Written project pitch

Full list of author information is available at the end of the article

1 Project title

Integrative Analysis of miRNA and mRNA Expression Profiles in Squamous Cell Carcinoma.

2 List of group members

Aakash Nepal (5229156) & Yunus Emre Kurnaz (5184877)

3 Project

3.0.1 What is the scientific question you want to answer?

The scientific question we want to answer with this project is whether we can determine a miRNA-mRNA interaction network based on the given miRNA and mRNA data and determine the key genes for squamous cell carcinoma.

3.0.2 What inspired you to work on this project?

Both group members have worked with similar datasets and analysis techniques in their research group as well as in their bachelor's thesis, so this project seemed to provide a familiar basis for this type of analysis.

3.0.3 What types of data will you analyze?

We mostly will use affymetrix microarray datasets and the gene expression microarray datasets aswell as TCGA data (miRNA and mRNA).

3.0.4 What computational/statistical methods do you plan to use to analyze the data?

Limma and DESeq2 $-->$ differential gene expression analysis

Elastic net and LASSO $-->$ Machine Learning algorithms for extracting the key genes

MirTarget $-->$ Data Integration of miRNA into mRNA

DAVID or STRING $-->$ Data Annotation

3.0.5 What diseases/organisms/tissues will you study?

We will study the microRNA and mRNA profiles of Esophageal squamous cell carcinoma (ESCC) patients in healthy controls and affected patients with ESCC. Therefore, the tissue we will work on is the esophageal.

3.0.6 List related work

[1] Arnold M, Sierra MS, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global patterns and trends in colorectal cancer incidence and mortality. Gut. 2017;66(4):683–91.

[2] Wu Y, Hu L, Liang Y, et al. Up-regulation of lncRNA CASC9 promotes esophageal squamous cell carcinoma growth by negatively regulating PDCD4 expression through EZH2. Mol Cancer 2017;16:150.

[3] VYang M, Liu R, Li X, et al. miRNA-183 suppresses apoptosis and promotes proliferation in esophageal cancer by targeting PDCD4. Mol Cells 2014;37:873–80.

[4] Chou CH, Shrestha S, Yang CD, et al. miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. Nucleic Acids Res 2018;46(D1):D296–302.

[5] Shibata A, Matsuda T, Ajiki W, et al. Trend in incidence of adenocarcinoma of the esophagus in japan, 1993–2001. Jpn J Clin Oncol 2008;38:464–8.

3.0.7 How does your project differ from related work?

Our project aims to identify an interaction network from miRNA and mRNA data of hub genes identified by advanced machine learning techniques. The dataset we are using is from esophageal squamous cell carcinoma. These will be merged with the TCGA data and then a survival analysis will be performed.

4 Data

4.0.1 Where will you get your data from?

The dataset can be downloaded from TCGA and the NCBI GEO repository.

4.0.2 How many samples are available?

For the miRNA data we have in total 819 (331 healthy patients and 488 diseased patients). As for the mRNA datasets we have 113 in total (55 healthy patients and 58 diseased patients).

4.0.3 Is the data multi-modal (multi-omics)? Is all data coming from the same study?

Yes the datasets are multi-omics, since we use both miRNA and mRNA expression data, but the datasets are from different studies.

4.0.4 In what format is the data available?

The datasets can be downloaded as txt files or can be directly loaded with the GEO library in R (GEOquery).

4.0.5 Do you expect to spend significant amount of time on pre-processing the data?

Since some of the datasets are not preprocessed, we will first try to filter them, but if according to our established deadline it does not work, we will use only the datasets that do work in order to keep our schedule.

4.0.6 Are you sure there are no restrictions on data access?

The miRNA and mRNA datasets are available for public download on the NCBI GEO website and The Cancer Genome Atlas Program (TCGA) is a public database.

5 Summary of the project plan

Esophageal squamous cell carcinoma (ESCC) is a malignancy that poses a serious threat to human health and has a high incidence rate and a low 5-year survival rate. MicroRNAs (miRNAs) are generally considered to have an important regulatory function in human cancer, but the but the potential regulatory mechanisms of miRNA-mRNA in the context of ESCC are still poorly understood.

Therefore, with this project, we aim to establish a miRNA and mRNA interaction network to reveal possible relationships. After the data exploration phase ends, we will start with the actual analysis. First, using Deseq2 and Limma, two different differential gene expression algorithms, we will try to identify the differentially expressed genes. This will be applied to both the miRNA and mRNA datasets.

Since both algorithms work based on different distributions, we want to use an intersection to filter out the most important genes of both analysis steps.

In order to use the miRNA data for further analysis, we need to perform a data integration using tools such as MirTarget. This will convert the miRNA into target mRNA. With the help of these, we will be able to create a miRNA and mRNA interaction network . Then the mRNA data from the previous step and the newly generated target mRNA are intersected again.

Based on this data, the next step is to apply various machine learning techniques such as Elastic net to determine the key genes, which are validated with ROC curves.

With the help of DAVID or STRING, these genes will then be annotated and also, the miRNA and mRNA interaction network will be created. However, if there is still time left, we will consider doing an additional survival analysis and look into methylation analysis for those key hub genes with the help of TCGA data.

6 Time line

6.1 WEEK 1

- Acquiring the data
- Download and install necessary packages
- Data exploration

6.2 WEEK 2

- Data cleaning, if needed.
- Start with the first DGE with Deseq2 and Limma for miRNA
- Start with the first DGE with Deseq2 and Limma for mRNA
- Intersect results of both analysis separately

6.3 WEEK 3

- Data Integration with MirTarget
- If previous step worked, then overlap these results with the Differential Expression Analysis of mRNA data (Data Integration)
- Apply the machine learning technique and identify the key miRNA and mRNAs

6.4 WEEK 4

- Validate the results using ROC validation
- Visualizations, Annotation and miRNA-mRNA interaction network

6.5 WEEK 5

- If possible try to create Kaplan Meier Curves with the TCGA clinical data and our results
- Work on the the report and presentation

6.6 WEEK 6

- Finalize the analysis
- Finalize presentation
- Finalize project report

Author details

References