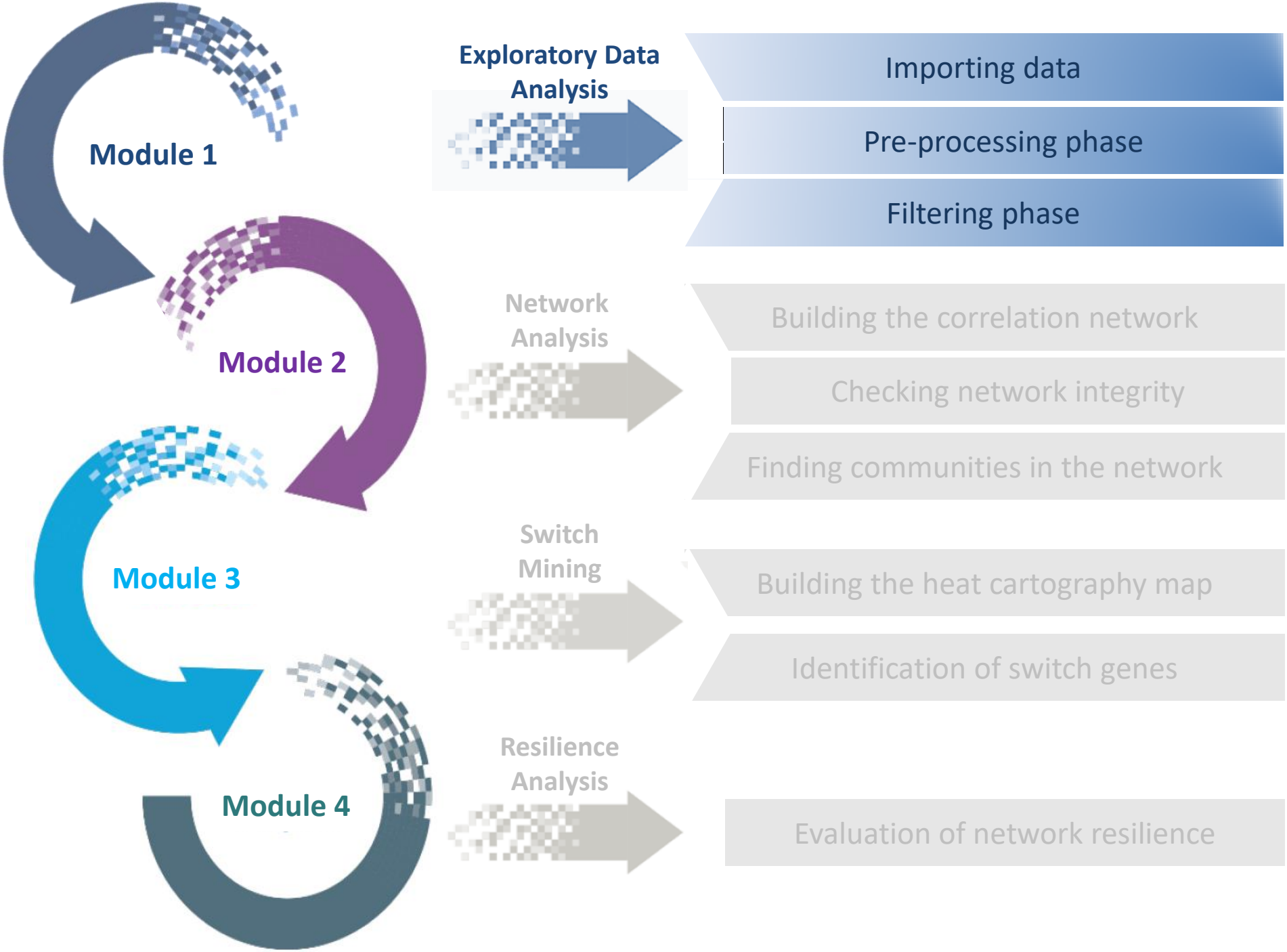
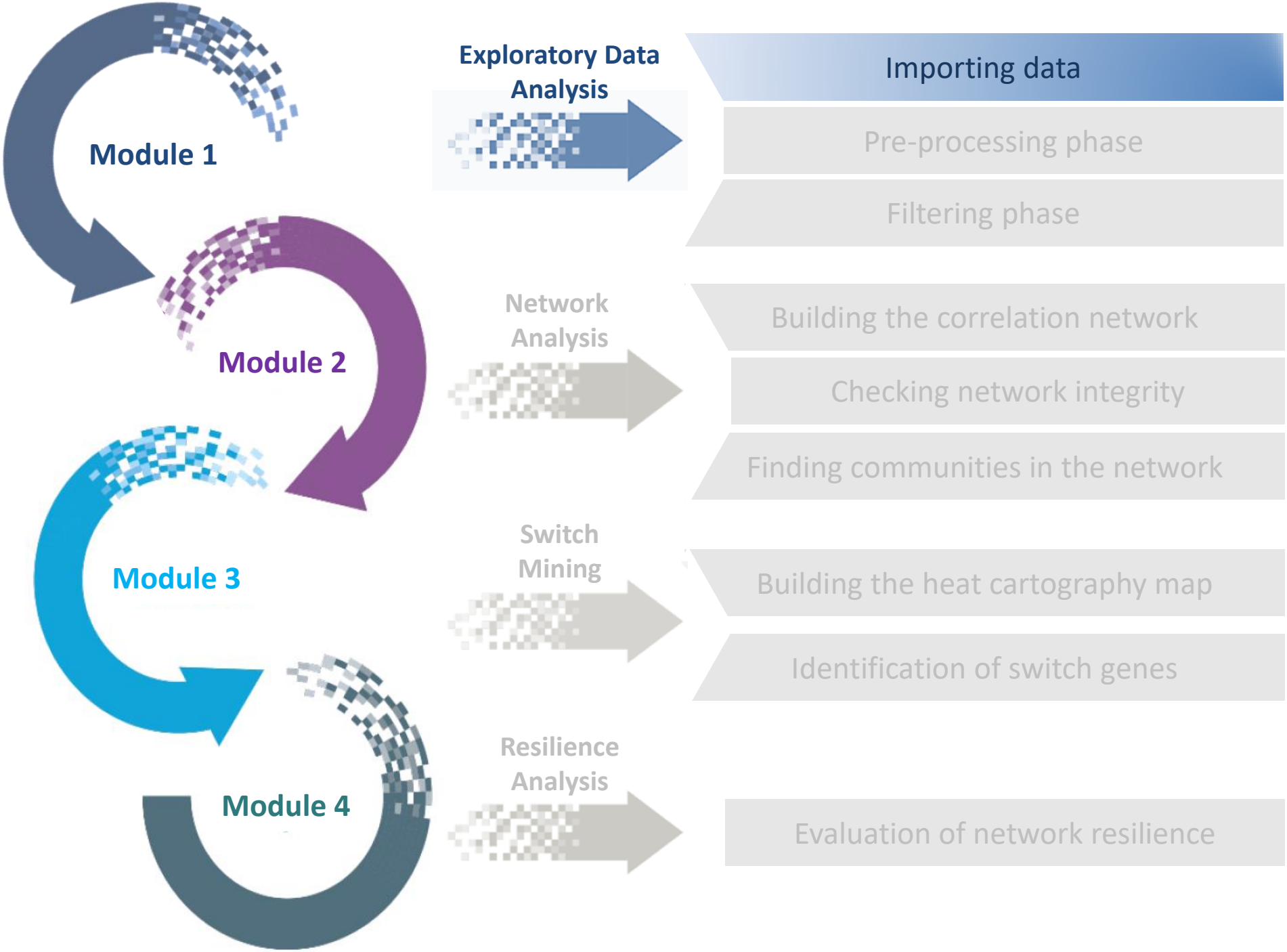
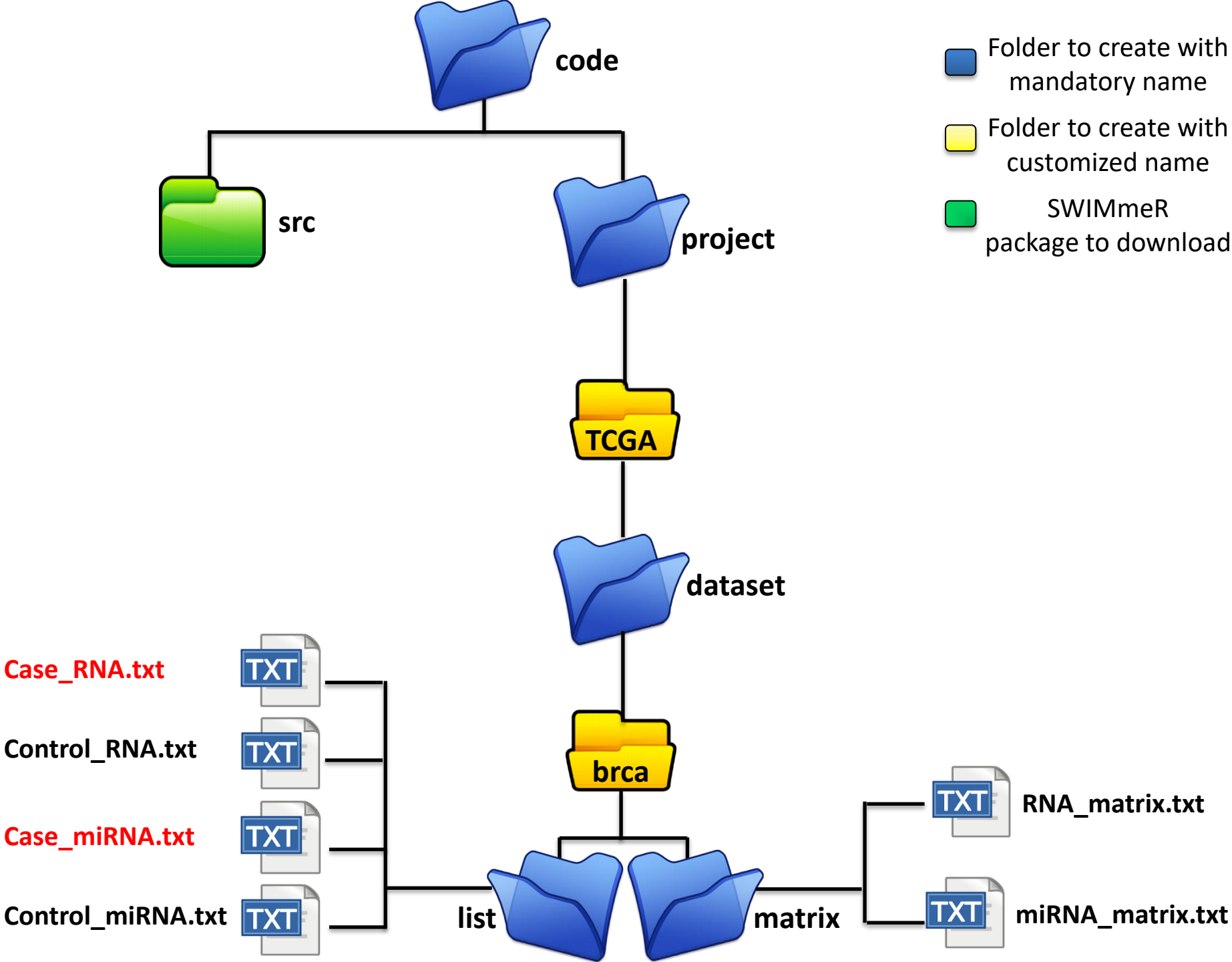
Module 1:
Exploratory Data Analysis

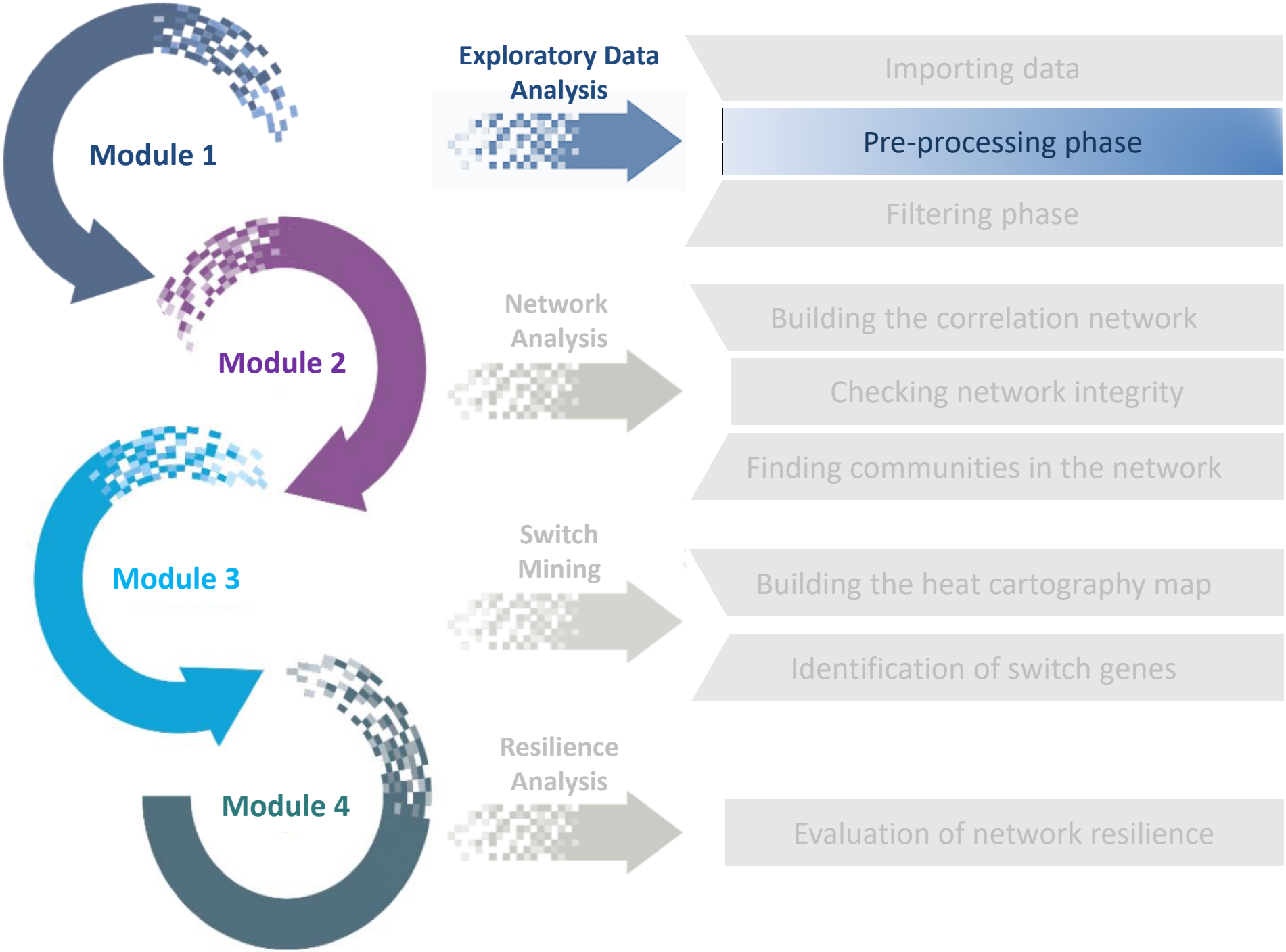

Module 1
Module 2
Module 3
Module 4
Exploratory Data Analysis
Importing data
Pre-processing phase
Filtering phase
Network Analysis
Building the correlation network
Checking network integrity
Finding communities in the network
Switch Mining
Building the heat cartography map
Identification of switch genes
Resilience Analysis
Evaluation of network resilience


Module 1
Module 2
Module 3
Module 4
Exploratory Data Analysis
Importing data
Pre-processing phase
Filtering phase
Network Analysis
Building the correlation network
Checking network integrity
Finding communities in the network
Switch Mining
Building the heat cartography map
Identification of switch genes
Resilience Analysis
Evaluation of network resilience

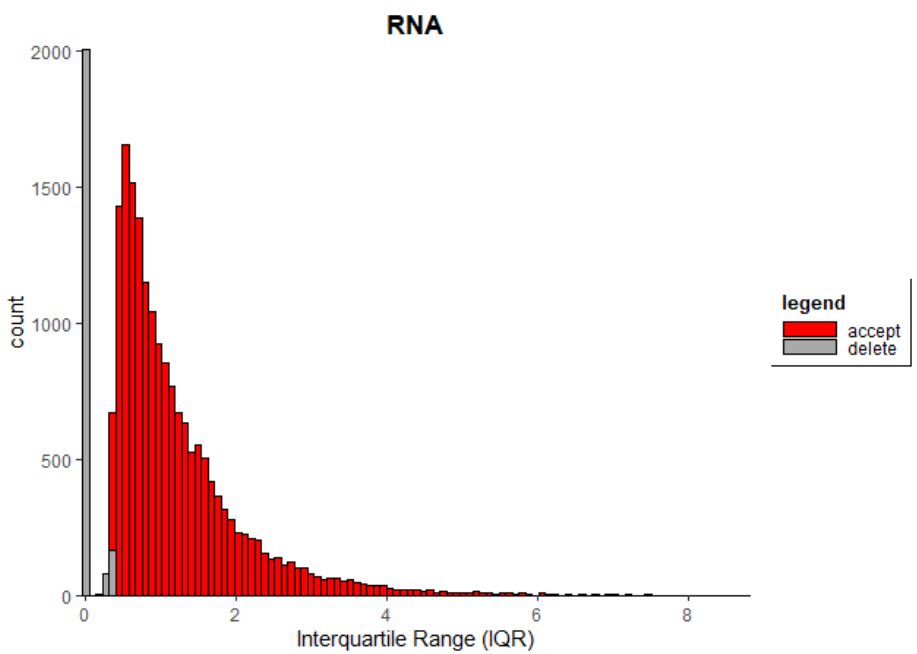
Example dataset - brca

- Data of **miRNA- and RNA-sequencing** samples of breast invasive carcinoma (**brca**), downloaded from **TCGA**:
  - 1182 RNA-sequencing samples
    - 1069 tumor samples
    - 113 normal samples
  - 1212 miRNA-sequencing samples
    - 1108 tumor samples
    - 104 normal samples
- The analysis was restricted to **103** tumor and matched-normal samples (i.e. tissues that are adjacent to the tumor and taken from the same patient) for both the RNA-sequencing and miRNA-sequencing


code
src
project
TCGA
dataset
brca
list
matrix
Case_RNA.txt
Control_RNA.txt
Case_miRNA.txt
Control_miRNA.txt
TXT
RNA_matrix.txt
miRNA_matrix.txt
Folder to create with mandatory name
Folder to create with customized name
SWIMmeR package to download


