Q2 Functional Genomics

Pierrot Van der Aa

2024-04-01

We first load the morphological counts and check if every columns only contain numbers ("double"). Then we load the clinical data. We replace the occurrences of 99 in the column DTHVNT by NA because that's what it means. We then take off patients with NA's because they are not handled further in the analyses. Given that there are only four of them, we can delete them without impacting too much the analysis. We turn SEX into a factor and scale the numerical clinical data (AGE, HGHT, WGHT and BMI). We apply similar procedure to technical features (i.e. factorising DTHVNT, COHORT and DTHHRDY and scaling TRISCHD).

```r
#Load morphological counts
count_data <- as.data.frame(read_tsv("Data_ThyroidGland/OG/morphological_counts_lunit_dino.tsv"))
sapply(count_data, typeof)
```

```
##                     SMPLID  Mophological.cluster.G4_0
##                "character"                   "double"
##  Mophological.cluster.G4_1  Mophological.cluster.G4_2
##                   "double"                   "double"
##  Mophological.cluster.G4_3  Mophological.cluster.G4_4
##                   "double"                   "double"
##  Mophological.cluster.G4_5  Mophological.cluster.G4_6
##                   "double"                   "double"
##  Mophological.cluster.G4_7  Mophological.cluster.G4_8
##                   "double"                   "double"
##  Mophological.cluster.G4_9 Mophological.cluster.G4_10
##                   "double"                   "double"
## Mophological.cluster.G4_11 Mophological.cluster.G4_12
##                   "double"                   "double"
## Mophological.cluster.G4_13 Mophological.cluster.G4_14
##                   "double"                   "double"
## Mophological.cluster.G4_15 Mophological.cluster.G4_16
##                   "double"                   "double"
## Mophological.cluster.G4_17 Mophological.cluster.G4_18
##                   "double"                   "double"
## Mophological.cluster.G4_19 Mophological.cluster.G4_20
##                   "double"                   "double"
## Mophological.cluster.G4_21 Mophological.cluster.G4_22
##                   "double"                   "double"
## Mophological.cluster.G4_23 Mophological.cluster.G4_24
##                   "double"                   "double"
## Mophological.cluster.G4_25 Mophological.cluster.G4_26
##                   "double"                   "double"
## Mophological.cluster.G4_27 Mophological.cluster.G4_28
##                    "double"                   "double"
## Mophological.cluster.G4_29 Mophological.cluster.G4_30
##                    "double"                   "double"
## Mophological.cluster.G4_31
##                    "double"
```

```
#Load clinical data
clinic_data <- as.data.frame(read_tsv("Data_ThyroidGland/OG/clinical_data.tsv"))
colnames(clinic_data)
```

```
##  [1] "SMPLID"   "SUBJID"   "COHORT"   "SEX"      "AGE"      "HGHT"
##  [7] "WGHT"     "BMI"      "SMPTHNTS" "TRISCHD"  "DTHVNT"   "DTHHRDY"
## [13] "SMPLID.1" "IMGURL"
```

```
#clinical data preprocessing
clinic_data$SEX <- as.factor(clinic_data$SEX)
clinic_data$AGE <- scale(clinic_data$AGE)
clinic_data$HGHT <- scale(clinic_data$HGHT)
clinic_data$WGHT <- scale(clinic_data$WGHT)
clinic_data$BMI <- scale(clinic_data$BMI)
#getting rid of the NA's
clinic_data$DTHVNT <- replace(clinic_data$DTHVNT, clinic_data$DTHVNT == 99, NA) #replace the 99 by NA g
sapply(clinic_data, function(x) which(is.na(x))) #see where are the NAs
```

```
## $SMPLID
## integer(0)
##
## $SUBJID
## integer(0)
##
## $COHORT
## integer(0)
##
## $SEX
## integer(0)
##
## $AGE
## integer(0)
##
## $HGHT
## integer(0)
##
## $WGHT
## integer(0)
##
## $BMI
## integer(0)
##
## $SMPTHNTS
## integer(0)
##
## $TRISCHD
## integer(0)
```

```
##
## $DTHVNT
## [1] 196 219 239 260
##
## $DTHHRDY
## integer(0)
##
## $SMPLID.1
## integer(0)
##
## $IMGURL
## integer(0)
```

```
NAs_ID <- c(clinic_data$SMPLID[[196]], clinic_data$SMPLID[[219]], clinic_data$SMPLID[[239]], clinic_data
count_data <- subset(count_data, !(count_data$SMPLID %in% NAs_ID)) #remove the four occurences of NA in
clinic_data <- drop_na(clinic_data) #remove the four occurences of NA in the clinic dataset
#technical data preprocessing
clinic_data$DTHVNT <- as.factor(clinic_data$DTHVNT)
clinic_data$COHORT <- as.factor(clinic_data$COHORT)
clinic_data$DTHHRDY <- as.factor(clinic_data$DTHHRDY)
clinic_data$TRISCHD <- scale(clinic_data$TRISCHD)
sapply(clinic_data, typeof)
```

```
##      SMPLID      SUBJID      COHORT         SEX         AGE        HGHT
## "character" "character"   "integer"   "integer"    "double"    "double"
##        WGHT         BMI    SMPTHNTS     TRISCHD      DTHVNT     DTHHRDY
##    "double"    "double" "character"    "double"   "integer"   "integer"
##    SMPLID.1      IMGURL
## "character" "character"
```

