REPORT

NAME – Pratham Dundannavar

PRN – 20220901083

B tech – Biomedical Engineering

Internship - Extion(Bioinformatics)

Supervisor – Shubangi Patil

**DNA Methylation Biomarker Discovery OF Colon cancer**

DNA methylation is the process which stops transcription due to addition of methyl groups to the DNA chains. The addition can also act as biomarkers which further provide a brief idea on diseases especially cancers. Specific DNA methylation patterns can help to provide the onset or even the type of cancer along with its severity.

1. **Downloading the dataset from GEO (**[**https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE53051**](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE53051)**) using R Bioconductor package.**
2. Install the Bioconductor package and GEOquery, minfi and ChAMP by running the following codes in the R environment:

install.packages(“BiocManager”)

BiocManager::install(“GEOquery”)

BiocManager::install(“minfi”)

BiocManager::install(“ChAMP”)

1. Download the Series Matrix File from the website: <https://ftp.ncbi.nlm.nih.gov/geo/series/GSE53nnn/GSE53051/matrix/>

This file contains beta values in a wide format meaning the “ID\_REF” (CpG probe IDs) and the rest columns are our SampleIDs having values between 0 to 1 indicating beta values.

1. **Loading the data into R**

Following codes help to load the data into R:

#Loading Methylation Data

matrix\_file = "C:/Users/Admin/Desktop/GSE53051\_series\_matrix.txt"

library(data.table) # Load the data.table package

library(GEOquery) # Load GEOquery for phenotype data

skip\_lines <- 74 #Skips lines starting with "!"

cat(paste0("\nLoading ", matrix\_file))

#TRY CATCH

beta\_values\_dt = tryCatch({

fread(

input = matrix\_file,

skip = skip\_lines,

header = TRUE,

data.table = FALSE, # Output as a standard data.frame

check.names = FALSE,

verbose = TRUE, # Keep verbose for detailed messages

nThread = 1

)

},

error = function(e) {

message(paste0("Error during fread"))

NULL

})

# Check if beta\_values\_dt was successfully created

if (is.null(beta\_values\_dt))

{

stop("fread failed to load the data.")

} else {

cat("\nSuccessfully loaded data using fread!\n")

}

# Renaming the first column to "ID\_REF" which are row names

if ("ID\_REF" %in% colnames(beta\_values\_dt))

{

rownames(beta\_values\_dt) <- beta\_values\_dt$ID\_REF

beta\_values\_dt$ID\_REF <- NULL # Remove the column as it's now row names

cat("Set 'ID\_REF' as row names and removed the column.\n")

}else{

first\_col\_name <- colnames(beta\_values\_dt)[1]

if (!is.null(first\_col\_name) && startsWith(first\_col\_name, "cg")) #CpG IDs start with "cg"

{

cat(paste0("First column is '", first\_col\_name, "'. Setting as row names.\n"))

rownames(beta\_values\_dt) <- beta\_values\_dt[[first\_col\_name]]

beta\_values\_dt[[first\_col\_name]] <- NULL

} else {

warning("Column 'ID\_REF' not found.")

}

}

#Analyzing the Data Frame

cat("\nDimensions of the beta value matrix:\n")

print(dim(beta\_values\_dt))

cat("\nFirst 5 rows and first 5 columns of the beta value matrix:\n")

print(head(beta\_values\_dt[, 1:min(5, ncol(beta\_values\_dt))]))

# Identify a sample ID from the header for summary

sample\_id\_for\_summary <- colnames(beta\_values\_dt)[1] # Take the first sample column

if (!is.null(sample\_id\_for\_summary))

{

cat(paste0("\nSummary of beta values for the first sample (", sample\_id\_for\_summary, "):\n"))

print(summary(beta\_values\_dt[, sample\_id\_for\_summary]))

} else {

warning("Could not identify a sample column for summary.")

