#2 Phylogenetics to Phylogenomics

For this second exercise we use the BUSCO completeness results to build phylogenetic trees and a species tree

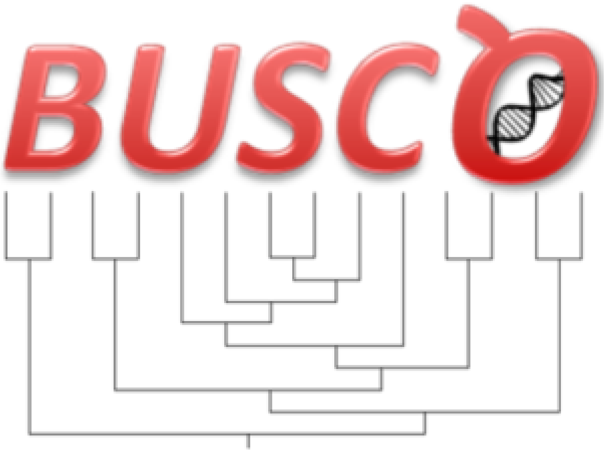
We will need to extract the sequence data from the BUSCO results, then align and trim the orthologous sequences, then concatenate them to perform the phylogenomics species tree reconstruction

By the end of this second exercise you should:

[1] have understood what BUSCO produces that we can use for building trees
[2] have produced some orthologue protein sequence alignments
[3] have understood why we need to trim these multiple sequence alignments
[4] have produced some gene trees from the alignments and visualised them
[5] have investigated the agreement or disagreement amongst individual gene trees
[6] have built a concatenated protein sequence alignment and used it to build the species phylogeny

NB: on the following pages, lines starting with a '*' are instructions or information, while lines starting with a '$' are commands to be typed into the terminal and executed

*Required



1. **My VM is up and running and I'm ready to proceed.** *

Mark only one oval.

○ Yes *Skip to question 2.*

○ No *Skip to "We're here to help!."*

[A] Extracting universal single-copy BUSCOs

[1] First we need to create a directory in which we will perform this exercise
* From your HOME directory in the terminal

```
$ mkdir rmw2
$ cd rmw2
```

[2] Then we need to download a simple extraction script
* If you are already skilled in programming then this sort of task is trivial, or if you are still learning or short on time (like now) then using such a script is the quick way forward
* From the Moodle site, find the folder under 'Day 2 Rob Waterhouse' called 'BUSCO_phylogenetics', inside you should see the Perl script file called 'extract_universal_single_copy_sequences.pl'

* Right click to get the full URL of the file (Copy Link Location) and then wget it to your VM
$ wget
https://edu.sib.swiss/pluginfile.php/6272/mod_folder/content/0/extract_universal_single_copy_sequences.pl
* NB: if the URL you copied ends with '?forcedownload=1' then delete this part

[3] Running the extraction script
* Before running 'extract_universal_single_copy_sequences.pl' open it to see more-or-less what it is going to try to do
* NB: the locations of the BUSCO results downloaded in exercise #1 are hard-coded in this script so if you did not follow the instructions exactly then you might need to edit this script so that it can find those results
* It expects to find the single_copy_busco_sequences.tar.gz tarballs in ~/rmw1/run_xxxxxx/ - where xxxxxx is the assembly code for each species you explored in exercise #1
* Running it is the easy part!
$ perl extract_universal_single_copy_sequences.pl

On the Moodle site (find this folder)

BUSCO_phylogenetics

2. **The script helpfully prints out the number of single-copy BUSCOs found in each assembly, and at the end how many were found in all assemblies - how many universal (all 10 species) single-copy BUSCOs were found?** *

Mark only one oval.

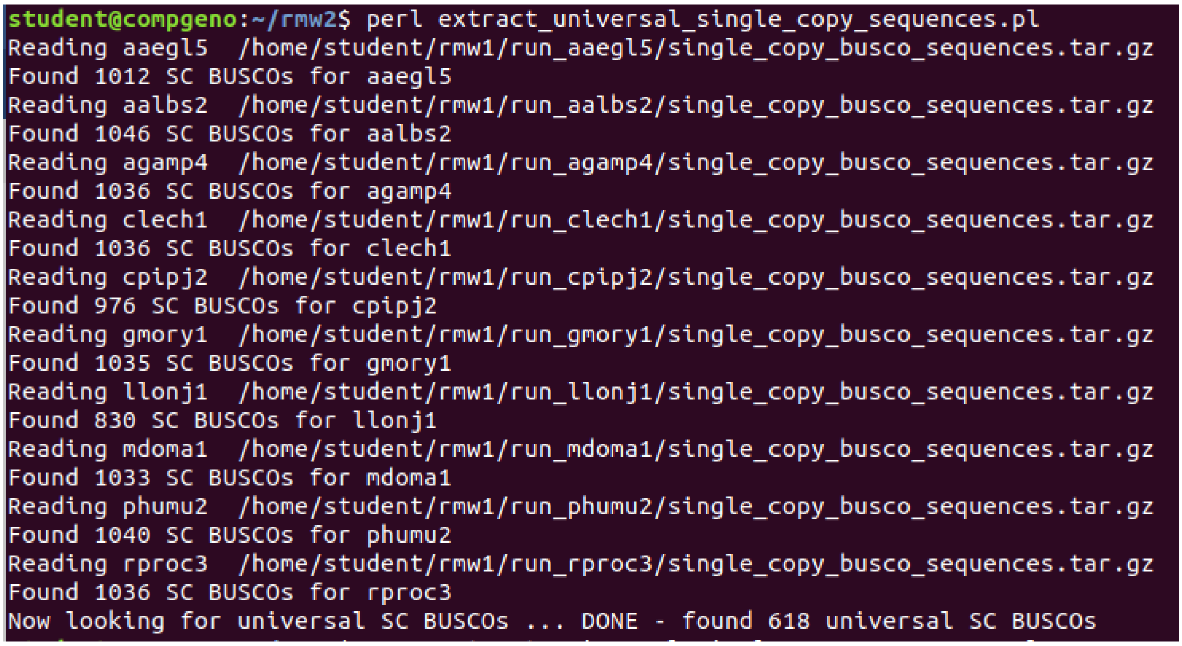
○ found 168 universal SC BUSCOs *Skip to question 3.*

○ found 618 universal SC BUSCOs *Skip to question 4.*

○ It is not working for me :-(*Skip to "We're here to help!."*

Are you sure?

Check the printed output of the script, it should look something like the following:



```
student@compgeno:~/rmw2$ perl extract_universal_single_copy_sequences.pl
Reading aaegl5  /home/student/rmw1/run_aaegl5/single_copy_busco_sequences.tar.gz
Found 1012 SC BUSCOs for aaegl5
Reading aalbs2  /home/student/rmw1/run_aalbs2/single_copy_busco_sequences.tar.gz
Found 1046 SC BUSCOs for aalbs2
Reading agamp4  /home/student/rmw1/run_agamp4/single_copy_busco_sequences.tar.gz
Found 1036 SC BUSCOs for agamp4
Reading clech1  /home/student/rmw1/run_clech1/single_copy_busco_sequences.tar.gz
Found 1036 SC BUSCOs for clech1
Reading cpipj2  /home/student/rmw1/run_cpipj2/single_copy_busco_sequences.tar.gz
Found 976 SC BUSCOs for cpipj2
Reading gmory1  /home/student/rmw1/run_gmory1/single_copy_busco_sequences.tar.gz
Found 1035 SC BUSCOs for gmory1
Reading llonj1  /home/student/rmw1/run_llonj1/single_copy_busco_sequences.tar.gz
Found 830 SC BUSCOs for llonj1
Reading mdoma1  /home/student/rmw1/run_mdoma1/single_copy_busco_sequences.tar.gz
Found 1033 SC BUSCOs for mdoma1
Reading phumu2  /home/student/rmw1/run_phumu2/single_copy_busco_sequences.tar.gz
Found 1040 SC BUSCOs for phumu2
Reading rproc3  /home/student/rmw1/run_rproc3/single_copy_busco_sequences.tar.gz
Found 1036 SC BUSCOs for rproc3
Now looking for universal SC BUSCOs ... DONE - found 618 universal SC BUSCOs
```

3. **How many universal (all 10 species) single-copy BUSCOs were found?** *

Mark only one oval.

○ found 168 universal SC BUSCOs *Skip to question 3.*

○ found 618 universal SC BUSCOs *Skip to question 4.*

○ It is not working for me :-(*Skip to "We're here to help!."*

[B] Aligning & trimming each set of BUSCOs

Now we have a 618 complete single-copy BUSCOs identified in all of our 10 assemblies

Do you understand why there are 618 rather than 830, which is the number of single-copy BUSCOs recovered from the assembly with the fewest BUSCOs (llonj1)?

=====
* If you were unable to make the extraction work you can download the pre-computed files
* From the Moodle site, find the folder under 'Day 2 Rob Waterhouse' called 'BUSCO_phylogenetics', inside you should see the tarball called 'busco_seqs.tar.gz'
* Right click to get the full URL of the file (Copy Link Location) and then wget it to your VM
$ wget https://edu.sib.swiss/pluginfile.php/6272/mod_folder/content/0/busco_seqs.tar.gz
* NB: if the URL you copied ends with '?forcedownload=1' then delete this part
* untarzip the downloaded tarball
$ tar -xf busco_seqs.tar.gz
=====

[1] First browse the busco_seqs folder and examine some of the files
* NB: these are not yet aligned, so they are in standard FASTA format, with different lengths and no gaps
$ more busco_seqs/EOG090X0TJE.fas

[2] Check that the alignment and trimming software are installed
* We will use MAFFT for multiple sequence alignments, you can find out more here: https://mafft.cbrc.jp/alignment/software/
* We will use TrimAl for filtering the multiple sequence alignments, you can find out more here: http://trimal.cgenomics.org/
$ mafft --version
* This should tell you that you're using: v7.407 (2018/Jul/23)
$ trimal --version
* This should tell you that you're using: trimAl 1.2rev59
* Checking the version of a software you want to use of course does not necessarily mean that it has been completely correctly installed, but at least by doing this you know it is installed, directly accessible, and responsive
* Also, when building comparative genomics workflows you will probably end up using lots of different tools, so it is a good idea to note the versions of each tool that you use because these can change and could mean slightly different results (or a broken workflow) if you or someone else tries to repeat the workflow in the future

[3] Then we need to download a simple script that will automate running MAFFT and TrimAl for each of the 618 BUSCO sets
* From the Moodle site, find the folder under 'Day 2 Rob Waterhouse' called 'BUSCO_phylogenetics', inside you should see the Perl script file called 'align_trim_universal_single_copy_sequences.pl'
* Right click to get the full URL of the file (Copy Link Location) and then wget it to your VM
$ wget
https://edu.sib.swiss/pluginfile.php/6272/mod_folder/content/0/align_trim_universal_single_copy_sequences.pl
* NB: if the URL you copied ends with '?forcedownload=1' then delete this part

