Birkbeck, University of London
MSc Bioinformatics
Omics (Spring 2022/2023)
Coursework Assignment
Authors: Dr Igor Ruiz and Dr Irilenia Nobeli

Date of the assignment:	14/04/2023
Date assignment is due:	05/05/2023
Marking:	Highest mark possible (for the full coursework): 33

General Guidelines

The aim of the NGS coursework is to give you experience in analyzing genomic data from Illumina next-generation sequencing technologies. Your role as a bioinformatician is not simply to run programs with recommended parameters and pass on the results. Your role is to run and re-run analyses, making informed judgements about the results and improving your methods and pipelines, making use of your experience, shared knowledge and other people's expertise. This coursework aims to train your ability to do just that.

In order to carry out NGS analysis you also need to be familiar with using the command line in unix, understanding shell scripts and creating reports that make your work easy to understand and share and reproducible (we recommend markdown and R markdown files for this).

The coursework below tries to address some of these skills. There are 2 parts to it. Altogether, this coursework counts as **33%** of the mark for the module.

Guidelines for Coursework Part 2

Part 2 of the coursework will contribute **25%** to the total module mark.

Aim

The goal of this assignment is to give you the opportunity to explore and familiarize yourselves with some aspects of analyzing RNA-seq data, with a focus on carrying out differential expression with one of the most popular programs available (DESeq2).

Introduction

Irimia et al.[1] studied the alternative splicing of micro-exons in neurons and have found that this highly conserved process is widely dysregulated in the autistic brain. As part of their work, they carried out NGS sequencing of 12 autistic and 12 neurotypical brain samples and analysed their RNA-seq data to identify differentially expressed genes. The original paper includes a very large number of experiments as is typical of such high impact factor journals, but we are only interested in the part described under "Comparison of autistic and control brains" in the "Extended experimental procedures" of the paper.

Reference

1. Irimia et al. (2014). A highly conserved program of neuronal microexons is misregulated in autistic brains. Cell 159, 1511. *http://dx.doi.org/10.1016/j.cell.2014.11.035* . This paper is open access and so available to read/download.

What you DO NOT need to do

You do not really need to read or understand the original paper (this would be a lot of work!). This assignment is confined to the questions I'm asking below. Some of them refer to the findings regarding differential expression of the gene nSR100 using DESeq2. This gene, otherwise known as SRRM4 (Ensembl id ENSG00000139767), is a splicing factor that the paper suggests is differentially expressed in the autistic brain and so is responsible for the observed dysregulation of microexon splicing (see Figure 7b in the paper). Irimia et al. have used a different approach to estimate differential expression (their R script is provided, in case you are interested, but you do not need to run or understand this to complete the coursework) and hence your results are likely to be different to theirs. I am providing you two files necessary to carry out the analysis: the raw counts table and the metadata table.

What you need to do

You need to upload ONE file on Moodle containing answers to the questions below: the file should be an **R markdown script** so that it contains your code, your comments and the output of the code. Embed the Shiny App in your R markdown document.

The files you need for this part of the coursework are under the directory:
`/d/in4/u/ubcg71a/teaching/omics/coursework/part2`

Questions for Part 2 (15 marks)

Using the DESeq2 vignette as a guide:

i) produce an R script that will carry out differential expression analysis at the gene level using the counts provided by Irimia et al. (these counts are in the file

GSE64018_countlevel_12asd_12ctl_edited.txt). Use a GLM design that assumes the counts for a gene in any given sample are only dependent on the “diagnosis” column associated with that sample in the metadata table (your design formula should only have one variable). Write a one-line comment to explain each step of the analysis in the R script.

[3 marks]

ii) Report how many genes are differentially expressed (up and down-regulated) at an adjusted *p*-value cut-off of 0.05.

[2 marks]

iii) Produce a volcano plot of the results of the differential expression and label the top 10 genes (with gene names or IDs) with highest log fold changes (in either direction) and adjusted *p*-values < 0.05.

[3 marks]

iv) Produce a figure showing the counts for gene SRRM4 for all samples using figure 7b in the Irimia et al. publication as a guide to what the figure should look like. Use your judgement to decide whether or not to show raw counts and explain your choice.

[3 marks]

v) Compare your results with the paper’s results for gene SRRM4 and suggest reasons for any discrepancies.

[2 marks]

vi) Given that the samples originate from humans (hence differ a lot biologically) and given that sequencing statistics reported in the metadata.txt file show variations across the samples, suggest how the approach you followed in (i) could be improved. You do not need to write actual code for this, just explain in words what could be changed.

[2 marks]

viii) Create a Shiny App in R that will allow a user to explore counts for any gene, the same way suggested in (iv). In other words, the user should be able to type a valid gene name or ID (you can pick whatever you want to use, if you don’t want to use both) and have a figure displayed showing the counts of this gene for all samples in this dataset.

[10 marks]

Explanation of the files provided

metadata.txt : text file with metadata (22 variables in columns) for the 24 samples (in rows). Row names (in the first column) are the sample names and there is a header row.

GSE64018_countlevel_12asd_12ctl_edited.txt : Counts table submitted by Irimia et al. to GEO. This is in the usual format, with samples as columns and gene ids as rows.

GSE64018_FPKMnormalization.R: the R script submitted by the authors to the GEO database (this won’t run“out” of the box as you don’t have the same directory and file set up as they do). *This script is not required to answer the questions and you may ignore it.*