ORIGINAL ARTICLE

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

MIFlowCyt

Consistent Quantitative Gene Product Expression: #1. Automated Identification of Regenerating Bone Marrow Cell Populations Using Support Vector Machines

Andrew P. Voigt,[1] Lisa Eidenschink Brodersen,[1] Laura Pardo,[2] Soheil Meshinchi,[2] Michael R. Loken[1]*

[1]HematoLogics, Inc., Seattle, WA

[2]Fred Hutchinson Cancer Research Center, Seattle, WA

Received 28 January 2016; Revised 20 April 2016; Accepted 10 June 2016

Grant sponsor: National Cancer Institute (NCI) Biomarker, Imaging, and Quality of Life Supplemental Funding Program (BIQSFP), Grant number: 1 U10 CA180886.

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Michael R. Loken, HematoLogics, Inc., 3161 Elliot Ave., Suite 200, Seattle, WA 98121, USA.
E-mail: mrloken@hematologics.com

Conflict of Interest: A.P.V. and L.E.B. are employed by HematoLogics, Inc. M.R.L. is an equity owner and President of HematoLogics, Inc.

Published online 14 July 2016 in Wiley Online Library (wileyonlinelibrary.com)

DOI: 10.1002/cyto.a.22905

ISAC
International Society for Advancement of Cytometry

• Abstract

Identification and quantification of maturing hematopoietic cell populations in flow cytometry data sets is a complex and sometimes irreproducible step in data analysis. Supervised machine learning algorithms present promise to automatically classify cells into populations, reducing subjective bias in data analysis. We describe the use of support vector machines (SVMs), a supervised algorithm, to reproducibly identify two distinctly different populations of normal hematopoietic cells, mature lymphocytes and uncommitted progenitor cells, in the challenging setting of pediatric bone marrow specimens obtained 1 month after chemotherapy. Four-color flow cytometry data were collected on a FACS Calibur for 77 randomly selected postchemotherapy pediatric patients enrolled on the Children's Oncology Group clinical trial AAML1031. These patients demonstrated no evidence of detectable residual disease and were divided into training ($n = 27$) and testing ($n = 50$) cohorts. SVMs were trained to identify mature lymphocytes and uncommitted progenitor cells in the training cohort before independent evaluation of prediction efficiency in the testing cohort. Both SVMs demonstrated high predictive performance (lymphocyte SVM: sensitivity >0.99, specificity >0.99; uncommitted progenitor cell SVM: sensitivity $= 0.94$, specificity >0.99) and closely mirrored manual cell classifications by two expert-analysts. SVMs present an efficient, automated methodology for identifying normal cell populations even in stressed bone marrows, replicating the performance of an expert while reducing the intrinsic bias of gating procedures between multiple analysts. © 2016 The Authors. Cytometry Part A published by Wiley Periodicals, Inc. on behalf of ISAC.

• Key terms

Key terms: flow cytometry; support vector machines; classification; lymphocytes; uncommitted progenitor cells; bone marrow

Introduction

DURING the process of monitoring response to acute myeloid leukemia (AML) therapy using flow cytometry, it was noted that patterns of antigen expression of the normal, regenerating hematopoietic cells were remarkably similar from patient to patient, even early postchemotherapeutic treatment. For three separate Children's Oncology Group AML trials over a period of 13 years, standardized four-color panels were used to pattern early stages of hematopoiesis (1,2). Using the "difference from normal" flow cytometric approach, low levels of neoplastic leukemia were distinguished from normal regenerating myeloid cells based on the detection of small populations of viable cells exhibiting aberrant antigen expression. These studies comprised the analysis of over 10,000 specimens obtained from >2,300 patients monitored throughout the course of treatment.

Table 1. Monoclonal antibody combinations

TUBE NO.	FITC	PE	PERCP	APC
1	HLA-DR	CD11b	CD45	CD34
Clone	L243 (BD)	D12 (BD)	2D1 (BD)	8G12 (BD)
2	CD36	CD38	CD45	CD34
Clone	FA6.152 (BC)	HB7 (BD)		
3	CD16	CD13	CD45	CD34
Clone	3G8 (BD)	L138 (BD)		
4	CD14	CD33	CD45	CD34
Clone	Mϕ/P9 (BD)	P67.6 (BD)		
5	CD7	CD56	CD45	CD34
Clone	4H9 (BD)	MY31 (BD)		
6	CD38	CD117	CD45	CD34
Clone	HIT2 (Invitro)	104D2 (BD)		
7	CD36	CD64	CD45	CD34
Clone	FA6.152 (BC)	22 (T)		
8	CD19	CD123	CD45	CD34
Clone	4G7 (BD)	9F5 (BD)		

Quantifying the variance of gene product (antigen) expression within and between individuals on these recovering bone marrows is complicated by the heterogenous composition of the specimens. There are 11 different lineages of cells that must be identified within the bone marrow, in addition to all stages of maturation from the hematopoietic stem cell to the mature blood cells. As a result, cell populations are not discrete and the proportions of different cell types vary depending upon the rate of hematopoietic reconstitution following therapy. However, the positions of all cell populations in the six-dimensional data space appeared to remain constant.

To quantify the variance of gene product expression without subjective gating bias, the supervised algorithms support vector machines (SVMs) were selected to identify key reference cell populations. This technique defines a multi-dimensional plane based on a teaching set created by an expert. This plane is then used to identify corresponding cell populations in new, independent patient data sets that combine the exact same parameters based on constancy of position in N-dimensional space. This article describes the efficacy of the SVMs to reproduce the expert-detection of two of the reference populations, mature lymphocytes and uncommitted progenitor cells. Two companion papers then use this approach to determine the variability of gene product expression among normal cells in postchemotherapy regenerating bone marrow specimens.

