**TCGA data wrangling for RNA-Seq transcriptome profiling data – A step-by-step tutorial**

Harry Smith

Step 0: Create a parent directory for data.

~/Documents/TCGA/TCGA-PRAD/

Step 1: Identify the dataset of interest and download the file manifest.

* Identify the Files to download:
* Navigate to the following link: <https://portal.gdc.cancer.gov/repository?facetTab=cases&filters=%7B%22op%22%3A%22and%22%2C%22content%22%3A%5B%7B%22op%22%3A%22in%22%2C%22content%22%3A%7B%22field%22%3A%22files.analysis.workflow_type%22%2C%22value%22%3A%5B%22HTSeq%20-%20Counts%22%5D%7D%7D%2C%7B%22op%22%3A%22in%22%2C%22content%22%3A%7B%22field%22%3A%22files.experimental_strategy%22%2C%22value%22%3A%5B%22RNA-Seq%22%5D%7D%7D%5D%7D>
* Then choose the appropriate project id filter to pick the project you would like to download
* Add these files to the cart by clicking the 'Add All Files to Cart' icon
* Navigate to the cart (the icon at the top right corner)
* Click on the download > manifest button. Put this in the parent folder for your cancer type (I labeled this the TCGA project ID (e.g. TCGA-PRAD)

Step 2: Download and merge counts data using the GDC-rnaseq-tool

* If you do not already have pip3 installed, you will need to do so now. You can do this with brewer. In the terminal type:
  + brew install python3
* You will then need to install 2 modules. In the terminal type:
  + pip3 install pandas
  + pip3 install requests
* Download the gdc-rna-seq-tool.py script (<https://github.com/cpreid2/gdc-rnaseq-tool/releases/download/1.0/gdc-rnaseq-tool.py>)
* Navigate to parent directory
  + cd ~/Documents/TCGA/TCGA-PRAD/
* In the terminal type
  + python3 ~/Documents/TCGA/TCGA-PRAD/gdc-rna-seq-tool.py ~/Documents/TCGA/TCGA-PRAD/gdc\_manifest\_20180419\_175110.txt
  + OPTIONAL: if you want HUGO gene symbols added you can add --hugo
* This will create a folder called Merged\_RNASeq\_SOMENUMBER. Inside you will find
  + Merged\_Counts.tsv – your merged count matrix
  + **RNA-Seq** – Folder containing the raw quantitation data from the download. You should only have data in the folder labled **HTSeq-Counts**

Things to note:

The merged dataset is a simple merge of all the files from the cart you downloaded. That is, some samples have multiple read files. If you read in your Merged\_Counts.tsv file into R, you will notice that the column names are different from the file names. These are the TCGA barcodes associated with the file. A description of TCGA barcodes can be found here (<https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode>) but basically what you will need is the 3rd element in the name. This is the subject ID. The number of unique values for these should match the number of cases there were on the TCGA website. You can then use whatever method you like to merge these columns in R based on those subject IDs.