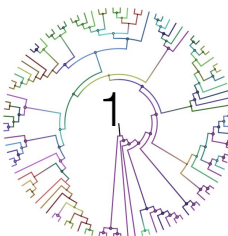


Generating phylogenomic data matrices: hands-on session

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



Generating phylogenomic data matrices: hands-on session

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

TOTAL: 16 samples



Siro
Luisa
Pepe
Juan



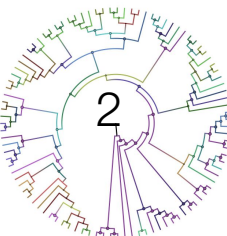
Noah
Oskar
Summer
Montana



Joseph
Margaret
Maripepa
Maria

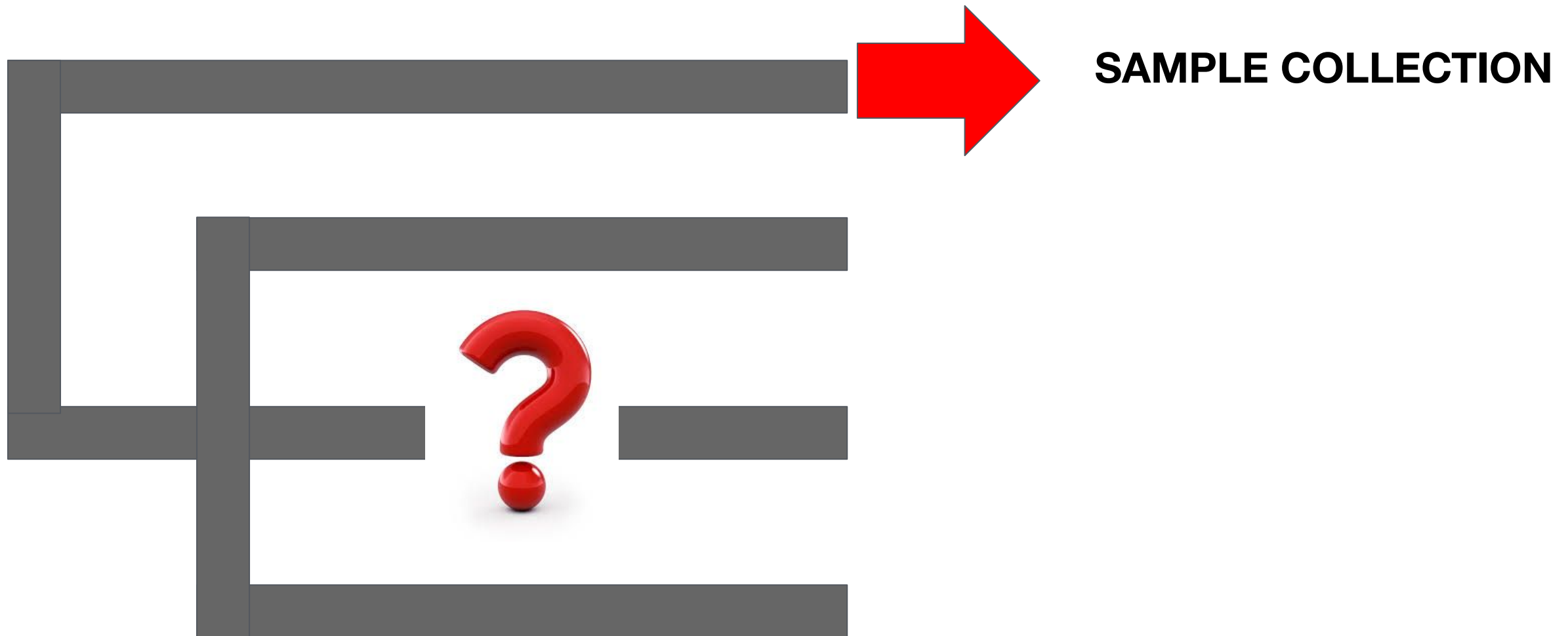


Amparo
Paco
Adelaide
Margo



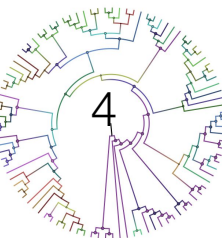
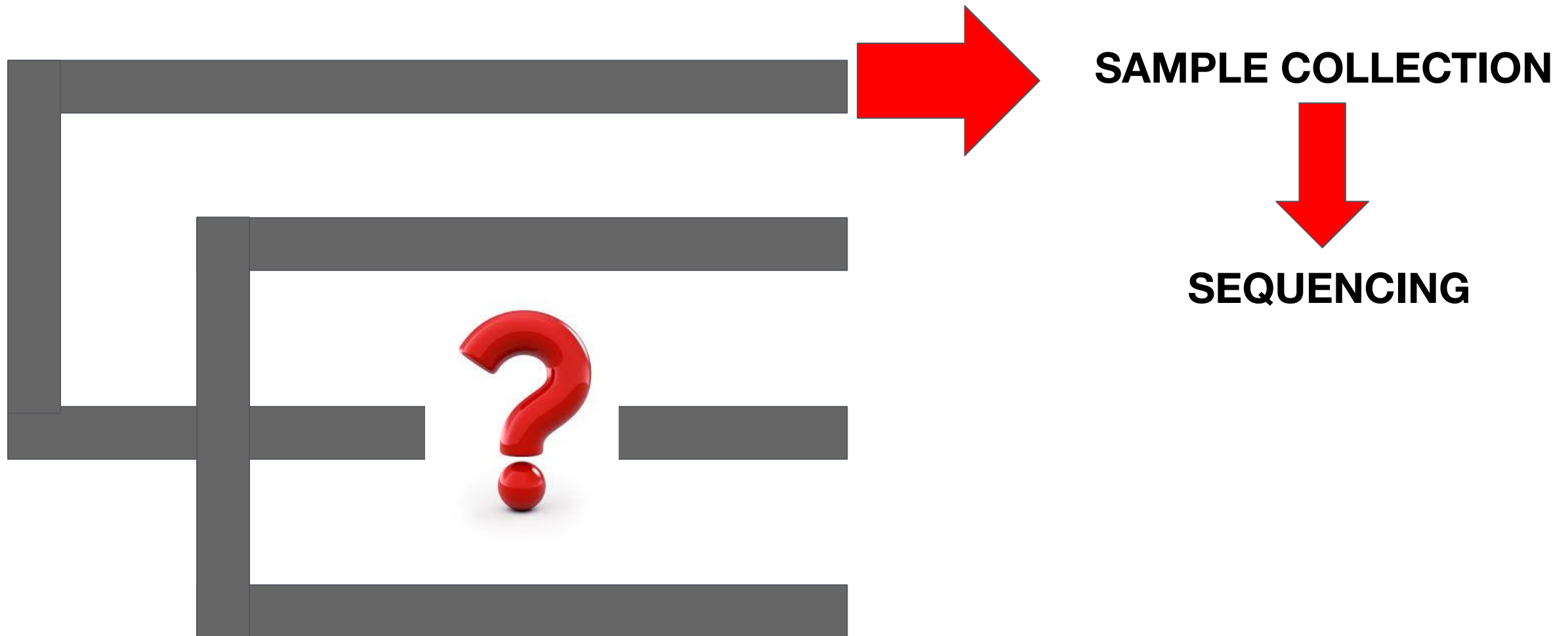
Generating phylogenomic data matrices: hands-on session

