STATISTICAL METHODS AND COMPUTATIONAL TOOLS FOR NORMALIZATION AND SPATIAL ANALYSIS OF MULTIPLEXED IMAGING DATA

By

Coleman Reed Harris

Dissertation
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
in

Biostatistics

May 31, 2022

Nashville, Tennessee

Approved:

Jonathan Schildcrout, Ph.D.

Simon Vandekar, Ph.D.

Hakmook Kang, Ph.D.

Ken Lau, Ph.D.

For my incredible son, Jamie.

ACKNOWLEDGMENTS

First, I would like to thank my advisor Simon Vandekar for his continued support and direction throughout my doctoral degree. I came to you with the idea of doing "spatial statistics" together, and we jumped into this work head-first – for that, I am extremely grateful. I would also like to thank Jonathan Schildcrout, Hakmook Kang, and Ken Lau for serving on my committee and providing valuable feedback on this body of work. I want to thank Julia Wrobel for her mentorship and feedback on many of these projects, and I want to thank my collaborators for welcoming me with open arms into the multiplexed imaging field and for supporting my work – Robert Coffey, Eliot McKinley, Joseph Roland, Cody Heiser, Qi Liu, and Martha Shrubsole.

The NIH Big Data to Knowledge training grant gave me the flexibility to pursue research questions that fulfill my passions – for that I want to thank Jeffrey Blume for his encouragement and Bradley Malin for his continued support. I appreciate the mentorship from you both more than you know.

I have been fortunate enough to spend a portion of my career in various internships and part-time positions over the past 5 years. I want to thank Jud Schneider for my internship with Nashville Biosciences and for taking a chance on me full-time after my PhD. I also want to thank Chase Spurlock, Lukasz Wylezinski, Jamieson Gray, and the Decode Health team for the years of experience in data science. And to the many people whose path I've crossed while freelance writing, exploring science policy, and collaborating – thank you.

Although it has been a few years since my time at the University of Alabama, I want to thank two people who played a major part in starting my research career. First, a big thank you to Adam Hauser for being my first research mentor and for the experience I gained working in his lab. I also want to thank Bruce Barrett for his guidance and mentorship as I began my career in statistics.

This degree would not have been possible without the incredible support from my family and friends. The past five years have been a whirlwind and you all have stuck around, supported me, and helped me through this rollercoaster. To my parents – Adam Harris & Beth Wollmuth, to the Woodsides – Caitlyn, Ginger, & Mark, to my girlfriend and best friends – Brecken Overly, David Solon, & Reid Jenkins, to my cohort – Alese Halvorson, Ben Lane, Cassie Hennessy, Jess Brown, Megan Murray, Michael Williams, & Yuqi Tian, and of course to my son Jamie: **Thank you**.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER 1

Introduction

1.1 Overview

This dissertation deals with problems arising within the multiplexed imaging pipeline, using traditional statistical methods and algorithms in a new type of biological imaging data. In Chapter 2, we demonstrate the utility of applying normalization algorithms in multiplexed imaging data to remove slide effects and provide an evaluation criteria for better data quality. In Chapter 3, we introduce the `mxnorm` R package to implement, evaluate, and visualize normalization techniques in multiplexed imaging, which also provides options for researchers to introduce and assess user-defined normalization methods. In Chapter 4, we build upon the removal of technical variation in this pipeline, adapting spatial statistics methods that leverage marked point processes for use in analyzing multiplexed imaging data.

Before we dive in, we will introduce the basics of multiplexed imaging and some of the statistical methods used in this thesis to equip the reader with background information to understand the later chapters.

1.2 Multiplexed imaging

Multiplexed imaging methods have only been developed and introduced within the last decade – these methods extend the idea of single-cell sequencing (e.g., assigning some quantitative value for a biomarker or protein to a specific cell) and seek to take this measure *in situ*, or within a tissue sample. These can be considered "stacked" images which provide detailed spatial information about interactions between cells and tissue types in complex biological processes like cancer, tumor development, and more. The "stacked" component is better defined as staining a sample for a biomarker of interest, generating a set of images, washing the sample, and re-staining it for a separate marker. This process is repeated dozens of times depending on the number of markers used in a given study. Biological samples are typically a section of tissue, usually related to a type of tumor or cancer. Images are collected as "regions of interest", or ROIs – subsets of the full tissue sample, where the entire tissue sample is not always imaged and images are not necessarily contiguous. This criteria for creating multiplexed images broadly sets the scene for the kind of studies and data we work with – the types of methods introduced to perform these processes vary across research labs and technologies, and will be discussed more in depth in the chapters below.

1.3 Pipelines and data

Multiplexed imaging experiments generate data across hundreds of slides and images, creating terabytes of data via imaging analysis pipelines. Data are collected in batches for multiple marker channels across slides, where, as noted above, the images comprise smaller regions of the tissue sample. Within a given slide or image, individual cells are identified using segmentation algorithms – this introduces the problem of a “ground truth” in this type of data, where cell- and tissue-labels are often identified in an unsupervised manner. Quantitative values for a marker channel are then assigned at the median- or mean-cell level, where some cell has coordinates (x, y) on ROI j, taken from slide i, for some marker channel c. This data structure allows for the *in situ* analysis of multiple marker channels over a large number of cells within a tissue sample.

One major issue in multiplexed imaging data is the presence of systematic noise at a variety of levels, including batch and slide effects, imaging and clinical variables, and optical effects (Berry et al., 2021). This complexity and the within-slide dependence structure of the data can disrupt inference, while technical variability can be confounded through this complicated pipeline. In general, it is difficult to develop standardized pre-processing pipelines because of substantial variability in markers used across different studies and differences in target proteins across organs and cancer types (Schapiro et al., 2022). Furthermore, the field lacks a set of unified tools to perform many discrete components in the pipeline – visualization, cell segmentation, normalization, spatial analysis, and more. Below we will introduce the two main components of the multiplexed imaging pipeline that are relevant to this thesis – normalization and spatial analysis methods.

1.4 Normalization methods

Thematically, this dissertation began as an exploration of applying traditional spatial statistics methodology in multiplexed imaging data. However, we quickly discovered the presence of severe batch effects in many of the data sources discussed later that recalibrated our research interests. Broadly, normalization is a statistical technique used to adjust the input data values to improve data quality and remove systematic noise. These methods have been introduced widely in other fields of interest, particularly with regards to genetic sequencing data and neuroimaging. The first algorithm of interest in this thesis is the ComBat algorithm – this method is a location-scale model that uses empirical Bayes methods to adjust for batch effects in genetic micro-array data (Johnson et al., 2007). We are also interested in functional data analysis (FDA), which is a set of statistical methods designed for curves, densities, and functions of some domain. Curve registration is a non-parametric tool from the functional data field that non-linearly transforms the domain of the input data to align curves across some covariate (e.g., subject, time, site, etc.) (Ramsay and Silverman, 2005). Both the ComBat algorithm and FDA registration method are further discussed in our implementation of normalization algorithms in multiplexed imaging data in Chapters 2 and 3.

1.5 Spatial statistics

After developing approaches for normalization in multiplexed imaging data, we then adapt methods from the spatial statistics field – the methods in this thesis largely rely on point process theory (Illian et al., 2008). Broadly, point processes are stochastic models of point patterns, which are an irregular collection of points in some area or set. Point processes describe data at some set of finite points, where we implement statistical methods to understand the distribution of this set, spatial relationships between the points, and more. In multiplexed imaging data, we are interested in the spatial relationship of these points (e.g., cell locations) with the value of some marker channel at those cells. This means that the point processes we discuss are *marked* point processes – here we consider some random number of cells in a multiplexed image as the *points*, and some set of quantitative marker values as the *marks*. These definitions are important for summarizing point processes, which is the motivation for the summary measures discussed and introduced in Chapter 4 of this thesis.

CHAPTER 2

Quantifying and correcting slide-to-slide variation in multiplexed immunofluorescence images

2.1 Summary

Multiplexed imaging is a nascent single-cell assay with a complex data structure susceptible to technical variability that disrupts inference. These *in situ* methods are valuable in understanding cell-cell interactions, but few standardized processing steps or normalization techniques of multiplexed imaging data are available. We implement and compare data transformations and normalization algorithms in multiplexed imaging data. Our methods adapt the ComBat and functional data registration methods to remove slide effects in this domain, and we present an evaluation framework to compare the proposed approaches. We present clear slide-to-slide variation in the raw, unadjusted data, and show that many of the proposed normalization methods reduce this variation while preserving and improving the biological signal. Further, we find that dividing multiplexed imaging data by its slide mean, and the functional data registration methods, perform the best under our proposed evaluation framework. In summary, this approach provides a foundation for better data quality and evaluation criteria in multiplexed imaging.

2.2 Background

Single-cell assays are increasingly valued for their ability to provide information about the cell micro-environment and cell population interactions in healthy and cancerous tissues (Islam et al., 2020; McKinley et al., 2022; Shrubsole et al., 2008). Multiplexed imaging methods like multiplexed immunofluorescence (MxIF) (Gerdes et al., 2013), multiplexed immunohistochemistry (IHC) (Tsujikawa et al., 2017) and CODEX (Goltsev et al., 2018) are *in situ* analyses of multiple marker channels over a large number of cells within a given tissue sample. These methods build upon dissociative single cell analysis methods like flow cytometry (Bradford et al., 2004) and single-cell RNA sequencing (Chen et al., 2019) to allow scientists to better understand spatial cell-cell interactions in biological samples.

One significant issue in multiplexed imaging data is the presence of systematic noise at a variety of levels, related to batch and slide effects, imaging variables, and optical effects (Berry et al., 2021; Chang et al., 2020). A single experiment may contain hundreds of slides and terabytes of data across which a researcher seeks to make inference (Maric et al., 2021). However, this data complexity and the within-slide dependencies induce complex effects that can disrupt inference. This technical variability can be compounded through the complex image pre-processing pipeline and may contribute to biases that increase type 1 or type 2 error.

	None	**ComBat**	**Registration (fda)**
$\log_{10}$	$\log_{10}(y+1)$	$\text{ComBat}\left(\log_{10}(y+1)\right)$	$\text{fda}\left(\log_{10}(y+1)\right)$
Mean division	$\frac{y}{\mu_{ic}}$	$\text{ComBat}\left(\frac{y}{\mu_{ic}}\right)$	$\text{fda}\left(\frac{y}{\mu_{ic}}\right)$
Mean division $\log_{10}$	$\log_{10}(\frac{y}{\mu_{ic}}+\frac{1}{2})$	$\text{ComBat}\left(\log_{10}(\frac{y}{\mu_{ic}}+\frac{1}{2})\right)$	$\text{fda}\left(\log_{10}(\frac{y}{\mu_{ic}}+\frac{1}{2})\right)$

Table 2.1: **Summary of normalization procedures implemented.** Transformations (rows) and normalization (columns) performed on the data. Here y is the median cell intensity values for an arbitrary marker channel c, and μ_{ic} is the slide mean for slide i of the median cell intensity values for marker channel c.

Furthermore, it is difficult to develop a standardized pre-processing pipeline because of substantial variability in the markers used across different studies, as target proteins differ across organs and cancer types (Schapiro et al., 2022; Yapp et al., 2021). Image normalization is a technique used to adjust the input pixel- or image-level values of an image to remove noise and improve image quality. Due to the nascent development of multiplexed imaging, there are few established statistical tools that address challenges related to technical variation in this data set (Chang et al., 2020). Normalization methods may improve similarity across images by removing the unknown effect of technical variability. Moreover, statistical methods for batch correction and image normalization can be modified to fit this complex data structure to ultimately reduce systematic noise and improve statistical inference.

Extensive work has been done in other fields to adjust for batch effects and systematic noise, particularly with regards to neuroimaging and genetic sequencing data. One primary method employed in both of these fields is the ComBat method, introduced for genetic micro-array data (Johnson et al., 2007) and then adapted to neuroimaging in the analysis of magnetic resonance imaging (MRI) data (Fortin et al., 2017; Yu et al., 2018). The ComBat method is a location-scale model that implements an empirical Bayes algorithm to adjust for batch effects, and is robust to outliers in small sample sizes. Curve registration, a non-parametric tool from functional data analysis (FDA), has been used in recent work to adjust for systematic variability in accelerometry and MRI data (Marron et al., 2015; Wrobel et al., 2020, 2019). In the neuroimaging context, curve registration is used to normalize the imaging data by non-linearly transform the image intensity domain so that it is similar across images from different subjects, potentially collected on different scanners. Multiplexed imaging data are further complicated because it is non-negative, which other groups have remarked upon in similar imaging applications like spatial transcriptomics (Elosua-Bayes et al., 2021) – this requires unique derivations and/or applications of normalization methods to ensure no contradictions arise from negative marker intensities.

While adaptable, existing methods for normalizing data from other domains cannot be directly applied within multiplexed imaging due to the unusual format of the data (cell populations can differ substantially across samples), and the heavy skewness of the image histogram. The few algorithms adapted specifically for normalizing multiplex imaging data still could benefit from upstream normalization using algorithms adapted from other domains (Chang et al., 2020; Raza et al., 2016). For example, the RESTORE algorithm is a method developed for multiplexed imaging that uses negative control cells to remove unwanted variation across slides (Chang et al., 2020). However, this method relies on clustering mutually exclusive marker pairs using cell-level labels that are defined using unnormalized marker intensities and thus embed biases as detailed in this chapter. Raza et al also introduced normalization methods in the multiplexed imaging that implement a procedure of image filters and transformations (Raza et al., 2016). These methods show improvements at the pixel and image level, but do not correct for slide or batch effects that are prevalent as detailed in this work. Hence, the normalization methods proposed here can be applied early in the image processing pipeline to reduce bias in subsequent steps like phenotyping and spatial correlation analyses.

In this chapter we introduce and compare normalization and data transformation methods for multiplexed imaging data introduced previously in Harris et al. (2022b). These techniques combine transformations of the scale of the data from its raw form with algorithms (namely, ComBat and functional data registration) adapted to remove slide effects from the data. We further develop multiple novel metrics to quantify and measure the removal of technical variation in these data, where cell populations can differ across slides. We use data from the Human Tumor Atlas Network to evaluate the methods we compare here (Rozenblatt-Rosen et al., 2020; Chen et al., 2021). While we apply the methods here to segmented and quantified single-cell data from multiplexed imaging, they can also be applied at the pixel level.

2.3 Methods

2.3.1 Implementation

We compare three data transformations: $\log_{10}$, mean division (division by the slide-level mean), and mean division with $\log_{10}$, and three normalization procedures: no normalization, ComBat, and functional data registration, for a total of nine potential multiplex image normalization algorithms (Table 2.1).

2.3.1.1 Transformations

Let $Y_{ic}(u)$ denote the raw intensity of unit u on slide i for marker channel c (here u corresponds to segmented cell intensities). We consider the following transformations: the $\log_{10}$ transformation, $\log_{10}(Y_{ic}(u)+1)$, where the addition of 1 follows since $Y_{ic}(u)$ is integer-valued; the mean division transformation: $\frac{Y_{ic}(u)}{\mu_{ic}}$, where μ_{ic} is the mean intensity value on slide i for channel c; and the mean division $\log_{10}$ transformation, $\log_{10}\left(\frac{Y_{ic}(u)}{\mu_{ic}} + \frac{1}{2}\right)$, where again μ_{ic} is the mean intensity value on slide i for channel c. Here the data are no longer integer-valued, and the addition of $\frac{1}{2}$ ensures values greater than $\frac{1}{2}$ are positive and less than $\frac{1}{2}$ are negative to properly adjust this scale of data. Other transformations that are less relevant to this thesis were performed in the Supplement of Harris et al. (2022b) that under-performed the optimal methods discussed here and are available online with that publication.

2.3.1.2 ComBat normalization

We adapted the empirical Bayes framework of the ComBat algorithm (Fortin et al., 2017; Johnson et al., 2007) for multiplexed imaging data. We parameterize mean and variance of the slide-level batch effects, with the location-scale model

$$Y_{ic}(u) = \alpha_c + \gamma_{ic} + \delta_{ic}\varepsilon_{ic}(u),$$

where we define $Y_{ic}(u)$ as the intensity of unit u on slide i for marker channel c and α_c as the the grand mean of $Y_{ic}(u)$ for channel c. Though in principle, units can be at the pixel or cell level, in our application, $Y_{ic}(u)$ is the median cell intensity (or its transformed counterpart) of a selected marker for a given segmented cell on a specific slide in the dataset. Here γ_{ic} is the the mean batch effect of slide i for channel c and we assume $\gamma_{ic} \sim N(\gamma_c, \tau_c^2)$, δ_{ic}^2 is the variance batch effect of slide i for channel c and we assume $\delta_{ic}^2 \sim IG(\omega_c, \beta_c)$, and we assume the random errors $\varepsilon_{ic}(u) \sim N(0,1)$. We use the data to estimate $\hat{\alpha}_c$ and then estimate $\hat{\gamma}_{ic} = \frac{1}{U_{ic}}\sum_u Y_{ic}(u)$, or the sample mean intensity on slide i for channel c. We further define $\hat{\sigma}_c = \frac{1}{N}\sum_{ic}(Y_{ic}(u) - \hat{\alpha}_c - \hat{\gamma}_{ic})^2$ and let:

$$Z_{ic}(u) = \frac{Y_{ic}(u) - \hat{\alpha}_c}{\hat{\sigma}_c^2},$$

where we assume $Z_{ic}(u) \sim N(\gamma_{ic}, \delta_{ic}^2)$. Based on the posterior conditional means, we find the following empirical Bayes estimators of the two batch effect parameters (a detailed derivation of these estimators can be found in the Appendix):

$$\delta_{ic}^{2*} = \frac{\bar{\beta}_c + \frac{1}{2}\sum_u (Z_{ic}(u) - \gamma_{ic}^*)^2}{\frac{U_{ic}}{2} + \bar{\omega}_c - 1}, \gamma_{ic}^* = \frac{U_{ic} \cdot \bar{\tau}_c^2 \cdot \hat{\gamma}_{ic} + \delta_{ic}^{2*} \cdot \bar{\gamma}_c}{U_{ic} \cdot \bar{\tau}_c^2 + \delta_{ic}^{2*}}$$

Where we define U_{ic} as the number of quantified cells present on a particular slide i for a given channel c. We calculate the hyper-parameter estimates of $\bar{\beta}_c, \bar{\omega}_c, \bar{\tau}_c^2, \bar{\gamma}_c$ using the method of moments and iterate between estimating the hyper-parameters and batch effect parameters until convergence (Dempster et al., 1977; Johnson et al., 2007). Upon convergence, we use these batch effects to adjust the data,

$$Y_{ic}^{*}(u) = \frac{\hat{\sigma}_c^2}{\hat{\delta}_{ic}^{*}}(Z_{ic}(u) - \hat{\gamma}_{ic}^{*}) + \hat{\alpha}_c.$$

This model adjusts the Z-normalized intensity data, $Z_{ic}(u)$, by the mean and variance batch effects, and rescales back to the initial scale of the data with the mean and variance of the raw marker intensity values. Note that zeroes were left in the data prior to the ComBat normalization, since for each scale transformation we perform on the data the zeroes are meaningful rather than an absence of signal.

2.3.1.3 Functional data registration

For the second normalization algorithm we implemented functional data registration using the `fda` R package (Ramsay and Silverman, 2005; Ramsay et al., 2020). This approach uses functional data analysis (FDA) methods to approximate the histograms for each slide and channel as smooth densities, and uses functional registration to align the densities to their average at the slide-level. Functional registration is performed by estimating a monotonic warping function for each density that stretches and compresses the intensities such that densities are aligned. These warping functions are then used to transform the marker intensity values in the images so that non-biological variability is reduced across slides.

