BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–9

A framework for analytical characterization of monoclonal antibodies based on reactivity profiles in different tissues

Elizabeth Rossin [1,†], Tsung-I Lin [2,3,†], Hsiu J. Ho [4], Steven J. Mentzer [5] and Saumyadipta Pyne [6,7,*]

[1]Health Sciences and Technology, Harvard Medical School, Boston, MA, USA, [2]Department of Applied Mathematics and Institute of Statistics, National Chung Hsing University, Taichung, Taiwan, [3]Department of Public Health, China Medical University, Taichung, Taiwan, [4]Department of Statistics, Tunghai University, Taichung, Taiwan, [5]Laboratory of Adaptive and Regenerative Biology, Brigham & Women's Hospital, Harvard Medical School, Boston, MA, USA [6]Broad Institute of MIT and Harvard University, Cambridge, MA, USA, and [7]Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Monoclonal antibodies (mAb) are among the most powerful and important tools in biology and medicine. MAb development is of great significance to many research and clinical applications. Therefore, objective mAb classification is essential for categorizing and comparing mAb panels based on their reactivity patterns in different cellular species. However typical flow cytometric mAb profiles present unique modeling challenges with their non-Gaussian features and inter-sample variations. It makes accurate mAb classification difficult to do with the currently used kernel based or hierarchical clustering techniques.

Results: To address these challenges, in the present study we developed a formal 2-step framework called mAbprofiler for systematic, parametric characterization of mAb profiles. Further, we measured the reactivity of hundreds of new antibodies in diverse tissues using flow cytometry, which we successfully classified using mAbprofiler.

First, mAbprofiler fits a mAb's flow cytometric histogram with a finite mixture model of skew t distributions that is robust against non-Gaussian features, and constructs a precise, smooth and mathematically rigorous profile. Then it performs novel curve clustering of the fitted mAb profiles using a nonlinear regression of skew t mixture models that is robust against inter-sample variation. Thus mAbprofiler provides a new framework for identifying robust mAb classes, all well-defined by distinct parametric templates, which can be used for classifying new mAb samples. We validated our classification results both computationally and empirically using mAb profiles of known classification.

Availability and Implementation: A demonstration code in R is available at the journal website. The R code implementing the full framework is available from the author website – http://amath.nchu.edu.tw/www/teacher/tilin/software

Contact: Saumyadipta_Pyne@dfci.harvard.edu

†These authors contributed equally to this work.

*To whom correspondence should be addressed.

1 INTRODUCTION

Monoclonal antibodies (mAb) are among the most powerful, popular and important tools in a biomedical laboratory for probing different cellular types, states and functions. Research in the past decades has led to the development of large collections of mAb for specific binding to cell surface antigens, which facilitated purification and functional characterization of a variety of cell populations. It also unlocked the great potential of using mAb for therapy in many serious diseases such as cancer. Using platforms such as flow cytometry, one can measure quantitatively the binding of a mAb, in single-cell resolution, to the corresponding antigen whose expression may serve as a marker of cellular characteristics for a given specimen, see Herzenberg *et al*., 2001. Therefore it is important to characterize mAb reactivity patterns in different cell types and tissues with analytical precision and rigor so that both known and new mAb can be categorized and compared accurately and objectively.

MAb classification is of great practical importance to many fields in bio-medicine such as immunology, hematology, pathology and clinical immunotherapy. Large-scale attempts at analyzing mAb to identify new molecules were pioneered in the human leukocyte differentiation antigens (HLDA) workshops (see review in Zola and Swart, 2005) where the reactivities of large panels of mAb were measured against widely available cell lines. The reactivity was given a binary assignment compared to a negative control – either the antibody bound to its antigen on a given cell, or it did not – as measured by fluorescence intensity. The frequency with which this occurred over a cell population was then recorded, and hierarchical clustering was employed to group similar reactivity - thus was born the "Clusters of Differentiation" (CD) classes, widely used today to identify various cell populations (Bernard and Boumsell, 1984).

In recent years, the workshop approach for identifying new molecules to define cell types has become less applicable due to the current capabilities of molecular identification at gene level (Zola and Swart, 2005). An alternative approach for mAb characterization involves the use of primary cell populations that are derived

systematically from different tissues in selected species (e.g. Pratt *et al*., 2009). Typically mAb reactivity patterns, as measured with cytometric density histograms, can present jagged non-smooth curves with features in the form of peaks and shapes that are difficult to characterize analytically. Inter-sample variation in cytometric data makes the modeling problem even more challenging. Not only do these make accurate binary percent positive/negative calls harder but also render ineffective the current clustering approaches that are poorly-suited to model or classify such noisy curve profiles.

In general, analytical characterization of mAb reactivity patterns has received limited attention in statistics and computer science (Spiegelhalter and Gilks, 1987; Gilks and Shaw, 1995; Kim *et al.*, 2002; Zeng *et al*., 2002; Salganik *et al*., 2005; Zeng *et al*., 2007; Pratt *et al*., 2009 and references therein). As shown in Pratt *et al*. (2009), mAb classification faces technical challenges at multiple levels. Single parameter flow cytometric histograms used for measuring mAb reactivity often have multiple peaks with non-Gaussian features and irregular shapes. Few of the known algorithms can model the underlying distributions and their key features precisely and robustly. In addition, due to cytometric platform noise, the measurements of peak features tend to vary in terms of both significance and location, making direct comparison of samples challenging. Moreover, standard clustering approaches meant for multivariate points, such as hierarchical clustering, are not well-suited for grouping curves, which in this case represent histogram profiles. Histogram profiles, when viewed as points, can vary considerably with different choices of binning parameters, producing jagged patterns. Hence a new clustering approach is necessary that can robustly detect the characteristic features lying within every mAb's noisy curve profile, which is not merely a multivariate point. Simultaneously the approach must also account for the cytometric inter-sample variation among the curve features across mAb profiles to achieve accurate classification.

To address these challenges, in the present study we generated (a) new data for a large collection of mAb, and (b) developed mAbprofiler, a new general framework to characterize and cluster mAb profiles systematically and rigorously. More than 1000 subcloned murine hybridomas, made against sheep cell membrane antigens, were considered for analysis. A subset of mAb were selected for inclusion in this study based on their distinctive staining profile and surface expression in six diverse tissues (splenocytes, lymph node cells, alveolar macrophages, efferent lymphocytes, fetal thymus and thymocytes). Further, in mAbprofiler, we present a 2-step framework based on new parametric modeling algorithms. In the first step, it profiles every mAb defined by its flow cytometric histogram with a finite mixture model of skew t distributions that is robust against both outliers and asymmetry, which are often responsible for producing non-Gaussian features. An Expectation-Maximization (EM) algorithm is used for fitting every pattern with a smooth and mathematically rigorous profile that specifies all key features precisely, with the help of a probability density function. In the second step, for each tissue, mAbprofiler performs curve clustering of the fitted mAb profiles with a novel nonlinear regression of skew t mixture models that is robust against inter-sample variation. We used an effective criterion, the Jump Statistic, for model selection with the optimal number of clusters (or mAb classes). In addition to these robust tissue-specific mAb classes, our framework uncovered new group structures among profiles undetected by traditional approaches like hierarchical clustering. Importantly, mAbprofiler also generates class-specific parametric signatures that can be used for (a) comparing and categorizing mAb classes, and (b) classifying new mAb panels. Finally, we validated our classification results for different tissues both computationally and empirically using mAb profiles of previously known classification.

2 MATERIALS AND METHODS

MAb production and sample generation: We followed the system of mAb cloning and harvesting using protocols developed in our lab and described in Li *et al.* (1995), Pratt *et al*. (2009) and references therein. The panel of anti-sheep mAb were tested for reactivity against 6 different sheep tissues: splenocytes, lymph node cells, alveolar macrophages, efferent lymphocytes, fetal thymus cells and thymocytes. The cells were washed and fixed before they were analyzed with an Epics XL flow cytometer (Beckman Coulter, Miami, FL). After quality control by human expert inspection, we generated 561 mAb reactivity patterns categorized by tissue: spleen, lymph node, lung lavage, bone marrow, fetal thymus, and thymocytes. The fitted mAb profiles are available from the authors upon request.