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



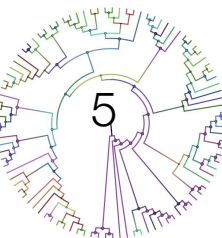
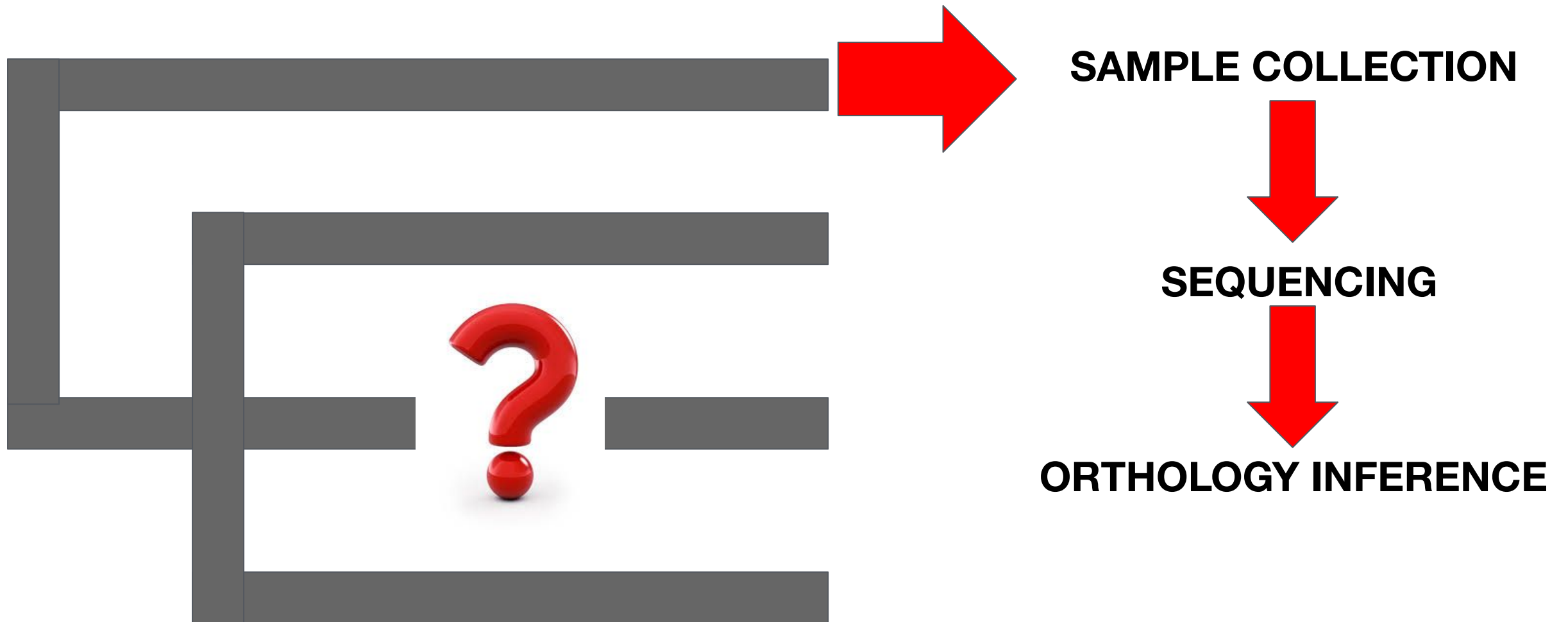
Generating phylogenomic data matrices: hands-on session

