Statistical analysis of single-cell protein data

Brooke L. Fridley, PhD
Department of Biostatistics and Bioinformatics, Moffitt Cancer Center
Tampa, FL 33612, USA
Biostatistics and Epidemiology Core, Children's Mercy Hospital
Kansas City, MO 64108, USA
Email: Brooke.Fridley@Moffitt.org; Fridley.Brooke@gmail.com

Simon Vandekar, PhD
Department of Biostatistics, Vanderbilt University Medical Center
Nashville, TN 37203, USA
Email: Simon.Vandekar@VUMC.org

Inna Chervoneva, PhD
Division of Biostatistics, Thomas Jefferson University
Philadelphia, PA 19107, USA
Email: Inna.Chervoneva@Jefferson.edu

Julia Wrobel, PhD
Department of Biostatistics and Bioinformatics, Emory University
Atlanta, GA 30322, USA
Email: Julia.Wrobel@Emory.edu

Siyuan Ma, PhD
Department of Biostatistics, Vanderbilt University Medical Center
Nashville, TN 37203, USA
Email: Siyuan.Ma@VUMC.org

Immune modulation is considered a hallmark of cancer initiation and progression, with immune cell density being consistently associated with clinical outcomes of individuals with cancer. Multiplex immunofluorescence (mIF) microscopy combined with automated image analysis is a novel and increasingly used technique that allows for the assessment and visualization of the tumor microenvironment (TME). Recently, application of this new technology to tissue microarrays (TMAs) or whole tissue sections from large cancer studies has been used to characterize different cell populations in the TME with enhanced reproducibility and accuracy. Generally, mIF data has been used to examine the presence and abundance of immune cells in the tumor and stroma compartments; however, this aggregate measure assumes uniform patterns of immune cells throughout the TME and overlooks spatial heterogeneity. Recently, the spatial contexture of the TME has been explored with a variety of statistical methods. In this PSB workshop, speakers will present some of the state-of-the-art statistical methods for assessing the TIME from mIF data.

Keywords: spatial biology, multiplex immunofluorescence, single-cell protein, tumor microenvironment, biostatistical analysis, spatial analysis

1. Introduction, Background and Motivation

The treatment of cancers has been revolutionized in recent years with the advent of immunotherapies[1-6]. However, not all patients respond to immunotherapies and a subset of patients that initially respond to immunotherapy go on to develop resistance. To understand why some patients do not respond to immunotherapies, much research has been devoted to understanding the role of the immune contexture of the tumor immune microenvironment (TIME) and its association with clinical outcomes[4,7-9]. Thus, immune profiling using a variety of approaches has become an important part of immuno-oncology.

Some commonly used approaches for studying the tumor immune microenvironment include (but are not limited to): flow cytometry[10], imaging mass cytometry[11], immunohistochemistry (IHC)[12], immune cell devolution of bulk RNA-seq data[13], single-cell RNA-seq[14], spatial transcriptomics[15] and multiplex immunofluorescence (mIF)[16]. Multiplex immunofluorescence microscopy combined with automated image analysis is a novel and increasingly used technique that allows for the assessment and visualization of the TME. This technology has been applied to a variety of sample types, from whole slide images to regions of interest (ROIs)[17] and tissue microarrays (TMAs)[18,19].

As with any new technology, there are inevitability challenges with the statistical analysis of the single-cell imaging data[20,21]. Some of the challenges come from cell phenotyping, which is labeling cells as positive or negative for each antibody of interest. This is a necessary preprocessing step that occurs before spatial data analysis that is critical for accurately estimating immune cell abundance in the TIME. After phenotyping, it is typical to measure immune cell abundance, typically calculated as percent or proportions of specific cell types in the tumor compartment of the tissue. A challenge of this task is that many cell types are often observed at low-abundance (i.e., zero-inflated), particularly in low immune infiltrated tumors (e.g., immune “cold” tumors).

Besides the protein markers used for phenotyping cells, it is often of important to quantify the actual levels of proteins of interest in all or some cell types. Such quantitative functional markers may include proliferation markers (e.g., Ki-67, PCNA), checkpoint proteins (e.g., PD-1, PD-L1, CTLA-4) and growth factors and receptors (e.g., EGFR, HER2). Traditionally, a single mean expression level across the cells of interest is computed and considered as a biomarker. This approach ignores important tumor heterogeneity and has low sensitivity for detecting high expression in some portion but not all cells of interest. Alternative approaches have been recently developed[22,23] using the entire distributions of single-cell protein expression levels in a tumor tissue to derive quantitative functional markers.

