

# Applications of Statistical Computing in Complex Social and Biological Systems Modeling

George G Vega Yon, Ph.D.

University of Southern California, Department of Preventive Medicine

The University of Utah

(virtual)

June 14, 2021

## Funding and support



National Cancer Institute Grant #1P01CA196596.



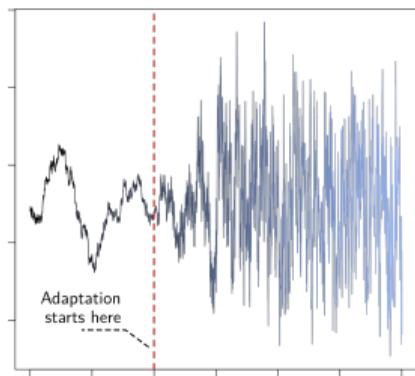
U.S. Army Research Laboratory and the U.S. Army Research Office under grant number W911NF-15-1-0577.



Advanced Research Computing  
Enabling scientific breakthroughs at scale

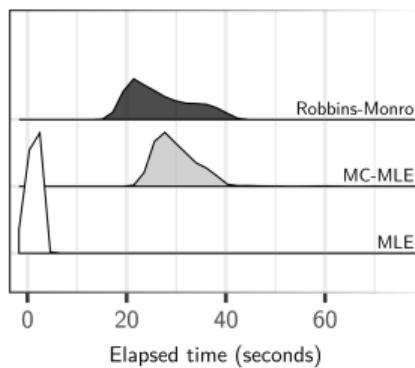
My work sits at the intersection between...

### Statistics



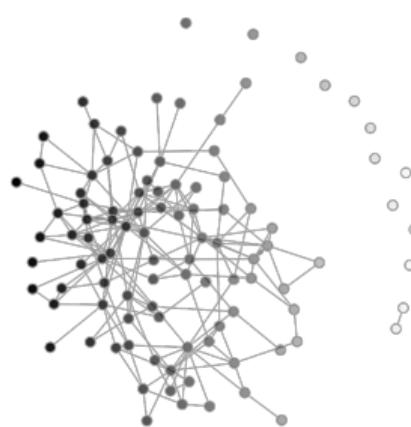
Bayesian, Non-parametric,  
Spatial

### Computer Science



parallel computing, HPC,  
software

### Complex Systems



social, biological, technical

Part I: A general framework for modeling functional evolution

Part 2: Exponential Random Graph Models for Small Networks

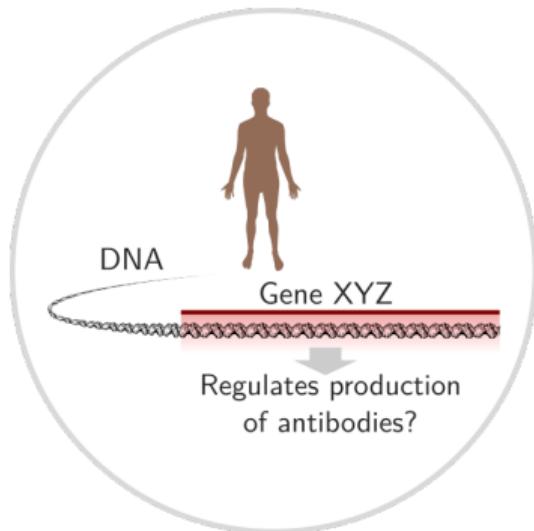
Part 3: Other Projects and Future Research

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

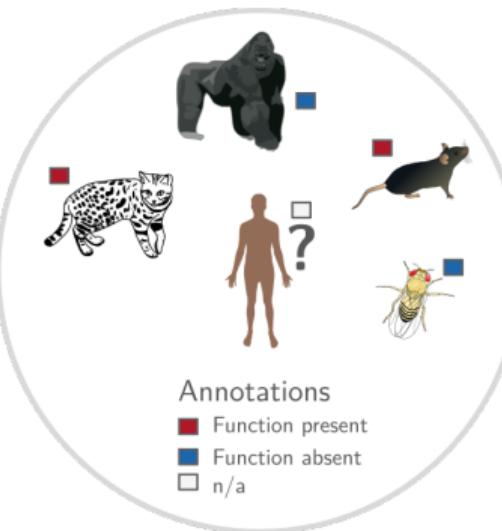
## A general framework for modeling functional evolution

*Joint with:* Paul D Thomas, Paul Marjoram, Huaiyu Mi, Duncan Thomas, and John Morrison

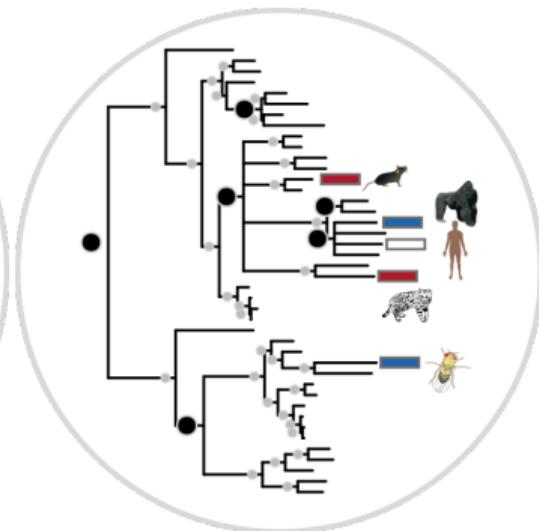
# Is gene XYZ involved in process ABC?



Complex to directly assess



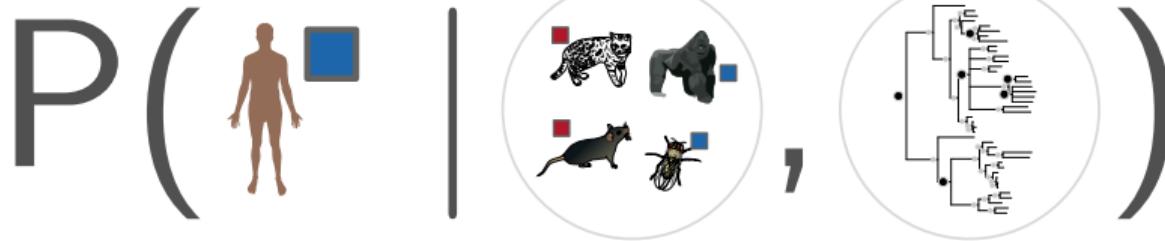
But we may know from other species



And we further know how these *genetically connected*

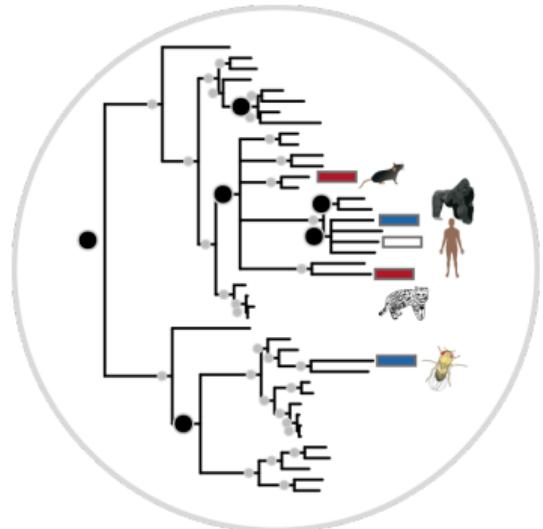
... let's rephrase the question.

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

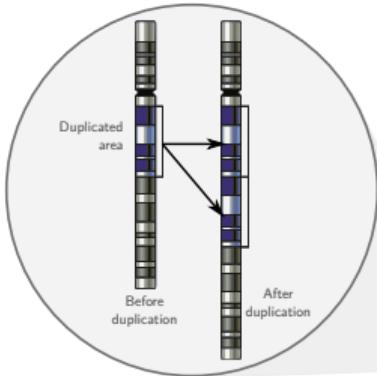


- ▶ ~ 15,000 phylogenetic trees
- ▶ ~ 8 million annotations
- ▶ ~ 600 thousand on human genes
- ▶ ~ < 10% are based on experimental evidence... Improving our knowledge on genetics is fundamental for advancing Biomedical Research

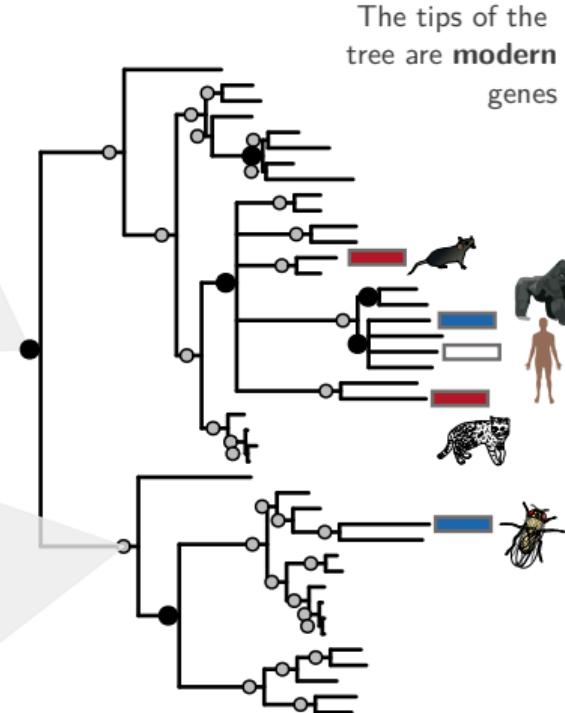
Only on 2020, 2,000+ COVID-19 papers using the GO (Google Scholar)

► more

- nodes are Duplication Events



- nodes are Speciation Events



[◀ go back](#)

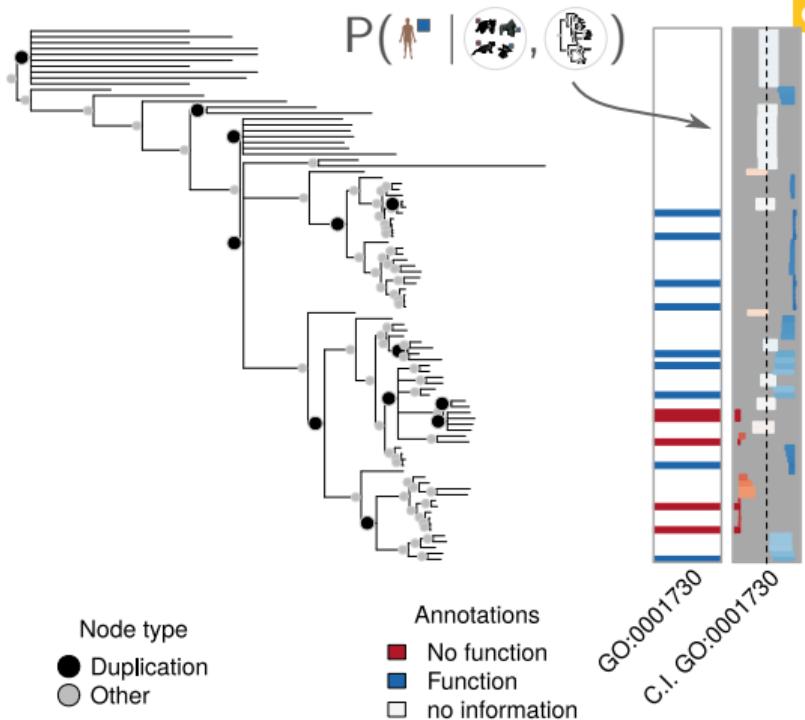
## Example: Molecular function in family PTHR1128

**Name:** 2'-5'-oligoadenylate synthetase activity

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

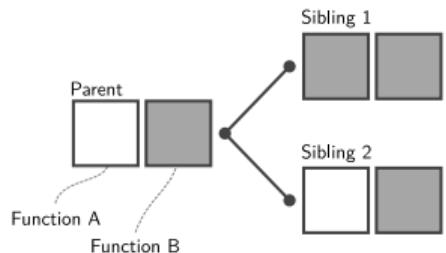
**MAE:** 0.34

**AUC:** 0.91

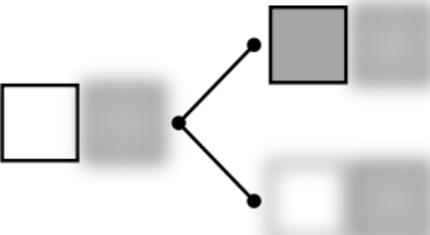


**Note:** Prediction made using **aphylo** (Vega Yon and *et al*, PLOS Comp. Bio 2021)

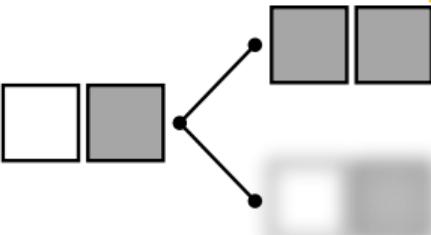
[see details](#)



