Biodiversity bioinformatics

From large-scale phylogenomics to gene families and functions

26 Aug 2024: Orthology with OMA

UNIL | Université de Lausanne

comparative genomics LAB

SIB

Learning objectives

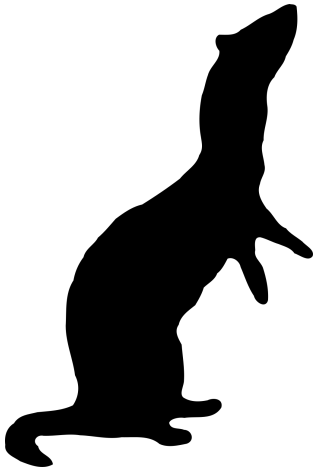
This course is centered around comparative genomics. After the course, you will be able to:

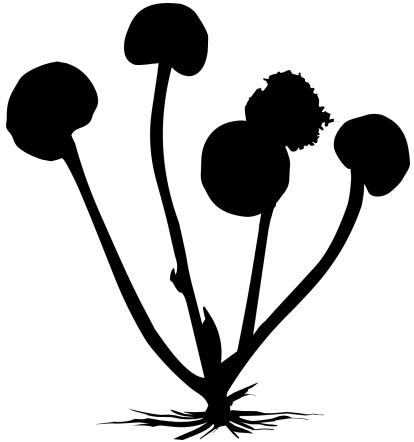
- Define orthology, paralogy and their subtypes
- Retrieve orthology information from the OMA database
- Map sequences quickly to their Hierarchical Orthologous Groups
- Infer orthologs on custom genomes using the FastOMA pipeline
- Construct and interpret phylogenetic species trees using OMA data

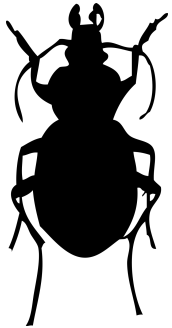
Schedule

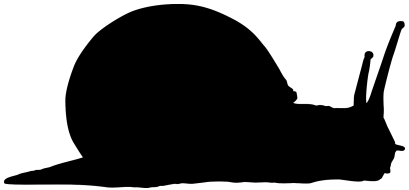
Schedule Aug 26 2024

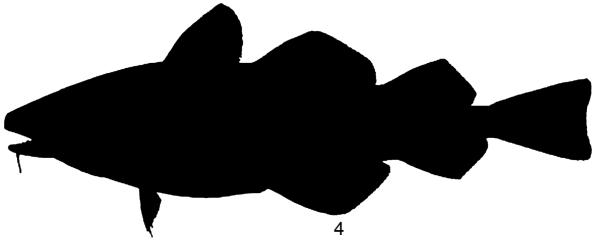
Time	Activity	In charge
9:00-9:30	Welcome, Introductions	Natasha Glover
9:30-10:00	Lecture: Overview, objectives, motivation, concept of orthologs, HOGs	Natasha Glover
10:00-10:45	**Module 1**: genes, groups, and genomes in the OMA Browser	Natasha Glover
10:45 - 11:15	Coffee break	
11:15-11:25	Go over results	Natasha Glover
11:25 - 11:30	Lecture: introduce OMAmer	Athina Gavriilidou
11:30 - 11:55	**Module 2**: Fast placement with OMAmer	Athina Gavriilidou
11:55-12:00	Go over results	Athina Gavriilidou
12:00-12:10	Lecture: FastOMA	Stefano Pascarelli
12:10-12:30	**Module 3 part 1**: FastOMA (launch it before lunch)	Stefano Pascarelli
12:30-13:30	Lunch	
13:30-14:00	**Module 3 part 2**: FastOMA	Stefano Pascarelli
14:00-14:15	Go over results	Stefano Pascarelli
14:15-14:30	Lecture: Gene trees and species trees	Christophe Dessimoz
14:30-15:00	Coffee Break	
15:00-16:00	**Module 4**: Building Species Trees	Christophe Dessimoz
16:00-16:15	Go over results	Christophe Dessimoz
16:15-16:45	OMA Clinic	Natasha, Athina, Stefano, Christophe
16:45-17:00	Wrap up	Natasha, Athina, Stefano, Christophe

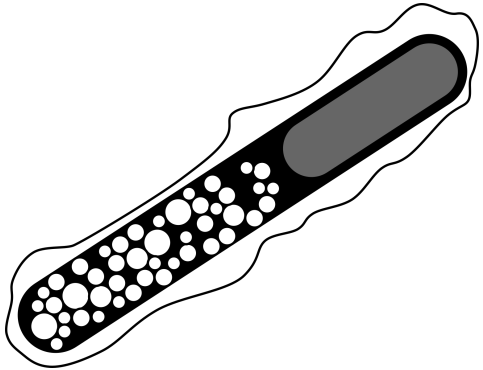


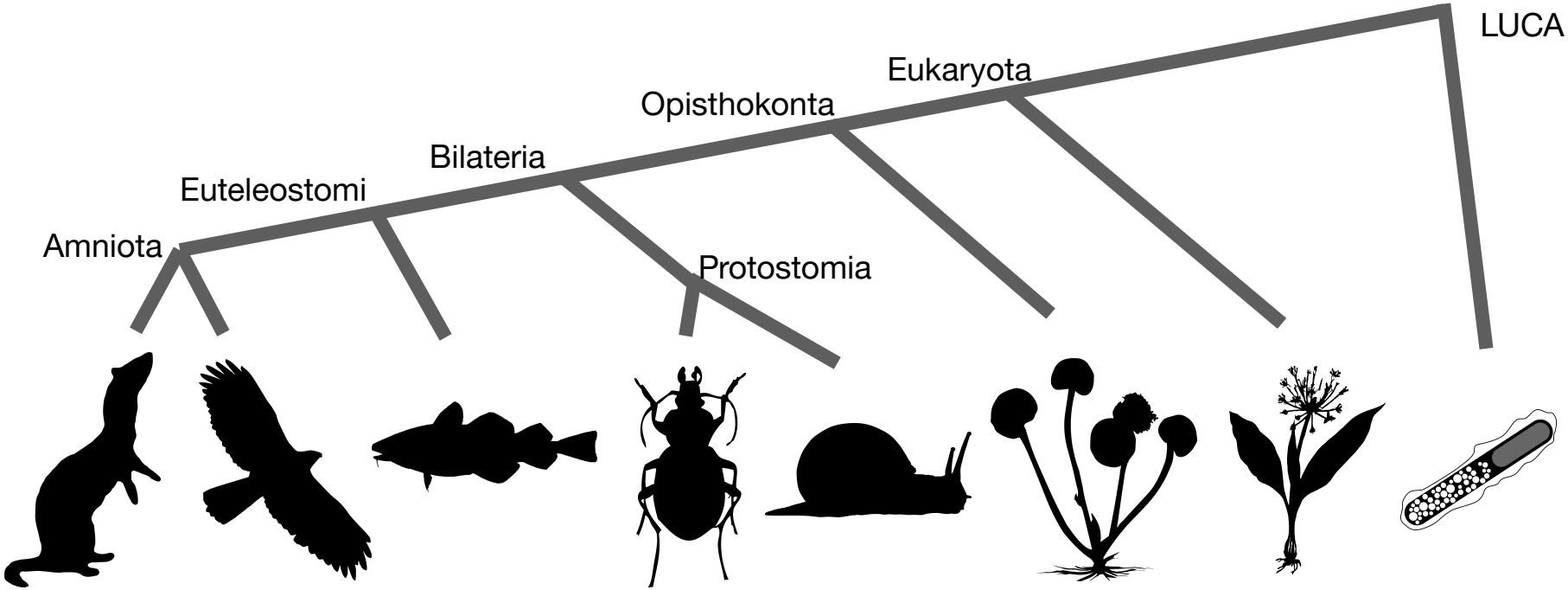


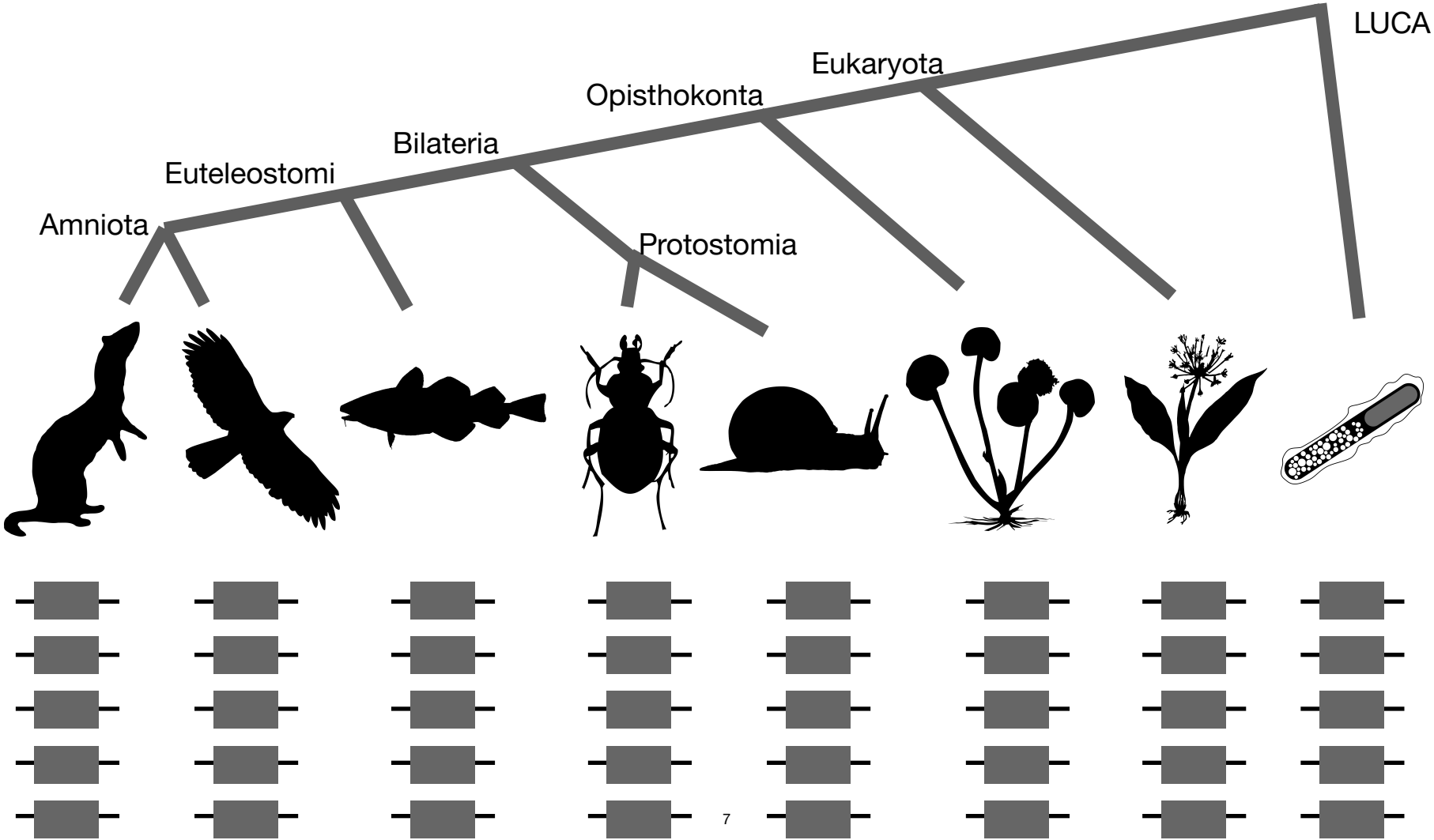


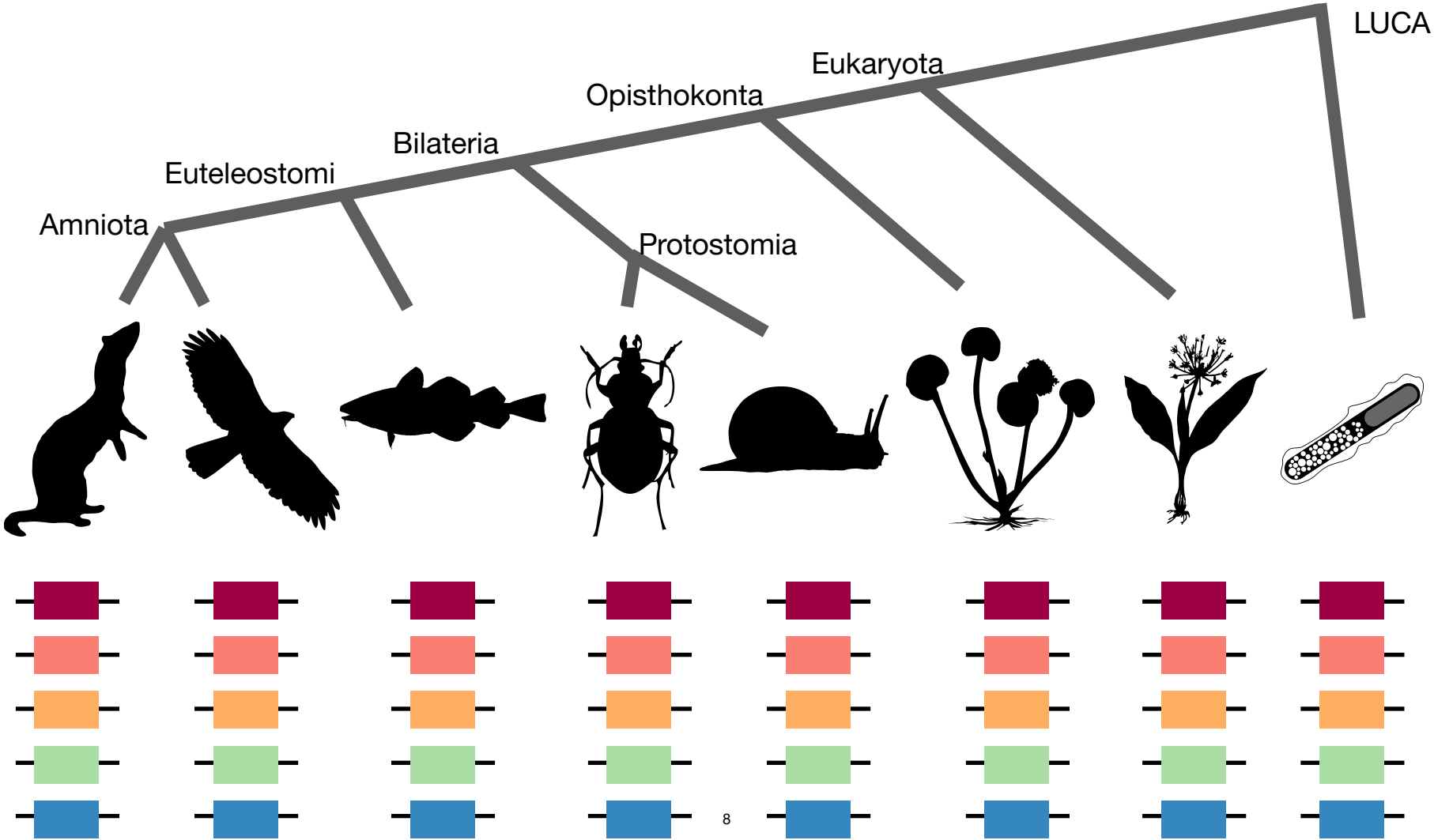


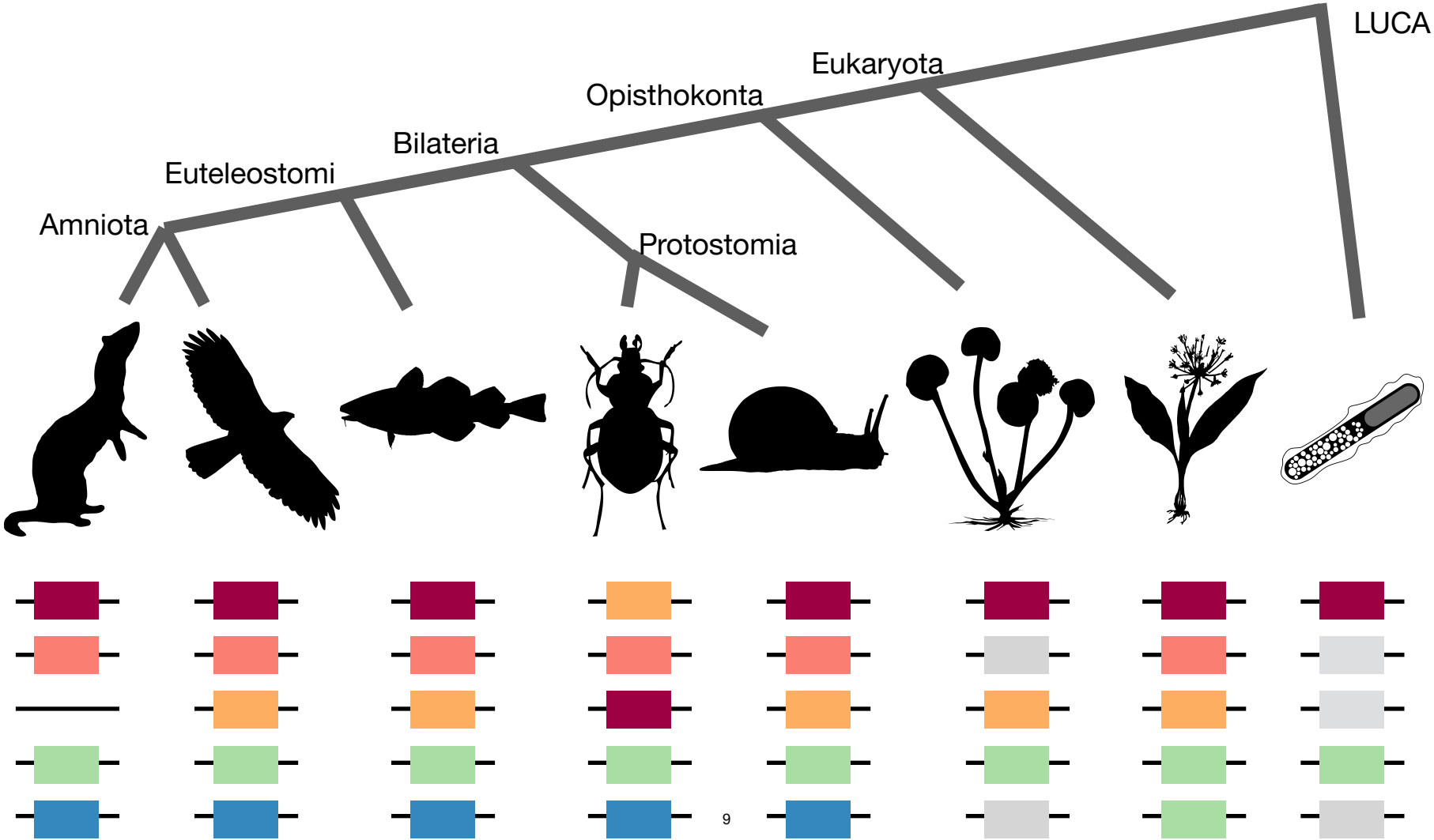





LUCA
Eukaryota
Opisthokonta
Bilateria
Euteleostomi
Amniota
Protostomia


LUCA
Eukaryota
Opisthokonta
Bilateria
Euteleostomi
Amniota
Protostomia

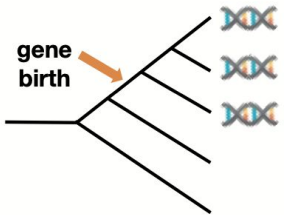

LUCA
Eukaryota
Opisthokonta
Bilateria
Euteleostomi
Amniota
Protostomia

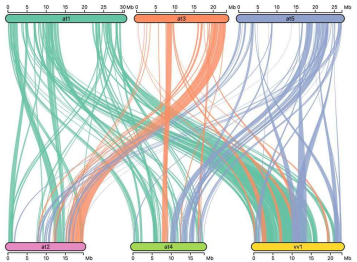

LUCA
Eukaryota
Opisthokonta
Bilateria
Euteleostomi
Amniota
Protostomia

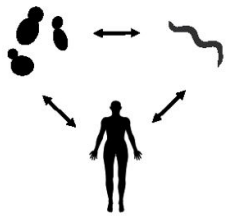
Orthology is fundamental to comparative genomics

Applications:

- Building phylogenetic trees
- Transferring gene function
- Finding lineage-specific genes
- Studying ancestral genomes
- Finding co-evolving genes
- Exploring synteny





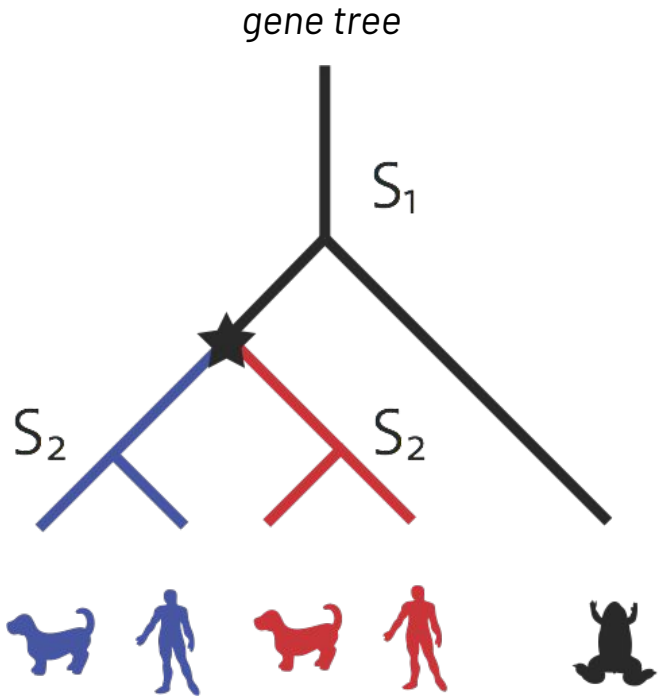


Orthology: What is it?

Homology

- The study of genetic material almost always starts with identifying, within or across species, **homologous regions**—regions of common ancestry.
- Homologs = **gene families**
- It is useful to distinguish between two classes of homologous genes.

