

Bayesian statistical
approaches to inferring
shared evolutionary events

Jamie Oaks

Auburn University

phyletica.org

[@jamoaks](https://twitter.com/jamoaks)

Scan for slides:



phyletica.org/slides/adaptive.pdf



© 2007 Boris Kulikov boris-kulikov.blogspot.com

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
- ▶ *Breanna Siple*
- ▶ *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 Bayesian
phylogenetics to infer patterns
predicted by processes of
diversification



- Phylogenetics is rapidly becoming the statistical foundation of biology

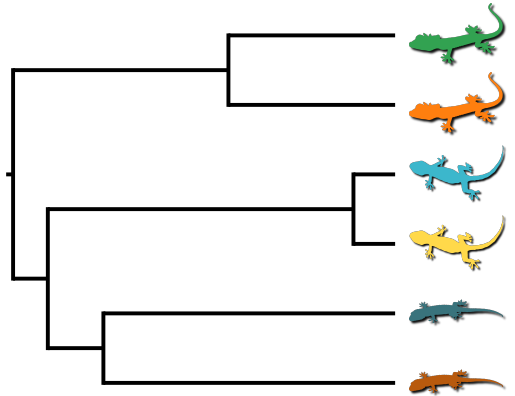


- ▶ Phylogenetics is rapidly becoming the statistical foundation of biology
- ▶ “Big data” present exciting possibilities and challenges

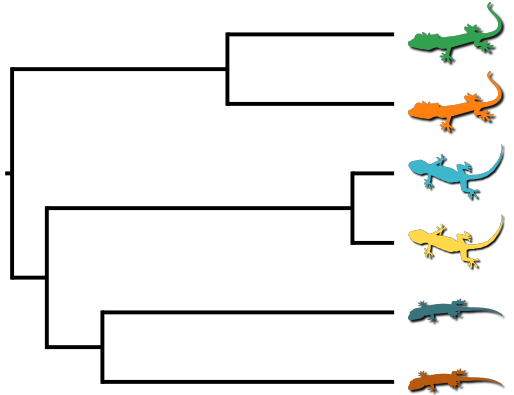


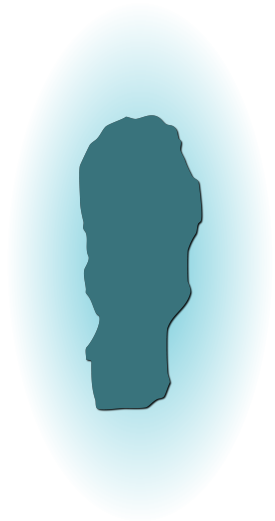
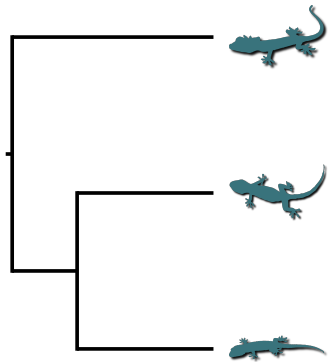
- ▶ Phylogenetics is rapidly becoming the statistical foundation of biology
- ▶ “Big data” present exciting possibilities and challenges
- ▶ Many opportunities to develop new ways to study biology in light of phylogeny

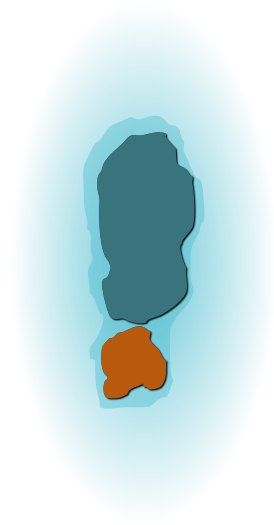
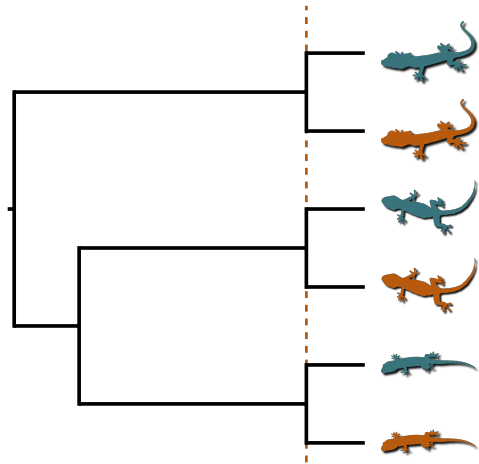




- **Assumption:** All processes of diversification affect each lineage independently

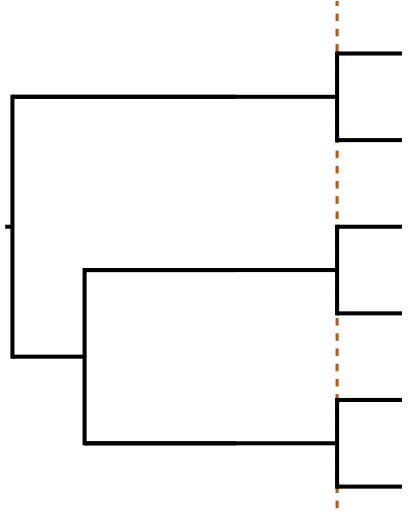






Biogeography

- ▶ Environmental changes that affect whole communities of species

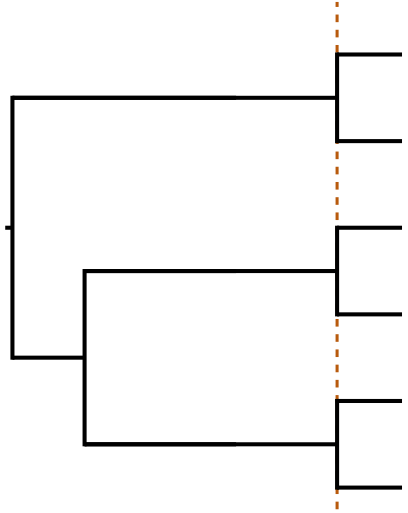


Biogeography

- ▶ Environmental changes that affect whole communities of species

Genome evolution

- ▶ Duplication of a chromosome segment harboring gene families



Biogeography

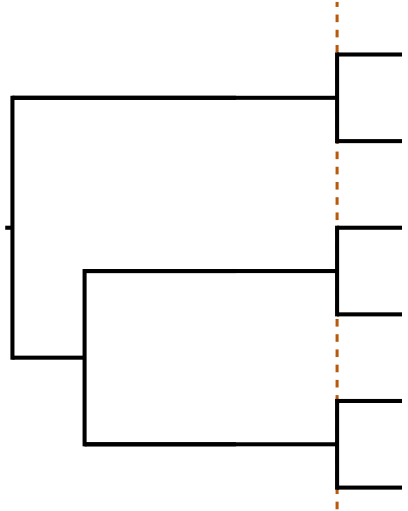
- ▶ Environmental changes that affect whole communities of species

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

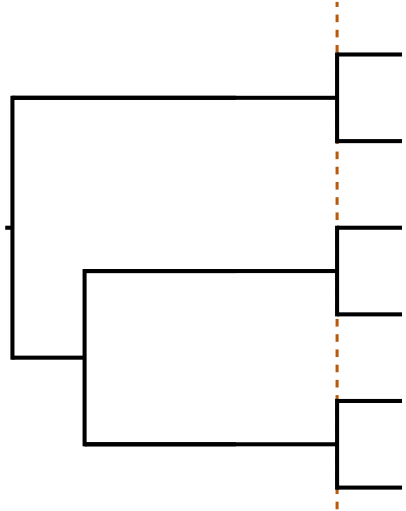
- ▶ Duplication of a chromosome segment harboring gene families

Epidemiology

- ▶ Transmission at social gatherings

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

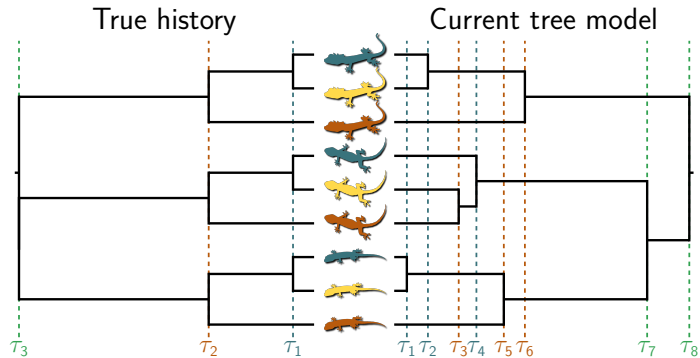
- ▶ Speciation of the host
- ▶ Co-colonization of new host species



Why account for shared divergences?

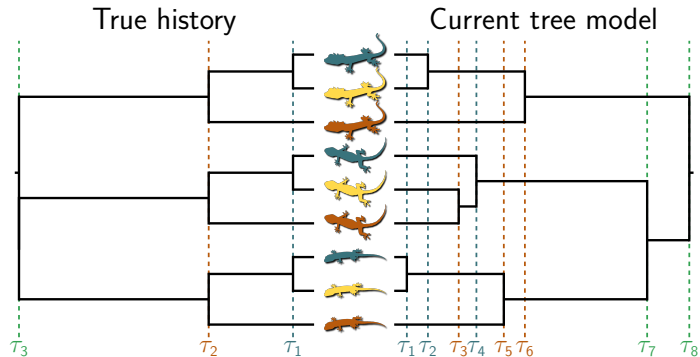
Why account for shared divergences?

1. Improve inference



Why account for shared divergences?

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



Biogeography

- ▶ Environmental changes that affect whole communities of species

Genome evolution

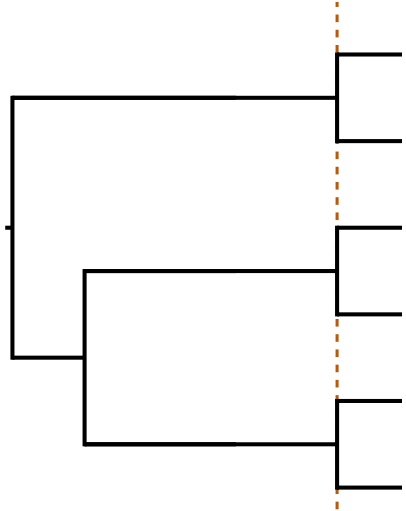
- ▶ Duplication of a chromosome segment harboring gene families

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”)

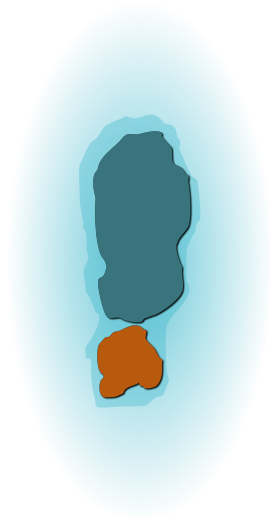
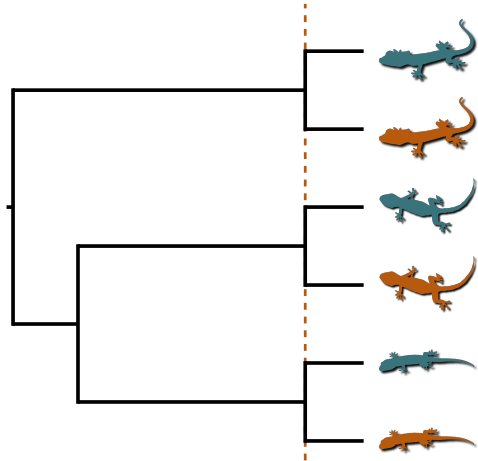
- A fully phylogenetic approach

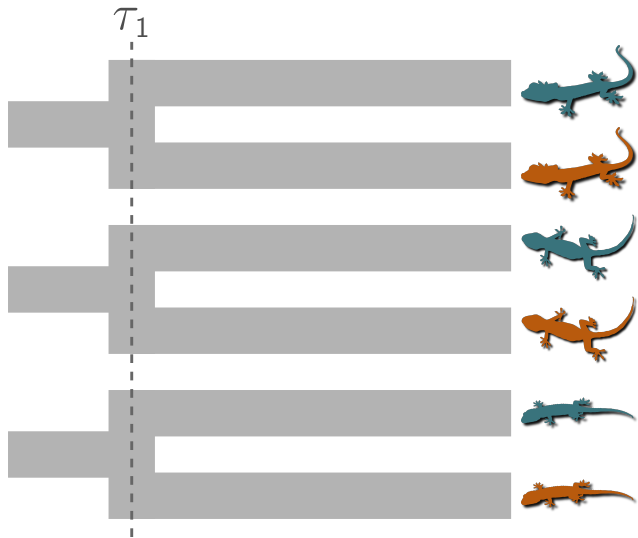
Approaches to the problem

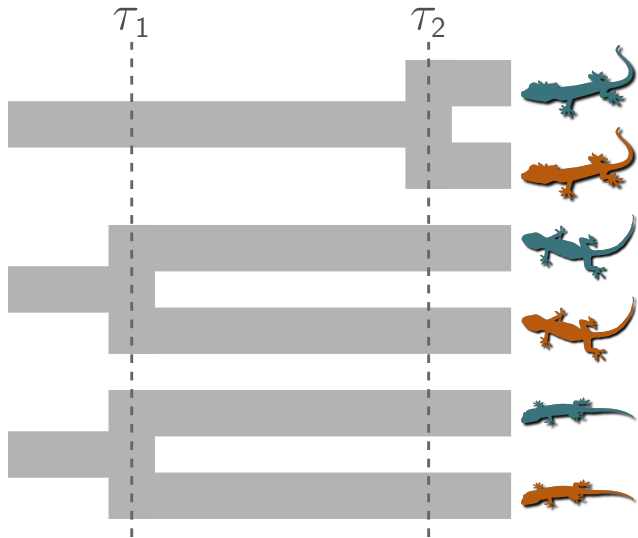
A pairwise approach (keep it “simple”)

