Computational Genomics Resource Development for Cancer Diagnostics

Zihe Zhang
Professor Subhajyoti De

Introduction

Cancer has always been one of the leading causes of death around the world. It is crucial that we detect the signals for cancer or tumor development early on. For the current common method to detect cancer progression, it is usually a painful and time-consuming process for the patient. For example, common detection methods include X-rays, CT scans, and biopsy. Through exposure, the first two options can induce serious damage to one's tissue over a long term. The biopsy detection is basically collecting skin samples from a particular surface. It usually involves sticking needles through a patient's skin. It is a painful process and also the analysis is only limited to certain surfaces of the skin and cannot be conclusive of all parts of organs. Also, the doctor cannot accurately decipher the signals of cancer through X-ray pictures, or other types of pixelated analysis. Those types of digital evidence could be misleading due to the blurriness and fragmentation. A more accurate, cost-friendly method to detect cancer in the early stage of growth or recurrence can have an impact on clinical management of cancer patients.

Epigenetic signature is a type of chemical modification to the DNA molecule. Fragmented cell-free DNA carries epigenetic signals that regulate RNA's expression. It can provide information about their tissue's origin. Epigenetic signals can also be decomposed into data in forms of matrices consisting of 0 and 1s. We would be using that as the primary source of data for our research of constructing a linear regression model. Then, we would be able to identify how much each variable can contribute to the signals of cancer.

There are many other institutions that are working on the similar type of approaches as us and there are already many cancer detecting methods through the analyzation of RNA such as RNA-sequencing, Epigenetic Profiling, and Non-Coding RNA Biomarkers. However, we would try to solve this problem by studying the individual and cumulative effects of different parts of the blood. For example, we would study DNA fragment, Epigenetic signals that regulate RNA's genetic encoding. We are focusing on a multi-factor level of analysis instead of deep diving into a singular factor and how it marks the cues of cancer. Although, it is possible that not each of the approaches would have excellent predictive power, the comparative analysis will identify which ones are effective and whether any of those or a combined predictor outperforms current approaches.Our approach by studying the patterns within a patient's blood gives us an advantage to get information about all types of organs in the body because there would be DNA fragments from all over the place, proving way more reliable data than the needle approach.

One key factor that distinguishes the epigenetic approach is its unique ability to predict the tissue's origin solely through RNA/epigenetic signals. While methods for studying DNA fragments and microbiomes can provide clues and help narrow down potential locations, they lack the precision to confirm a specific region without errors.

The primary distinction between the microbiomes and RNA approaches lies in the signals and data they utilize, originating from different components of blood. The precise patterns of these signals are still open for discussion. Nevertheless, we are applying the same approach when calculating the coefficients of the variables.

Related Work

During the initial approach of the research, we must first understand the background and implications by reading the paper of [1] “Signatures Beyond Oncogenic Mutations in Cell-Free DNA Sequencing for Non-Invasive, Early Detection of Cancer.” Then, we can further solidify our knowledge on how genetic information is encoded and processed through a guide of codebase on [2] “Analysis of epigenetic signals captured by fragmentation patterns of cell-free DNA.” A more advanced data processing pipeline including data cleaning is then introduced by the codebase of [3] “A comprehensive quality control and analysis pipeline for cell-free DNA high-throughput sequencing data.” Overall, these sources can help me quickly grasp the topic and dive into a quick analysis on genetic information using python.

The first paper, [1] Signatures Beyond Oncogenic Mutations in Cell-Free DNA Sequencing for Non-Invasive, Early Detection of Cancer, points out that non-invasive cancer detection has long been a paramount objective in oncology. Highlighting the inherent challenges and limitations of existing detection strategies, the research underlines the transformative potential of genomic profiling of cell-free DNA (cfDNA) derived from liquid biopsies. One of the paper's central ideas is the use of massively parallel sequencing as a technique to detect multiple cancer types in a single assay. Such an approach stands in stark contrast to traditional methods, offering a broader window into the molecular landscape of cancer. Beyond merely identifying oncogenic mutation statuses, this method delves deeper, uncovering an array of molecular signatures. These signatures, as the research indicates, can offer invaluable insights into the potential tissue of origin, the extent of clonal proliferation, and the specific states of malignant diseases.

