RESEARCH

Model-based cell clustering and population tracking for time-series flow cytometry data

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

Abstract

Motivation: Modern flow cytometry technology has enabled the simultaneous analysis of multiple cell markers at the single-cell level, and it is widely used in a broad field of research. The detection of cell populations in flow cytometry data has long been dependent on "manual gating" by visual inspection. Recently, numerous software have been developed for automatic, computationally guided detection of cell populations; however, they are not designed for time-series flow cytometry data. Time-series flow cytometry data are indispensable for investigating the dynamics of cell populations that could not be elucidated by static time-point analysis. Therefore, there is a great need for tools to systematically analyze time-series flow cytometry data.

Results: We propose a simple and efficient statistical framework, named CYBERTRACK (CYtometry-Based Estimation and Reasoning for TRACKing cell populations), to perform clustering and cell population tracking for time-series flow cytometry data. CYBERTRACK assumes that flow cytometry data are generated from a multivariate Gaussian mixture distribution with its mixture proportion at the current time dependent on that at a previous timepoint. Using simulation data, we evaluate the performance of CYBERTRACK when estimating parameters for a multivariate Gaussian mixture distribution, tracking time-dependent transitions of mixture proportions, and detecting change-points in the overall mixture proportion. The CYBERTRACK performance is validated using two real flow cytometry datasets, which demonstrate that the population dynamics detected by CYBERTRACK are consistent with our prior knowledge of lymphocyte behavior.

Conclusions: Our results indicate that CYBERTRACK offers better understandings of time-dependent cell population dynamics to cytometry users by systematically analyzing time-series flow cytometry data.

Keywords: flow cytometry; time-series; topic model; Baysian inference

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

Background

Flow cytometry is a widely used technology for identifying and quantifying cellular properties and cell populations by measuring expression levels of surface and intracellular proteins at the single-cell level. Modern flow cytometers allow the simultaneous detection of nearly 20 protein markers per cell with a throughput of thousands of cells per second. The flow cytometry technique has greatly contributed to understanding the cellular biological processes and supporting clinical diagnoses in fields including immunology, cancer biology, and regenerative medicine [1, 2, 3].

An important challenge in the analysis of flow cytometry data is the classification of individual cells into canonical cell types, that is, subset populations such as T and B cells. The traditional approach of "manual gating" is performed by visually inspecting a two-dimensional scatter plot, but it suffers from several major limitations, including subjectivity, operator bias, difficulties in detecting unknown cell populations, and difficulties in reproducibility [4, 5, 6].

To overcome these limitations, several methods have been proposed for the computationally guided or automated detection of unknown cell populations by unsupervised clustering, including FlowSOM, X-shift, PhenoGraph, Rclusterpp, and

flowMeans [7]. Although these methods have been successfully applied to identify both major and rare cell populations, they are not designed for modeling and analyzing time-series data and thus cannot capture the time-dependent properties and dynamics of cell populations. For example, in clinical applications such as cancer immunotherapies, we are interested in investigating drug effects on cell populations by monitoring their dynamics throughout the treatment period [8]. Time-series flow cytometry data offer information on longitudinal cell population dynamics that could not be elucidated by conventional static time-point data. However, such research is currently limited by a lack of a systematic mathematical framework to adequately model and analyze time-series flow cytometry data.

To address this problem, we propose a new statistical framework, named CYBERTRACK (CYtometry-Based Estimation and Reasoning for TRACKing cell populations), for the automatic clustering and tracking of a mixture proportion of cell populations in time-series flow cytometry data. Our contributions are summarized as follows:

- Our framework is based on the Topic Tracking Model proposed by Iwata *et al.,* 2009, which is designed for tracking topic distribution that changes over time. We extend their model to handle time-series flow cytometry data, which is assumed to follow a multivariate Gaussian mixture distribution.
- By assuming that the mixture proportion at the current time is dependent on that at a previous time, CYBERTRACK is capable of estimating the longitudinal transition of multiple cell populations and detecting the "change-point" in the overall mixture proportion.
- We provide a simple and efficient learning procedure for the proposed model by using a stochastic EM algorithm, which is an alternate iteration of Gibbs sampling and maximum a posteriori (MAP) estimation of parameters. CYBERTRACK is implemented in an R environment, and the implementation is available from `https://github.com/kodaim1115/CYBERTRACK`.

A conceptual view of an analysis by CYBERTRACK is shown in Figure 1.

Our model and algorithm are described in the "Methods" section. To validate its performance and practicability, we applied CYBERTRACK to both simulation and real time-series flow cytometry datasets for two immunological experiments.

