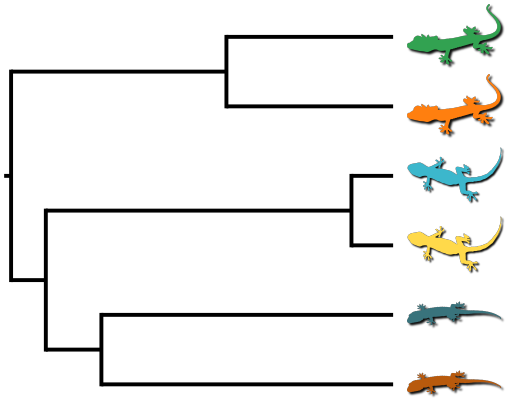
Generalizing phylogenetics to infer patterns of shared evolutionary events

MIC-Phy 2021

Jamie R. Oaks
Auburn University
phyletica.org
@jamoaks

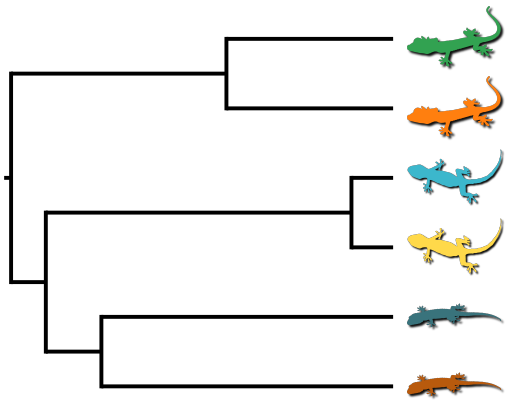
Perry L. Wood, Jr.
Auburn University
perryleewoodjr.com
@perryleewoodjr

Thanks to MIC-Phy organizers!

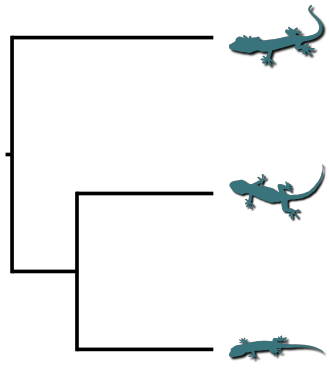


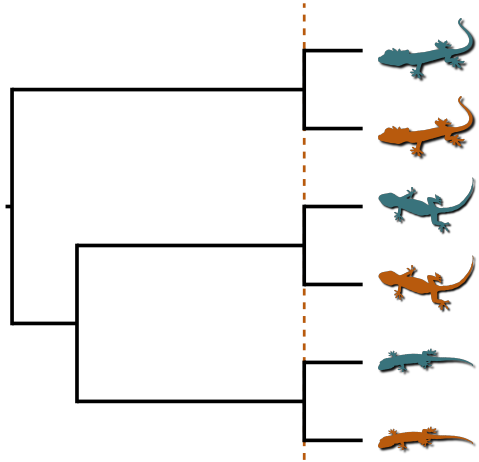
phylopic.org CC0

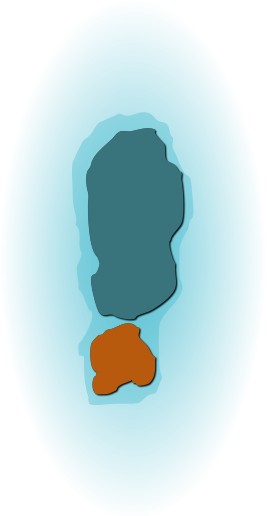
- **Assumption:** All processes of diversification affect each lineage independently and only cause bifurcating divergences.

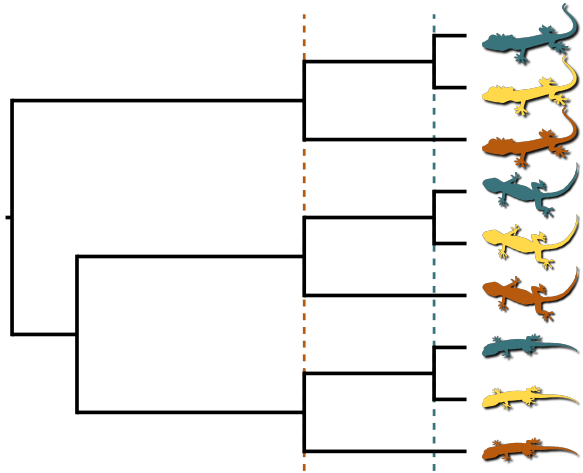


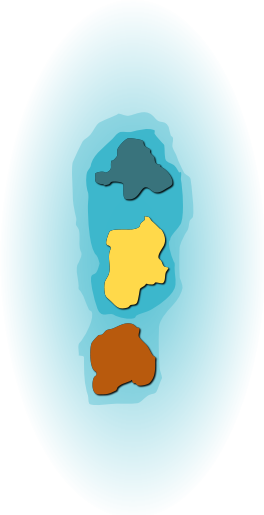
phylopic.org CC0

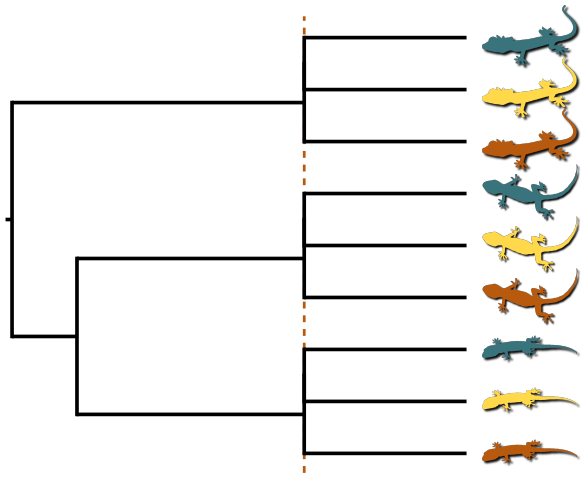


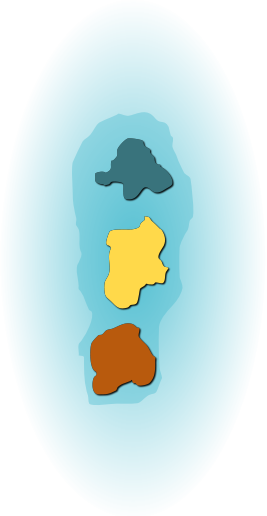






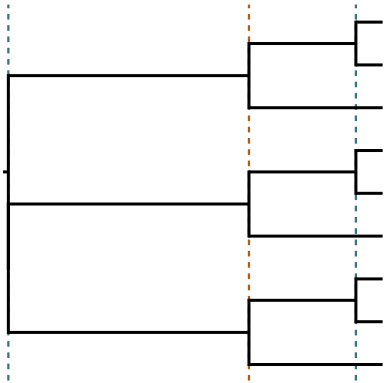






Biogeography

- Environmental changes that affect whole communities of species

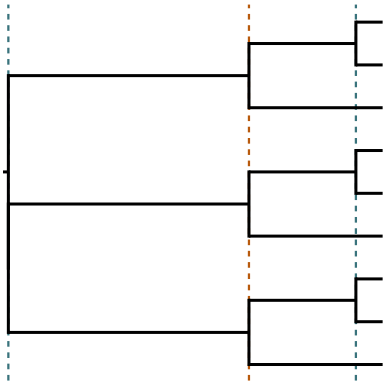


Biogeography

- Environmental changes that affect whole communities of species

Gene family evolution

- Chromosomal duplications



Biogeography

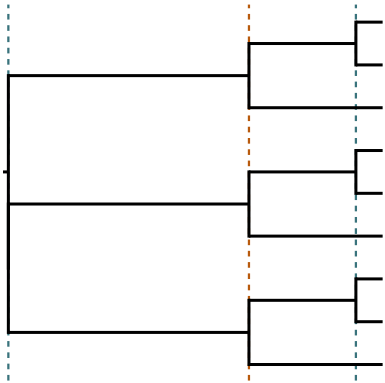
- Environmental changes that affect whole communities of species

