BIO407 / BIO627 – Biostatistics – Mini project

Rationale:

Learning by doing is far more effective than just listening to lectures, reading books or even having discussions about the topic. This mini-project is designed to give you ample freedom to explore your creative and problem solving abilities. Instead of focussing on example toy problems for which solutions can be found in a book, this project introduces a question that was recently solved and has various unsolved interesting open questions. To make the best out of this course you need to take initiative and try to explore new avenues.

One of the important points while using statistical tools is to be able to decide which tests are to be used for which kinds of experimental design and datasets. Although this is covered in the textbooks, it is suggested that you read this paper by Sullivan et al. (2016) that provides a flowchart representation of how to decide on the tests to be used. As the course progresses, you will be introduced to various concepts in statistics. The goal of this mini-project is for you to apply these concepts to a real world problem. To ensure that you start working on this project from day-1, you have been provided the material at the very beginning of the course.

Problem statement:

Nef (Negative Regulatory Factor) is a small 27-35 kDa protein that is coded by the HIV genome. This gene performs several different tasks during infection. However, despite being studied for more than two decades, it was not known how this protein increases the infectivity of new virus particles released from the cell. In the year 2015, two groups of scientists independently solved this 20 year old mystery and shed light on a pair of new genes that could restrict HIV in absence of the Nef protein (Rosa et al. 2015; Usami et al. 2015).

As part of this mini-project, first you will replicate the results from one of these papers (Rosa et al. 2015). The full text of the paper and data needed for the analysis can be found in the mini-project folder. This paper helped establish that Nef prevents the action of two host proteins with antiviral activity. Read the paper in detail to understand what was done. Briefly, they measured the expression level of all the genes in 15 cell lines using RNAseq. They also measured infectivity ratio of these same cell lines with Nef+ and Nef- HIV. Using this data, they looked at correlations between expression levels of each gene with the infectivity ratio's. Genes whose expression level showed a strong correlation with the infectivity ratio were chosen as candidate genes that were involved in restricting the virus in the absence of the Nef gene. Further experimental validation of these candidate genes lead to this interesting new discovery.

In order to replicate the result, you will need to "look at the data" using descriptive statistics and also visualise it using graphical tools. After this you will calculate correlations between gene expression level and infectivity ratio. Apply FDR corrections to make sure that multiple-testing does not lead significant results. Evaluate whether FDR correction has to be applied on the full dataset or you can think of a different approach to this issue. Represent your results in the form of scatter plots, boxplots, bar graphs, etc. using high resolution figures.

The next steps after replicating the results are largely dependent upon you. This will require you to understand the biology of the question and come up with hypothesis. You will then use your understanding of statistical tests to try and test your hypothesis.

Solution scheme:

You will be presenting your solution in class through a 10 minute slide presentation. Each individual will present their solution ***highlighting the novelty of their approach***. Email the final slide presentation before your presentation in class to ensure that it is graded.

The solution is to be presented in below scheme:

1. Hypothesise and justify your reasoning (see grading scheme below)
2. Approach used by you to test your hypothesis
3. What tests did you use and why?
4. Conclude with the final biological insight that you obtained

Grading scheme:

1. Replicate the result from earlier study (required to pass the course) – copied results will not be graded: Grade D
2. Come up with 1 new hypothesis based on biology & test it using statistical test. Grade C
3. Come up with many new hypothesis based on biology, test them (using statistical tests that demonstrate your familiarity with different tests). Grade B
4. Based on hypothesis testing find a biologically meaningful result & explain how you arrived at it. Grade A
5. Use additional publicly available data and integrate that in your analysis to obtain biologically "interesting" results. Grade O

REAMDE file from the Mini-Project folder:

```
1) CellLine_RNAseq_read_counts <- "Folder containing 15 files. Each file
corresponds to one of the cell lines used in the Rosa et al., paper."

Example: "SRR2166624.htseq"

This file contains two columns:

Column 1: Human Gene ID's from Ensemble website
(http://asia.ensembl.org/index.html)
Column 2: Number of reads sequenced for this gene using RNA sequencing

First few lines of the file are shown here as an example:

ENSG00000000003 0
ENSG00000000005 0
ENSG00000000419 535
ENSG00000000457 643
ENSG00000000460 885
ENSG00000000938 9

2) infect.txt <- "Infectivity ratio file Nef+/Nef- infectivity ratio" by
cell type
```

References

Rosa A, Chande A, Ziglio S, De Sanctis V, Bertorelli R, Goh SL, McCauley SM, Nowosielska A, Antonarakis SE, Luban J, et al. 2015. HIV-1 Nef promotes infection by excluding SERINC5

from virion incorporation. Nature 526:212–217.
Sullivan LM, Weinberg J, Keaney JF. 2016. Common statistical pitfalls in basic science research. J. Am. Heart Assoc. 5.
Usami Y, Wu Y, Göttlinger HG. 2015. SERINC3 and SERINC5 restrict HIV-1 infectivity and are counteracted by Nef. Nature 526:218–223.