Generalizing phylogenetics to infer patterns predicted by processes of diversification

Jamie Oaks
Auburn University
phyletica.org
@jamoaks

Scan for slides:

phyletica.org/slides/au-dbs-2021.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
- Saman Jahangiri
- 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*

Generalizing phylogenetics to infer patterns predicted by processes of diversification

Jamie Oaks
Auburn University
phyletica.org
@jamoaks

Scan for slides:

phyletica.org/slides/au-dbs-2021.pdf

© 2007 Boris Kulikov boris-kulikov.blogspot.com

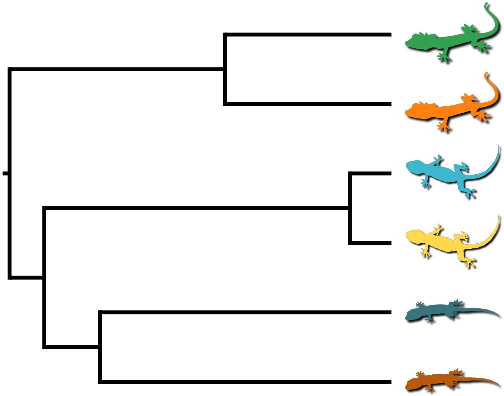
- Phylogenetics is rapidly becoming the statistical foundation of biology

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

© 2007 Boris Kulikov boris-kulikov.blogspot.com

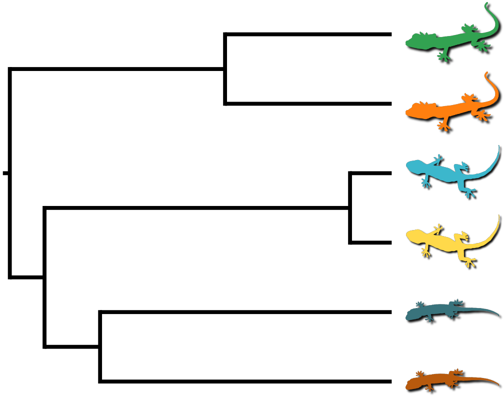
- Phylogenetics is rapidly becoming the statistical foundation of biology
- "Big data" present exciting possibilities and challenges
- Many opportunities to develop new ways to study biology in light of phylogeny

© 2007 Boris Kulikov boris-kulikov.blogspot.com

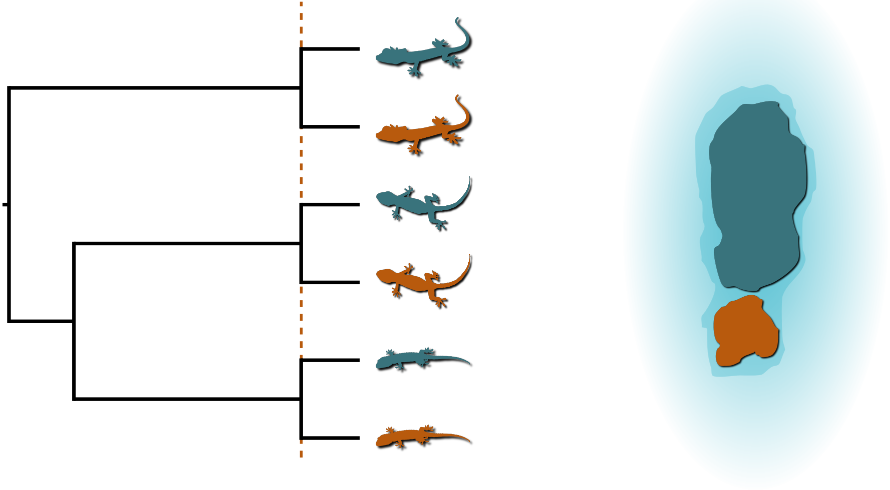


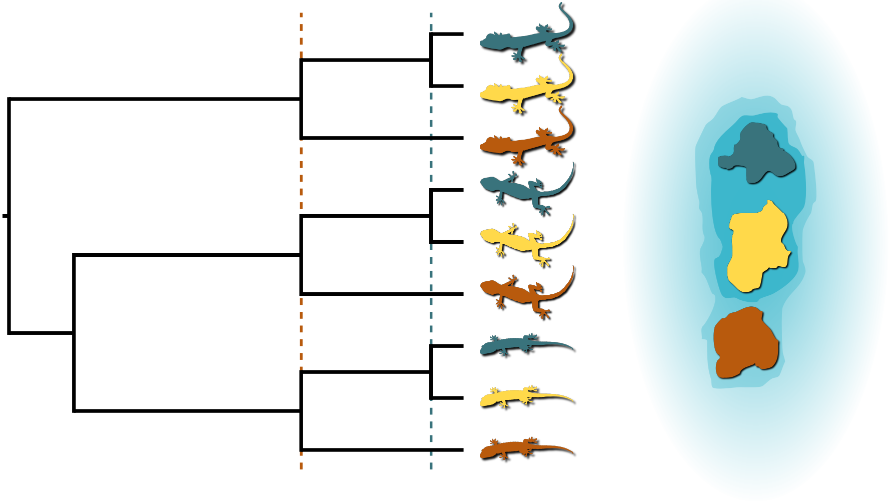
phylopic.org CC0

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



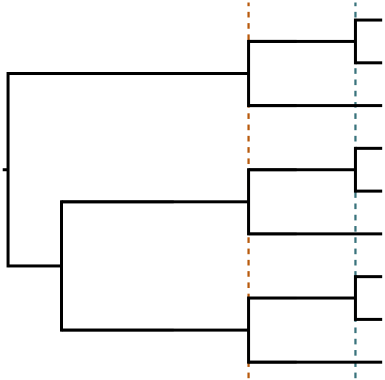
phylopic.org CC0





Biogeography

- Environmental changes that affect whole communities of species

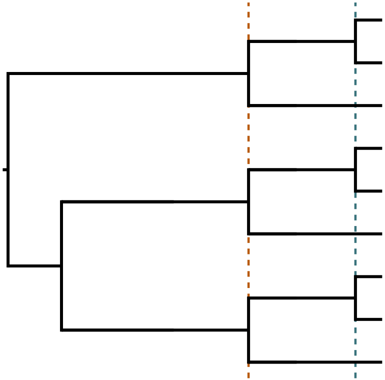


Biogeography

- Environmental changes that affect whole communities of species

Genome evolution

- Duplication of a chromosome segment harboring gene families



Biogeography

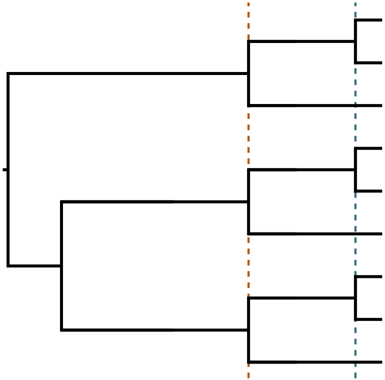
- Environmental changes that affect whole communities of species

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

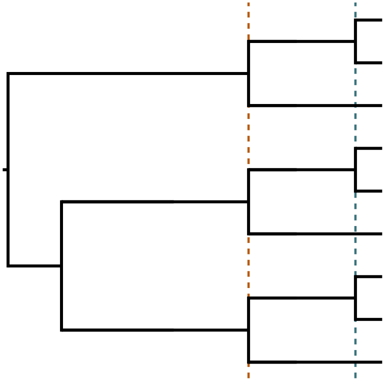
- Duplication of a chromosome segment harboring gene families

Epidemiology

- Transmission at social gatherings

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

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

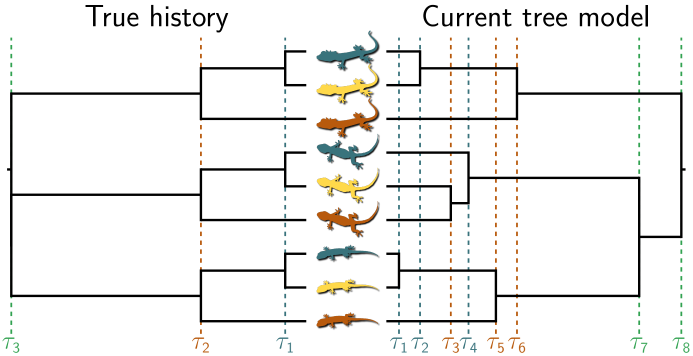


Why account for shared divergences?

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

Why account for shared divergences?

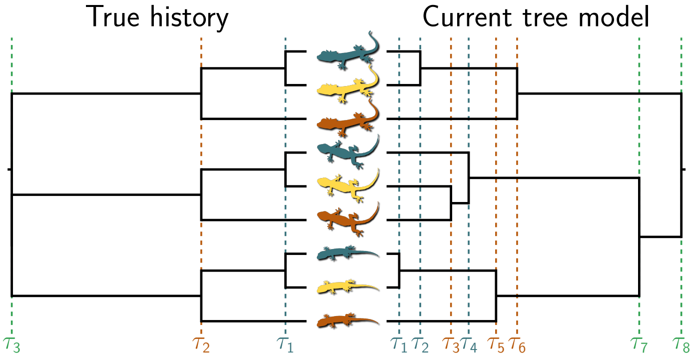
1. Improve inference



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

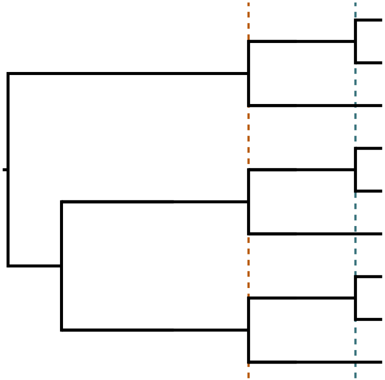
- Duplication of a chromosome segment harboring gene families

Epidemiology

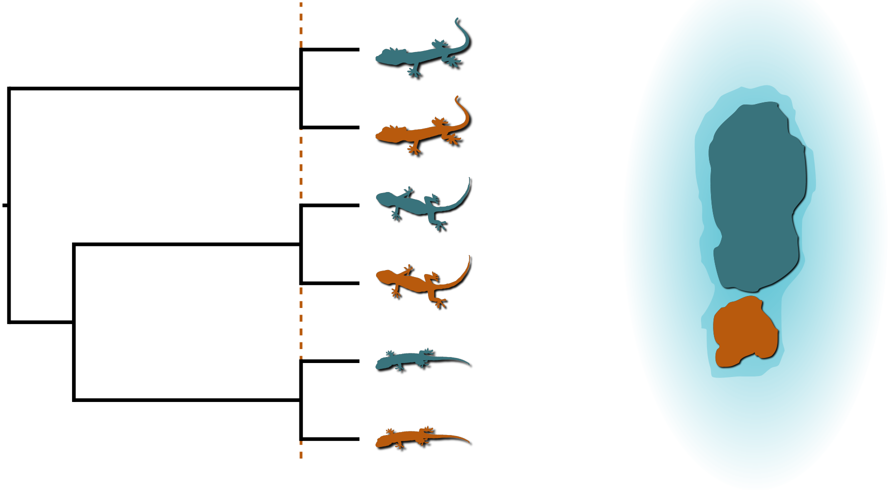
- Transmission at social gatherings

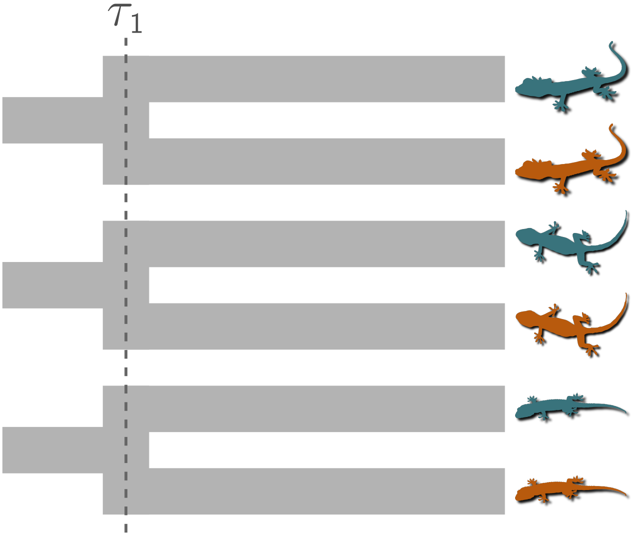
Endosymbiont evolution (e.g., parasites, microbiome)

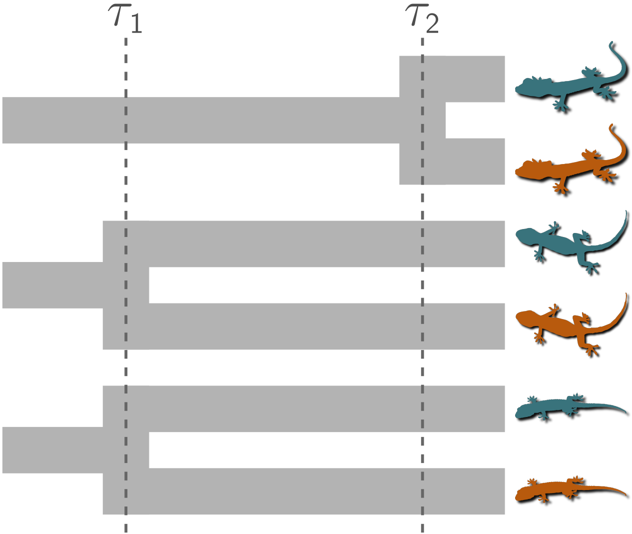
- Speciation of the host
- Co-colonization of new host species

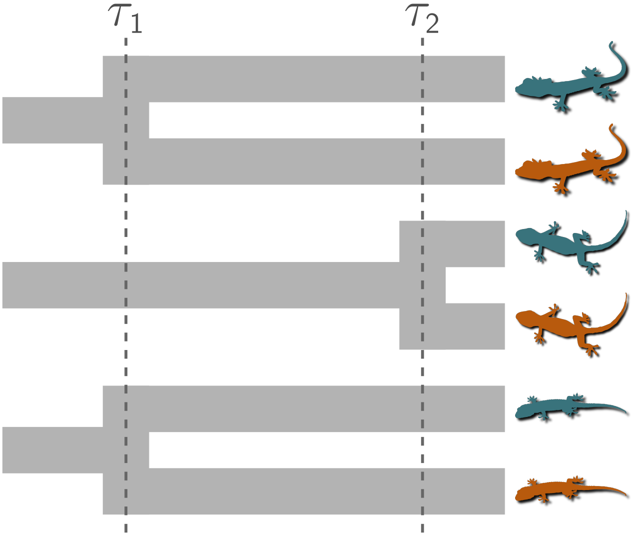


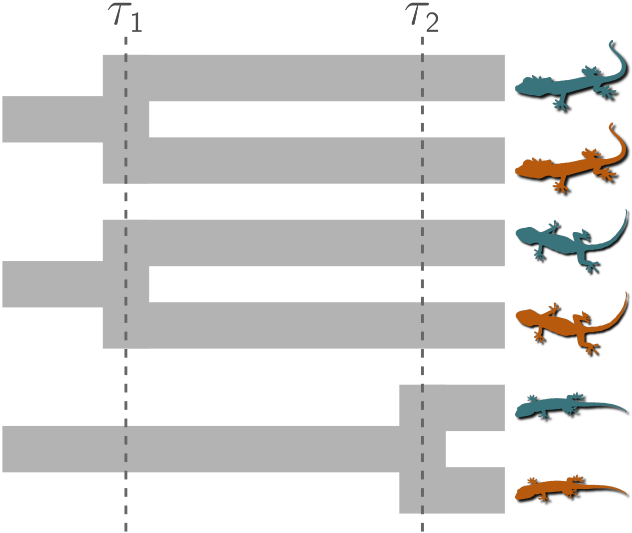
- Approaches to the problem
    - A pairwise approach (keep it “simple”)
    - An aside for a related problem with population demography
    - A fully phylogenetic approach
- The next 5 years in the Phyletica Lab

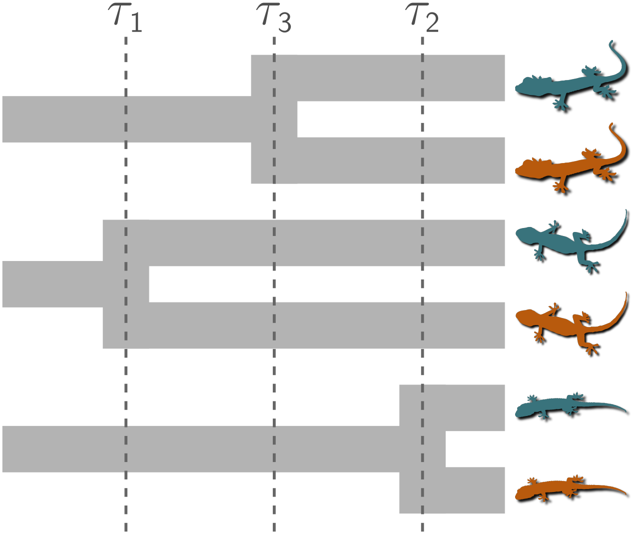



τ_1

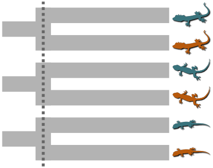

τ_1
τ_2


τ_1
τ_2

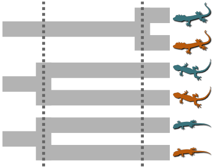

τ_1
τ_2


τ_1
τ_3
τ_2

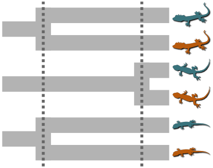
m_1



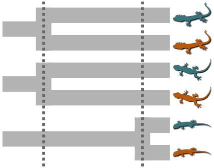
m_2



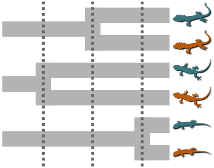
m_3



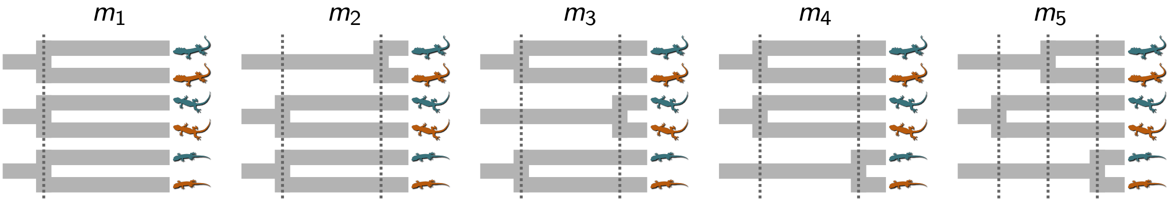
m_4



m_5



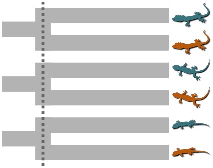
J. R. Oaks et al. (2013). *Evolution* 67: 991–1010, J. R. Oaks (2014). *BMC Evolutionary Biology* 14: 150