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



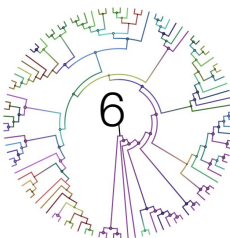
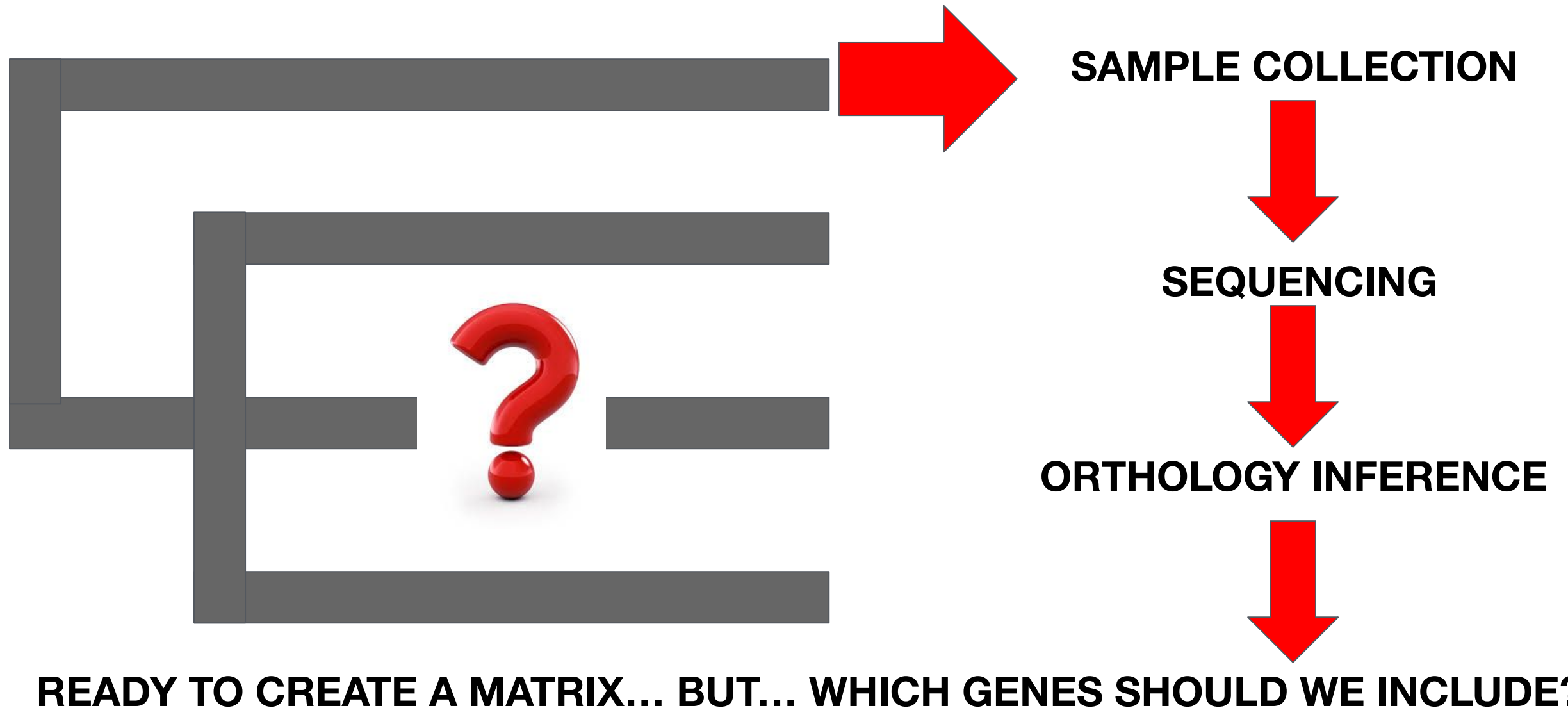
Generating phylogenomic data matrices: hands-on session

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



Generating phylogenomic data matrices: hands-on session

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



Generating phylogenomic data matrices: hands-on session

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

- **MISSING DATA**

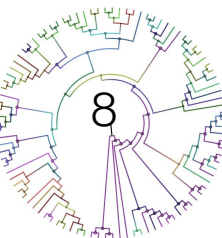


Generating phylogenomic data matrices: hands-on session

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

- **MISSING DATA**

Let's create **different matrices with different taxon occupancy** to account for the effect of missing data.



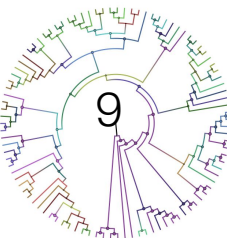
Generating phylogenomic data matrices: hands-on session

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

- **MISSING DATA**

Let's create **different matrices with different taxon occupancy** to account for the effect of missing data.

- 1) The data is located in the folder
Bear_dataset_SensitivityAnalyses/MISSING_DATA



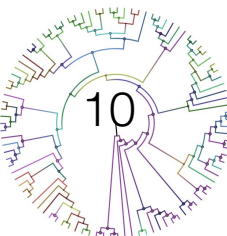
Generating phylogenomic data matrices: hands-on session

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

- **MISSING DATA**

Let's create **different matrices with different taxon occupancy** to account for the effect of missing data.

- 1) The data is located in the folder
Bear_dataset_SensitivityAnalyses/MISSING_DATA
- 2) If you check the list of files in the folder (ls), you'll see that there are 50 orthologous genes ('number.fa'). They're already aligned.



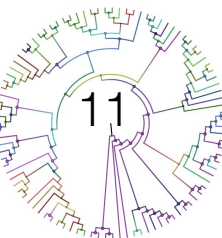
Generating phylogenomic data matrices: hands-on session

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

- **MISSING DATA**

Let's create **different matrices with different taxon occupancy** to account for the effect of missing data.

- 1) The data is located in the folder
Bear_dataset_SensitivityAnalyses/MISSING_DATA
- 2) If you check the list of files in the folder (ls), you'll see that there are 50 orthologous genes ('number.fa'). They're already aligned.
- 3) There are also 3 python scripts. For them to run, we'll need the python libraries **numpy** and **cogent** (already installed in our cluster).



Generating phylogenomic data matrices: hands-on session

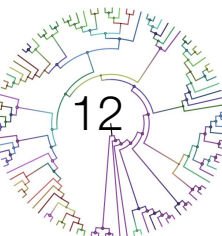
Is the polar bear the sister group to the American black bear or the brown bear?

