Transcriptomic preprocessing and Differential Expression analysis of GSE28360 (HTN Datasets)

Loranda_Calderon

2025-06-01

```
library(GEOquery)
```

```
## Loading required package: Biobase

## Loading required package: BiocGenerics

##
## 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, intersect, is.unsorted, lapply, Map, mapply,
##     match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
##     Position, rank, rbind, Reduce, rownames, sapply, setdiff, sort,
##     table, tapply, union, unique, unsplit, which.max, which.min

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

## Setting options('download.file.method.GEOquery'='auto')

## Setting options('GEOquery.inmemory.gpl'=FALSE)
```

```
library(oligo)
```

```
## Loading required package: oligoClasses

## Welcome to oligoClasses version 1.64.0
```

```
## Loading required package: Biostrings

## Loading required package: S4Vectors

## Loading required package: stats4

##
## Attaching package: 'S4Vectors'

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

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

## Loading required package: IRanges

## Loading required package: XVector

## Loading required package: GenomeInfoDb

##
## Attaching package: 'Biostrings'

## The following object is masked from 'package:base':
##
##     strsplit

## ========================================================================

## Welcome to oligo version 1.66.0

## ========================================================================
```

```
library(limma)
```

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

## The following object is masked from 'package:oligo':
##
##     backgroundCorrect

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

```
library(sva)
```

```
## Loading required package: mgcv
```

```
## Loading required package: nlme
```

```
##
## Attaching package: 'nlme'
```

```
## The following object is masked from 'package:Biostrings':
##
##     collapse
```

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

```
## This is mgcv 1.9-1. For overview type 'help("mgcv-package")'.
```

```
## Loading required package: genefilter
```

```
## Loading required package: BiocParallel
```

```
library(WGCNA)
```

```
## Loading required package: dynamicTreeCut
```

```
## Loading required package: fastcluster
```

```
##
## Attaching package: 'fastcluster'
```

```
## The following object is masked from 'package:stats':
##
##     hclust
```

```
##
```

```
##
## Attaching package: 'WGCNA'
```

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

```
## The following object is masked from 'package:S4Vectors':
##
##     cor
```

```
## The following object is masked from 'package:stats':
##
##     cor
```

```
library(ensembldb)
```

```
## Loading required package: GenomicRanges
```

```
## Loading required package: GenomicFeatures
```

```
## Loading required package: AnnotationDbi
```

```
## Loading required package: AnnotationFilter
```

```
##
## Attaching package: 'ensembldb'
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
library(biomaRt)
library(arrayQualityMetrics)
```

```
# Download GSE28360 data
gse_HTN <- getGEO("GSE28360", GSEMatrix =TRUE,getGPL=FALSE)
```

```
## Found 1 file(s)
```

```
## GSE28360_series_matrix.txt.gz
```

```
datMeta_HTN <- pData(gse_HTN[[1]])
rownames(datMeta_HTN) <- datMeta_HTN$geo_accession
```

```
# Read GSE28360 data
setwd("/Users/lorandacalderonzamora/GSE28360/")
celfiles <- list.files(pattern = ".CEL.gz$", full.names = TRUE)
rawData_HTN <- read.celfiles(celfiles)
```

```
## Loading required package: pd.hugene.1.0.st.v1
```

```
## Loading required package: RSQLite
```

```
## Loading required package: DBI
```

