Hierarchical modeling for rare event detection and cell subset alignment across flow cytometry samples

Andrew Cron\textsuperscript{1}, Jacob Frelinger\textsuperscript{2}, Lynn Lin\textsuperscript{3}, Cécile Gouttefangeas\textsuperscript{4}, Satwinder K. Singh\textsuperscript{6}, Cedrik M. Britten\textsuperscript{5}, Marij J.P. Welters\textsuperscript{6}, Sjoerd H. van der Burg\textsuperscript{6}, Mike West\textsuperscript{1}, Cliburn Chan\textsuperscript{7,*}

Abstract

Flow cytometry is the prototypical assay for multi-parameter single cell analysis, and is essential in vaccine and biomarker research for the enumeration of antigen-specific lymphocytes that are often found in extremely low frequencies (0.1% or less). Standard analysis of flow cytometry data relies on visual identification of cell subsets by experts, a process that is subjective and often difficult to reproduce. An alternative and more objective approach is the use of statistical models to identify cell subsets of interest. Critical features currently deficient in such models are the ability to detect extremely low frequency event subsets without biasing the estimate by pre-processing enrichment, and the ability to align cell subsets across multiple data samples for comparative analysis. In this manuscript, we develop hierarchical modeling extensions to the Dirichlet Process Gaussian Mixture Model (DPGMM) approach we have previously described for cell subset identification, and show that the hierarchical DPGMM (HDPGMM) naturally generates an aligned data model across multiple samples that respects both sample-specific and batch-specific characteristics. HDPGMM also increases the sensitivity to extremely low frequency events by “borrowing strength” across multiple samples analyzed simultaneously. We validate the accuracy and reproducibility of HDPGMM estimates of antigen-specific T cells on clinically relevant reference peripheral blood mononuclear cell (PBMC) samples with known frequencies of antigen-specific T cells. These cell samples take advantage of retrovirally TCR-transduced T cells spiked into autologous PBMC samples to give a defined number of antigen-specific T cells detectable by HLA-peptide multimer binding. We also provide highly-optimized open source software that can take advantage of both multiple processors (using MPI) and massively parallel GPU cores (using CUDA) to accelerate the numerical computations. By addressing the rare subset and cell subset alignment problems, HDPGMM greatly increases the usefulness of automated flow cytometry data analysis and its relevance for immune monitoring and the discovery of immune-based biomarkers.

1. Department of Statistical Science, Duke University, Durham, NC 27708, USA
2. Program in Computational Biology and Bioinformatics, Duke University, Durham, NC 27708, USA
3. Population Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA
4. Interfaculty Institute for Cell Biology, Department of Immunology, Eberhard Karls University, Tuebingen, Germany
5. Translational Oncology at the Johannes Gutenberg-University Medical Center Mainz, 55131 Mainz, Germany
6. Dept of Clinical Oncology, Leiden University Medical Center, Leiden, the Netherlands
7. Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, NC 27710, USA
* Corresponding author: cliburn.chan@iduke.edu
**Author Summary**

The use of flow cytometry to count antigen-specific T cells is essential for vaccine development, monitoring of immune-based therapies and the search for immune biomarkers. Analysis of such data is challenging because antigen-specific cells are often present in frequencies of less than 1 in 1,000 peripheral blood mononuclear cells. Standard analysis of flow cytometry data relies on visual identification of cell subsets by experts, a process that is subjective and often difficult to reproduce. Consequently, there is intense interest in automated approaches for cell subset identification. One popular class of such automated approaches is the use of statistical mixture models. For rare cell subsets however, mixture models typically require pre-processing of the data to enrich for the target cell type, and this can lead to bias in the estimate. We propose a hierarchical extension to statistical mixture models that increases the sensitivity to rare events by “borrowing strength” across multiple samples analyzed simultaneously, and validate it on clinically relevant reference peripheral blood mononuclear cell (PBMC) samples with known frequencies of CD8 T cells engineered to express T cell receptors specific for the cancer-testis antigen (NY-ESO-1). The hierarchical approach also naturally leads to data alignment, enabling the direct comparison of cell subsets across multiple samples, greatly facilitating downstream analysis such as biomarker discovery.

