On clustering for cell phenotyping in multiplex immunohistochemistry (mIHC) and multiplexed ion beam imaging (MIBI) Data

Souvik Seal (souvik.seal@cuanschutz.edu)
University of Colorado https://orcid.org/0000-0003-3268-610X
Julia Wrobel
University of Colorado
Amber M. Johnson
University of Colorado
Raphael A. Nemenoff
University of Colorado
Erin L. Schenk
University of Colorado
Benjamin G. Bitler
University of Colorado
Kimberly R. Jordan
University of Colorado
Debashis Ghosh
University of Colorado

Research Article

Keywords: Multiplex Tissue Imaging, Cell Phenotyping, Vectra Polaris, MIBI, Semi-supervised Learning

Posted Date: April 19th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-609920/v2

RESEARCH NOTE

On clustering for cell phenotyping in multiplex immunohistochemistry (mIHC) and multiplexed ion beam imaging (MIBI) Data

Souvik Seal[1*], Julia Wrobel[1], Amber M. Johnson[2], Raphael A. Nemenoff[2], Erin L. Schenk[3], Benjamin G. Bitler[4], Kimberly R. Jordan[5] and Debashis Ghosh[1]

Abstract

Objective: Multiplex immunohistochemistry (mIHC) and multiplexed ion beam imaging (MIBI) images are usually phenotyped using a manual thresholding process. The thresholding is prone to biases, especially when examining multiple images with high cellularity.

Results: Unsupervised cell phenotyping methods including PhenoGraph, flowMeans, and SamSPECTRAL, primarily used in flow cytometry data, often perform poorly or need elaborate tuning to perform well in the context of mIHC and MIBI data. We show that, instead, semi-supervised cell clustering using Random Forests, linear and quadratic discriminant analysis are superior. We test the performance of the methods on two mIHC datasets from the University of Colorado School of Medicine and a publicly available MIBI dataset. Each dataset contains numerous highly complex images.

Keywords: Multiplex Tissue Imaging; Cell Phenotyping; Vectra Polaris; MIBI; Semi-supervised Learning

Introduction

Several technologies have recently been developed for probing single-cell spatial biology, including multiparameter immunofluorescence, imaging mass cytometry (IMC), multiplex immunohistochemistry (mIHC) and multiplexed ion beam imaging (MIBI). Collectively these technologies are often referred to as multiplex tissue imaging. The spatial capabilities of these new technologies offer up the potential for researchers to develop a novel understanding of the biological mechanisms underlying cellular and protein interactions in a wide array of scientific contexts. Multiplex tissue imaging platforms are rapidly developing and all produce data of a similar structure: two dimensional images of tissue at the resolution of cells and nuclei, where proteins in the sample have been labeled with antibodies called "markers" that attach to cell membranes and proteins in the tissue sample to allow for identification of distinct cell types and functions. 'Multiplex' refers to the fact that multiple markers can be tagged for each image, with typically 6-8 markers for mIHC images and up to 40 markers for MIBI images.

*Correspondence: souvik.seal@cuanschutz.edu
[1]Department of Biostatistics and Informatics, University of Colorado CU Anschutz Medical Campus, Aurora, Colorado, USA
Full list of author information is available at the end of the article

Multiplex tissue imaging technologies have many data pre-processing and analyses steps that have not yet been uniformly implemented. Cell phenotyping, defined as identification of cell populations based on marker expression, is a challenging process in this context. Current cell phenotyping approaches for multiplex imaging data require researchers to manually set a threshold intensity value for each marker. Cells with intensity values greater or less than the threshold are labeled positive or negative for that marker, respectively. Other methods utilize algorithms trained on examples of positive and negative cells that are manually identified in an image while visualizing each marker individually. Cells are then phenotyped based on the expression of each marker. For example, CD4 T cells are positive for markers CD3 and CD4 and negative for CD8. This manual phenotyping approach is cumbersome for high parameter panels and depends on the reliability and expert knowledge of the user selecting positive cells or choosing thresholds, which may differ between users. Thus, manual gating is not only prone to human error but also time consuming and costly. Algorithms have already been developed to tackle these same phenotyping issues for multiplex technologies that analyze single cells in a liquid suspension without spatial resolution, namely flow and mass

cytometry [1]. In particular, automatic gating methods using machine learning algorithms have become more and more popular in flow and mass cytometry data as the number of analyzed parameters has increased [2] .

