Application of user-guided automated cytometric data analysis to large-scale immunoprofiling of invariant natural killer T cells

Xinli Hu[a,b,c,d,e,1], Hyun Kim[a,b,c,1], Patrick J. Brennan[a], Buhm Han[a,b,c,d], Clare M. Baecher-Allan[f], Philip L. De Jager[d,g], Michael B. Brenner[a,2], and Soumya Raychaudhuri[a,b,c,d,h,2]

[a]Division of Rheumatology, Immunology, and Allergy, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115; [b]Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115; [c]Partners Center for Personalized Genetic Medicine, Boston, MA 02115; [d]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142; [e]Harvard-MIT Division of Health Sciences and Technology, Harvard Medical School, Boston, MA 02115; [f]Department of Dermatology/Harvard Skin Disease Research Center, Brigham and Women's Hospital, Boston, MA 02115; [g]Program in Translational Neuropsychiatric Genomics, Department of Neurology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115; and [h]Faculty of Medical and Human Sciences, University of Manchester, Manchester M13 9PT, United Kingdom

Contributed by Michael B. Brenner, October 11, 2013 (sent for review July 7, 2013)

Defining and characterizing pathologies of the immune system requires precise and accurate quantification of abundances and functions of cellular subsets via cytometric studies. At this time, data analysis relies on manual gating, which is a major source of variability in large-scale studies. We devised an automated, user-guided method, X-Cyt, which specializes in rapidly and robustly identifying targeted populations of interest in large data sets. We first applied X-Cyt to quantify $CD4^+$ effector and central memory T cells in 236 samples, demonstrating high concordance with manual analysis ($r = 0.91$ and 0.95, respectively) and superior performance to other available methods. We then quantified the rare mucosal associated invariant T cell population in 35 samples, achieving manual concordance of 0.98. Finally we characterized the population dynamics of invariant natural killer T (iNKT) cells, a particularly rare peripheral lymphocyte, in 110 individuals by assaying 19 markers. We demonstrated that although iNKT cell numbers and marker expression are highly variable in the population, iNKT abundance correlates with sex and age, and the expression of phenotypic and functional markers correlates closely with CD4 expression.

automated analysis | flow cytometry

Flow cytometry is a technology widely used in clinical practice and in research, particularly in the field of immunology. It is capable of interrogating a wide variety of markers on many different cell types on a single-cell basis, using fluorophore-conjugated antibodies. Although molecular as well as genomic studies have advanced the understanding of immunological processes and autoimmune diseases, the components of the human immune system and their functions have yet to be comprehensively described. Without such a reference "catalog" of the immune system, it is ultimately difficult to interpret the pathogenic significance of genetic, molecular, or phenotypic variants observed in diseases.

Immunoprofiling is emerging as a means to establish the constituents, physiological roles, and population dynamics of the immune system (1). Specifically, it aims to define the cellular components of the immune system, the developmental processes and lineage relationship among the cell types, and the phenotypes and functions of each cell type at different physiological states. To profile such a complex and dynamic system in large sample sizes, high-throughput cytometric studies have become crucial.

Cytometric technologies are quickly advancing and outpacing analytical approaches. At this time, flow cytometers can measure up to 17 markers (2). Next-generation cytometers, such as cytometry by time of flight, will soon to be able to assay hundreds of markers (3). However, data analysis largely relies on manual gating by expert analysts. It is a simple, but slow, process that is dependent on 1D or 2D visualization and sequential gating, using software such as FlowJo. As the numbers of samples and markers in a study increase, gating becomes increasingly time consuming and inconsistent and does not fully exploit the power of high-dimensional information contained in these complex studies.

In recent years, a number of automated methods for cytometric data handling, particularly for cell population identification, have emerged and demonstrated their power to harness the rich information in large-scale data, minimize inconsistencies, and reduce analysis time (4). Current methods use parametric (5–7) or nonparametric (8–13) clustering to partition high-dimensional data. Some methods specialize in capturing difficult (such as rare or convex) cell populations (6, 7, 13) and delineating developmental and functional relationships among cell types (14). These methods make no assumptions about the underlying structure of the data and primarily aim to discover all discernible populations de novo in each sample. As a consequence, they have been used primarily for exploratory studies.

In contrast to exploratory studies, the goal of many immunoprofiling studies is to reliably and consistently identify the target cell population across many individuals. For example, a profiling study may aim to quantify regulatory T (Treg) cells in healthy controls and patients with autoimmune diseases, using antibodies specifically selected for identifying Tregs. In this case, the goal of the analysis is to accurately extract Tregs from all samples, using

Significance

Multimarker cytometry technologies, including cytometry by time of flight, are rapidly advancing and will soon produce large sets of rich immunoprofiling data. Here, we introduce X-Cyt, an automated cytometric data analysis tool for large-scale studies. Unlike existing software intended for exploratory analysis, it specializes in identifying populations of interest in large sample numbers. X-Cyt incorporates user guidance to allow customizable and targeted analysis, as well as template-guided partitioning to enable rapid batch processing. In this study, we demonstrate its performance in three data sets. With its versatility, reproducibility, and speed of analysis, as well as user-customized flexible analytical scheme, X-Cyt will certainly become a valuable tool for future cytometry-based profiling studies.

Author contributions: X.H., H.K., M.B.B., and S.R. designed research; X.H., H.K., and P.J.B. performed research; X.H., C.M.B.-A., and P.L.D.J. contributed new reagents/analytic tools; X.H. and H.K. analyzed data; C.M.B.-A. oversaw experimental design; P.L.D.J. provided samples; and X.H., H.K., P.J.B., B.H., M.B.B., and S.R. wrote the paper.

