ISAC CYTOMETRY
Journal of Quantitative Cell Science PART A

Automated Data Cleanup for Mass Cytometry

Charles Bruce Bagwell,[1*] Margaret Inokuma,[1] Benjamin Hunsberger,[1] Donald Herbert,[1] Christopher Bray,[1] Beth Hill,[1] Gregory Stelzer,[2] Stephen Li,[2] Avinash Kollipara,[3] Olga Ornatsky,[2] Vladimir Baranov[2]

[1]Verity Software House, Topsham, Maine

[2]Fluidigm Canada Inc., Markham, Ontario, Canada

[3]Department of Pediatrics and Department of Microbiology and Immunology, University of North Carolina, Chapel Hill, North Carolina

Received 6 June 2019; Revised 24 September 2019; Accepted 7 October 2019

Grant sponsor: NIH, Grant numberAI119164, Grant numberRO1

*Correspondence to: C. Bruce Bagwell, Verity Software House Topsham, Maine PO Box 247 Email: cbb@vsh.com

Published online in Wiley Online Library (wileyonlinelibrary.com)

DOI: 10.1002/cyto.a.23926

Abstract

Mass cytometry is an emerging technology capable of 40 or more correlated measurements on a single cell. The complexity and volume of data generated by this platform have accelerated the creation of novel methods for high-dimensional data analysis and visualization. A key step in any high-level data analysis is the removal of unwanted events, a process often referred to as data cleanup. Data cleanup as applied to mass cytometry typically focuses on elimination of dead cells, debris, normalization beads, true aggregates, and coincident ion clouds from raw data. We describe a probability state modeling (PSM) method that automatically identifies and removes these elements, resulting in FCS files that contain mostly live and intact events. This approach not only leverages QC measurements such as DNA, live/dead, and event length but also four additional pulse-processing parameters that are available on Fluidigm Helios™ and CyTOF® (Fluidigm, Markham, Canada) 2 instruments with software versions of 6.3 or higher. These extra Gaussian-derived parameters are valuable for detecting well-formed pulses and eliminating coincident positive ion clouds. The automated nature of this new routine avoids the subjectivity of other gating methods and results in unbiased elimination of unwanted events. © 2019 International Society for Advancement of Cytometry

Key terms

Gaussian parameters; probability state Modeling; quality control; unattended analysis

CyTOF instruments were invented as an alternative technology to flow cytometry (1,2). Instead of fluorescent molecules, heavy metals are conjugated to antibodies, which bind to specific epitopes on cells (3). Metal-labeled cells are detected and quantified by inductively coupled plasma mass spectrometry (ICP-MS) with time-of-flight detection (4). Because this technology avoids spectral overlap of fluorescent dyes and isotopic metal contamination is well less than 5%, the number of correlated measurements is mainly limited by the number of stable isotopes of rare earth metals, which realistically could soon reach 100 or more (5).

Whether a cytometer is fluorescence-based or metal-based, there are always undesired events that need to be eliminated prior to analysis. Typical flow cytometers have internal circuitry or logic that ignores signal-derived pulses that are partially formed or abnormally long. Flow cytometry often employs a forward by 90° light-scatter gate to eliminate debris and aggregates. Also, pulse processing features such as peak height, width, and area can be leveraged to reduce the number of aggregates.

Since mass cytometry atomizes particles into clouds of positively charged ions, its pulse processing capabilities are mainly targeted at detecting and eliminating coincident ion clouds or poorly formed pulses. Mass cytometry also has DNA intercalators (1) that can eliminate debris and some true aggregates. Both technologies

typically employ some membrane disruption measurement to eliminate dead cells (1).

This article first describes the parameters and measurements as well as the strategies used to clean up data and then describes in detail how PSM automatically eliminates unwanted events by using modeling techniques. It then presents data that demonstrate the reproducibility and unbiased nature of the method.

MATERIAL AND METHODS

PBMC Specimens

Multiple lots of cryopreserved PBMC from healthy donors were obtained from a commercial biological specimen supply source (Discovery Life Sciences, Huntsville, AL). Peripheral blood mononuclear cell (PBMC) samples were also collected by Canadian Blood Services (Vancouver, BC) using BD Vacutainer® Glass Mononuclear Cell Preparation Tubes (CPT™) (Becton Dickinson, Franklin Lakes, NJ). Human whole blood was collected in CPT tubes and then centrifuged at 1,800*g* for 15 min to separate the buffy coat. Tubes were shipped overnight on 4°C gel packs, and cells were collected, washed, and frozen the following day. (We are grateful to Canadian Blood Services and donors for providing research samples for completion of this project. The reporting and interpretation of the research findings are the responsibility of the authors, and the views expressed herein do not necessarily represent the views of Canadian Blood Services.)

PBMC Staining

Vials of cryopreserved PBMC were thawed and washed. The viability and cell count were determined, and the cells were washed in Maxpar® cell staining buffer (CSB) at cold temperatures ranging from 4 to 8°C. After the wash, the cells were resuspended in CSB to a concentration of 6×10^7 cells/ml. Fc receptors were blocked by adding 5 µl of Human TruStain FcX™ to 3×10^6 cells in 50 µl and incubated for 10 min. About 215 µl of CSB was then added to the PBMC and 270 µl of PBMC was directly added to each dried antibody tube for antibody staining. After a 30-min incubation, the cells were washed twice in CSB, followed by fixation in 1.6% paraformaldehyde for 10 min. Following fixation, the cells were centrifuged to a pellet, the fixative was removed, and the pellet was resuspended in 1 mL of the 125 nM Cell-ID™ Intercalator-Ir (1) and incubated overnight at 4°C.

Sample Acquisition

Following the overnight incubation, the fixed cells were washed twice in CSB with a final resuspension of the cells at 1×10^6 cells/ml in Maxpar Cell Acquisition Solution (CAS) containing 0.1× EQ™ Four Element Calibration Beads. Acquisition was performed on a Helios system utilizing CyTOF Software version 6.7.1016. All instruments were equipped with a WB Injector and samples were acquired in CAS. All instruments are routinely evaluated to ensure the performance at or above the minimum Helios system specifications for calibration. Following instrument tuning and bead sensitivity testing, the system was preconditioned with CAS. A minimum of 400,000 events were acquired per file at a typical acquisition rate of 250–500 events/second. The files that were used to generate Table 3 are available in the flow repository, Cleanup for Mass Cytometry (FR-FCM-Z29V).

Acquisition Rate Experiment

Fourteen files generated for the acquisition rate experiment (Figure 7 and Table 2) are available in the flow repository, Acquisition Rate Experiment (FR-FCM-Z29U).

Live/Dead Discrimination Experiment

The four files created for the live/dead Discrimination experiment (see Figure 9) are available in the flow repository, Live/Dead Discrimination Experiment (FR-FCM-Z2AZ).

Custom Panel Experiment

Eight FCS 3.0 files from a single control sample of PBMC were provided by University of North Carolina to further test the reproducibility of data after the automated Gaussian parameter cleanup procedure. The UNC Reproducibility data set was derived from a Human Peripheral Blood Leukapheresis Pack (half-sized) Cat # 70500 (Stem Cell Technologies, Canada). Cells were labeled with a phenotype panel comprising 34 antibodies. Vials of cryopreserved PBMCs from the leukapheresis pack were thawed and initially washed in Thaw buffer (RPMI 1640 medium containing 50 Units/ml Benzonase Nuclease, Ultrapure; Cat # E8263 [Sigma-Aldrich]) and then washed with RPMI media without Benzonase. The viability and cell counts were determined. The enumerated cells were resuspended in 1 ml of CyTOF 1× PBS (cPBS) to a concentration of 3×10^6 cells/ml. Cell-ID Cisplatin (198-Pt) (1:1,000) was added to cells for 5 min at RT to stain dead cells. Fc receptors were blocked by adding 20 µl of Fc block (eBioscience; Cat # 14-9161-73) to 3×10^6 cells in 50 µl and incubated for 10 min. After a wash in Maxpar Cell Staining Buffer, pellet was resuspended in 1 mL of CSB. About 200 µls of cells containing 0.6×10^6 cells were stained with CD45_115In and 800 µls of cells containing 2.4×10^6 cells were stained with CD45_89Y in separate tubes. Individually stained cells were washed twice in 1× CSB and cell pellets containing both CD45 stained cells were pooled and stained for the remaining 32 markers. Cells were washed twice and fixed in 2% cold paraformaldehyde for 30 min. Following fixation, the cells were incubated in 1 mL of the 125 nM Cell-ID™ Intercalator-Ir and incubated overnight at 4°C until acquisition.

The data set was comprised of eight FCS files that were generated across four time points. At each time point, a stained sample was halved and 200,000 cells from each half were acquired at UNC and NIEHS mass cytometry core facilities. This data set is available in the flow repository, Reproducibility 2 Experiment (FR-FCM-Z2AY).

Data Normalization

After acquisition, data were normalized using the CyTOF Software v.6.7.1016. This method normalizes data to a global

Table 1. Expression profiles properties for the cleaned cell type

	EXPRESSION PROFILES								
PROPERTIES	BEADS	OFFSET	WIDTH	CENTER	DNA1	RESIDUAL	EVENT_LENGTH	LIVE/DEAD	DNA2
VLog max range	16,384	8,192	8,192	16,384	16,384	8,192	8,192	8,192	16,384
Initial intensity value	7.2	39.3	36.5	68.2	69.0	38.5	21.1	5.8	76.2
Initial SD value	8.6	5.0	6.0	2.5	2.6	5.0	2.5	4.8	2.8
Estimation range	10	20	20	20	40	20	20	25	40
Intensity estimate	CP	LP	LP	LP	TP	TP	LP	CP	TP
Degrees of freedom	1	2	3	4	5	6	7	8	9
Final intensity value	8.1	38.5	36.9	68.6	73.3	37.3	21.7	3.75	80.7
%Cleaned	97.3	91.8	91.7	85.2	79.3	78.5	73.2	73.9	72.4

standard, called a bead passport, determined for each lot of EQ beads. This passport contains a profile of mean Dual instrument (Di) counts of all the masses for the lot of the beads as determined by multiple measurements during manufacture of the EQ beads. The normalization factor is the ratio of passport median Di values to bead singlet population median Di values of the encoding isotopes. EQ Four Element Beads contain the isotopes of cerium (140/142Ce), europium (151/153Eu), holmium (165Ho), and lutetium (175/176Lu). Major isotopes 140Ce and 175Lu, 165Ho and 151/153Eu are used in determining normalization factors and for mass channels between the encoding isotopes. The other isotope normalization factors are then either linearly interpolated or extrapolated. All mass channel event values are then multiplied by these normalization factors to obtain the normalized values and data are written to the normalized file.