Our aim in this paper is to compare automated unsupervised and semi-supervised clustering algorithms for phenotyping multiplex tissue imaging data. We will adapt approaches originally developed for two non-spatial technologies, flow and mass cytometry, and test our algorithms on two mIHC datasets [3, 4] obtained from the University of Colorado School of Medicine and one publicly available MIBI dataset [5]. The paper is structured as follows. First, we describe the automated cell phenotyping approaches based on unsupervised and semi-supervised clustering. Next, we describe our datasets and adaptation of existing phenotyping approaches to multiplex tissue imaging data. Then we present the results and conclude with a discussion. Our tools are publicly available on GitHub as the R package VectraMIBI.

Main text

Existing phenotyping algorithms

Unsupervised learning algorithms

Unsupervised cell phenotyping algorithms partition cells into different classes based on their multiplex marker expression without using any prior knowledge [6]. As such, these methods are initially unbiased and usually time and memory efficient as well. In addition, novel cell types and populations can be discovered by not biasing clustering algorithms with prior information about marker expression. However, these methods suffer from several major limitations. For example, once the cells have been classified by an unsupervised algorithm, researchers manually gate the obtained classes to compare meaningful cell types (e.g. CD4 T cell, CD68+ macrophages etc.). This step can be cumbersome and again prone to human error. PhenoGraph [7], flowMeans [8] and SamSPECTRAL [9] are some of the most popular unsupervised cell phenotyping algorithms [1, 2].

Semi-supervised learning algorithms

Semi-supervised cell phenotyping approaches typically involve building a predictive model using multiplex marker expression from a subset of cells in a dataset, called the training set, that have been manually phenotyped [10]. The built models are then used to phenotype the remaining cells, or the test set. Unlike unsupervised methods, the cells in this case are directly assigned to existing phenotypes which obviates the problem of matching arbitrary clusters to meaningful cell types. One can argue that the first step of manually phenotyping cells in the training set is subjected to human error. However, the size of the training set is usually just a fraction of the full dataset. Therefore, ensuring the purity of manual phenotyping of the training dataset should be easy relative to manually phenotyping all of the data; though this remains a practical limitation for all current approaches.

DeepCyTOF [11], CyTOF linear classifier [12] and ACDC [13] are popular semi-supervised methods in flow and mass cytometry [2]. CyTOF linear classifier, which is based on linear discriminant analysis (LDA), has been shown to outperform more complex algorithms like DeepCyTOF, ACDC on several CyTOF datasets [2, 12]. All the above methods are briefly described further in Table S1 (in the supplementary).

LDA assumes that the data has equal variance across groups and is normally distributed. Though these assumptions may hold for CyTOF data, in mIHC datasets both assumptions are violated. To address these problems, we consider more general machine-learning algorithms such as quadratic discriminant analysis (QDA) [14] and Random Forest [15]. QDA is similar to LDA but does not require equal variance across groups. The decision tree-based Random Forest method is robust for non-normal data and has several additional advantages demonstrated by [16]; these include minimal tuning parameters, excellent off-the-shelf prediction, honest estimates of classification through out-of-bag samples, and stable prediction behavior. Therefore, in the context of mIHC and MIBI data, we propose to use Random Forest and compare its performance with LDA and QDA.

