GSE50567 Tumor Sample

Xinru Qiu

September 5th, 2017

Contents

1 Background of GSE50567

Organism: Homo sapiens

Summary: We analyzed 35 breast cancer specimens. Surgical samples obtained during mastectomy were flash-frozen in liquid nitrogen and stored at -80°C. Only samples from patients without neoadjuvant chemotherapy were used in this study as chemotherapy may seriously affect gene expression profile. All tissue samples were collected at the Pomeranian Medical University in Szczecin. Seventeen tumor samples were collected from patients with hereditary breast cancer: 12 were derived from tumors affecting women with hereditary BRCA1 mutation, the only one from a woman with BRCA2 mutation, while another eight cases had familial history of breast/ovarian cancer, but were negative for the BRCA1/2 mutations (so called BRCAx cases). Proportion of BRCA1 and BRCA2 mutated tumors was typical for the Polish population. Ten samples were derived from patients with apparently sporadic disease (no familial history of cancer) while 4 patients had a history of familial cancer aggregation (FCA) but without prevalence of breast/ovarian cancers. Thus, these samples were merged with sporadic samples in most of the analyses. All BRCA1 mutation-linked tumors in our study were negative for estrogen receptor (by immunohistochemistry, standard procedures for ER, PGR and HER2 staining were applied), while the only BRCA2-mutated tumor was ER-positive. There were 26 ductal and 5 medullary carcinomas within the study group, which is consistent with the distribution of histopathological types in BRCA1 mutation carriers. Patients were diagnosed at stage T1-2, N0-1 and M0. Caution: this submission contains the data from 6 microarrays done on the normal/pathologically unchanged breast tissue from breast cancer patiets. The data from normal tissues was not analyzed in the paper BRCA1-related gene signature in breast cancer is strongly influenced by ER status and molecular type by Lisowska et al., 2011, Front Biosci (Elite Ed). 2011 Jan 1;3:125-36

```
## ---- message=FALSE, warning=FALSE, include=FALSE-----------------------
library(Biobase)
library(oligoClasses)
library(knitr)
```

```
library(BiocStyle)
library(oligo)
library(geneplotter)
library(ggplot2)
library(LSD)
library(gplots)
library(RColorBrewer)
library(ArrayExpress)
library(arrayQualityMetrics)
library(stringr)
library(matrixStats)
library(topGO)
library(genefilter)
library(pd.hg.u133.plus.2)
library(hgu133plus2.db)
library(pheatmap)
library(mvtnorm)
library(DAAG)
library(multcomp)
library(limma)
library(ReactomePA)
library(clusterProfiler)
library(devtools)
library(biomaRt)
library(reshape2)
library(EnrichmentBrowser)
library(tidyr)
library(dplyr)
list.files("GSE50567/CEL")
```

```
##  [1] "GSM1223653_HBC_t01.CEL.gz"  "GSM1223654_HBC_t02.CEL.gz"
##  [3] "GSM1223655_HBC_t09A.CEL.gz" "GSM1223656_HBC_t09B.CEL.gz"
##  [5] "GSM1223657_HBC_t11.CEL.gz"  "GSM1223658_HBC_t12.CEL.gz"
##  [7] "GSM1223659_HBC_t14.CEL.gz"  "GSM1223660_HBC_t17.CEL.gz"
##  [9] "GSM1223661_HBC_t21.CEL.gz"  "GSM1223662_HBC_t26.CEL.gz"
## [11] "GSM1223663_HBC_t28.CEL.gz"  "GSM1223664_HBC_t33.CEL.gz"
## [13] "GSM1223665_HBC_t10.CEL.gz"  "GSM1223666_HBC_t04.CEL.gz"
## [15] "GSM1223667_HBC_t19.CEL.gz"  "GSM1223668_HBC_t24.CEL.gz"
## [17] "GSM1223669_HBC_t32.CEL.gz"  "GSM1223670_HBC_t35.CEL.gz"
## [19] "GSM1223671_HBC_t08.CEL.gz"  "GSM1223672_HBC_t36.CEL.gz"
## [21] "GSM1223673_HBC_t37.CEL.gz"  "GSM1223674_HBC_t06.CEL.gz"
## [23] "GSM1223675_HBC_t07.CEL.gz"  "GSM1223676_HBC_t13.CEL.gz"
## [25] "GSM1223677_HBC_t20.CEL.gz"  "GSM1223678_HBC_t22.CEL.gz"
## [27] "GSM1223679_HBC_t27.CEL.gz"  "GSM1223680_HBC_t29.CEL.gz"
## [29] "GSM1223681_HBC_t31.CEL.gz"  "GSM1223682_HBC_t34.CEL.gz"
## [31] "GSM1223683_HBC_t05.CEL.gz"  "GSM1223684_HBC_t15.CEL.gz"
## [33] "GSM1223685_HBC_t18.CEL.gz"  "GSM1223686_HBC_t25.CEL.gz"
## [35] "GSM1223687_HBC_t30.CEL.gz"  "GSM1223688_HBC_n02.CEL.gz"
## [37] "GSM1223689_HBC_n03.CEL.gz"  "GSM1223690_HBC_n07.CEL.gz"
## [39] "GSM1223691_HBC_n11.CEL.gz"  "GSM1223692_HBC_n12.CEL.gz"
## [41] "GSM1223693_HBC_n21.CEL.gz"
```

```
celfiles <- list.files("GSE50567/CEL", full = TRUE)
```

```
SDRF <- read.delim(
  url("https://www.ebi.ac.uk/arrayexpress/files/E-GEOD-50567/E-GEOD-50567.sdrf.txt"))
rownames(SDRF) <- SDRF$Array.Data.File
SDRF <- AnnotatedDataFrame(SDRF)

raw_data <- read.celfiles(celfiles, verbose = FALSE, phenoData = SDRF)
```

