Swiss Institute of Bioinformatics

UNIL | Université de Lausanne

Département de Biologie Computationnelle

# Orthology inference at scale with FastOMA

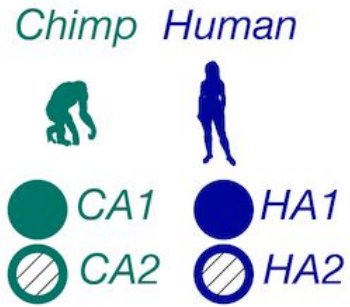

Sina Majidian

27 July 2023

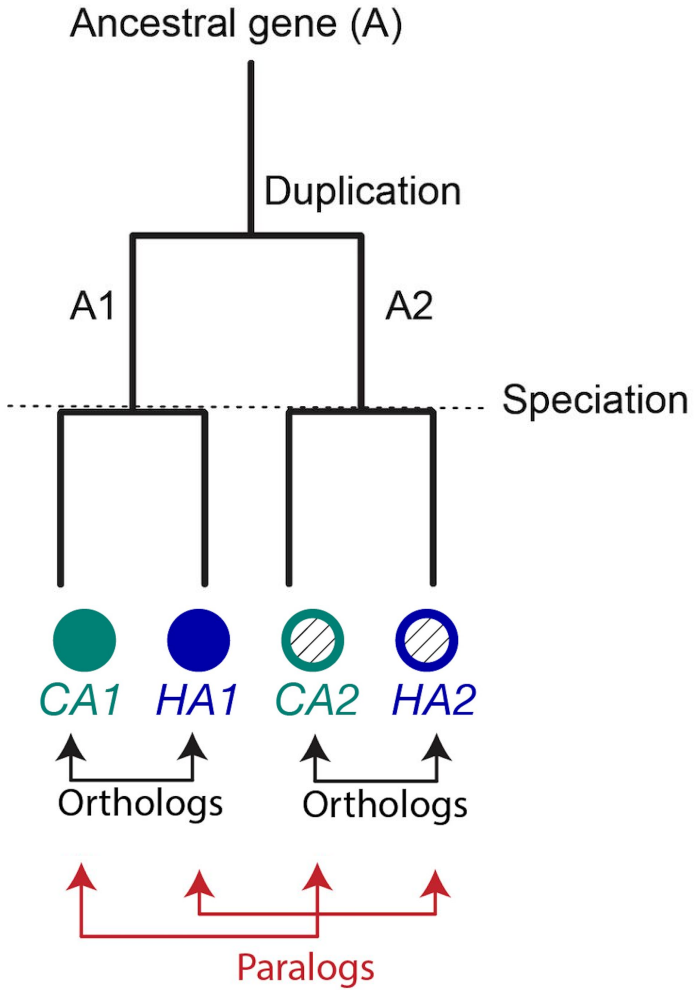
@DessimozLab
@SinaMajidian

# Introduction

- Orthology vs paralogy



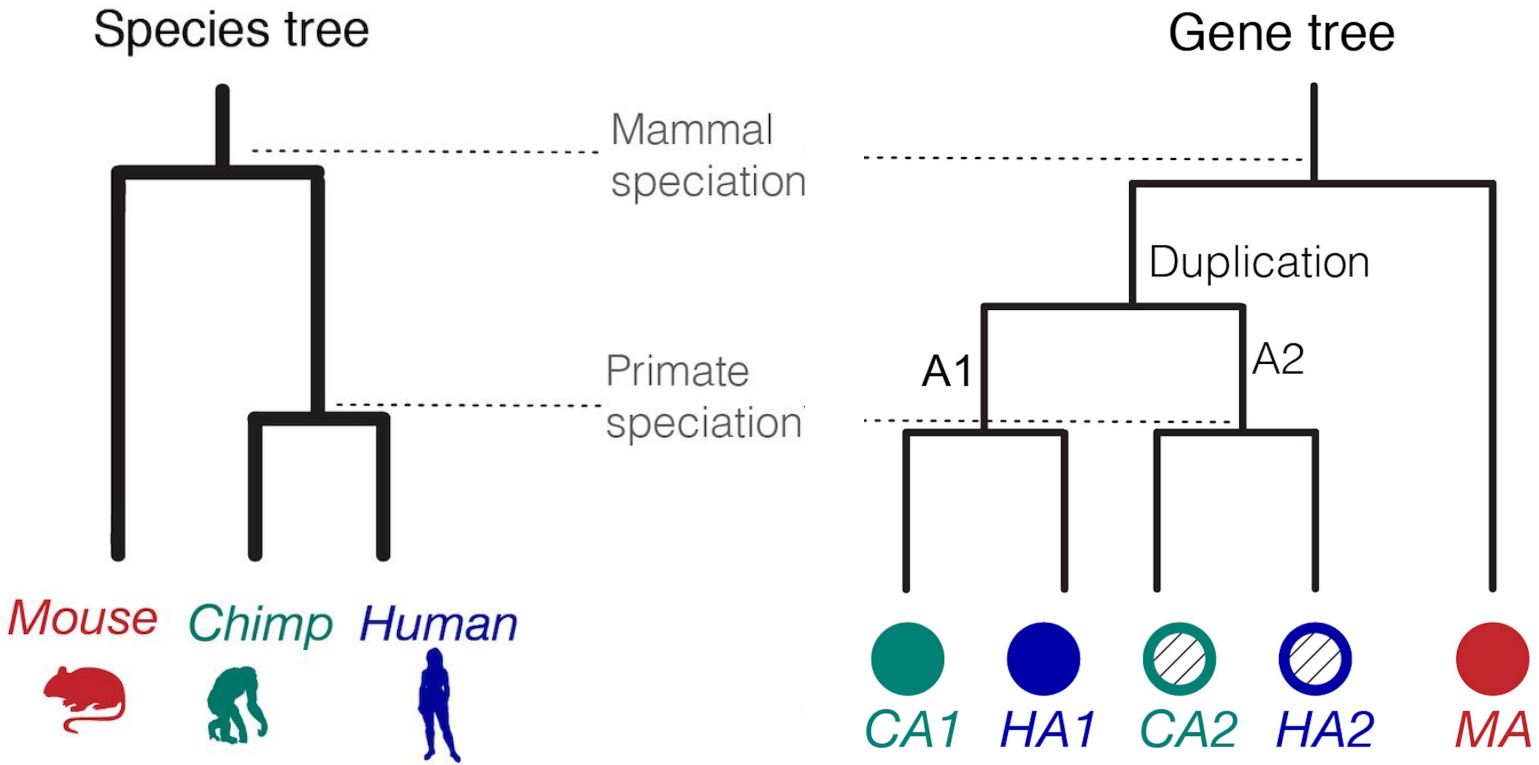

- Informative for
  - studying gene evolution
  - phylogeny inference, ...
- Found using all-vs-all alignment
  - **Not scalable**




Nevers, Orthology: promises and challenges, Evolutionary Biology. 2020.

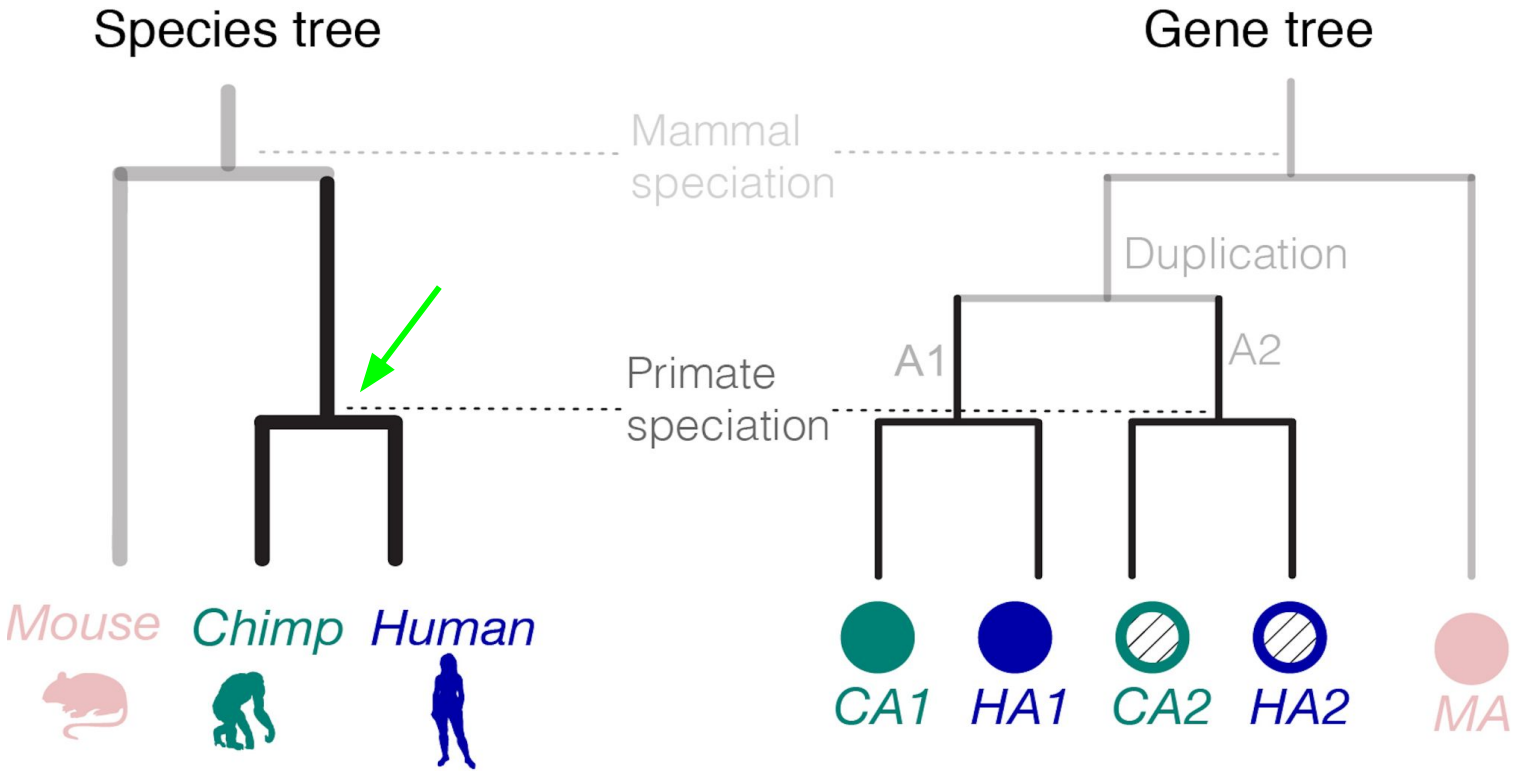
# **HOG** Hierarchical Orthologous Group