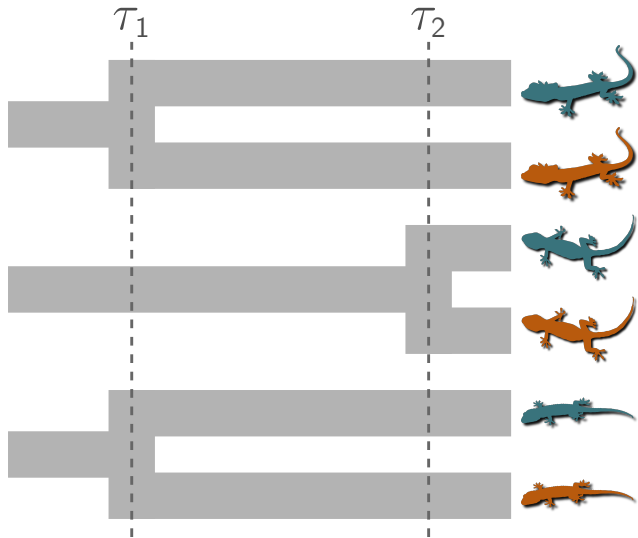
A fully phylogenetic approach

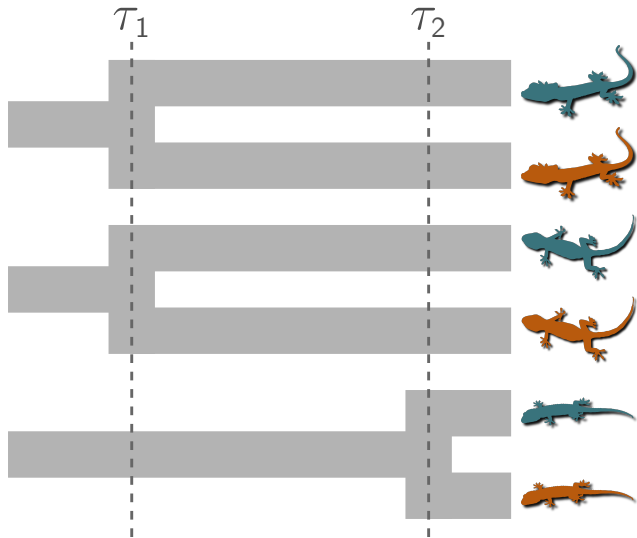


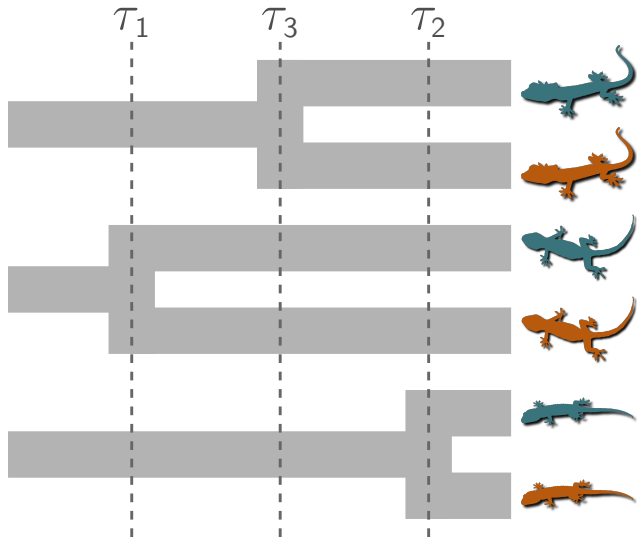


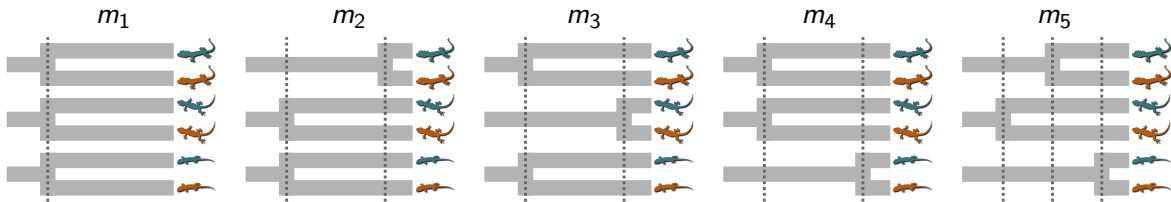


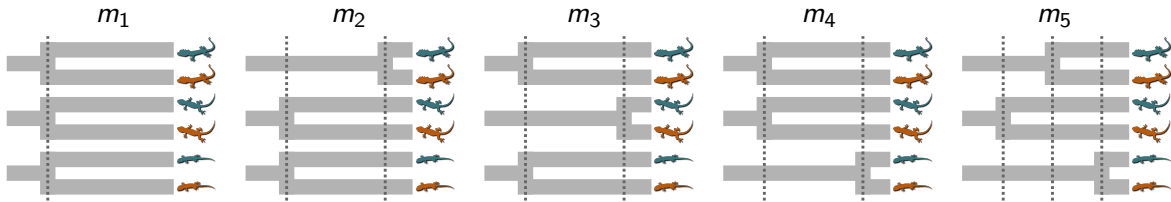






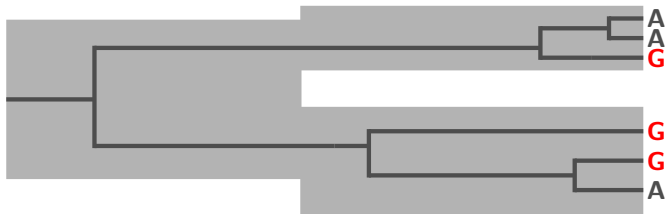


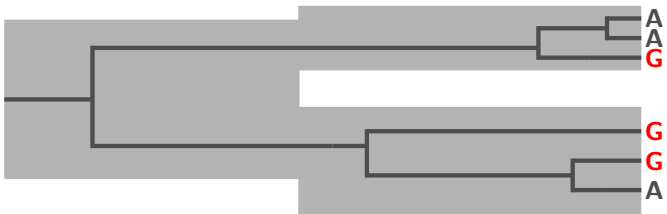




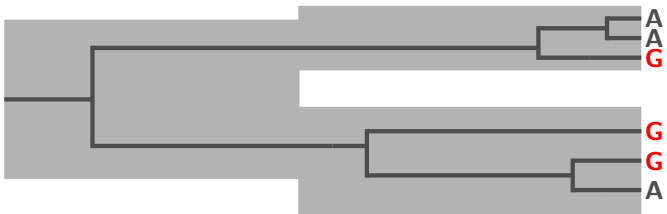
We want to infer the model and divergence times given genetic data



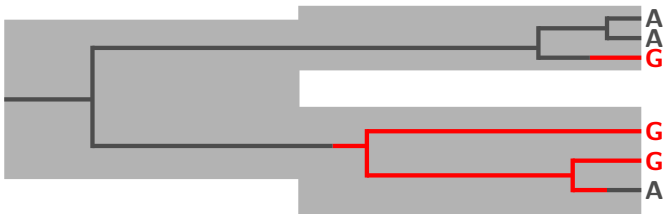




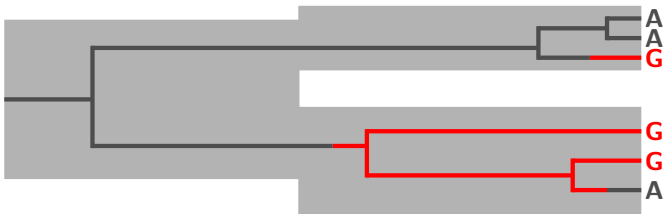
- ▶ Conditional on “population tree”, model “gene tree” using coalescent
 - ▶ Coalescent is a stochastic model of shared inheritance (continuous-time Markov chain = CTMC)



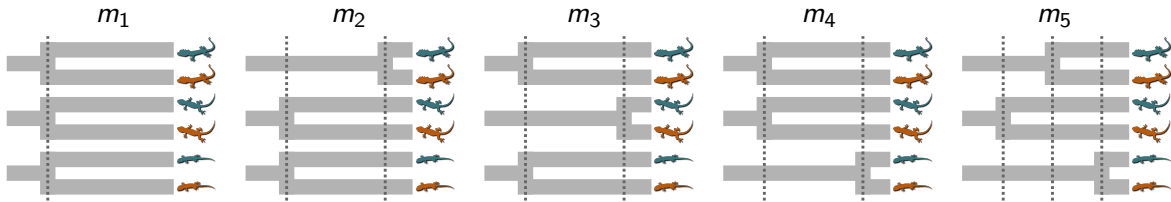
- ▶ Conditional on “population tree”, model “gene tree” using coalescent
 - ▶ Coalescent is a stochastic model of shared inheritance (continuous-time Markov chain = CTMC)
 - ▶ Branching pattern is a function of population size



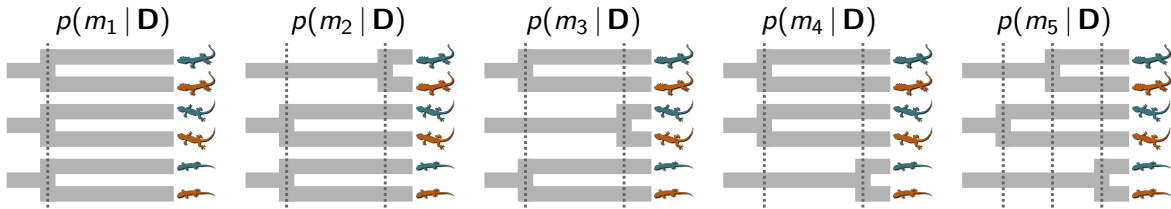
- ▶ Conditional on “population tree”, model “gene tree” using coalescent
 - ▶ Coalescent is a stochastic model of shared inheritance (continuous-time Markov chain = CTMC)
 - ▶ Branching pattern is a function of population size
- ▶ Conditional on gene tree, model mutation as a CTMC
- ▶ Genetic characters provide information about gene trees



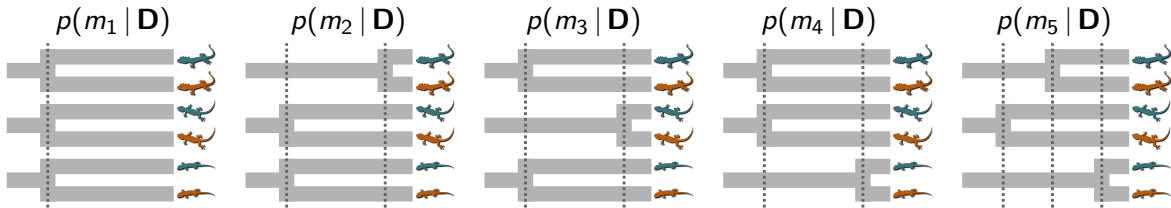
- ▶ Conditional on “population tree”, model “gene tree” using coalescent
 - ▶ Coalescent is a stochastic model of shared inheritance (continuous-time Markov chain = CTMC)
 - ▶ Branching pattern is a function of population size
- ▶ Conditional on gene tree, model mutation as a CTMC
- ▶ Genetic characters provide information about gene trees
- ▶ Gene trees inform population tree (population sizes and divergence time)



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

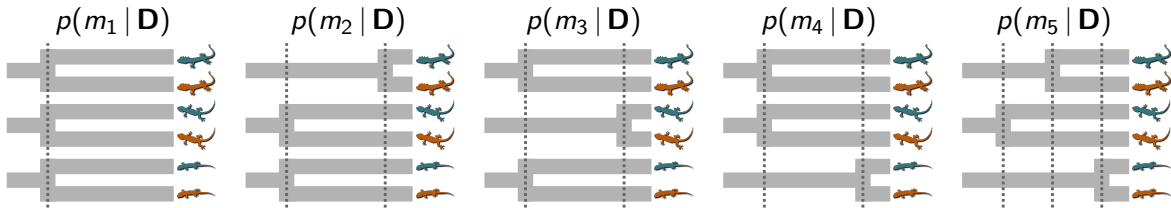


We want to infer the model and divergence times given genetic data



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

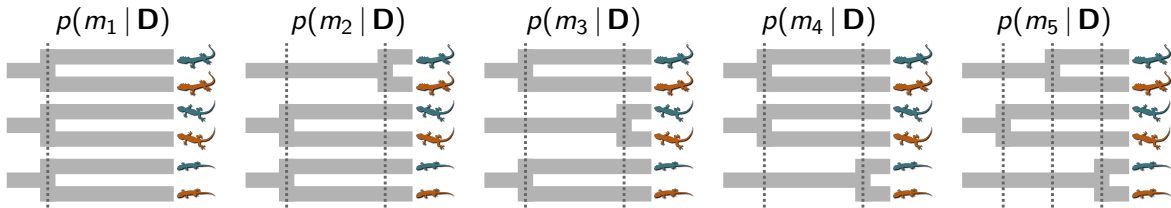
$$p(m_i | \mathbf{D}) \propto p(\mathbf{D} | m_i)p(m_i)$$



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

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

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

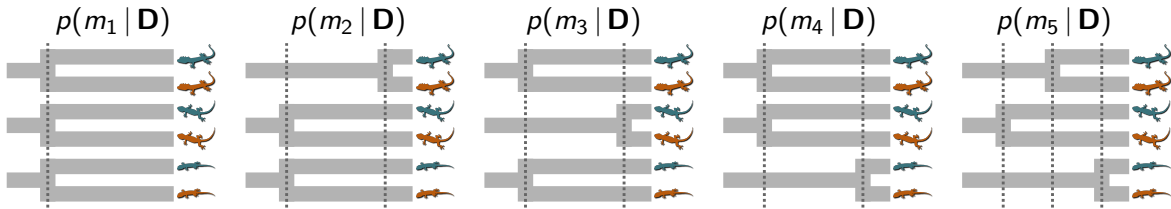


We want to infer the model and divergence times given genetic data

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

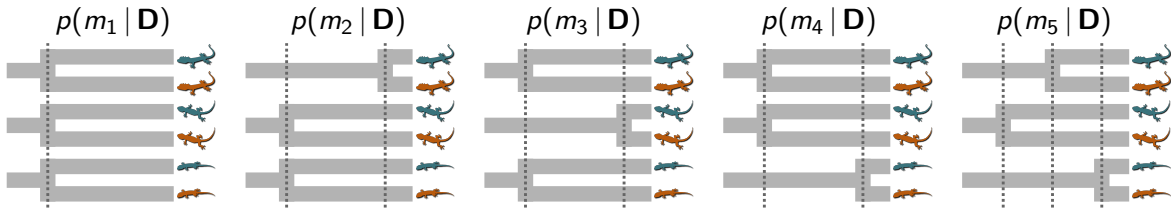
$$p(\mathbf{D} | m_i) = \int_{\theta} p(\mathbf{D} | \theta, m_i) p(\theta | m_i) d\theta$$

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



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

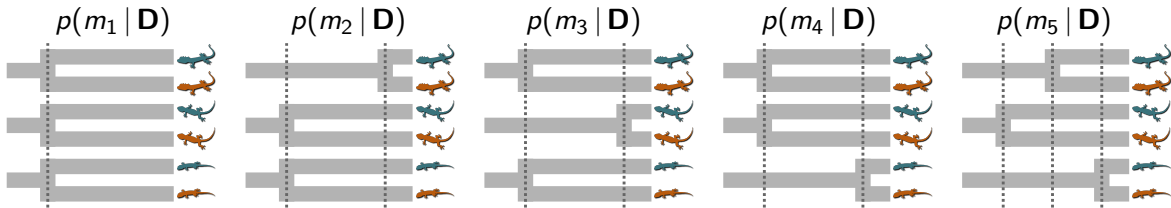
Challenges:



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

Challenges:

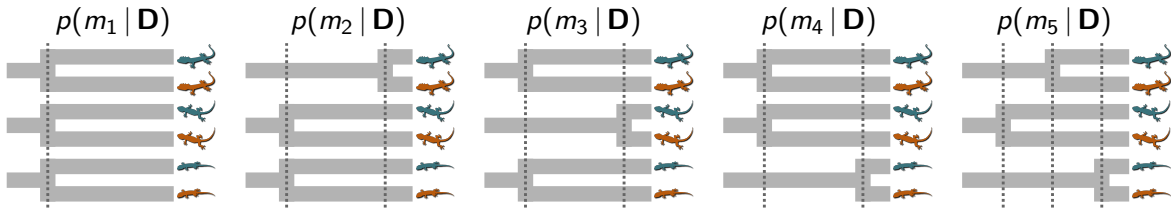
1. Likelihood is tractable, but gene trees are difficult



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

Challenges:

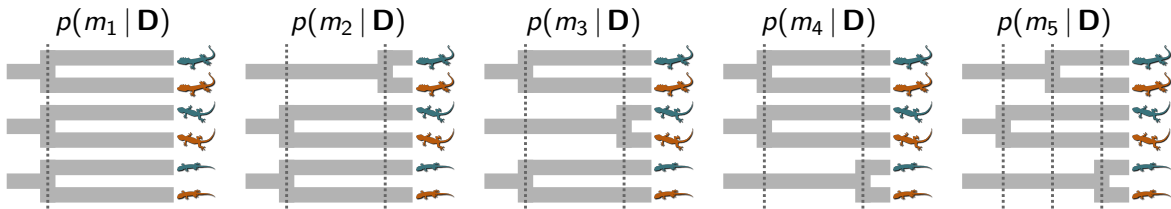
1. Likelihood is tractable, but gene trees are difficult
2. Sampling over all possible models



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

Challenges:

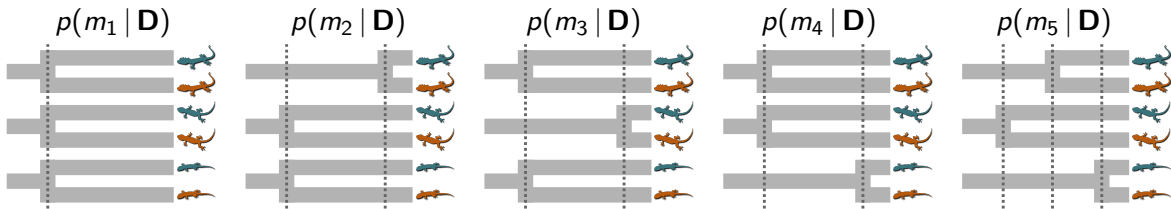
1. Likelihood is tractable, but gene trees are difficult
2. Sampling over all possible models
 - ▶ 3 taxa = 5 models



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

Challenges:

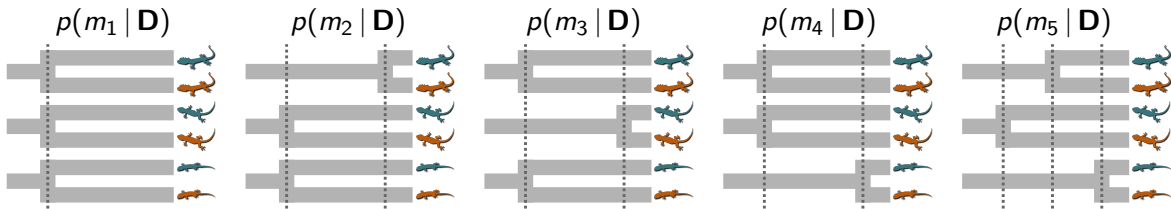
1. Likelihood is tractable, but gene trees are difficult
2. Sampling over all possible models
 - ▶ 3 taxa = 5 models
 - ▶ 10 taxa = 115,975 models



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!!

