DNA Methylation Analysis Manual

Supplementary Materials — Group 4

MSc in Bioinformatics — University of Bologna

June 2025

Contents

Preface

This document serves as an extensive technical guide for the DNA methylation analysis pipeline implemented by Group 4. It integrates theoretical foundations, detailed R code explanations, best practices, and troubleshooting tips to empower rigorous and reproducible Illumina HumanMethylation450K data analysis.

1 Glossary of Key Terms

Glossary (Part 1)

p-value The probability of obtaining a result as extreme as the observed one under the null hypothesis.

Detection p-value Metric indicating if the probe signal is distinguishable from background noise (values > 0.01 are often unreliable).

Beta-value Ratio of methylated to total signal, representing methylation level (range 0–1).

M-value Log_2 ratio of methylated to unmethylated signal; preferred for statistical testing due to improved variance properties.

CpG site DNA region where a cytosine nucleotide is followed by a guanine; key location for DNA methylation.

Heatmap A color-coded matrix visualization representing probe methylation across samples, often used for clustering analysis.

Workspace The R environment that stores all loaded objects during a session.

Object Any data structure stored in R, such as vectors, data frames, or lists.

Function A block of R code that performs a specific task.

Variable A name assigned to an object in R, enabling later reference.

Script A text file (.R) containing a sequence of R commands.

Console The interactive R prompt in RStudio or R.

Data frame A two-dimensional table-like structure with rows and columns.

Vector A one-dimensional sequence of data elements of the same type.

List A flexible container that can store multiple types of R objects.

False Positives Results that appear significant by chance but do not reflect true biological effects.

Benjamini-Hochberg (BH) A method for controlling the false discovery rate (FDR) when performing multiple statistical tests.

Bonferroni A conservative correction method for multiple testing that controls the family-wise error rate (FWER) by adjusting p-values.

PC number Principal Component (PC) number; indicates the order of variance explained by each component in PCA, where PC1 captures the most variance.

Variance A statistical measure of data spread; in PCA, it represents how much variation each principal component explains.

Linkage In hierarchical clustering, linkage defines how distances between clusters are calculated; methods include complete, average, and single linkage.

Glossary (Part 2)

Fold Change The difference in mean Beta-values between two groups (e.g. disease vs. control), indicating the magnitude of methylation change.

QC (Quality Control) A set of procedures to ensure the reliability and accuracy of experimental data before downstream analysis.

Loop A control structure that repeats a block of code multiple times.

Operator A symbol (e.g., `+`, `-`, `>`, `<`, `==`) used to perform operations.

False Discovery Rate (FDR) Correction method that controls the expected proportion of false positives among significant results.

Dye Bias Technical artifact introduced by different dye efficiencies; corrected via normalization.

Manifest file Metadata file that maps probe IDs to genomic coordinates and annotations.

RGChannelSet minfi object storing raw red and green channel intensities.

MethylSet minfi object after background correction, ready for downstream analysis.

Normalization Process that adjusts data to remove technical variation while preserving biological differences.

Batch effect Unwanted variation introduced by experimental conditions (e.g., processing date, operator).

Pipeline An ordered sequence of analysis steps.

Preprocessing Steps like background correction and normalization applied to raw data to prepare it for analysis.

Differential methylation Statistical comparison of methylation levels between groups (e.g., disease vs. control).

t-test A statistical test comparing the means of two groups to identify significant differences.

Volcano plot A scatter plot showing significance ($-\log_{10}$ p-value) versus effect size (e.g., delta beta).

Manhattan plot Genome-wide plot showing $-\log_{10}$ p-values by chromosome position.

PCA (Principal Component Analysis) A technique to explore sample variance and detect batch effects.

Pipe operator (`%>%`) An operator from magrittr/dplyr to chain commands for cleaner syntax (not used in this workflow).

2 Theoretical Background

DNA Methylation

DNA methylation is an epigenetic modification that occurs when a methyl group is added to the 5' position of cytosine residues within CpG dinucleotides. This modification is crucial for regulating gene expression and maintaining genome stability. Aberrant methylation is associated with diseases such as cancer and neurological disorders.

