Phylogenomics and Population Genomics:
Inference and Applications

ORTHOLOGY PREDICTION FOR PHYLOGENOMIC ANALYSES

Marina Marcet Houben
mmarcet@bsc.es
Barcelona Supercomputing Center

Before starting:

- Log in to your session

```
ssh username@ec2-34-242-61-70.eu-west-1.compute.amazonaws.com
```

- Copy the github session into your main folder:

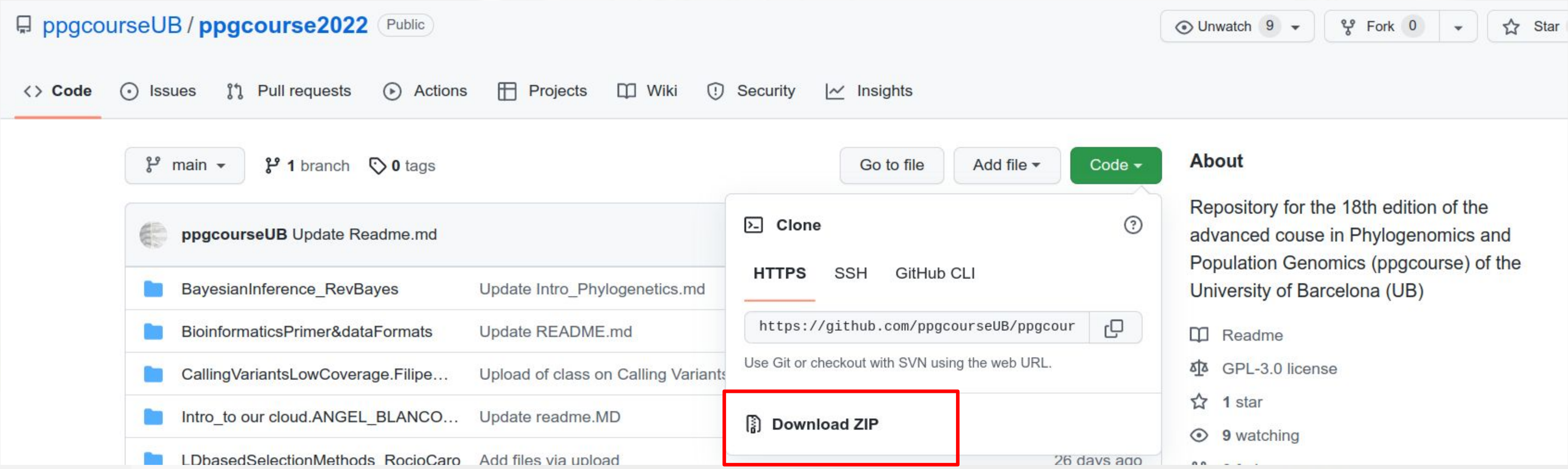
```
svn export
https://github.com/ppgcourseUB/ppgcourse2023/trunk/Orthology_prediction_for_phylogenomic_analyses.MARINA_MARCET
```

- Move into the folder of the session:

```
cd Orthology_prediction_for_phylogenomic_analyses.MARINA_MARCET/
```

Just in case the svn export does not work:

- Copy the main repository into your computer:



ppgcourseUB Update Readme.md

BayesianInference_RevBayes	Update Intro_Phylogenetics.md
BioinformaticsPrimer&dataFormats	Update README.md
CallingVariantsLowCoverage.Filipe…	Upload of class on Calling Variants
Intro_to our cloud.ANGEL_BLANCO…	Update readme.MD
LDbasedSelectionMethods_RocioCaro	Add files via upload

26 days ago

```
https://github.com/ppgcourseUB/ppgcour
```

Use Git or checkout with SVN using the web URL.

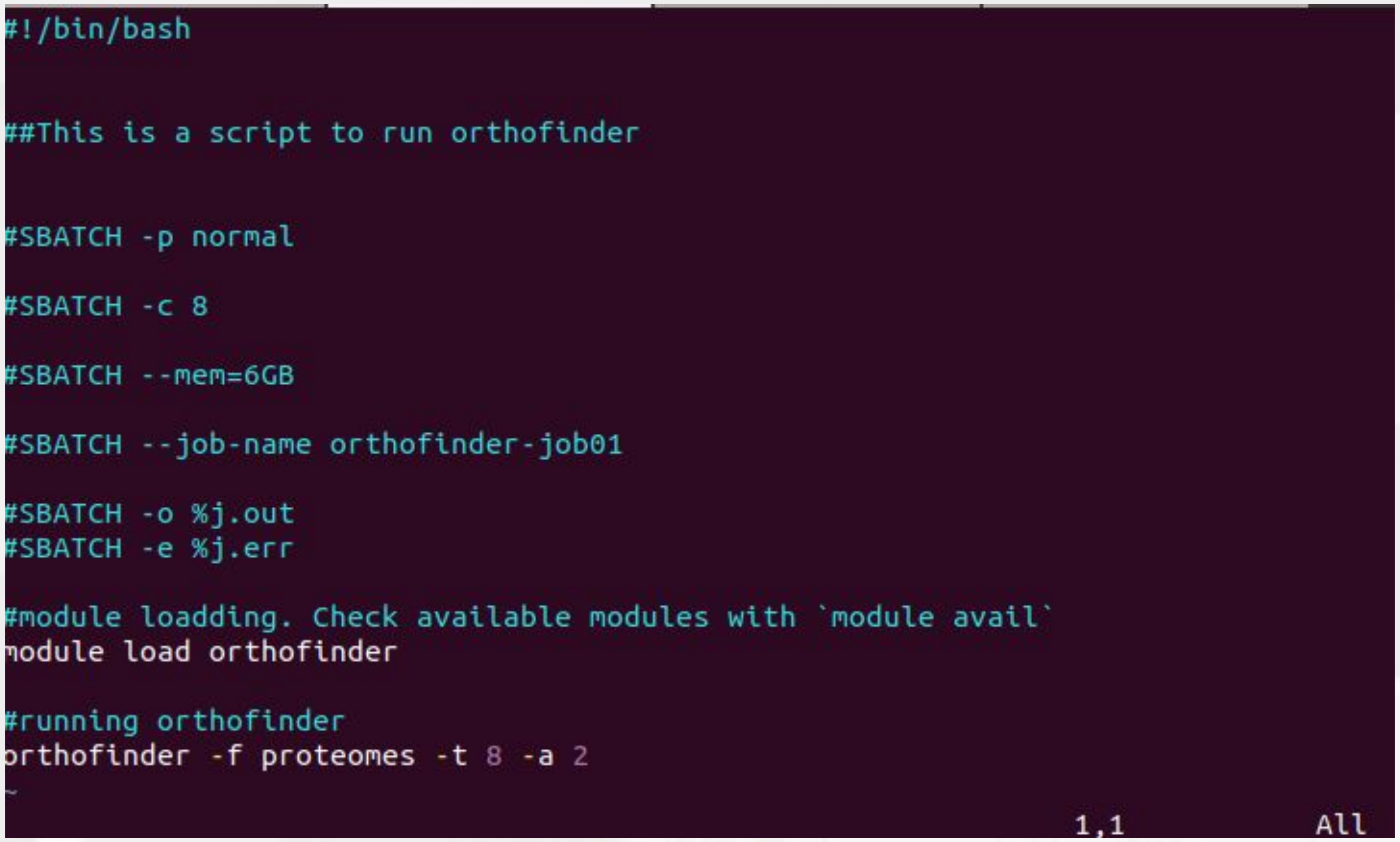
Repository for the 18th edition of the advanced couse in Phylogenomics and Population Genomics (ppgcourse) of the University of Barcelona (UB)

