CytoReason - Exercise for bioinformatics candidates

Contact: renaud@cytoreason.com

Skills assessed:

- Ability to analyze an RNA-Seq dataset and graphically present the outputs;
- Ability to explain, document and organize an analysis workflow and its results;
- Proficiency in using the R/Bioconductor ecosystem;
- This is *NOT* a timed exercise, but we expect it to be doable in no more than 2 days after the Q&A session with CytoReason.

Data: Public RNA-Seq data of lesional psoriatic and normal skin (source: GSE54456)

Provided data as text files:

- Gene-level raw count matrix
- Sample annotation file
- Gene annotation file

Setup and Outputs:

- The analysis should be performed in an R/Bioconductor environment.
- A fresh Github repository that you share with us, where you **regularly** commit/push your work as you progress through the exercise.
- A single script file that performs and document your analysis, as well as generate the output files.
- Make sure your code includes detailed explanations/comments.
- You are free/encouraged to use any package you deem relevant, however those should be publicly available to ensure that your analysis is reproducible.
  We recommend using the packages `edgeR` and/or `limma`. Other packages/functions you may find useful for this exercise are `ggplot2`, `pheatmap` or `NMF::aheatmap`, `affy`, `ggfortify`, `ggrepel`.

Instructions:

1. Read and understand the exercise.
2. Get in touch with us to setup a meeting (by skype or at the office), in order to ask questions and make sure everything is clear about the exercise before you start.
3. Setup your environment (R, Git).
4. When ready to start the exercise, create the Github repository and share it with us. Try performing the exercise all in one go.
5. Feel free to contact us if you have other questions while busy with the exercise.
6. Send us an email with the final commit hash when you are done.

Exercise: *Use RNA-seq data to find genes which are significantly differentially expressed between two conditions*

- Load the data into R and make sure the count and annotation data are consistent with each other.
- Filter the count data for lowly-expressed genes, for example, only keep genes with a CPM >= 1 in at least 75% samples, in at least one of the groups.
- Generate an object that contains the library-size normalized log-CPM data. Save it as a binary file (.rda or .rds).
- Generate basic plots of your choice to investigate its main properties (library sizes, densities, PCA coloured by group, etc...).
- The PCA plot may suggest the presence of outlier/mis-labeled samples in this dataset. Try to identify them and remove them from the downstream analysis.
- Run a differential expression analysis comparing lesional vs normal samples. This can be done according to your preference either on the count data or the normalized log-CPM data, using appropriate statistical method.
- Export the results in a tab-separated text file: a table with genes in rows along with gene annotations and any relevant statistic.
- Select the top 100 most significant *annotated* genes and generate a heatmap of the log-CPM data, with samples in columns, annotated with the group variable.
- Generate a volcano plot (x-axis is the effect size and y-axis is the p-value) for this analysis. The selected 100 most significant genes should be colored.

Select at least one of the following in addition to the above exercise:

- Use `ExpressionSet` objects to hold all the provided raw data. Do the same for the normalized log-CPM data.
- Use the `ggplot2` package to generate some of the plots
- Include other steps or plots you think are useful/interesting in getting insights from the data or the results.
- Write your analysis as an R-markdown script and generate a pdf report.
- Generalize you code by writing and using a function that takes the provided count matrix, a sample annotation data frame and a variable name, and returns a data frame of gene statistic of differential expression.
- Run a gene set enrichment analysis on the results of the differential expression analysis. Use only one collection of gene sets.