Generalizing Bayesian phylogenetics to infer shared evolutionary events

Jamie Oaks
Auburn University
phyletica.org
🐦 @jamoaks

phyletica.org/slides/lib.pdf

© 2007 Boris Kulikov boris-kulikov.blogspot.com

Phyletica Lab

The Phyleticians

Postdocs

- Perry Wood, Jr
- *Brian Folt*
- *Jesse Grismer*

Graduate students

- Tashitso Anamza
- Matt Buehler
- Kerry Cobb
- Kyle David
- Randy Klabacka
- Morgan Muell
- Tanner Myers
- Claire Tracy
- *Branna Sipley*
- *Aundrea Westfall*

Undergraduate students

- Laura Lewis
- Mary Wells
- Hailey Whitaker
- Noah Yawn
- *Charlotte Benedict*
- *Eric Carbo*
- *Ryan Cook*
- *Andrew DeSana*
- *Miles Horne*
- *Jacob Landrum*
- *Nadia L'Bahy*
- *Jorge Lopez-Perez*
- *Holden Smith*
- *Virginia White*
- *Kayla Wilson*

The last 5 years

Generalizing Bayesian phylogenetics to infer shared evolutionary events

The next 5 years

- My vision for a position at the LIB

- Phylogenetics is rapidly becoming the statistical foundation of biology

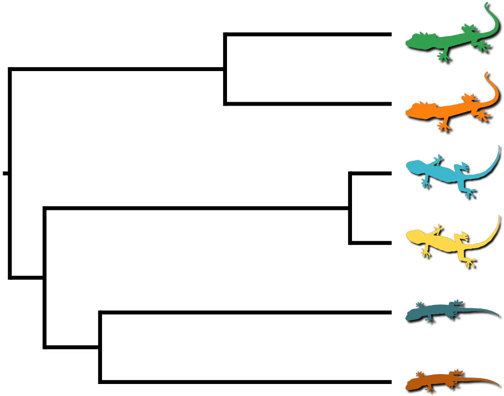
© 2007 Boris Kulikov boris-kulikov.blogspot.com

- Phylogenetics is rapidly becoming the statistical foundation of biology
- “Big data” present exciting possibilities and challenges

© 2007 Boris Kulikov boris-kulikov.blogspot.com

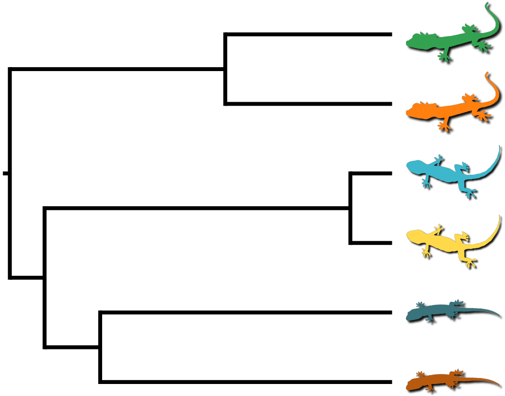
- Phylogenetics is rapidly becoming the statistical foundation of biology
- “Big data” present exciting possibilities and challenges
- Many opportunities to develop new ways to study biology in light of phylogeny

© 2007 Boris Kulikov boris-kulikov.blogspot.com

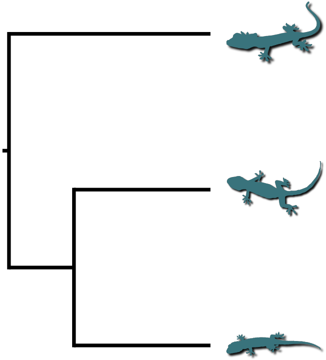


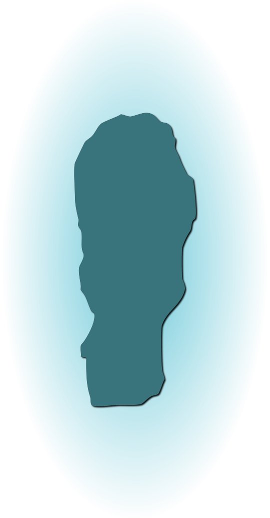
phylopic.org CC0

- **Assumption:** All processes of diversification affect each lineage independently



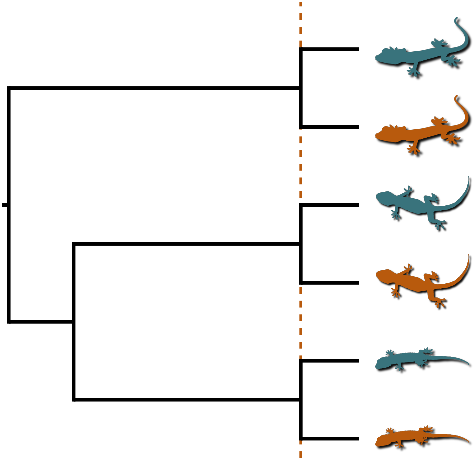
phylopic.org CC0

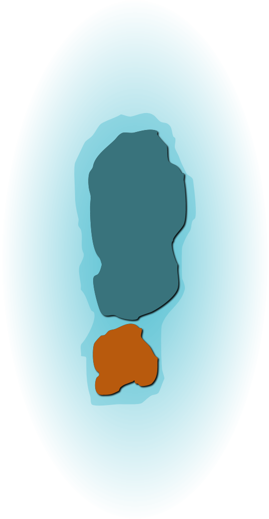




J. R. Oaks (2019). *Systematic Biology* 68: 371–395

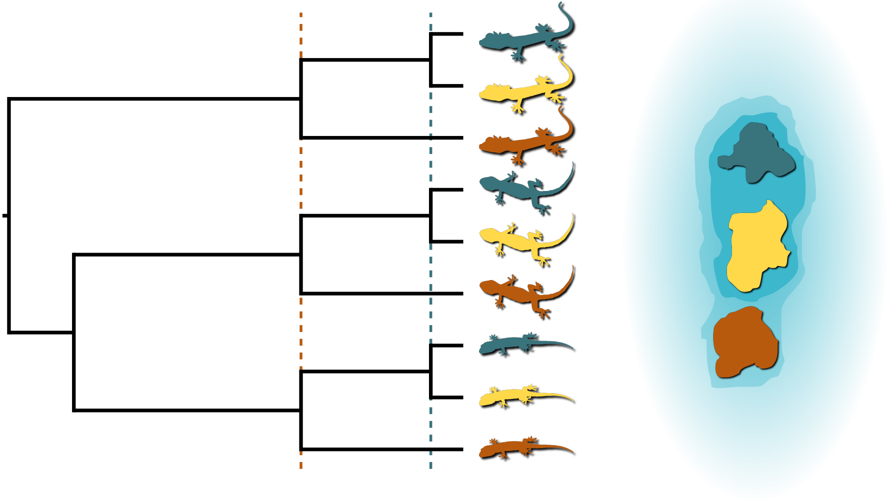
J. R. Oaks, C. D. Siler, and R. M. Brown (2019). *Evolution* 73: 1151–1167





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

J. R. Oaks, C. D. Siler, and R. M. Brown (2019). *Evolution* 73: 1151–1167

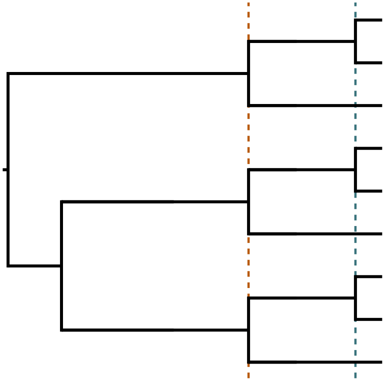


J. R. Oaks (2019). *Systematic Biology* 68: 371–395

J. R. Oaks, C. D. Siler, and R. M. Brown (2019). *Evolution* 73: 1151–1167

Biogeography

- Environmental changes that affect whole communities of species

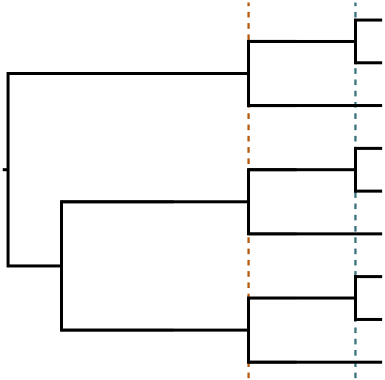


Biogeography

- Environmental changes that affect whole communities of species

Genome evolution

- Duplication of a chromosome segment harboring gene families



Biogeography

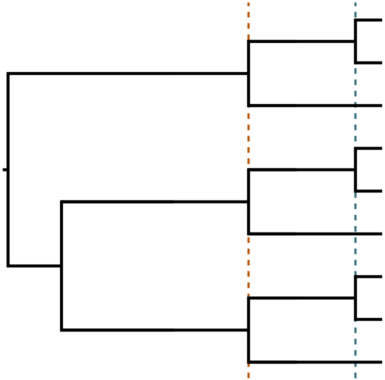
- Environmental changes that affect whole communities of species

Genome evolution

- Duplication of a chromosome segment harboring gene families

Epidemiology

- Transmission at social gatherings



Biogeography

- Environmental changes that affect whole communities of species

Genome evolution

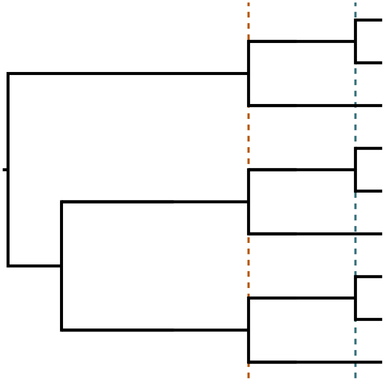
- Duplication of a chromosome segment harboring gene families

Epidemiology

- Transmission at social gatherings

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

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

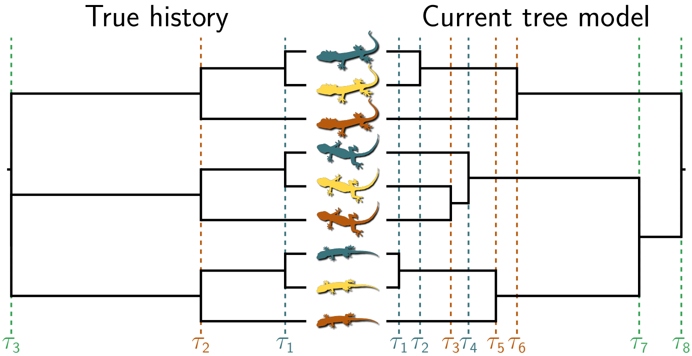


Why account for shared divergences?

J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*

Why account for shared divergences?

1. Improve inference

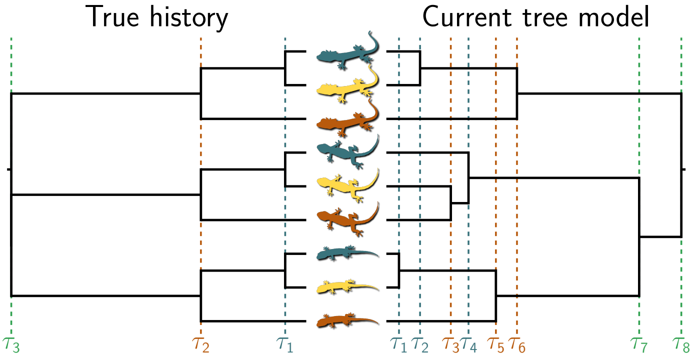


J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*

Why account for shared divergences?

1. Improve inference

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



J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*

Biogeography

- Environmental changes that affect whole communities of species

Genome evolution

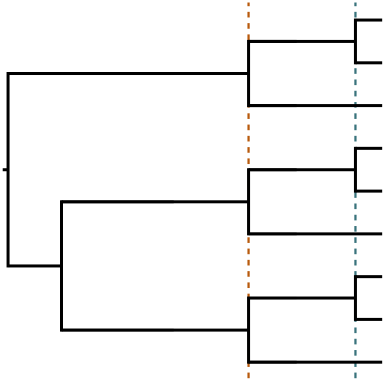
- Duplication of a chromosome segment harboring gene families

Epidemiology

- Transmission at social gatherings

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

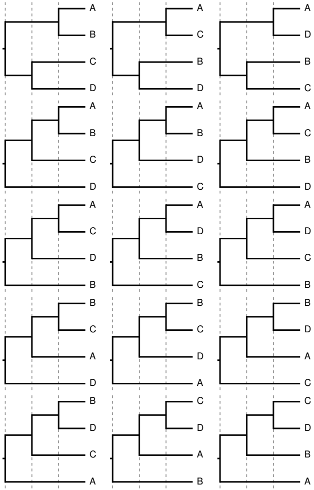
- Speciation of the host
- Co-colonization of new host species



