flowCut — An R package for precise and accurate automated removal of outlier events and flagging of files based on time versus fluorescence analysis

Justin Meskas*, Sherrie Wang† and Ryan Brinkman‡

Running head: flowCut — An R package for precise and accurate automated removal of outlier events and flagging of files based on time versus fluorescence analysis

*E-mail: justinmeskas@gmail.com
†E-mail: sherrie2865@gmail.com
‡E-mail: rbrinkman@bccrc.ca

Abstract

Technical artifacts that occur during the data acquisition process of cytometry data can impact the quality of the data and its analysis results. Common technical issues, such as clogging, can cause spurious events and fluorescence intensity shifting. These events should be identified and potentially removed before being passed to the next stage of the gating analysis. flowCut, an R package, automatically detects anomaly events and flags files for flow cytometry experiments. Results are on par with manual analysis and it outperforms existing automated approaches.

Availability: R package: https://github.com/jmeskas/flowCut. Test data uploaded to FlowRepository (Repository ID: FR-FCM-ZYPD) along with the manual results.

Key words: data quality control, flow cytometry

1 Introduction

Even when proper protocols are followed, technical issues such as clogging can result in spurious events in flow cytometry data. Removing those instrumental level problematic events can improve data quality and subsequent downstream analysis. We developed flowCut to address shortcomings and improve performance over the two available automated data cleaning tools (flowAI [1], flowClean [2]). flowClean has a minimum 30,000 cells requirements and, often times, is not able to capturing sufficient abnormal regions, as shown in Figure 1. flowAI can be aggressive, as seen in Figure 1 where flowAI removed a large amount of normal regions in addition to outlier regions. To address the current limitations of the two algorithms, we present a new algorithm, flowCut, as a new state-of-the-art flow cytometry quality control tool. We evaluated flowCut's performance against that of flowClean and flowAI.

[Figure 1 about here.]

2 Methods

flowCut separates events along the time axis into equally sized segments (parameter defaulted at 500 events) and calculates if any should be removed. This parameter as well as all other numerical user-definable parameters have robust defaults based on extensive testing (described fully in the supplemental materials). flowCut determines if a file needs cleaning based on three tests with user-definable parameters. These three threshold parameters are: the maximum range of the means of each segment across all channels (*MaxOfMeans*), the average range of the means of each segment over all channels (*MeanOfMeans*), and the maximum continuous jump between adjacent segments (*MaxContin*).

Figure 2 (a) shows how the mean drifts a proportion of 0.231 (proportion of the distance between the brown lines, the 2nd and 98th percentiles), where the points of the pink and brown jagged line are the means of the segments. The numbers on top of Figure 2 (a) indicate mean drift before cutting, mean drift after cutting, and the max one step mean change between adjacent segments after cutting. The mean drifts such as the 0.231 are the *Means* when calculating the *MaxOfMeans* and the *MeanOfMeans*. Since *MaxOfMeans* is defaulted at 0.15, the file shown in Figure 2 (a) will undergo cleaning because $0.231 > 0.15$.

If flowCut deems the file needs cleaning, then eight measures of each segment are calculated (See Figure 2 (c)): the mean, the median, the 5th, 20th, 80th and 95th percentiles, the second moment (variation) and the third moment (skewness). These measures are absolute Z-scores (absolute because we are only interested in the difference and not the direction). Since the Z-scores are naturally normalized, flowCut can sum the eight measures across all channels and all measures, resulting in each segment being represented by only one statistical number. We utilize the deGate function from flowDensity [3] to find the natural point in the distribution to remove significantly different segments (See Figure 2 (d)). flowCut then repeats the three tests explained above, and also validates if the file is monotonically increasing in time. If at least one of the four tests are not passed, the file is flagged.

[Figure 2 about here.]

3 Tests and Results

We chose 55 exemplary files from FlowRepository [4] to compare the analysis of the three algorithms and a random sampling to manual analysis (see supplementary materials for details). The random sampling was to compare the F1 scores of the algorithms to the F1 score a random cleaning. The percentage of events removed with random sampling was equal to the amount of events removed by manual analysis.

For each of the 55 files, we did blind manual analysis that visually identified the problematic regions in the fluorescence vs time domain in every channel after compensation, transformation and margins removal. Problematic regions include spikes, fluorescence drifting, discontinuity or low density.

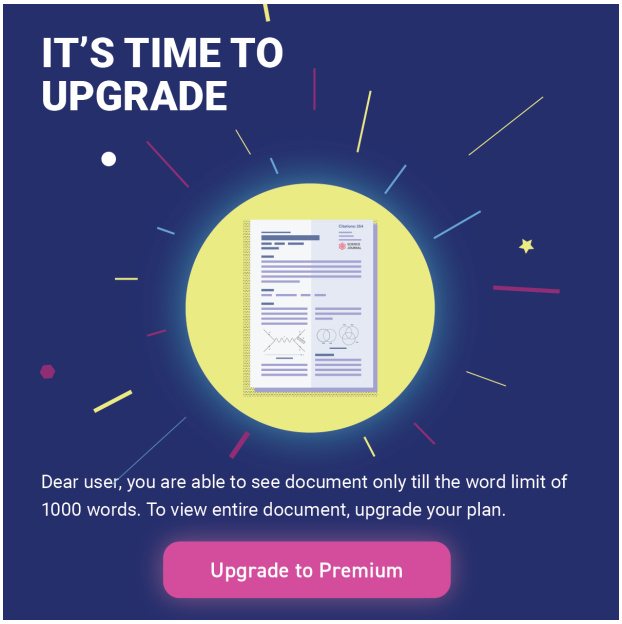
3.1 Comparing Algorithms with Manual Gating — File Based Evaluation

