Homework 3

Lorenzo Bozzi,
{lorenzo.bozzi}@aalto.fi
Course Name: High-Throughput Bioinformatics
Course Code: CS-E5875
Due date: 4/12/2023

***Abstract*—This experiment focuses on the analysis of methylation data obtained through bisulfite sequencing using both normal and beta-binomial models, comparing the results derived from these two approaches. The emphasis lies on the data related to chromosome 21, extracted from a set of 4 samples. Samples 1 and 2 represent specimens subjected to a specific treatment, while samples 3 and 4 serve as control samples. Especcially we want to comapre differencese between beta-binomial with full and the the reduced model without the test factor.**

I. Introduction

DNA methylation, a pivotal form of genetic regulation, operates without altering the DNA sequence itself. Predominantly found in CpG dinucleotides, it plays a crucial role within CpG islands of promoters, influencing gene expression and interacting with transcription factors.

Critical for development and DNA protection, methylation is influenced by environmental factors, potentially impacting health and disease predispositions.

While not directly heritable, methylation patterns exhibit similarities among family members, suggesting a combined influence of genetics and the environment.

DNA demethylation, once viewed as a passive process, now actively involves TET proteins in removing methyl groups from DNA.

The utilization of the beta-binomial model in DNA methylation analysis offers a sophisticated statistical perspective, allowing the assessment of significant differences between specific groups.

This analysis will deepen our understanding of epigenetic mechanisms, potentially revealing differences in methylation levels between treated and control samples, contributing to the comprehension of epigenetic modifications in pathological conditions.

II. Data Analysis

The experiments for this assignment were conducted on a local machine, specifically a MacBook Air M1 (2020) equipped with 8 GB of RAM and the Apple-designed M1 chip. We utilized Jupyter Notebook (version 6.5.2) as the programming environment and employed R as the primary programming language.

The dataset, stored in `rrbsData.RData`, can be loaded into R using the `load` function. Upon loading, explore the `BsRaw` object, containing slots: `assays`, `rowRanges`, and `colData`.

Table I: Summary of DNA Methylation Data - `BSraw` Class.

Attribute	Description
Class	BSraw
Dim	963 × 4
Assays	totalReads, methReads
Row names	NULL
RowData names	NULL
Col names	sample1, sample2, sample3, sample4
ColData names	group

Filtration was applied to remove CpG sites with extremely high coverages per sample to avoid PCR duplicates in RRBS, retaining about 88.68% of the original data.

The assessment of methylation fractions for each locus and sample revealed distinct patterns(see figure 1). Initially, a majority of loci showed remarkably high methylation levels, hinting at a significant portion of heavily methylated DNA.

Following this peak, a sharp decline in methylation values created a clear separation between two distinct loci populations. This separation suggests the existence of two distinct subgroups within the genome, each with unique epigenetic profiles.

Subsequently, a more gradual transition in methylation levels emerged, indicating progressive variations in methylation across genomic regions.

Notably, a sudden surge towards extremely high methylation values in the last bin highlighted a concentration of loci with distinct biological roles or particularly regulated epigenetically.

This complex distribution underscores the diversity and asymmetry in methylation dynamics within the genome.

The M-value histograms reveal a distinct peak in the first bin across all samples, suggesting a cluster of loci with markedly high M-values. Ther is a sort of oscillatory patterns signify variations in methylation levels among different loci within each sample.

An interesting observation is the absence or negligible representation of methylation in some loci within a sample, indicating distinct methylation patterns. These patterns are depicted in Figure 2.

We identified differentially methylated loci (DML) between treatment and control groups using the `limma` package. The moderated t-test in `limma` allowed for robust variance estimation by sharing information between genes. Constructing a design matrix utilizing `colData` ensured the inclusion of group information along with the intercept term.

