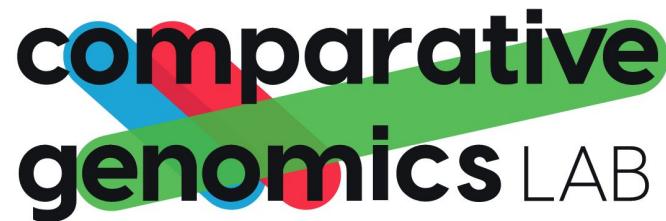


Biodiversity bioinformatics

From large-scale phylogenomics to gene families and functions

30 Aug 2023: Orthology with OMA



Learning objectives

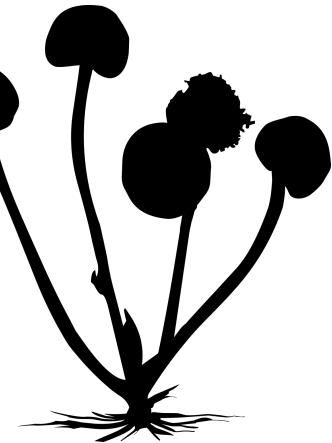
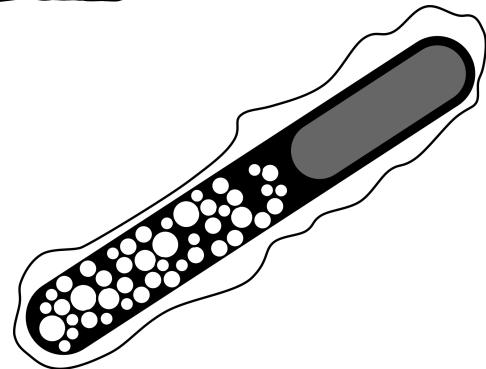
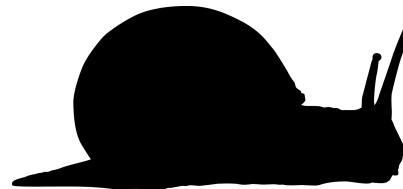
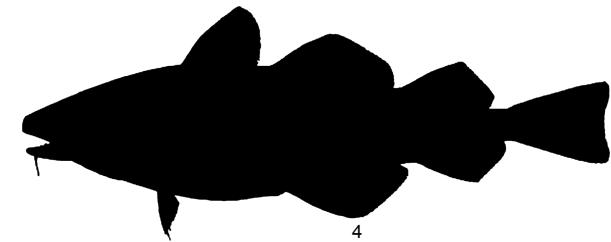
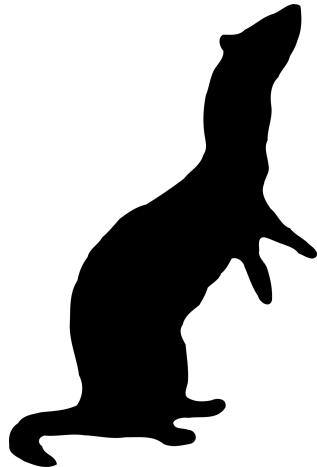
This course is centered around comparative genomics. After the course, you will be able to:

- 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

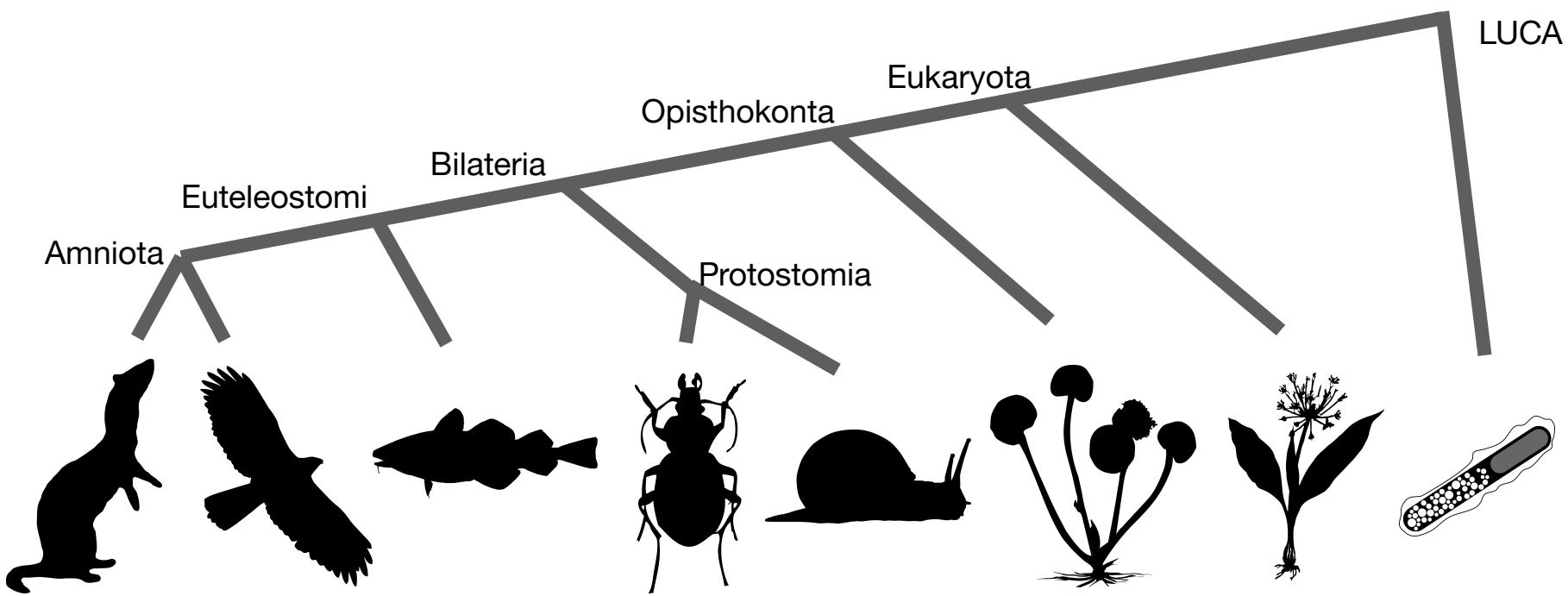
Schedule

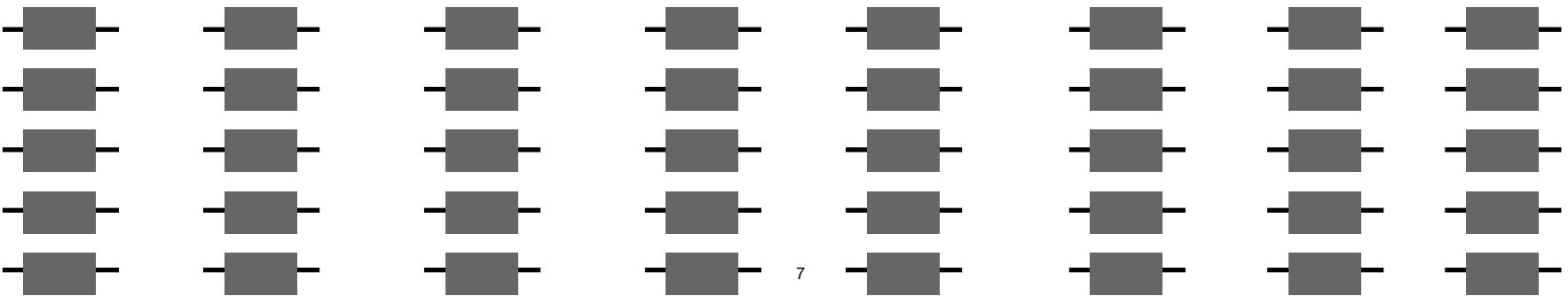
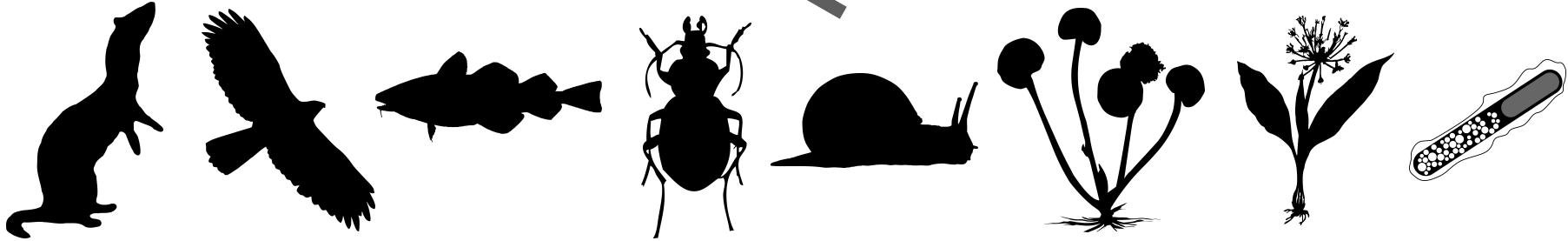
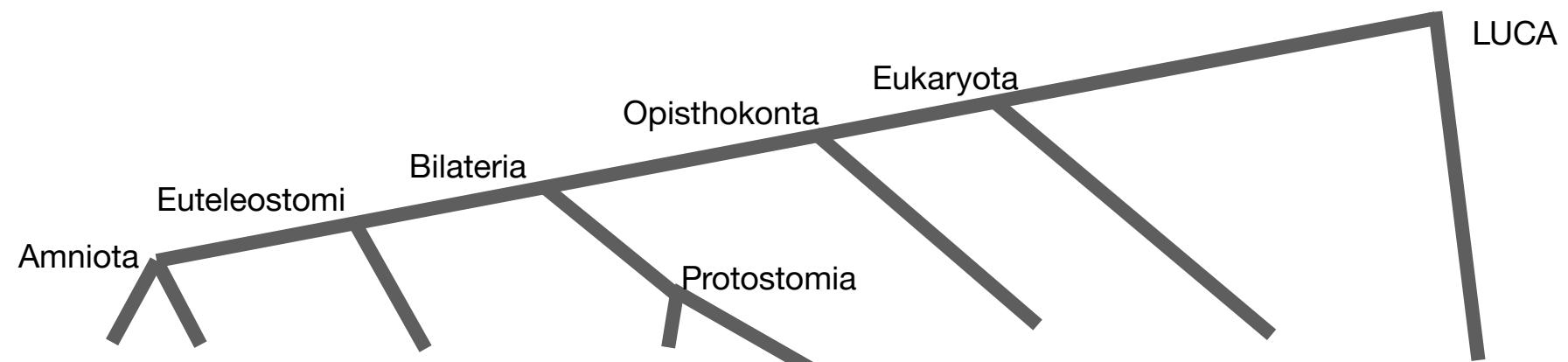
Schedule Aug 30 2023

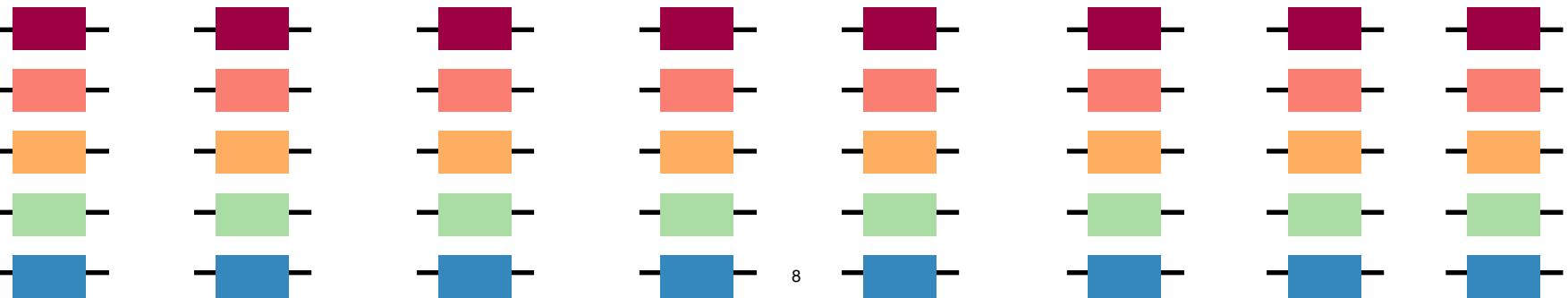
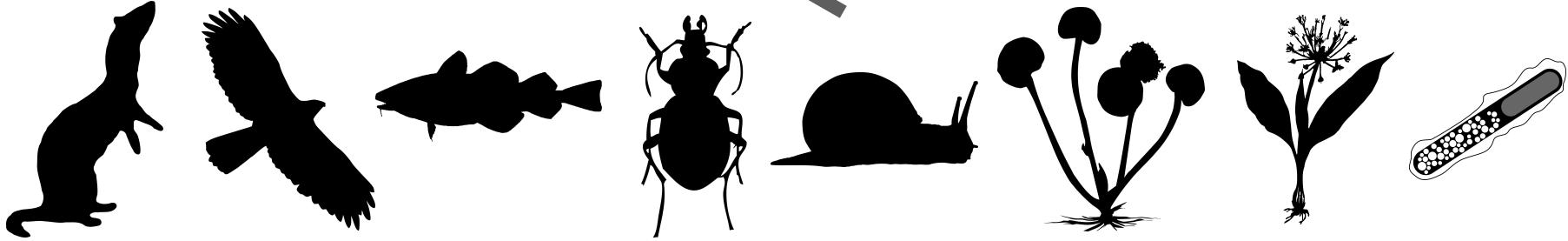
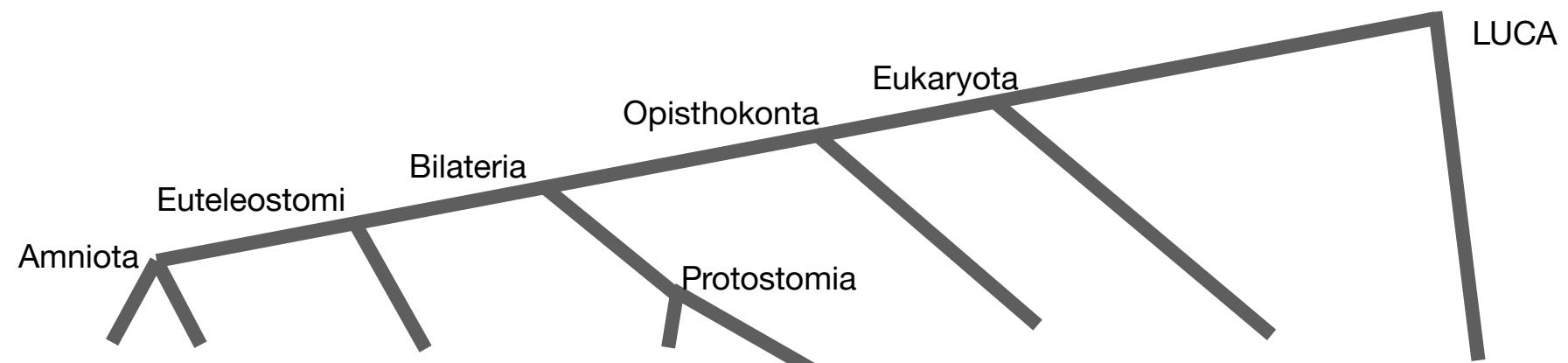
Time	Activity	In charge
9:00-9:30	Welcome, Introductions	Natasha Glover
9:30-10:00	Lecture: Overview, objectives, motivation, concept of orthologs, HOGs	Natasha Glover
10:00-10:45	Module 1: genes, groups, and genomes in the OMA Browser	Natasha Glover
10:45 - 11:15	Coffee break	
11:15-11:25	Go over results	Natasha Glover
11:25 - 11:30	Lecture: introduce OMAMer	Yannis Nevers
11:30 - 11:55	Module 2: Fast placement with OMAMer	Yannis Nevers
11:55-12:00	Go over results	Yannis Nevers
12:00-12:10	Lecture: FastOMA	Sina Majidian
12:10-12:30	Module 3 part 1: FastOMA (launch it before lunch)	Sina Majidian
12:30-13:30	Lunch	
13:30-14:00	Module 3 part 2: FastOMA	Sina Majidian
14:00-14:15	Go over results	Sina Majidian
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, Yannis, Christophe
16:45-17:00	Wrap up	Natasha, Yannis, Christophe

