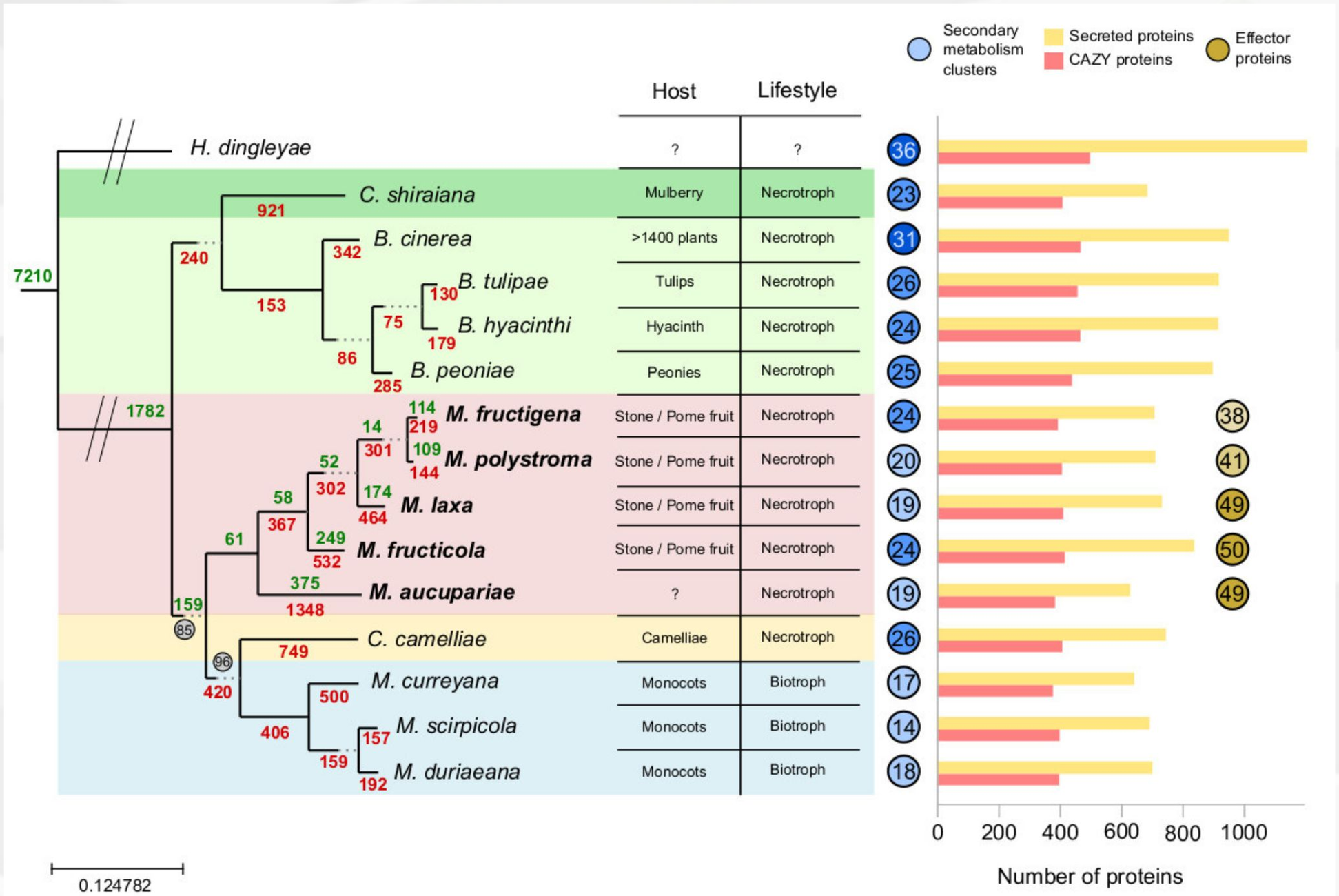
Phylogenomics and Population Genomics:
Inference and Applications

ORTHOLOGY PREDICTION FOR PHYLOGENOMIC ANALYSES

Marina Marcet Houben
mmarcet@bsc.es
Barcelona Supercomputing Center

- Degree in Biochemistry URV
- PhD in Fungal evolution Centro de investigación Principe Felipe and Center for Genomic regulation
- Center for Genomic regulation
- Barcelona Supercomputing Center & Institut de Recerca Biomedica


Secondary metabolism clusters
Secreted proteins
CAZY proteins
Effector proteins
Host
Lifestyle
H. dingleyae
?
?
C. shiraiana
Mulberry
Necrotroph
B. cinerea
>1400 plants
Necrotroph
B. tulipae
Tulips
Necrotroph
B. hyacinthi
Hyacinth
Necrotroph
B. peoniae
Peonies
Necrotroph
M. fructigena
Stone / Pome fruit
Necrotroph
M. polystroma
Stone / Pome fruit
Necrotroph
M. laxa
Stone / Pome fruit
Necrotroph
M. fructicola
Stone / Pome fruit
Necrotroph
M. aucupariae
?
Necrotroph
C. camelliae
Camelliae
Necrotroph
M. curreyana
Monocots
Biotroph
M. scirpicola
Monocots
Biotroph
M. duriaeana
Monocots
Biotroph
7210
921
240
342
153
130
75
179
86
285
1782
114
219
14
301
109
144
52
302
174
58
464
367
249
61
532
375
1348
159
85
96
749
420
500
406
157
159
192
36
23
31
26
24
25
24
38
20
41
19
49
24
50
19
49
26
17
14
18
0
200
400
600
800
1000
Number of proteins
0.124782