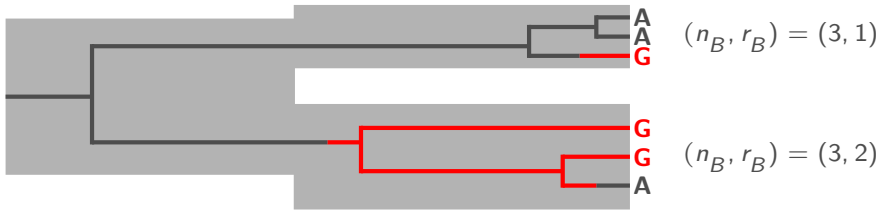


We want to infer the model and divergence times given genetic data

Challenges:

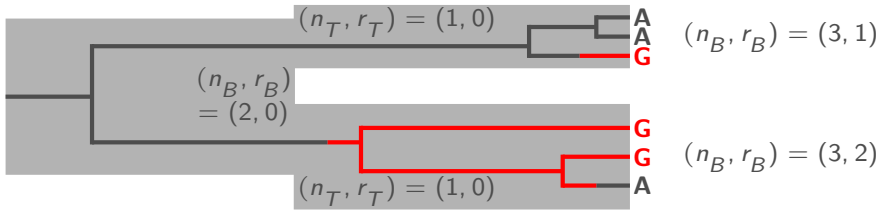
1. Likelihood is tractable, but gene trees are difficult
2. Sampling over all possible models
 - ▶ 3 taxa = 5 models
 - ▶ 10 taxa = 115,975 models
 - ▶ 20 taxa = 51,724,158,235,372 models!!

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



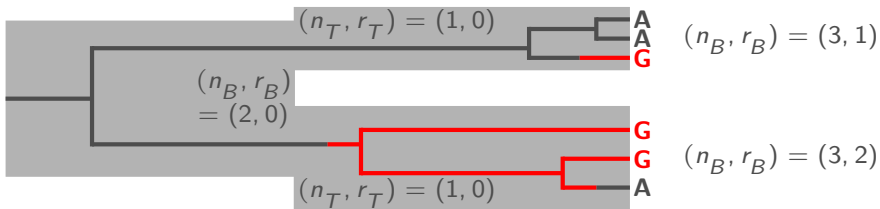
¹ T. Schmelzer and L. N. Trefethen (2007). *Electronic Transactions on Numerical Analysis* 29: 1–18

² D. Bryant et al. (2012). *Molecular Biology and Evolution* 29: 1917–1932



¹ T. Schmelzer and L. N. Trefethen (2007). *Electronic Transactions on Numerical Analysis* 29: 1–18

² D. Bryant et al. (2012). *Molecular Biology and Evolution* 29: 1917–1932

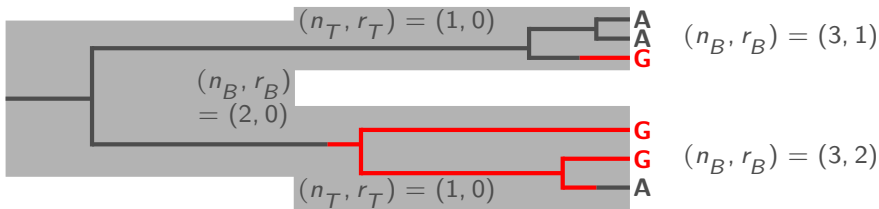


$Q =$

	$(1, 0)$	$(1, 1)$	$(2, 0)$	$(2, 1)$	$\dots$	(n, n)
$(1, 0)$	.	.	.	.		.
$(1, 1)$	.	.	.	.		.
$(2, 0)$	.	.	.	.		.
$(2, 1)$	.	.	.	.		.
$\vdots$						
(n, n)	.	.	.	.		.

¹ T. Schmelzer and L. N. Trefethen (2007). *Electronic Transactions on Numerical Analysis* 29: 1–18

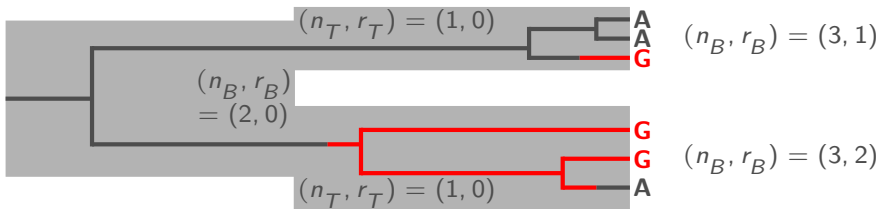
² D. Bryant et al. (2012). *Molecular Biology and Evolution* 29: 1917–1932



	$(1, 0)$	$(1, 1)$	$(2, 0)$	$(2, 1)$	$\dots$	(n, n)		
$Q =$	$(1, 0)$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$Q_{(n,r);(n,r-1)}$	$= (n - r + 1)v$, mutation,
	$(1, 1)$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$Q_{(n,r);(n,r+1)}$	$= (r + 1)u$, mutation,
	$(2, 0)$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$Q_{(n,r);(n-1,r)}$	$= \frac{(n-1-r)n}{2N_e(u+v)}$, coalescence,
	$(2, 1)$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$Q_{(n,r);(n-1,r-1)}$	$= \frac{(r-1)n}{2N_e(u+v)}$, coalescence,
	$\vdots$						$Q_{(n,r);(n,r)}$	$= -\frac{(n-1)n}{2N_e(u+v)} - (n - r)v - ru$.
	(n, n)	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$		

¹ T. Schmelzer and L. N. Trefethen (2007). *Electronic Transactions on Numerical Analysis* 29: 1–18

² D. Bryant et al. (2012). *Molecular Biology and Evolution* 29: 1917–1932

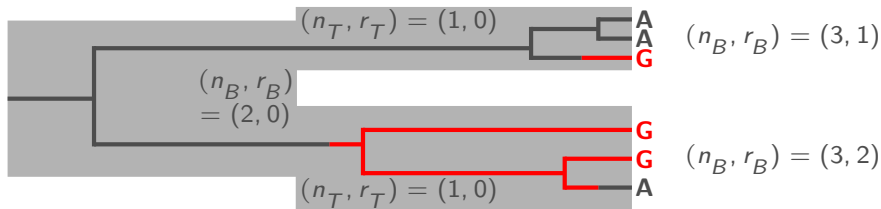


		$(1, 0)$	$(1, 1)$	$(2, 0)$	$(2, 1)$	$\dots$	(n, n)	
$Q =$	$(1, 0)$	.	.	.	.	.	.	$Q_{(n,r);(n,r-1)} = (n-r+1)v$, mutation,
	$(1, 1)$	.	.	.	.	.	.	$Q_{(n,r);(n,r+1)} = (r+1)u$, mutation,
	$(2, 0)$	.	.	.	.	.	.	$Q_{(n,r);(n-1,r)} = \frac{(n-1-r)n}{2N_e(u+v)}$, coalescence,
	$(2, 1)$	.	.	.	.	.	.	$Q_{(n,r);(n-1,r-1)} = \frac{(r-1)n}{2N_e(u+v)}$, coalescence,
	$\vdots$							
	(n, n)	.	.	.	.	.	.	$Q_{(n,r);(n,r)} = -\frac{(n-1)n}{2N_e(u+v)} - (n-r)v - ru$.

► e^{Qt} to keep track of all conditional probabilities along each branch (Carathéodory-Fejér method¹)

¹ T. Schmelzer and L. N. Trefethen (2007). *Electronic Transactions on Numerical Analysis* 29: 1–18

² D. Bryant et al. (2012). *Molecular Biology and Evolution* 29: 1917–1932

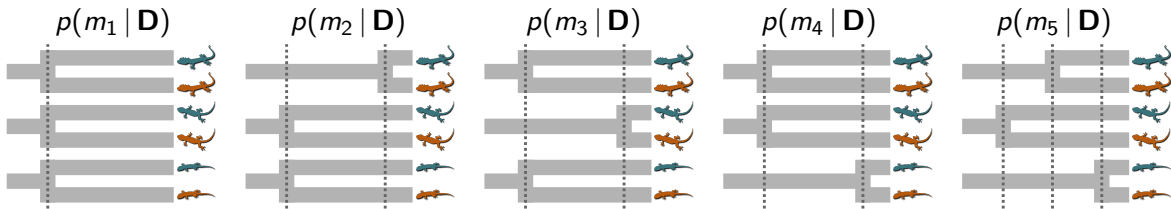


		$(1, 0)$	$(1, 1)$	$(2, 0)$	$(2, 1)$	$\dots$	(n, n)		
$Q =$	$(1, 0)$	.	.	.	.	.	.	$Q_{(n,r);(n,r-1)}$	$= (n - r + 1)v$, mutation,
	$(1, 1)$	.	.	.	.	.	.	$Q_{(n,r);(n,r+1)}$	$= (r + 1)u$, mutation,
	$(2, 0)$	.	.	.	.	.	.	$Q_{(n,r);(n-1,r)}$	$= \frac{(n-1-r)n}{2N_e(u+v)}$, coalescence,
	$(2, 1)$	.	.	.	.	.	.	$Q_{(n,r);(n-1,r-1)}$	$= \frac{(r-1)n}{2N_e(u+v)}$, coalescence,
	$\vdots$							$Q_{(n,r);(n,r)}$	$= -\frac{(n-1)n}{2N_e(u+v)} - (n - r)v - ru$.
	(n, n)	.	.	.	.	.	.		

- ▶ e^{Qt} to keep track of all conditional probabilities along each branch (Carathéodory-Fejér method¹)
- ▶ At root, get likelihood of population tree integrated over all possible gene trees and mutational histories²

¹ T. Schmelzer and L. N. Trefethen (2007). *Electronic Transactions on Numerical Analysis* 29: 1–18

² D. Bryant et al. (2012). *Molecular Biology and Evolution* 29: 1917–1932



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

- ▶ We need a distribution over all ways of partitioning the population pairs to divergence-time classes
- ▶ The Dirichlet process (DP) is a convenient and flexible solution



Peter Dirichlet

- ▶ We need a distribution over all ways of partitioning the population pairs to divergence-time classes
- ▶ The Dirichlet process (DP) is a convenient and flexible solution
 - ▶ Common Bayesian nonparametric approach to assigning variables to an unknown number of categories



Peter Dirichlet

- ▶ We need a distribution over all ways of partitioning the population pairs to divergence-time classes
- ▶ The Dirichlet process (DP) is a convenient and flexible solution
 - ▶ Common Bayesian nonparametric approach to assigning variables to an unknown number of categories
 - ▶ Controlled by “concentration” parameter



Peter Dirichlet

- ▶ We need a distribution over all ways of partitioning the population pairs to divergence-time classes
- ▶ The Dirichlet process (DP) is a convenient and flexible solution
 - ▶ Common Bayesian nonparametric approach to assigning variables to an unknown number of categories
 - ▶ Controlled by “concentration” parameter
 - ▶ Variables are exchangeable under the DP, allowing Gibbs sampling



Peter Dirichlet

Ecoevolity

Estimating evolutionary coeality

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

- ▶ Analytically integrate over gene trees and mutational histories¹
- ▶ Dirichlet-process prior across divergence models
- ▶ Gibbs sampling² to numerically sample models

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 **evol**utionary **coevality**

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

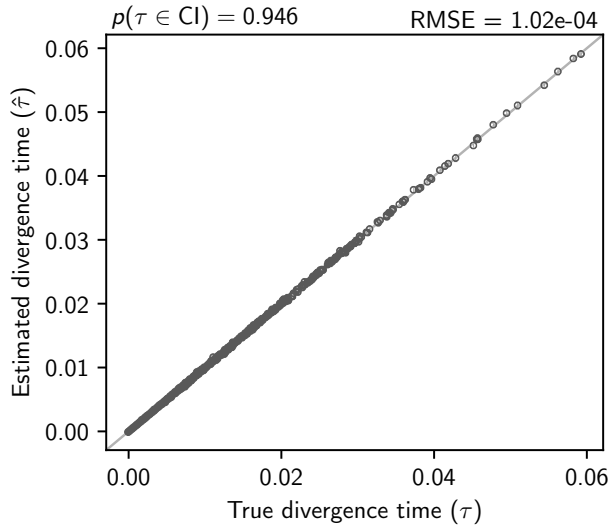
- ▶ Analytically integrate over gene trees and mutational histories¹
- ▶ Dirichlet-process prior across divergence models
- ▶ Gibbs sampling² to numerically sample models
- ▶ *Goal: Fast, full-likelihood Bayesian method to infer patterns of co-diversification from genome-scale data*

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

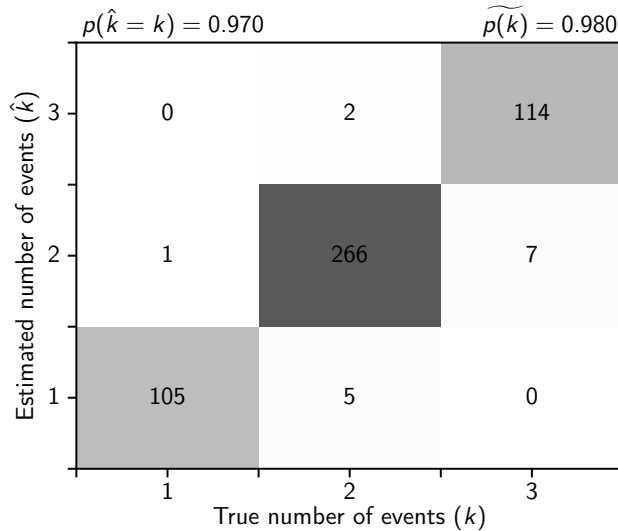
² R. M. Neal (2000). *Journal of Computational and Graphical Statistics* 9: 249–265

Does it work?