- **MISSING DATA**

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

`python count_genesPerSpecies.py`

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



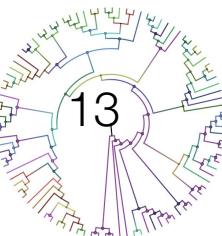
Generating phylogenomic data matrices: hands-on session

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

- **MISSING DATA**

5) Now let's select the genes that have a taxon 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). Let's run the script:

```
python select_taxon_occupancy.py
```



Generating phylogenomic data matrices: hands-on session

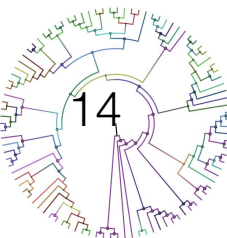
Is the polar bear the sister group to the American black bear or the brown bear?

- **MISSING DATA**

5) Now let's select the genes that have a taxon 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). Let's run the script:

`python select_taxon_occupancy.py`

It will ask you to select the minimum taxon occupancy. Let's start by 3. It will create a folder called '**`orthologs_min_[number]_taxa`**'. Open it and check how many genes were selected with this threshold.



Generating phylogenomic data matrices: hands-on session

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

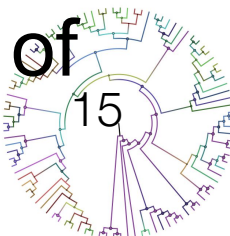
- **MISSING DATA**

5) Now let's select the genes that have a taxon 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). Let's run the script:

```
python select_taxon_occupancy.py
```

It will ask you to select the minimum taxon occupancy. Let's start by 3. It will create a folder called '**orthologs_min_[number]_taxa**'. 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.



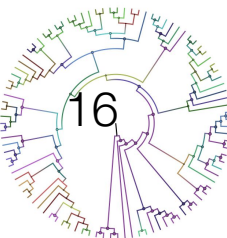
Generating phylogenomic data matrices: hands-on session

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

- **MISSING DATA**

6) Let's now think again on our goal: to resolve the interrelationships between *Ursus* species. If we select genes just based on taxon 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.



Generating phylogenomic data matrices: hands-on session

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

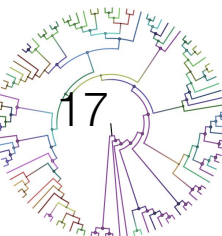
- **MISSING DATA**

7) Let's open the **decisive_genes.py** script and inspect it together.

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).

Run the script:

```
python decisive_genes.py
```



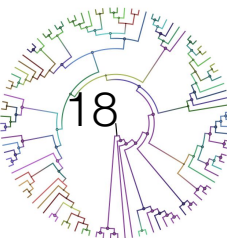
Generating phylogenomic data matrices: hands-on session

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

- **MISSING DATA**

8) 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.

How do you think this may affect the inferred phylogenetic relationships?



Generating phylogenomic data matrices: hands-on session

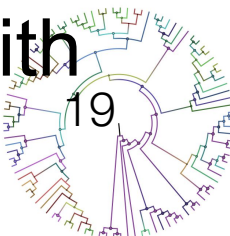
Is the polar bear the sister group to the American black bear or the brown bear?

- **MISSING DATA**

8) 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.

How do you think this may affect the inferred phylogenetic relationships?

9) Now (or at the end of the class) you can play with these scripts to create different matrices (see guidelines for instructions), run some trees and see how the topology and the support for each node/lineage changes. But first let's learn how to create a matrix with these genes.



Generating phylogenomic data matrices: hands-on session

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

- **MISSING DATA**

To create your matrix, first you will need to concatenate the genes selected for each matrix. We will use the software **catsequences**



Generating phylogenomic data matrices: hands-on session

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

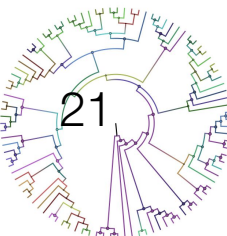
- **MISSING DATA**

We have to create a file with the list of genes that we'd like to concatenate. Let's create a file with the list of decisive genes. For that, go the folder containing your decisive genes

(orthologs_3_fasta) (**cd MISSING_DATA/orthologs_min_3_fasta**) and execute the following command: **ls *.fa > list_decisive_genes.txt**

To concatenate the genes, run the following command:
catsequences list_decisive_genes.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!!



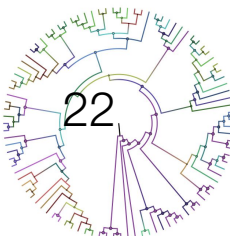
Generating phylogenomic data matrices: hands-on session

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

- **MISSING DATA**

Once the genes are concatenated, use one of the phylogenetic inference programs that you learned in the previous days to run a quick tree and test how the phylogeny varies depending on the genes that you include.

Do you see many differences? What is the factor that is affecting the most in the tests you run?



Generating phylogenomic data matrices: hands-on session

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

- **OTHER PROPERTIES: genesortR**

10) 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.

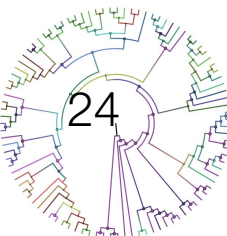
Generating phylogenomic data matrices: hands-on session

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

- **OTHER PROPERTIES: genesortR**

10) 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.



Generating phylogenomic data matrices: hands-on session

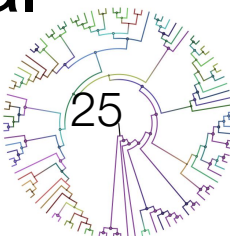
Is the polar bear the sister group to the American black bear or the brown bear?

- **OTHER PROPERTIES: genesortR**

10) 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.



Generating phylogenomic data matrices: hands-on session

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

- **OTHER PROPERTIES: genesortR**

11) Let's take our 50 orthogroups and analyze them with **genesortR** to see which ones are the most adequate to analyze. We will use species tree 1 for this analysis.

