### Title

Automated Analysis of Clinical Flow Cytometry Data A Chronic Lymphocytic Leukemia Illustration




### Journal

Clinics in Laboratory Medicine, 37(4)

### ISSN

0272-2712



### Authors

Scheuermann, Richard H
Bui, Jack
Wang, Huan-You
et al.



### Publication Date

2017-12-01

### DOI

10.1016/j.cll.2017.07.011

Peer reviewed

HHS Public Access

Author manuscript

*Clin Lab Med*. Author manuscript; available in PMC 2018 December 01.

Published in final edited form as:

*Clin Lab Med*. 2017 December ; 37(4): 931–944. doi:10.1016/j.cll.2017.07.011.

# Automated Analysis of Clinical Flow Cytometry Data: A Chronic Lymphocytic Leukemia (CLL) Illustration

**Richard H. Scheuermann, PhD**,

Director of La Jolla Campus and Professor, J. Craig Venter Institute, La Jolla, California, USA, and Professor, Department of Pathology, University of California, San Diego, California, USA

**Jack Bui, MD, PhD**,

Associate Professor, Department of Pathology, University of California, San Diego, La Jolla, California, USA

**Huan-You Wang, MD, PhD**, and

Professor, Department of Pathology, University of California, San Diego, La Jolla, California, USA

**Yu Qian, PhD**

Assistant Professor, J. Craig Venter Institute, La Jolla, California, USA

## SYNOPSIS

Flow cytometry is commonly used in cell-based diagnostic evaluation for blood-borne malignancies including leukemia and lymphoma. The current practice for cytometry data analysis relies on “manual gating” to identify cell subsets in complex mixtures, which is subjective, labor-intensive, and poorly reproducible. Here we review recent efforts to develop, validate, and disseminate automated computational methods and pipeline for cytometry data analysis that could help overcome the limitations of manual analysis and provide for efficient and data-driven diagnostic applications. We demonstrate the performance of an optimized computational pipeline in a pilot study of chronic lymphocytic leukemia data from our clinical diagnostic laboratory.

### Keywords

Chronic lymphocytic leukemia; minimal residual disease; cell-based diagnostics; automated gating; cluster analysis; flow cytometry; FLOCK

CORRESPONDING AUTHOR: Richard H. Scheuermann, rscheuermann@jcvi.org, +1-858-200-1876, J. Craig Venter Institute, 4120 Capricorn Lane, La Jolla, CA, 92037, USA.

Richard H. Scheuermann, rscheuermann@jcvi.org, +1-858-200-1876, J. Craig Venter Institute, 4120 Capricorn Lane, La Jolla, CA, 92037, USA.

Jack Bui, jbui@mail.ucsd.edu, +1-858-534-3890, Department of Pathology, University of California, San Diego, Biomedical Sciences Building Room 1028, 9500 Gilman Drive, La Jolla, CA, 92093-0612, USA.

Huan-You Wang, huw003@ucsd.edu, +1-858-822-2538, Department of Pathology, School of Medicine, University of California, San Diego, 3855 Health Sciences Drive, La Jolla, CA, 92093-0987, USA.

Yu Qian, mqian@jcvi.org, +1-858-200-1837, J. Craig Venter Institute, 4120 Capricorn Lane, La Jolla, CA, 92037, USA.



DISCLOSURE STATEMENT

The authors do not have a conflict of interest to claim.

## Background

Cells of the peripheral blood can serve as sentinels of the physiological and pathological state of an organism. Normal vascular recirculation and extravascular migration allows these cells to touch every part of the body. The number and phenotype of blood cells are also constantly influenced by the sea of cytokines, growth factors, hormones and other small molecules they are bathed in, such that the cellular constituents of blood also reflect its molecular constituents. Thus, a detailed, accurate and consistent representation of the qualitative and quantitative properties of blood cells can be used to understand the mechanistic underpinnings of disease and to identify potential biomarkers of disease diagnosis, prognosis and therapeutic response.

In the 1960's, Alexander Vastem recognized that the enumeration of blood cell types (Complete Blood Count, CBC) could be used diagnostically as evidence for certain kinds of infections and malignancies. Although the CBC has emerged as a critical laboratory assay, it cannot detect, with any detail, the phenotypes of the enumerated blood cells. Flow cytometry (FCM) represents an advancement in the analysis and characterization of blood cells, enabling researchers and clinicians to identify the surface antigen expression of blood cells using fluorochrome-conjugated antibodies, multiple lasers to provide specific excitation of fluorochromes, and several detectors to quantitate the emitted fluorescent signal. Indeed, using a simple cocktail of antibodies each conjugated to defined fluorochromes with known emission ranges, an investigator can identify specific lymphoid and myeloid populations in peripheral blood, as well as the expression of antigenic determinants, some of which can be diagnostic or prognostic.

In 2007, Davis et al. published recommendations from an expert panel regarding the medical indications for performing FCM-based diagnostic testing, including cytopenias, elevated leukocyte counts, atypical cells in bodily fluids, plasmacytosis or monoclocal gammopathy, and organomegaly.[1] Based on the individual indications, specific staining panels are selected to target the likely cell suspects. For example, at the University of California, San Diego (UCSD) Center for Advanced Laboratory Medicine (CALM), 10 different tube/panels are in routine use to aid in the diagnosis of acute and chronic leukemia. Table 1 shows an example of three such tubes, the fluorescent channels used, and the antigens detected.

Although the use of these panels has demonstrated accurate and clinically actionable diagnosis and classification of hematolymphoid neoplasms, there is room for improvement, given the genetic and phenotypic diversity of blood cell diseases. For example, while flow cytometry can accurate diagnose acute promyelocytic leukemia (APL) based on multiple surface antigens, molecular subclassification of APL using cytogenetic detection of the t(15;17) translocation is diagnostic for a subtype that responds to all transretinoic acid. The fact that this leukemia appears to be derived from a distinct population of immature granulocytes suggests that this important subclassification could potentially be achieved using more complex staining panels alone without the need for cytogenetics. However, despite the routine use of complex, i.e. > 8-color, staining panels in research laboratories in recent years, high complexity flow cytometry panels have not been incorporated into routine

use in the clinical flow cytometry laboratory. In part, the lack of consensus of diagnostically relevant antigens and the inability to standardize analysis templates have hindered progress in clinical laboratories.

