High-Speed Automatic Characterization of Rare Events in Flow Cytometric Data

Yuan Qi[*1,2], Youhan Fang[1], David R. Sinclair[3,4], Shangqin Guo[5], Meritxell Alberich-Jorda[6], Jun Lu[7], Daniel G. Tenen[8,9], Michael G. Kharas[10], and Saumyadipta Pyne[*3,11]

[1]Department of Computer Science, Purdue University, West Lafayette, IN, USA
[2]Department of Statistics, Purdue University, West Lafayette, IN, USA
[3]Public Health Dynamics Laboratory, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, USA
[4]Department of Health Policy and Management, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, USA
[5]Department of Cell Biology, Yale University School of Medicine, New Haven, CT, USA
[6]Institute Molecular Genetics of the AS CR, Prague, Czech Republic
[7]Department of Genetics, Yale University School of Medicine, and Yale Stem Cell Center, New Haven, CT, USA
[8]Center for Life Sciences, Harvard Medical School, and Harvard Stem Cell Institute, Boston, MA, USA
[9]Cancer Science Institute, National University of Singapore, Singapore
[10]Molecular Pharmacology Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[11]Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, USA
[*]Corresponding authors: spyne@pitt.edu, alanqi@cs.purdue.edu

Abstract

A new computational framework for FLow cytometric Analysis of Rare Events (FLARE) has been developed specifically for fast and automatic identification of rare cell populations in very large samples generated by platforms like multi-parametric flow cytometry. Using a hierarchical Bayesian model and information-sharing via parallel computation, FLARE rapidly explores the high-dimensional marker-space to detect highly rare populations that are consistent across multiple samples. Further it can zoom into specified regions of interest to extract subpopulations with desired precision.

Introduction

Studies focusing on rare cell populations are becoming increasingly common owing to technological advances such as high-speed, multi-parametric flow cytometry, and emerging biomedical applications like stem cell therapy, and single cell analysis. Researchers in fields such as hematology, cancer, immunology, pathology, stem cell biology, and regenerative medicine, have focused on many interesting, yet relatively rare, populations of cells in blood and other tissues and systems that have important biomedical functions and characteristics, e.g., long-term hematopoietic stem cells. Methods for accurate detection or automated isolation of rare therapy-resistant cells in tumors with stem cell like properties, tumor cells circulating in blood, or regulatory T cells, can have profound influence on basic and clinical research. Platforms like multi-color

flow cytometry, in conjunction with the development of rich panels of markers and antibodies, have been used to establish signatures for various rare cellular species and lineages in terms of the expressed surface and intracellular marker proteins [1]. Advances in mass cytometry have promised the ability to determine 50-100 features per cell [2, 3]. To address such increasingly multi-parametric and multiplexed immunoprofiling of each cell, studies have demonstrated the critical and timely need for systematic and automated multivariate analysis and visualization suitable for high-dimensional data [4, 5, 6]. As the number of potential sequences of marker-combinations continues to grow exponentially (with the number of markers), a thorough search for rare events in high-dimensional marker-space clearly gets difficult with the more subjective and painstaking approach of traditional manual gating [7].

Analytically, a population of cells having similar, characteristic expression of k (> 1) markers can be measured as events with similar fluorescence intensities, i.e., as a cluster of points located closely in k-dimensional marker-space [4]. However traditional clustering approaches may not be adequate for identification of rare populations for several technical reasons. The new data are not only high-dimensional (i.e., involving multi-parametric or multiplexed staining panels) but simultaneously are also high-resolution (single cell level) and considerably high-throughput (hundreds of thousands of cells per sample) by design. Typically, therefore, if a population of interest is rare and consists of, say, fewer than 1% or 0.1% of the total number of cells in a given sample, then for reliable detection of such a population, it is common to use a sample size (N) in the order of $10^5 - 10^6$ cells, each measured as a k-dimensional point. Thus a large cytometric sample can present a "needle in a haystack" scenario for the search of any rare population therein, resulting either in inefficient coverage of the k-dimensional marker-space (the volume of which increasing exponentially with k), or detection of a number of spurious small populations (often outliers of larger, noisy populations). In general, clustering methods like k-means or hierarchical clustering use some measure of distance between every pair of points to determine their closeness for clustering assignment. While effective for clustering, a few thousands genes in transcriptomic datasets, clearly such quadratic-time $O(N^2)$ approaches would be computationally inadequate for searching complex cytometric datasets with much larger N.

