Bayesian statistical approaches to inferring shared evolutionary events

Jamie Oaks
Auburn University
phyletica.org
@jamoaks

Scan for slides:

phyletica.org/slides/bms.pdf

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
- *Breanna Sipley*
- *Aundrea Westfall*

Undergraduate students

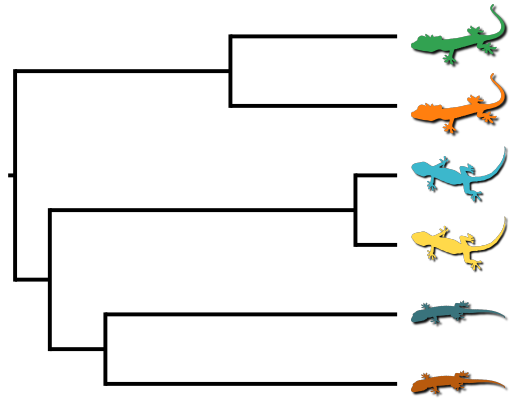
- Laura Lewis
- Mary Wells
- Hailey Whitaker
- Noah Yawn
- *Charlotte Benedict*
- *Eric Carbo*
- *Ryan Cook*
- *Andrew DeSana*
- *Miles Horne*
- *Jacob Landrum*
- *Nadia L'Bahy*
- *Jorge Lopez-Perez*
- *Holden Smith*
- *Virginia White*
- *Kayla Wilson*

Generalizing Bayesian phylogenetics to infer patterns predicted by processes of diversification

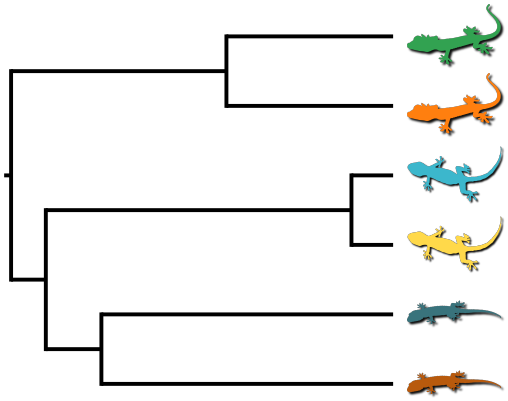
- Phylogenetics is rapidly becoming the statistical foundation of biology

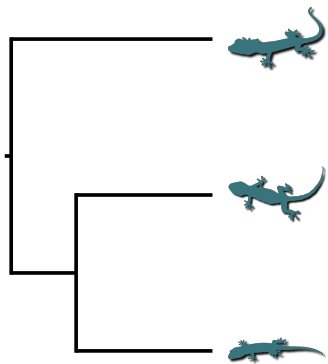
- Phylogenetics is rapidly becoming the statistical foundation of biology
- "Big data" present exciting possibilities and challenges

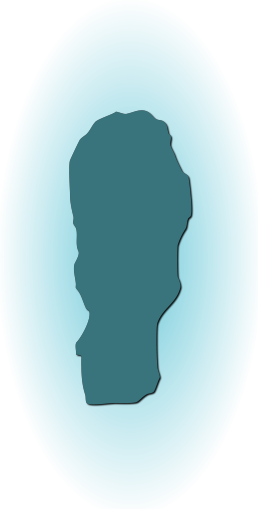
- Phylogenetics is rapidly becoming the statistical foundation of biology
- "Big data" present exciting possibilities and challenges
- Many opportunities to develop new ways to study biology in light of phylogeny

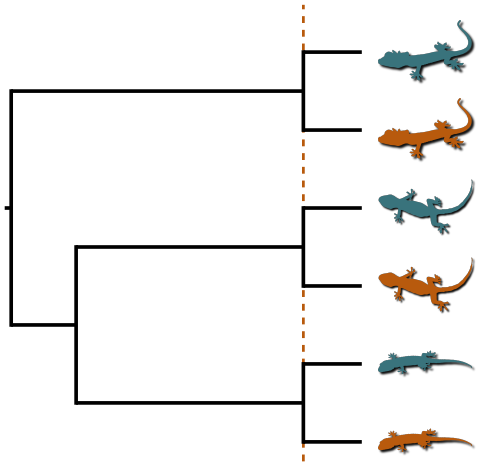


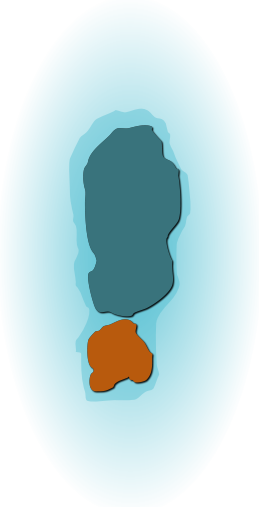
- **Assumption:** All processes of diversification affect each lineage independently





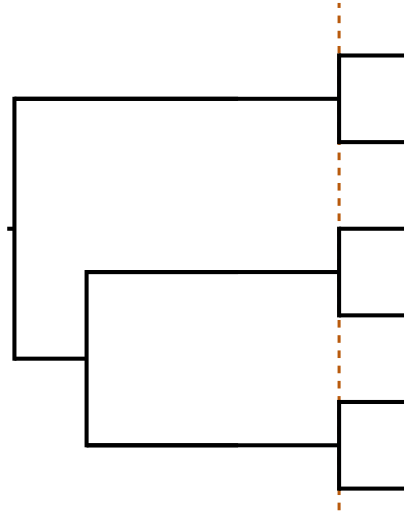






Biogeography

- Environmental changes that affect whole communities of species

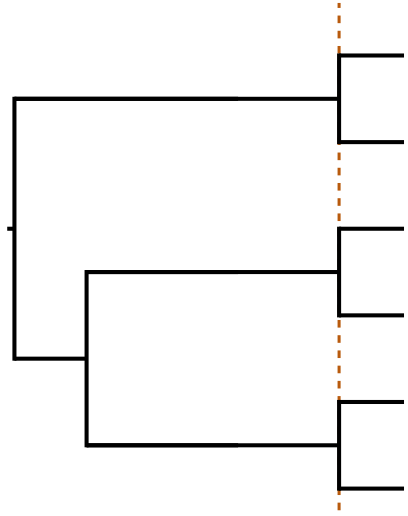


Biogeography

- Environmental changes that affect whole communities of species

Genome evolution

- Duplication of a chromosome segment harboring gene families



Biogeography

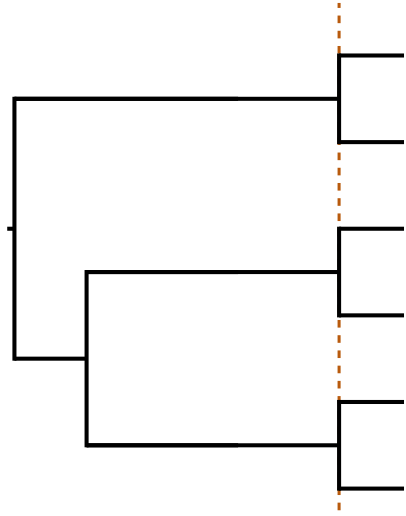
- Environmental changes that affect whole communities of species

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

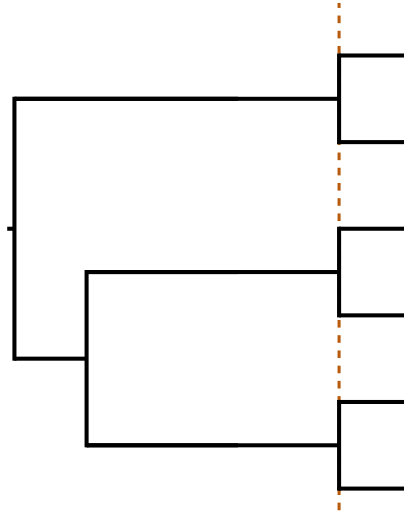
- Duplication of a chromosome segment harboring gene families

Epidemiology

- Transmission at social gatherings

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

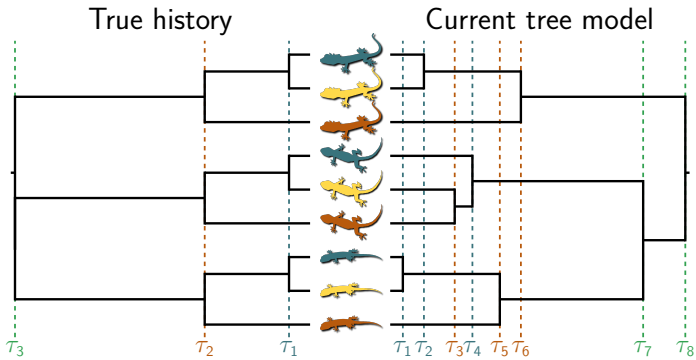
- Speciation of the host
- Co-colonization of new host species



Why account for shared divergences?

Why account for shared divergences?

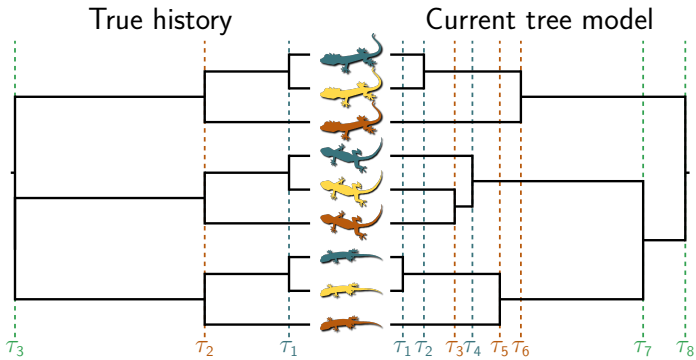
1. Improve inference



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

Why account for shared divergences?

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



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

Biogeography

- Environmental changes that affect whole communities of species

Genome evolution

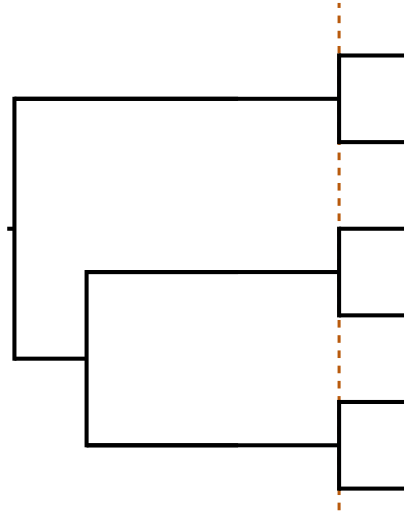
- Duplication of a chromosome segment harboring gene families

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”)
- A fully phylogenetic approach

Approaches to the problem

- A pairwise approach (keep it "simple")
- A fully phylogenetic approach

Tashitso Anamza

Tanner Myers

Randy Klabacka

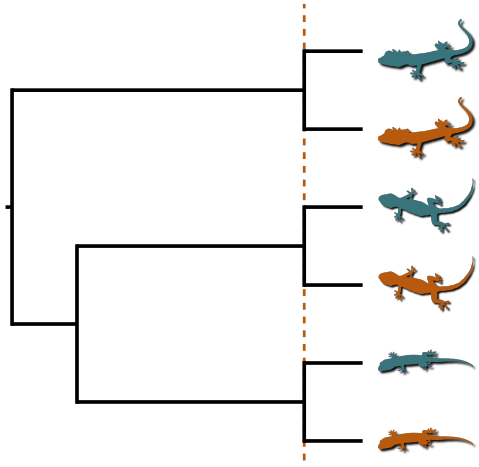
Perry Wood, Jr.

