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 Anschutz Medical Campus
Julia Wrobel
University of Colorado Anschutz Medical Campus
Amber M. Johnson
University of Colorado Anschutz Medical Campus
Raphael A. Nemenoff
University of Colorado Anschutz Medical Campus
Erin L. Schenk
University of Colorado Anschutz Medical Campus
Benjamin G. Bitler
University of Colorado Anschutz Medical Campus
Kimberly R. Jordan
University of Colorado Anschutz Medical Campus
Debashis Ghosh
University of Colorado Anschutz Medical Campus

Research Article

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

Posted Date: June 24th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-609920/v1

RESEARCH

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

Problem: 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

Background

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: 2-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.

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

*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

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.

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.

An important step is to perform the initial clustering and mapping simultaneously on a dataset pooled across all subjects. In particular, we cannot do the mapping on a per-subject basis. This is because labels are assigned in an arbitrary fashion during unsupervised clustering. For M clusters, there would be $M!$ possible ways of attaching the labels. Thus, there is a non-uniqueness in how cluster labels are assigned, which in turn leads to an alignment problem that increases exponentially in the number of clusters.

PhenoGraph [7], flowMeans [8] and SamSPECTRAL [9] are some of the most popular unsupervised cell phenotyping algorithms [1]. A brief description of the methods can be found in Table 2. On multiple mass cytometry by time-of-flight (CyTOF) datasets, [2] concluded that flowMeans outperforms PhenoGraph overall. In the Results section, we demonstrate some of the shortcomings of these methods on one of our mIHC datasets.

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], described further in Table 2. CyTOF linear classifier, which is based on linear discriminant analysis (LDA), has been shown to outperform more complex semi-supervised clustering algorithms like DeepCyTOF, ACDC on several CyTOF datasets [12]. LDA has also displayed superior performance over ACDC on six CyTOF datasets [2]. Proper performance of the LDA algorithm 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 also consider the machine-learning algorithms quadratic discriminant analysis (QDA) and Random Forests (RF). QDA is similar to LDA but does not require equal variance across groups, thereby eases the first assumption. The decision tree-based Random Forest method [14] is robust for non-normal data and has several additional advantages demonstrated by [14]; 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 Forests 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). Each dataset consists of multiple patients and corresponding images (one image per subject) that contain different numbers of cells. The two mIHC datasets were segmented and phenotyped using Inform (v2.4.8, Akoya Biosciences), commercially available software for Vectra data [15], 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.

mIHC ovarian cancer dataset
There are 302,147 cells from 132 patients. There are five different cell types: CD19+, CD3+/CD8-, CD3+/CD8+, CD68+, CK+/Ki67+. There are 6 mean marker measures: CD19, CD3, CK, CD8, Ki67, CD68 for 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 patients 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 5 mean marker measures CD19, CD3, CK, CD8, CD14 for each of the cells. 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 patients and one image per patient. It has six different cell groups: Immune, Endothellial, Mesenchymal-like, Tumor, Keratin-positive tumor and Unidentified. There are 34 mean marker measures available, such as CD3, CD8, CD63, Ki67, and Vimentin.

Table 1 lists the cell distribution of different datasets.

Clustering Algorithms

We now apply the Random Forest, LDA and QDA algorithms to the datasets described above. We use different random subsets (of different sizes) of the entire dataset to train predictive models. The prediction performance of the built models is evaluated on the held-out dataset. For each method, we compare predicted cell types with the true cell types and evaluate performance using prediction accuracy, adjusted rand index (ARI) [16] and normalized mutual information (NMI) [17].

Suppose there are C different cell populations (c_i) in the training dataset, each of which has cell frequency of n_i. Let the total number of cells in training dataset be $n = \sum_{i=1}^{C} n_i$. We briefly go through the algorithms: Random Forests, LDA and QDA below.

Random Forests

The Random Forest algorithm [18] builds multiple decision trees ($f_b, b = 1, \ldots, B$, where B is the number of trees in the forest) and merges them together to obtain an accurate and stable prediction. We chose $B = 500$ trees for our analyses. To predict the type of a new cell with marker vector x, majority vote is considered i.e the majority of $f_b(x), b = 1, \ldots, B$.

LDA and QDA

Linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) [19] are two classic classifiers, with a linear and a quadratic decision boundary, respectively. To predict the cell type of a new cell with marker vector x, both LDA and QDA assign it to cell population class c_i for which the posterior probability of x being part of c_i is maximum, across all cell populations,

$$\underset{c_i, i=1,\ldots,C}{argmax}\ p(x|c_i)p(c_i)$$

$$p(x|c_i) = \frac{1}{(2\pi)^{k/2}|\Sigma_i|^{1/2}} \exp(-(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i))$$

$$p(c_i) = \frac{n_i}{n}$$

Unlike QDA, LDA makes an additional homoscedasticity assumption that $\Sigma_i = \Sigma, \forall i = 1, \ldots, C$. The mean vector μ_i and the covariance matrix Σ_i are generally estimated using maximum likelihood approaches.

