OMA and **OMArk** for homology exploration and gene annotation quality control

UNIL | Université de Lausanne

comparative genomics LAB

SIB

Learning objectives

- ❖ Where to easily find orthology information for well-studied species?

  Query the OMA Browser and understanding HOGs

- ❖ Where to get quick homology estimate for my newly sequenced species?

  Run OMAmer for sequence placement into HOGs and interpret results

- ❖ How to know if a proteome is of good quality ?

  Run OMArk for proteome quality assessment and interpret results

Session plan

1. Hierarchical Orthologous Groups and the OMA Browser

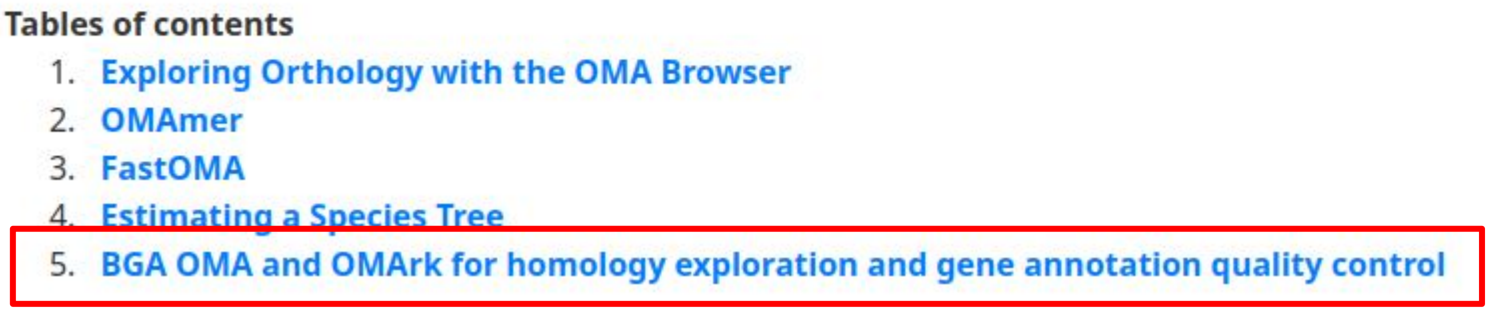
2. Fast sequence placement with OMAmer

3. Gene repertoire quality assessment with OMArk

OMA Academy website

https://omabrowser.org/oma/academy/

https://tinyurl.com/OMABGA24



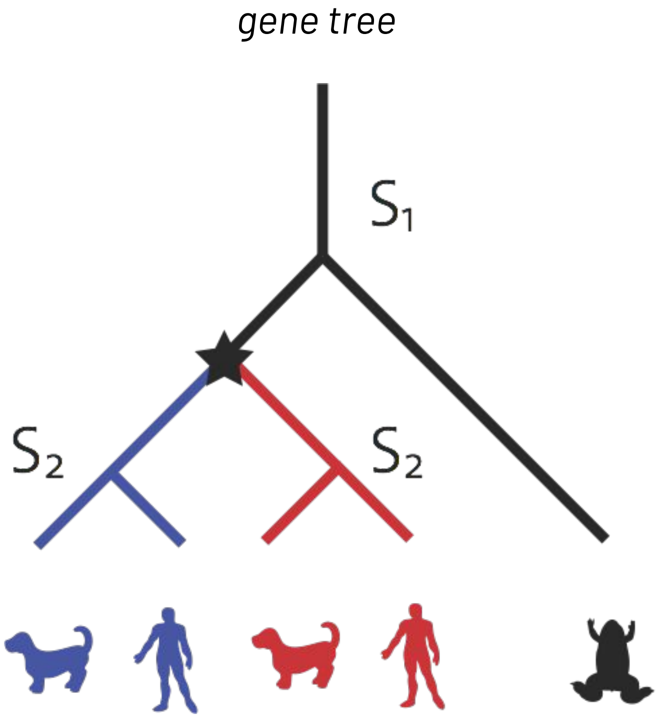
Tables of contents

1. Exploring Orthology with the OMA Browser
2. OMAmer
3. FastOMA
4. Estimating a Species Tree
5. BGA OMA and OMArk for homology exploration and gene annotation quality control

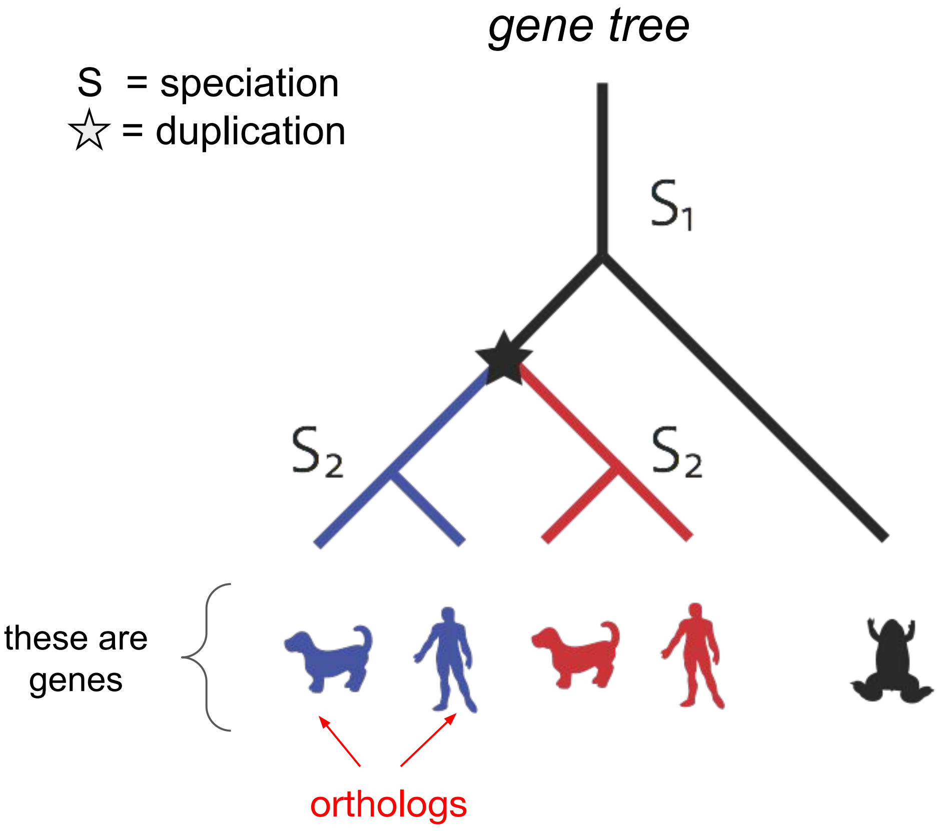
Orthology &
Hierarchical Orthologous Groups (HOGs)

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.



