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], Saumyadipta Pyne[3,11,12*],

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 of Molecular Genetics of the ASCR, 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 ICMR National Institute of Medical Statistics, New Delhi, India
12 Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, USA

*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 focus within specified regions of interest in marker-space to detect 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 diverse 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 need for systematic and automated multivariate analysis and visualization suitable for high-dimensional data [4–6]. As the number of potential combinations of markers 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 cell populations for several technical reasons. The new data are not only high-dimensional (i.e., involving multi-parametric or multiplexed 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 "searching for a needle in a haystack" scenario for the identification 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 or features in omic data, 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 and/or technological sources of inter-sample variation including single cell level heterogeneity, individual subjects, different time-points and conditions, and platform noise — all of 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. Such clusters may contain complex high-dimensional structures such as heavy tails or skewness, 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.

Model

In this section, we describe the our new hierarchical Bayesian model, FLARE, for FLow cytometric Analysis of Rare Events, to identify cell populations from multiple samples and detecting rare cell populations. Given the increasing high-dimensionality of cytometric data, there is a critical need to assist the manual gating procedure using unsupervised computational approaches to explore the marker-space, especially to identify specific cell populations that may appear at unknown locations under certain conditions such as drug-resistant cells or a rare signature of disease prognosis.

To this end, we designed a hierarchical Bayesian model that can share information across multiple samples to substantiate the occurrence of any genuine rare cluster of events. First, we model the cell populations in each sample by a mixture of probability distributions, say, multivariate Gaussian components, so that we can assign a probability score to associate each cell with a population, thus reflecting the underlying structures of individual samples. Second, we let the Gaussian components — corresponding to cell populations in different samples — be similar to each other via common prototype populations up to certain small variations, so that we can capture the minor differences between individual samples. Third, we allow some Gaussian components to appear only in certain — but not necessarily all — samples, and report these populations, even if they are rare events.

Let us denote the cytometric data by $\mathbf{X}$ and the cell memberships by $\mathbf{H} = h_{nk}^{(m)}$ where $h_{nk}^{(m)}$ denotes the membership of the n-th cell in the m-th sample to that sample's k-th Gaussian component with mean $\boldsymbol{\mu}_k^{(m)}$ and precision $\boldsymbol{\lambda}_k^{(m)}$. Then the data likelihood is

$$P(\mathbf{X}|\boldsymbol{\Theta}, \mathbf{H}) = \prod_{m=1}^{M} \prod_{n=1}^{N_m} \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}_n^{(m)} | \boldsymbol{\mu}_k^{(m)}, (\boldsymbol{\Lambda}_k^{(m)})^{-1})^{h_{nk}^{(m)}} \tag{1}$$

where M is the number of samples, N_m the number of cells in sample m, K the maximal number of cell populations for each sample, and $\boldsymbol{\Theta} = \{\boldsymbol{\mu}_k^{(m)}, \boldsymbol{\Lambda}_k^{(m)}\}_{m,k}$. Each data point $\mathbf{x}_n^{(m)}$ has dimension D.

The latent membership indicators $\mathbf{H}$ has a factorized discrete prior distribution:

$$P(\mathbf{H}|\boldsymbol{\pi}) = \prod_{m=1}^{M} \prod_{n=1}^{N_m} \prod_{k=1}^{K} (\pi_k^{(m)})^{h_{nk}^{(m)}} \tag{2}$$

where $\pi_k^{(m)}$ is the probability of the kth population appearing in sample m and $\sum_k \pi_k^{(m)} = 1$. If $\pi_k^{(m)} = 0$, then the kth population does not exist in the m-th sample. To model the uncertainty in $\boldsymbol{\pi}^{(m)} = [\pi_k^{(m)}, \ldots, \pi_k^{(m)}]$, we use a symmetric Dirichlet prior distribution:

$$P(\boldsymbol{\pi}) = \prod_{m=1}^{M} C(\alpha_0) \prod_{k=1}^{K} (\pi_k^{(m)})^{\alpha_0 - 1} \tag{3}$$

where α_0 is a hyperparameter and $C(\alpha_0) = \frac{\Gamma(M\alpha_0)}{\Gamma(\alpha_0)^M}$.