The current approach for identification of diagnostic cell populations from cytometry data is based on manual gating analysis, which is subjective, labor-intensive, and poorly reproducible, especially when dealing with higher-dimensional complex datasets. Over the last several years, our group and others have developed a suite of computational tools for the processing and analysis of cytometry data (reviewed previously).[2,3] These include computational tools and informatics resources for: i) data pre-processing to manipulate file formats, identify outlier events and samples, and adjust for batch effects; ii) automated gating for supervised and unsupervised cell population identification; iii) post-processing for cell-based biomarker identification, feature extraction, and data visualization; iv) data standards and database resources for cytometry data dissemination. The majority of the computational tools are made available as open source software with unrestrictive licensing, and the majority of resources are publicly available.

In order to assess the performance of different computational approaches for cytometry data analysis and to provide guidance to end-users on their use, an international consortium - FlowCAP (Flow Cytometry Critical Assessment of Population Identification Methods, http://flowcap.flowsite.org) - was assembled to assess and compare computational methods through a series of analysis challenges. FlowCAP-I tested whether automated algorithms could reproduce expert manual gating; FlowCAP-II focused on sample classification.[4] FlowCAP-I results showed that several algorithms were able to achieve similar results to expert manual analysis for five testing datasets from GvHD (graft-versus-host disease), DLBCL (diffuse large B-cell lymphoma), HSCT (hematopoietic stem cell transplant), WNV (symptomatic West Nile virus infection), and normal donors (ND). As an example of their performance, multiple methods achieved F-measure scores for event-level classification accuracy of >0.9 in comparison with expert manual gating for DLBCL in human peripheral blood mononuclear cell (PBMC) samples. In FlowCAP-II, several classification algorithms were able to effectively classify acute myeloid leukemia (AML) from non-AML samples with 100% accuracy. These computational methods have now been found to effectively identify novel cell-based biomarkers in a variety of research settings.

A small number of studies have now begun to evaluate the use of these methods in diagnostic settings. In 2012, Bashashati et al. used the flowClust method to identify a subtype of diffuse large B-cell lymphoma DLBCL) in which the lymphoma cells showed a unique high side scatter pattern reflecting internal cellular complexity.[5] Importantly, the subset of patients carrying this DLBCL subtype showed signfifcantly inferior overall and progression-free survival, suggesting that this lymphoma phenotype might serve as a useful biomarker to identify DLBCL patients at high risk for relapse. Zare et al. used the SamSPECTRAL and FeaLect methods to identify features in flow cytometry data that are useful in distinguishing between mantle cell lymphoma (MCL) and small lymphocytic lymphoma (SLL), and showed that the classification accuracy increased, for MCL from 64% to 100% and for SLL from 69% to 97%, in comparison with standard of care manual analysis of the diagnostic flow cytometry data.[6] Craig et al. used the flowType and

RchyOptimyx methods and found that a CD10+CD38− B cell population showed significantly different proportions in germinal center B-cell lymphoma versus germinal center hyperplasia, however the absolute proportion in any given patient was not considered specific enough to be used in a diagnostic setting.[7] Dorfman et al. used the FLOCK methods to identify discete mast cell populations in the majority of patients with systemic mastocytosis, with a sensitivity of 75% and a specificity of 86%.[8] Dorfman et al. also used FLOCK to identify discrete plasma cell populations in the bone marrow of patients with plasma cell neoplasms with a sensitivity of 97%, compared with only 81% for standard flow cytometric analysis.[9] Finally, Levine et al. used the PhenoGraph method, which algorithmically defines phenotypes in high-dimensional single-cell data, to stratify acute myeloid leukemia (AML) patients into subtypes with prognostic differences based on the activation of signaling pathways derived using mass cytometry data.[10] These examples clearly indicate the promise of using computational approaches of flow cytometry data in a clinical diagnostics setting.

As the complexity of cytometry data has increased in recent years, it has been recognized by the translational research community that computational support is becoming essential for accurate single cell phenotyping. With the advent of clinical genomics applications, the diagnostic lab environment will need to become familiar with the validation and use of computational pipelines in their diagnostic workflows. In the case of cytometry, the integration of computational methods into the laboratory workflow would allow for the use of more complex staining panels for traditional diagnostic tests (i.e. leukemia and lymphomas) where objective analysis methods could provide for more consistent disease characterization and subtype identification with therapeutic and prognostic implications. In addition, the ability to consistently manage and interpret data from complex staining panels would also allow for the validation of diagnostic tests for other diseases, including the monitoring of circulating tumors cells for solid tumors and the diagnosis and monitoring of other immune-mediated diseases (e.g. allergy, asthma and autoimmune disease). And while diagnostic flow analysis has historically focused on the identification of neoplastic cells within the patient sample, a comprehensive elucidation of both the neoplastic ***and*** normal cellular components may also help identify candidates for cancer immunotherapy and contribute to the goals of precision medicine.

## Illustration

In order to illustrate the potential use of these computational methods in a clinical diagnostic setting, we present the results of our preliminary attempts to optimize the application of a selected set of computational methods for the automated identification of chronic lymphocytic leukemia (CLL) cells stained with a newly implemented 10 marker panel. The goal is to illustrate the process of computational pipeline optimization, to highlight the promise that these methods bring for more objective identification of CLL cells in patient samples, and to explore their potential utility for monitoring of minimal residual disease (MRD).

The computational data processing and analysis workflow we have implemented for the CLL FCM data analysis (Figure 1A) consists of the following steps:

### Data preparation

20 FCS 3.0 files from peripheral blood samples of 20 subjects were received from UCSD clinical labs for CLL diagnostic evaluation. 11 subjects received a diagnosis of CLL; 5 subjects were reported as having no evidence of CLL (no-CLL); 4 subjects were evaluated for the presence of MRD following therapy. Protected health information (PHI) was scrubbed from the file headers and pseudo file names are used in the data analysis. Except for the corresponding subject disease status (CLL, non-CLL, MRD), no other clinical data about the subjects is disclosed. The reagents used in the 10 color CLL panel are: CD45-FITC, CD22-PE, CD5-PerCP55, CD19-PECy7, CD79b-APC, CD23-APC-R700, CD81-APC-H7, CD10-BV421, CD43-BV510, CD3-BV605. Cells were stained according to our standard protocol, acquisition was performed on a BD FACSCanto 10-color instrument, and manual analysis was done using FCS Express software (DeNovo).

### Logicle Transformation

The second step in our workflow is to apply FCSTrans[11] to convert the binary FCS files, compensate them using the compensation matrices in the file headers, and transform the cellular marker expression values for optimizing the segregation of cell populations for both visualization and data analysis purposes. FCSTrans reproduces the logicle transformation procedure used in the FlowJo™ software (TreeStar, Inc.) and generates consistent displays and transformed values.[11] The output of FCSTrans for each FCS file is a data matrix with each column a parameter measured in the FCM experiment and each row a cellular event.