Module 1
Module 2
Module 3
Module 4
Exploratory Data Analysis
Importing data
Pre-processing phase
Filtering phase
Network Analysis
Building the correlation network
Checking network integrity
Finding communities in the network
Switch Mining
Building the heat cartography map
Identification of switch genes
Resilience Analysis
Evaluation of network resilience

Interquartile range (IQR) frequency distribution

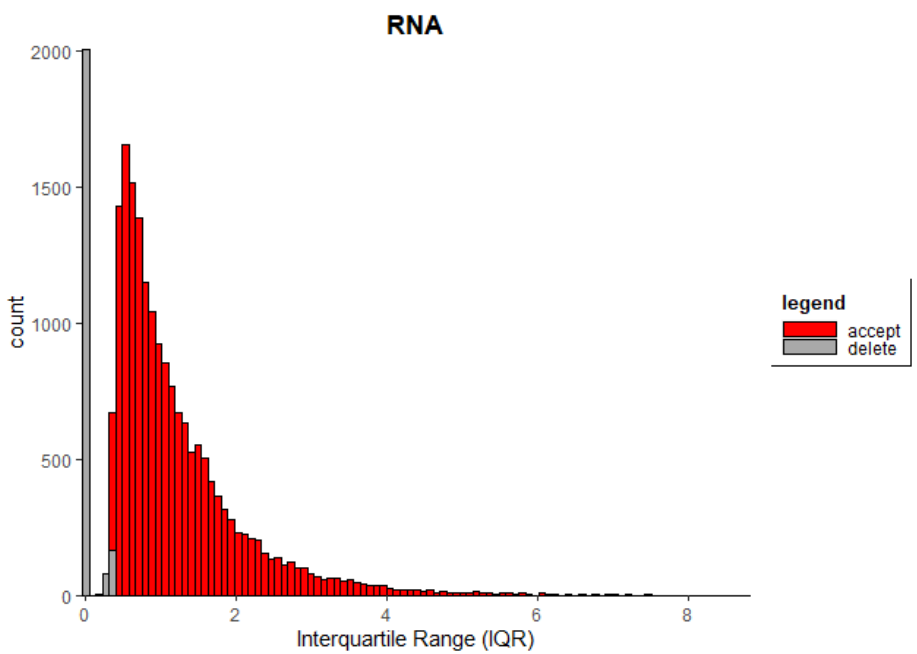


- x-axis represents the IQR
- y-axis represents the frequency of RNAs with that IQR
- grey bars represent the RNAs discarded, red bars the retained ones

Interquartile range (IQR) and Percentage of zeros

- This step aims to remove genes that have a small **Interquartile range (IQR)** and that have a **number of zero values** greater than a chosen threshold

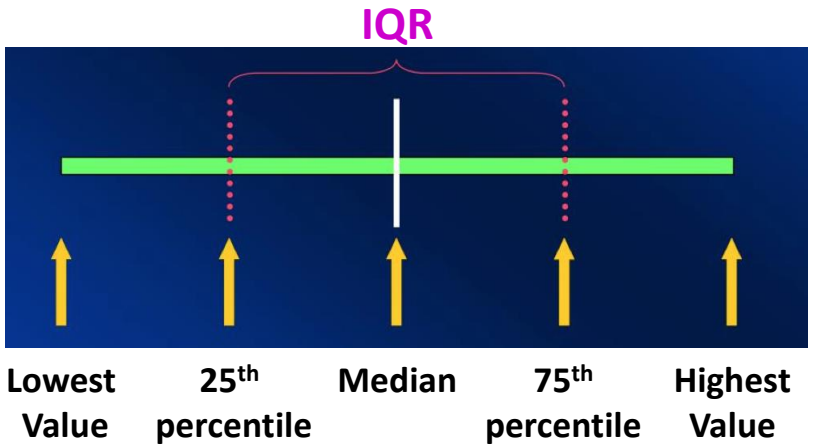
Interquartile range (IQR) frequency distribution



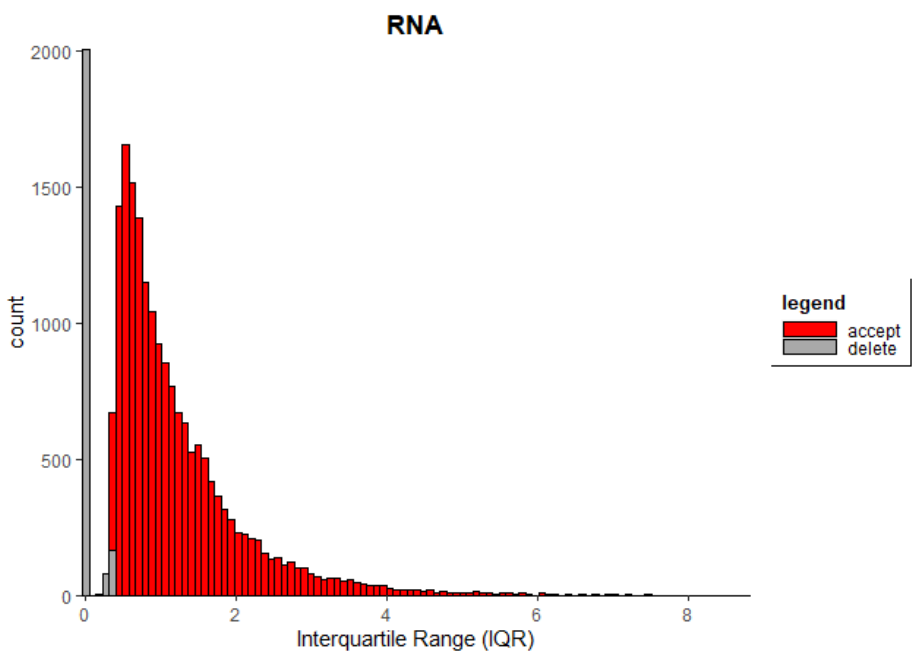
- x-axis represents the IQR
- y-axis represents the frequency of RNAs with that IQR
- grey bars represent the RNAs discarded, red bars the retained ones

Interquartile range (IQR)

- It is a measure of data variability (dispersion) around the median: IQR = 75th perc -25th perc
- It represents the range in which vary the 50% of the data when ordered from lowest to highest
- A small IQR indicates that values are less dispersed around the median.



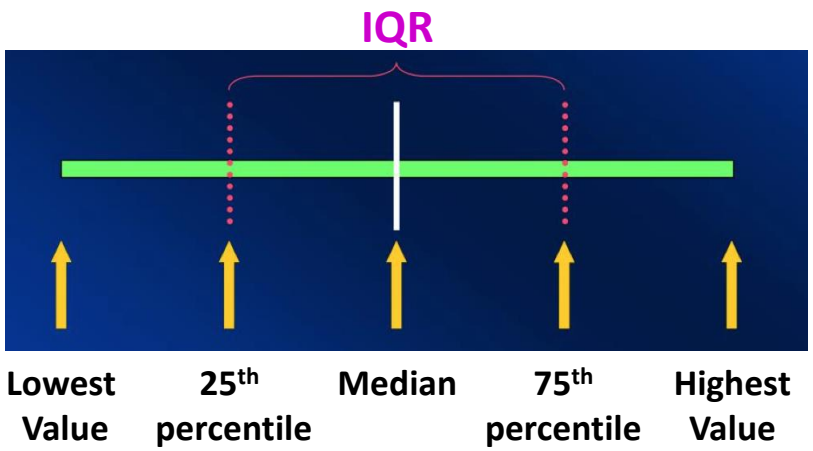
Interquartile range (IQR) frequency distribution



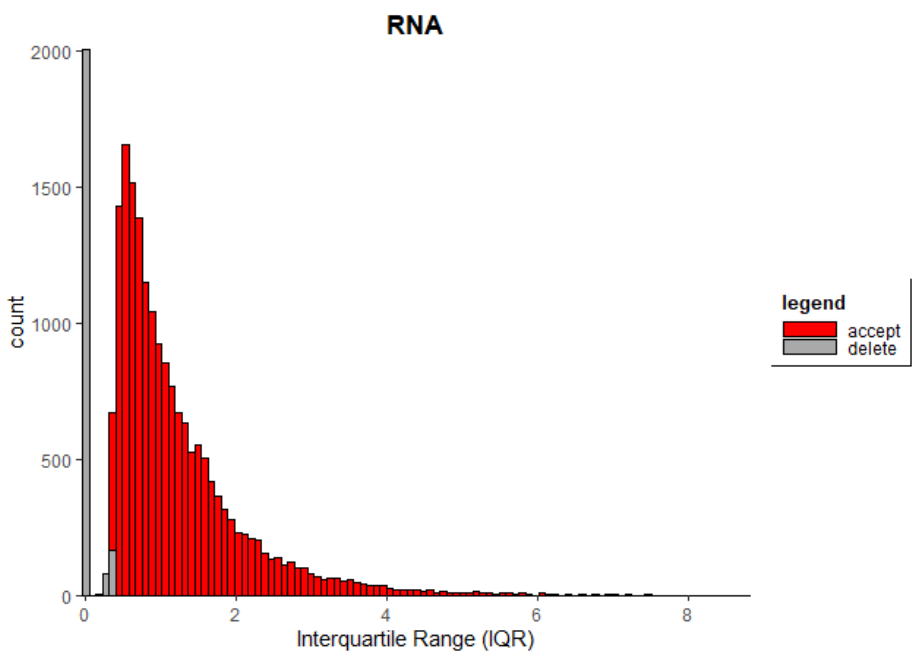
- x-axis represents the IQR
- y-axis represents the frequency of RNAs with that IQR
- grey bars represent the RNAs discarded, red bars the retained ones

Interquartile range (IQR)

```
computeStat <- function(log_data,N,M,paired,method,output_file_stat_dataORIG){

  ########################################
  # input parameters

  data <- log_data$data
  data_control <- log_data$data_control
  data_case <- log_data$data_case
  ########################################

  IQR <- apply(data,1,IQR,type=5)
```



Interquartile range (IQR) frequency distribution



- x-axis represents the IQR
- y-axis represents the frequency of RNAs with that IQR
- grey bars represent the RNAs discarded, red bars the retained ones

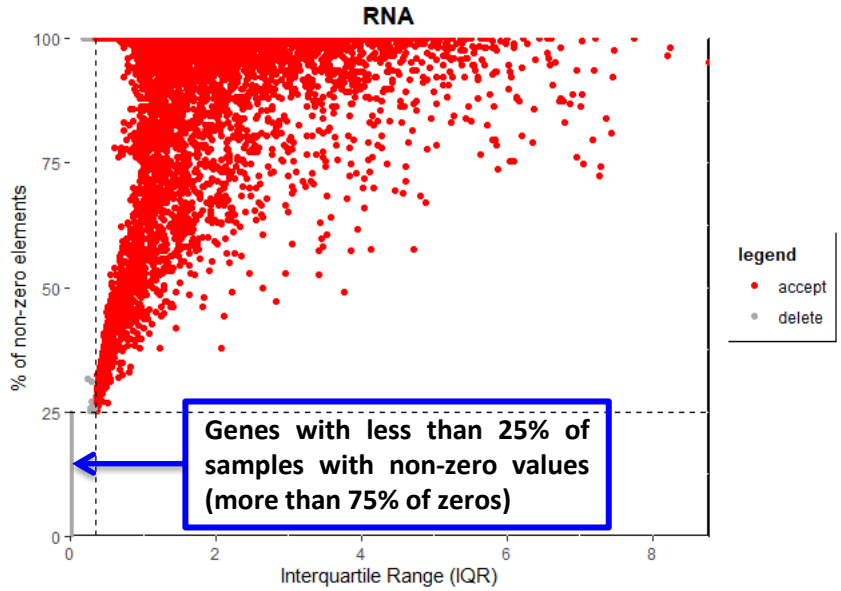
Histogram of IQR

```
getHistogram <- function(x,threshold,title,xlabel){

  df <- data.frame(variable = x)

  df$legend <- ifelse( (abs(x) <= threshold), "delete", "accept")

  w <- (max(x) - min(x)) / 100

  p = ggplot(df, aes(variable, fill = legend)) + geom_histogram
(binwidth = w, colour='black') +
    scale_x_continuous(expand = c(0, 0)) + scale_y_continuous
(expand = c(0, 0)) +
    scale_fill_manual(values = c("delete" = "darkgrey", "accept" =
"red")) +
    theme(panel.grid.major = element_blank(), panel.grid.minor =
element_blank(),
          panel.background = element_blank(), axis.line = element_l
ine(colour = "black"),
          plot.title = element_text(hjust = 0.5, face = "bold"),
          legend.title = element_text(colour = "black", size=10,
face="bold"),
          legend.key.height = unit(0.2, "cm"),legend.key.width =
unit(1, "cm"),
          legend.box.background = element_rect(colour = "black")) +
    labs(title = title, x = xlabel)

  print(p)

}
```

Scatter plot of the non-zeros values as function of the IQR

Percentage of zeros values

```
computeStat <- function(log_data,N,M,paired,method,output_file_stat_dataORIG){

  ######################################
  # input parameters

  data <- log_data$data
  data_control <- log_data$data_control
  data_case <- log_data$data_case
  ######################################

  IQR <- apply(data,1,IQR,type=5)

  perc_zeros <- apply(data, 1, function(x){ length(which(x == 0)) / length(x) *
100 })

  logFC <- rowMeans(data_case) - rowMeans(data_control)

  pval <- apply(data, 1, function(data){
    t.test(data[1:N], data[(N+1):M], paired = paired)$p.value
  })

  pval_adj <- p.adjust(pval, method = method)

  df_stat <- data.frame(IQR = IQR,
                        perc_zeros = perc_zeros,
                        logFC = logFC,
                        pval = pval,
                        pval_adj = pval_adj)

  write.table(df_stat, output_file_stat_dataORIG, row.names = T, col.names = NA,
sep = "\t", quote = F)

  return(df_stat)

}
```

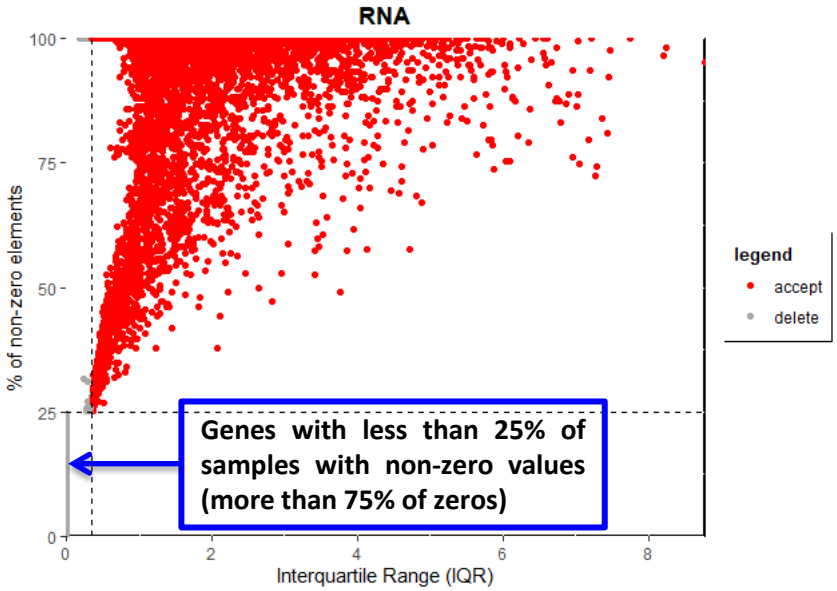


- x-axis represents IQR
- y-axis represents the % of non-zero values.
- vertical and horizontal lines mark the chosen thresholds
- grey circles represent the RNAs discarded, the red circles are the retained RNAs