Measurement Transformations

A key attribute for each mass cytometer measurement is the type of transformation used to convert linear intensity values to transformed values. For all the measurements involved in selecting cleaned events, the type of variance-stabilizing transform was VLog (15) with $\alpha = 1$ and $\beta = 10$. Since Helios selects the next power of two for the maximum observed measurement values encoded in the FCS 3.0 keywords, $PnR, it was decided not to use this value for calculating transformed data because it would be affected by outlier events. Instead, a fixed maximum value was used for each measurement value (see Table 1, first row). The maximum values were picked such that for all files in the study, no file had live intact event measurement values greater than this maximum rangse. It was also set to be a power-of-two in order that other third-party software could easily read the exported FCS files.

Automated Analysis

All analyses were done by Fluidigm Pathsetter™, powered by GemStone 2.0.41, Verity Software House, Topsham, Maine. The automated analysis used two model templates: Cleanup Model, Version: 31Jan19, and the MIP Model, Version: 01Feb19.

Event Detection and Event Length

Mass cytometry has the advantageous characteristic of allowing users to construct their own strategies to eliminate events formed from partial pulses or coincident event pulses. Most mass cytometry studies leverage a measurement labeled Event_length to filter out either partial or abnormally long signal pulses.

It is helpful to have a reasonably clear picture of how the instrument determines Event_length. When cells are injected into a hot argon torch, they form a plasma where electrons have been stripped from the atoms. The positively charged ions from the heavy metal labels of antibodies ultimately collide with an ion detector (secondary electron multiplier) to form electron pulses that are then converted into voltage pulses. An excellent and detailed description of the process is provided by Olsen et al. (6). Digitized signals from all selected mass channels are summed together to the form a total current signal that is then used for event length determination. Typical duration of a single-cell event is approximately 300 µs. Since time of flight (TOF) spectrum is recorded every 13 µs, approximately 20 spectra (pushes) are recorded for each single-cell event (see Figure 1).

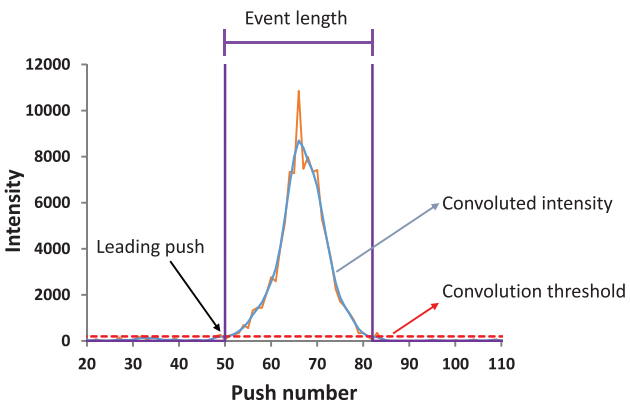

Figure 1. Event detection. The total current or intensity pulse is first smoothed with a convoluted Gaussian smoothing routine to eliminate unwanted noise. An event begins when smoothed intensities are higher than an internal threshold for at least 10 but no more than 150 consecutive pushes or digitizations. An event ends when the smoothed intensity drops below the internal threshold [Color figure can be viewed at wileyonlinelibrary.com]

The total current pulse is first smoothed with a convolution Gaussian routine to eliminate unwanted noise in the signal. An event begins when smoothed intensities are higher than an internal threshold for at least 10 but no more than 150 consecutive pushes. An event ends when the smoothed intensity drops below the internal threshold. The duration of this pulse in units of pushes is stored as the Event_length measurement in the FCS file. Since there are relatively few digitizations, a picket fence type of pattern is often apparent when viewing Event_length in histograms and dot plots (not shown). If noise in the signal causes a premature ending to the signal duration, the magnitude of Event_length is relatively small, whereas if two or more ion clouds are captured in the same pulse, it is relatively large.

DNA1 and DNA2

Typically, Event_length is coupled with one or two DNA content measurements labeled DNA1 and DNA2. These measurements are derived from cationic double-stranded nucleic acid intercalators that have a natural abundance of iridium (191Ir and 193Ir) or rhodium (103Rh) (1). These positively charged molecules normally do not pass through intact live cell membranes unless the cells are fixed.

If they are used prior to fixation, intercalators can select for live cells from dead cells, and if they are used after fixation, they can differentiate intact singlet nucleated cells from debris or aggregates. Since the iridium intercalator contains both 191Ir and 193Ir, users typically couple the two for their cleanup selection strategies. Care must be taken with these measurements not to bias the results, since different cell types have different intercalator stainabilities (7). Figure 2 shows a typical gating scheme that leverages DNA1, Event_length, and bead intensities. Often a spot gate in a DNA1 and DNA2 dot-plot is used to gate desirable events (not shown).

Live/Dead

If needed, a live/dead cell selection strategy is normally added to Event_length and the DNA intercalators. There are currently two classes of molecules that commonly distinguish live from dead cells. One method uses rhodium- or iridium-containing metallointercalators that form ionic bonds with double-stranded nucleic acids (1) after crossing disrupted cellular membranes. Because the bond is noncovalent, this reagent is normally added to live cells during antibody staining and must be used immediately prior to acquisition so that it is not washed away during the preparation procedure.

The other method involves staining the cells with a platinum-based chemotherapeutic agent called cisplatin, which ultimately forms covalent bonds with protein nucleophiles (8). Since cisplatin rapidly crosses compromised cell membranes, its absence can select for events representing live cells. Although several stable isotopes of platinum are present in cisplatin, 195Pt is the most abundant and is typically used for detection. The monoisotopic cisplatin 194Pt and 198Pt are also available for live/dead discrimination. Since cisplatin bonds with proteins, there may be a noticeable cell type difference in its intensity for dead cells.

Gaussian Parameters (Center, Width, Offset, and Residual)

The Helios system and some CyTOF 2 instruments with CyTOF Software of version 6.3 or higher have four additional embedded calculated measurements that provide information about the quality of the total ion current pulse. They are stored as FCS 3.0 measurements and are labeled Center, Width, Offset, and Residual.

The algorithm finds the average digitized and smoothed pulse value, subtracts it from each digitized value, and then divides the difference by the same average value. This procedure avoids the problem of finding a stable baseline. A typical normalized pulse is shown in Figure 3A.

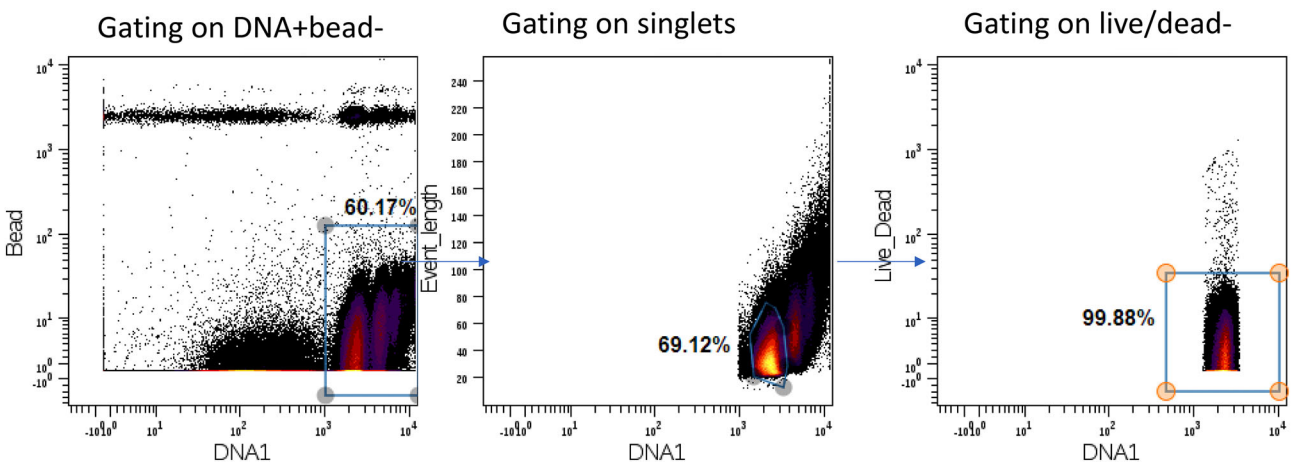

Figure 2. Typical gating strategy. Users generally use some combination of Bead, Event_length, DNA, and Live/Dead measurement intensities to gate out unwanted beads, debris, and aggregates. Care must be taken to avoid preferentially eliminating specific cell types like monocytes, since they have a higher stainability than cell types like T- and B-cells [Color figure can be viewed at wileyonlinelibrary.com]

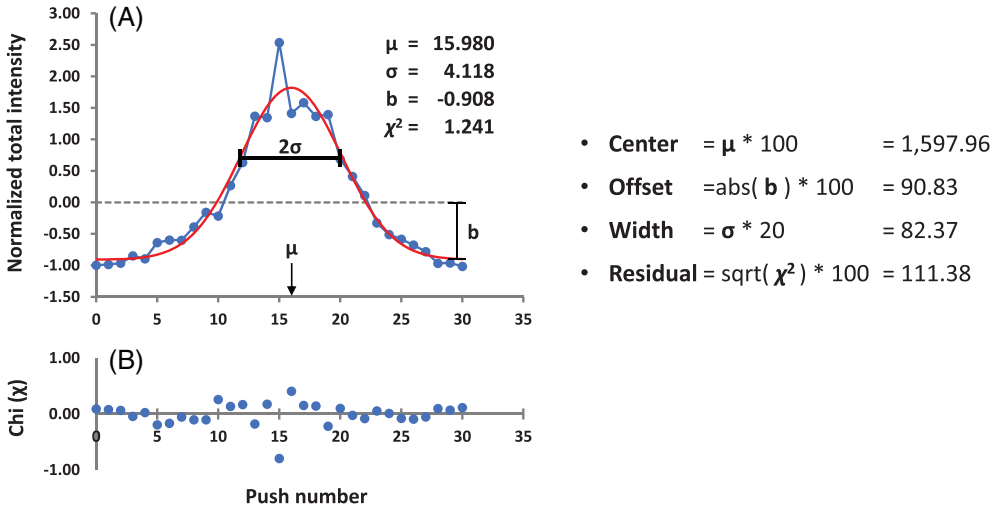

Figure 3. Gaussian discrimination parameters [Color figure can be viewed at wileyonlinelibrary.com]

Nonlinear least-squares analysis (9) are then performed on the normalized digital data with the Gaussian formula,

$$y = ae^{-\frac{(x-\mu)^2}{2\delta^2}} + b$$

The nonlinear least-squares algorithm finds the values for a, μ, δ, and b that best matches the observed ion current push data to the above formula. The mean of the fitted Gaussian is given by μ, the standard deviation, δ, the height above the zero line is "a,", and the distance below is "b." The residual data (see Figure 3B) are squared to eliminate signs and then summed to form a sum of chi-squares value.

