Automated Gating of Flow Cytometry Data via Robust Model-Based Clustering

Kenneth Lo,1* Ryan Remy Brinkman,2 Raphael Gottardo1

Abstract

The capability of flow cytometry to offer rapid quantification of multidimensional characteristics for millions of cells has made this technology indispensable for health research, medical diagnosis, and treatment. However, the lack of statistical and bioinformatics tools to parallel recent high-throughput technological advancements has hindered this technology from reaching its full potential. We propose a flexible statistical model-based clustering approach for identifying cell populations in flow cytometry data based on t-mixture models with a Box–Cox transformation. This approach generalizes the popular Gaussian mixture models to account for outliers and allow for nonelliptical clusters. We describe an Expectation-Maximization (EM) algorithm to simultaneously handle parameter estimation and transformation selection. Using two publicly available datasets, we demonstrate that our proposed methodology provides enough flexibility and robustness to mimic manual gating results performed by an expert researcher. In addition, we present results from a simulation study, which show that this new clustering framework gives better results in terms of robustness to model misspecification and estimation of the number of clusters, compared to the popular mixture models. The proposed clustering methodology is well adapted to automated analysis of flow cytometry data. It tends to give more reproducible results, and helps reduce the significant subjectivity and human time cost encountered in manual gating analysis.

Key terms

Box–Cox transformation; EM algorithm; mixture model; outliers; statistics; t distribution; flow cytometry; gating; clustering

Flow cytometry (FCM) can be applied to analyze thousands of samples per day. However, as each dataset typically consists of multiparametric descriptions of millions of individual cells, data analysis can present a significant challenge. As a result, despite its widespread use, FCM has not reached its full potential because of the lack of an automated analysis platform to parallel the high-throughput data-generation platform. As noted in a recent Communication to the Editor (1), in contrast to the tremendous interest in the FCM technology, there is a dearth of statistical and bioinformatics tools to manage, analyze, present, and disseminate FCM data. There is considerable demand for the development of appropriate software tools, as manual analysis of individual samples is error-prone, nonreproducible, nonstandardized, not open to reevaluation, and requires an inordinate amount of time, making it a limiting aspect of the technology (2–10).

One major component of FCM analysis involves gating, the process of identifying homogeneous groups of cells that display a particular function. This identification of cell populations currently relies on using software to apply a series of manually drawn gates (i.e., data filters) that select regions in 2D graphical representations of the data. This process is based largely on intuition rather than standardized statistical inference (3,11,12). It also ignores the high-dimensionality of FCM data, which may convey information that cannot be displayed in 1 or 2D projections. This
is illustrated in Supplementary Figure 1 with a synthetic data-
set, consisting of two dimensions, generated from a t-mixture
model (13) with three components. While the three clusters
can be identified using both dimensions, the structure is
hardly recognized when the dataset is projected on either
dimension. Such an example illustrates the potential loss of in-
formation if we disregard the multivariate nature of the data.
The same problem occurs when projecting three (or more)-
dimensional data onto two dimensions.

Several attempts have been made to automate the gating
process. Among those, the K-means algorithm (14) has found
the most applications (15–18). Demers et al. (17) have pro-
posed an extension of K-means allowing for nonspherical
clusters, but this algorithm has been shown to lead to per-
formance inferior to fuzzy K-means clustering (18). In fuzzy
K-means (19), each cell can belong to several clusters with dif-
f erent association degrees, rather than belonging completely
to only one cluster. Even though fuzzy K-means takes into con-
sideration some form of classification uncertainty, it is a heur-
istic-based algorithm and lacks a formal statistical foundation.
Other popular choices include hierarchical clustering algo-
rimths (e.g., linkage or Pearson coefficients method). How-
ever, these algorithms are not appropriate for FCM data, since
the size of the pairwise distance matrix increases in the order
of $n^2$ with the number of cells, unless they are applied to some
preliminary partition of the data (16), or they are used to clus-
ter across samples, each of which is represented by a few statis-
tics aggregating measurements of individual cells (20,21).
Classification and regression trees (22), artificial neural net-
works (23) and support vector machines (24,25) have also
been used in the context of FCM analyses (26–29), but these
supervised approaches require training data, which are not
always available.