The second source of codebase, [2] Analysis of epigenetic signals captured by fragmentation patterns of cell-free DNA, provides a collection of scripts pivotal in reproducing the analytical methods detailed in another research paper titled "Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin."

Constructed using Python, the codebase employs libraries such as pysam, bx, and numpy to achieve its objectives. The repository meticulously details various techniques, including primary data processing of sequencing data, simulated read generation, nucleosome peak calling, and a comprehensive analysis of DHS sites. Furthermore, it annotates TFBSs and other pivotal genomic features, underlining the extensive interplay between cfDNA fragmentation patterns and epigenetic signals.

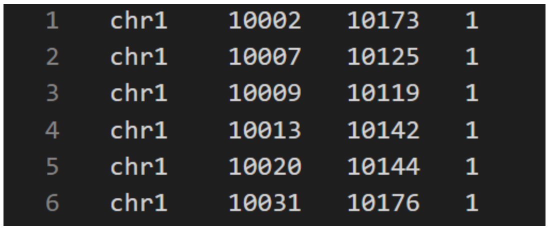
The third source of codebase, [3] A comprehensive quality control and analysis pipeline for cell-free DNA high-throughput sequencing data, also performs a similar procedure as source 2. However, the current codebase has a stronger compatibility for more formats of genetic information. A key feature of cfDNApipe is its capability to analyze bisulfite sequencing data. DNA methylation, a central epigenetic modification, plays a crucial role in gene regulation. Aberrant methylation patterns are frequently linked to diseases, especially cancer. Through its bisulfite sequencing analysis, cfDNApipe provides researchers with tools to explore these epigenetic changes in the context of cancer. This is a major difference between this and the second source since the second source is unable to differentiate the aberrant methylation within the data.

Although these sources have helped me significantly, my approach to detect epigenetic signals would be different from the previous codebases. With the help of existing genetic data processing codecase, I can compare my results with theirs to validify my algorithm's prediction accuracy. If the final prediction accuracy is not high enough to be considered as a practical approach, I will be closely comparing their methods and potentially improve my algorithms based on our differences.

Method

Data Preparation

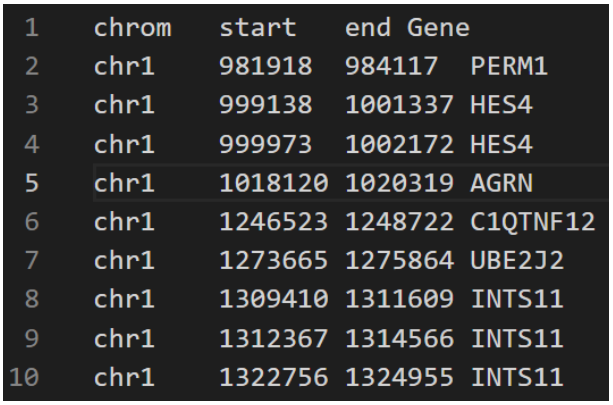
We have two sets of data. The first dataset, as shown in figure 1.1, comprises blood samples from individuals categorized as healthy, ovarian cancer patients, or gastric cancer patients. This dataset is structured into rows representing different blood samples, with four columns detailing the following: chromosome type, start location of the genome, end location of the genome, and the number of fragments (this one is not as important to our study).



```
chr1	10002	10173	1
chr1	10007	10125	1
chr1	10009	10119	1
chr1	10013	10142	1
chr1	10020	10144	1
chr1	10031	10176	1
```

Figure 1.1. A snippet of the first 5 rows of blood sample data for the gastric cancer group.

The second dataset, as shown in figure 1.2, lacks explicit health or disease labels. Instead, it is a standardized file that focuses on transcription start sites (TSS) and includes labeled gene types. Each row corresponds to a different transcription start site and contains the chromosome type, start and end locations of the genome segment, and crucially, the identified type of gene associated with that site.