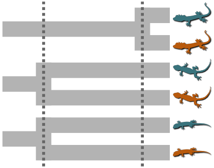
We want to infer the model and divergence times given genetic data

J. R. Oaks et al. (2013). *Evolution* 67: 991–1010, J. R. Oaks (2014). *BMC Evolutionary Biology* 14: 150

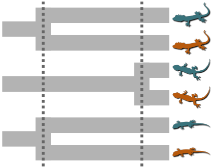
$p(m_1 \mid \mathbf{D})$



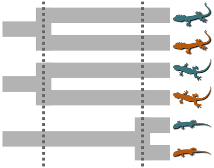
$p(m_2 \mid \mathbf{D})$



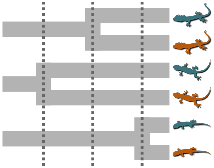
$p(m_3 \mid \mathbf{D})$



$p(m_4 \mid \mathbf{D})$

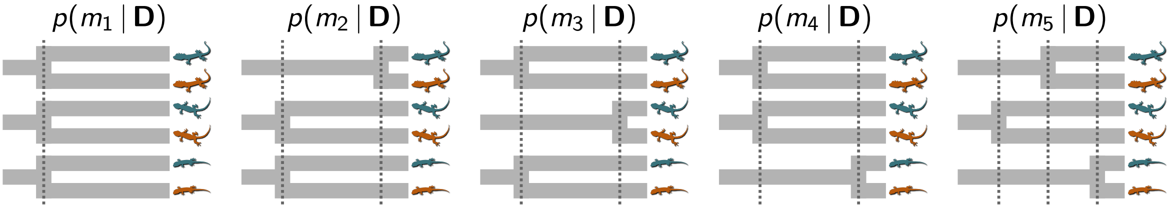


$p(m_5 \mid \mathbf{D})$



We want to infer the model and divergence times given genetic data

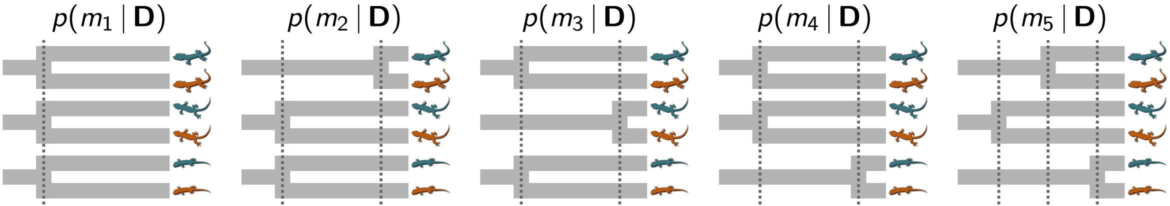
J. R. Oaks et al. (2013). *Evolution* 67: 991–1010, J. R. Oaks (2014). *BMC Evolutionary Biology* 14: 150



We want to infer the model and divergence times given genetic data

$$p(m_i \mid \mathbf{D}) \propto p(\mathbf{D} \mid m_i) p(m_i)$$

J. R. Oaks et al. (2013). *Evolution* 67: 991–1010, J. R. Oaks (2014). *BMC Evolutionary Biology* 14: 150

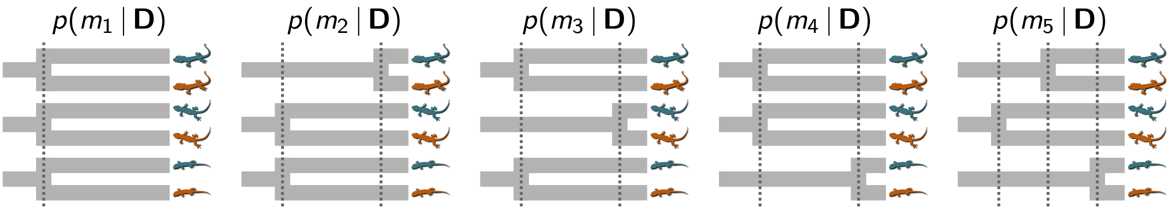


We want to infer the model and divergence times given genetic data

$$p(m_i \mid \mathbf{D}) \propto p(\mathbf{D} \mid m_i) p(m_i)$$

$$p(\mathbf{D} \mid m_i) = \int_\theta p(\mathbf{D} \mid \theta, m_i) p(\theta \mid m_i) d\theta$$

J. R. Oaks et al. (2013). *Evolution* 67: 991–1010, J. R. Oaks (2014). *BMC Evolutionary Biology* 14: 150



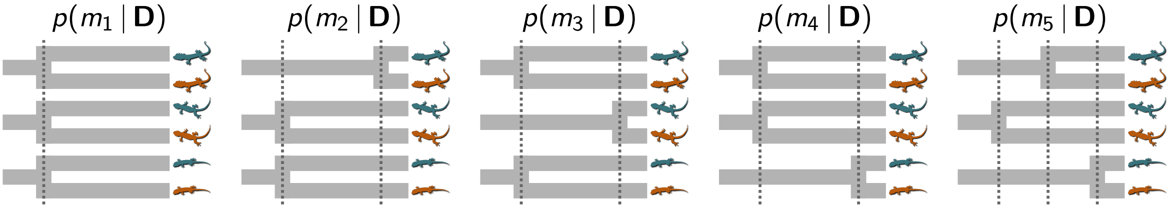
We want to infer the model and divergence times given genetic data

$$p(m_i \mid \mathbf{D}) \propto p(\mathbf{D} \mid m_i) p(m_i)$$

$$p(\mathbf{D} \mid m_i) = \int_{\theta} p(\mathbf{D} \mid \theta, m_i) p(\theta \mid m_i) d\theta$$

- Divergence times
- Gene trees
- Substitution parameters
- Demographic parameters

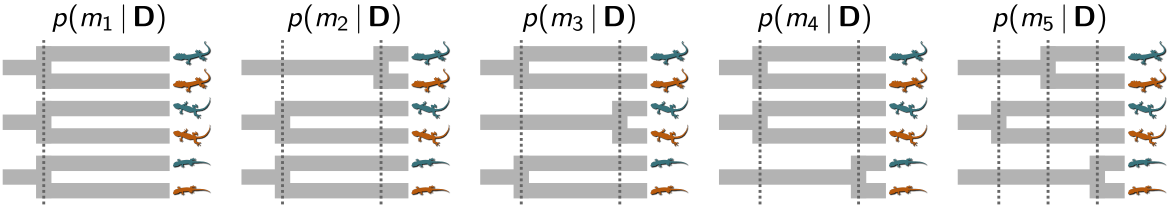
J. R. Oaks et al. (2013). *Evolution* 67: 991–1010, J. R. Oaks (2014). *BMC Evolutionary Biology* 14: 150



We want to infer the model and divergence times given genetic data

Challenges:

J. R. Oaks et al. (2013). *Evolution* 67: 991–1010, J. R. Oaks (2014). *BMC Evolutionary Biology* 14: 150

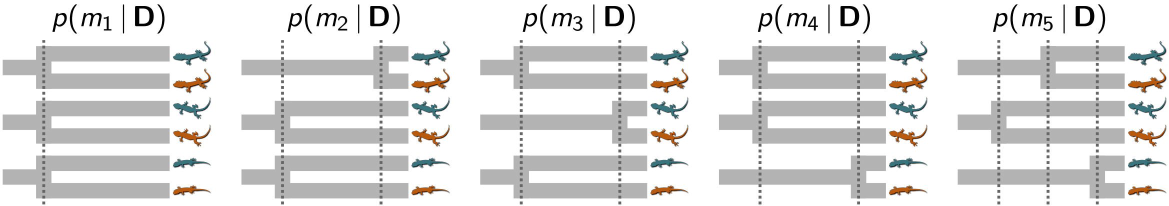


We want to infer the model and divergence times given genetic data

Challenges:

1. Likelihood is tractable, but gene trees are difficult

J. R. Oaks et al. (2013). *Evolution* 67: 991–1010, J. R. Oaks (2014). *BMC Evolutionary Biology* 14: 150

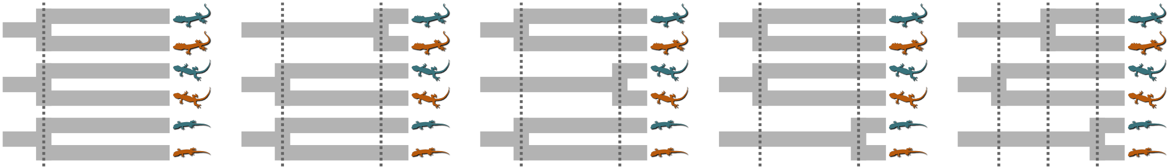


We want to infer the model and divergence times given genetic data

Challenges:

1. Likelihood is tractable, but gene trees are difficult
2. Sampling over all possible models

J. R. Oaks et al. (2013). *Evolution* 67: 991–1010, J. R. Oaks (2014). *BMC Evolutionary Biology* 14: 150



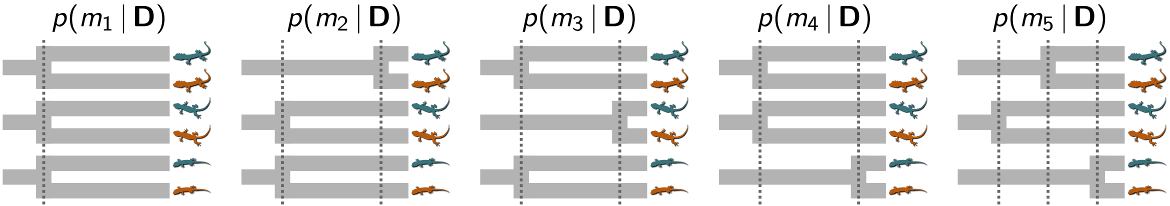
$p(m_1 \mid \mathbf{D})$ $p(m_2 \mid \mathbf{D})$ $p(m_3 \mid \mathbf{D})$ $p(m_4 \mid \mathbf{D})$ $p(m_5 \mid \mathbf{D})$

We want to infer the model and divergence times given genetic data

Challenges:

1. Likelihood is tractable, but gene trees are difficult
2. Sampling over all possible models
    - 3 taxa = 5 models

J. R. Oaks et al. (2013). *Evolution* 67: 991–1010, J. R. Oaks (2014). *BMC Evolutionary Biology* 14: 150

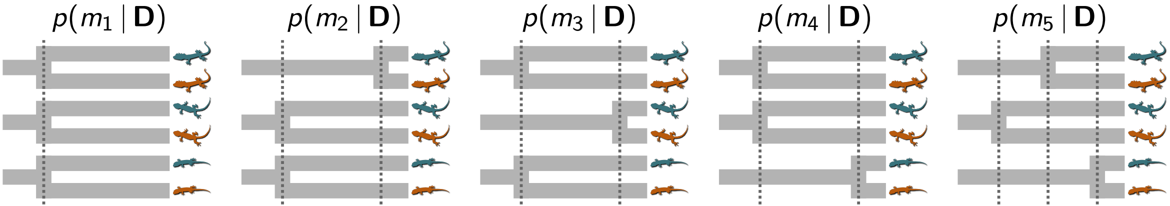


We want to infer the model and divergence times given genetic data

Challenges:

1. Likelihood is tractable, but gene trees are difficult
2. Sampling over all possible models
    - 3 taxa = 5 models
    - 10 taxa = 115,975 models

J. R. Oaks et al. (2013). *Evolution* 67: 991–1010, J. R. Oaks (2014). *BMC Evolutionary Biology* 14: 150

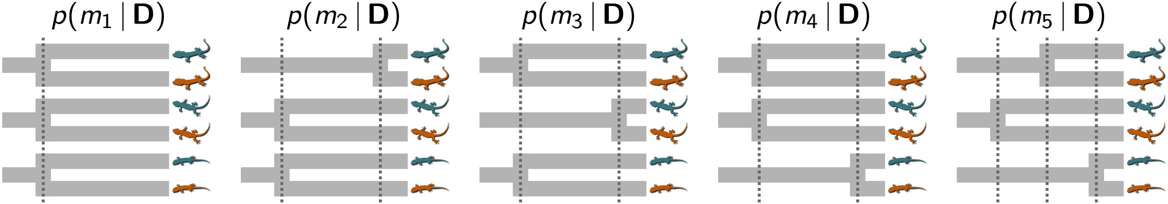


We want to infer the model and divergence times given genetic data

Challenges:

1. Likelihood is tractable, but gene trees are difficult
2. Sampling over all possible models
   - 3 taxa = 5 models
   - 10 taxa = 115,975 models
   - 20 taxa = 51,724,158,235,372 models!!

J. R. Oaks et al. (2013). *Evolution* 67: 991–1010, J. R. Oaks (2014). *BMC Evolutionary Biology* 14: 150



We want to infer the model and divergence times given genetic data

Challenges:

1. Likelihood is tractable, but gene trees are difficult
2. Sampling over all possible models
    - 3 taxa = 5 models
    - 10 taxa = 115,975 models
    - 20 taxa = 51,724,158,235,372 models!!

Approximate Bayesian computation (ABC) methods do not perform well for this model-choice problem

J. R. Oaks et al. (2013). *Evolution* 67: 991–1010, J. R. Oaks (2014). *BMC Evolutionary Biology* 14: 150

Ecoevolity

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

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

- *Goal: Fast, full-likelihood Bayesian method to infer patterns of co-diversification from genome-scale data*

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

[2] R. M. Neal (2000). *Journal of Computational and Graphical Statistics* 9: 249–265

Ecoevolity

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

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

- Analytically integrate over gene trees and mutational histories[1]
- Dirichlet-process prior across divergence models
- Gibbs sampling[2] to numerically sample models

- *Goal: Fast, full-likelihood Bayesian method to infer patterns of co-diversification from genome-scale data*

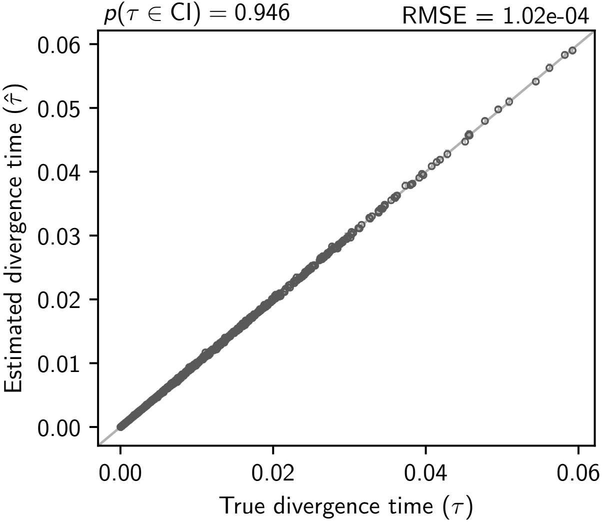
---

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

[2] R. M. Neal (2000). *Journal of Computational and Graphical Statistics* 9: 249–265

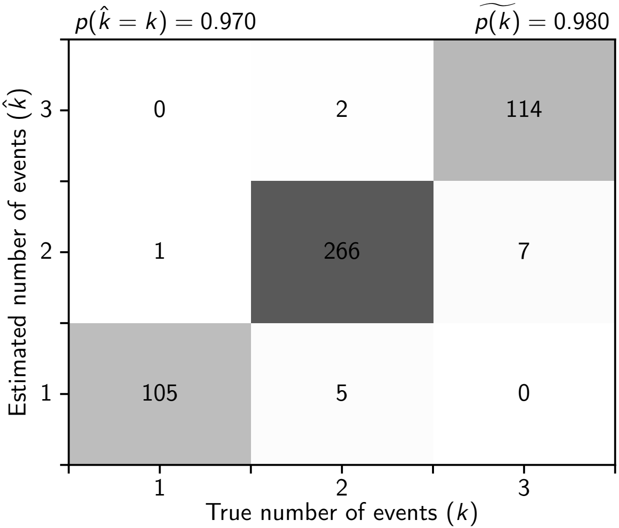
Does it work?

