Command Line for Data Filtering Answers

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).

   ```
   awk '$3 < 0.05' example.tsv | wc -l
   2031
   ```

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.

   ```
   cut -f1,3,4,10,11 example.tsv > q2.tsv
   ```

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

   ```
   cut -f10 example.tsv | grep si: | wc -l
   2381
   ```

4. How many genes have a biotype of “protein_coding”?

   ```
   cut -f9 example.tsv | grep protein_coding | wc -l
   18693
   ```

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.

   ```
   awk -F"\t" '$8 == -1 { print $1 "\t" $10 "\t" $5
   "\t" $8 }' example.tsv > q5.txt
   ```

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?

It turns out the version of `more` on the training room computers doesn't have an option for this. Instead, you need to use `less`, which is an updated version of `more`.

```
less -N example.txt
```

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!)

```
sort -g -k5 example.tsv | more
sort -V -k5 example.tsv | more
```

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

```
awk '$5 == "2" && $7 > 10000000 && $6 < 20000000'
example.tsv | wc -l
107
```