Cytometric data preparation: Each flow cytometric sample was represented as a 3-column matrix, where columns contained forward-scatter, side-scatter and the fluorescence intensity for a particular fluorophore tagged antibody, and each row represented a single cell (or event). The data were pre-processed to remove debris and dead cells. 98% of the data consisted of experiments where 10,000 events were captured. After log10 transformation of the data, we performed multi-step cleanup and filtering: first, we removed points from samples whose maximum intensity value was populated by more than 25% of cells. Such spikes are signs of poor calibration during data acquisition. Second, we filtered zero fluorescent values which might also represent possible calibration problems. Finally, we removed extreme outliers in data (points more than 3 standard deviations away from the sample mean) that are most likely due to platform noise. After filtering, the median number of events for each sample was approximately 9390.

Step 1 of mAbprofiler (histogram profiling): After filtering, the resulting histogram of each antibody's fluorescence intensity was fitted with a finite mixture model of univariate skew t distributions using an EM algorithm described in Supplementary material. Since Bayesian Information Criterion (BIC) is known to select restrictive models which may be inadequate for feature detection in our non-smooth data, we instead used the well known Integrated Completed Likelihood (ICL) criterion for our model selection (McLachlan and Krishnan, 2008). ICL scores for optimal models showed no further improvement for most samples beyond 10 components, which was the maximum number of components fit by the mixture model. Figure 1 shows a sample histogram and the fitted profile with a grey curve. Clearly the optimal model produced a smooth and accurate profile, and all the significant features and their locations are captured and specified by the model parameters. Since the model is a univariate version of Pyne *et al*. (2009) approach, we have described it along with its EM algortithm in the Supplementary Information for completeness.

Step 2 of mAbprofiler (profile clustering): Here we present a new and robust model-based curve clustering approach using non-linear

regression of skew t mixture models. The model along with the EM algorithm for clustering of the mAb profiles (which were fitted in Step 1) are described below.

Following Azzalini and Dalla-Valle (1996), a random vector $\boldsymbol{Z}$ is said to follow the multivariate skew normal (MSN) distribution, denoted by $\boldsymbol{Z} \sim SN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$, if its density takes the form

$$f(\boldsymbol{z}) = 2\phi_p(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\Phi(\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{z} - \boldsymbol{\mu})),$$

where $\phi_p(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the pdf of p-variate normal distribution with mean vector $\boldsymbol{\Sigma}$ and covariance matrix $\boldsymbol{\Sigma}$, $\Phi(\cdot)$ represents the cdf (short for cumulative distribution function) of the standard normal distribution and $\boldsymbol{\Sigma}^{-1/2}$ is the square root matrix of $\boldsymbol{\Sigma}^{-1}$ satisfying $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{\Sigma}^{-1}$. If $\boldsymbol{\lambda} = \boldsymbol{0}$, then the distribution of $\boldsymbol{Z}$ reduces to $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

For the ease of theoretical and computational developments, Arellano-Valle *et al.* (2005) gave the following stochastic representation for the MSN distribution:

$$\boldsymbol{Z} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\big[\boldsymbol{\delta}|U_0| + (\boldsymbol{I}_p - \boldsymbol{\delta}\boldsymbol{\delta}^{\top})^{1/2}\boldsymbol{U}_1\big], \quad U_0 \perp \boldsymbol{U}_1, \quad (1)$$

where $\boldsymbol{\delta} = \boldsymbol{\lambda}/\sqrt{1 + \boldsymbol{\lambda}^{\top}\boldsymbol{\lambda}}$, $U_0 \sim N(0, 1)$, $\boldsymbol{U}_1 \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$ and the symbol '$\perp$' indicates independence.

The multivariate skew t (MST) distribution was proposed by Azzalini and Capitaino (2003), which is related to the MSN distribution as follows:

$$\boldsymbol{Y} = \boldsymbol{\mu} + \tau^{-1/2}\boldsymbol{Z}, \quad \boldsymbol{Z} \perp \tau, \quad (2)$$

where $\boldsymbol{Z} \sim SN_p(\boldsymbol{0}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ and $\tau \sim \mathrm{Gamma}(\nu/2, \nu/2)$. It follows from (2) that $\boldsymbol{Y} \mid \tau \sim SN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/\tau, \boldsymbol{\lambda})$. By Proposition 1 of Lin *et al.* (2007), integrating out τ from the joint density of $(\boldsymbol{Y}, \tau)$ yields the marginal density of $\boldsymbol{Y}$

$$f(\boldsymbol{Y}) = 2\, t_p(\boldsymbol{Y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}; \nu)\, T\Big(\boldsymbol{\lambda}^{\top}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{Y} - \boldsymbol{\mu})\sqrt{\frac{\nu + p}{\nu + \Delta}}\,\Big|\nu + p\Big), \quad (3)$$

where $t_p(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ denotes the pdf of p-variate t distribution with location vector $\boldsymbol{\mu}$, scale matrix $\boldsymbol{\Sigma}$ and degrees of freedom (df) $\nu \in (0, \infty)$; $T(\cdot|\nu)$ represents the cdf of Student's t distribution with df ν, and $\Delta = (\boldsymbol{Y} - \boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu})$. We shall denote $\boldsymbol{Y} \sim ST_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu)$ if $\boldsymbol{Y}$ has density given in (3).

Suppose we have a set of m input profiles $\{\boldsymbol{y}_j\}_{j=1}^m$ and each response vector $\boldsymbol{y}_j$ consists of n_j consecutive observations. We assume the response vector $\boldsymbol{y}_j \in \mathbb{R}^{n_j}$ is generated from

$$\boldsymbol{y}_j = \boldsymbol{\mu}(\boldsymbol{\beta}, \boldsymbol{x}_j) + \boldsymbol{\varepsilon}_j; \quad (j = 1, \ldots, m),$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients related to design matrix $\boldsymbol{X}_j = [\boldsymbol{x}_{j1} \cdots \boldsymbol{x}_{jn_j}]^{\mathrm{T}}$ with $\boldsymbol{x}_{jk} = (x_{jk1}, \ldots, x_{jkp})^{\mathrm{T}}$; $\boldsymbol{\mu}_j \equiv \boldsymbol{\mu}(\boldsymbol{\beta}, \boldsymbol{x}_j)$ is a vector-valued nonlinear (differentiable) function of $\boldsymbol{\beta}$ governing within-profile behavior, and $\boldsymbol{\varepsilon}_j$ is the resulting error vector equal to the discrepancy between $\boldsymbol{y}_j$ and $\boldsymbol{\mu}_j$.

A skew-t based nonlinear regression model is defined by assuming $\boldsymbol{\varepsilon}_j \sim St_{n_j}(\boldsymbol{0}, \boldsymbol{\Sigma}_j, \boldsymbol{\lambda}_j, \nu)$. Depending on the context, various assumptions should be made on $\boldsymbol{\Sigma}_j$ and $\boldsymbol{\lambda}_j$ to reduce the number of parameters to be estimated. Following De la Cruz (2008), we set $\boldsymbol{\Sigma}_j = \sigma^2\boldsymbol{I}_{n_j}$ to reflect the assumption of exchangeable errors among individuals and $\boldsymbol{\lambda}_j = \lambda\boldsymbol{1}_{n_j}$, where $\boldsymbol{1}_{n_j}$ is an $n_j \times 1$ unit vector for ensuring an identifiable model. In some circumstances, it is quite common to assume a time series like dependence structure for $\boldsymbol{\Sigma}_j$, which is a function of a small number of free parameters and depends on j only through its dimension n_j. Note that the skew t can be reduced to the following particular models that enhance the ease of implementation: the *skew normal* ($\nu \to \infty$), *Student's t* ($\lambda \to 0$) and the most common *normal* ($\lambda \to 0; \nu \to \infty$) models.

