Bioinformatics, 37(11), 2021, 1632–1634
doi: 10.1093/bioinformatics/btaa873
Advance Access Publication Date: 14 October 2020
Applications Note

Systems biology

CYBERTRACK2.0: zero-inflated model-based cell clustering and population tracking method for longitudinal mass cytometry data

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

[1]Division of Systems Biology, [2]Division of Immunology, Graduate School of Medicine, Nagoya University, Nagoya 4668550, Japan and [3]Division of Cancer Immunology, Research Institute/EPOC, National Cancer Center, Tokyo, Chiba 1040045/2778577, Japan

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Martelli Pier Luigi

Received on March 24, 2020; revised on September 16, 2020; editorial decision on September 24, 2020; accepted on September 28, 2020

Abstract

Summary: Recent advancements in high-dimensional single-cell technologies, such as mass cytometry, enable longitudinal experiments to track dynamics of cell populations and identify change points where the proportions vary significantly. However, current research is limited by the lack of tools specialized for analyzing longitudinal mass cytometry data. In order to infer cell population dynamics from such data, we developed a statistical framework named CYBERTRACK2.0. The framework's analytic performance was validated against synthetic and real data, showing that its results are consistent with previous research.

Availability and implementation: CYBERTRACK2.0 is available at https://github.com/kodaim1115/CYBERTRACK2.

Contact: shimamura@med.nagoya-u.ac.jp

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

High-dimensional single-cell technology, such as mass cytometry or cytometry by time-of-flight, provides the ability to investigate the expression patterns of pre-defined sets of surface and intracellular proteins at single cell resolutions (Spitzer and Nolan, 2016). Recently, longitudinal analysis using mass cytometry has yielded important information that cannot be obtained using conventional analysis of static time points. In the field of cancer immunity, mass cytometry analysis of tumor samples of the same patient at different time points is increasingly being utilized to better understand response and resistance to immune checkpoint blockade (Chen *et al.*, 2016; Greenplate *et al.*, 2016). For example, longitudinal mass cytometry analysis of paired peripheral blood biopsies from before and after anti-PD-1 treatment has revealed that the frequency of a certain monocyte subset was strongly associated with the patients' responsiveness to the treatment (Krieg *et al.*, 2018).

One of the main objectives of analyzing longitudinal cytometry data is to identify the underlying dynamics of cell populations and to track their temporal fluctuation. Recently, we proposed a Topic Tracking Model-based statistical framework named CYBERTRACK designed for analyzing longitudinal flow cytometry data (Iwata *et al.*, 2009; Minoura *et al.*, 2019). Although it is a powerful tool to discover cell population dynamics from such data, it has some limitations. One limitation is that it cannot be used to analyze mass cytometry data directly due to the high proportion of zeros in the data, so it does not follow the assumed probability distribution in CYBERTRACK. A zero in the mass cytometry data indicates that the number of metal isotopes was below the detection limit of the instrument, as the amount of marker protein expression in the cells was low. These zero values are typically substituted by random numbers to avoid computational problems occurring from cells having the same value. Although this approach is commonly adopted for practical convenience, it underutilizes the information the data possesses. Another limitation is that CYBERTRACK uses a stochastic expectation-maximization (EM) algorithm, so it is not suitable for detecting rare cell populations (Naim and Gildea, 2012). Like the EM algorithm, the stochastic EM algorithm often misses very rare populations when they exist near large populations. In these cases, it tends to lump rare populations and larger populations, possibly leading to a misunderstanding of the data. Because studies using mass cytometry often aim to discover the dynamics of rare cell populations that consists of $<1\%$ of the total population, tools for correctly identifying such populations are extremely important. In order to address these problems; here, we present an updated version of CYBERTRACK, CYBERTRACK2.0, for the automatic clustering and tracking of proportionally mixed cell populations in longitudinal mass cytometry data.

