upf. Aula Global

You are logged in as **JOAN FRANCESC GILABERT NAVARRO** (**Log out**)

Course 2015-2016

cat esp eng **other** La Factoria Aula Global Help

My home ► 2015-31032-T1 ► General ► Project description (1st part)

Navigation

My home
- Site home
- Site pages
- My profile
- Current course

Administration

Course administration

My profile settings

2015-31032-T1 IEO. Infomation Extraction from OMICS Technologies

Project description (1st part)

The final result of the project should be a report in the form of a scientific publication re-analyzing the data set that you have chosen, with a supplementary material containing all the scripts, data and figures that support the contents of the report and that make the entire analysis **reproducible**.

Technical requirements

The report should be written in LaTeX (see http://www.latex-project.org) and therefore one milestone shoud be that you get acquainted with the basics of elaborating a document using this text processing system.

If you work in a Linux system, LaTeX should be already installed, otherwise try at all times to install it through the package installation system of your Linux distribution (because of the many files involved). If you work in a Windows or Mac OS system, then I recommend you to install the TeXLive distribution (http://www.tug.org/texlive).

Once you have got LaTeX installed in your computer you can try to process and modify the template document you will find under the folder 'latexTemplate' link from the moodle site of the course. That should allow you to quickly grasp the very basics steps of creating a LaTeX document, but you can find lots of introductory material to LaTeX by Googling "latex beginner". However, in this first part of the project you **do not** have to deliver the report, although you are welcome to start wokring on it.

In this first part of the project you should work to produce a first part of the supplementary material in the form of a web page which should be created using markdown (see http://en.wikipedia.org/wiki/Markdown) mixed with R. Such a type of document is known as an "R Markdown" file (see http:/markdown.rstudio.com) and has extension .Rmd and it is processed through an R package called knitr. The library has one copy of a book explaining in-depth how to use 'knitr' (see

http://cataleg.upf.edu/search*cat/X?knitr), although the book is not required to learn and use R markdown and 'knitr'.

The materials you have to submit should be a tar ball (or zip file) analogous to the one called 'projectTemplate.zip' which you should find in the moodle site of the course. This file contains a template which you can modify to do your project. As an example, it contains a basic analysis of a TCGA RNA-seq data set that is not among the ones available for you on the projects page. You can see the final result of this template by opening the file 'projectTemplate.html' with a web browser. You can also find the source .Rmd file that produced the web page, and which include the instructions to process the .Rmd into an .html file as well as a few comments in the source about what some of the R-markdown tags and options mean.

The analysis in the template covers the basic initial tasks to perform on a RNA-seq data set. There is some interpretation of the results with suggestions about decisions to be made. However, you should make your own decisions on your data set based on the results you obtain. Every decision is fine as long as it has a sensible justification behind.

Content requirements

As for the particular contents that you should develop for this first part of the project, they should include:

1. Choose a RNA-seq data set among the ones offered in the data sets folder.
2. Read introductory material about the corresponding type of tumor you are going to analyze. Try to understand the major tumorigenic mechanisms that participate in the growth and proliferation of such a tumor type. This essentially means to gather information about what are the most relevant genes to this cancertype. The goal of the project **is not** to reproduce previous findings about this cancer type, although you may choose to do so, but rather to find some simple question related to this cancer type that you can answer focusing on the contrast between tumor and normal samples, or on some other simple contrast of interest using the analysis techinques we have seen in class. In this respect, you can consider analyzing some of the available clinical variables such as the tumor stage encoded in 'ajcc_pathologic_tumor_stage'. Each clinical variable has a metadata with a so-called 'CDE' identifier which you can use to fetch futher information about at https://cdebrowser.nci.nih.gov.
3. Try to figure out factors that generate variability unrelated to the outcome of interest. By means of the diagnostics we have seen the lecture on batch identification, try to ensure that you do not have a major confounding with the outcome of interest.
4. To speed up some parts of the analysis, you may choose to work with a subset of the samples.
5. Carry out quality assessment and normalization of the data.
6. Search for differentially expressed (DE) genes using the simple F-test implemented in the package SVA to do a two-group comparison. For this first part of the project *do not attempt* to interpret the list of DE genes, just report how many of them do you find and how the distribution of p-values looks like. Consider estimating surrogate variables with SVA to see whether the number of

expression changes increases or decreases. Bear in mind that the actual DE analysis you will do in the second part will be more sophisticated since you will have to take into account aspects such as variance heterogeneity of log CPM values or the fact that normal samples were derived from the same pool of individuals as a fraction of the tumor samples. For this first part of the project, there is no need to address these issues.

7. The project template that is provided is not comprehensive and it just tries to help you in quickly learning how to work with R Markdown files, so you should try to make your supplementary material more readable and complete than the template provided.

Deadline

Using the submission link you will find at the top of the IEO moodle site, submit the tarball (or zip file) of the directory structure containing the source R Markdown file, data, figures and resulting HTML by the **10th of may**, each fraction of 24h delay will be penalized with 1 point (out of the 10 possible). There is a maximum file size submission limit of 400Mb. Although it is very unlikely, if your zip file is larger, let me know and we will find a way to upload the data.

Last modified: dimarts, 26 abril 2016, 9:15