

TOPIC: Bioinformatic analysis of publicly available RNAseq data**TAGS:** Capstone Project, Bioinformatics, RNAseq, Rstudio**BRIEF DESCRIPTION**

There are a growing number of RNAseq datasets that have been submitted to publically available databases. Often the data has been used to look at a specific subset of data within the Authors realm of interest. However, there may be useful data on gene expression in pathways more aligned to other researcher's areas of study. This provides scope to interrogate the data for projects and allow the person doing the work to develop skills in bioinformatics for which there is a growing demand within the life sciences.

The aim of the project is to interrogate publicly available RNAseq data that has been deposited in the Gene Expression Omnibus (GEO) database at NCBI (<https://www.ncbi.nlm.nih.gov/gds>) and reanalyse them for differential gene expression to identify genes and pathways that may be of interest. This is made possible because there is a publically available set of training videos on YouTube that go through the processes systematically, which could be easily modified to use with any of the datasets in the GEO database. Dr Simon Cockell of the Newcastle University Bioinformatics Support Unit has produced the video series. The series of videos are called lockdown-learning and are available here:

<https://www.youtube.com/channel/UC7aizSyonJqZI3O3U4SDTiA/videos>

The number of videos are growing, but the first 42 are required for this type of analysis.

The project would require the student to complete the training videos (42 hours) then apply the methods and scripting to a new data set from GeoDb. This would probably be better aligned to 40 credit project modules.

SKILLS DEVELOPED

The student would become familiar with the linux environment and the use of Bash scripting. Running the Xubuntu in Oracle virtualbox running on either a Mac or windows.

The student would become familiar with the statistical programming language R and running it in the integrated development environment Rstudio.

REQUIREMENTS

The student would need access to a PC or Mac with administrator rights or have the software installed on a machine.

The supervisor would have to familiar with R and ideally have run through the course himself or herself to help the student troubleshoot their modified scripts for new datasets.

AUTHOR

Dr Noel Carter, University of Sunderland – noel.carter@sunderland.ac.uk