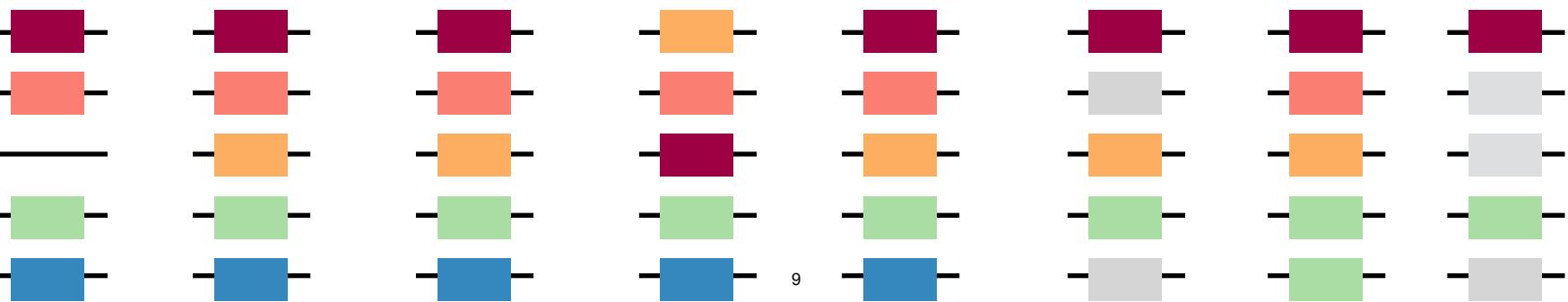
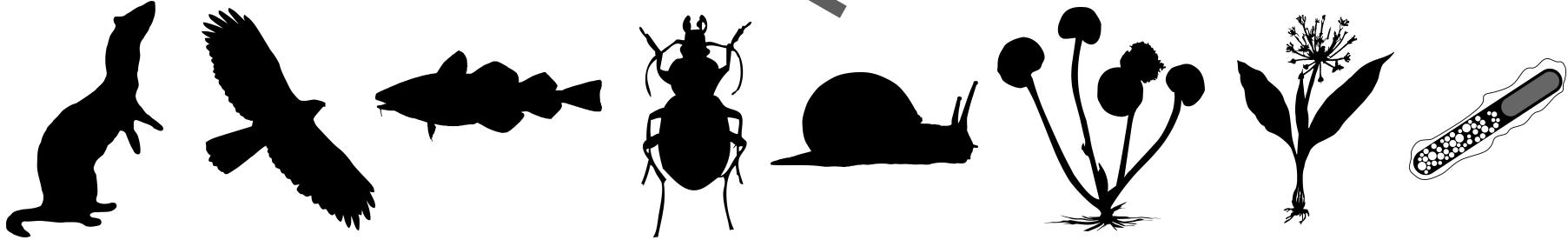
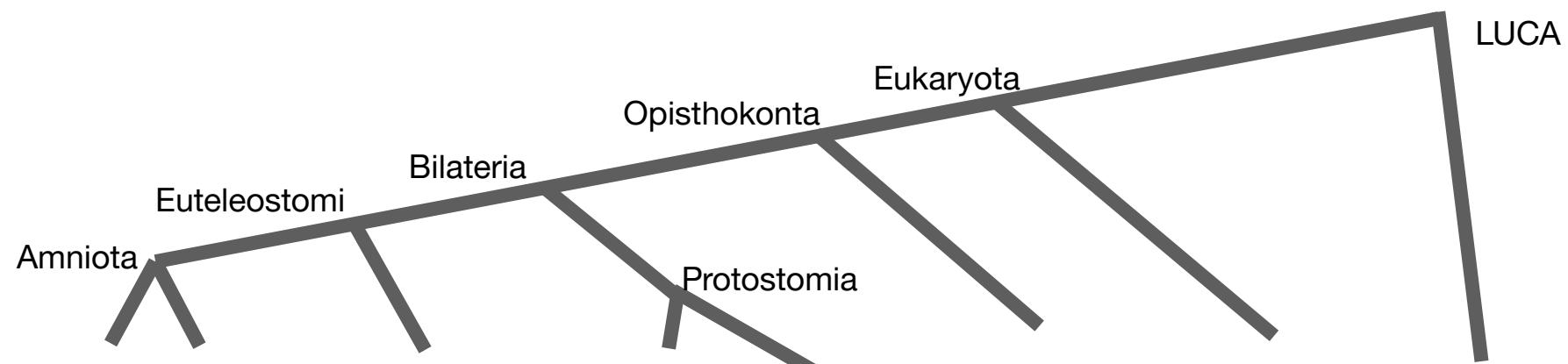








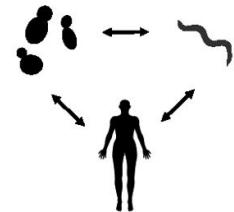
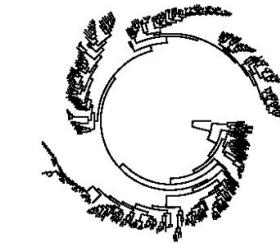
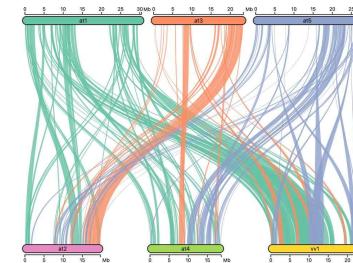
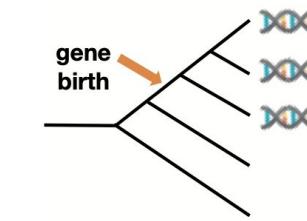




Orthology is fundamental to comparative genomics

Applications:

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

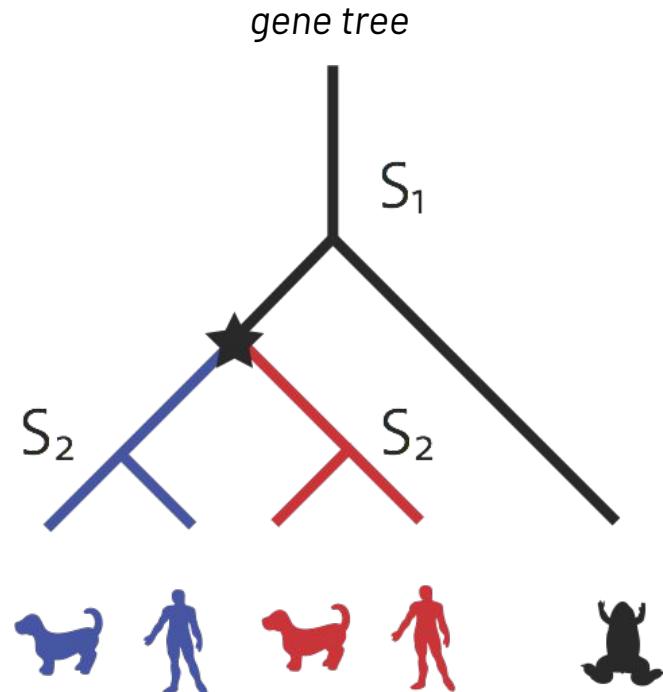


Orthology: What is it?

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**

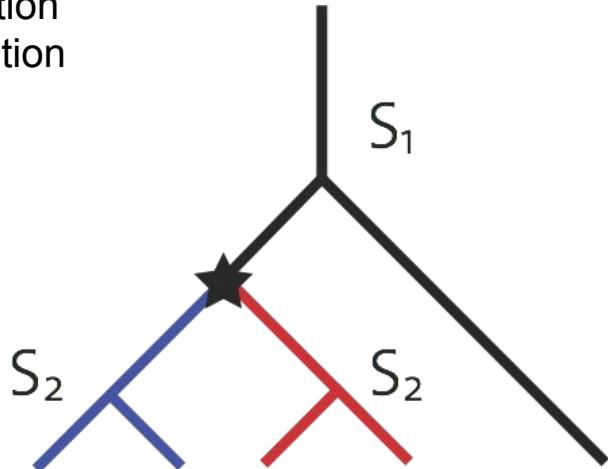
Ortho = exact
Para = beside/next to

Orthologs

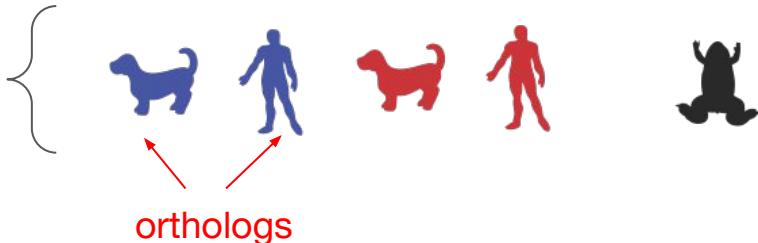
- Two genes in two species are **orthologous** if they derive from one gene in their last common ancestor
- Can be thought of as “corresponding genes” between species

S = speciation
★ = duplication

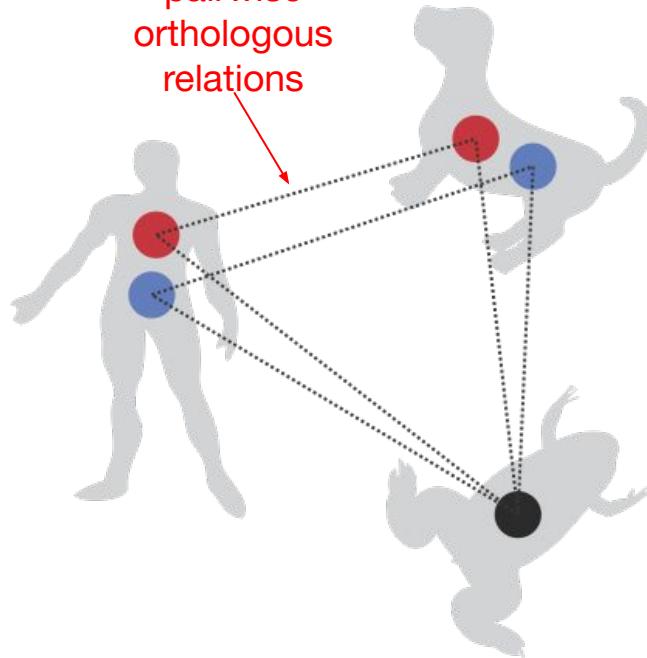
gene tree



these are genes



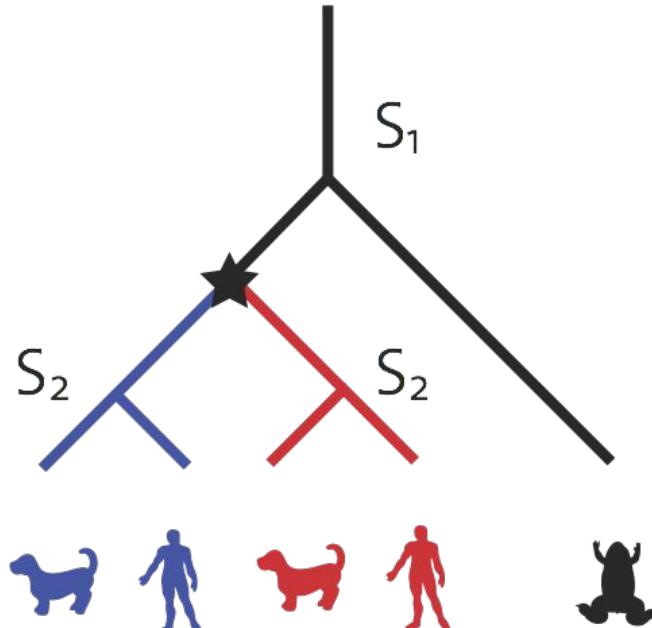
pairwise
orthologous
relations

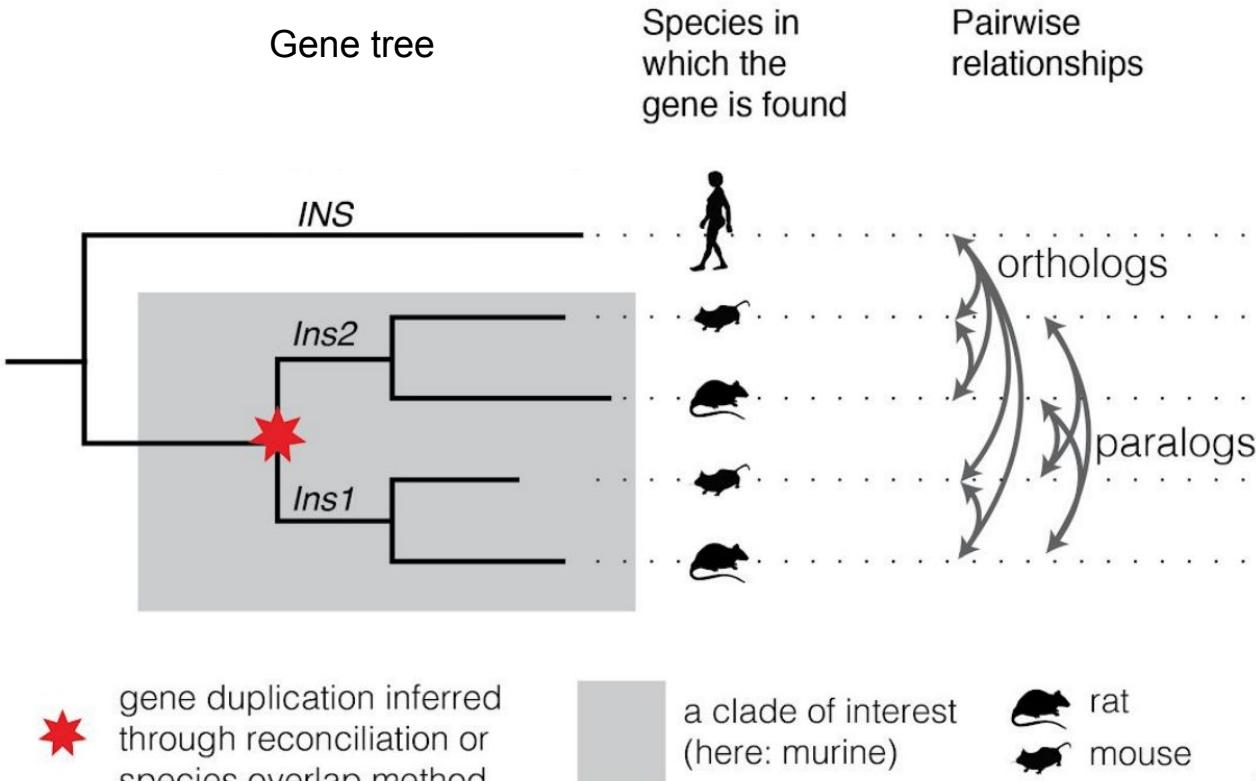


Common misconceptions

S = speciation
★ = duplication

- 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)

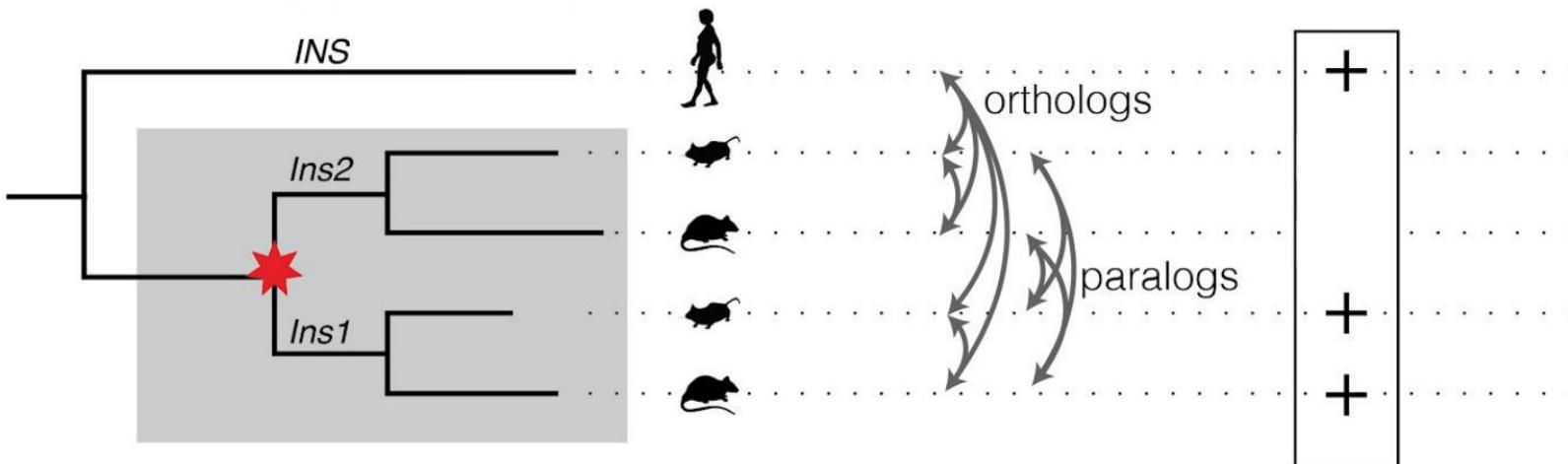
- Groups of genes which are all orthologous to each other
- Strict orthologous group
- Not necessarily 1:1 orthology, but each group contains at most one gene per species