Outline

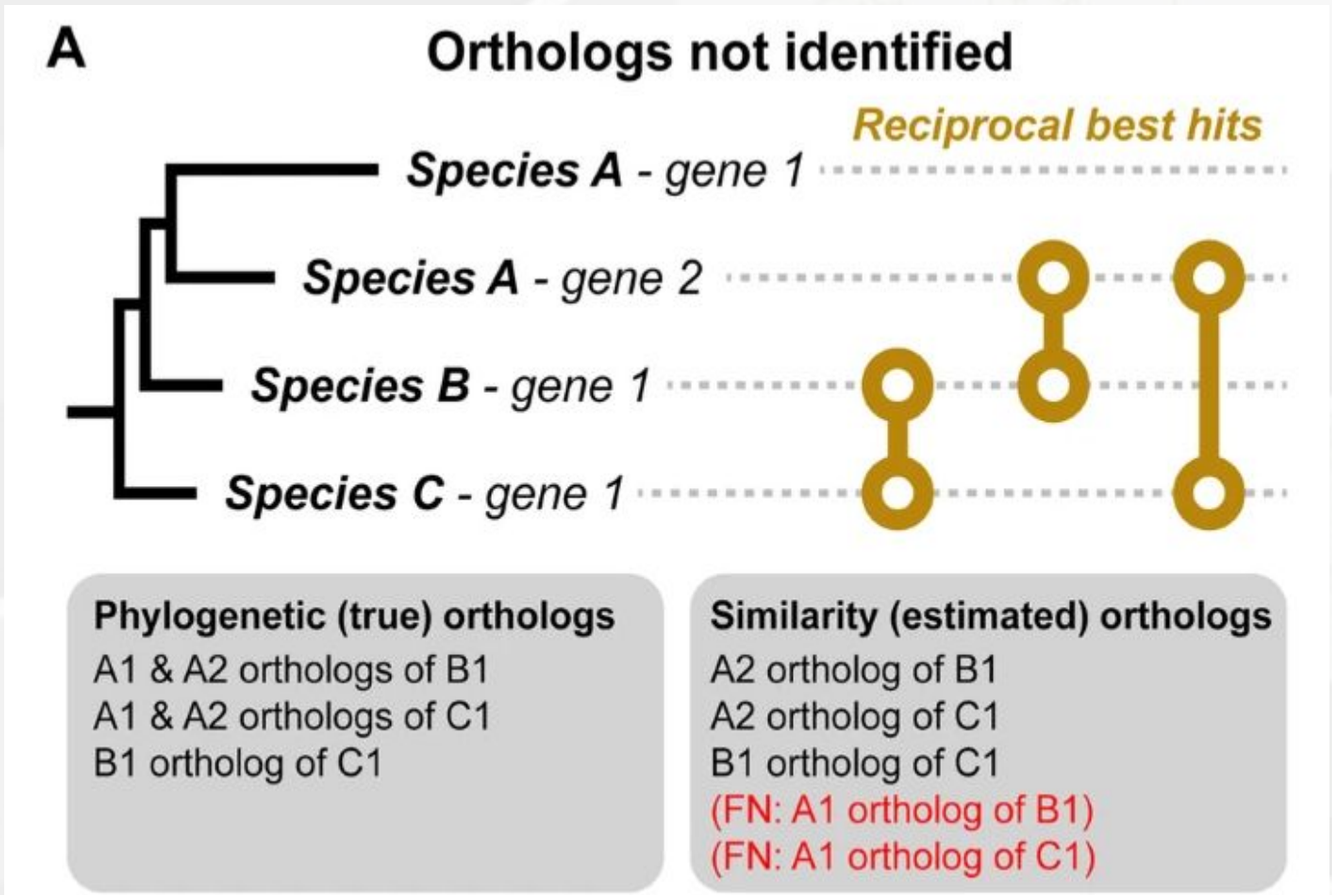
- Reminder
- Previous considerations
- OrthoFinder

Reminders

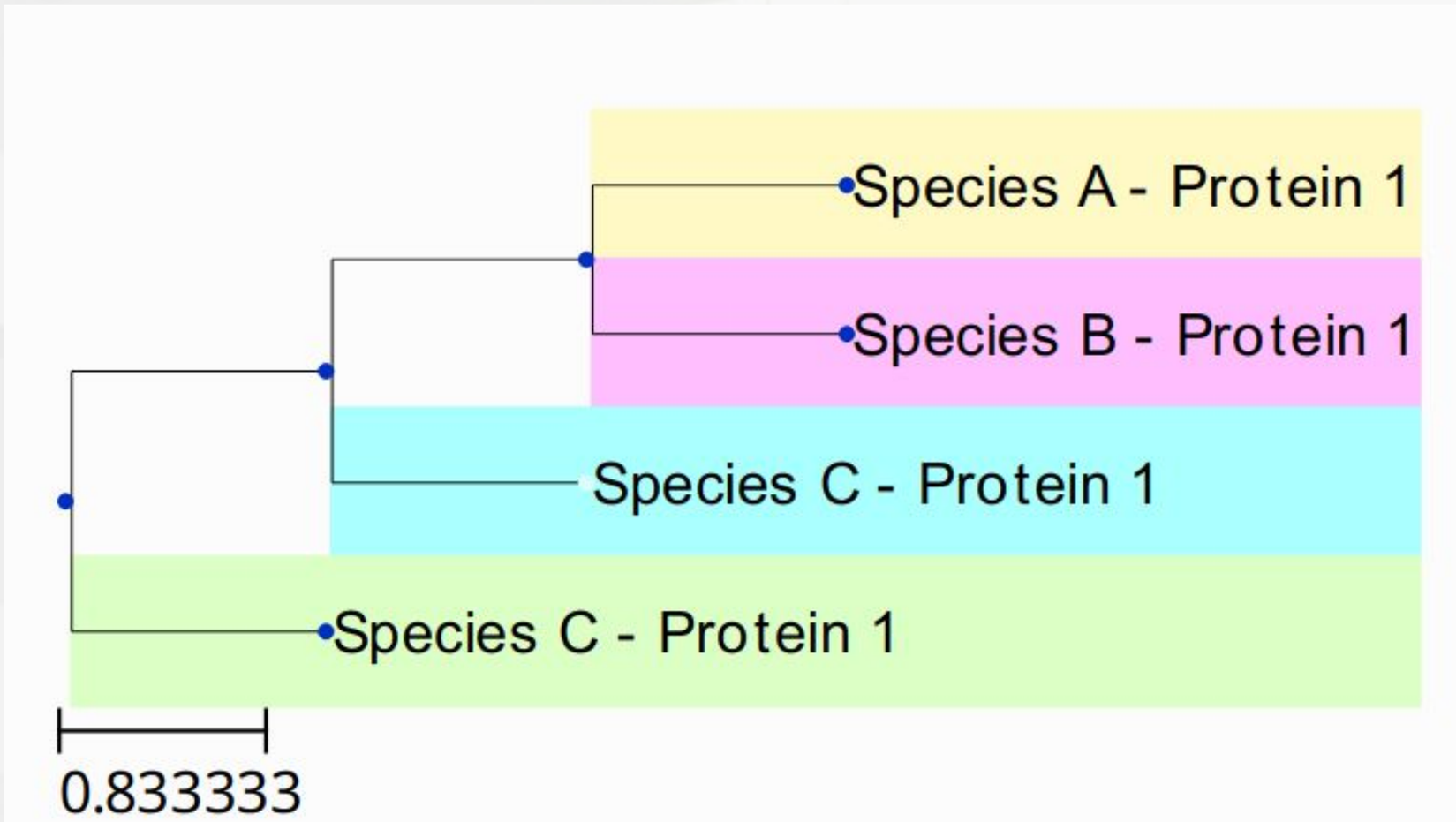
Homologs: Sequences that descend from a common ancestor.

