

BB512/BB612 - Week II

```
suppressPackageStartupMessages(library(ggpubr))
suppressPackageStartupMessages(library(BioBase))
suppressPackageStartupMessages(library(preprocessCore))
```

Data

We'll explore the `bodymap` expression dataset.

```
con <- url("http://bowtie-bio.sourceforge.net/recount/ExpressionSets/bodymap_eset.RData")
load(file=con)
close(con)

bm <- bodymap.eset

pdata <- pData(bm) # phenotype data
edata <- exprs(bm) # expression data
fdata <- fData(bm) # features data
```

Phenotype data

We'll first explore the phenotype data:

```
head(pdata,3)

##           sample.id num.tech.reps tissue.type gender age      race
## ERS025098   ERS025098            2    adipose     F  73 caucasian
## ERS025092   ERS025092            2    adrenal     M  60 caucasian
## ERS025085   ERS025085            2      brain     F  77 caucasian

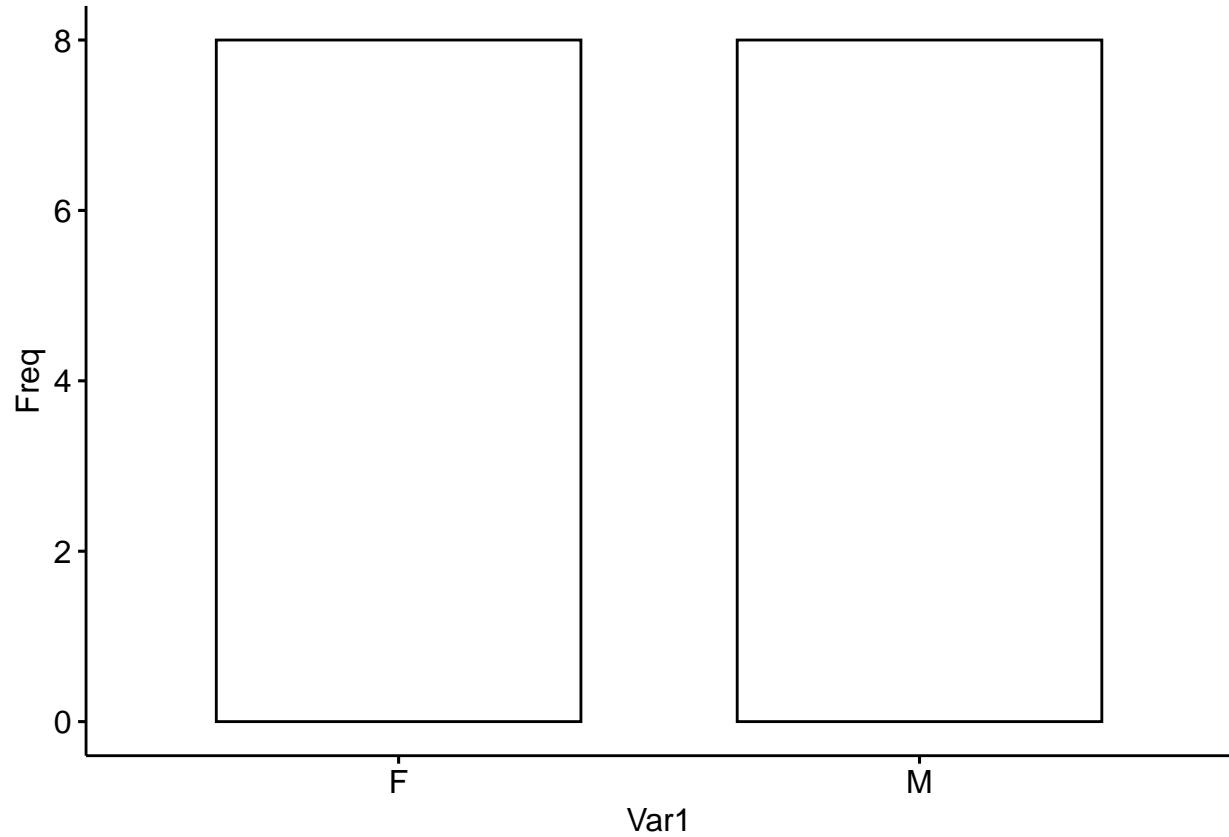
table(pdata$gender)

##
## F M
## 8 8

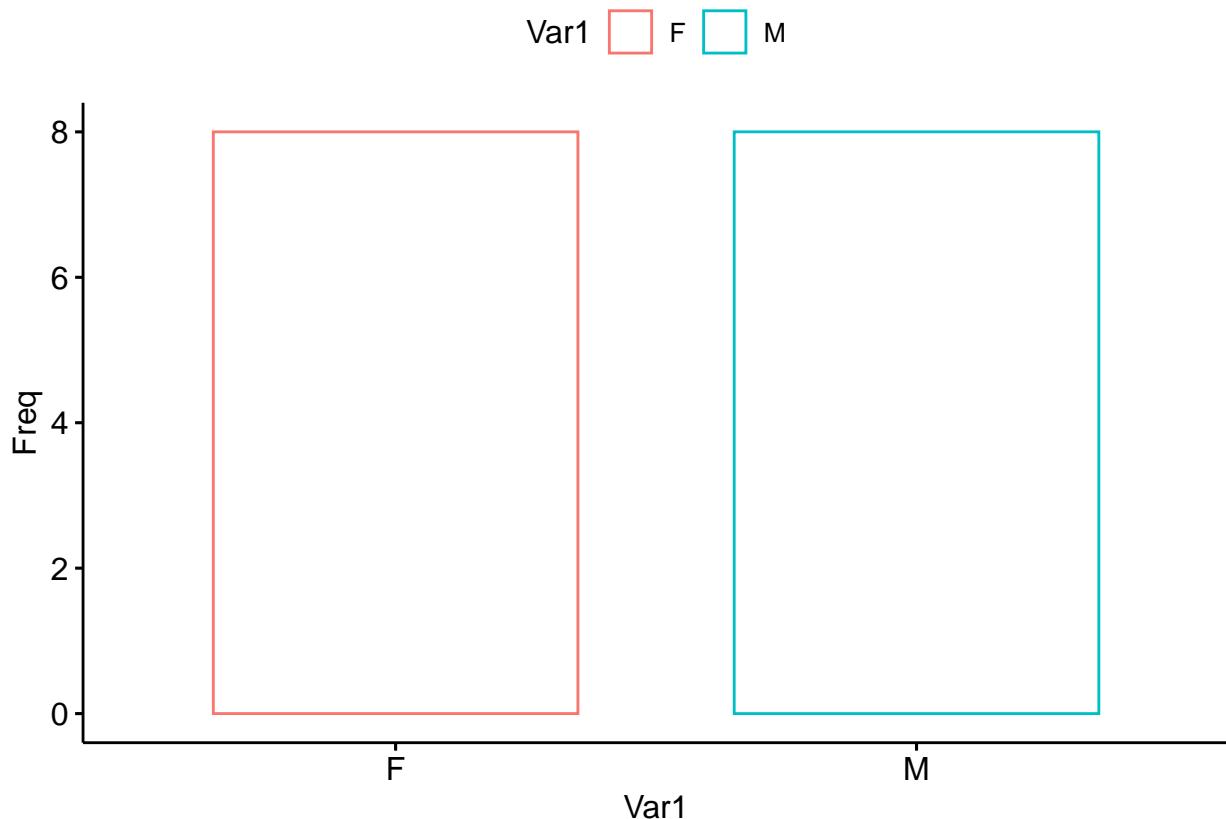
for_plot_df <- as.data.frame(table(pdata$gender))
head(for_plot_df)

##   Var1 Freq
## 1   F    8
## 2   M    8

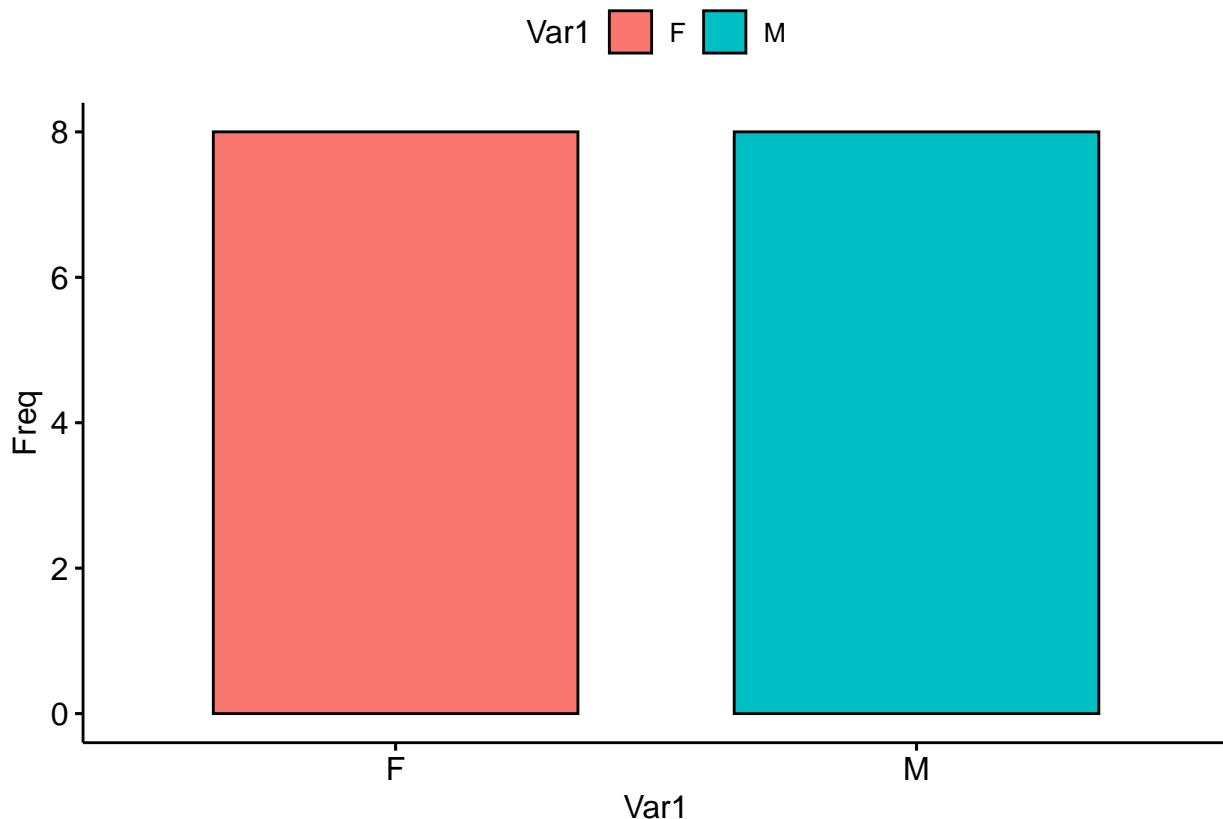
g <- ggbarplot(for_plot_df, x = "Var1", y = "Freq")
g
```



```
g2 <- ggbarplot(for_plot_df, x = "Var1", y = "Freq", color = "Var1")
g2
```



```
g3 <- ggbarplot(for_plot_df, x = "Var1", y = "Freq", fill = "Var1")  
g3
```