Illumina 450K BeadChip

The Illumina HumanMethylation450K BeadChip enables high-throughput analysis of over 485,000 CpG sites across the genome. It uses bisulfite conversion to differentiate between methylated and unmethylated cytosines:

- **Infinium I**: Utilizes two separate probes per CpG site—one specific for the methylated sequence and one for the unmethylated sequence. Each probe uses single-color detection in separate bead types.
- **Infinium II**: Utilizes a single probe per CpG site that hybridizes regardless of methylation status. The methylation state is distinguished using two-color detection in the same bead type.

Bisulfite Conversion

Bisulfite treatment converts unmethylated cytosines into uracils (read as thymines after PCR), while methylated cytosines remain unchanged. This differential conversion allows the array probes to detect methylation states.

Characteristic	**Infinium I**	**Infinium II**
Number of Probes per CpG Site	Two separate probes (one for methylated, one for unmethylated)	One single probe detects both methylated and unmethylated sequences
Probe Design	Two different bead types for methylated and unmethylated targets	One bead type per CpG site hybridizes regardless of methylation status
Single Base Extension (SBE)	Adds labeled nucleotide at the end of each probe; signals are detected by the same color channel (G = methylated, A = unmethylated)	Adds labeled nucleotide at the end of the single probe; signals are detected by two different color channels (G = methylated, A = unmethylated)
Detection Channel	Single color channel detects both methylated and unmethylated	Two separate color channels for methylated and unmethylated probes
CpG Coverage	Slightly lower overall coverage due to array space used by dual probes	Slightly higher overall coverage due to single-probe design
Signal Intensity and Dynamic Range	Higher dynamic range; less compression of Beta values	Lower dynamic range; more compression of Beta values
Dye Bias Susceptibility	Less dye bias, easier normalization	Higher dye bias, requires adjustment
Data Interpretation	Raw signals easier to interpret directly	Requires correction for dye bias and sequence context

Table 1: Comparison of Infinium I and Infinium II probe chemistries used in Illumina DNA methylation BeadChip arrays.

Figure 1: Overview of the Illumina Infinium HumanMethylation450 BeadChip platform. This array enables high-throughput analysis of over 485,000 methylation sites per sample at single-nucleotide resolution. It covers 99% of RefSeq genes, targeting an average of 17 CpG sites per gene region (including promoter, 5'UTR, first exon, gene body, and 3'UTR). The platform also covers 96% of CpG islands, as well as island shores and flanking regions. Additional features include coverage of non-CpG methylated sites (identified in human stem cells), differentially methylated sites between tumor and normal samples (multiple cancer types), CpG islands outside of coding regions, and miRNA promoter regions. These features make the platform ideal for epigenome-wide association studies (EWAS).

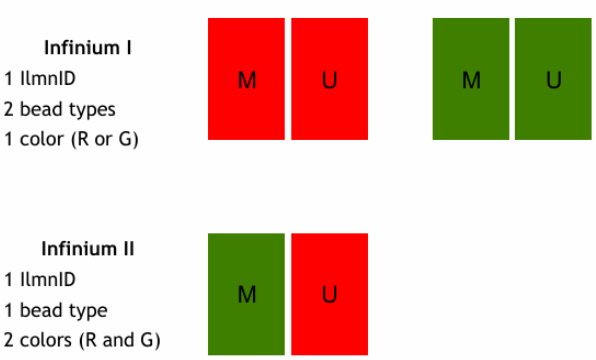

Figure 2: Comparison of Infinium I and Infinium II probe designs.

3 R — A Deep Dive

What is R?

R is a free, open-source programming language and environment specifically designed for statistical computing and data visualization. It is widely used in bioinformatics because of its:

- Flexibility in handling large and complex datasets typical of omics studies.
- Extensive ecosystem of packages dedicated to genomics and epigenetics (e.g. `minfi`, `limma`, `edgeR`).
- Integration with Bioconductor, providing standardized workflows for methylation analysis.
- Seamless interface with RStudio, an integrated development environment (IDE) that simplifies script writing, data visualization, and debugging.