[4] Running the align & trim script
* Before running 'align_trim_universal_single_copy_sequences.pl' open it to see more-or-less what it is going to try to do
* NB: the location of the extracted BUSCO sequences from the previous step is hard-coded in this script so if you did not follow the instructions exactly then you might need to edit this script so that it

can find these files
* It expects to find the busco_seqs directory in your current directory, i.e. ~/rmw2/busco_seqs/
* Running it is the easy part!
* As there are 618 alignments to run this can take a few minutes, so we shall redirect the output to a log file (>& aln-trm.log.txt) and we shall run the script in the background (&)

```
$ perl align_trim_universal_single_copy_sequences.pl >& aln-trm.log.txt &
```

[5] While running ...
* Go back to the MAFFT and TrimAl websites (URLs given above) to find out more about these tools
* In the align & trim script MAFFT was called as follows:

```
mafft --quiet busco_seqs/$busco > busco_alns/$out
```

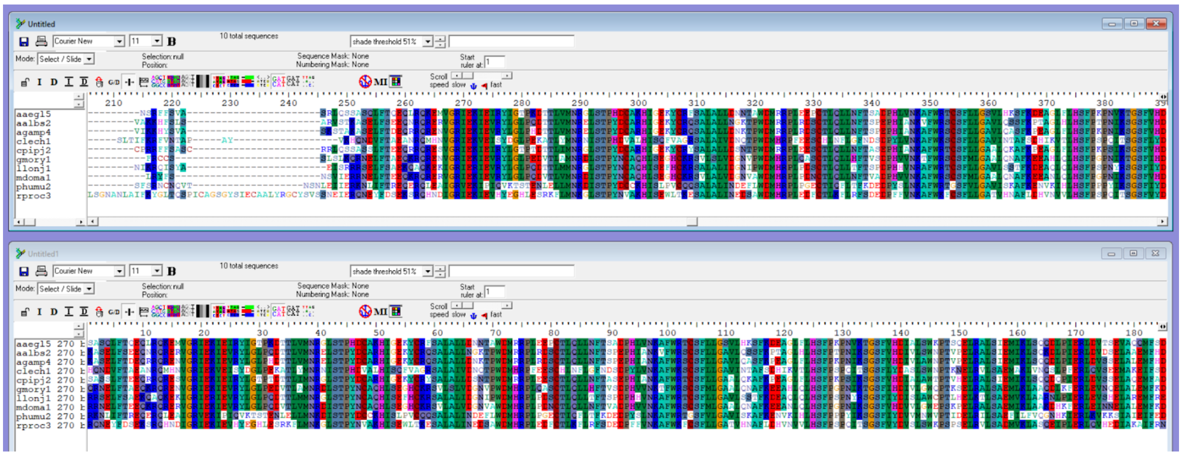
* The --quiet option is simply to stop MAFFT from printing progress messages
* In the align & trim script TrimAl was called as follows:

```
trimal -in busco_alns/$out -out busco_alns/$trm -strictplus
```

[6] Have a look at some completed alignments before and after trimming
* You can just browse the text files, or if you have your favourite alignment viewer to hand you can use that

```
$ more busco_alns/EOG090X0A3R.aln
$ more busco_alns/EOG090X0A3R.aln.trm
```

Before (top) and after (bottom) trimming an alignment



4. **Briefly, what does the '-strictplus' option used for TrimAl do? (hint, user guide can be found on TrimAl's website in the 'Publications' section)** *

[C] Gene trees with RAxML

=====
* If you were unable to make the align & trim steps work you can download the pre-computed files
* From the Moodle site, find the folder under 'Day 2 Rob Waterhouse' called 'BUSCO_phylogenetics', inside you should see the tarball called 'busco_alns.tar.gz'
* Right click to get the full URL of the file (Copy Link Location) and then wget it to your VM

```
$ wget https://edu.sib.swiss/pluginfile.php/6272/mod_folder/content/0/busco_alns.tar.gz
```

* NB: if the URL you copied ends with '?forcedownload=1' then delete this part
* untarzip the downloaded tarball

```
$ tar -xf busco_alns.tar.gz
```

=====

Just as there are many different tools for multiple sequence alignment (and several for trimming/filtering too), there are also many different methods and tools for building phylogenetic trees. Here we will use RAxML to build gene trees for each of the trimmed alignments we have already prepared.

RAxML => Randomized Axelerated Maximum Likelihood
Website: https://sco.h-its.org/exelixis/web/software/raxml/index.html
The userguide is 61 pages ... we will not go through everything now, but if you're going to be building trees as part of your research projects then you will probably have to read the userguides of several different tools

[1] Decide what kind of tree estimate we want to build
* first check RAxML installation

```
$ raxmlHPC-SSE3 -v
```

* you should see the version identifier and a list of people who have contributed to RAxML
* As well as the userguide you can use RAxML's help info

```
$ raxmlHPC-SSE3 -h
```

* You will see that the required inputs are listed first

```
-s sequenceFileName     - i.e. your input alignment
-n outputFileName       - i.e. a label you want to give the output
-m substitutionModel    - i.e. the sequence substitution model you want to use
```

* Input and output are easy to understand, if you are not familiar with sequence substitution models then you should probably read up on these at some point, you should not blindly be performing phylogenetic analyses without some understanding of the substitution models and where they come from, you can find a brief introduction on Wikipedia: https://en.wikipedia.org/wiki/Substitution_model
* The model we will use needs to be an amino acid (AA) model because we have aligned protein sequences
* From the userguide, available AA substitution models include: DAYHOFF, DCMUT, JTT, MTREV, WAG, RTREV, CPREV, VT, BLOSUM62, MTMAM, LG, MTART, MTZOA, PMB, HIVB, HIVW, JTTDCMUT, FLU, STMTREV, DUMMY, DUMMY2, AUTO, LG4M, LG4X, PROT_FILE, GTR_UNLINKED, GTR
* We will use the JTT model: Jones–Taylor–Thornton
* We will use a GAMMA model of rate heterogeneity
* So our call to RAxML will look something like the following:

```
raxmlHPC-SSE3 -s INPUTFILE -n OUTPUTLABEL -m PROTGAMMAJTT
```

* We will also need to specify a 'seed' as a seed is required for randomized stepwise addition order parsimony starting tree that is computed prior to the actual ML optimization

```
-p 12345
```

* And we will point to a directory (full path required) where we want all the results to go e.g.

```
 -w /home/student/rmw2/busco_tree
```

[2] Launch tree building for all your universal single-copy BUSCOs
* From the Moodle site, find the folder under 'Day 2 Rob Waterhouse' called 'BUSCO_phylogenetics', inside you should see the Perl script file called 'trees_universal_single_copy_sequences.pl'
* Right click to get the full URL of the file (Copy Link Location) and then wget it to your VM

```
$ wget
https://edu.sib.swiss/pluginfile.php/6272/mod_folder/content/0/trees_universal_single_copy_sequences.pl
```

* NB: if the URL you copied ends with '?forcedownload=1' then delete this part
* Before running 'trees_universal_single_copy_sequences.pl' open it to see more-or-less what it is going to try to do
* NB: the location of the aligned and trimmed BUSCO sequences from the previous step is hard-coded in this script so if you did not follow the instructions exactly then you might need to edit this script so that it can find these files
* It expects to find the busco_alns directory in your current directory, i.e. ~/rmw2/busco_alns/
* As there are 618 trees to build this will take quite some time, so we shall redirect the output to a log file (>& trees.log.txt) and we shall run the script in the background (&)

```
$ perl trees_universal_single_copy_sequences.pl >& trees.log.txt &
```

[3] While the trees are being built, try to build just one yourself
* First make a directory where you want the output to go, e.g.

```
$ mkdir mytree
```

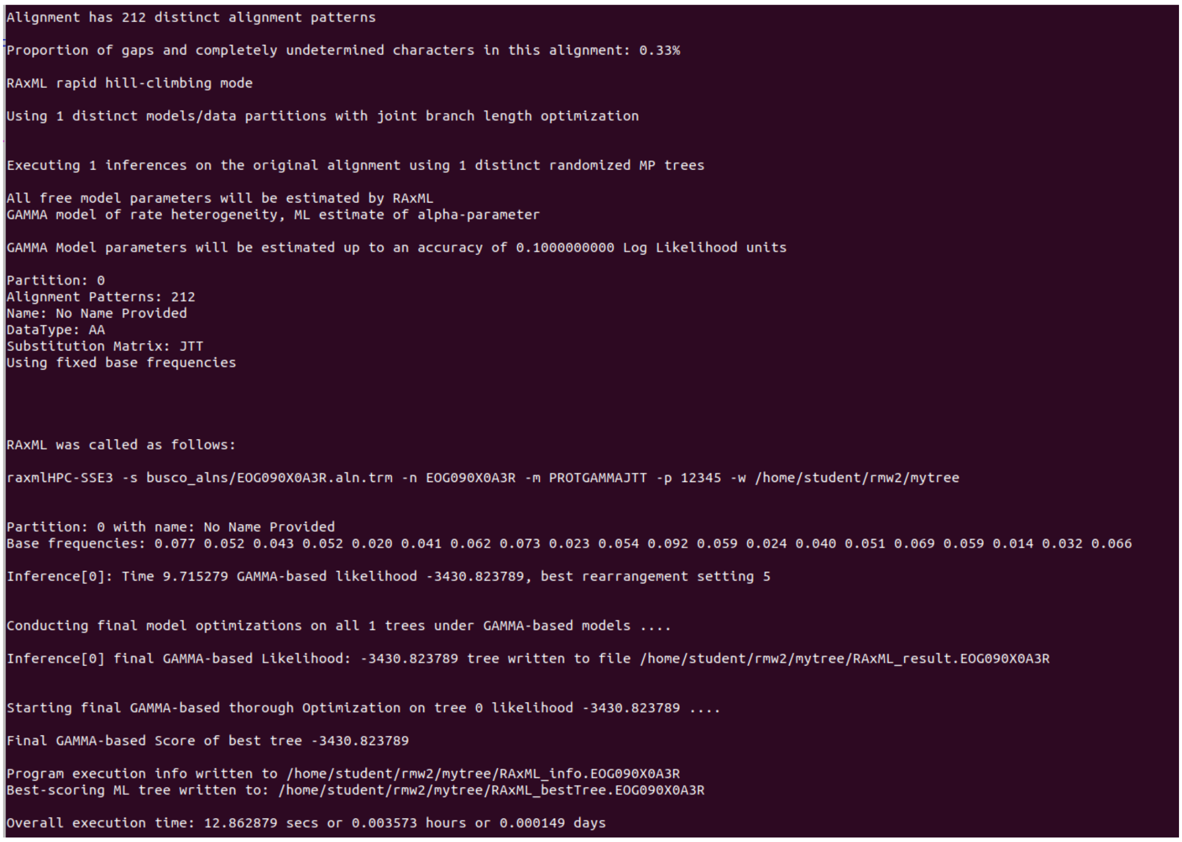
* Then select an aligned and trimmed BUSCO sequence set to use, and issue your command to RAxML, something like the following:

```
$ raxmlHPC-SSE3 -s busco_alns/EOG090X0A3R.aln.trm -n EOG090X0A3R -m PROTGAMMAJTT -p
12345 -w ~/rmw2/mytree
```

