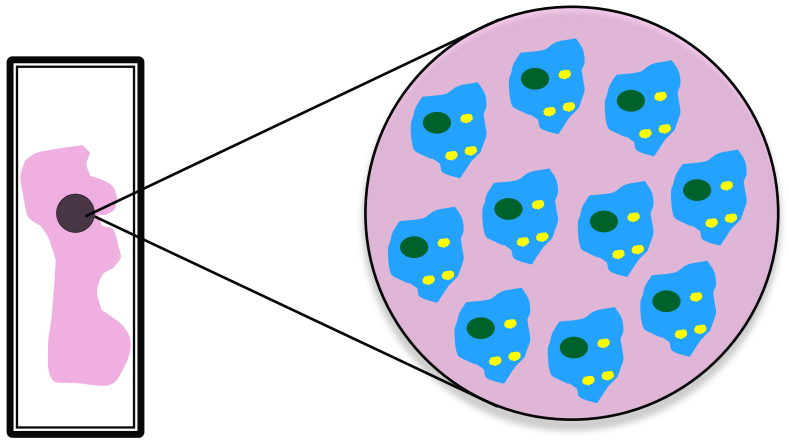
Statistical analysis of multiplex single-cell imaging data

Julia Wrobel, PhD

Department of Biostatistics and Bioinformatics

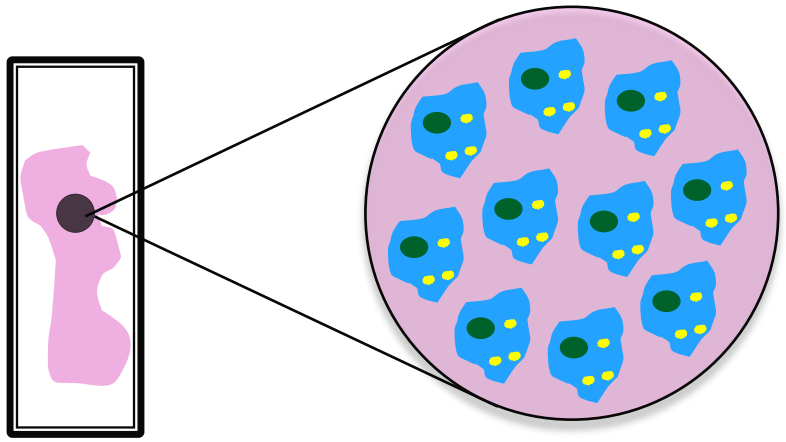
What is single cell multiplex imaging?

- High dimensional analysis of tissue samples at the resolution of individual cells



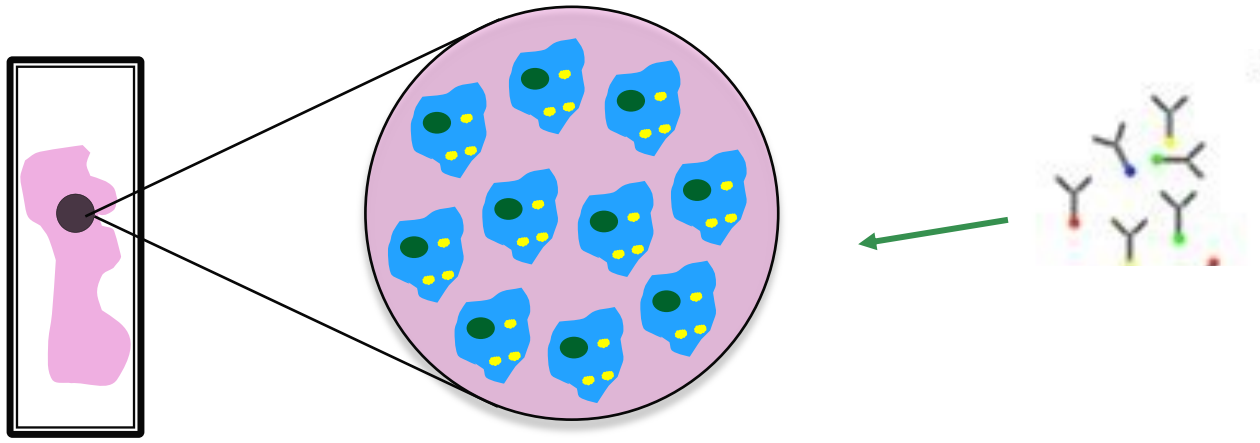
What is single cell multiplex imaging?

- *Single cell* refers to individual cell resolution



What is single cell multiplex imaging?

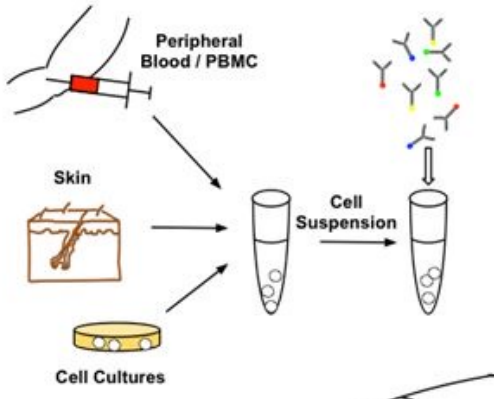
- ***Multiplex*** refers to multiple types of protein in the tissue that are tagged

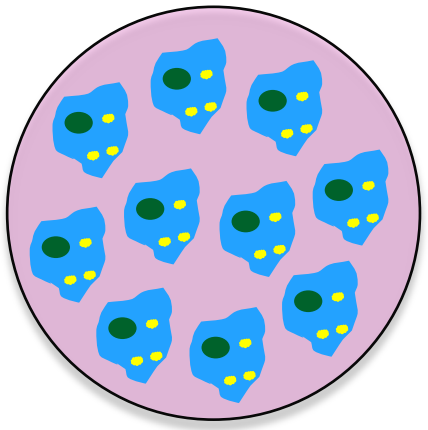


- Each protein label is called a *marker*
  - *Phenotypic markers*: used to define cell and/or tissue type
  - *Functional markers*: inform cell function
    - *Present across multiple cell types*

What is single cell multiplex imaging?

- *Imaging* means spatial relationships in the tissue are preserved

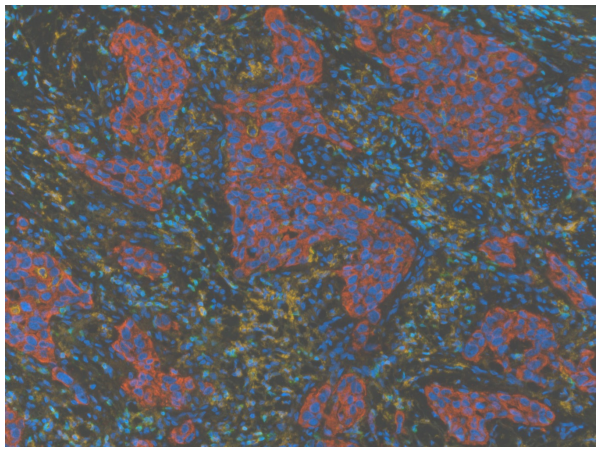
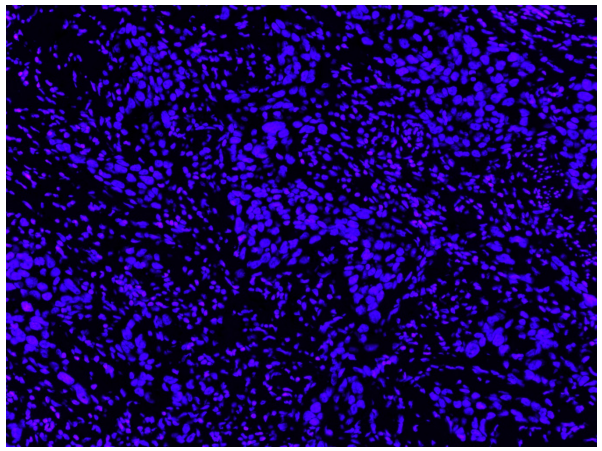
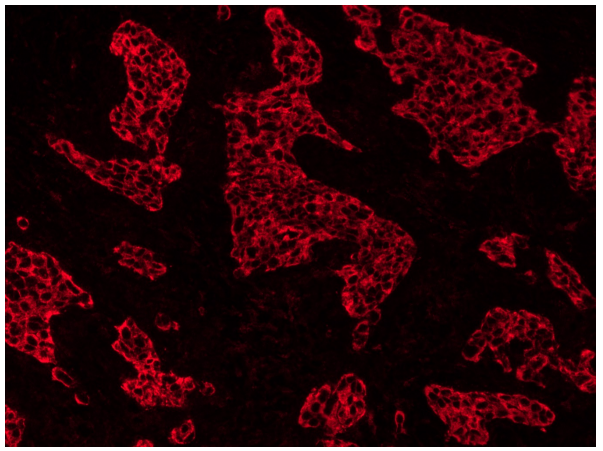
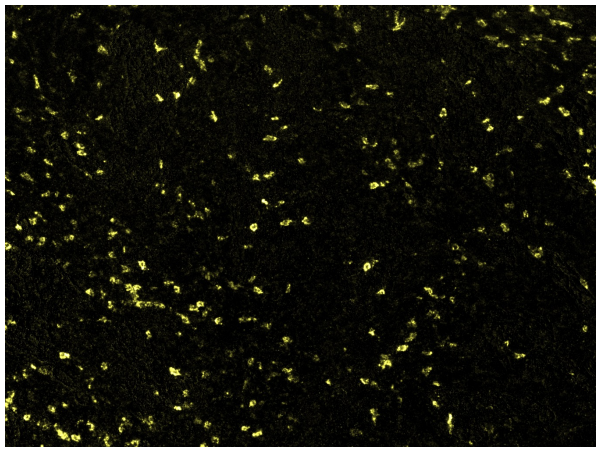




- Precursor technologies (left) required suspension of cells in solution, destroying spatial info

What is single cell multiplex tissue imaging?

