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

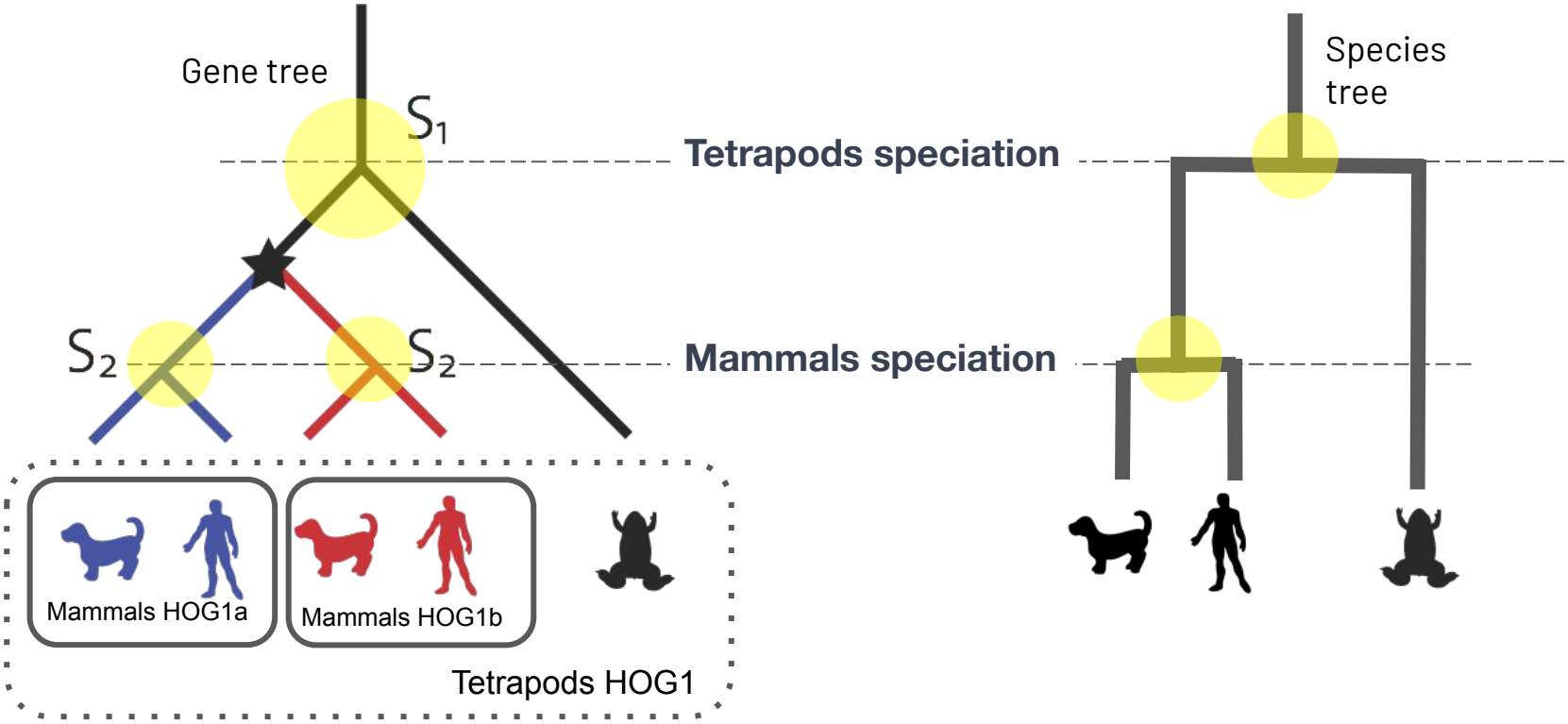
1. Hierarchical Orthologous Groups and the OMA Browser

2. Fast sequence placement with OMAmer

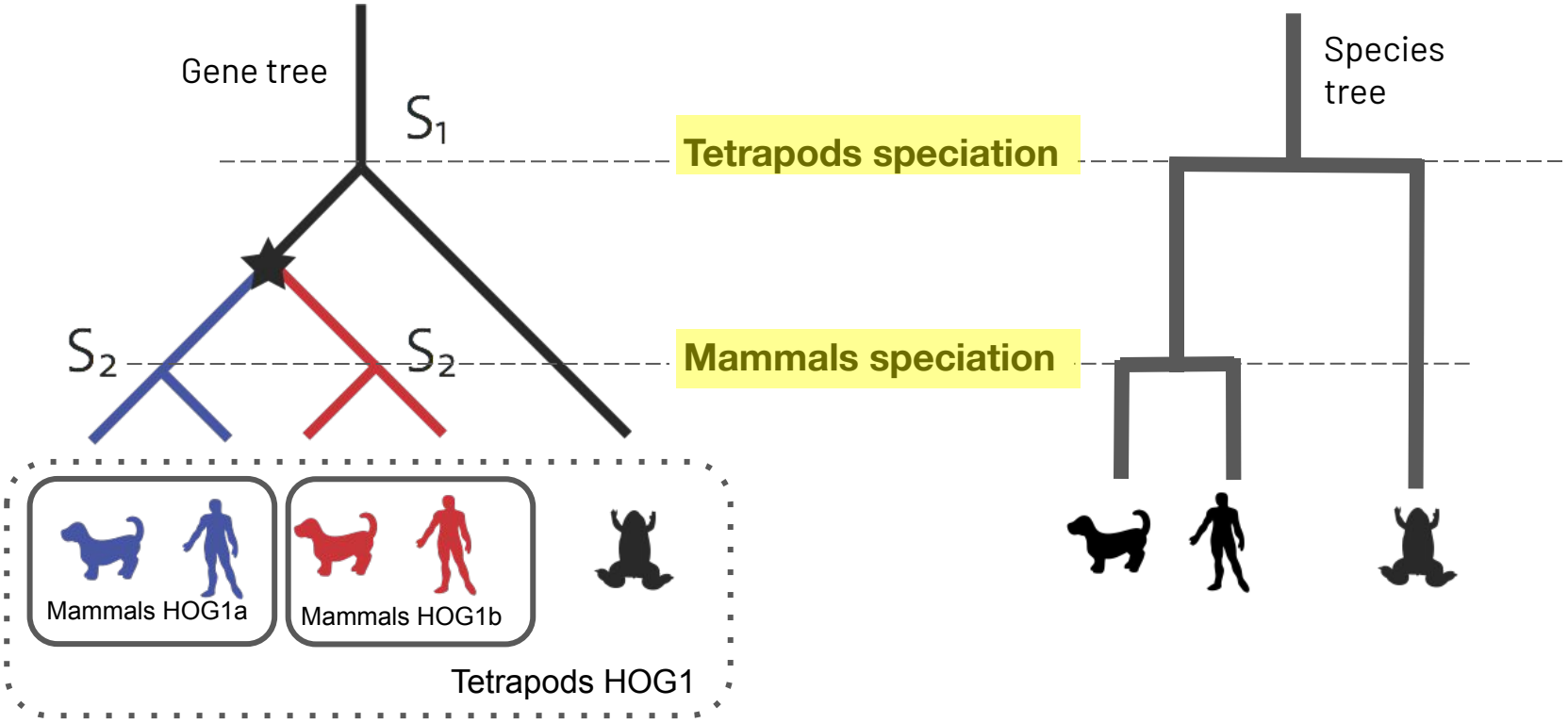
3. Gene repertoire quality assessment with OMArk