}

#Extracting Sample Information

cat("\nExtracting phenotype data using GEOquery\n")

gse <- getGEO("GSE53051", GSEMatrix = TRUE, getGPL = FALSE)

if (length(gse) > 1) {

gse <- gse[[1]] #first gms

}

pheno\_data <- pData(phenoData(gse))

cat("\nDimensions of the phenotype data:\n")

print(dim(pheno\_data))

cat("\nFirst few rows and selected columns of phenotype data:\n")

print(head(pheno\_data[, c("title", "source\_name\_ch1", "characteristics\_ch1.1", "characteristics\_ch1")]))

print(colnames(pheno\_data))

Creating a phenotypic Dataframe:

gse\_local <- getGEO(filename = "C:/Users/Admin/Desktop/GSE53051\_series\_matrix.txt", GSEMatrix = TRUE, getGPL = FALSE)

if (class(gse\_local) == "list") {

gse\_local <- gse\_local[[1]] # Select the first (and likely primary) GSMList if it's a list

}

# The phenotype data is stored in the pData slot of the ExpressionSet object

pheno\_data <- pData(phenoData(gse\_local))

cat("\nDimensions of the phenotype data:\n")

print(dim(pheno\_data))

cat("\nFirst few rows and selected columns of phenotype data:\n")

print(head(pheno\_data[, c("title", "source\_name\_ch1", "characteristics\_ch1.1", "characteristics\_ch1", "characteristics\_ch1.2")]))

cat("\nAvailable columns in pheno\_data:\n")

print(colnames(pheno\_data))

#Aligning Phenotypic Data columns with original dataset

sample\_names\_beta <- colnames(beta\_values\_dt)

sample\_names\_pheno <- rownames(pheno\_data)

common\_samples <- intersect(sample\_names\_beta, sample\_names\_pheno)

if (length(common\_samples) == 0) {

stop("No common samples found between beta values and phenotype data.")

}

if (length(common\_samples) < length(sample\_names\_beta)) {

warning(paste0("Number of common samples (", length(common\_samples), ") is less than total beta samples (", length(sample\_names\_beta), ")."))

}

# Reordering original dataset columns and pheno\_data rows to match

beta\_values\_dt <- beta\_values\_dt[, common\_samples]

pheno\_data <- pheno\_data[common\_samples, ]

print("\nSuccessfully aligned ")

\*Preparing pheno\_data for differential analysis

In this the phenotypic dataframe is prepared, cleaned and centred around two binary values which give the overall state of the disease which are “Normal” and “Cancer”. The “Adenoma”, “Metastasis” have not been considered because the overall sample size of the differential analysis reduces which would give inaccurate results.

# 1. Standardizing Disease State

pheno\_data$Condition <- tolower(pheno\_data$`characteristics\_ch1.1`)

pheno\_data$Condition[grepl("normal", pheno\_data$Condition)] <- "normal"

pheno\_data$Condition[grepl("cancer|tumor|adenocarcinoma", pheno\_data$Condition)] <- "cancer" # Added common cancer terms

pheno\_data$Condition <- factor(pheno\_data$Condition, levels = c("normal", "cancer")) # Only normal/cancer as main levels

# 2. Standardizing Source Tissue

pheno\_data$Tissue <- tolower(gsub(" tissue|\\s+sample|\\s+tissue", "", pheno\_data$source\_name\_ch1, ignore.case = TRUE))

pheno\_data$Tissue <- factor(pheno\_data$Tissue)

#Filtering and Realigning

filtered\_sample\_ids <- rownames(pheno\_data)[pheno\_data$Condition %in% c("normal", "cancer") &

!is.na(pheno\_data$Condition) & !is.na(pheno\_data$Tissue)]

pheno\_data\_filtered <- pheno\_data[filtered\_sample\_ids, ]

beta\_values\_dt\_filtered <- beta\_values\_dt[, filtered\_sample\_ids]

#Filtered dataframe

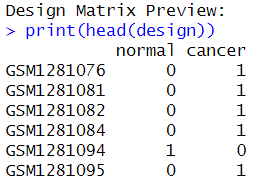
cat("\nSample Distribution (Normal vs. Cancer) by Tissue:\n")

print(table(pheno\_data\_filtered$Condition, pheno\_data\_filtered$Tissue))