Materials and Methods

Datasets

Our analysis incorporates three multiplex tissue imaging datasets: an ovarian cancer dataset [3] acquired on the mIHC Vectra Polaris platform (Akoya Biosciences), a lung cancer dataset [4] acquired on the mIHC Vectra 3.0 system (Akoya Biosciences), and a breast cancer dataset [5] collected on the MIBI platform (IonPath, Inc). The two mIHC datasets were segmented and phenotyped using Inform (v2.4.8, Akoya Biosciences), commercially available software for Vectra data [17], and the MIBI dataset was phenotyped in MATLAB using deep learning-based methods [5]. For each cell, the expression data is available for multiple markers. The datasets are described in detail below and Table 1 lists the overall distribution of the cell types in different datasets.

mIHC ovarian cancer dataset

There are 302,147 cells from 132 subjects. There are five different cell types: CD19+, CD3+/CD8-, CD3+/CD8+, CD68+, CK+/Ki67+. There are six

markers, CD19, CD3, CK, CD8, Ki67, CD68 observed in each of the cells. More details on this data can be found at [3].

mIHC lung cancer dataset

There are 1,590,327 cells from 153 subjects each with 3-5 images (in total, 761 images). There are six different cell types: CD14+, CD19+, CD4+, CD8+, CK+, Other+ (meaning they do not belong to any of the indicated phenotypes). There are five markers, CD19, CD3, CK, CD8, CD14. More details on this data can be found at [4].

MIBI breast cancer dataset

The triple-negative breast cancer (TNBC) MIBI dataset [5] has 201,656 cells from 43 subjects and one image per subject. It has six different cell groups: Immune, Endothellial, Mesenchymal-like, Tumor, Keratin-positive tumor and Unidentified. There are 44 markers available, such as CD3, CD8, CD63, Ki67, and Vimentin.

Table 1 The frequency of cells belonging to different cell types in different datasets.

Dataset	Cell Type	Total Cells
mIHC ovarian cancer	CD19+	15267 (5%)
	CD3+/CD8-	15952 (5.3%)
	CD3+/CD8+	41008 (13.6%)
	CD68+	57632 (19.1%)
	CK+/Ki67+	172288 (57%)
mIHC lung cancer	CD14+	175878 (11.1 %)
	CD19+	154045 (9.7 %)
	CD4+	232878 (14.6 %)
	CD8+	124102 (7.8 %)
	CK+	594140 (37.4 %)
	Other+	309284 (19.4 %)
MIBI breast cancer	Unidentified	1839 (1 %)
	Immune	83336 (41.3 %)
	Endothelial	2089 (1 %)
	Mesenchymal-like	8479 (4.2 %)
	Tumor	3177 (1.6 %)
	Keratin-positive tumor	102736 (50.9 %)

Clustering algorithms used

We use Random Forest, Linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) on the datasets described above. We use random subsets (of different sizes) of the entire dataset to train predictive models. The prediction performance of the built models are evaluated on the held-out dataset.

Results

We primarily focus on semi-supervised rather than unsupervised approaches since the latter have several limitations including post-clustering manual labeling of the cell types. First we briefly highlight some of the problems of the unsupervised methods. Then, we shift to semi-supervised methods and compare the usability of Random Forest with LDA and QDA in all three datasets.

Unsupervised Methods

In the mIHC lung cancer dataset, we clustered the cells of one subject at a time based on their multiplex marker expression using the unsupervised methods, PhenoGraph, SamSPECTRAL and flowMeans. T-distributed stochastic neighbor embedding (t-SNE) [18] has been used by researchers to visualize high-dimensional data in various contexts including flow and mass cytometry [19, 20]. In Figure 1, for a particular subject, we compared the true cell labels with the labels estimated using the unsupervised methods, overlaid on the first two t-SNEs of the marker data.

For most of the subjects, including the one depicted in Figure 1, PhenoGraph classified the cells into a large number of clusters (around 18) compared to the true number of cell types which is 6. flowMeans generated a more accurate number of clusters (around 7), but the clustering was mostly inaccurate. SamSPECTRAL was highly variable depending on the input values of the tuning parameters and none of the combinations yielded cell labels comparable to the true ones.