Finally, there is growing evidence that the spatial architecture of the TIME has high impact on disease progression and response to immunotherapy. Generally, mIF data has been used to examine the presence and abundance of immune cells in the TIME; however, this aggregate measure assumes uniform patterns of immune cells throughout the tumor and overlooks spatial heterogeneity. Recently, the spatial contexture of the TIME has been explored with a variety of spatial statistical methods, including those for assessing co-localization. In this session, speakers will present some of the state-of-the-art statistical methods for assessing the TIME from mIF data. All slides and R code presented during the workshop can be found at http://juliawrobel.com/PSB_scProteomics .

2. Speaker Abstracts

Overview of abundance-based and spatial-based analysis approaches for multiplex imaging data

Brooke L Fridley

With the advent of immunotherapies for the treatment of cancer, much research is being conducted to understand the tumor immune microenvironment (TIME). To date, much of the research completed has focused on understanding the abundance of different immune cell subsets in the TIME using either single-cell RNA-seq or multiplex immunofluorescence (mIF). One benefit in using mIF based technologies is that, in addition to abundance of immune cells, one is also able to get the spatial location of these cells within the TIME. Thus, researchers can answer question that relate to the spatial architecture or contexture of the TIME and how this might impact clinical outcomes. In this presentation, we provide an overview of how mIF data is generated and analysis methods used for assessing the non-spatial aspects of the TIME (i.e., abundance level analyses). After providing an overview of mIF data and abundance-based analysis approaches, we will review a variety of spatial statistical approaches for analyzing the spatial contexture. To facilitate spatial analyses, we will also present on an R package, *spatialTIME,* developed to generate these spatial statistics on large sets of samples[17,24].

Normalization and Cell Phenotyping for mIF data

Simon Vandekar

Normalization and cell phenotyping are critical steps in the multiplexed image analysis pipeline prior to performing downstream statistical analysis because they remove batch effects and identify consistent cell types across slides. These analysis steps are particularly challenging for mIF data due to the unique heterogeneity of the image intensities across slides and overlapping cell distributions. We review some recently proposed normalization methods [25,26] and discuss the three main procedures for cell phenotyping (marker gating, unsupervised clustering, and supervised algorithms), in the context of mIF imaging[27], including our recently developed semi-supervised algorithm, *GammaGateR*. The R package *GammaGateR* focuses on efficiently estimating the marginal distributions of single-cell marker intensities using a novel closed-form Gamma mixture model to identify marker positive cells. It incorporates biological constraints to improve consistency across a large number of slides and allows users to interactively curate the model fit. We compare several cell phenotyping algorithms developed for multiplexed imaging and demonstrate how to use the results to perform spatial analyses of mIF imaging data.

Quantile biomarkers based on single-cell multiplex immunofluorescence imaging data

Inna Chervoneva

Modern pathology platforms for multiplex fluorescence-based immunohistochemistry provide distributions of cellular signal intensity (CSI) levels of proteins across the entire cell populations within the sampled tumor tissue. However, heterogeneity of CSI levels is usually ignored, and the simple mean signal intensity (MSI) value is considered as a cancer biomarker. To account for tumor heterogeneity, we consider the entire CSI distribution as a predictor of clinical outcome. This allows retaining all quantitative information at the single-cell level by considering the values

of the quantile function (inverse of the cumulative distribution function) estimated from a sample of CSI levels in a tumor tissue.

A simple and intuitive approach is to select an optimal quantile of the CSI distribution as the best predictor of clinical outcome of interest. In Yi et al (2023)[23], we developed an algorithm, implemented in the R package *Qindex,* for selecting optimal CSI distribution quantiles as best predictors of outcome. The proposed algorithm was used to select optimal quantile biomarkers of progression-free survival in a large cohort of breast cancer patients and validated in an independent external validation cohort. The optimal quantile protein biomarkers yielded generally improved prognostic value as compared to the standard MSI biomarkers.

A more comprehensive approach is to derive new biomarkers as single-index predictors based on the entire CSI distribution summarized as a quantile function. [22] The proposed Quantile Index (QI) biomarker is defined as a linear or nonlinear functional regression predictor of outcome. The linear functional regression quantile Index (FR-QI) is the integral of subject-specific CSI quantile function multiplied by the common weight function[22]. The nonlinear functional regression quantile index (nFR-QI) is computed as the integral of unspecified bivariate twice differentiable function with probability p and subject-specific quantile function as arguments. The weight and nonlinear bivariate function are represented by penalized splines and estimated by fitting suitable functional regression models to a clinical outcome. The proposed QI biomarkers were derived for proteins expressed in cancer cells of malignant breast tumors and compared to the standard MSI predictors and optimal quantile protein biomarkers[23]. The R package *Qindex* implements the optimization of QI biomarkers and their evaluation in an independent test set.

