SysBio 2020 - Advanced Computational Systems Biology course

The following exercises are meant to be performed along the afternoon practical session.

You're encouraged to follow good coding practices and learn to find solutions to the problems independently (hints and examples will be given). You are also encouraged to set up a local `git` repository to keep track of your changes. You may do that by opening a terminal on your working directory and typing:

```
git init [name_of_your_repository]
```

You can commit your changes as you see fit (every functional change) or after each exercise for example, with:

```
git add [name_of_the_file/path]
git commit -m "[commit_message]"
```

Example solutions will be provided at the end of the day.

Exercises day 1 - Omics data types, resources and tools

1) Installing packages:

Open a new R script or notebook file and install the following packages from Bioconductor: `biomaRt`, `Biobase`, `GEOquery` and `qusage`. When finished, load these in the current session. *You can find instructions for installation in the packages' Bioconductor web page*.

2) Loading data set from GEO:

The `GEOquery` package provides functions to download and manage data sets stored in the GEO database. As an example, in the following days we will work with a subset of the NCI60 study[1] (GEO ID: `GSE116436`). Here they studied 60 different cancer cell-lines and their transcriptomic changes after treating them with 15 different drugs. We will focus on the Lapatinib treatment whose ID is `GSE116445`. Load this data set in R along the corresponding annotations. *You can do all this with the function* `getGEO()`

3) Explore the data object:

After downloading you may explore the data object to check if downloaded properly and familiarize with the object and data themselves. *Some functions you might find useful:* `class()`, `is()`, `dim()`, `head()` / `tail()`, `attributes()` ...

4) Subsetting the data:

In the following days we will work with a subset of this data set. Select those samples with highest drug concentration (10000nM), the minimum (0nM, will use

as controls) and latest time point (24h). *To do so, you may use the* `title` *attribute from the* `ExpressionSet` *object which contains the tags of the different samples*

5) Retrieve/locate the expression and annotation data:

`Biobase` package provides functions to work with `ExpressionSet` objects. Use `exprs()`, `fData()` and `pData()` to assign variables to the expression, feature and phenotype data. Take some time to familiarize with the information contained. *You can use some of the functions of the previous exercise as well as* `rownames()` *and* `colnames()`

6) Identifier conversion:

The identifiers of our expression data set are probe IDs. Since we will need Gene Symbols in the future and they're easier to read, you must convert them. First create a mapping table with `biomaRt` and then map the probe identifiers to Gene Symbols. You'll find that sometimes different probes correspond to a same gene, therefore when substituting the IDs you must also collapse all probes of a same gene (e.g. mean).

7) Saving data:

Save the expression data with the mapped IDs as a plain text file (csv, tsv, txt…). Remember to keep the column and row names. Save also the phenotype data matrix for the upcoming days.

8) (Extra) Retrieving pathway annotation:

Now we will map our measurements to a specific pathway. MSigDB provides information that maps genes/proteins to pathways (gene sets). Go to http://software.broadinstitute.org/gsea/msigdb/ and download the Biocarta curated pathways (with Gene Symbols) in `.gmt` format. Load the gene sets in R and subset the information for the MAPK pathway (`BIOCARTA_MAPK_PATHWAY`). `.gmt` *files can be read with the* `read.gmt()` *function from the* `qusage` *package.*

9) (Extra) Obtaining PPI network:

Now, download the protein-protein interaction network (PPI) from OmniPath with Gene Symbols (http://omnipathdb.org/interactions?genesymbols=1, you can read the table directly from R). Then subset the network to contain only interactions which involve the members of the pathway we extracted in the previous exercise.

10) (Extra) Mapping expression to pathway:

As an example, take the median expression of the nodes in the network and save it in a separate file. *You may iterate over the* ***unique*** *set of genes of the pathway*

11) (Extra) Visualizing with Cytoscape:

Open Cytoscape and import the pathway file as a network. Select the first column as source nodes and the second one as target nodes. Then import the median expression as a table of node attributes. Finally you can color the nodes according to the expression level.

References

1. Monks A et al. The NCI Transcriptional Pharmacodynamics Workbench: A Tool to Examine Dynamic Expression Profiling of Therapeutic Response in the NCI-60 Cell Line Panel. Cancer Res 2018 Dec 15;78(24):6807-6817. PMID: 30355619