Abe *et al. BMC Bioinformatics* 2020, **21**(Suppl 13):393
https://doi.org/10.1186/s12859-020-03671-7

BMC Bioinformatics

RESEARCH Open Access

Model-based clustering for flow and mass cytometry data with clinical information

Ko Abe[1], Kodai Minoura[1,2], Yuka Maeda[3], Hiroyoshi Nishikawa[2,3] and Teppei Shimamura[1*]

From The 18th Asia Pacific Bioinformatics Conference
Seoul, Korea. 18–20 August 2020

*Correspondence:
shimamura@med.nagoya-u.ac.jp
[1]Division of Systems Biology, Nagoya University Graduate School of Medicine, 65 Tsurumai-cho, Showa-ku, 4668550 Nagoya, Japan
Full list of author information is available at the end of the article

Abstract

Background: High-dimensional flow cytometry and mass cytometry allow systemic-level characterization of more than 10 protein profiles at single-cell resolution and provide a much broader landscape in many biological applications, such as disease diagnosis and prediction of clinical outcome. When associating clinical information with cytometry data, traditional approaches require two distinct steps for identification of cell populations and statistical test to determine whether the difference between two population proportions is significant. These two-step approaches can lead to information loss and analysis bias.

Results: We propose a novel statistical framework, called LAMBDA (Latent Allocation Model with Bayesian Data Analysis), for simultaneous identification of unknown cell populations and discovery of associations between these populations and clinical information. LAMBDA uses specified probabilistic models designed for modeling the different distribution information for flow or mass cytometry data, respectively. We use a zero-inflated distribution for the mass cytometry data based the characteristics of the data. A simulation study confirms the usefulness of this model by evaluating the accuracy of the estimated parameters. We also demonstrate that LAMBDA can identify associations between cell populations and their clinical outcomes by analyzing real data. LAMBDA is implemented in R and is available from GitHub (https://github.com/abikoushi/lambda).

Keywords: Flow cytomety, Mass cytometory, Bayesian mixture model, Stochastic EM algorithm

Background

The recent development of high-dimensional flow cytometry and mass cytometry (CyTOF) allows for characterizing cell types and states by detecting the expression levels of pre-defined sets of surface and intracellular proteins at single cell resolution [1]. For an individual subject, the modern flow cytometry data consist of 20 or more protein

measurements from millions of cells from the subject. The recent mass cytometry systems use antibodies tagged with heavy metal isotopes which reduce signal interference due to spectral overlap and autofluorescence and enable the detection of more than 40 proteins per cell [2]. This high-dimensional cytometry data contains useful information to diagnose diseases such as leukemia [3] and HIV [4], as well as to predict clinical outcomes such as the response to cancer immune-therapies [5].

A key challenge in the analysis of high-dimensional cytometry data is to identify unknown cell populations that relate as prognostic factors to clinical outcomes of interest. Traditional analysis is done by manual gating which suffers not only from the need to detect unknown cell populations, but also from the need to ensure reproducibility [6]. This lack of reproducibility has two subjective causes. The first concerns the order in which pairs of markers are explored. This order often allows some degree of freedom in the gating process, so it might lead to the selection of different cells by alternative gating strategies. The second concerns the boundaries of the gates being used. There is considerable diversity between operators in terms of gating strategy, in which some experts gate strictly and others gate generously. The subjective nature of manual gating allows for the production of too wide a variety of results to be accurately reproducible.

As an alternative to manual gating, researchers have developed several computational methods, including Citrus [7], cydar [8], and diffcyt [9], to infer cell populations or states associated with an outcome variable in high-dimensional cytometry data. However, these methods require two steps: a first step in which cell populations are identified using a clustering algorithm, and a second step in which the summary statistics of the identified cell populations are concatenated into a clinical outcome of interest which can lead to information loss and analysis bias. Furthermore, these methods do not consider the distinctive features of the expression values of mass cytometry data as opposed to flow cytometry data. Mass cytometry data is marked by a zero-inflated distribution (Fig. 1). That is, proteins can be either 'on' or 'off,' in which either a positive expression measure is recorded or the recorded expression is zero or negligible, and where a very high proportion of the data entries are zero. On the other hand, in flow cytometry, the boundary between 'on' and 'off' is more ambiguous, which leads to a bimodal Gaussian distribution (Fig. 1). Therefore, lack of consideration for this distribution difference in the existing methods masks the underlying difference in cell populations and gives rise to a misleading conclusion in both basic and clinical research.

To address the aforementioned problems, we propose a new probabilistic approach for identifying unknown cell populations associated with clinical outcomes of interest which we have named LAMBDA (Latent Allocation Model with Bayesian Data Analysis). The contributions of our proposed method are summarized as follows:

- Our method is a one-step procedure that directly uses cytometry data at the single cell level to simultaneously discover cell populations and to identify the associations of these populations with clinical outcomes of interest. Our model can also be used to find relationships between cell populations and a single clinical outcome as well as relationships between cell populations and multiple clinical outcomes.
- Our method is based on correctly specified probabilistic models that are designed for modeling the different distribution information of flow and mass cytometry data respectively. In the case of flow cytometry, LAMBDA assumes that the data is

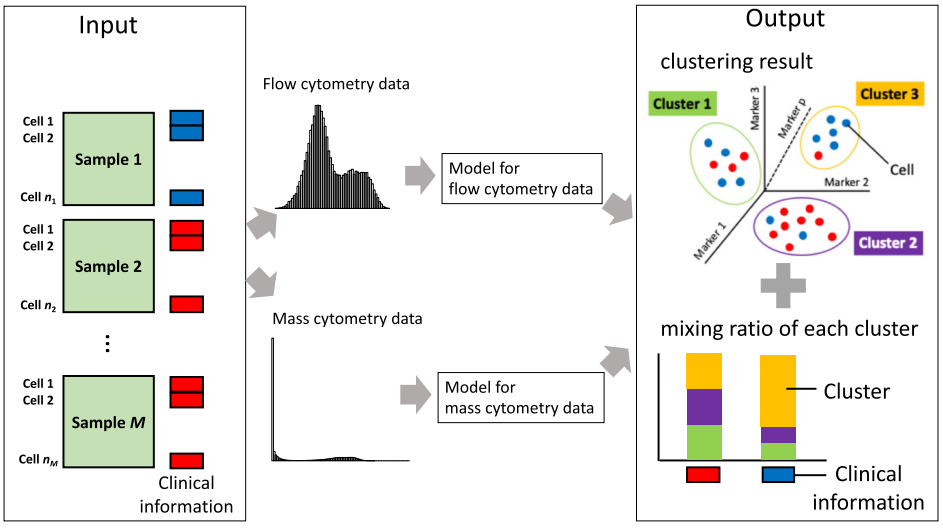

Fig. 1 Conceptual diagram. LAMBDA estimates composition ratio of clusters depends on the clinical information

generated from a mixture of multivariate normal distributions, each of which represents an unknown cell population. On the other hand, in the case of mass cytometry, LAMBDA assumes that the data is generated from a mixture of zero-inflated distributions that represent censoring of expression below a substantial limit of detection. In both models, the compositions of cell populations are assumed to vary with clinical outcomes.