Tools and software for functional data analysis of multiplexed imaging data
Julia Wrobel

The TME, which characterizes the tumor and its surroundings, plays a critical role in understanding cancer development and progression. Recent advances in imaging techniques enable researchers to study spatial structure of the TME at a single-cell level. Many popular approaches for analyzing spatial relationships between cell types or quantifying spatial co-expression of biological markers in multiplex imaging data are based on point process theory. The location of cells in mIF data are treated as following a point process, realizations of a point process are called “point patterns”, and point process models seek to understand correlations in the spatial distributions of cells. Under the assumption that the rate of a cell is constant over an entire region of interest a point pattern will exhibit complete spatial randomness (CSR), and it is often of interest to model whether cells deviate from CSR either through clustering or repulsion.

Spatial summary functions characterize the degree of spatial interaction among cells across different radii, however, these are often evaluated at a single arbitrarily chosen cellular distance. Using techniques from functional data analysis, we introduce an approach to model the association between these summary spatial functions and patient-level survival outcomes across all radii simultaneously, while controlling for other clinical scalar predictors such as age and disease stage. In addition, we introduce a novel hypothesis test to what level of model flexibility is most appropriate for a given multiplex imaging dataset. Finally, our methods are implemented in *mxfda,* a general-purpose R package for functional data analysis of multiplex imaging data.

A Flexible Generalized Linear Mixed Effects Model for Testing Cell-Cell Colocalization in Spatial Immunofluorescent Data
Siyuan Ma

mIF data analysis is interested in characterizing the nuanced spatial context of tissue microenvironments, such as the infiltration or exclusion of certain immune cell populations in tumor tissues. To test for cell colocalization or exclusion events, existing methods often rely on image-wide statistics to create null distributions for cell colocalization events and evaluate their statistical significance[28]. Given that tissue characteristics can be image-specific (i.e., size of images, the local topology of tissue organization), this type of approach does not generalize well for comparisons between images/conditions. We show that, by examining cell colocalization events on a per-cell basis, they can be modeled with common count-based distributions such as the binomial. As such, cell colocalization or exclusion can be practically analyzed with generalized linear mixed effects models with spatially correlated error terms. This allows flexible inclusion and testing of image/condition effects and subject-specific correlations, because they can be easily modeled as fixed or random regression effects. We demonstrate that this model relies on essentially the same assumptions as existing image-wide modeling approaches. In practice, it can be implemented with the readily available R package *spaMM*. We exemplify the utility of such a model with an application in protein immunofluorescent imaging of inflammatory bowel disease tissues[29].

3. Acknowledgments

This research was supported in part by the National Institutes of Health (R01 CA279065, R01 CA222847).

References

1. Ribas A, Wolchok JD. Cancer immunotherapy using checkpoint blockade. *Science*. Mar 23 2018;359(6382):1350-1355. doi:10.1126/science.aar4060
2. Couzin-Frankel J. Breakthrough of the year 2013. Cancer immunotherapy. *Science*. Dec 20 2013;342(6165):1432-3. doi:10.1126/science.342.6165.1432
3. Drake CG, Lipson EJ, Brahmer JR. Breathing new life into immunotherapy: review of melanoma, lung and kidney cancer. *Nat Rev Clin Oncol*. Jan 2014;11(1):24-37. doi:10.1038/nrclinonc.2013.208
4. Havel JJ, Chowell D, Chan TA. The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy. *Nat Rev Cancer*. Mar 2019;19(3):133-150. doi:10.1038/s41568-019-0116-x
5. Thorsson V, Gibbs DL, Brown SD, et al. The Immune Landscape of Cancer. *Immunity*. Apr 17 2018;48(4):812-830 e14. doi:10.1016/j.immuni.2018.03.023
6. Pardoll DM. The blockade of immune checkpoints in cancer immunotherapy. *Nat Rev Cancer*. Mar 22 2012;12(4):252-64. doi:10.1038/nrc3239
7. Martinez-Morilla S, Villarroel-Espindola F, Wong PF, et al. Biomarker Discovery in Patients with Immunotherapy-Treated Melanoma with Imaging Mass Cytometry. *Clin Cancer Res*. Apr 1 2021;27(7):1987-1996. doi:10.1158/1078-0432.CCR-20-3340
8. Thurin M, Cesano A, Marincola, eds. *Biomarkers for immunotherapy of cancer* Springer; 2020. *Methods in Molecular Biology*