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



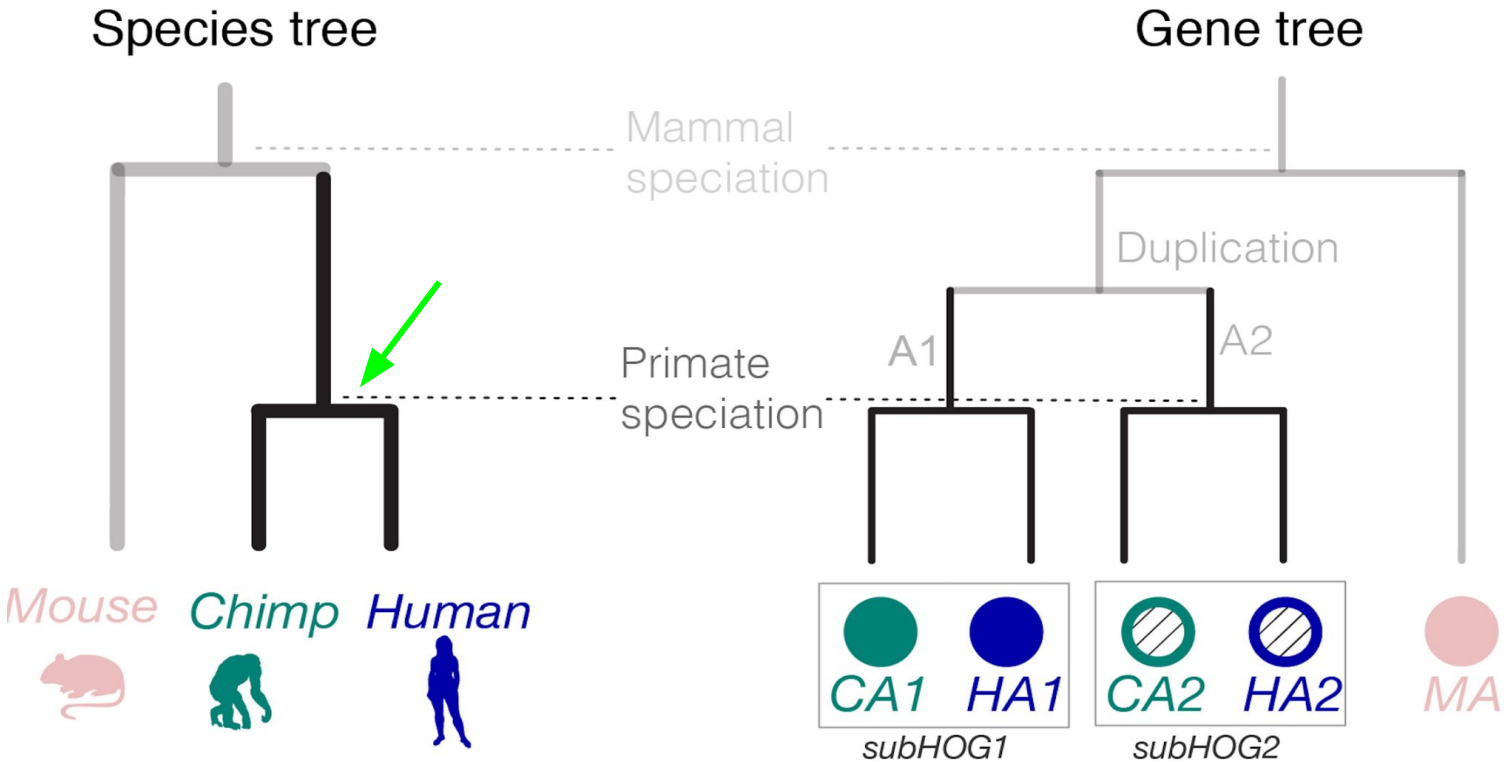
# **HOG** Hierarchical Orthologous Group

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



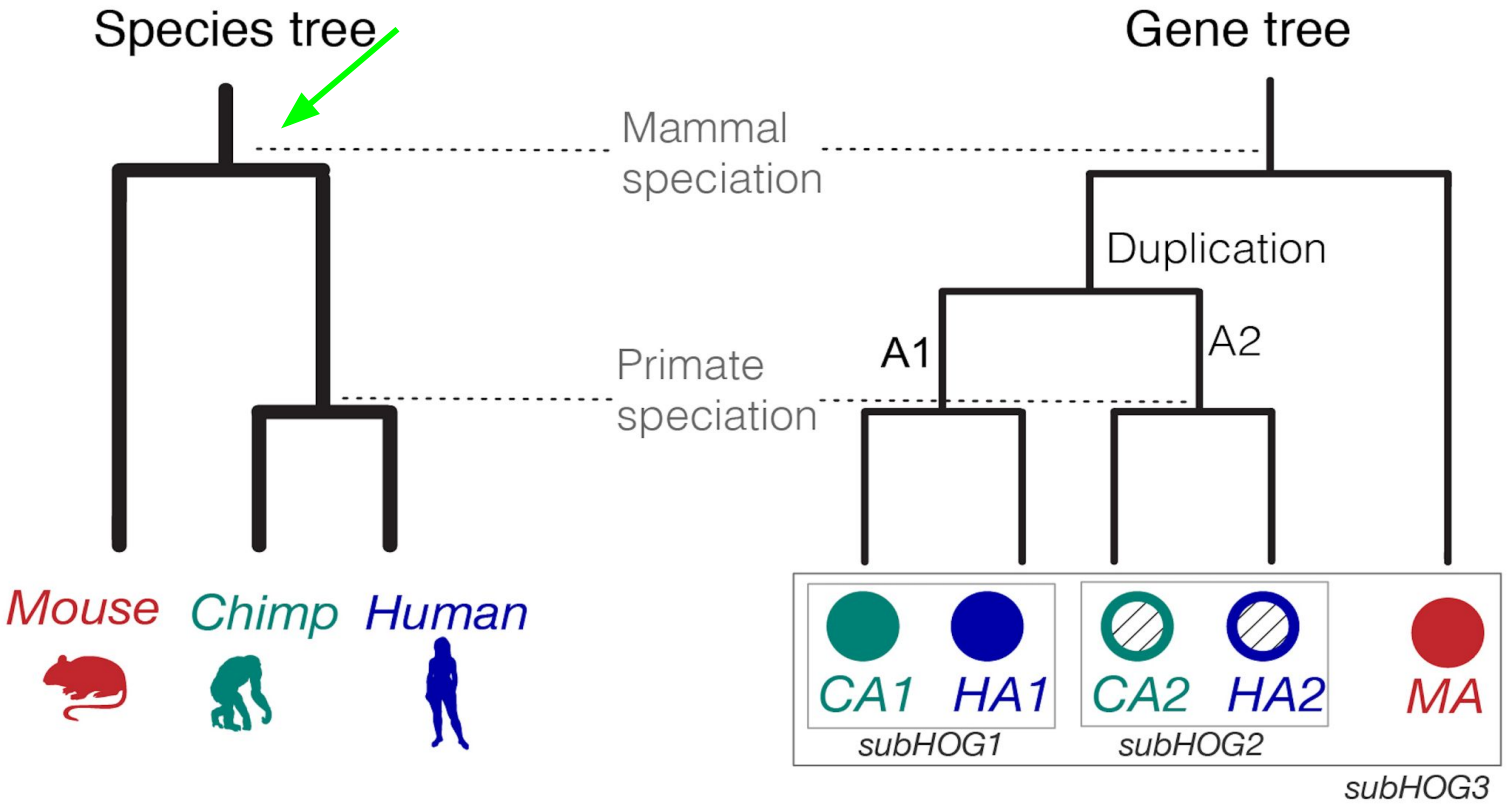
# **HOG** Hierarchical Orthologous Group