Orthologs vs. paralogs



Orthologs vs. paralogs

- **Orthologs:** pairs of genes that started diverging via evolutionary speciation
- **Paralogs:** pairs of genes that started diverging via gene duplication

- **Orthology:** A relation between pairs of genes that started diverging via evolutionary speciation
- **Paralogy:** A relation between pairs of genes that started diverging via gene duplication

Ortho = exact
Para = beside/next to

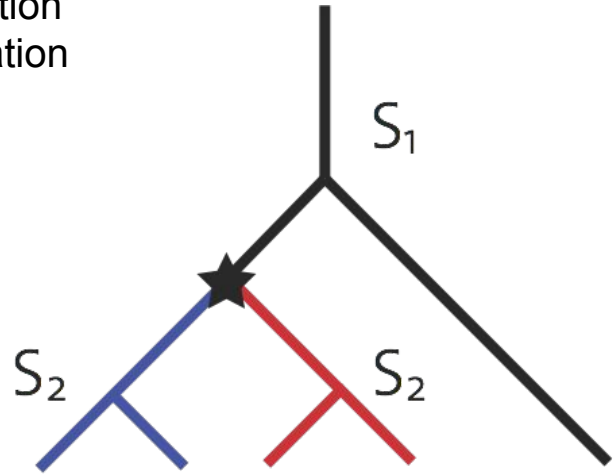
Orthologs

- Two genes in two species are orthologous if they derive from one gene in their last common ancestor

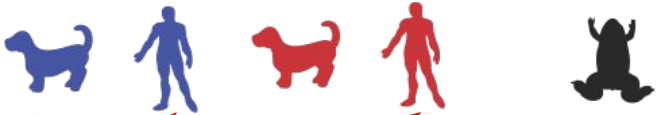
- Can be thought of as “corresponding genes” between species

gene tree

S = speciation
☆ = duplication



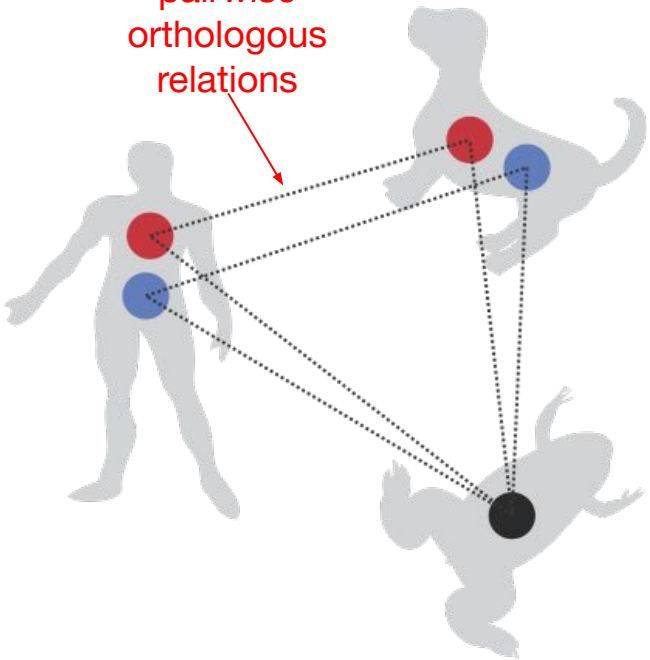
these are genes



orthologs

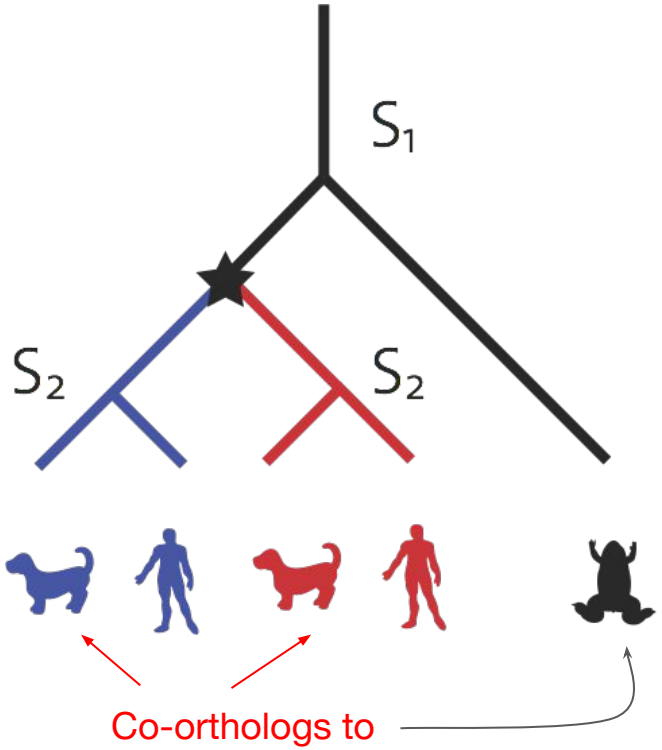
paralogs

pairwise orthologous relations



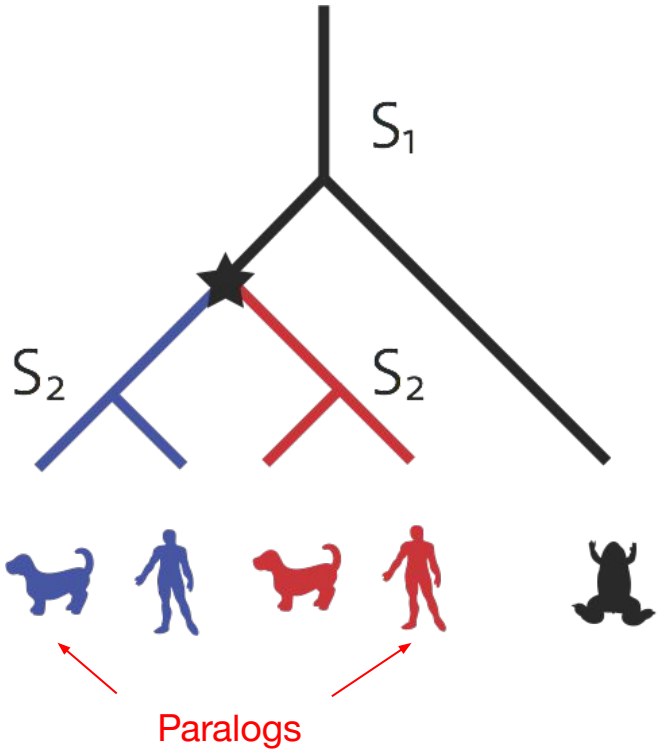
Common misconceptions

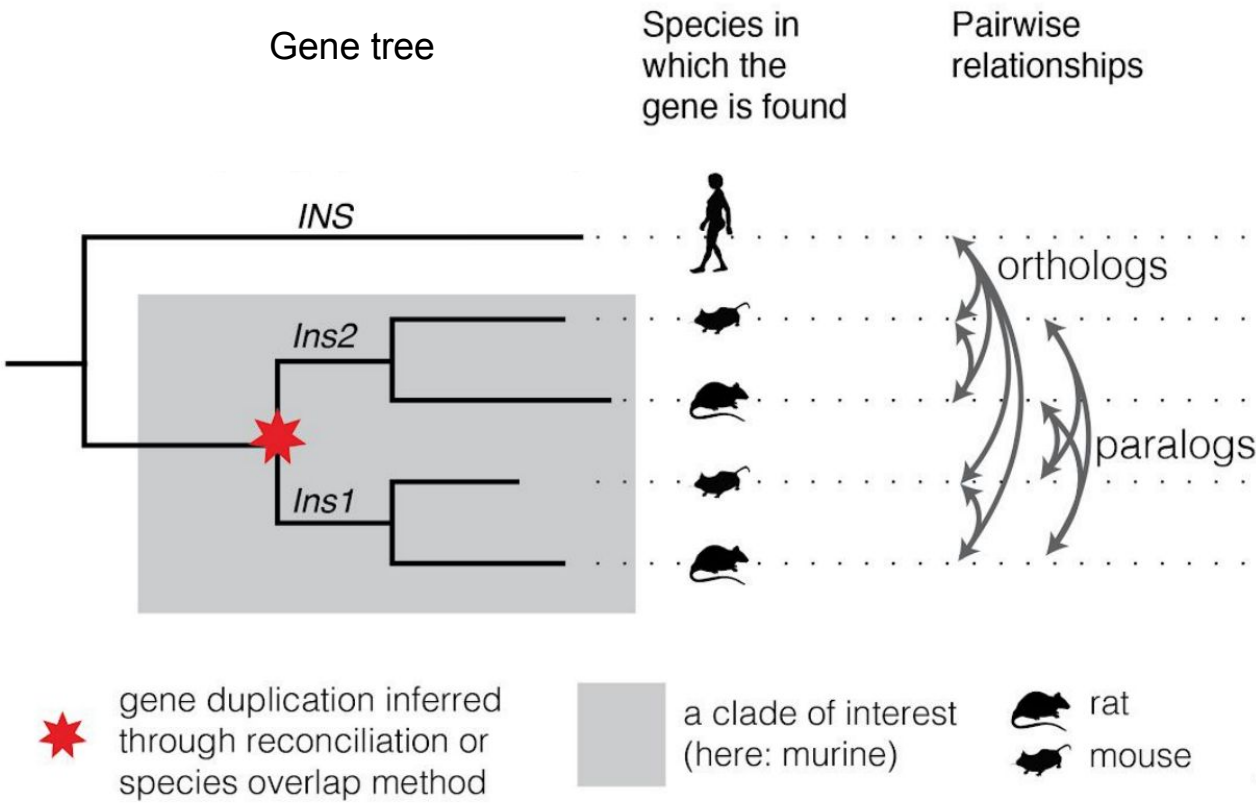
- Orthologs don't have to be one-to-one
- Orthology can also be a one-to-many, many-to-one, or many-to-many relationship
- Paralogs don't have to be in the same species



Common misconceptions

- Orthologs don't have to be one-to-one
- Orthology can also be a one-to-many, many-to-one, or many-to-many relationship
- Paralogs don't have to be in the same species

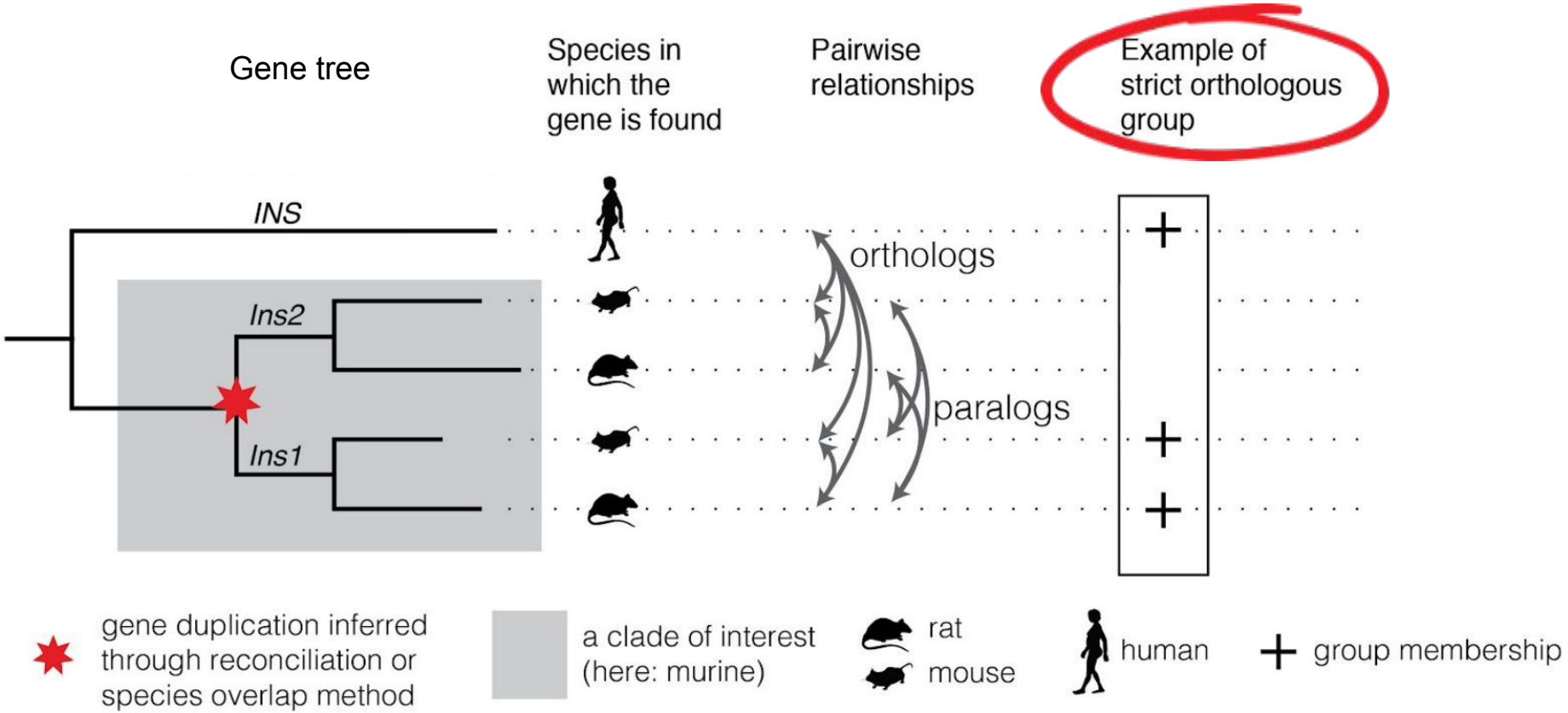




Difficult to interpret pairwise relationships when referring to groups of genes in several species...

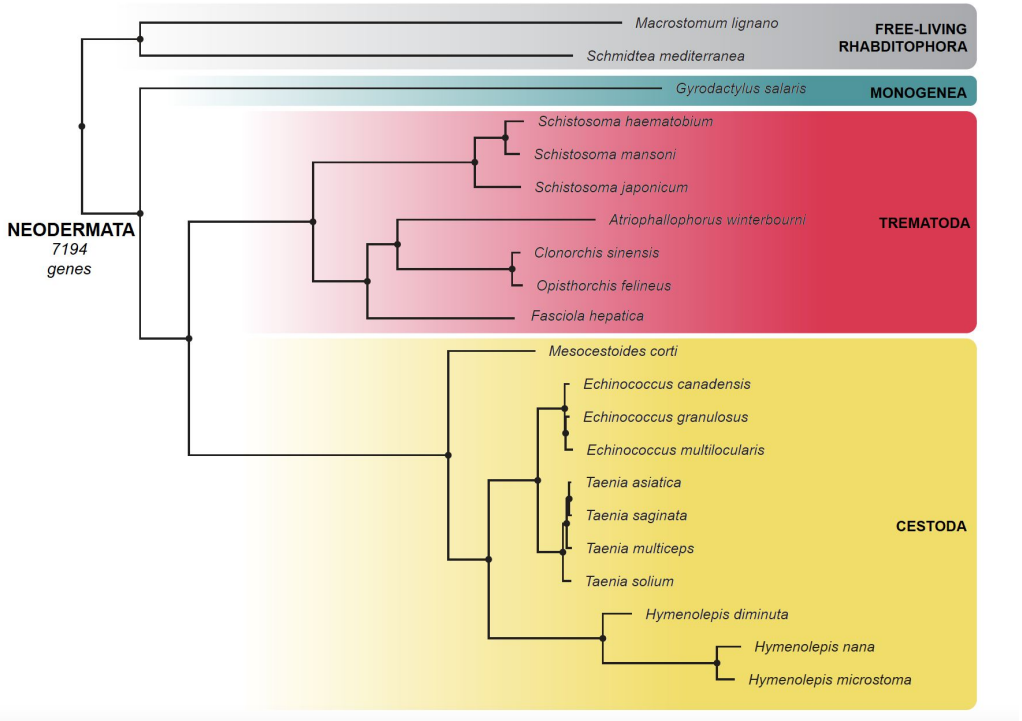
Orthologous Groups (i.e. OMA Groups/OGs)

- Groups of genes which are all orthologous to each other
- Strict orthologous group
- Not necessarily 1:1 orthology, but each group contains at most one gene per species


Gene tree
Species in which the gene is found
Pairwise relationships
Example of strict orthologous group
INS
Ins2
Ins1
orthologs
paralogs
gene duplication inferred through reconciliation or species overlap method
a clade of interest (here: murine)
rat
mouse
human
group membership

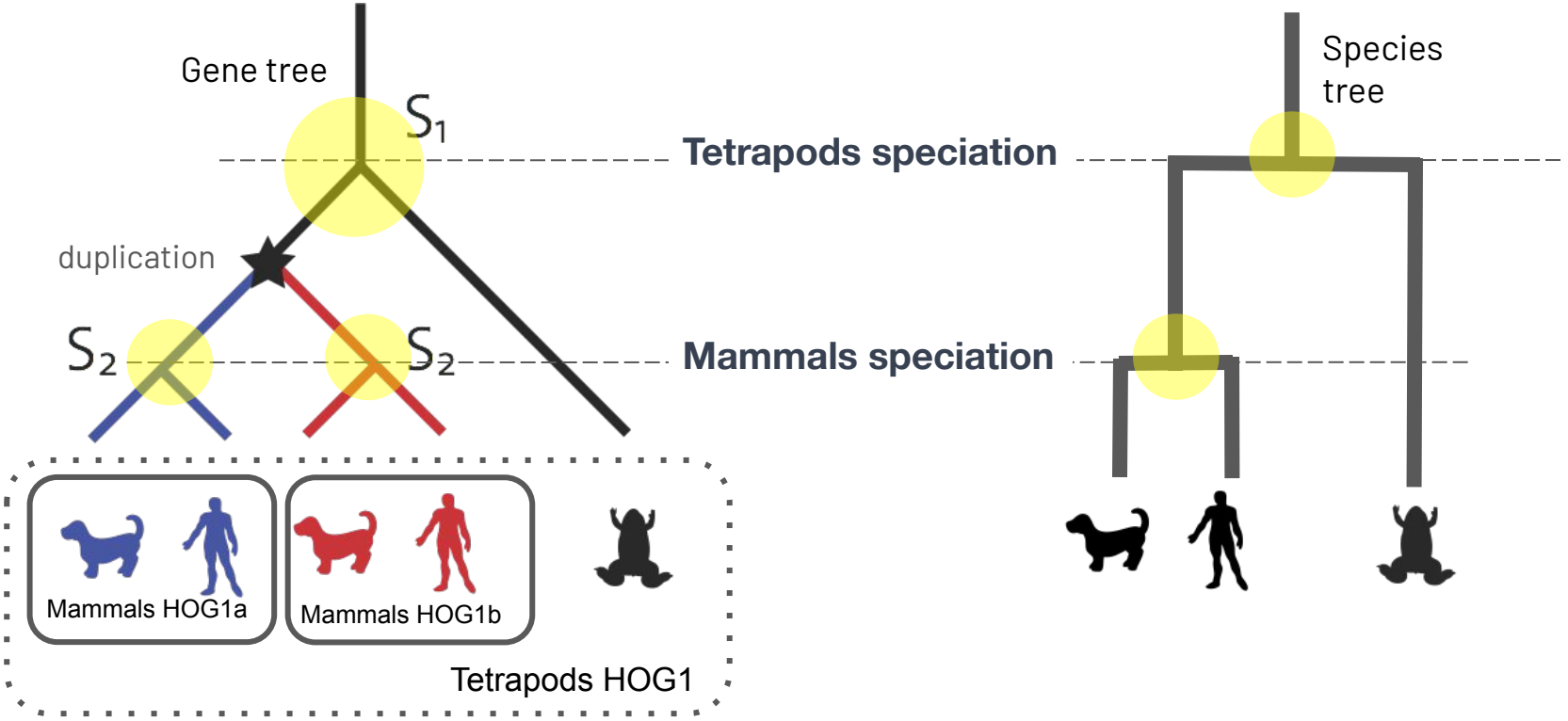
Making species trees

Create species trees for a clade of interest using the genes in the OMA Group (strict Orthologous Group) at that taxonomic level

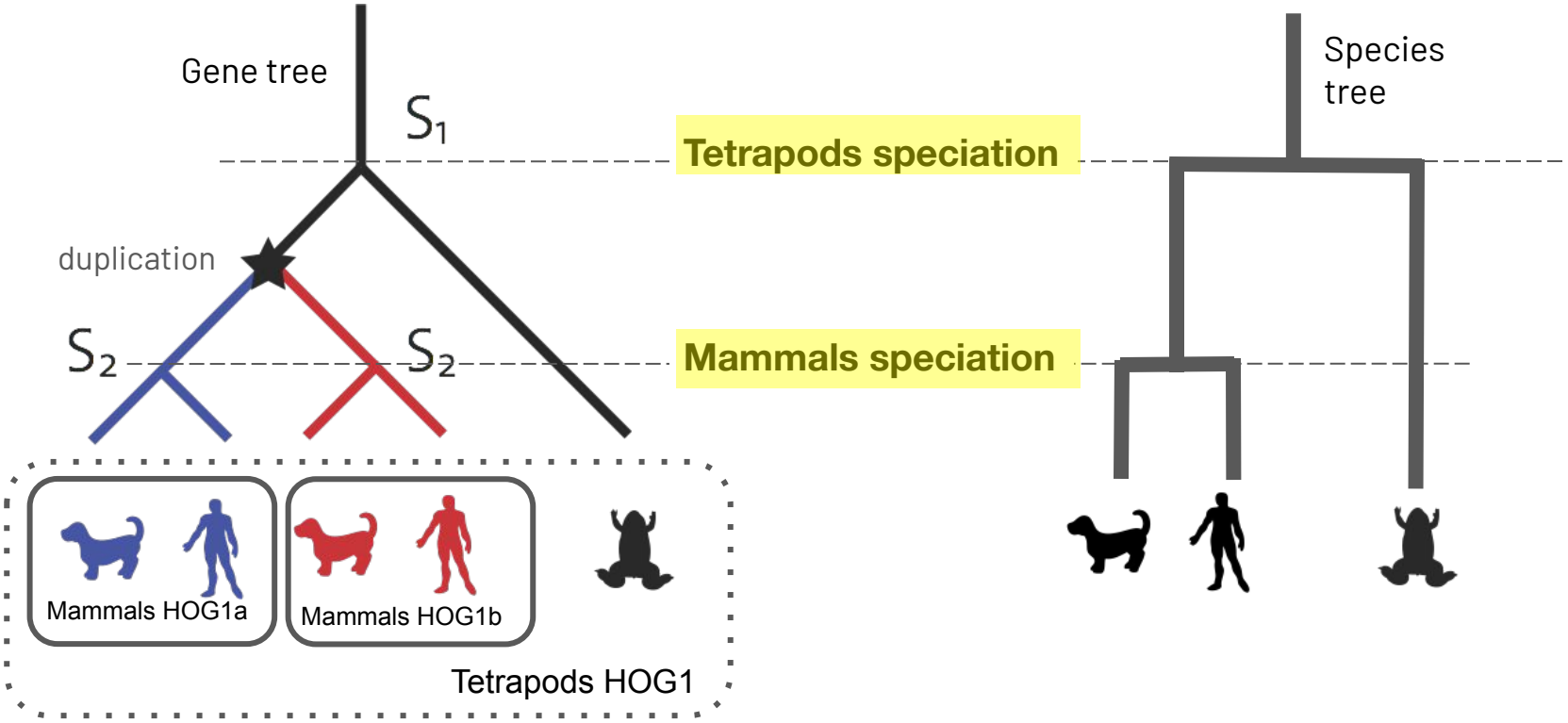


Hierarchical Orthologous Groups (HOGs)