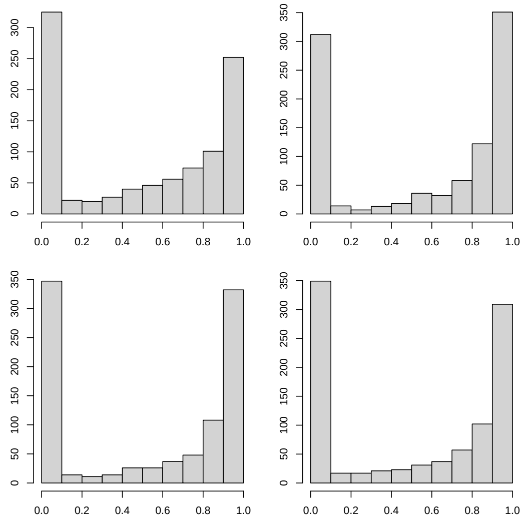

Figure 1: Visualisation of Methylation Fractions

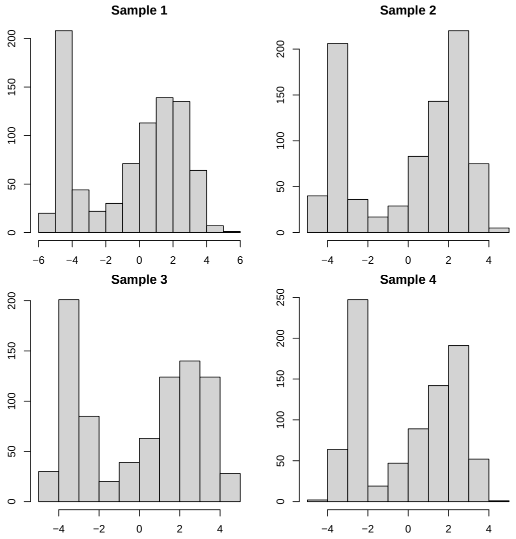

Figure 2: M-value histograms.

The obtained results represent the matrix of estimated coefficients for each locus or gene, highlighting the differences in methylation between the `treatment` and `control` groups. Each row of this matrix corresponds to a locus or gene, while the columns depict the differences in methylation among the four samples: two from the `treatment` group and two from the `control` group. To better illustrate the findings, Table II represents the estimated coefficients.

0.3184537	1.405057	1.7346011	1.8274776
0.7731899	1.127012	1.6094379	1.0881410
2.9704145	2.577688	4.4308168	3.0204249
0.4653632	1.876917	1.1526795	1.9169226
1.1370786	2.574519	0.4054651	0.9343092
1.7917595	3.601868	1.3350011	1.2039728

Table II: Estimated Coefficients Matrix

Furthermore, the results indicate that 807 loci exhibit a p-value below 0.05. This suggests a substantial difference in methylation between the two groups for these loci, signifying a significant impact of different treatments on the observed methylation differences.

These findings underscore the importance of the identified loci in the analysis, highlighting significant differences in methylation between the `treatment` and `control` groups in the dataset.

For further analysis using the beta-binomial model, we'll leverage the `aodml` function from the `aods3` package. Incorporating pseudocounts into both methylated and total read counts before model learning is crucial to mitigate numerical optimization issues, particularly in scenarios where $m_i = 0$ or $m_i = n_i$ across all samples.

The outcome furnishes a list of adjusted p-values for multiple testing (FDR). Notably, it has pinpointed 25 loci with a corrected FDR below 0.05, indicating a noteworthy discrepancy in methylation levels between the treatment and control groups at these specific loci.

III. Conclusion

In this section, we address points 6 and 7 of the assignment.

The analysis identified 25 loci exhibiting a statistically significant difference in methylation levels (FDR ¡ 0.05) between the treatment and control groups. These findings underscore the biological relevance of these specific genomic regions, suggesting substantial differences in methylation patterns associated with the studied conditions. The controlled false discovery rate correction indicates a high likelihood that these identified loci truly represent meaningful distinctions in methylation status between the experimental groups.

Comparing the results obtained through `limma` and `aodml` reveals a significant difference in the number of loci with differential methylation (DML). `limma` identified a total of 807 DML, while `aodml` identified only 25. Examining the specific differences, 794 DML were unique to `limma`, whereas `aodml` identified 12 DML exclusively. These disparities may stem from intrinsic analytical approaches of the two methods: `limma` could be more sensitive in detecting DML, while `aodml` might adopt a more conservative strategy.