In statistics, the problem of finding homogeneous groups
of observations is referred to as clustering. An increasingly
popular choice is model-based clustering (13,30–33), which
has been shown to give good results in many applied fields
involving high dimensions (greater than ten); see, for example
Refs. (33–35). In this paper, we propose to apply an unsup-
ervised model-based clustering approach to identify cell popu-
lations in FCM analysis. In contrast to previous unsupervised
methods (6–8,15–18), our approach provides a formal unified
statistical framework to answer central questions: How many
populations are there? Should we transform the data? What
model should we use? How should we deal with outliers (aberr-
ant observations)? These questions are fundamental to FCM
analysis, where one does not usually know the number of
populations, and where outliers are frequent. By performing
clustering using all variables consisting of fluorescent markers,
the full multidimensionality of the data is exploited, leading to
more accurate and more reproducible identification of cell
populations.

The most commonly used model-based clustering
approach is based on finite Gaussian mixture models
(13,31–33). However, Gaussian mixture models rely heavily
on the assumption that each component follows a Gaussian
distribution, which is often unrealistic. A common approach
is to look for transformations of the data that make the nor-


MATERIALS AND METHODS

Data Description

To demonstrate our proposed automated clustering, we
use the two publicly (44) available FCM datasets.

The Rituximab dataset. Flow cytometric high-content
screening (45) was applied in a drug-screening project to iden-
tify agents that would enhance the antilymphoma activity of
Rituximab, a therapeutic monoclonal antibody (46). One
thousand six hundred different compounds were distributed
into duplicate 96-well plates and then incubated overnight
with the Daudi lymphoma cell line. Rituximab was then added
to one of the duplicate plates, and both plates were incubated
for several more hours. In addition to cells treated with the
compound alone, other controls included untreated cells and
cells treated with Rituximab alone. During the entire culture
period, cells were incubated with anti-BrdU and the DNA binding dye 7-
AAD. The proportion of cells in various phases of the cell cycle
and undergoing apoptosis was measured with multiparameter
FACS analysis.

The GvHD dataset. Graft-versus-host disease (GvHD) occurs
in allogeneic hematopoietic stem cell transplant recipients
when donor–immune cells in the graft initiate an attack on the
skin, gut, liver, and other tissues of the recipient. It is one of
the most significant clinical problems in the field of
allogeneic blood and marrow transplantation. FCM was used to collect data on patients subjected to bone marrow transplant with a goal of identifying biomarkers to predict the development of GvHD. The GvHD dataset is a collection of weekly peripheral blood samples obtained from 31 patients following allogeneic blood and marrow transplant (47). Peripheral blood mononuclear cells were isolated using Ficoll-Hypaque and then cryopreserved for subsequent batch analysis. At the time of analysis, cells were thawed and aliquoted into 96-well plates at $1 \times 10^4$ to $1 \times 10^5$ cells per well. The 96-well plates were then stained with 10 different four-color antibody combinations. Stained cells could be stained in 96-well plates with optimally diluted fluorescently conjugated antibodies.

### Gaussian Mixture Models

The conventional model-based clustering approach is based on finite Gaussian mixture models (13,31–33), where each cluster can be described by a separate Gaussian distribution. Formally, given data $y$, with independent $p$-dimensional multivariate observations $y_1, y_2, \ldots, y_n$, the likelihood for a mixture model with $G$ components is

$$L(n, \ldots, \mu_G; \Sigma_1, \ldots, \Sigma_G; w_1, \ldots, w_G|y) = \prod_{i=1}^{G} \sum_{g=1}^{G} w_g \Phi_p(y_i|\mu_g, \Sigma_g), \quad \sum_{g=1}^{G} w_g = 1,$$

where $\Phi_p(|\mu_g, \Sigma_g)$ is the $p$-dimensional multivariate Gaussian distribution with mean $\mu_g$ and covariance matrix $\Sigma_g$, and $w_g$ is the probability that an observation belongs to the $g$th component. Estimates of the unknown parameters $\Psi = (\Psi_1, \ldots, \Psi_G)$, where $\Psi_g = (\mu_g, \Sigma_g, w_g)$ can be obtained conveniently using the Expectation-Maximization (EM) algorithm (32,48,49).

