Assignment 9

How to complete this assignment

- Read through the background information
- Each part of the assignment begins with more background pertaining to that part and a set of tasks to complete. Complete all the tasks and then answer the associated questions. **Note that this information will not be available in the Quercus quiz.**
- As you complete the questions, mark your answer to each question.
- Questions will be either:
    - multiple-choice questions that require you to provide either a single answer or to select multiple answers.
    - questions that require a short text answer
- Open the associated assignment quiz on Quercus and enter your answers to each question.
- You may only submit this quiz once, so be sure you answer all questions before submitting the quiz.

Before you begin

- Open a new terminal session from your JupyterHub (New > Terminal)
- Set the PWD to `/home/jovyan/Week.9/Assignment.9`
- Install BLAST

Mark breakdown

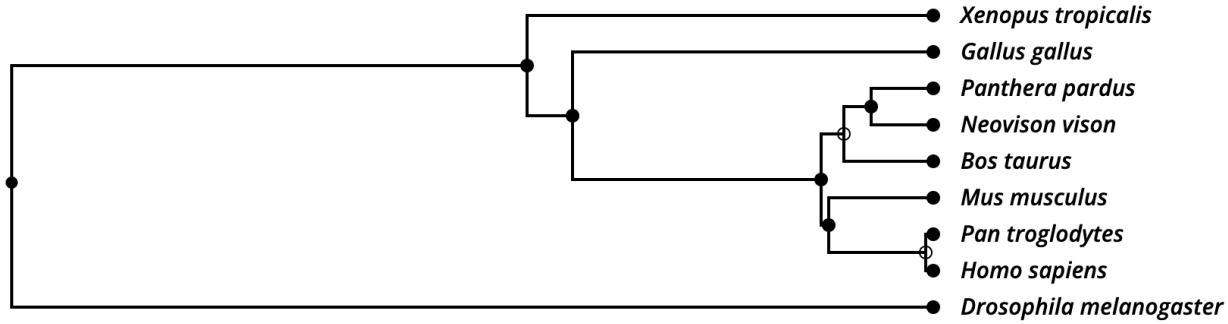
Part 1 – 4 questions – 6 marks
Part 2 – 4 questions – 4 marks
Part 3 – 6 questions – 10 marks
Bonus – 1 question – 1 mark

BACKGROUND

In this assignment you will be working with nucleotide and amino acid sequences from an undisclosed species, which will be referred to as "unknown_species". Questions will be testing your interpretation of the results of sequence alignment tasks.

Species (excluding the unknown species) you will work with in this assignment:

Species	Common Name	Divergence from Homo sapiens (millions of years ago)*
Homo sapiens	Human	N/A
Pan troglodytes	Chimpanzee	6.7
Mus musculus	Mouse	90
Neovison vison	Mink	94
Panthera pardus	Leopard	94
Bos taurus	Cow / Bovine	94
Gallus Gallus	Chicken	312
Xenopus tropicalis	Western clawed frog	352
Drosophila Melanogaster	Fruit fly	797



*Phylogenetic tree and divergence times are from http://www.timetree.org/

Data Sources:
Proteomes were downloaded from UniProt (https://www.uniprot.org/)
Protein and CDS sequences are downloaded from both Ensembl (https://ensembl.org/) and UniProt (https://www.uniprot.org/)

PART 1: ZNF330 (6 marks)

Set the PWD to: `/home/jovyan/Week.9/Assignment.9/part_1`

In this part we will examine the conservation of ZNF330*. Perform the following three alignments and then answer the questions below using the results:

- Align the CDS of ZNF330 in the unknown species (`unknown_species_ZNF330_CDS.fa`) to the CDSs of ZNF330 in *Homo sapiens*, *Pan troglodytes*, *Mus musculus*, *Gallus gallus*, *Xenopus tropicalis*, and *Drosophila melanogaster* (`ZNF330_orthologs_CDS.fa`).
- **Using a word size of 15**, align the CDS of ZNF330 in the unknown species (`unknown_species_ZNF330_CDS.fa`) to the CDSs of ZNF330 in *Homo sapiens*, *Pan troglodytes*, *Mus musculus*, *Gallus gallus*, *Xenopus tropicalis*, and *Drosophila melanogaster* (`ZNF330_orthologs_CDS.fa`).
- Align the amino acid sequence of ZNF330 in the unknown species (`unknown_species_ZNF330_protein.fa`) to the CDSs of ZNF330 in *Homo sapiens*, *Pan troglodytes*, *Mus musculus*, *Gallus gallus*, *Xenopus tropicalis*, and *Drosophila melanogaster* (`ZNF330_orthologs_protein.fa`).

*Note that the name of ZNF330 is Noa36 in *Drosophila melanogaster*.

Hint: perform the alignments with the regular output format and the tab delimited output format to help you answer all the questions.

Question 1

(1 mark)

What command did you use to align the amino acid sequence of ZNF330 in the unknown species to the CDSs of ZNF330 in *Homo sapiens*, *Pan troglodytes*, *Mus musculus*, *Gallus gallus*, *Xenopus tropicalis*, and *Drosophila melanogaster*. (Assuming you did not use the `-outfmt` option and did not redirect the standard output.)