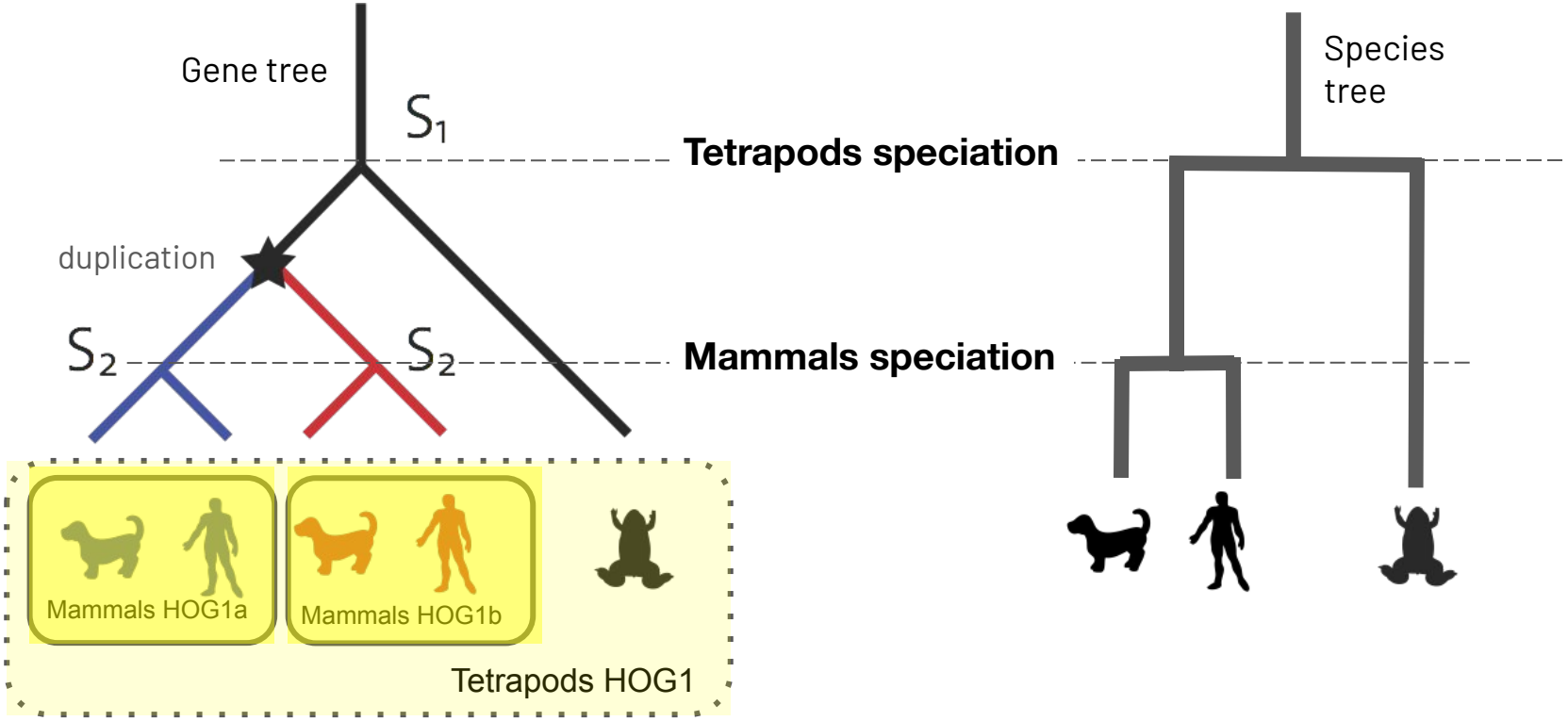
HOGs = Sets of genes that descended from a **common ancestral gene** in a given ancestral species



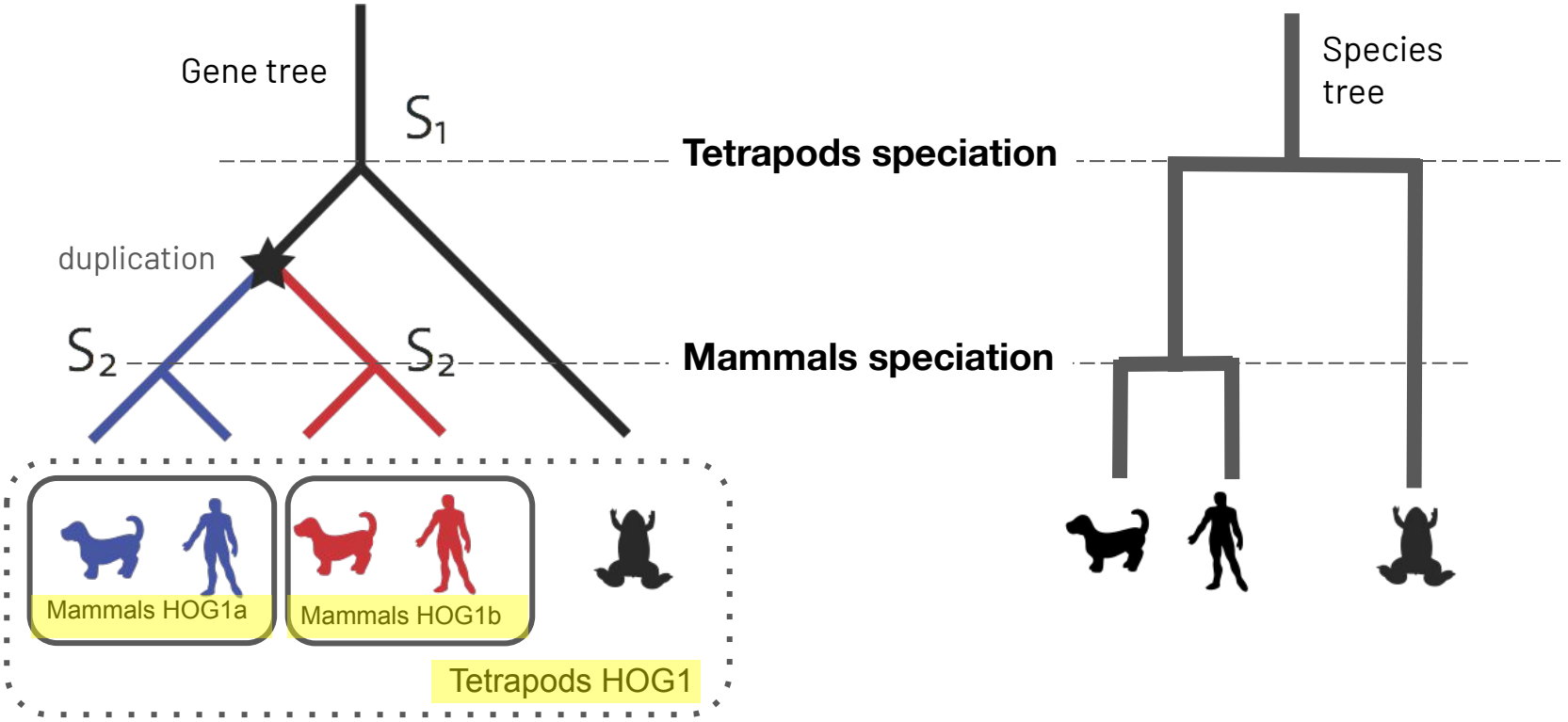
HOGs are defined with respect to **specific clades**

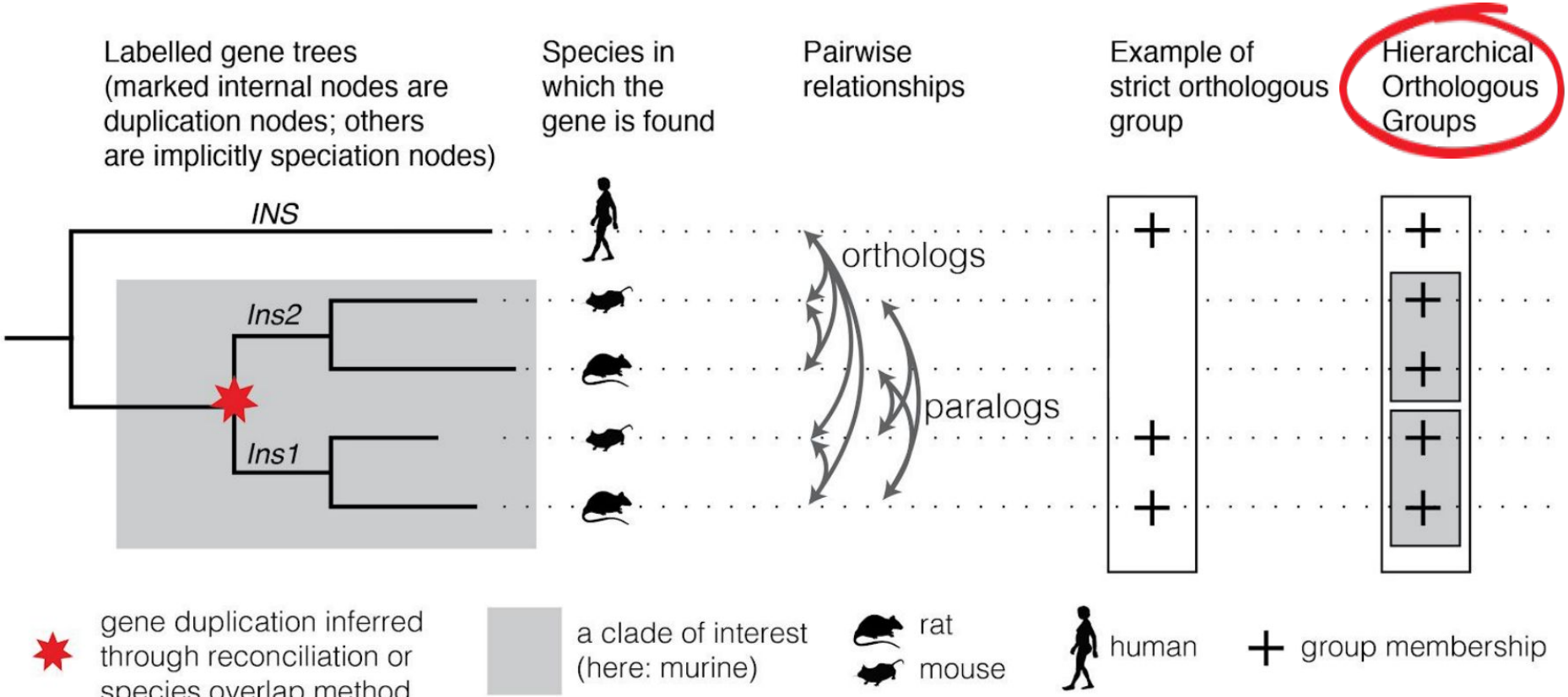


HOGs are **hierarchical** because the groups are defined with respect to deeper clades that **subsume** multiple groups defined on their descendants



HOGs are **gene families**; SubHOGs are nested subfamilies




Labelled gene trees (marked internal nodes are duplication nodes; others are implicitly speciation nodes)
Species in which the gene is found
Pairwise relationships
Example of strict orthologous group
Hierarchical Orthologous Groups
INS
Ins2
Ins1
orthologs
paralogs
gene duplication inferred through reconciliation or species overlap method
a clade of interest (here: murine)
rat
mouse
human
group membership

How to infer orthology?

Orthology Databases

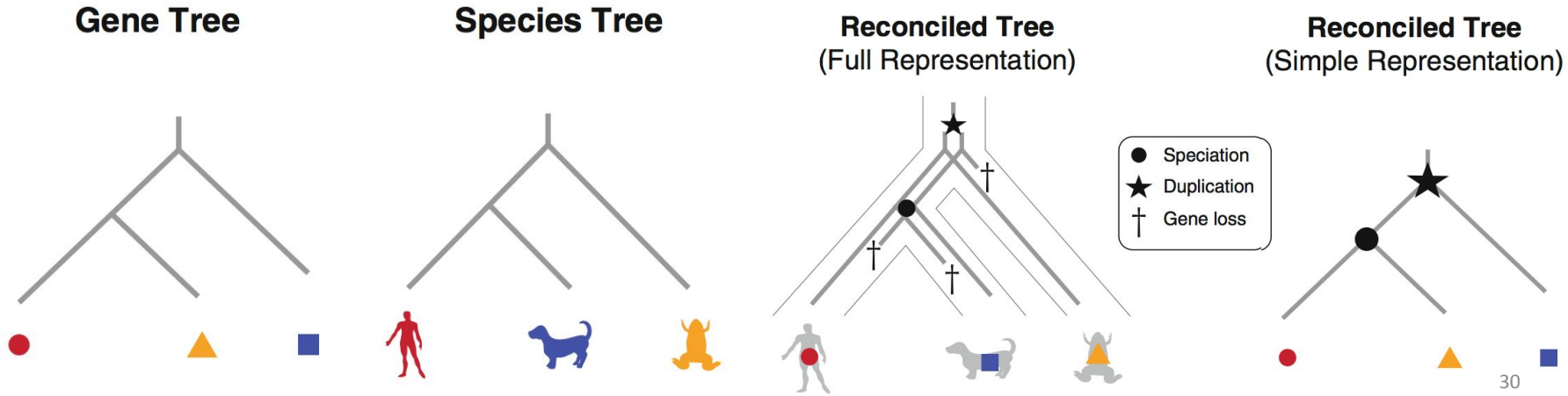
- OMA: https://omabrowser.org/
- OrthoDB: https://www.orthodb.org/
- EggNOG: http://eggnogdb.embl.de
- InParanoid/Hieranoid: http://hieranoidb.sbc.su.se/
- COG: https://www.ncbi.nlm.nih.gov/COG/
- PANTHER: http://www.pantherdb.org/genes/
- PhylomeDB: http://phylomedb.org/
- And more!

Many different methods:
Namely tree-based or graph-based

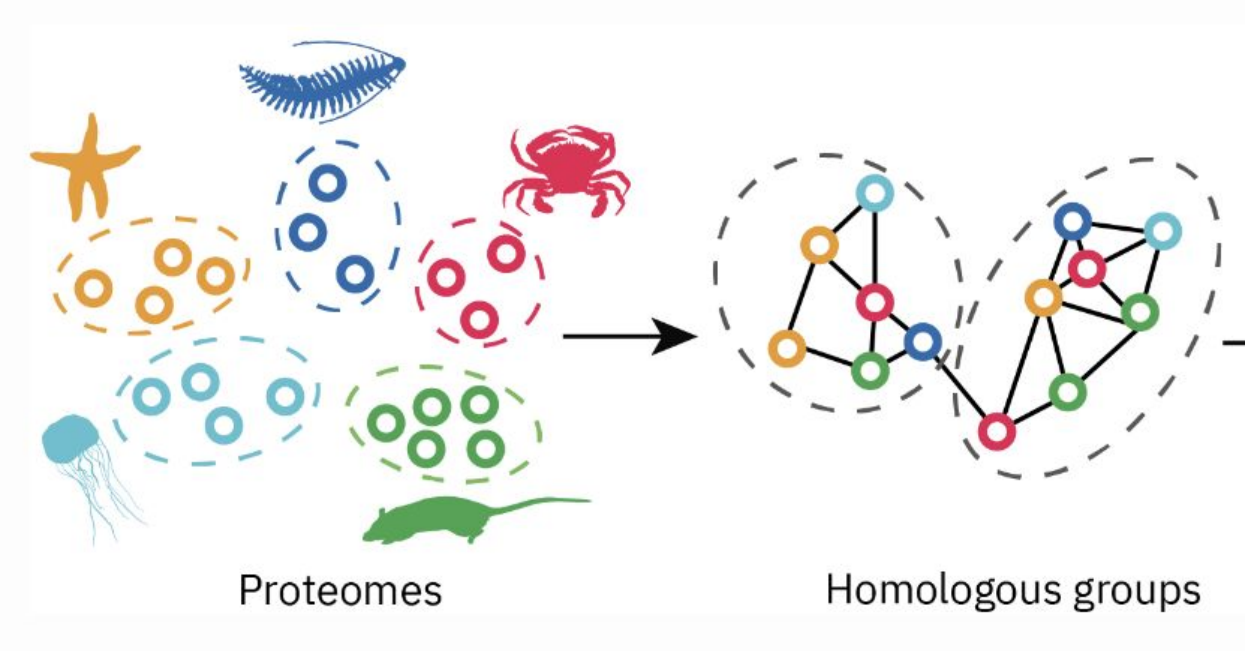
Tree-based methods

- Start with a group of homologous sequences
- Reconstruct a gene tree
- Infer the type of evolutionary event represented by each internal node of the tree

Gene tree/species tree reconciliation:



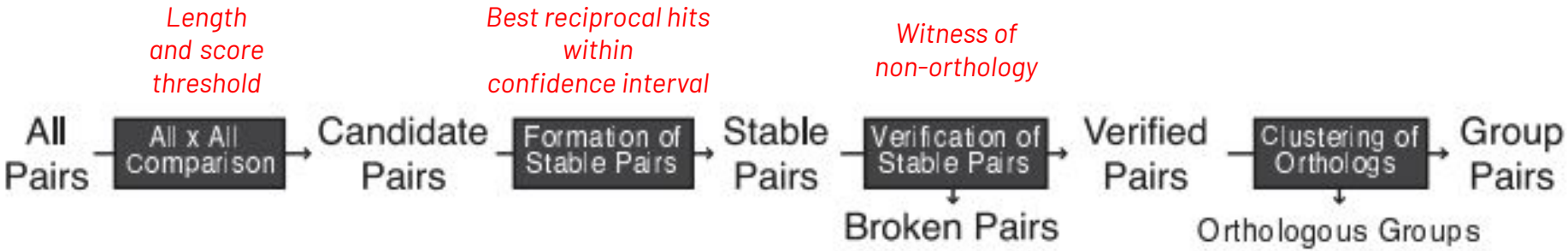
Graph-based methods



- Compare pairs of genes within and between species
- For pairs of genes between 2 species, orthologs tend to be the pairs of sequences that have diverged the least

Oma browser

Pipeline:



Oma browser — orthologous matrix

All protein sequences from >2800 full genomes

(22 million)

all vs. all comparison

(100 trillion alignments)

~12 million CPU hours

Sequences of shared ancestry

identify closest pairs

Orthologous pairs

(tens of billions)

GETHOGs algorithm

Hierarchical Orthologous Groups

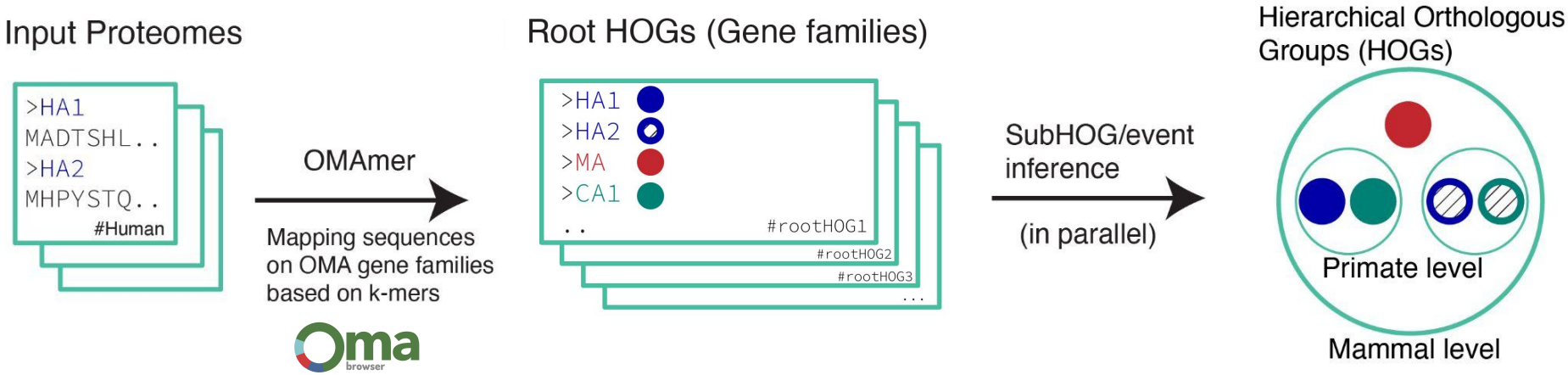
(hundreds of thousands)

- Browser (interactive)
- REST API / python+R libraries
- Flat files
- SPARQL endpoint

Dessimoz et al., RECOMB CG 2005
Dessimoz et al., Nucl Acids Res 2006
Roth et al., BMC Bioinformatics 2008
Altenhoff et al, Nucl Acids Res 2011
Altenhoff et al. PLoS ONE 2013
Altenhoff et al. NAR 2015
Train et al. Bioinformatics / ISMB 2017

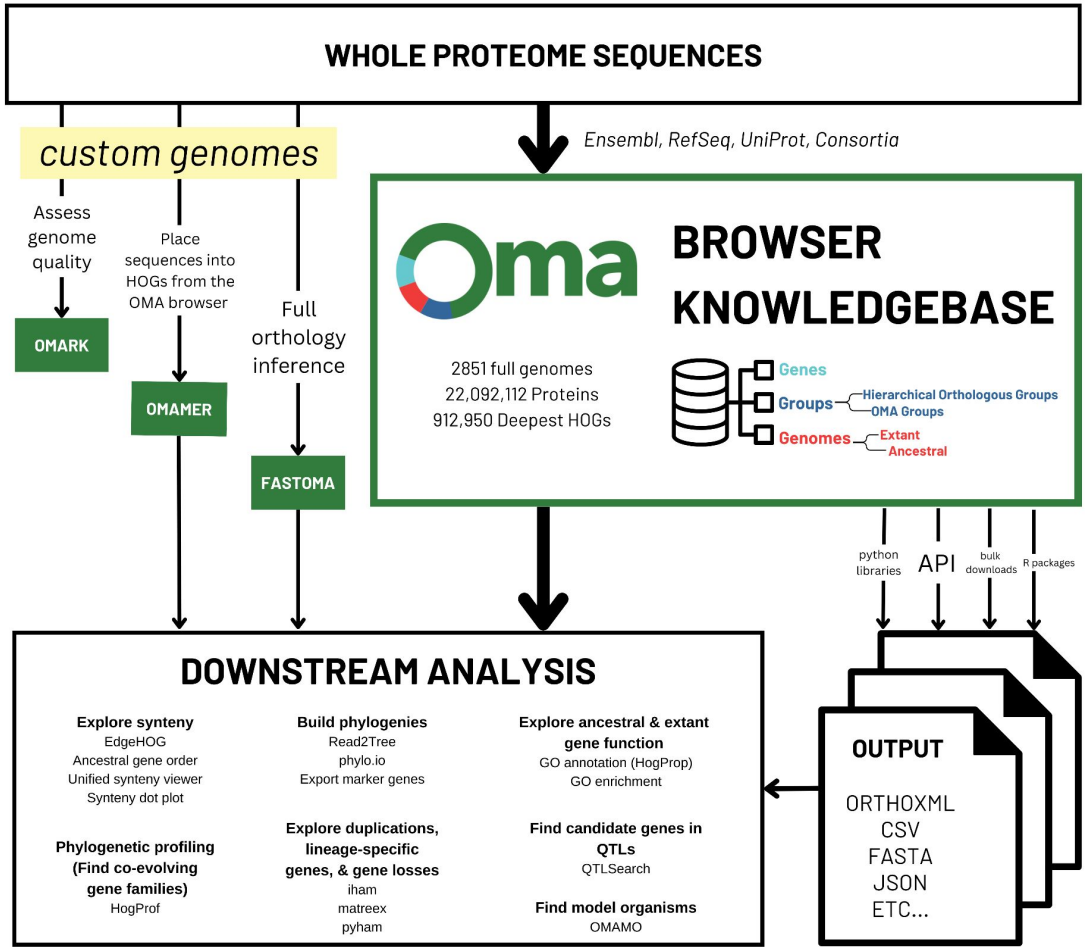
Altenhoff et al. NAR 2018
Altenhoff et al. Genome Research 2019
Zahn-Zabal et al. F1000Res 2020
Glover F1000Res 2020
Altenhoff et al. NAR 2021
Altenhoff et al. NAR 2024