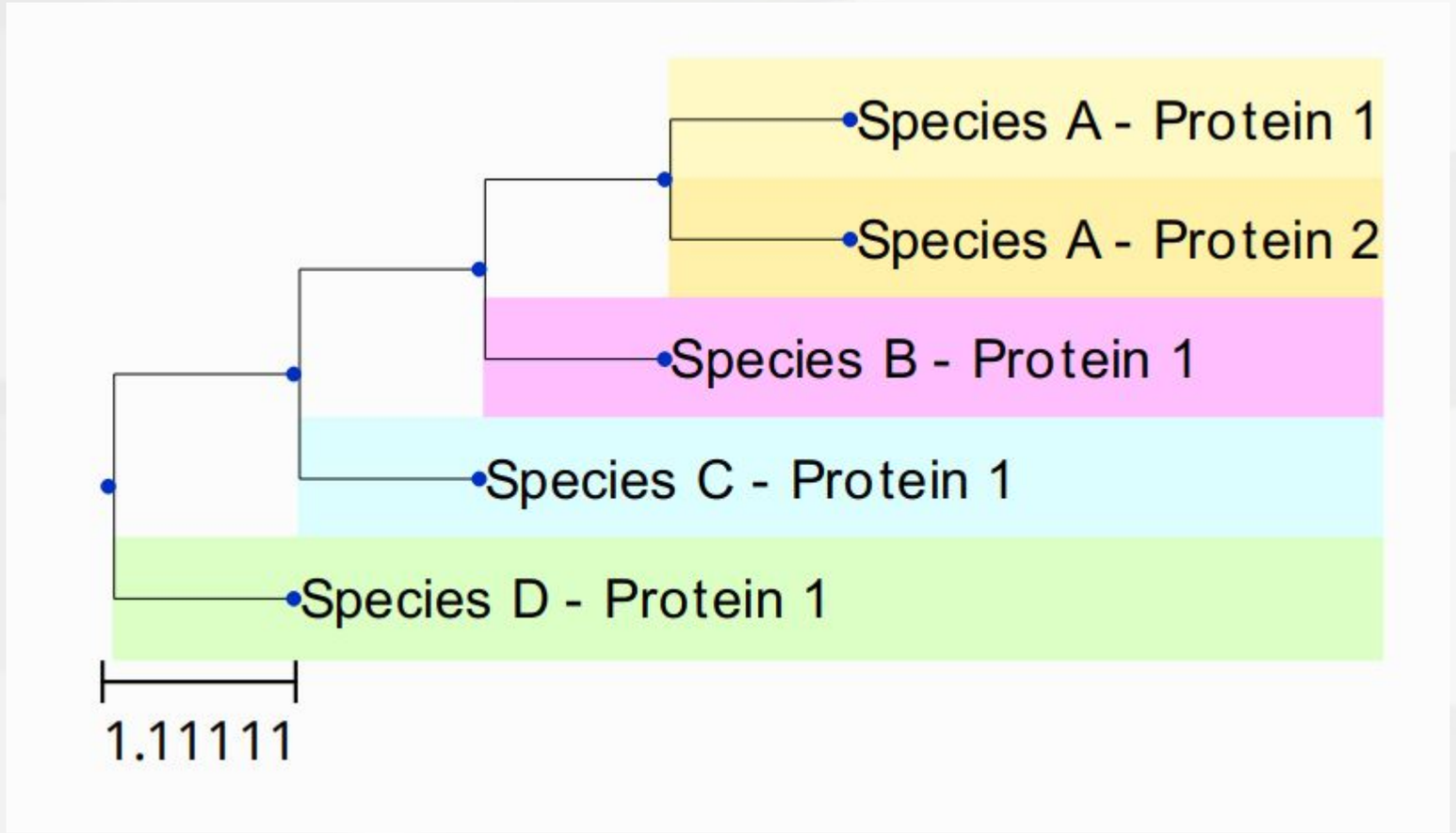
Orthologs: Sequences that come from a speciation event.