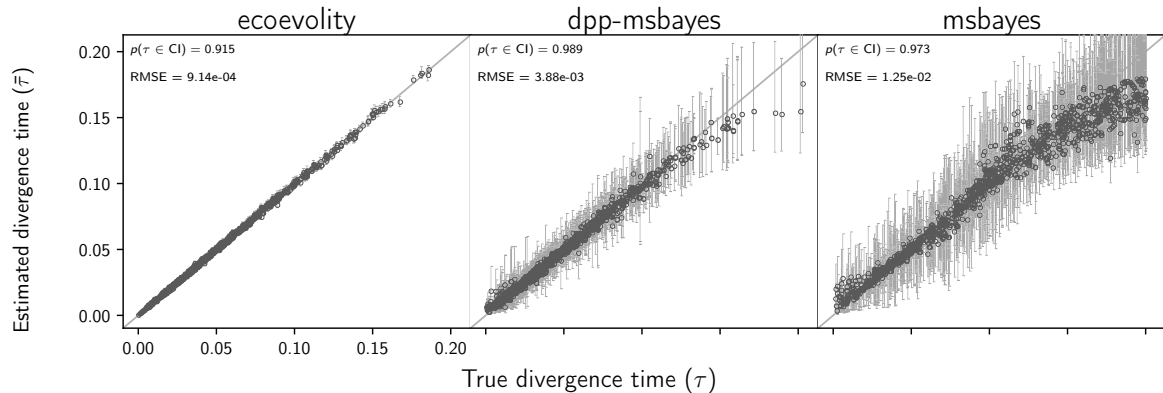
Simulation results



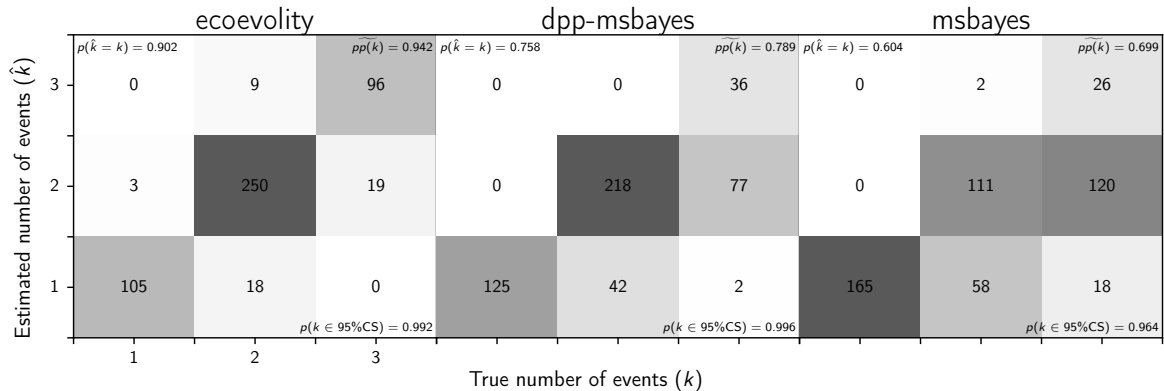
Simulation results



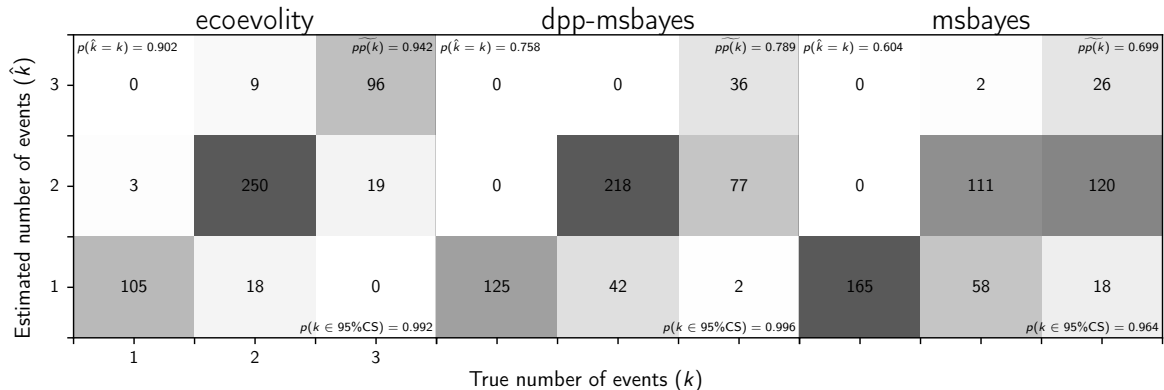
“Bake-off” results



“Bake-off” results



“Bake-off” results



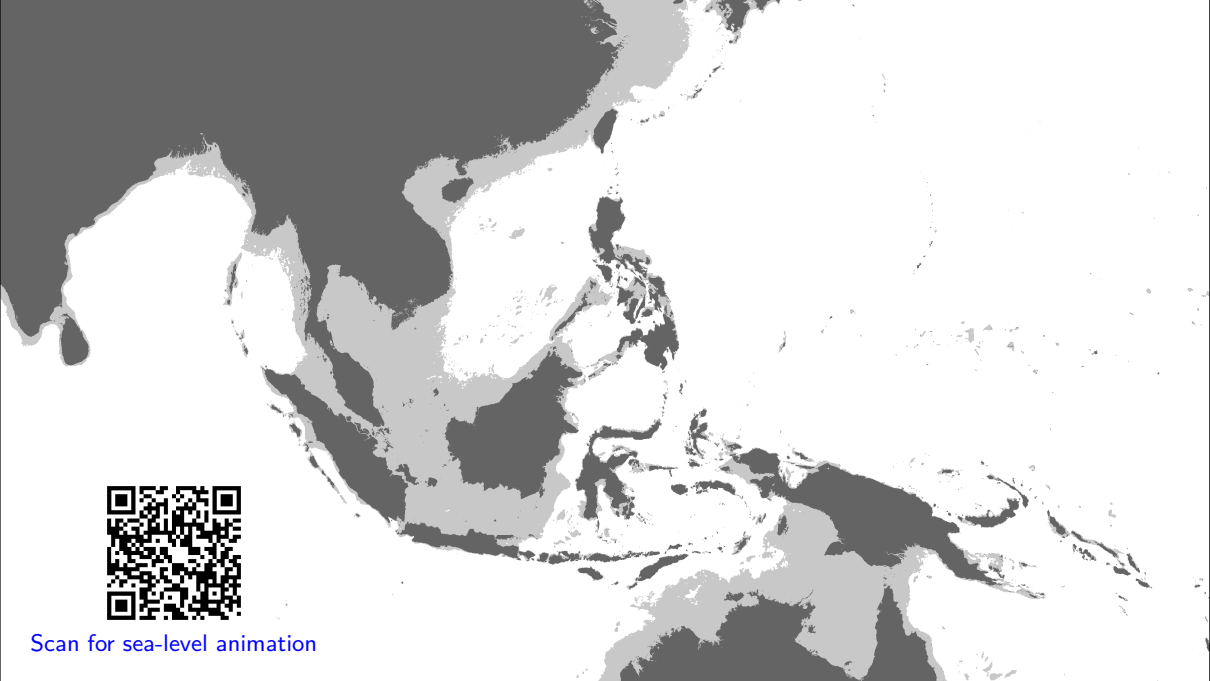
33.4 minutes

Average run time:

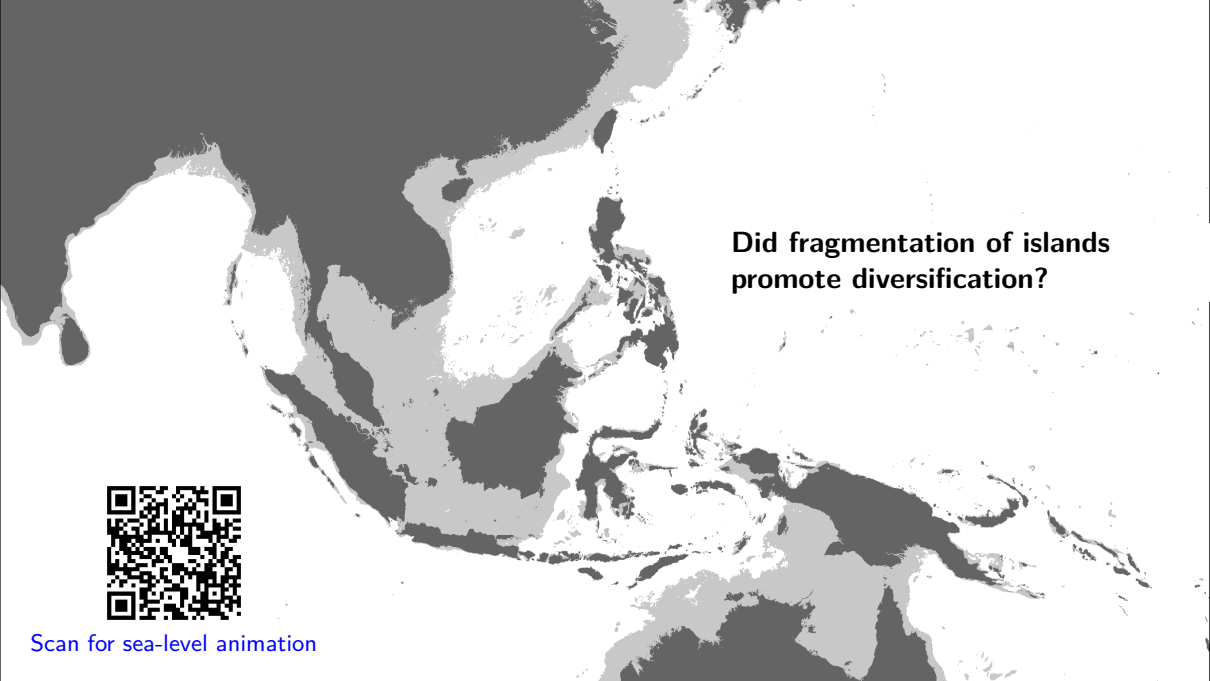
4.4 days



[Scan for sea-level animation](#)



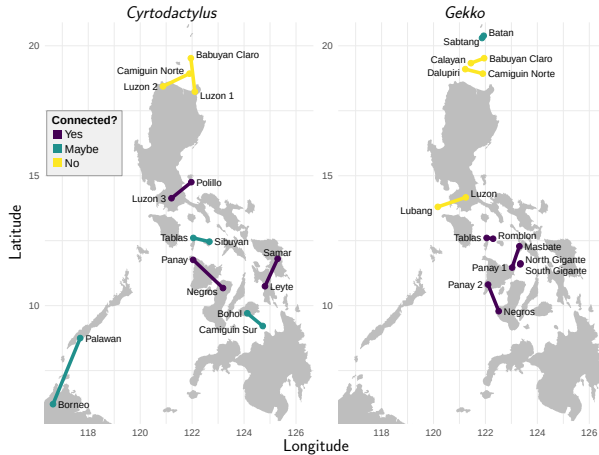
Scan for sea-level animation



**Did fragmentation of islands
promote diversification?**

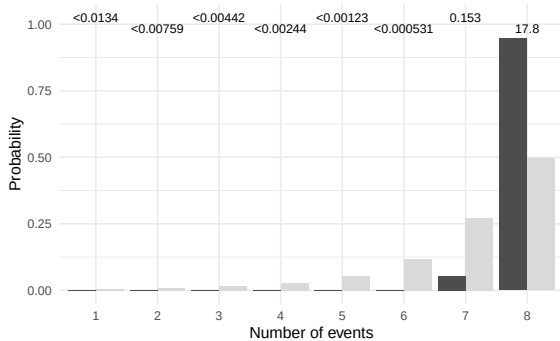


[Scan for sea-level animation](#)

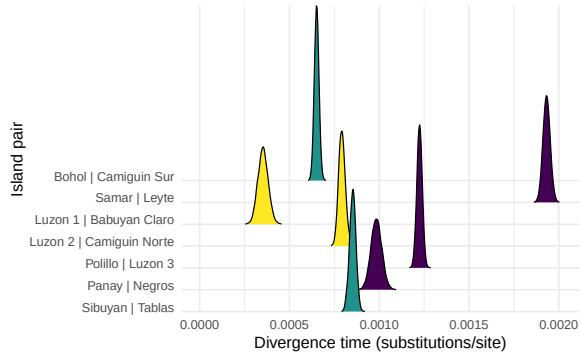
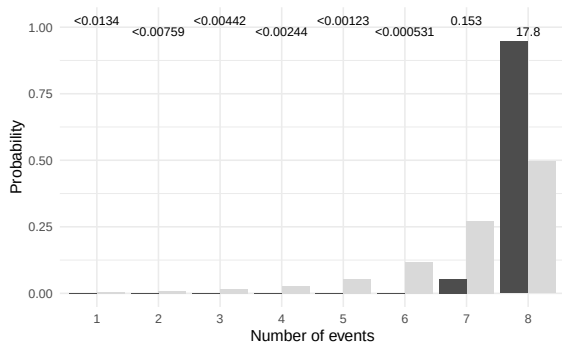


- ▶ Sampled 2–5 individuals from 8 pairs of populations for both *Cyrtodactylus* and *Gekko*
- ▶ Collected short DNA sequences (RADseq) from across genome of each individual

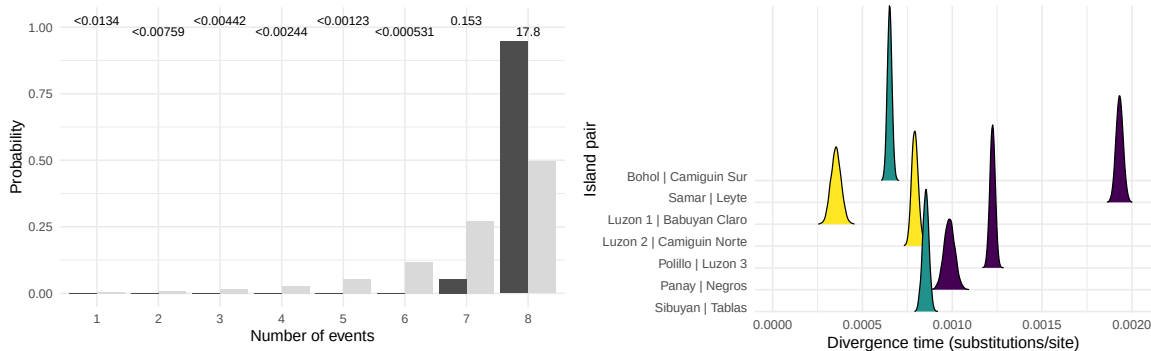
Results: *Cyrtodactylus*



Results: *Cyrtodactylus*



Results: *Cyrtodactylus*



Strong support for independent divergences

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 perhaps more important

Take-home points

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

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

Approaches to the problem

A pairwise approach (keep it “simple”)

A fully phylogenetic approach



Biogeography

- ▶ Environmental changes that affect whole communities of species

Genome evolution

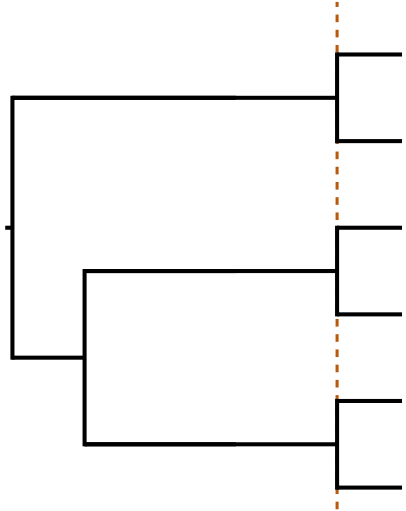
- ▶ Duplication of a chromosome segment harboring gene families

Epidemiology

- ▶ Transmission at social gatherings

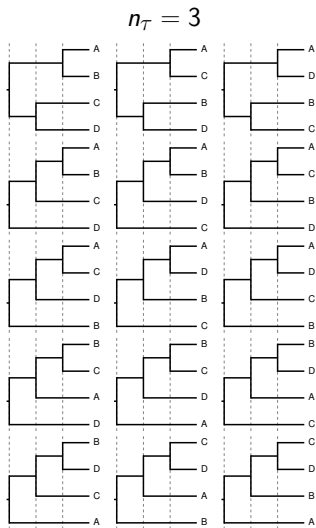
Endosymbiont evolution (e.g., parasites, microbiome)

