

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

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 ?

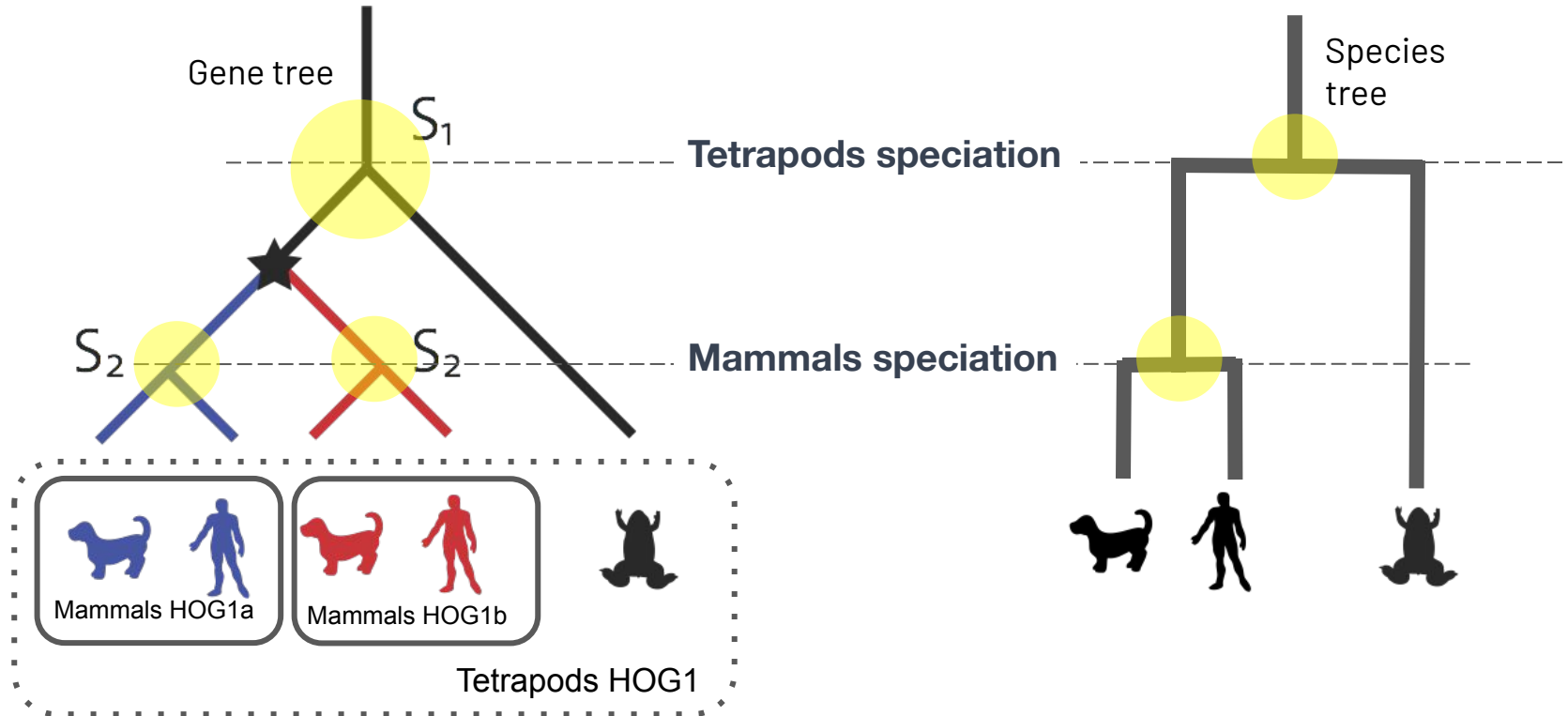
Run OMArk for proteome quality assessment and interpret results

Session plan

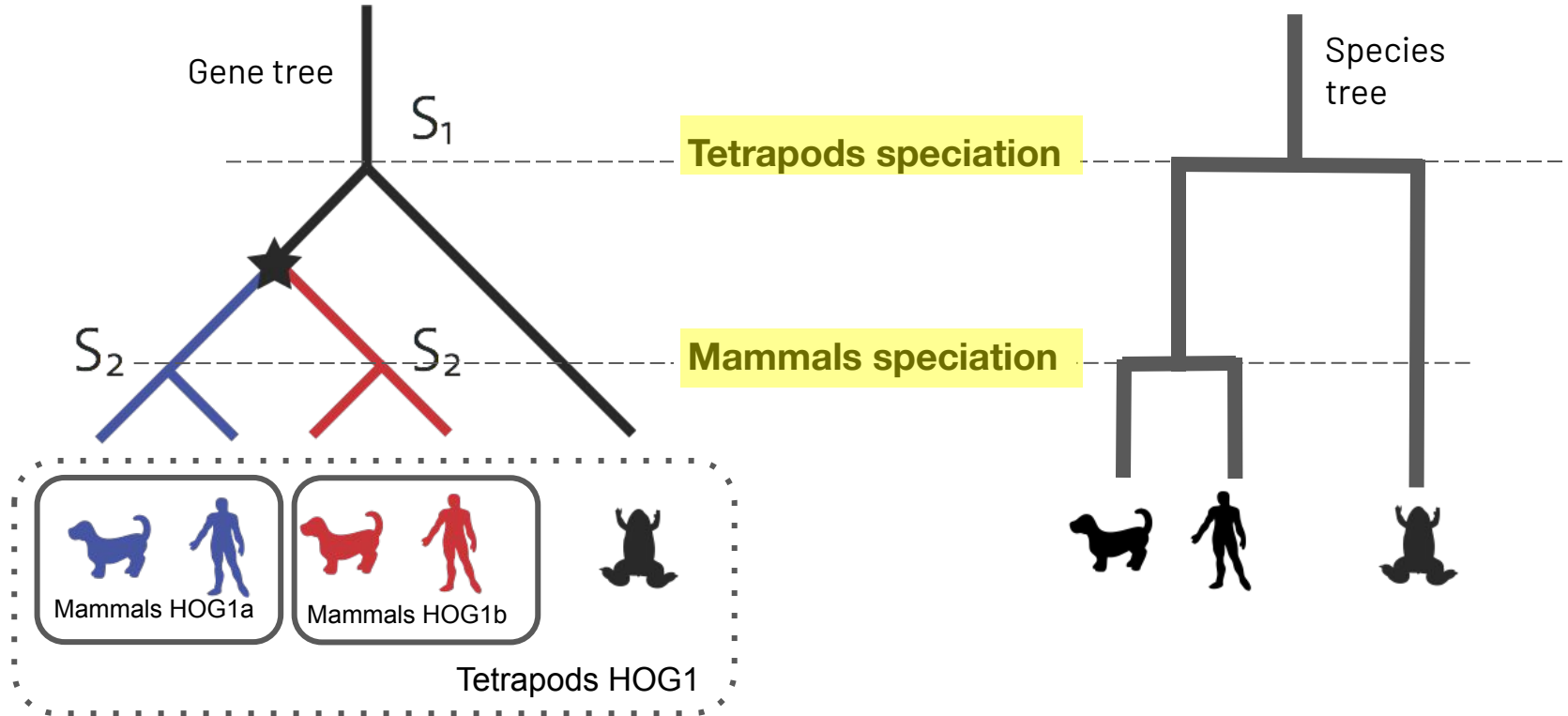
1. Hierarchical Orthologous Groups and the OMA Browser
2. Fast sequence placement with OMAmer
3. Gene repertoire quality assessment with OMArk