Another practical challenge stems from biological or technological sources of inter-sample variation such as single cell level heterogeneity, individual subjects, different time-points and conditions, or platform noise, which make consistent identification of particularly the rarer populations difficult. Moreover, as cells undergo state transitions, for instance during differentiation, the corresponding changes in marker-expressions result in hierarchies of inter-connected clusters with complex high-dimensional structures that present unique data modeling challenges for computational analysis [5, 8]. Therefore, we developed FLARE as a new computational framework that can simultaneously meet the somewhat conflicting requirements of (a) high speed, (b) high precision, and (c) robust data modeling.

Results and discussion

We developed a new computational framework FLARE for FLow cytometric Analysis of Rare Events (although it is applicable to any platform that generates per cell multi-marker data). FLARE is based on a hierarchical Bayesian model (not to be confused with common hierarchical clustering, which usually is not model-based), and employs parallel computation for its high-speed high-precision analysis. The Bayesian model (Fig. 1) of FLARE allows implementation of several distinct features to specifically address the challenges mentioned above. For consistent identification of a particular rare population C, FLARE's model parameters allow information about C to be shared across different samples. (In parallel computing, we implemented this via communication among nodes each of which analyzed a distinct sample.) The strategy builds repeated inter-sample consensus on the existence of C (or lack thereof), thus guarding against unsupervised detection of possibly numerous spurious small populations. Consequently, the model estimation is highly robust against high inter-sample variation and platform noise, which otherwise are known to affect the reproducibility or the quality of match between analogous populations across samples and replicates [9]. We found FLARE's information-sharing feature to be especially useful for rare cell populations (e.g., blue clusters in Fig. 1) which contain very few cells since it effectively increases the number of observations for estimation. Second, the estimation ambiguity (between the red and the green clusters in Sample 1, corresponding to Subject 1, in Fig. 1) is reduced since the information about a population (e.g., the red cluster of Sample 2, corresponding to Subject 2, in Fig. 1) can guide the estimation of its counterpart across samples (i.e., the red cluster of Sample 1 in Fig. 1) in FLARE's joint model. Third, the joint model also allows partial consistency such that some clusters can exist in one or more samples but not necessarily in all of them. Thus without needing any additional metaclustering step [4], we can identify also those clusters (e.g., the green cluster in Fig. 1b) that exist only in certain samples, a case that is not uncommon for rare populations (such as transient subsets that appear only in certain stages of cell differentiation and are absent otherwise).

FLARE offers many distinct advantages specifically for characterization of rare populations. First, FLARE shares information across multiple samples in a hierarchical Bayesian model (Supplementary Fig. 1) to identify cell populations in all samples or in only part of samples. Unlike common clustering methods, FLARE does not need a priori specification of the optimal number of clusters in data, which gives FLARE an advantage while searching samples which may contain populations ranging from significantly big to extremely rare. Instead, FLARE automatically sparsifies a large finite mixture models so that unnecessary clusters are removed and the actual number of clusters used to fit the data is learned accordingly. In practice our model can be viewed as an efficient approximation to Dirichlet Process Mixture models. The information

sharing and automatic identification of the number of cell populations greatly boost the estimation accuracy of FLARE. Simulation results showed that FLARE achieves favorable estimation performance over alternative methods (Supplementary Fig. 3-6). Second, our model is fit with a Variational Bayes (VB) approach (see Section A of the Supplementary) which provides computationally efficient and accurate estimation of all the latent variables (i.e., the output of the model). Furthermore, FLARE is parallelized with careful consideration on workload balancing in a distributed computing environment (Supplementary Algorithm 1 and Supplementary Fig. 2). It achieves almost linear speedup given more computational nodes (Supplementary Fig. 7), making it truly scalable on large data.