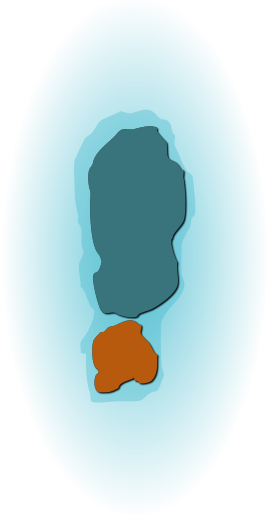
Claire Tracy

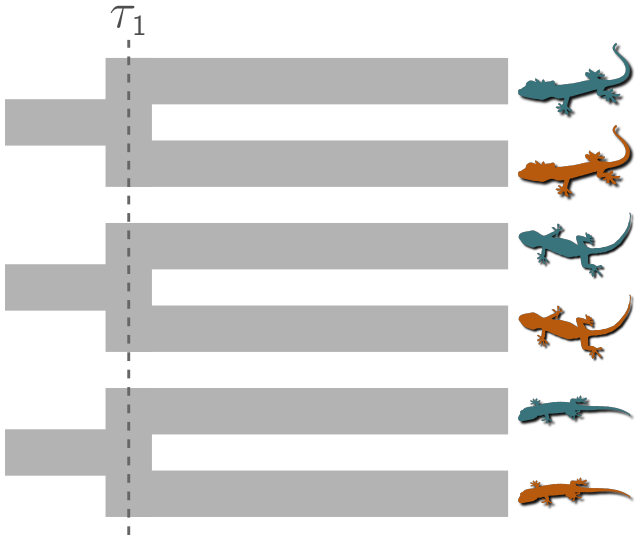
Kerry Cobb

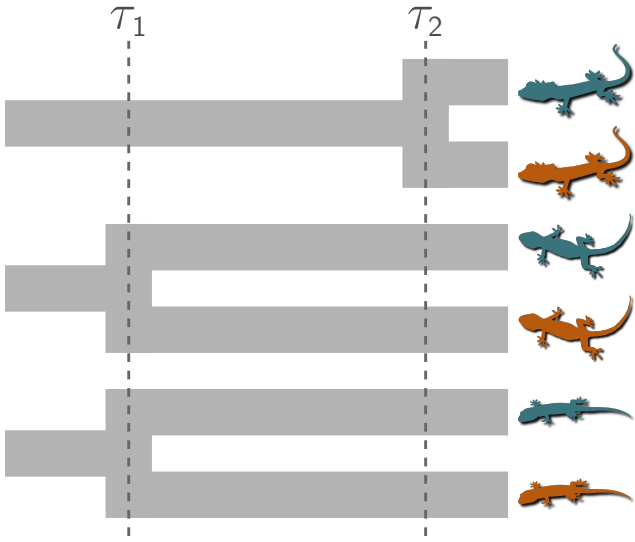
Matt Buehler

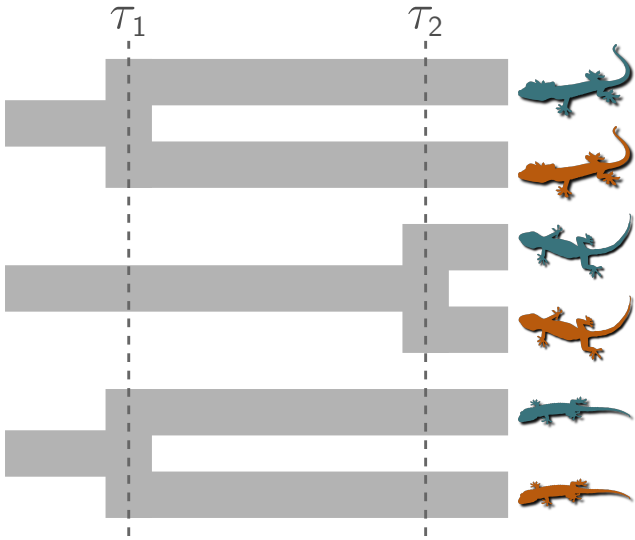
Nadia L'bahy

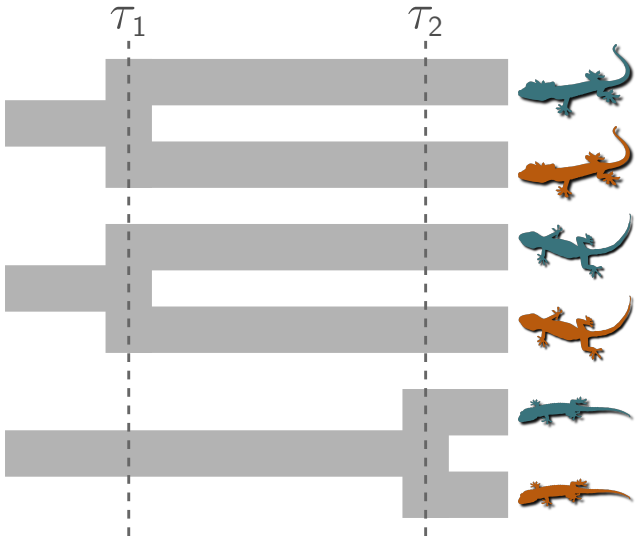


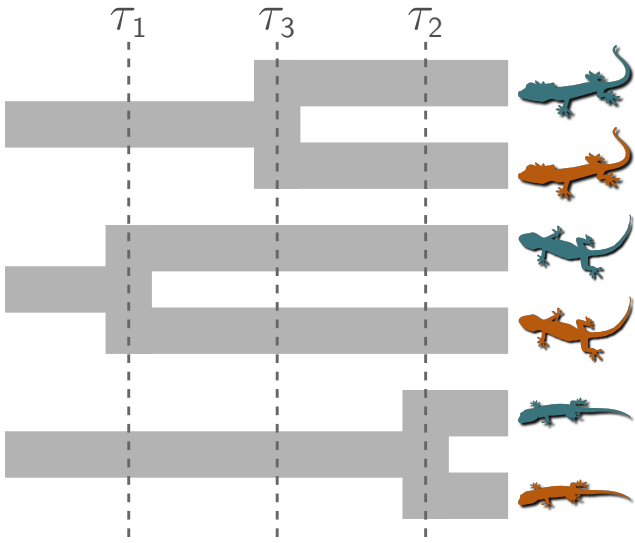


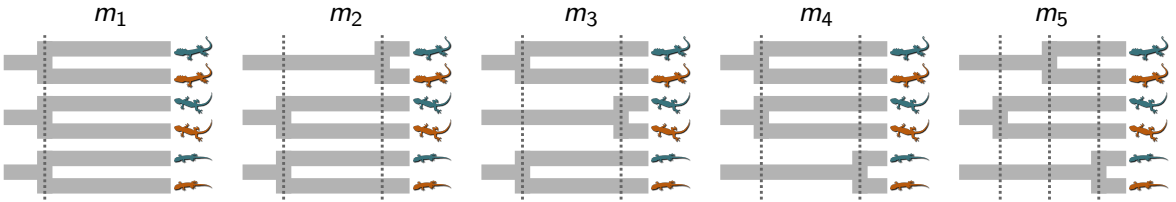

τ_1


τ_1
τ_2

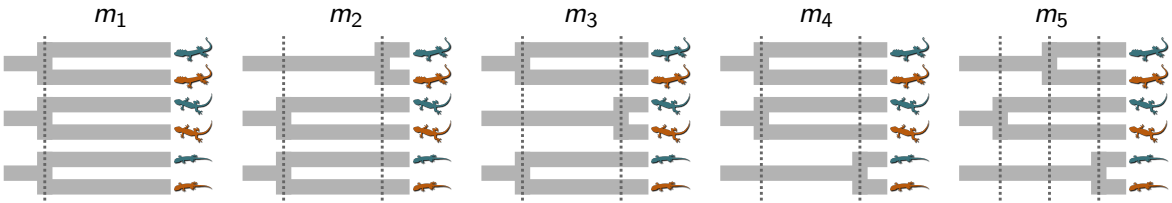

τ_1
τ_2


τ_1
τ_2


τ_1
τ_3
τ_2



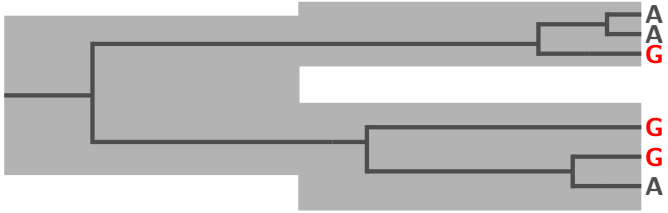
J. R. Oaks et al. (2013). *Evolution* 67: 991–1010, J. R. Oaks (2014). *BMC Evolutionary Biology* 14: 150

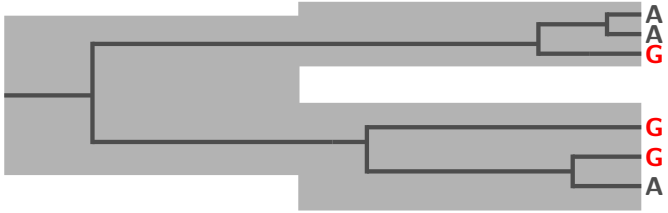


We want to infer the model and divergence times given genetic data

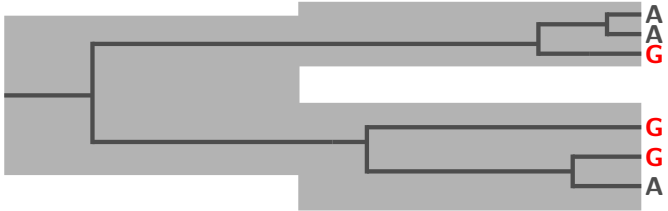
J. R. Oaks et al. (2013). *Evolution* 67: 991–1010, J. R. Oaks (2014). *BMC Evolutionary Biology* 14: 150


A
A
G
G
G
A

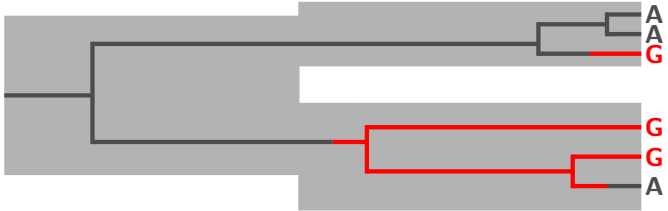

A
A
G
G
G
A



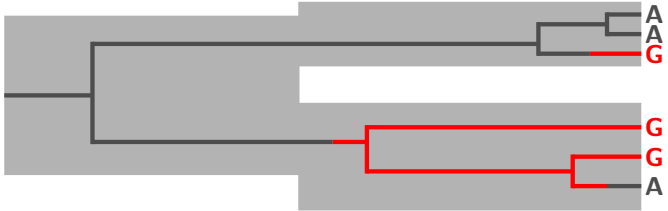
- Conditional on "population tree", model "gene tree" using coalescent
  - Coalescent is a stochastic model of shared inheritance (continuous-time Markov chain = CTMC)



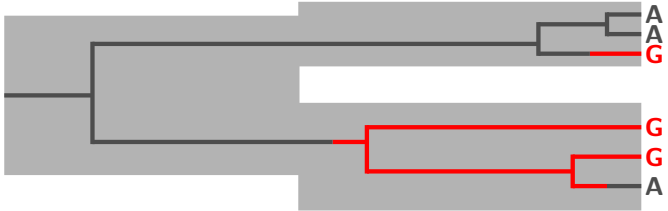
- Conditional on "population tree", model "gene tree" using coalescent
  - Coalescent is a stochastic model of shared inheritance (continuous-time Markov chain = CTMC)
  - Branching pattern is a function of population size