From (2), it can be verified that

$$\begin{aligned}
\boldsymbol{y}_j \mid (\gamma_j, \tau_j) &\sim N_n\Big(\boldsymbol{\mu}_j + \frac{\lambda\gamma_j}{(1 + n_j\lambda^2)}\boldsymbol{1}_{n_j}, \\
&\qquad \frac{\sigma^2}{\tau_j}\,(\boldsymbol{I}_n + \lambda^2\boldsymbol{1}_{n_j}\boldsymbol{1}_{n_j}^{\top})^{-1}\Big), \\
\gamma_j \mid \tau_j &\sim TN\Big(0, \frac{\sigma^2}{\tau_j}\,(1 + n_j\lambda^2); (0, \infty)\Big), \\
\tau_j &\sim \Gamma\Big(\frac{\nu}{2}, \frac{\nu}{2}\Big). \qquad (4)
\end{aligned}$$

Applying Bayes' rule yields

$$\begin{aligned}
\gamma_j \mid (\tau_j, \boldsymbol{y}_j) &\sim TN\Big(A_j, \frac{\sigma^2}{\tau_j}; (0, \infty)\Big), \qquad (5) \\
f(\tau_j \mid \boldsymbol{y}_j) &= \frac{\Phi(\tau_j^{1/2}\sigma^{-1}A_j)}{T(c_{0j}|\nu + n_j)} \\
&\qquad \times g\Big(\tau_j\Big|\frac{\nu + n_j}{2}, \frac{\nu + \sigma^{-2}\Delta_j}{2}\Big), \qquad (6)
\end{aligned}$$

where $\Delta_j = \boldsymbol{\varepsilon}_j^{\top}\boldsymbol{\varepsilon}_j$, $A_j = \lambda\boldsymbol{1}_{n_j}^{\top}\boldsymbol{\varepsilon}_j$ and $c_{rj} = A_j[(\nu + n_j + r)/(\sigma^2\nu + \Delta_j)]^{1/2}$. According to (5) and (6), it suffices to compute the following conditional expectations:

$$\begin{aligned}
E(\gamma_j \mid \boldsymbol{y}_j) &= A_j + \sigma\Big(\frac{\nu + n_j - 2}{\nu + \sigma^{-2}\Delta_j}\Big)^{-1/2} \\
&\qquad \times \frac{t(c_{-2,j}|\nu + n_j - 2)}{T(c_{0j}|\nu + n_j)}, \\
E(\tau_j \mid \boldsymbol{y}_j) &= \frac{\nu + n_j}{\nu + \sigma^{-2}\Delta_j}\,\frac{T(c_{2j}|\nu + n_j + 2)}{T(c_{0j}|\nu + n_j)}, \\
E(\tau_j\gamma_j \mid \boldsymbol{y}_j) &= A_j\,E(\tau_j \mid \boldsymbol{y}_j) \\
&\qquad + \sigma\Big(\frac{\nu + n_j}{\nu + \sigma^{-2}\Delta_j}\Big)^{1/2}\frac{t(c_{0j}|\nu + n_j)}{T(c_{0j}|\nu + n_j)}, \\
E(\tau_j\gamma_j^2 \mid \boldsymbol{y}_j) &= \sigma^2 + A_j\,E(\tau_j\gamma_j \mid \tau_j, \boldsymbol{y}_j), \qquad (7)
\end{aligned}$$

where $t(\cdot|\nu)$ is the pdf of the Student's t distribution with df ν.

Finite mixture models are commonly used for model-based clustering (McLachlan and Basford, 1988; Banfield and Raftery, 1993). Let a curve profile be given by a sequence $\boldsymbol{y}_j$ of observations at n_j (time) points $\boldsymbol{x}_j$ and assumed to be generated by one and only one cluster (i.e. a mAb class). Then our goal is to partition $\{\boldsymbol{y}_j\}_{j=1}^m$ into g homogeneous groups (or classes). For notational convenience, let $\boldsymbol{\mu}_{ij} = \boldsymbol{\mu}(\boldsymbol{\beta}_i, \boldsymbol{x}_j)$, $\boldsymbol{e}_{ij} = \boldsymbol{y}_j - \boldsymbol{\mu}_{ij}$, $A_{ij} = \lambda_i\boldsymbol{1}_{n_j}^{\top}\boldsymbol{e}_{ij}$, $\Delta_{ij} = \boldsymbol{e}_{ij}^{\top}\boldsymbol{e}_{ij}$ and $c_{rij} = A_{ij}[(\nu_i + n_j + r)/(\sigma_i^2\nu_i + \Delta_{ij})]^{1/2}$ for $i = 1, \ldots, g$ and $j = 1, \ldots, m$. Define

$$\psi_{n_j}\big(\boldsymbol{y}_j\big|\boldsymbol{x}_j, \boldsymbol{\theta}_i\big) = 2\,t_{n_j}\big(\boldsymbol{y}_j\big|\boldsymbol{\mu}_{ij}, \sigma_i^2\boldsymbol{I}_{n_j}, \nu_i\big)\,T\big(c_{0ij}\big|\nu_i + n_j\big),$$

the density of a cluster-specific skew-t nonlinear regression model that relates $(\boldsymbol{y}_j, \boldsymbol{x}_j)$ to $\boldsymbol{\theta}_i = (\boldsymbol{\beta}_i, \sigma_i^2, \lambda_i, \nu_i)$.

The mixture model for profile clustering is written as:

$$\boldsymbol{y}_j \sim \sum_{i=1}^{g} w_i \, \psi_{n_j}\big(\boldsymbol{y}_j | \boldsymbol{x}_j, \boldsymbol{\theta}_i\big), \tag{8}$$

where w_i's are mixing proportions which are constrained to be non-negative and $\sum_{i=1}^{g} w_i = 1$ and $\boldsymbol{\Theta} = (w_1, \ldots, w_{g-1}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g)$ represents all unknown parameters. The observed data log-likelihood function of $\boldsymbol{\Theta}$ is

$$\ell(\boldsymbol{\Theta}|\boldsymbol{y}) = \sum_{j=1}^{N} \log f(\boldsymbol{y}_j|\boldsymbol{\Theta}). \tag{9}$$

In general, there are no explicit analytical solutions for computing the ML estimator of $\boldsymbol{\Theta}$. The EM algorithm (Dempster *et al.*, 1977) is considered as a standard tool when applied for mixture models. In the EM framework for supporting the interpretation of incomplete data, it is convenient to introduce a set of allocation variables $\boldsymbol{Z}_j = (Z_{1j}, \ldots, Z_{gj})^{\mathrm{T}}$, $j = 1, \ldots, m$. The element Z_{ij} is taken to be one or zero to indicate if $\boldsymbol{y}_j$ does or does not come from the i-th component. This implies that $\boldsymbol{Z}_j$ follows a multinomial distribution with 1 trial and cell probabilities $w_1, \ldots, w_g$, denoted by $\boldsymbol{Z}_j \sim M(1; w_1, \ldots, w_g)$. Then, a hierarchical formulation of (8) obtained in conjunction with (4) is

$$\begin{aligned}
\boldsymbol{y}_j \mid (\gamma_j, \tau_j, Z_{ij} = 1) &\sim N_{n_j}\Big(\boldsymbol{\mu}_{ij} + \frac{\lambda_i \gamma_j \mathbf{1}_{n_j}}{(1 + n_j\lambda_i^2)}, \\
&\qquad \frac{\sigma_i^2}{\tau_j}\,(\boldsymbol{I}_{n_j} + \lambda_i^2 \mathbf{1}_{n_j}\mathbf{1}_{n_j}^{\top})^{-1}\Big), \\
\gamma_j \mid (\tau_j, Z_{ij} = 1) &\sim TN\Big(0, \frac{\sigma_i^2}{\tau_j}\,(1 + n_j\lambda_i^2); (0, \infty)\Big), \\
\tau_j \mid (Z_{ij} = 1) &\sim \Gamma\Big(\frac{\nu_i}{2}, \frac{\nu_i}{2}\Big), \\
\boldsymbol{Z}_j &\sim M(1; w_1, \ldots, w_g).
\end{aligned} \tag{10}$$