Code Availability

Open-source code and a tutorial for our semi-superivsed phenotyping approach are available as the VectraMIBI R package at
https://github.com/sealx017/VectraMIBI.
As input, the package requires a training dataset of segmented multiplex imaging data with fully annotated cells, and a test dataset that contains the marker data without phenotype annotations. The VectraMIBI pakage builds a Random Forest model on the training dataset, and uses predictions from that model to annotate (phenotype) the cells in the test dataset. The package also provides visualization tools including heatmaps of the average marker data over different cell types and image specific ridge plots of the marker data for different cell types that can be used to get an overview of the training dataset.

Results

We primarily focus on semi-supervised rather unsupervised approaches since the latter have several limitations including post-clustering manual labeling of cell types. First we highlight some of the issues with unsupervised methods by applying the unsupervised algorithms PhenoGraph, flowMeans, and SamSPECTRAL to the mIHC lung cancer dataset. Then we move to semi-supervised methods and compare the Random Forest algorithm with LDA and QDA on all three multiplex imaging datasets.

Unsupervised Methods

In the lung cancer dataset, we clustered the cells of one patient at a time based on their multiplex marker expression. We applied the unsupervised algorithms PhenoGraph, flowMeans, and SamSPECTRAL to the mIHC lung cancer dataset. These algorithms take one image at a time and clustered the cells in a given image based on their multiplex marker expression. T-distributed stochastic neighbor embedding (t-SNE) [20] has been used by researchers to visualize high-dimensional data in various contexts including flow and mass cytometry [21–24], and we compute the first two t-SNEs of the cell marker data to facilitate visualization of cluster labels from PhenoGraph, flowMeans, and SamSPECTRAL. Figure 1 demonstrates the scatter-plot of t-SNE1 and t-SNE2 of those cells with three different labels: on the left, the true cell types, in the middle, labels estimated using PhenoGraph and on the right, labels estimated using flowMeans.

For most of the subjects, PhenoGraph classified the cells into a large number of clusters (18) compared to the true number of cell types (6). flowMeans generated a more accurate number of clusters (7), but the clustering was inaccurate. SamSPECTRAL was highly variable depending on the input value for two algorithm tuning parameters. Figure 2 illustrates six different t-SNE scatter-plots with labels estimated using SamSPECTRAL with six different combinations of the tuning parameters. We saw that the tuning parameters massively affected the classification and none of the combinations were comparable to the true cell labels.

Semi-supervised Methods

The ability of the Random Forest, LDA, and QDA algorithms to accurately phenotype cells was assesed on our two mIHC datasets and the MIBI dataset. For each dataset, we randomly selected m images (out of the total, 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

The full mIHC ovarian cancer dataset contained 132 images. The training set sizes took the following values: $m = 7$ images (5% of the full data), $m = 13$ images (10% of the full data), $m = 20$ images (15% of the full data), and $m = 26$ images (20% of the full data). For each case, we considered 5 repetitions. Table 3 lists the mean (and standard deviation) of prediction accuracy, ARI, and NMI for the ovarian cancer mIHC data. Even for the smallest training set size (5%), all three methods performed well, with Random Forests having the highest mean prediction accuracy, ARI, and NMI. Random Forests also had significantly low standard deviation which accentuated its high robustness. As the training size increased, prediction accuracy, ARI, and NMI marginally improves across all three methods. Therefore, we conclude that it would be reasonable to use just 5% of the full data as the training set for this data.

mIHC lung cancer dataset

The full mIHC lung cancer dataset contained 761 images. The training set sizes m took the following values: 4 images (0.5% of the full data), 8 images (1% of the full data), 15 images (2% of the full data), 23 images (3% of the full data) and 76 images (10% of the full data). For each case, we considered 5 repetitions. Table 4 lists the mean (and standard deviation) of prediction accuracy, ARI, and NMI for the lung cancer mIHC data. The Random Forest algorithm again performed best across all the metrics. However, the prediction accuracy was significantly lower for the smaller training set sizes. Random Forest performance steadily increased as the training set size increased, whereas for LDA and QDA the performance stayed nearly the same across training set sizes.

To investigate the dip in overall performance in this dataset compared to the ovarian cancer dataset, we looked at the cell type specific prediction accuracy for a fixed training set. In Figure 3, we show a scatter-plot for each of the six different cell types depicting the cell frequency (on the x axis) and the corresponding prediction accuracy (on the y axis) in each of 761 images. For cell type CD19+, we observed that the prediction accuracy was quite low (median accuracy 0.298), especially in images where the CD19+ cell frequency was low. Figure 4 shows how many of the cells of a particular type are getting actually predicted to be of that type. For example, from Figure 4b, we see that a significant number of CD19+ cells are getting assigned