**Introduction**

**Model-based analysis for cell subset identification in flow cytometry**

Flow cytometry is the prototypical assay for multi-parameter single cell analysis, and is essential in vaccine development, monitoring of T cell-based immune therapies and the search for immune biomarkers. In many clinical research applications, the cell subsets of interest are antigen specific T lymphocytes that are often found in extremely low frequencies (0.1% or less). These antigen-specific T cells can be detected using HLA-peptide multimers or by their expression of effector proteins upon specific antigen stimulation in intracellular staining (ICS) assays. Current methods of flow cytometry analysis rely on visual gating of cell events to identify and quantify cell subsets of interest. However, the choice of sequence for the dot plots (gating strategy) and where to draw the gating boundaries is highly dependent on assay protocols and operator experience and may not be easily reproducible, as illustrated in recent international proficiency panels [1,2].

There has therefore been increasing interest in the use of objective, automated methods for cell subset identification. One approach that we and others have promoted is the use of statistical models to estimate the data distribution [3–5], followed by a mapping of summaries of the statistical distribution to cell subsets of biological interest. This model-based approach tends to be more numerically intensive than other ad hoc approaches to data clustering, but as we have previously demonstrated, this can be overcome by exploiting the cheap massively parallel capabilities of modern graphical processing units (GPUs). Importantly, the model-based approach has the advantage of using a declarative probabilistic framework that can be extended using well-established and understood mechanisms to improve discriminative power. In particular, hierarchical models that incorporate information from both the individual and group levels
when fitted to batched flow cytometry data samples increase both interpretability and sensitivity. These hierarchical models increase interpretability by aligning clusters in a way that enables direct comparison of cell subsets across data samples, and increase sensitivity for detecting very low frequency cell subsets by “borrowing strength” across multiple samples. Hierarchical models thus significantly improve the ability of model-based approaches to detect extremely low frequency event subsets, and enable the comparative analysis that is essential to any downstream analysis of multiple data samples.

With this in mind, the developments reported here concern the implementation of a highly optimized non-parametric hierarchical Gaussian mixture model based on a Dirichlet process prior, and extensions of the basic model to identify and quantify rare cell subsets in flow cytometry data. Simulated data is first used to demonstrate the advantages of hierarchical models over conventional clustering approaches. This is followed by validation of the model on experimental samples, using recently developed retrovirally TCR-transduced T cells that are spiked into autologous peripheral blood mononuclear cell (PBMC) samples to give a defined number of antigen-specific T cells [6]. Finally, the reproducibility and accuracy of this approach for rare cell quantification is compared to that of manual analysis performed by a group of experienced flow cytometry users.

**Statistical mixture models**

The basic concept in model-based approaches is to consider events in a flow cytometry data set as being random samples drawn from a multi-dimensional probability distribution. The objective of analysis is then to define the probability distribution model and evaluate inferences over the model parameters based on fit to the specific data set. Statistical mixture models are a standard approach for the construction of the underlying distribution, using the sum of many simpler probability distributions (e.g. multivariate Gaussian, Student-t or skewed distributions) to approximate arbitrary multi-dimensional distributions. For biological interpretation, fitted models are then used for clustering, i.e. using statistical properties of individual events to assign them to biological cell subsets. For example, with statistical mixture models, this can be done by grouping events with the highest probability of coming from a specific mixture component together, or merging of multiple components using specified criteria such as having a common mode [7,8].

Of course, the number of distinguishable cell subsets and Gaussian components necessary to fit the model satisfactorily is not known in advance. To avoid having to specify the number of mixture components needed in the model, we use a nonparametric approach with a Dirichlet process prior in which the number of components necessary is directly estimated from the data [9]. Computationally, this is much more efficient than fitting multiple models with different numbers of components and testing with some penalized likelihood (e.g. Akaike or Bayesian information criteria) to choose the best model, as only a single model fit is performed. Since we use multivariate Gaussian distributions as components, the overall approach is described as a Dirichlet process Gaussian mixture model (DPGMM). DPGMM are extremely flexible models that can fit flow data from flow cytometry experiments using different antibody-fluorochrome labels (e.g. 4-color HLA-peptide multimer and 11-color intracellular staining (ICS) panels), and a natural evolution of the fixed $k$ Gaussian mixture models we originally proposed [3]. Finally,
while the model uses Gaussian components, cell subsets are identified with merged components using the consensus modal clustering strategy described in Methods. As a result, cell subsets can have arbitrarily complex distributions and are not restricted to symmetric Gaussian clusters.

**Limitations of clustering approaches**

Clustering methods applied to data samples independently face two major limitations. The first is that cluster labels are not aligned across data samples, posing a problem for comparing subsets across multiple samples which is usually the purpose of the original experiment. The second is that there are limits to the ability of clustering models to identify very rare event clusters due to *masking* by abundant event clusters. In particular, this makes it difficult to identify clusters matching only antigen-specific HLA-peptide multimer labeled or polyfunctional T cells in ICS assays that may be biologically meaningful at frequencies of 0.1% or lower. We show in this paper that both issues are successfully addressed by the use of hierarchical Dirichlet process Gaussian mixture models (HDPGMM).

