Orals

Jiangmei (Ruby) Xiong

2023-08-15

Table of contents

Statistical Analysis in Multiplexed Immunofluorescence Imaging

This is the document for doctoral oral exam of Jiangmei Xiong. This document will walk you through the basics of multiplexed immunofluorescence image, Jiangmei's first dissertation paper, and a literature review for missing data imputation in multiplexed immunofluorescence imaging.

1 Chapter 1: Introduction to Multiplexed Immunofluorescence Images

1.1 Multiplexed Immunofluorescence Images

Multiplexed Immunofluorescence (mIF) Image is a recent development from Immunofluorescence (IF), a branch of Immunohistochemistry (IHC). The first structural conceptualization of IHC is established in 1941. Coons, Creech, and Jones (1941) described that in formalin-prefixed mammalian tissue, there is a type of antibody that can be identified by fluorescent antigens. Since then, IHC is developed into an important tool for cancer diagnosis (Duraiyan et al. 2012). Within the word "immunohistochemistry", "immuno" refers to the antigen-antibody reaction in the process, "histo" means tissue, and "chemistry" is the process. During IHC, antibody can be tagged with labels such as enzyme, fluorochromes, which reacts when the corresponding antigen-antibody bind is formed (Ramos-Vara 2005). Similarly, the word immunofluorescence can split into "immuno" and "fluorescence", and "fluorescence" corresponds to the fluorescent signal generated by the fluorocromes(Hussaini, Seo, and Rich 2022).

IHC/IF can only detect one biomarker for a tissue region. This limitation makes IHC/IF unable to identify more complicated expression patterns that require more than one biomarker (Sheng et al. 2023). The development of multiplexed IHC (mIHC)/IF(mIF) image resolved this issue. Multiplexed IHC/IF image display different protein information for each plex of the image, while retaining the spatial and morphological information of the tissue (Eng et al. 2022).

Figure **??** by Sheng et al. (2023) shows several different methods for creating mIF images. It can be seen that the first two methods both uses cycles of stain-photo-removal, and the last method is a one-off step where all labels are tagged at once. Sheng et al. (2023) also tabulated all multiplexd IHC/IF technologies, where the number of biomarkers that can be identified ranges from 4 to 100.

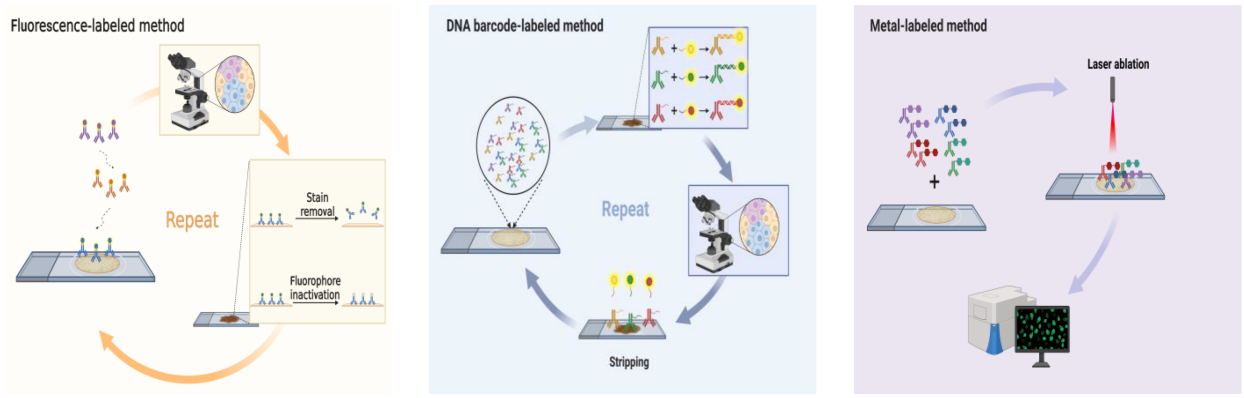

Figure 1.1: Different methods for creating mIF/mIHC images. Image courtesy of Sheng et al. (2023)

mIHC/mIF images are widely used in studies for immune tumor microenvironments (iTME). The studies often involves cell type proportions within a certain region, spatial clustering of immune cells or spatial interaction among different cell types (Wrobel, Harris, and Vandekar 2023). For example: Schürch et al. (2020) discovered that within granulcyte cell meightborhood, the enrichment PD-1+CD4+ T cells are correlated with the survival outcome of a subset of colorectal cancer patients; Chen et al. (2021) shows different immune cell proportion and clustering between different colorectal tumor types; Steinhart et al. (2021) found that certain immune cell proportions and spatial interactions are correlated with ovarian cancer patient survival outcomes.

