Generating phylogenomic data matrices: hands-on session

Generating phylogenomic data matrices: hands-on session

Generating phylogenomic data matrices: hands-on session

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

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

Generating phylogenomic data matrices: hands-on session

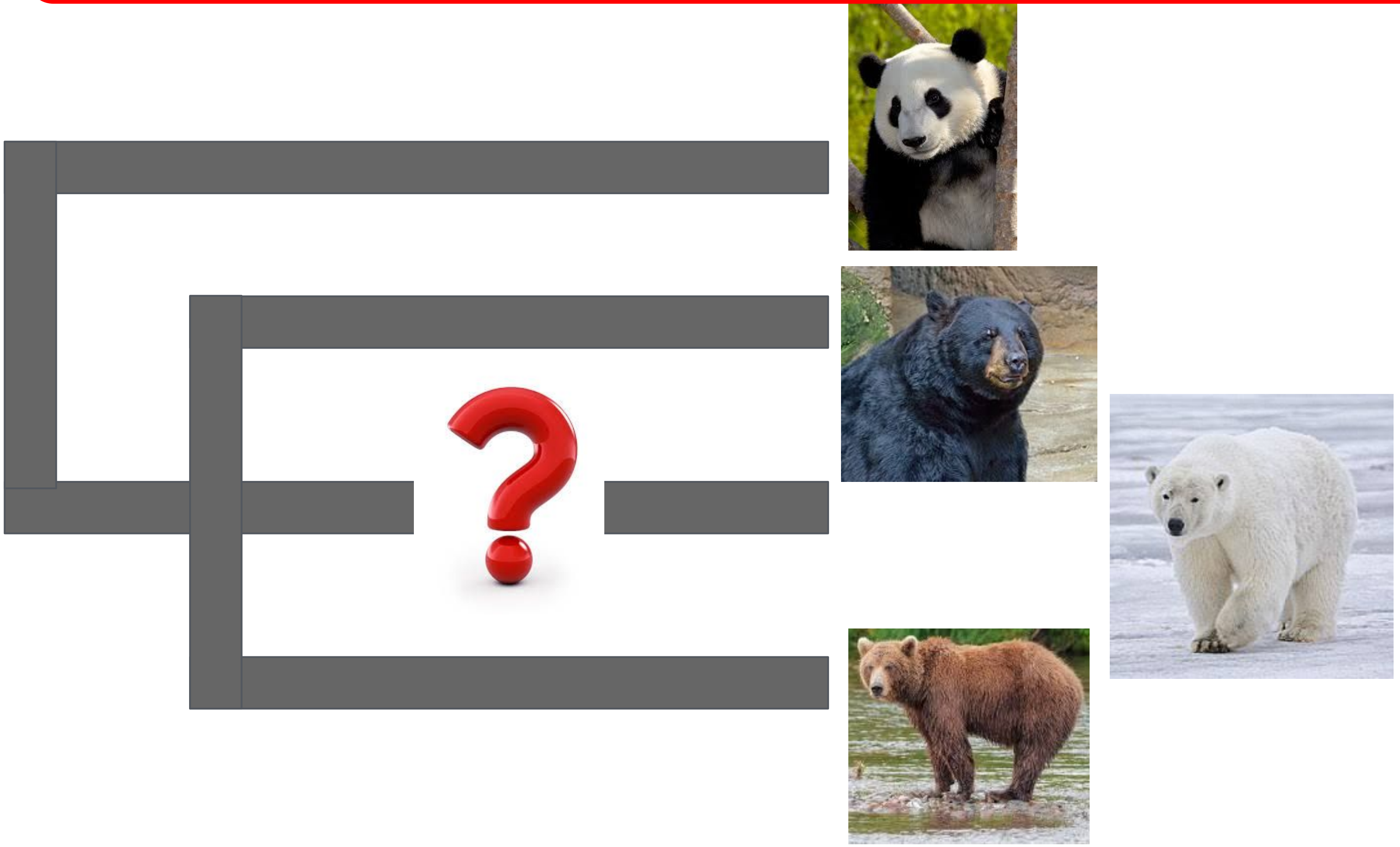
(Important piece of information (shared by Scott): Český Krumlov locals used to refer to the workshop participants as ‘**molekulos**’)

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!!

Generating phylogenomic data matrices: hands-on session

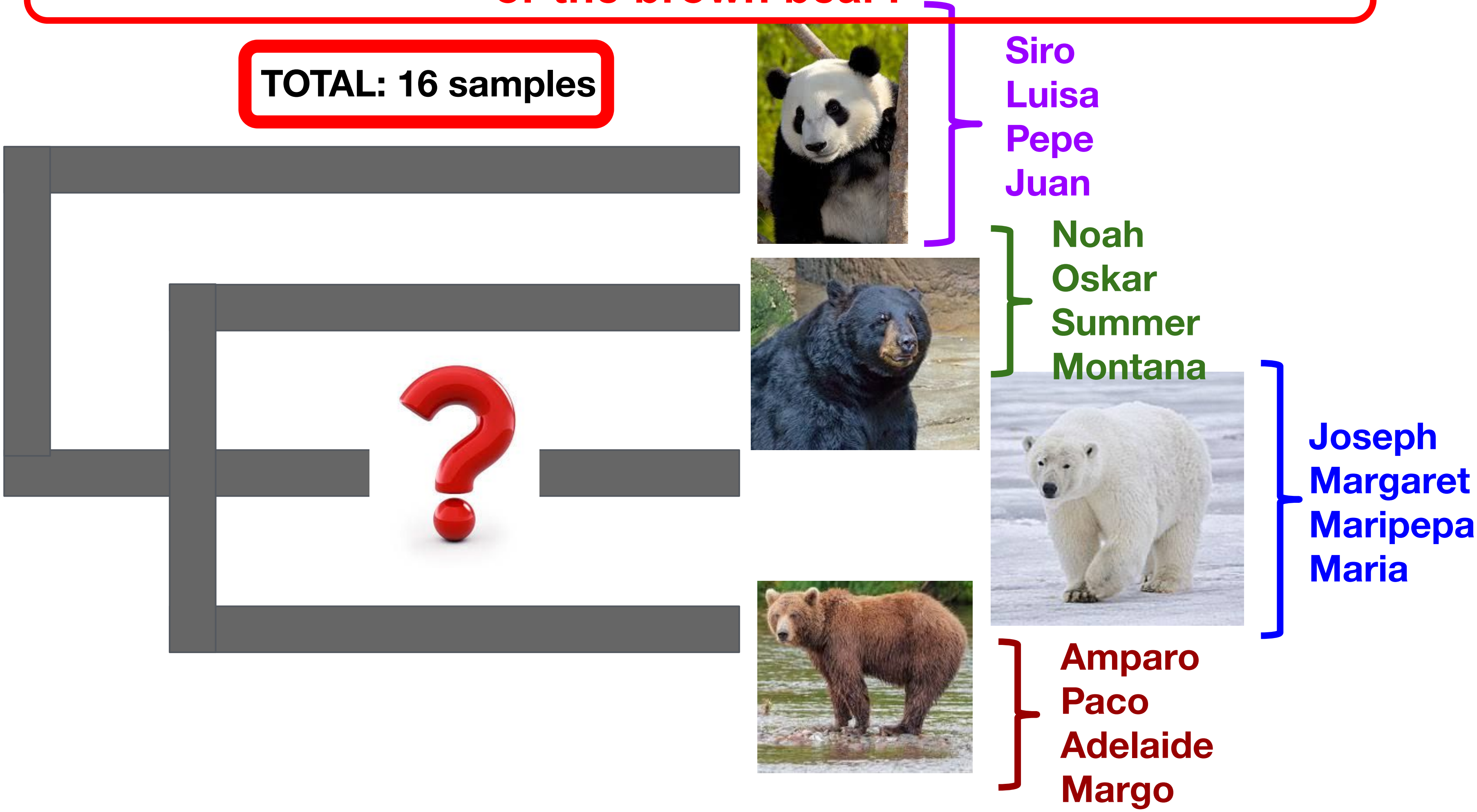
Is the polar bear the sister group to the American black bear or the brown bear?



