wellcome connecting science

Protein k-mers for protein family assignment and downstream applications

Dr Yannis Nevers
ICube laboratory - CSTB team
Université de Strasbourg

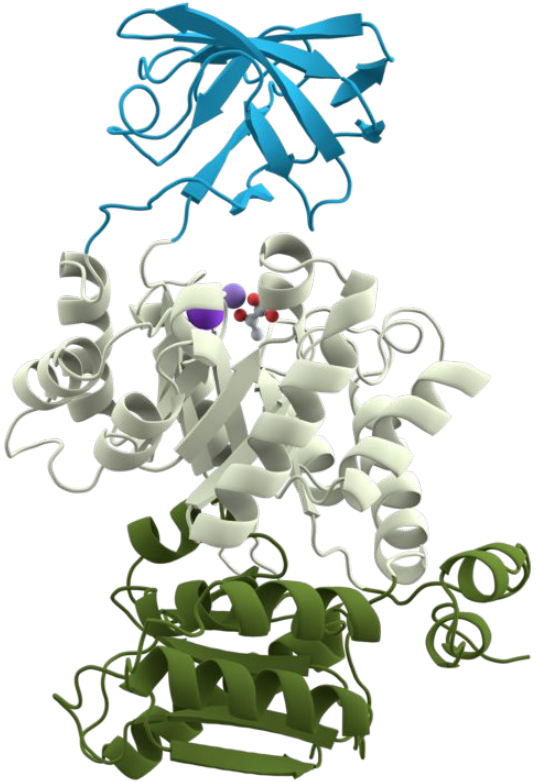
June 2025

Protein sequences

Protein :
Macromolecule composed of a chain of amino-acid residues.

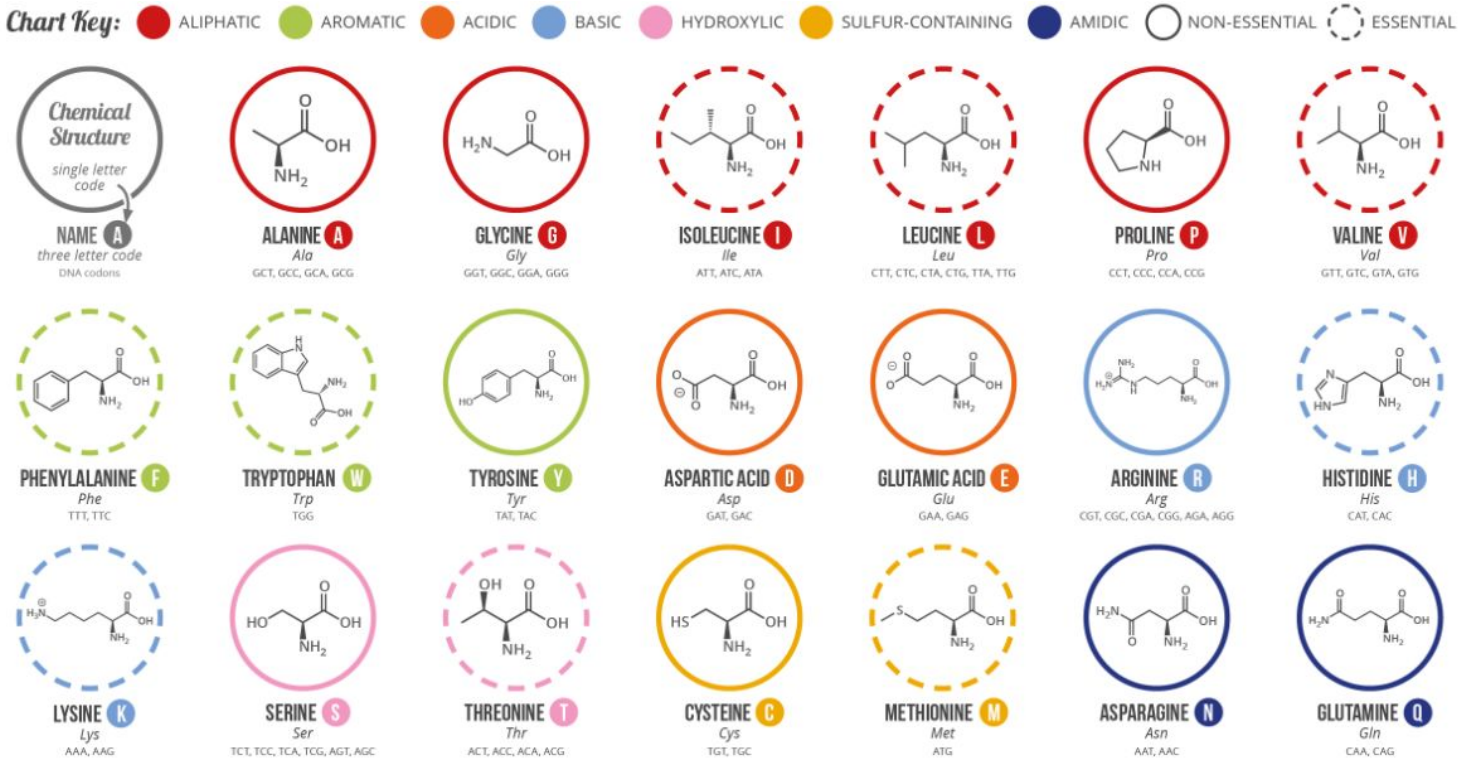
Its sequence is defined by the **coding sequence (CDS)** of a gene, encoded through the **genetic code.**

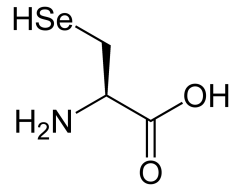
A protein sequence dictate its 3D structure and is often divided in **domains**



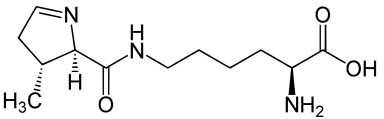
Amino-acid alphabet

- 20 canonical amino-acids + 2 non-canonical ones





Selenocysteine U

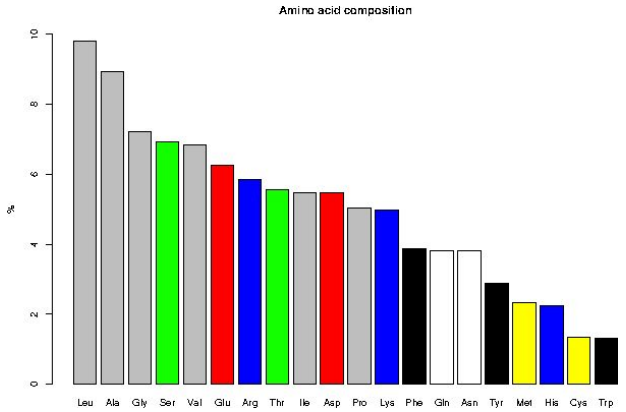


Pyrrolysine O

Image from Compound Interests - https://www.compoundchem.com/2014/09/16/aminoacids/

Amino acid alphabet

- 20 canonical amino-acids -> 20 lettere
- Degenerate genetic code :
  64 nucleotide triplets for 20 amino-acids
- Amino-acids have heterogeneous frequency

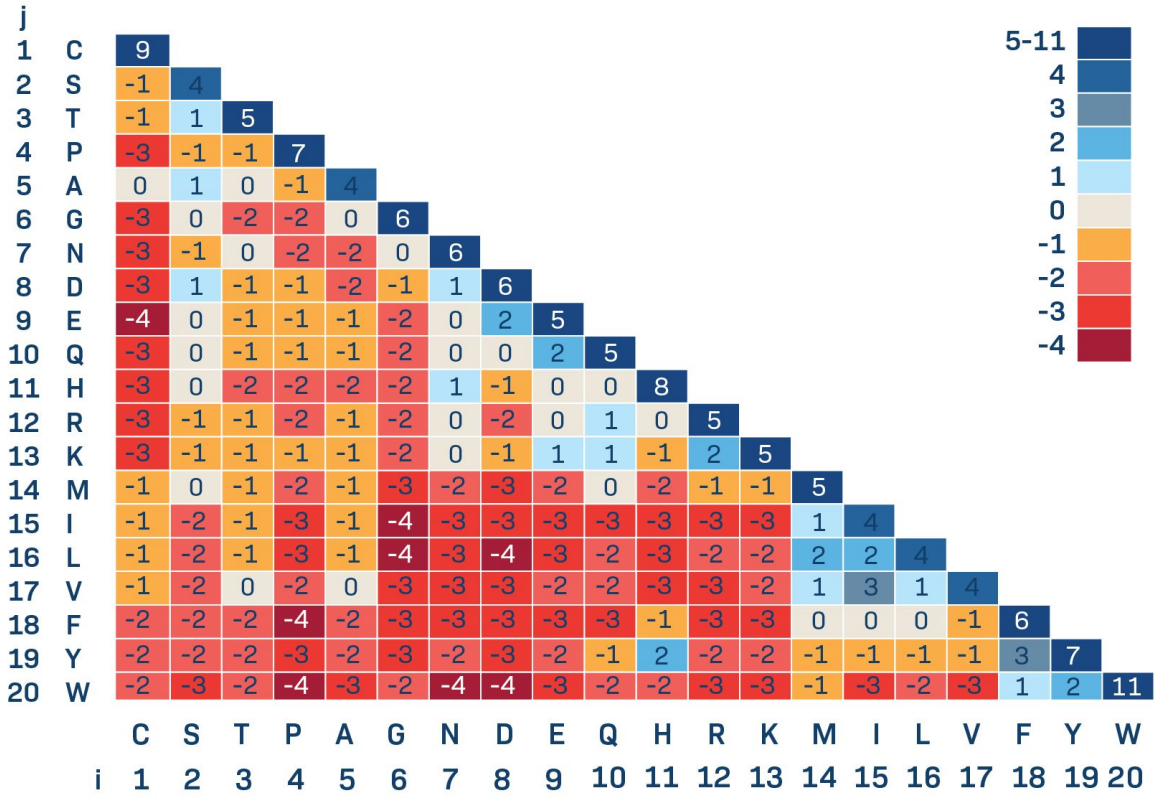


first base in codon	second base in codon: T	second base in codon: C	second base in codon: A	second base in codon: G	third base in codon
T	TTT Phe	TCT Ser	TAT Tyr	TGT Cys	T
	TTC Phe	TCC Ser	TAC Tyr	TGC Cys	C
	TTA Leu	TCA Ser	TAA stop	TGA stop	A
	TTG Leu	TCG Ser	TAG stop	TGG Trp	G
C	CTT Leu	CCT Pro	CAT His	CGT Arg	T
	CTC Leu	CCC Pro	CAC His	CGC Arg	C
	CTA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CTG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	ATT Ile	ACT Thr	AAT Asn	AGT Ser	T
	ATC Ile	ACC Thr	AAC Asn	AGC Ser	C
	ATA Ile	ACA Thr	AAA Lys	AGA Arg	A
	ATG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GTT Val	GCT Ala	GAT Asp	GGT Gly	T
	GTC Val	GCC Ala	GAC Asp	GGC Gly	C
	GTA Val	GCA Ala	GAA Glu	GGA Gly	A
	GTG Val	GCG Ala	GAG Glu	GGG Gly	G

Source: https://www.chemguide.co.uk

Amino acid alphabet

- Amino-acids are not equally replaceable across evolution



BLOSUM62 substitution matrix

(Image from https://www.labxchange.org/library/items/lb:LabXchange:20ec21:lx_imag)

Usage of protein k-mers

MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALT

MVLSPAD
VLSPADK
LSPADKT
...

First used in “protein sequencing”

COMPROTEIN: A COMPUTER PROGRAM TO AID PRIMARY PROTEIN STRUCTURE DETERMINATION*

Margaret Oakley Dayhoff and Robert S. Ledley
National Biomedical Research Foundation
Silver Spring, Maryland

Usage of protein k-mers

MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALT

MVLSPAD
VLSPADK
LSPADKT
...

Alignement-free sequence comparisons

- Fast sequence similarity search
- Protein clustering
- Sequence similarity estimation

Metagenomic applications

- Species classification
- Open Reading Frame identification (in reads)

k in protein k-mers

k (protein)	Number of mers	Eq in nucleotide k
2	4,00E+02	5
3	8,00E+03	7
4	1,60E+05	9
5	3,20E+06	11
6	6,40E+07	13
7	1,28E+09	16
8	2,56E+10	18
9	5,12E+11	20
10	1,02E+13	22
11	2,05E+14	24
12	4,10E+15	26
13	8,19E+16	29
14	1,64E+18	31
15	3,28E+19	33

Number of possible kmers

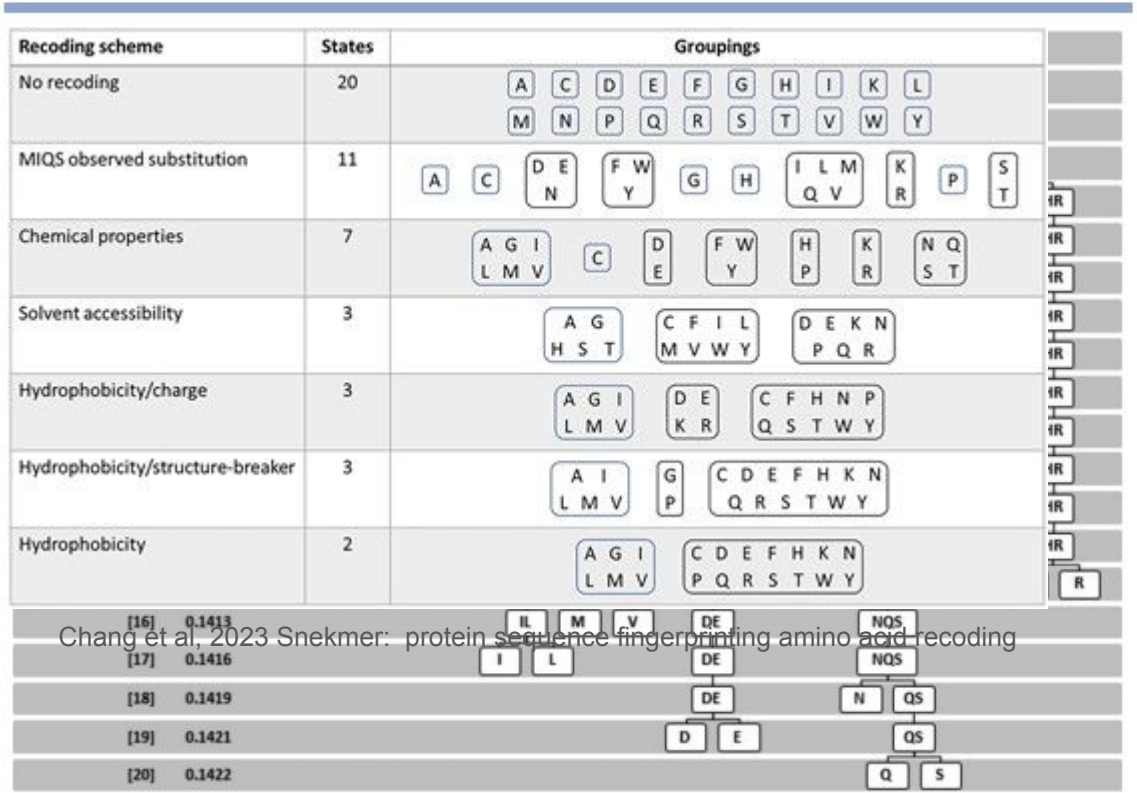
$$20^k \simeq 4^{2.16k}$$

More efficient than using CDS kmers (4^{3k})

One can use reduced alphabets to reduce it more…!