FastOMA, our new tool

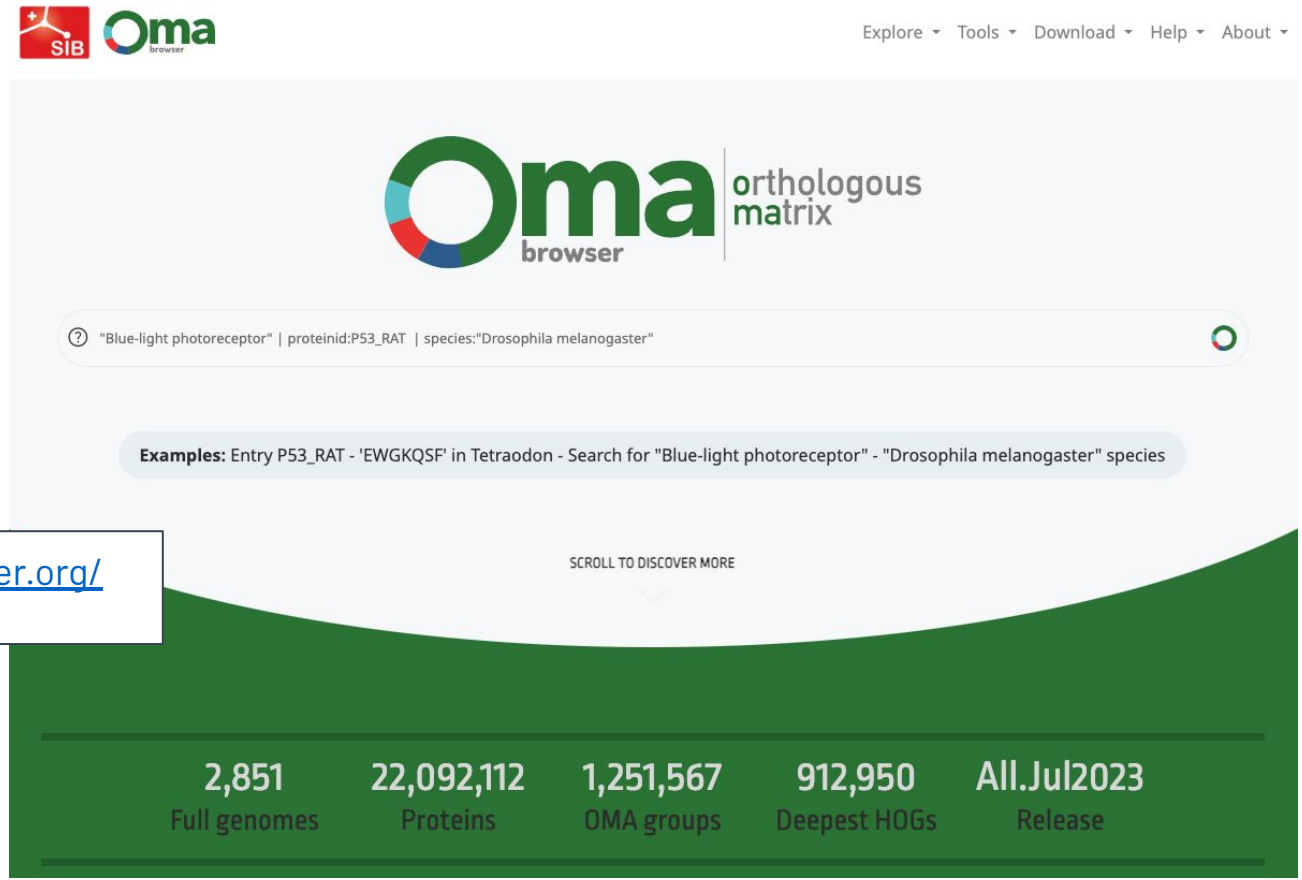


The OMA Ecosystem

From public and private genomes, orthology inference, a database of over 2800 species, and a variety of downstream analyses for comparative genomics

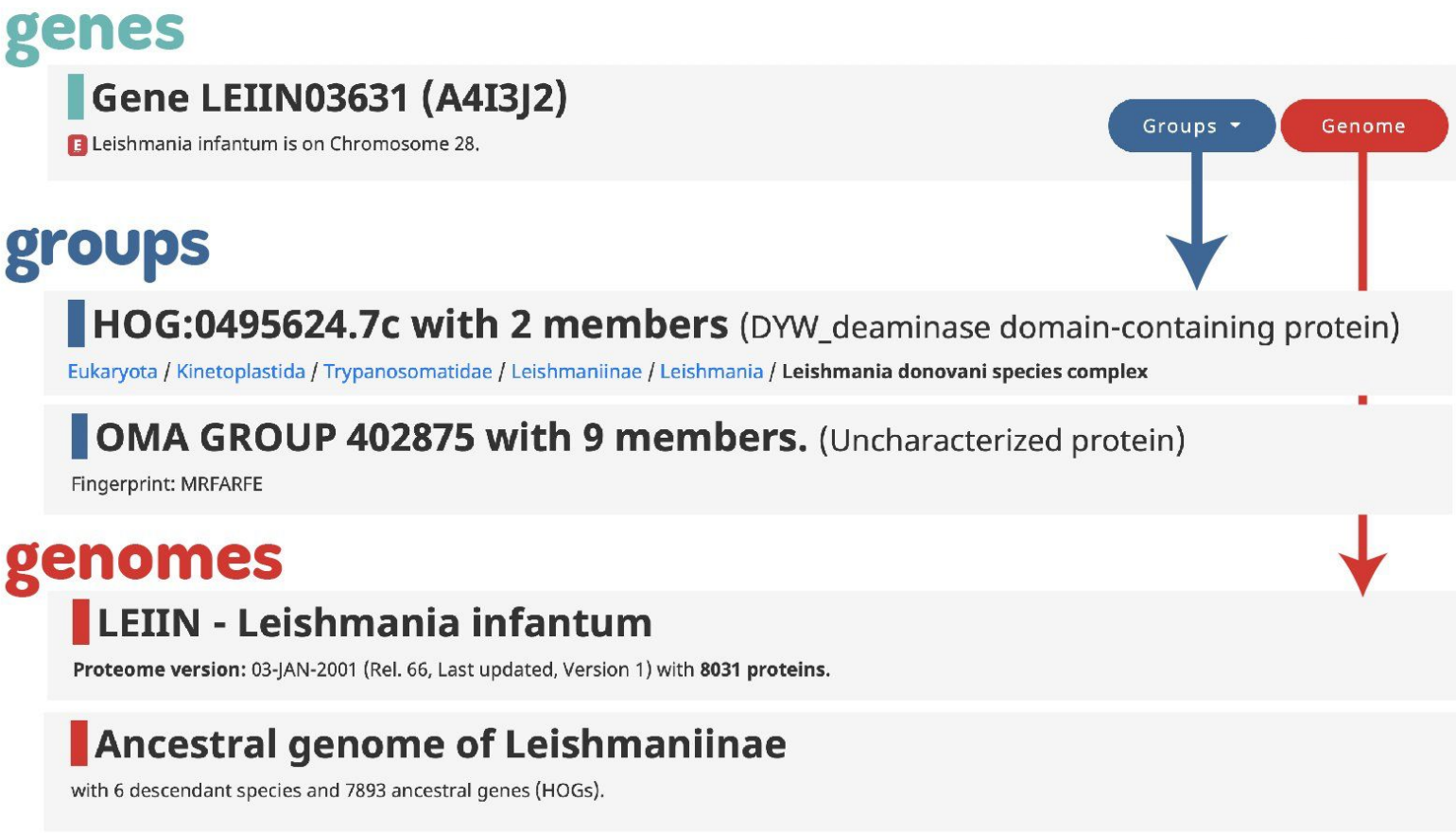

WHOLE PROTEOME SEQUENCES
custom genomes
Ensembl, RefSeq, UniProt, Consortia
Assess genome quality
Place sequences into HOGs from the OMA browser
Full orthology inference
OMARK
OMAMER
FASTOMA
Oma
BROWSER KNOWLEDGEBASE
2851 full genomes
22,092,112 Proteins
912,950 Deepest HOGs
Genes
Groups
Hierarchical Orthologous Groups
OMA Groups
Genomes
Extant
Ancestral
python libraries
API
bulk downloads
R packages
DOWNSTREAM ANALYSIS
Explore synteny
EdgeHOG
Ancestral gene order
Unified synteny viewer
Synteny dot plot
Phylogenetic profiling (Find co-evolving gene families)
HogProf
Build phylogenies
Read2Tree
phylo.io
Export marker genes
Explore duplications, lineage-specific genes, & gene losses
iham
matreex
pyham
Explore ancestral & extant gene function
GO annotation (HogProp)
GO enrichment
Find candidate genes in QTLs
QTLSearch
Find model organisms
OMAMO
OUTPUT
ORTHOXML
CSV
FASTA
JSON
ETC...

The OMA browser

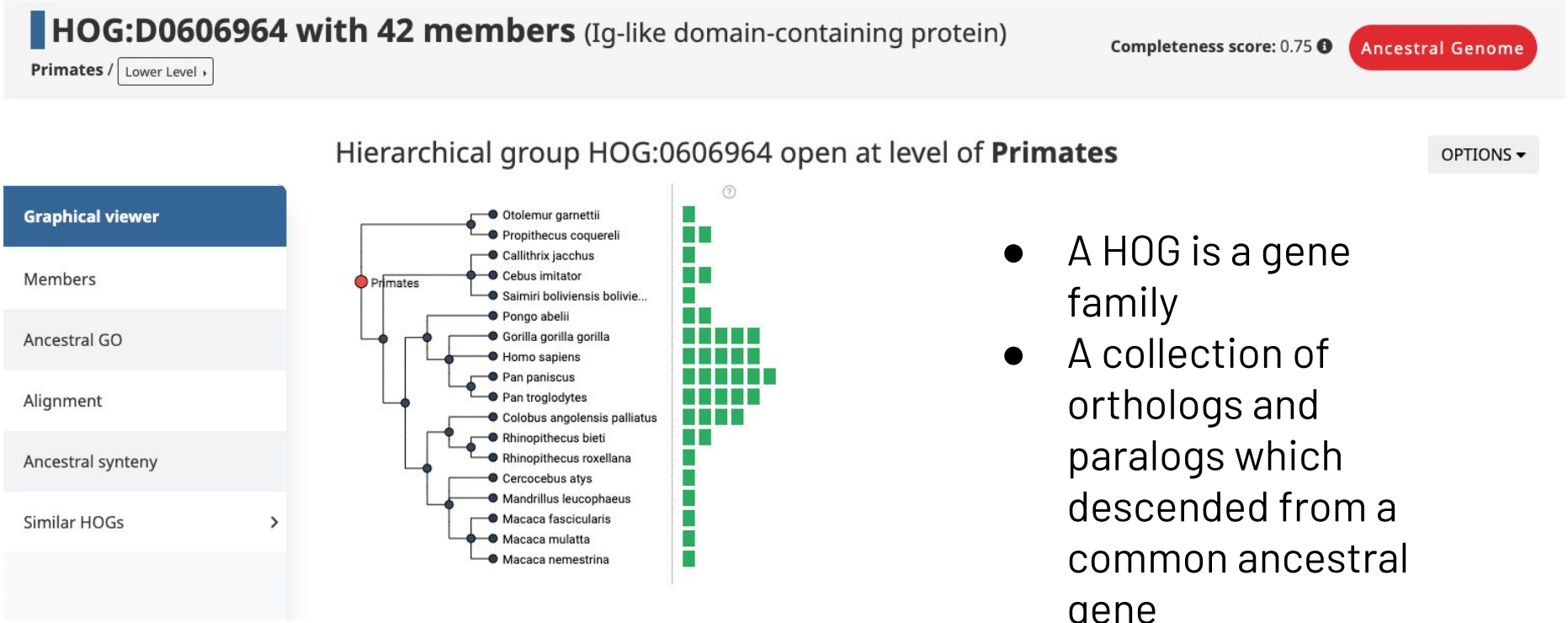


https://omabrowser.org/

The OMA Database: 3 kinds of pages

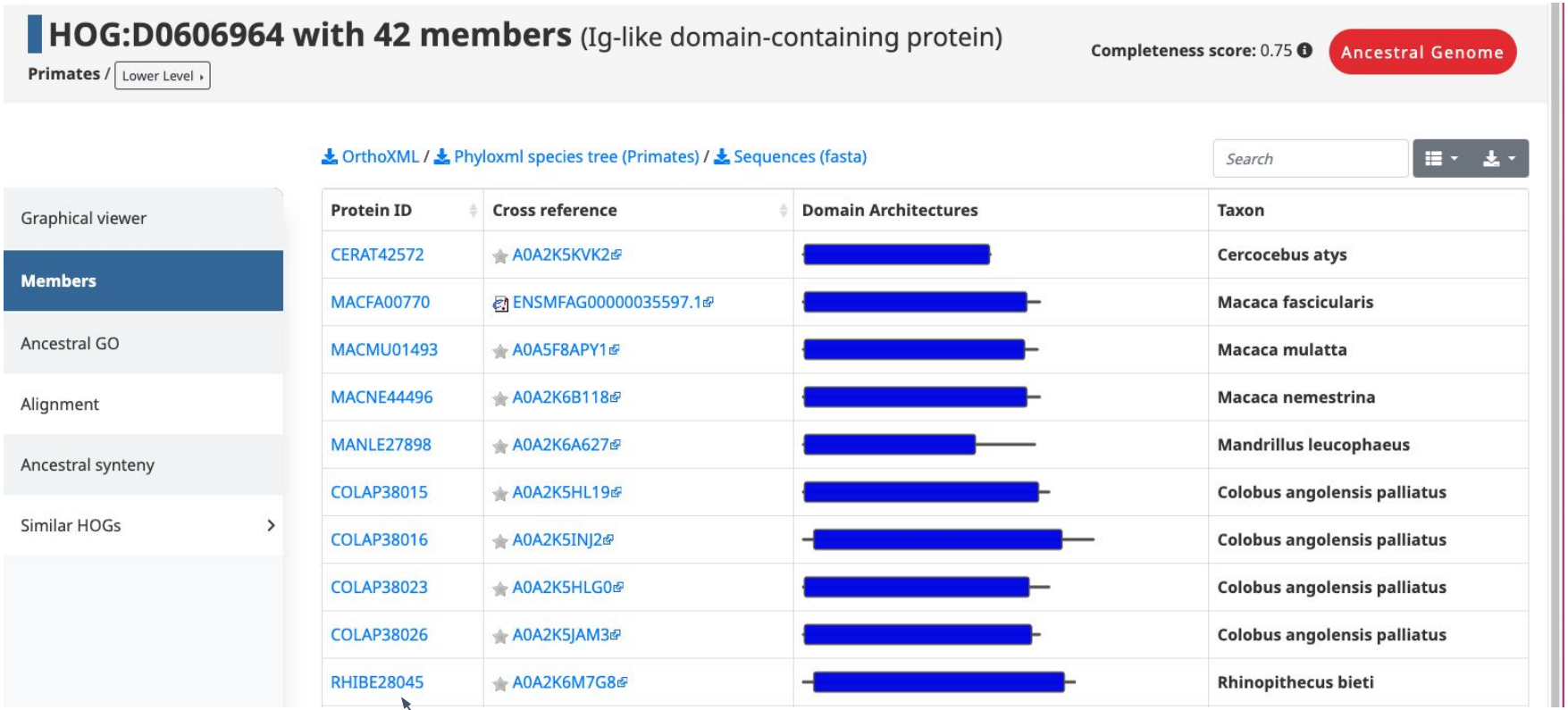


Hierarchical Orthologous Groups (HOGs)



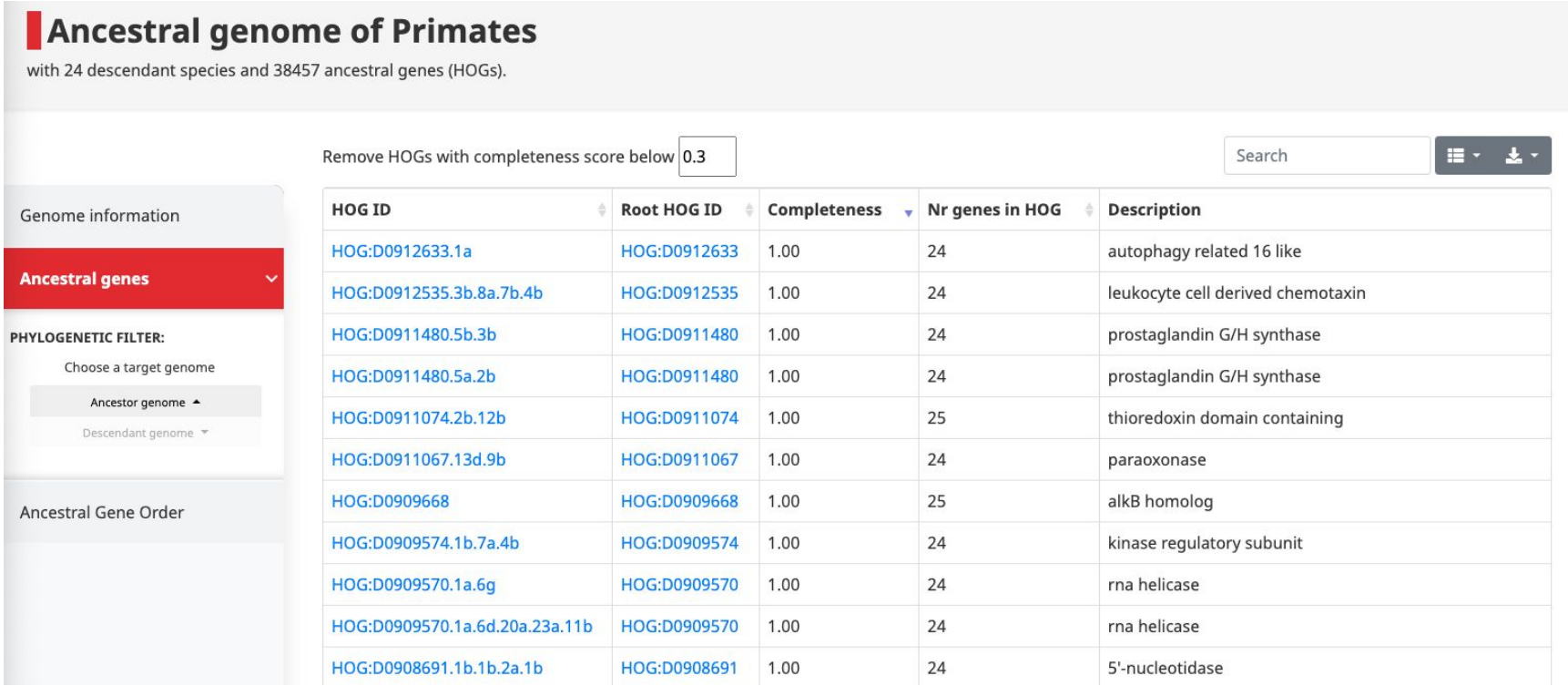
- A HOG is a gene family
- A collection of orthologs and paralogs which descended from a common ancestral gene

HOGs


HOG:D0606964 with 42 members (Ig-like domain-containing protein)
Primates / Lower Level
Completeness score: 0.75
Ancestral Genome
OrthoXML / Phyloxml species tree (Primates) / Sequences (fasta)
Search
Graphical viewer
Members
Ancestral GO
Alignment
Ancestral synteny
Similar HOGs
Protein ID
Cross reference
Domain Architectures
Taxon
CERAT42572 A0A2K5KVK2 Cercocebus atys
MACFA00770 ENSMFAG00000035597.1 Macaca fascicularis
MACMU01493 A0A5F8APY1 Macaca mulatta
MACNE44496 A0A2K6B118 Macaca nemestrina
MANLE27898 A0A2K6A627 Mandrillus leucophaeus
COLAP38015 A0A2K5HL19 Colobus angolensis palliatus
COLAP38016 A0A2K5INJ2 Colobus angolensis palliatus
COLAP38023 A0A2K5HLG0 Colobus angolensis palliatus
COLAP38026 A0A2K5JAM3 Colobus angolensis palliatus
RHIBE28045 A0A2K6M7G8 Rhinopithecus bieti
OMA ID