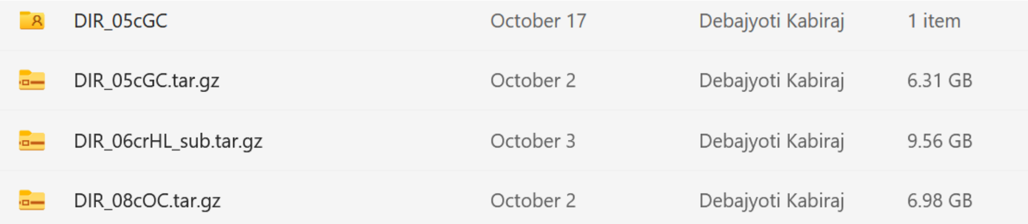


```
chrom	start	end Gene
chr1	981918	984117	PERM1
chr1	999138	1001337 HES4
chr1	999973	1002172 HES4
chr1	1018120 1020319 AGRN
chr1	1246523 1248722 C1QTNF12
chr1	1273665 1275864 UBE2J2
chr1	1309410 1311609 INTS11
chr1	1312367 1314566 INTS11
chr1	1322756 1324955 INTS11
```

Figure 1.2. A snippet of the first 10 rows of TSS data.

Computing Challenges with data operations

Since the data is already collected and preprocessed by previous researchers, we do not have to do any extensive data-cleaning. However, we still face run-time challenges when operating with such a huge amount of data set.



DIR_05cGC	October 17	Debajyoti Kabiraj	1 item
DIR_05cGC.tar.gz	October 2	Debajyoti Kabiraj	6.31 GB
DIR_06crHL_sub.tar.gz	October 3	Debajyoti Kabiraj	9.56 GB
DIR_08cOC.tar.gz	October 2	Debajyoti Kabiraj	6.98 GB

Figure 1.3. Blood sample data with their size in GB.

In our research, we have to go through each row of blood samples to get the most accurate genetic signal representation. If we decide to pick random samples from our large 10+ GB files, it would change the way we represent chromosome types because each piece of DNA fragment contributes to different types of chromosomes. So, to make sure our model is biologically valid overall, we need to carefully examine the entire dataset. Just as a reference, I tried creating a reference matrix with 3 blood samples and TSS data, and it took me about an hour for it to finish compiling by itself.

In order to efficiently analyze all of the data with a relatively reasonable run-time, a strong CPU, GPU, and high RAM is needed for the algorithms to compile and run properly. The solution for this is to use AMREAL which is the super computer designed for research purposes in rutgers. Their system requirement is definitely enough for our research and we can learn to use it to produce results quicker. One advantage is that we can let our tasks run in the background without constantly checking its progress.

Data Analysis

Upon obtaining our datasets, we begin with the analysis phase. The initial step involves examining the distribution and density of gene fragments across defined genomic intervals, which can be visualized through a bead-interval graph. This analysis utilizes only the first dataset, which includes blood samples from the healthy, ovarian cancer, and gastric cancer groups.

To conduct this analysis, we will calculate the number of gene fragments within each 100 base-pair window across the genome. This calculation will be represented as a matrix, where each element indicates the fragment count for a specific interval. After computing the matrix, we will plot it to visually inspect the distribution and density of gene fragments within these intervals across the different sample groups.

The objective is to replicate the analysis process for each group of blood samples: healthy, ovarian cancer, and gastric cancer. We will perform the same matrix calculations for each group, representing the quantity of gene fragments within each 100 base-pair window across the genome. We would want to create a density graph to better visualize the pattern of each blood group. One example graph for gastric cancer group is shown in figure 1.4.