Hierarchical Orthologous Groups (HOGs)

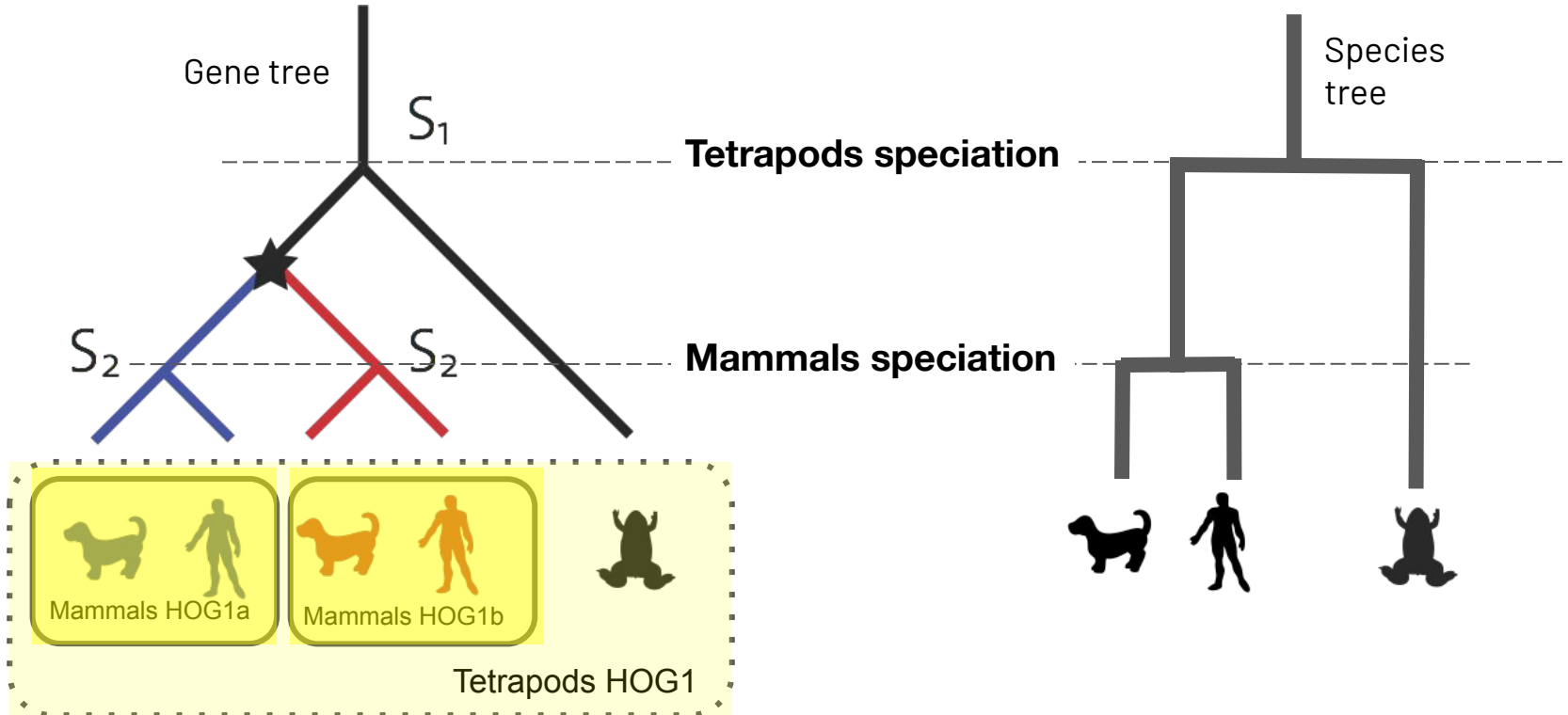
HOGs = Sets of genes that descended from a common ancestral gene in a given ancestral species



