Mapping Genotype to Phenotype through Joint Probabilistic Modeling of Single-cell Gene Expression and Chromosomal Copy Number Variation

Anabel Ojeda (Mentors: Dr. Elham Azizi, Joy Linyue Fan, 2023 IICD SRP)

Studying the genetic behavior of the complicated disease we call cancer is essential to gain more insight into how to tackle its treatment. A significant challenge to discovering the next breakthrough in cancer treatment is cell resistance. Large-scale copy number variation (CNV) is common in tumors and leads to differences in how cells within the tumor behave and respond to treatment. The traditional method for studying CNVs is by analyzing at the bulk level the entire genetic makeup of the tumor; however, the new method proposed within this project led by Elham Azizi, called ECHIDNA, uses single-cell analysis to understand these genetic changes more effectively. ECHIDNA utilizes two types of data, single-cell RNA sequencing (scRNA-seq) and population-level whole genome sequencing (WGS). Single-cell RNA sequencing provides information about gene expression within individual cells, and whole genome sequencing offers an overall picture of the tumor's genetic makeup. Integrating both data types enables the ability to identify distinct groups of cells within the tumor, called clones, and understand how their genetic variations influence their behavior. To accomplish this, ECHIDNA uses a probabilistic model that infers the clones from the bulk WGS data and incorporates the gene expression patterns drawn from the scRNA-seq data. The temporal dimension is also considered by analyzing samples taken throughout different stages of treatment. It is now possible to connect the stable genetic characteristics of the clones with the changes in cell behavior over time. ECHIDNA was tested on biopsies from melanoma patients undergoing anti-PD1 immunotherapy and proved to identify and analyze the different tumor clones accurately. This method provides valuable insights into how therapy affects various clones within the tumor, contributing to an overall enhanced understanding of the dynamics of tumor evolution and response to treatment. With my participation in this project, we will further study how copy number cell variation drives tumor heterogeneity to focus on a model that can group and cluster cells in a manner that does not depend on the sample cell and copy number data. For this project, I will investigate how to cluster the cells for those distinct groups to match the whole genome data. The clustering techniques I will incorporate are rooted in K-means and Gaussian mixture models. This investigation will consist of a continuous cycle of compiling the data, clustering the data, and running the model. Finding valuable pieces of literature and additional cancer-type data sets with single-cell DNA sequencing will test whether the model continues to be effective with other data types, specifically other cancer types.