Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

Gating Mass Cytometry Data by Deep Learning-Supplementary Materials

Huamin Li [1]*, Uri Shaham [2]*, Kelly P. Stanton [3], Yi Yao [4], Ruth Montgomery [4], Yuval Kluger [1,3,5]†

[1] Applied Mathematics Program, Yale University, 51 Prospect St., New Haven, CT 06511, USA
[2] Department of Statistics, Yale University, 24 Hillhouse Ave., New Haven, CT 06511, USA
[3] Department of Pathology and Yale Cancer Center, Yale University School of Medicine, New Haven, CT, USA
[4] Department of Internal Medicine, Yale School of Medicine, 333 Cedar St., New Haven, CT 06520, USA
[5] Interdepartmental Program in Computational Biology and Bioinformatics,Yale University, New Haven, CT, USA

† To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Mass cytometry or CyTOF is an emerging technology for high-dimensional multiparameter single cell analysis that overcomes many limitations of fluorescence-based flow cytometry. New methods for analyzing CyTOF data attempt to improve automation, scalability, performance, and interpretation of data generated in large studies. Assigning individual cells into discrete groups of cell types (gating) involves time-consuming sequential manual steps, untenable for larger studies.
Results: We introduce DeepCyTOF, a standardization approach for gating, based on deep learning techniques. DeepCyTOF requires labeled cells from only a single sample. It is based on domain adaptation principles and is a generalization of previous work that allows us to calibrate between a target distribution and a source distribution in an unsupervised manner. We show that DeepCyTOF is highly concordant (98%) with cell classification obtained by individual manual gating of each sample when applied to a collection of 16 biological replicates of primary immune blood cells, even when measured across several instruments. Further, DeepCyTOF achieves very high accuracy on the semi-automated gating challenge of the FlowCAP-I competition as well as two CyTOF datasets generated from primary immune blood cells: (i) 14 subjects with a history of infection with West Nile virus (WNV), (ii) 34 healthy subjects of different ages. We conclude that deep learning in general, and DeepCyTOF specifically, offers a powerful computational approach for semi-automated gating of CyTOF and flow cytometry data.
Availability: our codes and data are publicly available at `https://github.com/KlugerLab/deepcytof.git`.
Contact: yuval.kluger@yale.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Data Pre-processing

Given a blood samples A, we first perform an elementary logarithmic transformation

$$A_{i,j} \leftarrow \log(1 + A_{i,j}) \qquad (1)$$

In addition, we standardize each column of A to have zero mean and and unit variance.

*The first two authors contributed equally to this work.

2 Supplementary Materials

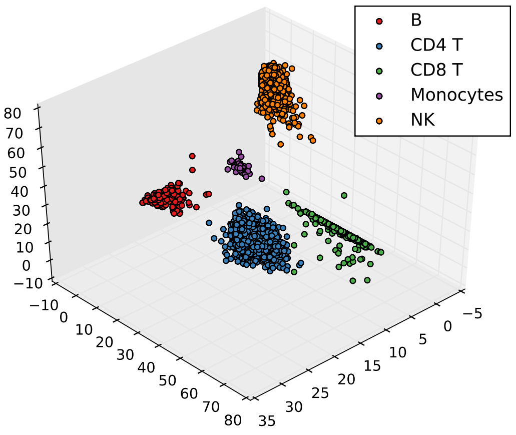

Fig. S1: Third hidden layer representation of a blood sample (the unlabeled cells are omitted), obtained from DeepCyTOF without a calibration step. Each color corresponds to a cell type. Different cell types are concentrated in different regions of the code space.

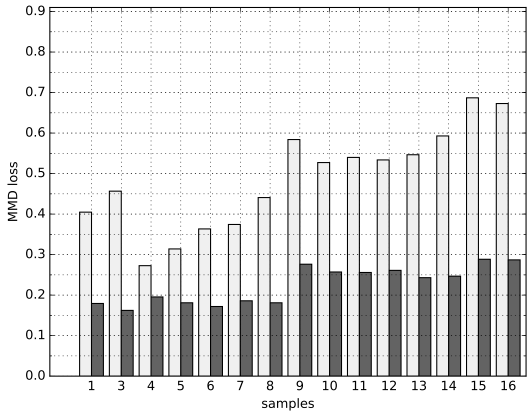

Fig. S2: MMD between the source (sample 2) and each of the of other 15 samples in the multi-center dataset before (white) and after (black) calibration. The MMD values were computed based on random batches of size 1000 from each sample.