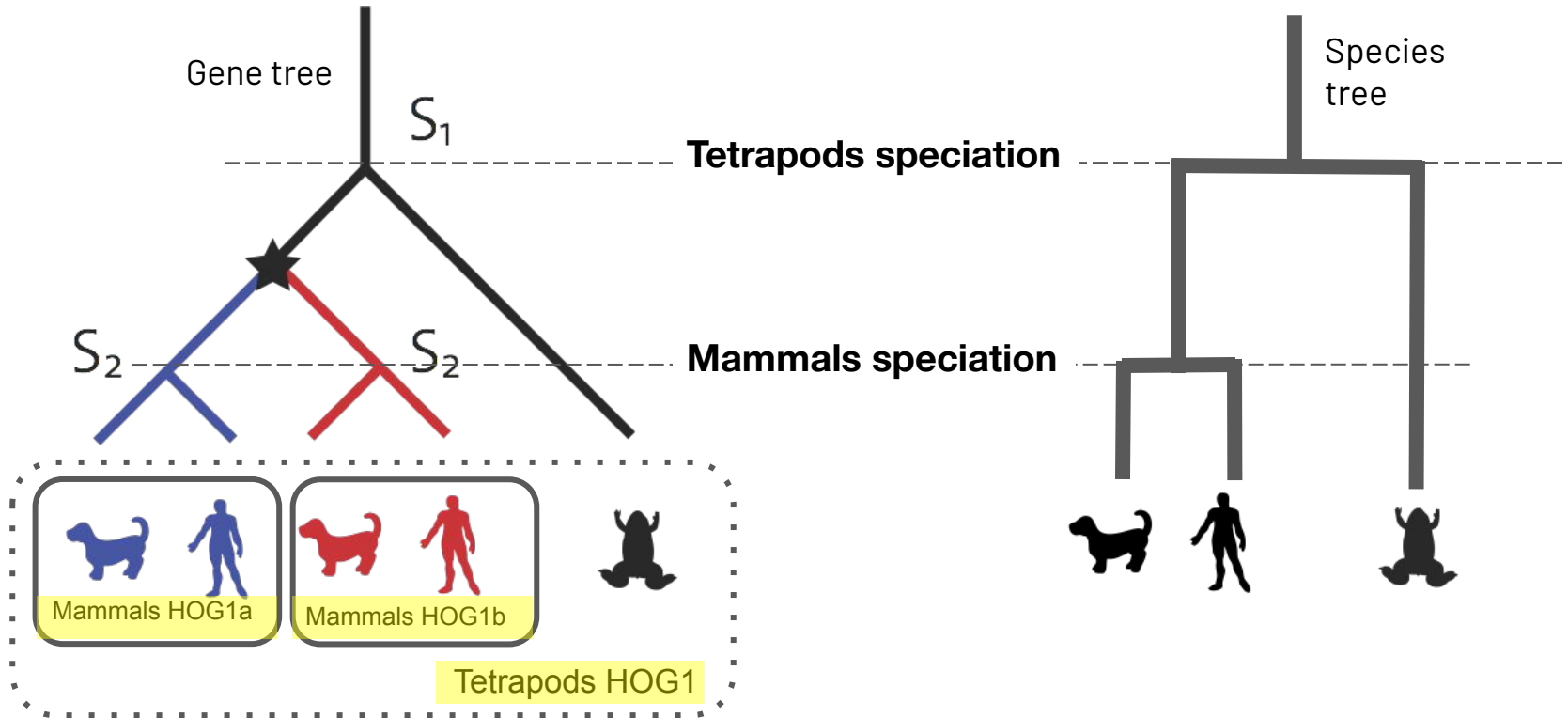
HOGs are defined with respect to specific clades



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



The OMA browser



Explore ▾ Tools ▾ Download ▾ Help ▾ About ▾



? "Blue-light photoreceptor" | proteinid:P53_RAT | species:"Drosophila melanogaster"

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

SCROLL TO DISCOVER MORE

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

<https://omabrowser.org/>

<https://oma-stage.vital-it.ch/>

(Please use this one for the exercises)

Hierarchical Orthologous Groups (HOGs)

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

Completeness score: 0.75 ⓘ

Ancestral Genome

Primates / Lower Level ▶

Hierarchical group HOG:0606964 open at level of **Primates**

OPTIONS ▼

Graphical viewer

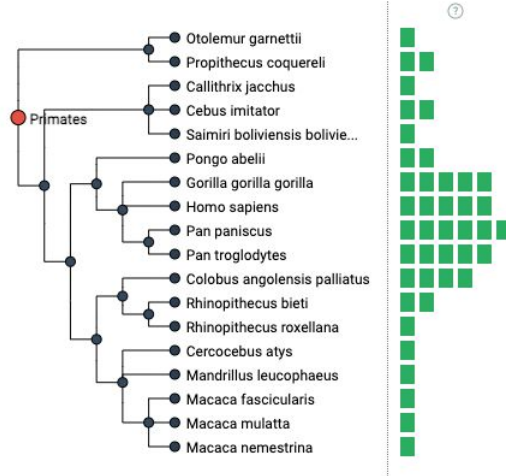
Members

Ancestral GO

Alignment

Ancestral synten

Similar HOGs >



- 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 38457 ancestral genes (HOGs).

Remove HOGs with completeness score below

Genome information

Ancestral genes

PHYLOGENETIC FILTER:

Choose a target genome

Ancestor genome

Descendant genome

Ancestral Gene Order

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

Hand-on exercises



<https://oma-stage.vital-it.ch/>

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

<https://tinyurl.com/BGAOMA>

Fast sequence placements with OM Amer

What is OMamer?