Here, let our observed cell intensity values $Y_{ic}(u)$ have density $Y_{ic}(u) \sim f(y \mid i,c)$. Our goal is to remove technical variation related to the slide by estimating a warping function, $\phi_{ic}(y)$, which is a monotonic transformation of the intensities. We first use a 21 degree of freedom cubic B-spline basis to approximate the densities of the median cell intensities for each slide and marker, $f(y \mid i,c) \approx \beta^T g(y)$ where $\beta \in \mathbb{R}^{21}$ is an unknown coefficient vector and $g(y)$ is a vector of known basis functions. We then register the approximated histograms to the average, restricting the warping function to be a 2 degree of freedom linear B-spline basis for some functions $h_1(y)$ and $h_2(y)$ and for constants C_0 and C_1 to be estimated from the data,

$$\phi_{ic}(x) = C_0 + C_1 \int_0^x \exp\{\beta_{1ic} h_1(y) + \beta_{2ic} h_2(y)\}\, dy,$$

such that the transformation is monotonic (Ramsay and Silverman, 2005). Unknown parameters β_{1ic} and β_{2ic} are estimated to minimize,

$$\int_y \|f_{ic}(\phi_{ic}(y)) - f(y)\|^2 dy$$

Where $f(y)$ is the average density across slides. We then use $\phi_{ic}(y)$ to calculate the normalized intensity values, $Y^*_{ic}(u)$:

$$Y^*_{ic}(u) = \phi_{ic}(Y_{ic}(u))$$

Note that the warping function $\phi_{ic}(y)$ is a map that takes in the raw median cell intensity value and outputs a new, normalized intensity value. Images are then normalized by taking the original intensity values in the image, and transforming them using the map defined by the warping function. This combined process can be summarized as first taking the raw data, smoothing the histogram of these data using a B-spline basis expansion, and then calculating a warping function to transform the smoothed data so that densities across slides within marker channel c are approximately aligned.

2.3.2 Evaluation framework

There is no accepted gold standard for evaluating normalization methods in multiplexed imaging because the same tissue sample cannot be imaged twice and there is substantial heterogeneity across samples (Nadarajan et al., 2019; Rozenblatt-Rosen et al., 2020). Here, our evaluation framework relies on the two following conditions to be deemed successful: (1) reduction in slide-to-slide variance in the cell intensity data and (2) preservation (and potential improvement) of existing biological signal in the data.

2.3.2.1 Alignment of marker densities

To determine if between-slide noise is visible when comparing densities, we visually inspect the changes in density curves for each transformation method. *A priori*, we expect that a successful transformation method will align the density curves across slides, and subsequently we inspect the placement of slide-level Otsu thresholds, a commonly used thresholding algorithm used in imaging analysis (Otsu, 1979), to confirm a reduction in variability between slides. To quantitatively measure the alignment of marker densities, we implement the *k*-sample Anderson-Darling statistic to quantify the likelihood that each slide is drawn from the same population (Scholz and Stephens, 1987). A higher value of this test statistic indicates greater evidence that the *k*-samples are drawn from different distributions.

2.3.2.2 Threshold discordance and accuracy

Otsu thresholding is a commonly used thresholding algorithm that defines an optimal threshold in gray-scale images and histograms, maximizing the between-class variance of pixel values to separate the data into two classes (Otsu, 1979). In this use case, we define Otsu thresholds at the slide-level for each of the markers in the study, where a cell with intensity value greater than the Otsu threshold is deemed marker positive. We then compare this to a global Otsu threshold, combining all slides, for each marker to calculate a mean

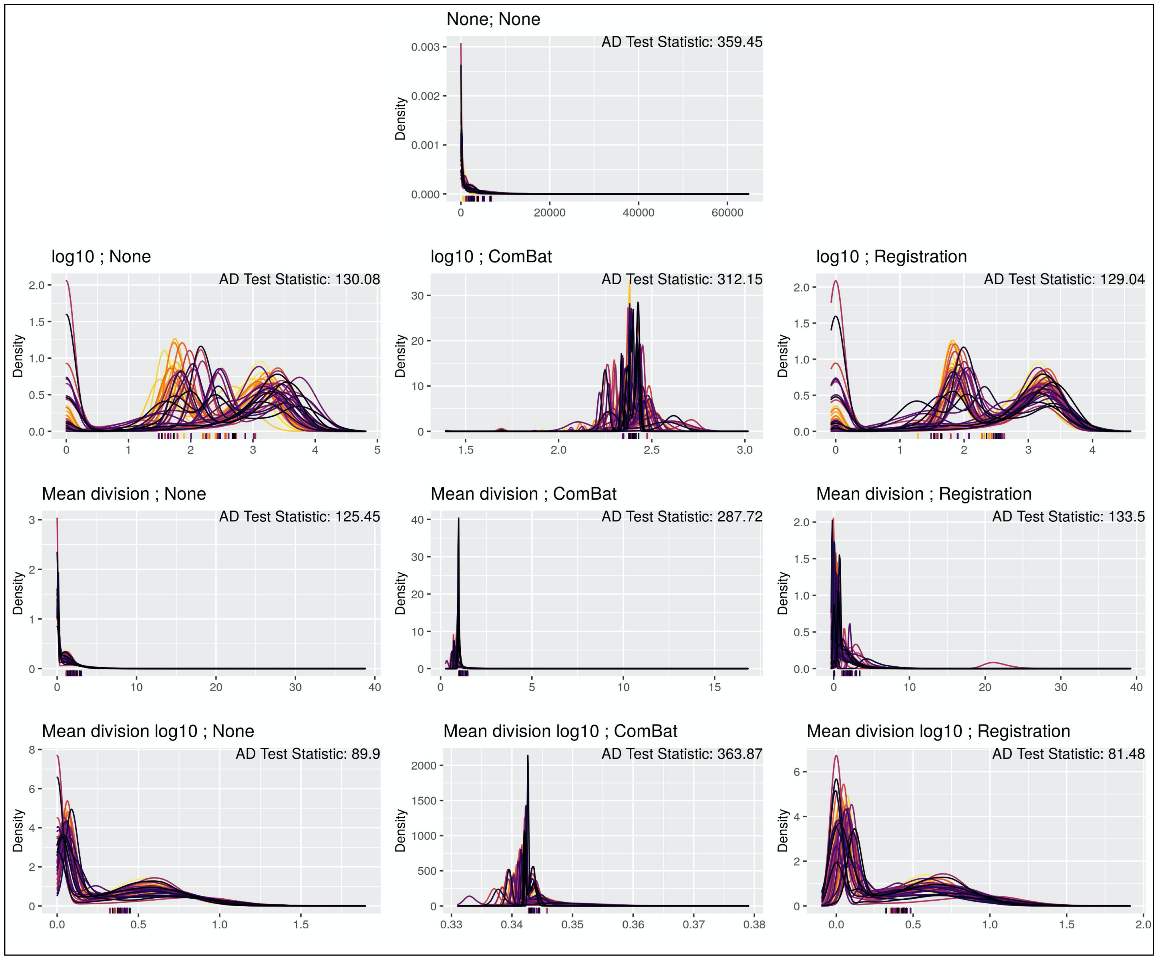

Figure 2.1: **Visual comparison of vimentin marker densities for each transformation method.** Density plots for the median cell intensity of the marker vimentin, where each color represents a different slide in the dataset. Each row is aligned with the scale transformations present in Table 2.1, where each column also matches with the normalization algorithms in Table 2.1. The ticks on the x-axis represent the Otsu thresholds for each slide for that transformed data, where the color again corresponds to the slide (such that the colors are one-to-one between threshold and density plot). Anderson-Darling test statistics for the marker vimentin are presented for each method in the top right corner.

discordance score across all slides for a given marker. For some marker channel c, slide i, and set of marker intensity values $Y_{ic}(u)$, define the indicator function for a given Otsu threshold o as $O_{ic}(u,o) = I(Y_{ic}(u) > o)$. Here, $O_{ic}(u,o)$ indicates which cells are in the expressed category using threshold o. The discordance metric is then defined as:

$$\frac{1}{N}\sum_{i}^{N}\left(\frac{\sum_{y} \mid O_{ic}(u,o_{ic}) - O_{c}(u,o_{c}) \mid}{U_{ic}}\right)$$

Where U_{ic} is the number of quantified cells present on a particular slide i for a given channel c, o_{ic} is the slide and channel specific Otsu threshold, and o_c is the threshold estimated across all slides for a given channel. Here we calculate a slide-level discordance score, e.g. the proportion of cells misclassified on each slide, and take an average of the score across slides for each marker channel. This measures the slide-to-slide discordance across all markers and transformation methods, to determine how similar Otsu thresholds are across slides following transformation. In this framework, a lower value of the threshold discordance score indicates better agreement across slides in the data.

2.3.2.3 Proportions of variance

To further assess the removal of slide related variance following each transformation of the data, we fit a random effects model using the `lme4` R package (Bates et al., 2015) with a random intercept for slide to assess what proportion of variance is present at the slide-level for each marker. A successful normalization algorithm will reduce the slide-level variance, ultimately removing technical variability to improve the quality of the data.

2.3.2.4 UMAP embedding

The Uniform Manifold Approximation and Projection (UMAP) is a technique for dimension reduction (McInnes et al., 2018) commonly used in the biological sciences to distinguish differences in cell populations between single-cell data (Becht et al., 2019). Here we reduce the data into two UMAP embeddings for each of the transformation methods using only four markers in the dataset: vimentin, collagen, pan-cytokeratin, and Na^{+}/K^{+}-ATPase. These markers were chosen for their ability to easily distinguish epithelial and stromal cells. We expect the UMAP embeddings to yield clear separation of the data when using the epithelium label in our dataset (see the Dataset Section). To quantify this separation of groups, we implement a k-means clustering model on the UMAP embeddings to predict the class label, and use the adjusted Rand index to measure the similarity with the true labels (Hartigan and Wong, 1979; Hubert and Arabie, 1985). Larger values of this index indicate better agreement between two sets of labels, adjusted for the chance grouping of elements. Note that across each slide in the dataset, approximately 10% of the data was used to derive the UMAP embeddings to reduce computational and visualization time.

Method	**Mean AD Test Statistic**	**Mean Otsu Discordance Score**	**Adj. Rand Index (Slide ID)**	**Mean Variance Proportion (Slide ID)**
None; None	275.019	0.085	0.033	0.138
$\log_{10}$; None	225.413	0.134	0.083	0.301
$\log_{10}$; ComBat	291.900	0.138	0.089	0.000
$\log_{10}$; Registration	217.649	0.110	0.037	0.232
Mean division ; None	138.774	0.041	0.007	0.000
Mean division ; ComBat	247.612	0.109	0.064	0.000
Mean division ; Registration	174.933	0.164	0.120	0.333
Mean division $\log_{10}$; None	114.653	0.055	0.010	0.046
Mean division $\log_{10}$; ComBat	321.810	0.132	0.071	0.000
Mean division $\log_{10}$; Registration	104.330	0.049	0.018	0.081

Table 2.2: **Quantitative metrics comparing normalization methods.** Results from the *k*-samples Anderson-Darling test statistic, the threshold discordance score, and the variance proportion at the slide level from the random effects modeling, all averaged across marker channels, as well as the adjusted Rand index for the slide identifiers comparing the raw data to the normalized data. For each of these metrics, small values indicate better performance for a given method.

2.3.3 Dataset

The data were collected from human colorectal cancer tissue samples from the Human Tumor Atlas Network (Rozenblatt-Rosen et al., 2020; Chen et al., 2021). The final dataset comprises over 2.2 million cells in the MxIF modality across over 2400 images on 43 different slides, with single-cell segmentation performed using an algorithm developed in-house (McKinley et al., 2022). Cell intensities for each marker were quantified as the median pixel value within the segmented cell, with tissue samples stained for 33 different marker channels. For the purpose of evaluating the algorithms compared in the chapter, we restricted our attention to the following markers: beta catenin (BCATENIN), CD3D (CD3), CD8 (CD8), collagen (COLLAGEN), Na^{+}/K^{+}-ATPase (NAKATPASE), olfactomedin 4 (OLFM4), pan-cytokeratin (PANCK), SRY-Box 9 (SOX9), vimentin (VIMENTIN). These markers were chosen because of their ability to distinguish between epithelial and stromal cells, PANCK, COLLAGEN, NAKATPASE, VIMENTIN (Blom et al., 2017; Ijsselsteijn et al., 2019); as immune markers, CD3, CD8 (Galon et al., 2006); as stem cell markers, OLFM4, SOX9 (Van der Flier et al., 2009; Scott et al., 2010); and as implicated in colon cancer, BCATENIN, (Shang et al., 2017).

We used epithelial and stromal cell labels and manually labeled marker positive cells as biological variables in order to quantify loss or improvement of biological signal due to each normalization method. The epithelial labels were created for each slide at the image level using a random forest trained on all of the markers included in the dataset. A cell was labeled as being in a particular cell class if that was the most likely class probability within the segmented cell area. We defined marker positive cells by first manually thresholding the immune marker images to create marker positive image masks. Then, for each segmented cell, the cell was defined as marker positive if more that 30% of its area contained marker-positive pixels. We refer to these as manual labels for CD3 and CD8. We also used a tumor image mask to denote whether a cell is in a tumor-containing region.

2.4 Results

2.4.1 Removal of slide-to-slide variation

2.4.1.1 Alignment of marker densities

Density curves of the marker vimentin for each transformation algorithm and corresponding slide-level Otsu thresholds, along with test statistics from the *k*-sample Anderson-Darling test were compared to determine alignment of curves across slides after transformation (Figure 2.1, Table 2.2). Beginning with the unnormalized transformed values, the $\log_{10}$ transformation produces density curves that are somewhat well-aligned (AD Test: 130.08), while the mean division and mean division $\log_{10}$ methods both compress the scale of the data and align well across slides (AD Test: 125.45, 89.90). Furthermore, each ComBat method performs poorly at aligning and reducing noise in the data, yielding the largest statistics from the Anderson-Darling test and visually noisy density curves. This is likely due to the Gaussian assumptions of the ComBat model that are not met in either the bi-modal ($\log_{10}$, mean division $\log_{10}$) or right-skewed (mean division) methods. The functional data registration aligns the $\log_{10}$ and mean division $\log_{10}$ well, and the algorithm yields marginal improvements for some of these transformations.

The best performing methods for this metric are the mean division, mean division $\log_{10}$, and mean division $\log_{10}$ combined with the functional data registration algorithm: the data is well-aligned across slides and when averaging Anderson-Darling statistics across all marker channels (Table 2.2), we see these methods yield the lowest values presenting stronger evidence these values are derived from the same parent distribution.

2.4.1.2 Threshold discordance score

In order to quantify how the normalization methods impact cell classification, we compared Otsu thresholding estimated at the slide level and across slides for each method to generate a discordance score and compare this to raw data (Figure 2.2A). Compared to the epithelium/stromal markers in the dataset, less identifiable markers like CD3 and CD8 yield the worst performance across nearly all methods, with large increases in the discordance score. Most methods increase the mean discordance score relative to the unadjusted data, with the exception of the mean division, mean division $\log_{10}$, and the mean division $\log_{10}$ with functional data registration. This evaluation again aligns with earlier assessments and suggests that these methods present improvements in the slide-to-slide agreement across all markers compared to the unadjusted data. We also observe that when comparing threshold discordance scores across all markers, these three methods yield the lowest values, and are the only methods to reduce this rate relative to the raw data (Table 2.2).

2.4.1.3 Proportions of variance

To understand how well each method removes slide-related variability, we fit a random effects model on the median cell intensities after applying each combination of transformation and normalization. The ComBat algorithm, by design, removed all of the variability related to slide across all methods (Figure 2.3, Table 2.2). The only other method that entirely removes all slide-to-slide variance across all marker channels is the mean division method – for the mean division $\log_{10}$ and mean division $\log_{10}$ with functional data registration methods, we also observe reduction in variance (though not completely removed) relative to the unnormalized data. And while ComBat reduces slide variability, it completely removes slide effects that may include biological differences. In short, the results of this metric suggest the utility of the mean division

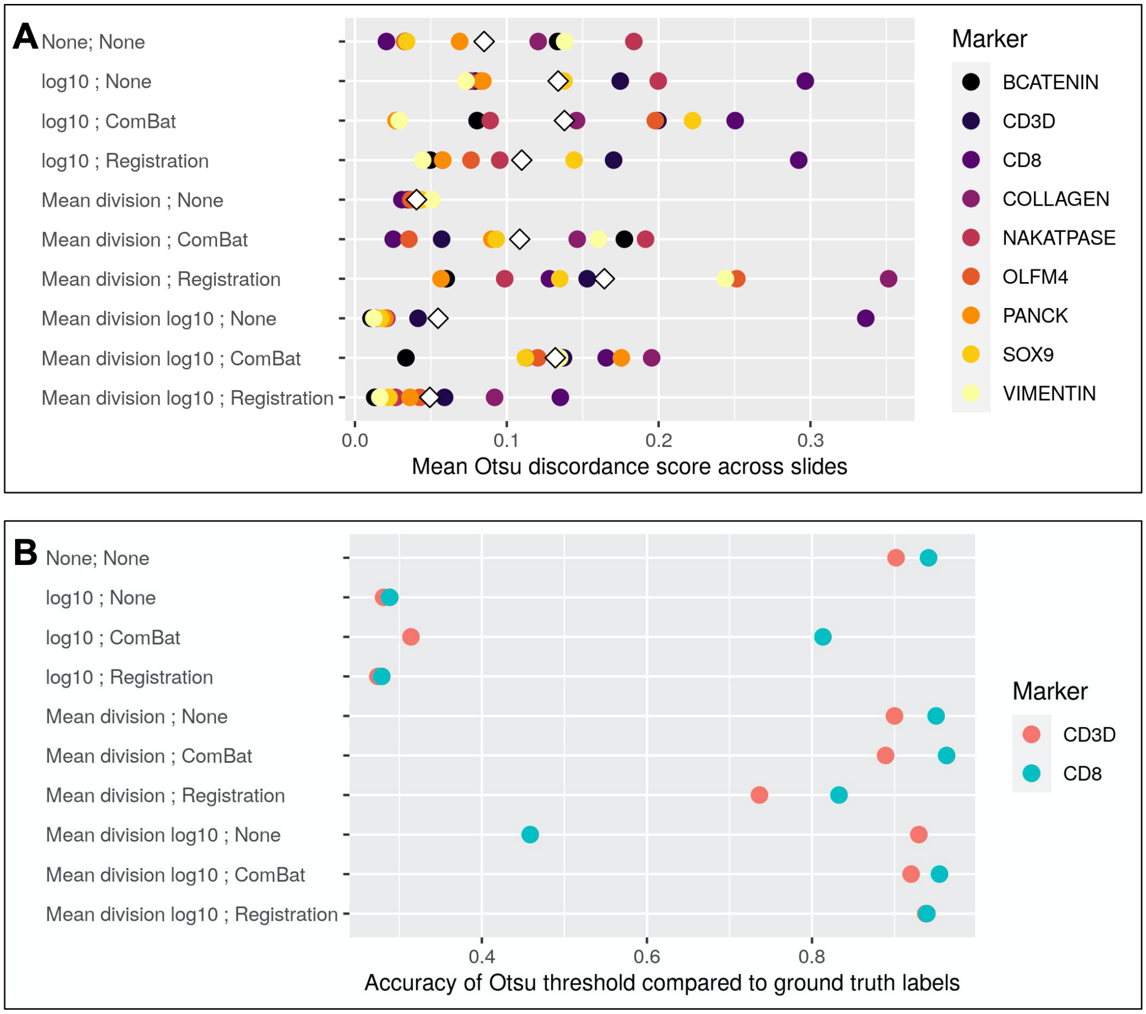

Figure 2.2: **Threshold discordance & accuracy.** (A) Otsu thresholds were calculated at the slide-level for each marker and compared to a global Otsu threshold for each marker to calculate a discordance score to compare transformation methods. The mean difference of the slide-level Otsu thresholds and the global Otsu threshold is then calculated for each marker, and presented as a point for each of the 9 markers, with the white diamond representing the mean discordance score across all markers for a given method. **Given that this is a discordance score, lower values indicate better agreement across slides.** (B) Otsu thresholds were calculated across slides for each marker to determine marker positive cells, which were then compared to the manual labels for the markers CD3 and CD8 to determine the accuracy of defining a cell as marker positive. This is presented as the accuracy rate of recapitulating the ground truth labels - **given that this is a measurement of accuracy, higher values indicate better agreement between the normalized data and labels.** Note also that for each of these plots, the top row indicates the results from the raw, unadjusted data.