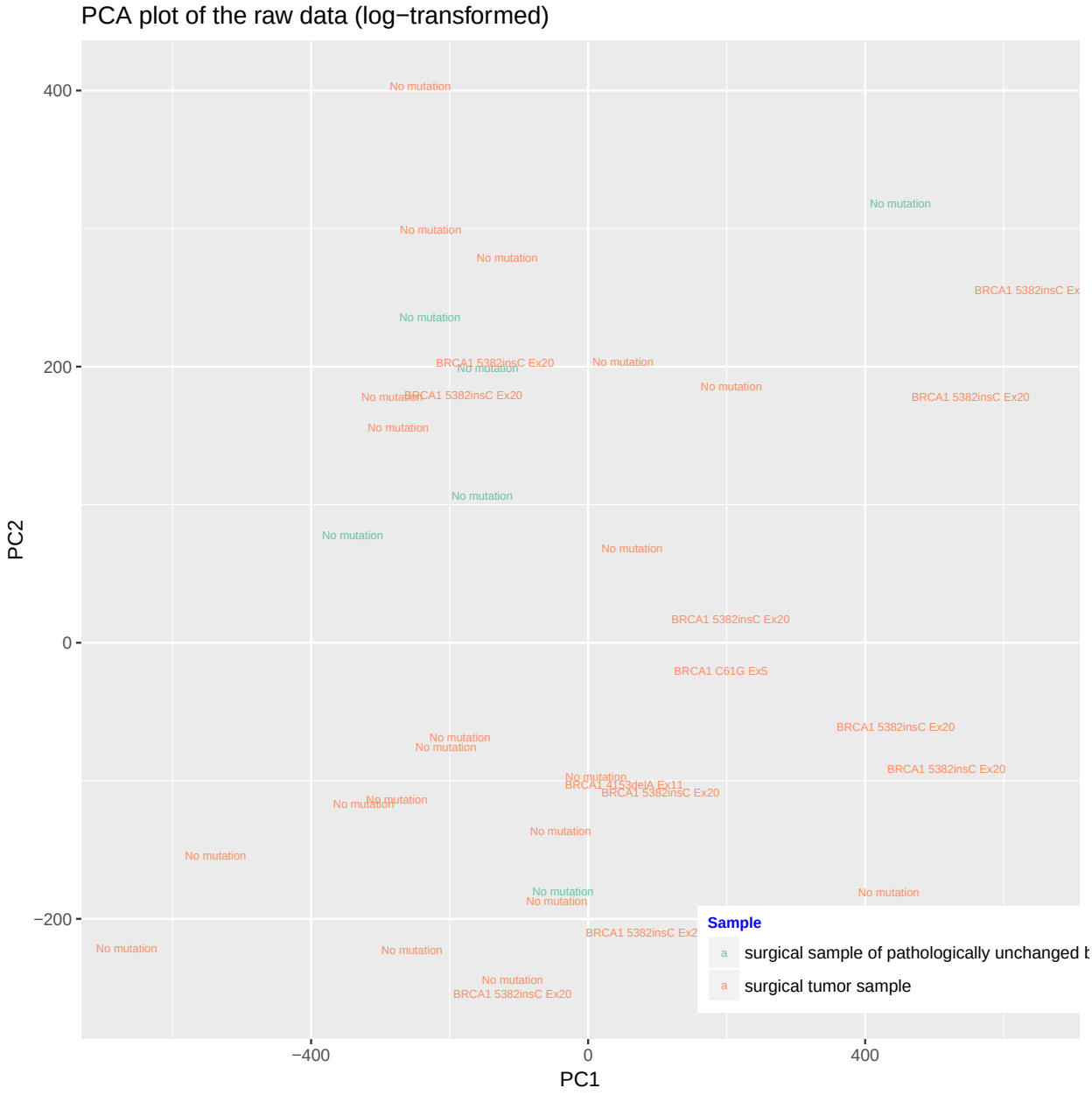
```
## Reading in : GSE50567/CEL/GSM1223653_HBC_t01.CEL.gz
## Reading in : GSE50567/CEL/GSM1223654_HBC_t02.CEL.gz
## Reading in : GSE50567/CEL/GSM1223655_HBC_t09A.CEL.gz
## Reading in : GSE50567/CEL/GSM1223656_HBC_t09B.CEL.gz
## Reading in : GSE50567/CEL/GSM1223657_HBC_t11.CEL.gz
## Reading in : GSE50567/CEL/GSM1223658_HBC_t12.CEL.gz
## Reading in : GSE50567/CEL/GSM1223659_HBC_t14.CEL.gz
## Reading in : GSE50567/CEL/GSM1223660_HBC_t17.CEL.gz
## Reading in : GSE50567/CEL/GSM1223661_HBC_t21.CEL.gz
## Reading in : GSE50567/CEL/GSM1223662_HBC_t26.CEL.gz
## Reading in : GSE50567/CEL/GSM1223663_HBC_t28.CEL.gz
## Reading in : GSE50567/CEL/GSM1223664_HBC_t33.CEL.gz
## Reading in : GSE50567/CEL/GSM1223665_HBC_t10.CEL.gz
## Reading in : GSE50567/CEL/GSM1223666_HBC_t04.CEL.gz
## Reading in : GSE50567/CEL/GSM1223667_HBC_t19.CEL.gz
## Reading in : GSE50567/CEL/GSM1223668_HBC_t24.CEL.gz
## Reading in : GSE50567/CEL/GSM1223669_HBC_t32.CEL.gz
## Reading in : GSE50567/CEL/GSM1223670_HBC_t35.CEL.gz
## Reading in : GSE50567/CEL/GSM1223671_HBC_t08.CEL.gz
## Reading in : GSE50567/CEL/GSM1223672_HBC_t36.CEL.gz
## Reading in : GSE50567/CEL/GSM1223673_HBC_t37.CEL.gz
## Reading in : GSE50567/CEL/GSM1223674_HBC_t06.CEL.gz
## Reading in : GSE50567/CEL/GSM1223675_HBC_t07.CEL.gz
## Reading in : GSE50567/CEL/GSM1223676_HBC_t13.CEL.gz
## Reading in : GSE50567/CEL/GSM1223677_HBC_t20.CEL.gz
## Reading in : GSE50567/CEL/GSM1223678_HBC_t22.CEL.gz
## Reading in : GSE50567/CEL/GSM1223679_HBC_t27.CEL.gz
## Reading in : GSE50567/CEL/GSM1223680_HBC_t29.CEL.gz
## Reading in : GSE50567/CEL/GSM1223681_HBC_t31.CEL.gz
## Reading in : GSE50567/CEL/GSM1223682_HBC_t34.CEL.gz
## Reading in : GSE50567/CEL/GSM1223683_HBC_t05.CEL.gz
## Reading in : GSE50567/CEL/GSM1223684_HBC_t15.CEL.gz
## Reading in : GSE50567/CEL/GSM1223685_HBC_t18.CEL.gz
## Reading in : GSE50567/CEL/GSM1223686_HBC_t25.CEL.gz
## Reading in : GSE50567/CEL/GSM1223687_HBC_t30.CEL.gz
## Reading in : GSE50567/CEL/GSM1223688_HBC_n02.CEL.gz
## Reading in : GSE50567/CEL/GSM1223689_HBC_n03.CEL.gz
## Reading in : GSE50567/CEL/GSM1223690_HBC_n07.CEL.gz
## Reading in : GSE50567/CEL/GSM1223691_HBC_n11.CEL.gz
## Reading in : GSE50567/CEL/GSM1223692_HBC_n12.CEL.gz
## Reading in : GSE50567/CEL/GSM1223693_HBC_n21.CEL.gz
```

```
pData(raw_data) <- pData(raw_data)[, c("Source.Name",
                                       "Characteristics..brca1.2.mutation.",
"Characteristics..clinical.sample.",                                        "FactorValue..ESTROGEN.RECEP
                                       "FactorValue..MOLECULAR.SUBTYPE.ACCORDING.TO.SORLIE.ET.AL...PNAS
```

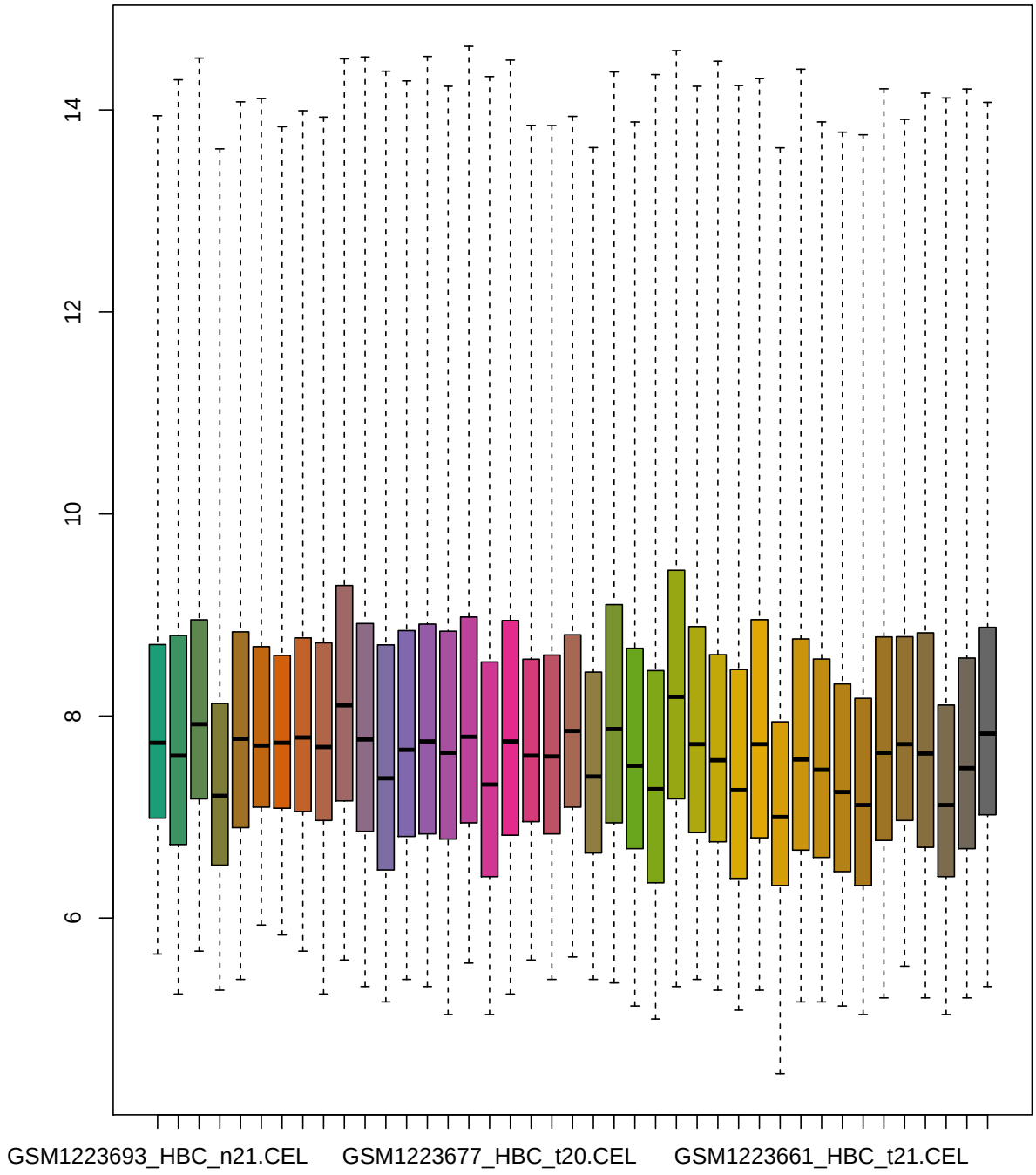
1.1 Quality control of the raw data

Here we check for outliers and try to see whether the data clusters as expected, by whether the sample came from a tumor sample. We use the identifiers of the individuals as plotting symbols.

```
exp_raw <- log2(exprs(raw_data))
PCA_raw <- prcomp(t(exp_raw), scale = FALSE)

dataGG <- data.frame(PC1 = PCA_raw$x[,1], PC2 = PCA_raw$x[,2],
                     Mutation = pData(raw_data)$Characteristics..brca1.2.mutation.,
                     Sample = pData(raw_data)$Characteristics..clinical.sample.,
                     Individual = pData(raw_data)$Source.Name)

p <- (qplot(PC1, PC2, data = dataGG, color =  Sample,
       main = "PCA plot of the raw data (log-transformed)", size = I(2),
       asp = 1.0, geom = "text",
       label = Mutation)
  + scale_colour_brewer(palette = "Set2"))

p + theme(legend.position = c(0.95, 0.08), legend.title = element_text(colour="blue", size=8,
                                    face="bold"))
```




```
boxplot(raw_data, target = "core",
        main = "Boxplots of log2-intensities for the raw data")
```