* When it is done have a look in your directory at the output files

```
$ ls -l mytree/
```

RAxML output from a single tree estimate



```
Alignment has 212 distinct alignment patterns

Proportion of gaps and completely undetermined characters in this alignment: 0.33%

RAxML rapid hill-climbing mode

Using 1 distinct models/data partitions with joint branch length optimization


Executing 1 inferences on the original alignment using 1 distinct randomized MP trees

All free model parameters will be estimated by RAxML
GAMMA model of rate heterogeneity, ML estimate of alpha-parameter

GAMMA Model parameters will be estimated up to an accuracy of 0.1000000000 Log Likelihood units

Partition: 0
Alignment Patterns: 212
Name: No Name Provided
DataType: AA
Substitution Matrix: JTT
Using fixed base frequencies



RAxML was called as follows:

raxmlHPC-SSE3 -s busco_alns/EOG090X0A3R.aln.trm -n EOG090X0A3R -m PROTGAMMAJTT -p 12345 -w /home/student/rmw2/mytree


Partition: 0 with name: No Name Provided
Base frequencies: 0.077 0.052 0.043 0.052 0.020 0.041 0.062 0.073 0.023 0.054 0.092 0.059 0.024 0.040 0.051 0.069 0.059 0.014 0.032 0.066

Inference[0]: Time 9.715279 GAMMA-based likelihood -3430.823789, best rearrangement setting 5


Conducting final model optimizations on all 1 trees under GAMMA-based models ....

Inference[0] final GAMMA-based Likelihood: -3430.823789 tree written to file /home/student/rmw2/mytree/RAxML_result.EOG090X0A3R


Starting final GAMMA-based thorough Optimization on tree 0 likelihood -3430.823789 ....

Final GAMMA-based Score of best tree -3430.823789

Program execution info written to /home/student/rmw2/mytree/RAxML_info.EOG090X0A3R
Best-scoring ML tree written to: /home/student/rmw2/mytree/RAxML_bestTree.EOG090X0A3R

Overall execution time: 12.862879 secs or 0.003573 hours or 0.000149 days
```

5. **Ready to continue?** *

Mark only one oval.

○ Yep, my trees are running and my single tree worked fine *Skip to question 6.*

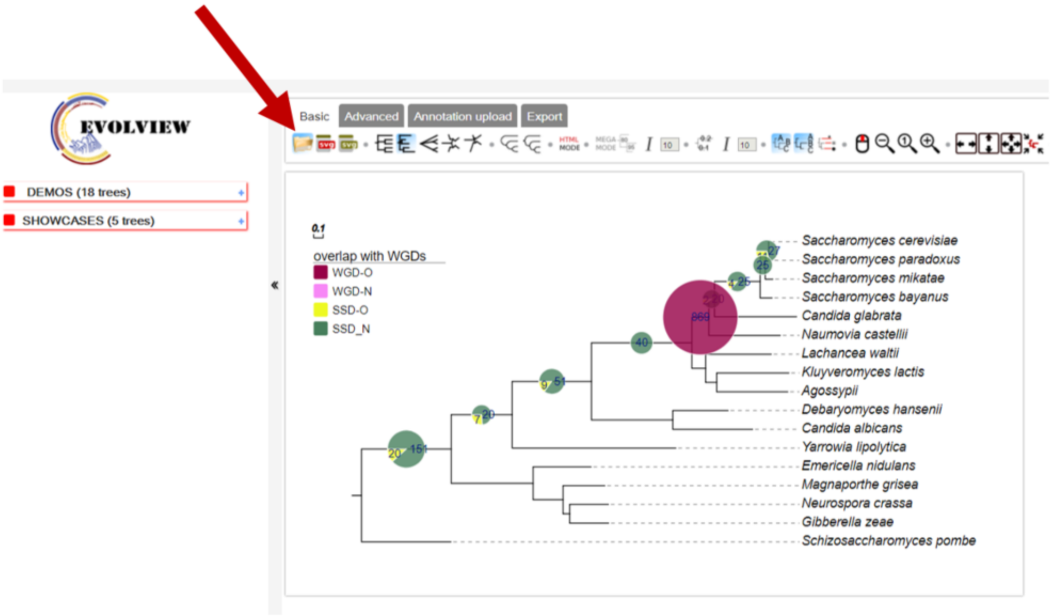
○ Nope, I need help! *Skip to "We're here to help!."*

[D] Tree visualisation

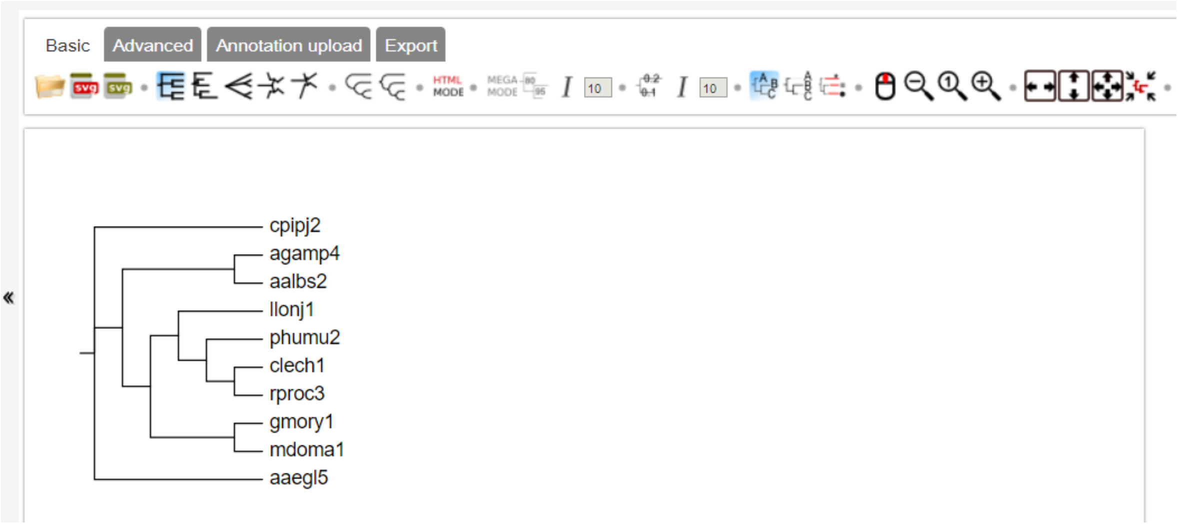
[1] Let's take a look at the single tree that you built (or one of the 618 trees that has finished computing)
* First have a look at the files produced by RAxML
$ ls -l mytree/
* There are several files produced for each tree estimate:
bestTree
info
log
parsimonyTree
result
* Feel free to explore these, but what we really want for now is the bestTree (i.e. the tree with the best maximum likelihood score)
* Copy the contents of the bestTree file onto your clipboard
* We will use an online tool to visualise it: http://www.evolgenius.info/evolview/
* Select 'Use without an account' (or log in if you have an account and want to log in)
* Select 'upload tree file' (see below for screenshot)
* Paste in your tree (newick format) in the 'Data:' box, give it a name e.g. mytree, and click the Submit button
* We're not going to explore all the different visualisation options today, but you can revisit EvolView later to learn about the various options etc.
* NB: if EvolView is not working (the website was down on Saturday) you can try with iTOL's Tree of Life viewer instead https://itol.embl.de/upload.cgi

EvolView upload tree file

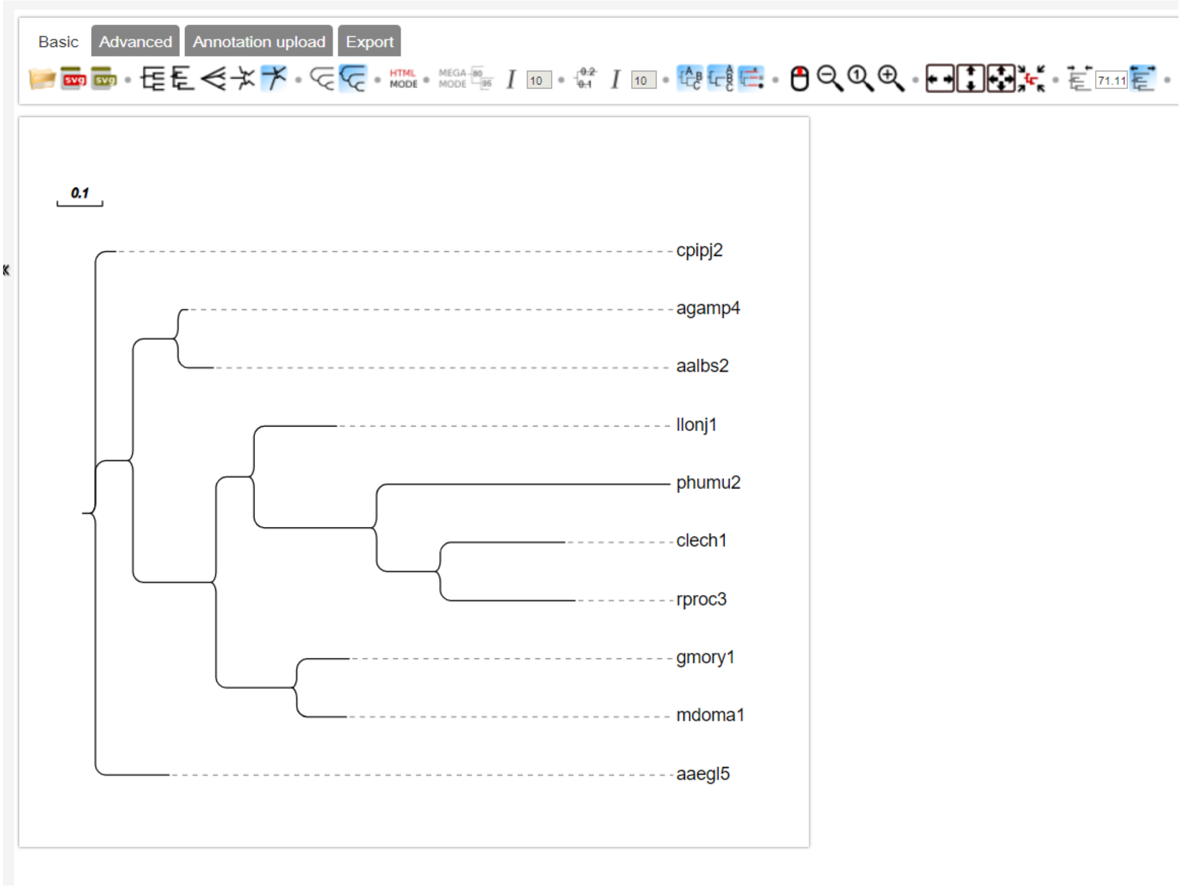
upload tree file



Uploaded tree



A different view after changing visualisation options



6. **Ready to continue?** *

Mark only one oval.