The authors declare no conflict of interest.

[1]X.H. and H.K. contributed equally to this work.

[2]To whom correspondence may be addressed. E-mail: mbrenner@rics.bwh.harvard.edu or soumya@broadinstitute.org.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1318322110/-/DCSupplemental.

a standardized definition. Automating this type of analysis is challenging because accurate intersample alignment of cell populations is required in addition to the partitioning of the cell populations within each sample.

We developed a user-guided analytical tool, X-Cyt, for automating targeted population identification in immunoprofiling studies. X-Cyt uses multivariate mixture modeling for partitioning cytometric data. Unlike unsupervised methods, X-Cyt allows the user to set up the optimal partitioning scheme. By applying a uniform scheme to all samples in a cohort, X-Cyt consistently identifies and aligns the targeted cell populations.

In this study, we aimed to identify and characterize invariant natural killer T (iNKT) cells. iNKT cells are lymphocytes with a nondiverse T-cell receptor repertoire that recognize CD1d-presented lipid antigens (15–17) and, in humans, normally make up less than 0.5% of circulating peripheral blood mononuclear cells (PBMCs) (18). They play important roles in host defense, autoimmunity, allergy, and cancer (19). Functional characterization of iNKT cells requires comprehensive assessment of surface expression of homing receptors, lectins, and cell adhesion molecules, as well as cytokine production. Immunoprofiling studies have yet to assay such a comprehensive set of markers in primary iNKT cells in a sufficiently large cohort (18, 20–24). Here, we profiled iNKT cells in 110 subjects with 19 surface and intracellular markers.

Results

Overview of the X-Cyt Method. X-Cyt identifies the populations of interest in a given sample by partitioning all events into clusters following a user-designed partitioning scheme. When more than one marker is used to define populations, X-Cyt partitions the data using multivariate mixture modeling via an expectation-maximization (EM) algorithm, as described in *Methods* and *SI Appendix*.

We make the assumption that in profiling studies, samples within a cohort share a general cell population structure. That is, similar cell populations are present in all samples, and their relative spatial configuration is conserved. X-Cyt therefore aims to follow the same user-defined partitioning scheme to analyze all samples while allowing for biological and technical variations. Population identification by X-Cyt is therefore accomplished in two major steps: a user-guided "trial" analysis to set up the partitioning scheme and a template-guided cohort analysis. Markers that describe the phenotype and function of cells are analyzed separately downstream of population identification.

Step 1. Set up the Partitioning Scheme. The goal of the initial trial phase of the analysis is to set up a partitioning scheme by optimizing two parameters for mixture modeling: a parsimonious combination of differentiation markers for defining the population or populations of interest and the number of clusters to adequately and intuitively partition the events. The user test-partitions a few representative samples, using different input parameters; evaluates the results; and then chooses one optimal scheme. The ideal resulting configuration is one that most accurately captures each target population as one coherent cluster of events (see *Methods* for a detailed description of parameter selection). The user-approved configuration (parameters of the mixture model components) is passed onto the second step as the template (Fig. 1*A*).

Step 2. Template-Guided Cohort Analysis. X-Cyt initializes the mixture model parameters of each sample to that of the template. The EM algorithm then iteratively updates the parameters describing the location, shape (covariance matrix), and proportion of each cluster. The EM algorithm indexes each emerging cluster according to the template, which automatically aligns across all samples simultaneously to clustering (Fig. 1*B*). Downstream to population extraction, markers that describe the phenotype and function of cells are analyzed separately (Fig. 1*C*).

We have made X-Cyt, along with a sample data set and user input files, available for download at www.broadinstitute.org/mpg/xcyt/.

Demonstration of X-Cyt's Performance in Two Data sets. We first assessed the performance of X-Cyt in identifying common cell populations by querying the proportions of memory cell subsets in $CD4^+$ T cells. We isolated $CD4^+$ T cells from PBMCs via magnetic-activated cell sorting depletion from a cohort of 236 healthy donors and labeled them with antibodies against CD45 isoforms RA and RO (CD45RA, CD45RO), and L-selectin (CD62L) (see *Methods* and *SI Appendix* for experimental methods).

To identify effector memory (T_{EM}) and central memory (T_{CM}) T cells, we partitioned each sample in two steps: bivariate normal mixture modeling using forward-scatter (FSC) and side-scatter (SSC) to obtain a purer $CD4^+$ T-cell population, and 3D normal