```
## Platform design info loaded.
```

```
## Reading in : ./GSM701161.CEL.gz
## Reading in : ./GSM701162.CEL.gz
## Reading in : ./GSM701163.CEL.gz
## Reading in : ./GSM701164.CEL.gz
## Reading in : ./GSM701165.CEL.gz
```

```
## Reading in : ./GSM701166.CEL.gz
## Reading in : ./GSM701167.CEL.gz
## Reading in : ./GSM701168.CEL.gz
## Reading in : ./GSM701169.CEL.gz
## Reading in : ./GSM701170.CEL.gz
## Reading in : ./GSM701171.CEL.gz
## Reading in : ./GSM701172.CEL.gz
## Reading in : ./GSM701173.CEL.gz
## Reading in : ./GSM701174.CEL.gz
```

```
datExpr_HTN <- exprs(rawData_HTN)
```

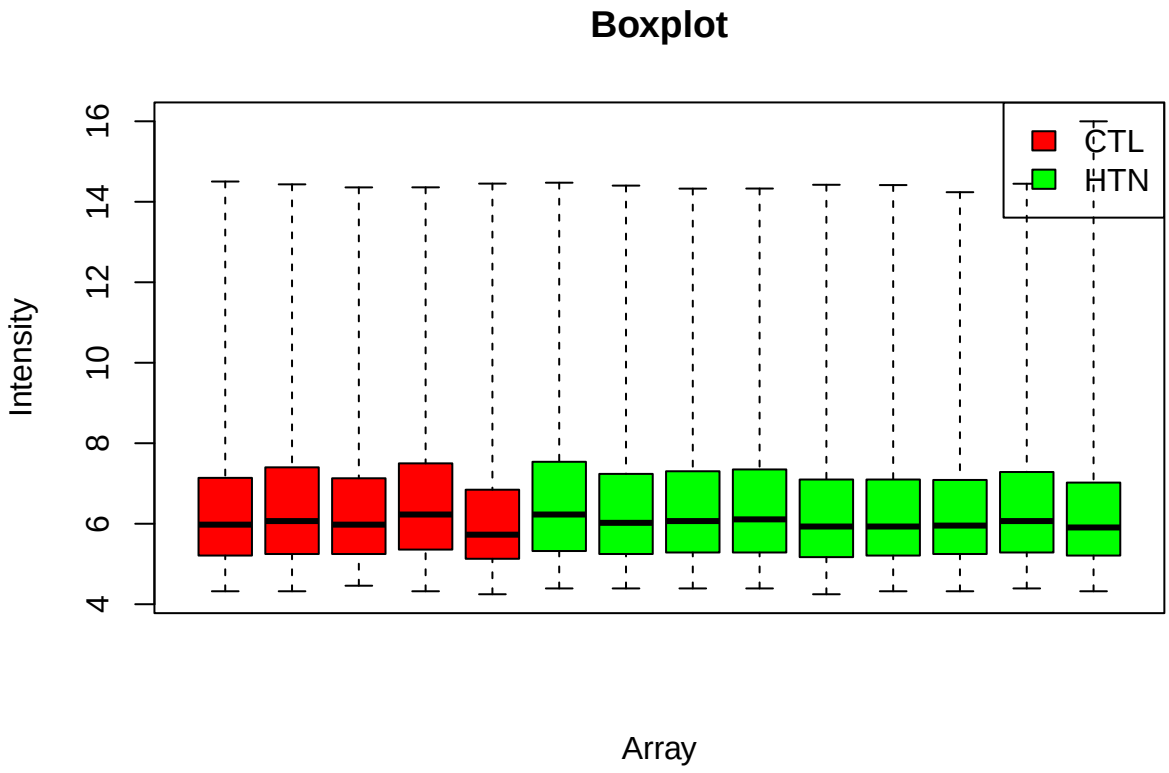
```
# Align datMeta_HTN and datExpr_HTN by sample identifiers
GSM_HTN <- rownames(pData(rawData_HTN))
GSM_HTN <- substr(GSM_HTN, 1, 9)
idx_HTN <- match(GSM_HTN, datMeta_HTN$geo_accession)
datMeta_HTN <- datMeta_HTN[idx_HTN, ]
colnames(datExpr_HTN)=rownames(datMeta_HTN)
```

```
# Cleaning and formatting of GSE28360 metadata
datMeta_HTN <- datMeta_HTN[, -c(3:7, 14:36)]