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

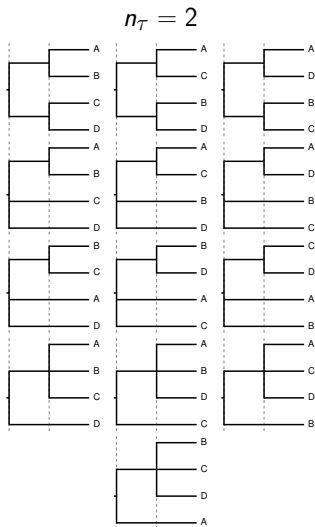
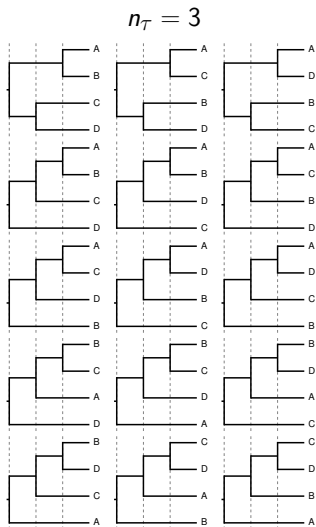


Generalizing tree space

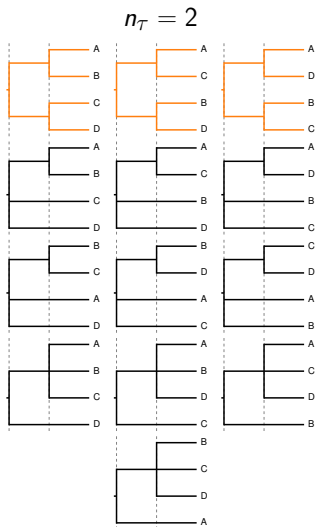
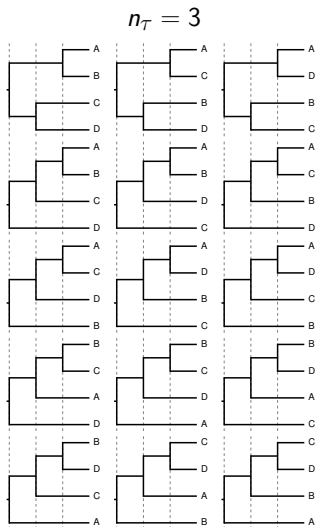
Generalizing tree space



Generalizing tree space



Generalizing tree space

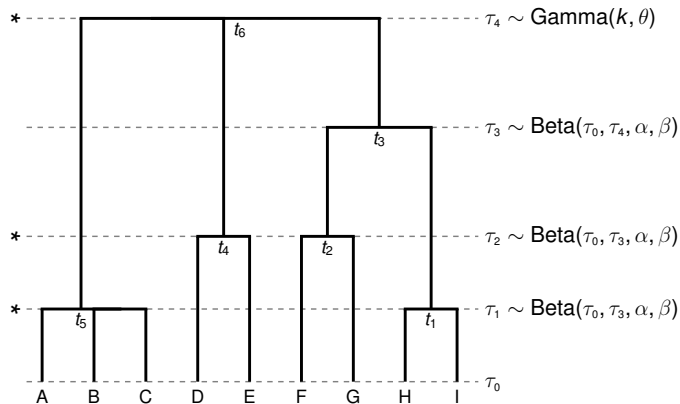


$n_T = 1$

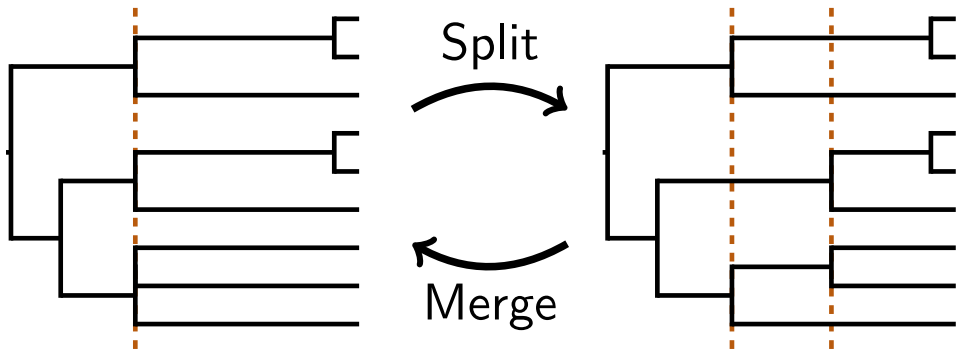


Generalized tree distribution

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

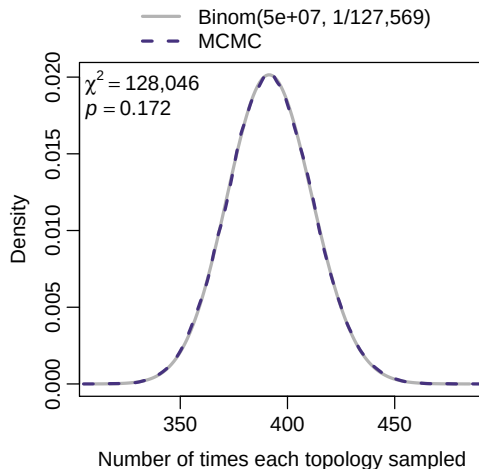


Inferring trees with shared divergences



Reversible-jump MCMC

Validating rjMCMC with 7-leaf tree



The rjMCMC algorithms sample the expected generalized tree distribution

PhycoEval

Phylogenetic coevality

J. R. Oaks et al. (2021). *bioRxiv*

Ecoevolity

Estimating evolutionary coevality

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

- ▶ **Tree model**

- ▶ rjMCMC sampling of generalized tree distribution

PhycoEval

Phylogenetic coevality

J. R. Oaks et al. (2021). *bioRxiv*

Ecoevolity

Estimating evolutionary 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 genealogies¹

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

PhycoEval

Phylogenetic coevality

J. R. Oaks et al. (2021). *bioRxiv*

Ecoevolity

Estimating evolutionary coevality

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

- ▶ **Tree model**

- ▶ rjMCMC sampling of generalized tree distribution

- ▶ **Likelihood model**

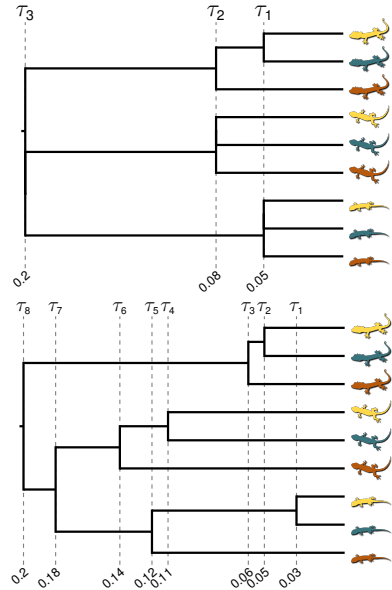
- ▶ CTMC model of characters evolving along genealogies
 - ▶ Infer species trees by analytically integrate over genealogies¹

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

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

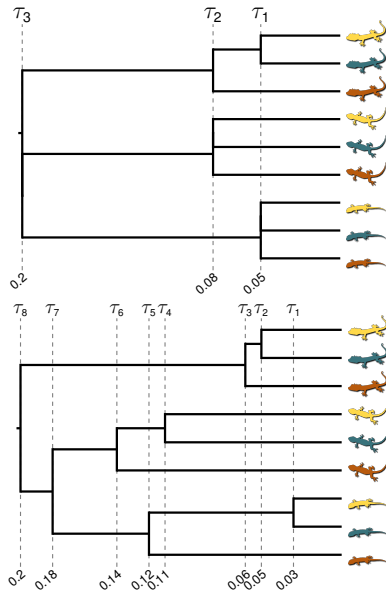
Methods: Simulations

- ▶ Simulated 100 data sets with 50,000 base pairs



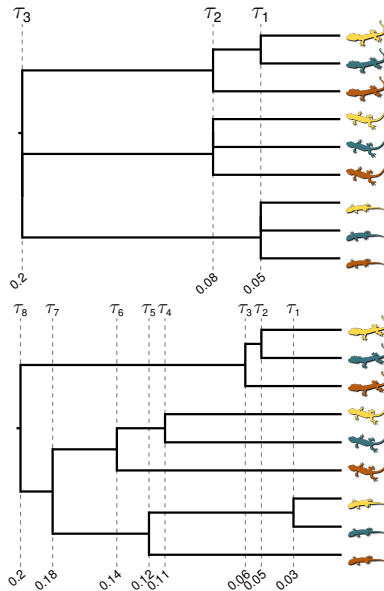
Methods: Simulations

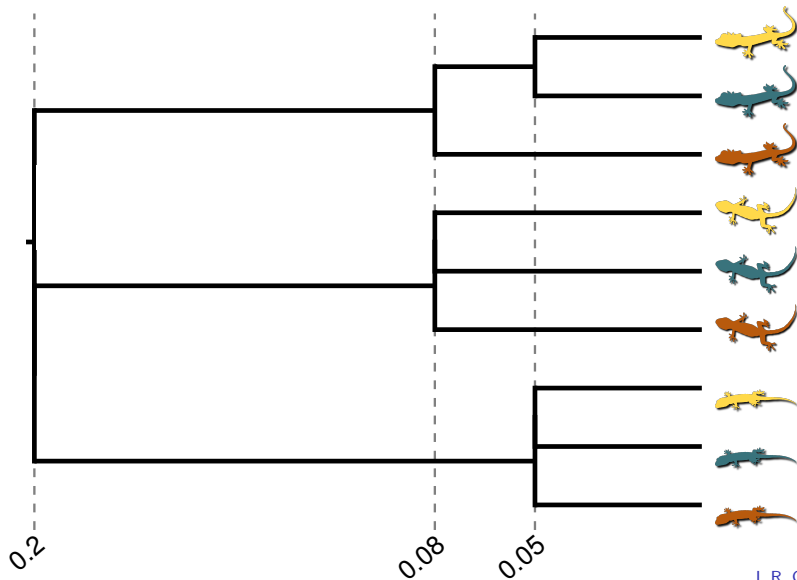
- ▶ Simulated 100 data sets with 50,000 base pairs
- ▶ Analyzed each data set with:
 - ▶ M_G = Generalized tree model
 - ▶ M_{IB} = Independent-bifurcating tree model



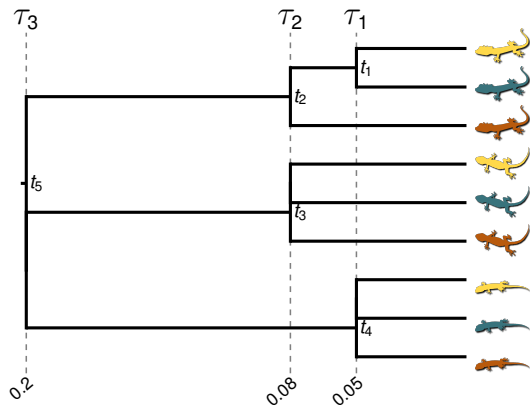
Methods: Simulations

- ▶ Simulated 100 data sets with 50,000 base pairs
- ▶ 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}

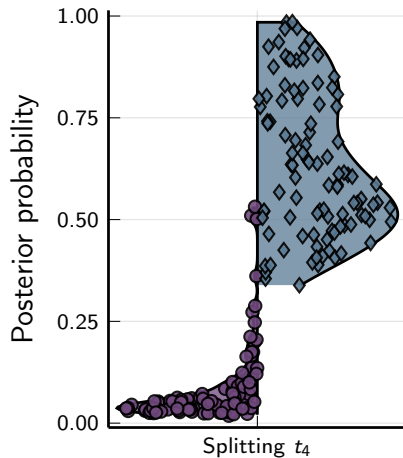
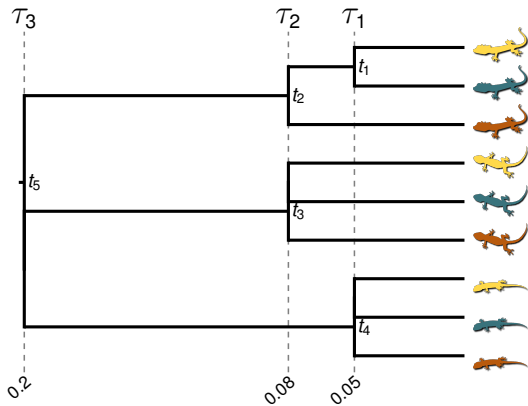




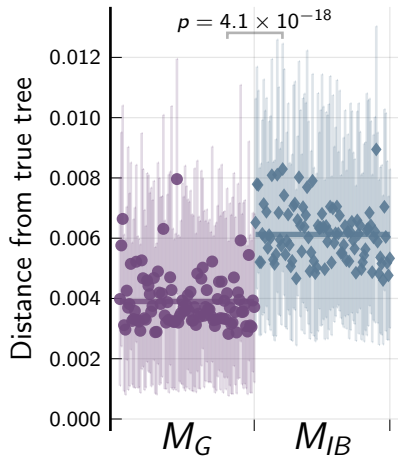
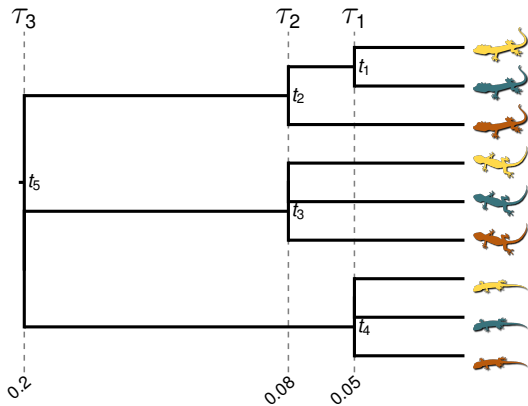




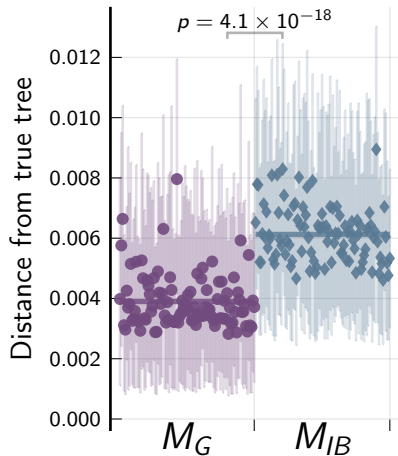
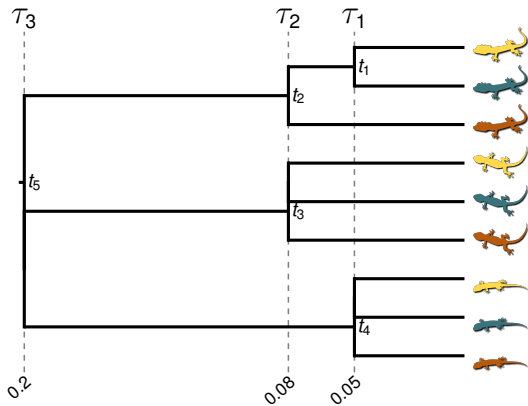
● M_G = Generalized model ◆ M_{IB} = Independent-bifurcating model



