Workshop – Introduction to R

Exercise

The data for this exercise is gene expression data from rat livers, downloaded from GEO, GSE32156.

For simplicity, I selected the genes with the highest standard deviation, and renamed them from “A” to “W”.
The data is available to you in the exercise folder, as “Expression data selected.csv”.

# Handling data in R

1. Import the data from the csv file to a data frame (type the full path, or change the working directory first to where the file is located).
2. Check the data type in each column. Are there factor columns where there should be character columns, or vice versa? Change the column types as needed.
3. Calculate the average (mean) expression for gene A.
4. Calculate the average (mean) expression for gene A, but do this separately for each age group.
5. Create a new data frame that contains only females who received the “high” pbde
treatment.

# Basic plots

For all following questions use the original data frame, not the subset.

1. Create a scatterplot of gene A vs. Gene B expression.
2. Color the plot markers by the different factors (one factor each time).
 Can you find which factor differentiates the two groups in the scatterplot?
3. The dataset has 23 genes. Create histograms for all of them using a loop.
 a. Split the plotting area using par(mfrow = …) into 4 rows and 6 columns.
 b. Write a loop that goes over all the numeric columns and creates a histogram
 from each one. Present the gene name (as the graph main title) over each
 histogram.
4. Create a scatterplot matrix to examine the correlation between five genes of your
choice using the “pairs” function.
5. Create a boxplot examining the effect of all three factors on gene B.
What groups are missing? Would you have planned this experiment differently?

# ggplot

1. Create a boxplot with data points over it (using geom\_boxplot and geom\_point), comparing the expression levels of gene E between the pbde treatments. Color the points by age.