- Images produced are multichannel TIFF (.tif) files
  - Each channel is a different protein marker
  - Each pixel contains a continuous intensity value for each marker
- Example below with non-small cell lung cancer data
  - 8 channels, 3 shown (Left to Right: composite image, nuclei, CK, CD8)

Cancer, multiplex imaging, and the tumor microenvironment

The tumor microenvironment (TME) is the area within and surrounding a tumor, including tumor cells, infiltrating immune cells, blood vessels, and other tissue

- What percentages of immune cell subtypes are present before and after chemotherapy?
- Do patients with high spatial clustering of B-cells and Macrophages survive longer?

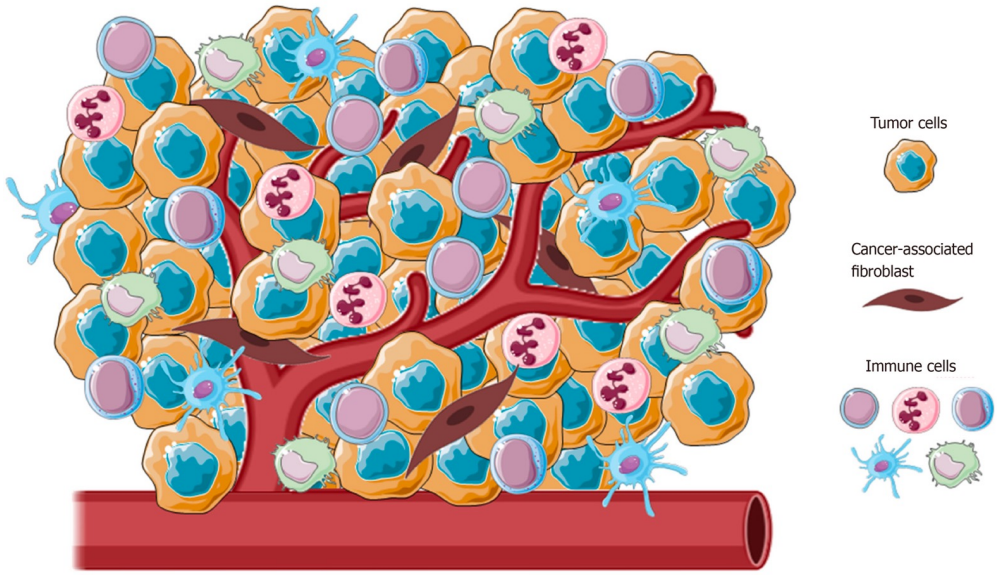


Image from Polidoro et. al. *World J. Gasteroentrol.*, 26(33) (2020)

A brief comparison of scMI techniologies

Platform	Class	Throughput	Multiplexicity	Software	Publications*
Vectra-Polaris	IF	high	~8 markers	Proprietary	> 100
Discovery Ultra	IF	high	5+ markers	Limited	< 10
CyTOF Imaging	IMC	low	37+	MatLab	> 35
MIBI	IMC	low	40+	Limited	>5
CODEX	Barcode-based	low	40+	Limited/proprietary	>1

* Up to April 2020

Table from Chang et. al. *Cancer Communications*, 40(4) (2020)

Comparing Spatial Transciptomics and scMI

Spatial Transciptomics

- Spatial analog of scRNA-seq
  - Targets mRNA
  - n << p, features are genes
- Lower spatial resolution
  - Not always single-cell
- Lower throughput

Within-sample focus

- Discovering spatially variable genes
- Understanding tumor topography

Multiplex Imaging

- Spatial analog of flow/mass cytometry
  - Targets protein
  - n > p, features are protein markers
- Higher spatial resolution
  - Truly single-cell
- High(er) throughput

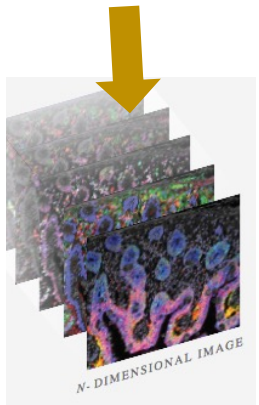
Across-image focus

- Modeling patient-level outcomes
- How well to image features predict overall survival?

scMI data processing and analysis pipeline

Image acquisition

Image filtering/ background correction



Image/ marker channel normalization

Tissue, cell, and nucleus segmentation

Cell phenotyping

Statistical image processing

Image ID	Cell ID	Tissue type	Phenotype	X	Y	Patient features
1	1	Stroma	CD4 T-cell	…	…	…
1	2	Stroma	B-cell	…	…	…
1	3	Tumor	Macrophage	…	…	…
1	4	Stroma	CD8 T-cell	…	…	…
2	1	Tumor	Tumor cell	…	…	…

Statistical analysis

Statistical software development

Statistical image processing

Statistical image processing

- Pixel level processing: work with multichannel tiff files directly
  - Operates on pixel intensity values
- Cell-level processing: work with tabular data after cell segmentation
  - Operates on median or mean intensity values aggregated at cell level

- Segmentation
  - Pixel-level
- Normalization
  - Pixel-level or cell-level
- Phenotyping
  - Pixel-level or cell-level

