Workshop on Genomics, Český Krumlov, Jan 2024

Rosa Fernández, Institute of Evolutionary Biology (CSIC-UPF), Spain

Phylogenomics Hands-on session - Evening Session: 19:00-22:00

rosa.fernandez@ibe.upf-csic.es, Lab webpage: www.metazomics.com, Twitter: @Rosamygale (personal), @metazomics (lab)

Contents

Software and Resources for this Practical

These links are provided for information only, they are not needed now for the practical but if you finish any exercises early you may want to investigate the tools/resources a little further by checking out their websites, user guides, and/or their main publications.

OrthoFinder - orthology inference
Github repository here; Publication here; Tutorials here

MUSCLE - multiple sequence alignment tool
Website here; Publication here; Userguide here

cogent - a toolkit for statistical analysis of biological sequences in python
Website here

catsequences - a tool to concatenate orthogroups to create supermatrices
Github repository here

IQ-TREE2 - inference of maximum likelihood phylogenies
Website here; Publication here; Userguide here

iTOL - online phylogenetic tree viewer
Website here; Publication here; Userguide here

genesortR - sorting and subsampling of phylogenomic datasets
Github repository here; Publication here

ASTRAL - species tree inference based on the multispecies coalescent
Github repository here; Publication here

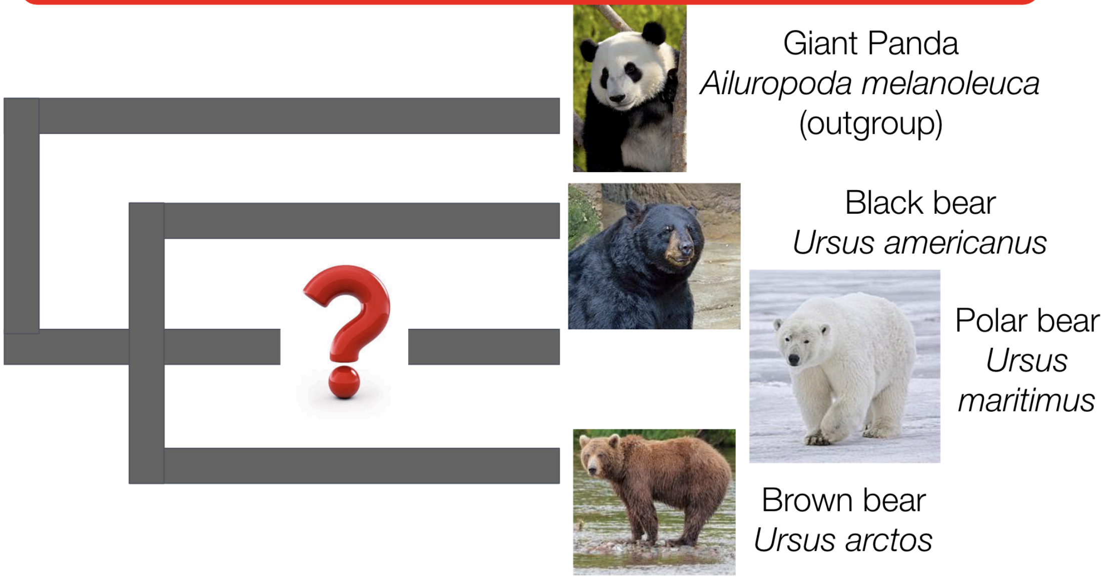
Generating phylogenomic data matrices: hands-on session

(Important piece of information (shared by Scott): Český Krumlov locals used to refer to the workshop participants as ‘**molekulos**’)

The Český Krumlov town hall decides to fund a project to understand whether the brown bear is more closely related to the polar bear or the American black bear

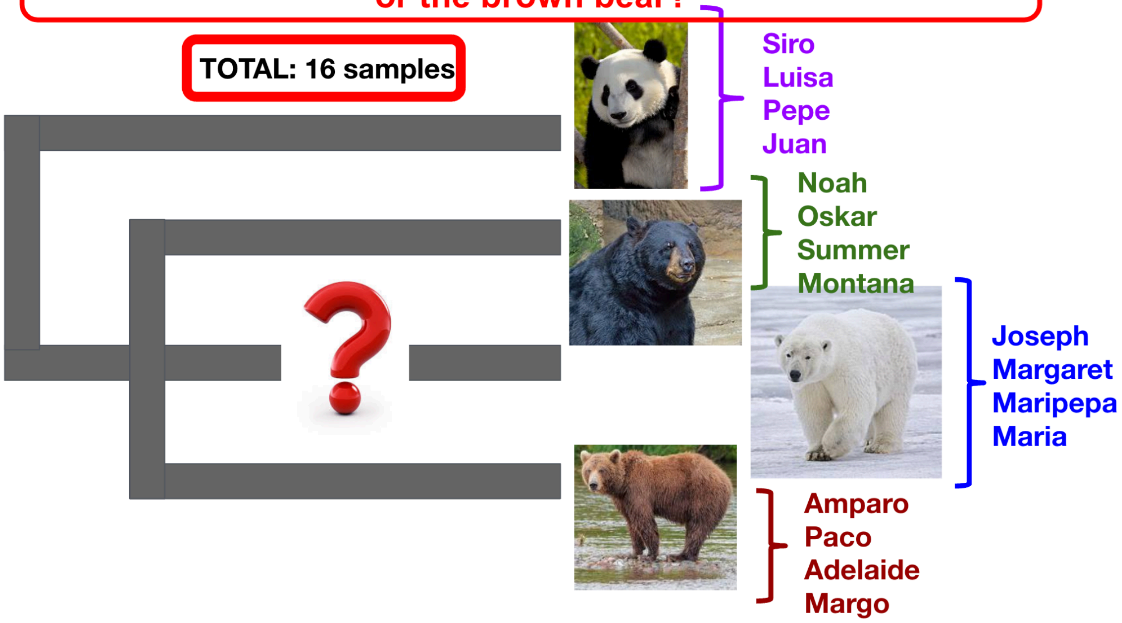
Let’s ask the ‘molekulos’ for help!!