Identification of rare cell subsets – while explicitly establishing their correspondence across samples – can (a) reveal, in an *unsupervised* way, the overall structure among the populations, both big and small, with respect to each other in every sample, and thereby (b) provide contextual information that helps in *supervised* zooming into chosen regions of interest in the marker-space to characterize or dissect (by sub-clustering) the rare populations with further precision. Such stepwise zooming approach is a known strength of sequential manual gating. An important advantage of FLARE's Bayesian design is that it can be made to systematically zoom into interesting regions or populations by a priori specifications. Thus FLARE can perform increasingly finer clustering using the same mathematical model, which can again match and verify the finer subpopulations across multiple samples. This allows FLARE to combine the benefits of an unsupervised clustering method with supervised analysis of manual sequential gating. We illustrate these aspects of FLARE using a multi-step analysis of a hierarchy of cell populations as observed in (i) normal hematopoiesis in mice and (ii) oncogenic progression in a mouse model.

For our first application of FLARE, we generated a 6-marker cytometric dataset to study cells from mouse bone-marrow (Supplementary Section C). In the first step, without any assumptions or human guidance, unsupervised analysis by FLARE revealed in a given set of 14 "training" samples, including multiple biological and technical replicates, a subset bearing a 6-marker signature of long-term murine hematopoietic stem cells (HSCs) (Fig. 2a-b). In the second step, using this estimated location parameter, FLARE zoomed into the corresponding region in an entirely new and much larger "test" sample (containing more than a million cells measured with the same 6 markers) to detect a very rare population (containing 0.045% of the total number of cells in the sample) with a more precise LT-HSC signature. This finding is supported by earlier analyses, using sequential 2D gating of 2 markers at any given time, according to which LT-HSCs are known to be Lineage$^-$Kit$^+$Sca$^+$CD34$^-$CD48$^-$CD150$^+$. Using parallel computational algorithm (see Section B of the Supplementary), the two steps took less than 10 minutes to finish. The consistency of the detected subsets across all 14 training samples, the small size of the final detected population and the precise marker-expressions of the cells therein all suggest how FLARE could be used for high-speed automated identification

of rare populations in cytometric data.

As a second application of FLARE we wanted to determine if we could track rare clusters during disease progression. Therefore, we utilized a model of myeloid leukemia that harbors the oncogenic fusion of PML and the retinoic acid receptor alpha (RARα). These mice succumb to a lethal acute promeylocytic leukemia (APL) that can be transplanted with increasing aggressiveness [10]. We have previously characterized the surface phenotype, which drives the APL and it closely resembles the normal promyelocyte population [11, 12]. Of note the mature granulocytes, which differentiate from the leukemic stem cell population (LSC) are unable to transplant the disease. The gating strategy for this population is challenging since these cells express low levels of lineage markers and are serially gated for CD34/c-Kit and then Gr1/ FcgRIIb (Lin^{lo} c-kit^{+} CD34^{hi} Gr^{mid} FcgRIIb^{+}). This population in a normal bone marrow is approximately 1% of the live bone marrow cells and increases to 5-6% in the leukemic mice (Fig. 3a).

By performing FLARE from live-gated cells stained for lineage, c-kit, CD34, Gr-1 and FcgRIIb we could identify 99 unique clusters that contained cells (Fig. 3b). To determine if FLARE could find the clusters that contained the LSC population, we further filtered the rows based on the clusters that had a greater than 1-fold change in proportion compared to the normal mice. This analysis then narrowed down the clusters to 44, which was further sorted by hierarchical clustering to create a heat map of the populations and their proportions allowing visualization of the hematopoietic cells that increased during leukemia progression. We identified a block of clusters that matched with the known surface phenotype of the promyelocytes and the LSCs, which we named the "Promyelocyte Signature" (Clusters 99, 57, 39, 25, 50 in the green box, Fig. 3c). We found the promyelocyte cluster to represent 1% of the live cells in the normal mice, which increased to 6-8% in the leukemic mice (Fig. 3c). FLARE successfully identified this rare population and demonstrates the utility for tracking changes within phenotypically defined populations during disease progression. Using parallel computing, this was accomplished in under 5 minutes.