Methods

Model

Suppose that we observe time series flow cytometry data $y_{t,d,n} \in \mathbb{R}^K$, where $t \in \{1, ..., T\}$ is a time index, $d \in \{1, ..., D\}$ is a case index, $n \in \{1, ..., N_{t,d}\}$ is a sample index, and K is the number of markers. Here, $N_{t,d}$ represents the number of samples observed at time t for case d. The objective of this study is to perform clustering of samples and track the time-dependent transition of cluster mixture proportion $\pi_{t,d,l}$ for each case, where $l \in \{1, ..., L\}$ is a cluster index. Our model is inspired by the Topic Tracking Model, and is an extension of the multivariate Gaussian mixture model. Topic Model is a Bayesian model which that was originally designed to extract latent semantics, or "topics", from text data. Topic Tracking Model is an extension of Topic Model specialized in tracking time-varying topic distribution [9]. Although the original Topic Tracking Model assumed each word was generated

from a multinomial distribution, this assumption does not apply to the case with flow cytometry data. Therefore, we assumed that flow cytometry data follow a multivariate Gaussian mixture distribution, and we constructed the algorithm to estimate parameters. Here, topics correspond to cell populations such as T cells or B cells. Figure 2 illustrates a plate diagram of our proposed model, where, $\boldsymbol{z}_{t,d,n}$ is a latent cluster vector of length L that holds 1 for the l-th element when a sample is generated from cluster l and holds 0 otherwise. We assume that each sample is generated from a multivariate Gaussian mixture distribution with the parameter vector $\boldsymbol{\mu}_l$ and $\boldsymbol{\Sigma}_l$, which represents the mean and the covariance matrix for cluster l, respectively. More specifically, the generative process of CYBERTRACK is defined by

$$\boldsymbol{y}_{t,d,n} \mid \boldsymbol{z}_{t,d,n} \sim \text{Gaussian}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z) \tag{1}$$
$$\boldsymbol{z}_{t,d,n} \mid \boldsymbol{\pi}_{t,d} \sim \text{Categorical}(\boldsymbol{\pi}_{t,d}) \tag{2}$$
$$\boldsymbol{\pi}_{t,d} \mid \alpha_{t,d}, \boldsymbol{\pi}_{t-1,d} \sim \text{Dirichlet}(\alpha_{t,d}\boldsymbol{\pi}_{t-1,d}) \tag{3}$$
$$\boldsymbol{\mu}_l \mid \tau, \boldsymbol{\Sigma}_l \sim \text{Gaussian}(0, \tau^{-1}\boldsymbol{\Sigma}_l) \tag{4}$$
$$\boldsymbol{\Sigma}_l^{-1} \mid \nu, \boldsymbol{\Lambda} \sim \text{Wishart}(\nu, \boldsymbol{\Lambda}^{-1}) \tag{5}$$

where z is a latent cluster of the n-th sample at time t for case d indicated by $\boldsymbol{z}_{t,d,n}$, $\boldsymbol{\mu}_z$ and $\boldsymbol{\Sigma}_z$ are the mean vector and covariance matrix of the latent cluster, respectively, $\boldsymbol{\pi}_{t,d} = \{\pi_{t,d,l}\}_{l=1}^{L}$ is the mixture proportion vector, and $\alpha_{t,d}$ represents the persistency parameter, which indicates how consistent the mixture proportion at time t is compared with that at the previous time $t-1$. A smaller $\alpha_{t,d}$ value indicates a larger discrepancy between the mixture proportion at time t and $t-1$. Thus, timepoints with relatively small persistency parameters could be considered as "change-points" in the mixture proportion. $\boldsymbol{\mu} = \{\boldsymbol{\mu}_l\}_{l=1}^{L}$ is the mean vectors of clusters, and $\boldsymbol{\Sigma} = \{\boldsymbol{\Sigma}_l\}_{l=1}^{L}$ is the covariance matrices of clusters. τ is the hyperparameter of the $\boldsymbol{\mu}$ prior distribution, and $\boldsymbol{\Lambda}$ and ν are the hyperparameters of the $\boldsymbol{\Sigma}$ prior distribution.

Parameter Estimation

Parameter estimation in CYBERTRACK is based on the stochastic EM algorithm, which is an alternate iteration of Gibbs sampling and maximum a posteriori estimation of parameters. Suppose t is the current time, and suppose we have flow cytometry data matrix $\mathbf{Y}_t = \{\mathbf{Y}_{t,d}\}_{d=1}^{D}$ and a mixture proportion matrix $\boldsymbol{\Pi}_t = \{\boldsymbol{\pi}_{t,d}\}_{d=1}^{D}$, where $\mathbf{Y}_{t,d} = \{\boldsymbol{y}_{t,d,n}\}_{n=1}^{N_{t,d}}$. We perform the inference of latent clusters based on Gibbs sampling. Let $\mathbf{Z}_t = \{\boldsymbol{z}_{t,d}\}_{d=1}^{D}$ be the set of latent clusters of all cases at time t, where $\boldsymbol{z}_{t,d} = \{\boldsymbol{z}_{t,d,n}\}_{n=1}^{N_{t,d}}$. The posterior distribution of $\mathbf{Z}_t$ given $\mathbf{Y}_t$, $\boldsymbol{\Pi}_t$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ can be written as follows:

$$\begin{aligned} & p(\mathbf{Z}_t \mid \mathbf{Y}_t, \boldsymbol{\Pi}_t, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ & \quad \propto p(\mathbf{Y}_t, \mathbf{Z}_t, \boldsymbol{\Pi}_t, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ & \quad \propto p(\mathbf{Y}_t \mid \mathbf{Z}_t, \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\mathbf{Z}_t \mid \boldsymbol{\Pi}_t) \\ & \quad = \prod_d \prod_n p(\boldsymbol{y}_{t,d,n} \mid \boldsymbol{z}_{t,d,n}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\boldsymbol{z}_{t,d,n} \mid \boldsymbol{\pi}_{t,d}). \end{aligned} \tag{6}$$