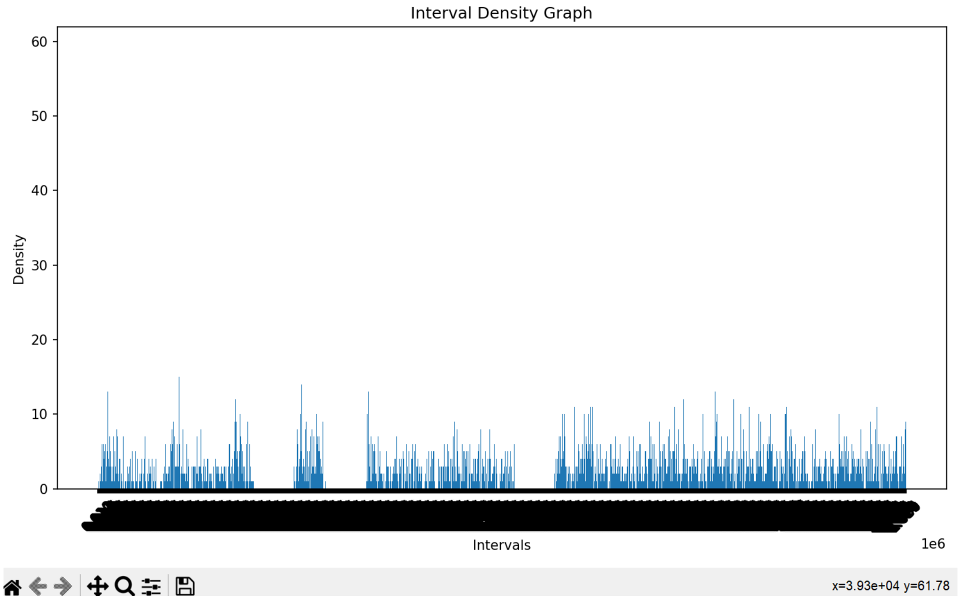


Figure 1.4 Interval graph for the first 10000 rows of blood data in a gastric cancer group.

Heatmap for Interval graph

To discern patterns within the vast and complex dataset, we can employ heatmaps, as shown in figure 1.5. These visual representations can effectively handle the large amount of data and numerous rows present in our study. By contrasting the color intensities on the heat maps generated for the healthy, ovarian cancer, and gastric cancer samples, we can identify genomic regions that warrant further investigation due to their differential expression.

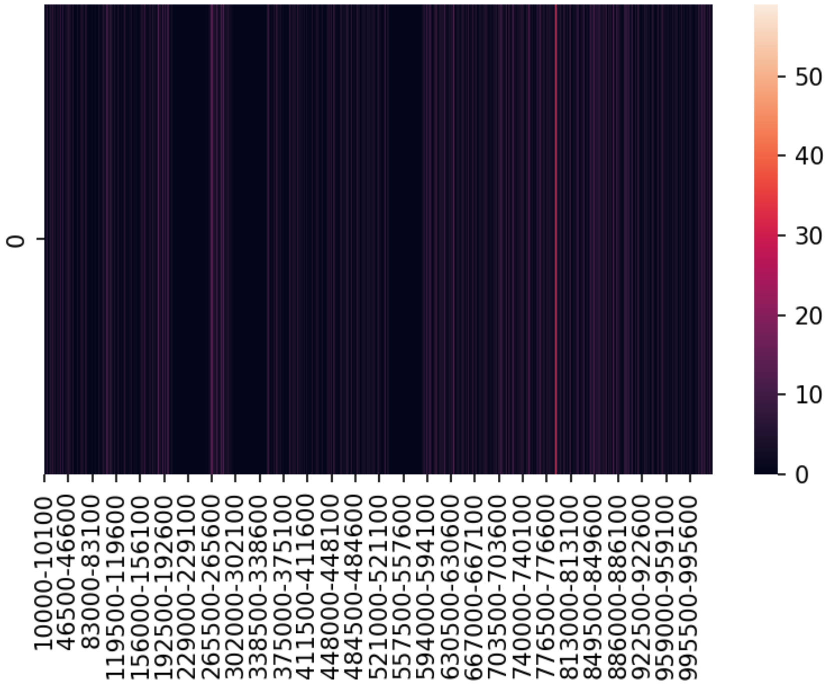
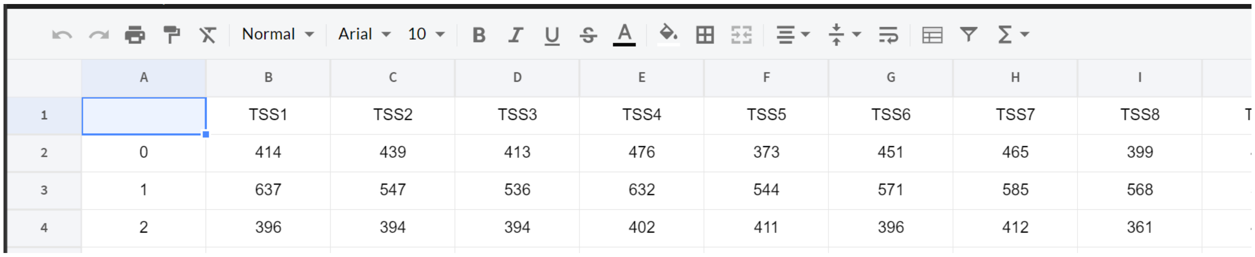


Figure 1.5. Heatmap for the previous interval data frame of the gastric cancer group.

