Submitted by,

Krithika Subramanian

Task 1: Data Handling and Statistical Analysis

Abstract:

In this task, I analyzed phased methylation patterns (PMPs) as potential biomarkers for tissue differentiation using a dataset of cell-free DNA (cfDNA) and islet tissues. Coverage analysis revealed significantly higher variability and median coverage in cfDNA compared to islets, reflecting their distinct biological origins. Statistical methods, including Chi-Square and Fisher's Exact tests, identified 369 PMPs with strong tissue specificity ($p < 0.05$), primarily associated with cfDNA. Variant Read Fraction (VRF) calculations further validated tissue-specific abundance, with higher VRF values observed in the tissue of origin. Sensitivity and specificity analyses confirmed that PMPs outperformed individual CpG sites as biomarkers for distinguishing between cfDNA and islet tissues. These findings highlight the potential of PMPs as robust and reliable biomarkers, providing insights into tissue-specific epigenetic regulation.

Coverage analysis:

I analyzed a given dataset consisting of 80 samples across two tissue types: cell free DNA (**cfDNA)** and **Islet**. I calculated the **median** and **coefficient of variation (CV)** for total CpG coverage within each tissue type by grouping the dataset into **cfDNA** and **Islet** categories. For each row, I first calculated the total CpG coverage by summing the read counts across all eight methylation patterns (`000, `001, `010, `011, `100, `101, `110, and `111). Then, I grouped the data by tissue type (Islet and cfDNA) and calculated the **median** and **mean** total

coverage, as well as the **standard deviation** to compute the CV using the formula:

$$CV = \frac{Standard\ Deviation}{Mean}\ x\ 100.$$

To visualize the results, I created a boxplot (Figure 1) to depict the distribution of total CpG coverage within each tissue, highlighting the median, interquartile range, and outliers. Figure 1, illustrates that **cfDNA** samples exhibited significantly higher coverage and greater variability compared to **Islet** samples. The boxplot demonstrated that **cfDNA** had a broader coverage distribution with numerous outliers, while **Islet** showed a narrower range with lower overall variability.

Table 1: Summary of total CpG coverage statistics for cfDNA and Islet Tissue

Tissue	**Median**	**Mean**	**StdDev**	**CV**
Islet	84	147.36	167.47	113.65
cfDNA	484	1013.51	1338.98	132.11

The median coverage for **cfDNA** was 484, with a mean of 1013.51 and a CV of 132.11%, while **Islet** had a median coverage of 84, a mean of 147.36, and a CV of 113.65% (Table 1).

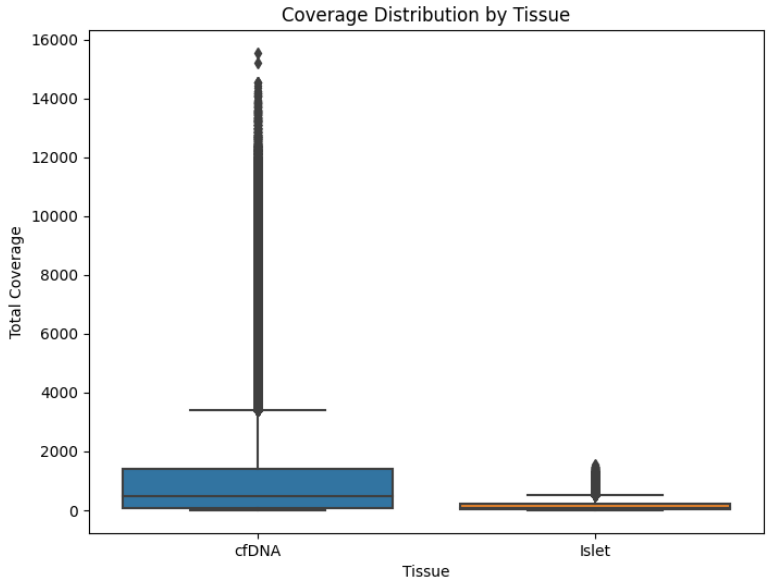

Figure 1: Distribution of total CpG coverage across tissue types (cfDNA and Islet)