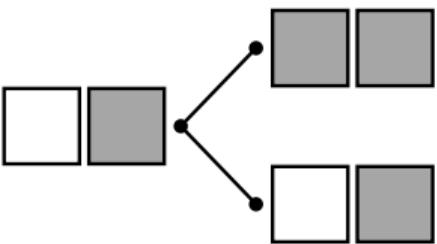
Has the function  
 Doesn't have the function



(a) Sibling and Function  
Conditional Independence



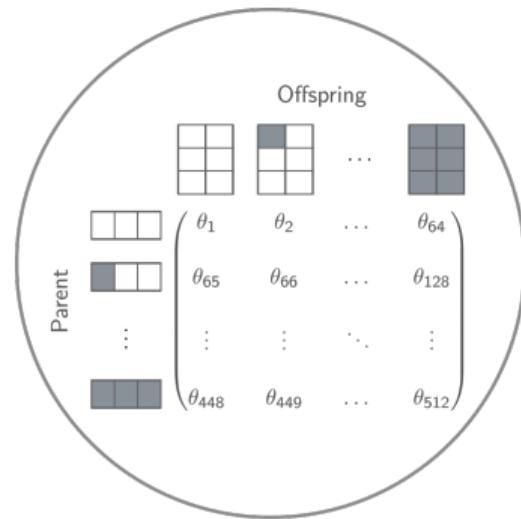
(b) Sibling Conditional  
Independence



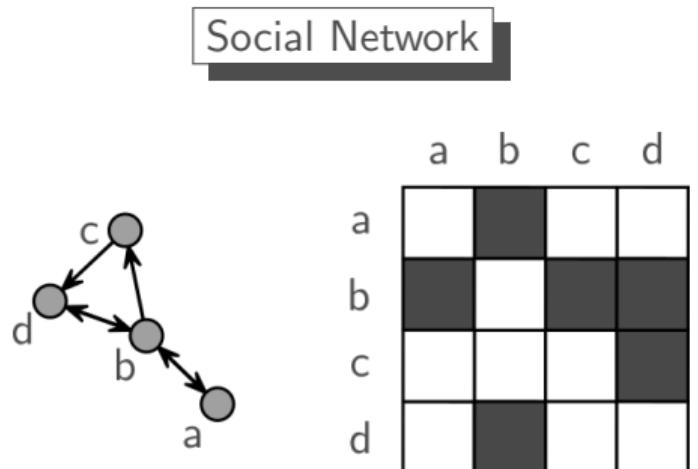
(c) No conditional  
independence

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

### Full Markov Transition Matrix



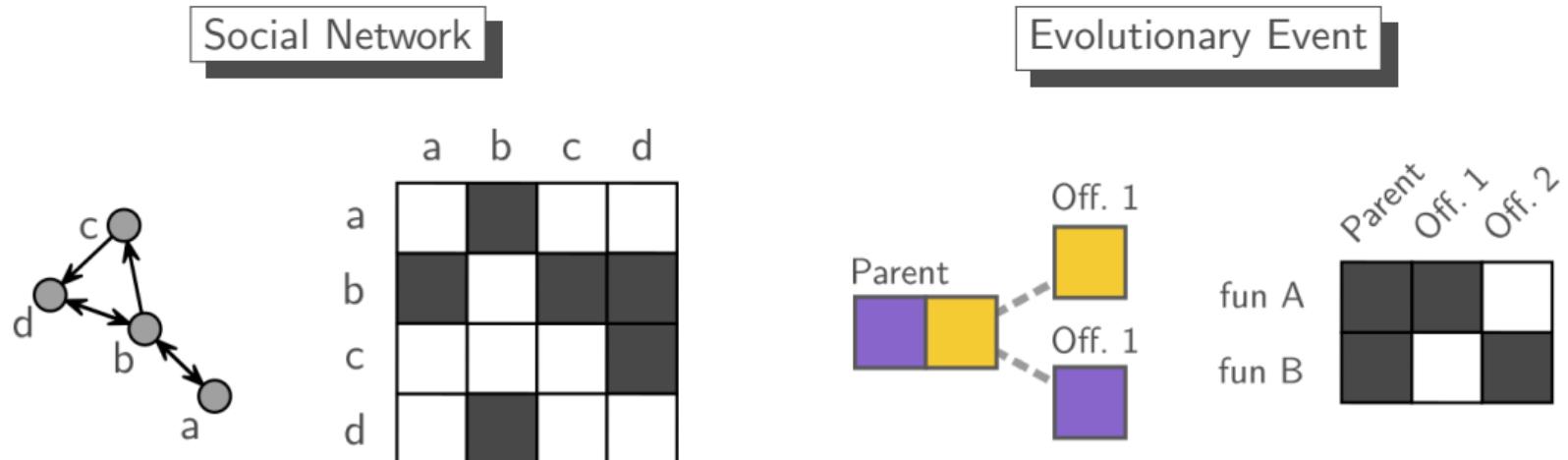
- ▶ 512 parameters
- ▶ Finding this many parameters not easy.
- ▶ Even if you can, interpretation is awkward.



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

Ultimately...

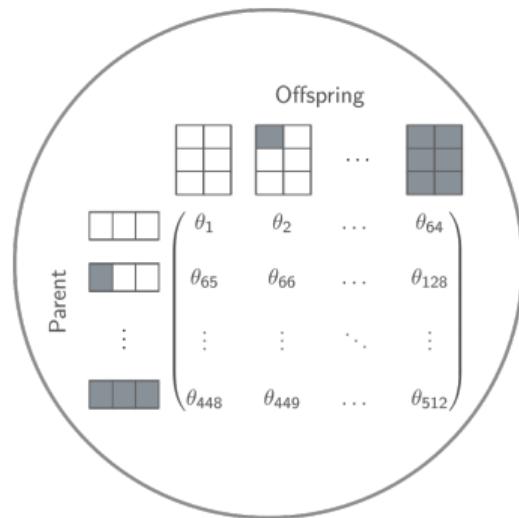
**ERGM ≡ Modeling binary arrays**



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



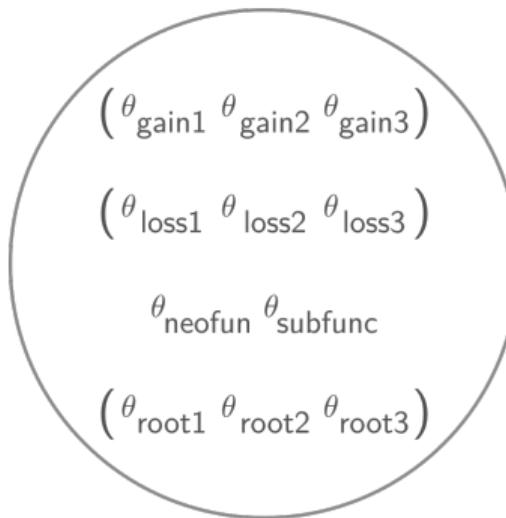
512 parameters

$$\begin{pmatrix} \theta_{\text{gain}1} & \theta_{\text{gain}2} & \theta_{\text{gain}3} \\ \theta_{\text{loss}1} & \theta_{\text{loss}2} & \theta_{\text{loss}3} \\ \theta_{\text{neofun}} & \theta_{\text{subfunc}} \\ (\theta_{\text{root}1} & \theta_{\text{root}2} & \theta_{\text{root}3}) \end{pmatrix}$$

Easier to fit

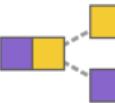
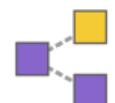
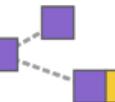
Easier to interpret

### Sufficient statistics



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

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

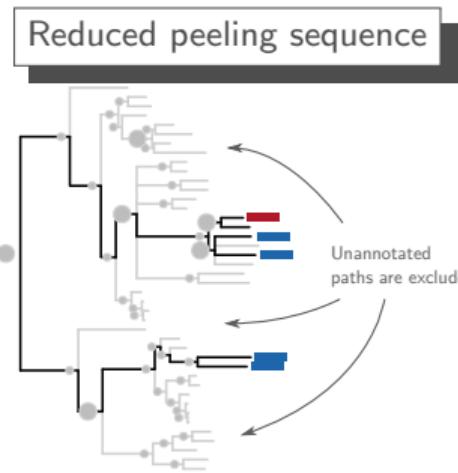
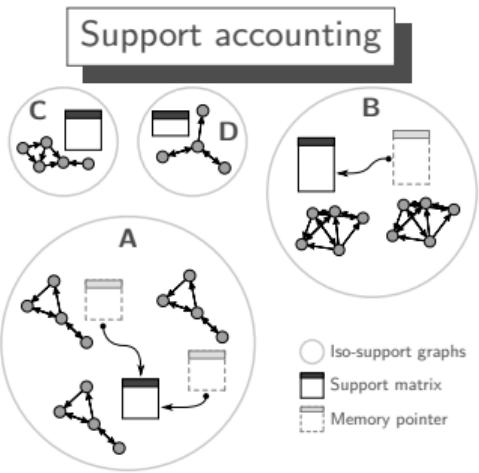
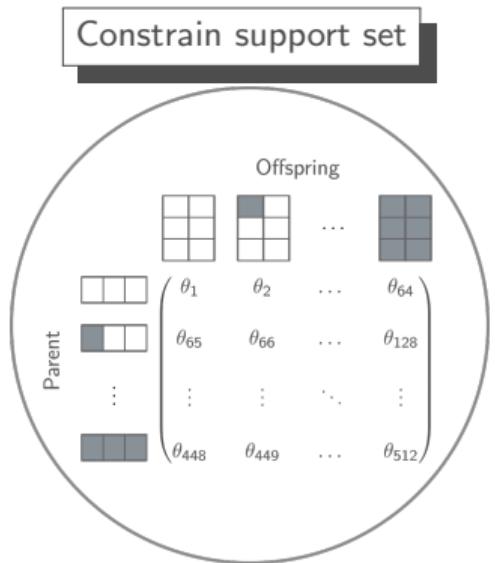
... I implemented this (and more) on `geese`



**GEne functional Evolution using SufficiEncy**

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

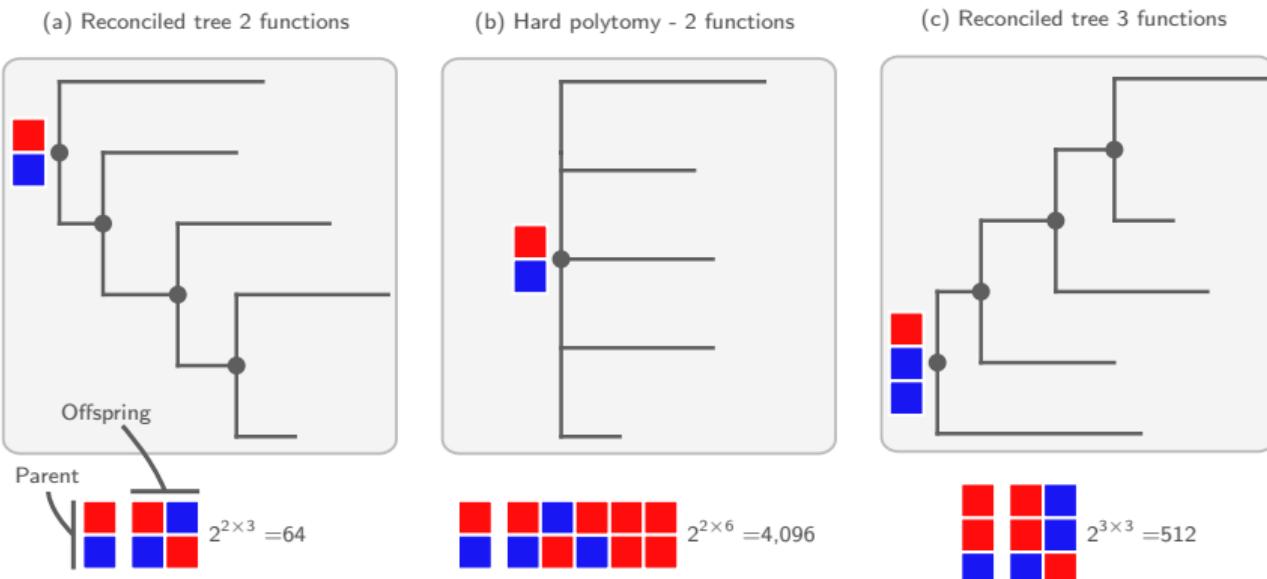


Computational Features of **geese**

... 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.)