After these Gaussian parameters and residual sum of chi-squares are derived from analyzing the pulse data, they are then scaled and renamed as shown in Figure 3 inset table. The Center measurement denotes the mean of the pulse, the Width is the standard deviation, Offset is the distance from the base of the pulse to the zero line, and Residual represents how well the Gaussian fitted the data.

The "a" parameter is not stored, presumably because it is likely to be highly correlated with "b." These Gaussian parameters can be leveraged by gating or modeling analysis strategies to eliminate unwanted non-Gaussian pulses as shown in Figure 4. Events from coincident clouds typically have either a low or high Center, low Width and Offset, and a high Residual. Selecting events using these Gaussian parameter distributions makes the final Event_length selection much easier

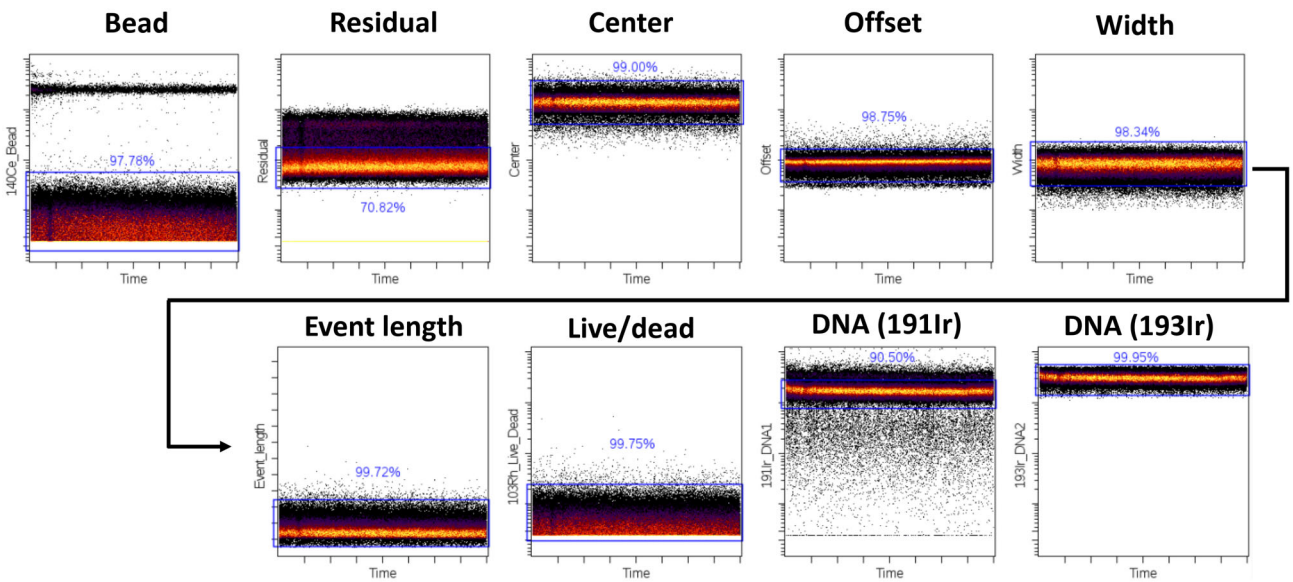

Figure 4. Typical gating strategy using Gaussian parameters (Gate GP+) [Color figure can be viewed at wileyonlinelibrary.com]

and therefore more reproducible since Event_length has a rather complicated distribution. These patterns along with Event_length provide a powerful means of eliminating undesired events based only on the shape of the pulse. Although Gaussian parameters are useful for eliminating events from merged ion clouds, they have no ability to eliminate true aggregates since an aggregate will be ionized into a single cloud of atoms.

Automated Probability State Modeling Strategy

Overview

Probability state models are composed of a set of cell types (10). A cell type is a specific type of population such as B cells or CD4 T-cells. For the Cleanup model, these cell types are Cleaned, Not Cleaned (no beads), and Plasma Temp QC (all beads). Within each cell type, there are a set of expression profiles (EPs) that control the fitting of specific measurements. The Cleaned cell type has EPs: Beads, Offset, Width, Center, DNA1, Residual, Event_length, Live/Dead, and DNA2 (Table 1). These EPs determine whether an event is to be exported as a live intact event or not.

The Plasma Temp QC cell type has EPs that evaluate the CeO+/Ce + ratio of intensities. Since the CeO+ covalent bond is one of the strongest for all heavy metals, it can be considered an upper range for all other heavy metal oxides. If the ratio falls below 4%, the temperature of the argon torch is considered hot enough to minimize the presence of other metal oxides. The modeling system is designed to illicit a warning if the ratio ends up being >4%. This threshold should not be confused with the Helios tuning solution, which has a LaO+/Tb + ratio of less than 3%.

Within each EP, there is at least one control definition point (CDP) that defines how the EP will fit the measurement data. Each CDP has three parameters associated with it: intensity (*y*-axis position), state (*x*-axis position), and standard deviation (SD or line-spread). The units for each of these parameters are relative transformed units that have a maximum positive range of 100. Since all EPs are constants, the state parameter is not a fitting parameter and therefore is irrelevant to the cleanup strategy.

Each of these parameters has one of seven ways it interacts with the detected peaks. These are User Estimate, Closest Peak (CP), Dimmest Peak (DP), Brightest Peak (BP), Smallest Peak (SP), Largest Peak (LP, greatest area), and Tallest Peak (TP, highest). Each of the EPs also has an estimation range property that controls how far away a selectable peak can be from the initial intensity position of the CDP for it to be chosen for estimates.

The fitting process generally begins by reading in a template model that is composed of cell types, EPs, and CDPs that are all preprogrammed to fit the data. For this section, an arbitrary FCS 3.0 Helios file will be used as an example. The algorithm finds the first enabled cell type and its first enabled EP. For the Cleanup model, these correspond to the cleaned cell type and beads EP (see Table 1 and Figure 5). The algorithm successively fits each EP within a cell type and then moves to the next enabled cell type. It continues in this manner until it has finished fitting all enabled cell types and their respective EPs.

The first step in fitting is a determination of a set of peaks for each EP (Figure 5). The locations of the detected peaks are shown as black triangles to the right of each EP. Peak parameters such as intensity and SD are controlled by a set of editable peak finder properties. For the Cleanup model, only the both sides type of peak-fitting method was used. The both sides option means that the SD estimation for the measurement line-spread uses both sides of the peak's frequency distribution for least-squares estimates of mean and SD. The initial positions of the events within an EP are their relative position in a file. However, once the algorithm has fit the first EP, the order of events along the *x*-axis is determined probabilistically. For constant EPs as found in the cleanup model, this process orders the events along the *x*-axis randomly.

Beads Expression Profile

This EP is currently linked to normalization beads that are coated with the metal 140Ce (Figure 5A). The model specifics for each involved EP are summarized in Table 1. Once the CDP has its final fitted intensity and SD values, it can convert each event to a chi-square value by the formula,

$$\chi^2_{i,0} = \frac{1}{2}\left(\frac{x_{i,0} - \mu_0}{\delta_0}\right)^2,$$

where $x_{i,0}$ is the ith event and zeroth transformed value, μ_0 is the first CDP intensity value, and δ_0 is the first CDP SD value.

At this point, the collection of chi-square values from the events form a chi-square probability distribution with a one degree of freedom. The probability of exclusion for the cleaned cell type is set to 0.01, which means that events that are outside the 0.99 high tail of this probability distribution are considered unclassified by that cell type. When events are classified into the cleaned cell type, their color changes from light gray to darker gray (Figure 5B). The Discussion section presents in more depth the engineering concepts involved in building and refining this PSM model. After the beads EP finished fitting this example data, the percentage of events classified into the cleaned cell type was 97.3% (Table 1, Beads column).

Offset Expression Profile

Offset is a Gaussian discrimination parameter that is low for multiple-peak pulses. In the template model the initial Intensity value is 39.3 and the Estimate is set for Largest Peak with an SD of 5.0 (Table 1, Offset column). After fitting, the Intensity parameter for the CDP was adjusted to 38.5 (Figure 5D). The two chi-squares were summed, forming a new chi-square distribution described by two degrees of freedom.

