

# Triadas, lazos y funciones genéticas cuando las redes sociales y la filogenética se encuentran

George G Vega Yon, Ph.D.

University of Southern California, Department of Preventive Medicine

Seminario de Data Science UAI

19 de Mayo, 2021

Parsimonious modeling of gene functional evolution

A general framework for modeling functional evolution

GEESE vs aphylo

You can download the slides from [ggv.cl/slides/uai-ds-may](http://ggv.cl/slides/uai-ds-may)



National Cancer Institute Grant #1P01CA196596.



U.S. Army Research Laboratory and the U.S. Army Research Office under grant number W911NF-15-1-0577.

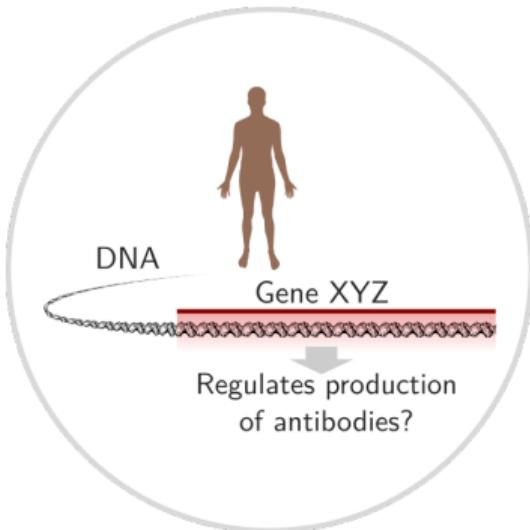


Advanced Research Computing  
Enabling scientific breakthroughs at scale

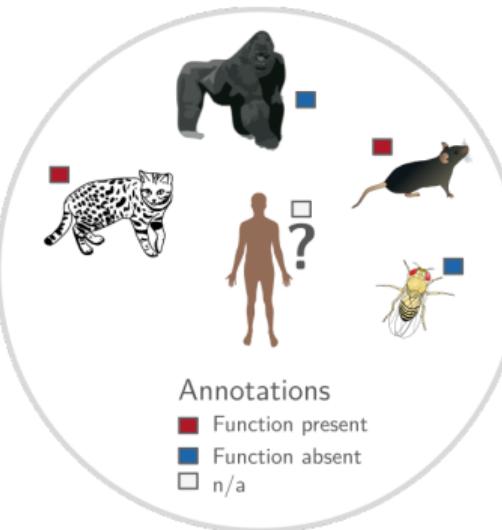
## Parsimonious modeling of gene functional evolution

*Joint with:* Paul D Thomas, Paul Marjoram, Huaiyu Mi, Duncan Thomas, and John Morrison  
(Published at *PLOS Computational Biology*)

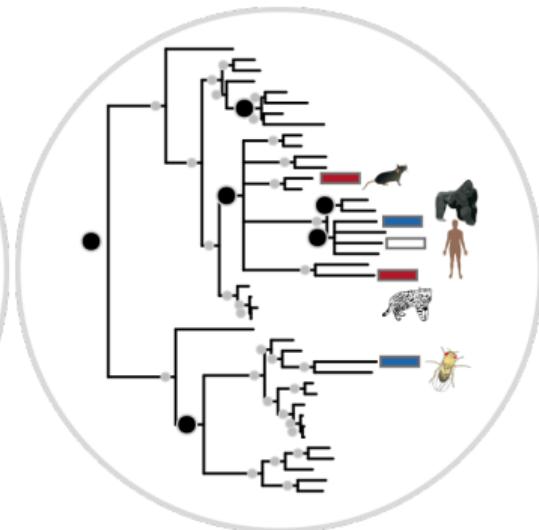
Is gene *XYZ* involved in process *ABC*?



Complex to directly assess



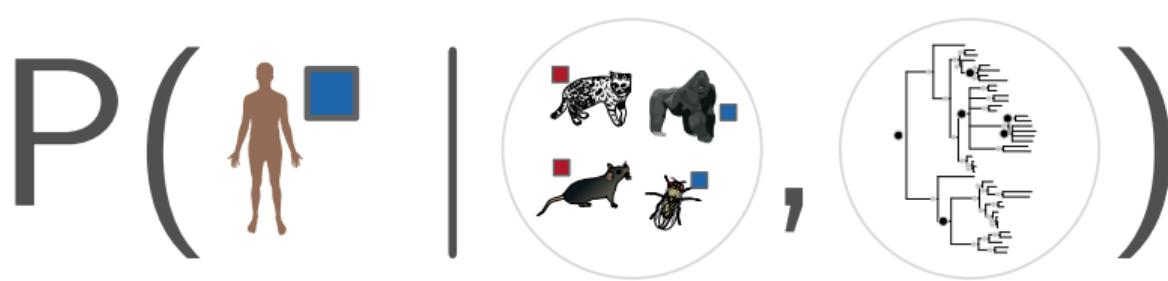
But we may know from other species



And we further know how these *genetically connected*

... let's rephrase the question.

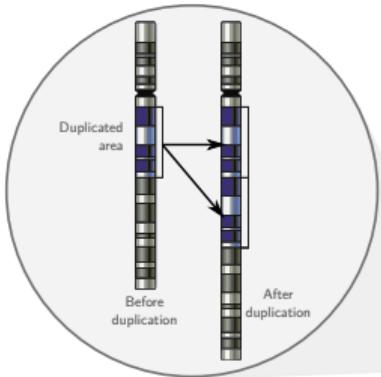
Is the human gene **XYZ** involved in process **ABC**, given what we know about that for other *related* species?



Annotations  
■ Function present  
■ Function absent  
□ n/a

▶ more

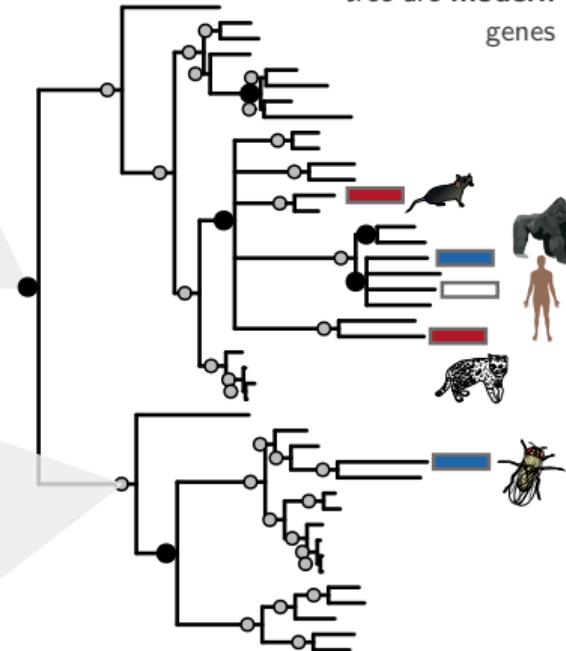
- nodes are Duplication Events



- nodes are Speciation Events



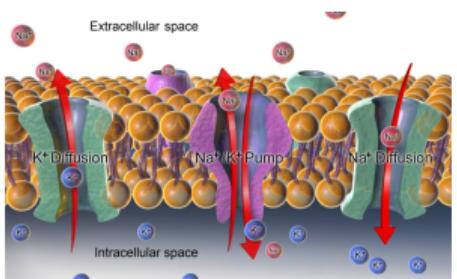
The tips of the tree are **modern** genes



Gene functions can be classified in three types:

## Molecular function

Active transport GO:0005215



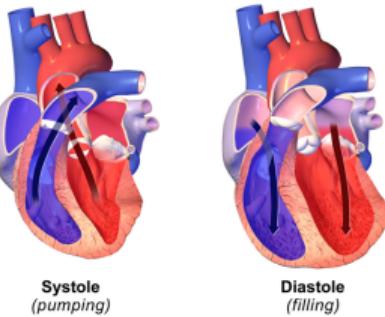
## Cellular component

Mitochondria GO:0004016

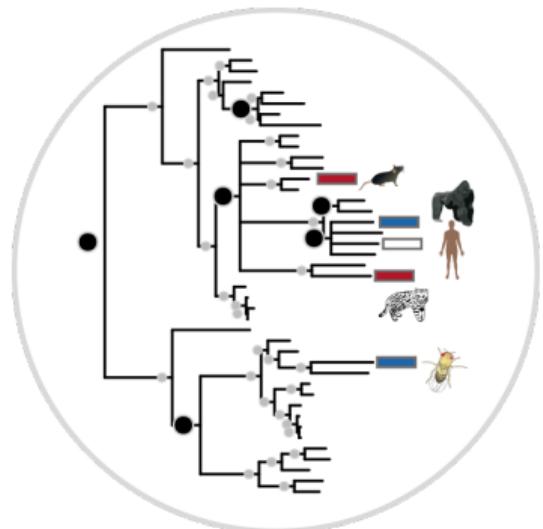


## Biological process

Heart contraction GO:0060047



... Where is all this data?