Generalizing tree space

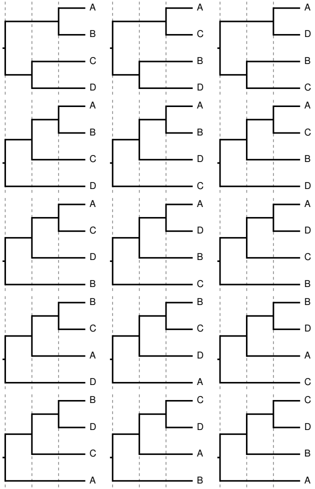
Generalizing tree space

$n_{\mathcal{T}} = 3$

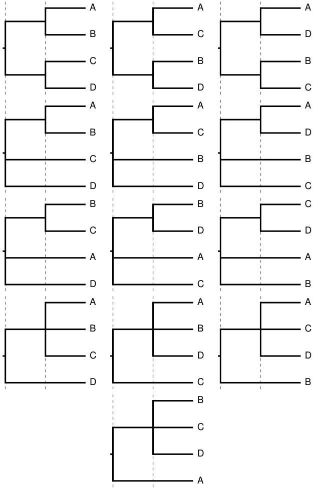


Generalizing tree space

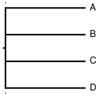
$n_{\mathcal{T}} = 3$



$n_{\mathcal{T}} = 2$

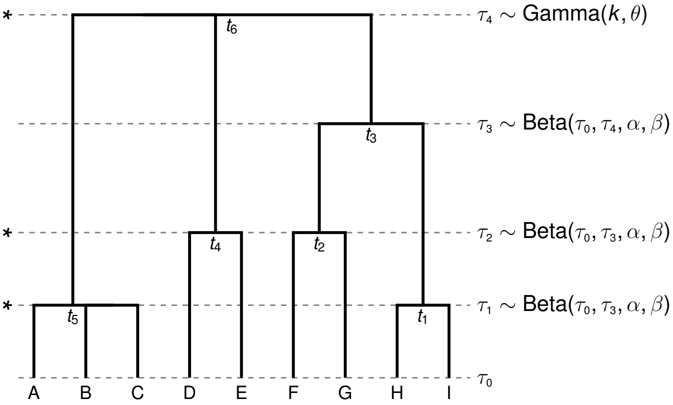


$n_{\mathcal{T}} = 1$



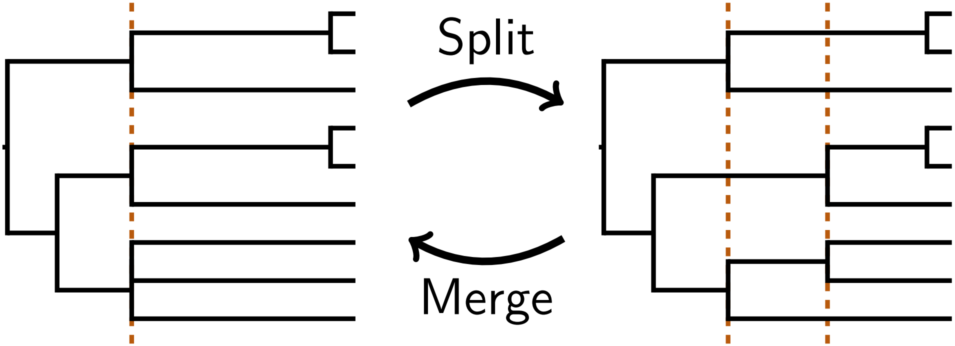
Generalized tree distribution

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


*
t_6
$\tau_4 \sim \text{Gamma}(k, \theta)$
t_3
$\tau_3 \sim \text{Beta}(\tau_0, \tau_4, \alpha, \beta)$
*
t_4
t_2
$\tau_2 \sim \text{Beta}(\tau_0, \tau_3, \alpha, \beta)$
*
t_5
t_1
$\tau_1 \sim \text{Beta}(\tau_0, \tau_3, \alpha, \beta)$
τ_0
A
B
C
D
E
F
G
H
I

J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*

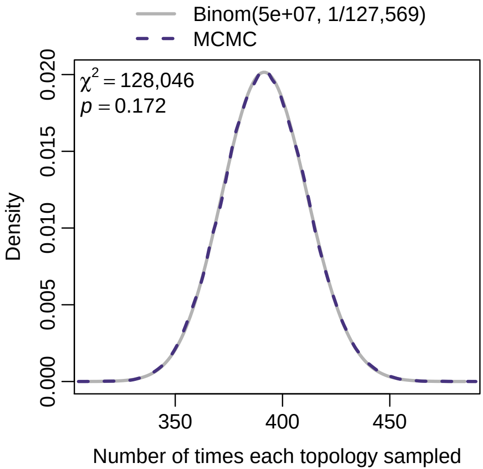
Inferring trees with shared divergences



Reversible-jump MCMC

J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*

Validating rjMCMC with 7-leaf tree



The rjMCMC algorithms sample the expected generalized tree distribution

J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*

PhycoEval

Phylogenetic coevality

J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*

Ecoevolity

Estimating **evo**lutionary **coevality**

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

- **Tree model**
  - rjMCMC sampling of generalized tree distribution

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

PhycoEval

Phylogenetic coevality

J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*

Ecoevolity

Estimating **evo**lutionary **coevality**

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

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

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

PhycoEval

Phylogenetic coevality

J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*

Ecoevolity

Estimating **evo**lutionary **coevality**

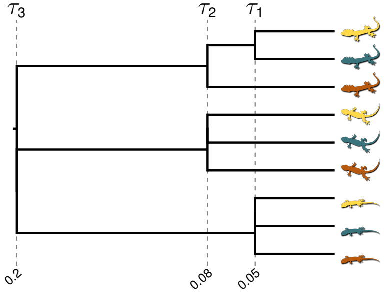
J. R. Oaks (2019). *Systematic Biology* 68: 371–395

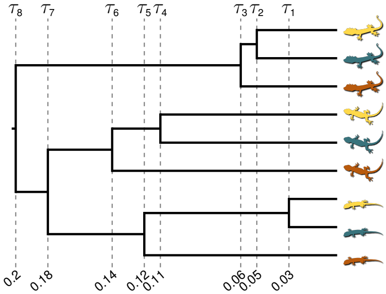
- **Tree model**
  - rjMCMC sampling of generalized tree distribution
- **Likelihood model**
  - CTMC model of characters evolving along genealogies
  - Infer species trees by analytically integrate over genealogies[1]
- *Goal: Co-estimation of phylogeny and shared divergences from genomic data*

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

Methods: Simulations

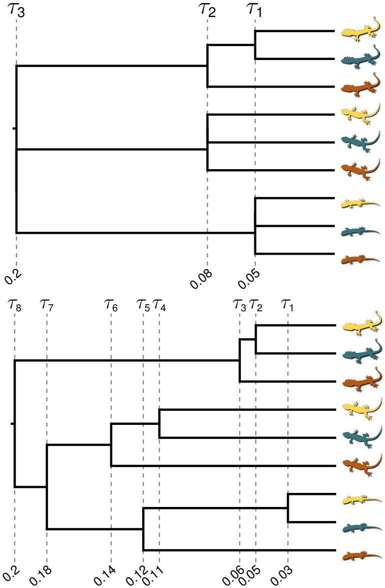
- Simulated 100 data sets with 50,000 characters





Methods: Simulations

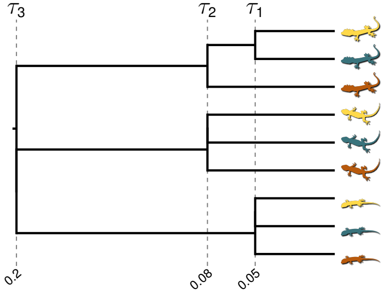
- Simulated 100 data sets with 50,000 characters
- Analyzed each data set with:
  - M_G = Generalized tree model
  - M_{IB} = Independent-bifurcating tree model

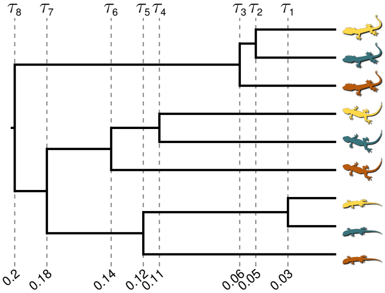


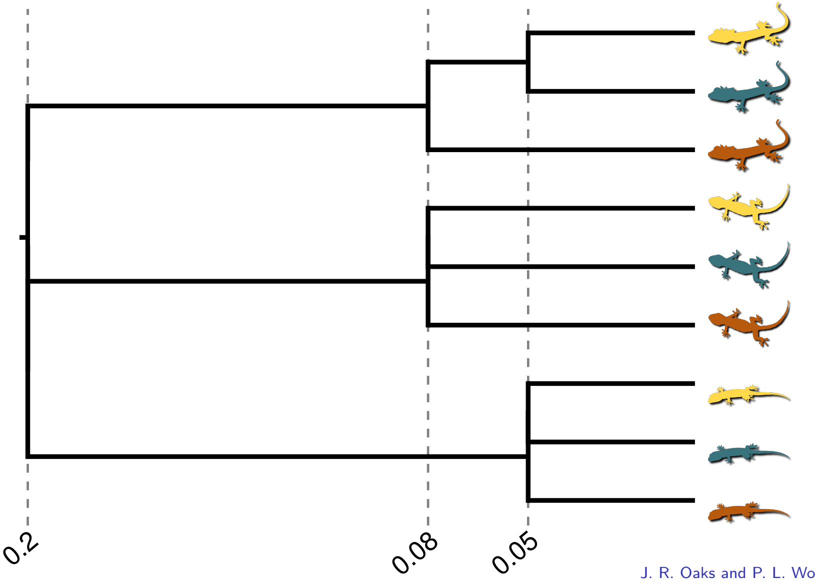
Methods: Simulations

- Simulated 100 data sets with 50,000 characters
- Analyzed each data set with:
  - M_G = Generalized tree model
  - M_{IB} = Independent-bifurcating tree model

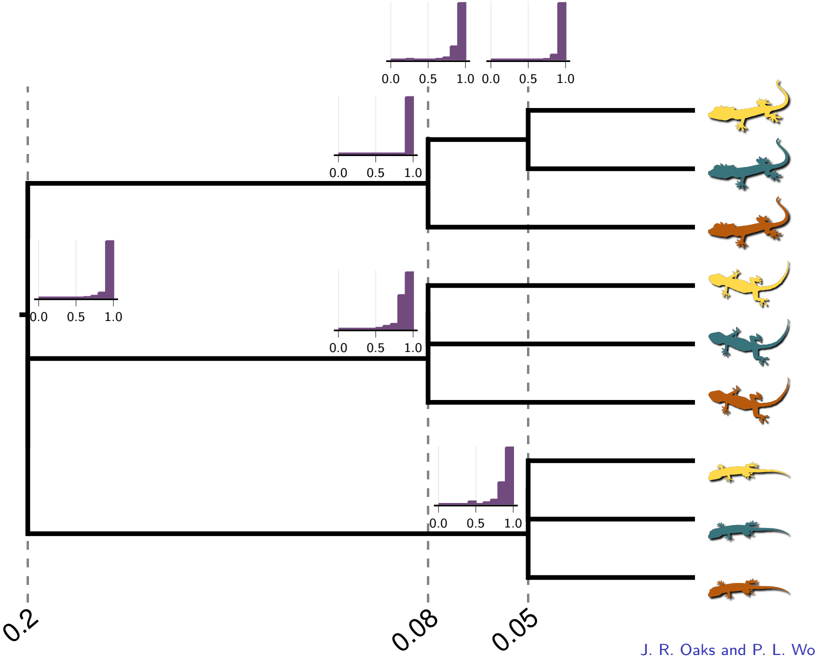
- Simulated 100 data sets where topology and div times randomly drawn from M_G and M_{IB}



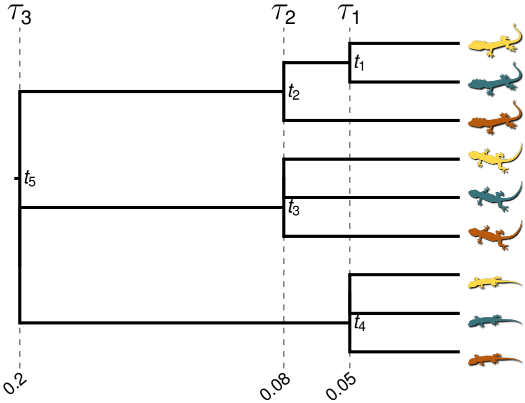




J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*

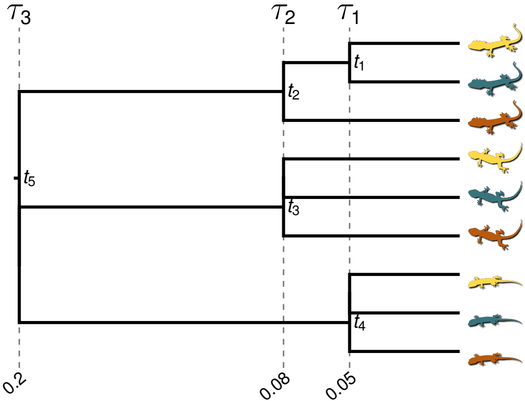


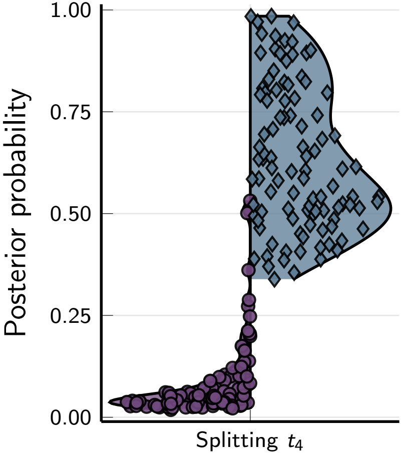
J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*



J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*

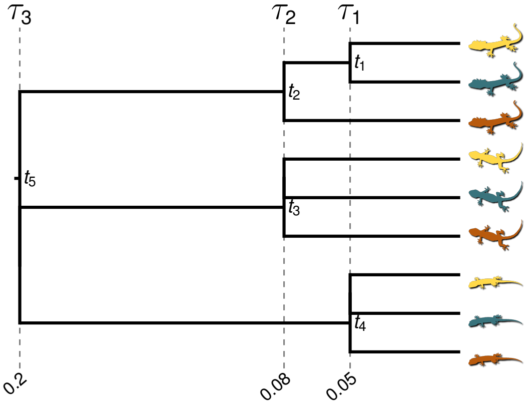
● M_G = Generalized model ◆ M_{IB} = Independent-bifurcating model