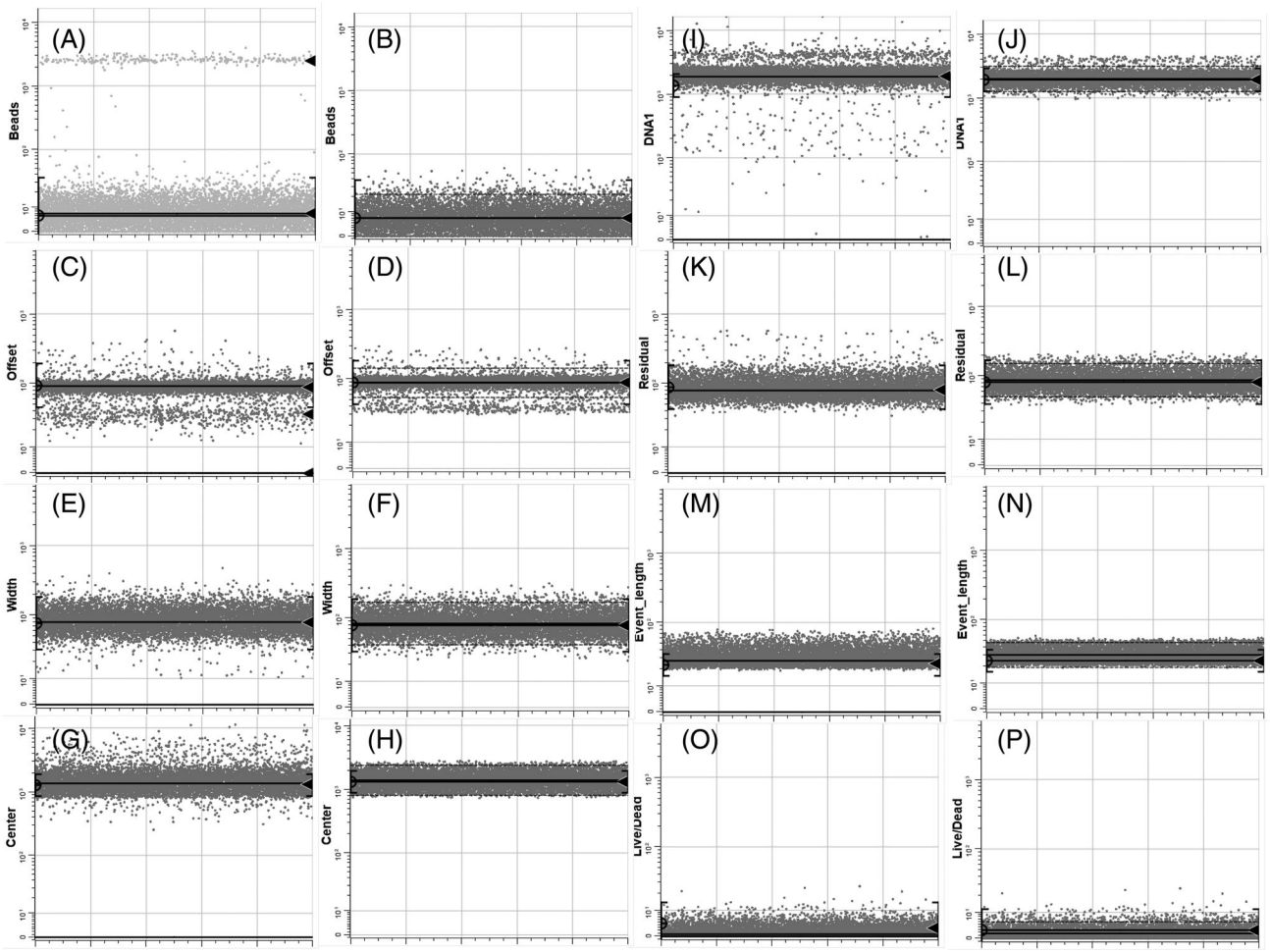

Figure 5. PSM fitting process. The cell type fitting process fits Beads, then Offset, then Width, then Center, then DNA1, then Residual, then Event_length, then Live/Dead, and then DNA2. A and B show before and after Bead (140Ce) fitting. C and D show before and after Offset fitting. E and F show before and after Width fitting. G and H show before and after Center fitting. I and J show before and after DNA1 fitting. K and L show before and after Residual fitting. M and N show before and after Event_length fitting. O and P show before and after Live/Dead fitting. Before and after DNA2 fitting is not shown. Events that that are within the chi-square distribution envelop from all these EP fits are considered for export as Live Intact Events

$$\chi_i^2 = \frac{1}{2}\sum_{j=0}^{1}\left(\frac{x_{i,j}-\mu_j}{\delta_j}\right)^2.$$

Events that were outside the 0.99 high tail of this new chi-square distribution with *two* degrees of freedom were considered unclassified. The percentage of total events in the cleaned cell type reduced to 91.8% (Table 1, Offset column, last row).

An important characteristic of this system is that events can be excluded based on the first EP fitting but can be reselected after the second EP is added, and the reverse is also true. As each EP is considered in the modeling process, the number of degrees of freedom increases by one for the classification chi-square distribution (see Table 1). This probabilistic selection process tends to be more and more refined as EPs are added.

Width Expression Profile

Width is a Gaussian discrimination parameter that is generally low for coincident ion clouds. The template model has 36.5 for its Intensity and Largest Peak for its estimate (Figure 5E). After fitting, the Intensity parameter became 36.9 (Figure 5F) reducing the cleaned percentage of total to 91.7% (see Table 1, Width column).

Center Expression Profile

Center is related to the mean of the Gaussian fit of the signal pulse. If a pulse has two peaks due to coincident ion clouds, where the first is the highest, Center is relatively low, whereas if the second peak in the pulse is the highest, it is relatively high. The template model has its intensity set to 68.2 with a largest peak (LP) Estimate and SD set to 2.5 (Figure 5G). After fitting, the Intensity parameter was set to 68.6 (Figure 5H) and the percentage of total reduced to 85.2% (Table 1).

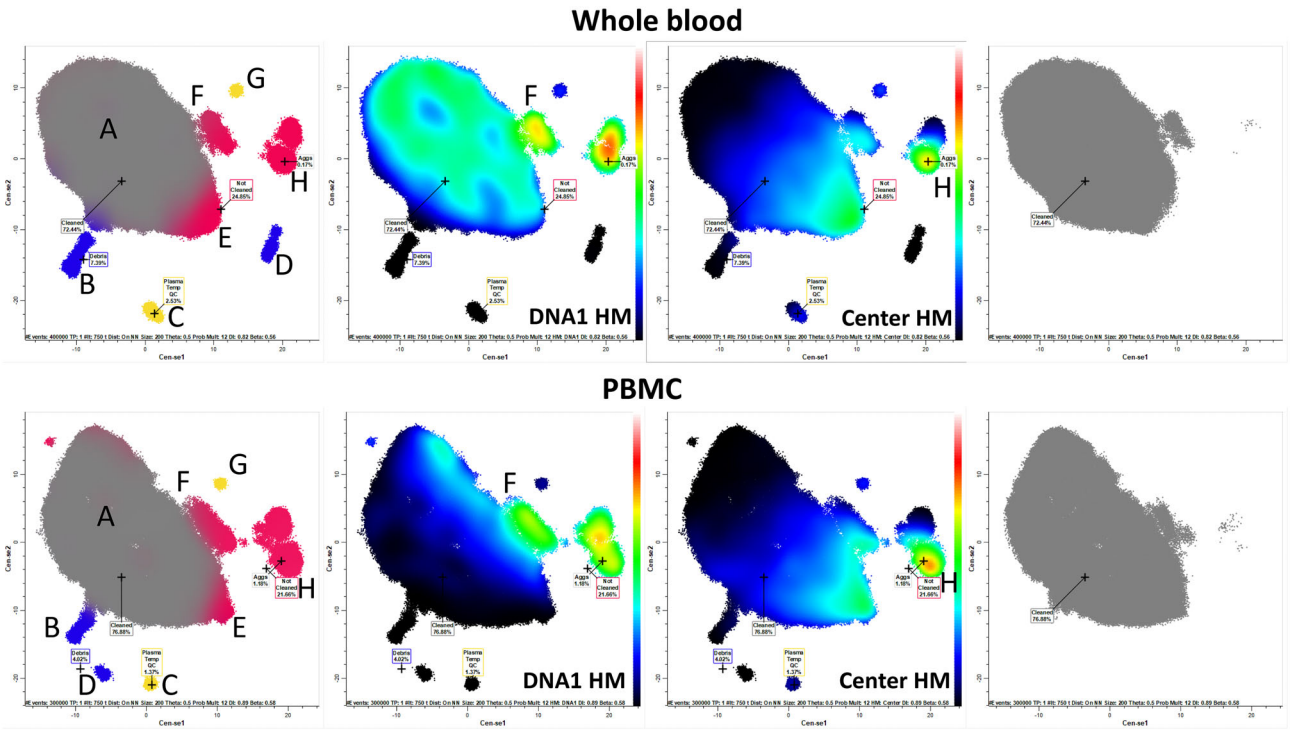

Figure 6. Cleanup Cen-se′ maps: The panels are Cen-se′ maps created from the QC measurements: DNA1, DNA2, Live/Dead, Beads, Event Length, Residual, Center, Width, and Offset on a whole blood (top row) and PBMC (bottom row) sample file. The top-left and bottom-left panels represent raw normalized data from sample files. Section A (dark gray) are the live intact events; B (blue) are the low DNA1 or debris events; C (yellow) are the normalization beads; D (blue) are events with zero valued pulse processing parameters (Residual, Center, Width, and Offset); E (red) are not cleaned events with high Residual and Event_lengths; F (red) are true aggregates with high DNA1 Intensities (see middle left panel DNA1 heat maps); G (yellow) are bead/cell aggregates; and H (red) are coincident ion clouds with low and high Center values (see right middle Center heat maps). Bimodal pulses often produce either low or high Center values depending on which peak is larger. The Center heat map panel shown at the top-right shows that some of the H events have high Center and others low, which is also consistent with this population being coincident ion clouds producing bimodal signal pulses. The right-most panels are maps of only the cleaned events. Even when classifying with both DNA1 and DNA2 some true aggregates can contaminate the cleaned events [Color figure can be viewed at wileyonlinelibrary.com]

DNA1 Expression Profile

DNA1 is the intercalator iridium 191 measurement. The template model sets the Intensity parameter to 69.0, Tallest Peak (TP) and SD to 2.6 (Figure 5I and Table 1). After fitting, the Intensity was adjusted to 73.3 (Figure 5J). The relative loss of doublets and triplets is apparent by comparing Figure 5I,J). The percentage of total reduced to 79.3% (see Table 1, DNA1 column).

In some samples with incomplete red cell lysis, sub-G1 peaks were observed with larger area than the G1 peak. However, the G1 peak was still the highest for these samples. By changing the estimate from Largest Peak to Tallest Peak, the algorithm properly interpreted the G1 peak for this outlier sample. This incremental and evolutionary nature of PSM models to handle worse-case scenarios makes them quite flexible and robust for real laboratory conditions.

