Intro to Epigenomics Assignment 3 Brukt

Bruk Tefera

11/03/24

```
knitr::opts_knit$set(root.dir = normalizePath("/home1/tefera/analysis_data"))
```

Package Download and Data-cleaning

```
if (!require("sesameData", quietly = TRUE))
BiocManager::install("sesameData")
```

```
##
## Attaching package: 'generics'
```

```
## The following objects are masked from 'package:base':
##
##     as.difftime, as.factor, as.ordered, intersect, is.element, setdiff,
##     setequal, union
```

```
##
## Attaching package: 'BiocGenerics'
```

```
## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs
```

```
## The following objects are masked from 'package:base':
##
##     anyDuplicated, aperm, append, as.data.frame, basename, cbind,
##     colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
##     get, grep, grepl, is.unsorted, lapply, Map, mapply, match, mget,
##     order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,
##     rbind, Reduce, rownames, sapply, saveRDS, table, tapply, unique,
##     unsplit, which.max, which.min
```

```
## Loading sesameData.
```

```
if (!require("sesame", quietly = TRUE))
BiocManager::install("sesame")
```

```
##
## ----------------------------------------------------------
## | SEnsible Step-wise Analysis of DNA MEthylation (SeSAMe)
## | --------------------------------------------------------
## | Please cache auxiliary data by "sesameDataCache()".
## | This needs to be done only once per SeSAMe installation.
## ----------------------------------------------------------
```

```
if (!require("limma", quietly = TRUE))
BiocManager::install("limma")
```

```
##
## Attaching package: 'limma'
```

```
## The following object is masked from 'package:BiocGenerics':
##
##     plotMA
```

```
if (!require("DESeq2", quietly = TRUE)) BiocManager::install("DESeq2")
```

```
##
## Attaching package: 'S4Vectors'
```

```
## The following object is masked from 'package:utils':
##
##     findMatches
```

```
## The following objects are masked from 'package:base':
##
##     expand.grid, I, unname
```

```
##
## Attaching package: 'MatrixGenerics'
## The following objects are masked from 'package:matrixStats':
##
##     colAlls, colAnyNAs, colAnys, colAvgsPerRowSet, colCollapse,
##     colCounts, colCummaxs, colCummins, colCumprods, colCumsums,
##     colDiffs, colIQRDiffs, colIQRs, colLogSumExps, colMadDiffs,
##     colMads, colMaxs, colMeans2, colMedians, colMins, colOrderStats,
##     colProds, colQuantiles, colRanges, colRanks, colSdDiffs, colSds,
##     colSums2, colTabulates, colVarDiffs, colVars, colWeightedMads,
##     colWeightedMeans, colWeightedMedians, colWeightedSds,
##     colWeightedVars, rowAlls, rowAnyNAs, rowAnys, rowAvgsPerColSet,
##     rowCollapse, rowCounts, rowCummaxs, rowCummins, rowCumprods,
##     rowCumsums, rowDiffs, rowIQRDiffs, rowIQRs, rowLogSumExps,
##     rowMadDiffs, rowMads, rowMaxs, rowMeans2, rowMedians, rowMins,
##     rowOrderStats, rowProds, rowQuantiles, rowRanges, rowRanks,
##     rowSdDiffs, rowSds, rowSums2, rowTabulates, rowVarDiffs, rowVars,
##     rowWeightedMads, rowWeightedMeans, rowWeightedMedians,
##     rowWeightedSds, rowWeightedVars
```

```
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
```

```
##
## Attaching package: 'Biobase'
```

```
## The following object is masked from 'package:MatrixGenerics':
##
##     rowMedians
```

```
## The following objects are masked from 'package:matrixStats':
##
##     anyMissing, rowMedians
```

```
## The following object is masked from 'package:ExperimentHub':
##
##     cache
```

```
## The following object is masked from 'package:AnnotationHub':
##
##     cache
```

```
library(DESeq2)
```