## 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...

**geese** featured example: Analyzing 77 experimentally annotated trees  
(Vega Yon, *et al.*, WIP)

## geese featured example: Analyzing 77 experimentally annotated trees

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

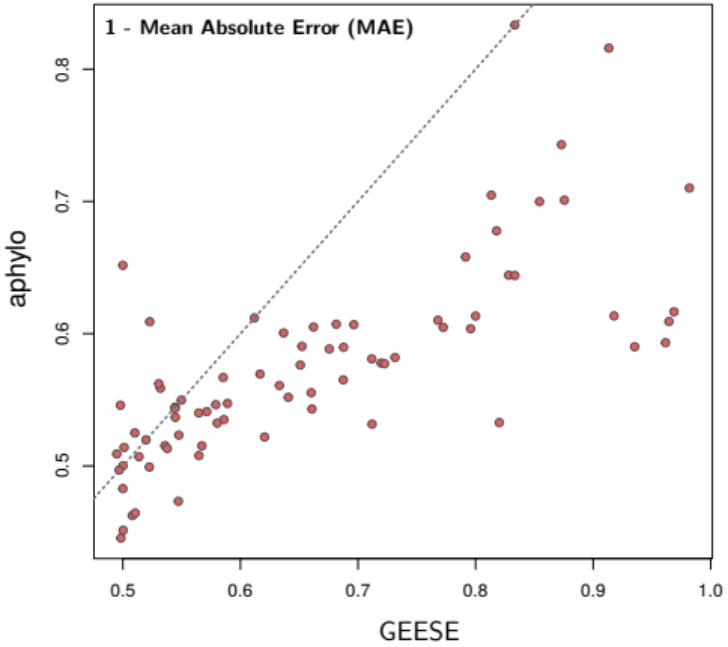
- ▶ Both models used "informative" priors.
- ▶ Both were fitted using Adaptive Metropolis (**fmcmc** R package).

more

## Example of code (R)

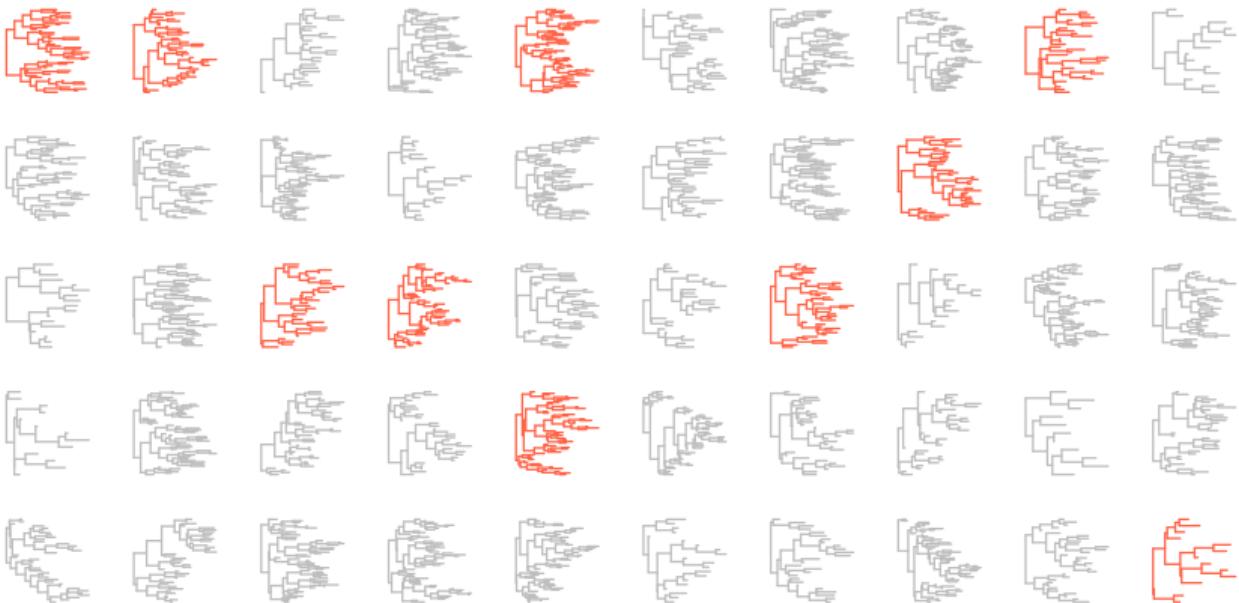
After initializing a geese object named model2fit:

```
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)
19
```



**Figure 1** 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 2** A dramatization of how a group of GEESE, i.e., a flock, looks like.

## Part 2: Exponential Random Graph Models for Small Networks

*Joint with:* Andrew Slaughter and Kayla de la Haye  
(published in the journal *Social Networks*)

- ▶ If COVID-19 has taught us something it is that networks matter.
- ▶ And especially small networks: Families, teams, friends, etc. The cornerstone of larger social systems.
- ▶ We can study networks using ERGMs.

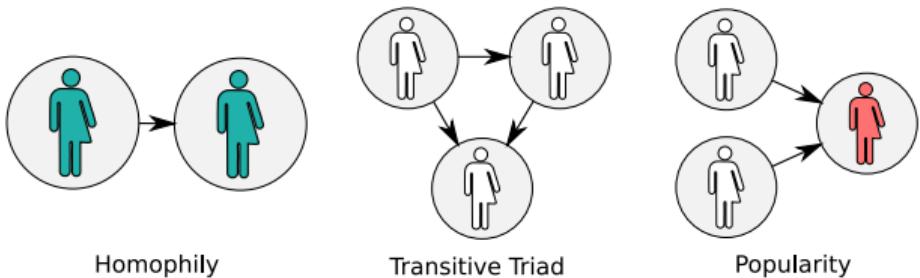
Data: Friendship network of a UK university faculty  
from `igraphdata`. Viz: R package `netplot` (yours truly,  
[github.com/usccana/netplot](https://github.com/usccana/netplot))



# What are Exponential Random Graph Models

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

- ▶ Statistical models of (social) networks.
- ▶ Not about individual ties, but about local structures (sufficient statistics).



$$\mathbb{P}(\mathbf{G} = \mathbf{g} \mid X = x) = \frac{\exp\{\theta^t s(\mathbf{g}, x)\}}{\sum_{\mathbf{g}' \in \mathcal{G}} \exp\{\theta^t s(\mathbf{g}', x)\}}, \quad \forall \mathbf{g} \in \mathcal{G}$$

A vector of  
model parameters
A vector of  
sufficient statistics


Observed data
The normalizing  
constant
All possible  
networks

- ▶ For any directed graph of size  $n$ , there are  $2^{n(n-1)}$  possible realizations.
  - ▶ A directed graph of size 5 has 1,048,576 possible configurations!
  - ▶ Most (all) applications use **approximations**... yet, for sufficiently small graphs we “can be exact.”

... I implemented this in the **ergmito** R package

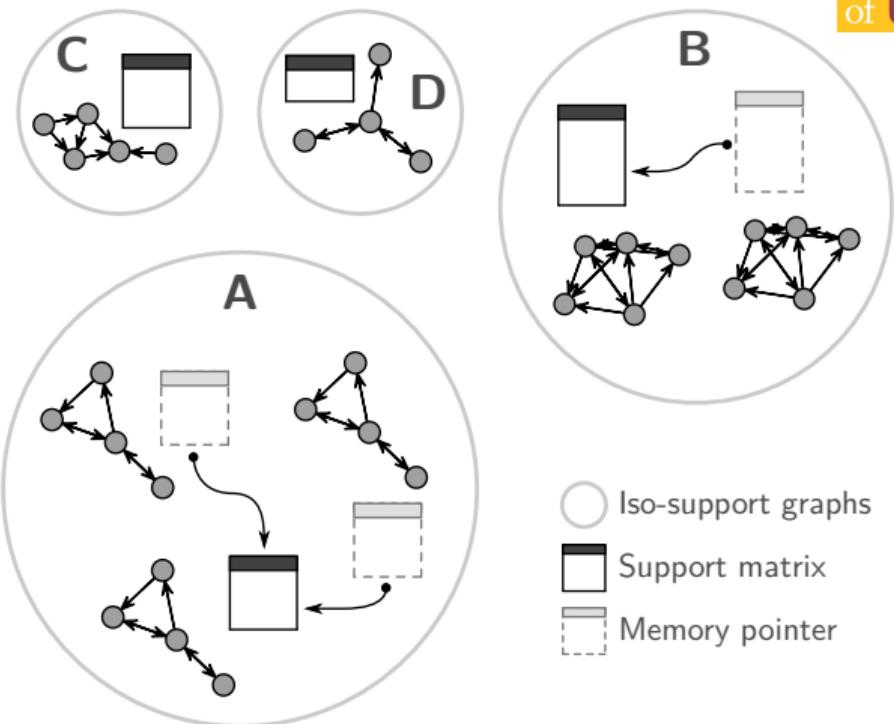
# Computational Complexity: ergmito

## Premise

- ▶ Full enumeration  $\neq$  All networks with  $k$  ties

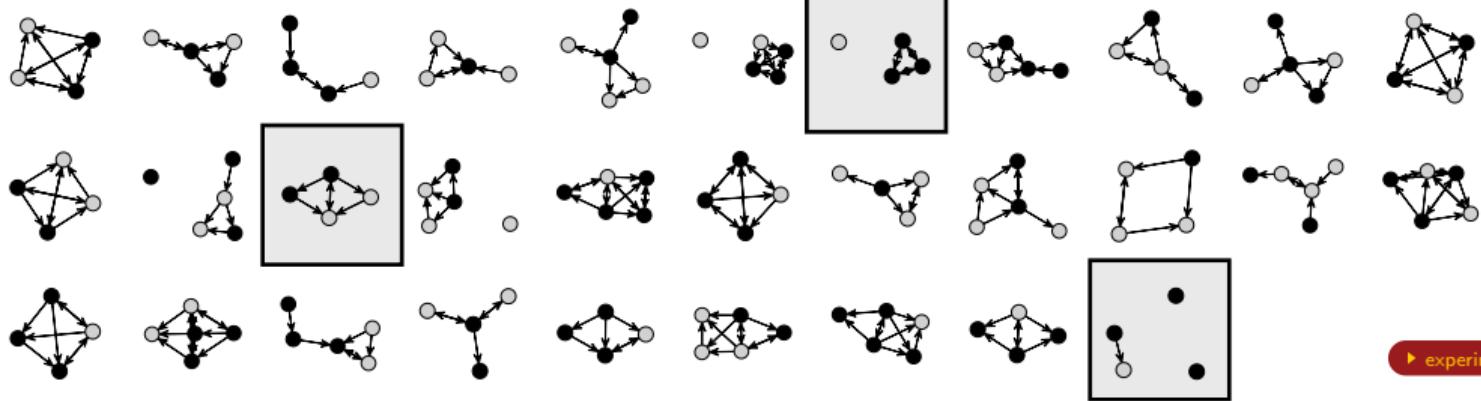
## Extra features

- ▶ *Iso-support* structures (model) recycled
- ▶ Core algorithm written in C++ (with **OpenMP**)



Fitting a pooled-data ERGMito with 80K nodes in 20K small nets  $\sim 4s$

**ergmito** featured example: Advice seeking networks in small teams  
(de la Haye et al, 2020)



### Key findings

- ▶ Low density.
- ▶ High balance (transitive triads).
- ▶ No evidence of gender homophily.
- ▶ Females are more likely to ask for advice.

