CIS 635 Final Report

Jacob Morrison

Due: 5 December 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 Introduction

Ovarian cancer comes in many varieties, including high grade serous ovarian carcinoma (HGSOC), clear cell ovarian carcinoma (CCOC), and endometrioid ovarian carcinoma (ENOC). HGSOC is the most prevalent form and patients generally respond well to treatment, though there is a high incidence of recurrence. CCOC and ENOC both occur in the endometrium; however, they have very different prognoses and responses to treatment. CCOC is usually diagnosed early, but has a relatively poor long term prognosis as the cancer does not respond well to treatment. ENOC, on the other hand, responds well to treatment and usually has a better prognosis than CCOC patients. Due to the prevalence of HGSOC and the interesting fact that CCOC and ENOC derive from the same tissue, but have different outlooks, our group at Van Andel Institute (VAI) focuses on these three types of ovarian cancer.

For this project, I used a dataset collected by one of the graduate students in the lab. The cohort this dataset is derived from includes patients with HGSOC, CCOC, or ENOC. Tissue samples were collected from each patient and then epithelial and stromal cell samples were subsequently dissected from each tissue sample. Then, RNA sequencing (RNA-seq) and whole genome bisulfite sequencing (WGBS) was performed on the epithelial and stromal cell samples.

Using this dataset, the graduate student found that differences in the stromal cell samples were not driven by the type of ovarian cancer [1], but were driven by some yet-to-be determined set of factors. This result was primarily found using the RNA-seq data, so this project focused on if I could recapitulate the results with the WGBS data. My approach focused on a basic statistical analysis of the available samples and a principal component analysis (PCA) to determine what drives differences between the samples. The project found the main driver of differences in the PCA to be cell type (epithelial versus stromal), rather than the stromal group found with the RNA-seq data. This makes sense as the epithelial cells are where the cancer is whereas the stromal cells are normal cells adjacent to the cancer tissue.

2 Related Work

Previous work by members of the lab showed that differences between ENOC and CCOC likely derived from a cell state (rather than cell type) that was “locked in” prior to development of the cancer [2]. This work, and most other work on ovarian cancer [1], was performed using bulk sequencing where all cells (or a large number of cells) in the sample are included in the preparation for sequencing. This can result in mixed signals from adjacent normal and cancer cells. To overcome this limitation, laser capture microdissection [3] was performed on the collected tissue samples to delineate between cancer cells (associated with epithelial cells) and normal cells (associated with stromal cells). This delineation provided the necessary framework to make better comparisons between normal and cancerous tissue to probe what may be going wrong in the normal tissue to give rise to ovarian cancer.

As described in the Introduction, this project is based on work done in the RNA-seq data and presented in a poster session at a conference on ovarian cancer earlier this year. The manuscript for these results is currently being written and will be submitted shortly for peer review.

3 Methods

This data was originally collected from 92 epithelial and stormal samples. However, 10 samples had contamination and were removed. Therefore, 82 samples were used as input to this project. It should be noted that, due to the nature of using human data samples, I will only be providing figures generated from this data and not providing the raw data itself. For testing the processing pipeline, a Python script has been provided in the GitHub repo to generate test data files.

Python version 3.14 was used throughout the project. Additionally, several libraries were relied on to aid in processing (Table 1). A brief description of the data mining pipeline is provided below. For more details, see the link to the GitHub repository provided in Section 6.

Library	Version
python	3.14
pandas	2.3.3
numpy	2.3.3
matplotlib	3.10.6
scikit-learn	1.7.2
scipy	1.16.3

Table 1: Python and associated libraries used within the project. Also included are the specific versions used.

Raw data files were read in using Python. The raw beta value (i.e., the fraction of methylated cytosines at a given location in the genome) was retained, as well as calculating a logit-transformed M-value for use in the PCA. Positions with low coverage (coverage < 10) or not on the canonical human chromosomes (chromosomes 1-22, X, Y, and the mitochondria) were removed from consideration. In order to avoid missing data, only the positions with data in all samples were retained. The processed data was then written to a file for use in other portions of the pipeline.

Next, basic statistics were calculated on the per-sample basis and the per-position (called "CpG" here to represent a cytosine followed by a guanine in the genome) basis. A box plot of the distribution of raw beta values for each sample was created, while a violin plot was created to show the average beta value for each position across either the epithelial or the stromal samples.