Scatter plot of the non-zeros values as function of the IQR

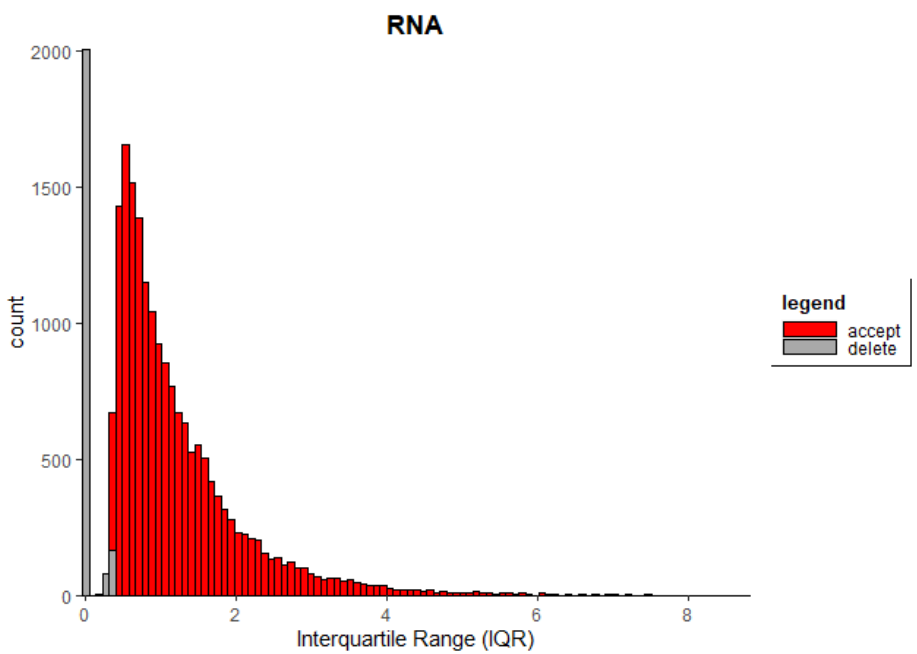
Scatter plot

```
getScatterPlot <- function(IQR,perc_zeros,threshold_prc_iqr,threshold_perc_zeros,title
){

  no_null_el <- 100 - perc_zeros

  df <- data.frame(IQR = IQR, no_null_el = no_null_el)

  thr_iqr <- as.numeric(quantile(IQR,threshold_prc_iqr))
  thr_no_null_el <- 100 - threshold_perc_zeros

  condition1 <- (IQR <= thr_iqr) | (no_null_el < thr_no_null_el)
  condition2 <- (IQR > thr_iqr) & (no_null_el >= thr_no_null_el)

  df$legend <- ifelse(condition1,"delete",ifelse(condition2,"accept","delete"))

  p <- ggplot(df, aes(x = IQR, y = no_null_el, color = legend)) + geom_point() +
    scale_x_continuous(expand = c(0, 0)) + scale_y_continuous(expand = c(0, 0)) +
    scale_color_manual(values = c("delete" = "darkgrey", "accept" = "red"))  +
    theme(panel.background = element_rect(fill = "white", colour = "black", size = 1),
          plot.title = element_text(hjust = 0.5, face = "bold"),
          legend.title = element_text(colour = "black", size=10, face="bold"),
          legend.key = element_rect(fill = "white", colour = "white"),
          legend.box.background = element_rect(colour = "black")) +
    labs(title = title, x = "Interquartile Range (IQR)", y = "% of non-zero elements")
+
    geom_hline(yintercept = thr_no_null_el, linetype = "dashed", color = "black") +
    geom_vline(xintercept = thr_iqr, linetype = "dashed", color = "black")

  print(p)

}
```

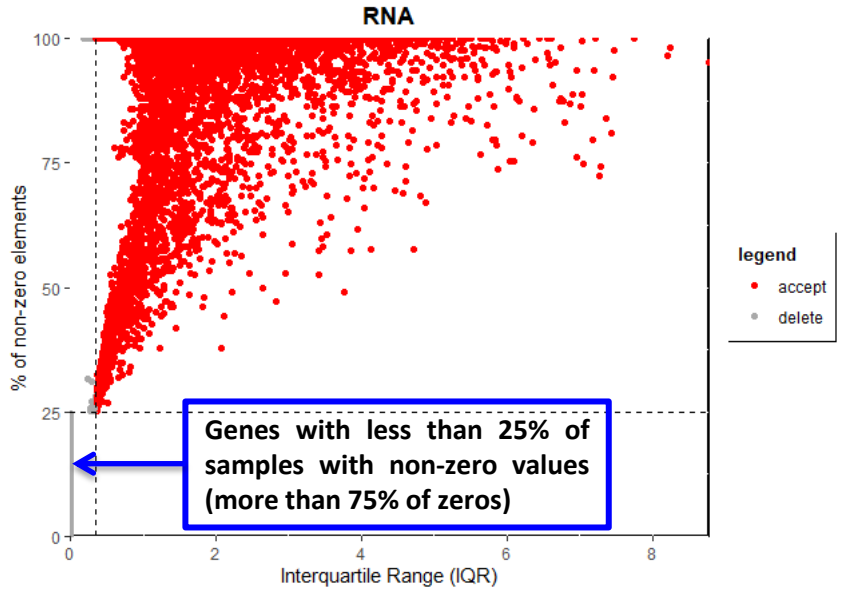


- x-axis represents IQR
- y-axis represents the % of non-zero values.
- vertical and horizontal lines mark the chosen thresholds
- grey circles represent the RNAs discarded, the red circles are the retained RNAs

IQR frequency distribution and scatter plot of the non-zeros values for RNAs

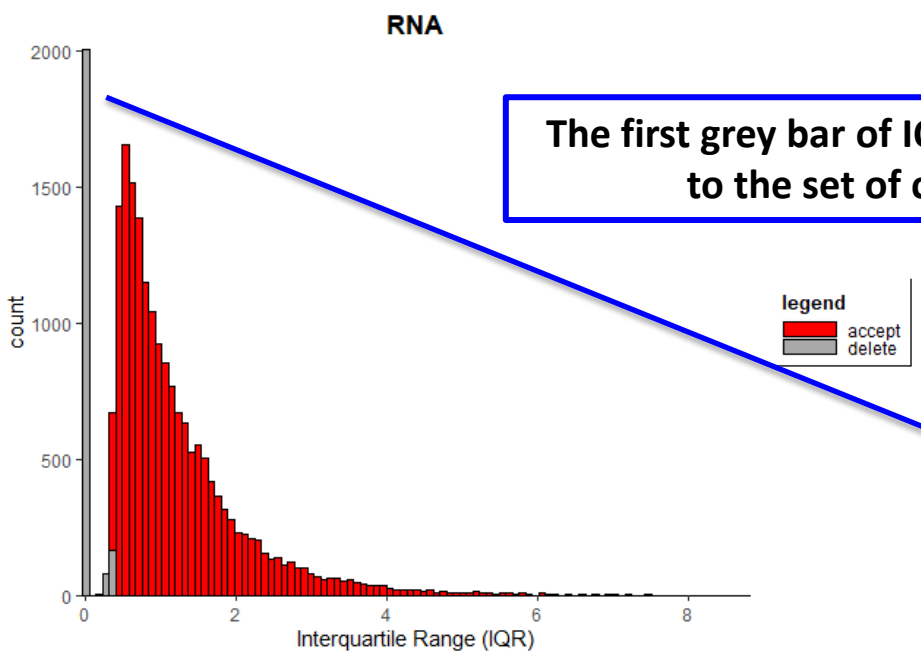


- x-axis represents the IQR
- y-axis represents the frequency of RNAs with that IQR
- grey bars represent the RNAs discarded, red bars the retained ones



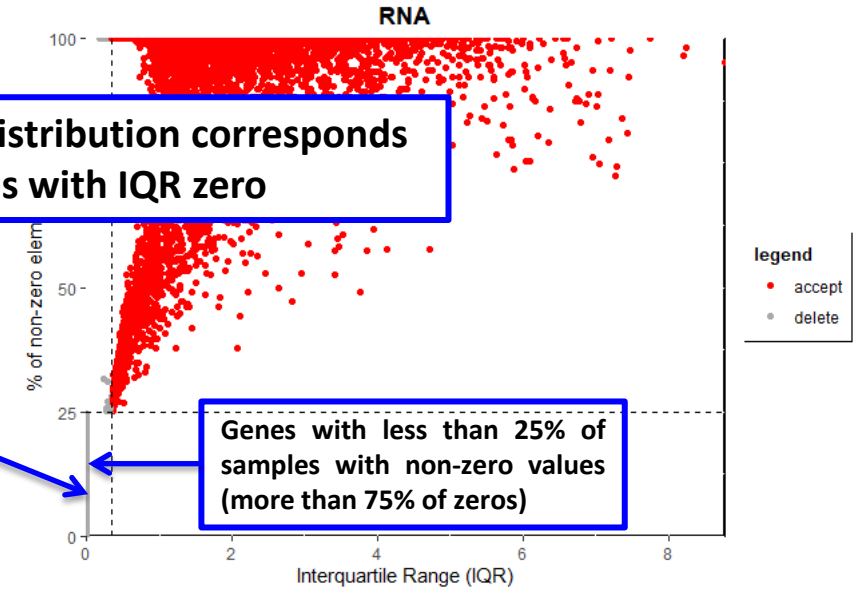
- x-axis represents IQR
- y-axis represents the % of non-zero values.
- vertical and horizontal lines mark the chosen thresholds
- grey circles represent the RNAs discarded, the red circles are the retained RNAs

IQR frequency distribution and scatter plot of the non-zeros values for RNAs



- x-axis represents the IQR
- y-axis represents the frequency of RNAs with that IQR
- grey bars represent the RNAs discarded, red bars the retained ones

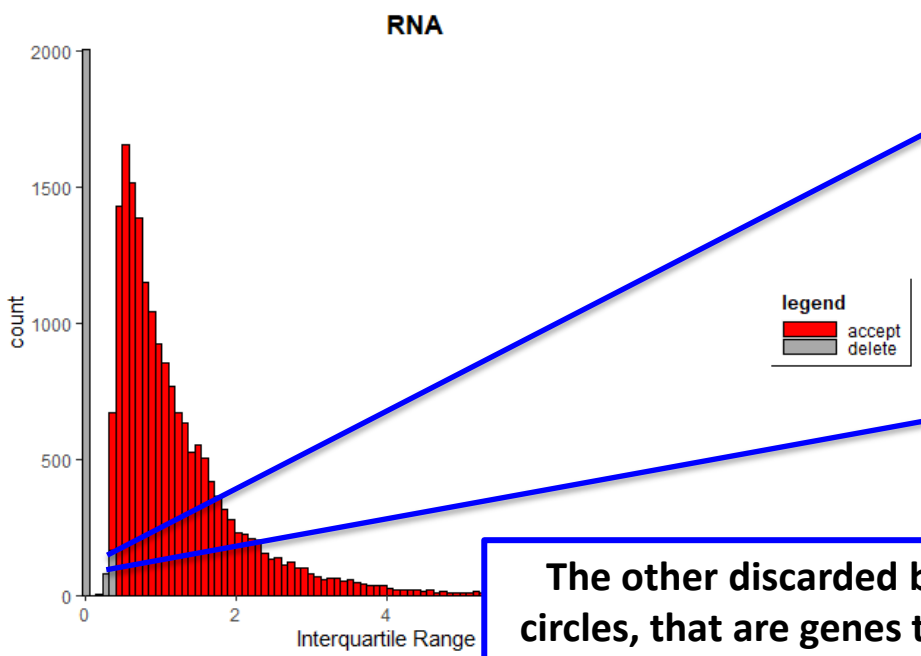
The first grey bar of IQR distribution corresponds to the set of circles with IQR zero

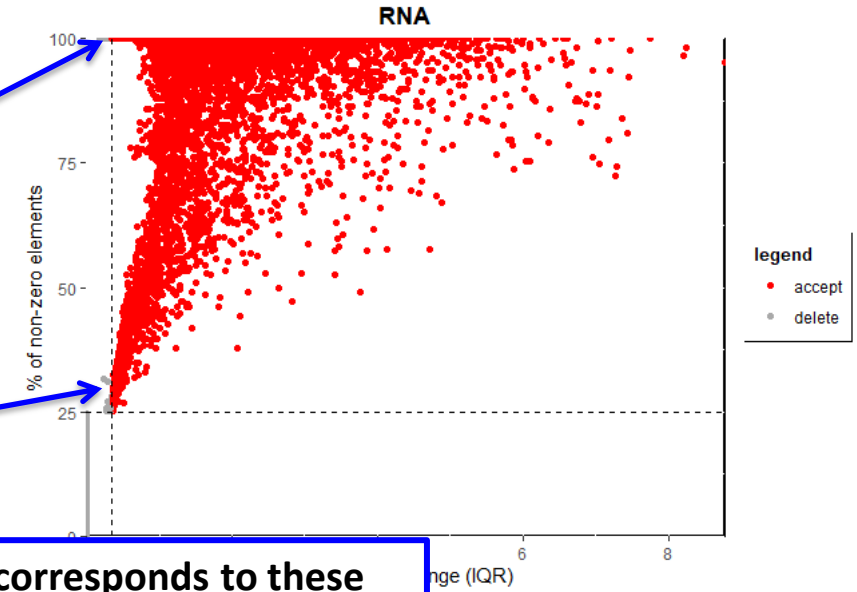


Genes with less than 25% of samples with non-zero values (more than 75% of zeros)

- x-axis represents IQR
- y-axis represents the % of non-zero values.
- vertical and horizontal lines mark the chosen thresholds
- grey circles represent the RNAs discarded, the red circles are the retained RNAs

IQR frequency distribution and scatter plot of the non-zeros values for RNAs





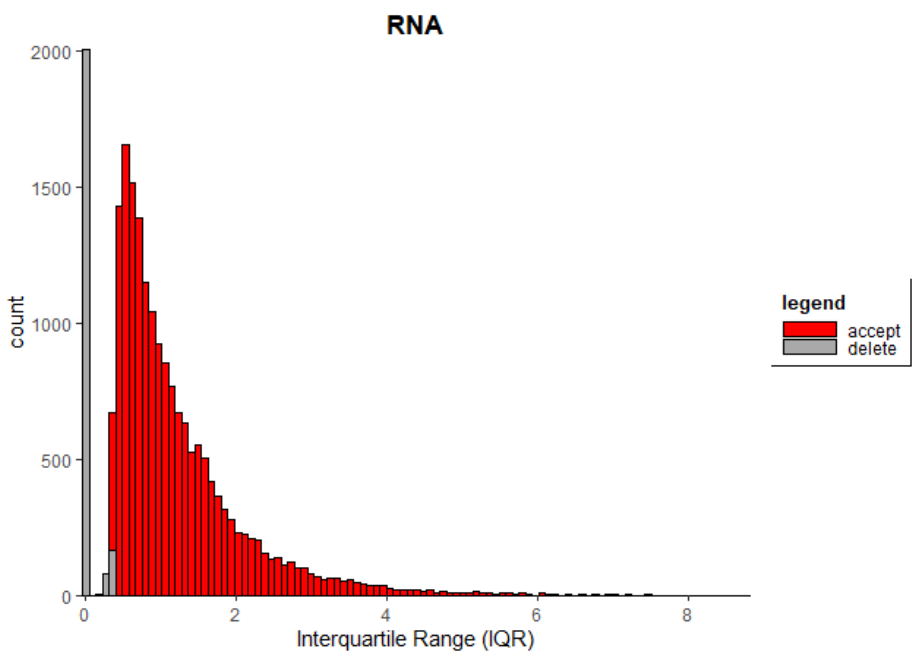
- x-axis represents the IQR
- y-axis represents the frequ
- grey bars represent the RNAs discarded, red bars the retained ones

The other discarded bars corresponds to these circles, that are genes that do not respect the IQR threshold but have an acceptable number of non-zero values

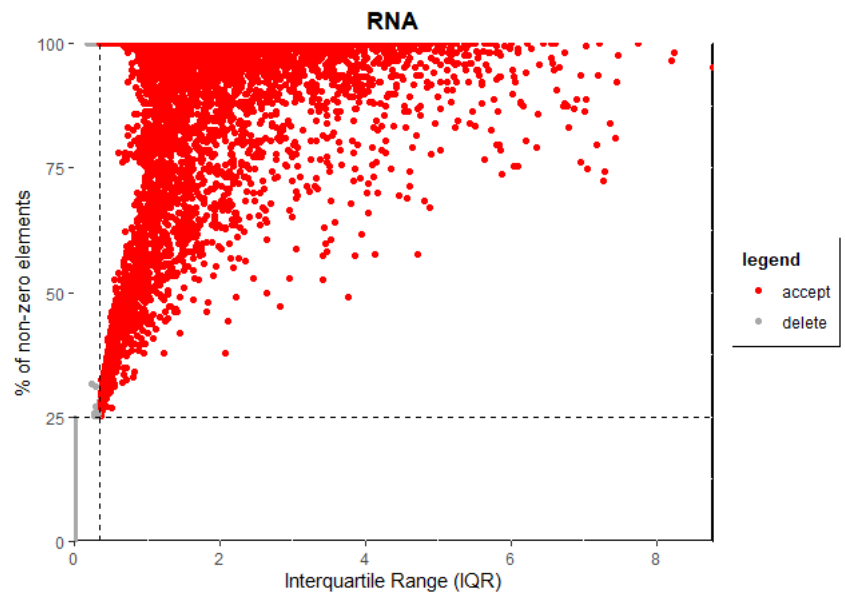
non-zero values.
s mark the chosen thresholds

- grey circles represent the RNAs discarded, the red circles are the retained RNAs

IQR frequency distribution and scatter plot of the non-zeros values for RNAs



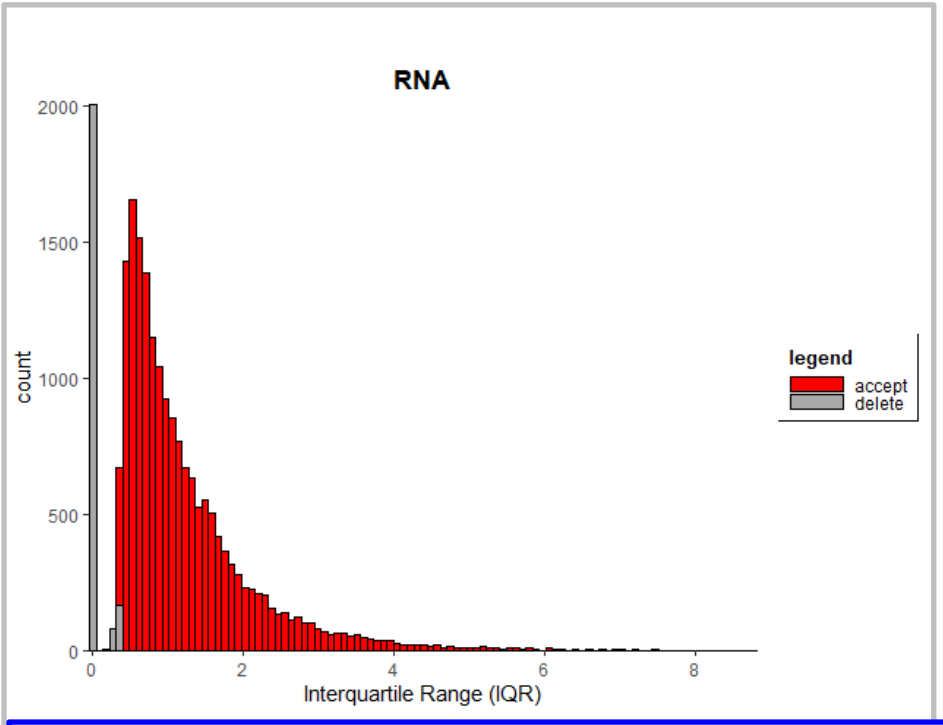
- x-axis represents the IQR
- y-axis represents the frequency of RNAs with that IQR
- grey bars represent the RNAs discarded, red bars the retained ones