Gene tree

Species in which the gene is found

Pairwise relationships

Example of strict orthologous group



gene duplication inferred through reconciliation or species overlap method



a clade of interest
(here: murine)



rat
mouse



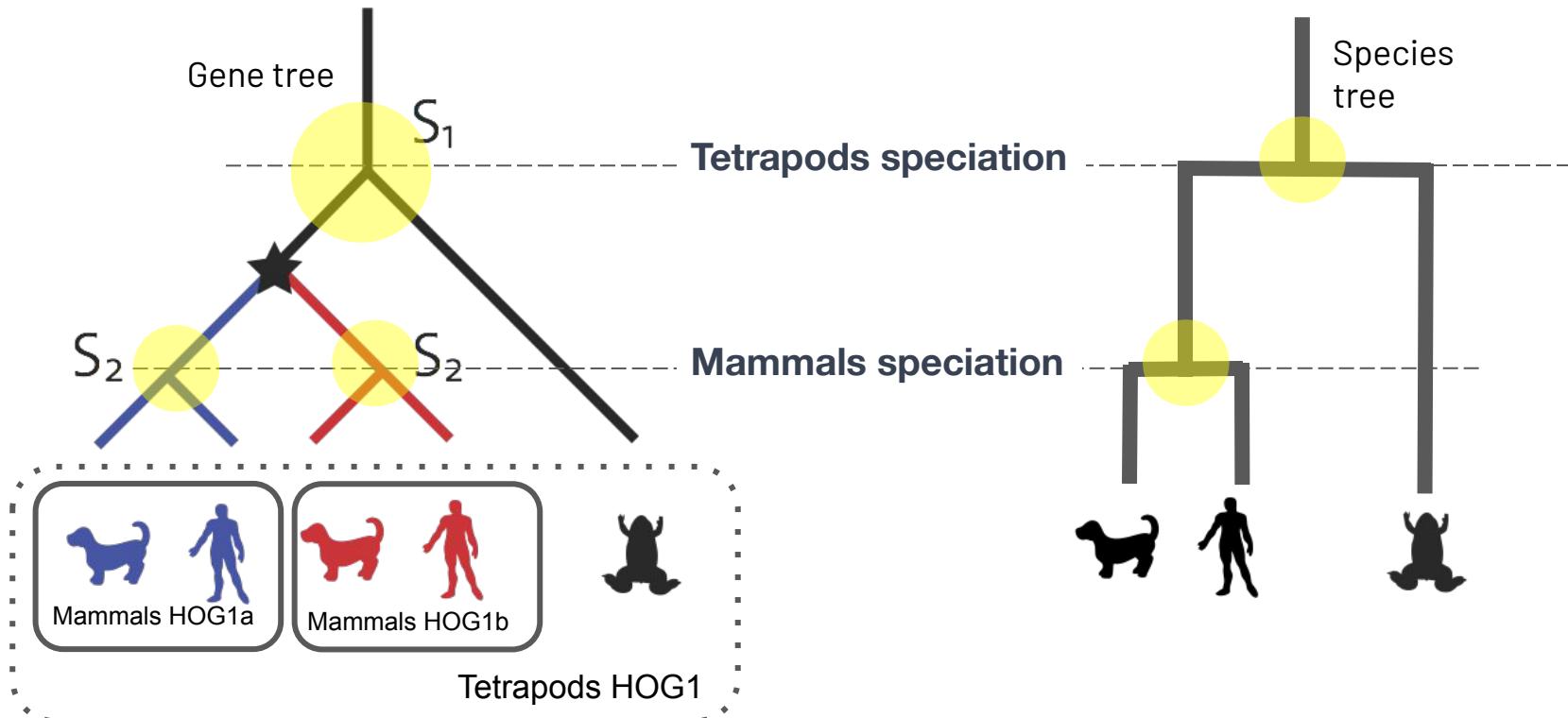
human



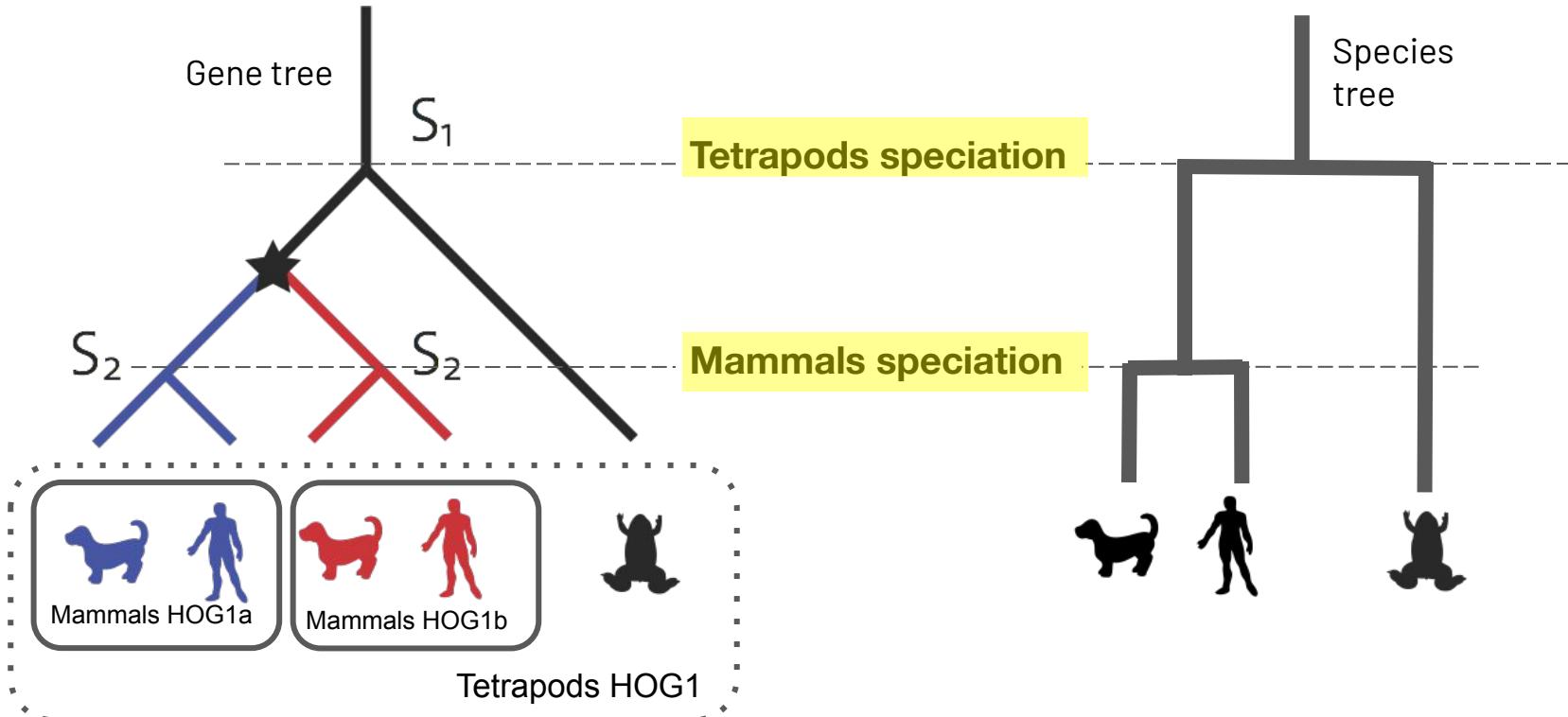
group membership

Hierarchical Orthologous Groups (HOGs)

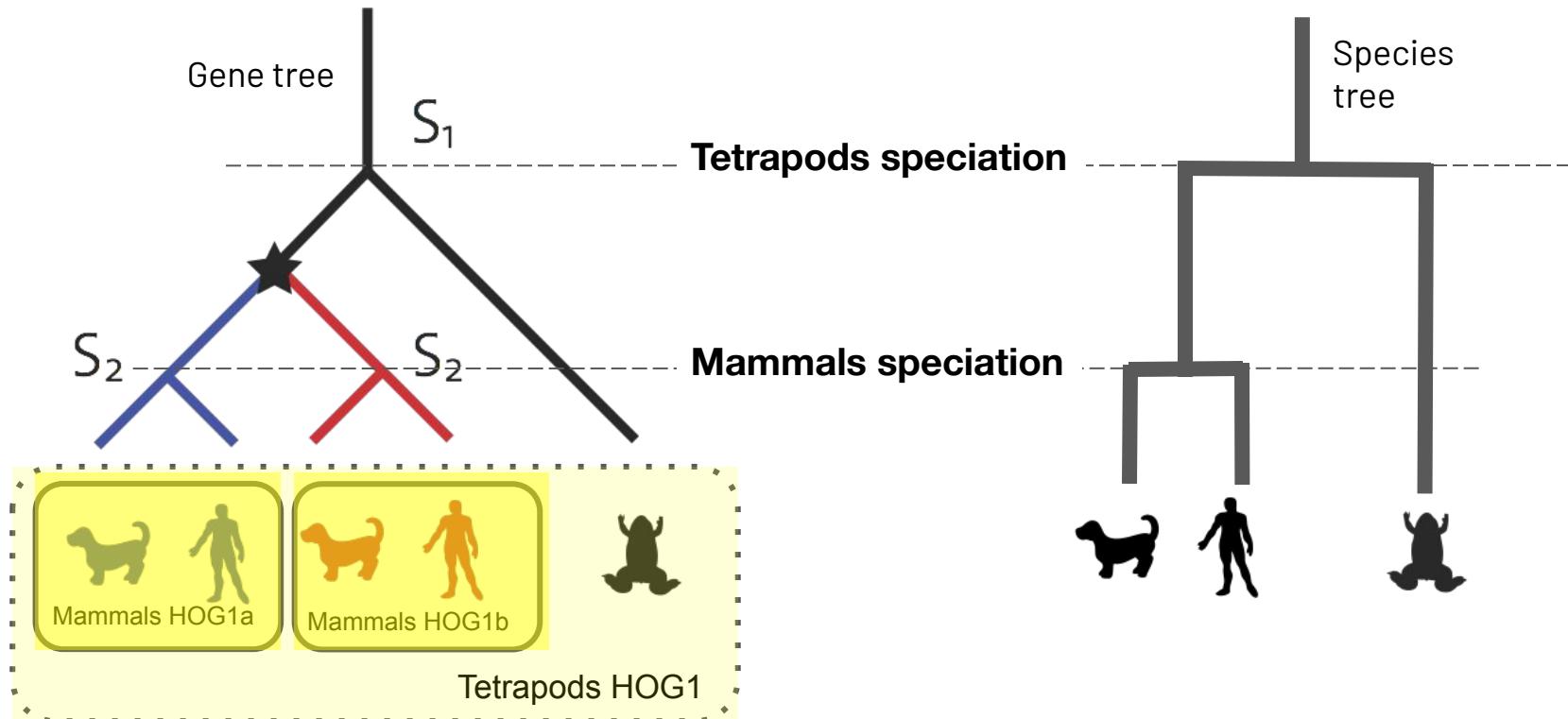
HOGs = Sets of genes that descended from a common ancestral gene in a given ancestral species



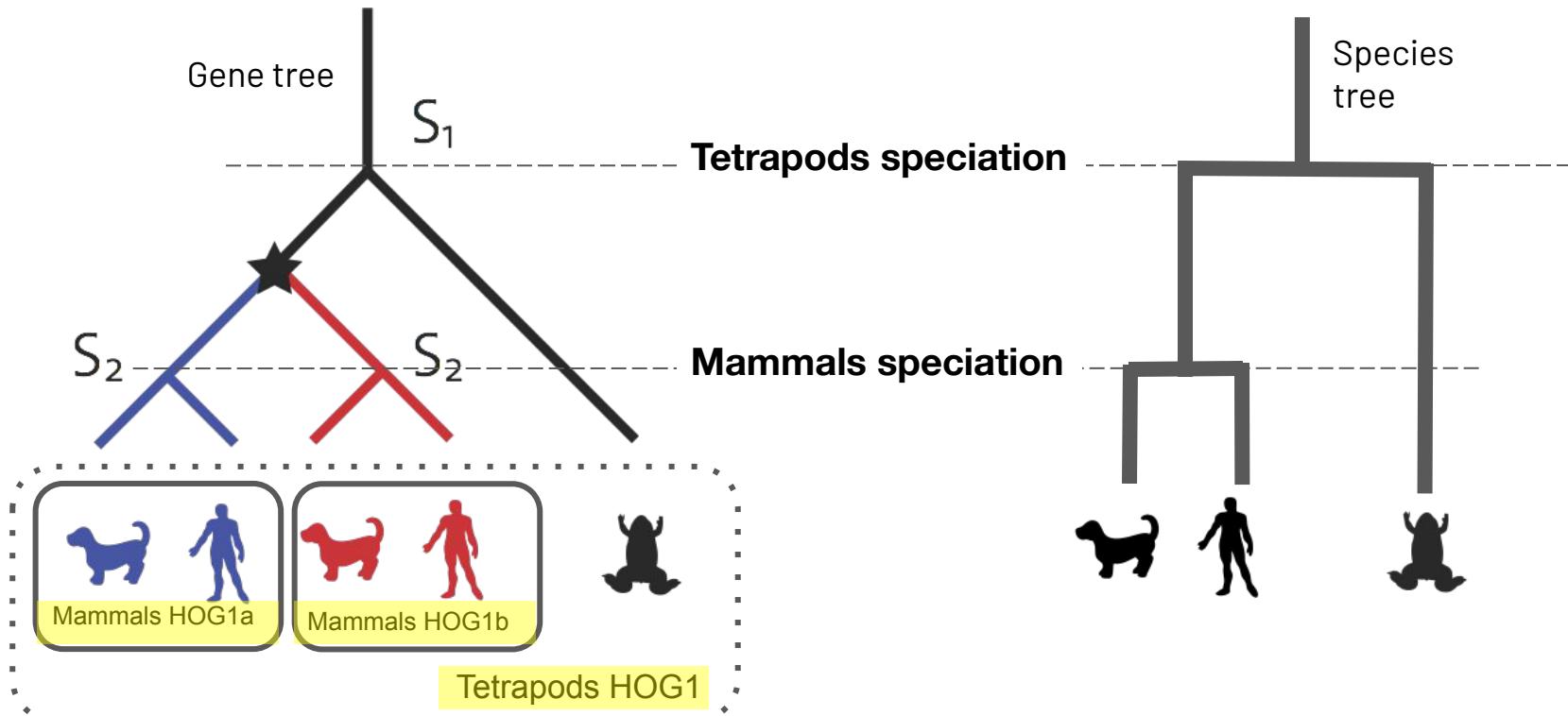
HOGs are defined with respect to specific clades



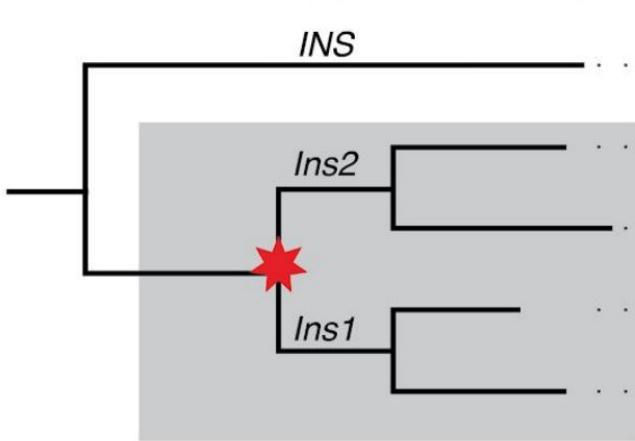
HOGs are hierarchical because groups defined with respect to deeper clades subsume multiple groups defined on their descendants



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

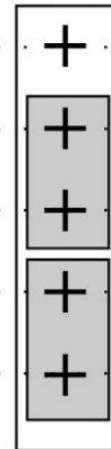
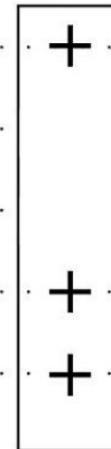


Pairwise
relationships

orthologs

paralogs

Example of
strict orthologous
group

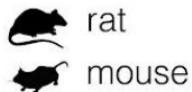


Hierarchical
Orthologous
Groups

★ gene duplication inferred
through reconciliation or
species overlap method



a clade of interest
(here: murine)



+ group membership

How to infer orthology?