- ▶ ~ 15,000 phylogenetic trees
- ▶ ~ 8 million annotations
- ▶ ~ 600 thousand on human genes
- ▶ ~ < 10% are based on experimental evidence... Improving our knowledge on genetics is fundamental for advancing Biomedical Research

Only on 2021, 2,500+ Cancer papers using the GO (Google Scholar)

# An evolutionary model of gene functions

Family: PTHR11258

Type: Molecular Function

Name: 2'-5'-oligoadenylate synthetase activity

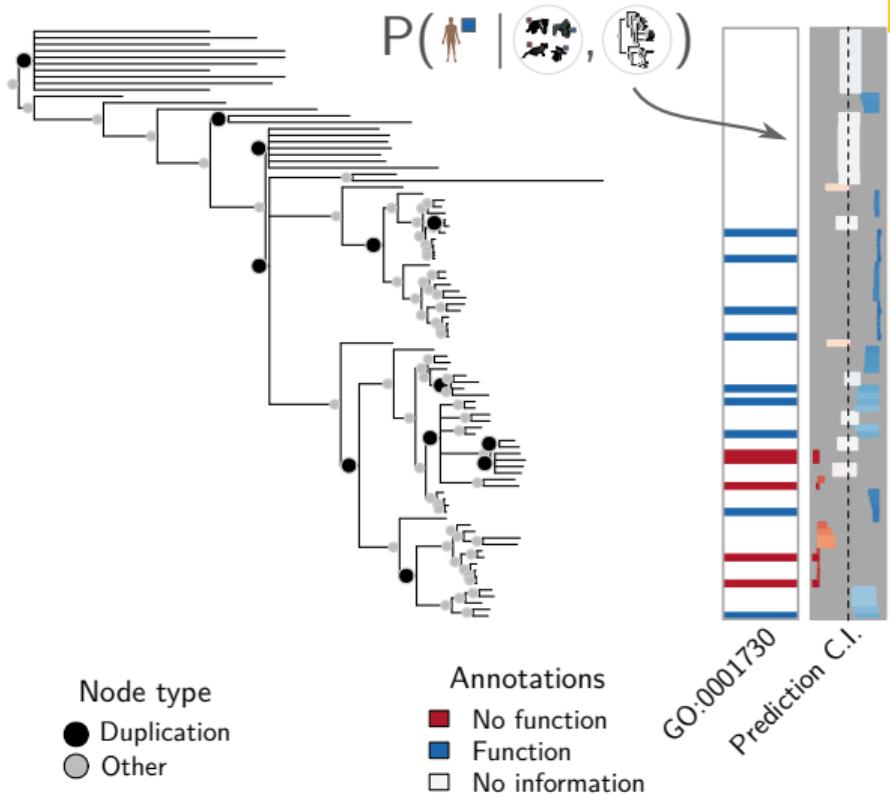
Desc: GO:0001730 involved in the process of cellular antiviral activity (wiki on [interferon](#)).

MAE: 0.34

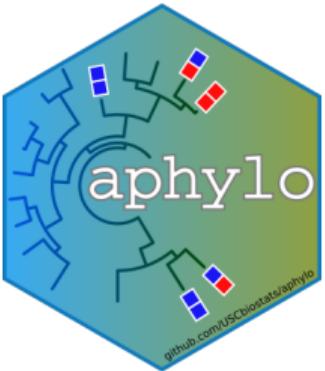
AUC: 0.91

I implemented this model in the **aphylo** R package

[see details](#)



## Results: What does aphylo brings to the table?



## Large scale

Estimate **pooled-data** models involving **hundreds of families** (1,300 genes at a time)

## Interpretable

Pooled-data model provides inference aligned with theoretical results (gene duplication is key)

## Fast

Computational efficiency allows making **inference and prediction fast** (1 second vs 2 hours)

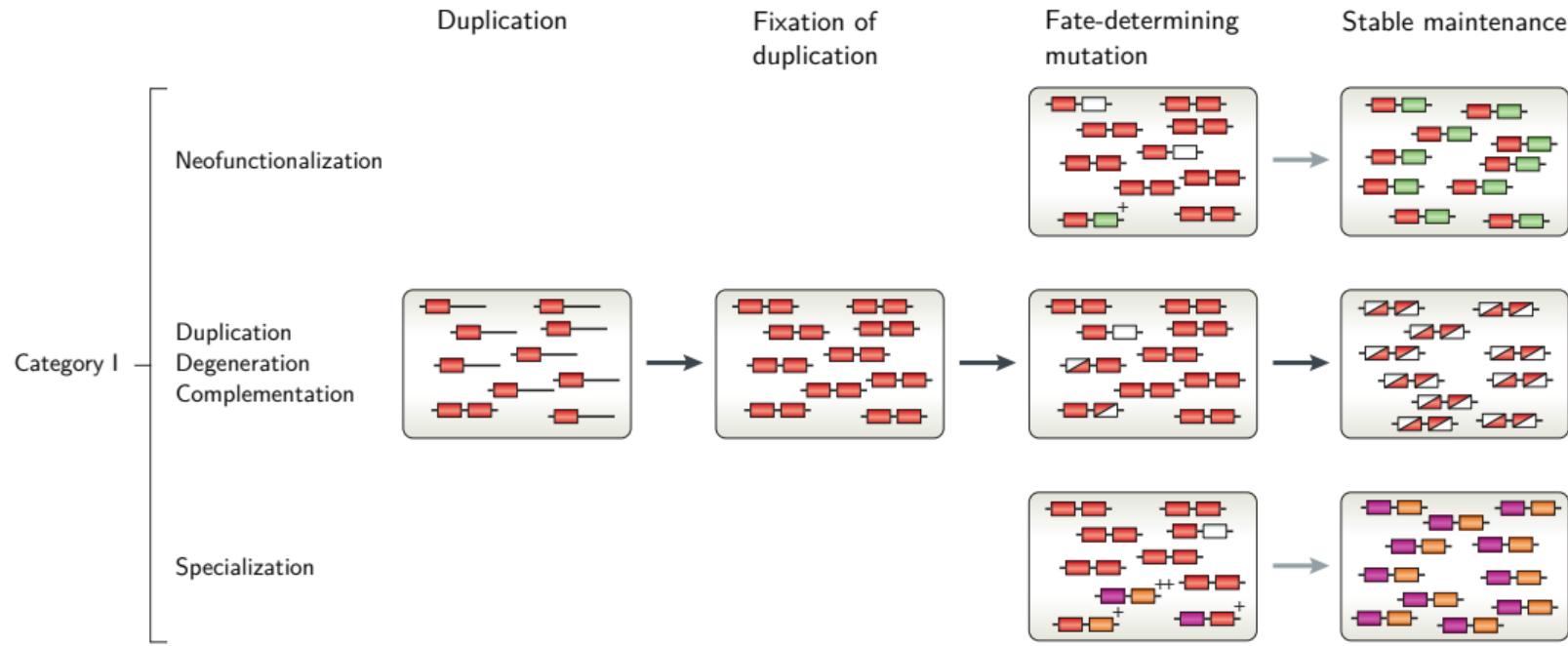
## Accuracy

Outperforms state-of-the-art phylo-models (0.72 vs 0.60 AUC)

▶ comp. feats

▶ details

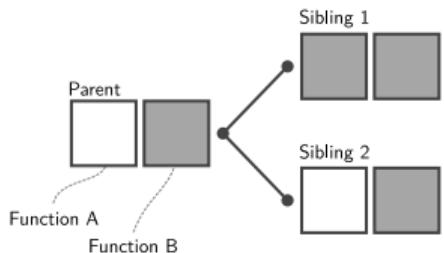
## A general framework for modeling functional evolution



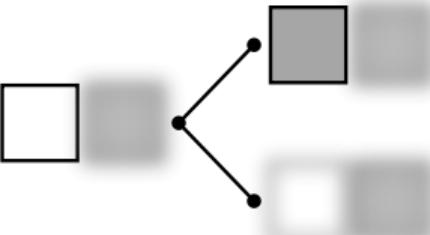
**Figure 1** Category I (*neutral fixation*). After the duplicated genes fixed, the new copy either: generate a new function, complements the original, or, if multiple functions originally present, specializes (including the original copy). Source: Adapted from Innan and Kod rashov (2010, nature rev. gen.)

▶ More

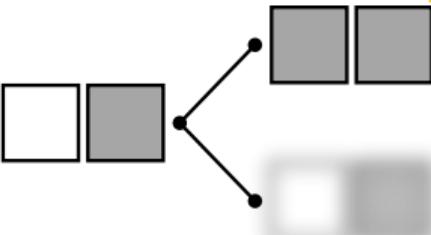
# Phylogenetics Modeling Strategies



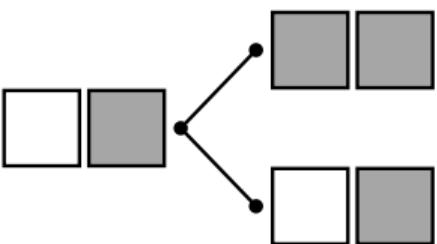
Legend:  
White square: Has the function  
Gray square: Doesn't have the function



(a) Sibling and Function Conditional Independence



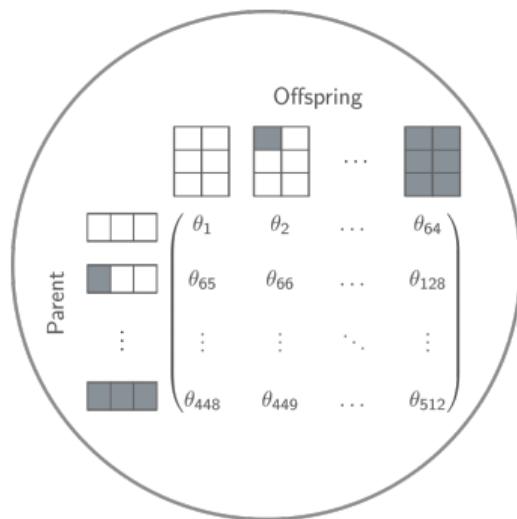
(b) Sibling Conditional Independence



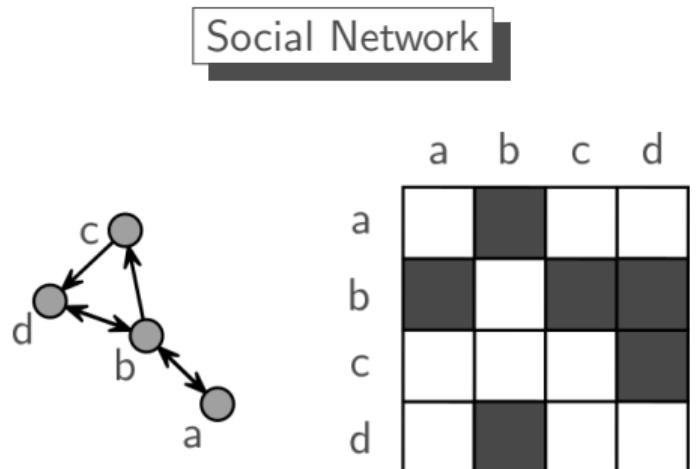
(c) No conditional independence

If we wanted to build a model with 3 functions, we would need to estimate...

