09:00-12:00 Genome annotation

- 09:00-09:30 Introduction to comparative genomics, and study system
    - Submit the first set of jobs
        i. Repeat mask
        ii. Mapping protein sets
        iii. *Ab initio* gene prediction
- 09:30-11:55 Introduction to genome annotation
    - Work through the rest of the programs
        i. EvidenceModeler
        ii. BUSCO
        iii. Functional annotation
- 11:55-12:00 Summary

12:00-13:00 Lunch

13:00-14:00 Comparative genomics

- Introduce Orthofinder
- Run Orthofinder

14:00-14:15 Break

14:15-16:00

- Basic visualization of Orthofinder results
- Introducing gene family analyses with CAFE5
- Running CAFE5

Genome assembly, annotation and comparative genomics

Day 2, afternoon

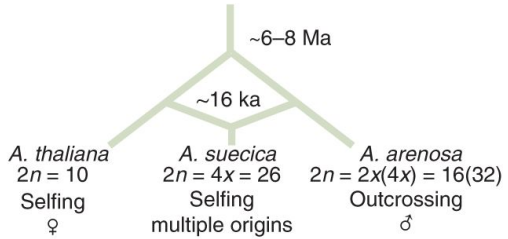
Teachers: Lars Grønvold, Thu-Hien To, Bram Danneels, Helle Tessand Baalsrud, Ole K. Tørresen

Norwegian Biodiversity & Genomics Conference 2024
9th April

Comparing gene content

- **Gene numbers can tell you something..**
  - E.g. genome duplications / polyploidizations

Burns et al. 2021



A. suecica

45k genes

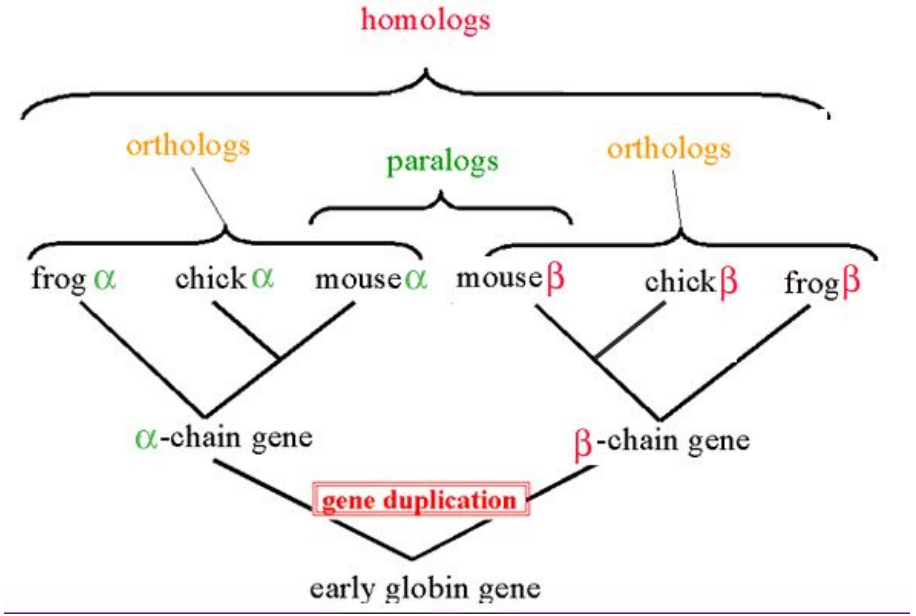
A. thaliana

22.5k genes

- ***Which genes* species share (orthologs), often more important**
  - The concept of orthology and homology

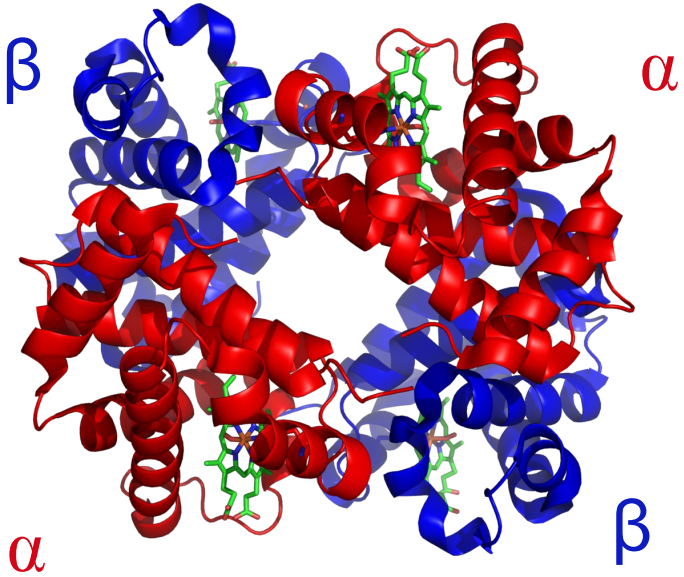
Detecting orthologs

- Identify gene duplications
- Reconstruct gene trees
- Infer natural selection on genes



