CIS 635 Project Proposal

Jacob Morrison

Due: 10 October 2025

- **Project Name:** Packing the Bits
- **Group Member(s):** Jacob Morrison
- **Project Title:** Investigating Tissue Differences in the Human Fallopian Tube across Ovarian Cancer Subtypes

1 Project Overview

I work at Van Andel Institute doing bioinformatics software development for a group that studies the origin of ovarian cancer. One of the graduate students in the group is working on a project that has collected two types of genomic data (RNA sequencing [RNA-seq] and whole genome bisulfite sequencing [WGBS]) in two different tissues (stromal and epithelial) across three different subtypes of ovarian cancer (high grade serous ovarian carcinoma [HGSOC], clear cell ovarian carcinoma [CCOC], and endometrioid ovarian carcinoma [ENOC]). One of her findings [1] is that differences in the stormal tissue are not driven by the type of cancer, but are instead driven by other, yet-to-be determined factors. This was done with the RNA sequencing data, but not with the WGBS data, so I would like to see if I can replicate, or at least approximate, those findings with the WGBS data. Whereas her work used a variety of tools dedicated to analyzing RNA-seq data, I will be using some of the techniques learned in class, including looking at the similarity and dissimilarity of the samples, analyzing the correlation and clustering of the results, and utilizing machine learning, to try and approximate those results. Depending on the timing of the project relative to finalizing of the paper manuscript, I may be able to work my results into the paper before it is submitted for peer review.

2 Related Work

As mentioned in Section 1, this project is building off a current project within the lab [1]. That project has utilized a variety of existing bioinformatic analysis tools to show that RNA expression differences between stromal tissues associated with three different ovarian cancer subtypes are not driven by the subtype, but rather by some other set factors, which are still being investigated. Two ways that my work will differ from hers is by using a related dataset (DNA methylation data rather than RNA expression data) and creating my own analysis pipeline. First, the driver of these differences was found using RNA expression data. As part of the experiment, DNA methylation data from WGBS was also collected. I will use this dataset to see if I can recover the same (or similar) results as in the RNA expression data. Second, the analysis performed in the lab used several pre-existing tools to perform the analysis. I, instead, will use a variety of techniques that we learned in class to perform a more rudimentary analysis.

3 Data Plan

I will be using a dataset of 92 samples collected from 58 tumors to study differences between epithelial and stromal tissue in HGSOC, CCOC, and ENOC. The data has already been processed, and I have been given access to the processed data files which contain information for every sequenced CpG (locations where DNA methylation primarily occurs in the genome). The information provided includes the chromsome, position on the chromosome, fraction of reads (sequences of DNA bases extracted from the genome) with methylated CpGs, and the number of reads that spans that position (called "coverage"). For preprocessing, I will need to remove samples that don't have matching RNA expression data, set minimum minimum coverage limits, and decide how I want to handle missing data.

NOTE: The data I will be using is protected human genetic data. Therefore, the results I show will be from my data, but I will be providing a mock dataset for testing how the project runs for grading.

4 Implementation Plan

As an overview, my pipeline will include the following steps: 1) reading and preprocessing data, 2) calculating descriptive statistics split by tissue and cancer subtype, 3) calculating a dissimilarity matrix between each of the samples, 4) clustering the data to find which samples are similar to one another, and 5) creating a basic machine learning pipeline. The overall pipeline will be written in Python and utilize Pandas (though I may try Polars instead) for data reading and descriptive statistics calculations. The dissimilarity matrix will be calculated by hand in Python, and the clustering and machine learning pipeline will use scikit-learn.

5 Evaluation Plan

I will evaluate the results of my algorithm in two ways: 1) confirming visually that the top two principal components in a principal component analysis (PCA) are are not driven by cancer subtype in the stromal tissue, and 2) computationally by seeing if a clustering algorithm finds a similar set of samples within each cluster as those that cluster together in the RNA expression data (as provided by the graduate student in the lab). If successful, I will see the PCA is not driven by cancer subtype and the clusters overlap between the two datasets. In terms of accuracy, I will treat the RNA expression data as ground truth and compare against results from that study.

6 Group Collaboration Plan

I will be working alone, so I will not need to manage group collaboration. That said, I will be utilizing GitHub for version control and for providing sample datasets to test how the data mining pipeline works.

7 Timeline

(On following page)

Week	Goals
7	Finalize and submit proposal
8	Read data with Python, Create mock dataset
9	Normalize, handle missing data
10	Calculate descriptive statistics
11	Calculate dissimilarity matrix
12	Perform clustering analysis
13	Set up and run ML models
14	Write up final report
15	Turn in

Table 1: Project timeline from midterm until the end of class.

References

[1] Djirackor, S., Heinze, K., Beddows, I., Sokol, D., de Souza, B.R., Koebel, M., Adams, M., Anglesio, M., and Shen, H. Parallel epigenetic and transcriptomic profiling of Carcinoma and Stroma Compartments across major Ovarian Cancer Histotypes. Poster presented at: AACR Special Conference in Cancer Research: Advances in Ovarian Cancer Research; 19–21 September 2025; Denver, CO.