### Full Markov Transition Matrix



- ▶ 512 parameters
- ▶ Finding this many parameters not easy.
- ▶ Even if you can, interpretation is awkward.

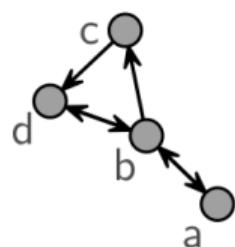


- ▶ Not about individual ties.
- ▶ Statistical inference on *motifs* (triangles, dyads, homophily, etc.)

Ultimately...

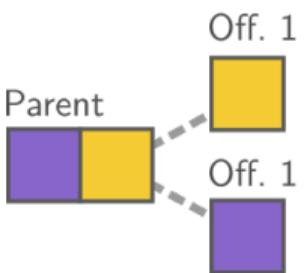
**ERGM**  $\equiv$  **Modeling binary arrays**

Social Network



	a	b	c	d
a	white	dark gray	white	white
b	dark gray	white	dark gray	dark gray
c	white	white	white	dark gray
d	white	dark gray	white	white

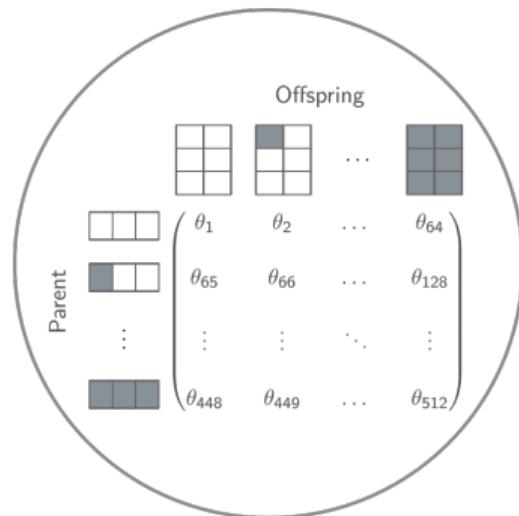
Evolutionary Event



Social Networks are usually represented as **adjacency matrices**, and so can evolutionary events!

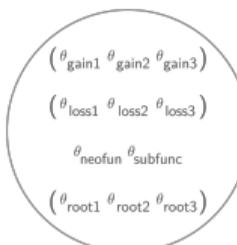
If we wanted to build a model with 3 functions, we would need to estimate...

### Full Markov Transition Matrix

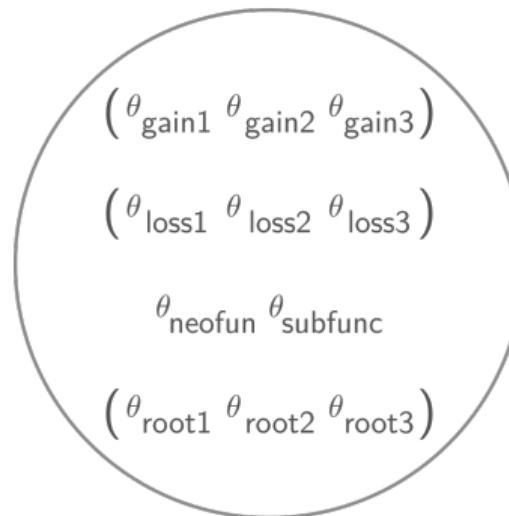


512 parameters

### Sufficient statistics

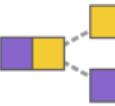
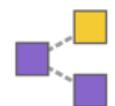
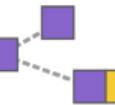


Easier to fit  
Easier to interpret



11 parameters (for example)

◀ numeric example

Representation	Description	Definition
	Gain of function	$(1 - x_p) \sum_{n:n \in Off} x_n$
	Loss of function	$x_p \sum_{n:n \in Off} (1 - x_n)$
	Subfunctionalization	$x_p^k x_p^j \sum_{n \neq m} x_n^k (1 - x_n^j) (1 - x_m^k) x_m^j$
	Neofunctionalization	$x_p^k (1 - x_p^j) \sum_{n \neq m} x_n^k (1 - x_n^j) (1 - x_m^k) x_m^j$
	Longest branch gains	$(1 - x_p^k) \mathbf{1} (x_m^k : m = \text{argmax}_n \text{blength}_n)$

**Table 1** Example of sufficient statistics for evolutionary transitions.

# Tree likelihoods: Felsenstein's Pruning algorithm

All possible transitions from  $\mathbf{x}_n$

Transition Probability (ERGM)

$$\mathbb{P}(\tilde{D}_n \mid \mathbf{x}_n, \Theta) = \sum_{\mathbf{x}} \mathbb{P}(\mathbf{x} \mid \mathbf{x}_n) \prod_{m \in O(n)} \mathbb{P}(\tilde{D}_m \mid \mathbf{x}_m)$$

Model Parameters

Vector of Sufficient Statistics

$$\mathbb{P}(\mathbf{x} \mid \mathbf{x}_n) = \frac{\exp \{ \Theta^t s(\mathbf{x}, \mathbf{x}_n) \}}{\sum_{\mathbf{x}'} \exp \{ \Theta^t s(\mathbf{x}', \mathbf{x}_n) \}}$$

Normalizing Constant

the *lingua franca* of SNA

It's parent state given the data

Gene state given the data

ways to get there

$$\mathbb{P}(x_p^p \mid \tilde{D}) = \sum_{x_p} \mathbb{P}(x_p \mid \tilde{D}) \left\{ \sum_{\{x^p: x_n^p = x\}} \mathbb{P}(x^p \mid x_p) \right\}$$

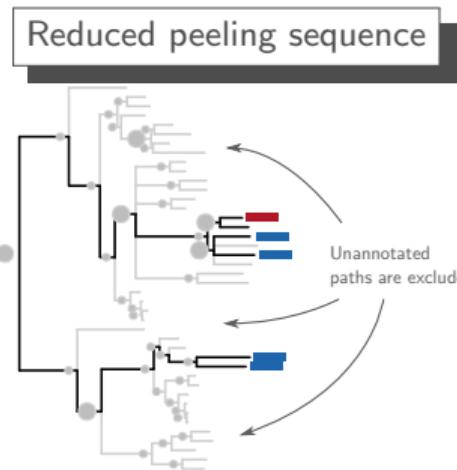
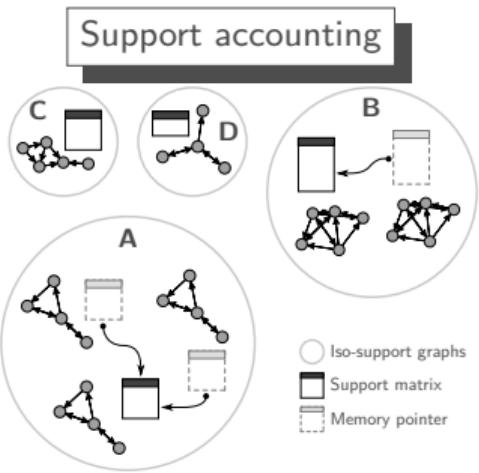
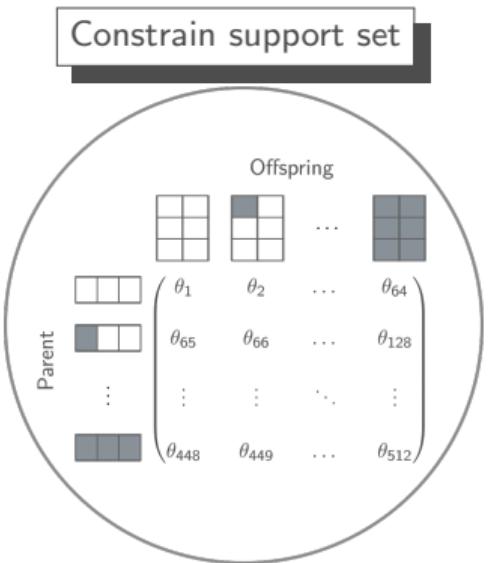
... I implemented this (and more) on `geese`



## GEne functional Evolution using SufficiEncy

... as part of **barry**, your to-go motif accountant

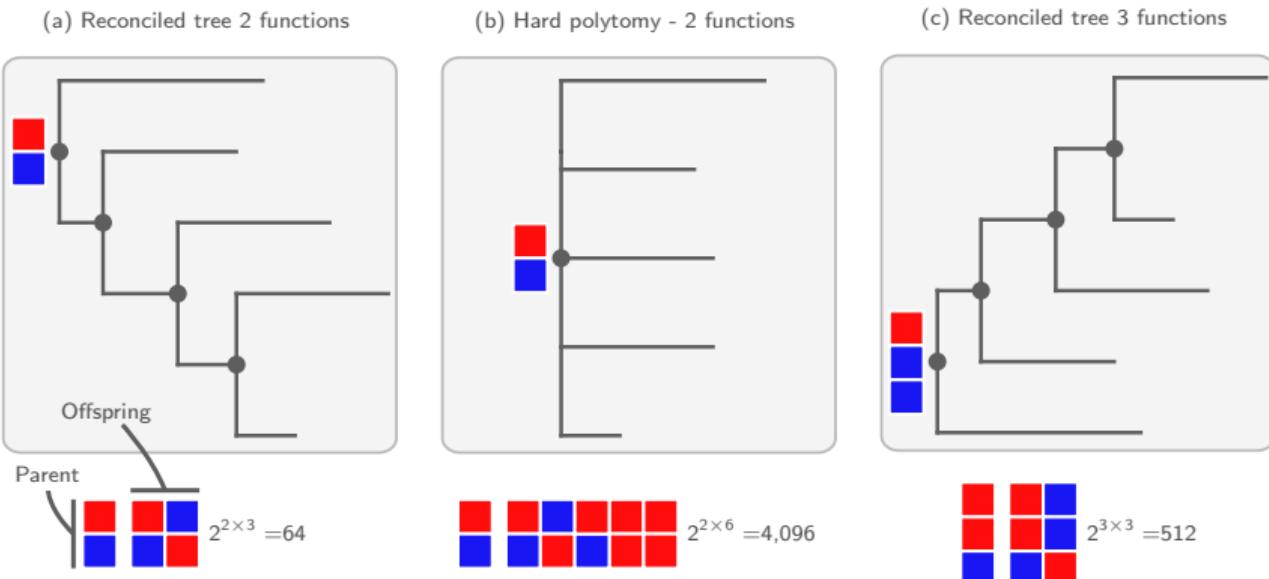


Computational Features of **geese**

... how big can we go?