Hierarchical Orthologous Groups (HOGs)

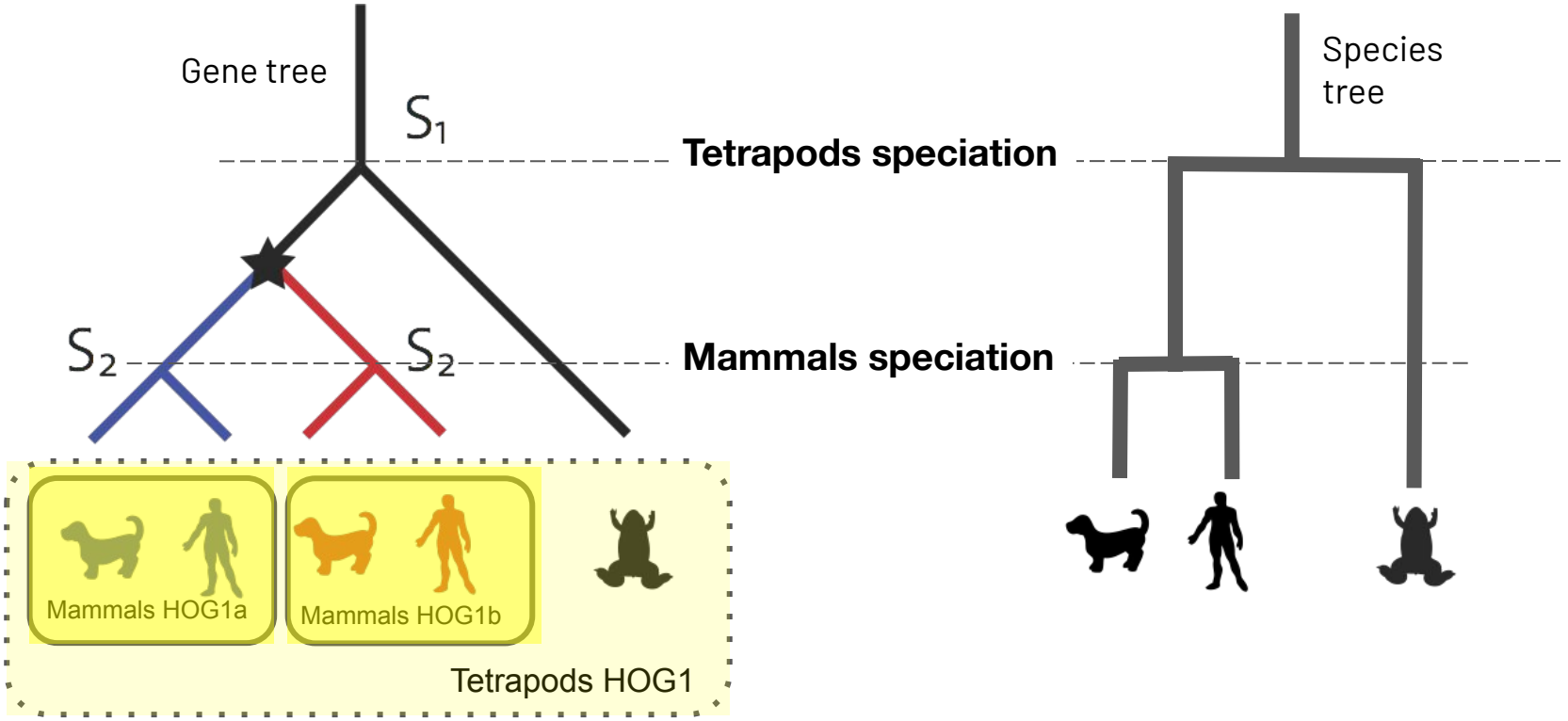
HOGs = Sets of genes that descended from a common ancestral gene in a given ancestral species



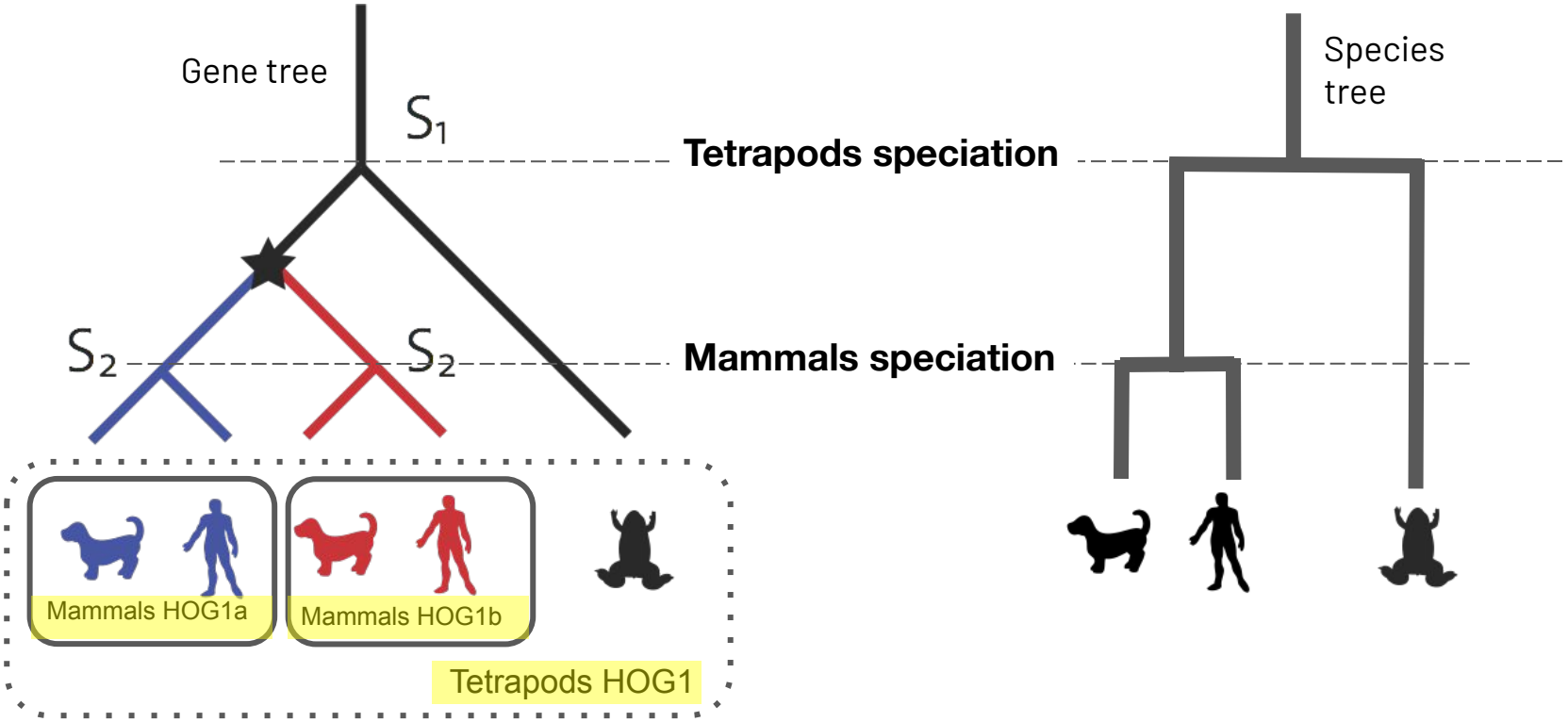
HOGs are defined with respect to specific clades