The logarithm of above will be:

$$\begin{aligned}&\log\{p(\boldsymbol{y}_{t,d,n} \mid \boldsymbol{z}_{t,d,n}, \boldsymbol{\mu}, \boldsymbol{\Sigma})p(\boldsymbol{z}_{t,d,n} \mid \boldsymbol{\pi}_{t,d})\} \\ &\quad = \sum_l z_{t,d,n,l}\{-\frac{1}{2}(\boldsymbol{y}_{t,d,n} - \boldsymbol{\mu}_l)^\top \boldsymbol{\Sigma}_l^{-1}(\boldsymbol{y}_{t,d,n} - \boldsymbol{\mu}_l) \\ &\quad + \frac{1}{2}\log \mid \boldsymbol{\Sigma}_l^{-1} \mid + \log \pi_{t,d,l}\} + \text{const.}\end{aligned} \tag{7}$$

Therefore, $\boldsymbol{z}_{t,d,n}$ is sampled from the following categorical distribution:

$$\tilde{\boldsymbol{z}}_{t,d,n} \sim \text{Categorical}(\boldsymbol{\eta}_{t,d,n}) \tag{8}$$

$$\begin{aligned}\eta_{t,d,n,l} \propto \exp\{&-\frac{1}{2}(\boldsymbol{y}_{t,d,n} - \boldsymbol{\mu}_l)^\top \boldsymbol{\Sigma}_l^{-1}(\boldsymbol{y}_{t,d,n} - \boldsymbol{\mu}_l) \\ &+ \frac{1}{2}\log \mid \boldsymbol{\Sigma}_l^{-1} \mid + \log \pi_{t,d,l}\} \\ \text{s.t.} \sum_l &\eta_{t,d,n,l} = 1,\end{aligned} \tag{9}$$

where $\boldsymbol{\eta}_{t,d,n} = \{\eta_{t,d,n,l}\}_{l=1}^L$. Suppose we have the mean of the previous mixture proportion $\hat{\boldsymbol{\pi}}_{t-1,d}$. The persistency parameter $\alpha_{t,d}$ is estimated by fixed point iteration.

$$\hat{\alpha}_{t,d} \leftarrow \hat{\alpha}_{t,d}\frac{\sum_l \hat{\pi}_{t-1,d,l}A_{t,d,l}}{\psi(N_{t,d} + \hat{\alpha}_{t,d}) - \psi(\hat{\alpha}_{t,d})}, \tag{10}$$

where $A_{t,d,l} = \psi(N_{t,d,l} + \hat{\alpha}_{t,d}\hat{\pi}_{t-1,d,l}) - \psi(\hat{\alpha}_{t,d}\hat{\pi}_{t-1,d,l})$, $\psi(\cdot)$ is the digamma function $\psi(x) = \frac{\partial \log \Gamma(x)}{\partial x}$, and $N_{t,d,l}$ is the number of samples assigned to cluster l at time t for case d. The mean of $\pi_{t,d,l}$ is then calculated as follows:

$$\hat{\pi}_{t,d,l} = \frac{N_{t,d,l} + \hat{\alpha}_{t,d}\hat{\pi}_{t-1,d,l}}{N_{t,d} + \hat{\alpha}_{t,d}}. \tag{11}$$

We substitute the E-step of the EM algorithm by Gibbs sampling, then $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are updated in the M-step is as follows:

$$\hat{\boldsymbol{\mu}}_l = \frac{\sum_n \boldsymbol{y}_{n,l}}{N_l + \tau} \tag{12}$$

$$\begin{aligned}&\hat{\boldsymbol{\Sigma}}_l \\ &= \frac{\boldsymbol{\Lambda} + \sum_n (\boldsymbol{y}_{n,l} - \hat{\boldsymbol{\mu}}_l)^\top (\boldsymbol{y}_{n,l} - \hat{\boldsymbol{\mu}}_l) + \tau \hat{\boldsymbol{\mu}}_l^\top \hat{\boldsymbol{\mu}}_l}{N_l + \nu - K - 1},\end{aligned} \tag{13}$$

where $\boldsymbol{y}_{n,l}$ is the n-th sample assigned to cluster l, and N_l is the number of samples assigned to cluster l.