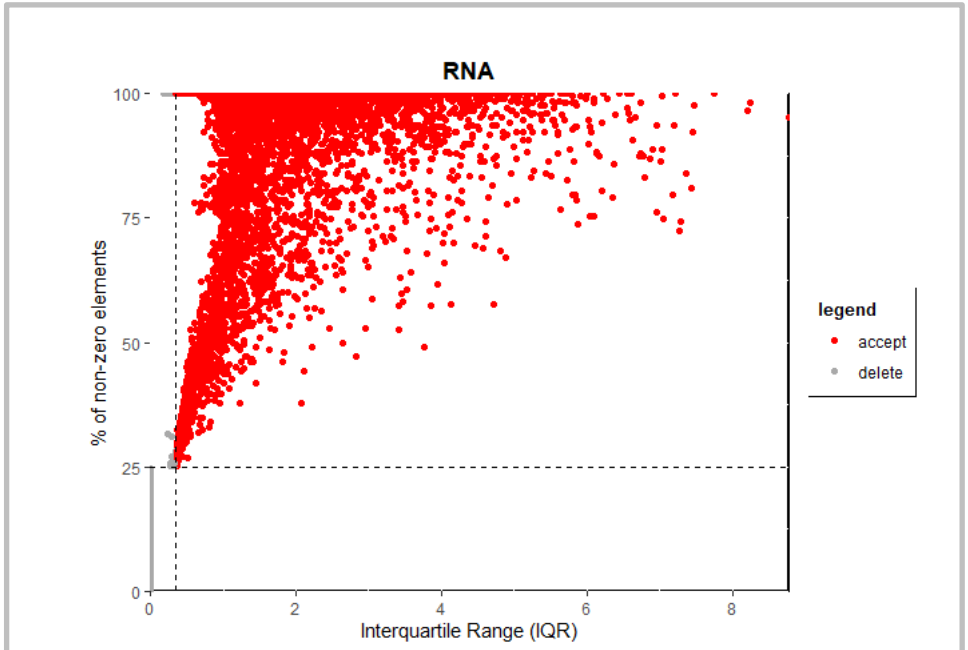


- x-axis represents IQR
- y-axis represents the % of non-zero values.
- vertical and horizontal lines mark the chosen thresholds
- grey circles represent the RNAs discarded, the red circles are the retained RNAs

⚠ **Caveat:** Reasonable thresholds correspond to a gap or discontinuity in the plot. In principle, one could select a greater IQR threshold in order to filter out other small red bars, even if they do not correspond to a gap or discontinuity.

IQR frequency distribution and scatter plot of the non-zeros values for RNAs

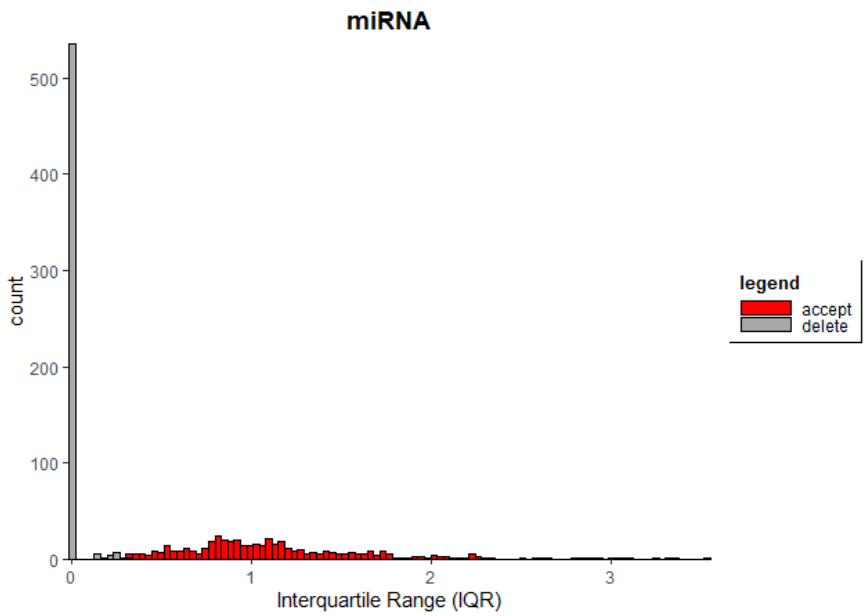




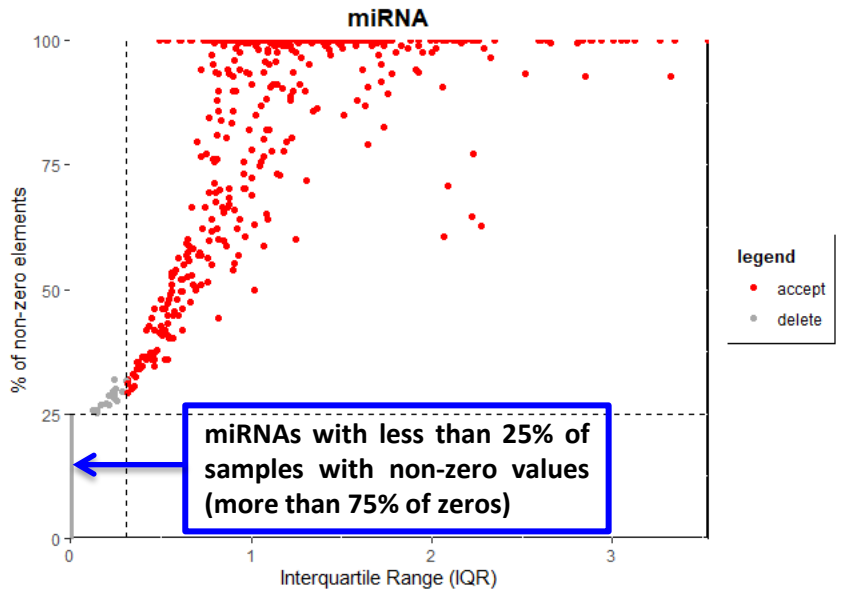
- x-axis represents IQR

This step is essential in order to shorten the list of genes to give as input to the the next a t-test and a standard correction procedure for multiple tests will be performed to adjust p-values. This correction depends on the length of input list: smaller is the list, less strict is the correction.

IQR frequency distribution and scatter plot of the non-zeros values for miRNAs

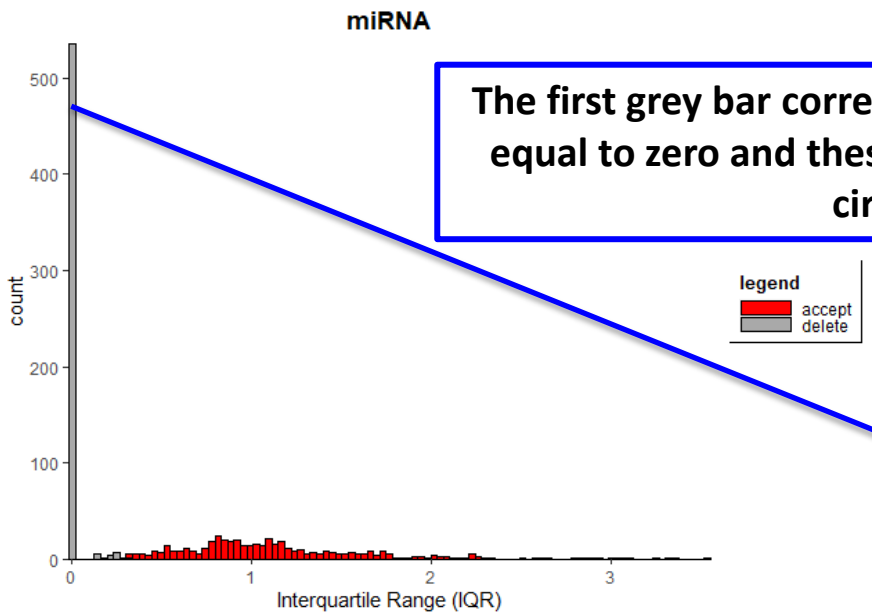


- x-axis represents the IQR
- y-axis represents the frequency of miRNAs with that IQR
- grey bars represent the miRNAs discarded, red bars the retained ones

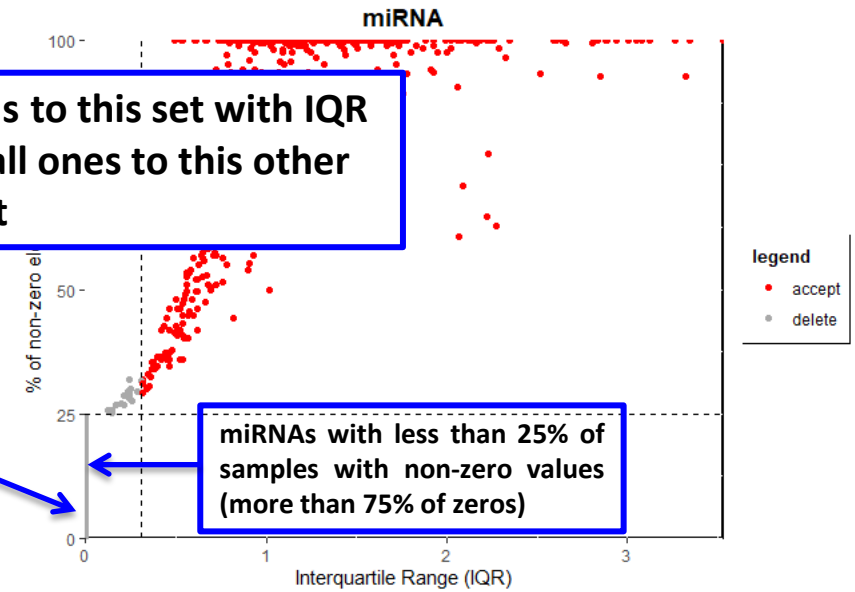


- x-axis represents IQR
- y-axis represents the % of non-zero values
- vertical and horizontal lines mark the chosen thresholds
- grey circles represent the miRNAs discarded, the red circles are the retained miRNAs

IQR frequency distribution and scatter plot of the non-zeros values for miRNAs



The first grey bar corresponds to this set with IQR equal to zero and these small ones to this other circle set

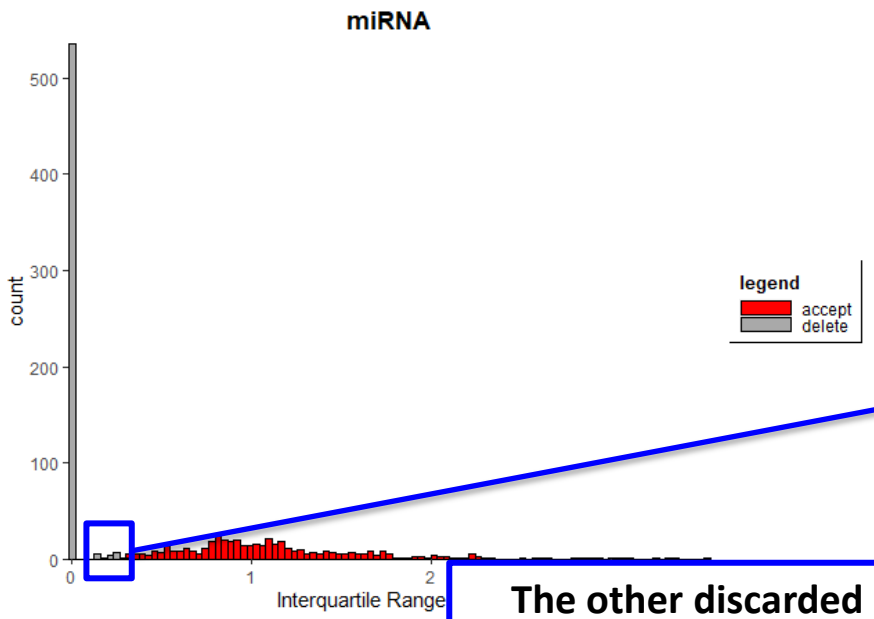


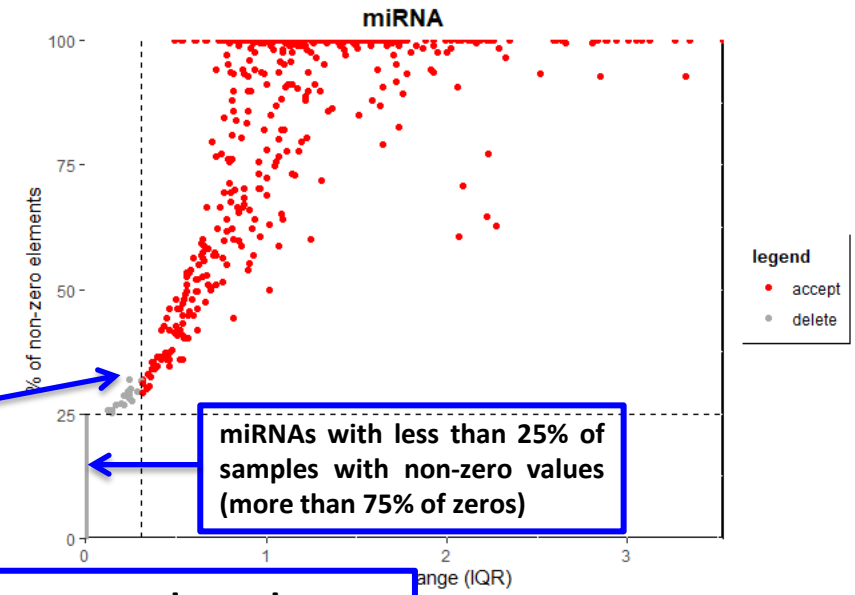
miRNAs with less than 25% of samples with non-zero values (more than 75% of zeros)

- x-axis represents the IQR
- y-axis represents the frequency of miRNAs with that IQR
- grey bars represent the miRNAs discarded, red bars the retained ones

- x-axis represents IQR
- y-axis represents the % of non-zero values
- vertical and horizontal lines mark the chosen thresholds
- grey circles represent the miRNAs discarded, the red circles are the retained miRNAs

IQR frequency distribution and scatter plot of the non-zeros values for miRNAs



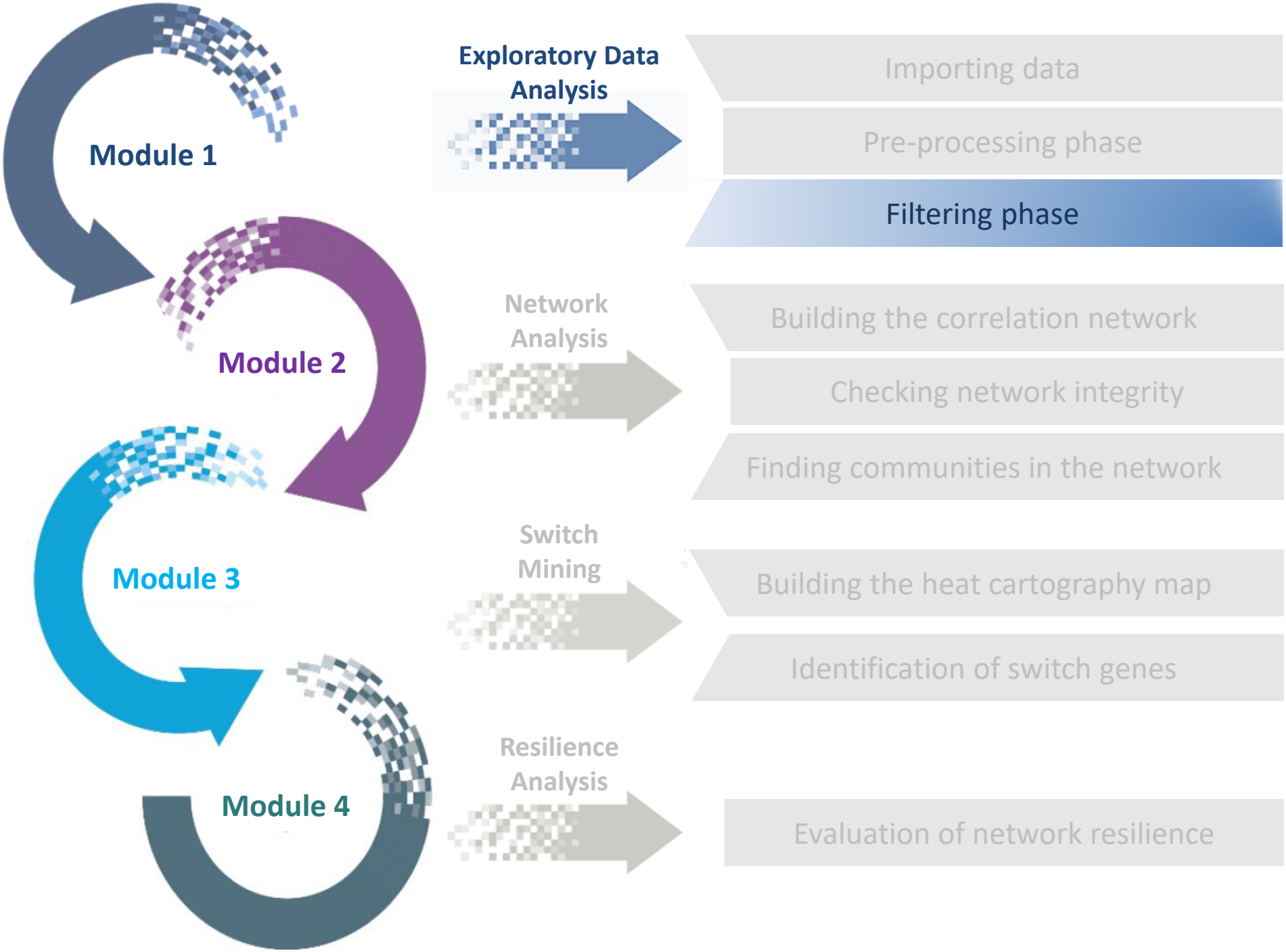


miRNAs with less than 25% of samples with non-zero values (more than 75% of zeros)