1.2 Data Structure

For mIHC/mIF images, each individual "plex" corresponds to a different immune protein identified by a type of stain. Each plex goes through analogous transition as shown in Figure **??** and form a table of cell expressions, and the tables of different protein expressions are combined in the end. Initially, greyscale intensity is assigned to each pixel. The greyscale intensity is taken as the intensity of marker expression. The image then goes through cell segmentation, which are usually based on machine learning or deep learning methods(McKinley et al. 2022; Schüffler et al. 2015). DAPI, a fluorescent stain typically used for cell morphology identification, is most often used for the initial cell segmentation. The same cell-segmentation will be used for all other marker channels. Next, the pixel intensities, pixel positions, and the cell that the pixel belongs to are entered into a dataframe. Finally, the pixel intensities and position will be averaged within the cell group. Often, the median of pixel intensities is used as well, to reduce the impact of pixel intensity outliers. The end result after combining all marker channels will be a dataset with each row representing an individual cell, columns of different marker expressions, and cell properties such as position, cell type (e.g. tumor cell or not, tissue type).

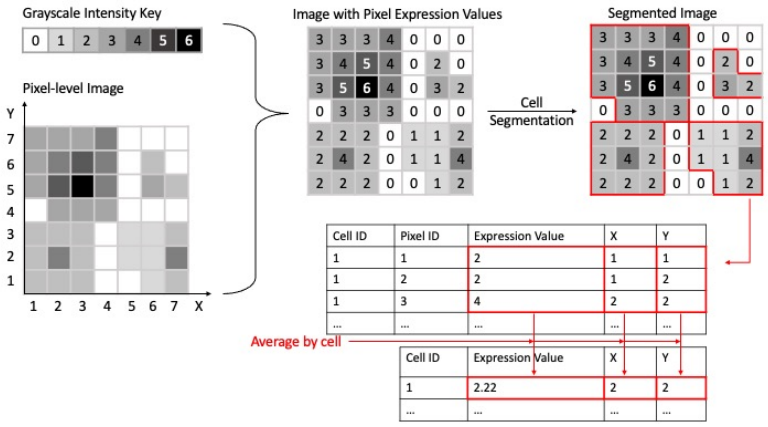

Figure 1.2: Transition from a single-plex mIHC/mIF image to a single-cell dataframe. The greyscale intensity range is only a demonstration. In application, the range of the greyscale intensity depends on the configuration of image softwares. The cell intensity of the bottom left cell is shown in the table as an example.

2 Chapter 2: GammaGateR

2.1 Research question

Before important spatial insights can be gleaned using statistical methods, mIF images undergo an intensive preprocessing pipeline to obtain single-cell measurements. While there are various steps included in the pipeline such as image registration, single-cell segmentation, quantification, and batch correction, cell phenotyping is typically the final step before downstream analyses on the cell-level data, similarly to other single-cell assays. Cell phenotyping identifies individual cell phenotypes from the measured marker expression values of the cell and directly affects the subsequent cell population analysis results.

The two most common approaches for cell phenotyping in mIF are manual gating and graph-based multivariate clustering. In manual gating, each sample is visualized separately to determine a threshold, and super-threshold cells are labeled as marker positive. This procedure is repeated for all marker channels and slides, and the phenotypes are determined by combining combinations of marker-positive cells. Alternatively, multivariate graph-based clustering is adapted from other single-cell assays. This approach first performs cell clustering, then assigns a phenotype to each cell group based on their average expression profile. Multivariate graph-based clustering is implemented with various modifications across many software packages. Unfortunately, both methods are labor intensive, and their accuracy suffers from image noise and spatial artifacts in mIF images that cause marker expression histograms to appear continuous or uni-modal . As a result, both phenotyping methods possess shortcomings that cannot be ignored. On one hand, manual gating can be subjective. On the other hand, graph-based clustering results are prone

to over-clustering and producing poor separation between clusters.

2.2 Previous works

The challenges described above are well recognized and there are a few methods and software developed that attempt to automate cell phenotyping for mIF images. For example, CellSighter is a recently proposed supervised deep-learning algorithm for cell phenotyping that requires a "gold standard" training dataset. Another recent solution, ASTIR (Automated assignment of cell identity from single-cell multiplexed imaging and proteomic data), is a fast unsupervised approach that defines cell phenotypes from segmented cell-level data by using a neural network-based mixture model assuming a multivariate log-normal distribution. Instead of binary outputs like in classification methods, ASTIR returns posterior probabilities of different cell types for each cell. This type of output is advantageous because it offers more information than nominal cell types and leaves cell labeling to the clinician's discretion. Lastly, Ahmadian et al. treat the analysis as a pixel classification problem and design a single-step framework for mIF phenotyping that is integrated with other preprocessing steps.