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

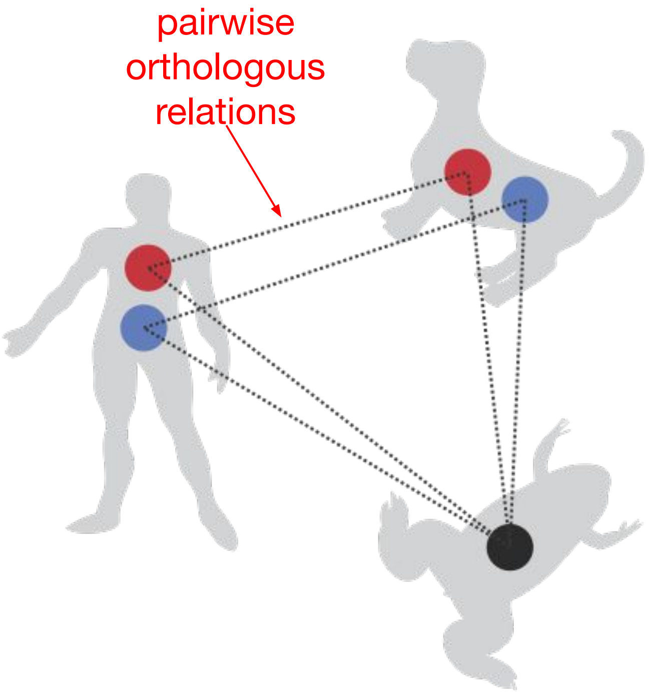


gene tree

S = speciation
☆ = duplication

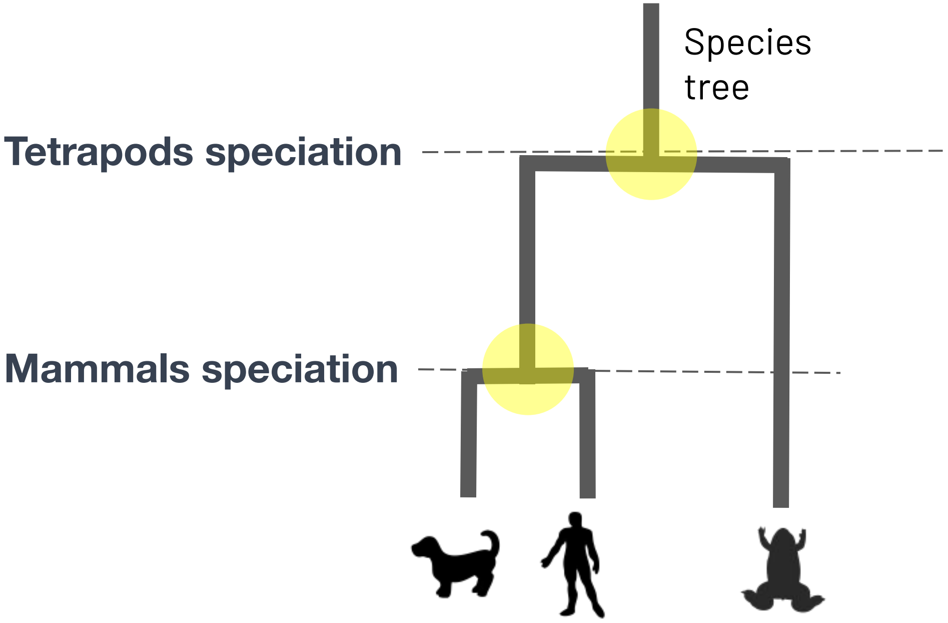
these are genes

orthologs

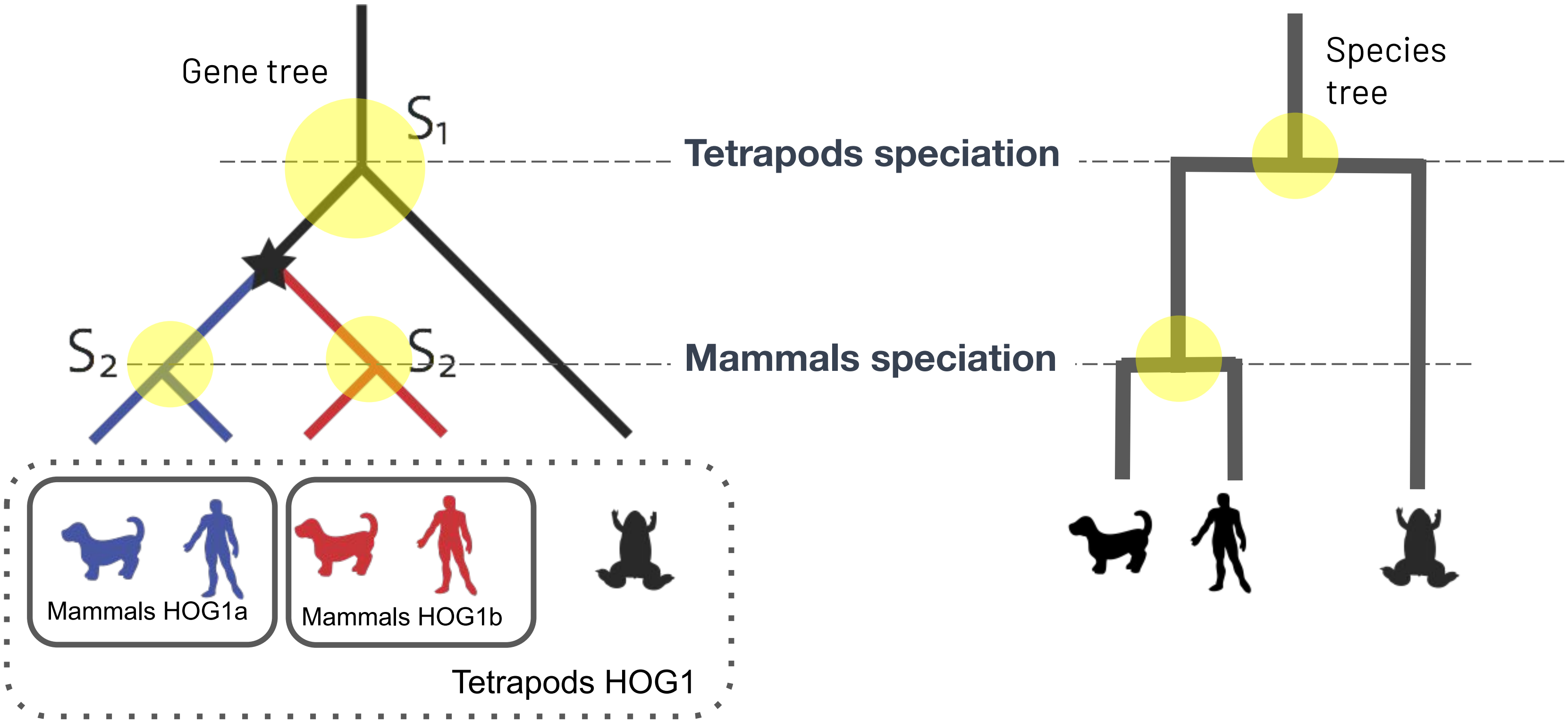


pairwise orthologous relations

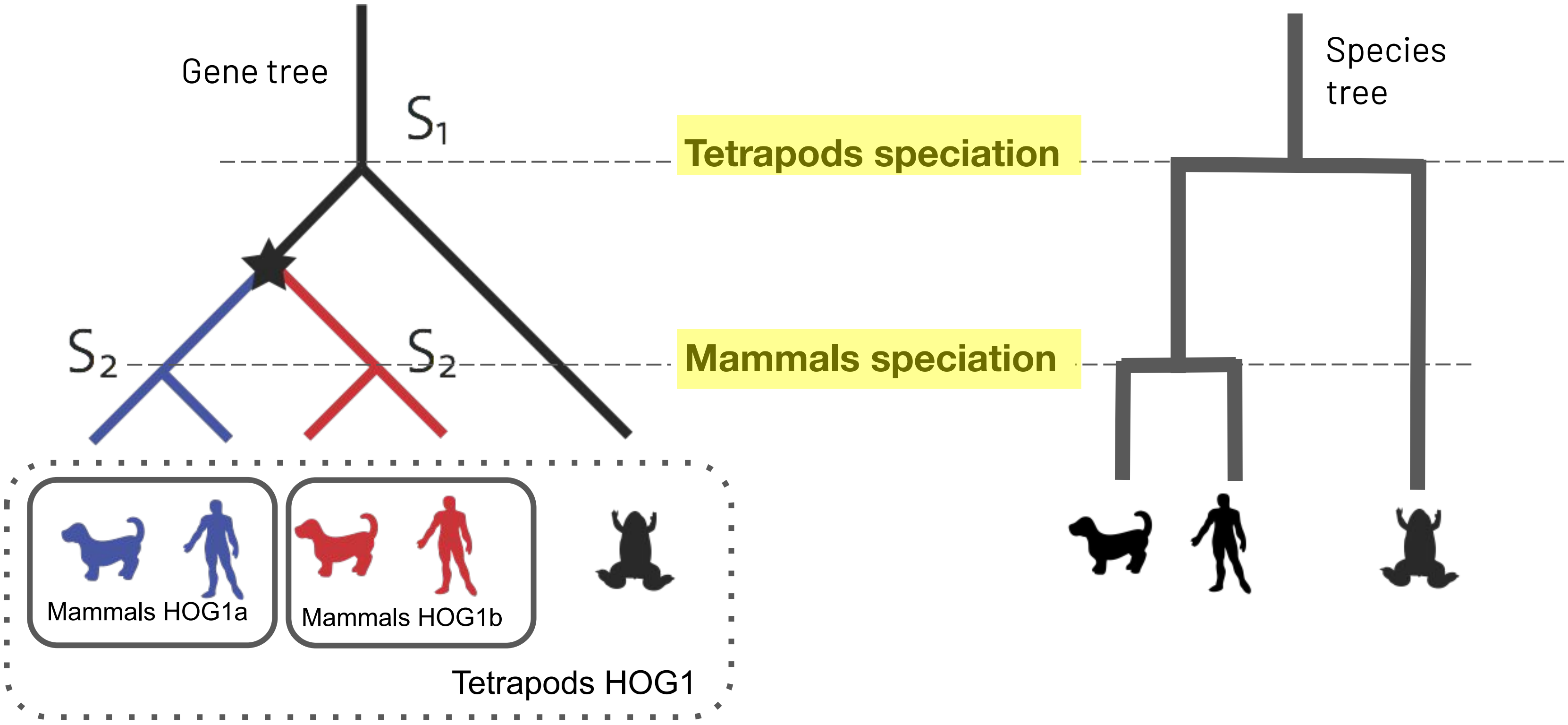
HOGs = Sets of genes that descended from a common ancestral gene in a given ancestral species



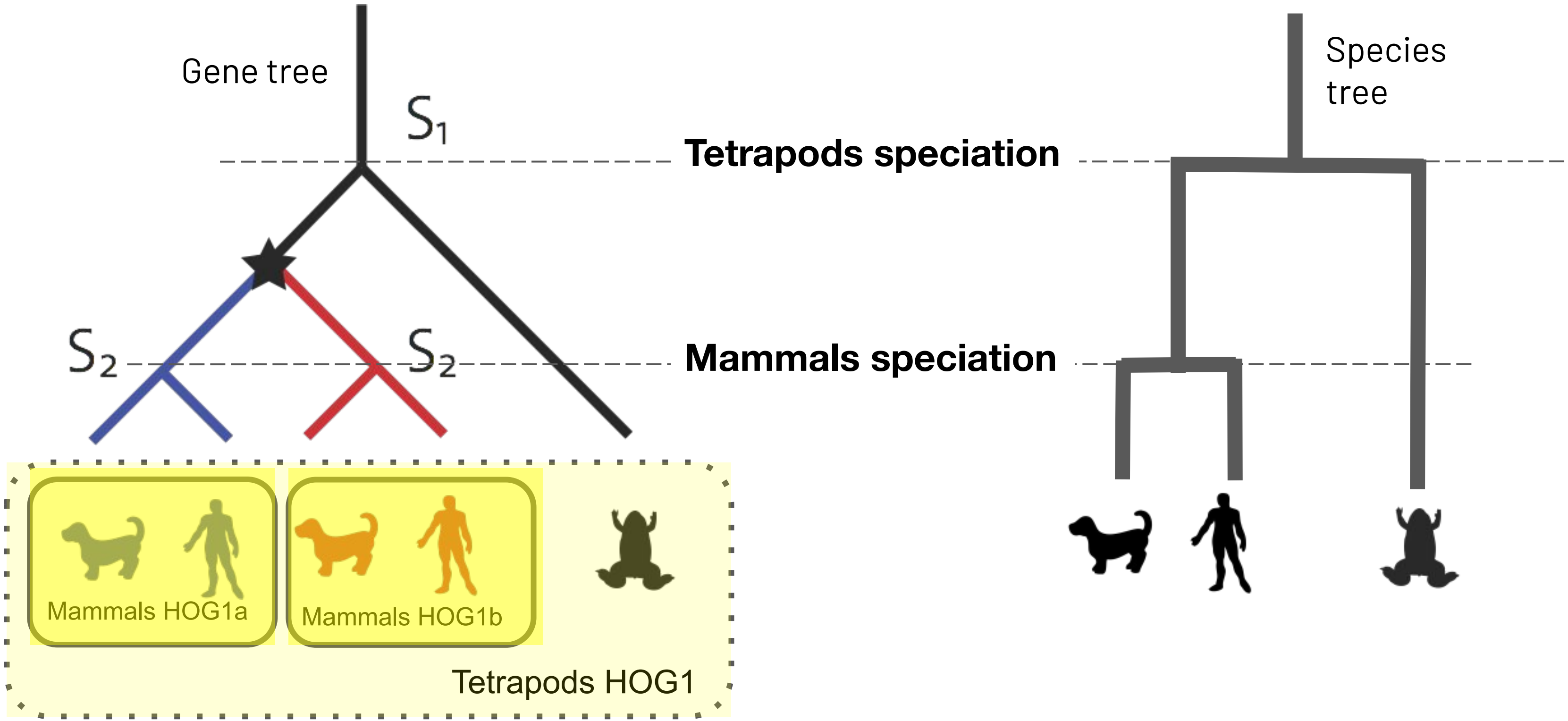
HOGs = Sets of genes that descended from a common ancestral gene in a given ancestral species



