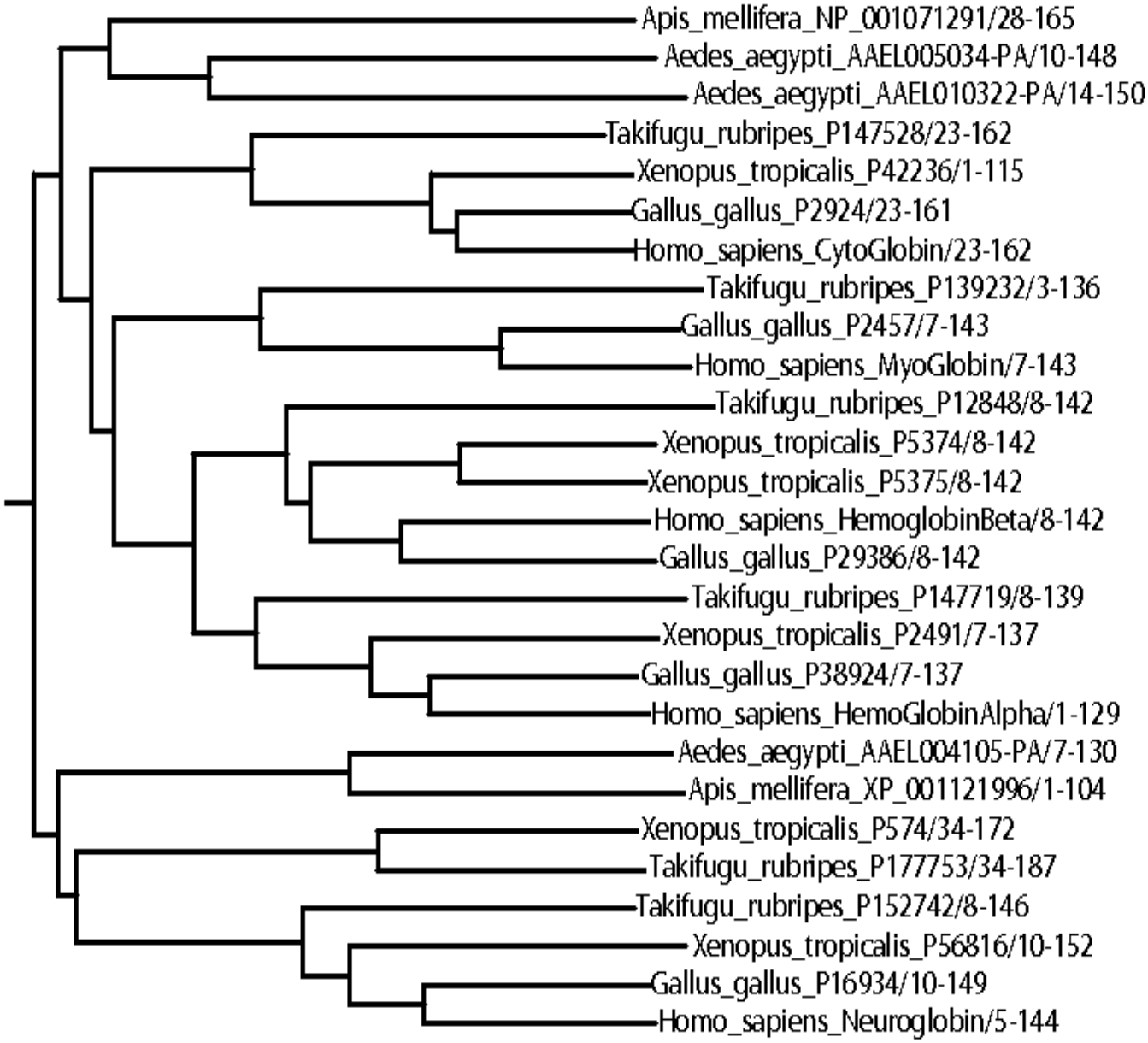
Question 4: globins, speciations & duplications



1. Look at the above tree of globin proteins. Annotate **all** internal nodes in terms of duplications and speciations using pen or pencil. You can use the species tree pictured at question one for reference of a species tree.

2. Mark / annotate in the tree the two places where it is likely that gene loss events occurred.

3. Use your now annotated globin tree. All internal nodes represent ancestral events. We want to map these events on a species tree. Draw a species tree (i.e. the relations between these species are given by the species tree in question 1) and on

this species tree mark (a) the number of globin genes in all ancestral and extant species (this is something you denote on the nodes), (b) the number of gene duplication (if any) on each **branch**, and (c) the number of gene losses (if any) on each **branch**.