### Prefiltering

While unsupervised data clustering methods can be applied to the whole data file for identification of cell populations, they usually generate a large number of data clusters as the number of parameters measured in an FCM experiment keeps increasing. Interpreting and annotating these data clusters is labor intensive. Some of the data clusters were found in debris, dead cells, and doublets. Including a data prefiltering step before the cluster analysis step allows the computational pipeline to focus on the cells of interest. Depending on the data clustering method used in the pipeline, the prefiltering step also helps the identification of small cell subsets, reduces the run time of the pipeline, as well as allows the population summary statistics (e.g., proportions) to be calculated based on the correct parent populations.

A data pre-filtering method we recently developed, called DAFi (Directed Automated Filtering and Identification of Cell Populations), is applied to identify the CLL cells from the input FCS files. The steps of DAFi are illustrated in the Figure 1B. In the first step of DAFi, the unsupervised FLOCK clustering method[12] is applied to partition the data into many small data clusters. Compared with unsupervised learning methods, one of the major features of DAFi is that it requires a manual gating strategy from the user. For the identification of the CLL cells, the 2D coordinates in user manual gating based on FSC/SSC, CD45, CD3, CD5, and CD19, are combined into a hyper-rectangle for the computational identification of the CLL cells in the high-dimensional space, through merging the FLOCK-identified data clusters with centroids located within the hyper-rectangle. Finally, the filtered data are visualized in 2D plots for visual examination. Unlike supervised approaches, this

way preserves the data-driven characteristics of unsupervised learning, which identifies data clusters using all data dimensions simultaneously without presumptions. On the other hand, through the use of a manual gating definition and plotting high-dimensional data clusters on 2D plots, DAFi facilitates the interpretation of the data analysis results in a way that the user predefines.

### Unsupervised identification of CLL subsets using FLOCK

The FLOCK-based computational pipeline[12] was applied to identify subsets in the CLL cells through unsupervised cluster analysis. In order to map the FLOCK-identified data clusters across the 20 files, we first normalized the CLL data across the files with the 0–1 min-max normalization method, and then merged the data together into a single file for FLOCK analysis. FLOCK returned a cluster membership file for each event in the merged data. We implemented a script to separate the events back into each original file together with their cluster membership. Thus, the same set of cluster IDs is used across the 20 files to indicate the mapping of these data clusters.

The FLOCK pipeline also outputs summary statistics tables with each row a data file (corresponding to one experiment sample) and each column a cell population identified in the sample. The contents of these tables include percentage values of cell subsets, mean fluorescence intensities (MFI) on each marker, cell phenotype profiles, and other predefined statistical analysis results. Then p-values based on statistical hypothesis testing can be calculated (e.g., through a non-parametric Wilcoxon rank sum test) to indicate if there is a significant difference between the subject groups.

## Results

The goal of this study was to determine if a computational pipeline could be developed for the identification of diagnostic cell populations in CLL patient samples using FCM data. We chose to use a new method we recently developed for directed unsupervised population identification called DAFi (unpublished). The DAFi approach uses “direction” from prior knowledge about the cell populations of interest that it typically used to drive manual gating analysis. Figure 2 shows a typical manual gating hierarchy used in the CALM diagnostic lab to identify CLL cells from FCM data with their new 10-color panel. A plot of Time vs SSC-A is used to determine if there are any instrument anomalies occurring during data acquisition; a plot of FSC-A vs FSC-H is used to gate on singlet cells with FSC-A = FSC-H; SSC-H vs CD45 is used to gate on leukocytes; FSC-A vs SSC-A is used to gate on lymphocytes; CD5 vs CD19 is used to gate on putative CLL cells (CD5+CD19+), normal B cells (CD5−CD19+), and normal T cells (CD5+CD19−). DAFi was initially configured to produce a hyper-rectangle to recapitulate this manual gating hierarchy.

Cell events retained by these initial DAFi filters for a single CLL blood sample are shown in Figure 3A, and for a composite of all 20 samples in the data set following unsupervised clustering using the FLOCK method in Figure 3B. The results from individual samples show that all three cell populations show natural distributions with the expected marker expression levels. The composite results show that although the general expected marker expression levels are observed for all samples, there also appears to be some heterogeneity in their

absolute fluorescence levels, especially in terms of CD5 expression in normal T cells (bottom middle) and both CD5 and CD19 expression in CLL cells (bottom right). The heterogeneity in CD5 expression can also be seen in plots from individual samples (Figure 4), with some samples showing relatively high levels of CD5 expression (bright - br), some showing intermediate levels (dim - di), and some showing a mixture of both (di/br).

In order to verify that the initial DAFi filters was specifically isolating CLL cells, we compared cells retained by the initial DAFi filters from both CLL and non-CLL patients. Unfortunately, cell retained by these initial DAFi filters were observed in three of the five non-CLL samples evaluated in the CD5 vs CD19 dot plots (Figure 5A). To investigate this discrepancy further, we examined the expression pattern of other cell surface markers that could be useful for improving the specificity of DAFi filtering and found that the majority of putative CLL cells from CLL patients were CD10− and CD79b dim, whereas the seemingly “CLL” cells retained by the initial DAFi filters in non-CLL patients were CD79b bright and CD10+ or −.

Based on these finding, we adjusted the configuration of DAFi to include additional filters for CD79b dim and CD10−. While the use of these additional filtering criteria had little effect on CLL cell population identification in CLL samples (Figure 6A), this revised definition eliminated the spurious retention of cell events in non-CLL samples (Figure 6B).

Finally, we examined the ability of the revised DAFi filtering to detect CLL cells in the setting of MRD, where distinguishing between CLL cells and normal B cells can be challenging. By including the CD10− and CD79b dim phenotypes in the revised CLL definition and DAFi filtering configuration, CLL cells could be clearly identified in two of the four MRD samples (Figure 7).

## Discussion and Conclusions

In this illustration, we examined the ability of an automated computational pipeline based on the DAFi method to identify CLL cells in multidimensional FCM data. Using an initial definition based solely on the co-expression of CD5 and CD19, we found that the cell populations identified were not specific to CLL patients, but that the populations identified in CLL versus non-CLL patients differed in their expression of CD10 and CD79b. Indeed, this pattern of CD10 and CD79b expression in CLL cells has been reported previously.[7,13] By changing the cell population definition and configuration of DAFi to retain cells that are CD5+CD19+CD10−CD79bdim, cells that are specific to the CLL patient cohort could be automatically identified without manual gating.

