Cytometry PART A
Journal of the International Society for Advancement of Cytometry

Sequential Univariate Gating Approach to Study the Effects of Erythropoietin in Murine Bone Marrow

Ram Achuthanandam,[1,2]* John Quinn,[3,4] Renold J. Capocasale,[2] Peter J. Bugelski,[2] Leonid Hrebien,[1] Moshe Kam[1]

[1]Electrical and Computer Engineering Department, Drexel University, Philadelphia, Pennsylvania

[2]Toxicology and Investigational Pharmacology, Centocor R&D Inc, Radnor, Pennsylvania

[3]Biomedical Engineering Department, Drexel University, Philadelphia, Pennsylvania

[4]Treestar Inc., Seattle, Washington

Received 6 August 2007; Revision Received 31 January 2008; Accepted 21 April 2008

*Correspondence to: Ram Achuthanandam, Research Scientist, Centocor R&D, 145 King of Prussia Road, Radnor, PA 19406, USA

Email: Rachutha@CNTUS.JNJ.com

Published online 21 May 2008 in Wiley InterScience (www.interscience.wiley.com)

DOI: 10.1002/cyto.a.20584

© 2008 International Society for Advancement of Cytometry

ISAC
International Society for Advancement of Cytometry

• Abstract

Analysis of multicolor flow cytometric data is traditionally based on the judgment of an expert, generally time consuming, sometimes incomplete and often subjective in nature. In this article, we investigate another statistical method using a Sequential Univariate Gating (SUG) algorithm to identify regions of interest between two groups of multivariate flow cytometric data. The metric used to differentiate between the groups of univariate distributions in SUG is the Kolmogorov-Smirnov distance (D) statistic. The performance of the algorithm is evaluated by applying it to a known three-color data set looking at activation of CD4+ and CD8+ lymphocytes with anti-CD3 antibody treatment and comparing the results to the expert analysis. The algorithm is then applied to a four-color data set used to study the effects of recombinant human erythropoietin (rHuEPO) on several murine bone marrow populations. SUG was used to identify regions of interest in the data and results compared to expert analysis and the current state-of-the-art statistical method, Frequency Difference Gating (FDG). Cluster analysis was then performed to identify subpopulations responding differently to rHuEPO. Expert analysis, SUG and FDG identified regions in the data that showed activation of CD4+ and CD8+ lymphocytes with anti-CD3 treatment. In the rHuEPO treated data sets, the expert and SUG identified a dose responsive expansion of only the erythroid precursor population. In contrast, FDG resulted in identification of regions of interest both in the erythroid precursors as well as in other bone marrow populations. Clustering within the regions of interest defined by SUG resulted in identification of four subpopulations of erythroid precursors that are morphologically distinct and show a differential response to rHuEPO treatment. Greatest expansion is seen in the basophilic and poly/orthochromic erythroblast populations with treatment. Identification of populations of interest can be performed using SUG in less subjective, time efficient, biologically interpretable manner that corroborates with the expert analysis. The results suggest that basophilic erythroblasts cells or their immediate precursors are an important target for the effects of rHuEPO in murine bone marrow. The MATLAB implementation of the method described in the article, both experimental data and other supplemental materials are freely available at http://web.mac.com/acidrap18. © 2008 International Society for Advancement of Cytometry

• Key terms

flow cytometry; multivariate data analysis; Kolmogorov-Smirnov statistic; FDG; erythropoietin; cluster analysis

Erythropoeisis is the process by which new red blood cells are formed in the bone marrow and erythropoietin (EPO) is the primary driver of this process (1–3). EPO acts by stimulating a cell surface receptor known as erythropoietin receptor (EPO-R) (1,2). Advances in recombinant technology have helped in the development of human recombinant EPO (rHuEPO) (4,5). The earliest morphologically identifiable erythroid progenitor cell is the proerythroblast (ProEB). ProEBs usually divide and make two daughter cells known as Basophilic Erythroblasts (BasoEB). BasoEBs divide and differentiate into polychromatophilic erythroblasts (PEB).

Orthochromatophilic erythroblasts (OEB) are formed from PEBs and have increased levels of hemoglobin. OEBs lose their nuclei and polyribosomes and emigrate from the marrow to circulate in blood as reticulocytes. Reticulocytes circulate in blood for about a day or two after which they mature into erythrocytes (6).

Recent developments in monoclonal antibody technology have facilitated the analysis of the different stages of late stage erythroid precursors (3,7–9) using flow cytometry. TER-119 is a monoclonal antibody that is associated with surface glycophorin A and is used as a marker for enumerating cells of erythroid lineage from ProEB to mature erythrocyte (9). CD71, the transferrin receptor (7), exists as a homodimer on the cell surface and is essential for cellular growth. Immature proliferating cells express CD71 and can thus serve as a marker of differentiation. Anti-TER-119 and anti-CD71 enable the enumeration of the various stages of erythroid precursor development using flow cytometry (3). Sca-1 (Ly-6A/E) is a marker expressed on the multipotent hematopoietic stem cell in mice bone marrow (10). CD11b (Mac-1) is expressed at varying levels on granulocytes, macrophages, myeloid-derived dendritic cells, natural killer cells, microglia, and B-1 cells (11). Gr-1 antigen expression is directly correlated with granulocyte differentiation and maturation and also expressed on the monocyte lineage in the bone marrow (12). IL-7Rα is expressed on common lymphoid progenitors and early stages of B lineage development in the bone marrow (13). The effect of rHuEPO in mouse bone marrow on erythroid, myeloid, common lymphoid progenitors, and hematopoietic stem cell populations was studied with the help of the aforementioned markers using multicolor flow cytometry.

Traditionally, multicolor flow cytometric data analysis by the expert involves examination of fluorescent intensity profiles, two colors at a time, to determine regions of interest in the data. This process is followed by comparison of regions of interest between data from groups of biological specimens arising from different physiological, toxicological, or pathological conditions. Expert analysis uses visual graphical methods to choose regions of interest by the expert's topical knowledge and experience. Though, the general location of the region of interest is based on markers used in the experiment, the exact region can vary depending on the expert. Concerns about consistency, time to perform analysis, and questions about reproducibility have led to the development of several statistical approaches to analyze flow cytometric data. Overton (14) suggested to quantify the channel-by-channel differences in event counts between a treated and control sample and using them to determine the percentage of population that responded to a treatment (percent positive). Cox et al. (15) used the chi-square test to perform a channel-by-channel comparison to detect variables that were significantly different between two samples. Lampirello (16,17) developed a model for background autofluorescence and assumed that the background distribution due to negative events does not change. The model thus estimated the "percent positive" events using a probability density function. Bagwell and Overton (14,18,19) have applied variations of the Kolmogorov-Smirnov (KS) test to identify differences between distributions. Finally, Roederer et al.'s (20–22) probability binning (PB) method uses a normalized chi-square statistic to identify differences between two samples. An equal-density binning algorithm is used to generate regions in the data from the control sample containing equal number of events per region. The same regions are then applied to the data from the treated sample to obtain the number of events in each region. A normalized chi-square statistic is then obtained to determine which regions are different in the treated sample. The process of obtaining regions of difference in such a manner is termed Frequency Difference Gating (FDG) (22).

Biological interpretability of the regions of interest identified with current statistical methods is a major limitation. We have tried to address this shortcoming by the use of a preprocessing step before performing the analysis and the use of a Sequential Univariate Gating algorithm (SUG). SUG employs the distance metric (*D*-value) calculated in a Kolmogorov-Smirnov two sample test to distinguish between univariate distributions (23). SUG begins by identifying if the distribution of events in any of the colors is different between the two groups. If the distribution is different for a color, regions where this difference exists is estimated. Once these regions have been identified in a single color, in the next step only, those events are considered. The next step involves determining the similarity of distribution of these events with respect to the other colors, which is tested again using SUG. Using such a nested univariate method to identify multivariate regions of interest, we retain the biological interpretation of the regions (the regions of interest are orthogonal). SUG is compared to the expert analysis and FDG on two data sets; one from murine spleen treated with anti-CD3 antibodies and the second from murine bone marrow treated with rHuEPO. We demonstrate that even using such a simple univariate statistical method we can perform a multivariate analysis. A preprocessing step is demonstrated to eliminate certain events that will enable better interpretability of the results. We also demonstrate the utility of postprocessing using a clustering algorithm to gain additional information.