- We provide a simple and efficient learning procedure for the proposed model using a stochastic EM algorithm that reduces computational cost. LAMBDA is implemented in the R environment, which is available from https://github.com/abikoushi/LAMBDA.

From here we will explain the method and its implications in detail. Figure 1 shows a conceptual view of analysis by LAMBDA. The "Methods" section details the proposed model and algorithm. The "Results" section includes an analysis of the efficiency of LAMBDA using synthetic and real data. The "Conclusion" section summarizes the data presented here and describes the possibility for future expansion of this model.

Methods

Model for flow cytometry data

Suppose that we observe the flow cytometry dataset $\boldsymbol{y}_n \in \mathbb{R}^K$, $(n = 1, \ldots, N)$ and clinical information $\boldsymbol{x}_n \in \mathbb{R}^D$. The dataset includes N cells, K markers and D-dimensional clinical information. Our goal is to identify cells populations from the data. Furthermore we seek to understand how these cell populations change depending on the clinical information. LAMBDA is a model based clustering method. Let L be the number of clusters. The data generative process of LAMBDA for **flow cytometry** data is defined as follows:

$$\boldsymbol{y}_n | w_n \sim \prod_{l=1}^{L} \text{Gaussian}\,(\boldsymbol{\mu_l}, \Sigma_l)^{w_{n,l}}$$

$$\boldsymbol{w}_n | \boldsymbol{x}_n \sim \text{Categorical}(\boldsymbol{\phi}_n)$$

$$\begin{aligned}\boldsymbol{\phi}_n &= \mathrm{softmax}(\boldsymbol{x}_n\boldsymbol{\beta})\\ \boldsymbol{\mu}_l &\sim \mathrm{Gaussian}\left(\mathbf{0}, \frac{1}{\tau}\boldsymbol{\Sigma}_l\right)\\ \boldsymbol{\Sigma}_l^{-1} &\sim \mathrm{Wishart}\,(\nu, \boldsymbol{\Lambda})\end{aligned} \tag{1}$$

where $w_{n,l}$ is l-th element of $\boldsymbol{w}_n$ and $D \times L$ matrix $\boldsymbol{\beta}$ is the effect of clinical information. For identifiability, the first column of $\boldsymbol{\beta}$ is always set as zero. Here, the softmax function is defined by $\mathrm{softmax}(\boldsymbol{x}) = \frac{\exp(\boldsymbol{x})}{\sum_{k=1}^{K}\exp(x_k)}$ for vector $\boldsymbol{x} = (x_1, \ldots, x_K)^\top$ using an element-wise exponential function. Figure 2 shows a plate diagram of this data generating process. This model is a kind of conditional Gaussian mixture model [10]. However, the details of the estimation method are not described in detail in the publication, so we will describe them here.

Parameter estimation for flow cytometry data

We find the maximum a posteriori probability (MAP) estimators, using an EM algorithm.

If the latent variable $w_{n,l}$ is given, the complete likelihood of this model is represented by the following formula:

$$L^{(c)} = \prod_{n=1}^{N}\prod_{l=1}^{L}\phi_{n,l}^{w_{n,l}}\mathcal{N}(\boldsymbol{y}_n|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)^{w_{n,l}}. \tag{2}$$

In the E-step, we calculate

$$w_{n,l}^{(i)} = \frac{\phi_{n,l}\mathcal{N}\left(\boldsymbol{y}_n|\boldsymbol{\mu}_l^{(i-1)}, \boldsymbol{\Sigma}_l^{(i-1)}\right)}{\sum_{l=1}^{L}\phi_{n,l}\mathcal{N}\left(\boldsymbol{y}_n|\boldsymbol{\mu}_l^{(i-1)}, \boldsymbol{\Sigma}_l^{(i-1)}\right)}, \tag{3}$$

where $\mathcal{N}(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the density function of the multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.

In the M-step, we update the parameters using:

$$\mu_{k,l}^{(i)} = \frac{\sum_{n=1}^{N} w_{n,l}^{(i)} y_{n,k}}{\sum_{n=1}^{N} w_{n,l}^{(i)} + \tau} \tag{4}$$

$$\boldsymbol{\Sigma}_l = \frac{\sum_n w_{n,l}^{(i)}\left(\boldsymbol{y}_n - \boldsymbol{\mu}_l^{(i)}\right)\left(\boldsymbol{y}_n - \boldsymbol{\mu}_l^{(i)}\right)^\top + \tau\boldsymbol{\mu}_l^{(i)\top}\boldsymbol{\mu}_l^{(i)} + \boldsymbol{\Lambda}}{\sum_n w_{n,l}^{(i)} + \nu - K}, \tag{5}$$

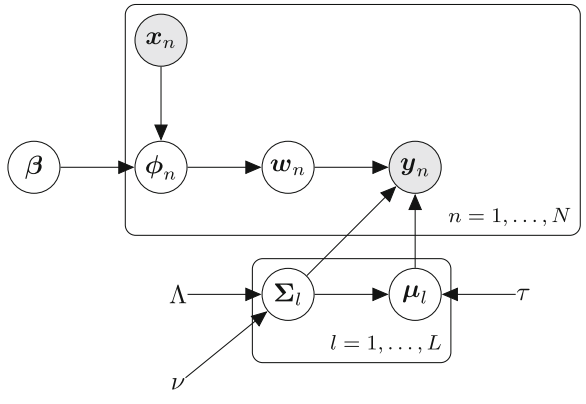

Fig. 2 Plate diagram of the data generating process in LAMBDA for flow cytometry data. The white nodes indicate latent variables and the gray nodes indicate observed variables

Because closed form solutions for $\boldsymbol{\beta}$ are unavailable, we use Newton's method to obtain estimates. We obtain estimates of $\boldsymbol{\beta}$ by maximizing the following equation, with respect to $\boldsymbol{\beta}$:

$$Q(\boldsymbol{\beta}) = \sum_{n=1}^{N}\sum_{l=1}^{L}\left\{ w_{nl}^{(i)}\left(\sum_{d=1}^{D} x_{nd}\beta_{d,l} - \log\sum_{l=1}^{L}\exp\left(\sum_{d=1}^{D} x_{n,d}\beta_{d,l}\right)\right)\right\} \tag{6}$$

First order derivative of the function $Q(\boldsymbol{\beta})$ is represented by:

$$\nabla Q(\boldsymbol{\beta}) = \sum_{n=1}^{N}(\boldsymbol{w}_n - \text{softmax}(X_n\boldsymbol{\beta}))\otimes \boldsymbol{x}_n. \tag{7}$$

Second order derivative of the function $Q(\boldsymbol{\beta})$ is represented by:

$$\nabla^2 Q(\boldsymbol{\beta}) = \sum_{n=1}^{N}(P_n - \text{softmax}(X_n\boldsymbol{\beta})\text{softmax}(X_n\boldsymbol{\beta})^\top)\otimes \boldsymbol{x}_n\boldsymbol{x}_n^\top \tag{8}$$

where $\otimes$ denotes the Kronecker product and P_n is defined as follows:

$$P_n = \begin{pmatrix} \phi_{n,1} & 0 & 0 & \cdots & 0 \\ 0 & \phi_{n,2} & 0 & \cdots & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & 0 & \cdots & \phi_{n,L} \end{pmatrix}. \tag{9}$$

Thus, we update the estimate of $\boldsymbol{\beta}$ using:

$$\text{vec}(\boldsymbol{\beta}^{(i+1)}) = \text{vec}(\boldsymbol{\beta}^{(i)}) - \text{vec}(\nabla Q(\boldsymbol{\beta}))\left(\nabla^2 Q(\boldsymbol{\beta})\right)^{-1}, \tag{10}$$

where vec is the vec operator.