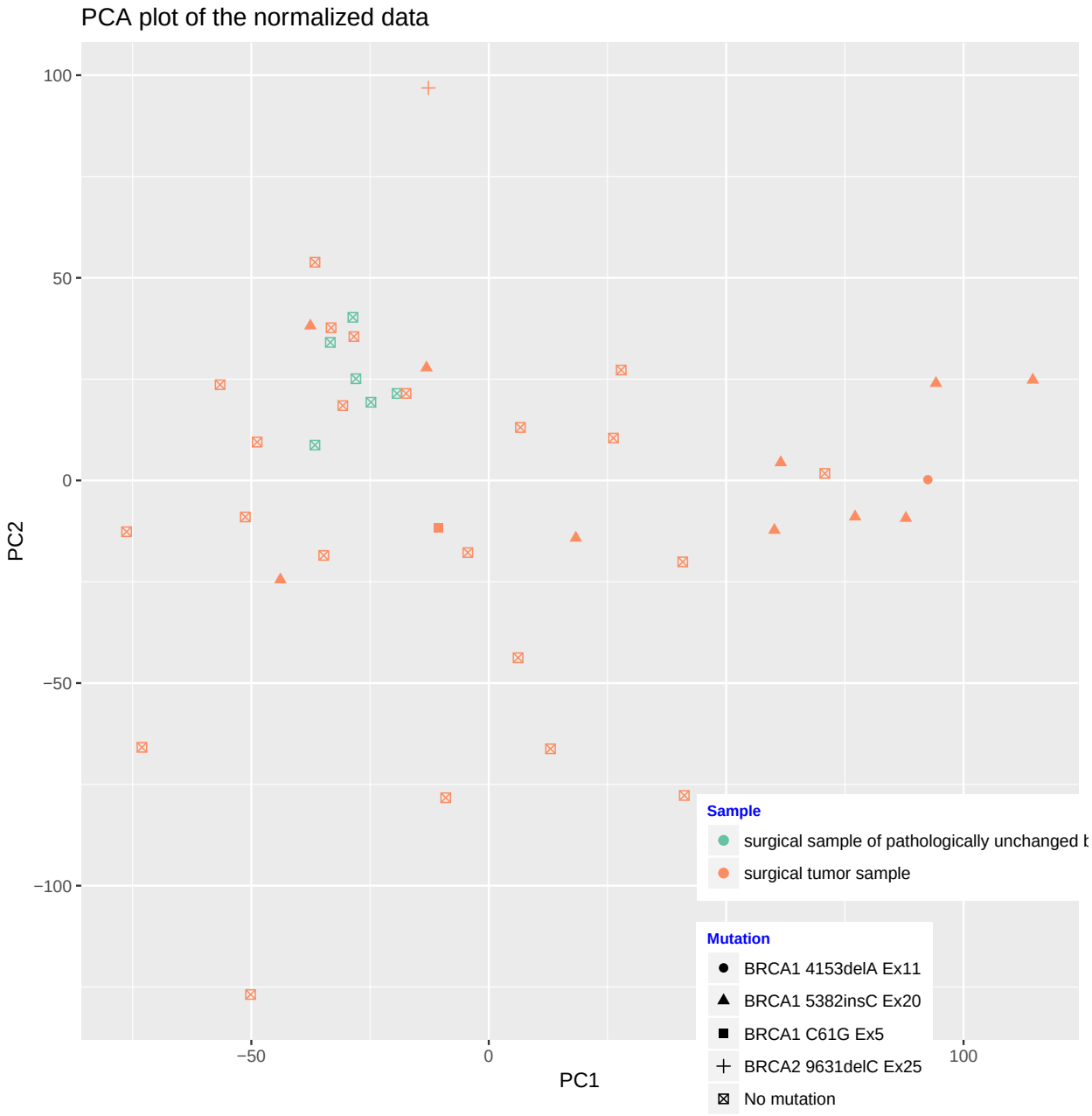
The PCA (performed on the log-intensity scale) plot of the raw data shows that the first principal component does not differentiates between the tissues nor wether it is a BRCA mutated sample. And the intensity boxplots show that the intensity distributions of the individual arrays are quite different, indicating the need of an appropriate normalization.

1.2 Normalization and Quality assessment of the calibrated data

```
BRCA_eset <- oligo::rma(raw_data)
```

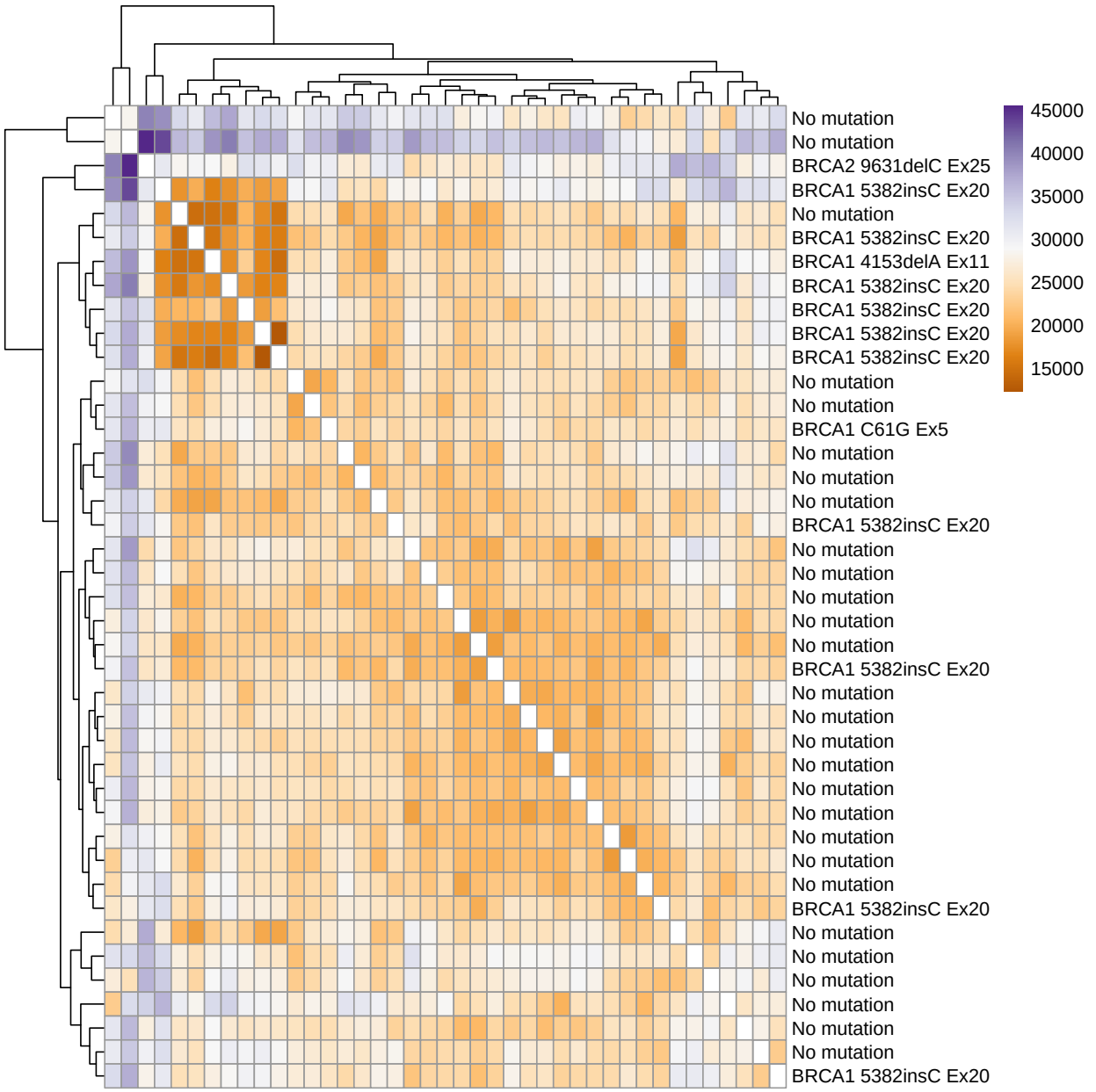
```
## Background correcting
## Normalizing
## Calculating Expression
```

```
exp_BRCA <- exprs(BRCA_eset)
PCA <- prcomp(t(exp_BRCA), scale = FALSE)

dataGG <- data.frame(PC1 = PCA$x[,1], PC2 = PCA$x[,2],
                     Sample = pData(BRCA_eset)$Characteristics..clinical.sample.,
                     Mutation = pData(BRCA_eset)$Characteristics..brca1.2.mutation.)

p <- (qplot(PC1, PC2, data = dataGG, color =  Sample, shape =  Mutation,
       main = "PCA plot of the normalized data", size = I(2), asp = 1.0)
  + scale_colour_brewer(palette = "Set2"))

p + theme(legend.position = c(0.95, 0.08), legend.title = element_text(colour="blue", size=8,
                                      face="bold"))
```




```
dists <- as.matrix(dist(t(exp_BRCA), method = "manhattan"))
colnames(dists) <- NULL
diag(dists) <- NA
rownames(dists) <-  pData(BRCA_eset)$Characteristics..brca1.2.mutation.
```

```
hmcol <- colorRampPalette(rev(brewer.pal(9, "PuOr")))(255)

pheatmap(dists, col = rev(hmcol), clustering_distance_rows = "manhattan",
         clustering_distance_cols = "manhattan")
```