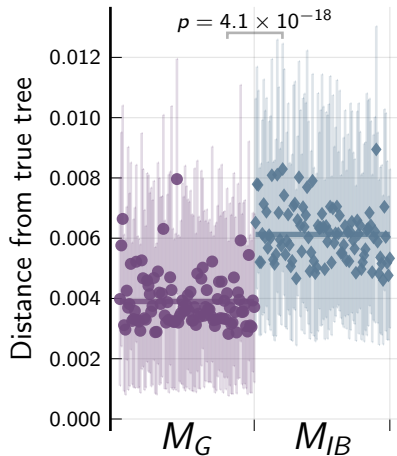
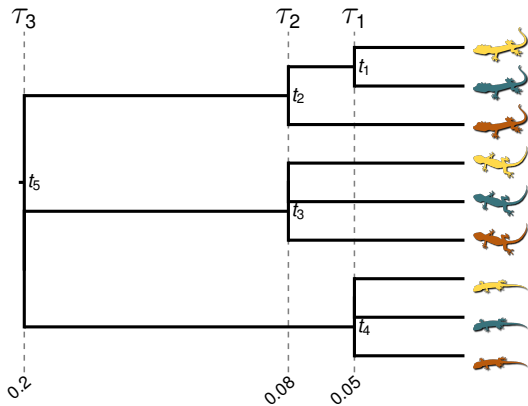
● M_G = Generalized model ◆ M_{IB} = Independent-bifurcating model



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

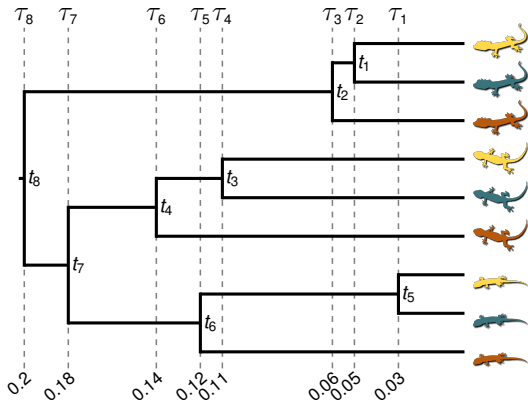


● M_G = Generalized model ◆ M_{IB} = Independent-bifurcating model

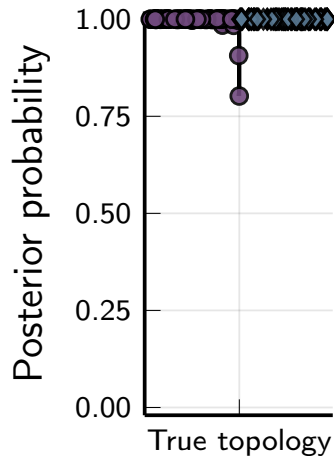
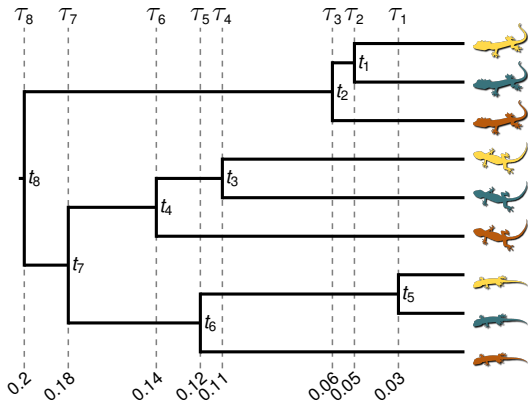


M_G significantly better at inferring trees with shared divergences

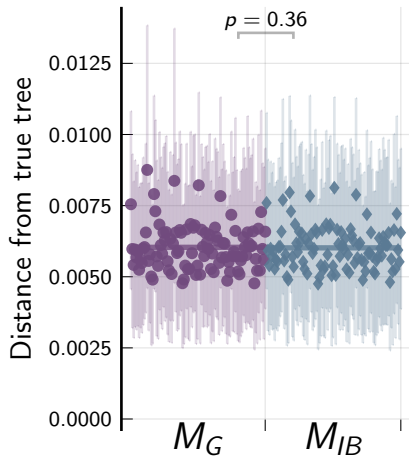
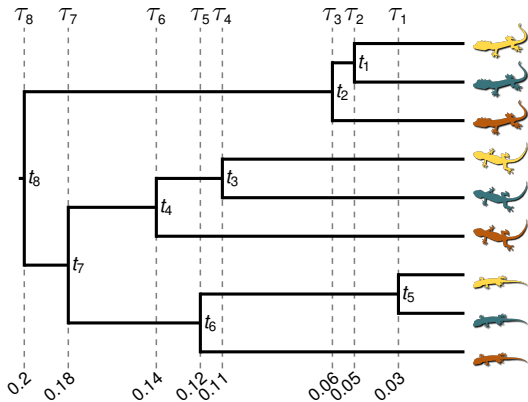
 M_G = Generalized model
  M_{IB} = Independent-bifurcating model



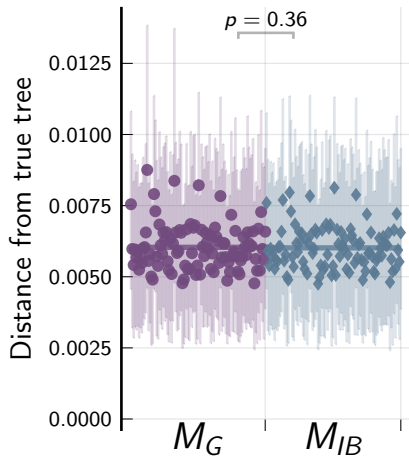
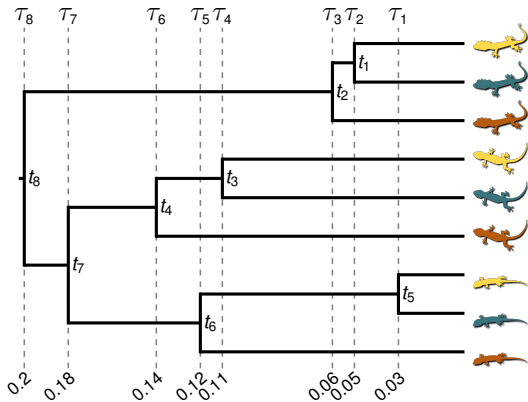
● M_G = Generalized model
 ◆ M_{IB} = Independent-bifurcating model



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

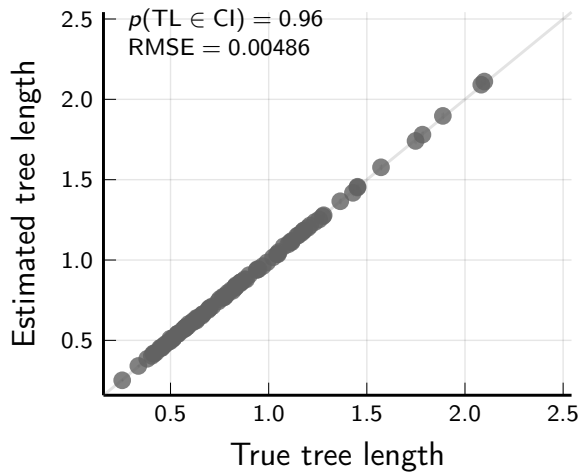
 $M_G = \text{Generalized model}$ M_{IB} = Independent-bifurcating model

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

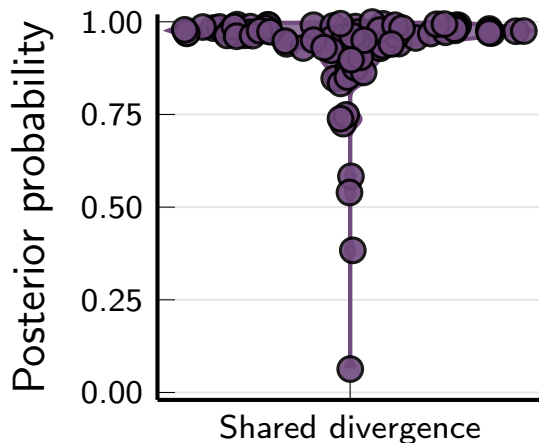
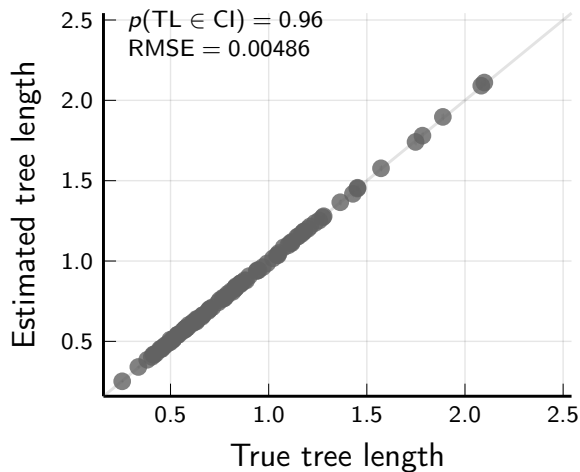


M_G performs as well as true model when divergences are independent

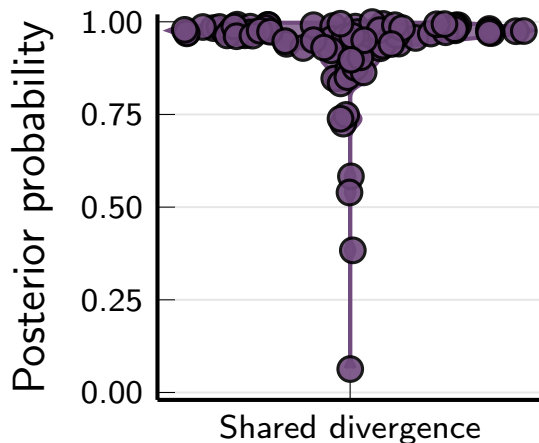
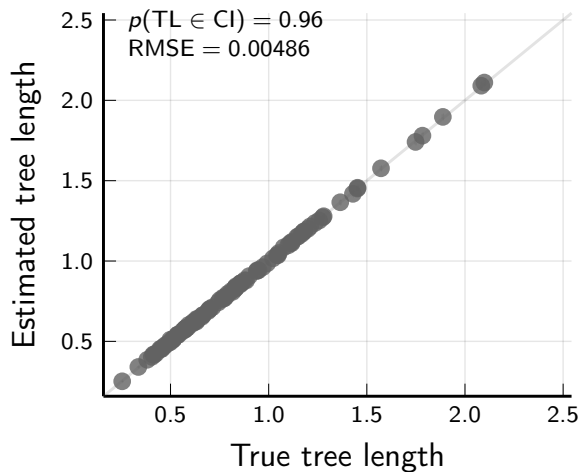
Results: random M_G trees



Results: random M_G trees



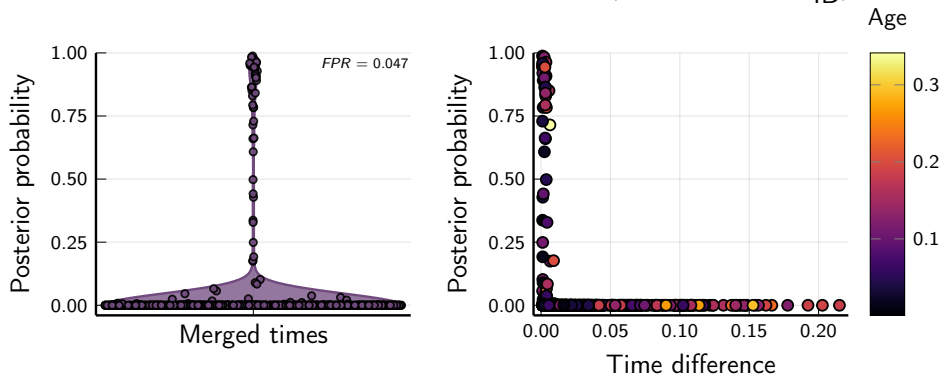
Results: random M_G trees



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

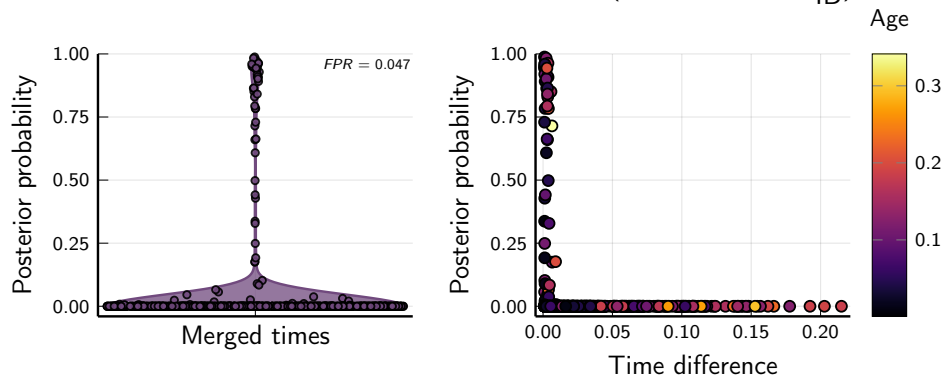
Results: random M_{IB} trees

Probability of incorrectly merged div times (true model = M_{IB})



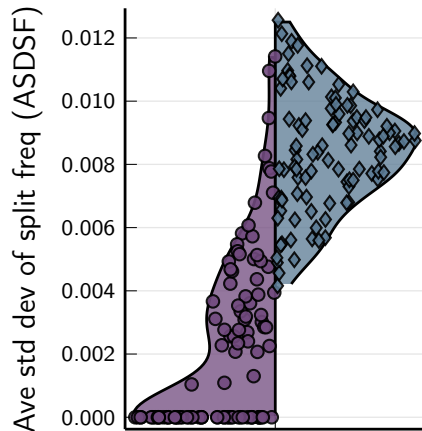
Results: random M_{IB} trees

Probability of incorrectly merged div times (true model = M_{IB})

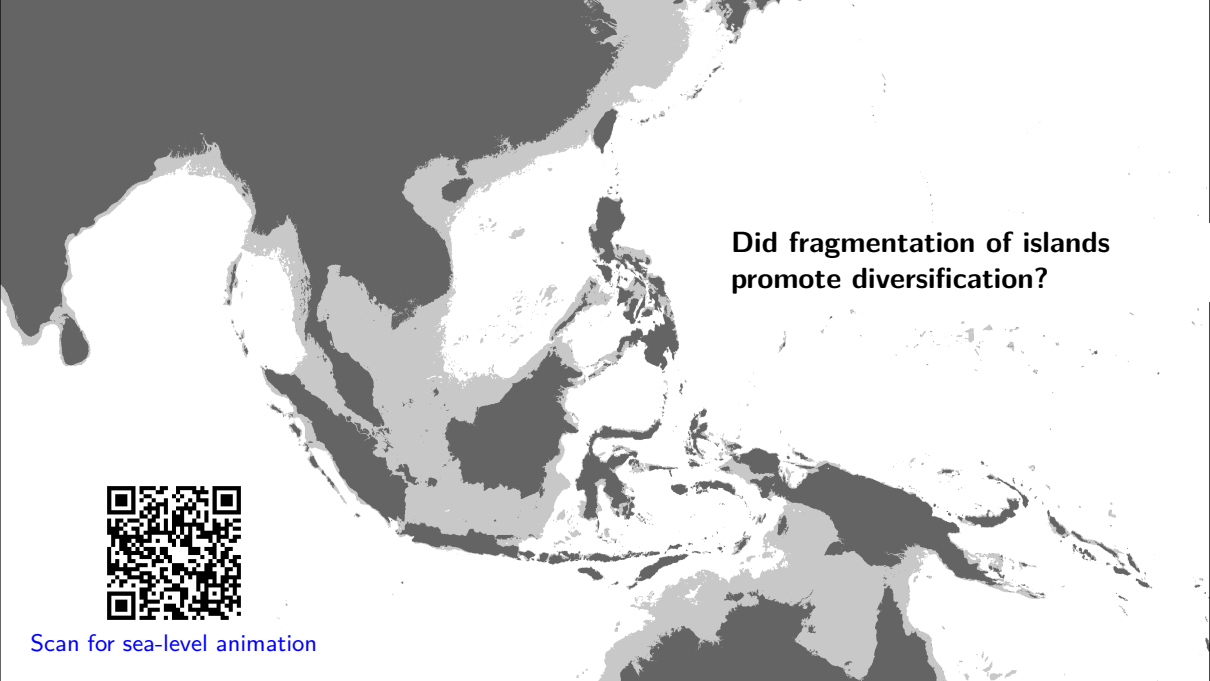


M_G has low false positive rate

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



Generalizing tree space improves MCMC convergence and mixing



**Did fragmentation of islands
promote diversification?**

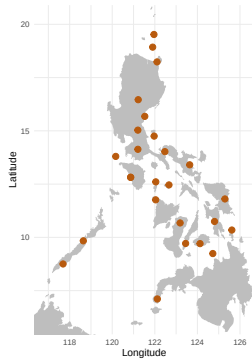


[Scan for sea-level animation](#)