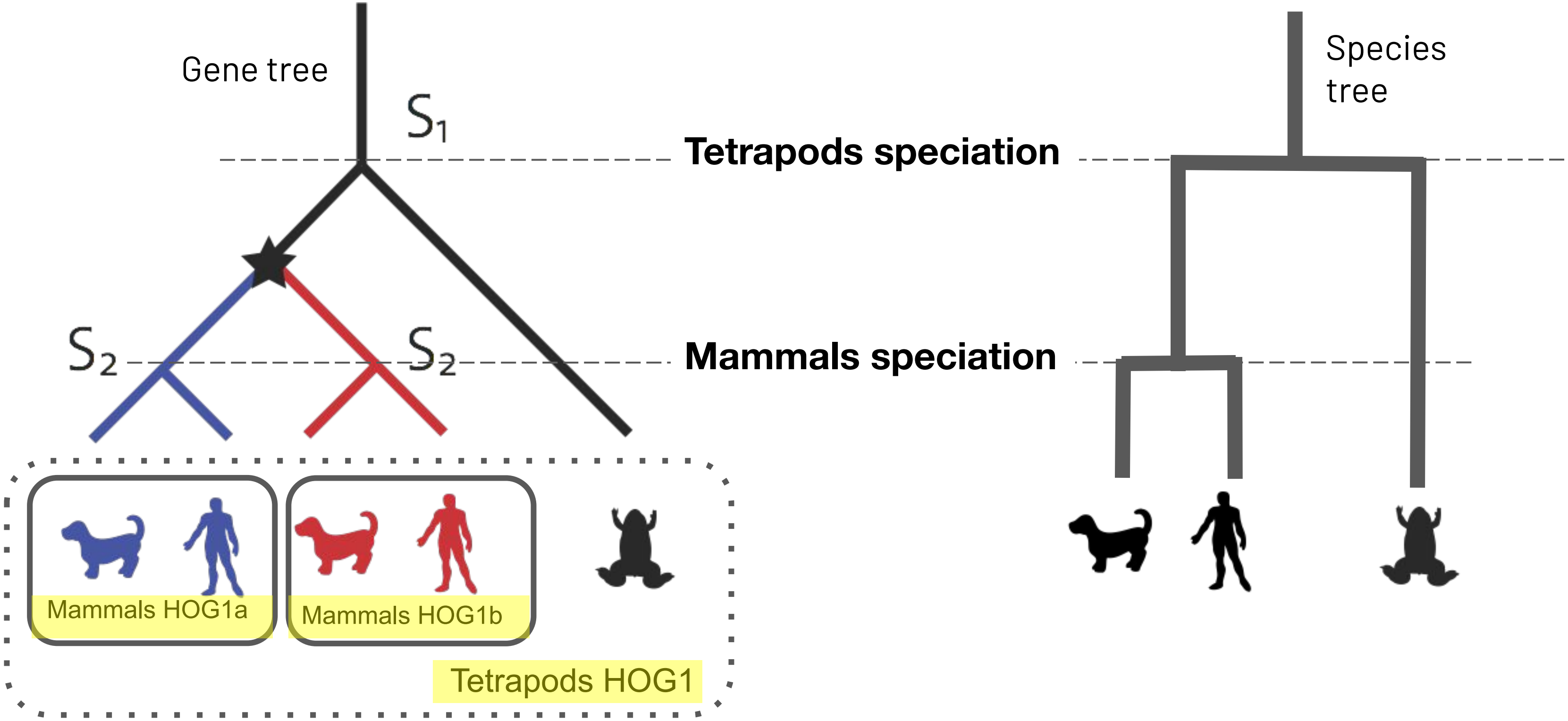
HOGs are defined with respect to specific clades



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

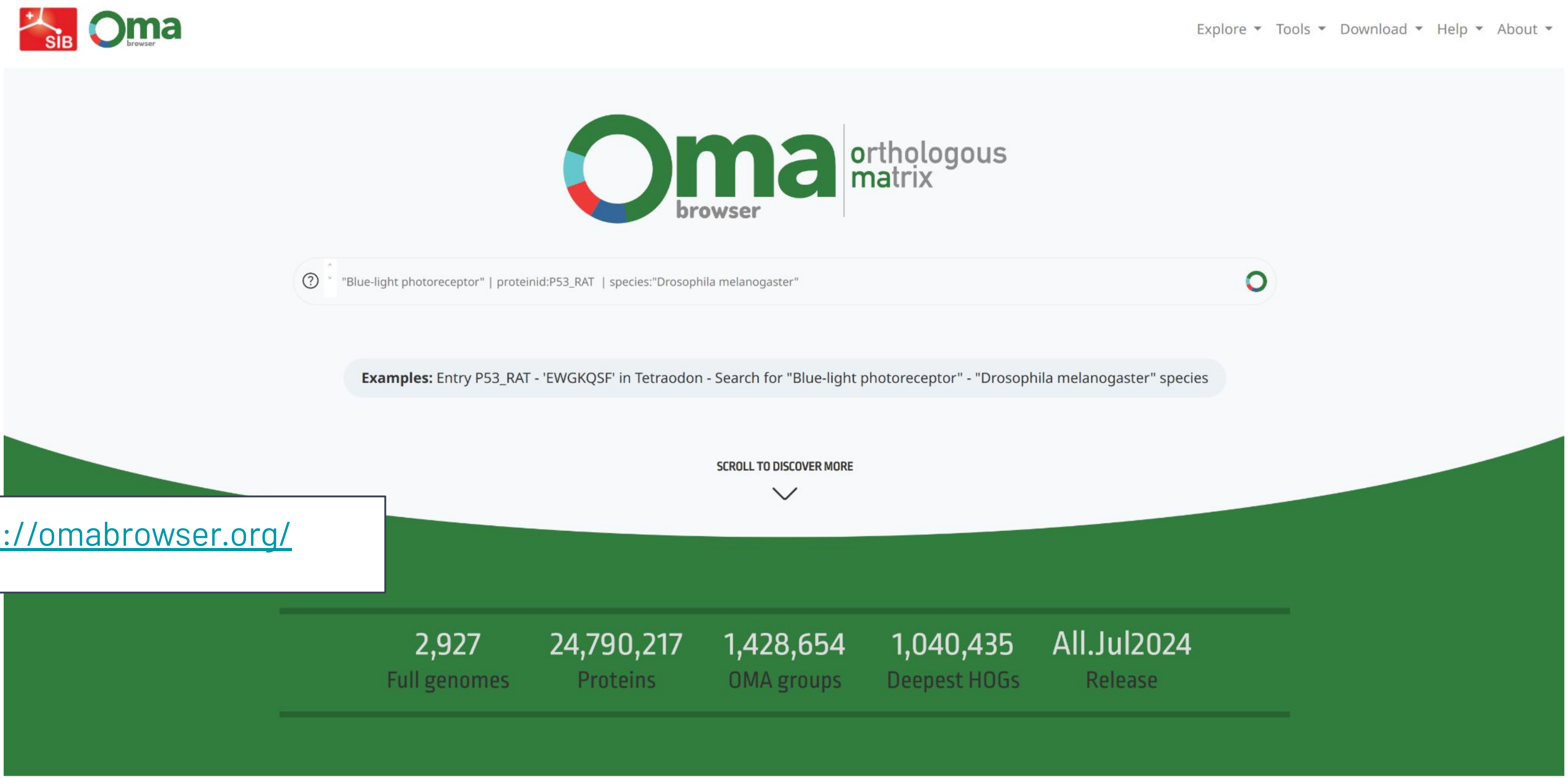


HOGs are gene families; SubHOGs are nested subfamilies



The OMA browser

The OMA browser



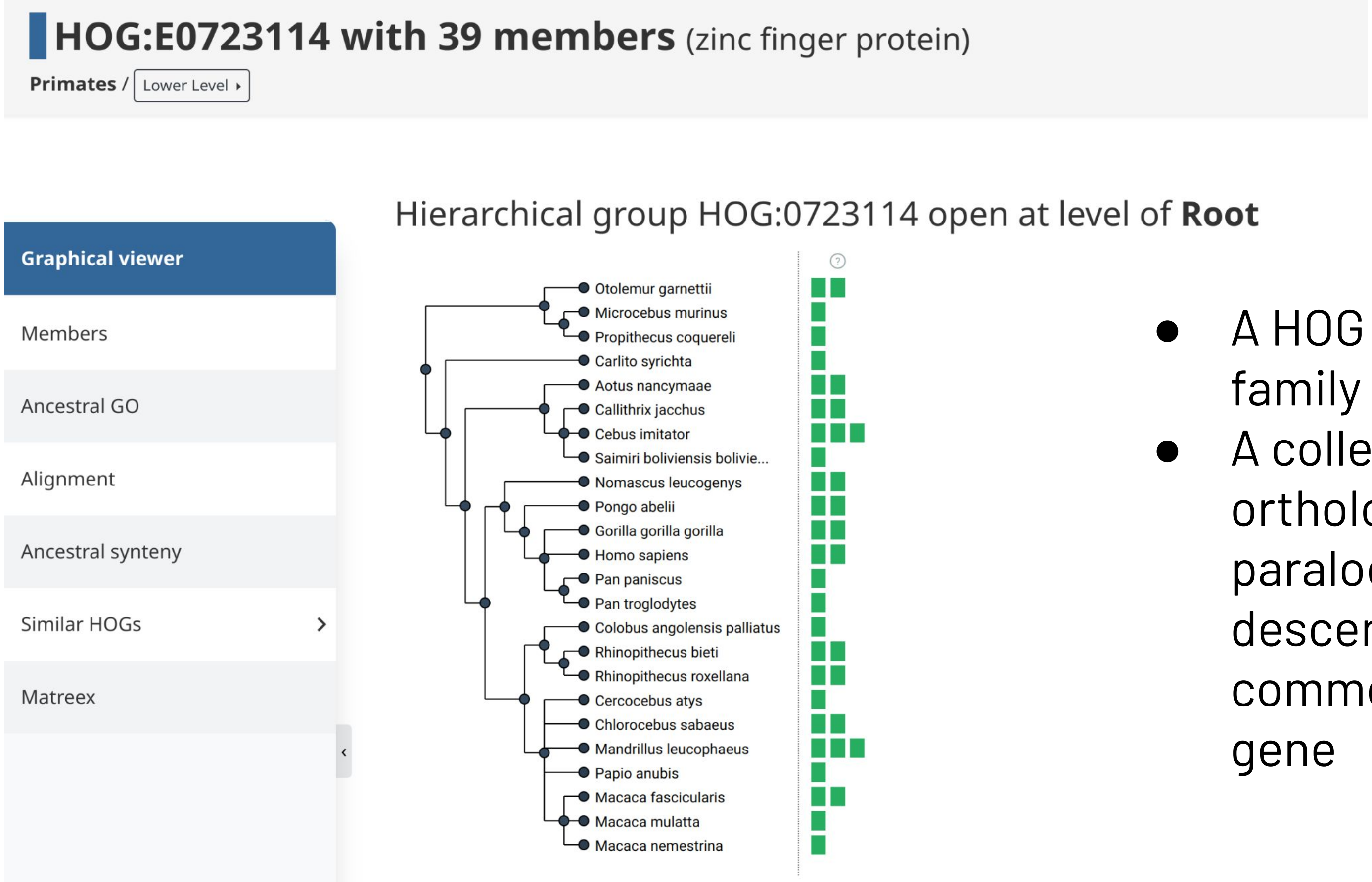
Oma browser
Explore
Tools
Download
Help
About
Oma browser orthologous matrix
"Blue-light photoreceptor" | proteinid:P53_RAT | species:"Drosophila melanogaster"
SCROLL TO DISCOVER MORE

Examples: Entry P53_RAT - 'EWGKQSF' in Tetraodon - Search for "Blue-light photoreceptor" - "Drosophila melanogaster" species

2,927	24,790,217	1,428,654	1,040,435	All.Jul2024
Full genomes	Proteins	OMA groups	Deepest HOGs	Release

https://omabrowser.org/

Hierarchical Orthologous Groups (HOGs)



HOG:E0723114 with 39 members (zinc finger protein)