Materials and Methods

Spleen from Agonistic Anti-CD3 Treatment

Ten to twelve-week-old mice were separated into two groups (three/group) with one group receiving phosphate buffered saline (PBS) and the other receiving a single dose of an agonistic anti-CD3 antibody. This antibody is known to induce activation of CD4+ and CD8+ cells measured by CD25+ levels (dose response experiment not shown). Mice were euthanized 24-h posttreatment, spleens excised, and single cell suspensions of fresh isolated splenocytes prepared. Cells in suspension were treated with ammonium chloride to lyze red blood cells and then resuspended in BD Staining buffer (BD Biosciences, San Jose, CA). Cells were labeled with an anti-CD4, anti-CD8, and anti-CD25 antibody conjugated with fluorescein (FITC), phycoerythrin (PE), and allophycocyanin (APC), respectively. Cells were read within 1 h and

Table 1. Animal and antibody panel description

RHuEPO DOSE LEVEL	# SAMPLES	PANEL 1	PANEL 2
PBS (0 IU/kg)	8	Anti-TER-119 PE, Anti-CD71 SA–APC	Anti-Gr-1 FITC, Anti-Sca-1 PE, Anti-IL-7Rα PE-Cy5, Anti-C11b APC
30 IU/kg	4	Anti-TER-119 PE, Anti-CD71 SA–APC	Anti-Gr-1 FITC, Anti-Sca-1 PE, Anti-IL-7Rα PE-Cy5, Anti-C11b APC
100 IU/kg	4	Anti-TER-119 PE, Anti-CD71 SA–APC	Anti-Gr-1 FITC, Anti-Sca-1 PE, Anti-IL-7Rα PE-Cy5, Anti-C11b APC
300 IU/kg	4	Anti-TER-119 PE, Anti-CD71 SA–APC	Anti-Gr-1 FITC, Anti-Sca-1 PE, Anti-IL-7Rα PE-Cy5, Anti-C11b APC
1,000 IU/kg	4	Anti-TER-119 PE, Anti-CD71 SA–APC	Anti-Gr-1 FITC, Anti-Sca-1 PE, Anti-IL-7Rα PE-Cy5, Anti-C11b APC
3,000 IU/kg	4	Anti-TER-119 PE, Anti-CD71 SA–APC	Anti-Gr-1 FITC, Anti-Sca-1 PE, Anti-IL-7Rα PE-Cy5, Anti-C11b APC
10,000 IU/kg	4	Anti-TER-119 PE, Anti-CD71 SA–APC	Anti-Gr-1 FITC, Anti-Sca-1 PE, Anti-IL-7Rα PE-Cy5, Anti-C11b APC

analysis performed on a Becton Dickinson 2—laser FACSCalibur© flow cytometer. All fluorescence data were acquired on 4-decade log scales (1024 channels). At least 10,000 events/sample were collected with debris and aggregates excluded using low forward and orthogonal light scatter values. The fluorescent intensity data collected from each animal was saved as "FCS 2.0" files (flow cytometry standard files).

Bone Marrow from rHuEPO Treatment

Ten to twelve-week-old C57BL/6 mice were used in our study. Six groups of mice (four/group) each received a single subcutaneous dose of rHuEPO (30, 100, 300, 1,000, 3,000, or 10,000 IU/kg) in PBS and the control group (eight mice) received only PBS. Mice were euthanized with CO_2 48 h post dosing and bone marrow flushed from the femurs and tibias and cell suspensions prepared. Cells in suspension were treated with ammonium chloride to lyze red blood cells. Cells were resuspended in BD Staining buffer. Cells were aliquoted into 96-well polystyrene round bottom tissue-culture plates at a concentration of 5×10^5 cells per well and preincubated with anti-murine CD16/CD32 (FcgRIII/II) 2.4G2 to reduce F_C receptor-mediated antibody binding. Following F_C block, cells were incubated for 20 min with panels of rat anti-murine monoclonal antibodies as shown in Table 1. Aliquots of cells were also incubated with appropriately matched-isotype control antibodies. Cells were read fresh within 1 h and analysis performed on a Becton Dickinson 2—laser FACSCalibur© flow cytometer. All fluorescence data were acquired on 4-decade log scales (1024 channels). At least 30,000 events/sample were collected with debris and aggregates excluded using low forward and orthogonal light scatter values. The fluorescent intensity data collected from each animal was saved as "FCS 2.0" files.

Reagents and Monoclonal Antibodies

Phosphate buffered saline (PBS) without Ca^{++} and Mg^{++} and fetal calf serum (FCS) were purchased from Invitrogen (Carlsbad, CA). Sodium azide (NaN_3) was purchased from Sigma (St Louis, MO). rHuEPO was obtained from Ortho Biologics (Raritan, NJ). Doses are expressed as IU/kg (the activity of rHuEPO was 120 IU/μg.). The following monoclonal antibodies (mAbs) against mouse cell surface markers were purchased from BD-Pharmingen (BD Biosciences, San Jose, CA): PE-conjugated anti-TER-119, biotinylated anti-CD71 (clone C2, APC-conjugated streptavidin), PE conjugated anti-Ly6A/E (Sca-1, clone E13-161.7), FITC conjugated anti-Ly-6G (Gr-1, clone RB6-8C5), biotin-conjugated anti-IL-7 receptor α (CD127, clone B12-1, PE-Cy5-conjugated streptavidin), APC-conjugated anti-CD11b (Mac-1, clone M1/70); FITC conjugated anti-CD4, PE conjugated anti-CD8, and APC conjugated anti-CD25. Appropriately labeled isotype-matched IgG controls were also used.

Data Analysis: Expert Method

For the spleen data set, cells were enumerated as either CD4+ or CD8+ as shown in Figure 1. The percentage of events that were CD25+ in each population was deemed as activated cells. The fraction of events that were CD4+\CD25+ and CD8+\CD25+ was determined. These fractions enumerate the fraction of CD4+ and CD8+ cells that are activated in the spleen with the anti-CD3 antibody.

For the bone marrow data set, fluorescent intensity data from each treatment group (PBS or rHuEPO treatment) were viewed as two-dimensional scatter plots. ProEB (R1), BasoEB (R2), and POEB (R3) were delineated using the two-dimensional plot of TER-119 vs. CD71 (Fig. 2A). The R4 gate as shown in Figure 2B identified sca-1+ hematopoietic stem cells. Events positive for IL-7Rα (R5) were identified as common lymphoid progenitors. Monocytoid (R6) and granulocytoid (R7) populations were delineated using a two-dimensional plot of Gr-1 vs. CD11b. The gating strategy for the expert analysis is shown in Figure 2. Event counts were found for each population and normalized to the PBS treated values and represented as a percentage of the control. The graphs in

