LSKD2016: Microarray Data Analysis with R

Shweta Bagewadi Kawalia

General: This lecture/exercise is meant to train you to be self-responsible and work independently. Before asking, try to solve problems by yourself or together with your collegues. It is highly necessary to understand things (not just to memorize them). If you don't understand something, read the help pages, the package vignette (after library(Biobase) type openVignette()) or search on the web. If after that there are still unclear issues ask the "experts of the day" or the teaching staff.

Task 1: Workflow preparation, test data retrieval, literature research

1. Make yourself familiar with the Editor and the R-environment on the Desktops provided in the lab. Or you can use R-environment/R Studio on your private laptops
2. Please define the OS-dependent library path ".libPaths("W:/R/win-library/2.13")" at the top of your script
3. Go to www.ncbi.nlm.nih.gov/geo/ and read about the GEO database
4. Who is the owner of the database?
5. What kind of data is inside of the database?
6. Search for the dataset "GSE9633" and download the raw data (CEL-files packed in tar.gz archive)
7. Unpack it on your computer
8. What kind of dataset is it?
9. What is the topic of this publication? When was the last update?
10. Use function "ReadAffy" in package "affy" to read the data into R. This should generate an AffyBatch object in your workspace
11. How is the data organized?
12. Gene expression values from this object can be accessed using function "exprs".
13. Information on the experimental design/"pheno data" (i.e. which sample is hybridized on which chip) can be accessed via function "pData"
14. Does the phenotype information you see here coincide with the information on the GEO webpage?
15. If not, generate yourself a csv-file in OpenOffice or Excel, in which you enter the corresponding information manually.
16. Read the file into R via "read.csv"
17. Assign the correct pheno data to your R object. Tip: Type "?pData" to access the manual page for "pData"

Optional task

1. Create a heatmap of the data and store it into a pdf-file
2. Discuss the heatmap properties and how to interpret
3. What are different distance measures and clustering methods that can be used to generate heatmaps?
4. Search and download some more dataset from the homepage and discuss their properties/heatmaps (e.g.: GSE1297)

Task 2: Implementing variance stabilization (VSN)

1. Create an “ExpressionSet” object of the expr/pheno data generated in Task 1, which is needed for MAplots.
2. Explain “ExpressionSet” object
3. Make MA-plot (bioconductorpackage “*lumi*”) and meanSdPlot (Tip: type “?MAplot” and “?meanSdPlot” for help) to compare “sensitive” (s) and “resistant” (r) among each other. Define the quality of the data using the plots (plotting such huge amount of data always takes some time, so be patient and wait)
4. Explain MA-plot and meanSDPlot. Do you know additional plots you could make to visualize/analyse the data?
5. Store all plots in a single pdf-file using function “pdf”. What do you see? Can you interpret the results? Remember, what you have learned in the lecture “Algorithmic Bioinformatics”. You can also have a look at http://compdiag.molgen.mpg.de/ngfn/pma2008mar.php and read about “Quality Control, Normalization and Design” to gain understanding. Moreover, you are free to search for any other type of information on the internet
6. Normalize your data using function “vsnrma” in package “vsn”. What does this method do?
7. Apply the same using “mas5” package. What does the method do? How is it different from vsnrma?
8. Create the above-mentioned plots for your normalized data (using vsnrma and mas5) and store them in a different pdf-file
9. Compare the plots of the normalized data with the ones you made before. What are the differences?
10. What are the differences when you try to compare the two normalization package results?
11. Now learn about the package “ArrayQualityMetrics” https://bioconductor.org/packages/release/bioc/html/arrayQualityMetrics.html
12. Apply the quality check for the normalized and raw data using this package

Optional tasks

1. Are there additional kinds of plots for normalized data?
2. Compute cellline distances via function “dist”. Use different method’s (Euclid, maximum ...) for distance measures
3. Generate a hierarchical clustering (average linkage method) from cellline distances (function “hclust”) and plot it
4. Compute gene distances via function “dist”. Use different method’s (Euclid, maximum ...) for distance measures
5. Plot the first 10-50 genes in a dplot (tip: use package “hopach”)
6. What do you see? Does it make sense to plot the first 10-50 genes only? If not, why?