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

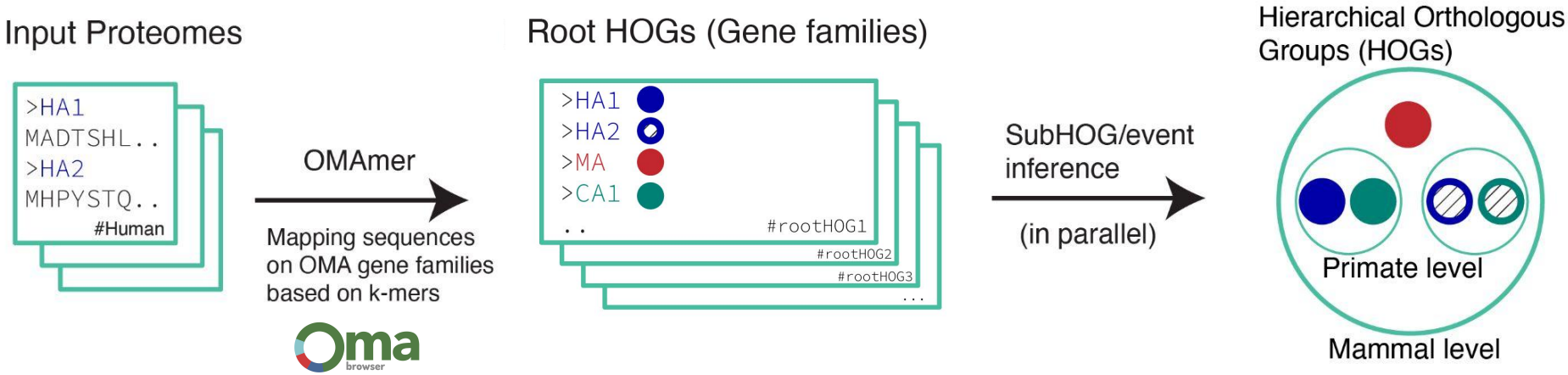


# **HOG** Hierarchical Orthologous Group

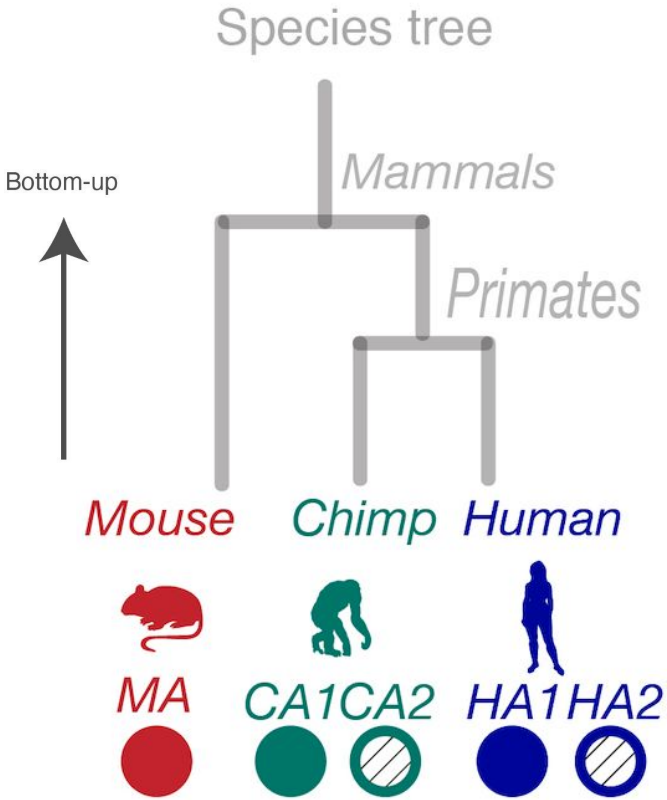
- Genes descended from a common ancestral gene at a specific taxonomic level

