

# Computational Analysis of gene Family Evolution

- 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.
- CAFE • 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.
- CAFE2 • 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*.
- CAFE3,4 • 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.
- CAFE5 • Fábio K Mendes, Dan Vanderpool, Ben Fulton, Matthew W Hahn, CAFE 5 models variation in evolutionary rates among gene families, *Bioinformatics*, 2020

Gene family defined by orthofinder

hierarchical orthogroups

orthogroups

gene family →

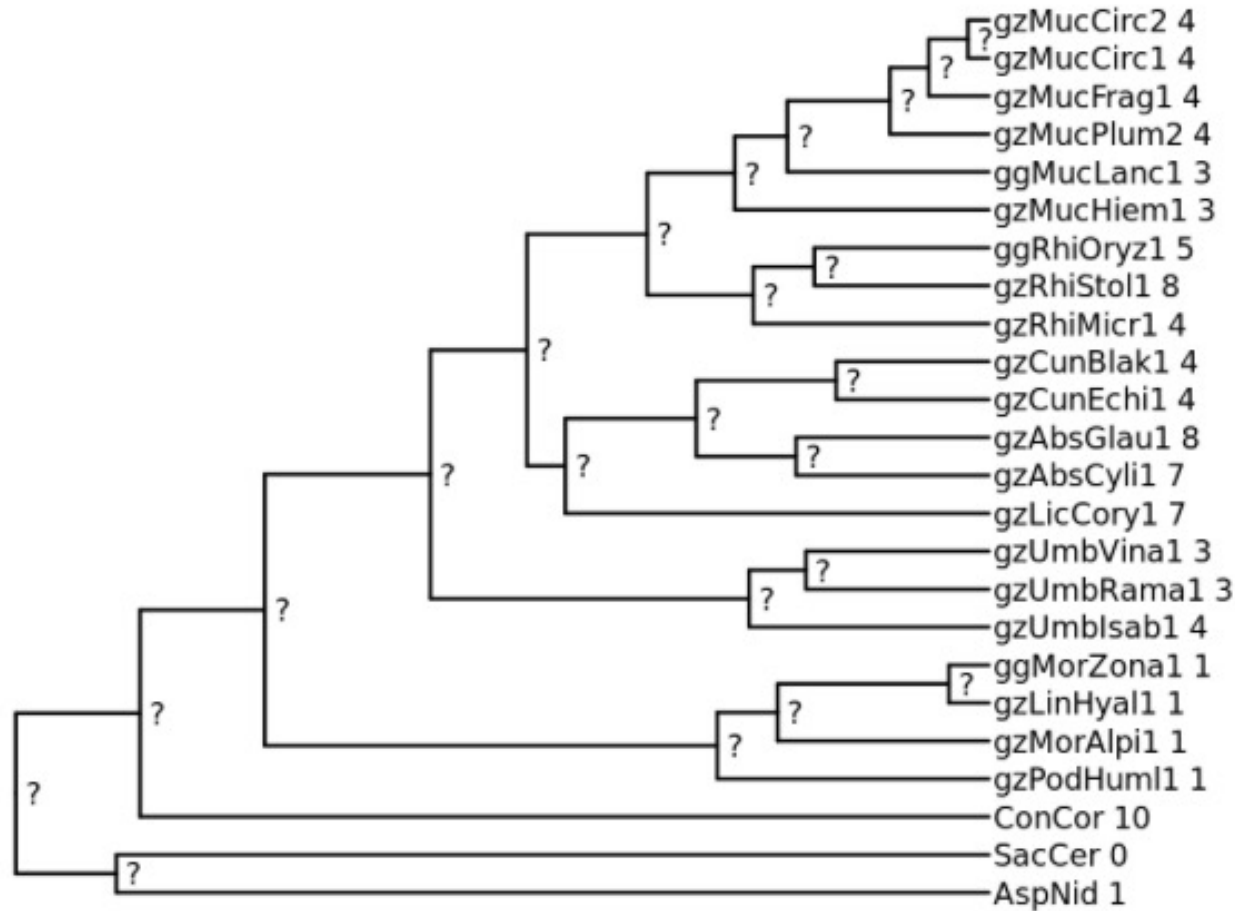
HOG	OG	AspNid.pep	ConCor.proteins	SacCer.proteins	ggMorZona1.proteins	ggMucLanc1.proteins	...
N0.HOG0000000	OG0000000	ANIA_05003		YPL230W, YMR182C	GGMOZO1EN_008077, GGMOZO1EN_001786	GGMULA1EN_008598, GGMULA1EN_000691, GGMULA1EN_000990, GGMULA1EN_007452	
N0.HOG0000001	OG0000000	ANIA_01251	CONCODRAFT_55634				
N0.HOG0000002	OG0000000						
N0.HOG0000003	OG0000000	ANIA_06195	CONCODRAFT_37485, CONCODRAFT_21796, CONCODRAFT_26205	YGL035C YER028C, YGL209W	GGMOZO1EN_000154, GGMOZO1EN_003935	GGMULA1EN_001958, GGMULA1EN_003830, GGMULA1EN_007797, GGMULA1EN_006125, GGMULA1EN_003642	
N0.HOG0000004	OG0000000						
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	
...							