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

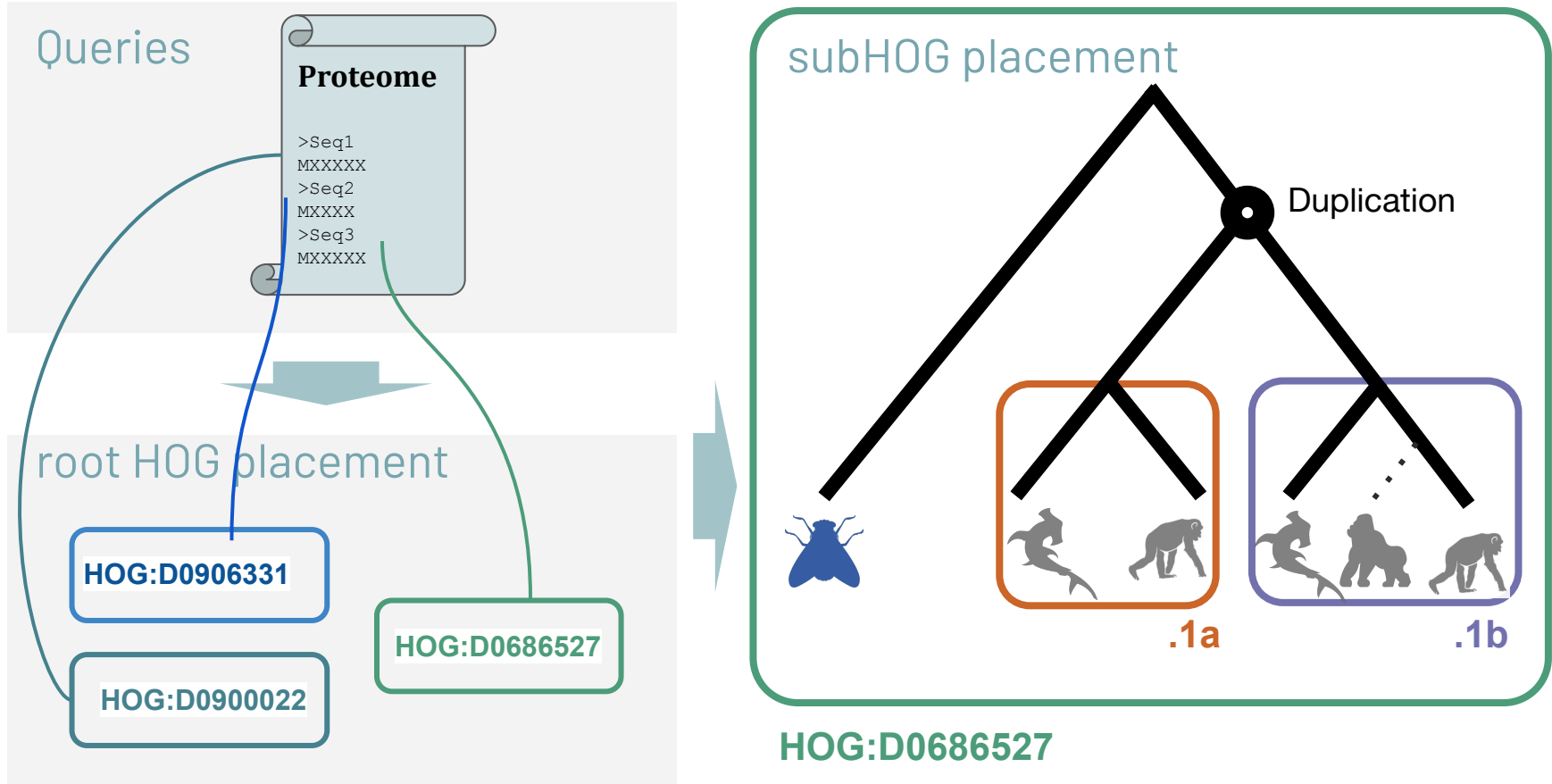
OMamer: tree-driven and alignment-free protein assignment to subfamilies outperforms closest sequence approaches

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>^{†4}

OMAmer placement - principle



k-mer based placement

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

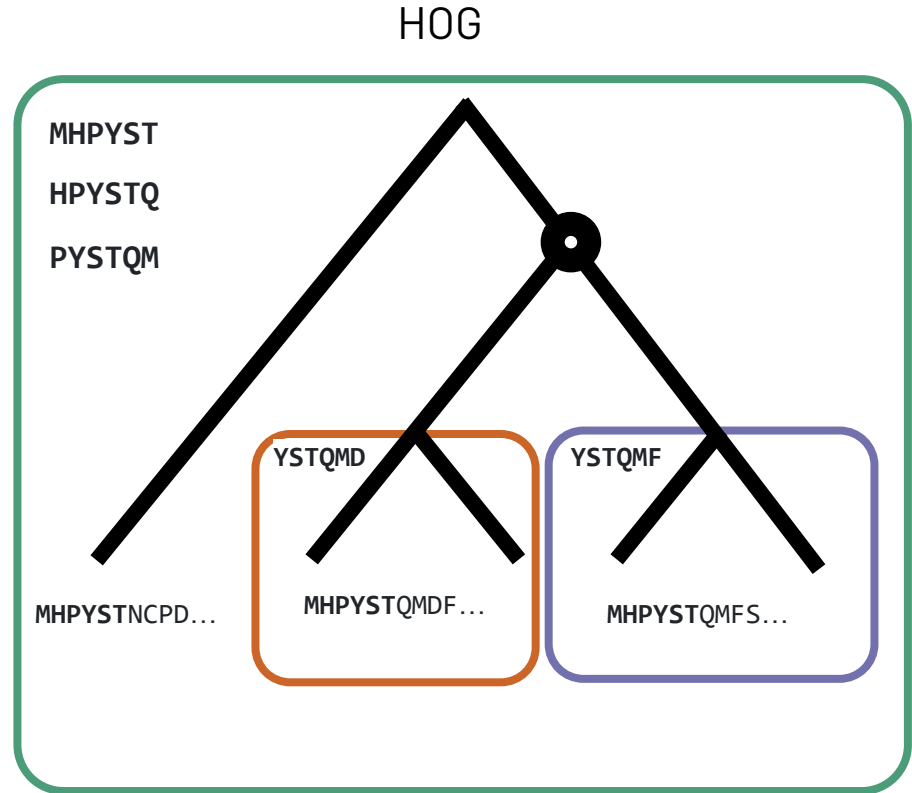
Query sequence

MHPYSTQMFS LQITVMEDSQ SDMSIELPLS

MHPYST
HPYSTQ
PYSTQM

...
...
...