Semi-supervised Methods

For each dataset, we randomly selected m images (out of the total size, M) to train the models on and evaluated their performance on the remainder of the images. We repeated this for different sizes of m, then 5 times each at each size of m with different randomly selected images. Results were aggregated across repetitions and summarized by prediction accuracy, adjusted rand index (ARI), and normalized mutual information (NMI).

mIHC ovarian cancer dataset

For the mIHC ovarian cancer dataset, we considered multiple training set-sizes which are just fractions of the full-data size M, $m = 7$ (5%), 13 (10%), 20 (15%), and 26 (20%). For each set-size m, we considered 5 repetitions. Table 2 lists the mean (and standard deviation) of prediction accuracy, ARI, and NMI. Even for the smallest training set-size, all three methods performed well, with Random Forest having the highest mean prediction accuracy, ARI, and NMI. Random Forest also had significantly low standard deviation which accentuated its high robustness. As the training size increased, prediction accuracy, ARI, and NMI marginally improved for all three methods.

mIHC lung cancer dataset

For the mIHC lung cancer dataset, we considered the training set-sizes, $m = 4$ (0.5%), 8 (1%), 15 (2%), 23 (3%) and 76 (10%). For each case, we considered 5 repetitions. Random Forest again outperformed LDA and QDA (Table 2). However, the prediction accuracy was significantly lower for the smaller training set-sizes.