Simulation results



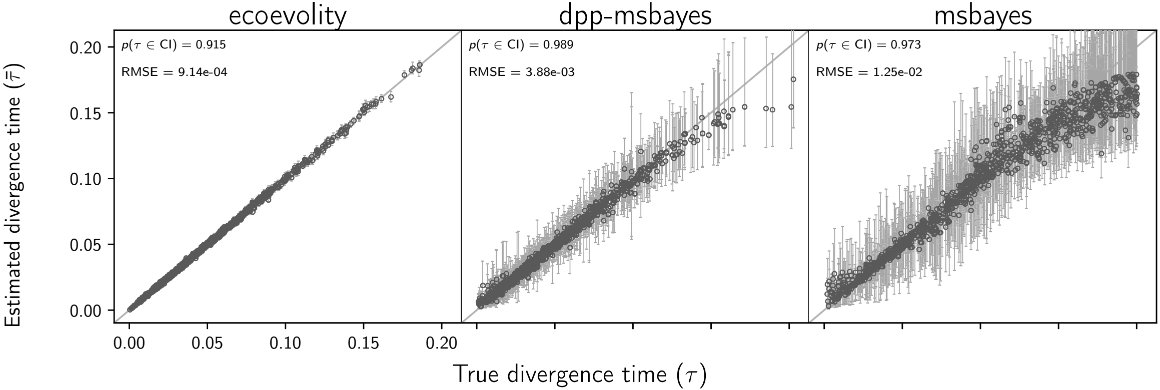
J. R. Oaks (2019). *Systematic Biology* 68: 371–395

Simulation results



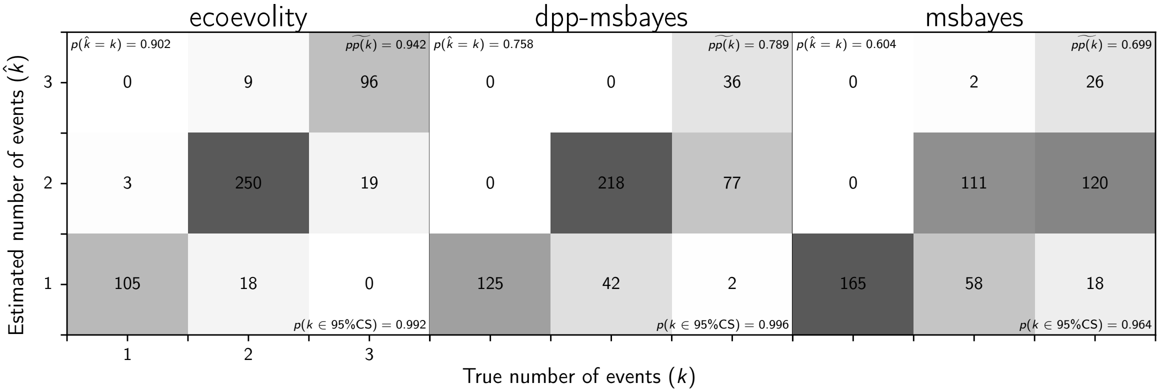
J. R. Oaks (2019). *Systematic Biology* 68: 371–395

“Bake-off” results



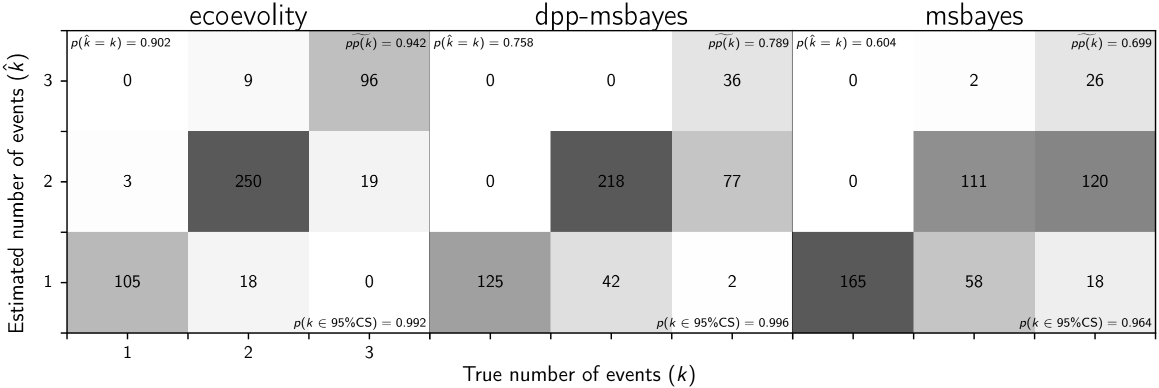
J. R. Oaks (2019). *Systematic Biology* 68: 371–395

“Bake-off” results



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

“Bake-off” results

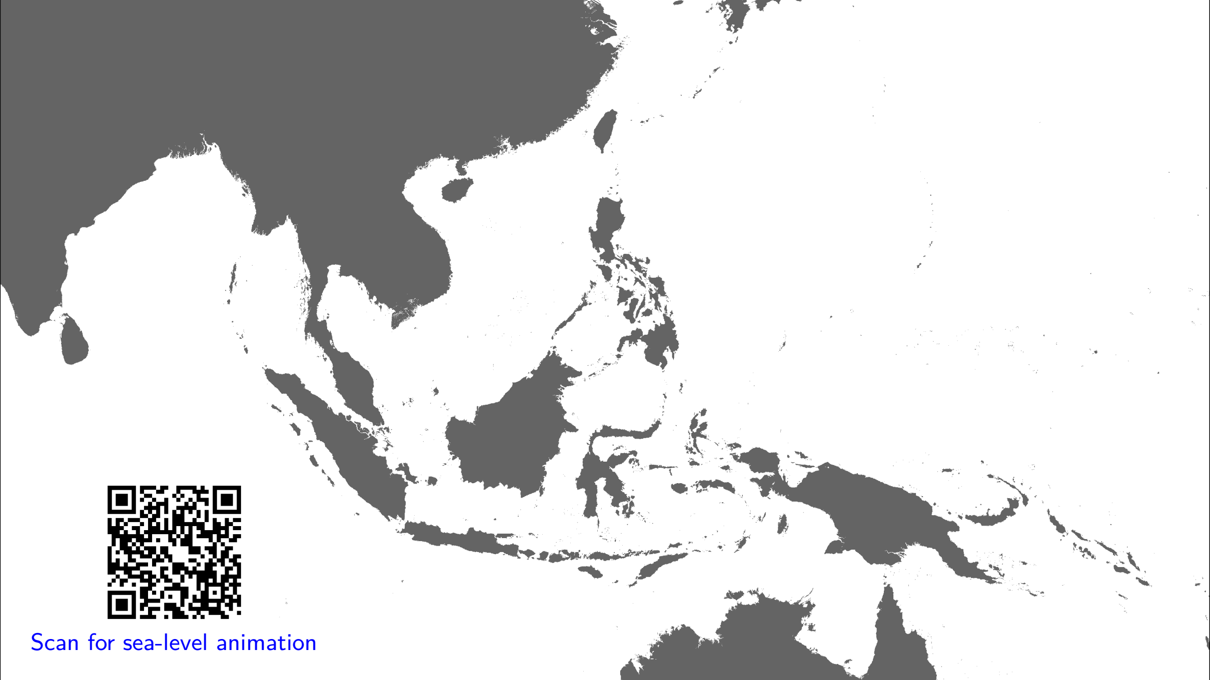


Average run time:

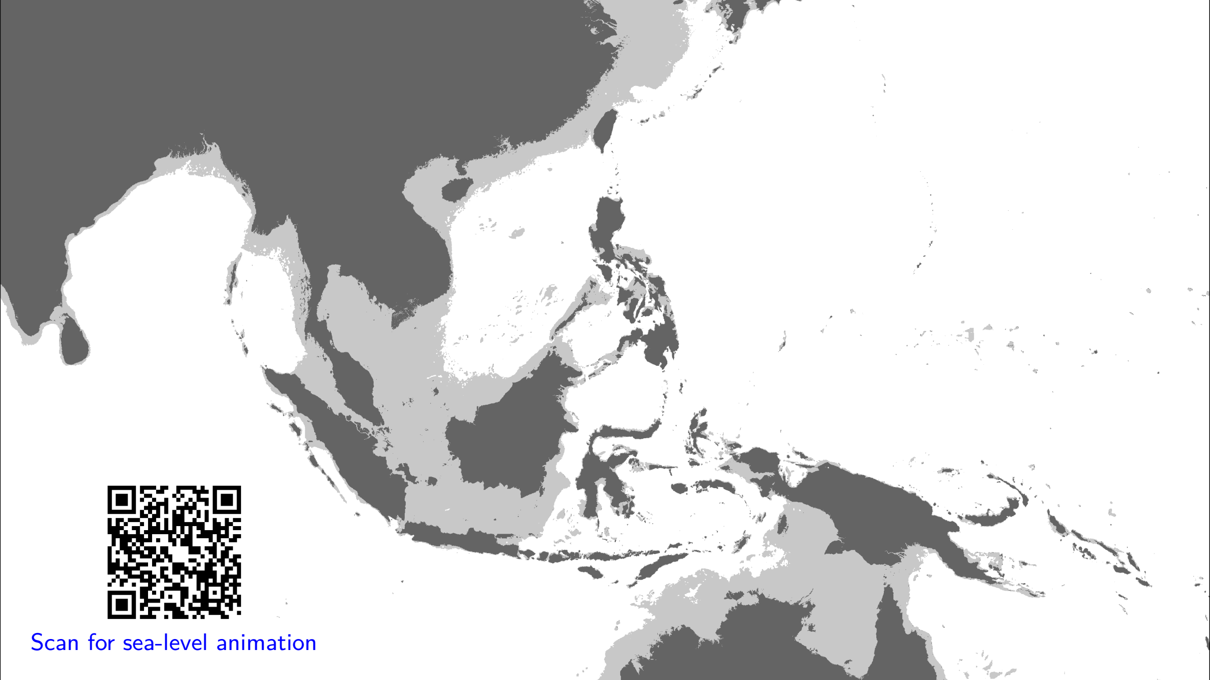
33.4 minutes

4.4 days

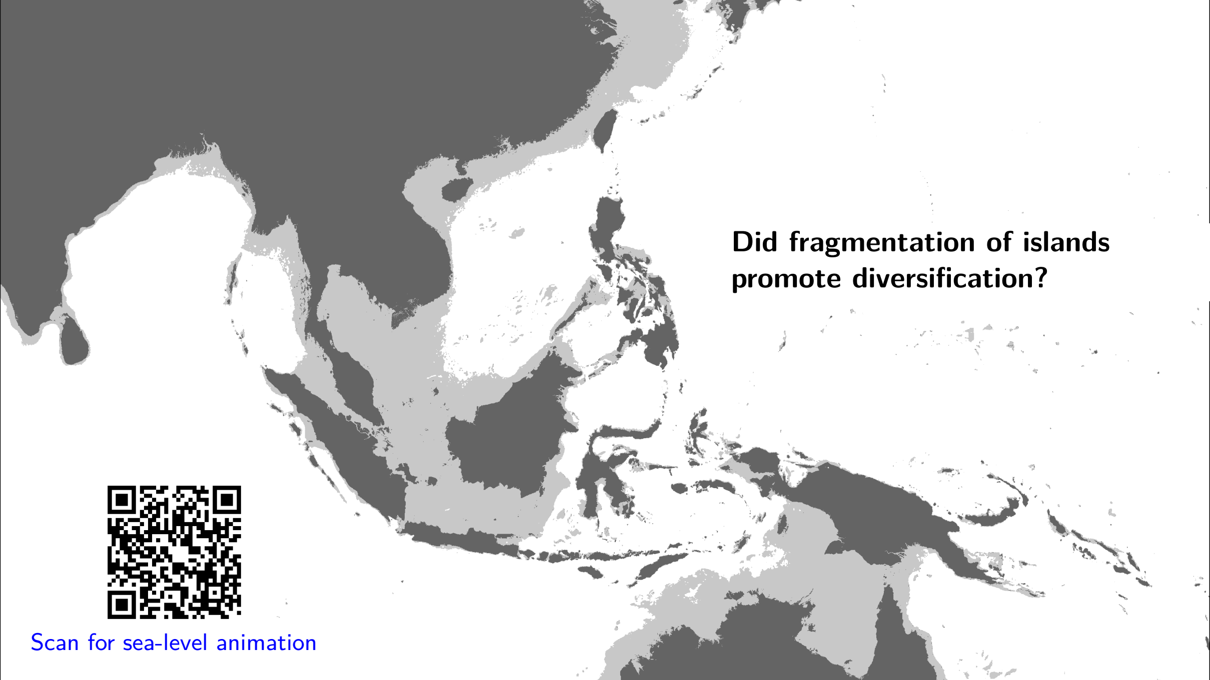
J. R. Oaks (2019). *Systematic Biology* 68: 371–395



Scan for sea-level animation

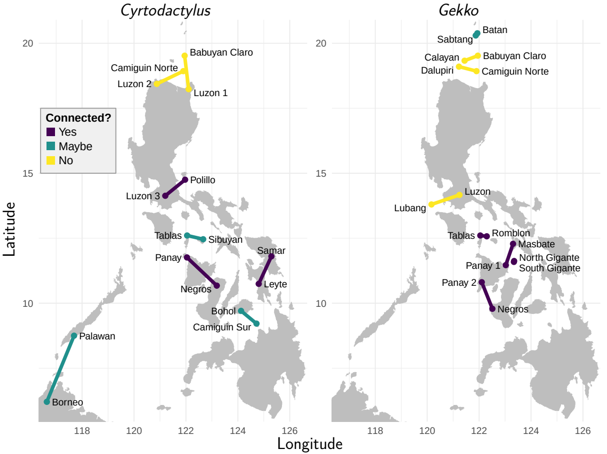


Scan for sea-level animation



Did fragmentation of islands promote diversification?

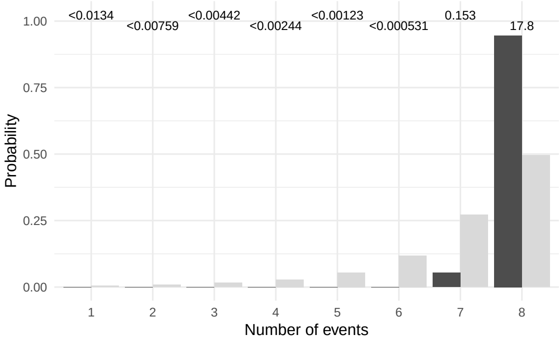
Scan for sea-level animation



- Sampled 2–5 individuals from 8 pairs of populations for both *Cyrtodactylus* and *Gekko*
- Collected short DNA sequences (RADseq) from across genome of each individual

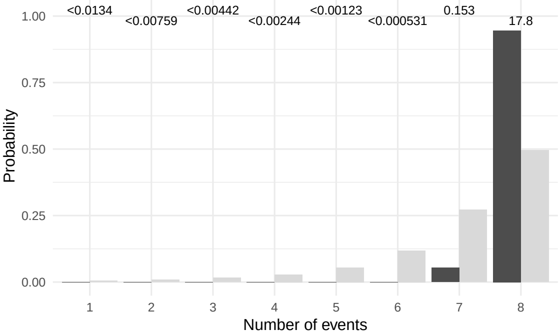
J. R. Oaks, C. D. Siler, and R. M. Brown (2019). *Evolution* 73: 1151–1167

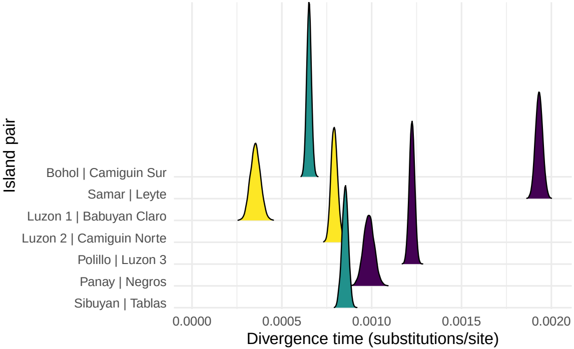
Results: *Cyrtodactylus*



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

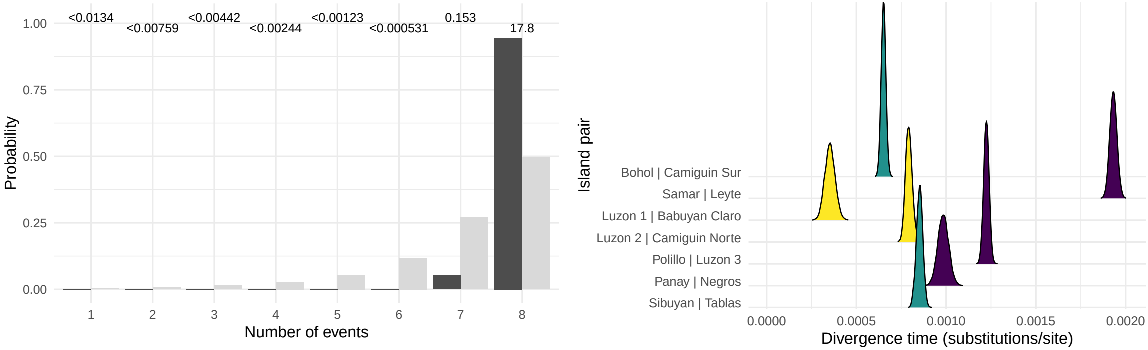
Results: *Cyrtodactylus*





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

Results: *Cyrtodactylus*



Strong support for independent divergences

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

Take-home points

- Support against the climate-driven “species-pump” hypothesis

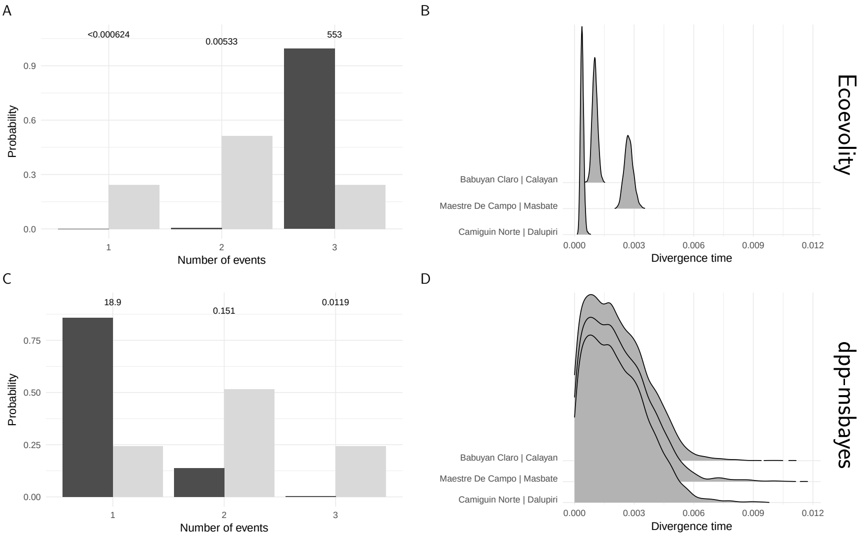
Take-home points

- Support against the climate-driven “species-pump” hypothesis
- Habitat heterogeneity and rare over-water dispersal via rafting on vegetation are likely more important

Take-home points

- Support against the climate-driven “species-pump” hypothesis
- Habitat heterogeneity and rare over-water dispersal via rafting on vegetation are likely more important
- Full-likelihood, Bayesian approach is faster and more accurate than ABC

ABC vs ecoevolity results



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

Why so different?

J. R. Oaks et al. (2019). *Systematic Biology* 68: 681–697

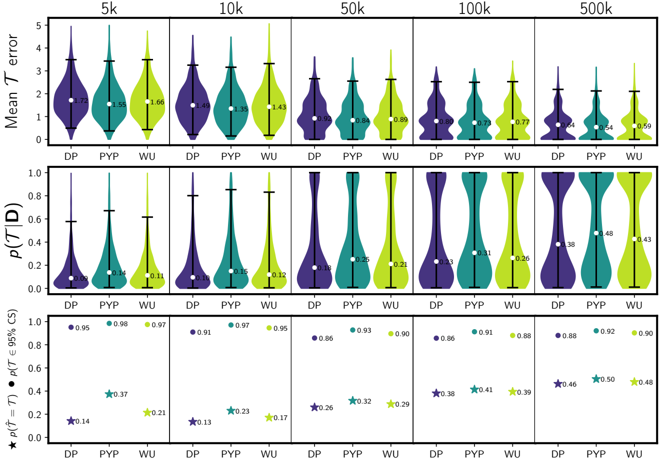
Interactive coin-flipping demo

tl;dr:

Averaging likelihood with respect to “diffuse” priors creates a strong penalty against divergence-time parameters

Better priors than the Dirichlet process?

Better priors than the Dirichlet process?