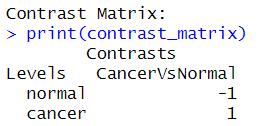
1. **Differential Analysis**

For this we will be using a Bioconductor package entitled “limma”, this package needs to know which binary group is reference (keep in mind that we have “Normal” and “Cancer” as our binary groups) so, we will set “Normal” group as our reference. This is done in order to let the ‘logFC’ (log fold change) represent the methylation level in the ‘cancer’ group (which is we need) in the form of (Methylation[Cancer] – Methylation[Normal]). A positive logFC means hypermethylation and a negative logFC means hypomethylation in ‘cancer’ group.

A design matrix is important to let limma understand the relationships between our samples. So, in short design matrix creates sampled groups of our colon cancer. Meaning our SampleIDs would have the number of samples which are cancerous or normal in a tabular format due to the design matrix.



The contrast matrix is defined to make specific comparisons between these groups meaning it highlights the difference in methylation between normal and cancer groups.



Note: Differential Analysis initially has been done on Colon cancer as it had largest samples.

Number of significant DMPs (adj. P < 0.05): 152910

Number of significant DMPs with |delta beta| > 0.1: 93354

1. Install Packages:

BiocManager::install(“limma”,force=TRUE)

BiocManager::install(“IlluminaHumanMethylation450kanno.ilmn12.hg19",force=TRUE)

library(limma)

library(minfi) # For getting probe annotations

library(IlluminaHumanMethylation450kanno.ilmn12.hg19) # Annotation data

cat("\nDifferential Methylation Analysis\n")

#Subset data for Colon tissue

chosen\_tissue = "colon"

colon\_pheno = pheno\_data\_filtered[pheno\_data\_filtered$Tissue == chosen\_tissue, ]

colon\_beta = beta\_values\_dt\_filtered[, rownames(colon\_pheno)]

cat(paste0("\nAnalyzing ", chosen\_tissue, " tissue:\n"))

print(table(colon\_pheno$Condition))

#Normal as reference

colon\_pheno$Condition = relevel(colon\_pheno$Condition, ref = "normal")

#Defining the Design Matrix

design = model.matrix(~0 + Condition, data = colon\_pheno)

colnames(design) = gsub("Condition", "", colnames(design))

cat("\nDesign Matrix Preview:\n")

print(head(design))

cat("\nDimension of Design Matrix:", dim(design), "\n")

#Contrast matrix for comparing Cancer with Normal

contrast\_matrix = makeContrasts(CancerVsNormal = cancer - normal, levels = design)

cat("\nContrast Matrix:\n")

print(contrast\_matrix)

#Limma

#Fitting the linear model

cat("\nFitting linear model\n")

fit = lmFit(colon\_beta, design)

#Applying contrasts

cat("Applying contrasts\n")

fit2 = contrasts.fit(fit, contrast\_matrix)

#Smoothening

cat("Performing empirical Bayes\n")

fit2 = eBayes(fit2)

print("\nLimma Completed\n")

# 5. Extract results and add annotations

topDMPs = topTable(fit2, coef = "CancerVsNormal", number = Inf, adjust.method = "BH") # Inf to get all results

# Adding annotations to the results

cat("\nAdding probe annotations...\n")

data(IlluminaHumanMethylation450kanno.ilmn12.hg19) #Gets annotation Data for Illumina 450k

anno = getAnnotation(IlluminaHumanMethylation450kanno.ilmn12.hg19)

#Merging

anno\_subset = anno[, c("chr", "pos", "UCSC\_RefGene\_Name", "Relation\_to\_Island", "Regulatory\_Feature\_Group")]

topDMPs\_annotated = merge(topDMPs, anno\_subset, by = "row.names", all.x = TRUE)

colnames(topDMPs\_annotated)[1] = "Probe\_ID" # Rename the first column back to Probe\_ID

cat("\nFirst few rows of Top Differentially Methylated Probes (DMPs):\n")

print(head(topDMPs\_annotated))

cat("\nSummary of results:\n")