to types: CD4+, Other+. Since we would expect the CD19 marker staining intensity to be the instrumental variable in separating CD19+ cells from the other cell types, we inspected its distribution in the images for which CD19+ prediction accuracy was above 0.9 and the images for which CD19+ prediction accuracy was less than 0.25, illustrated in two ridge plots in Figure 5. Figure 5a shows the distribution of the CD19 marker staining intensity (color coded by original cell type) of the cells from four images for which CD19+ prediction accuracy is higher than 0.9 and Figure 5b shows the ridge plot of CD19 marker staining intensity (color coded by original cell type) of the cells from five images for which CD19+ prediction accuracy is lower than 0.25. In Figure 5a, we see that the cells that were originally phenotyped as CD19+ have distinctively high staining intensities for the CD19 marker compared to the cells of other types. However, in Figure 5b, the CD19 marker value is not distinctively different across the cell types in any of the low accuracy images. We believe that the low staining intensity of the CD19 marker across different cell types is the cause of the poor prediction of CD19+ cells for those particular images.

MIBI Data

The full MIBI breast cancer dataset contained 43 images. The training set sizes m took the following values: 2 images (5% of the full data), 4 images (10% of the full data), 8 images (20% of the full data). For each case, we considered 5 repetitions. Table 5 lists the mean (and standard deviation) of prediction accuracy, ARI, and NMI for the MIBI data. Even for the smallest training set size (5%), the Random Forest achieved great prediction accuracy. For LDA, the accuracy increased steadily as the training set size increased but, Random Forests outperformed LDA in all cases. We have not reported 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 training dataset contained at least one of the six cell groups with number of cells less than the number of observed markers (34), since the sample covariance matrix of that group would become rank deficient.

Table 6 lists the computation time in minutes across different methods and datasets. All the methods were run on a *Mac* system with 32 GB DDR4 RAM and 2.4 GHz 8-Core Intel Core i9 processor. We shall note that LDA and QDA both took fractions of the time taken by Random Forests. Therefore, LDA and QDA both can be used for a preliminary analysis or when there is a time constraint.

Discussion

In this paper, we have investigated usability of machine learning based cell phenotyping methods that are popular in flow and mass cytometry, for mIHC and MIBI data. Since unsupervised methods require a burdensome step of matching identified cell clusters to meaningful cell phenotypes, we primarily focused on semi-supervised methods which directly classify the cells into defined cell phenotypes.

To use semi-supervised approaches, the cells of a relatively small training dataset need to be manually phenotyped. Next, predictive models are built based on the training dataset and then, the cells from the rest of the dataset can be phenotyped using those models.

In our comparison of semi-supervised algorithms, we found that the Random Forest method performed better than LDA and QDA across different performance methods, training set sizes, image sets. Specifically, we demonstrated the superiority of Random Forests in terms of measures like prediction accuracy, ARI and NMI over the other two methods in three datasets: mIHC ovarian cancer dataset, mIHC lung cancer dataset, MIBI triple negative breast cancer dataset. However, LDA and QDA were both computationally less demanding than the Random Forest.

We have noticed that cells of a few particular types can get incorrectly phenotyped if the corresponding markers are not informative enough. For example, in some patients from the lung cancer dataset, CD19 marker value is not distinctive across different cell types which makes identifying CD19+ cells hard and results in poor prediction performance. Therefore, we advise thorough investigation of the training images via heatmaps and ridge plots of the marker data across different cell types to determine if any of the cell types will potentially have 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.

We summarize our methods as a *R* package named as VectraMIBI (available at this link) that provides visualization tools like heatmaps and ridge plots of the marker data across different cell types and builds a predictive model based on Random Forests.

Declarations

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).

Funding
B.G.B. 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.

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.

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

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.

Ethics approval and consent to participate
Not Applicable.

Consent for publication
Not Applicable.

Authors' information
Not applicable.