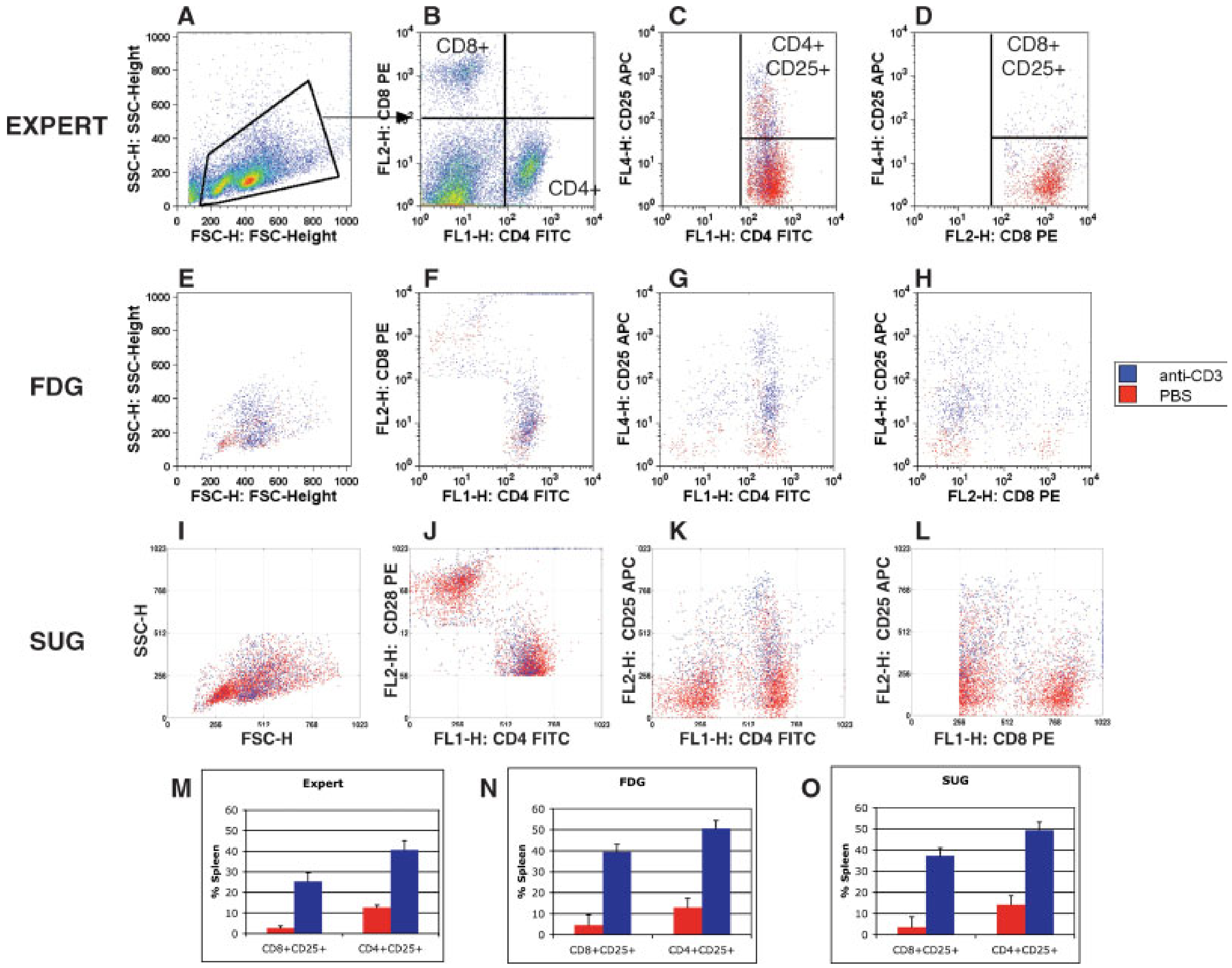

Figure 1. Figure 1 shows a comparison of analysis results of data obtained from spleen treated with an anti-CD3 antibody. Figure **A–D** shows the gating strategy to identify CD4+ and CD8+ cells and identifies activation of these cells by CD25 expression (Figure 1m). FDG (**E–H**) and SUG (**I–L**) analysis resulted in the identification of the same regions (gates) as set by the expert. Additionally, the change in activation levels based on CD25+ events were also very similar between the three analysis methods (**M–O**).

Figure 2E–2M indicate mean changes with respect to control with error bars indicating standard deviations. Statistical significance was determined by performing an ANOVA followed by a Tukey multiple hypothesis correction (23).

Data Preprocessing

Preprocessing involved removing (gating out) events negative for all the immunostains in a given data set. Each single color immunostain was inspected visually by the expert and the threshold below which an event was negative for that immunostain was ascertained. By identifying the threshold for all the immunostains, we can identify a region, which is negative for all the immunostains. Such a preprocessing step eliminates data that cannot be biologically interpreted if a region of interest is found in that area using a statistical analysis be it FDG or SUG. Preprocessing was used on both the data sets before employing either FDG or SUG.

Data Analysis: Frequency Difference Gating (FDG)

FDG analysis was performed on the data using its implementation in Flowjo 6.4.1 under the OS X (Apple) operating system ("Multi-sample Compare Platform"). The eight PBS controls were compared and the normalized chi-squared values calculated for all the samples with 1,000 bins per sample (standard in Flowjo). Regions corresponding to bins having a chi-square value greater than the largest chi-squared value from the controls were established as the region of interest for all samples (please refer to (22) for further details).

Exported Data Description

Data files obtained from the FACS Calibur containing fluorescent intensity and scatter values were converted to tab delimited text files using the freeware "WinMDI" (Joe Trotter, The Scripps Institute, Flow Cytometry Core Facility). The forward scatter (FSC), orthogonal scatter (SSC), and the various

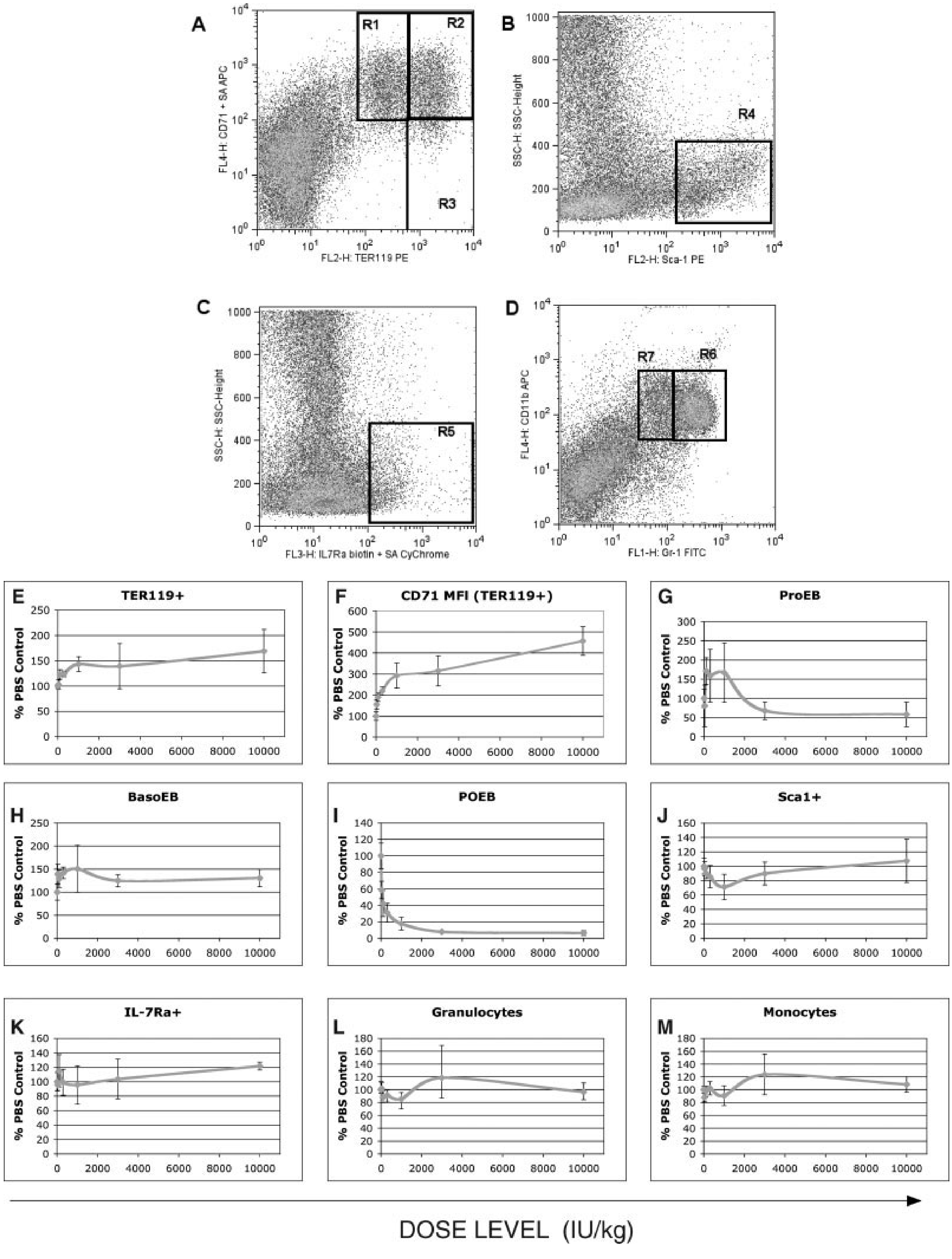

Figure 2. Depicts the gating strategy to delineate populations of interest from bone marrow data by the expert is shown in **A**–**D**. R1, R2, and R3 are the gates to identify TER-119+ events. R1 represents the proerythroblasts, R2 the basophilic erythroblasts, and R3 the poly/ortho chromatophillic erythroblasts. R4 identifies Sca-1+ stem cells, R5 the IL-7R population, R6 the monocytoid, and R7 the granulocytoid populations. **E**–**M**: The changes in the populations normalized to the PBS control values. In E, we see a dose responsive increase in the erythroid progenitors (TER-119+). F: The upregulation of the CD71 parameter as s function of dosage level. G–I depict changes in the different erythroid precursors and J–M depict changes in the non-erythroid populations.

immunostains are termed variables. Data obtained from one animal is termed a sample. Given a set of n samples, $\{X_i\}_{i=1}^n$, each sample, X_i, is a $u \times v$ matrix, where u is the number of events recorded, v the number of variables in the experiment and each element of the matrix, $x_{i,j,k}$, the fluorescent intensity value ($x_{i,j,k} \in [0, 1{,}023]$). We can organize all the k-th column values of a sample into a $1 \times u$ column vector, $P_{i,k} = [x_{i,1,k} \quad x_{i,2,k} \quad \ldots \quad x_{i,u,k}]^T$ comprising data points of the k-th variable. Each sample can be represented as:

$$X_i = \begin{bmatrix} x_{i,1,1} & x_{i,1,2} & \cdots & x_{i,1,v} \\ x_{i,2,1} & x_{i,2,2} & \cdots & x_{i,2,v} \\ \vdots & \vdots & \ddots & \vdots \\ x_{i,u,1} & x_{i,u,2} & \cdots & x_{i,u,v} \end{bmatrix} = [P_{i,1} \quad P_{i,2} \quad \ldots \quad P_{i,v}]. \tag{1}$$

We assign a class label to each sample, denoted by y_i, where $y_i \in \{+1 \ -1\}$. The set of all class labels can be organized into a $1 \times n$ class vector, $Y = [y_1 \ y_2 \ \ldots \ y_n]$, where n^+ is the number of samples from the positive class ($y_i = +1$) and n^- is the number of samples from the negative class ($y_i = -1$) and $n = n^+ + n^-$.