The normalized data's PCA plot shows that after normalization, we have separated tumor tissue from non-tumor tissue. However, the heatmap of the sample-to-sample distances roughly separates BRCA mutation and non-mutation samples and we can also see that the samples do not cluster strongly by whether the sample is a BRCA mutation, confirming the impression from the PCA plot that the separation between the mutation and non-mutations samples is not perfect.

1.3 Subset for only BRCA1 mutation vs No mutation in Tumor Samples

```
raw_data <- raw_data[, raw_data$Characteristics..clinical.sample. == "surgical tumor sample"]
raw_data <- raw_data[, !grepl("BRCA2", raw_data$Characteristics..brca1.2.mutation.)]
BRCA_eset <- oligo::rma(raw_data)
```

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

```
exp_BRCA <- exprs(BRCA_eset)
```

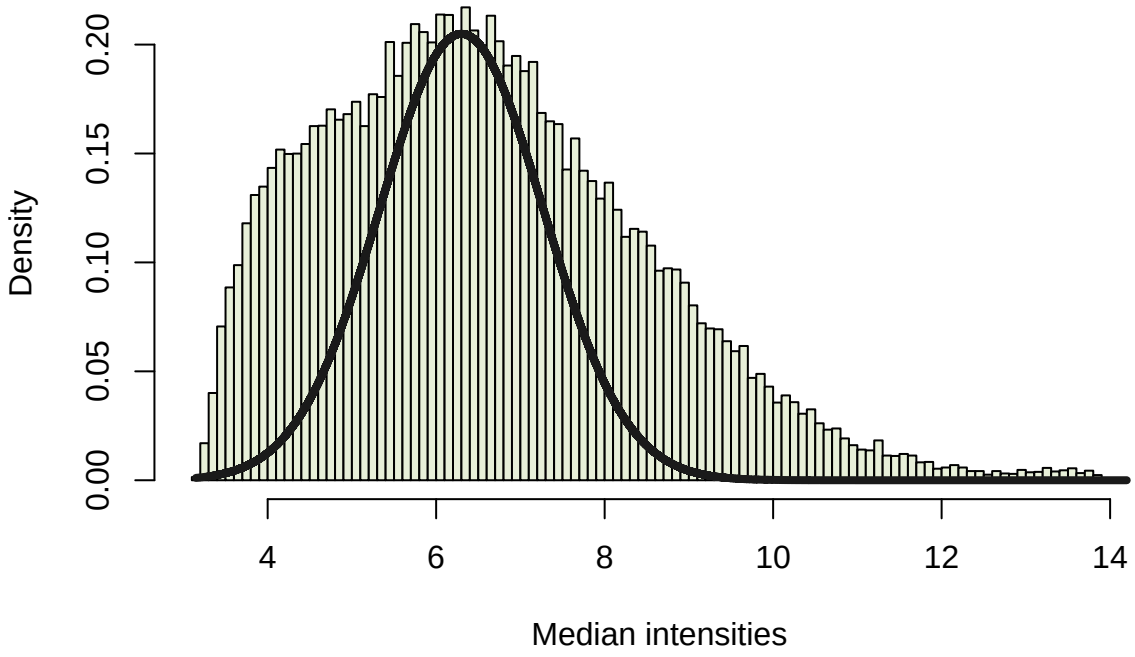
1.4 Filtering based on intensity

We now filter out lowly expressed genes. In the following histogram of the gene-wise medians, we can clearly see an enrichment of low medians on the left hand side. These represent the genes we want to filter. We will use the 5% quantile of this distribution as a threshold, then keep only those genes that show an expression higher than the threshold in at least as many arrays as in the smallest experimental group. First we check how many samples in each experimental group.

```
Mutation <- str_replace_all(pData(BRCA_eset)$Characteristics..brca1.2.mutation., " ", "_")
Mutation <- ifelse(Mutation == "No_mutation", "No_mutation", "Mutation")

no_of_samples <- table(paste0(
  pData(BRCA_eset)$FactorValue..ESTROGEN.RECEPTOR.STATUS.BY.IMMUNOHISTOCHEMISTRY., "_",
  Mutation))
no_of_samples
```

```
##
##    ER(-)_Mutation ER(-)_No_mutation ER(+)_No_mutation
##                12                17                 5
```

```
## ---- message=FALSE, warning=FALSE----------------------------------------
BRCA_medians <- rowMedians(exprs(BRCA_eset))
hist_res <- hist(BRCA_medians, 100, col="#e7efd8", freq = FALSE,
                 main = "Histogram of the median intensities",
                 xlab = "Median intensities")
emp_mu <- hist_res$breaks[which.max(hist_res$density)]
emp_sd <- mad(BRCA_medians)/2
prop_cental <- 0.50
lines(sort(BRCA_medians), prop_cental*dnorm(sort(BRCA_medians),
                                            mean = emp_mu , sd = emp_sd),
      col = "grey10", lwd = 4)
```




```
cut_val <- 0.05 / prop_cental
thresh_median <- qnorm(0.05 / prop_cental, emp_mu, emp_sd)
samples_cutoff <- min(no_of_samples)
idx_thresh_median <- apply(exprs(BRCA_eset), 1, function(x){
  sum(x > thresh_median) >= samples_cutoff})
table(idx_thresh_median)
```

```
## idx_thresh_median
## FALSE  TRUE
##  9757 44918
```

```
BRCA_filtered <- subset(BRCA_eset, idx_thresh_median)

## ---- message=FALSE, warning=FALSE----------------------------------------
anno_BRCA  <- AnnotationDbi::select(hgu133plus2.db,
                                    keys=(featureNames(BRCA_filtered)),
                                    columns = c("SYMBOL", "GENENAME"),
                                    keytype="PROBEID")
#Removing multiple mapping
probe_stats <- anno_BRCA   %>%
  group_by(PROBEID) %>%
  summarize(no_of_matches = n_distinct(SYMBOL)) %>%
  filter(no_of_matches > 1)
#probe_stats
#dim(probe_stats)
ids_to_exlude <- ((featureNames(BRCA_filtered) %in% probe_stats$PROBEID) |
                    featureNames(BRCA_filtered)  %in% subset(anno_BRCA ,is.na(SYMBOL))$PROBEID)
```

```
#table(ids_to_exlude)
BRCA_final <- subset(BRCA_filtered, !ids_to_exlude)
#validObject(BRCA_final)
fData(BRCA_final)$PROBEID <- rownames(fData(BRCA_final))
fData(BRCA_final) <- left_join(fData(BRCA_final), anno_BRCA)
fData(BRCA_final) <- fData(BRCA_final) %>% distinct(fData(BRCA_final)$PROBEID, .keep_all = T)
rownames(fData(BRCA_final)) <-fData(BRCA_final)$PROBEID
#validObject(BRCA_final)

## ----Using Limma to find DE genes--------------------------------------
Individual <- as.character(pData(BRCA_final)$Source.Name)
Individual <- gsub(" 1$", "", Individual)
Individual[]
```

```
##  [1] "GSM1223687" "GSM1223686" "GSM1223685" "GSM1223684" "GSM1223683"
##  [6] "GSM1223682" "GSM1223681" "GSM1223680" "GSM1223679" "GSM1223678"
## [11] "GSM1223677" "GSM1223676" "GSM1223675" "GSM1223674" "GSM1223673"
## [16] "GSM1223672" "GSM1223671" "GSM1223670" "GSM1223669" "GSM1223668"
## [21] "GSM1223667" "GSM1223666" "GSM1223664" "GSM1223663" "GSM1223662"
## [26] "GSM1223661" "GSM1223660" "GSM1223659" "GSM1223658" "GSM1223657"
## [31] "GSM1223656" "GSM1223655" "GSM1223654" "GSM1223653"
```

```
Mutation <- str_replace_all(pData(BRCA_final)$Characteristics..brca1.2.mutation., " ", "_")
Mutation <- ifelse(Mutation == "No_mutation", "No_mutation", "Mutation")
Mutation <- as.factor(Mutation)


ER <- str_replace_all(pData(BRCA_final)$FactorValue..ESTROGEN.RECEPTOR.STATUS.BY.IMMUNOHISTOCHEMISTRY.,


design_BRCA_tumor <- model.matrix(~0 + Mutation)
colnames(design_BRCA_tumor) <- c("Mutation", "No_mutation")
fit_BRCA_tumor <- lmFit(BRCA_final, design_BRCA_tumor)
contrast.matrix_tumor <- makeContrasts(Mutation_NoMutation = Mutation - No_mutation, levels=design_BRCA
contrast.matrix_tumor
```

```
##              Contrasts
## Levels        Mutation_NoMutation
##   Mutation                     1
##   No_mutation                 -1
```

```
Mutation_NoMutation_fits_tumor <- contrasts.fit(fit_BRCA_tumor, contrast.matrix_tumor)
Mutation_NoMutation_ebFit_tumor <- eBayes(Mutation_NoMutation_fits_tumor)
options(digits=2)
topTable(Mutation_NoMutation_ebFit_tumor, coef=1,  n=20, adjust="BH")
```

```