Gene family evolution

- Chromosomal duplications

Epidemiology

- E.g., transmission at social gatherings



Biogeography

- Environmental changes that affect whole communities of species

Gene family evolution

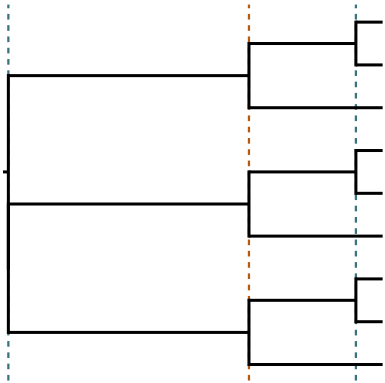
- Chromosomal duplications

Epidemiology

- E.g., transmission at social gatherings

Endosymbiont evolution (e.g., parasites, microbiome)

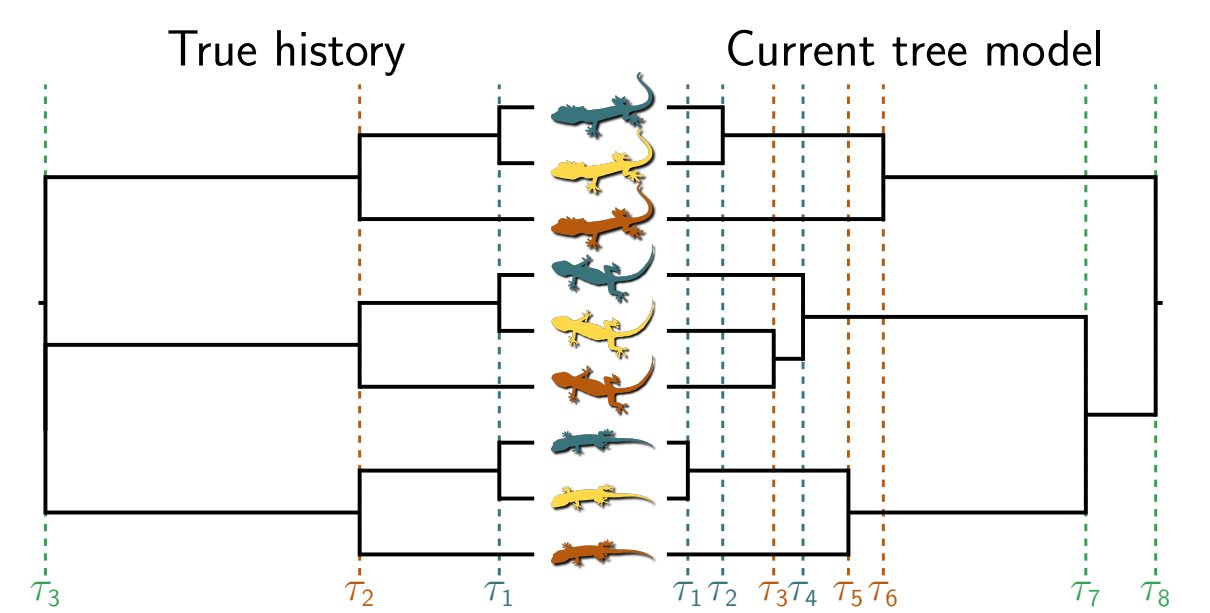
- Speciation of the host
- Co-colonization of new host species



Why account for shared divergences?

Why account for shared divergences?

1. Improve inference


True history
Current tree model
τ_3
τ_2
τ_1
τ_1 τ_2 τ_3 τ_4 τ_5 τ_6
τ_7
τ_8

Why account for shared divergences?

1. Improve inference

Why account for shared divergences?

1. Improve inference
2. **Provide a framework for studying processes of co-diversification**

Biogeography

- Environmental changes that affect whole communities of species

Gene family evolution

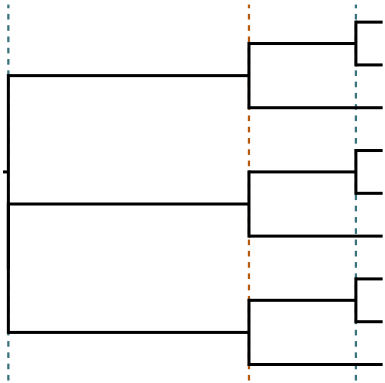
- Chromosomal duplications

Epidemiology

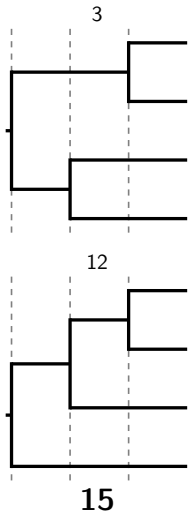
- E.g., transmission at social gatherings

Endosymbiont evolution (e.g., parasites, microbiome)

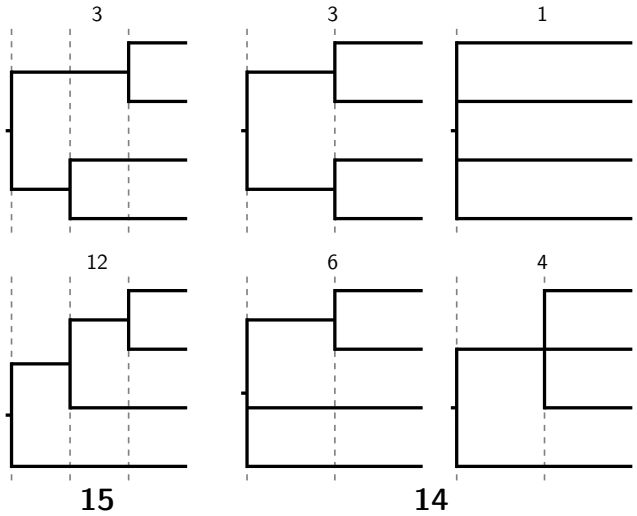
- Speciation of the host
- Co-colonization of new host species



Generalizing tree space

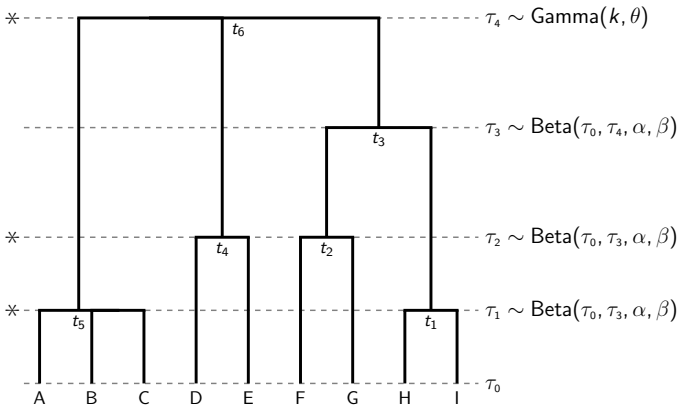


Generalizing tree space



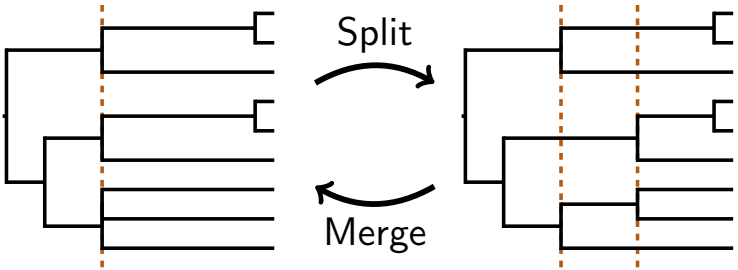
Generalized tree distribution

- All topologies equally probable
- Parametric distribution on age of root
- Beta distributions on other div times