## Support size: Computational feasibility

Whether the likelihood is tractable or not depends on how large is the “largest event”



(in practice, arrays up to 32 cells, i.e., 4.3 billion comb., are feasible.)

## Example: Simple model with two functions

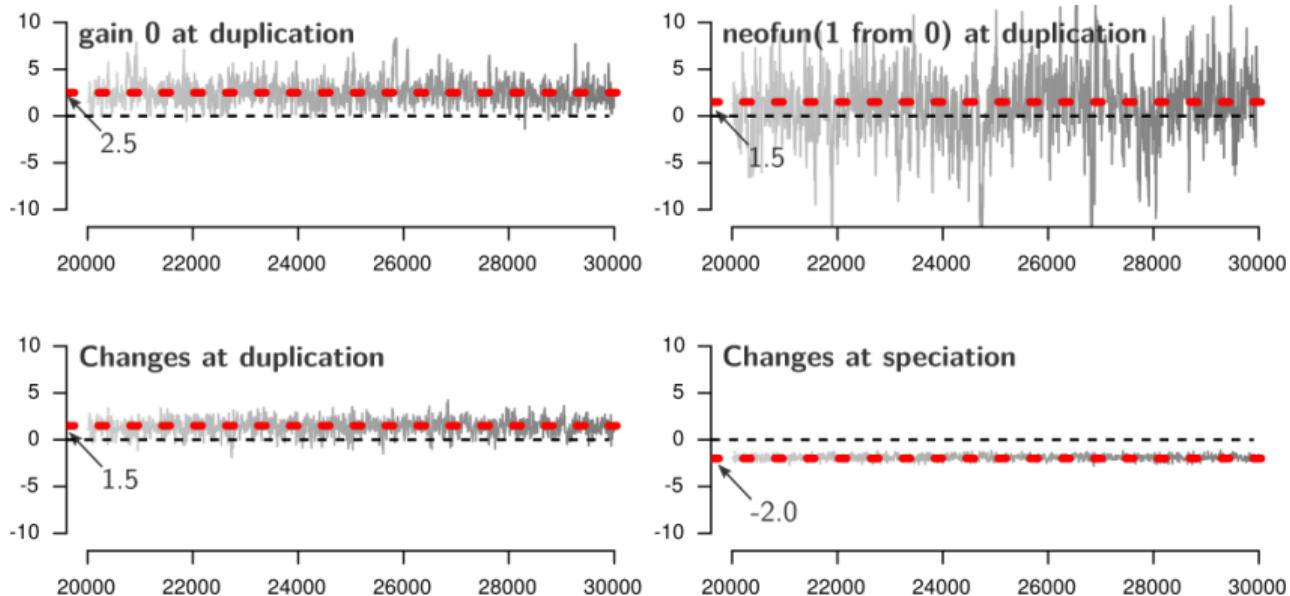
To illustrate, we will **simulate** and then **estimate** the parameters for the following process:

1. 100 genes on a simulated phylogenetic tree.
2. Two functions, 0 and 1,
3. Function 0 is gain with some prob. at a dupl. event,
4. Function 1 is gain as neofunctionalization (from 0) at a dupl. event,
5. There is a higher chance of changes at duplication.
6. There is low chance root node starts off with either 0 or 1.

We will fit the model using Robust Adaptive Metropolis with a logistic prior centered at 0 with scale 2.

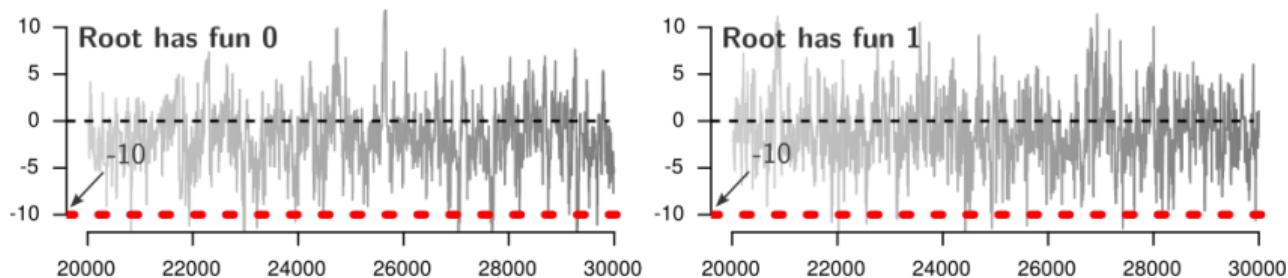
## Example: Simple model with two functions

posterior distributions



**Figure 2** MCMC Trace of the functional gain of 0, neofunctionalization (1 from 0), and change rate (by event type).

## Example: Simple model with two functions posterior distributions (contd')

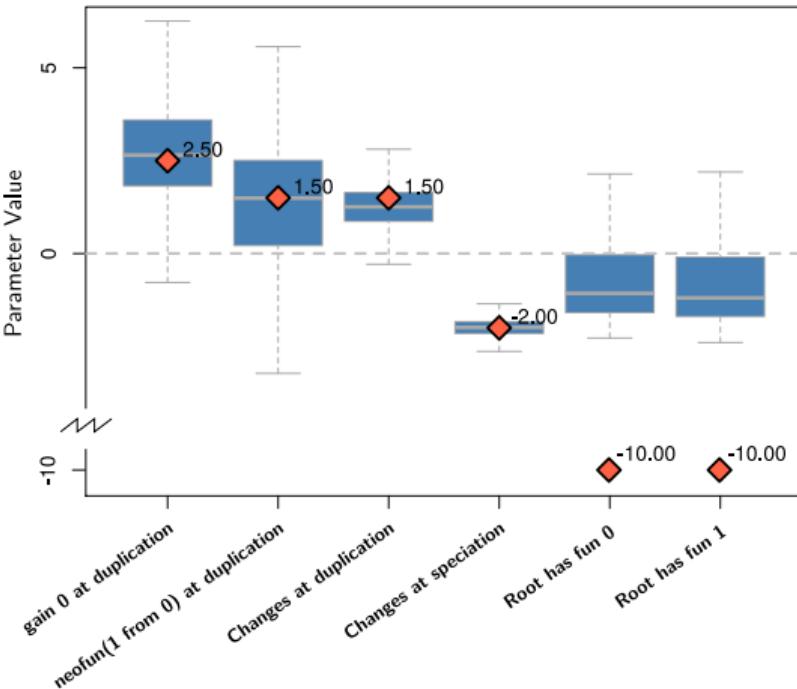


**Figure 3** MCMC Trace of root parameters. The true population parameters are  $(\theta_{root0}, \theta_{root1}) = (-10.0, -10.0)$ .  
Root node probabilities are always hard to get.

**Figure 4** Distribution of parameter estimates from 5,000 phylo trees  
w/ 100 leafs.

Repeated this experiment 5,000 times:

- ▶ MCMC for fitting.
- ▶ RAM kernel.
- ▶ Logistic prior at zero with scale two.
- ▶ Each tree took < 1min estimation.



## What questions?

With this modeling framework, we could tackle, e.g.,

- ▶ Potentially improve prediction accuracy.
- ▶ Make inferences of the sort of:
  - "Function A or function B, which came first?"
  - "When was subfunctionalization more likely to happen?"
  - "Where functions A and B gained at the same time?"
- ▶ and much more...

## Data Application

Using experimentally annotated phylogenies



## Comparison setup

Using data from Vega Yon et al. (2021), we re-estimated the models using **GEESE** and compared the results to those in the **aphylo** paper:

- ▶ 77 experimentally annotated trees
- ▶ We only used trees in which

$$2^{(\max \text{ Polytomy}+1) \times \text{nfun}} < 0.5 \times 10^9$$

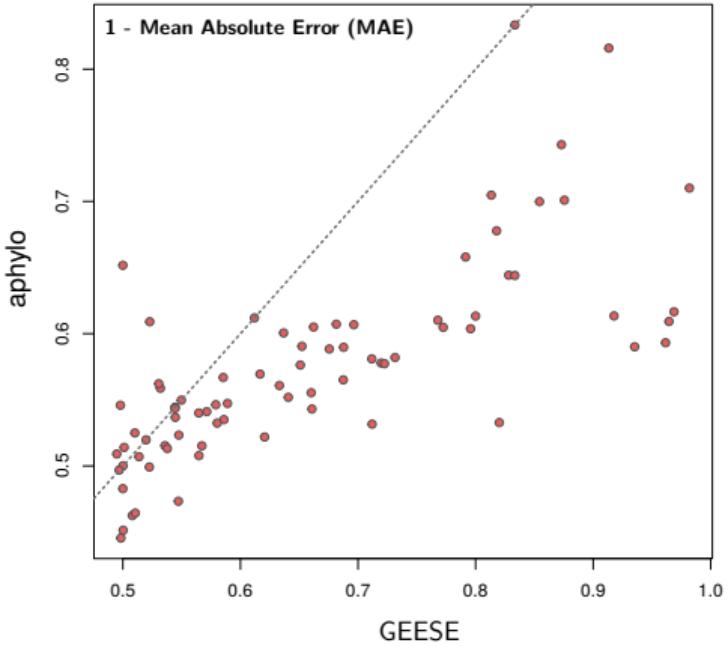
i.e., half billion

- ▶ Both models used "informative" priors.
- ▶ Both were fitted using Adaptive Metropolis (**fmcmc** R package).