Materials and Methods

Patient Data Set

Patients younger than 21 years with newly diagnosed de novo AML who were enrolled on Children's Oncology Group (COG) AAML1031 were eligible for this study. The trial was conducted in accordance with the Declaration of Helsinki and registered at www.clinicaltrials.gov as NCT01371981. A total of 77 randomly selected, pediatric AML patients obtained approximately 1 month postchemotherapy were identified as having no evidence of residual disease (3). Patients were randomly assigned to training ($n = 27$) and testing ($n = 50$) cohorts.

Specimen Collection

Bone marrow aspirates were collected in heparin (the preferred anticoagulant) or EDTA. The data were obtained over a period of 3 years and 6 months using three separate flow cytometers, multiple reagent lots, and processed by multiple technicians.

Flow Cytometry

Specimens were processed as routine clinical bone marrows as previously described (1). Briefly, 100 μL of bone marrow was added to cocktails of pretittered antibodies at room temperature in the dark. Red blood cells were lysed using 3.5 mL of buffered NH_4Cl (0.83%) at 37°C for 5 min before centrifugation at 300G. Cells were then washed with 3 mL of phosphate buffered saline containing 2% fetal calf serum and resuspended to 0.5 mL in 1% paraformaldehyde for analysis on one of three FACS Calibur instruments (Becton Dickinson Biosciences, San Jose, CA). 200,000 events were collected for each tube. The flow cytometers were cross standardized and calibrated using RCP-30A and RFP-30A beads (Spherotech, Lake Forest, IL) with spectral compensation performed using peripheral blood cells labeled with CD4 (SK3, BD) conjugated to fluorescein (FITC), phycoerythrin (PE), peridinin chlorophyll protein (PerCP), or allophycocyanin (APC). Eight combinations of antibodies are presented in Table 1. The eighth tube was added after the beginning of the study to identify immature B lymphoid precursors/plasmacytic dendritic cells and basophils, so this reagent combination was only implemented in 45 of the 50 patients in the testing cohort.

Support Vector Machines

An expert-analyst (MRL) classified mature lymphocyte and uncommitted progenitor cell populations using Winlist (Verity Software House, Topsham, ME) for all patients in both the training and testing cohorts. All flow-cytometry data

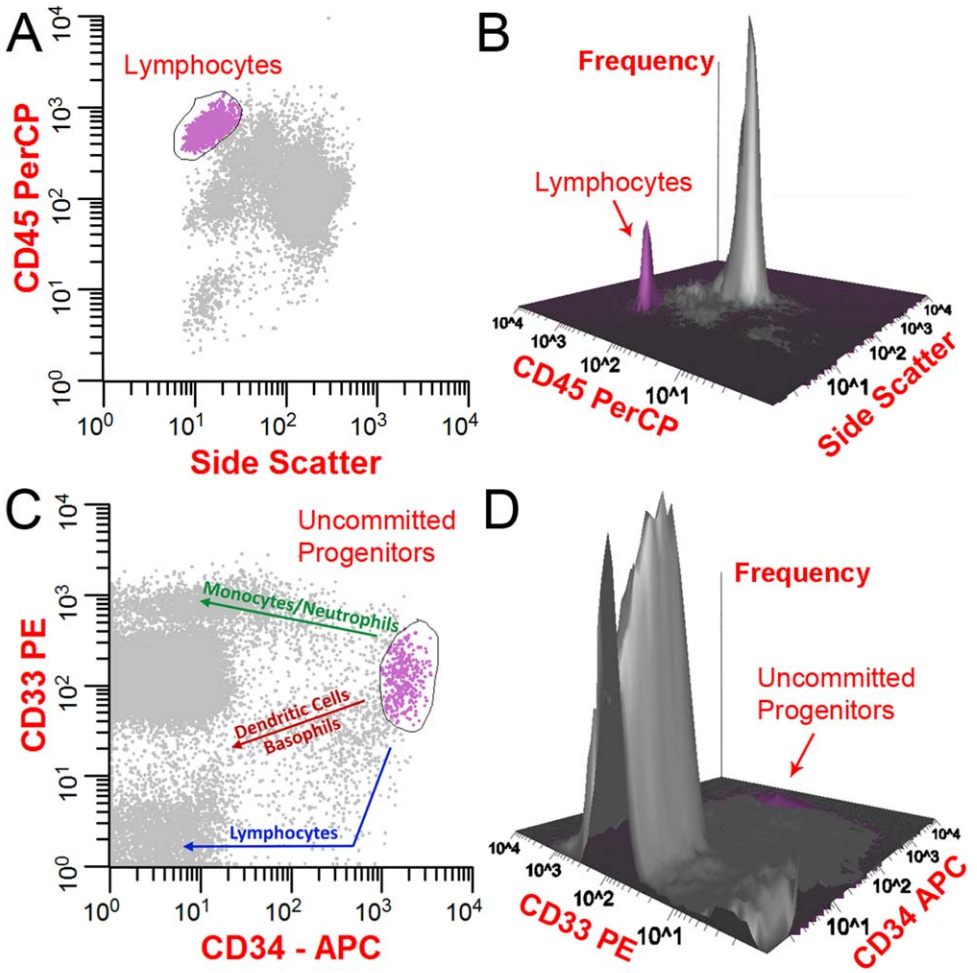

Figure 1. Expert cellular classifications for SVM training: (A) Lymphocytes (purple) were identified by an expert analyst as a discrete cluster of events with high CD45 intensity and low SSC. (B) The high relative frequency of the lymphocyte population is depicted on a 3D plot of CD45, SSC, and frequency. (C) Uncommitted progenitor cells (purple) were identified by an expert analyst as the cells with the brightest CD34 intensity before a gain or loss of CD33. Maturational pathways as these cells commit to monocyte, neutrophil, dendritic, basophil, and lymphocyte lineages are shown with arrows. (D) The low relative frequency of the uncommitted progenitor cell population is depicted on a 3D plot of CD33, CD34, and frequency.

were analyzed in log space, with the exception of forward light scatter (FSC).