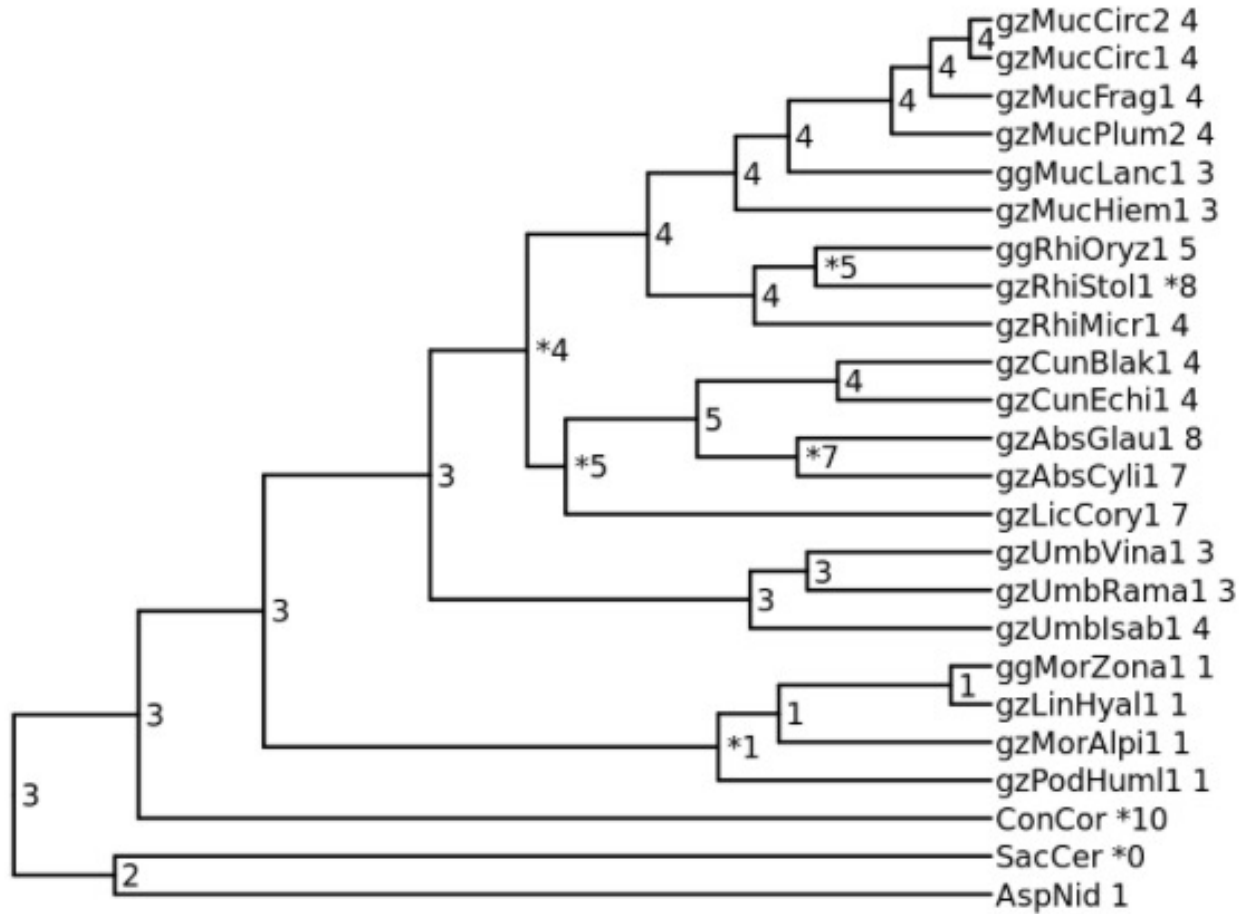
- How gene family sizes have evolved?
- Genes had been gained/lost randomly or under natural selection
- When the contractions/expansions happened?