Load in all necessary packages

```
library(TCGAbiolinks)
library(sesame)
library(sesameData)
library(limma)
library(ggplot2)
```

```
methylation_clinical <- read.csv("/project/rohs_1070/analysis_data/brca_methylation_clinical.cs
v", row.names = 1)
cpg_sites <- read.csv("/project/rohs_1070/analysis_data/brca_cpg_sites.csv", row.names = 1)

library(data.table)
```

```
##
## Attaching package: 'data.table'
```

```
## The following object is masked from 'package:SummarizedExperiment':
##
##     shift
```

```
## The following object is masked from 'package:GenomicRanges':
##
##     shift
```

```
## The following object is masked from 'package:IRanges':
##
##     shift
```

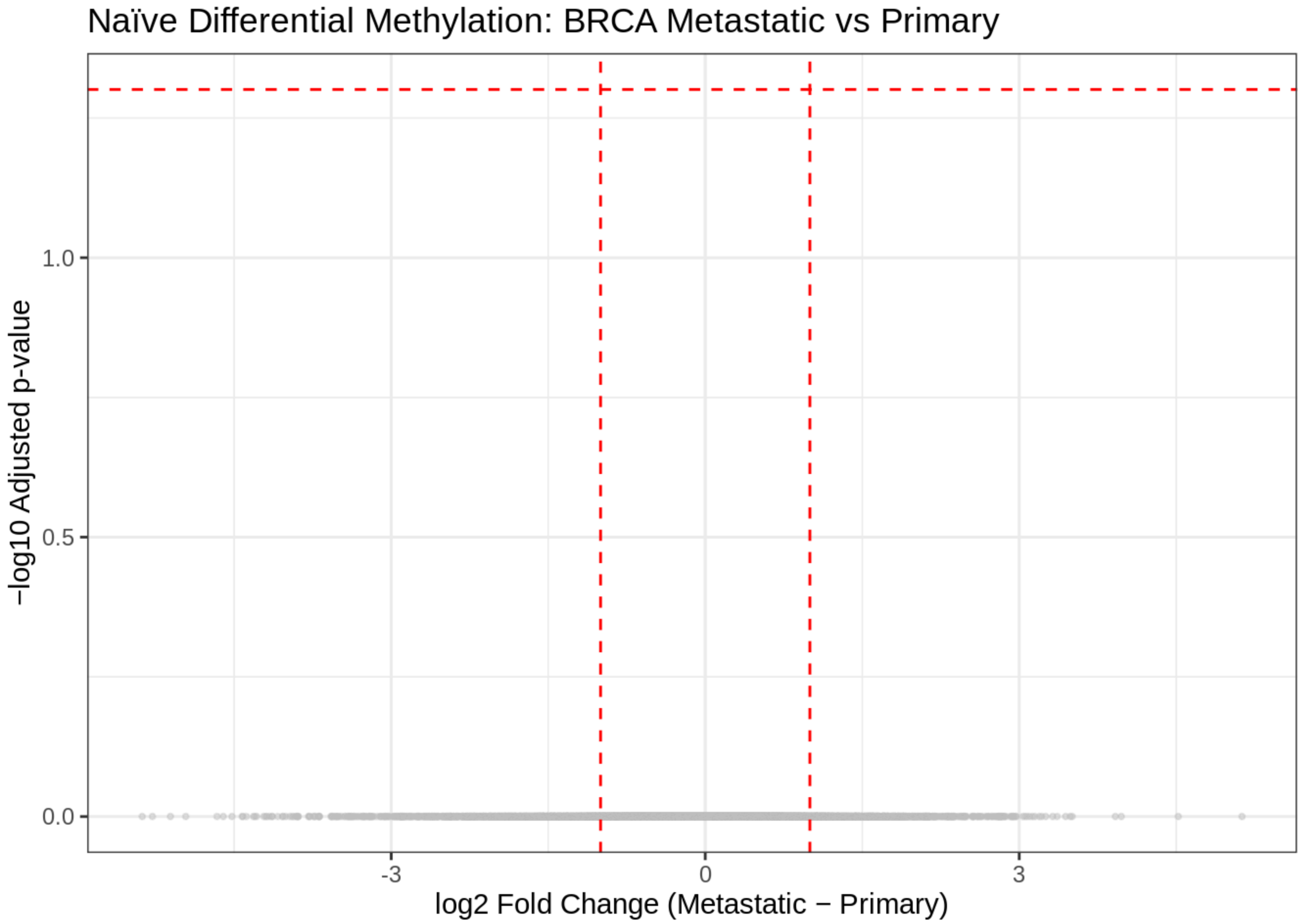
```
## The following objects are masked from 'package:S4Vectors':
##
##     first, second
```

```
betas <- fread("/project/rohs_1070/analysis_data/brca_methylation_betas.csv",
               nrows = 20000, showProgress = TRUE)
rownames(betas) <- betas[[1]]
betas[[1]] <- NULL
```

