PART IV

Additional Topics

Copy the file "practice_genes.txt" from /home/elavelle to your own linuxc directory.

Let's try a new means of file viewing:

```
less practice_genes.txt
```

It has many columns and entries of differet lengths, which make it difficult to view. We'll toss in the column command and a couple additional options:

```
column -t practice_genes.txt | less -S
```

That's better. But we probably don't need all these columns. How can we retain the names, log2 fold change, and adjusted p-value in tab-delimited format?

```
awk '{print $2 "\t" $4 "\t" $8}' practice_genes.txt > practice_genes_lite.tsv
```

Now, let's filter out genes which are missing data for.

```
awk '$3 != "NA" {print $0}' practice_genes_lite.tsv > practice_genes_lite_complete.tsv
```

It makes sense to order the file with the lowest p-value at the top.

```
sort -k3 practice_genes_lite_complete.tsv
```

This isn't quite what we want, however. It's sorting the values as character strings, not number values. Also, it dropped the header to the very end. Let's avoid this by saving the header line to another file.

```
head -1 practice_genes_lite_complete.tsv > genes_header.tsv
```

Now, we'll do a generalized sort on everything from the second line on:

```
tail -n +2 practice_genes_lite_complete.tsv | sort -g -k3 > \
practice_genes_lite_sorted.tsv
```

Next, we concatenate the new file back on to the header.

```
cat genes_header.tsv practice_genes_lite_sorted.tsv > \
practice_genes_lite_sorted_complete.tsv
```

Pretend we have a list of genes of interest and we want to cross reference it with those in our results. This will be easier if we can compare them directly with the contents of the first row of our file. To do so, we'll remove the quotes.

```
sed -i 's/"//g' practice_genes_lite_sorted_complete.tsv
```

Finally, let's actually do that! Take a guess at what the grep options are doing, then check the documentation with man or –help.

```
grep -wf /home/elavelle/cancer_genes.tsv practice_genes_lite_sorted_complete.tsv
```