colnames(datMeta_HTN)[3] <- "Dx"
datMeta_HTN$Dx[rownames(datMeta_HTN) %in% c("GSM701161", "GSM701162", "GSM701163", "GSM701164", "GSM701
datMeta_HTN$Dx[rownames(datMeta_HTN) %in% c("GSM701166", "GSM701167", "GSM701168", "GSM701169", "GSM701
datMeta_HTN$Dx <- as.factor(datMeta_HTN$Dx)
```

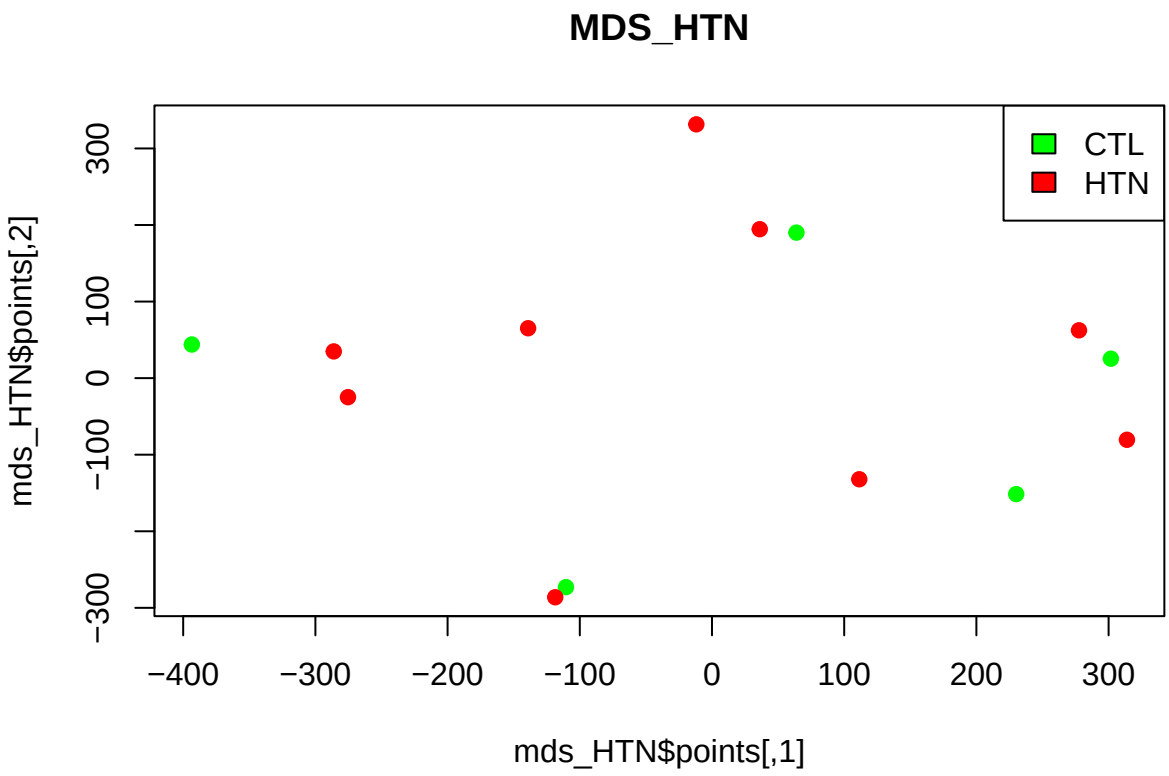
```
# Preprocessing and quality assessment of GSE28360 raw expression data
datExpr_HTN <- log2(datExpr_HTN)
dim(datExpr_HTN)
```

```
## [1] 1102500      14
```

```
# Exploratory visualization of GSE28360 raw data
boxplot(datExpr_HTN,range=0, col=c('red', 'green')[as.numeric(datMeta_HTN$Dx)], xaxt='n', xlab = "Array
legend("topright",legend = levels(datMeta_HTN$Dx),fill = c('red', 'green')[as.numeric(as.factor(levels(
```




```
mds_HTN = cmdscale(dist(t(datExpr_HTN)),eig=TRUE)
plot(mds_HTN$points,col=c('green', 'red')[as.numeric(datMeta_HTN$Dx)],pch=19,main="MDS_HTN")
legend("topright",legend = levels(datMeta_HTN$Dx),fill =c('green', 'red')[as.numeric(as.factor(levels(da