Ancestral genomes

The collection of HOGs at a given taxonomic level

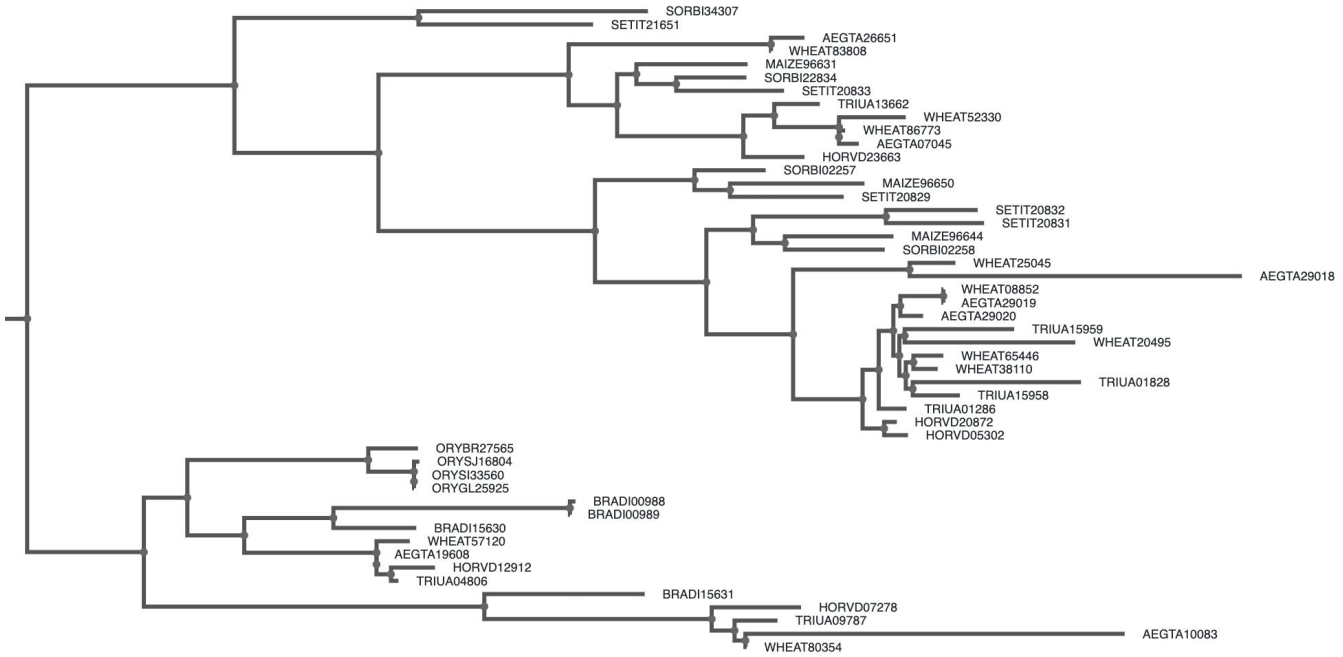


HOG ID	Root HOG ID	Completeness	Nr genes in HOG	Description
HOG:D0912633.1a	HOG:D0912633	1.00	24	autophagy related 16 like
HOG:D0912535.3b.8a.7b.4b	HOG:D0912535	1.00	24	leukocyte cell derived chemotaxin
HOG:D0911480.5b.3b	HOG:D0911480	1.00	24	prostaglandin G/H synthase
HOG:D0911480.5a.2b	HOG:D0911480	1.00	24	prostaglandin G/H synthase
HOG:D0911074.2b.12b	HOG:D0911074	1.00	25	thioredoxin domain containing
HOG:D0911067.13d.9b	HOG:D0911067	1.00	24	paraoxonase
HOG:D0909668	HOG:D0909668	1.00	25	alkB homolog
HOG:D0909574.1b.7a.4b	HOG:D0909574	1.00	24	kinase regulatory subunit
HOG:D0909570.1a.6g	HOG:D0909570	1.00	24	rna helicase
HOG:D0909570.1a.6d.20a.23a.11b	HOG:D0909570	1.00	24	rna helicase
HOG:D0908691.1b.1b.2a.1b	HOG:D0908691	1.00	24	5'-nucleotidase

OMA’s Downstream analyses

Make gene trees*

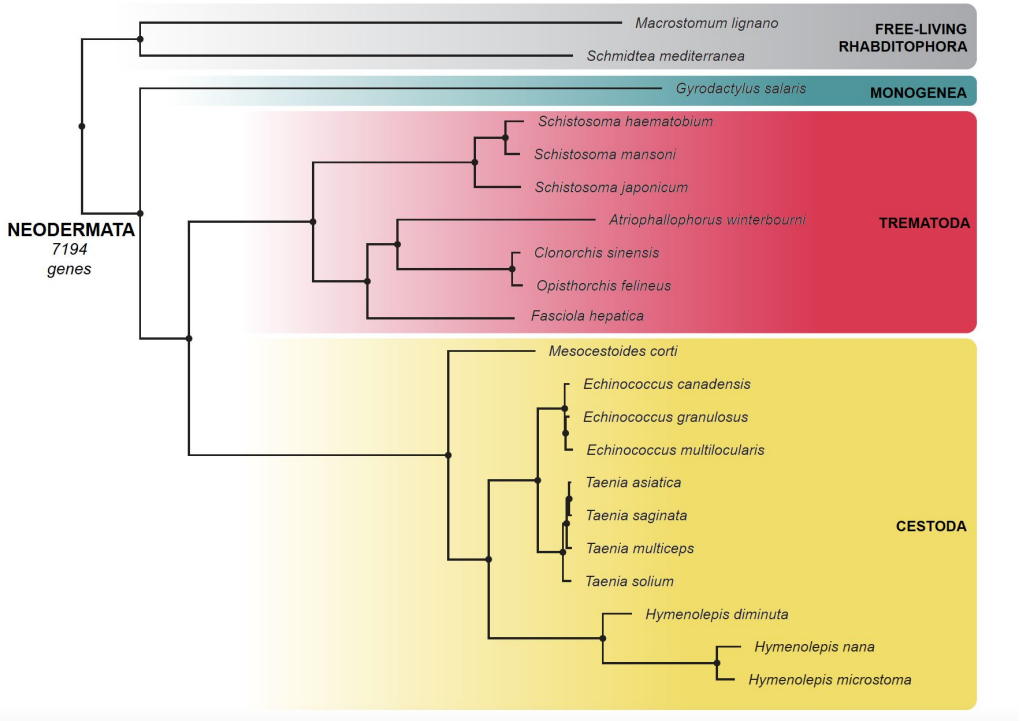
Create gene trees for a clade of interest using the genes in the HOG at that taxonomic level



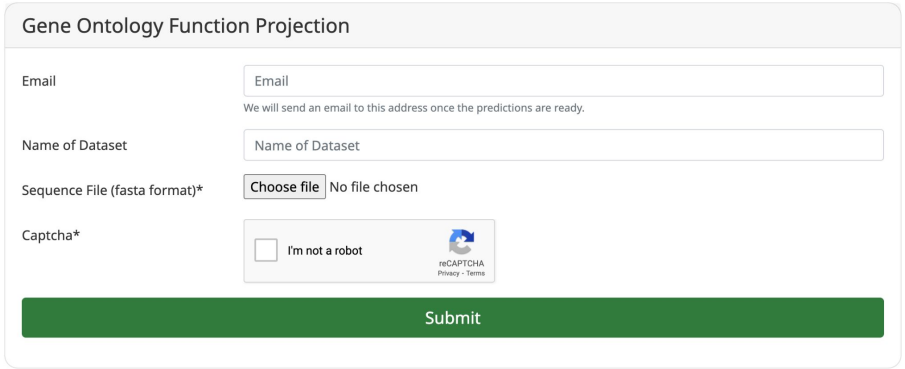
*Uses external software outside of the OMA Ecosystem

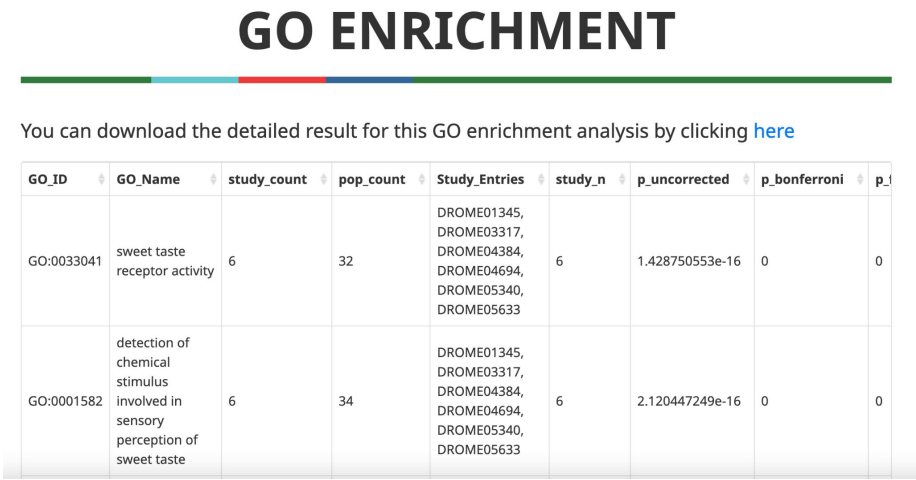
Making species trees*

Create species trees for a clade of interest using the genes in the OMA Group (strict Orthologous Group) at that taxonomic level



Infer gene function and perform GO enrichment analysis

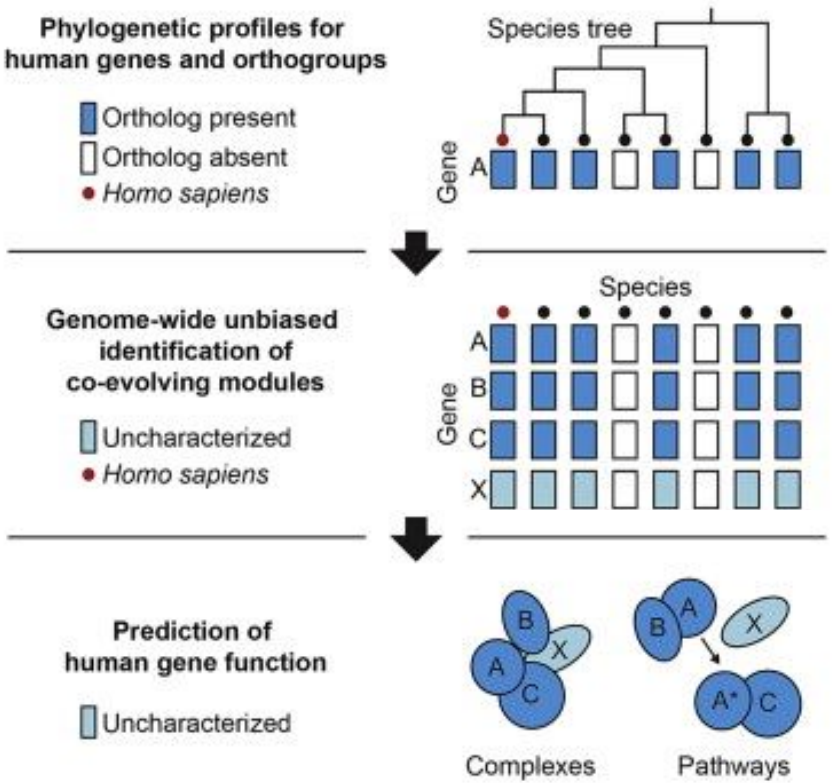




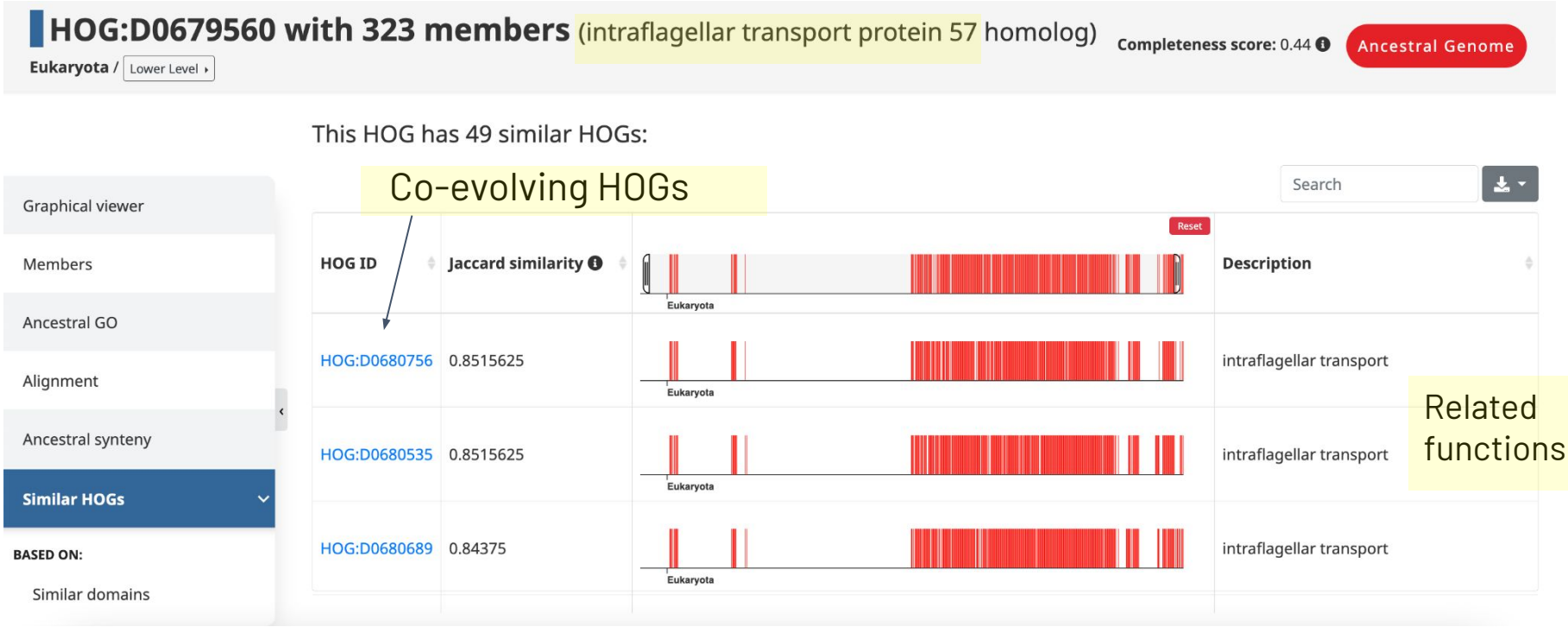
Transfer Gene Ontology annotations from genes in the same HOG from well-studied model species to other species

Perform a GO enrichment analysis to find overrepresented functions in a set of extant or ancestral genes

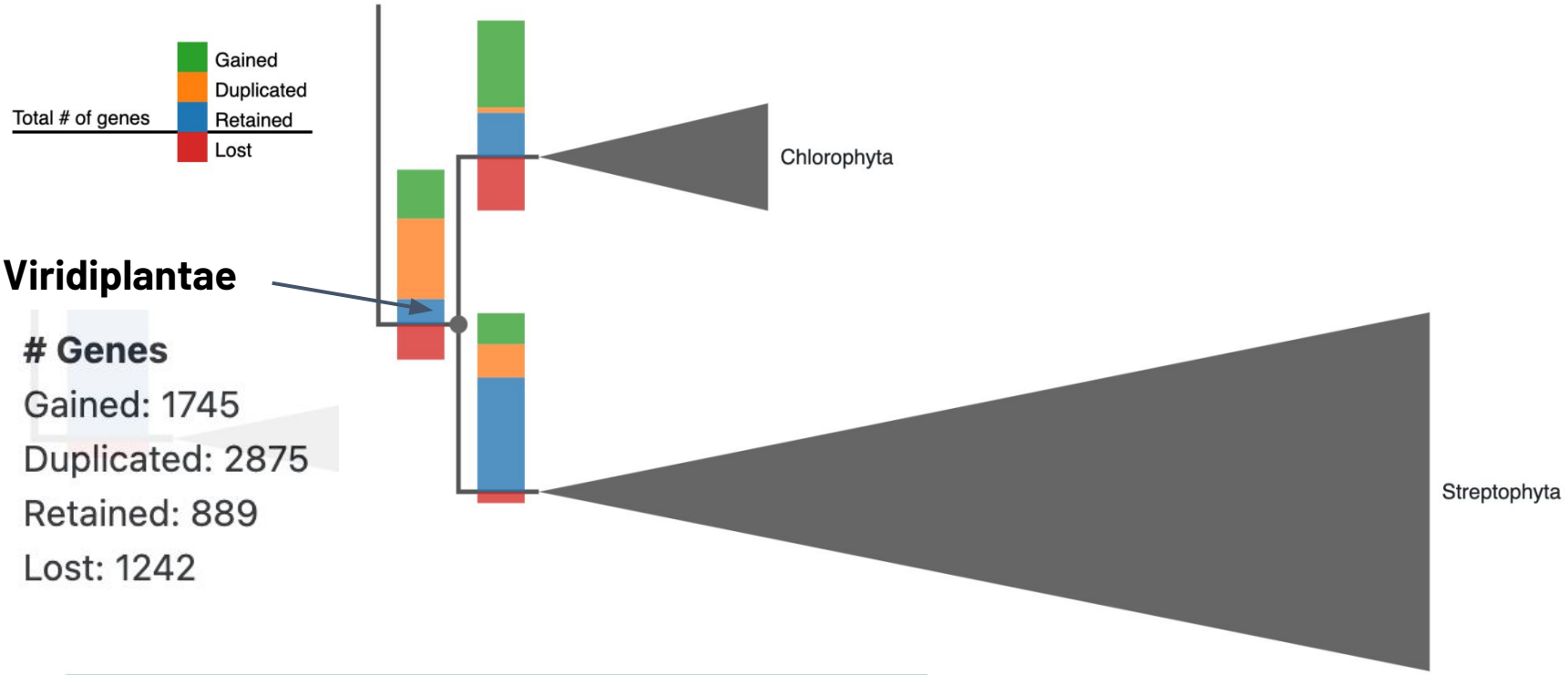
Phylogenetic profiling



Phylogenetic profiling

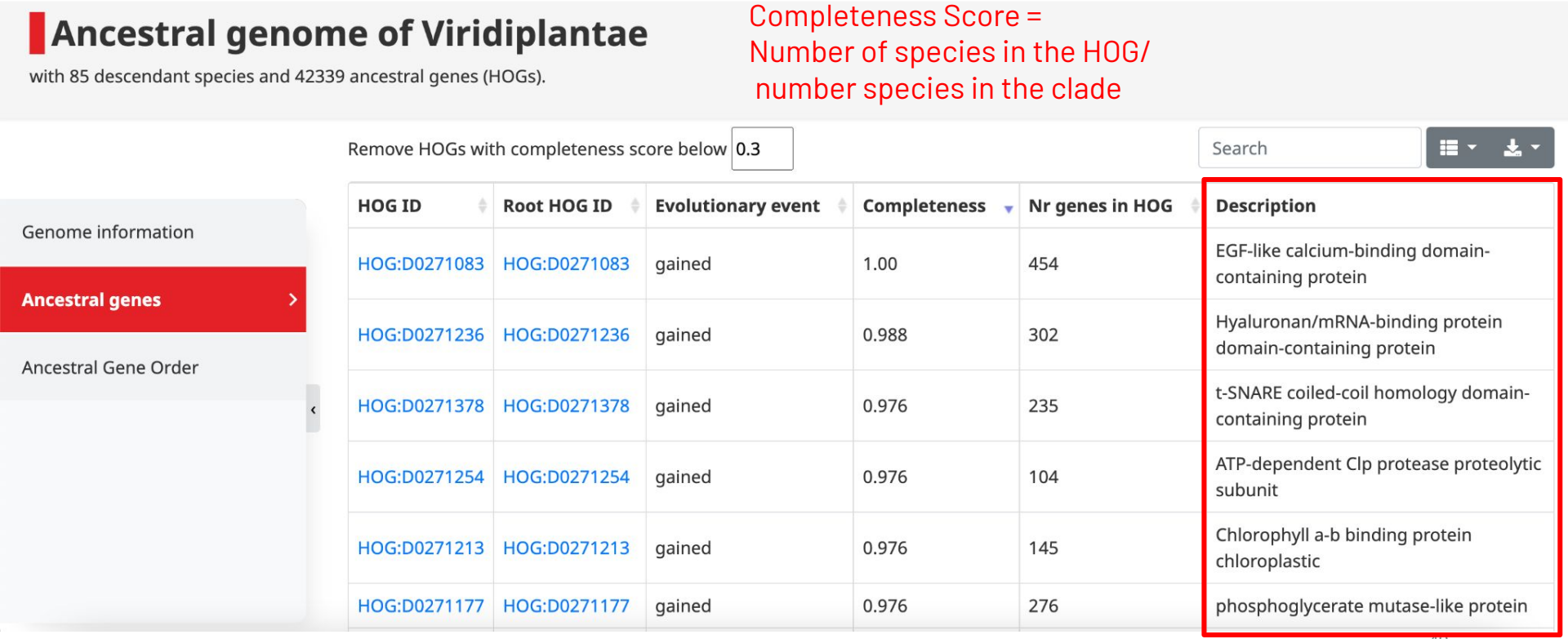


Tracking the evolutionary history of a gene family

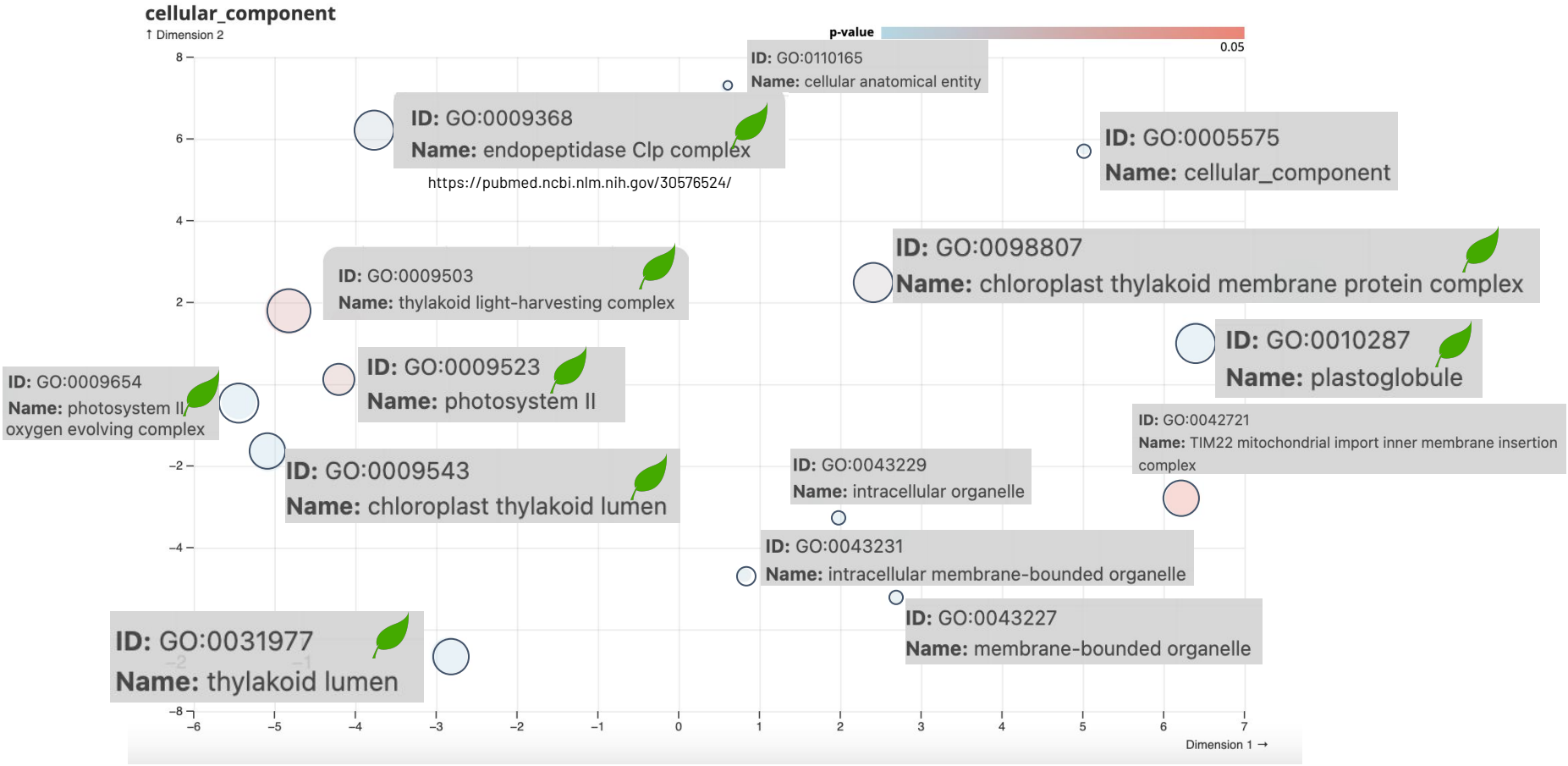


Filtered HOGs. We only show HOGs with a completeness score above 30%.