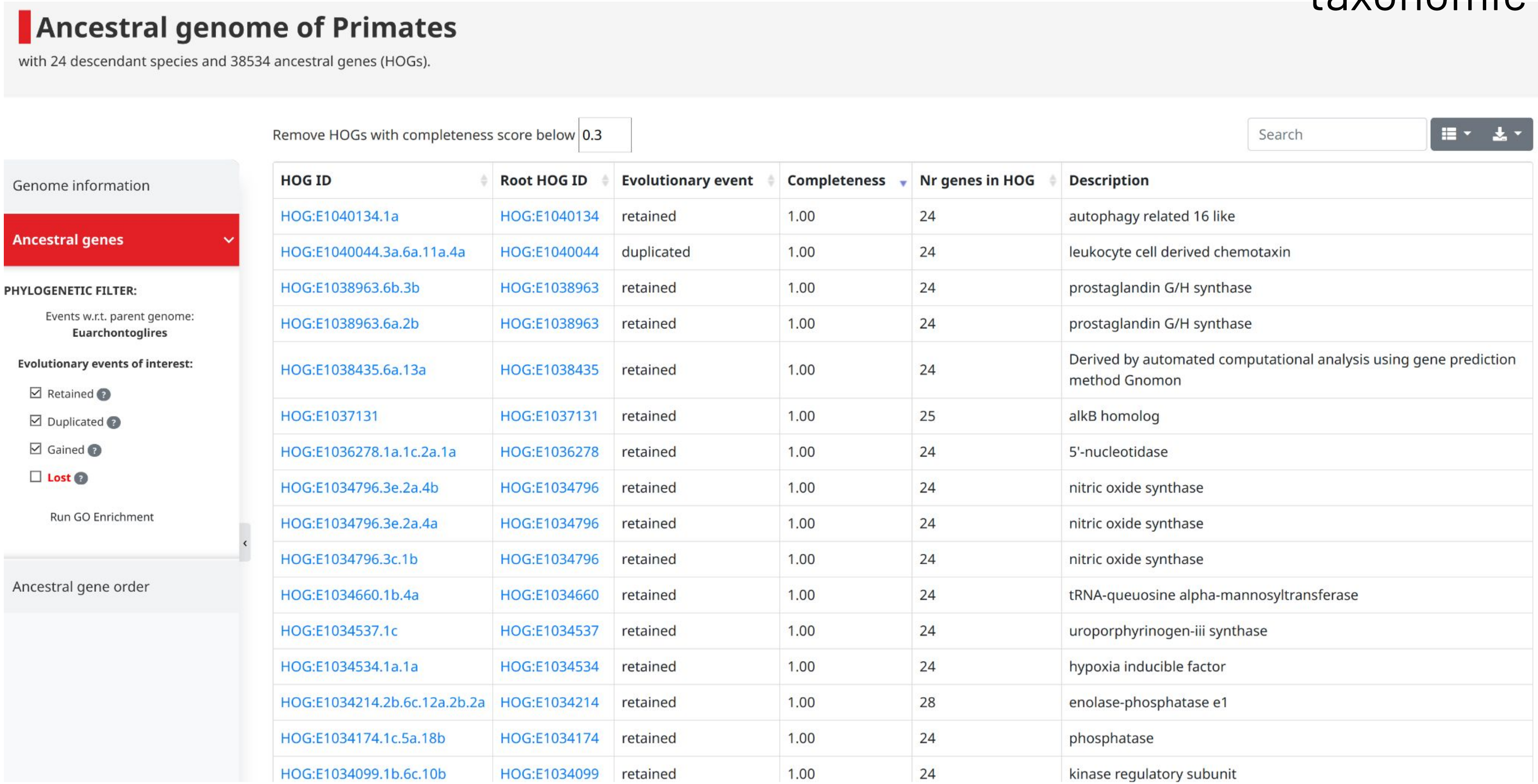
Primates / Lower Level ▸

Hierarchical group HOG:0723114 open at level of **Root**

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

Ancestral genomes

The collection of HOGs at a given taxonomic level



Ancestral genome of Primates

with 24 descendant species and 38534 ancestral genes (HOGs).

Genome information

Ancestral genes

PHYLOGENETIC FILTER:

Events w.r.t. parent genome:
Euarchontoglires

Evolutionary events of interest:

☑ Retained
☑ Duplicated
☑ Gained
□ **Lost**

Run GO Enrichment

Ancestral gene order

Remove HOGs with completeness score below 0.3

Search

HOG ID	Root HOG ID	Evolutionary event	Completeness	Nr genes in HOG	Description
HOG:E1040134.1a	HOG:E1040134	retained	1.00	24	autophagy related 16 like
HOG:E1040044.3a.6a.11a.4a	HOG:E1040044	duplicated	1.00	24	leukocyte cell derived chemotaxin
HOG:E1038963.6b.3b	HOG:E1038963	retained	1.00	24	prostaglandin G/H synthase
HOG:E1038963.6a.2b	HOG:E1038963	retained	1.00	24	prostaglandin G/H synthase
HOG:E1038435.6a.13a	HOG:E1038435	retained	1.00	24	Derived by automated computational analysis using gene prediction method Gnomon
HOG:E1037131	HOG:E1037131	retained	1.00	25	alkB homolog
HOG:E1036278.1a.1c.2a.1a	HOG:E1036278	retained	1.00	24	5'-nucleotidase
HOG:E1034796.3e.2a.4b	HOG:E1034796	retained	1.00	24	nitric oxide synthase
HOG:E1034796.3e.2a.4a	HOG:E1034796	retained	1.00	24	nitric oxide synthase
HOG:E1034796.3c.1b	HOG:E1034796	retained	1.00	24	nitric oxide synthase
HOG:E1034660.1b.4a	HOG:E1034660	retained	1.00	24	tRNA-queuosine alpha-mannosyltransferase
HOG:E1034537.1c	HOG:E1034537	retained	1.00	24	uroporphyrinogen-iii synthase
HOG:E1034534.1a.1a	HOG:E1034534	retained	1.00	24	hypoxia inducible factor
HOG:E1034214.2b.6c.12a.2b.2a	HOG:E1034214	retained	1.00	28	enolase-phosphatase e1
HOG:E1034174.1c.5a.18b	HOG:E1034174	retained	1.00	24	phosphatase
HOG:E1034099.1b.6c.10b	HOG:E1034099	retained	1.00	24	kinase regulatory subunit

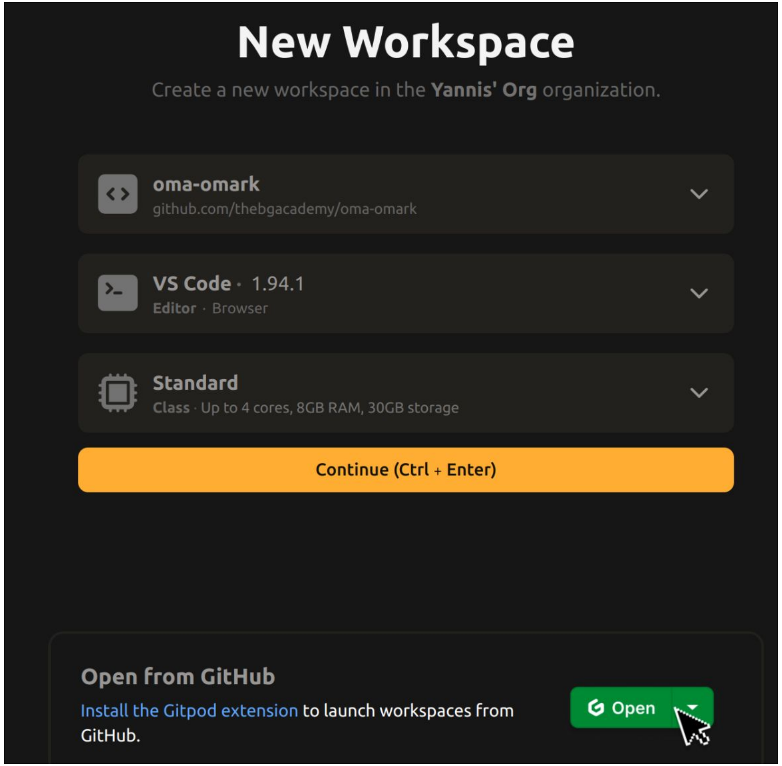
Hand-on exercices

https://omabrowser.org/oma/academy/

https://tinyurl.com/OMABGA24

https://omabrowser.org/

Gitpod



New Workspace

Create a new workspace in the **Yannis' Org** organization.

oma-omark
github.com/thebgacademy/oma-omark

VS Code · 1.94.1
Editor · Browser

Standard
Class · Up to 4 cores, 8GB RAM, 30GB storage

Continue (Ctrl + Enter)

Open from GitHub

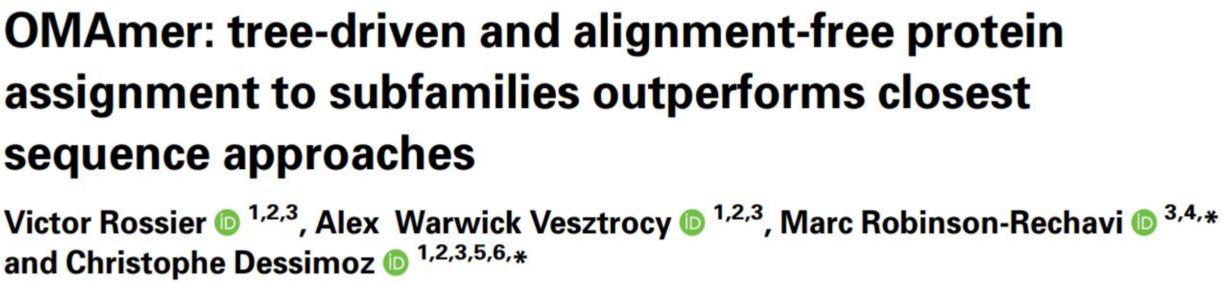
Install the Gitpod extension to launch workspaces from GitHub.

Fast sequence placements with OMAmer

What is OMAmer?

- Fast sequence placement into existing HOGs from the OMA Browser

- More accurate than closest sequence matching for subfamily placement!