```

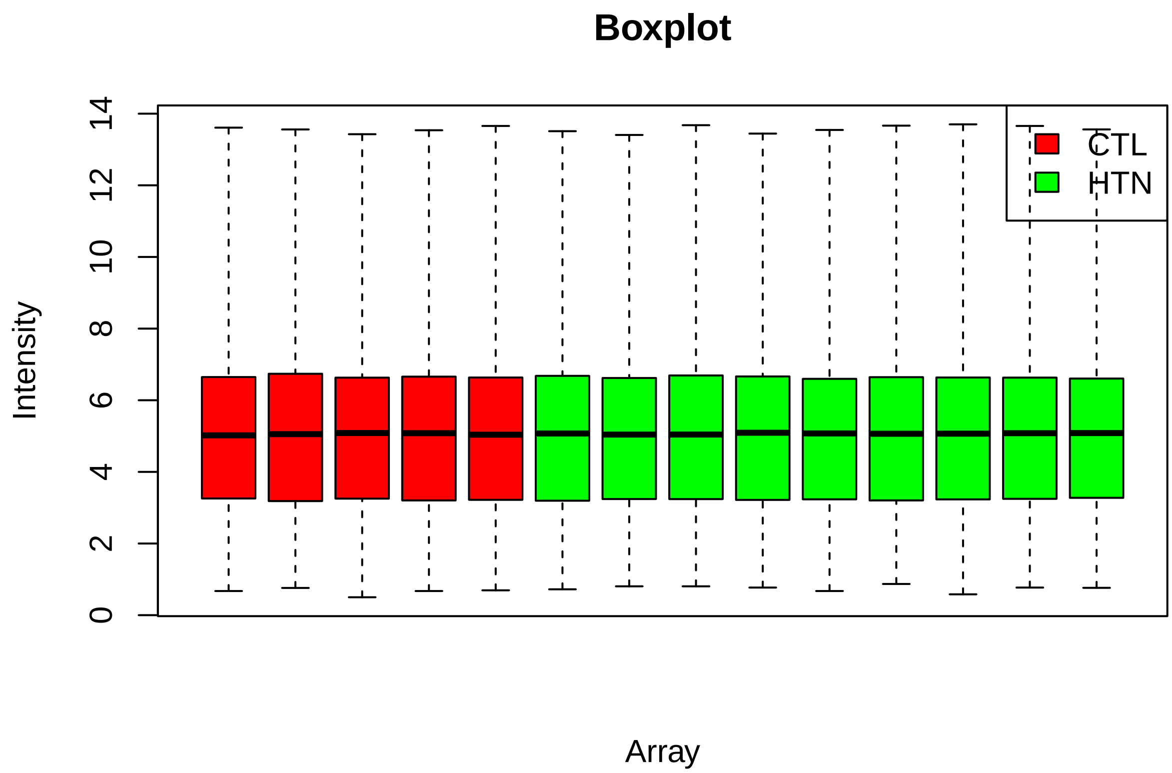



```
# Normalization using RMA
datExpr_HTN <- rma(rawData_HTN)
```

```
## Background correcting
## Normalizing
## Calculating Expression
```

```
datExpr_HTN <- exprs(datExpr_HTN)
```

```
# Exploratory visualization of GSE28360 normalized data
boxplot(datExpr_HTN,range=0, col=c('red', 'green')[as.numeric(datMeta_HTN$Dx)], xaxt='n', xlab = "Array
legend("topright",legend = levels(datMeta_HTN$Dx),fill = c('red', 'green')[as.numeric(as.factor(levels(
```




```r
# QC analysis with arrayQualityMetrics
datMeta_proc_HTN <- new("AnnotatedDataFrame", data = datMeta_HTN)
colnames(datExpr_HTN) <- rownames(datMeta_HTN)
eset_HTN <- new("ExpressionSet", exprs = datExpr_HTN, phenoData = datMeta_proc_HTN)


arrayQualityMetrics(expressionset = eset_HTN,
                    outdir = "/Users/lorandacalderonzamora/Downloads/QC_GSE28360_Report",
                    force = TRUE,
                    do.logtransform = FALSE)
```