In EM, we first define the unobserved cluster membership associated with each observation $y_i$, as $z_{ig} = (z_{i1}, \ldots, z_{ig})$ with

$$z_{ig} = \begin{cases} 1 & \text{if } y_i \text{ belongs to cluster } g \\ 0 & \text{otherwise.} \end{cases}$$

The E-step of the EM algorithm requires computing $\tilde{z}_{ig} \propto E_q(Z_{ig}|y_i)$, which is interpreted as the posterior probability that $y_i$ belongs to cluster $g$:

$$\tilde{z}_{ig} = \frac{w_g \Phi_p(y_i|\mu_g, \Sigma_g)}{\sum_{j=1}^{G} w_j \Phi_p(y_i|\mu_j, \Sigma_j)}$$

The M-step is filled by the following closed-form expressions for the unknown parameters:

$$\hat{w}_g \leftarrow \frac{n_g}{n} ; \quad \hat{\mu}_g \leftarrow \frac{\sum_{i=1}^{n} z_{ig} y_i}{n_g};$$

$$\hat{\Sigma}_g \leftarrow \frac{\sum_{i=1}^{n} z_{ig} (y_i - \hat{\mu}_g)(y_i - \hat{\mu}_g)^T}{n_g}$$

where $n_g \equiv \sum z_{ig}$. The EM algorithm alternates between the E and M steps until convergence. Observation $y_i$ may then be assigned to cluster $g$ associated with the largest $\tilde{z}_{ig}$ value, which corresponds to the maximum a posteriori classification.

### t-Mixture Models

The multivariate t distribution. In the presence of outliers, Gaussian distributions might give poor representations of clusters due to the large influence of outliers. One strategy is to replace the Gaussian distribution with a t distribution, of which the heavier tail provides a mechanism to handle outliers. The t mixture likelihood can be written as in Eq. (1), where the Gaussian density is replaced by the $t$ density with mean $\mu$ ($v > 1$), covariance matrix $\Sigma$ ($v > 2$) and $v$ degrees of freedom given by

$$\varphi(y; \mu, \Sigma, v) = \frac{\Gamma\left(\frac{v+p}{2}\right)\Gamma\left(\frac{v}{2}\right)}{(\pi v)^{p/2} \Gamma\left(\frac{v}{2}\right)} \left\{ 1 + \frac{(y - \mu)\Sigma^{-1}(y - \mu)}{v} \right\}^{-\frac{v+p}{2}}.$$  

(4)

As in the Gaussian case, estimates of the unknown parameters $\Psi = (\Psi_1, \ldots, \Psi_G, v)$ where $\Psi_g = (\mu_g, \Sigma_g, w_g, v)$ can be obtained using the EM algorithm (40,50,51). The algorithm uses the fact that we can parameterize a $t$ distribution using a normal-gamma compound distribution.

### Maximum likelihood estimation for a t-mixture model.

In EM for $t$-mixture models, we define the unobserved cluster membership $z_{ig} = (z_{i1}, \ldots, z_{ig})$ as in the Gaussian case. To facilitate the formulation of the $t$ distribution, we also define the weights $u_i$s, coming from the normal-gamma compound parameterization, with

$$Y_i|u_i, z_{ig} = 1 \sim N(\mu_g, \Sigma_g/u_i)$$

(5)

independently for $i = 1, \ldots, n$, and $U_i \sim Ga(v/2, v/2)$. The advantage of writing the model in this way is that, conditional upon the $U$s, the sampling errors are again normal but with different variances, and estimation becomes a weighted least squares problem. Now the E-step requires computing $\tilde{z}_{ig} \equiv E_q(Z_{ig}|y_i)$ and $\tilde{u}_{ig} \equiv E_q(U_i|y_i, z_{ig} = 1)$:

$$\tilde{z}_{ig} \leftarrow \frac{w_g \varphi_p(y_i|\mu_g, \Sigma_g, v)}{\sum_{j=1}^{G} w_j \varphi_p(y_i|\mu_j, \Sigma_j, v)}$$

(6)

and

$$\tilde{u}_{ig} \leftarrow \frac{v + p}{v + (y_i - \mu_g)\Sigma_g^{-1}(y_i - \mu_g)}$$

(7)

which lead to the following closed form estimates for the unknown parameters during the M-step:
Robust Model-Based Clustering for FCM Data