1. Naive Differential Methylation

```
### (1) Naive Differential Methylation: Metastatic vs Primary BRCA

betas <- as.data.frame(betas)

# Standardize sample IDs
colnames(betas) <- substr(gsub("\\.", "-", colnames(betas)), 1, 12)
methylation_clinical$barcode12 <- substr(methylation_clinical$barcode, 1, 12)

# Keep only primary and metastatic samples
methylation_clinical <- subset(
  methylation_clinical,
  definition %in% c("Primary solid Tumor", "Metastatic")
)
methylation_clinical$definition <- droplevels(factor(
  methylation_clinical$definition,
  levels = c("Primary solid Tumor", "Metastatic")
))

# Align samples present in both datasets
common <- intersect(colnames(betas), methylation_clinical$barcode12)
betas <- betas[, common, drop = FALSE]
methylation_clinical <- methylation_clinical[
  match(common, methylation_clinical$barcode12),
]

# Clean up any NA samples
na_mask <- complete.cases(methylation_clinical$definition)
betas <- betas[, na_mask, drop = FALSE]
methylation_clinical <- methylation_clinical[na_mask, ]

# Convert β-values to M-values safely
betas[betas <= 0] <- 1e-6
betas[betas >= 1] <- 1 - 1e-6
mval <- log2(betas / (1 - betas))

# limma model: metastatic vs primary
library(limma)
design <- model.matrix(~ 0 + methylation_clinical$definition)
colnames(design) <- c("Primary", "Metastatic")
contrast <- makeContrasts(MetMinusPrim = Metastatic - Primary, levels = design)

fit <- lmFit(mval, design)
fit <- contrasts.fit(fit, contrast)
fit <- eBayes(fit)
```

```
# Compile results
dm_results <- topTable(fit, coef = "MetMinusPrim", number = Inf, sort.by = "P")
dm_results$CpG <- rownames(dm_results)

# Map CpG to gene names (only if available)
if ("gene" %in% colnames(cpg_sites)) {
  dm_results$gene <- cpg_sites$gene[match(dm_results$CpG, rownames(cpg_sites))]
}

# Mark significant CpGs
dm_results$signif <- with(dm_results, abs(logFC) >= 1 & adj.P.Val < 0.05)

# Volcano plot
library(ggplot2)
ggplot(dm_results, aes(x = logFC, y = -log10(adj.P.Val), color = signif)) +
  geom_point(alpha = 0.4, size = 0.7) +
  geom_vline(xintercept = c(-1, 1), linetype = "dashed", color = "red") +
  geom_hline(yintercept = -log10(0.05), linetype = "dashed", color = "red") +
  scale_color_manual(values = c("TRUE" = "blue", "FALSE" = "grey")) +
  labs(
    x = "log2 Fold Change (Metastatic – Primary)",
    y = "–log10 Adjusted p-value",
    title = "Naïve Differential Methylation: BRCA Metastatic vs Primary"
  ) +
  theme_bw() +
  theme(legend.position = "none")
```