**Hierarchical Dirichlet process for information sharing**

Hierarchical, or multi-level models, represent individual events in flow cytometry data as being organized into successively higher units. For example, individual events belong to a sample, and a sample may belong to a batch of samples. The critical idea is that cell subset phenotypes that are common across data samples can be used to inform and hence better characterize events in individual samples. We can model this by placing all data samples under a common prior, such that the mean and covariance in any of the individual sample Gaussian components are shared across all samples, but the weight (proportion) of the component in each sample is unique. As described by Teh et al (2006) [10], this can be achieved by using a set of random measures $G_j$, one for each data sample, where $G_j$ is distributed according to a sample-specific Dirichlet process $DP(\alpha_0j, G_0j)$. The sample-specific DPs are then linked by a common discrete prior defined by another $DP(\gamma, H)$. This hierarchical model leaves the cluster locations and shapes constant across datasets, and hence aligns the clusters in that the location of the normal components is common to all data samples.

The hierarchical DP mixture model allows information sharing over data sets. In other words, if a rare cell subtype is found in more than one of the samples, we share this information across the datasets to detect the subtype even though the frequency in a particular data set may be vanishingly small. HDPGMM thus increases sensitivity for clustering cell subsets that are of extremely low frequency in one sample but common to many samples or present in high frequency in one or more samples. In principle, there is no lower limit to the size of a cluster that can be detected in a particular sample, so long the algorithm is able to “borrow strength” from other samples as described. In practice, vanishingly small clusters (e.g. 3-5 events out of 100,000) require expert interpretation to distinguish background from signal, but it is not uncommon for biologically significant antigen-specific cells to be present at such frequencies.
Results

Simulation study

We illustrate the ability of hierarchical modeling to simultaneously overcome the problem of masking of rare event clusters and provide an alignment of cell subsets over multiple data samples. Four simulated data sets were created, each with up to 4 bivariate normal clusters in 4 quadrants. Clusters in each quadrant may have different means or covariance matrices, or be absent entirely; see Figure 1. We compared four different approaches to clustering the data – independent fitting of DPGMM to each data sample, using a reference data set, using pooled data, and using hierarchical modeling.

Independent fitting of DPGMM to simulated data

In Figure 1 row 1, DPGMM were independently fitted to each of the data samples and modal clustering performed on the posterior distribution averaged over post burn-in iterations. Events were assigned to the modal clusters for which the posterior probability was highest and color coded by the identity of the modal cluster. The first obvious issue with this approach is that modal cluster labels are not coherent over data samples as shown by the different assignments of similar cell subsets in different samples in Figure 2. Consequently, it is not possible to directly compare modal cluster frequencies across data sets without further post-processing. The second more subtle issue is that the small (5 event) cluster in data sample 3 (circled in red) has been masked by the large green cluster even though it matches the distinct blue cluster in data sample 1 and the red cluster in data sample 2 and should be interpreted as a separate cell subset.

Using a reference data set

In row 2, we fitted a DPGMM to sample 3 (reference data set), then used the posterior distribution found to classify events in all the other samples. While this ensures that all clusters are aligned across the data sets, it has several limitations. The first issue is the need to choose a specific reference data sample, which introduces an element of subjectivity. A more worrying issue is that differences in distribution across data samples are simply ignored, and this can result in artifacts as shown with data sample 1 and data sample 2, where there is mixing of the red/green clusters because the mean or covariance matrices of those clusters deviated from that of the reference data sample 3. Also, because the small cluster (circled in red) is masked in data sample 3, it is also missed in all the other samples. While another data sample could have been chosen as the reference, it is clear from inspection of the variation across the simulated data samples that no single reference can give a satisfactory result.

Using pooled data

In row 3, we fitted a DPGMM to pooled data from all four data samples. Pooled data is problematic because the resulting distribution is for an “averaged” data sample, and may result in the loss of information specific to a particular sample. We observe artifacts from clusters present in the pooled distribution but not in the specific sample in data sample 2 (green events in blue
cluster) and data sample 4 (red events in blue cluster). A subtle issue is the over-counting of red cluster events in data sample 3 (9 events circled in red) due to the excessive influence of the red clusters in data samples 1 and 2.