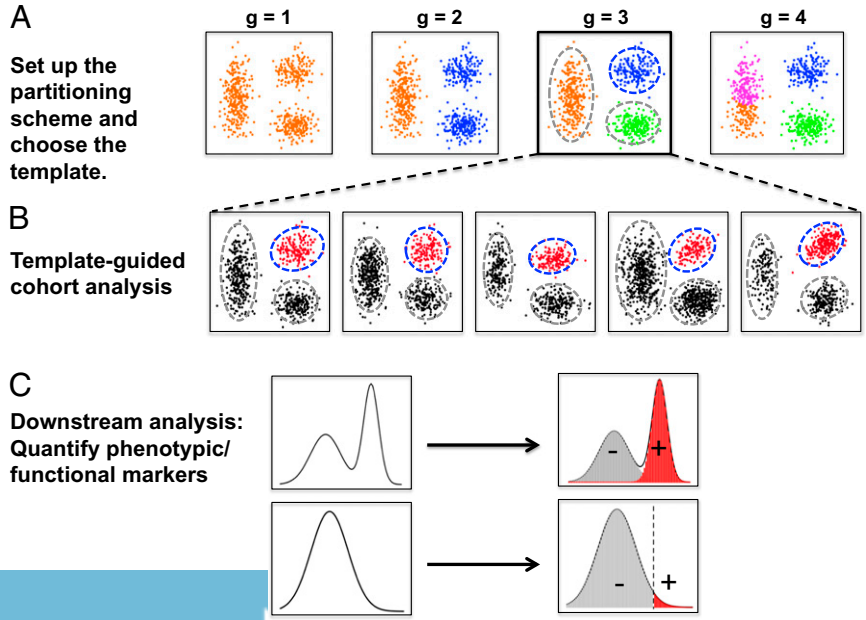

Fig. 1. Schematic of X-Cyt's analytical process (synthetic samples). (*A*) In a few representative samples, the user adjusts analytical parameters and evaluates the clustering outcome. Adjustable parameters include the differentiation markers to be used, the number of clusters in mixture modeling (g), distribution type, and SD cutoffs for continuous markers. The user selects one optimal set of parameters that most accurately identifies the cell populations of interest (here the blue cluster using g = 3). The clustering result of the representative sample is chosen as the template (dashed circles in the g = 3 panel). (*B*) X-Cyt applies the template to guide the partitioning of all samples in the study. The population of interest (shown in red dots and blue dashed circle) is consistently identified across all samples. (*C*) Downstream to population extraction, random samples are pooled to establish the distribution of phenotypic/functional markers. The percentage of cells positive for each marker is reported based on either mixture modeling (top) or SD cutoff (bottom).

mixture modeling using CD45RA, CD45RO, and CD62L. To determine the optimal partitioning scheme, an expert analyst assessed different sets of partitioning parameters in 10 random samples. In the first step, a two-component mixture model captured the $CD4^+$ T-cell population, which was extracted from each sample. In the second step, the analyst evaluated a range of four to nine clusters and selected the seven-cluster model, as it most accurately captured the T_{EM} and T_{CM} subsets (Fig. 2*A*). X-Cyt applied this partitioning scheme to all 236 samples and consistently identified the T_{EM} and T_{CM} subsets (representative samples shown in Fig. 2*B*). We compared X-Cyt results to proportions defined by an independent expert cytometry analyst with manual gating in FlowJo. We observed that the proportions for both populations were highly concordant with the manual analysis ($r = 0.91$ and $r = 0.95$; $P < 10^{-15}$, Pearson correlation test; Fig. 2*C*).

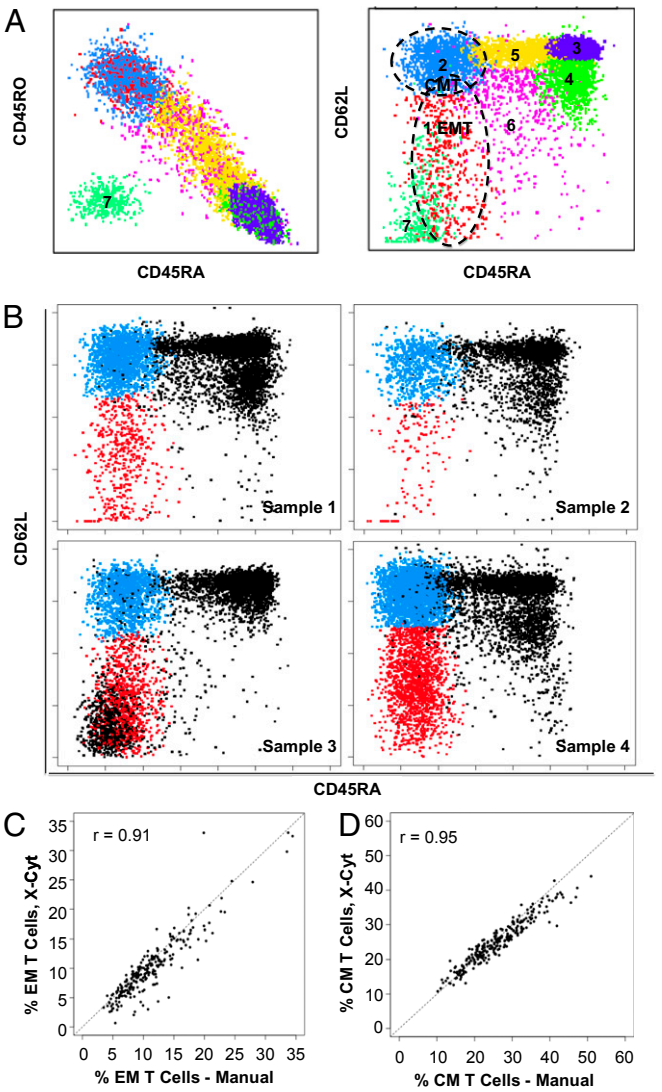

Fig. 2. $CD4^+$ memory T-cell subset identification. (*A*) An optimal model of seven clusters using CD45RA, CD45RO, and CD62L identified T_{EM} (red, cluster1) and T_{CM} (blue, cluster2) cells. Other clusters include naive and intermediate $CD4^+$ T cells, as well as impurities. (*B*) X-Cyt consistently identified the T_{EM} (red) and T_{CM} (blue) populations in all samples. Four random samples are shown here. (*C*) X-Cyt and manual gating in FlowJo returned highly concordant proportions of T_{EM} and (*D*) T_{CM} in 236 samples.