Orthology Databases

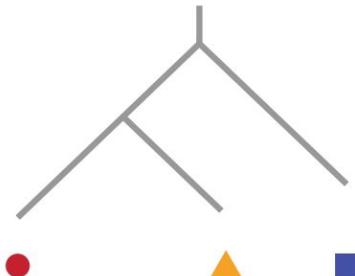
- OMA: <https://omabrowser.org/>
 - OrthoDB: <https://www.orthodb.org/>
 - EggNOG: <http://eggnogdb.embl.de>
 - InParanoid/Hieranoid:
<http://hieranoidb.sbc.su.se/>
 - COG: <https://www.ncbi.nlm.nih.gov/COG/>
 - PANTHER: <http://www.pantherdb.org/genes/>
 - PhylomeDB: <http://phylomedb.org/>
 - And more!
- Many different methods:
Namely *tree-based or graph-based*

Tree-based methods

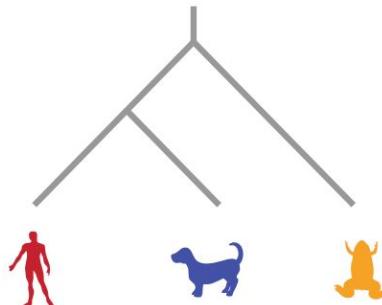
- Start with a group of homologous sequences
- Reconstruct a gene tree
- Infer the type of evolutionary event represented by each internal node of the tree

Gene tree/species tree reconciliation:

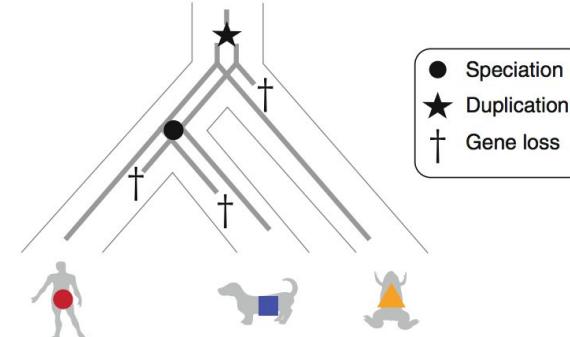
Gene Tree



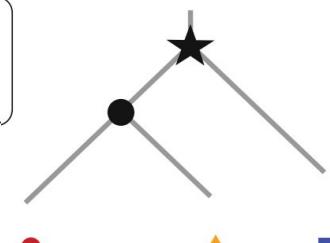
Species Tree



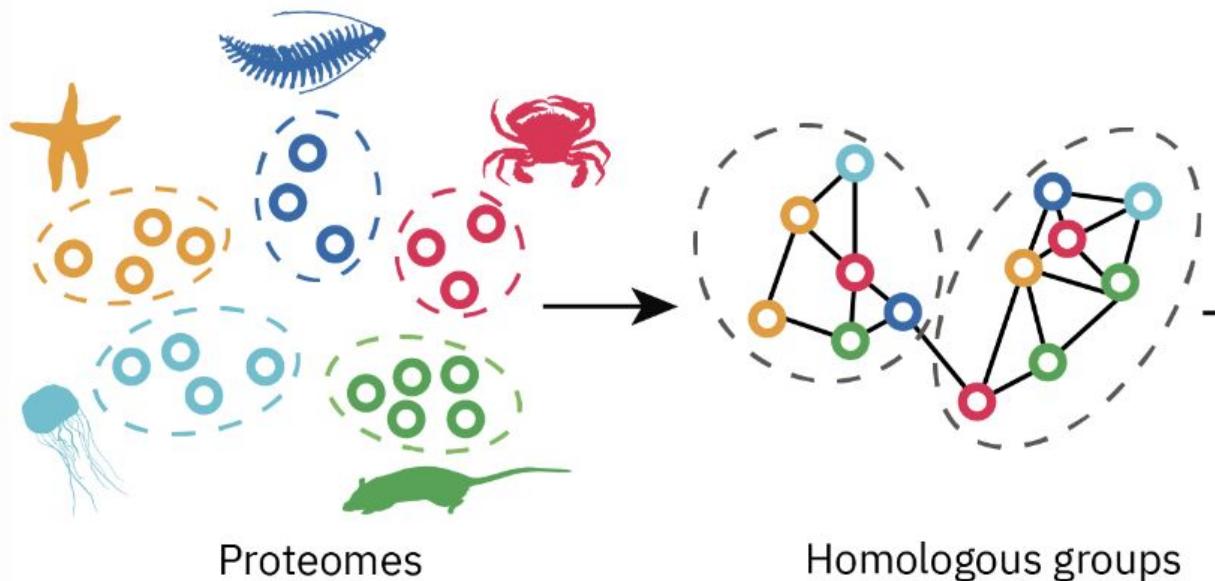
**Reconciled Tree
(Full Representation)**



**Reconciled Tree
(Simple Representation)**



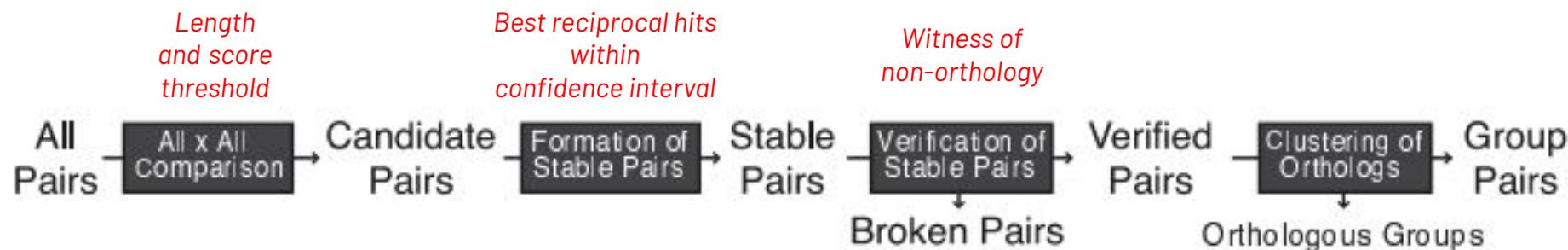
Graph-based methods



- 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

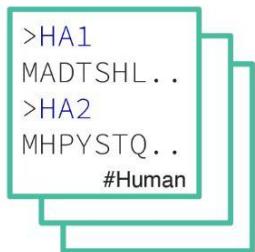


Pipeline:



FastOMA, our new tool

Input Proteomes

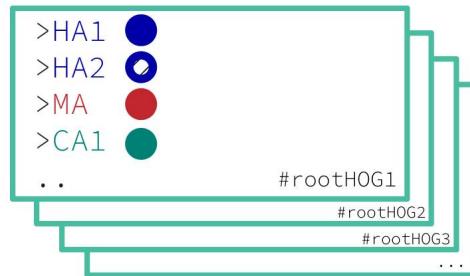


OMAmer

Mapping sequences
on OMA gene families
based on k-mers

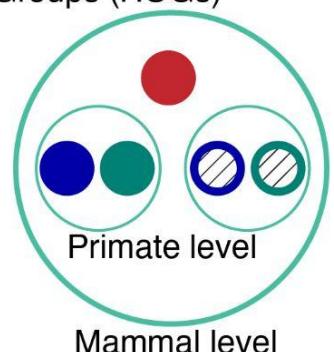


Root HOGs (Gene families)



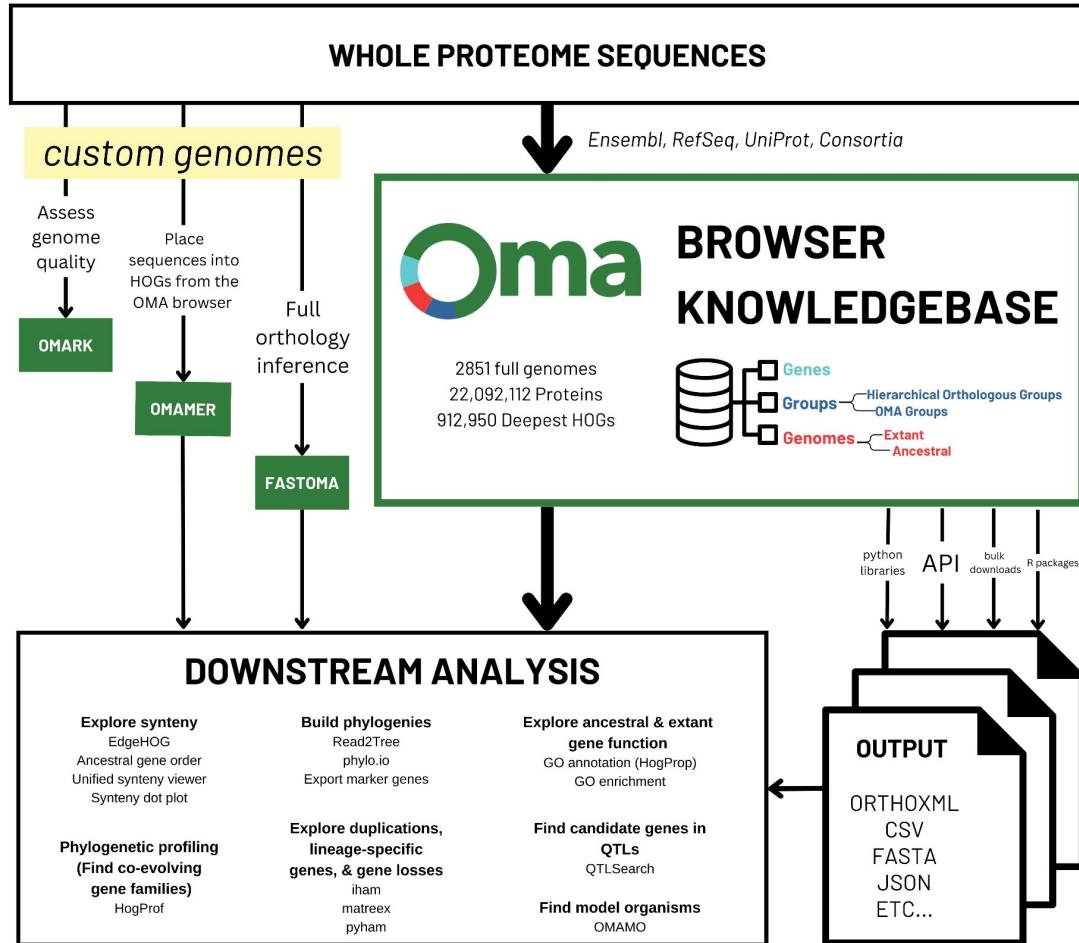
SubHOG/event
inference
(in parallel)

Hierarchical Orthologous Groups (HOGs)



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



The OMA browser

The screenshot shows the OMA browser interface. At the top left is the SIB logo and the Oma browser logo. At the top right are navigation links: Explore, Tools, Download, Help, and About. The main title "Oma orthologous matrix browser" is centered above a search bar. The search bar contains the query: "Blue-light photoreceptor" | proteinid:P53_RAT | species:"Drosophila melanogaster". Below the search bar is a "Examples" section with the text: "Entry P53_RAT - 'EWGKQSF' in Tetraodon - Search for "Blue-light photoreceptor" - "Drosophila melanogaster" species". A "SCROLL TO DISCOVER MORE" button is visible. At the bottom, a green footer bar displays statistics: 2,851 Full genomes, 22,092,112 Proteins, 1,251,567 OMA groups, 912,950 Deepest HOGs, and All.Jul2023 Release.

<https://omabrowser.org/>

<https://oma-stage.vital-it.ch/>
(Please use this one for the exercises)

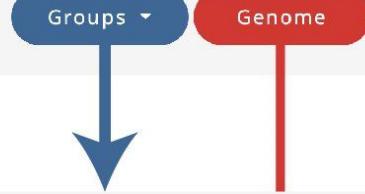
2,851 Full genomes 22,092,112 Proteins 1,251,567 OMA groups 912,950 Deepest HOGs All.Jul2023 Release

The OMA Database: 3 kinds of pages

genes

Gene LEIIN03631 (A4I3J2)

E Leishmania infantum is on Chromosome 28.



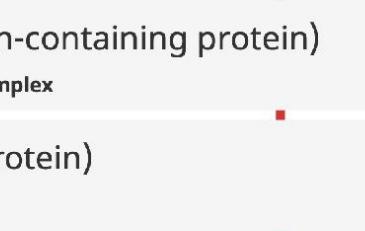
groups

HOG:0495624.7c with 2 members (DYW_deaminase domain-containing protein)

Eukarya / Kinetoplastida / Trypanosomatidae / Leishmaniinae / Leishmania / Leishmania donovani species complex

OMA GROUP 402875 with 9 members. (Uncharacterized protein)

Fingerprint: MRFARFE



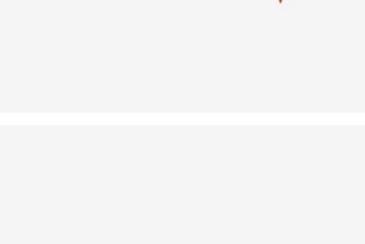
genomes

LEIIN - Leishmania infantum

Proteome version: 03-JAN-2001 (Rel. 66, Last updated, Version 1) with 8031 proteins.

Ancestral genome of Leishmaniinae

with 6 descendant species and 7893 ancestral genes (HOGs).



Hierarchical Orthologous Groups (HOGs)

HOG:D0606964 with 42 members (Ig-like domain-containing protein)

Primates / Lower Level ▾

Completeness score: 0.75 ⓘ

Ancestral Genome

Hierarchical group HOG:0606964 open at level of Primates

OPTIONS ▾

Graphical viewer

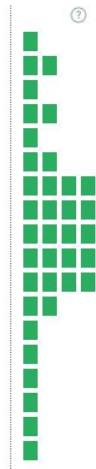
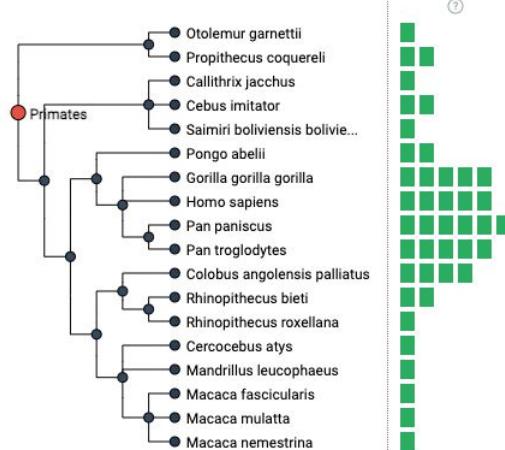
Members

Ancestral GO

Alignment

Ancestral synteny

Similar 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)

Completeness score: 0.75 ⓘ Ancestral Genome

Primates / Lower Level ▾

[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

Ancestral genome of Primates

with 24 descendant species and 38457 ancestral genes (HOGs).