Developed and tested 2 additional priors over models

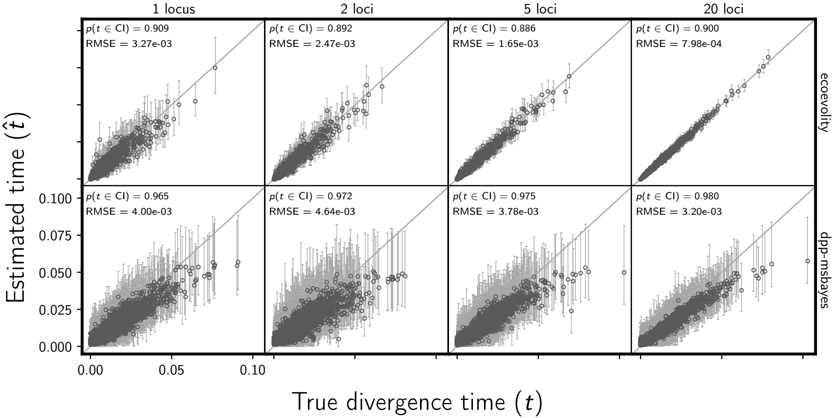
Pitman-Yor process performs better

What about smaller data sets?

What about smaller data sets?

Full-likelihood Bayesian approach performs better than ABC

Many “legacy” data sets can be re-evaluated



Open science: everything is available. . .

Software:

- Ecoevolity:
  phyletica.org/ecoevolity

Open-Science Notebooks:

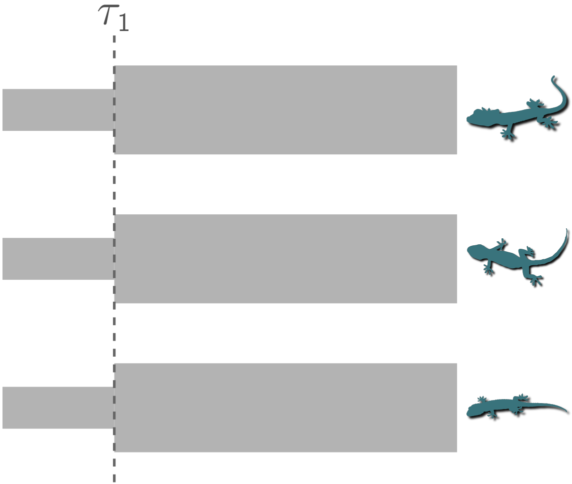
- Gecko RADseq:
  github.com/phyletica/gekgo
- Simulation analyses:
  github.com/phyletica/ecoevolity-experiments
  github.com/phyletica/ecoevolity-model-prior
  github.com/phyletica/codiv-sanger-bake-off

phyletica.org/codiv-sanger-bake-off

Approaches to the problem

- A pairwise approach (keep it “simple”)
- An aside for a related problem with population demography
- A fully phylogenetic approach

The next 5 years in the Phyletica Lab

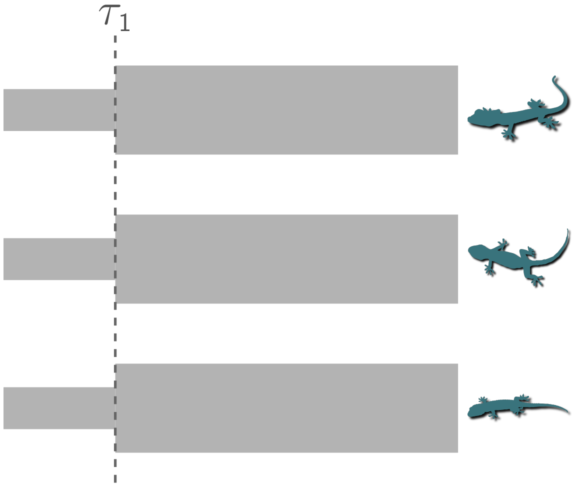


- Recent interest in testing shared demographic changes

[1] Y. L. Chan, D. Schanzenbach, and M. J. Hickerson (2014). *Molecular Biology and Evolution* 31: 2501–2515

[2] A. T. Xue and M. J. Hickerson (2015). *Molecular Ecology* 24: 6223–6240

[3] A. T. Xue and M. J. Hickerson (2017). *Molecular Ecology Resources* 17: e212–e224



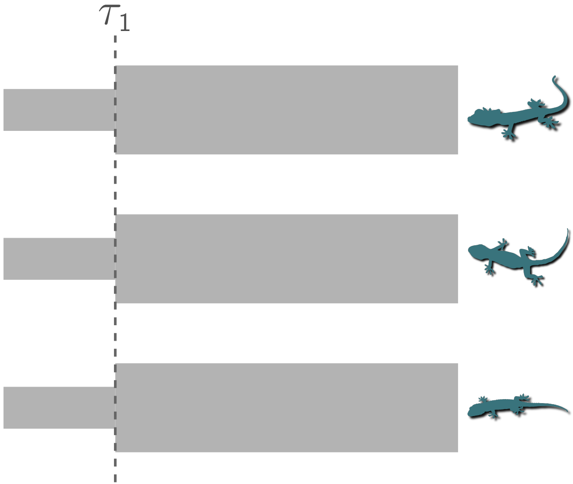
- Recent interest in testing shared demographic changes
- Several nice ABC approaches[1,2,3]
  - Inferred simultaneous expansion of 5 Alaskan populations of sticklebacks (posterior probability $= 0.99$)[2]

---

[1] Y. L. Chan, D. Schanzenbach, and M. J. Hickerson (2014). *Molecular Biology and Evolution* 31: 2501–2515

[2] A. T. Xue and M. J. Hickerson (2015). *Molecular Ecology* 24: 6223–6240

[3] A. T. Xue and M. J. Hickerson (2017). *Molecular Ecology Resources* 17: e212–e224

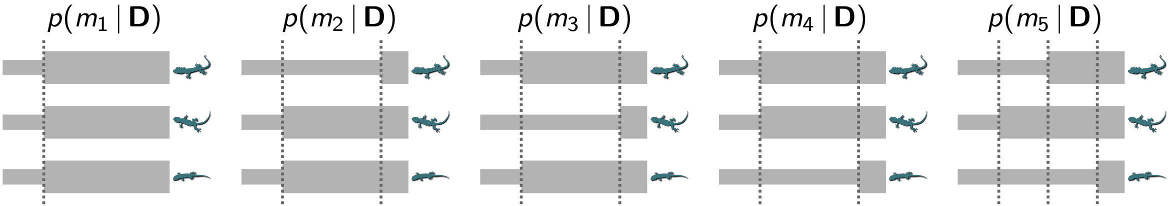


- Recent interest in testing shared demographic changes
- Several nice ABC approaches[1,2,3]
  - Inferred simultaneous expansion of 5 Alaskan populations of sticklebacks (posterior probability = 0.99)[2]
- It's a tricky inference problem
  - Change in population size can become unidentifiable in 3 ways

[1] Y. L. Chan, D. Schanzenbach, and M. J. Hickerson (2014). *Molecular Biology and Evolution* 31: 2501–2515

[2] A. T. Xue and M. J. Hickerson (2015). *Molecular Ecology* 24: 6223–6240

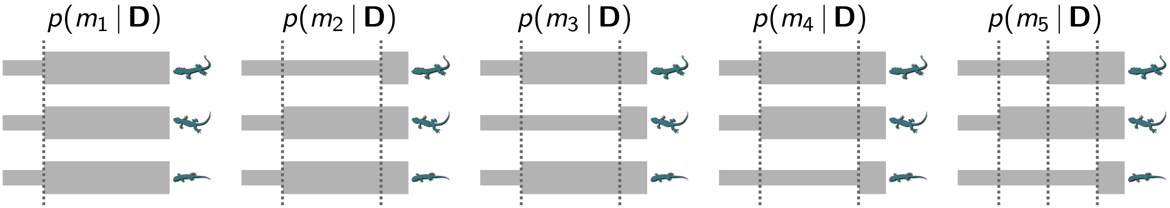
[3] A. T. Xue and M. J. Hickerson (2017). *Molecular Ecology Resources* 17: e212–e224



$p(m_1 \mid \mathbf{D})$ $p(m_2 \mid \mathbf{D})$ $p(m_3 \mid \mathbf{D})$ $p(m_4 \mid \mathbf{D})$ $p(m_5 \mid \mathbf{D})$

Given genomic data, can we infer the correct model and the timing of the demographic events?

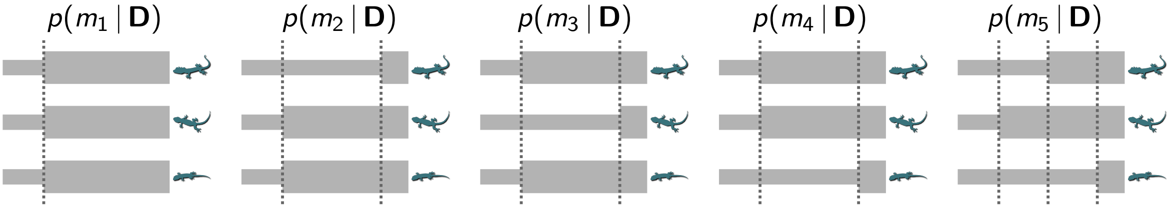
J. R. Oaks, N. L'Bahy, and K. A. Cobb (2020). *Evolution* 74: 2184–2206



$p(m_1 \mid \mathbf{D})$ $p(m_2 \mid \mathbf{D})$ $p(m_3 \mid \mathbf{D})$ $p(m_4 \mid \mathbf{D})$ $p(m_5 \mid \mathbf{D})$

Given genomic data, can we infer the correct model and the timing of the demographic events?

J. R. Oaks, N. L'Bahy, and K. A. Cobb (2020). *Evolution* 74: 2184–2206

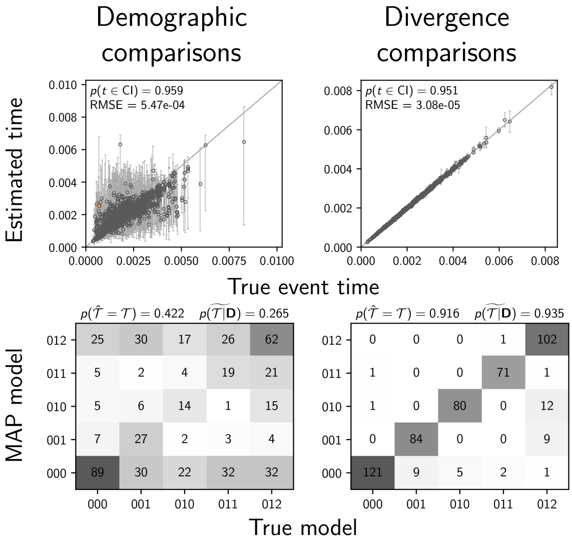


Given genomic data, can we infer the correct model and the timing of the demographic events?

- Implemented full-likelihood Bayesian approach in ecoevolity
- Assessed performance with simulations
- Applied to stickleback genomic data

J. R. Oaks, N. L'Bahy, and K. A. Cobb (2020). *Evolution* 74: 2184–2206

Simulation results



Demographic comparisons

Divergence comparisons

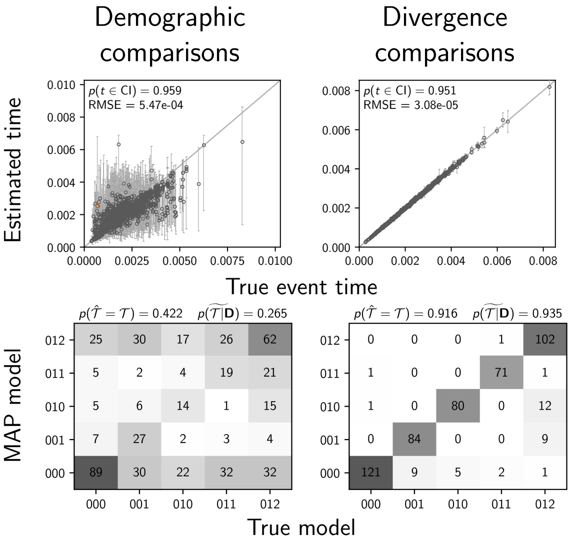
	000	001	010	011	012
012	25	30	17	26	62
011	5	2	4	19	21
010	5	6	14	1	15
001	7	27	2	3	4
000	89	30	22	32	32

	000	001	010	011	012
012	0	0	0	1	102
011	1	0	0	71	1
010	1	0	80	0	12
001	0	84	0	0	9
000	121	9	5	2	1

J. R. Oaks, N. L'Bahy, and K. A. Cobb (2020). *Evolution* 74: 2184–2206

Simulation results

Simulations confirm theoretical expectations that changes in population size are difficult to estimate

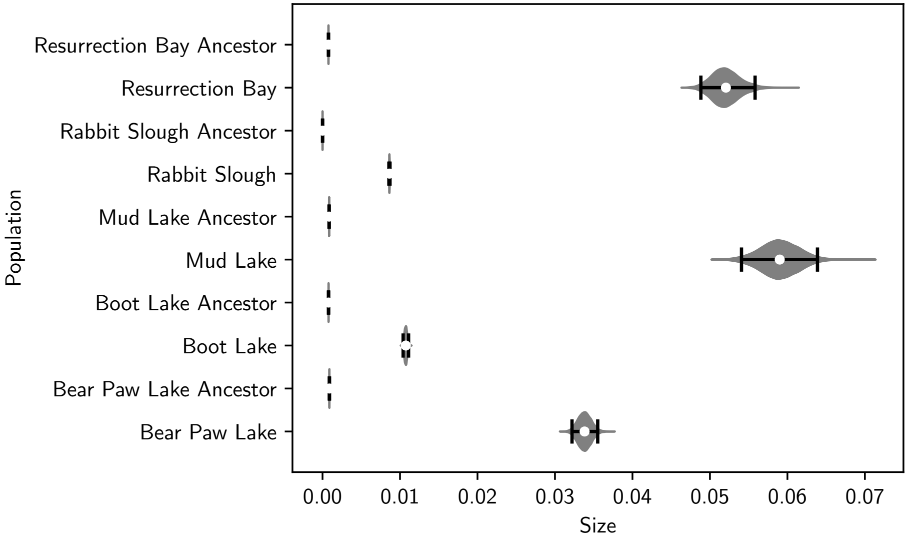


	000	001	010	011	012
012	25	30	17	26	62
011	5	2	4	19	21
010	5	6	14	1	15
001	7	27	2	3	4
000	89	30	22	32	32

	000	001	010	011	012
012	0	0	0	1	102
011	1	0	0	71	1
010	1	0	80	0	12
001	0	84	0	0	9
000	121	9	5	2	1

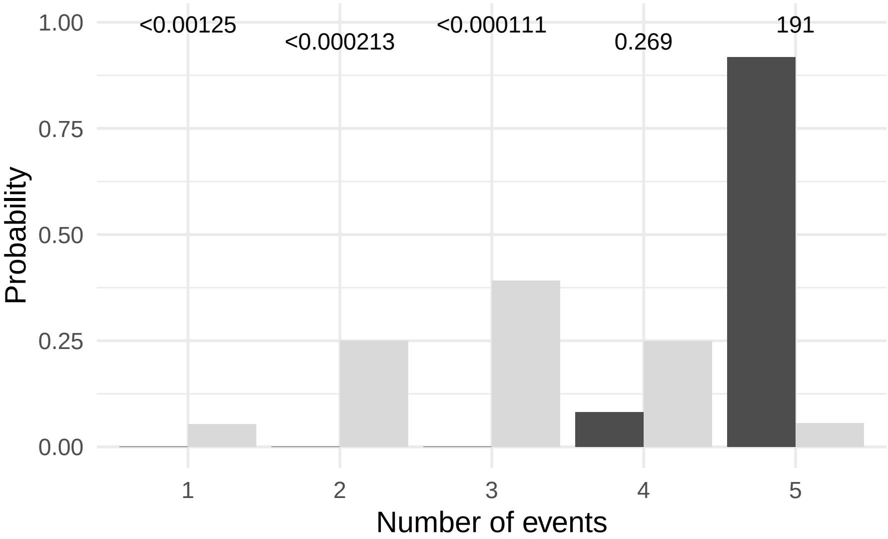
J. R. Oaks, N. L'Bahy, and K. A. Cobb (2020). *Evolution* 74: 2184–2206

Stickleback results



J. R. Oaks, N. L'Bahy, and K. A. Cobb (2020). *Evolution* 74: 2184–2206

Stickleback results



J. R. Oaks, N. L'Bahy, and K. A. Cobb (2020). *Evolution* 74: 2184–2206

Open science: everything is available. . .