Third, a dissimilarity matrix was calculated and plotted for the 82 samples. Equation 1 shows the dissimilarity metric used in calculating this matrix.

$$1 - \frac{\vec{x} \cdot \vec{y}}{||\vec{x}|| \times ||\vec{y}||} \tag{1}$$

To better define the distance between samples, the 10,000 positions with the highest variance were used in calculating the dissimilarity.

Finally, a PCA was performed and plotted with a variety of metadata used to label the data points. The M-values were used when performing the PCA to provided a dataset scaled between -1 and +1. Two components were retained in the PCA and accounted for 30% of the variance.

4 Experiments, Results, and Discussion

A TOML configuration file was used to readily assign variable inputs to the data mining pipeline. The default configuration is setup to process the example files, though these must be first generated by the user. Additional configuration files can be created for use with other datasets. Four values can be configured by the user:

1. The directory where the raw data lives (`data_dir`)
2. The path to the metadata TSV file (`meta_file`)
3. The number of processes to use when processing the raw files (`n_processes`)
4. A file name to save the preprocessed data to for downstream processing (`preprocessed_file`)

For parallel processing, the project code utilizes the `multiprocessing` library. Each raw input file is given its own process on which the preprocessing is performed.

The descriptive statistics are somewhat enlightening (Figure 1). There are several epithelial samples that show median beta values (which can also be described as methylation levels as shown) that are around 50%. In normal tissues, the median sits between 70–80%, though cancer samples have shown overall lower levels of methylation. So, while the decreased median is reasonable, the fact there are only a few samples is more surprising. The other noteworthy element is the few samples that have median beta values near 100%. I am not sure of why this may be the case, but it is something to look at. In terms of the mean and median methylation levels for each position across samples (Figure 1), it is further confirmation to see the epithelial (i.e., cancerous) cells have a lower beta level across the board when compared with the stormal (i.e., normal) cells.

The dissimilarity matrix (Figure 2) reveals a few epithelial cells that differ more from the other samples – both stromal and epithelial samples. The sample names are small, so it is difficult to confirm, but it would be an interesting next step to confirm if the higher dissimilarity samples match up with the samples with relatively low or high median sample beta values.

Unsurprisingly, the main driver of the variance observed in the PCA is cell type – whether the cells are epithelial or stromal cells (Figure 3). However, the distinction is along a diagonal axis across PC1 and PC2, so the difference is some linear combination of the two components. Therefore, I tried several other variables to see if any of those drove either PC1 or PC2. Of the three variables I checked, none of these were shown to drive either of the observed principal components.

5 Conclusion

This project suffers from several limitations compared to a traditional analysis that might be performed at VAI. The first is time. Typically, one analysis would be the focus of your work for many months. However, for this course, I spent much less time on the project than I would normally. Therefore, the depth of analysis is less than what would normally be expected in a publication. A second limitation is in how the PCA was performed. In order to confirm the results seen in the RNA-seq analysis, I should have looked at the epithelial and stromal samples separately in a PCA. Intuitively, it is obvious that normal and tumor samples would be different and be the main driver of differences across all the samples. In the future, I would perform two PCAs and then either project one into the other, or leave them separate and perform the analysis in parallel. Finally, the last limitation I want to address here is there is no connection between the raw data and any biological conclusions. Each analysis was performed across the entire genome and no attempt was made to look at specific genes or known regions of interest. This limitation is largely due to the exploratory nature of this project over against driving at a specific biological question.

The latter two limitations could easily be addressed with additional work. First, the tumor and normal samples could be separated into their own principal component analyses. This would allow me to confirm the result shown in the RNA-seq, while also allowing me to explore how the stromal clusters are related in the epithelial PCA. Secondly, I could work in biological regions of interest and see how samples clustered based on differences seen within those regions. These biological regions of interest would also allow me to investigate what biological functions (i.e., tying RNA expression and DNA methylation together) are driving differences between the tumor and normal samples.

All in all, the results shown here were more confirmatory of the quality of the data than noteworthy. Some of the samples exhibited differences from normal when looking at the descriptive statistics; however, these were largely contained to the cancer samples, which consistently behave differently than normal samples. This differing behavior was also seen in the dissimilarity matrix where the tumor samples were the samples that had the highest dissimilarity from the other samples. Finally, the PCA showed the main driver of differences between samples was due to differences between tumor and normal samples.