Expert-classified cell populations from the cohorts were exported from Winlist into R software (4). For both lymphocytes and uncommitted progenitor cells, all data from the training cohort patients were merged to create one file that differentiated cells within the population of interest (+1) from the remainder of cells (−1). Log base-10 transformations were computed for right-angle side scatter (SSC), FITC, PE, PerCP, and APC parameters prior to SVM training. The linear range of FSC was scaled to emulate a zero-to-four log scale. An SVM was trained to identify each target population with a radial kernel using the e1071 library (5). A basic overview of an SVM algorithm is provided in the supplementary data (Supplementary Fig. 1). To manage the computational intensity of training the lymphocyte SVM, 1000 random manually classified lymphocytes (+1) and 1,500 random manually classified nonlymphocytes (−1) from each patient were merged to form the lymphocyte training data set. Because only support vectors influence the SVM decision boundary, the 1,500 nonlymphocytes in each patient were selected to enrich the training cohort with potential support vectors. Therefore, the nonlymphocytes were selected based on similar properties to the lymphocytes: a SSC of <2 log units and a CD45 intensity >2 log units. Similar exclusion procedures have been utilized to enrich datasets for support vectors and reduce the training time of SVMs (6,7). To manage the computational intensity of the uncommitted progenitor cell SVM training, manually classified uncommitted progenitor cells (+1) and similar nonuncommitted progenitor cells (−1) with a CD34 intensity >2 log units from each patient were merged to form the uncommitted progenitor cell training set. Because 27 patients comprised each training data set, a ninefold leave-three-out cross validation was applied on the training data to determine optimal C and γ parameters for both SVMs (8).

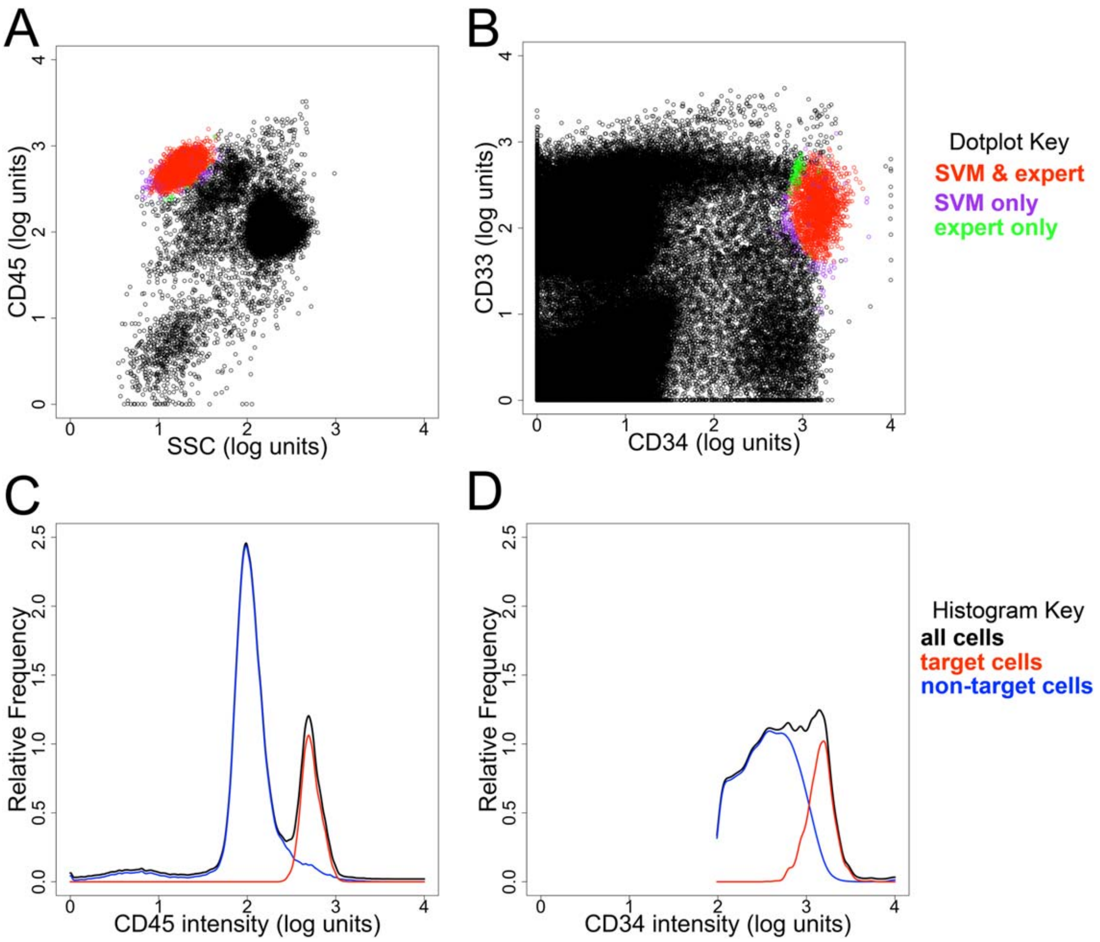

Figure 2. Qualitative evaluation of SVM predictions for a test cohort patient. (A, B) Each SVM prediction was compared to an independent manual classification of lymphocytes (A) and uncommitted progenitor cells (B). Cells colored in red were classified by both the expert analyst and the SVM. Discrepant classifications of events colored in green were identified only by the expert analyst, while events colored in purple were classified only by the SVM. The discrepant classifications occur at the outer boundaries of the target population. (C) A frequency curve of CD45 intensities for all cells (black) reveals a uniform subpopulation of cells with high-intensity of CD45, which is comprised almost entirely of SVM-classified lymphocytes (red). The majority of nonlymphocytes have a lower CD45 intensity (blue), with the remainder of bright CD45 cells classified as monocytes. (D) A frequency curve of CD34 intensities for all cells with a CD34 intensity greater than two log units (black) reveals a heterogeneous distribution of CD34 intensities. SVM-classification reveals a homogenous, high intensity CD34 peak for the uncommitted progenitor cells (red) compared to the lineage committed progenitor cells (blue).

