

## limma\_GSE29001

```
# assign samples to groups and set up design matrix
gs <- factor(sml)
groups <- make.names(c("normal","cancer"))
levels(gs) <- groups
gset$group <- gs
design <- model.matrix(~group + 0, gset)
colnames(design) <- levels(gs)

gset <- gset[complete.cases(exprs(gset)), ] # skip missing values

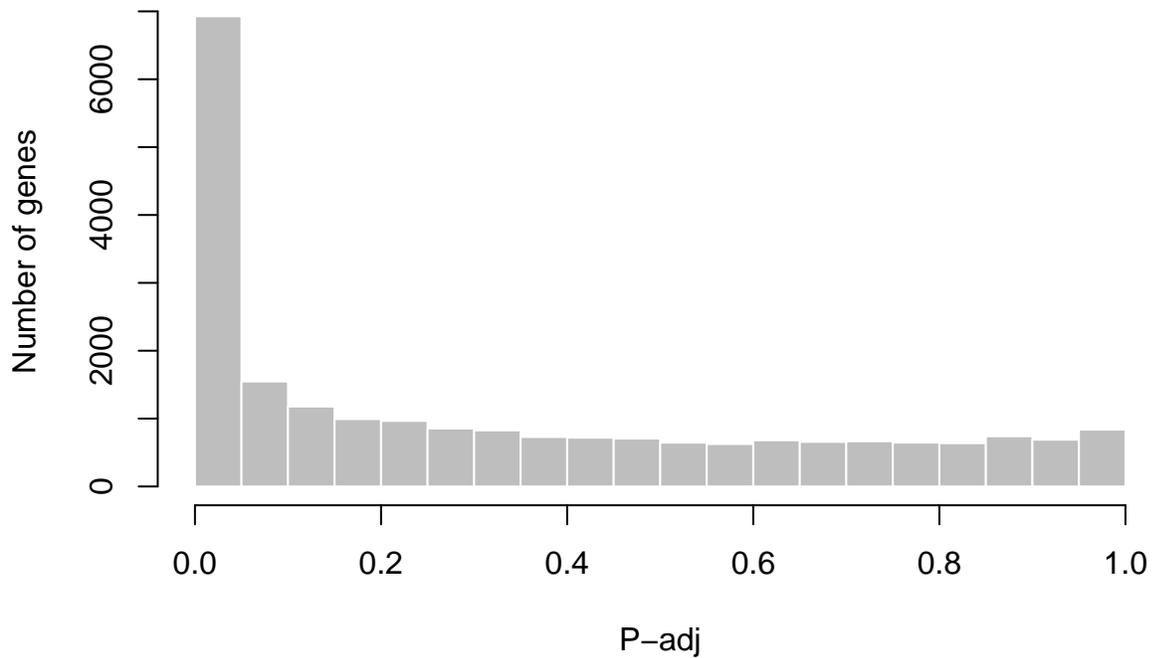
fit <- lmFit(gset, design) # fit linear model

# set up contrasts of interest and recalculate model coefficients
cts <- c(paste(groups[1],"-",groups[2],sep=""))
cont.matrix <- makeContrasts(contrasts=cts, levels=design)
fit2 <- contrasts.fit(fit, cont.matrix)

# compute statistics and table of top significant genes
fit2 <- eBayes(fit2, 0.01)
tT <- topTable(fit2, adjust="BH", sort.by="B", number=250)
# here an error could occur but we can delete the unnecessary columns and continue:
tT <- subset(tT, select=c("ID","adj.P.Val","P.Value","t","B","logFC","Gene.symbol","Gene.title"))
#write.table(tT, file=stdout(), row.names=F, sep="\t")

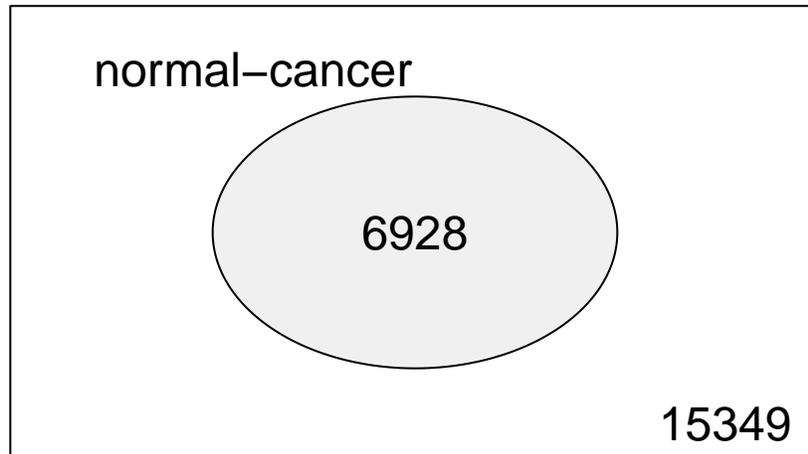
# Visualize and quality control test results.
# Build histogram of P-values for all genes. Normal test
# assumption is that most genes are not differentially expressed.
tT2 <- topTable(fit2, adjust="BH", sort.by="B", number=Inf)
hist(tT2$adj.P.Val, col = "grey", border = "white", xlab = "P-adj",
     ylab = "Number of genes", main = "P-adj value distribution")
```