- Unzip the file
- Use scp -r folderName username@ec2-34-242-61-70.eu-west-1.compute.amazonaws.com://home/username/

How to use VIM to edit files

(You can also use emacs if you prefer)

- To open a file: vim fileName

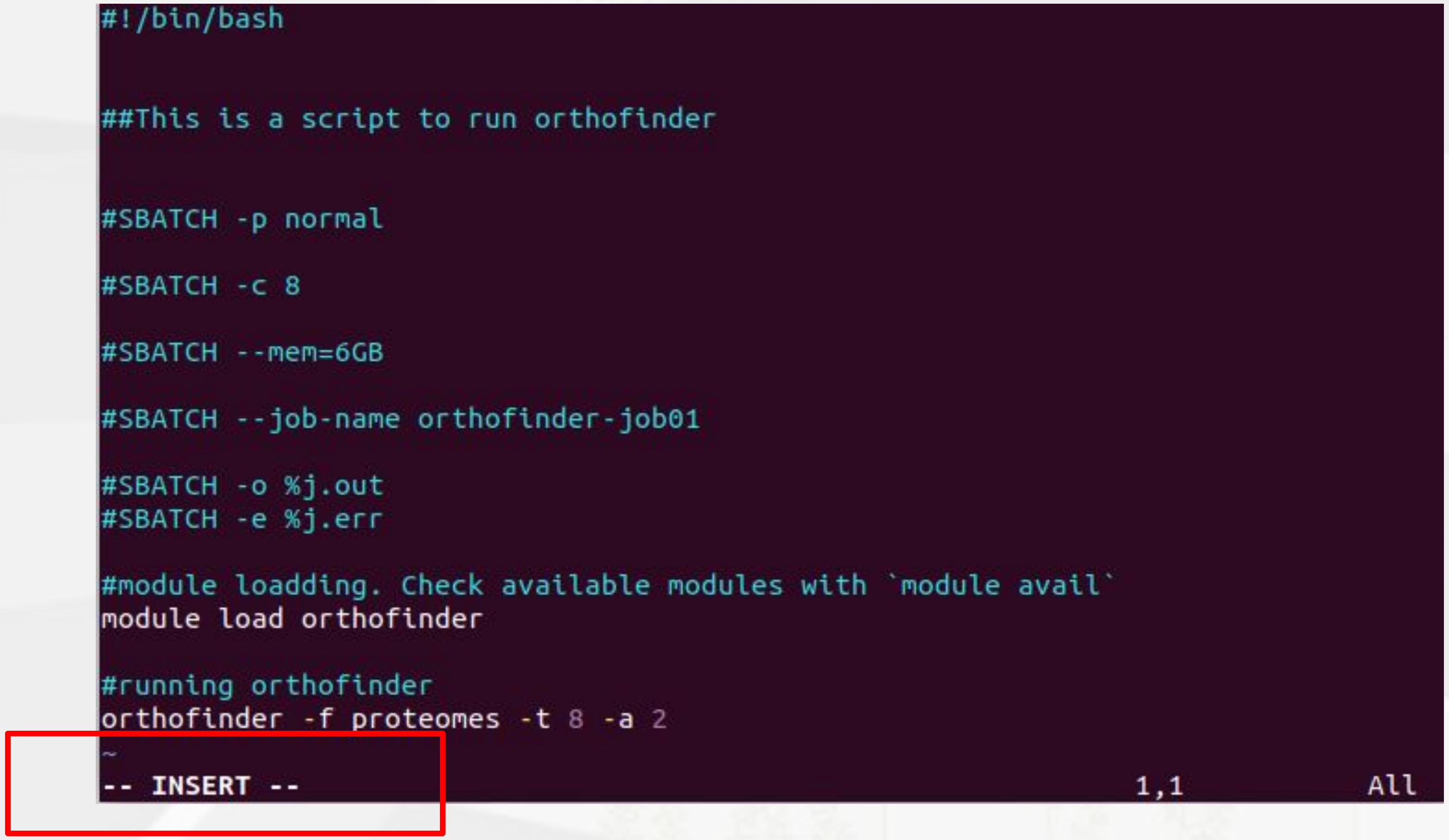


```
#!/bin/bash


##This is a script to run orthofinder


#SBATCH -p normal

#SBATCH -c 8

#SBATCH --mem=6GB

#SBATCH --job-name orthofinder-job01

#SBATCH -o %j.out
#SBATCH -e %j.err

#module loadding. Check available modules with `module avail`
module load orthofinder

#running orthofinder
orthofinder -f proteomes -t 8 -a 2
~
                                                            1,1           All
```

How to use VIM to edit files

(You can also use emacs if you prefer)