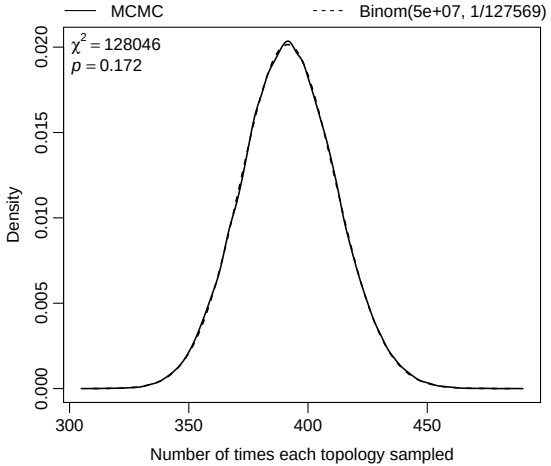
H. Kishino et al. (2001). *Molecular Biology and Evolution* 18: 352–361

Bayesian model averaging



Reversible-jump MCMC

Validating rjMCMC with 7-leaf tree



The rjMCMC algorithms sample the expected generalized tree distribution

Phycoeval (part of Ecoevolity[2])

[1] D. Bryant et al. (2012). *Molecular Biology and Evolution* 29: 1917–1932

[2] J. R. Oaks (2019). *Systematic Biology* 68: 371–395

Phycoeval (part of Ecoevolity[2])

- CTMC model of characters evolving along genealogies
- Infer species trees by analytically integrate over genealogies[1]
- rjMCMC sampling of generalized tree distribution

[1] D. Bryant et al. (2012). *Molecular Biology and Evolution* 29: 1917–1932

[2] J. R. Oaks (2019). *Systematic Biology* 68: 371–395

Phycoeval (part of Ecoevolity[2])

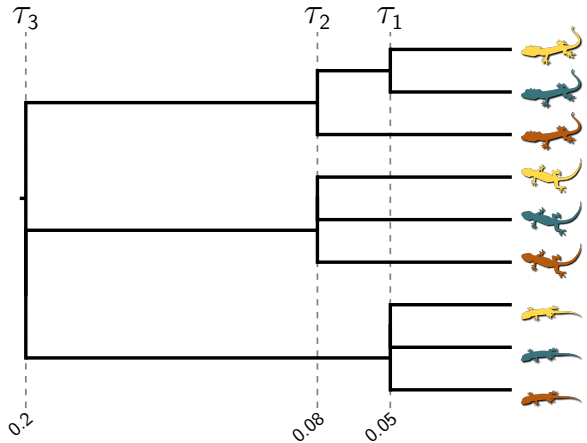
- CTMC model of characters evolving along genealogies
- Infer species trees by analytically integrate over genealogies[1]
- rjMCMC sampling of generalized tree distribution

- *Goal: Co-estimation of phylogeny and shared divergences from genomic data*

[1] D. Bryant et al. (2012). *Molecular Biology and Evolution* 29: 1917–1932
[2] J. R. Oaks (2019). *Systematic Biology* 68: 371–395

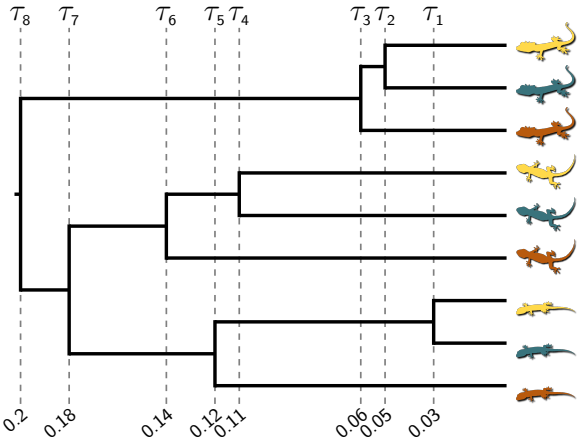
Methods: Simulations

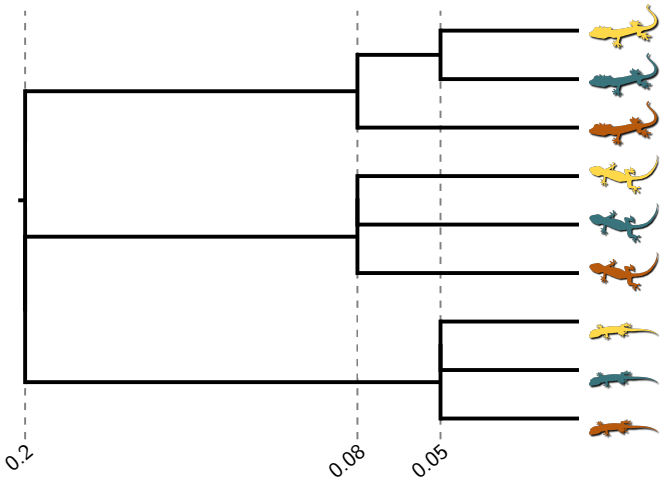
- Simulated 100 datasets with 50,000 characters
- Strict clock
- One population size
- We also did simulations where topology and div times drawn from prior