Reduced alphabet

- Reducing the amino-acid alphabet using physico-chemical properties
- Many reduction methods exist leading to different alphabet size
- Shown here: preserving long-range contact information in protein fold

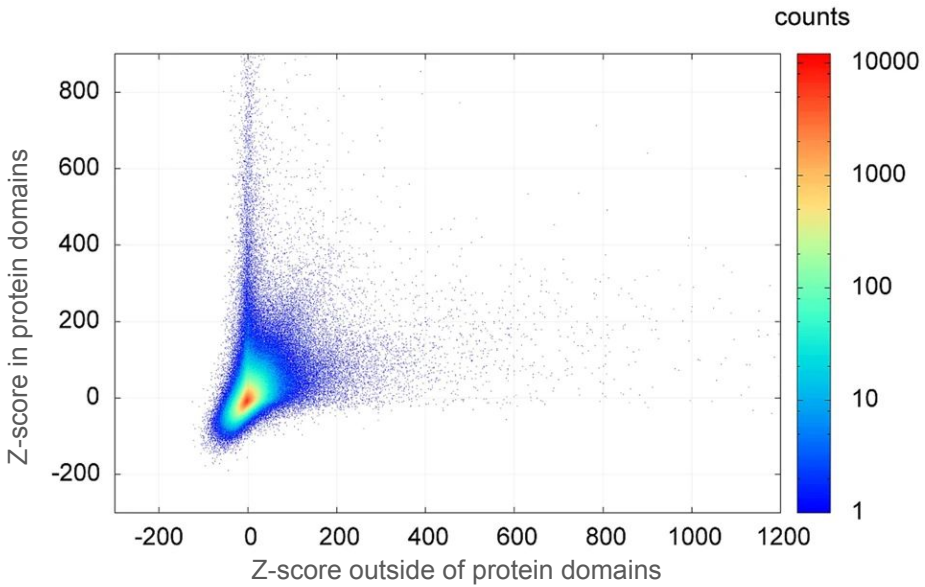


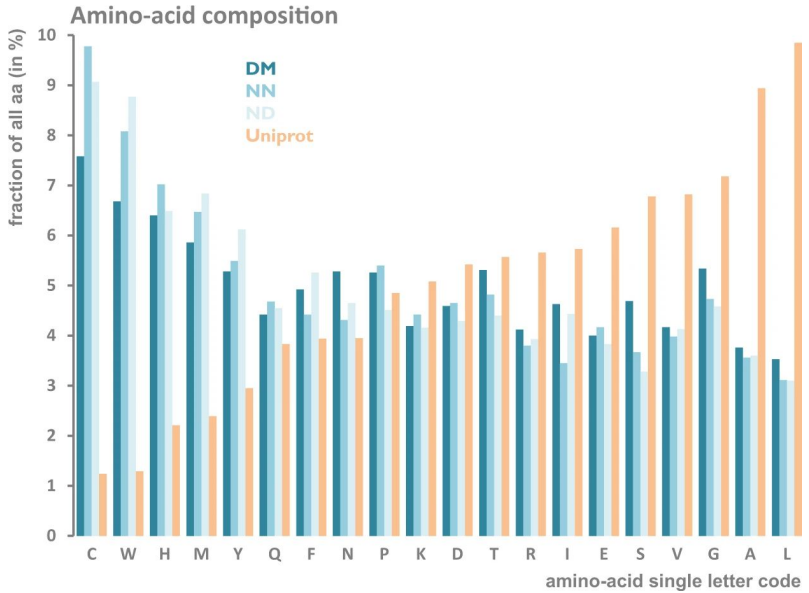
Chang et al, 2023 Snekmer: protein sequence fingerprinting amino acid recoding

Solis, 2015. Amino acid alphabet reduction contact interactions in proteins

Protein k-mer frequencies

- 5-mer frequencies are not only dependant of amino-acid frequency
- Related to protein domain organisation (Functional motifs)





From **Global pentapeptide statistics are far away from expected distributions**. Poznansky et al, 2018

Protein clustering

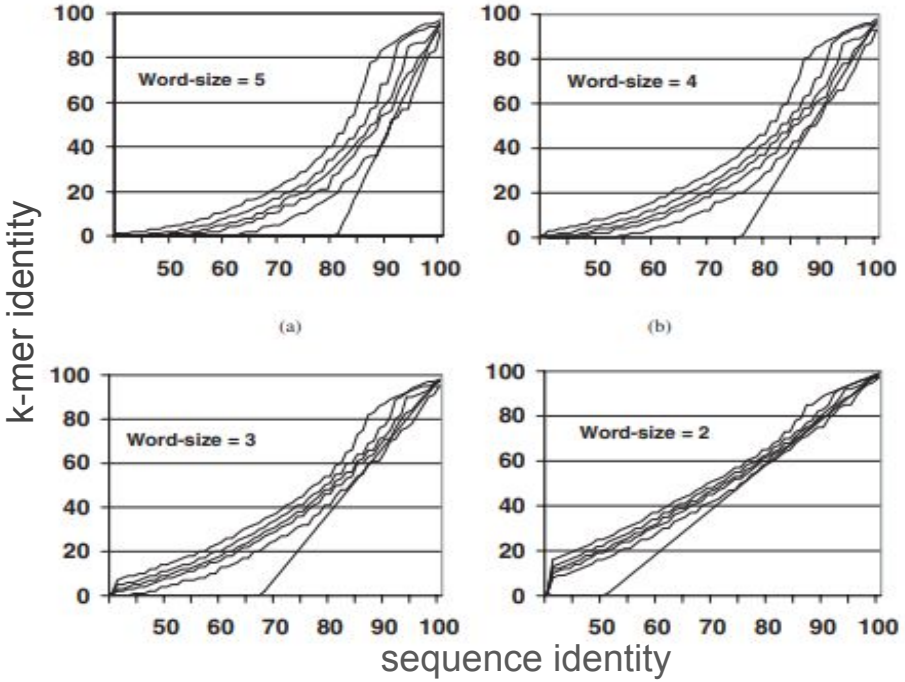
Goal : reduce redundancy in protein databases (e.g cluster proteins with >90% identity)

sequence identity should have at least a certain number of identical dipeptides, tripeptides and so on. For example, two sequences having 85% identical residues over a 100-residue window will have at least 70 identical dipeptides, 55 tripeptides and 25 pentapeptides. Therefore, pairs of sequences that don't satisfy these conditions don't have to be aligned, which allows speeding up the clustering of the database.

From **Clustering homologous sequences large databases.** Li et al, 2002

Strategy implemented in

- CD-Hit (Fu et al, 2012)
- UCLUST (Edgar,2010)



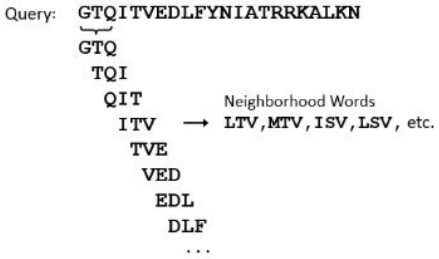
From **Tolerating redundancy speeds up clustering of large databases.** Li et al, 2002

K-mers in sequence similarity search

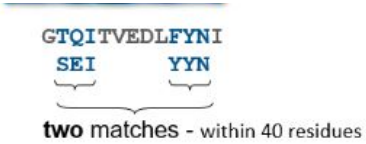
Goal : identify homologs and measure sequence similarity (with or without alignment)

Example : **BLASTP**

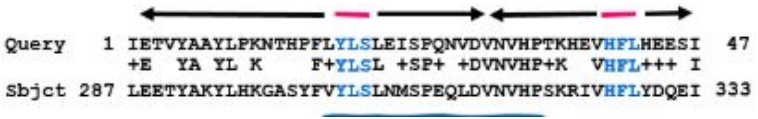
Decomposition into words



Words finding



Match extension for alignment

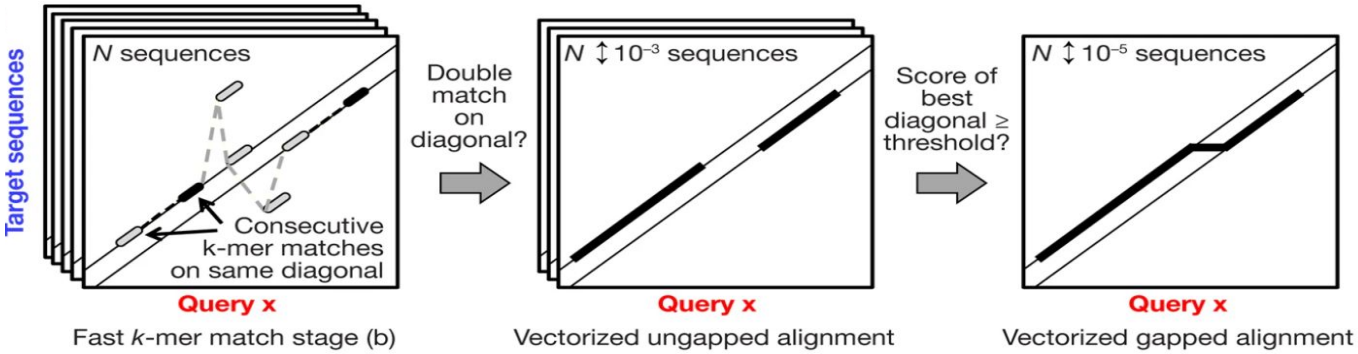


Basic Local Alignment Search Tools, Altschul et al, 1990
Illustration from https://www.nlm.nih.gov/ncbi/workshops/2022-10_Basic-Web-BLAST/how-blast-works.html

K-mers in sequence similarity search

Goal : identify homologs and measure similarity (with or without alignment)

MMSeqs 2



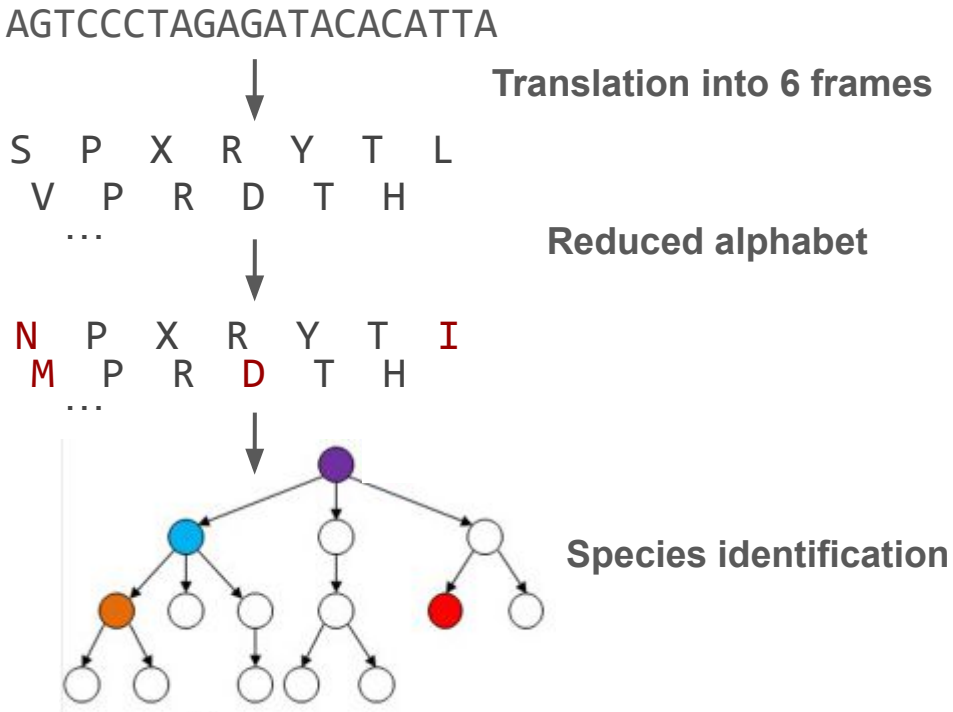
From **MMseqs2 sensitive protein sequence searching massive data sets**, Steinegger and Söding2017

Kmers in sequence similarity search

Tool	Type	Alignment	Default *k*	Reference
BLASTP	Similar words	Yes	5	(Altschul et al, 1990)
MMSeqs2	Similar words	Yes	7	(Steinegger et al, 2017)
USEARCH	Exact match	Yes	5	(Edgar, 2010)
RAPSearch	Exact match - Reduced alphabet	Yes	6	(Zhao et al, 2012)
DIAMOND	Exact match - Reduced alphabet - Spaced words	Yes	12 (15-24)*	(Buchfink et al, 2015)
Simrank	Exact match	No	7	(DeSantis et al, 2011)
kaamer	Exact match	No	7	(Deraspe et al, 2022)

Protein k-mers in metagenomics

Species identification in reads

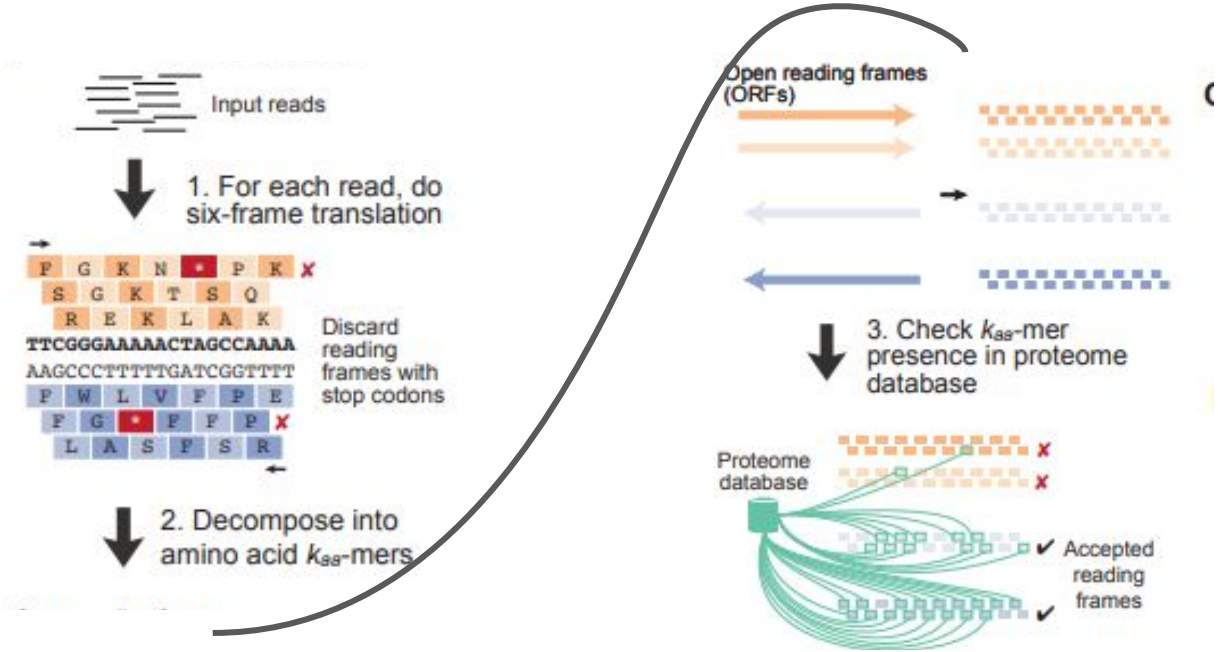


- Increase speed and reduce memory usage relative to nucleotide
- Implemented in KRAKEN2X (Wood et al, 2019)

Adapted from Wood et Salzberg, 2014. **Kraken**.