## Example of code (R)

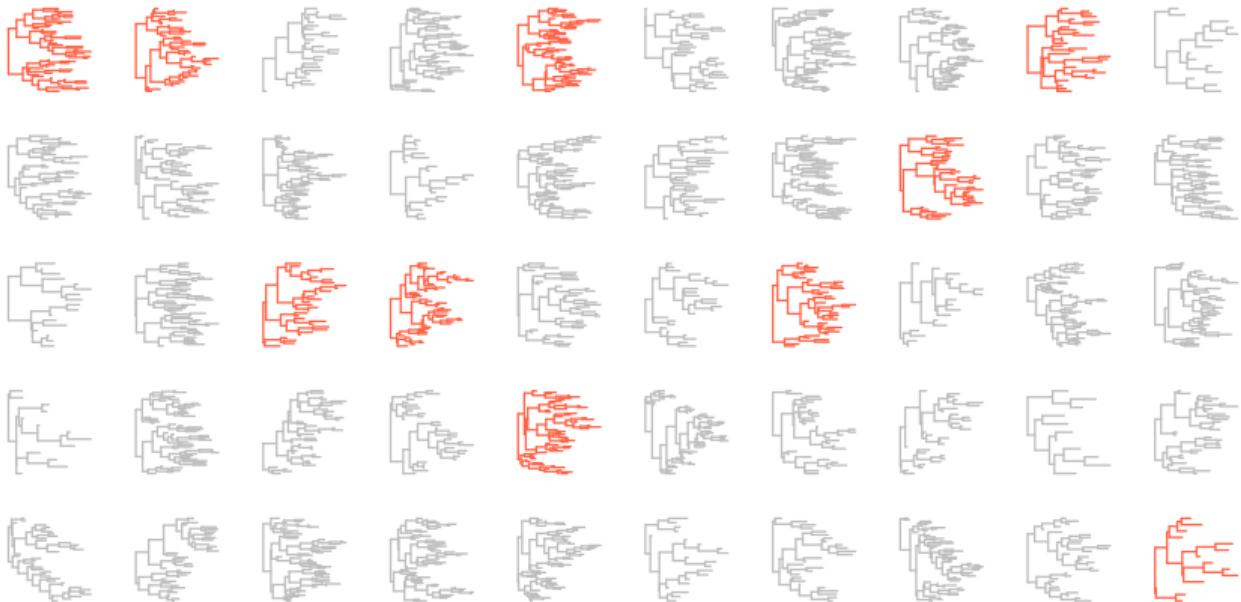
After initializing a geese object named model2fit:

```
1 # For later use (see last two lines)
2 term_overall_changes(model2fit, duplication = TRUE)
3 term_overall_changes(model2fit, duplication = FALSE)
4
5 # Couting how many genes change
6 term_genes_changing(model2fit, duplication = TRUE)
7
8 # Gain and loss at duplication
9 term_gains(model2fit, funs = 0:1, duplication = TRUE)
10 term_loss(model2fit, funs = 0:1, duplication = TRUE)
11
12 # Gain and loss at speciation
13 term_gains(model2fit, funs = 0:1, duplication = FALSE)
14 term_loss(model2fit, funs = 0:1, duplication = FALSE)
15
16 # Constraining the support set
17 rule_limit_changes(model2fit, id = 0, lb = 0, ub = 4, duplication = TRUE)
18 rule_limit_changes(model2fit, id = 1, lb = 0, ub = 4, duplication = FALSE)
```



**Figure 5** Performance as Mean Absolute Error (MAE). Each set of coordinates shows the value of 1 minus the MAE of **GEESE**, x-axis, and **aphylo** (y-axis). The statistics were computed using leave-one-out based on individual parameter estimates. Overall, GEESE performs better than aphylo in most cases.

## Tapping into computational scalability



**Figure 6** A dramatization of how a group of GEESE, i.e., a flock, looks like.

Goal

$$P(\text{ } | \text{ } , \text{ })$$

Annotations:  
■ Function present  
■ Function absent  
□ n/a

Today



Next steps



- ▶ We are in a race for uncovering **what genes do**.
- ▶ **Automatic algorithms** provide a way.
- ▶ Many alternatives... many unrealistic **assumptions**.
- ▶ **geese** (Sufficient Stats.): a general framework for gene function evolution.
- ▶ Further study its properties (bias, power, accuracy).
- ▶ Fit pooled data models (flocks).
- ▶ Find applications for this **model modeling framework**.

# Triadas, lazos y funciones genéticas cuando las redes sociales y la filogenética se encuentran

George G Vega Yon

<https://ggyv.cl>

vegayon@usc.edu



Gracias!

# The Gene Ontology Project

## Example of GO term

---

<b>Accession</b>	GO:0060047
<b>Name</b>	heart contraction
<b>Ontology</b>	biological_process
<b>Synonyms</b>	heart beating, cardiac contraction, hemolymph circulation
<b>Alternate IDs</b>	None
<b>Definition</b>	The multicellular organismal process in which the heart decreases in volume in a characteristic way to propel blood through the body. Source: GOC:dph

---

**Table 2** Heart Contraction Function. source: amigo.geneontology.org

You know what is interesting about this function?

◀ go back

These four species have a gene with that function... and two of these are part of the same evolutionary tree!



*Felis catus* pthr10037



*Oryzias latipes* pthr11521



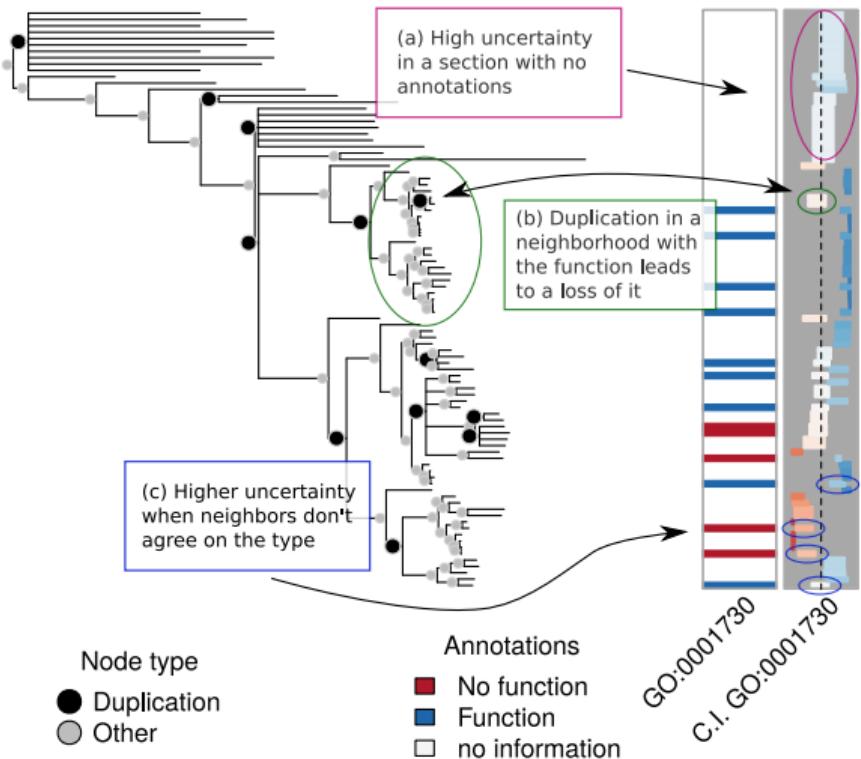
*Anolis carolinensis* pthr11521



*Equus caballus* pthr24356

[◀ go back](#)

## Example of Data + Predictions

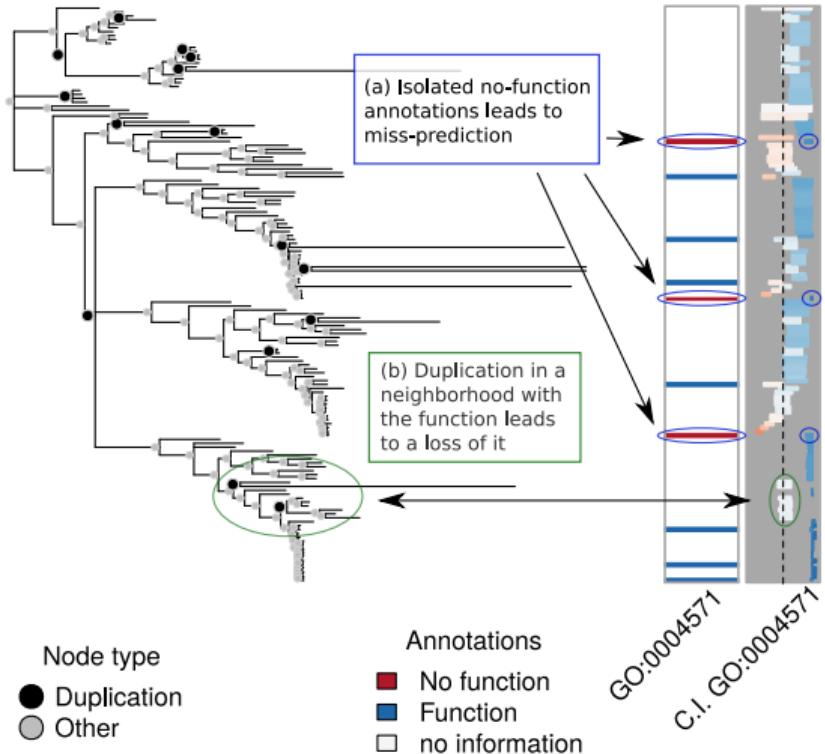
**Family: PTHR11258****Type:** Molecular Function**Name:** 2'-5'-oligoadenylate synthetase activity**Desc:** GO:0001730 involved in the process of cellular antiviral activity (wiki on [interferon](#)).**MAE:** 0.34**AUC:** 0.91[see a bad one](#)[◀ go back](#)