```
## The directory '/Users/lorandacalderonzamora/Downloads/QC_GSE28360_Report' has been created.
```

```
## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value
## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value
## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value
## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value
## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value
## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value
```

```
## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value
## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value
## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value
## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value
## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value
## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value
## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value
## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value
```

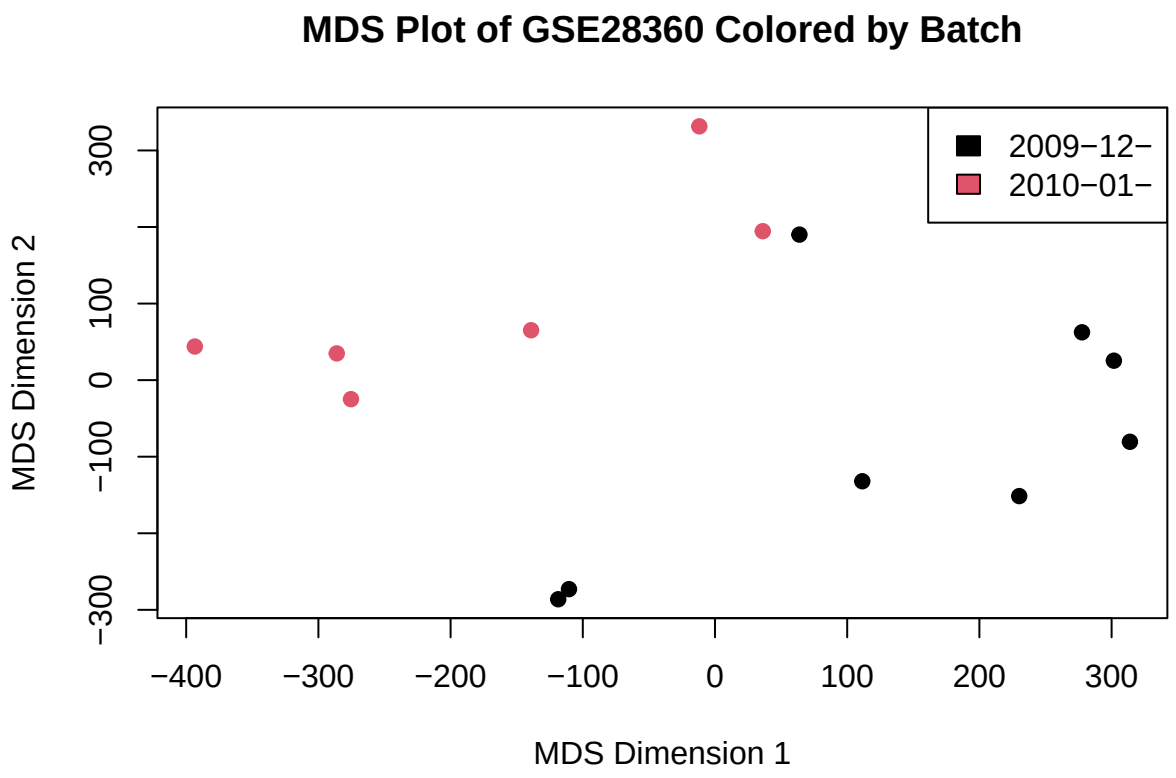
```
## (loaded the KernSmooth namespace)
```

```r
# Extract ScanDate from GSE28360 for batch effect correction
batch_HTN <- protocolData(rawData_HTN)$dates
batch_HTN <- substr(batch_HTN,1,8)
batch_HTN <- as.factor(batch_HTN)
table(batch_HTN)
```

```
## batch_HTN
## 2009-12- 2010-01-
##        8        6
```

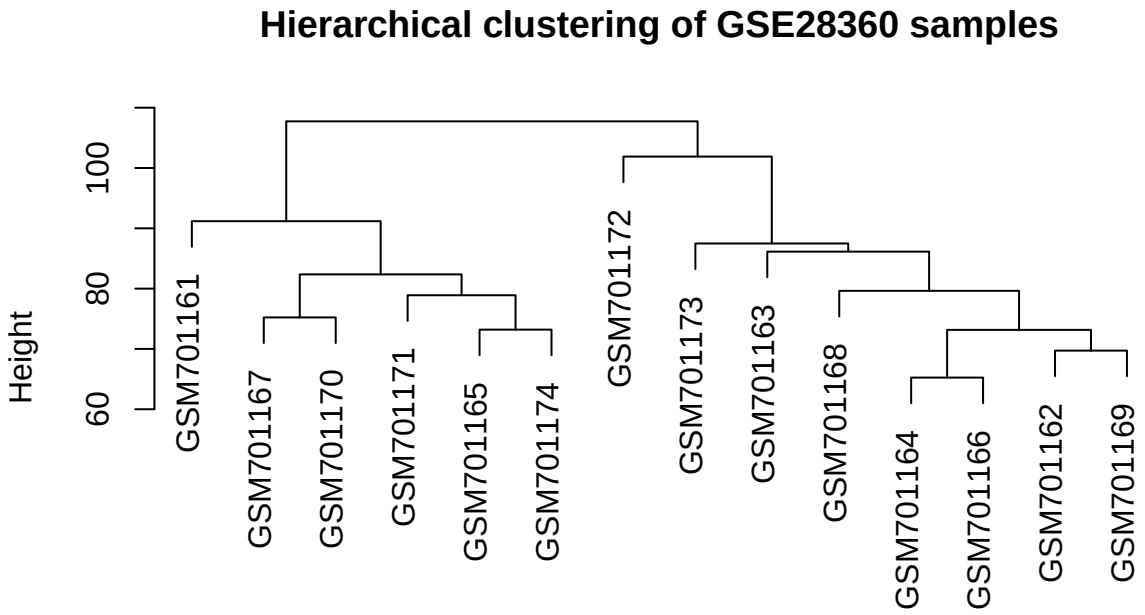
```r
datMeta_HTN$Batch <- batch_HTN
```

```r
# Visualization of ScanDate metadata from GSE28360 to identify potential batch effects
plot(mds_HTN$points,col = as.numeric(datMeta_HTN$Batch),pch=19,main="MDS Plot of GSE28360 Colored by Ba