Data Analysis: Sequential Univariate Gating (SUG)

The Kolmogorov-Smirnov two-sample (KS) (23) test is a statistical test used to determine if the distribution of a variable from two samples arises from the same underlying unknown distribution. The KS test is a nonparametric test that is capable of identifying differences in distributions (location or shape). Given $P_{a,k}$ and $P_{b,k}$, the D-value ($D_{a,b}$) in a KS two-sample test is calculated as

$$D_{k,a,b} = \max |F(P_{a,k}) - F(P_{b,k})|, \tag{2}$$

where $F\,(P_{i,k})$ is the cumulative distribution function of $P_{i,k}$. In the presence of multiple samples from each class, we can determine this critical D-value ($\hat{D}_k$) empirically for each parameter. The $\hat{D}_k$ is used as the minimum D-value that determines if a variable is different in a sample from the negative class compared to a sample from the positive class. To calculate $\hat{D}_k$ for the k-th variable, all the D-values are obtained by comparing the distribution of the k-th variable between samples from the negative class with each other and samples from the positive class with each other using the KS test. $\hat{D}_k$ is then calculated as the maximum D-value obtained from these comparisons. Let $D^-_{k,a,b}$ be the D-values obtained by comparing the k-th variable between samples from the negative class with each other and $D^+_{k,p,q}$ the values obtained by comparing samples from the positive class with each other. Mathematically,

$$\hat{D}_k = \max\left(D^-_{k,a,b}, D^+_{k,p,q}\right) y_a, y_b \in -1 \quad y_p, y_q \in +1 \tag{3}$$

Next, the D-values obtained by comparing the k-th variable distribution between samples from the negative class ($y = -1$) to samples from the positive class ($y = +1$) are recorded. The k-th variable is said to have a region of interest (ROI) if more than 90% of the D-values recorded are greater than the critical D-value ($\hat{D}_k$) for that variable.

$$ROI^k = \left\{ \begin{matrix} exists| & D_{k,a,p} > \widehat{D}_k \\ none| & D_{k,a,p} < \widehat{D}_k \end{matrix} \right\} \tag{4}$$
$$y_a = -1; y_p = +1$$

A ROI is defined as the region in the multivariable fluorescent intensity space that is different between samples from two classes. The identification of a ROI is performed in a nested fashion using SUG. First, we identify the variables that have different distributions in the two classes. We then identify the location (intensity/channel numbers) in each variable of this difference by employing a window of W consecutive fluorescent intensity levels and applying the same distance metric to each of the W windows. This window slides over the distribution, moving it $W/2$ intensity levels at a time from intensity level zero to the maximum intensity. By sliding the window and calculating differences in distribution each time, we can identify a range of intensity values where the distribution is different between the two classes. The process of identification of an ROI with one variable is shown graphically in Figure 3A.

To obtain a multivariate ROI, we extract the events obtained by performing the univariate analysis and then compare the distributions of the events in the other v-1 variables (conditional marginal distribution of the other v-1 variables). This analysis is performed for each variable containing a univariate ROI. The nested analysis can be depicted in a tree structure and at each level the presence of an ROI is determined. The final multivariate ROI is the union of all the ROIs ascertained at the last level in the analysis tree. The identification of the multiparameter ROI using this nested approach is depicted graphically in Figure 3B.

Results

Analysis of Anti-CD3 Treated Spleen Data

Data analysis by the expert was performed as shown in Figure 1. Expert analysis resulted in identification of increased levels of activation of both the CD4+ and CD8+ populations. FDG was performed with 1,000 bins and the threshold level, the chi-square value selected as 0.055. The results of FDG indicate a great number of events that are CD25+ in both CD4+ and CD8+ populations corroborating the expert analysis. For SUG, $v = 5$ (FSC, SSC, CD4, CD8, and CD25) and $u = 10{,}000$ events. The window $W = 256$ channels for determining where the distributions were different (ROI as a function of W is shown in the supplemental material). Samples treated with PBS are from the negative class ($y = -1$) and samples treated with anti-CD3 antibody from the positive class ($y = +1$). ROIs were determined by comparing the samples from the PBS treated group to the anti-CD3 treated group. SUG, also resulted in identifying both CD4+and CD8+ events that had higher levels of CD25+ events. These results show that FDG and SUG corroborate the expert analysis. The anti-CD3 agonistic antibody used in the experiment is known to activate CD4+ and CD8+ T-cells, thus showing that FDG and SUG resulted in identification of differences between two groups in

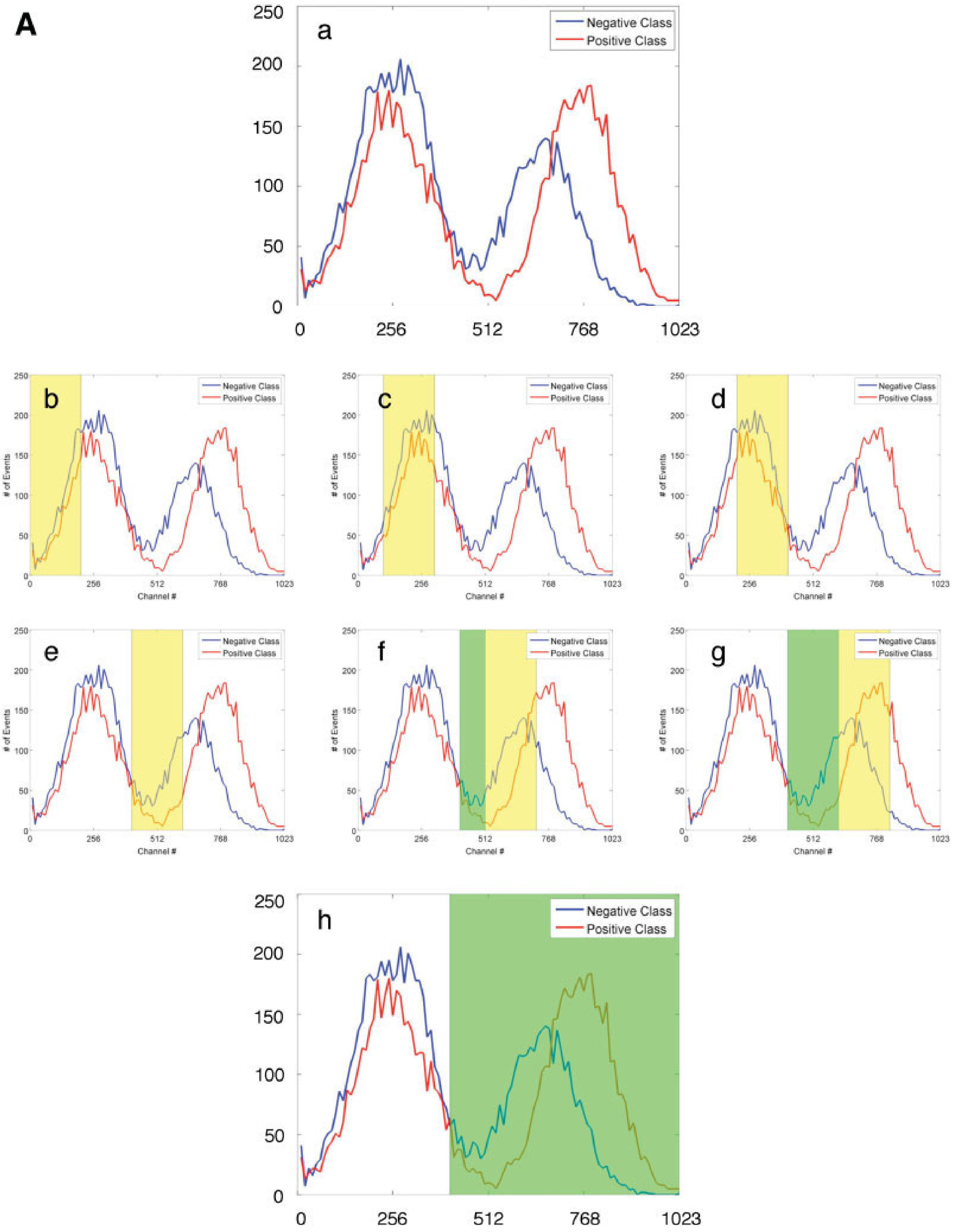

Figure 3. **A**: The histogram of a variable from a representative sample from the positive and negative classes (the variable is identified to be different in the two classes). (b)–(g) show the process of obtaining the location of the ROI by recursively using SUG in a window (shaded yellow) to look for differences in the distribution. Windows that are identified as having a different distribution are labeled as ROIs (shaded green). (h) The final ROI and thus identifies the location of difference between the events in the two classes. **B**: The process to obtain a two-dimensional ROI. (a) and (b) are the representative negative class sample and positive class sample. (c) and (d) represent the univariate ROI obtained using SUG in variable 1. (e) and (f) depict the histogram of the events from the ROI determined at the previous stage. (e) shows the histograms of events in the shaded region in (d) and (f) the histogram of shaded region in (c). SUG is used again to estimate an ROI exists and the final ROI in two dimensions is shown in (g).

