

Module 3: FastOMA



FastOMA, our new tool

Input Proteomes

```
>HA1  
MADTSHL..  
>HA2  
MHPYSTQ..  
#Human
```

OMAmer
Mapping sequences
on OMA gene families
based on k-mers

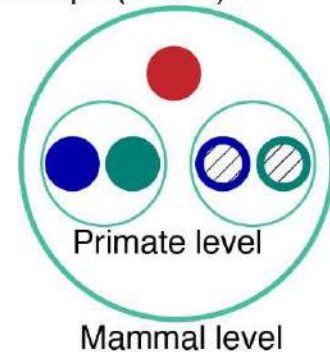


Root HOGs (Gene families)

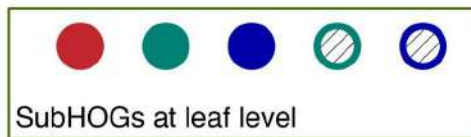
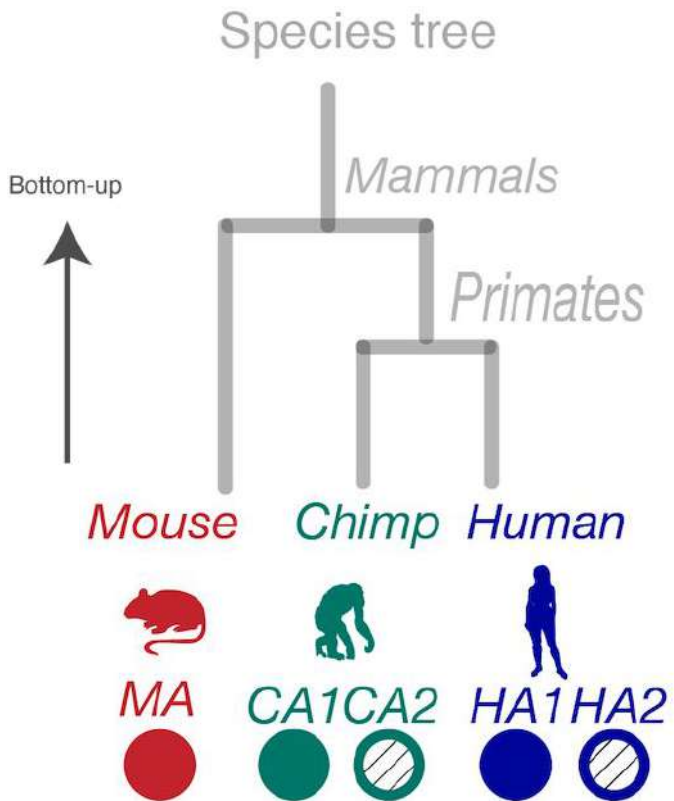
```
>HA1 ●  
>HA2 ●  
>MA ●  
>CA1 ●  
.. #rootHOG1  
.. #rootHOG2  
.. #rootHOG3  
...
```

SubHOG/event
inference
(in parallel)

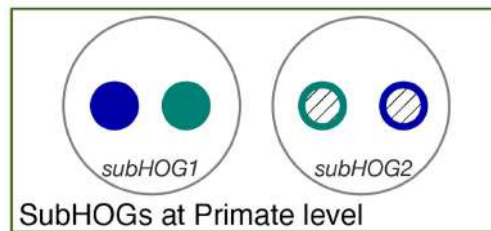
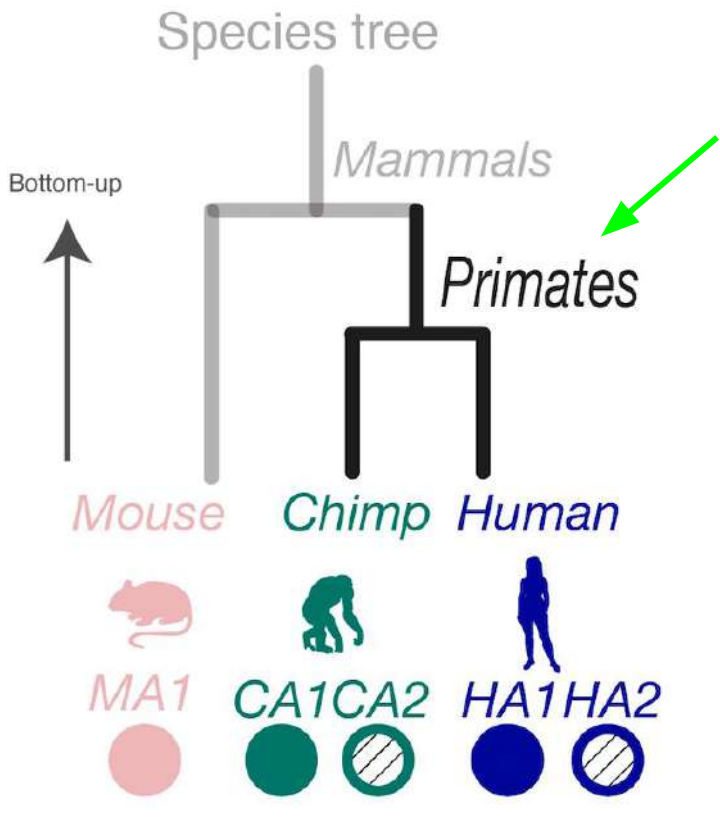
Hierarchical Orthologous
Groups (HOGs)