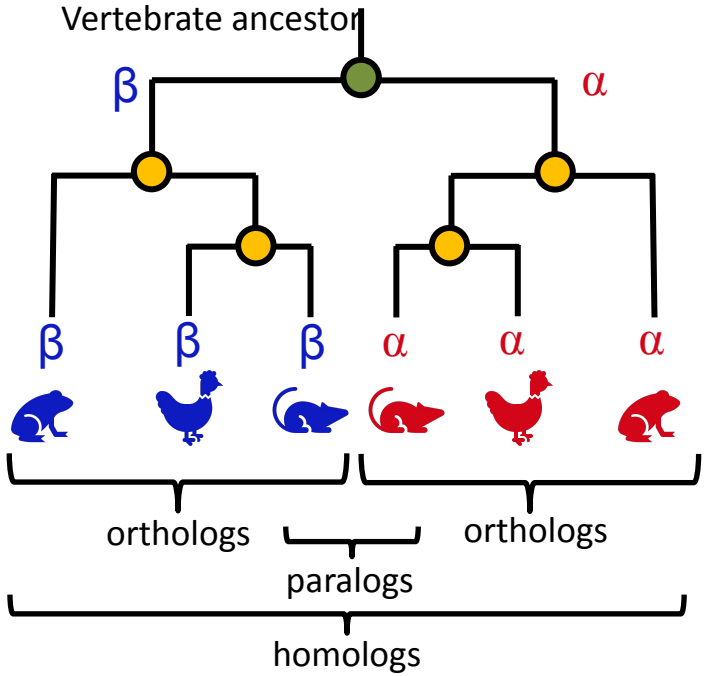
Speciation events

Homologs, parologs and orthologs



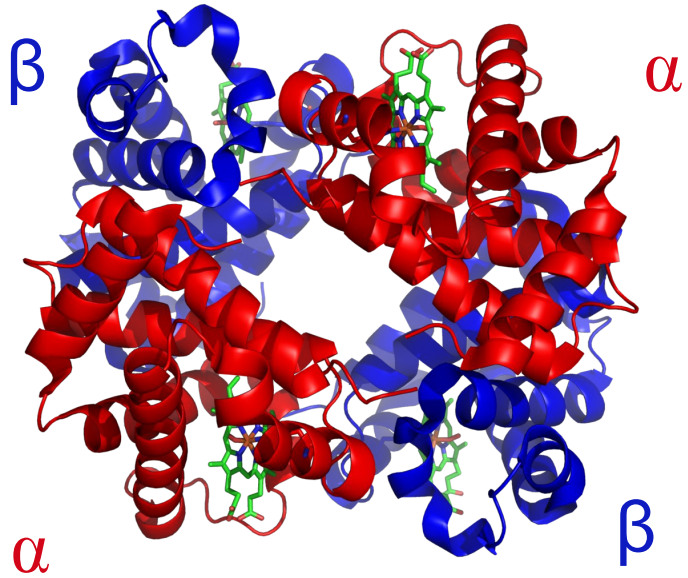
Hemoglobin tetramer

Homologs, parologs and orthologs



Gene duplication

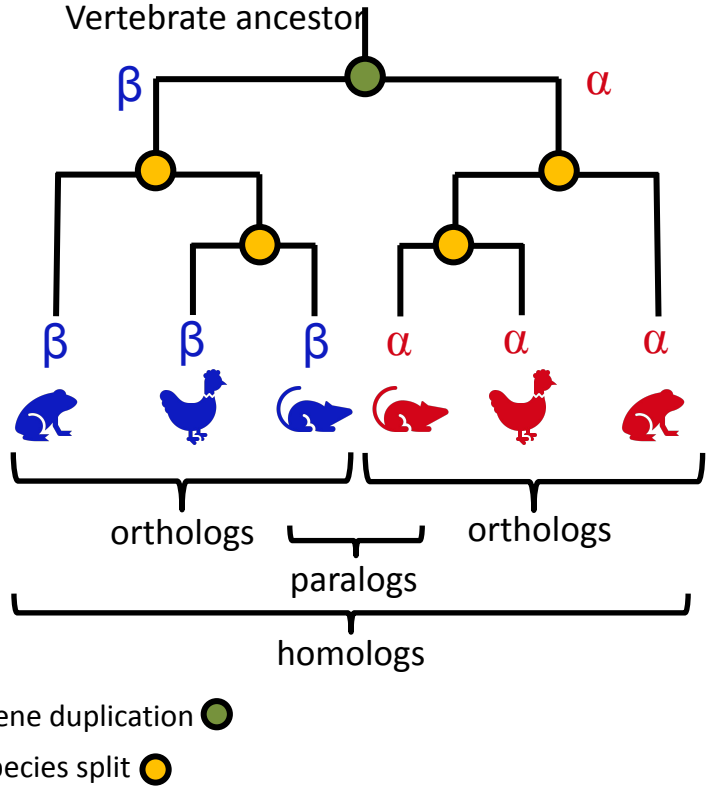
Species split



Hemoglobin tetramer

https://www.ncbi.nlm.nih.gov/books/NBK62051/

Homologs, parologs and orthologs



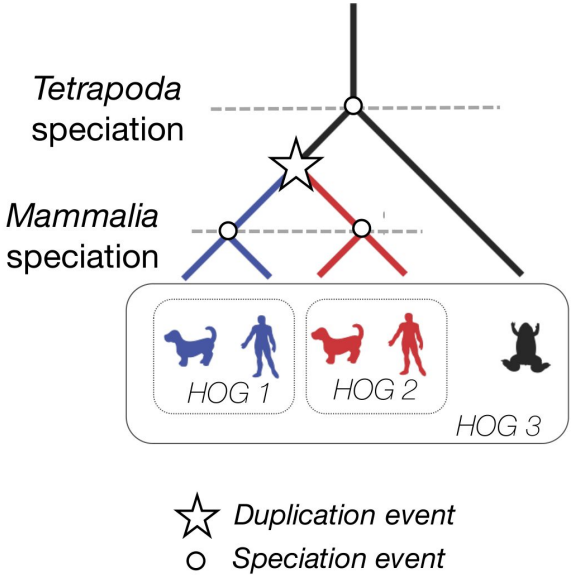
Homologs:
- Similar sequences
- Descent from a common ancestor
- May or may not have a similar function

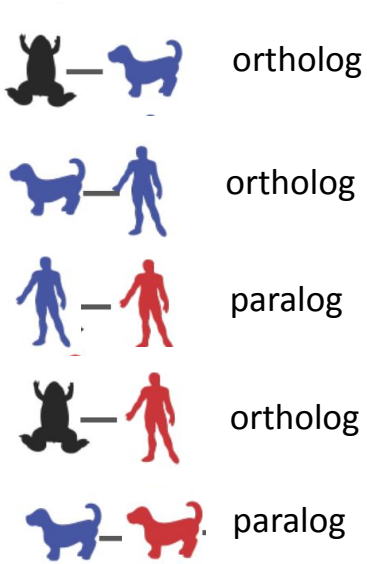
Paralogs:
- Homologous sequences within a species
- Arose by gene duplication
- May or may not have a similar function

Orthologs:
- Homologous sequences in different species
- Arose from a speciation event in a common ancestor
- May or may not have a similar function

https://www.ncbi.nlm.nih.gov/books/NBK62051/

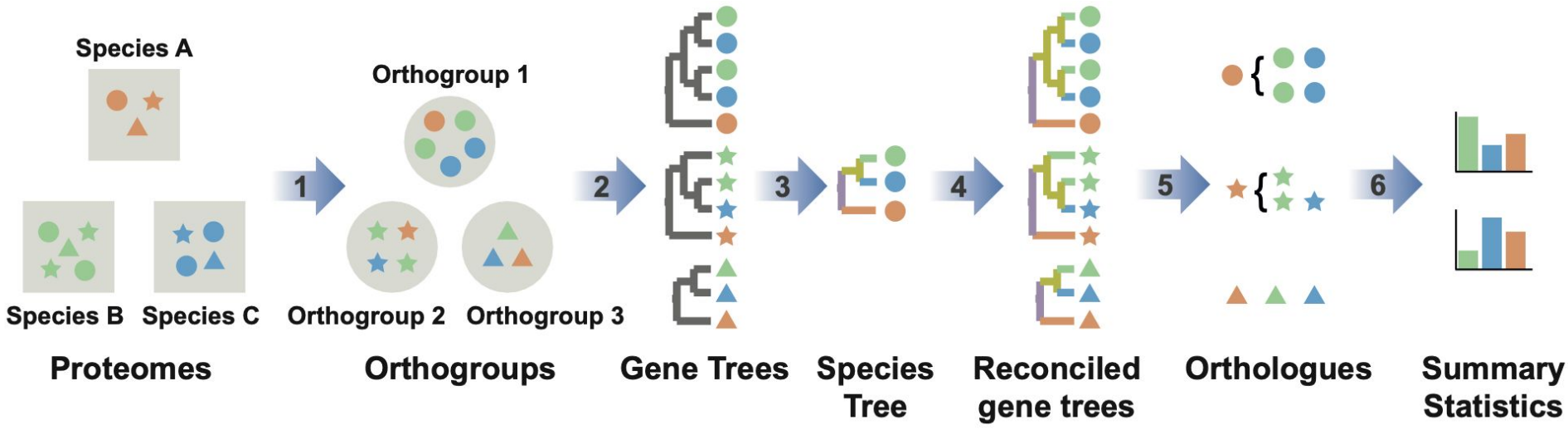
Homology: ortholog or paralog?