Software:

- Ecoevolity: phyletica.org/ecoevolity

Open-Science Notebooks:
github.com/phyletica/ecoevolity-demog-experiments

Approaches to the problem

- A pairwise approach (keep it “simple”)
- An aside for a related problem with population demography
- A fully phylogenetic approach

The next 5 years in the Phyletica Lab

Biogeography

- Environmental changes that affect whole communities of species

Genome evolution

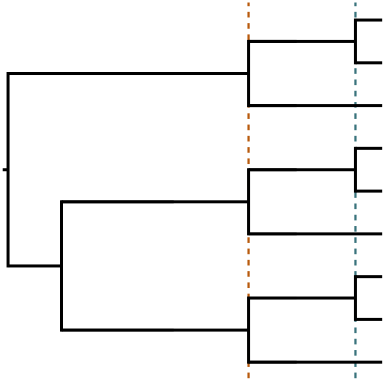
- Duplication of a chromosome segment harboring gene families

Epidemiology

- Transmission at social gatherings

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

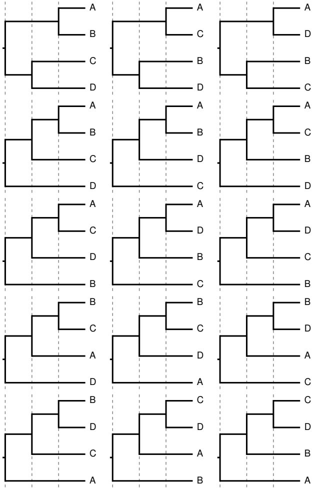
- Speciation of the host
- Co-colonization of new host species



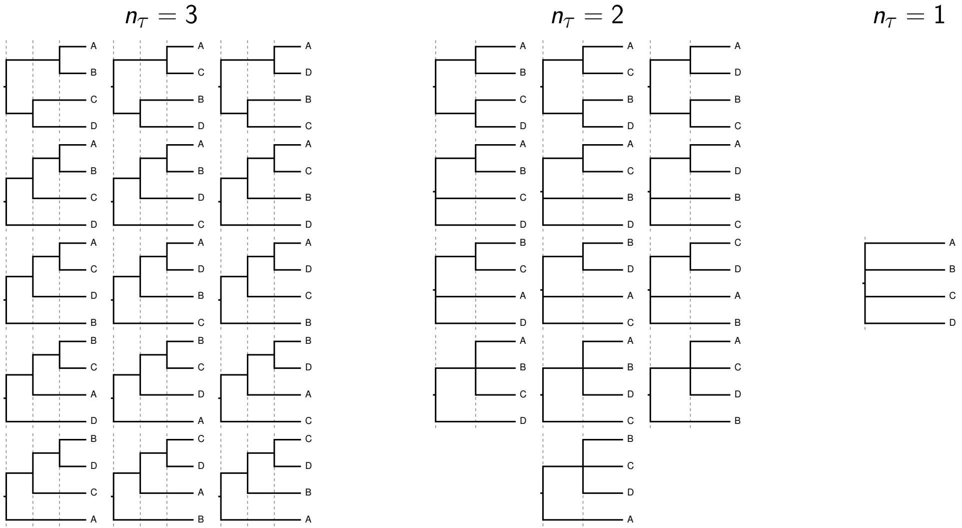
Generalizing tree space

Generalizing tree space

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

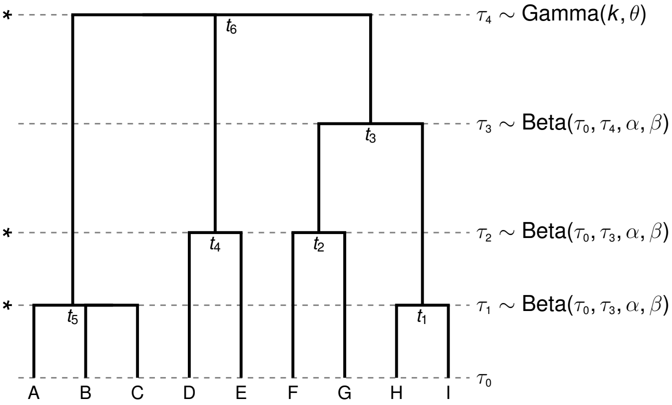


Generalizing tree space



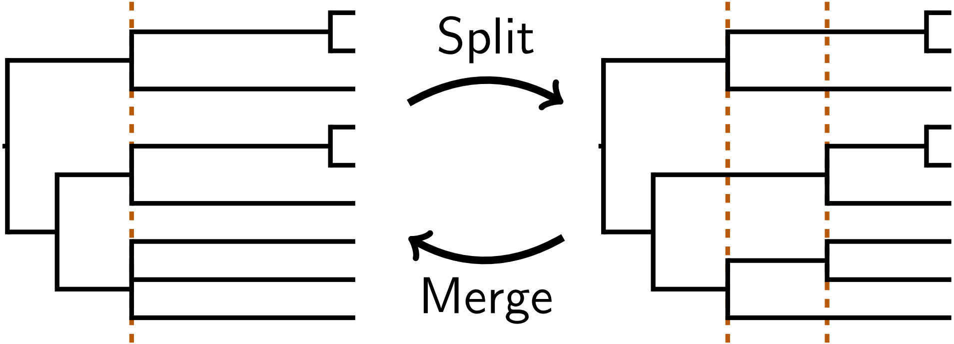
Generalized tree distribution

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



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

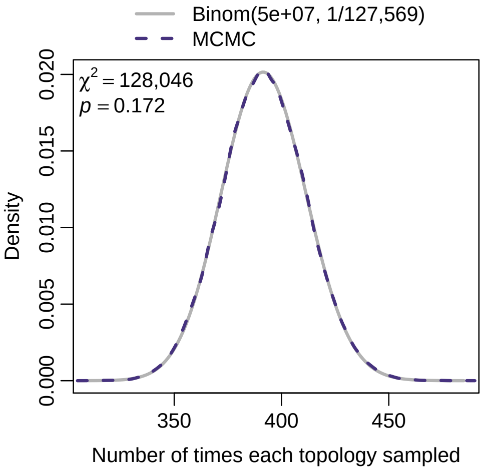
Inferring trees with shared divergences



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

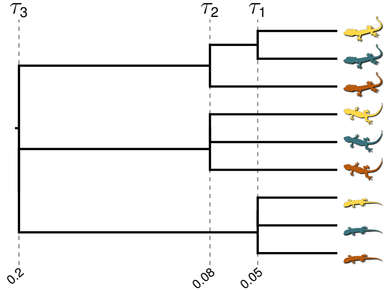
- **Tree model**
  - rjMCMC sampling of generalized tree distribution
- **Likelihood model**
  - CTMC model of characters evolving along genealogies
  - Infer species trees by analytically integrate over genealogies[1]
- *Goal: Co-estimation of phylogeny and shared divergences from genomic data*

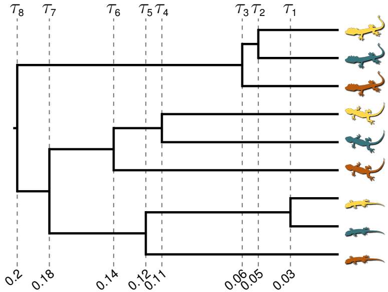
---

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

Methods: Simulations

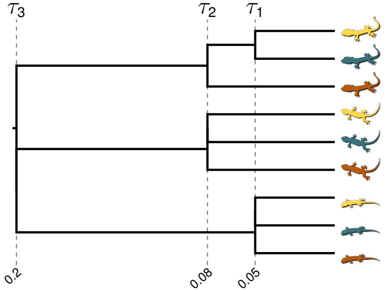
- Simulated 100 data sets with 50,000 characters

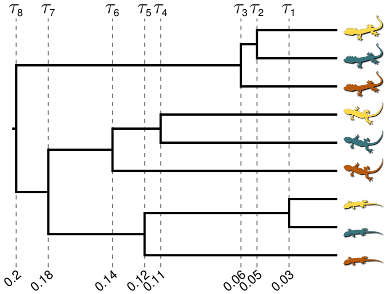




Methods: Simulations

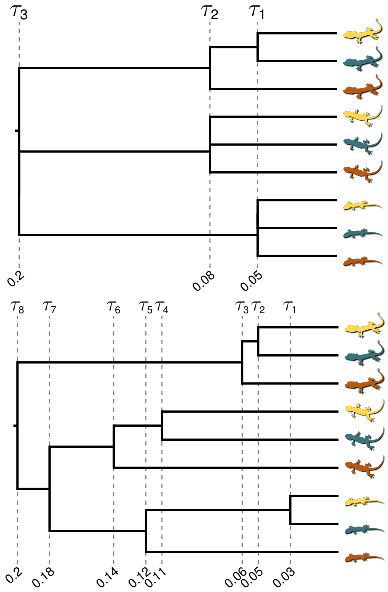
- Simulated 100 data sets with 50,000 characters
- Analyzed each data set with:
    - M_G = Generalized tree model
    - M_{IB} = Independent-bifurcating tree model

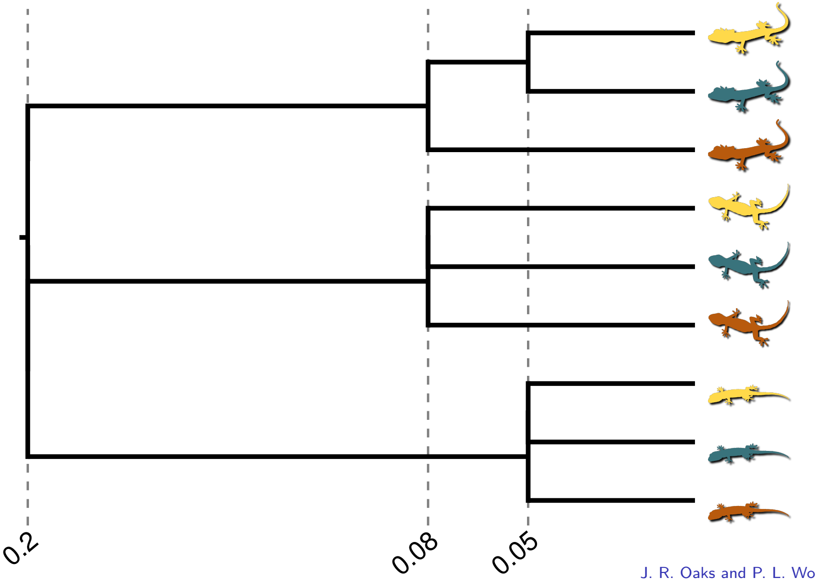




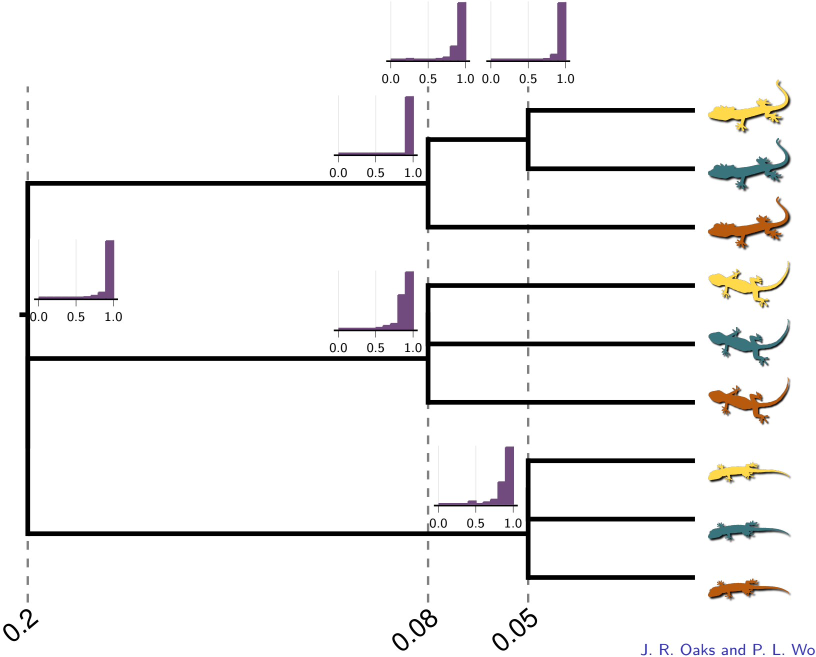
Methods: Simulations

- 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}

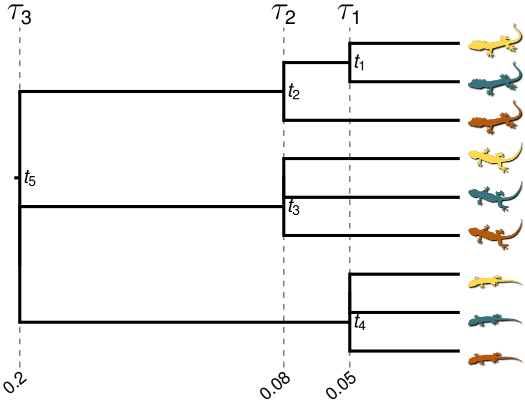




J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*

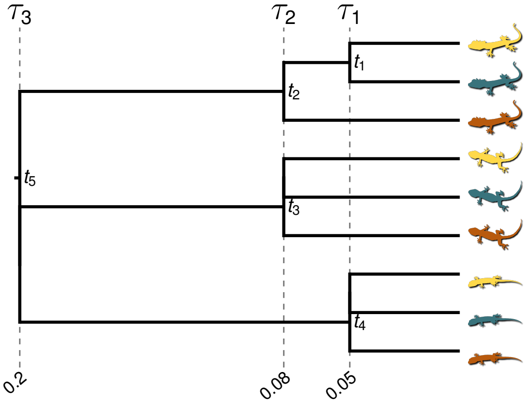


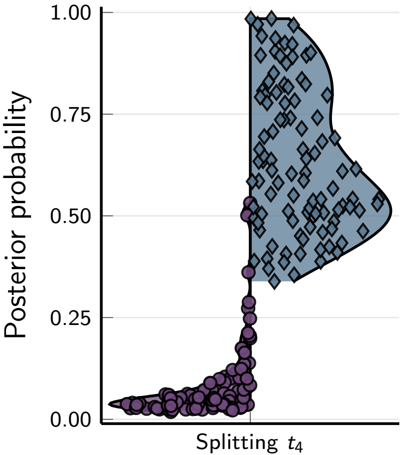
J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*



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

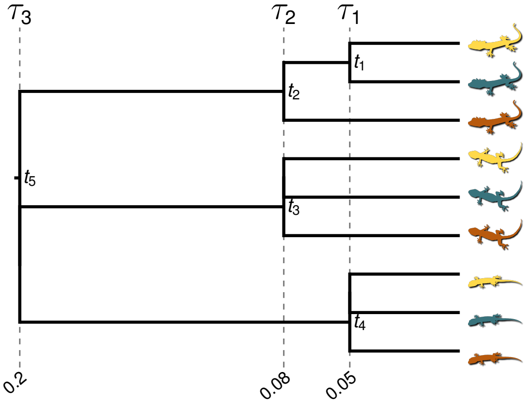
● M_G = Generalized model ◆ M_{IB} = Independent-bifurcating model

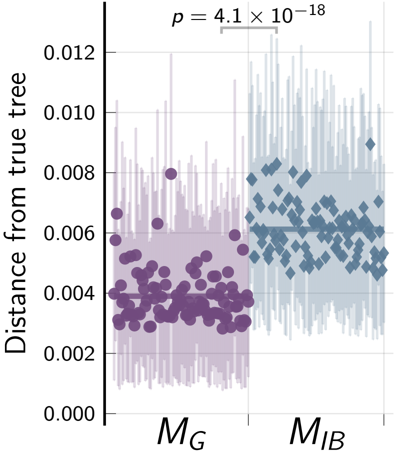




J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*

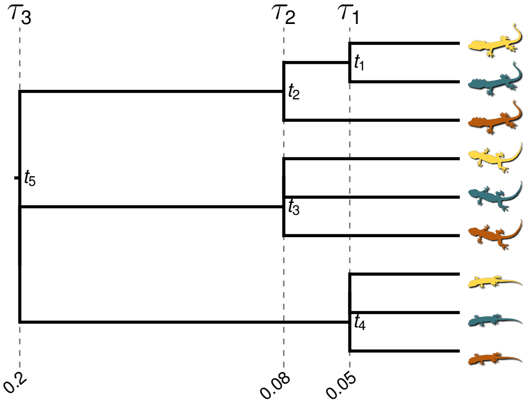
● M_G = Generalized model ◆ M_{IB} = Independent-bifurcating model

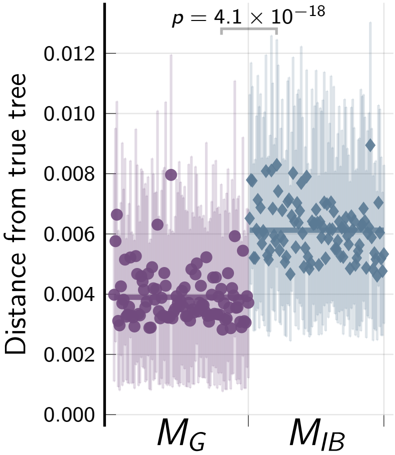




J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*

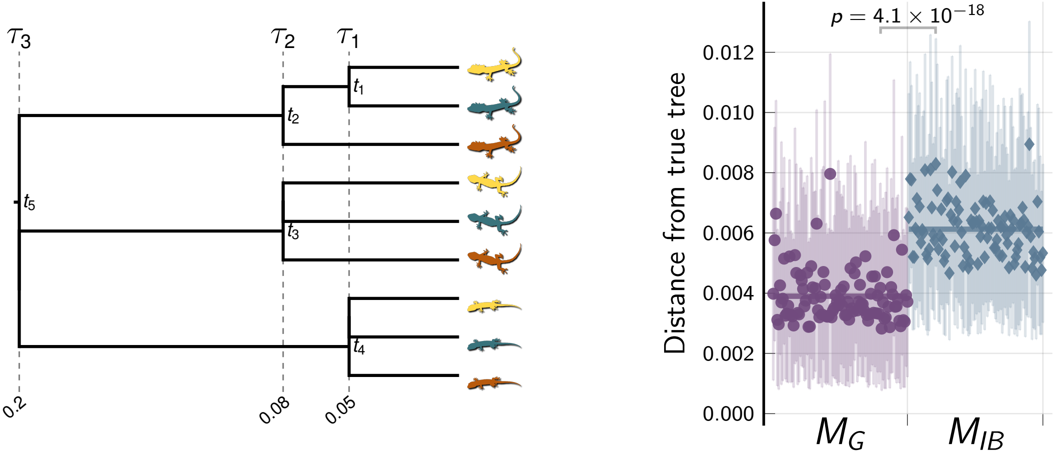
● M_G = Generalized model ◆ M_{IB} = Independent-bifurcating model





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

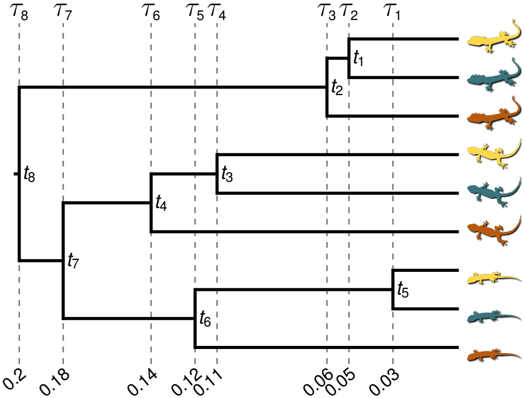
● M_G = Generalized model ◆ M_{IB} = Independent-bifurcating model



M_G significantly better at inferring trees with shared divergences

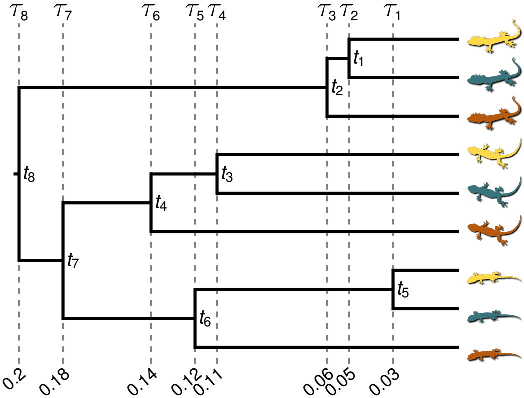
J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*

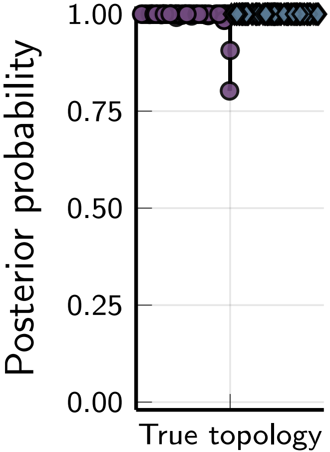
● M_G = Generalized model ◆ M_{IB} = Independent-bifurcating model