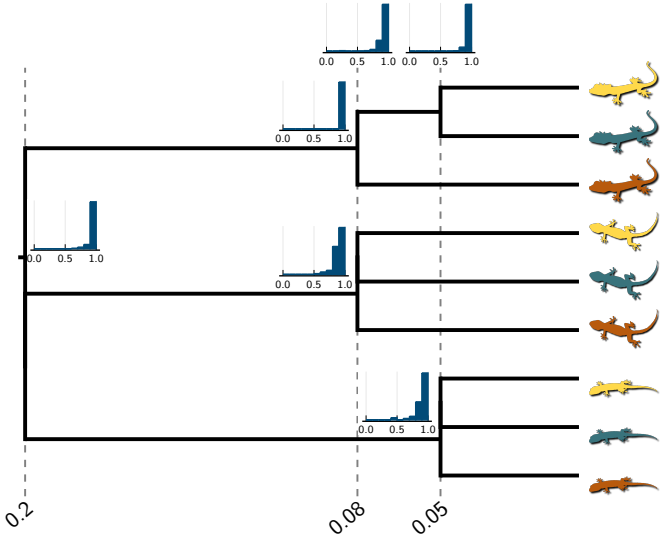


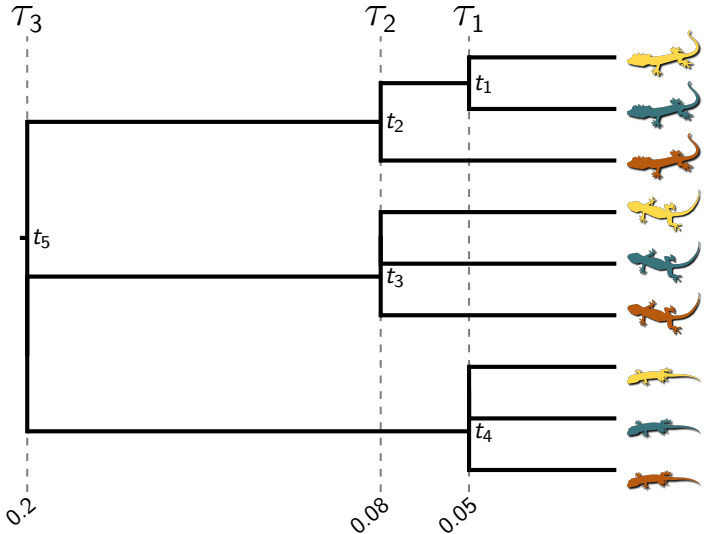
Methods: Simulations

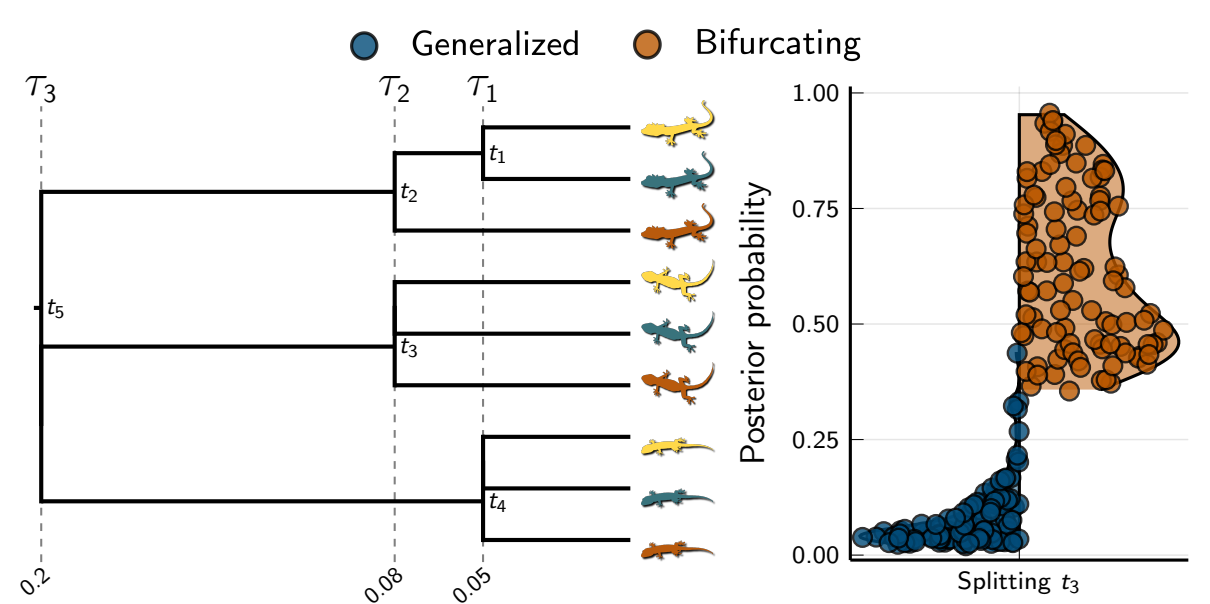
- Simulated 100 datasets with 50,000 characters
- Strict clock
- One population size
- We also did simulations where topology and div times drawn from prior

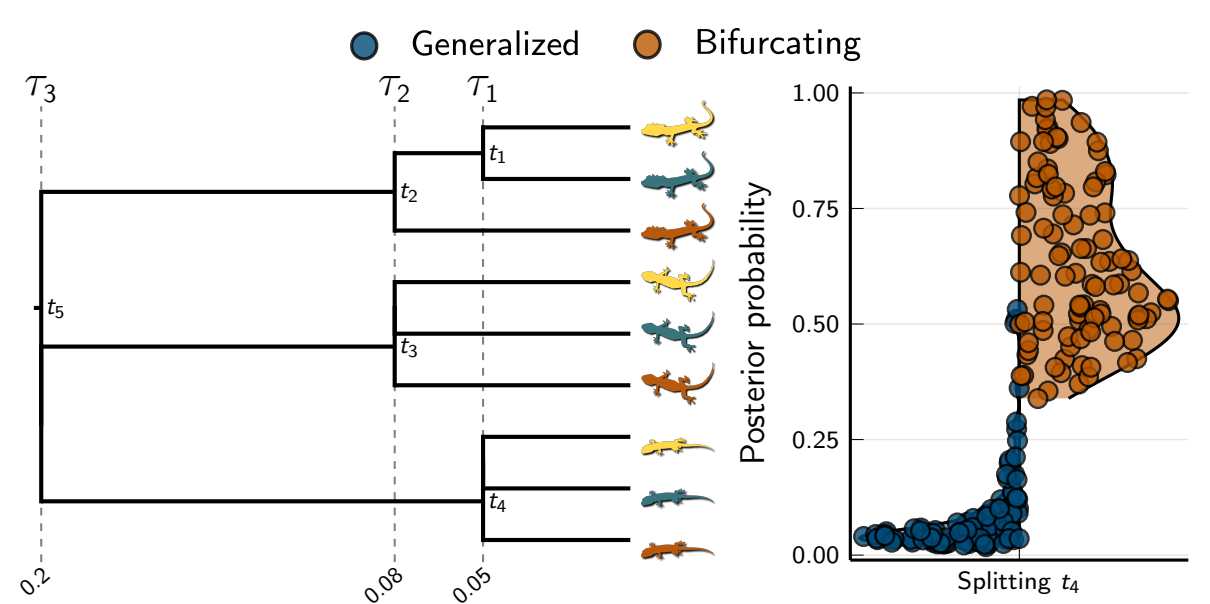


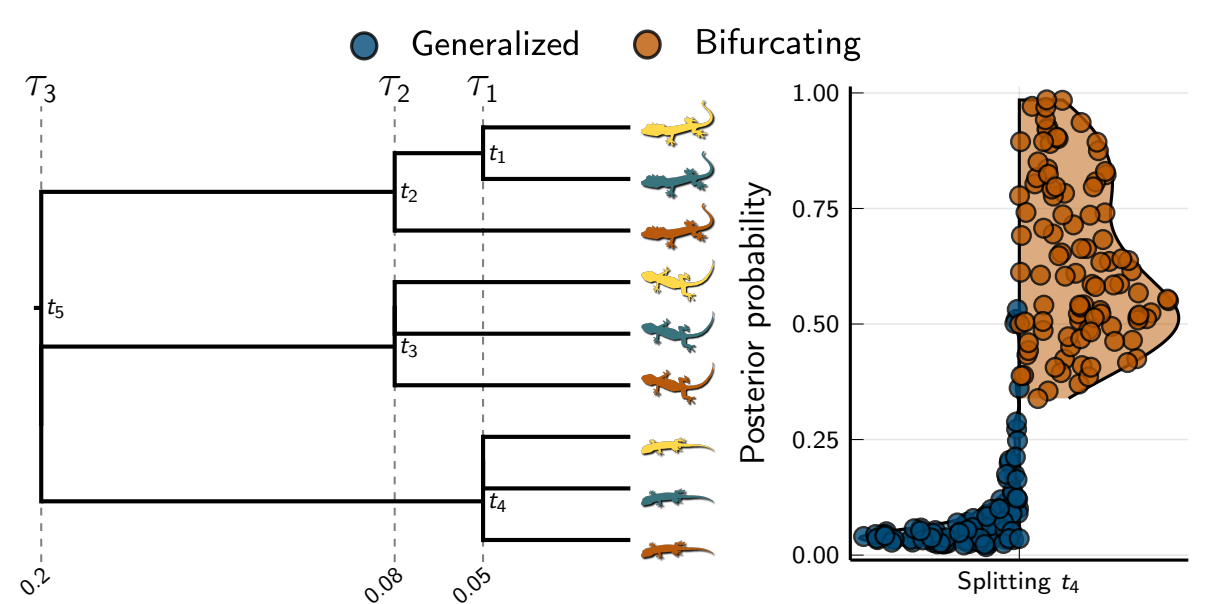

0.2
0.08
0.05


0.0 0.5 1.0
0.0 0.5 1.0
0.0 0.5 1.0
0.0 0.5 1.0
0.0 0.5 1.0
0.0 0.5 1.0
0.2
0.08
0.05