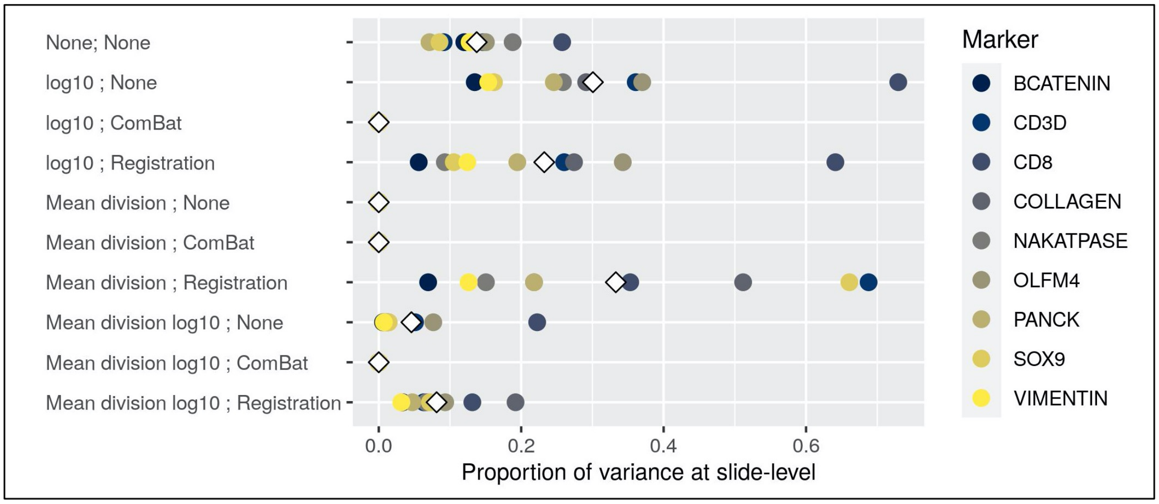

Figure 2.3: **Proportion of variance present at slide-level in random effects model.** Scatter plots that denote the proportion of variance at the slide-level for each normalization method for each of the marker channels in this dataset. Variance proportions were calculated using a random effects model with a random intercept for slide – methods that perform well should reduce the slide level variance. Note also that the top row indicates the results from the raw, unadjusted data.

methods in removing slide-level variance across marker channels.

2.4.2 Preservation of existing biological signal

2.4.2.1 Marker-positive accuracy using Otsu thresholds

We further utilized Otsu thresholding to identify marker positive cells and compared these to the manual labels for CD3 and CD8 to determine which normalization methods most accurately recapitulate the raw data (Figure 2.2B). Results suggest that the scale of the data is pivotal in whether a method maintains marker-positive accuracy, with each of the methods on the $\log_{10}$ scale demonstrating dramatic reductions in marker-positive accuracy compared to the raw data, while the mean division method performs the best across all methods. The methods that have performed well in the aforementioned evaluation metrics perform well here, namely the mean division method and the mean division $\log_{10}$ with functional data registration. This continues to suggest these methods reduce the slide-to-slide variation present in the data while accurately capturing marker-positive cells after transformation.

2.4.2.2 UMAP embedding

We compared UMAP embeddings of four related markers across normalization methods to compare the separation of epithelium and stromal tissue labels. In the raw data, the embeddings separate well (Adj. Rand Index: 0.82), however the data includes the presence of outliers that suggest mixing of the tissue classes in the UMAP embedding space (Figure 2.4A). Nearly all methods implemented improve upon the separation of groups based on the adjusted Rand index, yet many of these methods present co-localization that does not clearly depict separation as desired. We do observe distinct separation of the aforementioned methods of interest: mean division (Adj. Rand Index: 0.94), mean division $\log_{10}$ (Adj. Rand Index: 0.95), and the mean division $\log_{10}$ with functional data registration (Adj. Rand Index: 0.97) - each of these UMAP embeddings presents distinct groups that suggests these methods are improving the separation of these two tissue classes.

We also compared the distribution of the unique slide identifiers in the UMAP embeddings of these four markers, which in the raw data (Adj. Rand Index: 0.033) points to specific slide co-localization in the data (Figure 2.4B, Table 2.2). In this case, we desire low values of the adjusted Rand index, which suggest poor prediction of slide labels and indicate the removal of slide-level variance. Many of the methods, particularly those implementing the ComBat algorithm, worsen the distribution of these slide identifiers and increase the adjusted Rand index, suggesting additional slide-to-slide noise added to the data. This suggests that ComBat removes both biological signal and slide-to-slide effects that are exaggerated in the UMAP embedding space. In contrast, there is reduced slide-to-slide clustering in the UMAP embeddings for each of the following methods: mean division (Adj. Rand Index: 0.01), mean division $\log_{10}$ (Adj. Rand Index: 0.01), and mean division $\log_{10}$ with functional data registration (Adj. Rand Index: 0.02). These methods appear to both reduce the observed slide-to-slide variation noted here and in the aforementioned results, while maintaining the necessary biological signal of interest.

2.5 Discussion

In this chapter, we derived the ComBat algorithm for a new modality and employed a novel use of functional data registration to align histograms of multiplexed imaging data. In the absence of a gold standard for comparison in multiplexed imaging data, validating any normalization procedure is challenging. The suggested evaluation framework introduced here can be used to assess the presence and reduction of slide effects in multiplexed imaging data, which we implemented to evaluate 9 combinations of transformations and normalization methods. Further, our framework can be applied in the absence of a ground truth by quantifying the amount of slide related variability and comparing to manually labeled biological features, providing a foundation for further development of evaluation criteria in the multiplexed domain. Also note that since the proposed methods are applied within a given marker channel, this work can be extended into other imaging domains like IHC that do not involve multiplexing.

Similarly, the use of Otsu thresholding in this chapter is the standard procedure for imaging domains like IHC (Tsujikawa et al., 2019; Trinh et al., 2017). However, markers like the phosphorylated epidermal growth

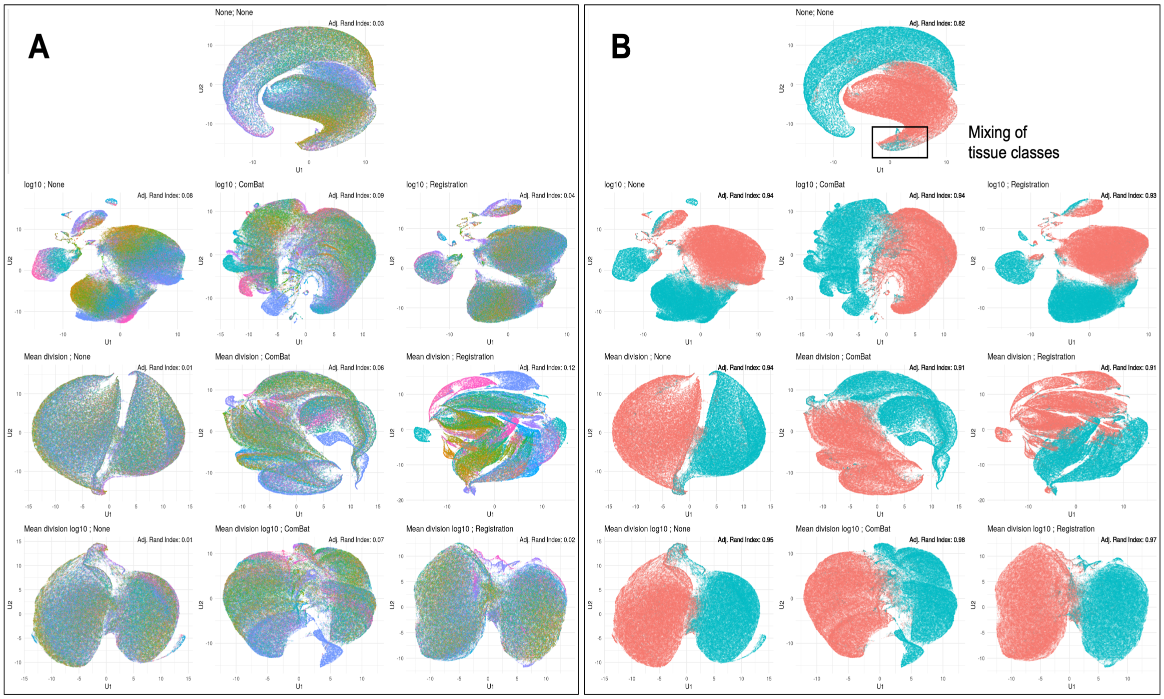

Figure 2.4: **UMAP embedding of data for each transformation method.** UMAP embedding of the transformed data with points colored by slide identifier (A) and tissue type (B). The rectangle in (B) denotes the mixing of tissue classes present in the raw, unadjusted data UMAP embedding. Adjusted Rand index values for each embedding are presented in the top right corner.

factor receptor (p-EGFR) are typically categorized into multiple groups based on staining intensity (Hashmi et al., 2018; Shan et al., 2017). While the Otsu threshold may not capture this categorization, it remains a reasonable proxy for these quantitative markers in the absence of a pathologist, and other metrics implemented here like the Anderson-Darling statistic may be more appropriate. Furthermore, future methods development could focus on implementing multi-Otsu thresholding methods into the threshold discordance score, or adapt marker-specific thresholding methods that better capture variability in the quantitative markers. Notably, the correspondence between a marker positive cell defined by an Otsu threshold and biological signal is not necessarily one-to-one. For example, the $\log_{10}$ transformation non-linearly compresses the domain, such that a larger proportion of the x-axis is allotted to cells that are marker negative (background and unexpressed cells), which may have led to greater variability in the Otsu thresholds.

We find that the raw data scale has clear slide-to-slide variation present, and that normalization methods can reduce slide level variation while preserving and improving biological signal relative to the raw, unadjusted data. These findings suggest that the mean division transformation method reduces slide variability and improves the biological signal. In addition, the mean division $\log_{10}$ scale (unnormalized) performs well across all evaluation metrics, with the noted exclusion of results for the marker CD8. This discrepancy is remedied with the functional data registration, which is a limitation of the mean division $\log_{10}$ transformation but points to the robustness of the registration algorithm to maintain and improve the quality of the data.

However, note that the registration algorithm does not perform well with skewed data, suggesting that improvements we see in data that appears bi-modal (e.g., better suited to the non-parametric assumption of functional data) is not necessarily transferable to right-skewed data that violates assumptions of smoothness in the B-spline basis – future work could explore this result. The ComBat method performs adequately, but appears to over normalize the data and relies heavily on a Gaussian assumption that is violated in this skewed-right dataset. The clear limitation of this normalization method and others is that when applied to whole tissue slides, any between slide variability is confounded with biological variability. Recent adaptations of ComBat like ComBat-seq for RNA-seq data may provide a better framework to implement in the multiplexed imaging space (Zhang et al., 2020), including future work that could address how the algorithm handles zeroes. Note also that recent advances applying deep learning in fluorescence microscopy analysis combine information across heterogeneous combinations of markers to ameliorate similar problems that we address in this chapter, namely technical variation and comparing disparate data sources (Gomariz et al., 2021) – this could be a valuable avenue for future normalization approaches.

In practice, the mean division method is “good enough” – it is simple, computationally efficient, and appears the least likely to introduce error while still reducing slide-to-slide variation and maintaining biological signal. The mean division $\log_{10}$ method may be necessary in the case of statistical modeling, since skewed distributions are not suitable for many statistical models, but may not be the best way to represent cell intensities as a predictor variable (as appears the case for the mean division method). We see that in the case of mean division $\log_{10}$ data, it may be necessary to use the registration algorithm to remedy discrepancies like those visible for the marker CD8.

CHAPTER 3

mxnorm: An R package to normalize multiplexed imaging data

3.1 Summary

As multiplexed imaging research develops at a rapid pace, there is a growing necessity for computational implementations of cutting-edge methods in the multiplexed imaging pipeline. Here we extend our previous work applying normalization methods in this data type and introduce the `mxnorm` R package, which provides two key services: (1) a collection of normalization methods and analysis metrics to implement and compare normalization in multiplexed imaging data, and (2) a foundation for storing multiplexed imaging data in R using S3. We adapt both previously introduced normalization algorithms and analysis methods like the Otsu discordance metric, and further introduce options for users to provide user-defined normalization algorithms. This allows users the ability to leverage our robust evaluation framework of normalization efficacy and develop optimal normalization frameworks (and analysis pipelines) in multiplexed imaging data, ultimately setting a foundation for evaluating these normalization methods in the field.

3.2 Package Overview

3.2.1 Background

Multiplexed imaging is an emerging single-cell assay that can be used to understand and analyze complex processes in tissue-based cancers, autoimmune disorders, and more. These imaging technologies, which include co-detection by indexing (CODEX), multiplexed ion beam imaging (MIBI), and multiplexed immunofluorescence imaging (MxIF), provide detailed information about spatial interactions between cells (Goltsev et al., 2018; Angelo et al., 2014; Gerdes et al., 2013). Multiplexed imaging experiments generate data across hundreds of slides and images, often resulting in terabytes of complex data to analyze through imaging analysis pipelines. Methods are rapidly developing to improve particular parts of the pipeline, including software packages in R and Python like `spatialTime`, `imcRtools`, `MCMICRO`, and `Squidpy` (Creed et al., 2021; Windhager et al., 2021; Schapiro et al., 2022; Palla et al., 2022). An important, but understudied component of this pipeline is the analysis of technical variation within this complex data source – intensity normalization is one way to remove this technical variability. The combination of disparate pre-processing pipelines, imaging variables, optical effects, and within-slide dependencies create batch and slide effects that can be reduced via normalization methods. Current state-of-the-art methods vary heavily across research labs and image acquisition platforms, without one singular method that is uniformly robust – optimal statistical methods seek to improve similarity across images and slides by removing this technical variability while maintaining the underlying biological signal in the data.

mxnorm is open-source software built with R and S3 methods that implements, evaluates, and visualizes normalization techniques for multiplexed imaging data. Extending methodology described in and Chapter 2 and Harris et al. (2022b), we intend to set a foundation for the evaluation of multiplexed imaging normalization methods in R. This easily allows users to extend normalization methods into the field, and provides a robust evaluation framework to measure both technical variability and the efficacy of various normalization methods. One key component of the R package is the ability to supply user-defined normalization methods and thresholding algorithms to assess normalization in multiplexed imaging data. This chapter builds upon previously published core features, usage details, and extensive tutorials from Harris et al. (2022a), the package documentation and vignette in the software repository and on CRAN (Harris, 2022).

3.2.2 Motivation

Multiplexed imaging measures intensities of dozens of antibody and protein markers at the single-cell level while preserving cell spatial coordinates. This allows single-cell analyses to be performed on biological samples like tissues and tumors, much like single-cell RNA sequencing, with the added benefit of *in situ* coordinates to better capture spatial interactions between individual cells (McKinley et al., 2022; Chen et al., 2021). Current research using platforms like MxIF and MIBI demonstrate this growing field that seeks to better understand cell-cell populations in cancer, pre-cancer, and various biological research contexts (Ptacek et al., 2020; Gerdes et al., 2013).

In contrast to the field of sequencing & micro-array data and the established software, analysis, and methods therein, multiplexed imaging lacks established analysis standards, pipelines, and methods. Recent developments in multiplexed imaging seek to address the broad lack of standardized tools – the MCMICRO pipeline seeks to provide a set of open-source, reproducible analyses to transform whole-slide images into single-cell data (Schapiro et al., 2022). Researchers in the field have also developed a ground truth dataset to evaluate differences in batch effects and normalization methods (Graf et al., 2022), while other open issues in the field that may produce open-source solutions include tissue segmentation, end-to-end image processing, and removal of image artifacts. With this diversity of open issues in multiplexed imaging, our work focuses specifically on normalization methods and evaluating these results in multiplexed imaging data. Namely, standard normalization software in the sequencing field includes open-source packages in R and Python like sva, limma, and Scanorama (Leek et al., 2012; Smyth, 2005; Hie et al., 2019), but an analogue for evaluating and developing normalization methods does not exist for multiplexed imaging data.

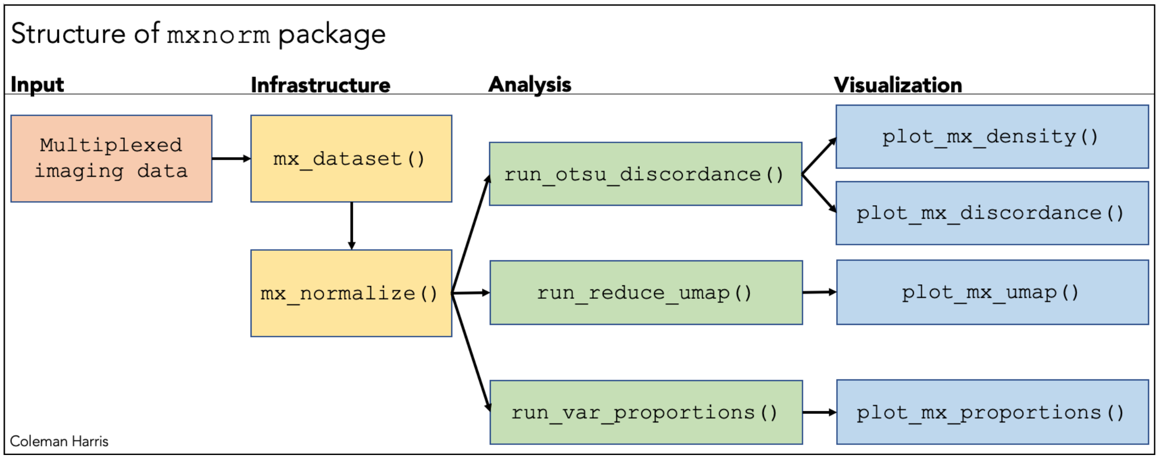

Figure 3.1: **`mxnorm` Package Structure:** Diagram demonstrating the basic structure of the `mxnorm` package and associated functions included in the software.

We recently proposed and evaluated several normalization methods for multiplexed imaging data, which along with other recent work shows that normalization methods are important in reducing slide-to-slide variation (Chang et al., 2020; Burlingame et al., 2021; Harris et al., 2022b). These recently developed algorithms are the beginning of contributions to normalization literature, but lack a simple, user-friendly implementation. Further, there is no software researchers can use to develop and evaluate normalization methods in their own multiplexed imaging data; multiplexed imaging software is limited mostly to Matlab, Python, and only a scattered few R packages exist. Two prominent packages, `cytomapper` and `giotto`, contain open-source implementations for analysis and visualization of highly multiplexed images (Eling et al., 2020; Dries et al., 2021b), but do not explicitly address normalization of the single-cell intensity data. Hence, there is a major lack of available tools for researchers to explore, evaluate, and analyze normalization methods in multiplexed imaging data. The `mxnorm` package provides this framework, with easy-to-implement and customizable normalization methods along with a foundation for evaluating their utility in the multiplexed imaging field.

3.2.3 Functionality

As shown in Figure 3.1, there are three main types of functions implemented in the `mxnorm` package – infrastructure, analysis, and visualization. The first infrastructure function, `mx_dataset()`, specifies and creates the S3 object used throughout the analysis, while the `mx_normalize()` function provides a routine to normalize the multiplexed imaging data, which specifically allows for normalization algorithms defined by the user. Each of the three analysis functions provides methods to run specific analyses that test for slide-to-slide variation and preservation of biological signal for the normalized and unnormalized data, while the four visualization functions provide methods to generate `ggplot2` plots to assess the results. We also extend the `summary()` generic function to the `mx_dataset` S3 object to provide further statistics and summaries.

The statistical methodology underlying the methods we implemented in `mxnorm` builds upon existing work in both R and Python. Normalization algorithms available in `mx_normalize()` leverage methodology derived in the ComBat paper, the `fda` package, and the `tidyverse` framework (Johnson et al., 2007; Ramsay et al., 2020; Wickham et al., 2019). The threshold discordance methods available in `run_otsu_discordance()` leverage methodology from Otsu's original paper and the `scikit-image` implementation of Otsu thresholding in Python (Otsu, 1979; van der Walt et al., 2014). Our implementation of the UMAP algorithm in `run_reduce_umap()` leverages both the UMAP paper and the `uwot` implementation of the UMAP algorithm in R (McInnes et al., 2018; Melville, 2021). The random effects modeling options available in `run_var_proportions()` leverage the `lme4` R package (Bates et al., 2015).

3.3 Basic Example

In general, we expect multiplexed imaging data in a `data.frame` format that includes columns for slide & image identifiers, separate columns for marker intensity values, and some set of metadata columns like tissue identifiers, phenotypic traits, medical conditions, etc. Alongside the `mxnorm` package itself, we introduce the `mx_sample` dataset that demonstrates the expected structure of multiplexed imaging data, and provides simulated marker intensity values that demonstrate strong slide effects. The first 3 rows of this dataset appear as follows:

```
#>   slide_id image_id marker1_vals marker2_vals marker3_vals metadata1_vals
#> 1   slide1   image1           15           17           28            yes
#> 2   slide1   image1           11           22           31             no
#> 3   slide1   image1           12           16           22            yes
```

This dataset consists of 3 markers across 4 slides (with 750 ”cells” on each slide) and 1 metadata column, and was specifically created to demonstrate the effect of normalization methods in multiplexed imaging data. To ensure a streamlined framework for the analysis of this type of data, we have created an S3 object `mx_dataset` to store the data and continue building upon as we normalize the data and analyze our results.