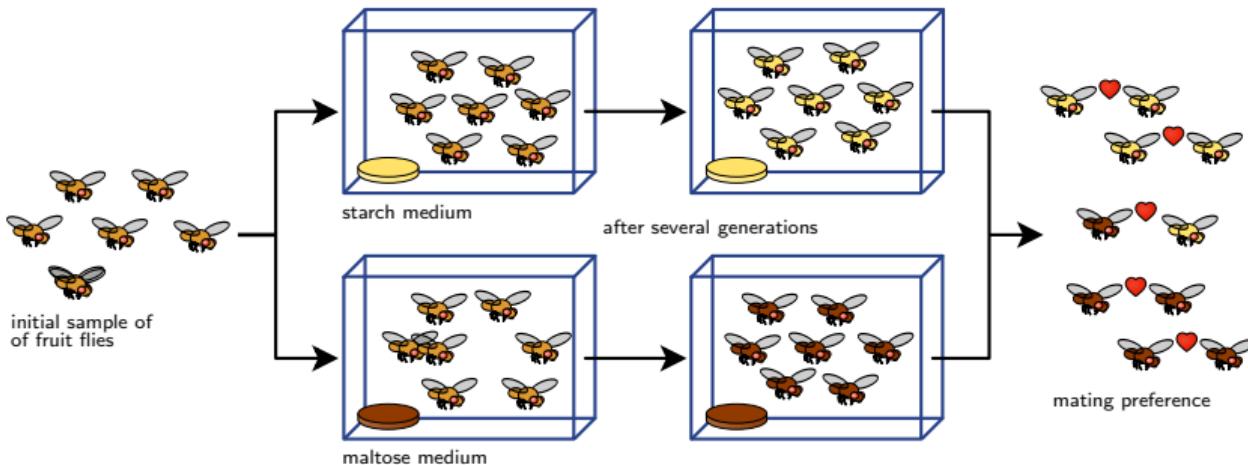
## Example 2: Bad quality prediction

MAE: 0.52

AUC: 0.33

Type: Molecular Function

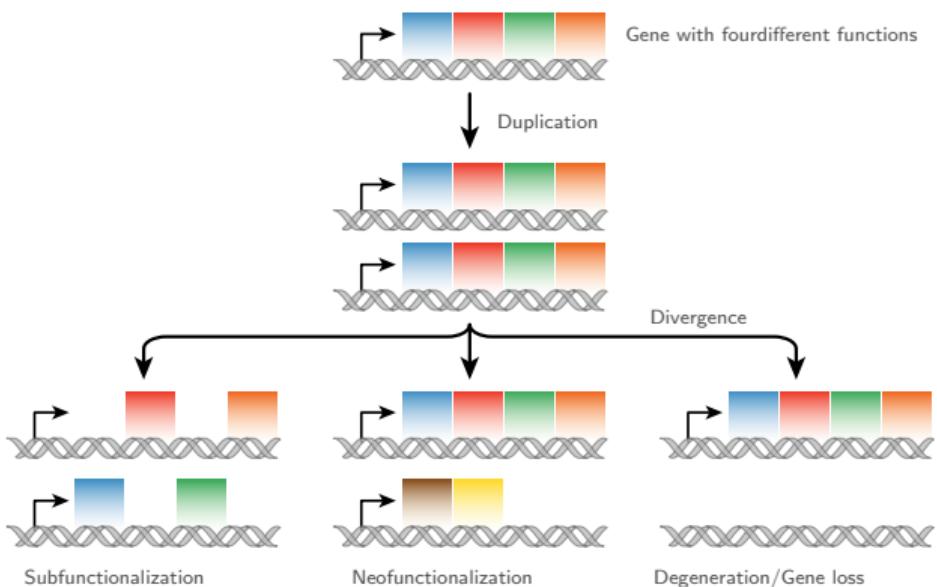
Name: mannosyl-oligosaccharide  
1,2-alpha-mannosidase activityDesc: GO:0004571 involved in  
synthesis of glycoproteins ([wiki](#)  
and [examples](#)).[◀ go back](#)



**Figure 7** Dodd (1989): After one year of isolation, flies showed a significant level of assortativity in mating (wikimedia)

◀ go back

# Duplication



**Figure 8** A key part of molecular innovation, gene duplication provides opportunity for new functions to emerge  
(wikimedia)

◀ go back

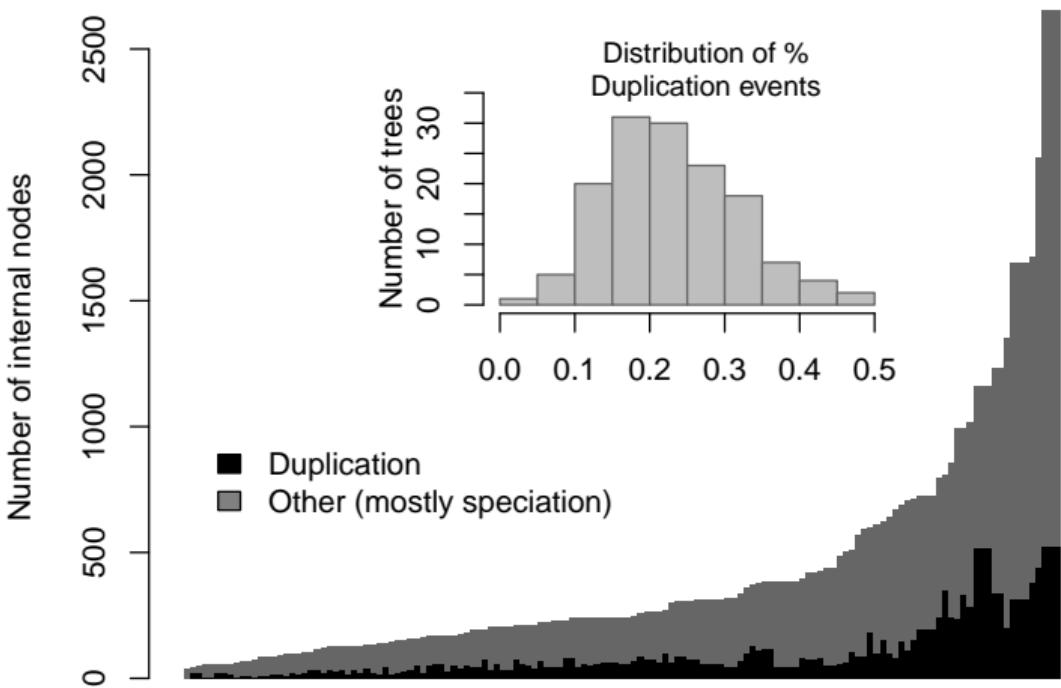
## Data: Phylogenetic trees

Sample of annotations (first 10 in a single tree, Phosphoserine Phosphatase [PTHR10000])

Internal id	Branch Length	type	ancestor
AN0		S	LUCA
AN1	0.06	S	Archaea-Eukaryota
AN2	0.24	S	Eukaryota
AN3	0.44	S	Unikonts
AN4	0.42	S	Opisthokonts
AN6	0.68	D	
AN9	0.79	S	Amoebozoa
AN10	0.18	D	
AN15	0.57	S	Dictyostelium
AN18	0.52	S	Alveolata-Stramenopiles

[◀ go back](#)

## Data: Node type (events)

[◀ go back](#)

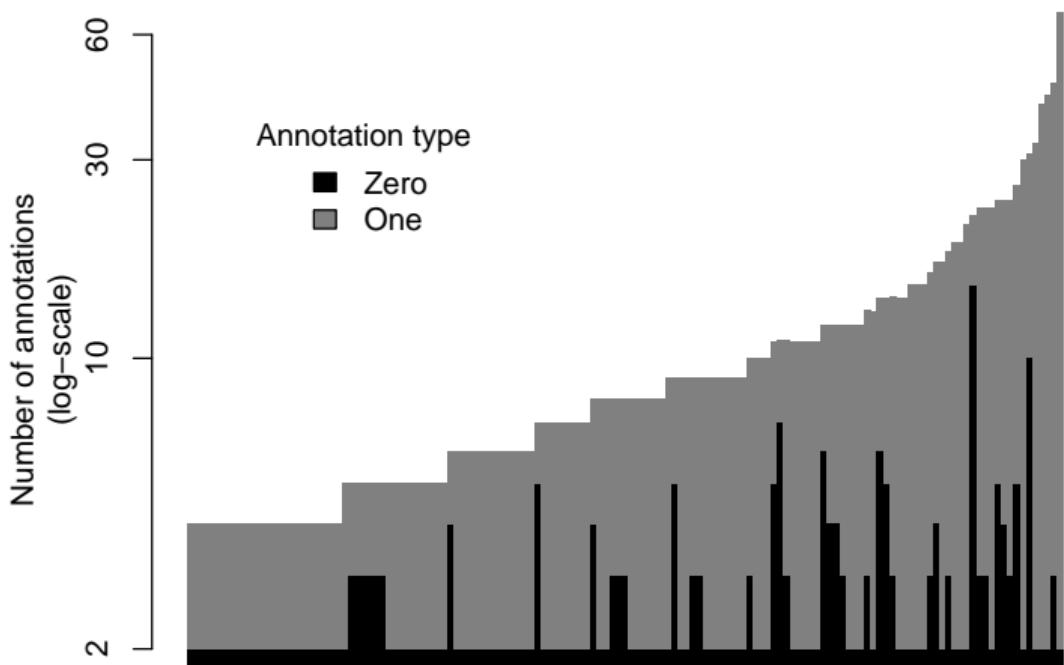
## Data: Annotations (example)