##                 PROBEID     SYMBOL
## 234312_s_at 234312_s_at      ACSS2
## 235635_at     235635_at    ARHGAP5
## 233914_s_at 233914_s_at       SBF2
## 212653_s_at 212653_s_at      EHBP1
## 209220_at     209220_at       GPC3
## 218913_s_at 218913_s_at       GMIP
## 227148_at     227148_at    PLEKHH2
## 205280_at     205280_at       GLRB
```

```
## 227526_at     227526_at       CDON
## 242358_at     242358_at RASSF8-AS1
## 227126_at     227126_at      PTPRG
## 225207_at     225207_at       PDK4
## 218872_at     218872_at       TESC
## 209311_at     209311_at     BCL2L2
## 204944_at     204944_at      PTPRG
## 218189_s_at 218189_s_at       NANS
## 225946_at     225946_at     RASSF8
## 203688_at     203688_at       PKD2
## 232555_at     232555_at      CREB5
## 37549_g_at   37549_g_at       BBS9
##                                                          GENENAME
## 234312_s_at           acyl-CoA synthetase short-chain family member 2
## 235635_at                            Rho GTPase activating protein 5
## 233914_s_at                                SET binding factor 2
## 212653_s_at                            EH domain binding protein 1
## 209220_at                                              glypican 3
## 218913_s_at                                GEM interacting protein
## 227148_at    pleckstrin homology, MyTH4 and FERM domain containing H2
## 205280_at                                  glycine receptor beta
## 227526_at                cell adhesion associated, oncogene regulated
## 242358_at                                 RASSF8 antisense RNA 1
## 227126_at               protein tyrosine phosphatase, receptor type G
## 225207_at                            pyruvate dehydrogenase kinase 4
## 218872_at                                                tescalcin
## 209311_at                                              BCL2 like 2
## 204944_at               protein tyrosine phosphatase, receptor type G
## 218189_s_at                          N-acetylneuraminate synthase
## 225946_at                     Ras association domain family member 8
## 203688_at   polycystin 2, transient receptor potential cation channel
## 232555_at                   cAMP responsive element binding protein 5
## 37549_g_at                               Bardet-Biedl syndrome 9
##             fData.BRCA_final..PROBEID logFC AveExpr    t P.Value adj.P.Val
## 234312_s_at               234312_s_at  0.75     9.1  6.0 8.6e-07     0.031
## 235635_at                 235635_at  1.20     6.3  5.5 3.1e-06     0.044
## 233914_s_at               233914_s_at  1.06     6.9  5.5 3.7e-06     0.044
## 212653_s_at               212653_s_at  1.09     8.9  5.4 5.2e-06     0.047
## 209220_at                 209220_at  1.91     7.1  5.3 7.5e-06     0.050
## 218913_s_at               218913_s_at -0.56     8.2 -5.2 9.6e-06     0.050
## 227148_at                 227148_at  1.40     5.1  5.1 1.1e-05     0.050
## 205280_at                 205280_at  1.03     5.0  5.1 1.3e-05     0.050
## 227526_at                 227526_at  1.19     7.4  5.1 1.3e-05     0.050
## 242358_at                 242358_at  0.82     4.7  5.1 1.4e-05     0.050
## 227126_at                 227126_at  1.48     7.7  5.0 1.6e-05     0.052
## 225207_at                 225207_at  2.63     7.6  4.9 1.9e-05     0.056
## 218872_at                 218872_at  1.16     7.4  4.9 2.0e-05     0.056
## 209311_at                 209311_at  0.78     8.5  4.9 2.3e-05     0.056
## 204944_at                 204944_at  0.88     6.8  4.9 2.3e-05     0.056
## 218189_s_at               218189_s_at -0.79     9.7 -4.9 2.6e-05     0.058
## 225946_at                 225946_at  1.38     6.6  4.8 2.9e-05     0.062
## 203688_at                 203688_at  1.37     7.7  4.7 3.6e-05     0.067
## 232555_at                 232555_at  0.88     7.5  4.7 3.9e-05     0.067
## 37549_g_at                37549_g_at  0.45     7.0  4.7 4.0e-05     0.067
```

```
##               B
## 234312_s_at 5.4
## 235635_at   4.3
## 233914_s_at 4.1
## 212653_s_at 3.8
## 209220_at   3.5
## 218913_s_at 3.3
## 227148_at   3.2
## 205280_at   3.1
## 227526_at   3.0
## 242358_at   3.0
## 227126_at   2.9
## 225207_at   2.7
## 218872_at   2.7
## 209311_at   2.5
## 204944_at   2.5
## 218189_s_at 2.5
## 225946_at   2.3
## 203688_at   2.1
## 232555_at   2.1
## 37549_g_at  2.1
```

```
table_tumor <-  topTable(Mutation_NoMutation_ebFit_tumor, number = Inf)
head(table_tumor)
```

```
##                 PROBEID  SYMBOL
## 234312_s_at 234312_s_at   ACSS2
## 235635_at     235635_at ARHGAP5
## 233914_s_at 233914_s_at    SBF2
## 212653_s_at 212653_s_at   EHBP1
## 209220_at     209220_at    GPC3
## 218913_s_at 218913_s_at    GMIP
##                                                GENENAME
## 234312_s_at acyl-CoA synthetase short-chain family member 2
## 235635_at                   Rho GTPase activating protein 5
## 233914_s_at                            SET binding factor 2
## 212653_s_at                     EH domain binding protein 1
## 209220_at                                      glypican 3
## 218913_s_at                         GEM interacting protein
##             fData.BRCA_final..PROBEID logFC AveExpr    t P.Value adj.P.Val
## 234312_s_at               234312_s_at  0.75     9.1  6.0 8.6e-07     0.031
## 235635_at                   235635_at  1.20     6.3  5.5 3.1e-06     0.044
## 233914_s_at               233914_s_at  1.06     6.9  5.5 3.7e-06     0.044
## 212653_s_at               212653_s_at  1.09     8.9  5.4 5.2e-06     0.047
## 209220_at                   209220_at  1.91     7.1  5.3 7.5e-06     0.050
## 218913_s_at               218913_s_at -0.56     8.2 -5.2 9.6e-06     0.050
##               B
## 234312_s_at 5.4
## 235635_at   4.3
## 233914_s_at 4.1
## 212653_s_at 3.8
## 209220_at   3.5
## 218913_s_at 3.3
```