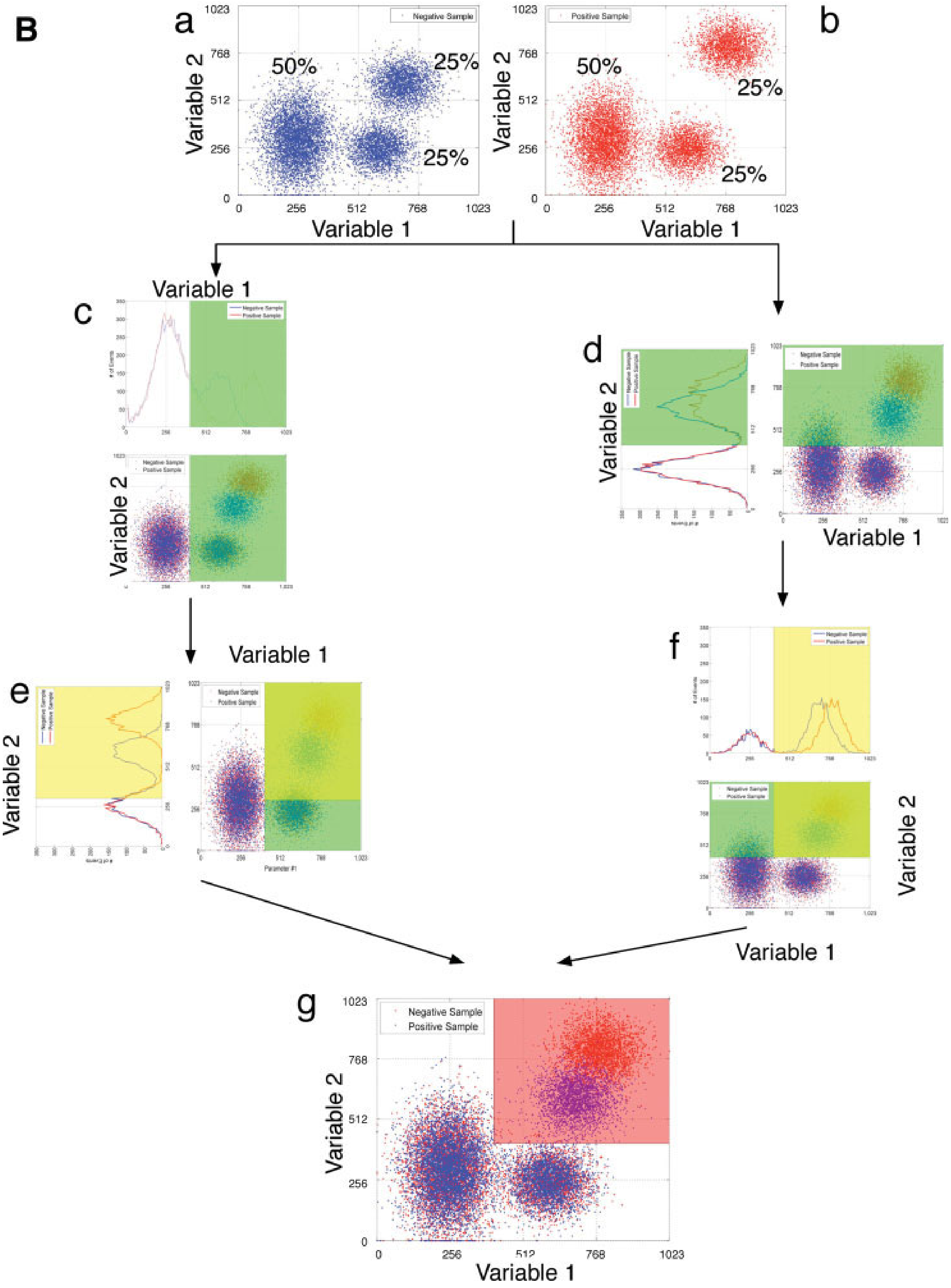

Figure 3. Continued.

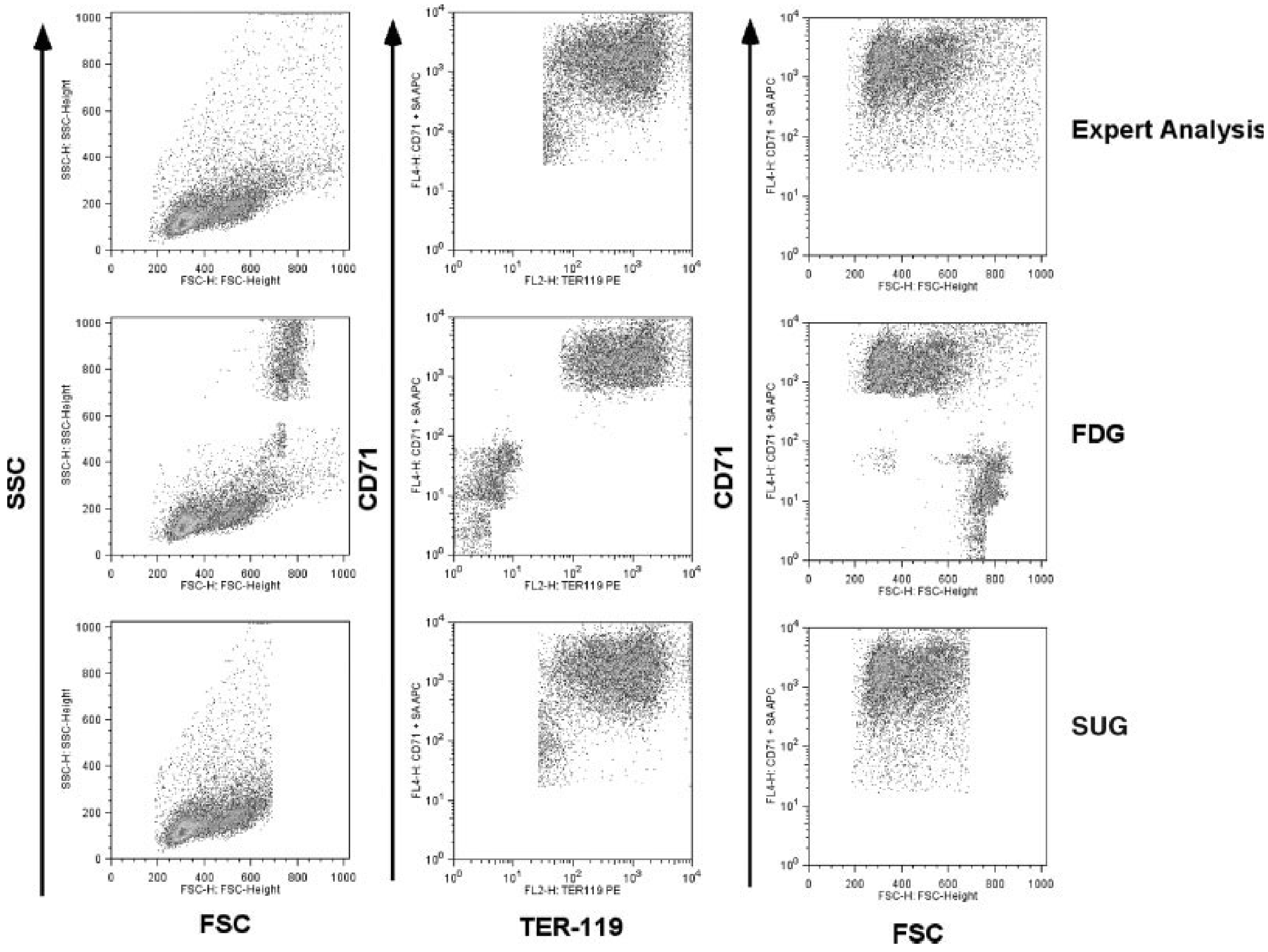

Figure 4. Depicts three two-dimensional projections of regions of interest obtained from the three different methods while comparing data from PBS controls to rHuEPO treatment at 10,000 U/kg.

the biologically relevant populations (CD4+CD25+ and CD8+CD25+).