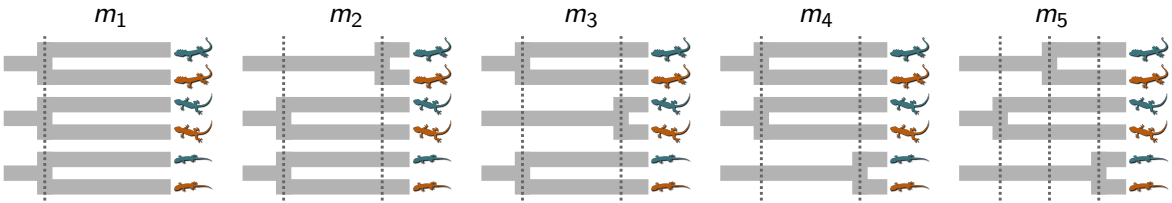
- Conditional on "population tree", model "gene tree" using coalescent
  - Coalescent is a stochastic model of shared inheritance (continuous-time Markov chain = CTMC)
  - Branching pattern is a function of population size
- Conditional on gene tree, model mutation as a CTMC



- 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

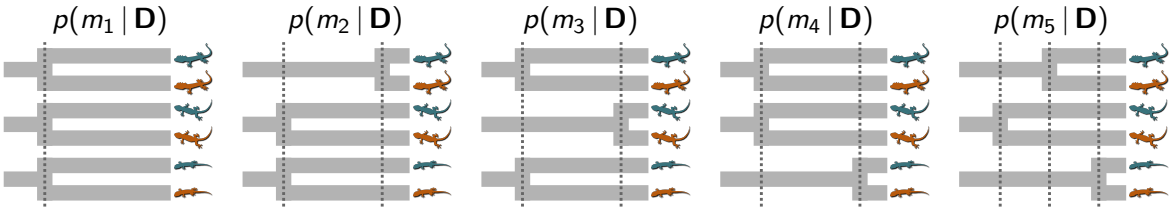


- Conditional on "population tree", model "gene tree" using coalescent
  - Coalescent is a stochastic model of shared inheritance (continuous-time Markov chain = CTMC)
  - Branching pattern is a function of population size
- Conditional on gene tree, model mutation as a CTMC
- Genetic characters provide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



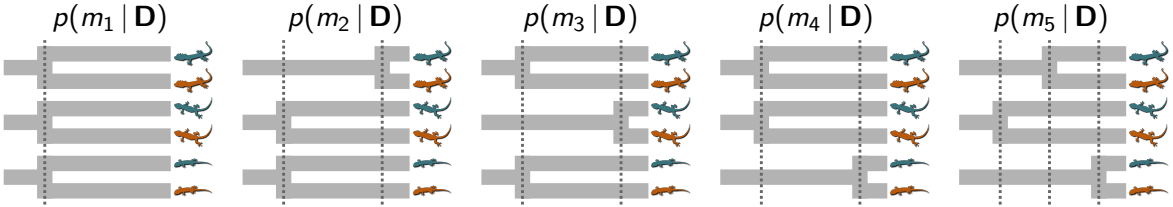
We want to infer the model and divergence times given genetic data

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

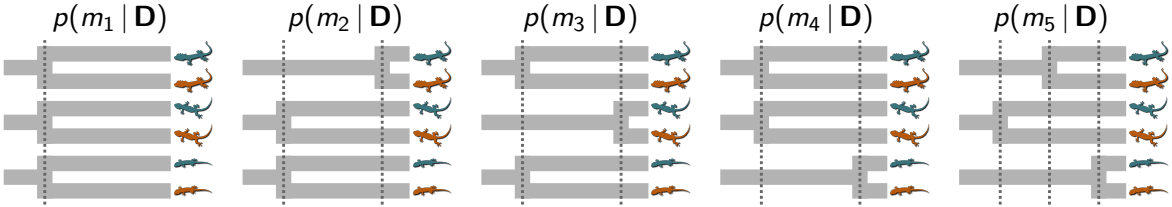
J. R. Oaks et al. (2013). *Evolution* 67: 991–1010, J. R. Oaks (2014). *BMC Evolutionary Biology* 14: 150



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

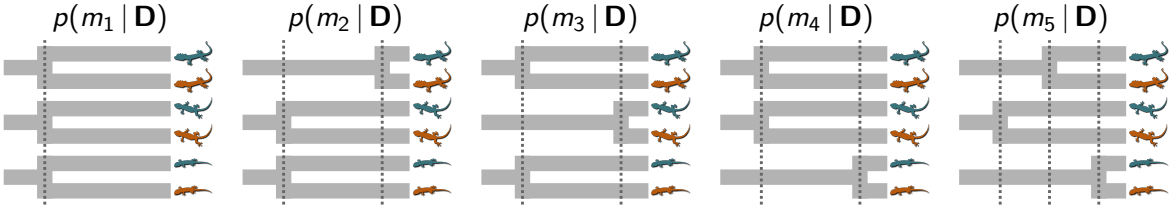


We want to infer the model and divergence times given genetic data

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

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

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



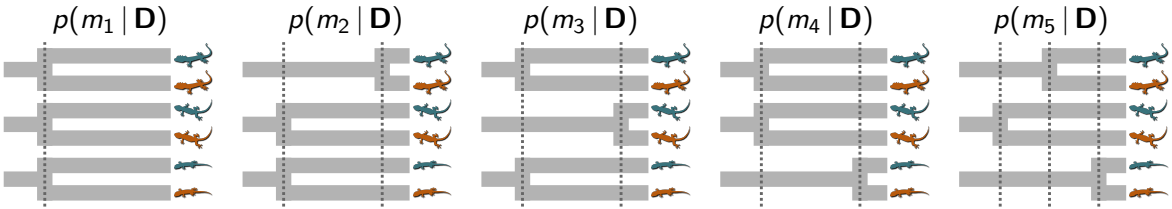
We want to infer the model and divergence times given genetic data

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

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

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

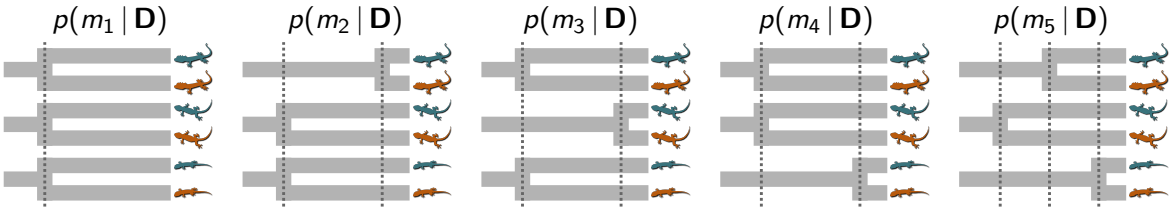
J. R. Oaks et al. (2013). *Evolution* 67: 991–1010, J. R. Oaks (2014). *BMC Evolutionary Biology* 14: 150



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

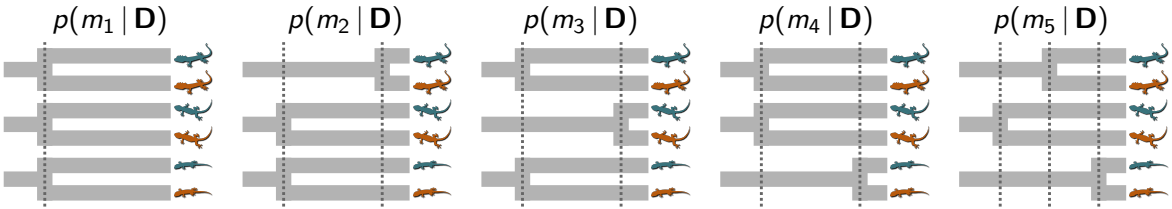


We want to infer the model and divergence times given genetic data

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

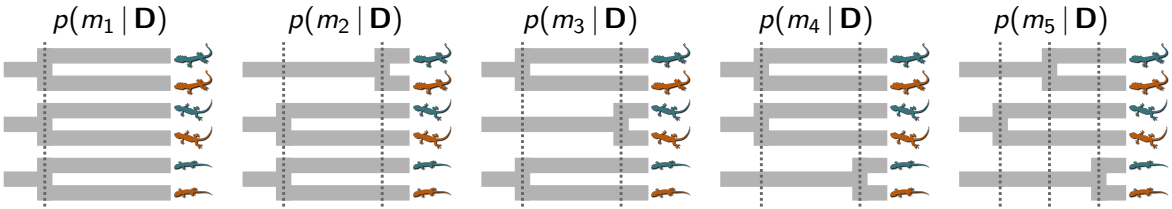


We want to infer the model and divergence times given genetic data

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

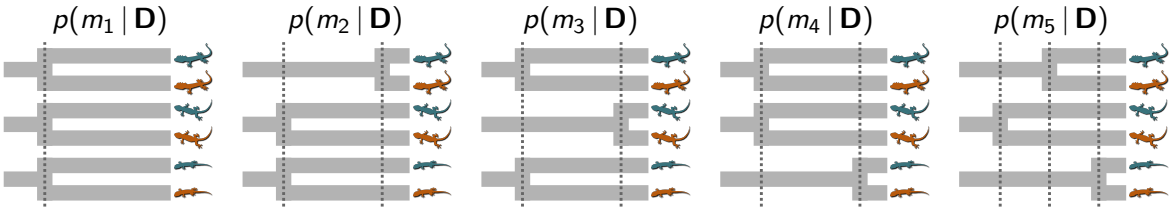


We want to infer the model and divergence times given genetic data

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

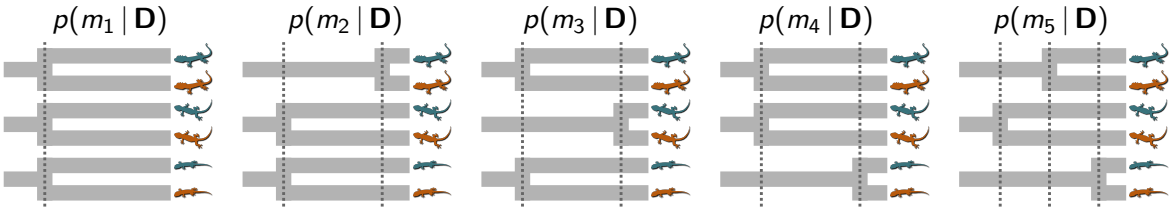


We want to infer the model and divergence times given genetic data

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

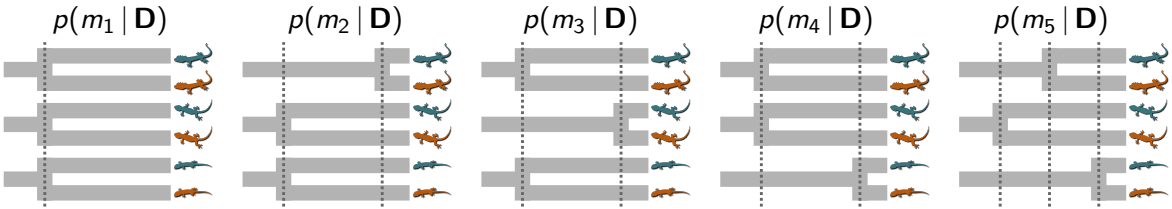


We want to infer the model and divergence times given genetic data

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



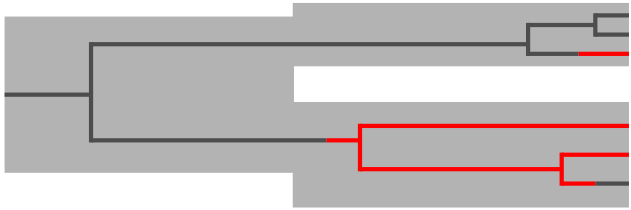
We want to infer the model and divergence times given genetic data