Residual Expression Profile

This last Gaussian parameter quantifies the differences between the Gaussian model and the pulse. Pulses that are bell shaped will have relatively low residual values. High

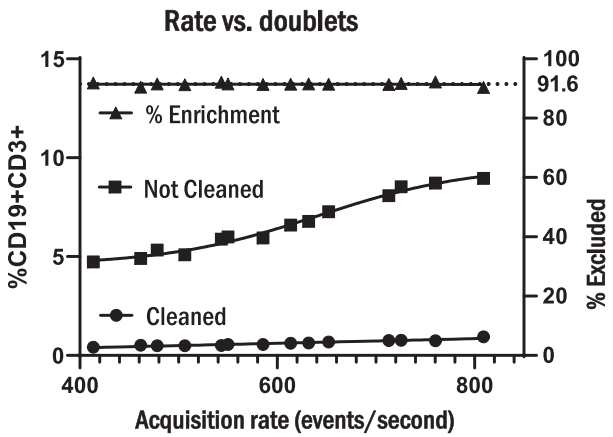

Figure 7. Acquisition rate versus doublets. Whole blood from one donor was diluted to form 14 different samples that had differing acquisition rates (events/second). The percentage of double-positive CD19 + CD3+ to single positive CD19 + CD3– and CD19-CD3+ was recorded for the Cleaned events (circles) and Excluded or Not Cleaned events (squares). The average percent enrichment of doublets excluded (solid triangles) was found to be 91.6%

values for this measurement are indicative of coincident ion clouds and poorly formed pulses. The template model sets the intensity to 38.5 with an estimate of tallest peak (see Figure 5K), and after fitting the intensity was set to 37.3 (see Figure 5L), reducing the percentage of total to 78.5% (Table 1, residual column).

Event_length Expression Profile

Since Event_length is normally a highly skewed distribution that does not have discreet peaks, it is modeled later so it can benefit from all the preceding fitted EPs. The template model sets the intensity to 21.1 with an estimate of Largest Peak (Figure 5M). After fitting, the Intensity was adjusted to 21.7 (Figure 5N), reducing the percentage of total to 73.2% (see Table 1, Event_length column).

Live/Dead Expression Profile

The template model sets the Live/Dead EP Intensity to 5.8 with estimate of Closest Peak and SD set to 4.8 (Figure 5O) and after fitting the Intensity was adjusted to 3.9 (see Figure 5P), increasing slightly the percentage of total to 73.9% (see Table 1, Live/Dead column).

DNA2 Expression Profile

The DNA2 EP has the same settings as the DNA1 EP and is not shown in Figure 5 but is tabulated in Table 1, DNA2 column.

Cen-se′ Maps

Cen-se′™ maps (Cauchy Enhanced Nearest-neighbor Stochastic Embedding) are high-resolution dimensionality-reduction mappings that generate a visual display of high-dimensional data labeled with the major cell populations and their percentages (11). Figure 6 shows Cen-se′ maps created from the QC measurements: DNA1, DNA2, Live/Dead, Beads, Event Length, Residual, Center, Width, and Offset for a whole blood (top row) and a PBMC (bottom row) sample. The left-most panels represent maps derived from raw normalized data. In the left-most panels, A (dark gray) are the live intact events; B (blue) are the low DNA1 or debris events; C (yellow) are the normalization beads; D (blue) are events with zero valued pulse processing parameters (Residual, Center, Width, and Offset); E (red) are Not Cleaned events with high Residual and Event_lengths; F (red) are true aggregates with high DNA1 intensities (see the DNA1 heat maps); G (yellow) are bead/cell aggregates; and H (red) are coincident ion clouds with low and high center values (see the Center heat maps). Bimodal pulses often produce either low or high center values depending on which peak is larger. The center heat map panels reveal that some of the H events have high Center and others low, which is also consistent with this population being considered as coincident bimodal ion clouds. The right-most panels are maps of only the cleaned events.

Data Compaction

Before the data are cleaned, the Cleanup algorithm removes unnecessary data segments from the raw FCS 3.0 files. The

Table 2. Analysis strategy reproducibility. The same data set for Figure 7 was subjected to different gating and modeling analysis strategies. The percentages of live intact cells were evaluated for five different populations (CD8 T-cells, CD4 T-cells, B-cells, NK-cells, and monocytes) in 14 samples. The six analysis strategy statistics are represented as columns in the table. The table shows that when Gaussian parameters (GP) are added to the Gating strategy, the average %CV improved from 3.45 to 2.41. For the Modeling strategies, leveraging both the DNA measurements (DNA+) and Gaussian parameters (GP+) resulted in the lowest %CV of 2.04

ANALYSIS STRATEGY REPRODUCIBILITY[1]													
	GATING						MODELING						
	GP−[2]			GP+			DNA+[3] GP+			DNA− GP+		DNA+ GP−		
POPULATIONS	MEAN	SD	%CV	MEAN	SD	%CV	MEAN	SD	%CV	MEAN	SD	MEAN	SD	%CV
CD8 T-cells	4.97	0.17	3.35	5.19	0.09	1.82	4.78	0.08	1.75	4.44	0.09	4.67	0.10	2.20
CD4 T-cells	8.26	0.23	2.80	8.67	0.18	2.13	8.37	0.14	1.65	7.83	0.15	8.23	0.16	1.94
B-cells	4.59	0.27	5.83	5.08	0.17	3.43	4.84	0.13	2.67	4.86	0.11	4.64	0.15	3.27
NK-cells	4.75	0.12	2.54	4.88	0.10	1.96	4.86	0.10	2.11	4.48	0.09	4.82	0.11	2.23
Monocytes	4.84	0.11	2.72	4.97	0.14	2.71	4.45	0.09	2.01	4.07	0.09	4.49	0.09	1.99
Average			3.45			2.41			2.04					2.33

[1] Results are percentage of live intact cells, $n = 14$.
[2] GP represents population selection using Gaussian parameters.
[3] DNA represents population selection using DNA1 and DNA2 measurements.

Table 3. Reproducibility experiment. This table summarizes the results of a reproducibility experiment that examined three instruments, three technicians, and three replicates generating 27 samples from a single PBMC donor. The right-most column shows the %CVs for each of the 37 population percentages. The overall mean and median %CVs were 14.0 and 10.2, respectively