MSIELP
SIELPL
IELPLS



How to use OMAMer

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

Proteome

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

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

Hand-on exercises



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

<https://tinyurl.com/BGAOMA>

Quality assessment with OMArk

How to use OMAMer

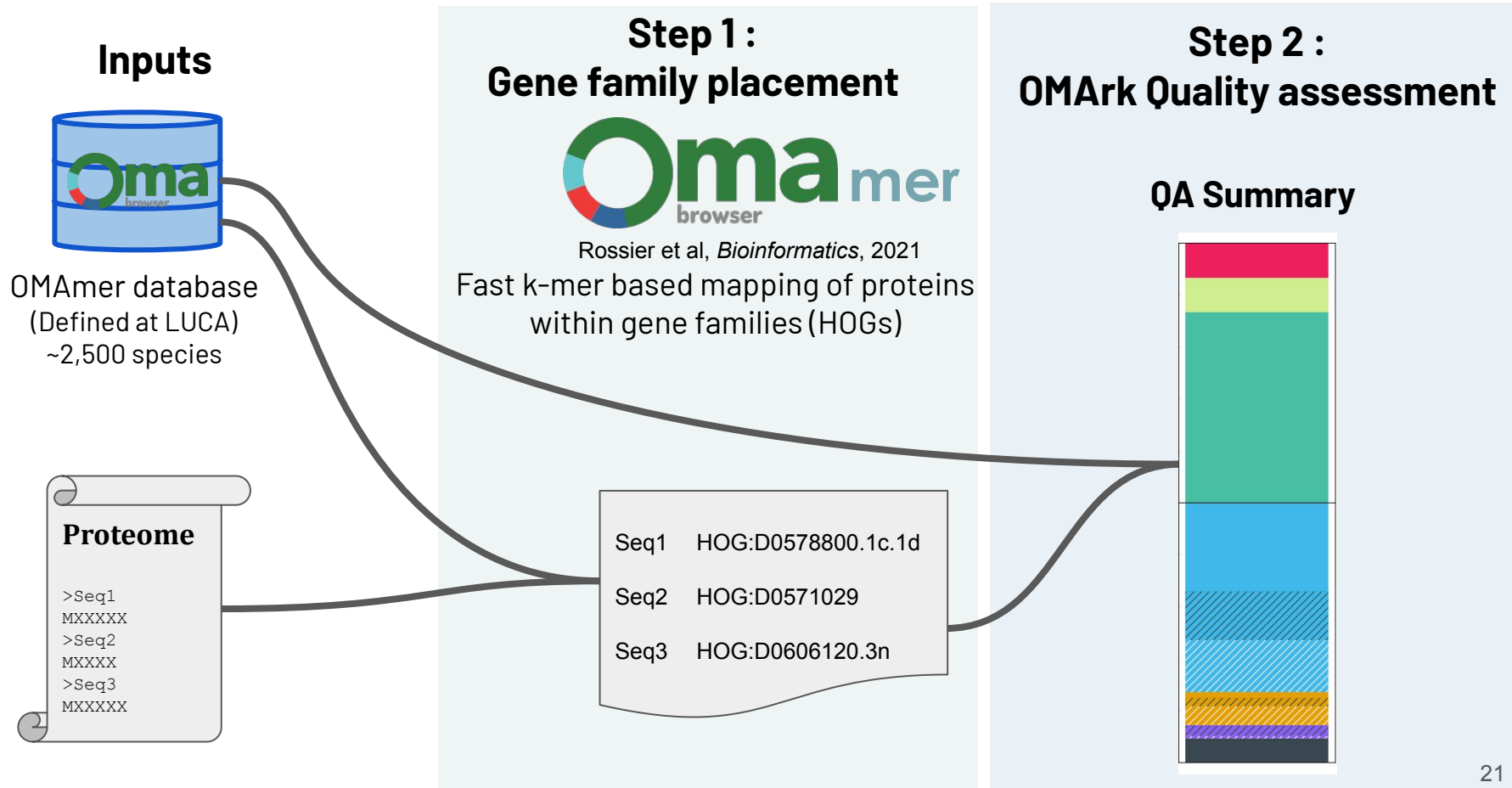
Coding-gene repertoire : set of coding-genes annotated on a given genome sequence

Available on database as **proteomes**

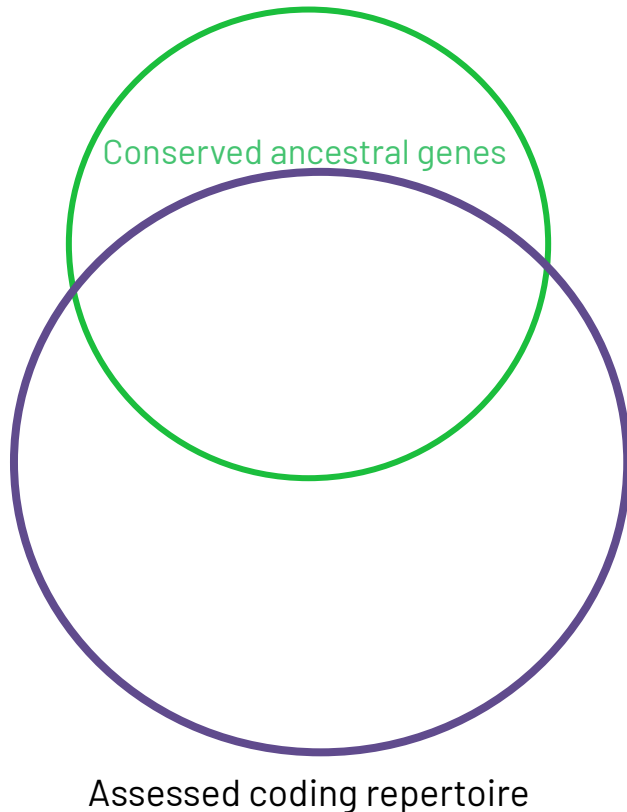
- Subject to quality issues
- Missing genes
 - Fragmented genes
 - Inclusion of non-coding regions
 - Contamination

Lack of tool to detect all these issues !