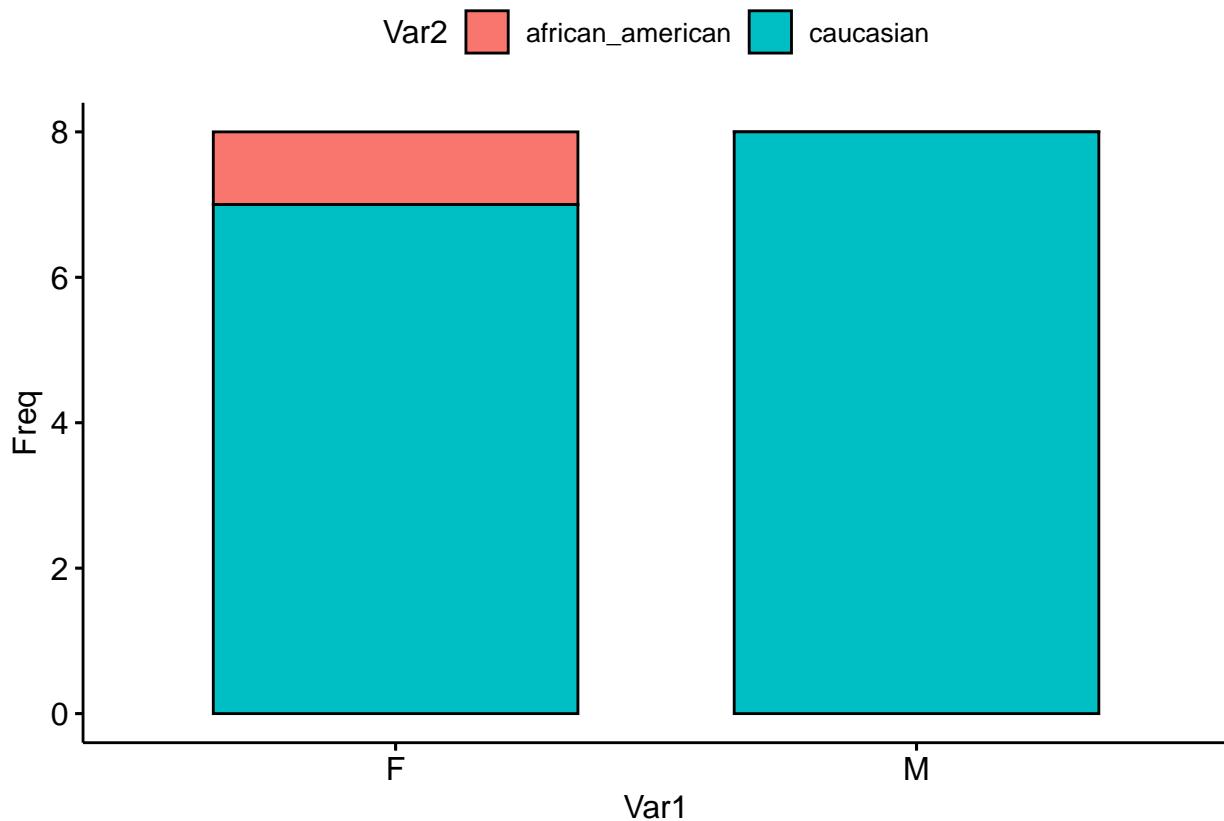
```
table(pdata$gender, pdata$race)

##
##      african_american caucasian
##      F                  1          7
##      M                  0          8

for_plot_df2 <- as.data.frame(table(pdata$gender, pdata$race))
head(for_plot_df2)

##   Var1           Var2 Freq
## 1   F  african_american    1
## 2   M  african_american    0
## 3   F      caucasian      7
## 4   M      caucasian      8

g4 <- ggbarplot(for_plot_df2, "Var1", "Freq", fill = "Var2")
g4
```