3.3.1 Creating the S3 object

3.3.1.1 Using **`mxnorm::mx_dataset()`**

Let's load the `mx_sample` dataset into the S3 object we'll use for our analyses, the `mx_dataset` object. Here we specify the following parameters,

- `data`: the input dataset in a `data.frame` format
- `slide_id` and `image_id`: the identifiers of interest in our dataset
- `marker_cols`: the set of marker columns in our input data that we want to include
- `metadata_cols`: metadata columns in our input data that we want to include

Now we make the following call:

```
mx_data = mx_dataset(data=mx_sample,
                     slide_id="slide_id",
                     image_id="image_id",
                     marker_cols=c("marker1_vals",
                                   "marker2_vals",
                                   "marker3_vals"),
                     metadata_cols=c("metadata1_vals"))
```

And now the `mx_dataset` S3 object becomes the foundation for each of the methods and analyses we have implemented in `mxnorm`. After we create this S3 object, we can then run the normalization, analysis, and visualize our results using the other exposed functions in `mxnorm`. First, we must run the normalization of the data itself via the `mx_normalize()` method. Here, we leverage the S3 structure of the `mx_dataset` object to build upon and add attributes to keep our analysis in one consistent object.

3.3.2 Normalization of multiplexed imaging data

3.3.2.1 Using **`mxnorm::mx_normalize()`**

Now that we've created the object, we can use the `mx_normalize()` function to normalize the imaging data. Here we specify:

- `mx_data`: the `mx_dataset` object with the data we want to normalize
- `transform`: the transformation method we want to perform, which in this case is `mean_divide`.
- `method`: the normalization method we want to implement, which in this case is `None`.

- `method_override`: an optional parameter to provide a user-defined normalization method (see the details below for an example)
- `method_override_name`: an optional parameter to re-name the `method` attribute when specifying user-defined normalization

Now we make the following call:

```
mx_data = mx_normalize(mx_data = mx_data,
                       transform = "mean_divide",
                       method="None",
                       method_override=NULL,
                       method_override_name=NULL)
```

The `mx_dataset` object now has normalized data in the following form in the `norm_data` attribute, with additional `transform` and `method` attributes added to the `mx_dataset` object as well:

```
#>   slide_id image_id marker1_vals marker2_vals marker3_vals metadata1_vals
#> 1   slide1   image1    0.6293173    0.4091531    0.5264357            yes
#> 2   slide1   image1    0.3063198    0.6621725    0.6367893             no
#> 3   slide1   image1    0.3870692    0.3585492    0.3057285            yes
```

Note that there are multiple normalization approaches implemented into the `mxnorm` package (including user-defined normalization) – namely, the ComBat algorithm and an adaptation of functional data analysis using the `fda` package.

3.3.2.2 Implementation of ComBat

The original ComBat algorithm is implemented in the Surrogate Variable Analysis (`sva`) Bioconductor package, which is a popular and well-maintained package "for removing batch effects and other unwanted variation in high-throughput experiment" (Leek et al., 2012). The ComBat function is well-documented and versatile for correcting batch effects using the method introduced originally in microarray data via `sva::ComBat()`, however, the assumptions made for this function are based largely on the expression matrices produced in microarray studies, not those typical to imaging or multiplexed studies.

Efforts to extend into the neouroimaging space provide a good foundation for adapting the ComBat algorithm to alternate modalities (Fortin et al., 2017), which inspired our extension into the multiplexed domain. Our implementation here is similar to that adapted by Fortin et al in the `neuroCombat` package, but is focused largely on datasets typical in the multiplexed imaging field.

There are a handful of distinctions to discuss regarding the ComBat implementation in `mxnorm`. As noted previously, we expect multiplexed imaging data to be marker-dependent and in the "long" format. This means that for some set of multiplexed *n* slides and *m* images, we don't expect a perfect $n \times m$ expression matrix for a given marker channel. We can also take advantage of working with "long" data to leverage `tidyverse` packages & functions like `dplyr` for easier/faster calculation of batch effects – this algorithm is detailed in the `/R/combat_helpers.R` file in the software repository. Ultimately, we then take the same approach with running the ComBat algorithm – initialize values of our parameters of interest, run the algorithm to calculate batch effects using empirical Bayes, and then standardize the data to correct for slide-to-slide variation.

3.3.2.3 Implementation of functional data registration

While the `fda` package is the basis of much functional data analysis in R (and the basis of the analyses performed in `mxnorm`), there are a handful of other implementations/extensions of this field that are relevant to the `mxnorm` package both for underlying methods and better understanding of functional data. Extensions of the FDA paradigm in R include the `refund` package (Goldsmith et al., 2021), which includes methods for regression of functional data and similar applications to imaging data, and the `registr` package (Wrobel et al., 2021) that focuses on the registration of functional data generated from exponential families. There are also similar extensions of registration algorithms like the `mica` package (Wrobel, 2021), which seeks to apply FDA registration algorithms to the harmonization of multi-site neuroimaging data.

Again as noted previously, we expect multiplexed imaging data to be marker-dependent and in the "long" format – hence we run the registration algorithm across slides for a given marker. Here we are using the `fda` package to setup the basis functions, run initial registration, generate the inverse warping functions, and then register the raw data to the mean registered curve to create a normalized intensity value. This process and the extensive hyper-parameters available are detailed in the `/R/registration_helpers.R` file in the software repository.

3.3.3 Otsu discordance scores

3.3.3.1 Using `mxnorm::run_otsu_discordance()`

Now that we've normalized the multiplexed imaging data, we can start to analyze the results and understand the performance of our normalization. Using the above normalized data, we can run an Otsu discordance score analysis to determine how well our normalization method performs. Broadly, this method calculates the distance of slide-level Otsu thresholds from the ”global” Otsu threshold for a given marker channel to quantify the slide-to-slide alignment of values via a summary metric (this is discussed in depth in Chapter 2). In this analysis, we look for lower discordance scores to distinguish better performing normalization methods, which indicates better agreement between slides for a given marker. To run this analysis we specify:

- `mx_data`: the `mx_dataset` object with the data we want to analyze
- `table`: the set of data we want to analyze using Otsu discordance, either `raw`, `normalized`, or `both`
- `threshold_override`: an optional parameter to provide a user-defined thresholding method (see the details below for example)
- `plot_out`: an optional parameter to output plots when running Otsu discordance

And to run this method we use:

```
mx_data = run_otsu_discordance(mx_data,
                        table="both",
                        threshold_override = thold_override,
                        plot_out = FALSE)
```

This method adds an `otsu_data` table to the `mx_dataset` object that contains the results of the discordance analysis, with an additional attribute `threshold` to denote the type of thresholding algorithm used and the `otsu_table` to denote which tables in our object we ran the analysis on:

```
#>   slide_id       marker table slide_threshold marker_threshold
#> 1   slide1 marker1_vals   raw        12.01758         54.89844
#> 2   slide2 marker1_vals   raw        20.01367         54.89844
#> 3   slide3 marker1_vals   raw        87.05664         54.89844
#>   discordance_score
#> 1         0.4506667
#> 2         0.4306667
#> 3         0.2573333
```

We see in the above table that for each slide and marker pair, we generate a `discordance_score` that summarizes the distance between the `slide_threshold` and `marker_threshold`. Since we have completed this analysis, we can also begin to visualize some of the results. First, we plot the densities of each marker before and after normalization, along with the associated Otsu thresholds visible as a ticks in the rug plot for each density curve in Figure 3.2.

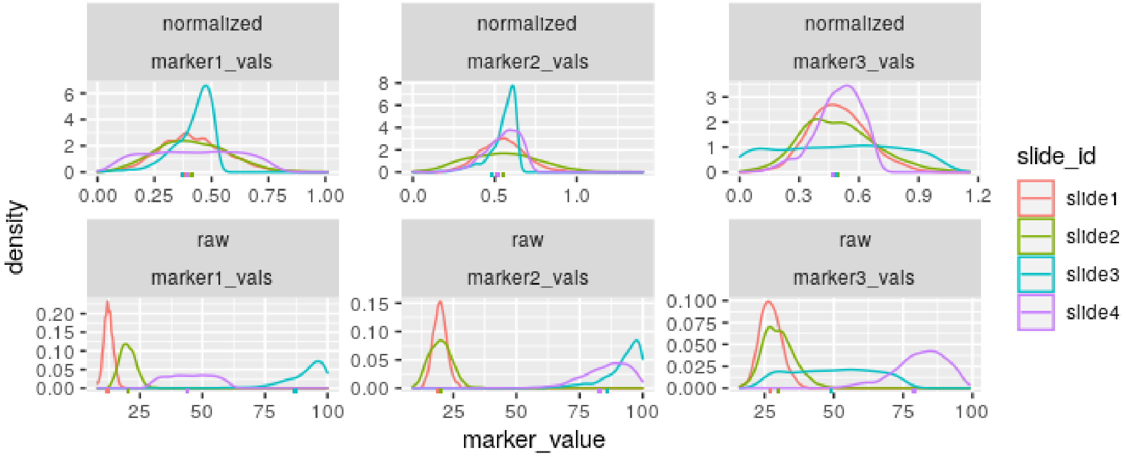

Figure 3.2: **Marker density alignment of the `mx_sample` dataset.** Demonstrates the alignment of marker values across simulated slides in the `mx_sample` dataset for both the raw and normalized datasets.

In Figure 3.2 we observe that not only are the density curves for each marker in the analysis far better aligned after normalization, we also see that the Otsu thresholds (ticks on the x-axis) have moved far closer than in the `raw` data. In general, also note that all plots generated using `mxnorm` are `ggplot2` plots and can be adjusted and adapted as needed given the `ggplot2` framework. We can also visualize the results of the threshold discordance analysis stratified by slide and marker, with slide means indicated by the white diamonds, as demonstrated in Figure 3.3 below.

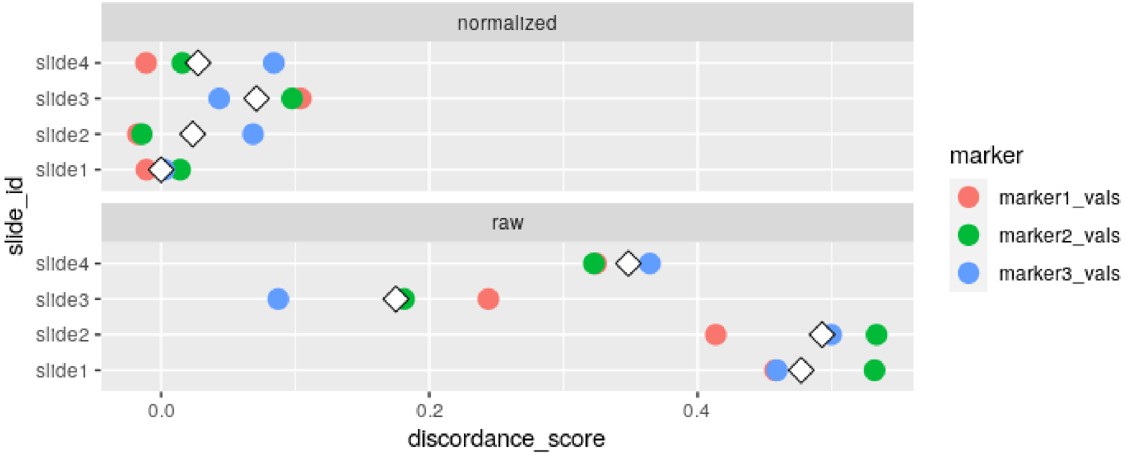

Figure 3.3: **Average Otsu discordance scores in the `mx_sample` dataset.** Demonstrates the Otsu discordance scores across all simulated marker and slide combinations in the `mx_sample` dataset, with slide-level averages depicted as white diamonds.

Note that for each slide and marker pair in the dataset (denoted as colored points in the above plot), we see a reduction in threshold discordance in the `normalized` data compared to the `raw` data. Further, we also see dramatic improvements in the mean threshold discordance denoted by the white diamonds for the normalized data.

3.3.3.2 Implementation of Otsu discordance scores

We implement this metric as defined in the Harris et al. (2022b) in the mxnorm package as an analysis method, e.g. run_otsu_discordance(), which takes in the mx_dataset object and produces an output table in the mx_dataset object called otsu_data which is shown above. The mean and SD of the discordance is also produced when summarizing the object using summary.mx_dataset() for a given mx_dataset object if the Otsu discordance analysis has already been run.

To calculate Otsu thresholds in our package, we use the thresholding options from the scikit-image.filters Python module which provide a notable speed increase on Otsu thresholding methods available in R (van der Walt et al., 2014). Note that thresholding options extend beyond just the Otsu threshold – the discordance score can be overridden to either accept an user-defined thresholding method or one of the univariate thresholds from scikit-image.filters.

3.3.4 UMAP dimension reduction

3.3.4.1 Using **mxnorm::run_reduce_umap()**

We can also use the UMAP algorithm to reduce the dimensions of our markers in the dataset as follows, using the metadata_col parameter for later (e.g., this metadata is similar to tissue type, medical condition, subject group, etc. in practice). The UMAP algorithm is stochastic, so we use set.seed() below to ensure results are reproducible. Here we specify:

- mx_data: the mx_dataset object with the data we want to analyze
- table: the set of data we want to analyze using UMAP dimension reduction, either raw, normalized, or both
- marker_list: the markers in the mx_dataset object we want to use for dimension reduction
- downsample_pct: UMAP embedding can be computationally expensive for big datasets, so we present a downsample percentage to reduce the input data size
- metadata_col: any metadata in mx_dataset to store for plotting later (see below for plotting using metadata1_vals)

And now run the following command:

```
set.seed(1234)
mx_data = run_reduce_umap(mx_data,
                          table="both",
                          marker_list = c("marker1_vals",
                           "marker2_vals",
                           "marker3_vals"),
                          downsample_pct = 0.5,
                          metadata_cols = c("metadata1_vals"))
```

This adds UMAP dimensions to our `mx_dataset` object in the following form (note the inclusion of `slide_id` as an identifier, which we'll use later) and the `umap_table` attribute to denote which tables in our object we ran the analysis on. We can observe this data, and note the inclusion of UMAP coordinates:

```
#> marker1_vals marker2_vals marker3_vals metadata1_vals slide_id table
#> 1004           22           22           30             no   slide2   raw
#> 623            12           19           28            yes   slide1   raw
#> 2953           60           89           91            yes   slide4   raw
#>             U1         U2
#> 1004  -3.63464 -0.9834719
#> 623  -10.35462 -2.0627188
#> 2953  10.10994 -4.7028897
```

We can further visualize the results of the UMAP dimension reduction as follows using the metadata column we specified above in Figure 3.4.

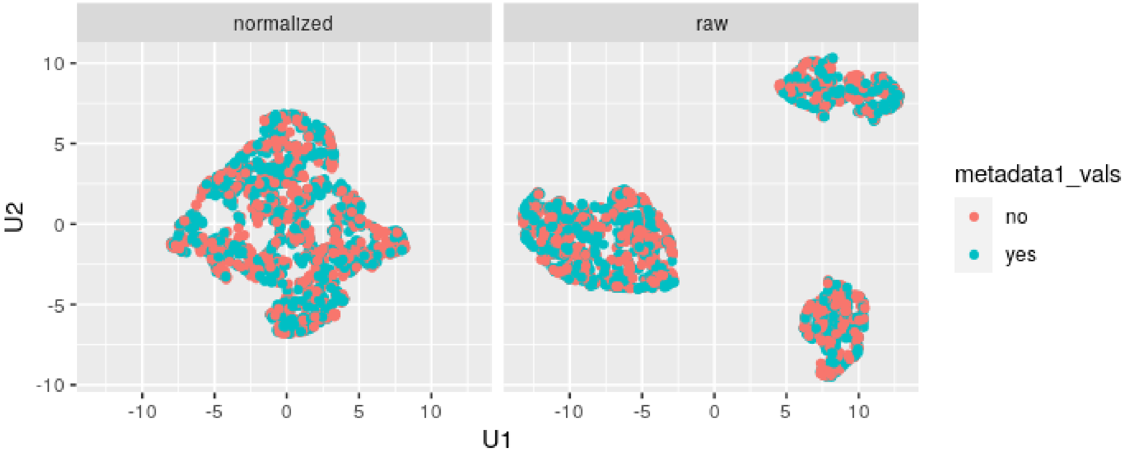

Figure 3.4: **UMAP embedding of `mx_sample` dataset for simulated metadata.** Demonstrates theoretical separation of groups in the raw and normalized datasets for UMAP embedding coordinates in the `mx_sample`.

Note that since the sample data is simulated, we don't see separation of the groups like we would expect with biological samples that have some underlying correlation. What we can observe, however, is the clear separation of slides in the `raw` data and subsequent mixing of these slides in the `normalized` data in Figure 3.5. This points to a removal of slide effects in the raw data when normalizing this dataset using the mean division method, as defined previously in Chapter 2.

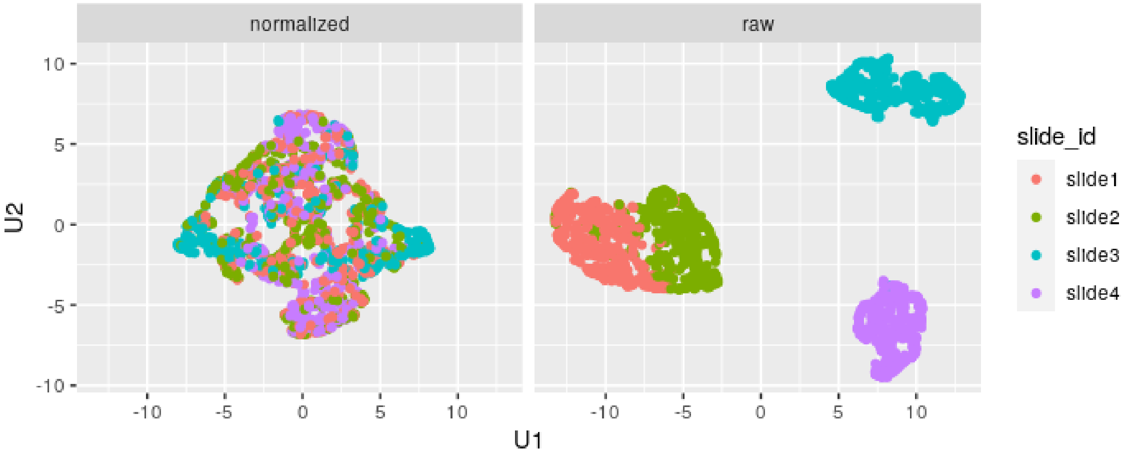

Figure 3.5: **UMAP embedding of `mx_sample` dataset for simulated slide identifiers.** Demonstrates separation of slides in the raw and normalized datasets for UMAP embedding coordinates in the `mx_sample`.

3.3.4.2 Implementation of UMAP embedding

The UMAP embedding algorithm McInnes et al. (2018), a dimension reduction commonly used in the biological sciences, is implemented here using the `uwot` R package (Melville, 2021). The method is often used to distinguish differences between groups and here can be used to highlight slide effects (clustering of slides) or determine biological separation of groups as shown in Figures 3.4 and 3.5. These options must be included in the `run_reduce_umap()` call using the `metadata_cols` parameter, and then can be visually inspected using the `plot_mx_umap()` method. Also note that the UMAP algorithm may take up significant computational time for large datasets – we've allowed for random downsampling of the data via the `downsample_pct` parameter to alleviate these concerns.

To further quantify the separation of groups for some given metadata, we implement the Cohen's kappa metric from the `psych` package and adjusted Rand index from the `fossil` package (Revelle, 2021; Vavrek, 2011). Each of these are executed using `summary.mx_dataset()` on an `mx_dataset` object that has already run a UMAP embedding analysis.