Naïve Differential Methylation: BRCA Metastatic vs Primary

```
### In my volcano plot, I observed that nearly all CpG sites clustered around a log2 fold change
of zero, with no points passing the significance thresholds. This indicates there were no strong
ly differentially methylated sites between metastatic and primary BRCA samples. The lack of sign
al is likely due to the very small metastatic sample size, which limits statistical power.
```

2. Direct comparison of methylation status to transcriptional activity

#…INSERT DESeq2 Stuff here to generate 'results'…

```
library(DESeq2)
library(limma)
library(matrixStats)

rna_clinical <- read.csv("/home1/tefera/analysis_data/BRCA_rna_clinical.csv")
rna_genes <- read.csv("/home1/tefera/analysis_data/BRCA_rna_genes.csv")
rna_counts   <- read.csv("/home1/tefera/analysis_data/BRCA_rna_counts.csv",
                         header = TRUE, row.names = 1, check.names = FALSE)

colnames(rna_counts) <- substr(colnames(rna_counts), 1, 12)
rna_clinical$barcode <- substr(rna_clinical$bcr_patient_barcode, 1, 12)
rna_clinical <- subset(rna_clinical, definition %in% c("Primary solid Tumor", "Metastatic"))

common <- intersect(colnames(rna_counts), rna_clinical$barcode)
rna_counts <- as.matrix(rna_counts[, common, drop = FALSE])
rna_clinical <- rna_clinical[match(colnames(rna_counts), rna_clinical$barcode), ]

rna_counts <- rna_counts[rowSums(rna_counts) >= 20, ]
rna_counts <- rna_counts[head(order(rowVars(rna_counts), decreasing = TRUE), 2000), ]

rna_clinical$definition <- factor(gsub(" ", "_", rna_clinical$definition),
                                  levels = c("Primary_solid_Tumor", "Metastatic"))

dds <- DESeqDataSetFromMatrix(rna_counts, rna_clinical, ~ definition)
vst_mat <- assay(vst(dds))
fit <- eBayes(lmFit(vst_mat, model.matrix(~ definition, rna_clinical)))
results <- topTable(fit, coef = 2, number = Inf)
results$gene_name <- rownames(results)
```

```
# Identify genes that are both downregulated and hypomethylated
downregulated <- results[results$logFC < -1, 'gene_name']
hypomethylated <- dm_results[dm_results$logFC < -1, 'gene']
interest_genes <- intersect(downregulated, hypomethylated)
```