## P-adj value distribution



The above plot shows the adjusted p-value distribution across the number of genes. The p-value for this experiment was adjusted using Benjamini Hochberg method(BH). Genes falling to the left of the significance threshold of 0.05 are considered statistically significant and may have potential biological relevance.

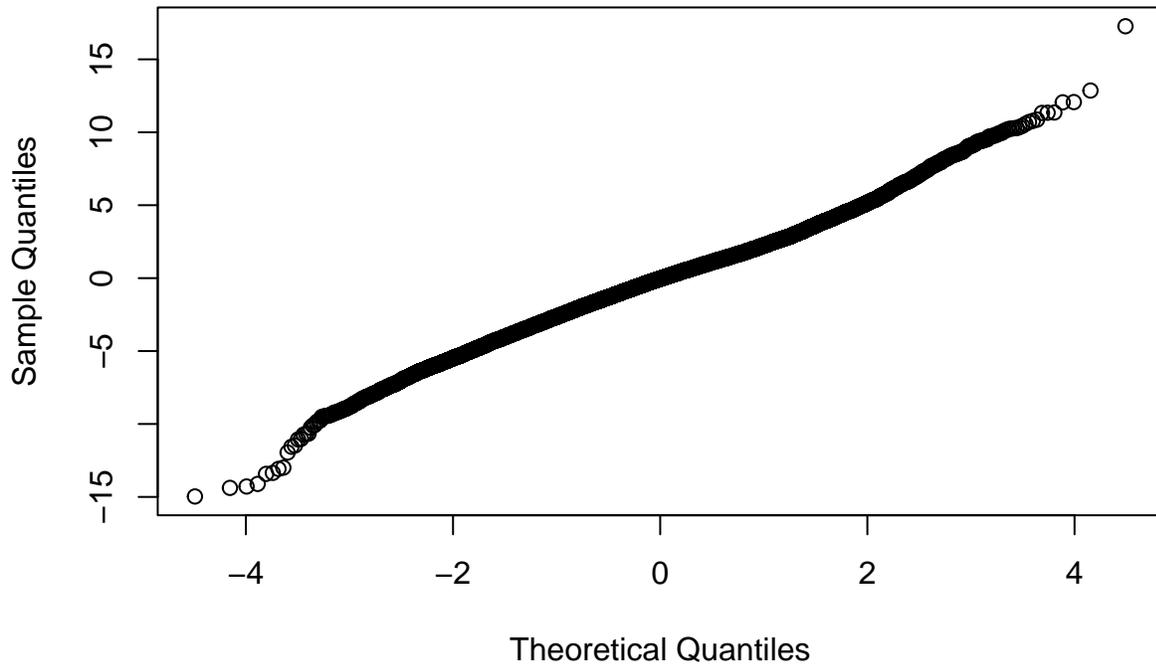
```
# summarize test results as "up", "down" or "not expressed"  
dT <- decideTests(fit2, adjust.method="BH", p.value=0.05, lfc=0.263)  
  
# Venn diagram of results  
vennDiagram(dT, circle.col=palette())
```



