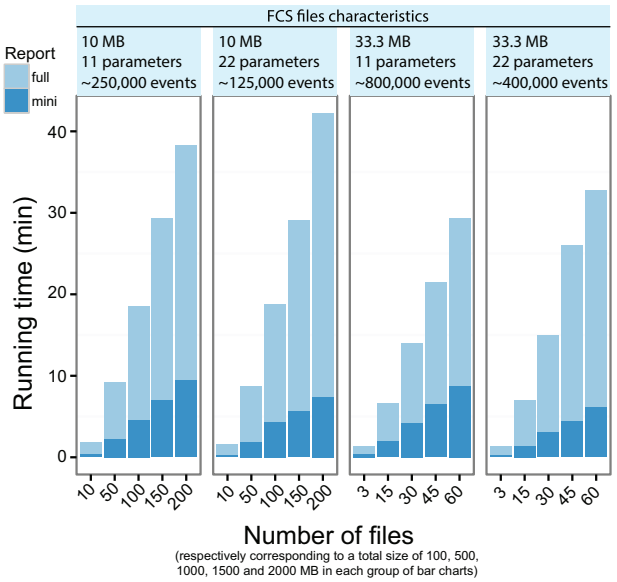

Fig. 4. Running time of a quality control analysis with the automatic method of flowAI. The graphics' creation for the full report, that is fundamental for an accurate examination, takes a considerable amount of time. Alternatively, a mini-report containing only the percentages of anomalies is produced without significant running time increase

The speed of flowAI is mostly influenced by the size of the FCS file rather than the number of parameters or events and the creation of the graphics for the full report takes the greatest amount of time. The possibility of creating a mini report containing only the percentages of anomalies is provided but it is discouraged for now, unless the user is sure of the nature of the anomalies in the entire dataset.

On the contrary, the running time for flowClean increases considerably with the number of parameters because of its way of defining cell populations through combinations of positive signals from the different parameters (Supplementary Fig. S9).

Overall, flowAI performance was faster for all the datasets used and, in particular, at least 3 times faster when using FCS files with 22 parameters (Fig. 4 and Supplementary Fig. S9).

4 Conclusion

Over the last few years, we have seen increasing efforts in automatizing pipelines of biomedical data analysis through computational algorithms. FCM is still one technique that hardly abandons the concept of manual analysis since usually the data produced has high variability that requires human interpretation. Often, the analysis demands high expertise and the results are still conditioned by a subjective evaluation. Our idea was born from the intention of removing the technical variability of FCM data in an objective way, thus reducing subjectiveness in interpretations and improving the performance of downstream computational analyses. This is especially the case when a high number of files is analyzed and when anomalies are generated by multiple sources.

We defined an approach and created an R package, flowAI, to automatically or interactively detect anomalies in FCM data. First, anomalous patterns and peaks are removed from the flow rate automatically by a method built upon time series decomposition and the

Generalized ESD test. Second, the tool checks the stability of the signal over time for each channel; here the automatic method uses a changepoint algorithm to detect durable shifts in the mean or variance of the acquisition values. Lastly, the dynamic range of the values acquired for each channel is refined. The upper limit is cleared of the margin events and the lower limit is cleared of the negative outliers.

From the use of the flowAI package, we expect a general improvement in the quality of research that employs FCM instruments. Removing events with erratic intensity values will facilitate different aspects of FCM analysis such as: (i) more effective compensation since the overlap signal is subtracted only from real values; (ii) more accurate detection of rare cells due to the removal of background noise; (iii) easier characterization of the nature of an ambiguous cell population (either as undefined cell type or as technical issue).

When using the automatic method for the quality control of a dataset of FCS files, it is preferable to infer the optimal settings for a dataset using a sample of few FCS files. In fact, because of the intuitiveness of the flowAI report, it is easy to infer the source of recurrent anomalies in a FCM experiment. Subsequently, the automatic method of flowAI can be run on the entire dataset with customized settings. Lastly, because the automatic quality control might still not meet the expectations for certain FCS files, the checking of the full reports reveals where it is necessary to intervene manually with the interactive method of flowAI or with another method. This last point is a limitation of flowAI that could be overcome by the dynamic adjustment of the settings of the automatic method, but for now it remains an open question that warrants further investigation. An additional consideration is that flowAI is designed to detect anomalies within a single FCS file, hence, other tools are necessary to check for anomalies between batches of FCS files. Also, another challenging task is the designing of a complete automatic pre-processing pipeline.

In conclusion, our quality control approach produces a comprehensive check of the FCM data implementing algorithms never employed before. We recommend the usage of flowAI as a first preprocessing step of the data right after they are obtained from the FCM instrument so that all the downstream analyses, from compensation to detection or rare cells, will benefit from it.

Acknowledgements

We thank Xavier Camous and Immanuel Kwok for the helpful discussions about the manual gating of B cell subpopulations, Bernett Lee for providing critical comments about FCM instrumentations and the flow cytometry facility at Singapore Immunology Network (SIgN) for processing the data.

Funding

This work was supported by the A*STAR/SIgN core grant, the A*STAR Joint Council Office DP grant [1434m00115] and the A*STAR Research Attachment Program (ARAP) to GM.

Conflict of Interest: none declared.

References

Aghaeepour,N. *et al.* (2013) Critical assessment of automated flow cytometry data analysis techniques. *Nat. Methods*, **10**, 228–238.