Create Reference Matrix

Subsequent to the comparative analysis of the bead-interval graphs, we integrate the blood sample data with the transcription start sites (TSS) data, which includes labeled gene types. In this synthesis, we construct a matrix where the window size corresponds to the start and end locations of each TSS row, as shown in figure 1.6. For instance, if the TSS data indicates a start at position 5 and an end at position 100, then the first column of our matrix will cover the window from 5 to 100, continuing in this manner for each subsequent row.



	A	B	C	D	E	F	G	H	I	J
1		TSS1	TSS2	TSS3	TSS4	TSS5	TSS6	TSS7	TSS8	T
2	0	414	439	413	476	373	451	465	399	
3	1	637	547	536	632	544	571	585	568	
4	2	396	394	394	402	411	396	412	361	

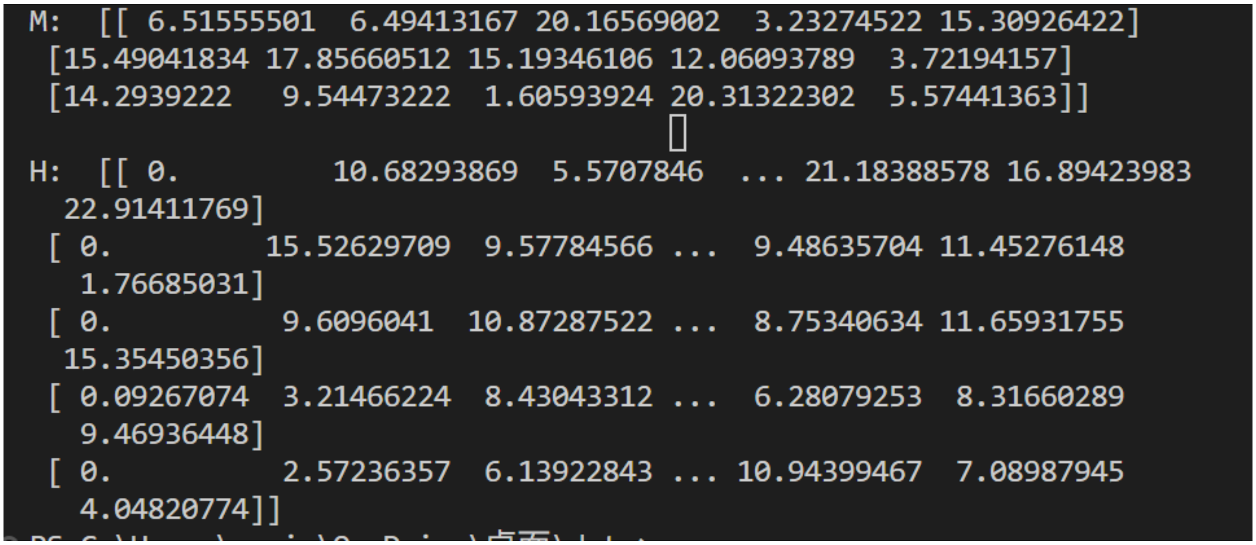
Figure 1.6. Resulting matrix from combining 3 blood samples in gastric cancer and TSS data.

This method differs from our initial graphing strategy in that it does not uniformly distribute window sizes but rather aligns them with the TSS data. While the first approach offers a broad overview of the gene fragment distribution across equal genomic intervals, it does not provide information about gene types. The second approach remedies this by specifically mapping the representation of each gene type, thereby granting us a detailed view of gene expression that could be pivotal for understanding the epigenetic variations between healthy and cancerous cells.

