Overview of the Assignment:

Task 1 was to Analyse the methylation Datasets by interpreting the phased methylation patterns using statistical tests.

Task 2 was to handle the NGS dataset by performing the Alignment and Mutation Calling which includes processing raw sequencing data and aligning the reads to the reference genome and identifying somatic mutations.

Task_1 : Data Handling and Statistical Analysis.

Data Overview: The data given is about CpG methylation. DNA methylation is an epigenetic changes. These are modifications on DNA that regulate whether genes are turned on or off. Its transfer of methyl group onto C5 position of Cytosine to form 5 -methylcytosine. CpG is Cytosine bonded to Guanosine with phosphate backbone. There are 28 million CpGs in Human Genome where 60% to 80% are generally methylated. CpG methylation serves as Epigenetic marker that varies across tissue types. Phased Methylation Pattern (PMP) is a unique set of coordinates that includes the DNA strand ('f' for forward (+) or 'r' for reverse (-)), the relative positions of three CpG sites on the same strand (e.g., x:y:z), and their methylation status (e.g., '000' for all unmethylated or '111' for all methylated). It represents a combined epigenetic signature across these CpG sites.

The Dataset contains phased methylation patterns from NGS results across two tissues that is cfDNA (Cell free DNA) and Islet. The Dataset has 15392183 Rows and 13 Columns which is very large Data which contains features like:

Strand : Indicates the DNA strand whether its is forward 'f' or reverse 'r'.

CpG Coordinates : It has relative positions of Three CpG sites(x:y:z)

Methylation Status : Eight possible patterns ('000' to '111').

Sample ID : Unique identifier for each sample.

Replicate : Indicates technical replicates

Tissue : Tissue type (cfDNA or Islet)

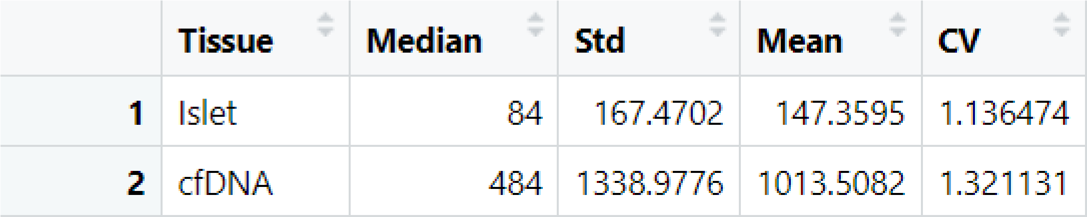
1. Coverage Analysis

   a. Calculate the median and coefficient of variation(CV) for single CpG coverage in each tissue.

      The coverage_stats table provides basic statistics about sequencing coverage for the two tissues (Islet and cfDNA), including:

      median_coverage: The median number of reads covering each genomic region.

      cv_coverage: The coefficient of variation (CV) of the coverage, which is calculated as the ratio of the standard deviation to the mean coverage, representing variability in sequencing depth.