We wanted to quantitatively compare the performance of X-Cyt with that of automated methods that were the top five performers in the FlowCAP consortium challenge (4): FLoCK, ADICyt, flowMeans, FLAME, and SamSPECTRAL. We were unable to run FLoCK, as it was not able to use standard FCS3.0 format files. We ran each method with their default parameters to identify $CD4^+$ T_{EM} cells from lymphocytes. We analyzed all 236 samples, using FLAME, and compared the $CD4^+$ T_{EM} cell percentages with those procured by an expert user via gating in FlowJo. FLAME achieved more modest concordance ($r = 0.50$) compared with that achieved by X-Cyt ($r = 0.91$; see *SI Appendix*, Fig. S1). Because flowMeans and SamSPECTRAL do not align clusters across samples, comparison of results with X-Cyt was not possible without manual intervention. Therefore, we manually inspected a random subset (20 samples) of the clustering results, and in each sample we selected the cluster most closely representing the T_{EM} cells to obtain a concordance. Even after manual selection of clusters, flowMeans and SamSPECTRAL achieved limited concordances of only 0.57 and 0.44, respectively. ADICyt had a high sample failure rate. In three separate attempts with the same 20 samples, we observed that on average, 50% of samples failed to cluster with different random seeds. We note, however, that ADICyt achieved high performance on the limited samples it did successfully analyze ($r = 0.98$, average of three runs). Representative clustering results by each method are shown in *SI Appendix*, Fig. S2.

Next we challenged X-Cyt to identify a rare population, mucosal associated invariant T (MAIT) cells, from PBMCs for 35 subjects. We labeled cells with antibodies against CD3, CD45, Vα7.2, and CD161. Following convention, we defined MAIT cells as $CD3^+$ $CD45^+V\alpha7.2^+CD161^+$. We first partitioned PBMCs into four clusters, using FSC and SSC to obtain lymphocytes. Subsequently, we partitioned in CD3 and CD45 dimensions to obtain a double-positive T-cell population. In five random samples, the analyst evaluated three to seven clusters and selected the six-cluster model as the best to identify MAIT cells. We then applied the template to all 35 samples. Comparing proportions obtained by X-Cyt with those procured by an independent manual analyst, we again observed high concordance ($r = 0.98$; $P < 10^{-15}$; *SI Appendix*, Fig. S3).

Characterizing Rare iNKT Cells. We applied X-Cyt to identify iNKT cell subsets from the peripheral blood samples of 110 individuals (see *SI Appendix* for experimental methods). We labeled PBMCs with a total of 19 surface and intracellular markers in nine separate panels. Each panel included the antibodies for CD3ε, CD4, and α-galactosylceramide-loaded CD1d tetramer, which are the standard markers used to identify iNKT cells (25), as well as two to three phenotypic or functional markers.

We configured X-Cyt to identify iNKT cells in three steps: a three-component bivariate normal mixture modeling using FSC and SSC to extract lymphocytes from PBMCs, a three-component bivariate normal mixture modeling using CD3ε and CD4 to identify $CD4^+$ and $CD4^-$ T cells, and a threshold cutoff of five standard deviations above the mean of all lymphocytes in CD1d tetramer to identify the iNKT cells (Fig. 3*A*).

We observed that iNKT cells were present in individuals at extremely low but highly variable abundances, ranging from 0.0033% to 0.89% of all $CD3\varepsilon^+$ cells (mean = 0.072%; median = 0.031%). The proportion of iNKT cells that are $CD4^+$ also ranged dramatically, from 1.4% to 87% (mean = 39.5%). For comparison, an expert user manually gated and quantified iNKT cells and the $CD4^+$ subset in 36 of the 110 subjects, using FlowJo. Automated and manual results were almost perfectly concordant for the percentages of both iNKT cells ($r = 0.99$; Fig. 3*B*) and the $CD4^+$ subset ($r = 0.99$; *SI Appendix*, Fig. S4).

Rapid and robust processing of cytometric data makes it feasible to discover population dynamics of immune cell subsets from profiling studies. We examined our cohort of 110 samples