○ Yep, I managed to view my tree using EvolView *Skip to question 7.*

○ Nope, I need help! *Skip to "We're here to help!."*

[E] Tree comparisons

We would like to know how much agreement or disagreement there is amongst the 618 BUSCO trees that we have built - i.e. is there a clear winner for what the species tree should actually look like?

[1] First download the results for all 618 BUSCOs
* From the Moodle site, find the folder under 'Day 2 Rob Waterhouse' called 'BUSCO_phylogenetics', inside you should see the tarball called 'busco_trees_all.tar.gz'
* Right click to get the full URL of the file (Copy Link Location) and then wget it to your VM

```
$ wget https://edu.sib.swiss/pluginfile.php/6272/mod_folder/content/0/busco_trees_all.tar.gz
```

* NB: if the URL you copied ends with '?forcedownload=1' then delete this part
* untarzip the downloaded tarball

```
$ tar -xf busco_trees_all.tar.gz
```

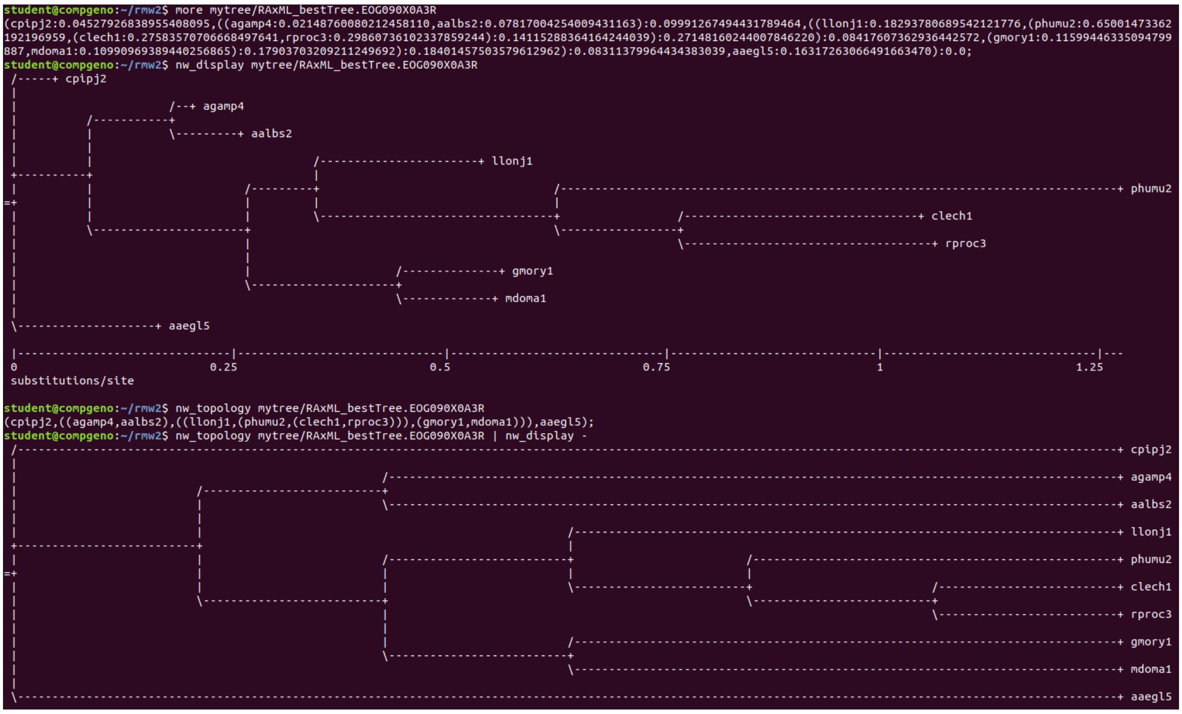
[2a] Compare the trees from each BUSCO
* Clearly uploading each tree to EvolView is not really an option!
* Instead we will use a command-line tool called newick utilities:
https://github.com/tjunier/newick_utils/wiki
* We aim to root and sort all the trees and then count the occurrences of each topology
* Why root and sort? We need to compare like-for-like trees, RAxML produces unrooted and unsorted trees, hence we must first make them comparable
* Why just topology? The branch lengths are clearly going to be very different for each tree as they were built from different genes (proteins), but here we want to explore the agreement or disagreement in the species phylogeny produced by each gene (and as they are universal single-copy orthologs in this case the gene trees will have the same leaves as the species tree), so we need just the topologies and not the branch lengths
* You can try it on your single tree first (if you chose a different BUSCO then make substitute with your identifier) ...
* First to view the newick file on the terminal:

```
$ more mytree/RAxML_bestTree.EOG090X0A3R
```

* Now to view it as a tree on the terminal:
$ nw_display mytree/RAxML_bestTree.EOG090X0A3R
* Now to strip out the branch lengths preserving only the topology
$ nw_topology mytree/RAxML_bestTree.EOG090X0A3R
* You can follow the topology command with the display command with a pipe '|'
$ nw_topology mytree/RAxML_bestTree.EOG090X0A3R | nw_display -
* From these species we know from the literature that Pediculus humanus is the undisputed outgroup species, so we can reroot the tree using phumu2 as the outgroup
$ nw_topology mytree/RAxML_bestTree.EOG090X0A3R | nw_reroot - phumu2
* Again we can pipe the commands to display it
$ nw_topology mytree/RAxML_bestTree.EOG090X0A3R | nw_reroot - phumu2 | nw_display -
* And finally we can sort alphabetically
$ nw_topology mytree/RAxML_bestTree.EOG090X0A3R | nw_reroot - phumu2 | nw_order -
* And visualise it
$ nw_topology mytree/RAxML_bestTree.EOG090X0A3R | nw_reroot - phumu2 | nw_order - | nw_display -

[2b] Compare the trees from each BUSCO
* Now let's look at all the trees rather than just our example
* First we will concatenate all the newick trees into a single file
$ cat busco_trees_all/RAxML_bestTree.* > bestTrees.txt
* Now get topologies, reroot, and sort all of these trees
$ nw_topology bestTrees.txt | nw_reroot - phumu2 | nw_order -
* Now count the number of different trees (here we can use basic unix commands 'sort' and 'uniq')
$ nw_topology bestTrees.txt | nw_reroot - phumu2 | nw_order - | sort | uniq -c | sort -nk1
* The first sort simply sorts the lines from the newick utilities output
* Then 'uniq -c' simply asks to count the number of times each unique line (in this case a tree) appears
* Then the final 'sort -nk1' sorts the output numerically (n) by the first column (k1)

Using newick utilities: nw_display, nw_topology

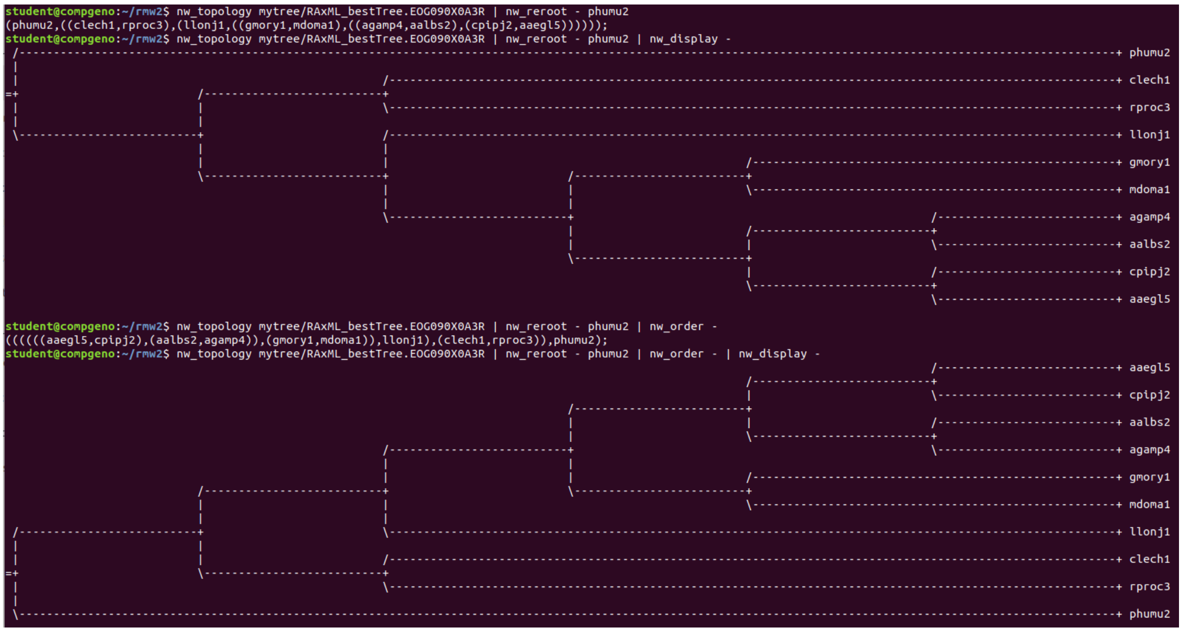


```
student@compgeno:~/rmw2$ more mytree/RAxML_bestTree.EOG090X0A3R
(cpipj2:0.04527926838955408095,((agamp4:0.02148760080212458110,aalbs2:0.07817004254009431163):0.09991267494431789464,((llonj1:0.18293780689542121776,(phumu2:0.65001473362
192196959,(clech1:0.27583570706668497641,rproc3:0.29860736102337859244):0.14115288364164244039):0.27148160244007846220):0.08417607362936442572,(gmory1:0.11599446335094799
887,mdoma1:0.10990969389440256865):0.17903703209211249692):0.18401457503579612962):0.08311379964434383039,aaegl5:0.16317263066491663470):0.0;
student@compgeno:~/rmw2$ nw_display mytree/RAxML_bestTree.EOG090X0A3R
 /-----+ cpipj2
 |
 |                      /--+ agamp4
 |          /-----------+
 |          |           \---------+ aalbs2
 |          |
 |          |                                /-----------------------+ llonj1
 +----------+                                |
 |          |                      /---------+                                  /---------------------------------------------------------------------------------+ phumu2
=+          |                      |         |                                  |
 |          |                      |         \----------------------------------+                 /----------------------------------+ clech1
 |          \----------------------+                                            \-----------------+
 |                                 |                                                              \------------------------------------+ rproc3
 |                                 |
 |                                 |                     /--------------+ gmory1
 |                                 \---------------------+
 |                                                       \-------------+ mdoma1
 |
 \--------------------+ aaegl5

 |-------------------------------|------------------------------|-------------------------------|------------------------------|-------------------------------|---
 0                            0.25                            0.5                            0.75                              1                            1.25
 substitutions/site

student@compgeno:~/rmw2$ nw_topology mytree/RAxML_bestTree.EOG090X0A3R
(cpipj2,((agamp4,aalbs2),((llonj1,(phumu2,(clech1,rproc3))),(gmory1,mdoma1))),aaegl5);
student@compgeno:~/rmw2$ nw_topology mytree/RAxML_bestTree.EOG090X0A3R | nw_display -
 /----------------------------------------------------------------------------------------------------------------------------------------------------------------+ cpipj2
 |
 |                                                     /----------------------------------------------------------------------------------------------------------+ agamp4
 |                          /--------------------------+
 |                          |                          \----------------------------------------------------------------------------------------------------------+ aalbs2
 |                          |
 |                          |                                                     /-------------------------------------------------------------------------------+ llonj1
 +--------------------------+                                                     |
 |                          |                          /--------------------------+                         /-----------------------------------------------------+ phumu2
=+                          |                          |                          |                         |
 |                          |                          |                          \-------------------------+                          /--------------------------+ clech1
 |                          \--------------------------+                                                    \--------------------------+
 |                                                     |                                                                               \--------------------------+ rproc3
 |                                                     |
 |                                                     |                          /-------------------------------------------------------------------------------+ gmory1
 |                                                     \--------------------------+
 |                                                                                \-------------------------------------------------------------------------------+ mdoma1
 |
 \----------------------------------------------------------------------------------------------------------------------------------------------------------------+ aaegl5
```