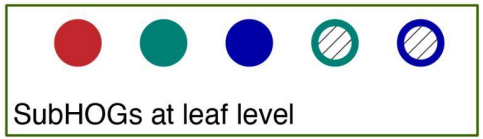


# **FastOMA**, our new tool

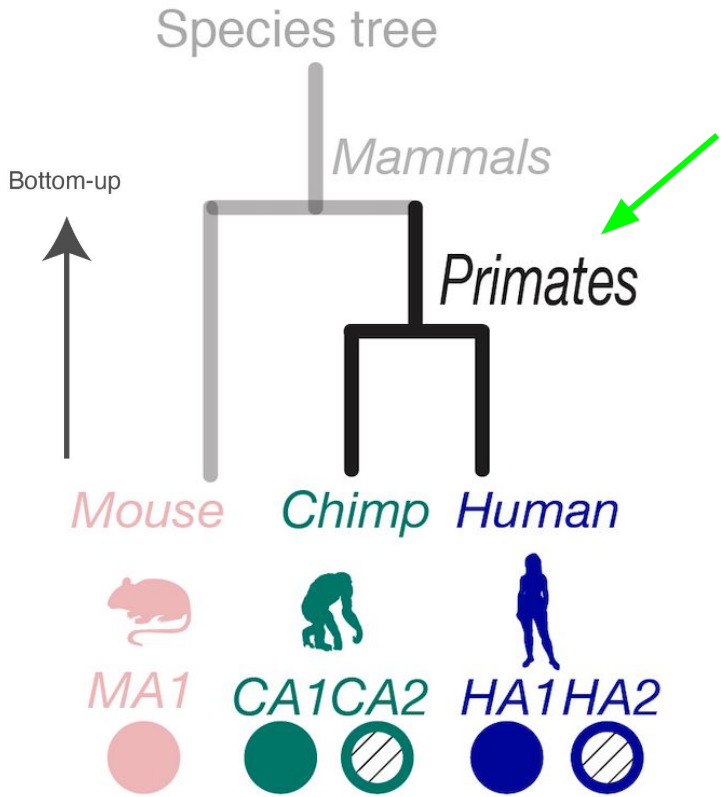


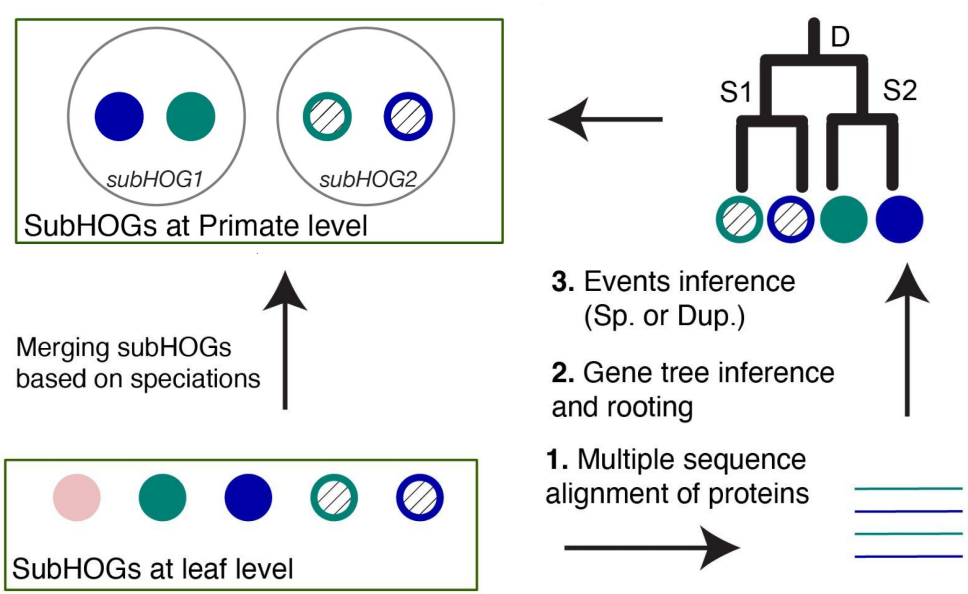
# HOG inference



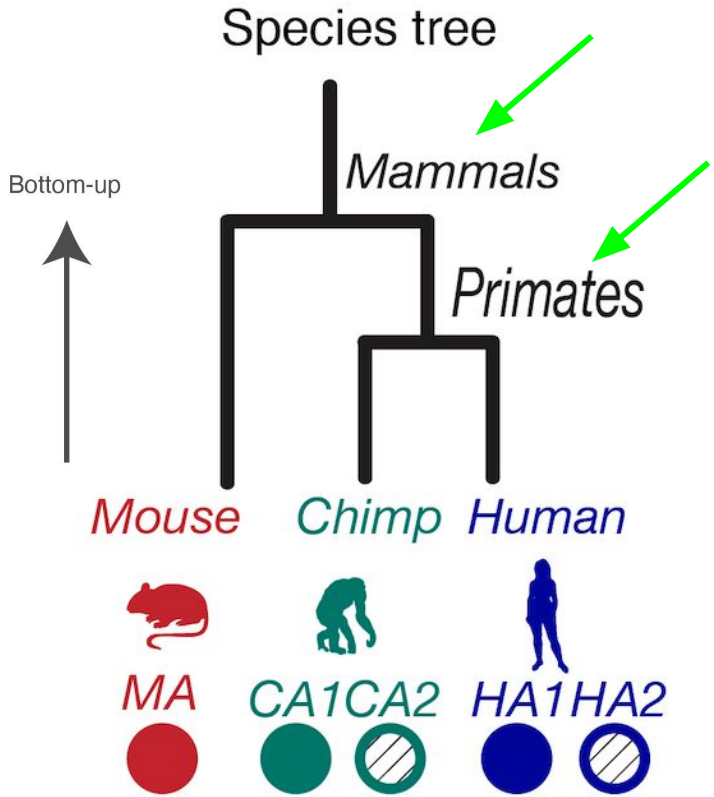



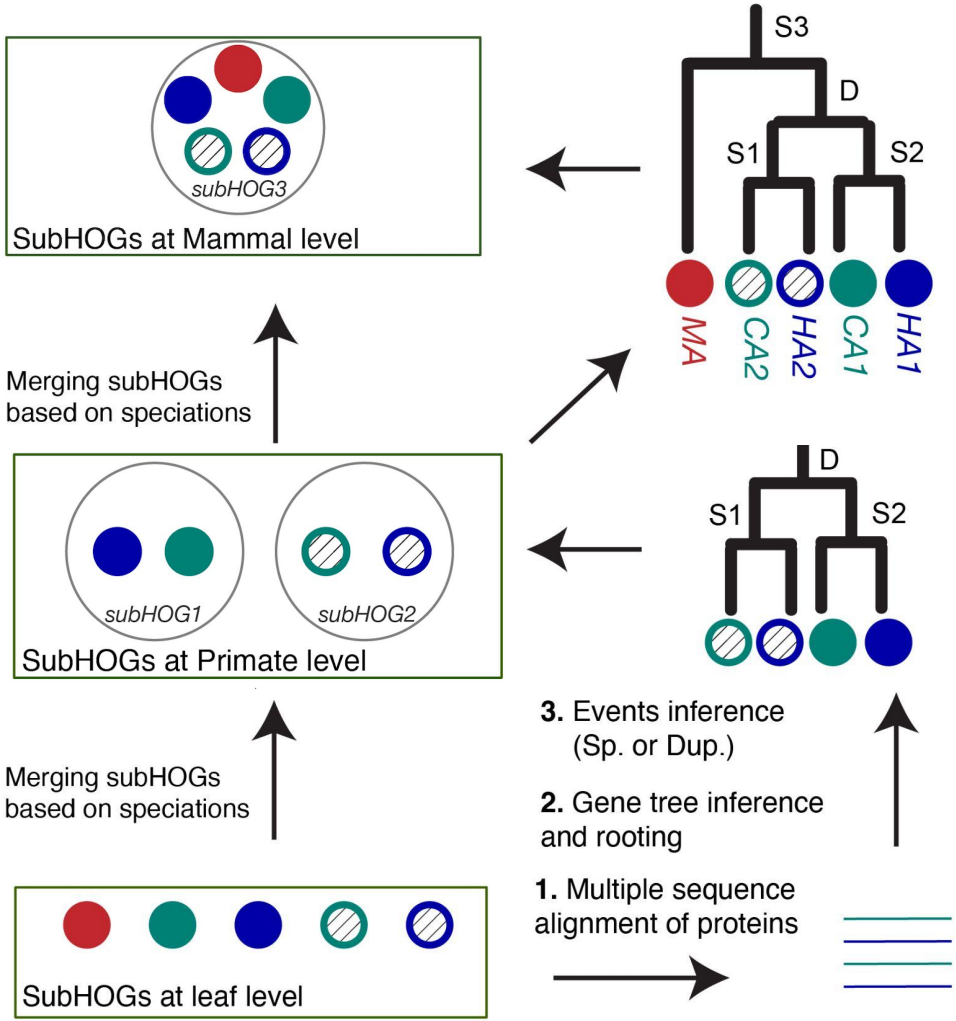
# HOG inference



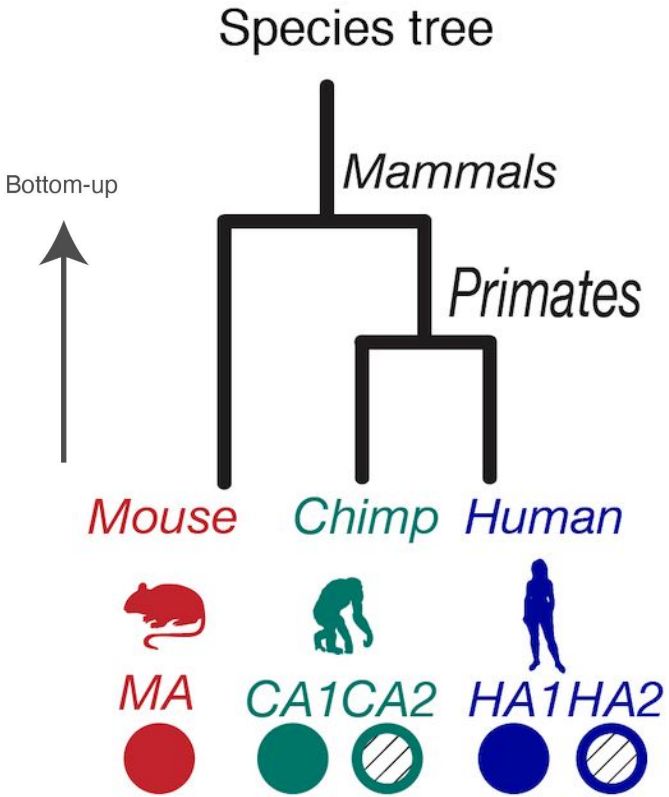



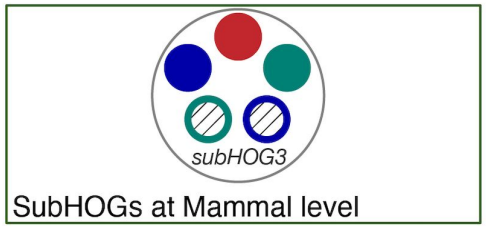
# HOG inference

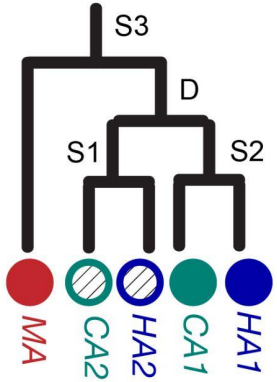


# HOG inference



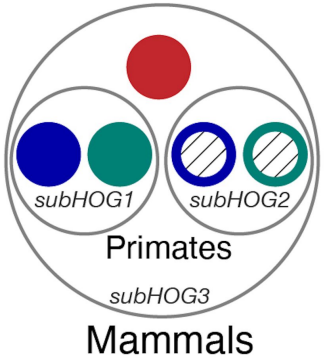



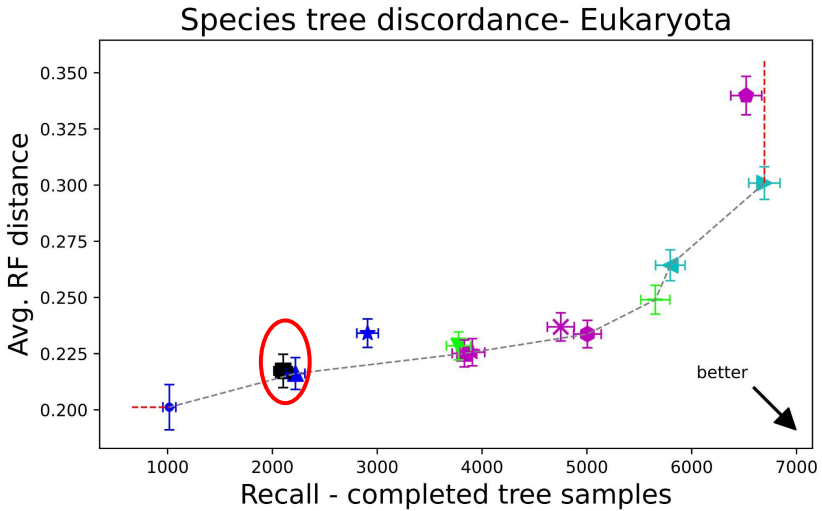

**HOG in orthoXML**

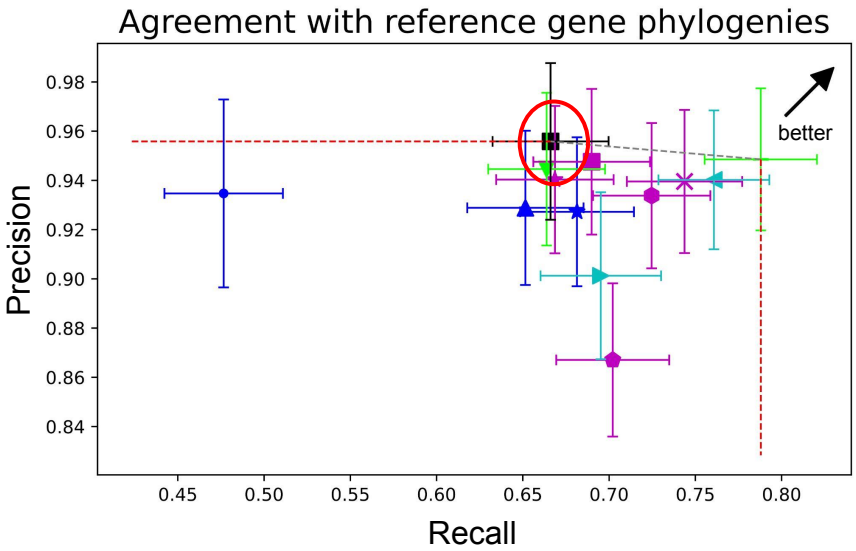
```
<orthoGroup3        Mammals>
  MA
  <paraGroup>
    <orthoGroup1    Primates>
      HA1
      CA1
    </orthoGroup1>
    <orthoGroup2    Primates>
      HA2
      CA2
    </orthoGroup2>
  </paraGroup>
</orthoGroup3>
```

