BIOS274 Problem Set 2 – Part 1
Due before 9:00 am December 9, 2019

This problem set is in two parts. Part one is due this week. Part two is due by 9:00 am on Friday December 16th.

In part 1 you will parse files accessed via a URL and write selected information to files. In part 2 you'll use the information in these files. Example code is provided on how to open the remote files and stream their contents to your script without downloading the file.

You are interested in differential mRNA expression. You find a set of RNA-seq experiments using several different tissues. The results from nine of these tissues are selected as being interesting. The goal is to identify interesting patterns of expression between these tissues. The RNA-seq gene quantification files can be accessed via a URL and processed on the fly, so you don't need to download the files. This makes your script more reusable, sharable, and saves space on your computer. You will also need to parse the GENCODE chromosome feature annotation file. The GENCODE information is provided in the GTF (Gene Transfer Format) file, a useful format but slightly more difficult to parse than a plan TSV (tab separated values) file. A specific GENCODE URL is in the ps2_gencode_url.tsv file. This is the version of GENCODE used by the ENCODE RNA-seq computational analysis pipeline in defining the gene quantifications.

1. The gene quantification TSV files and the GTF file have headers. The first line of the TSV files defines what is in each column. At least define constants to specify the index to that column, do not hardcode the index number within the script. The header lines of the GENCODE GTF start with '#' and includes release information.
2. For the quantification files you only need to store gene ID, TPM value, and file accession number. Save this information to a new TSV, one file per quant file. Only save the data for those genes that have a TPM value greater than **0.5**. The accession number is saved as a convenience in part 2.
3. From the quant files only Ensembl **gene** IDs are needed, those that start with 'ENSG'. What are those other IDs?
4. All these data were created by the ENCODE project. The quant file name is an accession number. You can explore the metadata for the experiment associated with the quant files. Enter the file accession number into the search box on https://encodeproject.org/, or go directly to the information on the experiment with the file's accession as in this URL, eg. https://www.encodeproject.org/search/?searchTerm=ENCFF408GFZ
5. For the GTF file you only want information from the rows that have Source = HAVANA and Feature = gene. You'll find the other information you want in column 9, the

Attribute column. However, the Attribute column contains tag-value pairs with additional information all together in one string. The tag-value pairs are separated with '; ' (colon space). You need to parse the attributes string and collect gene_id, gene_name and gene_type associated information. Only genes with all three of these values defined should be saved. Only those three pieces of information need to be saved in a TSV file. The description of the GTF file formation is at: https://uswest.ensembl.org/info/website/upload/gff.html

6. For the gene IDs, do not include the version number. Remove the decimal point and digits to the right. For example, ENSG00000269732.1 would be saved as ENSG00000269732
7. Do not hardcode the URLs for the quant files into your script, rather read a list of URLs from a file on your computer. The file ps2_gene_quant_URLs.tsv is provided on canvas. In class we talked about two ways that can be used to specify the file that contains this list, these options include importing the 'os' or 'sys' modules. For the quant file parsing have your script loop over the list of URLs.
8. The filenames used for your output files should be explicit, this highlight their difference from other files.
9. Sample code has been provided for retrieving the data from a URL. Retrieve example_URL_code.ipynb from canvas. Text and gzipped data are handled slightly differently.
10. In preparation for Part 2 you can explore how to find more details on genes starting with an Ensembl gene ID.
11. Test your output data.
12. Add appropriate comments to your script as if you want to share this with a colleague. Include your name and date on a comment early in the file. Save your script with a filename that's informative, perhaps parse_quant_tsv.ipynb and parse_gencode_gtf.ipynb, not Homework2.ipynb. Also create a HTML version of these as you did for PS1.
13. Upload your two .ipynb and two .html files to canvas as you did last week.

A file with the URLs of TSV (ps2_gene_quant_URLs.tsv) and GTF (ps2_gencode_url.tsv) files is provided on canvas.

ENCFF016CBS	omental fat pad	https://www.encodeproject.org/files/ENCFF016CBS/@@download/ENCFF016CBS.tsv
ENCFF365ZMW	sigmoid colon	https://www.encodeproject.org/files/ENCFF365ZMW/@@download/ENCFF365ZMW.tsv
ENCFF408GFZ	subcutaneous adipose	https://www.encodeproject.org/files/ENCFF408GFZ/@@download/ENCFF408GFZ.tsv
ENCFF505TUS	prostate gland	https://www.encodeproject.org/files/ENCFF505TUS/@@download/ENCFF505TUS.tsv
ENCFF633OSJ	suprapubic skin	https://www.encodeproject.org/files/ENCFF633OSJ/@@download/ENCFF633OSJ.tsv
ENCFF862LZL	heart left ventricle	https://www.encodeproject.org/files/ENCFF862LZL/@@download/ENCFF862LZL.tsv
ENCFF863ERP	testis	https://www.encodeproject.org/files/ENCFF863ERP/@@download/ENCFF863ERP.tsv
ENCFF916ODF	vagina	https://www.encodeproject.org/files/ENCFF916ODF/@@download/ENCFF916ODF.tsv
ENCFF918KPC	stomach	https://www.encodeproject.org/files/ENCFF918KPC/@@download/ENCFF918KPC.tsv

Output example from parsed TSV files. Output data from each TSV file into a separate file with three columns. The columns are separated by just a TAB character. The columns are gene ID, TPM value, quant file accession number.

gene_id	TPM	file_accession
ENSG00000000003	3.53	ENCFF016CBS
ENSG00000000005	1.24	ENCFF016CBS
ENSG00000000419	2.62	ENCFF016CBS
ENSG00000000457	1.1	ENCFF016CBS
ENSG00000000460	0.76	ENCFF016CBS
ENSG00000000938	7.05	ENCFF016CBS
ENSG00000000971	4.15	ENCFF016CBS

The output example of parsing the GTF file. The output is a TSV with three columns, this information: Ensembl gene ID, HGNC gene name, gene feature type

gene_id	gene_name	gene_type
ENSG00000223972	DDX11L1	transcribed_unprocessed_pseudogene
ENSG00000227232	WASH7P	unprocessed_pseudogene
ENSG00000243485	RP11-34P13.3	lincRNA
ENSG00000237613	FAM138A	lincRNA
ENSG00000268020	OR4G4P	unprocessed_pseudogene
ENSG00000240361	OR4G11P	unprocessed_pseudogene
ENSG00000186092	OR4F5	protein_coding