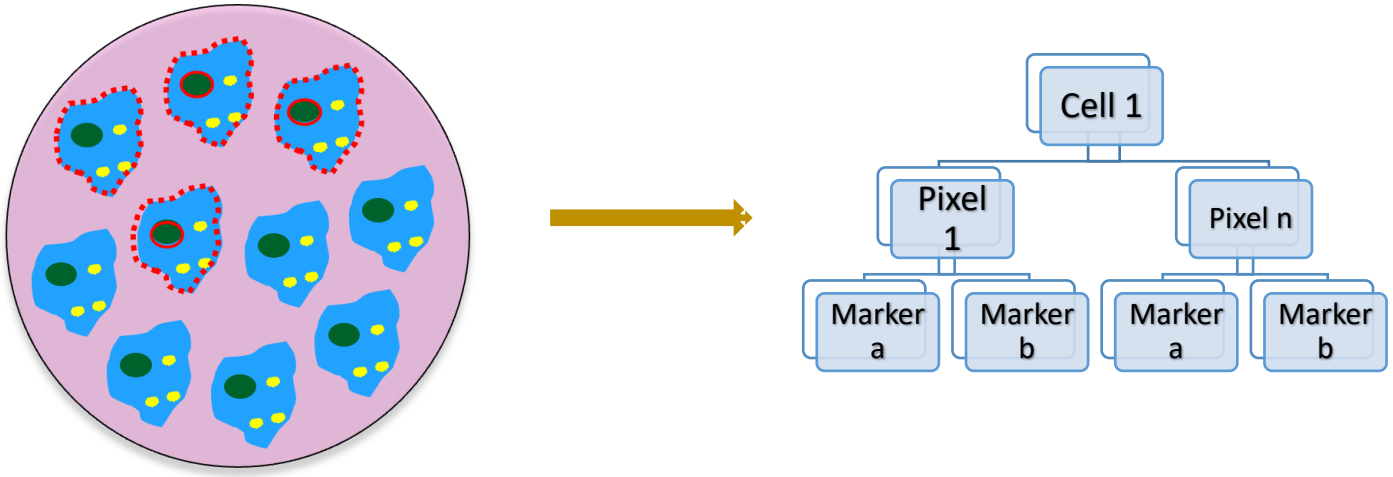
Statistical image processing

Cell segmentation

Cell segmentation

Identifies cells and nuclei in image

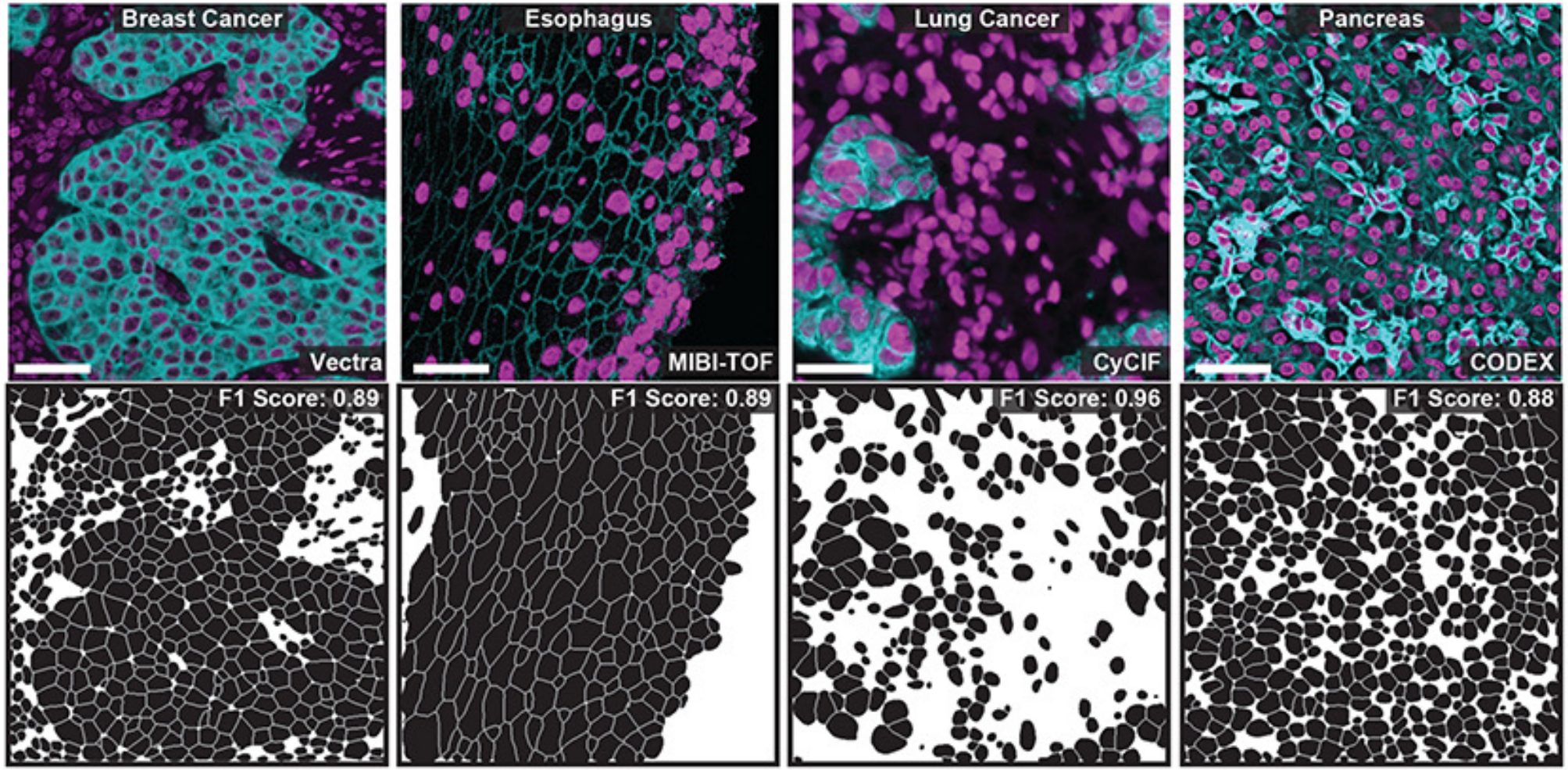
1. Nucleus channel used to identify nucleus
2. Cell membrane or cytoplasm markers used to draw boundary around nucleus



Segmentation approaches

- Proprietary software
  - inForm by Akoya, Halo automated image analysis software
  - GUI-based, user-friendly
  - No manually segmented data required
- GUI-based open-source software
  - CellProfiler, ilastik, QuPath
  - User-friendly for non-computer scientists/statisticians
  - No manually segmented data required*
- Deep-learning based open-source software
  - Best performance*
  - Need manually segmented data*
  - Hard to adapt without computational expertise

Segmentation- Mesmer



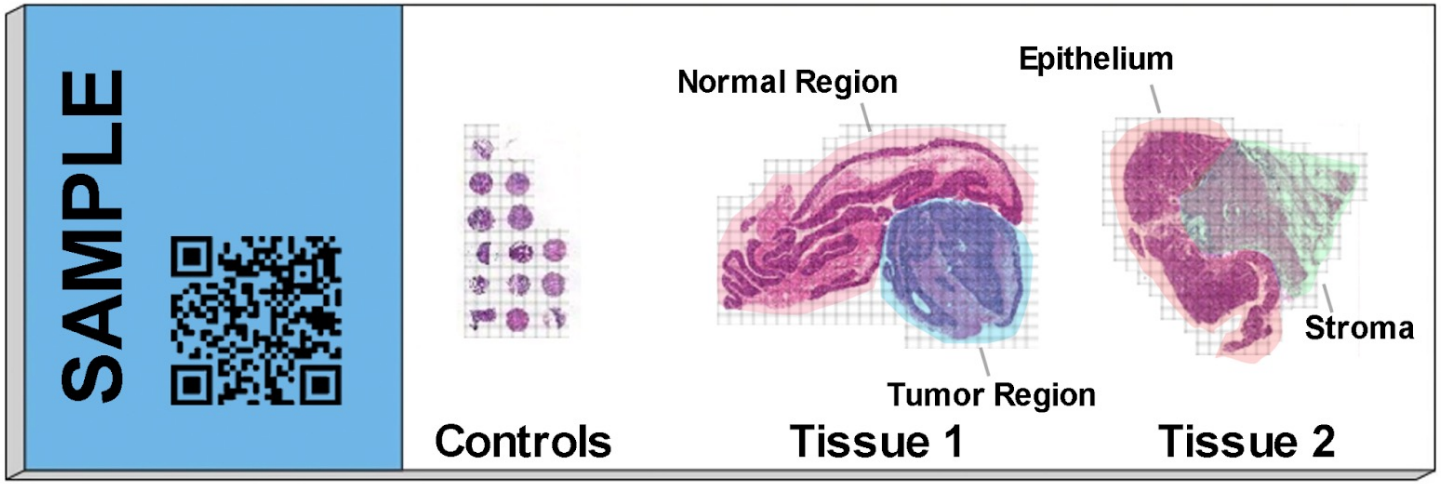
[16] Greenwald et. al. *Nature Biotechnology*, 2022.

Statistical image processing

Image intensity normalization, batch correction, and harmonization

The slide-to-slide problem

- Tissues placed on a slide, each contains *(10s to 100s)* of images
- Multiple slides are imaged in the same experiment
  - Multiple sources of noise introduced each time: optical effects, instrument parameter tuning, different times of staining for antibodies
- Large batch effects!



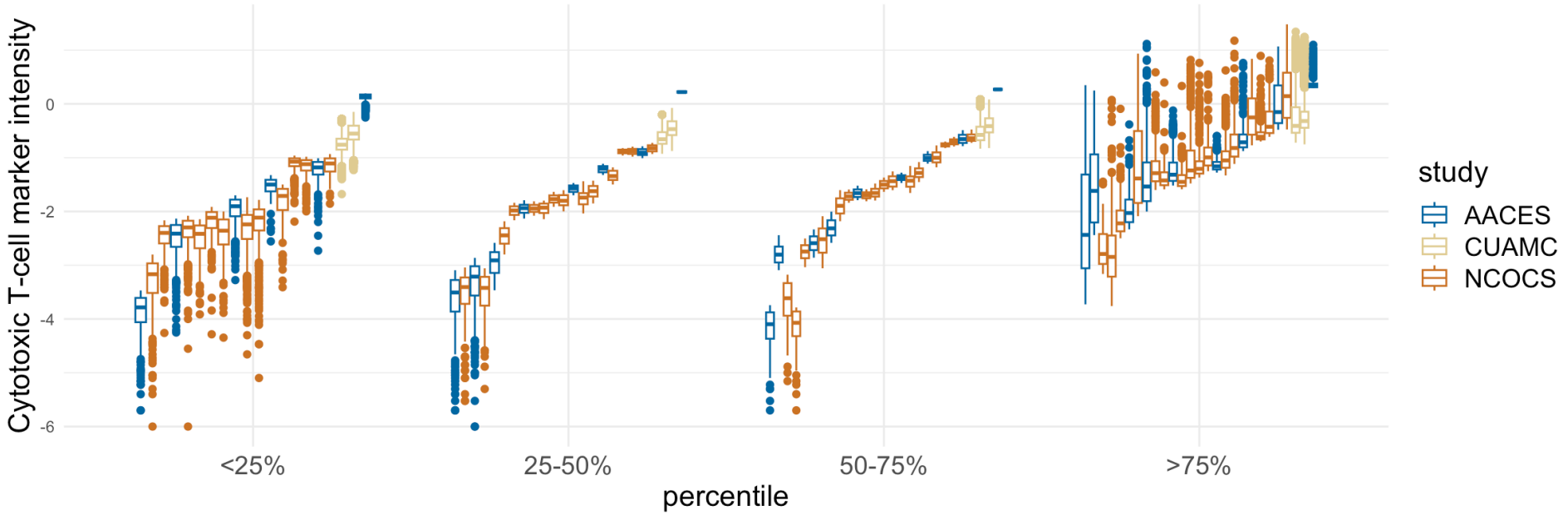
Transformation vs. normalizing vs. harmonization

Image transformation, normalization, and batch correction are used to make the data more appropriate for downstream analysis by removing non-biological biases in marker intensity distributions