τ_3
τ_2
τ_1
t_1
t_2
t_5
t_3
t_4
0.2
0.08
0.05

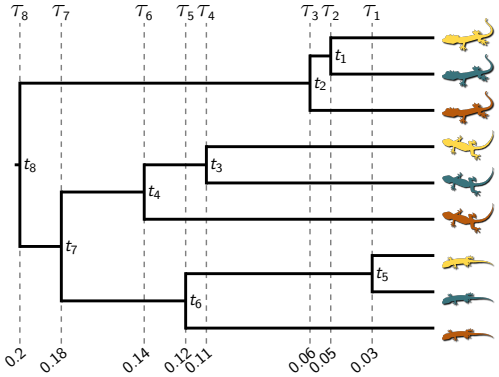

Generalized
Bifurcating
$\mathcal{T}_3$
$\mathcal{T}_2$
$\mathcal{T}_1$
t_1
t_2
t_3
t_4
t_5
0.2
0.08
0.05
Posterior probability
1.00
0.75
0.50
0.25
0.00
Splitting t_3


Generalized
Bifurcating
$\mathcal{T}_3$
$\mathcal{T}_2$
$\mathcal{T}_1$
t_1
t_2
t_3
t_4
t_5
0.2
0.08
0.05
Posterior probability
1.00
0.75
0.50
0.25
0.00
Splitting t_4

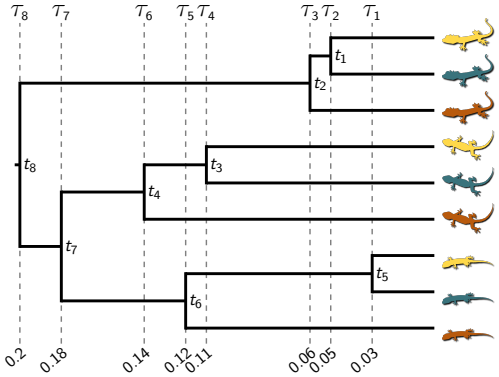


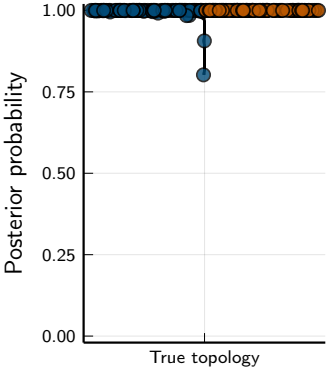
New method avoids spurious support for non-existant branches