Result

Simulation Study

We conducted a simulation experiment to examine the performance of CYBERTRACK. We set $K = 10$, $T = 5$, and $D = 2$. The $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ was randomly generated. $\boldsymbol{\pi}$ were set manually to have change-points for mixture proportions. For the hyperparameters, we set $\tau = 10^{-5}$ and $\nu = K + 2$, and $\boldsymbol{\Lambda}$ was set to an identity matrix, which is equivalent to giving weakly informative priors. With this parameter setting, we randomly generated 1000 samples for each timepoint in each case (10,000 samples in total). The simulation was repeated 10 times, and different data was synthesized each time. The mean and standard error (se) for the estimated $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Sigma}}$, $\hat{\boldsymbol{\pi}}$, and $\hat{\alpha}$ are shown in Figure 3. The results show that the parameters for a multivariate Gaussian mixture distribution were reasonably estimated by the stochastic EM algorithm, and that CYBERTRACK successfully tracked the time-dependent transition of the mixture proportion in multiple cases. As shown in Figure 3d, $\hat{\alpha}$ holds small values at $t = 3, 5$ for case 1 and $t = 2, 4$ for case 2, indicating the dramatic transition of the mixture proportion at that timepoint.

Results for Real Data

To validate the CYBERTRACK performance for cell clustering and tracking mixture proportions of cell populations, we applied CYBERTRACK to real world flow cytometry data uploaded to Cytobank (https://www.cytobank.org/). In Landrigan's study (https://community.cytobank.org/cytobank/experiments/35226), naive CD4+ T cells were purified and stimulated using anti-CD3 and anti-CD28 antibodies. Five cases were tested: unstimulated, stimulated by only anti-CD3 antibody, and stimulated by both anti-CD3 and anti-CD28 antibodies, with two dosages tested for the anti-CD3 antibody (0.3 μg/mL and 0.8 μg/mL). It is known that the stimulation of CD3 triggers the activation of naive CD4+ T cells, which accompanies the phosphorylation of SLP76/S6 and CD247 (pSLP76/pS6, pCD246) [10]. CD28 is the co-stimulatory factor that enhances and prolongs T cell activation [11]. Soon after activation, the levels of pSLP/pS6 and pCD247 decrease owing to negative feedback. Consequently, the cells become CD45RO+ memory T cells.

To determine the number of clusters, we used the elbow method, which involves plotting the sum of squared error (SSE) within each cluster against the number of clusters. For Landrigan's study, the number of clusters was determined as 16 by using the elbow method (Figure 4a). Figure 4b shows the heatmap generated from the estimated $\hat{\boldsymbol{\mu}}$; clusters 1, 7, 8, 9, 12, and 16 are the pSLP76/pS6+ pCD247+ activated naive T cells, and clusters 2, 4, 6, and 13 are pSLP76/pS6- pCD247- CD45RO+ memory T cells. The time-dependent transition of the mixture proportion is shown in Figure 5. While the mixture proportion remains stable over time in unstimulated cases, other cases show dynamic fluctuation, as expected. In stimulated cases, a high proportion of activated naive T cells (specifically, cluster 16 and 7 for dosages 0.3 and 0.8 μg/mL, respectively) was observed at $t = 3$ min. Their proportions decreased through $t = 6, 10$ min by the T cell's negative feedback mechanism. Figure 5 shows that as the number of activated T cells decreases, the memory T cell populations increase, indicating the transformation of naive T cells into memory T cells. This behavior was well represented by $\hat{\alpha}$ estimates, shown in

Figure 6; the $\hat{\alpha}$ for stimulated cases shows small values at $t = 6$ and $t = 10$ min compared with that of unstimulated cases, indicating dynamic changes in cell population constitution at those timepoints. Interestingly, in cases stimulated with both anti-CD3 and anti-CD28 antibodies, a prominent increase of clusters with moderate levels of pSLP76/pS6 and pCD247 (cluster 14 and 1 for dosages 0.3 μg/mL and 0.8 μg/mL, respectively) was observed at $t = 6$. These clusters can be interpreted as cell populations that are transitioning from a highly activated state to an inactivated memory state. This is consistent with the well-known prolonged T cell activation by stimulation of CD28, thus further indicating that CYBERTRACK is capable of illustrating dynamic biological processes from time-series flow cytometry data [11].

We also applied CYBERTRACK to the data in Huang's study (https://community.cytobank.org/cytobank/experiments/5002), where cells were collected from mice whose lymph nodes were stimulated with either interleukin 7 (IL7) or interferon alpha (IFNα). It is known that IL7 and IFNα interact with their receptors on the lyphocytes' surface and activate lymphocytes through the phosphorylation of STAT family proteins (e.g., pSTAT1 and pSTAT5), which promotes the transcription of immune-related genes [12, 13].