K-mers for protein sequence prediction

ORF Finding in RNA-seqs reads - Orpheum



Preprint : **Single-cell transcriptomics for the 99.9% of species without reference genomes,** Botvinnik et al, 2021

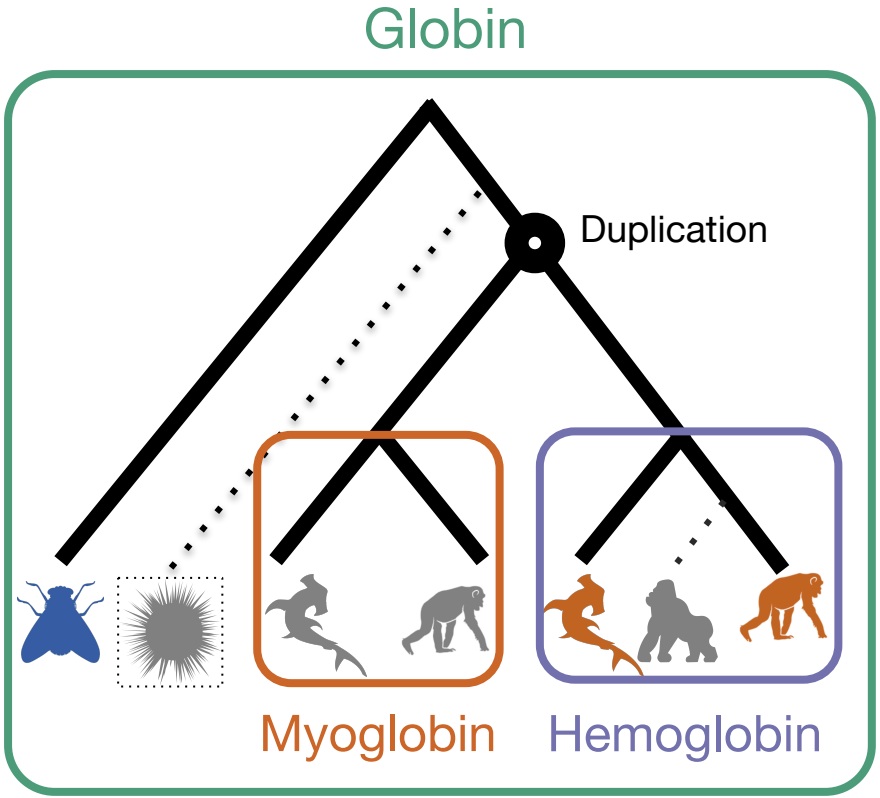
OMAmer

K-mer based fast protein placement into gene families

OMAmer

- Task: placing proteins into gene families and subfamilies
- Based only on k-mer content comparisons
- Relies on an appropriate data structure : Hierarchical Orthologous Groups (HOGs)

Victor Rossier



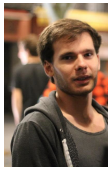
The OMA database

OMA : Orthology MAtrix

- Sequence-similarity based orthology inference algorithm
- Precomputed orthology relationship

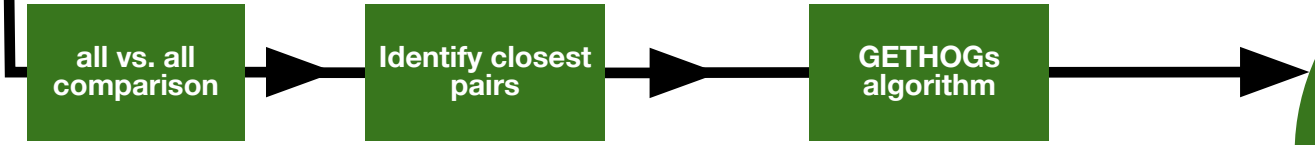

Christophe Dessimoz


Adrian Altenhoff


Clément Train


Natasha Glover

Inputs :
Protein sequences as FASTA file



Sequences of shared ancestry

Orthologous pairs (graph)

Hierarchical Orthologous Groups (HOGs)

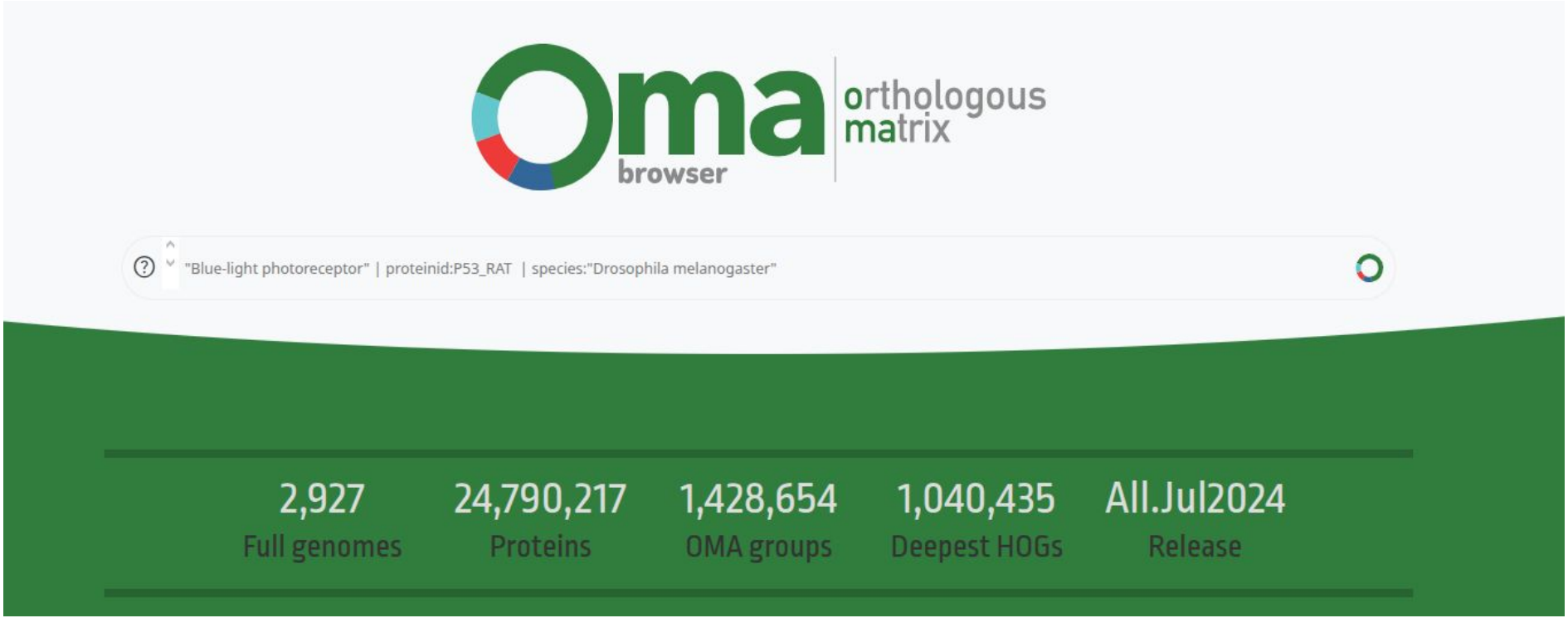
Protein k-mers

Oma orthologous matrix browser

https://omabrowser.org/oma/home/

Dessimoz et al., Nucl Acids Res 2006
Altenhoff et al, Nucl Acids Res 2011
Altenhoff et al. NAR 2024