Analysis of rHuEPO Treated Bone Marrow Data

Expert analysis was performed as shown in Figure 2. Mean Fluorescent Intensity (MFI) of the variables was found for each of the populations at each of six dosage levels and compared to controls. The analysis shows a dose responsive increase in erythroid precursors (TER-119+), an expansion of the BasoEB populations, and a decrease in the POEB population. The Sca-1+, IL-7Ra+, granulocytes and monocytes were not affected by rHuEPO treatment. Expert analysis resulted in identification only of an increase in the erythroid precursor populations (BasoEB) with rHuEPO treatment. Additionally, an increase in CD71 (a marker for transferrin receptor) was found in a dose responsive manner.

FDG identified ROIs in the rHuEPO treated samples compared to the controls from both panels. FDG also identified regions in the erythoid precursor populations (as expected). However, FDG spilt the erythroid precursor populations in a segmented or abrupt fashion (Fig. 4). FDG selected segmented regions in each of the erythroid precursors populations including ProEB, BasoEB, and P\OEB populations. FDG identified differences in data from Panel 2, but these differences were not dose responsive in nature. Thus, FDG was found to be very sensitive in identifying ROIs both within the erythroid precursor and other populations.

In SUG, for data from Panel 1 we have $v = 6$ parameters (FSC, SSC, Gr-1, CD11b, IL7-Rα and Sca-1) and $u = 30{,}000$ events. Data from Panel 2 has $v = 4$ parameters (FSC, SSC, TER-119 and CD71) and $u = 30{,}000$ events. The window $W = 256$ channels for determining where the distributions were different (ROI as a function of W is shown in the supplemental material). Samples treated with PBS are from the negative class ($y = -1$) and samples treated with a dose of rHuEPO are from the positive class ($y = +1$). ROIs are determined by comparing the samples from the PBS treated group to one of the dose levels of the rHuEPO treated group (e.g., PBS group vs. 300 IU/kg rHuEPO group).

Using SUG, ROIs were found comparing PBS treated sample to rHuEPO treated samples stained with antibodies from Panel 2 (anti-TER-119 and anti-CD71). ROIs were found

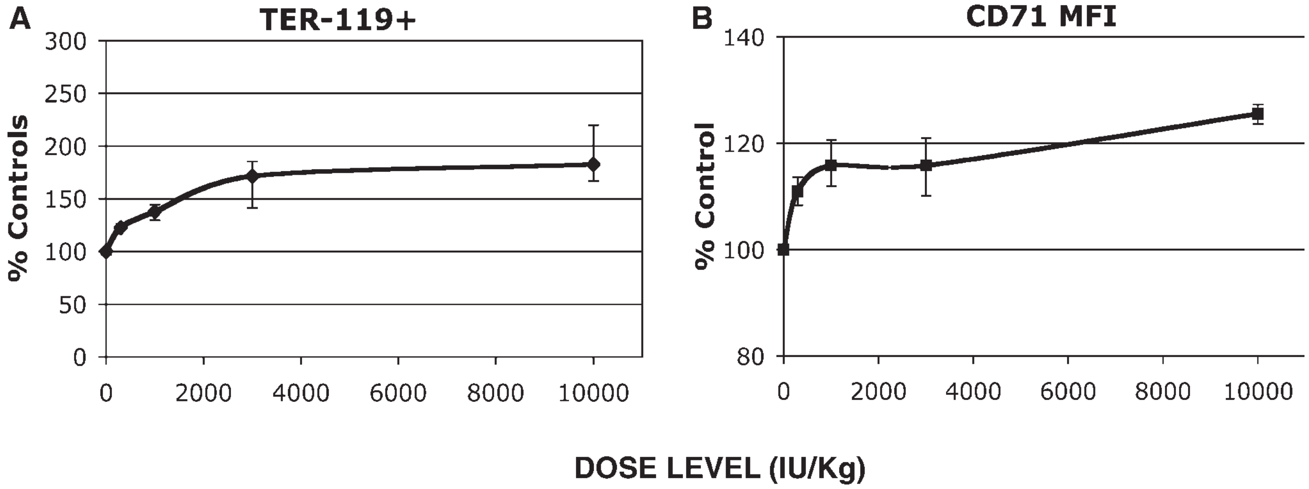

Figure 5. **A**: A dose responsive increase in erythroid precursors within the ROI with 30U/kg being the no effect dose of rHuEPO. **B**: The upregulation of CD71 on erythroid precursors.

when PBS treated samples were compared to rHuEPO treated samples only at dosage levels 300, 1,000, 3,000, and 10,000 IU/kg. At each dosage level the ROI was found at the same location and is shown in the two-dimensional plot of TER-119 vs. CD71 parameters in Figure 4. Our method identified a region, which is different between the PBS treated controls and the rHuEPO treated animals that contain late stage erythroid precursors (TER-119+). The comparison of the expert analysis, FDG and SUG is illustrated in Figure 4 for data from samples treated with 10,000 U/kg of rHuEPO. No other ROIs were found when comparing the PBS treated samples to the rHuEPO treated samples stained with antibodies from Panel 2 (parameters include - Sca-1, Gr-1, CD11b, and IL-7Rα) at any dosage level. These results indicate that SUG obtained ROIs only in the erythroid precursor populations, similar to the expert analysis and FDG, without a priori knowledge of the data. Moreover, our method did not require going over all data from all the samples manually, and was able to produce biologically interpretable results in a time efficient way.

RHuEPO Causes Dose Responsive Expansion of Erythroid Progenitors in the Bone Marrow

Using our method, we show the mean number of events within the ROI as a percentage of the controls as a function of dose level in Figure 5 with error bars indicating inter-quartile distance. The 30 and 100 IU/kg dose points were not included as no ROIs were identified at those levels. At the higher dosage levels we observed a dose responsive increase in the number of events in the ROI showing an expansion of the erythroid precursors with rHuEPO treatment. The mean fluorescent intensity of the CD71 parameter of the events in the ROI is shown as a function of dosage level in Figure 5B (normalized to controls). Here an increase in the Mean Fluorescent Intensity (MFI) is observed in the CD71 parameter in a dose responsive manner. The increases in erythroid progenitors and the corresponding effects on the CD71 parameter are similar to the expert data analysis (Fig. 2).

Postprocessing: Identification of Four (4) Subpopulations Within ROI

Postprocessing of the data is performed to extract more information from the results. This was performed using a cluster analysis to test for the presence of subpopulations within the ROI. The Robust Competitive Agglomeration (RCA) algorithm (24) was used to perform a cluster analysis. Briefly, RCA is a competitive agglomerative clustering algorithm. Unlike other popular partitioning clustering algorithms like *K*-means or Fuzzy C-means (25) clustering the number of clusters need not be specified in advance. If the number of clusters specified is larger than actually found within the data set, the above-mentioned algorithms partition the data by force splitting a single cluster into multiple smaller clusters. The RCA algorithm on the other hand begins by initializing the number of clusters in a data set to a large value. Typically this value is based on the type to data set that is being used. In flow data very often we begin with four clusters just within the light scatter distributions so a value between 10 and 20 to initialize will be sufficient depending on the number of immunostains.