This is the first 10 of ~ 400,000 experimental annotations used:

	Family	Id	GO term	Qualifier
1	PTHR12345	HUMAN HGNC=15756 UniProtKB=Q9H190	GO:0005546	
2	PTHR11361	HUMAN HGNC=7325 UniProtKB=P43246	GO:0016887	CONTRIBUTES_TO
3	PTHR10782	MOUSE MGI=MGI=3040693 UniProtKB=Q6P1E1	GO:0045582	
4	PTHR23086	ARATH TAIR=AT3G09920 UniProtKB=Q8L850	GO:0006520	
5	PTHR32061	RAT RGD=619819 UniProtKB=Q9EPI6	GO:0043197	
6	PTHR46870	ARATH TAIR=AT3G46870 UniProtKB=Q9STF9	GO:1990825	
7	PTHR15204	MOUSE MGI=MGI=1919439 UniProtKB=Q9Z1R2	GO:0045861	
8	PTHR22928	DROME FlyBase=FBgn0050085 UniProtKB=Q9XZ34	GO:0030174	
9	PTHR35972	HUMAN HGNC=34401 UniProtKB=A2RU48	GO:0005515	
10	PTHR10133	DROME FlyBase=FBgn0002905 UniProtKB=O18475	GO:0097681	

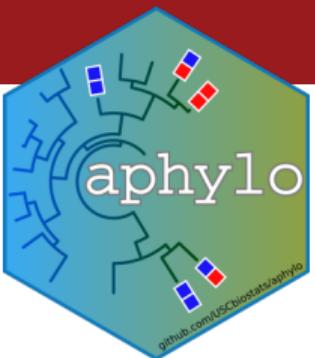
◀ go back

## Data: Experimental Annotations



◀ go back

# Computational features of aphylo

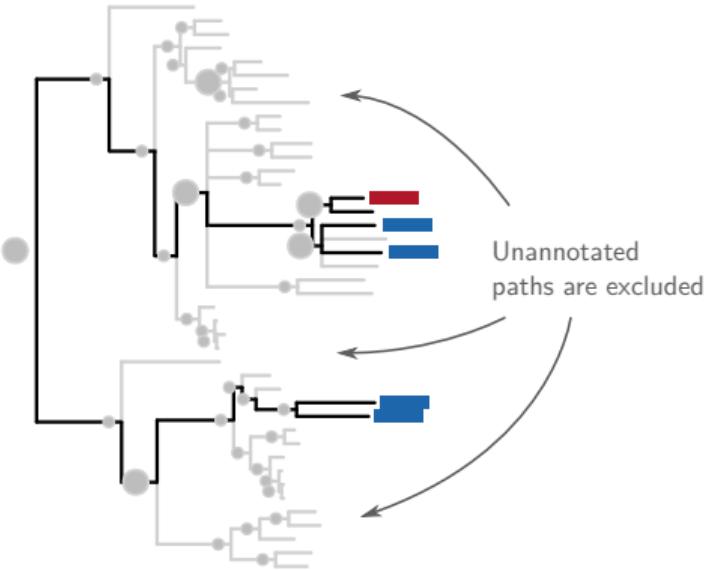


## Baseline features

- ▶ Parsimony: Conditional independence across functions/siblings.
- ▶ Post-order Tree traversal: Linear complexity  $O(|\text{tree}|)$ .

## Additional features

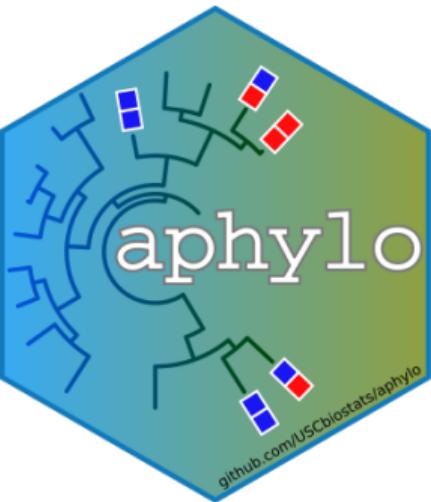
- ▶ Reduced pruning sequence: Induced sub-tree of nodes connected to annotated leafs  
     $\implies$  Complexity  $O(|\text{Induced sub-tree}|) \leq O(|\text{tree}|)$
- ▶ Implemented in C++ (**pruner** library)



◀ return

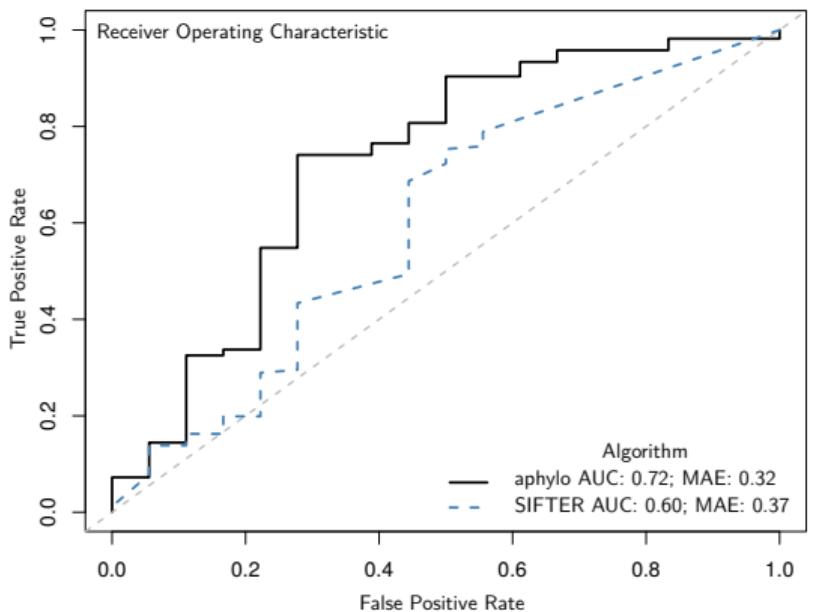
## Results: Implementation and Large scale study

- ▶ Simulation, estimation, and prediction: **aphylo** R package.
- ▶ Large simulation study (all known trees, about 15,000) on USC's HPC cluster.
- ▶ Prediction quality assessment on  $\sim 1,300$  genes involving  $\sim 130$  families... estimation of parameters using a pooled-data model (< 5 min). [◀ modeling](#) [◀ estimates](#)
- ▶ In a subset of  $\sim 200$  predictions we found 46 novel annotations

[▶ more](#)[◀ go back](#)

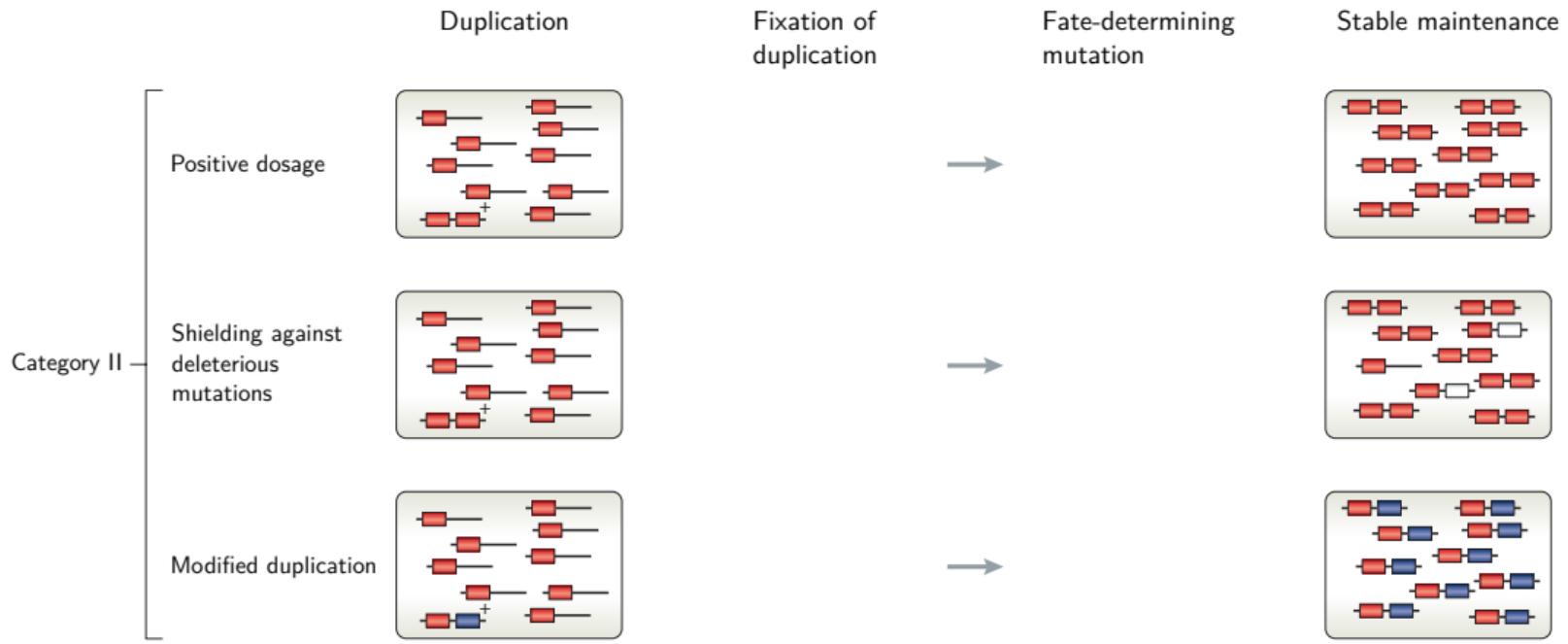
## Results: Performance and Scalability

aphylo vs SIFTER (state-of-the-art phylo-based model) on 147 genes.