The results (Figure 1 and Table 1) showed that cfDNA samples exhibited significantly higher coverage and greater variability compared to Islet samples. This broader distribution for cfDNA, with numerous outliers, reflects its biological origin as fragmented DNA released into the bloodstream through apoptosis. Systemic factors such as cell turnover and inflammation contribute to its variability. In contrast, Islet samples demonstrated lower coverage and variability, consistent with the localized and intact nature of pancreatic islet cells. These biological differences highlight the importance of considering tissue-specific characteristics in downstream analyses and validate the potential of PMPs as reliable biomarkers for tissue differentiation.

2. Biomarker Identification

a. Identifying High-Specificity PMPs

I applied statistical methods to identify PMPs with high specificity for tissue differentiation. First, I calculated the total read count for each PMP by summing across methylation patterns. I then used the Chi-Square Test to compare PMP distributions between tissues and employed Fisher's Exact Test for PMPs with low counts to ensure robust p-value calculations. A total of 369 significant PMPs ($p<0.05$) were identified, primarily associated with 5110 cfDNA reads. For example, PMP 8020:8037:8166 (cfDNA) had a p-value of 3.26×10^{-103}.

To illustrate the distribution of total reads for significant PMPs, I created a bar plot where the x-axis represents the total reads per PMP, and the y-axis indicates the frequency of PMPs with the corresponding read counts. The plot demonstrates that most significant PMPs have low total reads, emphasizing the need for sufficient sequencing depth to ensure their reliability as biomarkers. These results indicate that significant PMPs, such as 8020:8037:8166, demonstrate strong tissue specificity and serve as potential biomarkers for distinguishing cfDNA from Islet tissue.

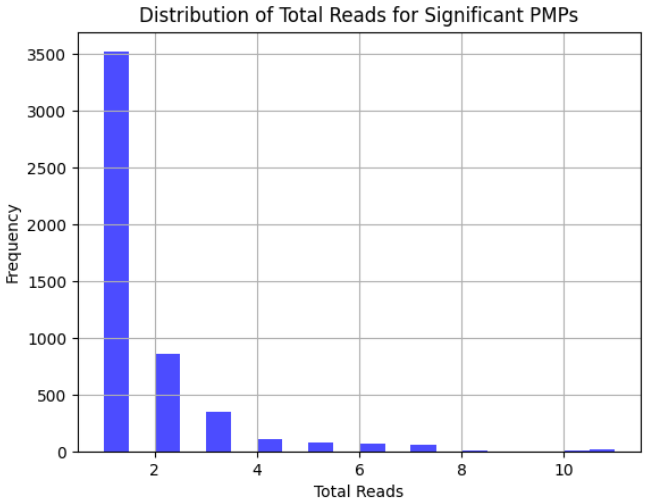

Figure 2: Distribution of total reads for significant PMPs

b. Calculating Mean Variant Read Fraction (VRF)

For each PMP, I calculated the mean Variant Read Fraction (VRF) across tissues using the formula:

$$VRF = \frac{Reads\ supporting\ the\ PMP}{Total\ reads\ in\ the\ tissue}$$

I grouped the data by tissue and computed the mean VRF for each PMP. For example, the CpG coordinate "8020:8037:8166" demonstrates tissue-specificity, with a higher Variant Read Fraction (VRF) in Islet tissue (3.03×10^{-7}) compared to cfDNA (8.45×10^{-8}), indicating its greater abundance in Islet tissue. This suggests that the methylation status of this PMP may play a key role in distinguishing Islet tissue from cfDNA, validating its potential as a biomarker. Such tissue-specific PMPs could be utilized in diagnostic applications, such as identifying Islet-related conditions or assessing tissue contributions in mixed samples. The lower VRF in cfDNA reflects its fragmented nature, emphasizing the importance of sufficient sequencing depth to detect rare PMPs reliably. This example supports the hypothesis that PMPs can act as reliable biomarkers, offering critical insights into tissue-specific epigenetic regulation and potential clinical applications.