In each iteration, through agglomeration and competition between the clusters for an individual data point, the number of clusters (initially large) is reduced to a number known as the "natural" number of clusters. The "natural" number of clusters depicts the number of clusters that can be expected in a data set. RCA utilizes a combination of two metrics to assign membership of a data point to a representative cluster. The first is a set of probabilistic constrained parameters that create good partitions between clusters and signifies the fuzzy association of a data point to a cluster center. It is the robust distance of the data point to each of the cluster centers. The second metric is an unconstrained possibilistic metric that assigns membership based on "typicality" of the cluster. The second weight accounts for the effects of outliers and noise in the data set and gives lower preference to data points "far away" from a cluster compared to ones that are "closer." (refer to Ref. 25, for further details). In our case to initialize the RCA algorithm to define ini-

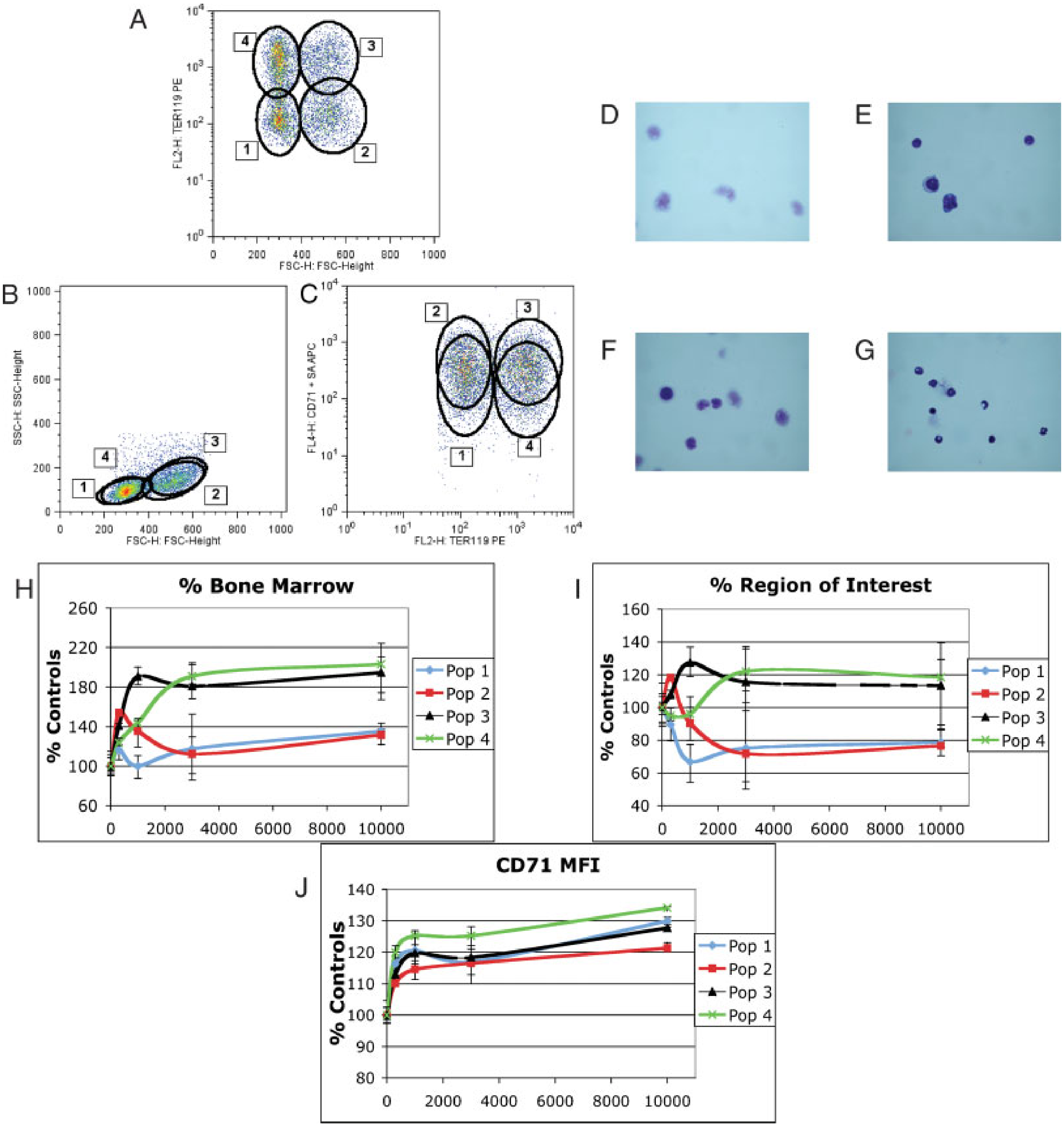

Figure 6. **A**–**C**: The results of the RCA clustering algorithm on one of the samples. RCA identified the presence of 4 clusters, which appear overlapping in B and C and appear most separated in A. Giemsa stained cytospin preparations of subpopulations of the clusters is shown in **D**–**G**. D is population 1, E is population 2, F is population 3, and G is population 4. **H**–**J**: The changes in the populations (Pop 1, Pop 2, Pop 3, and Pop 4) as a function of dosage level. Change in percentage of events within the marrow as a function of dose is shown in H. I shows the fraction of events in the ROI that belong to a subpopulation. J shows upregulation in CD71.

tial cluster centers and memberships we used a fuzzy C-means algorithm (25). Additionally, a Gaussian model for the data distribution was used to identify distance metrics in RCA. This resulted in hyper-ellipsoidal populations in the fluorescent intensity space. We used Matlab (R13.1, The Mathworks, Lowell, MA), to perform all the data analysis.

The inputs to the clustering algorithm were the FSC, SSC, TER-119 and CD71 parameter values of the ROI events. Clustering was performed 20 times for each sample resulting in four subpopulations within the ROI. Each time we set the initial cluster number to 10 clusters. Figure 6 shows three different two-dimensional projections of the four clusters, which

Table 2. Cluster classification

CLUSTER NUMBER	DESCRIPTION OF PARAMETERS (RELATIVELY)	NAME
1	Low FSC, Low SSC, Low TER119+, Low CD71+ (Small, simple, TER-119 dim, CD71 dim)	Population 1 (Pop 1)
2	High FSC, High SSC, Low TER119+, High CD71+ (Large, complex, TER-119 dim, CD71 bright)	Population 2 (Pop 2)
3	High FSC, High SSC, High TER119+, Low CD71+ (Large, complex, TER-119 bright, CD71 bright)	Population 3 (Pop 3)
4	Low FSC, Low SSC, High TER119+, High CD71+ (Small, simple, TER-119 dim, CD71 dim)	Population 4 (Pop 4)

include projections of FSC vs. TER-119 (Fig. 6A), FSC vs. SSC (Fig. 6B), and TER-119 vs. CD71 (Fig. 6C). The clusters appear most separated (nonoverlapping) in the FSC vs. TER-119 projection (Fig. 6A). In Table 2, we give the relative parameter values and label the clusters as Populations 1, 2, 3, and 4. Using all four parameters (FSC, SSC, TER-119, and CD71), we were able to sort the subpopulations. Microscopic examination of Giemsa stained cytocentrifuge preparations made from each subpopulation is shown in Figure 6D–6G. Identification of the four populations can be made based on the cytostain as: Population 1 consisting of pro-erythroblasts, Population 2 of early basophilic erythroblasts, Population 3 as late basophilic erythroblasts, and Population 4 as polychromatophilic and orthochromatophilic erythroblasts.

Differential Response of the Four Subpopulations to rHuEPO

The cluster analysis assigned each event within the ROI to belong to one of the four clusters (subpopulations) and the response of the subpopulations to rHuEPO as a function of dosage was determined. Figure 6H depicts the average change in the fraction of events within each of the subpopulations, normalized to the PBS treated samples. An increase in the number of events with dosage level is found in each of the subpopulations. A greater increase in the number of events is evident in Populations 3 and 4 relative to Populations 1 and 2. Figure 6I shows the fraction of events belonging to each of the subpopulations within the ROI. Even though the absolute number of events increases with dosage in each of the populations, a higher increase is seen in Populations 3 and 4 relative to Populations 1 and 2, indicating greater expansion. Figures 6J depicts the changes in the CD71 MFI of the events in the subpopulations as a function of dosage. A dose responsive increase in the CD71 MFI is seen in all the subpopulations.

Discussion

Though flow cytometry is a technology that has been in use for several decades, the commonly used data analysis methods are still primarily visual in nature. The expert analysis is rather subjective because it is predicated on topical knowledge and experience of the observer. The subjectivity in setting a region of interest can lead to different analysis results by different experts. For example, a simple process of estimating the percentage of positive events for a color can be skewed due to this visual analysis. The expert analysis can also be time consuming when a large number of samples are present, since every combination of two parameter plots is viewed when multiparameter data is present. The analysis time increases exponentially as the number of parameters increases in an experiment. As a result, FDG was proposed to reduce the analysis time and improve efficiency.

FDG is currently used to identify univariate and multivariate differences between samples. FDG is an algorithm that is capable of identifying a multivariate region of interest using the joint distributions rather than marginal distributions. Once that region is identified we still face the same problem of visualization to identify what the biological characteristics of that region are. Although it is very sensitive in identifying regions of difference, some of these may not be biologically interpretable.

Our approach, the Sequential Univariate Gating algorithm is shown to identify differences in univariate and multivariate distributions, in the presence of multiple (>2) samples per group. The *D*-value used in the Kolmogorov-Smirnov test is used as a metric to determine the existence of an ROI within a variable. To determine a multivariate region of interest, SUG is utilized in a nested manner. Performing the analysis in this fashion can identify differences that are biologically interpretable as the regions are orthogonal in nature. The location of the ROI in the marginal distributions can be ascertained very easily and thus lead to biological interpretability. SUG can only be used when there are more than two samples per group. The more the number of samples within each class, the greater the confidence with which a difference is ascertained.