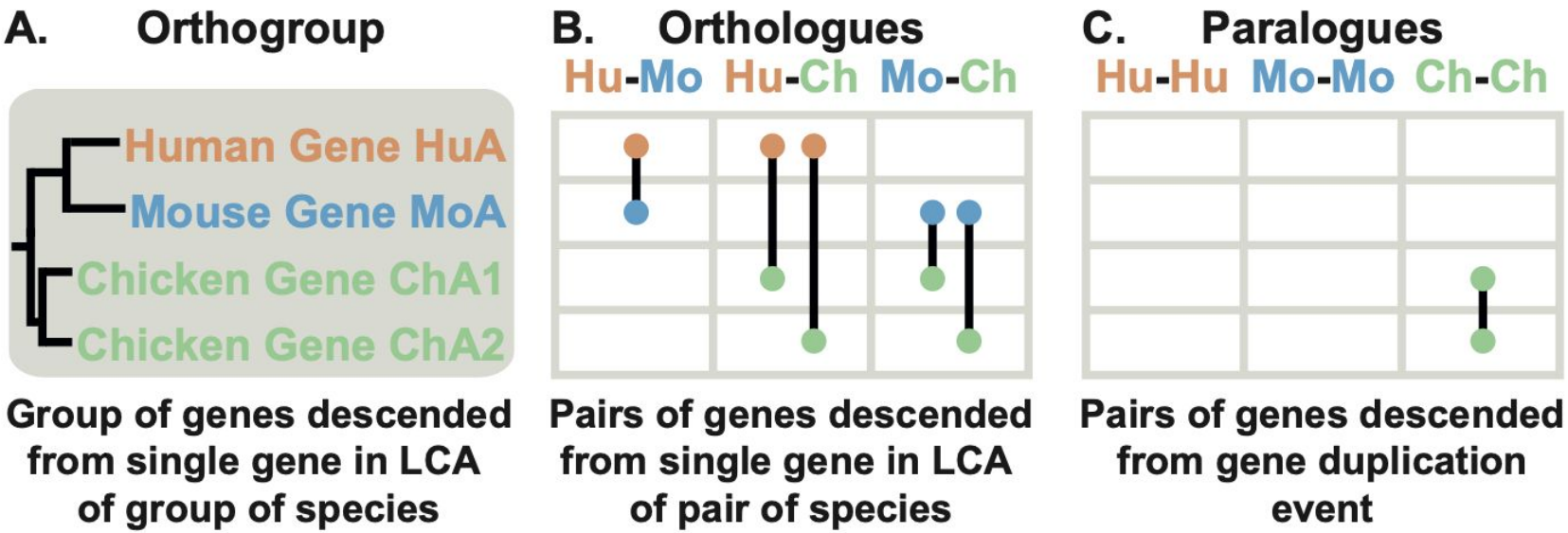


Orthofinder

Orthogroup: set of genes from multiple species descended from a single gene in the last common ancestor (LCA) of that set of species



Orthofinder



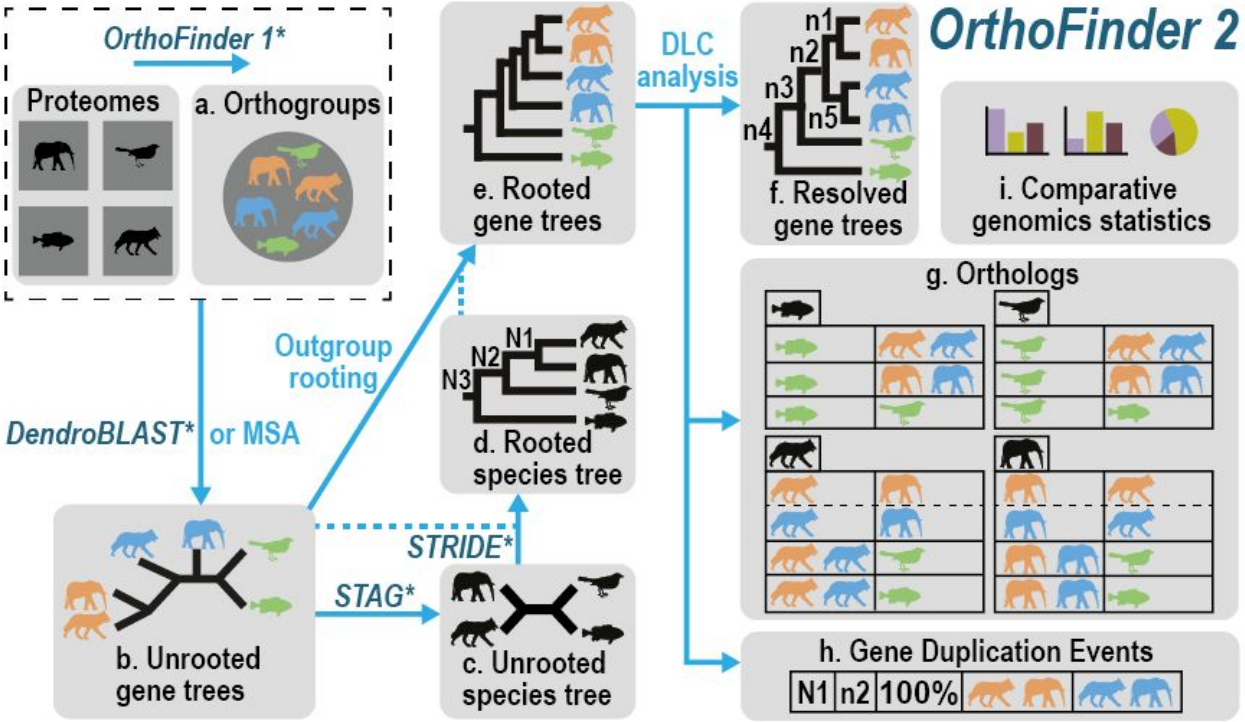
Orthofinder

A gene tree must be correctly rooted in order for it to show the correct evolutionary history of the gene family and thus to allow correct ortholog inference

Proteomes:
Fasta files with protein sequences for all genes, for each species

The Species Tree Root Inference from Duplication Events (STRIDE) algorithm

Species Tree from All Genes (STAG)



Orthofinder - resources

- Papers:

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1832-y

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-015-0721-2?ref=https://githubhelp.com

- Github:

https://github.com/davidemms/OrthoFinder Includes a talk you can watch to learn more!

- Tutorial:

https://davidemms.github.io/

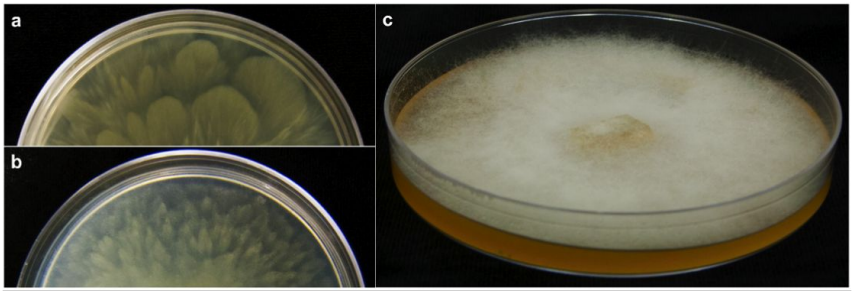
Orthofinder exercise

Objective:

Identifying orthogroups and determine the species tree for Mucoromycota fungi

Our data

- Annotated assemblies for 21 species
- Haploid genomes
- Small genomes: 22-60 Mb



Podila humilis

Outgroup species?