- Before you start to write, press **i**



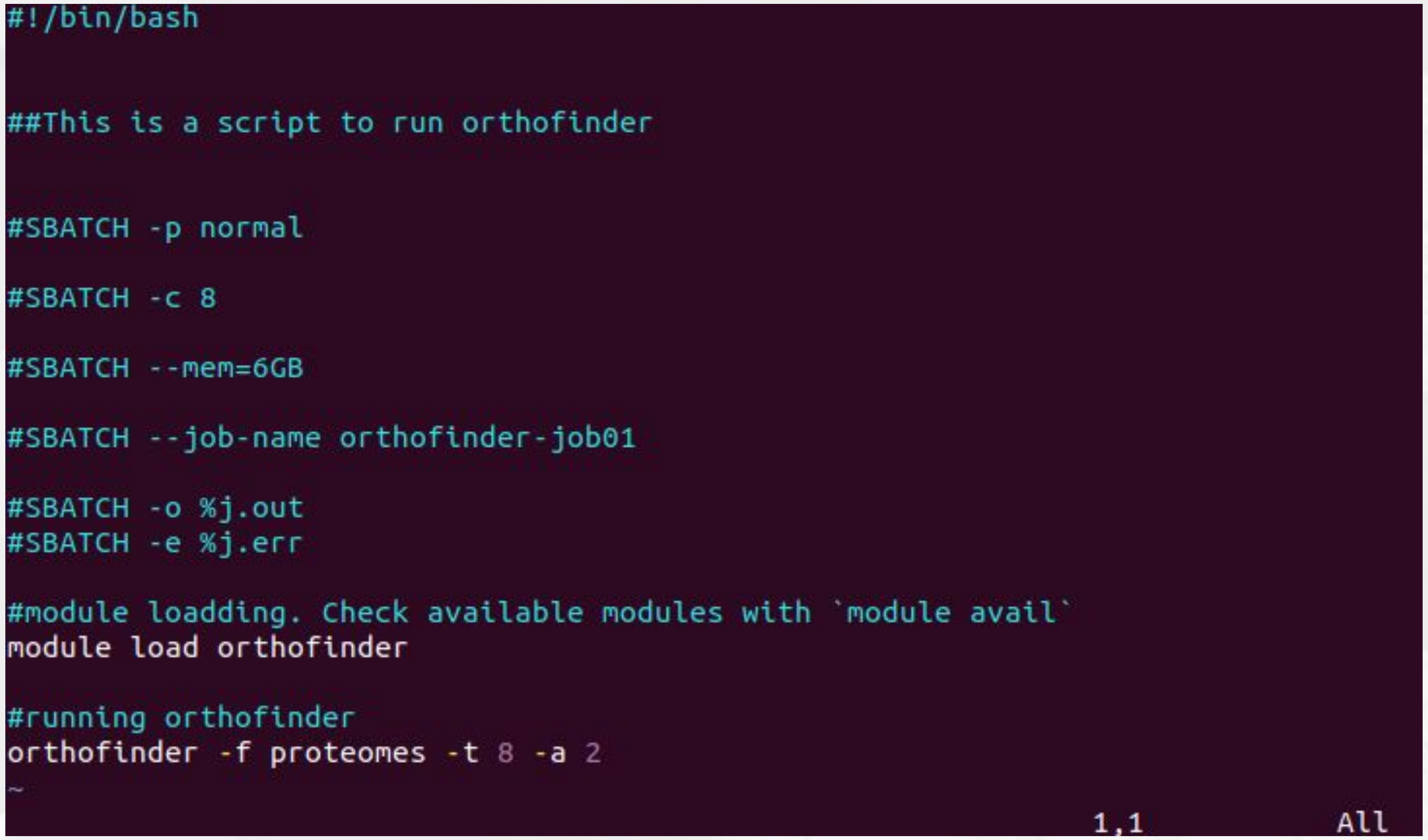
```
#!/bin/bash

##This is a script to run orthofinder

#SBATCH -p normal

#SBATCH -c 8

#SBATCH --mem=6GB

#SBATCH --job-name orthofinder-job01

#SBATCH -o %j.out
#SBATCH -e %j.err

#module loadding. Check available modules with `module avail`
module load orthofinder

#running orthofinder
orthofinder -f proteomes -t 8 -a 2
~
-- INSERT --                                              1,1           All
```

You now should have the word insert at the bottom

How to use VIM to edit files

(You can also use emacs if you prefer)

- Once you have edited what you wanted, press **ESC** (you will see that the --- insert --- will disappear)



```
#!/bin/bash

##This is a script to run orthofinder

#SBATCH -p normal

#SBATCH -c 8

#SBATCH --mem=6GB

#SBATCH --job-name orthofinder-job01

#SBATCH -o %j.out
#SBATCH -e %j.err

#module loadding. Check available modules with `module avail`
module load orthofinder

#running orthofinder
orthofinder -f proteomes -t 8 -a 2
~
                                                         1,1           All
```