The Venn diagram visualizes how many genes fall into each category (“up,” “down,” and “not expressed”), and if there are any overlapping genes between these categories. In the plot above we can see that , there were 6928 significant up and down regulated genes. others 15349 were not expressed.

```
# create Q-Q plot for t-statistic  
t.good <- which(!is.na(fit2$F)) # filter out bad probes  
qqt(fit2$t[t.good], fit2$df.total[t.good], main="Moderated t statistic")
```

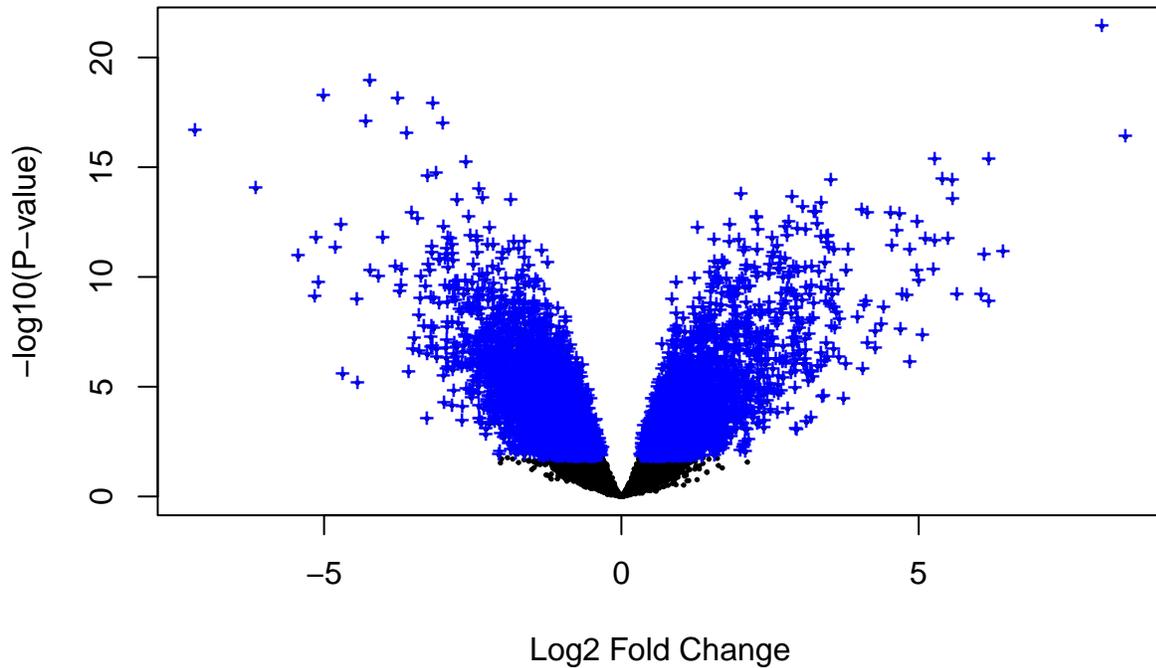
## Moderated t statistic



Q-Q plots (Quantile-Quantile plots) are useful for visually assessing whether the observed data follows an expected theoretical distribution, such as the normal distribution. In the plot above we can observe that the data points fall approximately along a straight line which suggests that the t-statistics are approximately normally distributed.

```
# volcano plot (log P-value vs log fold change)  
#colnames(fit2) # list contrast names  
ct <- 1 # choose contrast of interest  
volcanoplot(fit2, coef=ct, main=colnames(fit2)[ct], pch=20,  
            highlight=length(which(dT[,ct]!=0)), names=rep('+', nrow(fit2)))
```