- Transformation: log, arcsinh, square root
  - Make data more normally distributed, do not adjust for systematic effects
- Normalization: adjusts distribution of marker intensities in each slide, image, or channel separately to make distributions appear more similar
- Batch correction / harmonization: explicitly removes systematic bias using variables that account for processing steps

Batch effects in multiplex imaging data

- Distributions of marker intensity values can be **very** different across slides
  - Can lead to bias in downstream phenotyping and analyses
- Usually not accounted for in analysis!
  - Unlike neuroimaging, where normalization/harmonization well established



Challenges specific to multiplex imaging

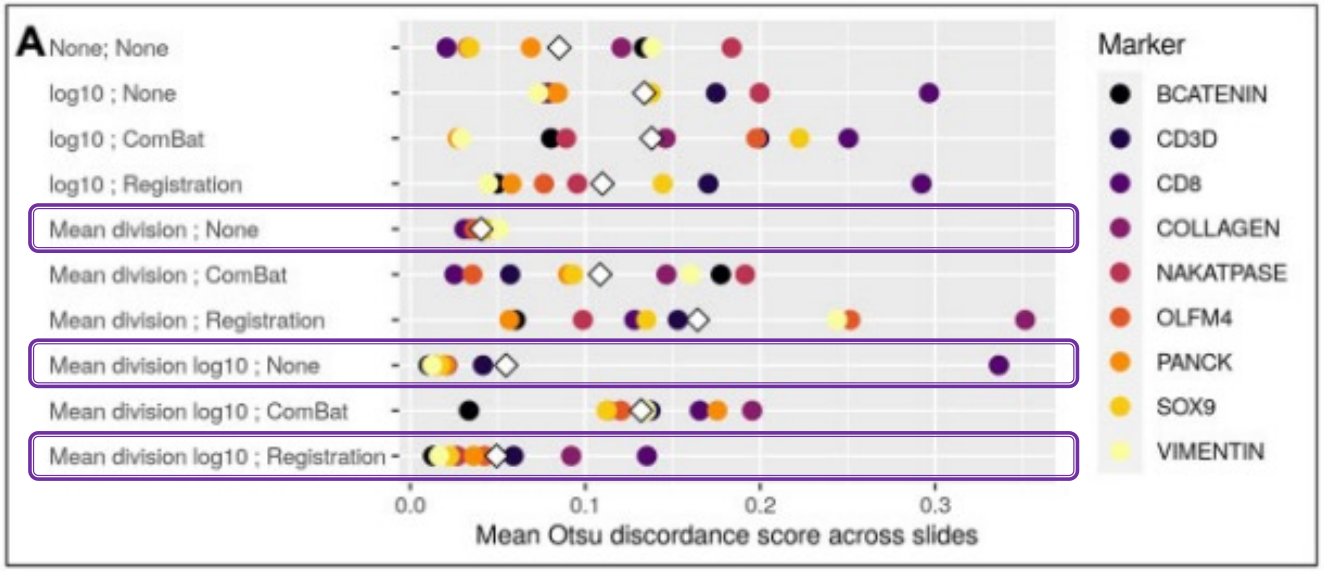
- Highly skewed, zero-inflated, and overlapping intensity distributions
- Slides differ in their composition
- Imperfect cell segmentation increases uncertainty
- Lack of "ground truth" to evaluate good normalization

- Existing batch correction methods (e.g., Combat) don't work well for this data

A batch correction evaluation framework for scMI

- Harris et. al. 2022 established metrics
  - Alignment of marker densities
  - Cell phenotyping discordance
  - Proportion of variance due to slide
- Compared 9 combinations of transformation/normalization
  - Mean division is simple and worked well

Cell phenotyping discordance

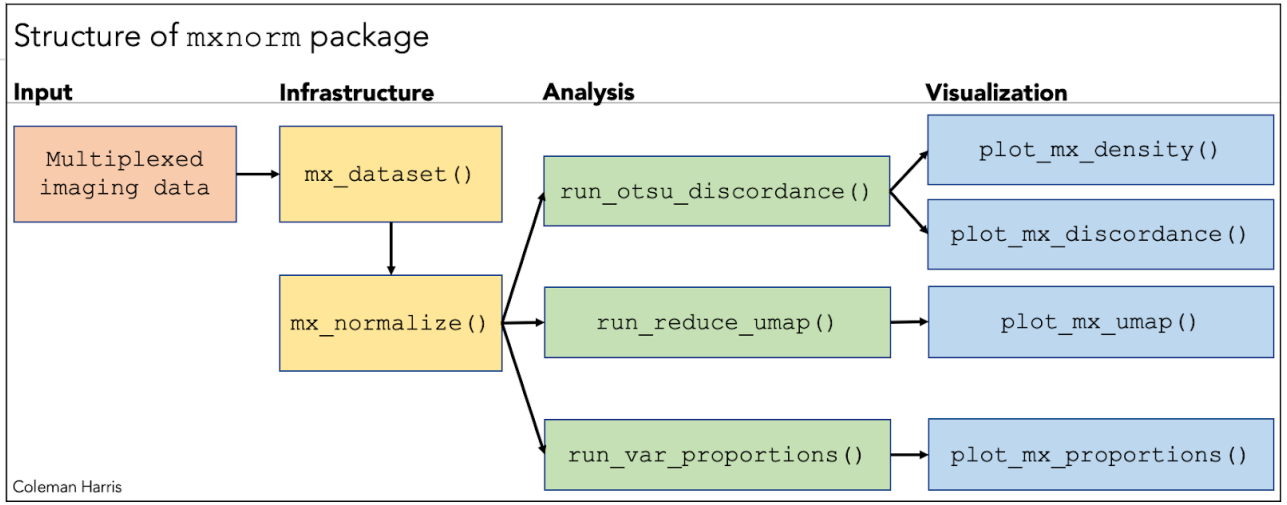


Harris, McKinley, Roland, Shrubsole, Lau, Coffey, Wrobel, Vandekar. *Bioinformatics*, 38(6) (2022)

README.md

mxnorm: An R package to normalize multiplexed imaging data.

CRAN 1.0.2 | downloads 3404 | JOSS 10.21105/joss.04180



- Allows users to easily allow evaluate normalization in their own data
- Default normalization options (from Harris et al. 2022) or user specified
  - mxnorm can be used to evaluate new methods in future papers

Harris, Wrobel, Vandekar. *Journal of Open Source Software*, 7(71) (2022)

Statistical image processing

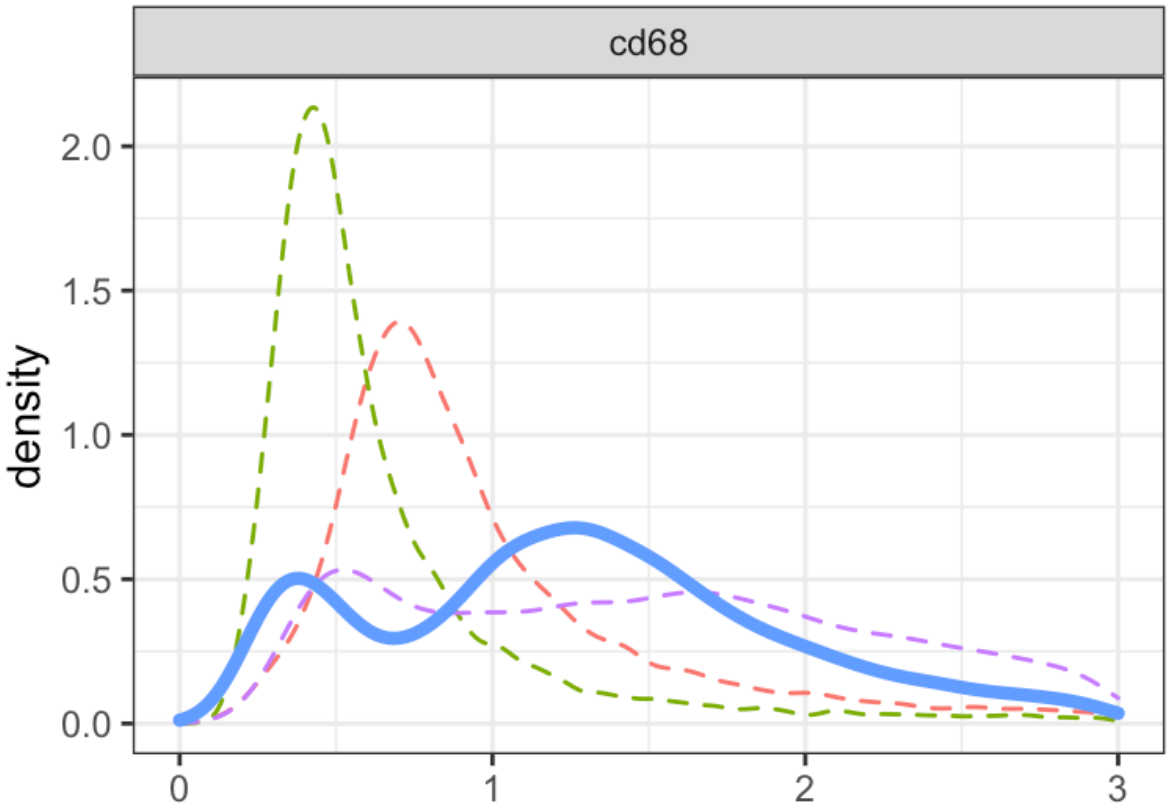
Cell phenotyping

Cell phenotyping

The process of identifying cell types from marker expression values

- Cell labeling

Conceptually, goal is to create a "cut point" in marker intensities where cells are either positive or negative for a marker