Hierarchical modeling

Finally, in row 4, we fitted a HDPGMM to all four data sets simultaneously with the consensus modal clustering approach to identify cell subsets as described in Methods. Clusters are aligned across data sets, there is no spurious mixing of clusters, and the rare event cluster in data sample 3 is correctly classified as having 5 events (circled in red).

Experimental study

To evaluate the utility of HDPGMM for identifying rare event clusters in real data, we analyzed reference cell samples containing a predefined number of T cells with known TCR specificity for the NY-ESO-1 cancer-testis antigen. TCR-transduced cells were added to autologous PBMC samples at concentrations of 0%, 0.013125%, 0.02625%, 0.0525%, 0.105% and 0.21%. There is also a small background contribution by antigen-specific T cells that are already present in the unspiked sample, which is estimated to be 0.01% using the mean frequency from expert manual gating by 10 flow practitioners. A total of 50,000 events was then collected from each sample for analysis. At the highest spike frequency, we would therefore expect to detect a maximum of 0.22%, or 110 antigen-specific T cell events out of 50,000 total events. This is a challenging clustering problem as the frequency of expected multimer-positive events is extremely low, but ideal for validation since the expected number of T cells that bind with high-affinity to the HLA-peptide multimer is known.

DPGMM and HDPGMM models were then separately fitted to these six data samples using the FSC, SSC, CD45, CD3 and HLA-multimer channels (5 dimensional), using a truncated Dirichlet process with 128 mixture components, 20,000 burn-in steps and 2,000 identified iterations to calculate the posterior distribution as described in Methods. After consensus modal clustering, the multimer positive clusters were visually selected and the frequency of multimer-positive events as a percentage of all 50,000 events calculated. A side-by-side comparison of manually gated, DPGMM and HDPGMM classifications is shown in Figure 2. The middle and right panels show a comparison of the cell labeling resulting from DPGMM and HDPGMM, highlighting the interpretability of the hierarchical modeling – cell subsets are consistently labeled across data samples, allowing direct comparison of any cell subset of interest, not just of the multimer positive events. For the detection of CD3+CD45+ multimer positive cells, the hierarchical model finds a more consistent cluster of events across all data samples, has fewer outlier events that are likely to be false positives (e.g. the CD45 negative to low events in the DPGMM fits shown in rows 1 and 4), and is also more sensitive for the samples with the lower spiked-in frequencies.

Finally, to evaluate the robustness of the frequency estimates, the fitting was repeated 10 times for each algorithm using different random number seeds, and the summary frequencies of multimer-positive events for the replicate runs is shown in Figure 3. HDPGMM (blue circles, right panel) estimates have equal or less variability at every spike dilution when compared with DPGMM (blue circles, left panel). In particular, HDPGMM is more accurate and robust than
DPGMM for very low spike dilutions (less than or equal to 0.02625%). HDPGMM estimates from different runs also compare favorably with manual gating estimates from 10 different human experts.

Discussion

We have shown that HDPGMM improves on fitting individual samples with DPGMM in two ways - 1) it aligns clusters, making direct comparison of cluster counts across samples possible, and 2) by sharing information across samples, it can identify biologically relevant cell subsets present at frequencies in the 0.01–0.1% range, since “real” cell subsets would naturally be expected to be present in multiple data samples. The hierarchical model is also preferable to using a reference data sample or pooling the data from all samples, since individual sample characteristics are lost with these alternative strategies.

One limitation of the HDPGMM model is that all the data to be fitted need to be simultaneously available. This is not an issue for most studies, but may be limiting for longitudinal studies that collect samples serially over an extended period where interim analyses need to be performed. Even in these cases, it may be useful to batch process cell samples in stages using a hierarchical model, then perform post-processing to align cell subsets over different stages. Because the HDPGMM “borrows strength” across data samples, cell subsets that are consistent across data samples will be extremely robust features in the posterior distribution. Hence, it is likely that features across batches will be more consistent and easier to align for HDPGMM-fitted batch samples than if every sample was independently fitted.

Technically, our implementation of the HDPGMM integrates several innovations necessary to make such hierarchical models a practical tool for flow cytometry analysis, including the use of a Metropolis-within-Gibbs step for sampling, an identification strategy to maintain consistent component labels across iterations that allows us to calculate the posterior distribution from multiple MCMC iterations, and a consensus modal map to merge components in such a way that non-Gaussian cell subsets are aligned across multiple data sets. To ensure scalability, we have implemented MPI and CUDA optimized code that can take advantage of multiple CPUs and GPUs from a cluster of machines to fit a single HDPGMM model to multiple data sets.