9. Fridman WH, Zitvogel L, Sautes-Fridman C, Kroemer G. The immune contexture in cancer prognosis and treatment. *Nat Rev Clin Oncol*. Jul 25 2017;doi:10.1038/nrclinonc.2017.101
10. Cossarizza A, Chang HD, Radbruch A, et al. Guidelines for the use of flow cytometry and cell sorting in immunological studies (second edition). *Eur J Immunol*. Oct 2019;49(10):1457-1973. doi:10.1002/eji.201970107
11. Baharlou H, Canete NP, Cunningham AL, Harman AN, Patrick E. Mass Cytometry Imaging for the Study of Human Diseases-Applications and Data Analysis Strategies. *Front Immunol*. 2019;10:2657. doi:10.3389/fimmu.2019.02657
12. Magaki S, Hojat SA, Wei B, So A, Yong WH. An Introduction to the Performance of Immunohistochemistry. *Methods in molecular biology*. 2019;1897:289-298. doi:10.1007/978-1-4939-8935-5_25
13. Sturm G, Finotello F, Petitprez F, et al. Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics*. Jul 15 2019;35(14):i436-i445. doi:10.1093/bioinformatics/btz363
14. Haque A, Engel J, Teichmann SA, Lonnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome medicine*. Aug 18 2017;9(1):75. doi:10.1186/s13073-017-0467-4
15. Burgess DJ. Spatial transcriptomics coming of age. *Nature reviews*. Jun 2019;20(6):317. doi:10.1038/s41576-019-0129-z
16. Tan WCC, Nerurkar SN, Cai HY, et al. Overview of multiplex immunohistochemistry/immunofluorescence techniques in the era of cancer immunotherapy. *Cancer Commun (Lond)*. Apr 2020;40(4):135-153. doi:10.1002/cac2.12023
17. Wilson C, Soupir AC, Thapa R, et al. Tumor immune cell clustering and its association with survival in African American women with ovarian cancer. *PLoS Comput Biol*. Mar 2022;18(3):e1009900. doi:10.1371/journal.pcbi.1009900
18. Hathaway CA, Wang T, Townsend MK, et al. Lifetime Exposure to Cigarette Smoke and Risk of Ovarian Cancer by T-cell Tumor Immune Infiltration. *Cancer Epidemiol Biomarkers Prev*. Jan 9 2023;32(1):66-73. doi:10.1158/1055-9965.EPI-22-0877
19. Hathaway CA, Conejo-Garcia JR, Fridley BL, et al. Measurement of ovarian tumor immune profiles by multiplex immunohistochemistry: implications for epidemiologic studies. *Cancer Epidemiol Biomarkers Prev*. Mar 20 2023;doi:10.1158/1055-9965.EPI-22-1285
20. Wilson CM, Ospina OE, Townsend MK, et al. Challenges and Opportunities in the Statistical Analysis of Multiplex Immunofluorescence Data. *Cancers (Basel)*. Jun 17 2021;13(12)doi:10.3390/cancers13123031
21. Wrobel J, Harris C, Vandekar S. Statistical Analysis of Multiplex Immunofluorescence and Immunohistochemistry Imaging Data. *Methods in molecular biology*. 2023;2629:141-168. doi:10.1007/978-1-0716-2986-4_8
22. Yi M, Zhan T, Peck AR, et al. Quantile Index Biomarkers Based on Single-Cell Expression Data. *Laboratory investigation; a journal of technical methods and pathology*. Aug 2023;103(8):100158. doi:10.1016/j.labinv.2023.100158
23. Yi M, Zhan T, Peck AR, et al. Selection of optimal quantile protein biomarkers based on cell-level immunohistochemistry data. *BMC Bioinformatics*. Jul 22 2023;24(1):298. doi:10.1186/s12859-023-05408-8
24. Creed JH, Wilson CM, Soupir AC, et al. spatialTIME and iTIME: R package and Shiny application for visualization and analysis of immunofluorescence data. *Bioinformatics*. Nov 4 2021;doi:10.1093/bioinformatics/btab757

25. Harris CR, McKinley ET, Roland JT, et al. Quantifying and correcting slide-to-slide variation in multiplexed immunofluorescence images. *Bioinformatics*. Mar 4 2022;38(6):1700-1707. doi:10.1093/bioinformatics/btab877
26. Graf J, Cho S, McDonough E, et al. FLINO: a new method for immunofluorescence bioimage normalization. *Bioinformatics*. Jan 3 2022;38(2):520-526. doi:10.1093/bioinformatics/btab686
27. Geuenich MJ, Hou J, Lee S, Ayub S, Jackson HW, Campbell KR. Automated assignment of cell identity from single-cell multiplexed imaging and proteomic data. *Cell Syst*. Dec 15 2021;12(12):1173-1186 e5. doi:10.1016/j.cels.2021.08.012
28. Schapiro D, Jackson HW, Raghuraman S, et al. histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data. *Nat Methods*. Sep 2017;14(9):873-876. doi:10.1038/nmeth.4391
29. Kondo A, Ma S, Lee MYY, et al. Highly Multiplexed Image Analysis of Intestinal Tissue Sections in Patients With Inflammatory Bowel Disease. *Gastroenterology*. Dec 2021;161(6):1940-1952. doi:10.1053/j.gastro.2021.08.055