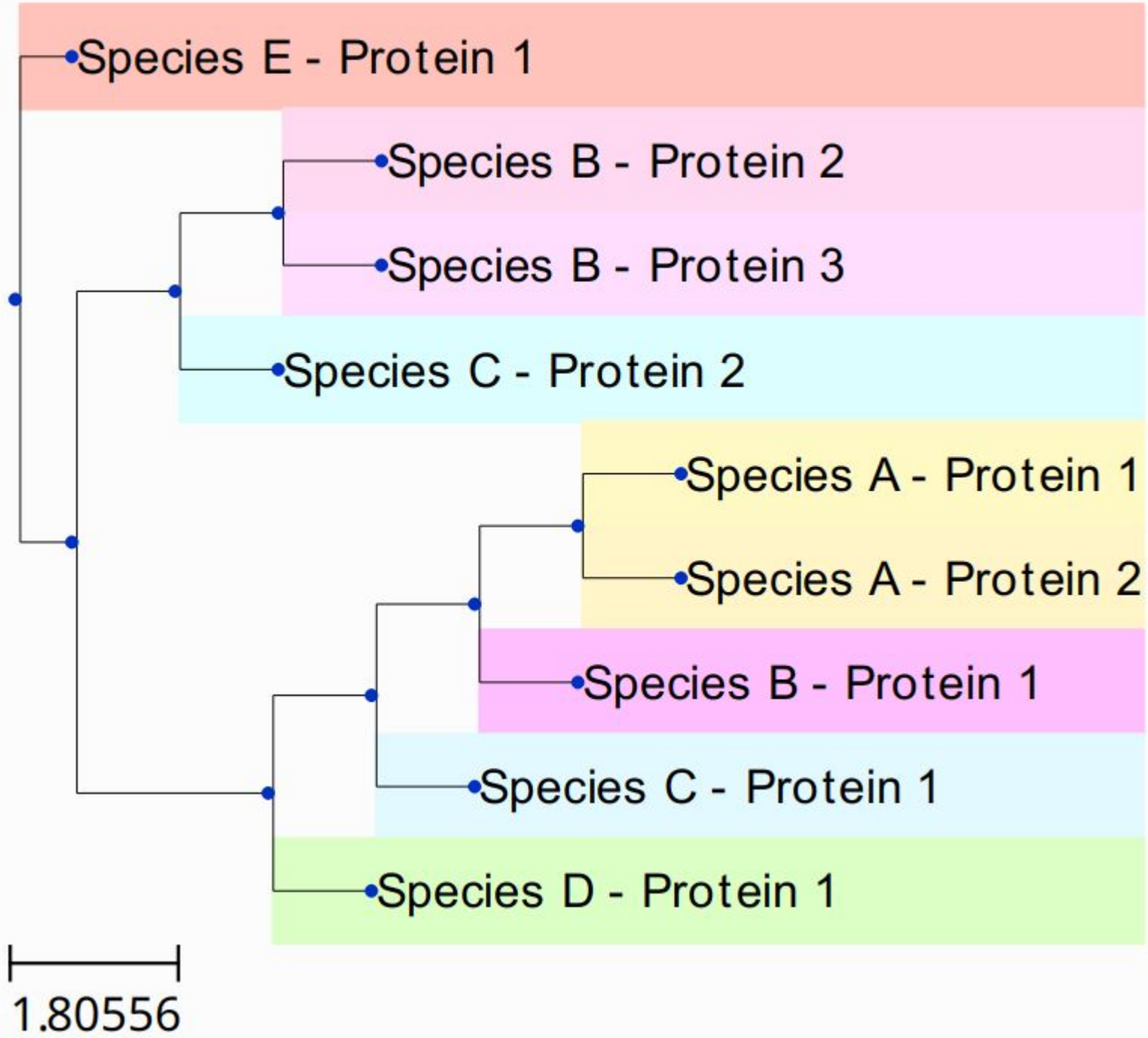
Paralogs: Sequences that come from a duplication event.



Orthogroups: Group of sequences that descend from a speciation event and can contain orthologs and in-paralogs.


Species A - Protein 1
Species B - Protein 1
Species C - Protein 1
Species C - Protein 1
0.833333


Species A - Protein 1
Species A - Protein 2
Species B - Protein 1
Species C - Protein 1
Species D - Protein 1
1.11111


Species E - Protein 1
Species B - Protein 2
Species B - Protein 3
Species C - Protein 2
Species A - Protein 1
Species A - Protein 2
Species B - Protein 1
Species C - Protein 1
Species D - Protein 1
1.80556

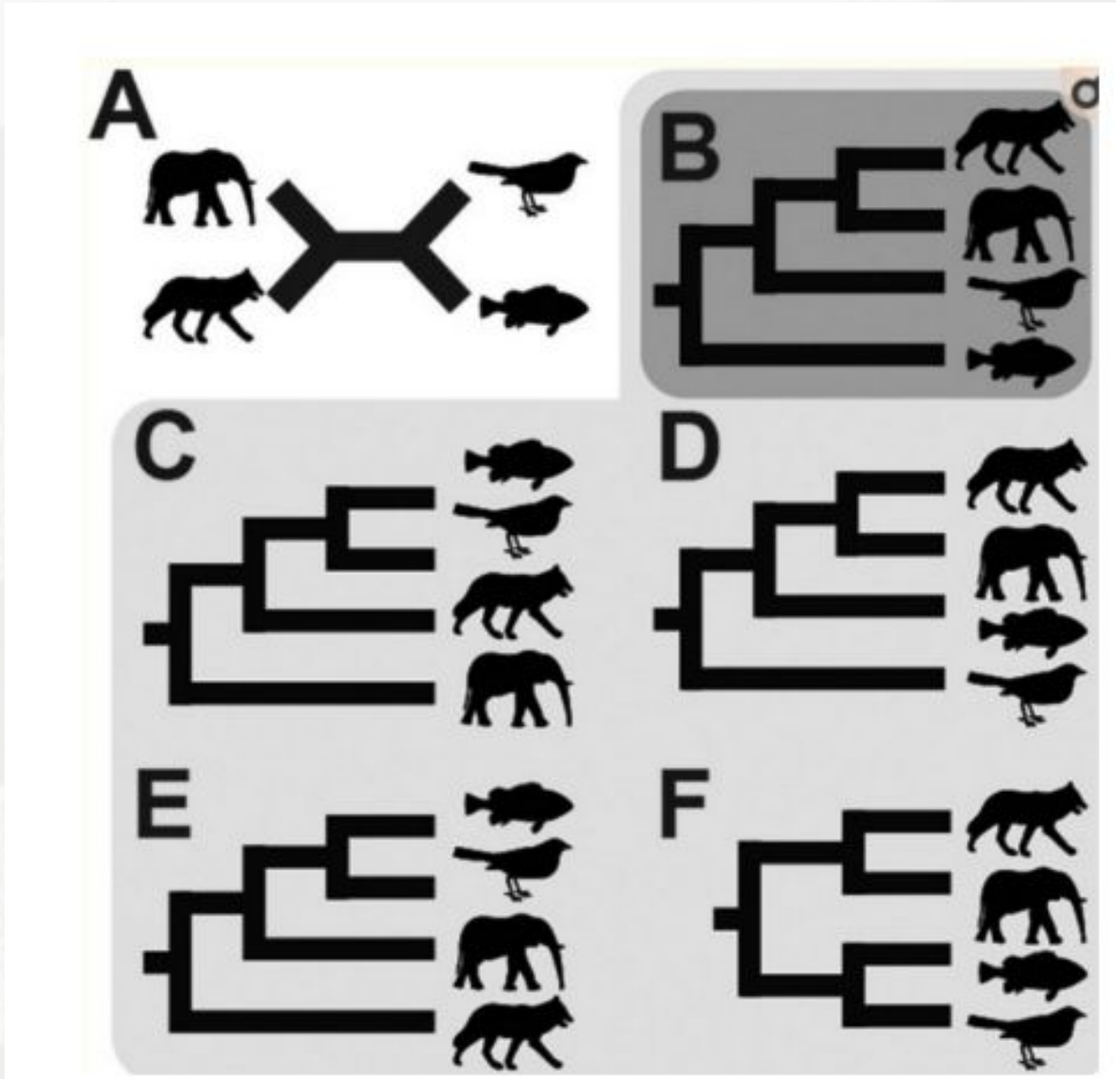
First considerations: What do you need to think about before starting.