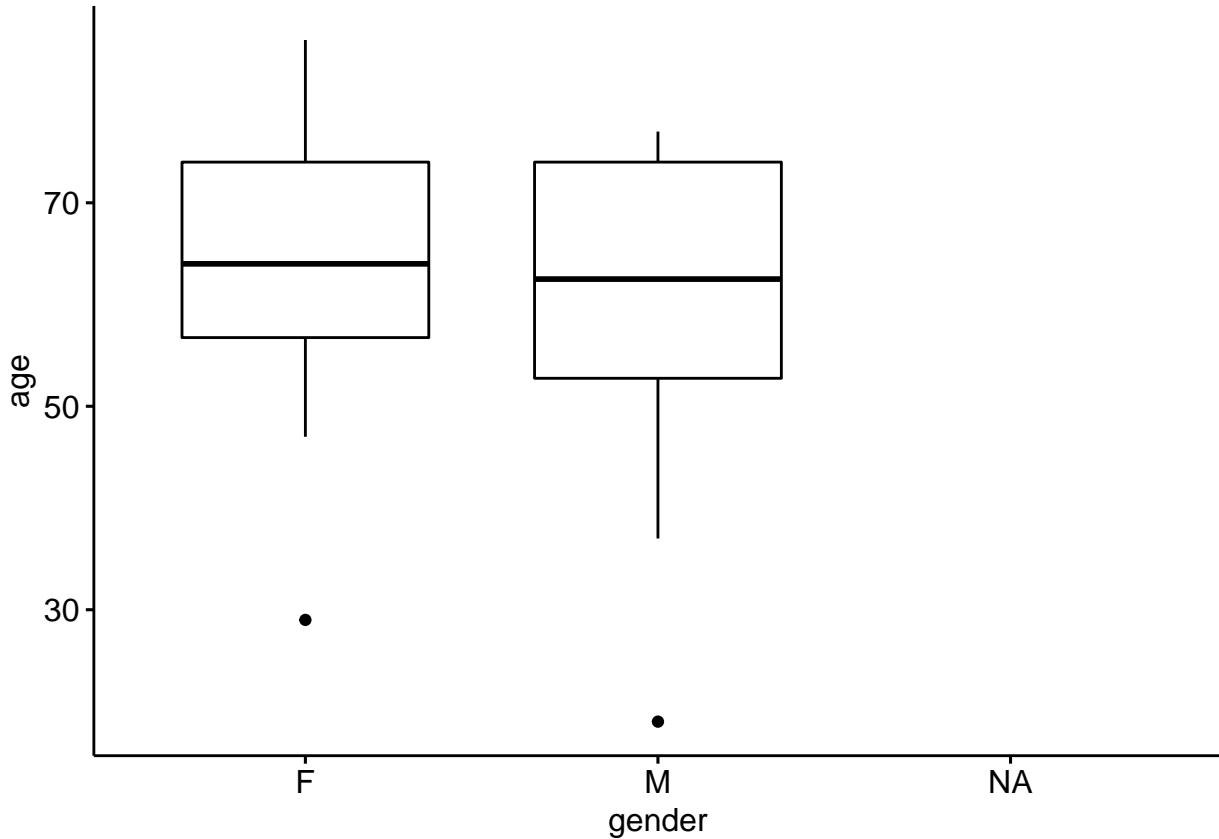


```
summary(pdata$age)
```

```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.    NA's
##     19.0    55.2   62.5    60.4    74.0    86.0       3
```