Statistics

Sensitivity and specificity were used to evaluate prediction efficiency, and were calculated according to the standard definitions, where the manual classifications of the expert analyst (MRL) were regarded as the true classifications.

Because predicted cell populations were often small when compared to the total number of collected events, prediction efficiency was also evaluated with the Matthews correlation coefficient (MCC) (9). The MCC is a balanced measure of the quality of a binary prediction algorithm even when classes are of different sizes. The MCC returns a value between −1 and 1, where 1 represents a perfect prediction, 0 represents a random prediction, and −1 represents complete disagreement between the prediction and observation, and is calculated by the following definition:

$$\text{MCC}= \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where $TP =$ # True Positives, $TN =$ # True Negatives, $FP =$ # False Positives, $FN =$ # False Negatives.

Results

SVM Training

Two different cell types were selected to test the performance of SVMs in a complex data set of normal regenerating bone marrow cells following chemotherapy. The first group of cells, mature lymphocytes, is relatively straightforward for an expert analyst to identify, and is presented to demonstrate the efficacy of classifying a homogenous, discrete, relatively frequent cell population with SVMs. The second group of cells, uncommitted progenitor cells, is challenging to identify as this population is a heterogeneous, nonlinearly separable population of infrequent cells.

SVMs were initially trained to identify each target population in regenerating bone marrow using the combination of

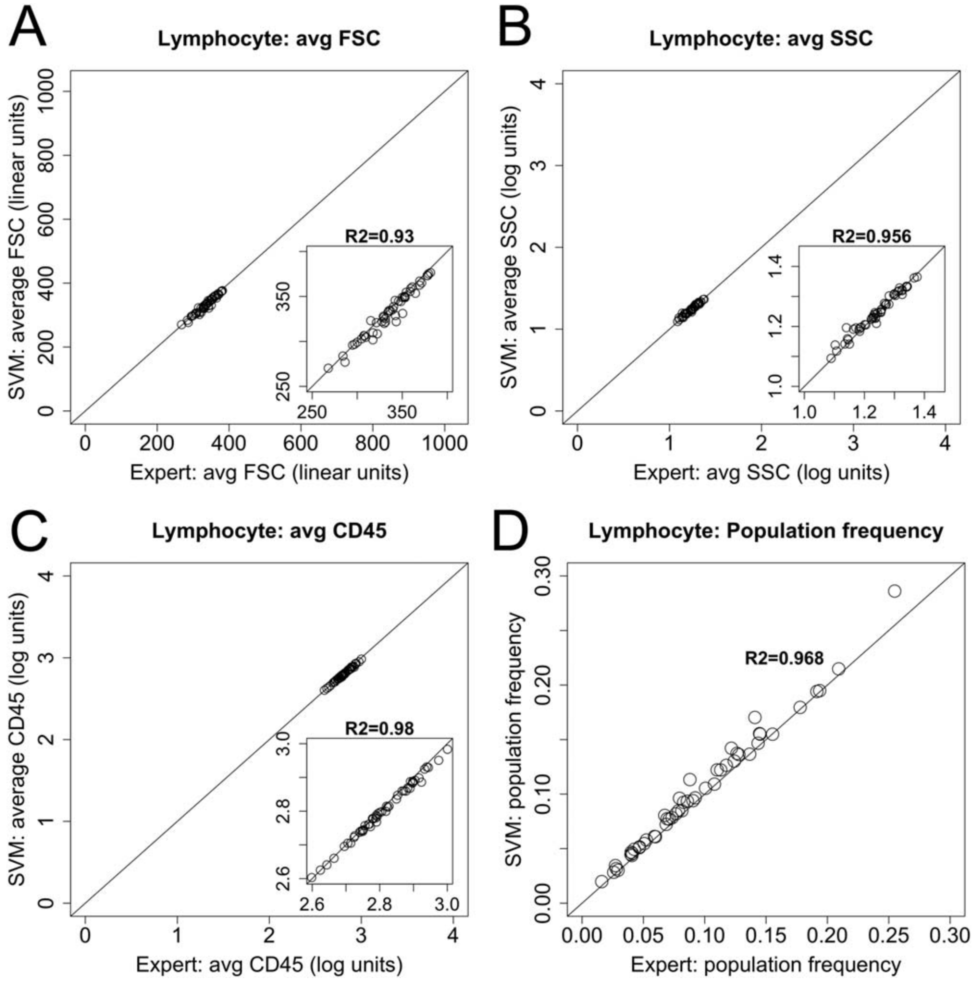

Figure 3. Intensity and frequency correlations between the lymphocyte SVM and expert. (A–C) The average intensities of FSC, SSC, and CD45 were computed for expert-identified lymphocytes (*x*-axis) and the SVM-identified lymphocytes (*y*-axis) for the test cohort. Each plot is scaled to display the range of intensity values in a manual analysis, and the lower-right quadrant of each plot provides a magnified view of the correlation. (D) The frequency of expert-identified lymphocytes (*x*-axis) versus the SVM-identified lymphocytes (*y*-axis) for the test cohort. R2 correlations were calculated according to the linear model SVM = Expert.

parameters found in Tube 4 of the antibody panel (Table 1): linear FSC, log SSC, log CD14 (FITC), log CD33 (PE), log CD45 (PerCP), and log CD34 (APC). Lymphocyte cell populations are identified as a distinct, homogenous cluster of events with high CD45 intensity and low SSC (Fig. 1A). This population of cells is of relatively high frequency, typically comprising 5–20% of cells in the bone marrow (Fig. 1B). An expert-analyst manually identified this discrete lymphocyte population by CD45 versus log SSC gating in combination with FSC in the training cohort patients, and these manual classifications were used to train and cross-validate the lymphocyte SVM. The inclusion of CD14 (FITC) and CD33 (PE) parameters did not further improve predictive performance in SVM training (data not shown). Consequently, mature lymphocyte SVM training was only performed with FSC, SSC, CD45, and CD34 parameters.