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. Liu, P., Liu, S., Fang, Y., Xue, X., Zou, J., Tseng, G., Konnikova, L.: Recent advances in computer-assisted algorithms for cell subtype identification of cytometry data. Frontiers in cell and developmental biology **8**, 234 (2020)
2. Liu, X., Song, W., Wong, B.Y., Zhang, T., Yu, S., Lin, G.N., Ding, X.: A comparison framework and guideline of clustering methods for mass cytometry data. Genome biology **20**(1), 1–18 (2019)
3. Jordan, K.R., Slansky, J., Minic, A., Richer, J.K., Costello, J.C., Clauset, A.J., Kumar, R.T., Behbakht, K., Sikora, M.J., Bitler, B.G.: Characterization of the ovarian tumor transcriptome and microenvironment in pre and post-chemotherapy treated patients. AACR (2020)
4. Johnson, A.M., Boland, J.M., Wrobel, J.L., Klezcko, E.K., Weisner-Evans, M.L., Heasley, L., Clambey, E.T., Nemenoff, R.L., Schenk, E.L.: Cancer cell-specific mhcii expression as a determinant of the immune infiltrate organization and function in the non-small cell lung cancer tumor microenvironment. bioRxiv (2021)
5. Keren, L., Bosse, M., Marquez, D., Angoshtari, R., Jain, S., Varma, S., Yang, S.-R., Kurian, A., Van Valen, D., West, R., *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. Chen, J., Lin, F.: Unsupervised clustering algorithms for flowmass cytometry data. Computational methods with applications in bioinformatics analysis. Singapore: World Scientific Publishing Company, 194 (2017)
7. Levine, J.H., Simonds, E.F., Bendall, S.C., Davis, K.L., El-ad, D.A., Tadmor, M.D., Litvin, O., Fienberg, H.G., Jager, A., Zunder, E.R., *et al.*: Data-driven phenotypic dissection of aml reveals progenitor-like cells that correlate with prognosis. Cell **162**(1), 184–197 (2015)
8. Aghaeepour, N., Finak, G., Hoos, H., Mosmann, T.R., Brinkman, R., Gottardo, R., Scheuermann, R.H.: Critical assessment of automated flow cytometry data analysis techniques. Nature methods **10**(3), 228–238 (2013)
9. Zare, H., Shooshtari, P., Gupta, A., Brinkman, R.R.: Data reduction for spectral clustering to analyze high throughput flow cytometry data. BMC bioinformatics **11**(1), 403 (2010)
10. Sassano, E.: Machine learning methods for flow cytometry analysis and visualization (2018)
11. Li, H., Shaham, U., Yao, Y., Montgomery, R., Kluger, Y.: Deepcytof: Automated cell classification of mass cytometry data by deep learning and domain adaptation. bioRxiv, 054411 (2016)
12. Abdelaal, T., van Unen, V., Höllt, T., Koning, F., Reinders, M.J., Mahfouz, A.: Predicting cell populations in single cell mass cytometry data. Cytometry Part A **95**(7), 769–781 (2019)
13. Lux, M., Krüger, J., Rinke, C., Maus, I., Schlüter, A., Woyke, T., Sczyrba, A., Hammer, B.: acdc–automated contamination detection and confidence estimation for single-cell genome data. BMC bioinformatics **17**(1), 1–11 (2016)
14. Breiman, L.: Random forests. Machine Learning **24**, 123–140 (2001)
15. Kramer, A.S., Latham, B., Diepeveen, L.A., Mou, L., Laurent, G.J., Elsegood, C., Ochoa-Callejero, L., Yeoh, G.C.: 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)
16. Rand, W.M.: Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association **66**(336), 846–850 (1971)
17. Amelio, A., Pizzuti, C.: Is normalized mutual information a fair measure for comparing community detection methods? In: Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, pp. 1584–1585 (2015)
18. Ho, T.K.: Random decision forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition, vol. 1, pp. 278–282 (1995). IEEE
19. McLachlan, G.J.: Discriminant Analysis and Statistical Pattern Recognition vol. 544. John Wiley & Sons, ??? (2004)
20. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
21. van Unen, V., Höllt, T., Pezzotti, N., Li, N., Reinders, M.J., Eisemann, E., Koning, F., Vilanova, A., Lelieveldt, B.P.: Visual analysis of mass cytometry data by hierarchical stochastic neighbour embedding reveals rare cell types. Nature communications **8**(1), 1–10 (2017)
22. Kimball, A.K., Oko, L.M., Bullock, B.L., Nemenoff, R.A., van Dyk, L.F., Clambey, E.T.: A beginner's guide to analyzing and visualizing mass cytometry data. The Journal of Immunology **200**(1), 3–22 (2018)
23. Kobak, D., Berens, P.: The art of using t-sne for single-cell transcriptomics. Nature communications **10**(1), 1–14 (2019)
24. Toghi Eshghi, S., Au-Yeung, A., Takahashi, C., Bolen, C.R., Nyachienga, M.N., Lear, S.P., Green, C., Mathews, W.R., O'Gorman, W.E.: Quantitative comparison of conventional and t-sne-guided gating analyses. Frontiers in immunology **10**, 1194 (2019)