Coding-gene repertoire quality



Coding-gene repertoire quality



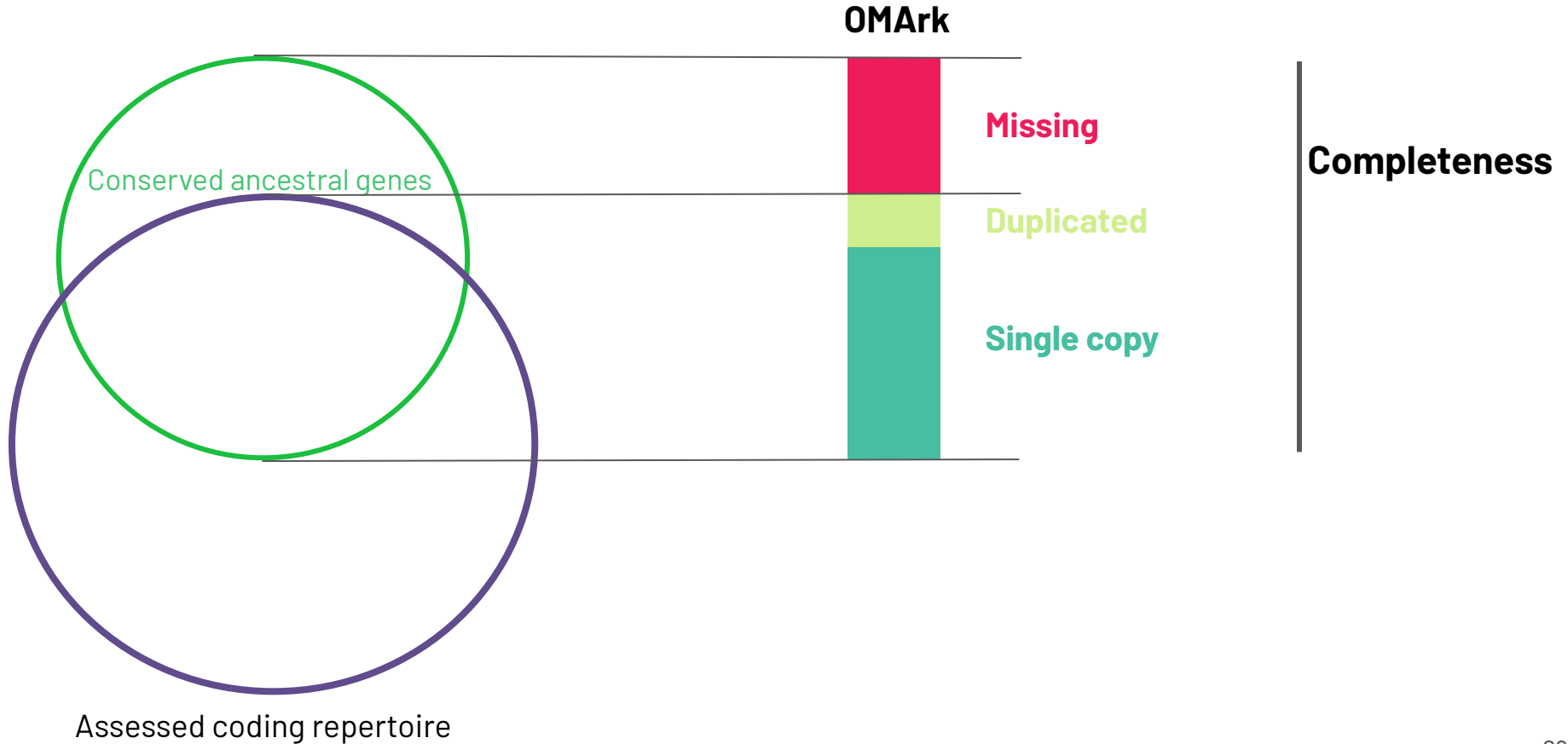
Ancestral lineage :

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

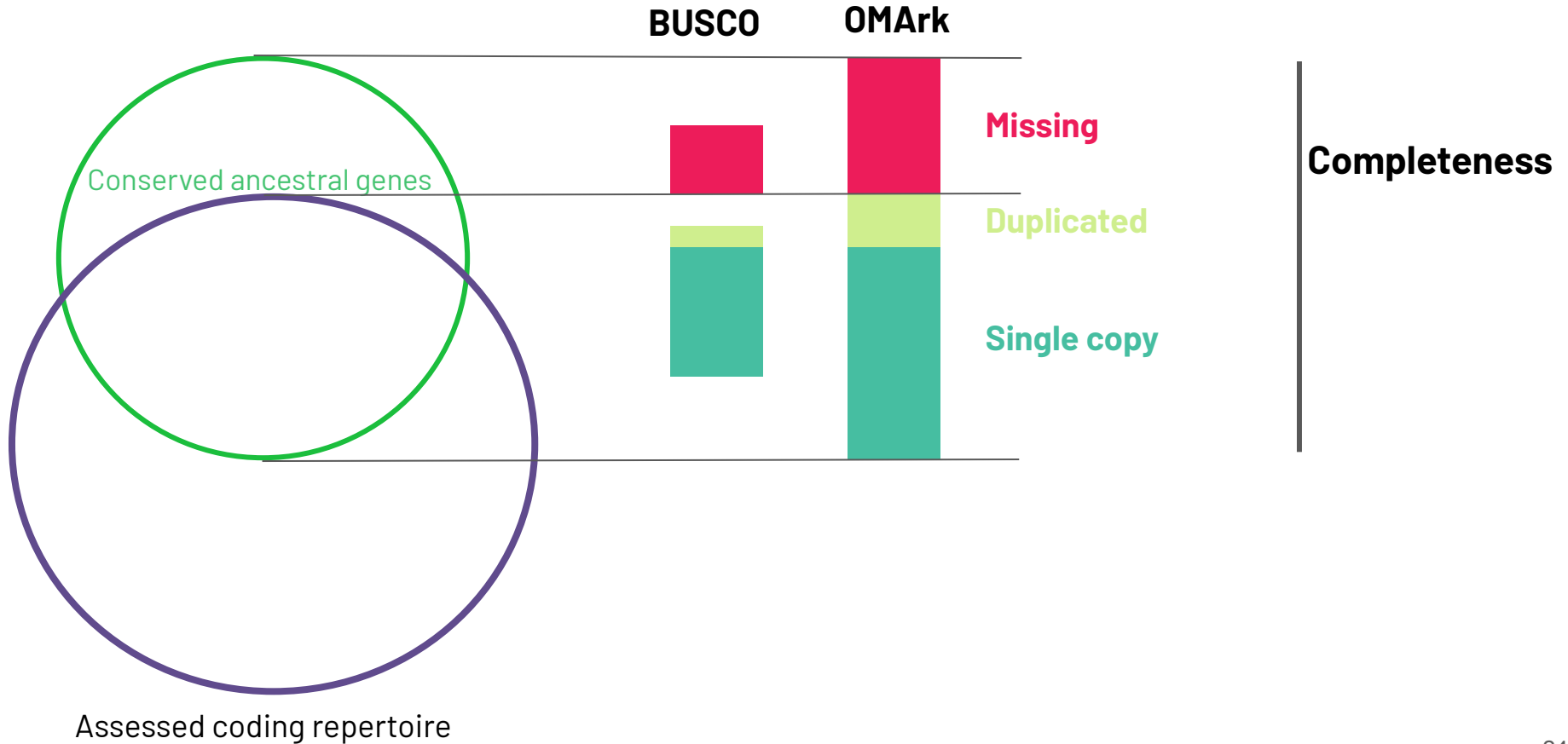
Conserved ancestral genes:

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

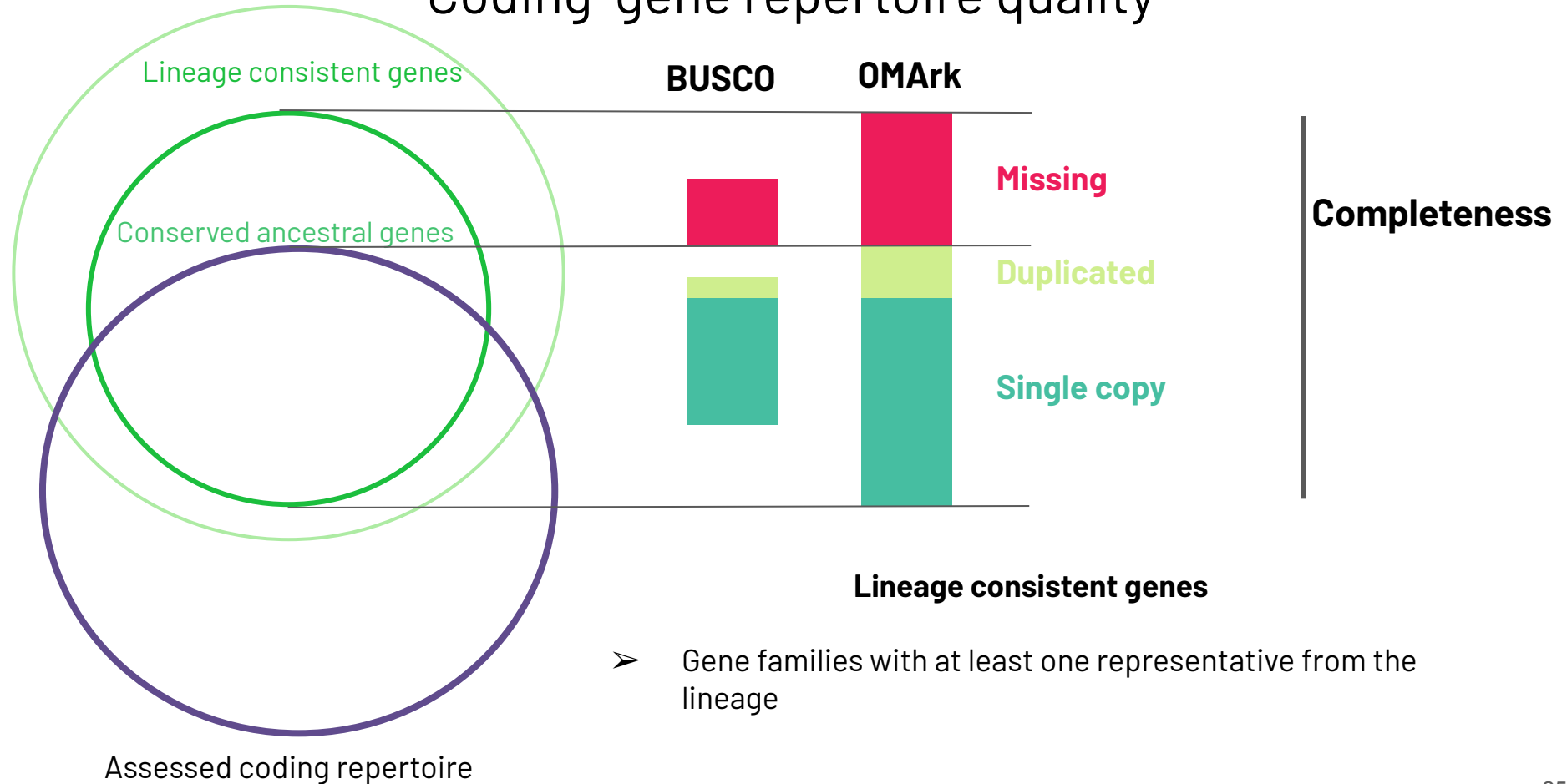
Coding-gene repertoire quality



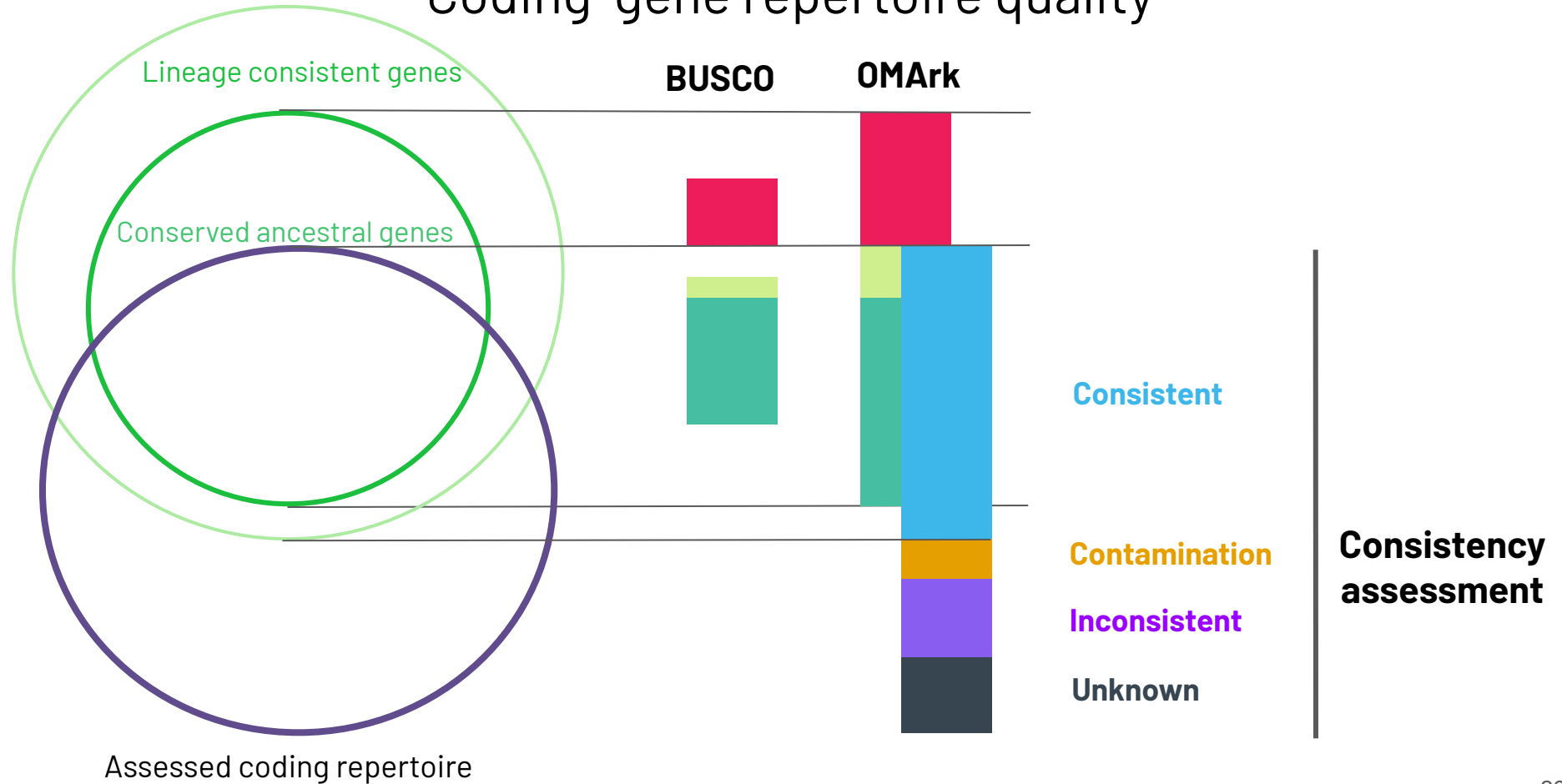
Coding-gene repertoire quality



Coding-gene repertoire quality



Coding-gene repertoire quality