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

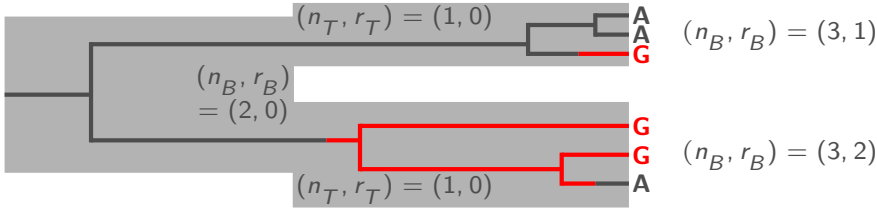


$(n_B, r_B) = (3, 1)$

$(n_B, r_B) = (3, 2)$

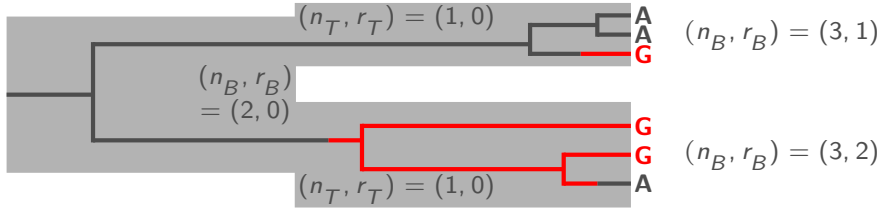
[1] T. Schmelzer and L. N. Trefethen (2007). *Electronic Transactions on Numerical Analysis* 29: 1–18

[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

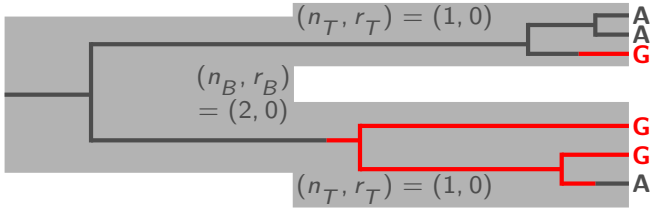
[2] D. Bryant et al. (2012). *Molecular Biology and Evolution* 29: 1917–1932



$$
Q = \begin{array}{c|ccccccc}
 & (1,0) & (1,1) & (2,0) & (2,1) & \cdots & (\mathbf{n},\mathbf{n}) \\
\hline
(1,0) & \cdot & \cdot & \cdot & \cdot & & \cdot \\
(1,1) & \cdot & \cdot & \cdot & \cdot & & \cdot \\
(2,0) & \cdot & \cdot & \cdot & \cdot & & \cdot \\
(2,1) & \cdot & \cdot & \cdot & \cdot & & \cdot \\
\vdots & & & & & & \\
(\mathbf{n},\mathbf{n}) & \cdot & \cdot & \cdot & \cdot & & \cdot
\end{array}
$$

[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

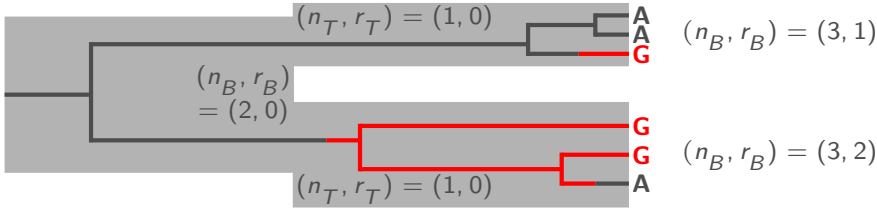


$$
Q = \begin{array}{c|ccccccc}
 & (1,0) & (1,1) & (2,0) & (2,1) & \cdots & (\mathbf{n},\mathbf{n}) \\
(1,0) & \cdot & \cdot & \cdot & \cdot & & \cdot \\
(1,1) & \cdot & \cdot & \cdot & \cdot & & \cdot \\
(2,0) & \cdot & \cdot & \cdot & \cdot & & \cdot \\
(2,1) & \cdot & \cdot & \cdot & \cdot & & \cdot \\
\vdots & & & & & & \\
(\mathbf{n},\mathbf{n}) & \cdot & \cdot & \cdot & \cdot & & \cdot
\end{array}
$$

$$
\begin{aligned}
Q_{(n,r);(n,r-1)} &= (n-r+1)v, && \text{mutation,} \\
Q_{(n,r);(n,r+1)} &= (r+1)u, && \text{mutation,} \\
Q_{(n,r);(n-1,r)} &= \tfrac{(n-1-r)n}{2N_e(u+v)}, && \text{coalescence,} \\
Q_{(n,r);(n-1,r-1)} &= \tfrac{(r-1)n}{2N_e(u+v)}, && \text{coalescence,} \\
Q_{(n,r);(n,r)} &= -\tfrac{(n-1)n}{2N_e(u+v)} - (n-r)v - ru. &&
\end{aligned}
$$

[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



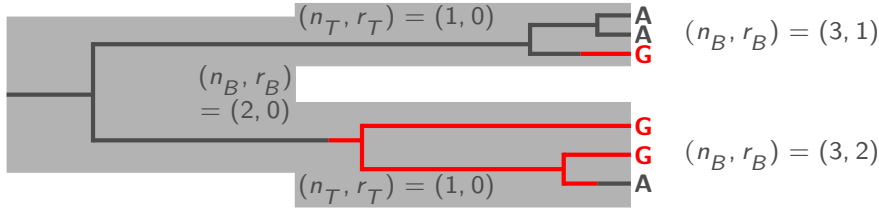
$$
Q = \begin{array}{c|ccccccc}
 & (1,0) & (1,1) & (2,0) & (2,1) & \cdots & (\mathbf{n},\mathbf{n}) \\
\hline
(1,0) & \cdot & \cdot & \cdot & \cdot & & \cdot \\
(1,1) & \cdot & \cdot & \cdot & \cdot & & \cdot \\
(2,0) & \cdot & \cdot & \cdot & \cdot & & \cdot \\
(2,1) & \cdot & \cdot & \cdot & \cdot & & \cdot \\
\vdots & & & & & & \\
(\mathbf{n},\mathbf{n}) & \cdot & \cdot & \cdot & \cdot & & \cdot
\end{array}
$$

$$
\begin{aligned}
Q_{(n,r);(n,r-1)} &= (n-r+1)v, && \text{mutation,} \\
Q_{(n,r);(n,r+1)} &= (r+1)u, && \text{mutation,} \\
Q_{(n,r);(n-1,r)} &= \tfrac{(n-1-r)n}{2N_e(u+v)}, && \text{coalescence,} \\
Q_{(n,r);(n-1,r-1)} &= \tfrac{(r-1)n}{2N_e(u+v)}, && \text{coalescence,} \\
Q_{(n,r);(n,r)} &= -\tfrac{(n-1)n}{2N_e(u+v)} - (n-r)v - ru. &&
\end{aligned}
$$

- e^{Qt} to keep track of all conditional probabilites along each branch (Carathéodory-Fejér method[1])

[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



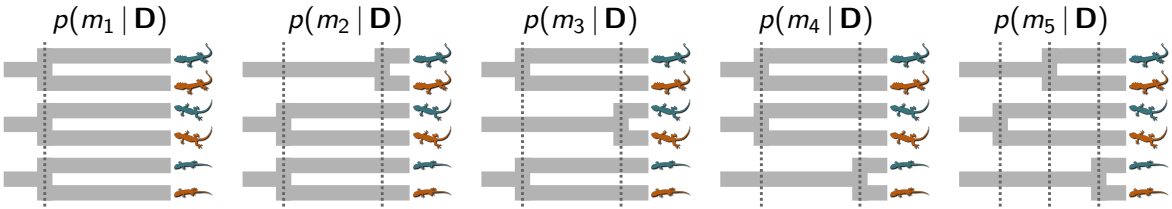
$$
Q = \begin{array}{c|ccccccc}
 & (1,0) & (1,1) & (2,0) & (2,1) & \cdots & (\mathbf{n},\mathbf{n}) \\
\hline
(1,0) & \cdot & \cdot & \cdot & \cdot & & \cdot \\
(1,1) & \cdot & \cdot & \cdot & \cdot & & \cdot \\
(2,0) & \cdot & \cdot & \cdot & \cdot & & \cdot \\
(2,1) & \cdot & \cdot & \cdot & \cdot & & \cdot \\
\vdots & & & & & & \\
(\mathbf{n},\mathbf{n}) & \cdot & \cdot & \cdot & \cdot & & \cdot
\end{array}
$$

$$
\begin{aligned}
Q_{(n,r);(n,r-1)} &= (n-r+1)v, && \text{mutation,} \\
Q_{(n,r);(n,r+1)} &= (r+1)u, && \text{mutation,} \\
Q_{(n,r);(n-1,r)} &= \tfrac{(n-1-r)n}{2N_e(u+v)}, && \text{coalescence,} \\
Q_{(n,r);(n-1,r-1)} &= \tfrac{(r-1)n}{2N_e(u+v)}, && \text{coalescence,} \\
Q_{(n,r);(n,r)} &= -\tfrac{(n-1)n}{2N_e(u+v)} - (n-r)v - ru. &&
\end{aligned}
$$

- e^{Qt} to keep track of all conditional probabilites along each branch (Carathéodory-Fejér method[1])
- At root, get likelihood of population tree integrated over all possible gene trees and mutational histories[2]

[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

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

- 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 **evo**lutionary **coeval**ity

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

[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

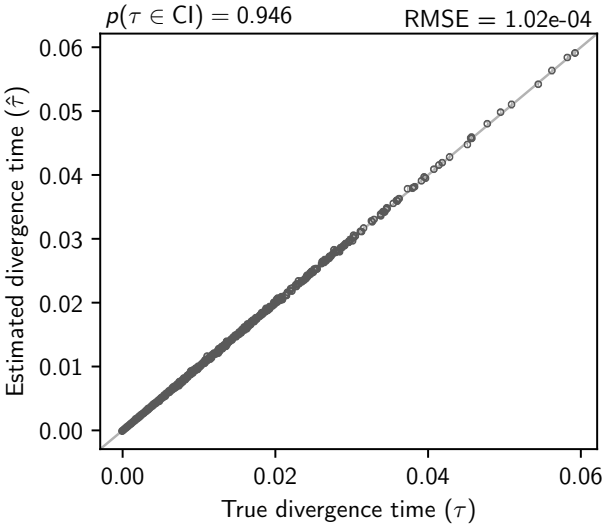
- *Goal: Fast, full-likelihood Bayesian method to infer patterns of co-diversification from genome-scale data*

[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

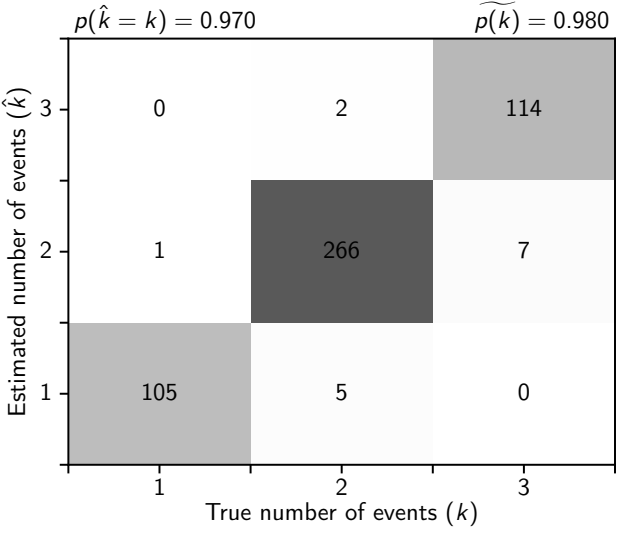
Does it work?