Box–Cox Transformation

To handle transformation and outlier identification simultaneously, we propose a $t$-mixture model with Box–Cox transformation. The Box–Cox transformation (36) of an observation $y$ is defined as follows:

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log(y) & \lambda = 0 \end{cases}$$

where $\lambda$ is referred to as the Box–Cox parameter. The function stated in Eq. (9) is defined for positive values of $y$ only. In view of the occasional need to handle negative-valued data in FCM analysis, here we adopt a modified version (52) of the Box–Cox transformation which is also defined for negative values:

$$y^{(\lambda)} = \frac{\operatorname{sgn}(y)|y|^\lambda - 1}{\lambda}, \quad \lambda > 0.$$  

Note that the allowable range of $\lambda$ in Eq. (10) is changed to be strictly positive to avoid discontinuity across zero, which would occur if a negative value for $\lambda$ was used to transform data. When all data values are positive, this modified Box–Cox transformation is the same as the original version. In general, for multivariate data, we may specify a Box–Cox parameter for each dimension. However, in the context of FCM data, since different variables used in each stage of our sequential clustering (see below) share similar characteristics, it is reasonable to set the Box–Cox parameter common to all variables. When we allowed for different Box–Cox parameters for different variables, we found that the Box–Cox parameter estimates are of similar magnitudes, justifying the use of one Box–Cox parameter for all variables in each stage (data not shown).

While the E step remains basically the same, as given by Eq. (2), replacing $y_i$ with $y_i^{(\lambda)}$, the incorporation of the Box–Cox parameter slightly complicates the M step. No closed-form solution is available for $\lambda$, which needs to be estimated by some numerical optimization technique. Please see Section 1 of Supplementary Material for a detailed account of EM for Gaussian or $t$-mixture models with transformation selection.

In each case, the EM algorithm needs to be initialized. Here we have chosen to use the algorithm of Fraley (53) for initialization; see Section 2 of Supplementary Material for details.

Density Estimation

To visualize FCM data, it may be convenient to project high-dimensional data on 1D or 2D density plots. One such application can be found in the analysis of the GvHD data, in which cells selected through the CD3+ gate were projected on the CD4 and CD8β dimensions to produce contour plots (see Fig. 1 and Supplementary Fig. 4). Usually, nonparametric methods are applied to produce such plots. However, all nonparametric methods require a tuning parameter [e.g., bandwidth for kernel density estimation (54)] to be specified to control the smoothness of these plots, and different software packages have different default settings. In the model-based clustering framework, such plots can be easily generated at a very low computational cost once estimates of the model parameters are available. The degree of smoothness is controlled by the number of components, which is chosen by the Bayesian Information Criterion (BIC) (55). Please see Section 3 of Supplementary Material for more details.

Selecting the Number of Clusters

When the number of clusters is unknown, we use the BIC. For Gaussian mixture models, it is defined as

$$\text{BIC}_G = 2 \log L_G - K_G \log(n)$$  

where $L_G$ is the maximized likelihood value of Eq. (1) for a $G$-component Gaussian mixture model, $n$ is the sample size, and $K_G$ is the number of independent parameters. BIC would then be computed for a range of possible values for $G$ and the one with the largest BIC (or relatively close to it) would be selected.

The BIC formula introduced in Eq. (11) can still be used for $t$-mixture models even with Box–Cox transformation. Note that since we do not estimate the degrees of freedom parameter, a $t$-mixture model has the same number of parameters, $K_G$, as a Gaussian mixture model. However, when the Box–Cox transformation is included in the model, we have one more parameter.
Sequential Approach to Clustering

In practice, gating is often done on a preselected subset of the data chosen by projecting the data on the forward light scatter (FSC) and sideward light scatter (SSC) dimensions. These two variables, which measure the relative morphological properties (corresponding roughly to cell size and shape) of the cells, are often used to distinguish basic cell types (e.g., monocytes and lymphocytes) or to remove dead cells and cell debris. As a consequence, similar to Hahne et al. (56), we have adopted a sequential approach to clustering. We first use the FSC and SSC variables to cluster the data and find basic cell populations, and then perform clustering on one or more populations of interest using all other variables consisting of fluorescent markers. However, our methodology could also be applied to any subset or the entire set of variables.

