Unsupervised Learning Methods for Investigating Feature Associations in Lung Cancer Data

Clara Craiu[1], Cassandra Hung[2], Owen Yoo[3]

[1]McGill University, [2]Wake Forest University, [3]University of Michigan

SCHOOL OF PUBLIC HEALTH
BIOSTATISTICS
UNIVERSITY OF MICHIGAN

Background

Multiplex Imaging

- Recent work in the field of cancer data science has explored the use of spatial biomarkers within a tumor microenvironment (TME) to provide insights into patient characteristics and outcomes.
- Multiplex imaging (MI) allows for the analysis of TMEs at single-cell resolution while preserving spatial relationships between cells.
- Using MI, individual cells are tagged with fluorescent antibodies that attach to specific proteins on a cell's surface. These markers allow cells to be classified by phenotype and function.

The Data

- The VectraPolarisData package[1] contains a lung cancer MI dataset collected at the University of Colorado Anschutz Medical Campus, which includes 761 images from 153 patients.
- The dataset covers information such as x and y coordinates of each cell in an image, cell phenotypes, and patients' characteristics ("patient metadata"), including age at diagnosis, pack-years, time to recurrence, and gender.

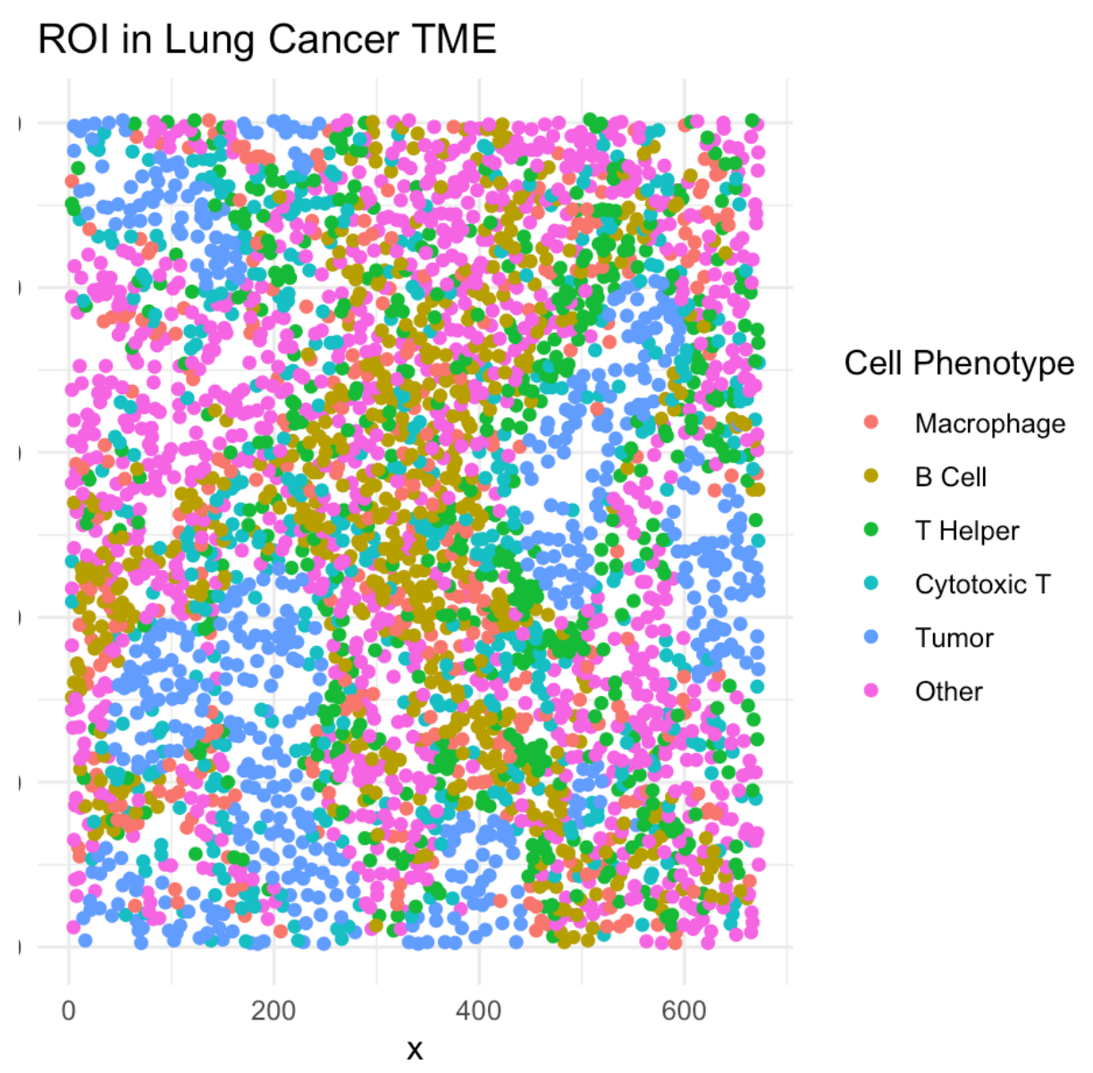