Simulation results



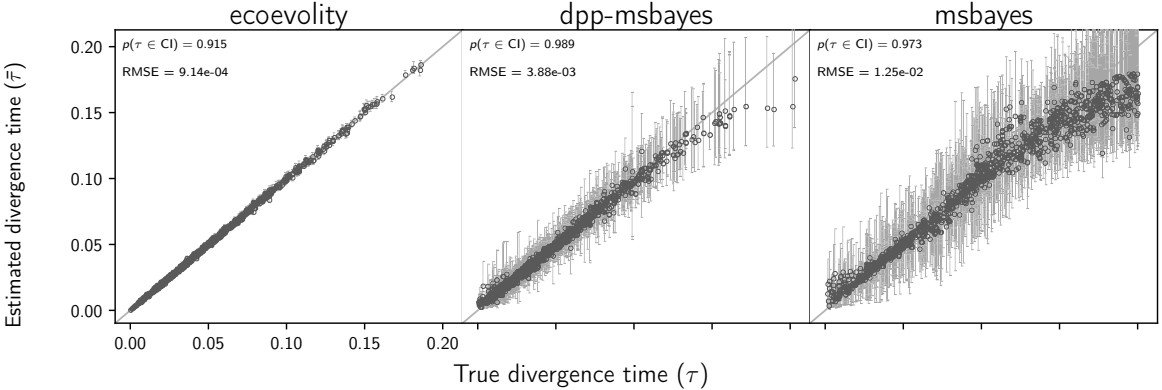
J. R. Oaks (2019). *Systematic Biology* 68: 371–395

Simulation results



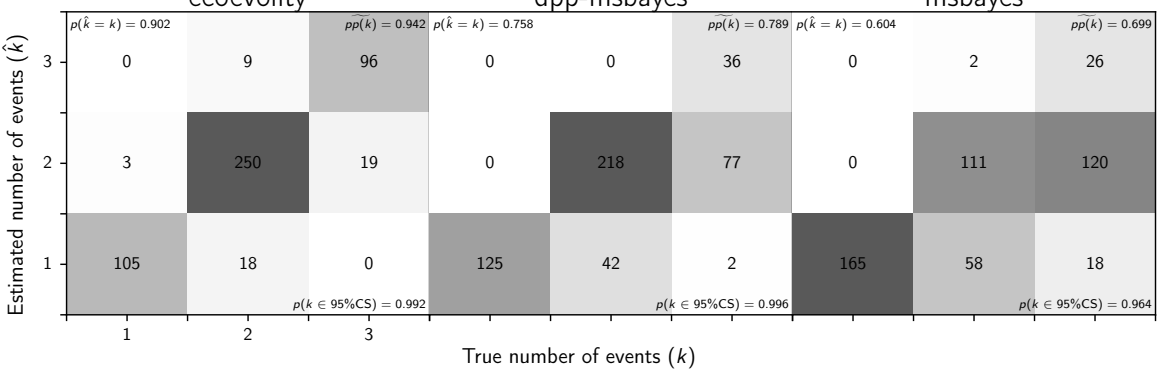
J. R. Oaks (2019). *Systematic Biology* 68: 371–395

“Bake-off” results



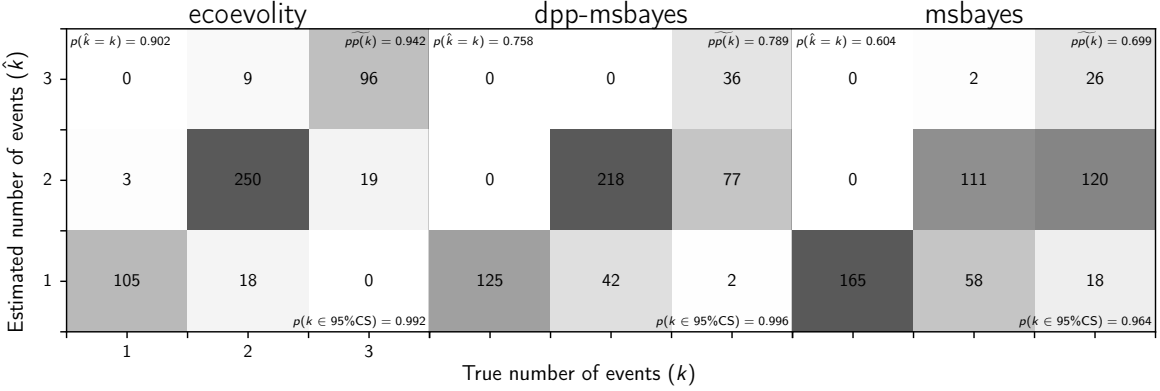
J. R. Oaks (2019). *Systematic Biology* 68: 371–395

“Bake-off” results



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

“Bake-off” results

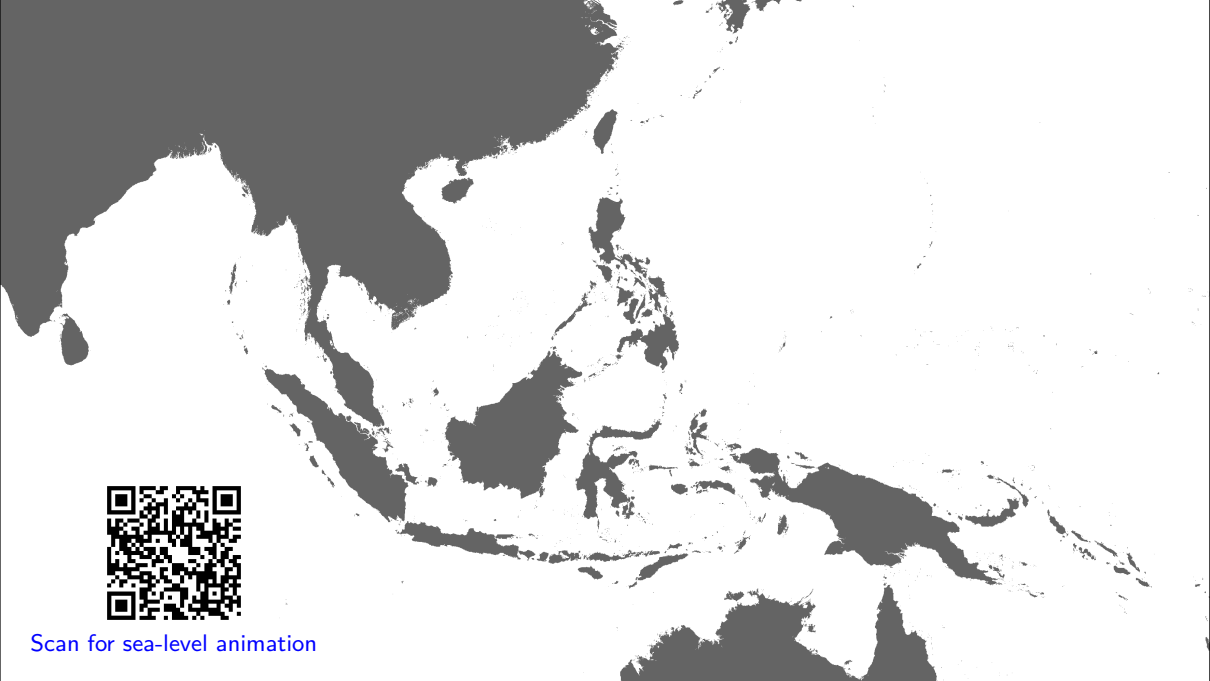


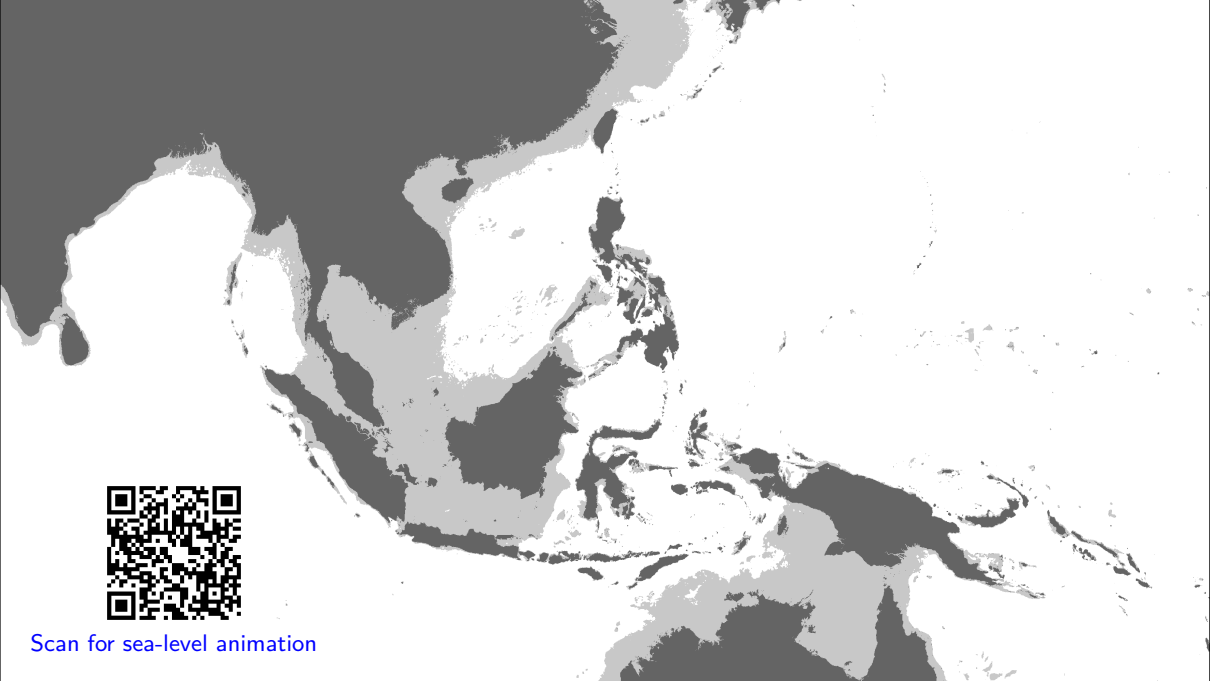
Average run time:

33.4 minutes

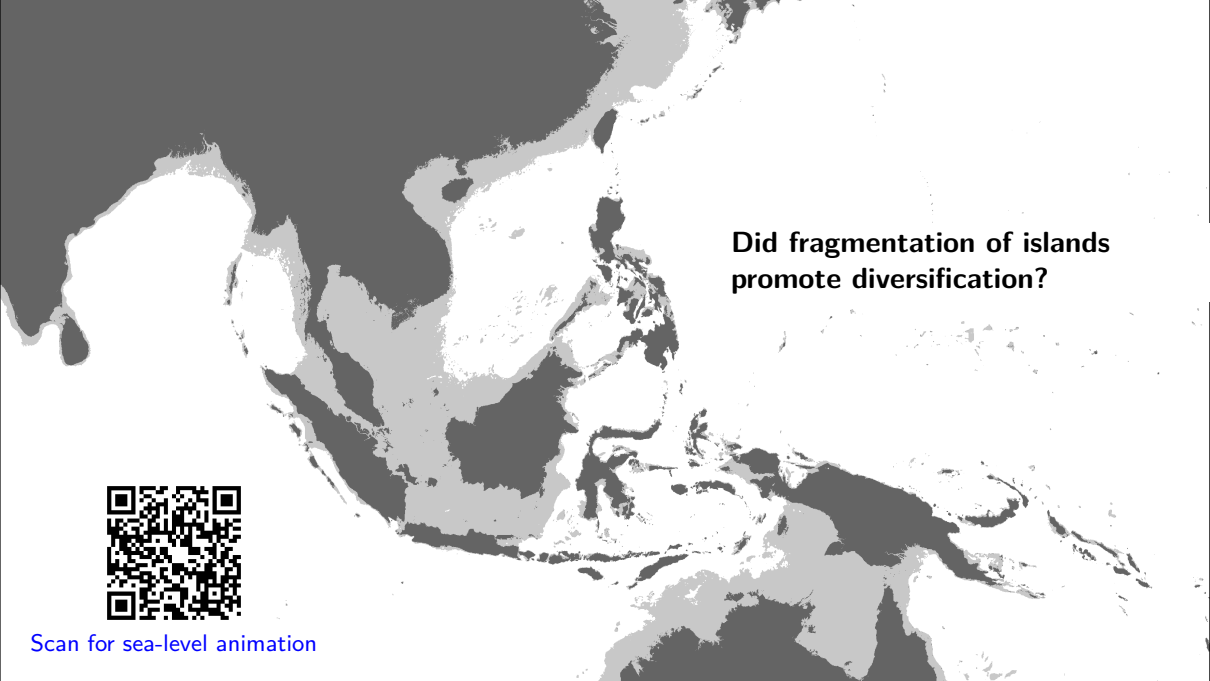
4.4 days

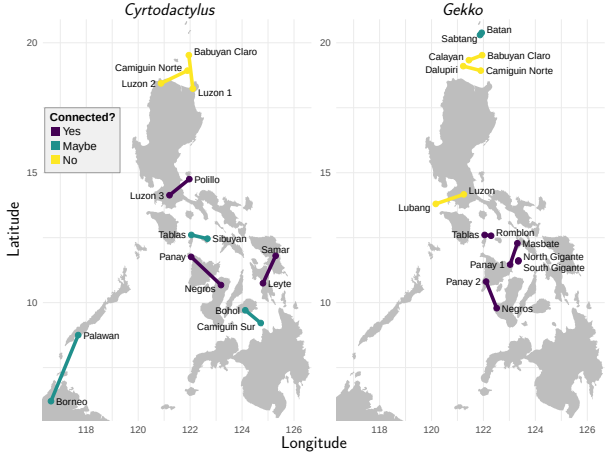
J. R. Oaks (2019). *Systematic Biology* 68: 371–395