In contrast, uncommitted progenitor cells are not discrete but a continuous population. These cells include the hematopoietic stem cells and multipotent progenitor cells that have not yet expressed any lineage-associated surface gene products (10,11). Uncommitted progenitor cells are defined by a homogenous high expression of the key antigen, CD34, and coexpression of CD33 (Fig. 1C). CD33 changes in intensity once the uncommitted progenitors decide a maturational path, increasing in CD33 expression for monocytes and neutrophils, decreasing in CD33 expression for plasmacytic dendritic cells, or rapidly losing CD33 for lymphoid progenitor cells. An expert analyst manually identified the uncommitted progenitor population by gating the brightest intensity CD34 cells before the gain or loss of CD33 (Fig. 1C). This population is of noticeably lower frequency than the lymphocyte population, typically comprising 0.5–2% of all cells in the bone marrow (Fig. 1D). This manual classification was completed for all training cohort patients and used to train and cross-validate the uncommitted progenitor cell SVM. All six parameters were necessary to train this SVM with maximal prediction performance (data not shown).

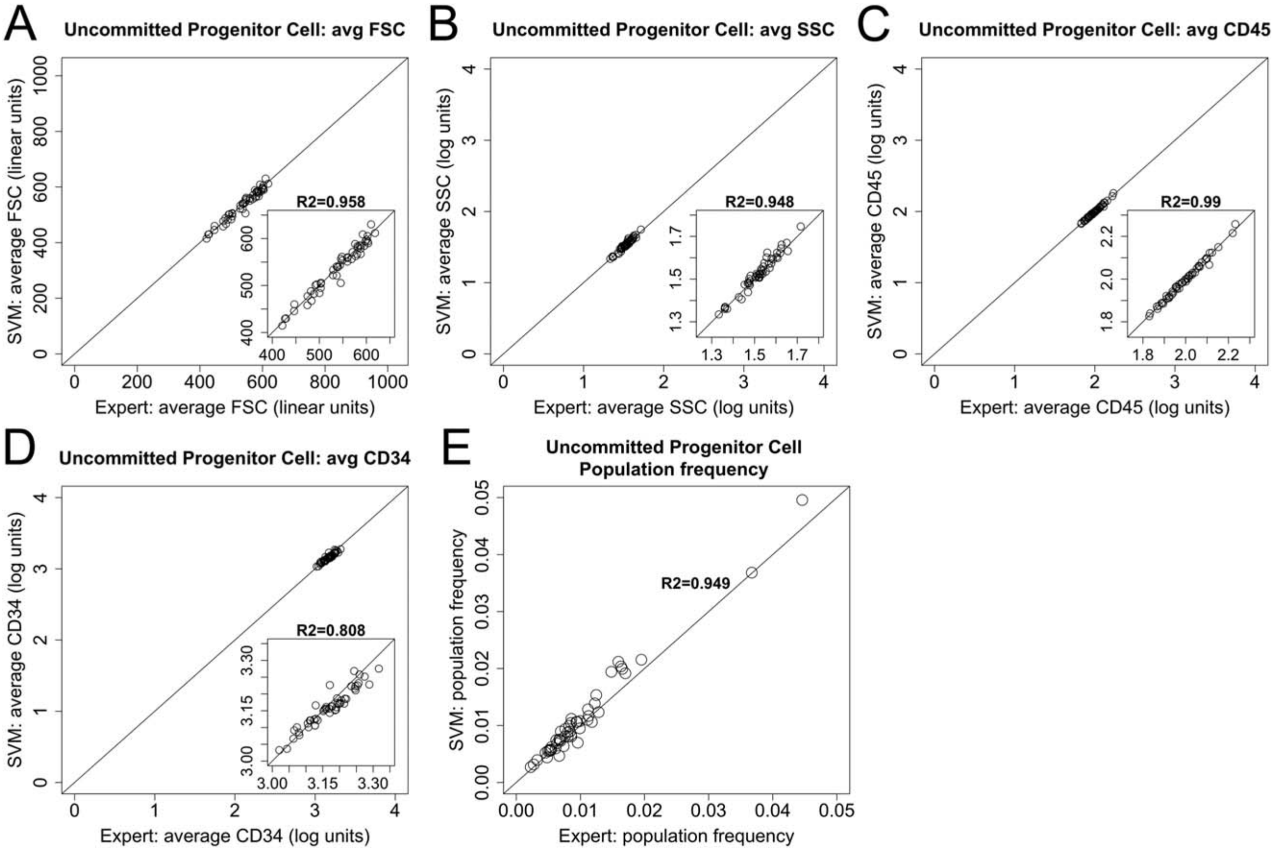

Figure 4. Intensity and frequency correlations between uncommitted progenitor cell SVM and expert. (A–D) The average intensities of FSC, SSC, CD45, and CD34 were computed for expert-identified uncommitted progenitor cells (*x*-axis) and the SVM-identified uncommitted progenitor cells (*y*-axis) for the test cohort. Each plot is scaled to display the range of intensity values in a manual analysis, and the lower-right quadrant of each plot provides a magnified view of the correlation. (E) The frequency of expert-identified uncommitted progenitor cells (*x*-axis) versus the SVM-identified uncommitted progenitor cells (*y*-axis) for the test cohort. R2 correlations were calculated according to the linear model SVM = Expert.