R is particularly valued for its ability to chain together data preprocessing, statistical testing, visualization, and result export, making it a complete solution for DNA methylation analysis.

The RStudio Interface

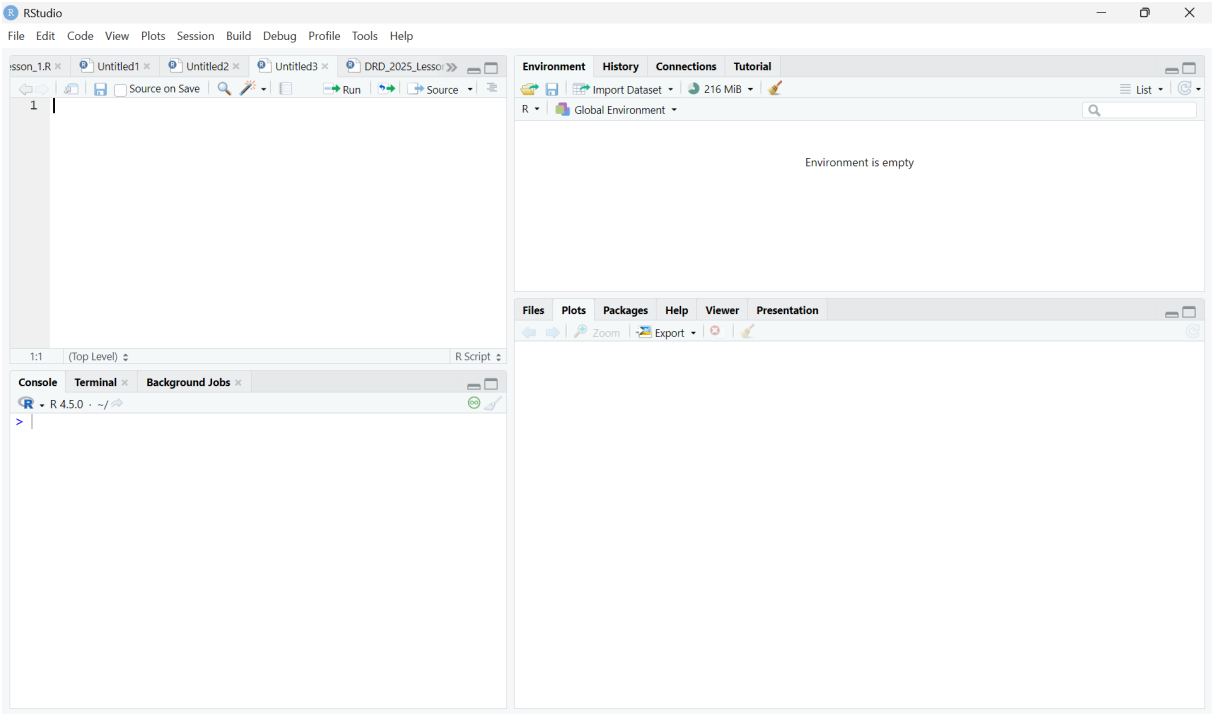

Figure 3: RStudio Interface — Console, Script Editor, Workspace, Plots.

Tips and Useful Commands in R

To efficiently work with R, it is helpful to know some basic tips and commands:

- Set the working directory with `setwd("your/path")` and check it with `getwd()`.
- Use the assignment operator `<-` to create objects: `mydata <- read.csv("data.csv")`.
- Clean the workspace to avoid object conflicts: `rm(list = ls())`.
- Use the `#` symbol to comment and annotate code.
- Check installed packages with `installed.packages()` and load them with `library(pkgname)`.
- Explore data with commands like `str()`, `summary()`, `head()`, and `tail()`.
- Use pipes (`%>%`) from the `magrittr` or `dplyr` package to chain commands and improve code readability.
- For quick plotting, try `plot()`, while for more advanced graphics use `ggplot2`.