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

Outgroups

An outgroup is a species that falls outside of our group of interest.

Do we want to add any to our analysis?

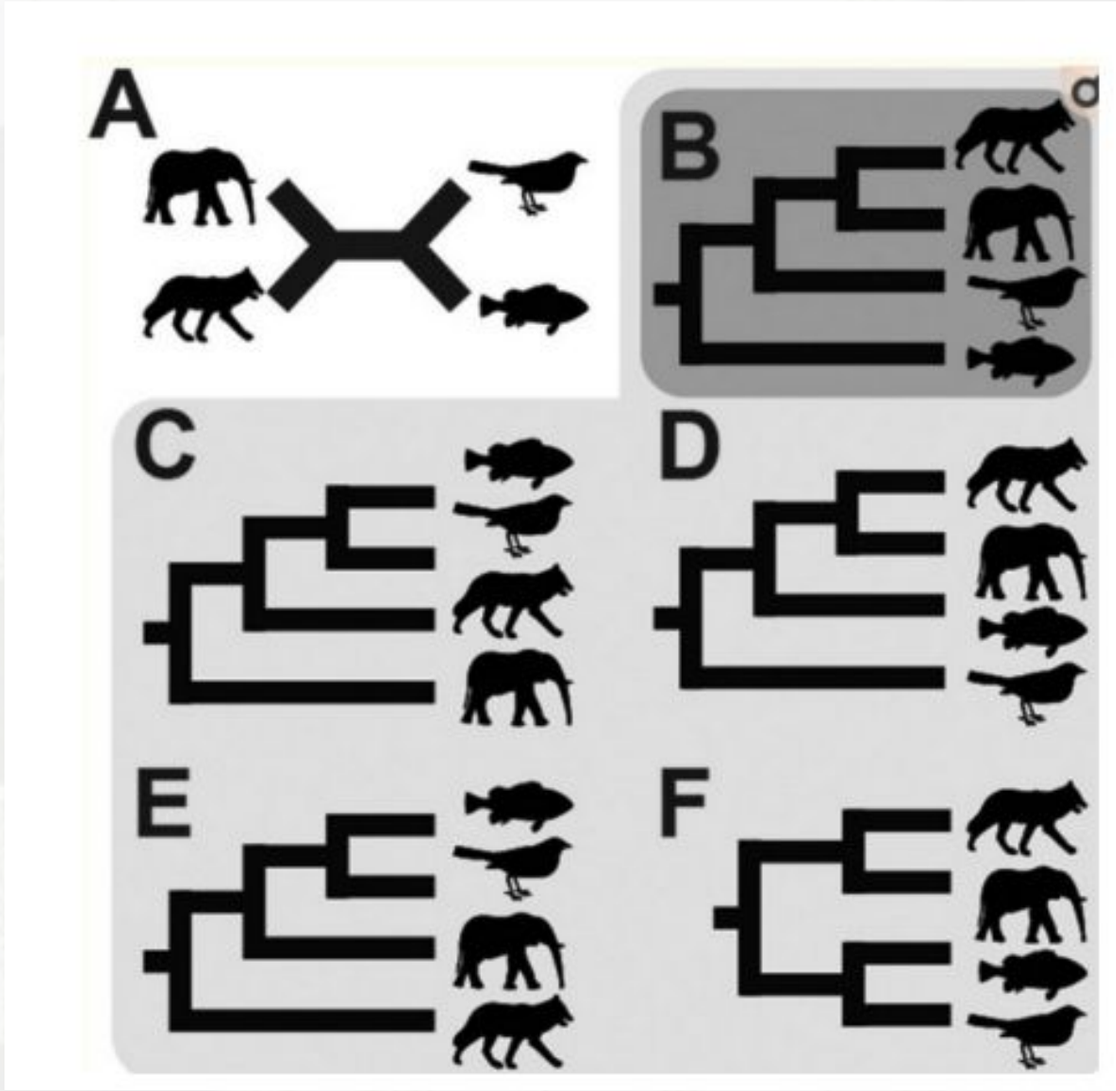


Outgroups

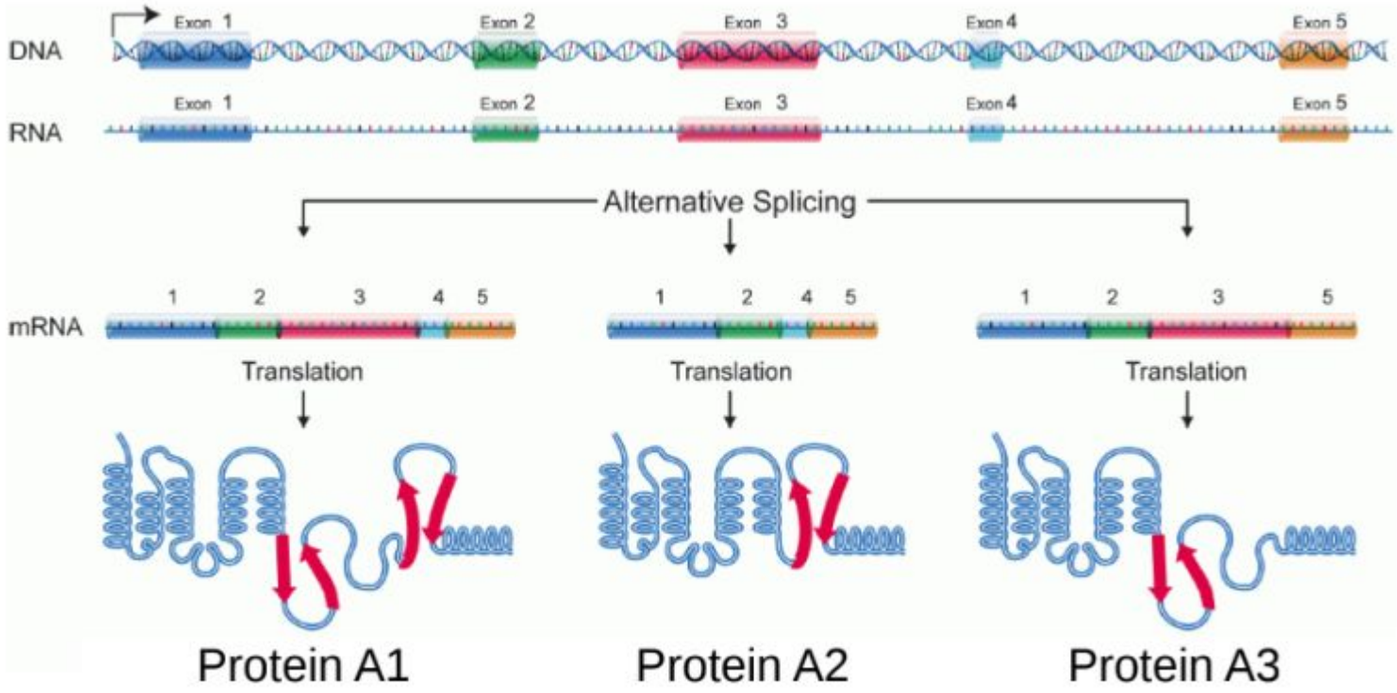
- When building a species tree, it is very important to use an outgroup in order to give directionality to the tree.

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

- If possible add at least two outgroups.