Qualitative Evaluation of SVM Population Predictions

After training, the two SVMs were applied to identify lymphocytes and uncommitted progenitor cells in the independent test cohort. Algorithmic predictions were first qualitatively evaluated. The SVM-predicted populations were compared to the expert analyst's test cohort classifications in each test patient for the lymphocytes (Fig. 2A) and uncommitted progenitor cells (Fig. 2B). The population predictions agreed with the corresponding expert-classifications, and the majority of cells were correctly identified in each target population. However as expected, some misclassifications occurred, and these discrepant predictions were located at the edges, rather than the center, of the SVM decision boundary. Additionally, histograms of key antigens were compared between SVM-predictions and all other events in each test patient. In the display of CD45 intensities of all cells, a uniform, bright CD45 population can be readily visualized (Fig. 2C). This population is comprised almost entirely of SVM-classified lymphocytes (with the remainder of bright CD45 positive cells identified as predominantly monocytes). In contrast, initial analysis of all CD34 positive cells (with an intensity greater than two log units) revealed a heterogeneous distribution of CD34 intensities (Fig. 2D). The SVM-classified uncommitted progenitor cells can be identified as a homogenous high intensity CD34 peak clearly distinguishable within this heterogeneous population.

Quantitative Evaluation of SVM Population Predictions

Classification performance of each SVM was quantitatively evaluated. Each SVM prediction was compared to the independent expert classification for all patients in the testing cohort. The average sensitivity, specificity, and MCC of the lymphocyte SVM and the uncommitted progenitor cell SVM demonstrated remarkably high predictive performance by all three metrics (Table 2). A similar comparison was made to a second expert (LP) who analyzed the same data set with instructions to identify the lymphocytes and uncommitted progenitor cells, without disclosing the details of how boundaries were established. The average sensitivity, specificity, and MCC of the classifications of the second expert were computed in comparison to the first expert (Table 2). The SVMs and second expert (LP) demonstrated similar performance replicating the gates of the first expert (MRL).

Table 2. Prediction efficiency of each SVM

	LYMPHOCYTE SVM	UNCOMMITTED PROGENITOR CELL SVM	LYMPHOCYTE EXPERT 2	UNCOMMITTED PROGENITOR CELL EXPERT 2
Sensitivity	0.994	0.944	0.934	0.974
Specificity	0.991	0.998	0.996	0.998
MCC	0.948	0.904	0.940	0.921

Average sensitivity, specificity, and MCC values were computed by calculating the mean of all sensitivity, specificity, and MCC measurements for the 50 test predictions for both the SVM and expert 2 (LP). Classifications from expert 1 (MRL) were designated as the true classifications.

Classification performance was additionally evaluated by correlating population frequency and average surface gene product intensities between each SVM-predicted cell population and the expert-identified cell population in the test cohort. R^2 values were calculated to evaluate how well the data fit a hypothesized linear model (SVM = Expert) (Figs. (3) and (4)). On a four-decade scale, the relationship between the SVM and the expert analyst are essentially indistinguishable. Within the intensity and frequency regions of interest, small discrepancies between the SVM and expert can be identified. However, 97% of lymphocyte SVM intensities and 92% of uncommitted progenitor cell SVM intensities were within 0.05 log units of the expert-characterized intensity, demonstrating the high-functional agreement between the SVM and the expert. A similar high correlation between the two experts analyzing the same dataset was observed (Supplemental Figs. 2 and 3), suggesting that the SVM and second expert were comparable in replicating the classifications of the first expert.

Replicate Identification of Lymphocytes and Uncommitted Progenitor Cells

In addition to the initial panel of CD14/CD33/CD45/CD34 (Table 1, Tube 4), seven other combinations of FITC and PE conjugated antibodies were used to study cell surface gene product relationships in each regenerating bone marrow specimen. Collectively, these unique reagent combinations provided eight replicate analyses in each patient to identify the lymphocytes and uncommitted progenitor cells with SVMs.

The initial lymphocyte SVM was trained using only FSC, SSC, CD45, and CD34 parameters, characteristics which were conserved between the eight reagent combinations. Therefore, the original lymphocyte SVM was applied to classify this cellular population in each of the additional combinations of FITC and PE conjugated-antibodies. In contrast, the initial uncommitted progenitor cell SVM training required both CD14 FITC and CD33 PE parameters to achieve optimal predictive performance. Therefore, this specific SVM could not be applied to any of the other combinations of reagents. Consequently, eight unique SVMs were trained (one for each regent combination) to identify uncommitted progenitor cells. These independently trained SVMs were then applied to identify the uncommitted progenitor cells in each of the eight tubes with different FITC and PE conjugated-antibodies, providing replicate analyses identifying the same cell population but with different SVMs.

For both lymphocytes and uncommitted progenitor cells, replicate intensities of FSC, SSC, and CD45 and replicate cell frequency measurements between the eight SVM-predictions were analyzed in the test cohort. Replicate CD34 intensities were additionally analyzed for uncommitted progenitor cells but not lymphocytes, as the lymphocytes do not express this gene product. The variability in average intensity and frequency between the eight predictions was computed for each patient, and averaged for the test cohort (Table 3). The lymphocytes, identified by the same SVM in all eight combinations of antibodies, demonstrated minimal variation of FSC, SSC, and CD45 between the eight populations identified in each patient. In addition, the replicate lymphocyte frequency variation is low, suggesting that a similar proportion of cells were identified in each of the eight replicates. Collectively, this data demonstrates that the lymphocyte-SVM trained using Tube 4 of the reagent panel identified a similar population of cells in each combination of reagents studied.

Table 3. Target population variability between replicate analyses

	LYMPHOCYTES	UNCOMMITTED PROGENITOR CELLS
Replicate FSC SD (linear units)	3.28	25.6
Replicate SSC SD (log units)	0.022	0.053
Replicate CD45 SD (log units)	0.017	0.032
Replicate CD34 SD (log units)	N.A.	0.031
Replicate frequency SD	0.0053	0.0019

In each test patient, the average FSC, SSC, and CD45 intensities were calculated for the eight lymphocyte and uncommitted progenitor cell predictions. Replicate CD34 intensities were additionally computed for uncommitted progenitor cells but not lymphocytes, as the lymphocytes do not express this gene product. The variation (standard deviation) between these eight measurements was calculated for each test patient and averaged for the test patient cohort. Additionally, the predicted population frequency (predicted events/total events) was calculated for each of the eight predictions. The variation (standard deviation) of the predicted population frequency between the eight predictions was calculated for each patient and averaged for the test patient cohort.

