Zebrafish Dataset Practical 2

All of these exercises will be done on the significantly differentially expressed genes from the 3 dpf comparison. Before you begin, copy "uninf_3dpf_hom_vs_sib.sig.tsv" (which just contains the significant genes, sorted by adjusted p-value) from "penelopeprime" to your home directory. Alternatively, you can download this file from:
https://funcgen2019.buschlab.org/downloads/practical2/uninf_3dpf_hom_vs_sib.sig.tsv

1. Use the `cut` command to get a list of just the Ensembl IDs for all the significantly differentially expressed genes in the 3 dpf comparison. Use this list with BioMart to get the human orthologues of these genes. Your output should include the zebrafish Ensembl ID, the zebrafish gene name, the human Ensembl ID, the human gene name, the human orthology type, the percentage identity (both target to query and query to target) and the human orthology confidence. How many of the zebrafish genes have a human orthologue? How many have a high confidence human orthologue?

   This problem was harder than we intended. Sorry! To get the IDs:

   ```
   cut -f1 uninf_3dpf_hom_vs_sib.sig.tsv | grep
   ENSDARG > uninf_3dpf_hom_vs_sib.sig.ids.tsv
   ```

   But getting just the genes with a human orthologue is quite hard. If the zebrafish Ensembl ID (ENSDARG) is in the 1st column and the human Ensembl ID (ENSG) is in the 3rd column then this will work:

   ```
   cut -f1,3 mart_export.txt | grep ENSG | cut -f1 |
   sort -u | wc -l
   ```

   The first `cut` just gets the IDs, the `grep` will remove any lines without a human ENSG ID (i.e. no human orthologue), the second `cut` gets the zebrafish ENSDARG IDs and then `sort -u` (where -u stands for "unique") will remove any duplicates.

   To get the number of high confidence human orthologues, you just need to additionally filter on the 6th column:

```
awk '$6 == 1' mart_export.txt | cut -f1,3 | grep
ENSG | cut -f1 | sort -u | wc -l
```

2. Use BioMart to get the sequence of the 1000 bp upstream of each transcript of the top 20 (extracted using `head`) most significantly differentially expressed genes. The header information should include the gene Ensembl ID, the gene name, the transcript Ensembl ID, the transcript type and the transcript length.

   ```
   head -20 uninf_3dpf_hom_vs_sib.sig.ids.tsv >
   uninf_3dpf_hom_vs_sib.sig.top20.ids.tsv
   ```

3. Use `awk` (not `cut`, which doesn't allow you to change the column order) to make a new file that contains the chromosome, start, end and Ensembl ID (in that order) for all the significant genes. Name the file something like "`uninf_3dpf_hom_vs_sib.sig.bed`". Congratulations - you've made a BED file. See https://genome.ucsc.edu/FAQ/FAQformat.html#format1 for more information about the BED format. Try viewing this file in Ensembl by adding it as a custom track. Have a look at the distribution of the genes in the "Whole genome", "Chromosome summary" and "Region in detail" views.

   ```
   awk -F"\t" '{ print $5 "\t" $6 "\t" $7 "\t" $1 }'
   uninf_3dpf_hom_vs_sib.sig.tsv | grep ENSDARG >
   uninf_3dpf_hom_vs_sib.sig.bed
   ```

4. In the "Region in detail" view, go to "Configure this page" and add the merged RNA-seq models, including the merged intron-spanning reads. Can you find any evidence for alternative splicing in any of the significant genes?

   http://www.ensembl.org/Help/Faq?id=472 has some useful info about viewing RNA-seq data in Ensembl.