Packages and Libraries

Packages extend R's functionality. Use:

Installing and Loading Packages

```
install.packages("ggplot2")
library(ggplot2)
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install("minfi")
```

Data Structures in R

In R, data structures are fundamental for organizing and analyzing data. Here are the most common structures you will encounter in methylation analysis:

- **Vector**: A one-dimensional sequence of elements of the same type (numeric, character, logical).
- **Factor**: A special type of vector for categorical data with predefined levels.
- **Matrix**: A two-dimensional array of elements of the same type.
- **Data Frame**: A table-like structure with rows and columns, where each column can have a different type.
- **List**: A flexible container that can hold objects of different types and structures (e.g. data frames, vectors, even other lists).

Example:

```
# Vector
my_vector <- c(1, 2, 3, 4)

# Factor
my_factor <- factor(c("Control", "Treatment", "Control"))

# Matrix
my_matrix <- matrix(c(1, 2, 3, 4, 5, 6), nrow = 2, byrow = TRUE)

# Data Frame
my_df <- data.frame(SampleID = c("A1", "A2"), Value = c(0.5, 0.7)
   )

# List
my_list <- list(Numeric = my_vector, Category = my_factor, Matrix
    = my_matrix, Table = my_df)
```

Understanding these data structures is essential for efficient data manipulation and analysis in R.

4 About `minfi`

The `minfi` package is a comprehensive and essential R/Bioconductor toolkit for analyzing Illumina HumanMethylation arrays (450K and EPIC). It enables efficient processing and robust analysis of DNA methylation data by providing:

- Import of raw IDAT files using `read.metharray.exp()`, supporting flexible sample sheet integration.
- Quality control metrics, such as probe detection p-values using `detectionP()` and negative control plots, to ensure data reliability.
- Preprocessing methods, including background correction and functional normalization (`preprocessFunnorm()`), to correct technical biases and batch effects.
- Extraction of Beta-values and M-values using `getBeta()` and `getM()`, facilitating downstream statistical testing.

Overall, `minfi` supports a complete methylation analysis workflow, from raw data import to preprocessing and quality assessment, making it an essential package for reproducible and high-quality methylation studies.

Note. Additional R packages were also used to support data visualization, statistical testing, and probe annotation, including:

- `gplots` for generating heatmaps and performing hierarchical clustering.
- `qqman` for generating Manhattan plots to visualize genome-wide significance.
- `factoextra` and `factoMineR` for computing and visualizing Principal Component Analysis (PCA); `ggplot2`, `ggpubr`, and `cluster` for enhanced graphical representation and clustering overlays.

- `genefilter` for efficient high-throughput statistical testing using `rowttests()`.
- `IlluminaHumanMethylation450kanno.ilmn12.hg19` — a Bioconductor annotation package used to map probes to genomic coordinates, gene regions, and CpG features (required for biological interpretation within the `minfi` workflow).

All Bioconductor packages were installed via `BiocManager` to ensure compatibility and consistency across releases.