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

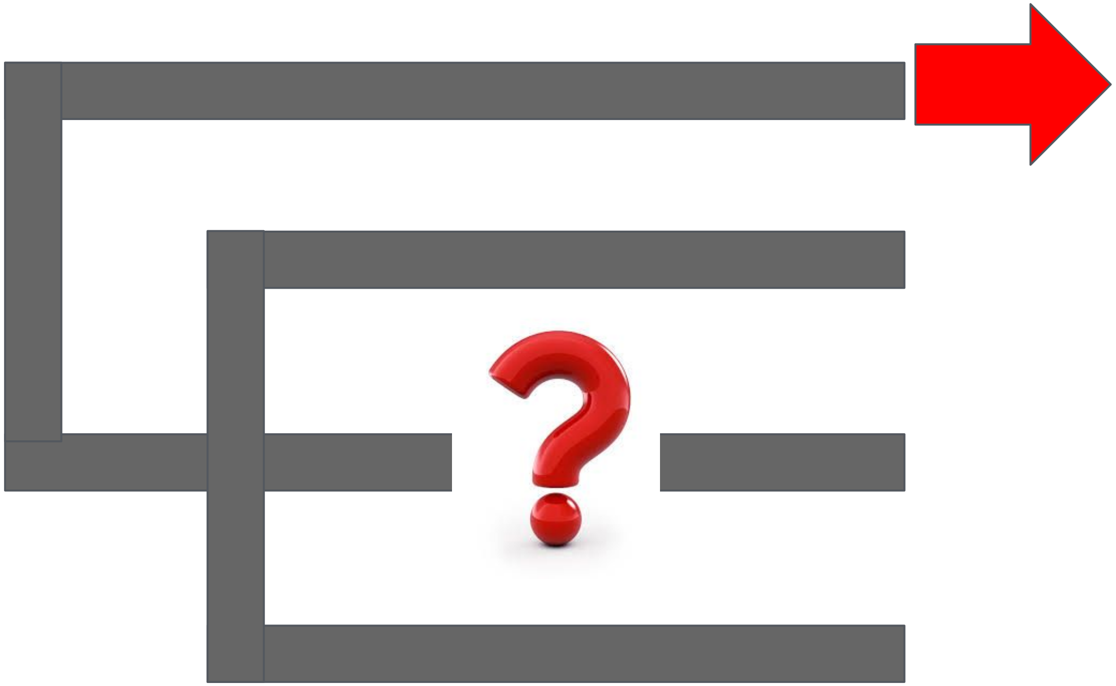
Noah
Oskar
Summer
Montana

Joseph
Margaret
Maripepa
Maria

Amparo
Paco
Adelaide
Margo

Generating phylogenomic data matrices: hands-on session

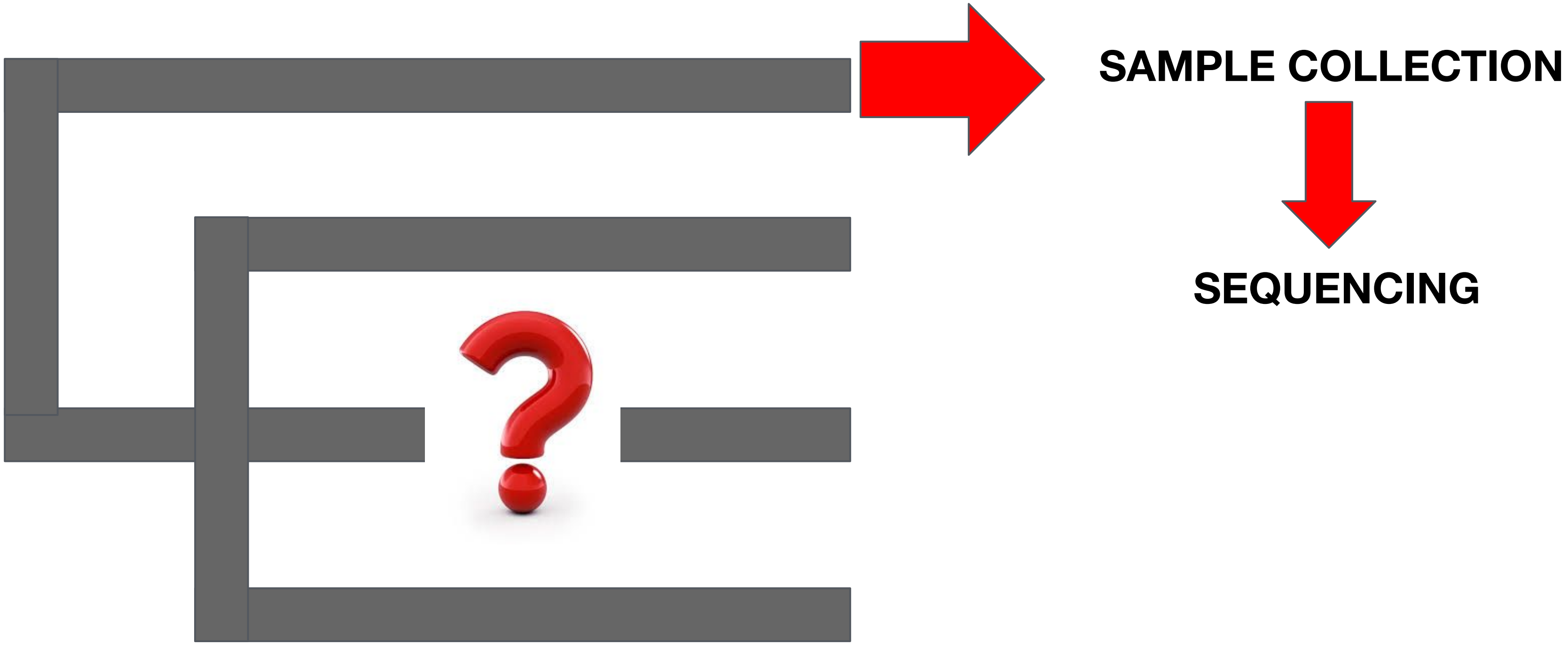
Is the polar bear the sister group to the American black bear or the brown bear?



SAMPLE COLLECTION

Generating phylogenomic data matrices: hands-on session

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

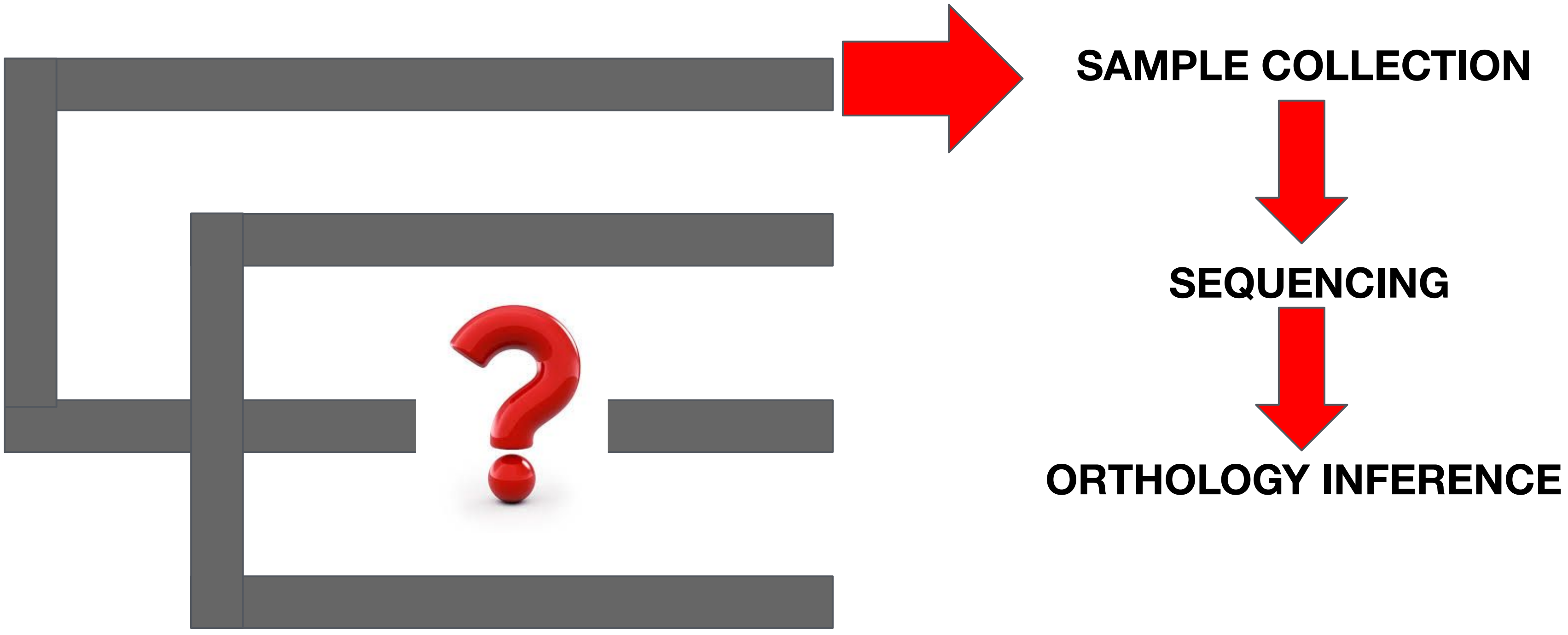


SAMPLE COLLECTION

SEQUENCING

Generating phylogenomic data matrices: hands-on session

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



SAMPLE COLLECTION

SEQUENCING

ORTHOLOGY INFERENCE

Generating phylogenomic data matrices: hands-on session

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

- **ORTHOLOGY INFERENCE**

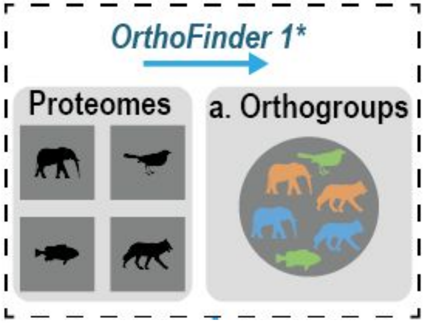
1) Let’s infer orthologous groups (OGs) in our bear dataset. We will use a tool called **OrthoFinder**.

Generating phylogenomic data matrices: hands-on session

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

- **ORTHOLOGY INFERENCE**

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

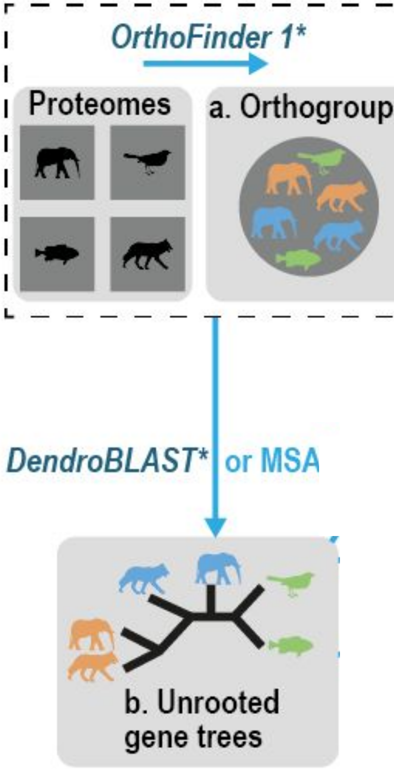


Generating phylogenomic data matrices: hands-on session

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

- **ORTHOLOGY INFERENCE**

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

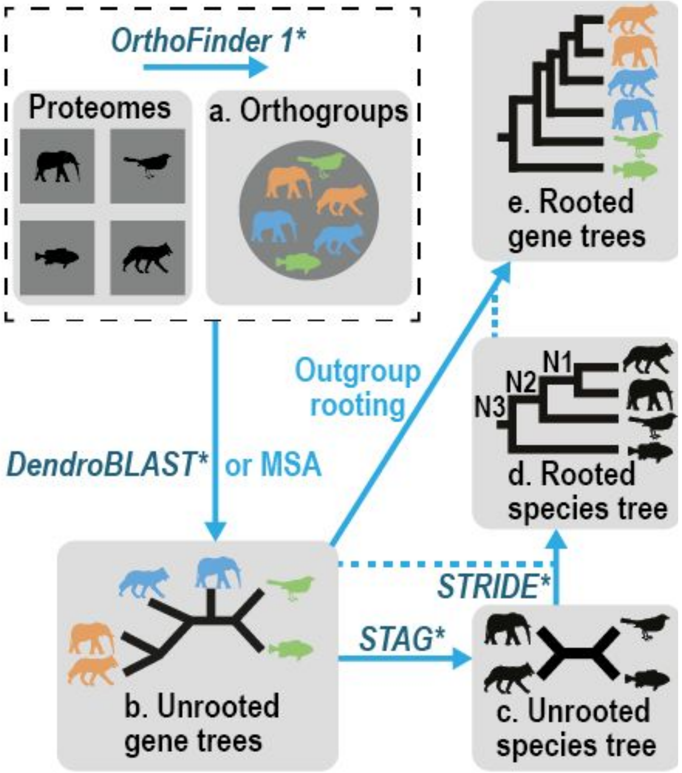


Generating phylogenomic data matrices: hands-on session

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

- **ORTHOLOGY INFERENCE**

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

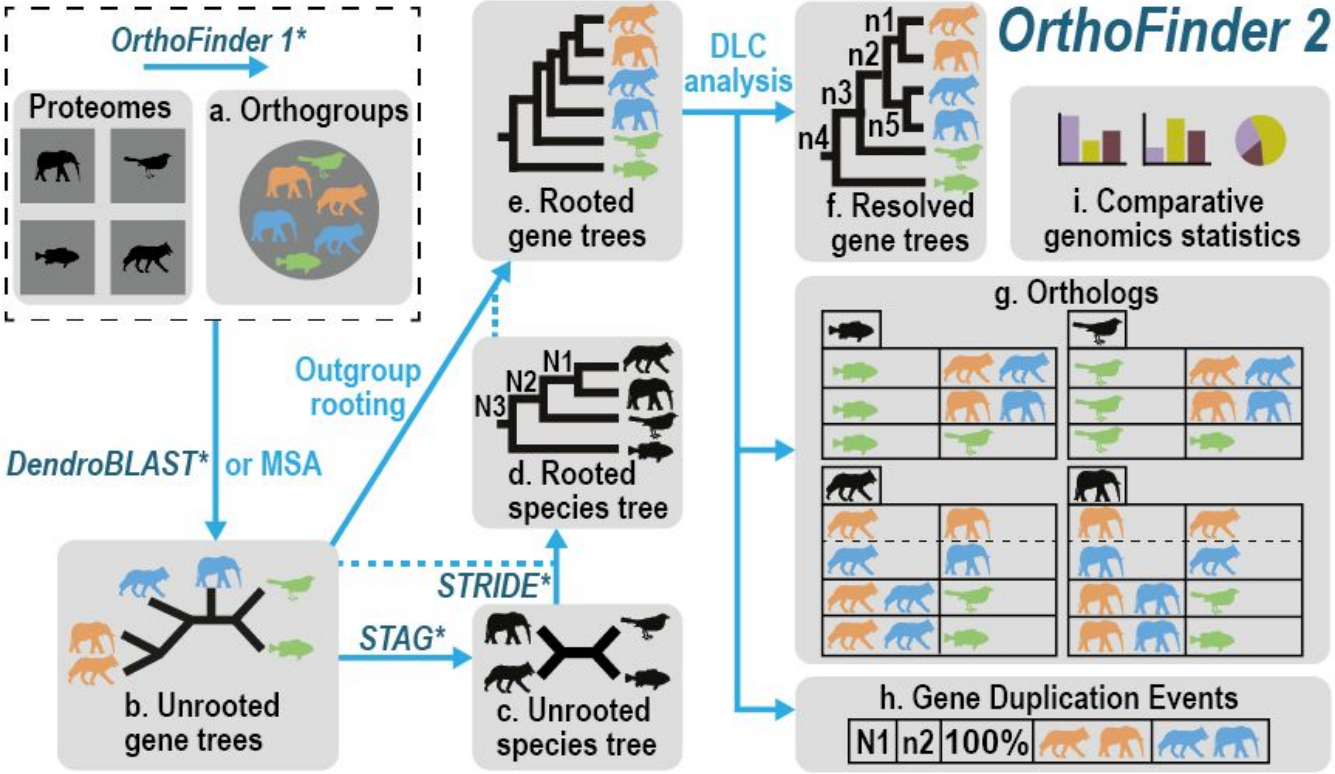


Generating phylogenomic data matrices: hands-on session

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

- **ORTHOLOGY INFERENCE**

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



Generating phylogenomic data matrices: hands-on session

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

- **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

Generating phylogenomic data matrices: hands-on session

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

- **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
```

Lert's run OrthoFinder:

```
conda activate orthofinder
```

Generating phylogenomic data matrices: hands-on session

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

- **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
```

Lert’s run OrthoFinder:

```
conda activate orthofinder

orthofinder -f ORTHOLOGY_INFERENCE/
```

Generating phylogenomic data matrices: hands-on session

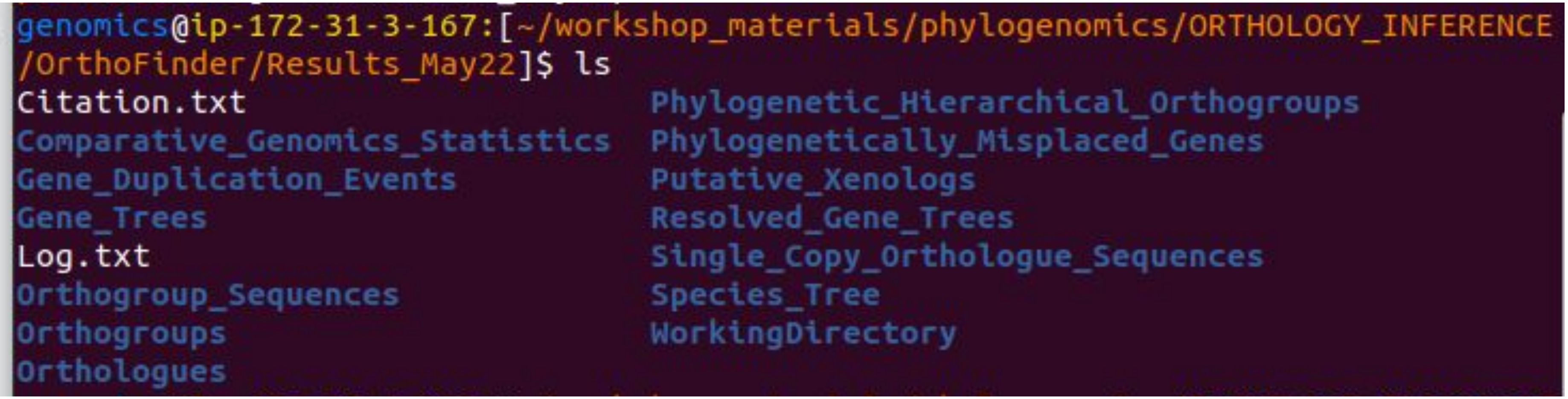
Is the polar bear the sister group to the American black bear or the brown bear?

- **ORTHOLOGY INFERENCE**

2) Let's inspect the output of OrthoFinder.

The results are in the folder
ORTHOLOGY_INFERENCE/OrthoFinder/Results_May22

Some important files and folders that you may want to check:



```
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
```

Generating phylogenomic data matrices: hands-on session

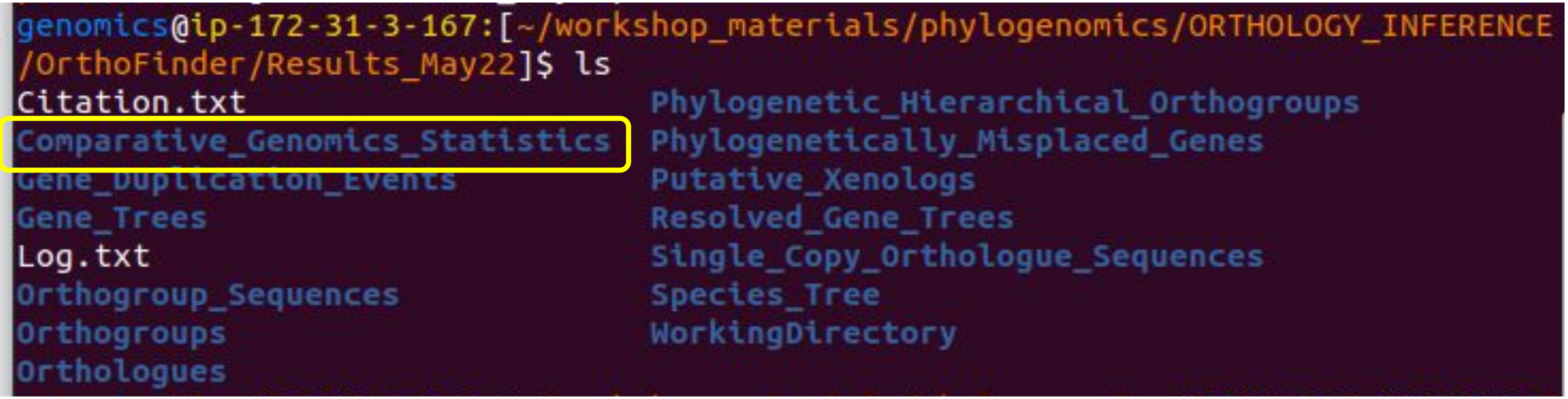
Is the polar bear the sister group to the American black bear or the brown bear?

- **ORTHOLOGY INFERENCE**

2) Let’s inspect the output of OrthoFinder.

The results are in the folder
ORTHOLOGY_INFERENCE/OrthoFinder/Results_May22

Some important files and folders that you may want to check:



```
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
```

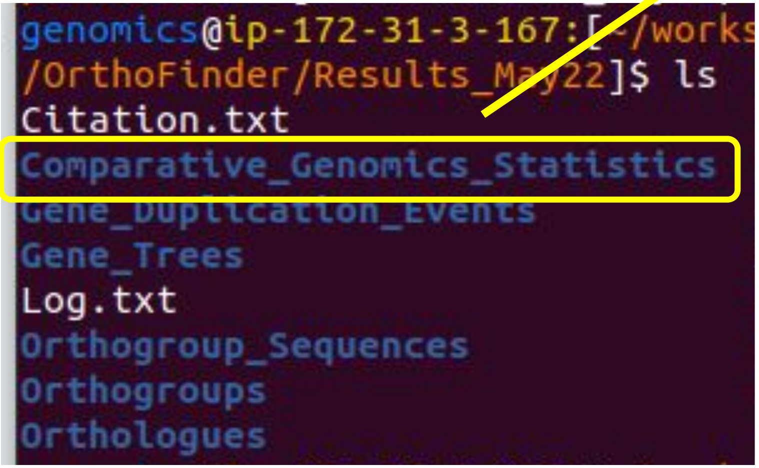
Generating phylogenomic data matrices: hands-on session

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

- **ORTHOLOGY INFERENCE**

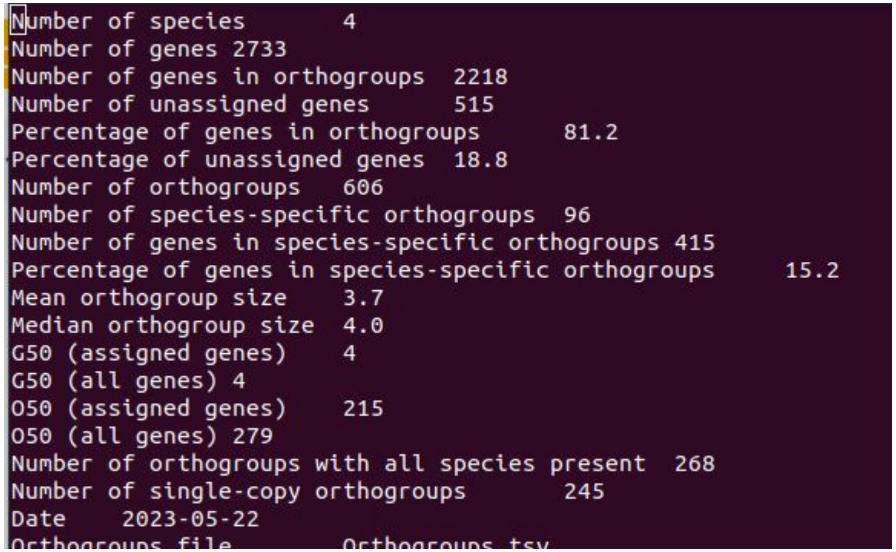
2) Let’s inspect the output of OrthoFinder.

The results are in the folder
ORTHOLOGY_INFERENCE/OrthoFinder/Results_May22



```
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
```

Generating phylogenomic data matrices: hands-on session

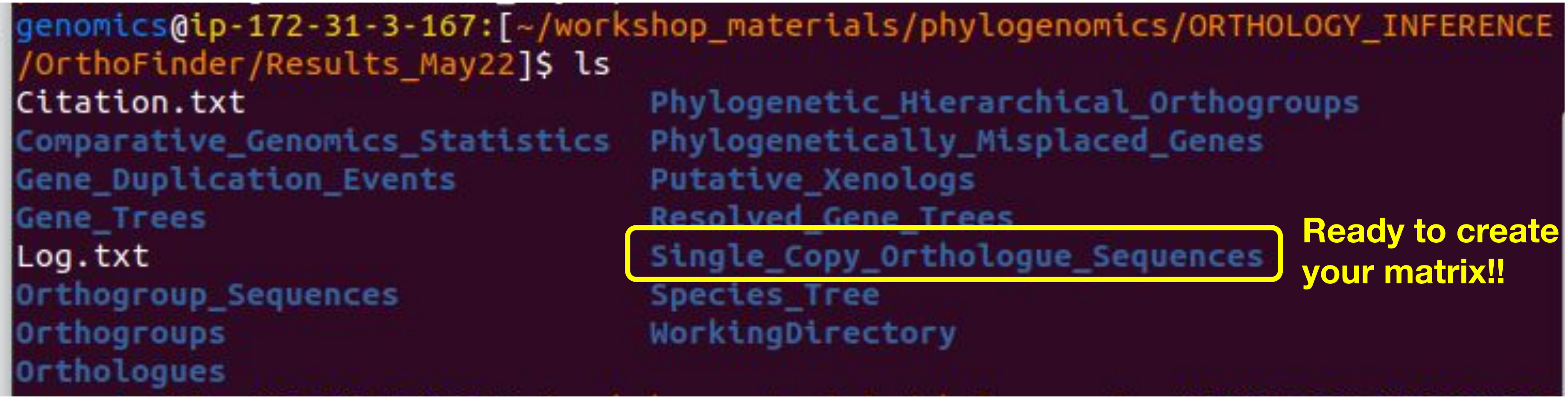
Is the polar bear the sister group to the American black bear or the brown bear?