In spite of the independent algorithm-training utilizing different combinations of reagents, each of the eight uncommitted progenitor cell SVMs identified a population of cells with similar low variation of FSC, SSC, CD45, and CD34 within each patient. Although the uncommitted progenitor cells are infrequent, on average comprising 1.2% of cells in the marrow in the test cohort, the variation in frequency between the eight replicate predictions was only 0.2%. Again, this provides strong evidence that a similar proportion of cells were identified in each of the eight replicate tests. Collectively, these data suggests that despite independent SVM training, a remarkably similar population of uncommitted progenitor cells were identified in each tube of the reagent panel, further validating the high efficacy of SVM-identification of this population.

Discussion

To quantify the variance of gene product expression of hematopoietic cell populations in the bone marrow, reference populations must be identified with limited analytical bias. In this manuscript, we show that SVMs present an automated methodology to identify such reference populations with high agreement to the expert. By testing SVMs on pediatric bone marrows postchemotherapy, we studied the efficacy of SVMs in a challenging analytical setting, comprised of stressed, heterogeneous, maturing, multilineage populations of cells. This challenging setting replicates the difficulties of performing clinical flow cytometry analysis for the detection of residual disease based on "difference from normal" (1). The location of the normal cells in *n*-dimensional data space is crucial in this approach to detect low levels of cells expressing phenotypic aberrancies (12–14).

Each metric of predictive performance evaluated in this manuscript indicates that SVMs closely mirror expert gating designations. First, high sensitivity, specificity, and MCC measurements demonstrated strong agreement between SVM and expert classifications. Second, frequency and mean fluorescent intensity measurements were remarkably correlated between expert- and SVM-identified populations. Third, the low variation of frequency and mean fluorescent intensity measurements for replicate analysis within each individual patient indicates that SVMs reproducibly characterize the same cell population studied with different combinations of reagents, illustrating the power of this algorithm in a stable analytical system. Fourth, the comparison between two experts analyzing the same data set illustrates that the SVM replicates expert-trained classifications as well as a second, independent expert. Collectively, this data shows that SVMs can effectively replicate the expert identification of lymphocytes and uncommitted progenitor cells while reducing analytical bias inherent to manual gating procedures.

The success of population classification using SVMs is dependent on two assumptions. First, quality control of the system must be rigorous to maintain identical data collection on multiple instruments with multiple lots of reagents over an extended period of time, such as the length of a clinical trial. In the data presented here, three separate, cross standardized flow cytometers were used in data collection spanning three and a half years, employing multiple lots of titered reagents. Second, the variability of cellular characteristics between patients must be minimal. As SVMs are trained using only fluorescent and light scattering intensity measurements, each population's collective surface gene product expression must reside in a unique and constant region within the six-dimensional space to be efficiently identified by the SVM. In this test system, the location of the mature lymphocytes and the normal uncommitted progenitor cells were sufficiently stable to identify the target populations even in stressed regenerating bone marrow specimens after chemotherapeutic treatment. Although, the cellular frequency of these populations varied significantly, the location of these populations with regard to their assayed cellular characteristics remained stable enough for the trained SVM to replicate the expert analysis for each test patient.

There are limitations associated with SVM analysis of flow cytometry data. SVMs by definition only perform predictions in the two-class setting, distinguishing between the target population and all other events. Hence, identifying multiple cell populations within the same patient requires the training of multiple unique SVMs. In addition, several unique SVMs are sometimes needed to identify the same population of cells in different combinations of reagents, as demonstrated by the identification of the uncommitted progenitor cells. The training of multiple SVMs can be a time-consuming process for the expert analyst and computationally intensive. Further, the selection of kernel and cost parameters can affect the performance of an SVM. Cross-validation of training data can be used to estimate predictive performance for numerous combinations of input variables, allowing for the identification of kernel and cost parameters that lead to better performance. Yet cross-validation can likewise be computationally intensive for multiple SVMs in a large training cohort. Because only support vectors influence the SVM-decision boundary, computational time can be reduced by training the algorithm on a subset of the data with characteristics similar to the population of interest (6,7). For example, the uncommitted progenitor cell SVM was trained only on cells with CD34 intensities >2 log units. In the testing cohort, SVM-identification of both the lymphocytes and uncommitted progenitor cells averaged 42 sec on a machine with 16GB of RAM in patients with 200,000 collected events.

The field of automated flow cytometry analysis contains numerous supervised, semisupervised, and unsupervised algorithms to identify populations of hematopoetic cells (15–21). However, the majority of existing algorithms include unsupervised or clustering components to assemble high-density groups of data points into populations (22–25). SVMs cannot reproducibly identify cell populations in which gene product expression varies substantially between patients. In such a setting, unsupervised algorithms may better identify clusters of cells independent of consistent fluorescent intensity measurements. Instead, SVMs are predicated on identifying constant quantitative surface gene product expression patterns found on biological cell populations independent of cell frequency. SVMs have shown success in classifying cell populations in peripheral blood (26) and murine bone marrow (27), and even at identifying acute lymphoblastic leukemia with reproducible phenotypes (28). In this manuscript, we extend this approach to identify crucial reference populations in postchemotherapy pediatric bone marrow specimens with reagent combinations used to study over 10,000 patients in three large clinical trials.

SVMs present an automated methodology to identify reference populations with high expert agreement. This expert-trained algorithm exhibits high predictive performance in the challenging specimen type of pediatric bone marrow after chemotherapy. Hence, SVMs present a powerful tool to precisely quantify the variation of gene product expression with limited analytic bias.