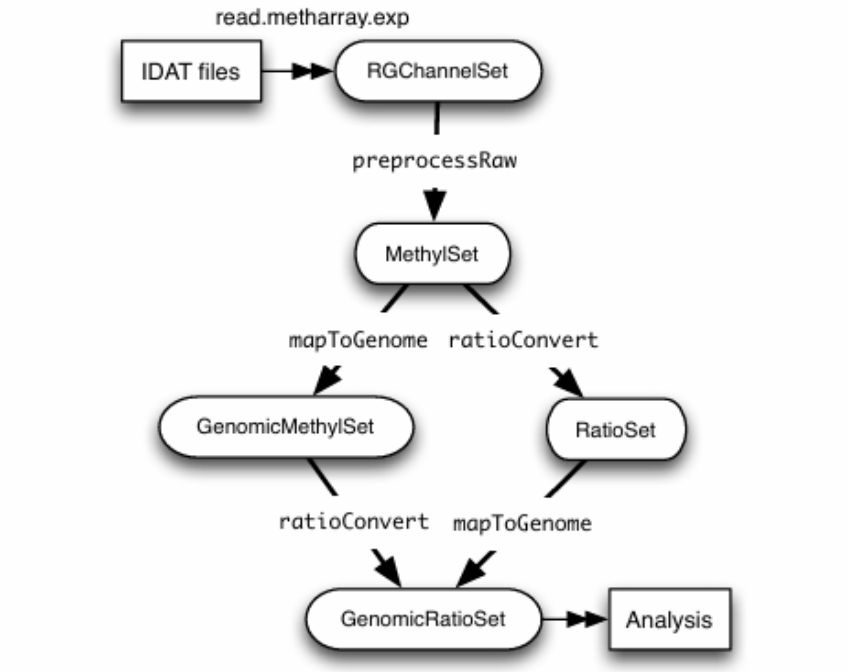

Figure 2: Flow chart of the `minfi`'s functions

Figure 4: Flow chart of the `minfi` pipeline: data objects evolve from raw IDAT files (`RGChannelSet`) to genome-annotated methylation matrices (`GenomicRatioSet`). Each transformation adds structure and annotation, enabling quality control, normalization, and biological interpretation.

Each object plays a defined role within the `minfi` pipeline:

- `RGChannelSet`: Raw intensity data (red and green channels) from IDAT files, unfiltered and unprocessed.
- `MethylSet`: Background-corrected methylated and unmethylated signals, still unmapped to the genome.
- `RatioSet`: Includes Beta-values (methylated / total) and/or M-values (log2 ratio), useful for differential methylation analysis.
- `GenomicMethylSet`: Like `MethylSet`, but with genome mapping information.
- `GenomicRatioSet`: Final analysis-ready object: Beta or M-values mapped to genome coordinates, ideal for visualization and statistical testing.

Function name	**Input**	**Output**	**Purpose**
`preprocessNoob`	RGset	Mset	Corrects background and dye bias (single-sample method).
`preprocessSWAN`	RGset	Mset	Adjusts for probe type bias between Infinium I and II probes.
`preprocessQuantile`	RGset or Mset	GenomicRatioSet	Equalizes distributions across samples, often used in general omics.
`preprocessFunnorm`	RGset	Mset	Uses internal control probes to remove unwanted variation while preserving biological signal. *(Used in our analysis)*

Table 2: Comparison of preprocessing functions available in the `minfi` package.

5 Project Data

For this project, we analyzed DNA methylation data from the Illumina HumanMethylation450K array. Below is a summary of the key files and data objects used in the analysis:

File/Data Object	Description
`SampleSheet.csv`	Sample metadata including ID, group (CTRL/DIS), sex, and batch.
`.idat files`	Raw fluorescence intensity data (red and green channels).
`Illumina450Manifest_clean.RData`	Cleaned manifest with probe IDs, genomic coordinates, and design type.
`RGChannelSet`	Raw data object created with `read.metharray.exp()`.
`MSet.raw`	Object created after background correction using `preprocessRaw()`, used for quality control prior to normalization.
`MSet.norm`	Normalized data obtained using functional normalization (`preprocessFunnorm()`).
`Beta/M values`	Matrices of methylation quantification (`getBeta()`, `getM()`).
`Annotation`	Genomic mapping of probes using the `IlluminaHumanMethylation450kanno.ilmn12.hg19` annotation package.
`DetectionP`	Detection p-values per probe/sample (`detectionP()`).
`final_ttest`	Data frame with raw p-values from group comparison (computed using `rowttests()`).
`final_ttest_corr`	Data frame with corrected p-values (BH and Bonferroni).
`diff_meth_df`	Summary table of differentially methylated probes.
`pca_results`	PCA results object for sample clustering.
`Plots`	Figures including QC plots, density and boxplots, PCA, scree plot, volcano and Manhattan plots, and hierarchical clustering heatmaps.

Table 3: Key files and data objects used in the Group 4 DNA methylation analysis.

Additionally, the following **assigned parameters** were used for the analysis:

Parameter	Value
Group ID	4
Probe Address	44666390
Detection p-value cut-off	0.01
Normalization method	`preprocessFunnorm()`
Clustering methods	Average, Complete, Single linkage

Table 4: Assigned parameters used in the analysis.

6 Analytical Pipeline Overview

A successful DNA methylation study requires a well-designed workflow that systematically addresses each analytical step, from biological hypothesis to data visualization. Below is an expanded summary of the pipeline we followed in this project.