Using newick utilities: nw_reroot, nw_order



```
student@compgeno:~/rmw2$ nw_topology mytree/RAxML_bestTree.EOG090X0A3R | nw_reroot - phumu2
(phumu2,((clech1,rproc3),(llonj1,((gmory1,mdoma1),((agamp4,aalbs2),(cpipj2,aaegl5))))));
student@compgeno:~/rmw2$ nw_topology mytree/RAxML_bestTree.EOG090X0A3R | nw_reroot - phumu2 | nw_display -
 /----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+ phumu2
 |
 |                                                          /-----------------------------------------------------------------------------------------------------------------------+ clech1
=+                            /-----------------------------+
 |                            |                             \-----------------------------------------------------------------------------------------------------------------------+ rproc3
 |                            |
 \----------------------------+                             /-----------------------------------------------------------------------------------------------------------------------+ llonj1
                              |                             |
                              |                             |                                                           /-----------------------------------------------------------+ gmory1
                              \-----------------------------+                             /-----------------------------+
                                                            |                             |                             \-----------------------------------------------------------+ mdoma1
                                                            |                             |
                                                            \-----------------------------+                                                           /-----------------------------+ agamp4
                                                                                          |                             /-----------------------------+
                                                                                          |                             |                             \-----------------------------+ aalbs2
                                                                                          \-----------------------------+
                                                                                                                        |                             /-----------------------------+ cpipj2
                                                                                                                        \-----------------------------+
                                                                                                                                                      \-----------------------------+ aaegl5

student@compgeno:~/rmw2$ nw_topology mytree/RAxML_bestTree.EOG090X0A3R | nw_reroot - phumu2 | nw_order -
((((((aaegl5,cpipj2),(aalbs2,agamp4)),(gmory1,mdoma1)),llonj1),(clech1,rproc3)),phumu2);
student@compgeno:~/rmw2$ nw_topology mytree/RAxML_bestTree.EOG090X0A3R | nw_reroot - phumu2 | nw_order - | nw_display -
                                                                                                                                                      /-----------------------------+ aaegl5
                                                                                                                        /-----------------------------+
                                                                                                                        |                             \-----------------------------+ cpipj2
                                                                                          /-----------------------------+
                                                                                          |                             |                             /-----------------------------+ aalbs2
                                                                                          |                             \-----------------------------+
                                                            /-----------------------------+                                                           \-----------------------------+ agamp4
                                                            |                             |
                                                            |                             |                             /-----------------------------------------------------------+ gmory1
                              /-----------------------------+                             \-----------------------------+
                              |                             |                                                           \-----------------------------------------------------------+ mdoma1
                              |                             |
 /----------------------------+                             \-----------------------------------------------------------------------------------------------------------------------+ llonj1
 |                            |
 |                            |                             /-----------------------------------------------------------------------------------------------------------------------+ clech1
=+                            \-----------------------------+
 |                                                          \-----------------------------------------------------------------------------------------------------------------------+ rproc3
 |
 \----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+ phumu2
```

7. **How many times does the most frequent tree topology occur?** *

Mark only one oval.

○ 141 *Skip to question 8.*

○ 143 *Skip to question 9.*

○ 125 *Skip to question 8.*

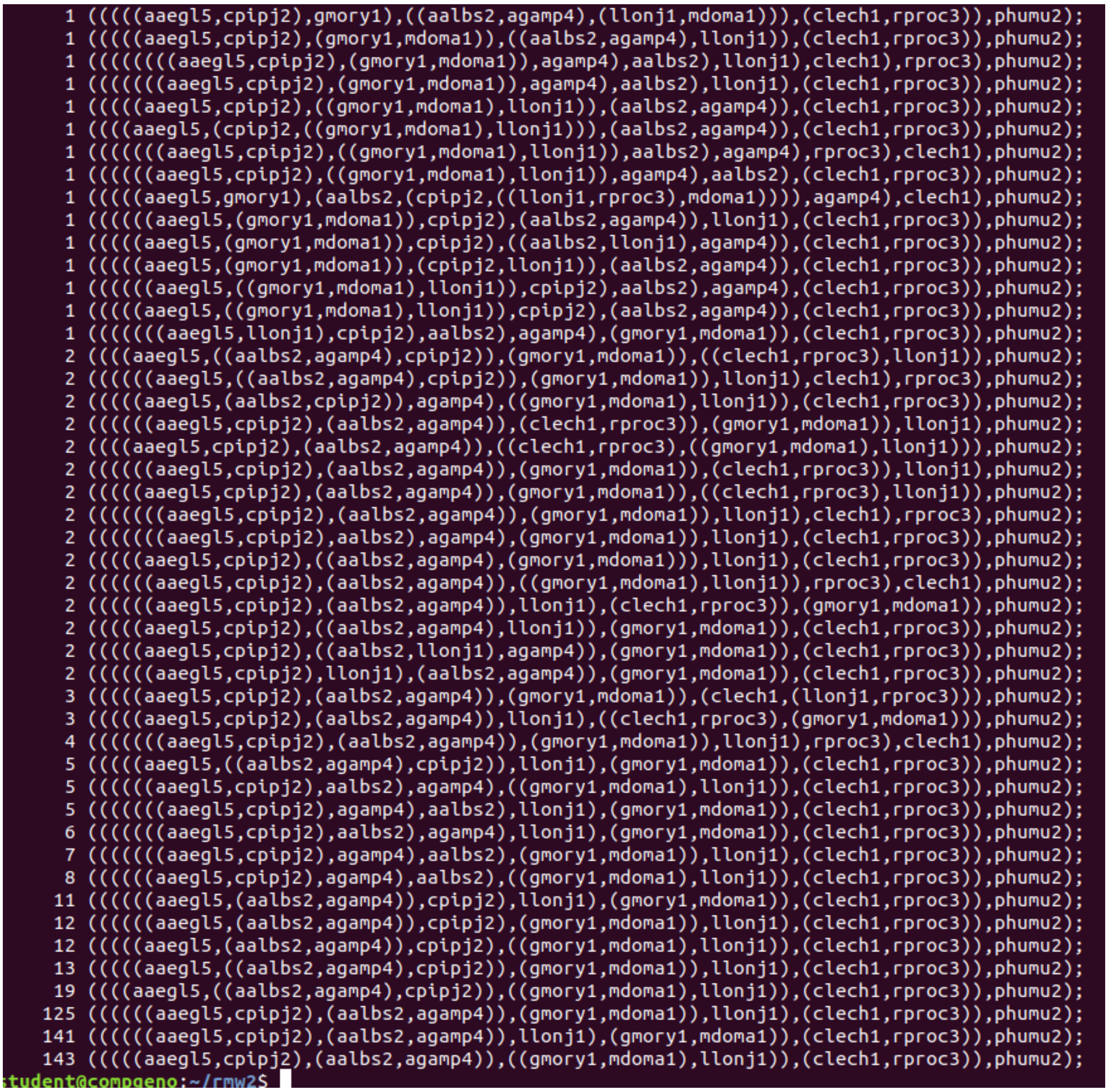
Are you sure?

Reminder:

* First we will concatenate all the newick trees into a single file
$ cat busco_trees_all/RAxML_bestTree.* > bestTrees.txt
* Now get topologies, reroot, and sort all of these trees
$ nw_topology bestTrees.txt | nw_reroot - phumu2 | nw_order -
* Now count the number of different trees (here we can use basic unix commands 'sort' and 'uniq')
$ nw_topology bestTrees.txt | nw_reroot - phumu2 | nw_order - | sort | uniq -c | sort -nk1
* The first sort simply sorts the lines from the newick utilities output
* Then 'uniq -c' simply asks to count the number of times each unique line (in this case a tree) appears
* Then the final 'sort -nk1' sorts the output numerically (n) by the first column (k1)

So the multi-component command should be something like:
$ nw_topology bestTrees.txt | nw_reroot - phumu2 | nw_order - | sort | uniq -c | sort -nk1