To share information between clusters of different samples, we let the mean parameter, $\boldsymbol{\mu}_k^{(m)}$, of each cluster in a sample follow a Gaussian prior distribution common to all samples:

$$P(\boldsymbol{\mu}_k^{(m)} | \boldsymbol{\eta}_k) = \mathcal{N}(\boldsymbol{\mu}_k^{(m)} | \boldsymbol{\eta}_k, (\beta_0 \mathbf{I})^{-1}) \tag{4}$$

where $\boldsymbol{\eta}_k$ is the mean parameter of the kth prototype cluster — which is estimated from data as $\boldsymbol{\mu}_k^{(m)}$ — and β_0 is a hyperparameter. Similar, the covariance matrix, $\boldsymbol{\Lambda}_k^{(m)}$, of each cluster in a sample follows a Wishart prior distribution common to all samples:

$$P(\boldsymbol{\Lambda}_k^{(m)}|\boldsymbol{\Omega}_k) = \mathcal{W}(\boldsymbol{\Lambda}_k^{(m)}|\boldsymbol{\Omega}_k, \sigma_0) \tag{5}$$

where $\boldsymbol{\Omega}_k$ is a symmetric, positive definite matrix — estimated from data just as $\boldsymbol{\Lambda}_k^{(m)}$ — and σ_0 is the degree of freedom.

Since we need to estimate the parameters of the prototype clusters from data as well, we assign a Gaussian hyper-prior distribution over the mean of each prototype cluster, $\boldsymbol{\eta_k}$:

$$P(\boldsymbol{\eta_k}) = \mathcal{N}(\boldsymbol{\eta}_k|\mathbf{0}, \mathbf{I})^{-1}) \tag{6}$$

Also, we assign an Inverse-Wishart hyper-prior distribution over the shape of each each prototype cluster, $\boldsymbol{\Omega}$:

$$P(\boldsymbol{\Omega}_k) = \mathcal{W}^{-1}(\boldsymbol{\Omega}_k|\boldsymbol{\Phi}_0, \nu_0) \tag{7}$$

where $\boldsymbol{\Phi}_0$ and ν_0 are hyperparameters. In our experiments, we set $\boldsymbol{\Phi}_0 = \mathbf{I}$ and $\nu_0 = 6D$ to obtain a diffuse prior over $\boldsymbol{\Omega}_k$.

Combining the data likelihood, the priors and the hyper-priors, we obtain the following joint distribution for our model:

$$\begin{aligned}
&P(\mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\eta}, \boldsymbol{\Omega}, \mathbf{h}, \boldsymbol{\pi}) \\
=&P(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \mathbf{h})P(\mathbf{h}|\boldsymbol{\pi})P(\boldsymbol{\mu}|\boldsymbol{\eta})P(\boldsymbol{\Lambda}|\boldsymbol{\Omega})P(\boldsymbol{\eta})P(\boldsymbol{\Omega})P(\boldsymbol{\pi}) \\
=&\Big(\prod_{m=1}^{M}\prod_{n=1}^{N_m}\prod_{k=1}^{K}\mathcal{N}(\mathbf{x}_n^{(m)}|\boldsymbol{\mu}_k^{(m)}, (\boldsymbol{\Lambda}_k^{(m)})^{-1})^{h_{nk}^{(m)}}\prod_{m=1}^{M}\prod_{n=1}^{N_m}\prod_{k=1}^{K}(\pi_k^{(m)})^{h_{nk}^{(m)}}\Big) \\
&\Big(\prod_{k=1}^{K}\mathcal{N}(\boldsymbol{\eta}_k|\mathbf{0}, \mathbf{I})^{-1})\mathcal{W}^{-1}(\boldsymbol{\Omega}_k|\boldsymbol{\Phi}_0, \nu_0)\Big)
\end{aligned} \tag{8}$$

The joint distribution is depicted in Fig. 1.

With the priors specified over the cluster parameters in multiple samples and the hyper-priors over the parameters of the prototype clusters, we constructed a hierarchical Bayesian model, FLARE. The model allows the cluster locations (given by the means) and the shapes (given by the covariance matrices) of each sample to be similar to those of their prototype cluster so that the information from multiple samples could be combined for accurate and robust estimation of clusters. At the same time, FLARE allows the clusters of each sample to be slightly different from their prototypes, accounting for the variations among different biological samples. In our experiments, we set $\beta_0 = 500$ and $\sigma_0 = 6D$ so that the stochastic variation between a sample cluster mean and its prototype cluster mean is reasonable small.

Notably, in our model, a cluster can also not contain any data point in a particular sample, and thus, the cluster may be absent in certain samples. From our estimation results, we can easily distinguish which clusters are common to all samples and which only appear in certain samples. The Variational Bayes (VB) approach to compute the approximate posterior distribution of the FLARE parameters is described in Section A of the Supplementary Materials. The algorithm for parallel computing, and specifically for for sharing information across samples to detect the consistent clusters is described in Section B of the Supplementary Materials.