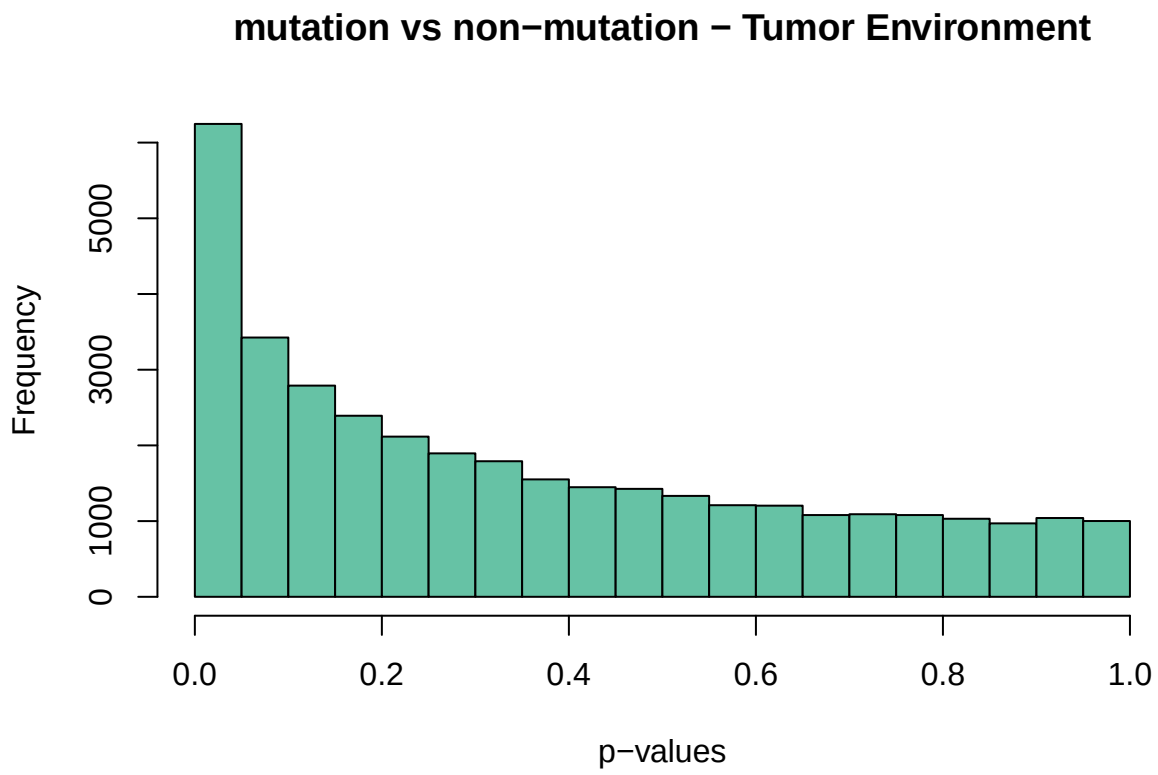
```r
table(table_tumor$adj.P.Val < 0.1)
```

```
##
## FALSE  TRUE
## 35384   738
```

```r
table(table_tumor$P.Value < 0.001)
```

```
##
## FALSE  TRUE
## 35708   414
```

```r
hist(table_tumor$P.Value, col = brewer.pal(3, name = "Set2")[1],
     main = "mutation vs non-mutation - Tumor Environment", xlab = "p-values")
```




```r
## ----Gene ontology (GO) based enrichment analysis-----------------------------------------------
DE_genes_tumor <- subset(table_tumor, adj.P.Val < 0.1)$PROBEID
back_genes_idx <- genefinder(BRCA_final, as.character(DE_genes_tumor),
                             method="manhattan", scale="none")
back_genes_idx <- sapply(back_genes_idx, function(x)x$indices)
back_genes <-featureNames(BRCA_final)[back_genes_idx]
back_genes <- setdiff(back_genes, DE_genes_tumor)
intersect(back_genes, DE_genes_tumor)
```

```
## character(0)
```

```r
length(back_genes)
```

```
## [1] 6666
```

```
#multidensity(list(
#  tumor=  table_tumor[,"AveExpr"] ,
#  fore= table_tumor[DE_genes_tumor , "AveExpr"],
#  back= table_tumor[rownames(table_tumor) %in% back_genes, "AveExpr"]),
#  col = c("#e46981", "#ae7ee2", "#a7ad4a"),
#  xlab="mean expression",
#  main = "DE genes for tumor - background - matching")

gene_IDs <- rownames(table_tumor)
in_universe <- gene_IDs %in% c(DE_genes_tumor,  back_genes)
inSelection <-  gene_IDs %in% DE_genes_tumor
tumor_genes <- factor(as.integer(inSelection[in_universe]))
names(tumor_genes) <- gene_IDs[in_universe]

top_GO_data <- new("topGOdata", ontology = "BP", allGenes = tumor_genes,
                   nodeSize = 10, annot=annFUN.db, affyLib = "hgu133plus2.db")

result_top_GO_elim <- runTest(top_GO_data, algorithm = "elim", statistic = "Fisher")
result_top_GO_classic <- runTest(top_GO_data, algorithm = "classic", statistic = "Fisher")


res_top_GO <- GenTable(top_GO_data, Fisher.elim = result_top_GO_elim,
                       Fisher.classic = result_top_GO_classic,
                       orderBy = "Fisher.elim" , topNodes = 100)
genes_top_GO <- printGenes(top_GO_data, whichTerms = res_top_GO$GO.ID,
                           chip = "hgu133plus2.db", geneCutOff = 1000)
res_top_GO$sig_genes <- sapply(genes_top_GO, function(x){
  str_c(paste0(x[x$'raw p-value' == 2, "Symbol.id"],";"), collapse = "")
})
```

2 Gene ontology (GO) based enrichment analysis

Using FDR under 10% for function analysis

Top 20 GO enrichment terms

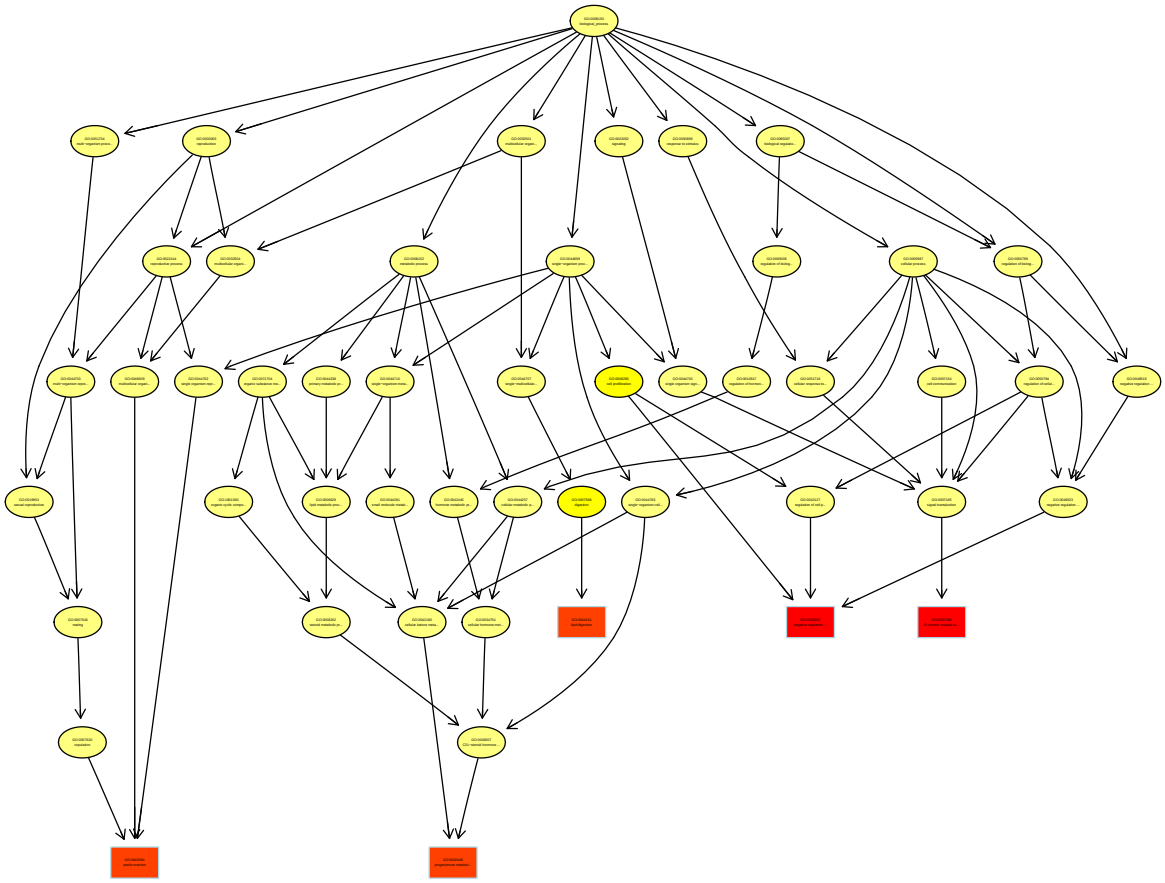
```
head(res_top_GO[,1:8], 20)
```

```
##         GO.ID                                          Term Annotated
## 1  GO:0008285 negative regulation of cell proliferatio...       352
## 2  GO:0007186 G-protein coupled receptor signaling pat...       275
## 3  GO:0044241                             lipid digestion        11
## 4  GO:0043084                             penile erection        11
## 5  GO:0042448              progesterone metabolic process        11
## 6  GO:0034383 low-density lipoprotein particle clearan...        12
## 7  GO:0045923 positive regulation of fatty acid metabo...        12
## 8  GO:0008360                    regulation of cell shape       102
## 9  GO:1900119 positive regulation of execution phase o...        10
## 10 GO:0010737                  protein kinase A signaling        10
## 11 GO:0035024 negative regulation of Rho protein signa...        18
## 12 GO:0042304 regulation of fatty acid biosynthetic pr...        18
## 13 GO:0034113              heterotypic cell-cell adhesion        33
```

```
## 14 GO:0043268 positive regulation of potassium ion tra...        19
## 15 GO:0035815 positive regulation of renal sodium excr...        11
## 16 GO:0042311                                 vasodilation        11
## 17 GO:0070542                     response to fatty acid        42
## 18 GO:0072203 cell proliferation involved in metanephr...        15
## 19 GO:0019934                    cGMP-mediated signaling        15
## 20 GO:0050892                      intestinal absorption        15
##    Significant Expected Rank in Fisher.classic Fisher.elim Fisher.classic
## 1           63     36.3                     14     6.0e-06        6.0e-06
## 2           51     28.4                     23     1.8e-05        1.8e-05
## 3            7      1.1                     28     2.7e-05        2.7e-05
## 4            7      1.1                     29     2.7e-05        2.7e-05
## 5            7      1.1                     30     2.7e-05        2.7e-05
## 6            7      1.2                     40     6.0e-05        6.0e-05
## 7            7      1.2                     41     6.0e-05        6.0e-05
## 8           24     10.5                     45     7.8e-05        7.8e-05
## 9            6      1.0                     62     0.00017        0.00017
## 10           6      1.0                     63     0.00017        0.00017
## 11           8      1.9                     67     0.00021        0.00021
## 12           8      1.9                     68     0.00021        0.00021
## 13          11      3.4                     78     0.00030        0.00030
## 14           8      2.0                     81     0.00033        0.00033
## 15           6      1.1                     85     0.00034        0.00034
## 16           6      1.1                     86     0.00034        0.00034
## 17          17      4.3                      1     0.00037        2.9e-07
## 18           7      1.6                     89     0.00037        0.00037
## 19           7      1.6                     90     0.00037        0.00037
## 20           7      1.6                     91     0.00037        0.00037
```