Model for mass cytometry data

In the same manner as the previous subsection, we observe a mass cytometry dataset, in this case $\boldsymbol{y}_n \in \mathbb{R}^K$, $(n = 1, \ldots, N)$ and clinical information $\boldsymbol{x}_n \in \mathbb{R}^D$. An important feature of mass cytometry data, is that a very high proportion of the data entries are zero. The ordinary mixture of the Gaussian model can not explain these zeroes. Here, LAMBDA steps in to properly assess the data. The data generative process of LAMBDA for **mass cytometry** data is defined as follows:

$$\begin{aligned} \boldsymbol{y}_n &= \begin{cases} z_{n,k} & z_{n,k} > 0 \\ 0 & z_{n,k} \leq 0 \end{cases} \\ \boldsymbol{z}_n|w_n &\sim \prod_{l=1}^{L}\text{Gaussian}\,(\boldsymbol{\mu_l}, \Sigma_l)^{w_{n,l}} \\ \boldsymbol{w}_n|\boldsymbol{x}_n &\sim \text{Categorical}(\boldsymbol{\phi}_n) \\ \boldsymbol{\phi}_n &= \text{softmax}(\boldsymbol{x}_n\boldsymbol{\beta}) \\ \boldsymbol{\mu}_l &\sim \text{Gaussian}\left(\mathbf{0}, \frac{1}{\tau}\boldsymbol{\Sigma}_l\right) \\ \boldsymbol{\Sigma}_l^{-1} &\sim \text{Wishart}\,(\nu, \boldsymbol{\Lambda}) \end{aligned} \tag{11}$$

Figure 3 shows a plate diagram of this data generating process.

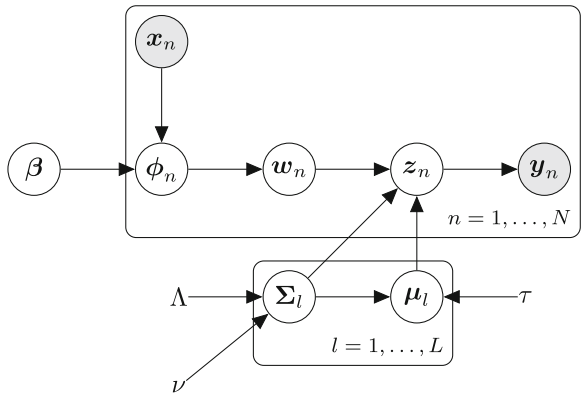

Fig. 3 Plate diagram of the data generating process in LAMBDA for mass cytometry data

Parameter estimation for mass cytometry data

Looking again at Eq. (11), if the latent variables w_n and z_n are given, the complete likelihood of this model is represented by the following formula:

$$L^{(c)} = \prod_{n=1}^{N} \prod_{l=1}^{L} \phi_{n,l}^{w_n,l} \mathcal{N}(\boldsymbol{z}_n | \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)^{w_n,l}. \tag{12}$$

For mass cytometry data, we use a stochastic EM algorithm. In the E step, Monte Carlo samples $\tilde{\boldsymbol{w}}_n$ and $\tilde{\boldsymbol{z}}_n$ replace missing data $\boldsymbol{w}_n$ and $\boldsymbol{z}_n$.

If an arbitrary value for $\boldsymbol{z}_n$ is given, we can sample $\tilde{\boldsymbol{w}}_n$ from following categorical distribution:

$$\tilde{\boldsymbol{w}}_n \sim \text{Categorical}(\boldsymbol{\eta}_n) \tag{13}$$

where the l-th element of $\boldsymbol{\eta}_n$ is represented by the following formula:

$$\eta_{nl} = \frac{\phi_{n,l} \mathcal{N}(\tilde{\boldsymbol{z}}_n | \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}{\sum_{l=1}^{L} \phi_{n,l} \mathcal{N}(\tilde{\boldsymbol{z}}_n | \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}. \tag{14}$$

In contrast, if an arbitrary value for $\boldsymbol{w}_n$ is given, we can sample $\tilde{\boldsymbol{z}}_n$ from truncated Normal distribution. The steps of the Gibbs sampler for generating $\tilde{\boldsymbol{z}}_n$ are:

- If $y_{n,k} > 0$, $\tilde{z}_{n,k} = y_{n,k}$, othewise,
- let $\mu_k^{(n)} = \prod_{l=1}^{L} \mu_{k,l}^{\tilde{w}_{n,l}}$ and $\Sigma_{i,j}^{(n)} = \prod_{l=1}^{L} \Sigma_{i,j,l}^{w_{n,l}}$, where $\Sigma_{i,j,l}$ is (i,j)-element of $\boldsymbol{\Sigma}_l$
- let $\mu_{k,-k}^{(n)} = \mu_k^{(n)} - \Sigma_{k,k}^{(n)} (\boldsymbol{\Sigma}^{(n)})_{k,-k}^{-1} (z_{n,-k} - \boldsymbol{\mu}_{-k}^{(n)})$
- let $b = \Phi(0 | \mu_{k,-k}^{(n)}, \Sigma_{k,k}^{(n)})$
- let $u \sim \text{Uniform}(0,1)$
- $\tilde{z}_{n,k} = \mu_{k,-k}^{(n)} + \sqrt{\Sigma_{k,k}^{(n)}} \Phi^{-1}(ub | 0,1)$

where $\Phi(y | \mu, \Sigma)$ denotes the distribution function of a univariate Gaussian distribution with mean μ and variance Σ and x_{-i} is the set of all variables in x except for the i-th variable.

In the M-step, by replacing $y_{n,k}$ and $w_{n,l}^{(i)}$ with samples $\tilde{z}_{n,k}$ and $\tilde{w}_{n,l}$, Eqs. (4), (5), and (10) can be used.