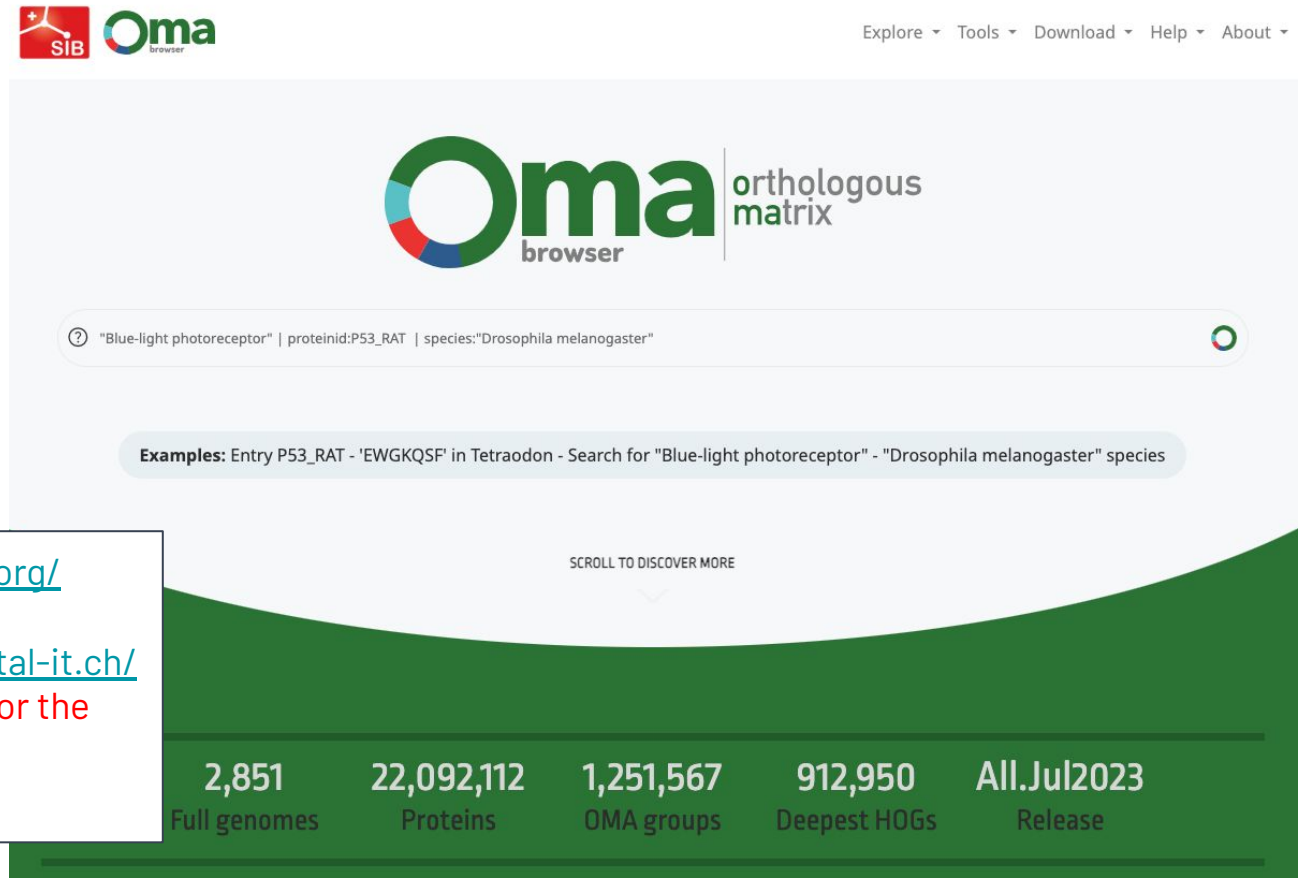
HOGs are hierarchical because groups defined with respect to deeper clades subsume multiple groups defined on their descendants