### What's different (from regular ERGMs)?

- ▶ Interaction effects: edges  $\times \mathbf{1}$  ( $n = 5$ ).
- ▶ Constrained support: edge  $> 4$ .
- ▶ Transformed variables:  $(\text{Homophily})^{1/2}$ .
- ▶ Bootstrapping: 1,000 replicates in less than 1.5 minutes...  
... if you are lucky, using “regular” ERGMs would take you about 5 hours.

▶ details

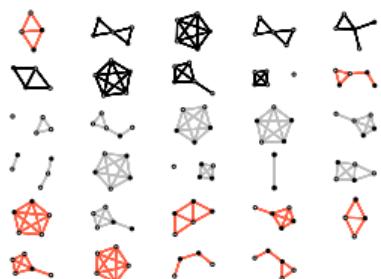
▶ gof

▶ data

## Part 3: Other Projects and Future Research

## Sociology

Using ERGMitos to model  
political discussion networks in  
Romania



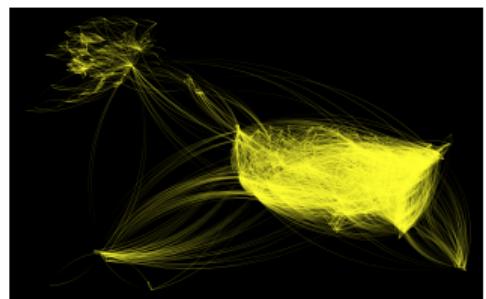
## Criminology

The role of social networks on  
police use of force

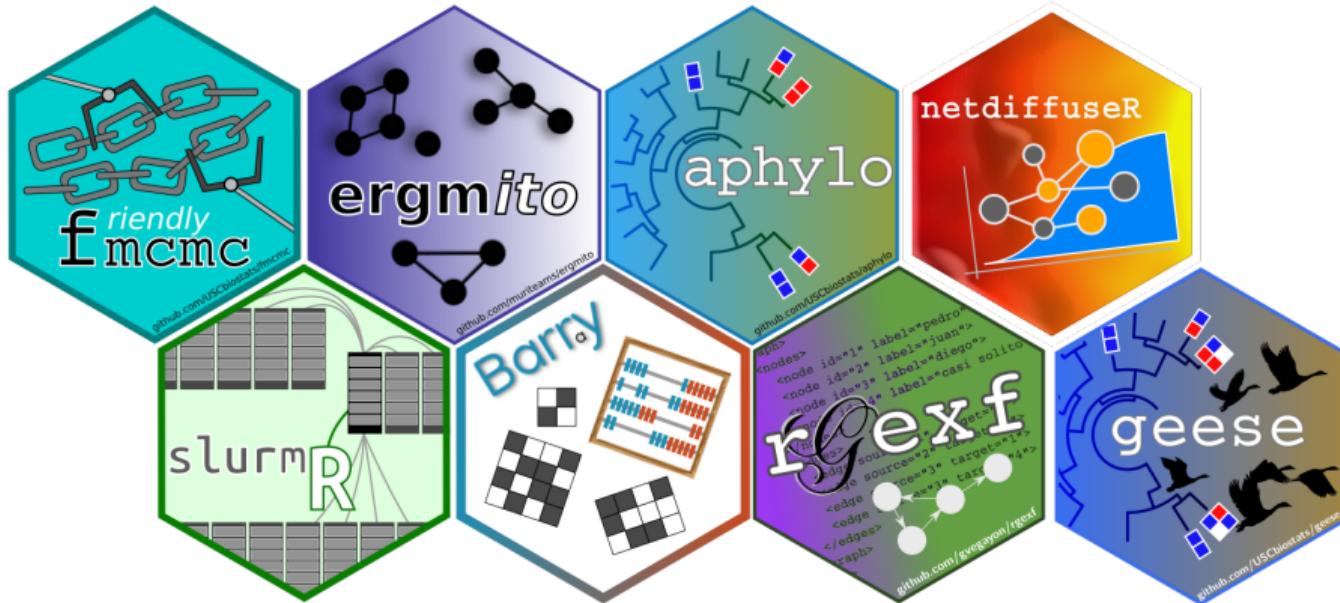


## Statistical Computing

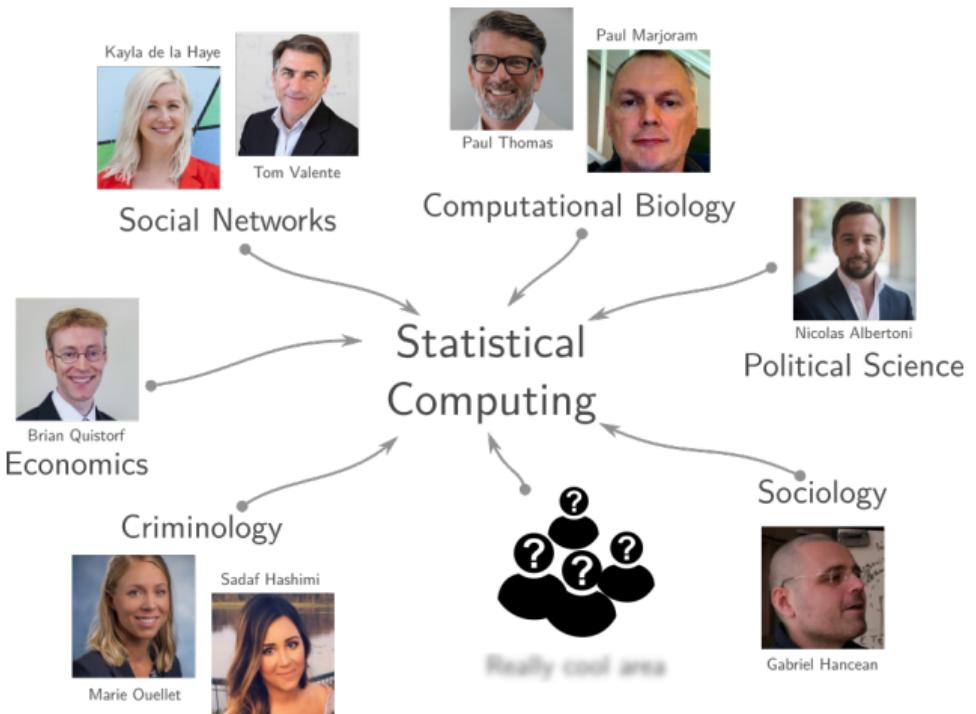
Leverage computing power to  
advance theory and application in  
studying complex systems...



Continue developing scientific software...



# The future: Keep building bridges



...you could be next!

# **Statistical and Computational Methods for Complex Systems**

**George G Vega Yon**

<https://ggyv.cl>

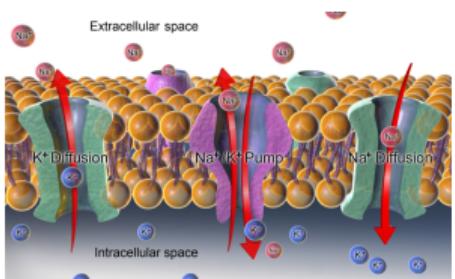
**vegayon@usc.edu**

# **Thank you!**

Gene functions can be classified in three types:

## Molecular function

Active transport GO:0005215



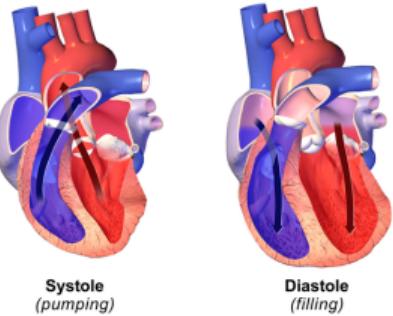
## Cellular component

Mitochondria GO:0004016



## Biological process

Heart contraction GO:0060047



◀ go back

# 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

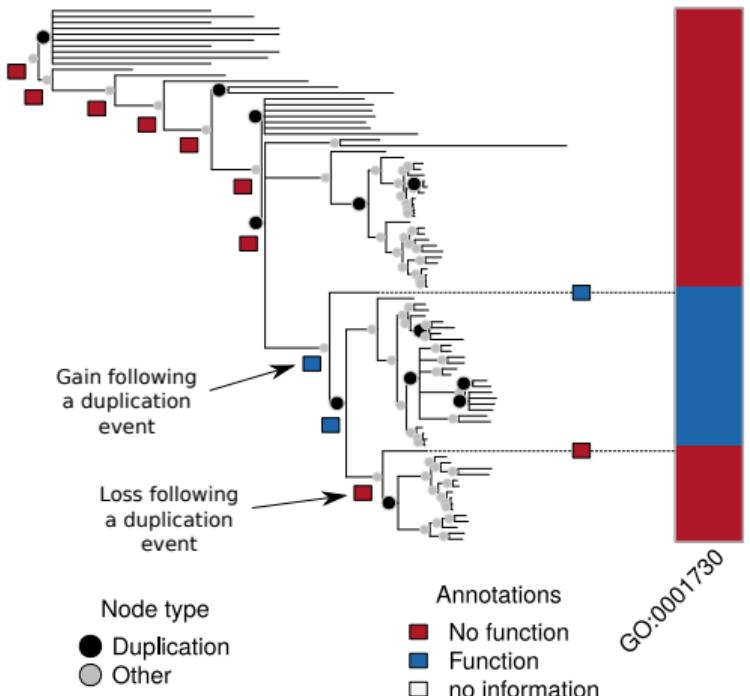
---

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

You know what is interesting about this function?

◀ go back

# An evolutionary model of gene functions



- ▶ Starting with the root node (no function in this case).
- ▶ Passes the baton to its offspring.
- ▶ Possibly without change (on this particular function).
- ▶ Or, with some probability, gaining... or loosing the function.
- ▶ Until it reaches the end of the tree (modern genes).

▶ more on duplication

▶ duplication vs speciation

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



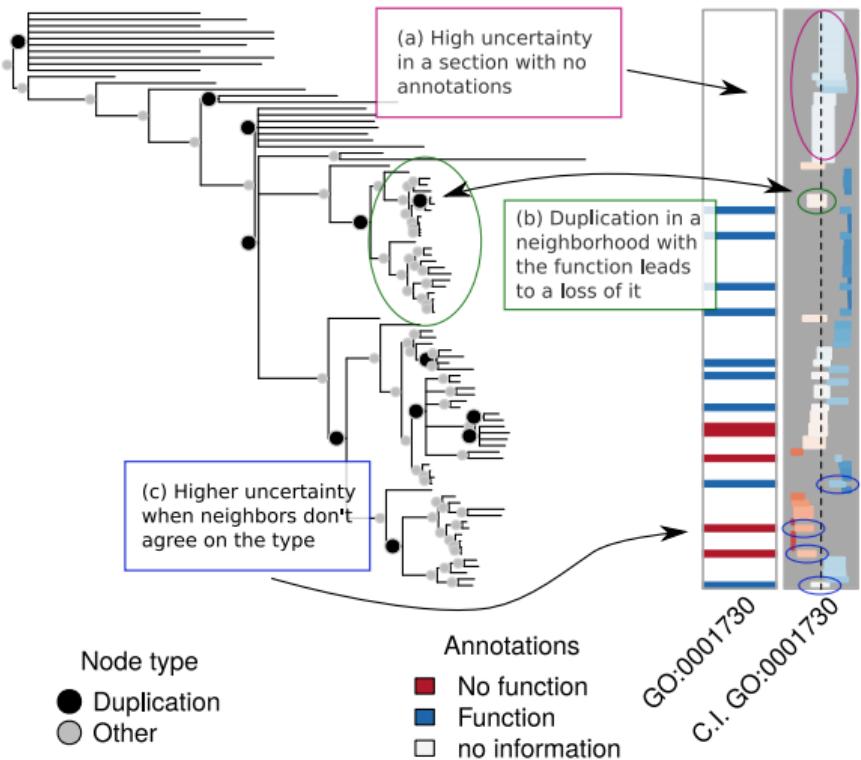
*Anolis carolinensis* pthr11521



*Equus caballus* pthr24356

[◀ go back](#)