Generating phylogenomic data matrices: hands-on session

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

- **OTHER PROPERTIES: genesortR**

11) Let's take our 50 orthogroups and analyze them with **genesortR** to see which ones are the most adequate to analyze. We will use species tree 1 for this analysis.

Data and scripts are located in:

Bear_dataset_SensitivityAnalyses/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 newick gene trees concatenated (50_genes.nwk).



Generating phylogenomic data matrices: hands-on session

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

- **OTHER PROPERTIES: genesortR**

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

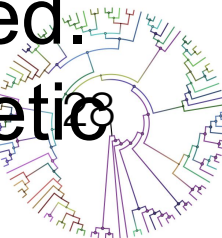
module purge

module load r-adephylo r-phytools r-phangorn r-cowplot

Rscript genesortR.R

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. How do gene properties vary according to their phylogenetic usefulness?



Generating phylogenomic data matrices: hands-on session

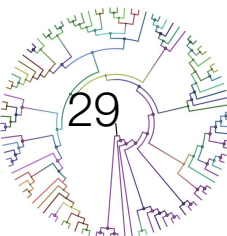
Is the polar bear the sister group to the American black bear or the brown bear?

- **OTHER PROPERTIES: genesortR**

13) 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**



Generating phylogenomic data matrices: hands-on session

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

- **OTHER PROPERTIES: genesortR**

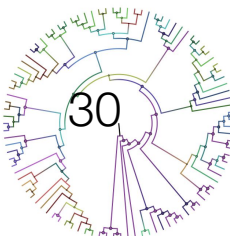
13) We will now test how selecting the *most* and the *least* phylogenetically useful genes affects the tree inferred.

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**

Once we have obtained the smaller alignments, use one of the phylogenetic inference programs that you have used before to run a quick tree and test how the phylogeny varies when using genes with different phylogenetic usefulness.

Do you see any differences?



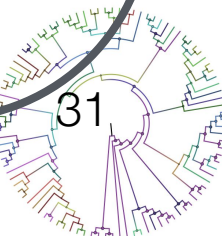
Analyzing gene tree/species tree conflict: hands-on session

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

ASTRAL is a tool for estimating an unrooted species tree given a set of unrooted gene trees.

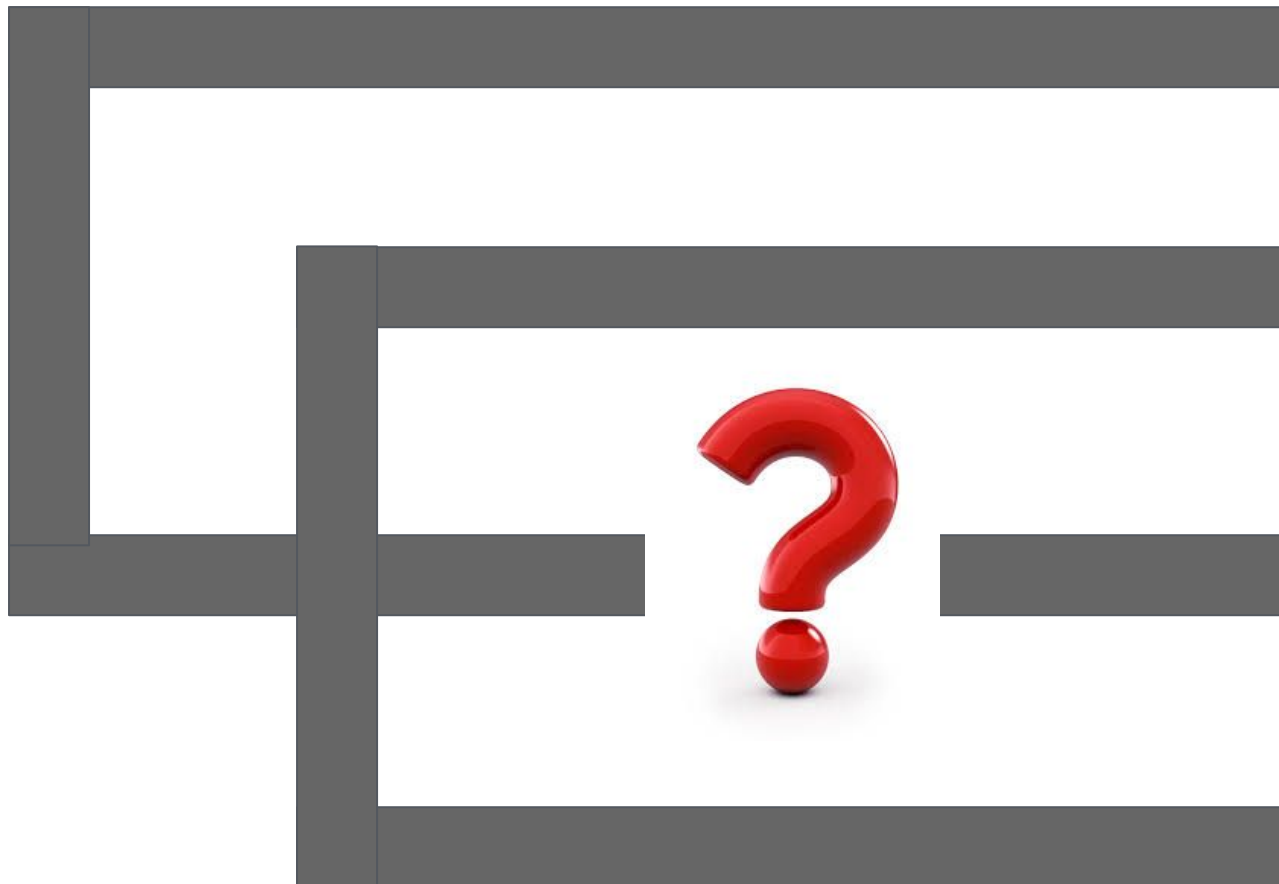
ASTRAL is statistically consistent under the multi-species coalescent model (and thus is useful for handling incomplete lineage sorting, i.e., ILS).