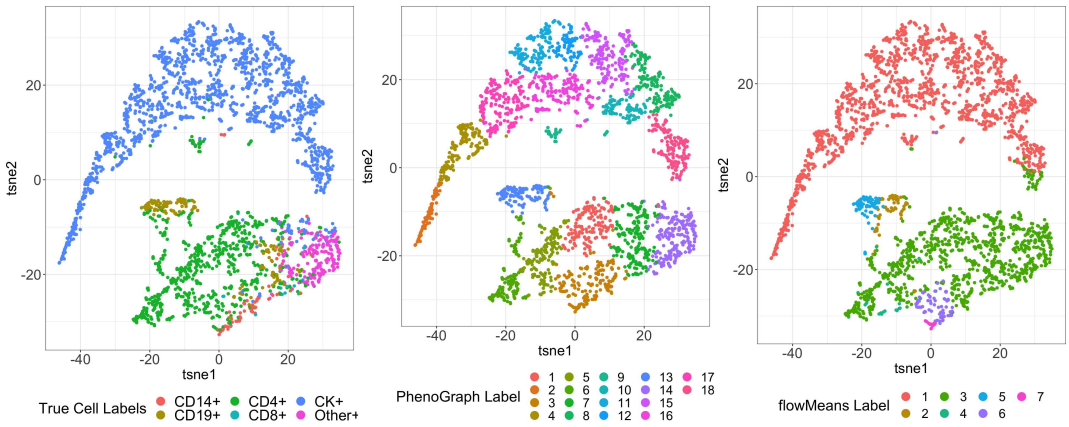

Figure 1 Comparison of cell labels estimated by PhenoGraph and flowMeans with the true cell labels. Left: scatter-plot of TSNE1 and TSNE2 for different cells colored by their true labels. Middle: scatter-plot of TSNE1 and TSNE2 for different cells colored by estimated labels using PhenoGraph. Right: scatter-plot of TSNE1 and TSNE2 for different cells colored by estimated labels using flowMeans.

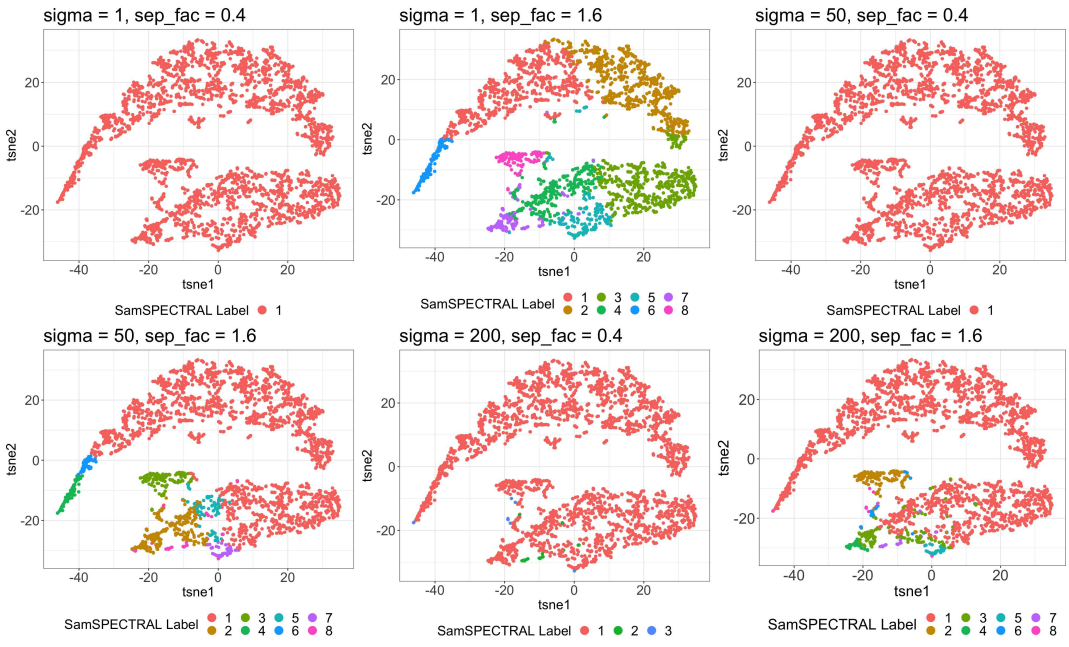

Figure 2 Scatter-plot of TSNE1 and TSNE2 for different cells colored by estimated labels using SamSPECTRAL for different values (from low to high) of sigma and separation factor (`sep_fac`).