HOGs are gene families; SubHOGs are nested subfamilies



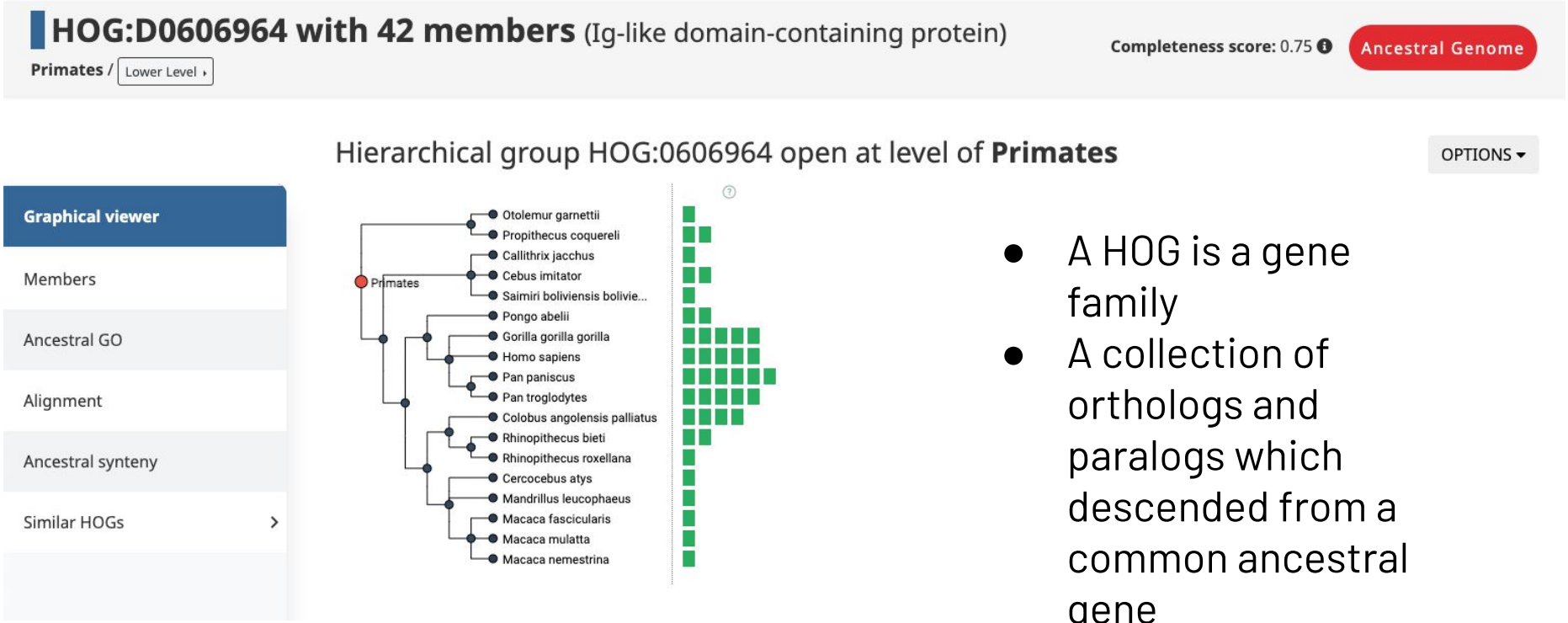
The OMA browser



https://omabrowser.org/

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

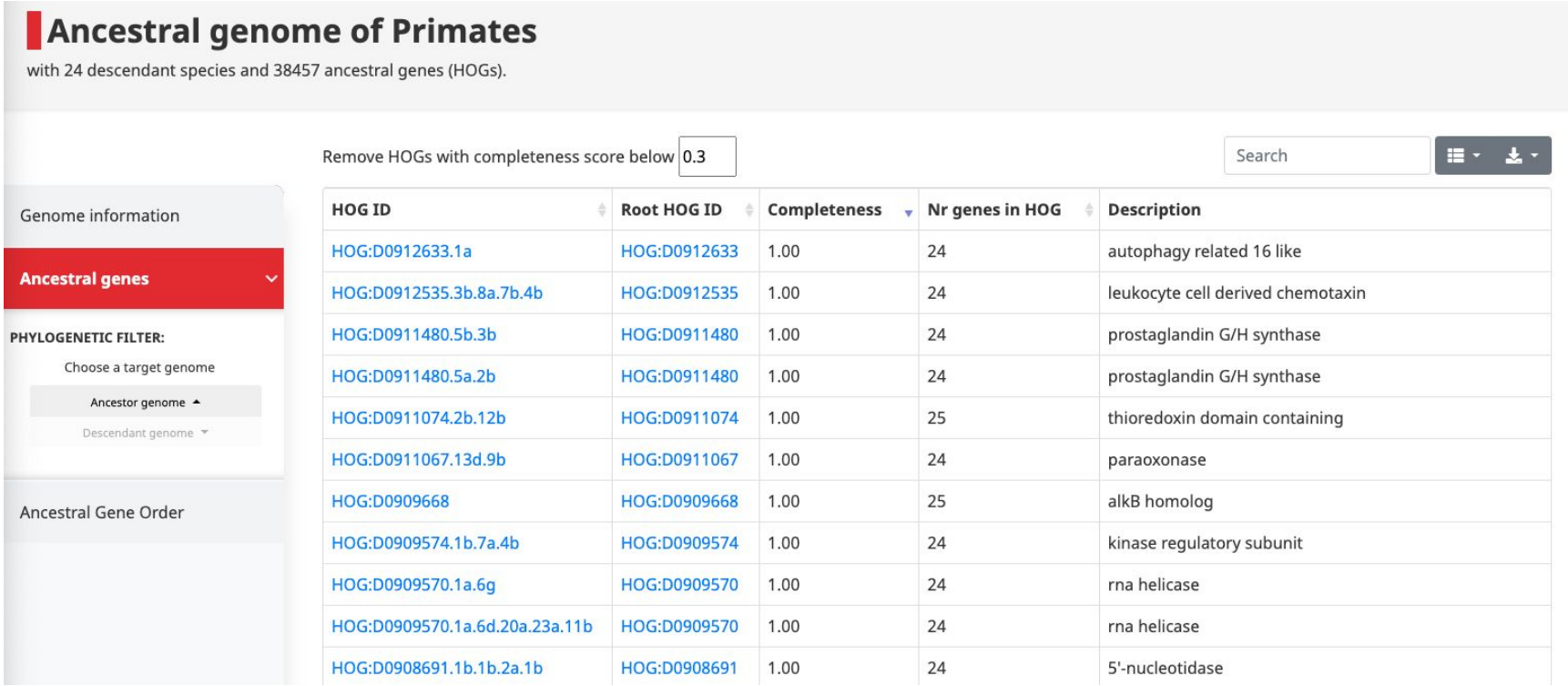
Hierarchical Orthologous Groups (HOGs)



- 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



Hand-on exercices

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

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

https://tinyurl.com/BGAOMA

Fast sequence placements with OMAmer

What is OMAmer?