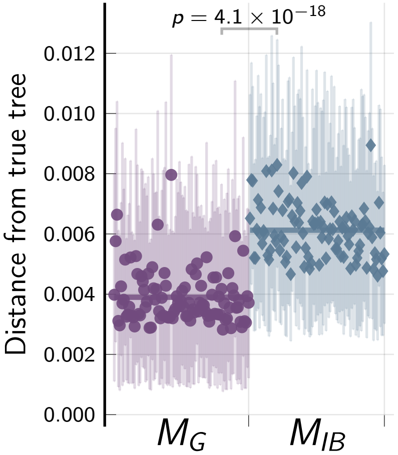




J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*

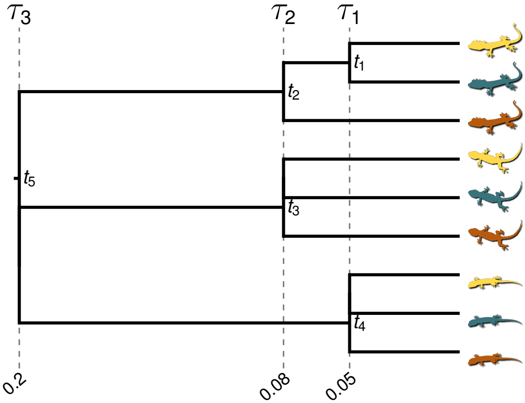
● M_G = Generalized model ◆ M_{IB} = Independent-bifurcating model

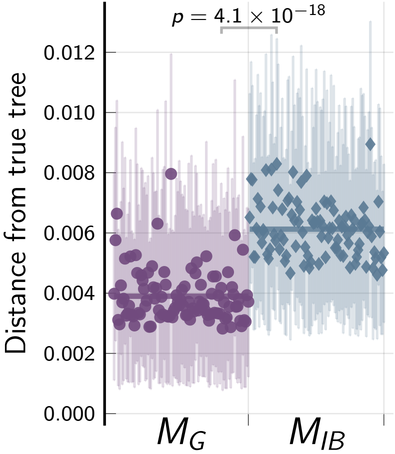




J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*

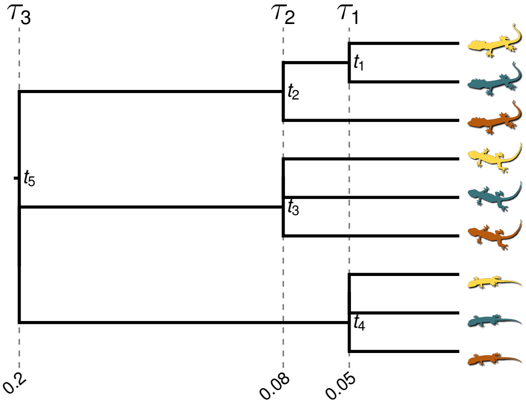
● M_G = Generalized model ◆ M_{IB} = Independent-bifurcating model

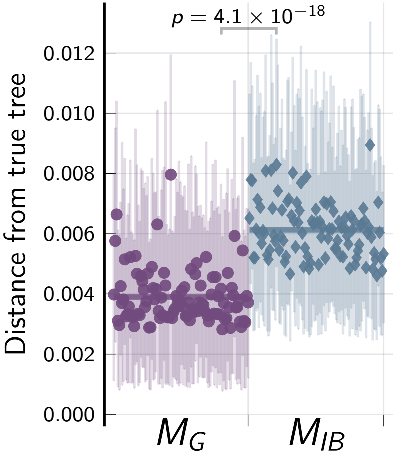




J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*

● M_G = Generalized model ◆ M_{IB} = Independent-bifurcating model

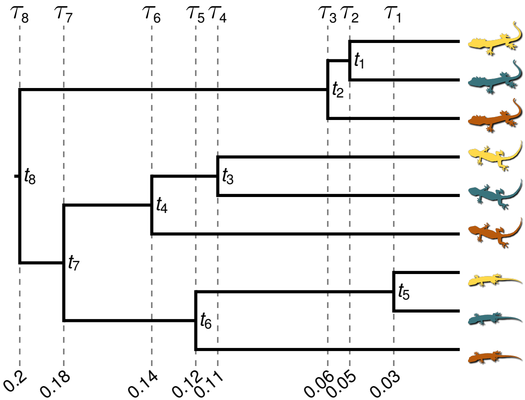




M_G significantly better at inferring trees with shared divergences

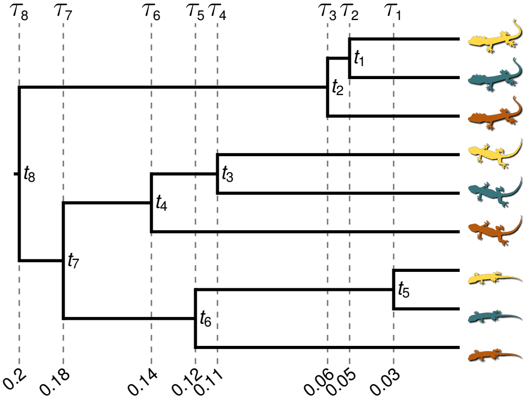
J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*

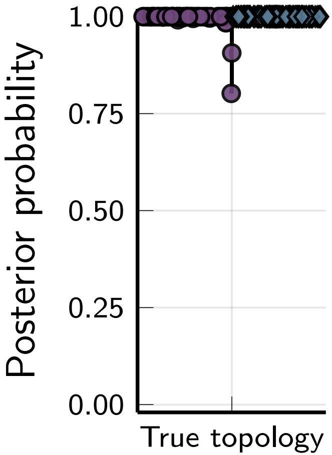
● M_G = Generalized model ◆ M_{IB} = Independent-bifurcating model



J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*

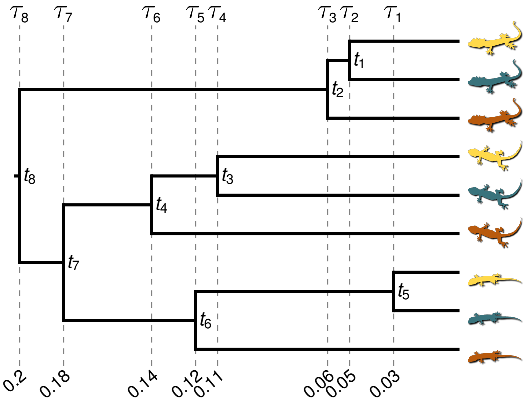
● M_G = Generalized model ◆ M_{IB} = Independent-bifurcating model

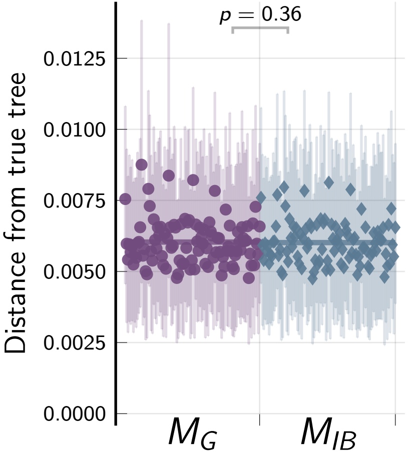




J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*

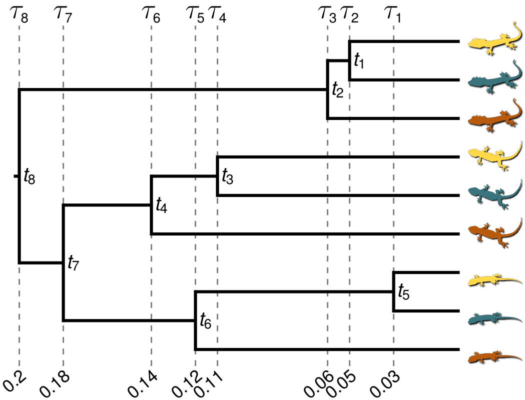
● M_G = Generalized model ◆ M_{IB} = Independent-bifurcating model

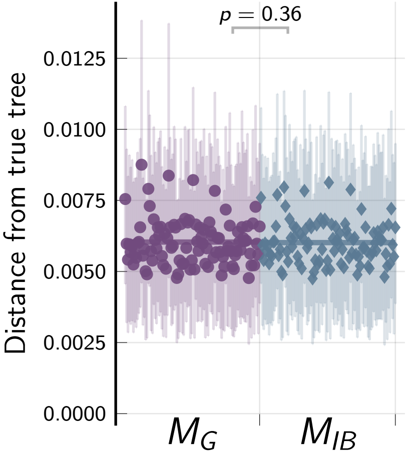




J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*

● M_G = Generalized model ◆ M_{IB} = Independent-bifurcating model

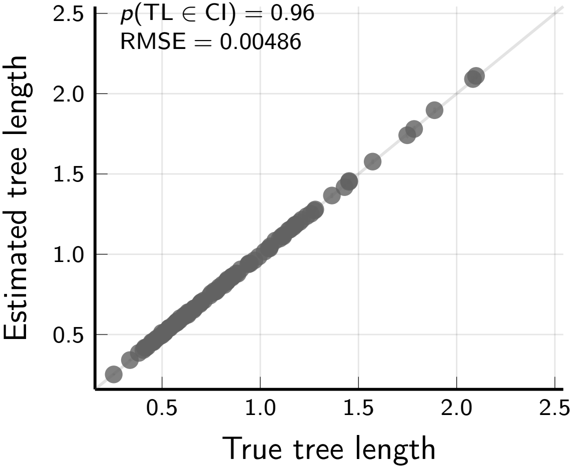




M_G performs as well as true model when divergences are independent

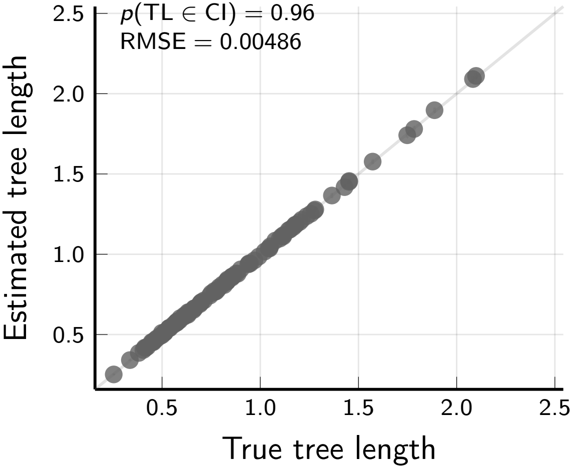
J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*

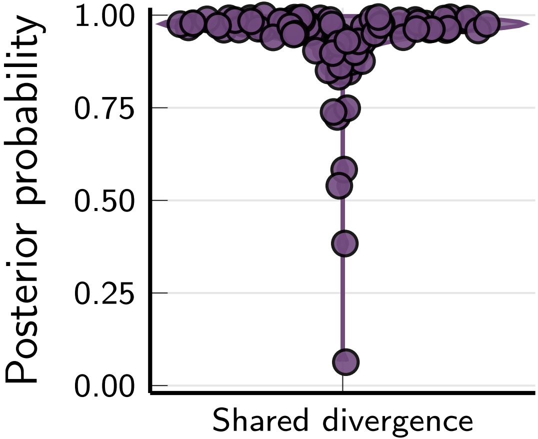
Results: random M_G trees



J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*

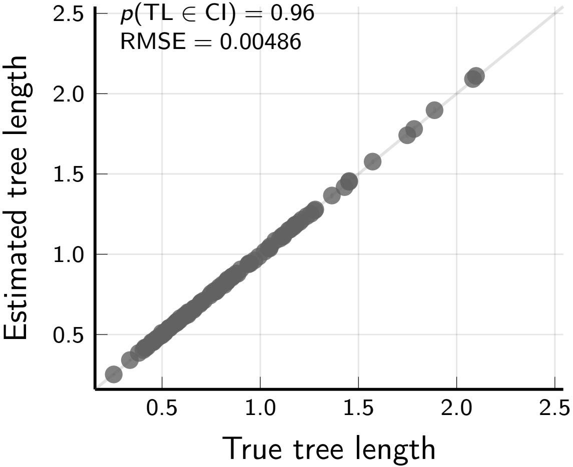
Results: random M_G trees

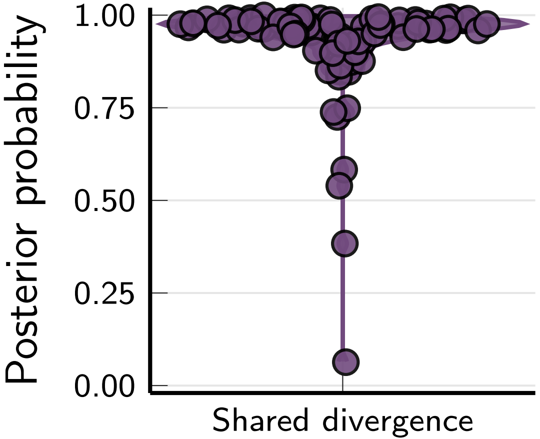




J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*

Results: random M_G trees

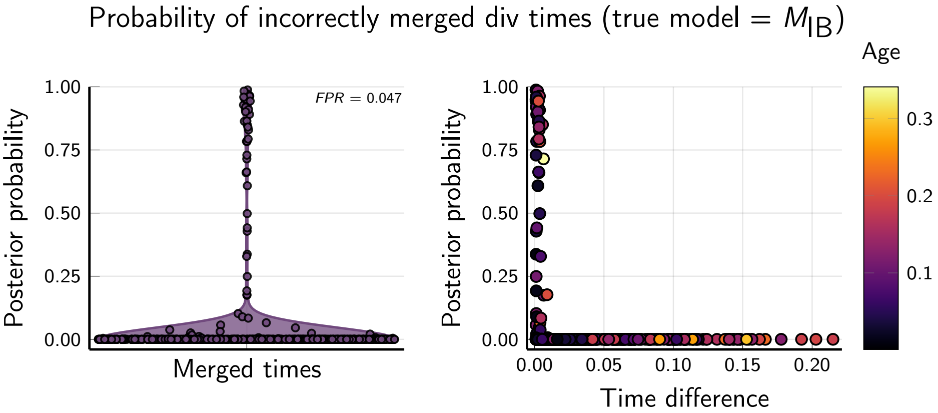




M_G performs well with data simulated on random trees with shared divergences

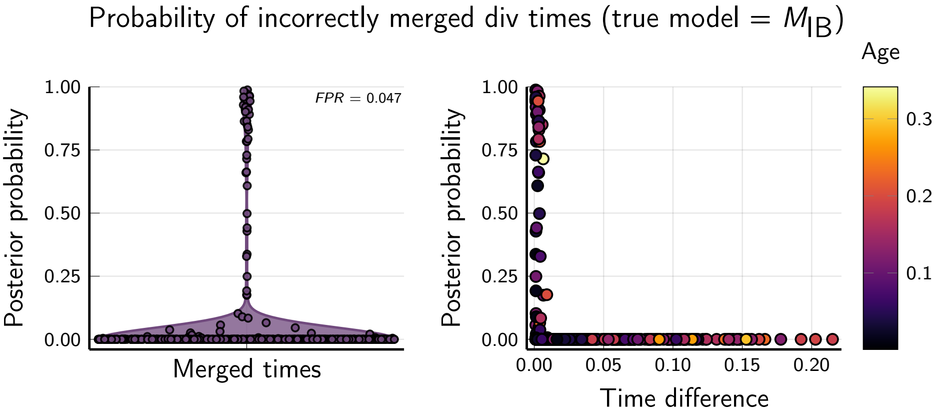
J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*