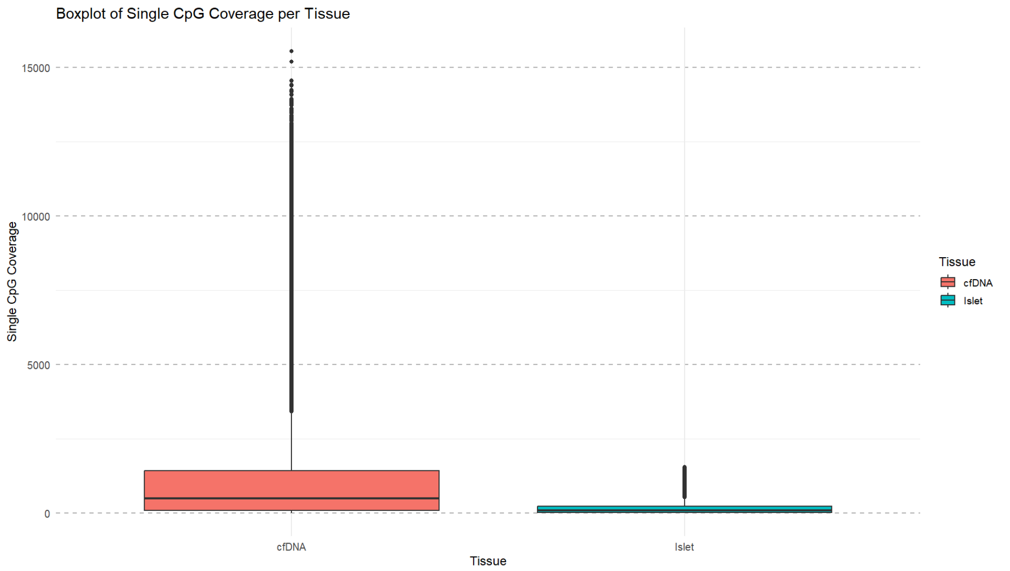
	Tissue	Median	Std	Mean	CV
1	Islet	84	167.4702	147.3595	1.136474
2	cfDNA	484	1338.9776	1013.5082	1.321131

Both **Islet** and **cfDNA** tissues show **high variability** in single CpG coverage (CVs above 100%), indicating that CpG coverage is not consistent across samples within each tissue type.

cfDNA has higher median coverage than **Islet**, but the coverage is still quite variable in both tissues, suggesting that there are areas with much higher and much lower coverage in both tissues.

The relatively high CVs in both tissues suggest that while the tissues may have regions of consistent coverage, there are also regions where CpG sites have very low or very high coverage, pointing to potential issues in data consistency or biological heterogeneity.

b. Generate plots summarizing the coverage statistics

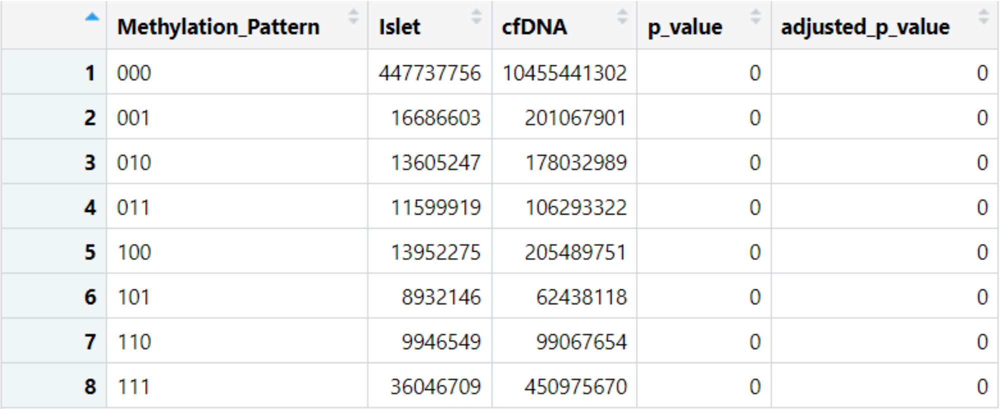


Boxplot of Single CpG Coverage per Tissue

2.Biomarker Identification

a. Identify PMPs with high specificity for tissue differentiation, minimizing false positives for Tissue #1 while allowing some false negatives. Use statistical or machine learning approaches to assign confidence (e.g., p-values) to each PMP.

This analysis uses a statistical approach to identify significant methylation patterns that differentiate between two tissue types (`cfDNA` and `Islet`). First, the data is summarized by aggregating the total coverage for each methylation pattern across tissues. A contingency table is created to compare these coverage values. Chi-squared tests are performed for each pattern to test whether the distribution of methylation between the tissues is significantly different. The p-values from these tests are adjusted using the False Discovery Rate (FDR) method to account for multiple comparisons, ensuring robust identification of significant patterns. Patterns with adjusted p-values below 0.05 are considered significant.