This is a chunck to prepare the data to merge them into a DESeq object.

```
# First we use the sample ID as rownames in the two dataframes
count <- count_data[,-1]
rownames(count) <- count_data[,1]
clinic <- clinic_data[,-1]
rownames(clinic) <- clinic_data[,1]
# Then we make sure that all the rows are in the same order in both dataframes
all(rownames(count) == rownames(clinic))
```

```
## [1] TRUE
```

```
# For the sake of the formula, we need to transpose the count data
count <- t(count)
```

Q2.1

Compute systematically associations between clinical variables and morphological cluster counts. The purpose is to compare the magnitude of the associations of the different variables with morphology.

For each clinical variable (age, sex, height, weight and BMI), we compute the association of the variable with the cluster count. So the association between a clinical variable and the morphology.

```r
health_names <- c("AGE", "SEX", "HGHT", "WGHT", "BMI")
formul <- ''
res <- ''
for(param in health_names){
  print(param)
  formul <- as.formula(paste("~",param, sep = ""))
  res <- paste("res", param, sep = "")
  dds <-  DESeqDataSetFromMatrix(countData = count, colData = clinic, design = formul)
  dds_res <- DESeq(dds)
  assign(res, results(dds_res))
}
```

```
## [1] "AGE"
## [1] "SEX"
## [1] "HGHT"
## [1] "WGHT"
## [1] "BMI"
```

The following plots show, in blue, the cluster which are up- or down-regulated. However, given that we talk about morphological cluster counts and not gene expression, the blue dot represent morphological clusters which present a significant correlation with the clinical variable. The closer a point is to the 0 log fold change line, the more likely is this point to have a non-significant interaction with the clinical variable.

(copy paste from chat GPT for information:)

"In differential expression analysis, an MA plot is a commonly used visualization tool to assess the relationship between the magnitude of gene expression differences (M) and the average expression level (A) across conditions or treatments. Here's how to interpret an MA plot:

M-Axis (Magnitude of Differential Expression):

- The M-axis represents the magnitude of differential expression between two conditions or treatments. It is calculated as the log fold change (logFC) in expression levels between conditions.
- Each point on the M-axis corresponds to a gene or transcript, with genes upregulated in one condition relative to another plotted above the zero line and genes downregulated plotted below the zero line.
- The higher the absolute value of M, the greater the magnitude of differential expression. Genes with larger M-values are considered to have more substantial changes in expression between conditions.

A-Axis (Average Expression Level):

- The A-axis represents the average expression level of genes across conditions or treatments. It is calculated as the average of the log-transformed expression values for each gene.
- Each point on the A-axis corresponds to a gene or transcript, with genes plotted based on their average expression level across conditions.
- The A-axis provides information about the abundance or intensity of gene expression. Genes with higher average expression levels are typically located towards the center of the plot, while genes with lower expression levels are towards the edges.

Patterns and Features in the MA Plot:

- Vertical Spread: The vertical spread of points along the M-axis indicates the variability or dispersion in gene expression differences across the dataset. A wider spread suggests greater variability in fold changes between conditions.