Results: random M_{IB} trees



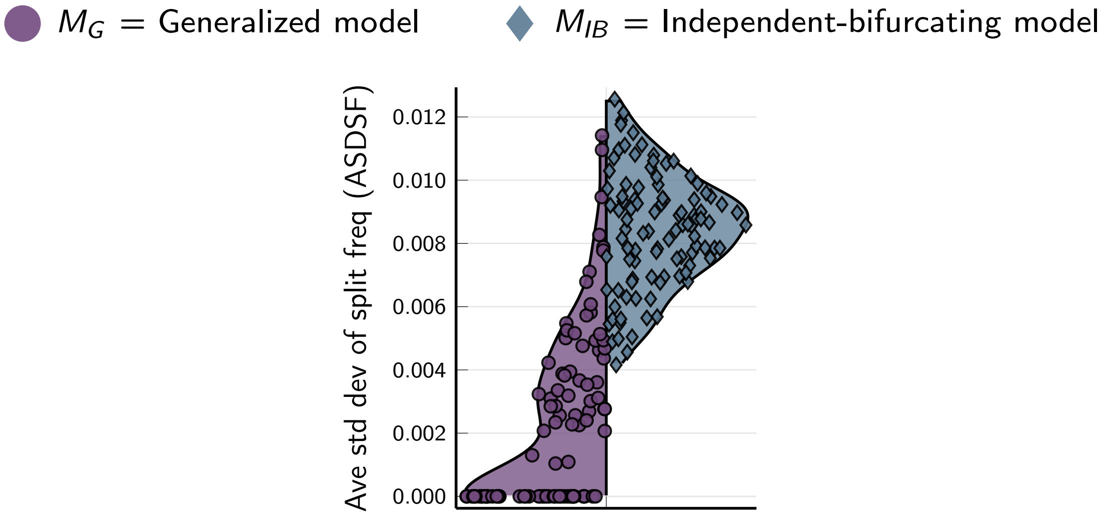
J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*

Results: random M_{IB} trees



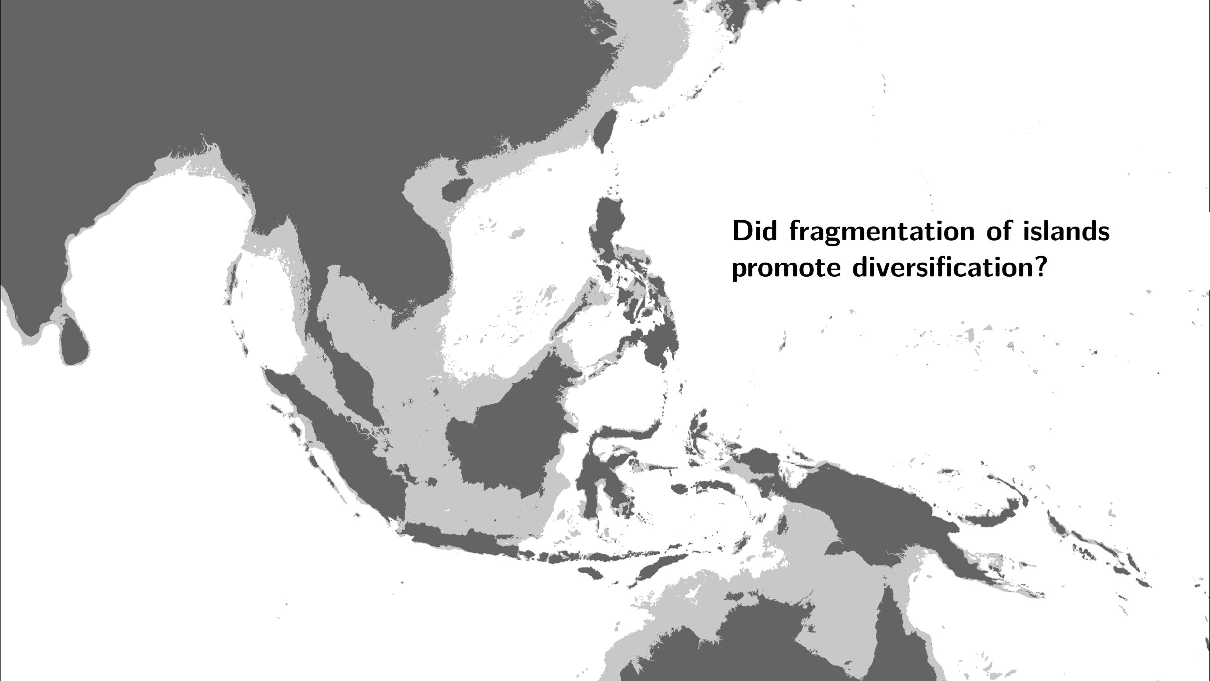
M_G has low false positive rate

J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*



Generalizing tree space improves MCMC convergence and mixing

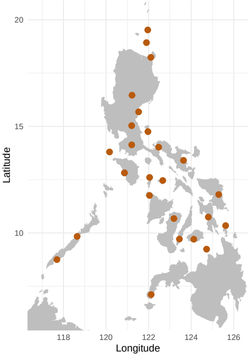
J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*



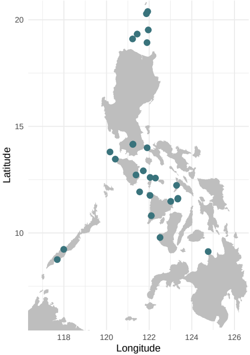
Did fragmentation of islands promote diversification?

Cyrtodactylus

©Rafe M. Brown



Gekko

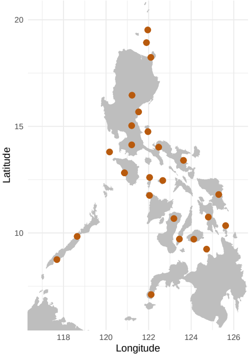


©Rafe M. Brown

J. R. Oaks, C. D. Siler, and R. M. Brown (2019). *Evolution* 73: 1151–1167

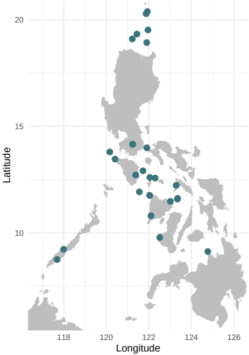
Cyrtodactylus

©Rafe M. Brown



1702 loci
155,887 sites

Gekko

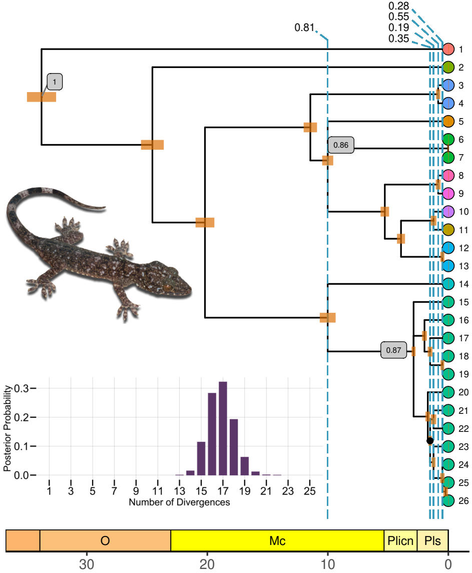


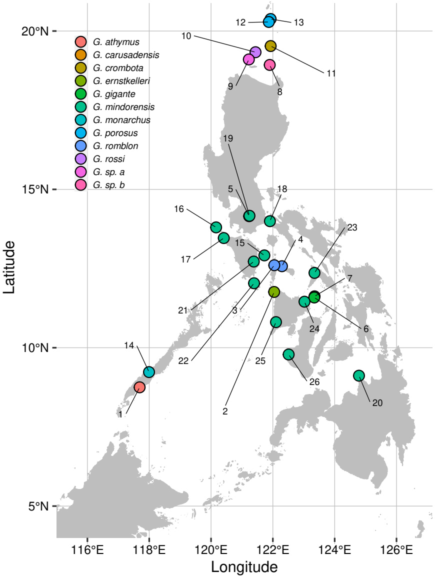
©Rafe M. Brown

1033 loci
94,813 sites

J. R. Oaks, C. D. Siler, and R. M. Brown (2019). *Evolution* 73: 1151–1167

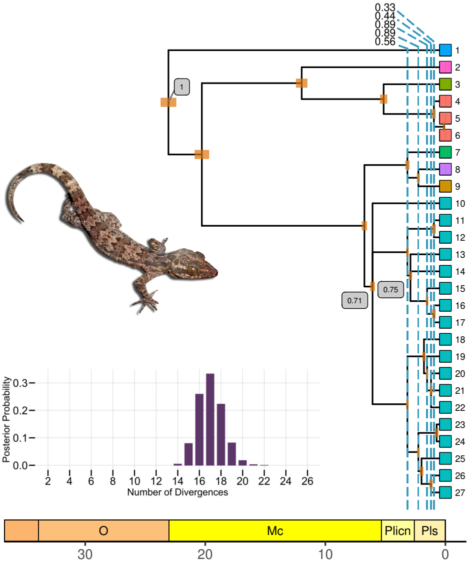
Gekko

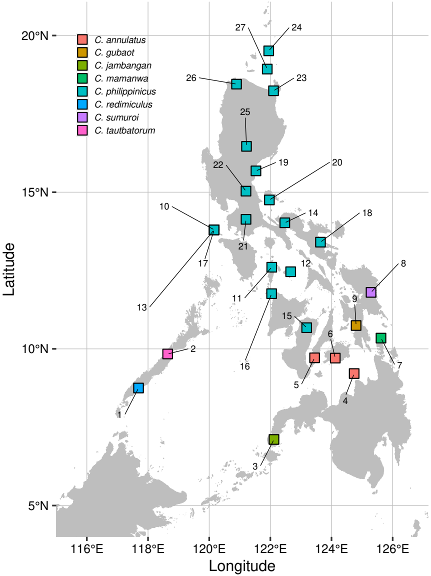




J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*

Cyrtodactylus





J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*

Take-home points

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

Take-home points

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

Take-home points

- We can accurately infer phylogenies with shared divergences with moderately sized data sets
- Generalizing tree space avoids spurious support and improves MCMC mixing
- Among Philippine gekkonids, we found support for shared divergences predicted by sea-level changes

Open science: everything is available. . .