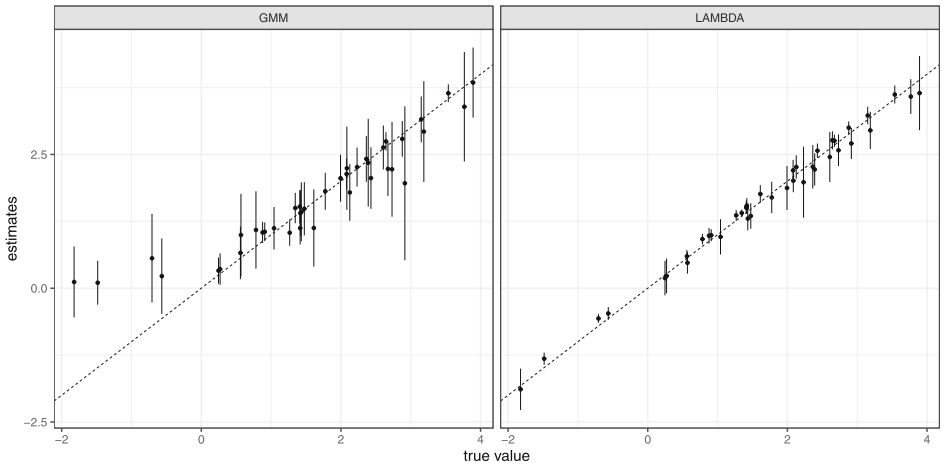

Fig. 4 Simulation result of $\boldsymbol{\mu}$. The comparison true $\boldsymbol{\mu}$ and the mean of estimates $\hat{\boldsymbol{\mu}}$. The error bars indicates SE. Left panel show the result of GMM, right panel show the result of LAMBDA

Model selection

In fitting the model, it is important to choose an appropriate number for L. It is well known that the number of cluster L with the lowest Bayesian information criterion (BIC) is an appropriate number. The BIC is defined as follows:

$$\mathrm{BIC} = -2\log(\mathcal{L}) + f\log(N), \tag{15}$$

where $\mathcal{L}$ is the likelihood and f is the number of estimated parameters. However, for mass cytometry data, it requires a high computational cost to calculate the exact likelihood in the stochastic EM algorithm. Thus, in this article, we use BIC for flow cytometry data, and *elbow* method for mass cytometry data to choose L. Elbow method chooses a number of clusters that adding another cluster doesn't give a better fit to the data. The goodness of fit of the model to data is evaluated by the sum of squared error (SSE). SSE is defined by following:

$$\mathrm{SSE} = \sum_{n=1}^{N}\sum_{k=1}^{K}\left(y_{n,k} - \max\left(0, \mu_k^{(n)}\right)\right)^2. \tag{16}$$

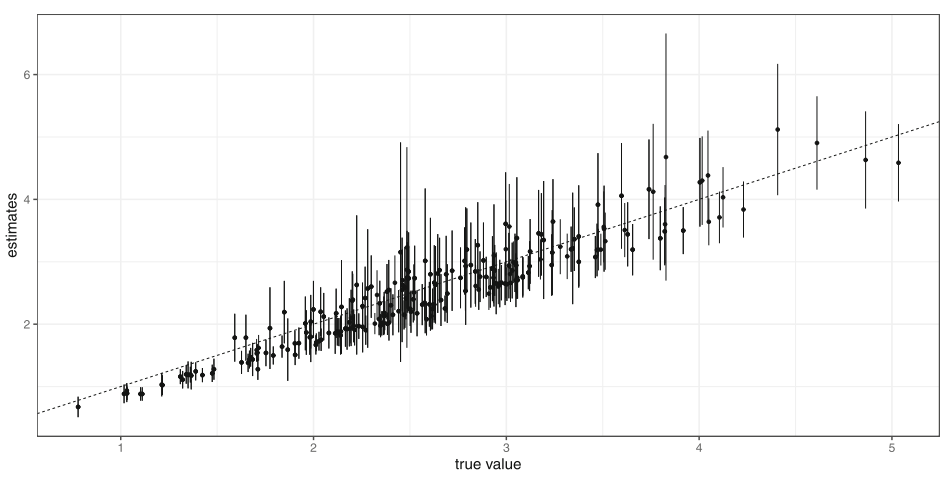

Fig. 5 Simulation result of $\boldsymbol{\Sigma}$. The comparison true $\boldsymbol{\Sigma}$ and the mean of estimates $\hat{\boldsymbol{\Sigma}}$. The error bars indicates SE

Table 1 Simulation result of mixture proportion

cluster	1	2	3	4
	category 1			
true value	0.1	0.2	0.3	0.4
mean (LAMBDA)	0.10	0.20	0.30	0.40
mean (GMM)	0.10	0.20	0.31	0.39
SE (LAMBDA)	0.01	0.02	0.00	0.01
SE (GMM)	0.01	0.04	0.03	0.03
	category 2			
true value	0.25	0.25	0.25	0.25
mean (LAMBDA)	0.24	0.26	0.25	0.25
mean (GMM)	0.25	0.26	0.24	0.24
SE (LAMBDA)	0.04	0.04	0.03	0.01
SE (GMM)	0.03	0.04	0.03	0.06

Results

Simulation study

To evaluate the standard error (SE) and the bias of the estimations, we conducted simulation experiments. The bias of $\hat{\theta}$ is defined by the difference between the true value and the estimated value ($E[\hat{\theta}] - \theta$). The synthetic data was naturally produced via the data generating process given by Eq 11. We set $K = 10$. The μ and Σ were randomly generated.

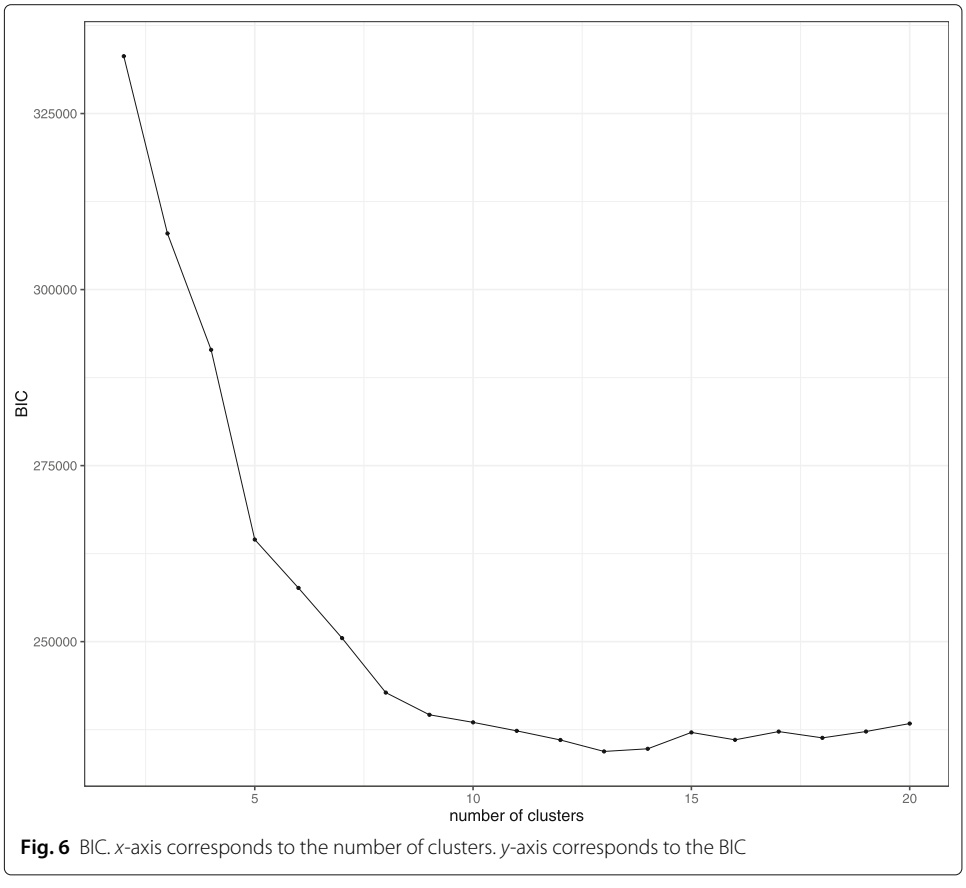

Fig. 6 BIC. *x*-axis corresponds to the number of clusters. *y*-axis corresponds to the BIC

We used a multivariate normal distribution to generate the synthetic data, and values less than 0 were replaced by 0.