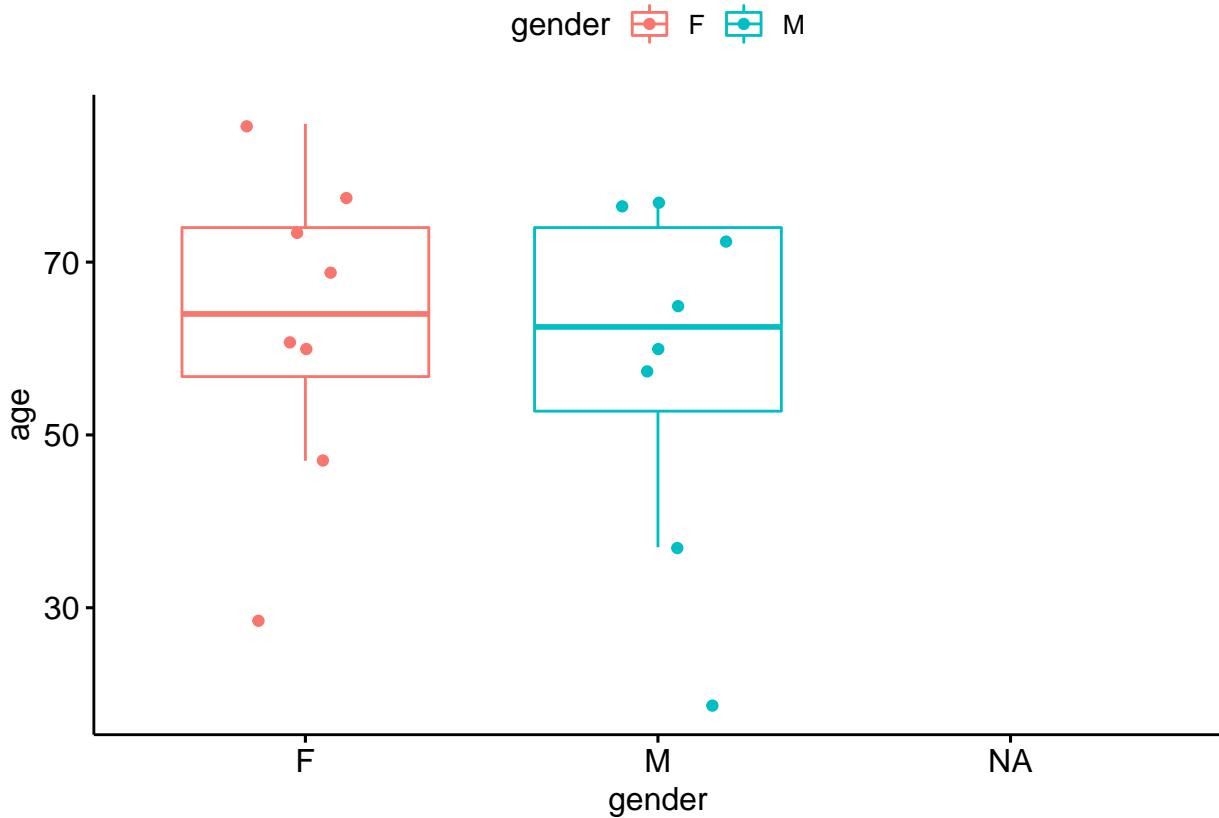
```
ggboxplot(pdata, "gender", "age")
```

```
## Warning: Removed 3 rows containing non-finite values (stat_boxplot).
```



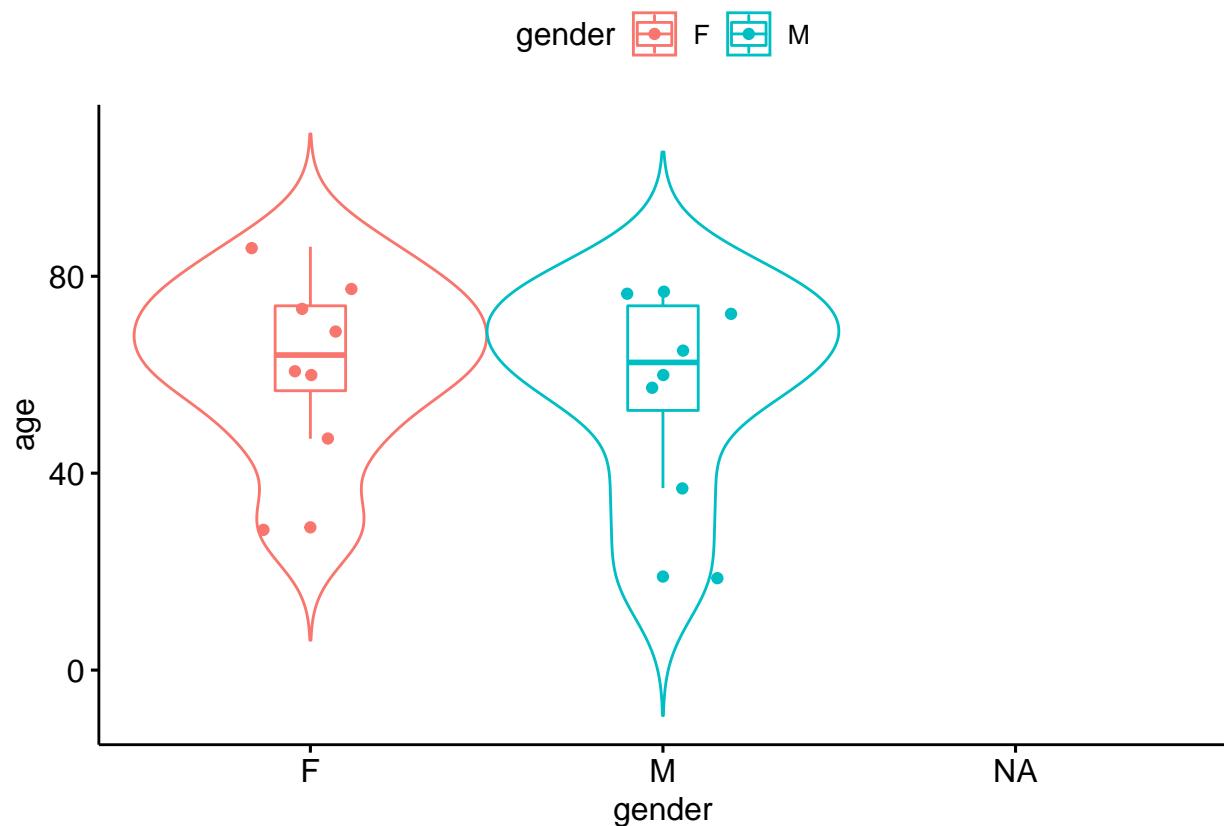
```
g5 <- ggboxplot(pdata, "gender", "age", color = "gender", add = "jitter")  
g5
```

```
## Warning: Removed 3 rows containing non-finite values (stat_boxplot).  
## Warning: Removed 3 rows containing missing values (geom_point).
```



```
g6 <- ggviolin(pdata, "gender", "age", color = "gender", add = c("boxplot", "jitter"))
g6
```