**Nested structure of HOG**



# Quest for Orthologs Benchmarking set



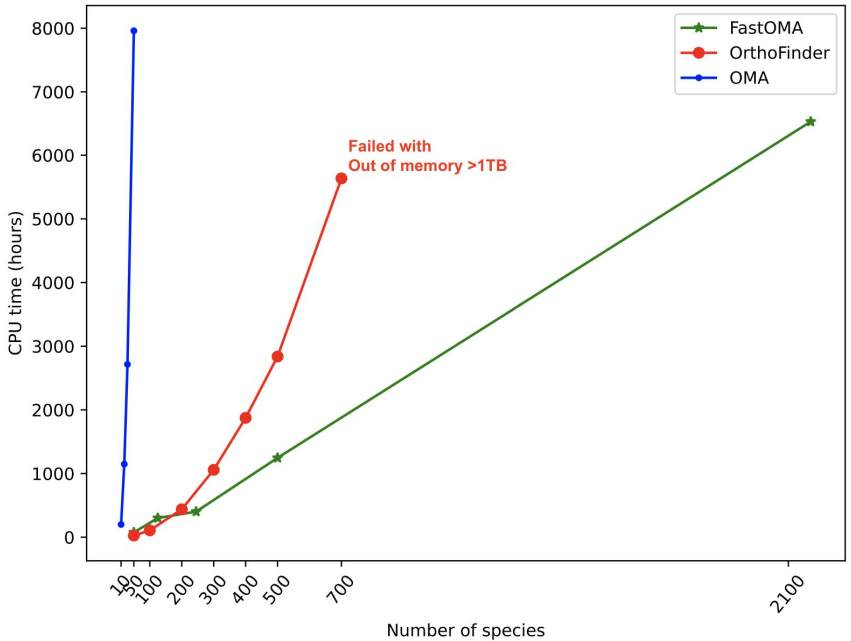




OMA_Pairs
OMA_Groups
OMA_GETHOGs
Domainoid+
InParanoid_Xenfix
OrthoMCL
OrthoInspector 3
sonicparanoid
PANTHER
Ensembl_Compara
Hieranoid_2
Orthofinder
FastOMA

**State-of-the-art methods**

# **Orthology inference** for Eukaryote dataset

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




**github.com/DessimozLab/FastOMA**

# Thank you !

**Our lab is hiring postdocs!**
**lab.dessimoz.org**

FNSNF
Swiss National Science Foundation

SIB
Swiss Institute of Bioinformatics

Unil
UNIL | Université de Lausanne

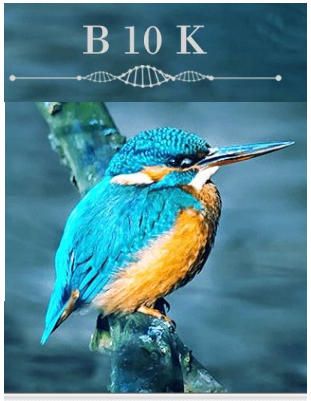
# Outline

- Introduction to orthology
- FastOMA
- OMA use cases
  - Phylogenetic profiling
  - Gene copy number
- Evolution of phenotypes
- Open questions