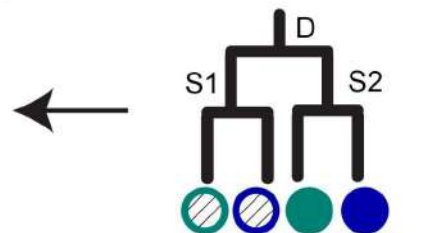
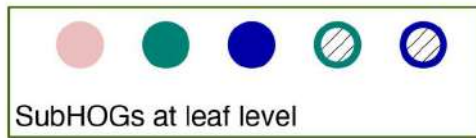
HOG inference



HOG inference



Merging subHOGs based on speciations



3. Events inference (Sp. or Dup.)

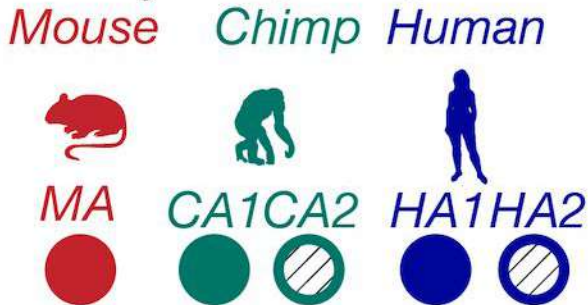
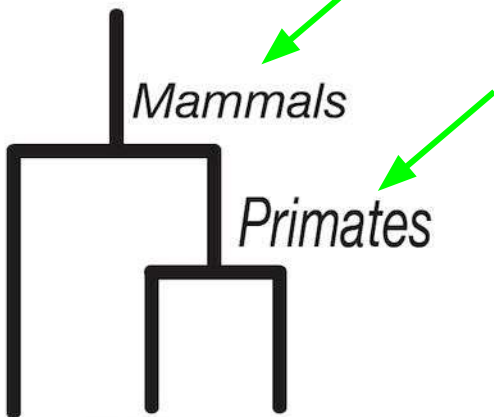
2. Gene tree inference and rooting

1. Multiple sequence alignment of proteins

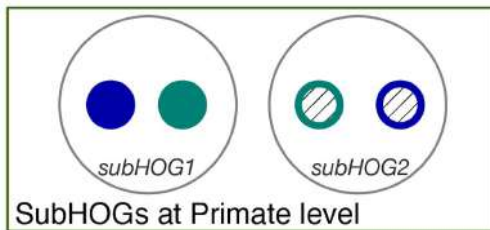


HOG inference

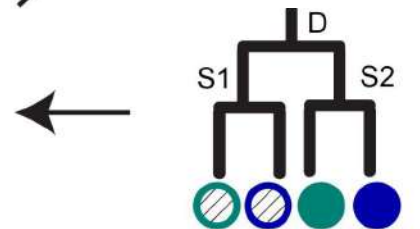
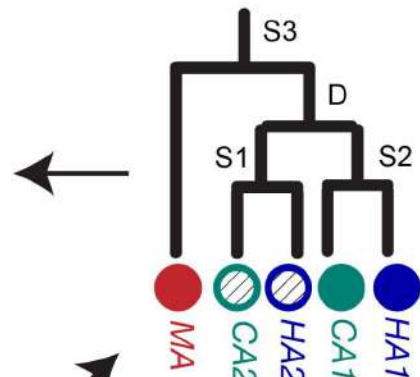
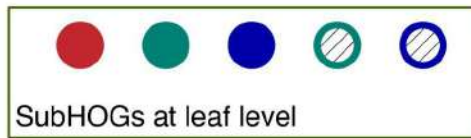
Species tree