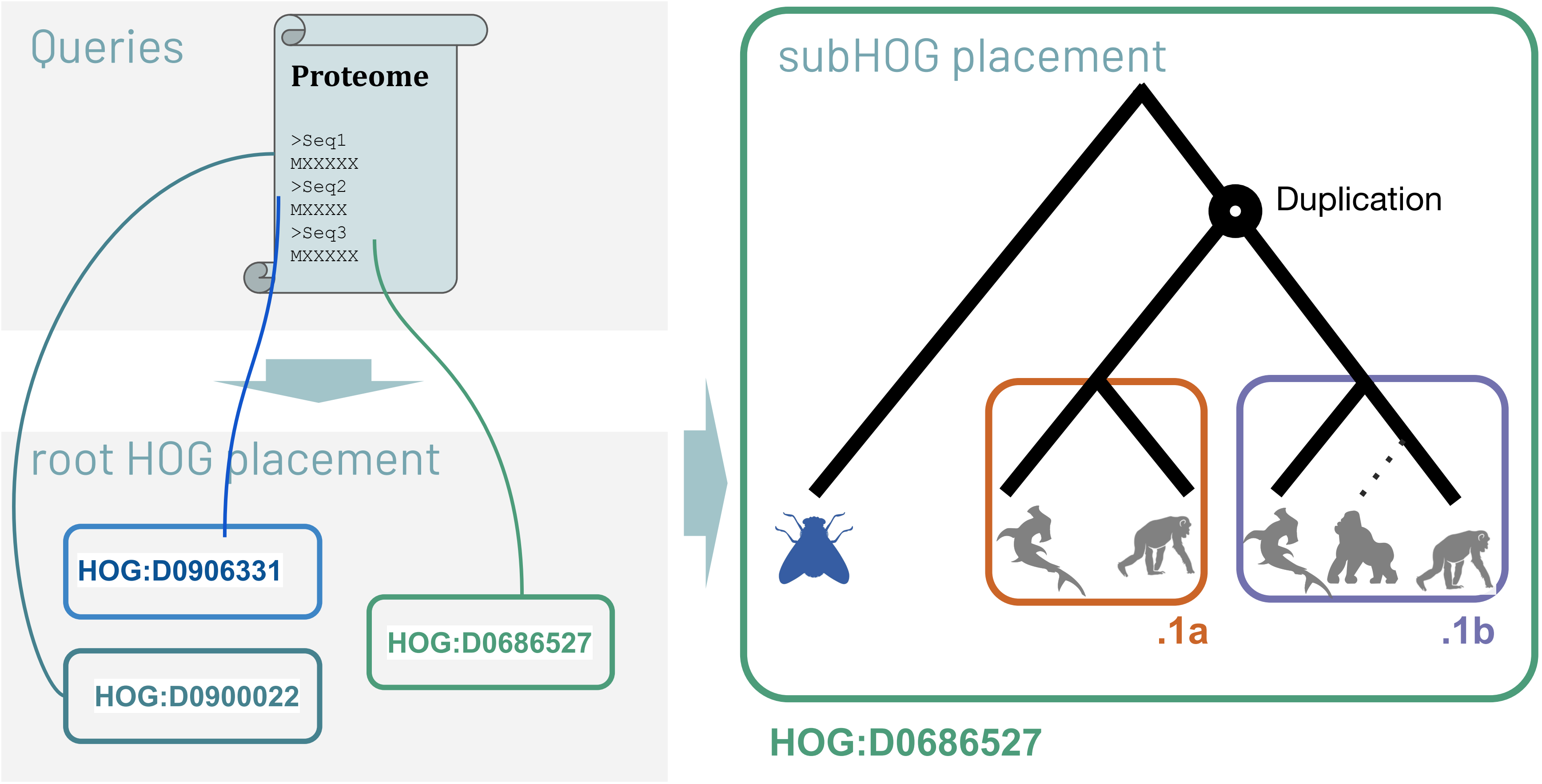


OMAmer: tree-driven and alignment-free protein assignment to subfamilies outperforms closest sequence approaches

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



Queries

Proteome

```
>Seq1
MXXXXX
>Seq2
MXXXX
>Seq3
MXXXXX
```

root HOG placement

HOG:D0906331

HOG:D0900022

HOG:D0686527

subHOG placement

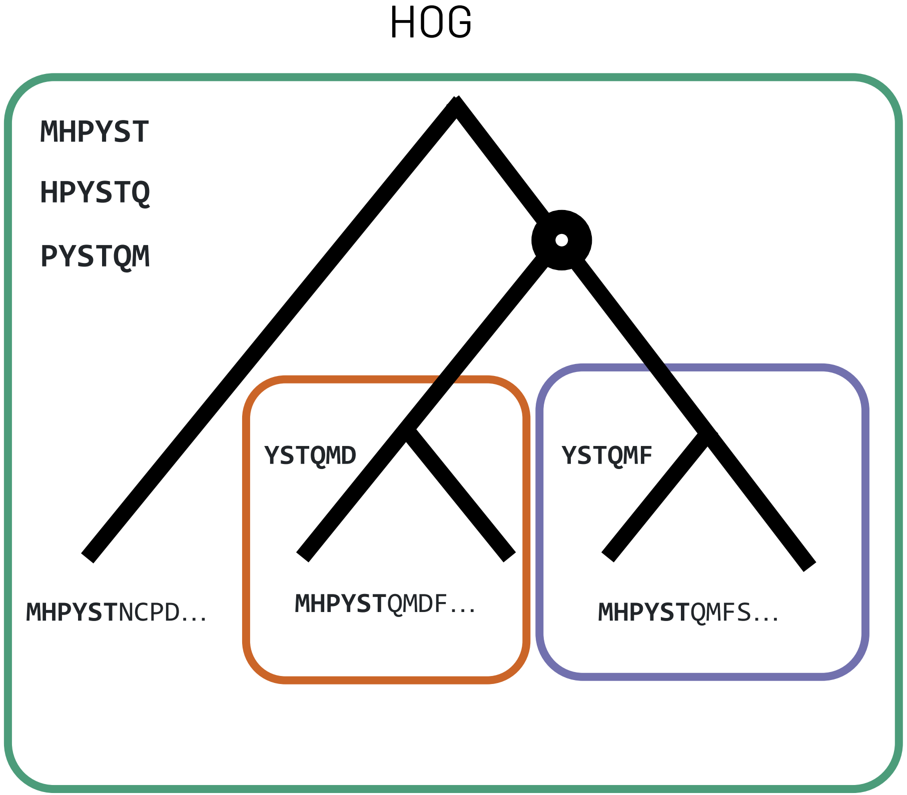
HOG:D0686527

k-mer based placement

❖ **k-mers** : words of k characters in a sequences

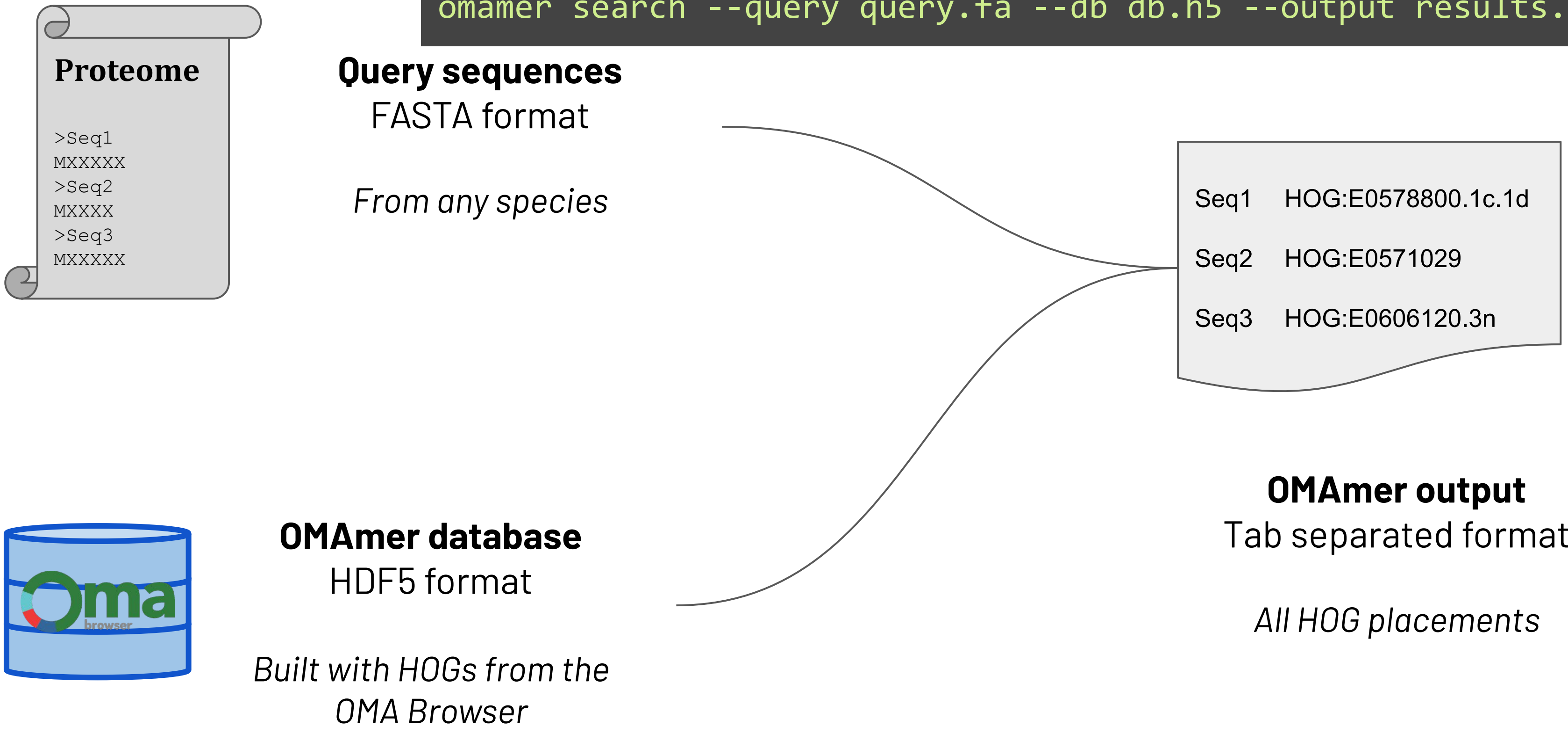
Query sequence

```
MHPYSTQMFS LQITVMEDSQ SDMSIELPLS

MHPYST
 HPYSTQ
  PYSTQM

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



How to use OMAmer

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



Proteome

```
>Seq1
MXXXXX
>Seq2
MXXXX
>Seq3
MXXXXX
```

Query sequences
FASTA format

From any species

OMAmer database
HDF5 format

Built with HOGs from the OMA Browser

```
Seq1    HOG:E0578800.1c.1d
Seq2    HOG:E0571029
Seq3    HOG:E0606120.3n
```

OMAmer output
Tab separated format

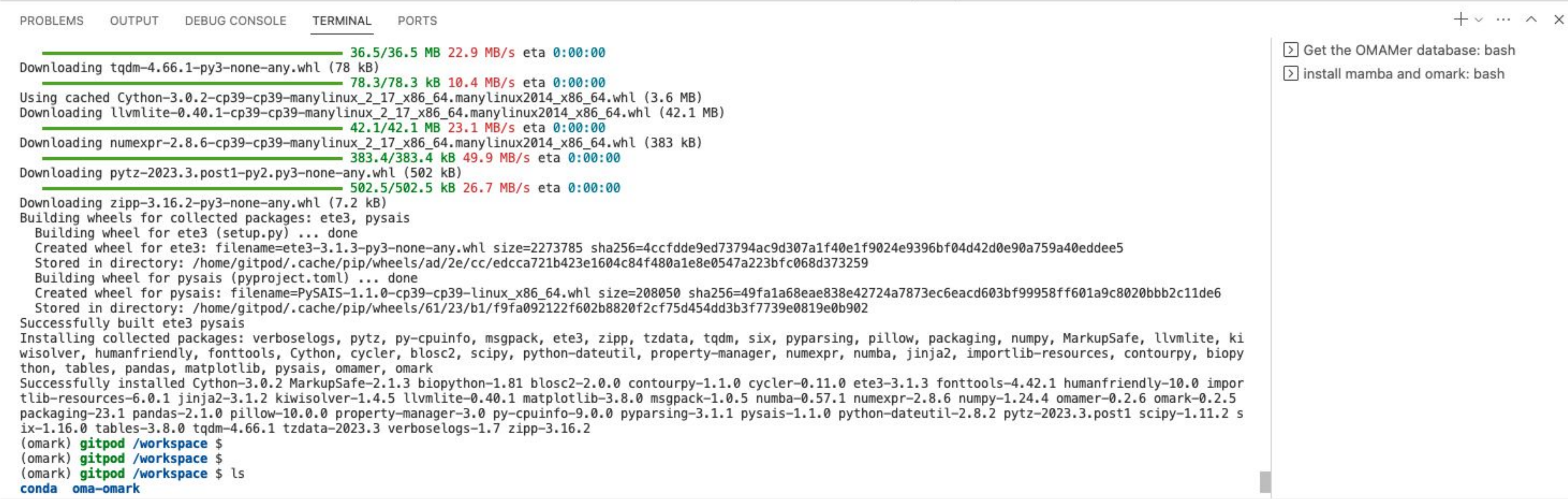
All HOG placements

Hand-on exercices

Gitpod

https://omabrowser.org/oma/academy/

https://tinyurl.com/OMABGA24