Figure 1: An example plot of a region-of-interest (ROI).

Motivation and Question

The question under consideration is: What insights can unsupervised clustering of lung cancer patients using spatial biomarkers provide about associations in patient metadata?

Lung cancer is the leading cause of cancer deaths both in the US and globally. Researching the potential associations between spatial patterns and other features in patient data can help improve patient outcomes while also finding potential prognostic and predictive biomarkers.

Methods

Ripley's K Function

- Ripley's K function is an essential tool that measures the standardized average number of neighbors of cell within a specific radius.

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

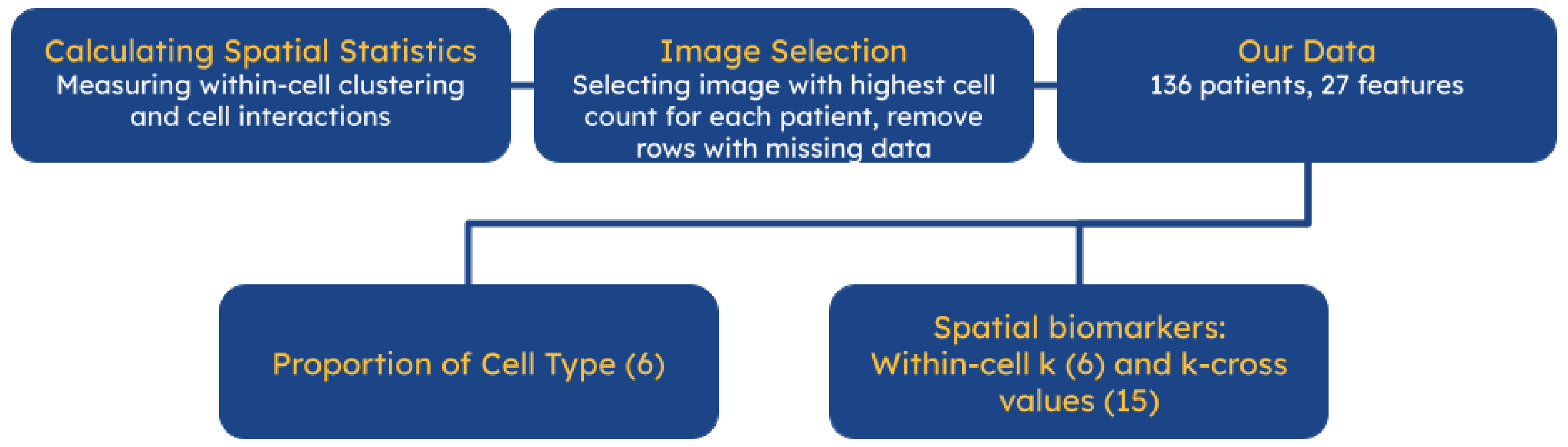
- A curve for theoretical and observed Ripley's K functions can be used to visualize the degree of cell clustering and repulsion. The difference between these observed and theoretical curves at specific radii (in this project, 40 μm) can be used to assess cell clustering.
- One of the main assumptions of the function is spatial homogeneity, meaning that the point intensity is constant across the study area. To avoid violating this assumption, we can calculate the permuted k values.