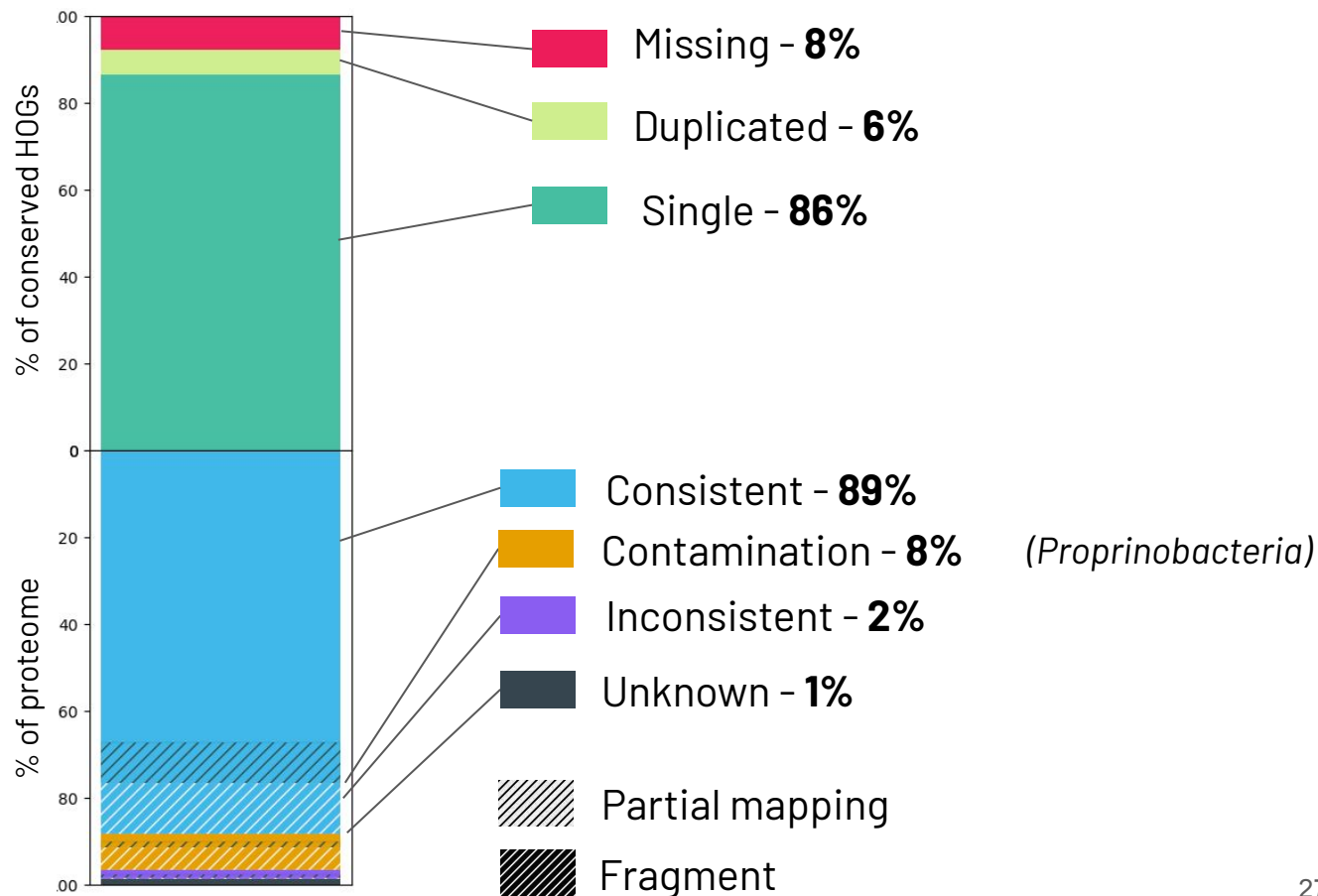
Big-headed turtle
Platysternon megacephalum

Clade : **Archelosauria**

10,514 conserved HOGs

Number of genes : **21,371**

Results - Graph summary



Hand-on exercises



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

<https://tinyurl.com/BGAOMA>