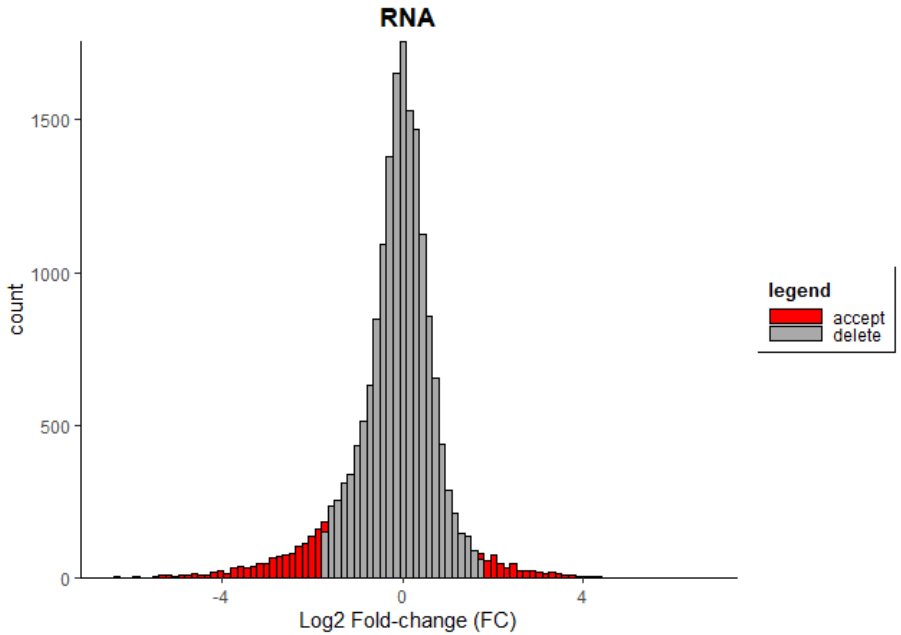
The other discarded bars corresponds to these other circles with an IQR greater than 0

- x-axis represents the IQR
- y-axis represents the frequency of miRNAs with that IQR
- grey bars represent the miRNAs discarded, red bars the retained ones

- y-axis represents the % of non-zero values
- vertical and horizontal lines mark the chosen thresholds
- grey circles represent the miRNAs discarded, the red circles are the retained miRNAs


Module 1
Module 2
Module 3
Module 4
Exploratory Data Analysis
Importing data
Pre-processing phase
Filtering phase
Network Analysis
Building the correlation network
Checking network integrity
Finding communities in the network
Switch Mining
Building the heat cartography map
Identification of switch genes
Resilience Analysis
Evaluation of network resilience

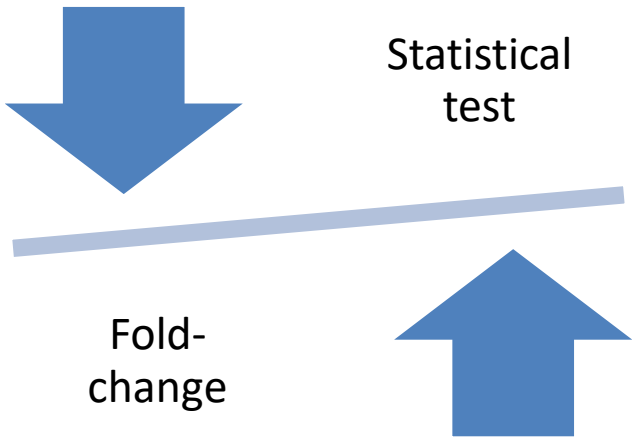
Fold-change (FC) frequency distribution



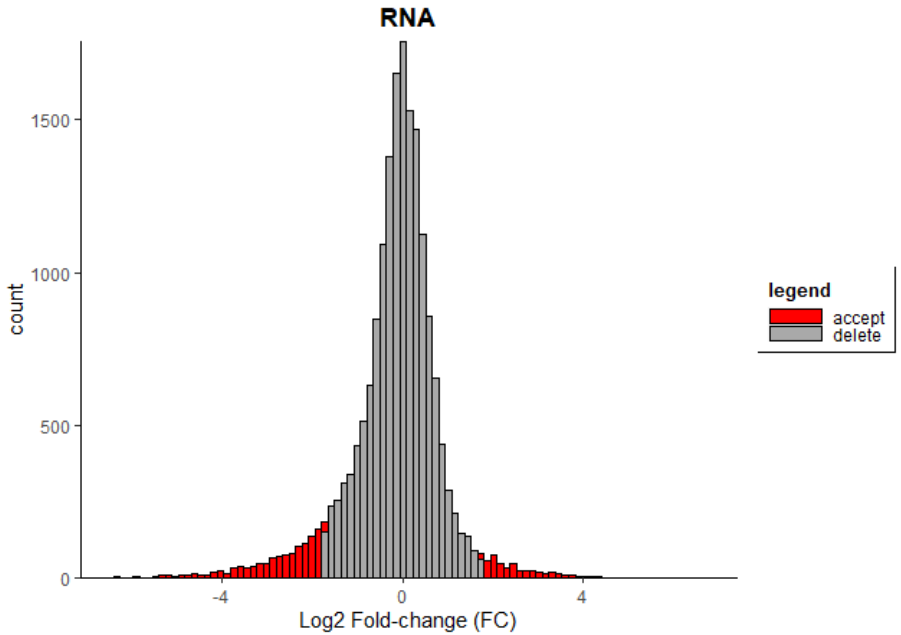
- x-axis represents the fold-change (logarithmic scale)
- y-axis represents the frequency of the obtained fold-change values
- grey bars represent the RNAs discarded, red bars the retained ones

Fold-change (FC) and Statistical test

- The filtering step aims to select the genes that are varying on average a lot and in a statistically significant way between the two conditions



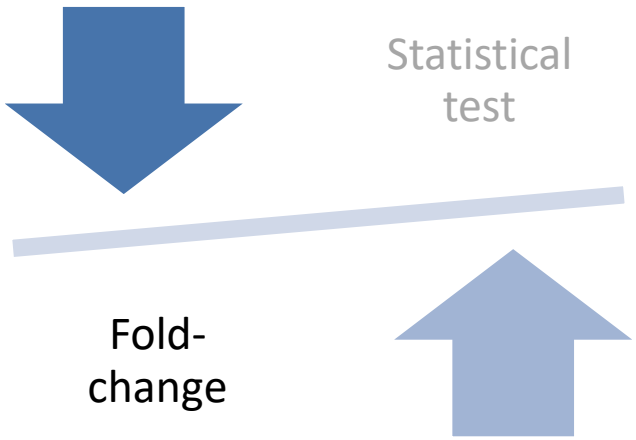
Fold-change (FC) frequency distribution



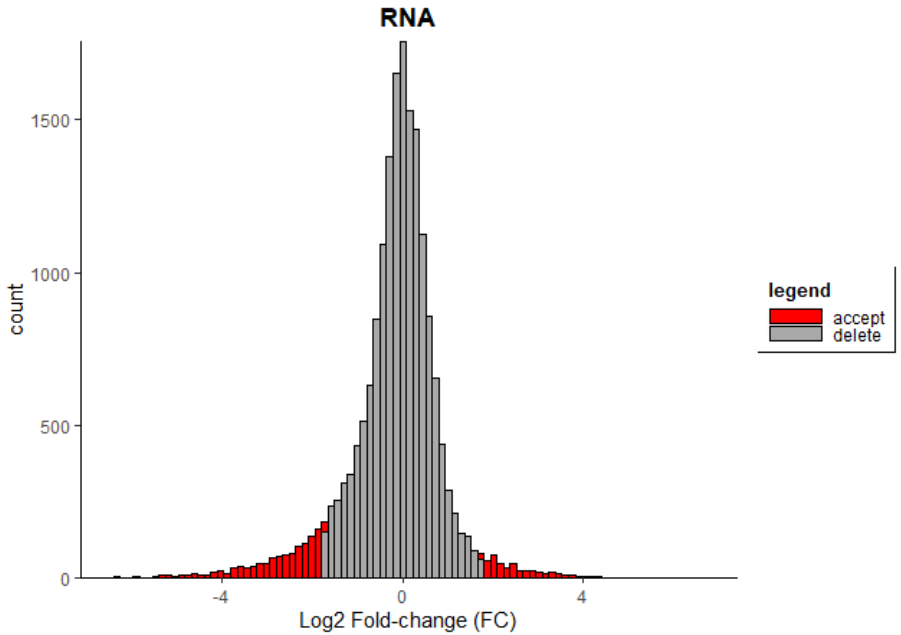
- x-axis represents the fold-change (logarithmic scale)
- y-axis represents the frequency of the obtained fold-change values
- grey bars represent the RNAs discarded, red bars the retained ones

Fold-change (FC)

- The filtering step aims to select the genes that **are varying on average a lot** and in a statistically significant way between the two conditions
- **FC =** mean case/mean control



Fold-change (FC) frequency distribution

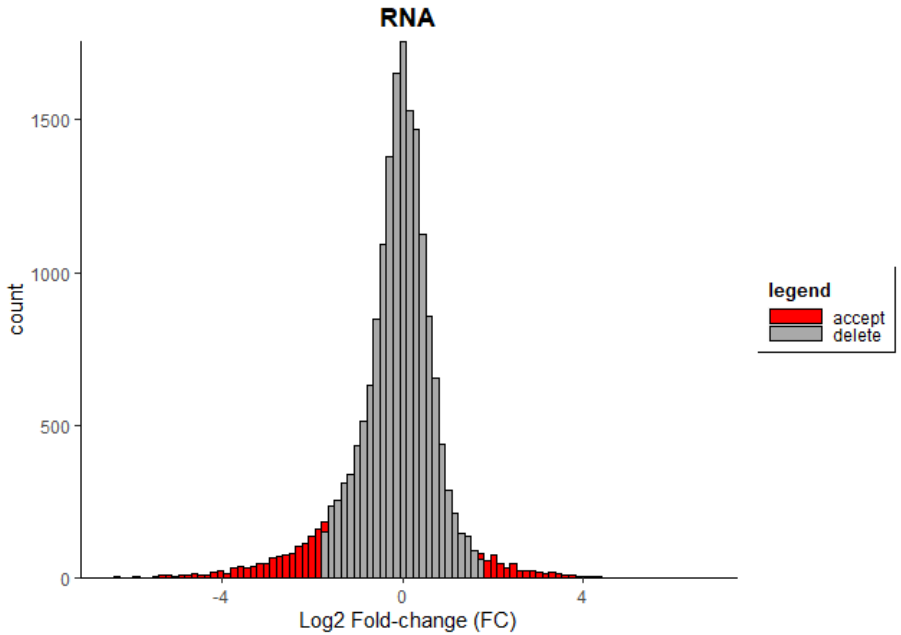


- x-axis represents the fold-change (logarithmic scale)
- y-axis represents the frequency of the obtained fold-change values
- grey bars represent the RNAs discarded, red bars the retained ones

Fold-change (FC)

```
computeStat <- function(log_data,N,M,paired,method,output_file_stat_dataORIG){

  ######################################
  # input parameters

  data <- log_data$data
  data_control <- log_data$data_control
  data_case <- log_data$data_case
  ######################################

  IQR <- apply(data,1,IQR,type=5)

  perc_zeros <- apply(data, 1, function(x){ length(which(x == 0)) / length(x) *
100 })

  logFC <- rowMeans(data_case) - rowMeans(data_control)

  pval <- apply(data, 1, function(data){
    t.test(data[1:N], data[(N+1):M], paired = paired)$p.value
  })

  pval_adj <- p.adjust(pval, method = method)

  df_stat <- data.frame(IQR = IQR,
                        perc_zeros = perc_zeros,
                        logFC = logFC,
                        pval = pval,
                        pval_adj = pval_adj)

  write.table(df_stat, output_file_stat_dataORIG, row.names = T, col.names = NA,
sep = "\t", quote = F)

  return(df_stat)

}
```

Fold-change (FC) frequency distribution

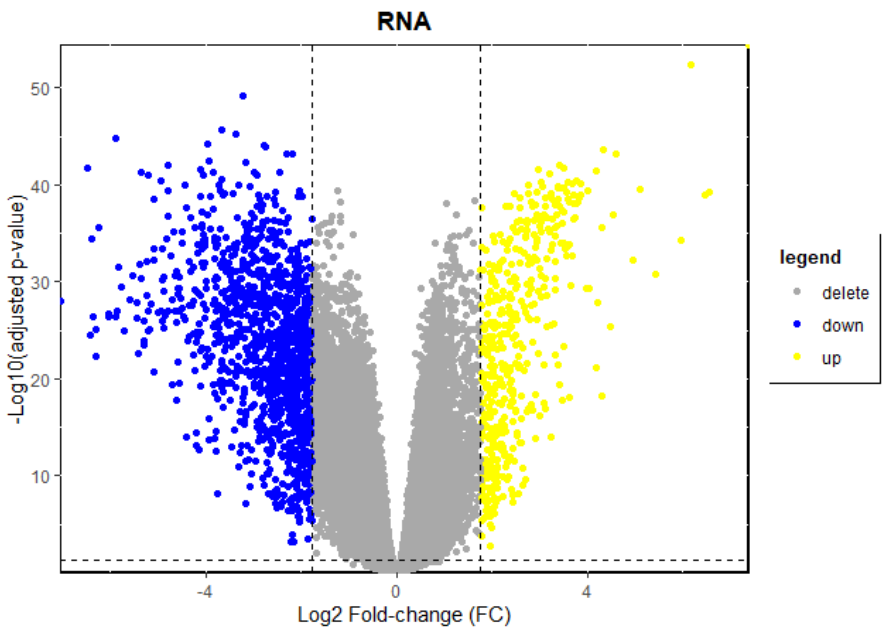


- x-axis represents the fold-change (logarithmic scale)
- y-axis represents the frequency of the obtained fold-change values
- grey bars represent the RNAs discarded, red bars the retained ones

Histogram of Fold-change (FC)

```
getHistogram <- function(x,threshold,title,xlabel){

  df <- data.frame(variable = x)

  df$legend <- ifelse( (abs(x) <= threshold), "delete", "accept")

  w <- (max(x) - min(x)) / 100

  p = ggplot(df, aes(variable, fill = legend)) + geom_histogram
(binwidth = w, colour='black') +
    scale_x_continuous(expand = c(0, 0)) + scale_y_continuous
(expand = c(0, 0)) +
    scale_fill_manual(values = c("delete" = "darkgrey", "accept" =
"red")) +
    theme(panel.grid.major = element_blank(), panel.grid.minor =
element_blank(),
          panel.background = element_blank(), axis.line = element_l
ine(colour = "black"),
          plot.title = element_text(hjust = 0.5, face = "bold"),
          legend.title = element_text(colour = "black", size=10,
face="bold"),
          legend.key.height = unit(0.2, "cm"),legend.key.width =
unit(1, "cm"),
          legend.box.background = element_rect(colour = "black")) +
    labs(title = title, x = xlabel)

  print(p)

}
```

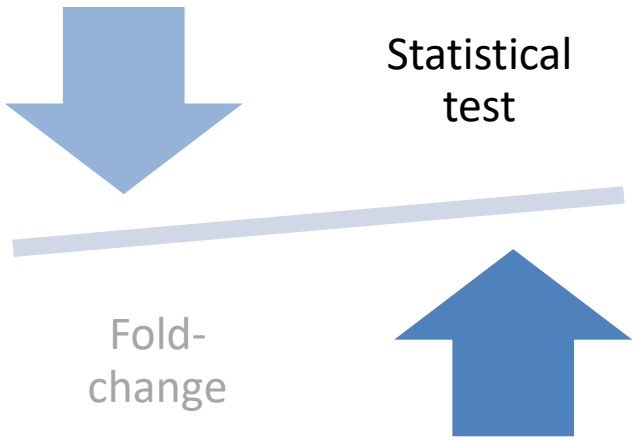
Volcano plot



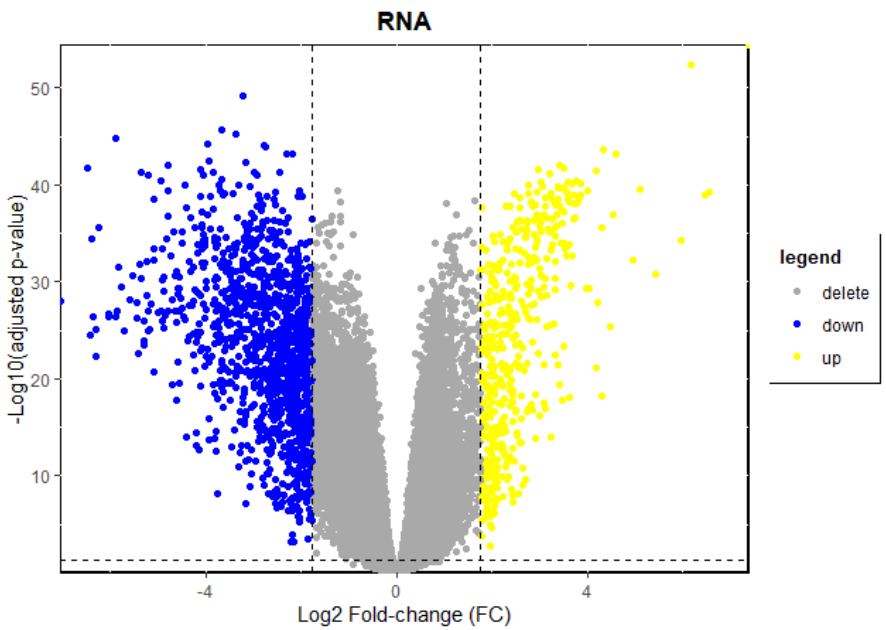
- x-axis represents the FC
- y-axis represents the FDR adjusted p-values (-log10 of the p-values) of the Student's t-test
- vertical dashed red lines and the horizontal ones correspond to the selected thresholds for the fold-change and the adjusted p-values
- grey points represent discarded according to the selected threshold
- blue points are the down-regulated RNAs
- yellow points are the up-regulated RNAs