RESULTS
Application to Real Datasets
The Rituximab dataset. We have reanalyzed a part of the Rituximab dataset using our sequential clustering approach. This data contains 1,545 cells and four variables: FSC, SSC, and two fluorescent markers, namely, 7-AAD and anti-BrdU. We compared the different models described in the Materials and Methods section (t mixture with Box–Cox, t mixture, Gaussian mixture with Box–Cox, and Gaussian mixture) with the results obtained through expert manual analysis using the commercial gating software FlowJo (Tree star, Ashland, Oregon) and the K-means clustering algorithm (14). As mentioned in the Materials and Methods section, we use a sequential approach where we first cluster the FSC vs. SSC variables to select basic cell populations (first stage), and then cluster the selected population(s) using all remaining variables (second stage).

Figure 2a shows the initial gating performed by a researcher using FlowJo on the FSC and SSC variables. To facilitate the comparison of our clustering approach with manual analysis at the second stage, we tried to mimic this analysis. In order to do so, we used a t-mixture model with Box–Cox transformation, fixing the number of components at one, and removed points with weights [as given by Eq. (7)] less than 0.5, corresponding to outliers (see Materials and
Methods for details). As shown in Figure 2, the selected cells are not exactly the same but close enough to allow us to compare our clustering approach to manual gating results when using the two fluorescent markers.

At the second stage, we compare the different clustering models on the selected cells. Since the number of clusters is unknown in advance, we make use of the BIC. The BIC curves shown in Supplementary Figure 2, corresponding to the different models, peak around three to four clusters, motivating us to examine the results obtained using three (Fig. 3) and four (Supplementary Fig. 3) clusters respectively. As expected, $K$-means performs poorly, as spherical clusters do not provide a good fit. Similarly, untransformed mixture models ($t$ and Gaussian), constrained by the assumption of elliptical clusters, are not flexible enough to capture the top cluster. Furthermore, Gaussian mixture models (even with Box–Cox transformation) are very sensitive to outliers, which can result in poor classification. For example, when four clusters are used, the Gaussian mixture model breaks the larger cluster into two to accommodate outliers, while the Gaussian mixture model with Box–Cox transformation also has a large spread-out cluster to accommodate outliers. Finally, Figure 3b and Supplementary Figure 3b show that our $t$-mixture model-based clustering approach with Box–Cox transformation can provide comparable results with the manual gating analysis by identifying three of the four clusters with well-fit boundaries. Note, however, that none of the four clustering methods detect the left rectangular gate seen on Figure 3a, which is most likely because of its lower cell density compared to the other gates and the lack of clear separation along the "7 AAD" dimension. This gate, which corresponds to apoptotic cells (46), contains a loose assemblage of cells located at the left of the three far right gates. Our methodology permits the identification of the three right clusters with well-fit boundaries, and thus could be combined with expert knowledge in order to identify apoptotic cells. For example, one could compute a one-dimensional boundary at the left-end border of the two largest clusters, and automatically label cells on the left of that line apoptotic.

Having shown the superiority of our clustering framework in terms of flexibility and robustness compared to common approaches, we now turn to a larger dataset to demonstrate further its capability.

The GvHD dataset. Two samples of the GvHD dataset (47) have been reanalyzed, one from a patient who eventually developed acute GvHD, and one from a control. Both datasets consist of more than 12,000 cells and four markers, namely, anti-CD4, anti-CD8β, anti-CD3, and anti-CD8, in addition to the FSC and SSC variables. One objective of the analysis is to look for the CD3+CD4+CD8β+ cells (47). To demonstrate the capability of our proposed automated clustering approach, we try to mimic the gating strategy stated in Ref. (47). Figures 1a–1c and Supplementary Figures 4a–4c shows the gating performed by an expert researcher using FlowJo.