6 Data and Software Availability

Software for the project is available on GitHub at: https://github.com/GVSU-CIS635/projects-packingthebits. The original data shown in this report is not available due to patient confidentiality. For testing of the pipeline, instructions for generating example data are given in the repository's README.

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.

[2] Beddows, I., Fan, H., Heinze, K., Johnson, B.K., Leonova, A., Senz, J., Djirackor, S., Cho, K.R., Pearce, C.L., Huntsman, D.G., Anglesio, M.S., and Shen, H. Cell State of Origin Impacts

Development of Distinct Endometriosis-Related Ovarian Carcinoma Histotypes. Cancer Res. 2024; 84 (1): 26–38. doi: 10.1158/0008-5472.CAN-23-1362.

[3] Emmert-Buck, M.R., Bonner, R.F., Smith, P.D., Chuaqui, R.F., Zhuang, Z., Goldstein, S.R., Weiss, R.A., Liotta, L.A. Laser capture microdissection. Science. 1996; 274 (5289): 998–1001. doi: 10.1126/science.274.5289.998. PMID: 8875945.

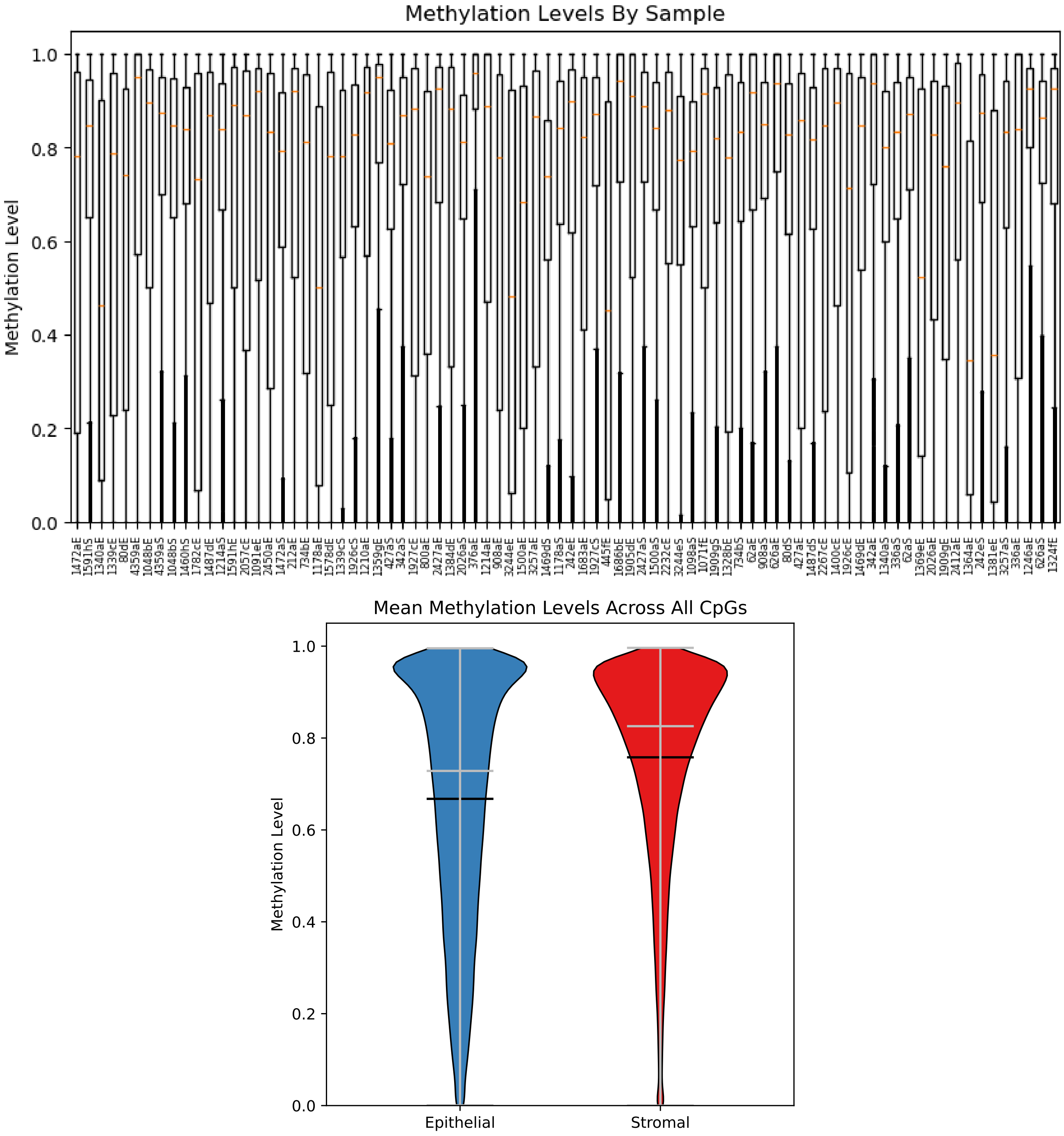

Figure 1: Descriptive statistics of per-sample beta value distributions (top) and the mean beta value for each position (CpG) across samples by cell type (bottom). The horizontal gray line shows the median beta value, while the black line shows the mean beta value.

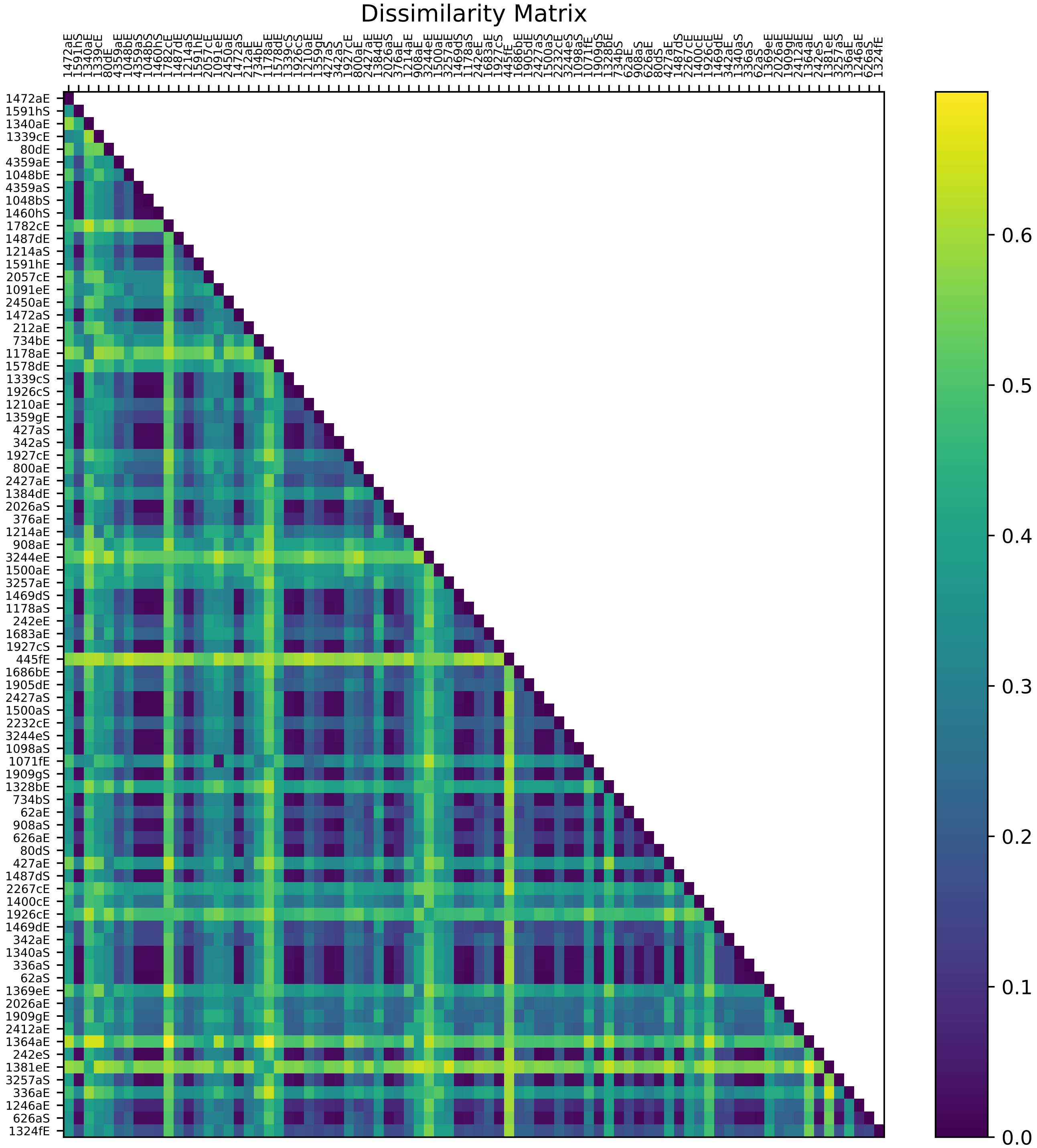

Figure 2: Dissimilarity matrix across samples.