© The Author(s) 2020. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

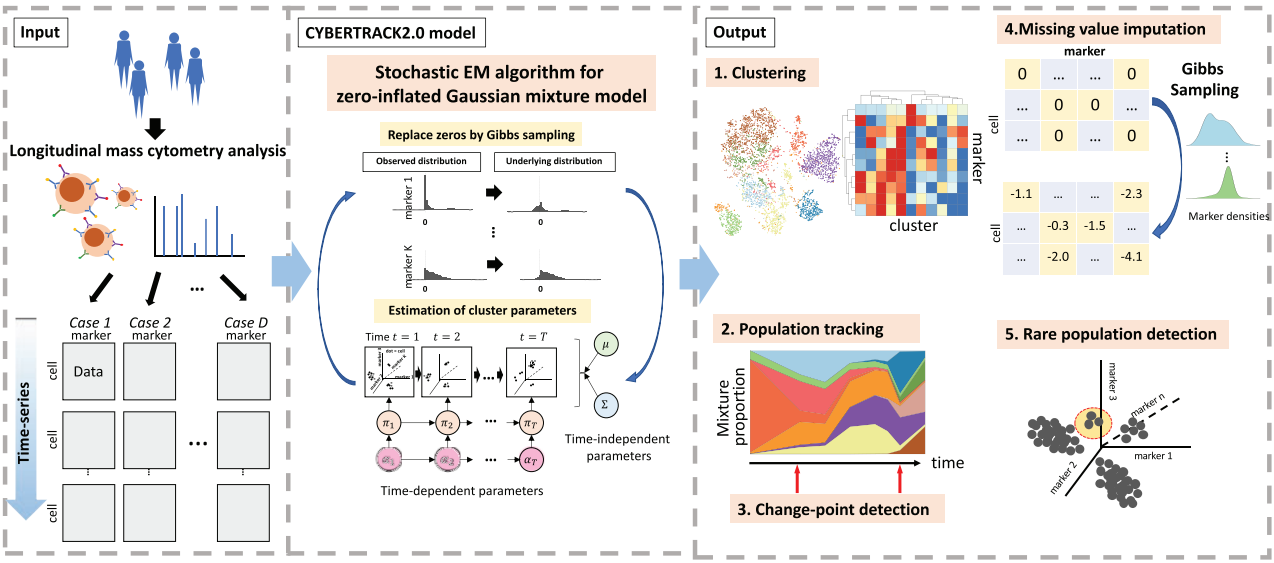

Fig. 1. Conceptual view of CYBERTRACK2.0. Our method takes longitudinal mass cytometry data as an input. Inference process of CYBERTRACK2.0 is based on stochastic EM algorithm for zero-inflated GMM, which consists of (i) replacing zeros by Gibbs sampling from underlying distributions and (ii) estimation of cluster parameters. As an output, CYBERTRACK2.0 provides information on cell clustering, cell population tracking, and change-points in overall mixture proportion. It can impute missing values in mass cytometry data by Gibbs sampling from estimated probability distributions. Also, it implements modified weighted iterative sampling algorithm to find very rare cell populations

2 Materials and methods

The improvements of CYBERTRACK2.0 are summarized as follows: (i) a new probabilistic model for generating mass cytometry data based on a zero-inflated multivariate Gaussian mixture distribution that can handle the high amount of zeros in mass cytometry data. (ii) A new algorithm for detecting rare cell populations that uses the stochastic EM algorithm combined with weighted iterative sampling (Naim *et al.*, 2014).

Figure 1 shows a conceptual view of CYBERTRACK2.0 analysis flow. We provided an efficient and straightforward algorithm for estimating parameters of the proposed model. A detailed explanation of our model and estimation procedure is described in Supplementary Material.

3 Results and discussion

Using simulation and real experimental mass cytometry data, we validated the cell clustering, cell population tracking and change point detection performance of CYBERTRACK2.0. First, we conducted a simulation study by generating a synthetic longitudinal mass cytometry dataset, which includes rare cell populations with larger populations (from 1% to 30%) (Supplementary Figs S3 and S4). Adding to it, we show that imputation of missing values (zeros) by Gibbs sampling provides approximate mean expression levels below detection limit of mass cytometry. Using this synthetic data, we compared the performance of CYBERTRACK2.0 with the original version of CYBERTRACK and a Gaussian mixture model (GMM). As a result, we confirmed that CYBERTRACK2.0 performs better in clustering cells when compared to the other methods (Supplementary Fig. S4).

In addition, using pseudo-longitudinal data generated from ground truth mass cytometry data, we compared clustering performance of our model with FlowSOM and PhenoGraph (Levine *et al.*, 2015; Van Gassen *et al.*, 2015). We show that clustering performance of CYBERTRACK2.0 is better or comparable to these state-of-the-art methods (Supplementary Figs S7 and S8). These simulation studies validated that CYBERTRACK2.0 has high clustering performance. Furthermore, the ability of CYBERTRACK2.0 is not restricted to clustering; our method produces reasonable estimates for the zero-inflated multivariate Gaussian mixture distribution, and accurately tracks cell population dynamics, and can detect change-points (Supplementary Fig. S6). Also, zero replacement by Gibbs sampling provides imputed data for other downstream analysis. For detailed information on the simulation study, see Supplementary Material.