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.



Analyzing gene tree/species tree conflict: hands-on session

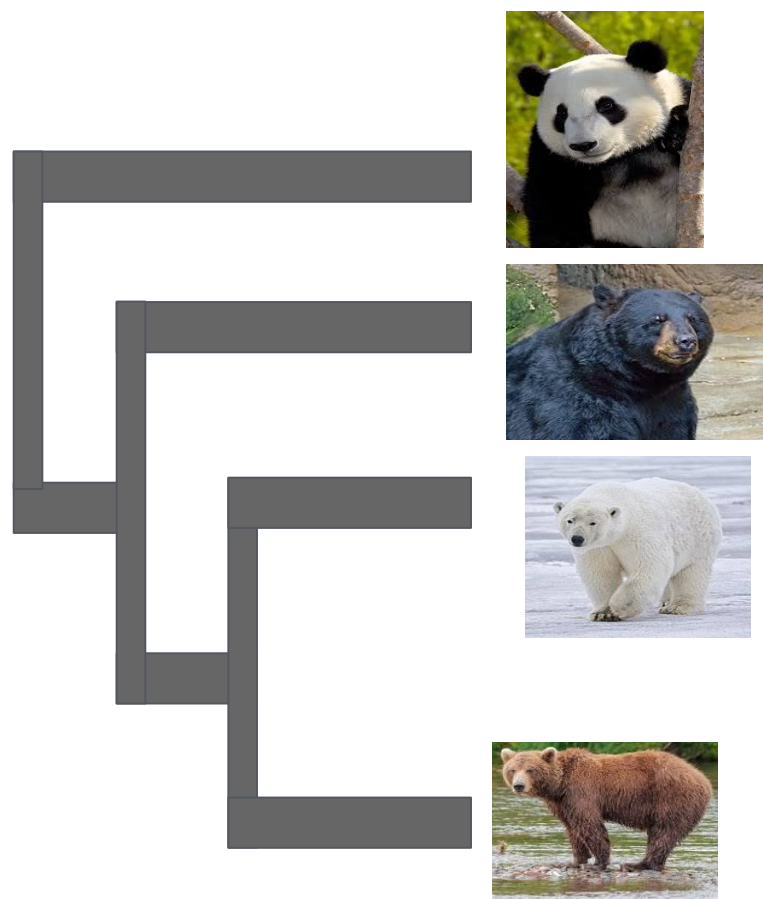
Is the polar bear the sister group to the American black bear or the brown bear?



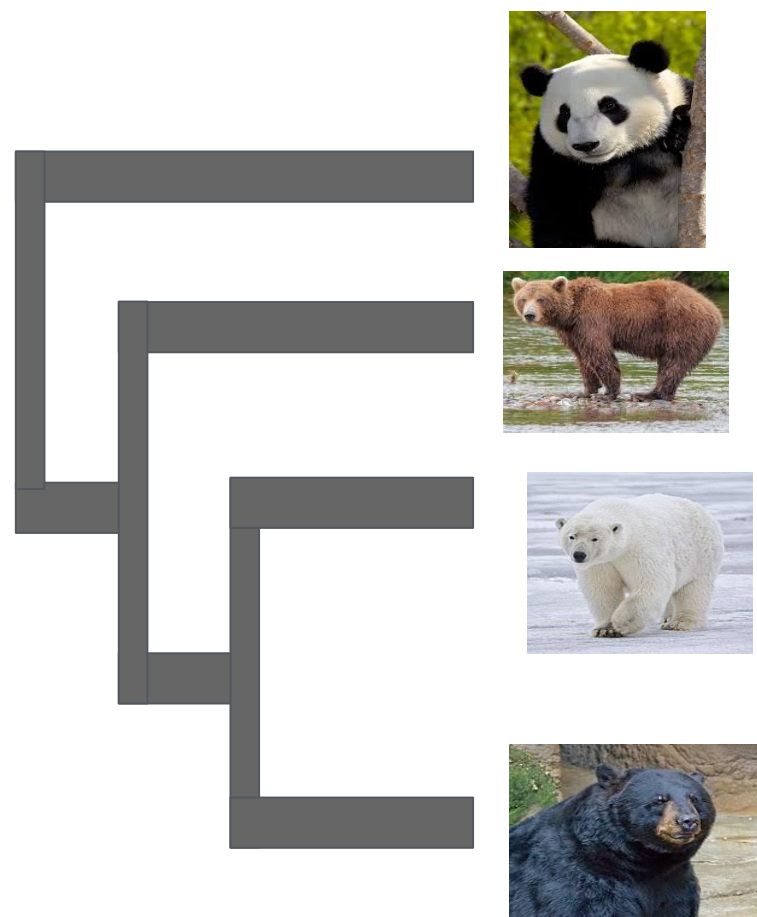
Analyzing gene tree/species tree conflict: hands-on session