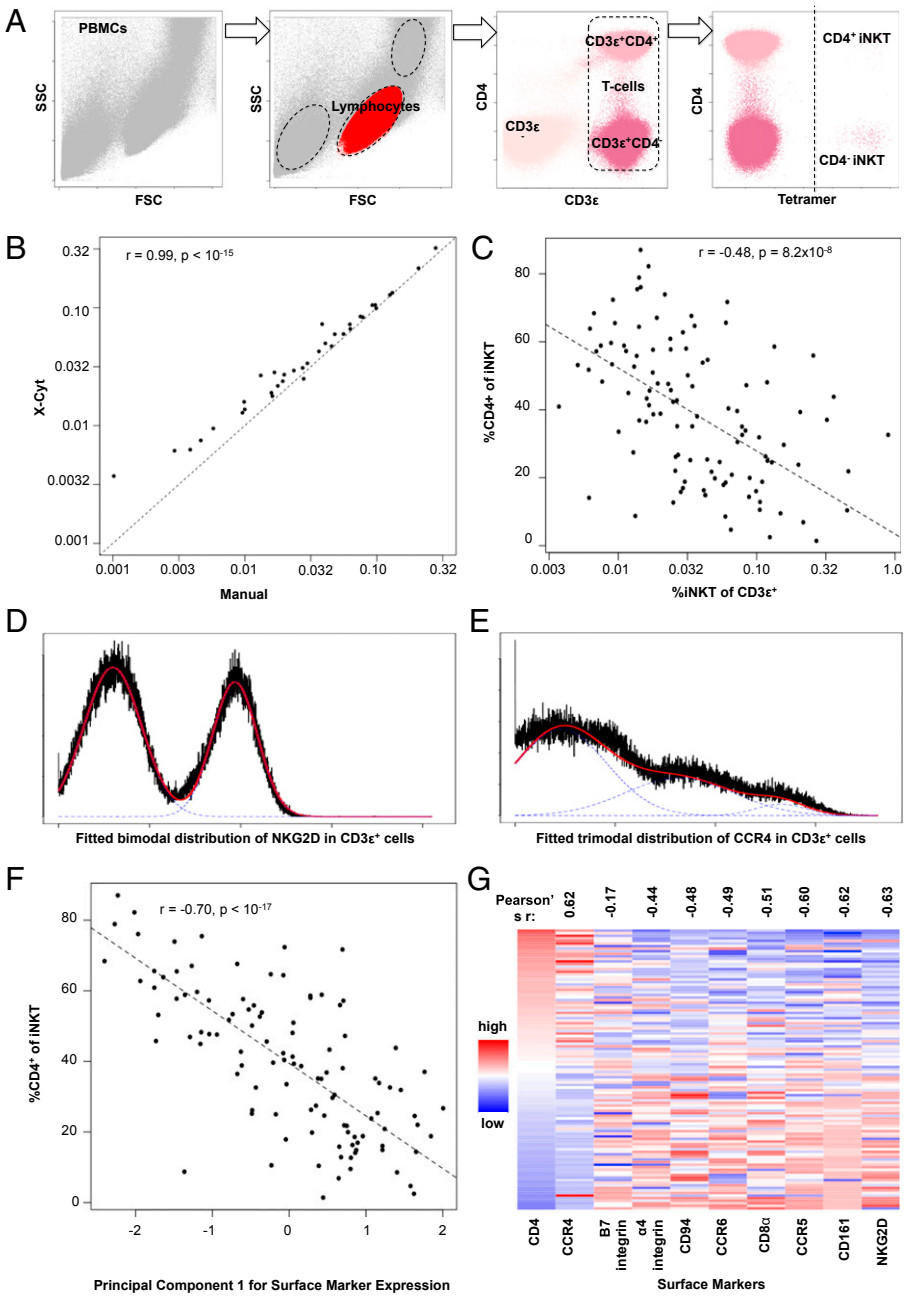

Fig. 3. iNKT cell identification. (*A*) The partitioning scheme: FSC and SSC were clustered into three components to identify the lymphocyte population (red). Lymphocytes were subsequently clustered using CD3ε and CD4 into three components; namely, the $CD3\varepsilon^-$, $CD3\varepsilon^+CD4^+$, and $CD3\varepsilon^+CD4^-$ populations. A cutoff of five SDs above the mean in α-galactosylceramide-loaded CD1d tetramer isolated the $tetramer^+$ iNKT cells (either $CD4^+$ or $CD4^-$). (*B*) X-Cyt returned iNKT cell proportions highly concordant with manual gating (Pearson's $r = 0.99$). (*C*) The $CD4^+$ proportion of iNKT cells correlates negatively with total iNKT abundance. Randomly sampled $CD3\varepsilon^+$ cells from all 110 samples were pooled to establish the intensity distribution of each phenotypic marker. Fitted distributions of (*D*) NKG2D (bimodal) and (*E*) CCR4 (trimodal) are shown. (*F*) The first principal component of expression levels of the nine surface markers, which captured 31.8% of total variation, correlates strongly with the proportion of $CD4^+$ iNKT cells. (*G*) The heat map shows the correlations (also indicated by Pearson's r on top) of the nine surface markers' expression with $CD4^+$ proportion in iNKT cells. Each row represents one sample.

for interesting population dynamics of iNKT cellular subsets. First, we note that 11 of the 110 subjects had two visits separated by at least 2 mo. In these subjects, we observed stable iNKT abundances and $CD4^+$ proportions over time ($r = 0.99$ and 0.98, respectively; see *SI Appendix*, Fig. S5 and Fig. S6). We observed a negative correlation between the proportion of $CD4^+$ iNKT cells and the ($\log_{10}$) proportion of total iNKT cells ($r = -0.48$; $P = 8.2 \times 10^{-8}$, Pearson correlation; Fig. 3*C*). Also, women had significantly higher amounts of iNKT cells than men ($\text{median}_{\text{female}} = 0.038\%$; $\text{median}_{\text{male}} = 0.022\%$; $P = 8.7 \times 10^{-3}$, Wilcoxon test; *SI Appendix*, Fig. S7). Finally, we observed that iNKT cell abundance correlated negatively with age ($P = 0.014$, Pearson correlation; *SI Appendix*, Fig. S8). The correlations between iNKT cell abundance and age, sex, and $CD4^+$ proportion are independent of each other; they remain significant in a multivariate regression (*SI Appendix*, Table S2). However, the proportion of $CD4^+$ iNKT cells was not correlated with sex ($P = 0.12$) or age ($P = 0.75$). Some of these trends had been observed in previous data sets (24). With a larger sample size, we confirmed the correlations with statistical significance.