Is the polar bear the sister group to the American black bear or the brown bear?



Is the polar bear the sister group to the American black bear or the brown bear?

TOTAL: 16 samples



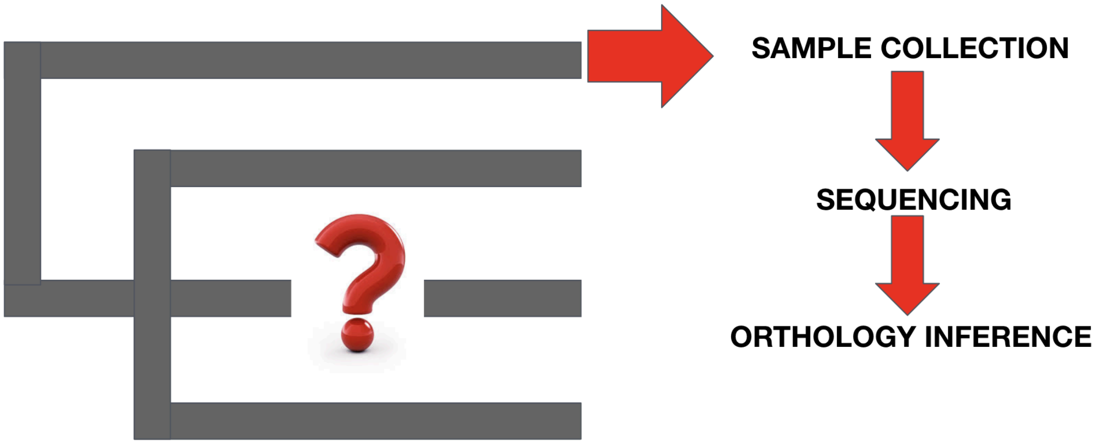
Siro
Luisa
Pepe
Juan

Noah
Oskar
Summer
Montana

Joseph
Margaret
Maripepa
Maria

Amparo
Paco
Adelaide
Margo

Is the polar bear the sister group to the American black bear or the brown bear?