Hierarchical Clustering

- Clustering method where data is categorized based on similarities defined by an algorithm and distance metric
- Ward's algorithm is an agglomerative method which aims to minimize the total error sum of square increase when clustering. In this case, all patients start in their own cluster and two clusters are merged based on the minimizing function. [2]
- Ward's Minimizing Function: $D(c_1, c_2) = \frac{|c_1||c_2|}{|c_1|+|c_2|} d(\mu_{c_1}, \mu_{c_2})^2$

c_n is the cluster, $|c_n|$ is the size of the cluster, μ_{c_n} is the center of the cluster

Selecting Data for Clustering Algorithm



Calculating Spatial Statistics
Measuring within-cell clustering and cell interactions

Image Selection
Selecting image with highest cell count for each patient, remove rows with missing data

Our Data
136 patients, 27 features

Proportion of Cell Type (6)

Spatial biomarkers:
Within-cell k (6) and k-cross values (15)

Dendrogram

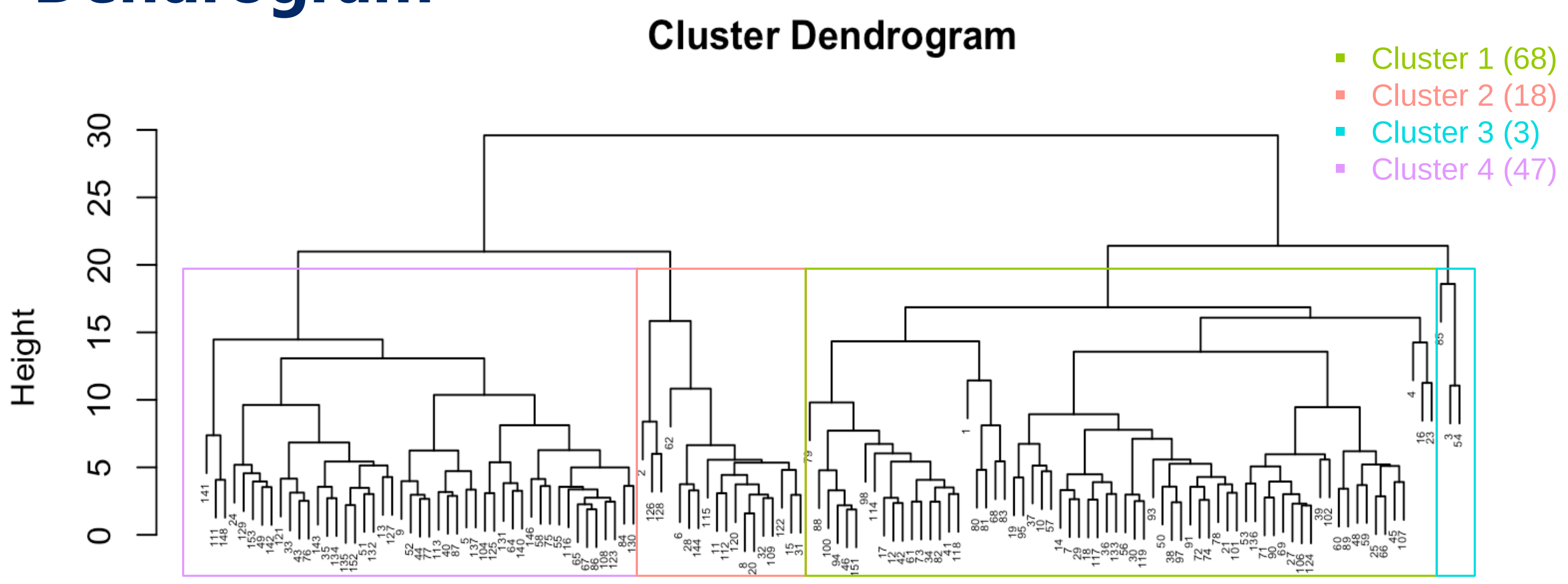


Figure 2: Dendrogram depicting the results of using hierarchical clustering to group patients by proportions of cell types and spatial biomarkers. Based on the dendrogram, four clusters were produced.