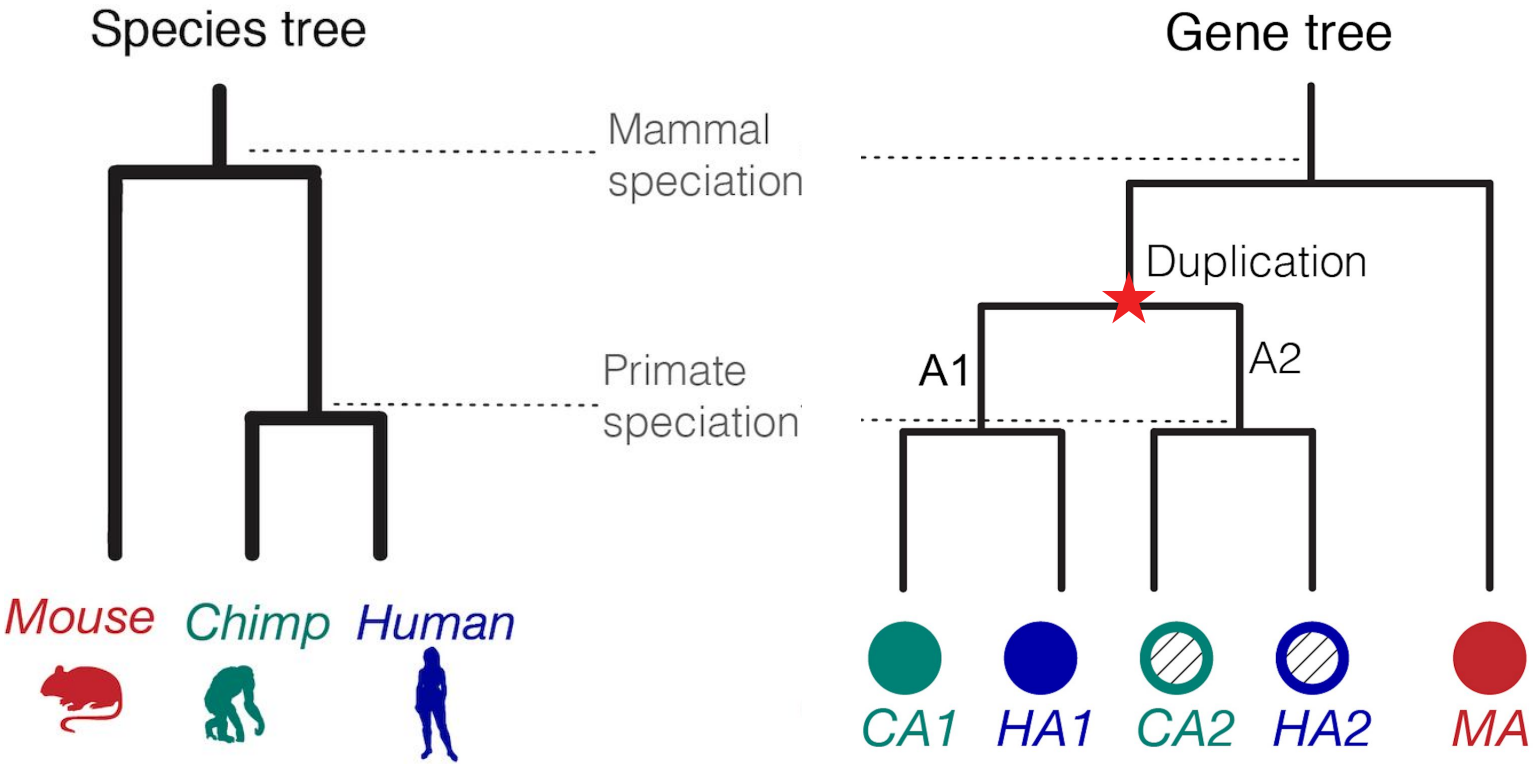
The OMA database



https://omabrowser.org

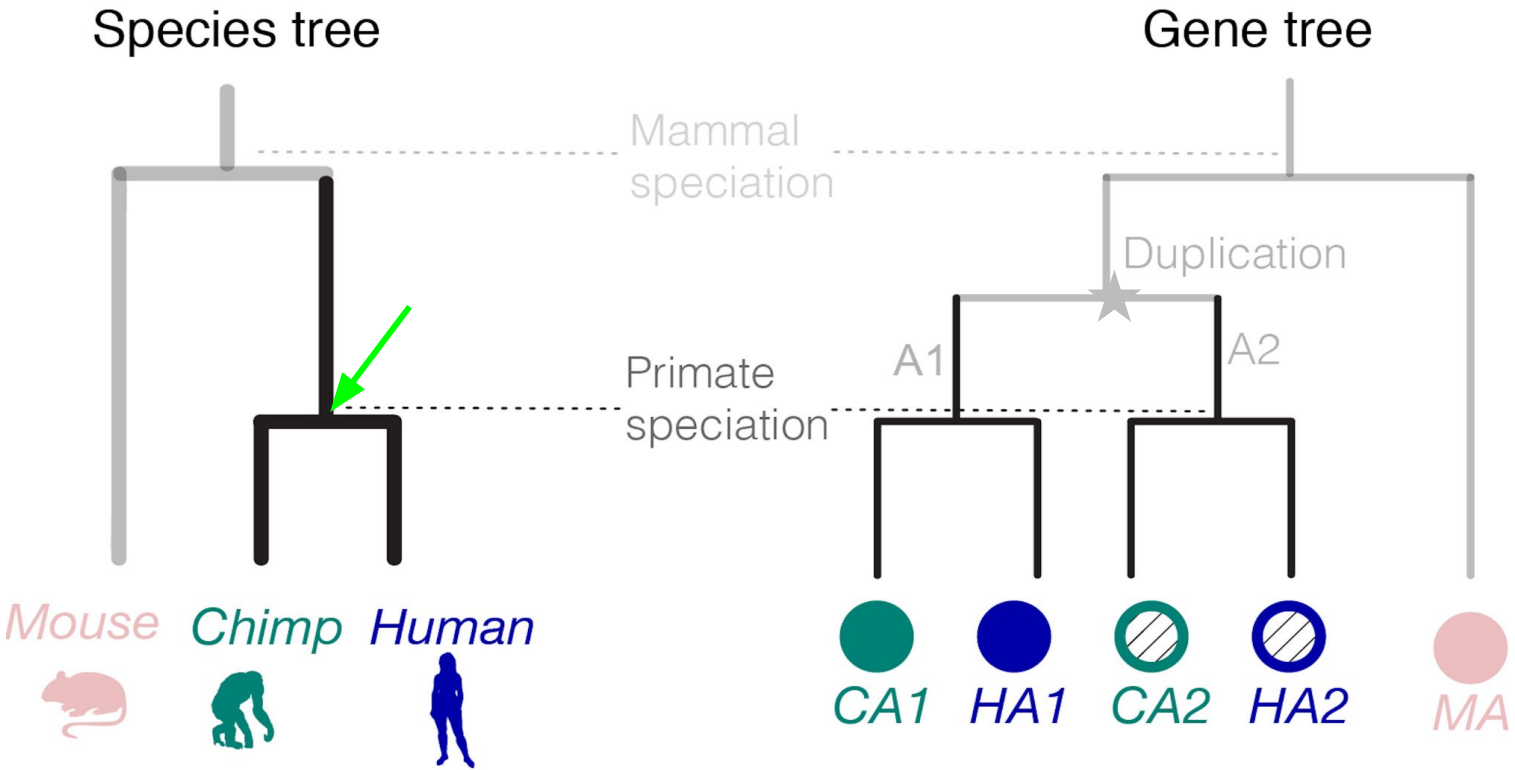
Hierarchical Orthologous Group

➢ Genes descended from a common ancestral gene at a specific taxonomic level



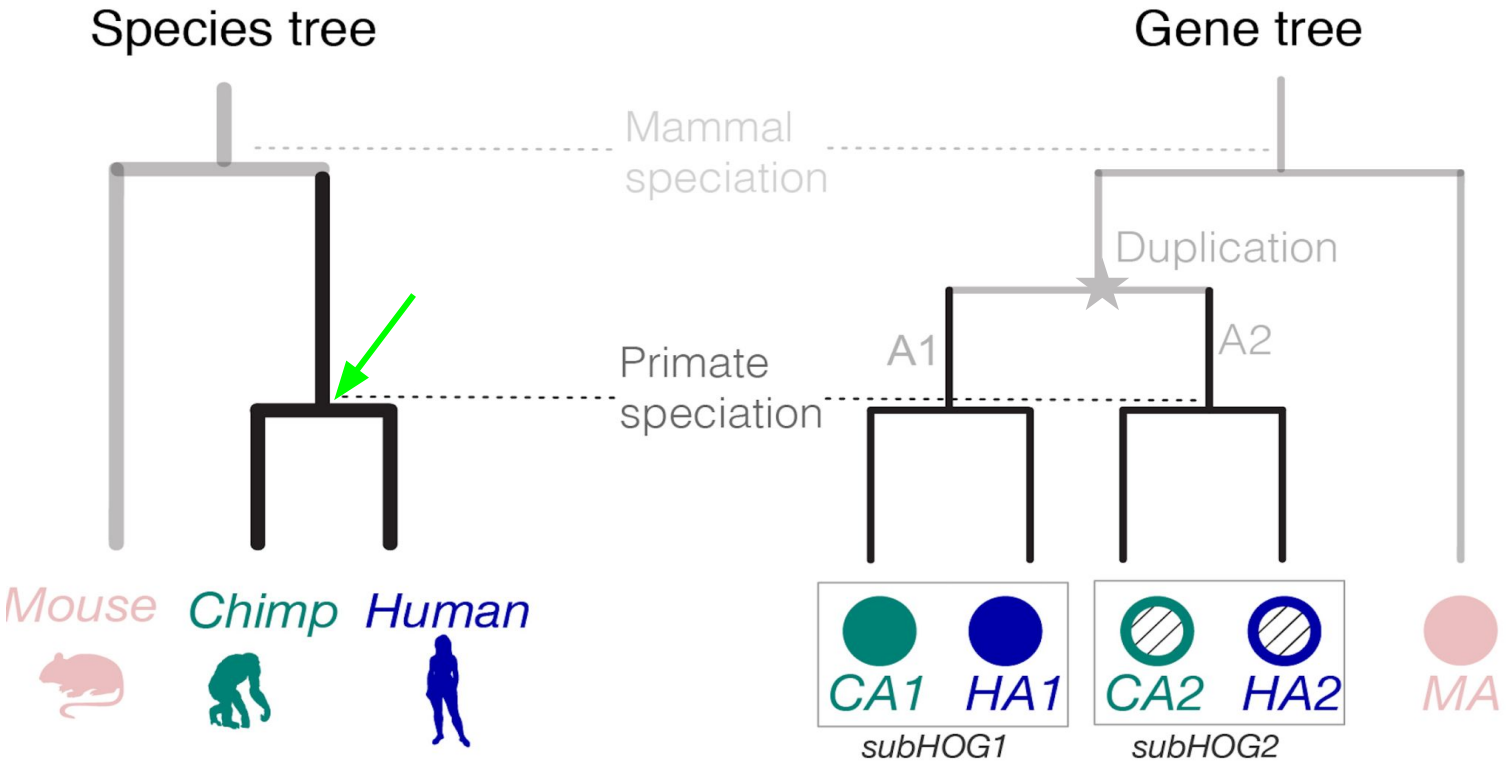
Hierarchical Orthologous Group

➢ Genes descended from a common ancestral gene at a specific taxonomic level



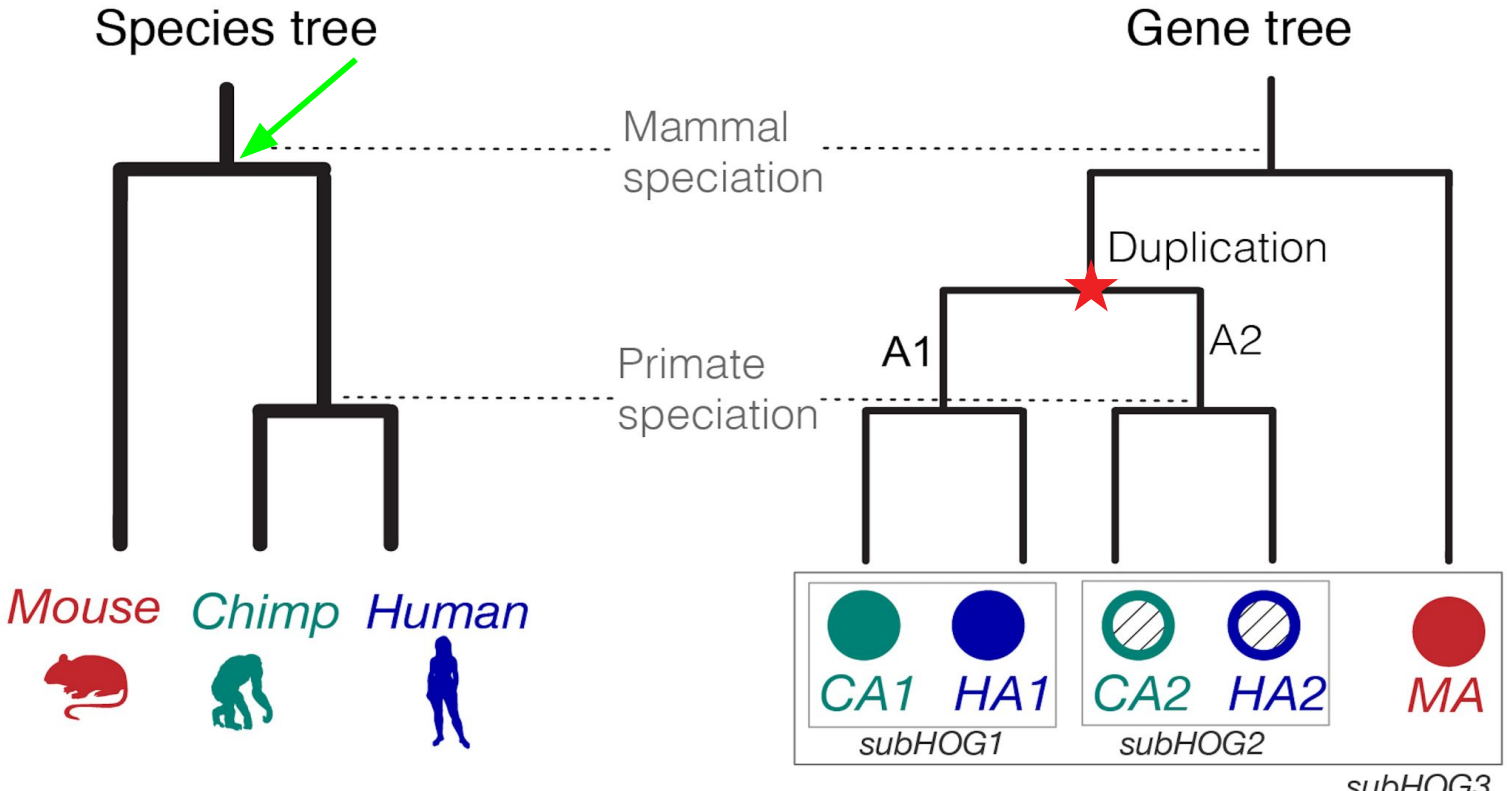
Hierarchical Orthologous Group

➢ Genes descended from a common ancestral gene at a specific taxonomic level

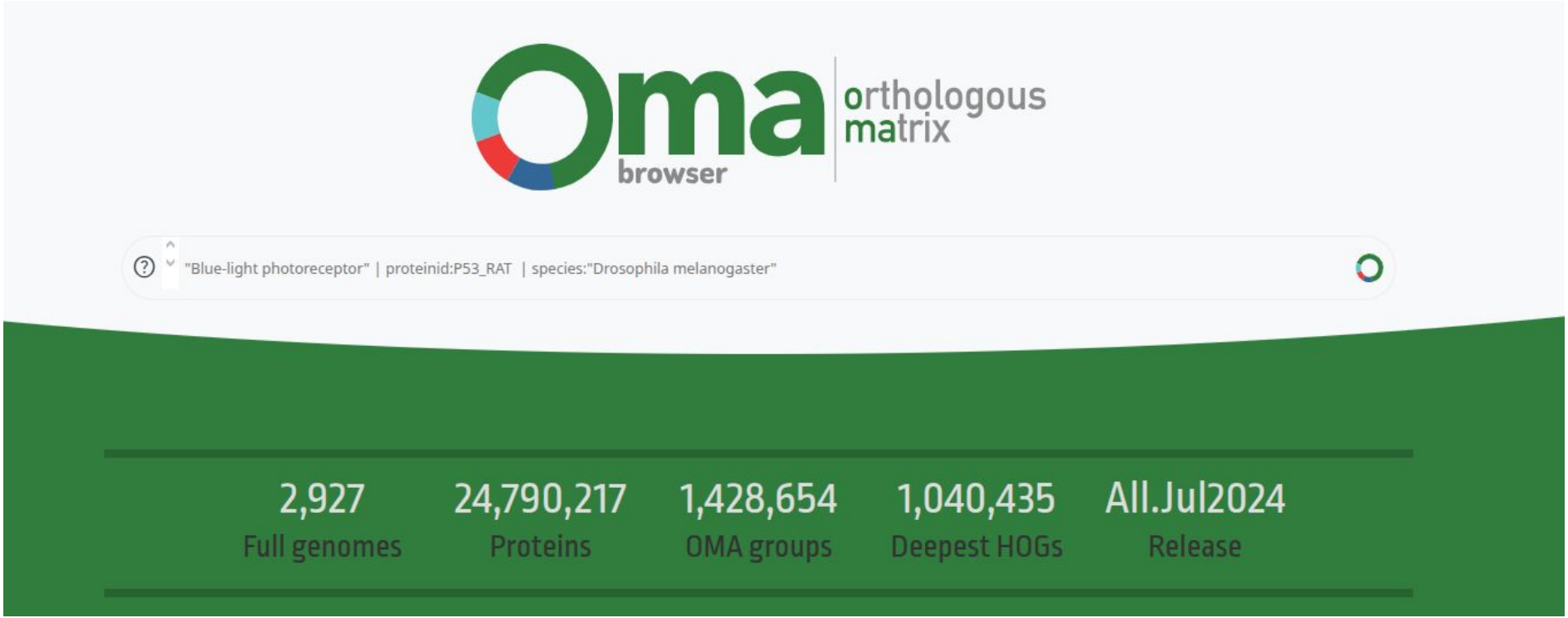


Hierarchical Orthologous Group

➢ Genes descended from a common ancestral gene at a specific taxonomic level



The OMA database



https://omabrowser.org

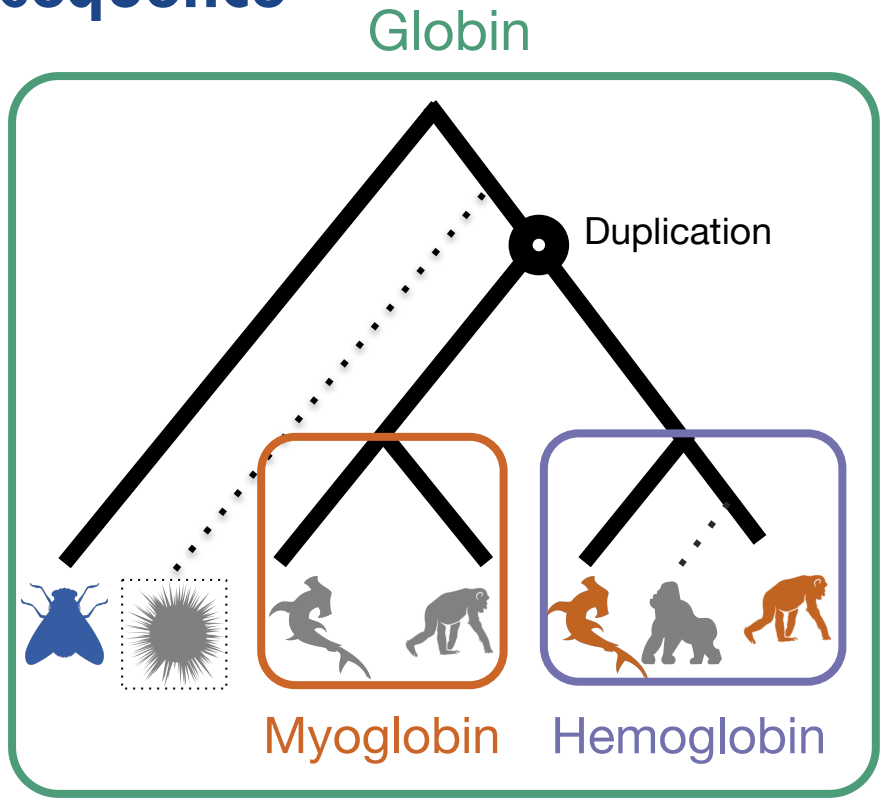
Family placement vs closest sequence

- Gene families reduce the number of comparisons

  $N = \text{query seq. nr.}$
  $M = \text{HOG. nr.}$

  - $N \cdot M$ is much lower than N^2
  - M grows slowly with new genomes

- The closest sequence does not always yield the correct subfamily !