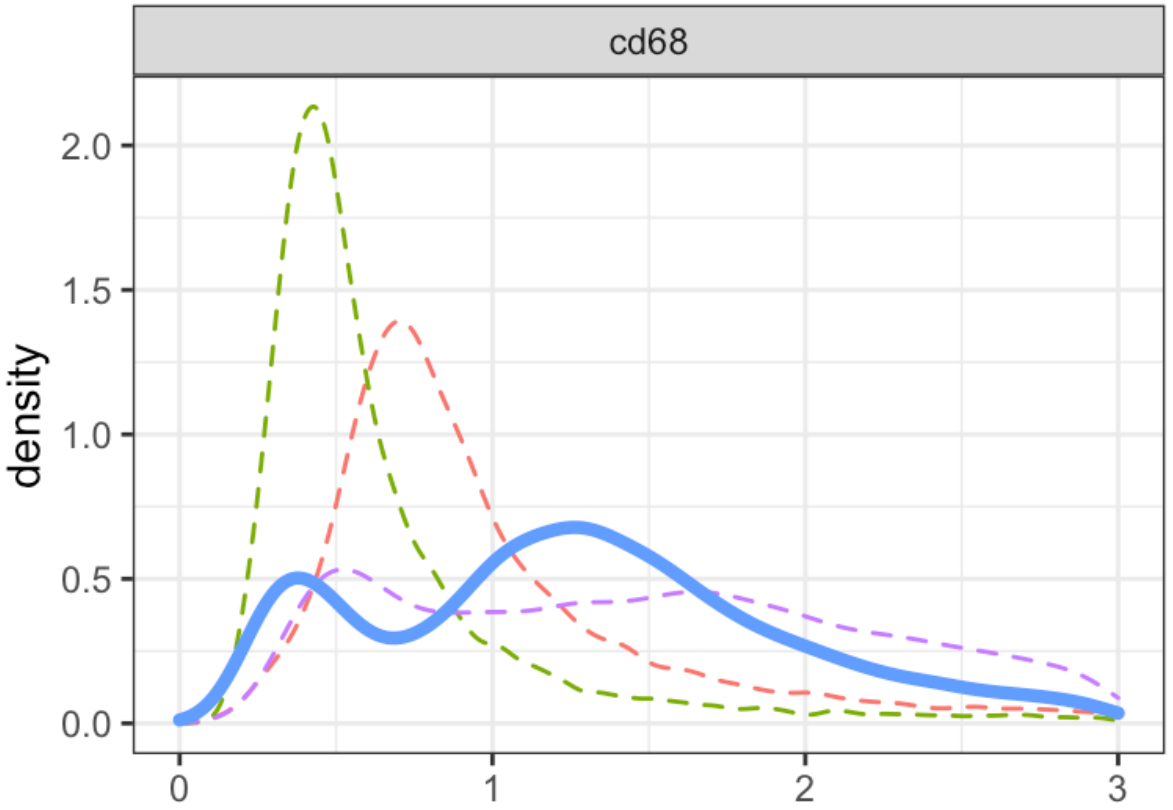
Cell types by immune markers

Type of cell	CD markers
stem cells	CD34+, CD31-, CD117
all leukocyte groups	CD45+
Granulocyte	CD45+, CD11b, CD15+, CD24+, CD114+, CD182+[17]
Monocyte	CD4, CD45+, CD14+, CD114+, CD11a, CD11b, CD91+,[17] CD16+[18]
T lymphocyte	CD45+, CD3+
T helper cell	CD45+, CD3+, CD4+
T regulatory cell	CD4, CD25, FOXP3 (a transcription factor)
Cytotoxic T cell	CD45+, CD3+, CD8+
B lymphocyte	CD45+, CD19+, CD20+, CD24+, CD38, CD22
Thrombocyte	CD45+, CD61+
Natural killer cell	CD16+, CD56+, CD3-, CD31, CD30, CD38

Thanks to Brooke Fridley and Alex Soupir at the Moffitt Cancer Center for this image

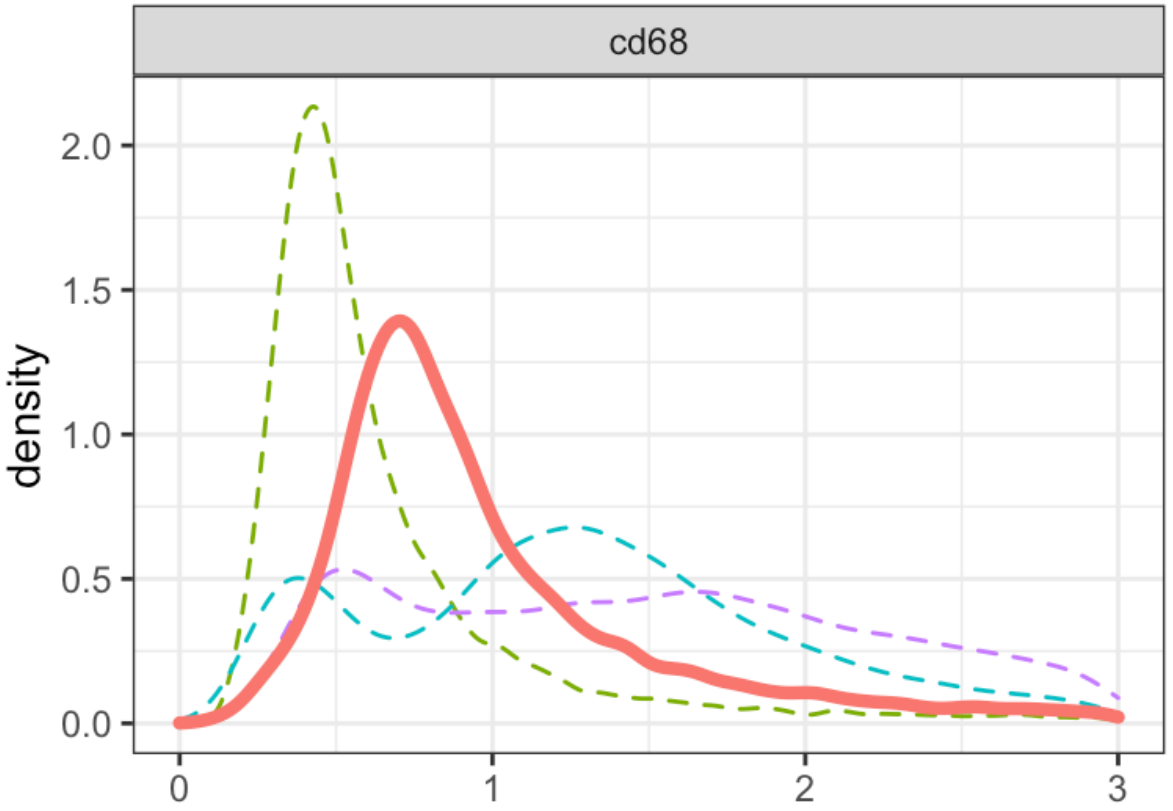
Phenotyping challenges for MI data

- Marker intensity distributions are highly right skewed
  - Often no clear bimodality
- Sensitive to upstream image processing
  - Normalization
  - Cell segmentation



Phenotyping challenges for MI data

- Marker intensity distributions are highly right skewed
  - Often no clear bimodality
- Sensitive to upstream image processing
  - Normalization
  - Cell segmentation