	Methylation_Pattern	Islet	cfDNA	p_value	adjusted_p_value
1	000	447737756	10455441302	0	0
2	001	16686603	201067901	0	0
3	010	13605247	178032989	0	0
4	011	11599919	106293322	0	0
5	100	13952275	205489751	0	0
6	101	8932146	62438118	0	0
7	110	9946549	99067654	0	0
8	111	36046709	450975670	0	0

The results identified eight methylation patterns with highly significant differences (adjusted p-value = 0), indicating strong tissue-specific methylation. For example, the pattern `000` showed a stark difference in coverage between `cfDNA` (10.45 billion reads) and `Islet` (447 million reads). These findings highlight potential biomarkers for distinguishing between tissues based on methylation patterns, providing a foundation for further validation and exploration.

Potential Biomarkers:

From the results:

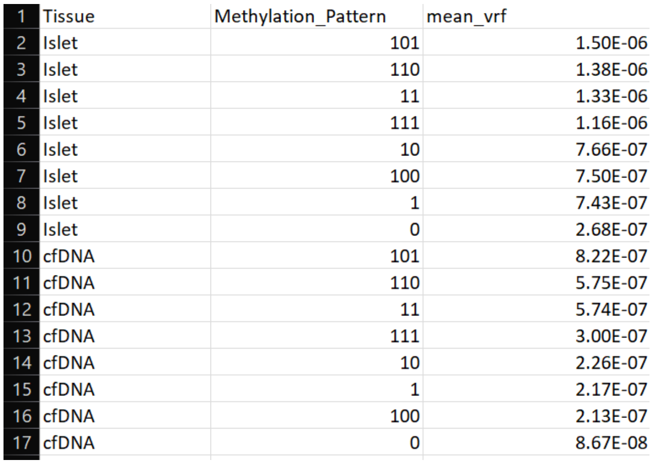
- **Pattern `000`**: High coverage in `cfDNA` (10.45 billion reads) compared to `Islet` (447 million reads).
- **Pattern `001`**: Strongly enriched in `cfDNA` (201 million reads) versus `Islet` (16.6 million reads).
- **Pattern `011`**: Significantly higher coverage in `cfDNA` (106 million reads) compared to `Islet` (11.6 million reads).

  These three patterns can be considered as potential **biomarkers** for differentiating between tissues, specifically between **cfDNA** and **Islet** because of :

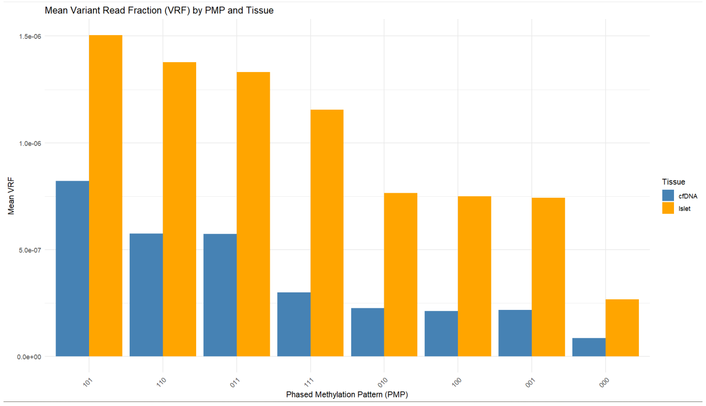
  High specificity: These patterns show substantial coverage differences between the two tissues, with cfDNA having much higher coverage in these methylation patterns compared to Islet.

  Statistical significance: The chi-squared tests for these patterns returned p-values of 0, which are highly significant and, after adjustment for multiple testing, remain below the typical threshold (adjusted p-value < 0.05).

b. Calculate the mean variant read fraction (VRF) for each PMP in both tissues.