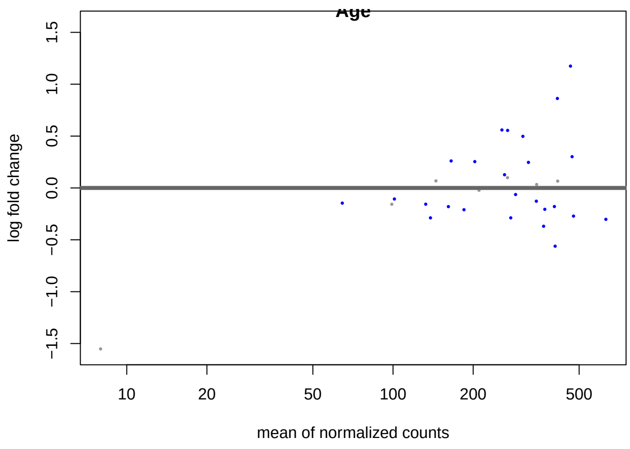
- Horizontal Position: The horizontal position of points along the A-axis reflects the average expression level of genes. Patterns such as curvature or slope in the plot may indicate systematic biases or technical artifacts in the data.
- Central Tendency: The central tendency of points in the MA plot can provide insights into the overall distribution of gene expression changes. Clustering of points around the zero line suggests no significant differential expression, while points deviating from the zero line indicate genes with differential expression.
- Outliers: Outlying points in the plot represent genes with unusually large or small fold changes compared to the majority of genes. These outliers may represent biologically interesting candidates for further investigation.

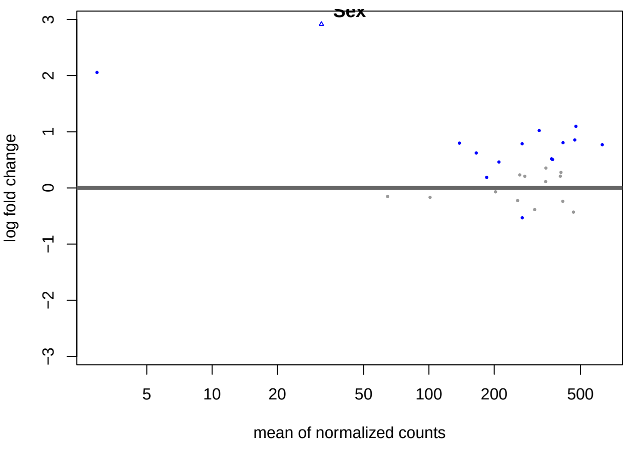
Interpretation:

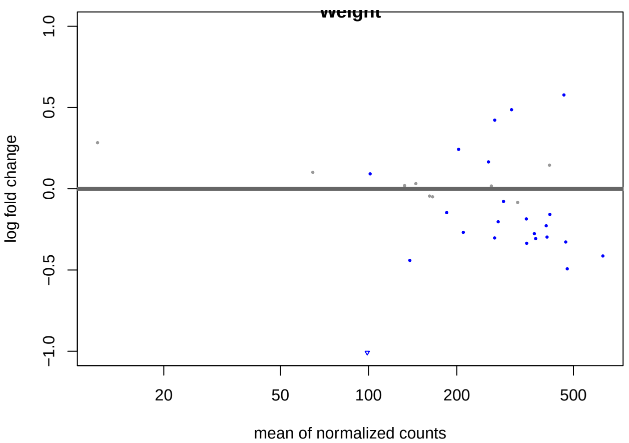
- Upregulated genes are represented by points above the zero line on the M-axis, indicating higher expression in one condition compared to another.
- Downregulated genes are represented by points below the zero line on the M-axis, indicating lower expression in one condition compared to another.
- Genes with no significant differential expression will cluster around the zero line on the M-axis.

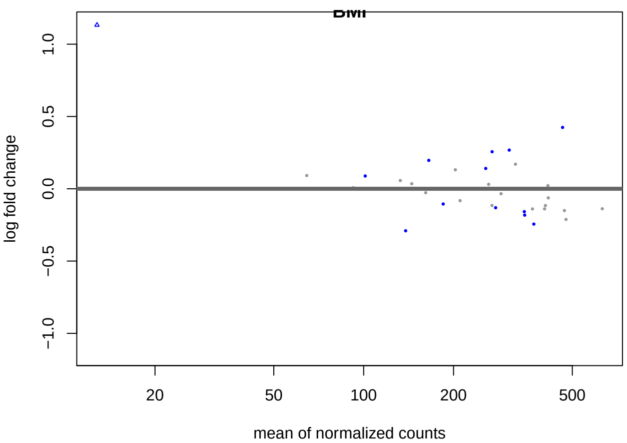
Overall, an MA plot provides a visual summary of gene expression changes between conditions, allowing researchers to identify differentially expressed genes, assess the magnitude of expression changes, and detect systematic biases or technical artifacts in the data.”