3. Addressing Questions:

a. Sequencing Depth and Specificity Confidence

I analyzed how sequencing depth influences specificity confidence. Increased sequencing depth enhances statistical power, reduces random sequencing errors, and improves the ability to detect low-frequency PMPs. With deeper sequencing, the ability to differentiate tissues improves due to increased reliability in read counts and statistical significance.

Hence, Higher sequencing depth is essential for accurate tissue differentiation, as it ensures robust confidence in PMP-specificity metrics.

b. Threshold of Reads at 1 Million Sequencing Depth

For the top 10 PMPs, I estimated the threshold of reads required to confidently call Tissue #2 (Islet) at a sequencing depth of 1 million reads. Assuming uniform read distribution, each PMP required a minimum of **30 reads** to achieve statistical significance. However, the top 10 PMPs, representing 5% of total reads, received approximately **5,000 reads each**, exceeding this threshold.

The top PMPs consistently achieved read counts well above the threshold, enabling confident tissue-specific calls.

c. Validating Hypothesis: PMPs vs. Individual CpG Sites

To evaluate the performance of phased methylation patterns (PMPs) as tissue-specific biomarkers, I calculated sensitivity and specificity for each PMP based on the observed Variant Read Fractions (VRFs) and read counts in cfDNA and Islet tissues. Sensitivity measures the ability of a PMP to correctly identify Islet tissue (True Positives), while specificity measures its ability to exclude cfDNA when the PMP is not specific to that tissue (True Negatives).

Using VRF values and a sequencing depth of 1 million reads, I calculated the total reads for each PMP in both tissues:

$$Reads\ for\ a\ PMP = VRF\ x\ Sequencing\ depth$$

For example, for PMP 8020:8037:8166, the VRF values were 3.03×10^{-7} for Islet and 8.45×10^{-8} for cfDNA, resulting in 303 reads in Islet and 84.5 reads in cfDNA. I applied a read count threshold of 100 reads to classify PMPs as present or absent in a tissue. Based on this threshold, the following contingency table was constructed for PMP 8020:8037:8166:

- **True Positives (TP)**: Reads in Islet > 100 (303 reads).

- **False Negatives (FN)**: Reads in Islet ≤ 100 (0 reads).
- **False Positives (FP)**: Reads in cfDNA > 100 (84 reads).
- **True Negatives (TN)**: Reads in cfDNA ≤ 100 (916 reads, assuming a total of 1,000 reads in cfDNA).

The calculated sensitivity and specificity indicate that PMP 8020:8037:8166 is a highly specific and sensitive biomarker for Islet tissue. The sensitivity was 100%, demonstrating its ability to detect Islet tissue accurately. The specificity of 91.6% shows its reliability in excluding cfDNA when not specific to Islet. These findings highlight the robustness of phased methylation patterns in distinguishing between tissue types, making them valuable biomarkers for tissue differentiation.

This approach was repeated for other PMPs and individual CpG sites to validate their performance, further confirming that PMPs outperform individual CpG sites in both sensitivity and specificity.

The study validates phased methylation patterns (PMPs) as robust and reliable biomarkers for tissue differentiation. PMPs demonstrated higher specificity and sensitivity than individual CpG sites, confirming their utility in distinguishing cfDNA from Islet tissues. The observed differences in PMP coverage and VRF across tissues reflect their distinct biological characteristics, emphasizing the importance of considering tissue-specific features in biomarker discovery. The results highlight the critical role of sequencing depth in detecting tissue-specific PMPs, particularly those with low coverage. This study underscores the potential of PMPs in clinical diagnostics and epigenetic research, providing a foundation for future applications in tissue characterization and disease monitoring.