RDS Class Assignment- A Custom Workflow

Simon Tomlinson Nov 2023

In RDS we have built a number of workflows that load and process data and generate outputs that are typically simple visualisations or summary tables. In the class assignment you will build a novel processing pipeline for some individual data that you will be given. The aim is to

1. Build a custom workflow in RMarkdown to process the data
2. Process the data and generate a custom report

Requirements

1. Write a script that achieves all the objectives of the data processing (see below) **[50% of the marks]**
2. The script will be expected to comply with the style guidelines as well adhere to defensive coding techniques. **[20% of the marks]**
3. Generate a formatted document report from the R Markdown script. This document is only required to be a very brief description of the work. It should include images of the final heatmaps. It must be formatted as a University coursework submission **[20% of the marks].**
4. Provide an archive of the script the coursework and all of data required to run the script. This archive should contain the report **[10% of the marks].**

The report should be no more than **six pages long including figures**.

Detailed Data Processing Instructions

- You are provided with a set of data to process as well as several other files of data and annotation.
- The data set consists of an input data table “data_all.csv”. The data is numerical data with row 1 containing sample names and column 1 containing gene names. There is also a table “gene_annotation.csv” and “sample_annotation.csv” containing gene and sample annotation. Finally, a file is provided that provides a list of genes to use for plotting called “genelist_uun.txt”. This list is unique to you UUN, which is your University “s number” and each list is different.
- Load the input data tables into R. The aim is to load the data then convert the dataall to log scaled data. Select the required data for the plot -only those genes on the “genelist_unn.txt” need to be plotted. Each gene should only appear once on the final plot.

- In the plot annotate each gene row with the type of gene it is (XA, XB or XC) and annotate each sample with the treatment group that the sample was derived from.
- Before plotting the samples, rename the gene names with the “LongName” from the gene annotation table.
- Create two heatmaps, one where both the genes and the samples are clustered. Create one were only the genes are clustered
- Provide any other information that you think is informative about the processing or visualization and interpretation of the data
- At the end of the report, provide a brief few sentences to outline the interpretation of the heatmap plots.
- You can obtain files for the assignment on bioinfmsc6 on the path `/shared_files/RDS_assignment_files/<snumber>_files.zip` where <snumber> is your own student number.