We provide software for HDPGMM fitting to flow cytometry data sets, together with pre-specified robust default parameters and hyper-parameters that make practical usage simple. In our experience, we have never needed to adjust these parameters for data sets ranging from 3-color to 11-color flow cytometry data sets. The only parameters we individually set are the number of burn-ins, the number of iterations to collect for the posterior distribution, and the maximal number of components for the truncated DP algorithm. These parameters are tuned mainly for computational efficiency since conservative defaults that would be expected to be effective in all common use cases can be given, with the trade-off being longer run times. Of course, as with all Bayesian models, informative priors (e.g. from fits to previously collected similar data samples) can be provided to greatly reduce the compute time.

The fitting of HDPGMM is computationally demanding but can be tremendously accelerated with cheap commodity graphics cards as previously described [11]. For example, running an MCMC sampler for 20,000 burn-in and 2,000 identified iterations to fit a 128-component
HDPGMM to the six multimer data sets shown in Figure 2 took less than 6 hours on a Linux workstation using a single NVidia GTX 580 card costing under USD 500. Open source code for fitting DPGMM and HDPGMM models to flow cytometry data is available from [http://code.google.com/p/py-fcm/](http://code.google.com/p/py-fcm/). The code is written in the Python programming language, and will run on regular CPUs, but is optimized for massively parallel computing using the CUDA interface (a suitable Nvidia GPU is required for CUDA).

In summary, we describe and provide highly optimized code for a hierarchical modeling extension to statistical mixture models that improves on the robustness, sensitivity and interpretability of model-based approaches for automated flow cytometry analysis. We demonstrate the consistency of frequency of HDPGMM estimates on reference data samples spiked with extremely low frequencies of antigen-specific cells, a scenario directly relevant to many clinical research studies in vaccine development, immune monitoring and immune biomarker discovery where the frequency of rare antigen-specific T cells is of interest.

**Materials and Methods**

**Hierarchical modeling**

**Dirichlet process mixture of Gaussians**

Assume we observe flow cytometry measurements \( x = \{x_1, \ldots, x_n\} \) where each \( x_i \) is a \( p \) dimensional vector. Let the probability density function for \( x \) be

\[
f(x|\Theta) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k)
\]

(1)

where \( \Theta \) is the complete set of parameters in the model, \( K \) is the number of possible clusters, and \( N(x_i|\mu_k, \Sigma_k) \) is the \( p \) dimensional multivariate normal density evaluated at \( x_i \) with mean \( \mu_k \) and covariance matrix \( \Sigma_k \). The mixture weights \( \pi_k \) are all positive and sum to one. The \( \pi_k \) are modeled as random probabilities from a so-called “stick-breaking” prior process. Specifically,

\[
\tilde{\pi}_k \sim \text{Beta}(1, \alpha) \quad \text{and} \quad \pi_k = \tilde{\pi}_k \prod_{l=1}^{k-1} (1 - \tilde{\pi}_l) \quad \text{for} \quad k = 2, \ldots, K - 1,
\]

(2)

where \( \text{Beta} \) denotes a beta distribution [12]. Note that \( \pi_1 = \tilde{\pi}_1 \) and \( \pi_K = 1 - \sum_{l=1}^{K-1} \pi_l \). A key advantage of the (truncated) Dirichlet process specification is that it results in automated inference on the number of clusters based on a pre-specified large value \( K \). That is, with such an encompassing \( K \), many of the \( \pi_k \) will be inferred as very close to zero, leaving a reduced set of effective clusters. A complete Bayesian model specification is completed by putting priors on \( \alpha, \mu, \) and \( \Sigma \).

An alternative and equivalent representation of (1) is to assume that for each observation \( x_i \) we have an unknown label \( z_i \). If we assume \( P(z_i = k) = \pi_k \) and \( x_i|z_i = k \sim N(\mu_k, \Sigma_k) \), marginalizing the \( z_i \) yields (1). This parametrization makes posterior computation more tractable, and inference about \( z_i \) is equivalent to inferring the cluster assignment for \( x_i \).
Hierarchical Dirichlet process mixture of Gaussians

We now generalize DPGMM to simultaneously classify T cells across multiple datasets. Assume we observe \( J \) different sets of FCM measurements \( x_j = \{x_{j1}, \ldots, x_{jn_j}\} \). Each dataset then has its own probability density function given by

\[
f_j(x|\Theta) = \sum_{k=1}^{K} \pi_{jk} N(x|\mu_k, \Sigma_k).
\] (3)

Note the primary difference between DPGMM and HDPGMM is \( \pi_{jk} \). The sets of cluster locations and shapes, \( \mu \) and \( \Sigma \), are assumed to be the same across datasets while the prevalence of the clusters \( \pi \) is allowed to vary across datasets. A similar two level parametrization holds here as in DPGMM. The approach now introduces the inherent, latent component indicators \( z_{ji} \) such that, for each observation \( i \) and component \( j \), \( P(z_{ji} = k) = \pi_{jk} \); this leads to conditional distributions \( x_{ji}|z_{ji} = k \sim N(\mu_k, \Sigma_k) \) and opens the path to routine computational methods.