- Fast sequence placement into existing HOGs from the OMA Browser

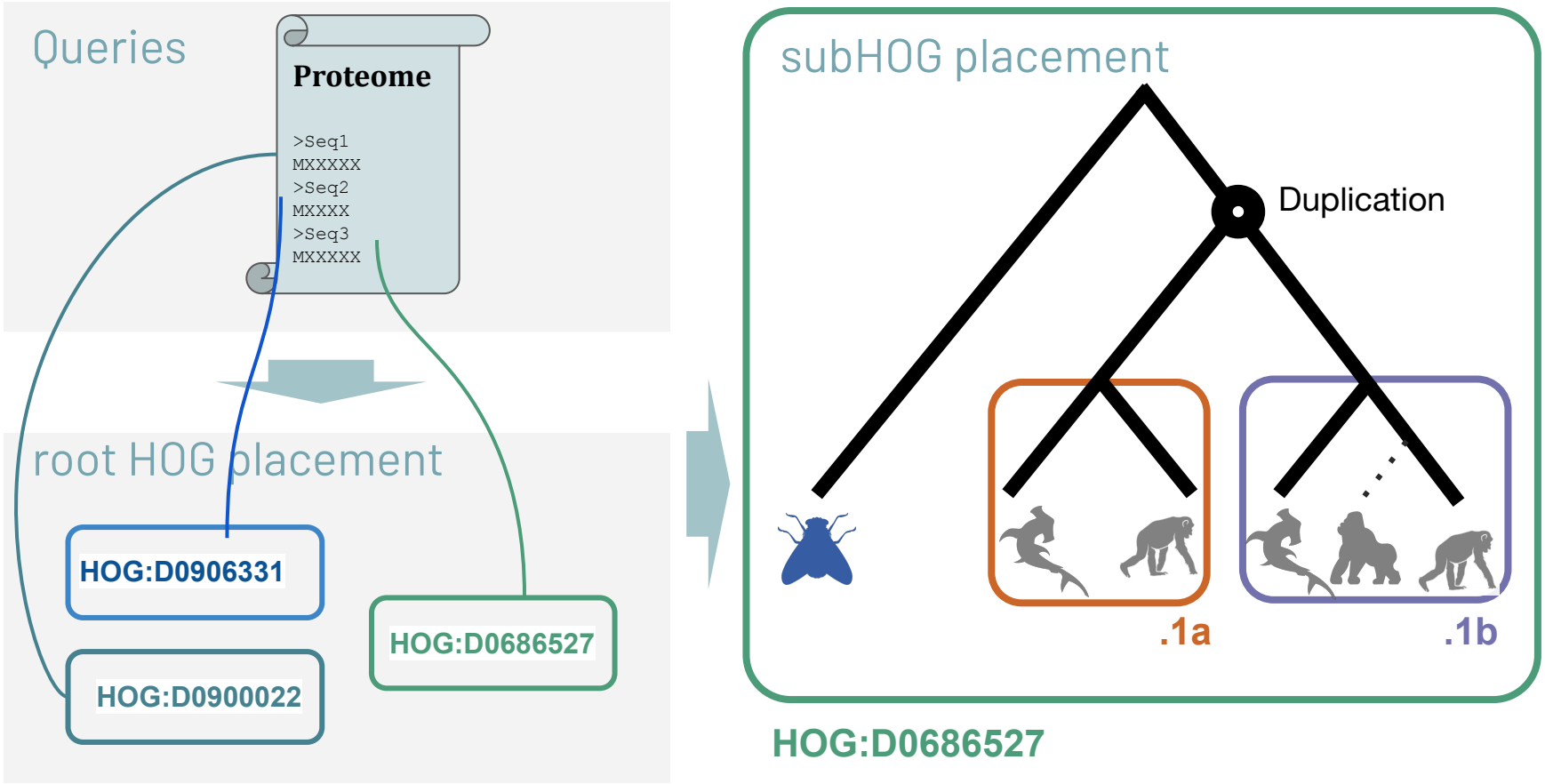
- More accurate than closest sequence matching for subfamily placement!