- Now to save write **:wq!** and press enter

Reminder: How to move through the terminal

- To go to a folder:

```
cd folderName/folderName1
```

- To move back to the previous folder:

```
cd ..
```

- If you’re completely lost:

```
cd
```

This will just bring you to your home folder

Outline

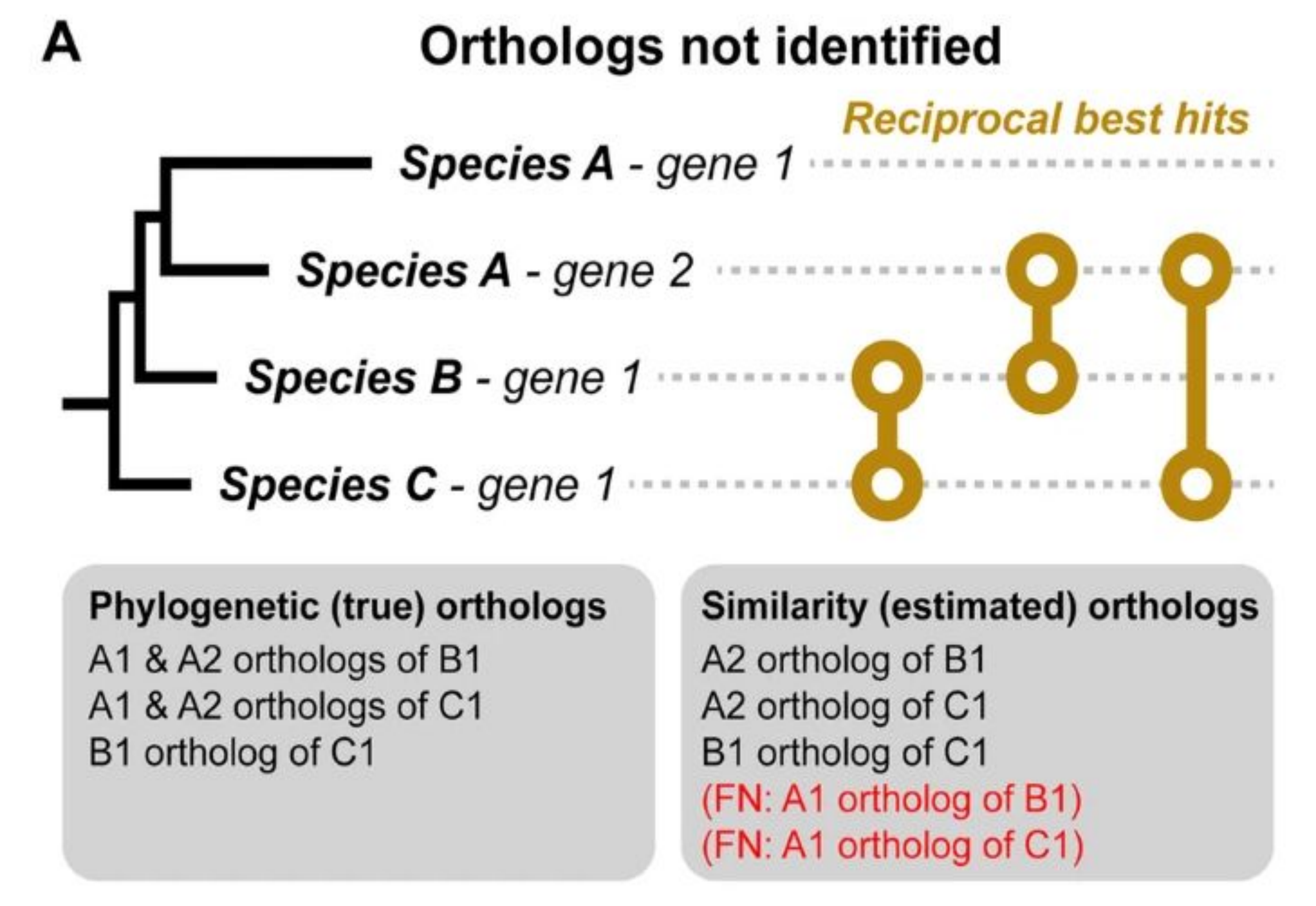
- Reminder
- Previous considerations
- OrthoFinder

Reminders

Homologs: Sequences that descend from a common ancestor.

Orthologs: Sequences that come from a speciation event.

Paralogs: Sequences that come from a duplication event.



Phylogenetic (true) orthologs
A1 & A2 orthologs of B1
A1 & A2 orthologs of C1
B1 ortholog of C1

Similarity (estimated) orthologs
A2 ortholog of B1
A2 ortholog of C1
B1 ortholog of C1
(FN: A1 ortholog of B1)
(FN: A1 ortholog of C1)

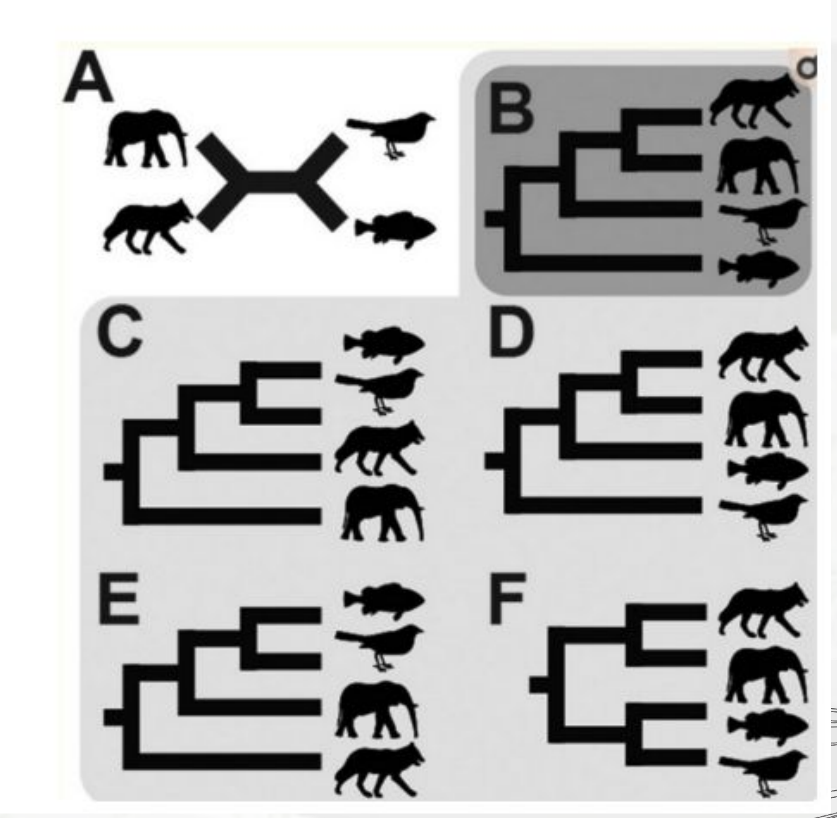
Orthogroups: Group of orthologous genes that can contain inparalogs

First considerations: What do you need to think about before starting.

- Species selection, specially outgroups
- Filtering of isoforms
- Fasta headers
- Computational resources

Species selection, specially outgroups

- How many species should we use?
- Genomes? Transcriptomes?
- Outgroups? How many?