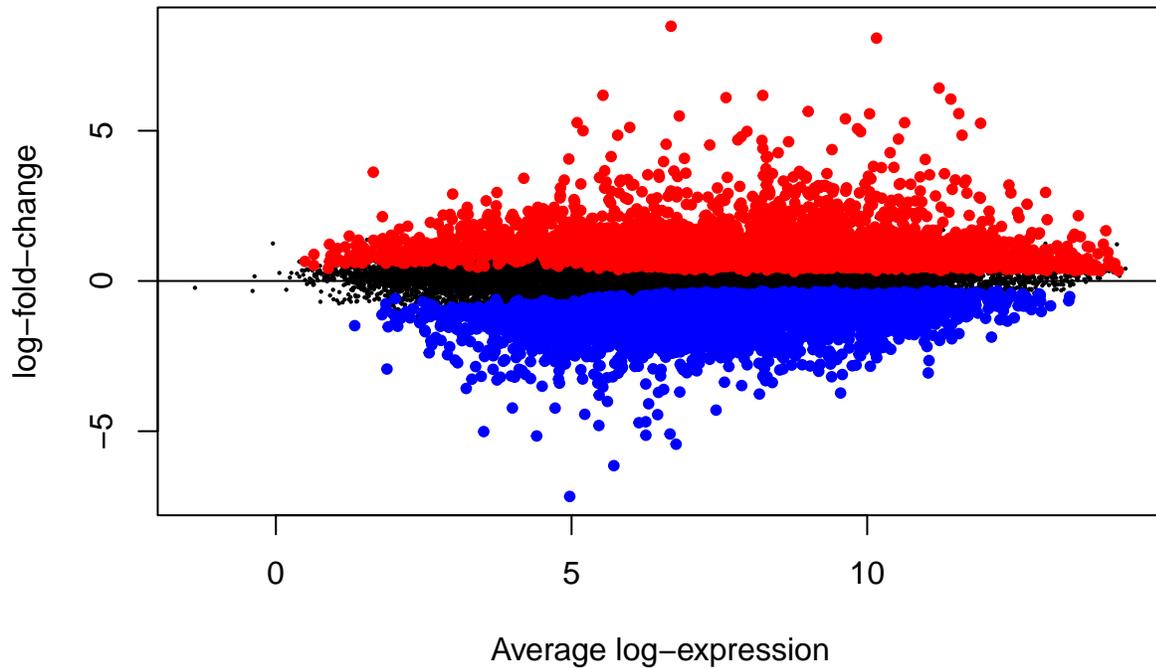
## normal-cancer



Volcano plots are commonly used in genomics to identify differentially expressed genes based on their statistical significance and magnitude of change. The resulting volcano plot has the log-fold change (x-axis) i.e. 0.263 plotted against the negative logarithm of the adjusted p-value (y-axis) i.e. 0.05. Genes with a significant change in expression (based on adjusted p-values) appear as points far away from the center along the y-axis, while genes with a substantial fold change appear farther away from the center along the x-axis.

```
# MD plot (log fold change vs mean log expression)  
# highlight statistically significant (p-adj < 0.05) probes  
plotMD(fit2, column=ct, status=dT[,ct], legend=F, pch=20, cex=1)  
abline(h=0)
```