HOGs and subHOGs k-mers

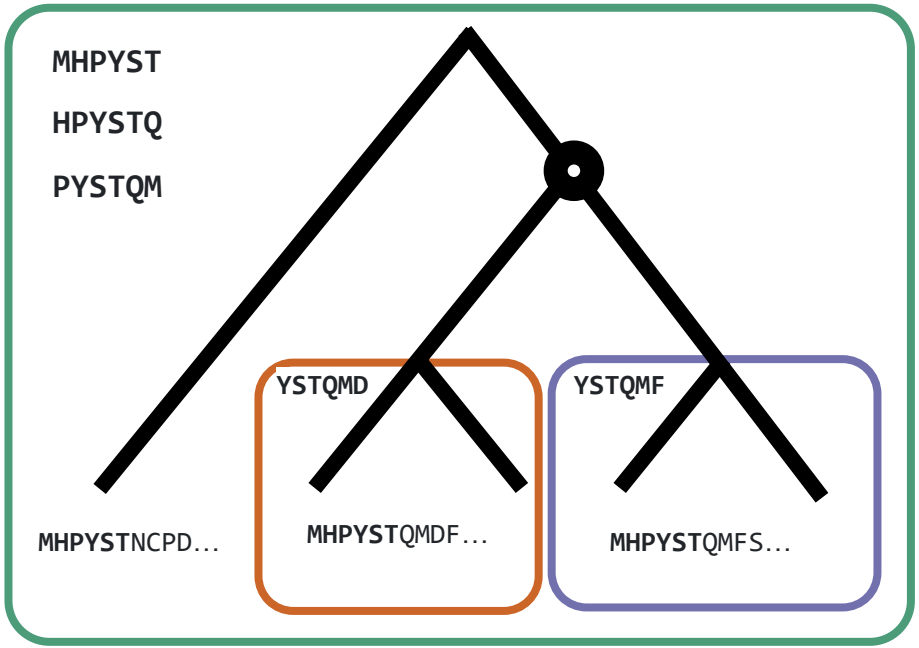
Query sequence

```
MHPYSTQMFS LQITVMEDSQ SDMSIELPLS

MHPYST
 HPYSTQ
  PYSTQM

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

HOG

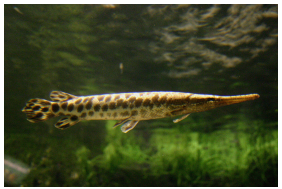


Process of HOG attribution

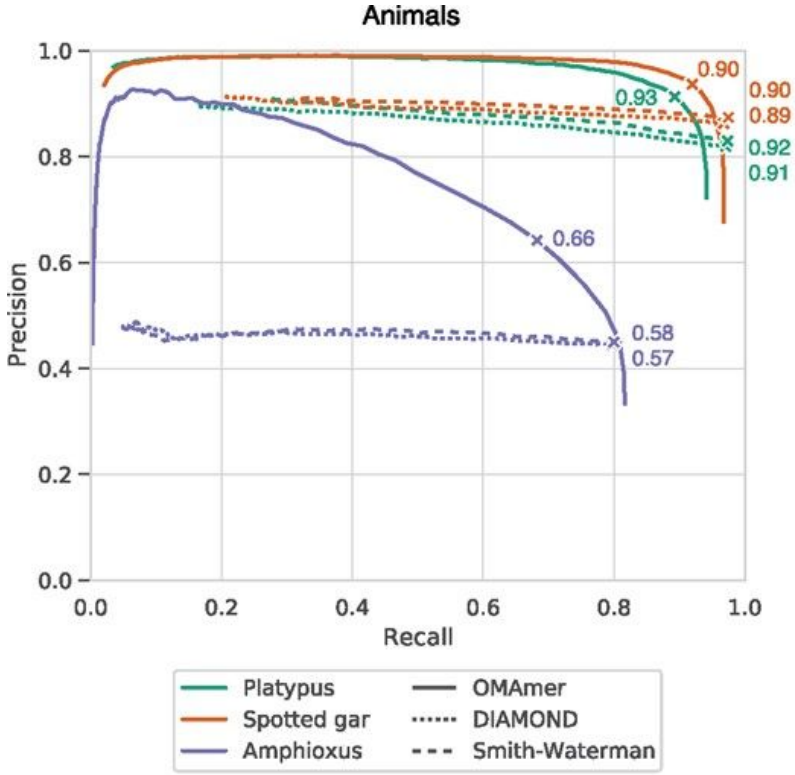
1. Find gene families (HOGs) sharing **kmers** with query
2. Filter by a “p-value” threshold **α**
3. Order by excess shared kmers over the expected (*normcount*)
4. Keep the best *n* results
5. Place into **subfamilies** of the selected HOGs - higher subfamily normcount
6. Return the placement

Results - subfamily placement

Benchmarking over 3 *outgroup* species


- Species removed from the database
- **Positive** if placed into a correct family and subfamily (not overspecific)



OMAmer subfamily placement is more accurate than using the closest sequence !

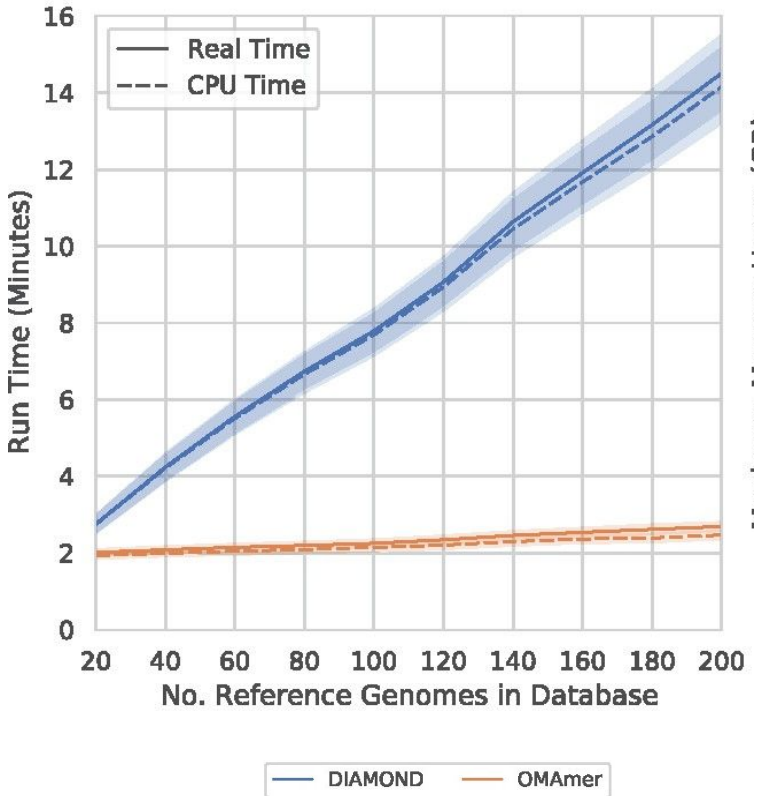
OMAmer performances

Runtime for placing all proteins from 20 metazoan proteomes relative to database size

OMAmer scales better than similarity search softwares !

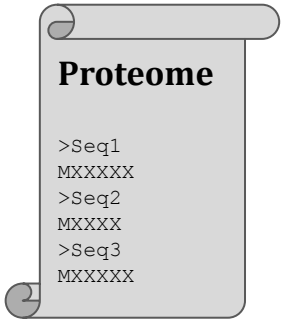
OMAmer 2.1.0 (1 thread)

20,000+ proteins placed in a database of 2,927 diverse genomes in **4:30** minutes using **13.3 Gb**



OMAmer usage

```
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

OMAmer 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

OMAmer references

Victor Rossier

Sequence analysis

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,*]

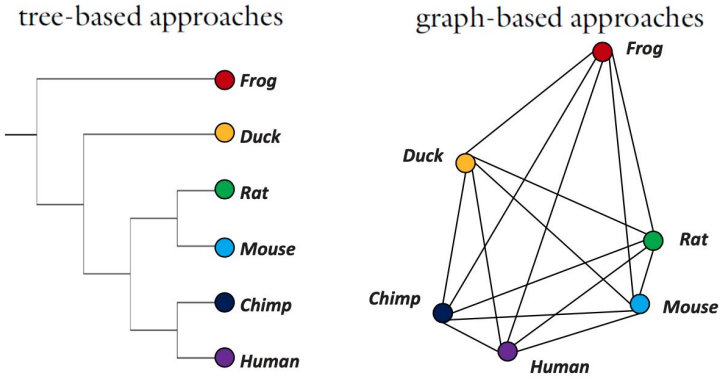
GitHub : https://github.com/DessimozLab/omamer

OMAmer use-case

Scalable orthology assignment with FastOMA

OMAmer use case: speeding up orthology inference

Specie number in orthology databases



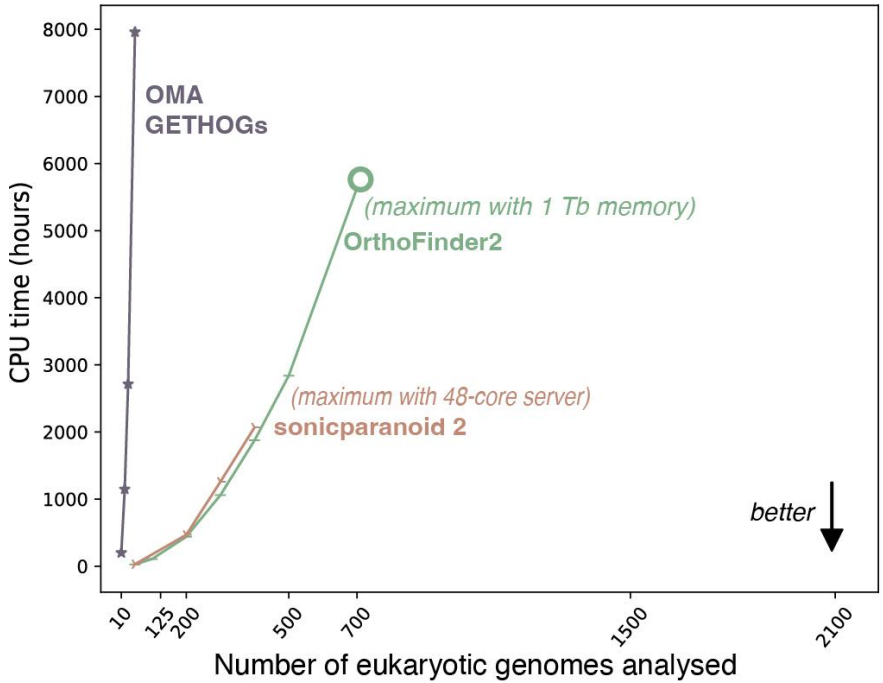
Method/resource	# species	
Oma orthologous matrix browser	2,851	Graph based
OrthoDB	20,110	Graph based
DB phylome	6,000	Tree based
NOG EggNOG 6.0.0	12,535	Graph + tree
OrthoFinder 2	-	Graph + tree

❌ All-vs-all comparison

R. Waterhouse. Biodiversity Bioinformatics. SIB course.
The Quest for Orthologs orthology benchmark service in 2022. NAR.

Bottleneck in orthology inference

- **Dataset** : UniProt reference eukaryotic proteomes

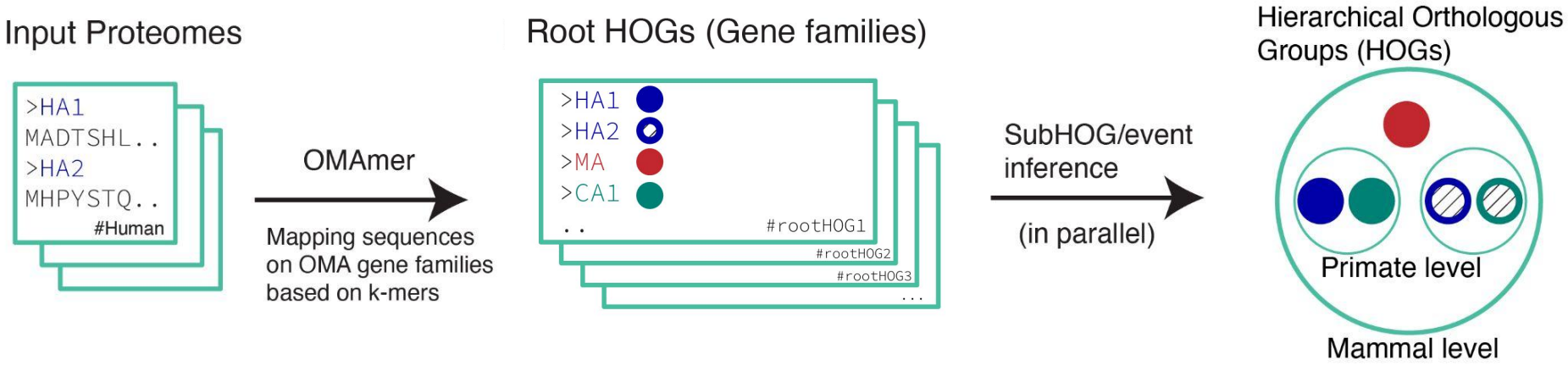


Existing methods have a hard time scaling to thousands of proteomes !

FastOMA


Sina Majidian

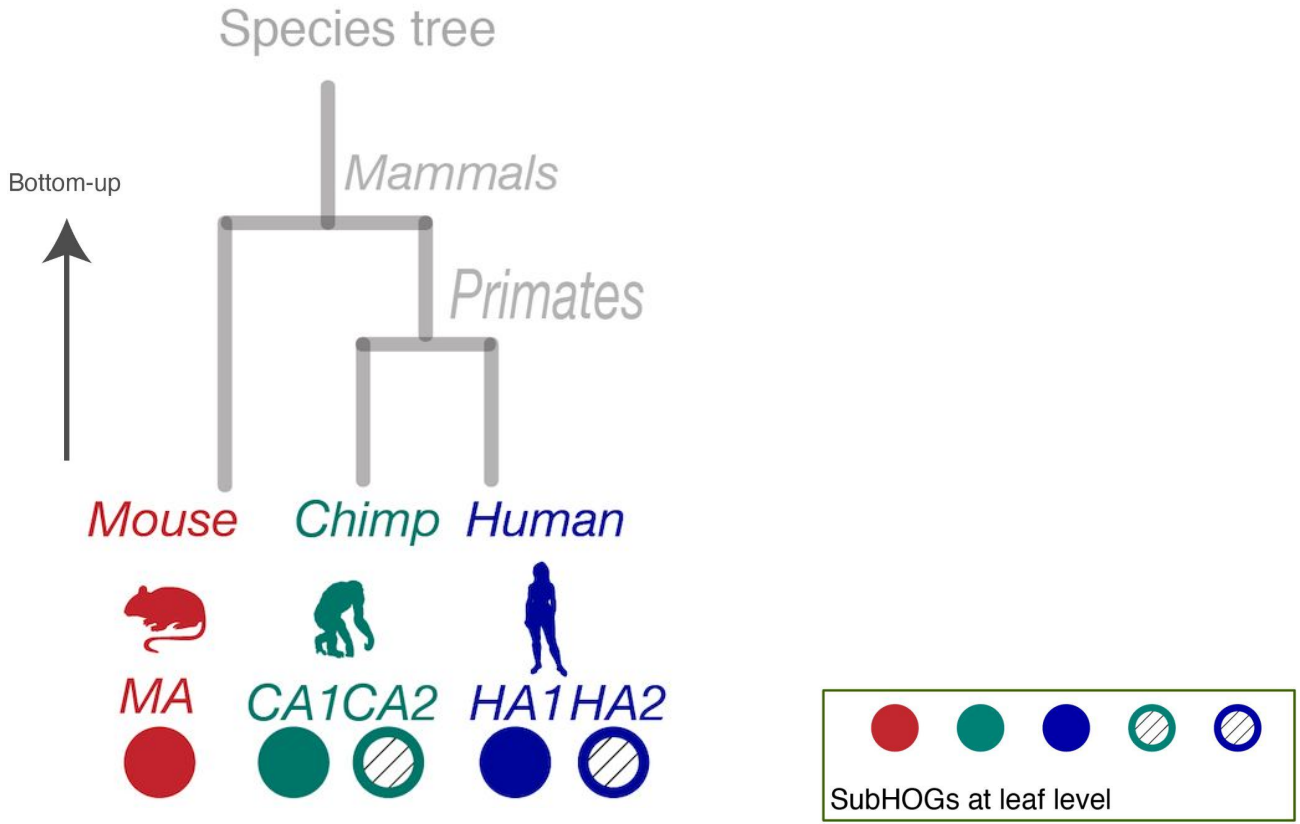
- OMAmer family placement is used as pre-assignment of orthologous groups