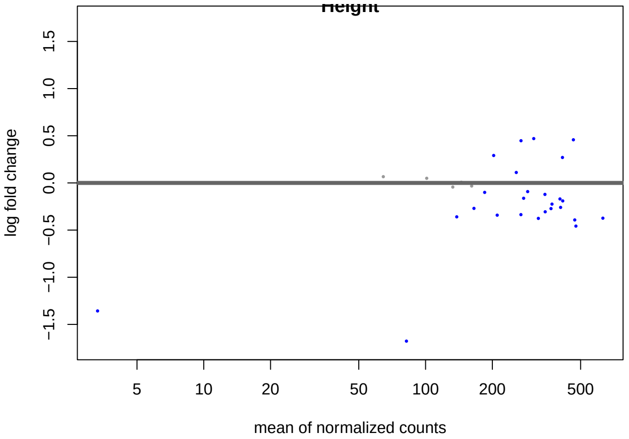
```r
par(mar = c(4, 4, .1, .1))
plotMA(resAGE, main = "Age")
plotMA(resSEX, main = "Sex")
plotMA(resWGHT, main = "Weight")
plotMA(resBMI, main = "BMI")
plotMA(resHGHT, main = "Height")
```











On all the MA plots here above, we can observe that one cluster has much lower mean normalized counts value.

```
resAGE[which.min(resAGE$baseMean),]
```

```
## log2 fold change (MLE): AGE
## Wald test p-value: AGE
## DataFrame with 1 row and 6 columns
##                              baseMean log2FoldChange     lfcSE      stat
##                             <numeric>      <numeric> <numeric> <numeric>
## Mophological.cluster.G4_23   7.97346       -1.55183  0.323588  -4.79569
##                                pvalue      padj
##                             <numeric> <numeric>
## Mophological.cluster.G4_23         NA        NA
```

```
resBMI[which.min(resBMI$baseMean),]
```

```
## log2 fold change (MLE): BMI
## Wald test p-value: BMI
## DataFrame with 1 row and 6 columns
##                              baseMean log2FoldChange     lfcSE      stat
##                             <numeric>      <numeric> <numeric> <numeric>
## Mophological.cluster.G4_23   12.7838        1.30147  0.345581   3.76604
##                                 pvalue       padj
##                              <numeric>  <numeric>
## Mophological.cluster.G4_23 0.00016586 0.00356575
```

```
resHGHT[which.min(resHGHT$baseMean),]
```

```
## log2 fold change (MLE): HGHT
## Wald test p-value: HGHT
## DataFrame with 1 row and 6 columns
##                             baseMean log2FoldChange     lfcSE      stat
##                            <numeric>      <numeric> <numeric> <numeric>
## Mophological.cluster.G4_23   3.32137        -1.3574  0.320407  -4.23648
##                                 pvalue        padj
##                              <numeric>   <numeric>
## Mophological.cluster.G4_23 2.27051e-05 0.000181641
```

```
resSEX[which.min(resSEX$baseMean),]
```

```
## log2 fold change (MLE): SEX 2 vs 1
## Wald test p-value: SEX 2 vs 1
## DataFrame with 1 row and 6 columns
##                             baseMean log2FoldChange     lfcSE      stat
##                            <numeric>      <numeric> <numeric> <numeric>
## Mophological.cluster.G4_23   2.94561        2.05723  0.688388   2.98848
##                                pvalue       padj
##                             <numeric>  <numeric>
## Mophological.cluster.G4_23 0.00280368 0.00815615
```

```
resWGHT[which.min(resWGHT$baseMean),]
```

```
## log2 fold change (MLE): WGHT
## Wald test p-value: WGHT
## DataFrame with 1 row and 6 columns
##                             baseMean log2FoldChange     lfcSE      stat
##                            <numeric>      <numeric> <numeric> <numeric>
## Mophological.cluster.G4_23   11.8861       0.283465  0.354501  0.799615
##                               pvalue      padj
##                            <numeric> <numeric>
## Mophological.cluster.G4_23        NA        NA
```

The code chunck here above allows to see that the morphological cluster concerned by this lower mean normalized count is always cluster G4_23. When we look at the pictures of that cluster on the morphological atlas, it appears that this cluster is made of portions of muscular tissue adjacent to the thyroid itself and does not display the purple coloration of the nuclei leading to a lower mean normalized count.