Acknowledgment

The authors wish to thank Abel Wolman, Ph.D. for his valuable discussions leading to the successful use of this approach to flow cytometry.

Literature Cited

1. Loken M, Alonzo T, Pardo L, Gerbing R, Raimondi S, Hirsch B, Ho P, Franklin J, Cooper T, Gamis A, et al.. Residual disease detected by multidimensional flow cytometry signifies high relapse risk in patients with de novo acute myeloid leukemia: A report from Children's Oncology Group. Blood 2012; 120:1581–1588.
2. Gamis A, Alonzo T, Meshinchi S, Sung L, Gerbing R, Raimondi S, Hirsch B, Kahwash S, Heerma-McKenney A, Winter L, et al. Gemtuzumab ozogamicin in children and adolescents with de novo acute myeloid leukemia improves event-free survival by reducing relapse risk: Results from the randomized phase III children's oncology group trial AAML0531. J Clin Oncol 2014; 32:3021–3032.
3. Sievers E, Lange B, Alonzo T, Gerbing R, Bernstein I, Smith F, Arceci R, Woods W, Loken M. Immunophenotypic evidence of leukemia after induction therapy predicts relapse: Results from a prospective children's cancer group study of 252 patients with acute myeloid leukemia. Blood 2003; 101:3398–3406.
4. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2015.
5. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.6-7. 2015. Available at: http://CRAN.R-project.org/package=e1071.
6. Wang J, Neskovic P, Cooper L. Training data selection for support vector machines. Adv Nat Comput 2005; 3610:554–564.
7. Tsang I, Kwok T, Cheung P. Core vector machines: Fast SVM training on very large data sets. J Mach Learn Res 2005; 6:363–392.
8. Golub G, Heath M, Wahba G. Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics 1979; 21:215–223.
9. Matthews B. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 1975; 405:442–451.
10. Terstappen L, Huang S, Safford M, Lansdorp P, Loken M. Sequential generations of hematopoietic colonies derived from single non-lineage committed CD34+CD38- progenitor cells. Blood 1991; 77:1218–1227.
11. Terstappen L, Loken M, Huand S, Olweus J, Lund-Johansen F. Phenotypic Characterization of the Hematopoietic Stem Cell. Becton Dickinson and Company, assignee. US Patent 5,840,580A 1998.
12. Hurwitz C, Loken M, Graham M, Karp J, Borowitz M, Pullen D, Civin C. Asynchronous antigen expression in B lineage acute lymphoblastic leukemia. Blood 1988; 72: 299–307.
13. Terstappen L, Loken M. Myeloid cell differentiation in normal bone marrow and acute myeloid leukemia assessed by multi-dimensional flow cytometry. Anal Cell Path 1990; 2:229–240.
14. Terstappen L, Loken M. Multi-dimensional flow cytometric characterization of myeloid maturation in normal bone marrow and acute myeloid leukemia. In: Burger G, Oberholzer M, Vooijs G, editors. Advances in Analytical Cellular Pathology. Amsterdam, The Netherlands: Excerpta Medica; 1990: 209–210.
15. Aghaeepour N, Finak G, The FlowCAP Consortium, the DREAM Consortium, Hoos H, Mosmann T, Brinkman R, Gottardo R, Scheuermann R. Critical assessment of automated flow cytometry data analysis techniques. Nat Methods 2013; 10:228–238.
16. Murphy R. Automated identification of subpopulations in flow cytometric list mode data using clustering analysis. Cytometry 1985; 6:302–309.
17. Aghaeepour N, Nikolic R, Hoos H, Brinkman R. Rapid cell population identification in flow cytometry data. Cytometry Part A 2011; 79A:6–13.
18. Ge Y, Sealfon S. flowPeaks: A fast unsupervised clustering for flow cytometry data via K-means and density peak finding. Bioinformatics 2012; 28:2052–2058.
19. Bierre P, Thiel D, Mickaels R. Cellular multiparameter cluster-analysis using hierarchical attractors. FASEB J 1994; 8:A1000.
20. Frankel D, Frankel S, Binder B, Vogt R. Application of neural networks to flow cytometry data analysis and real-time cell classification. Cytometry 1996; 23:290–302.
21. Malek M, Taghiyar M, Chong l, Finak G, Gottardo R, Brinkman R. flowDensity: Reproducing manual gating of flow cytometry data by automated density-based cell population identification. Bioinformatics 2015; 31:606–607.
22. Bashashati A, Brinkman R. A survey of flow cytometry data analysis methods. Adv Bioinform 2009; 584603
23. Naim I, Datta S, Rebhahn J, Cavenaugh J, Mosmann T, Sharma G. SWIFT-scalable clustering for automated identification of rare populations in large, high-dimensional flow cytometry datasets, Part 1: Algorithm design. Cytometry Part A 2014; 85A:408–421. A
24. Walther G, Zimmerman N, Moore W, Parks D, Meehan S, Belitskaya I, Pan J, Herzenberg L. Automatic clustering of flow cytometry data with density-based merging. Adv Bioinform 2009; 686759
25. Zare H, Shooshtari P, Gupta A, Brinkman R. Data reduction for spectral clustering to analyze high throughput flow cytometry data. BMC Bioinform 2010; 11:403
26. Lee G, Stoolman L, Scott C. Transfer learning for auto-gating of flow cytometry data. J Mach Learn Res 2012; 27:155–166.
27. Quinn J, Fisher P, Capocasale R, Achuthanandam R, Kam M, Bugelski P, Hrebien L. A statistical pattern recognition approach for determining cellular viability and lineage phenotype in cultured cells and murine bone marrow. Cytometry Part A 2007; 71A:612–624.
28. Toedling J, Rhein P, Ratei R, Karawajew L, Spang R. Automated in-silico detection of cell populations in flow cytometry readouts and its application to leukemia disease monitoring. BMC Bioinform 2006; 7:282