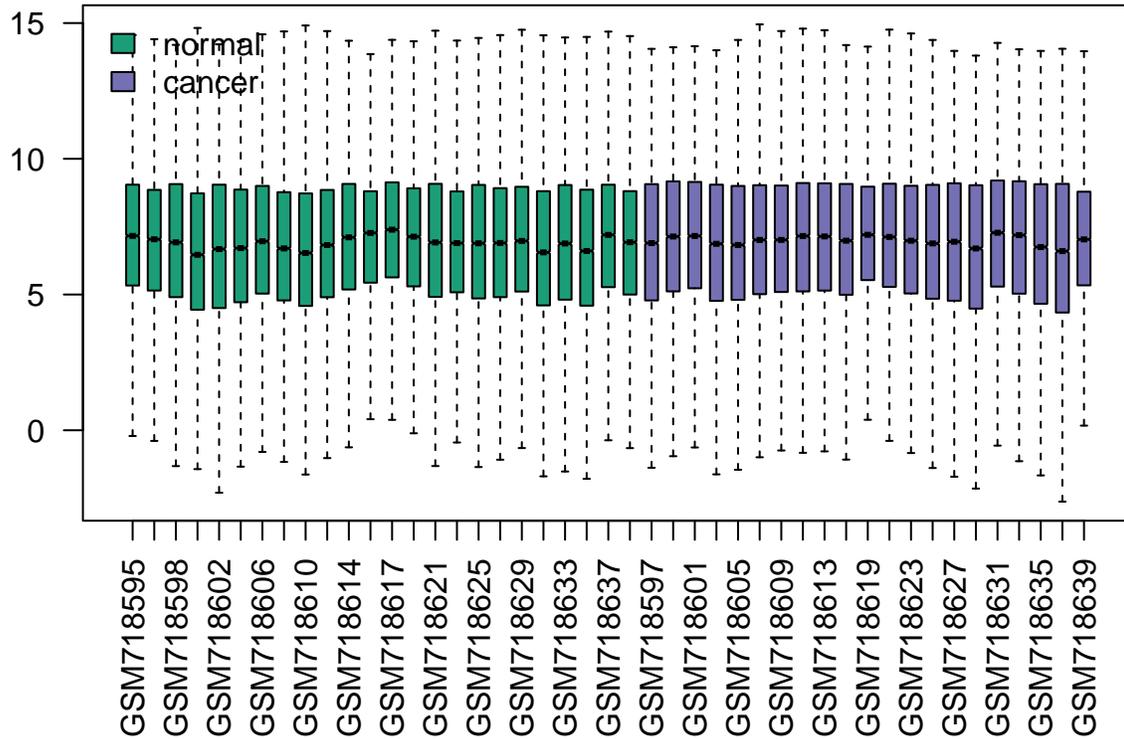
## normal-cancer



MD plots are commonly used to assess the magnitude and direction of gene expression changes between groups. Each point on the plot represents a gene, and the position of the point indicates the log-fold change and the mean log expression level for that gene.