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

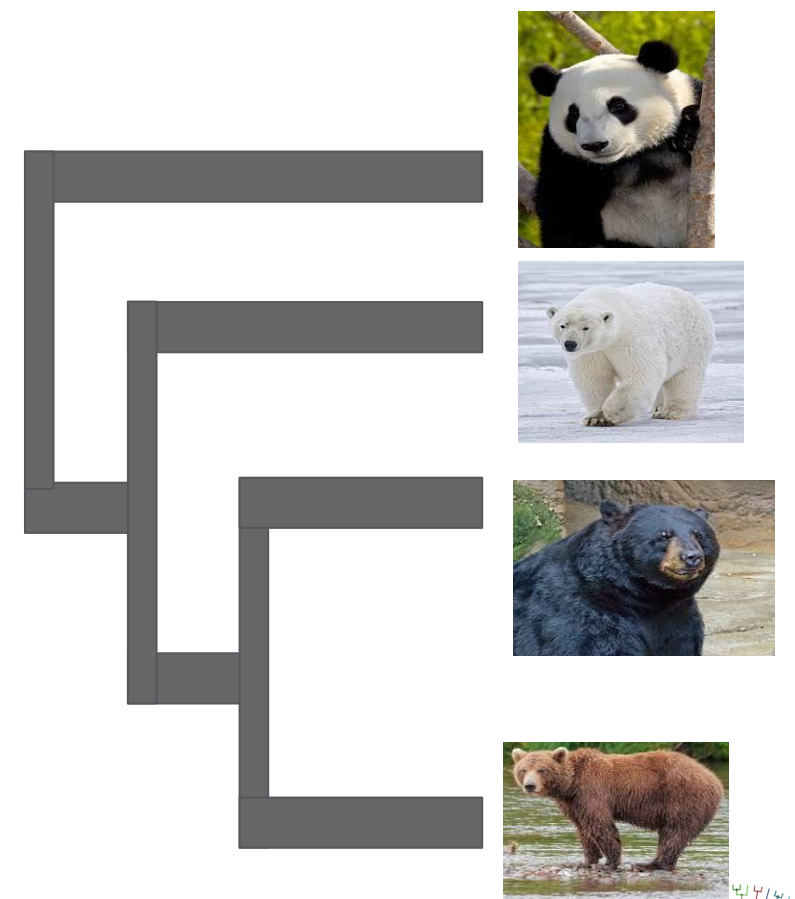
1) Let's analyze conflict between individual gene trees to see which phylogenetic hypothesis is the most robustly supported:



Species Tree 1



Species Tree 2



Species Tree 3



Analyzing gene tree/species tree conflict: hands-on session

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

2) We have selected 50 orthologous genes and have run individual gene trees with IQTree. Let's have a look at them here:

Bear_dataset_SensitivityAnalyses/PHYLIP (.tree files)



Analyzing gene tree/species tree conflict: hands-on session

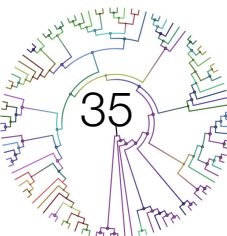
Is the polar bear the sister group to the American black bear or the brown bear?

2) We have selected 50 orthologous genes and have run individual gene trees with IQTree. Let's have a look at them here:

Bear_dataset_SensitivityAnalyses/PHYLIP (.tree files)

3) ASTRAL-III needs all gene trees in the same file. For that, let's concatenate them:

```
cat *trees > bears_allTrees.tre
```



Analyzing gene tree/species tree conflict: hands-on session

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

2) We have selected 50 orthologous genes and have run individual gene trees with IQTree. Let's have a look at them here:

Bear_dataset_SensitivityAnalyses/PHYLIP (.tree files)

3) ASTRAL-III needs all gene trees in the same file. For that, let's concatenate them:

```
cat *trees > bears_allTrees.tre
```

And create a folder called ASTRAL and move the files there:

```
mkdir ASTRAL
```

```
mv bears_allTrees.tre ASTRAL
```

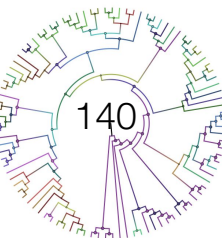
And go to that folder: **cd ASTRAL**



Analyzing gene tree/species tree conflict: hands-on session

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

4) Let's now run an analysis on the 50 individual gene trees:
astral -i bears_allTrees.tre

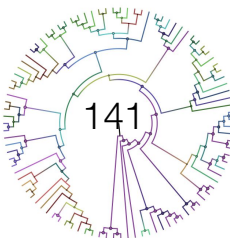


Analyzing gene tree/species tree conflict: hands-on session

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

4) Let's now run an analysis on the 50 individual gene trees:
astral -i bears_allTrees.tre

You'll see the output in the screen. To save it in an output file run:
astral -i bears_allTrees.tre 2> output_ASTRAL.txt



Analyzing gene tree/species tree conflict: hands-on session

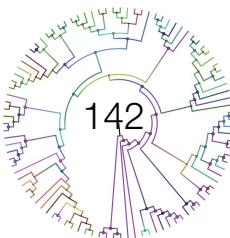
Is the polar bear the sister group to the American black bear or the brown bear?

4) Let's now run an analysis on the 50 individual gene trees:
astral -i bears_allTrees.tre

You'll see the output in the screen. To save it in an output file run:
astral -i bears_allTrees.tre 2> output_ASTRAL.txt

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.



Analyzing gene tree/species tree conflict: hands-on session

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

4) Let's now run an analysis on the 50 individual gene trees:
astral -i bears_allTrees.tre

You'll see the output in the screen. To save it in an output file run:
astral -i bears_allTrees.tre 2> output_ASTRAL.txt

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.

Visualize it in your preferred software or online in **phylo.io** (just copy-paste the newick tree and click 'Render').