The number of clusters was determined as 26 by using the elbow method (Figure 7a), and the heatmap is shown in Figure 7b. In Huang's study, T cells were identified by CD4 and/or TCRβ, and B cells were identified by B220. As shown in the heatmap, CYBERTRACK clustered cells into canonical cell types, which include CD4+ TCRβ+ T cells (clusters 4, 6, 9, 10, 17, and 26), CD4- TCRβ+ T cells (clusters 2, 3, 15, 24), and B220+ B cells (clusters 1, 3, 5, and 21). Clusters with extremely high levels of both B220 and TCRβ are thought to be debris; therefore, they were excluded from further interpretation. Figure 8 shows the time-dependent transition of the mixture proportion for each cell population. CYBERTRACK detected cell populations that increased over time in both cases. These cell populations include pSTAT1+ pSTAT5+ T cell (cluster 3) and pSTAT1+ B cell (cluster 5), which are typical cell populations that are known to emerge upon IL7 and IFNα stimulation. Furthermore, CYBERTRACK also illustrated cell population dynamics that differed in two cases; pSTAT5+ T cells (clusters 9 and 24) increased only when stimulated by IL7, whereas pSTAT1+ T cells (clusters 6 and 15) increased only in the IFNα-stimulated case. Although IL7 and IFNα are known to induce the phosphorylation of a variety of STAT family proteins, the result shown here may reflect the preferential upregulation of STAT5 and STAT1 by IL7 and IFNα, respectively [14, 15]. The estimated $\hat{\alpha}$ shows that the change-points are located at $t = 2, 4$ min for IL7 stimulation and $t = 2$ min for IFNα stimulation (Figure 9). Furthermore, analysis by CYBERTRACK revealed that stimulation by IL7 induces more dramatic changes in cell population constitution at an early stage (until $t = 4$ min), as indicated by the small $\hat{\alpha}$ values.

Discussion

Here, we propose a model-based cell clustering and population-tracking algorithm called CYBERTRACK. The aim of CYBERTRACK is to discover the underlying dynamics of cell populations in time-series flow cytometry data. Our model is inspired by the Topic Tracking Model [9], and we modified it for the parameter estimation of a multivariate Gaussian mixture distribution. CYBERTRACK is capable

of (*i*) cell clustering, (*ii*) tracking the mixture proportion of each cell population, and (*iii*) detecting the change-point in the overall mixture proportion.

Recently, a tool called mass cytometry was introduced to the field of biomedical research. Mass spectrometry-based detection of marker genes by mass cytometry has enabled the investigation of more than 40 markers simultaneously, providing much more informative data with higher-dimensions compared with fluorescence-based conventional flow cytometry. Recent research trends in single-cell biology highly depend on mass cytometry, and it has contributed to many important discoveries [16]. One limitation of CYBERTRACK is that it is inapplicable to mass cytometry data, because the data generated by mass cytometry do not follow a multivariate Gaussian distribution. Our future aim is to extend CYBERTRACK for application to time-series mass cytometry data.

The application of CYBERTRACK to simulation and real flow cytometry data has validated its performance for cell clustering and tracking mixture proportions in multiple cases. The results of CYBERTRACK analysis using two immunological experiments were consistent with our prior knowledge, which validates CYBERTRACK's ability to analyze time-series flow cytometry data. We believe that CYBERTRACK will be a powerful tool in various fields involving the investigation of cell population dynamics. For instance, in the field of cancer immunotherapy, the longitudinal immune monitoring of patients has become increasingly important as it provides information on the impact of therapeutic treatment on certain cell populations, or in finding cell populations that can be used as prognostic markers. Furthermore, CYBERTRACK will also be useful in basic research as it can give insights into flow cytometry time-series data in an unbiased manner. Because CYBERTRACK is capable of clustering cells from different cases, it is easy for researchers to compare population dynamics in experiments with a control and several cases.

Ethics approval and consent to paticipate
Not applicable.

Consent to publish
Not applicable.

Availability of data and materials
Our software is available from GitHub (https://github.com/kodaim1115/CYBERTRACK). Data for the immunological experiments used in this paper are available from Cytobank (https://cytobank.org/).

Funding
This work was supported by JSPS Grant-in-Aid for Young Scientists A (15H05325), and JSPS Grant-in-Aid for Scientific Research on Innovative Areas (15H05912 and 18H04798). The super-computing resources were provided by Human Genome Center, University of Tokyo.

Competing interests
The authors declare that they have no competing interests.

Author's contributions
KM, KA, TS designed the proposed algorithm. All authors reviewed and approved the final manuscript.

Acknowledgements
Not applicable.