Heatmap

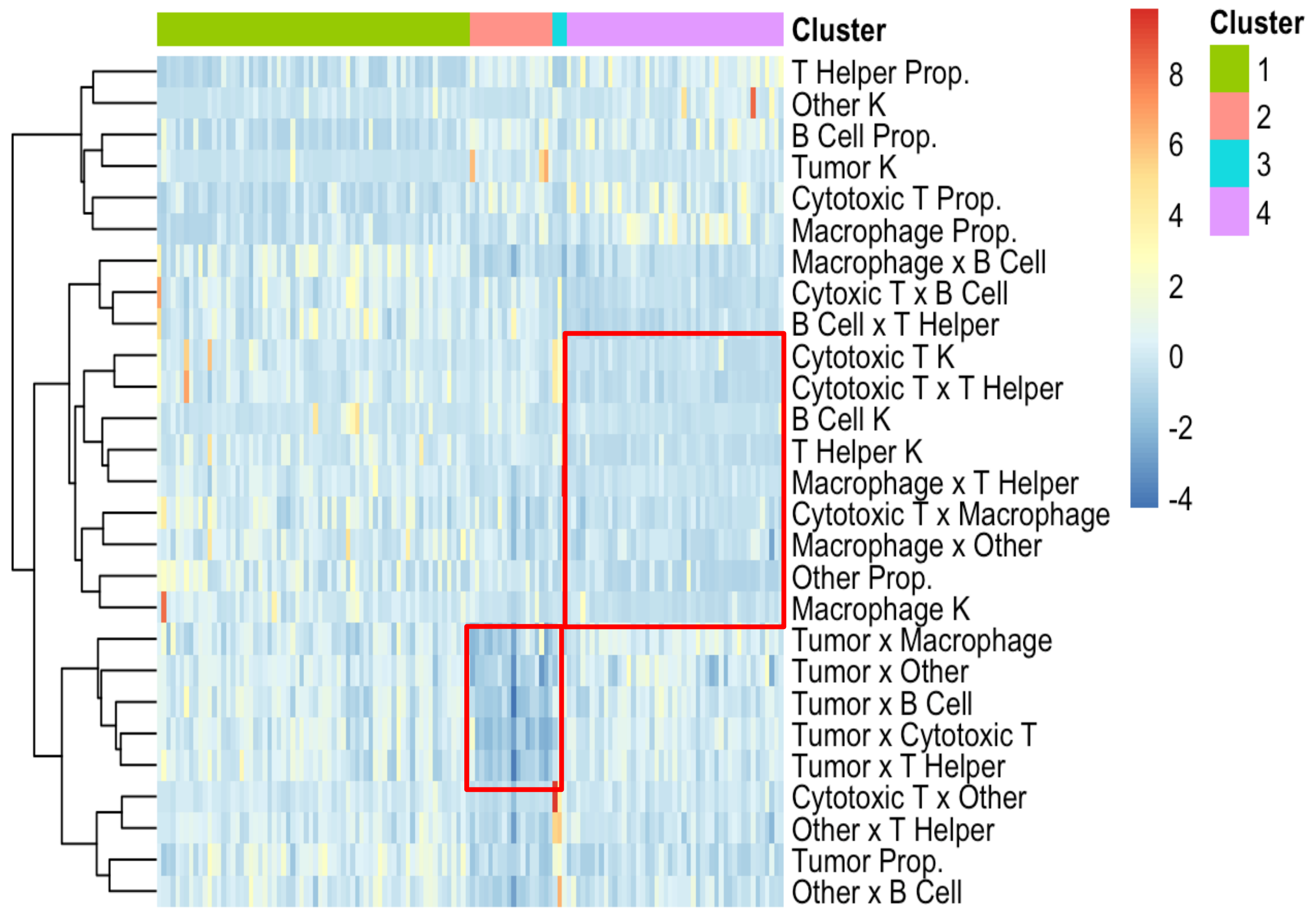


Figure 3: Heatmap showing relative scaled values of spatial features and cell type proportions for each cluster. Notable features include generally lower k-cross values for immune cell interactions in cluster 4, and cluster 2 has lower k-cross values for tumor cell interactions.

Results

- All clusters have no notable differences in demographic features.
- Cluster 1 has a higher proportion of tumor cells.
- Cluster 2 has less interaction between tumor and immune cells.
- Cluster 4 has less interaction between cytotoxic T cells and both T helper cells and B cells.
- Cluster 4 has a lower probability of patients experiencing recurrence.