Remove HOGs with completeness score below

Search grid icon download icon

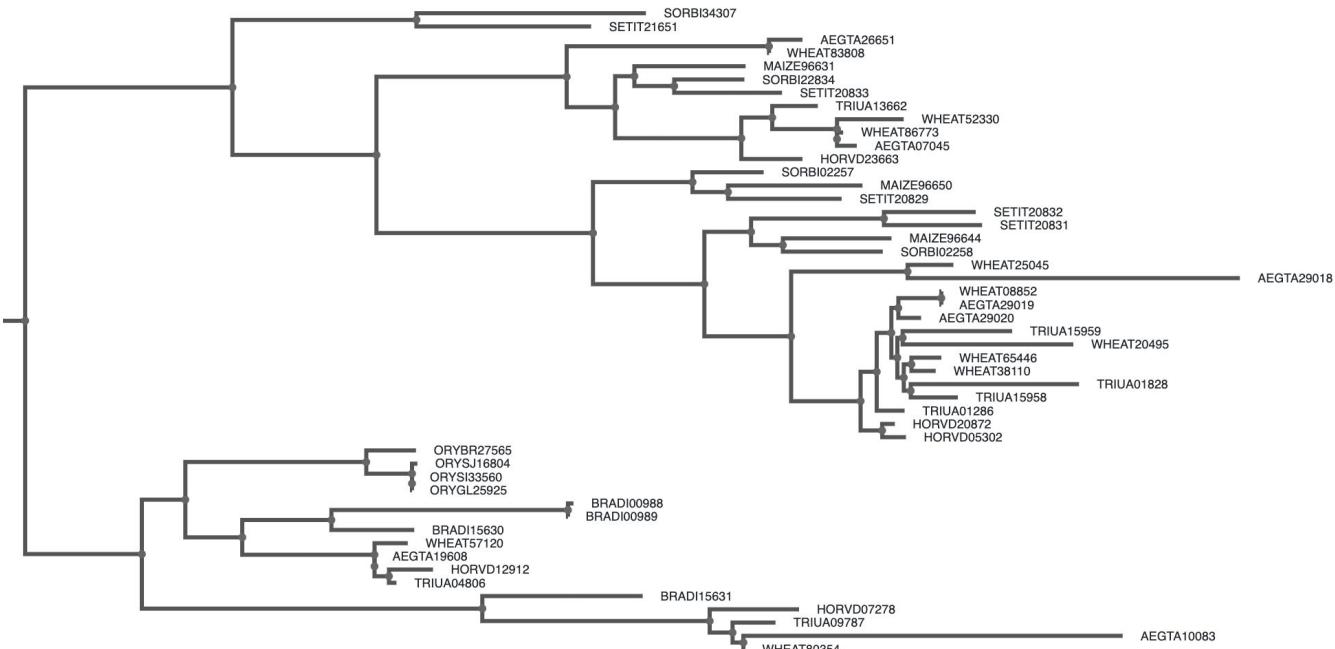
HOG ID	Root HOG ID	Completeness	Nr genes in HOG	Description
HOG:D0912633.1a	HOG:D0912633	1.00	24	autophagy related 16 like
HOG:D0912535.3b.8a.7b.4b	HOG:D0912535	1.00	24	leukocyte cell derived chemotaxin
HOG:D0911480.5b.3b	HOG:D0911480	1.00	24	prostaglandin G/H synthase
HOG:D0911480.5a.2b	HOG:D0911480	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

38

OMA's Downstream analyses

Make gene trees*

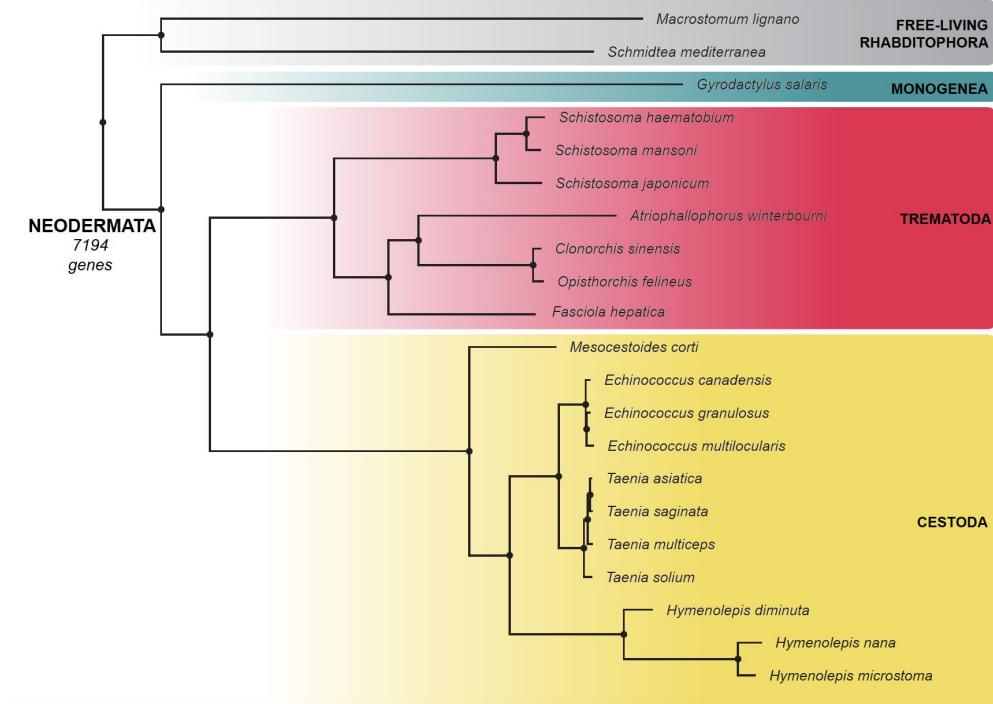
Create **gene** trees
for a clade of
interest using the
HOG at
that taxonomic level



*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

GO ENRICHMENT

Gene Ontology Function Projection

Email Email
We will send an email to this address once the predictions are ready.

Name of Dataset Name of Dataset

Sequence File (fasta format)* Choose file | No file chosen

Captcha* I'm not a robot  reCAPTCHA
Privacy · Terms

Submit

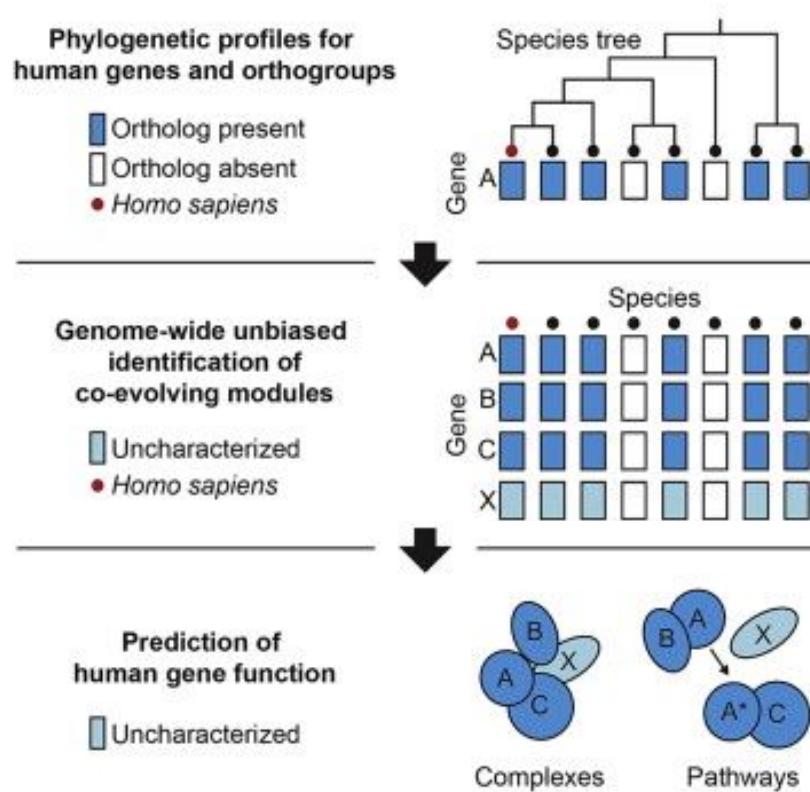
Transfer Gene Ontology annotations from genes in the same HOG from well-studied model species to other species

You can download the detailed result for this GO enrichment analysis by clicking [here](#)

GO_ID	GO_Name	study_count	pop_count	Study_Entries	study_n	p_uncorrected	p_bonferroni	p_fdr_bh
GO:0033041	sweet taste receptor activity	6	32	DROME01345, DROME03317, DROME04384, DROME04694, DROME05340, DROME05633	6	1.428750553e-16	0	0
GO:0001582	detection of chemical stimulus involved in sensory perception of sweet taste	6	34	DROME01345, DROME03317, DROME04384, DROME04694, DROME05340, DROME05633	6	2.120447249e-16	0	0

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

Phylogenetic profiling



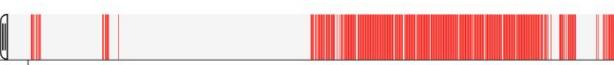
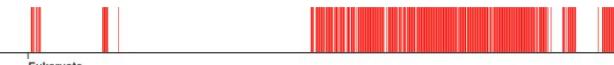
Phylogenetic profiling

HOG:D0679560 with 323 members (intraflagellar transport protein 57 homolog) Completeness score: 0.44 ⓘ Ancestral Genome

Eukaryota / Lower Level ▾

This HOG has 49 similar HOGs:

Co-evolving HOGs

HOG ID	Jaccard similarity ⓘ	Barcode	Description
HOG:D0680756	0.8515625		intraflagellar transport
HOG:D0680535	0.8515625		intraflagellar transport
HOG:D0680689	0.84375		intraflagellar transport

Graphical viewer

Members

Ancestral GO

Alignment

Ancestral synteny

Similar HOGs

BASED ON:

Similar domains

Search

Reset

Related functions

Gene content of ancestral genomes

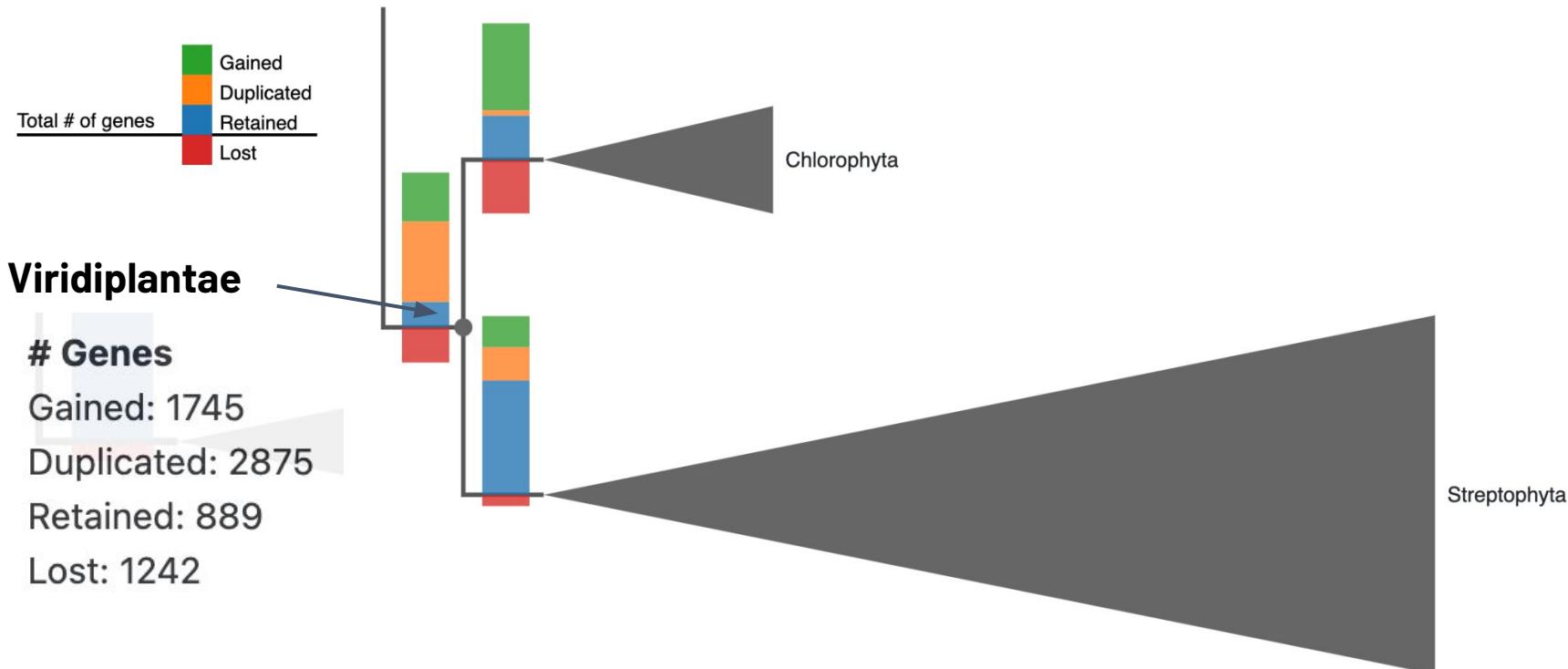
Ancestral genome of Viridiplantae

with 85 descendant species and 42339 ancestral genes (HOGs).

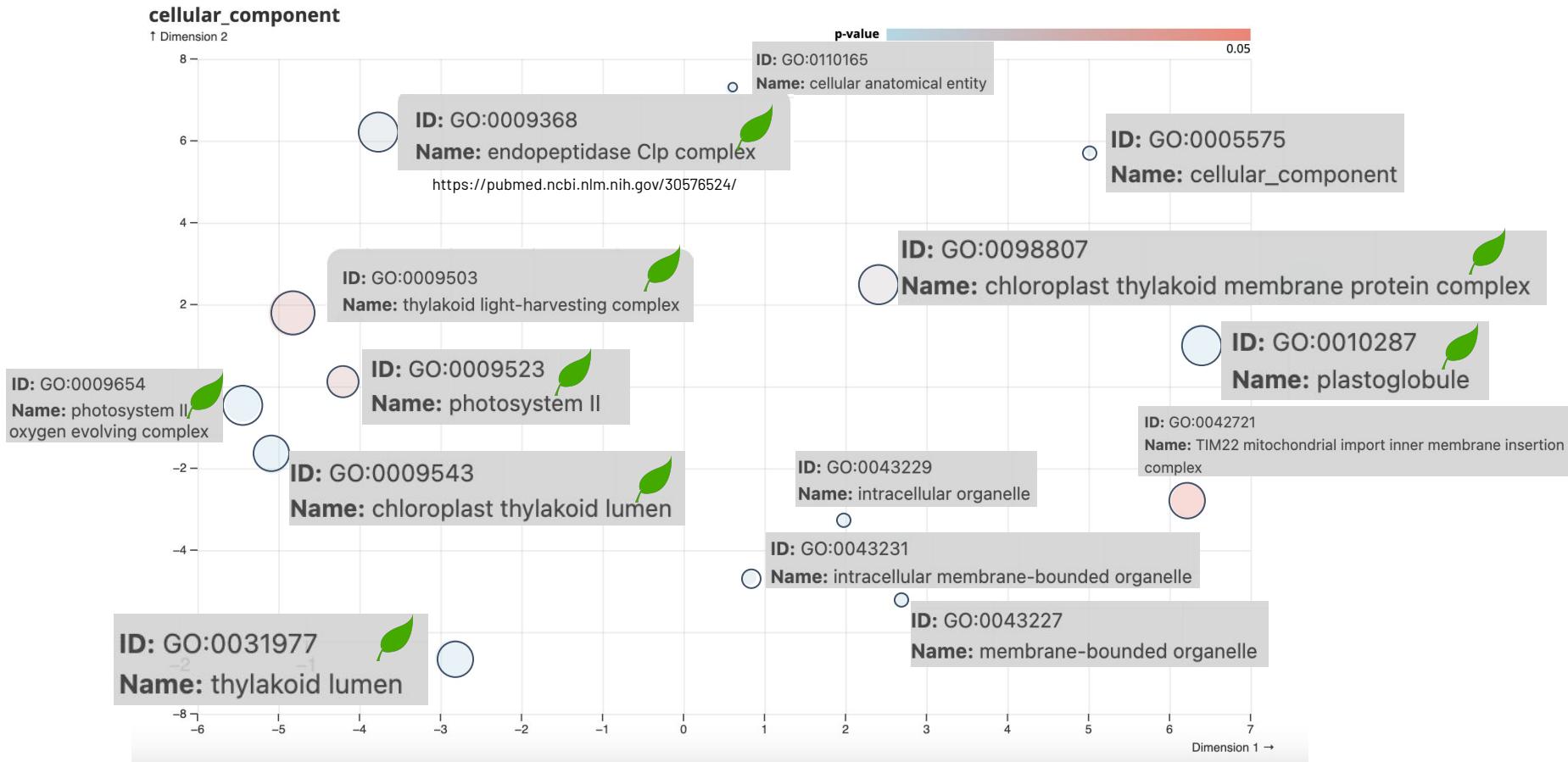
Completeness Score =
Number of species in the HOG/
number species in the clade

Remove HOGs with completeness score below 0.3						Search	grid icon	download icon	
Genome information	HOG ID	Root HOG ID	Evolutionary event	Completeness	Nr genes in HOG	Description			
Ancestral genes	HOG:D0271083	HOG:D0271083	gained	1.00	454	EGF-like calcium-binding domain-containing protein			
	HOG:D0271236	HOG:D0271236	gained	0.988	302	Hyaluronan/mRNA-binding protein domain-containing protein			
	HOG:D0271378	HOG:D0271378	gained	0.976	235	t-SNARE coiled-coil homology domain-containing protein			
	HOG:D0271254	HOG:D0271254	gained	0.976	104	ATP-dependent Clp protease proteolytic subunit			
	HOG:D0271213	HOG:D0271213	gained	0.976	145	Chlorophyll a-b binding protein chloroplastic			
	HOG:D0271177	HOG:D0271177	gained	0.976	276	phosphoglycerate mutase-like protein			

Tracking the evolutionary history of a gene family



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://oma-stage.vital-it.ch/oma/academy/>

Module 1: Finding orthology with the OMA Browser

The OMA browser serves as an access point for the OMA database, which contains precomputed homology data for over extant and ancestral genomes for over 2800 species (see the [latest list of species](#)).