By identifying CLL cells in multidimensional space simultaneously, DAFi filtering was robust to slight changes in marker expression. This robustness was especially important for CD5 expression, which showed considerable variability between CLL patients, ranging from bright to dim. In addition, the use of a precise multidimensional definition allowed for the sensitive identification of CLL cells in the setting of MRD.

We also identified considerable heterogeneity in CD5 expression in normal T cells (Figure 3B). It should be noted that this population of normal T cells is not typically the focus of

manual analysis and is considered “irrelevant” in the current diagnosis of CLL; however, the T cell phenotype of CLL patients could stratify responses to immune therapy and could find utility in the future. We postulate that an unbiased, automated analysis platform, as we have described herein, could be used to capture all of the populations in peripheral blood, thereby realizing the vision that peripheral blood samples could indeed reflect the physiology of the patient. We argue that as flow cytometry platforms become more complex and multidimensional, it becomes impossible to examine and interpret all of the data manually.

The work presented here was a pilot study to demonstrate the feasibility of applying automated computational pipelines to diagnostic data using a relatively small number of samples and a subjective evaluation of the results. Based on the positive findings from this pilot, a prospective study is underway to expand the sample size and apply objective evaluation criteria to quantitative assess the sensitivity, specificity and predictive value of this approach for CLL diagnosis in comparison with standard of care practices. The protocol being used will allow us to follow up on any discrepancies to examine any addition diagnostic testing results and patient outcome information to try and determine the correct result. By also examining the outcome of patients evaluated in the setting of MRD, we will also be able to determine if monitoring CLL cell population levels using automated computational pipelines would have prognostic utility. In the pilot study, we also obtain some evidence that distinct sub-types of CLL might exist based on the expression of the other markers (e.g. CD22 and CD81) in the 10-color staining panel used (data not shown). However, the small sample size made it difficult to draw any definitive conclusions. The prospective study should allow us to determine if the detection of distinct subtypes is reproducible and if show different response to therapy and prognostic outcome.

The ultimate goal of these efforts is to achieve adoption of automated computational pipelines for routine use in the clinical diagnostic laboratory. This goal presents a challenge in determining how best to validate computational pipelines for FCM data analysis for patient management use in a CLIA environment. Emerging experience with the application of computational methods for the processing and interpretation of next generation sequencing data for diagnostic use[14] could be used to establish the necessary computational method validation plan for FCM data.

Flow cytometry analysis of patient samples has become an indispensable component of the modern diagnostic toolkit for a variety of different hematopoietic diseases, especially leukemias, lymphomas, and myeloproliferative disorders. Advancements in cytometry instrumentation and the availability of more complex staining panel collections promise to allow for more accurate identification and monitoring of disease for improved diagnosis and prognosis of patients. However, these advances result in an increase in the complexity of data that feed into the diagnostic process. The validation and use of computational pipelines to assist in the processing and interpretation of these data will be essential to realize their true potential in the clinical laboratory.

## Acknowledgments

This work was supported by the U.S. National Institutes of Health - 1U01TR001801 and CA157885, and The Hartwell Foundation.

## References


1. Davis BH, Holden JT, Bene MC, Borowitz MJ, Braylan RC, Cornfield D, Gorczyca W, Lee R, Maiese R, Orfao A, Wells D, Wood BL, Stetler-Stevenson M. 2006 Bethesda International Consensus recommendations on the flow cytometric immunophenotypic analysis of hematolymphoid neoplasia: medical indications. Cytometry B Clin Cytom. 2007; 72(Suppl 1):S5–13. [PubMed: 17803188]
2. O'Neill K, Aghaeepour N, Spidlen J, Brinkman R. Flow cytometry bioinformatics. PLoS Comput Biol. 2013; 9(12):e1003365. [PubMed: 24363631]
3. Kvistborg P, Gouttefangeas C, Aghaeepour N, Cazaly A, Chattopadhyay PK, Chan C, Eckl J, Finak G, Hadrup SR, Maecker HT, Maurer D, Mosmann T, Qiu P, Scheuermann RH, Welters MJ, Ferrari G, Brinkman RR, Britten CM. Thinking outside the gate: single-cell assessments in multiple dimensions. Immunity. 2015 Apr 21; 42(4):591–2. DOI: 10.1016/j.immuni.2015.04.006 [PubMed: 25902473]
4. Aghaeepour N, Finak G, FlowCAP Consortium; DREAM Consortium. Hoos H, Mosmann TR, Brinkman R, Gottardo R, Scheuermann RH. Critical assessment of automated flow cytometry data analysis techniques. Nat Methods. 2013 Mar; 10(3):228–38. Epub 2013 Feb 10. Erratum in: Nat Methods. 2013 May;10(5):445. DOI: 10.1038/nmeth.2365 [PubMed: 23396282]
5. Bashashati A, Johnson NA, Khodabakhshi AH, Whiteside MD, Zare H, Scott DW, Lo K, Gottardo R, Brinkman FS, Connors JM, Slack GW, Gascoyne RD, Weng AP, Brinkman RR. B cells with high side scatter parameter by flow cytometry correlate with inferior survival in diffuse large B-cell lymphoma. Am J Clin Pathol. 2012 May; 137(5):805–14. [PubMed: 22523221]
6. Zare H, Bashashati A, Kridel R, Aghaeepour N, Haffari G, Connors JM, Gascoyne RD, Gupta A, Brinkman RR, Weng AP. Automated analysis of multidimensional flow cytometry data improves diagnostic accuracy between mantle cell lymphoma and small lymphocytic lymphoma. Am J Clin Pathol. 2012 Jan; 137(1):75–85. [PubMed: 22180480]
7. Craig FE, Foon KA. Flow cytometric immunophenotyping for hematologic neoplasms. Blood. 2008 Apr 15; 111(8):3941–67. [PubMed: 18198345]
8. Dorfman DM, LaPlante CD, Pozdnyakova O, Li B. FLOCK cluster analysis of mast cell event clustering by high-sensitivity flow cytometry predicts systemic mastocytosis. Am J Clin Pathol. 2015 Nov; 144(5):764–70. [PubMed: 26486741]
9. Dorfman DM, LaPlante CD, Li B. FLOCK cluster analysis of plasma cell flow cytometry data predicts bone marrow involvement by plasma cell neoplasia. Leuk Res. 2016 Sep.48:40–5. [PubMed: 27479652]
10. Levine JH, Simonds EF, Bendall SC, Davis KL, Amir ED, Tadmor MD, Litvin O, Fienberg HG, Jager A, Zunder ER, Finck R, Gedman AL, Radtke I, Downing JR, Pe'er D, Nolan GP. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. Cell. 2015; 162:184–197. [PubMed: 26095251]
11. Qian Y, Liu Y, Campbell J, Thomson E, Kong YM, Scheuermann RH. FCSTrans: an open source software system for FCS file conversion and data transformation. Cytometry A. 2012 May; 81(5): 353–6. [PubMed: 22431383]
12. Qian Y, Wei C, Eun-Hyung Lee F, Campbell J, Halliley J, Lee JA, Cai J, Kong YM, Sadat E, Thomson E, Dunn P, Seegmiller AC, Karandikar NJ, Tipton CM, Mosmann T, Sanz I, Scheuermann RH. Elucidation of seventeen human peripheral blood B-cell subsets and quantification of the tetanus response using a density-based method for the automated identification of cell populations in multidimensional flow cytometry data. Cytometry B Clin Cytom. 2010; 78(Suppl 1):S69–82. [PubMed: 20839340]
13. Hulkkonen J, Vilpo L, Hurme M, Vilpo J. Surface antigen expression in chronic lymphocytic leukemia: clustering analysis, interrelationships and effects of chromosomal abnormalities. Leukemia. 2002 Feb; 16(2):178–85. [PubMed: 11840283]
14. Rehm HL, Bale SJ, Bayrak-Toydemir P, Berg JS, Brown KK, Deignan JL, Friez MJ, Funke BH, Hegde MR, Lyon E, Working Group of the American College of Medical Genetics and Genomics Laboratory Quality Assurance Commitee. ACMG clinical laboratory standards for next-generation sequencing. Genet Med. 2013 Sep; 15(9):733–47. [PubMed: 23887774]