(Extra) Making Boxplots

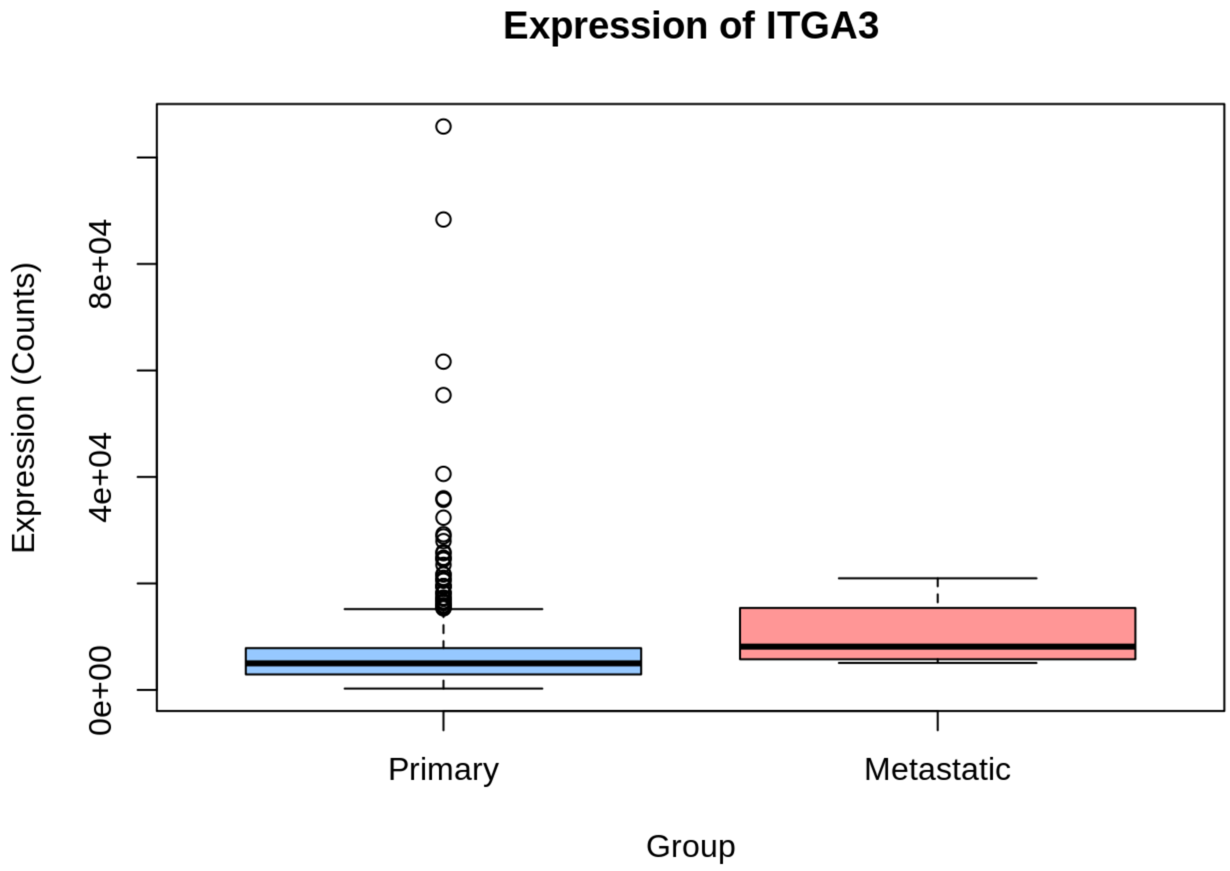
```
GENE <- "ITGA3"  # you can swap for PLXND1 or LAMP2

# find its Ensembl ID
ensembl_id <- rna_genes$gene_id[rna_genes$gene_name == GENE]
cat("Gene:", GENE, "| Ensembl ID:", ensembl_id, "\n")
```

```
## Gene: ITGA3 | Ensembl ID: ENSG00000005884.18
```

```
# extract tumor and metastatic samples
rna_clinical_tumor  <- rna_clinical$definition == "Primary_solid_Tumor"
rna_clinical_metast <- rna_clinical$definition == "Metastatic"

rna_tumor  <- as.numeric(rna_counts[rownames(rna_counts) == ensembl_id, rna_clinical_tumor])
rna_metast <- as.numeric(rna_counts[rownames(rna_counts) == ensembl_id, rna_clinical_metast])

cat("Primary samples:", length(rna_tumor),
    "| Metastatic samples:", length(rna_metast), "\n")
```

```
## Primary samples: 1091 | Metastatic samples: 4
```

```
par(mfrow = c(1, 1))
par(mar = c(5, 5, 4, 2) + 0.1)
boxplot(rna_tumor, rna_metast,
        xlab = "Group", ylab = "Expression (Counts)",
        names = c("Primary", "Metastatic"),
        col = c("#99CCFF", "#FF9999"),
        main = paste("Expression of", GENE))
```



```
### I observe that ITGA3 expression tends to be higher in metastatic samples, while primary tumors show a wider spread with several outliers. Because there are only a few metastatic cases, the differences cannot be considered definitive, but the overall trend is consistent with previous studies linking ITGA3 to increased cell adhesion and metastatic potential in breast cancer.
```