- All other computations are done within a HOG - no all-against-all

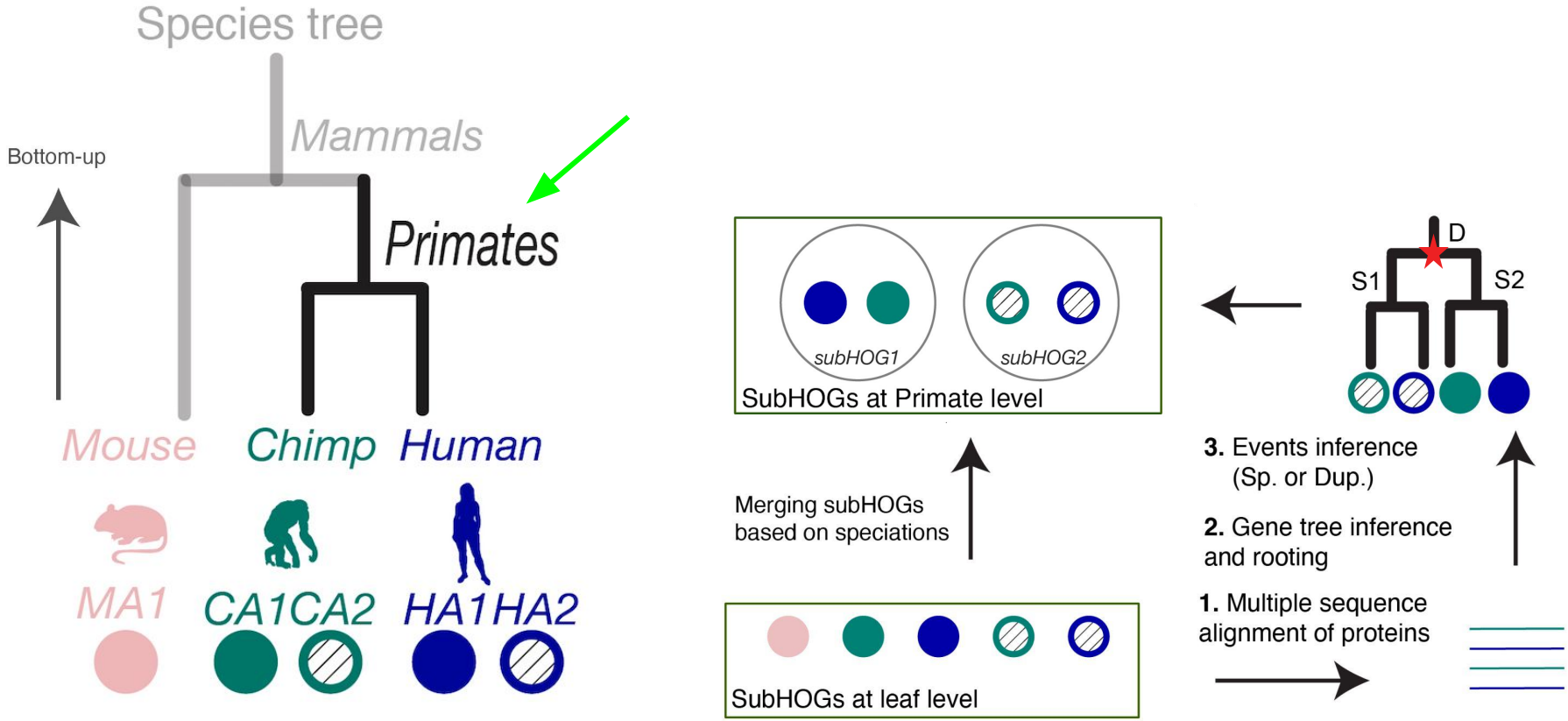
Majidian et al, Orthology inference at scale with FastOMA, 2025

FastOMA - Orthology inference



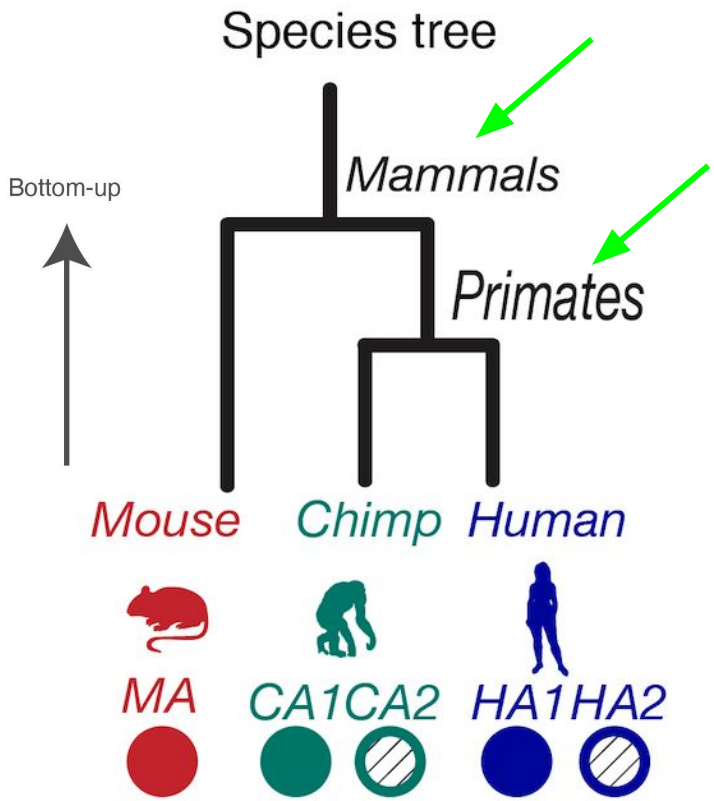
Majidian et al, Orthology inference at scale with FastOMA, 2025

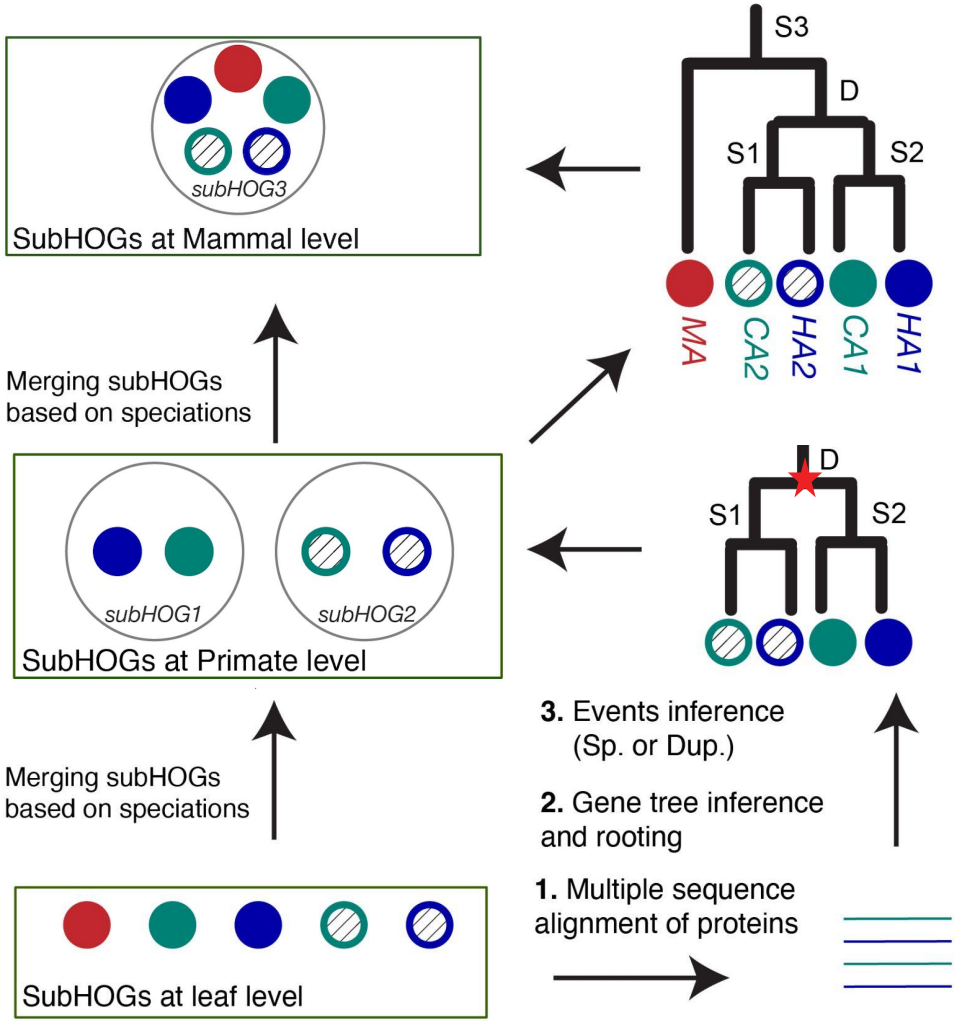
FastOMA - Orthology inference



Majidian et al, Orthology inference at scale with FastOMA, 2025

FastOMA

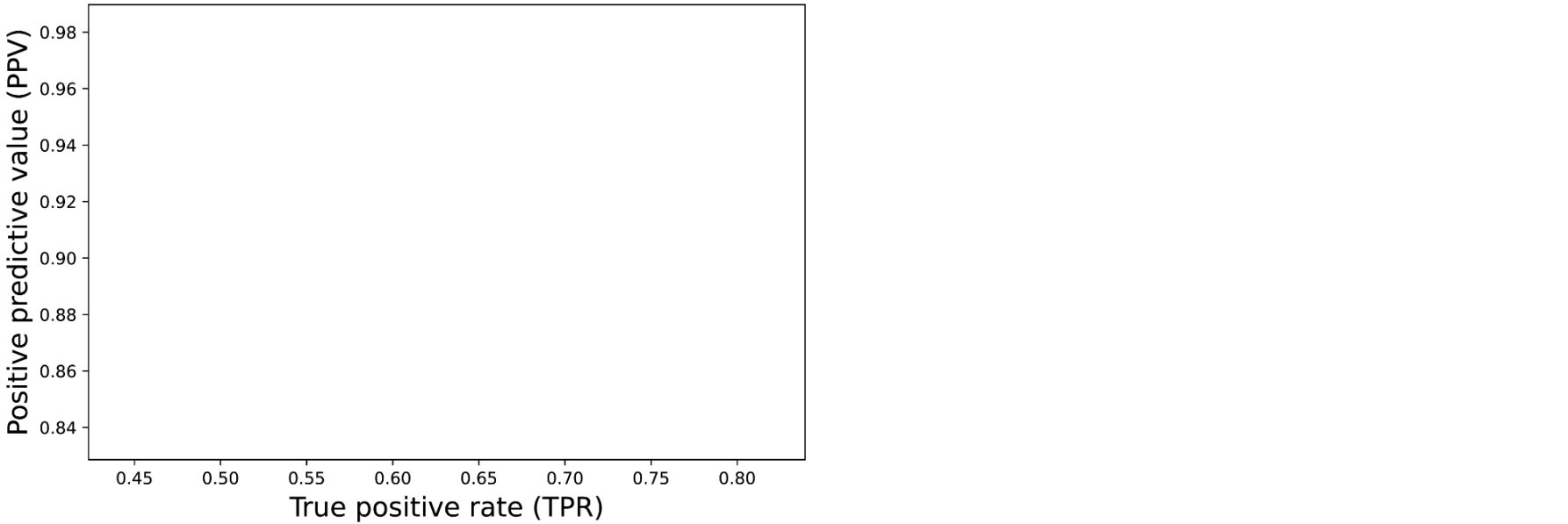




Majidian et al, Orthology inference at scale with FastOMA, 2025

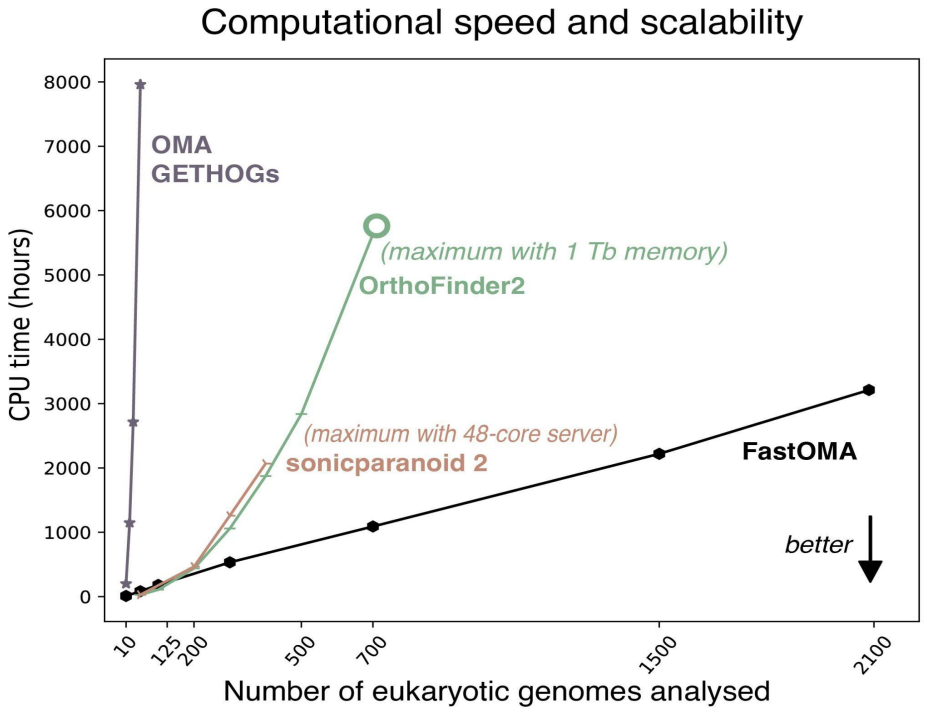
FastOMA orthology benchmarking

Agreement with reference phylogeny (SwissTrees)



FastOMA speed

- **Dataset** : UniProt reference eukaryotic proteomes
- In a single day using 300 CPUs



FastOMA references

Sina Majidian

Brief Communication

https://doi.org/10.1038/s41592-024-02552-8

Orthology inference at scale with FastOMA

Received: 16 January 2024

Accepted: 29 October 2024

Published online: 3 January 2025

Sina Majidian[1,2], Yannis Nevers[1,2], Ali Yazdizadeh Kharrazi[1], Alex Warwick Vesztrocy[1,2], Stefano Pascarelli[1,2], David Moi[1,2], Natasha Glover[1,2], Adrian M. Altenhoff[2,3] & Christophe Dessimoz[1,2]

GitHub : https://github.com/DessimozLab/FastOMA

OMAmer use-case

Homology-based quality control of structural gene annotation (proteome)

OMAmer use-case : proteome quality

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

How to use gene family assessment to detect these issues?