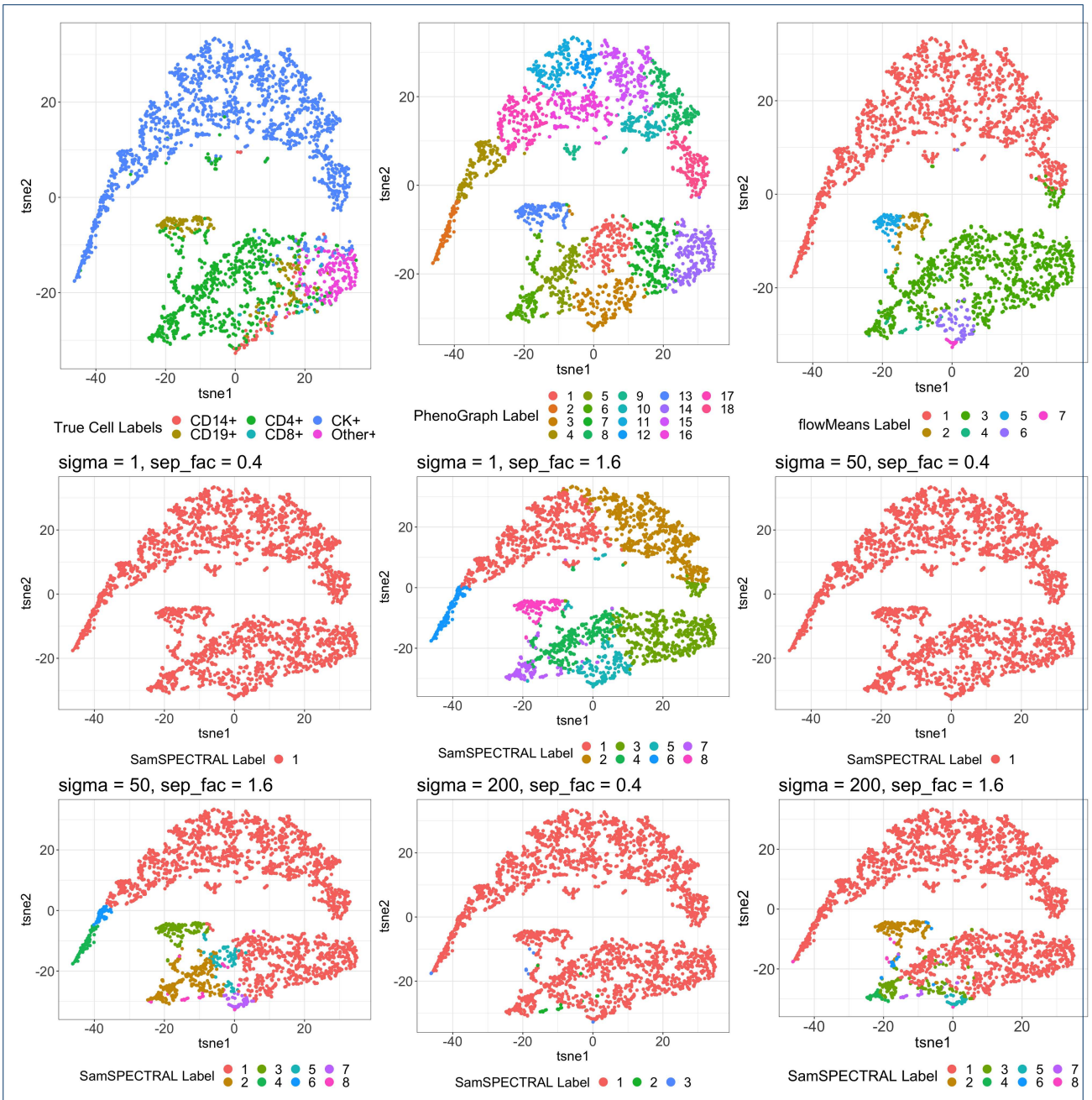

Figure 1 Comparison of cell labels estimated by PhenoGraph, flowMeans and SamSPECTRAL with the true cell labels for a particular subject. The top row (from left to right) shows the scatter-plot of TSNE1 and TSNE2 for different cells colored by three different labels, true labels, estimated labels using PhenoGraph and flowMeans. The second and third rows show the scatter-plot of TSNE1 and TSNE2 for different cells colored by the estimated labels using SamSPECTRAL for different values (from low to high) of sigma and separation factor (`sep_fac`).

Random Forest's performance steadily improved as the training set size increased, whereas for LDA and QDA, the performance stayed nearly the same across training set sizes. We noticed a dip in the overall performance of all the methods in this dataset compared to the ovarian cancer dataset which we further investigated and the details are provided in the supplementary.

MIBI breast cancer dataset

For the mIHC lung cancer dataset, we considered the training set-sizes, $m = 2$ (5%), 4 (10%) and 8 (20%). Even with the smallest training set-size, Random Forest achieved great prediction accuracy (Table 2). For LDA, the accuracy increased steadily as the training set size increased but, Random Forest outperformed LDA in all cases. We did not report the performance of QDA on this dataset since it often encountered error due to "rank deficiency" for small training sizes. The error originated if the frequency of at least one of the six cell types was less than the number of observed markers (44) in the training dataset making the sample covariance matrix rank-deficient.

Table S3 (in the supplementary) lists the run-times in minutes of the different methods in different datasets.

Limitations

We have noticed that cells of certain types can get incorrectly phenotyped if the corresponding markers are not informative enough. For example, in some subjects from the lung cancer dataset, CD19 marker intensity is not distinctive across different cell types which makes identifying CD19+ cells hard and results in poor prediction performance (refer to the supplementary). Therefore, we advise thorough investigation of the training images via heat-maps and ridge-plots of the intensity of the markers across different cell types. If will help to understand if any of the cell types may encounter poor prediction performance. It shall also be kept in mind that the mIHC datasets we analyzed were originally phenotyped using the InForm software. It is a possibility that the original phenotyping was inaccurate and our "ground truth" itself is biased.

Abbreviations

mIHC = Multiplex Immuno Histochemistry, MIBI = Multiplex Ion Beam Imaging, LDA = Linear Discriminant Analysis, QDA = Quadratic Discriminant Analysis.

Supplementary Information

Additional file 1 - Supplementary material. Here, we provide a section explaining the overall dip in the performance of the methods in the mIHC lung cancer dataset. Figure S1, S2, and S3 focus on the mIHC lung cancer dataset, and respectively show the scatter-plot of accuracy of the Random Forest approach for predicting different cell types, the bar-plot of proportion of cell types getting correctly predicted, and the ridge-plot of CD19 marker intensity in the cells of different images. Table S1 and S2 respectively list the summary of a few existing methods and the run-times of the methods in different datasets.

