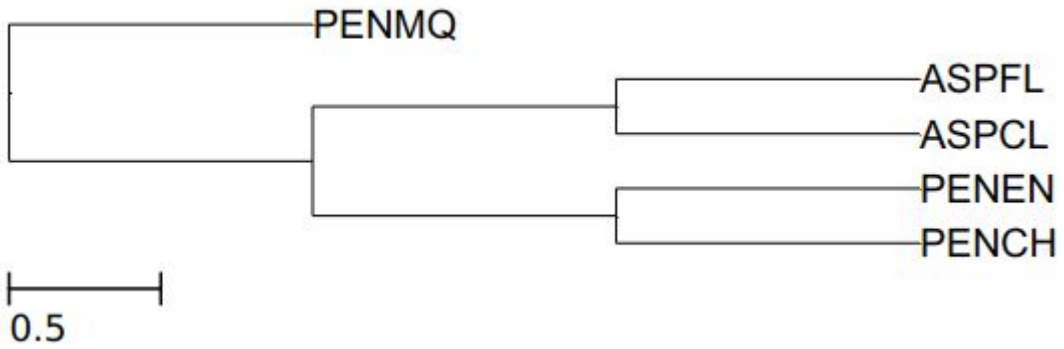
Session 5: Phylogenomics

B.- Phylogeny based orthology assignment

REMINDER: Gene tree reconstruction is divided in three steps: homology search, multiple sequence alignment and model selection plus tree reconstruction. There are many ways and many programs to perform each of the steps so it is often up to the person building the tree to decide how to build them. Here are some of the programs you can use to perform the different steps.
Homology search: blast, hmm, and similar programs as seen before
Alignment reconstruction: mafft, muscle, t-coffee
Tree reconstruction: phyml, raxml, fasttree, iqtree

In these exercises we are going to start with pre-build trees which you can find in the folder called exercise 4. To build this tree we have performed a blast search using protein ASPCL_0083_04882 as a starting point, then we have filtered the blast results to keep only those that have an evalue below 1e-05. The multiple sequence alignment was done with muscle and the resulting fasta alignment was converted to phylip format using readAl. Finally RAxML was used to build the phylogenetic tree using the PROTGAMMALG model and 100 rapid bootstrap repetitions.

Species tree:



4.- Obtain orthologs based on gene trees manually:

Open in a browser the following link: http://phylo.io/ Now insert the contents of the file ASPCL_0083_04882.tree.txt into the tree data part of the web and visualize the tree. Using this tree as example we are going to discuss how orthology inference is done using two kinds of algorithms: reconciliation and species overlap (see slides of session 5).

4.1.- Based on what we have discussed for the tree ASPCL_0083_04882.tree.txt, now do the same for PENEN_0144_10558.tree.txt. And fill in the following table of orthology and paralogy relationships when referring to PENEN_0144_10558:

	Orthologs	Paralogs
BRH		
InParanoid		
orthoMCL	ASPCL_0077_01917 ASPFL_0017_13218 PENCH_0037_09950 PENMQ_0029_09206	None
Tree based (Species overlap)		
Tree based (Reconciliation)		

Discuss the differences between the different predictions and why they happened. Which method do you think is the most reliable?

C.- Phylogenomics

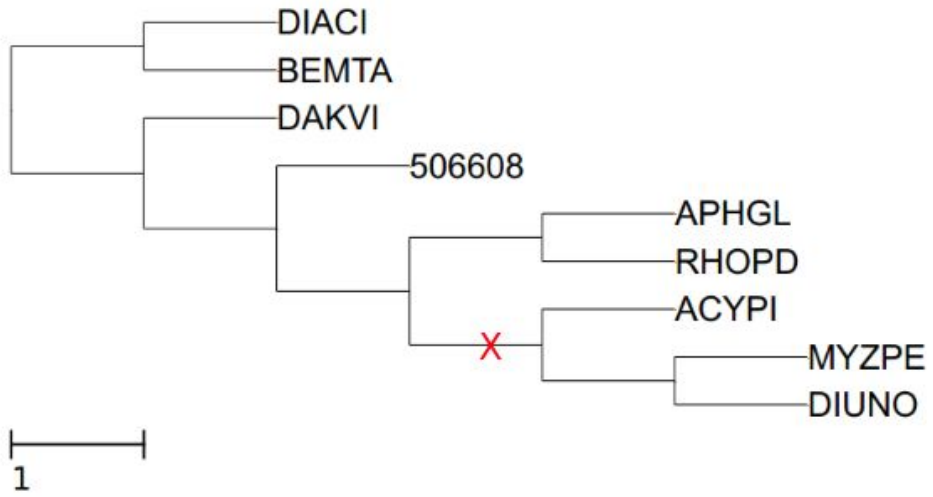
What we have done above for one single tree can be done for multiple trees at the same time. To do that we can use tools such as ape (http://ape-package.ird.fr/) or ete (http://etetoolkit.org/). Both are programming libraries that allow the user to work with trees, manipulate them and extract information. In the folder called “extra” in session 5 you can see an example of a script that uses ETE to calculate orthology relationships based on the species overlap algorithm.
While orthology prediction is often one of the objectives of working with collections of trees, there are other kind of analyses that can be done such as:

1.- Search for evolutionary events
2.- Test hypothesis based on topology
3.- Population genetics analyses

We are going to focus on the first two points.

5.1.- Go to the exercise5 folder. There you will find a file called tree_collection.txt. In there you will find a collection of 25 trees in the following format:
tree_number <tab> newick tree
In this exercise the species to which each sequence belongs is indicated at the end of the code (i.e. sequence Phy00BX4MK_ACYPI belongs to species ACYPI).

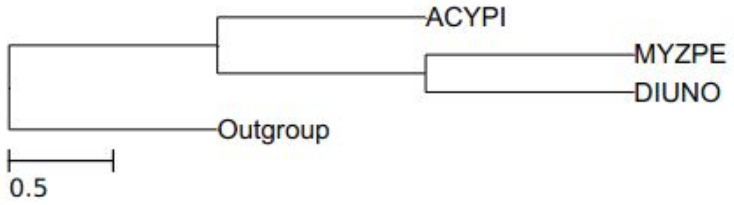
Given the species tree below:



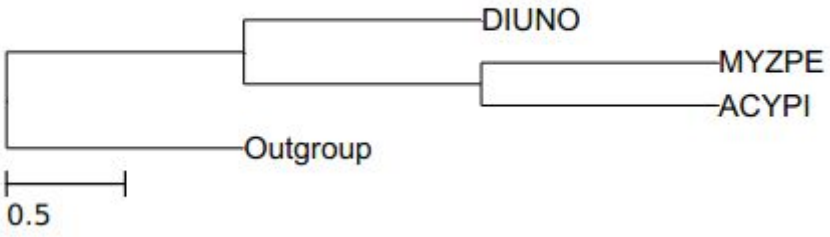
Visualize each tree in the tree_collection.txt file using phylo.io and match each tree to the conditions below. Each tree can be matched to multiple conditions. Trees may need to be rooted properly.

a.- Search for trees that have species specific duplications.
b.- If we assume that a duplication happened at the common ancestor of the species involved in the duplication, how many duplications happened at the node X? (use the species overlap algorithm to infer duplications)
c.- If we assume that species DIACI and BEMTA are very far related from species ACYPI, search for trees that could show a horizontal gene transfer event.
d.- Search for trees where RHOPD and ACYPI form a monophyletic clade.
e.- See the two topologies below, we are interested in knowing how many trees support each of the topologies shown. (Note: When doing this kind of analysis we never consider paralogy relationships, we only count orthologs. So if a tree has only orthologs between two of the species and the third is a paralog this tree is not considered. In addition notice that the tree needs to have an outgroup and there cannot be other species in between our species of interest).

TOPOLOGY 1:



TOPOLOGY 2:



f.- Search for trees that are identical to the species tree.
g.- Search for trees that are congruent with the species tree.
h.- One of the main applications for this kind of methodology is to build a species tree. When building species trees, we need groups of orthologous genes that have a one-to-one relationship in all the species of interest. Search among your trees which ones would be suitable for such analysis.
i.- Search for a tree that shows a protein family that was created de novo in ACYPI.

5.2.- Having done the analysis above, answer the following questions:

A.- Which is the percentage of gene trees that follow the species tree? Is this more or less trees than you were expecting?
B.- Given the results obtained in point e, which of the two topologies is the most represented in the trees? Is it the same we find in the species tree?
C.- How many of the trees that can be used to construct the species tree are actually congruent with it? Do you think this may affect the species tree reconstruction?
D.- There was one tree of a family that appeared de novo in ACYPI. How sure are we that the protein was really created in ACYPI? Are there any ways we could check it out?