Statistical test

- The filtering step aims to select the genes that are varying on average a lot and **in a statistically significant way** between the two conditions
- **Student's t-test**



Volcano plot

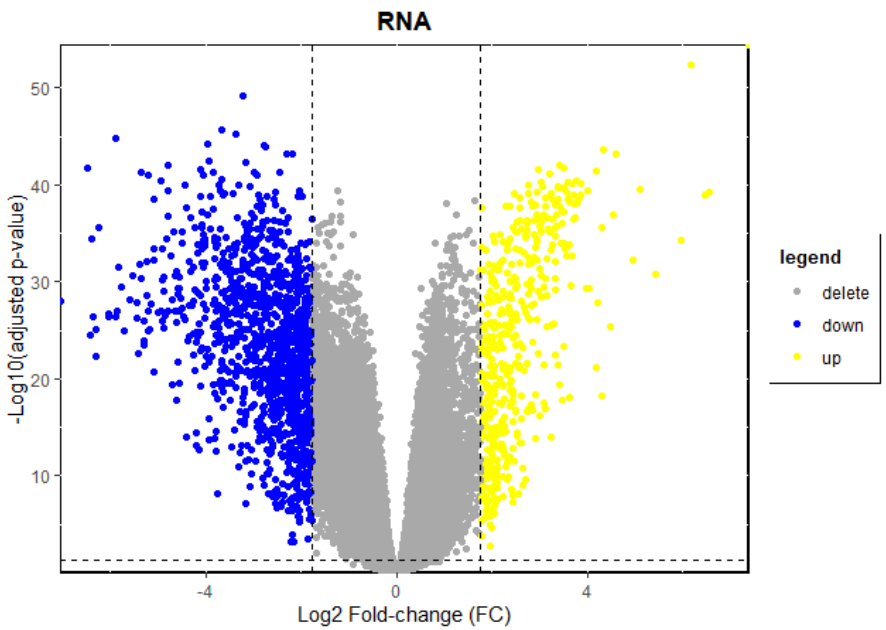


- x-axis represents the FC
- y-axis represents the FDR adjusted p-values (-log10 of the p-values) of the Student's t-test
- vertical dashed red lines and the horizontal ones correspond to the selected thresholds for the fold-change and the adjusted p-values
- grey points represent discarded according to the selected threshold
- blue points are the down-regulated RNAs
- yellow points are the up-regulated RNAs

Statistical test

- Statistical tests are necessary to accept or reject a hypothesis
- Each statistical test starts from a **"null" hypothesis (H_0)** of absence of differences
- **Example:** "null" hypothesis = absence of differences between average values of normal and cancer distributions
- The **goal of a statistical test** is to reject the null hypothesis, and then accepting the alternative hypothesis (H_1) that differences exist

Volcano plot

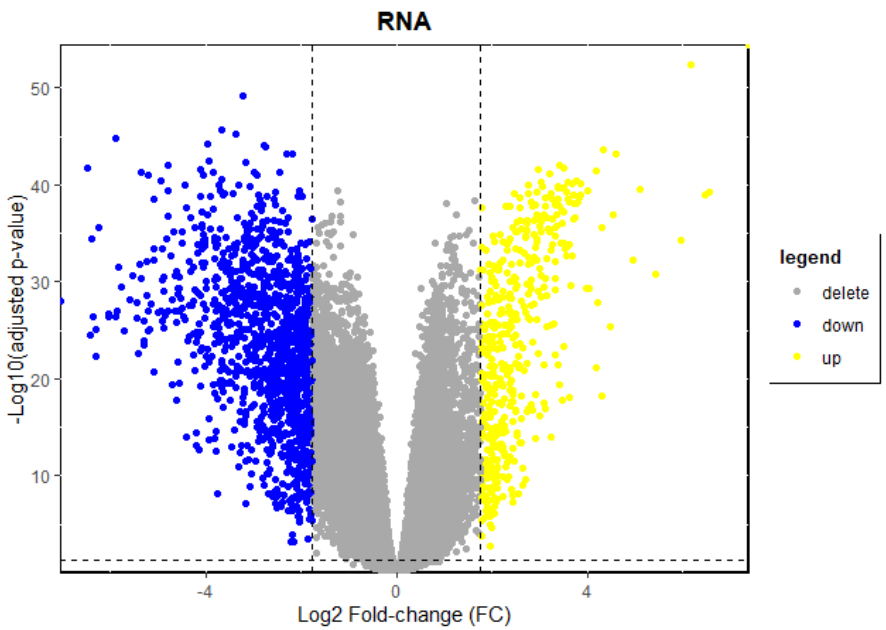


- x-axis represents the FC
- y-axis represents the FDR adjusted p-values (-log10 of the p-values) of the Student's t-test
- vertical dashed red lines and the horizontal ones correspond to the selected thresholds for the fold-change and the adjusted p-values
- grey points represent discarded according to the selected threshold
- blue points are the down-regulated RNAs
- yellow points are the up-regulated RNAs

Statistical test

- Since studied subjects, for many they are, always represent a small sample of all those potentially eligible and due to biological variability, **the results of a statistical test will always be expressed in terms of probability**
- This is because the sample we studied may not be representative of the universe of patients and the conclusions we reach may therefore be erroneous
- In accepting or reject the null hypothesis it is therefore always possible to make a mistake
- However, if the probability of making such an error is very low, we will accept with sufficient confidence the conclusions we arrived

Volcano plot



- x-axis represents the FC
- y-axis represents the FDR adjusted p-values (-log10 of the p-values) of the Student's t-test
- vertical dashed red lines and the horizontal ones correspond to the selected thresholds for the fold-change and the adjusted p-values
- grey points represent discarded according to the selected threshold
- blue points are the down-regulated RNAs
- yellow points are the up-regulated RNAs

Statistical test: p-value

There is always the risk that the decision to reject H_0 is wrong

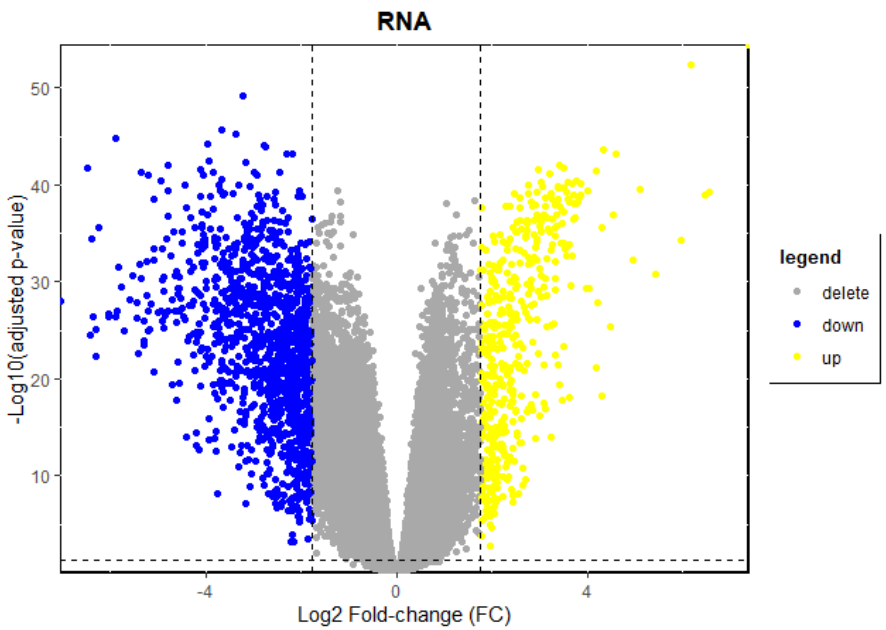
That is, there is always the risk of stating that **"the difference exists"**

When **"the difference does not exist"**

The measurement of this risk is called **level of significance**

The level of significance can be chosen arbitrarily (usually chosen equal to 0.05 or 0.01)	The lowest value at which H_0 can be rejected is called the **p-value**

Volcano plot



- x-axis represents the FC
- y-axis represents the FDR adjusted p-values (-log10 of the p-values) of the Student's t-test
- vertical dashed red lines and the horizontal ones correspond to the selected thresholds for the fold-change and the adjusted p-values
- grey points represent discarded according to the selected threshold
- blue points are the down-regulated RNAs
- yellow points are the up-regulated RNAs

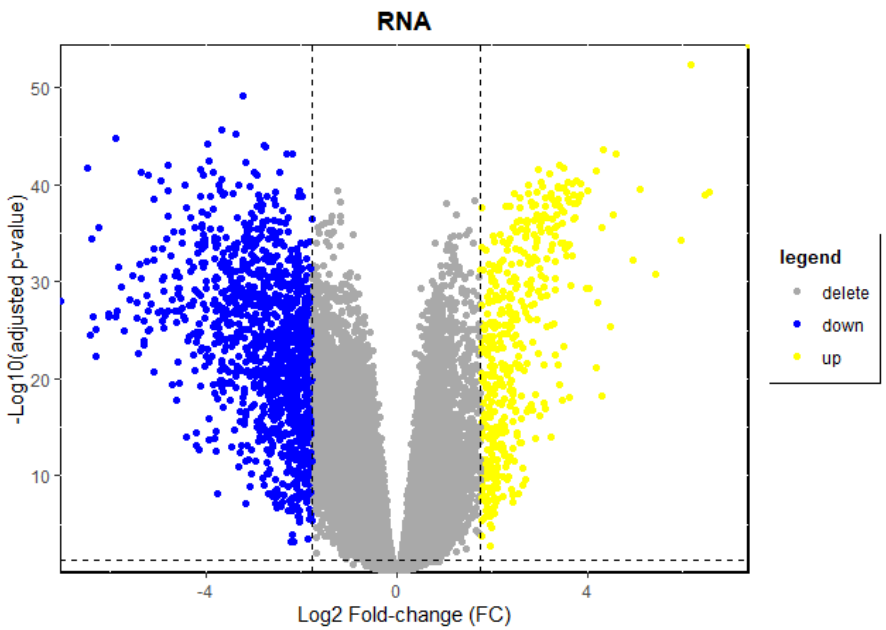
Statistical test: p-value

- Indicates the smallest probability of making a mistake by rejecting the null hypothesis (H_0), i.e. the probability of making a mistake by stating that there is a difference between the groups being compared

$p \leq 0.05$ → Statistically significant result → Reject H_0

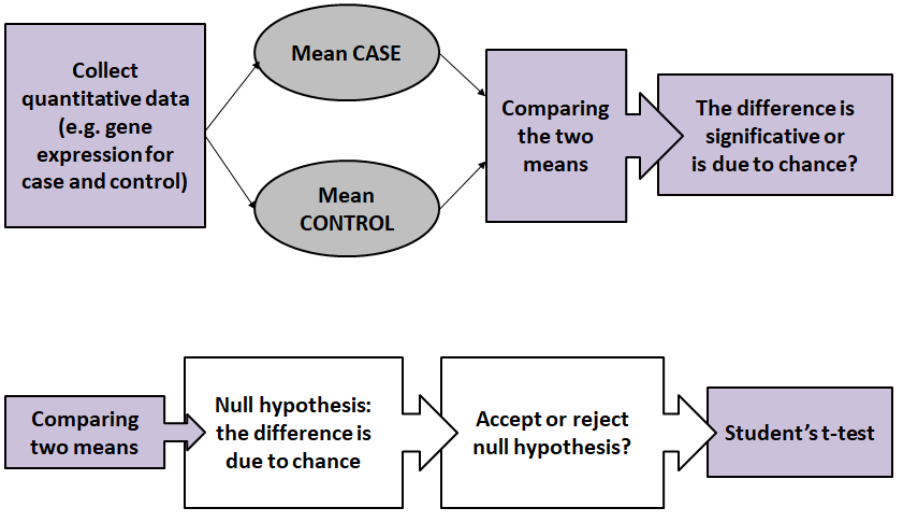
Probability of making a mistake rejecting H_0 is lower than 5%

Volcano plot

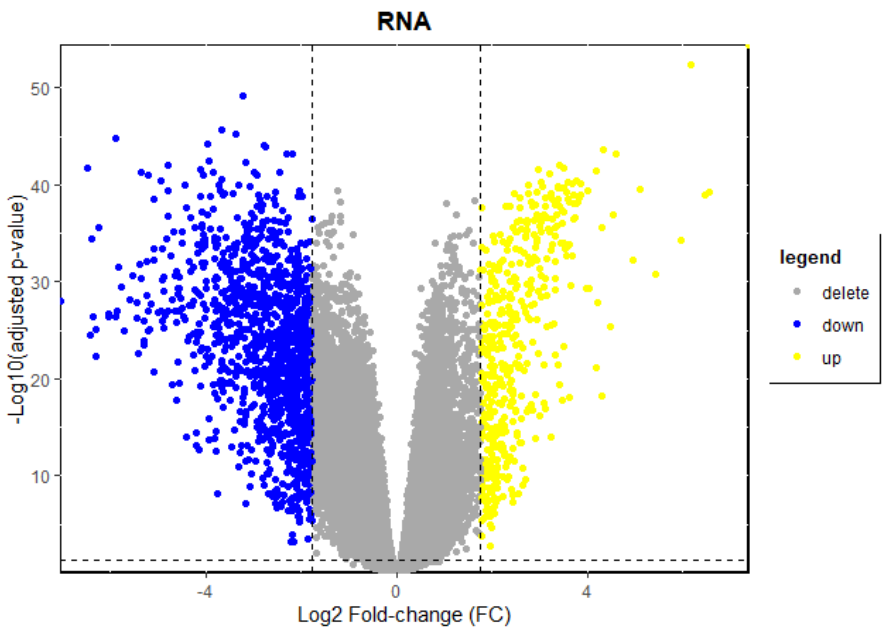


- x-axis represents the FC
- y-axis represents the FDR adjusted p-values (-log10 of the p-values) of the Student's t-test
- vertical dashed red lines and the horizontal ones correspond to the selected thresholds for the fold-change and the adjusted p-values
- grey points represent discarded according to the selected threshold
- blue points are the down-regulated RNAs
- yellow points are the up-regulated RNAs

Statistical test: Student's t-test



Volcano plot



- x-axis represents the FC
- y-axis represents the FDR adjusted p-values (-log10 of the p-values) of the Student's t-test
- vertical dashed red lines and the horizontal ones correspond to the selected thresholds for the fold-change and the adjusted p-values
- grey points represent discarded according to the selected threshold
- blue points are the down-regulated RNAs
- yellow points are the up-regulated RNAs

Statistical test: Student's t-test

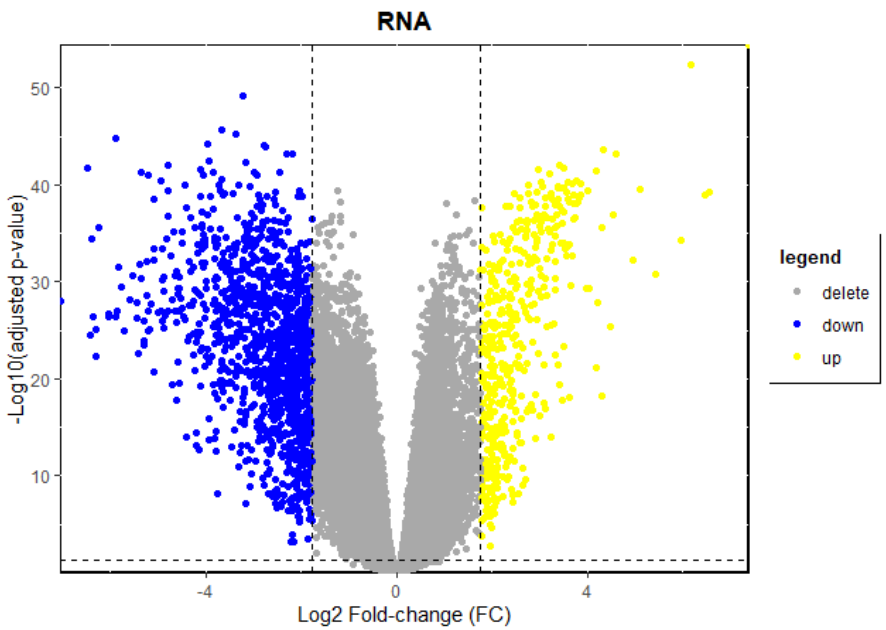
$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

is the standard deviation of $\bar{X}$

- Used If the distribution of X is approximately normal
- Used when the standard deviation is unknown
- Used if small sample size
- Can also be used for comparing two samples

$$t = \frac{\bar{x}_a - \bar{x}_b}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}}$$