Software:

- Phycoeval:
  https://github.com/phyletica/ecoevolity
  (release coming soon)

Open-Science Notebooks:

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

phyletica.org/codiv-sanger-bake-off

Vision for LIB position

Phylogenetic theory/methods

- Develop process-based and trait-dependent distributions over the space of generalized trees

Empirical work

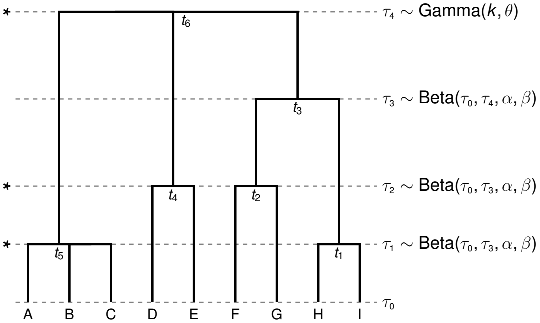
- Did the evolution of habitat preference affect the diversification of bent-toed geckos?
- Epidemiological dynamics of “super-spreading” events during the COVID-19 pandemic

Teaching

- Coding to learn evolution

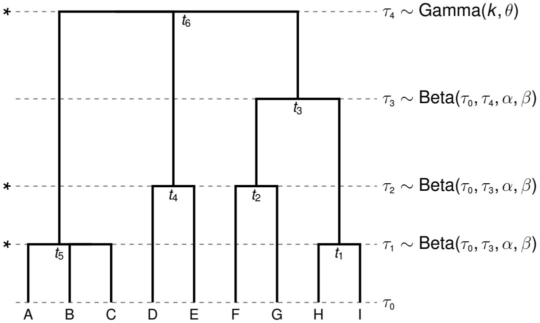
Generalized tree distribution

- Our current distribution over trees is motivated by mathematical convenience


*
*
*
t_6
t_3
t_4
t_2
t_5
t_1
$\tau_4 \sim \text{Gamma}(k, \theta)$
$\tau_3 \sim \text{Beta}(\tau_0, \tau_4, \alpha, \beta)$
$\tau_2 \sim \text{Beta}(\tau_0, \tau_3, \alpha, \beta)$
$\tau_1 \sim \text{Beta}(\tau_0, \tau_3, \alpha, \beta)$
τ_0
A
B
C
D
E
F
G
H
I

Generalized tree distribution

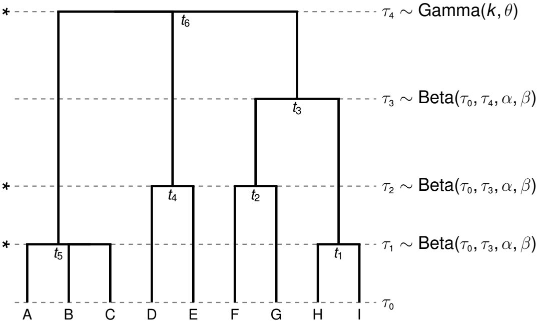
- Our current distribution over trees is motivated by mathematical convenience
- A process-based distribution would allow us to learn about parameters that control diversification processes



*
*
*
t_6
t_3
t_4
t_2
t_5
t_1
$\tau_4 \sim$ Gamma(k, θ)
$\tau_3 \sim$ Beta$(\tau_0, \tau_4, \alpha, \beta)$
$\tau_2 \sim$ Beta$(\tau_0, \tau_3, \alpha, \beta)$
$\tau_1 \sim$ Beta$(\tau_0, \tau_3, \alpha, \beta)$
τ_0
A
B
C
D
E
F
G
H
I

Generalized tree distribution

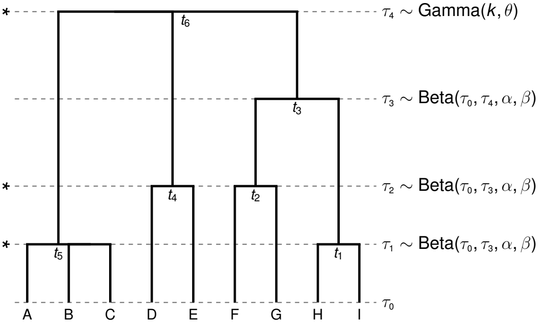
- Our current distribution over trees is motivated by mathematical convenience
- A process-based distribution would allow us to learn about parameters that control diversification processes
- Goal: port M_G algorithms to `RevBayes` and develop generalized birth-death model


*
*
*
t_6
t_3
t_4
t_2
t_5
t_1
A
B
C
D
E
F
G
H
I
$\tau_4 \sim \text{Gamma}(k, \theta)$
$\tau_3 \sim \text{Beta}(\tau_0, \tau_4, \alpha, \beta)$
$\tau_2 \sim \text{Beta}(\tau_0, \tau_3, \alpha, \beta)$
$\tau_1 \sim \text{Beta}(\tau_0, \tau_3, \alpha, \beta)$
τ_0

Generalized tree distribution

- Our current distribution over trees is motivated by mathematical convenience
- A process-based distribution would allow us to learn about parameters that control diversification processes
- Goal: port M_G algorithms to `RevBayes` and develop generalized birth-death model

Sebastian Höhna
LMU Munich

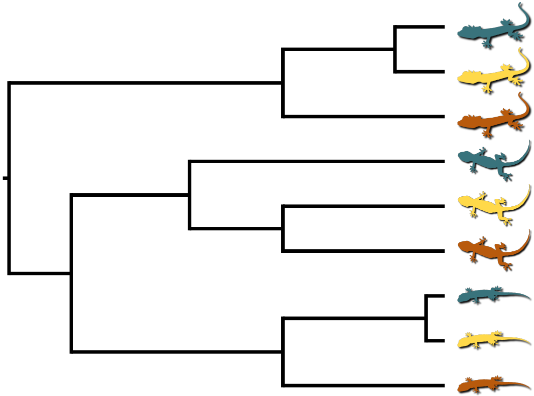

*
t_6
$\tau_4 \sim \text{Gamma}(k, \theta)$
t_3
$\tau_3 \sim \text{Beta}(\tau_0, \tau_4, \alpha, \beta)$
*
t_4
t_2
$\tau_2 \sim \text{Beta}(\tau_0, \tau_3, \alpha, \beta)$
*
t_5
t_1
$\tau_1 \sim \text{Beta}(\tau_0, \tau_3, \alpha, \beta)$
τ_0
A
B
C
D
E
F
G
H
I

Generalizing the birth-death process

Generalizing the birth-death process

Birth-death basics:

- Lineages speciate at rate λ
- Lineages go extinct at rate μ
- We sample extant lineages with probability ρ



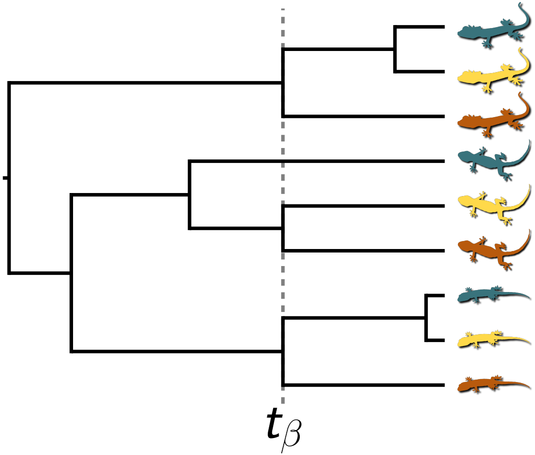
Generalizing the birth-death process

Birth-death basics:

- Lineages speciate at rate λ
- Lineages go extinct at rate μ
- We sample extant lineages with probability ρ

"Birth-death-burst" (BDB) process:

- Include "burst events" that occur at rate λ_β
- Each lineage diverges with probability β



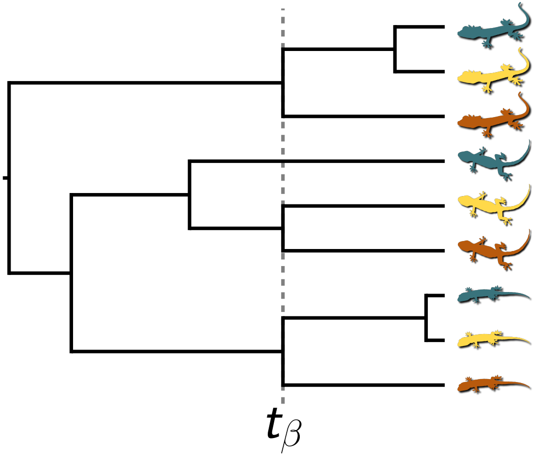
Generalizing the birth-death process

Birth-death basics:

- Lineages speciate at rate λ
- Lineages go extinct at rate μ
- We sample extant lineages with probability ρ

"Birth-death-burst" (BDB) process:

- Include "burst events" that occur at rate λ_β
- Each lineage diverges with probability β
- Allow λ, μ, λ_β, & β to vary depending on the traits of lineages across the tree



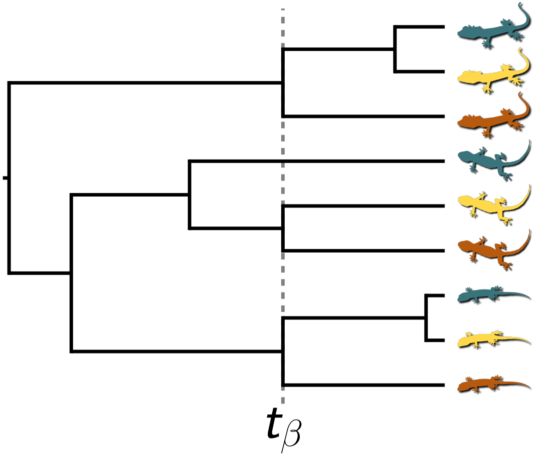
Generalizing the birth-death process

Birth-death basics:

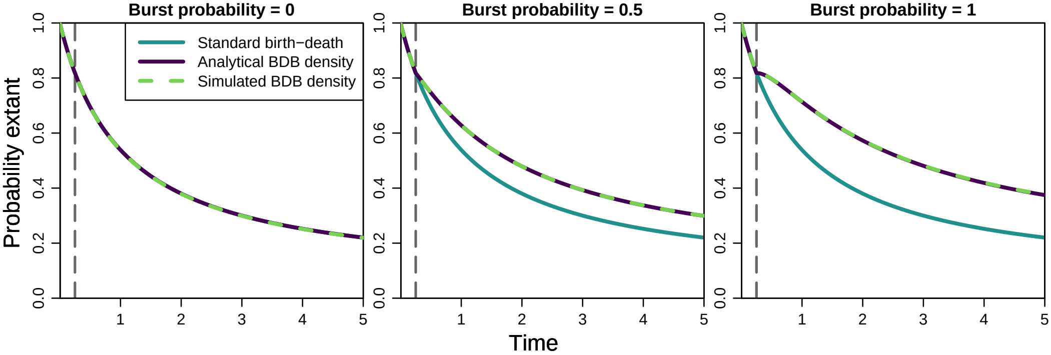
- Lineages speciate at rate λ
- Lineages go extinct at rate μ
- We sample extant lineages with probability ρ

"Birth-death-burst" (BDB) process:

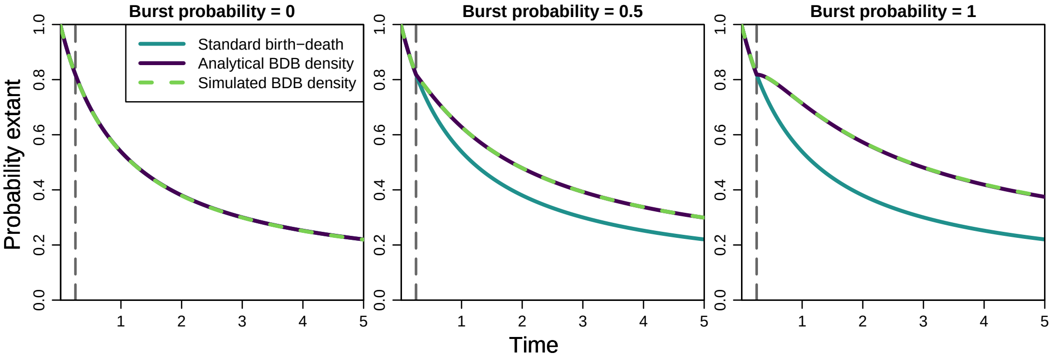
- Include "burst events" that occur at rate λ_β
- Each lineage diverges with probability β
- Allow λ, μ, λ_β, & β to vary depending on the traits of lineages across the tree
- Bayesian model-averaging to infer set of trait-dependent BDB models that best explain data



Birth-death-burst validation



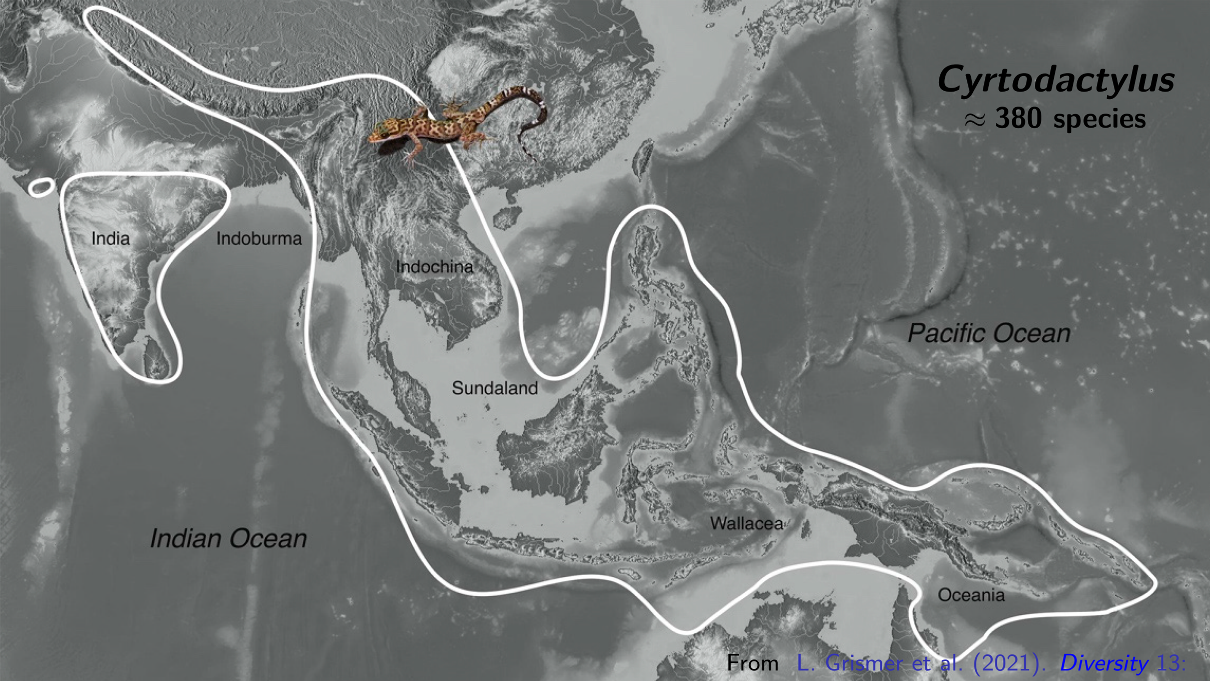
Birth-death-burst validation



We have correctly derived the likelihood of trees under the BDB model

Cyrtodactylus

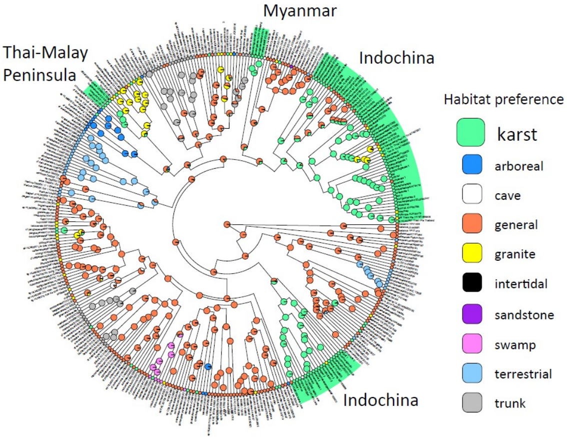
$\approx$ 380 species



From L. Grismer et al. (2021). *Diversity* 13:

Karst endemism in *Cyrtodactylus*

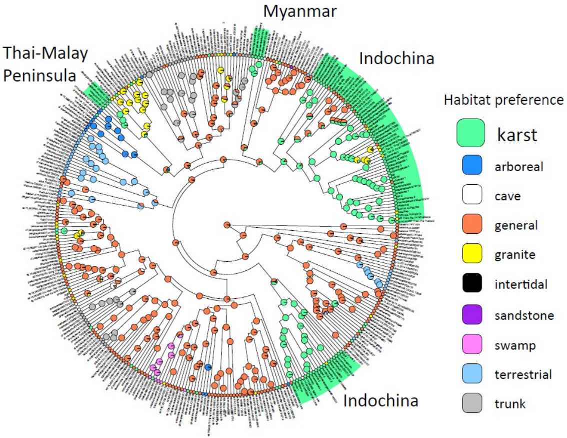
- *Cyrtodactylus* are ecologically diverse, ranging from generalists to microhabitat specialists



L. Grismer et al. (2021). *Diversity* 13:

Karst endemism in *Cyrtodactylus*

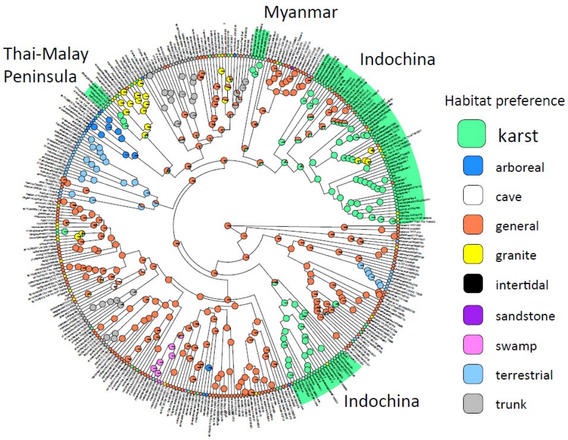
- *Cyrtodactylus* are ecologically diverse, ranging from generalists to microhabitat specialists
- Karst-specificity evolved 24 times



L. Grismer et al. (2021). *Diversity* 13:

Karst endemism in *Cyrtodactylus*

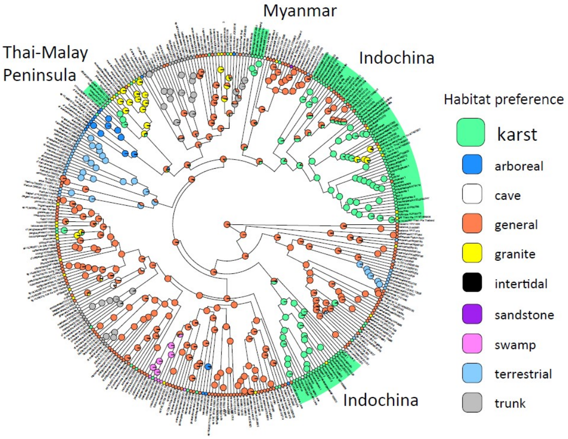
- *Cyrtodactylus* are ecologically diverse, ranging from generalists to microhabitat specialists
- Karst-specificity evolved 24 times
- Comprise 25% of species despite tiny fraction of landscape being karst



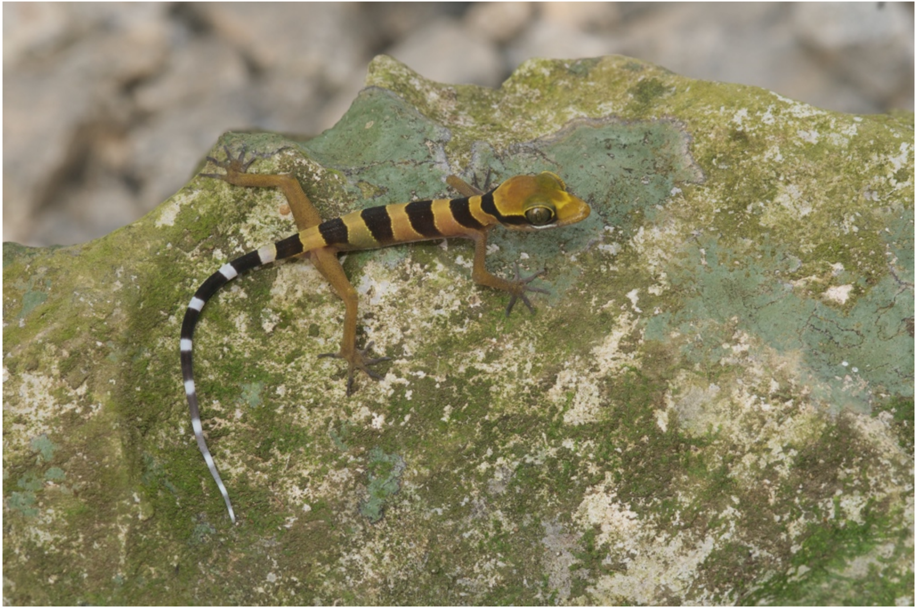
L. Grismer et al. (2021). *Diversity* 13:

Karst endemism in *Cyrtodactylus*

- *Cyrtodactylus* are ecologically diverse, ranging from generalists to microhabitat specialists
- Karst-specificity evolved 24 times
- Comprise 25% of species despite tiny fraction of landscape being karst
- Karst-specific species show remarkable levels of micro-endemism



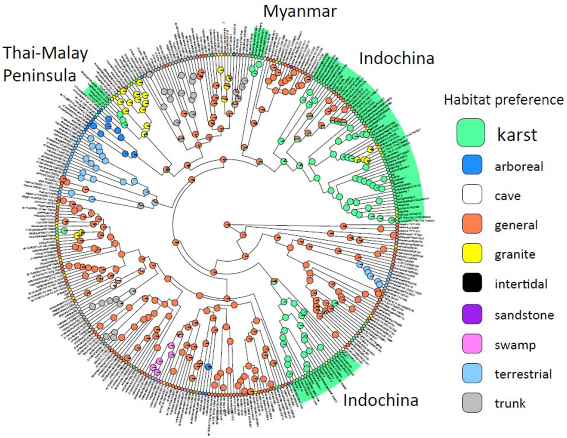
L. Grismer et al. (2021). *Diversity* 13:



©Perry Wood, Jr.

Why high levels of diversity and endemism on karst?

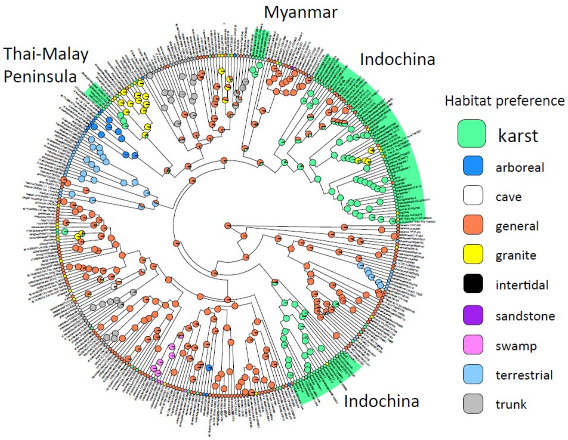
- "Rapid" fragmentation of karst habitat caused by the uplift and subsequent erosion of limestone sediment over the last 30my



L. Grismer et al. (2021). *Diversity* 13:

Why high levels of diversity and endemism on karst?

- “Rapid” fragmentation of karst habitat caused by the uplift and subsequent erosion of limestone sediment over the last 30my
- *E.g.*, Major river systems carved through and isolated limestone karst formations (Ayeyarwady, Chiang Mai, Mekong, Red, and Salween)

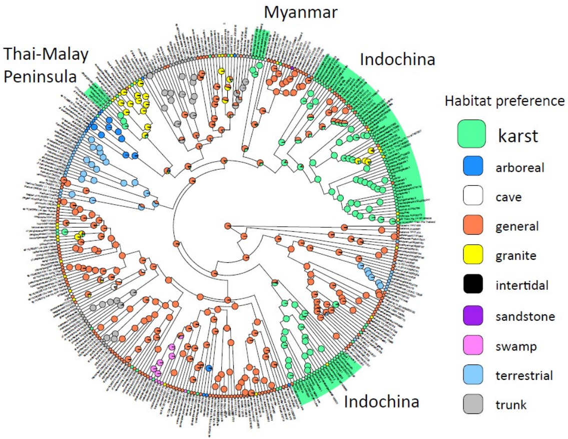


L. Grismer et al. (2021). *Diversity* 13:

Why high levels of diversity and endemism on karst?

- “Rapid” fragmentation of karst habitat caused by the uplift and subsequent erosion of limestone sediment over the last 30my
- *E.g.*, Major river systems carved through and isolated limestone karst formations (Ayeyarwady, Chiang Mai, Mekong, Red, and Salween)

- **Hypothesis**: The fragmentation of limestone karst habitat drove diversification of karst-specific lineages of *Cyrtodactylus*

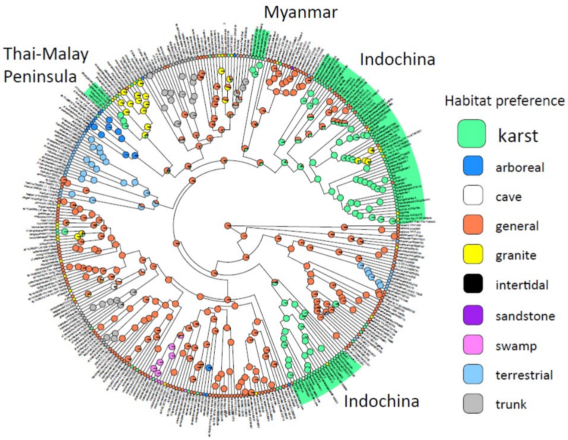


L. Grismer et al. (2021). *Diversity* 13:

Why high levels of diversity and endemism on karst?

- "Rapid" fragmentation of karst habitat caused by the uplift and subsequent erosion of limestone sediment over the last 30my
- *E.g.*, Major river systems carved through and isolated limestone karst formations (Ayeyarwady, Chiang Mai, Mekong, Red, and Salween)

- **Hypothesis**: The fragmentation of limestone karst habitat drove diversification of karst-specific lineages of *Cyrtodactylus*
- **Prediction**: Increased rate of ***shared divergences*** in karst-adapted lineages

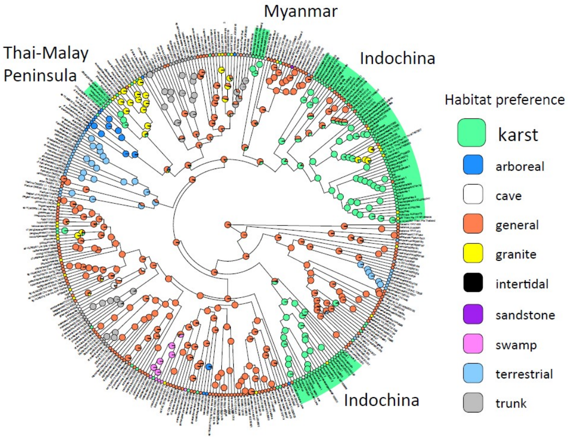


L. Grismer et al. (2021). *Diversity* 13:

Why high levels of diversity and endemism on karst?

Plan:

- Access to tissue samples of 368 of the 380 *Cyrtodactylus* species

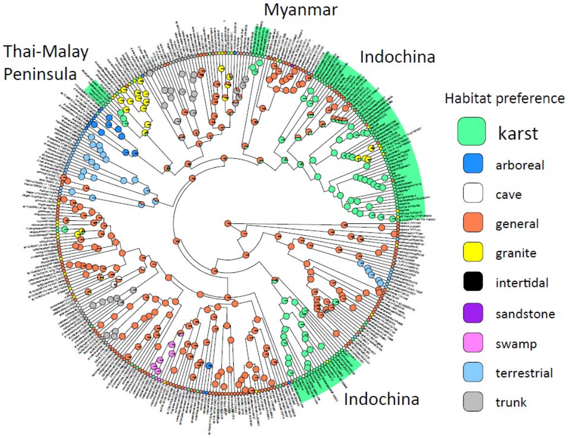


L. Grismer et al. (2021). *Diversity* 13:

Why high levels of diversity and endemism on karst?

Plan:

- Access to tissue samples of 368 of the 380 *Cyrtodactylus* species
- Sequence 5,060 UCE loci

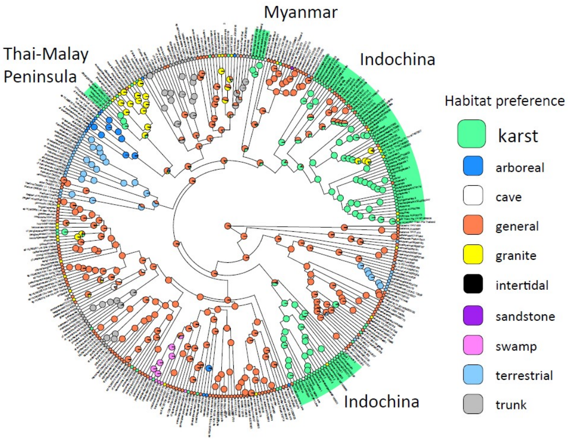


L. Grismer et al. (2021). *Diversity* 13:

Why high levels of diversity and endemism on karst?