Let $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m), \boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_m)$, $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_m)$ and $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_m)$. It follows from (10) that the complete data log-likelihood function of $\boldsymbol{\Theta}$ given $(\boldsymbol{\gamma}, \boldsymbol{\tau}, \boldsymbol{Z}, \boldsymbol{y})$ is

$$\begin{aligned}
&\ell_c(\boldsymbol{\Theta} \mid \boldsymbol{Z}, \boldsymbol{\gamma}, \boldsymbol{\tau}, \boldsymbol{y}) \\
&= \sum_{j=1}^{m}\sum_{i=1}^{g} Z_{ij}\Big\{ \log w_i - \frac{n_j + 1}{2}\log\sigma_i^2 - \frac{1}{2\sigma_i^2}\Big[\Upsilon_{1ij} + \Upsilon_{2ij}\Big] \\
&\quad + \Big(\frac{\nu_i}{2}\Big)\log\Big(\frac{\nu_i}{2}\Big) - \log\Gamma\Big(\frac{\nu_i}{2}\Big) + \Big(\frac{\nu_i}{2}\Big)(\log\tau_j - \tau_j)\Big\},
\end{aligned} \tag{11}$$

where $\Upsilon_{1ij} = \tau_j(\boldsymbol{y}_j - \boldsymbol{\mu}_{ij})^{\top}(\boldsymbol{y}_j - \boldsymbol{\mu}_{ij})$ and $\Upsilon_{2ij} = \tau_j\big[\gamma_j - \lambda_i \mathbf{1}_{n_j}^{\top}(\boldsymbol{y}_j - \boldsymbol{\mu}_{ij})\big]^2$.

The EM algorithm proceeds by alternately repeating the E- and M- steps where, at the k-th iteration, the E-step involves the calculation of the Q-function, which is the expected value of the complete data log-likelihood (11) conditional on $\boldsymbol{y}$ and the current estimate $\hat{\boldsymbol{\Theta}}^{(k)}$ for $\boldsymbol{\Theta}$, is given by

$$Q(\boldsymbol{\Theta} \mid \hat{\boldsymbol{\Theta}}^{(k)}) = E\big(\ell_c(\boldsymbol{\Theta} \mid \boldsymbol{y}, \boldsymbol{\gamma}, \boldsymbol{\tau}, \boldsymbol{Z}) \mid \boldsymbol{y}, \hat{\boldsymbol{\Theta}}^{(k)}\big). \tag{12}$$

To evaluate (12), the necessary conditional expectations include

$$\begin{aligned}
\hat{\tau}_{ij}^{(k)} &= E(\tau_j|\cdots), \qquad \hat{\kappa}_{ij}^{(k)} = E(\log\tau_j|\cdots), \\
\hat{\gamma}_{1ij}^{(k)} &= E(\tau_j\gamma_j|\cdots), \quad \hat{\gamma}_{2ij}^{(k)} = E(\tau_j\gamma_j^2|\cdots),
\end{aligned} \tag{13}$$

where the symbol "$|\cdots$" stands for conditioning on $Z_{ij} = 1, \boldsymbol{Y}_j = \boldsymbol{y}_j$ and $\boldsymbol{\Theta} = \hat{\boldsymbol{\Theta}}^{(k)}$ and they are directly obtainable through using identities (7) and the law of iterative expectations. Moreover, we define

$$\hat{z}_{ij}^{(k)} = \Pr(Z_{ij} = 1 \mid \boldsymbol{y}, \hat{\boldsymbol{\Theta}}^{(k)}) = \frac{\hat{w}_i^{(k)}\,\psi_{n_j}\big(\boldsymbol{y}_j|\boldsymbol{x}_j, \hat{\boldsymbol{\theta}}_i^{(k)}\big)}{f(\boldsymbol{y}_j|\hat{\boldsymbol{\Theta}}^{(k)})}, \tag{14}$$

which is the posterior probability that the j-th curve belongs to the i-th component evaluated at the $(k+1)$-st iteration. Therefore, the Q-function (12) can be written as

$$\begin{aligned}
Q(\boldsymbol{\Theta}|\hat{\boldsymbol{\Theta}}^{(k)}) &= \sum_{j=1}^{m}\sum_{i=1}^{g}\hat{z}_{ij}^{(k)}\Big\{\log w_i - \Big(\frac{n_j+1}{2}\Big)\log\sigma_i^2 \\
&\quad - \frac{1}{2\sigma_i^2}\Big[\Upsilon_{1ij}^{(k)}(\boldsymbol{\beta}_i) + \Upsilon_{2ij}^{(k)}(\boldsymbol{\beta}_i)\Big] + \Big(\frac{\nu_i}{2}\Big)\log\Big(\frac{\nu_i}{2}\Big) \\
&\quad - \log\Gamma\Big(\frac{\nu_i}{2}\Big) + \Big(\frac{\nu_i}{2}\Big)(\hat{\kappa}_{ij}^{(k)} - \hat{\tau}_{ij}^{(k)})\Big\}.
\end{aligned} \tag{15}$$

where $\Upsilon_{1ij}^{(k)}(\boldsymbol{\beta}_i) = \hat{\tau}_{ij}^{(k)}(\boldsymbol{y}_j - \boldsymbol{\mu}_{ij})^{\top}(\boldsymbol{y}_j - \boldsymbol{\mu}_{ij})$ and $\Upsilon_{2ij}^{(k)}(\boldsymbol{\beta}_i) = \hat{\gamma}_{2ij}^{(k)} - 2\lambda_i\hat{\gamma}_{1ij}^{(k)}\mathbf{1}_{n_j}^{\top}(\boldsymbol{y}_j - \boldsymbol{\mu}_{ij}) + \lambda_i^2\hat{\tau}_{ij}^{(k)}[\mathbf{1}_{n_j}^{\top}(\boldsymbol{y}_j - \boldsymbol{\mu}_{ij})]^2$.

In summary, the implementation of the EM algorithm proceeds as follows:

E-step: Given $\boldsymbol{\Theta} = \hat{\boldsymbol{\Theta}}^{(k)}$, compute $\hat{\tau}_{ij}^{(k)}$, $\hat{\kappa}_{ij}^{(k)}$, $\hat{\gamma}_{1ij}^{(k)}$, $\hat{\gamma}_{2ij}^{(k)}$ and $\hat{z}_{ij}^{(k)}$, for $i = 1, \ldots, g$ and $j = 1, \ldots, n$, by using Eqs. (13) and (14), respectively.

M-step: Calculating $\hat{\boldsymbol{\Theta}}^{(k+1)}$ by optimizing (15) over $\boldsymbol{\Theta}$, the updating formulae are given by

$$\begin{aligned}
\hat{w}_i^{(k+1)} &= \frac{1}{m}\sum_{j=1}^{m}\hat{z}_{ij}^{(k)}, \\
\hat{\boldsymbol{\beta}}_i^{(k+1)} &= \arg\min_{\boldsymbol{\beta}_i}\Big\{\sum_{j=1}^{m}\frac{\hat{z}_{ij}^{(k)}}{\sigma_j^{2(k)}}\Big[\Upsilon_{1ij}^{(k)}(\boldsymbol{\beta}_i) + \Upsilon_{2ij}^{(k)}(\boldsymbol{\beta}_i)\Big]\Big\},
\end{aligned}$$

