ELSEVIER

Blood 142 (2023) 905–906

The 65th ASH Annual Meeting Abstracts

ORAL ABSTRACTS

803.EMERGING TOOLS, TECHNIQUES AND ARTIFICIAL INTELLIGENCE IN HEMATOLOGY

Machine Learning (ML)-Enabled Automation for High-Throughput Data Processing in Flow Cytometry

Anna L. Kamysheva, MSc[1], *Dmitrii V. Fastovets, MSc*[1], *Roman N. Kruglikov, MSc*[1], *Arseniy A. Sokolov, BSc*[1], *Anastasiya S. Fefler, MSc*[1], *Anastasiia A. Bolshakova, MSc*[1], *Anastasia Radko, BSc*[1], *Ilya E. Krauz, MSc*[1], *Sheila T. Yong*[1], *Michael Goldberg, PhD*[1], *Ravshan Ataullakhanov, PhD*[1], *Aleksandr Zaitsev, MSc*[1]

[1] BostonGene, Corp., Waltham, MA

Introduction

Flow cytometry is widely used in clinical and research laboratories for diagnostics, biomarker discovery, and immune system monitoring. Flow cytometry data processing still uses gating- and clustering-based approaches that are highly time-consuming and subjective. Data processing time increases with panel size and number of detected populations, posing challenges to the search for new biomarkers. Low reproducibility and method limitations have thus far hindered efforts to automate and standardize flow cytometry data processing; hence, these efforts have not yielded any significant advancements in data processing methods. Here we present a new ML-based algorithm for automated cell-type labeling. Our supervised ML approach allows us to classify every event in a flow cytometry data file solely based on the presence and absence of markers, without the need for prior knowledge or assumption about cell population content in the sample. This approach enables the detection of rare and/or new cell populations with a high average quality metric (f1-score). The rapid and high-quality analysis our algorithm can perform renders it applicable in clinical settings, particularly for detecting hematological abnormalities and cancers.

Methods

We processed 500 blood samples from a cohort of healthy donors and patients with various cancer diagnoses using 10 different 18-channel multicolor flow cytometry panels. We then used data from either the entire or a portion of these 500 samples in a 3:1 split for training:test datasets to train and test our algorithm on each cytometry panel. To do this, we manually matched cells with certain cellular phenotypes to create 10 high-quality training sets for supervised learning and 10 test datasets, one pair for each of the 10 panels.

To train the cell type classifier, we set up a two-level boosting-based model. The first-level model filters out outliers, including dead cells, cellular debris, beads, and other undefined particles, in order to hone in on the target population.

The second-level model for predicting cell types within a target population is defined by two approaches. The population-based approach detects major subpopulation types in a target population and predicts the precise population labels. This approach is useful for labeling a small number of previously known or predicted subpopulations. The marker-based approach is useful for target populations with large numbers of subpopulations, such as T cells harboring different combinations of cell-surface receptors. It predicts the presence or absence of specific markers on each cell to assign its phenotype. It also allows us to construct complex hierarchies in order to detect new populations that are challenging to identify manually. Figure 1 outlines our workflow.

Results

We validated our final set of 10 trained models on our test dataset. The summarized number of detected cell populations in the test dataset was 221, which corresponds to the number of unique cell types predicted by our models. Table 1 shows the evaluation metrics for our algorithm for populations with > 0.1% whole blood cells (WBCs).The average quality metric (f1-score) for all antibody panels used is 0.86. This value is the mean of all f1-scores calculated for all cell populations identified by our algorithm. Mean f1-score is the highest (0.96) for large populations, lower (0.87) for mid-sized populations, and lowest but acceptable (0.77) for small populations. Mean quality score for the marker-based models is also high (0.96). Compared to manual evaluation that took approximately 1 hour to analyze one data file, the algorithm completed analysis within 10 seconds.

Conclusion

Our new algorithm automates cell labeling and produces high-quality outputs that are comparable to manual processing, but with a much shorter turnaround time (TAT) and without the need for prior knowledge or expert competence from the user.

Importantly, it allows us to effectively and accurately filter out outliers, identify the target population, and divide this target population into multiple cell subtypes including new and rare cell subpopulations, all without a priori assumptions about cell population content in the sample. Given its ability to perform high-quality cell population analysis and its short TAT, our algorithm provides rapid, unbiased, and precise cell typing that will have utility for the diagnosis of heme malignancies and immunoprofiling.