Plan:

- Access to tissue samples of 368 of the 380 *Cyrtodactylus* species
- Sequence 5,060 UCE loci
- Apply habitat-dependent BDB model: Model averaging to infer the posterior set of habitat-dependent models

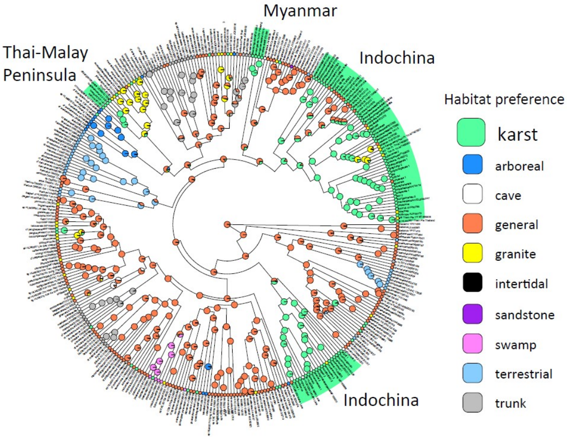


L. Grismer et al. (2021). *Diversity* 13:

Why high levels of diversity and endemism on karst?

Plan:

- Access to tissue samples of 368 of the 380 *Cyrtodactylus* species
- Sequence 5,060 UCE loci
- Apply habitat-dependent BDB model: Model averaging to infer the posterior set of habitat-dependent models
- Approximate posterior probability that karst-specific lineages have higher rate of shared divergences (λ_{β})

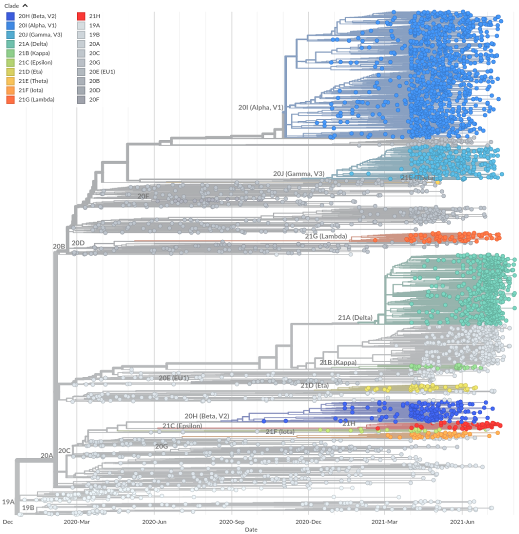


L. Grismer et al. (2021). *Diversity* 13:

Epidemiological dynamics of COVID-19 pandemic

Questions:

- What is the relative contribution of social gatherings to the spread of SARS-CoV-2?

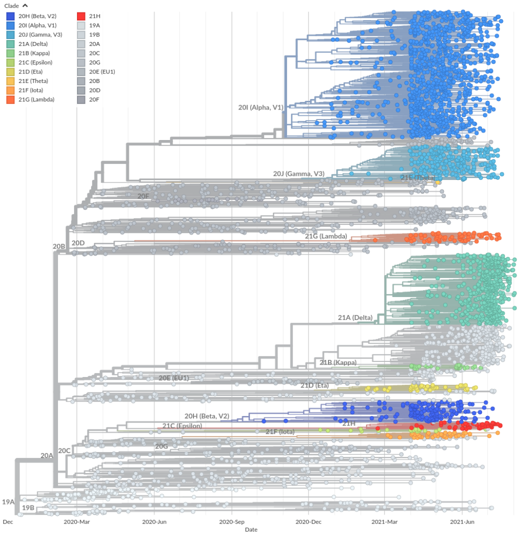


nextstrain.org J. Hadfield et al. (2018). *Bioinformatics* 34: 4121–4123

Epidemiological dynamics of COVID-19 pandemic

Questions:

- What is the relative contribution of social gatherings to the spread of SARS-CoV-2?
- Does this vary among variants of the virus?

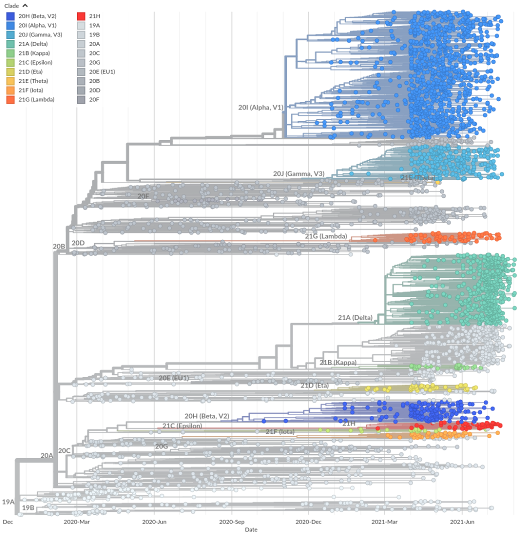


nextstrain.org J. Hadfield et al. (2018). *Bioinformatics* 34: 4121–4123

Epidemiological dynamics of COVID-19 pandemic

Questions:

- What is the relative contribution of social gatherings to the spread of SARS-CoV-2?
- Does this vary among variants of the virus?
- Does this increase during holidays?

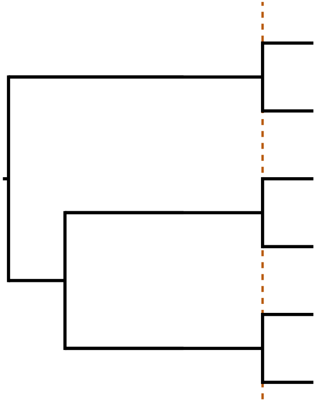


nextstrain.org J. Hadfield et al. (2018). *Bioinformatics* 34: 4121–4123

Divergence patterns predicted by gatherings

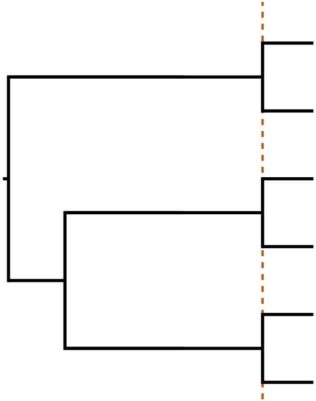
Divergence patterns predicted by gatherings

- Multiple infected people spreading SARS-CoV-2 at a gathering will create shared divergences across "transmission tree"



Divergence patterns predicted by gatherings

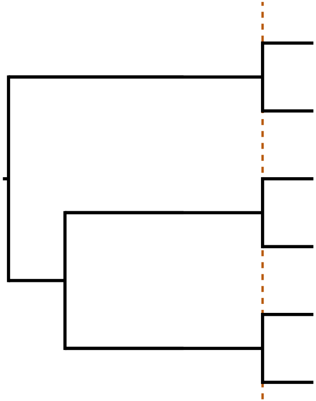
- Multiple infected people spreading SARS-CoV-2 at a gathering will create shared divergences across "transmission tree"
- ***Shared divergences*** are a good proxy for spread at gatherings



Divergence patterns predicted by gatherings

- Multiple infected people spreading SARS-CoV-2 at a gathering will create shared divergences across "transmission tree"
- ***Shared divergences*** are a good proxy for spread at gatherings

Plan:

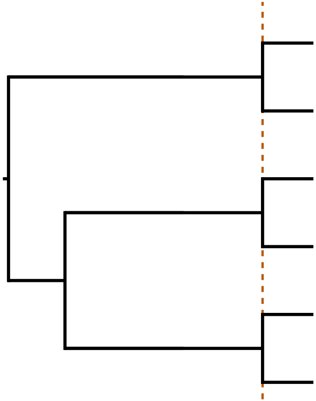


Divergence patterns predicted by gatherings

- Multiple infected people spreading SARS-CoV-2 at a gathering will create shared divergences across “transmission tree”
- ***Shared divergences*** are a good proxy for spread at gatherings

Plan:

- Apply strain-dependent BDB model to regional SARS-CoV-2 sequence datasets

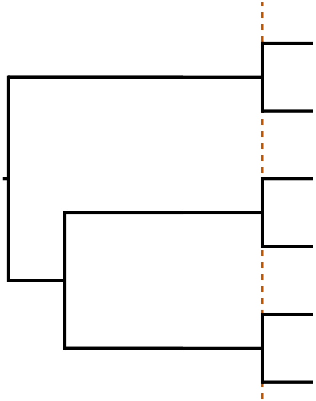


Divergence patterns predicted by gatherings

- Multiple infected people spreading SARS-CoV-2 at a gathering will create shared divergences across “transmission tree”
- ***Shared divergences*** are a good proxy for spread at gatherings

Plan:

- Apply strain-dependent BDB model to regional SARS-CoV-2 sequence datasets
- Estimate relative rate of shared divergences (λ_β/λ)

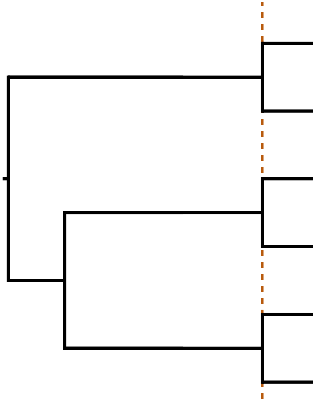


Divergence patterns predicted by gatherings

- Multiple infected people spreading SARS-CoV-2 at a gathering will create shared divergences across "transmission tree"
- ***Shared divergences*** are a good proxy for spread at gatherings

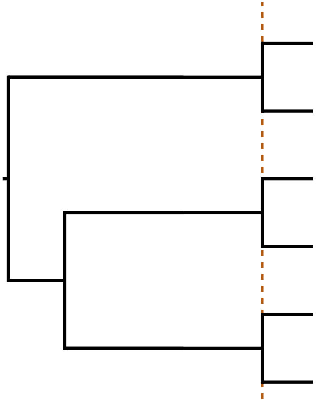
Plan:

- Apply strain-dependent BDB model to regional SARS-CoV-2 sequence datasets
- Estimate relative rate of shared divergences (λ_β/λ)
- Approximate posterior probability that λ_β varies among variants



Divergence patterns predicted by gatherings

- Multiple infected people spreading SARS-CoV-2 at a gathering will create shared divergences across “transmission tree”
- ***Shared divergences*** are a good proxy for spread at gatherings

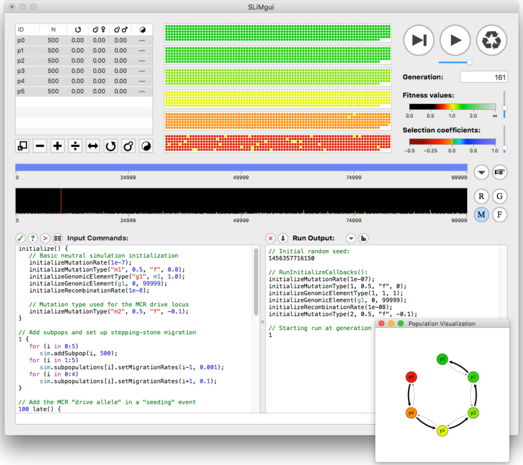


Plan:

- Apply strain-dependent BDB model to regional SARS-CoV-2 sequence datasets
- Estimate relative rate of shared divergences (λ_β/λ)
- Approximate posterior probability that λ_β varies among variants
- Summarize λ_β over time to quantify the effect of holidays

Teaching: Coding to learn evolution

- Develop coding-to-learn evolution course



ID	N	↺	♂♀	♂♂	◕
p0	500	0.00	0.00	0.00	—
p1	500	0.00	0.00	0.00	—
p2	500	0.00	0.00	0.00	—
p3	500	0.00	0.00	0.00	—
p4	500	0.00	0.00	0.00	—
p5	500	0.00	0.00	0.00	—

Generation: 161

Input Commands:

```
initialize() {
   // Basic neutral simulation initialization
   initializeMutationRate(1e-7);
   initializeMutationType("m1", 0.5, "f", 0.0);
   initializeGenomicElementType("g1", m1, 1.0);
   initializeGenomicElement(g1, 0, 99999);
   initializeRecombinationRate(1e-8);

   // Mutation type used for the MCR drive locus
   initializeMutationType("m2", 0.5, "f", -0.1);
}

// Add subpops and set up stepping-stone migration
1 {
   for (i in 0:5)
      sim.addSubpop(i, 500);
   for (i in 1:5)
      sim.subpopulations[i].setMigrationRates(i-1, 0.001);
   for (i in 0:4)
      sim.subpopulations[i].setMigrationRates(i+1, 0.1);
}

// Add the MCR "drive allele" in a "seeding" event
100 late() {
```

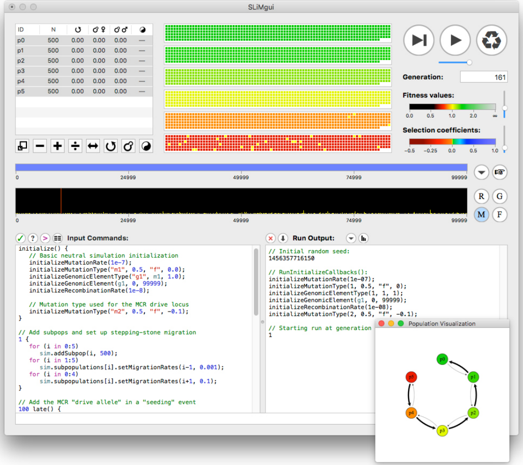
Run Output:

```
// Initial random seed:
1456357716150

// RunInitializeCallbacks():
initializeMutationRate(1e-07);
initializeMutationType(1, 0.5, "f", 0);
initializeGenomicElementType(1, 1, 1);
initializeGenomicElement(g1, 0, 99999);
initializeRecombinationRate(1e-08);
initializeMutationType(2, 0.5, "f", -0.1);

// Starting run at generation
1
```