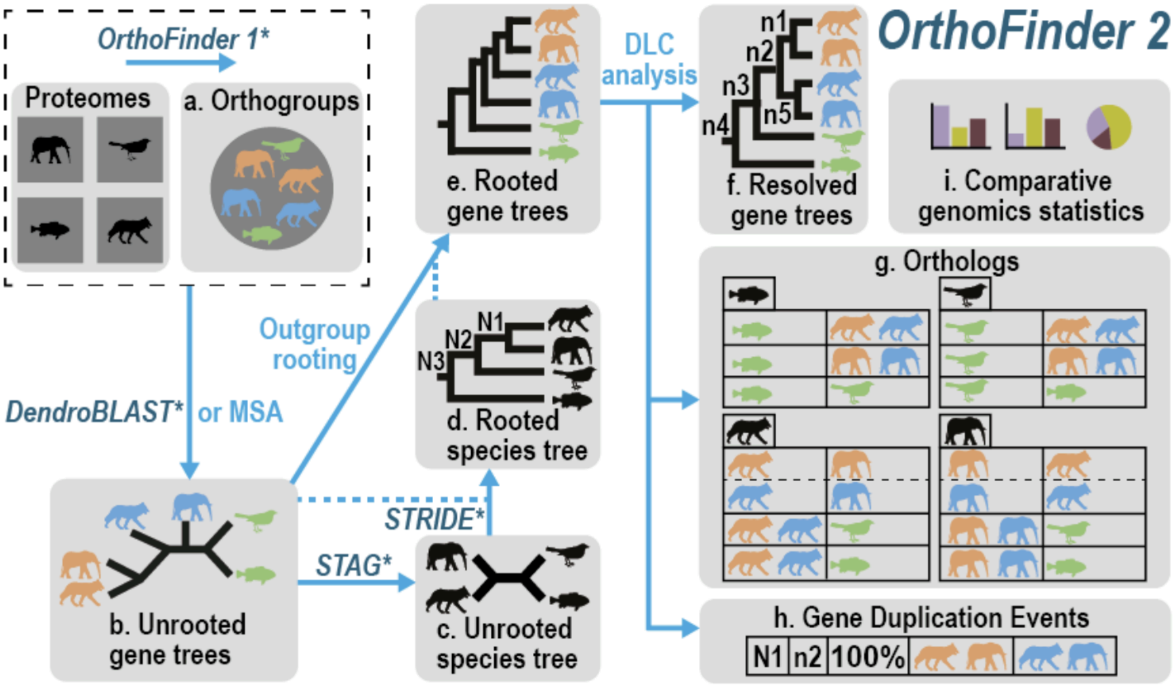
SAMPLE COLLECTION

SEQUENCING

ORTHOLOGY INFERENCE

STEP 1: Orthology Inference

1) Let's infer orthologous groups (OGs) in our bear dataset. We will use a tool called OrthoFinder.



We'll run all our analyses from the folder:

```
/home/genomics/workshop_materials/phylogenomics
```

2) Let's run OrthoFinder. For that, first we have to activate an environment for it to work:

```
conda activate orthofinder
```

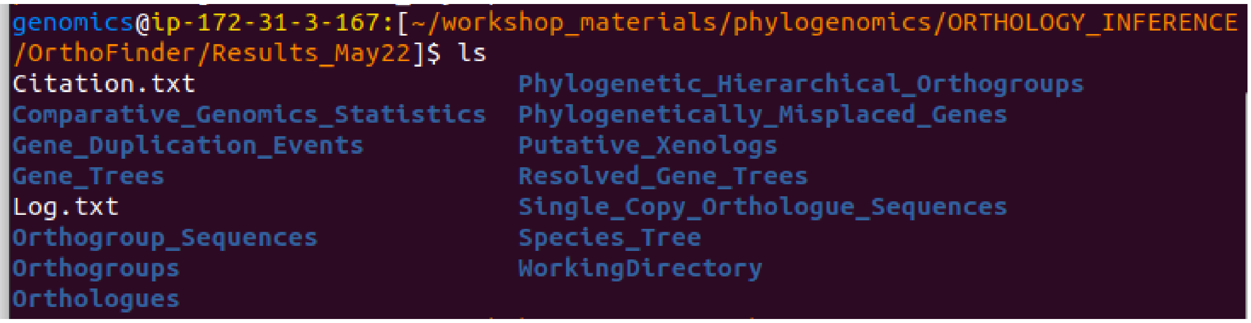
(New to conda environments? An environment is a directory that contains a specific collection of packages that you have installed. For example, you may have one environment with NumPy 1.7 and its dependencies, and another environment with NumPy 1.6 for legacy testing. If you change one environment, your other environments are not affected. You can easily activate or deactivate environments, which is how you switch between them. You can also share your environment with someone. More info about environments here).

3) Now we're ready to run OrthoFinder! It's very simple, just run the command 'orthofinder' followed by 'f' and the name of the species where you have all your datasets, in this case 'ORTHOLOGY_INFERENCE'. *(Note that since orthology inference may take a while, we are just going to test it with 4 out of our 16 samples, one per species).*

```
orthofinder -f ORTHOLOGY_INFERENCE
```

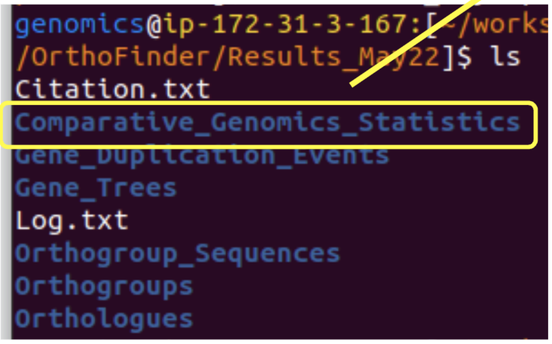
4) Let's inspect the output of OrthoFinder. You'll see that a new folder has been automatically created, it's called 'Results+[today's date]', in our case this is the path:

ORTHOLOGY_INFERENCE/OrthoFinder/Results_Jan18



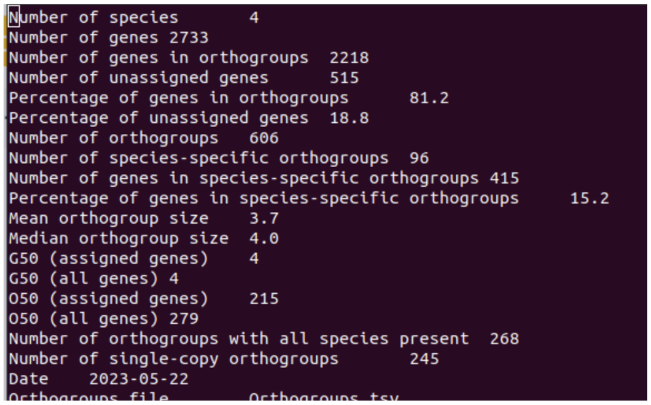
```
genomics@ip-172-31-3-167:[~/workshop_materials/phylogenomics/ORTHOLOGY_INFERENCE
/OrthoFinder/Results_May22]$ ls
Citation.txt                    Phylogenetic_Hierarchical_Orthogroups
Comparative_Genomics_Statistics Phylogenetically_Misplaced_Genes
Gene_Duplication_Events         Putative_Xenologs
Gene_Trees                      Resolved_Gene_Trees
Log.txt                         Single_Copy_Orthologue_Sequences
Orthogroup_Sequences            Species_Tree
Orthogroups                     WorkingDirectory
Orthologues
```

