Zebrafish Dataset Practical 2

Before you begin, you'll need the files from yesterday's practical in your home directory.

All of these exercises should be done on the 3 dpf comparison. You can either use the files you created yesterday or the appropriate columns of the full deseq2-results.tsv dataset. (The former approach is probably easier!)

1. Use the `cut` command to get a list of 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?

2. Use BioMart to get the sequence of the 1000 bp upstream of each transcript of the top 20 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.

3. Use `awk` 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.

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?