Darwin TREE of LIFE

BAT 1K

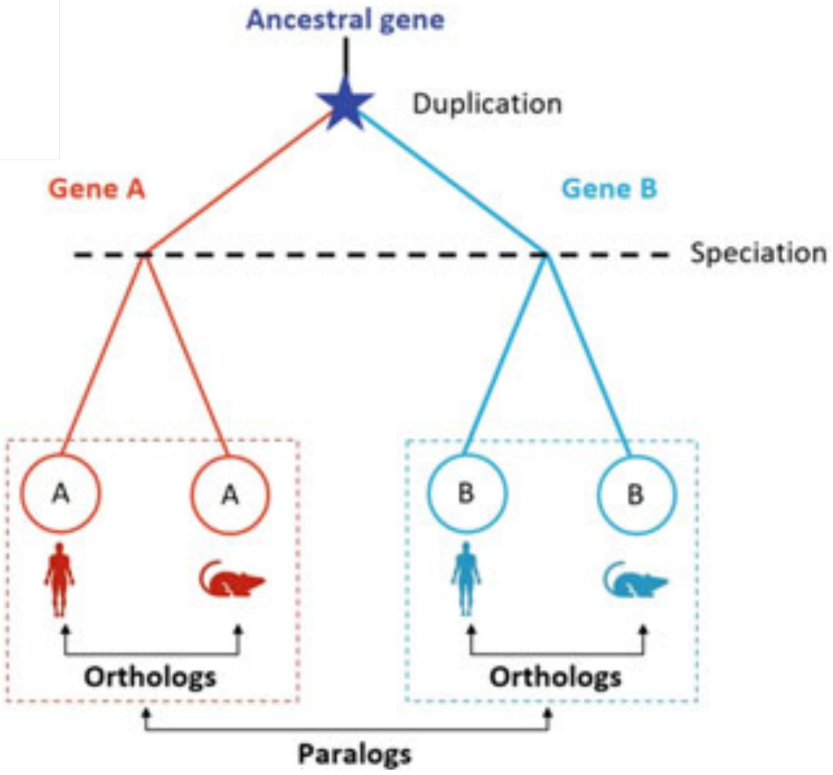
AQUATIC SYMBIOSIS GENOMICS



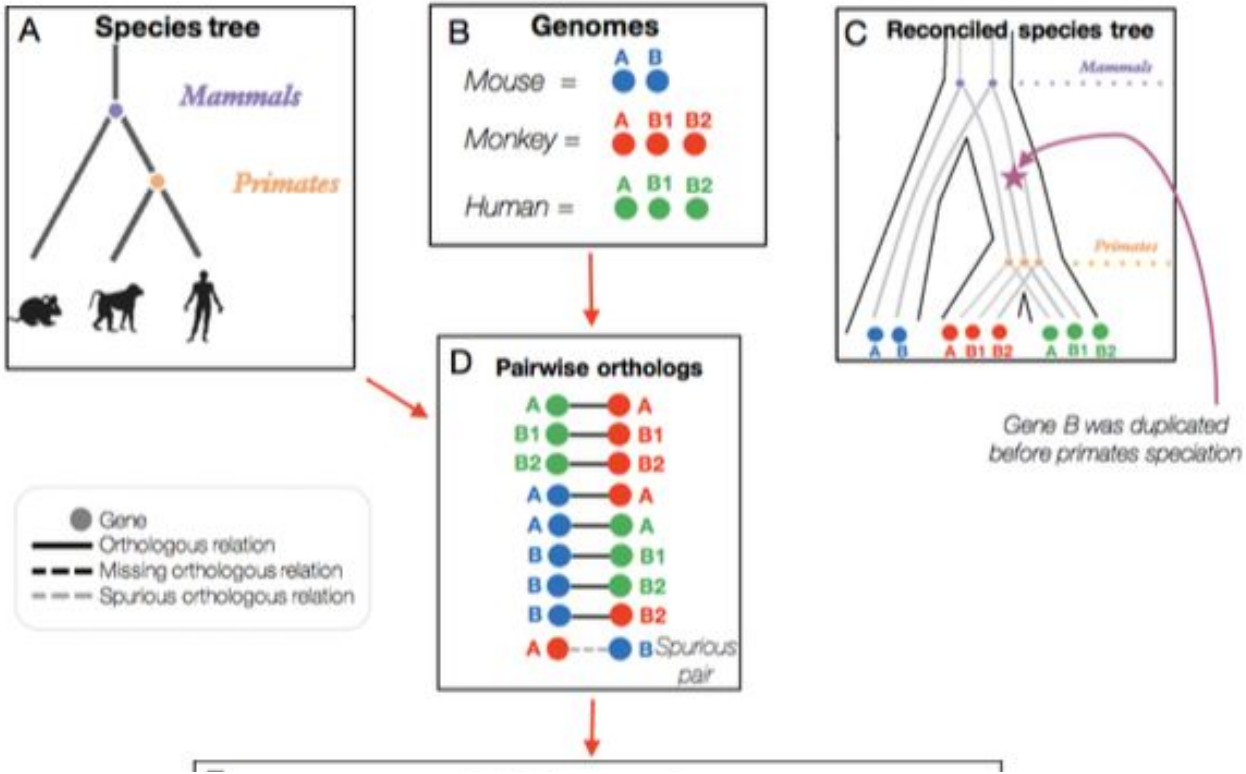
# Introduction

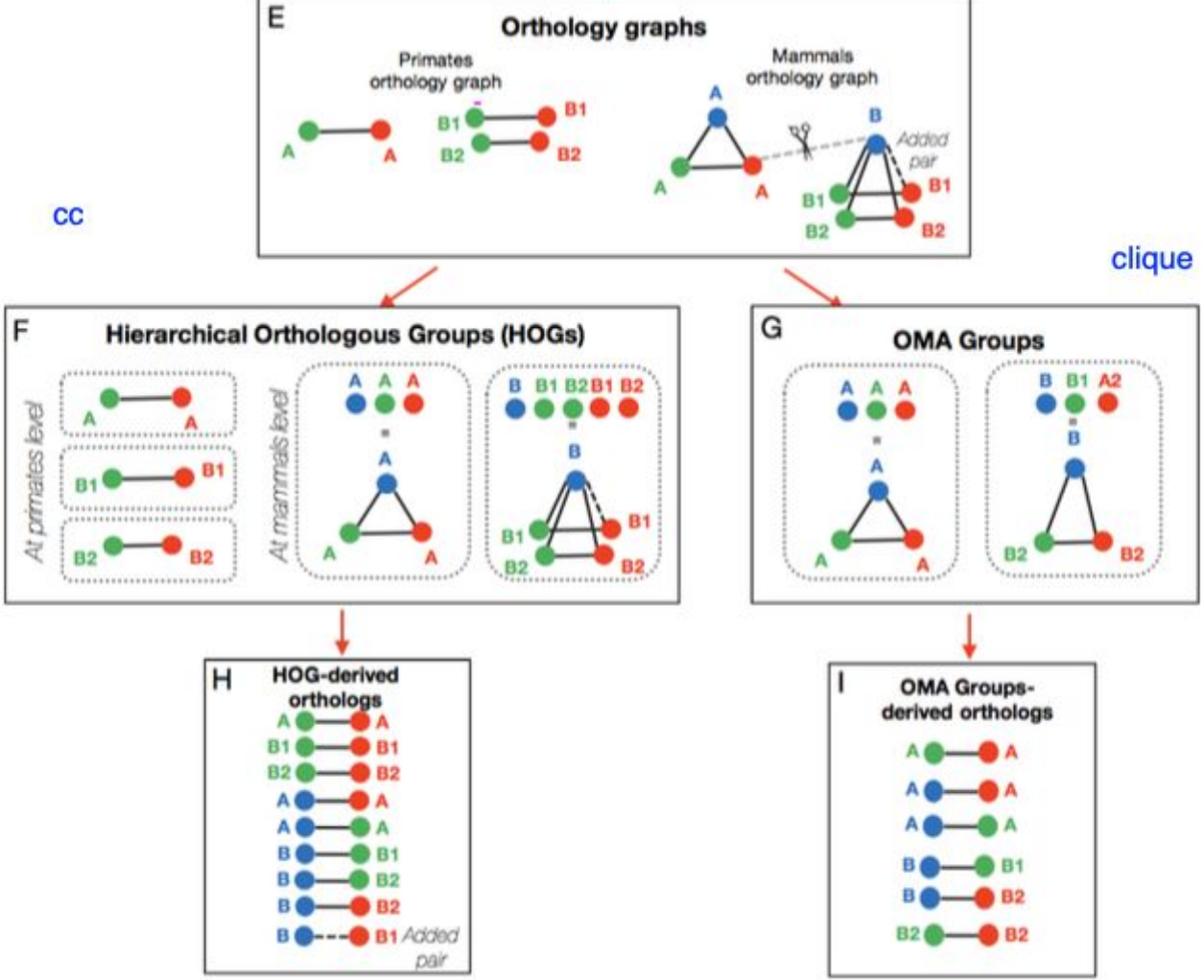
- Homologs
  - two proteins that descend from the same ancestor.
- Orthologs
  - two proteins in two species evolved form one gene in the same common ancestor
  - diverged from a speciation event
  - generally with same function

- Found using all-vs-all alignment
- Informative for studying gene family evolution, phylogeny inference, ...

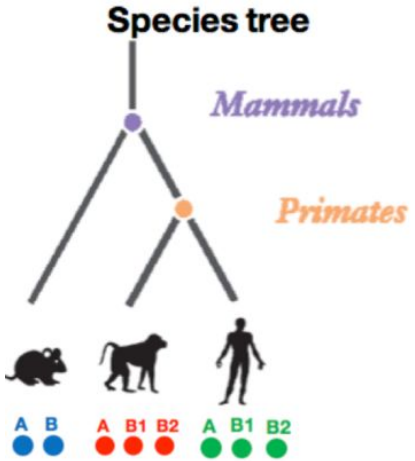



Nevers, Orthology: promises and challenges -Evolutionary Biology. 2020.


A Species tree
Mammals
Primates
B Genomes
Mouse =
Monkey =
Human =
A B
A B1 B2
A B1 B2
C Reconciled species tree
Mammals
Primates
Gene B was duplicated
before primates speciation
D Pairwise orthologs
Spurious pair
Gene
Orthologous relation
Missing orthologous relation
Spurious orthologous relation


E
Orthology graphs
Primates orthology graph
Mammals orthology graph
Added pair
cc
clique
F
Hierarchical Orthologous Groups (HOGs)
At primates level
At mammals level
G
OMA Groups
H
HOG-derived orthologs
Added pair
I
OMA Groups-derived orthologs

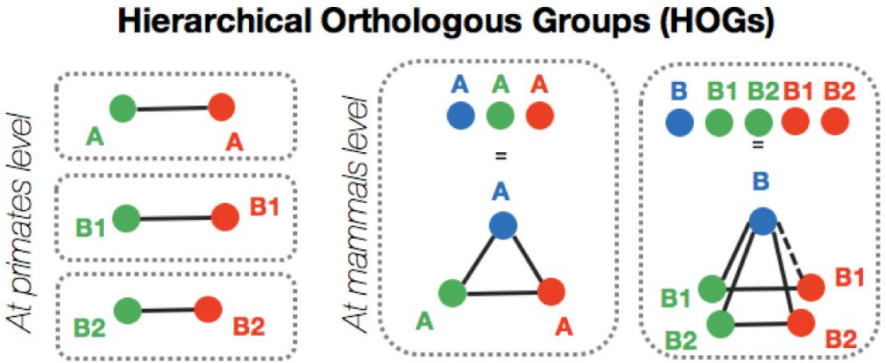
# **HOG** Hierarchical Orthologous Group




- sets of genes  descended from a common ancestral gene at a specific taxonomic level

- Informative for

- map gene loss and duplication onto species trees
- Reconstruct ancestral genomes
- Predict of gene function in new species
- Identify genetic changes enabling organisms to adapt to different environments

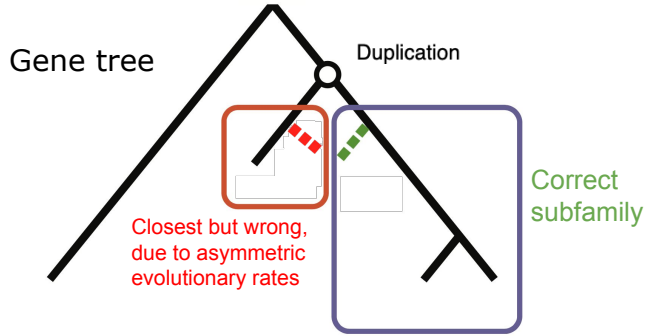


# Basis of the proposed pipeline (FastOMA)

- Map proteins to the database of subfamily of genes

- Traditionally achieved by finding the closest sequence (by BLAST or DIAMOND).

  ❌ the closest sequence might belong to a different subfamily (not ortholog)