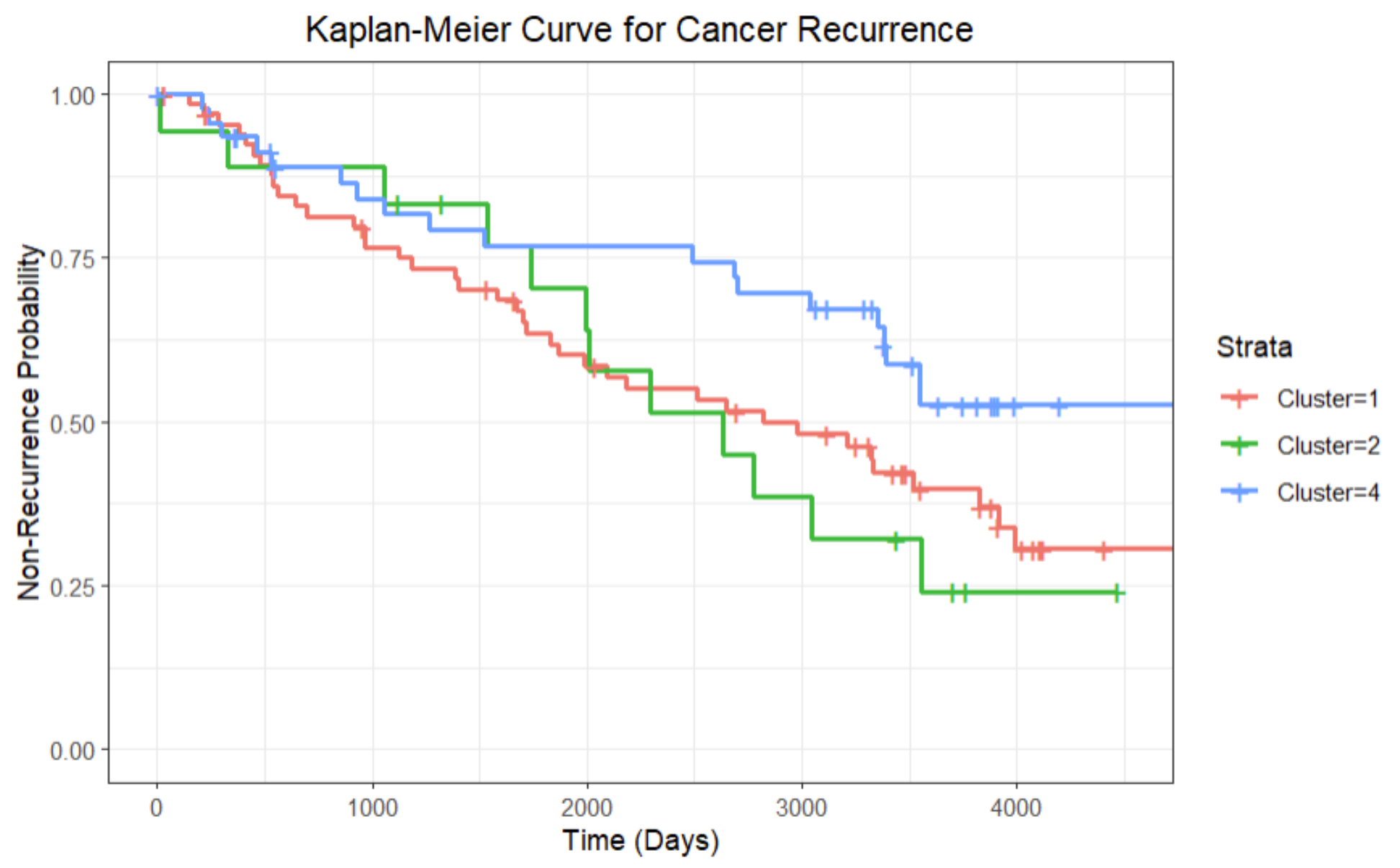


Figure 4: Kaplan-Meier depicting non-recurrence probability over time for all three curves. Based on these results, clusters 1 and 2 have a higher probability of recurrence than cluster 4 (p = 0.04).