In summary, the hierarchical design and distributed variational estimation allows FLARE to share information about corresponding clusters across samples, and quickly detect various populations in an unsupervised manner. In the process, it efficiently searches the high-dimensional marker-space to reveal the overall population structure. Thereupon it can zoom into regions of interest and also perform supervised analysis of subpopulations similar in principle to manual gating except FLARE does it in high-dimensions and with mathematical rigor. In the future we look forward to embedding this step into FACS systems for real-time sorting of the desired cells. Thus multi-parametric population signatures determined by FLARE are quantitative and precise, which helps to standardize definitions of specific cellular species, allow objective extraction, and facilitate reproducible cytometric analysis. Finally, the distributed estimation algorithm in FLARE is currently implemented using Message Passing Interface (MPI) and can be readily adapted to

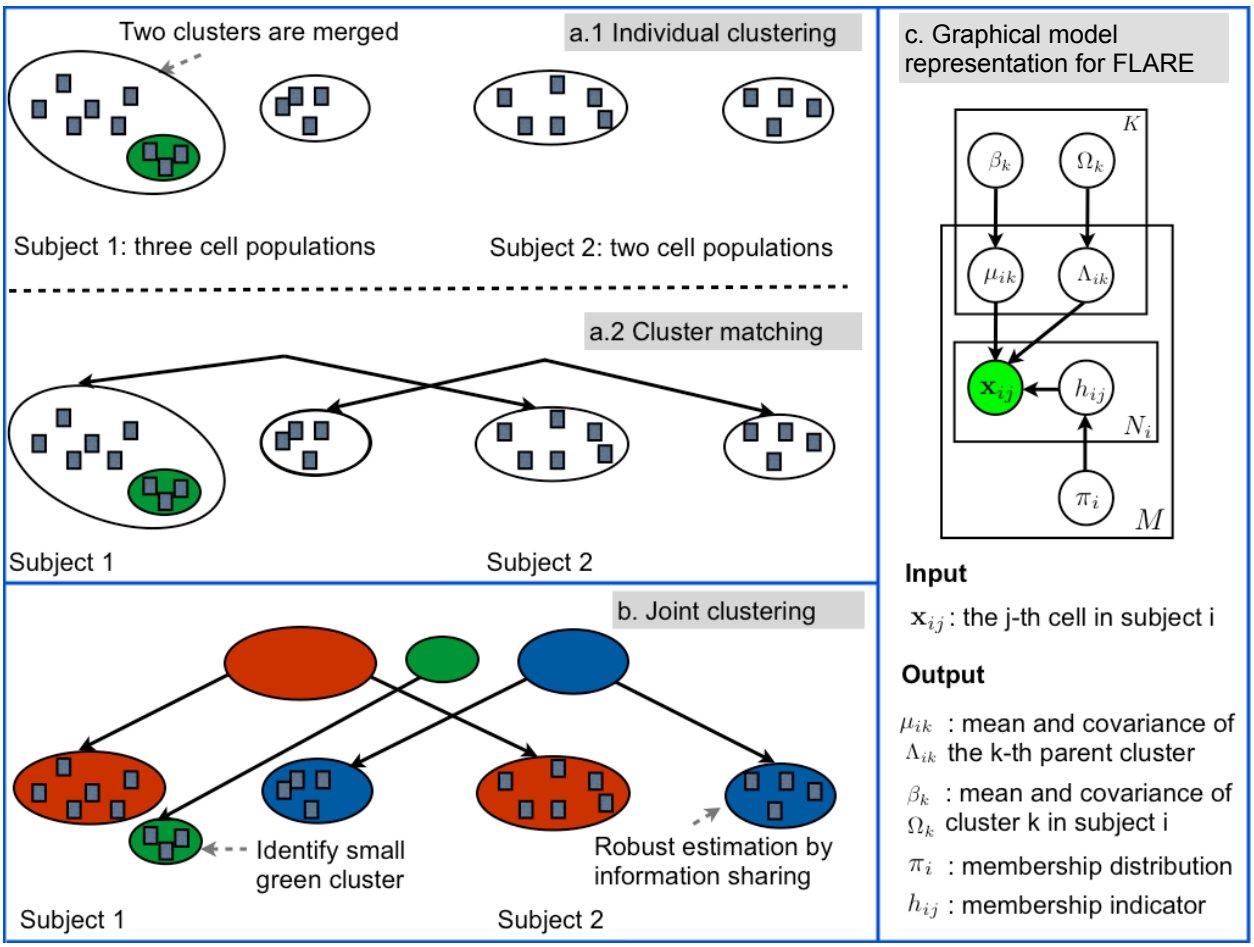


Figure 1: **Illustrative example for FLARE and its graphical model representation.** Panels a.1 and a.2 show the limitation of separate population analysis on individual samples: it misses the detection of the rare cell population in green. Panel b shows that by sharing population information via the parental nodes, FLARE can more accurately estimate the big cell population in the red cluster and detect the rare population in the green cluster. Panel c visualizes the hierarchical Bayesian model of FLARE (See Section A of the Supplementary for details) .

popular distributed computing platforms.