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



a gene family

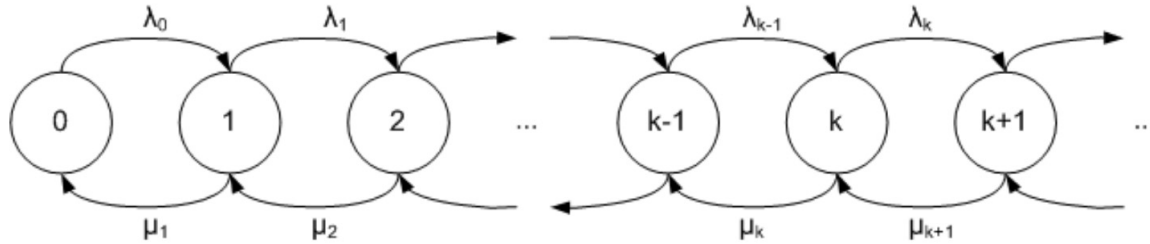
CAFE estimates gene family size evolutions over a **phylogeny** using a **statistical model**



a gene family

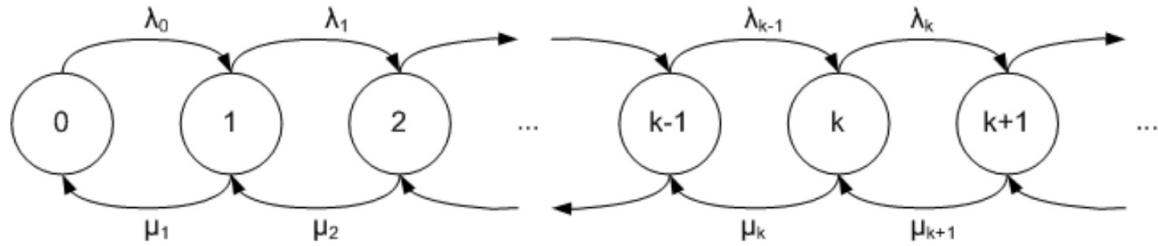
# Statistical model

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



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

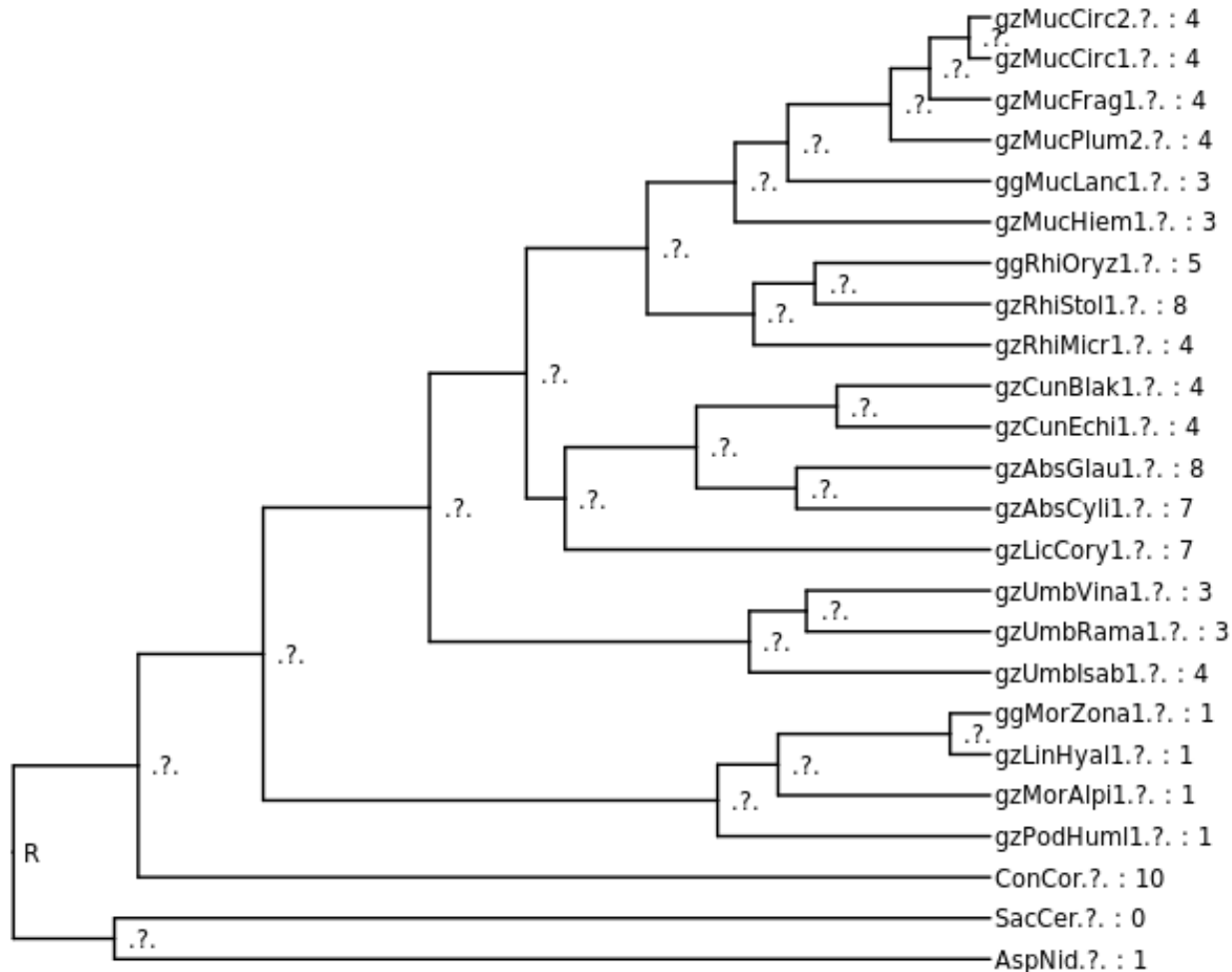
# Conditional likelihood computation



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

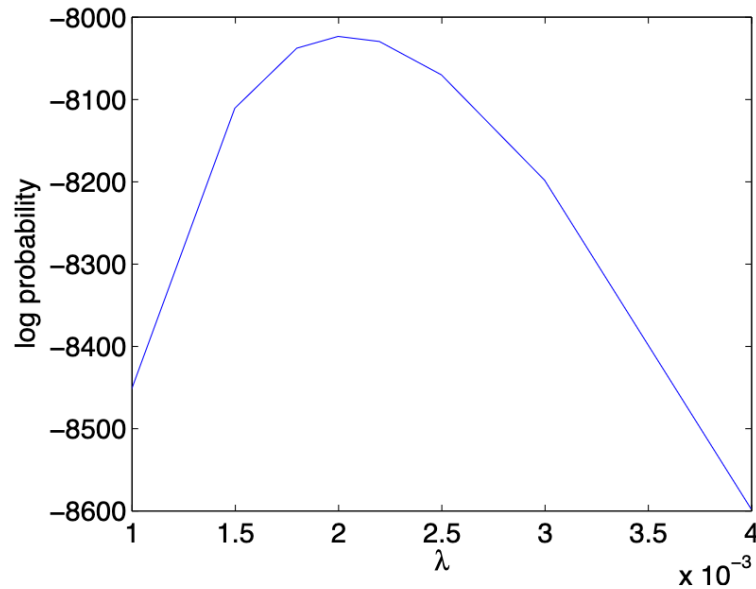