Generalized Bifurcating

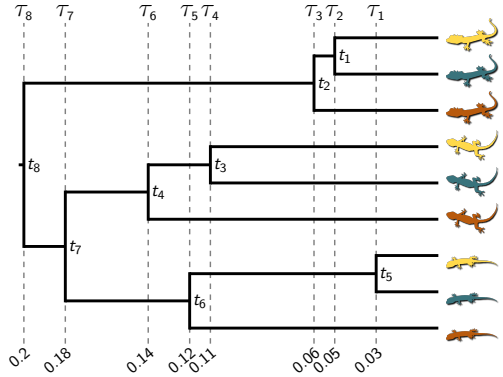


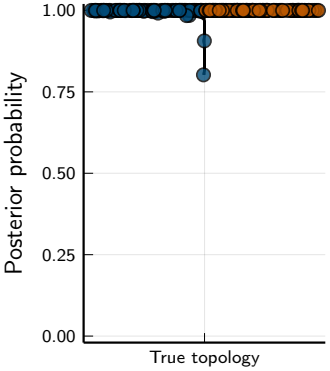
Generalized Bifurcating

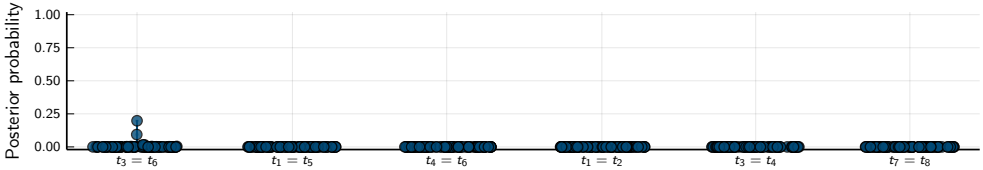




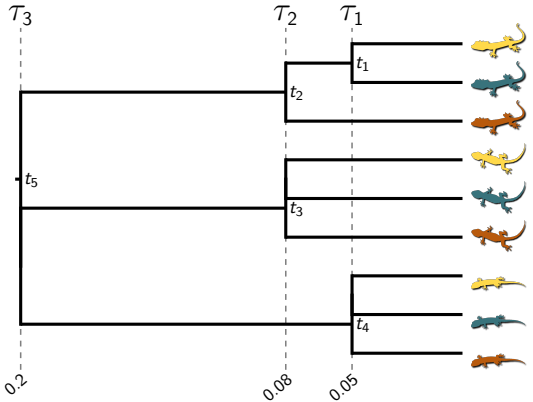
Generalized Bifurcating

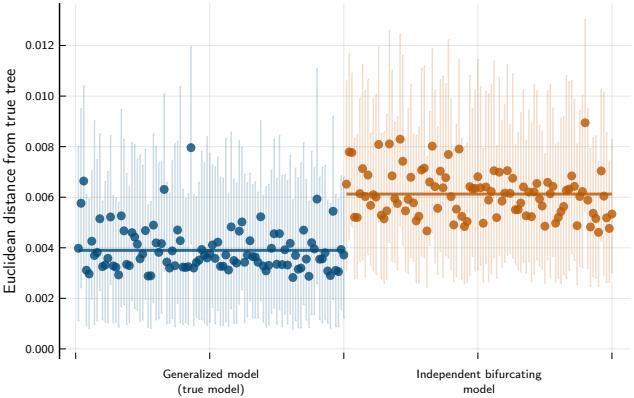






Generalized Bifurcating

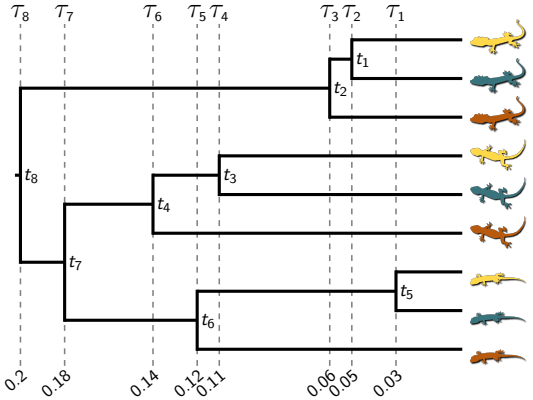


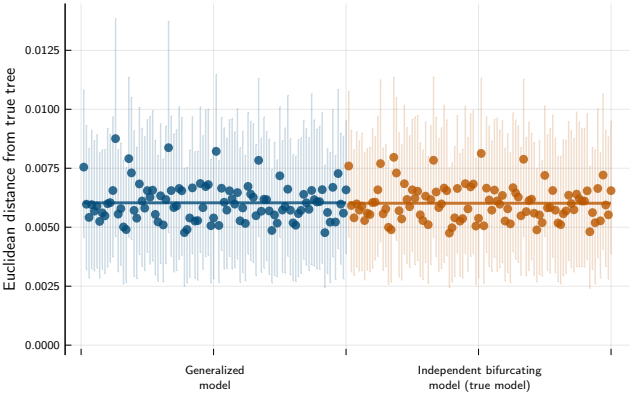


Wilcoxon signed-rank test P-value $= 4.08e^{-18}$

M. K. Kuhner et al. (1994). *Molecular Biology and Evolution* 11: 459–468

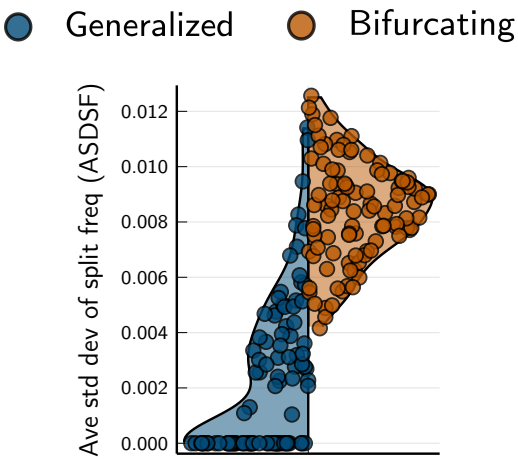
Generalized Bifurcating





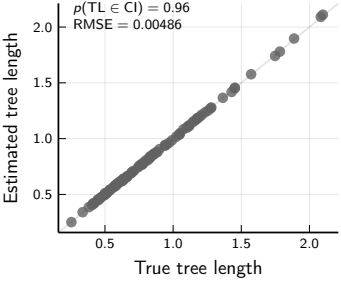
Wilcoxon signed-rank test P-value = 0.36

M. K. Kuhner et al. (1994). *Molecular Biology and Evolution* 11: 459–468

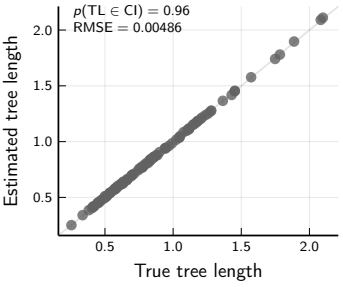


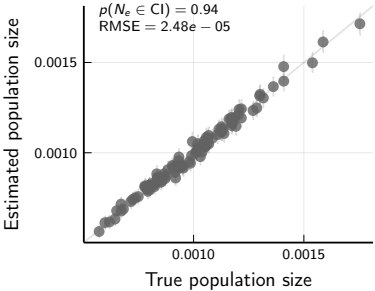
New method improves MCMC convergence and mixing