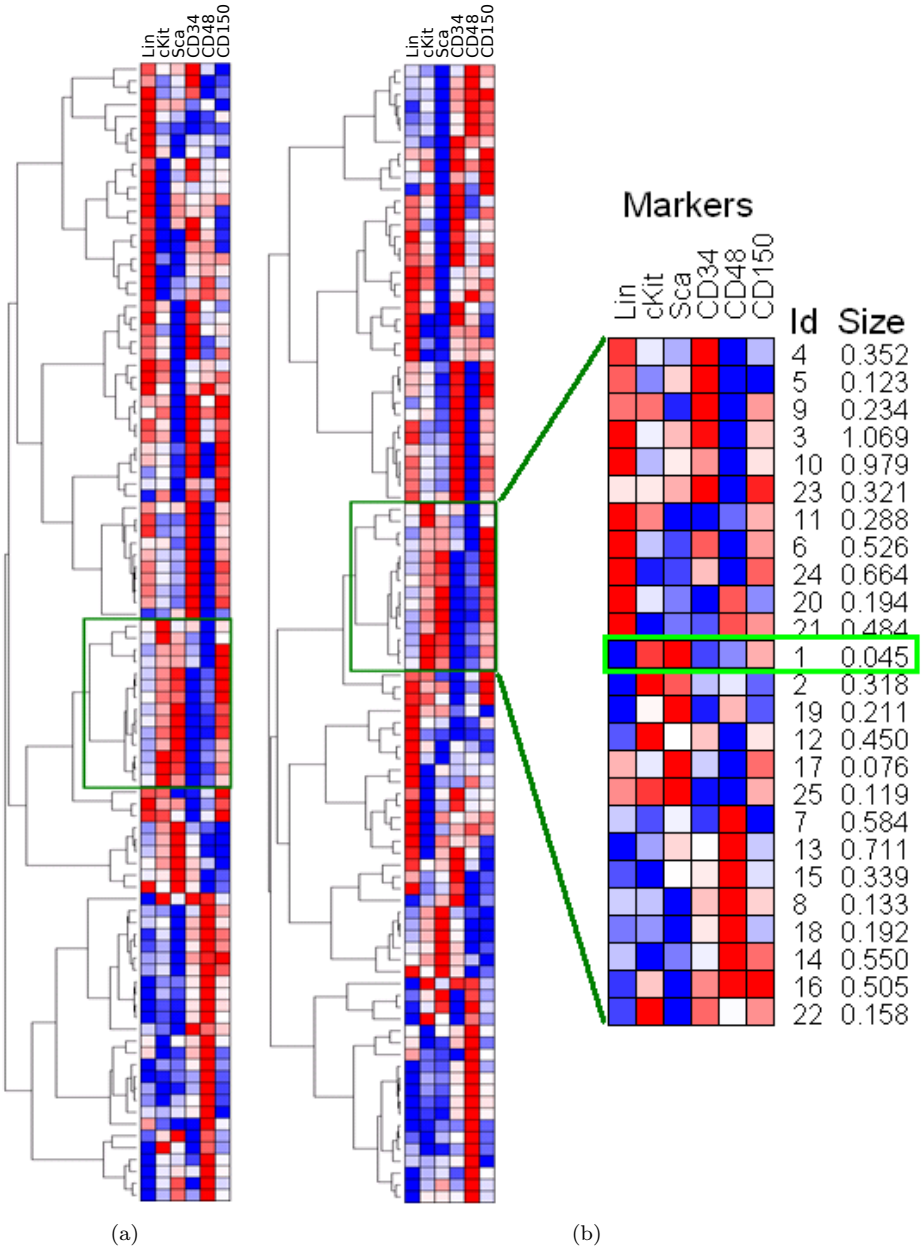
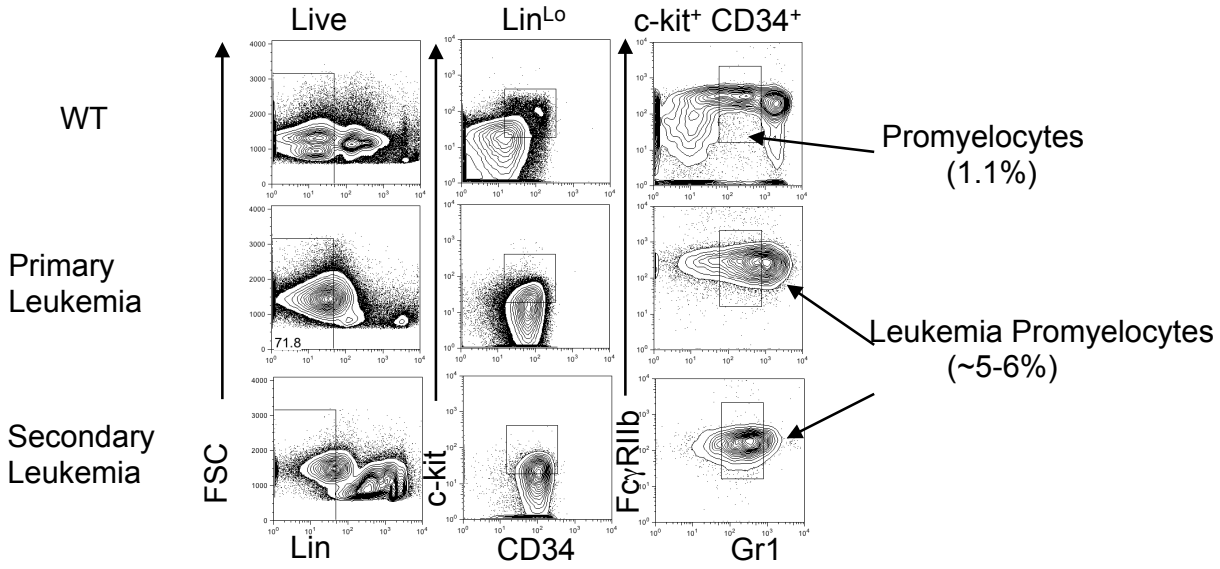


