UM PhD vacancy | UNIQUE project

Bioinformatics test

In this test, we will be working with DNA methylation data obtained via Reduced Representation Bisulfite Sequencing (RRBS) from semen samples of males undergoing endurance training.

Please follow the instructions and provide your answers to the questions below. Indications are included if you are using R, but feel free to use any other programming language to complete the test as you see fit.

Please share your report and all scripts employed in the completion of this test.

Go to Gene Expression Omnibus (GEO) and download the data with GEO ID: GSE109474 (GSE109474_methReads.csv.gz, GSE109474_predictedMeth.csv.gz and GSE109474_totalReads.csv.gz). These three files contain either the total number of reads, reads corresponding to the methylated epiallele or the predicted methylation ratio.

Q1) What is the experimental design of the study (number of samples per condition, replicates, controls)? What is the difference between RRBS and WGBS? Which restriction enzymes did they use in this assay and what are their target sequences?

Q2) Why does the *predictedMeth* file have a size of 146.0 Mb, while the *methReads* and *totalReads* files have a size of only 18.5 and 20.1 Mb?

First, explore the *methReads* and *totalReads* .csv files. When using R, you can import the *methReads* and *totalReads* .csv files (*clue: it will be much faster if you use the function fread from the "data.table" R-package*). The data contains 4 columns storing characters: "seqnames", "start", "end" and "cluster.id". These will be a nuisance for any downstream analysis. Create a identifier vector combining chromosome start and end position with the following format: "chr_i:start-end". Then, remove the 4 character columns from the imported matrices, convert to numeric matrix data type and use the created ID vector as row.names. Perform this processing step for both *methReads* and *totalReads.* Finally, compute DNA methylation ratio as in:

$$\beta = \frac{\#\ Reads_{Methylated}}{\#\ Reads_{total}}$$

You may round the matrix to 3 decimals to reduce storage constraints.

Q3) Count the number of missing values (NAs) in the obtained methylation ratio matrix. How were the NAs originated in the methylation ratio? Clue: count the number of NAs in *methReads* and *totalReads*.

Q4) Visualize the methylation density distribution for each sample in a single plot. You may employ the convenient densityPlot function from the minfi R-package. Attach the resulting figure and comment about it.

Q5) Perform a multi-dimensional scaling (MDS) plot on the top 1000 most variable positions. You may employ the mdsPlot function from the minfi R-package. Employ the rainbow palette to colour each individual differently, use the filled circle point shape and display the legend in three columns. Describe briefly the basis of an MDS plot, attach and interpret the obtained visualization.

Extra points: label each axis with the percentage of variance explained.

Q6) Extract all positions included in the region chr12:739000-740500, that corresponds to a strong methylation quantitative trait locus (meQTL). How many positions are included in this region? Visualize the data in these positions via heatmap. Use the heatmap.2 function from the R-package gplots. Use ColSideColors argument to add a colour guide in which each individual has a different colour. Describe and interpret the output.

Now, import *predictedMeth* .csv file in R (*again with the "data.table" R-package*) and perform the same processing as for *methReads* and *totalReads* .csv files (e.g. remove character columns and add as row.names and identifier to each position as in "chr_i:start-end". You may round the matrix to 3 decimals to reduce storage resources.

Q7) Extract the methylation values for the row corresponding to "chr6:209819-209819" from "*predictedMeth*.csv". Plot this row vector against the methylation ratio we manually computed in the prior section for this same row. Does anything strike from the visualization? Explain why this is so. How does the densityPlot look for "*predictedMeth*.csv"? Use again the function densityPlot from the minfi R-package. Comment on the output plot.