The OMA browser focuses on three main data types: genes, groups, and genomes. Gene-centric pages provide detailed information about a specific gene, including its sequence, cross-references, functional annotations, and evolutionary data. Group-centric pages classify genes into OMA Groups (Orthologous Groups; OGs) and Hierarchical Orthologous Groups (HOGs) to define families and subfamilies. and Hierarchical Orthologous Groups (HOGs) to define families and subfamilies. Genome-centric pages offer information about extant or ancestral species, associated genes, related genomes, and a synteny viewer.

[Back to home / Reset](#)

1.1. Browsing the gene page

1.2. Exploring Hierarchical Orthologous Groups

1.3. Browsing the Genomes page

- Break at 10:45



Gartner: 60% of cloud workloads will be built using C



Gitpod

CDEs

Resources ▾

Solutions ▾

Platform ▾

Cust

Always ready-to-code.

The developer platform for on-demand cloud
development environments. Create software faster



☰ README.md

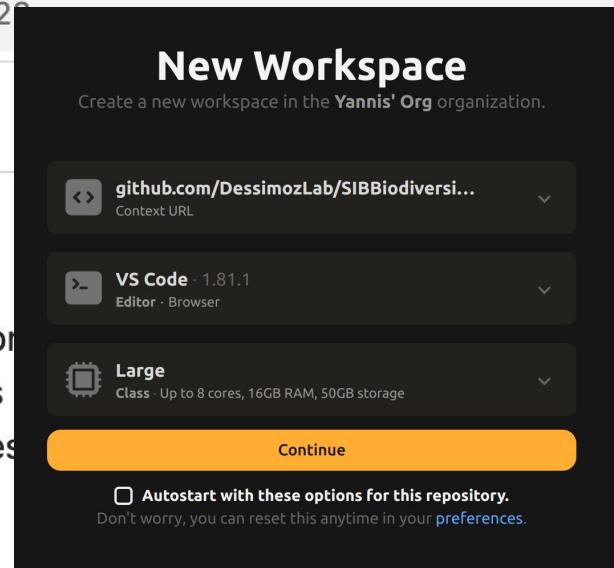
Software

We will be mainly working on an [GitPod](#), an online integrated development environment that allows users to write, edit, and run code directly in a web browser. GitPod is a browser-based IDE. All the software, code, and files needed for the course are stored and processed in the cloud, so you won't 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 SIBBiodiversityGenomics2023, 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



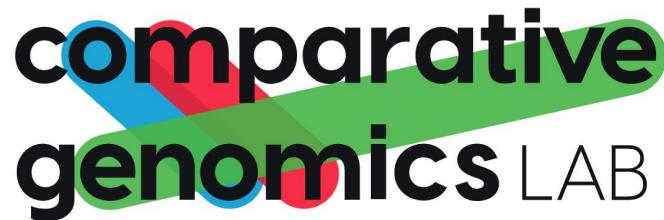
Continue



Autostart with these options for this repository.

Don't worry, you can reset this anytime in your [preferences](#).

Module 2: OMAmer for sequence placement into HOGs



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 OM Amer?

- ❖ Fast sequence placement into existing HOGs from the OMA Browsers
- ❖ More accurate than closest sequence matching for subfamily placement!

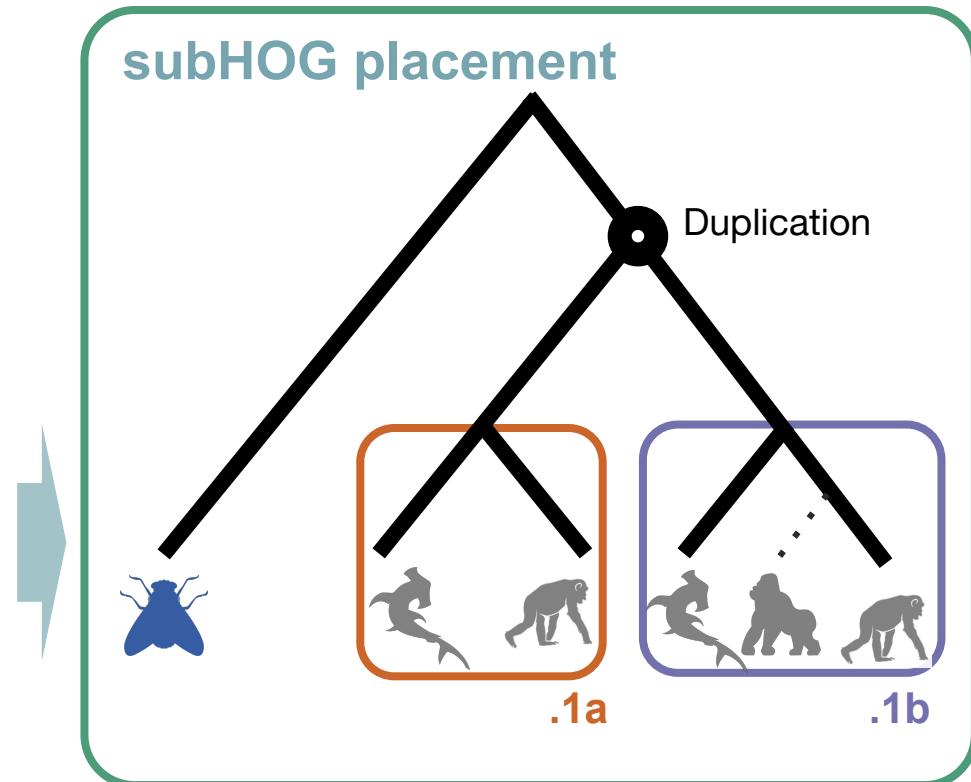
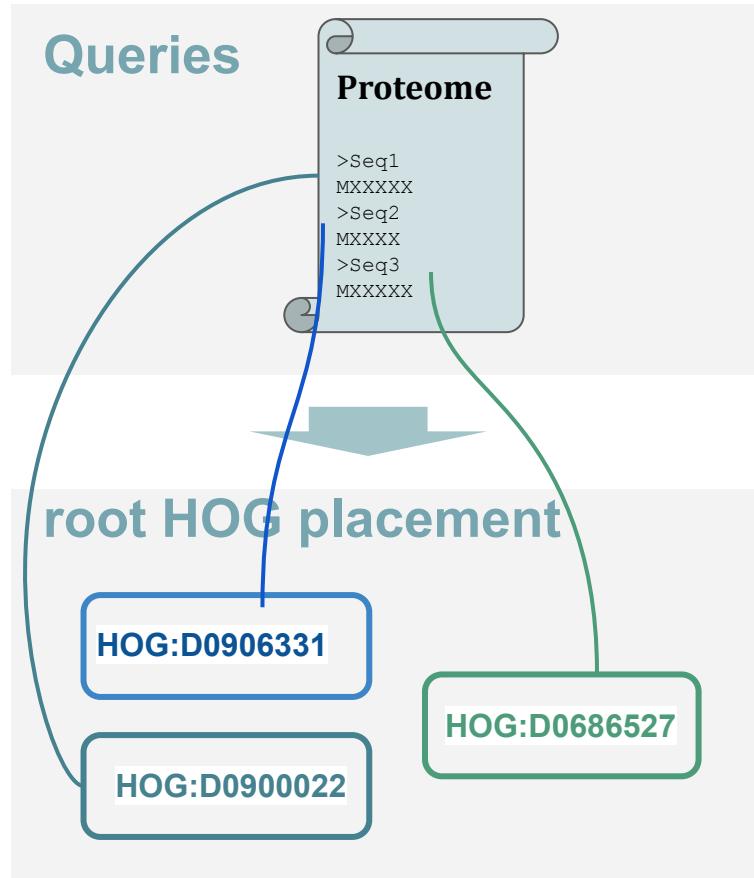
OM Amer: tree-driven and alignment-free protein assignment to subfamilies outperforms closest sequence approaches

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 sequences

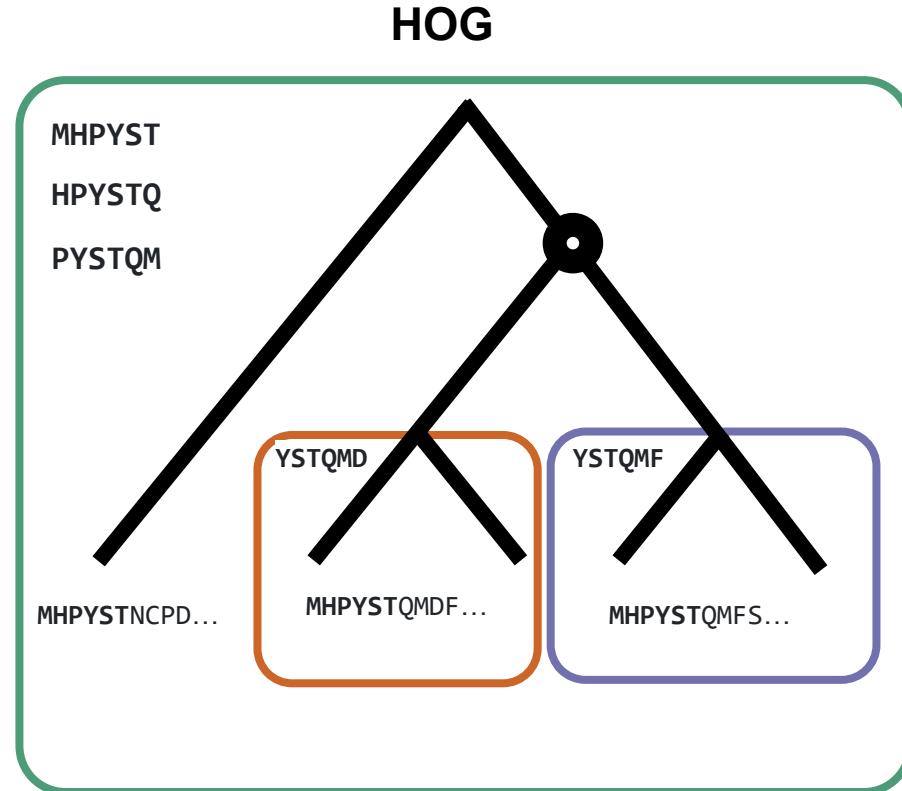
Query sequence

MHPYSTQMFS LQITVMEDSQ SDMSIELPLS

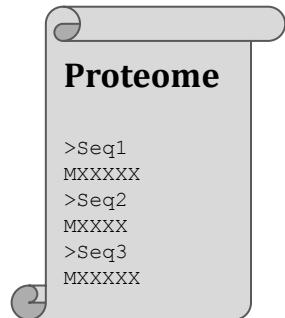
MHPYST
HPYSTQ
PYSTQM

...
...
...

MSIELP
SIELPL
IELPLS



How to use OMAMer



OMAMer database
HDF5 format
Built with HOGs from the OMA Browser

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

Query sequences
FASTA format

From any species

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	overlap	family-score	subfamily-score	qseqlen	subfamily-medianseqlen
Seq1	HOG:D0681219.3c	1.0	0.99701[...]	0.99701[...]	966	1065
Seq2	HOG:D0433152	1.0	0.99967[...]	0.99967[...]	303	334

- ❖ qseqid : Query identifier

Interpreting the output

qseqid	hogid	overlap	family-score	subfamily-score	qseqlen	subfamily-medianseqlen
Seq1	HOG:D0681219.3c	1.0	0.99701[...]	0.99701[...]	966	1065
Seq2	HOG:D0433152	1.0	0.99967[...]	0.99967[...]	303	334

- ❖ qseqid: Query identifier
- ❖ **hogid : Where the query is placed**

Interpreting the output

qseqid	hogid	overlap	family-score	subfamily-score	qseqlen	subfamily-medianseqlen
Seq1	HOG:D0681219.3c	1.0	0.99701[...]	0.99701[...]	966	1065
Seq2	HOG:D0433152	1.0	0.99967[...]	0.99967[...]	303	334

- ❖ qseqid: Query identifier
- ❖ hogid : Where the query is placed
- ❖ **Family score/subfamily score :** Proportion of k-mer of the sequence in common with HOG
(- Proportion expected by chance)

Interpreting the output

qseqid	hogid	overlap	family-score	subfamily-score	qseqlen	subfamily-medianseqlen
Seq1 HOG:D0681219.3c		1.0	0.99701[...]	0.99701[...]	966	1065
Seq2 HOG:D0433152		1.0	0.99967[...]	0.99967[...]	303	334

- ❖ qseqid: Query identifier
- ❖ hogid : Where the query is placed
- ❖ Family score/subfamily score : Proportion of k-mer of the sequence in common with HOG
(- Proportion expected by chance)
- ❖ **Overlap: distance between the first and last matching kmer on the sequence (0 to 1)**

Interpreting the output

qseqid	hogid	overlap	family-score	subfamily-score	qseqlen	subfamily-medianseqlen
Seq1 HOG:D0681219.3c		1.0	0.99701[...]	0.99701[...]	966	1065
Seq2 HOG:D0433152		1.0	0.99967[...]	0.99967[...]	303	334

- ❖ qseqid: Query identifier
- ❖ hogid : Where the query is placed
- ❖ Family score/subfamily score : Proportion of k-mer of the sequence in common with HOG (- Proportion expected by chance)
- ❖ Overlap: distance between the first and last matching kmer on the sequence (0 to 1)
- ❖ **qseqlen and subfamily-medianseqlen : length of proteins in amino acids**

To remember

- ❖ Placement into HOGs allows 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 to still **take advantage of OMA Browser** wealth of data and features



OMA Academy

Welcome to the OMA Academy! Here, you will find online exercises which will aid you in becoming more familiar with orthology, phylogenies, and comparative genomics.

BACKGROUND

OMA ("Orthologous MAtrix") is a method and database for the inference of orthologs among complete genomes. It can be found at omabrowser.org. Many of the exercises use the OMA browser as a starting point. The OMA pipeline can also run on custom genomic/transcriptomic data using the OMA stand-alone software, and it is even possible to combine precomputed data with custom data by exporting parts of the OMA database.

Tables of contents

1. [Exploring Orthology with the OMA Browser](#)
2. [OMAmer](#)
3. [FastOMA](#)
4. [Estimating a Species Tree](#)

Module 2: Fast placement of sequences into HOGs with OMAmer

Sometimes you might have a few protein sequences from a genome which is not in the OMA database and you want to quickly find out which genes they share homology with. Or perhaps you even want to do this with a whole proteome.

OMAmer is a command-line software that places a given protein sequence onto one of the gene families available in the input OMA database. In other words, OMAmer finds the most likely HOG where the input protein belongs. OMAmer is based on comparing *k*-mers (substring of the sequence of *k* length) between a query sequence and HOGs. Since it only searches for *k*-mers that are in common between sequences, it does not need a sequence alignment (which is usually computationally intensive) and is a very fast alternative to high-resolution homology determination when one is simply looking for the gene family a sequence belongs to.

[Back to home / Reset](#)

2.1 OMAmer setup and requirements	▼
2.2 Placing a few sequences into Hierarchical Orthologous Groups	▼
2.3 Placing a whole proteome	▼

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/SIBBiodiversityBioinformatics2023/Module3_FastOMA/working_dir/in_folder/proteome.