5) Some important files and folders that you may want to check are the following:



```
genomics@ip-172-31-3-167:[~/works
/OrthoFinder/Results_May22]$ ls
Citation.txt
Comparative_Genomics_Statistics
Gene_Duplication_Events
Gene_Trees
Log.txt
Orthogroup_Sequences
Orthogroups
Orthologues
```

File 'Statistics_Overall.tsv'



```
Number of species       4
Number of genes 2733
Number of genes in orthogroups  2218
Number of unassigned genes      515
Percentage of genes in orthogroups      81.2
Percentage of unassigned genes  18.8
Number of orthogroups   606
Number of species-specific orthogroups  96
Number of genes in species-specific orthogroups 415
Percentage of genes in species-specific orthogroups     15.2
Mean orthogroup size    3.7
Median orthogroup size  4.0
G50 (assigned genes)    4
G50 (all genes) 4
O50 (assigned genes)    215
O50 (all genes) 279
Number of orthogroups with all species present  268
Number of single-copy orthogroups       245
Date    2023-05-22
Orthogroups file        Orthogroups.tsv
```

Here you can see the number of species and genes, the number of genes in orthogroups, species-specific orthogroups, etc. One important number to have a look at for species tree inference is the **number of single-copy genes**, as this will be our start to build our matrix, and this is what we will do in this hands-on session.

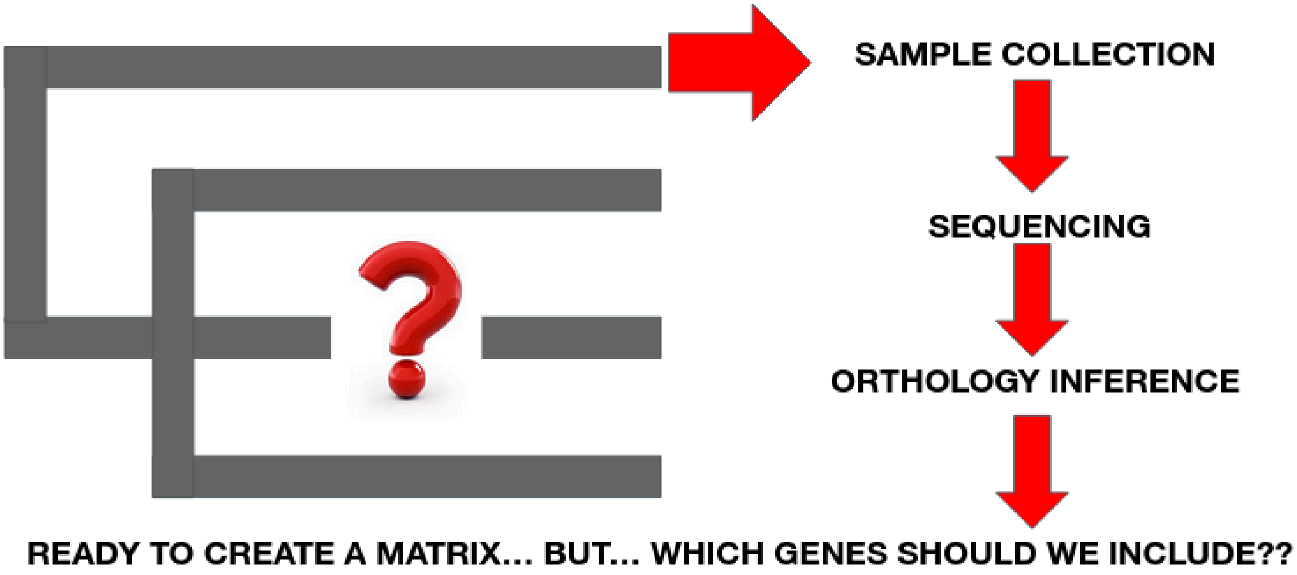
But let me warn you, the more species we include, the lower this number will be. If you have a very large dataset (eg, 100 or more terminals or so), this number may be super low or even ***zero***! So… what do we do then? If this is the case, you can use some tools such as PhyloPyPruner to prune monophyletic groups containing your species of interest from larger orthogroups that include paralogs (we won't be doing this during this lab, but feel free to explore it later - please note that you will need to install PhyloPyPruner since it's not in the instance).