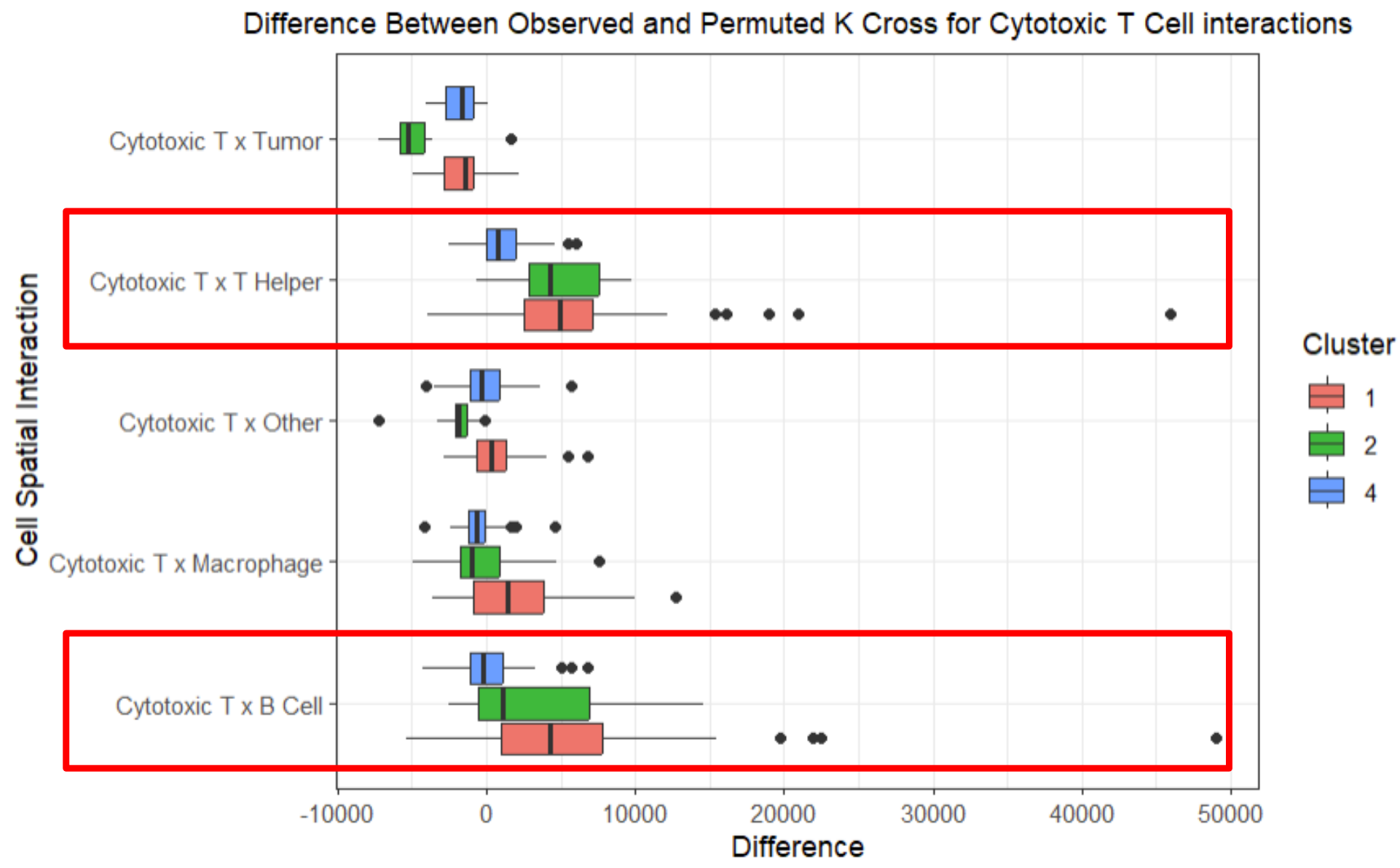


Figure 5: Boxplot depicting k-cross values of cytotoxic T cell interactions with other cells. Cluster 4 has lower k-cross values than the other two clusters with certain immune cells.

Limitations

- Small sample size and limited diversity in patient demographic features reduced scope of association study.
- Clustering by patient only allows for use of one image per patient.

Future Research

- Incorporate methods that allow multiple images for each patient to be used in the clustering algorithm.
- Use other clustering methods and incorporate more and different features in the algorithm.
- Further investigate the association between reduced cytotoxic T and T helper cells clustering with cancer recurrence.

Acknowledgement: We would like to thank Grant Carr, Dr. Junsouk Choi, and Dr. Veera Baladandayuthapani for their generous help and support throughout this project, as well as Dr. Bhramar Mukherjee and the entire staff for the opportunity to be part of the Big Data Summer Institute.

References

1. Amber M. Johnson, Jet al., 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, Volume 16, Issue 10, 2021, Pages 1694-1704, ISSN 1556-0864, https://doi.org/10.1016/j.jtho.2021.05.004.
2. Joe H. Ward Jr. (1963) Hierarchical Grouping to Optimize an Objective Function, Journal of the American Statistical Association, 58:301, 236-244, DOI: 10.1080/01621459.1963.10500845