```
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 36.5/36.5 MB 22.9 MB/s eta 0:00:00
Downloading tqdm-4.66.1-py3-none-any.whl (78 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 78.3/78.3 kB 10.4 MB/s eta 0:00:00
Using cached Cython-3.0.2-cp39-cp39-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (3.6 MB)
Downloading llvmlite-0.40.1-cp39-cp39-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (42.1 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 42.1/42.1 MB 23.1 MB/s eta 0:00:00
Downloading numexpr-2.8.6-cp39-cp39-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (383 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 383.4/383.4 kB 49.9 MB/s eta 0:00:00
Downloading pytz-2023.3.post1-py2.py3-none-any.whl (502 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 502.5/502.5 kB 26.7 MB/s eta 0:00:00
Downloading zipp-3.16.2-py3-none-any.whl (7.2 kB)
Building wheels for collected packages: ete3, pysais
  Building wheel for ete3 (setup.py) ... done
  Created wheel for ete3: filename=ete3-3.1.3-py3-none-any.whl size=2273785 sha256=4ccfdde9ed73794ac9d307a1f40e1f9024e9396bf04d42d0e90a759a40eddee5
  Stored in directory: /home/gitpod/.cache/pip/wheels/ad/2e/cc/edcca721b423e1604c84f480a1e8e0547a223bfc068d373259
  Building wheel for pysais (pyproject.toml) ... done
  Created wheel for pysais: filename=PySAIS-1.1.0-cp39-cp39-linux_x86_64.whl size=208050 sha256=49fa1a68eae838e42724a7873ec6eacd603bf99958ff601a9c8020bbb2c11de6
  Stored in directory: /home/gitpod/.cache/pip/wheels/61/23/b1/f9fa092122f602b8820f2cf75d454dd3b3f7739e0819e0b902
Successfully built ete3 pysais
Installing collected packages: verboselogs, pytz, py-cpuinfo, msgpack, ete3, zipp, tzdata, tqdm, six, pyparsing, pillow, packaging, numpy, MarkupSafe, llvmlite, ki
wisolver, humanfriendly, fonttools, Cython, cycler, blosc2, scipy, python-dateutil, property-manager, numexpr, numba, jinja2, importlib-resources, contourpy, biopy
thon, tables, pandas, matplotlib, pysais, omamer, omark
Successfully installed Cython-3.0.2 MarkupSafe-2.1.3 biopython-1.81 blosc2-2.0.0 contourpy-1.1.0 cycler-0.11.0 ete3-3.1.3 fonttools-4.42.1 humanfriendly-10.0 impor
tlib-resources-6.0.1 jinja2-3.1.2 kiwisolver-1.4.5 llvmlite-0.40.1 matplotlib-3.8.0 msgpack-1.0.5 numba-0.57.1 numexpr-2.8.6 numpy-1.24.4 omamer-0.2.6 omark-0.2.5
packaging-23.1 pandas-2.1.0 pillow-10.0.0 property-manager-3.0 py-cpuinfo-9.0.0 pyparsing-3.1.1 pysais-1.1.0 python-dateutil-2.8.2 pytz-2023.3.post1 scipy-1.11.2 s
ix-1.16.0 tables-3.8.0 tqdm-4.66.1 tzdata-2023.3 verboselogs-1.7 zipp-3.16.2
(omark) gitpod /workspace $
(omark) gitpod /workspace $
(omark) gitpod /workspace $ ls
conda  oma-omark
```

cd oma-omark/working_dir/

Quality assessment with OMArk

How to use OMAmer

Coding-gene repertoire : set of coding-genes annotated on a given genome sequence

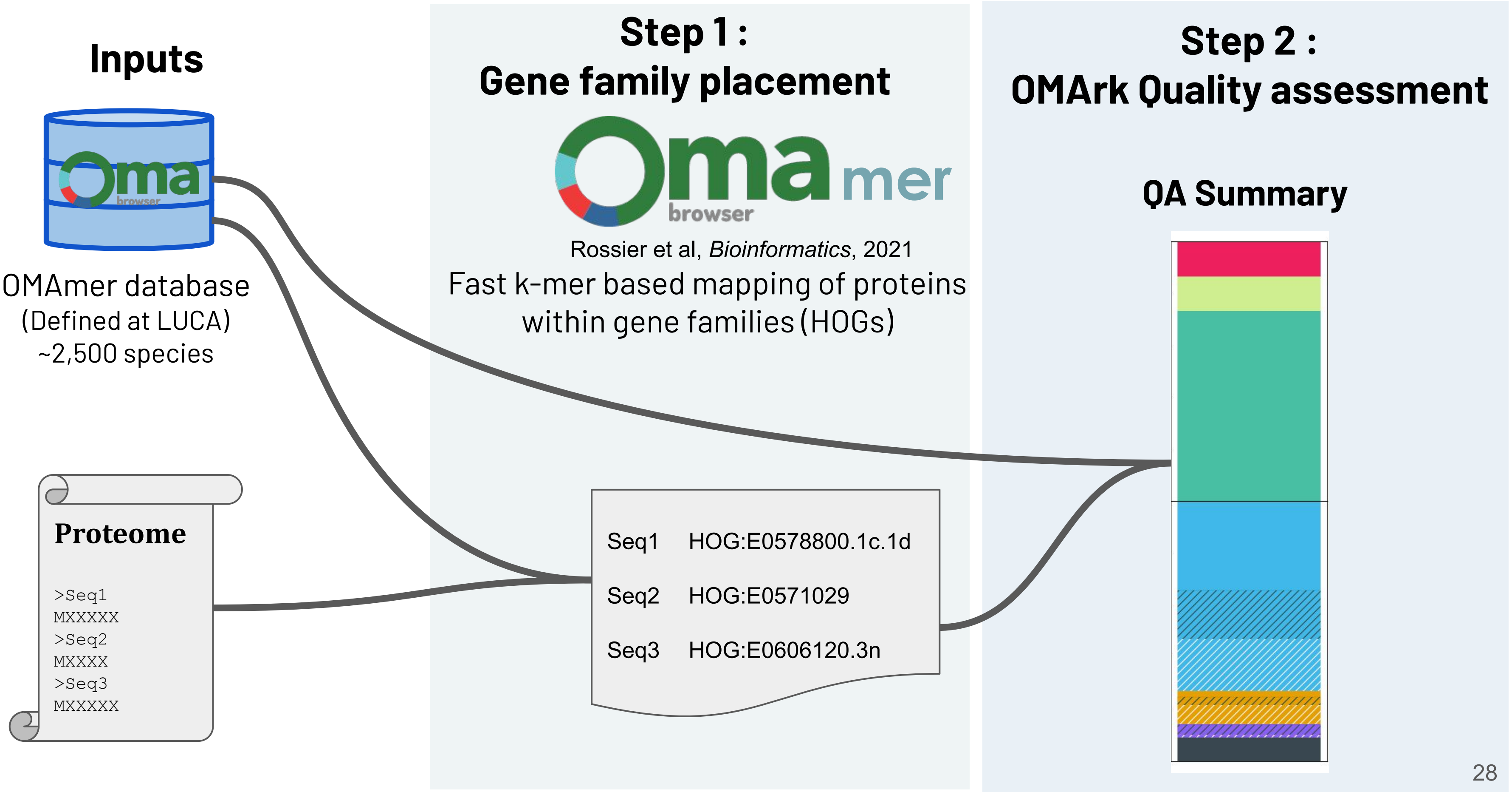
Available on database as **proteomes**

Subject to quality issues {

- Missing genes
- Fragmented genes
- Inclusion of non-coding regions
- Contamination

Lack of tool to detect all these issues !

Coding-gene repertoire quality



Inputs

OMAmer database
(Defined at LUCA)
~2,500 species

Proteome