Disclosures Kamysheva: *BostonGene:* Current Employment, Current equity holder in private company, Current holder of *stock options* in a privately-held company, Patents & Royalties: Patents. **Fastovets:** *BostonGene:* Current Employment, Patents & Royalties: Patents. **Kruglikov:** *BostonGene:* Current Employment, Patents & Royalties: Patents. **Sokolov:** *BostonGene:* Current Employment, Patents & Royalties: Patents. **Fefler:** *BostonGene:* Current Employment. **Bolshakova:** *BostonGene:* Current Employment. **Radko:** *BostonGene:* Current Employment. **Krauz:** *BostonGene:* Current Employment, Current equity holder in private company, Current holder of *stock options* in a privately-held company, Patents & Royalties: Patents; *Merck:* Current equity holder in publicly-traded company; *Gilead:* Current equity holder in publicly-traded company; *Pfizer:* Current equity holder in publicly-traded company; *Bayer:* Current equity holder in publicly-traded company. **Yong:** *BostonGene:* Current Employment. **Goldberg:** *BostonGene:* Current Employment, Current equity holder in private company, Current holder of *stock options* in a privately-held company, Patents & Royalties: Patents. **Ataullakhanov:** *BostonGene:* Current Employment, Current equity holder in private company, Current holder of *stock options* in a privately-held company, Patents & Royalties: Patents. **Zaitsev:** *BostonGene:* Current Employment, Current equity holder in private company, Current holder of *stock options* in a privately-held company, Patents & Royalties: Patents; *Illumina:* Current equity holder in publicly-traded company; *Adaptive Biotechnology:* Current equity holder in publicly-traded company.

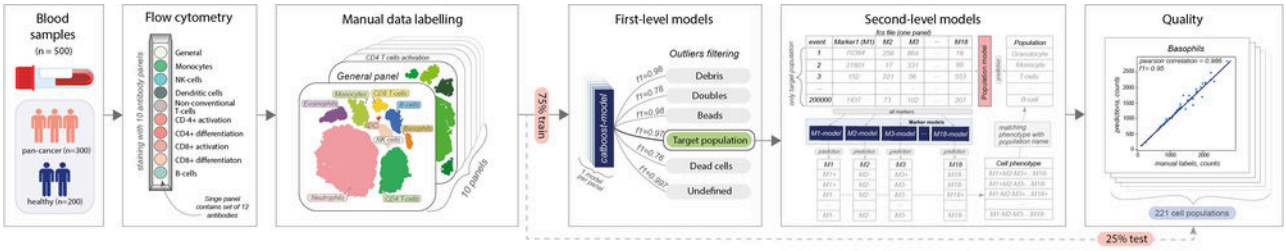

Figure. 1. *A schematic that outlines the process of training and validating cell-labeling models.*

Group	Population name	Median abundance in blood, % of WBCs	Mean f1-score	Mean Pearson correlation
	Large populations (18 populations)	**> 5%**	**0.96**	**0.98**
Granulocytes	Granulocytes	66.8	0.98	0.997
CD4+ T-cells	CD4+ T-cells	12.96	0.99	0.998
CD8+ T-cells	CD8+ T-cells	5.64	0.98	0.987
Monocytes	Monocytes	6.68	0.92	0.957
Monocytes	Classical monocytes	5.79	0.93	0.98
	Mid-size populations (67 populations)	**< 5 % and > 0.5 %**	**0.87**	**0.94**
Granulocytes	Basophils	0.57	0.95	0.986
CD4+ T-cells	CD4+ Central Memory CXCR3+	1.36	0.85	0.941
CD8+ T-cells	CD8+ TEMRA	0.69	0.85	0.96
NK-cells	NK-cells	2.66	0.97	0.995
B-cells	Naive B-cells	1.55	0.91	0.972
	Small populations (136 populations)	**< 0.5 % and > 0.1 %**	**0.77**	**0.9**
CD4+ T-cells	CD4+ Central Memory CCR4+ CCR6+ CXCR3- CXCR5- (Th17)	0.27	0.77	0.93
CD8+ T-cells	CD8+ Central Memory PD-1+	0.14	0.71	0.84
Monocytes	HLA-DR-low Monocytes	0.31	0.8	0.978
NK-cells	Mature NK cells CD158+ CD57+	0.23	0.91	0.988
Dendritic cells	cDC2	0.19	0.83	0.937
	Marker models (22 markers)		**0.96**	**0.975**
CD4+ T-cells	CCR4		0.93	0.96
CD4+ T-cells	CD4		0.99	0.998
CD4+ T-cells	CD45RA		0.95	0.986
CD4+ T-cells	TIGIT		0.94	0.97
NK-cells	CD56		0.95	0.99

Table 1. *Representative model prediction quality scores for small, mid-sized, and large populations of different immunological cell groups and representative marker model quality scores.*

Figure 1

https://doi.org/10.1182/blood-2023-180146