Downstream of the successful identification and quantification of $CD4^+$ and $CD4^-$ iNKT cell subsets, we characterized the expression pattern of phenotypic markers in each. We quantified the expression of each marker in each subset by measuring the proportion of events with positive expression. We randomly sampled and pooled $CD3\varepsilon^+$ cells from all subjects to display the natural intensity distribution of each marker. Two examples of phenotypic markers are shown in Fig. 3 *D* and *E*. Eight of the 11 surface markers [α4, β7, CCR6 (chemokine receptor 6), CCR5 (chemokine receptor 5), CD8α, CD94, CD161, and NKG2D (killer cell lectin-like receptor subfamily K, member 1)] and two of five cytokines (TNFα and IFNγ) followed bimodal distributions. For each of these 10 markers, we fitted a two-component mixture model. Using the mean and SD of the pooled

distribution, we calculated the proportions of iNKT cells belonging to the positive component in each sample, using maximum a posteriori estimation (*SI Appendix*). CCR4 followed a trimodal distribution, which we fitted with a three-component mixture model. We considered the sum of the higher two components to be the positive portion. Two surface markers [CD103 and IL23R (interleukin 23 receptor)] and three intracellular markers [IL4 (interleukin 4), IL13 (interleukin 13), and IL17A (interleukin 17A)] showed negligible staining in all $CD3\varepsilon^+$ cells. These five markers were excluded from subsequent expression analyses.

After assessing the global pattern of phenotypic marker expression among the 110 subjects, we then applied principal component analysis to look for general trends. We observed that the first principal component captured 31.8% of the total variation (*SI Appendix*, Fig. S9) and correlated tightly with the proportion of $CD4^+$ iNKT cells ($r = -0.70$; $P < 10^{-17}$; Fig. 3*F*). We then examined the expression level of individual markers in all iNKT cells and confirmed that each was correlated with the proportion of $CD4^+$ iNKT cells, indicating biased expression in either the $CD4^+$ or the $CD4^-$ subset (Fig. 3*G* and Table 1). Specifically, CCR4 was preferentially expressed by the $CD4^+$ subset, whereas all other surface markers were $CD4^-$-biased. Similarly, functional markers also showed iNKT subtype bias, where $CD4^-$ iNKT cells released much higher levels of TNFα and IFNγ on PMA-ionomycin stimulation (Table 1). These results suggest that variation in iNKT cell abundance, phenotypic marker expression, and functional response are all captured by CD4 expression, which is therefore a critical biomarker for iNKT function.

Discussion

In this study, we profiled human iNKT cells, a rare immune cell type, in 110 samples of peripheral blood. In this large cohort, we showed that the quantity of iNKT cells was low but variable in the population, showing increased quantity in females and a decreased quantity with age. Subsequently, we extracted patterns of expression of surface phenotypic markers and intracellular cytokines, observing differences between $CD4^+$ and $CD4^-$ iNKT subsets. By applying X-Cyt to characterize iNKT cells, we demonstrated the potential for robust and efficient automated population identification in a large-scale immunoprofiling study.

X-Cyt reliably discovers targeted populations with important advantages in terms of consistency and speed, which result from user guidance and template-guided partitioning. We make the distinction between the goal of X-Cyt and that of existing automated cytometric analysis tools that are, in general, designed for exploratory studies. In exploratory studies, for example, those aiming to map the developmental lineage of cell populations are defined de novo in each sample. However, targeted studies focus on a specific cell type, often in large sample sizes. In such studies, we can assume samples in a cohort share a population structure defined by selected markers. X-Cyt allows the user to choose markers for defining cell types, the sequence of partitioning, and the resolution at which to partition, thus catering the analysis to the original intent of the experiment.

Both biological and technical variations often create notable shifts in fluorescence intensities, which complicate batch data analysis. However, the shifts rarely alter the relative spatial arrangement of populations. X-Cyt uses a template to capture this conserved structure and uses an EM algorithm to optimize the fit for each sample independently, which gives the method substantial tolerance for intensity shifts. Via EM, the corresponding populations across samples are allowed to vary from the template in terms of the "site" (the location parameter), "shape" (the covariance matrix), and "size" (the mixing proportion). In *SI Appendix*, Fig. S10, we illustrate samples in which a gate (i.e., in FlowJo) requires manual adjustment in each sample, but X-Cyt automatically detects the shifted location via parameter optimization.

The use of a template confers two additional advantages over de novo clustering. First, the template serves as a guide for indexing emerging clusters (e.g., the T_{EM} and T_{CM} clusters are indexed as clusters 1 and 5, respectively, in every sample), which eliminates the need for a separate alignment step that could potentially introduce additional error. If a population is present in the template but missing from a given sample, no event in the sample will be assigned, and its proportion in that sample becomes "0." Next, by initializing the parameters to a close approximation of the optimal solution, the number of iterations needed to reach convergence in the EM algorithm decreases by several orders of magnitude, substantially reducing computation time. To demonstrate, we compared the run times of clustering, using X-Cyt with and without initialization by a template. Using ~200 megabytes of physical memory, X-Cyt was able to partition the $CD4^+$ T-cell subset (~8,000 cells) into four clusters using three markers (CD62L, CD45RA, and CD45RO) in about 5 s per sample compared with about 0.4 s with a template. In the MAIT cell study containing 500,000 cells per sample, X-Cyt partitioned a random subset of 100,000 cells into four clusters in two dimensions in ~45 s without a template. In contrast, clustering a full sample of 500,000 cells required about 10 seconds when guided by a template.

Emerging cytometric technologies, such as cytometry by time of flight, can simultaneously measure more than 30 markers in a cell (26, 27) and facilitate precise characterization of the human immune system. For these studies, robust and versatile analytical methods will become indispensible. In addition to algorithms well suited for exploratory studies, there is a strong need for tools to replace gating-based manual analysis when conducting focused characterization of targeted cell types. X-Cyt presents an efficient and robust method for analyzing such high-throughput immunoprofiling data sets.