NMF

Upon constructing the matrix from the transcription start sites (TSS) data, we recognized the absence of labeled rows (Y-axis). Our challenge is to extract patterns and conduct analyses solely based on the matrix content. Non-negative Matrix Factorization (NMF) is a compelling first choice for this task. NMF excels in feature extraction and is widely applied in fields such as image processing, text mining, and bioinformatics, particularly for gene expression analysis. It is especially beneficial when dealing with large datasets containing thousands of TSS, as it can significantly reduce computational load by decreasing the number of variables needed to discern general data patterns.

A critical consideration in NMF is determining the appropriate number of components (N) to extract. To address this, we might construct heatmaps and employ clustering algorithms that focus on the rows and columns to guide the selection of N.



```
M:  [[ 6.51555501  6.49413167 20.16569002  3.23274522 15.30926422]
 [15.49041834 17.85660512 15.19346106 12.06093789  3.72194157]
 [14.2939222   9.54473222  1.60593924 20.31322302  5.57441363]]

H:  [[ 0.         10.68293869  5.5707846  ... 21.18388578 16.89423983
  22.91411769]
 [ 0.         15.52629709  9.57784566 ...  9.48635704 11.45276148
   1.76685031]
 [ 0.          9.6096041  10.87287522 ...  8.75340634 11.65931755
  15.35450356]
 [ 0.09267074  3.21466224  8.43043312 ...  6.28079253  8.31660289
   9.46936448]
 [ 0.          2.57236357  6.13922843 ... 10.94399467  7.08987945
   4.04820774]]
```

Figure 1.7. NMF trials on the previous resulting matrix of 3 gastric cancer blood samples with N=5.

PCA (planning to finish before end of this semester)

Parallel to NMF, we consider using Principal Component Analysis (PCA). PCA serves to eliminate data noise by retaining only the principal components with the highest variance, thereby filtering out minor variations that might obscure the significant patterns necessary for classification. By emphasizing the differences between each data category, PCA supports our ultimate objective of distinguishing between cancer types based on gene clusters.

After applying both NMF and PCA, we will compare the outcomes to assess their consistency and whether they display similar trends. This comparative analysis will enhance our confidence in the results and contribute to a more robust understanding of the underlying epigenetic signatures that differentiate various cancer types.

There is no example of running PCA on our samples yet. I will try to compute this before the fall semester ends.

Heatmap for Resulting TSS Matrix

A heatmap is an important tool for visualizing which parts of our columns have the most similar chromosome expression signals throughout all blood samples. By creating a heatmap of three gastric cancer blood samples versus TSS, we would get the following:

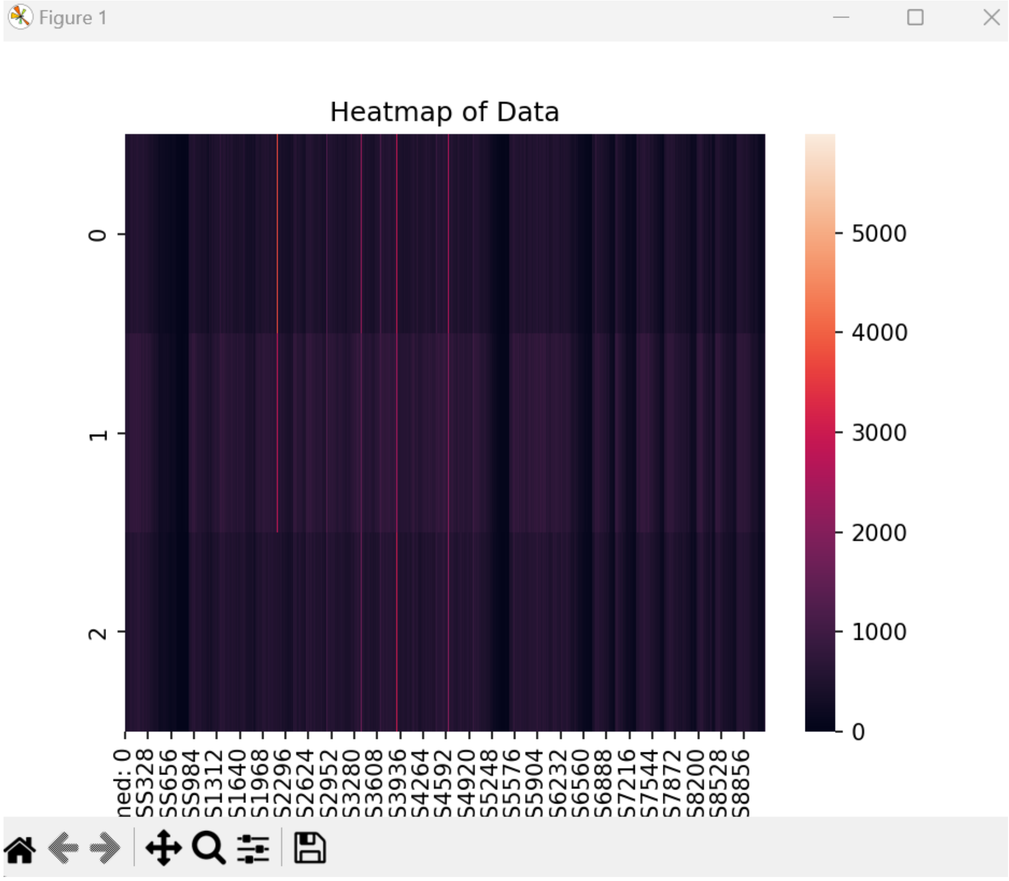


Figure 1.8. Heatmap of the resulting matrix from gastric cancer blood samples.

We have observed that certain transcription start sites (TSS) are more prominently represented in the gastric cancer samples. For example, the darker area from TSS656 to TSS984 indicates a lack of representation for those specific chromosomes. These dark spots might suggest that either that specific piece of signal is difficult to detect or it is missing. By focusing on the most expressed spots, such as those around TSS3930 (just before TSS3936 in the graph), we can conclude that for this subgroup of cancer data, one important feature is the sudden increase in signal from TSS3900 to TSS4000.

Clustering

Additionally, clustering algorithms such as hierarchical clustering or k-means can be utilized to categorize similar genomic windows. This method could highlight regions that exhibit consistent behavior across all samples, as well as pinpoint those that are unique to either healthy or cancerous states.