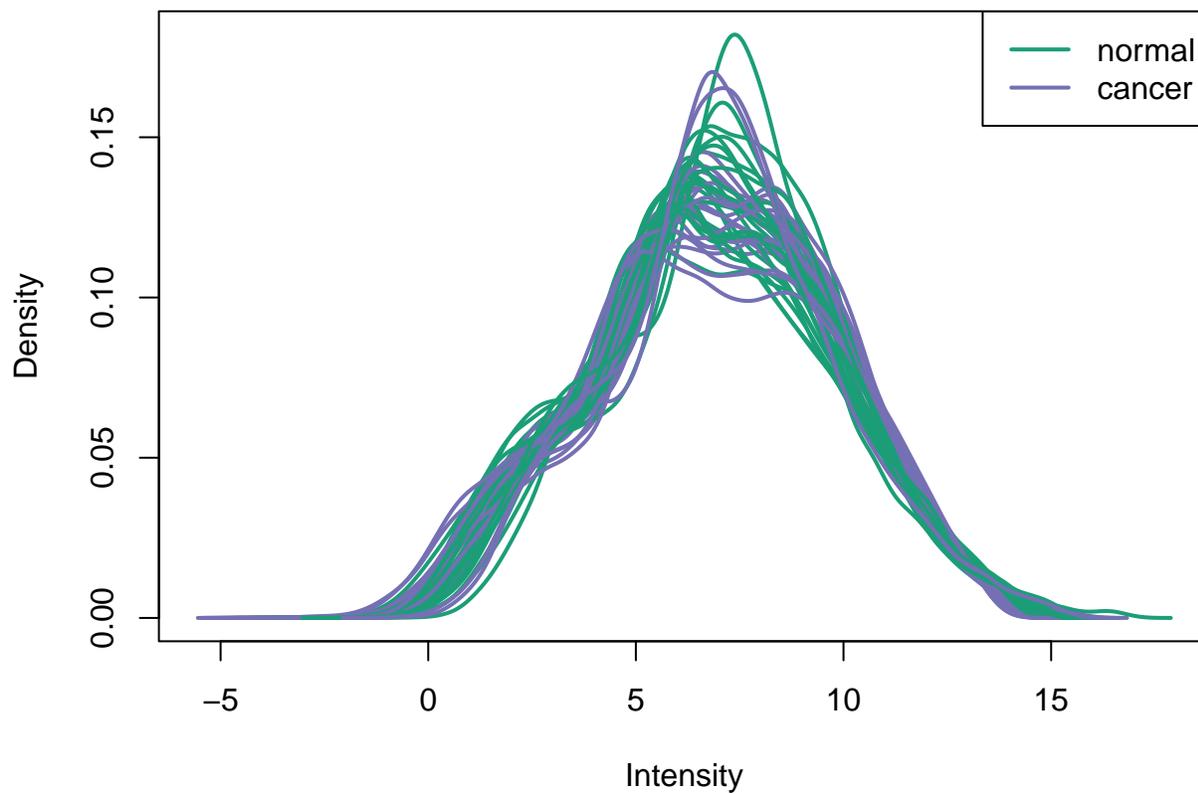
```
#####  
# General expression data analysis  
ex <- exprs(gset)  
  
# box-and-whisker plot  
#dev.new(width=3+ncol(gset)/6, height=5)  
ord <- order(gs) # order samples by group  
palette(c("#1B9E77", "#7570B3", "#E7298A", "#E6AB02", "#D95F02",  
          "#66A61E", "#A6761D", "#B32424", "#B324B3", "#666666"))  
par(mar=c(7,4,2,1))  
title <- paste ("GSE29001", "/", annotation(gset), sep = "")  
boxplot(ex[,ord], boxwex=0.6, notch=T, main=title, outline=FALSE, las=2, col=gs[ord])  
legend("topleft", groups, fill=palette(), bty="n")
```

## GSE29001/GPL571



```
# expression value distribution  
par(mar=c(4,4,2,1))  
title <- paste ("GSE29001", "/", annotation(gset), " value distribution", sep = "")  
plotDensities(ex, group=gs, main=title, legend ="topright")
```

## GSE29001/GPL571 value distribution



```
# UMAP plot (dimensionality reduction)
ex <- na.omit(ex) # eliminate rows with NAs
ex <- ex[!duplicated(ex), ] # remove duplicates
ump <- umap(t(ex), n_neighbors = 15, random_state = 123)
par(mar=c(3,3,2,6), xpd=TRUE)
plot(ump$layout, main="UMAP plot, nbrs=15", xlab="", ylab="", col=gs, pch=20, cex=1.5)
legend("topright", inset=c(-0.15,0), legend=levels(gs), pch=20,
      col=1:nlevels(gs), title="Group", pt.cex=1.5)
library("mapprotools") # point labels without overlaps
```

```
## Loading required package: sp
```

