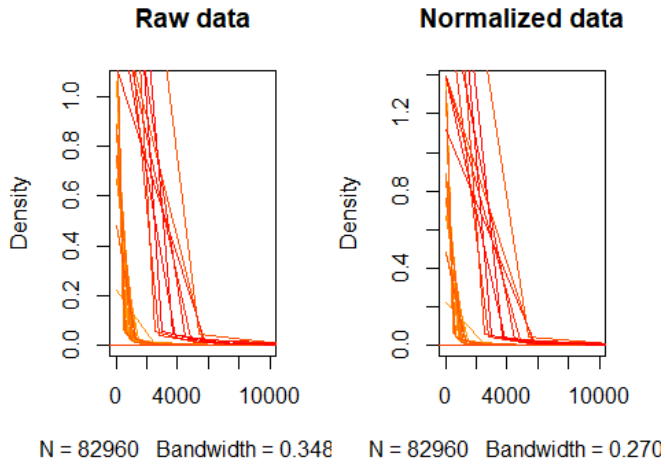
Re-analysis of fetal and adult brain raw RNA-seq data from the study "Developmental regulation of human cortex transcription and its clinical relevance at base resolution" (Jaffe et al, 2015 Jan PMID:25501035)

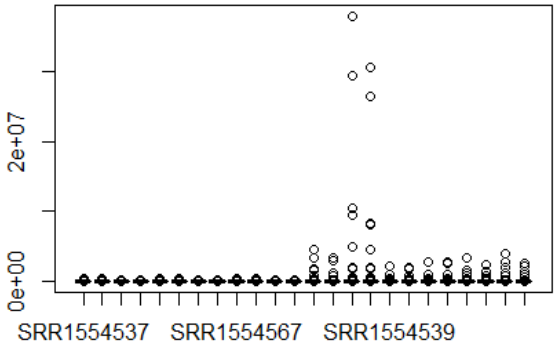
Task 5: Exploratory analysis

The first step was to normalize counts data. I used counts obtained with featureCounts on the 24 samples (12 fetal, 12 adult) and I performed normalisation using the DESeq2 package algorithm that, according to literature (e.g. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4484837/), seems much more reliable for RNA-seq data than the number of reads mapped to a feature per milion reads mapped.

```
library(DESeq2)
counts<-read.table("counts.txt",sep="\t",header=TRUE)
pdata<-read.table("PhenoData.txt",sep="\t",header=TRUE)
row.names(pdata)<-pdata$TecReplicate
de<-DESeqDataSetFromMatrix(counts,pdata,design=~1)
size<-estimateSizeFactors(de)
normalized<-counts(size,normalized=TRUE)
row.names(normalized)<-row.names(counts)
colnames(normalized)<-names(counts)
```

Here are shown the density plots for raw and normalized data. Since they're very similar, I infer that library size has not a big impact on these experimental data.





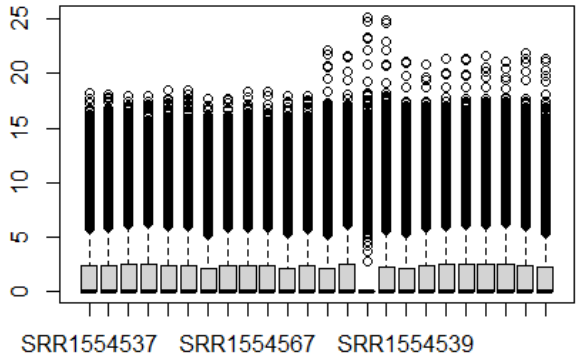
As shown by this boxplot and by the following representative summaries, all the samples are very left-skewed, with the majority of values equal to 0 and the median always lower than the mean.

```
summary(normalized[,1:3])
```

```
##    SRR1554537          SRR2071348         SRR1554538
##  Min.   :     0.00   Min.   :     0.0   Min.   :     0.00
##  1st Qu.:     0.00   1st Qu.:     0.0   1st Qu.:     0.00
##  Median :     0.00   Median :     0.0   Median :     0.00
##  Mean   :   182.74   Mean   :   183.7   Mean   :   175.52
##  3rd Qu.:     3.88   3rd Qu.:     4.1   3rd Qu.:     4.41
##  Max.   :299457.80
Max.   :285164.3   Max.   :255313.03
```