Declarations

Ethics approval and consent to participate

Not Applicable.

Consent for publication

Not Applicable.

Availability of data and materials

The MIBI breast cancer dataset used in the paper can be found at this link, https://www.angelolab.com/mibi-data. The mIHC datasets are available from the corresponding author on reasonable request. Our methods can be found as a *R* package named as VectraMIBI at this link. The package builds a Random Forest model on a given training dataset, and uses predictions from that model to annotate (phenotype) the cells of a test dataset. The package also provides visualization tools including heat-maps of the mean marker intensity over different cell types and image specific ridge-plots of the marker intensity for different cell types for basic exploration of the training dataset.

Competing interests

The authors declare that they have no competing interests.

Funding

B.G. is supported by the Department of Defense Award (OC170228) and an American Cancer Society Research Scholar Award (134106-RSG-19-129-01-DDC). E.L.S. is supported by NIH grant K12 CA086913 and ACS IRG #16-184-56 from the American Cancer Society to the University of Colorado Cancer Center, and a grant from the Cancer League of Colorado. S.S. is funded by the Grohne-Stepp Endowment from the University of Colorado Cancer Center. J.W. is supported by the NIH/NCATS Colorado CTSA (UL1TR002535).

Authors' contributions

S.S., J.W. and D.G. were involved with the conceptualization of the project, methodological development, analysis and writing of the first draft of the manuscript. All authors (S.S., J.W., A.M.J., R.A.N., E.L.S., B.G.B., K.R.J., D.G.) participated in the writing process and approved the final manuscript.

Acknowledgments

We thank the Human Immune Monitoring Shared Resource and support of the University of Colorado Human Immunology and Immunotherapy Initiative for their expert assistance in multiplex IHC and generation of the ovarian and lung datasets. We acknowledge the support of the University of Colorado Cancer Center Support Grant (P30CA046934).

Author details

[1]Department of Biostatistics and Informatics, University of Colorado CU Anschutz Medical Campus, Aurora, Colorado, USA. [2]Department of Medicine, School of Medicine, University of Colorado CU Anschutz Medical Campus, Aurora, Colorado, USA. [3]Division of Medical Oncology, School of Medicine, University of Colorado CU Anschutz Medical Campus, Aurora, Colorado, USA. [4]Department of Obstetrics and Gynecology, School of Medicine, University of Colorado CU Anschutz Medical Campus, Aurora, Colorado, USA. [5]Department of Immunology and Microbiology, School of Medicine, University of Colorado CU Anschutz Medical Campus, Aurora, Colorado, USA.

References

1. Peng Liu, Silvia Liu, Yusi Fang, Xiangning Xue, Jian Zou, George Tseng, and Liza Konnikova. Recent advances in computer-assisted algorithms for cell subtype identification of cytometry data. *Frontiers in cell and developmental biology*, 8:234, 2020.

Table 2 Prediction accuracy, ARI and NMI mean (± standard deviation) for different training set sizes in mIHC ovarian and lung cancer datasets and MIBI breast cancer dataset.