Fill in the command below to match the command you used (do not include what is already there!).

```
_________ -subject ZNF330_orthologs_protein.fa
```

Question 2

(1 mark)

Based on the results of the alignments, the unknown species is most likely to be:

a. An invertebrate (like Drosophila melanogaster)
b. A vertebrate (like the other 5 species)

Question 3 (SELECT ALL THAT APPLY)
(2 marks)
Compare the results of the two CDS alignments (the alignment with the default word size and the alignment with the word size set to 15). Which of the following statements are true?
(**Note:** The number of marks for this question does not necessarily reflect the number of options that should be selected.)

a. Setting the word size to 15 increased the number of hits to 5.
b. The *Drosohphila melanogaster* CDS and the unknown sequence CDS have 15 consecutive matching bases.
c. Only *Homo sapiens*, *Pan troglodytes*, and *Mus musculus* CDSs have 28 consecutive bases that align perfectly with the unknown species CDS.
d. The alignment between the unknown species CDS and the *Mus musculus* CDS has fewer gaps than the alignment between the unknown species CDS and the *Gallus gallus* CDS.
e. The first 417 bases in the unknown species CDS do not align with any of the other CDSs.

Question 4 (SELECT ALL THAT APPLY)
(2 marks)
Examine the results of the amino acid sequence alignments. (Hint: Look at the full alignment of the unknown species protein and the *Homo sapiens* protein.) Which of the following statements are true?
(**Note:** The number of marks for this question does not necessarily reflect the number of options that should be selected.)

a. All 6 protein sequences in `ZNF330_orthologs_protein.fa` align with the unknown species protein sequence, but not all 6 CDS sequences in `ZNF330_orthologs_protein.fa` align with the unknown species CDS sequence in the alignments performed.
b. The percent identities of the amino acid sequence alignments are higher than the corresponding CDS alignments due to non-synonymous mutations.
c. *Mus musculus* is likely more closely related to the unknown species than *Gallus gallus*.
d. Based on **percent positive substitutions**, all 6 species amino acid sequence alignments have $> 80\%$ similarity.
e. When valine (V) is substituted with glycine (G) it is counted as a positive substitution.
f. When serine (S) is substituted with threonine (T) it is counted as a positive substitution.

PART 2: PPP2CB & KPTN (4 marks)

Set the PWD to: `/home/jovyan/Week.9/Assignment.9/part_2`

In this part we will examine two proteins: PPP2CB (protein phosphatase 2 catalytic subunit beta) and KPTN (kaptin, actin binding protein). Both proteins are present in our unknown species and an ortholog is present in the following species: *Homo sapiens*, *Pan troglodytes*, *Mus musculus*, *Panthera pardus*, *Neovison vison*, *Gallus gallus*, and *Xenopus tropicalis*. (Orthologs are also present in other species, but we will only use these 7 species.)

Perform the following alignments and then answer the questions below using the results:

- Align the unknown species PPP2CB amino acid sequence (`unknown_species_PPP2CB_protein.fa`) and the amino acid sequences of the 7 orthologs (`PPP2CB_orthologs_protein.fa`).
- Align the unknown species KPTN amino acid sequence (`unknown_species_KPTN_protein.fa`) and the amino acid sequences of the 7 orthologs (`KPTN_orthologs_protein.fa`).

Question 5

(1 mark)
Which gene is more likely to be essential?

Question 6

(1 mark)
The unknown species is more likely to be a mammal than an amphibian or bird.

a. True
b. False

Question 7

(1 mark)
The unknown species KPTN protein is more similar to the *Homo sapiens* protein than the *Panthera pardus* protein.

a. True
b. False

Question 8

(1 mark)
Based on the alignment, it is possible that the PPP2CB protein sequence is 100% conserved across mammals.

a. True
b. False

PART 3: Functional Analysis (5 marks)

Set the PWD to: `/home/jovyan/Week.9/Assignment.9/part_3`

In the `proteomes` directory there are two files. One contains the *Homo sapiens* proteome (protein sequences in `homo_sapiens_proteome.fa`), one contains the *Bos taurus* proteome (protein sequences in `bos_taurus_proteome.fa`). Make a BLAST database for each proteome. (Make sure to use the option to parse sequence IDs!)

A scientist has been studying the unknown species and identified a gene that is associated with a phenotype that is often seen in the species. The amino acid sequence of the protein product of the gene is in the file:
`unknown_species_unknown_protein_common_phenotype.fa`

The scientist has also identified a gene that is associated with a disease that occurs in the species. The amino acid sequence of the protein product of the gene is in the file:
`unknown_species_unknown_protein_disease.fa`

Perform the following analysis and use your results to answer the following questions:

- BLAST each of the unknown species protein sequences against the *Bos taurus* proteome.
- BLAST each of the unknown species protein sequences against the *Homo sapiens* proteome.
- Look up the protein IDs on UniProt (https://www.uniprot.org/).

Question 9
(2 marks)
The protein ID of the most likely ortholog of the **unknown protein involved in the common phenotype** in *Bos taurus* is (6 capital letters and numbers):

Question 10
(2 marks)
The protein ID of the most likely ortholog of **unknown protein involved in the common phenotype** in *Homo sapiens* is (6 capital letters and numbers):

Question 11
(1 mark)
Based on the function of the orthologs you found, which of the following phenotypes is it most likely the scientist was studying in relation to the **unknown protein involved in the common phenotype**?

a. Neurological behaviour differences
b. Limb abnormalities
c. Pigmentation differences
d. Size differences

Question 12
(2 marks)
The protein ID of the most likely ortholog of **unknown protein involved in disease** in *Bos taurus* is (6 capital letters and numbers):

Question 13
(2 marks)
The protein ID of the most likely ortholog of **unknown protein involved in disease** in *Homo sapiens* is (6 capital letters and numbers):

Question 14
(1 mark)
Based on the function of the orthologs you found, which of the following organs is most likely to be affected in the disease caused by **unknown protein involved in disease**?

a. Heart
b. Kidney
c. Spleen
d. Brain

Bonus Question
(1 mark)
Based on the results throughout the assignment, how many millions of years ago would you estimate that the unknown species diverged from *Homo sapiens*?

a. 312
b. 6.7
c. 94
d. > 500