$$\begin{aligned}
\hat{\lambda}_i^{(k+1)} &= \frac{\sum_{j=1}^{m}\hat{z}_{ij}^{(k)}\hat{\gamma}_{1ij}^{(k)}\mathbf{1}_{n_j}^{\top}(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_{ij}^{(k+1)})}{\sum_{j=1}^{m}\hat{z}_{ij}^{(k)}\hat{\tau}_{ij}^{(k)}\big[\mathbf{1}_{n_j}^{\top}(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_{ij}^{(k+1)})\big]^2}, \\
\hat{\sigma}_i^{2(k+1)} &= \frac{\sum_{j=1}^{m}\hat{z}_{ij}^{(k)}\big[\hat{\Upsilon}_{1ij}^{(k)} + \hat{\Upsilon}_{2ij}^{(k)}\big]}{\sum_{j=1}^{m}\hat{z}_{ij}^{(k)}(n_j+1)},
\end{aligned}$$

where $\hat{\boldsymbol{\mu}}_{ij}^{(k+1)} = \boldsymbol{\mu}_{ij}(\hat{\boldsymbol{\beta}}_i^{(k+1)}, \boldsymbol{x}_j)$ and $\hat{\Upsilon}_{1ij}^{(k)}$ and $\hat{\Upsilon}_{2ij}^{(k)}$ are $\Upsilon_{1ij}^{(k)}(\boldsymbol{\beta}_i)$ and $\Upsilon_{2ij}^{(k)}(\boldsymbol{\beta}_i)$ in (15) with $\boldsymbol{\beta}_i$ replaced by $\hat{\boldsymbol{\beta}}_i^{(k)}$. Consequently, we

obtain $\hat{\nu}_i^{(k+1)}$ by solving the root of the following equation:

$$\log\left(\frac{\nu_i}{2}\right)+1-\mathrm{DG}\left(\frac{\nu_i}{2}\right)+\frac{1}{\sum_{j=1}^{m}\hat{z}_{ij}^{(k)}}\sum_{j=1}^{m}\hat{z}_{ij}^{(k)}\left(\hat{\kappa}_{ij}^{(k)}-\hat{\tau}_{ij}^{(k)}\right)=0.$$

This can be easily done with the help of the R routine '`uniroot`'. The E- and M- steps are alternately repeated until a suitable convergence rule is satisfied, e.g., the Aitken acceleration based stopping criterion $|\ell^{(k+1)} - \ell_\infty^{(k+1)}| < \epsilon$, where $\ell^{(k+1)}$ is the observed log-likelihood evaluated at $\hat{\boldsymbol{\theta}}^{(k)}$, $\ell_\infty^{(k+1)}$ is the asymptotic estimate of the log-likelihood at iteration $k+1$ (McLachlan and Krishnan 2008; Chap. 4.9) and ϵ is the desired tolerance.

Model selection for Step 2: Let $\boldsymbol{X}$ be a p-dimensional random sample drawn from a mixture distribution of g components, each with homogeneous covariance matrix $\boldsymbol{\Gamma}$, and let $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_g$ be a set of candidate cluster centers with $\boldsymbol{c}_r$ being the one closet to $\boldsymbol{X}$. Sugar and James (2003) developed an alternative simple approach to identify the optimal number of clusters based on the "distortion function", defined as

$$d_g = \frac{1}{p}\min_{\boldsymbol{c}_1,\ldots,\boldsymbol{c}_g} E(\boldsymbol{X}-\boldsymbol{c}_r)^\top\boldsymbol{\Gamma}^{-1}(\boldsymbol{X}-\boldsymbol{c}_r), \qquad (16)$$

which is a quantity that measures the average Mahalanobis distance between $\boldsymbol{X}$ and its closest cluster center $\boldsymbol{c}_r$. The Jump function due to Sugar and James (2003) is defined as

$$J_g = \hat{d}_g^{-C} - \hat{d}_{g-1}^{-C},$$

where C is an appropriate positive constant that makes a sharp jump at the true number of clusters and $\hat{d}_g$ is the minimum distortion obtained by the clustering algorithms. They have proven that an appropriate number of clusters can be identified at the peak of jump based on information-theoretic ideas. Their simulation studies have also empirically shown that the jump plot has good performance in finding the true number of clusters.

We applied the Jump function approach to the problem of curve clustering analysis. Each curve is assigned to the component with the largest posterior probability obtained by fitting model (8) for $g = 1, \ldots, g_{\max}$, a pre-specified maximum number of components. We chose $g_{\max}$ to be 12 for all tissues, except for two (spleen and bone marrow) where the model did not converge for $g > 10$. Let $\hat{\boldsymbol{y}}_{ij}$ be the fitted vector of $\boldsymbol{y}_j$ if $\boldsymbol{y}_j$ has been assigned outright to ith cluster, say $\boldsymbol{y}_j \in \mathcal{C}_i$. This gives

$$\begin{aligned}\hat{\boldsymbol{y}}_{ij} &= E\left(\boldsymbol{\mu}_j + \frac{\lambda_i\gamma_j\,\mathbf{1}_{n_j}}{(1+n_j\lambda_i^2)}\middle|\boldsymbol{y}_j, Z_{ij}=1, \hat{\boldsymbol{\Theta}}\right)\Big|_{\boldsymbol{\Theta}=\hat{\boldsymbol{\Theta}}}\\ &= \hat{\boldsymbol{\mu}}_{ij} + \frac{\hat{\lambda}_i\hat{\gamma}_{ij}\,\mathbf{1}_{n_j}}{(1+n_j\hat{\lambda}_i^2)},\end{aligned}$$

where $\hat{\gamma}_{ij} = E(\gamma_j|\boldsymbol{y}_j, Z_{ij}=1, \hat{\boldsymbol{\Theta}})$. Then, the mean squared error for $\boldsymbol{y}_j \in \mathcal{C}_i$ is given by

$$\hat{\Delta}_{ij} = \frac{1}{n_j}(\boldsymbol{y}_j - \hat{\boldsymbol{y}}_{ij})^\top(\boldsymbol{y}_j - \hat{\boldsymbol{y}}_{ij}),$$

which is the scaling squared distance from $\boldsymbol{y}_j$ to $\hat{\boldsymbol{y}}_{ij}$. It follows from (16) that the associated distortion function is empirically defined as

$$\hat{d}_g = \frac{1}{m-g}\sum_{j=1}^{m}\left\{\min\left[\hat{\Delta}_{1j}, \hat{\Delta}_{2j}, \ldots, \hat{\Delta}_{gj}\right]\right\}.$$

Theoretically, the distortion curve, $\hat{d}_g$ versus g, is always monotone decreasing. A simple way of choosing the optimal g is to look for the point at which the magnitude of change in $\hat{d}_g$'s becomes negligible, especially when the subclasses are well-separated. However, using the raw distortion curve could fail in certain cases. As suggested by Sugar and James (2003), the Jump plot method performs extremely well, provided that some suitable values for C are chosen. The optimal number of clusters in data can be visually determined from the peak patterns on the jump plot. Empirical studies show that the point with largest or secondary largest jump is often the best choice.

Quality of curve clustering: To determine the quality of our clustering of mAb profiles, modeled as probability density functions, we measured mean intra- and inter-cluster distances using a symmetric form of Kullback-Leibler distance, denoted by $sKL(p,q)$, between a pair of profiles (p,q), defined as follows:

$$sKL(p,q) = (KL(p,q) + KL(q,p))/2,$$

where $KL(p,q) = \sum_t p_t \log_2(p_t/q_t)$ at each observation point t.

To determine the quality of hierarchical clustering of mAb data, we used the R functions `hclust` (with Euclidean distance metric) and `asw` (average silhouette width). The R package `ks` is used for SiZer plot.

3 RESULTS

Following data preparation and preprocessing, in step 1, mAbprofiler modeled cytometric histograms for 561 mAb samples from 6 tissues using skew t mixture models. Figure 1 illustrates how a profile (shown as grey curve) constructed by mAbprofiler offers a smooth and precise representation of mAb density histograms. This can be contrasted with the original cytometric input in the form of highly non-smooth patterns as shown in Supplementary Fig. 1. To rigorously assess the precision of modeling with our skew t mixture models (STMIX), we computed log-likelihood maxima, Bayesian Information Criterion (BIC) values, the distances D_n between the data and the fitted model (based on Kolmogorov-Smirnov test), and CPU times for STMIX as well as for two competing models of more commonly used mixtures of Gaussian (NMIX) and t distributions (TMIX), and compared them in Supplementary Table 1. Clearly, as shown by BIC, mAbprofiler gives the best fit.