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

References

1. Maeda, Y., Nishikawa, H., Sugiyama, D., Ha, D., Hamaguchi, M., Saito, T., Sakaguchi, S. *et al.*, (2014). Detection of self-reactive CD8+ T cells with an anergic phenotype in healthy individuals. *Science*, **346**(6216), 1536-1540.
2. Saito, T., Nishikawa, H., Wada, H., Nagano, Y., Sugiyama, D., Atarashi, K., Nagase, H. *et al.*, (2016). Two FOXP3+ CD4+ T cell subpopulations distinctly control the prognosis of colorectal cancers. *Nature medicine*, **22**(6), 679.
3. Tober, J., Maijenburg, M. M., Li, Y., Gao, L., Hadland, B. K., Gao, P., Speck, N. A. *et al.*, (2018). Maturation of hematopoietic stem cells from prehematopoietic stem cells is accompanied by up-regulation of PD-L1. *Journal of Experimental Medicine*, **215**(2), 645-659.
4. Finak, G., Frelinger, J., Jiang, W., Newell, E. W., Ramey, J., Davis, M. M., Gottardo, R. *et al.*, (2014). OpenCyto: an open source infrastructure for scalable, robust, reproducible, and automated, end-to-end flow cytometry data analysis. *PLoS computational biology*, **10**(8), e1003806.
5. Finak, G., Langweiler, M., Jaimes, M., Malek, M., Taghiyar, J., Korin, Y., Pontikos, N. *et al.*, (2016). Standardizing flow cytometry immunophenotyping analysis from the Human ImmunoPhenotyping Consortium. *Scientific reports*, **6**, 20686.
6. Saeys, Y., Van Gassen, S., & Lambrecht, B. N. (2016). Computational flow cytometry: helping to make sense of high-dimensional immunology data. *Nature Reviews Immunology*, **16**(7), 449.
7. Weber, L. M., & Robinson, M. D. *et al.*, (2016). Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. *Cytometry Part A*, **89**(12), 1084-1096.
8. Kurose, K., Ohue, Y., Wada, H., Iida, S., Ishida, T., Kojima, T., Kakimi, K. *et al.*, (2015). Phase Ia study of FoxP3+ CD4 Treg depletion by infusion of a humanized anti-CCR4 antibody, KW-0761, in cancer patients. *Clinical Cancer Research*, **21**(19), 4327-4336.
9. Iwata, T., Watanabe, S., Yamada, T., & Ueda, N. (2009, June). Topic tracking model for analyzing consumer purchase behavior. *In Twenty-First International Joint Conference on Artificial Intelligence*.
10. Gaud, G., Lesourne, R., & Love, P. E. (2018). Regulatory mechanisms in T cell receptor signalling. *Nat Rev Immunol*, **18**(8), 485-497.
11. Alegre, M. L., Frauwirth, K. A., & Thompson, C. B. (2001). T-cell regulation by CD28 and CTLA-4. *Nature Reviews Immunology*, **1**(3), 220.
12. Mackall, C. L., Fry, T. J., & Gress, R. E. (2011). Harnessing the biology of IL-7 for therapeutic application. *Nature Reviews Immunology*, **11**(5), 330.
13. O'Brien, T. R. (2009). Interferon-alfa, interferon-lambda and hepatitis C. *Nature genetics*, **41**(10), 1048.
14. Khodarev, N. N., Roizman, B., & Weichselbaum, R. R. (2012). Molecular pathways: interferon/stat1 pathway: role in the tumor resistance to genotoxic stress and aggressive growth. *Clinical cancer research*, **18**(11), 3015-3021.
15. Foxwell, B. M., Beadling, C., Guschin, D., Kerr, I., & Cantrell, D. (1995). Interleukin-7 can induce the activation of Jak 1, Jak 3 and STAT 5 proteins in murine T cells. *European journal of immunology*, **25**(11), 3041-3046.
16. Yao, Y., Liu, R., Shin, M. S., Trentalange, M., Allore, H., Nassar, A., Montgomery, R. R. *et al.*, (2014). CyTOF supports efficient detection of immune cell subsets from small samples. *Journal of immunological methods*, **415**, 1-5.

Figures

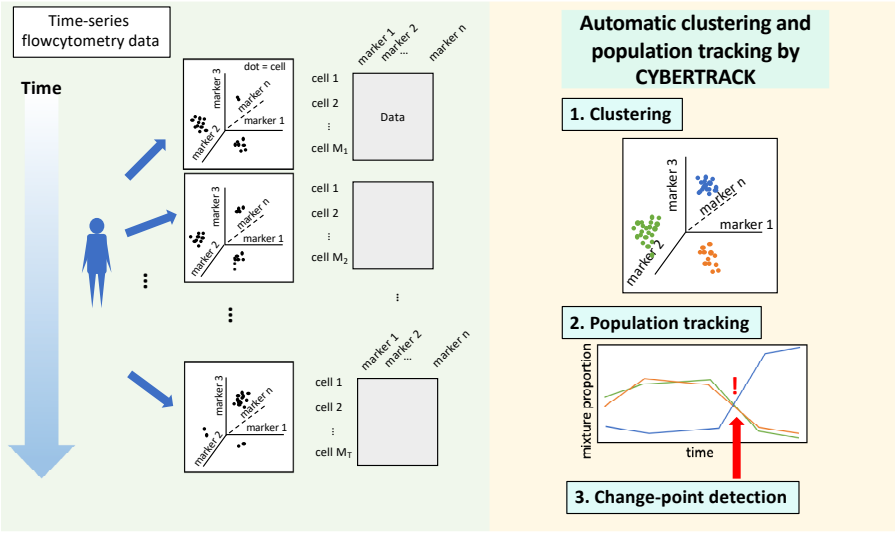

Figure 1 Conceptual view of the analysis by CYBERTRACK. The aim of CYBERTRACK is to model and analyze time-series flowcytometry data to understand dynamic cell population behavior that spans certain period of time. In time-series flow cytometry analysis, cells are acquired sequentially and their expression levels of marker proteins are analyzed, giving data matrices of cells and markers for each timepoint. CYBERTRACK takes these data matrices as an input and performs i) clustering, ii) tracking cell population dynamics, and iii) detecting change-points in cell population constitution.