A graphical representation of the topGO results.

```
showSigOfNodes(top_GO_data, score(result_top_GO_elim), firstSigNodes = 5,
               useInfo = 'def')
```



```
## $dag
## A graphNEL graph with directed edges
## Number of Nodes = 50
## Number of Edges = 87
##
## $complete.dag
## [1] "A graph with 50 nodes."
```

3 A pathway enrichment analysis using Reactome

Using FDR under 10% for pathway enrichment analysis

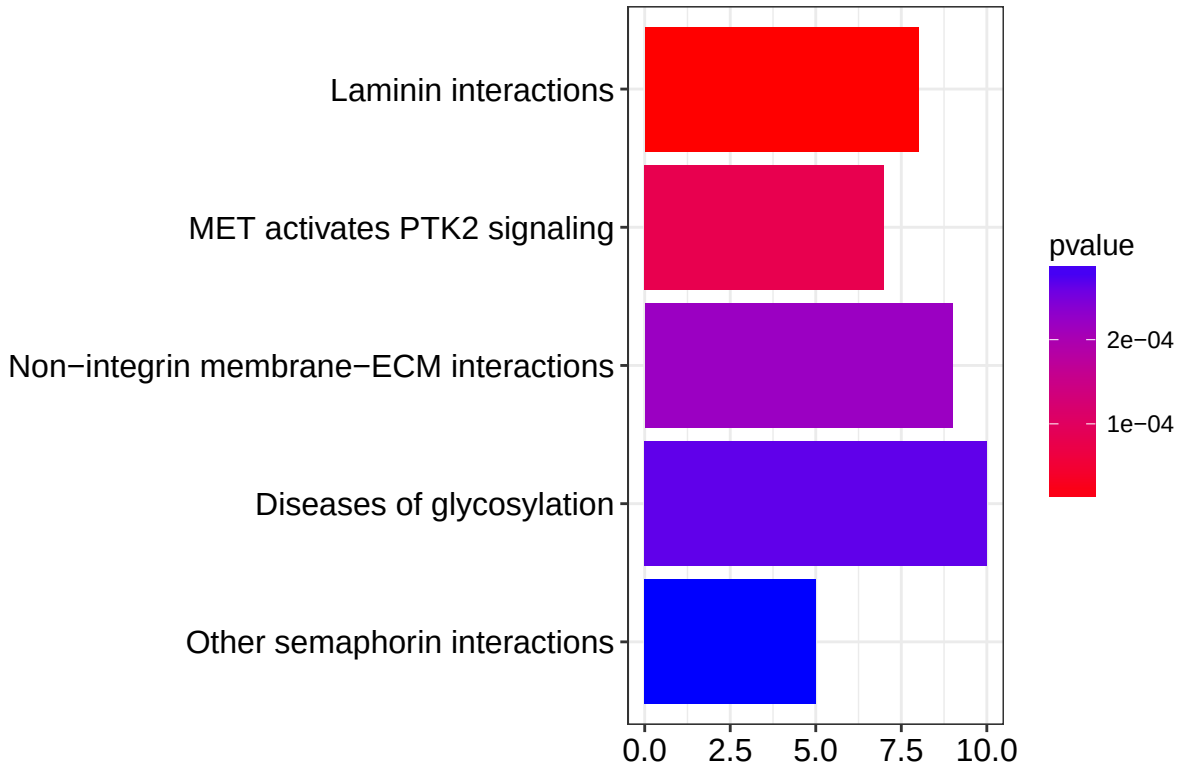
Enriched Reactome pathways and their p-values as a bar chart. - The top pathways can be displayed as a bar char that displays all categories with a p-value below the specified cutoff.

```
entrez_ids <- mapIds(hgu133plus2.db,
                     keys = rownames(table_tumor),
                     keytype="PROBEID",
                     column = "ENTREZID")

genelist = entrez_ids[DE_genes_tumor]
```

```
names(genelist) <- NULL
reactome_enrich <- enrichPathway(gene = genelist, organism = "human",
                                 pvalueCutoff = 0.2,
                                 readable = TRUE)

head(summary(reactome_enrich))[1:6]
```

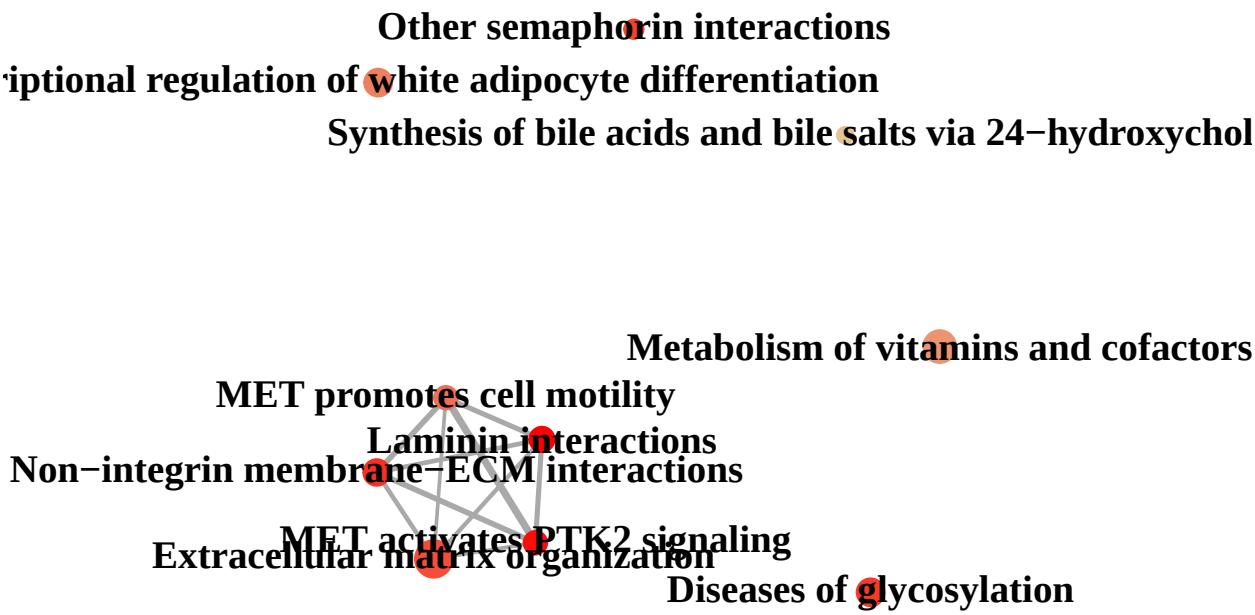
```
##                            ID                          Description
## R-HSA-3000157 R-HSA-3000157                   Laminin interactions
## R-HSA-8874081 R-HSA-8874081           MET activates PTK2 signaling
## R-HSA-3000171 R-HSA-3000171 Non-integrin membrane-ECM interactions
## R-HSA-3781865 R-HSA-3781865              Diseases of glycosylation
## R-HSA-416700   R-HSA-416700          Other semaphorin interactions
## R-HSA-1474244 R-HSA-1474244      Extracellular matrix organization
##                GeneRatio   BgRatio  pvalue p.adjust
## R-HSA-3000157     8/337  33/10281 8.3e-06   0.0073
## R-HSA-8874081     7/337  33/10281 7.8e-05   0.0346
## R-HSA-3000171     9/337  64/10281 2.2e-04   0.0515
## R-HSA-3781865    10/337  80/10281 2.7e-04   0.0515
## R-HSA-416700      5/337  19/10281 2.9e-04   0.0515
## R-HSA-1474244    22/337 300/10281 3.5e-04   0.0515
```

```
barplot(reactome_enrich)
```