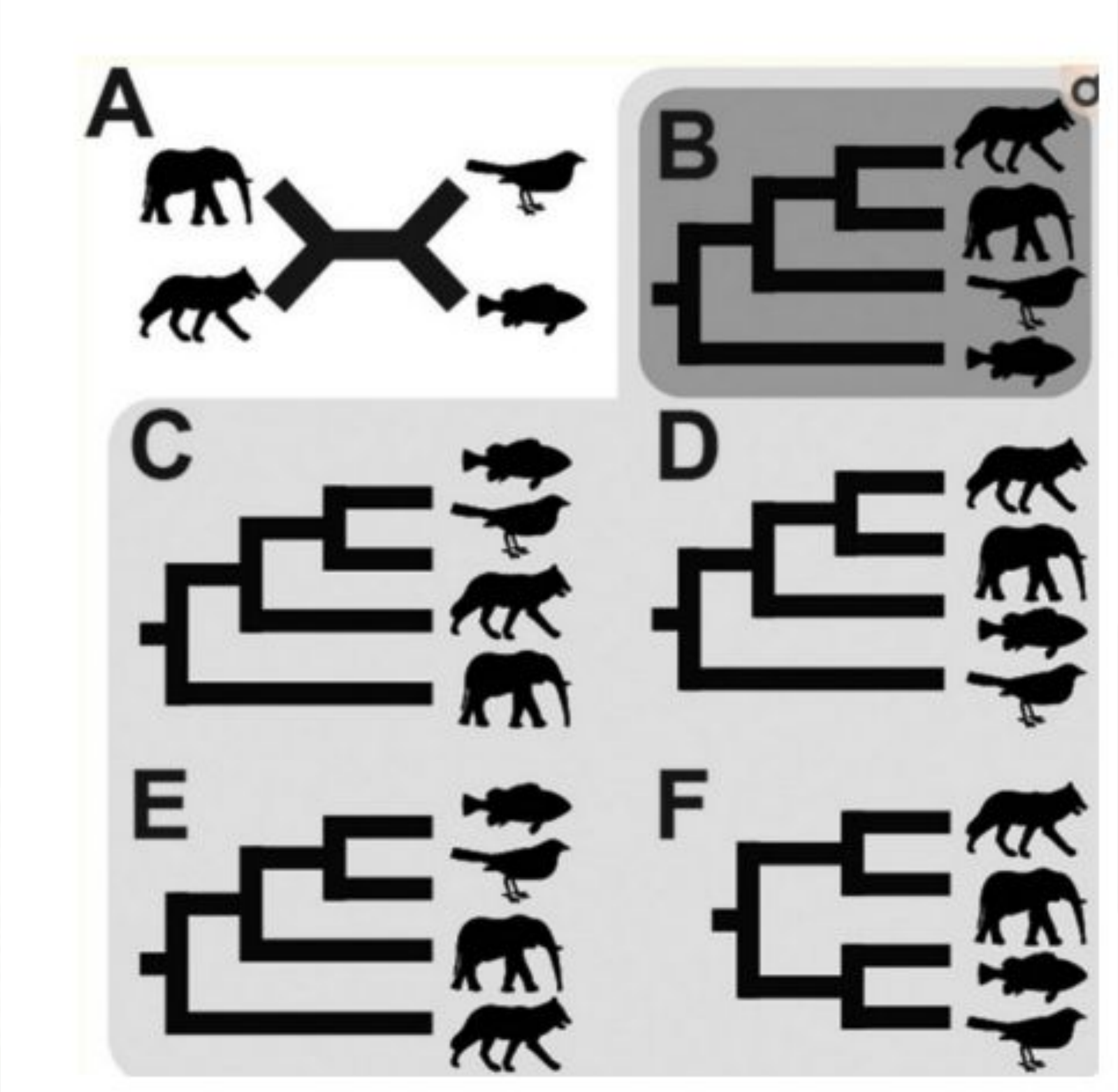


Outgroups

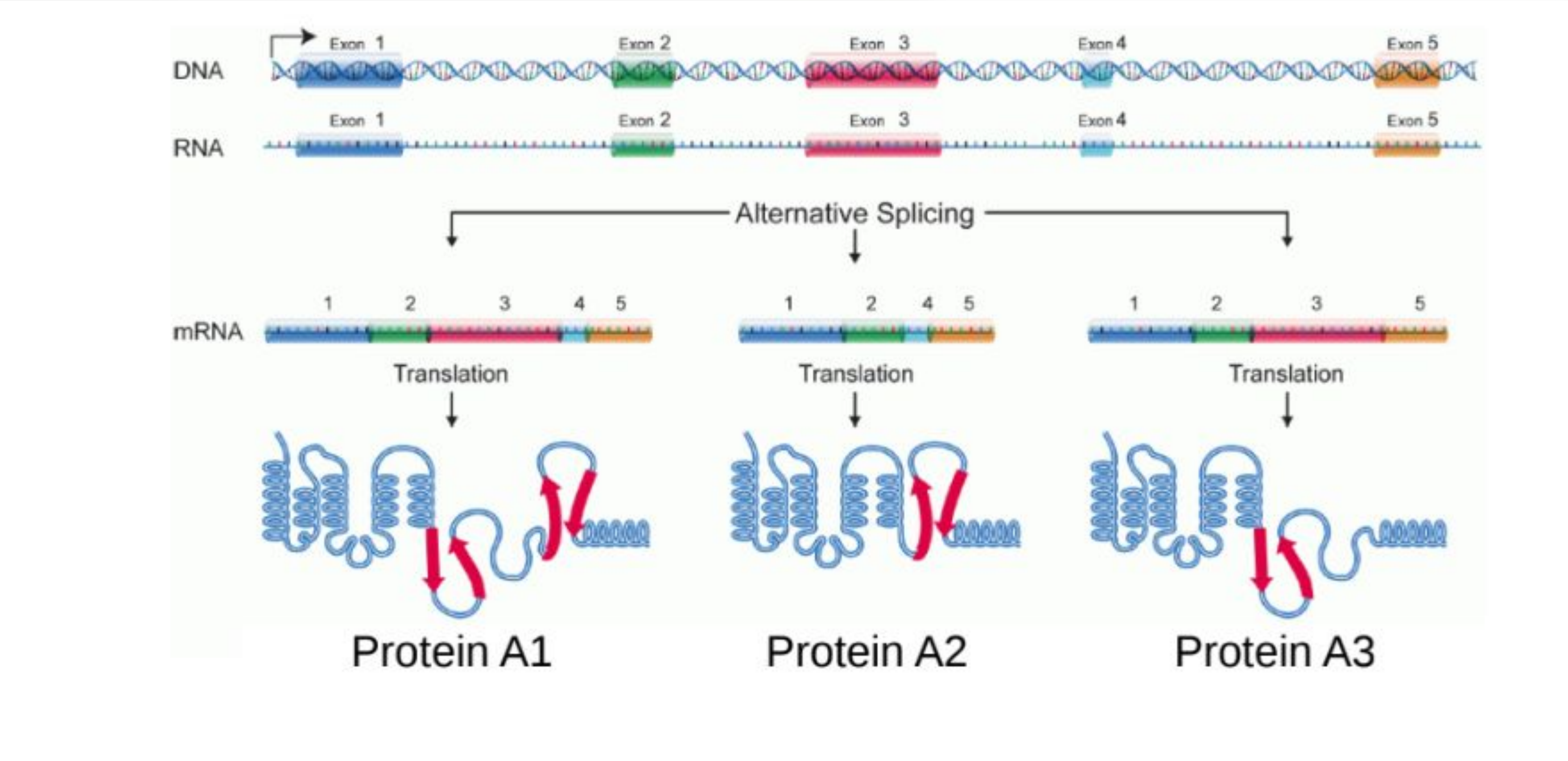
● When building a species tree, it is very important to use an outgroup in order to give directionality to the tree.