Q2.2

Discuss the association with technical variables.

NB: I base my answer on the file description_ClinicalScale.docx

From Q1, we know that the following associations exist between clinical and technical variables.

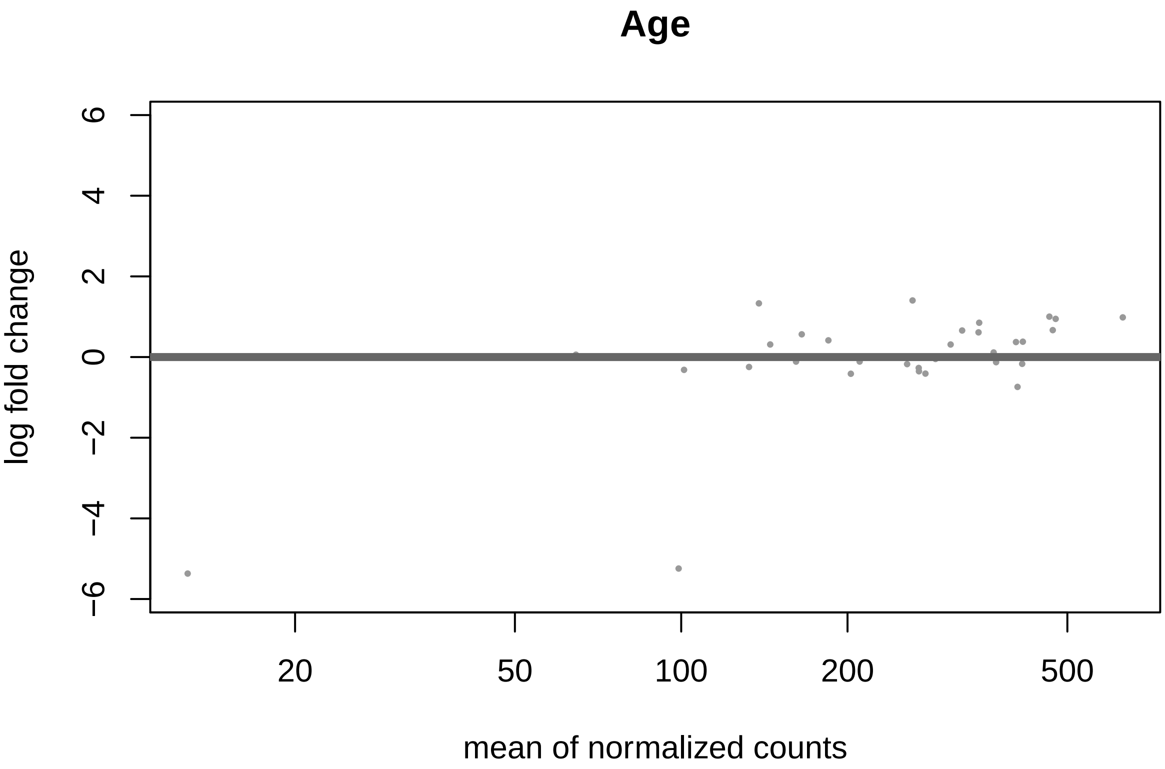
- for age, we have to correct for the ischemic time (TRISCHD), the presence of ventilator prior to death (c), the cohort (COHORT) and the type of death (DTHHRDY).

- for BMI, we do not have to correct for technical variables.
- for height, we have to correct for the cohort (COHORT) but we will consider that the mild correlation between height and ischemic time and between height and the presence of ventilator prior to death are not relevant enough to be included as confounding factors.
- for sex, we do not have to correct for the effects of technical variables. Again, we will ignore the mild correlations with the ischemic time and the presence of ventilator prior to death.
- for weight, we have to correct for the effect of the cohort (COHORT) but will neglect the mild correlation with the presence of a ventilator prior to death and the type of death.