Flow cytometric data sometimes suffers from changes in distributions when samples from the same source are run on different days or processed slight differently. The sources of variability can arise due to various factors including laser stability, laser alignment, antibody stability, and user handling. In our case, we assume that the data are sufficiently similar from all the samples within a group. However, due to the fact that we calculate the threshold for the *D*-value empirically it is a function of how similar the data are within each group. If at all there were shifts between data from the same group, SUG will not identify those differences as significant. SUG cannot be used to compare between only two samples.

The subjectivity within SUG arises from the window size *W* that is used to identify where the distribution between the two samples are different. After utilizing the algorithm on multiple datasets we have ascertained empirically that $W = 256$ provides a good enough resolution to identify the differences. However, it may not be necessary that the same fixed window size would be applicable to multiple datasets and maybe even different for the different variables. Even though we get good results with a fixed window size, a more efficient way to identify the location of the difference needs to be determined. In multivariate data it is possible that two data sets can have the same marginal distribution but different joint distributions. We could potentially use the KS distance and estimate a multivariate cumulative distribution and estimate the regions of difference in a similar manner. The problem lies in determining such a multivariate cumulative distribution. The methods that are present to identify this distribution are time consuming, such as a Parzen Windowing system (24). Computational complexity as well as taking into account sparse data sets makes it problematic to perform such cumulative distribution estimation. Thus we use a univariate method involving the KS distance metric and have demonstrated that a simple method like SUG can be used to determine multivariate regions of interest. Because we estimate the location of the ROI in an orthogonal fashion we can provide biological interpretability of the regions.

Most of the algorithmic methods are only tools to identify where the regions of interest exists. Currently, the data analysis can only be performed in a semi-supervised manner where the first stage involves using mathematical methods to look at the entire multivariate space to identify regions of differences. This should be followed at the next stage by a supervised expert analysis that can delve further into characterizing the biological differences.

Using our method we identified only regions that contained erythroid precursor populations to be effected by rHuEPO treatment. Previous studies have also shown this selectivity of rHuEPO to erythroid precursors (1,2). Regions specific only to erythroid lineage were identified and regions corresponding to the other lineages in the bone marrow were ignored. A dose responsive increase in the number of erythroid progenitors was observed in the marrow, indicating an expansion of the erythroid progenitors with rHuEPO treatment. No significant change in the TER-119 MFI was found, but a dose responsive increase in CD71 MFI was found, a marker for transferrin receptor on the cell surface. Increased expression possibly enhances the uptake of hemoglobin in the newly created erythroid precursors cells, as transferrin is required for the utilization of iron in the production of hemoglobin in these cells (1,2).

The presence of four subpopulations within the erythroid precursors was identified. An increase in the percentage of late basophilic, polychromatophilic and orthochromatophilic erythroblasts in a dose dependent manner is observed with rHuEPO treatment. The greater expansion of the more mature erythroid progenitors in a dose responsive manner suggests that rHuEPO not only stimulates the expansion of erythroid precursors but also moves the precursors along the path of differentiation and maturation faster, finally leading to more erythrocytes in blood. Another possibility is that rHuEPO rescues the late stage precursors from apoptosis, resulting in a greater number of mature precursors (1–3).

Conclusions

An algorithm employing sequential univariate gating is presented that can be used to identify differences between samples arising from two different conditions without a priori knowledge of the markers. SUG identifies differences between samples in the presence of multiple samples per group in a time efficient manner that correlates with the expert analysis. RHuEPO is shown to specifically drive a dose dependent expansion of erythroid precursor population in murine bone marrow and without any effects on other bone marrow populations studied. Morphologically different subpopulations within the erythroid progenitors respond differently to rHuEPO treatment.

Literature Cited

1. Fisher JW. Erythropoietin: Physiology and pharmacology update. Exp Biol Med 2003;228:1–14.
2. Spivak JL. The mechanism of action of erythropoietin. Int J Cell Cloning 1986;3:139–166.
3. Socolovsky M, Nam H, Fleming MD, Haase VH, Brugnara C, Lodish HF. Ineffective erythropoiesis in Stat5a(−/−)5b(−/−) mice due to decreased survival of early erythroblasts. Blood 2001;98:3261–3273.
4. Lin FK, Suggs S, Lin CH, Browne JK, Smailing R, Eric JC, Chen KK, Fox GM, Martin F, Wasser Z. Cloning and expression of the human erythropoietin gene. Proc Natl Acad Sci USA 1985;92:7850–7884.
5. Jacobs K, Shoemaker C, Rudersdorf R, Neill SD, Kaufman RJ, Mufson A, Seehra A, Jones SS, Hewick R, Fritsch EF. Isolation and characterization of genomic cDNA clones of human erythropoietin. Nature 1985;313:806–810.
6. Hoffman R, Benz EJ, Shattil SJ, Furie B, Cohen H, Silberstein LE, McGlave P. Hematology: Basic Principles and Practice. Elsevier; 2000.
7. Lesley J, Hyman R, Schulte R, Trotter J. Expression of transferrin receptor on murine hematopoietic progenitors. Cell Immunol 1984;83:14–25.
8. Zhang J, Socolovsky M, Gross AW, Lodish HF. Role of ras signaling in erythroid differentiation of mouse fetal liver cells: Functional analysis by a flow cytometry-based novel culture system. Blood 2003;102:3938–3946.
9. Kina T, Ikuta K, Takayama E, Wada K, Majumdar AS, Weissman IL, Katsura Y. The monoclonal antibody TER-119 recognizes a molecule associated with glycophorin A and specifically marks the late stages of murine erythroid lineage. Br J Hematol 2000;109:280–287.
10. Ito M, Anan K, Misawa M, Kai S, Hara J. In vitro differnetiation of Sca-1+Lin- cells into myeloid. B cell and T cell lineages. Stem Cells 1996;14:412–418.
11. Lagasse E, Weissman IL. Flow Cytometric identification of murine neutrophils and monocytes. J Immunol Methods 1996;197:139–150.
12. Fleming TJ, Fleming ML, Malek TR. Selective expression of Ly-6G on myeloid lineage cells in mouse bone marrow. RB6-8C5 mAb to granulocyte-differentiation antigen (Gr-1) detects members of the Ly-6 family. J Immunol Methods 1993;151:2399–2408.
13. Goodwin RG, Friend D, Ziegler SF, Jerzy R, Falk BA, Gimpel S, Cosman D, March SKDJ, Namen AE, Park LS. Cloning of the human and murine interleukin-7 receptors: Demonstration of a soluble form and homology to a new receptor superfamily. Cell 1990;60:941–951.
14. Overton RW. Modified histogram subtraction technique for analysis of flow cytometry data. Cytometry 1999;9:619–626.
15. Cox C, Reeder JE, Robinson RD, Suppes SB, Wheeless LL. Comparison of frequency distributions in flow cytometry. Cytometry 1988;9:291–298.
16. Lampariello F. Evaluation of the number of positive cells from flow cytometryic immunoassays by mathematical modelling of cellular autofluorescence. Cytometry 1994;15:294–301.
17. Lampariello F, Aiello A. Complete mathematical modeling method for the analysis of immunofluorescence distribution composed of negative and weakly positive cells. Cytometry 1994;32:241–254.
18. Bagwell CB. A journey through flow cytometric immunofluorescence analyses. Clin Immunol Newsl 1996:33–37.
19. Bagwell CB, Hudson JL, Irvin GL. Nonparametric flow cytometry analysis. J Histochem Cytochem 1979;27:293–296.
20. Roederer M, Treister A, Moore W, Herzenberg L. Probability binning comparison: A metric for quantitating univariate distribution differences. Cytometry 2001;45:37–46.
21. Roederer M, Treister A, Moore W, Hardy R, Herzenberg L. Probability binning comparison: A metric for quantitating multivariate distribution differences. Cytometry 2001;45:47–55.
22. Roederer M, Hardy R. Frequency difference gating: A multivariate method for identifying subsets that differ between samples. Cytometry 2001;45:56–64.
23. Siegel S, Castellan NJ. Nonparametric Statistics for The Behavioral Sciences. McGraw-Hill; 1988.
24. Frigui H, Krishnapuram R. A robust competitive clustering algorithm with applications in computer vision. IEEE Trans Pattern Anal 1999;21:450–465.
25. Duda R, Hart P, Stork D. Pattern Classification. Wiley Interscience; 2000.