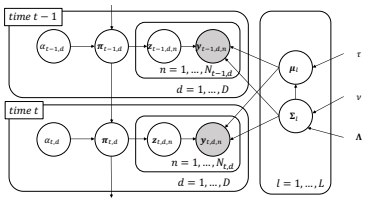

Figure 2 Graphical model of CYBERTRACK.

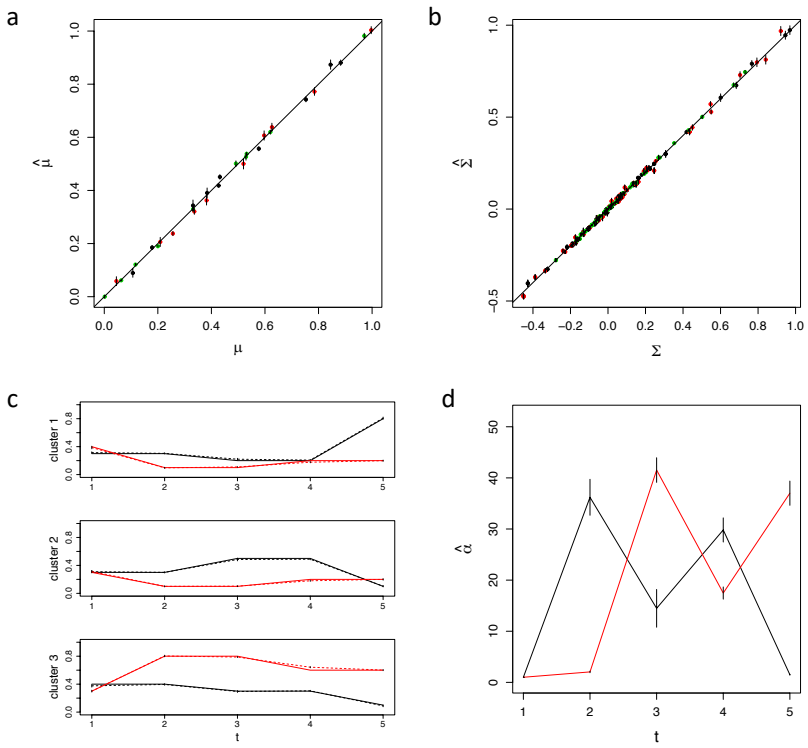

Figure 3 Simulation result for CYBERTRACK analysis. **a**, Estimated $\hat{\boldsymbol{\mu}}$ were plotted against true $\boldsymbol{\mu}$. Each dot represents elements of $\hat{\boldsymbol{\mu}}$ and was color-coded by cluster. **b**, Estimated $\hat{\boldsymbol{\Sigma}}$ values were plotted against true $\boldsymbol{\Sigma}$. Each dot represents elements of $\hat{\boldsymbol{\Sigma}}$ was were color-coded by cluster. **c**, Estimated $\hat{\boldsymbol{\pi}}$ for simulation data. Black and red lines represent the proportion mixture for case 1 and 2, respectively. Solid lines indicate the true mixture proportion and dashed lines indicate estimated proportion. **d**, Black and red lines represent proportion mixtures for cases 1 and 2, respectively. Timepoints where the alpha values decrease substantially indicate change-points in the overall mixture proportion. Error bars represent standard error.

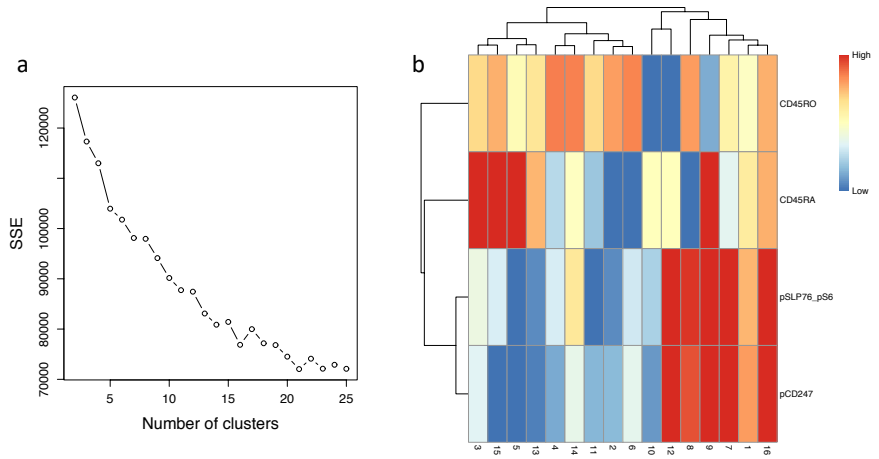

Figure 4 Clustering result for CYBERTRACK analysis on Landrigan's study. a, Elbow plot for Landrigan's study. SSE saturated at 16 clusters. b, Heatmap for Landrigan's study. Cluster number is shown on the x axis and markers are shown on the y axis.