## Example of Data + Predictions

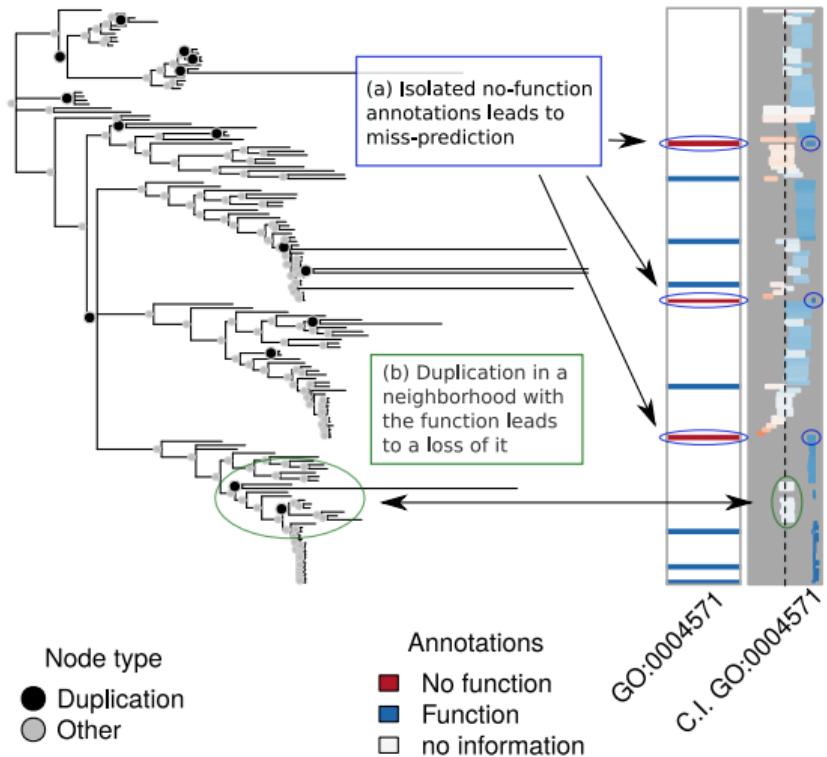
**Family: PTHR11258****Type:** Molecular Function**Name:** 2'-5'-oligoadenylate synthetase activity**Desc:** GO:0001730 involved in the process of cellular antiviral activity (wiki on [interferon](#)).**MAE:** 0.34**AUC:** 0.91[see a bad one](#)[◀ go back](#)

## Example 2: Bad quality prediction

MAE: 0.52

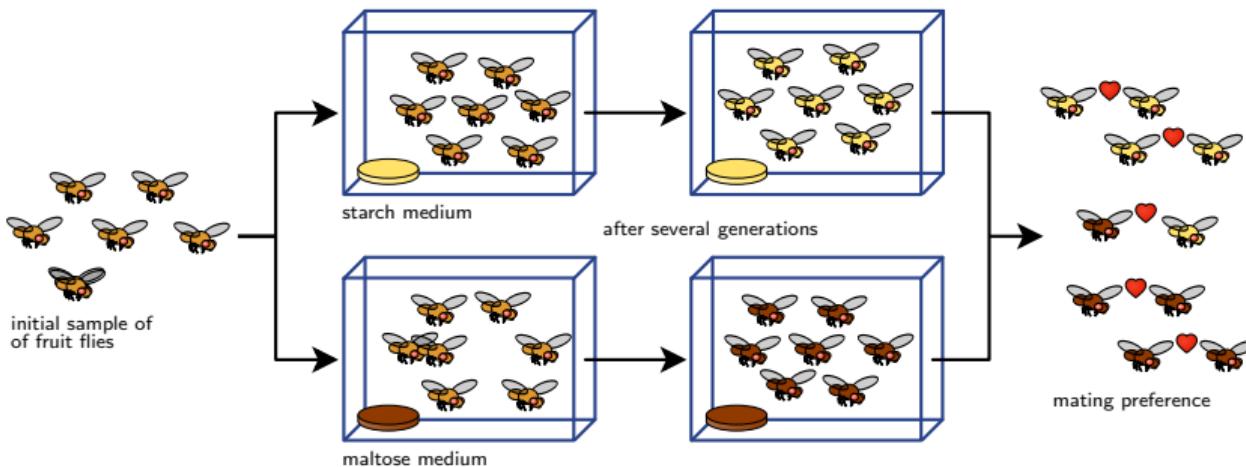
AUC: 0.33

Type: Molecular Function

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

		Pooled-data	One-at-a-time	
		Beta prior	Unif. prior	Beta Prior
Pooled-data				
Unif. prior	Beta prior	[-0.02,-0.01]	[-0.14,-0.10]	[-0.06,-0.03]
	Beta prior	-	[-0.12,-0.09]	[-0.04,-0.01]
One-at-a-time				
Unif. prior	Beta prior	-	-	[ 0.06, 0.09]

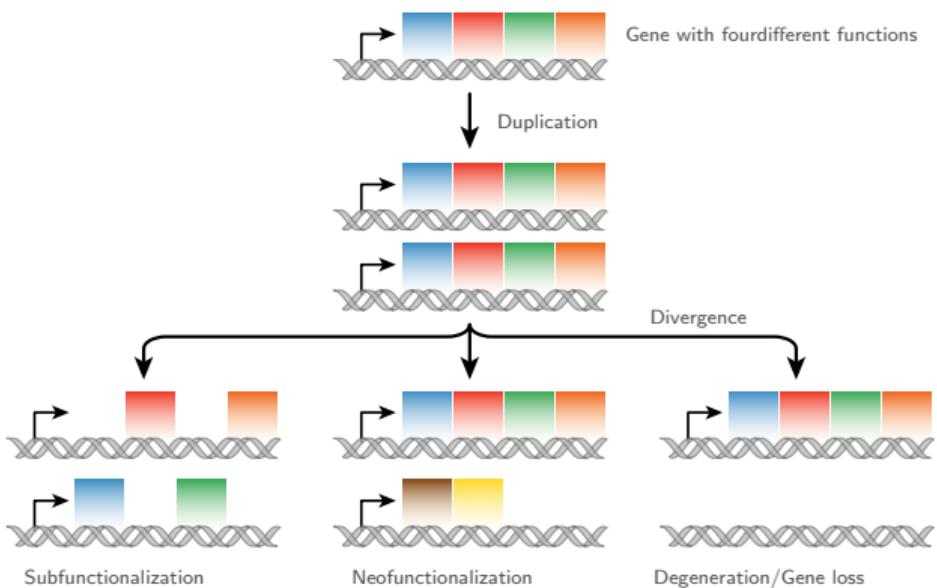
**Table 3** Differences in Mean Absolute Error [MAE]. Each cell shows the 95% confidence interval for the difference in MAE resulting from two methods (row method minus column method). Cells are color coded blue when the method on that row has a significantly smaller MAE than the method on that column; Conversely, cells are colored red when the method in that column outperforms the method in that row. Overall, predictions calculated using the parameter estimates from *pooled-data* predictions outperform *one-at-a-time*.



**Figure 3** Dodd 1989: After one year of isolation, flies showed a significant level of assortativity in mating (wikimedia)

◀ go back

# Duplication



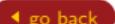
**Figure 4** A key part of molecular innovation, gene duplication provides opportunity for new functions to emerge  
(wikimedia)

◀ go back

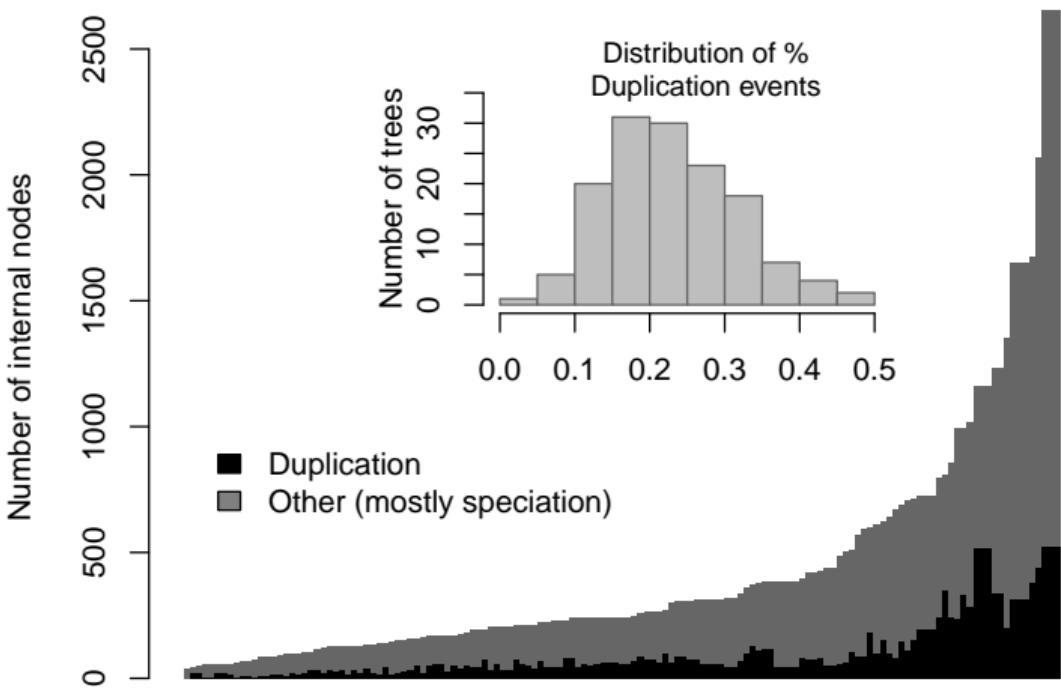
## Data: Phylogenetic trees

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](#)

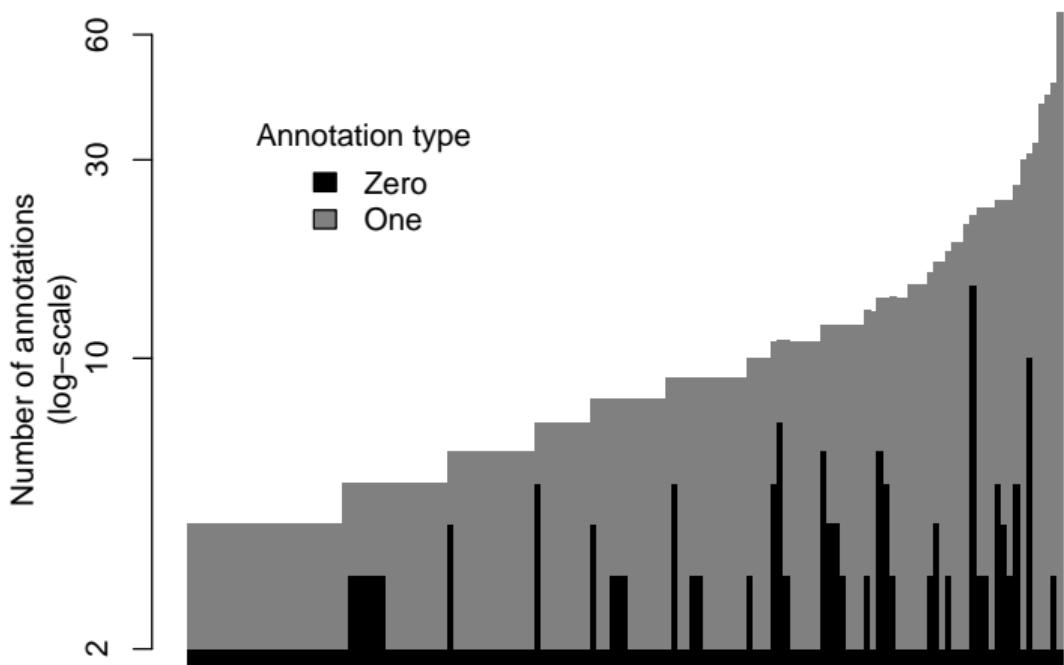
## Data: Annotations (example)