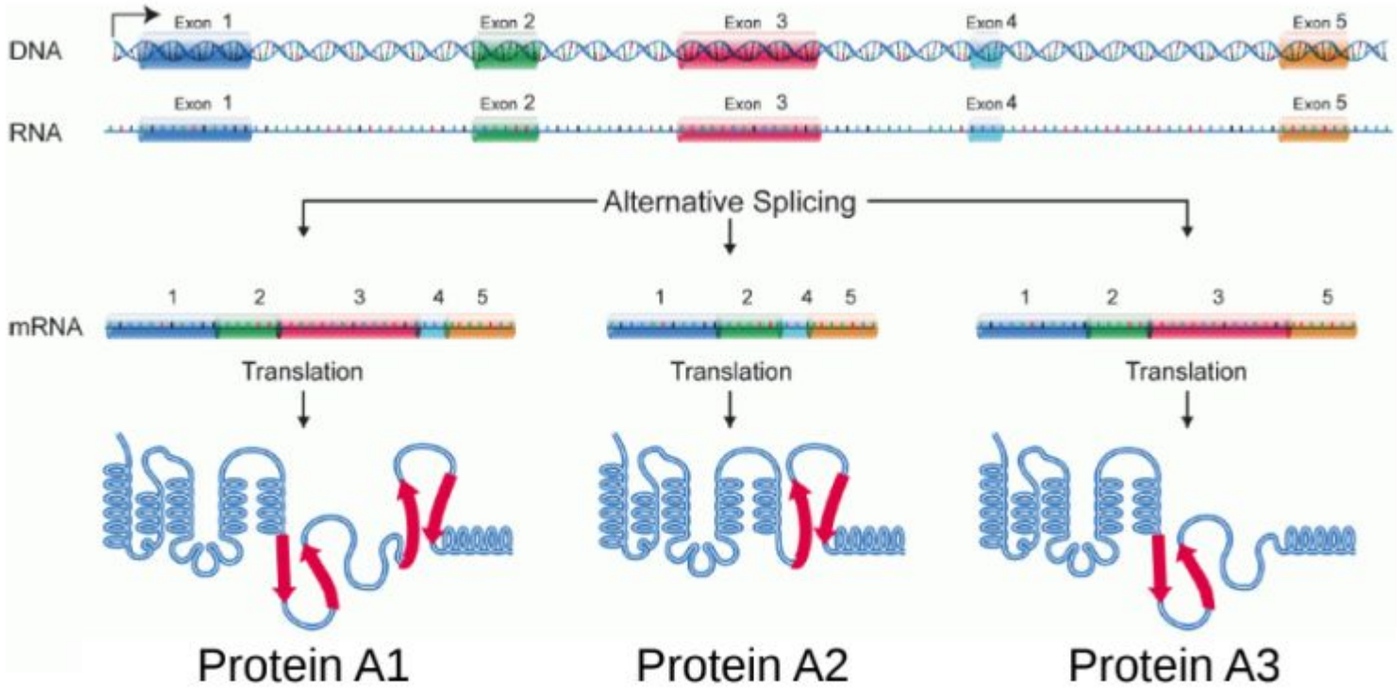


Isoforms

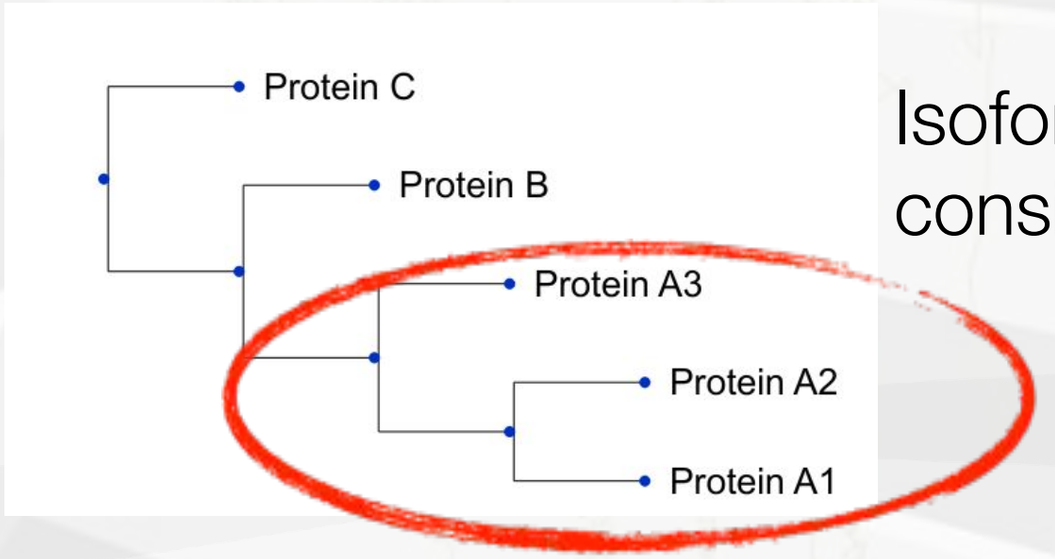


Should we add isoforms in our analysis?