- **ORTHOLOGY INFERENCE**

2) Let’s inspect the output of OrthoFinder.

The results are in the folder
ORTHOLOGY_INFERENCE/OrthoFinder/Results_May22

Some important files and folders that you may want to check:



```
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
```

Ready to create 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?

- **ORTHOLOGY INFERENCE**

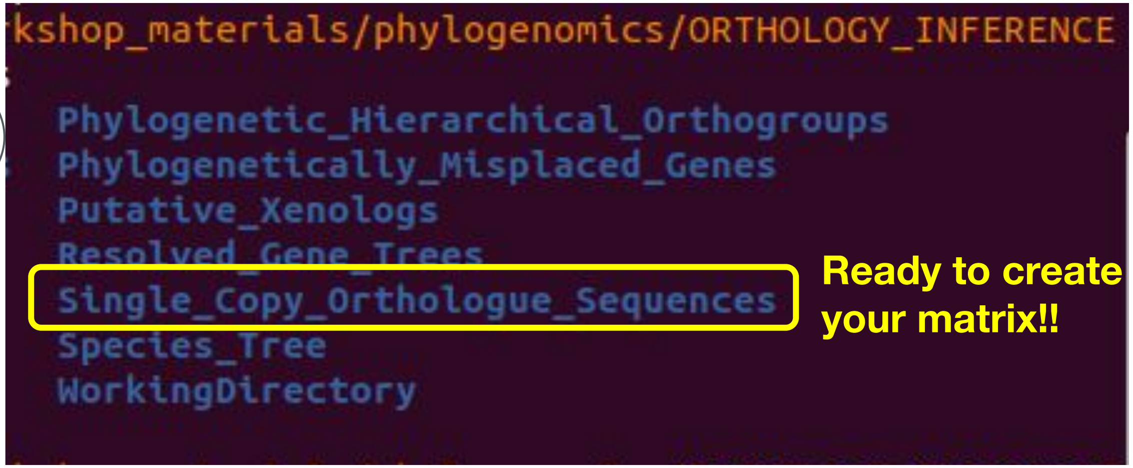
2) Let's inspect the output of **OrthoFinder**.

The results are in the folder
ORTHOLOGY_INFERENCE/OrthoFinder/Results_May22

Some important files and folders that you may want to check:

'Real world is horrible'

(Rob Waterhouse, 21st May 2023)



```
kshop_materials/phylogenomics/ORTHOLOGY_INFERENCE

  Phylogenetic_Hierarchical_Orthogroups
  Phylogenetically_Misplaced_Genes
  Putative_Xenologs
  Resolved_Gene_Trees
  Single_Copy_Orthologue_Sequences
  Species_Tree
  WorkingDirectory
```

Ready to create 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?

- **ORTHOLOGY INFERENCE**

2) Let’s inspect the output of OrthoFinder.

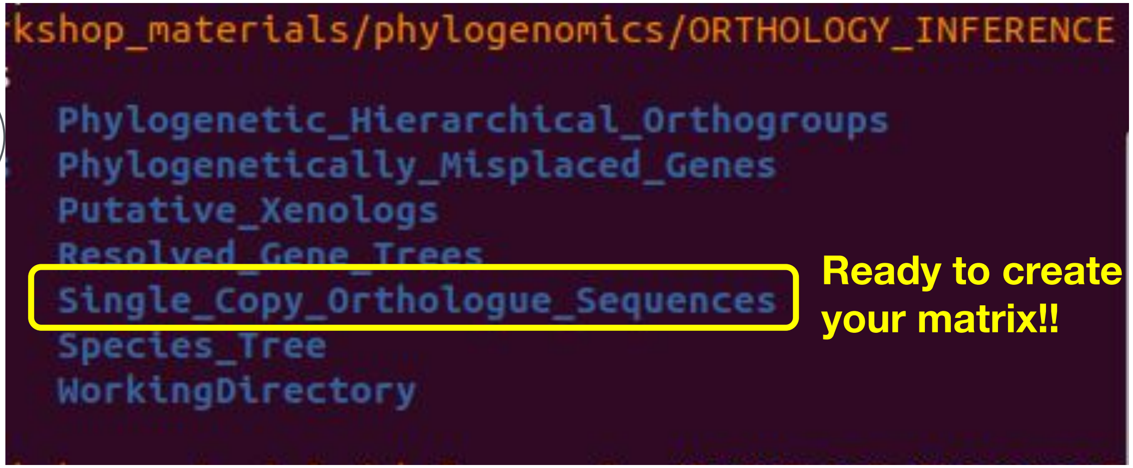
The results are in the folder
ORTHOLOGY_INFERENCE/OrthoFinder/Results_May22

Some important files and folders that you may want to check:

‘Real world is horrible’

(Rob Waterhouse, 21st May 2023)

The more species you have (and the more divergent), the less single copy OGs



```
kshop_materials/phylogenomics/ORTHOLOGY_INFERENCE

  Phylogenetic_Hierarchical_Orthogroups
  Phylogenetically_Misplaced_Genes
  Putative_Xenologs
  Resolved_Gene_Trees
  Single_Copy_Orthologue_Sequences
  Species_Tree
  WorkingDirectory
```

Ready to create your matrix!!

Generating phylogenomic data matrices: hands-on session

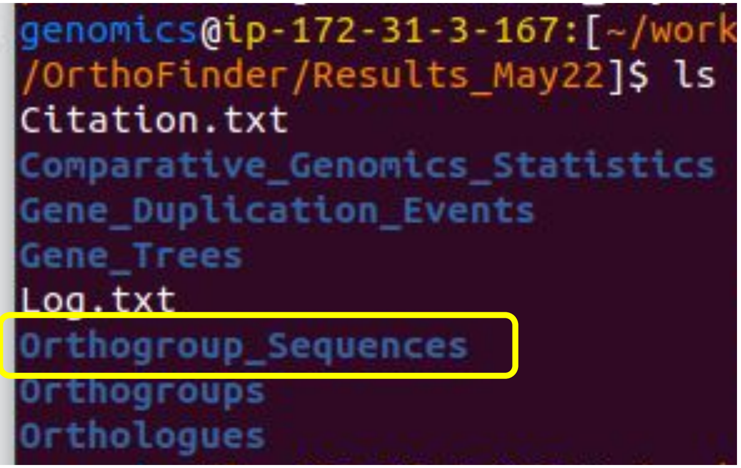
Is the polar bear the sister group to the American black bear or the brown bear?

- **ORTHOLOGY INFERENCE**

2) Let’s inspect the output of OrthoFinder.

You can prune paralogs and get subsets of OGs that resemble single copy ones -> **adequate for species tree inference!!**

PhyloPyPruner



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

Generating phylogenomic data matrices: hands-on session

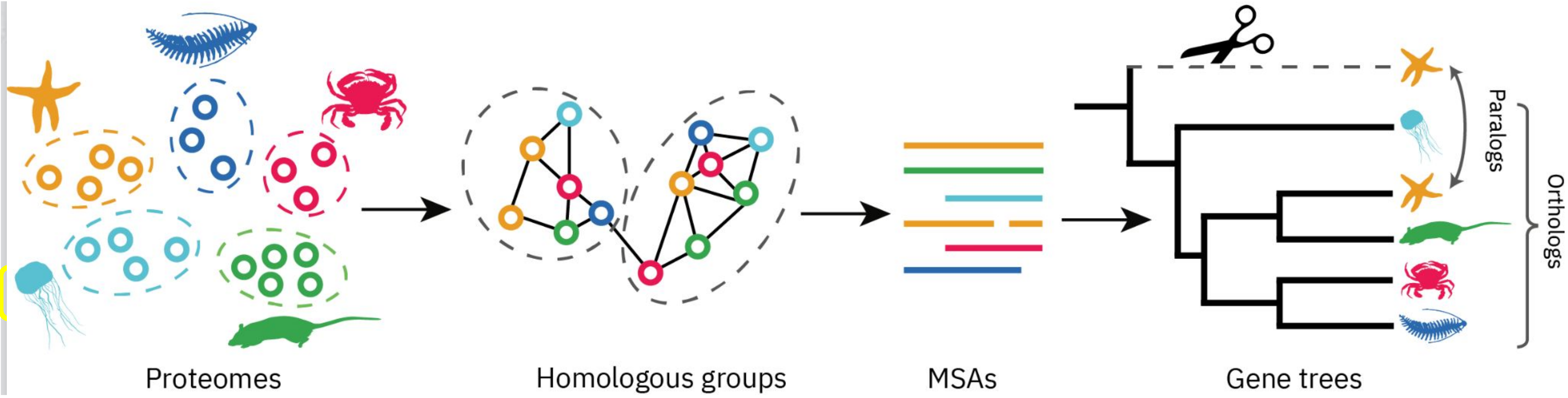
Is the polar bear the sister group to the American black bear or the brown bear?

- **ORTHOLOGY INFERENCE**

2) Let’s inspect the output of OrthoFinder.

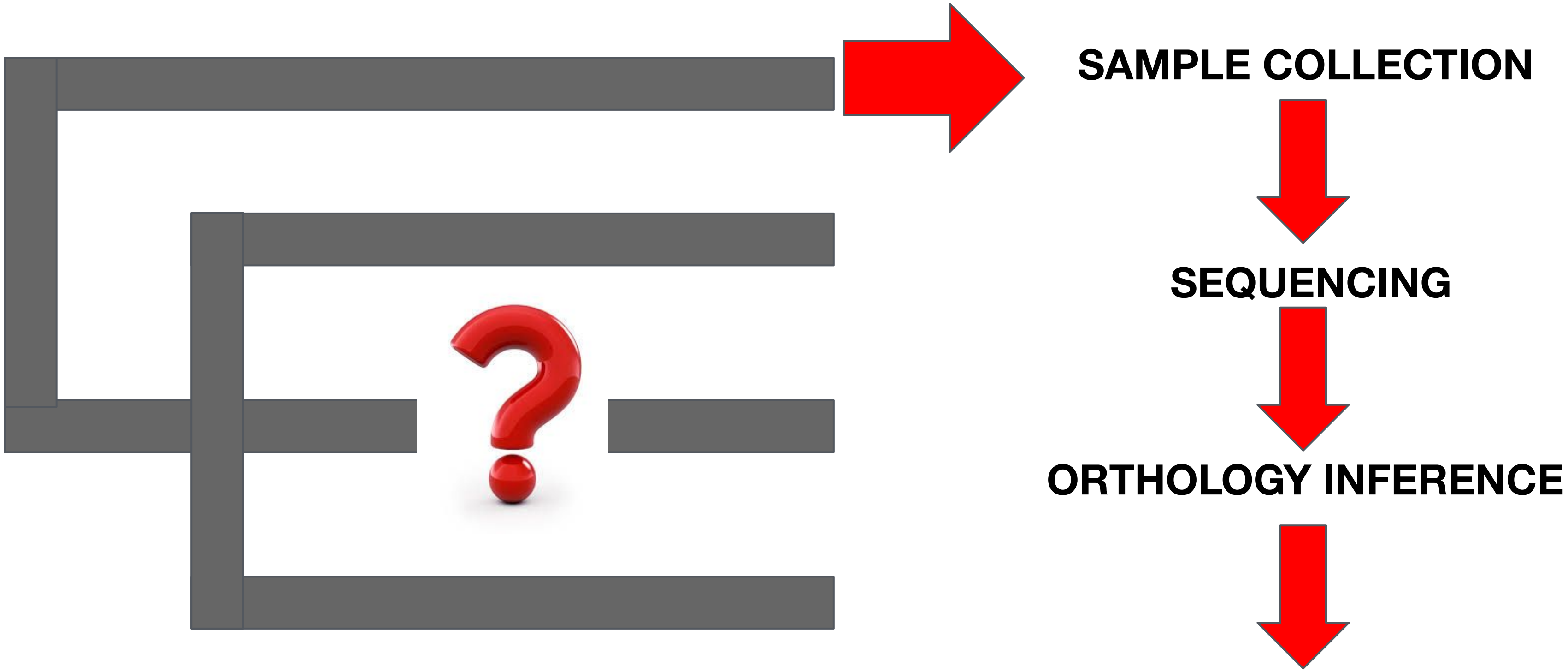
You can prune paralogs and get subsets of OGs that resemble single copy ones -> **adequate for species tree inference!!**

PhyloPyPruner



Generating phylogenomic data matrices: hands-on session

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??