In the initial gating, we first extracted the lymphocyte population using the FSC and SSC variables by applying a $t$-mixture model with Box–Cox transformation, fixing the number of clusters from one to eight in turn. Supplementary Figure 5a shows that the BIC for the positive sample has a large increase from three to four clusters and remains relatively constant afterwards, suggesting that a model fit using four clusters...
Figure 3. Second-stage clustering of the Rituximab data using all the fluorescent markers. (a) Four gates were drawn by a researcher to define four populations of interest. (b–f) Clustering was performed on the cells preselected from the first stage as shown in Figure 2b. The number of clusters was set to be three. (b–c) Points outside the boundary drawn in black have weights less than 0.5 and are labeled with “●” when t distributions were used. (d–f) For clustering performed without using t distributions, for comparison sake, boundaries are drawn in a way such that they correspond to the region of the same percentile which the boundaries drawn in (b–c) represent. Different symbols are used for the different clusters. The numbers shown in all plots are the percentages of cells assigned to each cluster. The K-means algorithm is equivalent to the classification EM algorithm (49,57) for a Gaussian mixture model assuming equal proportions, and a common covariance matrix being a scalar multiple of the identity matrix. The spherical clusters with equal volumes drawn in (f) correspond to such a constrained model.
is appropriate. Supplementary Figure 5b is the corresponding scatterplot showing the cluster assignment of the points on removing those with weights less than 0.5, regarded as outliers. It is clear that the region combining three of the clusters formed matches closely with the gate drawn by the researcher as shown in Figure 1a, corresponding to the lymphocyte population.

The next two stages in the manual gating strategy consist of locating the CD3⁺ cells by placing a range gate in the CD3 density plot (Fig. 1b), and then identifying the CD3⁺CD4⁺CD8β⁺ cells through the upper right gate in the CD4 vs. CD8β contour plot (Fig. 1c). When applying our proposed clustering approach, we can combine these two stages by handling all the variables consisting of fluorescent markers at once, fully utilizing the multidimensionality of FCM data.

The fitted model with 12 clusters seems to provide a good fit as suggested by the BIC (Fig. 4a). We compared our results with those obtained through the manual gating approach by first examining the estimated density projected on the CD3 dimension. The unimodal, yet skewed, density curve suggests that it is composed of two populations with substantially different proportions superimposed on each other (Fig. 1e). At a level of around 280, we can well separate the 12 cluster means along the CD3 dimension into two groups, and use the group with high cluster means in the CD3 dimension to represent the CD3⁺ population. The unimodal nature of the density curve (Figs. 1b and 1e) implies that the two underlying populations somewhat mix together, and therefore setting a fixed cutoff to classify the cells is likely inappropriate. The merit of our automated clustering
approach is shown here, that, instead of setting a cutoff, it makes use of the information provided by the other dimensions to help classify the cells into CD3⁺/CD3⁻ populations. The group with high cluster means in the CD3 dimension consists of five clusters, and among these five clusters, we can easily identify the two clusters at the upper right in the CD4 vs. CD8β scatterplot (Fig. 4b) as the CD3⁺CD4⁺CD8β⁻ population.

We have applied the same strategy to the control sample; see Supplementary Figure 4 and Figures 4c–4d). Figure 4c suggests that, this time, only seven clusters are necessary as the BIC is relatively flat after that. The associated gating results for the control sample is characterized by an absence of the CD3⁺CD4⁺CD8β⁻ cells, a distinct difference from the positive sample. This feature is also captured using our automated clustering approach; the fitted model contains no clusters at the upper right of the CD4⁺ vs. CD8β⁺ scatterplot (Fig. 4d). This cell population was of specific interest, as it was identified vs. CD8⁺ as one possibly predictive of GvHD, based on the manual gating analysis (47).

**Simulation Studies**

We have conducted a series of simulations to study the performance of different model-based clustering approaches under different model specifications. Model performance is compared using the following two criteria: (a) the accuracy in cluster assignment; (b) the accuracy in selecting the number of clusters. We performed two simulation studies, one where we set the dimension to two resembling the Rituximab dataset, and one where the dimension was set to four resembling the GvHD dataset. In each case, we generated data from each of the following models: t-mixture with Box–Cox, t-mixture, Gaussian mixture with Box–Cox, and Gaussian mixture, using the parameter estimates obtained at the second stage in the Rituximab and GvHD (positive sample) analyses. For the GvHD, to reduce computational burden, we only selected the five clusters with the largest means in the CD3 dimension, corresponding to the CD3⁺ population. We refer to the simulation experiments as the Rituximab and the GvHD settings, respectively. We fixed the number of cells at 500 and generated 1,000 datasets under each of the aforementioned models. To study the accuracy in selecting the number of clusters using BIC, we generated 100 datasets from the same GvHD setting with 1,000 cells. Here, we used 1,000 cells to avoid numerical problems with small clusters when the number of clusters used is significantly larger than the true number, while we decreased the number of datasets to 100 because of the increase in computation when estimating the number of clusters.