1. **Biological Question:**
   - Define the scientific goal, e.g., "Does methylation differ significantly between disease and control groups?"
   - Establish hypotheses and experimental design (number of samples, replicates, conditions).
2. **Sample Collection and Preparation:**
   - Extract DNA from relevant tissues or cell types.
   - Perform bisulfite conversion to differentiate methylated from unmethylated cytosines.
3. **Data Acquisition:**
   - Use the Illumina HumanMethylation450K array (or EPIC) to profile methylation.
   - Generate raw data files (`.idat`) containing red and green channel intensities.
4. **Data Import into R:**
   - Use `read.metharray.exp()` to read `.idat` files into an `RGChannelSet` (`RGSet`).
   - Inspect sample metadata from `SampleSheet.csv`.
5. **Annotation:**
   - Link probe IDs to genomic features using the cleaned manifest file `Illumina450Manifest_clean.RData` and the annotation package `IlluminaHumanMethylation450kanno.ilmn12.hg19`.
6. **Quality Control and Preprocessing:**
   - Calculate detection p-values with `detectionP()` to filter unreliable probes.

- Assess sample quality with `plotQC()` and `controlStripPlot()`.
- Apply background correction using `preprocessRaw()` to obtain `MSet.raw`, used for initial QC assessment.
- Normalize data with `preprocessFunnorm()` to obtain `MSet.norm`, minimizing technical variation while preserving biological signal.

7. **Methylation Quantification:**
   - Calculate Beta-values using `getBeta()` and M-values using `getM()`.
8. **Statistical Testing:**
   - Perform group comparisons with `t.test()`.
   - Use `rowttests()` from the `genefilter` package to efficiently perform a vectorized t-test across all probes.
9. **Multiple Testing Correction:**
   - Adjust p-values using `p.adjust()` (methods: BH, Bonferroni).
10. **Visualization:**
    - Generate density plots and boxplots (`ggplot2`), PCA plots (`prcomp()`, `fviz_eig()` from `factoextra`), volcano and Manhattan plots (`ggplot2`, `qqman`), and hierarchical heatmaps (`heatmap.2()` from `gplots`).
11. **Biological Interpretation:**
    - Map significant CpGs to genes and pathways.
    - Integrate methylation results with external datasets (e.g., gene expression, clinical data).

Figure 5: Pipeline Overview

6.1 Function Reference

Data Import

```
read.metharray.exp() | Imports raw .idat files into an RGChannelSet object
(red and green channel intensities).
getRed() | Extracts red channel intensity matrix from an RGChannelSet.
getGreen() | Extracts green channel intensity matrix from an RGChannelSet.
```

Annotation

```
Illumina450Manifest_clean | Cleaned manifest file mapping probe IDs to
genomic coordinates.
IlluminaHumanMethylation450kanno.ilmn12.hg19 | Bioconductor annotation
package for gene and region mapping of probes.
```

Quality Control

```
detectionP() | Computes detection p-values per probe/sample.
plotQC() | Visualizes signal intensities to assess sample quality.
controlStripPlot() | Displays negative control probe intensities for quality
check.
```

Preprocessing

```
preprocessRaw() | Performs initial background correction on raw signals.
preprocessFunnorm() | Applies functional normalization using internal
control probes.
```

Beta/M-value Calculation

```
getBeta() | Extracts Beta-values (methylated / total intensity).
getM() | Extracts M-values (log2 ratio of methylated to unmethylated signal).
```

Statistical Testing

```
t.test() | Standard t-test for comparing group means (used manually per
probe).
rowttests() | Vectorized high-throughput t-test across all probes (from
genefilter).
```

Multiple Testing Correction

```
p.adjust() | Adjusts p-values using FDR or FWER methods (e.g., BH,
Bonferroni).
```

Visualization

```
heatmap.2() | Generates heatmaps with hierarchical clustering (gplots
package).
fviz_eig() | Plots variance explained by principal components (from
factoextra).
manhattan() | Creates Manhattan plots for genome-wide significance (qqman
package).
```

7 Interpretation of Outputs and Parameter Choices

This section provides a high-level technical discussion that complements the analytical pipeline, focusing on the rationale behind key parameter choices and their impact on data quality, statistical validity, and interpretability. By integrating theoretical principles with observed results, it bridges methodological steps and biological insight.