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 sample 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 sample occupancy** to account for the effect of missing data.

1) The data is located in the folder **phylogenomics/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 sample occupancy** to account for the effect of missing data.

1) The data is located in the folder **phylogenomics/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 sample occupancy** to account for the effect of missing data.

1) The data is located in the folder **phylogenomics/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').

3) There are also 3 python scripts. For them to run, we'll need the python libraries **numpy** and **cogent** (already installed in the AMI).

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:

```
conda deactivate

conda activate cogent

python count_genesPerSample.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**

python count_genesPerSample.py

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

Sample	No. OGs	No. total OGs	Proportion of total OGs
UrsusAmericanus_Montana	44	50	0.88
Ailuropoda_Siro	47	50	0.94
UrsusAmericanus_Noah	41	50	0.82
UrsusMaritimus_Joseph	12	50	0.24
UrsusMaritimus_Maria	48	50	0.96
UrsusMaritimus_Maripepa	47	50	0.94
Ailuropoda_Luisa	22	50	0.44

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 run a tree with all the data (we know that we are going to have missing data in our matrix).

From folder ‘MISSING_DATA’, let’s first align each gene, create a matrix with all the genes (concatenation) and then run a tree.

To align the genes:

```
for i in *
do
muscle -align $i -output $i.aln
done
```

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 concatenate the genes and generate the matrix, let’s use a software called **catsequences**:

```
ls *aln > list_all_genes.txt
catsequences list_all_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**

To infer the ML tree, let's try IQTREE2 (very interesting as it allows to run some complex mixture models).

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

(if you don't specify the model it will do model testing, but it takes a while, so feel free to try it at home)

You can see the best-fitted model in the file **allseqs.fas.iqtree,** and the maximum likelihood tree in the file **allseqs.fas.treefile.** You can visualize it in iTOL (just copy-paste the tree in the web server where it says 'Upload Tree').

Which topology is this matrix supporting?

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) 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). Let’s run the script:

```
conda deactivate
python2 select_sample_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**

6) 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 samples/specimens per species). Let's run the script:

```
conda deactivate
python2 select_sample_occupancy.py
```

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

6) 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 samples/specimens per species). Let's run the script:

```
conda deactivate
python2 select_sample_occupancy.py
```

It will ask you to select the minimum sample occupancy. Let's start by 3. 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.

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 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.

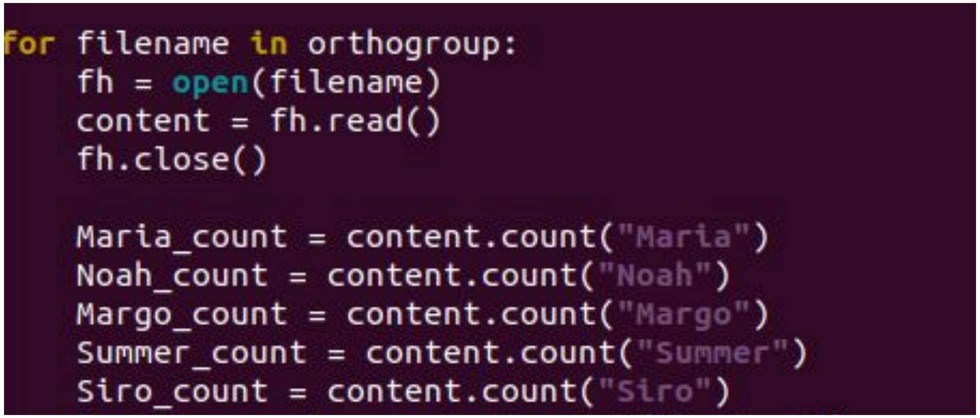
Let’s then select genes that have an homogeneous representation of all the four species.

Generating phylogenomic data matrices: hands-on session

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

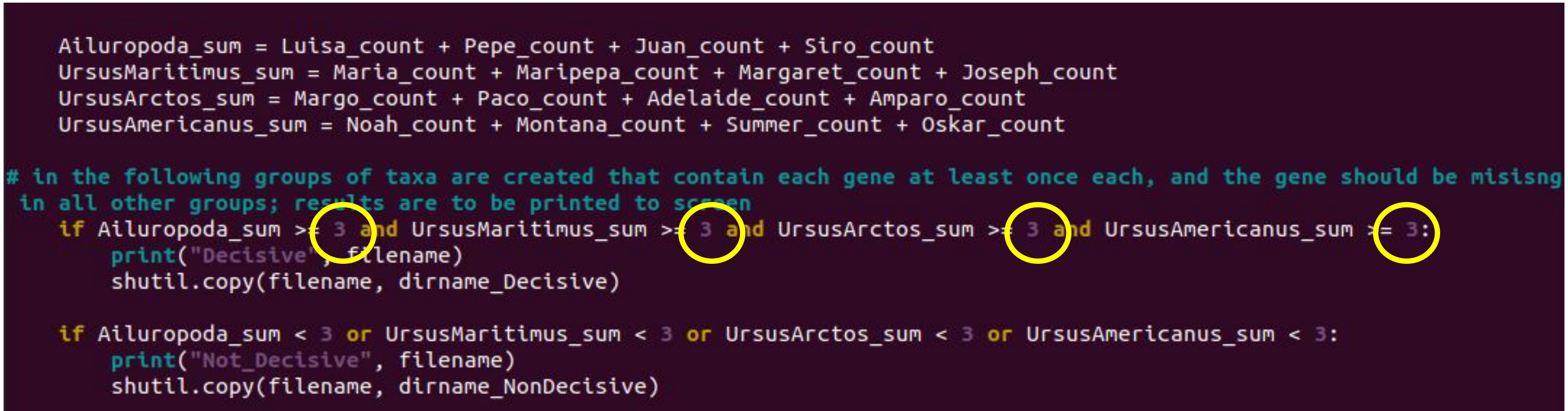
- **MISSING DATA**

8) Let’s open the **decisive_genes.py** script and inspect it together.