3.3.5 Variance components analysis

3.3.5.1 Using `mxnorm::run_var_proportions()`

We can also leverage `lmer()` from the `lme4` package to perform random effects modeling on the data to determine how much variance is present at the slide level. The default model specified is as follows for each marker in the `mx_dataset` object (e.g. a random intercept model where the intercept is `slide_id` for each marker), with any specifications of `metadata_cols` in the `run_var_proportions()` call adding fixed effects into the model below:

$$\text{marker} \sim \text{metadata\_cols} + (1|\text{slide\_id})$$

Note that the model we fit below sets the `metadata_cols` to `NULL`, implying the following basic random intercepts model:

$$\text{marker} \sim (1|\text{slide\_id})$$

In general, for an effective normalization algorithm we seek a method that reduces the proportion of variance at the `slide` level after normalization. Here we specify the following to run this analysis:

- `mx_data`: the `mx_dataset` object with the data we want to analyze
- `table`: the set of data we want to analyze using random effects, either `raw`, `normalized`, or `both`
- `metadata_cols`: any metadata in `mx_dataset` to add as fixed effects covariates,

- formula_override: an optional parameter to provide a user-defined random effects formula (see the details below for example)

- save_models: an optional parameter to save the lme4 models in the mx_dataset object

And now we run the following command:

```
mx_data = run_var_proportions(mx_data,
                              table="both",
                              metadata_cols = NULL,
                              formula_override = NULL,
                              save_models = FALSE)
```

After running the analysis, we see the addition of variance proportions to our mx_dataset object in the following form:

```
#>   proportions    level       marker table
#> 1:  0.97044933    slide marker1_vals   raw
#> 2:  0.02955067 residual marker1_vals   raw
#> 3:  0.97345576    slide marker2_vals   raw
#> 4:  0.02654424 residual marker2_vals   raw
```

These values summarize the proportion of variance explained by the random effect for slide, and any residual variance in the model. To understand how normalization impacts these values, we can further visualize these proportions in Figure 3.6.

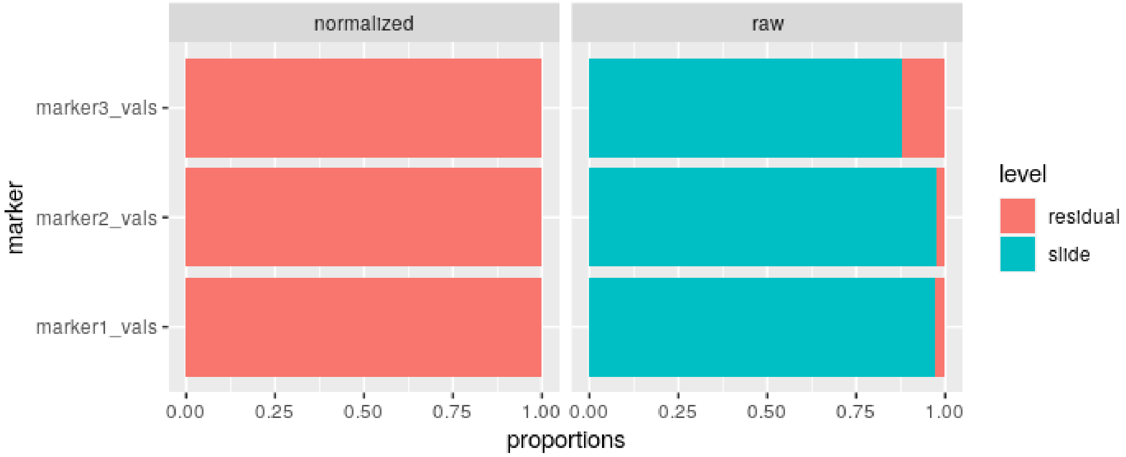

Figure 3.6: **Variance proportions for mx_sample dataset.** Demonstrates reduction in slide-to-slide variance proportions in the normalized data compared to raw mx_sample values.

In Figure 3.6 see that most of the variance in these models is due to slide-level effects in the `raw` data, but after normalization, nearly all of the variance in these random effects models due to slide-level effects is removed. This points to a normalization method that is performing well and removing the slide-to-slide variation in this type of data.

3.3.5.2 Implementation of variance components

Here we utilize random effects modeling in the `lme4` package (Bates et al., 2015). The default analysis fits a model for each marker in the dataset using only a slide-level intercept – this model can include additional covariates when using the `metadata_cols` parameter or re-define the modeling formula using `formula_override`.

3.4 User-defined normalization

As discussed in this chapter, one of the most important contributions of the `mxnorm` R package is the ability for users to define their own normalization methods. The goal of this functionality is to rapidly accelerate the development and evaluation of normalization methods in multiplexed imaging data, and ultimately add a vital tool to the multiplexed imaging pipeline. Here, we demonstrate a brief example of user-defined normalization – of less relevance to this thesis is the ability for users to define custom thresholding algorithms and random effects modeling using `mxnorm`. These are explored further in the package vignette.

As discussed in both Harris et al. (2022b) and Harris et al. (2022a), we find that the mean division normalization method performs the best across all evaluation metrics. However, let us consider a user-defined normalization method that instead of dividing the marker values by the slide mean, we divide by the median value. First, let us define this normalization function as follows:

```
quantile_divide <- function(mx_data, ptile=0.5){
    ## data to normalize
    ndat = mx_data$data

    ## marker columns
    cols = mx_data$marker_cols

    ## slide id
    slide = mx_data$slide_id

    ## get column length slide medians
    y = ndat %>%
        dplyr::group_by(.data[[slide]]) %>%
        dplyr::mutate(dplyr::across(all_of(cols),quantile,ptile))

    ## divide to normalize
    ndat[,cols] = ndat[,cols]/y[,cols]

    ## rescale
    ndat = ndat %>%
        dplyr::mutate(dplyr::across(all_of(cols),function(a){a + -min(a)}))

    ## set normalized data
    mx_data$norm_data = ndat

    ## return object
    mx_data
}
```

We first note a handful of important aspects of the `quantile_divide` function. First, we take in the `mx_data` object and some quantile that we wish to divide by (in this case, 0.5 since we are considering the median). We then use `tidyverse` methods to calculate a slide-level median, normalize all values, and re-scale. The `mx_data` object is then returned, with the added normalized data as an attribute. In general, applying user-defined normalization is as simple as this – any method and/or computation can be performed on the data, as long as the function takes in the `mx_dataset` object as input, and returns the same object with newly-normalized data. This is then passed to the `mx_normalize()` function, which would look something like the following:

```
## setup object
mx_user = mx_dataset(data=mx_sample,
                     slide_id="slide_id",
                     image_id="image_id",
                     marker_cols=c("marker1_vals","marker2_vals","marker3_vals"),
                     metadata_cols=c("metadata1_vals"))
## normalize with user-defined function
mx_user = mx_normalize(mx_user,
                       method_override = quantile_divide,
                             method_override_name = "median_divide")
```

Hence, we have now normalized the multiplexed imaging data using our own normalization technique. As noted above, analogous approaches exist to define `threshold_override` and `formula_override` for the Otsu discordance scores and variance components analysis respectively. In short, this added flexibility provides additional control to users of `mxnorm` to best handle the normalization of multiplexed imaging data within their respective processing pipelines.

CHAPTER 4

Applying spatial statistics methods to multiplexed imaging data

4.1 Summary

While multiplexed imaging methods are relatively new, cutting-edge research contributions in this field introduce a handful of the spatial analysis methods one might utilize when analyzing spatial relationships in multiplexed imaging data. These include visualization and exploration tools, spatial modeling approaches similar to differential expression testing in sequencing data analysis, and statistical learning models. Here we provide a brief survey of these spatial analysis methods, with a particular focus on three methods that leverage the statistical framework of point process theory. We then adapt each of these methods to compare statistical measures of spatial co-expression in multiplexed imaging data, including a new statistic that we define as the cumulative mark cross-correlation (CMCC). Finally, we develop a novel evaluation criteria using correlation analysis and predictive modeling in a non-small cell lung cancer (NSCLC) dataset to compare spatial analysis methods and determine which spatial index is best suited for multiplexed imaging data.

4.2 Introduction

Multiplexed imaging is a rapidly growing field of research that combines single-cell information with spatial coordinates to better understand complex biological processes like cancer development and tumor growth. Multiplexed imaging experiments generate data across hundreds of slides and images, creating terabytes of complex, spatial data via imaging analysis pipelines. One major development in this field is the application of spatial data analysis methods to visualize, explore, and analyze relationships between markers and tissue classes in this complex data source (Dries et al., 2021a; Wilson et al., 2021).

The first class of spatial methods introduced in the multiplexed imaging field is high-level visualization and exploration tools. These include packages like Seurat and Giotto (Hao et al., 2021b; Dries et al., 2021b), which function as toolkits that incorporate dozens of visualization methods and analysis functions. Both of these packages began as single-cell data analysis tools for use in sequencing data, and have evolved to adapt many of these methods for imaging and spatial data. Methods specific to visualizing and exploring multiplexed imaging data have also recently been introduced – the histoCAT toolbox provides an interactive exploration of cell phenotypes and neighborhood analyses (Schapiro et al., 2017), cytoMAP is a user-friendly, comprehensive platform for the spatial analysis of multiplexed tissues (Stoltzfus et al., 2020), and Squidpy is a tool for the visualization and analysis of spatial molecular data (Palla et al., 2022). Each of these toolkits provides similar features – a unified framework to visualize and perform basic spatial analysis within an interactive application or programming environment (typically R or Python), for a handful of multiplexed imaging types.

Considering that many of the methods in multiplexed imaging have evolved and taken inspiration from the single-cell sequencing field, analysis methods have also been developed to generate spatial models for differential expression. This was first introduced by Edsgärd et al. (2018) with the Trendsceek method, which is based on marked-point processes to identify genes with statistically significant spatial expression trends. More recently, Sun et al. (2020) introduce SPARK as a generalized linear spatial model for identifying spatial expression patterns of genes, and the recently introduced SPARK-X adapts this method non-parametrically to detect spatially expressed genes (Zhu et al., 2021). Further developments include other differential expression methods applied in the spatial domain like SpatialDE and SOMDE, which both model spatial data as Gaussian processes to identify spatially variable genes (Svensson et al., 2018; Hao et al., 2021a). The biggest distinction for each of these differential expression methods is that they identify spatial trends at the *gene* level, rather than for a given region or marker, which is most applicable for technologies like spatial transcriptomics. While relevant, we ultimately seek to make inference beyond the gene level for many types of multiplexed imaging studies.

A handful of these methods have also adapted various statistical learning models to either identify spatial gene expression or spatial neighborhoods. These include hidden Markov random fields (HMRF) as applied in sequential fluorescence *in situ* hybridization data (Zhu et al., 2018), and has been adapted for use in the Giotto package (Dries et al., 2021b). The BayesSpace method further implements a fully Bayesian model with Markov random field for use in spatial transcriptomic studies (Zhao et al., 2021). staNMF is a method that implements non-negative matrix factorization (Wu et al., 2016), while recent developments include applying spatial latent Dirichlet allocation to multiplexed imaging analysis (Chen et al., 2020). Again, these methods provide a foundation for the breadth of methods applied in multiplexed imaging and similar fields, but ultimately do not address our question of interest in this work.

Broadly, we are interested in a statistical method to quantify spatial co-expression of biological markers in multiplexed imaging data. A few relevant methods have leveraged marked point process theory to introduce measures of spatial co-expression. The SpatialTIME package includes many first-order summary statistics used in spatial analysis like Ripley's K-function that are applied to multiplexed imaging data, including some basic neighborhood analysis functions that introduce the idea of co-expression (Creed et al., 2021). However, of particular relevance to this work are second-order characteristics that incorporate both the spatial coordinates of multiplexed imaging data and the values of quantitative markers, which in this case we will consider the “marks” in the marked point processes. Keren et al. (2018) first introduced a spatial proximity measure and permutation testing procedure that is widely adopted in multiplexed imaging, while Chervoneva et al. (2021) introduced a method based on marked point processes to create a spatial index as a predictor for outcomes of interest.

In this chapter, we explore the methods introduced by Keren and Chervoneva, and adapt these methods into more reasonable summary statistics for spatial co-expression in multiplexed imaging data – a normalized version of the Keren’s statistic and the introduction of a new spatial index, the cumulative mark cross-correlation (CMCC), based broadly on Chervoneva’s approach. Furthermore, the Keren’s statistic is uncompared to other spatial statistics methods in the multiplexed imaging literature. We also introduce the Lee’s L-statistic for bivariate spatial data to compare the aforementioned spatial statistics with a more classical statistical summary, and to present the first comparison of this kind for spatial co-expression in multiplexed imaging data. We then evaluate this comparison in a NSCLC dataset using a correlation analysis to understand the amount of shared information between these quantities and implement a cross-validated prediction model to quantify which method best summarizes the available spatial information.

4.3 Background

Point processes are stochastic models of point patterns, an irregular collection of points in some area (or set). These models are commonly used in many different fields, for example, to understand the distribution of plants in ecology (Law et al., 2009), determining optimal use of neurophysiological measurements in neuroscience (Brown et al., 2004), and exploring the impact of quotes on financial trades in economics (Engle and Lunde, 2003). In multiplexed imaging, a handful of methods have implemented point process methods (Chervoneva et al., 2021; Keren et al., 2018; Edsgärd et al., 2018), which we seek to define, extend, and evaluate to explore bivariate spatial relationships among marker channels in multiplexed imaging data.

Mathematically we define some stationary point process N as a random counting measure observed in a bounded region $S \subset \mathbb{R}^2$, where we interpret some observed point pattern of cell coordinates as a random realization of N. Let $N(B)$ be defined as the number of points falling into any Borel set $B \subset S$, and the random set $N_p = \{x_n, \ldots\}$ as the set of all points in the process. For this definition we assume both additivity (for disjoint sets) and simplicity (that all points are different) (Chervoneva et al., 2021; Illian et al., 2008).

In multiplexed imaging, we are interested in the spatial relationship of some set of points $\{x_1, x_2, \ldots\}$ that represent cell locations, and the value of some marker channel at each of those cells $\{m_1, m_2, \ldots\}$. Marked point process methods consider a stationary process defined previously as a random counting measure, $M(B)$, for $B \subset \mathbb{R}^d$ and $C \subset \mathbb{R}$. $M(B \times C)$ thus denotes the random number of marked points $[x_n; m(x_n)]$ with $x_n \in B$ and $m(x_n) \in C$. We can also consider the *random set* of all points in the process, $M_p = \{[x_n; m(x_n)], \ldots\}$ (Illian et al., 2008).

We are often interested in summarizing the relationship between points $\{x_n\}$ and marks $\{m_n\}$, for example, understanding how markers are co-expressed within tissue types or tumors, which are spatially dependent biological processes. Example summary statistics at the point-level include the point process intensity $\Lambda(B) = \mathbb{E}(N(B))$ that measures the number of units per region, common summary statistics like Ripley's K-function that measures the average number of points within some distance from the typical point (Ripley, 1976), and Moran's index for spatial autocorrelation (Moran, 1948). Of interest in this chapter is Ripley's K-function, as it relates to Keren's permutation test as described below. The classic Ripley's K (which does not count the starting point u) can be defined by:

$$\lambda K(r) = \mathbb{E}\left[N(b(u,v) \setminus \{u\})\right] \text{ for } r \geq 0$$

Hence, $\lambda K(r)$ is mean number of points in a sphere of radius r centered at the starting point u (irrespective of mark values). Note here that the K-function is a *first*-order summary characteristic, e.g., one that concerns only one value of the process, such as the locations of the points or the probability that a point has some mark value. To study relationships between cell types in multiplexed imaging, we are interested in *second*-order summary statistics that are calculated using both the coordinates of the point process as well as the values of the marks. First, let us introduce the conditional mean of a mark, as defined in Chervoneva et al. (2021), given there is another point of the process a distance r away, can be written as (Schlather et al., 2004):

$$cMean(r) = \mathbb{E}\left[m(\mathbf{u}) | \mathbf{u}, \mathbf{v} \in M_p, ||\mathbf{u} - \mathbf{v}|| = r\right]$$

Hence $cMean(r)$ denotes the conditional mean of the marks at some set of points $\mathbf{u} = \{[x_u; m(x_u)], \ldots\}$, given some set $\mathbf{v} = \{[x_v; m(x_v)], \ldots\}$ of points at distance r away, where the expectation is taken over different values of $\mathbf{u}$. Further, the mark correlation function is useful for quantifying the relationship between quantitative marks – for example in multiplexed imaging data, the co-expression of two biological markers. We first define the non-normalized mark correlation function given some test function, $t(m_1,m_2)$:

$$c_t(r) = \mathbb{E}\left[t(m(\mathbf{u}),m(\mathbf{v}))|\mathbf{u},\mathbf{v} \in M_p, ||\mathbf{u}-\mathbf{v}|| = r\right]$$

This is the mean of the test function $t(m(\mathbf{u}),m(\mathbf{v}))$, at some set of points $\mathbf{u} = \{[x_u;m(x_u)],\dots\}$ with some set $\mathbf{v} = \{[x_v;m(x_v)],\dots\}$ of points at radius r (Illian et al., 2008). For the mark correlation function, we define the test function as $t(m_1,m_2) = m_1 \cdot m_2$ and normalize by the normalizing factor

$$\begin{aligned} c_t(\infty) &= \mathbb{E}\left[m(\mathbf{u}) \cdot m(\mathbf{v})|\mathbf{u},\mathbf{v} \in M_p, ||\mathbf{u}-\mathbf{v}|| = \infty\right] \\ &= \mathbb{E}\left[m(\mathbf{u})\right] \cdot \mathbb{E}\left[m(\mathbf{v})\right] \\ &= \mu^2, \end{aligned}$$

which is the value the function takes at very large distances r when the marks are effectively independent. We can now define the mark correlation function,

$$k_{mm}(r) = \frac{c_t(r)}{c_t(\infty)} = \frac{\mathbb{E}\left[m(\mathbf{u})|\mathbf{u},\mathbf{v} \in M_p, ||\mathbf{u}-\mathbf{v}|| = r\right]}{\mu^2}$$

The numerator is the conditional mean of the mark product of points in marked point process M with distance r from the starting point – we then divide by μ^2 to determine departures from theoretical independence (Illian et al., 2008). Hence, if $k_{mm}(r) \approx 1$ we conclude effective independence of the marks at some distance r, if $k_{mm}(r) > 1$ we conclude that the marks are spatially correlated at distance r, and if $k_{mm}(r) < 1$ we conclude that the marks are spatially anti-correlated at distance r.

We can further manipulate the definition of some arbitrary $c_t(r)$ and its normalized counterpart $k_t(r)$ with alternate test functions to provide quantities useful for describing papers in the **Evaluation** section. For example, if we define some test function $t(m_1,m_2) = (m_1-\mu)(m_2-\mu)$, we generate a function similar to the Moran's I-statistic (Shimatani, 2002), defined as (Illian et al., 2008):

$$I(r) = \frac{\mathbb{E}\left[(m(\mathbf{u})-\mu)(m(\mathbf{v})-\mu)\right]}{\sigma_\mu^2}$$

This becomes especially useful when considering Lee's L test for bivariate spatial data as defined below. We are interested in the following three methods that use the point process statistics defined above to understand spatial relationships of marker values in multiplexed imaging data. Here we introduce these methods and how we have adapted them for comparison in the **Evaluation** section.

4.4 Methods

4.4.1 Keren's statistic for spatial proximity

Keren et al. (2018) recently introduced a permutation test for assessing spatial proximity enrichment for pairs of markers that accounts for differential tissue structure across varying cell numbers and composition. They describe quantifying the number of marker-positive cells for marker X that are located within some radius r to marker-positive cells for marker Y, which we will define as Keren's statistic. The authors then randomized the locations of Y-positive cells to generate a null distribution of empirical Keren's statistics and calculated a Z-score representing the enrichment of X-positive cells close to Y-positive cells. This approach can be considered as a bootstrapped test of the Keren's statistic for some empirical null distribution of spatial proximity in the multiplexed images.