```
>Seq1
MXXXXX
>Seq2
MXXXX
>Seq3
MXXXXX
```

Step 1 : Gene family placement

Rossier et al, *Bioinformatics*, 2021

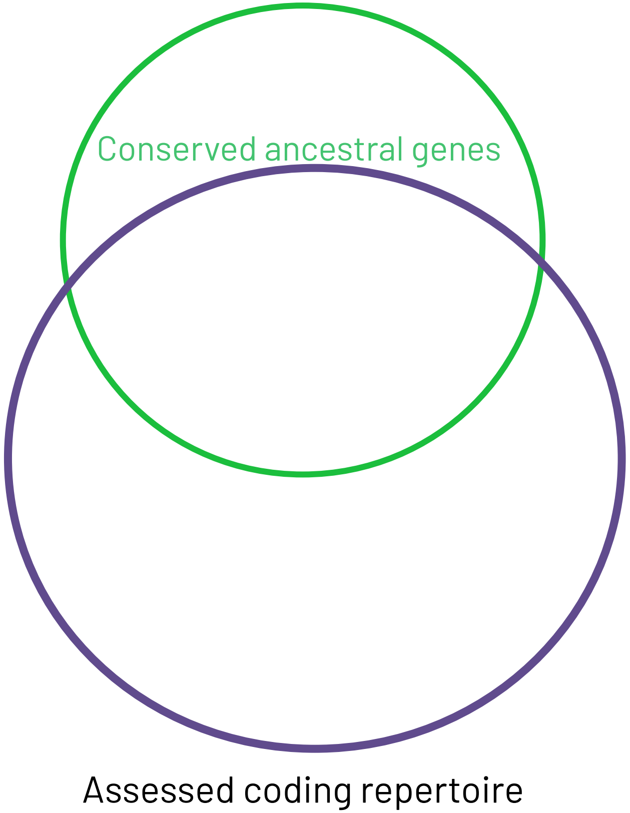
Fast k-mer based mapping of proteins within gene families (HOGs)

Seq1	HOG:E0578800.1c.1d
Seq2	HOG:E0571029
Seq3	HOG:E0606120.3n

Step 2 : OMArk Quality assessment

QA Summary

Coding-gene repertoire quality



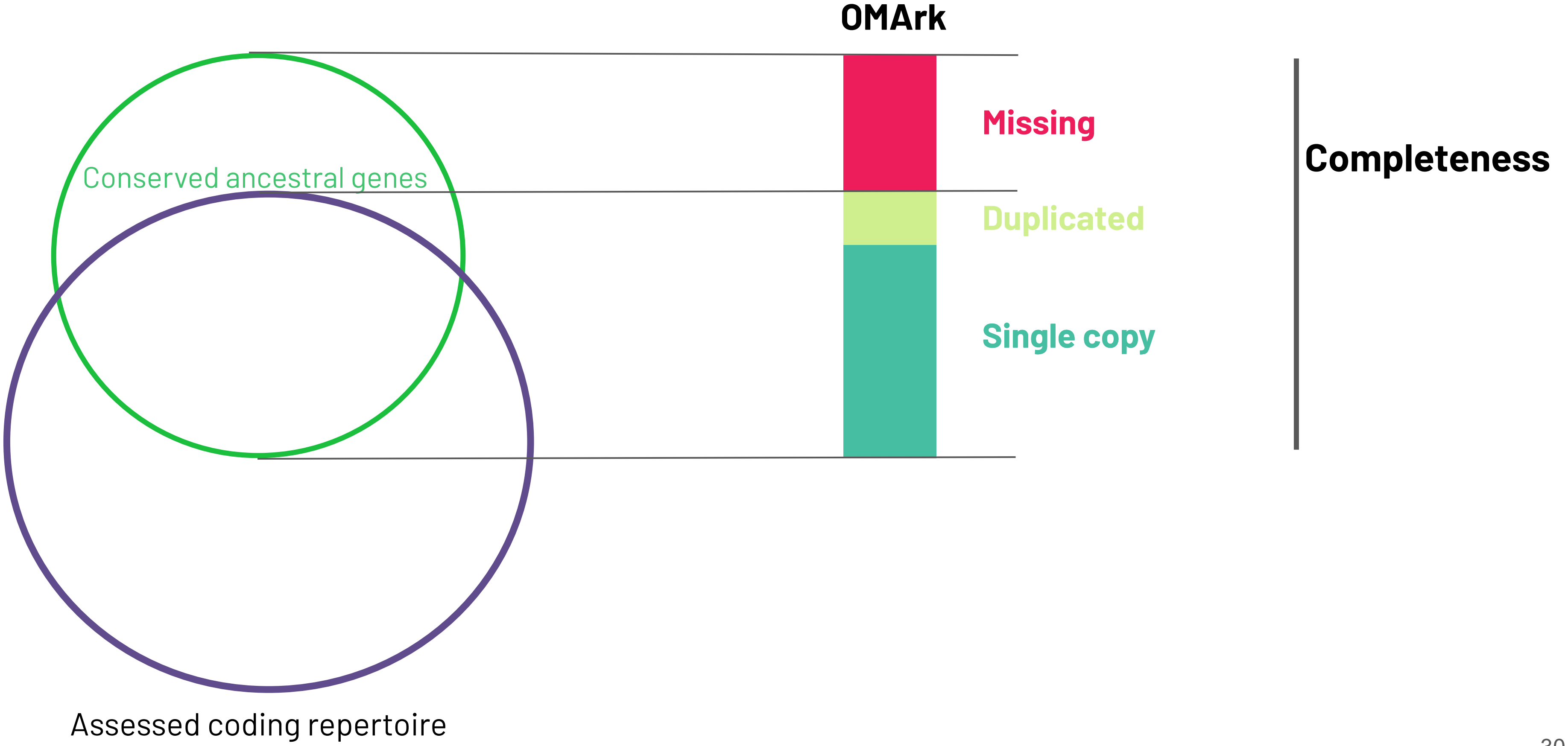
Ancestral lineage :

- ➢ Latest ancestor clades in with 5+ representatives in OMA
- ➢ Dynamically selected from taxid or from the placements

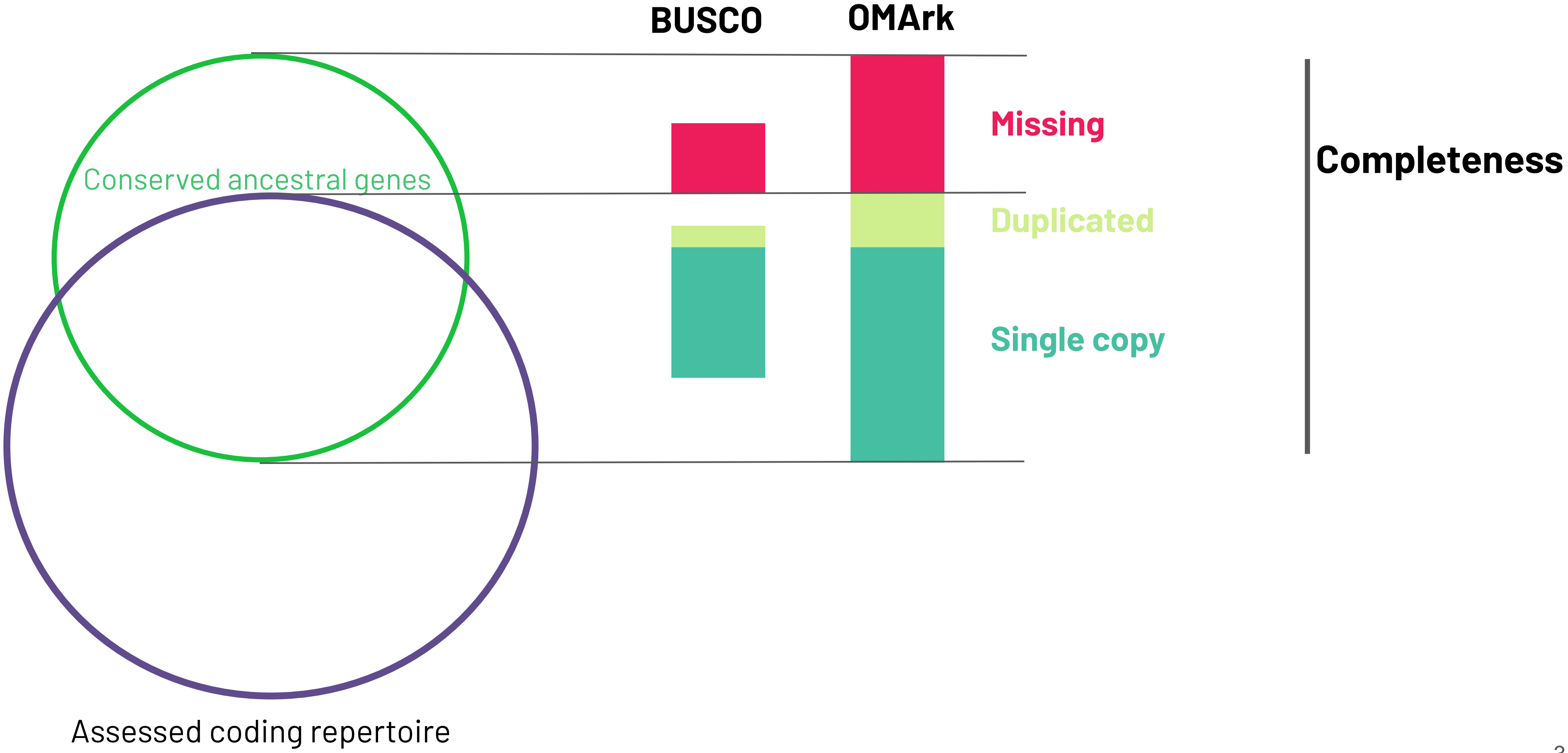
Conserved ancestral genes:

- ➢ Gene families defined at the ancestral lineage level (ancestral gene repertoire)
- ➢ Present in at least 80% species

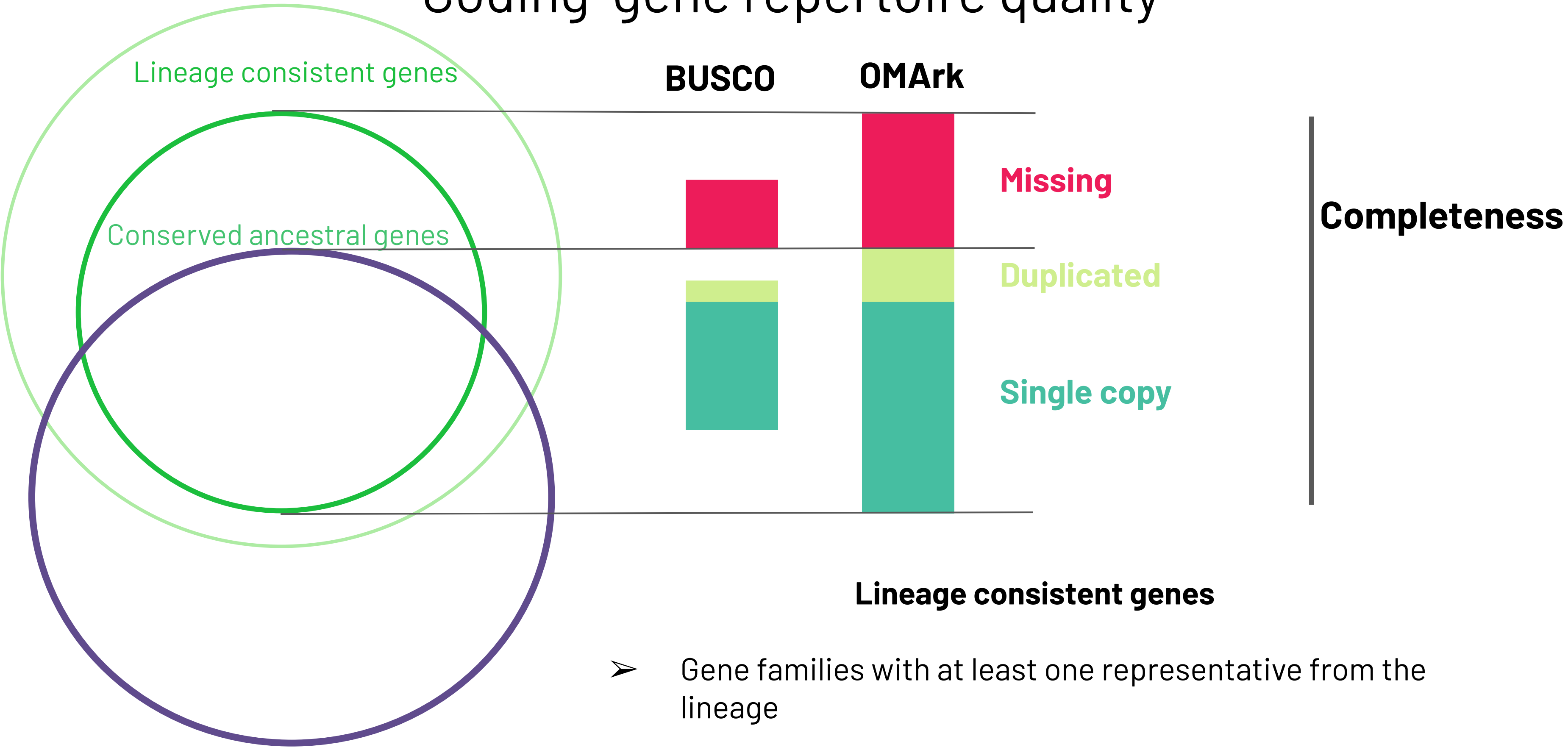
Coding-gene repertoire quality



Coding-gene repertoire quality



Coding-gene repertoire quality



Lineage consistent genes

Conserved ancestral genes

BUSCO

OMArk

Missing

Duplicated

Single copy

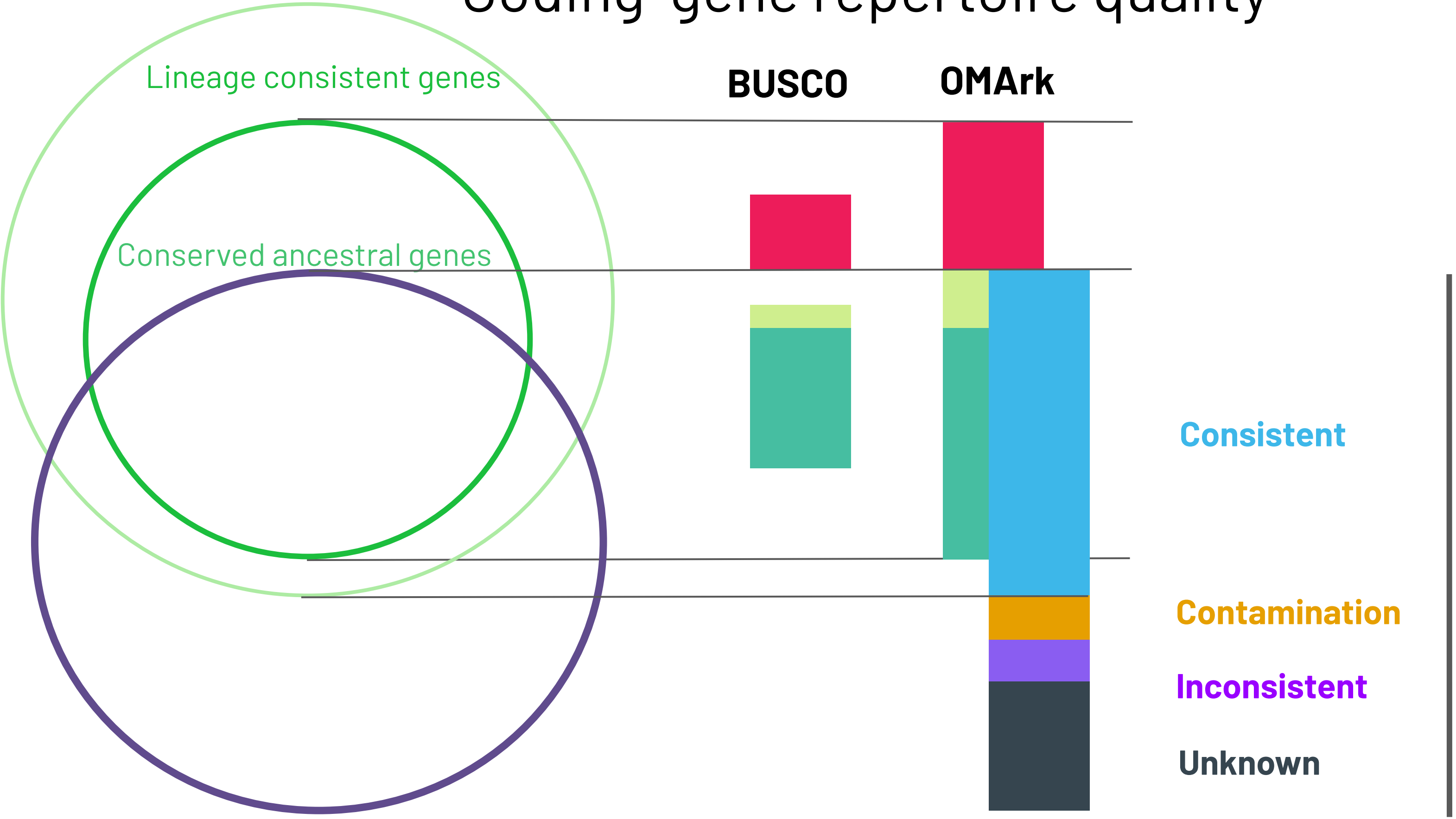
Completeness

Assessed coding repertoire

Lineage consistent genes

➢ Gene families with at least one representative from the lineage

Coding-gene repertoire quality