Task 3: Statistical Significance Analysis

Part 1:

1. Search for the dataset of Golub et. al (integrated in package *multtest*). What kind of dataset is it? What are phenotypes?
2. Load data from Golub et. al into R
3. Apply SAM-Analysis to the dataset using the "*siggenes*" package. Use "control = samControl(delta = seq(start, end, step))" to find a delta for which around 100 genes are called. Please use 500 permutations and 123 randomizations to make the results reproducible. Use the function "summary" to view the results.
4. What does rand-parameter do? How does it help in terms of reproducibility?
5. Plot your results for different deltas, save them to a pdf file and try to interpret/discuss them
6. Use the function "list.siggenes" to get a list of the called genes for the delta you chose.
7. Apply SAM to the normalized GSE9633 data (generated on Day 1) (derived from both raw using the load() function)
8. Is there a quicker way to find delta value for top 100 genes?
9. Search for a package to retrieve further information on the called genes, namely Entrez gene ID, gene symbol and gene name. Tip: Use "entrezID <- hgu133plus2ENTREZID[genesSAM]" and the function "toTable" to view your results
10. Write out your result table via "write.csv"
11. Please discuss the results by comparing them to the original publication. What is the significance of the obtained results?

Part 2:

1. Apply linear model using lmFit function from *limma* package and eBayes to normalized GSE9633 data (derived from raw)
2. Obtain top 100 genes using *limma*
3. Please discuss the results and compare them with the results from SAM

Task 4: Implementing Gene Set Enrichment Analysis (GSEA)

1. Register in MSigDB (http://www.broadinstitute.org/gsea/register.jsp). Browse the website and search for different gene sets and gene sets collections.
2. Do a short literature search about the Bioconductor *PGSEA* package and install it
3. Also install and load the package *GSEABase*
4. Browse GEO for the details of the data set GSE7023 and describe it briefly (disease, phenotype, ...)
5. Load the data set GSE7023 from GEO into R using "library(GEOquery)" and "gse <- getGEO("GSE7023", GSEMatrix=TRUE)"
6. The data retrieved by the getGEO query is already pre-processed and normalized
7. Apply GSEA to the loaded data (use the VAIgsc gene-set collection available in the PGSEA package, load it via data(VAIgsc) and use the subtype "No" (Normal) as the reference "ref=which(subtype=="NO")")
8. Write details about VAIgsc, what does it contain? Who generated it?
9. Create a smcPlot, store it to a pdf-file, and discuss the plotted results
10. Plot your results using smcPlot and store it to a pdf-file
11. Discuss the results (Tip: there are PSGEA tutorials available: Google for PGSEA.pdf and/or PGSEA2.pdf)

Task 5: Co-Expression Analysis

1. Install the WGCNA Package (Weighted Gene Coexpression Network Analysis) and step through the tutorials on network construction and analysis provided on: http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/Rpackages/WGCNA/
2. Download the Female data set
3. Clean the data and check for any missing values. Perform a hierarchical cluster analysis and plot the dendrograms using *flashClust* package (http://danifold.net/fastcluster.html), using *Euclidean* as the distance metric. What is the number of modules you would infer from the dendrograms for each data set? Are there any outliers?
4. Perform a TOM analysis (Topological Overlap Measure) using *TOMplot* and plot the TOM diagrams. What do these diagrams suggest as the numbers of modules?
5. Perform a correlation analysis for selected genes from the prospective modules and plot their coexpression diagrams: what is the number of modules you would suggest for each data set using these correlation diagrams? Are they different from the number of modules suggested by cluster analysis? Why?