Our interest is in extensions of this basic framework to hierarchical models on the \( \pi_{jk} \) that effectively picks the number of clusters in the model, but shares information across datasets to facilitate rare subtype discovery. Teh et al (2006) [10] give multiple representations for a hierarchical Dirichlet process for clustering across multiple datasets. We take their stick breaking approach:

\[
\tilde{\beta}_k \sim \text{Beta}(1, \alpha) \quad \text{and} \quad \beta_k = \tilde{\beta}_k \prod_{l=1}^{k-1} (1 - \tilde{\beta}_l) \quad \text{for} \quad k = 2, \ldots, K - 1.
\]

\[
\tilde{\pi}_{jk} \sim \text{Beta} \left( \alpha_0 \beta_k, \alpha_0 \left( 1 - \sum_{l=1}^{k} \beta_l \right) \right) \quad \text{and} \quad \pi_{jk} = \tilde{\pi}_{jk} \prod_{l=1}^{k-1} (1 - \tilde{\pi}_{jl}) \quad \text{for} \quad k = 2, \ldots, K - 1.
\] (4)

As before, \( \beta_1 = \tilde{\beta}_1, \beta_K = 1 - \sum_{l=1}^{K-1} \beta_l, \pi_{j1} = \tilde{\pi}_{j1}, \) and \( \pi_{jK} = 1 - \sum_{l=1}^{K-1} \pi_{jl}. \) Teh et al (2006) [10] show that this construction is equivalent to letting each dataset have its own Dirichlet process where the base measures also arise from a common base measure. We complete the prior specification by placing multivariate normal and inverse Wishart priors on each \( \mu_j \) and \( \Sigma_j \) respectively. We also use gamma priors on \( \alpha \) and \( \alpha_0 \) [9].

Posterior computation

We perform posterior inference by sampling via a Markov chain Monte Carlo (MCMC) algorithm using the latent classification variable \( z \). However, several challenges arise. First, the usual Gibbs sampling approach does not work because the conditional distributions of the \( \tilde{\beta}_k \) are difficult to work with. Crepe et al (2011) [13] use a similar model, but do not give details for sampling these key parameters. Secondly, the naming, or labeling, of the clusters is not well defined, so we need to deal with relabeling issues. Finally, the computation within each sampling iteration is very expensive.
**Metropolis within Gibbs**

Since the conditional distributions for \( \beta \) and \( \alpha \) are not conjugate, we propose a Metropolis within Gibbs approach. For each MCMC iteration, all other parameters are sampled via their full conditional distributions given in the appendix. We then propose a new \( \alpha' \) from a normal distribution centered at \( \alpha \) where we reflect negative values onto the positive half line and accept or reject the move according to the Metropolis Hastings (MH) ratio. We take the same approach for each \( \tilde{\beta}_k \) except that we reflect onto the unit interval. The variability in the random walks is tuned during the burn-in period to target a 50% acceptance rate supported by Gelman et al. [14]. Further details can be found in the appendix.

**Identification**

To address the label switching issue, we use the method of Cron and West [15] that maintains a coherent classification of the data across the MCMC iterates. This is enabled by defining a “reference” classification taken from the last iteration of the burn-in phase of the MCMC; this is simply the most likely cluster assignment for each event in all the datasets at that iterate. Then, at every further iteration the clusters are relabeled to minimize the misclassification rate when compared to the reference. This method is used because of its computational efficiency and good performance in other settings. Critically, this allows us to estimate the true posterior by component-wise averaging over multiple iterations after the burn-in phase, instead of using a point estimate as is typically done.

**GPU computation**

In each iteration of the MCMC, the multivariate normal distribution must be evaluated at every event (in every dataset) for each of the \( K \) clusters to get assignment probabilities. Without parallel computing, this takes the majority of the computation time. Therefore, we adapted the GPU computing ideas by Suchard et al. [11, 16] used in the “gpustats” python library to massively accelerate the computation. We also employed MPI techniques that use multiple GPUs simultaneously to optimize performance.