Q2.3

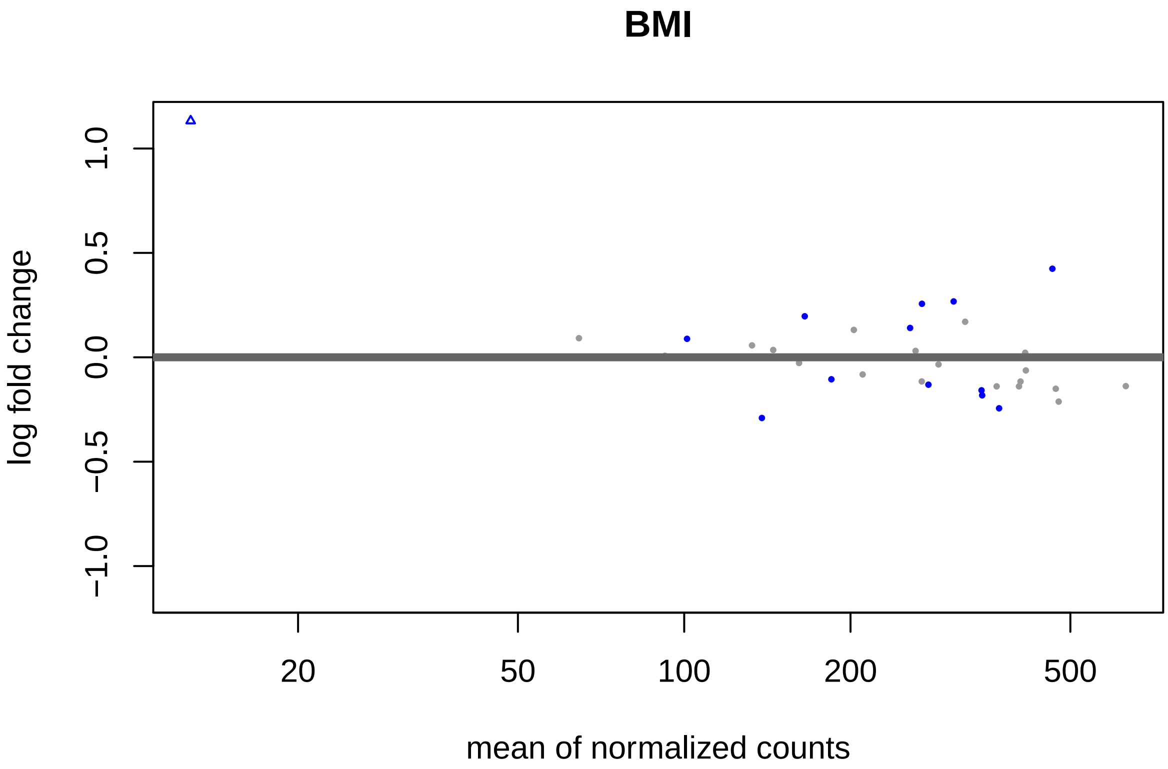
For non-technical variables, redo the analysis with adjustment for the confounding technical variables, if any is reported in Q2.2. Report and discuss significant associations.

```r
formul <- as.formula("~AGE+TRISCHD+DTHVNT+COHORT+DTHHRDY")
dds <-  DESeqDataSetFromMatrix(countData = count, colData = clinic, design = formul)
dds_res <- DESeq(dds)
resAGEcorr <- results(dds_res)
plotMA(resAGEcorr, main = "Age")
```