EXPERIMENTS	REPRODUCIBILITY OF PBMC CELLULAR POPULATIONS DERIVED FROM A SINGLE SAMPLE[1]																												
INSTRUMENTS	INST 1									INST 2									INST 3										
TECHNICIANS	TECH 1			TECH 2			TECH 3			TECH 1			TECH 2			TECH 3			TECH 1			TECH 2			TECH 3					
REPLICATES	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3			
COLUMNS	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	AV	SD	%CV
Lymphocytes	61.56	61.77	62.10	65.99	64.59	66.49	62.25	62.79	62.34	60.59	60.46	61.19	66.43	65.99	64.43	60.25	59.98	61.01	60.48	61.41	63.82	69.12	68.02	66.71	61.81	62.56	63.52	63.25	2.59	4.09
CD3 T-cells	38.21	39.45	39.16	41.44	39.28	39.91	38.88	39.05	38.43	38.02	38.31	38.65	42.03	41.55	40.48	37.63	37.55	38.60	37.79	38.77	40.13	41.92	41.35	41.58	38.16	38.63	39.37	39.42	1.40	3.56
CD8 T-cells	12.72	12.82	12.83	13.77	13.68	14.16	12.63	12.99	12.91	12.46	12.32	12.44	13.66	13.87	13.65	12.31	12.12	12.64	12.17	12.67	13.09	15.14	14.89	14.00	12.62	12.69	13.22	13.13	0.80	6.08
CD8 Naïve	0.84	0.93	0.87	1.01	0.92	0.91	1.16	1.22	1.04	1.06	0.94	0.94	1.05	1.01	1.03	1.05	1.06	1.23	1.03	0.93	1.23	1.10	1.04	0.98	1.09	0.95	1.08	1.03	0.10	10.16
CD8 CM	1.10	1.16	1.20	1.20	1.10	1.03	0.80	0.91	0.99	1.00	1.16	1.16	1.16	1.14	1.16	0.95	1.05	0.85	0.89	1.18	1.07	1.18	1.21	1.13	0.99	1.07	1.13	1.07	0.11	10.69
CD8 EM	5.61	5.74	5.72	6.07	6.08	6.35	5.65	5.67	5.79	5.53	5.33	5.46	6.06	6.33	6.13	5.49	5.26	5.55	5.49	5.64	5.74	6.73	6.56	6.30	5.67	5.76	5.85	5.84	0.38	6.44
CD8 TE	5.17	4.99	5.05	5.50	5.57	5.86	5.03	5.18	5.09	4.87	4.88	4.88	5.39	5.39	5.32	4.81	4.74	5.00	4.77	4.92	5.06	6.13	6.08	5.60	4.87	4.92	5.16	5.19	0.38	7.39
CD4 T-cells	22.74	23.83	23.55	24.59	22.56	22.55	23.51	23.31	22.84	22.76	23.23	23.36	25.22	24.63	23.75	22.58	22.69	23.32	22.97	23.52	24.06	23.36	22.96	24.41	22.89	23.25	23.41	23.40	0.69	2.97
CD4 Naïve	5.49	5.98	5.89	6.54	6.33	5.74	7.04	7.06	6.87	6.15	5.65	6.26	6.58	6.70	6.77	6.69	7.03	7.11	3.18	6.53	6.89	6.53	6.76	6.86	7.24	7.70	7.44	6.48	0.85	13.17
CD4 CM	6.66	8.13	7.78	7.56	6.37	7.09	5.97	5.77	5.99	6.77	8.21	7.22	8.31	7.33	7.07	6.45	5.65	6.25	10.22	6.88	7.15	7.13	6.69	7.36	5.975.75	6.01	6.95	1.01	14.54	
CD4 EM	7.48	7.02	7.14	7.66	7.15	6.86	7.63	7.58	7.01	6.89	6.62	6.95	7.42	7.79	7.30	6.74	7.31	7.18	6.67	7.30	7.36	6.94	6.84	7.41	7.16	7.30	7.44	7.19	0.31	4.36
CD4 TE	3.11	2.69	2.74	2.82	2.71	2.87	2.87	2.91	2.98	2.96	2.75	2.94	2.92	2.82	2.62	2.69	2.71	2.79	2.90	2.80	2.66	2.76	2.67	2.79	2.52	2.50	2.52	2.78	0.15	5.34
γδ T-cells	1.90	1.93	1.90	2.14	2.13	2.28	1.92	1.96	1.91	1.91	1.83	1.94	2.12	2.07	2.11	1.83	1.86	1.82	1.89	1.92	2.04	2.39	2.42	2.24	1.85	1.90	1.94	2.01	0.17	8.50
MAIT/NKT Cells	0.86	0.87	0.88	0.93	0.92	0.92	0.82	0.79	0.76	0.89	0.93	0.91	1.03	0.98	0.97	0.91	0.88	0.81	0.76	0.67	0.94	1.02	1.08	0.92	0.80	0.79	0.80	0.88	0.09	10.42
B-cells	12.00	11.33	11.89	11.29	11.38	10.36	12.30	12.53	12.58	11.68	11.49	11.66	11.39	11.37	10.98	12.04	12.04	11.99	11.73	11.74	11.73	11.37	10.61	10.44	12.27	12.48	12.43	11.67	0.61	5.20
B Naïve	9.31	8.72	9.16	8.81	9.07	8.32	9.35	9.57	9.57	8.99	8.90	9.00	9.01	9.00	8.64	9.16	9.11	9.08	9.05	9.05	9.11	9.33	8.76	8.38	9.39	9.66	9.49	9.07	0.33	3.69
B Mem	2.63	2.54	2.64	2.40	2.25	1.97	2.84	2.86	2.92	2.61	2.50	2.56	2.28	2.28	2.25	2.81	2.86	2.81	2.60	2.60	2.53	1.98	1.79	2.00	2.79	2.73	2.86	2.52	0.32	12.63
Plasmablasts	0.07	0.07	0.09	0.07	0.05	0.07	0.10	0.10	0.09	0.08	0.09	0.09	0.09	0.09	0.09	0.07	0.08	0.10	0.09	0.09	0.09	0.07	0.06	0.07	0.09	0.09	0.09	0.08	0.01	16.55
NK-cells	11.35	10.99	11.05	13.27	13.92	16.23	11.07	11.22	11.33	10.90	10.66	10.88	13.02	13.07	12.97	10.58	10.39	10.41	10.96	10.89	11.96	15.83	16.06	14.69	11.38	11.45	11.72	12.16	1.78	14.68
NK Early	5.35	5.21	5.15	6.36	6.52	7.67	4.70	4.78	4.78	5.19	5.03	5.18	6.34	6.30	6.20	4.50	4.45	4.47	5.18	5.26	5.65	7.53	7.56	7.09	4.88	4.94	5.00	5.60	0.99	17.66
NK Late	5.99	5.77	5.90	6.91	7.41	8.56	6.37	6.44	6.55	5.71	5.63	5.70	6.68	6.77	6.78	6.07	5.94	5.94	5.77	5.63	6.31	8.30	8.50	7.60	6.50	6.51	6.72	6.55	0.86	13.11
Monocytes	22.05	21.41	21.90	18.15	19.73	16.77	22.18	22.02	22.69	21.77	21.51	21.69	18.07	18.87	18.57	21.91	21.87	21.77	21.88	21.12	21.73	15.57	16.48	16.47	22.50	21.93	21.52	20.45	2.19	10.71
Mono Class	19.00	18.53	18.86	15.37	17.07	13.95	19.11	19.05	19.62	18.52	18.22	18.27	14.72	15.55	15.66	18.54	18.45	18.34	18.71	17.99	18.66	13.04	13.80	13.37	19.05	18.20	18.29	17.33	2.04	11.77
Mono Int	2.04	1.88	2.05	1.76	1.84	1.82	2.11	2.03	2.15	2.21	2.18	2.25	2.11	2.17	1.79	2.29	2.33	2.28	2.09	2.04	2.08	1.61	1.74	1.92	2.37	2.55	2.22	2.07	0.22	10.60
Mono NC	1.01	1.00	0.99	1.02	0.82	0.99	0.96	0.95	0.92	1.04	1.11	1.17	1.24	1.16	1.12	1.08	1.08	1.15	1.08	1.10	1.00	0.92	0.93	1.18	1.08	1.19	1.01	1.05	0.10	9.46
DCs	1.25	1.22	1.11	1.20	1.27	1.34	1.18	1.18	1.28	1.19	1.13	1.17	1.14	1.19	1.18	1.17	1.14	1.16	1.22	1.20	1.20	1.29	1.31	1.20	1.21	1.17	1.12	1.20	0.06	4.85
pDCs	0.45	0.46	0.42	0.45	0.50	0.51	0.48	0.48	0.49	0.44	0.45	0.46	0.42	0.47	0.47	0.48	0.46	0.47	0.47	0.46	0.48	0.52	0.50	0.50	0.47	0.47	0.43	0.47	0.03	5.46
mDCs	0.80	0.76	0.69	0.75	0.77	0.83	0.71	0.71	0.79	0.75	0.69	0.71	0.72	0.72	0.71	0.69	0.68	0.69	0.75	0.74	0.71	0.77	0.80	0.70	0.74	0.70	0.69	0.73	0.04	5.70
Granulocytes	1.59	2.67	1.88	1.94	1.47	3.15	1.39	2.09	1.72	2.24	2.61	2.12	1.91	1.48	3.22	2.01	2.50	2.35	1.63	1.97	1.33	1.50	1.17	3.30	1.43	1.48	1.38	1.98	0.60	30.38
Neutrophils	0.22	0.56	0.63	0.18	0.20	0.48	0.17	0.15	0.46	0.61	0.56	0.21	0.50	0.40	0.43	0.24	0.74	0.60	0.22	0.24	0.15	0.17	0.25	0.16	0.24	0.20	0.22	0.34	0.18	54.02
Basophils	0.06	0.24	0.08	0.18	0.17	0.05	0.07	0.05	0.17	0.20	0.24	0.24	0.05	0.13	0.06	0.22	0.06	0.21	0.30	0.25	0.31	0.15	0.18	0.06	0.20	0.23	0.23	0.16	0.08	50.85
Eosinophils	0.10	0.06	0.11	0.05	0.05	0.04	0.05	#	0.05	0.05	0.07	0.07	0.03	0.01	0.01	0.04	0.04	0.05	0.06	0.08	0.07	0.04	0.22	0.05	0.06	0.05	0.09	0.06	0.04	65.11

(Continues)

Table 3. Continued

REPRODUCIBILITY OF PBMC CELLULAR POPULATIONS DERIVED FROM A SINGLE SAMPLE[1]																															
EXPERIMENTS																															
INSTRUMENTS	INST 1									INST 2										INST 3										
	TECH 1			TECH 2			TECH 3			TECH 1			TECH 2			TECH 3			TECH 1			TECH 2			TECH 3					
TECHNICIANS	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3				
REPLICATES	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	AV	SD	%CV	
COLUMNS	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27				
CD66b- Neuts	**1.21**	**1.81**	**1.05**	**1.54**	**1.06**	**2.58**	**1.10**	**1.85**	**1.05**	**1.39**	**1.74**	**1.59**	**1.33**	**0.94**	**2.71**	**1.52**	**1.65**	**1.50**	**1.05**	**1.40**	**0.79**	**1.14**	**0.52**	**3.04**	**0.94**	**0.99**	**0.84**	**1.42**	**0.59**	**41.74**	
Tregs	**0.74**	**0.81**	**0.79**	**0.78**	**0.78**	**0.74**	**0.81**	**0.74**	**0.70**	**0.74**	**0.80**	**0.82**	**0.84**	**0.83**	**0.75**	**0.74**	**0.75**	**0.74**	**0.77**	**0.80**	**0.84**	**0.82**	**0.85**	**0.81**	**0.83**	**0.80**	**0.82**	**0.79**	**0.04**	**#**	
Th1-like	**0.92**	**1.00**	**0.98**	**0.95**	**0.91**	**1.01**	**0.77**	**0.78**	**0.79**	**0.96**	**0.95**	**0.98**	**1.05**	**1.07**	**0.95**	**0.85**	**0.76**	**0.83**	**1.04**	**0.93**	**1.02**	**0.99**	**0.98**	**0.94**	**0.70**	**0.76**	**0.84**	**0.92**	**0.10**	**11.25**	
Th2-like	**4.59**	**4.92**	**4.77**	**4.90**	**4.51**	**4.62**	**4.73**	**4.65**	**4.45**	**4.31**	**4.45**	**4.41**	**4.86**	**4.80**	**4.47**	**4.50**	**4.73**	**4.46**	**4.92**	**4.61**	**4.76**	**4.78**	**4.54**	**4.81**	**4.54**	**4.41**	**4.58**	**4.63**	**0.18**	**3.80**	
Th17-like	**2.02**	**2.13**	**2.08**	**2.23**	**2.08**	**2.03**	**2.20**	**2.09**	**2.06**	**2.09**	**2.15**	**2.16**	**2.33**	**2.22**	**2.19**	**1.98**	**1.96**	**2.07**	**2.19**	**2.13**	**2.19**	**2.23**	**2.23**	**2.40**	**2.21**	**2.18**	**2.10**	**2.14**	**0.10**	**4.69**	
																												Mean:		**13.97**	
																												Median:		**10.16**	

[1] Percentages of live intact events.

results of this data compaction procedure are files that are approximately 33% the original size. A specific example best demonstrates how file sizes change with data processing. The original size of a file after acquisition was 321.8 MB. After the compaction stage, it reduces to 105.9 MB and after data cleanup becomes 62.0 MB. This compaction process eases many storage issues associated with long-term archival of mass cytometry files.

RESULTS

Acquisition Rate Experiment

Whole blood from one donor was diluted to form 14 different samples that had differing acquisition rates (events/second). Figure 7 shows the percentage of double-positive CD19 + CD3+ to single-positive CD19 + CD3− and CD19-CD3+ for the cleaned events (solid circles) and not cleaned events (solid squares). The average enrichment of these doublets in the excluded cell type (solid triangles) was found to be 91.6%.