legend("topright",legend = levels(datMeta_HTN$Batch),fill = as.numeric(as.factor(levels(datMeta_HTN$Bat
```



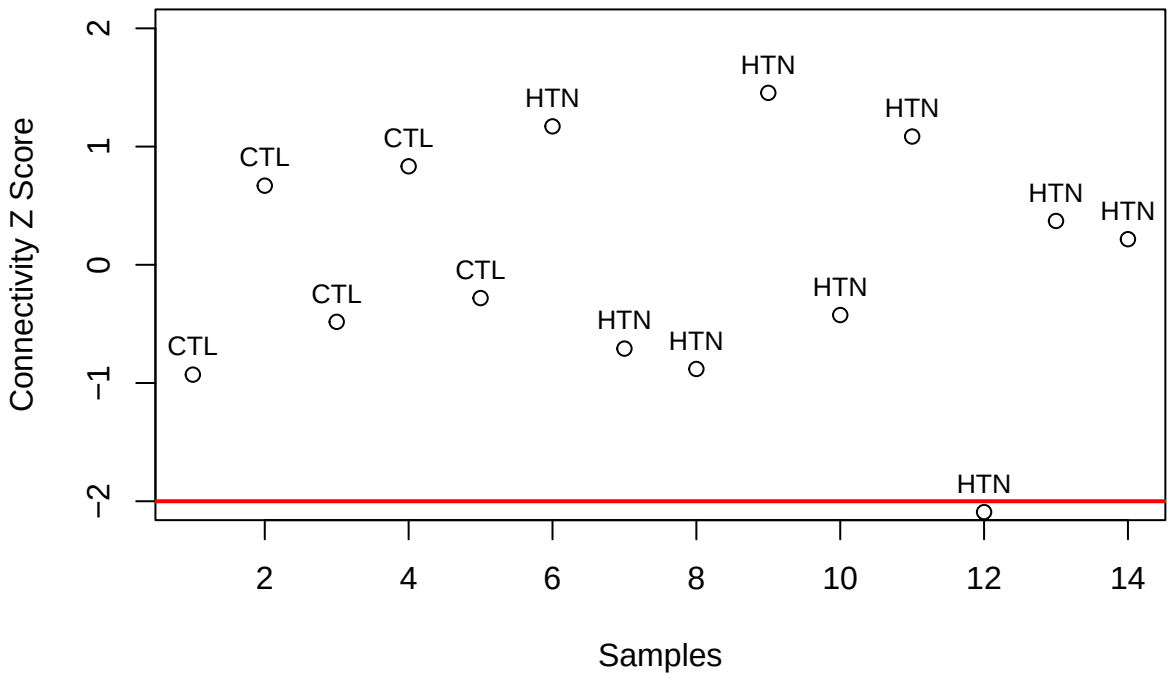

```r
# Create ExpressionSet object after Batch effect assessment
datMeta_HTN$Batch <- batch_HTN
datMeta_proc_HTN <- new("AnnotatedDataFrame", data = datMeta_HTN)
colnames(datExpr_HTN) <- rownames(datMeta_HTN)
datAll_HTN <- new("ExpressionSet", exprs = datExpr_HTN, phenoData = datMeta_proc_HTN)
# No singular batch was detected in the GSE28360 dataset.
# Therefore, batch correction with ComBat is technically feasible.
# However, as no evident batch effect was observed in exploratory analyses (MDS),
# ComBat was not applied, and no batch removal was necessary.
```

```r
# Sample Clustering and outlier detection
tree_HTN <- hclust(dist(t(exprs(datAll_HTN))), method = "average")
plot(tree_HTN, main = "Hierarchical clustering of GSE28360 samples", xlab = "", sub = "")
```




```
normadj_HTN <- (0.5 + 0.5*bicor(exprs(datAll_HTN)))^2
netsummary_HTN <- fundamentalNetworkConcepts(normadj_HTN)
C_HTN <- netsummary_HTN$Connectivity
Z.C_HTN <- (C_HTN - mean(C_HTN)) / sqrt(var(C_HTN))

datLabel_HTN <- pData(datAll_HTN)$Dx
plot(1:length(Z.C_HTN),Z.C_HTN,main="Outlier plot of GSE283604 samples ",xlab = "Samples",ylab="Connecti
text(1:length(Z.C_HTN),Z.C_HTN,label=datLabel_HTN,pos=3,cex=0.8)
abline(h= -2, col="red", lwd = 2)
```




```
# Identify and remove potential outlier from GSE28360 samples based on connectivity Z-score
# No samples exceeded the threshold (Z < -2), so none were removed
to_keep_HTN <- abs(Z.C_HTN) < 2
table(to_keep_HTN)
```