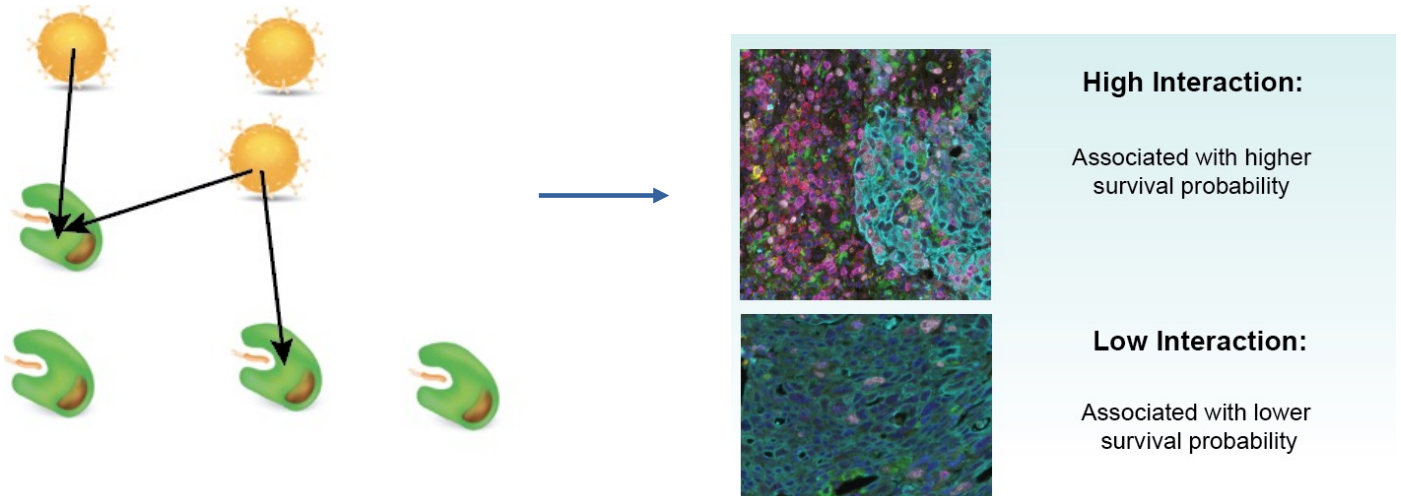
Phenotyping approaches

- Marker gating
  - Determine cutpoint in marker intensity histogram to designate “marker positive” and “marker negative” cells
- Unsupervised clustering methods
  - Seurat, Phenograph contain built-in software
  - Most developed for other single-cell analysis
- Semi-supervised
  - inForm/Halo proprietary software use manual gating to guide phenotyping
  - MAUI/CU-Anschutz: Deep-learning based pixel-level
  - Astir-: Deep-learning based cell-level
  - GammaGateR: new approach by Xiong, Vandekar, 2023+

Statistical analysis

How do multiplex images relate to patient outcomes?

1. Quantify characteristics / features of an image
   - Features: cell type proportions, degree of immune cell clustering
2. Relate to patient-level or clinical outcome
   - Features -> model covariates
   - Outcomes: disease progression, tumor subtype, patient survival time



B Steinhart, KR Jordan, J Bapat, MD Post, LW Brubaker, BG Bitler, and J Wrobel. *Molecular Cancer Research*, 19(12) (2021)

Image feature extraction

- Why can't we compare the images themselves directly (akin to neuroimaging)?
  - Highly heterogeneous structure
  - No spatial registration possible to obtain pixelwise correspondence

- Extract features instead to obtain correspondence across images
  - Cell type proportions
    - Common, but do not typically include spatial information
  - Spatial features
    - Typically measure clustering of one or more cell types

Statistical analysis

Spatial analysis

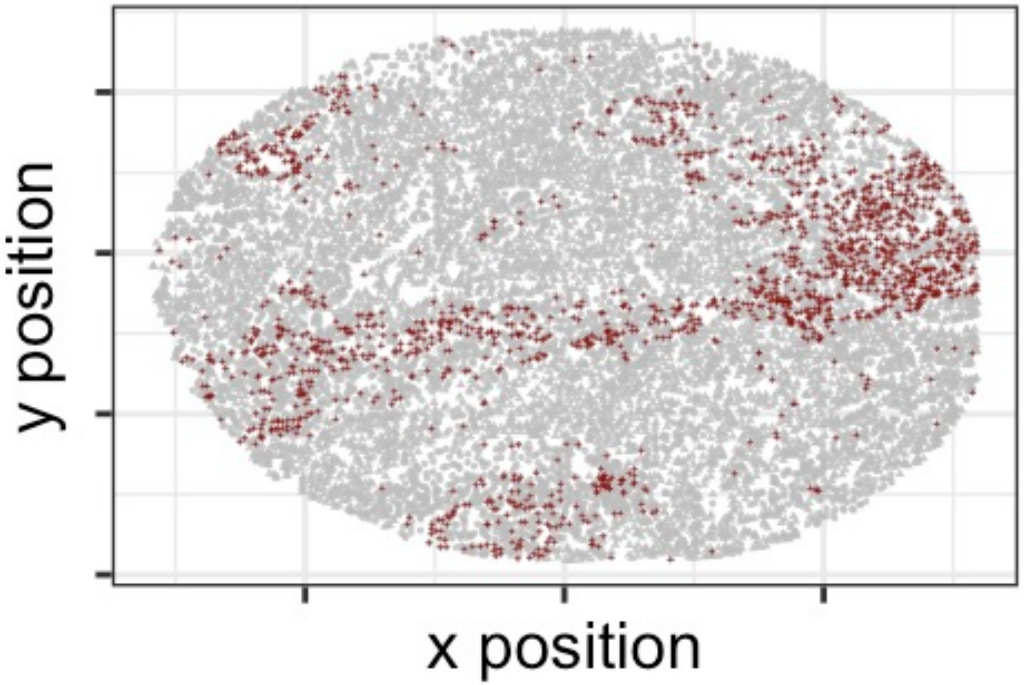
Spatial summary statistics

- Extracted to describe spatial relationship amongst cells in an image
  - Often separated by tumor/stroma
- Univariate spatial summary statistics
  - Clustering or dispersion of cells of one type
- Bivariate spatial summary statistics
  - Co-expression or co-clustering of two cell types (e.g. T-cells and B-cells)

- Methods based on spatial point processes
  - Analyze number of neighbors: K-function, L-function
  - Analyze distance to nearest neighbor: G-function

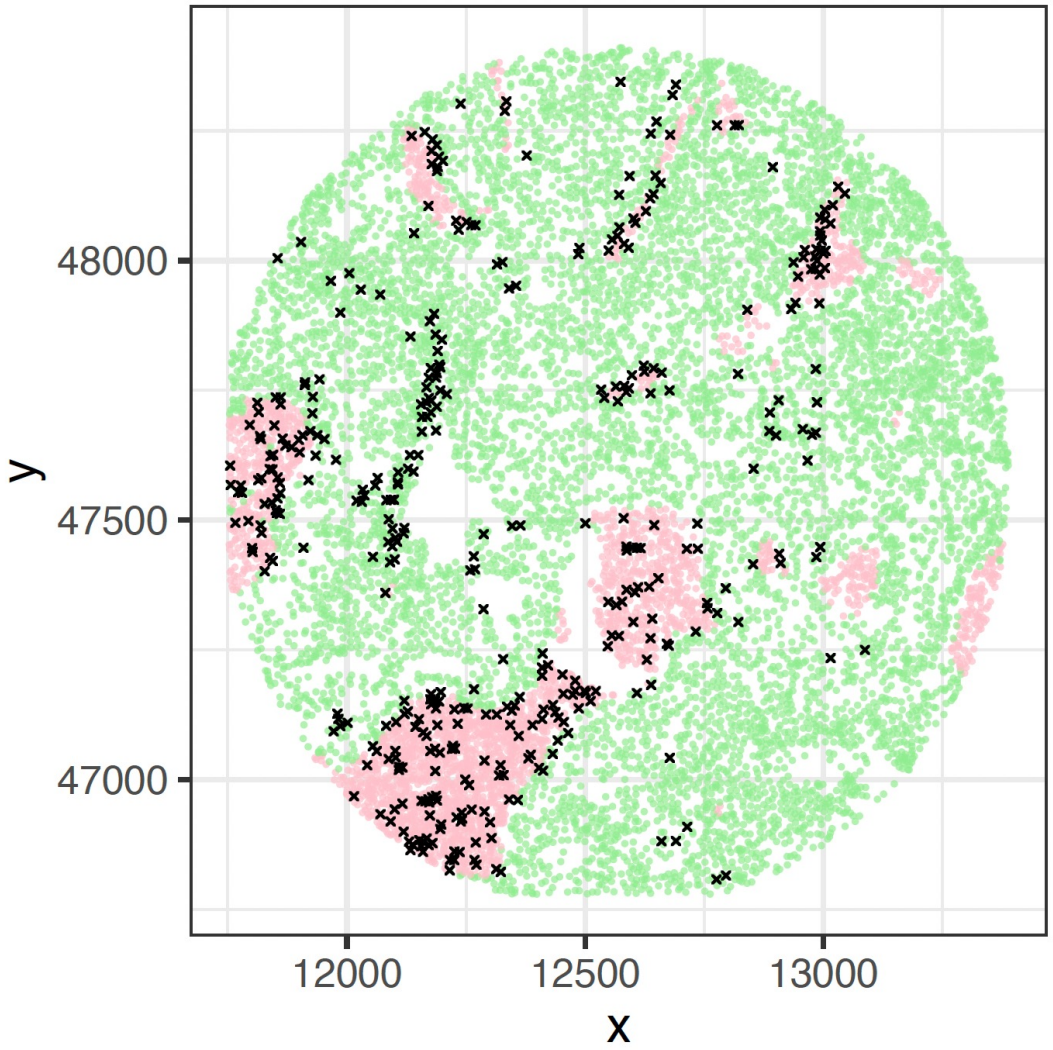
Spatial point patterns

- Locations of cells are considered random and to follow a *point process*
- *Marked point patterns* have covariates (cell shape, area, expression of CD3) associated with each point
- *Multitype point patterns* have multiple types of points
  - Red cells are immune cells
  - Interested in spatial arrangement of immune cells in the tumor