Cyrtodactylus



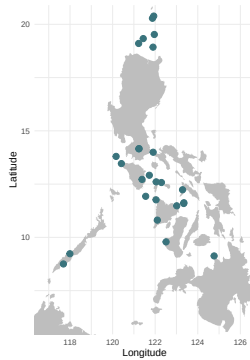
©Rafe M. Brown



Gekko



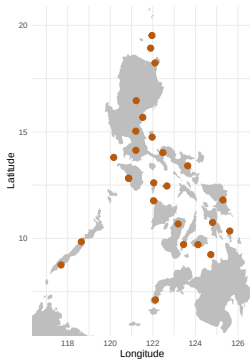
©Rafe M. Brown



Cyrtodactylus



©Rafe M. Brown

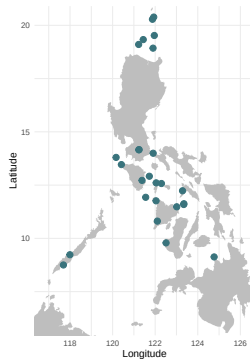


1702 loci
155,887 sites

Gekko

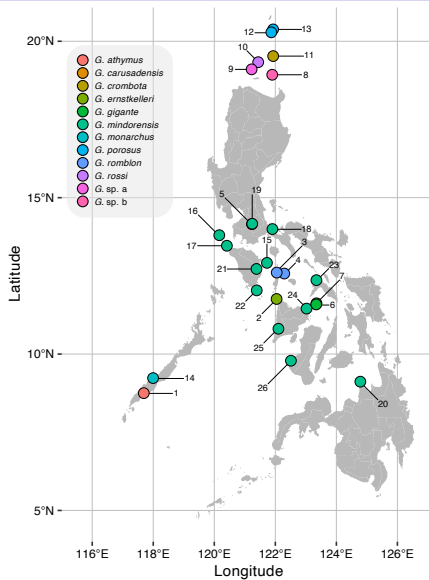
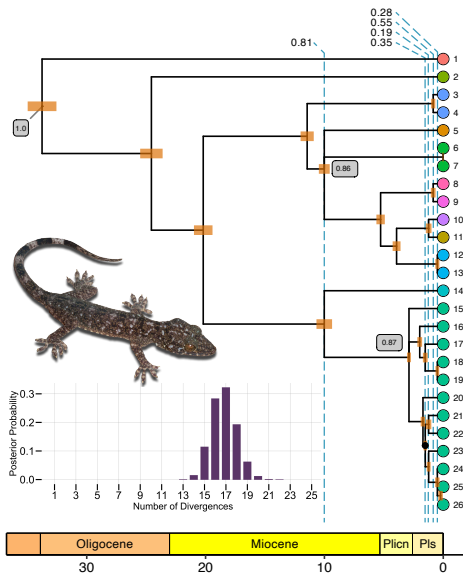


©Rafe M. Brown

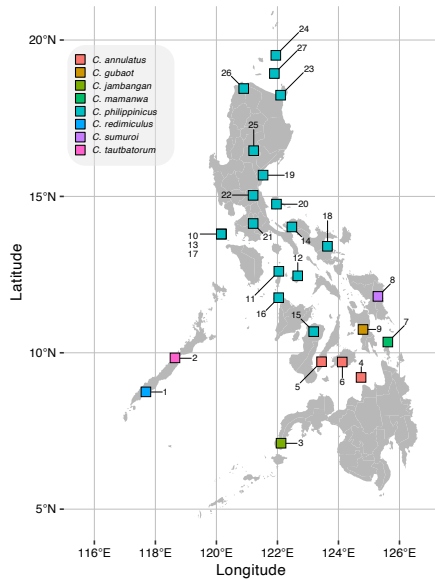
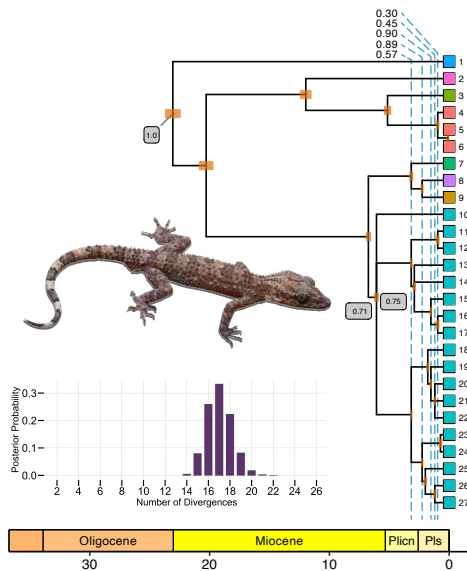


1033 loci
94,813 sites

Gekko



Cyrtodactylus



Take-home points

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

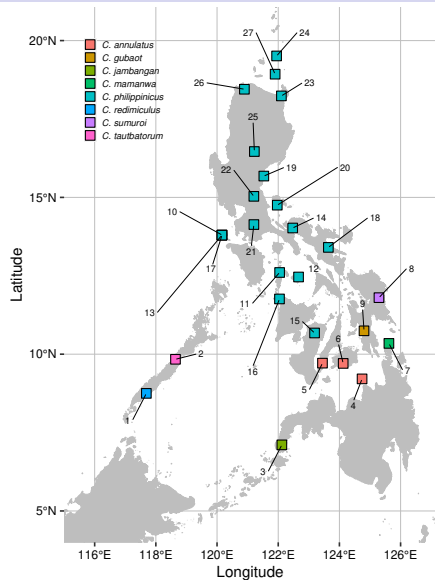
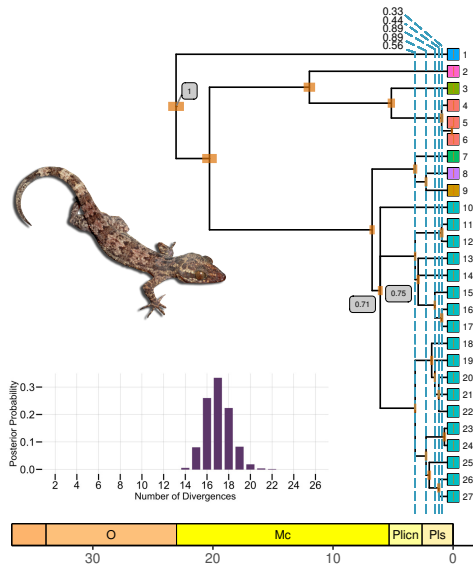
Take-home points

- ▶ 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

Cyrtodactylus



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

Ongoing work

Theory/methods

- ▶ Develop process-based and trait-dependent distributions over the space of generalized trees
 - ▶ “Birth-death-burst” model

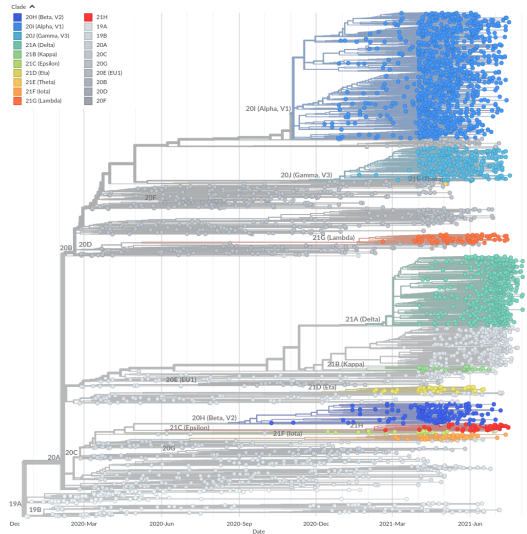
Ongoing work

Theory/methods

- ▶ Develop process-based and trait-dependent distributions over the space of generalized trees
 - ▶ “Birth-death-burst” model

Applications

- ▶ Epidemiological dynamics of “super-spreading” events during the COVID-19 pandemic



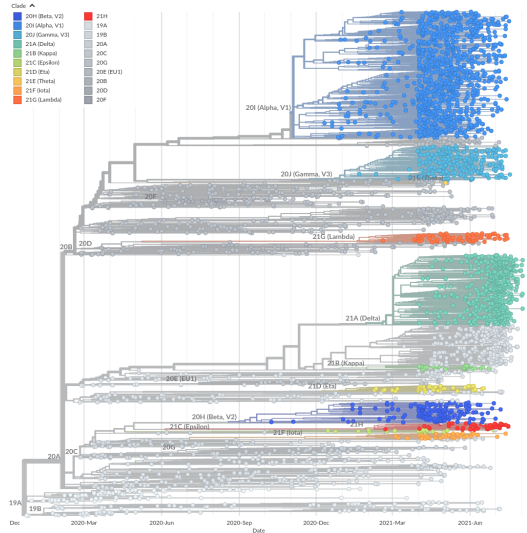
Ongoing work

Theory/methods

- ▶ Develop process-based and trait-dependent distributions over the space of generalized trees
 - ▶ “Birth-death-burst” model

Applications

- ▶ Epidemiological dynamics of “super-spreading” events during the COVID-19 pandemic
 - ▶ Estimate rate of spread via social gatherings



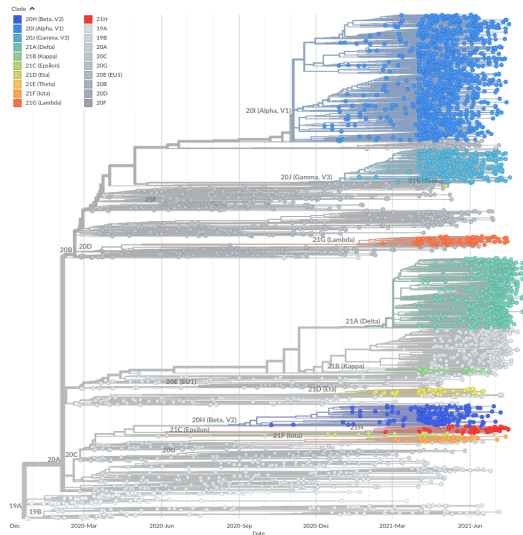
Ongoing work

Theory/methods

- ▶ Develop process-based and trait-dependent distributions over the space of generalized trees
 - ▶ “Birth-death-burst” model

Applications

- ▶ Epidemiological dynamics of “super-spreading” events during the COVID-19 pandemic
 - ▶ Estimate rate of spread via social gatherings
 - ▶ Test if this differs among variants



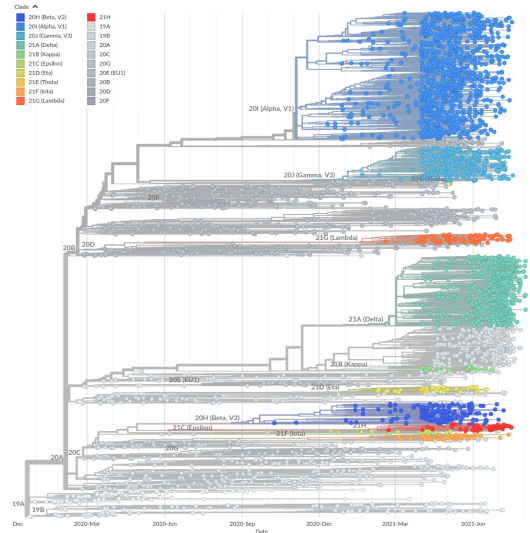
Ongoing work

Theory/methods

- ▶ Develop process-based and trait-dependent distributions over the space of generalized trees
 - ▶ “Birth-death-burst” model

Applications

- ▶ Epidemiological dynamics of “super-spreading” events during the COVID-19 pandemic
 - ▶ Estimate rate of spread via social gatherings
 - ▶ Test if this differs among variants
 - ▶ Quantify the effect of holidays



Acknowledgments

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

Computation:

- ▶ Alabama Supercomputer Authority
- ▶ Auburn University Hopper Cluster

Funding:



AUBURN
UNIVERSITY

Photo credits:

- ▶ Rafe Brown
- ▶ Perry Wood, Jr.
- ▶ [PhyloPic](#)

Questions?

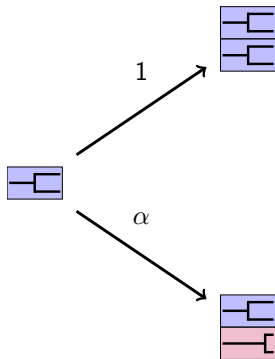
joaks@auburn.edu

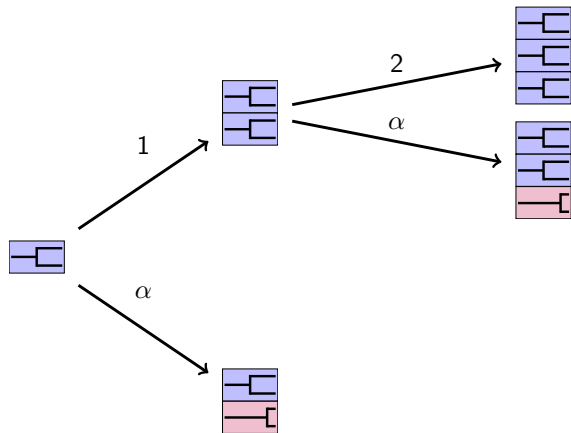
phyletica.org

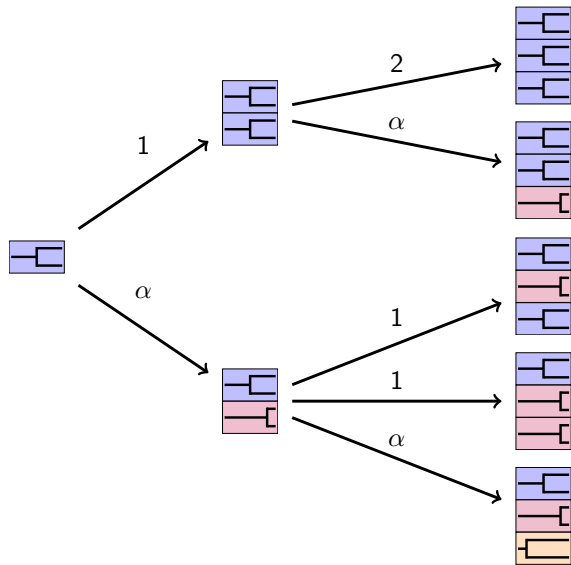


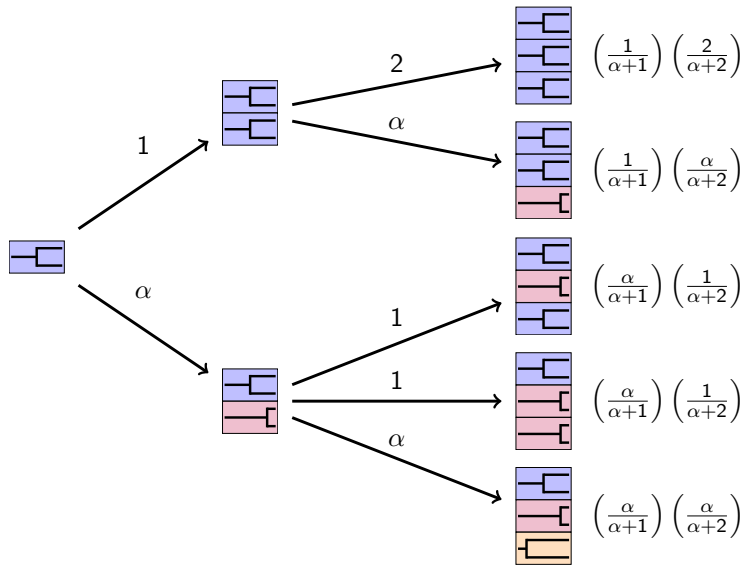
© 2007 Boris Kulikov boris-kulikov.blogspot.com