Applying a log2 transformation to the data (with a previously added pseudocount of 1) somewhat improves the distribution (see boxplot), but all the 0 values are preserved:

```
normlog<-log2(normalized+1)
```




```
summary(normlog[,1:3])
```

```
##    SRR1554537       SRR2071348
SRR1554538
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000
##  Median : 0.000   Median : 0.000   Median : 0.000
##  Mean   : 1.635   Mean   : 1.666   Mean   : 1.685
##  3rd Qu.: 2.286   3rd Qu.: 2.350   3rd Qu.: 2.436
##  Max.   :18.192   Max.   :18.121   Max.   :17.962
```

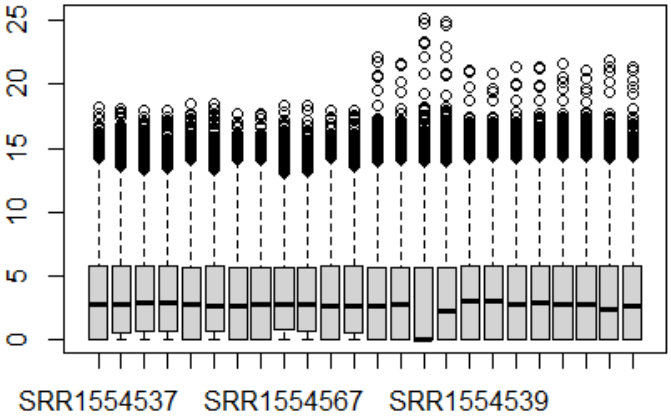
Subsequent filtering for features with mean >0 doesn't trigger any loss of information (as it only excludes genes with counts in every sample equal to 0), reduces by more than 50% the number of features and considerably improves the distribution, as shown in the following plots. So, this will be the tranformation used in further analysis.

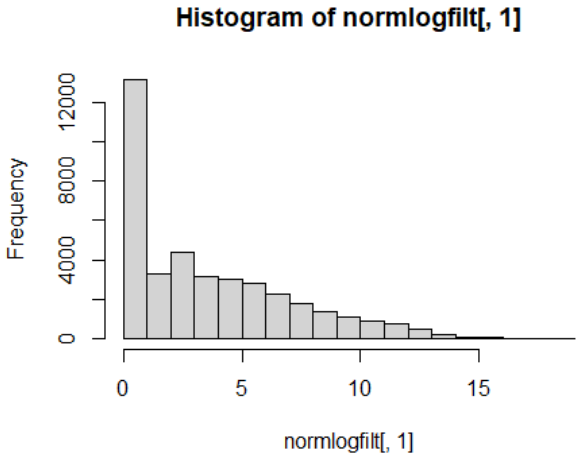
```
normlogfilt<-normlog[rowMeans(normlog)>0,]
dim(normalized)
```

```
## [1] 82960    24
```

```
dim(normlogfilt)
```

```
## [1] 38756    24
```



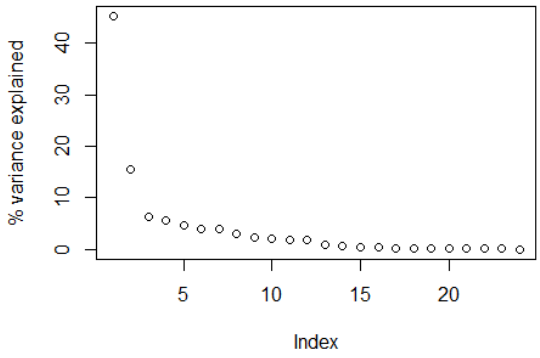



```
summary(normlogfilt[,1:3])
```

```
##    SRR1554537       SRR2071348        SRR1554538
##  Min.   : 0.000   Min.   : 0.0000   Min.   : 0.0000
##  1st Qu.: 0.000   1st Qu.: 0.4568   1st Qu.: 0.7052
##  Median : 2.685   Median : 2.7197   Median : 2.8687
##  Mean   : 3.499   Mean   : 3.5656   Mean   : 3.6078
##  3rd Qu.: 5.706   3rd Qu.: 5.7116   3rd Qu.: 5.7537
##  Max.   :18.192   Max.   :18.1214   Max.   :17.9619
```