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	

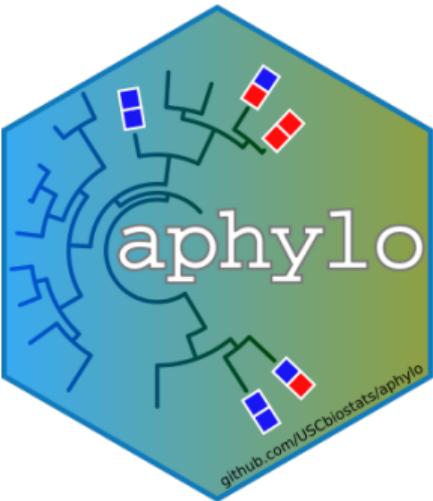
◀ go back

## Data: Experimental Annotations



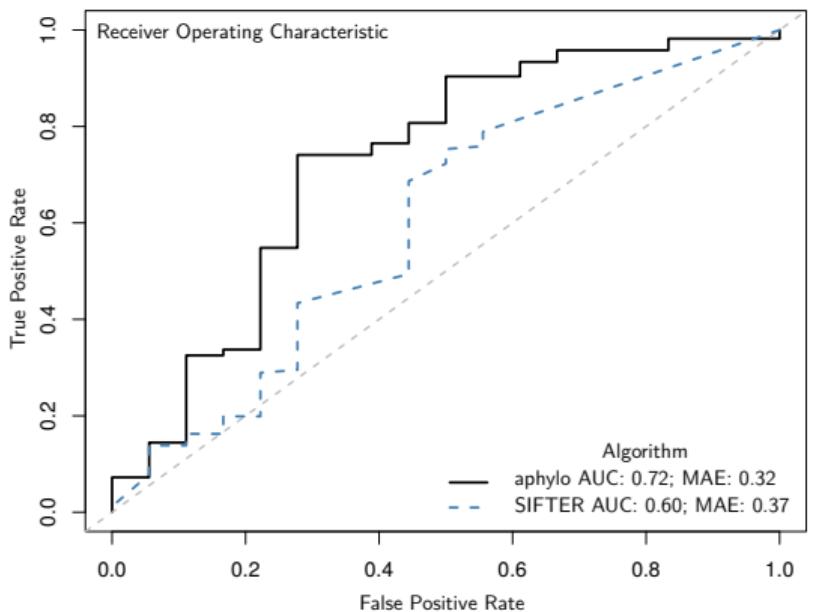
## Results: Implementation and Large scale study

- ▶ 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](#)

## Results: Performance and Scalability

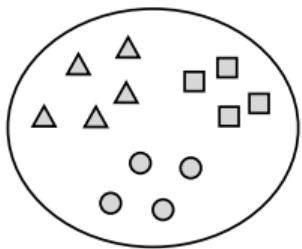
aphylo vs SIFTER (state-of-the-art phylo-based model) on 147 genes.



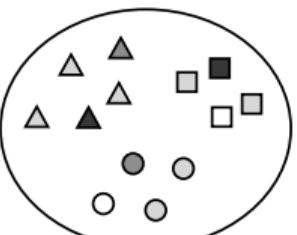
**Fast** 110 minutes (SIFTER) to calculate the posterior probabilities, aphylo took 1 second.

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

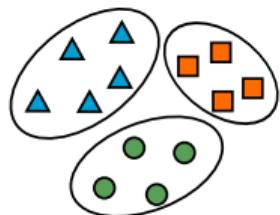
# Phylogenetics Modeling: Pooling data



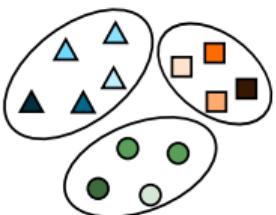
(a) Fixed rate accross functions



(b) Random rate accross functions



(c) Fixed rate within type



(d) Random rate within type

- (a) Featured in the first version of the model.
- (b) “Full glory” Hierarchical Bayes (1,001 parameters for the 141 functions).
- (c) Distilled version (a), improves accuracy.
- (d) Model estimated for Molecular Function (using Empirical Bayes) without significant improvements.

All methods are now available in the `aphylo` package: `aphylo_mle`, `aphylo_mcmc`, and `aphylo_hier`.

[◀ go back](#)

# 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	

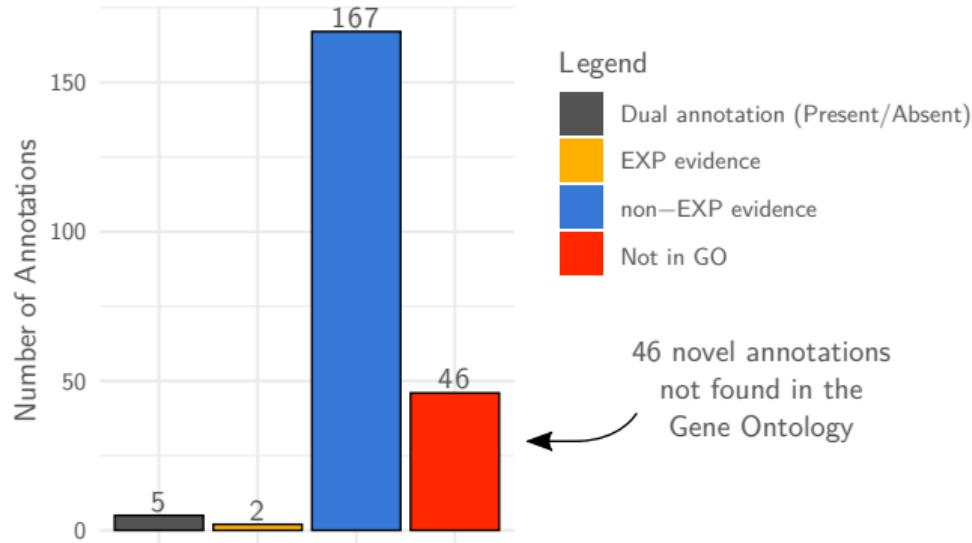
Previously, joint estimates out-performed one-at-a-time

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

Molecular Function  $\neq$  Biological Process ? Cellular Component

▶ data

▶ go back



**Figure 5** Distribution of predictions

◀ go back

## Asymptotic Behavior of ERGMs

- ▶ In the case that  $s_l = s(\mathbf{g}, x)$  is on the boundary:  $s_l \rightarrow \pm\infty$
- ▶ Since the support space of  $s(\mathbf{g}, x) \in \mathcal{S}$  is bounded, e.g. # edges  $\in [0, n \times (n - 1)]$ , we have:

$$\lim_{\theta_l \rightarrow \infty} l(\theta), \quad \lim_{\theta_l \rightarrow \infty} \nabla l(\theta), \quad \lim_{\theta_l \rightarrow \infty} \mathbf{H}(\theta)$$

log-likelihood, its gradient, and hessian are finite.

- ▶ The direct implication is that, while  $s(\mathbf{g}, x)$  is on the boundary, the MLE for the other statistics exists.<sup>1</sup>
- ▶ All equations ultimately involve realizations of  $s(\mathbf{g}', x)$  that equal  $s_l$ , relevant in: Simulations, Bootstrapping, etc.

◀ go back

<sup>1</sup>Handcock 2003 briefly mentions this

# Conditional Graph Distributions

- ▶ Long history in (soc.) network science.
- ▶ Common usage: Hypothesis test prevalence of a feature.

*Is the observed count of XYZ within the expected in a Bernoulli graph?*

*Are statistics A, B, and C different from graphs with 5 triangles?*

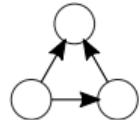
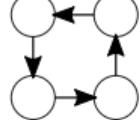
- ▶ Different names, same thing, e.g. CUG tests and rewiring algorithms.
- ▶  $\{\text{CUG, Rewiring}\} \subset \text{ERGM}$
- ▶ We can talk about *Conditional ERGMs*.

$$\mathbb{P}(s(\mathbf{G})_k = s_k \mid s(\mathbf{G})_l = s_l, \theta) = \frac{\exp\{\theta_{-l}^t s(\mathbf{g})_{-l}\}}{\sum_{\mathbf{g}' : s(\mathbf{g}')_l = s_l} \exp\{\theta_{-l}^t s(\mathbf{g}')_{-l}\}}$$

In this equation  $\theta_l$  becomes a nuisance parameter.

◀ go back

## Sufficient statistics have various forms

Representation	Description
	Mutual Ties (Reciprocity) $\sum_{i \neq j} y_{ij} y_{ji}$
	Transitive Triad (Balance) $\sum_{i \neq j \neq k} y_{ij} y_{jk} y_{ik}$
	Homophily $\sum_{i \neq j} y_{ij} \mathbf{1}(x_i = x_j)$
	Attribute-receiver effect $\sum_{i \neq j} y_{ij} x_j$
	Four Cycle $\sum_{i \neq j \neq k \neq l} y_{ij} y_{jk} y_{kl} y_{li}$

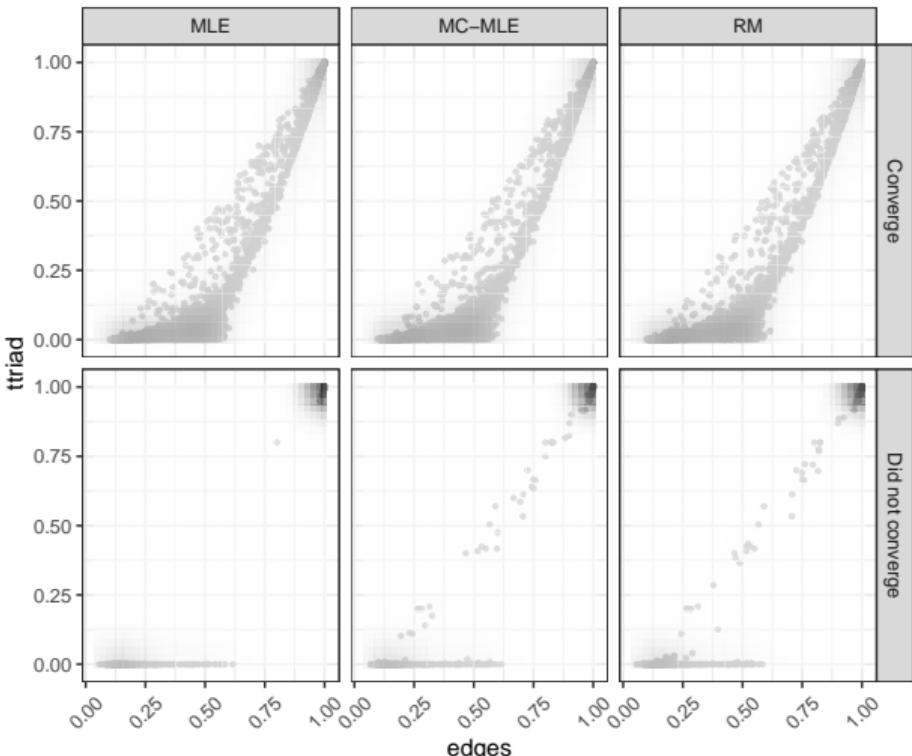
**Figure 6** Besides the common edge count statistic (number of ties in a graph), ERGMs allow measuring other more complex structures that can be captured as sufficient statistics.

◀ go back

# Simulation Study

## 1. Higher convergence rate

◀ return



# Simulation Study

1. Higher convergence rate
2. **Smaller bias**

◀ return

	MLE	MC-MLE	RM
edges	[0.27, 0.36]	[1.23, 1.65]	[0.55, 1.54]
ttriads	[-0.05, -0.03]	[-0.22, -0.16]	[-0.15, 0.48]

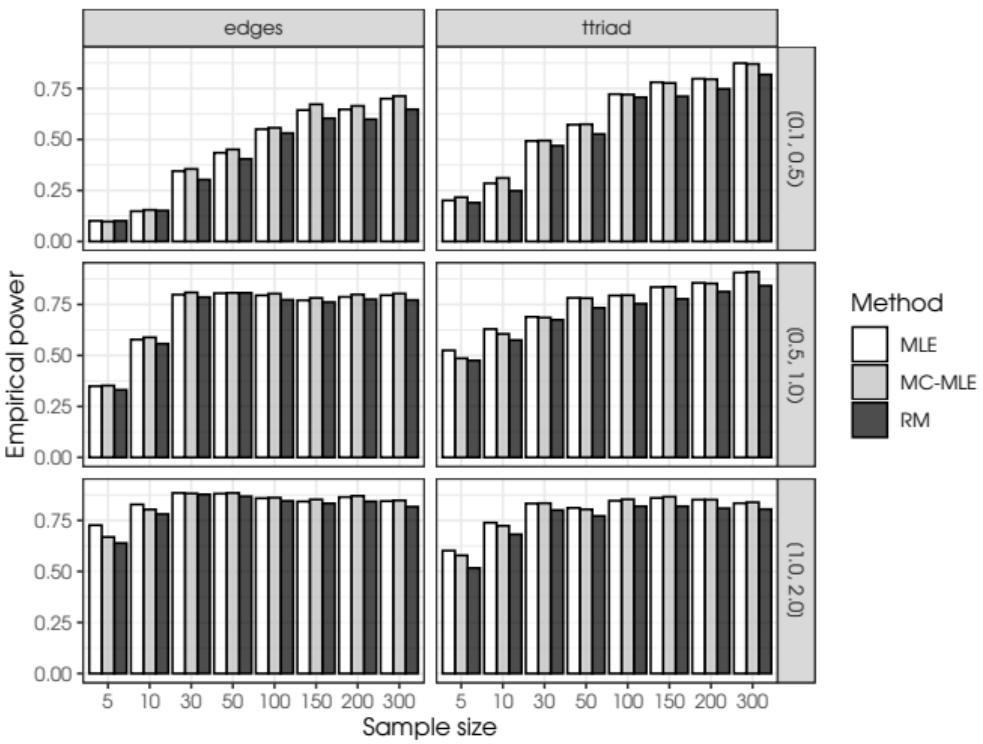
**Table 4** Empirical bias. Each cell shows the 95% confidence interval of each methods' empirical bias.