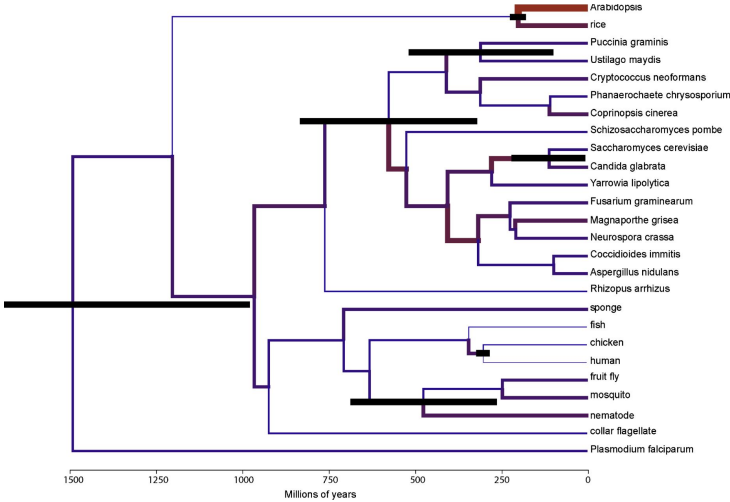
Conidiobolus coronatus

- Phylum Entomophthoromycota

Aspergillus nidulans

Saccharomyces cerevisiae

- Phylum Ascomycota
- Extensive functional annotation, model species
  - beneficial to interpret putative function of genes in orthogroups
- Divergence time from Mucoromycota is 750 MY
  - should add more outgroup species to ensure proper grouping of orthogroups



Outgroup species?

On the command line:

```
wget
http://ftp.ensemblgenomes.org/pub/fungi/release-54/fasta
/aspergillus_nidulans/pep/Aspergillus_nidulans.ASM1142v1
.pep.all.fa.gz -O AspNid.pep-transcripts.fa.gz

wget
http://ftp.ensemblgenomes.org/pub/fungi/release-54/fasta
/fungi_entomophthoromycota1_collection/conidiobolus_coro
natus_nrrl_28638_gca_001566745/pep/Conidiobolus_coronatu
s_nrrl_28638_gca_001566745.Conidiobolus_coronatus_NRRL28
638.pep.all.fa.gz
```

Outgroup species?

Advice:

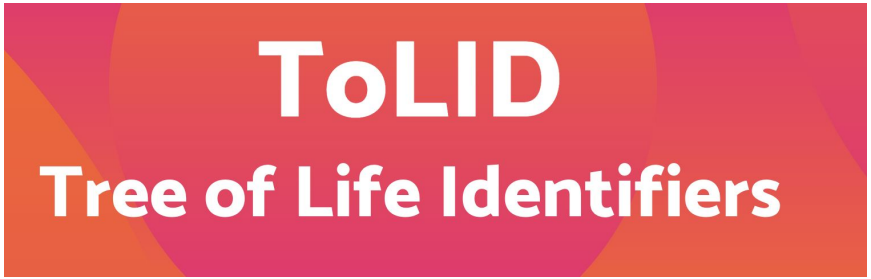
- pick species as close to your ingroup as possible
  - from public databases, or sequence it yourself
- include a species with extensive functional annotation, a “model species” close to your ingroup (if possible)

GENEONTOLOGY
Unifying Biology

Our data

We use tolIDs for our data

https://id.tol.sanger.ac.uk/



AspNid.pep-transcripts
ggMorZona1.proteins
gzLinHyal1.proteins
gzMorAlpi1.proteins
gzPodHuml1.proteins
gzUmbIsab1.proteins
gzUmbRama1.proteins
gzUmbVina1.proteins
.
.
.

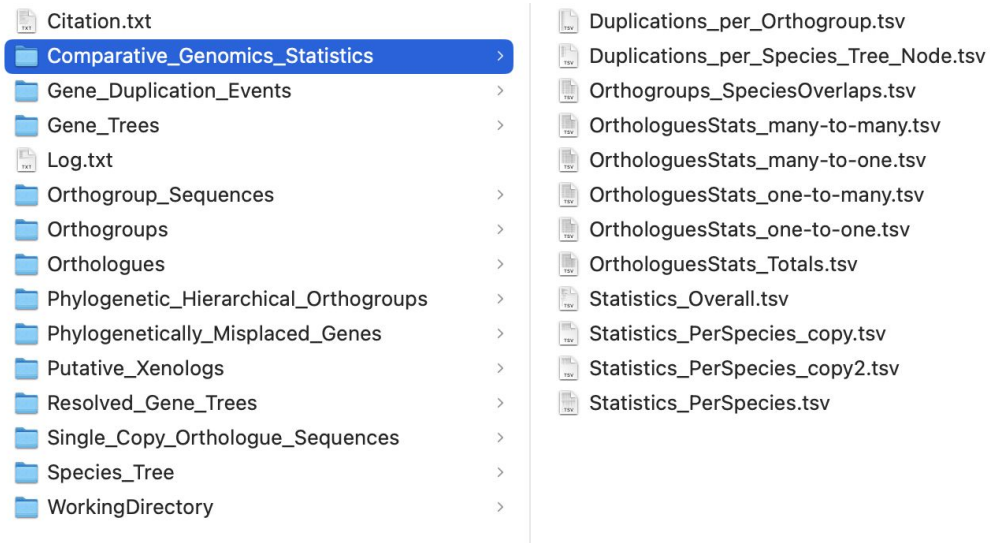
Run Orthofinder

https://github.com/ebp-nor/workshop-2024/blob/main/day2_genome_annotation/orthofinder.md

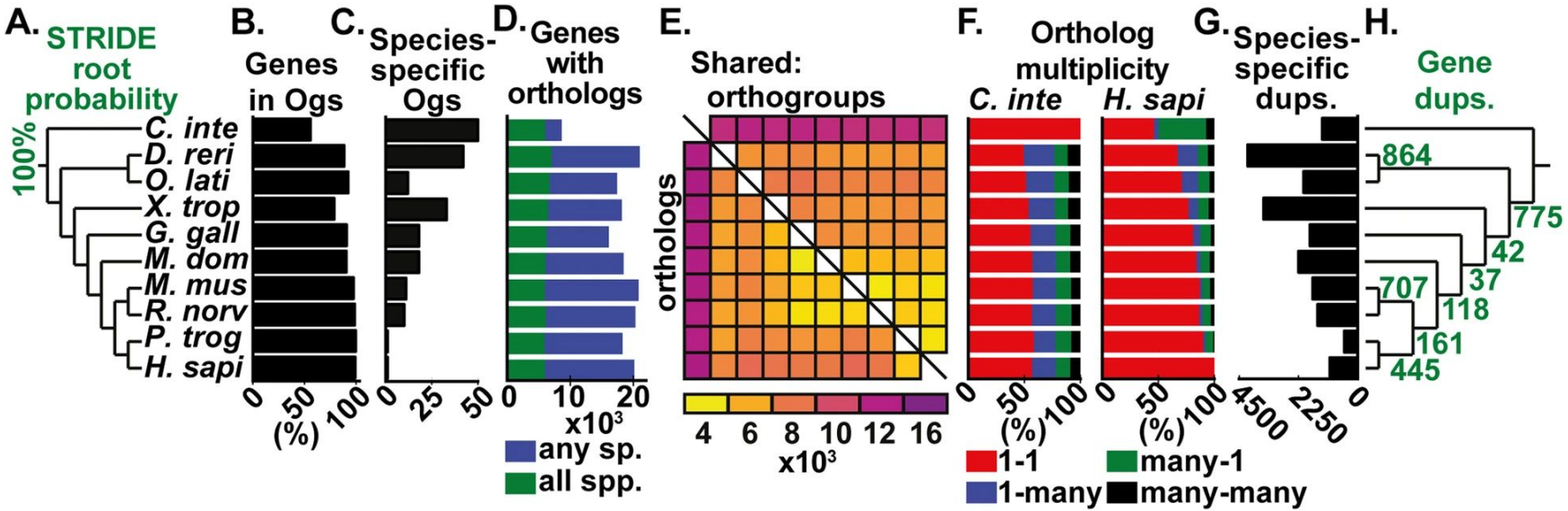
Orthofinder - exploring the results

- Hopefully by now you have some results!
  (If not, we will provide some;)