We then subdivided the 55 files into three categories based on confidence in the manual analysis. The first category includes 17 files that have removal regions with clearly defined boundaries such as discontinuity, low density regions and large spikes. The second category has 35 files that have fuzzy boundaries for removal regions. Examples include small spikes and cutoff in fluorescence drifting regions. This category also includes files

that have overlapping problematic regions detected by at least two algorithms but not by manual analysis. The third category has 3 files which manual analysis is arbitrary. The files look abnormal but the regions for removal are not clear (see supplementary material). Category three files contain no region of truth and are removed from being used as standards for evaluation. For the remaining 52 files, we calculated F1 scores for two categories for each algorithm.

The F1 score is , where precision is defined as the proportion of each algorithm's intersection with manually removed events in total manually removed events; recall is defined as the proportion of each algorithm's intersection

Edit this article on Typeset | Reformat this article



www.typeset.io | Get in touch with us | Looking for a Discount?

Acknowledgements & Funding

Special thanks to Sibyl Drissler and Jessica Tuengel for their testing, feedback and criticism.

Research reported in this publication was supported in part by NSERC, NIGMS [R01 GM118417-01A1], an NIH Infrastructure and Opportunity Fund Award linked to Human Immunology Project Consortium Award Number [U19AI118610] (IOF Pilot), and the Precision Vaccines Program.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

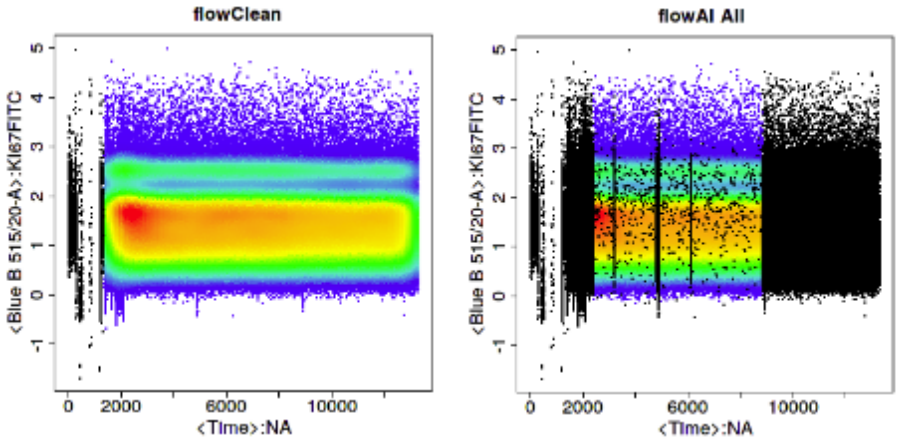


FIGURE 1: Quality control by flowClean (left) and flowAI (right) on the same file. The colored regions are fluorescence intensity signals. Red is the most dense, followed by yellow, green, and blue. Black is the removed regions.

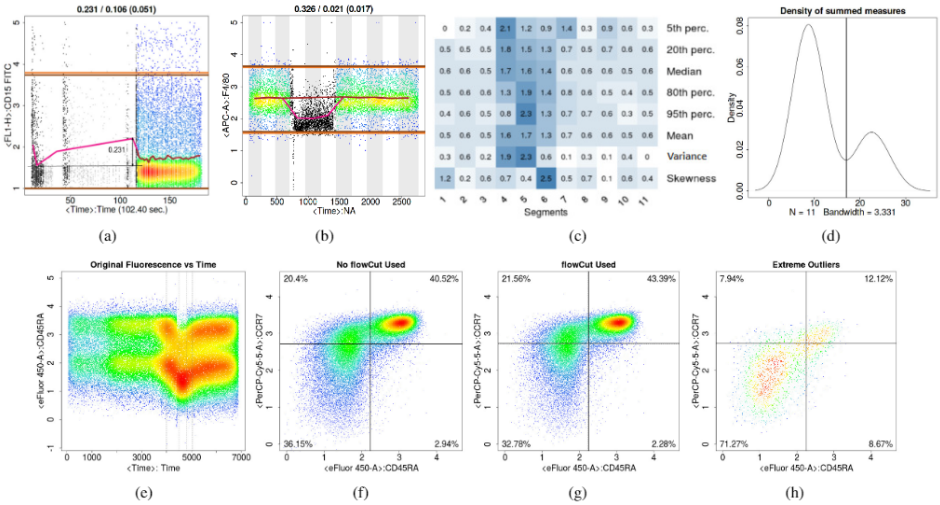


FIGURE 2: flowCut methodology. (a) flowCut removes low density regions (grey) and statistically different segments (black). (b)-(d) represent the same file. flowCut removes the problematic regions (black in (b)) that have higher statistical values indicated by the darker blue colour in (c) for all eight measures. The 11 columns in (c) correspond to the 11 segments shown in (b) by alternating background colour. We sum over all the measures across all channels to get a statistical score for each segment, and use these 11 values to create a density curve shown in (d). We find a threshold (vertical line in (d)) where the segments with high statistical values, which reside to the right, are removed. (e) shows a new file with fluorescence drifting for CD45RA (Similarly for CCR7 — not shown). (e)-(h) show the difference between (f) not using and (g) using flowCut. (h) shows only the gated events between the middle two grey vertical lines in (e) (the outside grey vertical lines are where flowCut removed events).