▶ alt take

# Simulation Study

1. Higher convergence rate
2. Smaller bias
3. **Higher power**

◀ return



1. Higher convergence rate
2. Smaller bias
3. Higher power
4. **Smaller type I error**

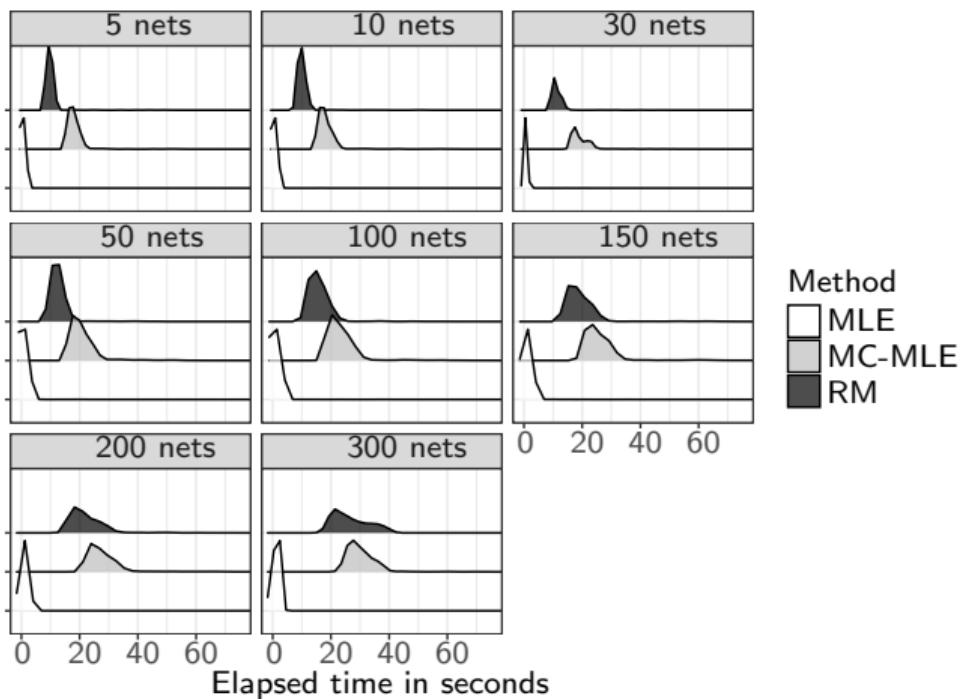
	Sample size	N. Sims.	P(Type I error)		$\chi^2$ (vs MLE)	
			MLE	MC-MLE	RM	MC-MLE
5	4,325	0.066	0.086	0.086	11.36 ***	11.36 ***
10	4,677	0.063	0.078	0.073	8.44 ***	3.73 *
15	4,818	0.060	0.072	0.063	5.50 **	0.41
20	4,889	0.054	0.065	0.061	5.30 **	2.05
30	4,946	0.053	0.059	0.055	1.60	0.07
50	4,987	0.053	0.055	0.047	0.16	1.67
100	4,999	0.054	0.054	0.050	0.00	0.81

◀ return

# Simulation Study

1. Higher convergence rate
2. Smaller bias
3. Higher power
4. Smaller type I error
5. **Elapsed time**

◀ return



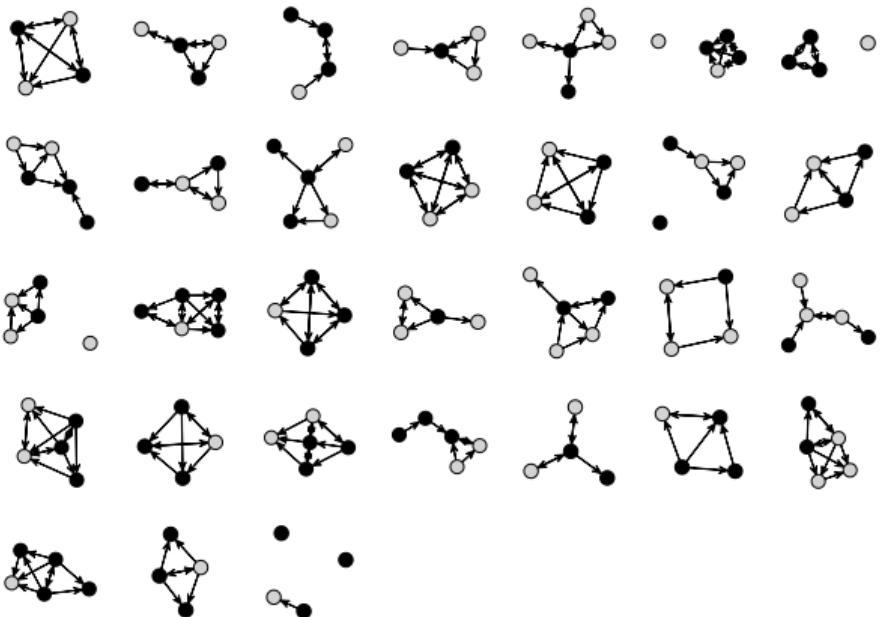
## Featured : Small Teams

### Sample

- ▶ 31 mixed-gender teams (4-5 members).
- ▶ University students.
- ▶ Don't know each other.
- ▶ At least one female/male per team.

### Experiment

- ▶ Complete 1 hour of group tasks.
- ▶ Captured network data using name generator survey: *Who did you go to for advice, information or help to complete the group task?*



Is Gender Homophily a feature of these graphs?

◀ go back

	(1)	(2)	(3)	(4)	(5)	(4b)
edges	-0.52** (0.17)	-0.91*** (0.23)	-0.54** (0.18)	-0.72*** (0.19)	-0.48* (0.19)	-0.72*** (0.17)
ttriads	0.36*** (0.06)	0.46*** (0.06)	0.37*** (0.06)	0.36*** (0.06)	0.36*** (0.06)	0.36*** (0.05)
Homophily (gender)	-0.03 (0.20)	-0.01 (0.21)	-0.20 (0.46)	-0.12 (0.20)	-0.01 (0.20)	-0.12 (0.20)
edges × 1 ( $n = 5$ )	-0.53*** (0.12)	-0.47** (0.16)	-0.52*** (0.13)	-0.53*** (0.13)	-0.53*** (0.12)	-0.53*** (0.13)
(Homophily) $^{1/2}$			0.54 (1.32)			
Sender (female)				0.46* (0.18)		0.46* (0.18)
Receiver (female)					-0.08 (0.18)	
<i>Constraint (offset)</i>						
edge > 4		Yes				
AIC	639.26	569.93	641.08	634.68	641.07	634.68
BIC	655.99	586.66	661.99	655.59	661.98	655.59
Num. networks	31	28	31	31	31	31
Time (seconds)	2.26	2.32	2.28	5.10	5.19	83.97
N replicates					1000	

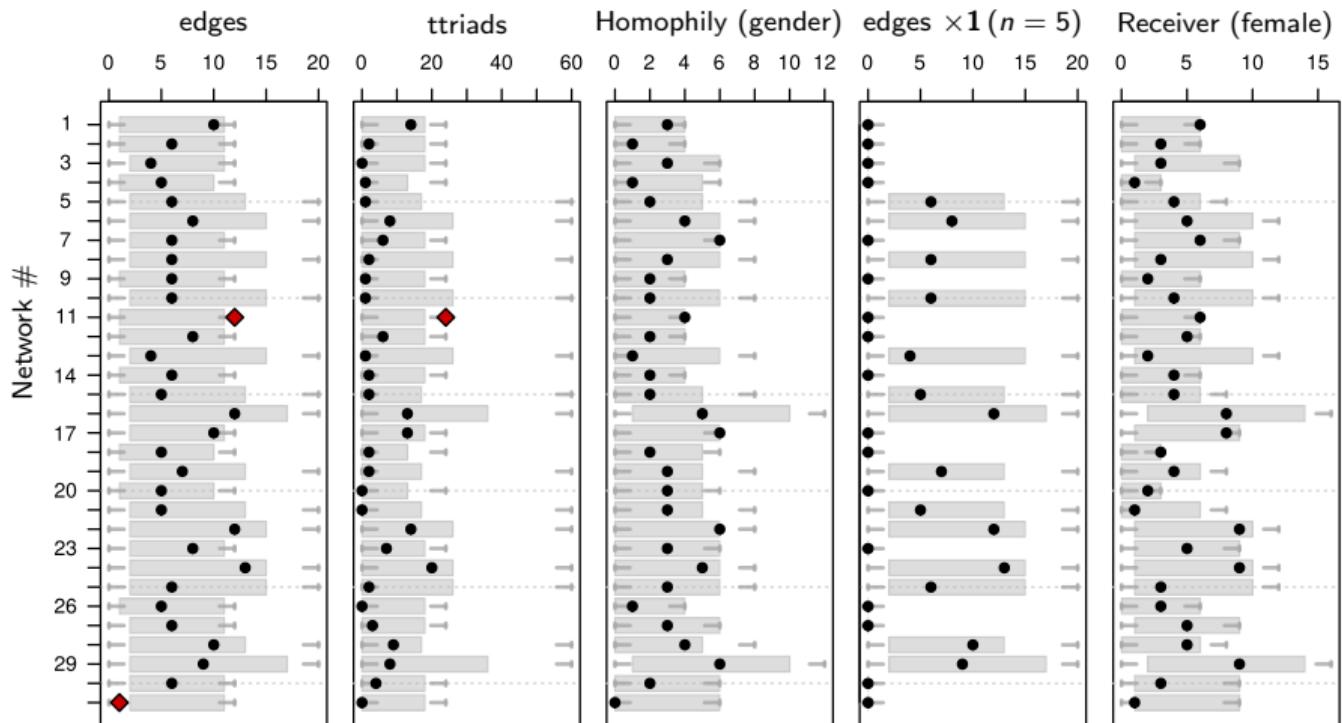
\*\*\*  $p < 0.001$ , \*\*  $p < 0.01$ , \*  $p < 0.05$

1. Interaction effects: seemingly included.
2. Transformed variables: also easy to add.
3. Using offset terms, we can constrain the support.
4. Each 1,000 bootstrap replicates took roughly 0.08 secs.
5. No support for gender homophily, but evidence of females sending more ties.

What about goodness-of-fit?

◀ go back

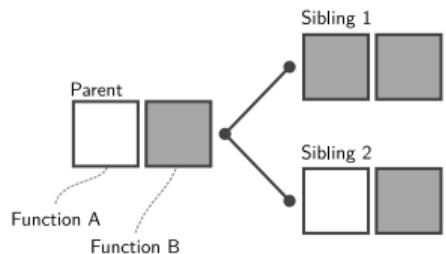
## What About Goodness-of-fit?