OMArk pipeline

Input

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

Proteome

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

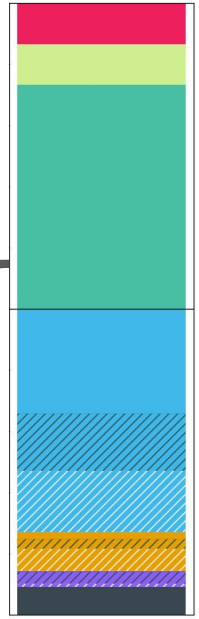
Step 1 : Gene family placement

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

Seq1	HOG:B0578800.1c.1d
Seq2	HOG:B0571029
Seq3	HOG:B0606120.3n

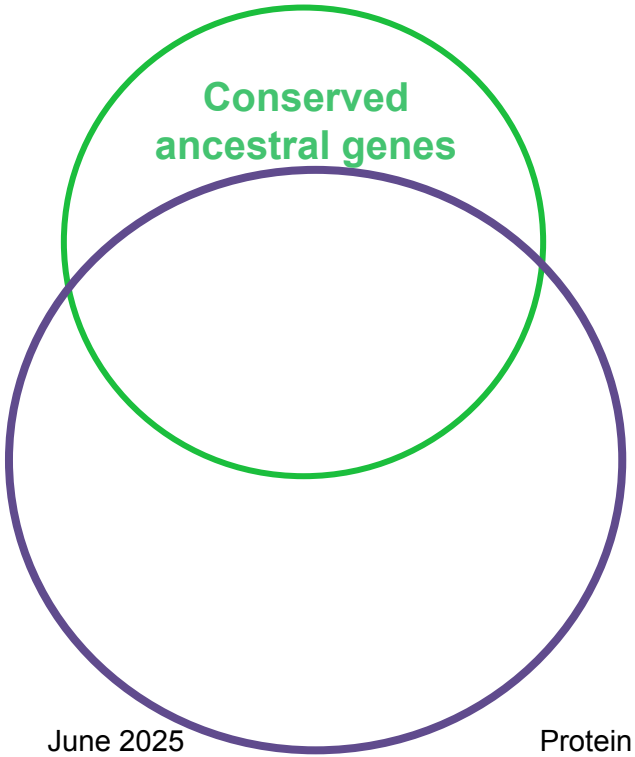
Step 2 : OMArk Quality assessment

QA Summary



Protein k-mers

Gene families for quality assessment



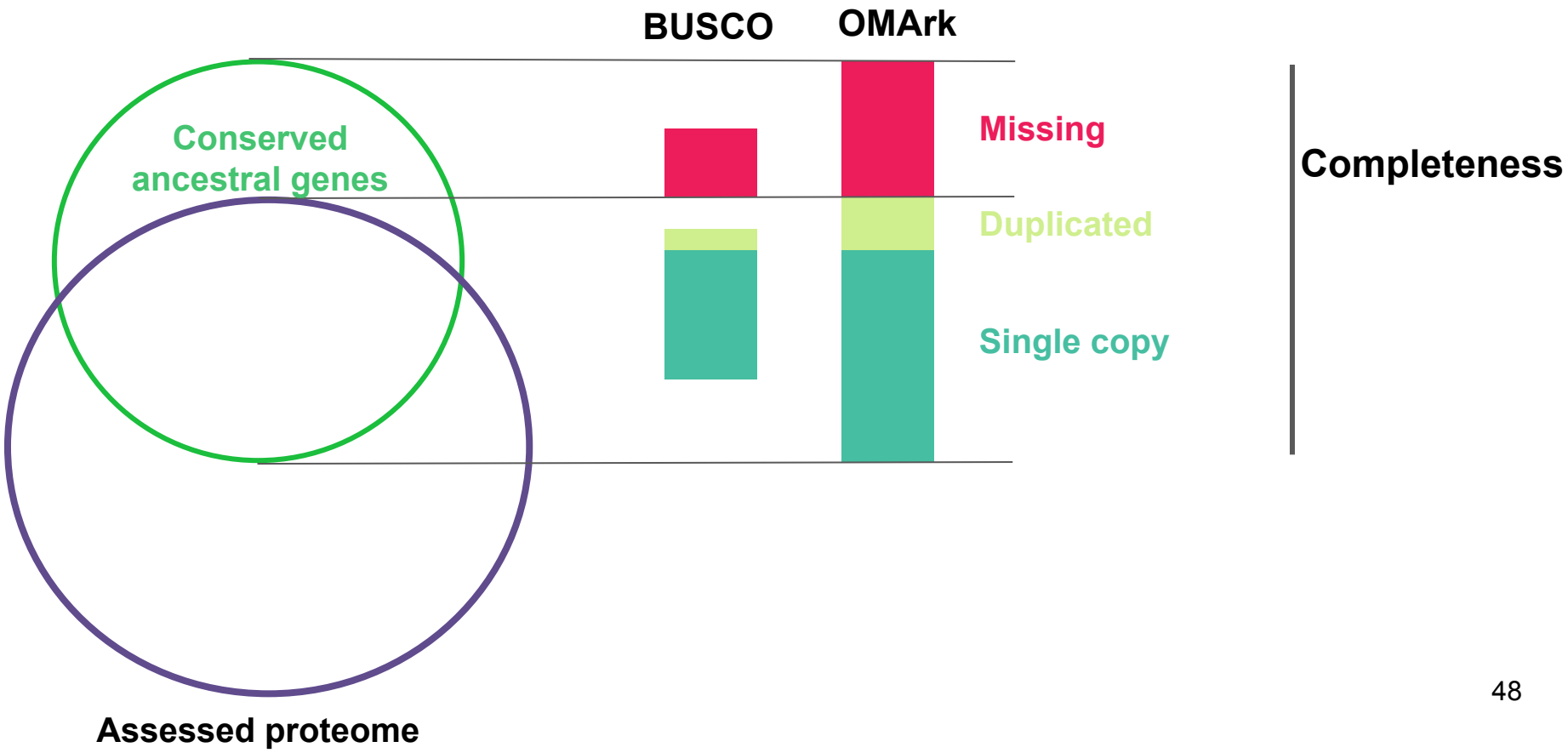
Ancestral lineage :

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

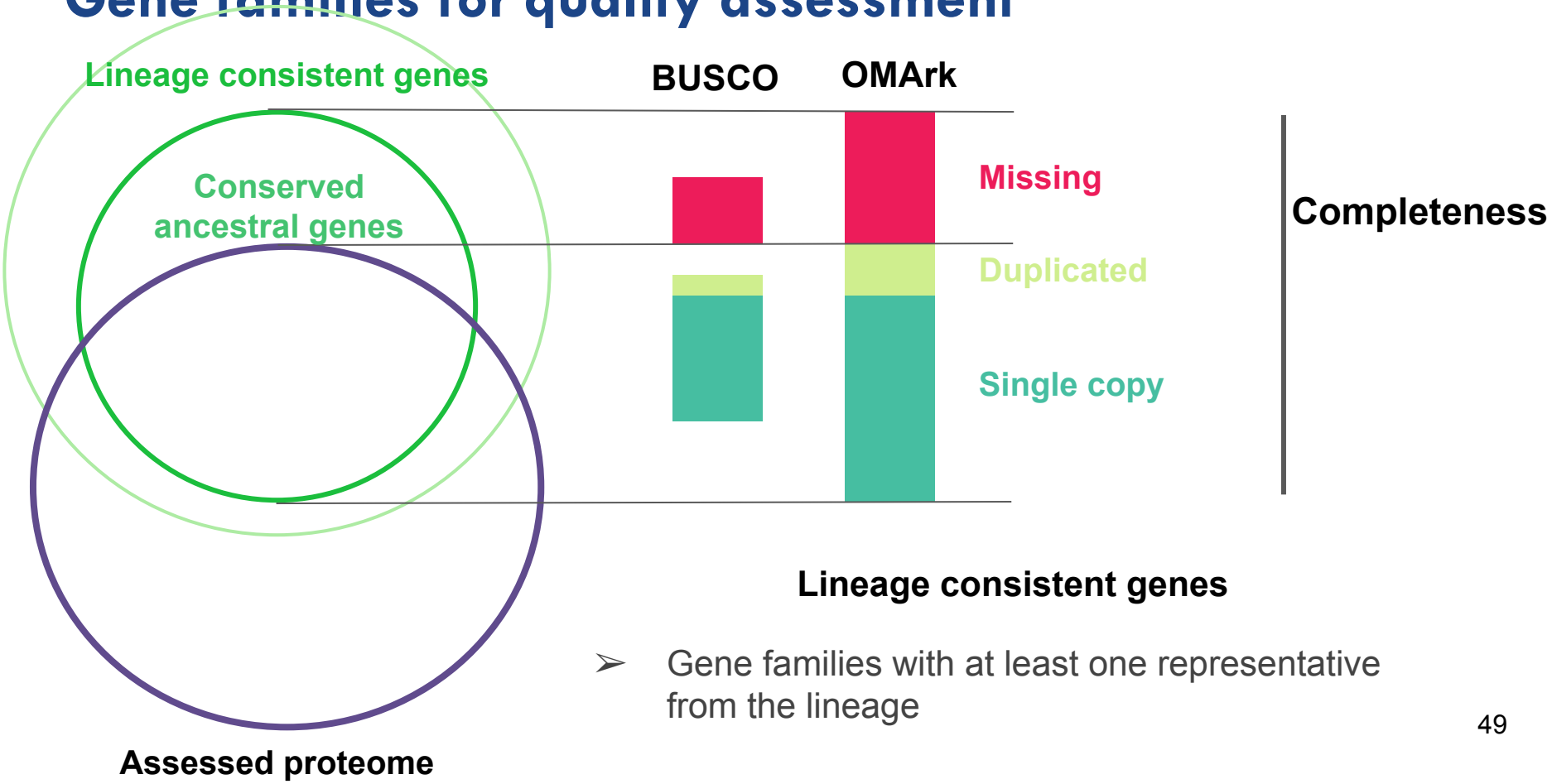
Conserved ancestral genes:

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

Gene families for quality assessment

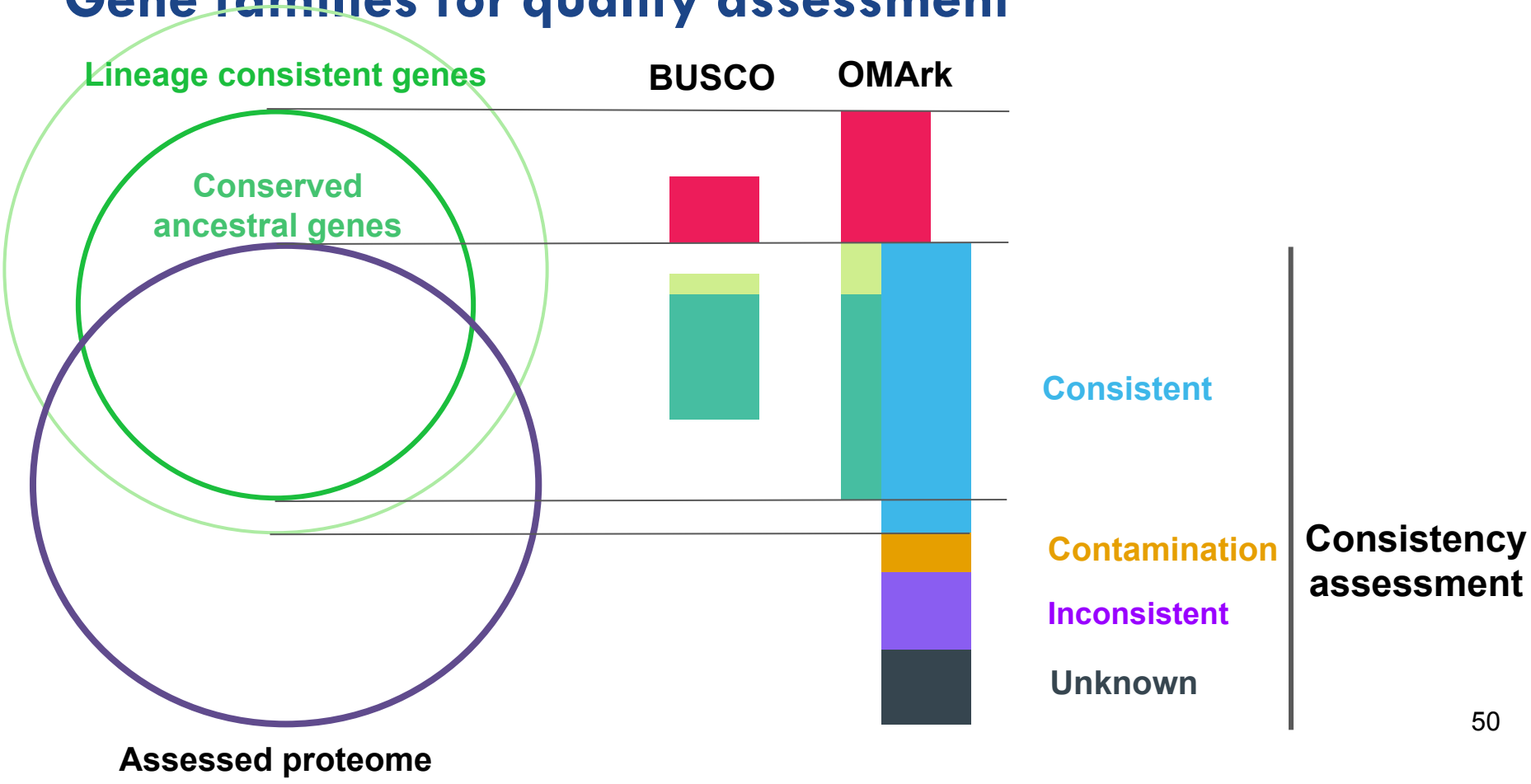


Gene families for quality assessment

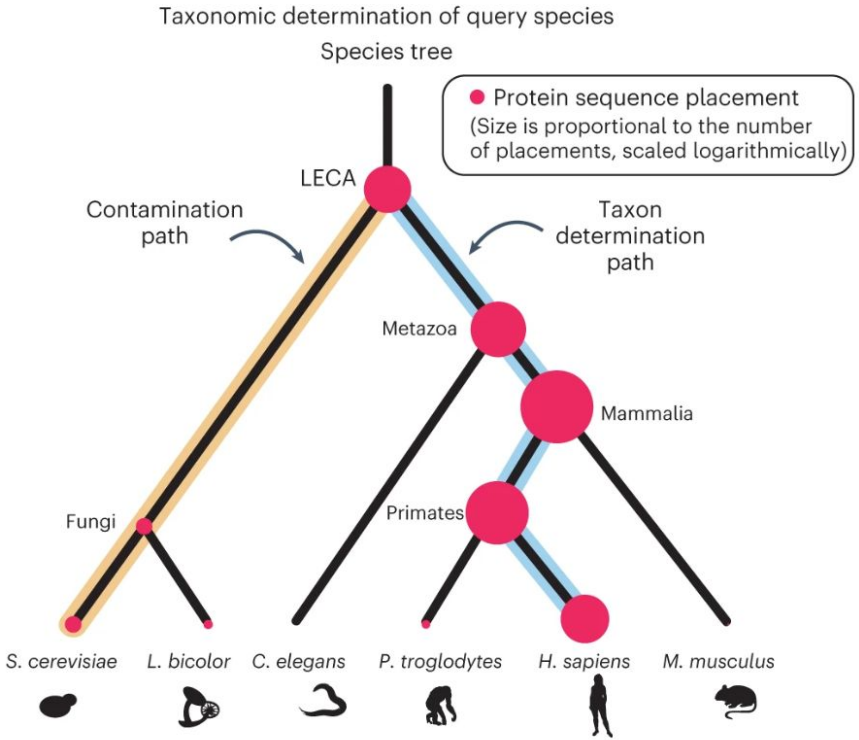


- Gene families with at least one representative from the lineage

Gene families for quality assessment



Species detection

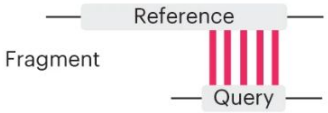


- HOG placement is associated to the taxon at which a subHOG is located
- This can be used to identify the species in the dataset
- Select the most specific clade in a branch with highest HOG representation
  - Control for database species imbalance
  - Control for duplication rate