We estimated the parameters from 100 replicates of the experiment. We set the sample size $N = 2000$, the number of clusters $L = 4$, and the categories $D = 2$. One category has the mixture proportion $\boldsymbol{\phi}_1 = (0.1, 0.2, 0.3, 0.4)^\top$, the other has the mixture proportion $\boldsymbol{\phi}_2 = (0.25, 0.25, 0.25, 0.25)^\top$. To differentiate these two categories, we set $X = (\mathbf{1}, \boldsymbol{x})$, where $\mathbf{1}$ is a vector of ones. The variable $\boldsymbol{x}$ is a dummy variable to indicate the category. When estimating parameters, we set $\tau = 0.01$, $\nu = K + 2$, and Λ is an identity matrix, which is equivalent to a weakly-informative prior distribution. To avoid the problem of label switching [11], the estimated parameters are rearranged as $\phi_{1,1} \leq \phi_{1,2} \leq \phi_{1,3} \leq \phi_{1,4}$.

Synthetic data was analyzed by LAMBDA along with ordinaly Gaussian mixture model (GMM) which cannot incorporate explanatory variables. using R package "mclust". We estimated the clusters by GMM and calculate the mixture proportion of the estimated clusters by category and the median for each cluster was used as the estimates of μ.

The mean and SE for the estimated $\hat{\mu}$ and $\hat{\Sigma}$ are shown in Figs. 4 and 5, respectively. We observed that the points were arranged diagonally, indicating that the estimator of LAMBDA is unbiased. In contrast, GMM estimates often have a large bias. The mean and SE for the estimated $\hat{\phi}$ is shown in Table 1. In the case of synthetic data, the algorithm of

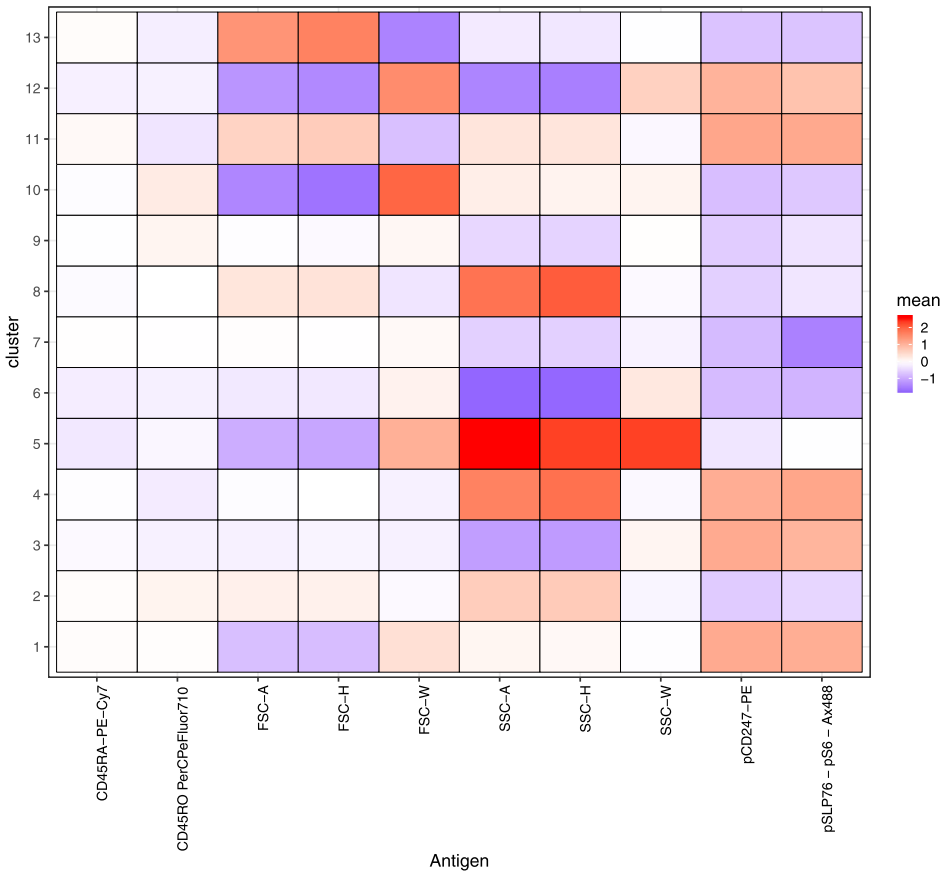

Fig. 7 Estimated $\hat{\mu}$ for Landrigan's study. The *x*-axis corresponds to the markers, and the *y*-axis corresponds to the clusters. Red, white, and blue indicates high, middle, and low marker intensity, respectively

LAMBDA uses parameters estimated with small biases and is able to produce reasonable estimates.

Results on real data

We applied LAMBDA to real world flow and mass cytometry data. When estimating parameters, we set $\tau = 0.01$, $\nu = K + 2$, and Λ is an identity matrix.

For the case of flow cytometry we turn to Landrigan's study (https://community.cytobank.org/cytobank/experiments/35226), in which naive CD4+ T cells were purified and stimulated by anti-CD3 and anti-CD28 antibodies.

Five cases were tested: unstimulated, stimulated by only the anti-CD3 antibody, stimulated by both the anti-CD3 and anti-CD28 antibodies, and two cases with different dosages for the anti-CD3 antibody (0.3 μg/mL and 0.8 μg/mL). The purpose of this study is identifying the associations of cell populations with elapsed time from the stimulations start point.

Thus, we use time, dosage, anti-CD3, and anti-CD28 as the covariate $\boldsymbol{X}$. All variables are treated as dummy variables.

It is known that stimulation of CD3 triggers activation of naive CD4+ T cells, which accompany the phosphorylation of SLP76/S6 and CD247 (pSLP76/pS6, pCD246) [12].