Results and discussion

We developed a new computational framework FLARE for FLow cytometric Analysis of Rare Events, although it may be applicable to other platforms that generate multi-marker data per cell.

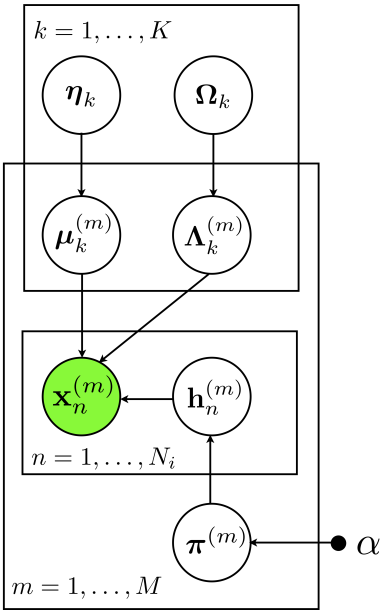

Fig 1. The probabilistic graphical model representation of FLARE. x is the cell, μ and Λ are the mean and covariance of the parent clusters, η and Ω are the mean and covariance of cluster k, π is the membership distribution, h is the membership indicator and α is the hyperparameter for π. The superscript m indicates the mth sample.

FLARE is based on a hierarchical Bayesian model, and employs parallel computation for its high-speed high-precision analysis. The Bayesian model (Fig 2) of FLARE allows implementation of several distinct features to specifically address the challenges mentioned above. For consistent identification of a particular rare population C, the model parameters allow information about C to be shared across different samples. In our parallel computing framework, 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 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 2) which contain very few cells since it effectively pools together more observations for estimation. Second, the estimation ambiguity (between the red and the green clusters in Sample 1, corresponding to Subject 1, in Fig 2) is reduced since the information about a population (e.g., the red cluster of Sample 2, corresponding to Subject 2, in Fig 2) can guide the estimation of its counterpart across samples (i.e., the red cluster of Sample 1 in Fig 2) 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 cluster alignment [4,13], we can identify also those clusters (e.g., the green cluster in Fig 2b) that exist only in certain samples, a situation that is not uncommon for rare populations e.g., transient subsets that are present only during certain stages of cell differentiation and are absent otherwise).

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 human guidance, unsupervised analysis by FLARE was run on 14 "training" samples, including multiple biological and technical replicates, and it identified a subset bearing a 6-marker signature of long-term murine

hematopoietic stem cells (LT-HSCs) (Fig 3a-b). In the second step, using this signature location parameter, FLARE focused on 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 two-dimensional gating, according to which LT-HSCs are known to be Lineage$^-$Kit$^+$Sca$^+$CD34$^-$CD48$^-$CD150$^+$. Using parallel computation (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 demonstrate how FLARE could be used for high-speed automated identification of rare populations in cytometric data.

Identifying a rare disease signature as a second application of FLARE during disease progression. Therefore, we utilized a model of myeloid leukemia that harbors the oncogenic fusion of the PML gene and the retinoic acid receptor alpha (RARα). These mice succumb to a lethal acute promeylocytic leukemia (APL) that can be subsequently transplanted with increasing aggressiveness [10]. We have previously characterized the cell surface phenotype which drives the APL and it closely resembles the normal promyelocyte population [11, 12]. Notably, the mature granulocytes, which differentiate from the leukemic stem cell population (LSC) are unable to transplant the disease. The manual gating strategy for this population is challenging since these cells express low levels of lineage markers and are sequentially 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 4a).

By running FLARE on live-gated cells stained with markers for lineage, c-kit, CD34, Gr-1 and FcgRIIb we could identify 99 unique clusters that contained cells (Fig 5a). To determine if the found clusters contained the LSC population, we focused on the populations with greater than 1-fold change in their proportions compared to the normal mice. This allowed us to "zoom" into 44 clusters of hematopoietic cells that increased during leukemia progression which were visualized with a heatmap (Fig. 5b). 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 5b). 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.