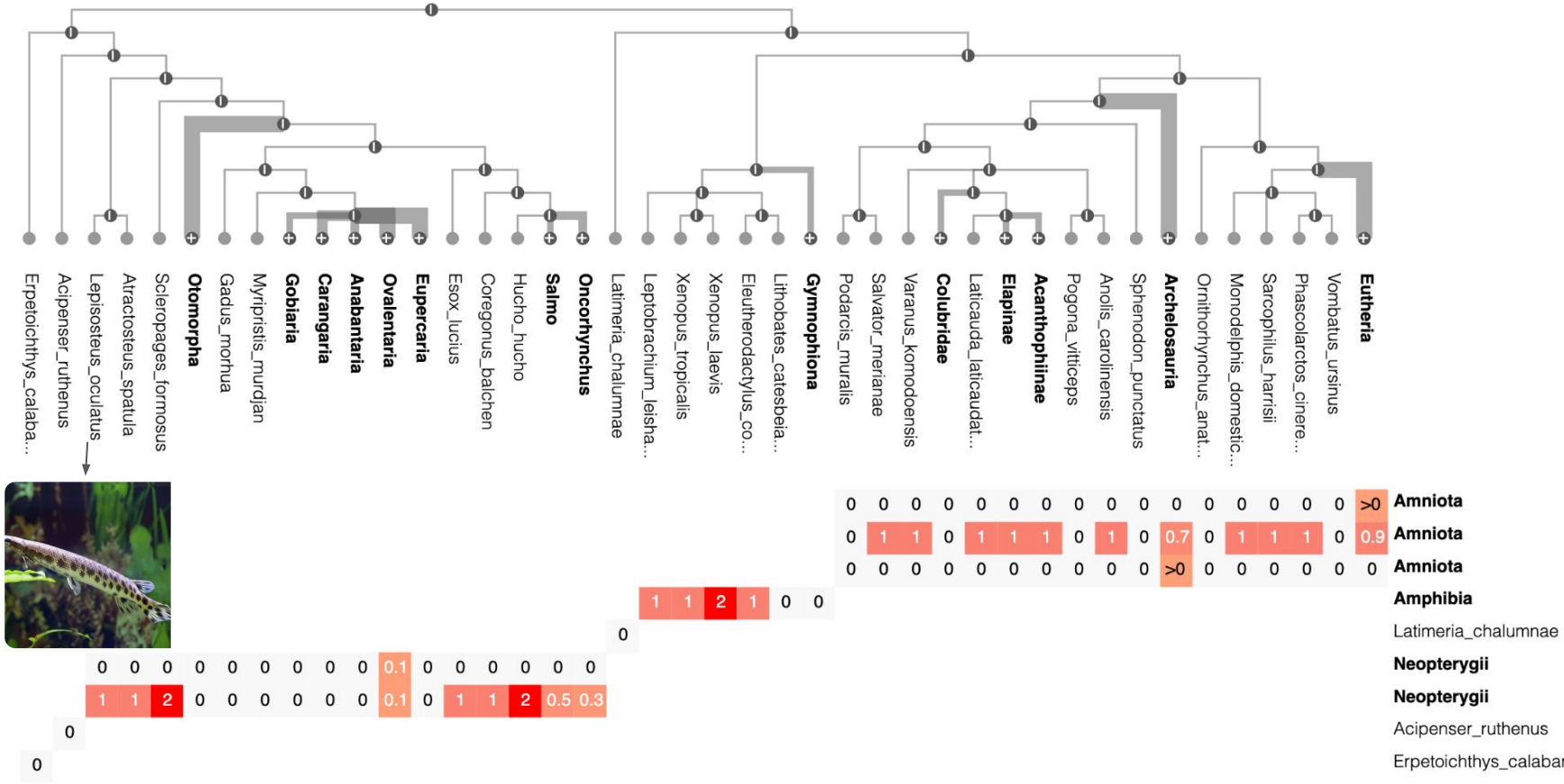

Our solution is OMAmer:

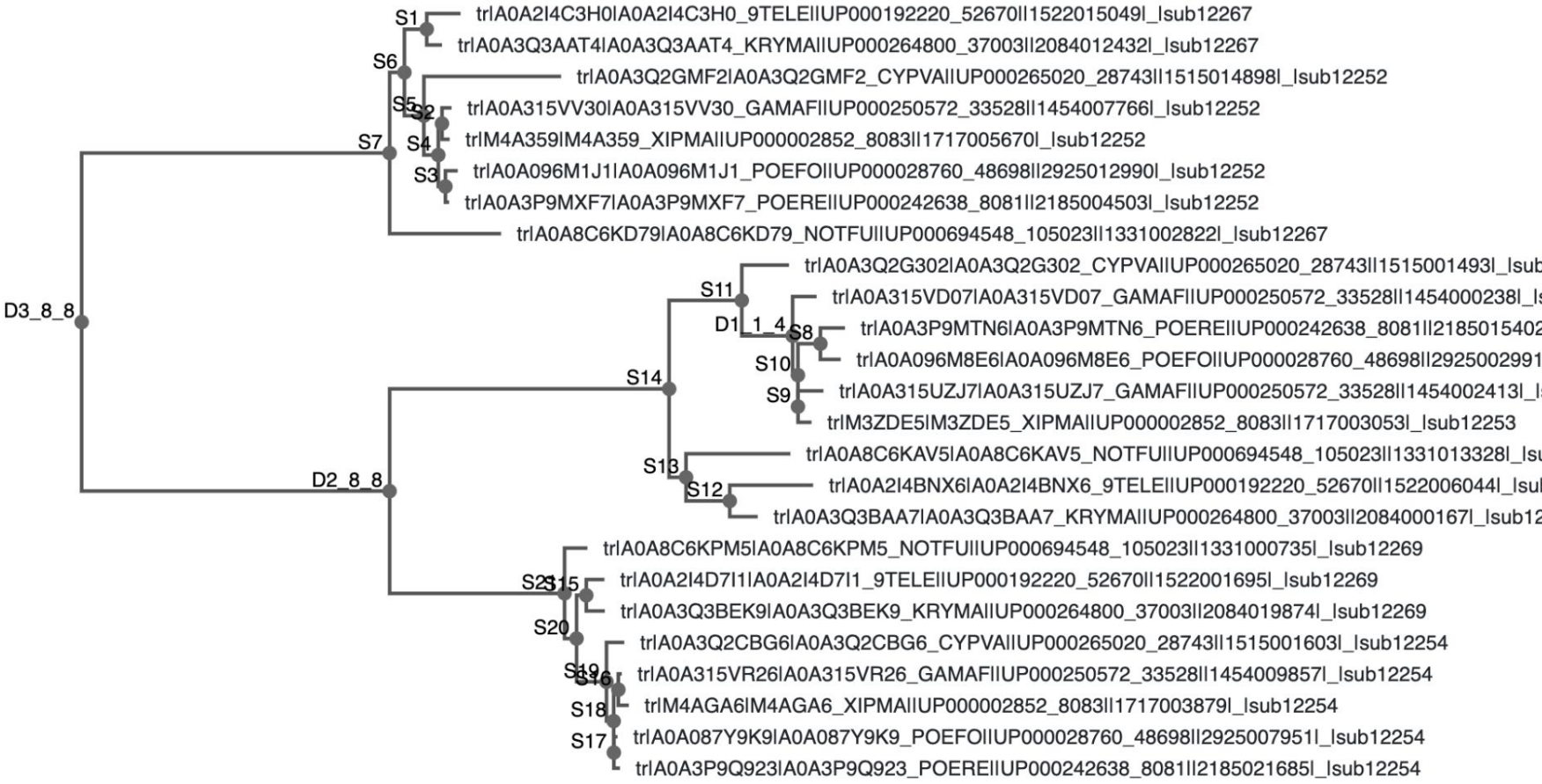
A subfamily-level classifier
using subfamily-informed k-mers.

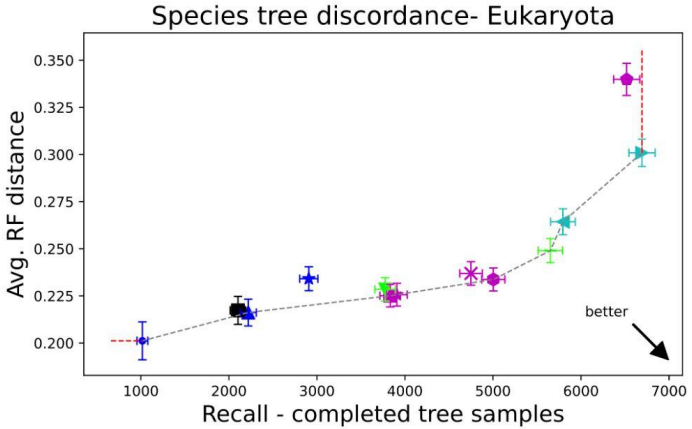
Rossier, V., et al. "OMAmer: tree-driven and alignment-free protein assignment to subfamilies outperforms closest sequence approaches." Bioinformatics 2021.