Figure 2: **FLARE identifies rare HSC population in murine bone marrow samples undergoing normal hematopoiesis.** (a) In the first step, unsupervised clustering by the hierarchical Bayesian model is used to explore the different subsets in 6-dimensional marker space. In 12 blood samples (3 biological replicate mice, 4 technical replicates per specimen), FLARE identified 96 populations matched across all samples. Plots (a) and (b) show the heatmaps for two representative samples, where each row represents a population's mean intensities for the 6 markers represented by the columns. Red/blue is used to depict high/low intensities. Thus a common region of interest, shown in green rectangle, was identified. In the 2nd step, in a new and much larger sample, FLARE zoomed into the specified region to detect the clusters therein. The uncovered hierarchy of populations is shown with a heatmap in plot (zoomed in right panel). We identified one particular population (denoted by cluster #1; light green rectangle) that has the size (0.045%) and marker-signature (Lineage^{-}c-kit^{+}Sca^{+}CD34^{-}CD48^{-}CD150^{hi}) consistent with the LT-HSC cells.


Live
Lin^Lo
c-kit+ CD34+
WT
Primary Leukemia
Secondary Leukemia
FSC
Lin
c-kit
CD34
FcγRIIb
Gr1
71.8
Promyelocytes (1.1%)
Leukemia Promyelocytes (~5-6%)

(a)