FLARE offers several distinct advantages specifically for characterization of rare populations. First, FLARE shares information across multiple samples in a hierarchical Bayesian model (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 it an advantage while searching samples which may contain populations ranging from significantly big to extremely rare. Instead, FLARE automatically allows an initial mixture model with a large number of components to become sparse as the inconsistent clusters are removed and the actual number of clusters used to fit the data is learned in the process. In practice, FLARE can be viewed as an efficient approximation to Dirichlet Process Mixture (DPM) models, which have been used in the past [13]. Of course, FLARE uses the strategy of information sharing for fitting robust models by verifying the rare clusters across samples. Simulation results showed that FLARE achieves favorable estimation performance over alternative methods (Supplementary Fig S2-S5).

Further, FLARE 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 modeling is parallelized with careful consideration on workload balancing in a distributed computing environment (Supplementary Algorithm 1 and Supplementary Fig S1). It achieves almost linear speedup given more computational nodes (Supplementary Fig S6), making it truly scalable for large datasets.

Identification of rare cell subsets – while establishing their correspondence across multiple 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 dissection of the chosen regions of interest in the marker-space to characterize the rare populations with further precision. Such progressive "zooming in" capability of FLARE mimics the 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 illustrated these aspects of FLARE using a multi-step analysis of a hierarchy of cell populations as observed in 2 datasets based on (i) normal hematopoiesis in mice, and (ii) oncogenic progression in a mouse model. Further examples of FLARE analyses of secondary (Treg) and simulated datasets along with the performance results are described in the Supplementary Materials.

Conclusion

In summary, the hierarchical design and distributed variational estimation allows FLARE to share information about corresponding clusters across samples, and quickly detect a variety of populations, including considerably rare ones, in an unsupervised manner. In the process, it efficiently searches the high-dimensional marker-space to reveal the underlying population structure. Thereupon it can progressively concentrate its search within 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 our future work we look forward to embedding this step into FACS systems for real-time sorting of the desired cells. Since the multi-parametric population signatures reported by FLARE are quantitative and precise, however rare the underlying events may be, it helps to verify and eventually standardize definitions of specific cellular species, allow objective extraction, and facilitate reproducible cytometric analysis. Finally, the parallel estimation algorithm in FLARE is currently implemented using Message Passing Interface (MPI) and can be readily adapted to popular distributed computing platforms.

Supporting information

S1 Fig. Topography of data storage among cluster nodes. Each rectangle represents a cluster node, with each column consisting of nodes from a particular sample. The edges of this graph connect nodes which must communicate with each other. The nodes of each column form a fully connected subgraph to show the communication done within each sample. Similarly, each column has representative node that participates in summations over all samples. The dotted edges represent the fact that based on the data, there can be an arbitrary number of samples and nodes per sample.

S2 Fig. The Adjusted Rand Index of each method on the synthetic datasets. We use the hard clustering results of the subject who has the small clusters to compute the ARIs against true clustering assignment.

S3 Fig. Visualization of clustering results in synthetic data.

S4 Fig. Maximum Jaccard index on Treg dataset.

S5 Fig. Maximum detection accuracy on Treg dataset.

S6 Fig. Speedup rate and load balancing efficiency. The top panel shows the speedup rate of our parallel inference algorithm using increasingly more cluster nodes. The bottom panel shows the load balancing efficiency. The balancing efficiency is calculated using equation (S32). With more nodes, the data are more evenly distributed so that the balancing efficiency keeps increasing.

S7 Fig. Maximum detection accuracy on Treg dataset.

File. Supplemental Materials for 'High-Speed Automatic Characterization of Rare Events in Flow Cytometric Data'. Further details on Variational Inference, Parallel Inference, and Experimental Results.

References

1. Preffer F, Dombkowski D. Advances in complex multiparameter flow cytometry technology: Applications in stem cell research. Cytometry Part B: Clinical Cytometry. 2009;76B(5):295–314. doi:10.1002/cyto.b.20480.

2. Tanner SD, Bandura DR, Ornatsky O, Baranov VI, Nitz M, Winnik MA. Flow cytometer with mass spectrometer detection for massively multiplexed single-cell biomarker assay. Pure and Applied Chemistry. 2009;80(12):2627–2641.

3. Bendall SC, Simonds EF, Qiu P, Amir EaD, Krutzik PO, Finck R, et al. Single-Cell Mass Cytometry of Differential Immune and Drug Responses Across a Human Hematopoietic Continuum. Science. 2011;332(6030):687–696. doi:10.1126/science.1198704.

4. Pyne S, Hu X, Wang K, Rossin E, Lin TI, Maier LM, et al. Automated high-dimensional flow cytometric data analysis. Proceedings of the National Academy of Sciences. 2009;106(21):8519–8524. doi:10.1073/pnas.0903028106.