The same data set for Figure 7 was subjected to different gating and modeling analysis strategies. Table 2 shows percentages of live intact cells were evaluated for five different populations (CD8 T-cells, CD4 T-cells, B-cells, NK-cells, and Monocytes) in 14 samples. The six analysis strategy statistics are represented as columns in the table. The table shows that when Gaussian parameters (GP) are added to the gating strategy, the average %CV improved from 3.45 to 2.41. For the modeling strategies, leveraging both the DNA measurements (DNA+) and Gaussian parameters (GP+) resulted in the lowest %CV of 2.04.

Reproducibility Experiment

A desirable attribute of an automated cleanup method is that it generates reproducible results. One way of assessing reproducibility is to examine live-intact percentages of canonical populations derived from the same PBMC donor for data produced by several cytometers and preparations. Table 3 summarizes the results of a reproducibility experiment that examined three instruments, three technicians, and three replicates generating 27 samples. The results were automatically generated from Maxpar Pathsetter™ software analysis using the herein described PSM Cleanup model and a deep immune phenotyping MIP model (Bagwell et al, in prep). The rightmost column shows the %CVs for each of the 37 populations. The sample data for three populations—CD4 T-cells, CD8 T-cells, and B-cells—are shown in Figure 8. The %CVs of CD4 T-cells, CD8 T-cells, and B-cells were found to be 3.0%, 6.1%, and 5.2%, respectively.

Live/Dead Experiment

In order to better assess the capability of the cleanup model to detect and eliminate dead cells, fixed and permeabilized PBMC preparations were added to whole blood at three different amounts (100,000, 500,000, and 1 million). Figure 9 shows the control population with no fixed cells in the upper-left panel. The increasing doses of fixed cells are in the other three panels. The red distribution are events that were

eliminated, and the dark gray events were exported as live and intact cells. The percentages of excluded events for each sample were 22.75% (A), 37.56% (B), 40.07% (C), and 43.11% (D).

Custom Panel Experiment

To test the generality of the cleanup method with a custom panel, eight PBMC files derived from a single control sample were cleaned by the described method and cellular populations were evaluated for reproducibility (see Table 4).

DISCUSSION

The initial insight on how best to leverage the Gaussian parameters was suggested by one of the authors, Vladimir Baranov, in 2017. He wondered what the Cen-se′ maps would look like if they were created with just the QC measurements: DNA1, DNA2, Live/Dead, Event_length, Center, Width, Offset, and Residual (Figure 6). Once it was realized that the map showed distinct clusters of events for wanted and unwanted events, it was relatively easy to encircle each one with animated color event regions and examine and model the expression profile (EP) patterns for each QC measurement.

The original order of the EPs was determined by their ability to divide events into desired and undesired categories. The general model development idea implicit in PSM is to begin with the EPs that best classify events for a cell type and end with those that have more complicated or subtle distributions such as Event_length. Although DNA1 and DNA2 are choices that a model builder might be tempted to put first in a series of cell type EPs, they were defined later since a user may elect to inactivate or modify them if they need to detect DNA aneuploid oncological samples.

After thousands of data files were subjected to the PSM cleanup analyses, successfully analyzing outlier samples necessitated slight changes to both the order of the EPs and how they interacted with the peak detection system. The result of this incremental refinement of the model design is a surprisingly robust method for selecting live intact events for further analyses.

Even with both DNA1 and DNA2 classifications, there were still some true aggregates present after cleanup (see Figure 6 right-most panels). During Cleanup model design, the SDs of the DNA1 and DNA2 EPs had to be narrow enough to eliminate most of the true aggregates but not too narrow that they selectively eliminated the slightly higher staining monocytes.

By leveraging biologically exclusive CD19 and CD3 cellular markers, the number of aggregates involving T-cells and B-cells can be enumerated. The percentages shown in Figure 7 are the double-positive events times 100 divided by the presumed number of singlet CD19 + CD3- and CD19-CD3+ events. It is assumed that the number of events

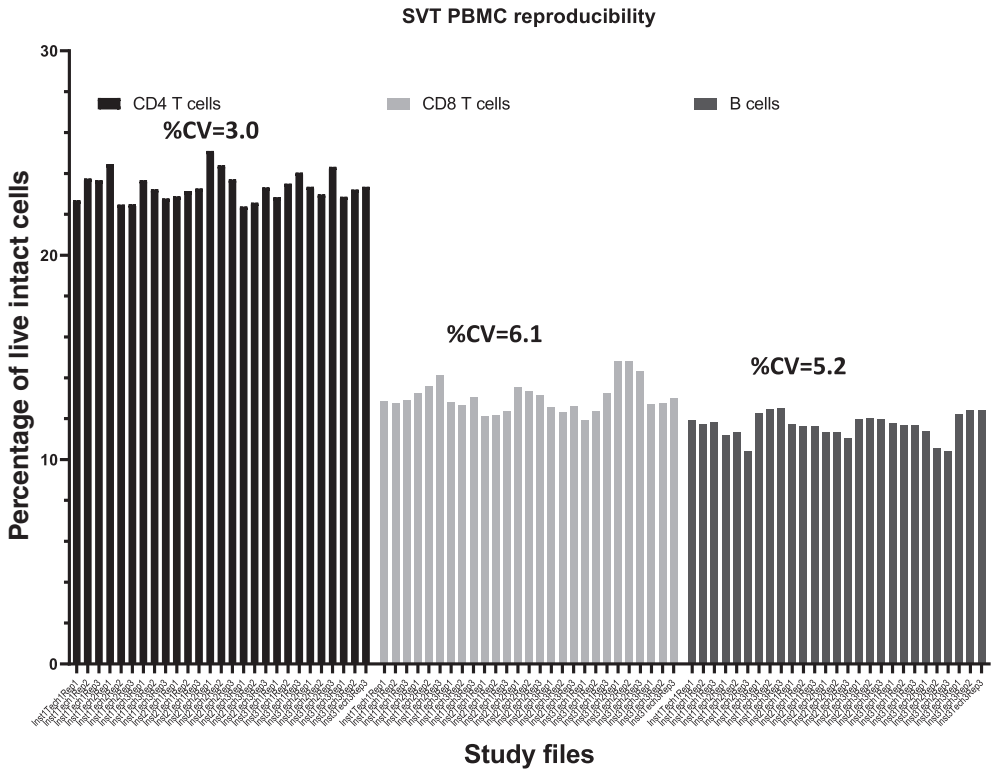

Figure 8. Population reproducibility. The same PBMC sample was run on three different instruments and stained by three different technicians with three different replicates. The populations CD4 T-cells, CD8 T-cells, and B-cells were selected to evaluate their reproducibility

in this double-positive population is primarily the summation of the initial number of aggregates in the sample plus the number of aggregate signals formed by coincident ion clouds. It is also assumed that these specific aggregates can be viewed as a kind of surrogate for aggregation in general.

Figure 7 shows that as acquisition rate was increased by increasing the concentration of cells in the sample, the percentage of detected CD19 + CD3+ double-positive events increased in both the cleaned (circles) and not cleaned (squares) cell types. The %ratio yields the relative enrichment of these double-positive events in the not cleaned cell type, which averaged 91.6% for this set of samples.

This same data set for was also subjected to different gating and modeling analysis strategies in order to show that leveraging Gaussian parameters for either gating or modeling can result in better reproducibility. Table 2 also suggests that it is possible to use the Gaussian parameters instead of DNA for analyzing samples that may have an aberrant DNA content.

Table 3 and Figure 8 show reasonably good reproducibility for most cellular population percentages. Some of the higher %CVs are from populations that have lower frequencies, which tend to be affected more by counting error. The granulocyte populations also tend to have poor reproducibility for PBMC samples since they are essentially contaminants of the sample preparation. The average %CV for all non-granulocyte population percentages was low, 8.6%. The data shown in Table 4 provide evidence that the described cleanup methodology is generalizable to other custom panels.