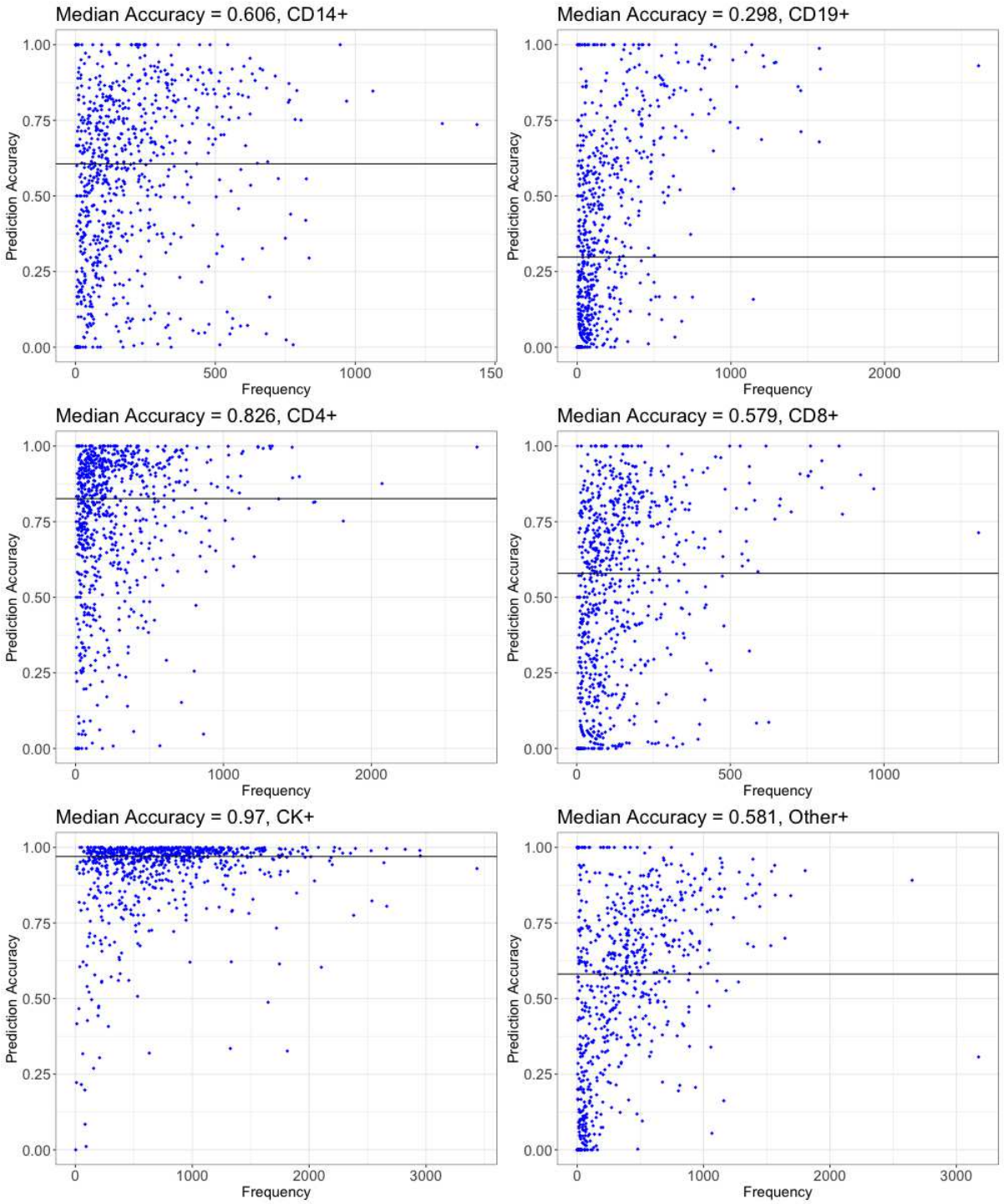

Figure 3 For each cell type, we show scatter-plot of its frequency in each of the images and the corresponding prediction accuracy. Median accuracy for each of the cell type is mentioned above the corresponding sub-plot. For cell type CD19+, the prediction accuracy is quite low, especially in the images where CD19+ cell frequency is low.