**Consensus modal clustering for cell subset identification and alignment**

As cell subsets may have non-Gaussian distributions, it is often necessary to merge several mixture components to represent a single cell subtype. An intuitively appealing concept is to cluster components together when the components share a common mode, since the mode is an objective feature of the posterior distribution that links multiple components. We have previously described efficient numerical methods for mode identification [8]; here we adapt the procedure to find a coherent modal assignment across data sets. We first create a consensus Gaussian mixture model distribution whose components have the same means and covariance matrices as the fitted HDPGMM model, but whose component weights are averaged over all data sets. The components in this consensus GMM are then merged via modal clustering to create a mapping of Gaussian components to modal clusters representing discrete cell subsets. The mapping is then used for all the fitted data sets, resulting in cell subset (modal cluster) alignment across
multiple data sets. Note that only the mapping of component to modal cluster is shared by all data sets, the component weights for each data set remain unique.

**Generation of experimental data**

The generation of the standard samples with a defined number of antigen-specific CD8 T cells spiked into autologous PBMC for use in HLA-peptide multimer has been described [6]. Briefly, Phytohemagglutinin (PHA; 0.5 µg/ml) and IL-2 (20 U/ml) stimulated HLA-A*0201 positive PBMC were retrovirally transduced with an HLA-A*0201 restricted NY-ESO157-165 specific TCR construct after the CD4 T cells were depleted using Dynabeads (Invitrogen). After 5 days, the transduced cells were harvested and purified using APC-conjugated NY-ESO-1 specific HLA multimer and magnetic cell sorting. Purified cells were clonally expanded, harvested and spiked at the desired percentage of NY-ESO-1 specific TCR expressing CD8 T cells into autologous PBMC. These samples were analyzed using a FACSCalibur (BD) flow cytometer and flow cytometry standard (FCS) format data files were collected for analysis of multimer positive events by manual gating performed independently by 10 experienced independently by 10 flow cytometry users as well as by the hierarchical modeling approach described in this manuscript.

**Acknowledgments**

Research partially supported by grants from the Wallace Coulter Foundation, the U.S. National Science Foundation (DMS 1106516), and the National Institutes of Health (P50-GM081883, RC1 AI086032). Any opinions, findings and conclusions or recommendations expressed in this work are those of the authors and do not necessarily reflect the views of the Wallace Coulter Foundation, the NIH or the NSF.

**References**

1. Britten C, Gouttefangeas C, Welters M, Pawelec G, Koch S, et al. (2008) The cimtmonitoring panel: a two-step approach to harmonize the enumeration of antigen-specific cd8+ t lymphocytes by structural and functional assays. Cancer Immunology, Immunotherapy 57: 289–302.

2. Welters MJP, Gouttefangeas C, Ramwadhdoebe TH, Letsch A, Ottensmeier CH, et al. (2012) Harmonization of the intracellular cytokine staining assay. Cancer Immunology, Immunotherapy : 1–12.

3. Chan C, Feng F, Oettinger J, Foster D, West M, et al. (2008) Statistical mixture modeling for cell subtype identification in flow cytometry. Cytometry Part A 73A: 693–701.

4. Pyne S, Hu X, Wang K, Rossin E, Lin T, et al. (2009) Automated high-dimensional flow cytometric data analysis. Proceedings of the National Academy of Sciences 106: 8519.

5. Lo K, Hahne F, Brinkman R, Gottardo R (2009) flowClust: a Bioconductor package for automated gating of flow cytometry data. BMC Bioinformatics 10: 145.
6. Singh S, Tummers B, Schumacher T, Gomez R, Franken K, et al. (2012) The development of standard samples with a defined number of antigen-specific t cells to harmonize t cell assays: a proof-of-principle study. Cancer Immunology, Immunotherapy: 1–13.

7. Finak G, Bashashati A, Brinkman R, Gottardo R (2009) Merging mixture components for cell population identification in flow cytometry. Advances in Bioinformatics 2009: Article ID 247646.

8. Chan C, Lin L, Frelinger J, Hebert V, Gagnon D, et al. (2010) Optimization of a highly standardized carboxyfluorescein succinimidyl ester flow cytometry panel and gating strategy design using discriminative information measure evaluation. Cytometry Part A 77: 1126–1136.