Tree-varying simulation results

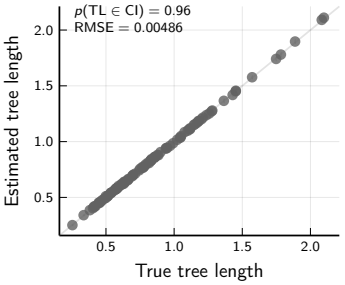


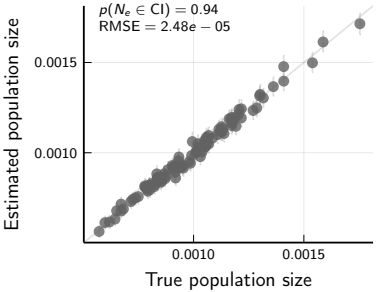
Tree-varying simulation results

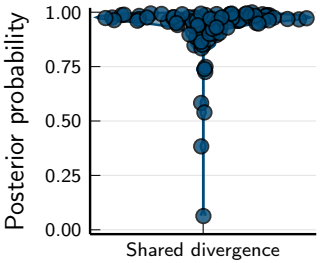




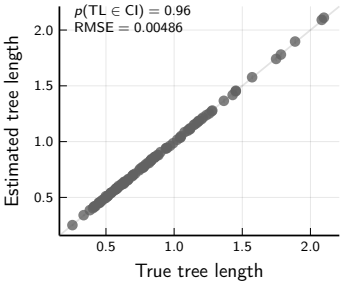
Tree-varying simulation results

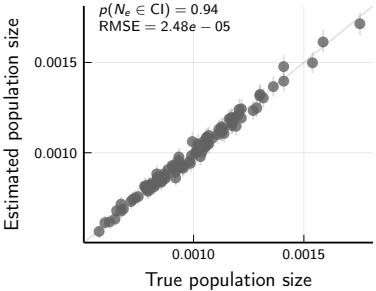


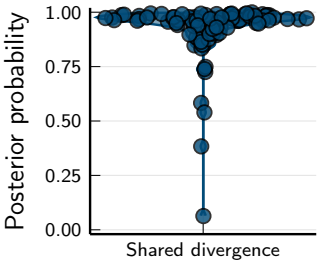




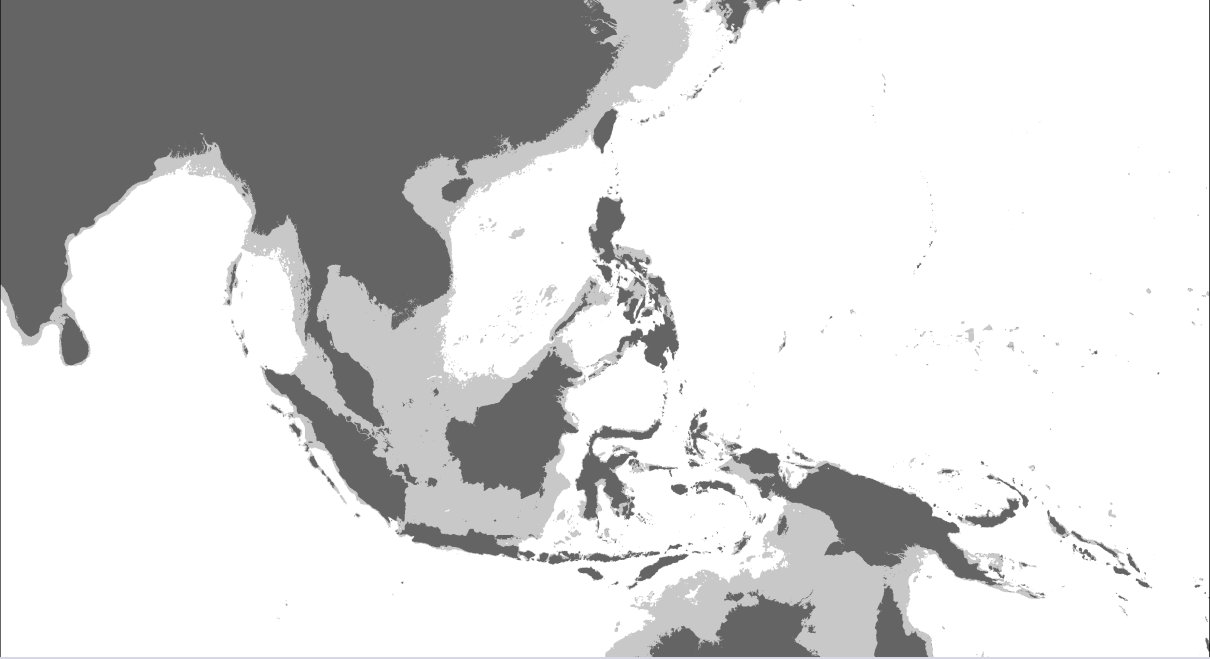
Tree-varying simulation results

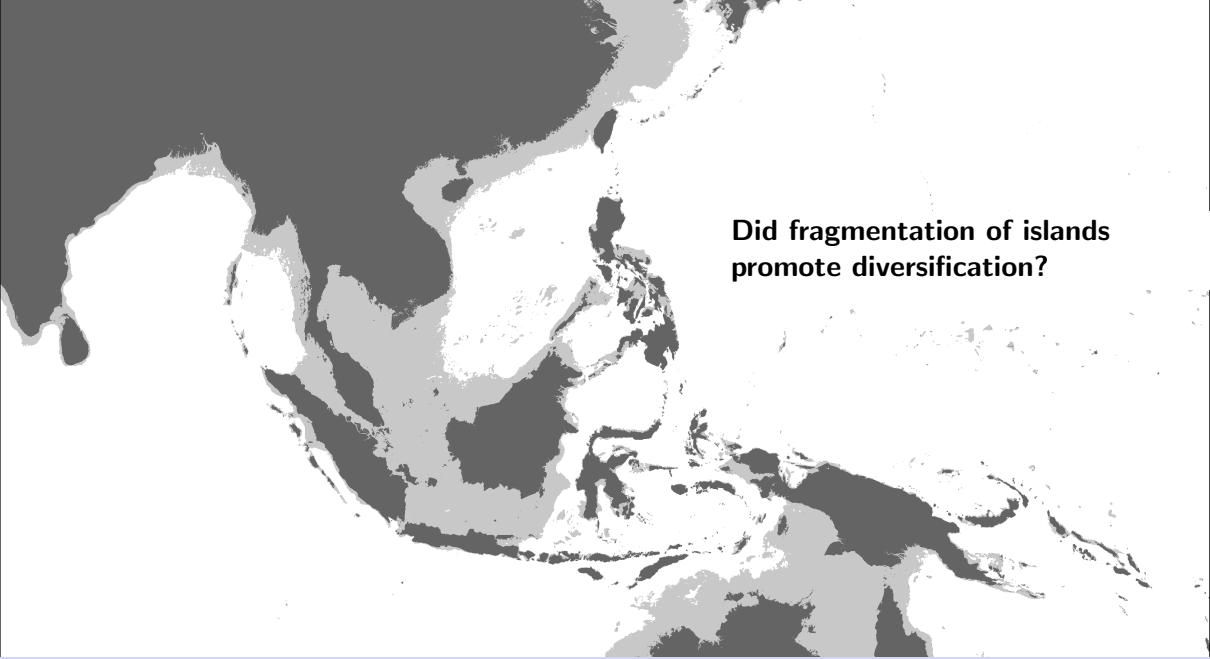


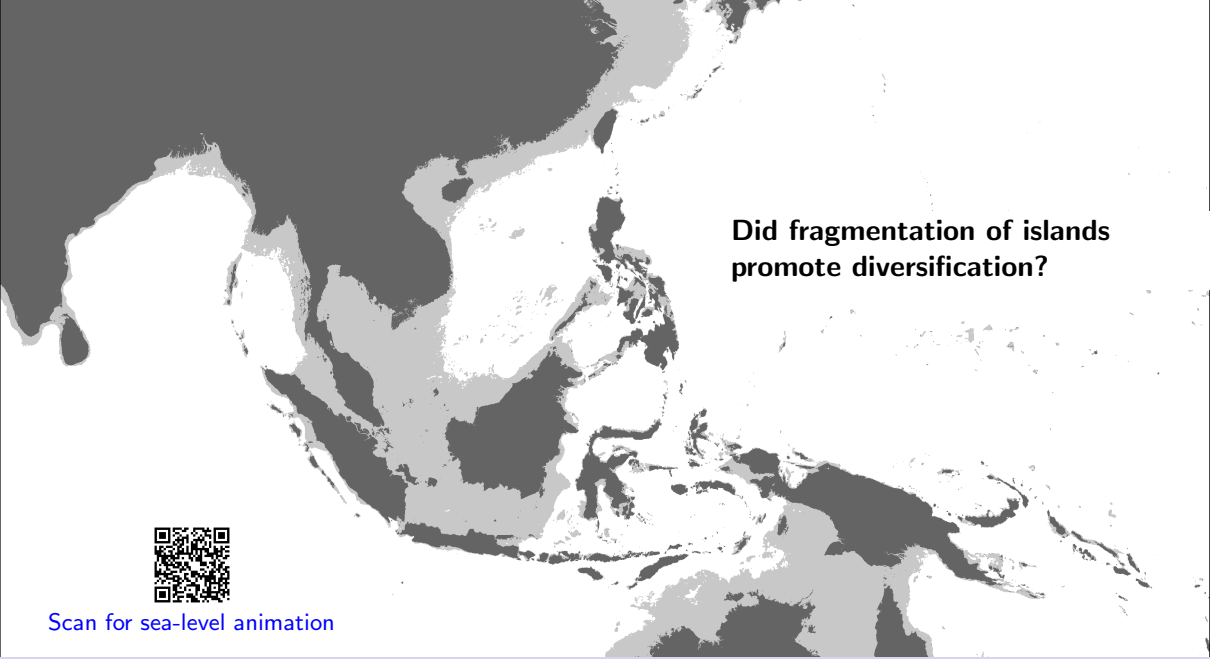




New method perform well with data simulated on random trees




Did fragmentation of islands
promote diversification?



Did fragmentation of islands promote diversification?

Scan for sea-level animation

©Rafe M. Brown

©Rafe M. Brown

J. R. Oaks et al. (2019). *Evolution* 73: 1151–1167

©Rafe M. Brown

- Sampled individuals from 27 and 26 populations across Philippines for *Cyrtodactylus* and *Gekko*, respectively

©Rafe M. Brown

J. R. Oaks et al. (2019). *Evolution* 73: 1151–1167

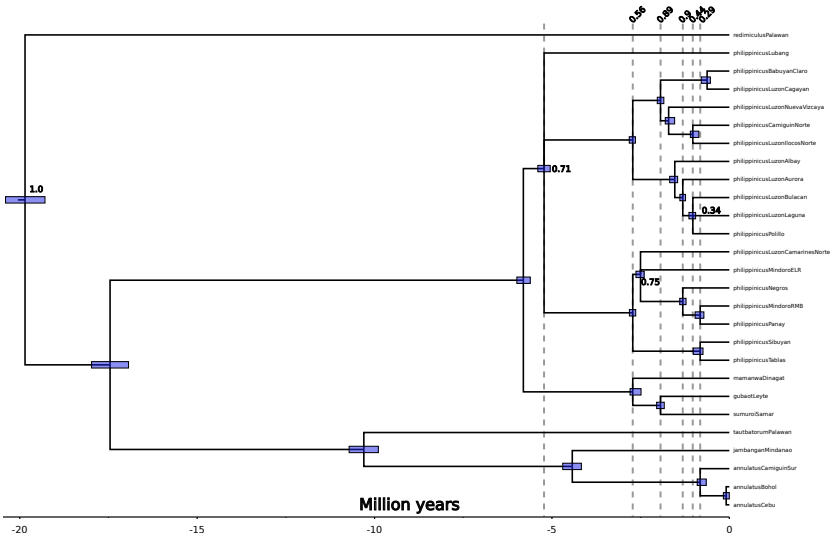
©Rafe M. Brown

- Sampled individuals from 27 and 26 populations across Philippines for *Cyrtodactylus* and *Gekko*, respectively
- Collected short DNA sequences (RADseq) from across genome of each individual
  - *Cyrtodactylus*: 1702 loci & 155,887 sites
  - *Gekko*: 1033 loci & 94,813 sites