Merging subHOGs based on speciations



Merging subHOGs based on speciations



3. Events inference (Sp. or Dup.)

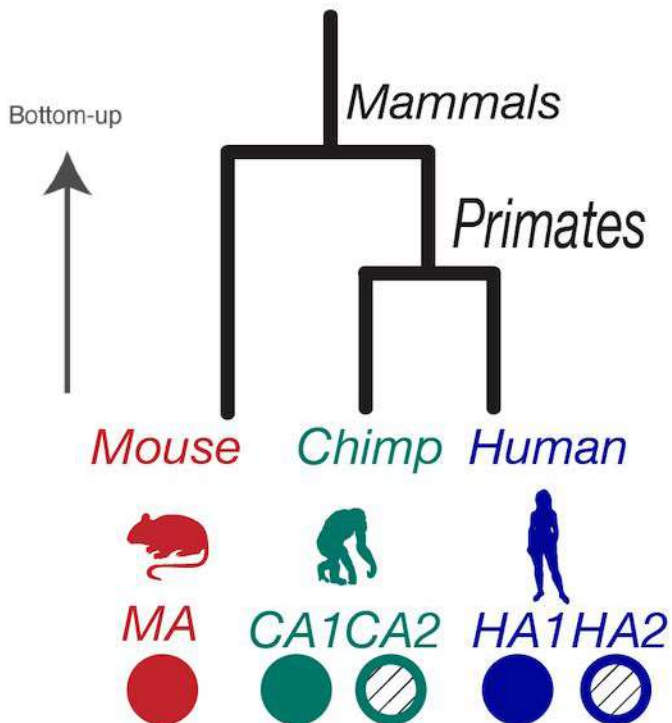
2. Gene tree inference and rooting

1. Multiple sequence alignment of proteins



HOG inference

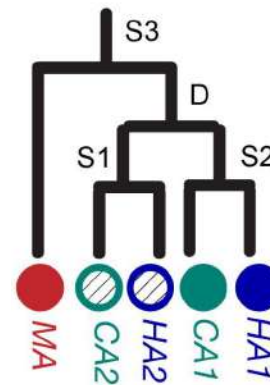
Species tree



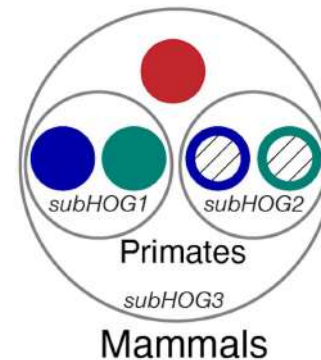
HOG in orthoXML

```

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

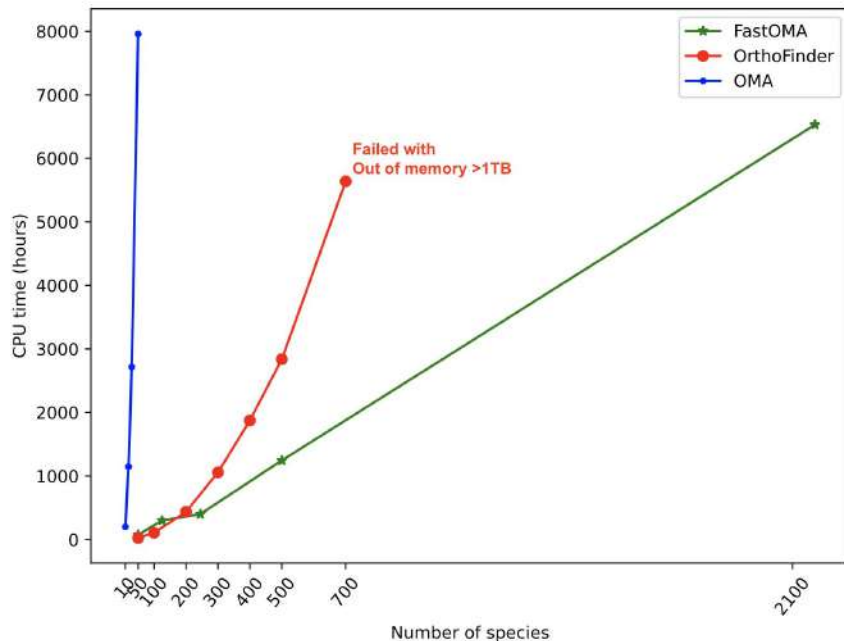


Nested structure of HOG



Orthology inference for Eukaryote dataset

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



[github.com/DessimozLab/
FastOMA](https://github.com/DessimozLab/FastOMA)

Module 3.3

- 4. How many Root HOGs are in the HOG file?

× Hint

× Answer



Each line in the output file denotes a gene family. After running, check the end of the file rootHOGs.tsv. Note that the indexing starts from 0.



There are 6793 rootHOG (gene families) in this file.

- 5. Consider the gene “60S ribosomal protein L15-A” in *Schizosaccharomyces pombe* with protein ID: RL15A_SCHPO. How many proteins are in the gene family (for these 5 species of interest)?