1	Tissue	Methylation_Pattern	mean_vrf
2	Islet	101	1.50E-06
3	Islet	110	1.38E-06
4	Islet	11	1.33E-06
5	Islet	111	1.16E-06
6	Islet	10	7.66E-07
7	Islet	100	7.50E-07
8	Islet	1	7.43E-07
9	Islet	0	2.68E-07
10	cfDNA	101	8.22E-07
11	cfDNA	110	5.75E-07
12	cfDNA	11	5.74E-07
13	cfDNA	111	3.00E-07
14	cfDNA	10	2.26E-07
15	cfDNA	1	2.17E-07
16	cfDNA	100	2.13E-07
17	cfDNA	0	8.67E-08



The analysis of the mean Variant Read Fraction (VRF) across Phased Methylation Patterns (PMPs) for the two tissues, Islet and cfDNA, reveals distinct trends. The Islet tissue consistently exhibits higher mean VRF values across all PMPs compared to cfDNA, suggesting that methylation patterns are more prevalent or detectable in Islet tissue.

Among the PMPs, `101` shows the highest mean VRF in both tissues, indicating its dominance or higher detection rate, while PMP `000` has the lowest mean VRF, reflecting its relative scarcity. This pattern demonstrates a gradual decline in VRF values from PMP `101` to PMP `000`.

The observed differences may be attributed to biological factors, such as the higher integrity and abundance of DNA in Islet tissue compared to the fragmented nature of cfDNA. These findings highlight the variability in methylation patterns between tissues and the potential use of specific PMPs, such as `101`, as biomarkers for distinguishing tissue types.

3. Address the following questions:

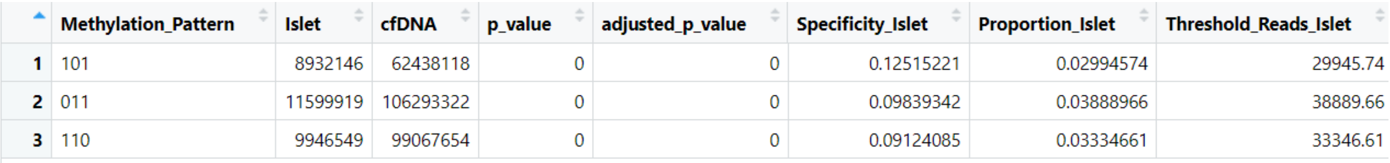
a. How does sequencing depth affect specificity confidence?

Adequate sequencing depth is critical to ensure sufficient coverage of PMPs, enhancing reliability and specificity in differentiating tissue types.

Sequencing depth significantly impacts specificity confidence by influencing the reliability of assigning methylation patterns or CpG sites to specific tissues. Higher sequencing depth improves coverage, reduces variability, and enhances the detection of rare patterns, resulting in greater confidence in tissue-specific assignments. This reduces noise and narrows confidence intervals, making statistical tests more robust and reliable.

Conversely, lower sequencing depth increases the likelihood of false negatives and false positives, compromising the accuracy of specificity measurements. In the current study, the higher median coverage in cfDNA (484 reads) compared to Islet (84 reads) demonstrates how greater depth supports more confident tissue differentiation, particularly for phased methylation patterns (PMPs).