15. Parks DR, Roederer M, Moore WA. A new “Logicle” display method avoids deceptive effects of logarithmic scaling for low signals and compensated data. Cytometry A. 2006 Jun; 69(6):541–51. [PubMed: 16604519]

## KEY POINTS

- Traditional manual gating analysis of cytometry data cannot effectively address the scale and complexity of data generation from modern cytometry instrumentation.
- Bioinformatics investigators have developed a collection of computational methods for automated identification of cell populations from high-dimensinal flow cytometry data; a small subset of these methods have been evaluated for their use in diagnostics applications of leukemia and lymphoma with promising results.
- By applying computational pipelines to classify CLL samples from healthy controls, the pilot study reported in this article illustrates the use of these methods to determine that traditional CLL definition based on CD5 and CD19 alone can be improved by also examining the expression levels of CD10 and CD79b in an automated fashion.
- Clincial validation of these computational approaches is ongoing and essential to realize the true potential of these methods for use in the clinical diagnostic laboratory.

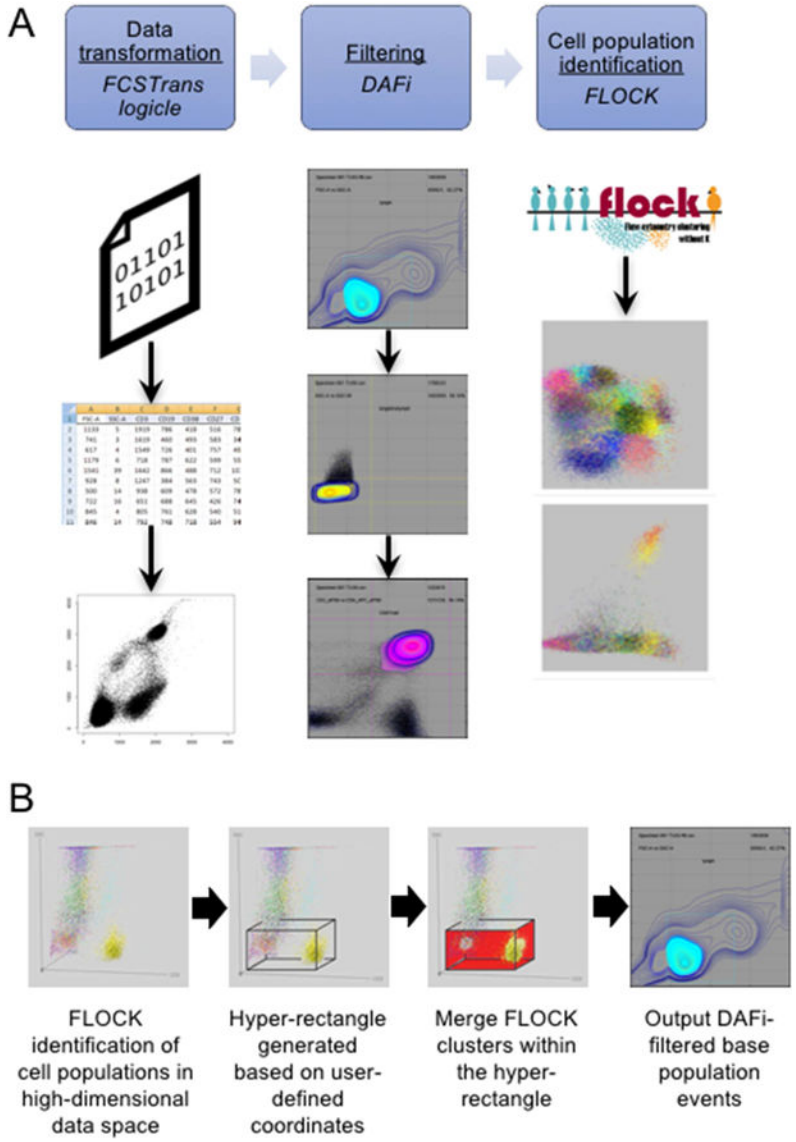


**Figure 1. Computational workflow for FCM data analysis**

A) The computational workflow used to analyze the CLL study data. Initial data transformation uses FCS file format as input to the FCSTrans algorithm[11], which applies a logicle transformation[15] to the fluorescence intensity values in order to obtain more normal distributions. The cell events are then filtered based on intensity values of selected parameters (e.g. FSC and SSC to capture lymphocyte events based on size and complexity) using the DAFi-filtering method (unpublished). Filtered events from all individual sample files are merged into a single file and cell populations identified using the FLOCK method[12] for unsupervised, density-based clustering. Cell events are then segregated back into sample specific files while retaining cell population membership annotations to facilitate cross-sample comparison. B) Details of DAFi filtering step. The DAFi filtering method begins by clustering cells into population in high dimensional space using FLOCK. A hyper-rectangle is defined by the user to define the spatial regions that contain the cell populations of interest. The cell events of cell populations with centroids located within the hyper-rectangle

of interest are then merged into a single base population for further downstream analysis using a cell population identification method, e.g. FLOCK.