Then, I calculated the singular values on the centered data and plotted the % of variance explained by each of the 24 principal components. First PC accounts for 45% of variance, second PC for 15%.

```
centered<-normlogfilt-rowMeans(normlogfilt)
svd<-svd(centered)
plot(svd$d^2/sum(svd$d^2)*100,ylab="% variance
explained")
```



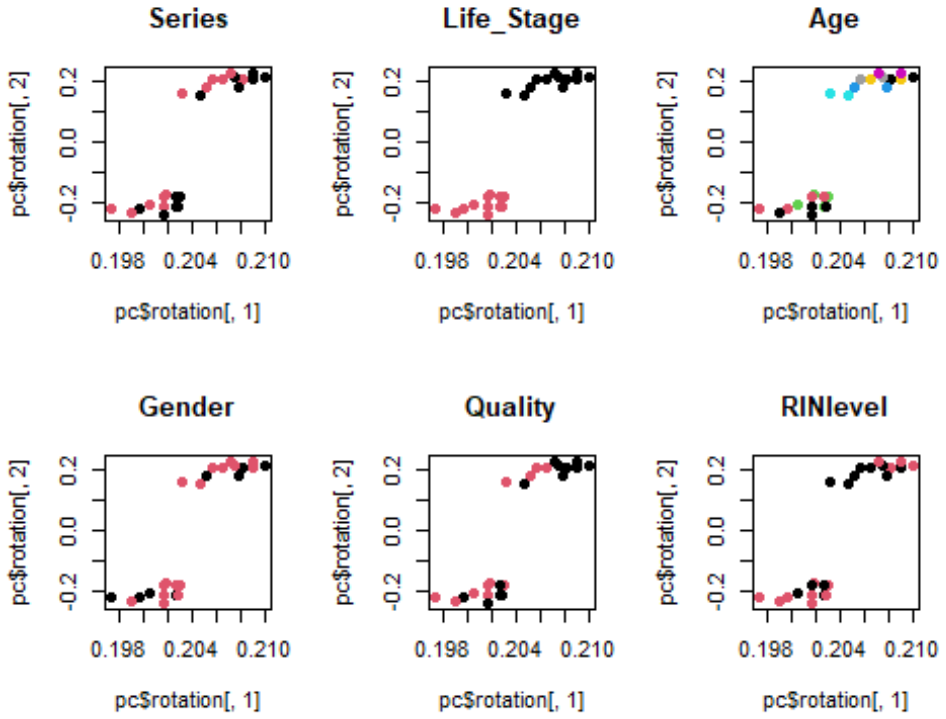
I eventually performed PCA and looked if any of the phenotype variables tend to cluster with respect to the first and second PC. I considered Series (reads SRR155… or SRR207…), Lifestage (Fetal or Adult), Age, Gender, Quality (Good or Poor according the my quality control), and RINlevel ("low" if less than RIN median, "high" if greater).

Altough Age plot is not very informative, I kept it because it shows that the 2 replicates of each sample tend to cluster together.

Different colours show different values of the phenotype.

```
pc<-prcomp(normlogfilt)
```

– *Elena Civati*



I've also tabled the correlation coefficients between each variable and the first and second PC:

```
##                First PC    Second PC
## Series      -0.27698452  0.005765973
## Life_Stage -0.84649948 -0.994170395
## Age          0.83930284  0.887029809
## Gender       0.03609562  0.020764261
## Quality     -0.48085503 -0.248825031
## RINlevel    -0.18605157 -0.316103898
```

As hoped, the strongest (and more evident from the plot) correlation is between PCs and Life Stage. Possible confounders are Quality and RINlevel, while I chose not to consider Age and Series as confounders, despite their high correlation with PCs, because each of them is already strongly correlated (statistically and biologically) with some of the other variables.