Scan for sea-level animation



Scan for sea-level animation

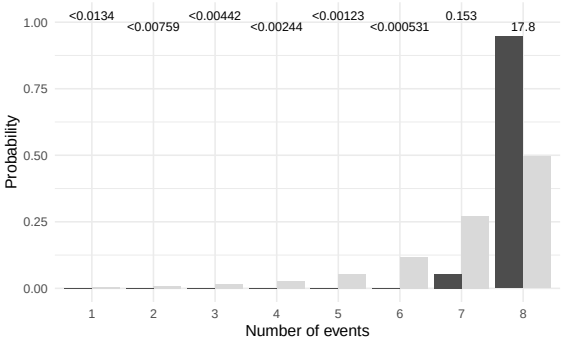

Did fragmentation of islands promote diversification?
Scan for sea-level animation



- 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

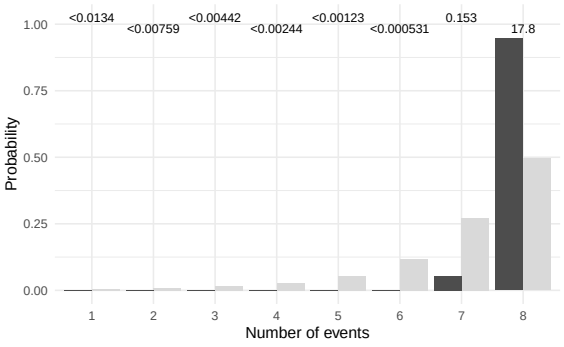
J. R. Oaks, C. D. Siler, and R. M. Brown (2019). *Evolution* 73: 1151–1167

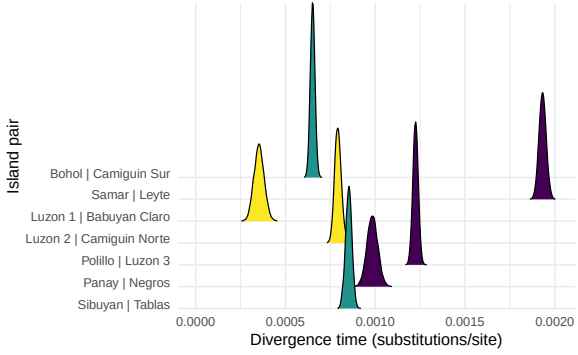
Results: *Cyrtodactylus*



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

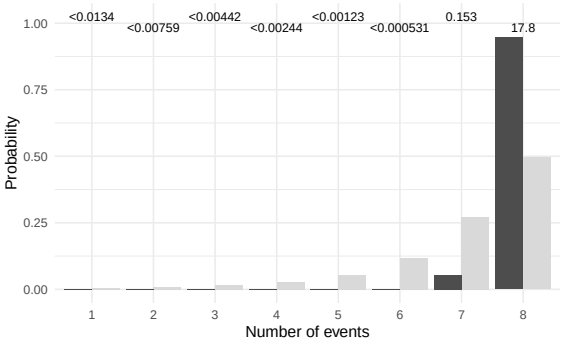
Results: *Cyrtodactylus*

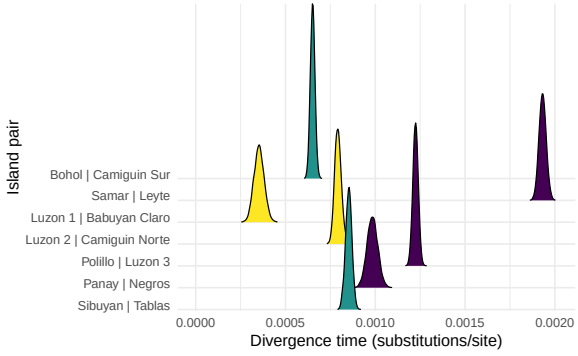




J. R. Oaks, C. D. Siler, and R. M. Brown (2019). *Evolution* 73: 1151–1167

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 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

Dr. Perry Wood, Jr.

Biogeography

- Environmental changes that affect whole communities of species

Genome evolution

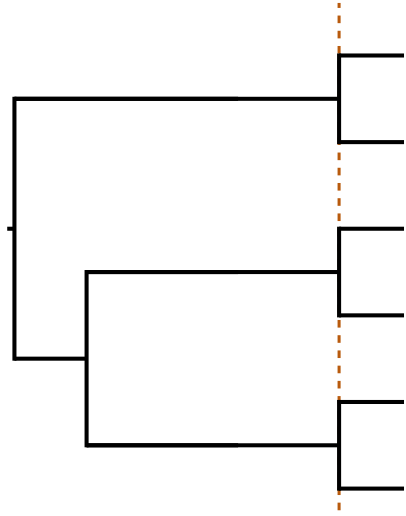
- Duplication of a chromosome segment harboring gene families

Epidemiology

- Transmission at social gatherings

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

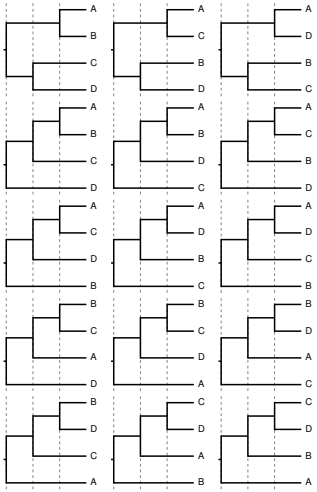
- Speciation of the host
- Co-colonization of new host species



Generalizing tree space

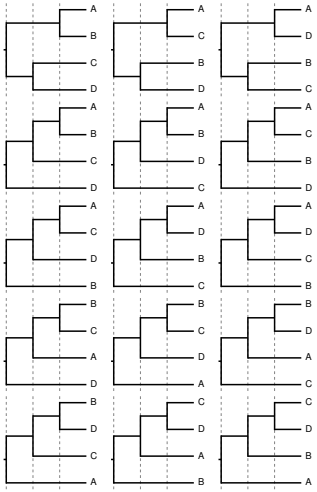
Generalizing tree space

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

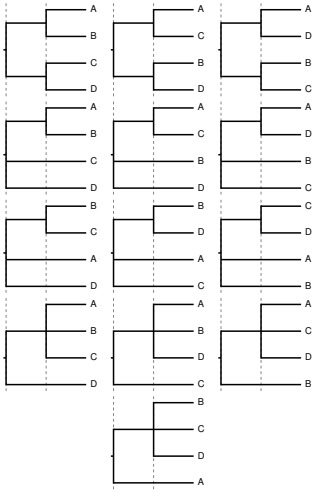


Generalizing tree space

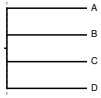
$n_\tau = 3$



$n_\tau = 2$

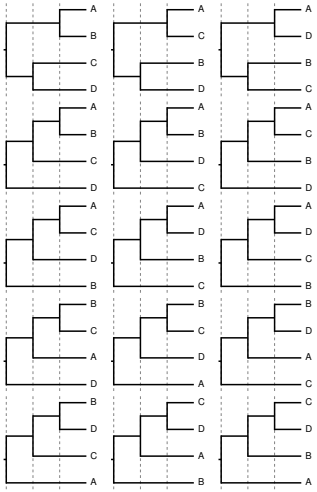


$n_\tau = 1$

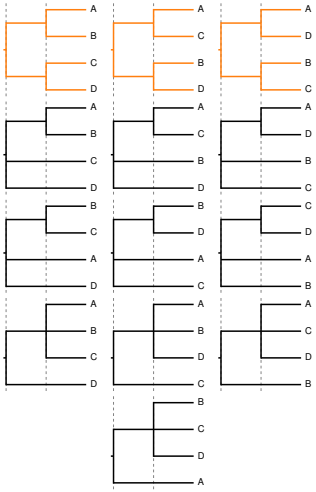


Generalizing tree space

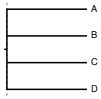
$n_\tau = 3$



$n_\tau = 2$

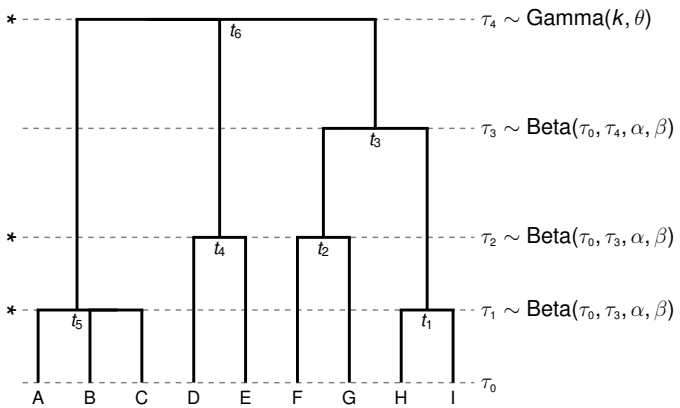


$n_\tau = 1$



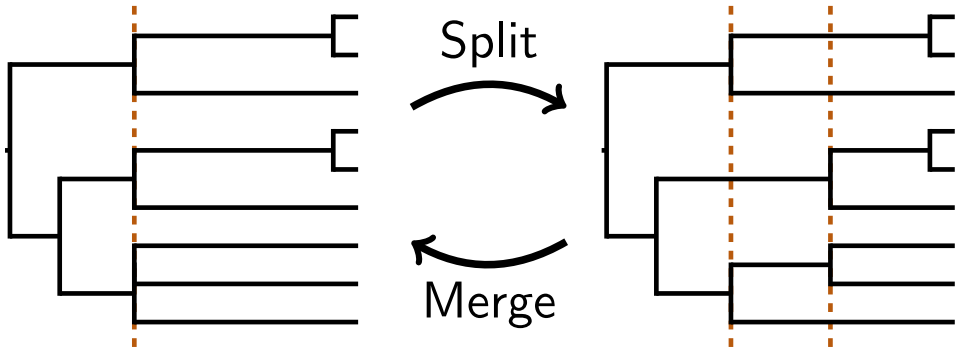
Generalized tree distribution

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



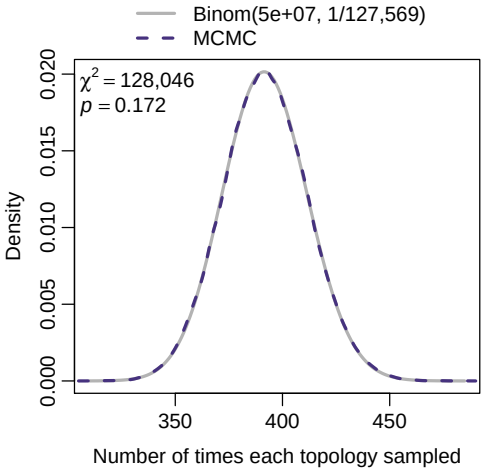
J. R. Oaks et al. (2021). *bioRxiv*

Inferring trees with shared divergences



Reversible-jump MCMC

Validating rjMCMC with 7-leaf tree



The rjMCMC algorithms sample the expected generalized tree distribution

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

PhycoEval

Phylogenetic coevality

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

Ecoevolity

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

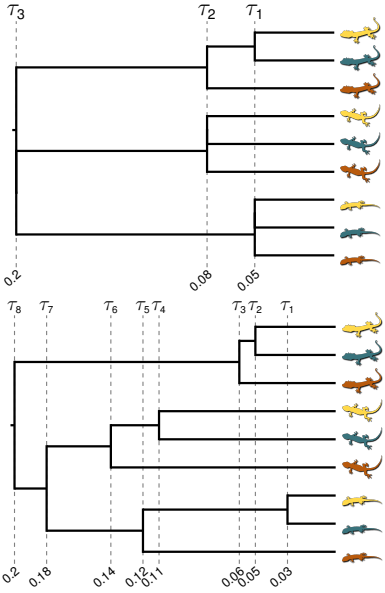
J. R. Oaks (2019). *Systematic Biology* 68: 371–395