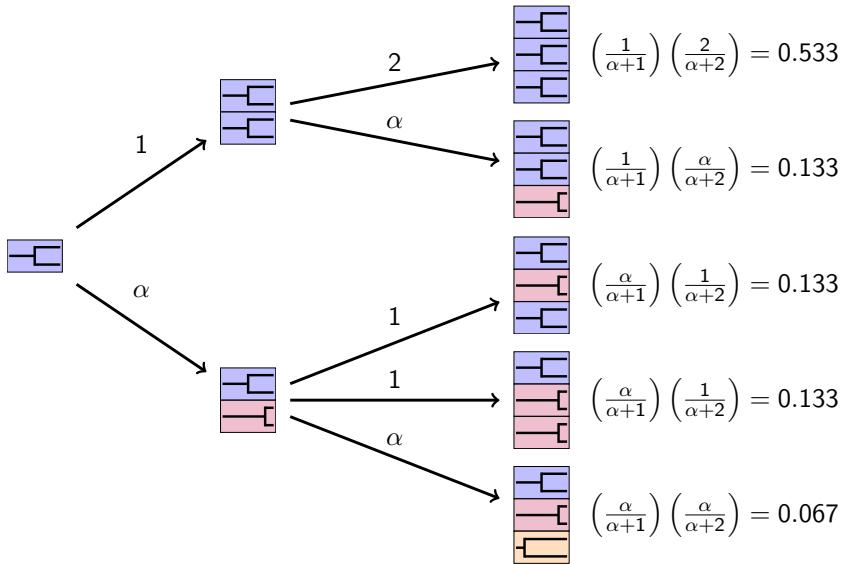




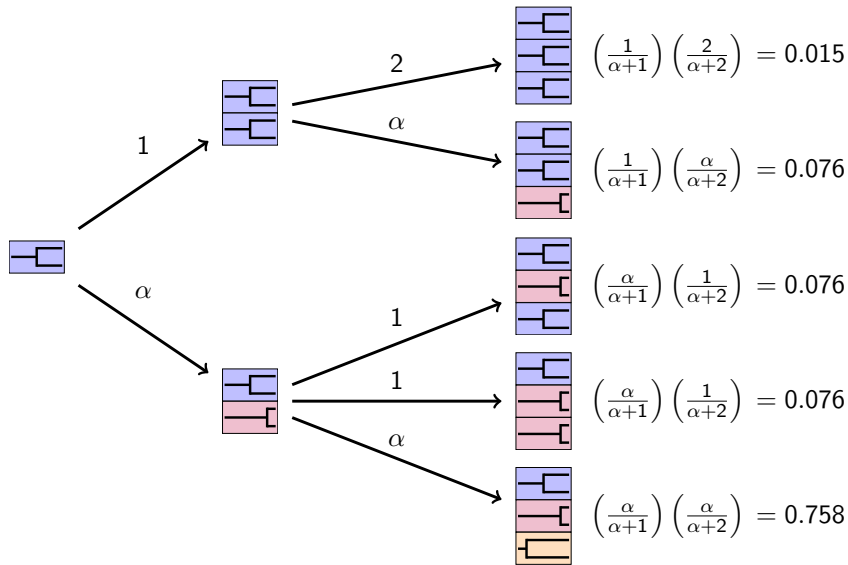




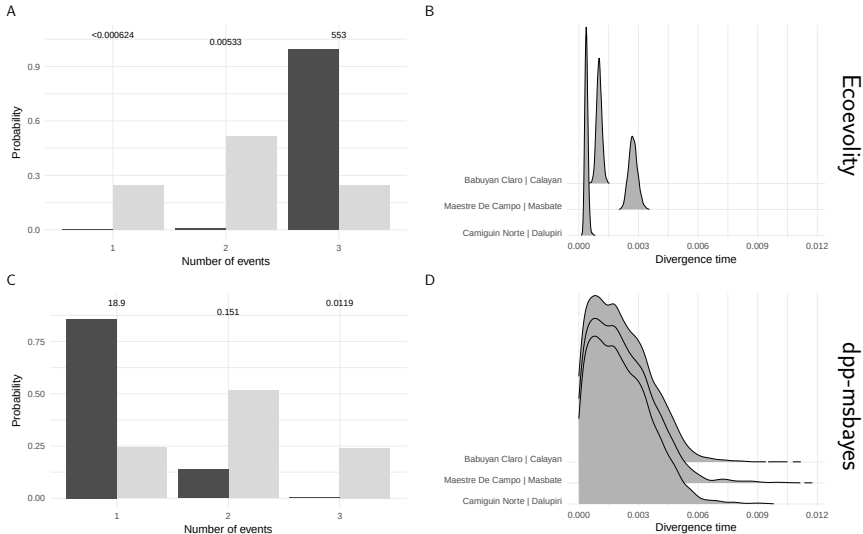
$\alpha = 0.5$



$$\alpha = 10.0$$



ABC vs ecoevolity results



Why so different?



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

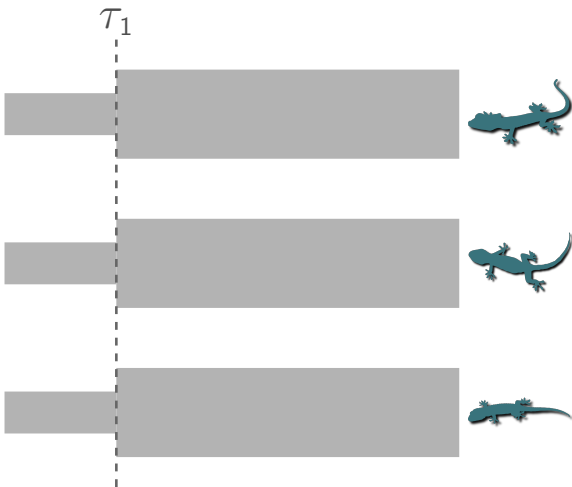


Interactive coin-flipping demo

tl;dr:

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

An aside for a related problem with population demography

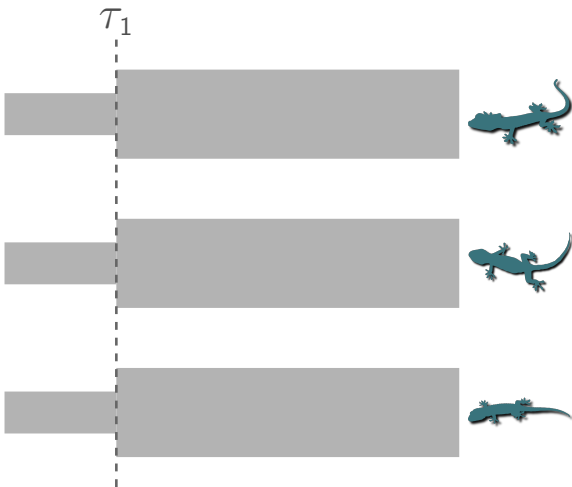


► Recent interest in testing shared demographic changes

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

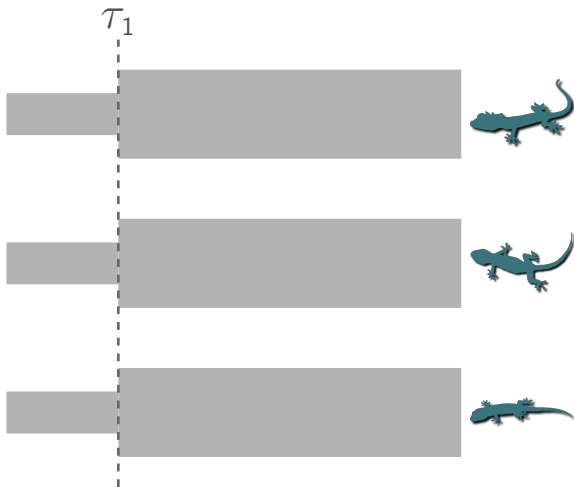


- ▶ Recent interest in testing shared demographic changes
- ▶ Several nice ABC approaches^{1,2,3}
 - ▶ Inferred simultaneous expansion of 5 Alaskan populations of sticklebacks (posterior probability = 0.99)²

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

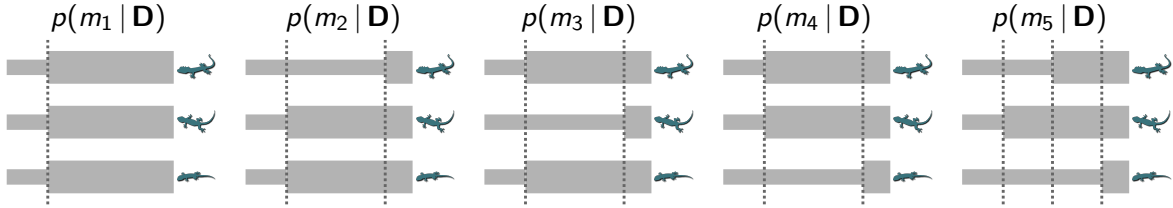


- ▶ Recent interest in testing shared demographic changes
- ▶ Several nice ABC approaches^{1,2,3}
 - ▶ Inferred simultaneous expansion of 5 Alaskan populations of sticklebacks (posterior probability = 0.99)²
- ▶ It's a tricky inference problem
 - ▶ Change in population size can become unidentifiable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

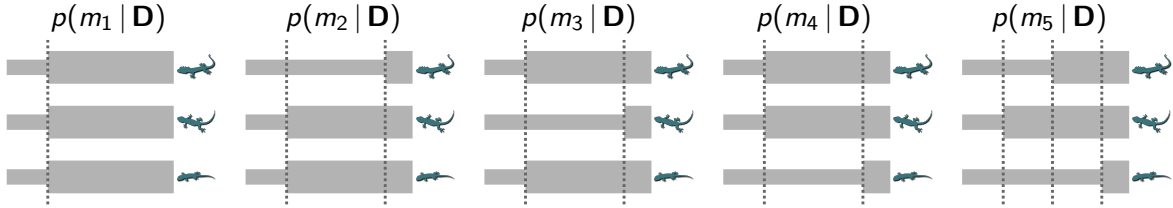
¹ Y. L. Chan, D. Schanzenbach, and M. J. Hickerson (2014). *Molecular Biology and Evolution* 31: 2501–2515

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

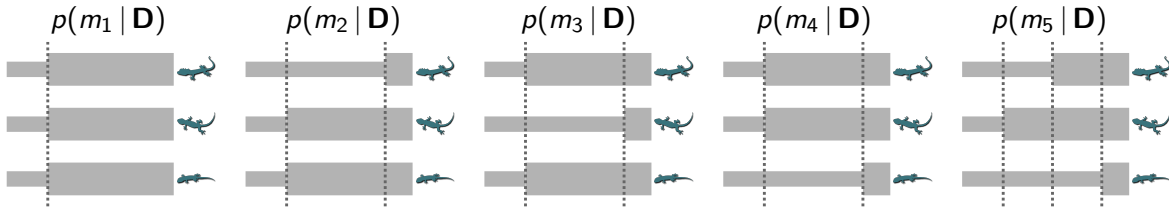


Given genomic data, can we infer the correct model and the timing of the demographic events?



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



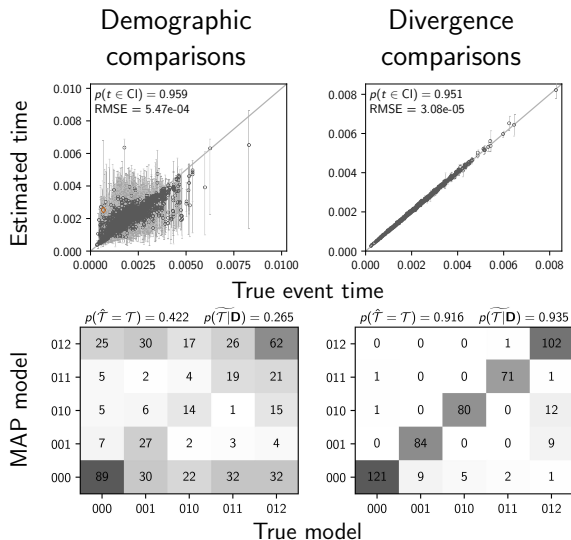


Given genomic data, can we infer the correct model and the timing of the demographic events?



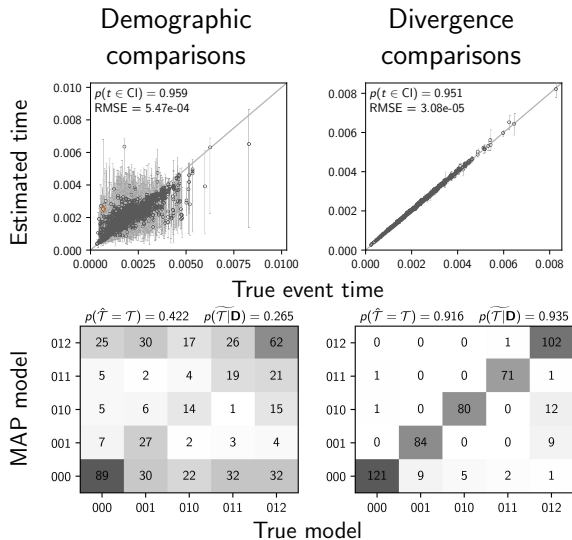
- ▶ Implemented full-likelihood Bayesian approach in evoevolity
- ▶ Assessed performance with simulations
- ▶ Applied to stickleback genomic data

Simulation results



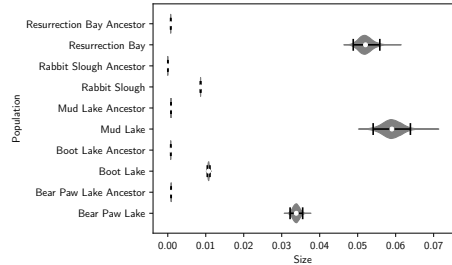
Simulation results

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



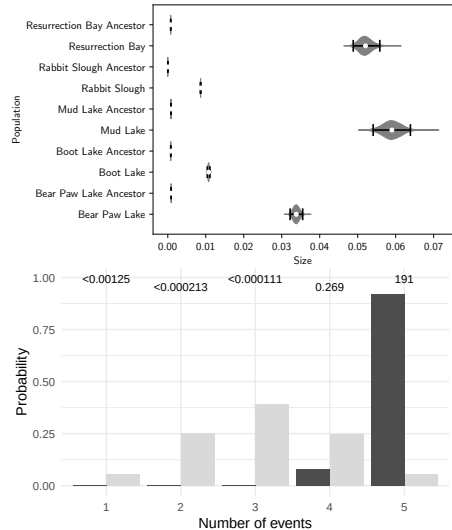
Stickleback results

- Strong support for expansions in all 5 populations; consistent with ABC results



Stickleback results

- ▶ Strong support for expansions in all 5 populations; consistent with ABC results
- ▶ Strong support all 5 expansions were independent; exact opposite of ABC results



Open science: everything is available...

Software:

► Ecoevolity: phyletica.org/ecoevolity

Open-Science Notebooks:

github.com/phyletica/ecoevolity-demog-experiments