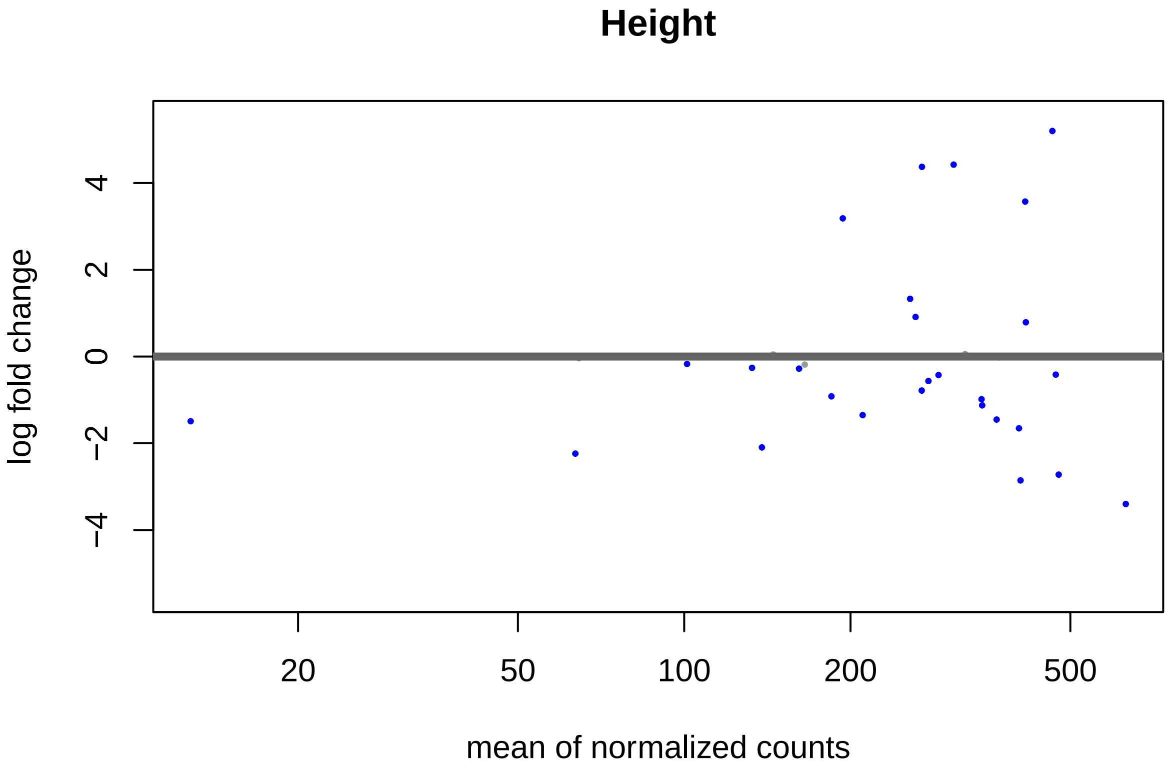


There are no significant clusters anymore.

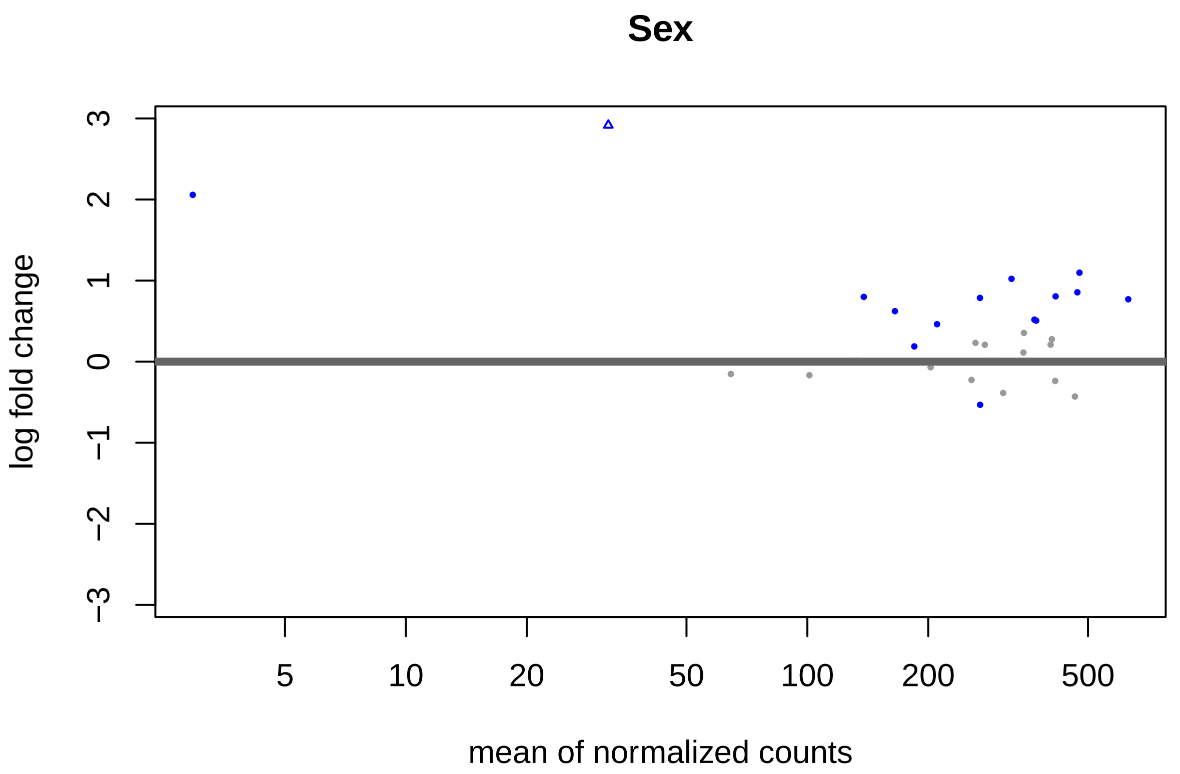
```r
#formul <- as.formula("~BMI")
resBMIcorr <- resBMI
plotMA(resBMIcorr, main = "BMI")
```