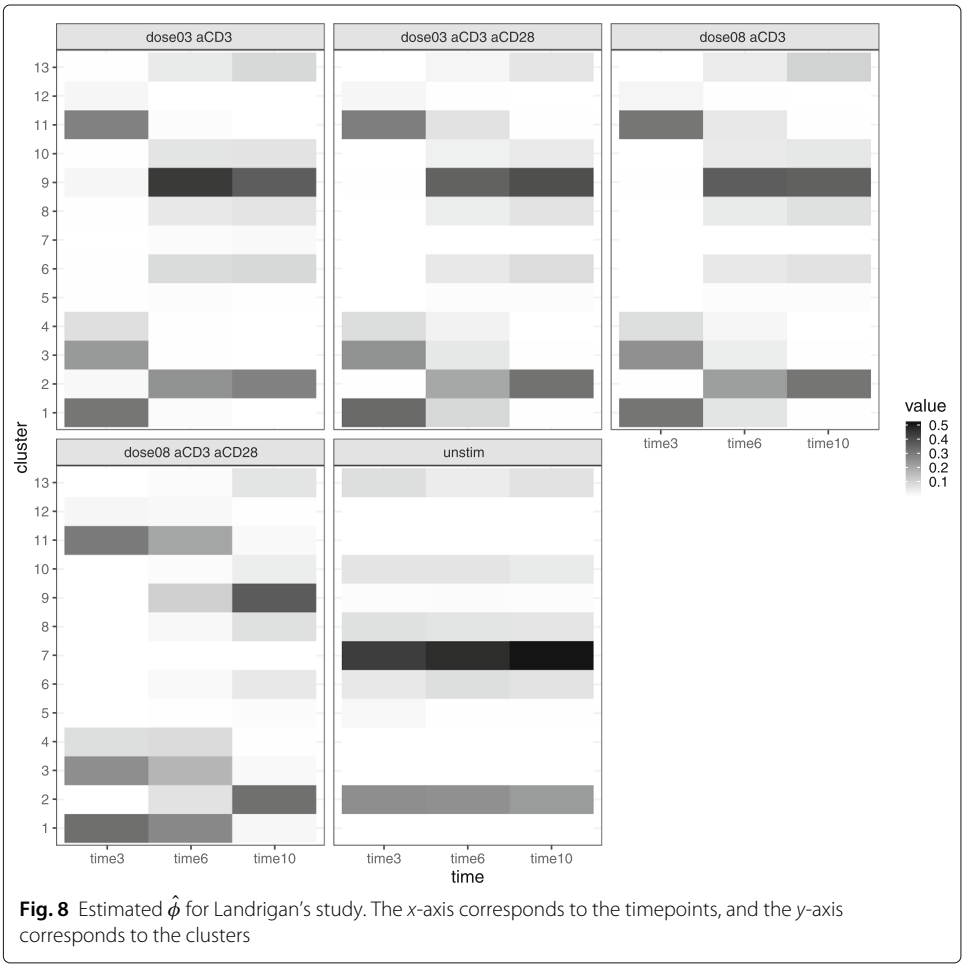

Fig. 8 Estimated $\hat{\phi}$ for Landrigan's study. The *x*-axis corresponds to the timepoints, and the *y*-axis corresponds to the clusters

CD28 is the co-stimulatory factor that enhances and prolongs T cell activation [13]. Soon after activation, the levels of phosphorylated SLP/S6 and CD247 decrease by negative feedback. Then the cells become CD45RO+ memory T cells. BIC (Fig. 6) determined the setting of 13 clusters. Figure 7 shows the $\hat{\boldsymbol{\mu}}$. Also as shown in Fig. 7, clusters 1, 3, 4, 11, and 12 are the pSLP76/pS6+ pCD247+activated naive T cells, and clusters 2, 9, and 10 are pSLP76/pS6- pCD247- CD45RO+ memory T cells. The mixture proportion is shown in Fig. 8. While the mixture proportion remains stable over time in unstimulated cases, other cases show a high proportion of activated naive T cells at 3 min and their proportion decreases at 6 and 10 min by deactivation through negative feedback. Figure 8 shows that as activated T cells decrease, the memory T cell population increases, indicating the transformation of naive T cells to memory T cells. This shows that in the case of flow cytometry data the method is able to provide a reasonable interpretation of the cell population clusters.

We also applied LAMBDA to mass cytometry data from Jin's study [14]. This data is available in FlowRepository (https://flowrepository.org/) under Repository ID: FR-FCM-ZY6C. The purpose of this study is discovering cell population related to clinical responses. In the case of mass cytometry data, for determination of the number of clusters, we used the elbow method, which is performed by plotting the SSE within each cluster against the number of clusters. In this case, the elbow method determined 14 clusters (Fig. 9).

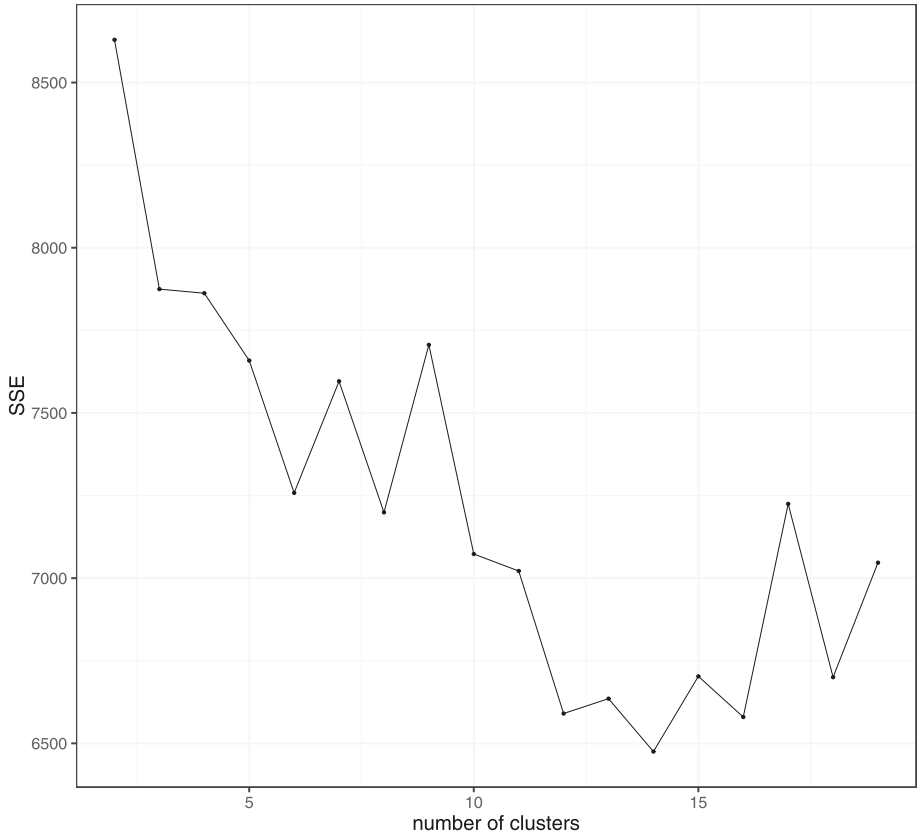

Fig. 9 Elbow plot for Jin's study. Decrease in SSE saturated at 8 clusters

Arranged by severity, clinical responses include healthy donors (ND), partial response (PR), stable disease (SD), and progressive disease (PD). ND were used as the baseline for ICB samples. Figure 10 shows the estimated mixture proportion $\hat{\phi}$. We observed that the value of the mixture proportion for cluster 2 increases as cancer progresses from PR to PD. Figure 11 shows the estimated $\hat{\mu}$. We denote hi and lo that the marker is high and low level expressed respectively. Cluster 2 is characterized by CD8+, T-bet lo, EOMES hi, PD1hi, and Ki67 lo. T-bet lo, EOMES hi, PD1hi, and Ki67 lo are exhaustion markers of the T cell. “Exhaustion" refers to cases where a T cell becomes dysfunctional due to the long-term induction of various co-repressive molecules such as PD-1, CTLA-4, and TIM-3. Pauken & Wherry [15] reported that the CD8+ T cells of T-bet hi and EOMES lo become T-bet lo, EOMES hi, and PD1 hi through exhaustion. The marker KI67 indicate cell mass culturing. The exhausted T cells have a low expression level in KI67. Blackburn [16] reported that the cell populations of T-bet lo, EOMES hi, and PD1 hi are not activated by blocking the PD1 / PDL-1 pathway with immune checkpoint inhibitors. LAMBDA shows that the cluster 2 cell population is high in patients who underwent a PDL1 inhibitor treatment with a poor prognosis. This finding is consistent with Pauken and Blackburn’s study, showing the effectiveness of LAMBDA in interpreting high dimensional mass cytometry data in real situations.