9. Escobar MD, West M (1995) Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association: 577–588.

10. Teh YW, Jordan MI, Beal MJ, Blei DM (2006) Hierarchical Dirichlet processes. Journal of the American Statistical Association 101: 1566-1581.

11. Suchard MA, Wang Q, Chan C, Frelinger J, Cron AJ, et al. (2010) Understanding GPU programming for statistical computation: Studies in massively parallel massive mixtures. Journal of Computational and Graphical Statistics 19: 419-438.

12. Ishwaran H, James L (2001) Gibbs sampling methods for stick-breaking priors. Journal of the American Statistical Association 96: 161-173.

13. Crepet A, Tressou J (2011) Bayesian nonparametric model for clustering individual co-exposure to pesticides found in the French diet. Bayesian Analysis 6: 127-144.

14. Gelman A, Roberts GO, Gilks WR (1996) Efficient Metropolis jumping rules. Bayesian Statistics 5: 599-607.

15. Cron AJ, West M (2011) Efficient classification-based relabeling in mixture models. The American Statistician 65: 16-20.

16. Suchard MA, Holmes C, West M (2010) Some of the What?, Why?, How?, Who? and Where? of graphics processing unit computing for Bayesian analysis. Bulletin of the International Society for Bayesian Analysis 17: 12-16.

**Figure Legends**

**Tables**
Figure 1. HDPGMM results in more accurate classification of events in simulated data than other statistical mixture model approaches. (Left) Row 1 shows independent fitting of DPGMMs to each data set; row 2 shows the use of reference posterior distribution from data set 3 to classify events in other data set; row 3 shows a DPGMM fitted to pooled data from all data sets; and row 4 shows fitting of an HDPGMM to all 4 other data sets. Results are described in the text. All models used a truncated DPGMM base with 16 components, a burn-in of 10,000 iterations, and sampling of 100 post burn-in iterations for the calculation of the posterior distribution. (Right) Contour plots of the log posterior distribution. The HDPGMM distributions (Row 4) are most similar to the independently fitted distributions (Row 1), with the advantage that the small cluster in data set 3 masked by its larger neighboring cluster on top has a distinct mode. In contrast, the reference and pooled distribution strategies have the exact same distribution for all data sets and lack the flexibility to model sample-specific features.
Figure 2. Comparison of manual, DPGMM and HDPGMM detection of rare antigen-specific events. The panels show the estimated frequencies (yellow boxes) of antigen-specific cells (large red dots) found using manual gating (left), DPGMM (middle) and HDPGMM (right) for data samples spiked with 0%, 0.013125%, 0.02625%, 0.0525%, 0.105% and 0.21% of retrovirally transduced T cells (from top to bottom). The red polygons in the left panel are gates used for identifying antigen-specific cells by manual gating; the exact shape, sequence and location of these gates is determined by the operator and may vary between different operators depending on their training, experience and expertise. With the DPGMM approach, cell subsets across the samples from top to bottom are not directly comparable as indicated by the event colors, posing a problem for quantification of the same cell subset in different samples. In contrast, with the HDPGMM approach, cell subsets are aligned and directly comparable across all samples. HDPGMM is also clearly more sensitive at detecting antigen-specific cells when the frequency is extremely low (first 3 rows). HDPGMM is also more consistent in labeling events across different samples, while DPGMM is prone to detect likely false positive antigen-specific cells that are CD45-low or negative (arrows in rows 1 and 4 of middle panel). HDPGMM improves on the accuracy and consistency because the model incorporates both sample-specific and group-specific information, in contrast to DPGMM which only has access to sample-specific information. For both DPGMM and HDPGMM, model fitting was done with an MCMC sampler running 20,000 burn-in and 2,000 averaged iterations.
Figure 3. HDPGMM estimates of rare antigen-specific cell frequencies are robust. For both DPGMM and HDPGMM, 10 MCMC runs with unique random number seeds were performed to evaluate the reproducibility of antigen-specific cell frequency estimates. Estimates of the antigen-specific frequencies form manual, DPGMM and HDPGMM approaches are shown as open blue circles, with the blue bar representing the mean of all 10 estimates at each spike frequency. The red crosses represent the “true” frequency of antigen-specific cells combining the known spiked-in frequencies and the average background from 10 manual evaluations. As shown in the figure, HDPGMM (right panel) estimates have equal or less variability at every spike dilution when compared with DPGMM (middle panel). In particular, HDPGMM is clearly more accurate and robust than DPGMM for very low spike dilutions (less than or equal to 0.02625%). For comparison, estimates from manual gating analyses performed independently by 10 experienced flow cytometry operators are shown in the left panel.