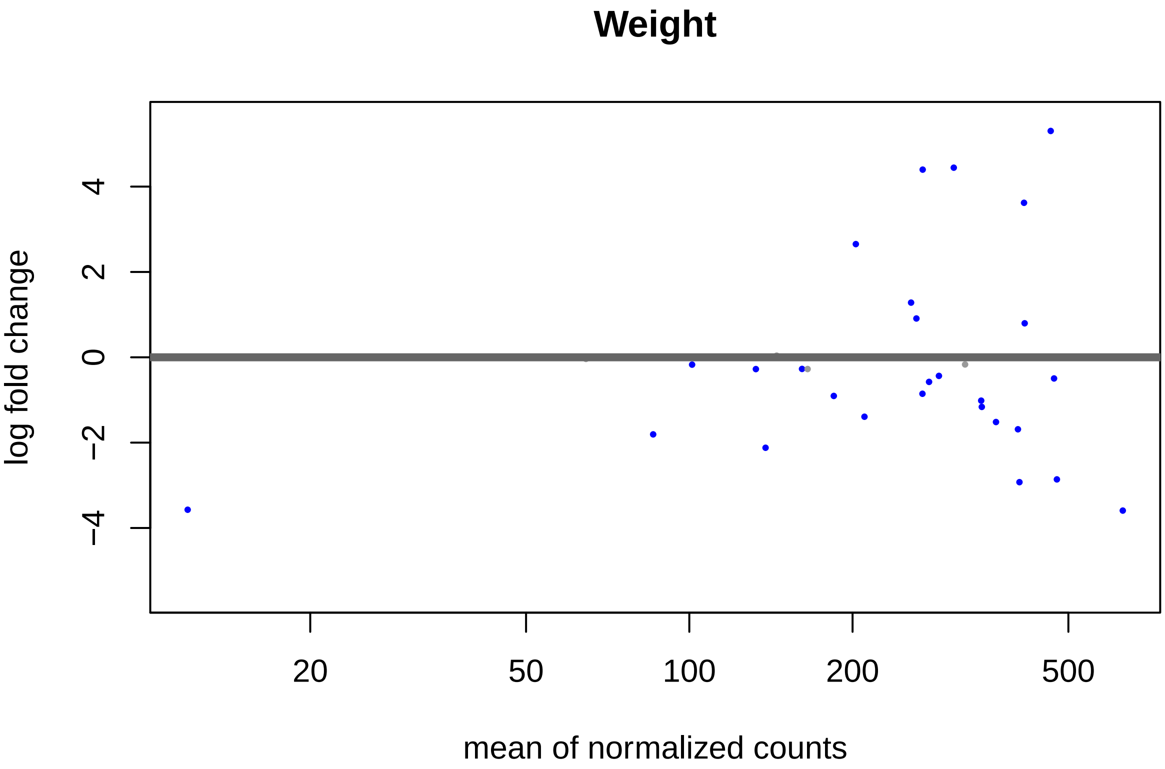
```
formul <- as.formula("~HGHT+COHORT")
dds <-  DESeqDataSetFromMatrix(countData = count, colData = clinic, design = formul)
dds_res <- DESeq(dds)
resHGHTcorr <- results(dds_res)
plotMA(resHGHTcorr, main = "Height")
```



```r
#formul <- as.formula("~SEX")
resSEXcorr <- resSEX
plotMA(resSEXcorr, main = "Sex")
```



```r
formul <- as.formula("~WGHT+COHORT")
dds <-  DESeqDataSetFromMatrix(countData = count, colData = clinic, design = formul)
dds_res <- DESeq(dds)
resWGHTcorr <- results(dds_res)
plotMA(resWGHTcorr, main = "Weight")
```



The MA plots haven't changed much except the one which relates the morphological cluster count to age. It seems that there is no correlation between the age and the morphological cluster count when we correct for technical variables.