```
## to_keep_HTN
## FALSE  TRUE
##     1    13
```

```
colnames(exprs(datAll_HTN))[!to_keep_HTN]
```

```
## [1] "GSM701172"
```

```
datAll_HTN <- datAll_HTN[, to_keep_HTN]
```

```
# Annotating Probes for GSE28360 dataset
ensembl <- useEnsembl(biomart = "ensembl", dataset = "hsapiens_gene_ensembl")
identifier_HTN <- "affy_hugene_1_0_st_v1"
getinfo_HTN <- c("affy_hugene_1_0_st_v1", "ensembl_gene_id", "entrezgene_id", "external_gene_name")
geneDat_HTN <- getBM(attributes = getinfo_HTN,
                     filters = identifier_HTN,
                     values = rownames(exprs(datAll_HTN)),
                     mart = ensembl)
idx_HTN <- match(rownames(exprs(datAll_HTN)), geneDat_HTN$affy_hugene_1_0_st_v1)
```

```r
geneDat_HTN <- geneDat_HTN[idx_HTN, ]
table(is.na(geneDat_HTN$ensembl_gene_id))
```

```
##
## FALSE  TRUE
## 29120  4177
```

```r
to_keep_HTN <- !is.na(geneDat_HTN$ensembl_gene_id)
geneDat_HTN <- geneDat_HTN[to_keep_HTN, ]
datAll_HTN <- datAll_HTN[to_keep_HTN, ]
```

```r
# Collapse Rows for GSE28360 by Ensembl Gene ID
table(duplicated(geneDat_HTN$affy_hugene_1_0_st_v1))
```

```
##
## FALSE
## 29120
```

```r
table(duplicated(geneDat_HTN$ensembl_gene_id))
```

```
##
## FALSE  TRUE
## 24657  4463
```

```r
CR_HTN <- collapseRows(exprs(datAll_HTN),
                       rowGroup = geneDat_HTN$ensembl_gene_id,
                       rowID = geneDat_HTN$affy_hugene_1_0_st_v1)
CRdata_HTN <- CR_HTN$datETcollapsed
idx_HTN <- match(CR_HTN$group2row[,"selectedRowID"], geneDat_HTN$affy_hugene_1_0_st_v1)
geneDat_HTN <- geneDat_HTN[idx_HTN, ]
rownames(geneDat_HTN) <- geneDat_HTN$ensembl_gene_id