7.1 Parameter Choices and Data Quality

Functional Normalization (`preprocessFunnorm()`) was selected as the normalization strategy to correct for dye bias, background noise, and moderate batch effects. This method uses internal control probes to capture and regress out technical variation, ensuring that true biological differences are preserved.

Detection p-value Cutoff (0.01) was used as a stringent threshold to exclude unreliable probes whose signal may not significantly exceed background noise. Although conservative, this filtering step ensures high confidence in the retained data and minimizes the inclusion of false positives.

7.2 Quality Control and Normalization

QC metrics were assessed through visual and statistical tools. The negative control strip plot showed low background intensities across all samples, indicating an absence of systemic artifacts. The QC plot of median methylated versus unmethylated signal intensities confirmed that all samples passed quality thresholds. Density and boxplots of Beta-values—stratified by Infinium probe type and experimental group (CTRL vs. DIS)—exhibited substantial improvement after normalization, with better alignment and reduced inter-sample variability. These observations confirm that functional normalization effectively removed technical variation and preserved the underlying biological signal.

7.3 Methylation Distributions

As expected, the Beta-value distribution was bimodal with peaks near 0 (unmethylated) and 1 (fully methylated), while M-values showed a symmetric distribution centered around zero. Although overall shapes were consistent across groups, subtle shifts between CTRL and DIS distributions suggested group-specific methylation changes, warranting formal statistical testing.

7.4 Principal Component Analysis (PCA)

The first two principal components (PC1 and PC2) captured a substantial portion of the total variance, but did not clearly separate the CTRL and DIS groups. Although CTRL samples tended to cluster in the negative PC1 space, this region also contained half of the DIS samples, indicating poor group discrimination.

In the PCA plot by sex, no major separation was observed along PC1. However, PC2 showed a strong sex-associated pattern: female samples clustered in the lower half of the graph, while male samples occupied the upper half, with only one exception.

A pronounced batch effect was detected. Samples from a single batch clustered closely together, while those from other batches appeared as distant outliers along PC1 and PC2. This suggests that the Sentrix ID variable contributes significantly to variance and that normalization did not fully correct for batch effects.

7.5 Differential Methylation and Multiple Testing

The volcano plot revealed a dense cluster of non-significant CpGs centered around zero, with no CpG sites passing the adjusted significance threshold (BH-adjusted $p < 0.01$). While several sites showed moderate methylation differences ($\Delta\beta > 0.1$), none exhibited strong statistical support after correction. The p-value distribution was approximately uniform, with only a slight enrichment near zero—suggesting weak global signal and a low probability of widespread differential methylation between CTRL and DIS groups. These results highlight the importance of multiple testing correction, which effectively filtered out signals not robust enough to reach significance.

P-Value Boxplots (Raw, BH, Bonferroni)

Boxplots comparing raw, BH-adjusted, and Bonferroni-adjusted p-values illustrate the drastic impact of multiple testing correction. While over 12,000 CpG sites appeared significant based on nominal p-values ($p \leq 0.05$), none remained below the threshold after applying either Benjamini-Hochberg or Bonferroni correction. This underscores the importance of correction methods in high-dimensional data to reduce false discoveries. The distribution of adjusted p-values shifted markedly toward 1, reflecting their conservative nature in controlling for type I errors across thousands of tests.
Note: A p-value threshold of 0.05 was used for initial hypothesis testing with `t.test()` and `rowttests()`, whereas a stricter threshold of 0.01 was applied to detection p-values during quality control. The former supports biological interpretation, while the latter ensures technical data reliability.

7.6 Genome-Wide Overview

The Manhattan plot provided a genome-wide visualization of methylation differences, highlighting a small number of CpG sites that surpassed the significance threshold after multiple testing correction. These top hits, spread across different chromosomes, may represent biologically relevant differentially methylated positions (DMPs). Although no large clusters were observed, the plot underscores the genomic distribution of statistically significant signals and supports their prioritization for downstream analysis and annotation.

