Name:

Comparative Genomics Fall 21

Orthology Lab

BLAST – refer to the instructions in the Spring21\_BLAST\_lab document.

1. What sequence file format are we dealing with in the downloaded mouse and zebrafish files this lab and how do you know?
2. What type of sequence data are in the mouse and zebrafish files and how do you know?
3. After you have run the first blast search, what is it about the standard output format that makes it difficult to parse and make sense of?
4. After you run the blast search to produce the tabular output, how many matches are there in total? Show your command and your answer.
5. After sorting the output by bitscore (column 12) and printing the first and last 25 lines, what is the relationship between column 12 and column 11?
6. How many hits are there with 65% or more protein identity? (hint: use wc -l on the output)
7. What is the relationship between evalue, bitscore, and percent identity?
8. Why do you think there are hits with 65% and above protein identity, but low bitscores and high evalues? (hint: check column 4 – and its header to see what it’s showing you)
9. How would you continue this analysis to make better predictions about orthology?

Orthologs and Paralogs with Ensembl

The photoreceptor cells in the retina of the human eye contain a number of different photoreceptors. The rod cells contain rhodopsin, which is responsible for monochromatic vision in the dark. The cone cells all contain one of three types of opsins, which respond to long-wave (red), middle-wave (green) and short-wave (blue) light, respectively, and are responsible for trichromatic color vision (see for instance <http://en.wikipedia.org/wiki/Opsin>).

Find the gene encoding the red-sensitive opsin.

Go to the Ensembl homepage (http://www.ensembl.org).

Select ‘Search: Human’ and type ‘long-wave-sensitive opsin gene’ in the ‘for’

text box.

Click [Go].

Click on ‘Gene’ on the page with search results.

Click on ‘Homo sapiens’.

Click on ‘OPN1LW (Human Gene) ENSG00000102076 X:154144243-154159032:1 Opsin 1, long wave sensitive [Source:HGNC Symbol;Acc:HGNC:9936]'. Note that the ‘LW’ in the gene symbol OPN1LW stands for ‘long-wave’.

Click on ‘Comparative Genomics - Paralogues’ in the side menu.

1. How many within-species paralogs have been identified for this gene?
2. Hover over the "Target %id" column header. What is the definition of "Target %id"?

Sort the table by "Target %id" from highest to lowest.

1. Which paralogs have the highest "Target %id" and "Query %id"?
2. Have a look at the genomic location of the red-, green- and blue-sensitive opsin genes. Does this explain why red-green color blindness is much more prevalent in males than in females (e.g. in the US population 7% vs 0.4%)?

Have a look at the gene tree for the red-sensitive opsin gene (OPN1LW).

First, click on ‘ENSG00000102076’ in case you are not already on the ‘Gene: OPN1LW’ tab.

Click on ‘Comparative Genomics - Gene Tree (image)’ in the side menu.

Click on ‘View paralogs of current gene’ below the gene tree image.

Click on the nodes (red squares) for the duplication events that have given rise to the various paralogs.

1. Which of its paralogs is due to the most recent duplication event? Give the gene name.
2. Is this reflected in the sequence similarity between the red-sensitive opsin and this paralog when compared with the other paralogs (see question 11)? Explain your answer.
3. At which taxonomic level did this duplication take place (kingdom, phylum, class, order, family, genus, species)? What is the name of this taxon? How long ago is it estimated to have occurred?

Click on the node that includes the common ancestor of OPN1LW, OPN1MW, and OPN1SW genes (you have to much further back in the phylogeny).

1. On which taxonomic level did this duplication take place (kingdom, phylum, class, order, family, genus, species)? What is the name of this taxon?