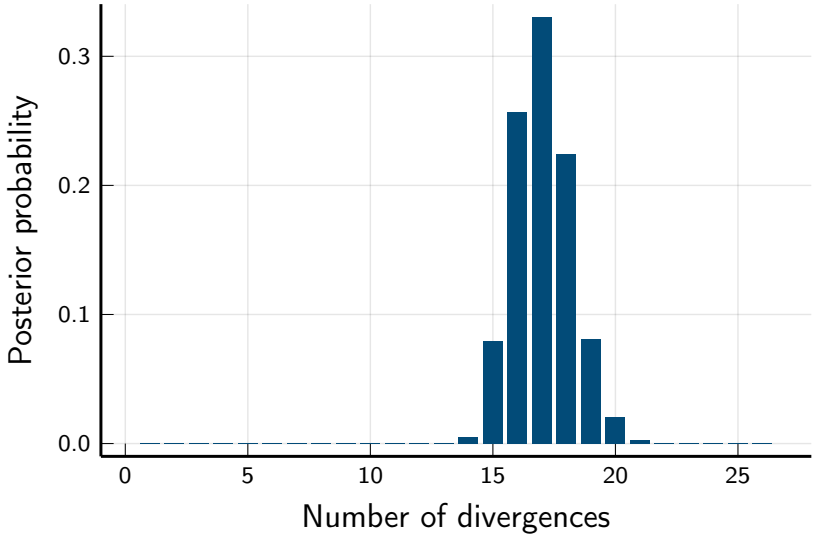
©Rafe M. Brown

J. R. Oaks et al. (2019). *Evolution* 73: 1151–1167

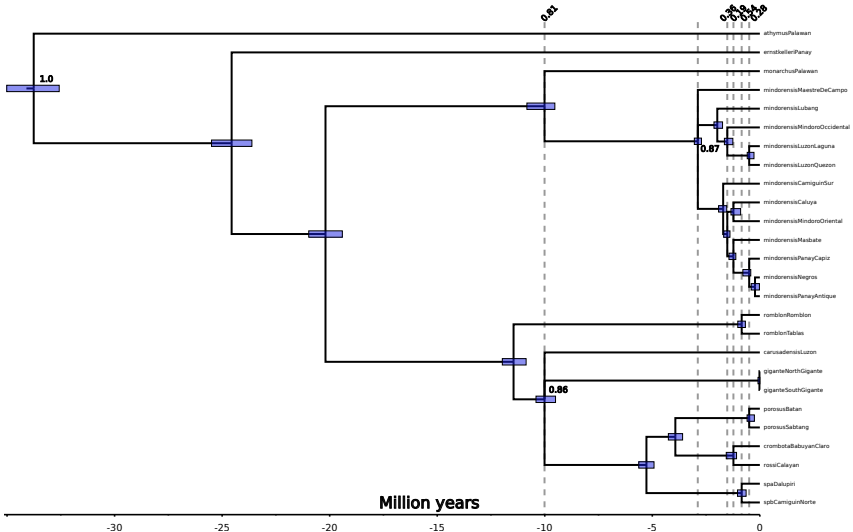
Cyrtodactylus



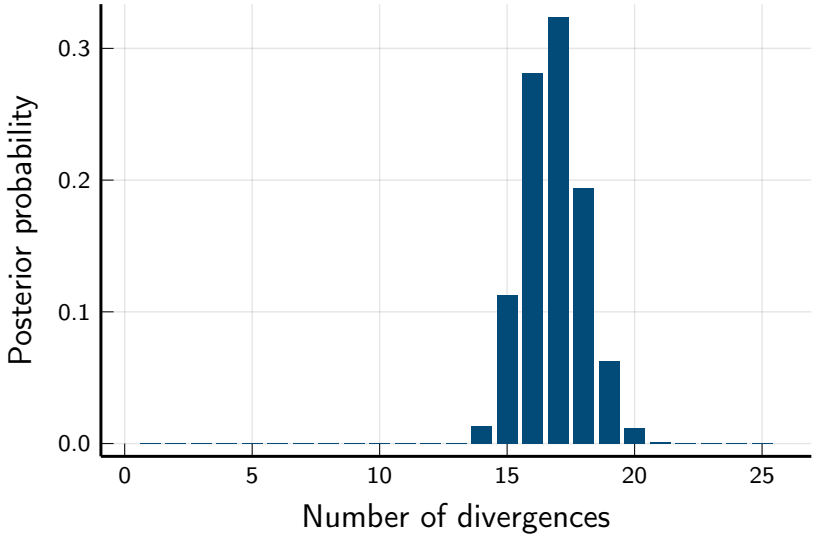
Cyrtodactylus



Gekko


1.0
0.87
0.86
athymusPalawan
ernstkelleriPanay
monarchusPalawan
mindorensisMaestreDeCampo
mindorensisLubang
mindorensisMindoroOccidental
mindorensisLuzonLaguna
mindorensisLuzonQuezon
mindorensisCamiguinSur
mindorensisCaluya
mindorensisMindoroOriental
mindorensisMasbate
mindorensisPanayCapiz
mindorensisNegros
mindorensisPanayAntique
romblonRomblon
romblonTablas
carusadensisLuzon
giganteNorthGigante
giganteSouthGigante
porosusBataan
porosusSabtang
crombotaBabuyanClaro
rossiCalayan
spaDalupiri
spbCamiguinNorte
Million years
-30
-25
-20
-15
-10
-5
0

Gekko


Posterior probability
0.0
0.1
0.2
0.3
0
5
10
15
20
25
Number of divergences

Take-home points

- We can accurately infer phylogenies with shared divergences with moderately sized genomic data sets

Take-home points

- We can accurately infer phylogenies with shared divergences with moderately sized genomic data sets
- Generalizing tree space can avoid spurious support and improve MCMC mixing

Take-home points

- We can accurately infer phylogenies with shared divergences with moderately sized genomic data sets
- Generalizing tree space can avoid spurious support and improve MCMC mixing
- We found support for shared divergences among Philippine gekkonids

So much to do...

- Theory on generalized tree space
- Better algorithms to take advantage of new ways to explore this space
- Port generalized tree prior to RevBayes to couple with other data models, relaxed-clock models, biogeographic models, etc.
- Couple with paleogeographically explicit model of range evolution
- Develop process-based priors

Everything is on GitHub...

Software:

- Phycoeval: https://github.com/phyletica/ecoevolity

Open-Science Notebooks:

- Phycoeval simulations: https://github.com/phyletica/phycoeval-experiments
- Gecko RADseq: https://github.com/phyletica/gekgo

Acknowledgments

- Phyletica Lab (the Phyleticians)
- Mark Holder
- Rafe Brown
- Cam Siler

Computation:

- Alabama Supercomputer Authority
- Auburn University Hopper Cluster

Funding:

DEB 1656004

Photo credits:

- Rafe Brown
- PhyloPic

Thanks to the organizers of MIC-Phy!

Questions?

joaks@auburn.edu

phyletica.org