Investigating Developmental Origins of Sarcomas through Computational Modeling and Machine Learning

Griffin Thompson, Shlien Lab, The Hospital for Sick Children in Toronto

July 2023

Abstract

This report outlines my contributions to a study investigating the developmental origins of sarcomas through computational analysis, machine learning, and single-cell RNA sequencing (scRNA-seq). Conducted at the Shlien Lab at SickKids, this multi-year grant-supported project explored hypotheses about sarcoma origins, with findings that challenge conventional cancer models and contribute to translational oncology

Note: This document summarizes my personal contributions to the research project. The code and data referenced herein are proprietary and cannot be shared, as the research publication has not yet been officially published.

Introduction

During my tenure at the Shlien Lab at SickKids in Toronto, I contributed to a project that aimed to investigate the developmental origins of sarcomas. This work was an integral part of a broader study intended for publication and supported by a multi-year grant. While my contributions laid the groundwork for this endeavor, I have not received updates on the progress since leaving my post. My efforts were primarily focused on computational analysis, machine learning implementation, and data-driven insights to support hypotheses about sarcoma origins.

Sarcomas are a heterogeneous group of cancers arising from mesenchymal tissues. Recent findings have suggested that sarcomas may recapitulate transcriptional programs from embryonic or fetal development, leading to the hypothesis that high-stemness (STEMhigh) sarcomas arise from embryonic-like mesenchymal progenitors, while low-stemness (STEMlow) sarcomas derive from adult mesenchymal stem cells. To address these questions, our work combined single-cell RNA sequencing (scRNA-seq), bulk RNA sequencing (RNA-seq), and advanced machine learning models to infer sarcoma origins and their transcriptional characteristics.

My Contributions

1. Data Preparation and Preprocessing

A significant portion of my role involved preparing large-scale sequencing datasets for analysis. I worked with both single-cell and bulk RNA sequencing data, ensuring it was

clean, normalized, and ready for downstream analysis. This included removing noise, filtering low-quality reads, and normalizing expression values across diverse datasets. Metadata was annotated to include tissue types, developmental stages, and sarcoma subtypes. These tasks were critical in maintaining data quality and ensuring accurate results.

2. Single-Cell Analysis Using R

I utilized R extensively, leveraging advanced packages such as `Seurat` and `SingleCellExperiment` to analyze single-cell sequencing data. Specific tasks included:

- **Clustering cells:** Identifying distinct groups based on transcriptional profiles to uncover patterns of differentiation and lineage.
- **Marker gene identification:** Detecting key genes that distinguished STEMhigh and STEMlow clusters, providing insights into their developmental trajectories.
- **Dimensionality reduction:** Applying PCA and UMAP to visualize high-dimensional data in lower dimensions, facilitating cluster identification and interpretation.

3. Implementation of the OTTER Pipeline

I employed the OTTER (Optimal Tumor Tissue Ensemble Recognition) pipeline, a machine learning model designed to classify tumor samples by assigning probabilities to specific tissue classes. My responsibilities included:

- Feeding preprocessed sarcoma RNA-seq data into the OTTER ensemble of convolutional neural networks.
- Evaluating output probabilities to assign sarcoma samples to their most likely developmental origins.
- Comparing OTTER's results with alternative alignment methods, such as Celligner, to ensure consistency and robustness.

4. Machine Learning and Projection of Sarcoma Data

To better understand sarcoma origins, I applied additional machine learning techniques to project bulk tumor samples onto single-cell transcriptional clusters. Key tasks included:

- Training k-nearest neighbors (KNN) classifiers to map tumors to the most similar single-cell populations.
- Embedding both bulk and single-cell data into shared UMAP spaces to enhance tumor-cell alignments.
- Running clustering algorithms to group sarcoma samples and identify shared transcriptional signatures.

5. Labeling and Interpretation of Sarcoma Clusters

Using the results from OTTER and other models, I labeled sarcoma clusters with biologically relevant annotations, such as dedifferentiated liposarcoma, undifferentiated pleomorphic sarcoma, and others. These labels were informed by external datasets and clinical annotations, providing a more comprehensive understanding of sarcoma subtypes and their potential origins.

6. Visualizing Developmental Trajectories

I created high-impact visualizations to communicate findings effectively. These included:

- UMAP plots illustrating sarcoma developmental trajectories from embryonic mesenchyme to adult mesenchymal progenitors.
- Heatmaps displaying gene expression profiles across STEMhigh and STEMlow clusters.
- Temporal plots mapping sarcoma samples to specific developmental stages, highlighting transitions along the differentiation axis.

7. Linking Sarcoma Origins to Stem Cell Biology

I analyzed transcriptional profiles of sarcoma samples in the context of in vitro stem cell differentiation data. This involved mapping sarcoma clusters to developmental trajectories derived from iPSC-to-mesenchymal progenitor differentiation pathways. The results demonstrated shifts in transcriptional profiles corresponding to differentiation timing, reinforcing the hypothesis of distinct developmental origins for STEMhigh and STEMlow sarcomas.

8. Integration of Public and In-House Datasets

To maximize the scope of the analysis, I integrated public datasets (e.g., GTEx) with in-house sarcoma data. This allowed for a broader comparison of sarcoma samples with normal and cancerous mesenchymal tissues, providing additional support for our findings.

9. Innovation with the Comitani Method

I contributed to the application of the "Comitani Method," an innovative approach for projecting bulk tumor data onto single-cell clusters. This involved preprocessing single-cell data, combining it with bulk tumor data, and implementing dimensionality reduction and clustering techniques to align the datasets effectively. This method provided novel insights into sarcoma classification and developmental lineage relationships.

10. Collaborative and Translational Impact

Throughout the project, I collaborated with biologists and clinicians to ensure the computational findings were biologically and clinically meaningful. My work contributed to discussions about experimental validation, including the potential use of CRISPR-Cas9 to test key genes identified in sarcoma clusters. While validation experiments were beyond the scope of my role, these suggestions highlighted the translational potential of our findings.

Discussion and Impact

My contributions to this project laid the foundation for understanding the developmental origins of sarcomas through computational modeling and machine learning. By integrating single-cell and bulk RNA-seq data with advanced analytical tools, I provided critical insights into the distinct transcriptional profiles of STEMhigh and STEMlow sarcomas. These findings have implications for improving diagnostic accuracy, refining disease models, and developing targeted therapies.

While my role primarily involved computational analysis, the broader study represents a significant step forward in the field of developmental oncology. The integration of machine learning and stem cell biology underscores the potential for interdisciplinary approaches to unravel complex biological questions. This experience not only deepened my understanding of computational genomics but also reinforced my commitment to leveraging engineering principles to solve biomedical challenges.

Conclusion

My work at the Shlien Lab exemplifies the power of computational tools and machine learning in advancing our understanding of cancer biology. Through innovative analysis and collaboration, I contributed to a project with significant potential for clinical impact. This experience has prepared me to pursue a master's degree in Biomedical Engineering, where I aim to further explore the intersection of computational biology, tissue engineering, and regenerative medicine.