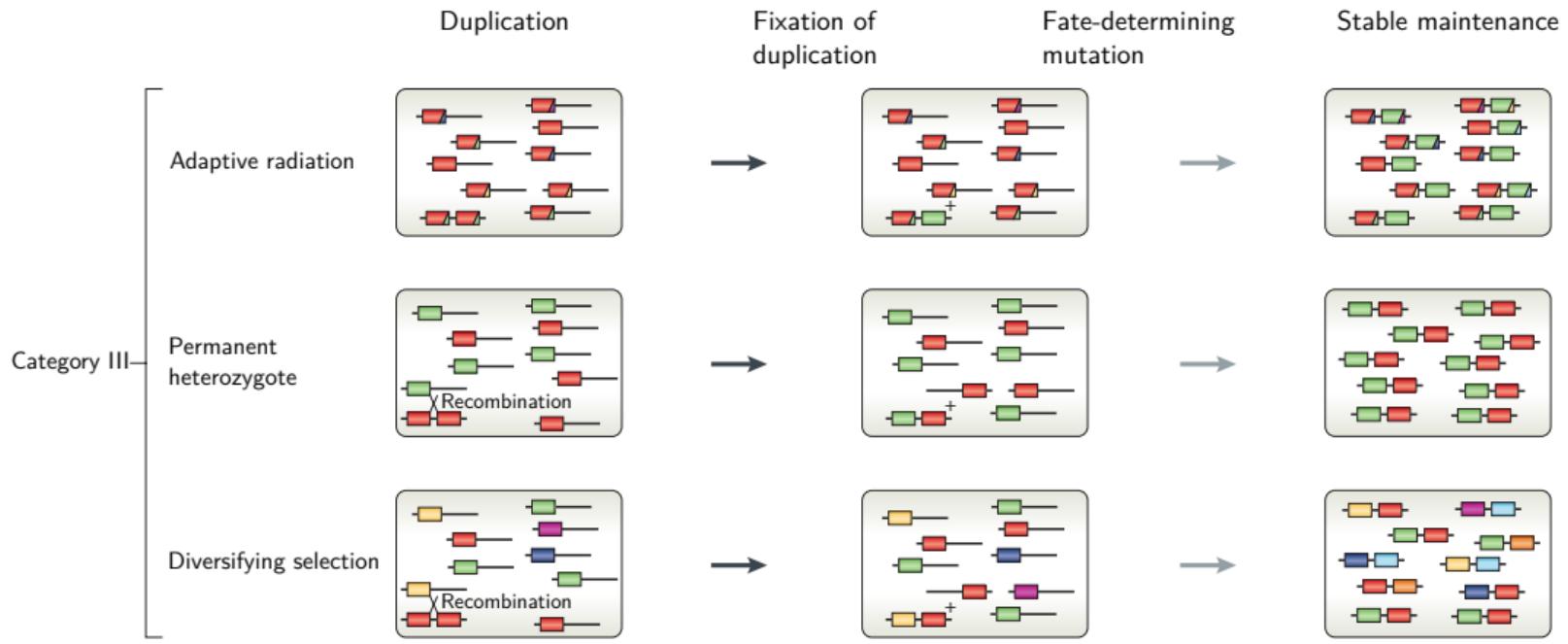
**Fast** 110 minutes (SIFTER) to calculate the posterior probabilities, aphylo took 1 second.

**Accurate** aphylo reported higher accuracy levels in LOO cross-validation (0.72 vs 0.60 AUC).



**Figure 9** Category II (*advantageous duplication*). Source: Adapted from Innan and Kondrashov (2010, *nature rev. gen.*)

◀ return



**Figure 10** Category III (advantageous duplication). Source: Adapted from Innan and Kondrashov (2010, nature rev. gen.)

◀ return

# Overview of Prediction Results

	Pooled	Type of Annotation		
		Molecular Function	Biological Process	Cellular Comp.
<b>Mislabeling</b>				
$\psi_{01}$	0.23	0.18	0.09	
$\psi_{10}$	0.01	0.01	0.01	
<b>Duplication Events</b>				
$\mu_{d01}$	0.97	0.97	0.10	
$\mu_{d10}$	0.52	0.51	0.03	
<b>Speciation Events</b>				
$\mu_{s01}$	0.05	0.05	0.05	
$\mu_{s10}$	0.01	0.01	0.02	
<b>Root node</b>				
$\pi$	0.79	0.71	0.88	
Trees	141	74	45	22
<b>Accuracy under the by-aspect model</b>				
AUC	-	0.77	0.83	
MAE	-	0.34	0.26	
<b>Accuracy under the pooled-data model</b>				
AUC	-	0.77	0.75	
MAE	-	0.35	0.34	

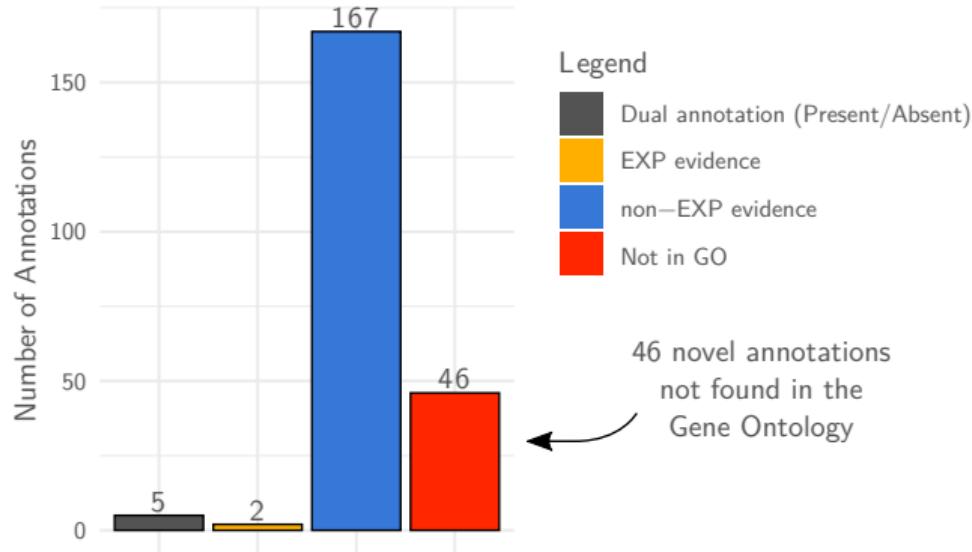
Previously, joint estimates out-performed one-at-a-time

- ▶ **Molecular Function** No change.
- ▶ **Biological Process** Significantly better.
- ▶ **Cellular Component** Does not converge.

Molecular Function  $\neq$  Biological Process ? Cellular Component

▶ data

▶ go back



**Figure 11** Distribution of predictions

◀ go back

# What Drives Evolution

Imagine that we have 3 functions (rows) and that each node has 2 siblings (columns)

		Transitions to	
		Case 1	Case 2
Parent	A	$\begin{bmatrix} 0 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$
	B	$\begin{bmatrix} 1 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$
	C	$\begin{bmatrix} 1 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$

## Sufficient statistics

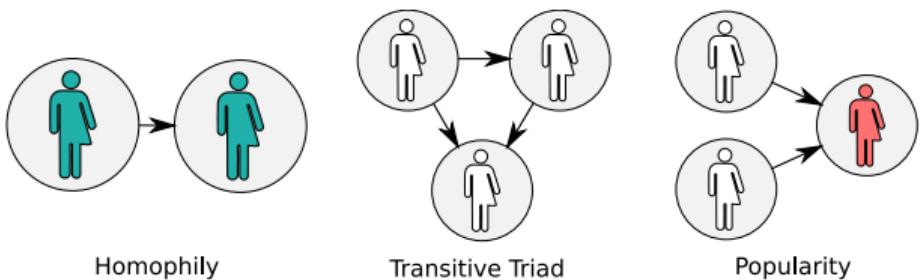
# Gains	1	1
Only one offspring changes (yes/no)	1	0
# Changes (gain+loss)	2	3
Subfunctionalizations (yes/no)	0	1

▶ return

# What are Exponential Random Graph Models

Exponential Family Random Graph Models, aka **ERGMs** are:

- ▶ Statistical models of (social) networks.
- ▶ Social Network Analysis: What drives social connections?
- ▶ Not about individual ties, but about local structures (sufficient statistics).



- ▶ Social Networks  $\equiv$  Adjacency Matrix  $\equiv$  Binary arrays

# What Drives Evolution: a game changer

In the model with 3 functions and 2 offspring per node:

- ▶ Full Markov transition matrix:  $2^3 \times 2^6 = 512$
- ▶ Using sufficient statistics:

Pairwise co-evolution: 3 terms,

Pairwise Neofunctionalization: 3 terms,

Pairwise Subfunctionalization: 3 terms,

Function specific gain: 3 terms,

Function specific loss: 3 terms,

Total: 15 parameters.

- ▶ Easier to fit and interpret.