Table 1. Differential expression of surface markers in $CD4^+$ and $CD4^-$ iNKT cells

Marker	Δ($CD4^-$ - $CD4^+$)	*P* value
NKG2D	66.5%	2.8×10^{-19}
α4 integrin	58.4%	8.6×10^{-17}
CCR5	58.0%	2.7×10^{-18}
CCR6	38.4%	4.6×10^{-15}
CD161	36.8%	9.5×10^{-17}
CD8	31.5%	2.0×10^{-19}
CD94	30.1%	3.6×10^{-18}
β7 integrin	21.6%	1.5×10^{-7}
CCR4	−33.4%	4.9×10^{-18}
ΔTNFα*	27.8	2.6×10^{-9}
ΔIFNγ*	23.3%	1.4×10^{-7}

Δ denotes the differential expression upon administration of PMA-ionomycin vs. DMSO (PMA-ionomycin – DMSO).

Methods

Overview of Flow Cytometry Data Sets. ***$CD4^+$ Memory T-cell Subset Study.*** PBMCs were isolated from the whole blood of 236 healthy volunteers and depleted of non-$CD4^+$ T cells, using magnetic-activated cell sorting kits. Cells were then stained with fluorophore-conjugated antibodies against CD45RO, CD45RA, and CD62L.

Mucosal Associated Invariant T-cell Study. PBMCs were isolated from the whole blood of 40 healthy. Cells were then stained with fluorophore-conjugated antibodies against CD3, CD45, Vα7.2, and CD161.

iNKT Cell Study. PBMCs were obtained from the blood of 110 healthy volunteers. From PBMCs, iNKT cells were stained with fluorophore-conjugated antibodies against 14 cell surface markers and five intracellular cytokines, after PMA-ionomycin administration.

Detailed experimental protocols can be found in *SI Appendix, Experimental Methods*. Flow cytometric data were exported from FlowJo as text files after compensation and transformation in the "channel number" format.

Data Partitioning. X-Cyt partitions the data with user-designated differentiation markers. At each step of partitioning, the user can opt to use multivariate mixture-modeling or univariate cutoffs to identify outliers.

Multivariate Mixture Modeling. X-Cyt fits a given number of multivariate components to a sample via an EM algorithm, as previously described (5). The user can specify three input parameters: the markers used for clustering, the number of expected clusters, and the distribution type (multivariate normal, skew-normal, *t*, or skew-*t*; default is normal). Given *m* differentiation markers and *g* clusters, X-Cyt models a given sample as an *m*-variate mixture of *g* components using EM algorithm initiated by *k*-means clustering. On convergence, each cluster is described by a location parameter, a covariance matrix that describes its multidimensional shape, as well as a mixing proportion. Each event in the sample is assigned membership to one of the *g* clusters. We describe the multivariate normal distribution and the EM algorithm in *SI Appendix*.

Trial Analysis in Representative Samples. Using a small test set of random samples, the user sets up the partitioning scheme, optimizes input parameters, and chooses a template.

Select Test Samples. The user randomly selects a small subset of samples from the cohort to serve as test samples. Assuming that a target population is present in *f*% of all samples, the chances of encountering this population at least once among *N* random test samples at a 95% confidence is described by $(1 - f)^N = 0.05$. Therefore, there is a 95% chance that a 20% population will be observed at least once in 14 samples, a 50% population will be observed at least once in five samples, and a 90% population will be observed at least once in two samples. A table of recommended size of test-sets is available in *SI Appendix*, Table S3.

Select Differentiation Markers. The user should select the subset of markers that most efficiently distinguishes the cell type or types of interest from the rest of the events. The user often already has selected a set of differentiating markers while designing the marker panel for the experiment. For example, one would use CD3, CD45 (RA/RO), and CD62L to identify naive T cells in PBMCs. In contrast, if certain markers in the panel are assayed for the purpose of characterizing the phenotype and function, rather than differentiating cell types (e.g., certain intracellular cytokines and chemokine receptors), they should be excluded in this step.

Select the Number of Clusters. In the trial analysis, the user should evaluate the partitioning result of each sample from testing a range of *g*. For example, given *k* differentiation markers, it is reasonable to test a range of *k* to 2^k clusters. The user reviews the output clusters and defines the optimal *g* as one that most accurately captures the population of interest as one cluster without including undesired events or spuriously splitting the population into more than one cluster. Often, the target population remains stable as one coherent cluster over a small range of *g*. In this case, the lowest *g* is recommended to minimize computation time.

SD Thresholds. For rare cell types with extreme intensities in one marker *M*, it is most efficient to first partition the sample to coarser-grained clusters, based on other differentiation markers, and then distinguish the rare events in *M* using a cutoff by SD threshold. For example, to identify Tregs from PBMCs using a panel of CD3, CD4, CD25, and Foxp3 (forkhead box P3), one may first partition all events with CD3, CD4, and CD25 to extract $CD3^+CD4^+CD25^+$-activated T cells and then apply a threshold cutoff in Foxp3 to extract the Tregs.

Guided Cohort Analysis by Template. X-Cyt uses a user-approved template selected from the trial analysis to guide the partitioning of all subsequent samples. The template serves as the initial parameters and as the indexing guide. Instead of using a *k*-means initialization, X-Cyt initializes each sample's mixture model parameters to that of the template's, on which EM algorithm iterates and converges quickly.

Phenotypic and Functional Marker Characterization. For each marker, we report the percentage of cells with positive expression. We construct a pooled sample of random events from all samples in the data set, thus establishing a "reference" fluorescence intensity distribution for each marker.

For multimodal markers, we assume the intensity distribution comprising *n* normal components. We fit a 1D mixture model on the pooled sample and then estimate the proportion of cells in each cell population positive for the marker in each sample. Details are provided in *SI Appendix*.