OrthoFinder also infers **putative xenologs** (ie, a type of orthologs that result from horizontal gene transfer; they should not be trusted 100% but are a good start to test more in depth if they are if you're interested), it gives you the gene trees (but be aware that they are inferred in a 'fast' manner; feel free to check the manual to if you'd like OrthoFinder to infer the gene trees with more appropriate software or models such as IQ-TREE), **gene duplication events** (interesting if you're interested in understanding gene duplication) and it also infers a **species tree!** For this, it uses a method called STAG that leverages all orthologous and also paralogous genes (ie, it takes into account the information of gene duplications as well). Feel free to inspect the species tree, it's written in Newick format. You can visualize it in iTOL (just copy-paste the tree in Newick format in the web server where it says 'Upload Tree').

What is the topology of the bear tree inferred by STAG?

In any case, it is strongly recommended to follow the supermatrix approach and build some matrices for yourself to infer the tree as well (you can add the results of STAG to your paper too, but otherwise the reviewers will complain). So let's learn how to do so!

Is the polar bear the sister group to the American black bear or the brown bear?



SAMPLE COLLECTION

SEQUENCING

ORTHOLOGY INFERENCE

READY TO CREATE A MATRIX... BUT... WHICH GENES SHOULD WE INCLUDE??

STEP 2: Supermatrix construction

6) As mentioned during the lecture, there are some parameters and factors that are important to take into account when preparing our matrix. The first one is **missing data** (see slides from the lecture to understand why). To learn to check how much we have, let's go to the folder **MISSING_DATA** (full path **/home/genomics/workshop_materials/phylogenomics/MISSING_DATA**).

If you check the list of files in the folder (command ls), you'll see that there are 50 orthologous genes (called 'number.fa', eg 100.fa, ..., n.fa).

There are also 3 python scripts. For them to run, we'll need the python library **cogent** (a tool for statistical analysis of biological sequences) (already installed in the AMI, but if you're trying this in your computer/cluster you will need to install it).

Let's deactivate our current environment and activate the one that has cogent installed

```
conda deactivate
conda activate cogent
```

7) Let's explore the amount of missing data that we have in each taxon. Let's run the script:

```
python count_genesPerSample.py
```

(Note: the script doesn't tell you what is each column, so the headings are the following: ***Sample_name / no. orthogroups in that sample / no. total orthogroups in the dataset /proportion of orthogroups present in that sample****)*

Explore the amount of missing data in each taxon. Which individuals are poorly represented in each species?

We're going to go ahead and construct a matrix with all the samples, but in your own project you may want to exclude the ones that have a lot of missing data before doing so.

8) To create the matrix, we first have to align the genes within each orthogroup, otherwise the length of the matrix will vary per taxon. To do so, we will use the software **MUSCLE,** with the command:

```
muscle -align [your_gene] -output [your_gene.aln]
```

9) Let's align all the orthogroups with a loop, and let's have an output name with this nomenclature: your_gene.fa.aln (hint: check how to do it *here*). Now if you check the content of the folder you'll see that you have 50 new files called '100.fa.aln' to 'n.fa.aln'.

10) The next (and final!) step for generating our matrix is to concatenate the aligned orthogroups. For that, we will use a software called **catsequences**. We first need to create a list of all the files we want to concatenate, i.e. our aligned orthogroups. Please do so (you can use bash for that, or check here how to do it). Please call your file **list_all_orthogroups.txt**

11) Now, let's run catsequences:

```
catsequences list_all_orthogroups.txt
```

It will create two files: one with the information of the partitions (**allseqs.partitions.txt**) and the other one with a concatenated fasta with all genes (**allseqs.fas**). ***This is your matrix!***

12) Let's now infer a phylogenetic tree with our matrix. For that, one of the most heavily used softwares is **IQ-TREE,** as it provides very interesting options to play with different substitution models, partitions, etc. This software allows quite complex analyses, here we will just test some of the most basic options, but feel free to check the documentation and play with some of the most advanced options.

First of all, we will infer a maximum likelihood tree with one of the most general substitution models for amino acid data, called LG *(note we are using amino acid data, if you have nucleotidic data you will need to use different models, which can be checked in the software documentation).* For that, just run (it will take a few minutes):

```
iqtree2 -s allseqs.fas -m LG
```

Note as well that if you don't specify the model, it will run a first step of model testing, which is ideal but may take a while, so feel free to do it at home or when you're done with the tutorial. Note also that the default model selection does not include mixture models (in which the sites in sequences can undergo different substitution processes along the same or different trees). If you want to try model selection with mixture models, you have to specify which models you'd like to test with the '-madd' flag (eg, -madd C10,C20,C60,LG4X; see more info in the documentation).

You can see the best-fitted model in the file **allseqs.fas.iqtree,** and the maximum likelihood tree in the file **allseqs.fas.treefile.**

Which topology does this tree support? Is it the same or different to the STAG tree we inferred with Orthofinder?

