*Chair of Experimental Bioinformatics*
*TUM School of Life Sciences Weihenstephan*

**Lecture: Advanced Bioinformatics**
**SoSe 2022**

# Exercise Sheet 05 - non-coding RNA

Total: 15.0 points 13.06.22 - 21.06.22

## 1. microRNAs… (2.5 points, 0.5 each)

- [x] … repress gene expression
- [ ] … enhance gene expression

…find their target…

- [ ] … based on random binding
- [ ] … by binding in GC-rich regions
- [x] … by binding in a seed match region
- [ ] … do not need to bind to be effective

…act…

- [ ] … alone
- [x] … in a complex with an Argonaute protein

…differ from siRNAs which show

- [x] … perfect complementarity and high specificity
- [ ] … imperfect complementarity and low specificity

target prediction

- [x] … suffers from false positives
- [ ] … suffers from false negatives

## 2. miRBase (4.5 points)

Go to http://mirbase.org which is the reference database for microRNAs. Find the entry for hsa-miR-23a-3p and answer the following questions

a) What is the accession number (0.5):

MIMAT0000078

b) What is the accession of the stem loop (0.5)?

MI0000079

c) What is the genomic position of this miRNA (0.5)?

chr19: 13836587-13836659 [-] (in GRCH38)

d) hsa-miR-23 is part of a microRNA cluster. Which other miRNAs are found there (0.5)?

hsa-mir-27a, hsa-mir-24-2

e) hsa-miR-23a-5p has only 1547 reads reported in contrast to over 3 million reads for the -3p variant. Can you explain this discrepancy? (1.0)

f) What is the top-scoring target gene of hsa-miR-23a-3p according to TargetScan (0.5)?

ZNF225

g) What is a miRNA family? (1.0)

- x … miRNAs with the same stem loop
- … miRNAs with identical seed sequence

## 3. Competing endogenous RNAs (1 point)

In the original paper, competing endogenous RNAs were explained as a messaging system. Read the landmark paper found at A ceRNA Hypothesis: The Rosetta Stone of a Hidden RNA Language? - ScienceDirect and answer the following questions:

a) miRNA binding sites are called (0.5)

- … miRNA response elements
- x … miRNAs control elements

b) According to the ceRNA hypothesis, miRNAs act (0.5)

- … in cis
- x … in trans

## 4. SPONGE (2 points)

The method SPONGE that was presented in the lecture can be used to infer competing endogenous RNA networks. Follow the vignette found at SPONGE .
(if the regular installation for this package as mentioned in this vignette fails, try out 'devtools::install_github('biomedbigdata/SPONGE')' inside your R console)

a) How does SPONGE identify gene-miRNA interactions (0.5)

by using user-provided miRNA target annotation, e.g. from TargetScan
by computing an elastic net regression model with the gene as response variable and the miRNA expression vectors as explanatory variables.
x both

b) How does SPONGE assign p-values to interactions (0.5)

it uses a standard t-test
it uses a Wald test
it uses a null model in which mscor values are simulated under the null hypothesis that the partial correlation differs from the correlation of the genes
x it uses a null model in which mscor values are simulated under the null hypothesis that the partial correlation equals the correlation of the genes

c) Follow the individual steps of SPONGE but use mircode_symbol instead of targetscan_symbol. Do the results change? If yes, how? (1 point)

Yes the results change.
As the prior knowledge of predicted targets change, the gene-miRNA interaction candidates also change.
This leads to different predicted interactions and therefore a completely different interaction network.

# 5. SpongEffects (5 points + Bonus)

SpongEffects is a novel method that infers subnetworks from pre-calculated ceRNA networks and can calculate sample specific scores related to their regulatory activity. It comes as an R package and is installed together with the SPONGE package from the previous task 4.
Here we will look into subnetworks of Testicular germ cell tumor (TGCT) using SPONGEdb (https://doi.org/10.1093/narcan/zcaa042) and spongEffects (https://doi.org/10.1101/2022.03.29.486212).

For this task, you will **use R markdown to generate a HTML report** that includes each step of the analysis along with visualisations. You can find a template in the moodle exercise directory. Use this spongEffects vignette as a guide: spongeEffects.Rmd

a) **Data acquisition and preparation (1.0)**
   Your first task will be data cleaning. Download the oldCohort.RData file from Moodle, which contains three tables: a gene expression table, a miRNA expression table and a metadata table, in which the cancer subtypes are indicated. The annotated dataset consists of 84 samples. We will not use the miRNA expression table in this task.
   As spongEffects uses machine learning in a later stage, you will also need to create a training and testing subset for the metadata and gene expression. Be sure that the subsets contain the same samples in the metadata and expression table and that the cancer subtype proportions are similar to the full dataset. The train and test sets should be of roughly equal size (50-50 split).
b) **Sponge modules (1.0)**
   Download the correct ceRNA network from SPONGEdb. Load it into R (see the R markdown template for which files you need) and perform the network filtering, get 100 central modules for lncRNA and define the Sponge modules by following the steps in the spongEffects vignette.
c) **Module enrichment and Machine learning (1.0)**
   Enrich the Sponge models using your train and test gene expression tables. Calibrate the training module and use it to predict the labels of the scaled test modules. Fill out the print chunk, where you give the sensitivity, specificity and balanced accuracy for your predictions.
d) **Visualisation and Interpretation (2.0 + Bonus)**
   Use the three visualisation methods `plot_top_modules` , `plot_density_scores` and `plot_heatmaps` to interpret your results. Add the interpretation as plain text below the corresponding plot in the markdown file.
   You can also try out different parameter settings in the steps above (e.g. train/test split size, filtering parameters, number of modules,...) and explain how they affect your results.