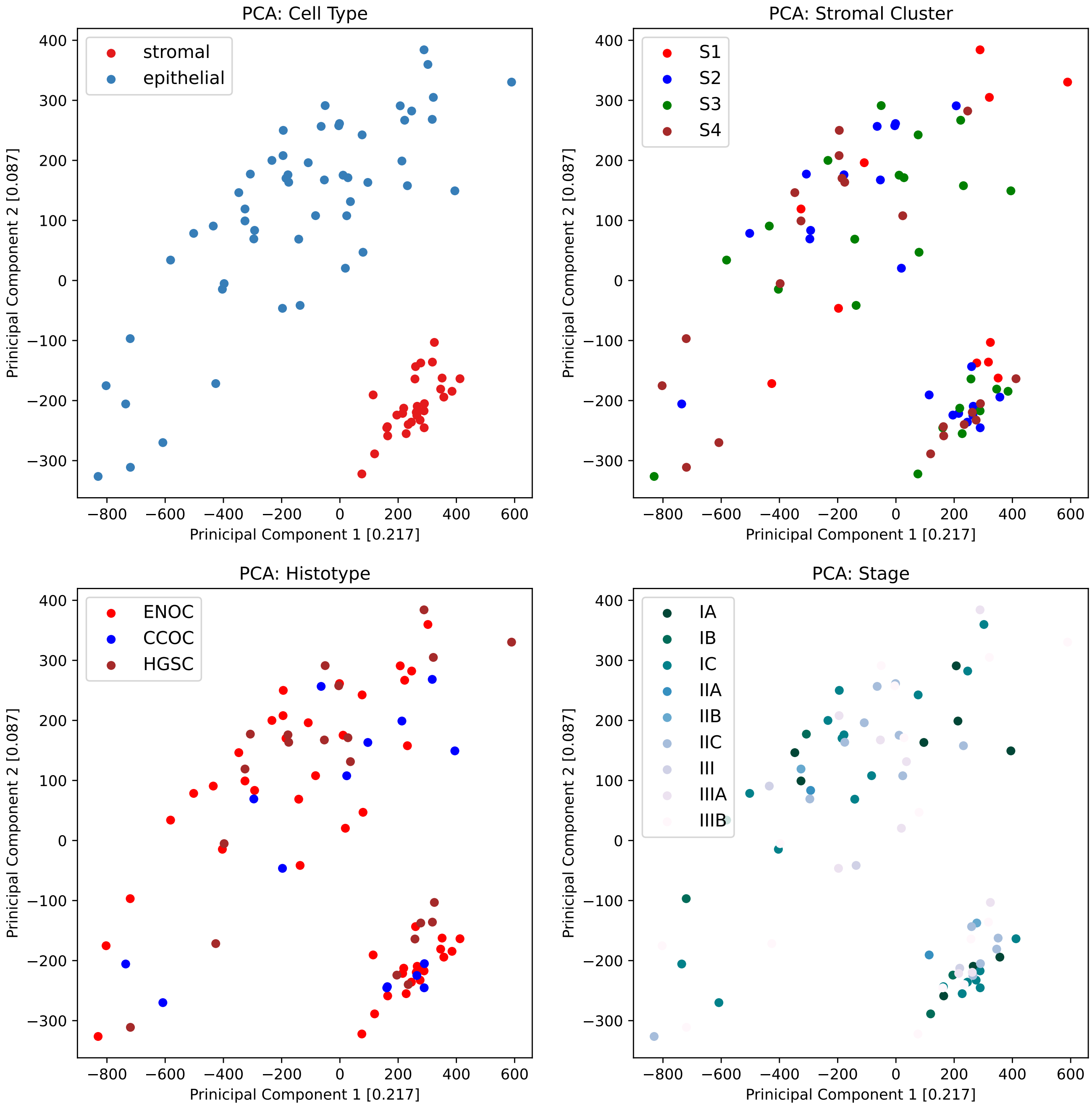

Figure 3: PCA shown for a variety of metadata, including cell type (top left), stromal cluster found in the RNA-seq analysis (top right), type of ovarian cancer (bottom left), and stage of ovarian cancer (bottom right).