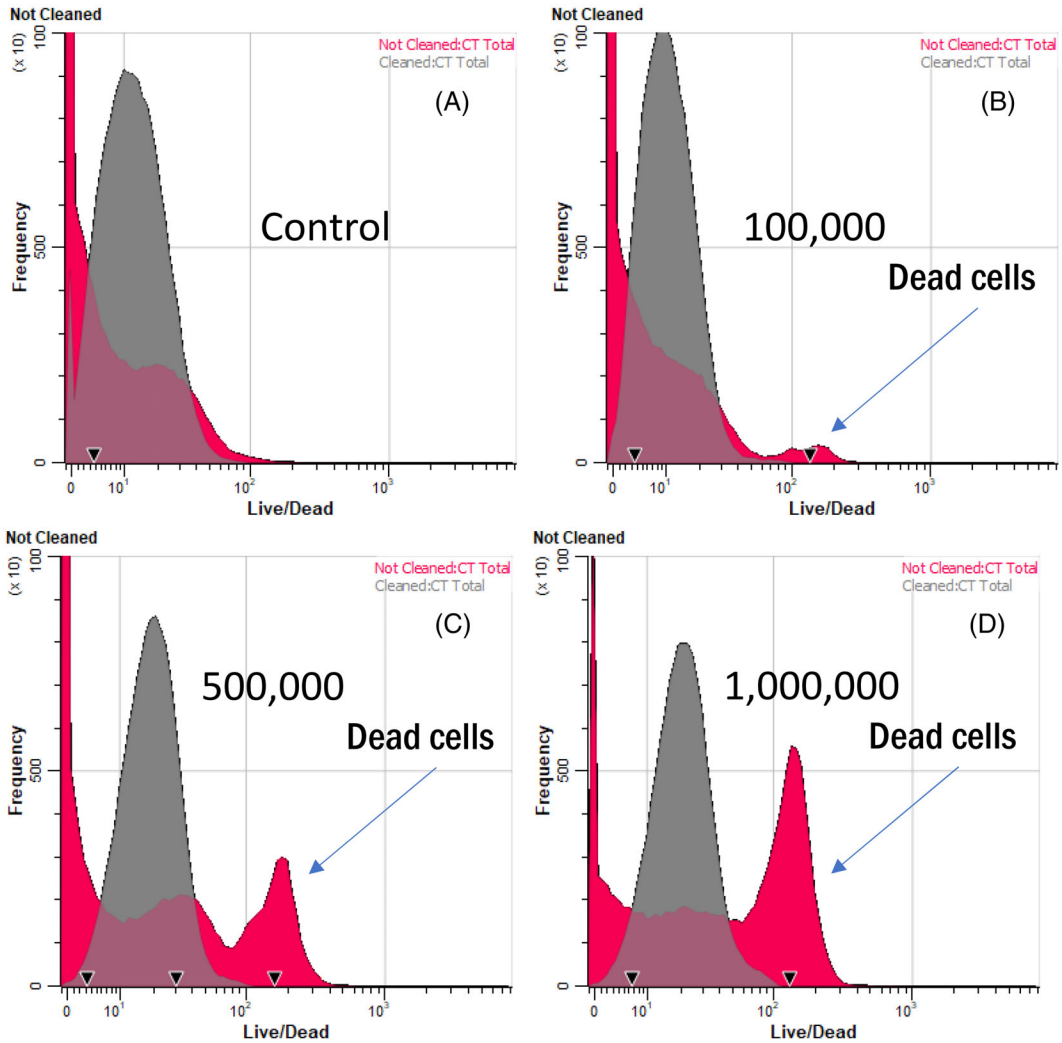

Figure 9. Live/dead discrimination. Fixed and permeabilized PBMC preparations were added to whole blood at three different amounts (100,000, 500,000, and 1 million) to test the ability of the Cleanup model to discriminate live from dead cell events. The control population with no fixed cells is shown in the upper-left panel. The increasing doses of fixed cells are in the other three panels. The red distribution are events that were eliminated, and the dark gray are events that were exported as live and intact cells. The percentages of excluded events for each sample were 22.75% (A), 37.56% (B), 40.07% (C), and 43.11% (D) [Color figure can be viewed at wileyonlinelibrary.com]

Table 4. Reproducibility experiment from custom panel. This table summarizes the results of a reproducibility experiment that examined two instruments and four replicates generating eight samples from a single PBMC peripheral blood leukapheresis pack. The right-most column shows the %CVs for each of the 28 population percentages. The overall reproducibility of all populations has mean and median % CVs of 12.8 and 10.6, respectively

	UNC REPRODUCIBILITY STUDY: POPULATION PERCENTAGES[1]										
REPLICATES	1	2	3	4	5	6	7	8	MEAN	SD	%CV
Lymphocytes	47.84	50.34	51.17	52.96	48.28	49.74	51.01	56.58	50.99	2.79	5.47
CD3 T-cells	41.25	43.82	43.06	45.11	41.43	43.62	42.19	47.35	43.48	2.03	4.67
CD8 T-cells	13.78	14.16	14.58	15.18	14.15	14.37	14.25	16.21	14.59	0.77	5.29
CD8 Naïve	9.87	10.21	9.58	9.65	10.18	10.35	9.20	10.93	10.00	0.54	5.37
CD8 CM	1.17	1.23	1.44	1.81	1.24	1.11	1.55	1.60	1.39	0.24	17.54
CD8 EM	1.08	1.26	2.23	2.18	1.16	1.50	2.11	2.22	1.72	0.51	29.91
CD8 TE	1.66	1.46	1.33	1.54	1.57	1.42	1.39	1.46	1.48	0.11	7.17
CD4 T-cells	26.13	28.16	27.07	28.69	26.18	27.90	26.61	29.71	27.56	1.28	4.64
CD4 Naïve	8.72	9.33	6.45	6.75	8.95	9.04	6.54	8.12	7.99	1.22	15.27
CD4 CM	5.55	5.22	6.85	7.11	5.48	5.07	7.31	6.82	6.18	0.93	15.03
CD4 EM	6.67	8.21	9.32	9.73	6.84	8.50	8.12	10.05	8.43	1.25	14.81
CD4 TE	5.20	5.39	4.45	5.10	4.91	5.28	4.65	4.71	4.96	0.34	6.76
γδ T-cells	1.33	1.50	1.41	1.24	1.10	1.34	1.33	1.44	1.34	0.12	9.34
B-cells	3.81	4.01	3.91	4.02	4.08	3.40	3.65	4.32	3.90	0.28	7.17
B Naïve	3.62	3.79	3.67	3.81	3.77	3.22	3.42	4.02	3.67	0.25	6.76
B Mem	0.19	0.22	0.25	0.21	0.31	0.18	0.23	0.30	0.23	0.05	19.78
NK cells	2.78	2.51	4.21	3.84	2.77	2.71	5.16	4.91	3.61	1.06	29.43
Monocytes	38.44	35.62	33.53	29.95	38.71	36.58	33.54	27.54	34.24	3.95	11.55
Mono Class	25.90	24.92	26.75	22.41	25.85	24.20	26.53	21.38	24.74	1.96	7.92
Mono Int	7.63	5.97	3.24	3.87	8.20	7.51	3.62	3.30	5.42	2.14	39.58
Mono NC	4.91	4.73	3.53	3.67	4.66	4.88	3.39	2.86	4.08	0.80	19.67
DCs	1.07	1.01	0.86	0.82	0.94	0.96	0.89	0.82	0.92	0.09	9.90
pDCs	0.31	0.37	0.31	0.29	0.29	0.37	0.30	0.27	0.31	0.04	11.37
mDCs	0.76	0.65	0.54	0.54	0.65	0.59	0.59	0.55	0.61	0.08	12.39
Tregs	1.01	1.30	1.54	1.21	1.16	1.27	1.43	1.40	1.29	0.17	12.81
Th1-like	3.13	3.05	2.94	3.55	3.06	3.33	3.24	3.34	3.20	0.20	6.25
Th2-like	4.22	4.77	4.99	4.66	4.56	4.48	4.71	5.18	4.70	0.30	6.31
Th17-like	0.54	0.75	0.80	0.73	0.50	0.73	0.62	0.77	0.68	0.11	16.23
										Mean	12.80
										Median	10.64

[1] Population percentages are of live intact cells.

Most of the validation testing for the cleanup model used cellular preparations with relatively few dead cells. In order to assess how well this cleanup routine eliminated events from dead cells, fixed and permeabilized PBMC were employed to mimic their leaky membrane behavior. Figure 9 shows that the Live/Dead EP was effective in removing these events from the exported live intact events.

This study demonstrates that probability state modeling is effective in automating the removal of unwanted events by means of a probabilistic chi-square-based selection process. This nonsubjective process generated cleaned FCS 3.0 data files that, when appropriately analyzed, demonstrated highly reproducible estimates for population frequencies. This study also shows how Gaussian-derived parameters that are normally stored as additional measurements for mass cytometry can be leveraged to help eliminate unwanted partial as well as coincident cloud waveforms.

ACKNOWLEDGMENTS

The authors wish to acknowledge the helpful editing from Dmitry Bandura, Tony Shuga, and Dorothy McDuffie. This work could not have been done without the generous resources of both Verity Software and Fluidigm. We would also like to acknowledge Marie Iannone, Director at UNC Mass Cytometry Core and Kevin Katen at NEIHS Mass Cytometry Facility that helped us generate UNC reproducibility dataset. UNC Mass Cytometry Core has been supported by University Cancer Research Fund (UCRF) and UNC Cancer Center Core Support Grant #P30CA016086. Lab work to generate this data set was supported by NIH RO1 Grant # AI119164.

CONFLICTS OF INTEREST

All the authors except Avinash Kollapara are currently employed or were employed by either Verity Software

House or Fludigm Corporation. This article describes a component of the product, Fluidigm Pathsetter™, which was a collaborative effort between these two companies. Fluidigm, Cell-ID, CyTOF, EQ, Helios, Maxpar and Pathsetter are trademarks and/or registered trademarks of Fluidigm Corporation in the United States and/or other countries. All other trademarks are the sole property of their respective owners. For Research Use Only. Not for use in diagnostic procedures.

LITERATURE CITED

1. Ornatsky OI, Lou X, Nitz M, Sheldrick WS, Baranov VI, Bandura DR, Tanner SD. Study of cell antigens and intracellular DNA by identification of element-containing labels and metallointercalators using inductively coupled plasma mass spectrometry. Anal Chem 2008;80:2539–2547.
2. Ornatsky O, Bandura D, Baranov V, Nitz M, Winnik MA, Tanner S. Highly multiparametric analysis by mass cytometry. J Immunol Methods 2010;361:1–20.
3. Lou X, Zhang G, Herrera I, Kinach R, Ornatsky O, Baranov V, Nitz M, Winnik MA. Polymer-based elemental tags for sensitive bioassays. Angew Chem Int Ed. (in English) 2007;46:6111–6114.
4. Bandura DR, Baranov VI, Ornatsky OI, Antonov A, Kinach R, Lou X, Pavlov S, Vorobiev S, Dick JE, Tanner SD. Mass cytometry: Technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. Anal Chem 2009;81:6813–6822.
5. Tanner SD, Baranov VI, Ornatsky OI, Bandura DR, George TC. An introduction to mass cytometry: Fundamentals and applications. Cancer Immunol Immunother 2013;62:955–965.
6. Olsen LR, Leipold MD, Pedersen CB, Maecker HT. The anatomy of single cell mass cytometry data. Cytometry 2018;95A(2):156–172.
7. Marie JI, Ellegaard J, Hokland P. Differences in relative DNA content between human peripheral blood and bone marrow subpopulations–consequences for DNA index calculations. Cytometry 1993;14:936–942.
8. Fienberg HG, Simonds EF, Fantl WJ, Nolan GP, Bodenmiller B. A platinum-based covalent viability reagent for single-cell mass cytometry. Cytometry 2012;81A:467–475.
9. Bevington PR. Data reduction and error analysis for the physical sciences. New York, NY: McGraw-Hill Book Company, 1969.
10. Bagwell CB, Hunsberger BC, Herbert DJ, Munson ME, Hill BL, Bray CM, Preffer FI. Probability state modeling theory. Cytometry 2015;87A:646–660.
11. Bagwell CB, Bray C, Herbert D, Hill BL, Inokuma MS, Stelzer G, Hunsberger B. Improving the t-SNE algorithms for cytometry and other technologies: Cen-se' mapping. J Biom Biostat 2019;10:1–13.