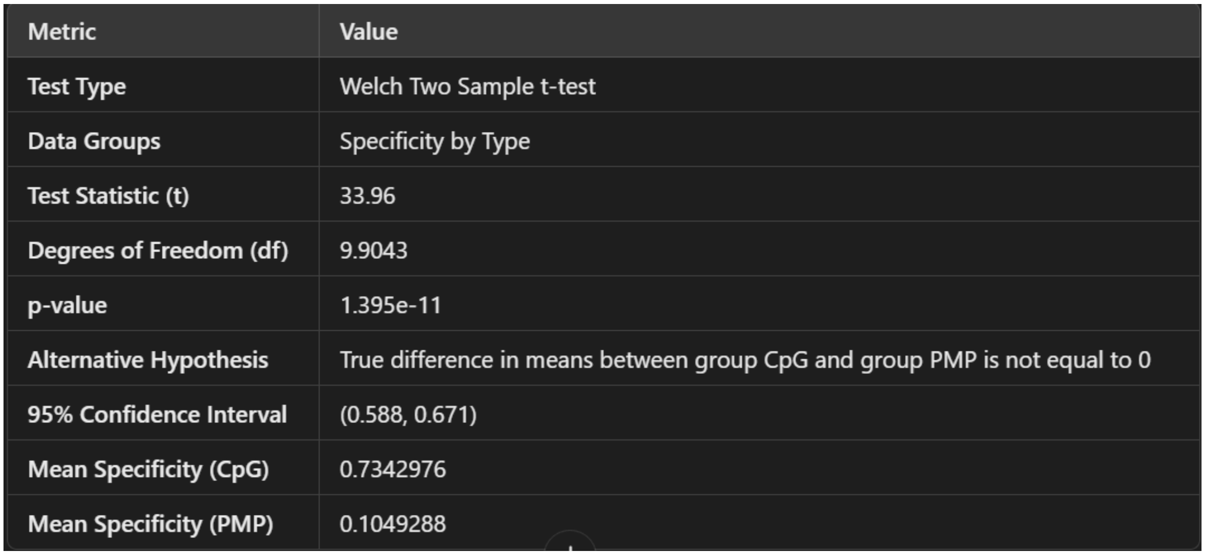
b. For the top 10 PMPs, estimate the threshold of reads required to confidently call Tissue #2 at a sequencing depth of 1 million reads.



	Methylation_Pattern	Islet	cfDNA	p_value	adjusted_p_value	Specificity_Islet	Proportion_Islet	Threshold_Reads_Islet
1	101	8932146	62438118	0	0	0.12515221	0.02994574	29945.74
2	011	11599919	106293322	0	0	0.09839342	0.03888966	38889.66
3	110	9946549	99067654	0	0	0.09124085	0.03334661	33346.61

The **threshold reads** depend on the **specificity** and **proportion of reads** for the Islet tissue. Patterns with higher specificity require fewer reads to confidently assign them to Islet tissue.
Methylation Pattern **101** is the most efficient biomarker among the three, as it requires the least number of reads (29,946) to confidently call Islet tissue at a sequencing depth of 1 million reads.
These thresholds guide the sequencing efforts needed for accurate tissue differentiation and biomarker validation.

c. Validate the hypothesis by comparing the specificity of the top 10 PMPs against individual CpG sites.



Metric	Value
Test Type	Welch Two Sample t-test
Data Groups	Specificity by Type
Test Statistic (t)	33.96
Degrees of Freedom (df)	9.9043
p-value	1.395e-11
Alternative Hypothesis	True difference in means between group CpG and group PMP is not equal to 0
95% Confidence Interval	(0.588, 0.671)
Mean Specificity (CpG)	0.7342976
Mean Specificity (PMP)	0.1049288

Conclusion from the t-test:

- The p-value indicates a statistically significant difference in specificity between CpG sites and PMPs.
- The 95% confidence interval for the difference in means is **[0.588, 0.671]**, confirming the substantial gap in specificity.
- CpG sites have significantly higher specificity than PMPs.

Hypothesis Validation

The hypothesis stated:

"Phased methylation patterns (PMPs) can act as reliable biomarkers to differentiate tissue types, providing higher specificity compared to individual CpG sites."

Based on these results:

The hypothesis is rejected. Individual CpG sites show significantly higher specificity compared to PMPs, contradicting the assumption that PMPs provide superior tissue differentiation specificity.

Task 2: NGS Data Analysis

Dataset: The dataset consists of NGS samples in FASTQ format, including one sample from normal tissue and one from cancer tissue.

1. Quality Control

a. Perform quality checks using tools like FastQC and summarize quality metrics (e.g., sequence counts, per-base quality, read duplication levels).
Per Base Sequence Content (FAIL)

- High or low frequencies of specific bases (A, T, C, G) at certain positions in sequences may suggest contamination, sequencing errors, or improper library preparation.
- Further investigation is needed to identify and filter these problematic bases.

b. Per Sequence GC Content (FAIL)

- Variations in GC content can lead to uneven read mapping and affect the performance of downstream analyses. High or low GC content could indicate biases in sequencing or contamination.

c. Sequence Duplication Levels (FAIL)

- High duplication rates can reduce the diversity of the dataset and lead to overestimation of variant frequencies or other metrics.
- Consider filtering or removing duplicates using tools like `picard` or `MarkDuplicates`.

d. Overrepresented Sequences (FAIL)

- Overrepresented sequences suggest possible presence of adapter sequences, contaminations, or PCR artifacts.
- These sequences need to be identified and removed for improved data quality.

e. Adapter Content (FAIL)

- Presence of adapter sequences indicates that library preparation might not have been optimized, requiring trimming or removal of adapters.