Despite its utility, Keren did not describe their method statistically. Here, we derive Keren's statistic in the context of marked point processes by comparing it to Ripley's K-function. Let $M(B)$ and M_p be as defined above in **Section 4.3**; the estimator of $K(r)$ is of the form (Baddeley et al., 2015; Ripley, 1988):

$$\hat{K}(r) = \frac{a}{n(n-1)} \sum_i \sum_j I(d_{ij} \leq r) \cdot e_{ij}$$

where a is the area of the window, n is the number of points, and the sum is over all points i and j in the point pattern. Here d_{ij} is the distance between two points, e_{ij} as the edge correction weight, and $I(d_{ij} \leq r)$ is an indicator that the distance between two points is less than some value r. Empirically, this estimator is biased for $K(r)$ due to the edge correction method and considering we only record observed points. Translating Keren's statistic as the number of "close" interactions between marker-positive cells for two given markers into a mathematical formula gives:

$$KS(r) = \sum_i \sum_j I(d_{ij} \leq r) I\left(m_1(x_i) \geq k\right) I(m_2(x_j) \geq k)$$

where $m_1(x_i)$ is the quantitative mark of some marker channel for point x_i, $m_2(x_j)$ is the quantitative mark of some marker channel for point x_j, and k is some threshold of marker positivity. We recognize this as an estimator of the mark-weighted K-function, as defined elsewhere including Illian et al. (2008) **Section 5.3.4**. Hence, we can consider Keren's raw statistic as a mark-weighted, un-normalized K-function. Note that this summary is a second order characteristic, as it depends on the mark values of the process. Noting the similarity between $\hat{K}(r)$ above and $\widehat{KS}(r)$, we normalize Keren's statistic as follows:

$$\widehat{KS}(r) = \frac{a}{n(n-1)} \sum_i \sum_j I(d_{ij} \leq r) I(m_1(x_i) \geq k) I(m_2(x_j) \geq k).$$

Hence $\widehat{KS}(r)$ is a consistent estimator for $KS(r)$, a form of the mark-weighted K-function to identify "close" interactions between two marker channels in multiplexed imaging data.

4.4.2 Lee's L test for bivariate spatial data

It is of interest to compare the popular Keren's statistic with more classical measures of spatial relationships like Moran's I-statistic (Moran, 1948). However, even a bivariate Moran's I-statistic is a global spatial summary that summarizes the spatial auto-correlation, rather than spatial co-dependence between two variables. Pointing to the drawbacks of using the classical Moran's I and Pearson's r correlation in spatial data, Lee (2001) introduces a bivariate spatial association measure (the L-statistic) to capture spatial co-patterning between two variables that incorporates both the point-to-point relationship and the spatial relationship between the variables. Although defined in context without marked point process theory, let us re-formulate the L-statistic using the marked point process $M(B)$ defined above, with two quantitative mark vectors of interest, $\underline{m}_1(x_n)$ and $\underline{m}_2(x_n)$.

$$L_{m_1,m_2} = \frac{\sum_i \left[\left(\sum_j w_{ij}(m_{1j} - \bar{m}_1) \right) \cdot \left(\sum_j w_{ij}(m_{2j} - \bar{m}_2) \right) \right]}{\sqrt{\sum (m_{1i} - \bar{m}_1)^2} \sqrt{\sum (m_{2i} - \bar{m}_2)^2}}$$

Here, note that the $\mathbf{W}$ is the row-standardized spatial weights matrix for the point locations $\{x_n\}$ of the point process M. We also recognize that the formulation for Lee's L is quite similar to the bivariate Moran's I, however, Lee (2001) shows further that the derivation of the bivariate Moran's I from Wartenberg (1985) is vulnerable as a bivariate measure of spatial relationships due to its calibration and susceptibility to false negatives.

Hence, we prefer the Lee's L-statistic as a spatially weighted and smoothed correlation of the quantitative marks, which improves upon classical spatial statistics like Moran's I that we seek to apply in this work to multiplexed imaging data.

4.4.3 Cumulative mark summaries

Recall again the definition of the conditional mark mean, $cMean(r)$, as defined above:

$$cMean(r) = \mathbb{E}\left[m(\mathbf{u}) | \mathbf{u}, \mathbf{v} \in M_p, ||\mathbf{u} - \mathbf{v}|| = r \right].$$

In this paper, the authors estimate the quantity non-parametrically using the following form from Schlather et al. (2004); Schlather (2001):

$$\widehat{cMean}(r) = \frac{1}{N_d} \sum_{|||\mathbf{u}-\mathbf{v}||-d| \leq \varepsilon/2} m(\mathbf{u}),$$

where d is the distance between two points, $\varepsilon > 0$ is a fixed bin width, and N_d is the number of pairs $(\mathbf{u}, \mathbf{v})$ such that the distance between the points meets the criterion: $|||\mathbf{u}-\mathbf{v}||-d| \leq \varepsilon/2$. Since these are functions of the distance r between points in the marked point process pattern, the authors adapt the function into a suitable cumulative index that can be investigated as a predictor of outcomes of interest. This quantity, $AUcMean_i(r)$ is defined as

$$\widehat{AUcMean}_i(d_{max}) = \frac{1}{d_{max}} \int_0^{d_{max}} \widehat{cMean}(r)dr$$

Here, $AUcMean$ is the average conditional mean of the mark in the set of marked points between 0 and the value of d_{max}. However, in this chapter, we are interested in bivariate spatial relationships in multiplexed imaging data, e.g. the relationship between some marker m_i and another marker m_j in regards to biological variables of interest like tissue type and tumor class. Hence, we re-derive a new index following the methods of Chervoneva et al. (2021) using the mark cross-correlation function, similar to the $k_{mm}(r)$ as defined above. First, let us define the mark cross-correlation $k^*_{mm}(r)$ in the following form similar to the mark correlation defined in **4.3**:

$$k^*_{mm}(r) = \frac{\mathbb{E}\left[m_i(\mathbf{u}) \cdot m_j(\mathbf{v})\right]}{\mu_i \cdot \mu_j}$$

for some set of two marks, $m_i(x_n) = m_{in}$ and $m_j(x_n) = m_{jn}$, and the mean mark values across the set of data μ_i and μ_j respectively. Hence we can derive an estimator of this function using many of the same quantities defined in Chervoneva et al. (2021):

$$\hat{k}^*_{mm}(r) = \frac{1}{N_d} \sum_{|||\mathbf{u}-\mathbf{v}||-d| \leq \varepsilon/2} \frac{m_i(\mathbf{u}) \cdot m_j(\mathbf{v})}{\hat{\mu}_i \cdot \hat{\mu}_j},$$

where again d is the distance between two points, $\varepsilon > 0$ is a fixed bin width, N_d is the number of pairs $(\mathbf{y}, \mathbf{v})$ such that the distance between the points meets the criterion: $|||\mathbf{u}-\mathbf{v}||-d| \leq \varepsilon/2$, and $\hat{\mu}_i$ and $\hat{\mu}_j$ are the mean mark values across the set of data respectively. Lastly, we define

$$CMCC(d_{max}) := AUk^*_{mm}(d_{max}) = \frac{1}{d_{max}} \int_0^{d_{max}} \hat{k}^*_{mm}(r)dr$$

as the mean mark cross-correlation over the set of marked points between 0 and the value of d_{max}, or similar to the cumulative mark product function (Shimatani and Kubota, 2004). Since we are considering this function between two sets of marks m_i and m_j, we define this statistic as the cumulative mark cross-correlation (CMCC). Hence, we have taken the approach from Chervoneva et al in the context of Shimatani and Kubota, and adapted a new spatial index to compare bivariate relationships of marker values in multiplexed imaging.

4.4.4 Data source

The NSCLC data was collected to analyze the relationship between tumor-infiltrating immune cells and major histocompatibility II expressing cancer cells in the NSCLC tumor microenvironment (TME), and consists of 761 mIF-imaged regions of interest (ROIs) from 153 patients (Johnson et al., 2021), e.g. roughly 5 ROIs per patient. Images were stained for DAPI, five phenotypic markers (CD3, CD8, CD14, CD19, cytokeratin) and one functional marker HLADR (MHCII), allowing for identification of CD4+, CD8+, CD14+, and CD19+ immune cells. The dataset has values at the cell-level, where we focus on four markers – two immune cell markers CD8 & CD14, a cancer-cell marker cytokeratin, and the functional marker HLADR. Marker values were normalized using the `mxnorm` R package (Harris et al., 2022a). We then use a cell-level identifier of tumor vs. stromal cells to infer the proportion of tumor cells within a region/image, which we use as an outcome of interest in the **Evaluation**.

4.4.5 Evaluation

4.4.5.1 Pairwise comparisons of spatial results

We quantify pairwise spatial relationships between marks by quantifying each of the three spatial summary statistics in a given ROI for all possible combinations of the markers discussed above. We then summarize these values across all ROIs to best understand the spatial co-expression of each marker pair of interest, and how each of the three statistical summaries captures that relationship.

4.4.5.2 Correlation analysis

After calculating each spatial summary statistic in a given ROI as discussed above, we can then further explore these statistical methods by comparing the values of each statistic produced for a given region/image to determine if any of the explored methods captures the same information from the NSCLC data. We consider this a correlation analysis, e.g., quantifying the amount of shared information between the statistical summaries within a ROI for a particular marker comparison.

4.4.5.3 Cross-validated prediction accuracy

To explore the prediction accuracy of each of the introduced spatial statistics summaries, we now formulate a prediction model. This model uses the proportion of tumor cells on a given image as the outcome of interest, and we formulate three models (one for each spatial statistic) with the following covariates: the spatial statistic summary (defined below as X), a factor variable `marker_comparison` that denotes which quantitative markers we are comparing like `CD8 vs. CD14`, and an interaction between these two variables. Hence, in `R` modeling notation we write each of these models as follows:

```
tumor_proportion ~ X + marker_comparison + X * marker_comparison
```

where the `tumor_proportion` and each statistic are both calculated at the image level, with the values of each statistic varying depending on the `marker_comparison` implemented. To estimate prediction accuracy, we holdout 25% of the ROIs and train a linear regression model as formulated above on the training set, with the spatial statistics scaled (e.g., divided by their standard deviation). We then predict the proportion of tumor in the holdout set using this model, and compute the sum of squared errors (SSE) to measure prediction accuracy. Lastly, we repeat this process 100 times to estimate the average SSE of these prediction models.

4.5 Evaluation

4.5.1 Methods of interest

In summary of the above, here we are interested in the following methods:

- **Lee's L-statistic**, which we define as the correlation between the spatial lag vectors of two random variables X and Y, weighted by a spatial smoothing scalar. **We interpret this quantity as a spatially weighted and smoothed correlation coefficient of two marker values.**

- **Keren's statistic**, which we define as the average number of X-positive points within some distance r of Y-positive points. Here we have normalized this statistic using the area and number of points akin to the estimator $\hat{K}(r)$. **We interpret this quantity as the number of close interactions in a region between marker-positive cells.**

- **CMCC (adapted from Chervoneva)**, which we define as the average mark cross-correlation over the set of marked points between o and some distance r for two marks, X and Y. **We interpret this quantity as the local departure of observed marker co-expression from theoretical independence.**

The Keren's statistic for measuring spatial proximity is widely used for spatial-omics data, and we showed it is analogous to a form of the mark-weighted K-function (Keren et al., 2018). It has not been compared to other spatial analysis methods. The first of these we chose to evaluate in comparison is the Lee's statistic for spatial autocorrelation, which combines the classical Pearson's correlation with the popular Moran's

I statistic into a bivariate spatial association measure (Lee, 2001). We adapted the univariate marker summaries introduced by Chervoneva et al. (2021) to handle bivariate spatial associations via the cumulative mark cross-correlation function (CMCC). Below we evaluate these three approaches using a correlation analysis and predictive modeling to understand if these methods capture similar spatial information from multiplexed imaging data and whether each provides a valuable summary in terms of predicting biological variables of interest.

Notably, both Keren's statistic and Chervoneva's $AUcMean$ methods have a distance parameter that is optimized for the data sources in these respective papers. However, in this chapter we maintain indices with some fixed d for continuity between the three measures discussed – for computing both Keren's statistic and for CMCC, we set $d = 40\mu$m. This choice is deliberate, and future work could address these methods over some varying set of distances, as well as implement Lee's local bivariate spatial association measure, to derive parameters as a function of some distance d and potentially implement a modeling comparison using functional data methods as performed by Vu et al. (2021) for example.

4.5.2 Comparison of spatial results across images

Using the four markers identified above, we first explore the three statistical summaries outlined above at the image/region level for each of the marker comparisons of interest. In the original analysis, the authors computed co-expression of the CD8 and CD14 markers in MHCII (HLADR) expressing cells (Johnson et al., 2021), using the `phenoptrReports` R package which generates reports and visualizations from data created by Akoya Biosciences' inForm software (Johnson, 2022). These researchers also point to a spatial interaction between cells expressing higher levels of cytokeratin and lower marker co-expression levels of immune cell markers like CD8 and CD14.

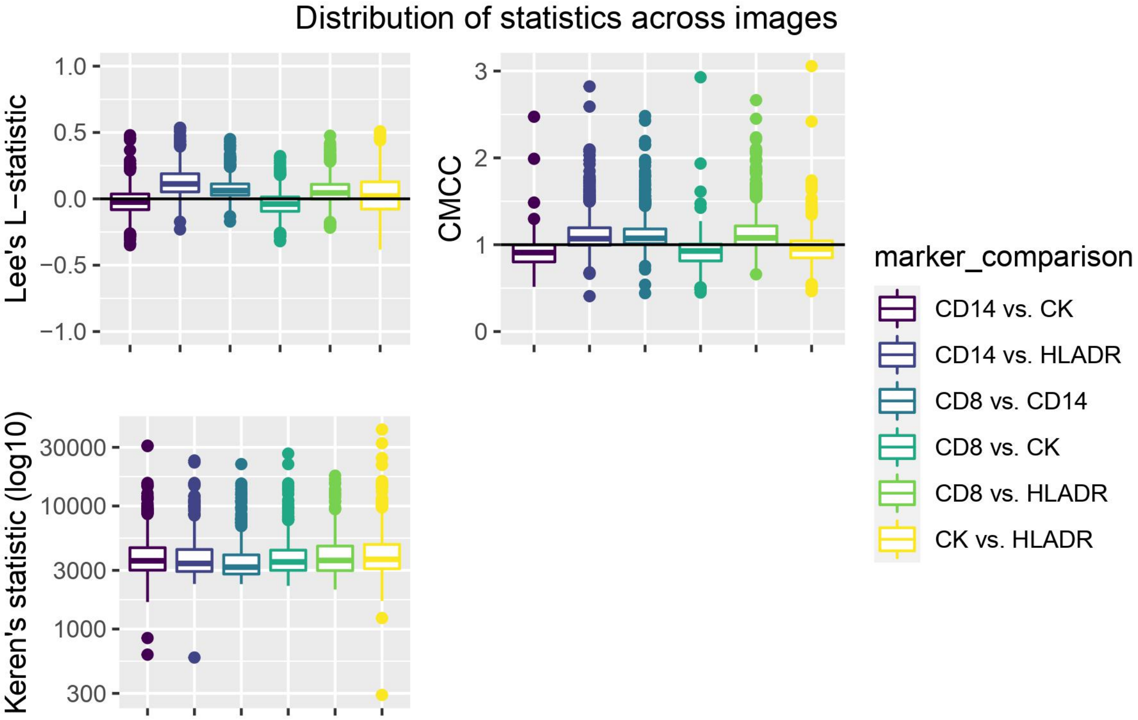

Figure 4.1: **Distribution of statistics across images in NSCLC dataset.** Boxplots comparing values of each statistic implemented across images/regions in the NSCLC multiplexed dataset.

The results in Figure 4.1 from Lee's L-statistic point to a positive spatial correlation between both CD14 and CD8 marker expression with HLADR expression, and co-expression patterns between the two immune cell markers CD8 and CD14. However, all other expression patterns are mostly inconclusive – in contrast, we can interpret the CMCC summary as the departure of the average marker co-expression from the theoretical null within some radius. Across images, we see that both CD14 and CD8 marker expression are positively co-expressed with the functional marker HLADR as expected, and again note that the immune markers CD8 and CD14 exhibit positive spatial co-expression of marker values for cells that are within the radius.

We further note that the CMCC points to reduced co-expression patterns between the two immune markers, CD8 and CD14, with the cancer-cell marker cytokeratin, which suggests a *reduction* in immune cell activity when in close proximity to cancer cells. This result, and the more reasonable interpretation of CMCC compared to Lee's L-statistic, provide a strong case for using CMCC going forward. Finally, we note that although Keren's statistic can be interpreted as an average number of "close" interactions between marker positive cells, it is quite difficult to interpret the results in Figure 4.1 and distinguish marker activity from the summary even when normalized by the number of cells. Below we will explore the efficacy of the Keren's statistic in both a correlation analysis and in regards to predictive modeling, but it is of note that the raw values of the statistic are difficult to interpret, especially when looking at distinct marker-marker interactions.

4.5.3 Correlation analysis

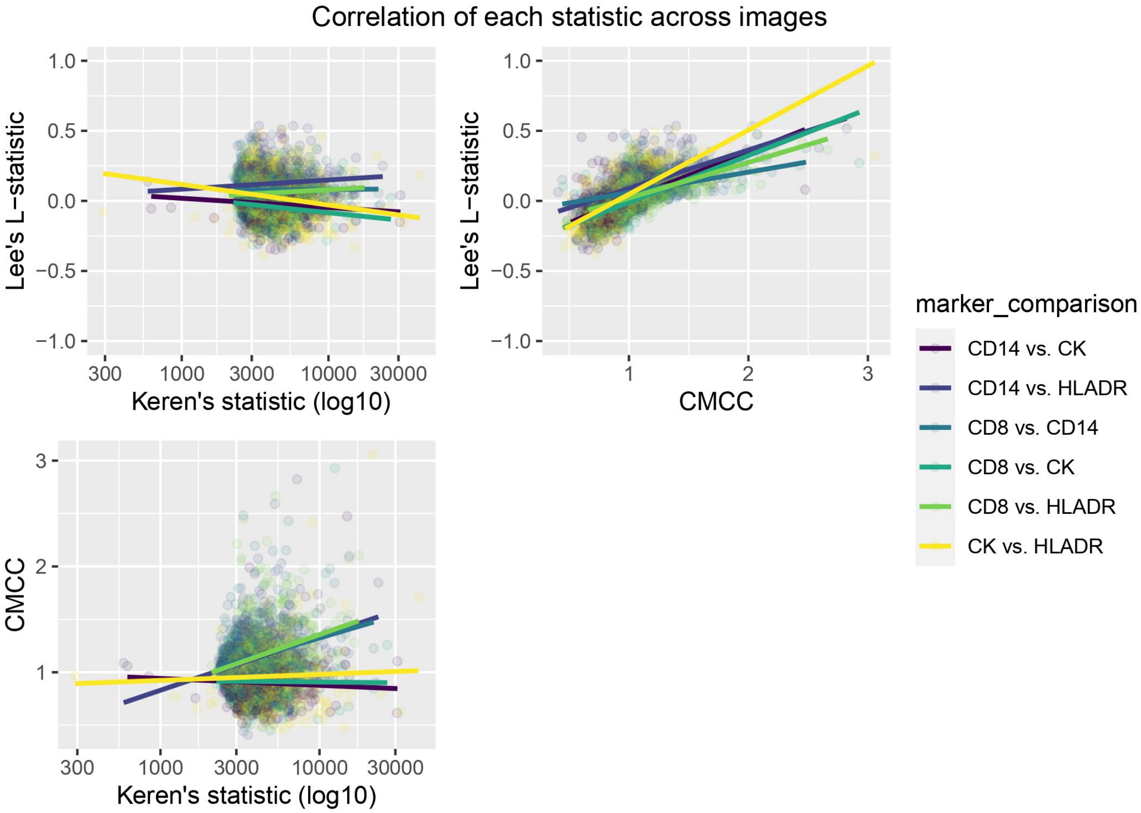

Figure 4.2: **Correlation of each statistic implemented in the NSCLC dataset.** Scatter plots comparing values of each summary across images/regions in the NSCLC multiplexed dataset with lines of best fit for each marker comparison.

We first note in the top-left panel of Figure 4.2 that Lee's L-statistic and Keren's statistic are largely uncorrelated, while Keren's statistic is only correlated with the CMCC for marker comparisons that do not include cytokeratin. This may perhaps be a result of the thresholding for marker-positivity in the Keren's method, effectively dichotomizing a continuous relationship, or may also point to the inability of Keren's method to provide a relative statistic that can be quickly interpreted for marker-marker interactions beyond a simple count measure.

We also note that for all marker comparisons, the CMCC is strongly positively correlated with Lee's L-statistic. This result is important, considering that above we point to the improvement in interpretability provided by CMCC in multiplexed imaging, and further conclude that the methods appear to be capturing similar spatial information from the data. Hence, in this correlation analysis we again provide support for the interpretability of CMCC that captures underlying spatial interactions between markers in multiplexed imaging data.