Question 5: From BLAST to trees to duplications to orthology

1. Our next goal is to infer the evolutionary history of a human protein starting from its sequence. This evolutionary history should reveal the orthologs in other species and the timing of the duplicates of our protein. Go to http://blast.ncbi.nlm.nih.gov/Blast.cgi and select protein blast. As a query, we will use the sequence of human 6-phosphofructo-2-kinase / fructose-2,6-bisphosphatase given here http://www.uniprot.org/uniprot/Q16877.fasta . Change the target database to refseq_proteins. We want to find the relationships (i.e. duplications, orthology) across a diverse set of important model organisms amongst the different homologues genes / proteins of this enzyme. Restrict your search to a few species in the textbox next "**organism**" by typing the name and selecting when it appears. Add more textboxes by clicking on the plus sign next to the textbox. Select the following organisms *Homo sapiens*, *Drosophila melanogaste*r *Caenorhabditis elegans* (worm), *Schizosaccharomyces pombe* (fission yeast) and *Saccharomyces cerevisiae S288c* (budding yeast/ bakers yeast). Run blast. How many significant hits with a query coverage >80% are reported in the organisms that you selected?

2. Look at all the hits in human, do you think all of these are the result of duplication or what, if anything, could some also represent? (In other words what does isoform mean in refseq.) If necessary open and inspect entries of some of the hits; or inspect all human hits in the original blast output by the appropriate ctrl-F. Given that we want to find the relations between genes, explain why you want to take one protein sequence / isoform per gene / locus for making a tree later on. How many 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase genes do you think human and Drosophila have?

3. Obtain all the multiple paralogous (if any) protein sequences (one isoform per gene, see question 2 above) for all of the genes that are hit with >80% query coverage in human, *D. melanogaster*, and *C. elegans*,. For yeast (*S. cerevisiae*) and *S. pombe* only collect the best hit, since for this question we are not interested in dynamics within fungi, but we do want to use them as outgroup. The easiest way (out of many

possible) to achieve this to select those sequences you want in the blast results output page. After you have selected the sequences you want to retrieve press the "download " button at the top of the list of hits and select fasta complete sequences from the drop-down menu. Look at this file in any text editor like wordpad, word or notepad++ but NOT normal notepad/kladblok. If this somehow resulted in you having downloaded not a raw text file but an html file remove the html header and footer. Why are the identifiers (e.g. `gi|64762406|ref|NP_006203.2|`) as you obtain them from NCBI “not suitable” to make a tree for determining **gene duplications?** The identifier is everything after the “>” up until the first space. Take also into consideration that most alignment, tree making or tree display programs cut-off too long gene identifiers and in any case do not display nor parse the characters after the first space.

4. Adjust the identifiers for your sequences to something more easy to read: try to make the gene names somehow reflect function and species. Also always keep different names for different sequences and keep in mind that the programs stop reading the names of your sequences after the first space. (This is also necessary because for many sequences in NR or refseq a single sequence can represented the same gene in different closely related species resulting in complicated fasta description line. See also https://en.wikipedia.org/wiki/FASTA_format) **Save this file as a plain text file!** Go to clustal omega upload your sequences with by now reformatted names of the homologs. Align the sequences and make a tree using clustal omega. (if somehow clustal does not like your file, you can try copy and paste into the text box of clustal). After you have obtained the tree, go to iTOL, upload the tree and inspect it, reroot it if necessary on a biological logical internal branch. Explain why you rooted the tree the way you did.

5. Sketch the resulting tree. Annotate the tree in terms of duplications and speciations. How many duplications does this tree imply?

6. Check the function of the different human genes, and the reconstruction according to literature from the following article https://bmcbiol.biomedcentral.com/articles/10.1186/1741-7007-4-16 (our proteins are in the left most panel of figure 2). What type of functional differentiation have the genes undergone.

7. According to your tree, which human gene(s) are orthologs of which genes in *D. melanogaster* and to which genes in *C. elegans*?

8. Given that bioinformatics is also about making life easier (and not just more complicated) we are also going to look at a precomputed set of relations for our gene and its orthologs and paralogs. One of the best places to that is ENSEMBL (as discussed in the lectures). We are going to query ENSEMBL with one of the proteins that we used in our tree. Query human ENSEMBL (http://www.ensembl.org/Homo_sapiens/Info/Index) for PFKFB2 and go to its gene view (ENSG00000123836). Click on "gene tree (image)" in the left panel. Inspect the tree. You should not immediately see the paralogous. To see the paralogs click on "View paralogs of current gene" (below the gene tree picture). Find the duplication nodes that separate the different human genes (including the duplication in the super tree).

9. Try to find *D. melanogaster* and *C. elegans* in the ensembl tree. Is the branching of the genes of these two genes the same or different as the one you obtained in your gene tree?

10. In the Ensembl gene entry ENSG00000123836, click on "orthologues"" in the left panel. Compare the results to your own answers regarding orthology for *D. melanogaster*, yeast and *C. elegans*. How does Ensembl classify the different "types" of orthology?