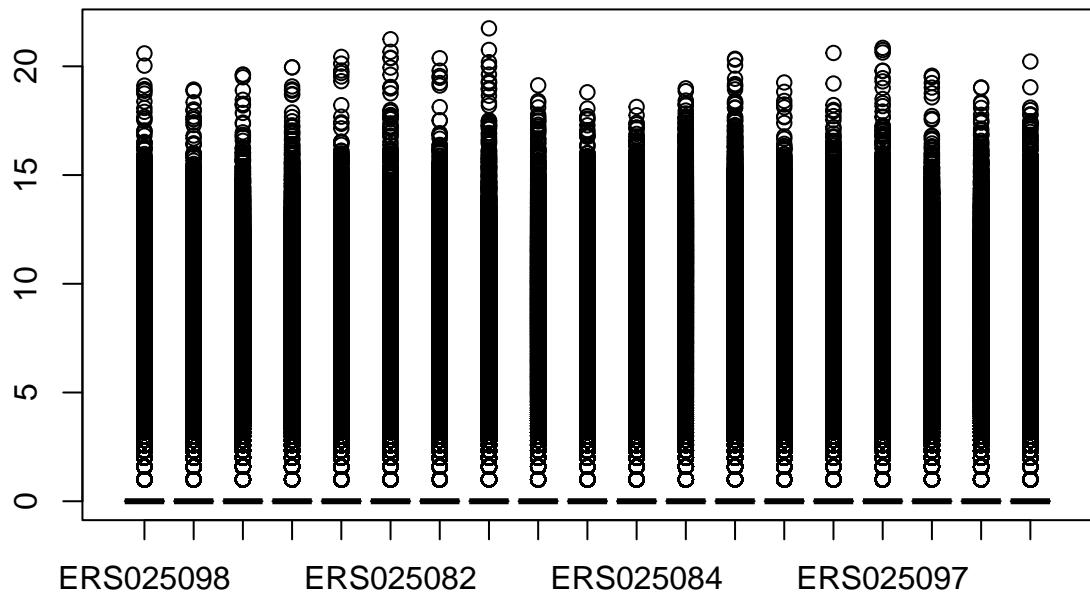
```
## Warning: Removed 3 rows containing non-finite values (stat_ydensity).
## Warning: Removed 3 rows containing non-finite values (stat_boxplot).
## Warning: Removed 3 rows containing missing values (geom_point).
```



Expression data

Look at overall distributions:

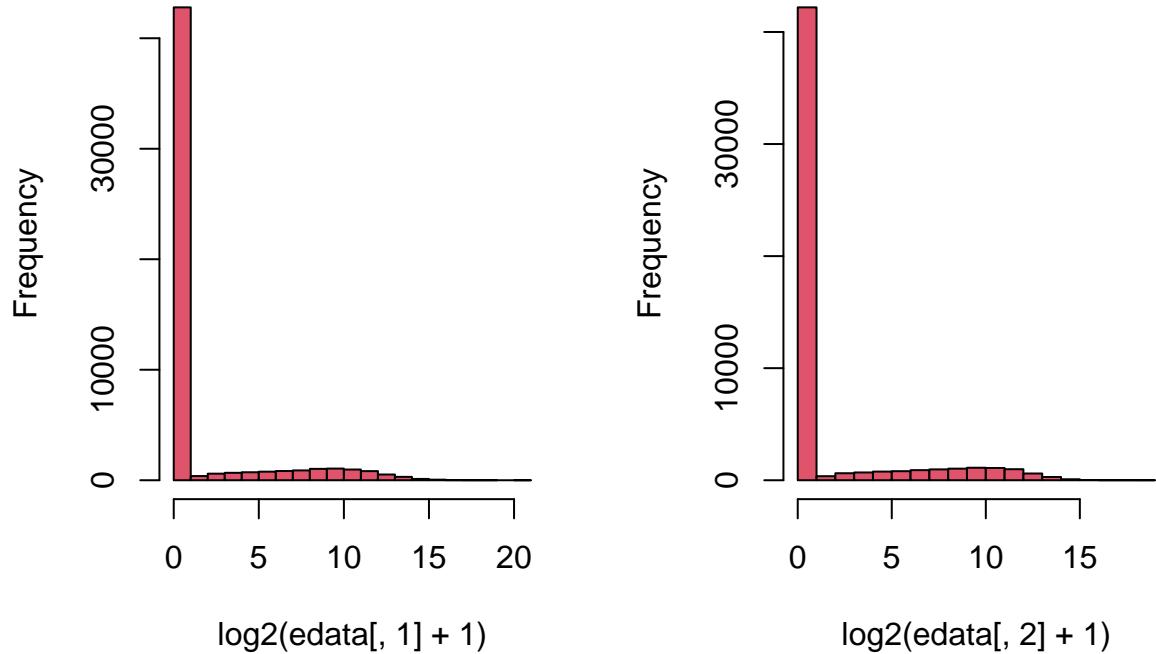
```
boxplot(log2(edata + 1))
```



alternatives:

```
# sample by sample histograms
par(mfrow = c(1, 2))
hist(log2(edata[, 1] + 1), col = 2)
hist(log2(edata[, 2] + 1), col = 2)
```

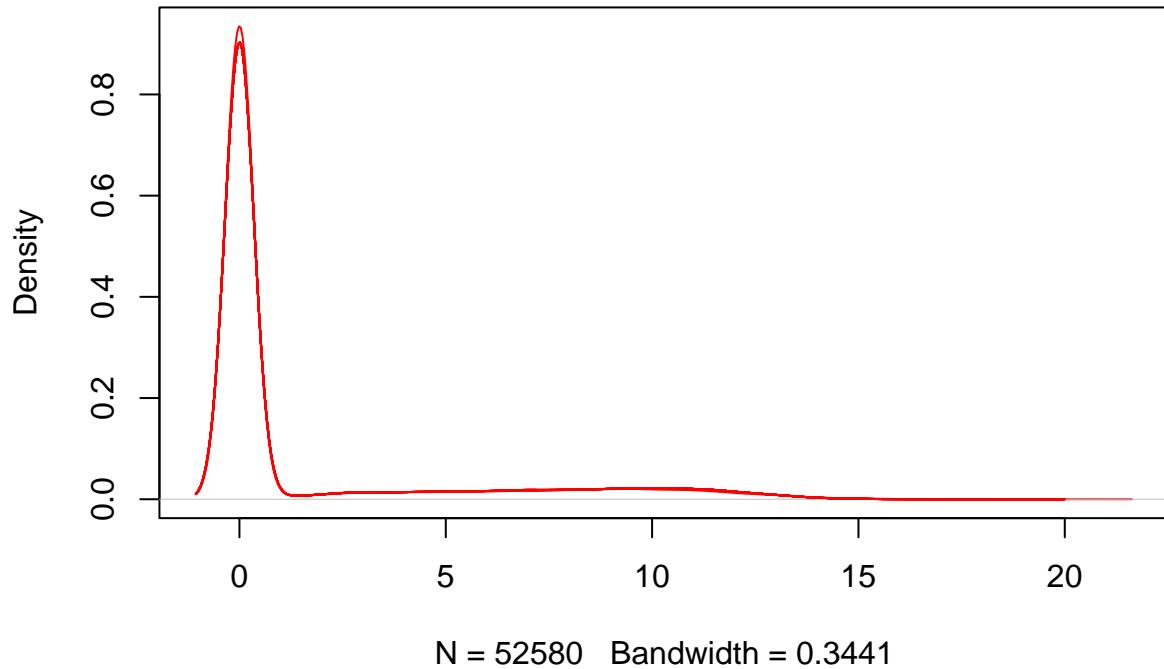
Histogram of $\log_2(\text{edata[, 1]} + 1)$ Histogram of $\log_2(\text{edata[, 2]} + 1)$



```
par(mfrow = c(1, 1))

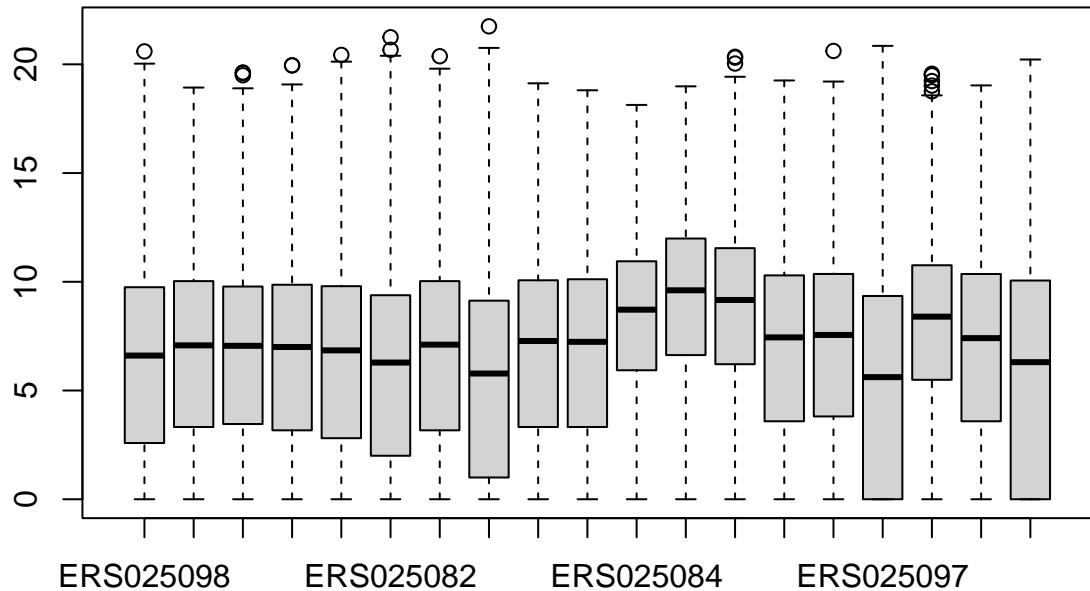
### or with density plots
plot(density(log2(edata[, 1] + 1)), col = "red")
for (i in 2:ncol(edata)) {
  lines(density(log2(edata[, 2] + 1)), col = "red")
}
```

```
density.default(x = log2(edata[, 1] + 1))
```