In step 2, for each tissue, mAbprofiler clustered the mAb profiles, specified as density curves, with our new algorithm for fitting skew t mixture of nonlinear regression models. It selected the model that corresponded to the optimal tissue-specific group structure using the maximal value of the Jump statistic over a range of clusters ($g = 1, 2 \ldots, g_{\max}$). Table 1 summarizes the results of this clustering step. The optimal choice of g over the values for which the EM converged is marked in Fig. 3 and Supplementary Fig. 1 (right panel). In Figure 2 (a)-(f), we show each of the six clusters of the spleen profiles. Notably, the profiles were grouped by their significant features overcoming inter-sample variation. Thus the clustering was both accurate and robust. Further, the bottom plot (g) shows the mean profile of every cluster in its own color, thereby contrasting the signature templates for every class while summarizing the characteristic features within each of them. Similar joint plots for every tissue are shown in Supplementary Fig. 2, which also includes the Jump statistics that help in the determination of the optimal group structures.

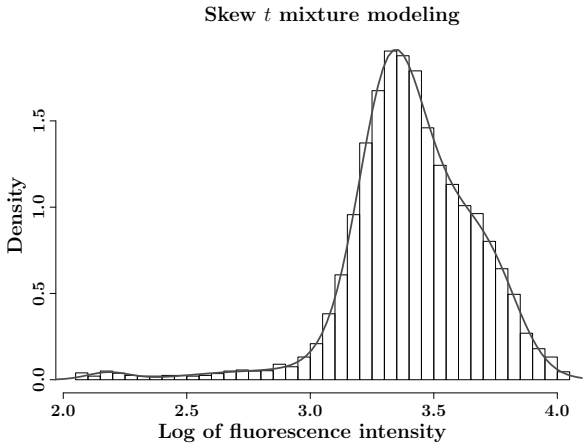

Fig. 1. *Profiling of a mAb reactivity against the surface of sheep cells*: A cytometric histogram measuring the reactivity of a particular mAb tested against the surface of sheep efferent lymphocytes is shown as log of fluorescence intensity of surface expression. In Step 1 of mAbprofiler, the surface expression pattern is profiled with a skew t mixture model as depicted by the smooth and precisely fit grey curve. It captures non-Gaussian features such as skewness and outliers common in cytometric distributions. For the original non-smooth pattern, see Supplementary Fig. 1.

Table 1. *Clustering statistics for Step 2*: For each tissue-type, its count of mAb profiles, number of profile clusters (i.e. mAb classes), IIR (average Intra-cluster distance to average Inter-cluster distance Ratio) and computing time (CT, in minutes) of the EM algorithm for $g = 1, 2 \ldots, g_{\max}$.

Tissue-type	No. of samples	No. of classes	IIR	CT	$g_{\max}$
Spleen	59	6	0.078	27.09	10
Thymocyte	111	12	0.068	118.04	12
Lung lavage	123	11	0.017	95.06	12
Bone marrow	48	8	0.014	23.94	10
Fetal thymus	89	12	0.038	81.24	12
Lymph node	131	8	0.072	111.21	12

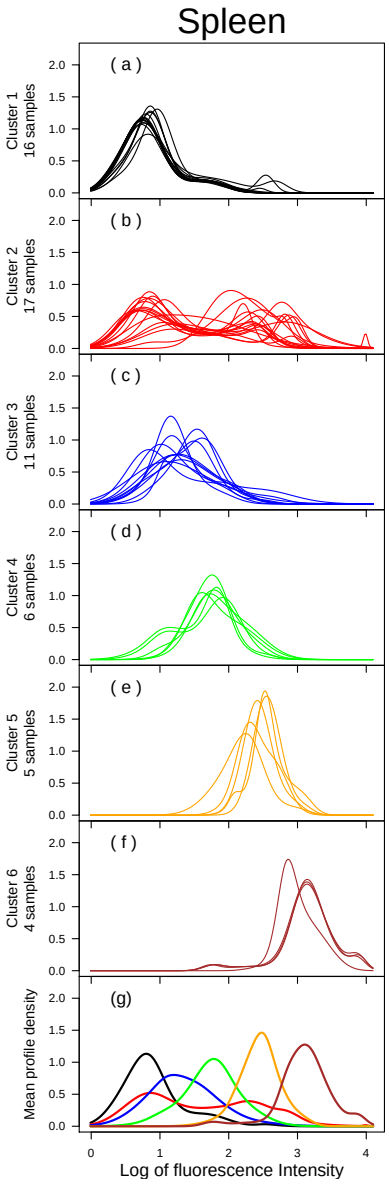

Fig. 2. *Classification of mAb profiles for Spleen*: In Step 2 of mAbprofiler, the profiles of all 59 mAb samples in Spleen were clustered with skew t mixture of nonlinear regression models. The profiles belonging to each of the 6 clusters (a)-(f) are shown in specific colors. The joint plot (g) of all 6 mean profiles in cluster-specific colors allows visual comparison of the cluster templates.

We validated our mAb classification both computationally and empirically. Since every fitted profile is defined by a probability distribution, we computed a symmetric form of Kullback-Leibler distance (sKL) between all pairs of profiles, and observed that the average intra-cluster distances between profiles are considerably lower than the average inter-cluster distances. The ratio (IIR) in every tissue is shown in Table 1. For illustration, the distance matrix for the 6 clusters for spleen is shown in Supplementary Fig. 3.

For empirical validation, for 4 tissue-types (spleen, thymocytes, bone marrow and lymph node) and 2 classes of antibodies (class I mAb T2/39 and anti-LFA-1 mAb F10-150 as described in Pratt *et al.*, 2009), we generated data for 2 pairs of mAb such that the mAb within each pair were known to target molecules of the same class, but across pairs, they targeted molecules from distinct classes. As shown in Supplementary Fig. 4, indeed all profile pairs (black thin curves) cluster together within each class (class I in left panel or anti-LFA-1 in right panel), but separately across 2 distinct classes, providing experimental evidence for precise and objective classification by our framework. As in Fig. 2, the mean profiles are shown in cluster-specific colors for distinguishing the 2 classes in each tissue.

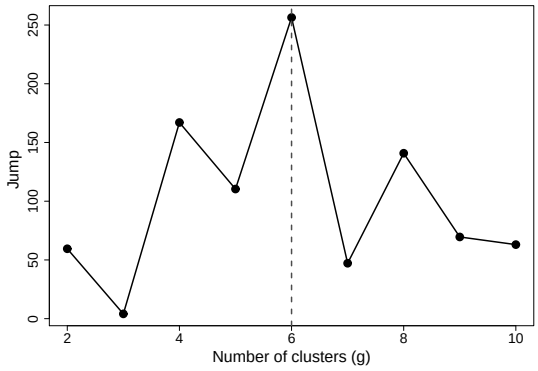

Fig. 3. *Model selection for curve clustering*: Maximum value of Jump for Spleen indicates that 6 is the optimal number of mAb classes in that tissue.

3.1 Comparative analysis with other methods

Besides internal validation, we also compared the performance of mAbprofiler with other established methods. We began with hierarchical clustering, which is the most commonly used approach for mAb classification (Bernard and Boumsell, 1984). When we used hierarchical clustering on our mAb profiles, then the method clearly failed to capture the complex class structure and detected few clusters. Based on Average Silhouette Width (ASW), a common measure for determining the quality of hierarchical clustering, we noted that the optimal number of mAb classes according to hierarchical clustering of our data was typically restricted to 4 or even fewer for all tissues other than Thymocytes. Moreover, little difference among the ASW scores for different number of clusters indicated that the hierarchical clusters had low separation (Supplementary Fig. 5).