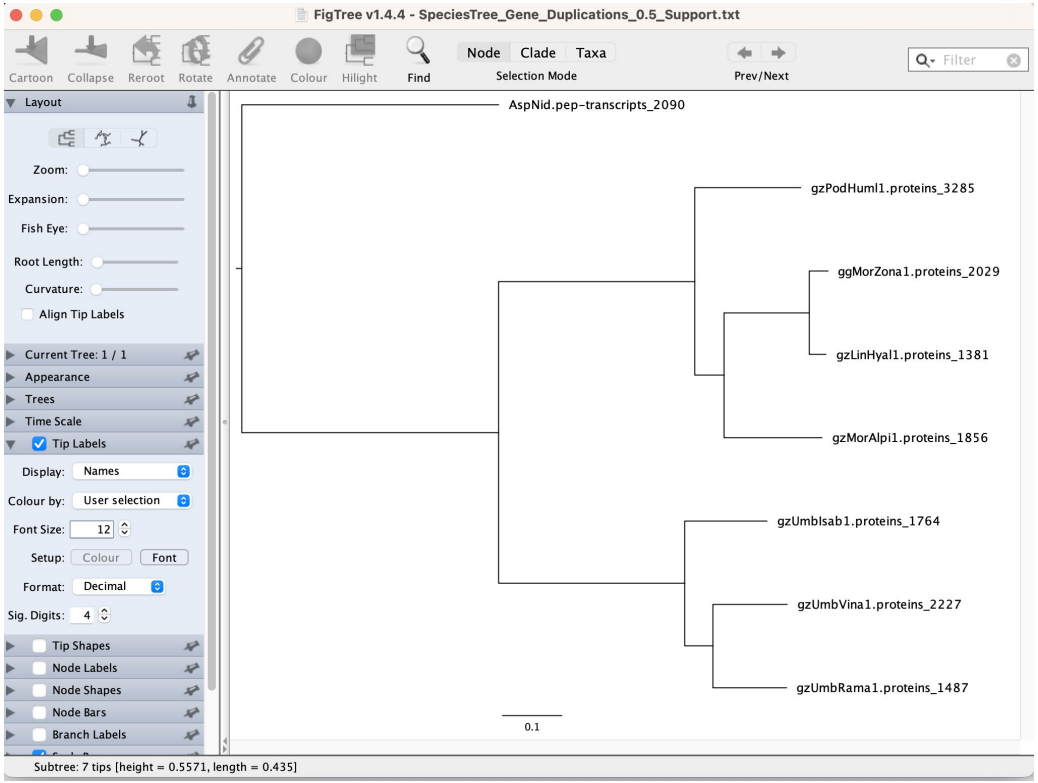
- But where do we go from here?

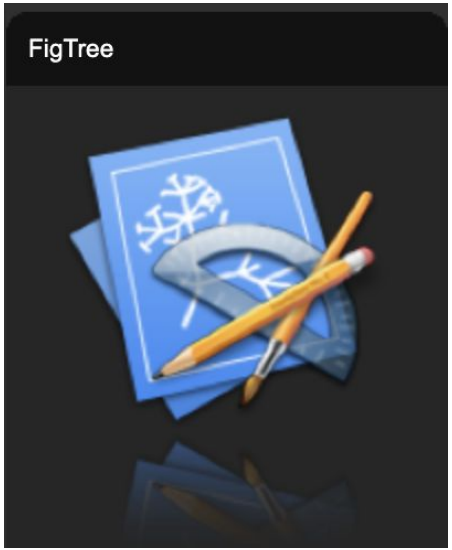


Orthofinder - exploring the results



Orthofinder - exploring the results

- 



Orthofinder - exploring the results

Objective: plot results from Orthofinder on a phylogenetic tree using R

(for quickly looking at trees I would still use Figtee, faster)

Icon for R

Icon for RStudio

Visualizing Orthofinder results

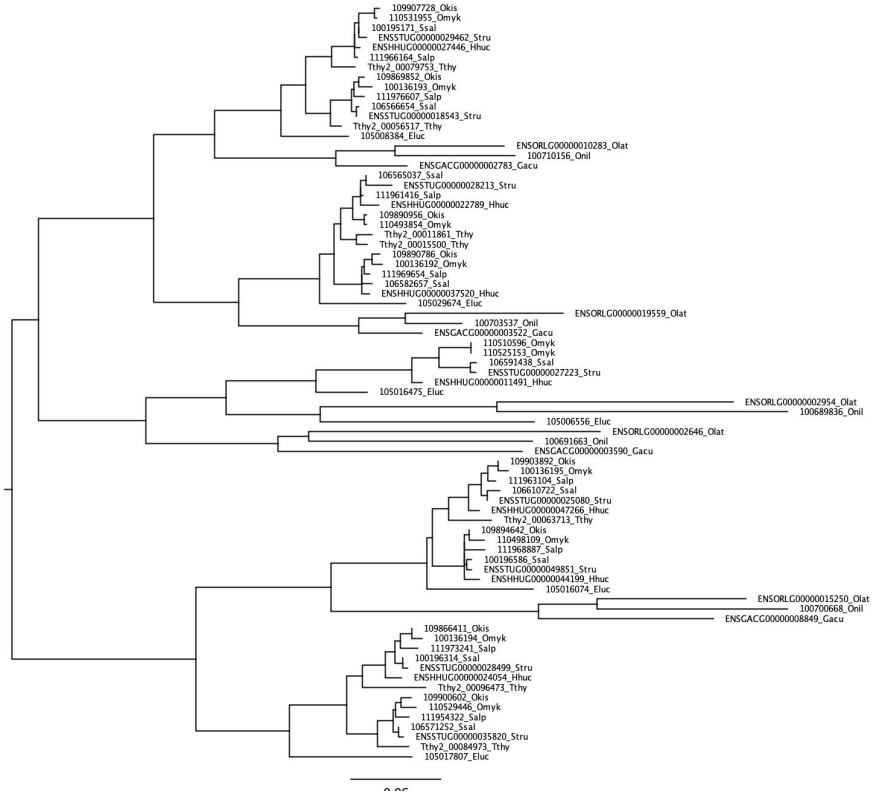
https://github.com/ebp-nor/workshop-2024/blob/main/day2_genome_annotation/Orthofinder_stats_2024.html