5. Qiu P, Simonds EF, Bendall SC, Gibbs KD, Bruggner RV, Linderman MD, et al. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. Nature Biotechnology. 2011;29(10):886–891. doi:10.1038/nbt.1991.

6. Lee SX, McLachlan GJ, Pyne S. Modeling of inter-sample variation in flow cytometric data with the joint clustering and matching procedure. Cytometry Part A. 2016;89(1):30–43. doi:10.1002/cyto.a.22789.

7. Lugli E, Roederer M, Cossarizza A. Data analysis in flow cytometry: The future just started. Cytometry Part A. 2010;77A(7):705–713. doi:10.1002/cyto.a.20901.

8. Pyne S, Lee SX, Wang K, Irish J, Tamayo P, Nazaire MD, et al. Joint Modeling and Registration of Cell Populations in Cohorts of High-Dimensional Flow Cytometric Data. PLoS ONE. 2014;9(7):e100334. doi:10.1371/journal.pone.0100334.

9. Maecker HT, McCoy JP, Nussenblatt R. Standardizing immunophenotyping for the Human Immunology Project. Nature Reviews Immunology. 2012;12(3):191–200. doi:10.1038/nri3158.

10. Brown D, Kogan S, Lagasse E, Weissman I, Alcalay M, Pelicci PG, et al. A PML–RARα transgene initiates murine acute promyelocytic leukemia. Proceedings of the National Academy of Sciences of the United States of America. 1997;94(6):2551–2556.

11. Guibal FC, Alberich-Jorda M, Hirai H, Ebralidze A, Levantini E, Di Ruscio A, et al. Identification of a myeloid committed progenitor as the cancer-initiating cell in acute promyelocytic leukemia. Blood. 2009;114(27):5415–5425. doi:10.1182/blood-2008-10-182071.

12. Wojiski S, Guibal FC, Kindler T, Lee BH, Jesneck JL, Fabian A, et al. PML–RARα initiates leukemia by conferring properties of self-renewal to committed promyelocytic progenitors. Leukemia. 2009;23(8):1462–1471. doi:10.1038/leu.2009.63.

13. Cron A, Gouttefangeas C, Frelinger J, Lin L, Singh, SK, Britten, CM. et al. Hierarchical Modeling for Rare Event Detection and Cell Subset Alignment across Flow Cytometry Samples. PLOS Computational Biology. 2013;9(7):1–14. doi:10.1371/journal.pcbi.1003130.