Thereafter, we adopted the established protocol of Pratt *et al*. (2009) in which mAb histograms were first smoothed with SiZer, and then hierarchical clustering was performed with those smoothed profiles. We show the results of that approach on our data using SiZer plots for the different tissues in Supplementary Fig. 6(a)-(f). As depicted with the dendrograms, while the larger clustering-structures were detected with smoothing, the finer structures were often ignored, thus resulting in highly heterogeneous classes. This can be seen clearly in the largest clusters in Spleen, Lung lavage and Bone marrow.

Finally, we also studied a combination our approach with that of Pratt *et al*. (2009) in which we clustered the SiZer-smoothed profiles (i.e. we replaced Step 1 of mAbprofiler) using our NLRST algorithm (i.e. we retained Step 2 of mAbprofiler). We observed that while NLRST could identify more classes in the same SiZer profiles for some tissues, the overall gain was not significant. In other words, the fact that mAbprofiler identified a much richer class structure could be attributed to the dual contributions of both Steps 1 as well as 2 of the new framework. While NLRST tackles the inter-sample variation along x-direction (feature location), the skew t mixture pdf provides a precise and continuous representation of the y-direction (feature significance). The resulting effectiveness of mAbprofiler's 2-step approach is illustrated, for example, in the 8 class-templates for the Lymph node which are distinctive along both x- and y-directions (Supplementary Fig. 2 left panel bottom plot). In contrast, the other methods failed to capture that dual complexity and identified only few dominant clusters. The full comparison of classes detected by all 4 methods is shown in Supplementary Table 2.

4 DISCUSSION

Monoclonal antibodies play an immensely important role in molecular biology, biochemistry and medicine. Their utility for probing, stimulating or inhibiting specific target molecules supports numerous diagnostic and immunotherapeutic applications (Zola, 2006). Further, design and development of new mAb are also of great industrial significance. Therefore, objective mAb classification is essential for categorizing and comparing the known as well as the newly developed mAb panels. Besides biochemical methods like immunoprecipitation, this is achieved by clustering flow cytometric reactivity patterns of mAb in different cell types. Unlike traditional HLDA workshops which classified leukocyte surface (CD) molecules (Zola and Swart, 2005) using cell lines, Pratt *et al*. (2009) recently described a system to facilitate practical mAb characterization in animal tissues. This approach is consistent with the new HCDM (Human Cell Differentiation Molecules) focus on various non-hematopoeitic cell types (Zola, 2006). In the present study, we enhanced that approach further by (a) generating a new, larger and more varied collection of mAb patterns in 6 different tissues, and importantly, by (b) constructing a new analytical framework, mAbprofiler, to formally address the technical challenges of mAb characterization. Our 2-step framework provides precise profiling of cytometric histograms (step 1) followed by novel clustering of these curve profiles (step 2). In addition to characterizing mAb for the present study, mAbprofiler can also provide a general framework to allow users to search for archived class signatures or to construct and classify new mAb profiles in a systematic way.

Previous mAb classification studies (e.g. Salganik *et al*., 2005; Pratt *et al*., 2009) have used non-parametric kernel density estimation techniques for detection of significant features in cytometric histograms, typically followed by hierarchical clustering based on Euclidean distances between the features. While being practical, such approaches may not always be precise or robust. For instance, the accuracy of density estimates by kernel-based methods are known to be strongly influenced by bandwidth selection (Jones *et al*., 1996). As observed in Supplementary Fig. 1, significance of the peak features in a given sample, as detected by the program SiZer, is clearly dependent on the choice of bandwidth. This poses a key practical problem, especially since we seek to do unsupervised classification of new mAb profiles. While recent advances in kernel-based techniques have addressed different aspects of cytometric analysis (e.g. Duong *et al*., 2009; Naumann *et al*., 2010), we followed the parametric approach developed by Pyne *et al*. (2009) and Frühwirth-Schnatter and Pyne (2010), which use finite mixtures of skew t distributions, for our purposes. Observations of non-Gaussian features in cytometric data made by these and other recent studies (Lo *et al.*, 2008; Ho *et al*., 2011; Pyne *et al*., 2011) led us to use this more general parametric family of distributions, which also the includes Gaussian distribution as a special case.

Finite mixture models have been extensively used in biology and medicine (McLachlan and Peel, 2000; Frühwirth-Schnatter, 2006). In step 1 of mAbprofiler, we presented a univariate version of the

Pyne *et al*.'s (2009) approach for profiling asymmetric and noisy mAb patterns with finite mixture model of skew t distributions fit via our own EM algorithm (see Supplementary Information). The EM algorithm converges fast in practice, and supports multiple well-known model selection criteria such as AIC, BIC and ICL. In the resulting smooth and precise profiles (see illustrative sample in Fig. 1), every component is specified by rigorous model parameters such as location, size, shape, variance and degrees of freedom. Further, the parametric design enables mAbprofiler to specify the significance of every mAb feature with a smooth and continuous probability density function, which can be represented as a curve that is well-defined at any resolution. Importantly, in step 2, mAbprofiler's skew t mixture of nonlinear regression models can cluster these curve profiles accurately for every cell type. While step 1 follows the approach of Pyne *et al*. (2009), step 2 introduces novel methodology and the EM algorithm implementing it.

A key challenge for cytometric data analysis is inter-sample variation. Similar mAb profiles can vary considerably in both their significance and location, which must be addressed by any algorithm designed for classifying cytometric data. While it is possible to transform or shift and align the data (e.g. Lo *et al.*, 2008; Hahne *et al*., 2010), we want to cluster the mAb profiles precisely in terms of the distinctive features that they present as curves with a robust approach. To systematically model that inter-sample variation, in Step 2, mAbprofiler presented a new nonlinear regression algorithm. It is also a solution for the more generic problem of curve clustering, an important topic in the field of pattern recognition which has not received much attention in the past (e.g. Gaffney, 2004; Gaffney *et al*., 2007; Liu and Yang, 2009). Here we extended the work of Gaffney (2004) to introduce skew t mixture of nonlinear regression models for robust curve clustering with asymmetric variation among the curve features. In our comparative analysis with other methods, we observed that hierarchical clustering is not as well suited for such clustering probably because it critically relies on precise pairwise distances between points. Trying to reduce a curve profile to a point – albeit a multi-dimensional point – can lead to loss of information about features due to binning of the data as specified by a cytometric histogram. That leads to fewer and less well-separated hierarchical clusters, as illustrated in Supplementary Fig. 5.

The problem of using hierarchical clustering for mAb classification gets further compounded with the issue of bandwidth selecton in smoothing of cytometric histograms such as in the protocol of Pratt *et al*. (2009). For our data, the SiZer-smoothed features for a pre-determined bandwidth led to mAb classes with high heterogeneity. While our NLRST (Step 2) clustering could improve detection of the classes with the same SiZer profiles, the net gain was not significant. Therefore the identification of a much richer group structure by mAbprofiler, as shown in Supplementary Table 2 (and the class templates in Fig. 2(g) and Supplementary Fig. 2 left panel) may be attributed to the dual advantage of both Steps 1 and 2 of the new framework. Hierarchical clustering fails to capture the complexity of data when presented in the form of noisy curve profiles in which the true significance of features is not apparent. This is even more difficult if there are few significant features, which, in turn, might suffer from inter-sample variation. By addressing these issues, mAbprofiler produced robust mAb classes – specified as curves of probability density functions – even in the presence of non-Gaussian variation. It achieves this without any need for transforming the profiles or reducing them to points as required for hierarchical clustering.