The result of ordering and counting the trees



```
      1 (((((aaegl5,cpipj2),gmory1),((aalbs2,agamp4),(llonj1,mdoma1))),(clech1,rproc3)),phumu2);
      1 (((((aaegl5,cpipj2),(gmory1,mdoma1)),((aalbs2,agamp4),llonj1)),(clech1,rproc3)),phumu2);
      1 ((((((((aaegl5,cpipj2),(gmory1,mdoma1)),agamp4),aalbs2),llonj1),clech1),rproc3),phumu2);
      1 (((((((aaegl5,cpipj2),(gmory1,mdoma1)),agamp4),aalbs2),llonj1),(clech1,rproc3)),phumu2);
      1 (((((aaegl5,cpipj2),((gmory1,mdoma1),llonj1)),(aalbs2,agamp4)),(clech1,rproc3)),phumu2);
      1 ((((aaegl5,(cpipj2,((gmory1,mdoma1),llonj1))),(aalbs2,agamp4)),(clech1,rproc3)),phumu2);
      1 (((((((aaegl5,cpipj2),((gmory1,mdoma1),llonj1)),aalbs2),agamp4),rproc3),clech1),phumu2);
      1 ((((((aaegl5,cpipj2),((gmory1,mdoma1),llonj1)),agamp4),aalbs2),(clech1,rproc3)),phumu2);
      1 (((((aaegl5,gmory1),(aalbs2,(cpipj2,((llonj1,rproc3),mdoma1)))),agamp4),clech1),phumu2);
      1 ((((((aaegl5,(gmory1,mdoma1)),cpipj2),(aalbs2,agamp4)),llonj1),(clech1,rproc3)),phumu2);
      1 (((((aaegl5,(gmory1,mdoma1)),cpipj2),((aalbs2,llonj1),agamp4)),(clech1,rproc3)),phumu2);
      1 (((((aaegl5,(gmory1,mdoma1)),(cpipj2,llonj1)),(aalbs2,agamp4)),(clech1,rproc3)),phumu2);
      1 ((((((aaegl5,((gmory1,mdoma1),llonj1)),cpipj2),aalbs2),agamp4),(clech1,rproc3)),phumu2);
      1 (((((aaegl5,((gmory1,mdoma1),llonj1)),cpipj2),(aalbs2,agamp4)),(clech1,rproc3)),phumu2);
      1 (((((((aaegl5,llonj1),cpipj2),aalbs2),agamp4),(gmory1,mdoma1)),(clech1,rproc3)),phumu2);
      2 ((((aaegl5,((aalbs2,agamp4),cpipj2)),(gmory1,mdoma1)),((clech1,rproc3),llonj1)),phumu2);
      2 ((((((aaegl5,((aalbs2,agamp4),cpipj2)),(gmory1,mdoma1)),llonj1),clech1),rproc3),phumu2);
      2 (((((aaegl5,(aalbs2,cpipj2)),agamp4),((gmory1,mdoma1),llonj1)),(clech1,rproc3)),phumu2);
      2 ((((((aaegl5,cpipj2),(aalbs2,agamp4)),(clech1,rproc3)),(gmory1,mdoma1)),llonj1),phumu2);
      2 ((((aaegl5,cpipj2),(aalbs2,agamp4)),((clech1,rproc3),((gmory1,mdoma1),llonj1))),phumu2);
      2 ((((((aaegl5,cpipj2),(aalbs2,agamp4)),(gmory1,mdoma1)),(clech1,rproc3)),llonj1),phumu2);
      2 (((((aaegl5,cpipj2),(aalbs2,agamp4)),(gmory1,mdoma1)),((clech1,rproc3),llonj1)),phumu2);
      2 (((((((aaegl5,cpipj2),(aalbs2,agamp4)),(gmory1,mdoma1)),llonj1),clech1),rproc3),phumu2);
      2 (((((((aaegl5,cpipj2),aalbs2),agamp4),(gmory1,mdoma1)),llonj1),(clech1,rproc3)),phumu2);
      2 (((((aaegl5,cpipj2),((aalbs2,agamp4),(gmory1,mdoma1))),llonj1),(clech1,rproc3)),phumu2);
      2 ((((((aaegl5,cpipj2),(aalbs2,agamp4)),((gmory1,mdoma1),llonj1)),rproc3),clech1),phumu2);
      2 ((((((aaegl5,cpipj2),(aalbs2,agamp4)),llonj1),(clech1,rproc3)),(gmory1,mdoma1)),phumu2);
      2 (((((aaegl5,cpipj2),((aalbs2,agamp4),llonj1)),(gmory1,mdoma1)),(clech1,rproc3)),phumu2);
      2 (((((aaegl5,cpipj2),((aalbs2,llonj1),agamp4)),(gmory1,mdoma1)),(clech1,rproc3)),phumu2);
      2 ((((((aaegl5,cpipj2),llonj1),(aalbs2,agamp4)),(gmory1,mdoma1)),(clech1,rproc3)),phumu2);
      3 (((((aaegl5,cpipj2),(aalbs2,agamp4)),(gmory1,mdoma1)),(clech1,(llonj1,rproc3))),phumu2);
      3 (((((aaegl5,cpipj2),(aalbs2,agamp4)),llonj1),((clech1,rproc3),(gmory1,mdoma1))),phumu2);
      4 (((((((aaegl5,cpipj2),(aalbs2,agamp4)),(gmory1,mdoma1)),llonj1),rproc3),clech1),phumu2);
      5 (((((aaegl5,((aalbs2,agamp4),cpipj2)),llonj1),(gmory1,mdoma1)),(clech1,rproc3)),phumu2);
      5 ((((((aaegl5,cpipj2),aalbs2),agamp4),((gmory1,mdoma1),llonj1)),(clech1,rproc3)),phumu2);
      5 (((((((aaegl5,cpipj2),agamp4),aalbs2),llonj1),(gmory1,mdoma1)),(clech1,rproc3)),phumu2);
      6 (((((((aaegl5,cpipj2),aalbs2),agamp4),llonj1),(gmory1,mdoma1)),(clech1,rproc3)),phumu2);
      7 (((((((aaegl5,cpipj2),agamp4),aalbs2),(gmory1,mdoma1)),llonj1),(clech1,rproc3)),phumu2);
      8 ((((((aaegl5,cpipj2),agamp4),aalbs2),((gmory1,mdoma1),llonj1)),(clech1,rproc3)),phumu2);
     11 ((((((aaegl5,(aalbs2,agamp4)),cpipj2),llonj1),(gmory1,mdoma1)),(clech1,rproc3)),phumu2);
     12 ((((((aaegl5,(aalbs2,agamp4)),cpipj2),(gmory1,mdoma1)),llonj1),(clech1,rproc3)),phumu2);
     12 (((((aaegl5,(aalbs2,agamp4)),cpipj2),((gmory1,mdoma1),llonj1)),(clech1,rproc3)),phumu2);
     13 (((((aaegl5,((aalbs2,agamp4),cpipj2)),(gmory1,mdoma1)),llonj1),(clech1,rproc3)),phumu2);
     19 ((((aaegl5,((aalbs2,agamp4),cpipj2)),((gmory1,mdoma1),llonj1)),(clech1,rproc3)),phumu2);
    125 ((((((aaegl5,cpipj2),(aalbs2,agamp4)),(gmory1,mdoma1)),llonj1),(clech1,rproc3)),phumu2);
    141 ((((((aaegl5,cpipj2),(aalbs2,agamp4)),llonj1),(gmory1,mdoma1)),(clech1,rproc3)),phumu2);
    143 (((((aaegl5,cpipj2),(aalbs2,agamp4)),((gmory1,mdoma1),llonj1)),(clech1,rproc3)),phumu2);
tudent@compgeno:~/rmw2$
```

8. **How many times does the most frequent tree topology occur?** *

Mark only one oval.

○ 141 *Skip to question 8.*

○ 143 *Skip to question 9.*

○ 125 *Skip to question 8.*

[F] Is there really a clear winner?

1st = 143 trees
2nd = 141 trees
3rd = 125 trees
After that the counts of different topologies are all very low, so which of these top 3 would you think is correct?

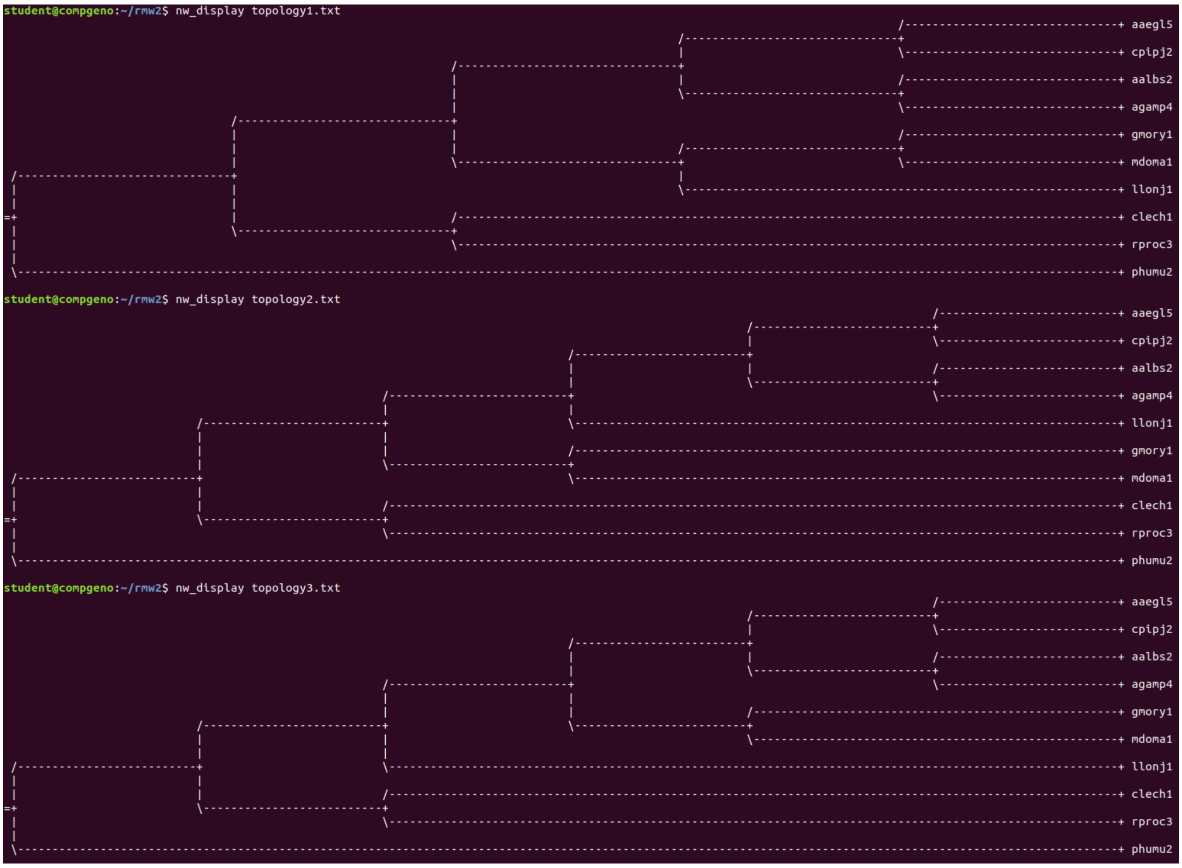
[1] Visualise the top 3 topologies
* First copy each topology into its own file, e.g.

```
$ echo '(((((aaegl5,cpipj2),(aalbs2,agamp4)),((gmory1,mdoma1),llonj1)),(clech1,rproc3)),phumu2);' > topology1.txt
$ echo '((((((aaegl5,cpipj2),(aalbs2,agamp4)),llonj1),(gmory1,mdoma1)),(clech1,rproc3)),phumu2);' > topology2.txt
$ echo '((((((aaegl5,cpipj2),(aalbs2,agamp4)),(gmory1,mdoma1)),llonj1),(clech1,rproc3)),phumu2);' > topology3.txt
```

* Then you can use nw_display again

```
$ nw_display topology1.txt
$ nw_display topology2.txt
$ nw_display topology3.txt
```

