Babraham Bioinformatics

Exercises:
Advanced R

Version 3.0

Licence

Record Keeping

For all of these exercises create a blank script called `methylation_expression.r` and add the critical parts of the analysis to this script to keep a record of what you did in this course. You will need this script for the last part of the course.

Exercise 1: Filtering and Deduplicating

- In the Advanced R Data folder you should have 2 tab delimited text files
  - `expression.txt` contains RPKM expression values
  - `methylation.txt` contains %methylation for gene bodies and promoters
- Load all of these datasets into separate data frames using `read.delim`
- In the methylation dataset remove any rows where the promoter methylation level is -1. These are regions which didn't contain enough data to allow for proper measurement. How many rows are left? To remove rows, simply select the rows you want to keep and then overwrite the original data frame.
- The methylation data is already deduplicated, but you should deduplicate the expression data (you can leave one instance of any repeated names).
- Using `grep` or `grepl` find out how many probes in the expression data set come from olfactory receptors (they contain "Olfr" in their names).

Exercise 2: Extending, Intersecting and Merging

- In the expression data change the chromosome names from being 1,2,3 etc to chr1,chr2,chr3. To do this you will need to use `paste` to add chr with an empty string as the separator. You can assign the modified names over the existing Chromosome column to replace the existing names.
- Use the `%in%` operator to get a list of gene names which are in the methylation but not the expression data set.
- Merge the expression and methylation data together. You want to merge based on the Probe column. As this column is already named the same between the two data frames you can just use `merge` with no additional options to merge the datasets together into a new combined data frame.

Exercise 3: Looping

- Use an `apply` statement to calculate the `range` of values in the expression and methylation columns in your merged dataset (probably columns 6,7 and 8).
- Use `tapply` to calculate the `mean` gene body methylation for genes on each chromosome. [Optional] Plot these results as a barplot.

Exercise 4: Functions

- There is no direct function in R to calculate the standard error of the mean (SEM). This value is simply the standard deviation divided by the square root of the number of observations. Write a function called `sem` to perform this calculation.
- Add error checking to your function so that it will return an `NA` value if passed in a vector which either contains an `NA` value.
- Add an extra optional parameter called "`absent`" to your function which specifies what value should be returned if an SEM calculation can't be performed. Have this default to NA, but test that it works with other values.

- Rerun the `tapply` calculation from the end of exercise 2 to calculate the SEM of the mean gene body methylation levels.

Exercise 5: Packages

- Locate the package you'd use for constructing violin plots (like an improved version of a boxplot). Install and load this package. Read the documentation to find out how it works.
- Construct violin plots for the Expression and Methylation datasets. Put these into a 1x3 combined plot area. Add a title to each plot.
- [optional] See if you could perform the operation above by using `sapply` through the column names of your merged data frame.

Exercise 6: Knitr

- Turn the script you've been developing during these exercises into a `.Rmd` file.
- Use knitr create an explanatory report from it which completely documents the analysis you did.