CAFE

Computational **A**nalysis of gene **F**amily **E**volution

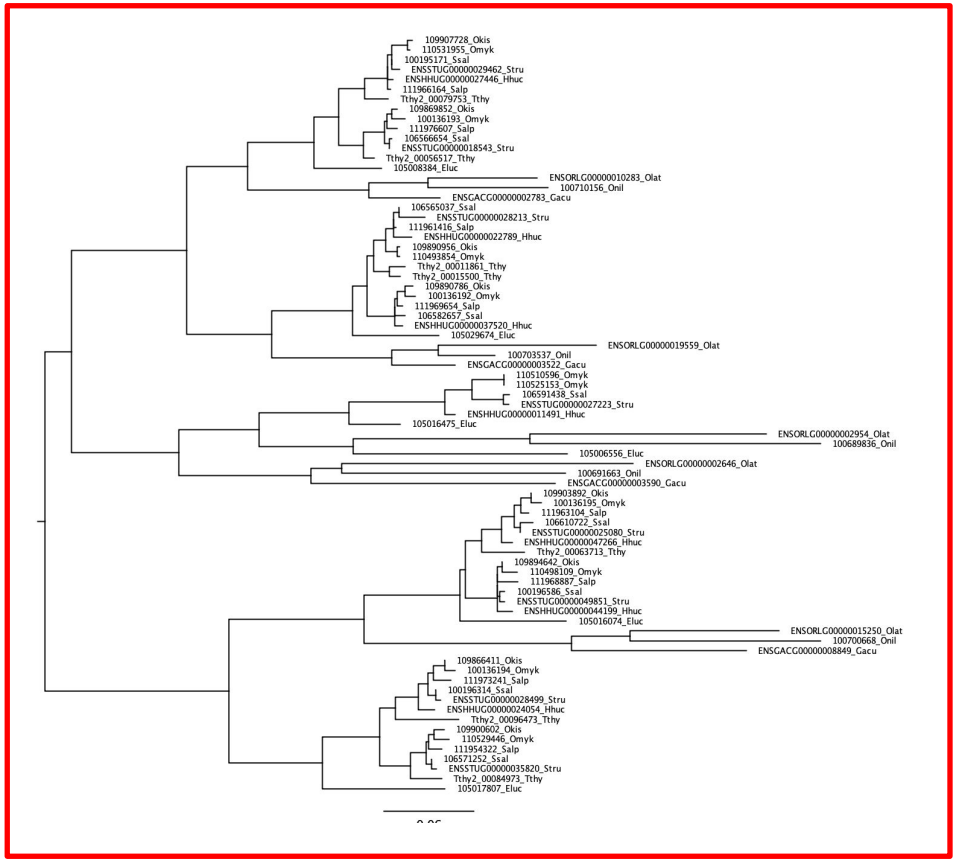
Gene family

Set of similar genes evolved from the same common ancestor gene



Gene family defined by orthofinder

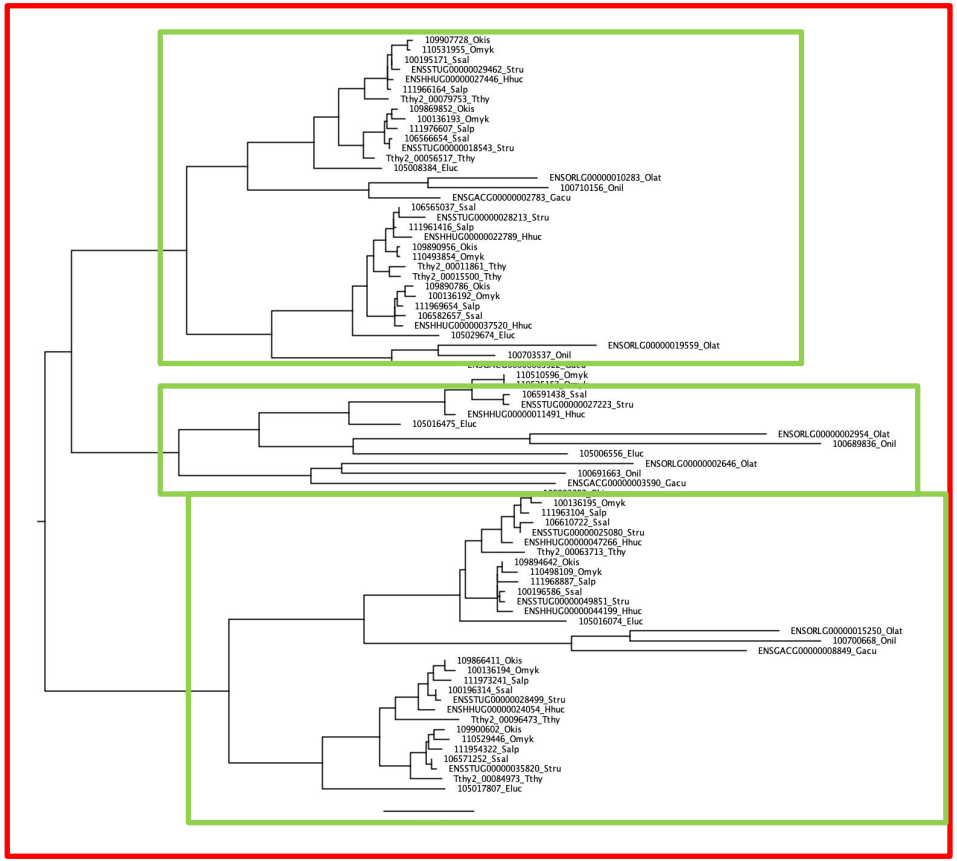
Set of similar genes evolved from the same common ancestor gene