#Adjusted p Values to find out significance

significant\_DMPs = subset(topDMPs\_annotated, adj.P.Val < 0.05)

cat(paste0("Number of significant DMPs (adj. P < 0.05): ", nrow(significant\_DMPs), "\n"))

# Biologically Significant |delta beta| > 0.1

significant\_and\_large\_effect\_DMPs = subset(significant\_DMPs, abs(logFC) > 0.1) # delta beta > 0.1 (10% difference)

cat(paste0("Number of significant DMPs with |delta beta| > 0.1: ", nrow(significant\_and\_large\_effect\_DMPs), "\n"))

#Beta Values for significant DMPs

cat("\nSummary of delta beta values for significant DMPs:\n")

print(summary(significant\_DMPs$logFC))

Save the files:

library(openxlsx)

output\_file1 = "C:/Users/Admin/Desktop/Colon Cancer(Binary) All probes.xlsx"

output\_file2 = "C:/Users/Admin/Desktop/Colon Cancer(Binary) significant beta.xlsx"

write.xlsx(topDMPs\_annotated, file = output\_file1, rownames=FALSE)

write.xlsx(significant\_and\_large\_effect\_DMPs, file= output\_file2, rownames=FALSE)

1. **Visualisation**
2. **Volcano Plots**

BiocManager::install(“EnhancedVolcano”)

install.packages(“dplyr”)

library(openxlsx)

library(dplyr) # For data manipulation like filter

file\_path = "C:/Users/Admin/Desktop/Colon Cancer(Binary) All probes.xlsx"

topDMPs\_annotated = read.xlsx(file\_path)

# Ensuring logFC and adj.P.Val are numeric, as read.xlsx might sometimes read them as character

topDMPs\_annotated$logFC = as.numeric(topDMPs\_annotated$logFC)

topDMPs\_annotated$adj.P.Val = as.numeric(topDMPs\_annotated$adj.P.Val)

# Re-creating significant\_and\_large\_effect\_DMPs for use in the plot caption

significant\_and\_large\_effect\_DMPs = topDMPs\_annotated %>%

filter(adj.P.Val < 0.05, abs(logFC) > 0.1)

cat("topDMPs\_annotated file loaded successfully. Dimensions: ")

print(dim(topDMPs\_annotated))

#Volcano Plot

cat("\nGenerating Volcano Plot...\n")

library(EnhancedVolcano)

library(ggplot2)

volcano\_plot = EnhancedVolcano(

toptable = topDMPs\_annotated,

lab = topDMPs\_annotated$Probe\_ID, # Label with Probe IDs

x = 'logFC',

y = 'adj.P.Val',

pCutoff = 0.05,

FCcutoff = 0.1,

title = 'Colon Cancer vs. Normal Colon: Differential Methylation',

subtitle = 'Identified by limma-eBayes',

# Using the re-created significant\_and\_large\_effect\_DMPs for the caption counts

caption = paste0('Total DMPs = ', nrow(topDMPs\_annotated),

'\nSignificant (adj. P < 0.05) = ', nrow(subset(topDMPs\_annotated, adj.P.Val < 0.05)),

'\nSignificant & Delta Beta > 0.1 = ', nrow(significant\_and\_large\_effect\_DMPs)),

pointSize = 1.0,

labSize = 3.0,

legendPosition = 'right',

col = c('plum4', 'maroon', 'mediumpurple2', 'coral2'),

colAlpha = 0.5,

drawConnectors = TRUE,

widthConnectors = 0.5,

colConnectors = 'grey30'

)

print("\nPrinting Volcano Plot to RStudio Plots pane\n")

print(volcano\_plot) # This line makes the plot appear!

tryCatch({

ggsave("C:/Users/Admin/Desktop/Colon\_Volcano\_All\_probes.png", plot = volcano\_plot, width = 10, height = 8, dpi = 300)