```

```r
# Differential Expression Analysis from GSE28360
mod <- model.matrix(~pData(datAll_HTN)$Dx)
fit <- lmFit(CR_HTN$datETcollapsed,mod)
fit <- eBayes(fit)
tt <- topTable(fit,coef = 2,n = Inf,genelist = geneDat_HTN)
head(tt)
```

```
##                 affy_hugene_1_0_st_v1 ensembl_gene_id entrezgene_id
## ENSG00000137959               7902541 ENSG00000137959         10964
## ENSG00000236398               8136844 ENSG00000236398        259285
## ENSG00000276192               7981708 ENSG00000276192            NA
## ENSG00000284182               8109157 ENSG00000284182        406935
## ENSG00000281935               7959696 ENSG00000281935        196385
## ENSG00000199788               7969091 ENSG00000199788            NA
##                 external_gene_name      logFC  AveExpr         t      P.Value
## ENSG00000137959              IFI44L -0.7745074 6.350643 -4.446962 0.0005236413
## ENSG00000236398            TAS2R39 -0.3603632 2.241008 -4.216344 0.0008223336
## ENSG00000276192               IGHE  0.6918890 3.897220  4.124704 0.0009851986
```

```
## ENSG00000284182             MIR143 -0.3485161 4.275881 -4.107648 0.0010189744
## ENSG00000281935             DNAH10  0.4074155 3.879245  4.050686 0.0011405871
## ENSG00000199788             RNY3P2 -0.7034725 3.019876 -3.965422 0.0013508915
##                 adj.P.Val         B
## ENSG00000137959 0.9998188 -2.927027
## ENSG00000236398 0.9998188 -3.024032
## ENSG00000276192 0.9998188 -3.064056
## ENSG00000284182 0.9998188 -3.071598
## ENSG00000281935 0.9998188 -3.096996
## ENSG00000199788 0.9998188 -3.135617
```