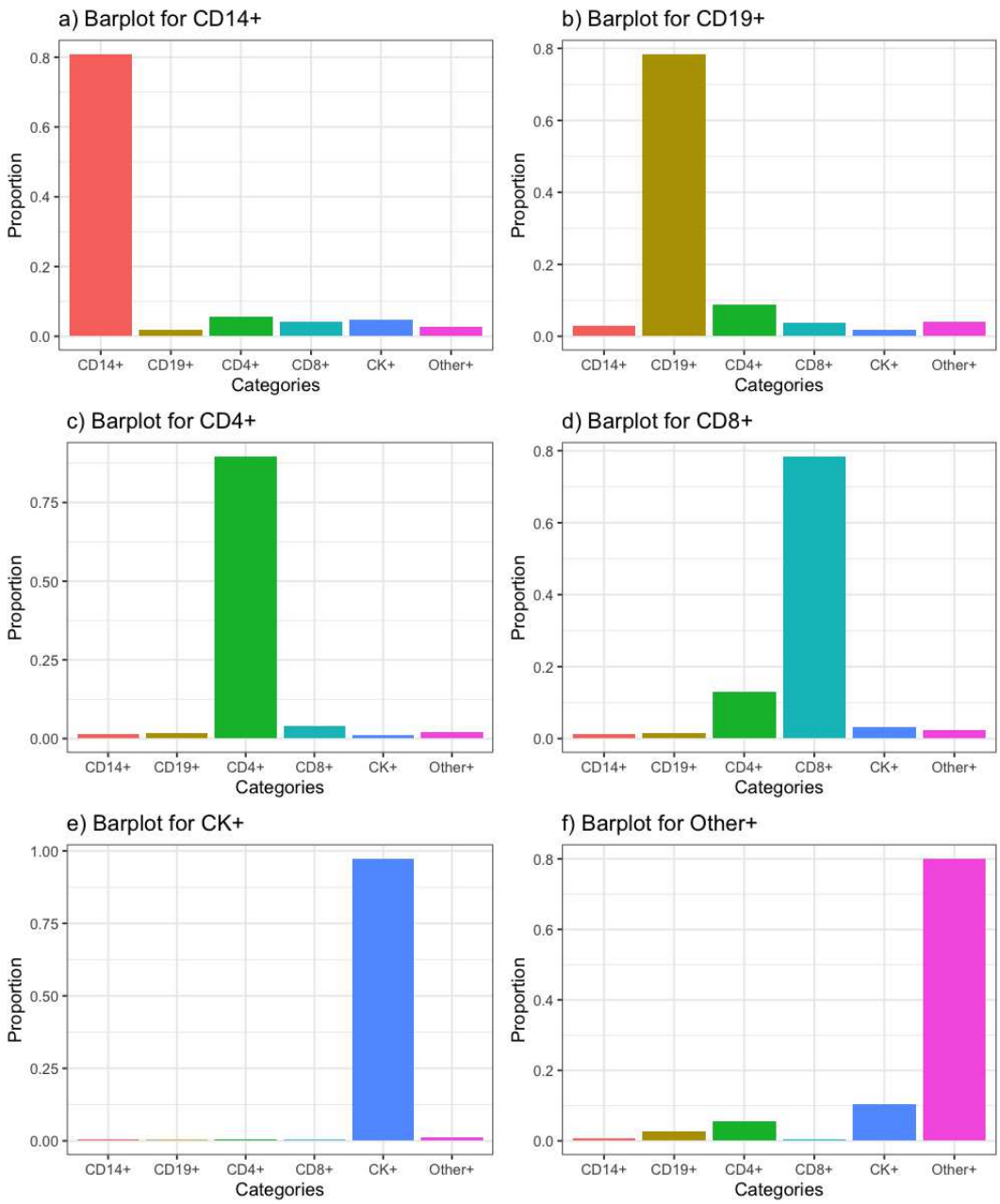

Figure 4 The bar-plots show how many of the cells of a particular type are actually getting predicted to be in that category. For example, (4b) shows that a significant number of CD19+ cells get assigned to CD4+ and Other+ categories.