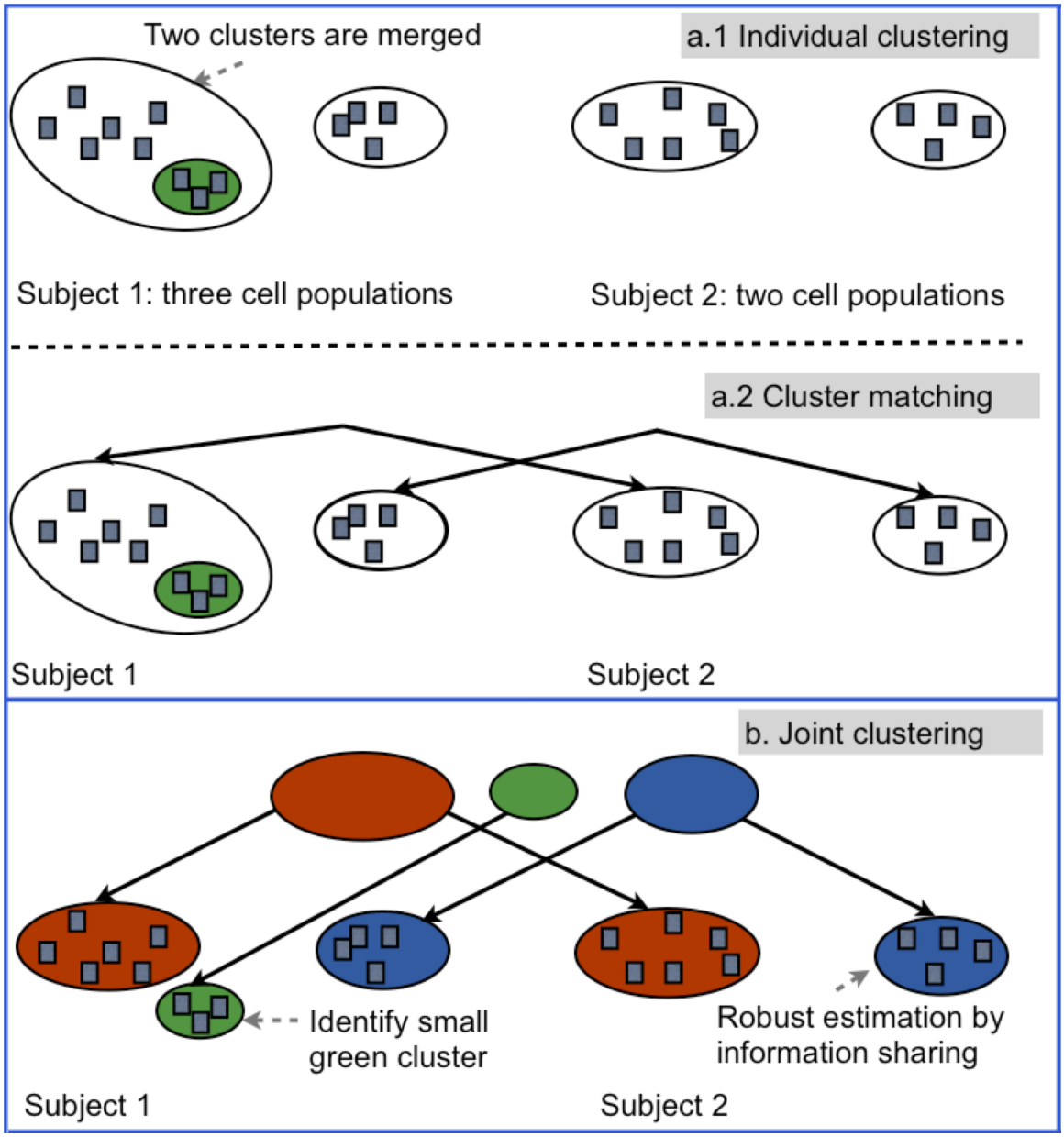

Fig 2. Illustrative example for FLARE and its graphical model representation. Panels a.1 and a.2 show the limitation of separate population analysis on individual samples: it misses the detection of the rare cell population in green. Panel b shows that by sharing population information via the parental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