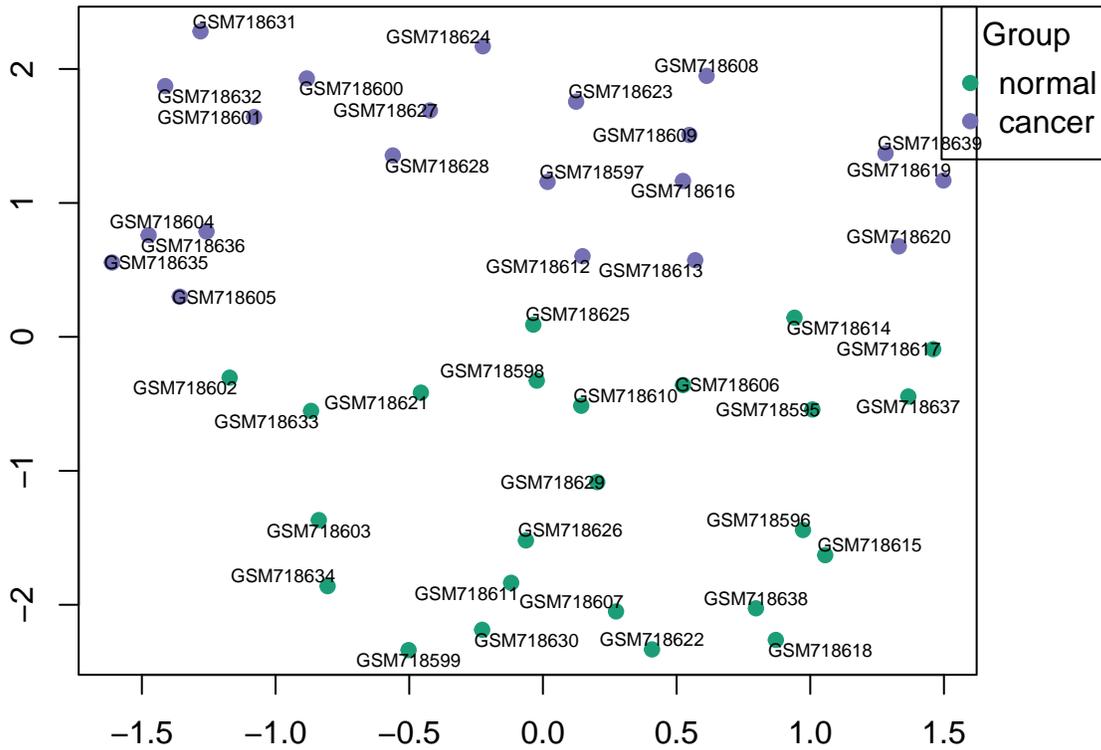
```
## The legacy packages mapprotools, rgdal, and rgeos, underpinning the sp package,
## which was just loaded, will retire in October 2023.
## Please refer to R-spatial evolution reports for details, especially
## https://r-spatial.org/r/2023/05/15/evolution4.html.
## It may be desirable to make the sf package available;
## package maintainers should consider adding sf to Suggests:.
## The sp package is now running under evolution status 2
## (status 2 uses the sf package in place of rgdal)
```

```
## Please note that 'mapprotools' will be retired during October 2023,
## plan transition at your earliest convenience (see
## https://r-spatial.org/r/2023/05/15/evolution4.html and earlier blogs
## for guidance); some functionality will be moved to 'sp'.
## Checking rgeos availability: FALSE
```

```
pointLabel(ump$layout, labels = rownames(ump$layout), method="SANN", cex=0.6)
```