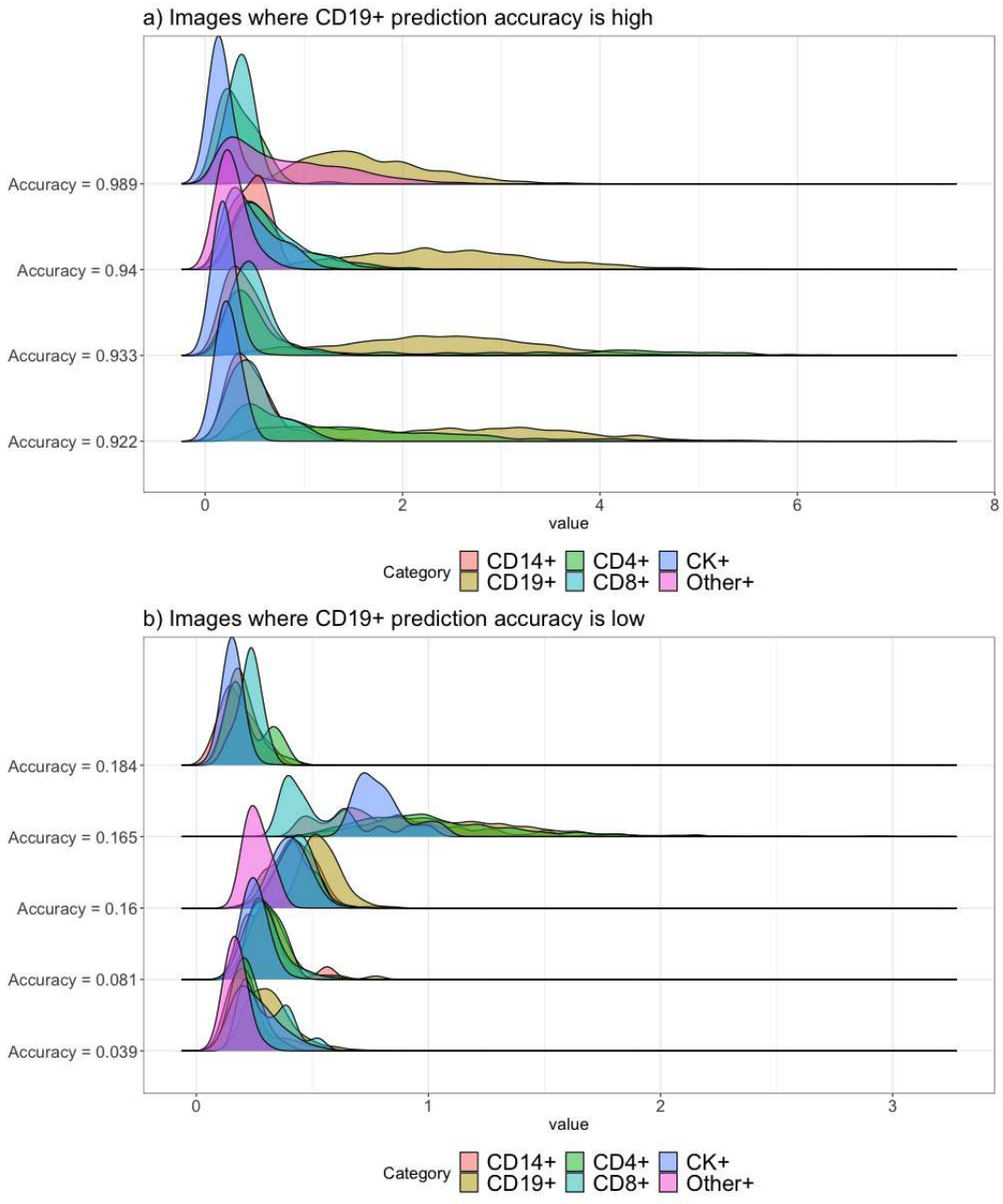

Figure 5 (5a) shows the ridge plot of mean CD19 marker intensity of cells (color coded by original phenotype) from some of the images in which CD19+ prediction accuracy is greater than 0.9. (5b) shows the ridge plot of mean CD19 marker intensity of cells from some of the images in which CD19+ prediction accuracy is less than 0.25.

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 %)

Table 2 Brief description of some of the existing methods in the field of Flow and Mass Cytometry

Type	Methods	Implementation tools	Brief Description
Unsupervised	PhenoGraph	R and Python	Detection of k-nearest neighbors of each cell, Jaccard similarity coefficient as connectivity, community detection based on connection density
	SamSPECTRAL	R	Employ a careful data reduction scheme if the sample size is huge, apply Spectral clustering on the reduced (or, full) data
	flowMeans	R	Modified k-means clustering, merging clusters by distance metrics
Semi-supervised	LDA	R and MATLAB	Linear discriminant analysis with training datasets
	DeepCyTOF	Python	Uses deep learning techniques on training dataset to build up a model that can be used for prediction
	ACDC	Python	Uses a cell type-marker table to determine landmark points based on which classification via random walks is performed

Table 3 Prediction accuracy, ARI and NMI mean ($\pm$ standard deviation) for different training set sizes in mIHC ovarian cancer dataset

Training size	Method	Accuracy	ARI	NMI
5%	Random Forests	0.944 $\pm$ 0.004	0.888 $\pm$ 0.007	0.783 $\pm$ 0.010
	LDA	0.899 $\pm$ 0.017	0.779 $\pm$ 0.047	0.642 $\pm$ 0.051
	QDA	0.909 $\pm$ 0.007	0.821 $\pm$ 0.023	0.699 $\pm$ 0.018
10%	Random Forests	0.949 $\pm$ 0.002	0.896 $\pm$ 0.004	0.795 $\pm$ 0.006
	LDA	0.889 $\pm$ 0.010	0.748 $\pm$ 0.027	0.609 $\pm$ 0.028
	QDA	0.919 $\pm$ 0.003	0.842 $\pm$ 0.007	0.720 $\pm$ 0.008
15%	Random Forests	0.951 $\pm$ 0.002	0.899 $\pm$ 0.003	0.802 $\pm$ 0.006
	LDA	0.898 $\pm$ 0.006	0.772 $\pm$ 0.018	0.633 $\pm$ 0.020
	QDA	0.920 $\pm$ 0.001	0.848 $\pm$ 0.005	0.724 $\pm$ 0.006
20%	Random Forests	0.952 $\pm$ 0.002	0.902 $\pm$ 0.002	0.806 $\pm$ 0.006
	LDA	0.899 $\pm$ 0.007	0.774 $\pm$ 0.018	0.634 $\pm$ 0.023
	QDA	0.922 $\pm$ 0.001	0.853 $\pm$ 0.003	0.727 $\pm$ 0.006

Table 4 Prediction accuracy, ARI and NMI mean (± standard deviation) for different training set sizes in mIHC lung cancer dataset

Training size	Method	Accuracy	ARI	NMI
0.5%	Random Forests	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 Forests	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 Forests	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 Forests	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 Forests	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

Table 5 Prediction accuracy, ARI and NMI mean (± standard deviation) for different training set sizes in MIBI breast cancer dataset

Training size	Method	Accuracy	ARI	NMI
5%	Random Forests	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 Forests	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 Forests	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

Table 6 Computation time (in minutes) across methods, datasets, and training set sizes.

Dataset	Training size	Random Forests	LDA	QDA
mIHC ovarian cancer	5%	6.15	0.798	0.751
	10%	11.01	1.059	1.029
	15%	19.86	1.453	1.395
	20%	23.49	1.585	1.516
mIHC lung cancer	0.5%	4.04	0.82	0.77
	1%	7.55	0.86	0.83
	2%	13.75	1.16	1.05
	3%	21.04	1.37	1.29
	10%	87.02	2.80	2.66
MIBI breast cancer	5%	13.54	0.001	X
	10%	30.93	0.001	X
	20%	68.87	0.002	X