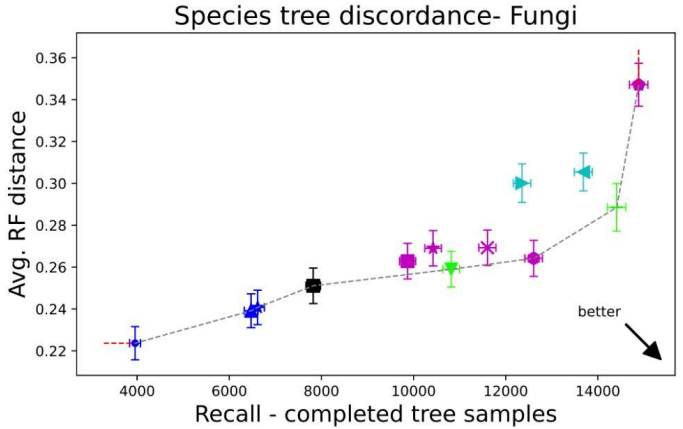
# Versican Core Protein

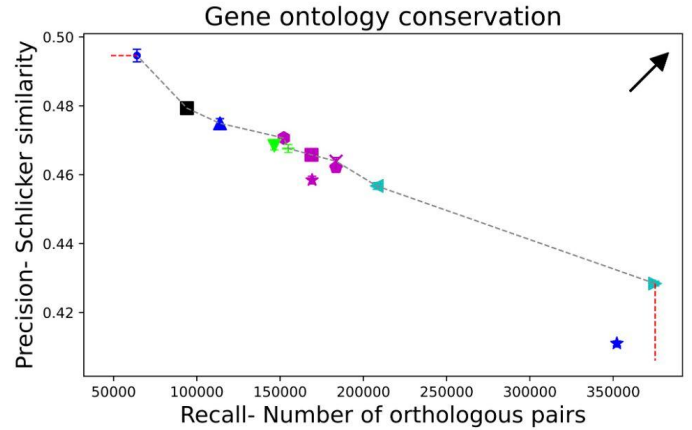
403 proteins
from 367 species

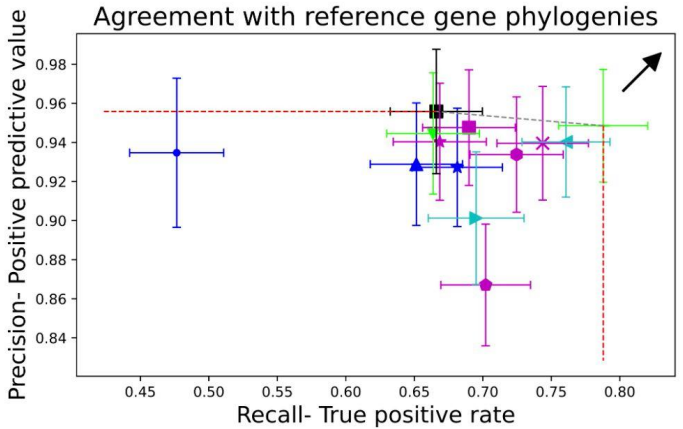



S1
trIA0A2I4C3H0IA0A2I4C3H0_9TELEIIUP000192220_52670II1522015049I_Isub12267
trIA0A3Q3AAT4IA0A3Q3AAT4_KRYMAIIUP000264800_37003II2084012432I_Isub12267
S6
trIA0A3Q2GMF2IA0A3Q2GMF2_CYPVAIIUP000265020_28743II1515014898I_Isub12252
S5
S2
trIA0A315VV30IA0A315VV30_GAMAFIIUP000250572_33528II1454007766I_Isub12252
S7
S4
trIM4A359IM4A359_XIPMAIIUP000002852_8083II1717005670I_Isub12252
S3
trIA0A096M1J1IA0A096M1J1_POEFOIIUP000028760_48698II2925012990I_Isub12252
trIA0A3P9MXF7IA0A3P9MXF7_POEREIIUP000242638_8081II2185004503I_Isub12252
trIA0A8C6KD79IA0A8C6KD79_NOTFUIIUP000694548_105023II1331002822I_Isub12267
trIA0A3Q2G302IA0A3Q2G302_CYPVAIIUP000265020_28743II1515001493I_Isub
S11
trIA0A315VD07IA0A315VD07_GAMAFIIUP000250572_33528II1454000238I_Is
D3_8_8
D1_1_4
S8
trIA0A3P9MTN6IA0A3P9MTN6_POEREIIUP000242638_8081II218501540
S10
trIA0A096M8E6IA0A096M8E6_POEFOIIUP000028760_48698II292500299
S14
S9
trIA0A315UZJ7IA0A315UZJ7_GAMAFIIUP000250572_33528II1454002413I_Is
trIM3ZDE5IM3ZDE5_XIPMAIIUP000002852_8083II1717003053I_Isub12253
trIA0A8C6KAV5IA0A8C6KAV5_NOTFUIIUP000694548_105023II1331013328I_Isu
S13
D2_8_8
S12
trIA0A2I4BNX6IA0A2I4BNX6_9TELEIIUP000192220_52670II1522006044I_Isub
trIA0A3Q3BAA7IA0A3Q3BAA7_KRYMAIIUP000264800_37003II2084000167I_Isub12
trIA0A8C6KPM5IA0A8C6KPM5_NOTFUIIUP000694548_105023II1331000735I_Isub12269
S21
S15
trIA0A2I4D7I1IA0A2I4D7I1_9TELEIIUP000192220_52670II1522001695I_Isub12269
trIA0A3Q3BEK9IA0A3Q3BEK9_KRYMAIIUP000264800_37003II2084019874I_Isub12269
S20
trIA0A3Q2CBG6IA0A3Q2CBG6_CYPVAIIUP000265020_28743II1515001603I_Isub12254
S19
S16
trIA0A315VR26IA0A315VR26_GAMAFIIUP000250572_33528II1454009857I_Isub12254
S18
trIM4AGA6IM4AGA6_XIPMAIIUP000002852_8083II1717003879I_Isub12254
S17
trIA0A087Y9K9IA0A087Y9K9_POEFOIIUP000028760_48698II2925007951I_Isub12254
trIA0A3P9Q923IA0A3P9Q923_POEREIIUP000242638_8081II2185021685I_Isub12254

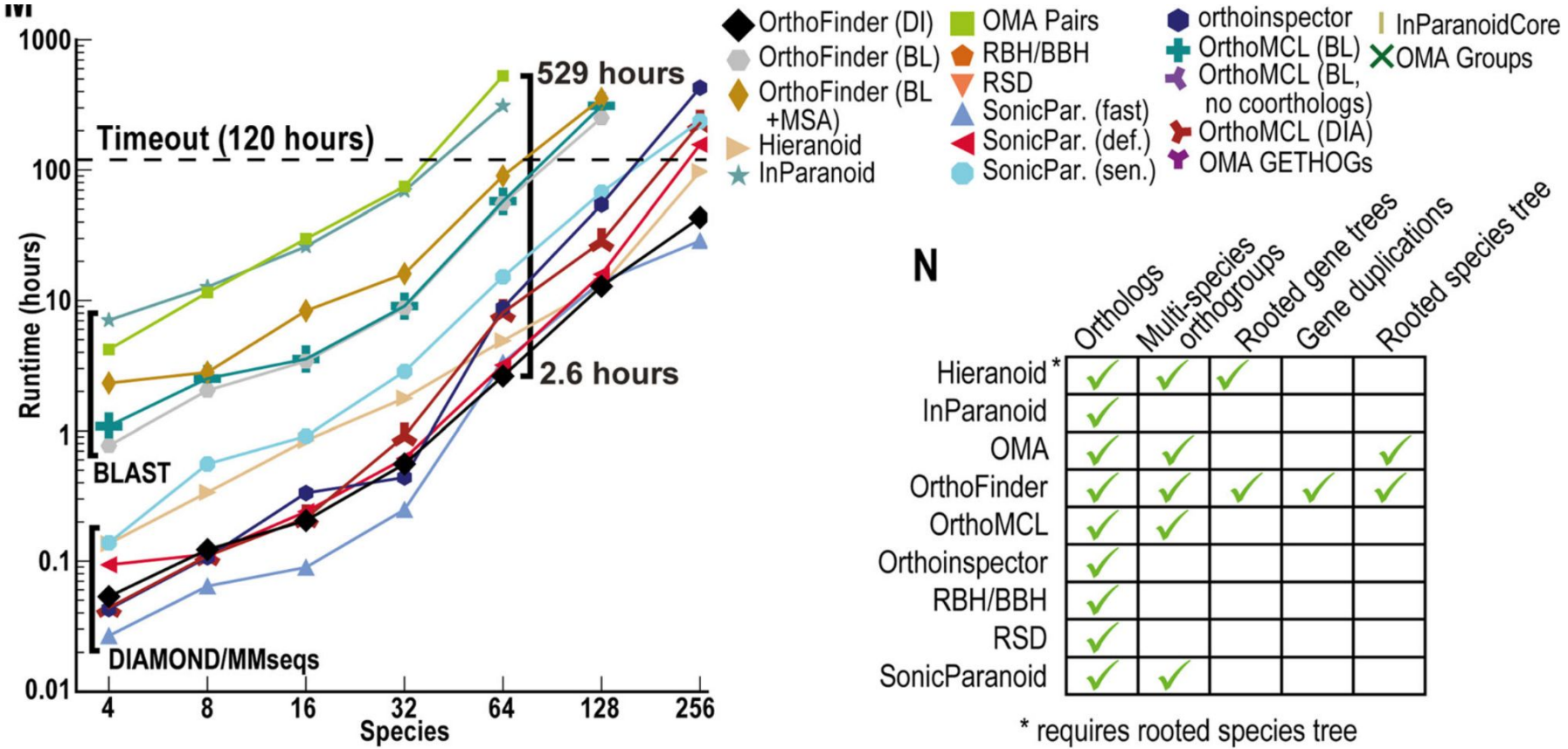

Species tree discordance- Eukaryota
Avg. RF distance
Recall - completed tree samples
better


Species tree discordance- Fungi
Avg. RF distance
Recall - completed tree samples
better


Gene ontology conservation
Precision- Schlicker similarity
Recall- Number of orthologous pairs


Agreement with reference gene phylogenies
Precision- Positive predictive value
Recall- True positive rate


OMA_Pairs
OMA_Groups
OMA_GETHOGs
Domainoid+
InParanoid_Xenfix
OrthoMCL
OrthoInspector 3
sonicparanoid
PANTHER
Ensembl_Compara
Hieranoid_2
Orthofinder
FastOMA


M
1000
100
10
1
0.1
0.01
Runtime (hours)
Timeout (120 hours)
529 hours
2.6 hours
BLAST
DIAMOND/MMseqs
4
8
16
32
64
128
256
Species
OrthoFinder (DI)
OrthoFinder (BL)
OrthoFinder (BL +MSA)
Hieranoid
InParanoid
OMA Pairs
RBH/BBH
RSD
SonicPar. (fast)
SonicPar. (def.)
SonicPar. (sen.)
orthoinspector
OrthoMCL (BL)
OrthoMCL (BL, no coorthologs)
OrthoMCL (DIA)
OMA GETHOGs
InParanoidCore
OMA Groups
N
Orthologs
Multi-species orthogroups
Rooted gene trees
Gene duplications
Rooted species tree
Hieranoid *
InParanoid
OMA
OrthoFinder
OrthoMCL
Orthoinspector
RBH/BBH
RSD
SonicParanoid
* requires rooted species tree