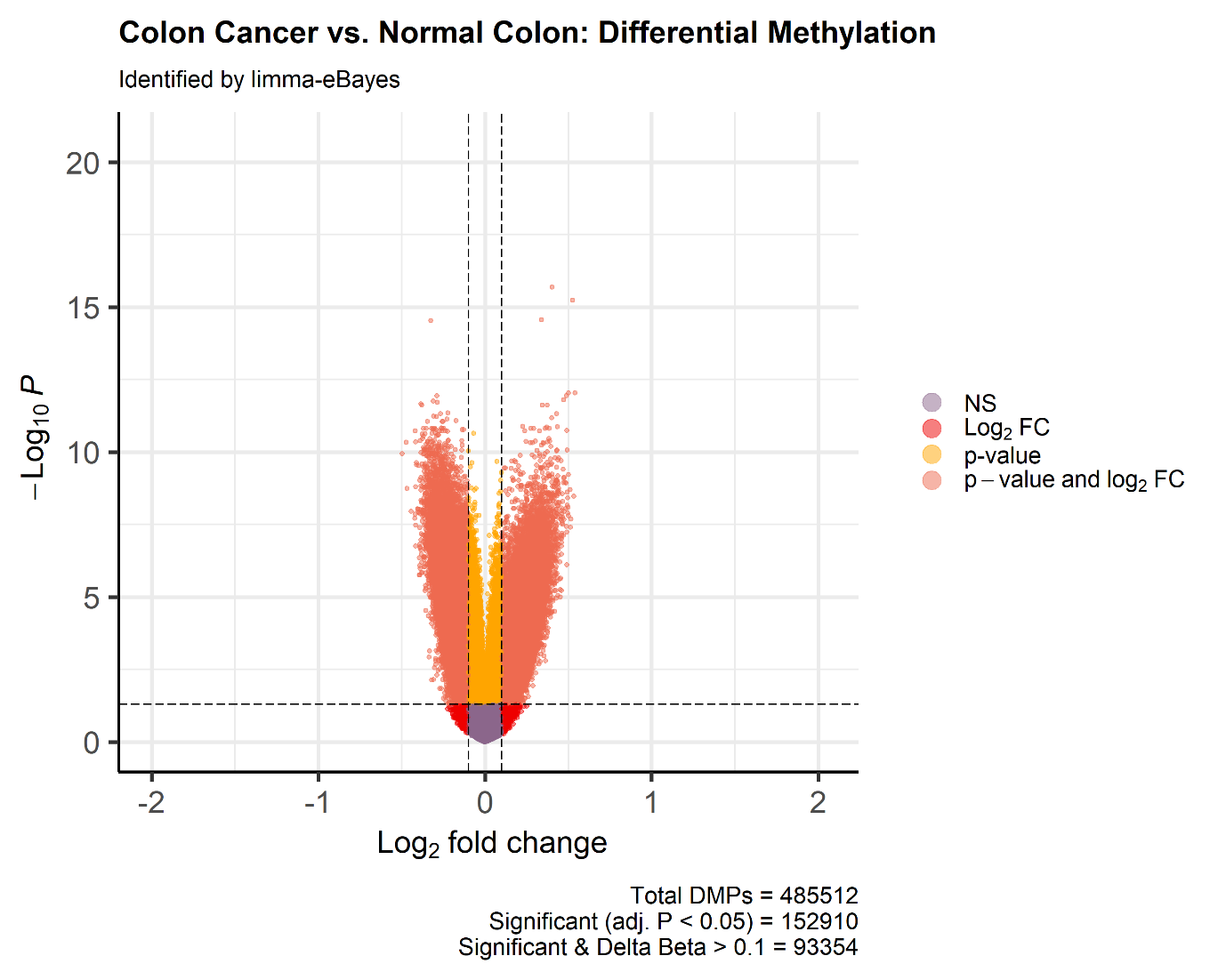
}, error = function(e) {

message("Error saving the plot")

})

cat("\nFinished Volcano Plot operations.\n")

Plots



* Query successful

This image is a **volcano plot** visualizing differential methylation between Colon Cancer and Normal Colon samples. Here's an interpretation of its components and what it tells us:

**Overall Purpose:** The plot aims to identify differentially methylated probes (DMPs) between colon cancer and normal colon tissues. Differential methylation refers to changes in the level of DNA methylation, a crucial epigenetic modification that can influence gene expression.