```
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)
```

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) 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**

9) 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.

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**

9) 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.

10) Now (or at the end of the class) 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.

Is missing data affecting the topology of your Maximum Likelihood tree?

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) 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**

11) 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**

11) 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**

12) 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**

12) 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:

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 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**

13) 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)

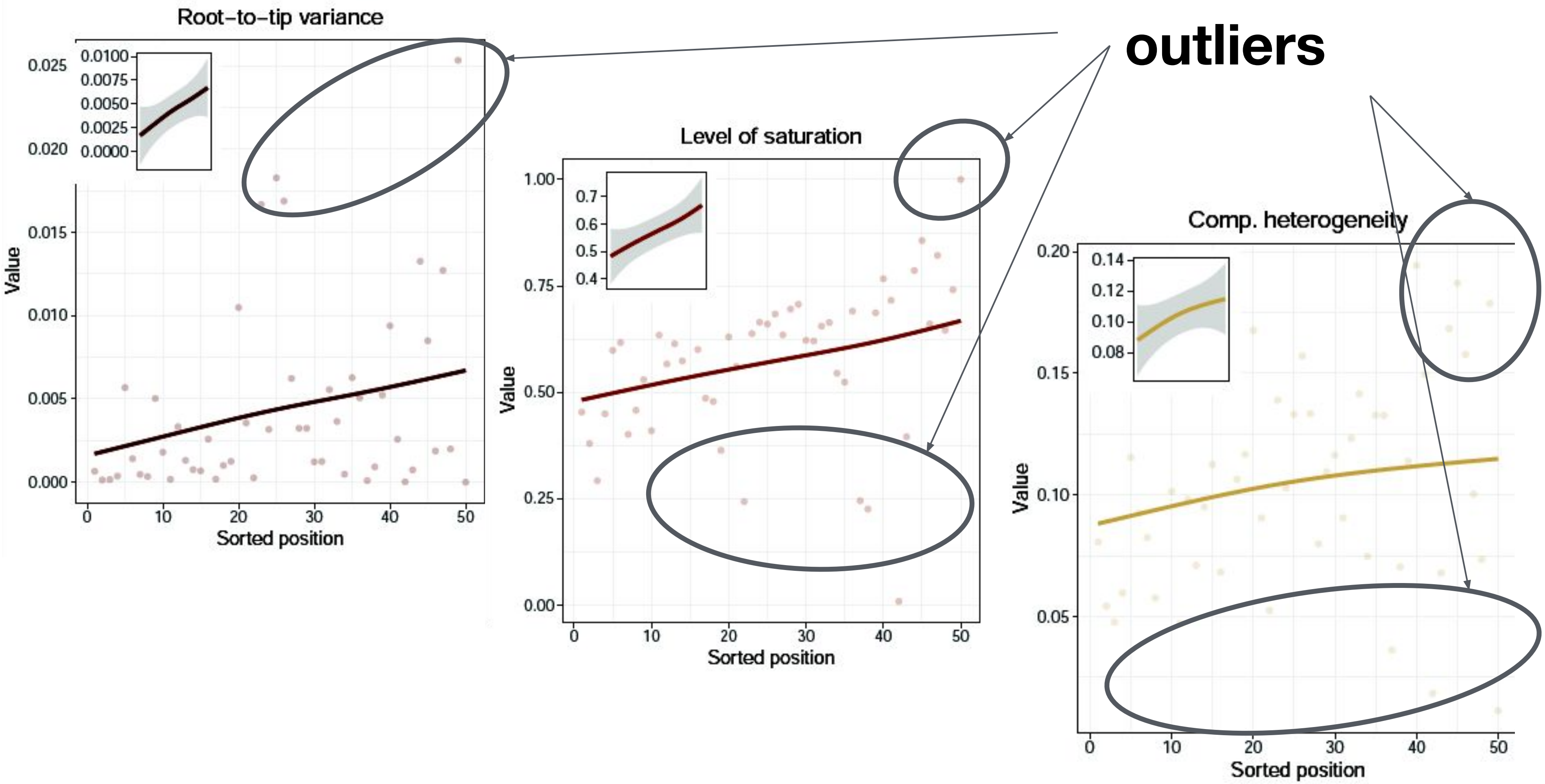
We'll obtain a copy of our concatenated alignment, partition file and gene trees sorted by their phylogenetic usefulness, from most to least useful.

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

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**



outliers

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**

14) 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**

14) 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 a tree and test how the phylogeny varies when using genes with different phylogenetic ‘usefulness’.

Do you see any differences in the topology?

Generating phylogenomic data matrices: hands-on session

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

SCIENCE & NATURE

Phylogenomics illuminate the interrelationships of the genus *Ursus* and supports the brown bear as sister group to the polar bear

Authors: Participants of the 2023 Workshop on Genomics český Krumlov (‘molekulos’)

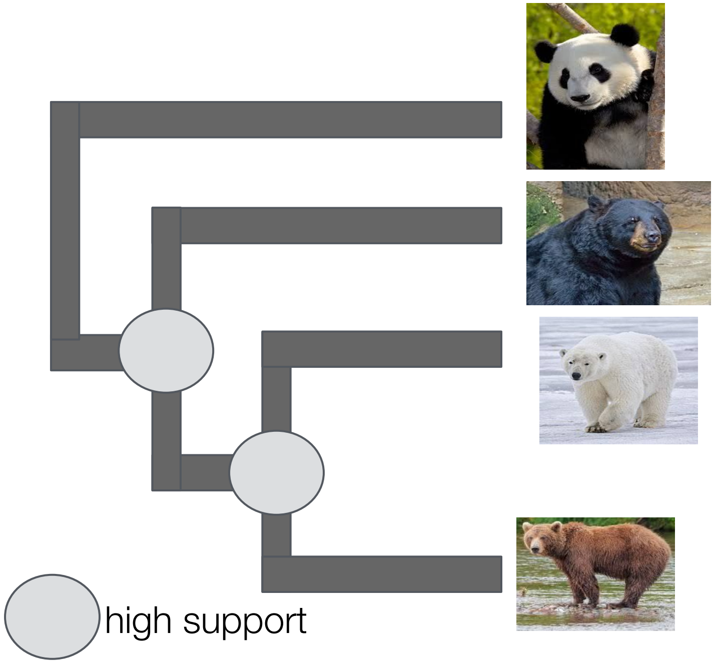
Generating phylogenomic data matrices: hands-on session

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

SCIENCE & NATURE

Phylogenomics illuminate the interrelationships of the genus *Ursus* and supports the brown bear as sister group to the polar bear

Authors: Participants of the 2023 Workshop on Genomics český Krumlov (‘molekulos’)



Abstract:

The interrelationships within the genus *Ursus* have been contentious based on the analysis of a limited amount of molecular markers. Here, we sequenced full genomes of 16 specimens of the American black bear, brown bear, polar bear and giant panda and explored their phylogenetic relationships through a phylogenomic spyglass. Our results, based on the analysis of multiple supermatrices to account for the effect of missing data, compositional heterogeneity and other confounding factors, strongly support a sister relationship of the brown bear to the polar bear. Our findings pave te road towards understanding bear evolution at a deeper level.

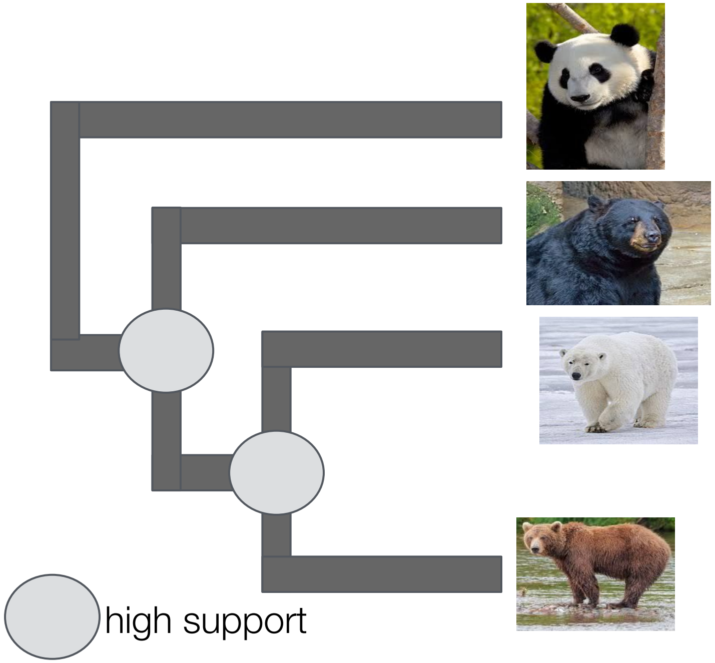
Generating phylogenomic data matrices: hands-on session

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

SCIENCE & NATURE

Phylogenomics illuminate the interrelationships of the genus *Ursus* and supports the brown bear as sister group to the polar bear

Authors: Participants of the 2023 Workshop on Genomics český Krumlov (‘molekulos’)



Abstract:

The interrela been
contentious unt of
molecular m nes of 16
specimens r, polar bear
and giant pa elationships
through a pl sed on the
analysis of multiple supermatrices to account for the effect of missing data, compositional heterogeneity and other confounding factors, strongly support a sister relationship of the brown bear to the polar bear. Our findings pave te road towards understanding bear evolution at a deeper level.

REJECTED

Generating phylogenomic data matrices: hands-on session

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

SCIENCE & NATURE

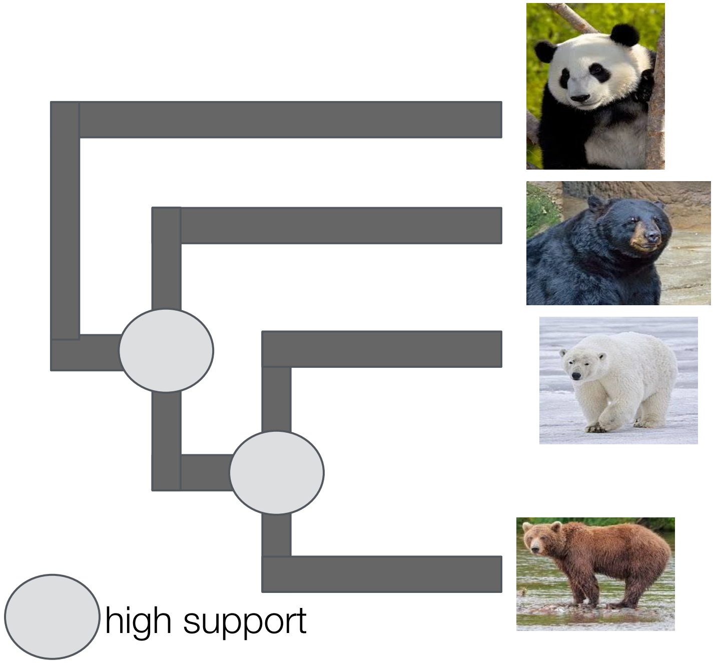
Phylogenomics illuminate the interrelationships of the genus *Ursus* and supports the brown bear as sister group to the polar bear

Authors: Participants of the 2023 Workshop on Genomics český Krumlov ('molekulos')

Abstract:

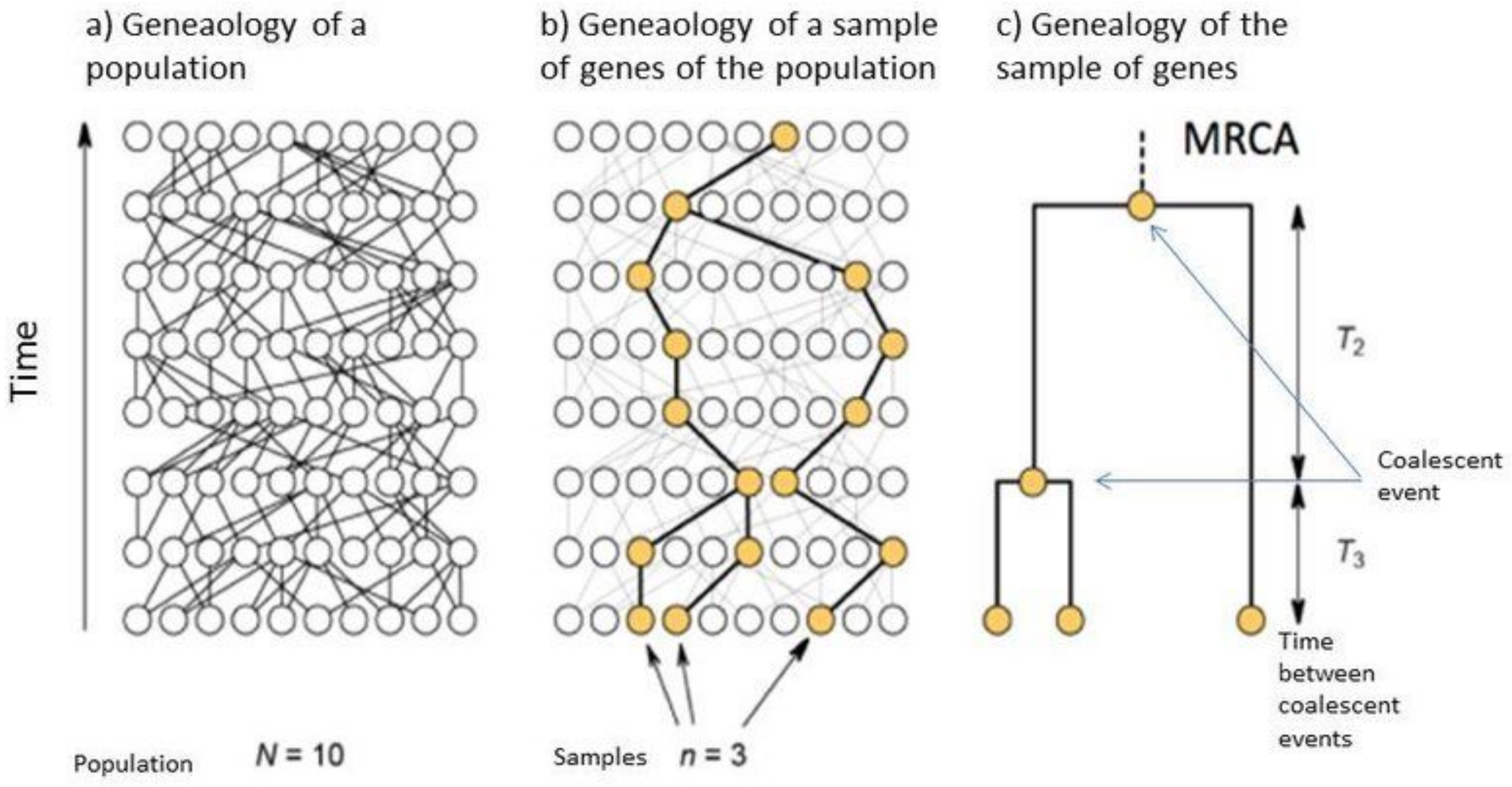
The interrela been
contentious unt of
molecular m nes of 16
specimens r, polar bear
and giant pa elationships
through a pl sed on the
analysis of multiple supermatrices to account for the effect of

REJECTED



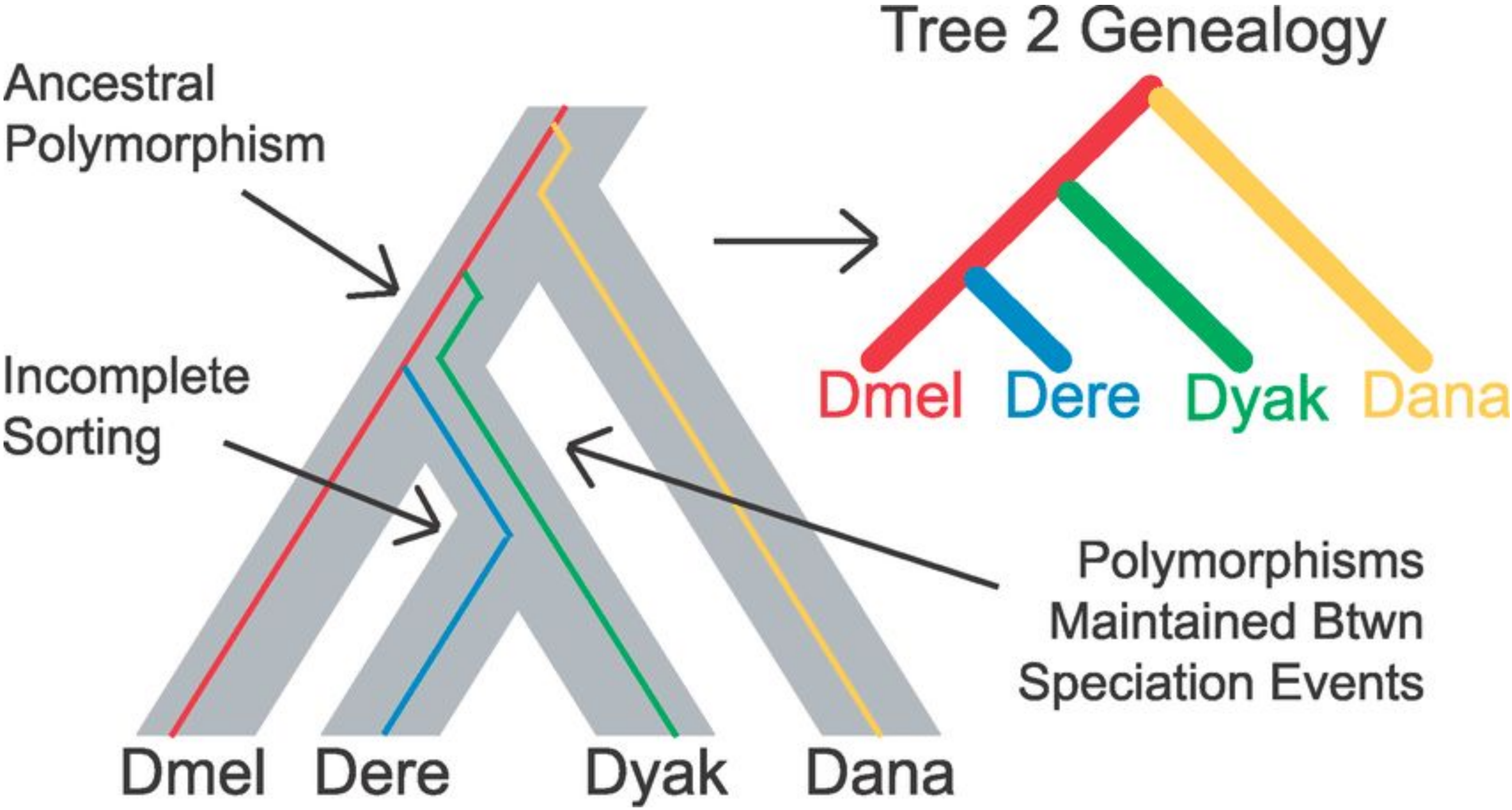
Reviewer #3: although I appreciate the efforts of the authors to account for confounding factors and test the robustness of their results, they failed to test whether their hypothesis was driven by incongruence between individual gene evolutionary trajectories.

Brief introduction to coalescent theory


a) Geneaology of a population
b) Geneaology of a sample of genes of the population
c) Genealogy of the sample of genes
Time
MRCA
T_2
T_3
Coalescent event
Time between coalescent events
Population
N = 10
Samples
n = 3

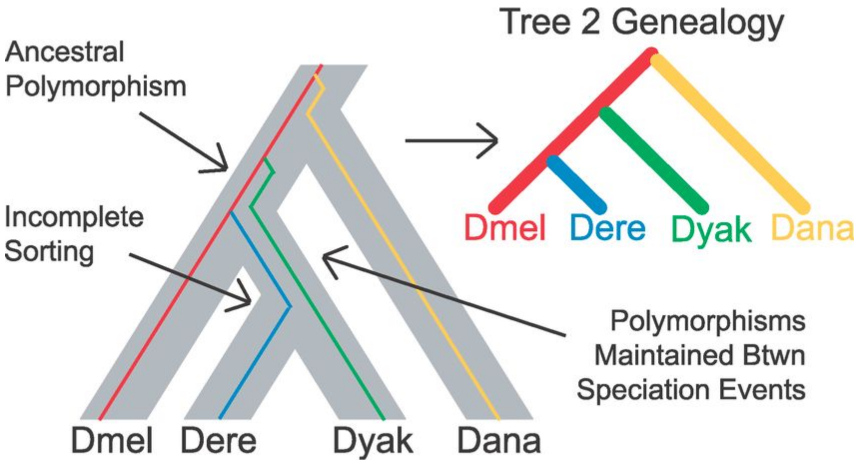
Analyzing gene tree/species tree conflict: hands-on session

Brief introduction to coalescent theory

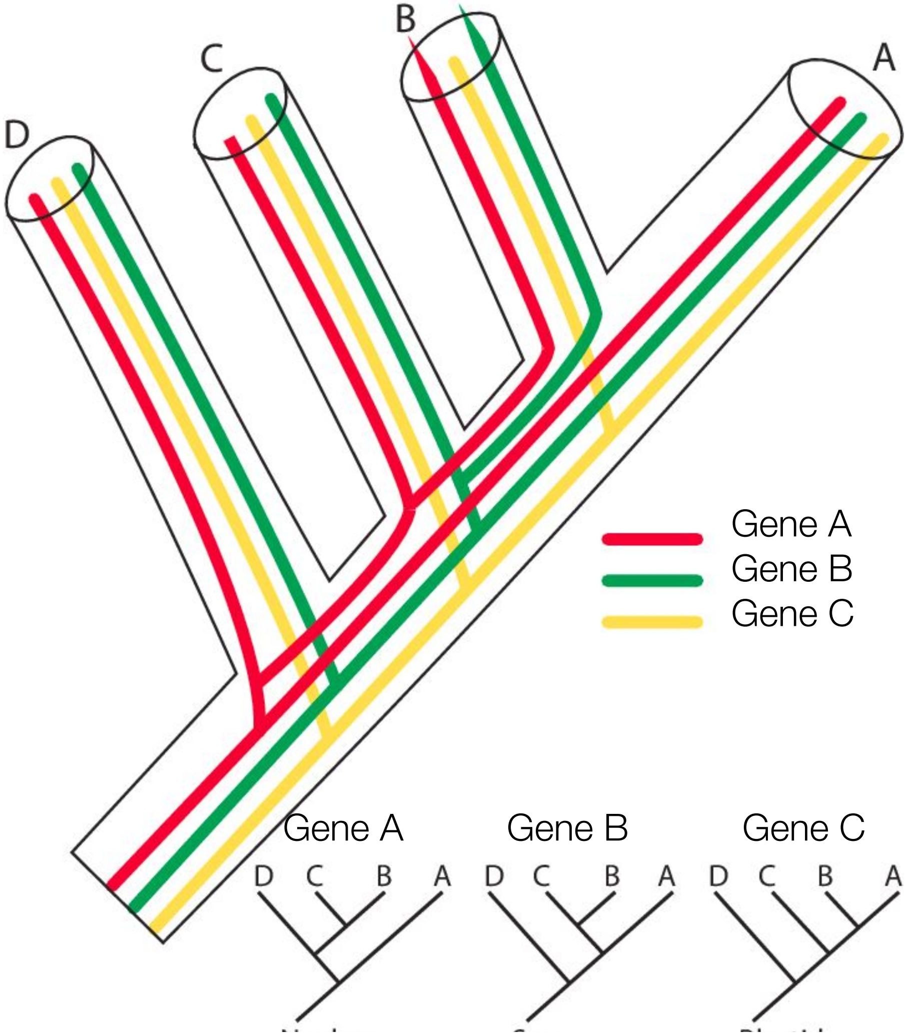


Pollard et al. 2006

Brief introduction to coalescent theory



Pollard et al. 2006



Naciri and Li 2015

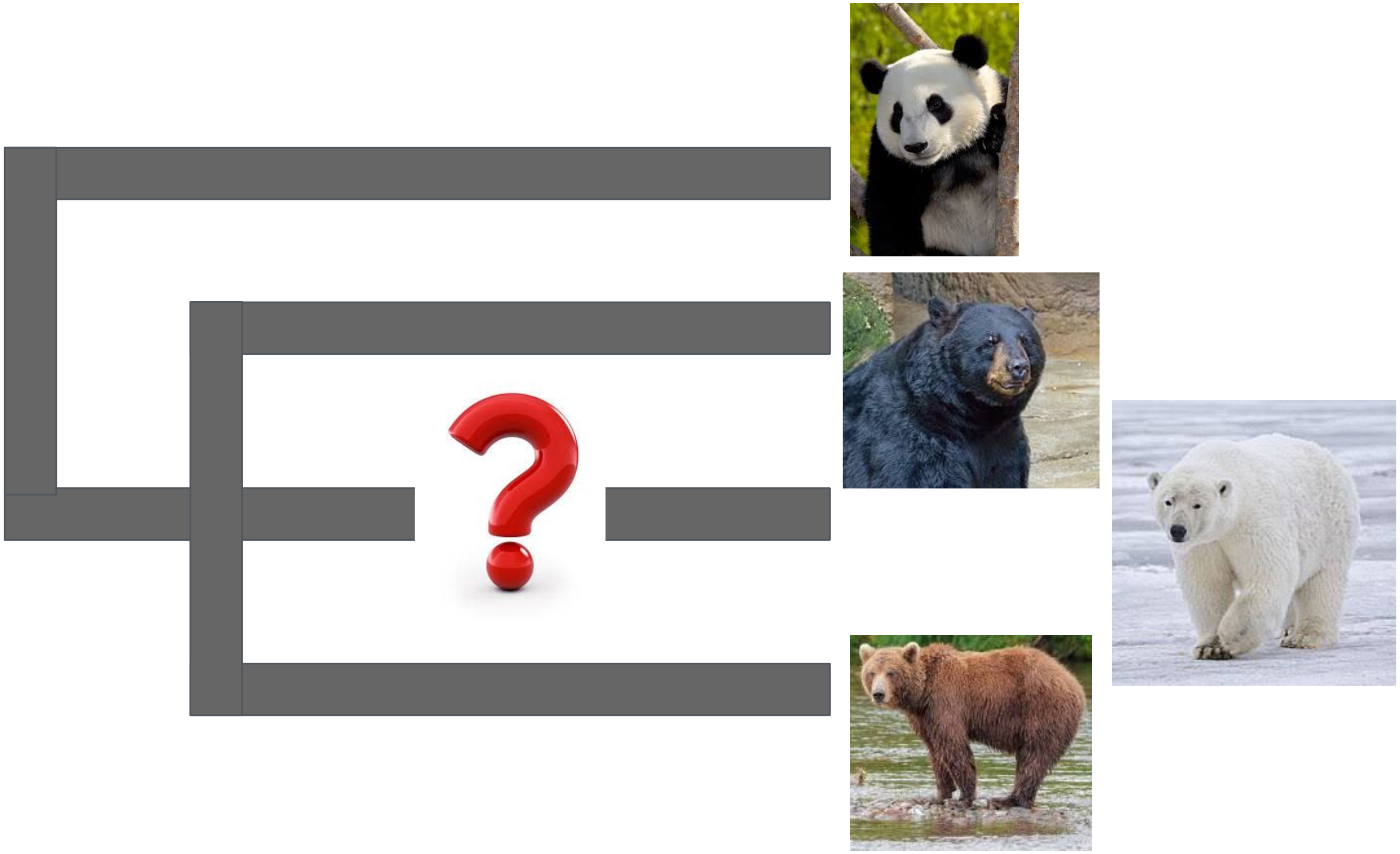
Analyzing gene tree/species tree conflict: hands-on session

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

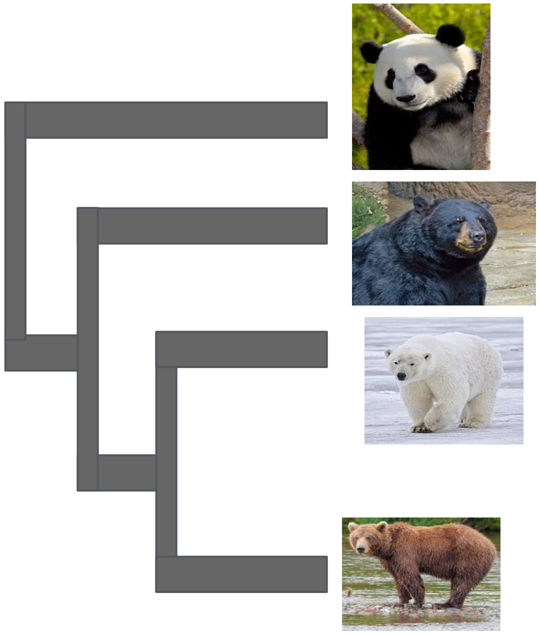
Is the polar bear the sister group to the American black bear or the brown bear?



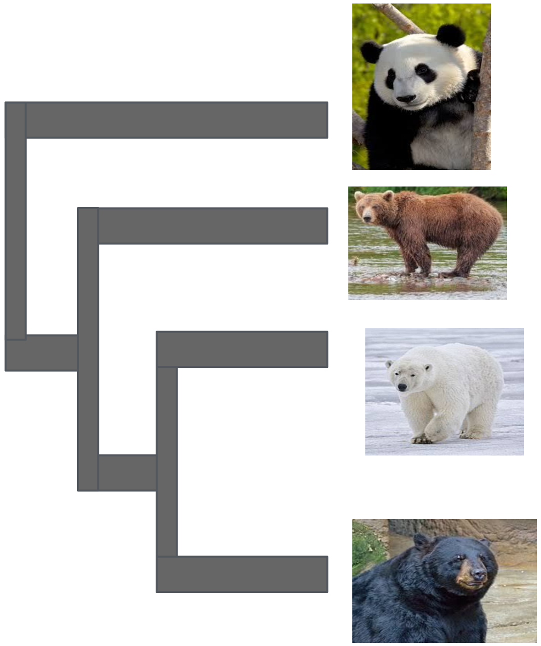
Analyzing gene tree/species tree conflict: hands-on session

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

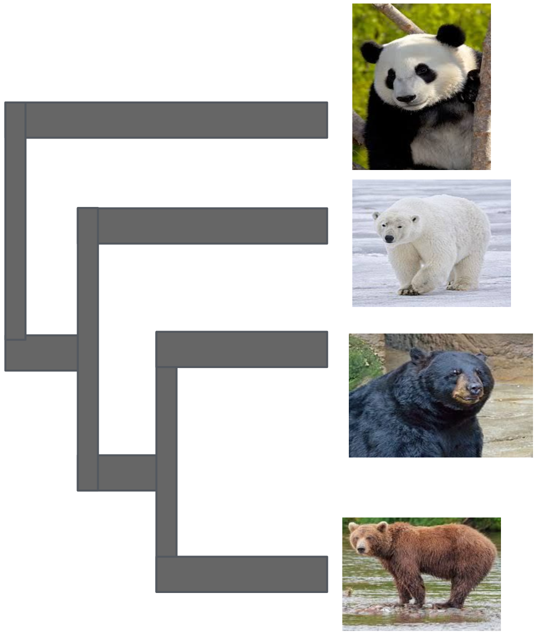
1) Let’s analyze conflict between individual gene trees to see which phylogenetic hypothesis is the most robustly supported:



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:

phylogenomics/ASTRAL (.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:

phylogenomics/ASTRAL (.tree files)

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