Top 3 tree topologies



```
student@compgeno:~/rmw2$ nw_display topology1.txt
                                                                                                                                   /-------------------------------+ aaegl5
                                                                                                  /--------------------------------+
                                                                                                  |                                \-------------------------------+ cpipj2
                                                                 /--------------------------------+
                                                                 |                                |                                /-------------------------------+ aalbs2
                                                                 |                                \--------------------------------+
                                                                 |                                                                 \-------------------------------+ agamp4
                                 /-------------------------------+
                                 |                               |                                                                 /-------------------------------+ gmory1
                                 |                               |                                /--------------------------------+
                                 |                               \--------------------------------+                                \-------------------------------+ mdoma1
 /-------------------------------+                                                                |
 |                               |                                                                \----------------------------------------------------------------+ llonj1
 |                               |
=+                               |                               /-------------------------------------------------------------------------------------------------+ clech1
 |                               \-------------------------------+
 |                                                               \-------------------------------------------------------------------------------------------------+ rproc3
 |
 \-----------------------------------------------------------------------------------------------------------------------------------------------------------------+ phumu2
student@compgeno:~/rmw2$ nw_display topology2.txt
                                                                                                                                        /--------------------------+ aaegl5
                                                                                                             /--------------------------+
                                                                                                             |                          \--------------------------+ cpipj2
                                                                                  /--------------------------+
                                                                                  |                          |                          /--------------------------+ aalbs2
                                                                                  |                          \--------------------------+
                                                       /--------------------------+                                                     \--------------------------+ agamp4
                                                       |                          |
                            /--------------------------+                          \--------------------------------------------------------------------------------+ llonj1
                            |                          |
                            |                          |                          /--------------------------------------------------------------------------------+ gmory1
                            |                          \--------------------------+
 /--------------------------+                                                     \--------------------------------------------------------------------------------+ mdoma1
 |                          |
 |                          |                          /-----------------------------------------------------------------------------------------------------------+ clech1
=+                          \--------------------------+
 |                                                     \-----------------------------------------------------------------------------------------------------------+ rproc3
 |
 \-----------------------------------------------------------------------------------------------------------------------------------------------------------------+ phumu2
student@compgeno:~/rmw2$ nw_display topology3.txt
                                                                                                                                        /--------------------------+ aaegl5
                                                                                                             /--------------------------+
                                                                                                             |                          \--------------------------+ cpipj2
                                                                                  /--------------------------+
                                                                                  |                          |                          /--------------------------+ aalbs2
                                                                                  |                          \--------------------------+
                                                       /--------------------------+                                                     \--------------------------+ agamp4
                                                       |                          |
                                                       |                          |                          /-----------------------------------------------------+ gmory1
                            /--------------------------+                          \--------------------------+
                            |                          |                                                     \-----------------------------------------------------+ mdoma1
                            |                          |
 /--------------------------+                          \-----------------------------------------------------------------------------------------------------------+ llonj1
 |                          |
 |                          |                          /-----------------------------------------------------------------------------------------------------------+ clech1
=+                          \--------------------------+
 |                                                     \-----------------------------------------------------------------------------------------------------------+ rproc3
 |
 \-----------------------------------------------------------------------------------------------------------------------------------------------------------------+ phumu2
```

9. **If you look carefully you will see that most relationships are stable, except for one species that appears in three different places - which species is misbehaving?** *

Mark only one oval.

○ aaegl5 Aedes aegypti mosquito *Skip to question 10.*

○ aalbs2 Anopheles albimanus mosquito *Skip to question 10.*

○ agamp4 Anopheles gambiae mosquito *Skip to question 10.*

○ clech1 Cimex lectularius bed bug *Skip to question 10.*

○ cpipj2 Culex quinquefasciatus mosquito *Skip to question 10.*

○ gmory1 Glossina morsitans tsetse fly *Skip to question 10.*

○ llonj1 Lutzomyia longipalpis sandfly *Skip to question 11.*

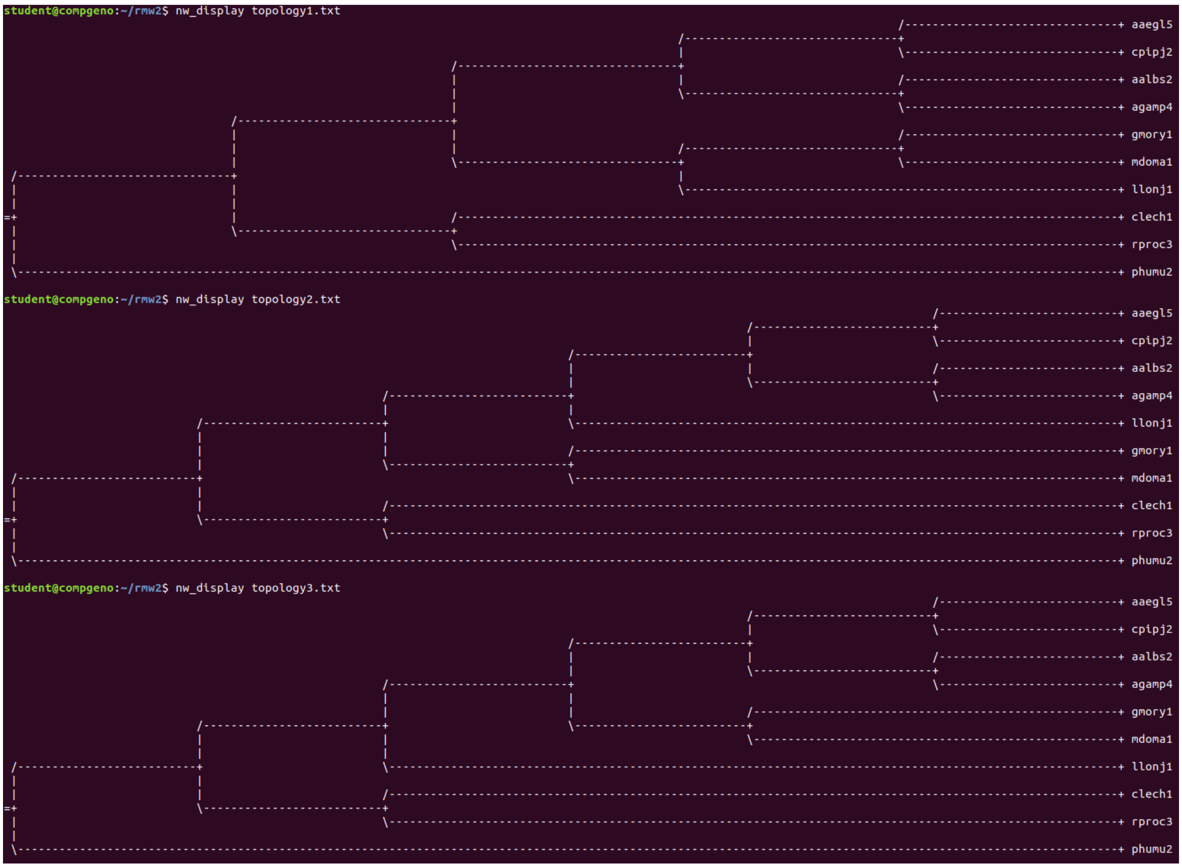
○ mdoma1 Musca domestica house fly *Skip to question 10.*

○ phumu2 Pediculus humanus body louse *Skip to question 10.*

○ rproc3 Rhodnius prolixus kissing bug *Skip to question 10.*

Are you sure?

Take a close look at the placement of llonj1



```
student@compgeno:~/rmw2$ nw_display topology1.txt
                                                                                                                                     /--------------------------------+ aaegl5
                                                                                                    /--------------------------------+
                                                                                                    |                                \--------------------------------+ cpipj2
                                                                   /--------------------------------+
                                                                   |                                |                                /--------------------------------+ aalbs2
                                                                   |                                \--------------------------------+
                                                                   |                                                                 \--------------------------------+ agamp4
                                  /--------------------------------+
                                  |                                |                                                                 /--------------------------------+ gmory1
                                  |                                |                                /--------------------------------+
                                  |                                \--------------------------------+                                \--------------------------------+ mdoma1
 /--------------------------------+                                                                 |
 |                                |                                                                 \-----------------------------------------------------------------+ llonj1
 |                                |
=+                                |                                /--------------------------------------------------------------------------------------------------+ clech1
 |                                \--------------------------------+
 |                                                                 \--------------------------------------------------------------------------------------------------+ rproc3
 |
 \--------------------------------------------------------------------------------------------------------------------------------------------------------------------+ phumu2
student@compgeno:~/rmw2$ nw_display topology2.txt
                                                                                                                                           /--------------------------+ aaegl5
                                                                                                               /---------------------------+
                                                                                                               |                           \--------------------------+ cpipj2
                                                                                    /--------------------------+
                                                                                    |                          |                           /--------------------------+ aalbs2
                                                                                    |                          \---------------------------+
                                                        /---------------------------+                                                      \--------------------------+ agamp4
                                                        |                           |
                             /--------------------------+                           \---------------------------------------------------------------------------------+ llonj1
                             |                          |
                             |                          |                           /---------------------------------------------------------------------------------+ gmory1
                             |                          \---------------------------+
 /---------------------------+                                                      \---------------------------------------------------------------------------------+ mdoma1
 |                           |
 |                           |                          /-------------------------------------------------------------------------------------------------------------+ clech1
=+                           \--------------------------+
 |                                                      \-------------------------------------------------------------------------------------------------------------+ rproc3
 |
 \--------------------------------------------------------------------------------------------------------------------------------------------------------------------+ phumu2
student@compgeno:~/rmw2$ nw_display topology3.txt
                                                                                                                                           /--------------------------+ aaegl5
                                                                                                               /---------------------------+
                                                                                                               |                           \--------------------------+ cpipj2
                                                                                    /--------------------------+
                                                                                    |                          |                           /--------------------------+ aalbs2
                                                                                    |                          \---------------------------+
                                                        /---------------------------+                                                      \--------------------------+ agamp4
                                                        |                           |
                                                        |                           |                          /------------------------------------------------------+ gmory1
                             /--------------------------+                           \--------------------------+
                             |                          |                                                      \------------------------------------------------------+ mdoma1
                             |                          |
 /---------------------------+                          \-------------------------------------------------------------------------------------------------------------+ llonj1
 |                           |
 |                           |                          /-------------------------------------------------------------------------------------------------------------+ clech1
=+                           \--------------------------+
 |                                                      \-------------------------------------------------------------------------------------------------------------+ rproc3
 |
 \--------------------------------------------------------------------------------------------------------------------------------------------------------------------+ phumu2
```

10. **Which species is misbehaving?** *

Mark only one oval.

○ aaegl5 Aedes aegypti mosquito *Skip to question 10.*

○ aalbs2 Anopheles albimanus mosquito *Skip to question 10.*

○ agamp4 Anopheles gambiae mosquito *Skip to question 10.*

○ clech1 Cimex lectularius bed bug *Skip to question 10.*

○ cpipj2 Culex quinquefasciatus mosquito *Skip to question 10.*

○ gmory1 Glossina morsitans tsetse fly *Skip to question 10.*

○ llonj1 Lutzomyia longipalpis sandfly *Skip to question 11.*

○ mdoma1 Musca domestica house fly *Skip to question 10.*

○ phumu2 Pediculus humanus body louse *Skip to question 10.*

○ rproc3 Rhodnius prolixus kissing bug *Skip to question 10.*

[G] Yep, that sandfly once again!