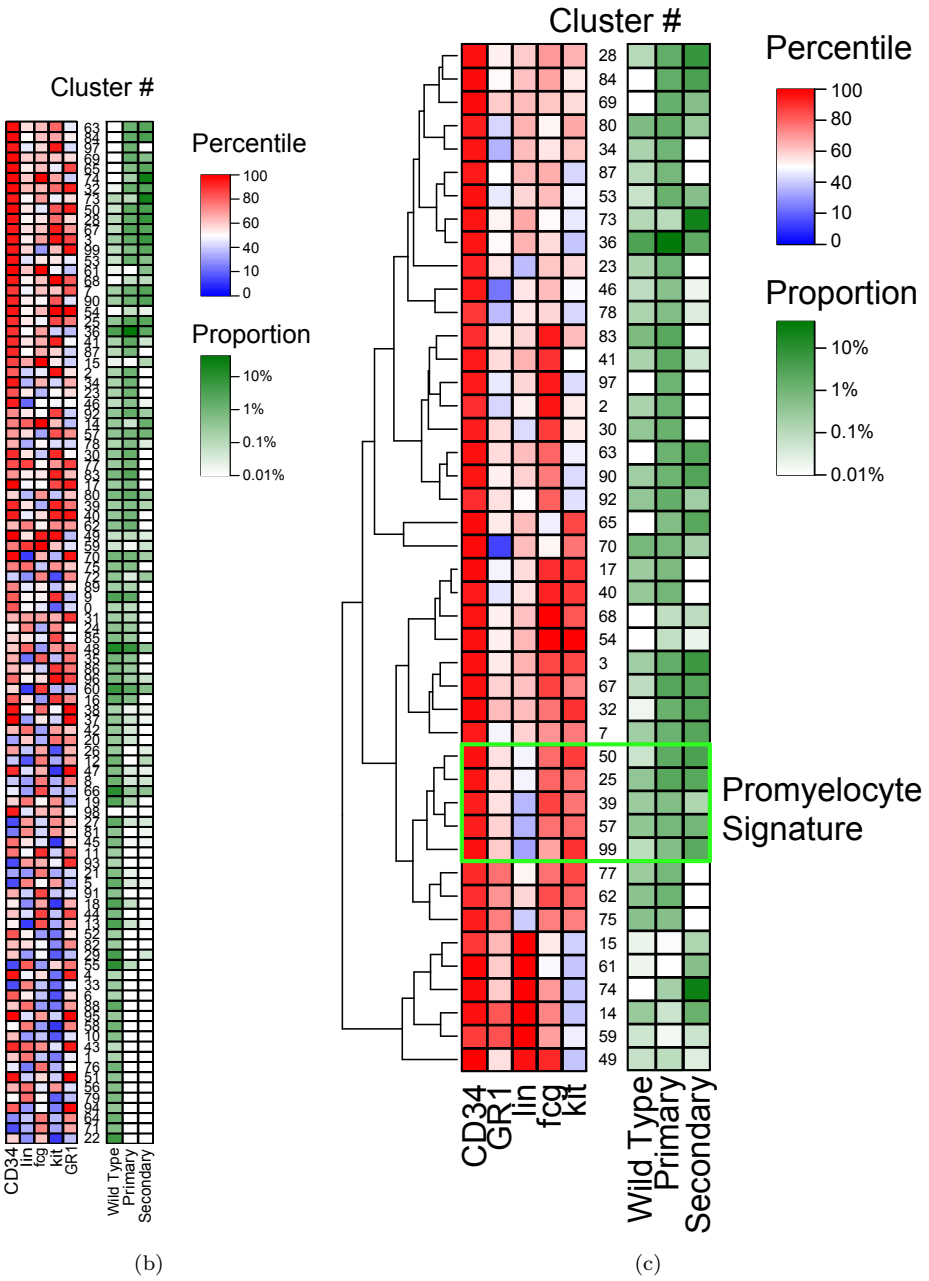


Figure 3: **FLARE in PML-RARα transgenic mouse model.** (a) Wildtype (C57BL6) bone marrow, primary leukemic mice with PML-RARα transgenic and secondarily transplanted mice were analyzed by flow cytometry gated as previously described [11]. Forward scatter (FSC), Lineage staining (Lin) are gated serially from left panels to right. Experimentally determined surface phenotype of leukemic promyleocytes are gated (right panel) and frequency of this population is shown among live cells. (b) Flow cytometric data from mice in (a) were analyzed with FLARE and a heat map of all 99 clusters with a surface phenotype and the proportion of the particular clusters among the live cells. (c) Data from (b) was further filtered based on populations with greater than 1-fold change in proportions from primary or secondary leukemia compared to wild type bone marrow cells. Green box indicated clusters (50, 25, 39, 57 and 99) that contain the previously experimentally derived known surface phenotype of the leukemic promyelocytes ($Lin^{lo}CD34^{hi}$c-$Kit^{hi}Gr1^{mid}FcgRIIb^{hi}$). Left panel represents the percentile of each flow staining parameter ranked among the 99 clusters. Right panel indicates proportion of populations within each cluster among the live-gated cells.

