

---

## DIGITAL EPIDEMIOLOGY AND PRECISION MEDICINE

### Project - part 1

---

## Differential Analyses of Gene Expression<sup>1</sup>

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/> [2] (data category: *Transcriptome Profiling*; data type: *Gene Expression Quantification*; workflow type: *HTSeq – FPKM*) selecting the project in GDC data portal according to the shared spreadsheet [3] (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

---

<sup>1</sup> a new title can be proposed (it is not required). A good title should be concrete, clear, and brief.

<sup>2</sup> available R script: <https://drive.google.com/file/d/1yyh2d9hH0ilBIQn3D8CMMiKaHRujMtzT/view?usp=sharing>

<sup>3</sup> [https://docs.google.com/spreadsheets/d/1CXiY\\_-QdBjGsiDhW5w169rLJdNaAjRElUtAXZ7Ninv4/edit#gid=0](https://docs.google.com/spreadsheets/d/1CXiY_-QdBjGsiDhW5w169rLJdNaAjRElUtAXZ7Ninv4/edit#gid=0)

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;
  - Binary adjacency matrix where  $a_{ij}=0$  if  $|\rho| < 0.7$  (where  $\rho$  is the Pearson correlation coefficient) [4].
- *Analysis:*
  - Compute the degree index, check if the network is a scale free network and 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 characterizing only cancer tissue.

### 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\_6\_Diff Co-Expression Analysis.pdf). 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 and find the hubs (5% of the nodes with highest degree values)
  - Compare the identified hubs set with those obtained in task 3.

**NB Check the steps in *Project steps.jpg* 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)

---

<sup>4</sup> 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

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: Monday 30 November 2020

---