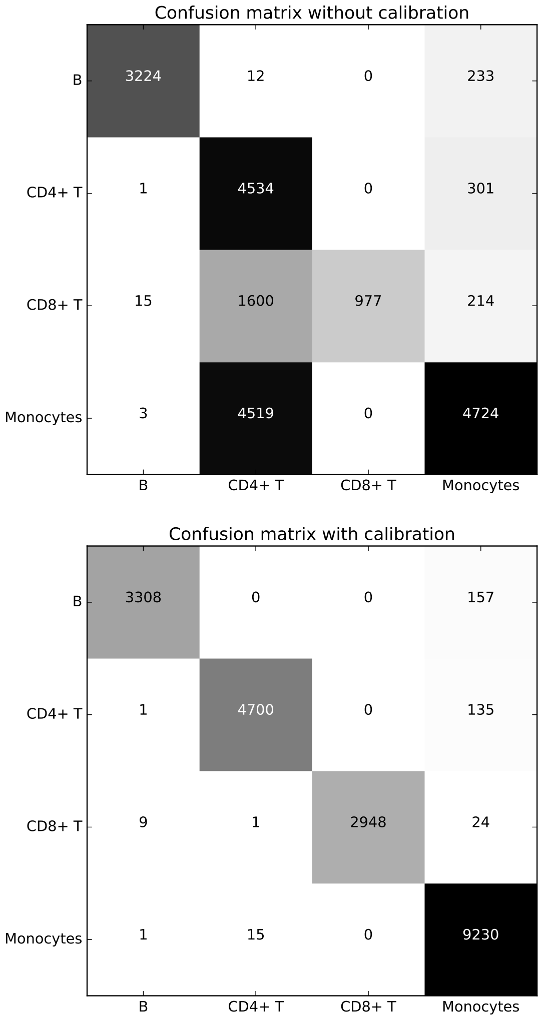

Fig. S3: Confusion matrix for visualizing the performance of DeepCyTOF when applied to sample 15 without (top) and with (bottom) a calibration step (the unlabeled cells are omitted). The rows represent the actual cell type label and the columns represent the predicted cell type label. The F-measure associated with the top panel (0.8857) is significantly lower than the F-measure associated with the bottom panel (0.9609).

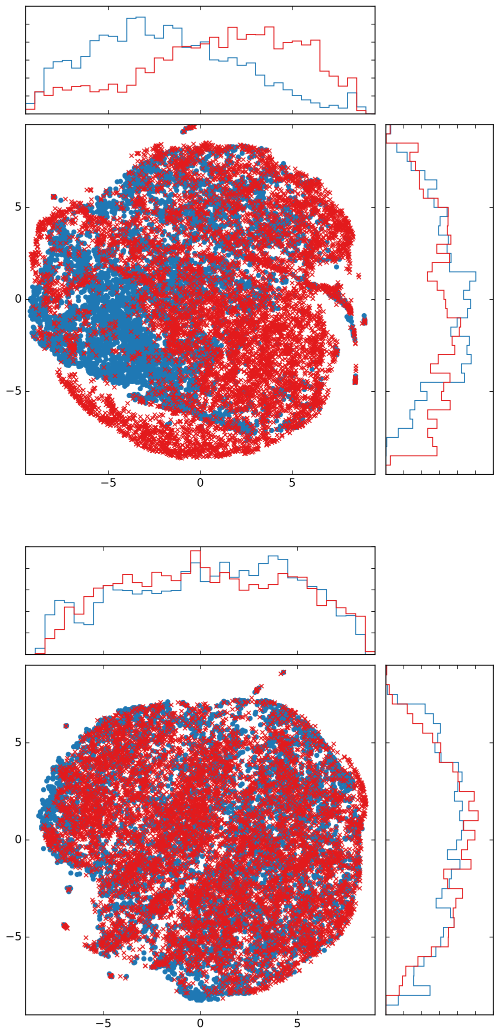

Fig. S4: t-SNE plot of CD8+ T cell from sample 15 before (top) and after (bottom) with DeepCyTOF.

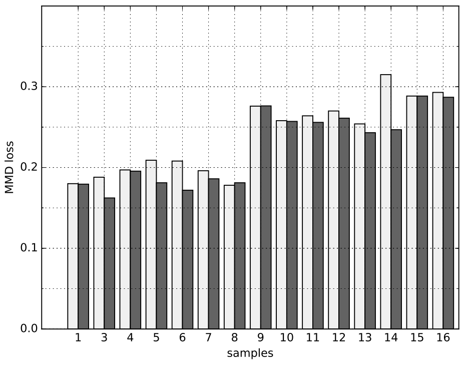

Fig. S5: MMD between the reference sample and each of the remaining 15 samples in the multi-institute cytometry dataset. White: without denoising. Black: with denoising.