Click to go back, hold to see history

.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
```

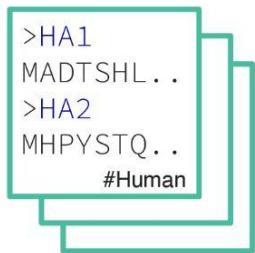


Module 3: FastOMA



FastOMA, our new tool

Input Proteomes

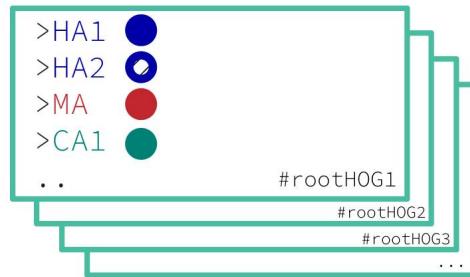


OMAmer

Mapping sequences
on OMA gene families
based on k-mers

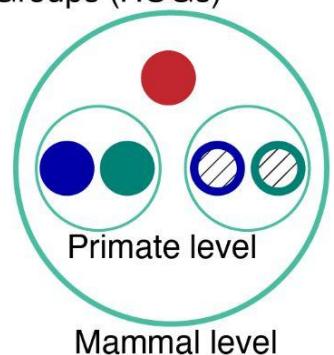


Root HOGs (Gene families)

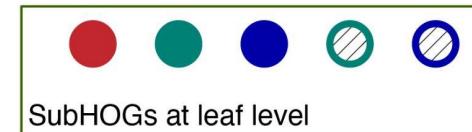
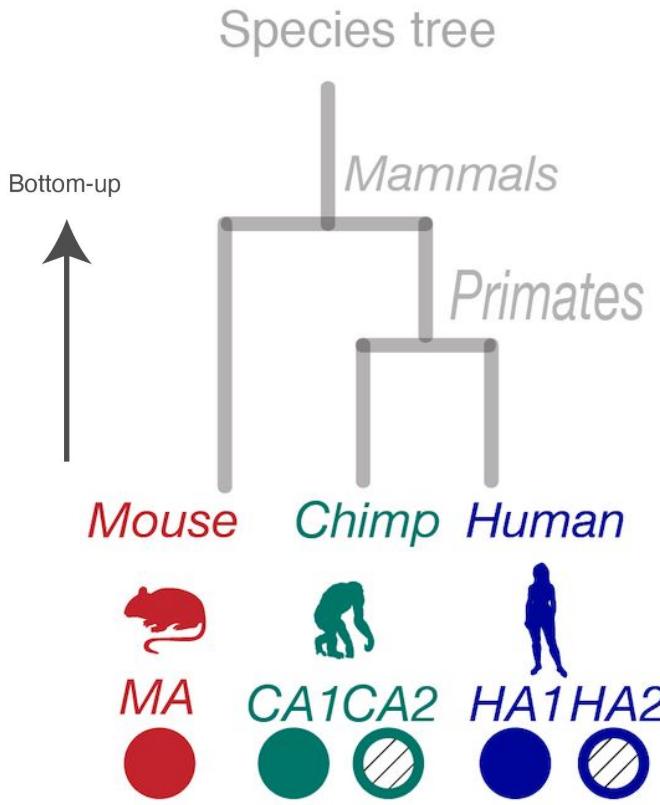


SubHOG/event
inference
(in parallel)

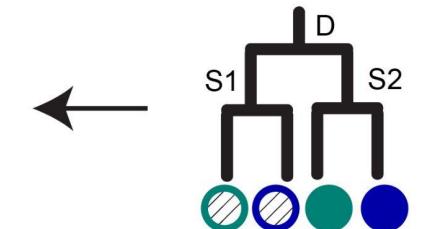
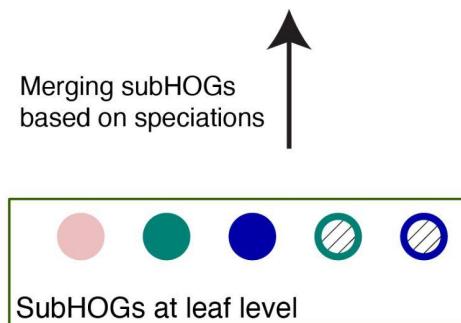
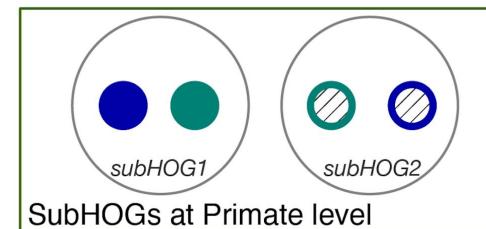
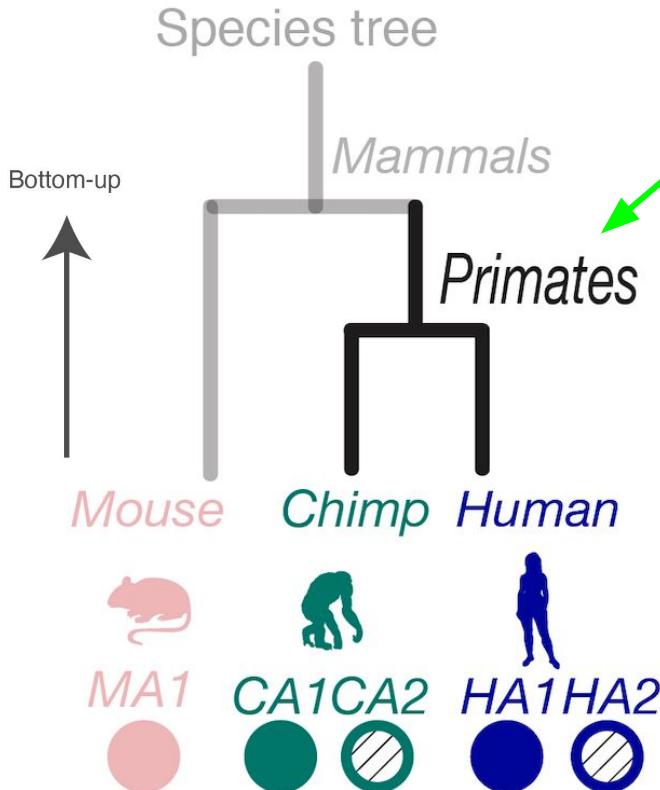
Hierarchical Orthologous Groups (HOGs)



HOG inference



HOG inference



Merging subHOGs
based on speciations

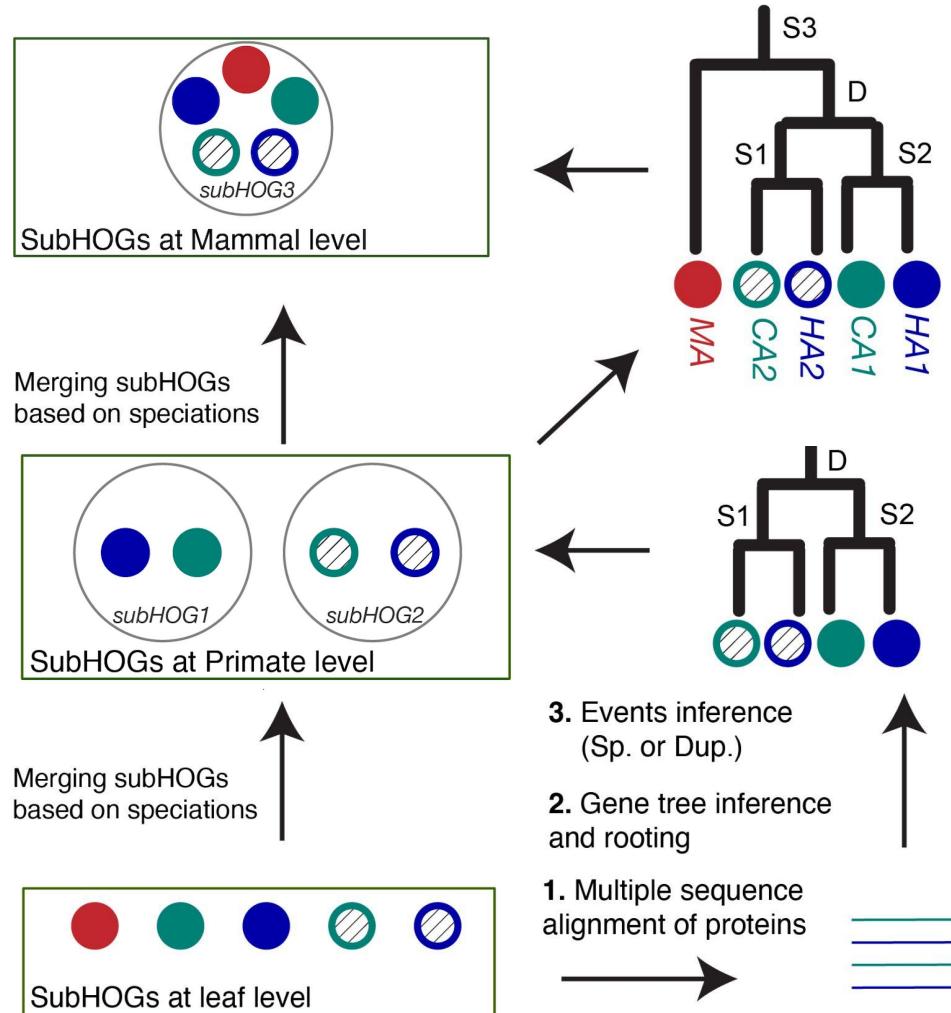
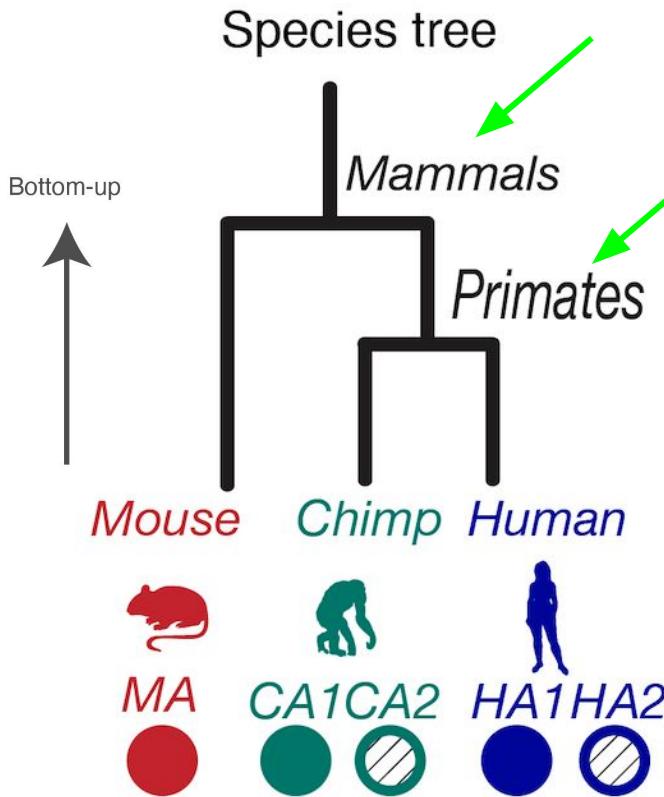
3. Events inference
(Sp. or Dup.)

2. Gene tree inference
and rooting

1. Multiple sequence
alignment of proteins

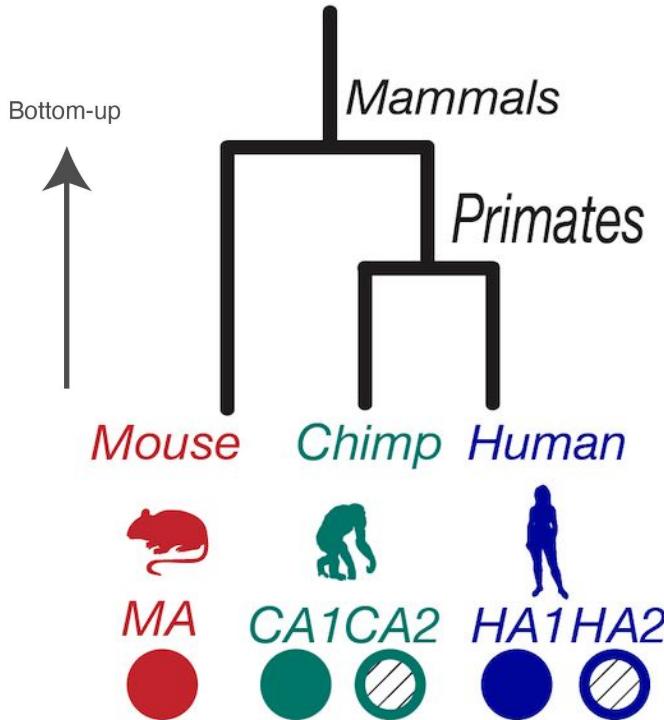


HOG inference



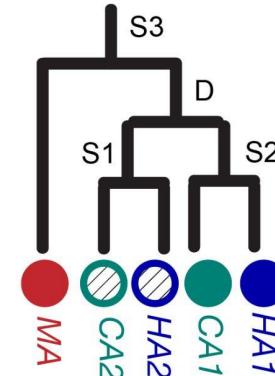
HOG inference

Species tree

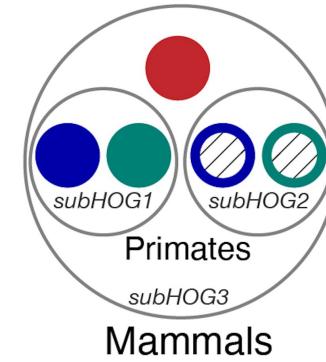


HOG in orthoXML

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



Nested structure of HOG

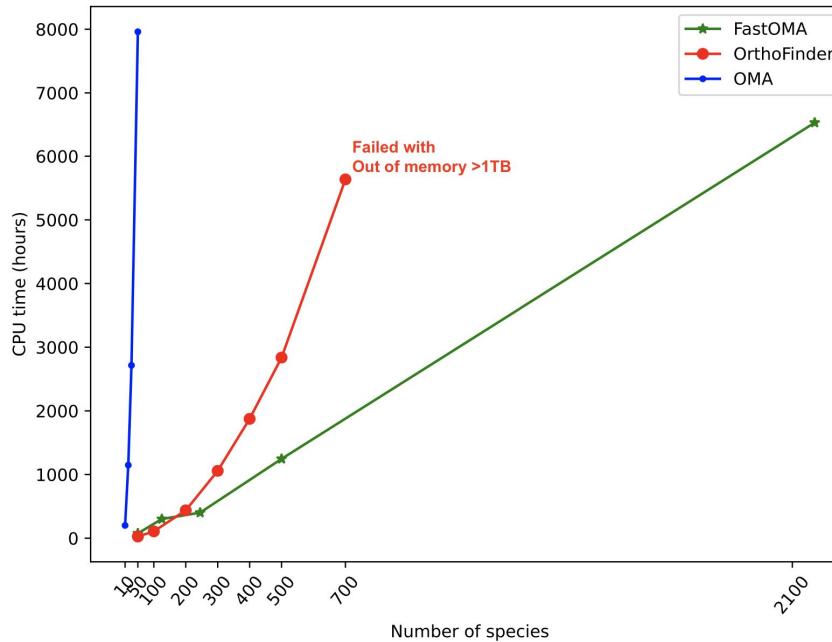


Orthology inference for Eukaryote dataset

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



[github.com/DessimozLab/
FastOMA](https://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.

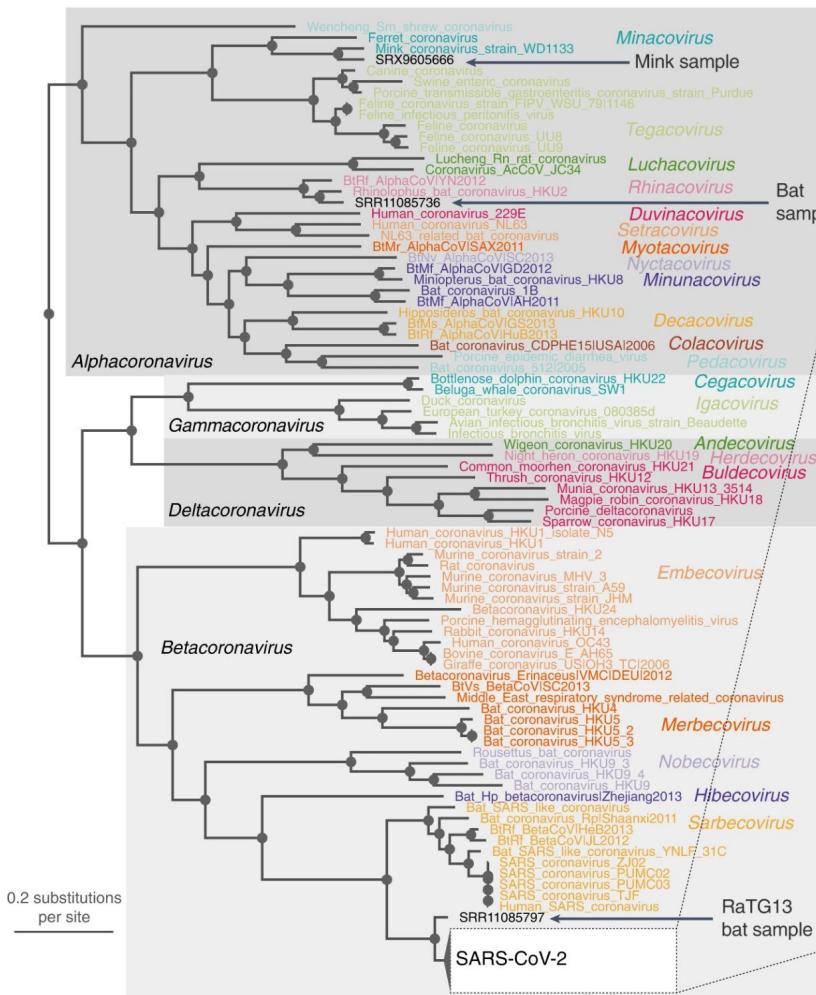
- 5. Consider the gene “60S ribosomal protein L15-A” in *Schizosaccharomyces pombe* with protein ID: RL15A_SCHPO. How many proteins are in the gene family (for these 5 species of interest)?

Module 4 : Building a species tree

A tree of Coronaviridae viruses

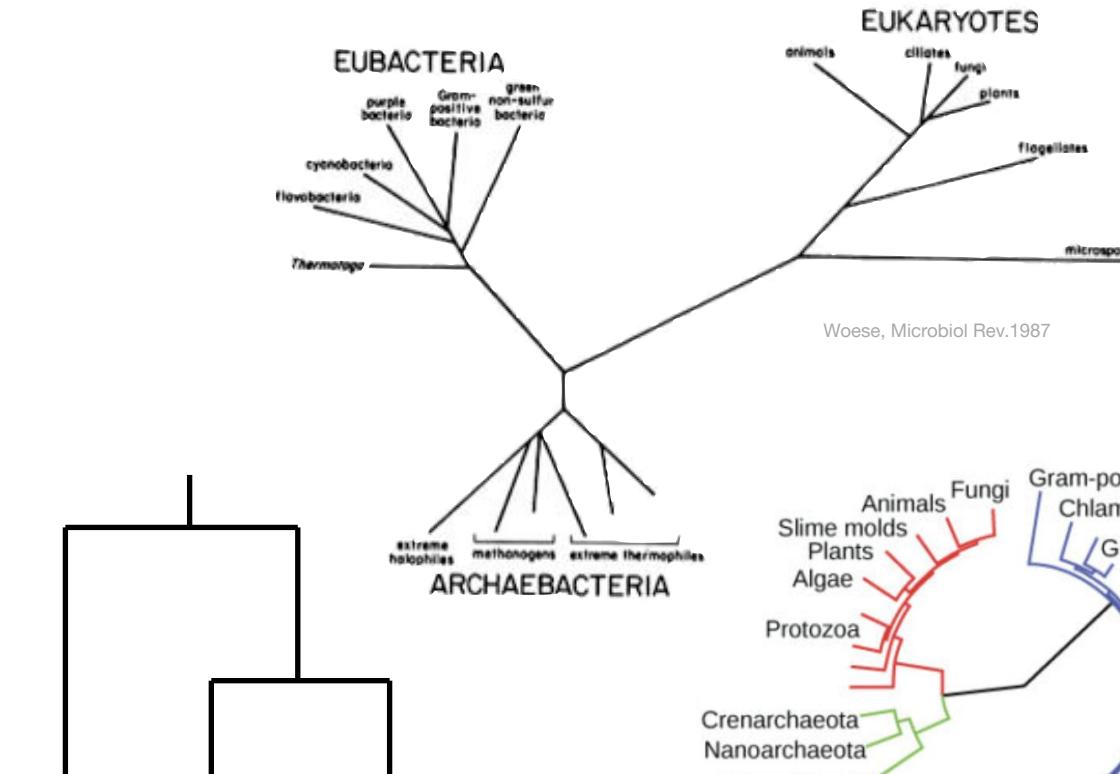
Concepts:

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

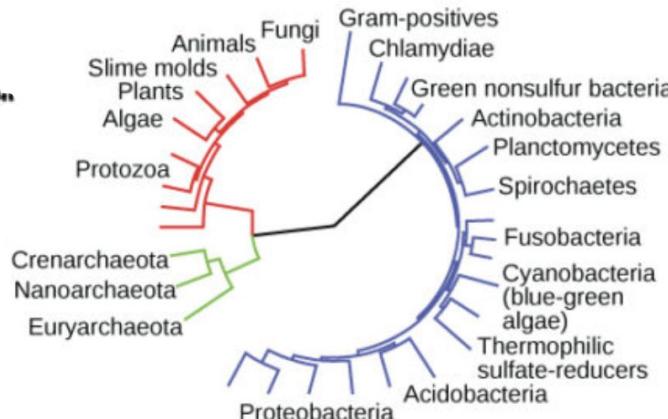


phylo.io

Other tree representations

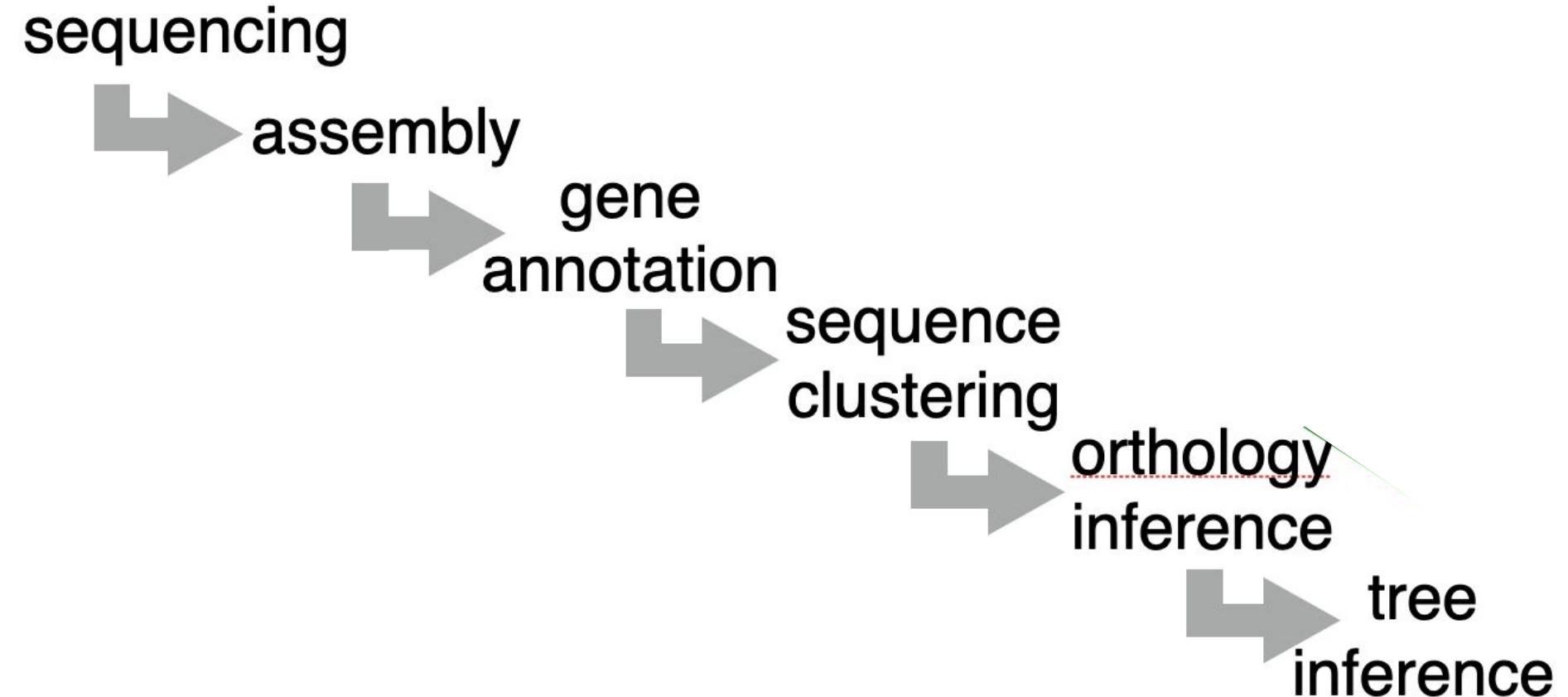


Gorilla Chimp Human



Concepts:

- Rooted vs. unrooted
- Scaled vs. unscaled



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

Vertebrata / Gnathostomata / Euteleostomi / Sarcopterygii / Tetrapoda / Amniota / Mammalia / Lower Level ▾



Graphical viewer

Members

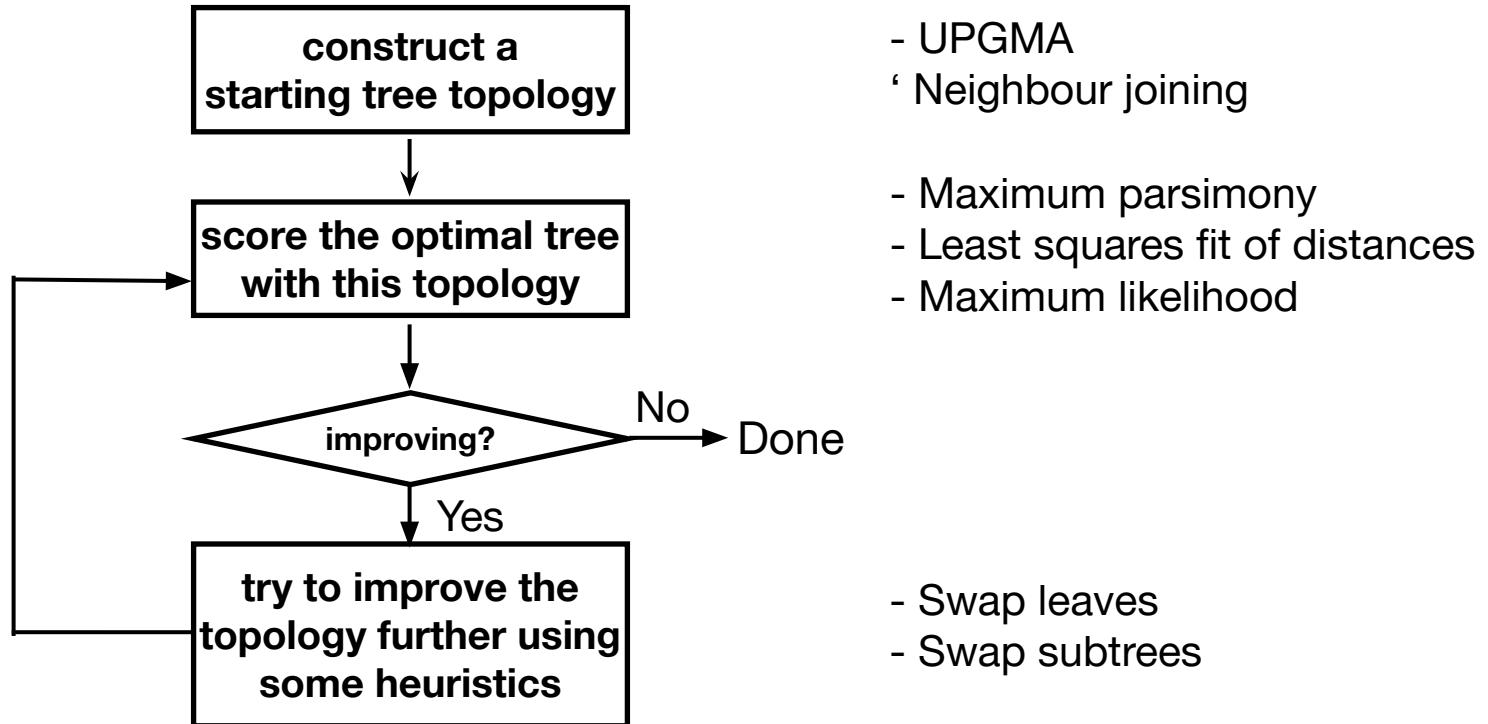
Alignment

Ancestral synteny

Similar HOGs

Label	2 . 4 . 6 . 8 . 10 . 12 . 14 . 16 . 18 . 20 . 22 . 24 . 26 . 28 . 30 . 32 . 34 . 36 . 38 . 40 . 42 . 44 . 46 . 48 . 50 . 52 . 54 . 56 . 58 . 60
CHIGR1/53/	- - - - - M A L W M K L L P L L A L L A L W E P N P A Q A F Y N Q H L C G S H L V E A L Y L V C G E R G F F Y T
MOUSE26691	- - - - - M A L L V H F L P L L A L L A L W E P K P T Q A F V K Q H L C G P H L V E A L Y L V C G E R G F F Y T
MOUSE57227	- - - - - M A L W M R F L P L L A L L A L F L W E S H P T Q A F V K Q H L C G S H L V E A L Y L V C G E R G F F Y T
