DIGITAL EPIDEMIOLOGY AND PRECISION MEDICINE
Project

Differential Analyses of Gene Expression[1]

A. Student, B. Student, C. Student

GROUP <nn>

Abstract. Brief and clear paragraph: it should be a miniature of the manuscript including a brief description of: scientific issue and aim, methods, main result and conclusion.

INTRODUCTION

This section includes: the definition of the scientific issue, the summary of main results in the scientific literature (state of the art) and of principles needed for the comprehension of the new hypothesis; the identification of the hypothesis, the disclosure of the relevance of the results and the overview of contents of following sections. In the introduction the aims of the study must be clearly stated.

It does not include conclusions and recommendations

MATERIALS AND METHODS

Describe in this section the experimental procedures and resources, data analysis procedures and statistical methods. Give enough detail to replicate the experiment but do not overwhelm the reader with too many details.

To do:

1. Data

   Download gene expression data from https://portal.gdc.cancer.gov/ (data category: *Transcriptome Profiling*; data type: *Gene Expression Quantification*; workflow type: *STAR - Counts*) selecting the project in GDC data portal according to the shared spreadsheet [2] (for example group 01: Breast Invasive Carcinoma **Project ID**: TCGA-BRCA) and select only patients for whom cancer and normal tissue files are available.

2. Differentially Expressed Genes (DEGs)

   Identify DEGs specifying the thresholds setting. Select the thresholds to obtain a subset of hundreds of genes: p-value threshold should be less than or equal to

[1] a new title can be proposed (it is not required). A good title should be concrete, clear, and brief.
[2] https://docs.google.com/spreadsheets/d/11wDGYtbHUF9LNzUvMv_T5R3xcwgCes13QH2Z7jHgg7w/edit?usp=sharing

0.05 (it is suggested to apply a correction for multiple comparisons) and Fold Change (FC) threshold: $|FC| \geq 1.2$.

3. Co-expression networks
    - *Computation.* Using only DEGs, compute the gene co-expression networks related to the 2 conditions (cancer, normal) considering:
        - Pearson's correlation (or another measure of similarity);
        - Binary adjacency matrix where $a_{ij}=0$ if $|\rho| < threshold$ [3].
    - *Analysis*:
        - Compute the degree index and check if the network is a scale free network
        - If so, find the hubs (5% of the nodes with highest degree values), compare hubs sets related to the two condition (cancer, normal) and identify the hubs selectively characterizing each network
4. Differential Co-expressed Network
    - *Computation.* Using only DEGs, compute the differential co-expression network (Cancer vs. Normal) following the procedure described in the slides (DEPM_5). Binary adjacency matrix with $a_{ij}=0$ if $|Z|<3$ [4].
    - *Analysis*:
        - Compute the degree index, check if the network is a scale free network
        - If so, find the hubs (5% of the nodes with highest degree values) and compare the identified hubs set with those obtained in task 3.
5. Patient Similarity Network (PSN)
    - Compute the Patient Similarity Network using cancer gene expression profile
    - Perform the community detection (e.g. apply Louvain algorithm to the PSN)
6. OPTIONAL TASKS
    - Compute a different centrality index (CI) and check the overlap between the 5% of the nodes with highest CI values and the degree-based hubs
    - Perform the study using a different similarity measure (Spearman correlation, biweight midcorrelation, etc.)
    - Perform gene set enrichment analysis (e.g., cancer hubs)

[3] For example, $|\rho|<0.7$ (where ρ is the Pearson correlation coefficient): this is a suggestion, but if with this threshold the obtained network is composed of many disconnected components or if it is too dense, a different value can be applied

- Perform task 5 using gene expression profiles related to normal condition and compare the community structures of the 2 conditions
- Perform PSN communities characterization (survival analysis, enrichment analysis using clinical data from the GDC data portal)

NB Check the steps in the annex1-*Project steps* to apply appropriate pre-processing of data

RESULTS AND DISCUSSION

Describe the obtained results with the help of figures and tables.

Expected figures:

- Volcano plot
- Degree distribution of the networks
- Subnetwork plot of the *most relevant* genes (e.g the node with the highest degree value in differential co-expression network)
- Network with highlighted community structure

Provide a discussion of the results (e.g. opinions, interpretation, relationship of your findings with other published findings) and implications that can be drawn from your findings.

(if possible) for one or two genes identified in task 3 (*hubs characterizing only cancer tissue*) and/or 4 (*hubs of the differential network*), provide some reference papers where that(those) gene(s) is(are) studied in the context of the same disease.

Produce a short report (up to 6 pages) following the guidelines. No template is provided. Use the freedom in choosing your format wisely. Body text font sizes smaller than 11pt should be avoided.

Please specify the used programming language, the available functions and, if necessary, those implemented. Do acknowledge any source you used, such as software code, third party figures, cited text, contribution by non-authors, etc.

Evaluation criteria

The mark will be based on: clarity of writing, accuracy of the methods' description, completeness and appropriate presentation of results (including quality of figures and tables), appropriate discussion of the outcome of comparisons, overall structure and format of the report.

Submission procedure

Each group will submit (by email with all the co-authors in cc) two files:

i. a PDF file containing the report;

ii. a compressed archive (zip, rar, 7z) containing the software code used to perform the analyses. The latter archive may contain intermediate results (unedited figures, raw output files, etc), if appropriate. Do not include downloaded data in the archive.

The files must be named according to the following scheme:

DEPM-1_proj_group-<nn>.<ext>,

where <nn> is the two-digit group number and <ext> is either 'pdf' or {'zip'|'rar'|'7z'}, e.g. DEPM-1_proj_group-01.pdf

Delivery date: Tuesday 12 December 2023

Annex1

Project steps

1. **Selection of the disease→** C, N data: genes x subjects; erase the genes (rows) with at least one zero value

2. **Differentially expressed genes (DEGs) C vs N:**
   - Fold Change: FC = log2(dataC/dataN);
   - statistical test C vs N
   - correction for multiple comparison
   - thresholds selection

3. **CoExpression Nets** (C, N square matrices: genes x genes): log-transform data using log2(x+1) before calculating the correlation

4. **Differential co-expression net** (Fisher z-transformation, z-score calculation and threshold selection)

5. **Patient Similarity Network**

NB: if you use a package (e.g. DESeq(2)) pay attention to the type of input data (raw or normalized counts)