Next, we validated the performance of CYBERTRACK2.0 using two real longitudinal mass cytometry datasets on cancer immunology and hematopoietic development (Krieg *et al.*, 2018; Palii *et al.*, 2019). Overall, the cell populations detected using our method were in agreement with the well-known cell lineages. An important result is that it could capture major to very rare cell populations, verifying the effectiveness of using our method in practical situations. In cancer immunology data, CYBERTRACK2.0 illustrated the enrichment of HLA-DR+ myeloids in patients responsive to anti-PD-1 treatment (Supplementary Figs S9–S11). Furthermore, analysis by CYBERTRACK2.0 discovered that the treatment triggers different dynamics among HLA-DR+ myeloid clusters, which may lead to more precise characterization of this potential prognostic marker population (Supplementary Figs S9–S11). For the hematopoietic development data, CYBERTRACK2.0 was able to systematically analyze dynamic emergence of cell lineages from hematopoietic stem and progenitor cells to erythrocytes and megakaryocytes (Supplementary Fig. S14), consistent with the original report (Palii *et al.*, 2019). For detailed explanation of these results, see Supplementary Material.

In summary, we proposed CYBERTRACK2.0, a novel statistical framework for longitudinal mass cytometry data analysis. It is based on topic tracking model and zero-inflated multivariate Gaussian mixture distribution to deal with the previously unsolved problems, such as (i) clustering of cells with longitudinal constraints and (ii) utilization of zeros in mass cytometry data. In addition, weighted iterative sampling was implemented in our method to maximize the chances of detecting rare cell populations of interest. Furthermore, users can use data imputed by CYBERTRACK2.0 for other downstream analysis such as pseudotime estimation or batch effect removal. We believe that CYBERTRACK2.0 is a powerful tool for researchers aiming to obtain biological or clinical insights from longitudinal mass cytometry data.

Funding

This research was supported by JSPS Grant-in-Aid for Scientific Research under grant No. 18H04798, 19H05210, 20H04841, and 20H04281. It was

also supported by the Japan Agency for Medical Research and Development (AMED) under grant No. JP19dm0107087h0004, JP19km0405207h9904, and JP19ek0109281h0003. The super-computing resources were provided by Human Genome Center, the University of Tokyo.

Conflict of Interest: none declared.

Data availability

Our codes are available at https://github.com/kodaim1115/CYBERTRACK2. Pseudo-longitudinal data was generated from data provided at https://flowrepository.org/id/FR-FCM-ZZPH. Mass cytometry data on cancer immunity is available at https://flowrepository.org/experiments/1124. Mass cytometry data on hematopoiesis is available at https://flowrepository.org/id/FR-FCM-ZYPT.

References

Chen,P.L. *et al.* (2016) Analysis of immune signatures in longitudinal tumor samples yields insight into biomarkers of response and mechanisms of resistance to immune checkpoint blockade. *Cancer Discov.*, **6**, 827–837.

Greenplate,A.R. *et al.* (2016) Systems immune monitoring in cancer therapy. *Eur. J. Cancer*, **61**, 77–84.

Iwata,T. *et al.* (2009) Topic tracking model for analyzing consumer purchase behavior. In: *Twenty-First International Joint Conference on Artificial Intelligence,Pasadena, California, pp. 1427–1439.*

Krieg,C. *et al.* (2018) High-dimensional single-cell analysis predicts response to anti-PD-1 immunotherapy. *Nat. Med.*, **24**, 144–153.

Levine,J.H. *et al.* (2015) Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell*, **162**, 184–197.

Minoura,K. *et al.* (2019) Model-based cell clustering and population tracking for time-series flow cytometry data. *BMC Bioinformatics*, **20**, 1–10.

Naim,I. *et al.* (2014) SWIFT-scalable clustering for automated identification of rare cell populations in large, high-dimensional flow cytometry datasets, Part 1: algorithm design. *Cytometry A*, **85**, 408–421.

Naim,I. and Gildea,D. (2012) Convergence of the em algorithm for gaussian mixtures with unbalanced mixing coefficients. *In Proceedings of the 29th International Conference on Machine Learning (ICML-12),* Edinburg, Scotland, pp. 1655–1662.

Palii,C.G. *et al.* (2019) Single-cell proteomics reveal that quantitative changes in co-expressed lineage-specific transcription factors determine cell fate. *Cell Stem Cell*, **24**, 812–820.

Spitzer,M.H. and Nolan,G.P. (2016) Mass cytometry: single cells, many features. *Cell*, **165**, 780–791.

Van Gassen,S. *et al.* (2015) FlowSOM: using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry A*, **87**, 636–645.