Balcilar,M. (2007) mFilter: Miscellaneous time series filters. *R Packag. version 0.1-3*.

Barnett,D. and Reilly,J.T. (2007) Quality control in flow cytometry. In: *Flow Cytometry: Principles and Applications*. Humana Press, Totowa, NJ, pp. 113–131.

Bashashati,A. and Brinkman,R.R. (2009) A survey of flow cytometry data analysis methods. *Adv. Bioinf.*, **2009**, 584603.

Becher,B. *et al.* (2014) High-dimensional analysis of the murine myeloid cell system. *Nat. Immunol.*, **15**, 1181–1191.

Brown,M. and Wittwer,C. (2000) Flow cytometry: principles and clinical applications in hematology. *Clin. Chem.*, **46**, 1221–1229.

Chandola,V. *et al.* (2009) Anomaly detection: a survey. *ACM Comput. Surv.*, **41**, 1–58.

Chang,W. *et al.* (2105) shiny: Web Application Framework for R.

Christiano,L.J. and Fitzgerald,T.J. (2003) The band pass filter. *Int. Econ. Rev. (Philadelphia)*, **44**, 435–465.

Edwards,W.F. and Cavalli-Sforza,L.L. (1965) A method for cluster analysis. *Biometrics*, **21**, 362–375.

Finak,G. *et al.* (2014) OpenCyto: an open source infrastructure for scalable, robust, reproducible, and automated, end-to-end flow cytometry data analysis. *PLoS Comput. Biol.*, **10**, e1003806.

Fletez-Brant,K. (2014) flowClean: flowClean. *R Packag. version 1.6.0*.

Gentleman,R. *et al.* (2006) flowQ: Quality control for flow cytometry. *R Packag. version 1.30.0*.

Hahne,F. *et al.* (2009) flowCore: a Bioconductor package for high throughput flow cytometry. *BMC Bioinformatics*, **10**, 1–8.

Hahne,F. *et al.* (2009) flowStats: Statistical methods for the analysis of flow cytometry data. *R Packag. version 3.28.1*.

Iglewicz,B. and Hoaglin,D.C. (1993) *How to Detect and Handle Outliers*. ASQC Basic References in Quality Control, ASQC, Milwaukee, WI, Vol. 16.

Jaye,D.L. *et al.* (2012) Translational applications of flow cytometry in clinical practice. *J. Immunol.*, **188**, 4715–4719.

Killick,R. and Eckley,I. (2014) changepoint: an R Package for changepoint analysis. *J. Stat. Softw*, **58**, 1–19.

Leys,C. *et al.* (2013) Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. *J. Exp. Soc. Psychol.*, **49**, 764–766.

Maaten,L.V.D. and Hinton,G. (2008) Visualizing data using t-SNE. *J. Mach. Learn. Res*, **9**, 2579–2605.

Mulley,W.R. and Kanellis,J. (2011) Understanding crossmatch testing in organ transplantation: a case-based guide for the general nephrologist. *Nephrology*, **16**, 125–133.

Newell,E.W. *et al.* (2012) Cytometry by time-of-flight shows combinatorial cytokine expression and virus-specific cell niches within a continuum of CD8+ T cell phenotypes. *Immunity*, **36**, 142–152.

Novo,D. and Wood,J. (2008) Flow cytometry histograms: transformations, resolution, and display. *Cytometry A*, **73A**, 685–692.

Oldaker,T.A. (2007) Quality control in clinical flow cytometry. *Clin. Lab. Med.*, **27**, 671–685.

Parks,D.R. *et al.* (2006) A new 'Logicle' display method avoids deceptive effects of logarithmic scaling for low signals and compensated data. *Cytometry A*, **69A**, 541–551.

Perfetto,S.P. *et al.* (2006) Quality assurance for polychromatic flow cytometry. *Nat. Protoc.*, **1**, 1522–1530.

Pozarowski,P. and Darzynkiewicz,Z. (2004) Analysis of cell cycle by flow cytometry. *Methods Mol. Biol.*, **281**, 301–311.

Qian,Y. *et al.* (2012) FCSTrans: an open source software system for FCS file conversion and data transformation. *Cytometry A*, **81A**, 353–356.

Qiu,P. *et al.* (2011) Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat. Biotechnol.*, **29**, 886–891.

Rosner,B. (1983) Percentage points for a generalized ESD many-outlier procedure. *Technometrics*, **25**, 165–172.

Sarkar,D. *et al.* (2008) Using flowViz to visualize flow cytometry data. *Bioinformatics*, **24**, 878–879.

Seamer,L.C. *et al.* (1997) Proposed new data file standard for flow cytometry, Version FCS 3.0. *Cytometry*, **28**, 118–122.

Shekhar,K. *et al.* (2014) Automatic classification of cellular expression by nonlinear stochastic embedding (ACCENSE). *Proc. Natl. Acad. Sci. USA*, **111**, 202–207.

Spidlen,J. *et al.* (2012) FlowRepository: a resource of annotated flow cytometry datasets associated with peer-reviewed publications. *Cytometry A*, **81A**, 727–731.

Van Gassen,S. *et al.* (2016) FloReMi: flow density survival regression using minimal feature redundancy. *Cytometry A*, **89**, 22–29.

Vermes,I. *et al.* (2000) Flow cytometry of apoptotic cell death. *J. Immunol. Methods*, **243**, 167–190.