© 2019 Philipp Messer messerlab.org/slim

Teaching: Coding to learn evolution

- Develop coding-to-learn evolution course
- Students use graphical modeling software, like `SLiM`, to gain intuition for how processes of evolution work and interact



ID	N	↺	♂♀	♂♂	◕
p0	500	0.00	0.00	0.00	—
p1	500	0.00	0.00	0.00	—
p2	500	0.00	0.00	0.00	—
p3	500	0.00	0.00	0.00	—
p4	500	0.00	0.00	0.00	—
p5	500	0.00	0.00	0.00	—

Generation: 161

Input Commands:

```
initialize() {
   // Basic neutral simulation initialization
   initializeMutationRate(1e-7);
   initializeMutationType("m1", 0.5, "f", 0.0);
   initializeGenomicElementType("g1", m1, 1.0);
   initializeGenomicElement(g1, 0, 99999);
   initializeRecombinationRate(1e-8);

   // Mutation type used for the MCR drive locus
   initializeMutationType("m2", 0.5, "f", -0.1);
}

// Add subpops and set up stepping-stone migration
1 {
   for (i in 0:5)
      sim.addSubpop(i, 500);
   for (i in 1:5)
      sim.subpopulations[i].setMigrationRates(i-1, 0.001);
   for (i in 0:4)
      sim.subpopulations[i].setMigrationRates(i+1, 0.1);
}

// Add the MCR "drive allele" in a "seeding" event
100 late() {
```

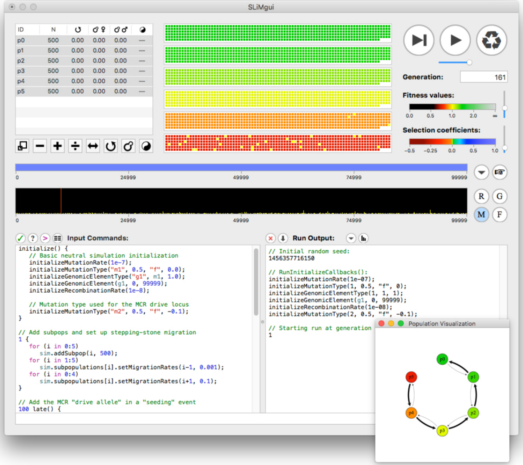
Run Output:

```
// Initial random seed:
1456357716150

// RunInitializeCallbacks():
initializeMutationRate(1e-07);
initializeMutationType(1, 0.5, "f", 0);
initializeGenomicElementType(1, 1, 1);
initializeGenomicElement(g1, 0, 99999);
initializeRecombinationRate(1e-08);
initializeMutationType(2, 0.5, "f", -0.1);

// Starting run at generation
1
```

© 2019 Philipp Messer messerlab.org/slim

Teaching: Coding to learn evolution

- Develop coding-to-learn evolution course
- Students use graphical modeling software, like `SLiM`, to gain intuition for how processes of evolution work and interact
- **Capstone activity**: Design macroevolutionary scenarios and write code to simulate genetic data under them



ID	N	↺	♂♀	♂♂	◕
p0	500	0.00	0.00	0.00	—
p1	500	0.00	0.00	0.00	—
p2	500	0.00	0.00	0.00	—
p3	500	0.00	0.00	0.00	—
p4	500	0.00	0.00	0.00	—
p5	500	0.00	0.00	0.00	—

Generation: 161

Input Commands:

```
initialize() {
   // Basic neutral simulation initialization
   initializeMutationRate(1e-7);
   initializeMutationType("m1", 0.5, "f", 0.0);
   initializeGenomicElementType("g1", m1, 1.0);
   initializeGenomicElement(g1, 0, 99999);
   initializeRecombinationRate(1e-8);

   // Mutation type used for the MCR drive locus
   initializeMutationType("m2", 0.5, "f", -0.1);
}

// Add subpops and set up stepping-stone migration
1 {
   for (i in 0:5)
      sim.addSubpop(i, 500);
   for (i in 1:5)
      sim.subpopulations[i].setMigrationRates(i-1, 0.001);
   for (i in 0:4)
      sim.subpopulations[i].setMigrationRates(i+1, 0.1);
}

// Add the MCR "drive allele" in a "seeding" event
100 late() {
```

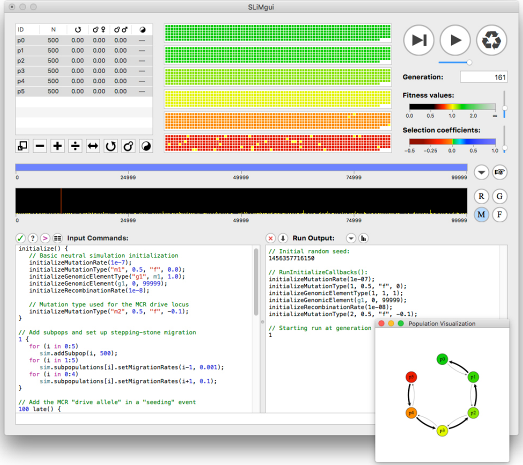
Run Output:

```
// Initial random seed:
1456357716150

// RunInitializeCallbacks():
initializeMutationRate(1e-07);
initializeMutationType(1, 0.5, "f", 0);
initializeGenomicElementType(1, 1, 1);
initializeGenomicElement(g1, 0, 99999);
initializeRecombinationRate(1e-08);
initializeMutationType(2, 0.5, "f", -0.1);

// Starting run at generation
1
```

© 2019 Philipp Messer messerlab.org/slim

Teaching: Coding to learn evolution

- Develop coding-to-learn evolution course
- Students use graphical modeling software, like `SLiM`, to gain intuition for how processes of evolution work and interact
- **Capstone activity**: Design macroevolutionary scenarios and write code to simulate genetic data under them
- Use these data for phylogenetic “bake-off”



ID	N	↺	♂♀	♂♂	◑
p0	500	0.00	0.00	0.00	—
p1	500	0.00	0.00	0.00	—
p2	500	0.00	0.00	0.00	—
p3	500	0.00	0.00	0.00	—
p4	500	0.00	0.00	0.00	—
p5	500	0.00	0.00	0.00	—

```
initialize() {
   // Basic neutral simulation initialization
   initializeMutationRate(1e-7);
   initializeMutationType("m1", 0.5, "f", 0.0);
   initializeGenomicElementType("g1", m1, 1.0);
   initializeGenomicElement(g1, 0, 99999);
   initializeRecombinationRate(1e-8);

   // Mutation type used for the MCR drive locus
   initializeMutationType("m2", 0.5, "f", -0.1);
}

// Add subpops and set up stepping-stone migration
1 {
   for (i in 0:5)
      sim.addSubpop(i, 500);
   for (i in 1:5)
      sim.subpopulations[i].setMigrationRates(i-1, 0.001);
   for (i in 0:4)
      sim.subpopulations[i].setMigrationRates(i+1, 0.1);
}

// Add the MCR "drive allele" in a "seeding" event
100 late() {
```

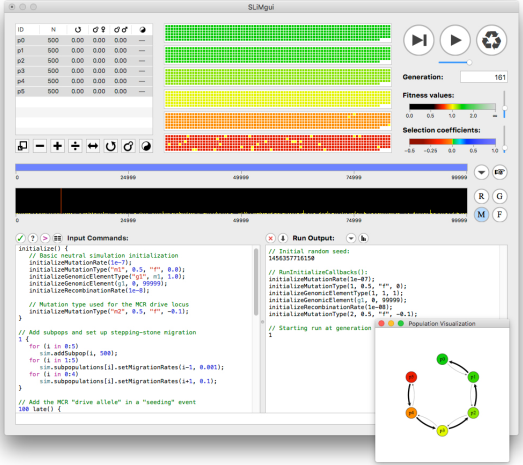
```
// Initial random seed:
1456357716150

// RunInitializeCallbacks():
initializeMutationRate(1e-07);
initializeMutationType(1, 0.5, "f", 0);
initializeGenomicElementType(1, 1, 1);
initializeGenomicElement(g1, 0, 99999);
initializeRecombinationRate(1e-08);
initializeMutationType(2, 0.5, "f", -0.1);

// Starting run at generation
1
```

© 2019 Philipp Messer messerlab.org/slim

Teaching: Coding to learn evolution

- Develop coding-to-learn evolution course
- Students use graphical modeling software, like `SLiM`, to gain intuition for how processes of evolution work and interact
- **Capstone activity**: Design macroevolutionary scenarios and write code to simulate genetic data under them
- Use these data for phylogenetic "bake-off"

- How robust are estimates under the BDB model when applied to data generated under ***very*** different models?



ID	N	↺	♂♀	♂♂	◐
p0	500	0.00	0.00	0.00	—
p1	500	0.00	0.00	0.00	—
p2	500	0.00	0.00	0.00	—
p3	500	0.00	0.00	0.00	—
p4	500	0.00	0.00	0.00	—
p5	500	0.00	0.00	0.00	—

```
initialize() {
   // Basic neutral simulation initialization
   initializeMutationRate(1e-7);
   initializeMutationType("m1", 0.5, "f", 0.0);
   initializeGenomicElementType("g1", m1, 1.0);
   initializeGenomicElement(g1, 0, 99999);
   initializeRecombinationRate(1e-8);

   // Mutation type used for the MCR drive locus
   initializeMutationType("m2", 0.5, "f", -0.1);
}

// Add subpops and set up stepping-stone migration
1 {
   for (i in 0:5)
      sim.addSubpop(i, 500);
   for (i in 1:5)
      sim.subpopulations[i].setMigrationRates(i-1, 0.001);
   for (i in 0:4)
      sim.subpopulations[i].setMigrationRates(i+1, 0.1);
}

// Add the MCR "drive allele" in a "seeding" event
100 late() {
```

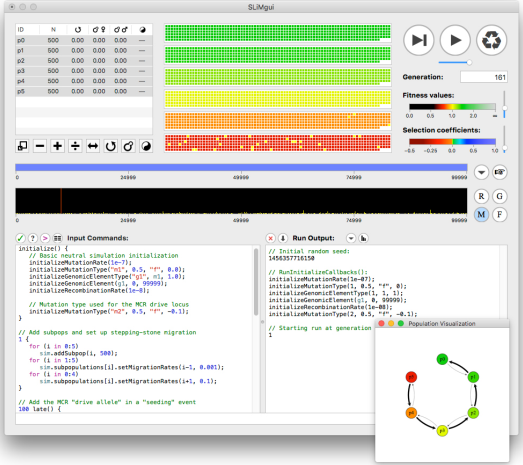
```
// Initial random seed:
1456357716150

// RunInitializeCallbacks():
initializeMutationRate(1e-07);
initializeMutationType(1, 0.5, "f", 0);
initializeGenomicElementType(1, 1, 1);
initializeGenomicElement(g1, 0, 99999);
initializeRecombinationRate(1e-08);
initializeMutationType(2, 0.5, "f", -0.1);

// Starting run at generation
1
```

© 2019 Philipp Messer messerlab.org/slim

Teaching: Coding to learn evolution

- Develop coding-to-learn evolution course
- Students use graphical modeling software, like `SLiM`, to gain intuition for how processes of evolution work and interact
- **Capstone activity**: Design macroevolutionary scenarios and write code to simulate genetic data under them
- Use these data for phylogenetic “bake-off”

- How robust are estimates under the BDB model when applied to data generated under *very* different models?
- Students co-author paper



ID	N	↺	⚦♀	⚦♂	◐
p0	500	0.00	0.00	0.00	—
p1	500	0.00	0.00	0.00	—
p2	500	0.00	0.00	0.00	—
p3	500	0.00	0.00	0.00	—
p4	500	0.00	0.00	0.00	—
p5	500	0.00	0.00	0.00	—

Generation: 161

Input Commands:

```
initialize() {
   // Basic neutral simulation initialization
   initializeMutationRate(1e-7);
   initializeMutationType("m1", 0.5, "f", 0.0);
   initializeGenomicElementType("g1", m1, 1.0);
   initializeGenomicElement(g1, 0, 99999);
   initializeRecombinationRate(1e-8);

   // Mutation type used for the MCR drive locus
   initializeMutationType("m2", 0.5, "f", -0.1);
}

// Add subpops and set up stepping-stone migration
1 {
   for (i in 0:5)
      sim.addSubpop(i, 500);
   for (i in 1:5)
      sim.subpopulations[i].setMigrationRates(i-1, 0.001);
   for (i in 0:4)
      sim.subpopulations[i].setMigrationRates(i+1, 0.1);
}

// Add the MCR "drive allele" in a "seeding" event
100 late() {
```

Run Output:

```
// Initial random seed:
1456357716150

// RunInitializeCallbacks():
initializeMutationRate(1e-07);
initializeMutationType(1, 0.5, "f", 0);
initializeGenomicElementType(1, 1, 1);
initializeGenomicElement(g1, 0, 99999);
initializeRecombinationRate(1e-08);
initializeMutationType(2, 0.5, "f", -0.1);

// Starting run at generation
1
```

© 2019 Philipp Messer messerlab.org/slim

Acknowledgments

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

Computation:

- Alabama Supercomputer Authority
- Auburn University Hopper Cluster

Funding:

DEB 1656004

Photo credits:

- Rafe Brown
- Perry Wood, Jr.
- PhyloPic

Questions?

joaks@auburn.edu

phyletica.org

© 2007 Boris Kulikov boris-kulikov.blogspot.com