Another example: macrophages in ovarian cancer

- Green cells are in tumor tissue area
- Pink cells are in stromal tissue area
- Black cells are macrophages
  - Quantify macrophage clustering in tumor and stromal areas



Spatial summary functions based on point processes

- Capture important features of the point pattern
- Assume location of points in point pattern are random variables
  - Typically number of points *n* is also a random variable

- Often to detect deviations from complete spatial randomness (CSR)
  - Clustering or repulsion
  - Ripley's K

Complete Spatial Randomness

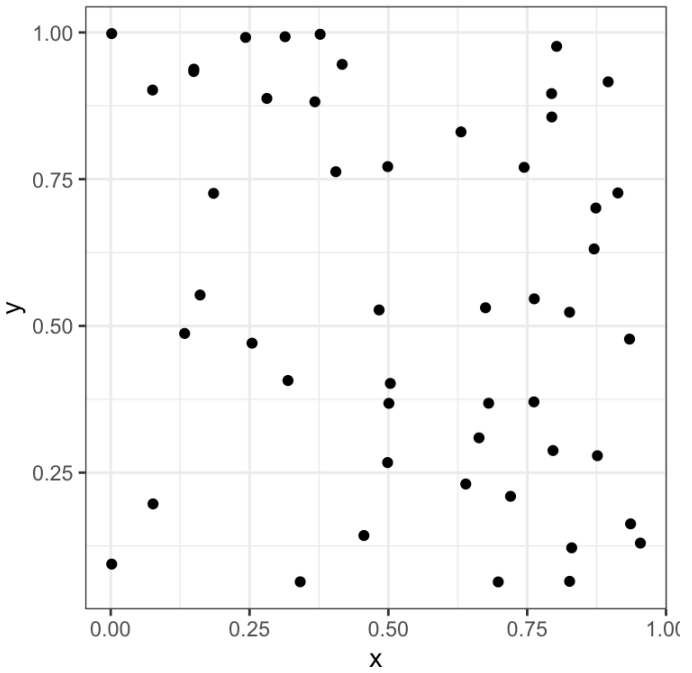
- Homogeneity: $E[n(X \cap A)] = \lambda|A|$
  - The expected number of points falling in a given region *A* should be proportional to the area of the region

- Spatial independence
  - The counts in disjoint subregions of *R* are independent random variables

Then, the number of points falling in *R* follows a **Poisson distribution**

- Homogeneous Poisson point process
  - Deviations from Homogeneous Poisson point process represent spatial clustering

$$P(n(X \cap A) = k) = e^{\lambda|A|} \frac{(\lambda|A|)^k}{k!}$$



λ is the *intensity*, or expected number of random points per unit area

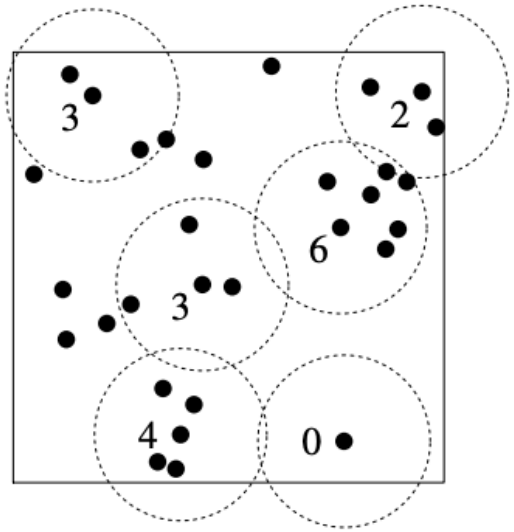
Ripley's K-function

- K-function is a popular metric for analyzing spatial correlation in point patterns
  - Essentially the standardized average number of neighbors of a cell within radius *r*

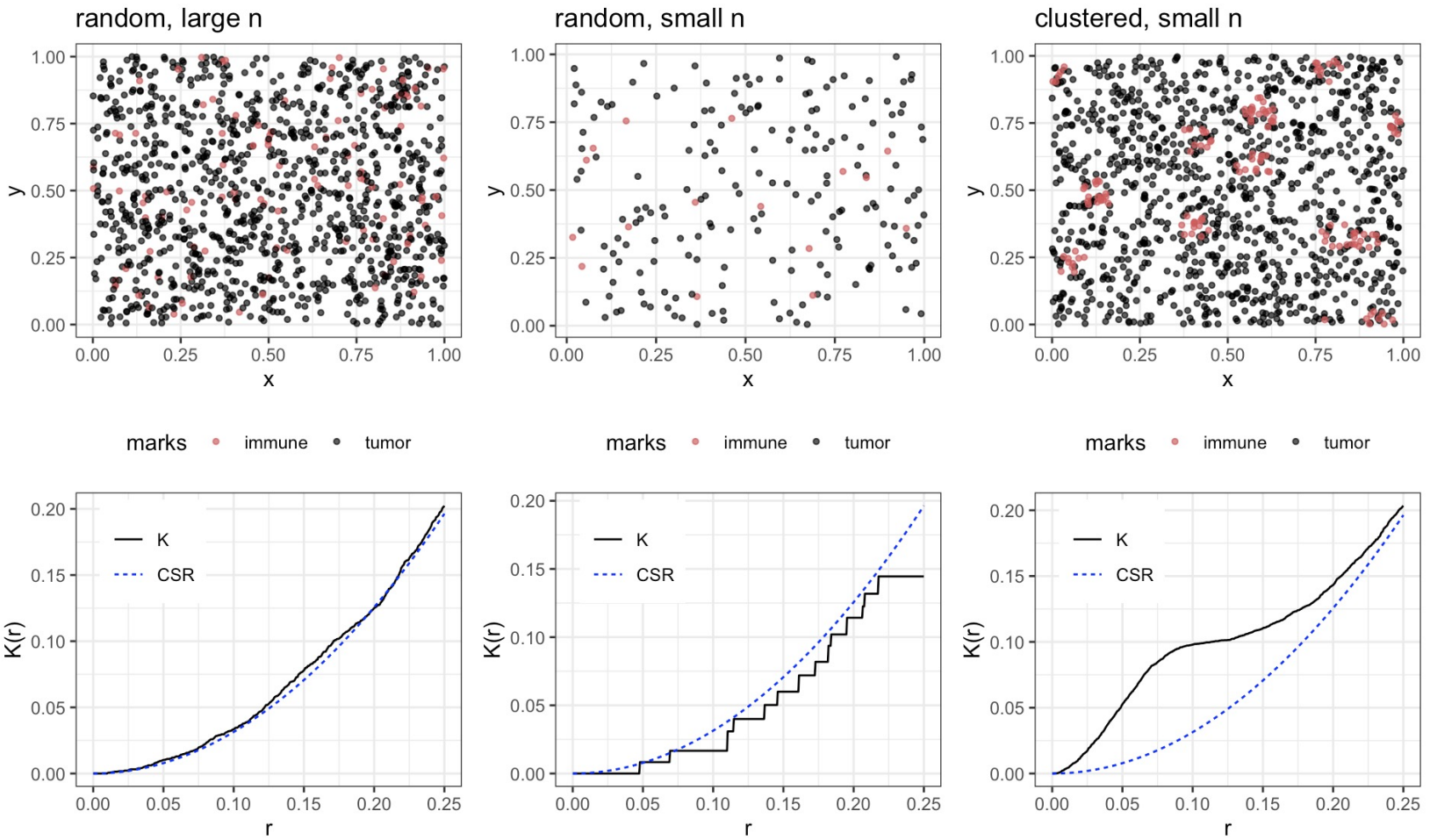
$$\widehat{K}(r) = \frac{|A|}{n(n-1)} \sum_{i}^{n} \sum_{i \neq \mathrm{j}} I\left(d(c_i, c_j) \leq r\right) e_{ij}$$

- Has theoretical value under CSR
  - Compare observed to theoretical value to assess clustering

$$E_{CSR}\left(\widehat{K}(r)\right) = \pi r^2$$



Ripley's K function for scMI data



Other spatial summary statistics

- Mark correlation function
  - Derivative of K function
- Nearest-neighbor G function

- Moran's I
  - Can be used to quantify continuous marker intensity values
  - Local and global versions
  - Univariate and bivariate

Issues that arise in multiplex imaging data

- Rare cell types: spatially summary metrics cannot be computed for images with no or few cells of a certain subtype
  - Makes it challenging to compare across images
- Multiple images per subject
  - Needs to be accounted for in analysis
- Violation of homogeneity assumptions bias estimation of K, G functions
  - Inhomogeneity can give appearance of spatial clustering