- **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

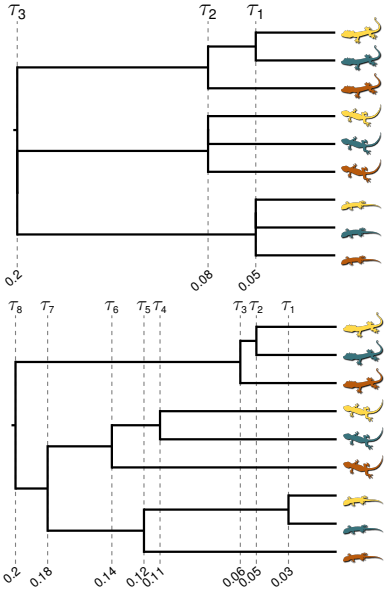
Methods: Simulations

- Simulated 100 data sets with 50,000 base pairs



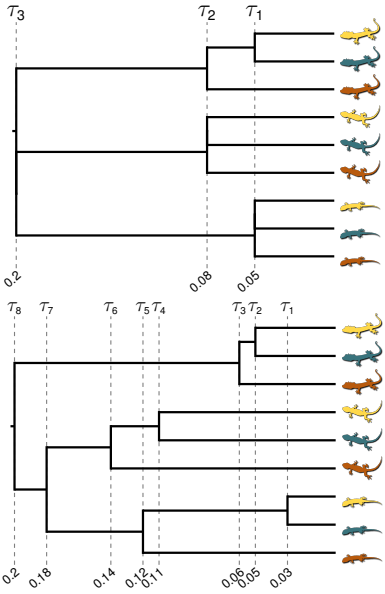
Methods: Simulations

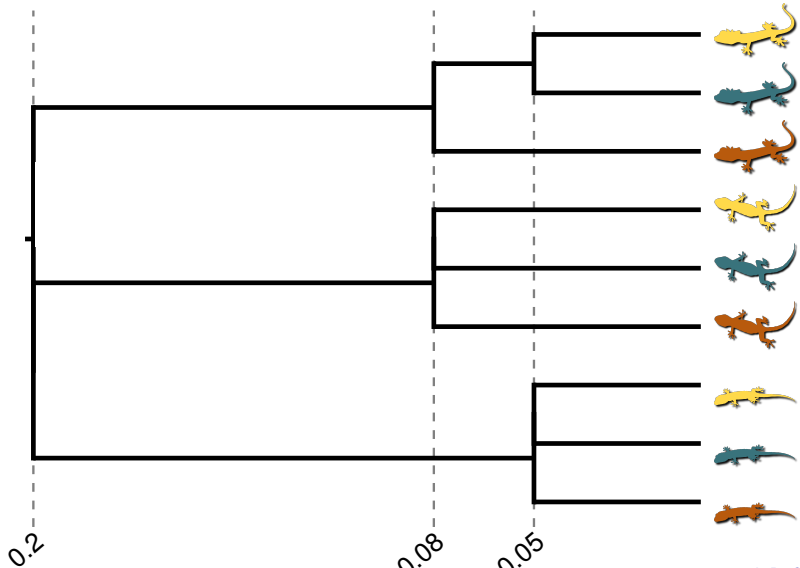
- Simulated 100 data sets with 50,000 base pairs
- Analyzed each data set with:
  - M_G = Generalized tree model
  - M_{IB} = Independent-bifurcating tree model



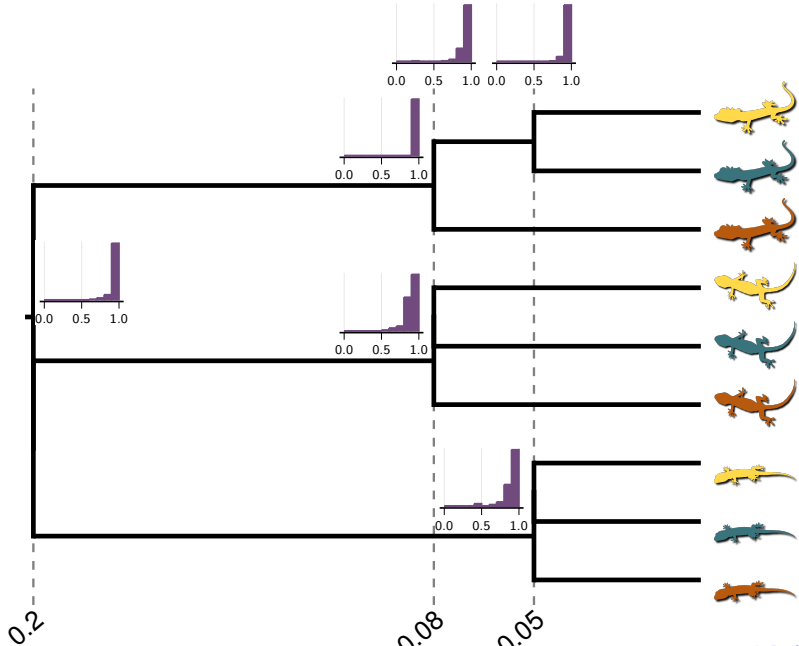
Methods: Simulations

- 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}

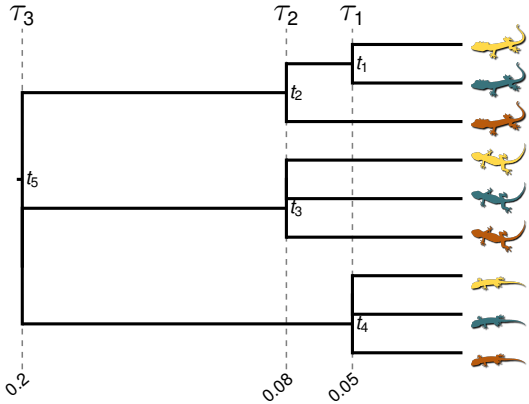




J. R. Oaks et al. (2021). *bioRxiv*

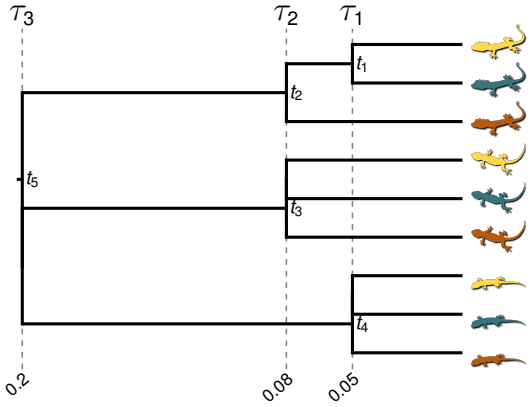


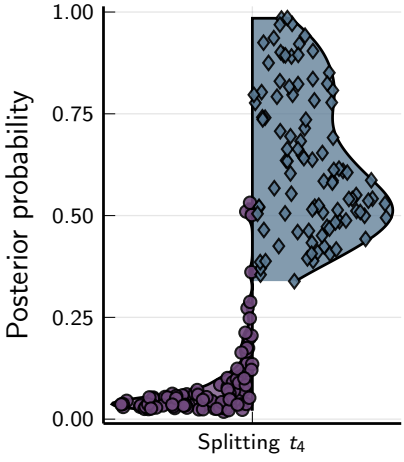
J. R. Oaks et al. (2021). *bioRxiv*



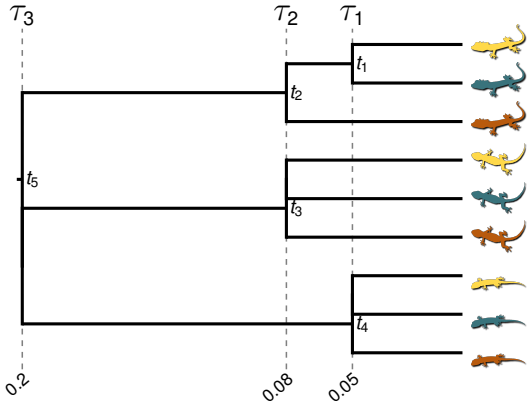
J. R. Oaks et al. (2021). *bioRxiv*

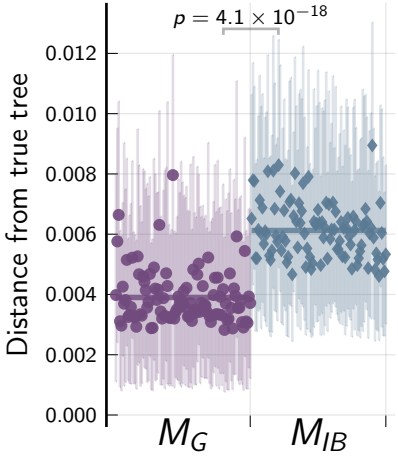
M_G = Generalized model M_{IB} = Independent-bifurcating model





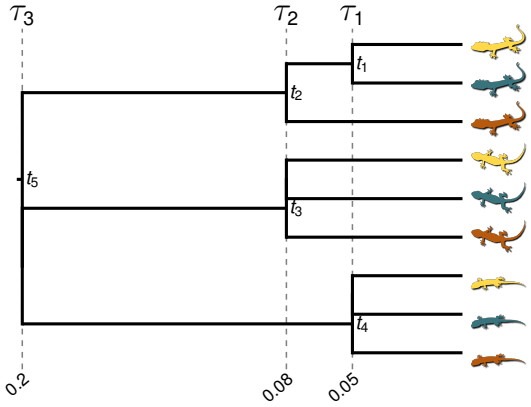
M_G = Generalized model M_{IB} = Independent-bifurcating model

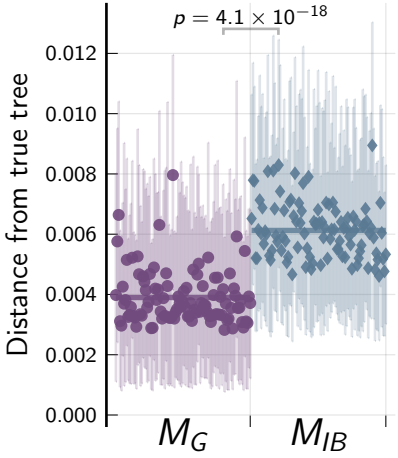




J. R. Oaks et al. (2021). *bioRxiv*

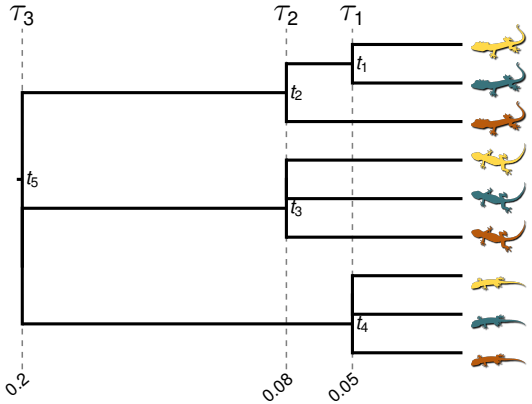
M_G = Generalized model M_{IB} = Independent-bifurcating model