Gene content of ancestral genomes



Ancestral GO enrichment of genes gained at the **Viridiplantae** level

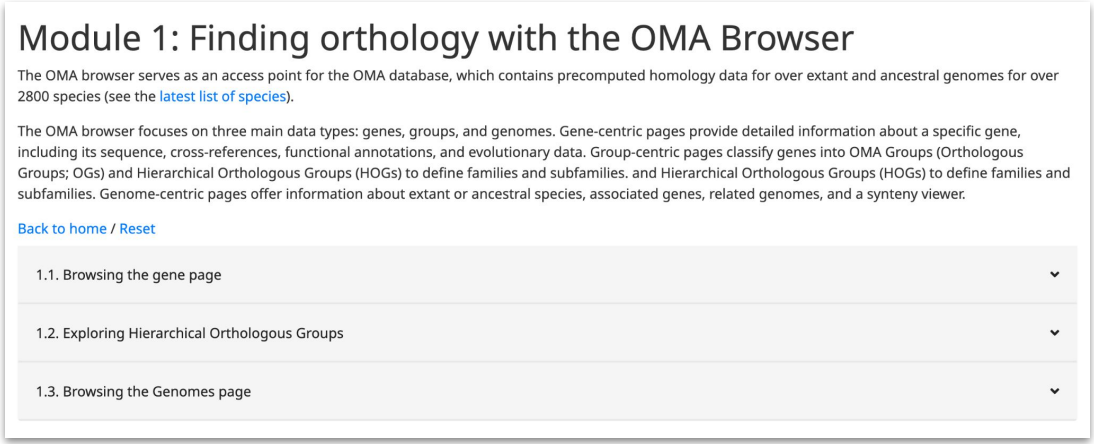


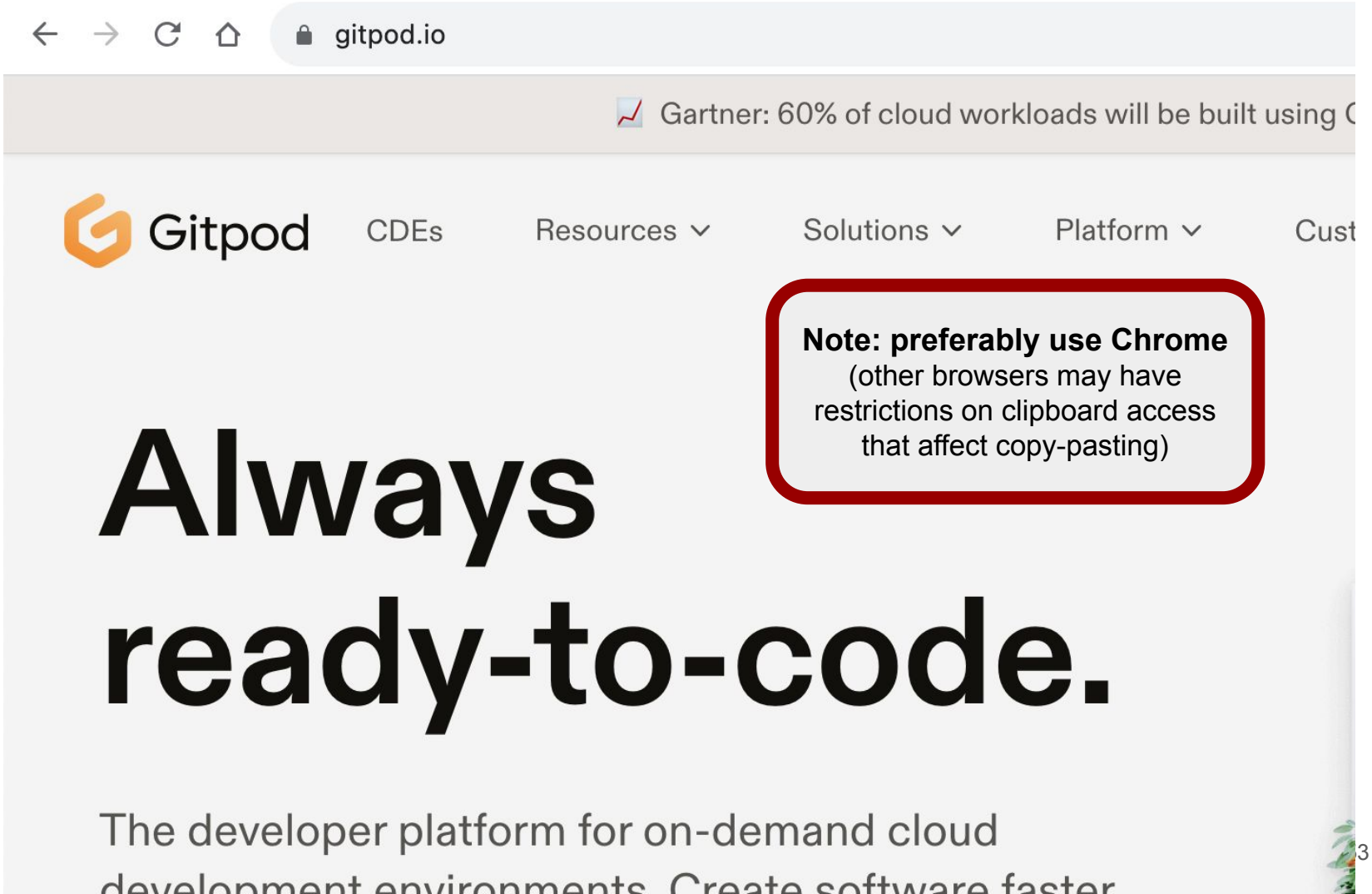
Resources and References

- Orthology: definitions, inferences and impact on species phylogeny inference (Fernández et al., 2019) https://arxiv.org/abs/1903.04530
- Inferring orthology and paralogy (Altenhoff and Dessimoz, 2012) https://people.inf.ethz.ch/adriaal/orthology-bookchapter.pdf
- Quest for Orthologs (consortium): https://questfororthologs.org/

Before the break

- Do Module 1 of the OMA Academy
  - https://omabrowser.org/oma/academy/
- Break at 10:45

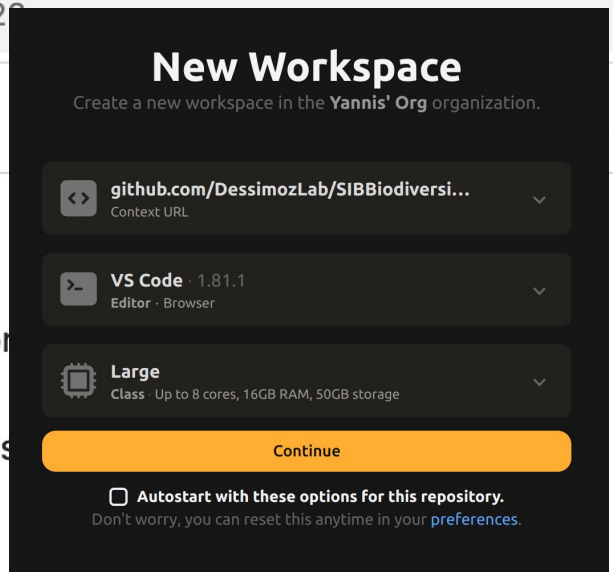



gitpod.io
Gartner: 60% of cloud workloads will be built using
Gitpod
CDEs
Resources
Solutions
Platform
Cust
Note: preferably use Chrome
(other browsers may have
restrictions on clipboard access
that affect copy-pasting)
Always
ready-to-code.
The developer platform for on-demand cloud

README.md

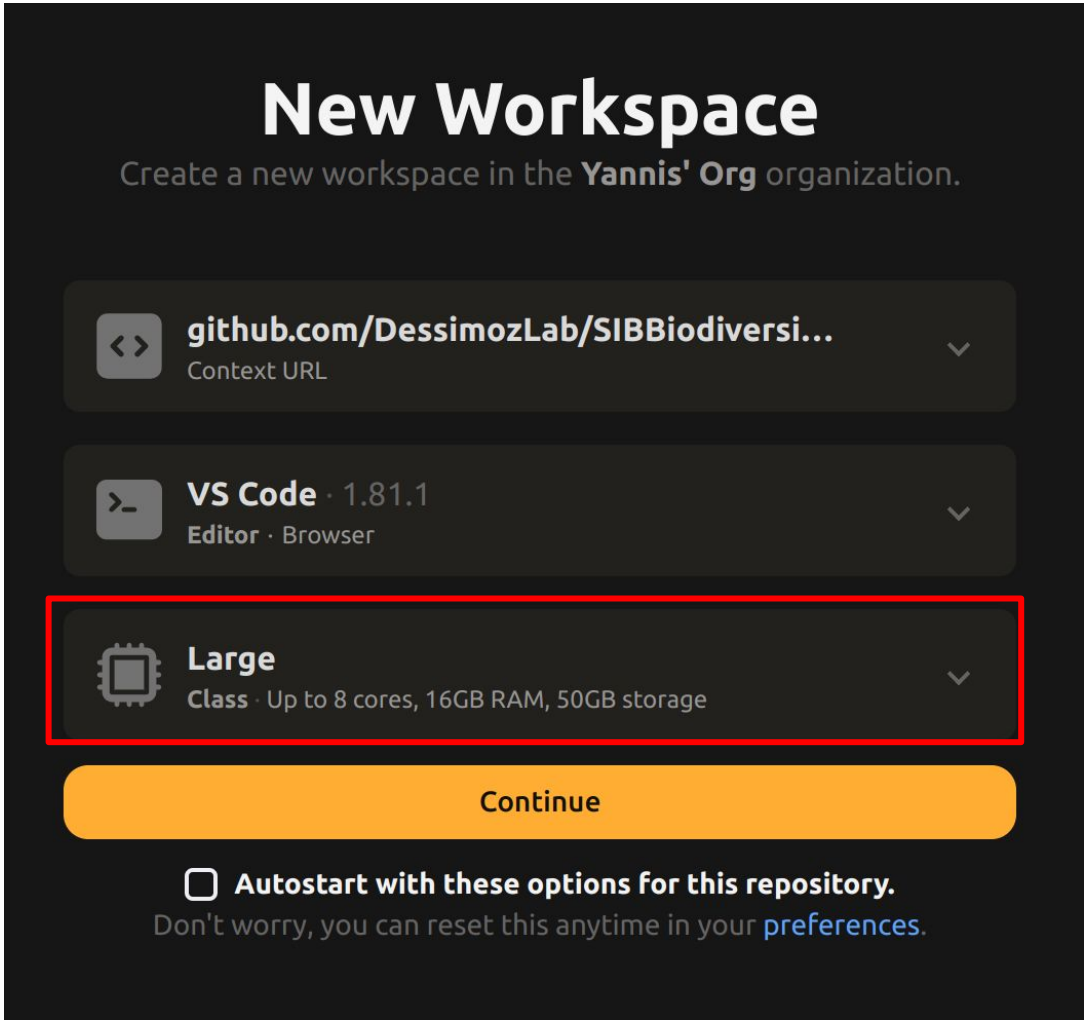
Software

We will be mainly working on an GitPod, an online integrated develop
users to write, edit, and run code directly in a web browser. GitPod is
software, code, and files needed for the course are stored and proces
need to install or configure anything locally.



You can access the GitPod here:
https://gitpod.io/#https://github.com/DessimozLab/SIBBiodiversityBioinformatics2023

Participants need to sign up for a GitPod account via Github and/or LinkedIn to access 50 hours per month for free, which is ample time to complete the exercises. After logging in, create a new workspace by choosing SIBBiodiverstyGenomics2023, Browser Editor, and Large configuration (8 cores, 16 GB RAM, 50 GB storage).


New Workspace
Create a new workspace in the Yannis' Org organization.
github.com/DessimozLab/SIBBiodiversi...
Context URL
VS Code · 1.81.1
Editor · Browser
Large
Class · Up to 8 cores, 16GB RAM, 50GB storage
Continue
Autostart with these options for this repository.
Don't worry, you can reset this anytime in your preferences.

After the installation, check if everything works

```
(omacademy) gitpod /workspace $ omamer -h
usage: omamer [-h] [--version] {mkdb,search,info} ...

OMAmer - tree-driven and alignment-free protein assignment to sub-families.

optional arguments:
  -h, --help     show this help message and exit
  --version, -v  Show version and exit.

Commands:
```

omamer -h

```
(omacademy) gitpod /workspace/SIBBiodiversityBioinformatics2024/Module3_FastOMA (main) $ nextflow -h
Usage: nextflow [options] COMMAND [arg...]

Options:
  -C
     Use the specified configuration file(s) overriding any defaults
  -D
     Set JVM properties
  -bg
     Execute nextflow in background
  -c, -config
     Add the specified file to configuration set
  -config-ignore-includes
     Disable the parsing of config includes
  -h
     Print this help
  -log
     Set nextflow log file path
  -q, -quiet
     Do not print information messages
```

nextflow -h

iqtree -h

Timeouts

Workspaces will stop after a period of inactivity without any user input.

Default Workspace Timeout

1h

Save

Use minutes or hours, like **30m** or **2h**

Module 2: OMAmer for sequence placement into HOGs

UNIL | Université de Lausanne

comparative genomics LAB

SIB

Why sequence placement ?

	Speed	Input	Proteomes
Orthology database	Instantaneous	Known identifiers	Only in the database
Placement into HOGs	Few minutes	Few sequences or whole proteome	Any proteome
Orthology inference	Hours	Whole proteomes	Any proteomes (one or several)

What is OMAmer?

- Fast sequence placement into existing HOGs from the OMA Browsers

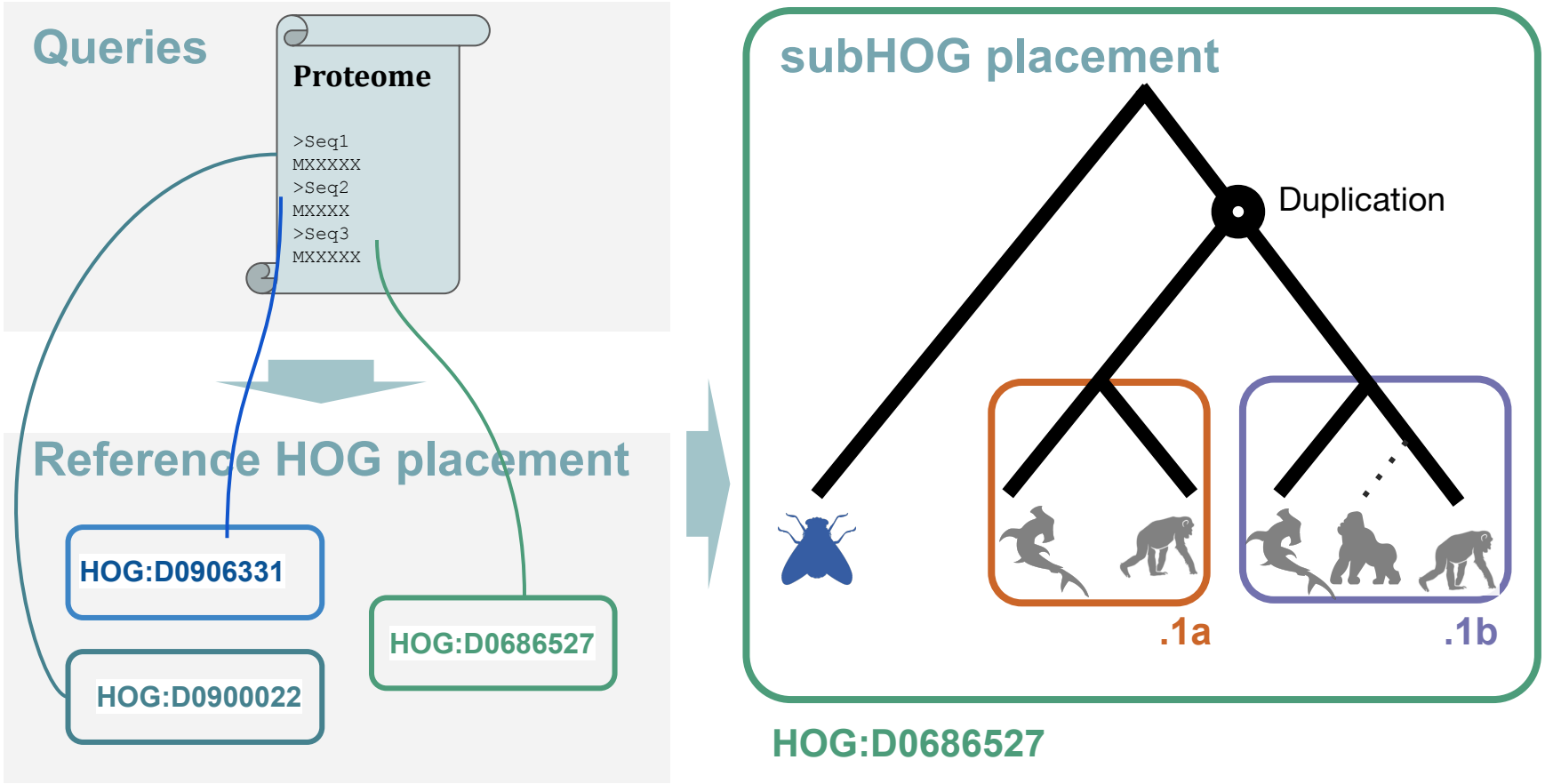
- More accurate than closest sequence matching for subfamily placement!

OMAmer: tree-driven and alignment-free protein assignment to subfamilies outperforms closest sequence approaches

Victor Rossier [1,2,3], Alex Warwick Vesztrocy [1,2,3], Marc Robinson-Rechavi [3,4,*] and Christophe Dessimoz [1,2,3,5,6,*]

https://github.com/DessimozLab/omamer

OMAmer placement - principle



k-mer based placement

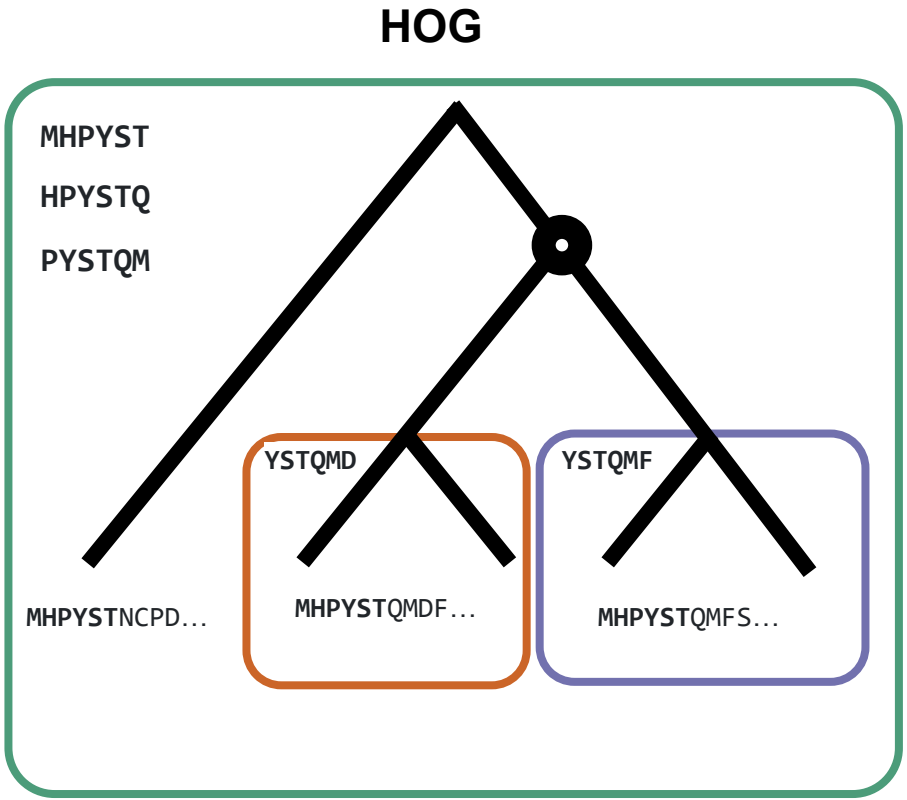
- **k-mers** : words of k characters in a sequence

Query sequence

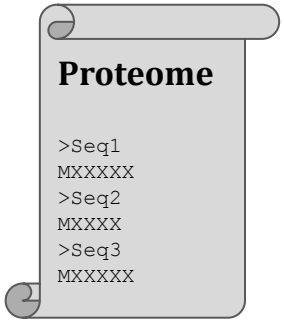
```
MHPYSTQMFS LQITVMEDSQ SDMSIELPLS

MHPYST
 HPYSTQ
  PYSTQM

          ...
            ...
              ...          MSIELP
                             SIELPL
                              IELPLS
```



How to use OMAmer




```
omamer search --query query.fa --db db.h5 --output results.txt
```

Query sequences
FASTA format

From any species

OMAmer database
HDF5 format

Built with HOGs from the OMA Browser

Seq1	HOG:D0578800.1c.1d
Seq2	HOG:D0571029
Seq3	HOG:D0606120.3n

OMAmer output
Tab separated format

All HOG placements

Interpreting the output

qseqid	hogid	hoglevel	family_p	family_ count	family_ normcount
Seq1	HOG:D0630083.1g.9c.31d	Theria	858.4562422	133	0.970765763
Seq2	HOG:D0630583	Gnathostomata	1002.895597	122	1

subfamily_ score	subfamily_count	qseqlen	subfamily_medianseqlen	qseq_overlap
0.9298	53	143	143	1
1	122	128	155	1

https://github.com/DessimozLab/omamer

Interpreting the output

qseqid	hogid	family_p	family_normcount	subfamily_score	qseq_overlap
Seq1	HOG:D0630083.1g.9c.31d	858.4562422	0.970765763	0.929822828	1
Seq2	HOG:D0630583	1002.895597	1	0.99999372	1