The new framework has several additional advantages. Its use of Jump statistic provides a suitable criterion for optimal model selection in profile clustering. Each step of mAbprofiler can be performed independently with its own EM algorithm, which offers the flexibility of pipelining the framework with external algorithms. As output, not only does mAbprofiler produce a smooth profile for a mAb histogram, it also generates a mean template for the "signature" pattern of every mAb class, along with parametric description of significant features therein. As a result, class-templates can be archived, and later searched for information on overall or specific characteristics of such known mAb classes. Thus it facilitates pattern matching with newly constructed mAb profiles, which can be grouped with classes having the most similar templates. Our computational and empirical validation of mAbprofiler classification shows how this is achieved. Another feature of our approach is that it does not require a clonal population. Expression can be analyzed both on individual cells and within a complex cell population. Moreover, our non-Gaussian model can be easily extended to temporal mAb profiling (Pyne *et al*., 2011), e.g., for measurements over the course of dampening of an inflammation in a certain tissue. The strength of mAbprofiler lies in providing a much-needed robust and objective framework for mAb characterization in different cell types and tissues.

ACKNOWLEDGEMENT

SP thanks Massimo Loda for financial support.

REFERENCES

Arellano-Valle, R.B., Bolfarine, H., Lachos, V.H. (2005) Skew-normal linear mixed models. *Journal of Data Science*, **3**, 415–438.

Azzalini, A., and Capitaino, A. (2003) Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution. *Journal of the Royal Statistical Society, Series B*, **65**, 367–389.

Azzalini, A., and Dalla-Valle, A. (1996) The multivariate skew-normal distribution. *Biometrika*, **83**, 715–726.

Banfield, J.D., and Raftery, A.E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, **49**, 803–821.

Bernard, A. and Boumsell, L. (1984) The clusters of differentiation (CD) defined by the First International Workshop on Human Leucocyte Differentiation Antigens. *Human Immunology*, **11**(1), 1–10.

De la Cruz, R. (2008) Bayesian non-linear regression models with skew-elliptical errors: Applications to the classification of longitudinal profiles. *Computational Statistics & Data Analysis*, **53**, 436–449.

Dempster, A.P., Laird, N.M., Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Duong, T., Koch, I., and Wand, M.P. (2009) Highest density difference region estimation with application to flow cytometric data. *Biom. Journal*, **51**(3), 504–21.

Frühwirth-Schnatter, S. (2006) *Finite Mixture and Markov Switching Models.* Springer, New York.

Frühwirth-Schnatter, S. and Pyne, S. (2010) Bayesian inference for finite mixtures of univariate and multivariate skew normal and Skew-t Distributions. *Biostatistics*, **11**, 317–336.

Gaffney, S. (2004). Probabilistic curve-aligned clustering and prediction with mixture models. Ph.D. Dissertation. Department of Computer Science, University of California, Irvine.

Gaffney, S.J., Robertson, A.W., Smyth, P., Camargo, S.J. and Ghil, M. (2007) Probabilistic clustering of extratropical cyclones using regression mixture models. *Climate Dynamics*, **29**, 423–440.

Gilks, W. R. and Shaw, S. (1995) Statistical analysis. *Leucocyte Typing V* (eds. Schlossman S. *et al.*). Oxford University Press, Oxford, 8–13.

Hahne, F., Khodabakhshi, A.H., Bashashati, A., Wong, C.J., Gascoyne, R.D., Weng, A.P., Seyfert-Margolis, V., Bourcier, K., Asare, A., Lumley, T., Gentleman, R. and Brinkman, R.R. (2010) Per-channel basis normalization methods for flow cytometry data. *Cytometry Part A*, **77**, 121–131.

Herzenberg, L.A., De Rosa, S.C. and Herzenberg, L.A. (2000) Monoclonal antibodies and the FACS: complementary tools for immunobiology and medicine. *Immunology Today*, **21**, 383–390.

Ho, H.J., Pyne, S. and Lin, T.I. (2011) Maximum likelihood inference for mixtures of skew Student-t-normal distributions through practical EM-type algorithms. *Statistics and Computing*, DOI: 10.1007/s11222-010-9225-9.

Jones, P. N. and McLachlan, G.J. (1992) Fitting finite mixture models in a regression context. *Australian Journal of Statistics*, **34**, 233–240.

Jones, M.C., Marron, J.S. and Sheather, S.J. (1996) A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, **91**, 401–407.

Kim, E.Y., Zeng, Q., Rawn,, J., Wand, M., Young, A.J., Milford, E., Mentzer, S.J., and Greenes, R.A. (2002) Using a neural network with flow cytometry histograms to recognize cell surface protein binding patterns. *Proc AMIA Symp*, 380–384.

Li, X., Abdi, K., and Mentzer, S.J. (1995) Hybridoma screening using an amplified fluorescence microassay to quantify immunoglobulin concentration. *Hybridoma*, **14**, 75–78.

Lin, T.I., Lee, J.C., Hsieh, W.J. (2007) Robust mixture modeling using the skew t distribution. *Statistics and Computing*, **17**, 81–92.

Liu, C.H., and Rubin, D.B. (1994) The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. *Biometrika*, **81**, 633–648.

Liu, X. and Yang, M.C.K. (2009) Simultaneous curve registration and clustering for functional data *Computational Statistics & Data Analysis*, **53**, 1361–1376.

Lo, K., Brinkman, R.R., Gottardo, R. (2008) Automated gating of flow cytometry data via robust model-based clustering. *Cytometry Part A*, **73**, 321–332.

McLachlan, G.J. and Basford, K.E. (1988) *Mixture Models: Inference and Application to Clustering.* Marcel Dekker, New York.

McLachlan, G.J. and Peel, D. (2000) *Finite Mixture Models.* Wiley, New York.

McLachlan, G.J. and Krishnan T. (2008) *The EM algorithm and extensions.* 2nd edn. John Wiley & Sons, New York.

Naumann, U., Luta, G., and Wand, M.P. (2010) The curvHDR method for gating flow cytometry samples. *BMC Bioinformatics*, **11**, 44.

Pratt, J.P., Zeng, Q., Ravnic, D., Huss, H., Rawn, J. and Mentzer S.J. (2009) Hierarchical clustering of monoclonal antibody reactivity patterns in nonhuman species. *Cytometry. Part A: The Journal of the International Society for Analytical Cytology*, **75**(9), 734–742.

Pyne, S., Hu, X., Wang, K., Rossin, E., Lin, T.I., Maier, L., Baecher-Allan, C., McLachlan, G.J., Tamayo, P., Hafler, D.A., De Jager, P.L. and Mesirov, J.P. (2009) Automated high-dimensional flow cytometric data analysis. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(21), 8519–8524.

Pyne, S., Ho, H.J., Haase, S.B. and Lin, T.I. (2011) Parametric modeling of cellular state transitions as measured with flow cytometry. *Proc. IEEE 1st International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)*, 147–152.

Salganik M.P., Milford, E.L., Hardie, D.L., Shaw, S. and Wand M.P. (2005) Classifying antibodies using flow cytometry data: class prediction and class discovery. *Biometrical Journal*, **91**, 785–800.

Spiegelhalter, D. J. and Gilks, W. R. (1987) Statistical analysis. *Leucocyte Typing I* (eds. McMichael, A. J. *et al.*). Oxford University Press., Oxford, 14–24.

Sugar, C.A. and James, G.M. (2003) Finding the number of clusters in a dataset: an information-theoretic approach. *Journal of the American Statistical Association*, **98**, 750–763.

Zeng, Q.T., Pratt, J.P., Pak, J., Ravnic, D., Huss, H. and Mentzer, S.J. (2007) Feature-guided clustering of multi-dimensional flow cytometry datasets. *Journal of Biomedical Informatics*, **40**, 325–331.

Zeng, Q., Wand, M., Young, A.J., Rawn, J., Milford, E.L., Mentzer, S.J., and Greenes, R.A. (2002) Matching of flow-cytometry histograms using information theory in feature space. *Proc AMIA Symp*, 929–933.

Zola, H. and Swart, B. (2005) The human leucocyte differentiation antigens (HLDA) workshops: the evolving role of antibodies in research, diagnosis and therapy. *Cell Research*, **15**, 691–694.

Zola, H. (2006) Medical applications of leukocyte surface molecules–the CD molecules. *Molecular Medicine*, **12**, 312–316.