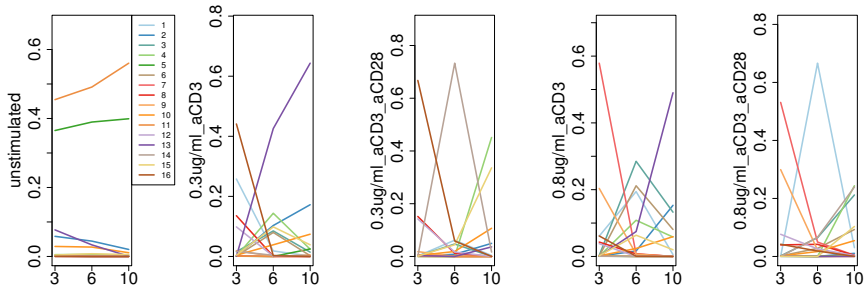

Figure 5 Cell population tracking result for Landrigan's study. Mixture proportions for five cases were drawn in different graphs. Each colored line represents the mixture proportion for a different cluster. t is on the x axis and the mixture proportion is on the y axis.

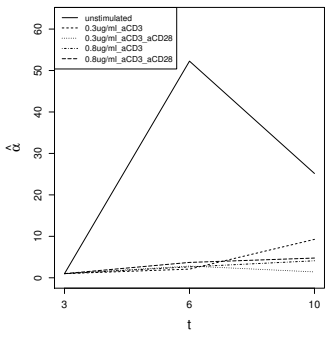

Figure 6 Estimated $\hat{\alpha}$ for Landrigan's study.

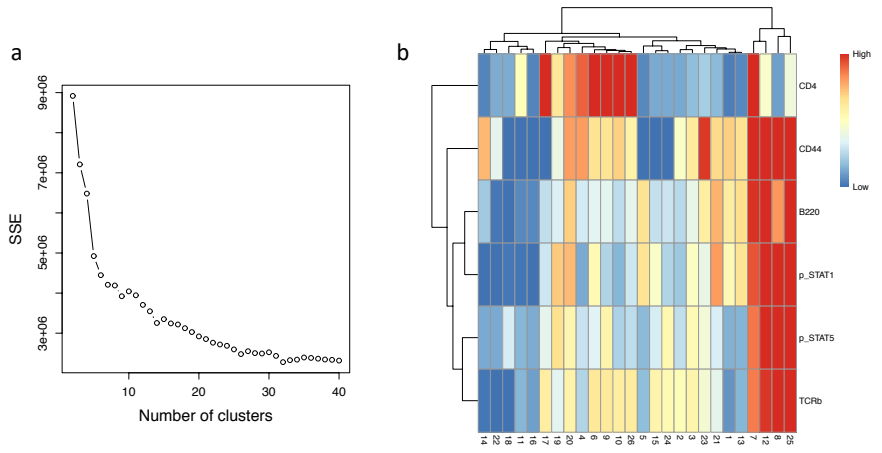

Figure 7 Clustering result for Huang's study. **a**, Elbow plot for Landrigan's study. SSE saturated at 16 clusters. **b**, Heatmap for Landrigan's study. Cluster number is shown on the x axis and markers are shown on the y axis.

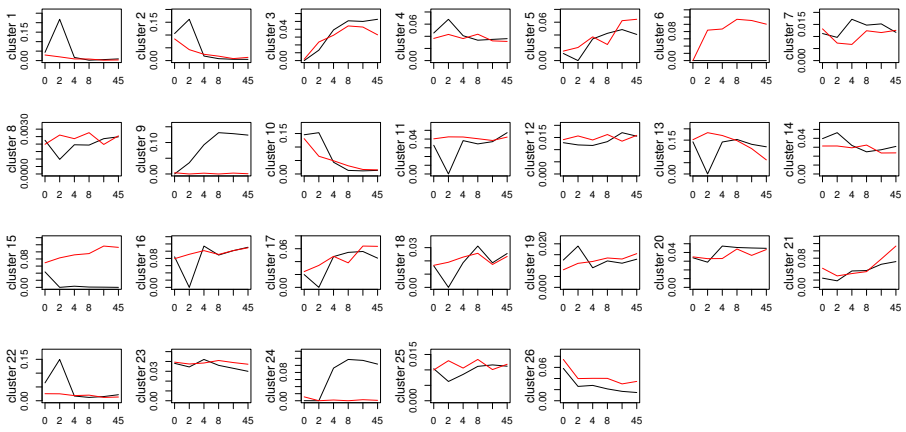

Figure 8 Cell population tracking result for Huang's study. Mixture proportions for 26 clusters were drawn in different graphs. Black and red lines represent IL7 stimulation and IFNα stimulation, respectively. t is on the x axis and the mixture proportion is on the y axis.

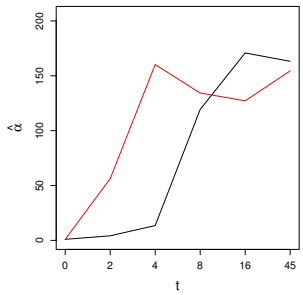

Figure 9 Estimated $\hat{\alpha}$ for Huang's study. Black and red lines represent IL7 stimulation and IFNα stimulation, respectively.