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

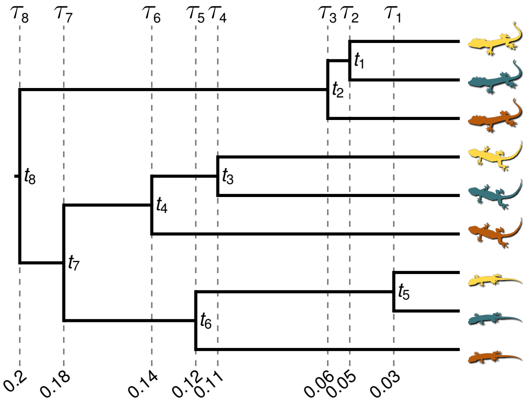
● M_G = Generalized model ◆ M_{IB} = Independent-bifurcating model

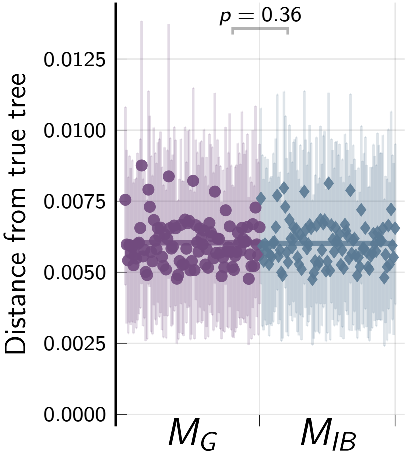




J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*

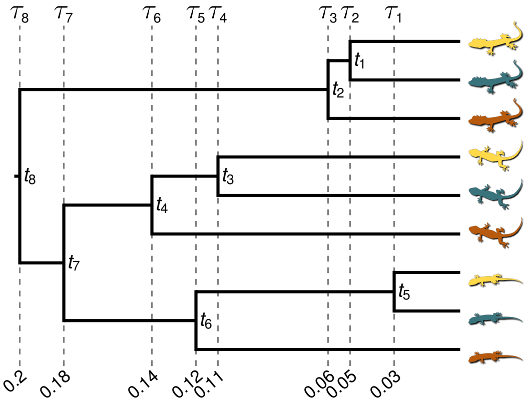
● M_G = Generalized model   ◆ M_{IB} = Independent-bifurcating model

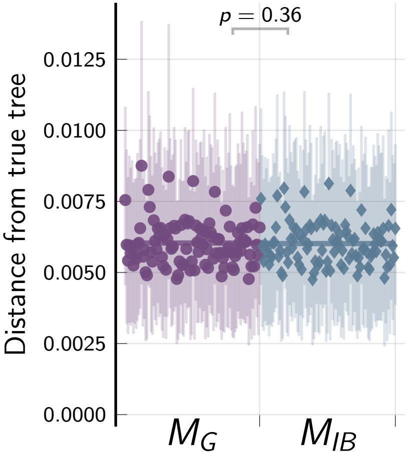




J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*

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

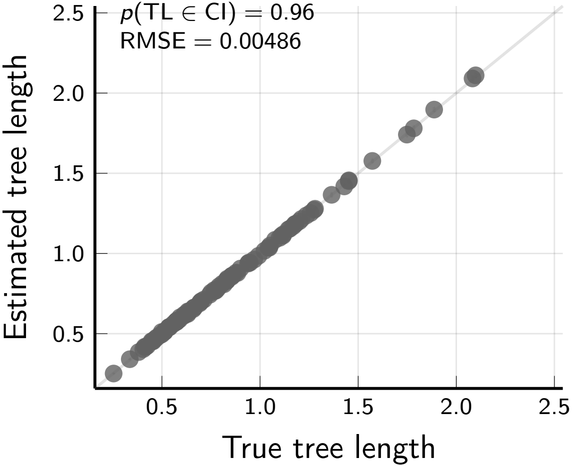




M_G performs as well as true model when divergences are independent

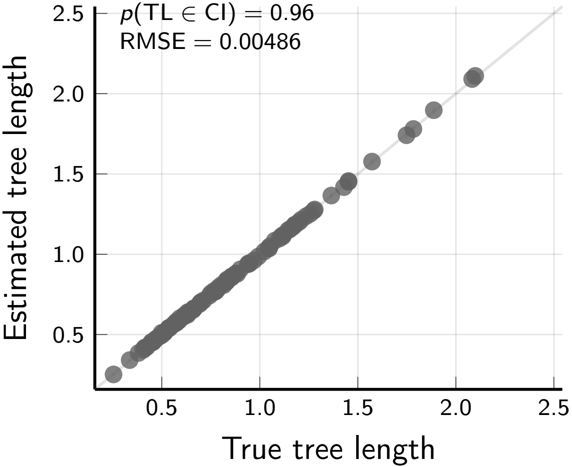
J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*

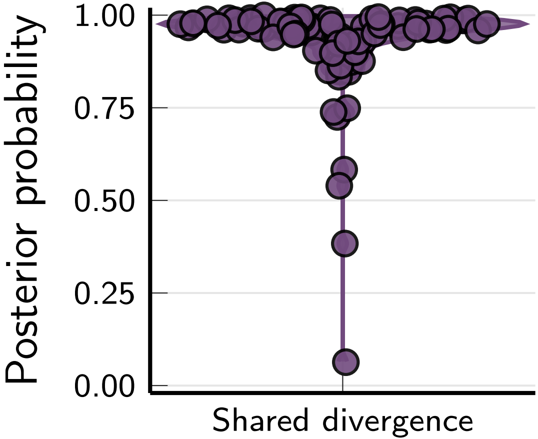
Results: random M_G trees



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

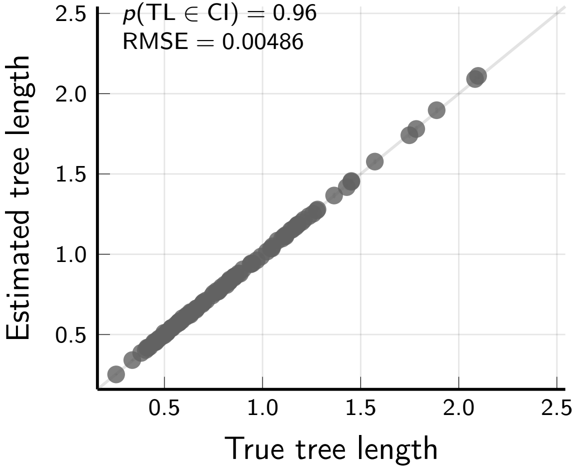
Results: random M_G trees

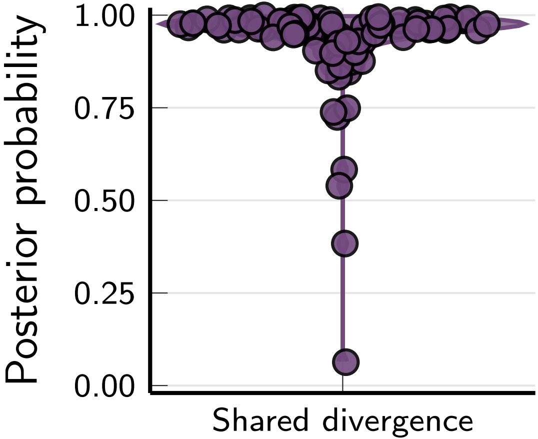




J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*

Results: random M_G trees

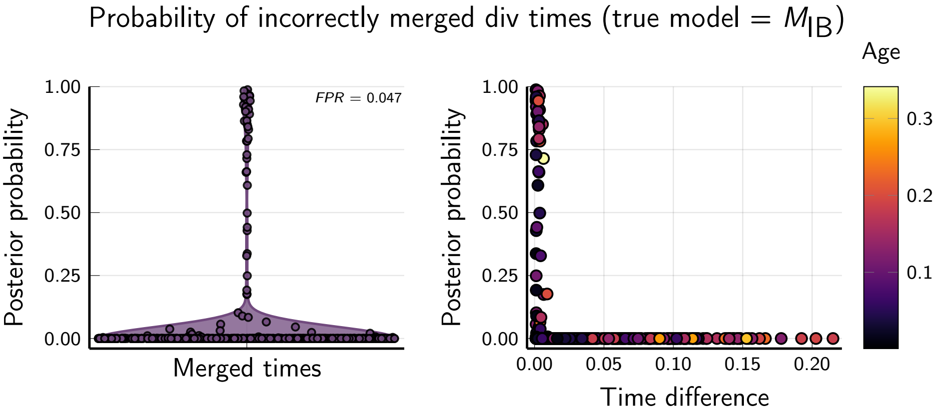




M_G performs well with data simulated on random trees with shared divergences

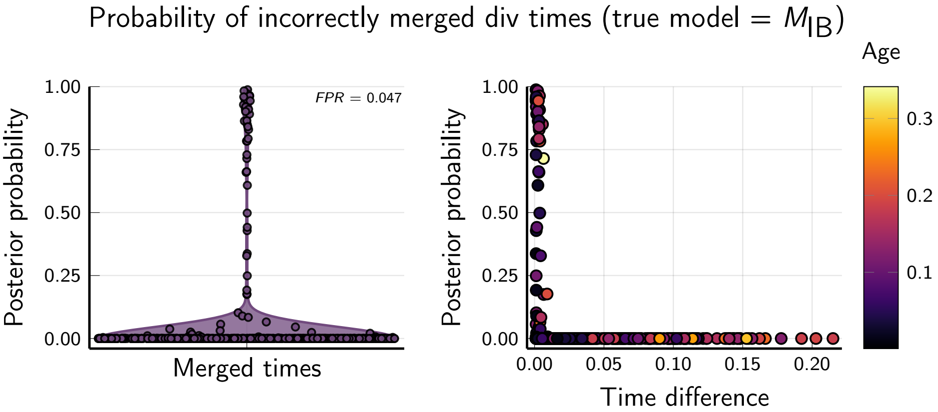
J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*

Results: random M_{IB} trees



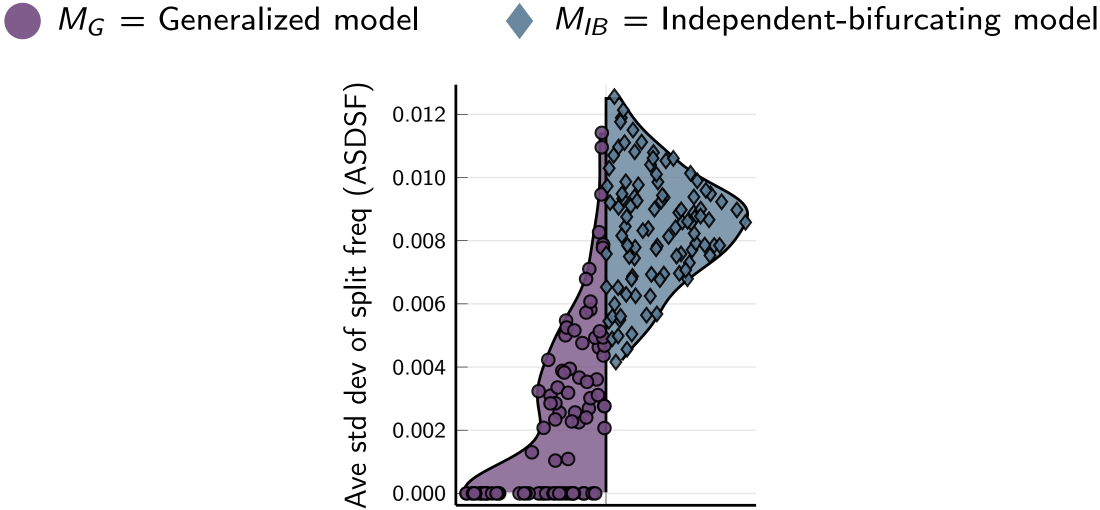
J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*

Results: random M_{IB} trees



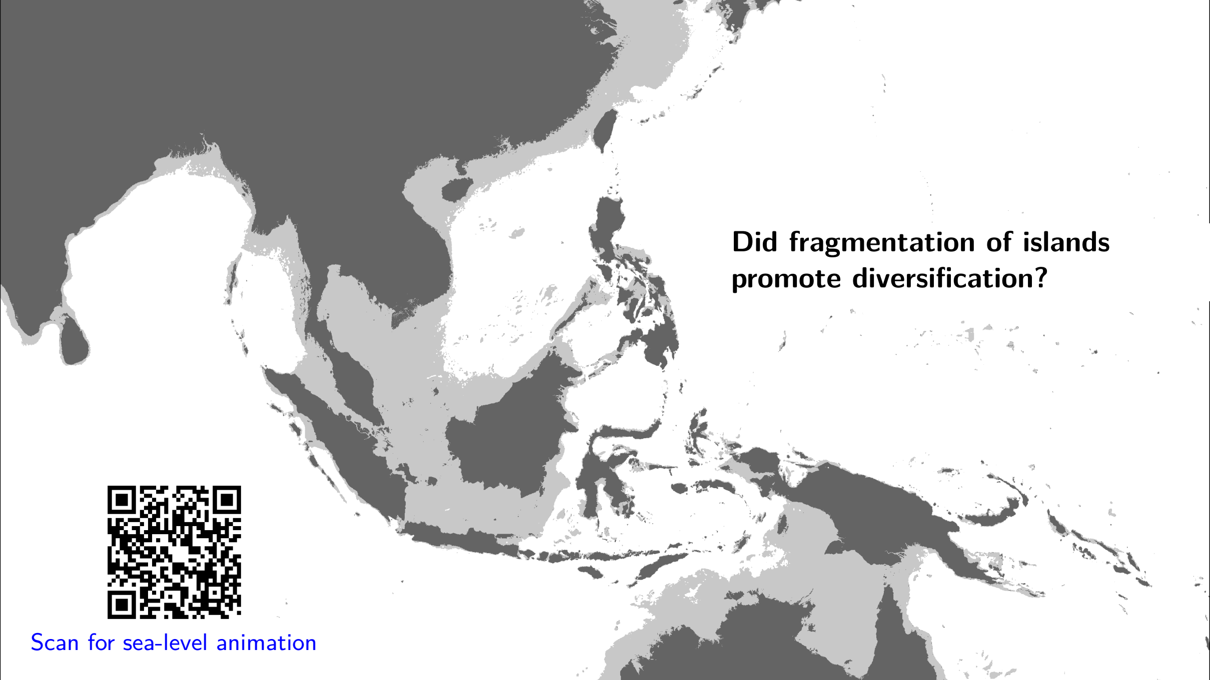
M_G has low false positive rate

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*

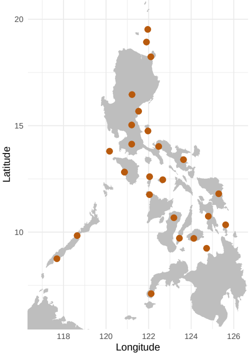


Did fragmentation of islands promote diversification?

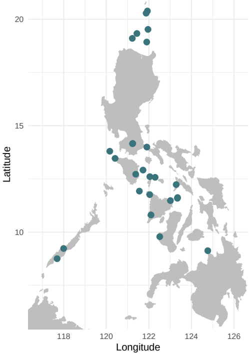
Scan for sea-level animation

Cyrtodactylus

©Rafe M. Brown



Gekko

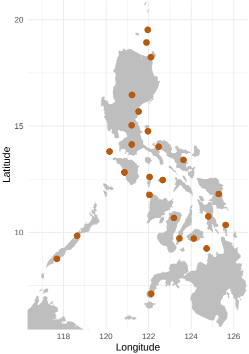


©Rafe M. Brown

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

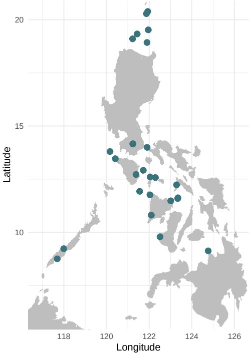
Cyrtodactylus

©Rafe M. Brown



1702 loci
155,887 sites

Gekko

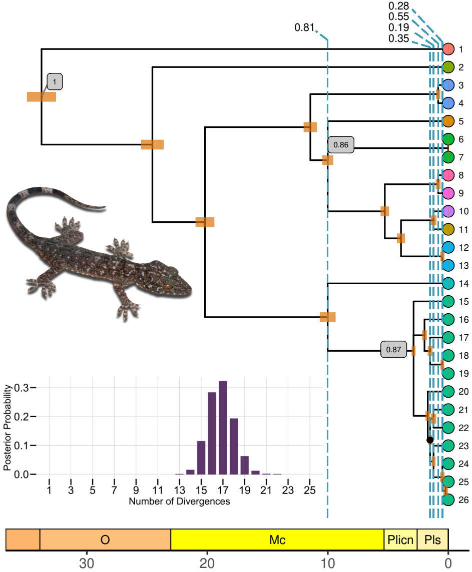


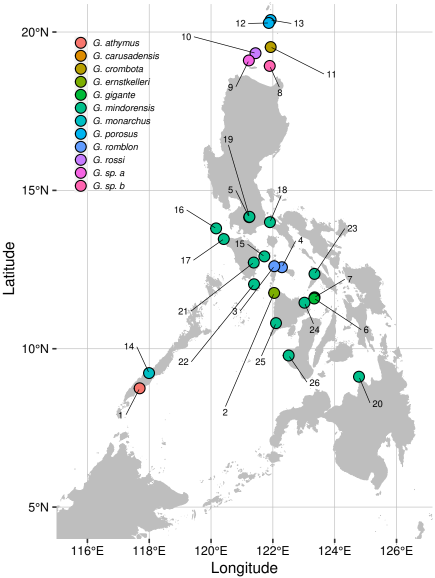
©Rafe M. Brown

1033 loci
94,813 sites

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

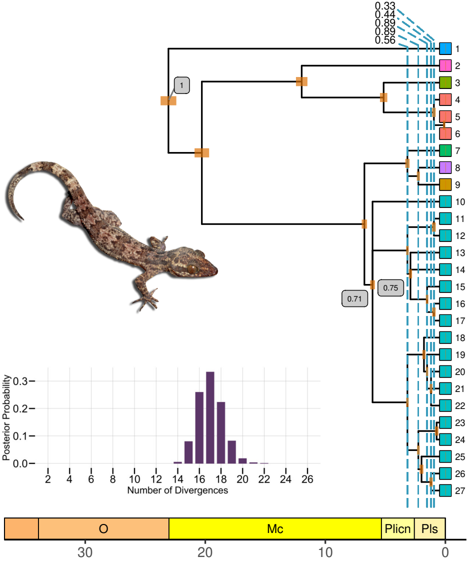
Gekko

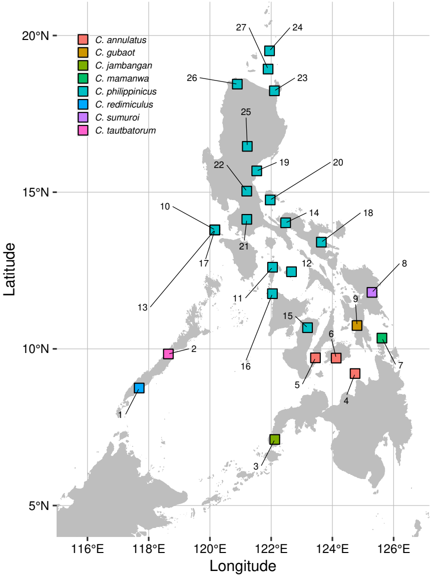




J. R. Oaks and P. L. Wood, Jr. (2021). *bioRxiv*

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

Open-Science Notebooks:

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

Approaches to the problem

- A pairwise approach (keep it “simple”)
- An aside for a related problem with population demography
- A fully phylogenetic approach

The next 5 years in the Phyletica Lab

Looking forward

Phylogenetic theory/methods

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

Empirical work

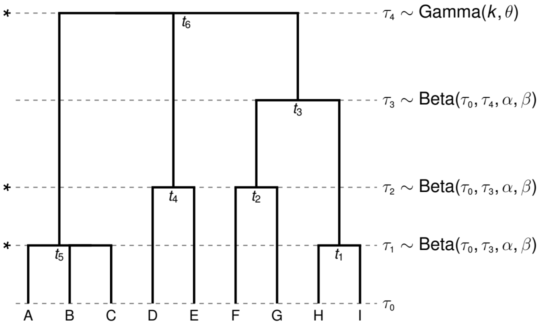
- Did the evolution of habitat preference affect the diversification of bent-toed geckos?
- Epidemiological dynamics of "super-spreading" events during the COVID-19 pandemic

Prison education

- Coding to learn evolution

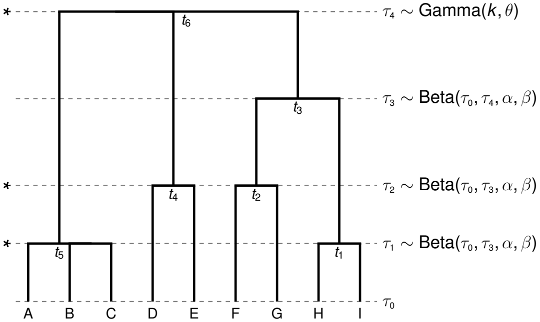
Generalized tree distribution

- Our current distribution over trees is motivated by mathematical convenience



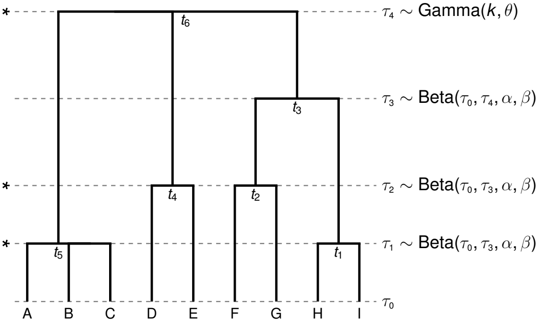
Generalized tree distribution

- 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



Generalized tree distribution

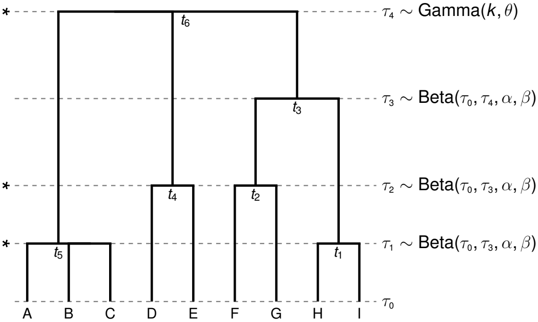
- Our current distribution over trees is motivated by mathematical convenience
- A process-based distribution would allow us to learn about parameters that control diversification processes
- Goal: port M_G algorithms to `RevBayes` and develop generalized birth-death model



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

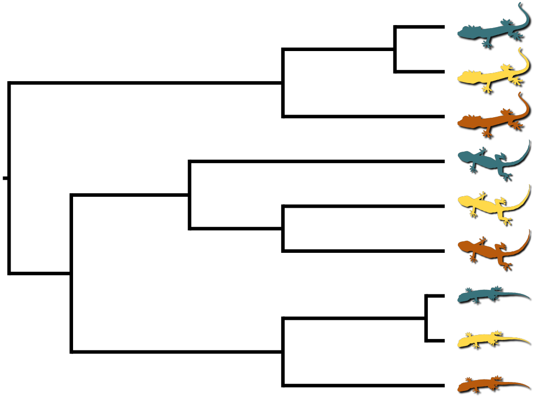


Generalizing the birth-death process

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



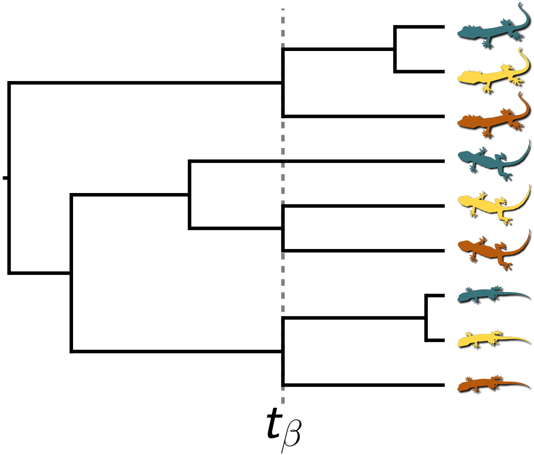
Generalizing the birth-death process

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



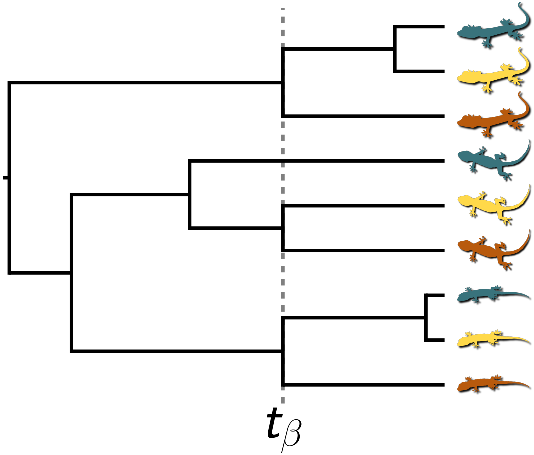
Generalizing the birth-death process

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



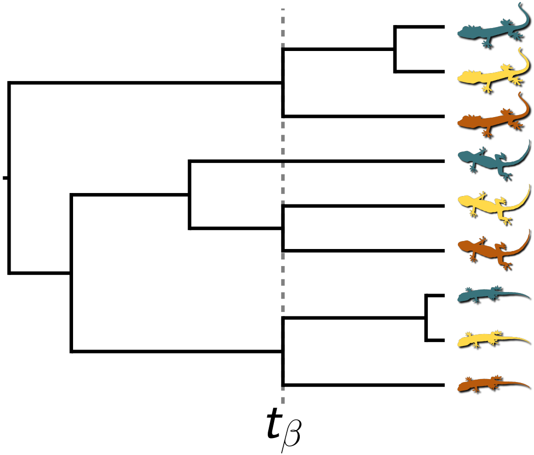
Generalizing the birth-death process

Birth-death basics:

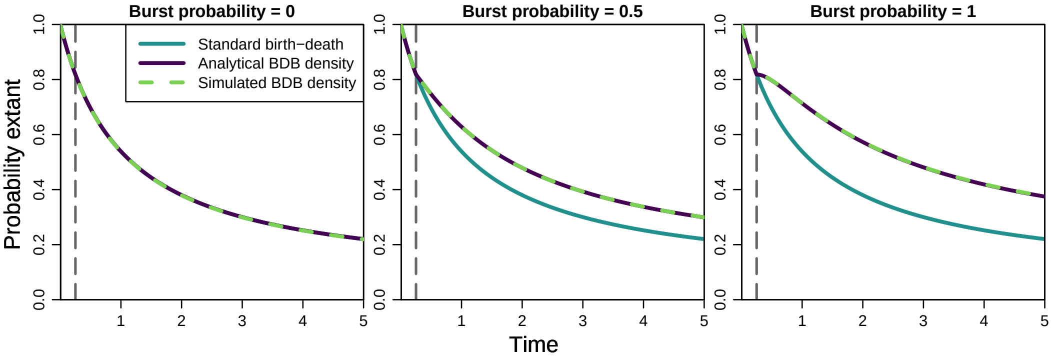
- Lineages speciate at rate λ
- Lineages go extinct at rate μ
- We sample extant lineages with probability ρ

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

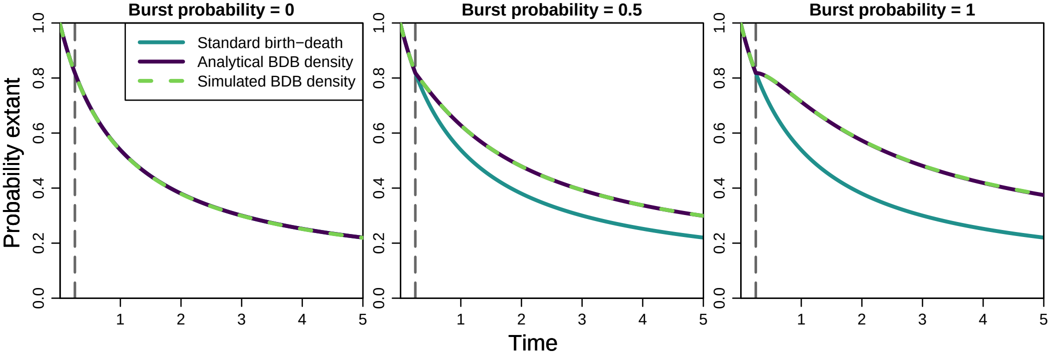
- Include "burst events" that occur at rate λ_β
- Each lineage diverges with probability β
- Allow λ, μ, λ_β, & β to vary depending on the traits of lineages across the tree
- Bayesian model-averaging to infer set of trait-dependent BDB models that best explain data



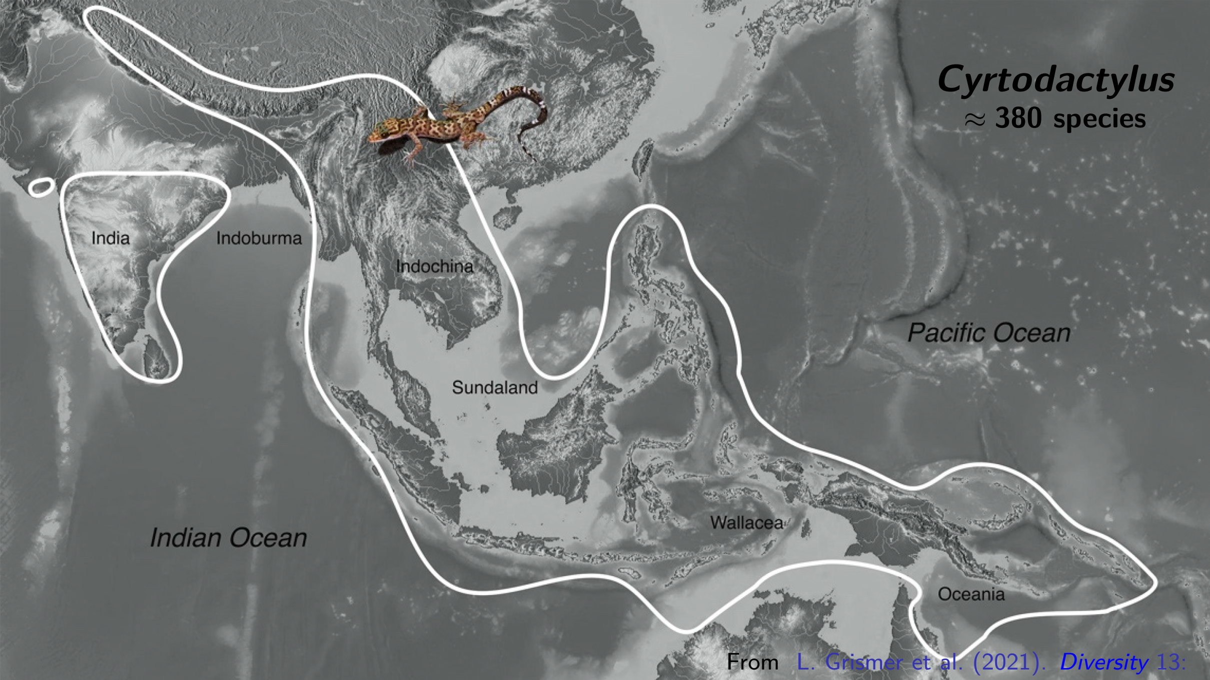
Birth-death-burst validation



Birth-death-burst validation



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



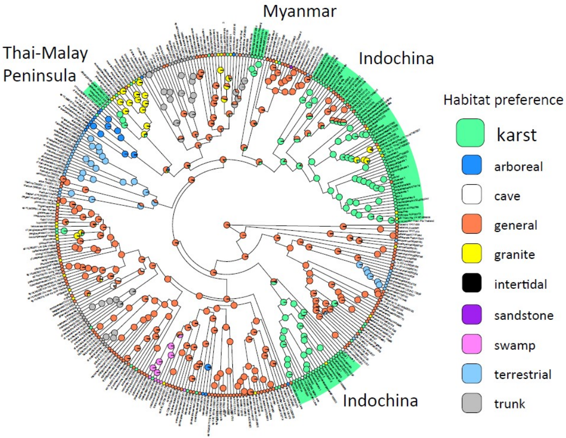
Cyrtodactylus

$\approx$ 380 species

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

Karst endemism in *Cyrtodactylus*

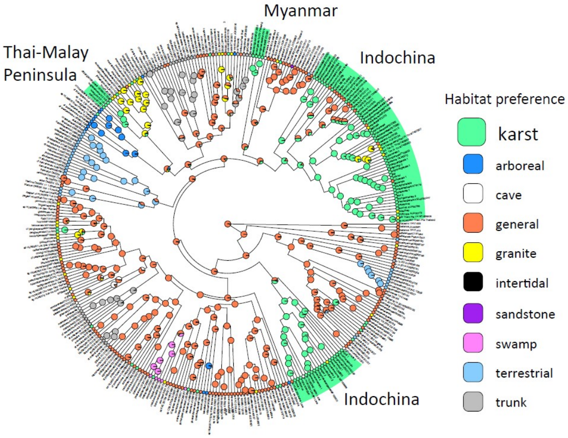
- *Cyrtodactylus* are ecologically diverse, ranging from generalists to microhabitat specialists



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

Karst endemism in *Cyrtodactylus*

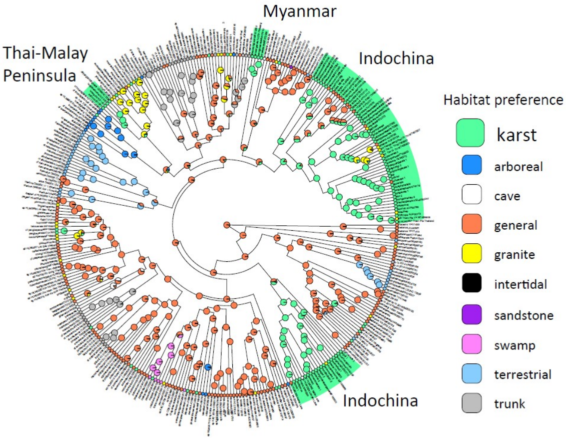
- *Cyrtodactylus* are ecologically diverse, ranging from generalists to microhabitat specialists
- Karst-specificity evolved 24 times



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

Karst endemism in *Cyrtodactylus*

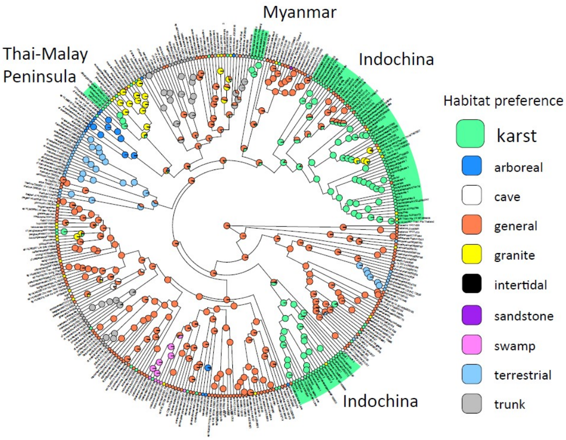
- *Cyrtodactylus* are ecologically diverse, ranging from generalists to microhabitat specialists
- Karst-specificity evolved 24 times
- Comprise 25% of species despite tiny fraction of landscape being karst



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

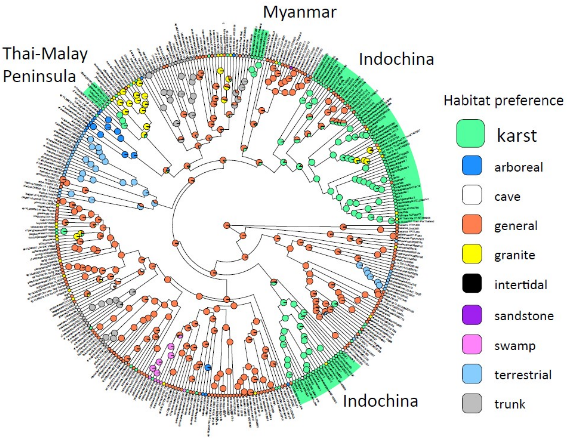


L. Grismer et al. (2021). *Diversity* 13:

©Perry Wood, Jr.

Why high levels of diversity and endemism on karst?

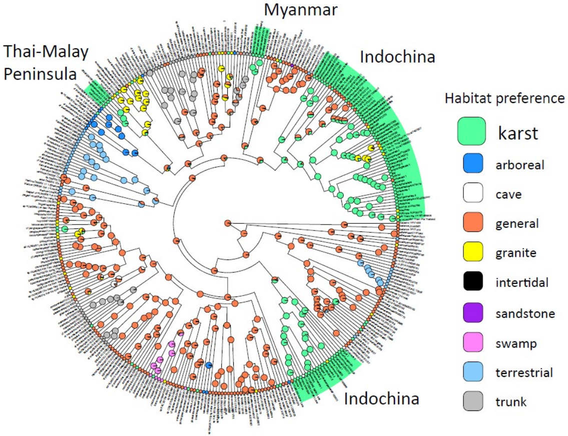
- “Rapid” fragmentation of karst habitat caused by the uplift and subsequent erosion of limestone sediment over the last 30my



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

Why high levels of diversity and endemism on karst?