OG: orthogroup

Gene family defined by orthofinder

Set of similar genes evolved from the same common ancestor gene

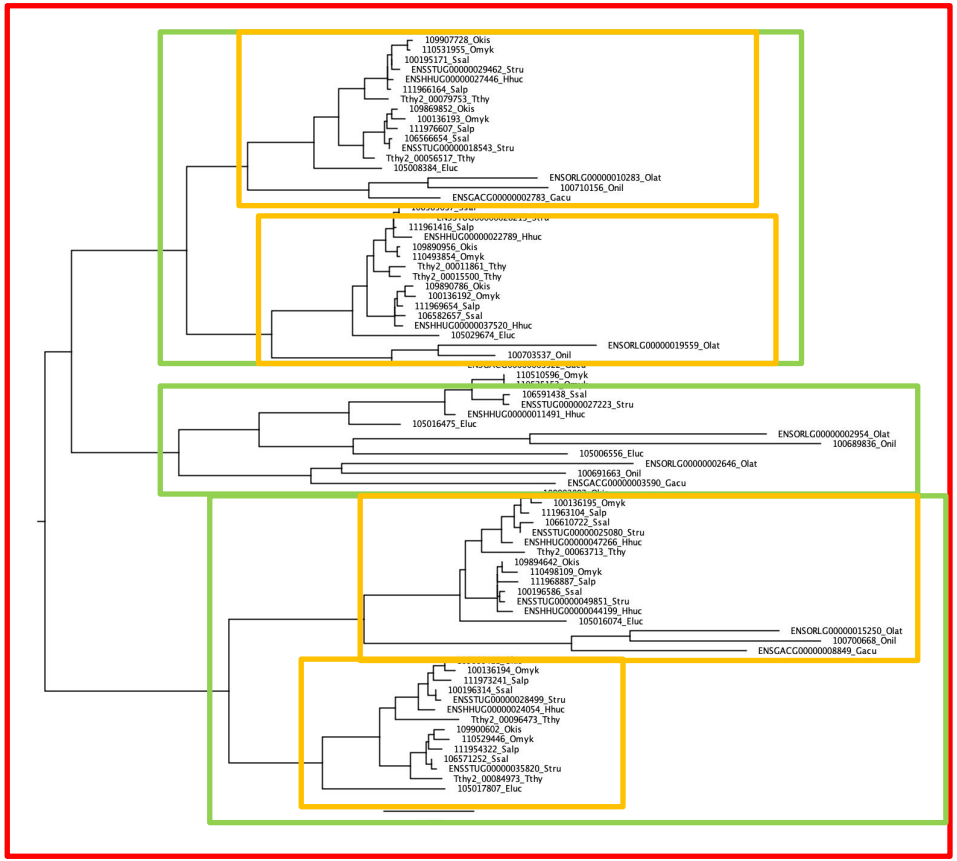


OG: orthogroup

HOG: N0

Gene family defined by orthofinder

Set of similar genes evolved from the same common ancestor gene



OG: orthogroup

HOG: N0

HOG: N1

Gene family defined by orthofinder

hierarchical orthogroups

orthogroups

gene family →

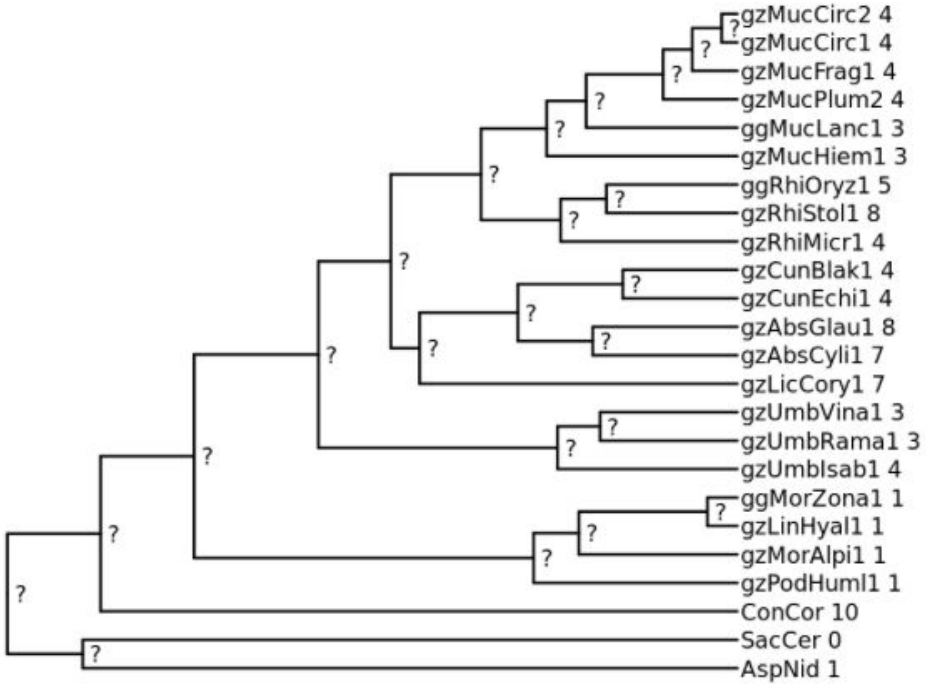
HOG	OG	AspNid.pep	ConCor.proteins	SacCer.proteins	ggMorZona1.proteins	ggMucLanc1.proteins	…
N0.HOG0000000	OG0000000	ANIA_0500 3		YPL230W, YMR182C	GGMOZO1EN_008077, GGMOZO1EN_001786	GGMULA1EN_008598, GGMULA1EN_000691, GGMULA1EN_000990, GGMULA1EN_007452	
N0.HOG0000001	OG0000000	ANIA_0125 1	CONCODRAFT_55634				
N0.HOG0000002	OG0000000						
N0.HOG0000003	OG0000000	ANIA_0619 5	CONCODRAFT_37485, CONCODRAFT_21796, CONCODRAFT_26205	YGL035C	GGMOZO1EN_000154, GGMOZO1EN_003935	GGMULA1EN_001958, GGMULA1EN_003830, GGMULA1EN_007797, GGMULA1EN_006125, GGMULA1EN_003642	
N0.HOG0000004	OG0000000			YER028C, YGL209W			
N0.HOG0000005	OG0000000						
N0.HOG0000006	OG0000000					GGMULA1EN_002316	
…							

Variation in the number of gene copies in each species for each gene family

HOG	OG	AspNid.pep	ConCor.proteins	SacCer.proteins	ggMorZona1.proteins	ggMucLanc1.proteins	…
N0.HOG0000000	OG0000000	1	0	2	2	4	
N0.HOG0000001	OG0000000	1	1	0	0	0	
N0.HOG0000002	OG0000000	0	1	0	0	0	
N0.HOG0000003	OG0000000	1	3	1	2	5	
N0.HOG0000004	OG0000000	0	0	2	0	0	
N0.HOG0000005	OG0000000	0	0	0	0	0	
N0.HOG0000006	OG0000000	0	0	0	0	1	
…							

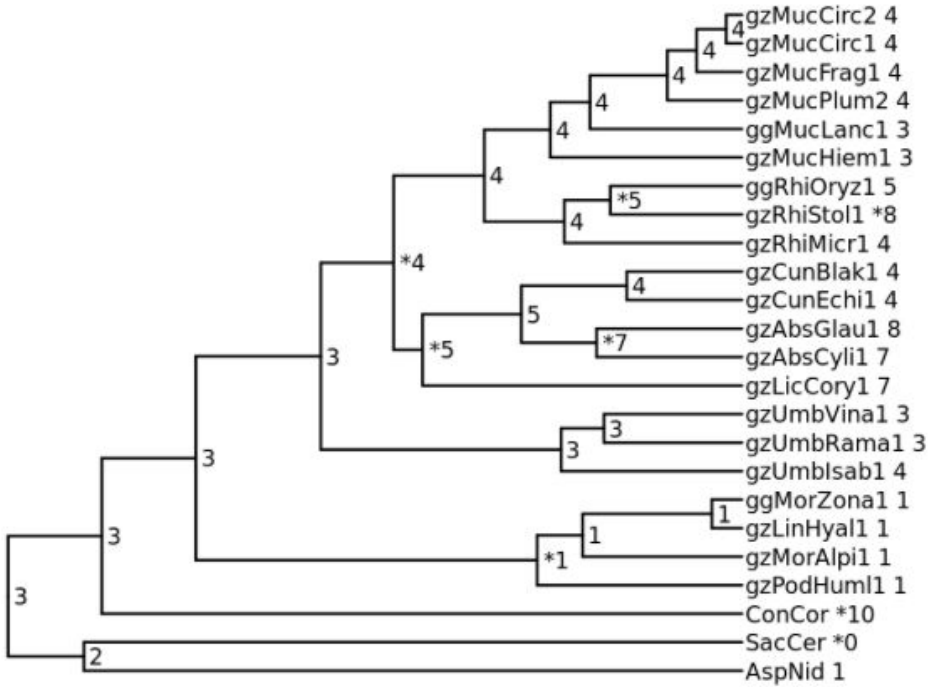
- How gene family sizes have evolved?
- Genes had been gained/lost randomly or under natural selection
- When the contractions/expansions happened?

CAFE estimates gene family size evolutions over a **phylogeny** using a **statistical model**



a gene family

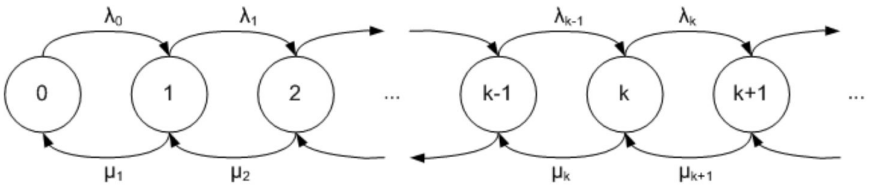
CAFE estimates gene family size evolutions over a **phylogeny** using a **statistical model**