So now let us try to use all the alignment data together in a phylogenomic analysis to estimate the species tree, rather than just looking for the most frequent topology amongst all the individual gene trees

[1] Concatenate the trimmed alignments
* * From the Moodle site, find the folder under 'Day 2 Rob Waterhouse' called 'BUSCO_phylogenetics', inside you should see the Perl script file called 'concatenate_universal_single_copy_sequences.pl'
* Right click to get the full URL of the file (Copy Link Location) and then wget it to your VM
$ wget https://edu.sib.swiss/pluginfile.php/6272/mod_folder/content/0/concatenate_universal_single_copy_sequences.pl
* NB: if the URL you copied ends with '?forcedownload=1' then delete this part
* NB: the location of the aligned and trimmed sequences is hard-coded in this script so if you did not follow the instructions exactly then you might need to edit this script so that it can find these files
* It expects to find the busco_alns directory in your current directory, i.e. ~/rmw2/busco_alns/
$ perl concatenate_universal_single_copy_sequences.pl

* This should produce the file: concatenated_buscos.aln
* If this does not work for you, then you can fetch the already concatenated sequences from the Moodle site, under 'Day 2 Rob Waterhouse' in 'BUSCO_phylogenetics', inside you should see the gzipped file called 'concatenated_buscos.aln.gz'
* Right click to get the full URL of the file (Copy Link Location) and then wget it to your VM

```
$ wget https://edu.sib.swiss/pluginfile.php/6272/mod_folder/content/0/concatenated_buscos.aln.gz
```

* NB: if the URL you copied ends with '?forcedownload=1' then delete this part
* Then unzip it

```
$ gunzip concatenated_buscos.aln.gz
```

[2] Filter the concatenated alignment

* To keep only the best-aligned regions we will filter the alignment with TrimAl again

```
$ trimal -in concatenated_buscos.aln -out concatenated_buscos.aln.trm -strictplus
```

* To be even more strict we will remove all positions that contain any gaps

```
$ trimal -in concatenated_buscos.aln.trm -out concatenated_buscos.aln.trm.nogaps -nogaps
```

* See how the file size decreases after filtering, list the files to see their sizes:

```
$ ls -lh concatenated*
```

[3] Now build the species phylogeny with RAxML

* This time we will perform bootstrap sampling to estimate the support values for the branching of the tree
* If you are not familiar with bootstrapping then once you have started running RAxML you can take a quick look at this paper: http://www.pnas.org/content/93/23/13429
* We will use the '-f a' option to perform a rapid bootstrap analysis and search for bestscoring ML tree in one program run
* We need to specify how many bootstrap samples to perform, using the '-N' option
* To carry out an ML search after a rapid BS inference you must specify a random number seed with '-x'
* We also need to specify a random number seed with the '-p' option for the randomized stepwise addition
* And like previously, we need to specify which amino acid substitution model to use with the '-m' option

```
$ raxmlHPC-SSE3 -s concatenated_buscos.aln.trm.nogaps -n speciestree -f a -N 3 -x 12345 -p 12345 -m PROTGAMMAJTT >& raxml.log.txt &
```

* Remember that you can check that is is running using the 'ps -uf' command
* Remember also that the '>& raxml.log.txt &' part sends the output to a log file and runs the process in the background
* This can take quite some time (even with only 3 bootstrap samples), so while yours is still running you can fetch the pre-computed results that were computed with 50 bootstrap samples
* From the Moodle site, under 'Day 2 Rob Waterhouse' in 'BUSCO_phylogenetics', inside you should see the tarball file called 'speciestree100.tar.gz'
* Right click to get the full URL of the file (Copy Link Location) and then wget it to your VM

```
$ wget https://edu.sib.swiss/pluginfile.php/6272/mod_folder/content/0/speciestree50.tar.gz
```

* NB: if the URL you copied ends with '?forcedownload=1' then delete this part
* Then untarzip it

```
$ tar -xf speciestree50.tar.gz
```

Lutzomyia longipalpis (photo Dr Ray Wilson)

11. **Ready to continue?** *

Mark only one oval.

○ Yep, my tree is running *Skip to "[H] The molecular phylogenomic species tree."*

○ Nope, I need help! *Skip to "We're here to help!."*

[H] The molecular phylogenomic species tree

[1] The RAxML results
* After untarzipping the pre-computed speciestree50.tar.gz you will see that because this time we ran several (50) bootstrap samples the output from RAxML is a bit different and includes files labelled 'bipartitions'
* These simply map the bootstrap support values onto the newick formatted tree - either on the nodes (RAxML_bipartitions...) or the branches (RAxML_bipartitionsBranchLabels...) because different tree visualisation tools can treat the formats of these labels differently
* You can copy the tree and view it in EvolView
* Or you can use newick utilities to view it on the terminal e.g.

```
$ nw_reroot RAxML_bipartitions.speciestree50 phumu2 | nw_order - | nw_display -
```

* Notice which node does NOT have 100% bootstrap support
* You might also like to see the agreement / disagreement in the bootstrap trees (the RAxML_bootstrap... file)

```
$ nw_reroot RAxML_bootstrap.speciestree50 phumu2 | nw_order - | sort | uniq -c | sort -nk1
```

* How many different topologies are there now? Why could this be?
* Is there now a much clearer winner?

[2] Making an ultrametric tree from the molecular phylogeny
* An ultrametric tree is a tree where all the path-lengths from the root to the tips are equal
* Note from your molecular phylogeny the branch lengths are in substitutions per site, so different rates of substitution in different lineages will result in different path-lengths from the root to the tips
* Instead, we would like to 'convert' this molecular phylogeny, using some basic assumptions, into a 'time-tree' where the branch lengths instead represent the times between each speciation and then of course all the root-to-tip lengths become the same (all species have the same amount of time to go back to their last common ancestor)
* For this we will use the chronos function of the ape package in R
https://www.rdocumentation.org/packages/ape/versions/5.1/topics/chronos
* First let's reroot and order our molecular phylogeny

```
$ nw_reroot RAxML_bestTree.speciestree50 phumu2 | nw_order - >
RAxML_bestTree.speciestree50RO
```

* Now in R we can convert the tree to an ultrametric tree setting the root age to 366 million years

(which is the estimated divergence time of Pediculus humanus from the rest of the species)
* NB depending on if you made your own rooted+ordered (RO) tree or if you downloaded it there may or may not be '.txt' and then of the 'RAxML_bestTree.speciestree50RO' tree file, so edit or copy&paste accordingly

```
====== R code ======
library(ape)
library(phytools)
moltree<-read.tree("RAxML_bestTree.speciestree50RO")
calib<-makeChronosCalib(moltree,node="root",age.min=366)
timtree<-chronos(moltree,calibration=calib,lambda=1,model="discrete")
is.ultrametric(timtree)
write.tree(timtree)
timtree2<-roundBranches(timtree,digits=0)
is.ultrametric(timtree2)
write.tree(timtree2)
==================
```

* You can now view your 'timetree2' using EvolView or any other viewer, e.g.

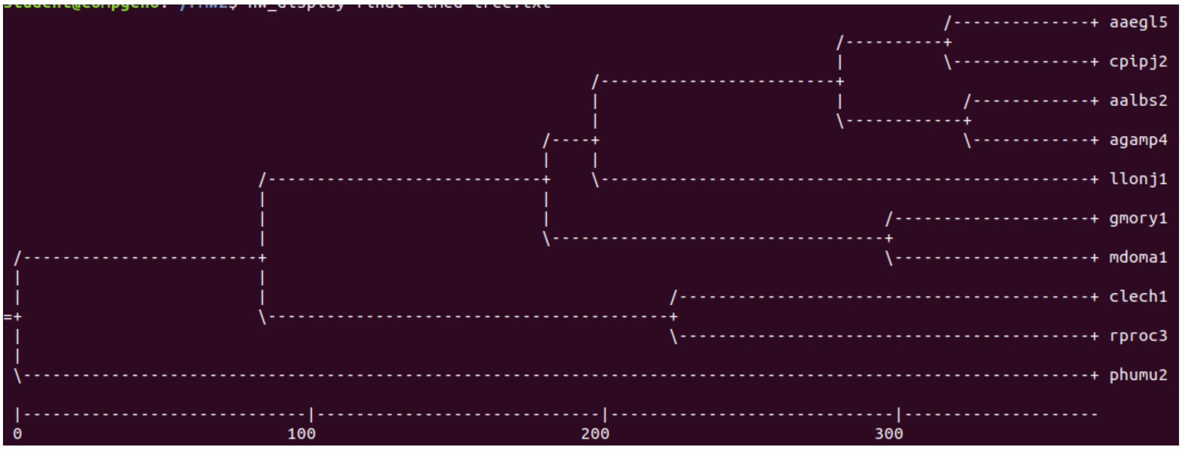
```
$ echo '((((((aaegl5:49,cpipj2:49):37,(aalbs2:43,agamp4:43):43):83,llonj1:169):16,(gmory1:71,mdoma1:71):114):98,(clech1:143,rproc3:143):140):83,phumu2:366);' > final-timed-tree.txt
$ nw_display final-timed-tree.txt
```

Hopefully by completing this exercise you:

[1] have understood what BUSCO produces that we can use for building trees
[2] have produced some orthologue protein sequence alignments
[3] have understood why we need to trim these multiple sequence alignments
[4] have produced some gene trees from the alignments and visualised them
[5] have investigated the agreement or disagreement amongst individual gene trees
[6] have built a concatenated protein sequence alignment and used it to build the species phylogeny

Click SUBMIT below to finish this exercise.

Final ultrametric tree



```
student@compgen0:~/TMP2$ nw_display final-timed-tree.txt
                                                                                                /--------------+ aaegl5
                                                                                     /----------+
                                                                                     |          \--------------+ cpipj2
                                                            /------------------------+
                                                            |                        |            /------------+ aalbs2
                                                            |                        \------------+
                                                       /----+                                     \------------+ agamp4
                                                       |    |
                          /----------------------------+    \--------------------------------------------------+ llonj1
                          |                            |
                          |                            |                                  /--------------------+ gmory1
                          |                            \----------------------------------+
 /------------------------+                                                               \--------------------+ mdoma1
 |                        |
 |                        |                                         /------------------------------------------+ clech1
=+                        \-----------------------------------------+
 |                                                                  \------------------------------------------+ rproc3
 |
 \-------------------------------------------------------------------------------------------------------------+ phumu2

 |-----------------------------|-----------------------------|-----------------------------|-------------------
 0                           100                           200                           300
```

Stop filling out this form.

We're here to help!

If you're stuck, please raise your hand and hopefully one of us will be able to help you.

Hit BACK below to return to the practical.