4.5.4 Cross-validated prediction accuracy

In Figure 4.3, we observe the average model performance using each of the three spatial statistics as covariates in predicting tumor proportion. We first note that the Lee statistic does not perform as well as either the CMCC or Keren statistic in terms of prediction, with a much higher average SSE over the cross-validated models.

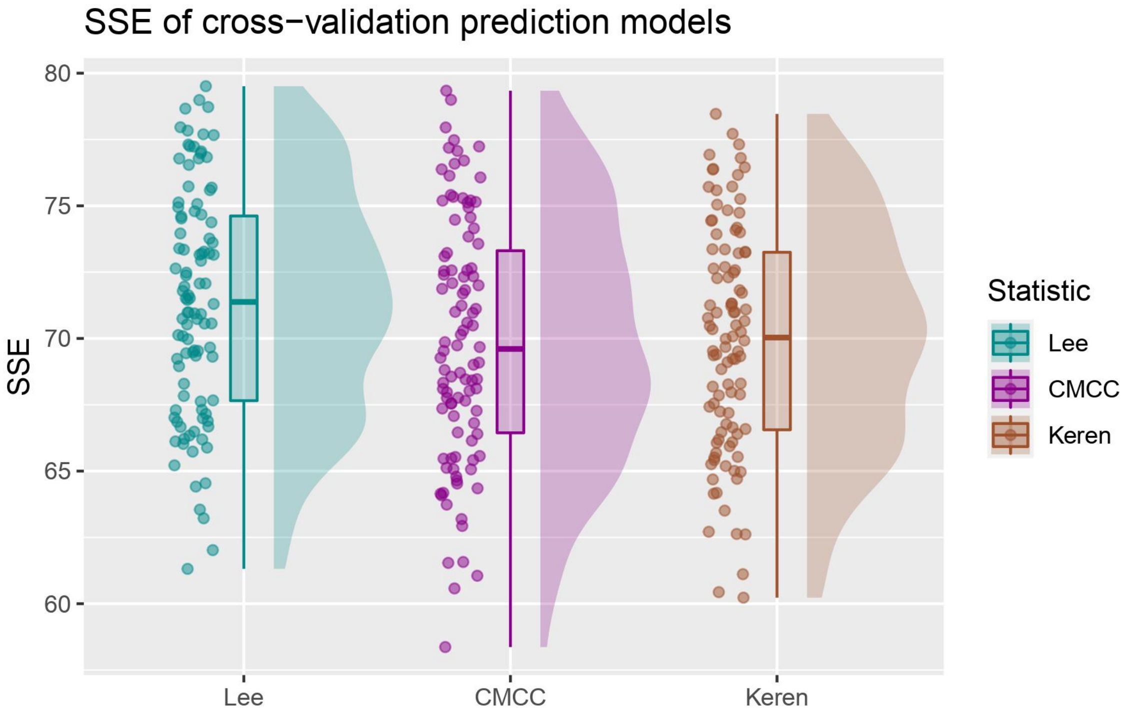

Figure 4.3: **Cross-validated prediction accuracy of models in the NSCLC dataset.** Raincloud plots comparing the distribution of SSEs for each spatial statistic model, for 100 cross-validated prediction models.

Further, we note similar performance by both the CMCC and Keren statistic, with a slightly longer tail and lower average error in the models that use the CMCC as a covariate. While inconclusive about a *better* spatial summary measure between the CMCC and Keren statistics, in terms of SSE it is clear these two methods both provide strong spatial summaries in predicting tumor proportion at the ROI level.

To further explore these models, we can also compare the absolute and relative importance of each variable as presented in Table 4.1.

Method	**Variable**	**Avg. Absolute Importance**	**Avg. Relative Importance**
CMCC	C	0.51	0.104
	C:marker_comparison		0.881
	marker_comparison		0.014
Keren	KS	-0.127	0.348
	KS:marker_comparison		0.650
	marker_comparison		0.002
Lee	L	-0.005	0.075
	L:marker_comparison		0.912
	marker_comparison		0.013

Table 4.1: **Variable importance of each statistic in prediction models.** Average absolute and relative importance of each group of variables in the linear regression models over the 100 cross-validated prediction models.

As noted previously, we have scaled the spatial statistics to allow for comparison of average absolute importance – on this metric, clearly the CMCC is the most absolutely important coefficient in the cross-validated models out of the statistics explored here, while Keren's statistic is the most important relative to other variables in those models. We hypothesize that this is because the Keren's statistic is a counting measure conditional on some set of marks, while the CMCC is a quantitative summary of mark values. Hence, we see a larger impact of which mark values are compared and their interaction term with the value of CMCC, in contrast to the Keren's statistic.

Method	Variables	Percent of Models
CMCC	C	1.00
	C:marker_comparisonCD14 vs. HLADR	1.00
	C:marker_comparisonCD8 vs. CD14	1.00
	C:marker_comparisonCD8 vs. HLADR	1.00
	C:marker_comparisonCK vs. HLADR	1.00
	marker_comparisonCD14 vs. HLADR	1.00
	marker_comparisonCD8 vs. CD14	1.00
	marker_comparisonCD8 vs. HLADR	1.00
	marker_comparisonCK vs. HLADR	1.00
Keren	KS:marker_comparisonCD8 vs. CK	0.85
	KS:marker_comparisonCD8 vs. CD14	0.97
	KS:marker_comparisonCK vs. HLADR	0.98
	marker_comparisonCD8 vs. CD14	0.98
	KS	1.00
	KS:marker_comparisonCD14 vs. HLADR	1.00
	KS:marker_comparisonCD8 vs. HLADR	1.00
	marker_comparisonCD14 vs. HLADR	1.00
	marker_comparisonCD8 vs. HLADR	1.00
Lee	marker_comparisonCD8 vs. CD14	0.83
	L:marker_comparisonCD14 vs. HLADR	0.86
	L:marker_comparisonCD8 vs. CD14	1.00

Table 4.2: **Relevant covariates across prediction models.** Percent of total models that each variable listed is identified as significant ($p < 0.05$) in the cross-validated linear regression models (note that only coefficients that arise in more than 80% of models are included here).

Lastly, we can compare significant model coefficients and the percentage of cross-validated models in which they arise in Table 4.2. The first result of note is that Lee's L-statistic does not arise as a significant coefficient in those models, while both the CMCC and Keren's statistics are significant in all 100 regression models. This, along with the previously discussed results, points to the lack of biologically relevant information maintained in Lee L-statistic and it's underperformance as a predictor of tumor proportion. Focusing on the model coefficients from the CMCC and Keren models, we observe that the following marker co-expression pairs are of most importance: co-expression of all markers CD14 & CD8 & CK with HLADR and co-expression of CD8 with CD14. These not only re-affirm the results of Johnson et al. (2021) but also support the initial correlation results presented in Figure 4.1 for the CMCC.

4.6 Discussion

Here we provided a comprehensive review of spatial statistics methods as implemented in multiplexed imaging, and identified and adapted three methods based on marked point processes for further evaluation. Each of these methods – Lee's L-statistic, the normalized Keren's statistic, and the CMCC – were introduced to capture bivariate spatial measures of marker co-expression in multiplexed imaging data. Specifically, we first showed that Keren's raw statistic is a form of the mark-weighted K-function and introduced a normalized estimate of the Keren statistic to better compare with the other spatial measures discussed here. Further, we adapted Chervoneva's approach of taking the average conditional mark mean in the set of some marked point pattern within a distance d_{max} and introduced an average mark cross-correlation taken over the same set that we interpret as the mean local departure of observed marker co-expression from theoretical independence. Both this CMCC and the normalized Keren's statistic are introduced for the first time here, and compared with the classical Lee's L-statistic for the first time in multiplexed imaging data.

In the Evaluation section we find that Lee's L-statistic is reasonably interpreted as a smoothed spatial correlation coefficient between marker values and captures the marker-marker interactions as expected in the NSCLC dataset. However, the prediction accuracy of this measure is much worse than the other two methods introduced here, and further explorations of the L-statistic in these prediction models points to a lack of importance of this variable as a predictor of the biological outcome of interest. In contrast, we find that Keren's statistic performs quite well in terms of prediction accuracy, and further explorations of this variable in the prediction models show this statistic is important in these models. However, even in its normalized form, we find that Keren's statistic is difficult to interpret and does not reasonably distinguish strong marker-marker interactions like either Lee's L-statistic or the CMCC.

This points to two potential drawbacks of using the Keren's statistic in biological research. The first is that while the Keren's statistic is an intuitive measure of marker-marker interactions, it was originally implemented as a testing procedure to compare the observed value of the statistic to a bootstrapped null distribution. While this is a reasonable method of testing (the `spatstat` R package offers methods for bootstrapping Lee's L-statistic in a similar fashion (Baddeley et al., 2015)), Keren's statistic lacks a reasonable method to distinguish bivariate spatial co-expression patterns between markers *without* some method of testing. Further, the method also relies heavily on thresholding biological markers as marker-positive – broadly, thresholding is an open question in multiplexed imaging data (Bortolomeazzi et al., 2022), and likely requires a discussion about normalization methods rather than a blanket marker-positive threshold (Harris et al., 2022b; Chang et al., 2020).

Due to the concerns with prediction accuracy using Lee's L-statistic and interpretability when using Keren's statistic, we introduced the CMCC to best summarize bivariate spatial co-expression of marker values in multiplexed imaging data. This method combines Chervoneva's approach with the mark cross-correlation statistic to introduce a new spatial index for marker co-expression. Here we have shown that CMCC is correlated, and therefore captures similar spatial information, with the Lee's L-statistic and maintains a similarly simple interpretation of marker-marker interactions. However, the CMCC does not suffer from the same reduced predcition accuracy of the L-statistic and instead achieves a slight improvement in prediction modeling when compared to Keren's statistic. Further, we find that the interpretability of the CMCC as a marker co-expression summary supercedes the unclear definition of Keren's statistic in practice. Hence, we have introduced a new bivariate spatial summary statistic that identifies marker co-expression patterns of interest, provides a good predictive summary of biological outcomes of interest, and ultimately serves as an easy to communicate index for marker-marker interactions in multiplexed imaging data.

There are a handful of reasonable limitations of this chapter that are also worth surfacing. First, due to the difficulty with simulating multiplexed imaging data, we use real data from the NSCLC study to cross-validate results rather than generating theoretical scenarios. In the future, developing simulation methods for multiplexed imaging data would allow for better analysis and comparison of these methods. Further, predictive modeling was performed using linear regression of tumor proportion on a given region (image) – functional data regression as some function of radius or distance could be explored in the future to better understand how these methods interaction. Finally, future work could explore more markers of interest and utilize other biological outcomes like survival endpoints or cancer type to allow for more diverse predictive modeling scenarios.

CHAPTER 5

Conclusion

In Chapter 2, we implemented and compared data transformations and normalization algorithms in multiplexed imaging data. Our methods adapted the ComBat algorithm and functional data registration methods to remove slide effects in this data, and we developed an evaluation framework to compare the proposed approaches. This framework introduced new ideas to the multiplexed imaging field, including the threshold discordance score and multiple methods of measuring slide-to-slide variation. We then present clear systematic variation in the raw, unadjusted data and show that normalizing multiplexed imaging data by its slide mean reduces this variation while preserving and improving the biological signal.

In Chapter 3, we developed the R package, `mxnorm`, to implement, evaluate, and visualize normalization techniques for multiplexed imaging data. This software allows users to extend normalization methods in multiplexed imaging, and provides our robust evaluation framework to measure both technical variability and the efficacy of various normalization methods. Further, the package allows users to supply user-defined normalization methods and thresholding algorithms, introducing a platform for comparing the utility of different normalization techniques.

Chapter 4 introduces spatial statistics methods in multiplexed imaging data to understand the spatial relationship of different biological markers following the removal of slide-to-slide variation. Here we leveraged marked point process theory to derive and compare three spatial statistics methods: Keren's permutation test for spatial proximity, Lee's L test for bivariate spatial data, and a proposed cumulative mark cross-correlation (CMCC) statistic. We then provide an evaluation of the adapted methods – a correlation analysis to quantify the amount of shared information between the statistics and a cross-validated prediction modeling analysis to determine each measure's prediction accuracy.

Altogether, this dissertation addresses two major components of the multiplexed imaging pipeline – removing systematic noise from this data and adapting spatial analysis methods in this field. The methods introduced here are important for determining and maintaining data quality to leverage this valuable data resource for future biological research. We hope this work provides the foundation (and inspiration) for introducing and evaluating normalization techniques and spatial analysis methods across the multiplexed imaging pipeline.

CHAPTER 6

Appendix

6.1 Application of ComBat

Note from Chapter 2 of this thesis that we have assumed that the standardized data $Z_{ic}(u) \sim N(\gamma_{ic}, \delta^2_{ic})$ with the following priors on the batch effects:

$$\gamma_{ic} \sim N(\gamma_c, \tau^2_c), \delta^2_{ic} \sim IG(\omega_c, \beta_c)$$

Recall that i denotes the slide from which the data was collected, c denotes the marker of interest, and u defines the unit of measuring intensity, which for this study is the median quantified marker intensity of the segmented cell. Note also that we defined $U_{ic} = \sum_{ic} u$, or the number of quantified cells present on a particular slide i for a given channel c.

6.1.1 Posterior Derivation for γ_{ic}

Using the empirical Bayes methodology, we must derive the posterior mean estimator of γ_{ic} to utilize in the ComBat model. Hence:

$$\begin{aligned}
\pi\left(\gamma_{ic}|Z_{ic}(u), \delta^2_{ic}\right) &= L\left(Z_{ic}(u)|\gamma_{ic}, \delta^2_{ic}\right) \cdot \pi(\gamma_{ic}) \\
&\propto \exp\left\{-\frac{1}{2\delta^2_{ic}}\sum_u (Z_{ic}(u) - \gamma_{ic})^2\right\} \cdot \exp\left\{-\frac{1}{2\tau^2_c}(\gamma_{ic} - \gamma_c)^2\right\} \\
&= \exp\left\{-\frac{1}{2\delta^2_{ic}}\left(\sum_u Y^2_{ic}(u) - 2\sum_u Z_{ic}(u)\gamma_{ic} + U_{ic} \cdot \gamma^2_{ic}\right) - \frac{1}{2\tau^2_c}\left(\gamma^2_{ic} - 2\gamma_{ic}\gamma_c + \gamma^2_c\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\frac{U_{ic}\tau^2_c + \delta^2_{ic}}{\delta^2_{ic}\tau^2_c}\right)\left[\gamma^2_{ic} - 2\left(\frac{\tau^2_c \sum_u Z_{ic}(u) + \delta^2_{ic}\gamma_c}{U_{ic}\tau^2_c + \delta^2_{ic}}\right)\gamma_{ic}\right]\right\}
\end{aligned}$$

Which after we complete the square, we see this is posterior follows the Normal distribution with the following expectation:

$$E\left[\gamma_{ic}|Z_{ic}(u), \delta^2_{ic}\right] = \frac{\tau^2_c \sum_u Z_{ic}(u) + \delta^2_{ic}\gamma_c}{U_{ic}\tau^2_c + \delta^2_{ic}}$$

To derive an estimator of the batch effect parameter, we must define the following estimators of the hyperparameters:

$$\bar{\gamma}_c = \frac{1}{U_{ic}}\sum_i \hat{\gamma}_{ic} \text{ and, } \bar{\tau}^2_c = \frac{1}{U_{ic} - 1}\sum_i (\hat{\gamma}_{ic} - \bar{\gamma}_c)^2$$

Hence we now derive the following estimator of γ_{ic}:

$$\gamma_{ic}^{*} = \frac{\bar{\tau}_c^2 U_{ic}\hat{\gamma}_{ic} + \delta_{ic}^{2*}\bar{\gamma}_c}{U_{ic}\bar{\tau}_c^2 + \delta_{ic}^{2*}}$$

6.1.2 Posterior Derivation for δ_{ic}^2

We employ the same methodology to derive the posterior mean estimator of δ_{ic}^2:

$$\begin{aligned}
\pi\left(\delta_{ic}^2|Z_{ic}(u),\gamma_{ic}\right) &= L\left(Z_{ic}(u)|\gamma_{ic},\delta_{ic}^2\right)\cdot\pi(\delta_{ic}^2) \\
&\propto \delta_{ic}^{2^{-\frac{U_{ic}}{2}}} \exp\left\{-\frac{1}{2\delta_{ic}^2}\sum_u (Z_{ic}(u)-\gamma_{ic})^2\right\}\cdot\delta_{ic}^{2^{-(\omega_c+1)}}\exp\left\{-\frac{\beta_c}{\delta_{ic}^2}\right\} \\
&= \delta_{ic}^{2^{-\left(\left[\frac{U_{ic}}{2}+\omega_c\right]+1\right)}}\exp\left\{-\frac{1}{2\delta_{ic}^2}\left(\sum_u Y_{ic}^2(u) - 2\sum_u Z_{ic}(u)\gamma_{ic} + U_{ic}\cdot\gamma_{ic}^2\right)-\frac{\beta_c}{\delta_{ic}^2}\right\} \\
&\propto \delta_{ic}^{2^{-\left(\left[\frac{U_{ic}}{2}+\omega_c\right]+1\right)}}\exp\left\{-\frac{1}{\delta_{ijc}^2}\left(\beta_c+\frac{1}{2}\sum_u (Z_{ic}(u)-\gamma_{ic})^2\right)\right\}
\end{aligned}$$

Which we note is an Inverse Gamma distribution with the following expectation:

$$E\left[\delta_{ic}^2|Z_{ic}(u),\gamma_{ic}\right] = \frac{\beta_c + \frac{1}{2}\sum_u (Z_{ic}(u)-\gamma_{ic})^2}{\frac{U_{ic}}{2}+\omega_c - 1}$$

To derive an estimator of the batch effect parameter, we must define the following estimators:

$$\hat{\delta}_{ic}^2 = \frac{1}{U_{ic}-1}\sum_u (Z_{ic}(u)-\hat{\gamma}_{ic})^2$$

We then calculate the sample mean of the $\hat{\delta}_{ic}^2$, $\bar{M}_c$ and $\bar{S}_c^2$ and set these equal to the moments of an Inverse Gamma distribution to yield the following estimators:

$$\bar{\omega}_c = \frac{\bar{M}_c + 2\bar{S}_c^2}{\bar{S}_c^2} \text{ and, } \bar{\beta}_c = \frac{\bar{M}_c^3 + \bar{M}_c\bar{S}_c^2}{\bar{S}_c^2}$$

Hence we now derive the following estimator of δ_{ic}^2:

$$\delta_{ic}^{2*} = \frac{\bar{\beta}_c + \frac{1}{2}\sum_u (Z_{ic}(u)-\hat{\gamma}_{ic})^2}{\frac{U_{ic}}{2}+\bar{\omega}_c - 1}$$

References

Angelo, M., Bendall, S. C., Finck, R., Hale, M. B., Hitzman, C., Borowsky, A. D., Levenson, R. M., Lowe, J. B., Liu, S. D., Zhao, S., et al. (2014). Multiplexed ion beam imaging of human breast tumors. *Nature Medicine*, 20(4):436–442.

Baddeley, A., Rubak, E., and Turner, R. (2015). *Spatial Point Patterns: Methodology and Applications with R.* Chapman and Hall/CRC Press, London.

Bates, D., Maechler, M., and Bolker, B. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W., Ng, L. G., Ginhoux, F., and Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology*, 37(1):38–44.

Berry, S., Giraldo, N. A., Green, B. F., Cottrell, T. R., Stein, J. E., Engle, E. L., Xu, H., Ogurtsova, A., Roberts, C., Wang, D., et al. (2021). Analysis of multispectral imaging with the astropath platform informs efficacy of pd-1 blockade. *Science*, 372(6547).

Blom, S., Paavolainen, L., Bychkov, D., Turkki, R., Mäki-Teeri, P., Hemmes, A., Välimäki, K., Lundin, J., Kallioniemi, O., and Pellinen, T. (2017). Systems pathology by multiplexed immunohistochemistry and whole-slide digital image analysis. *Scientific Reports*, 7(1):1–13.

Bortolomeazzi, M., Montorsi, L., Temelkovski, D., Keddar, M. R., Acha-Sagredo, A., Pitcher, M. J., Basso, G., Laghi, L., Rodriguez-Justo, M., Spencer, J., et al. (2022). A simpli (single-cell identification from multiplexed images) approach for spatially-resolved tissue phenotyping at single-cell resolution. *Nature communications*, 13(1):1–14.