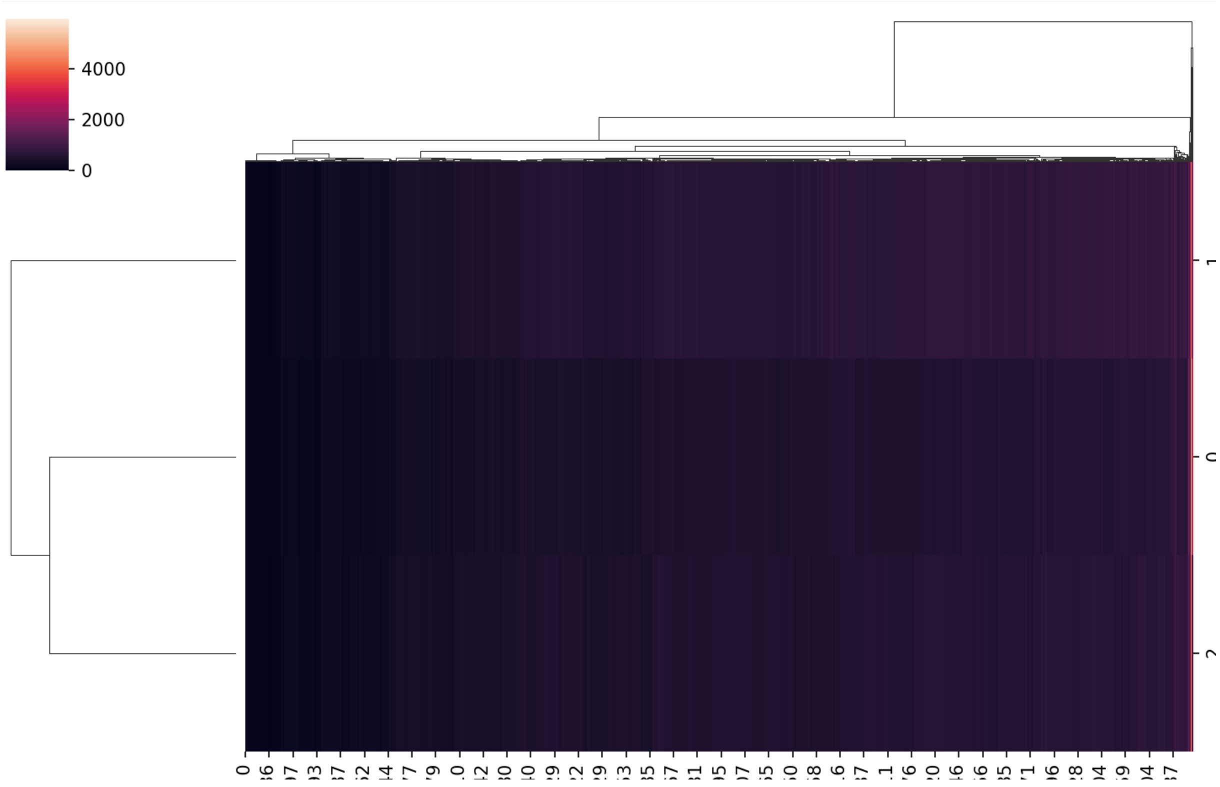


Figure 1.9. Hierarchical clustering to 3 gastric cancer samples.

Predictive Analysis (Planning to finish before the midpoint of spring semester)

Moving forward, although our matrices lack a Y-axis, we are aware of the sample classification—whether they pertain to healthy or diseased individuals. This information allows us to employ supervised learning techniques to classify the expression profiles. We will label the data within each matrix according to its known condition and use that to train a predictive model, such as a support vector machine or a random forest, to classify the condition of new samples.

A crucial preliminary step in this process involves data preprocessing to ensure that the gene expression levels are normalized and consistent across all matrices. This may necessitate normalization techniques and batch effect adjustments, particularly if the data were collected in different sessions or experiments. A significant challenge we face is establishing a method for normalizing data across various sample groups, which may require a deeper understanding of biomedical factors and the specifics of our dataset.

This part of the work will be expected to be done during the spring semester as we get more solid results from the previous PCA and NMF through all healthy, gastric cancer, and ovarian cancer groups.

Evaluation

To evaluate the results, we need to compare our predictions with the ground truth. This is crucial to establish the accuracy of our model from the original data. To examine the accuracy of a linear regression model, we can calculate the MSE (Mean Squared Error). It is extremely important in our study because we can easily identify if there is any outlier data that is negatively influencing our results. Another method is to find the PPS (Predictive Power Score). PPS is similar to the procedures in the linear model where we test how much can each variable contribute to the Y. We can ensure our method has no essential flaws if PPS has similar trends as our results. To avoid overfitting to the original data, we must also validate our estimations using results from other commonly employed cancer detection methods.

The evaluation will encompass the assessment of running time and the accuracy of cancer detection. Additionally, the cost of the test is a necessary consideration. We aim for an economically feasible model and a program that doesn't require high-end hardware settings. Preferably, the model should be budget-friendly and not demand substantial hardware resources. To further compare our accuracy across different population groups, we can employ techniques like cross-validation, address batch effects, and down-sampling.

Furthermore, we must evaluate how effectively our model can predict the sample tissue's origins when compared to alternative methods. Our model, based on epigenetic patterns, should, at the very least, demonstrate significantly better tissue location detection capabilities compared to approaches based on DNA and microbiomes.

Bibliography

[1]
Paper
De, S. (2021). Signatures Beyond Oncogenic Mutations in Cell-Free DNA Sequencing for Non-Invasive, Early Detection of Cancer. Frontiers in Genetics, 12, 759832. https://doi.org/10.3389/fgene.2021.759832

[2]
Codebase

Jay Shendure's lab at the University of Washington. (2020). Analysis of epigenetic signals captured by fragmentation patterns of cell-free DNA (Version 1.0.12) [Codebase]. GitHub. https://github.com/shendurelab/cfDNA

[3]
Codebase
XWangLabTHU. (2022). cfDNApipe: A comprehensive quality control and analysis pipeline for cell-free DNA high-throughput sequencing data (Version 1.0.81) [Codebase]. Github. https://github.com/XWangLabTHU/cfDNApipe