**Key Components of the Volcano Plot:**

* **X-axis (Log2​ fold change):** This represents the magnitude and direction of the methylation difference.
  + A positive Log2​ fold change indicates higher methylation in colon cancer compared to normal colon.
  + A negative Log2​ fold change indicates lower methylation in colon cancer compared to normal colon.
  + The further a point is from 0 on the x-axis, the larger the difference in methylation.
* **Y-axis (−Log10​P):** This represents the statistical significance of the methylation difference.
  + The P-value is a measure of the probability that the observed difference occurred by chance.
  + Taking the negative Log10​ of the P-value means that smaller P-values (more significant differences) will result in larger (higher) values on the y-axis.
  + For example, a P-value of 0.05 corresponds to −Log10​(0.05)≈1.3, while a P-value of 0.001 corresponds to −Log10​(0.001)=3.

**Color Coding of Points:**

* **Purple (NS - Not Significant):** These probes do not meet the criteria for statistical significance or fold change. They are likely not differentially methylated.
* **Red (Log2​ FC - Log2​ fold change only):** These probes show a significant Log2​ fold change but might not meet the P-value threshold for significance. In this specific plot, it seems these are probes that have a large fold change but fall below the horizontal P-value threshold.
* **Orange (P-value - P-value only):** These probes are statistically significant (above the P-value threshold) but may not have a large enough Log2​ fold change to be considered biologically meaningful.
* **Dark Red/Brown (P-value and Log2​ FC - P-value and Log2​ fold change):** These are the most interesting probes. They are both statistically significant (high on the y-axis) and have a substantial Log2​ fold change (far from 0 on the x-axis). These are the truly differentially methylated probes (DMPs).

**Lines on the Plot:**

* **Horizontal Dashed Line:** This line represents the P-value significance threshold. Any point above this line is considered statistically significant (e.g., P < 0.05, which is roughly at −Log10​(0.05)≈1.3). The specific P-value cutoff used here appears to be 0.05 given the annotation "Significant (adj. P < 0.05)".
* **Vertical Dashed Lines:** These lines represent the Log2​ fold change thresholds. Points outside these lines (either to the left or right) are considered to have a biologically significant change in methylation. Based on the annotation "Significant & Delta Beta > 0.1", it implies a fold change threshold might be related to a Delta Beta value, which for methylation data is often preferred over simple Log2​ fold change, though Log2​ fold change is plotted here. The exact Log2​ fold change corresponding to Delta Beta > 0.1 isn't directly visible but it defines the region for "significant Log2​ FC".