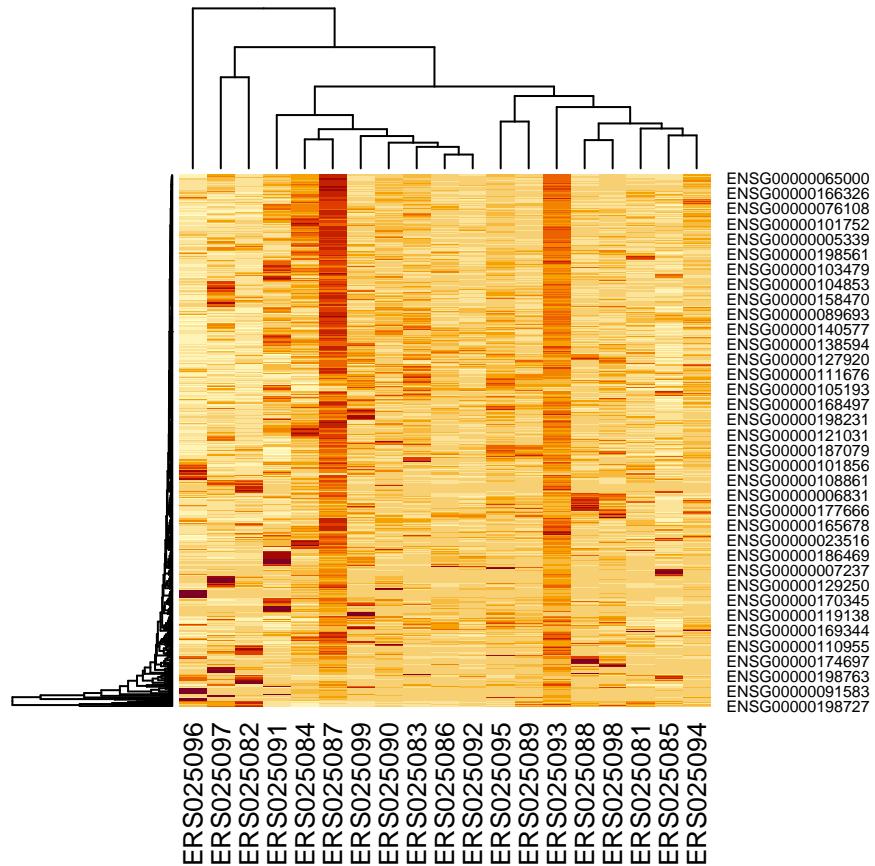
We can remove rows that are mostly zero and notice any differences in the distributions across samples.

```
filt_edata <- edata[rowMeans(edata) > 1, ]  
boxplot(log2(filt_edata + 1))
```



A common type of plot for genomics data is a heatmap. They are usually used for visualizing matrices. For example we can look at all genes with an average number of counts greater than 10000:

```
ematrix <- as.matrix(filt_edata)[rowMeans(filt_edata) > 10000, ]
heatmap(ematrix)
```



Quantile normalization:

```
lt_filt_edata <- log2(filt_edata + 1)
norm_edata <- normalize.quantiles(lt_filt_edata)
colnames(norm_edata) <- colnames(lt_filt_edata)
rownames(norm_edata) <- rownames(lt_filt_edata)
boxplot(norm_edata)
```