● Outgroups will also be necessary to root gene trees and perform orthology and paralogy predictions.

● If possible add at least two outgroups.

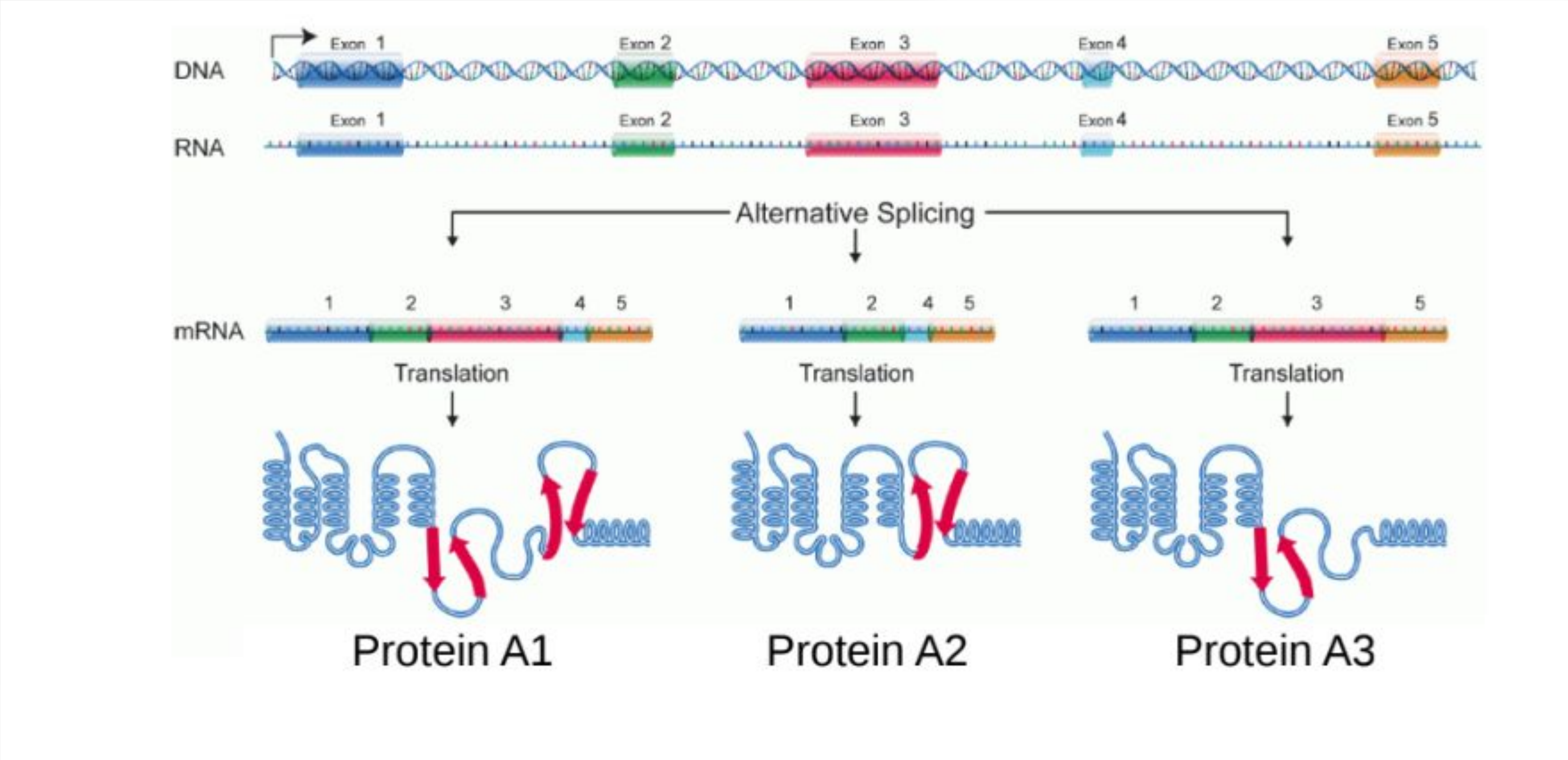


Isoforms

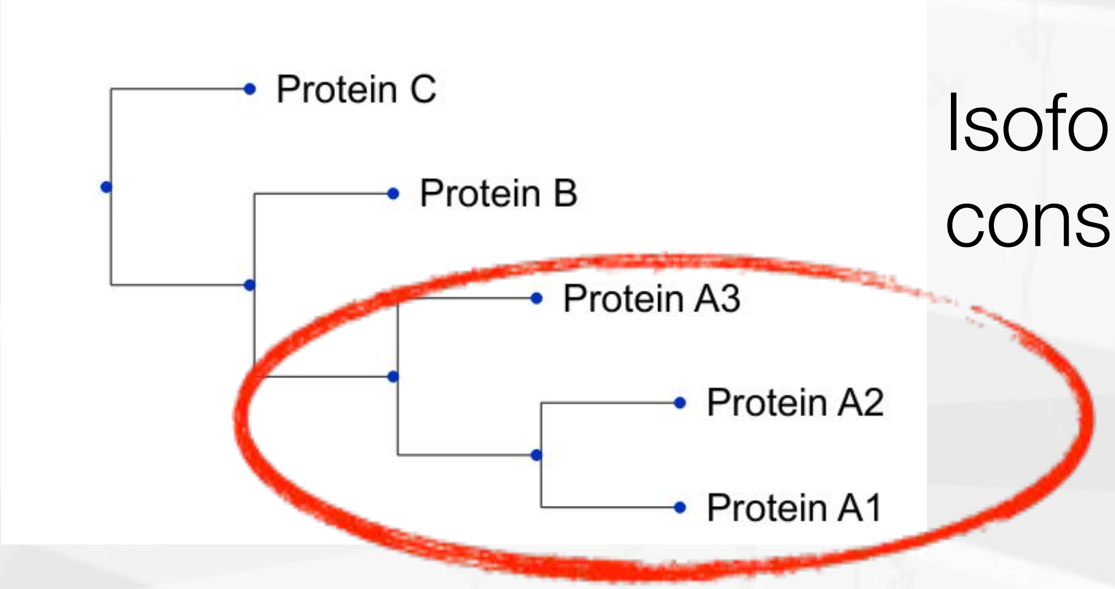


Should we add isoforms in our analysis?