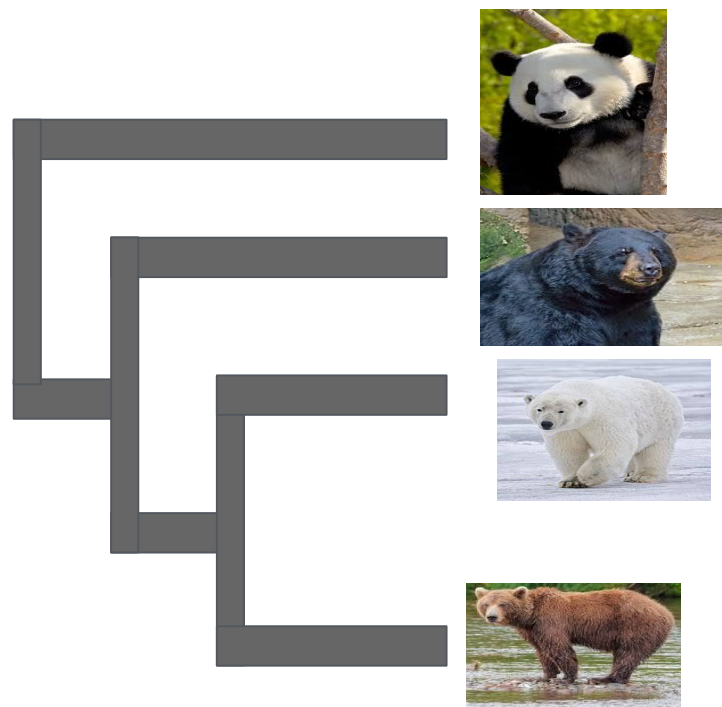


Analyzing gene tree/species tree conflict: hands-on session

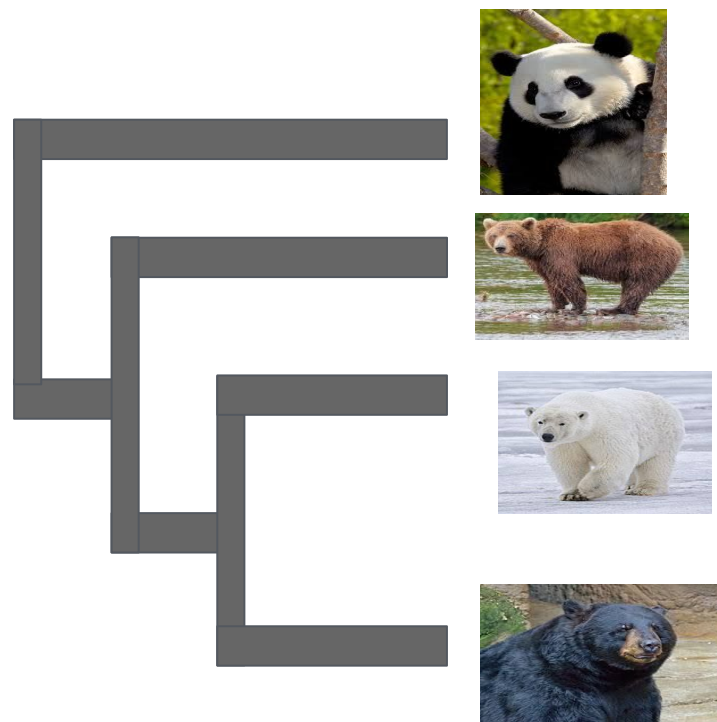
Is the polar bear the sister group to the American black bear or the brown bear?

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

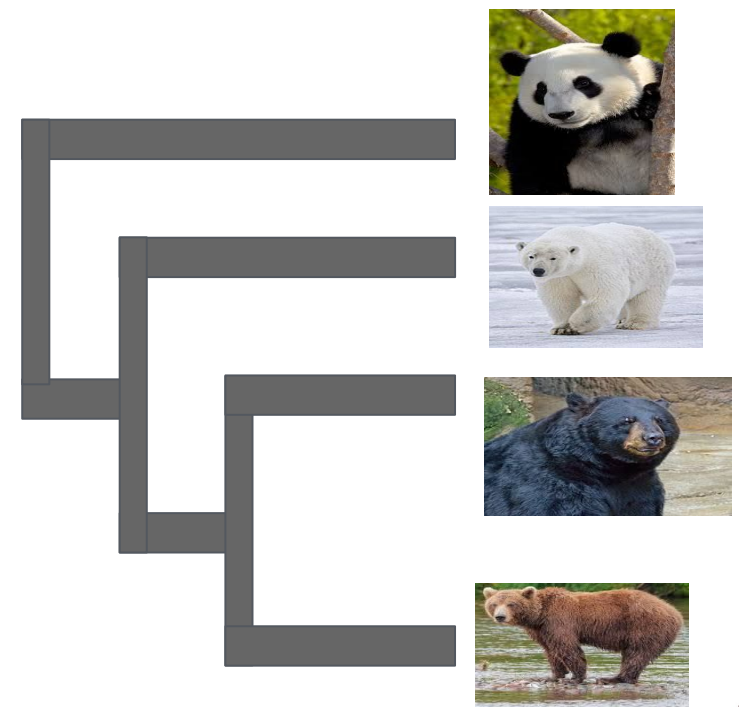
For that, let's score each species tree topology and compare the normalize quartet score for each one.



Species Tree 1



Species Tree 2



Species Tree 3



Analyzing gene tree/species tree conflict: hands-on session

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

5) 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 normalize quartet score for each one.

Go to the folder

Bear_dataset_SensitivityAnalyses/SPECIES_TREES

and check the three provided species trees. Visualize them and identify the differences.



Analyzing gene tree/species tree conflict: hands-on session

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

5) 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 normalize quartet score for each one.

Go to the folder

Bear_dataset_SensitivityAnalyses/SPECIES_TREES

and check the three provided species trees. Visualize them and identify the differences.

Let's now score them with ASTRAL.



Analyzing gene tree/species tree conflict: hands-on session

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

6) Let's score the first species tree. From the ASTRAL folder, run:

```
astral -i bears_allTrees.tre -q  
../SPECIES_TREES/bear_species_tree1.tre 2>  
score_speciesTree1.txt
```



Analyzing gene tree/species tree conflict: hands-on session

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

6) Let's score the first species tree. From the ASTRAL folder, run:

```
astral -i bears_allTrees.tre -q  
../SPECIES_TREES/bear_species_tree1.tre 2>  
score_speciesTree1.txt
```

Do the same with the species trees 2 and 3.



Analyzing gene tree/species tree conflict: hands-on session

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

6) Let's score the first species tree. From the ASTRAL folder, run:

```
astral -i bears_allTrees.tre -q  
../SPECIES_TREES/bear_species_tree1.tre 2>  
score_speciesTree1.txt
```

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*?