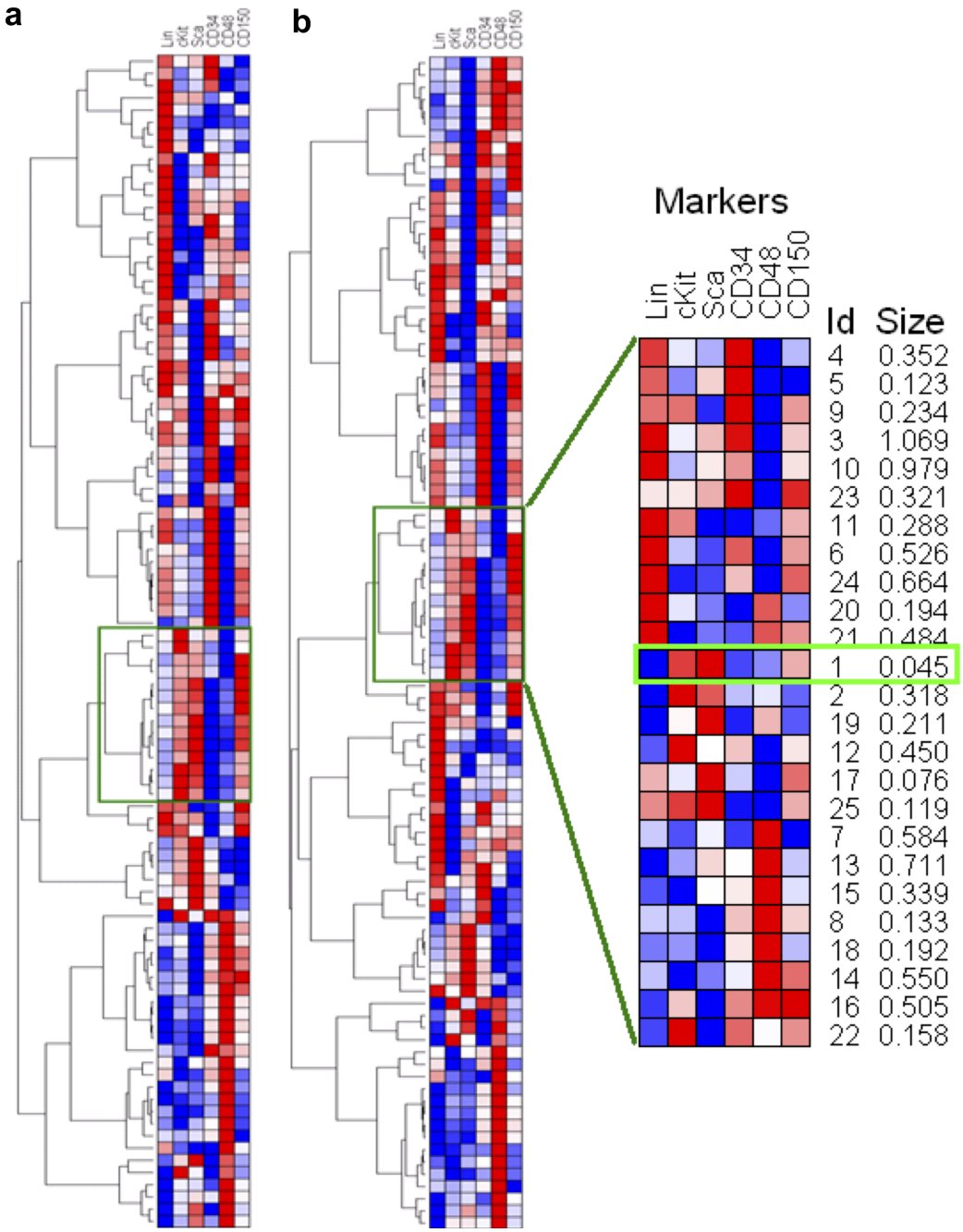

Fig 3. FLARE identifies rare HSC population in murine bone marrow samples undergoing normal hematopoiesis. (a) In the first step, unsupervised clustering by the hierarchical Bayesian model is used to explore the different subsets in 6-dimensional marker space. In 12 blood samples (3 biological replicate 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.

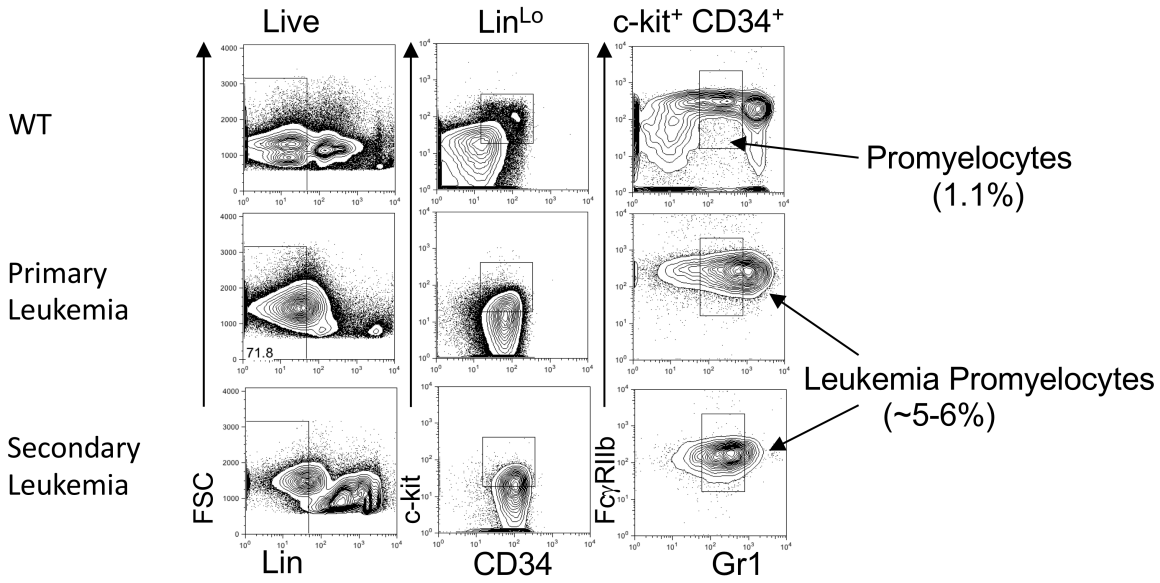

Fig 4. FLARE in PML-RARα transgenic mouse model. Wildtype (C57BL6) bone marrow, primary leukemic mice with PML-RARα transgenic and secondarily transplanted mice were analyzed by flow cytometry (and gated as previously described [11]). Forward scatter (FSC), Lineage staining (Lin) are gated serially from 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.