Nevertheless, inconsistencies persist in the results rendered by these learning-based methods when applied across markers, slides, batches, and datasets. These inconsistencies result from the immense variation in the cell-level distribution of phenotyping markers that are often too nuanced to be removed by existing batch correction methods. For these reasons, it is difficult to fully automate the cell phenotyping process, despite the availability of automated tools, and manual gating is still used to perform cell phenotyping because it is easy to visualize and evaluate the quality of the phenotype.

2.3 Methods

Because automated methods cannot be run without evaluation and supervised methods require a gold-standard dataset, no

method is truly fully automated. As a solution, we develop an explicitly semi-automated algorithm called GammaGateR. GammaGateR allows the user to easily perform cell phenotyping, visualize results, and conduct interpretable quality control while reducing manual labor. Based on a novel closed-form Gamma mixture model (cfGMM), GammaGateR is a probabilistic model that is fitted to each channel and slide separately, and outputs positive-component probabilities for each marker. These can then be easily thresholded and combined for semi-automated marker gating or input directly into downstream analysis. GammaGateR has important technical advantages, including 1) improved computation time and model convergence due to its novel closed-form property, and 2) high consistency and reproducibility for phenotyping results across mIF data batches due to incorporation of parameter boundary constraints. In applications on real-world mIF data, we find that GammaGateR has fast and consistent results across many slides and markers. We provide an open-source implementation of our method in the new GammaGateR R package (https://github.com/jiangmeirubyxiong/gammagater).

2.3.1 GammaGateR

The GammaGateR algorithm is unique to existing methods for its focus on parsimoniously modeling cell-level marker expression densities. This approach yields tailored-to-slide model estimation in cell-level mIF data where marker expression distributions can vary substantially across slides. The algorithm uses a zero-inflated two-component GMM to model marker expression for each slide. The Gamma mixture model naturally identifies marker-positive and marker-negative cell distributions and returns the probability of belonging to the marker-positive cell distribution for each cell. The returned probabilities can either be used directly in subsequent analysis or combined and dichotomized to define cell phenotypes. GammaGateR incorporates user-specified constraints to provide consistent model fit across a large number of slides. The model evaluation methods included in GammaGateR allow the user to evaluate the constraints and quality check results. The power source of GammaGateR is the closed-form Gamma mixture model, which is

a novel approach to phenotyping for mIF data that makes it more computationally efficient than traditional GMMs.

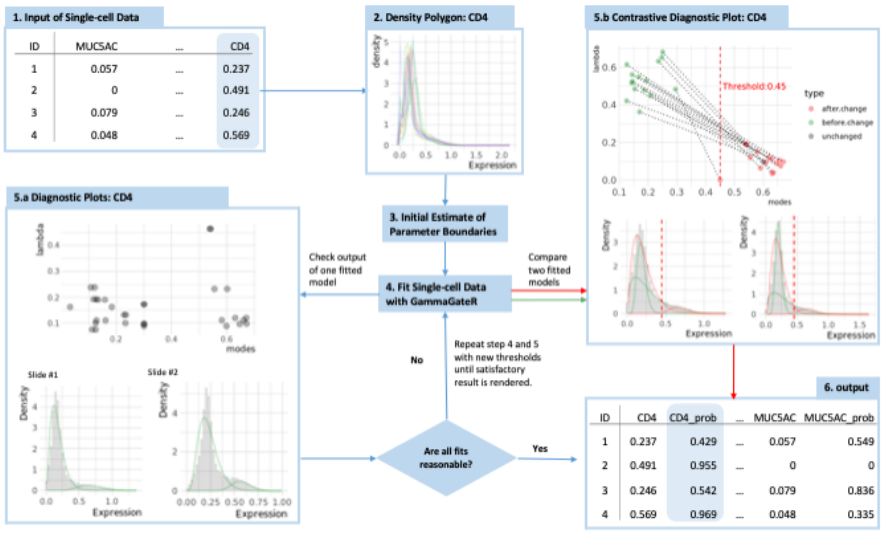

Figure 2.1: Figure 1

2.3.2 cfGMM

For mIF data, we use the GMM to fit cell marker expression values as a weighted sum of different probability distributions that represent unique cell populations [35]. The Gamma distribution is an excellent model for marker values because the domain of the Gamma distribution is strictly positive and it has the flexibility to model the varying skewed densities seen in mIF marker values (Figure 1.5.a). However, GMMs are not scalable for mIF image data, because they rely on computationally inefficient numerical methods to obtain the maximum likelihood estimator (MLE). The slow convergence of the MLE for the GMM makes it prohibitive to apply across a large number of channels, slides, and cells. As a solution, we develop a closed-form GMM (cfGMM; https://github.com/jiangmeirubyxiong/cfgmm) estimation procedure based on a recently developed estimator for the Gamma distribution [36]. In addition, to improve computational efficiency, the cfGMM has the benefit of allowing prior constraints on model parameters. With the cfGMM in