Final exam for the course: Introduction to R for Bioinformatics – BC203

General instructions

The final exam consists in analyzing one transcriptomics dataset – either microarray or RNA-seq data – producing all the results indicated below.

The final report must contain both (a) the results of the analysis, (b) the code used for producing these results and (c) a discussion for each step of the analysis. The report can be either (i) a single Rmarkdown file, containing all code, results and discussions, or (ii) an R script containing the code along with a PDF document containing discussions and results.

The analysis for each dataset comprises the following steps:

1. Downloading both expression values and phenotype data from the appropriate repository
2. Check the quality of the data and eventually exclude samples whose quality is not adequate. Use one or more of:
    2.1. Boxplot / violin plots
    2.2. Correlation analysis
    2.3. Principal Component Analysis
3. Identify the expression quantities associated to one or more phenotype, using the appropriate covariate when needed (see instructions for each dataset)
    3.1. Present the results with both tables and the appropriate graphs (at least volcano plot and heatmaps)
4. Perform a functional analysis for each contrast (see instructions for each dataset)

Selected datasets

GDS3929

This dataset is assigned to the group Maria Vasilarou, Maria Tsochatzidou, Nefeli Paschou

For this dataset it is necessary to find the differentially expressed genes across the following contrasts. Find also the number of common and contrast-specific genes.

1. Non-smoker vs. smoker in Neonatal cord blood
2. Non-smoker vs. smoker in Maternal peripheral blood
3. Non-smoker vs. smoker in term placenta

For each contrast perform a KEGG GSEA. Finally, produces three scatter plots, one scatter plot for each pair of GSEA analyses. Each scatterplot should have on the x-axis the -log10(adj. pvalue) for one GSEA analysis and on the y-axis the -log10(adj. pvalue) for the other GSEA analysis.

GDS4206

This dataset is assigned to the group Haris Zafiropoulos, Paschalis Natsidis, Cristina Chatzipantsiou, Kleio Verrou

For this dataset, find the probesets differentially expressed between No-relapse vs. Early relapse subjects.

Then, use both (a) the appropriate bioconductor annotation package and (b) ensemble-biomart for finding the genes (Entrez id) corresponding to each probeset.

For both (a) and (b) quantify the number of genes for which there is at least one probeset differentially expressed. Quantify both the number of common and contrast-specific genes.

Finally, perform a KEGG GSEA separately for both (a) and (b). Quantify the intersections among the significant pathways.

SRP018008

This dataset is assigned to the group Antonis Kioukis, Vaggelhs Theodorakis, Akis Linardos.

For this dataset do as follow:

1. Find the genes differentially expressed between Cancer and Normal subjects
2. Find the GO categories enriched for the differentially expressed genes (restrict the analysis to GO level 4)

Repeat step 1 and 2 for 100 times, each time resampling with replacement and with equal probability as many subjects as contained in the original dataset. For each GO level 4 category count how many times it is found significant in the 100 repetitions. Use a bar plot for visualizing the result.

SRP029262

This dataset is assigned to the group Manouela Kapsetaki, Chrysovalantou Kalaitzidou, Dimitris Kyriakis, Zacharias Papadovasilakis.

For this dataset consider only subject for which the HbA1c is reported. For those subjects:

1. Find the genes associated to HbA1c using age, BMI and gender as covariates
2. Perform three KEGG GSEA analyses, one for deregulated pathways, one for upregulated pathways and one for one for downregulated pathways

Is there any overlap in the three GSEA analyses?

Evaluation criteria

The project will be evaluated on two main criteria:

1. Correctness
2. Completeness
3. Clarity in the presentation of the results

Each member of each groups will be asked to answer questions about the project and about all subjects discussed in the course.