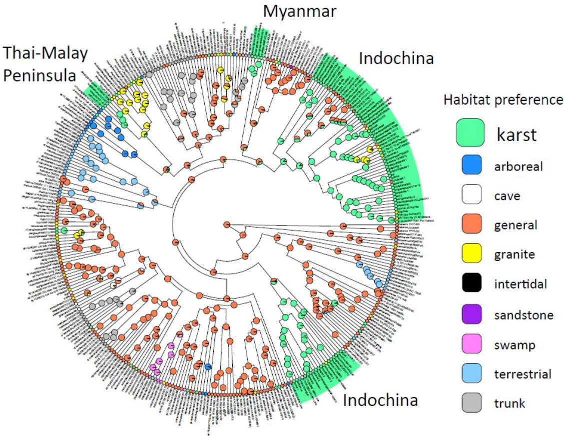
- “Rapid” fragmentation of karst habitat caused by the uplift and subsequent erosion of limestone sediment over the last 30my
- *E.g.*, Major river systems carved through and isolated limestone karst formations (Ayeyarwady, Chiang Mai, Mekong, Red, and Salween)



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*

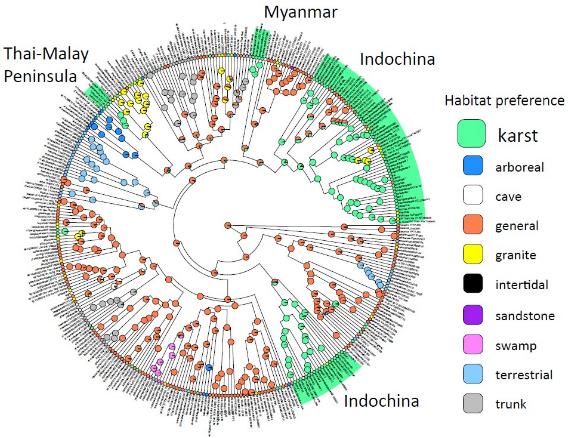


L. Grismer et al. (2021). *Diversity* 13:

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

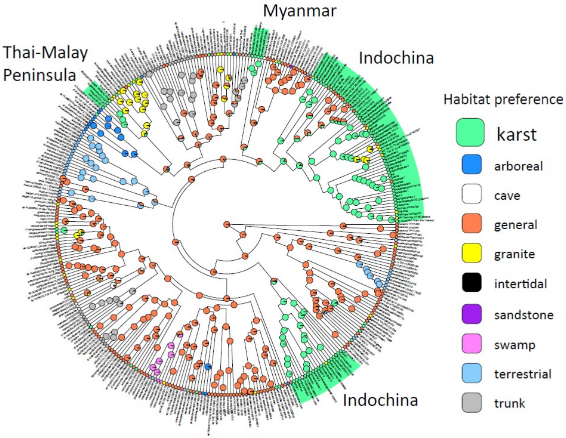


L. Grismer et al. (2021). *Diversity* 13:

Why high levels of diversity and endemism on karst?

Plan:

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

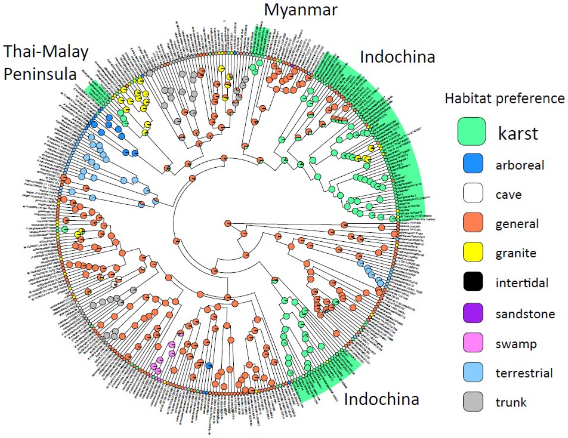


L. Grismer et al. (2021). *Diversity* 13:

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

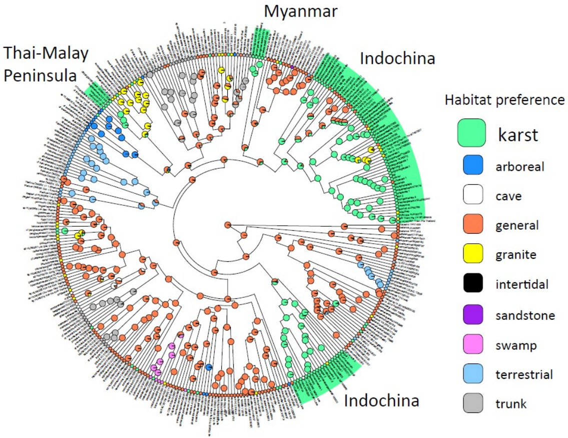


L. Grismer et al. (2021). *Diversity* 13:

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

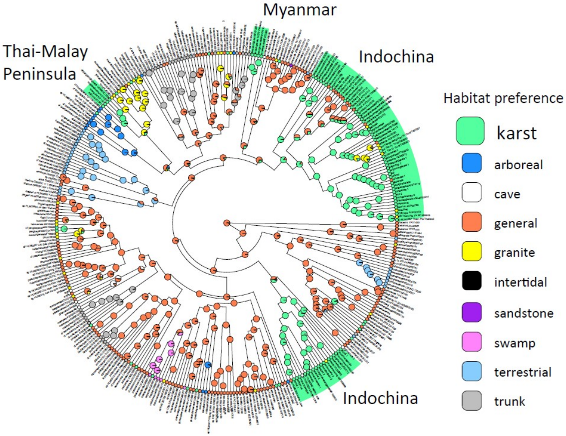


L. Grismer et al. (2021). *Diversity* 13:

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)

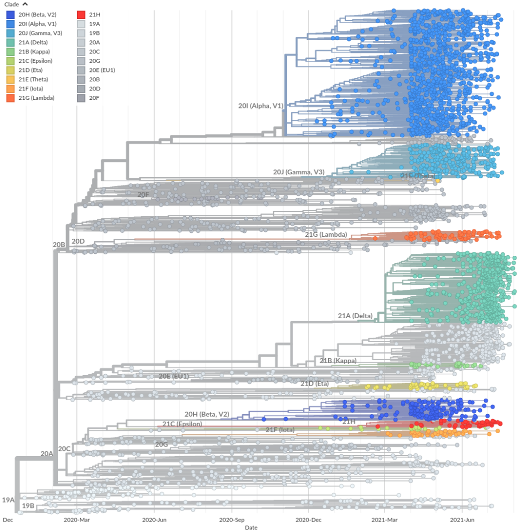


L. Grismer et al. (2021). *Diversity* 13:

Epidemiological dynamics of COVID-19 pandemic

Questions:

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

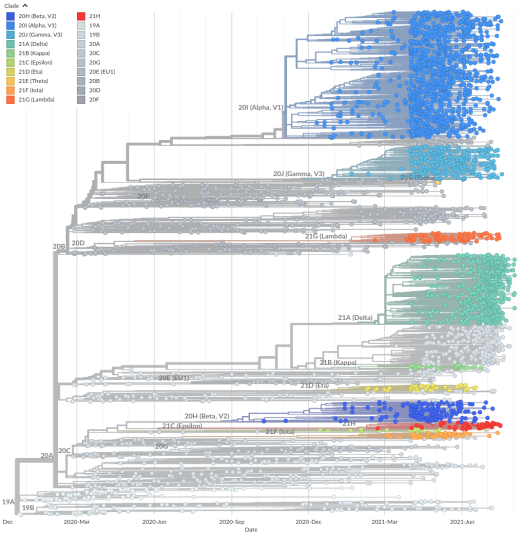


nextstrain.org J. Hadfield et al. (2018). *Bioinformatics* 34: 4121–4123

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?

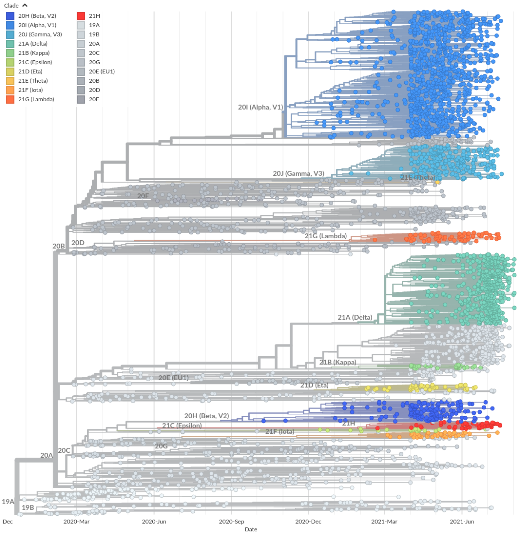


nextstrain.org J. Hadfield et al. (2018). *Bioinformatics* 34: 4121–4123

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?

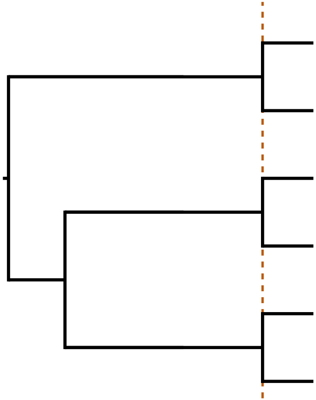


nextstrain.org J. Hadfield et al. (2018). *Bioinformatics* 34: 4121–4123

Divergence patterns predicted by gatherings

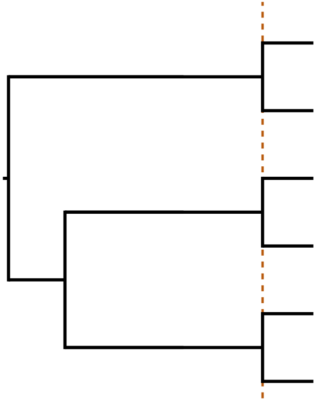
Divergence patterns predicted by gatherings

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



Divergence patterns predicted by gatherings

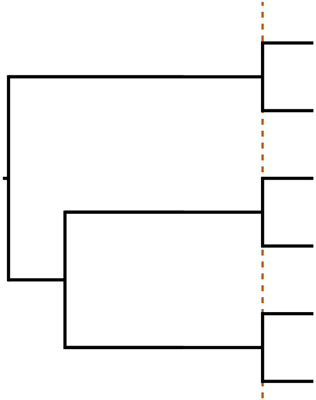
- Multiple infected people spreading SARS-CoV-2 at a gathering will create shared divergences across “transmission tree”
- ***Shared divergences*** are a good proxy for spread at gatherings



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:

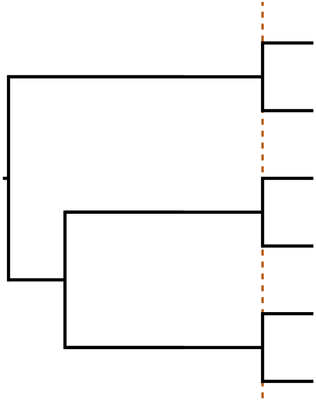


Divergence patterns predicted by gatherings

- 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

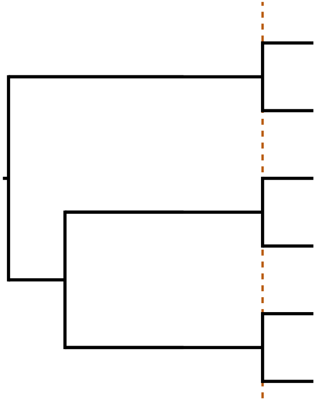


Divergence patterns predicted by gatherings

- 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)

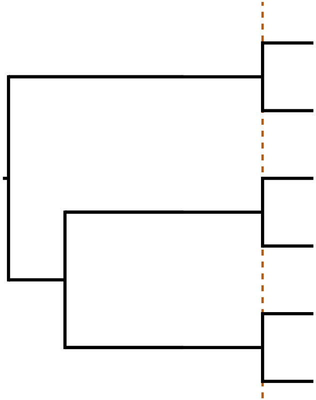


Divergence patterns predicted by gatherings

- 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

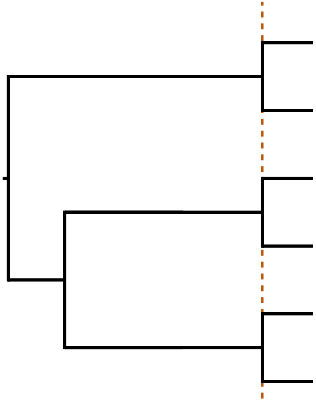


Divergence patterns predicted by gatherings

- 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

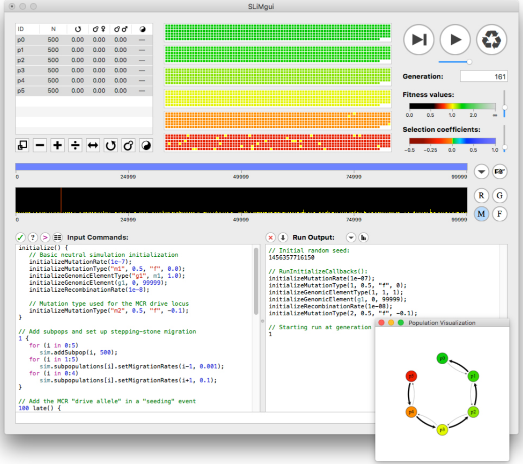
Plan:

- 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



Prison Ed: Coding to learn evolution

- Develop coding-to-learn evolution course (APAEP)



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() {
```

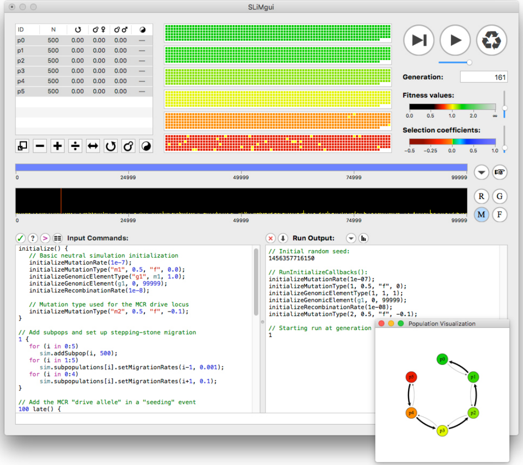
Run Output:

```
// 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

Prison Ed: Coding to learn evolution

- Develop coding-to-learn evolution course (APAEP)
- 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() {
```

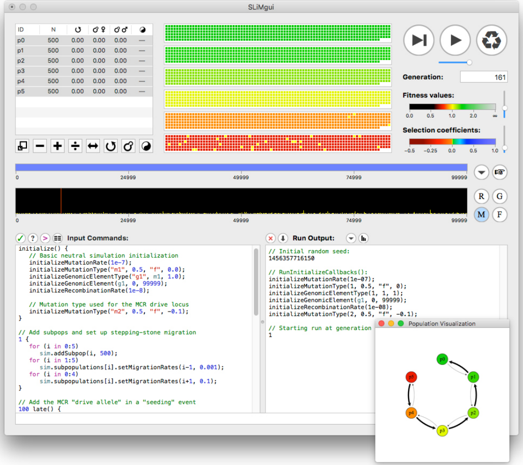
Run Output:

```
// 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

Prison Ed: Coding to learn evolution

- Develop coding-to-learn evolution course (APAEP)
- 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() {
```

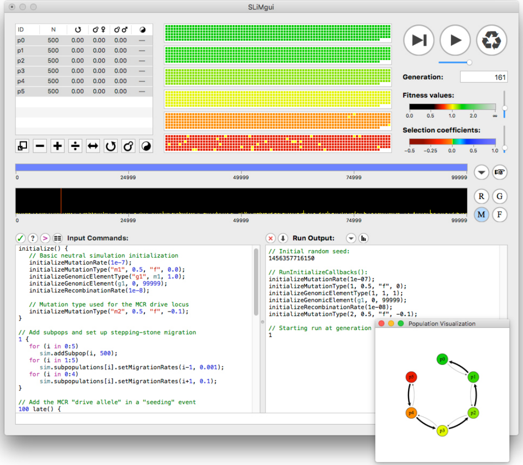
Run Output:

```
// 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
```

messerlab.org/slim

Prison Ed: Coding to learn evolution

- Develop coding-to-learn evolution course (APAEP)
- 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	—

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:

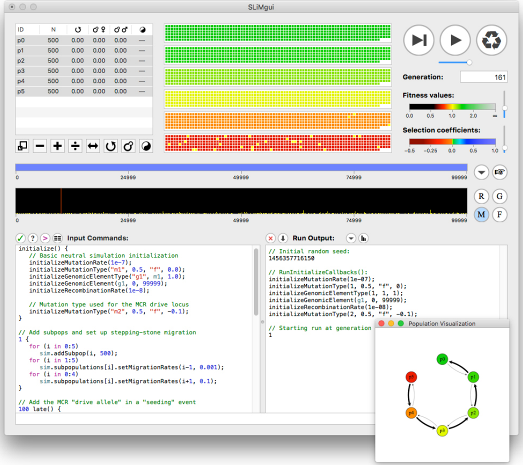
```
// Initial random seed:
1456357716150

// 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

Prison Ed: Coding to learn evolution

- Develop coding-to-learn evolution course (APAEP)
- 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	—

Generation: 161

```
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() {
```

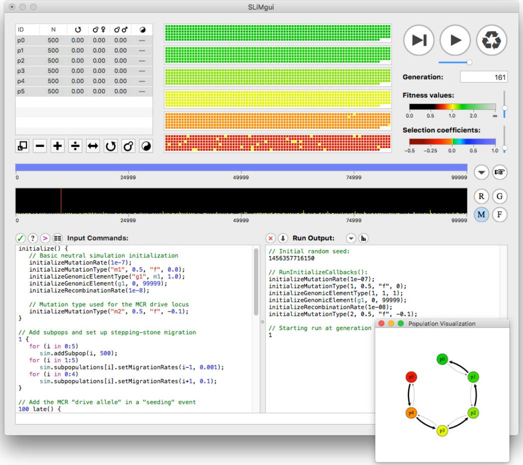
```
// Initial random seed:
1456357716150

// 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

Prison Ed: Coding to learn evolution

- Develop coding-to-learn evolution course (APAEP)
- 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

```
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

Acknowledgments

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

Computation:

- Alabama Supercomputer Authority
- Auburn University Hopper Cluster

Funding:

NSF

AUBURN
UNIVERSITY

Photo credits:

- Rafe Brown
- Perry Wood, Jr.
- PhyloPic

Acknowledgments

Questions?

joaks@auburn.edu

phyletica.org

© 2007 Boris Kulikov boris-kulikov.blogspot.com