Consistency assessment

Assessed coding repertoire

Results - Graph summary

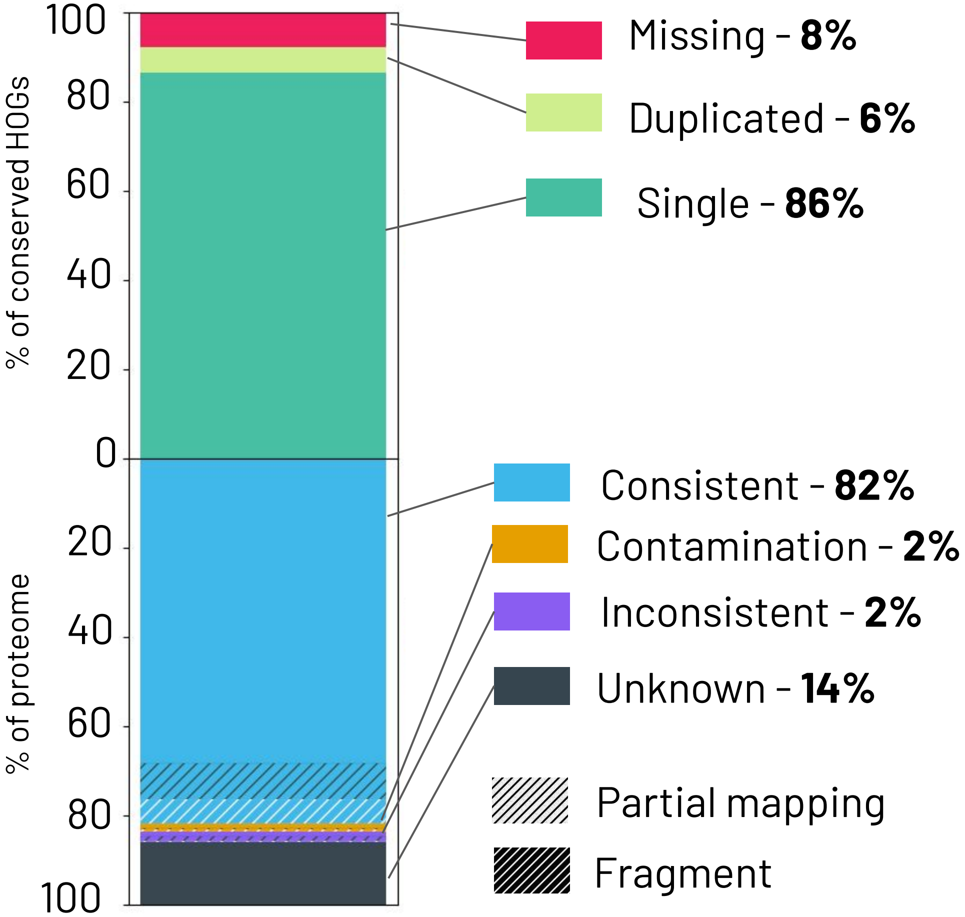
Source: wikipedia

Big-headed turtle
Platysternon megacephalum

Clade : **Archelosauria**

10,514 conserved HOGs

Number of genes : **21,371**



Hand-on exercices

Gitpod

https://omabrowser.org/oma/academy/

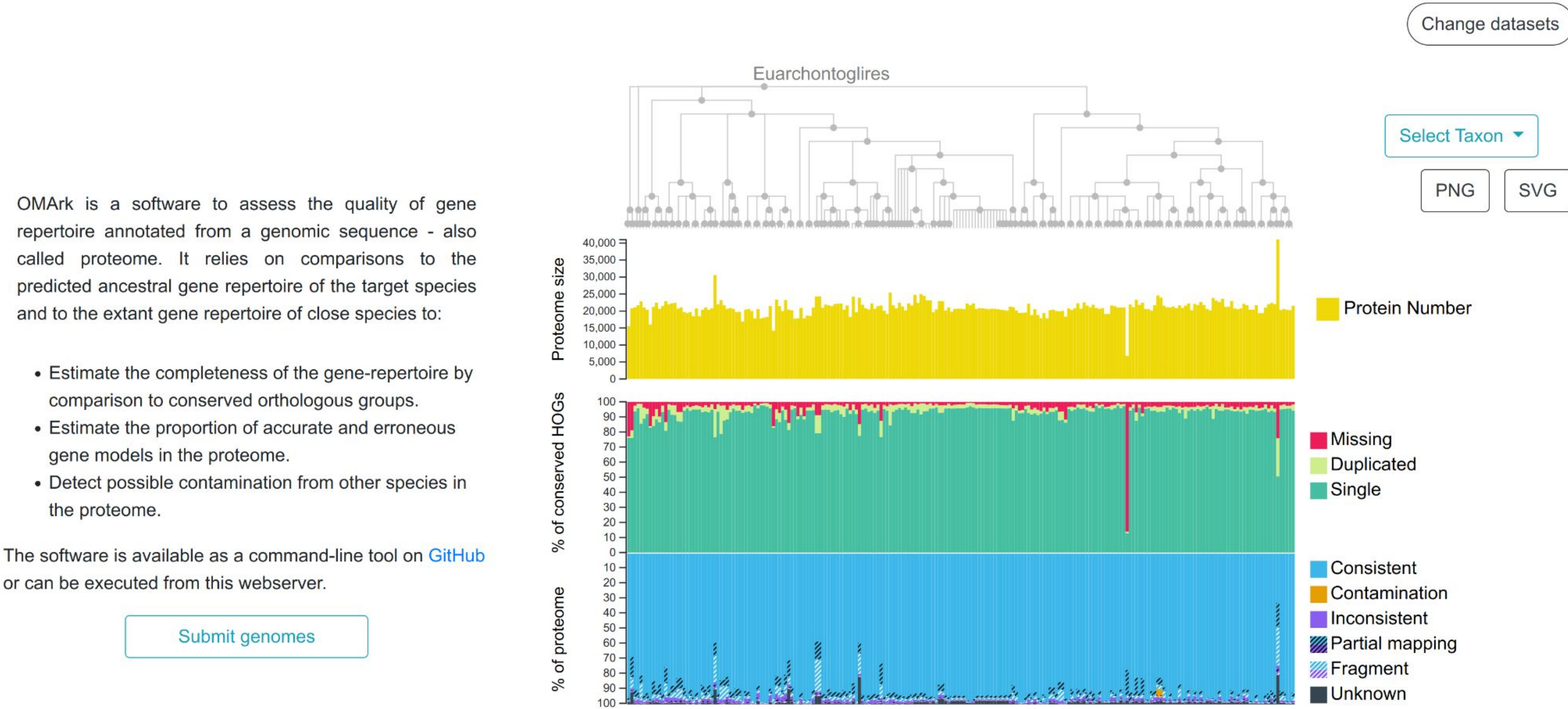
https://tinyurl.com/OMABGA24

using output of omamer from the expected_outputs

```
$ cd working_dir

$ cp ../expected_outputs/omamer/Monmon.omamer.txt omamer/
```

Hint: use tab to navigate between files!

https://omark.omabrowser.org/



OMArk is a software to assess the quality of gene repertoire annotated from a genomic sequence - also called proteome. It relies on comparisons to the predicted ancestral gene repertoire of the target species and to the extant gene repertoire of close species to:

- Estimate the completeness of the gene-repertoire by comparison to conserved orthologous groups.
- Estimate the proportion of accurate and erroneous gene models in the proteome.
- Detect possible contamination from other species in the proteome.

The software is available as a command-line tool on GitHub or can be executed from this webserver.