Isoforms



Should we add isoforms in our analysis?



Isoforms will be considered as paralogs

Headers

Fasta files contain headers that can be complicated. At first it will not bother you, but the downstream analysis can become much more complicated.

```
>sp|D2H788|RN182_AILME E3 ubiquitin-protein ligase RNF182
OS=Ailuropoda melanoleuca OX=9646 GN=RNF182 PE=3 SV=1
```

This is a typical Uniprot header.

Do you think it’s a good idea to use it as such?

Computational resources

There are many ways to calculate orthology relationships, and some are more computationally expensive than others.

What would you use to do a homology search?

Computational resources

There are many ways to calculate orthology relationships, and some are more computationally expensive than others.

Homology search: **Blast** is the tool by default, yet **Diamond** is much faster when the database is big.

Computational resources

There are many ways to calculate orthology relationships, and some are more computationally expensive than others.

Homology search: **Blast** is the tool by default, yet **Diamond** is much faster when the database is big.

Orthology prediction: **Tree based** orthology prediction is more accurate, yet **similarity based** methods are faster.

Species selection: **More species** give more resolution, yet everything becomes more computationally expensive.

Before running an analysis always consider what you need and if you have the resources to get it.

OrthoFinder

OrthoFinder is a fast, accurate and comprehensive pipeline for comparative genomics. It finds orthogroups and orthologs, infers rooted gene trees for all orthogroups and identifies all of the gene duplication events in those gene trees.

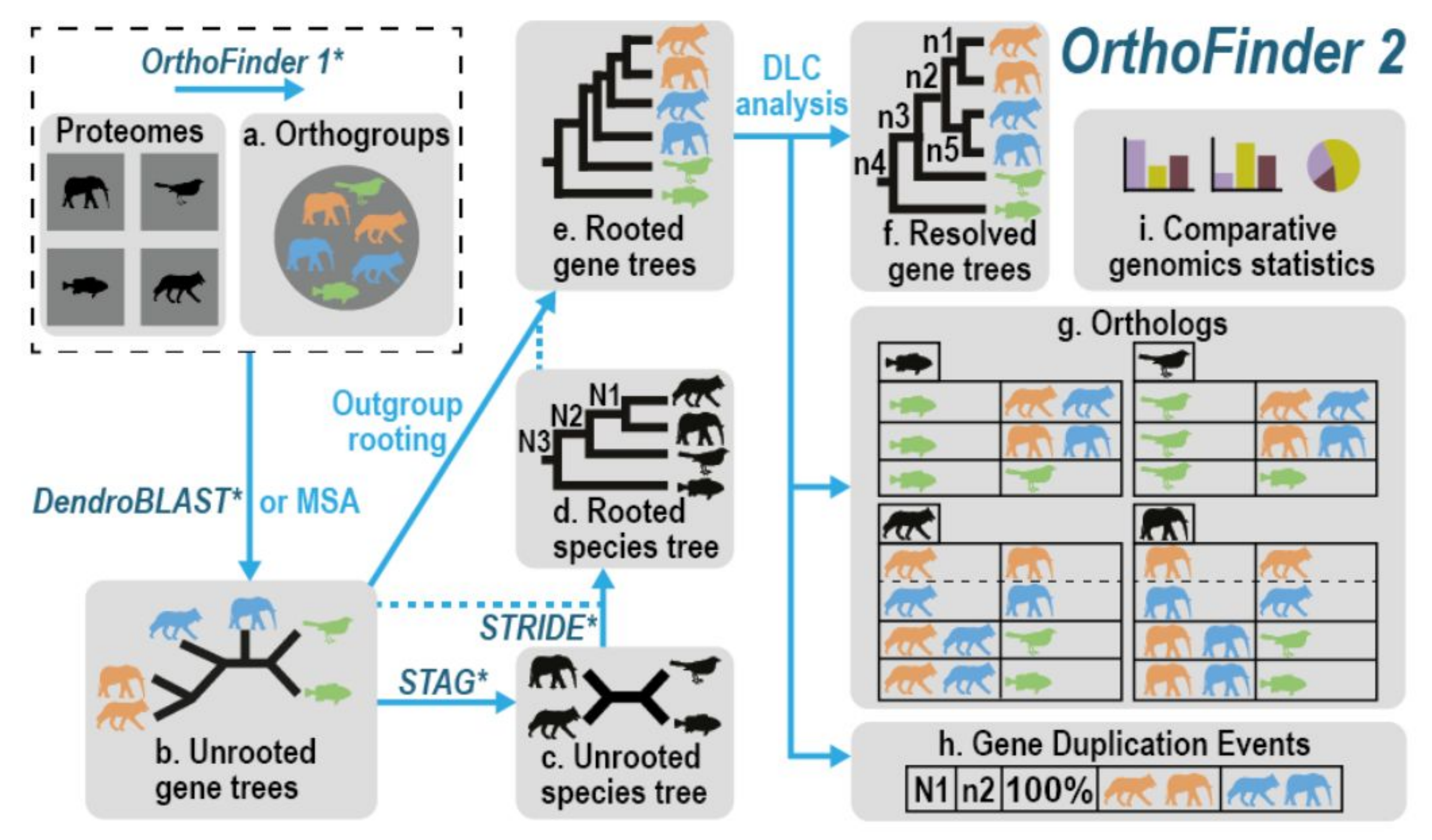
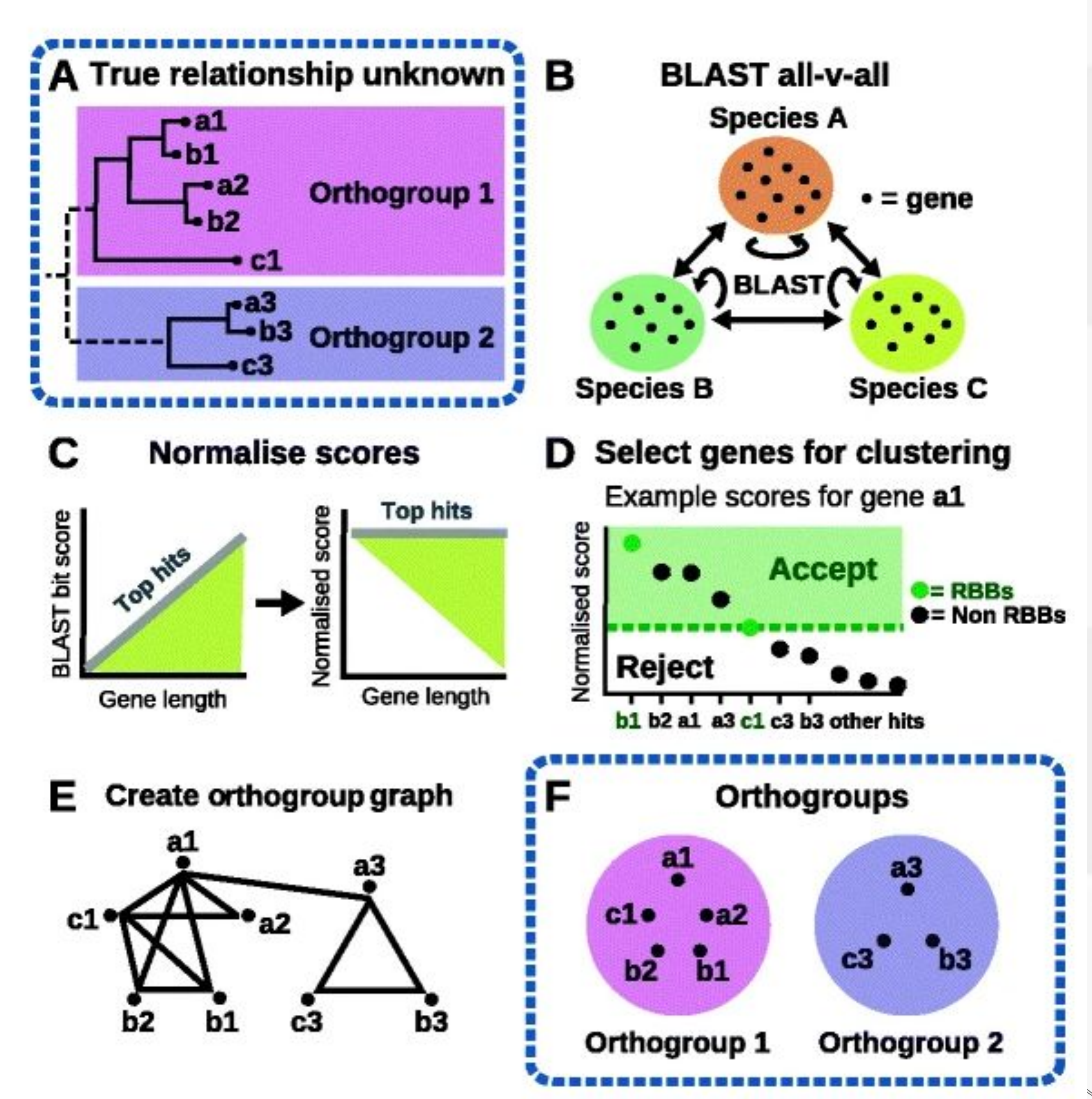