Volcano plot

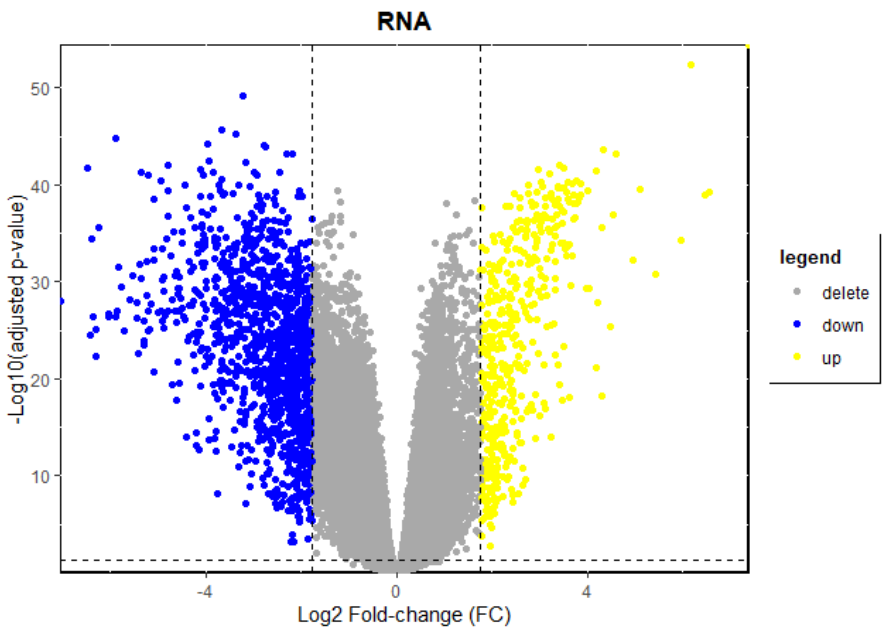


- x-axis represents the FC
- y-axis represents the FDR adjusted p-values (-log10 of the p-values) of the Student's t-test
- vertical dashed red lines and the horizontal ones correspond to the selected thresholds for the fold-change and the adjusted p-values
- grey points represent discarded according to the selected threshold
- blue points are the down-regulated RNAs
- yellow points are the up-regulated RNAs

Statistical test: Student's t-test

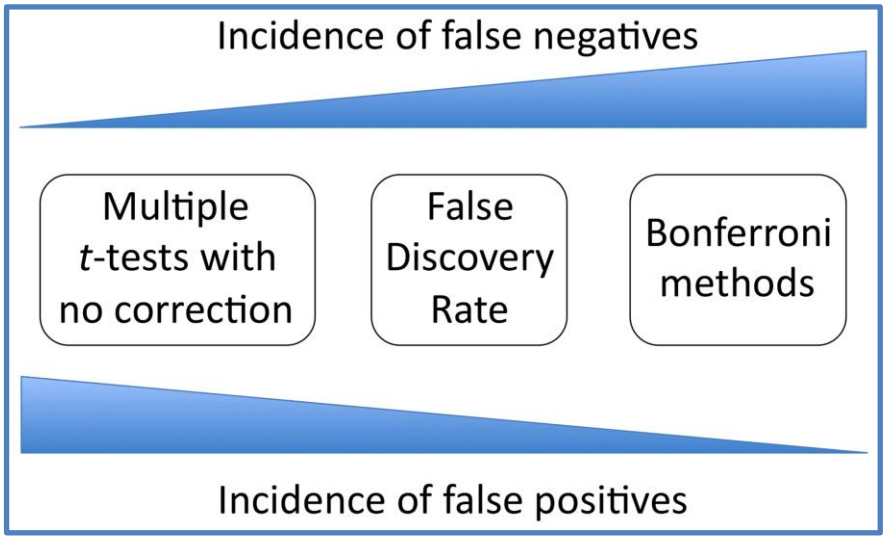
```
computeStat <- function(log_data,N,M,paired,method,output_file_stat_dataORIG){

  ########################################
  # input parameters

  data <- log_data$data
  data_control <- log_data$data_control
  data_case <- log_data$data_case
  ########################################

  IQR <- apply(data,1,IQR,type=5)

  perc_zeros <- apply(data, 1, function(x){ length(which(x == 0)) / length(x) *
100 })

  logFC <- rowMeans(data_case) - rowMeans(data_control)

  pval <- apply(data, 1, function(data){
    t.test(data[1:N], data[(N+1):M], paired = paired)$p.value
  })

  pval_adj <- p.adjust(pval, method = method)

  df_stat <- data.frame(IQR = IQR,
                        perc_zeros = perc_zeros,
                        logFC = logFC,
                        pval = pval,
                        pval_adj = pval_adj)

  write.table(df_stat, output_file_stat_dataORIG, row.names = T, col.names = NA,
sep = "\t", quote = F)

  return(df_stat)

}
```

Volcano plot

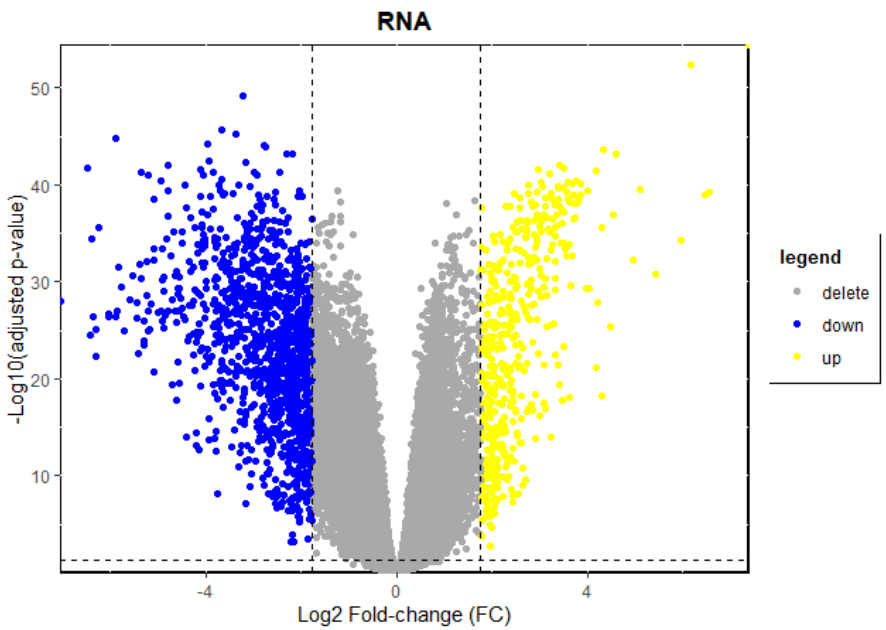


- x-axis represents the FC
- y-axis represents the FDR adjusted p-values (-log10 of the p-values) of the Student's t-test
- vertical dashed red lines and the horizontal ones correspond to the selected thresholds for the fold-change and the adjusted p-values
- grey points represent discarded according to the selected threshold
- blue points are the down-regulated RNAs
- yellow points are the up-regulated RNAs

Statistical test: Multiple tests correction



Volcano plot

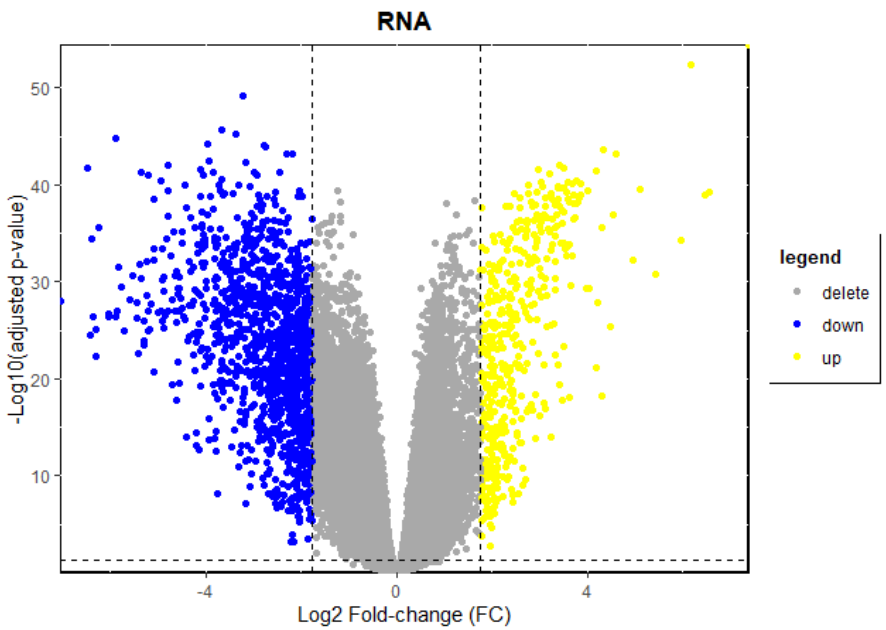


- x-axis represents the FC
- y-axis represents the FDR adjusted p-values (-log10 of the p-values) of the Student's t-test
- vertical dashed red lines and the horizontal ones correspond to the selected thresholds for the fold-change and the adjusted p-values
- grey points represent discarded according to the selected threshold
- blue points are the down-regulated RNAs
- yellow points are the up-regulated RNAs

Statistical test: Multiple tests correction

```
computeStat <- function(log_data,N,M,paired,method,output_file_stat_dataORIG){

  ########################################
  # input parameters

  data <- log_data$data
  data_control <- log_data$data_control
  data_case <- log_data$data_case
  ########################################

  IQR <- apply(data,1,IQR,type=5)

  perc_zeros <- apply(data, 1, function(x){ length(which(x == 0)) / length(x) *
100 })

  logFC <- rowMeans(data_case) - rowMeans(data_control)

  pval <- apply(data, 1, function(data){
    t.test(data[1:N], data[(N+1):M], paired = paired)$p.value
  })

  pval_adj <- p.adjust(pval, method = method)

  df_stat <- data.frame(IQR = IQR,
                        perc_zeros = perc_zeros,
                        logFC = logFC,
                        pval = pval,
                        pval_adj = pval_adj)

  write.table(df_stat, output_file_stat_dataORIG, row.names = T, col.names = NA,
sep = "\t", quote = F)

  return(df_stat)

}
```

Volcano plot

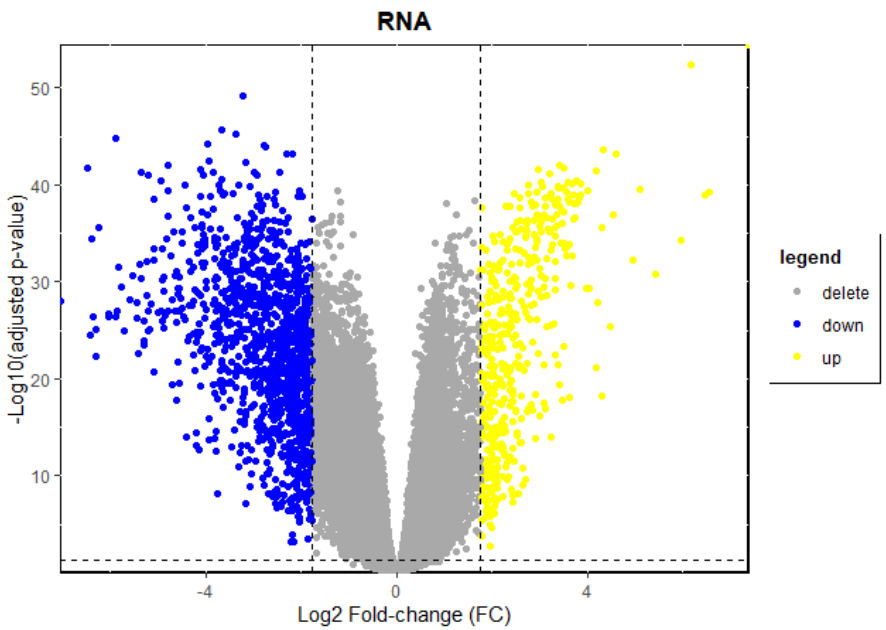


- x-axis represents the FC
- y-axis represents the FDR adjusted p-values (-log10 of the p-values) of the Student's t-test
- vertical dashed red lines and the horizontal ones correspond to the selected thresholds for the fold-change and the adjusted p-values
- grey points represent discarded according to the selected threshold
- blue points are the down-regulated RNAs
- yellow points are the up-regulated RNAs

Volcano plot

- A **volcano plot** is a type of scatterplot that shows statistical significance (p- value) versus magnitude of change (fold change)
- It enables quick visual identification of genes with large fold changes that are also statistically significant

Volcano plot

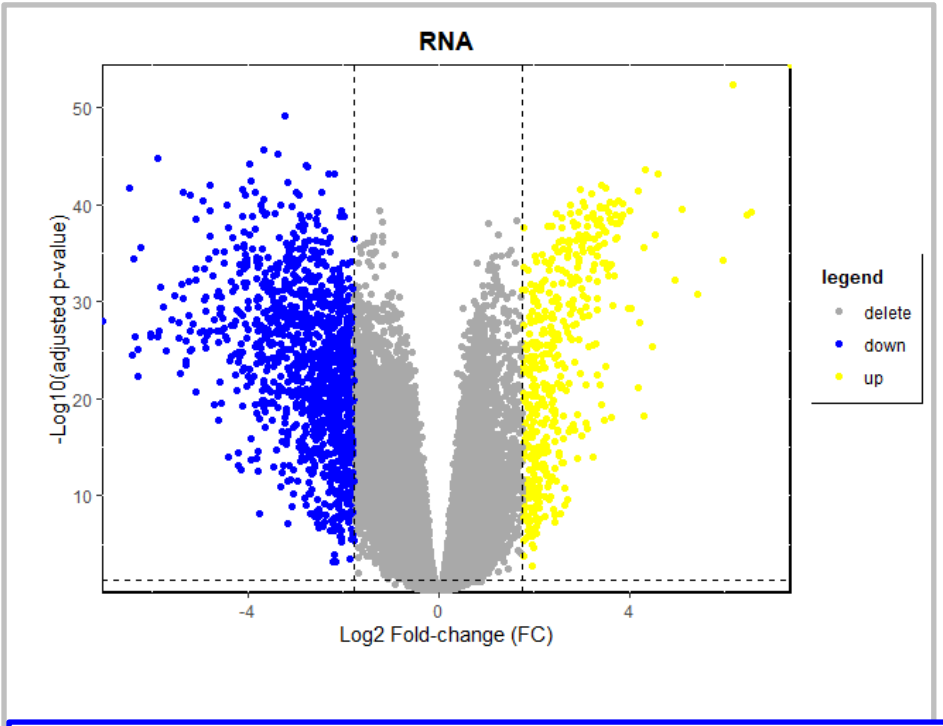


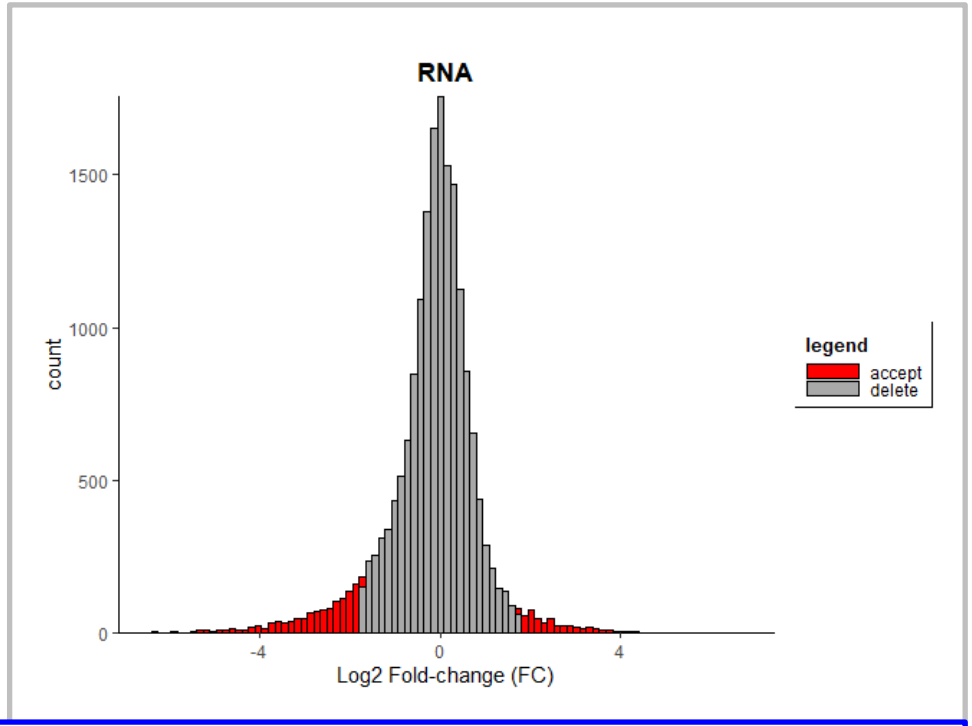
- x-axis represents the FC
- y-axis represents the FDR adjusted p-values (-log10 of the p-values) of the Student's t-test
- vertical dashed red lines and the horizontal ones correspond to the selected thresholds for the fold-change and the adjusted p-values
- grey points represent discarded according to the selected threshold
- blue points are the down-regulated RNAs
- yellow points are the up-regulated RNAs