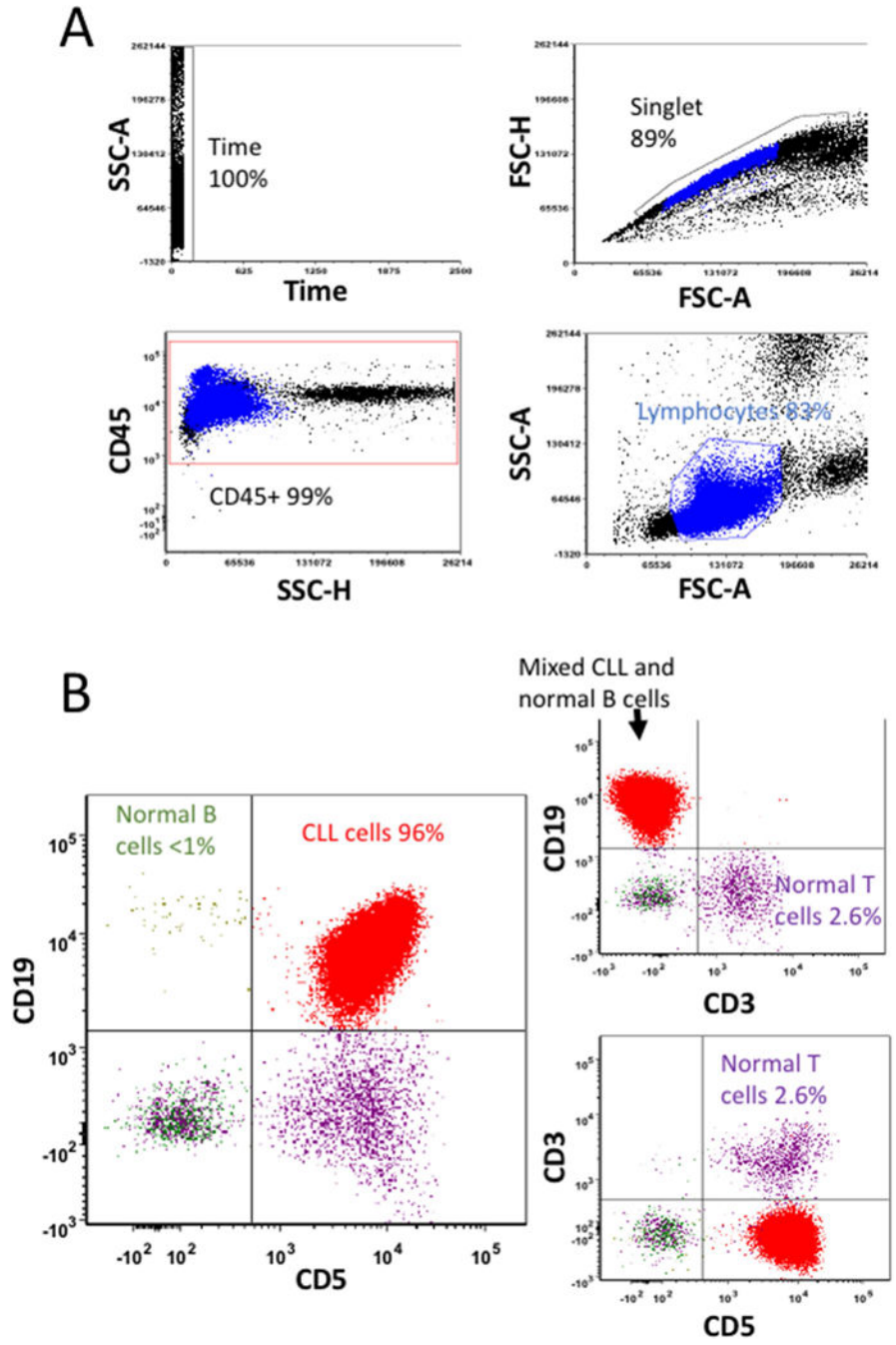


**Figure 2. Manual analysis for CLL**

A) Manual filtering of the lymphocyte base population. The singlet viable lymphocyte base population is captured using as FSC-H=FSC-A, CD45+, SSC-H low, SSC-A low, and FSC-A intermediate. B) CLL gate. CLL cells (red) are classically distinguished from normal B lymphocytes (CD19+, CD3−, CD5−; green) and T lymphocytes (CD19−, CD3+, CD5+; maroon) as being uniquely CD19+, CD3− and CD5+.

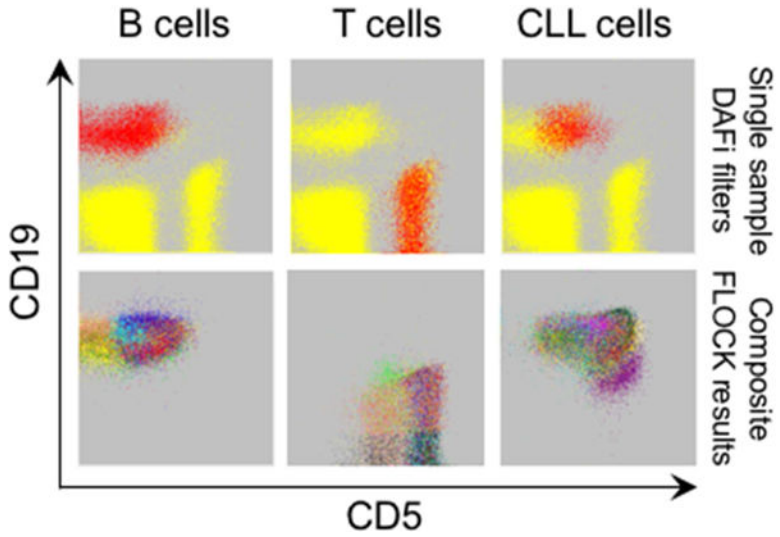


**Figure 3. DAFi filtering and FLOCK clustering of CLL FCM data**

The DAFi algorithm was configured to recapitulate the manual gating strategy for capturing normal B cells (upper left), normal T cells (upper middle) and CLL cells (upper right) as depicted in Figure 2, with the cell events obtained colored in red and cell events excluded colored in yellow, from a single CLL peripheral blood sample. The cell events for each of these three cell types from all samples were then merged and sub-population clusters identified using FLOCK-based unsupervised clustering, with cell events from each sub-population colored in a different color. Considerable heterogeneity between individual samples can be observed.