13) Now let's select the genes that have a sample occupancy above a certain threshold (i.e., we want to create a matrix only with the genes that have a minimum of, let's say, 3 species, otherwise we have the risk of having some species represented by a low number of genes in our matrix, which could create some artifacts and biases). Let's first deactivate the conda environment we were using for the previous script requiring cogent, and then run a second python script:

```
conda deactivate
```

```
python2 select_sample_occupancy.py
```

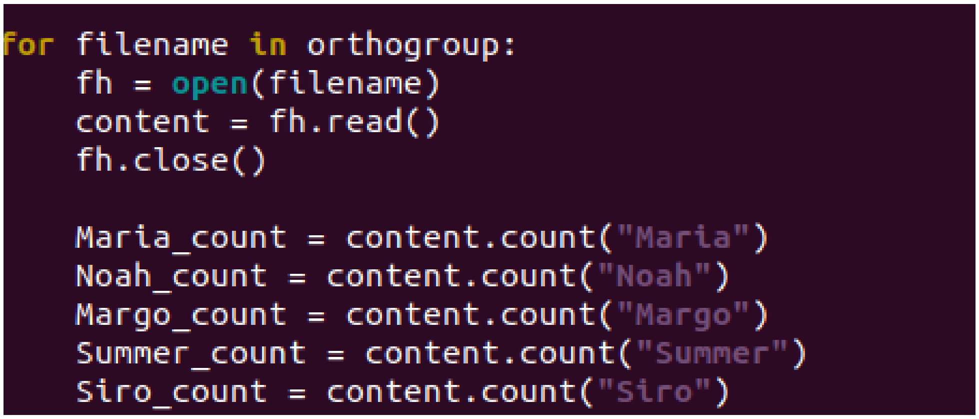
It will ask you to select the minimum sample occupancy. Let's start by 3 (ie, let's parse the orthogroups that have a minimum of three species). It will create a folder called **'orthologs_min_[number]_samples'**. Open it and check how many genes were selected with this threshold.

Run the script with different thresholds and check how the number of selected genes varies. Choose on the thresholds you've used, concatenate the aligned orthogroups and run a maximum likelihood tree. **Which topology are you obtaining?**

14) Let’s now think again on our goal: to resolve the interrelationships between *Ursus* species. If we select genes just based on sample occupancy, we may select some that do not include representatives of one or more of the species, and we’ll have a strongly biased dataset.

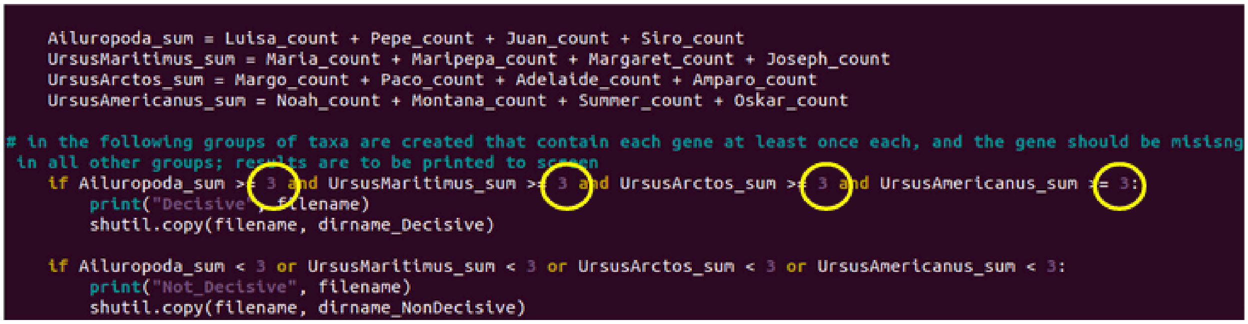
Let’s then select genes that have an homogeneous representation of all the four species. For that, we have a custom script named ‘**decisive_genes.py**’. Feel free to open it and inspect it.

Notice that at the end of the script we’re defining our four species and choosing a minimum number of individuals representing each species in the genes that will be selected (3 in this case).



```
for filename in orthogroup:
    fh = open(filename)
    content = fh.read()
    fh.close()

    Maria_count = content.count("Maria")
    Noah_count = content.count("Noah")
    Margo_count = content.count("Margo")
    Summer_count = content.count("Summer")
    Siro_count = content.count("Siro")
```



```
    Ailuropoda_sum = Luisa_count + Pepe_count + Juan_count + Siro_count
    UrsusMaritimus_sum = Maria_count + Maripepa_count + Margaret_count + Joseph_count
    UrsusArctos_sum = Margo_count + Paco_count + Adelaide_count + Amparo_count
    UrsusAmericanus_sum = Noah_count + Montana_count + Summer_count + Oskar_count

# in the following groups of taxa are created that contain each gene at least once each, and the gene should be misisng
 in all other groups; results are to be printed to screen
    if Ailuropoda_sum >= 3 and UrsusMaritimus_sum >= 3 and UrsusArctos_sum >= 3 and UrsusAmericanus_sum >= 3:
        print("Decisive", filename)
        shutil.copy(filename, dirname_Decisive)

    if Ailuropoda_sum < 3 or UrsusMaritimus_sum < 3 or UrsusArctos_sum < 3 or UrsusAmericanus_sum < 3:
        print("Not_Decisive", filename)
        shutil.copy(filename, dirname_NonDecisive)
```

15) Now run the script:

```
python decisive_genes.py
```

We now have 2 folders called **‘Decisive_genes3’** and **‘NonDecisive_genes3’**. Check how many genes you have in the ‘Decisive_genes3’ one. Change the threshold in the script, rerun it and check how the selected (=decisive) genes change.