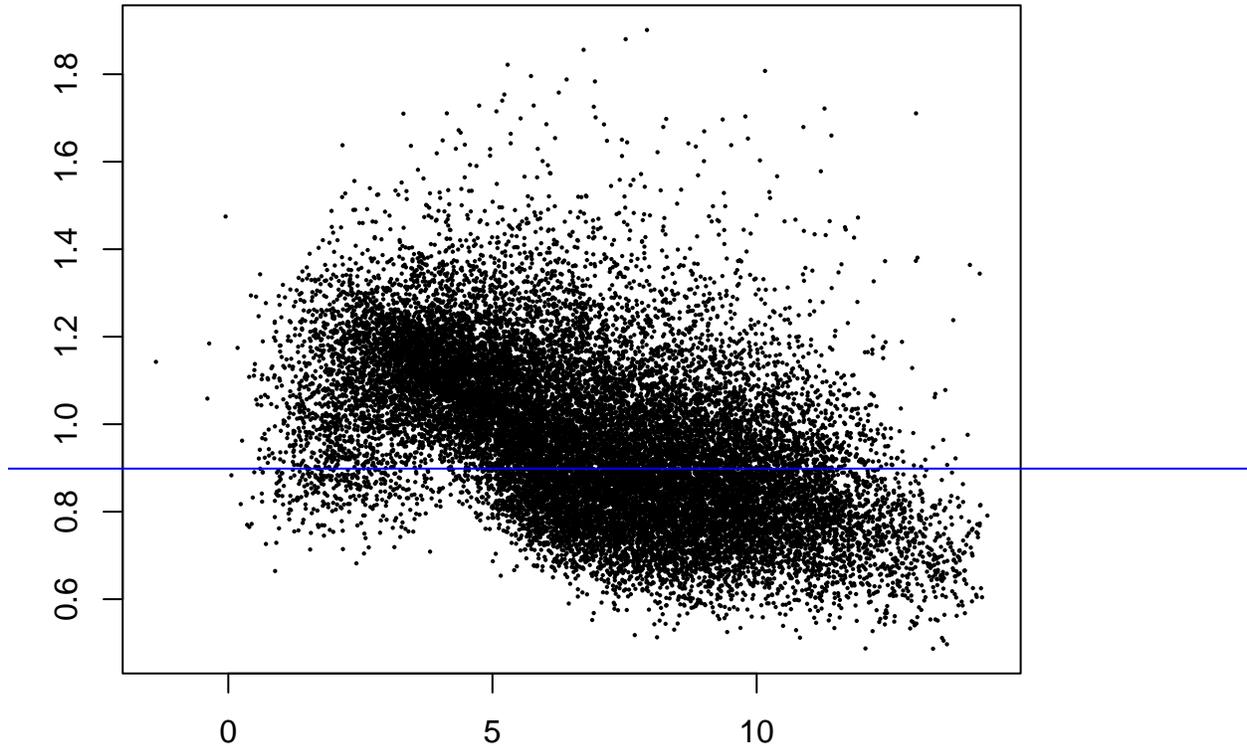
```
## Warning: Function moved to the car package because mapproj is retiring in 2023
```

### UMAP plot, nbrs=15



```
# mean-variance trend, helps to see if precision weights are needed  
plotSA(fit2, main="Mean variance trend, GSE29001")
```

## Mean variance trend, GSE29001



```
#dev.off()
```

1. Box-and-Whisker Plot: depicts the distribution of gene expression values across distinct sample groups (here, normal and cancer). The boxes show the interquartile range (IQR), while the centre line inside the box reflects the median expression value. The whiskers extend from the margins of the boxes and represent the variability of the data. Outliers are points that are not within the whiskers. The figure allows us to examine the expression distributions of the two groups and discover any variations in their central tendencies and spread.
2. Expression Value Distribution Plot: shows the density of gene expression levels for each sample group. The plot shows how gene expression values are spread among each group. It enables us to determine if the distributions of the normal and cancer groups are similar or dissimilar. Denser patches imply higher levels of gene expression, whereas sparser regions indicate lower levels of expression.
3. UMAP Plot (Dimensionality Reduction): displays a reduced-dimensional depiction of gene expression data using the UMAP method. The UMAP method uses dimensionality reduction to project high-dimensional gene expression data onto a 2D space. Each point on the plot represents a sample, and the colors indicate whether the sample is normal or cancer. The map allows us to see the separation or grouping of samples depending on their gene expression patterns.
4. Mean-Variance Trend Plot: depicts the connection between mean expression and variance in gene expression data. The plot shows us the variation of gene expression varies with the mean expression level.