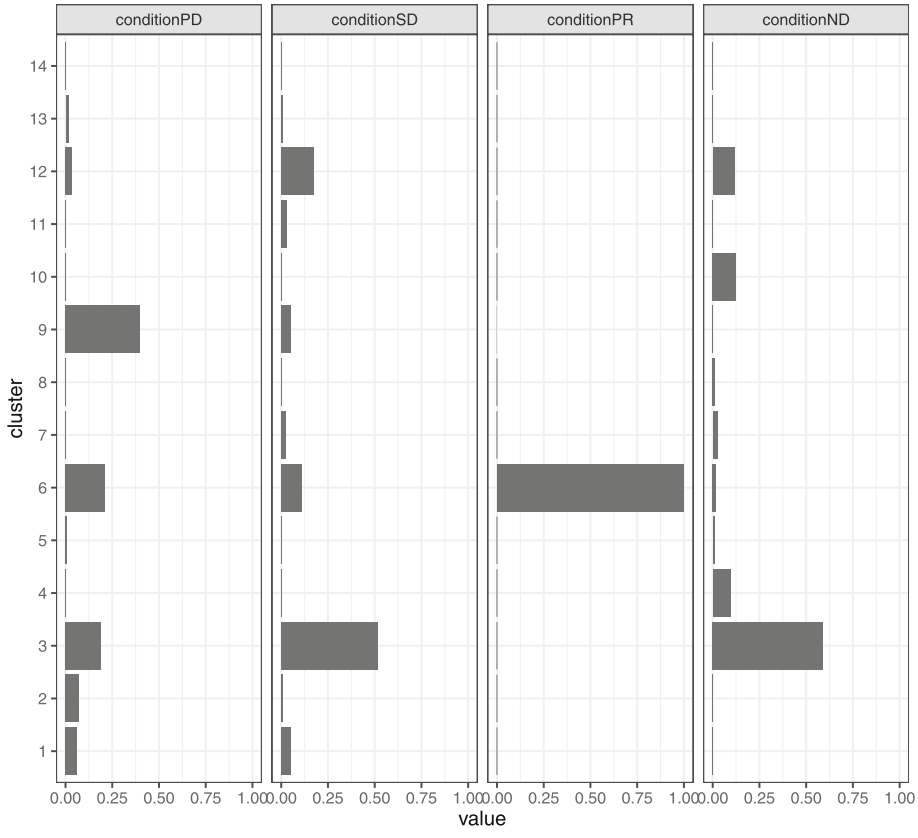

Fig. 10 Estimated $\hat{\phi}$ for Jin’s study. In the left figure, *x*-axis corresponds to the mixture proportion $\hat{\phi}$, and the *y*-axis corresponds to the clusters

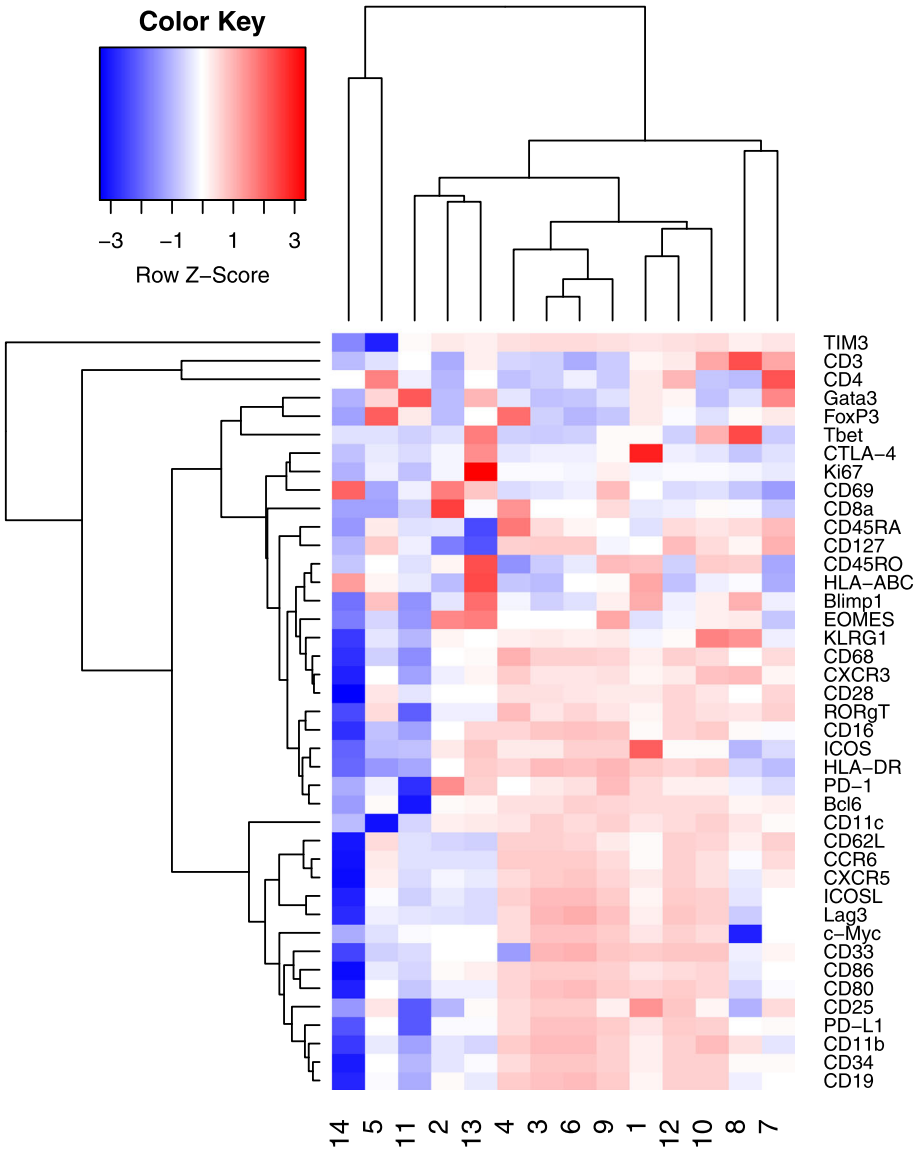

Fig. 11 Heatmap for Jin's study. *x*-axis corresponds to the markers, and the *y*-axis corresponds to the clusters

Through this analysis we can see that LAMBDA is a method that can efficiently estimate various clusters within cell populations and identify the associations between these cell clusters and their clinical outcomes in cases of both flow and mass cytometry data.