Dataset	Training size	Method	Accuracy	ARI	NMI
mIHC ovarian cancer	5%	Random Forest	0.944 ± 0.004	0.888 ± 0.007	0.783 ± 0.010
		LDA	0.899 ± 0.017	0.779 ± 0.047	0.642 ± 0.051
		QDA	0.909 ± 0.007	0.821 ± 0.023	0.699 ± 0.018
	10%	Random Forest	0.949 ± 0.002	0.896 ± 0.004	0.795 ± 0.006
		LDA	0.889 ± 0.010	0.748 ± 0.027	0.609 ± 0.028
		QDA	0.919 ± 0.003	0.842 ± 0.007	0.720 ± 0.008
	15%	Random Forest	0.951 ± 0.002	0.899 ± 0.003	0.802 ± 0.006
		LDA	0.898 ± 0.006	0.772 ± 0.018	0.633 ± 0.020
		QDA	0.920 ± 0.001	0.848 ± 0.005	0.724 ± 0.006
	20%	Random Forest	0.952 ± 0.002	0.902 ± 0.002	0.806 ± 0.006
		LDA	0.899 ± 0.007	0.774 ± 0.018	0.634 ± 0.023
		QDA	0.922 ± 0.001	0.853 ± 0.003	0.727 ± 0.006
mIHC lung cancer	0.5%	Random Forest	0.734 ± 0.179	0.575 ± 0.022	0.426 ± 0.018
		LDA	0.668 ± 0.052	0.413 ± 0.102	0.363 ± 0.070
		QDA	0.669 ± 0.048	0.459 ± 0.076	0.365 ± 0.036
	1%	Random Forest	0.755 ± 0.057	0.594 ± 0.021	0.450 ± 0.013
		LDA	0.704 ± 0.057	0.486 ± 0.116	0.395 ± 0.068
		QDA	0.692 ± 0.040	0.482 ± 0.067	0.387 ± 0.031
	2%	Random Forest	0.768 ± 0.009	0.608 ± 0.016	0.468 ± 0.011
		LDA	0.686 ± 0.063	0.440 ± 0.133	0.374 ± 0.083
		QDA	0.696 ± 0.019	0.472 ± 0.030	0.387 ± 0.010
	3%	Random Forest	0.777 ± 0.002	0.620 ± 0.008	0.480 ± 0.005
		LDA	0.674 ± 0.064	0.424 ± 0.134	0.355 ± 0.084
		QDA	0.687 ± 0.024	0.452 ± 0.044	0.373 ± 0.024
	10%	Random Forest	0.805 ± 0.001	0.665 ± 0.003	0.524 ± 0.003
		LDA	0.709 ± 0.008	0.500 ± 0.024	0.393 ± 0.011
		QDA	0.705 ± 0.011	0.475 ± 0.027	0.386 ± 0.011
MIBI breast cancer	5%	Random Forest	0.951 ± 0.016	0.869 ± 0.037	0.772 ± 0.055
		LDA	0.781 ± 0.135	0.618 ± 0.111	0.47 ± 0.065
	10%	Random Forest	0.971 ± 0.010	0.915 ± 0.027	0.853 ± 0.04
		LDA	0.836 ± 0.038	0.632 ± 0.045	0.492 ± 0.045
	20%	Random Forest	0.983 ± 0.002	0.948 ± 0.008	0.903 ± 0.011
		LDA	0.877 ± 0.010	0.714 ± 0.020	0.569 ± 0.018