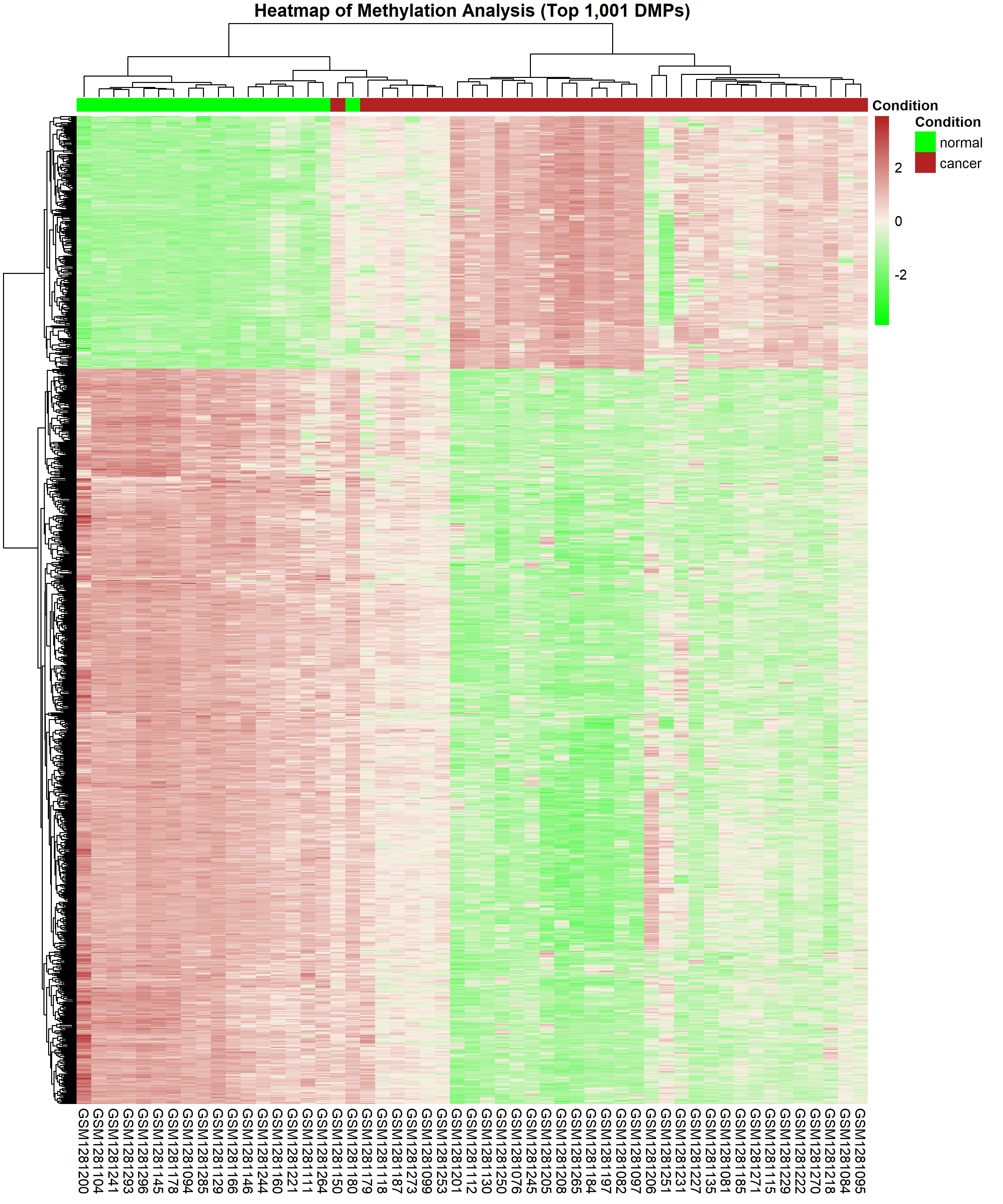
**Summary Statistics (at the bottom):**

* **Total DMPs = 485512:** This is the total number of methylation probes analyzed in the study.
* **Significant (adj. P < 0.05) = 152910:** This is the number of probes that are statistically significant based on their adjusted P-value (i.e., they fall above the horizontal dashed line). These are the orange and dark red/brown points.
* **Significant & Delta Beta > 0.1 = 93354:** This is the most crucial number. It represents the number of probes that are both statistically significant (adj. P < 0.05) AND have a biologically meaningful change in methylation (Delta Beta > 0.1, corresponding to the dark red/brown points). These are the true differentially methylated probes.

**Interpretation of Findings from the Plot:**

* **Many Differentially Methylated Probes:** There are a very large number of probes (93,354) that are both statistically and biologically significant, indicating widespread methylation changes between colon cancer and normal colon.
* **Both Hyper- and Hypomethylation:** The cluster of dark red/brown points on both the left and right sides of the plot suggests that there are probes that become significantly *hypomethylated* (left side, negative Log2​ fold change) and probes that become significantly *hypermethylated* (right side, positive Log2​ fold change) in colon cancer compared to normal colon.
* **Higher Significance for some DMPs:** The points extending very high on the y-axis (e.g., up to 20) indicate some probes have extremely low P-values, meaning the observed methylation differences for these probes are highly unlikely to be due to chance.
* **Variability in Fold Change:** While many points are clustered around small fold changes, there are also probes with substantial positive and negative Log2​ fold changes.