Volcano plot

```
getVolcanoPlot <- function(logFC,pval_adj,threshold_fc,threshold_pval_adj,title,output_fil
e){

  df <- data.frame(logFC = logFC, pval = -log10(pval_adj))

  condition1 <- (pval_adj <= threshold_pval_adj) & (logFC > log2(threshold_fc))
  condition2 <- (pval_adj <= threshold_pval_adj) & (logFC < -log2(threshold_fc))

  df$legend <- ifelse(condition1,"up",ifelse(condition2,"down","delete"))

  p <- ggplot(df, aes(x = logFC, y = pval, color = legend)) + geom_point() +
    scale_x_continuous(expand = c(0, 0)) + scale_y_continuous(expand = c(0, 0)) +
    scale_color_manual(values = c("delete" = "darkgrey", "up" = "yellow", "down" = "blue"
))  +
    theme(panel.background = element_rect(fill = "white", colour = "black", size = 1),
          plot.title = element_text(hjust = 0.5, face = "bold"),
          legend.title = element_text(colour = "black", size=10, face="bold"),
          legend.key = element_rect(fill = "white", colour = "white"),
          legend.box.background = element_rect(colour = "black")) +
    labs(title = title, x = "Log2 Fold-change (FC)", y = "-Log10(adjusted p-value)") +
    geom_hline(yintercept =  -log10(threshold_pval_adj), linetype = "dashed", color =
"black") +
    geom_vline(xintercept = log2(threshold_fc), linetype = "dashed", color = "black") +
    geom_vline(xintercept = -log2(threshold_fc), linetype = "dashed", color = "black")

  print(p)

  savePDF(p,output_file)

}
```

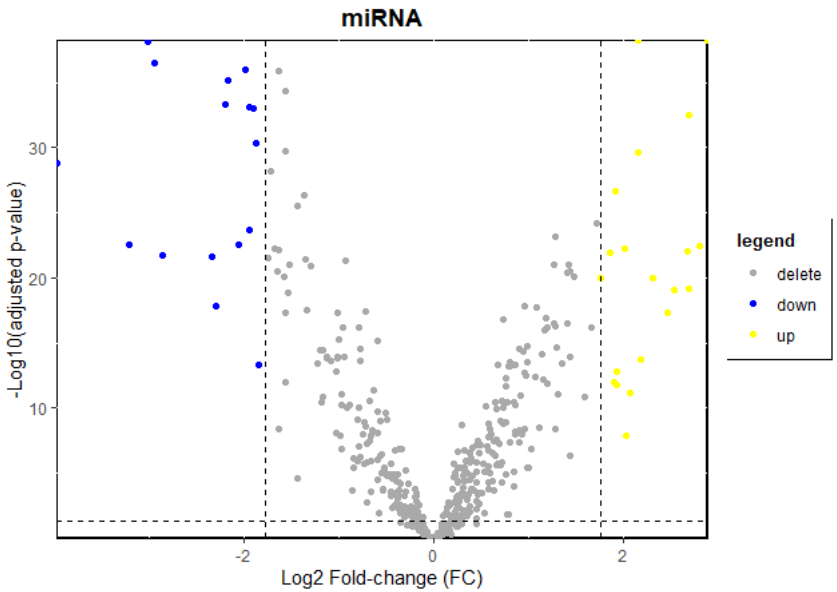
Volcano plot and FC distribution for RNAs



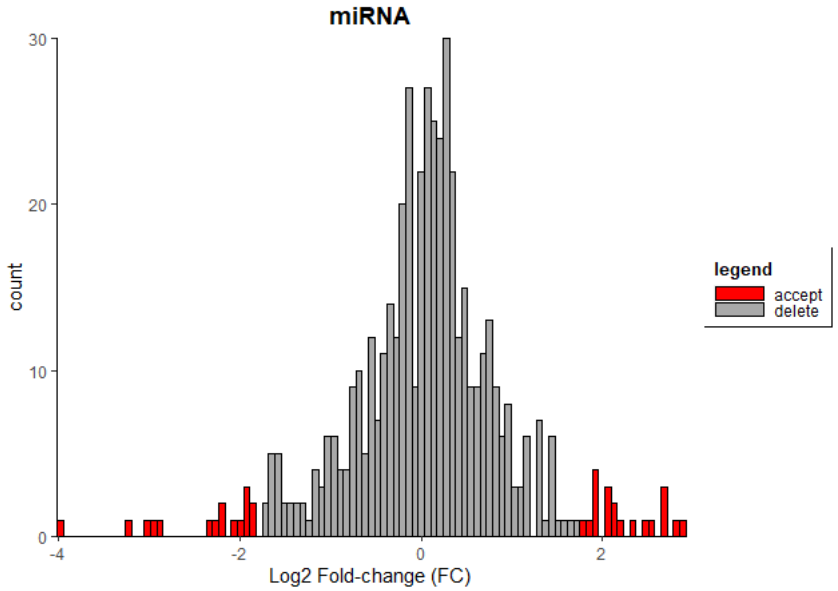


Looking at these plots you can understand which thresholds modify to have more or less differentially expressed genes. A reasonable choice can be to obtain about the 10% of the starting number of genes

Volcano plot and FC distribution for miRNAs

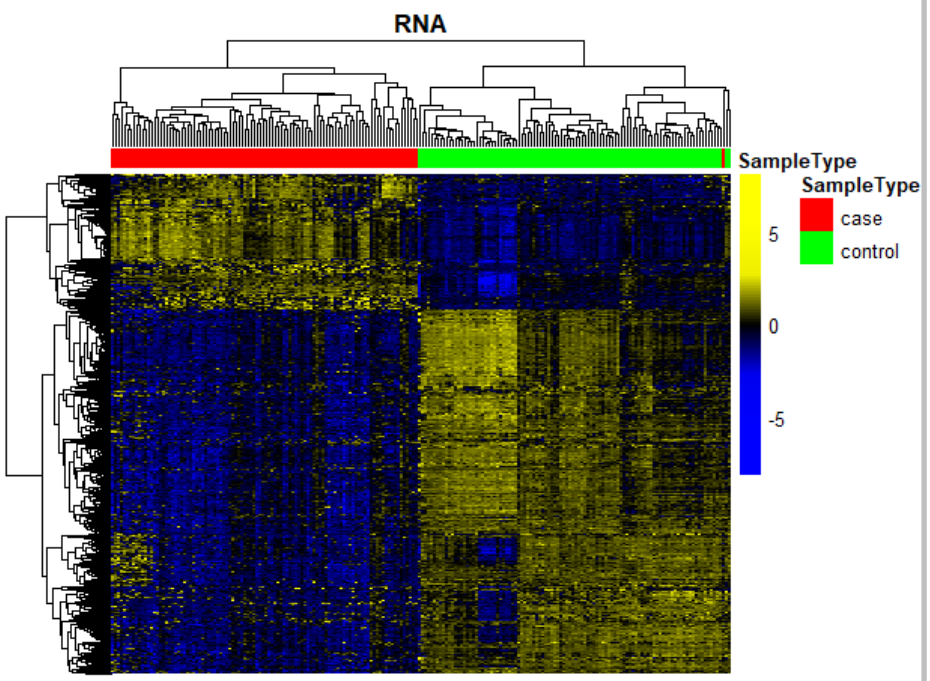


- x-axis represents the FC
- y-axis represents the FDR adjusted p-values (-log10 of the p-values) of the Student's t-test
- vertical dashed red lines and the horizontal ones correspond to the selected thresholds for the fold-change and the adjusted p-values
- grey points represent discarded according to the selected threshold
- blue points are the down-regulated miRNAs
- yellow points are the up-regulated miRNAs



- x-axis represents the fold-change (logarithmic scale)
- y-axis represents the frequency of the obtained fold-change values
- grey bars represent the miRNAs discarded, red bars the retained ones

Heatmap of differentially expressed RNAs



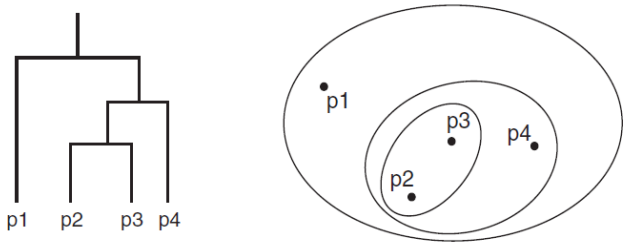
- Row refers to differentially expressed RNAs
- Columns refers to samples
- Colors represent different expression levels that increase from blue to yellow

Heatmap

- a data visualization in the form of a map in which data matrix values are represented as colors
- It is used to represent expression level of genes across samples

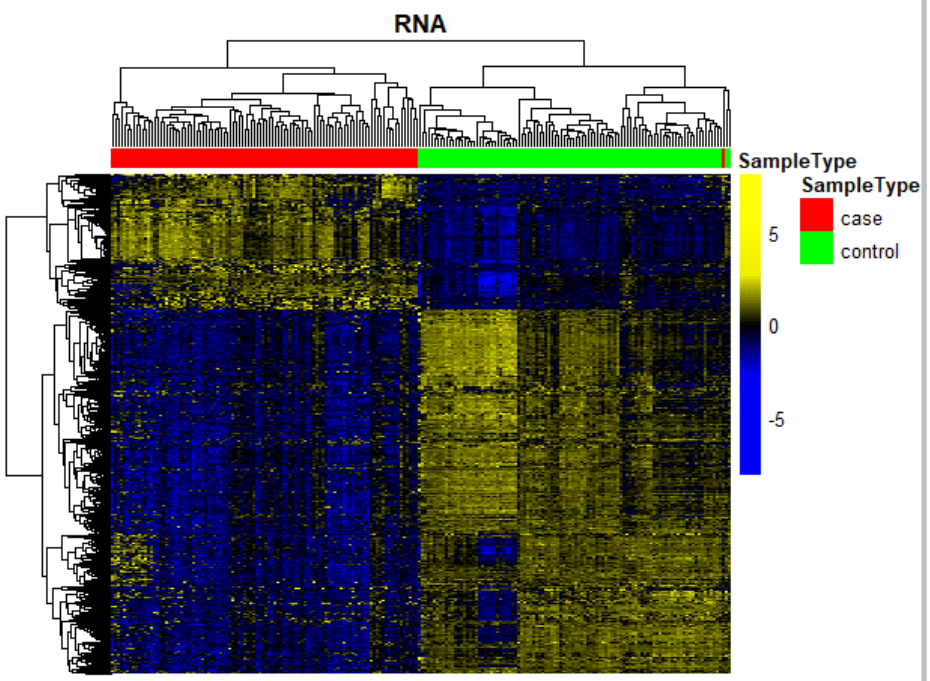
Dendrogram

- Diagram representing a tree that illustrates the arrangement of the clusters produced by the hierarchical clustering



(a) Dendrogram. (b) Nested cluster diagram.

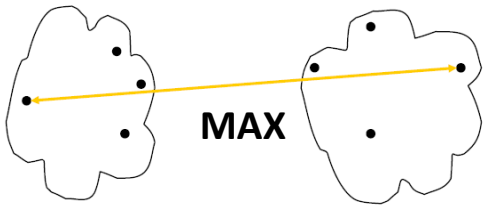
Heatmap of differentially expressed RNAs



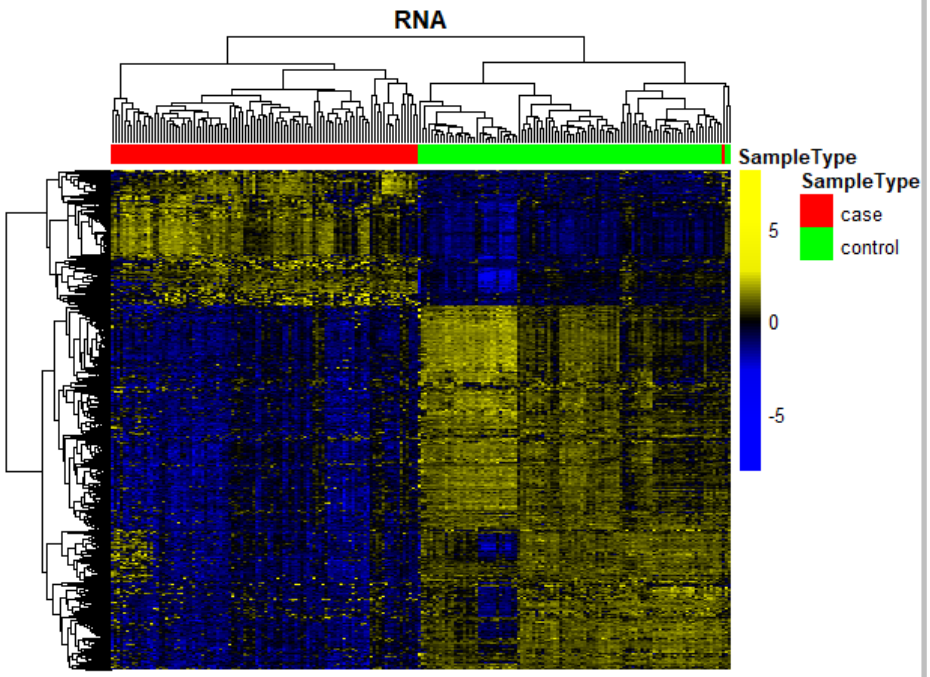
- Row refers to differentially expressed RNAs
- Columns refers to samples
- Colors represent different expression levels that increase from blue to yellow

Hierarchical clustering

- Hierarchical clustering for rows and columns of data matrix by using:
  - ❖ **Pearson correlation** as distance metric
  - ❖ linkage **complete** as clustering method (where distance is measured between the farthest pair of observations in two clusters)



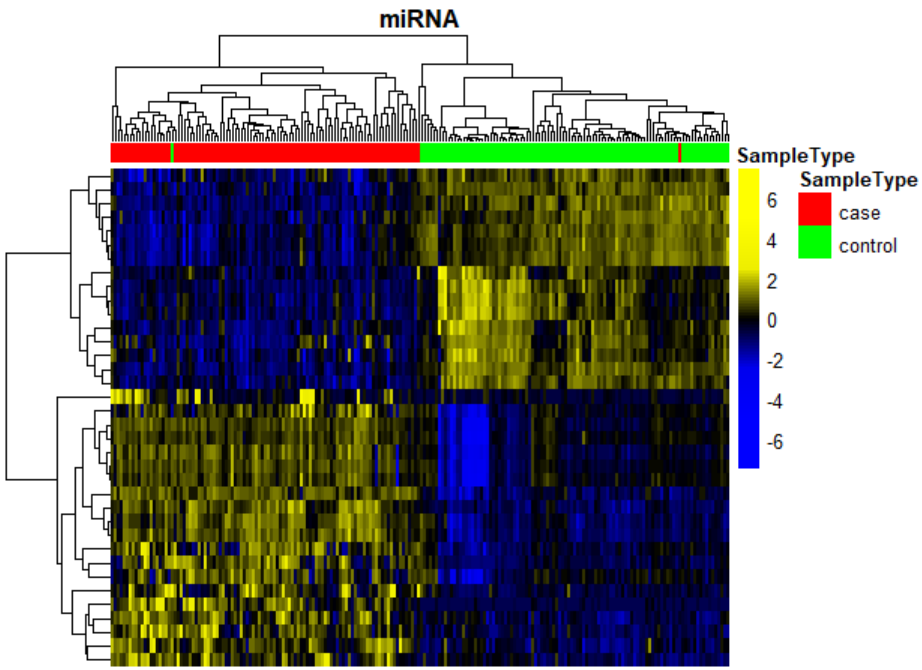
Heatmap of differentially expressed RNAs



- Row refers to differentially expressed RNAs
- Columns refers to samples
- Colors represent different expression levels that increase from blue to yellow

```
getHeatmap <- function(data.Filtered,output_file,title){

  ########################################
  # input parameters

  data <- data.Filtered$data
  control <- colnames(data.Filtered$data_control)
  case <- colnames(data.Filtered$data_case)
  ########################################
  samples <- ifelse( (colnames(data) %in% control), "control", "case"
)
  annotation <- data.frame(SampleType = samples)
  rownames(annotation) <- colnames(data)

  annotation_colors <- list(SampleType = c(case = "red", control =
"green"))

  colorbar <- colorRampPalette(colors = c("blue","blue1","blue2"
,"black","yellow2","yellow1","yellow"))(100)

  out <- pheatmap(data, scale = "row",
                  border_color = NA,
                  clustering_distance_rows = "correlation",
                  clustering_distance_cols = "correlation",
                  clustering_method = "complete",
                  cluster_cols = T,
                  cluster_rows = T,
                  annotation_col = annotation,
                  annotation_colors = annotation_colors,
                  color = colorbar,
                  show_rownames = F,
                  show_colnames = F,
                  main = title
                  #width = 10,
                  #height = 10,
                  #treeheight_row = 30,
                  #fontsize = 10,
                  #cellwidth = 0.3,
                  #cellheight = 0.3
  )

  saveHeatmapPDF(out,output_file)

}
```

Heatmap of differentially expressed miRNAs



- Row refers to differentially expressed miRNAs
- Columns refers to samples
- Colors represent different expression levels that increase from blue to yellow

```
getHeatmap <- function(data.Filtered,output_file,title){

  ########################################
  # input parameters

  data <- data.Filtered$data
  control <- colnames(data.Filtered$data_control)
  case <- colnames(data.Filtered$data_case)
  ########################################
  samples <- ifelse( (colnames(data) %in% control), "control", "case"
)
  annotation <- data.frame(SampleType = samples)
  rownames(annotation) <- colnames(data)

  annotation_colors <- list(SampleType = c(case = "red", control =
"green"))

  colorbar <- colorRampPalette(colors = c("blue","blue1","blue2"
,"black","yellow2","yellow1","yellow"))(100)

  out <- pheatmap(data, scale = "row",
                  border_color = NA,
                  clustering_distance_rows = "correlation",
                  clustering_distance_cols = "correlation",
                  clustering_method = "complete",
                  cluster_cols = T,
                  cluster_rows = T,
                  annotation_col = annotation,
                  annotation_colors = annotation_colors,
                  color = colorbar,
                  show_rownames = F,
                  show_colnames = F,
                  main = title
                  #width = 10,
                  #height = 10,
                  #treeheight_row = 30,
                  #fontsize = 10,
                  #cellwidth = 0.3,
                  #cellheight = 0.3
  )

  saveHeatmapPDF(out,output_file)

}
```

At the end of Module 1, you will obtain the differentially expressed genes