a gene family

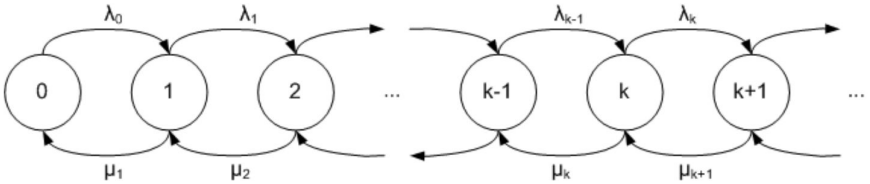
Statistical model

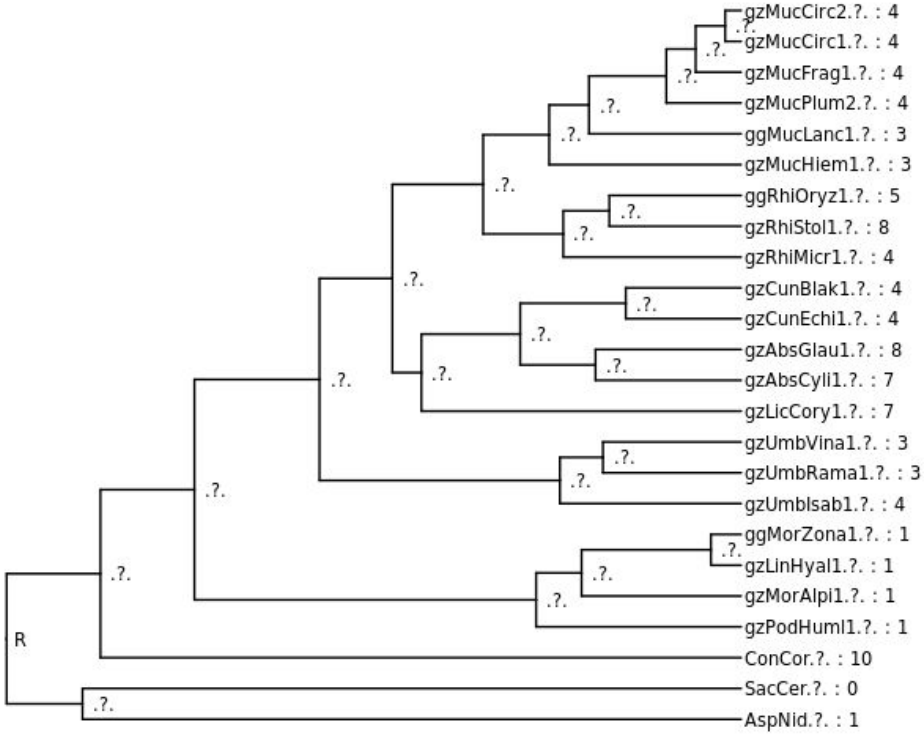
- Gene gain/loss process -> random **birth/death** model



the **rate** of gene gain/loss: $\lambda = \mu$

Conditional likelihood computation





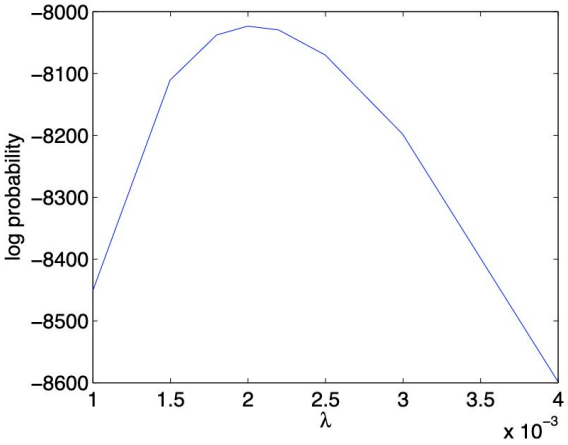
For each gene family, compute the likelihood of observing the gene family size for the leaf nodes, conditioned on the root size.

A likelihood function depends on :

- Tree toplogy
- branch lengths
- root size R
- gain/loss rate λ

Testing hypotheses about gene family evolution

Inferring λ



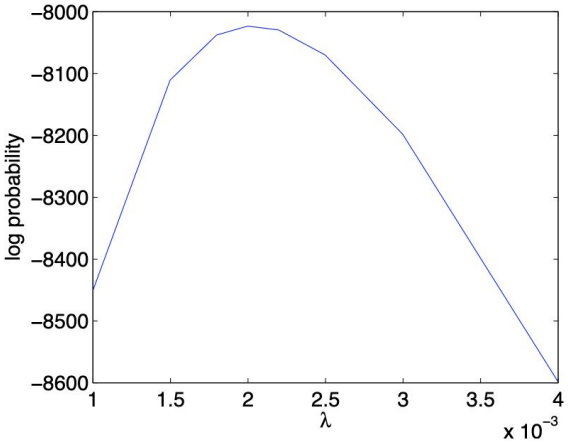
Maximum log likelihood

We have p-values conditioned on each value of the root node

- Choose the **largest** of these conditional p-values

- Significant p-value (ex. <0.05):
  - unlikely gene families
  - do not follow birth/death model
  - have undergone unusual (*natural selection*) expansions or contractions

Inferring λ



Maximum log likelihood

Testing hypotheses about gene family evolution

We have p-values conditioned on each value of the root node

- Choose the **largest** of these conditional p-values

- Significant p-value (ex. <0.05):
  - unlikely gene families
  - do not follow birth/death model
  - have undergone unusual (*natural selection*) expansions or contractions

Identifying the unlikely branches

If **removal** of 1 branch results in a large p-value (compared to a threshold), i.e. the remaining trees cannot reject the birth/death model, then this branch may be responsible for violating the model.

Several versions of CAFE

CAFE

CAFE2

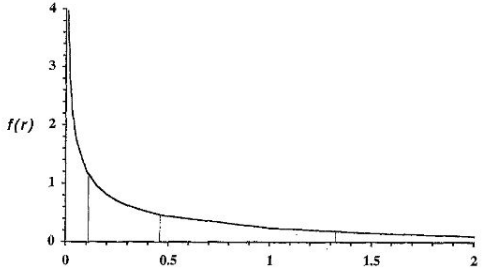
CAFE3,4

CAFE5

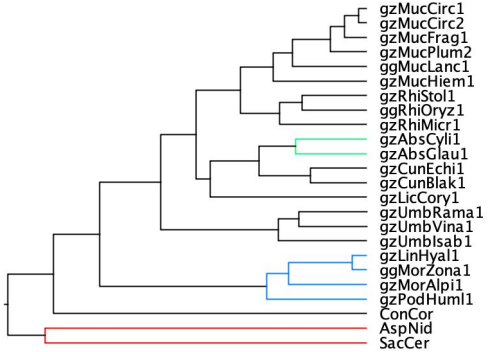
- Hahn, M. W., T. De Bie, J. E. Stajich, C. Nguyen, and N. Cristianini. 2005. Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Research* 15:1153–1160.
- De Bie, T., N. Cristianini, J. P. Demuth, and M. W. Hahn. 2006. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* 22:1269–1271.
- Hahn, M. W., J. P. Demuth, and S.-G. Han. 2007. Accelerated rate of gene gain and loss in primates. *Genetics* 177:1941–1949. Genetics.
- Han, M. V., G. W. C. Thomas, J. Lugo-Martinez, and M. W. Hahn. 2013. Estimating Gene Gain and Loss Rates in the Presence of Error in Genome Assembly and Annotation Using CAFE 3. *Mol. Biol. Evol.* 30:1987–1997.
- Fábio K Mendes, Dan Vanderpool, Ben Fulton, Matthew W Hahn, CAFE 5 models variation in evolutionary rates among gene families, Bioinformatics, 2020

Advanced features in the new version

- Allow **rate variation** among families using gamma-distributed rate categories



- Allow different rates for **different branches** *using user-defined branch partition*



- Account **errors** in gene family counts (*should give an error model for each gene family*)

Running cafe5 on the fungi data

- Go to github page: https://github.com/ebp-nor/workshop-2024/blob/main/day2_genome_annotation/CAFE5.md