Bradford, J. A., Buller, G., Suter, M., Ignatius, M., and Beechem, J. M. (2004). Fluorescence-intensity multiplexing: Simultaneous seven-marker, two-color immunophenotyping using flow cytometry. *Cytometry Part A: The Journal of the International Society for Analytical Cytology*, 61(2):142–152.

Brown, E. N., Kass, R. E., and Mitra, P. P. (2004). Multiple neural spike train data analysis: state-of-the-art and future challenges. *Nature neuroscience*, 7(5):456–461.

Burlingame, E. A., Eng, J., Thibault, G., Chin, K., Gray, J. W., and Chang, Y. H. (2021). Toward reproducible, scalable, and robust data analysis across multiplex tissue imaging platforms. *Cell Reports Methods*, 1(4):100053.

Chang, Y. H., Chin, K., Thibault, G., Eng, J., Burlingame, E., and Gray, J. W. (2020). Restore: Robust intensity normalization method for multiplexed imaging. *Communications Biology*, 3(1):1–9.

Chen, B., Cherie'R, S., McKinley, E. T., Simmons, A. J., Ramirez-Solano, M. A., Zhu, X., Markham, N. O., Heiser, C. N., Vega, P. N., Rolong, A., et al. (2021). Differential pre-malignant programs and microenvironment chart distinct paths to malignancy in human colorectal polyps. *Cell*, 184(26):6262–6280.

Chen, G., Ning, B., and Shi, T. (2019). Single-cell rna-seq technologies and related computational data analysis. *Frontiers in genetics*, 10:317.

Chen, Z., Soifer, I., Hilton, H., Keren, L., and Jojic, V. (2020). Modeling multiplexed images with spatial-lda reveals novel tissue microenvironments. *Journal of Computational Biology*, 27(8):1204–1218.

Chervoneva, I., Peck, A. R., Yi, M., Freydin, B., and Rui, H. (2021). Quantification of spatial tumor heterogeneity in immunohistochemistry staining images. *Bioinformatics*, 37(10):1452–1460.

Creed, J. H., Wilson, C. M., Soupir, A. C., Colin-Leitzinger, C. M., Kimmel, G. J., Ospina, O. E., Chakiryan, N. H., Markowitz, J., Peres, L. C., Coghill, A., et al. (2021). spatialTIME and iTIME: R package and shiny application for visualization and analysis of immunofluorescence data. *Bioinformatics*, 37(23):4584–4586.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Dries, R., Chen, J., Del Rossi, N., Khan, M. M., Sistig, A., and Yuan, G.-C. (2021a). Advances in spatial transcriptomic data analysis. *Genome Research*, 31(10):1706–1718.

Dries, R., Zhu, Q., Dong, R., Eng, C.-H. L., Li, H., Liu, K., Fu, Y., Zhao, T., Sarkar, A., Bao, F., et al. (2021b). Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biology*, 22(1):1–31.

Edsgärd, D., Johnsson, P., and Sandberg, R. (2018). Identification of spatial expression trends in single-cell gene expression data. *Nature methods*, 15(5):339–342.

Eling, N., Damond, N., Hoch, T., and Bodenmiller, B. (2020). cytomapper: an R/Bioconductor package for visualization of highly multiplexed imaging data. *Bioinformatics*, 36(24):5706–5708.

Elosua-Bayes, M., Nieto, P., Mereu, E., Gut, I., and Heyn, H. (2021). Spotlight: seeded nmf regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic acids research*, 49(9):e50–e50.

Engle, R. F. and Lunde, A. (2003). Trades and quotes: a bivariate point process. *Journal of Financial Econometrics*, 1(2):159–188.

Fortin, J.-P., Parker, D., Tunç, B., Watanabe, T., Elliott, M. A., Ruparel, K., Roalf, D. R., Satterthwaite, T. D., Gur, R. C., Gur, R. E., et al. (2017). Harmonization of multi-site diffusion tensor imaging data. *Neuroimage*, 161:149–170.

Galon, J., Costes, A., Sanchez-Cabo, F., Kirilovsky, A., Mlecnik, B., Lagorce-Pagès, C., Tosolini, M., Camus, M., Berger, A., Wind, P., et al. (2006). Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science*, 313(5795):1960–1964.

Gerdes, M. J., Sevinsky, C. J., Sood, A., Adak, S., Bello, M. O., Bordwell, A., Can, A., Corwin, A., Dinn, S., Filkins, R. J., et al. (2013). Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue. *Proceedings of the National Academy of Sciences*, 110(29):11982–11987.

Goldsmith, J., Scheipl, F., Huang, L., Wrobel, J., Di, C., Gellar, J., Harezlak, J., McLean, M. W., Swihart, B., Xiao, L., Crainiceanu, C., and Reiss, P. T. (2021). *refund: Regression with Functional Data*. R package version 0.1-24.

Goltsev, Y., Samusik, N., Kennedy-Darling, J., Bhate, S., Hale, M., Vazquez, G., Black, S., and Nolan, G. P. (2018). Deep profiling of mouse splenic architecture with codex multiplexed imaging. *Cell*, 174(4):968–981.

Gomariz, A., Portenier, T., Helbling, P. M., Isringhausen, S., Suessbier, U., Nombela-Arrieta, C., and Goksel, O. (2021). Modality attention and sampling enables deep learning with heterogeneous marker combinations in fluorescence microscopy. *Nature machine intelligence*, 3(9):799–811.

Graf, J., Cho, S., McDonough, E., Corwin, A., Sood, A., Lindner, A., Salvucci, M., Stachtea, X., Van Schaeybroeck, S., Dunne, P. D., et al. (2022). Flino: a new method for immunofluorescence bioimage normalization. *Bioinformatics*, 38(2):520–526.

Hao, M., Hua, K., and Zhang, X. (2021a). Somde: a scalable method for identifying spatially variable genes with self-organizing map. *Bioinformatics*, 37(23):4392–4398.

Hao, Y., Hao, S., Andersen-Nissen, E., III, W. M. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zagar, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. B., Yeung, B., Rogers, A. J., McElrath, J. M., Blish, C. A., Gottardo, R., Smibert, P., and Satija, R. (2021b). Integrated analysis of multimodal single-cell data. *Cell*.

Harris, C. (2022). *mxnorm: Apply Normalization Methods to Multiplexed Images*. R package version 1.0.1.

Harris, C., Wrobel, J., and Vandekar, S. (2022a). mxnorm: An r package to normalize multiplexed imaging data. *Journal of Open Source Software*, 7(71):4180.

Harris, C. R., McKinley, E. T., Roland, J. T., Liu, Q., Shrubsole, M. J., Lau, K. S., Coffey, R. J., Wrobel, J., and Vandekar, S. N. (2022b). Quantifying and correcting slide-to-slide variation in multiplexed immunofluorescence images. *Bioinformatics*, 38(6):1700–1707.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108.

Hashmi, A. A., Hussain, Z. F., Aijaz, S., Irfan, M., Khan, E. Y., Naz, S., Faridi, N., Khan, A., and Edhi, M. M. (2018). Immunohistochemical expression of epidermal growth factor receptor (egfr) in south asian head and neck squamous cell carcinoma: association with various risk factors and clinico-pathologic and prognostic parameters. *World journal of surgical oncology*, 16(1):1–9.

Hie, B., Bryson, B., and Berger, B. (2019). Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nature Biotechnology*, 37(6):685–691.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1):193–218.

Ijsselsteijn, M. E., van der Breggen, R., Farina Sarasqueta, A., Koning, F., and de Miranda, N. F. (2019). A 40-marker panel for high dimensional characterization of cancer immune microenvironments by imaging mass cytometry. *Frontiers in immunology*, 10:2534.

Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. (2008). *Statistical analysis and modelling of spatial point patterns*, volume 70. John Wiley & Sons.

Islam, M., Chen, B., Spraggins, J. M., Kelly, R. T., and Lau, K. S. (2020). Use of single-cell-omic technologies to study the gastrointestinal tract and diseases, from single cell identities to patient features. *Gastroenterology*, 159(2):453–466.

Johnson, A. M., Boland, J. M., Wrobel, J., Klezcko, E. K., Weiser-Evans, M., Hopp, K., Heasley, L., Clambey, E. T., Jordan, K., Nemenoff, R. A., et al. (2021). Cancer cell-specific major histocompatibility complex ii expression as a determinant of the immune infiltrate organization and function in the nsclc tumor microenvironment. *Journal of Thoracic Oncology*, 16(10):1694–1704.

Johnson, K. S. (2022). *phenoptrReports: Create reports using Phenoptics data*. https://akoyabio.github.io/phenoptrReports/, https://github.com/akoyabio/phenoptrReports/.

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127.

Keren, L., Bosse, M., Marquez, D., Angoshtari, R., Jain, S., Varma, S., Yang, S.-R., Kurian, A., Van Valen, D., West, R., et al. (2018). A structured tumor-immune microenvironment in triple negative breast cancer revealed by multiplexed ion beam imaging. *Cell*, 174(6):1373–1387.

Law, R., Illian, J., Burslem, D. F., Gratzer, G., Gunatilleke, C., and Gunatilleke, I. (2009). Ecological information from spatial patterns of plants: insights from point process theory. *Journal of Ecology*, 97(4):616–628.

Lee, S.-I. (2001). Developing a bivariate spatial association measure: an integration of pearson's r and moran's i. *Journal of geographical systems*, 3(4):369–385.

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., and Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6):882–883.

Maric, D., Jahanipour, J., Li, X. R., Singh, A., Mobiny, A., Van Nguyen, H., Sedlock, A., Grama, K., and Roysam, B. (2021). Whole-brain tissue mapping toolkit using large-scale highly multiplexed immunofluorescence imaging and deep neural networks. *Nature communications*, 12(1):1–12.

Marron, J. S., Ramsay, J. O., Sangalli, L. M., and Srivastava, A. (2015). Functional data analysis of amplitude and phase variation. *Statistical Science*, pages 468–484.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

McKinley, E. T., Shao, J., Ellis, S. T., Heiser, C. N., Roland, J. T., Macedonia, M. C., Vega, P. N., Shin, S., Coffey, R. J., and Lau, K. S. (2022). Miriam: A machine and deep learning single-cell segmentation and quantification pipeline for multi-dimensional tissue images. *Cytometry Part A*.

Melville, J. (2021). *uwot: The Uniform Manifold Approximation and Projection (UMAP) Method for Dimensionality Reduction*. R package version 0.1.11.

Moran, P. A. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):243–251.

Nadarajan, G., Hope, T., Wang, D., Cheung, A., Ginty, F., Yaffe, M. J., and Doyle, S. (2019). Automated multi-class ground-truth labeling of h&e images for deep learning using multiplexed fluorescence microscopy. In *Medical Imaging 2019: Digital Pathology*, volume 10956, page 109560J. International Society for Optics and Photonics.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66.

Palla, G., Spitzer, H., Klein, M., Fischer, D., Schaar, A. C., Kuemmerle, L. B., Rybakov, S., Ibarra, I. L., Holmberg, O., Virshup, I., et al. (2022). Squidpy: a scalable framework for spatial omics analysis. *Nature methods*, 19(2):171–178.

Ptacek, J., Locke, D., Finck, R., Cvijic, M.-E., Li, Z., Tarolli, J. G., Aksoy, M., Sigal, Y., Zhang, Y., Newgren, M., et al. (2020). Multiplexed ion beam imaging (mibi) for characterization of the tumor microenvironment across tumor types. *Laboratory Investigation*, 100(8):1111–1123.

Ramsay, J., Graves, S., and Hooker, G. (2020). Package 'fda'. *CRAN*.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer.

Raza, S. E. A., Langenkämper, D., Sirinukunwattana, K., Epstein, D., Nattkemper, T. W., and Rajpoot, N. M. (2016). Robust normalization protocols for multiplexed fluorescence bioimage analysis. *BioData mining*, 9(1):1–13.

Revelle, W. (2021). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 2.1.9.

Ripley, B. D. (1976). The second-order analysis of stationary point processes. *Journal of applied probability*, 13(2):255–266.

Ripley, B. D. (1988). *Statistical inference for spatial processes*. Cambridge university press.

Rozenblatt-Rosen, O., Regev, A., Oberdoerffer, P., Nawy, T., Hupalowska, A., Rood, J. E., Ashenberg, O., Cerami, E., Coffey, R. J., Demir, E., et al. (2020). The human tumor atlas network: charting tumor transitions across space and time at single-cell resolution. *Cell*, 181(2):236–249.

Schapiro, D., Jackson, H. W., Raghuraman, S., Fischer, J. R., Zanotelli, V. R., Schulz, D., Giesen, C., Catena, R., Varga, Z., and Bodenmiller, B. (2017). histocat: analysis of cell phenotypes and interactions in multiplex image cytometry data. *Nature methods*, 14(9):873–876.

Schapiro, D., Sokolov, A., Yapp, C., Chen, Y.-A., Muhlich, J. L., Hess, J., Creason, A. L., Nirmal, A. J., Baker, G. J., Nariya, M. K., et al. (2022). Mcmicro: A scalable, modular image-processing pipeline for multiplexed tissue imaging. *Nature methods*, 19(3):311–315.

Schlather, M. (2001). On the second-order characteristics of marked point processes. *Bernoulli*, pages 99–117.

Schlather, M., Ribeiro Jr, P. J., and Diggle, P. J. (2004). Detecting dependence between marks and locations of marked point processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):79–93.

Scholz, F. W. and Stephens, M. A. (1987). K-sample anderson–darling tests. *Journal of the American Statistical Association*, 82(399):918–924.

Scott, C. E., Wynn, S. L., Sesay, A., Cruz, C., Cheung, M., Gaviro, M.-V. G., Booth, S., Gao, B., Cheah, K. S., Lovell-Badge, R., et al. (2010). Sox9 induces and maintains neural stem cells. *Nature neuroscience*, 13(10):1181–1189.

Shan, Z.-Z., Chen, P.-N., Wang, F., Wang, J., and Fan, Q.-X. (2017). Expression of p-egfr and p-akt protein in esophageal squamous cell carcinoma and its prognosis. *Oncology letters*, 14(3):2859–2863.

Shang, S., Hua, F., and Hu, Z.-W. (2017). The regulation of β-catenin activity and function in cancer: therapeutic opportunities. *Oncotarget*, 8(20):33972.

Shimatani, K. (2002). Point processes for fine-scale spatial genetics and molecular ecology. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 44(3):325–352.

Shimatani, K. and Kubota, Y. (2004). Quantitative assessment of multispecies spatial pattern with high species diversity. *Ecological Research*, 19(2):149–163.

Shrubsole, M. J., Wu, H., Ness, R. M., Shyr, Y., Smalley, W. E., and Zheng, W. (2008). Alcohol drinking, cigarette smoking, and risk of colorectal adenomatous and hyperplastic polyps. *American journal of epidemiology*, 167(9):1050–1058.

Smyth, G. K. (2005). Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer.

Stoltzfus, C. R., Filipek, J., Gern, B. H., Olin, B. E., Leal, J. M., Wu, Y., Lyons-Cohen, M. R., Huang, J. Y., Paz-Stoltzfus, C. L., Plumlee, C. R., et al. (2020). Cytomap: a spatial analysis toolbox reveals features of myeloid cell organization in lymphoid tissues. *Cell reports*, 31(3):107523.

Sun, S., Zhu, J., and Zhou, X. (2020). Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nature methods*, 17(2):193–200.

Svensson, V., Teichmann, S. A., and Stegle, O. (2018). Spatialde: identification of spatially variable genes. *Nature methods*, 15(5):343–346.

Trinh, A., Trumpi, K., Felipe De Sousa, E. M., Wang, X., De Jong, J. H., Fessler, E., Kuppen, P. J., Reimers, M. S., Swets, M., Koopman, M., et al. (2017). Practical and robust identification of molecular subtypes in colorectal cancer by immunohistochemistry. *Clinical Cancer Research*, 23(2):387–398.

Tsujikawa, T., Kumar, S., Borkar, R. N., Azimi, V., Thibault, G., Chang, Y. H., Balter, A., Kawashima, R., Choe, G., Sauer, D., et al. (2017). Quantitative multiplex immunohistochemistry reveals myeloid-inflamed tumor-immune complexity associated with poor prognosis. *Cell reports*, 19(1):203–217.

Tsujikawa, T., Thibault, G., Azimi, V., Sivagnanam, S., Banik, G., Means, C., Kawashima, R., Clayburgh, D. R., Gray, J. W., Coussens, L. M., et al. (2019). Robust cell detection and segmentation for image cytometry reveal th17 cell heterogeneity. *Cytometry Part A*, 95(4):389–398.

Van der Flier, L. G., Haegebarth, A., Stange, D. E., Van de Wetering, M., and Clevers, H. (2009). Olfm4 is a robust marker for stem cells in human intestine and marks a subset of colorectal cancer cells. *Gastroenterology*, 137(1):15–17.

van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., Yu, T., and the scikit-image contributors (2014). scikit-image: image processing in Python. *PeerJ*, 2:e453.

Vavrek, M. J. (2011). fossil: palaeoecological and palaeogeographical analysis tools. *Palaeontologia Electronica*, 14(1):1T. R package version 0.4.0.

Vu, T., Wrobel, J., Bitler, B. G., Schenk, E. L., Jordan, K. R., and Ghosh, D. (2021). Spf: A spatial and functional data analytic approach to cell imaging data. *bioRxiv*.

Wartenberg, D. (1985). Multivariate spatial correlation: a method for exploratory geographical analysis. *Geographical Analysis*, 17(4):263–283.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

Wilson, C. M., Ospina, O. E., Townsend, M. K., Nguyen, J., Moran Segura, C., Schildkraut, J. M., Tworoger, S. S., Peres, L. C., and Fridley, B. L. (2021). Challenges and opportunities in the statistical analysis of multiplex immunofluorescence data. *Cancers*, 13(12):3031.

Windhager, J., Bodenmiller, B., and Eling, N. (2021). An end-to-end workflow for multiplexed image processing and analysis. *bioRxiv*.

Wrobel, J. (2021). *mica: Multi Image CDF Alignment (or multi site intensity harmonization by CDF alignment)*. R package version 0.1.0.

Wrobel, J., Bauer, A., Goldsmith, J., and McDonnell, E. (2021). *registr: Curve Registration for Exponential Family Functional Data*. R package version 2.0.7.

Wrobel, J., Martin, M., Bakshi, R., Calabresi, P., Elliot, M., Roalf, D., Gur, R., Gur, R., Henry, R., Nair, G., et al. (2020). Intensity warping for multisite mri harmonization. *NeuroImage*, 223:117242.

Wrobel, J., Zipunnikov, V., Schrack, J., and Goldsmith, J. (2019). Registration for exponential family functional data. *Biometrics*, 75(1):48–57.

Wu, S., Joseph, A., Hammonds, A. S., Celniker, S. E., Yu, B., and Frise, E. (2016). Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks. *Proceedings of the National Academy of Sciences*, 113(16):4290–4295.

Yapp, C., Novikov, E., Jang, W.-D., Chen, Y.-A., Cicconet, M., Maliga, Z., Jacobson, C. A., Wei, D., Santagata, S., Pfister, H., et al. (2021). Unmicst: Deep learning with real augmentation for robust segmentation of highly multiplexed images of human tissues. *bioRxiv*.

Yu, M., Linn, K. A., Cook, P. A., Phillips, M. L., McInnis, M., Fava, M., Trivedi, M. H., Weissman, M. M., Shinohara, R. T., and Sheline, Y. I. (2018). Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fmri data. *Human brain mapping*, 39(11):4213–4227.

Zhang, Y., Parmigiani, G., and Johnson, W. E. (2020). Combat-seq: batch effect adjustment for rna-seq count data. *NAR genomics and bioinformatics*, 2(3):lqaa078.

Zhao, E., Stone, M. R., Ren, X., Guenthoer, J., Smythe, K. S., Pulliam, T., Williams, S. R., Uytingco, C. R., Taylor, S. E., Nghiem, P., et al. (2021). Spatial transcriptomics at subspot resolution with bayesspace. *Nature biotechnology*, 39(11):1375–1384.

Zhu, J., Sun, S., and Zhou, X. (2021). Spark-x: non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies. *Genome Biology*, 22(1):1–25.

Zhu, Q., Shah, S., Dries, R., Cai, L., and Yuan, G.-C. (2018). Identification of spatially associated subpopulations by combining scrnaseq and sequential fluorescence in situ hybridization data. *Nature biotechnology*, 36(12):1183–1190.