2. Xiao Liu, Weichen Song, Brandon Y Wong, Ting Zhang, Shunying Yu, Guan Ning Lin, and Xianting Ding. A comparison framework and guideline of clustering methods for mass cytometry data. *Genome biology*, 20(1):1–18, 2019.
3. Kimberly R Jordan, Matthew J Sikora, Jill E Slansky, Angela Minic, Jennifer K Richer, Marisa R Moroney, Junxiao Hu, Rebecca J Wolsky, Zachary L Watson, Tomomi M Yamamoto, et al. The capacity of the ovarian cancer tumor microenvironment to integrate inflammation signaling conveys a shorter disease-free interval. *Clinical Cancer Research*, 26(23):6362–6373, 2020.
4. Amber M Johnson, Bonnie L Bullock, Alexander J Neuwelt, Joanna M Poczobutt, Rachael E Kaspar, Howard Y Li, Jeff W Kwak, Katharina Hopp, Mary CM Weiser-Evans, Lynn E Heasley, et al. Cancer cell–intrinsic expression of mhc class ii regulates the immune microenvironment and response to anti–pd-1 therapy in lung adenocarcinoma. *The Journal of Immunology*, 204(8):2295–2307, 2020.
5. Leeat Keren, Marc Bosse, Diana Marquez, Roshan Angoshtari, Samir Jain, Sushama Varma, Soo-Ryum Yang, Allison Kurian, David Van Valen, Robert West, et al. A structured tumor-immune microenvironment in triple negative breast cancer revealed by multiplexed ion beam imaging. *Cell*, 174(6):1373–1387, 2018.
6. Jinmiao Chen and Feng Lin. Unsupervised clustering algorithms for flowmass cytometry data. *Computational methods with applications in bioinformatics analysis. Singapore: World Scientific Publishing Company*, page 194, 2017.
7. Jacob H Levine, Erin F Simonds, Sean C Bendall, Kara L Davis, D Amir El-ad, Michelle D Tadmor, Oren Litvin, Harris G Fienberg, Astraea Jager, Eli R Zunder, et al. Data-driven phenotypic dissection of aml reveals progenitor-like cells that correlate with prognosis. *Cell*, 162(1):184–197, 2015.
8. Nima Aghaeepour, Greg Finak, Holger Hoos, Tim R Mosmann, Ryan Brinkman, Raphael Gottardo, and Richard H Scheuermann. Critical assessment of automated flow cytometry data analysis techniques. *Nature methods*, 10(3):228–238, 2013.
9. Habil Zare, Parisa Shooshtari, Arvind Gupta, and Ryan R Brinkman. Data reduction for spectral clustering to analyze high throughput flow cytometry data. *BMC bioinformatics*, 11(1):403, 2010.
10. Emily Sassano. Machine learning methods for flow cytometry analysis and visualization. 2018.
11. Huamin Li, Uri Shaham, Yi Yao, Ruth Montgomery, and Yuval Kluger. Deepcytof: Automated cell classification of mass cytometry data by deep learning and domain adaptation. *bioRxiv*, page 054411, 2016.
12. Tamim Abdelaal, Vincent van Unen, Thomas Höllt, Frits Koning, Marcel JT Reinders, and Ahmed Mahfouz. Predicting cell populations in single cell mass cytometry data. *Cytometry Part A*, 95(7):769–781, 2019.
13. Markus Lux, Jan Krüger, Christian Rinke, Irena Maus, Andreas Schlüter, Tanja Woyke, Alexander Sczyrba, and Barbara Hammer. acdc–automated contamination detection and confidence estimation for single-cell genome data. *BMC bioinformatics*, 17(1):1–11, 2016.
14. Geoffrey J McLachlan. *Discriminant analysis and statistical pattern recognition*, volume 544. John Wiley & Sons, 2004.
15. L Breiman, J H Freidman, R A Olshen, and C J Stone. *Classification and Regression Trees*. Belmont, CA: Wadsworth, 1984.
16. L Breiman. Random forests. *Machine Learning*, 24:123–140, 2001.
17. Anne S Kramer, Bruce Latham, Luke A Diepeveen, Lingjun Mou, Geoffrey J Laurent, Caryn Elsegood, Laura Ochoa-Callejero, and George C Yeoh. Inform software: a semi-automated research tool to identify presumptive human hepatic progenitor cells, and other histological features of pathological significance. *Scientific reports*, 8 (1):1–10, 2018.
18. Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.
19. Vincent van Unen, Thomas Höllt, Nicola Pezzotti, Na Li, Marcel JT Reinders, Elmar Eisemann, Frits Koning, Anna Vilanova, and Boudewijn PF Lelieveldt. Visual analysis of mass cytometry data by hierarchical stochastic neighbour embedding reveals rare cell types. *Nature communications*, 8(1):1–10, 2017.
20. Abigail K Kimball, Lauren M Oko, Bonnie L Bullock, Raphael A Nemenoff, Linda F van Dyk, and Eric T Clambey. A beginner's guide to analyzing and visualizing mass cytometry data. *The Journal of Immunology*, 200(1):3–22, 2018.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- suppfile.pdf