

Computer Lab 4

Bioinformatics

Linköpings Universitet, IDA, Statistik

2024 XII 05

Kurskod och namn:	732A51 Bioinformatics
Datum:	2024 XII 05—2024 XII 09 (lab session 5 XII 2024 SU04)
Delmomentsansvarig:	Krzysztof Bartoszek, Hao Chi Kiang
Instruktioner:	<p>This computer laboratory is part of the examination for the Bioinformatics course</p> <p>Create a group report, on the solutions to the lab as a .PDF file. Be concise and do not include unnecessary printouts and figures produced by the software and not required in the assignments.</p> <p>All R code should be included as an appendix into your report.</p> <p>In the report reference ALL consulted sources and disclose ALL collaborations. The report should be handed in via LISAM (or alternatively in case of problems e-mailed to ying.luo@liu.se), by 23:59 9 December 2024 at latest.</p> <p>Notice there is a deadline for corrections 23:59 2 February 2025 and a final deadline of 23:59 2 March 2025 after which no submissions nor corrections will be considered and you will have to redo the missing labs at the next course opportunity.</p> <p>The report has to be written in English.</p>

This assignment is based on the material provided from the R Bio: Untangling Genomes course, by Martin Morgan, Houtan Noushmehr, Tathiane Maistro Malta and others that took place in Montevideo (2015.10.05 – 2015.10.10) <https://www.bioconductor.org/help/course-materials/2015/Uruguay2015/> . You should go through all the material of this tutorial. The following page, <https://gist.github.com/hckiang/ba6bdabdc0b401c29077d752d6cc11c> , might be useful in case of problems with installing some of the required software.

On the webpage <https://www.bioconductor.org/help/course-materials/2015/Uruguay2015/day3-gene.expression.html> you can find a step by step analysis of gene expression data from HUVEC¹ and Ocular Vascular Endothelial² Cells. The data itself (as indicated in the analysis) can be found at <ftp://ftp.ncbi.nlm.nih.gov/geo/series/GSE20nnn/GSE20986/suppl/> . The description of the data can be found at <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE20986> .

It might be helpful if you also go through https://www.bioconductor.org/help/course-materials/2015/Uruguay2015/day5-data_analysis.html and <https://www.bioconductor.org/help/course-materials/2015/Uruguay2015/V6-RNASeq.html> . The latter for graphic producing code.

Question 1

Run all the R code and reproduce the graphics. Go carefully through the R code and explain in your words what each step does.

HINT Recall what a design/model matrix is from linear regression.

Question 2

In the presented analysis, there are no plots of raw paired data. In the section where the contrasts are defined find the three contrasts. Present the variables versus each other original, log-scaled and MA-plot for each considered pair both before and after normalization. A cluster analysis is performed on the page but not reported. Present plots and also draw heatmaps.

Question 3

The volcano plot is only for `huvec` versus `choroid`. Provide volcano plots for the other pairs. Indicate significantly differentially expressed genes. Explain how they are found.

Question 4

Try to find more information on the genes that are reported to be significantly differentially expressed. The place to start off is <https://www.ncbi.nlm.nih.gov/gene/>, remember that the data is from the species human. Try to look also for other databases where (some) information on the genes may be found. Try to follow on some of the provided links. Report in your own words on what you find.

¹https://en.wikipedia.org/wiki/Human_umbilical_vein_endothelial_cell

²<https://en.wikipedia.org/wiki/Endothelium>

Report all the Gene Ontology (GO) terms associated with each gene. Are any of the GO terms common between genes? If so do the common GO terms seem to be related to anything particular? Try to present GO analysis in an informative manner, if possible visualize.