[◀ go back](#)

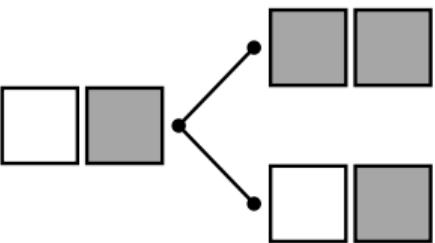
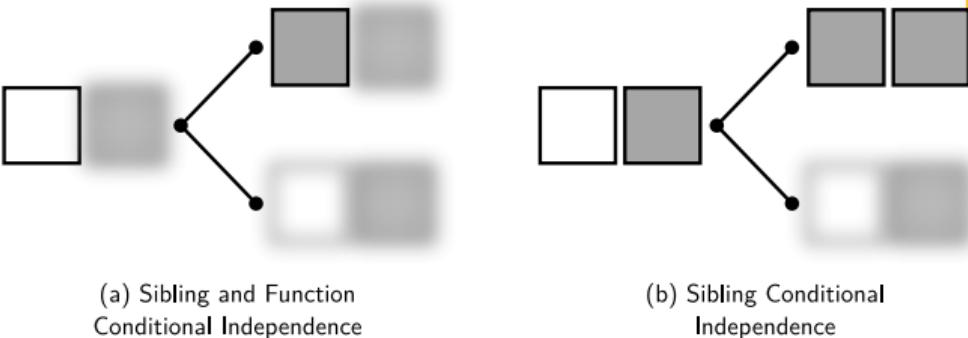
(1)	(2)	(3)	(4)	(5)	(6)
Size ( $n$ )	edges	ttriads	edges $\times$ $\mathbf{1} (n = 5)$	ttriads $\times$ $\mathbf{1} (n = 5)$	edges $\times$ $\log \{1/n\}$
4	10	14	0	0	-13.86
4	6	2	0	0	-8.32
4	4	0	0	0	-5.55
5	6	1	6	1	-9.66
5	8	8	8	8	-12.88
5	6	2	6	2	-9.66
... 25 more rows ...					

**Table 5** Example of observed sufficient statistics for the team advice networks. Pooled-data ERGMs have multiple observed sufficient statistics (also known as target statistics). Furthermore, as shown here, we can manipulate common statistics as *edges* (2) and *ttriads* (3) to include, e.g. interaction effects (4) and (5), or more complex transformations, e.g. (6).

# Phylogenetics Modeling Strategies



Has the function  
 Doesn't have the function

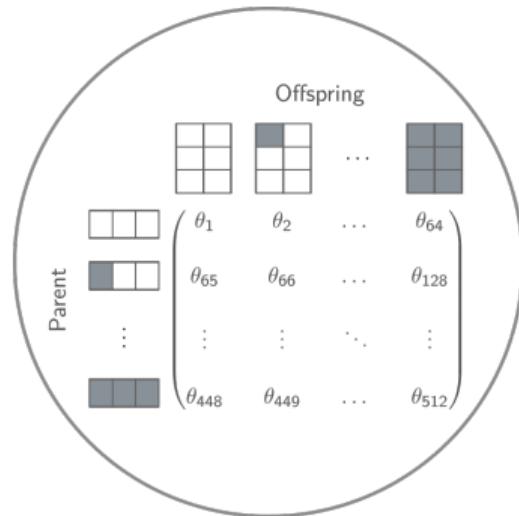


(c) No conditional  
independence

◀ go back

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

### Full Markov Transition Matrix



512 parameters

$$\begin{aligned} & (\theta_{\text{gain}1} \theta_{\text{gain}2} \theta_{\text{gain}3}) \\ & (\theta_{\text{loss}1} \theta_{\text{loss}2} \theta_{\text{loss}3}) \\ & \theta_{\text{neofun}} \theta_{\text{subfunc}} \\ & (\theta_{\text{root}1} \theta_{\text{root}2} \theta_{\text{root}3}) \end{aligned}$$

Easier to fit

Easier to interpret

### Sufficient statistics

$$\begin{aligned} & (\theta_{\text{gain}1} \theta_{\text{gain}2} \theta_{\text{gain}3}) \\ & (\theta_{\text{loss}1} \theta_{\text{loss}2} \theta_{\text{loss}3}) \\ & \theta_{\text{neofun}} \theta_{\text{subfunc}} \\ & (\theta_{\text{root}1} \theta_{\text{root}2} \theta_{\text{root}3}) \end{aligned}$$

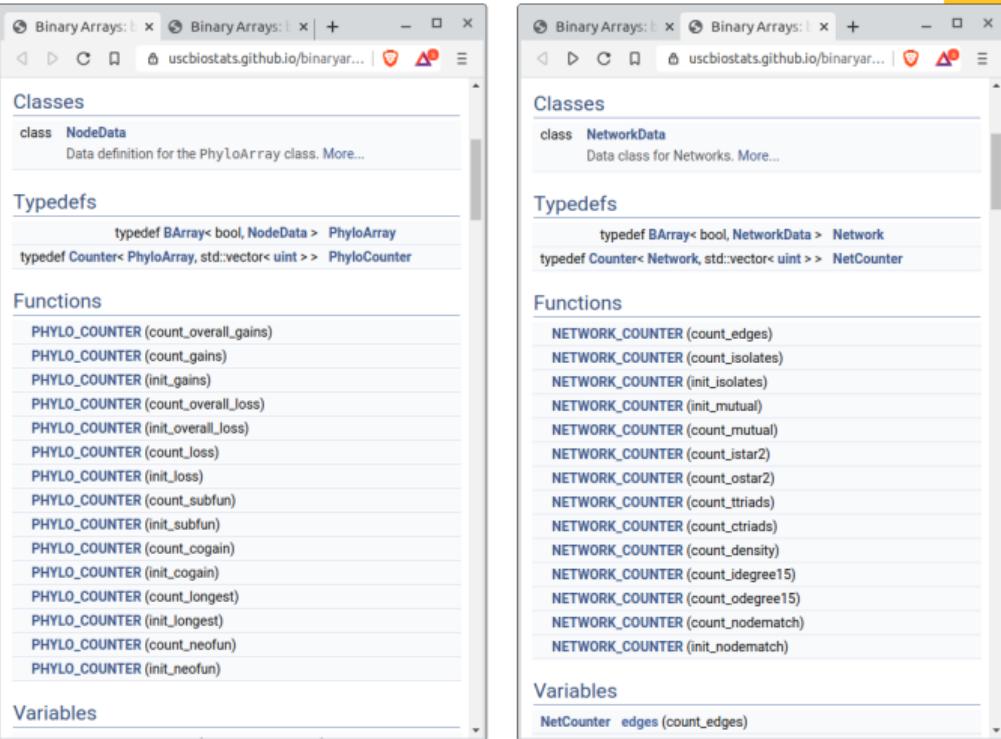
5 parameters

◀ numeric example

## Barry: your go-to *motif* accountant

- ▶ Sparse matrix represented using double hashmaps (fast row/column access).
- ▶ Template implementation for flexible weights and metadata.
- ▶ Fast counting using change statistics (Ch. 4).
- ▶ Calculation of support for sufficient stats.

[https://USCbiostats.github.io/  
binaryarrays](https://USCbiostats.github.io/binaryarrays)



**Figure 7** Screenshots from the project's website on GitHub.

# 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 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,

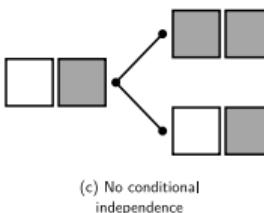
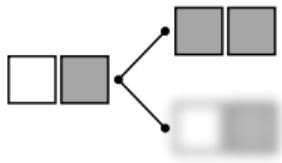
Pairwise Subfunctionalization: 3 terms,

Function specific gain: 3 terms,

Function specific loss: 3 terms,

Total: 15 parameters.

- ▶ Easier to fit and interpret.



## References |

-  Dodd, Diane M. B. (1989). "Reproductive Isolation as a Consequence of Adaptive Divergence in *Drosophila pseudoobscura*". In: Evolution 43.6, pp. 1308–1311. ISSN: 00143820, 15585646. URL: <http://www.jstor.org/stable/2409365>.
-  Handcock, Mark S. (2003). "Assessing Degeneracy in Statistical Models of Social Networks". In: Working Paper No. 39 76.39, pp. 33–50. ISSN: 1936900X. DOI: 10.1.1.81.5086. URL: <http://citeseerrx.ist.psu.edu/viewdoc/summary?doi=10.1.1.81.5086>.

## Example: Simple model with two functions

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

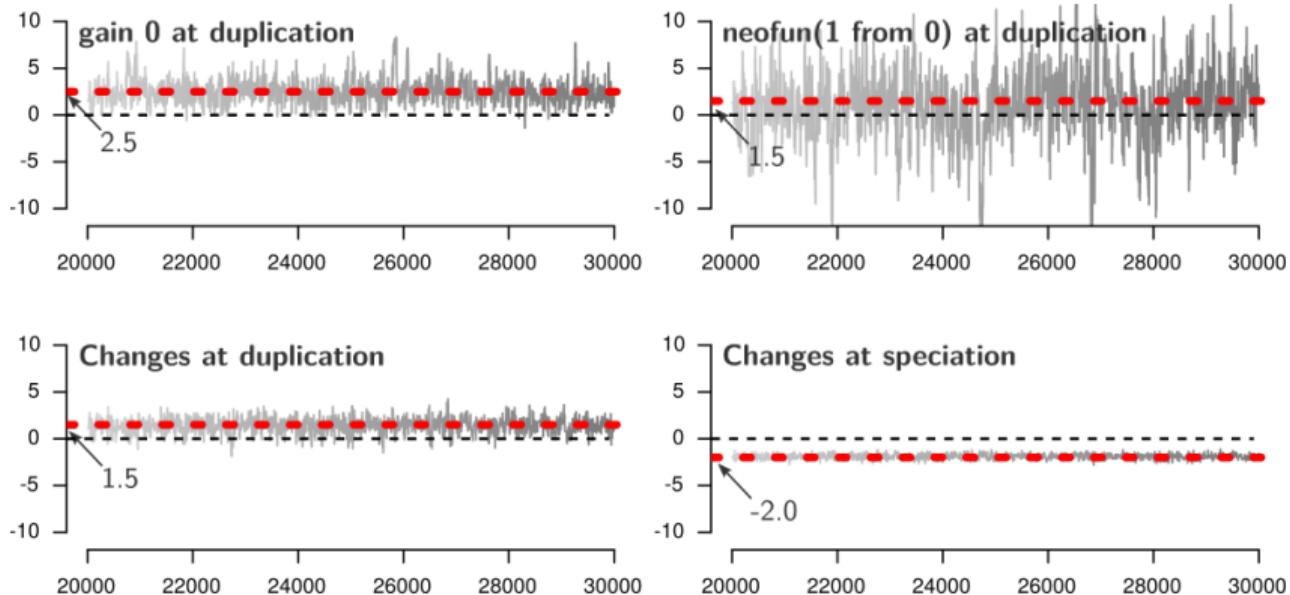
1. 100 genes on a simulated phylogenetic tree.
2. Two functions, **0** and **1**,
3. **Function 0** is likely to be gained at a dupl. event,
4. **Function 1** is gained as neofunctionalization (**from 0**) at a dupl. event,
5. There is a higher chance of **changes at duplication** (explicit).
6. Root node starts off without either function (i.e. prob  $\rightarrow 0$ ).

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

◀ go back

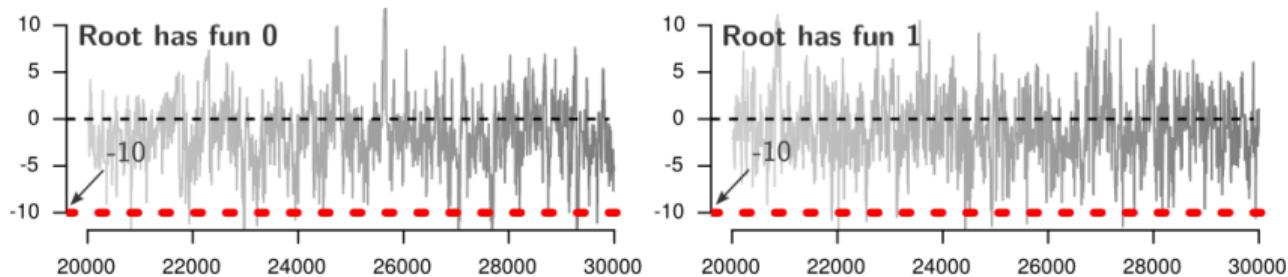
## Example: Simple model with two functions

posterior distributions



**Figure 8** 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 9** 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 10** 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.