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

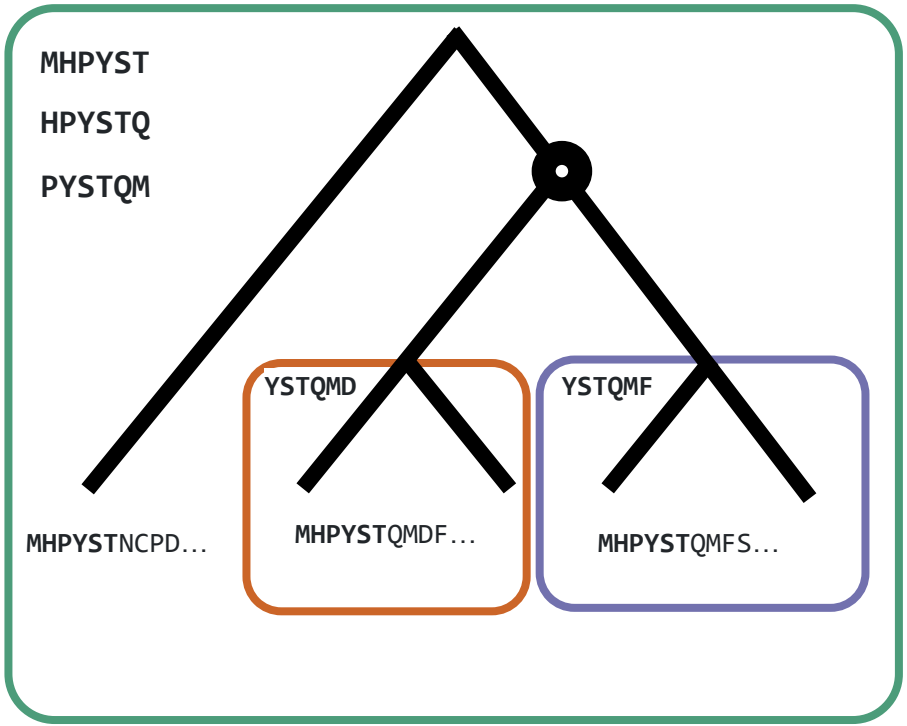


k-mer based placement

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

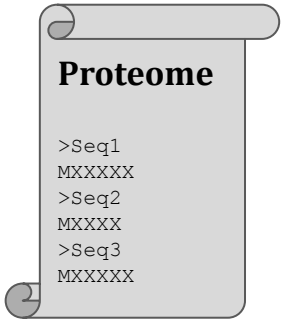
Query sequence

```
MHPYSTQMFS LQITVMEDSQ SDMSIELPLS

MHPYST
 HPYSTQ
  PYSTQM

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



How to use OMAmer

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



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

OMAmer database
HDF5 format

Built with HOGs from the OMA Browser

Hand-on exercices

Gitpod

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

Inputs