7.7 Clustering and Heatmaps

Hierarchical clustering using complete, average, and single linkage methods revealed a consistent global separation between CTRL and DIS samples, supporting the presence of methylation-based group differences. Among the methods, complete and average linkage produced compact and biologically interpretable clusters. In contrast, single linkage resulted in elongated branches with reduced resolution, likely due to chaining effects. Overall, average linkage provided a balanced trade-off between sensitivity and interpretability, and was therefore preferred for downstream analysis. The heatmaps also showed clear methylation gradients (green to red), reflecting levels of hypo- and hypermethylation across samples and probes.

7.8 Limitations and Potential Improvements

While the selected preprocessing and analysis strategy yielded robust results, potential improvements include:

- Evaluating less stringent detection p-value cutoffs (e.g., 0.05) to recover borderline CpGs of potential interest.
- Comparing alternative normalization methods (e.g., SWAN, quantile) to assess robustness.
- Considering ComBat or other batch correction methods to further minimize residual batch effects.

- Performing sensitivity analyses (e.g., resampling) to test the stability of differentially methylated signatures.

These refinements may enhance generalizability and biological relevance in future studies.

8 Tips and Troubleshooting

- **Object Not Found**: Check spelling and case sensitivity. Use `ls()` to list defined objects.
- **Package Not Loaded**: Load required packages with `library()`.
  Use `BiocManager::install()` for Bioconductor packages.
- **Dimension Mismatch**: Use `dim()`, `str()`, and `head()` to inspect object structure.
- **Warning Messages**: Use `warnings()` to review non-fatal alerts and investigate potential causes.
- **Missing Values**: Detect missing entries with `is.na()` and decide whether to impute or exclude.
- **Normalization Errors**: Ensure both `preprocessRaw()` and `preprocessFunnorm()` have been correctly applied.
- **Dye Bias**: Especially relevant in Infinium II probes; always normalize using control-based methods.
- **Version Conflicts**: Check package versions using `packageVersion("pkgname")` to avoid deprecated features.
- **Plotting Issues**: Verify input data format and completeness before using `ggplot2` or `gplots`.
- **Unexpected Results**: Confirm correct metadata assignment (e.g., group, batch) using `table()` and `colData()`.

9 Conclusion

This manual provides a comprehensive and reproducible framework for analyzing DNA methylation data using R and Bioconductor. Through the use of rigorous preprocessing (e.g., `preprocessFunnorm()`), statistical testing, and visualization techniques, it was possible to identify differentially methylated regions (DMRs) with potential biological relevance.

The results support the application of this pipeline in epigenome-wide association studies (EWAS), biomarker discovery, and disease mechanism exploration. In particular, the use of robust tools like `minfi`, combined with stringent quality control and normalization, enhances the interpretability and reproducibility of results.

This work exemplifies the integration of bioinformatics and molecular biology, and we hope it serves as a valuable resource for future research in epigenetics, disease genomics, and precision medicine.

10 Acknowledgements and Contact Information

10.1 Acknowledgements

We would like to thank **Professor Francesco Ravaioli** and the teaching staff of the course of **DNA-RNA Dynamics** at the University of Bologna. for their support and guidance throughout this project.

10.2 Contact and Support

This document and the full pipeline are maintained by **Group 4**. For any questions, bug reports, or suggestions, feel free to contact us:

- **Martina Castellucci** — martina.castellucci@studenti.unibo.it
- **Alessia Corica** alessia.corica@studenti.unibo.it
- **Sofia Natale** sofia.natale@studenti.unibo.it
- **Andrea Pusiol** andrea.pusiol@studenti.unibo.it+
- **Perla Lucaboni** perla.lucaboni@studenti.unibo.it
- **Aurora Mazzoni** aurora.mazzoni2@studenti.unibo.it
- **Bianca Mastroddi** bianca.mastroddi@studenti.unibo.it

Please open an issue on GitHub for reproducibility concerns or technical questions.