Now you can play with these scripts to create different matrices, run some trees and see how the topology and the support for each node/lineage changes.

What topology is the most robustly supported one? Is missing data affecting the topology of your Maximum Likelihood tree?

16) To further test the robustness of your phylogenomic hypothesis you should also generate matrices accounting for other confounding factors, such as evolutionary rate, compositional heterogeneity, heterotachy, etc.

There are many softwares to do so that you can explore: **BMGE** (compositional heterogeneity at the level of site), **BaCoCa** (compositional heterogeneity at the level of gene), **TIGER2** (order genes by evolutionary rate), etc.

We are going to try **genesortR**, an R package that explores several of these properties at the same time.

17) Let's take our 50 orthogroups and analyze them with **genesortR** to see which ones are the most adequate to analyze. We will need to specify one species tree, so let's use the one that is the one most highly supported by our previous analyses, and that is written in Newick format and named 'bear_species_tree.tre', for this analysis. (new to Newick format? You can find more info here).

Data and scripts are located in: **phylogenomics/GENESORTR**. Go to that folder.

You will see 3 R scripts, the species tree, the 50 gene alignments concatenated (50_genes.fa), its correspondent partitions file (50_genes.partitions.txt), and the gene trees concatenated in Newick format (50_genes.nwk).

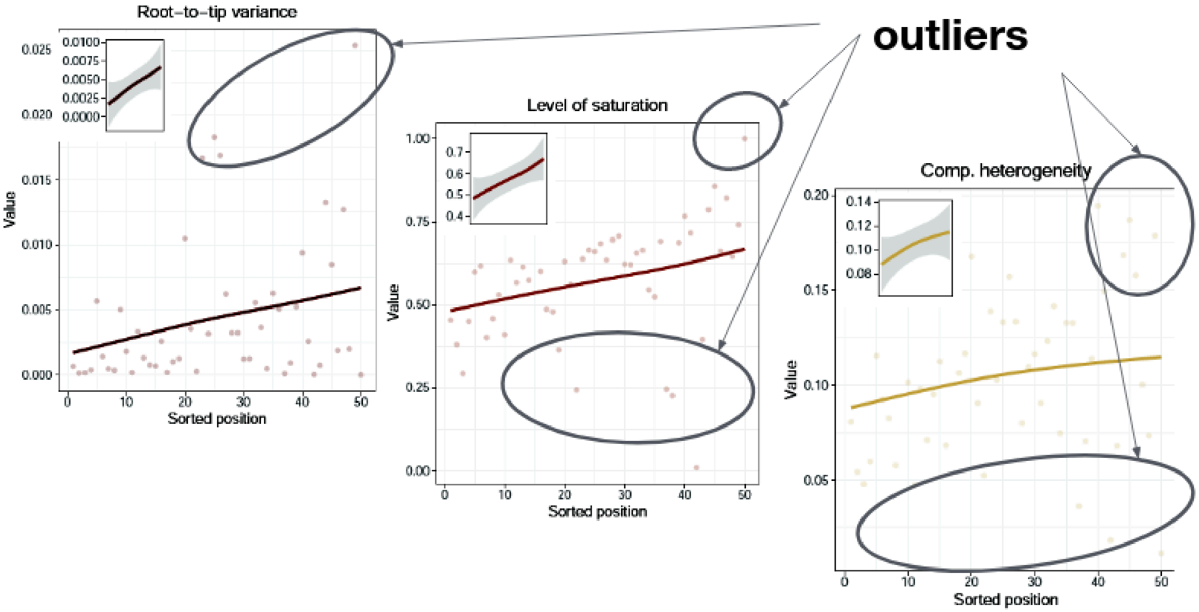
18) We will execute genesortR with default parameters on our 50 genes with this command:

```
Rscript genesortR.R
```

The names of the files are specified in the script, feel free to open it and inspect.

After running the script, we'll obtain a copy of our concatenated alignment, partition file and gene trees sorted by their phylogenetic usefulness, from most to least useful.

Take a look at the **sorted_figure_50_genes.pdf** file obtained. Which genes are most adequate for phylogenomic inference?



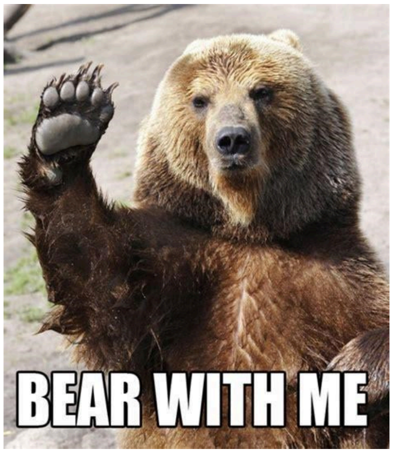
19) We will now test how selecting the *most* and the *least* phylogenetically useful genes affects the tree inferred.