**Classification results.** The four clustering methods in comparison were applied to each of the 1,000 datasets generated from each model. Model fitting was done by presuming that the number of clusters is known, i.e. four clusters for the Rituximab setting and five for GvHD. We compared the models via misclassification rates, i.e., the proportions of cells assigned to incorrect clusters. When computing the misclassification rates, all permutations of the cluster labels were considered, and the lowest misclassification rate was determined.

The scatterplot of one of the datasets (GvHD setting) generated from the t-mixture model with Box–Cox transformation can be found in Supplementary Figure 7. Overall results are shown in Table 1. As expected, the Gaussian mixture models perform poorly when data were generated from the t-mixture models because of a lack of mechanisms to handle outliers. When a transformation was applied during data generation, the mixture models without Box–Cox transformation fail to perform well. On the contrary, the flexibility of the t-mixture model with Box–Cox transformation does not penalize too much for model misspecification. This is illustrated by the results from the GvHD setting: the t-mixture model with Box–Cox transformation gives the lowest misclassification rates when the true model is instead the t-mixture model without transformation or the Gaussian mixture model with Box–Cox transformation.

<table>
<thead>
<tr>
<th>Model used to generate data under the Rituximab setting</th>
<th>Model used to fit data</th>
<th>t + Box–Cox</th>
<th>t</th>
<th>Gaussian + Box–Cox</th>
<th>Gaussian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model used to generate data under the GvHD setting</td>
<td>t + Box–Cox</td>
<td>0.187</td>
<td>0.211</td>
<td>0.279</td>
<td>0.251</td>
</tr>
<tr>
<td>GvHD + Box–Cox</td>
<td>0.255</td>
<td>0.263</td>
<td>0.339</td>
<td>0.315</td>
<td></td>
</tr>
<tr>
<td>Gaussian</td>
<td>0.321</td>
<td>0.400</td>
<td><strong>0.251</strong></td>
<td>0.352</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.344</td>
<td>0.329</td>
<td>0.317</td>
<td><strong>0.301</strong></td>
<td></td>
</tr>
<tr>
<td>Model used to generate data under the Rituximab setting</td>
<td>t + Box–Cox</td>
<td>0.112</td>
<td>0.116</td>
<td>0.205</td>
<td>0.230</td>
</tr>
<tr>
<td>GvHD + Box–Cox</td>
<td>0.107</td>
<td>0.111</td>
<td>0.191</td>
<td>0.221</td>
<td></td>
</tr>
<tr>
<td>Gaussian</td>
<td>0.135</td>
<td>0.143</td>
<td>0.139</td>
<td>0.152</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.134</td>
<td>0.132</td>
<td>0.132</td>
<td><strong>0.126</strong></td>
<td></td>
</tr>
</tbody>
</table>

The best results are shown in bold.

**Selecting the number of clusters.** In this part of the study, the four models in comparison were applied to each of the 100 datasets generated, setting the number of clusters from 1 to 10 in turn. The number of clusters that delivered the highest BIC was selected. We compared the models via the mode and the 80% coverage interval of the number of clusters selected out of the 100 repetitions. As shown in Table 2, the t-mixture models can select the correct number of clusters in the majority of repetitions, even in case of model misspecification. In
In many FCM clustering applications, the number of clusters is usually unknown and requires estimation. There are several approaches for choosing the number of clusters in model-based clustering, including resampling, cross validation, and various information criteria (58). Our approach to the problem is based on the BIC, which gives good results in the context of mixture models (33,59). BIC is computationally cheap to compute once maximum likelihood estimation for the model parameters has been completed, an advantage over other approaches, especially in the context of FCM where datasets tend to be very large. While computationally cheap, BIC relies heavily on an approximation of marginal likelihoods, which might not be very accurate for some data. Currently, we are looking for alternatives, for example, the integrated completed likelihood (60), to improve the estimation of the number of clusters. Nevertheless, combined with expert knowledge, we view BIC as a useful tool that can provide guidance on choosing a reasonable value, as supported by our simulation study of assessing the accuracy in selecting the number of clusters.

