Command Line for Data Filtering Exercises

Before you begin, copy "example.tsv" from "penelopeprime" to your home directory and then open Terminal. Alternatively, you can download the example file from:
https://funcgen2019.buschlab.org/downloads/command-line/example.tsv

1. Using the `awk` and `wc` commands (and a pipe), find out how many genes are significantly differentially expressed (i.e. adjusted p-value < 0.05).

2. Using the `cut` command, make a new file that just contains the Ensembl ID, the adjusted p-value, the $\log_2$ fold change and the gene name and description.

3. Search for all the genes whose name begins with "si:". How many are there?

4. How many genes have a biotype of "protein_coding"?

5. Using just the `awk` command, make a new file that contains the Ensembl ID, gene name, chromosome and strand (in that order) for all the genes on the reverse strand.

6. Use the `man` command to find out about the `more` command. What option do you need to use with `more` to see line numbers in the `example.tsv` file?

7. Use the `sort` command to order the file by chromosome. Does the order of the non-numeric "chromosomes" make sense? Try using the -V option of sort, instead of -g. Is the order now better? (The -V option is technically for sorting version numbers, but it's also really useful for sorting chromosome names!)

8. How many genes are between 10,000,000 bp and 20,000,000 bp on chromosome 1?