Structural consistency

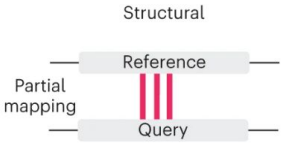
OMArk use database knowledge and kmer location information to evaluate gene structure

Fragment

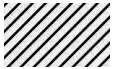


Query is less than half the median size of the gene family

Partial mapping



Matching k-mers are concentrated only to part of the query (20% or more not covered)



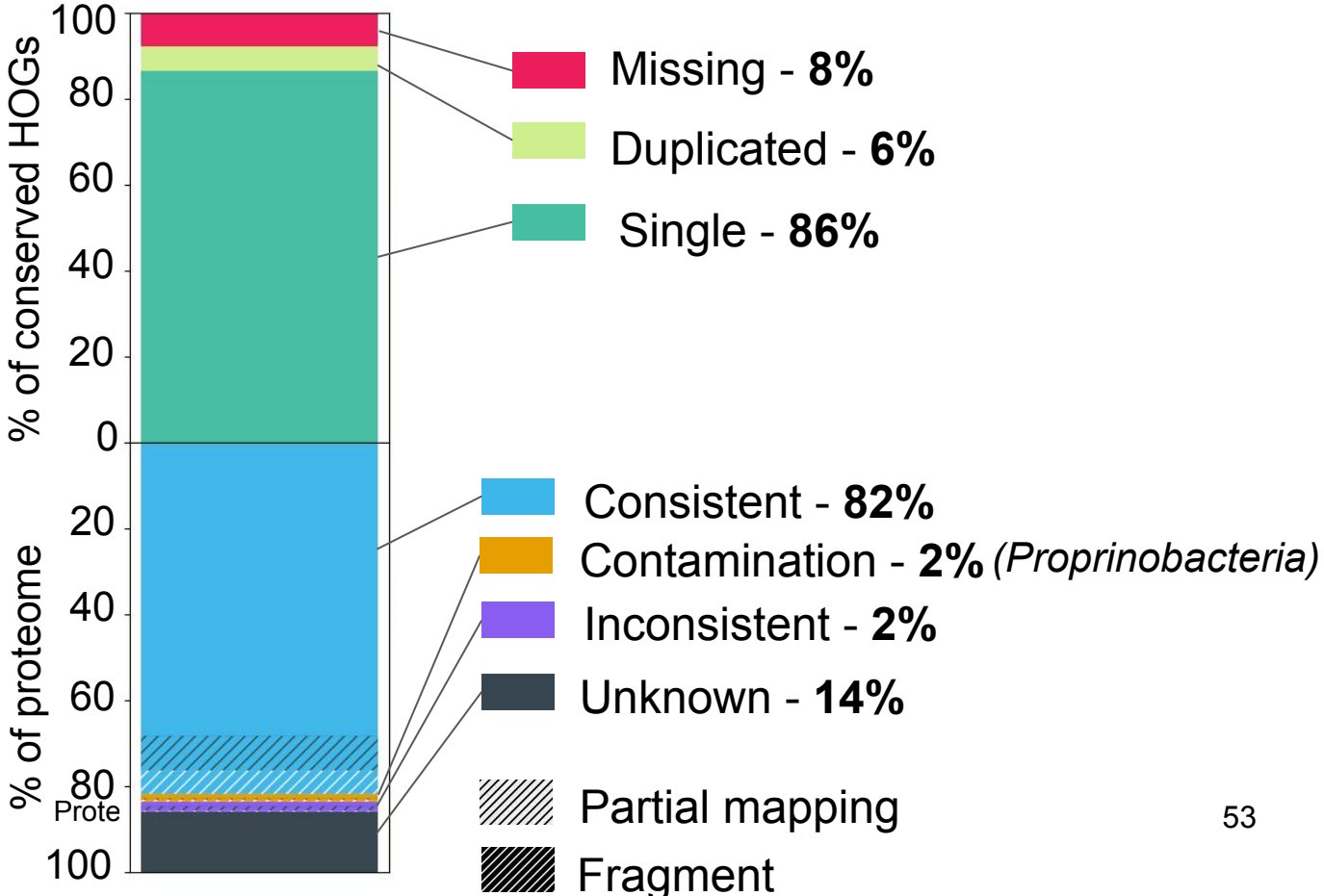
Example result

Platysternon megacephalum

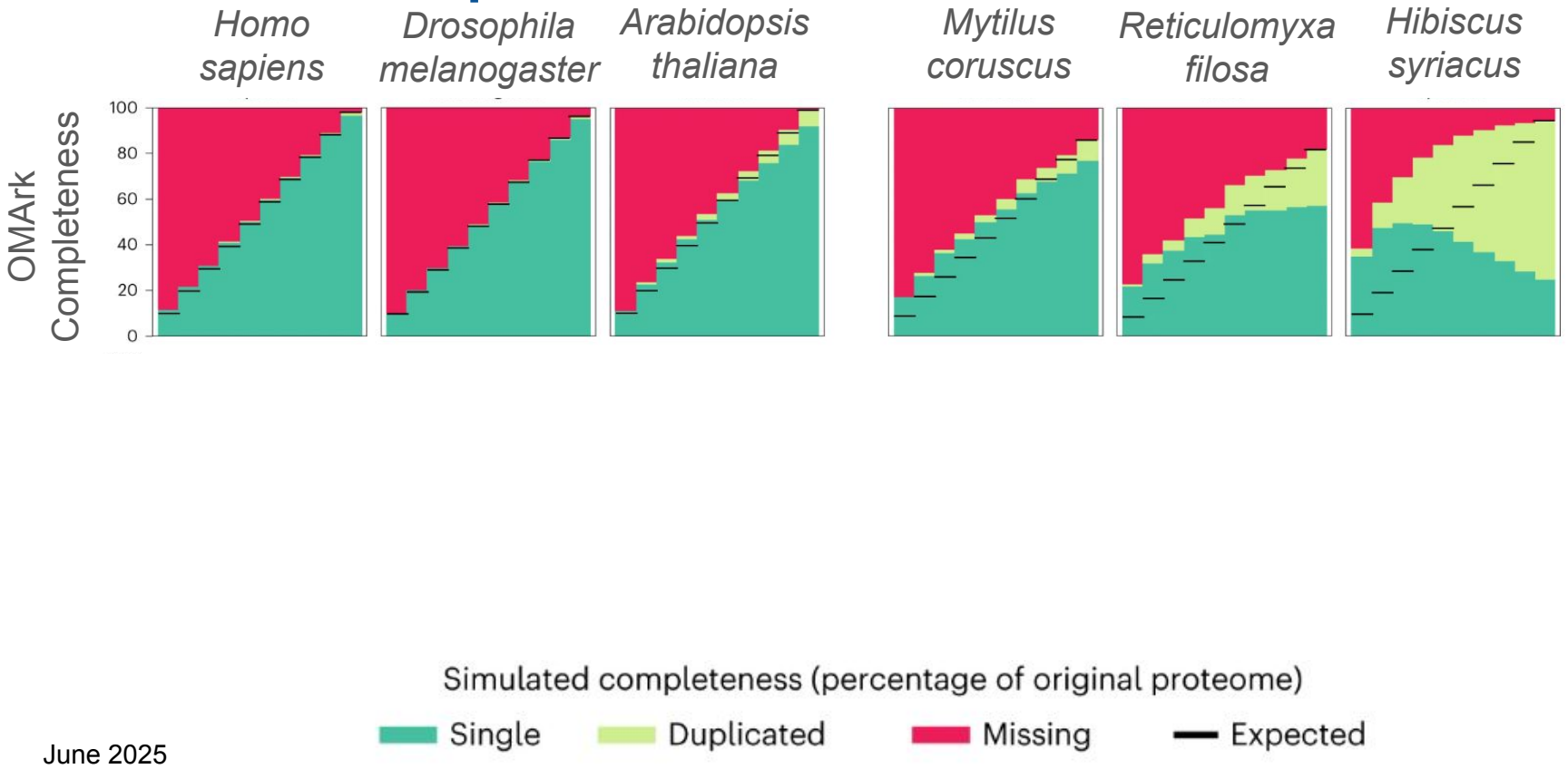
Clade : **Archelosauria**

10,514 conserved HOGs

Number of genes :
21,371



Simulation : completeness



Simulation : erroneous sequences

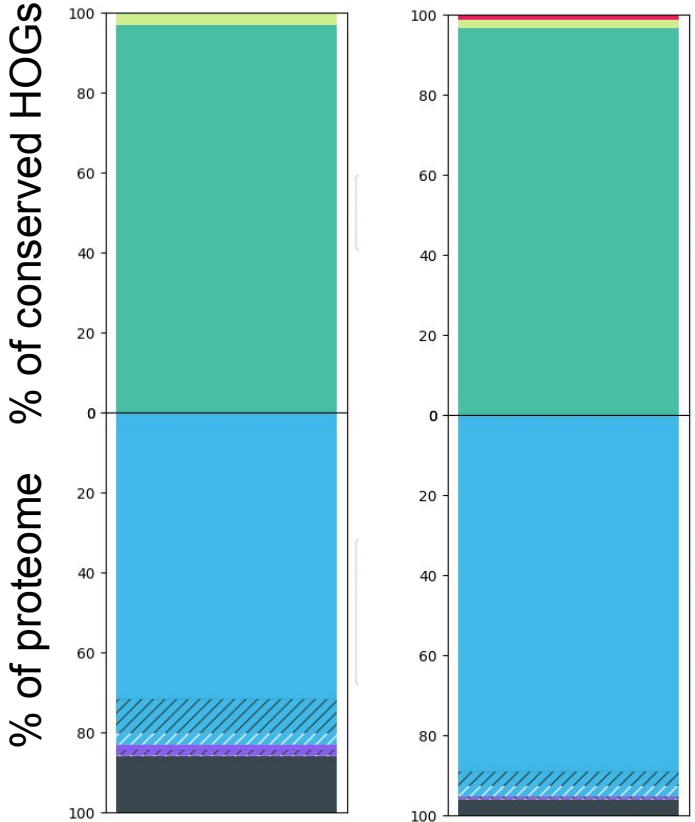
Protocol : Add randomly generated sequences; by increments of 10% of its size

Homo sapiens

Use-case : annotation comparison

Ensembl Metazoa 52
15,896 proteins
July 2022

Bumblebee
(*Bombus impatiens*)



Ensembl Metazoa 53+
10,632 proteins
Current

Results on public proteomes

Results on public data:

- 2,615 UniProt Reference proteomes
- 1,751 Ensemb proteomes
- 4,528 NCBI (Refseq-GenBank) proteomes
- User-uploaded proteomes

https://omark.omabrowser.org/

Take-home messages

- Protein k-mers is an efficient way to compare protein-coding genes and gene set
- They carry evolutionary and functional information
- Mainly used for fast similarity search and sequence classification
- Untapped potential for sequence analysis and global gene set comparisons ?

Take-home messages

- OMAmer combine a structured homology database and k-mer composition comparisons for fast mapping to gene family

- Fast and accurate mapping allows :
  - Scaled-up orthology inference
  - Whole proteome quality assessment
  - Species identification and gene structure evaluation

Acknowledgements

SIB Swiss Institute of Bioinformatics

UNIL | Université de Lausanne

comparative genomics LAB

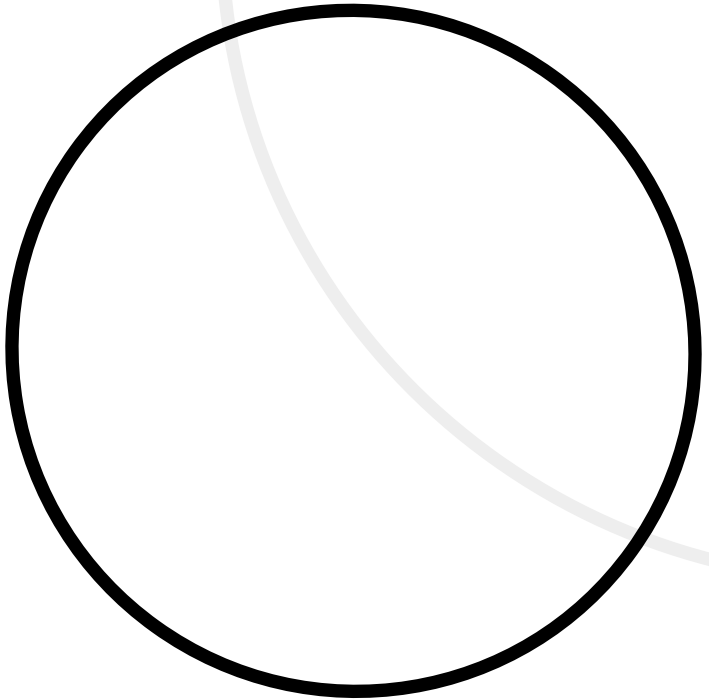
Christophe Dessimoz
Natasha Glover

Victor Rossier
Alex Warwick
Sina Majidian
David Moi
Silvia Prieto
Stefano Pascarelli
Irene Julca

Université de Strasbourg

iCUBE

thank you



Additional slides

“P-values” in OMAmer

- Probability of having *n* matches to a HOG at random depends on :
  - The number of k-mers in a HOG
  - The frequencies of these k-mers

- Modeled as a drawing with replacement problem
  - **What is the probability of n hits to a HOG given q (number of unique k-mers in the query) trials ?**
  - **Success** : drawing a k-mer that belongs to the target HOG in the database
  - *p* sampled from the binomial distribution

Assumption 1 : k-mers are independants (╳)

Assumption 2 : a same k-mer can be found twice *(but is q really small)*

https://github.com/DessimozLab/omamer