There exist several modified versions of the Box–Cox transformation to handle negative-valued data, for example, the log-shift transformation, which was also proposed in the paper for the original Box–Cox transformation (36). The advantage of our choice, given by Eq. (10), is that while continuity is maintained across the whole range of the data, it retains the simplicity of the form of the transformation without introducing any additional parameters; when all data are positive, it reduces to the same form of the original Box–Cox transformation.

It is well known that the convergence of the EM algorithm depends on the initial conditions used. A bad initialization may incur slow convergence or convergence to a local minimum. In the real-data examples and the simulation studies, we used a deterministic approach called hierarchical clustering (30,53) for initialization. We have found this approach to perform well in the datasets explored here. However, better initialization, perhaps incorporating expert knowledge, might
be needed for more complex datasets. For example, if there is a high level of noise in the data, it might be necessary to use an initialization method that accounts for such outliers; see Ref. (33) for an example.

To estimate how long it takes to analyze a sample of size typical for an FCM dataset, we have carried out a test run on a synthetic dataset, which consists of one million events and 10 dimensions. To complete an analysis with 10 clusters, it took about 20 min on a 3-GHz Intel Xeon processor with 2 GB of RAM. This illustrates that the algorithm should be quick enough for analyzing a large flow dataset. In general, the computational time increases linearly with the number of events and increases in the order of $p^2$ with the number of variables, $p$, per EM iteration. This is an advantage over hierarchical clustering in which the computational time and memory space required increase in the order of $n^2$. Making a hierarchical approach impractical when a sample of a moderate size, say, $>5,000$, is investigated. Meanwhile, we realize the need of high computational speed in FCM analysis, and are currently investigating means to speed up the EM algorithm for parameter estimation.

Like all clustering approaches, the methodology we have developed includes assumptions, which may limit the applicability of this approach, and it will not identify every cell population in every sample. If the distribution of the underlying population is highly sparse without a well-defined core, our approach may not properly identify all subpopulations. This is illustrated in the Rituximab analysis where the loosely structured group of apoptotic cells was left undetected. This in turn has hindered the capability of the approach from giving satisfactory estimates of the G1 and S frequencies for the identified clusters that would be desired for normal analysis of a 7-AAD DNA distribution for cultured cells. On the other hand, identification of every cluster may not always be important. The Rituximab study was designed as a high-throughput drug screen to identify compounds that caused a $>50\%$ reduction in S-phase cells (46), as would be captured by both the manual gates and our automated analysis should it occur. Furthermore, the exact identification of every cluster through careful manual analysis may not always be possible, especially in high-throughput experiments. For instance, in the manual analysis of the GvHD dataset, a quadrant gate was set in Figure 1c in order to identify the CD3$^+$CD4$^+$CD8$^β^+$ population, which was of primary interest. For convenience sake, this gate was set at the same level across all the samples being investigated. While five clusters can be clearly identified on the graph, it would be time consuming to manually adjust the positions of each of the gates for all the samples in a high-throughput environment, as well as identify all novel populations. Contrariwise, our automated approach can identify these clusters in short order without the need for manual adjustment. To complete the analysis of the GvHD dataset (>12,000 cells, six dimensions) to identify the CD3$^+$CD4$^+$CD8$^β^+$ population (Fig. 1), it took less than 5 min, using the aforementioned sequential approach to clustering, on an Intel Core 2 Duo with 2 GB of RAM running Mac OS X 10.4.10.

A rigorous quantitative assessment is important before implementing this, or any approach, as a replacement for expert manual analysis. The availability of a wide variety of example data would aid in the development and evaluation of automated analysis methodologies. We are therefore developing such a public resource, and would welcome contributions from the wider FCM community.

An R (61) package called flowClust is being developed to implement the clustering methodology proposed in this paper. The source code is built in C for optimal utilization of system resources and makes use of the BLAS library (62), which enables multithreaded processes. The R package will be available from Bioconductor (63) at http://www.bioconductor.org.

**ACKNOWLEDGMENTS**

Datasets were kindly provided by Maura Gasparetto and Clayton Smith. We thank Maura Gasparetto for her assistance on the FlowJo plots. RRB is an International Society for Analytical Cytology Scholar and a Michael Smith Foundation for Health Research Scholar.

**LITERATURE CITED**

41. Carroll RJ. Prediction and power transformation when the choice of power is restricted to a finite set. J Am Statist Assoc 1982;77:908–915.