For a unimodal marker, we specify a SD threshold. We report the proportion of cells that express the marker above the threshold.

ACKNOWLEDGMENTS. We thank Dr. Joshua Randall for assistance in preparing the software package. We thank Dr. Vijay Kuchroo for helpful discussions. We would like to acknowledge the PhenoGenetic Project at Brigham and Women's Hospital for providing blood samples. Support provided by the IDEA^2 program at the Massachusetts Institute of Technology (to X.H.); a career development award from the American Academy of Allergy, Asthma & Immunology ARTrust (to P.J.B.); research grants from the National Institutes of Health (AI063428, AI028973, and DK057521 to M.B.B.) and the American Diabetes Association (7-12-IN-07 to M.B.B.); and research grants from the Arthritis Foundation, the National Institutes of Health (U01HG0070033 and 5K08AR055688), and the Harvard University Milton Foundation (to S.R.).

1. Blumberg RS, Dittel B, Hafler D, von Herrath M, Nestle FO (2012) Unraveling the autoimmune translational research process layer by layer. *Nat Med* 18(1):35–41.
2. Perfetto SP, Chattopadhyay PK, Roederer M (2004) Seventeen-colour flow cytometry: Unravelling the immune system. *Nat Rev Immunol* 4(8):648–655.
3. Cheung RK, Utz PJ (2011) Screening: CyTOF-the next generation of cell detection. *Nat Rev Rheumatol* 7(9):502–503.
4. Aghaeepour N, et al.; FlowCAP Consortium; DREAM Consortium (2013) Critical assessment of automated flow cytometry data analysis techniques. *Nat Methods* 10(3): 228–238.
5. Pyne S, et al. (2009) Automated high-dimensional flow cytometric data analysis. *Proc Natl Acad Sci USA* 106(21):8519–8524.
6. Lo K, Brinkman RR, Gottardo R (2008) Automated gating of flow cytometry data via robust model-based clustering. *Cytometry A* 73(4):321–332.
7. Finak G, Bashashati A, Brinkman R, Gottardo R (2009) Merging mixture components for cell population identification in flow cytometry. *Adv Bioinforma*, 10.1155/2009/247646.
8. Chan C, et al. (2008) Statistical mixture modeling for cell subtype identification in flow cytometry. *Cytometry A* 73(8):693–701.
9. Qian Y, et al. (2010) 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* 78(Suppl 1):S69–S82.
10. Aghaeepour N, Nikolic R, Hoos HH, Brinkman RR (2011) Rapid cell population identification in flow cytometry data. *Cytometry A* 79(1):6–13.
11. Sugár IP, Sealfon SC (2010) Misty Mountain clustering: Application to fast unsupervised flow cytometry gating. *BMC Bioinformatics* 11:502.
12. Naumann U, Luta G, Wand MP (2010) The curvHDR method for gating flow cytometry samples. *BMC Bioinformatics* 11:44.
13. Zare H, Shooshtari P, Gupta A, Brinkman RR (2010) Data reduction for spectral clustering to analyze high throughput flow cytometry data. *BMC Bioinformatics* 11:403.
14. Qiu P, et al. (2011) Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat Biotechnol* 29(10):886–891.
15. Bendelac A, Savage PB, Teyton L (2007) The biology of NKT cells. *Annu Rev Immunol* 25:297–336.
16. Brigl M, Brenner MB (2004) CD1: Antigen presentation and T cell function. *Annu Rev Immunol* 22:817–890.
17. Kronenberg M (2005) Toward an understanding of NKT cell biology: Progress and paradoxes. *Annu Rev Immunol* 23:877–900.
18. Gumperz JE, Miyake S, Yamamura T, Brenner MB (2002) Functionally distinct subsets of CD1d-restricted natural killer T cells revealed by CD1d tetramer staining. *J Exp Med* 195(5):625–636.
19. Lawson V (2012) Turned on by danger: Activation of CD1d-restricted invariant natural killer T cells. *Immunology* 137(1):20–27.
20. Snyder-Cappione JE, et al. (2010) A comprehensive ex vivo functional analysis of human NKT cells reveals production of MIP1-α and MIP1-β, a lack of IL-17, and a Th1-bias in males. *PLoS ONE* 5(11):e15412.
21. Carvalho KI, et al. (2010) Skewed distribution of circulating activated natural killer T (NKT) cells in patients with common variable immunodeficiency disorders (CVID). *PLoS ONE* 5(9): pii: e12652.
22. O'Reilly V, et al. (2011) Distinct and overlapping effector functions of expanded human CD4+, CD8α+ and CD4-CD8α- invariant natural killer T cells. *PLoS ONE* 6(12): e28648.
23. Pariente B, et al. (2011) Activation of the receptor NKG2D leads to production of Th17 cytokines in CD4+ T cells of patients with Crohn's disease. *Gastroenterology* 141(1): 217–226, 226.e211–e212.
24. Montoya CJ, et al. (2007) Characterization of human invariant natural killer T subsets in health and disease using a novel invariant natural killer T cell-clonotypic monoclonal antibody, 6B11. *Immunology* 122(1):1–14.
25. Kawano T, et al. (1997) CD1d-restricted and TCR-mediated activation of valpha14 NKT cells by glycosylceramides. *Science* 278(5343):1626–1629.
26. Ornatsky O, et al. (2010) Highly multiparametric analysis by mass cytometry. *J Immunol Methods* 361(1-2):1–20.
27. Bendall SC, et al. (2011) Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* 332(6030):687–696.