Statistics for RNA-Seq

Computer Lab with R

Introduction. This computer lab makes you familiar with elementary statistics of RNA-Seq. The basic tool is the R system, so it is assumed that you are familiar with it. Almost any version will work, but it makes sense to use a recent version. This lab was tested with R version 2.14. No additional packages (libraries) are needed.

Software and data files. The files to be used in the lab will be put in a place on the network. There will also be a USB memory stick that can be passed along. There is one file containing step-by-step instructions for you to give to R. Make a copy of the files, so that you can change instructions freely, without losing the original ones. Also copy the data files to your working directory.

How to work. If you are lazy, you copy individual instructions or blocks of instructions to R and see what happens. However, in the initial phase it is advisable to type the instructions yourself. It is better way to learn and to remember. Start R and change to directory in which you stored the files.

Comments. The instruction file contains many comments to document what is going on. A large number of comments contain section numbers, like S1 and S9. They refer to the explanations that follow below.

1. Read in the data in the file 'Yeast_counts.Rdata', copy the data frame to a matrix, and print the names of the variables.
2. Make a scatterplot of the second column against the first column.
3. Improve the scatterplot with a title and label the axes. Also choose another and smaller symbol for the dots.
4. Use variables to indicate the columns. That makes it easier to try different combinations. Compare several columns to each other.
5. Change to logarithms of the counts (plus one, to avoid problems with zeros).
6. Do the same for square roots.
7. Make a histogram on the linear scale.
8. Do the same for the logarithms.
9. Improve the histogram, using more (an more narrow) bins and labels. You can play with other numbers in 'breaks = 50' to see the effect.
10. To make checking the Poisson assumption easy, a special function is defined. Copy it to R. It is instructive to follow the steps in the function and compare it to the computations in lecture

slides. You will see that the function ‘apply’ is used several times. Look in the R help to understand it better. You might also want to look at the help for ‘outer’.

11. Once the function definition has been sent to R, we can use it.
12. How many zero counts are in the data? You will get four fractions, one for each column of the data.
13. The same for the number of counts above 10, and above 100.
14. Read in the Marioni data from the text file that accompanies the paper, and perform the same analysis as above. You can type in all of the instructions completely, or copy them and make the changes that are needed. You don’t have to define the function ‘check_poisson’ again.
15. Similar to 14, but for the Montgomery data. Unfortunately there is a small error in the file name (it is ‘Mongomery.Rdata’), but the contents are OK.

Once you understand this analysis, you can apply it to your own data.