- **qseqid: Query identifier**

Interpreting the output

qseqid	**hogid**	family_p	family_normcount	subfamily_score	qseq_overlap
Seq1	HOG:D0630083.1g.9c.31d	858.4562422	0.970765763	0.929822828	1
Seq2	HOG:D0630583	1002.895597	1	0.99999372	1

- qseqid: Query identifier
- **hogid : HOG ID where the query is placed**

Interpreting the output

qseqid	hogid	**family_p**	family_normcount	subfamily_score	qseq_overlap
Seq1	HOG:D0630083.1g.9c.31d	858.4562422	0.970765763	0.929822828	1
Seq2	HOG:D0630583	1002.895597	1	0.99999372	1

- qseqid: Query identifier
- hogid : HOG ID where the query is placed
- **family_p : p-value for query matching this family/rootHOG - expressed in neg log units (*higher is better*)**

Significance threshold:
-log(family_p) <= α
α = -log(10^-6)

Interpreting the output

qseqid	hogid	family_p	**family_normcount**	subfamily_score	qseq_overlap
Seq1	HOG:D0630083.1g.9c.31d	858.4562422	0.970765763	0.929822828	1
Seq2	HOG:D0630583	1002.895597	1	0.99999372	1

- qseqid: Query identifier
- hogid : HOG ID where the query is placed
- family_p : p-value for query matching this family/rootHOG - expressed in neg log units (*higher is better*)
- **family_normcount : count of query k-mers in common with family/rootHOG (normalized)**

Significance threshold:
-log(family_p) <= α
α = -log(10^-6)

Interpreting the output

qseqid	hogid	family_p	family_normcount	**subfamily_score**	qseq_overlap
Seq1	HOG:D0630083.1g.9c.31d	858.4562422	0.970765763	0.929822828	1
Seq2	HOG:D0630583	1002.895597	1	0.99999372	1

- qseqid: Query identifier
- hogid : HOG ID where the query is placed
- family_p : p-value for query matching this family/rootHOG - expressed in neg log units (*higher is better*)
- family_normcount : count of query k-mers in common with family/rootHOG (normalized)
- **subfamily_score: subfamily OMAmer score (similarity between query and subHOG)**

Significance threshold:
-log(family_p) <= α
α = -log(10^-6)

Interpreting the output

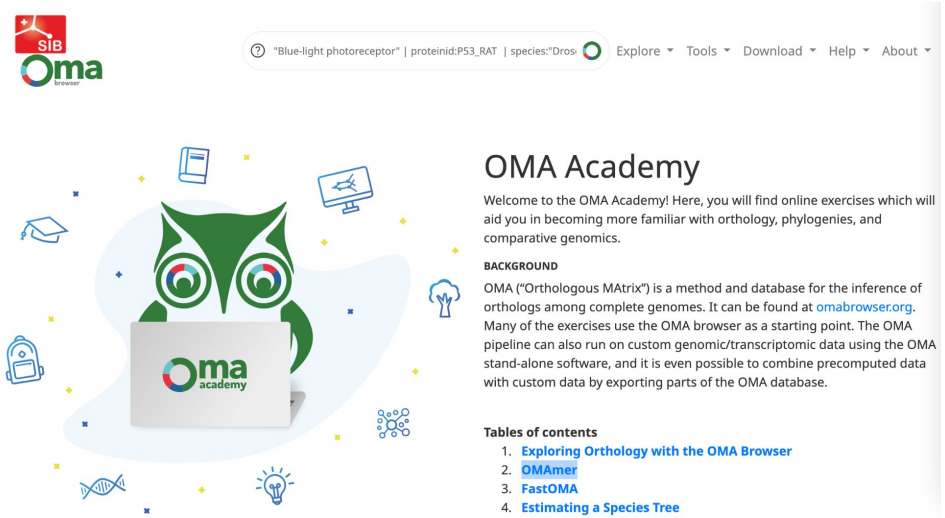
qseqid	hogid	family_p	family_normcount	subfamily_score	**qseq_overlap**
Seq1	HOG:D0630083.1g.9c.31d	858.4562422	0.970765763	0.929822828	1
Seq2	HOG:D0630583	1002.895597	1	0.99999372	1

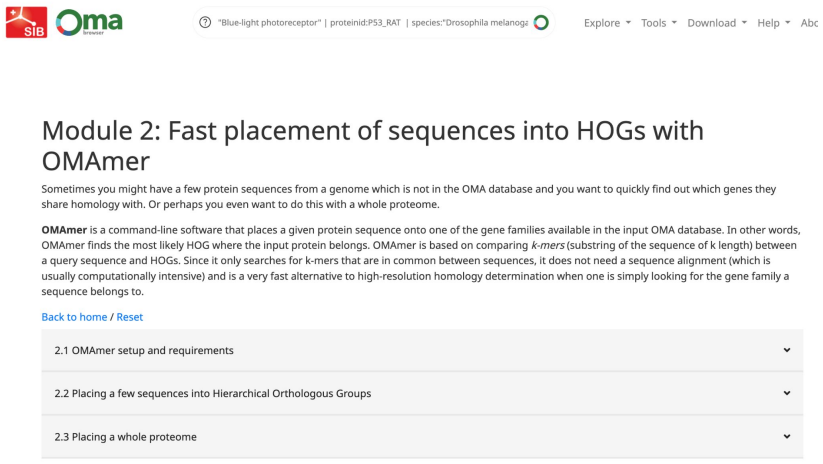
- qseqid: Query identifier
- hogid : HOG ID where the query is placed
- family_p : p-value for query matching this family/rootHOG - expressed in neg log units (*higher is better*)
- family_normcount : count of query k-mers in common with family/rootHOG (normalized)
- subfamily_score: subfamily OMAmer score (similarity between query and subHOG)
- **qseq_overlap : query k-mers (%) overlapping with k-mers of reference root-HOGs**

Significance threshold:
-log(family_p) <= α
α = -log(10^-6)

To remember

- Placement into HOGs allows us to find gene families for **species not in the database**
- Can be used on **any number of sequences** - from one to whole proteomes
- Precise to the subfamily level, but not a definitive proof of orthology
- Allows us to still **take advantage of OMA Browser** wealth of data and features





Module 2: work until 12:10

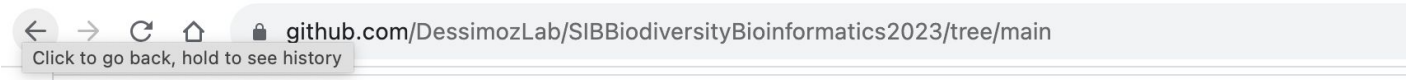
```
source /workspace/conda/bin/activate
conda activate omacademy
```

```
nextflow FastOMA/FastOMA_light.nf --input_folder in_folder --output_folder out_folder
-resume
```

```
cd /workspace/SIBBiodiversityBioinformatics2023/Module3_FastOMA/expected_output/
```

Typo alert! (Module 3.1):

In this exercise, we will run FastOMA standalone to infer the orthology information for five yeast species. We already provided the proteomes of five species in the GitPod environment, located at `/workspace/SIBBiodiversityBionformatics2023/Module3_FastOMA/working_dir/in_folder/proteome`.



.gitpod.yml	Changed path in the github file to reflect changing name of repo	yesterday
LICENSE	Initial commit	5 days ago
README.md	Update README.md	now

README.md

SIB Biodiversity Bioinformatics 2023

Teachers

- Natasha Glover
- Yannis Nevers
- Sina Majidian
- Christophe Dessimoz

FastOMA (temp)

FastOMA command line

```
cd /workspace/SIBBiodiversityBioinformatics2023/Module3_FastOMA/working_dir/
nextflow FastOMA_light.nf --input_folder in_folder --output_folder out_folder
```

expected output structure for test data

Then, following files and folders should appear in the folder `out_folder` which was the argument.

```
$ls out_folder
hogmap  OrthologousGroupsFasta  OrthologousGroups.tsv  output_hog.orthoxml
```

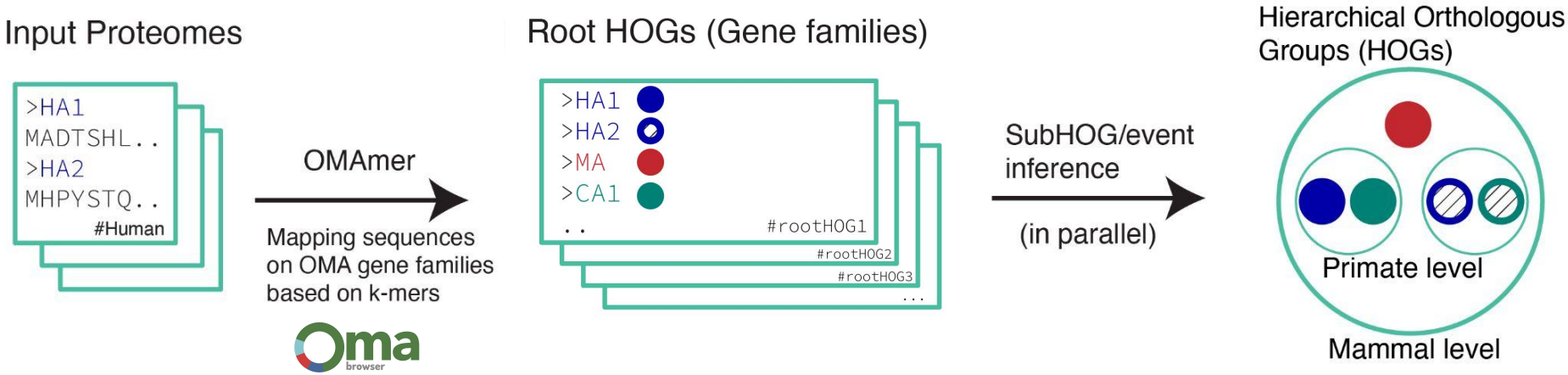
Module 3: FastOMA

UNIL | Université de Lausanne

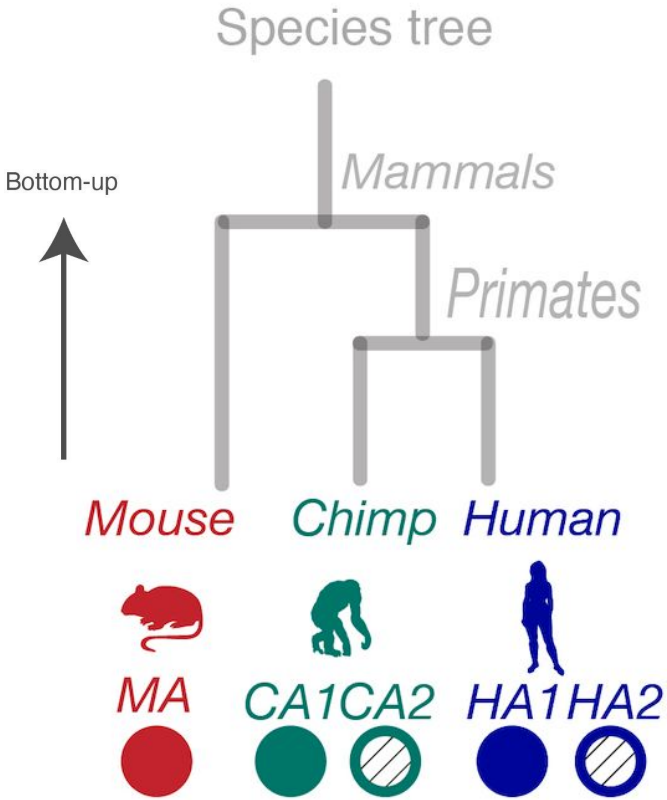
comparative genomics LAB

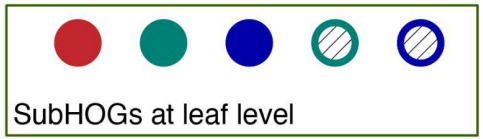
SIB

FastOMA, our new tool

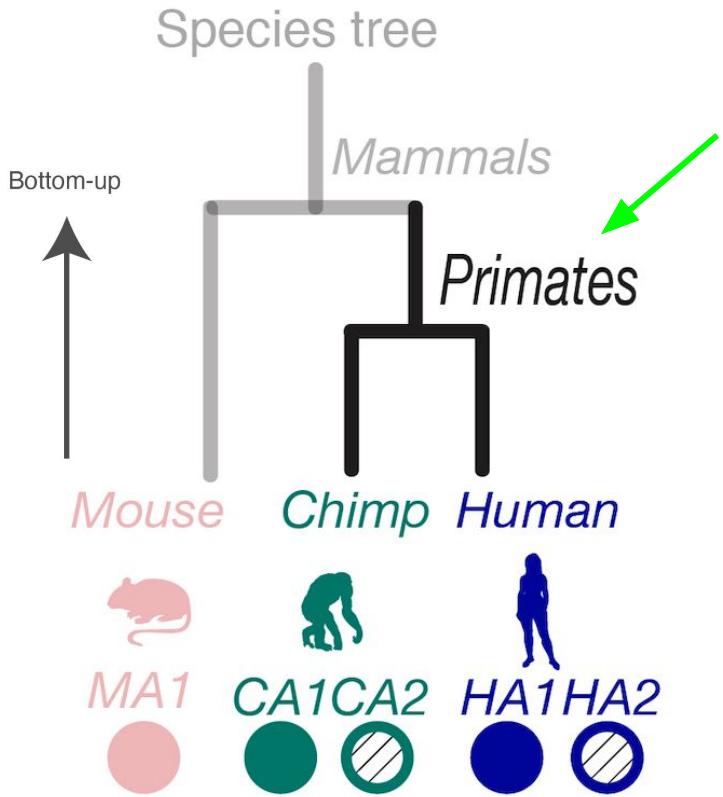


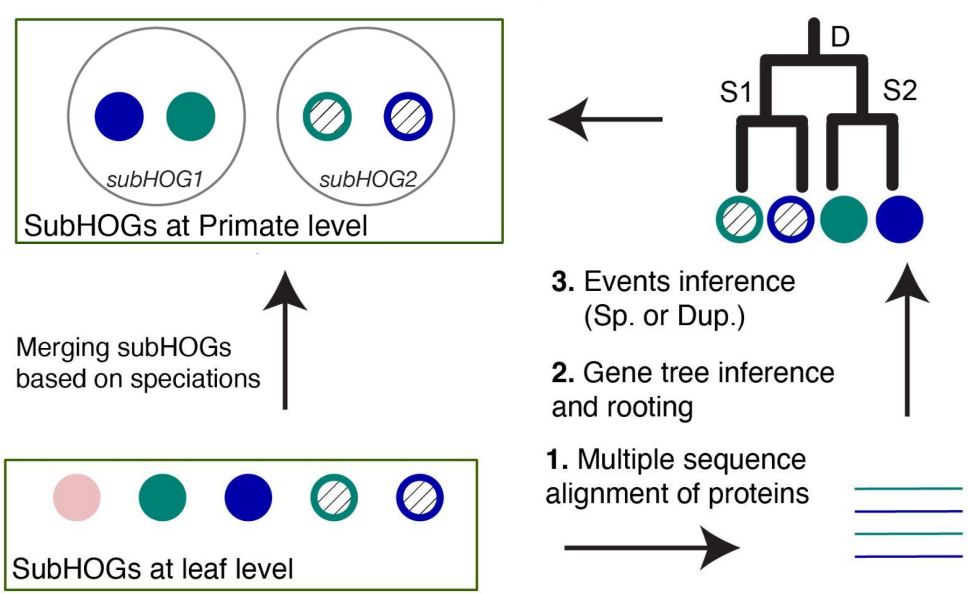
HOG inference



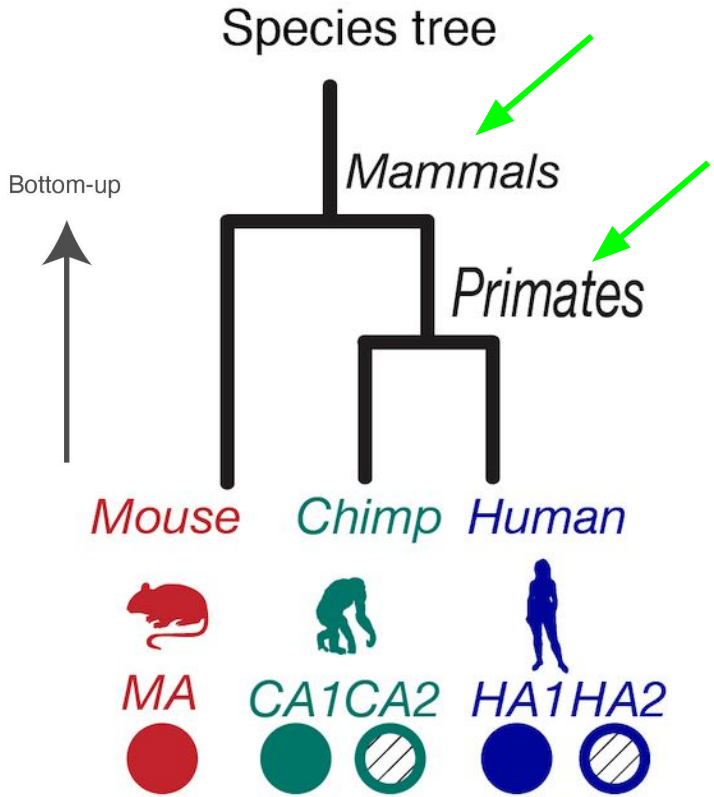


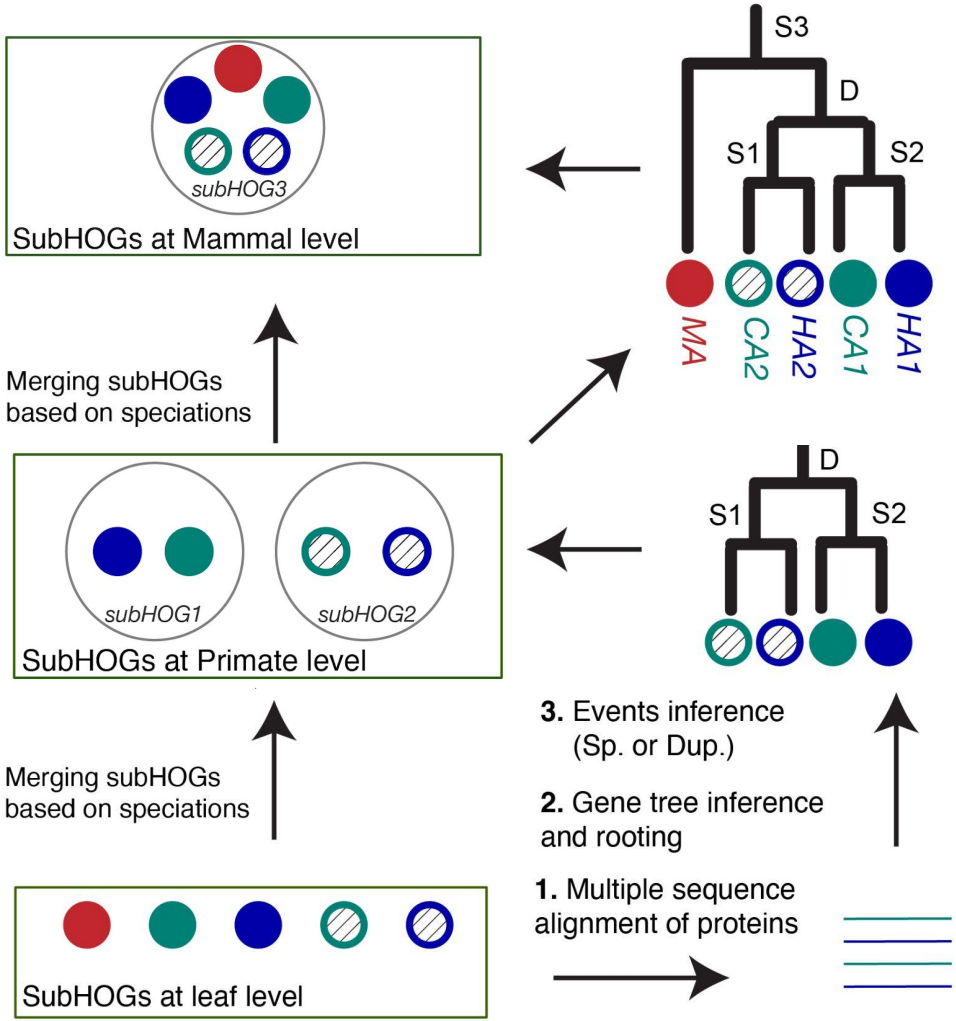
HOG inference



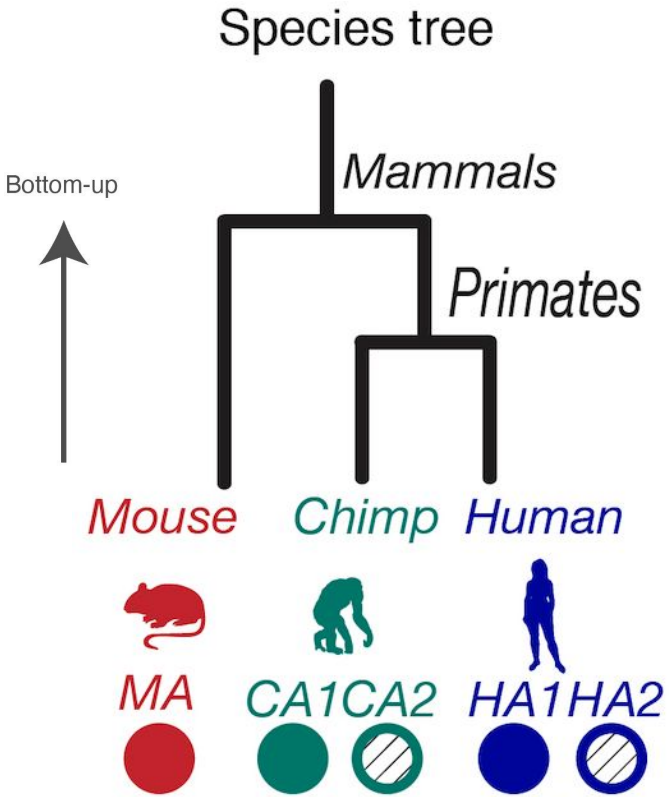


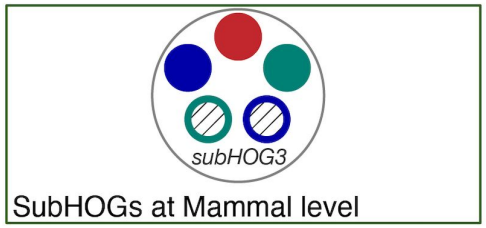
HOG inference





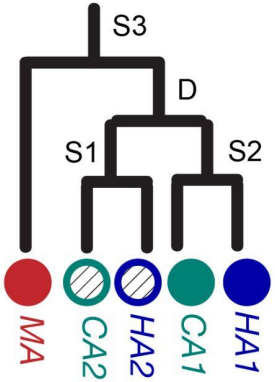
HOG inference



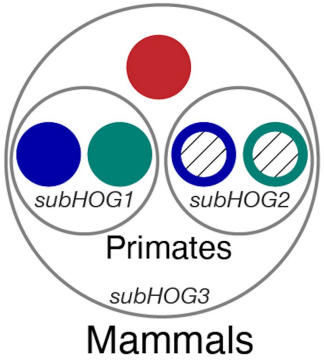


HOG in orthoXML

```
<orthoGroup3        Mammals>
  MA
  <paraGroup>
    <orthoGroup1    Primates>
      HA1
      CA1
    </orthoGroup1>
    <orthoGroup2    Primates>
      HA2
      CA2
    </orthoGroup2>
  </paraGroup>
</orthoGroup3>
```