1. Alignment and Mutation Calling

a. Align the samples to the human genome using tools like Bowtie2 or BWA
The fastq files of Normal and Tumor were given which had both reverse and forward files.
The reference genome used to align was hg19. Initially the reference genome was indexed and using the alignment tool Bowtie2 the samples were aligned to the human genome.

b. Identify somatic mutations present in the cancer sample but absent in the normal tissue.
i. Benchmark Software: Use established tools such as Mutect2, Strelka2, or VarScan2 for somatic mutation identification and background mutation estimation.

c. Custom Code Development: Write your own scripts, leveraging tools like Samtools, bcftools, or Python/R libraries, to perform mutation detection and calculate the required metrics.

```
# Index the Reference Genome
samtools faidx reference.fasta
# Align the Reads to the Reference Genome
bwa mem reference.fasta reads.fastq > aligned_reads.sam
# Convert SAM to BAM and Sort
samtools view -Sb aligned_reads.sam | samtools sort -o sorted_reads.bam
# Index the Sorted BAM File
samtools index sorted_reads.bam
# Call Variants
samtools mpileup -f reference.fasta sorted_reads.bam | bcftools call -mv -Ob -o variants.bcf
# Convert BCF to VCF
bcftools view variants.bcf > variants.vcf
```

```
# Load necessary libraries

if (!requireNamespace("VariantAnnotation", quietly = TRUE)) {
install.packages("BiocManager")
BiocManager::install("VariantAnnotation")
}
library(VariantAnnotation)
# Function to calculate Ti/Tv ratio
calculate_titv_ratio <- function(vcf_file) {
vcf <- readVcf(vcf_file, "hg19")
 # Extract reference and alternate alleles
 ref <- as.character(ref(vcf))
 alt <- as.character(unlist(alt(vcf)))
 # Define transitions and transversions
 transitions <- c("AG", "GA", "CT", "TC")
transversions <- c("AC", "CA", "AT", "TA", "CG", "GC", "GT", "TG")
```

```
 # Count transitions and transversions
ti_count <- sum(paste0(ref, alt) %in% transitions)
tv_count <- sum(paste0(ref, alt) %in% transversions)
# Calculate Ti/Tv ratio
 titv_ratio <- if (tv_count != 0) ti_count / tv_count else NA
return(list(ti_count = ti_count, tv_count = tv_count, titv_ratio = titv_ratio))
}
# Usage example
vcf_file <- "variants.vcf"
result <- calculate_titv_ratio(vcf_file)
cat("Transitions (Ti):", result$ti_count, "\n")
cat("Transversions (Tv):", result$tv_count, "\n")
cat("Ti/Tv Ratio:", result$titv_ratio, "\n")
```

This repository provides a comprehensive pipeline for variant detection and analysis, leveraging tools such as Samtools, bcftools, and R. It guides users through the process of indexing the reference genome, aligning sequencing reads, converting and sorting BAM files, and calling variants to generate a VCF file. The pipeline also includes an R script for calculating the Transition/Transversion (Ti/Tv) ratio from the VCF file. The R script uses the VariantAnnotation package to read and process the VCF file, count transitions and transversions, and compute the Ti/Tv ratio. Users are required to have Samtools, bcftools, BWA, and R installed, along with the necessary R packages. This workflow provides a robust framework for performing variant analysis and calculating key metrics, essential for genomic studies. The repository includes detailed instructions and example commands to facilitate easy implementation and reproducibility of the analysis.