Enriched Reactome pathways enrichment results as a graph. - The "enrichment map" displays the results of the enrichment analysis as a graph, where the color represents the p-value of the pathway and the edge-thickness is proportional to the number of overlapping genes between two pathways.

```
enrichMap(reactome_enrich, n = 10, vertex.label.font = 2)
```



4 A pathway enrichment analysis using GSEA

Enriched GSEA pathways enrichment results as a graph.

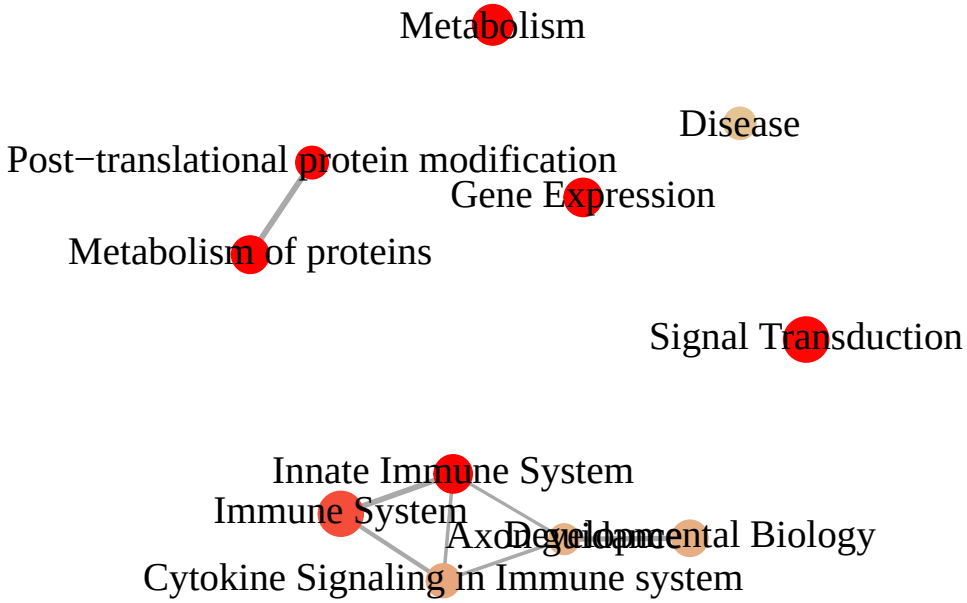
```
entrez_ids <- mapIds(hgu133plus2.db,
                     keys = rownames(table_tumor),
                     keytype="PROBEID",
                     column = "ENTREZID")
entrez_ids <- as.data.frame(entrez_ids)
de_tumor <- merge(entrez_ids, table_tumor, by=0)
de_tumor <- as.data.frame(de_tumor)
gene_avg_exp <- de_tumor %>%
  filter(adj.P.Val < 0.2) %>%
  group_by(entrez_ids) %>%
  summarise_all(funs(mean))%>%
  dplyr::select(entrez_ids,AveExpr) %>%
  arrange(desc(AveExpr))

gene_avg_exp$entrez_ids <- as.character(gene_avg_exp$entrez_ids)
gene_avg_exp2 <- structure(gene_avg_exp$AveExpr, names= gene_avg_exp$entrez_ids)
```

```
y <- gsePathway(gene_avg_exp2, nPerm=1000,
                minGSSize=120, pvalueCutoff=0.2,
                pAdjustMethod="BH", verbose=FALSE)
res <- as.data.frame(y)
res
```

```
##                           ID                            Description
## R-HSA-168249   R-HSA-168249                   Innate Immune System
## R-HSA-74160     R-HSA-74160                        Gene Expression
## R-HSA-392499   R-HSA-392499                 Metabolism of proteins
## R-HSA-1430728 R-HSA-1430728                             Metabolism
## R-HSA-597592   R-HSA-597592 Post-translational protein modification
## R-HSA-162582   R-HSA-162582                  Signal Transduction
## R-HSA-168256   R-HSA-168256                        Immune System
## R-HSA-1280215 R-HSA-1280215     Cytokine Signaling in Immune system
## R-HSA-1266738 R-HSA-1266738                 Developmental Biology
## R-HSA-422475   R-HSA-422475                       Axon guidance
## R-HSA-1643685 R-HSA-1643685                             Disease
##               setSize enrichmentScore NES pvalue p.adjust qvalues rank
## R-HSA-168249      235            0.19 1.8  0.001   0.0050  0.0028    1
## R-HSA-74160       234            0.21 1.9  0.001   0.0050  0.0028    1
## R-HSA-392499      221            0.25 2.3  0.001   0.0050  0.0028    3
## R-HSA-1430728     283            0.19 1.9  0.002   0.0061  0.0034    1
## R-HSA-597592      146            0.23 2.0  0.002   0.0061  0.0034    1
## R-HSA-162582      415            0.16 1.7  0.004   0.0100  0.0056    2
## R-HSA-168256      408            0.15 1.5  0.026   0.0557  0.0312    1
## R-HSA-1280215     163            0.17 1.5  0.055   0.0875  0.0491    6
## R-HSA-1266738     198            0.16 1.5  0.057   0.0875  0.0491   49
## R-HSA-422475      125            0.18 1.5  0.059   0.0875  0.0491   55
## R-HSA-1643685     142            0.17 1.5  0.064   0.0875  0.0491    6
##                              leading_edge
## R-HSA-168249  tags=0%, list=0%, signal=0%
## R-HSA-74160   tags=0%, list=0%, signal=0%
## R-HSA-392499  tags=0%, list=0%, signal=0%
## R-HSA-1430728 tags=0%, list=0%, signal=0%
## R-HSA-597592  tags=1%, list=0%, signal=1%
## R-HSA-162582  tags=0%, list=0%, signal=0%
## R-HSA-168256  tags=0%, list=0%, signal=0%
## R-HSA-1280215 tags=1%, list=0%, signal=1%
## R-HSA-1266738 tags=5%, list=2%, signal=5%
## R-HSA-422475  tags=6%, list=2%, signal=7%
## R-HSA-1643685 tags=1%, list=0%, signal=1%
##                                        core_enrichment
## R-HSA-168249
## R-HSA-74160
## R-HSA-392499
## R-HSA-1430728
## R-HSA-597592
## R-HSA-162582
## R-HSA-168256
## R-HSA-1280215
## R-HSA-1266738 5063/2252/2494/152831/53616/3672/6332/9365
## R-HSA-422475        6328/5063/2252/152831/3672/6332/9365
```

```
## R-HSA-1643685
```

```
enrichMap(y)
```



5 Conclusion:

5.1 Function Analysis

Top GO terms are:
1. negative regulation of cell proliferation
2. G-protein coupled receptor signaling pathway
3. lipid digestion
4. penile erection
5. progesterone metabolic process
6. low-density lipoprotein particle clearance
7. positive regulation of fatty acid metabolic process
8. regulation of cell shape
9. positive regulation of execution phase of apoptosis
10. protein kinase A signaling

5.2 Pathway enrichment analysis using Reactome

Top Affected Pathways:
1. Laminin interactions
2. MET activates PTK2 signaling
3. Non-integrin membrane-ECM interactions
4. Diseases of glycosylation
5. Other semaphorin interactions
6. Extracellular matrix organization

5.3 Pathway enrichment analysis using GSEA

Top Affected Pathways:
1. Metabolism of proteins
2. Metabolism
3. Gene Expression
4. Post-translational protein modification
5. Innate Immune System
6. Signal Transduction
7. Immune System
8. Developmental Biology
9. Cytokine Signaling in Immune system
10. Disease