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.

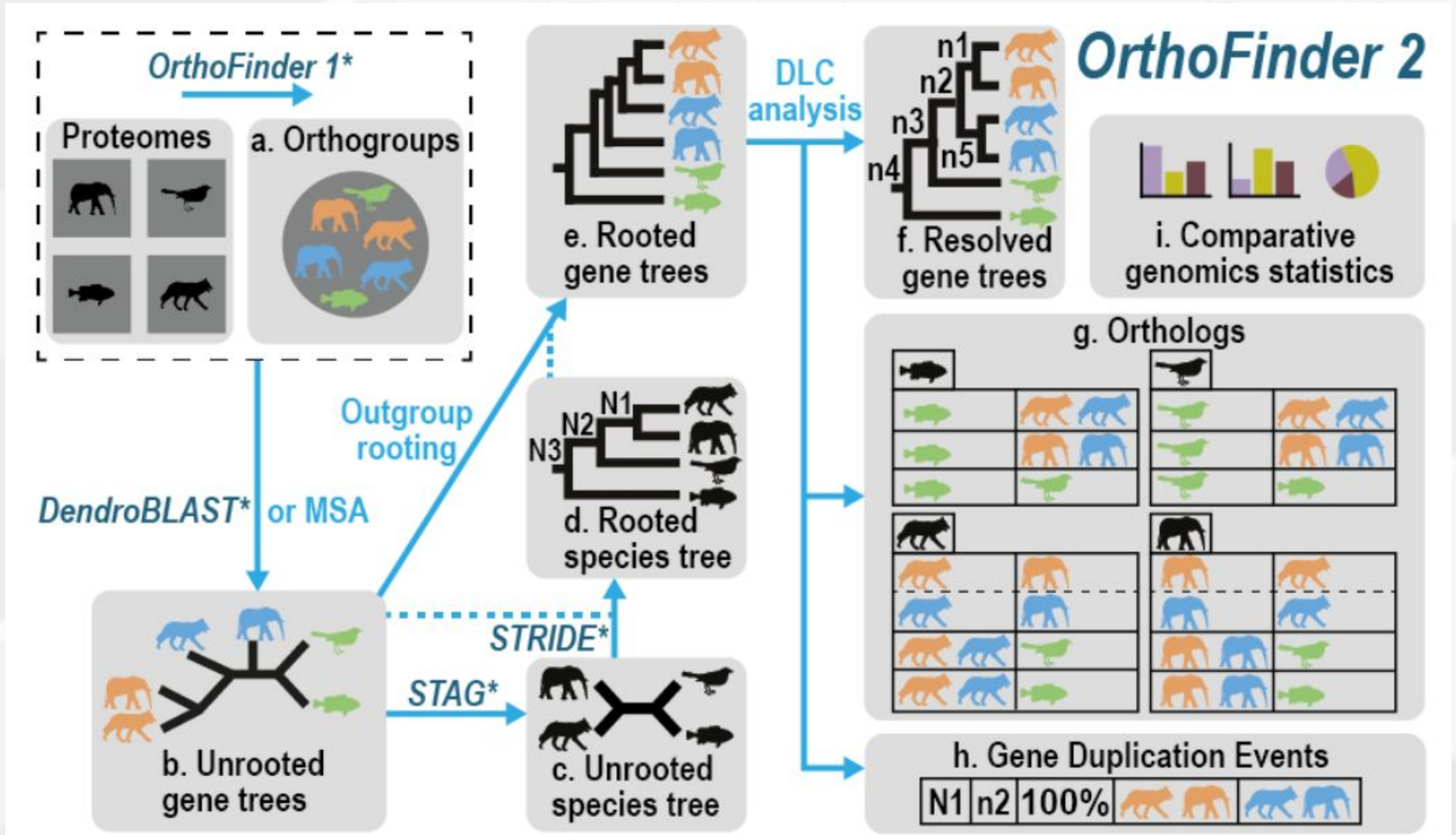
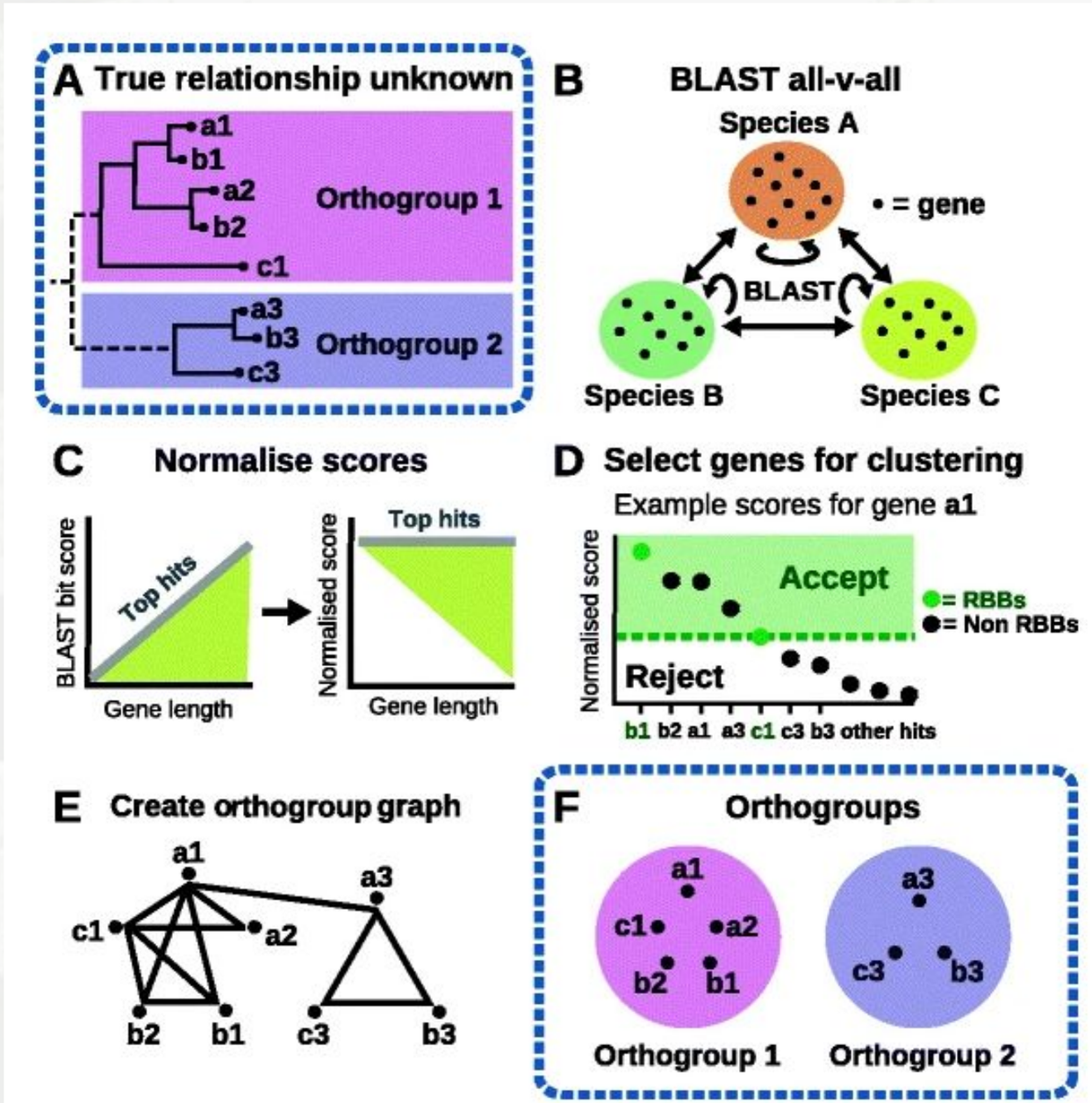