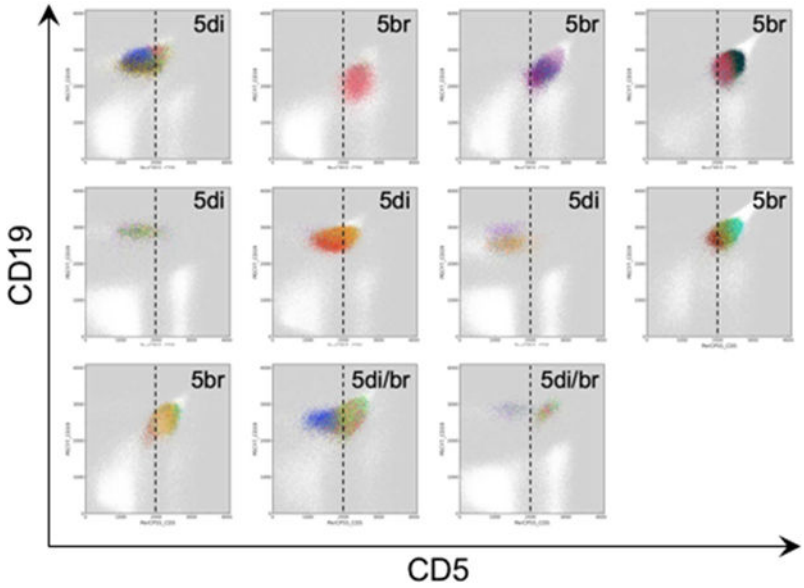


**Figure 4. Initial CLL candidate cell populations from CCL and non-CLL patients based on CD5 and CD19 expression only**

DAFi filtering of CD5+CD19+ cells (putative CLL cells) from peripheral blood samples of eleven different CLL patients are shown. The putative CLL population was further subdivided into sub-population clusters using FLOCK-based unsupervised clustering, with cell events from each putative CLL sub-population colored in a different color. Considerable variability in absolute CD5 expression between patients can be observed, with some patients carrying putative CLL cells with relatively dim (5di) and some with relatively bright (5br) expression of CD5.

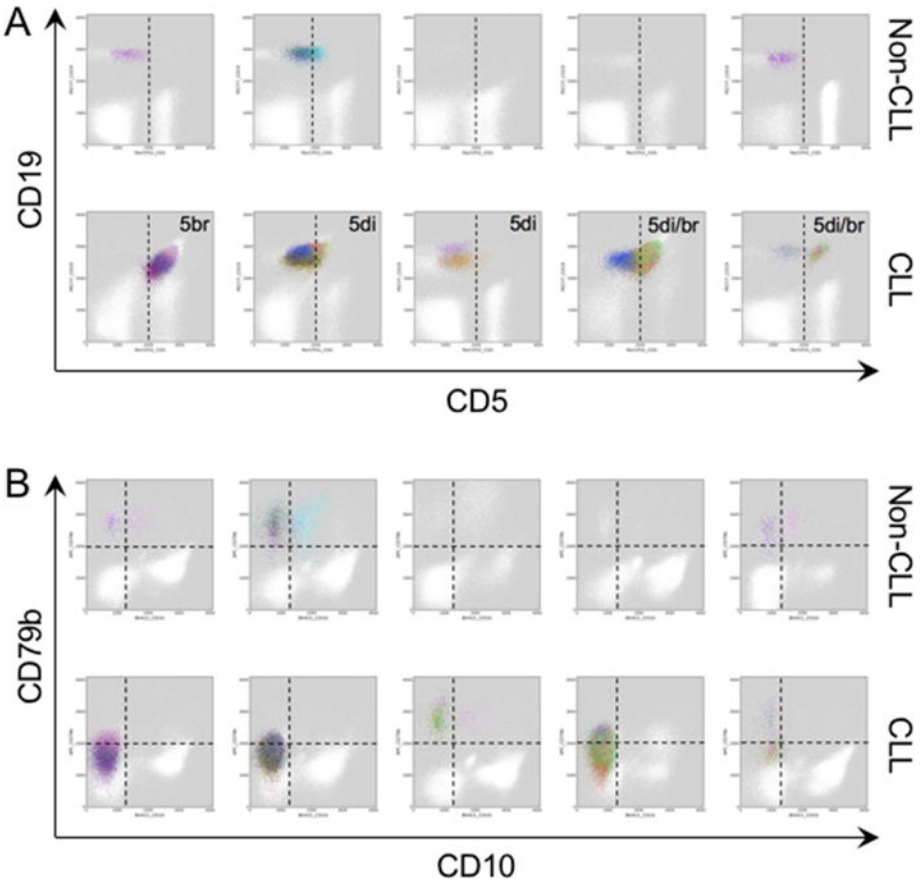


**Figure 5. Initial CLL candidate subset from non-CCL patients based on CD5 and CD19 expression only**

DAFi filtering of CD5+CD19+ cells (putative CLL cells) from peripheral blood samples of five different patients with (CLL, lower rows) and without (non-CLL, upper rows) a definitive diagnosis of CLL from a combination of clinical, cellular and molecular diagnostic tests are shown. The putative CLL population was further subdivided into sub-population clusters using FLOCK-based unsupervised clustering, with cell events from each putative CLL sub-population colored in a different color. A) CD5 versus CD19 dot plots. Based on the CD5+CD19+ CLL definition, cell events captured using DAFi filtering in non-CLL patients can be observed (upper row). B) CD10 versus CD79b dot plots. However, and examination of the expression of the expression of CD10 and CD79b in the CD5+CD19+ population reveals that these cells in CLL patients generally lack expression of CD10 and exhibit dim expression of CD79b (lower row). Whereas in non-CLL patients the CD5+CD19+ cell express high levels of CD79b with or without CD10 expression (upper row). These results suggest that a revised CLL definition of CD5+CD19+CD10−CD79dim could be more specific for the identification of authentic CLL cells.

