

Exercises

Exercise 1 - Introduction

We have an Excel file with the expression of 1000 genes in 18 patients with NASH (a liver disease) and we want to plot the distribution of gene expression per patients, by using BoxPlotR.

Since BoxPlotR does not support columns with text, we must remove the column that contains the names of the genes. To do this, open "GSE48452_msel.xlsx" with a spreadsheet (LibreOffice Calc, Excel, etc), remove the column and save the result as a comma-separated file with the name "GSE48452_msel_noname.csv"

Exercise 2 - BoxPlotR

We want to see the distribution of the expression of the genes of the patients with NASH of Exercise 1 in order to check that there are not any problem with the samples. For that, we must upload the file "GSE48452_msel_noname.csv", which we have previously created (there is also a copy of it in the "results" folder), to BoxPlotR.

1. The names of the samples are very large and overlap each other, to see them correctly mark the "Rotate sample names" option under "Modify labels and title".
2. Change the color of the boxes. To do this you can write the name of the colors (or color) in the box "Color (s)" or look for the HEX code of the color in this web: <http://colorbrewer2.org/>
3. Go to "Figure legend template" and observe what it says.

Questions:

1. Adding the data points you get more information or "dirty" the image? Why? ("Bee swarm" representation can take a while to be shown)
2. Remove data points. All samples should have a value to all (1000) genes? Set a restriction for the plot to show only the samples with 1000 genes.
3. Try to upload "GSE48452_msel.csv". What happen when plot the data? Which message it is shown?

Exercise 3 - Heatmapper

We have performed a RNA-seq experiment with the aim to detect differences in the expression of the genes of two groups of mice: wild type (WT) and treated with the hormone T3 (WT_T3).

Once the data were normalized, we performed a differential expression analysis and we saved these normalized expression data of the genes significantly differentially expressed in the file "rna_tmmfdr_DE.txt".

1. Open the file with a spreadsheet or notepad and note its content. The column "NAME" contains the name of the genes significantly differentially expressed and is necessary for Heatmapper to put the name of each of the genes. The other columns pertain to the samples analyzed and the rows. So each column is a sample and each row is a gene.
2. Upload the file to Heatmapper and observe the result. The samples are in the same order that in the file. Apply clustering to columns (and show the dendrogram) to see better the differences of the expression.
3. Reduce the number of shades. What happens if you use a very low number of shades (3 or 4)?
4. Restore the number of shades to 50 and change the color scheme for another one that you like. (You can use <http://colorbrewer2.org> (for diverging data) to choose a good combination of colors.)
5. Add a title and a name to the axes and download the figure in PNG format.

Questions

1. By default, the expression is normalized by rows (genes). What happens when you remove this normalization? Is there a gene much more expressed than the rest? Could you find out what expression value WT2 has in the sample?

Extra exercises

Exercise 4 - BoxPlotR

The file "distributions.txt" is a semicolon-separated file containing 3 groups (A, B and C) of random numbers following different distributions (bimodal, uniform and normal).

1. Visualize the data using a boxplot. Could you find out which distribution corresponds to each sample?
2. Add data points to the image and try it again. You can use default, bee swarm and jittered options. Is it easier now to find the distributions?
3. Now use violin plot and bean plot to see the distribution of the data.

Questions:

1. What do you think is the best representation to visualize this set of data?
2. Which distribution corresponds to each sample?

Exercise 5 - BoxPlotR

In this exercise we are going to see how outliers can affect mean. The file "mean.csv" contains two series of eight values.

1. Upload the file to BoxPlotR and see the values (only one is different in each serie).
2. Visualize the data as a boxplot. Log scale could be a good option.
3. What happens with the median? (Exact values are displayed at "Box plot statistics" table)
4. Now display sample means and see the result. What is the effect of the outlier on the mean?

Exercise 6 - BoxPlotR

Upload again the file "GSE48452_msel_noname.csv" and see the distribution using violin plot. It gives you more information than the boxplot?

The file "GSE48452_mall_noname.csv" contains all the samples (67) of the experiment GSE48452 (both NASH and no NASH). Upload this file and compare and see what happens when you have many samples.

Change the plot width to 1000, the axis font size to 6 and compare box plot against violin plot. Which one seems more useful? Why?

Exercise 7 - Heatmapper

After a gut metagenomics analysis of obese and lean people we have a matrix containing the abundance of different bacterial genera of each sample (patient) (gut_abundance_abr.txt).

As the program does not allow long names, bacterial genera are abbreviated according to the following table:

Synt.	<i>Syntrophococcus</i>
Tera.	<i>Terasakiella</i>
Rumi.	<i>Ruminococcus</i>
Mari.	<i>Marinilabilia</i>
Coll.	<i>Collinsella</i>
Bact.	<i>Bacteroides</i>
Palu.	<i>Paludibacter</i>
Bran.	<i>Brantella</i>
Desu.	<i>Desulfovibrio</i>

- Apply cluster by samples (columns).
- Remove any scale type. Which is the most abundant genus? In which group of patients?
- In the case of non-scaled data. Are the results better visualized with sequential color palettes or divergent color palettes? (<http://colorbrewer2.org>)