Figure 1: Automatic OrthoFinder analysis

OrthoFinder

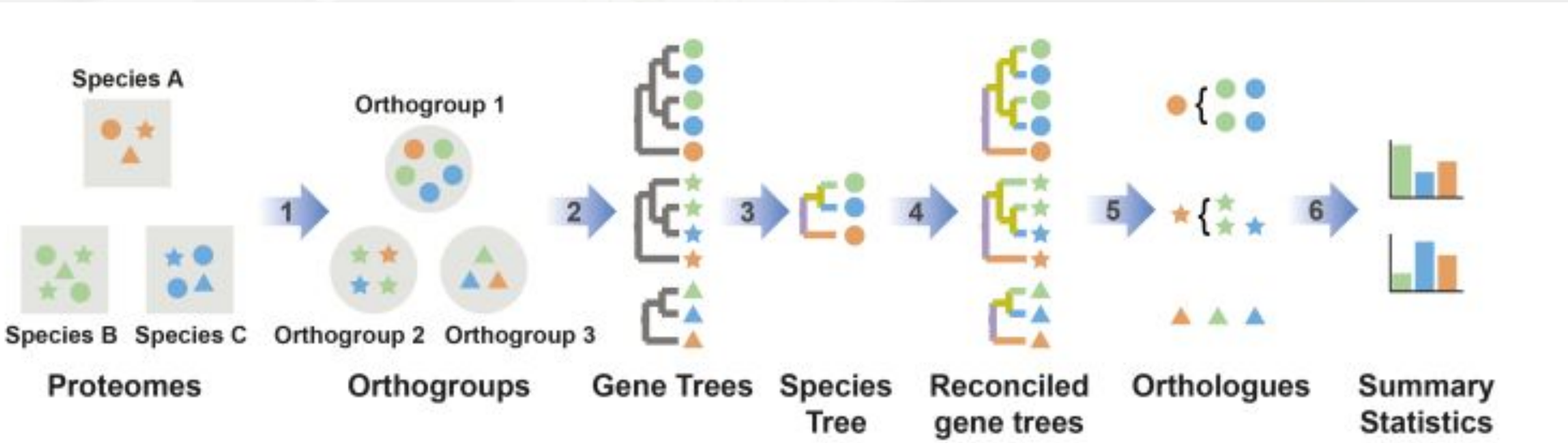
Things that Orthofinder solves compared to other algorithms:

- 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/ppgcourse2022/tree/main/Orthology