RATNO17730	M P S C G H C S N M A L W I R F L P L L A L L I L W E P R P A Q A F V K Q H L C G S H L V E A L Y L V C G E R G F F Y T
RATNO18008	- - - - - M A L W M R F L P L L A L L A L V L W E P K P A Q A F V K Q H L C G P H L V E A L Y L V C G E R G F F Y T
NANGA19616	- - - - - M A L W M R L L P L L A L L A F W G P N P G Q A F V N Q H L C G S H L V E A L Y L V C G E R G F F Y T
ICTTR07348	- - - - - M A L W T R L L P L L A L L A L L G P D P A Q A F V N Q H L C G S H L V E A L Y L V C G E R G F F Y T
CERAT20372	- - - - - M A L W M R L L P L L A L L A L W G P D P V P A F V N Q H L C G S H L V E A L Y L V C G E R G F F Y T
CHLSB00063	- - - - - M A L W M R L L P L L A L L A L W G P D P V P A F V N Q H L C G S H L V E A L Y L V C G E R G F F Y T
MACFA08228	- - - - - M A L W M R L L P L L A L L A L W G P D P A P A F V N Q H L C G S H L V E A L Y L V C G E R G F F Y T
MACMU09924	- - - - - M A L W M R L L P L L A L L A L W G P D P A P A F V N Q H L C G S H L V E A L Y L V C G E R G F F Y T
MACNE29303	- - - - - M A L W M R L L P L L A L L A L W G P D P A P A F V N Q H L C G S H L V E A L Y L V C G E R G F F Y T
MANLE26401	- - - - - M A L W M R L L P L L A L L A L W G P D P V P A F V N Q H L C G S H L V E A L Y L V C G E R G F F Y T
PAPAN06025	- - - - - M A L W M R L L P L L A L L A L W G P D P V P A F V N Q H L C G S H L V E A L Y L V C G E R G F F Y T
COLAP22086	- - - - - M A L W M R L L P L L A L L A L W G P D P V P A F V N Q H L C G S H L V E A L Y L V C G E R G F F Y T
RHIBE04930	- - - - - M A L W M R L L P L L A L L A L W G P D P V P A F V N Q H L C G S H L V E A L Y L V C G E R G F F Y T
RHIRO07577	- - - - - M A L W M R L L P L L A L L A L C G P D P V P A F V N Q H L C G S H L V E A L Y L V C G E R G F F Y T
GORG003436	- - - - - M A L W M R L L P L L A L L A L W G P D P A A A F V N Q H L C G S H L V E A L Y L V C G E R G F F Y T
HUMAN03911	- - - - - M A L W M R L L P L L A L L A L W G P D P A A A F V N Q H L C G S H L V E A L Y L V C G E R G F F Y T
PANPA05915	- - - - - M A L W M R L L P L L A L L A L W G P D P A S A F V N Q H L C G S H L V E A L Y L V C G E R G F F Y T
PANTR02017	- - - - - M A L W M R L L P L L V L L A L W G P D P A S A F V N Q H L C G S H L V E A L Y L V C G E R G F F Y T
PONAB02480	- - - - - M A L W M R L L P L L A L L A L W G P D P A - A F V N Q H L C G S H L V E A L Y L V C G E R G F F Y T

General Scheme to Build Phylogenetic Trees

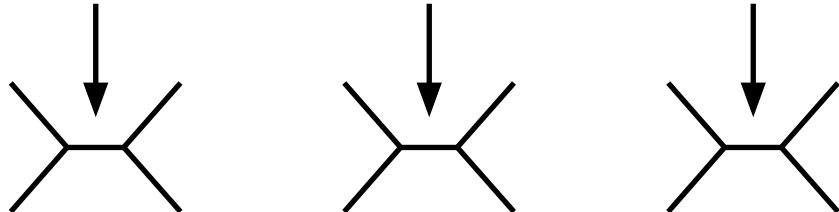
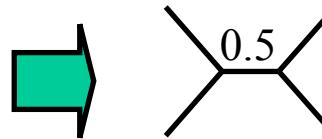


Measuring confidence with the Bootstrap

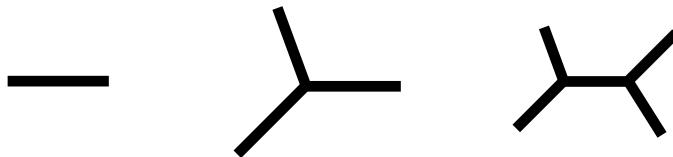
How would our tree inference change if we repeated the experiment with other data that came about through the same process?

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

Data	Replicate 1	Replicate 2
A: ALTF ^C G	A: LCGC ^A L	A: ATF ^A L ^F
B: NL ^T FCG	B: LCGC ^N L	B: NT ^F NLF
C: ALSFRG	C: LRGRAL	C: ASF ^A L ^F
D: NLSFRG	D: LRGRNL	D: NSF ^N LF



**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”**

3

**Number of binary trees
Unrooted Rooted**

1