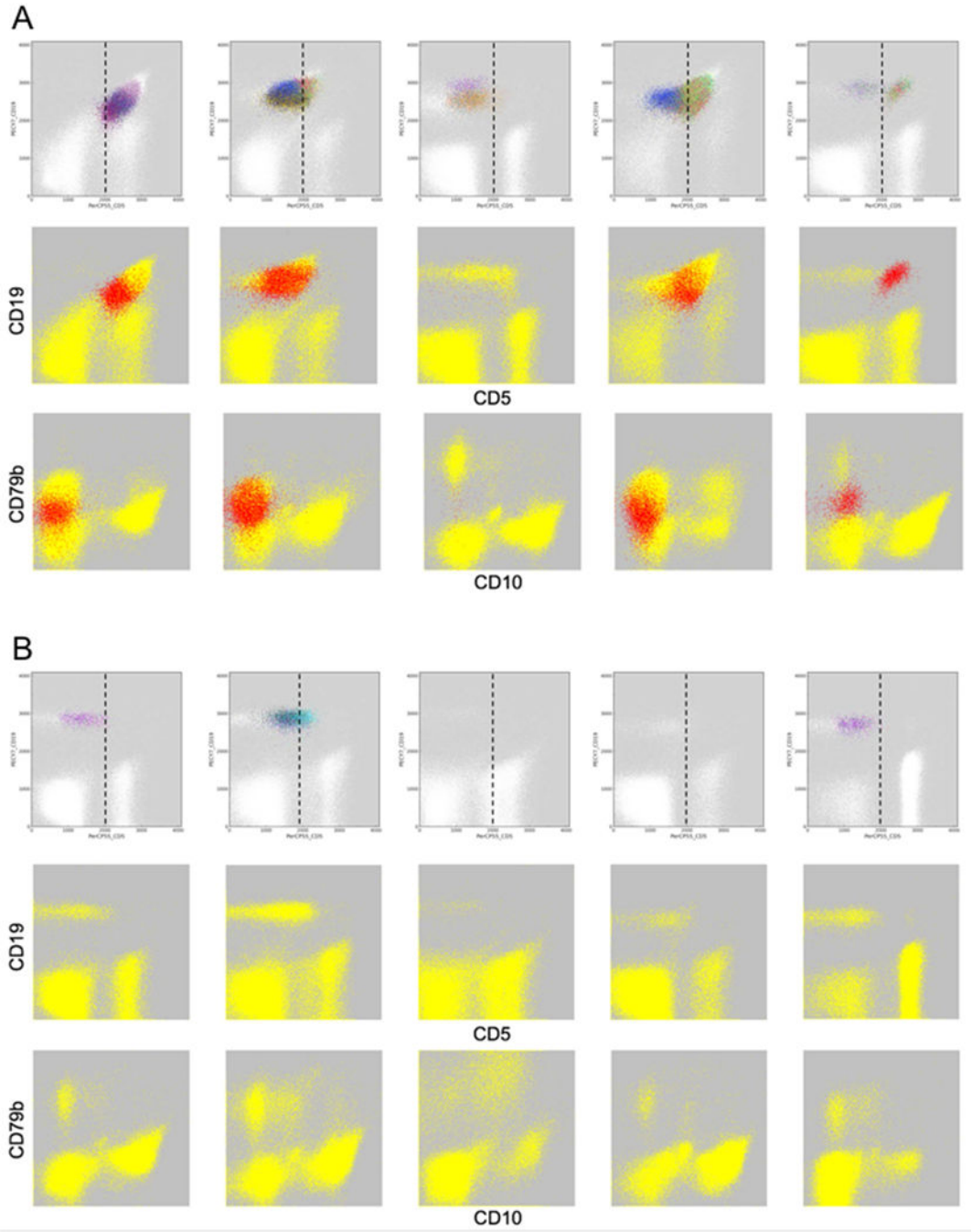


**Figure 6. Detection of CLL cells based on the revised CD5+CD19+CD10−CD79bdim CLL definition**

The analysis presented in Figure 5 was repeated with the addition of additional DAFi filtering criteria to further filter based on the lack of CD10 expression and dim CD79b expression. Each of the top rows in A and B show the results using the previous definition based on CD5+CD19+ only. The bottom two rows in each section show the DAFi filtering results using the revised definition with the putative CLL cells colored in red. A) Samples from five different CLL patients. B) Samples from five different non-CLL patients.

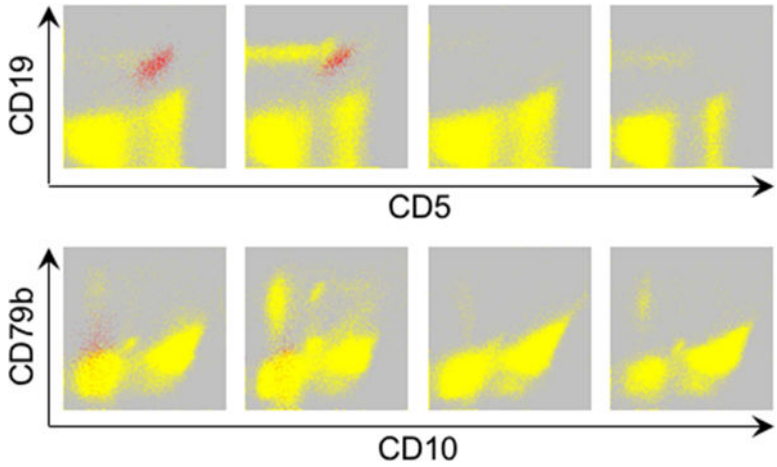


**Figure 7. Detection of CLL cells based on the revised CD5+CD19+CD10−CD79bdim definition in the setting of minimal residual disease**

Peripheral blood samples from four different CLL patients in clinical remission following treatment were analyzed using DAFi filtering based on the revised CD5+CD19+CD10-CD79bdim CLL definition, with the CLL cells colored in red. Minimal residual disease is evident in two of the four patient peripheral blood samples.

**Table 1**

Diagnosis of acute and chronic leukemia

| Acute Myeloid Leukemia (AML) Panel | | | | | Chronic Lymphocytic Leukemia (CLL) Panel | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **FL1** | **FL2** | **FL3** | **FL4** | | **FL1** | **FL2** | **FL3** | **FL4** |
| Tube 1 | CD15 | CD33 | CD45 | CD34 | Tube 1 | CD45 | CD5 | CD3 | CD19 |
| Tube 2 | CD2 | CD117 | CD45 | CD34 | Tube 2 | CD43 | CD79a | CD5 | CD19 |
| Tube 3 | HLA-DR | CD7 | CD13 | CD34 | Tube 3 | CD20 | CD38 | CD5 | CD19 |