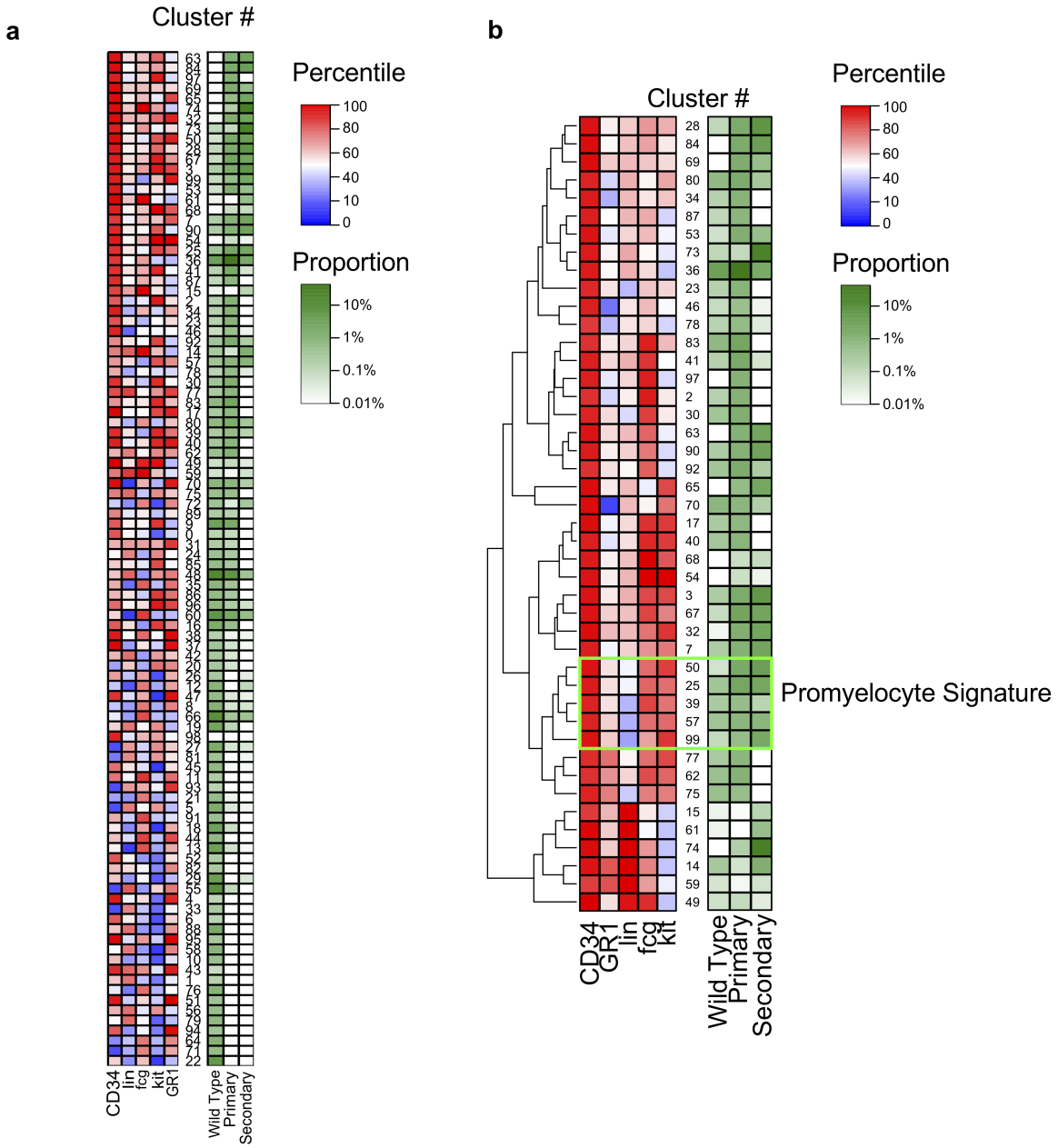

Fig 5. FLARE in PML-RARα transgenic mouse model. Flow cytometric data from mice in panel (a) were analyzed with FLARE and a heat map of all 99 clusters with a surface phenotype and the proportion of the particular clusters among the live cells. The populations with greater than 1-fold change in their proportions from primary or secondary leukemia compared to wild type bone marrow cells are shown in panel (b). Green box indicates clusters (50, 25, 39, 57 and 99) that contain the previously experimentally derived known surface phenotype of the leukemic promyelocytes ($Lin^{lo}CD34^{hi}c\text{-}Kit^{hi}Gr1^{mid}FcgRIIb^{hi}$). The left panel represents the percentile of each flow staining parameter ranked among the 99 clusters. The right panel indicates proportion of live-gated cells within each cluster.