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

Proteome

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

Step 1 : Gene family placement

OMAmer browser

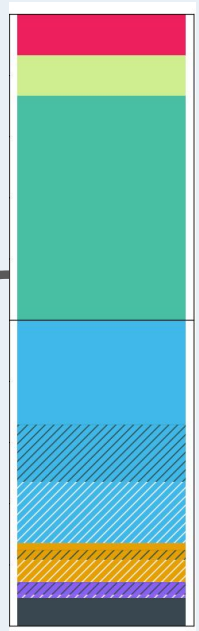
Rossier et al, *Bioinformatics*, 2021

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

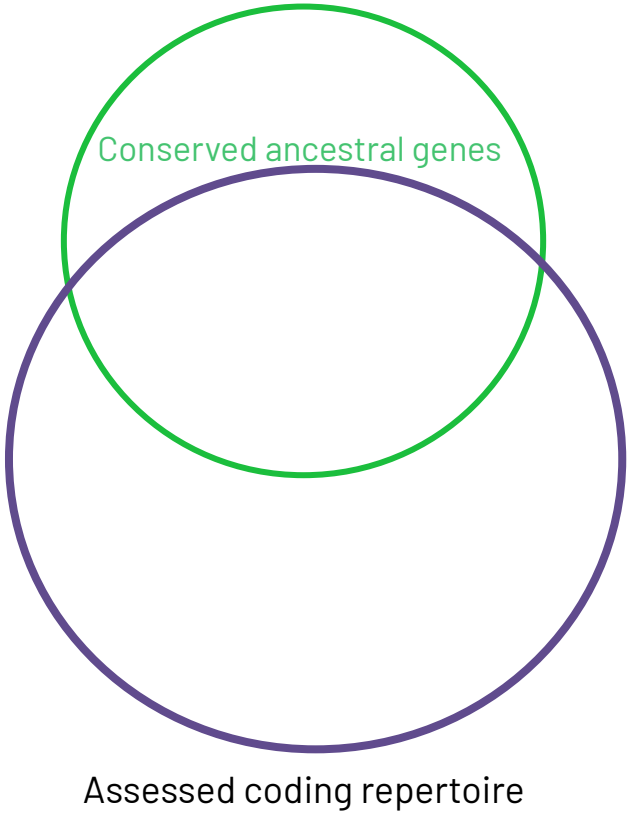
Seq1	HOG:D0578800.1c.1d
Seq2	HOG:D0571029
Seq3	HOG:D0606120.3n

Step 2 : OMArk Quality assessment

QA Summary



Coding-gene repertoire quality



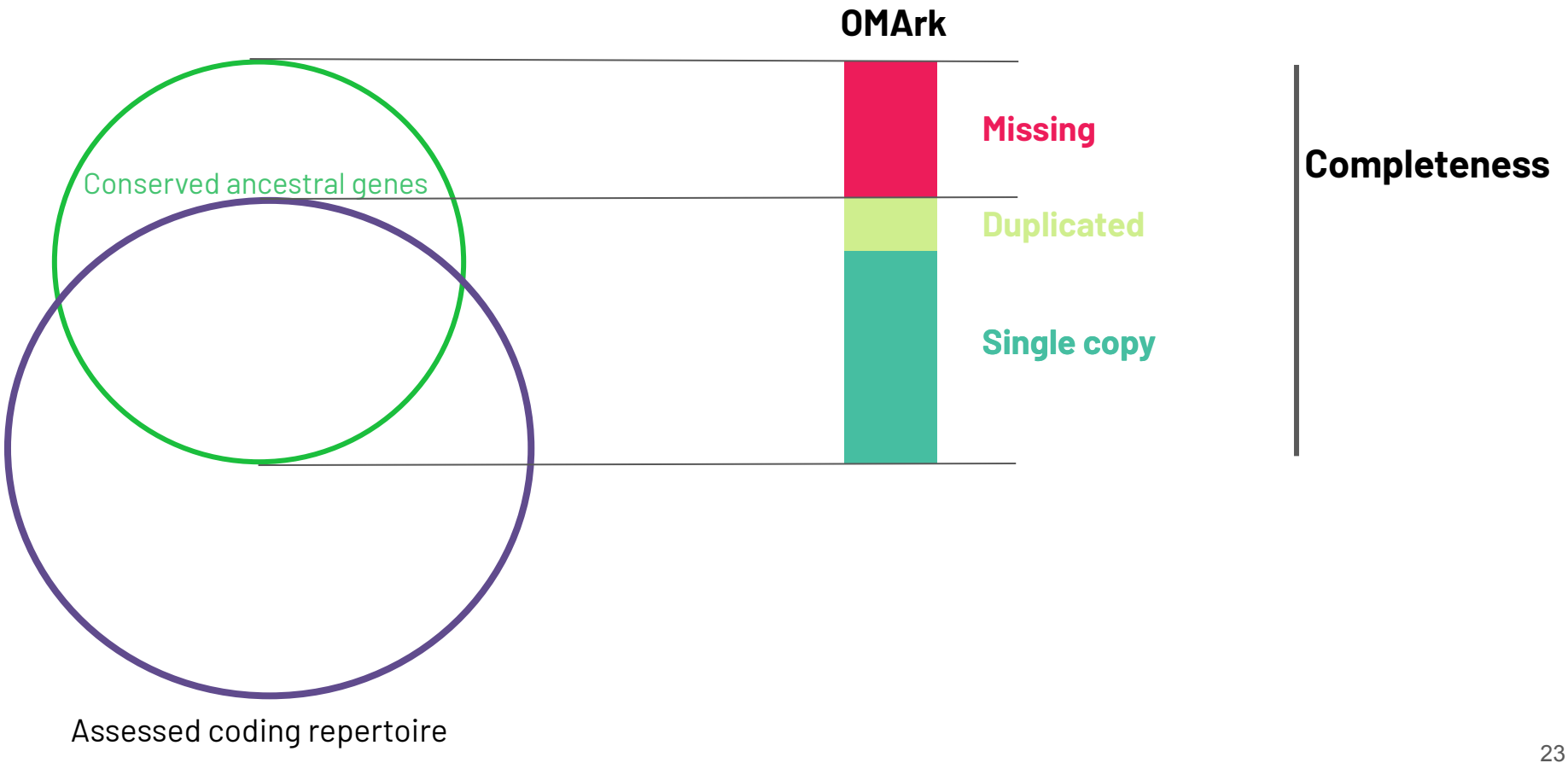
Ancestral lineage :

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

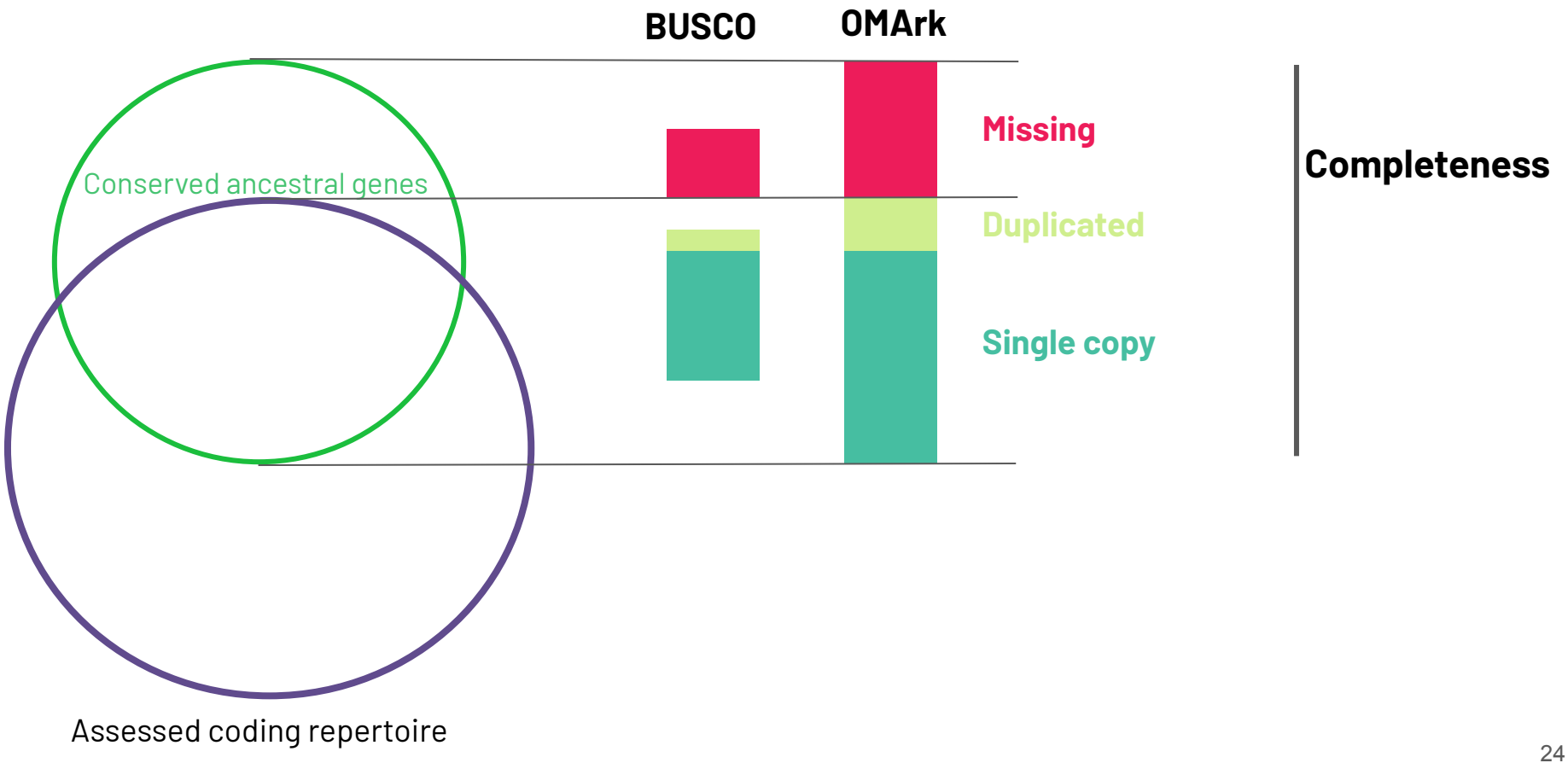
Conserved ancestral genes:

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

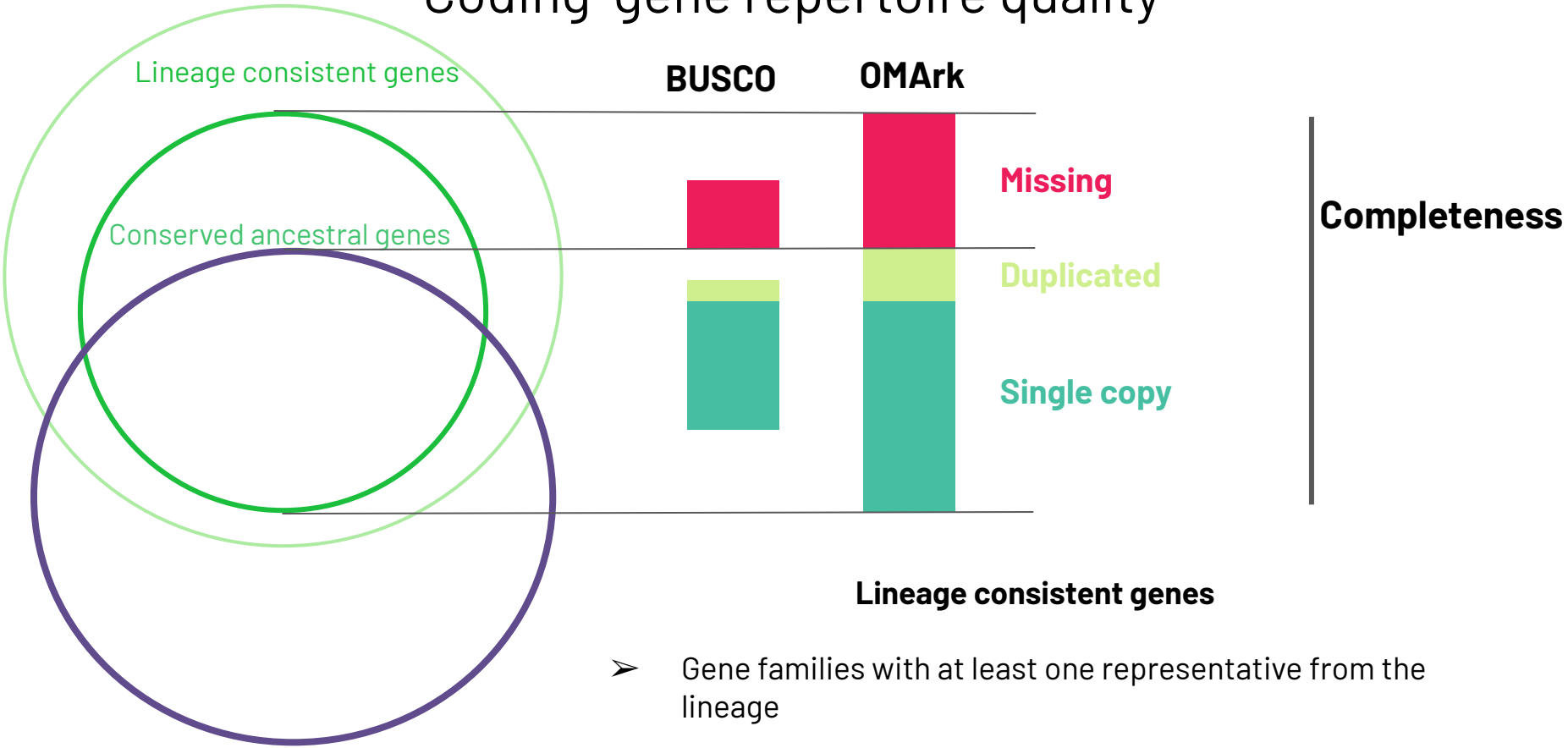
Coding-gene repertoire quality



Coding-gene repertoire quality

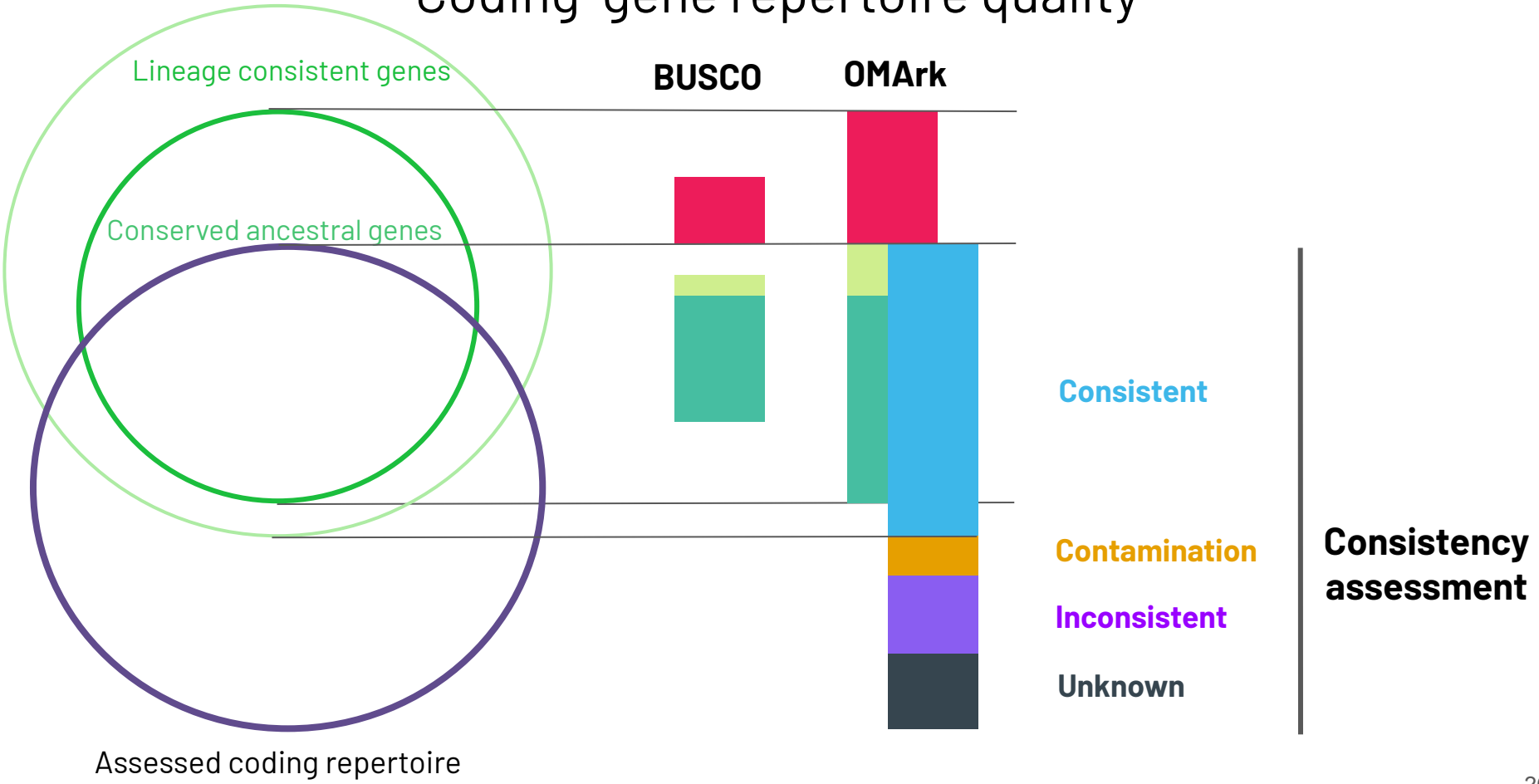


Coding-gene repertoire quality



➢ Gene families with at least one representative from the lineage

Coding-gene repertoire quality



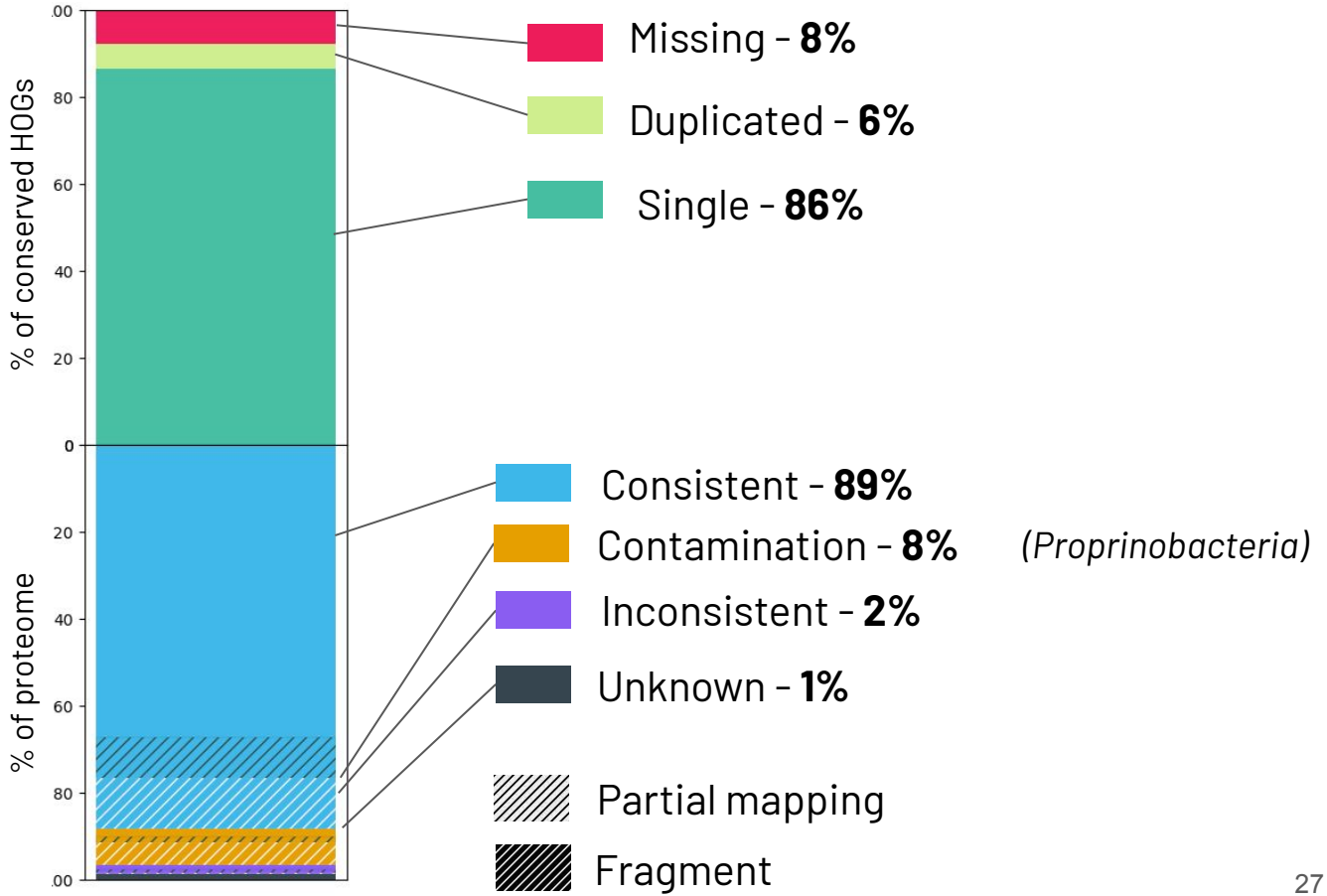
Results - Graph summary

Big-headed turtle
Platysternon megacephalum

Clade : **Archelosauria**

10,514 conserved HOGs

Number of genes : **21,371**



Hand-on exercices

Gitpod

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

https://tinyurl.com/BGAOMA