To obtain the 10 best genes run: **Rscript select_10_best_genes.R**

To obtain the 10 worst genes run: **Rscript select_10_worst_genes.R**

Now use one of the phylogenetic inference programs that you have used before to run some trees and test how the phylogeny varies when using genes with different phylogenetic 'usefulness'. **Do you see any differences in the inferred topology compared to the other trees you inferred already?**

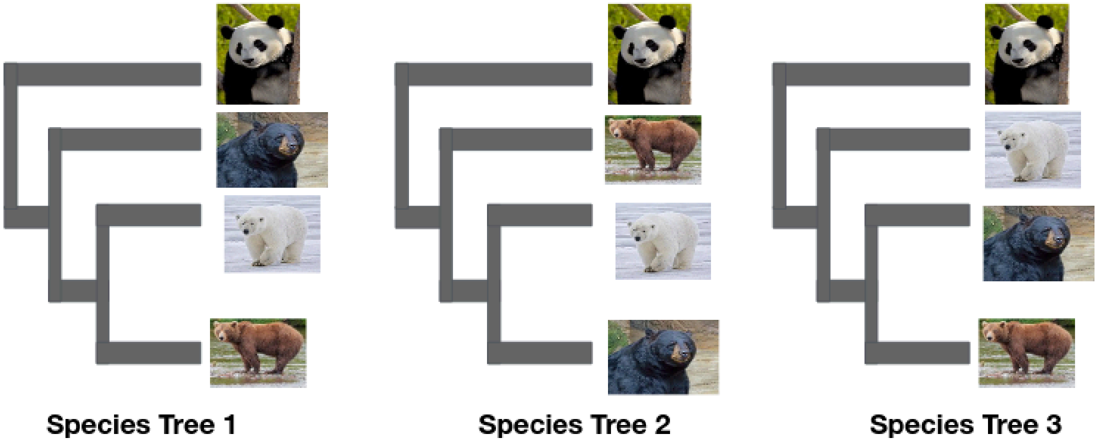
20) OK, so until now we've been doing analyses at the level of supermatrix. At this point, you should have inferred several phylogenetic trees and you're starting to have an idea of the robustness of how the species tree looks like. Let's wrap up this part before continuing with the tutorial, as there are a few theoretical slides about the next part that I'd like to share with you all about what's coming next: the multispecies coalescent. But before that, feel free to take a well-deserved break!



Section II: Species tree inference under the multispecies coalescent

21) **Ready for the last part of the tutorial?** Let's now explore which species tree is most robustly supported by each individual gene tree considered individually instead of concatenated. For that, we are going to use ASTRAL, a tool for estimating an unrooted species tree given a set of unrooted gene trees. ASTRAL is statistically consistent under the multi-species coalescent model (and thus is useful for handling incomplete lineage sorting, i.e., ILS). ASTRAL finds the species tree that has the maximum number of shared induced quartet trees with the set of gene trees, subject to the constraint that the set of bipartitions in the species tree comes from a predefined set of bipartitions.

22) Let's analyze conflict between individual gene trees to see which phylogenetic hypothesis is the most robustly supported:



23) We have selected 50 orthologous genes and have run individual gene trees with IQ-TREE. Let's have a look at them here:

phylogenomics/ASTRAL (.tree files)

24) ASTRAL needs all gene trees in the same file. For that, **let's concatenate them**, feel free to use the cat command and name the file 'bears_allTrees.tre' (hint: you can check here how to do it).

25) Let's now run an analysis on the 50 individual gene trees:

```
java -jar $HOME/software/Astral/astral.5.7.8.jar -i bears_allTrees.tre 2> output_ASTRAL.txt
```

26) Examine the output. What is the optimal tree inferred by ASTRAL? What is the final **normalized quartet score**?

(The normalized quartet score is the proportion of input gene tree quartet trees satisfied by the species tree. This is a number between zero and one; the higher this number, the *less* discordant your gene trees are).

27) So far ASTRAL showed us the preferred topology. Let's now check how our individual gene trees support the alternatives topologies.

For that, let's score each species tree topology and compare the normalized quartet score for each one.

28) Check the three provided species trees (bear_species_tree1.tre, bear_species_tree2.tre, bear_species_tree3.tre). Visualize them and identify the differences.

Let's now score them with ASTRAL, starting with the first one. From the ASTRAL folder, run:

```
java -jar $HOME/software/Astral/astral.5.7.8.jar -i bears_allTrees.tre -q bear_species_tree1.tre 2> score_speciesTree1.txt
```

29) Do the same with the species trees 2 and 3. Compare the results.

Which phylogenetic hypothesis is the most robustly supported?

Which branches are not supported by many genes in each analyses? Does this affect the overall preferred phylogeny of *Ursus*?

30) You've made it! This is the end of the tutorial. Let's now reconvene to wrap up this lab together. Feel free to play with the advanced options mentioned throughout the tutorial, as we are just covering the basics here and phylogenomic inference can be pretty complex.