References

[1] Preffer, F. and Dombkowski, D. Advances in complex multiparameter flow cytometry technology: Applications in stem cell research. *Cytometry Part B: Clinical Cytometry* **76B**(5), 295–314, Sep (2009).

[2] Tanner, S. D., Bandura, D. R., Ornatsky, O., Baranov, V. I., Nitz, M., and Winnik, M. A. Flow cytometer with mass spectrometer detection for massively multiplexed single-cell biomarker assay. *Pure and Applied Chemistry* **80**(12), 2627–2641 (2009).

[3] Bendall, S. C., Simonds, E. F., Qiu, P., Amir, E.-a. D., Krutzik, P. O., Finck, R., Bruggner, R. V., Melamed, R., Trejo, A., Ornatsky, O. I., and et al. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* **332**(6030), 687–696, May (2011).

[4] Pyne, S., Hu, X., Wang, K., Rossin, E., Lin, T.-I., Maier, L. M., Baecher-Allan, C., McLachlan, G. J., Tamayo, P., Hafler, D. A., and et al. Automated high-dimensional flow cytometric data analysis. *Proceedings of the National Academy of Sciences* **106**(21), 8519–8524, May (2009).

[5] Qiu, P., Simonds, E. F., Bendall, S. C., Gibbs, K. D., Bruggner, R. V., Linderman, M. D., Sachs, K., Nolan, G. P., and Plevritis, S. K. Extracting a cellular hierarchy from high-dimensional cytometry data with spade. *Nature Biotechnology* **29**(10), 886–891, Oct (2011).

[6] Lee, S. X., McLachlan, G. J., and Pyne, S. Modeling of inter-sample variation in flow cytometric data with the joint clustering and matching procedure. *Cytometry Part A* **89**(1), 30–43, Jan (2016).

[7] Lugli, E., Roederer, M., and Cossarizza, A. Data analysis in flow cytometry: The future just started. *Cytometry Part A* **77A**(7), 705–713, Apr (2010).

[8] Pyne, S., Lee, S. X., Wang, K., Irish, J., Tamayo, P., Nazaire, M.-D., Duong, T., Ng, S.-K., Hafler, D., Levy, R., and et al. Joint modeling and registration of cell populations in cohorts of high-dimensional flow cytometric data. *PLoS ONE* **9**(7), e100334, Jul (2014).

[9] Maecker, H. T., McCoy, J. P., and Nussenblatt, R. Standardizing immunophenotyping for the human immunology project. *Nature Reviews Immunology* **12**(3), 191–200, Feb (2012).

[10] Brown, D., Kogan, S., Lagasse, E., Weissman, I., Alcalay, M., Pelicci, P. G., Atwater, S., and Bishop, J. M. A PML–RARα transgene initiates murine acute promyelocytic leukemia. *Proceedings of the National Academy of Sciences of the United States of America* **94**(6), 2551–2556, 03 (1997).

[11] Guibal, F. C., Alberich-Jorda, M., Hirai, H., Ebralidze, A., Levantini, E., Di Ruscio, A., Zhang, P., Santana-Lemos, B. A., Neuberg, D., Wagers, A. J., and et al. Identification of a myeloid committed

progenitor as the cancer-initiating cell in acute promyelocytic leukemia. *Blood* **114**(27), 5415–5425, Oct (2009).

[12] Wojiski, S., Guibal, F. C., Kindler, T., Lee, B. H., Jesneck, J. L., Fabian, A., Tenen, D. G., and Gilliland, D. G. PML–RARα initiates leukemia by conferring properties of self-renewal to committed promyelocytic progenitors. *Leukemia* **23**(8), 1462–1471, Mar (2009).