Figure 1: Automatic OrthoFinder analysis

OrthoFinder

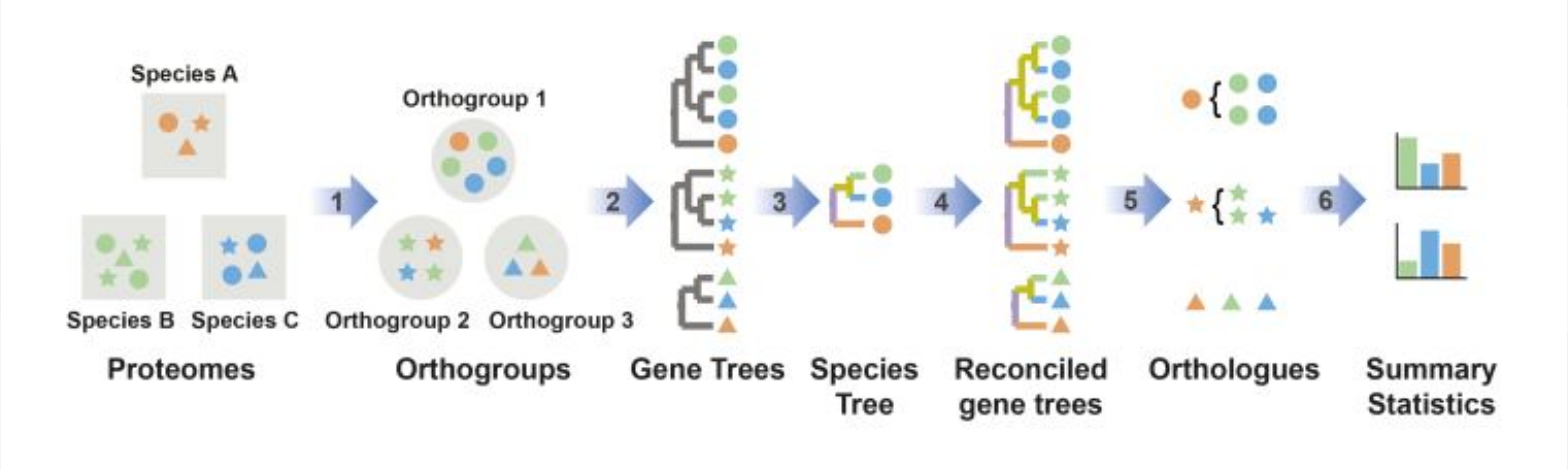
Things that Orthofinder solves compared to other algorithms:

- Bias towards gene length.
- Bias towards distantly related species.



OrthoFinder

The pipeline goes from a set of proteomes to fully resolved gene trees and their orthologs and paralogs



Time for the practical!

https://github.com/ppgcourseUB/ppgcourse2023/tree/main/Orthology_prediction_for_phylogenomic_analyses.MARINA_MARCET