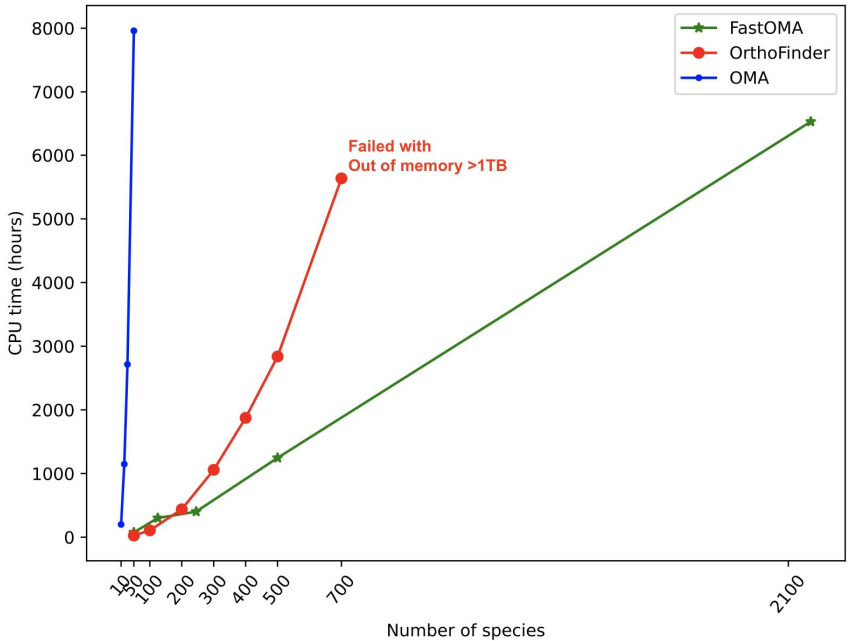


Nested structure of HOG



Orthology inference for Eukaryote dataset

- 2180 eukaryotic species
- Uniprot reference proteomes
- in a single day using 300 CPUs



github.com/DessimozLab/FastOMA

Module 3.3

- **4. How many Root HOGs are in the HOG file?**

✖ Hint ✖ Answer

Each line in the output file denotes a gene family. After running, check the end of the file rootHOGs.tsv. Note that the indexing starts from 0.

There are 6793 rootHOG (gene families) in this file.

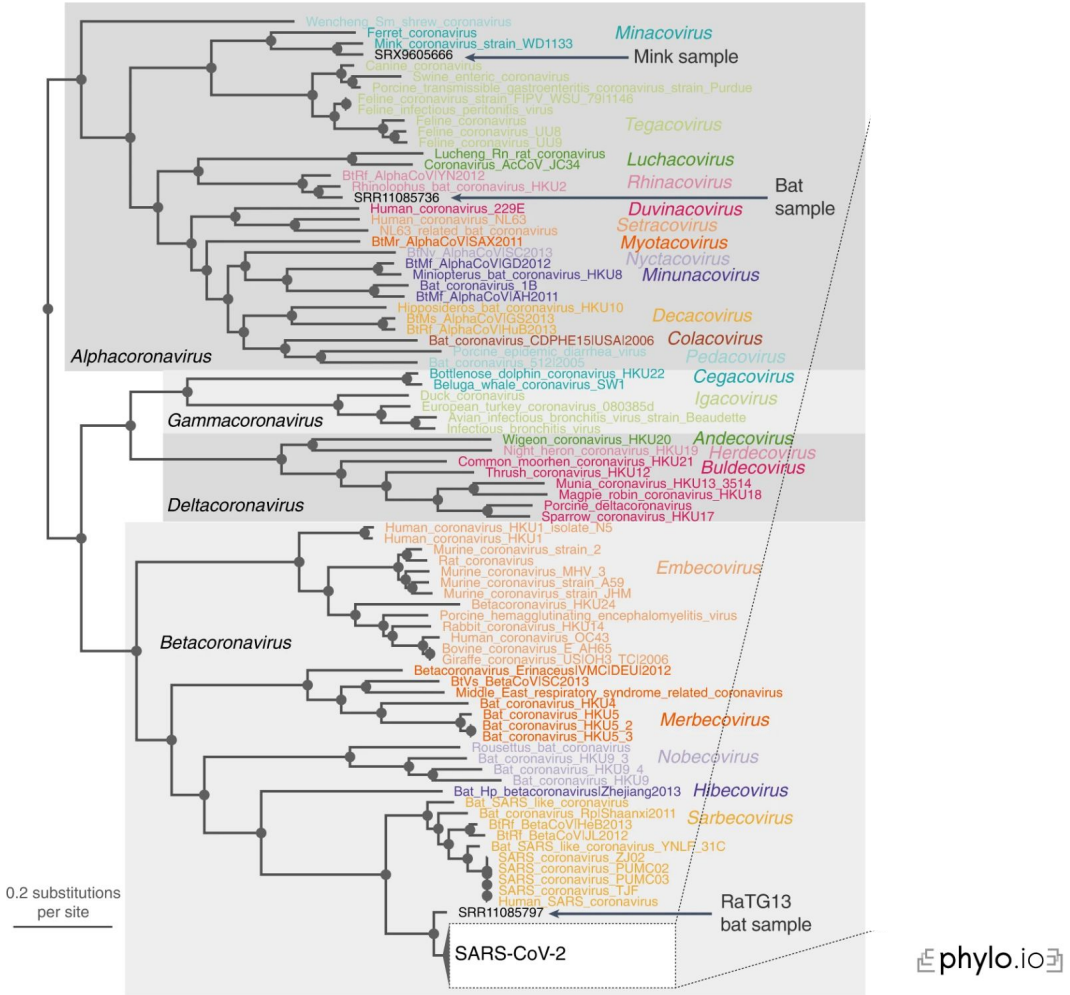
- **5. Consider the gene "60S ribosomal protein L15-A" in *Schizosaccharomyces pombe* with protein ID: RL15A_SCHPO. How many proteins are in the gene family (for these 5 species of interest)?**

Module 4 : Building a species tree

A tree of *Coronaviridae* viruses

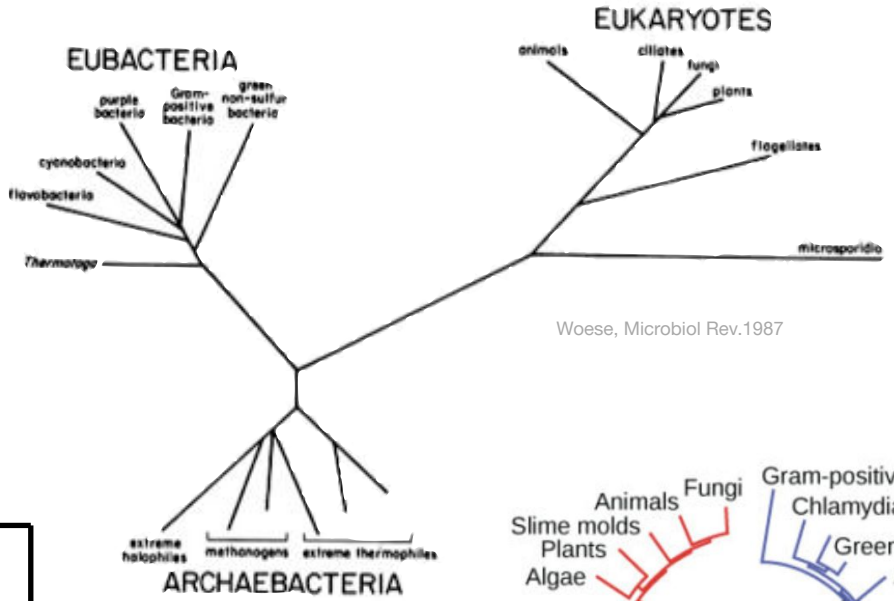
Concepts:

- Topology
- Branches lengths
- Internal nodes
- Leaves
- Root
- Clades / lineages



Dylus et al. Nat Biotech 2023

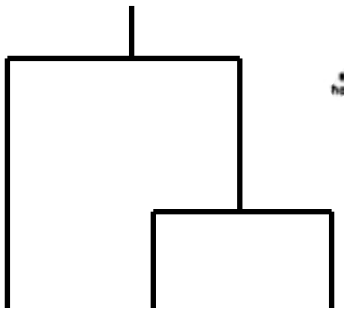
Other tree representations



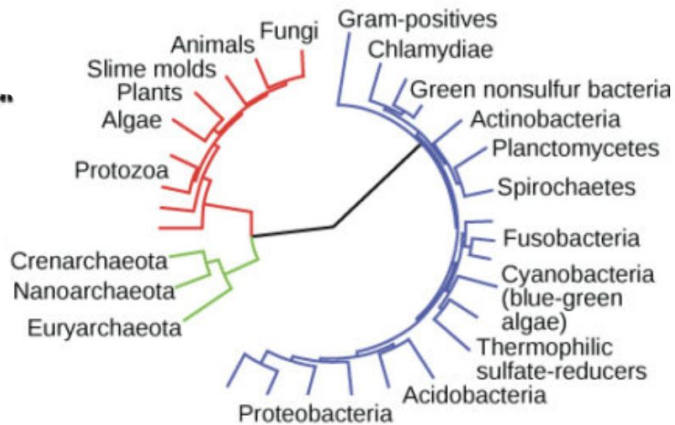
Woese, Microbiol Rev.1987

Concepts:

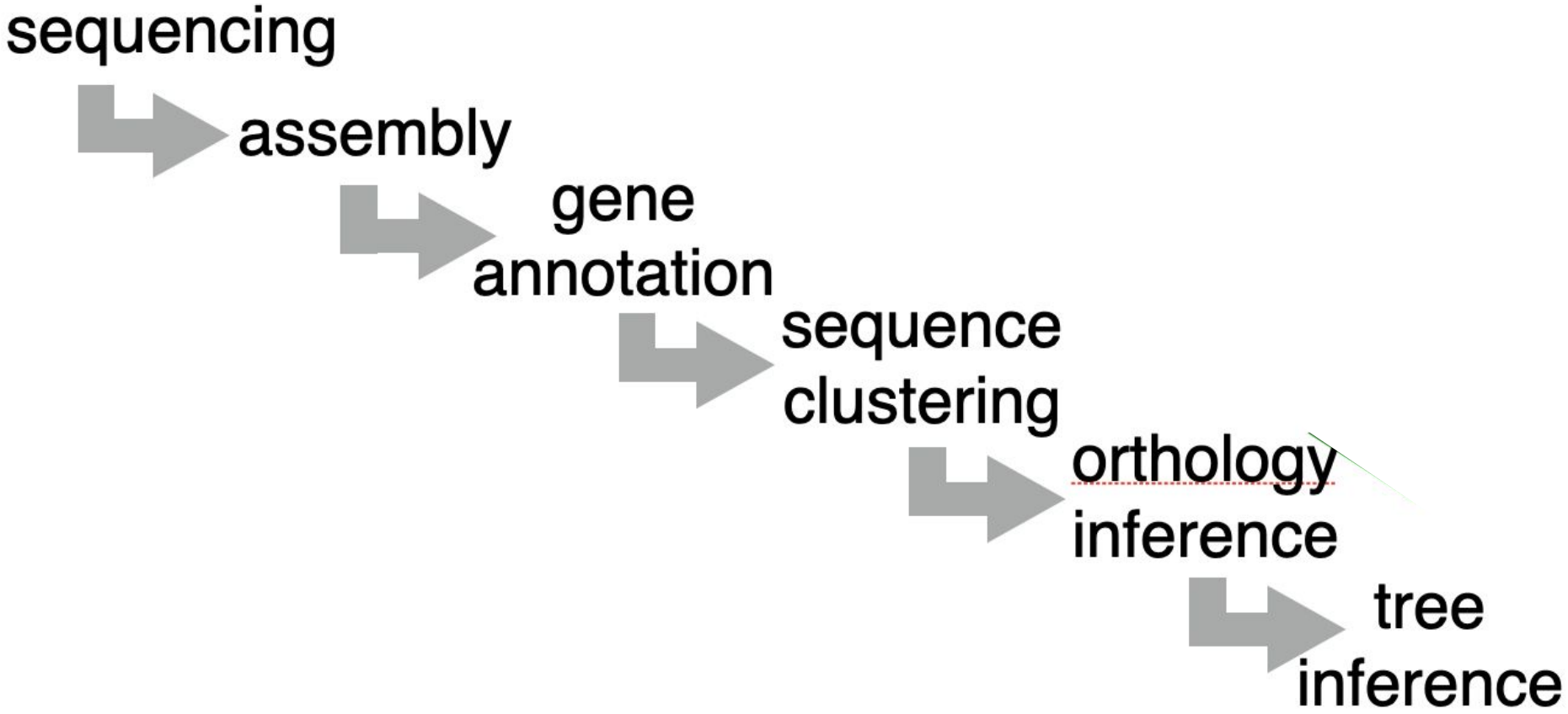
- Rooted vs. unrooted
- Scaled vs. unscaled



Gorilla Chimp Human



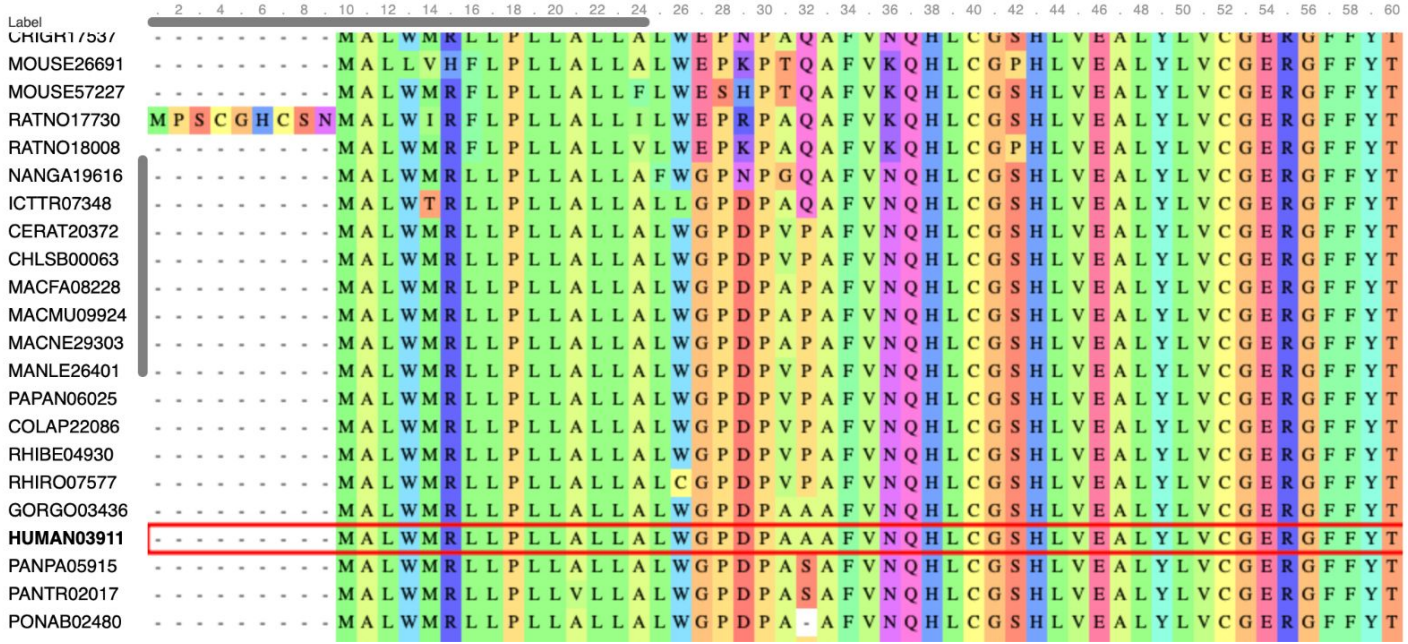
https://courses.lumenlearning.com/wm-biology2/chapter/phylogenetic-trees/


sequencing
assembly
gene annotation
sequence clustering
orthology inference
tree inference

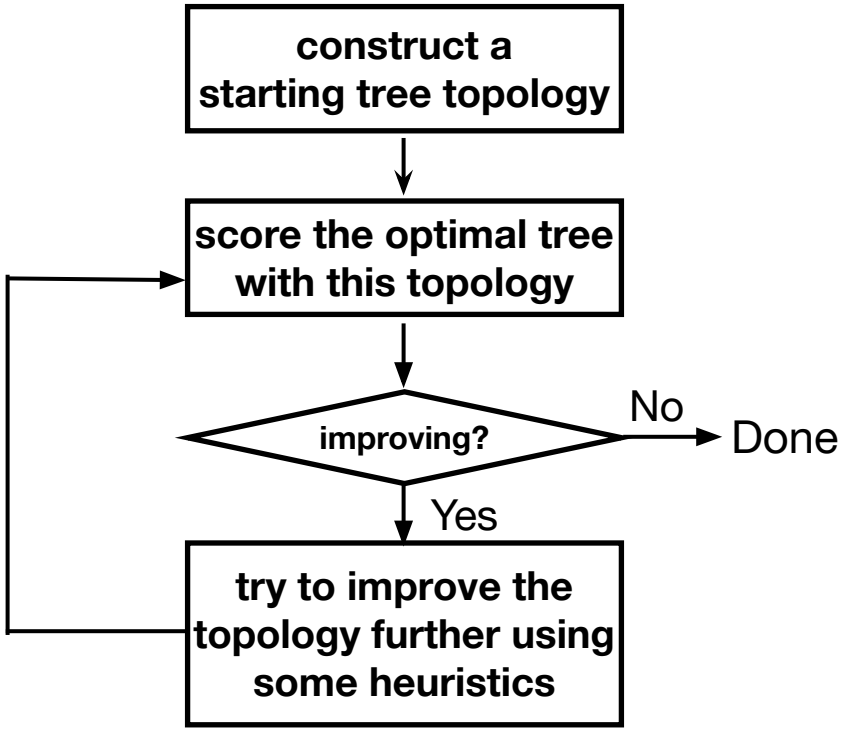
HOG:C0620879.1a with 61 members (insulin)

Vertebrata / Gnathostomata / Euteleostomi / Sarcopterygii / Tetrapoda / Amniota / **Mammalia** / Lower Level ▸

- Graphical viewer
- Members
- **Alignment**
- Ancestral synteny
- Similar HOGs



General Scheme to Build Phylogenetic Trees



- UPGMA
‘ Neighbour joining

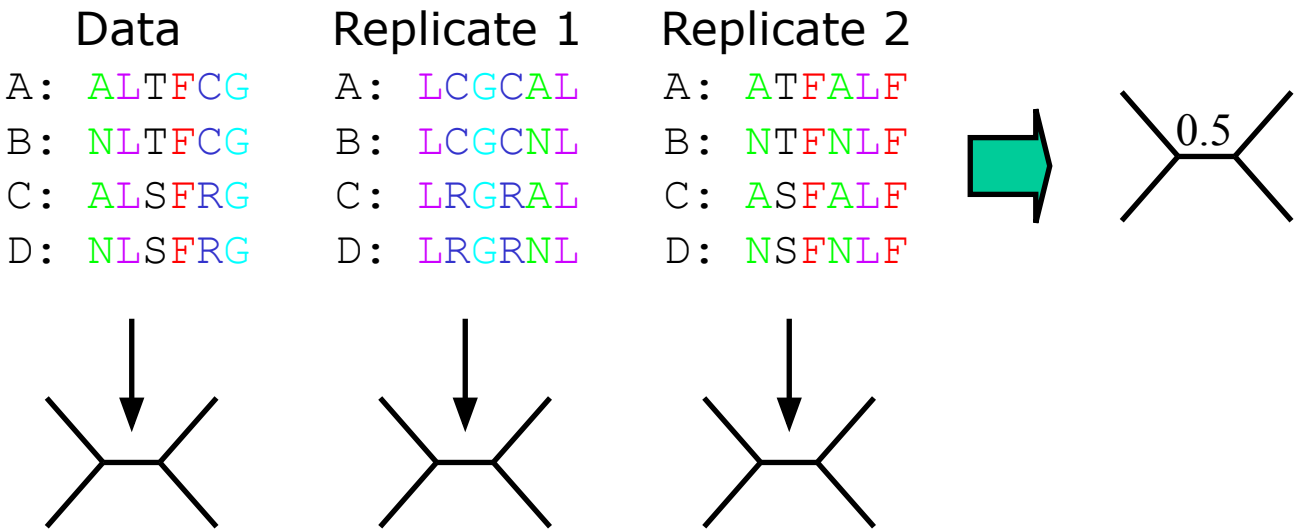
- Maximum parsimony
- Least squares fit of distances
- Maximum likelihood

- Swap leaves
- Swap subtrees

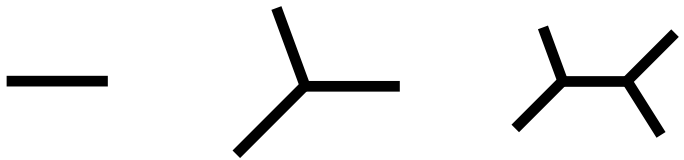
Measuring confidence with the Bootstrap

How would our tree inference change if we repeated the experiment with other data that came about through the same process?

Use “pseudo-replicates”: resample alignment columns (with replacement), same length.



How many branches are there in an unrooted bifurcating tree of n taxa?



2 3 4

1 3 5

How many topologies?

Number of "taxa"	Number of binary trees	
	Unrooted	Rooted
3	1	