Statistical analysis

Spatial analysis of scMI data using functional data analysis

Modeling scalar spatial features- general workflow

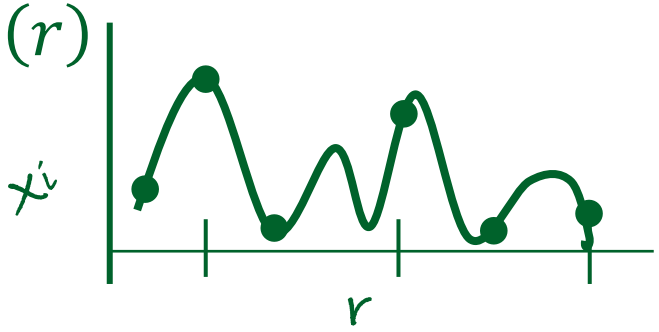
1. Choose a cell spatial relationship you want to analyze
   - e.g. co-occurrence of B-cells and macrophages
2. Select a spatial summary measure (e.g. bivariate Ripley's K).
3. Choose a particular radius ***r*** at which to evaluate Ripley's K.
   - This can be selected based on clinical knowledge.
4. Evaluate spatial summary measure at radius ***r*** for all images in the dataset.
5. Select a patient-level outcome of interest (e.g. overall survival)
6. Model the outcome, using and the spatial summary measure as a covariate
   - Include other potential confounders

Caveat: selection of single radius is arbitrary and results in loss of information

- Methods from functional data analysis can help

Functional data analysis

- Outcome or covariate is curve or function $X_i(r)$
  - Conceptually r is continuous
  - In practice r is discrete

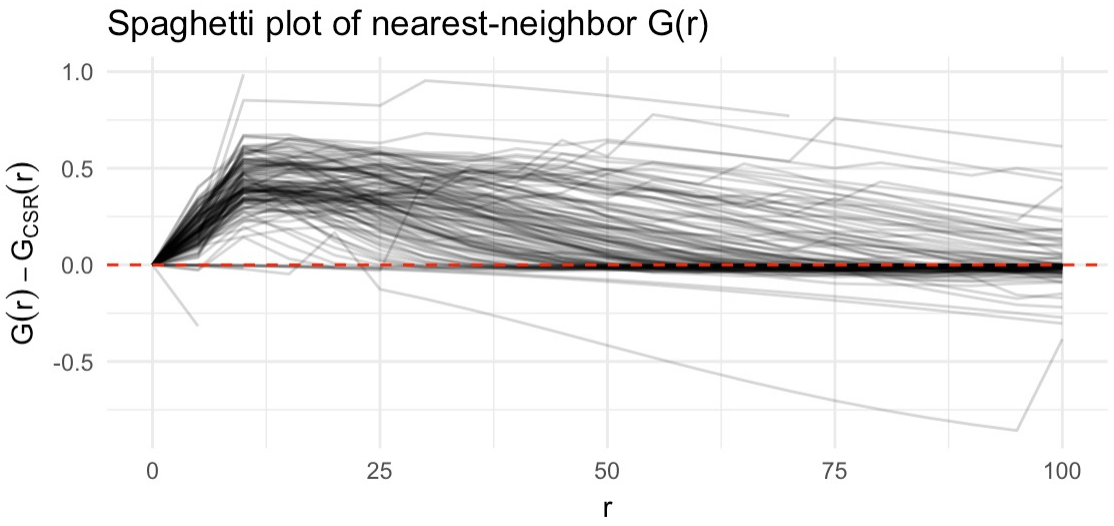


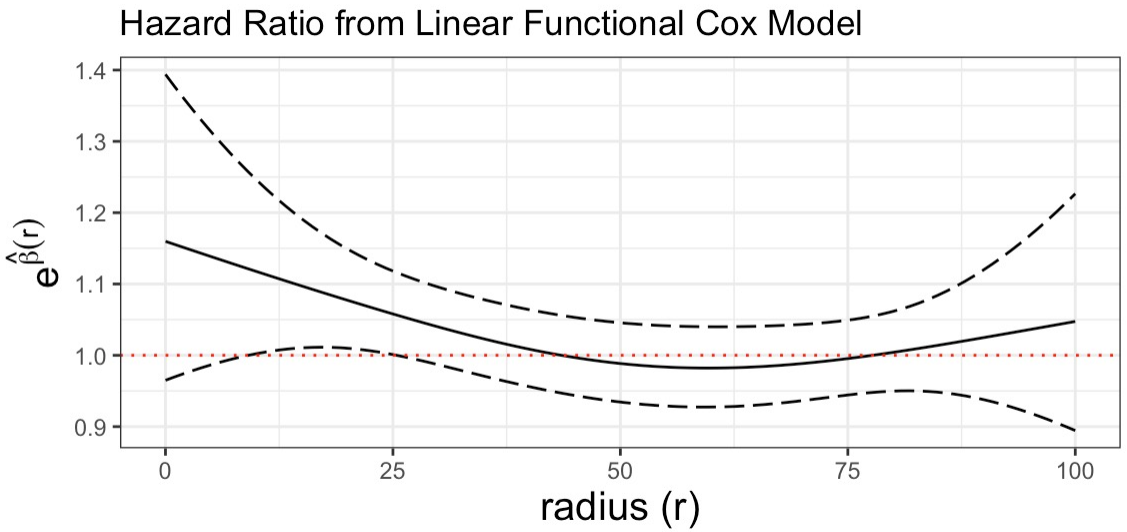
- Great framework for studying spatial summary functions
  - For our application $X_i(r)$ is a spatial summary function (e.g. K)

- There are functional analogs of common tools like regression, PCA

Linear functional Cox model (lfcm)

$$\log \lambda_i(t; Z_i, X_i) = \log \lambda_0(t) + \sum_{k=1}^{p} \gamma_k Z_{ik} + \int \beta(r) X_i(r) dr$$





Gellar et al 2015. "Cox regression models with functional covariates for survival data." *Statistical Modelling.*
Vu, Wrobel, Ghosh. *PLoS Computational Biology,* 18(6). *2022*

Resources

`VectraPolarisData` Package

- Surprisingly little multiplex-imaging data is publicly available
- VectraPolarisData package addresses this
  - Bioconductor `ExperimentHub` package as of April 2022
- 2 large datasets from my collaborators at CU-Anschutz

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Installation

To install this package, start R (version "4.2") and enter:

```
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("VectraPolarisData")
```

`VectraPolarisData` datasets

HumanLungCancerV3

- Non-small cell lung carcinoma ROIs
- Vectra3 mIHC, CU-AMC
- 761 images from 153 patients
- 1,604,786
- 7 markers
  - 5 phenotypic (CD3, CD8, CD14, CD19, CD68, ck)
  - 1 functional (HLADR)
- Patient-level outcomes

HumanOvarianCancerVP

- High-grade serous ovarian cancer tumor microarray
- VectraPolaris mIHC, CU-AMC
- 128 patients, 1 image each
- 1,610,431 cells
- 8 markers
  - 5 phenotypic (CD3, CD8, CD68, CD19, CK)
  - 2 functional (pstat3, IER3)
- Patient-level outcomes

Other resources

- Short course on multiplex imaging
  - http://juliawrobel.com/MI_tutorial
  - Code examples using VectraPolarisData, fully open-source
- Textbook chapter on multiplex imaging (Wrobel, Harris, Vandekar)
  - http://juliawrobel.com/Downloads/mIF_chapter.pdf
- Monthly national working group for scMI and spatial transcriptomics
  - Run by Simon Vandekar
  - Email me or Simon if you'd like to be added

Acknowledgements, and thanks!

Colorado SPH Biostatistics

- Andrew Leroux
- Thao Vu
- Souvik Seal
- Tushar Ghosh
- Debashis Ghosh

Vanderbilt Biostatistics

- Simon Vandekar
- Ruby Xiong
- Coleman Harris

Moffit Cancer Center

- Brooke Fridley
- Lauren Peres
- Alex Soupir

Contact Info

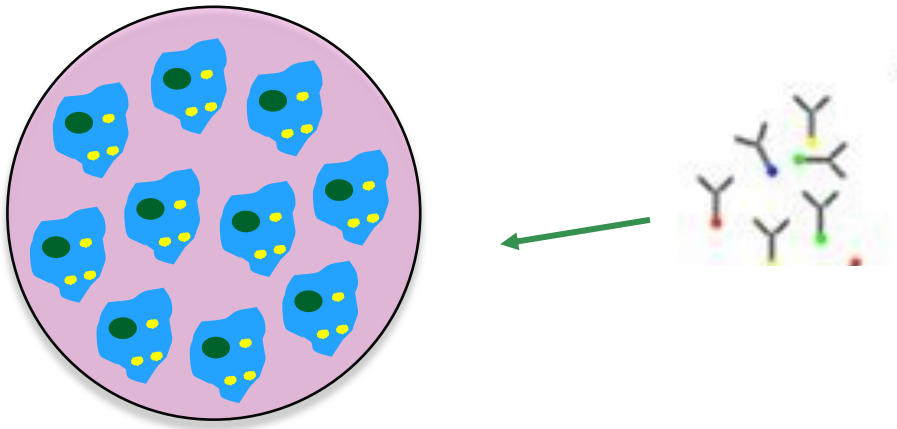
julia.wrobel@emory.edu

juliawrobel.com

github.com/julia-wrobel/mxfda

What is single cell multiplex imaging?

- ***Multiplex*** refers to multiple types of protein in the tissue that are tagged



- *Immunofluorescence* based
  - Proteins stained with fluorescent antibodies then imaged using fluorescence microscopy
- *Mass cytometry based*
  - Proteins tagged with metal isotypes (IMC, MIBI)