**In conclusion, this volcano plot powerfully illustrates that colon cancer is associated with extensive and statistically significant differential methylation, involving both increased (hypermethylation) and decreased (hypomethylation) methylation at a large number of genomic loci compared to normal colon tissue.** These findings are crucial for understanding the epigenetic landscape of colon cancer and for identifying potential biomarkers or therapeutic targets.



* Query successful

This image is a **heatmap** visualizing the methylation levels of the **top 1,001 differentially methylated probes (DMPs)** between colon cancer and normal colon samples. It provides a more detailed, sample-specific view of the methylation changes identified by the volcano plot.

Here's a breakdown of the heatmap components and its interpretation:

**Overall Purpose:** The heatmap visually clusters samples and methylation probes based on their methylation patterns, allowing for the identification of distinct groups and consistent methylation changes in the top DMPs.

**Key Components of the Heatmap:**

* **Rows (Methylation Probes):** Each row represents one of the top 1,001 differentially methylated probes (DMPs). The dendrogram on the left clusters probes with similar methylation patterns across samples.
* **Columns (Samples):** Each column represents an individual sample. The labels at the bottom (e.g., GSM128104, GSM128120) are unique identifiers for each sample. The dendrogram at the top clusters samples with similar methylation profiles across the selected probes.
* **Color Scale (Right Side):** This scale indicates the methylation level:
  + **Red:** Higher methylation.
  + **Green:** Lower methylation.
  + The numbers 2 and -2 indicate the range or intensity of methylation values, likely representing Log2​ ratios or normalized methylation beta values, where 2 is highest and -2 is lowest.
* **Top Bar (Condition):** This colored bar above the sample labels indicates the "Condition" of each sample:
  + **Green:** Normal colon sample.
  + **Red:** Colon cancer sample. This bar is crucial for observing how methylation patterns segregate by disease status.

**Interpretation of Findings from the Heatmap:**

1. **Clear Sample Segregation:**
   * The most striking observation is the nearly perfect segregation of samples based on their "Condition." The top dendrogram clearly separates the green (normal) samples from the red (cancer) samples into two main clusters.
   * This indicates that the top 1,001 DMPs are highly effective in distinguishing between normal colon tissue and colon cancer tissue. This strong separation confirms the robustness of the differential methylation findings.
2. **Distinct Methylation Patterns in Cancer vs. Normal:**
   * **Upper Block of Probes:** There's a large block of probes (upper part of the heatmap) that are predominantly **green** in the normal samples and predominantly **red** in the cancer samples. This signifies widespread **hypermethylation** (increased methylation) in colon cancer at these specific probes compared to normal tissue.
   * **Lower Block of Probes:** Conversely, there's another large block of probes (lower part of the heatmap) that are predominantly **red** in the normal samples and predominantly **green** in the cancer samples. This indicates widespread **hypomethylation** (decreased methylation) in colon cancer at these probes.
3. **Consistency within Groups:**
   * Within the "normal" group, the methylation patterns (green/red balance for specific probes) are largely consistent across different normal samples.
   * Similarly, within the "cancer" group, there's a strong consistency in methylation patterns across different cancer samples, although there might be some minor heterogeneity, as seen by slight variations in color intensity or small sub-clusters within the cancer samples.
4. **Individual Sample Exceptions (if any):**
   * While the segregation is almost perfect, look closely for any "misplaced" samples (e.g., a green bar sample clustering with red bar samples, or vice versa). In this specific heatmap, the clustering appears remarkably clean, suggesting the DMPs are highly discriminatory. There might be a very slight mixed cluster of normal samples around GSM128113 - GSM128109 area, but overall, it's very distinct.

**In conclusion:** This heatmap visually validates the findings from the volcano plot. It clearly demonstrates that the top 1,001 differentially methylated probes exhibit distinct methylation patterns that accurately discriminate between colon cancer and normal colon samples. The presence of large blocks of both hypermethylated and hypomethylated probes in cancer highlights the extensive epigenetic reprogramming that occurs in this disease, which is a hallmark of cancer. This visual representation reinforces the idea that these methylation changes could serve as powerful biomarkers for colon cancer diagnosis or prognosis.