A likelihood function depends on :

- Tree topology
- branch lengths
- root size  $R$
- gain/loss rate  $\lambda$



# Testing hypotheses about gene family evolution

## Inferring $\lambda$



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.  $<0.05$ ):

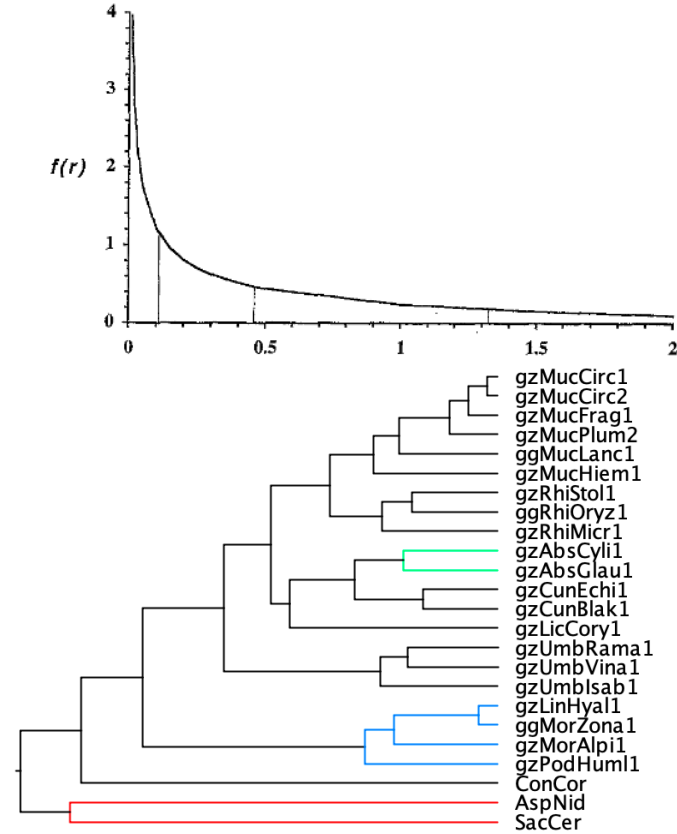
- unlikely gene families
- do not follow birth/death model
- have undergone unusual (*natural selection*) expansions or contractions

## 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.

# 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](#)