Discussion

LAMBDA should prove useful as it is described in this paper, but there is room for future study and improvement. Recently, with the development of next generation sequencing technologies, single cell sequencing was introduced to the field of biomedical research. Sequencing the DNA provides a higher resolution of cellular differences and a better understanding of the function of an individual cell in the context of its microenvironment. In this context, our future aim is the extension of LAMBDA for application to single cell

DNA data. This will allow us to understand the condition of the cell on a fundamental level, contributing to our overall understanding of biology and its processes.

Conclusion

With the development of high dimensional flow and mass cytometry data, researches have been challenged with the need to properly identify and interpret data about cell populations. To meet this challenge, we proposed a statistical framework that uses flow and mass cytometry data to discover cell clusters and the associations between individual clusters and clinical information.

As described in the "Methods" section, this model uses a stochastic EM to estimate parameters. In terms of computation, this parameter-estimation procedure offers an advantage over procedures that use an ordinary EM algorithm. This is because, in an algorithm that uses ordinary EM, the computational cost is large due to calculating the high-dimensional conditional expectation. By contrast, our procedure involves a Gibbs sampling that substitutes for this requirement, significantly reducing the computational cost.

In addition to being computationally efficient, our framework also has useful properties from the perspective of data analysis. Usual methods of clustering are not able to support the inclusion of explanatory variables. However, LAMBDA can include any explanatory variables. This property allows LAMBDA to analyze experimental results with various settings. Because of this novel feature, we expect that LAMBDA will be efficiently applied to studies that seek an association between cell populations and clinical information, advancing our ability to predict disease and predict outcomes of treatment.

Abbreviations

BIC: Bayesian information criterion; EM: Expectation-maximization; GMM: Gaussian mixture model; HIV: Human immunodeficiency virus; MAP: Maximum a posteriori probability; SE: Standard error

Acknowledgements

Not applicable.

About this supplement

This article has been published as part of *BMC Bioinformatics Volume 21 Supplement 13, 2020: Selected articles from the 18th Asia Pacific Bioinformatics Conference (APBC 2020): bioinformatics*. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-21-supplement-13.

Authors' contributions

KA and TS designed the proposed algorithm. KM, YM and HN designed the experiments. All authors have read and approved the final manuscript.

Funding

This research was supported by JSPS Grant-in-Aid for Scientific Research on Innovative Areas [No. 15H05912, 18H04798, and 19H05210]. It was also supported by the Japan Agency for Medical Research and Development (AMED) under the Strategic Research Program for Brain Sciences [No. JP18dm0107087], and under Practical Research Project for Rare / Intractable Diseases [No. JP17ek0109281]. The super-computing resources were provided by Human Genome Center, University of Tokyo. Publication costs are funded by all of the funding. The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

LAMBDA is implemented with R and is available from GitHub (https://github.com/abikoushi/LAMBDA).

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details
[1]Division of Systems Biology, Nagoya University Graduate School of Medicine, 65 Tsurumai-cho, Showa-ku, 4668550 Nagoya, Japan. [2]Division of Immunology, Nagoya University Graduate School of Medicine, 65 Tsurumai-cho, Showa-ku, 4668550 Nagoya, Japan. [3]Division of Cancer Immunology, Research Institute/EPOC, National Cancer Center, Chuo-ku tsukiji 5-1-1/Kashiwa-shi kashiwanoha 6-5-1, 1040045/2778577 Tokyo/Chiba, Japan.

Published: 17 September 2020

References

1. Spitzer MH, Nolan GP. Mass Cytometry: Single Cells, Many Features. Cell. 2016;165(4):780–91.
2. Bendall SC, Simonds EF, Qiu P, Amir E, Krutzik PO, Finck R, Bruggner RV, Melamed R, Trejo A, Ornatsky OI, Balderas RS, Plevritis SK, Sachs K, Pe'er D, Tanner SD, Nolan GP, et al. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. Sci (N Y). 2011;332(6030):687–96.
3. Rawstron AC, Kreuzer KA, Soosapilla A, Spacek M, Stehlikova O, Gambell P, Milani R, et al. Reproducible diagnosis of chronic lymphocytic leukemia by flow cytometry: An European Research Initiative on CLL (ERIC) & European Society for Clinical Cell Analysis (ESCCA) Harmonisation project. Cytom B Clin Cytom. 2018;94(1):121–8.
4. Abraham RS, Aubert G. Flow Cytometry, a Versatile Tool for Diagnosis and Monitoring of Primary Immunodeficiencies. Clin Vaccine Immunol. 2016;23(4):254–71. https://doi.org/10.1128/CVI.00001-16.
5. Smid M, Rodríguez-González FG, Sieuwerts AM, Salgado R, Prager-Van der Smissen WJ, van Der Vlugt-Daane M, Van de Vijver MJ, et al. Breast cancer genome and transcriptome integration implicates specific mutational signatures with immune cell infiltration. Nat Commun. 2016;7(1):2910.
6. Saeys Y, Van Gassen S, Lambrecht BN. Computational flow cytometry: helping to make sense of high-dimensional immunology data. Nat Rev Immunol. 2016;16(7):449.
7. Bruggner RV, Bodenmiller B, Dill DL, Tibshirani RJ, Nolan GP. Automated identification of stratifying signatures in cellular subpopulations. Proc Natl Acad Sci. 2014;111(26):E2770–7.
8. Lun A, Lun MA, BiocParallel D, biocViews FlowCytometry M. Package 'cydar'. 2017.
9. Weber LM, Nowicka M, Soneson C, Robinson MD. diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering. BioRxiv. 2018349738.
10. Fahey MT, Thane CW, Bramwell GD, Coward WA. Conditional Gaussian mixture modelling for dietary pattern analysis. J R Stat Soc Ser A Stat Soc. 2007;170(1):149–66.
11. Stephens M. Dealing with label switching in mixture models. J R Stat Soc Ser B Stat Methodol. 2000;62(4):795–809.
12. Gaud G, Lesourne R, Love PE. Regulatory mechanisms in T cell receptor signalling. Nat Rev Immunol. 2018;18(8):485–97.
13. Alegre ML, Frauwirth KA, Thompson CB. T-cell regulation by CD28 and CTLA-4. Nat Rev Immunol. 2001;1(3):220.
14. Jin G, Xue G, Wang RS, Wu LY, Lance M, Lu Y, Zhang W. Single-Cell Modeling of CD8+ T Cell Exhaustion Predicts Response to Cancer Immunotherapy. bioRxiv. 2018459867.
15. Pauken KE, Wherry EJ. Overcoming T cell exhaustion in infection and cancer. Trends Immunol. 2015;36(4):265–76.
16. Blackburn SD, Shin H, Freeman GJ, Wherry EJ. Selective expansion of a subset of exhausted CD8 T cells by α PD-L1 blockade. Proc Natl Acad Sci. 2008;105(39):15016–21.

Publisher's Note