```
cat *tree > 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:

```
java -jar $HOME/software/Astral/astral.5.7.8.jar -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:

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

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

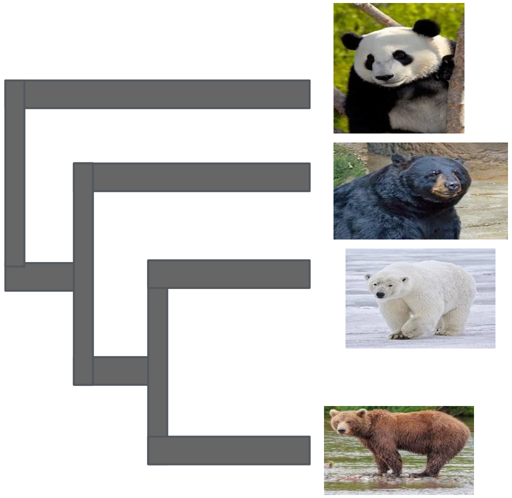
->The normalized quartet score is the proportion of input gene tree quartet trees satisfied by the species tree. This is a number between zero and one; the higher this number, the *less* discordant your gene trees are.

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

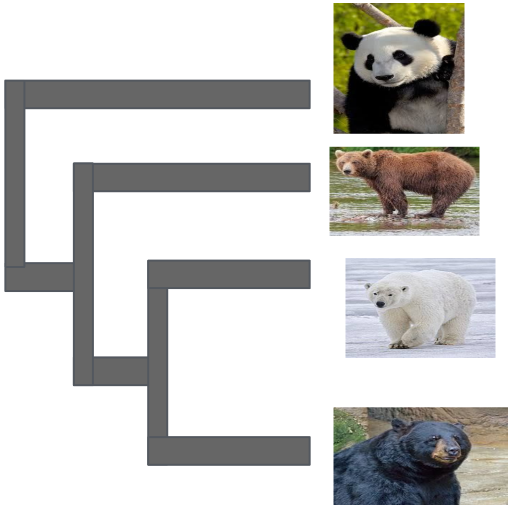
Is the polar bear the sister group to the American black bear or the brown bear?

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

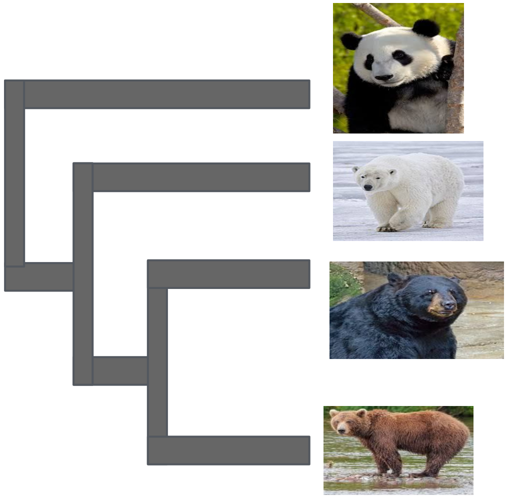
For that, let’s score each species tree topology and compare the normalize quartet score for each one.



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 togology. 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.

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.

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:

```
java -jar $HOME/software/Astral/astral.5.7.8.jar -i
bears_allTrees.tre -q 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:

```
java -jar $HOME/software/Astral/astral.5.7.8.jar -i
bears_allTrees.tre -q 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:

```
java -jar $HOME/software/Astral/astral.5.7.8.jar -i
bears_allTrees.tre -q 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*?

Generating phylogenomic data matrices: hands-on session

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

SCIENCE & NATURE

Phylogenomics illuminate the interrelationships of the genus *Ursus* and supports the brown bear as sister group to the polar bear

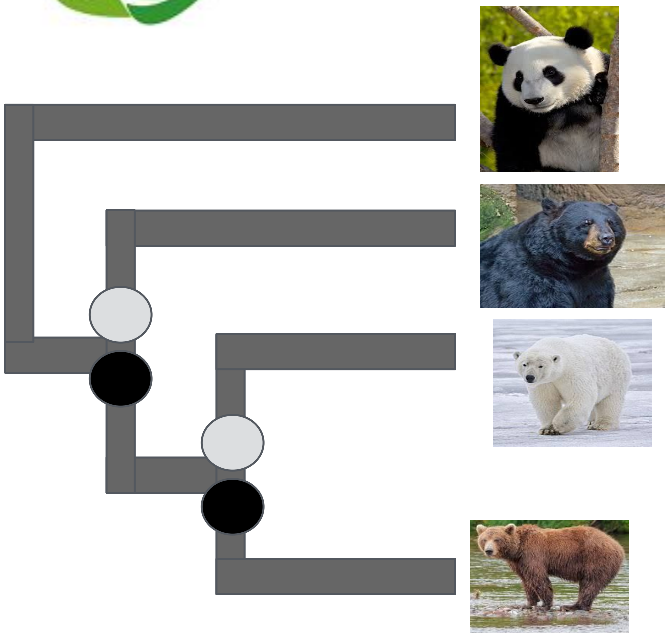
Authors: Participants of the 2023 Workshop on Genomics český Krumlov (‘molekulos’)

Generating phylogenomic data matrices: hands-on session

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

Phylogenomics illuminate the interrelationships of the genus *Ursus* and supports the brown bear as sister group to the polar bear

Authors: Participants of the 2023 Workshop on Genomics český Krumlov (‘molekulos’)



○ high support supermatrix

● high support indiv. gene trees (multispecies coalescent)

Abstract:

The interrelationships of the species within the genus *Ursus* has been contentious based on the analysis of a limited amount of molecular markers. Here, we sequenced full genomes of 16 specimens of the American black bear, brown bear, polar bear and giant panda and explored their phylogenetic relationships through a phylogenomic spyglass. Our results, based on the analysis of multiple supermatrices to account for the effect of missing data, compositional heterogeneity and other confounding factors, **as well as accounting for incongruence between individual gene trees under the multispecies coalescent model**, strongly support a sister relationship of the brown bear to the polar bear. Our findings pave te road towards understanding bear evolution at a deeper level.

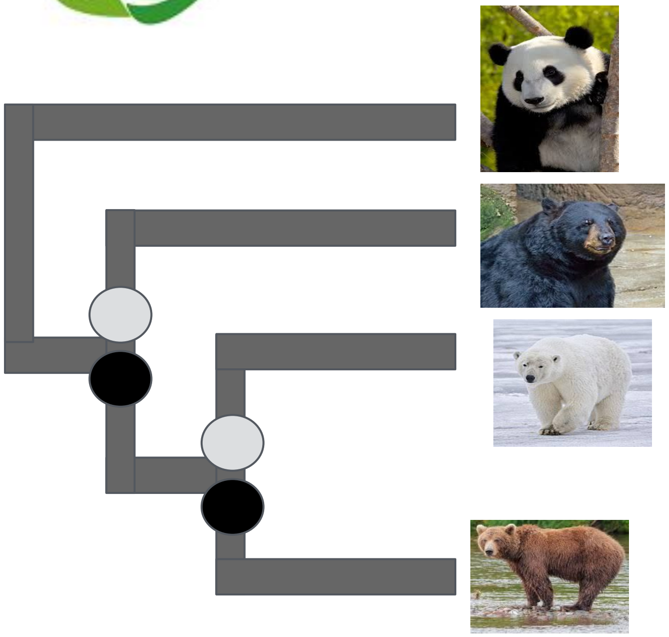
Generating phylogenomic data matrices: hands-on session

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

SCIENCE & NATURE

Phylogenomics illuminate the interrelationships of the genus *Ursus* and supports the brown bear as sister group to the polar bear

Authors: Participants of the 2023 Workshop on Genomics český Krumlov ('molekulos')



○ high support supermatrix

● high support indiv. gene trees (multispecies coalescent)

Abstract:

The int nus *Ursus*
has be imited
amour ed full
genom bear, brown
bear, p eir
phylog c spyglass.
Our res ermatrices to
accour nal
heterog **well as**
accou lual gene
trees under the multispecies coalescent model, strongly support a sister relationship of the brown bear to the polar bear. Our findings pave te road towards understanding bear evolution at a deeper level.

ACCEPTED