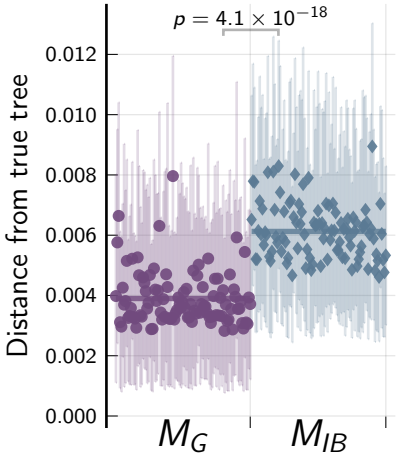




J. R. Oaks et al. (2021). *bioRxiv*

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

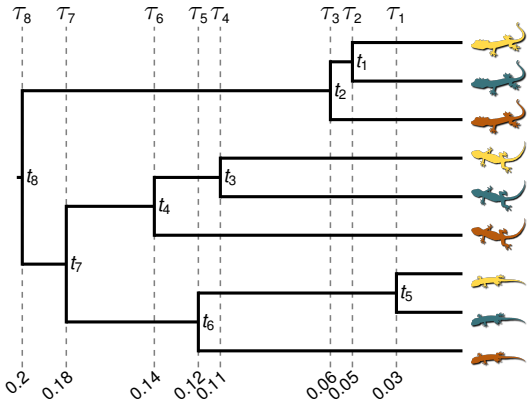




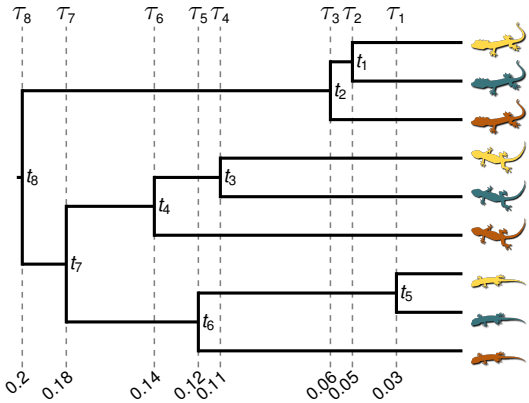
M_G significantly better at inferring trees with shared divergences

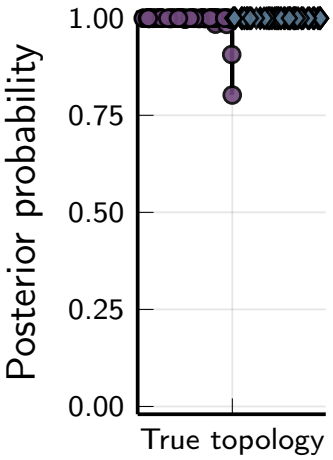
M_G = Generalized model

M_{IB} = Independent-bifurcating model



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

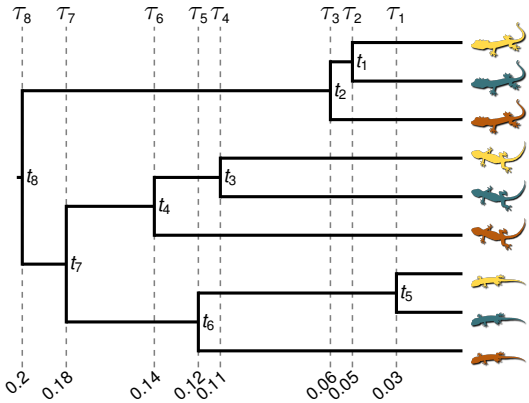


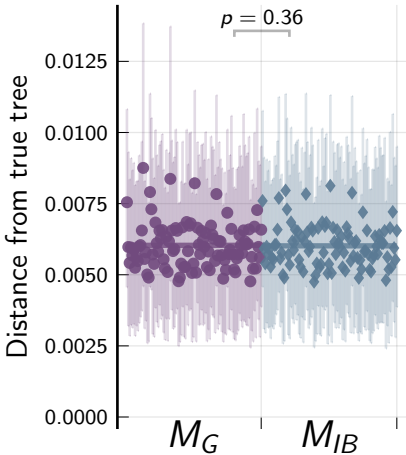


J. R. Oaks et al. (2021). *bioRxiv*

M_G = Generalized model

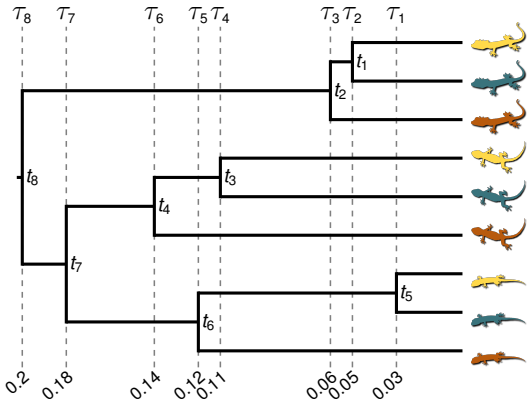
M_{IB} = Independent-bifurcating model

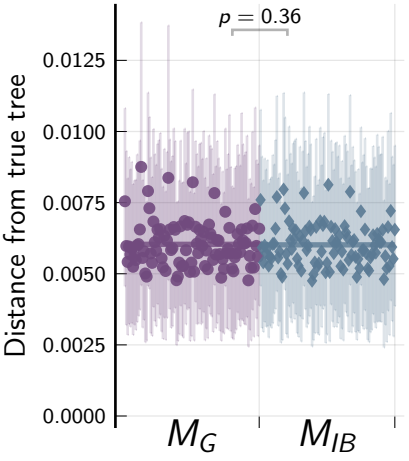




J. R. Oaks et al. (2021). *bioRxiv*

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

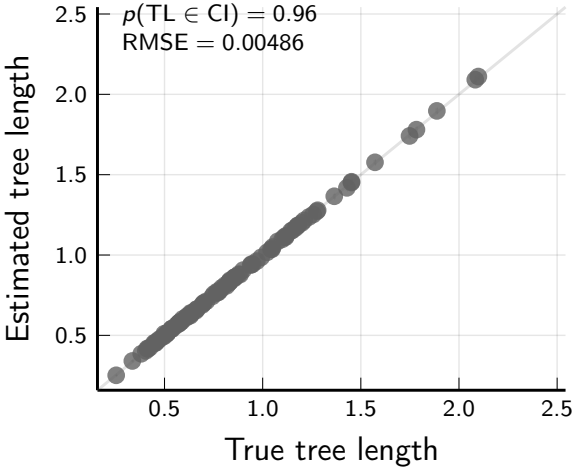




M_G performs as well as true model when divergences are independent

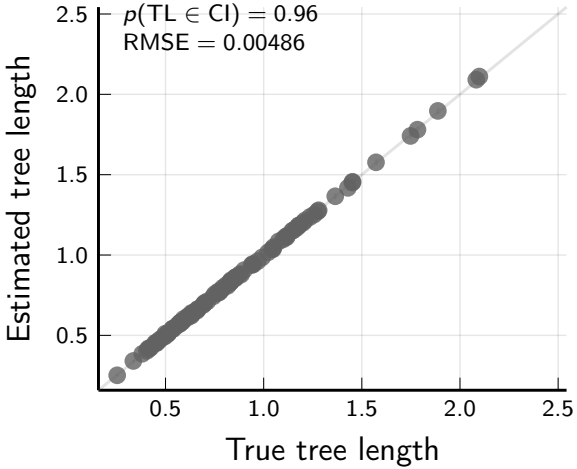
J. R. Oaks et al. (2021). *bioRxiv*

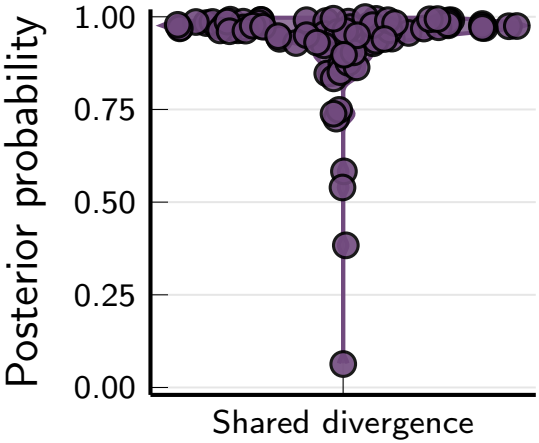
Results: random M_G trees



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

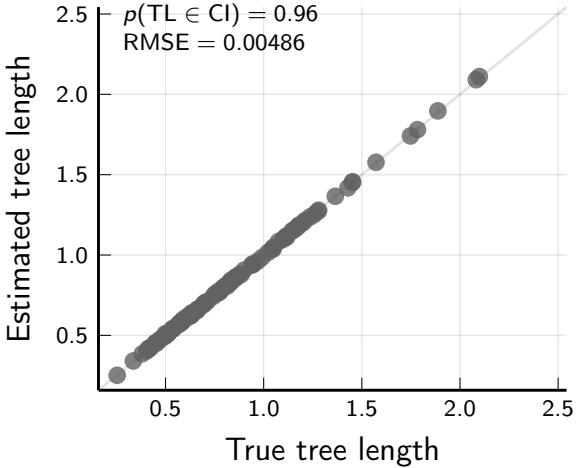
Results: random M_G trees

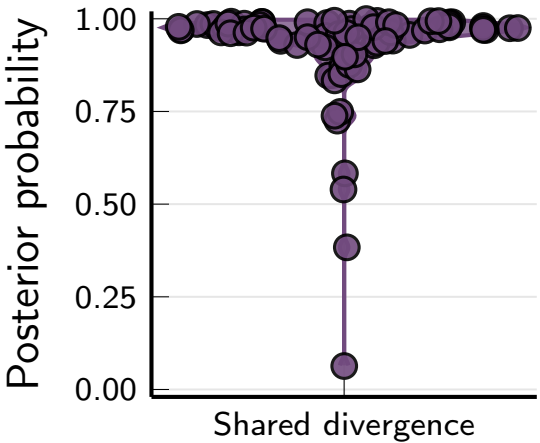




J. R. Oaks et al. (2021). *bioRxiv*

Results: random M_G trees

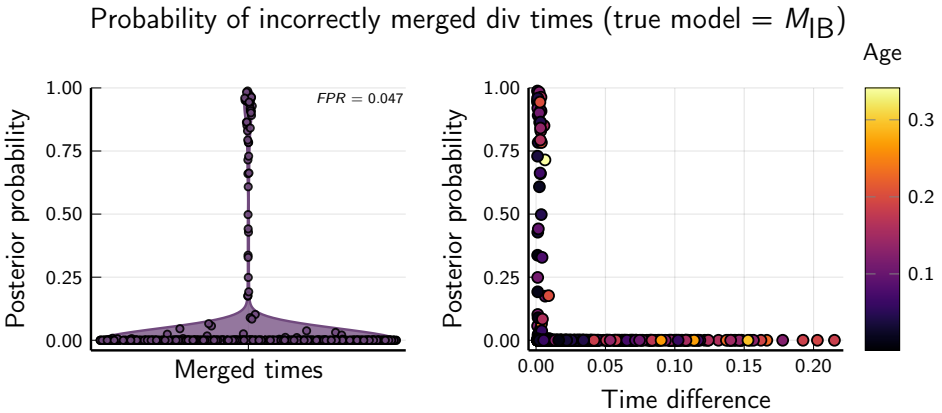




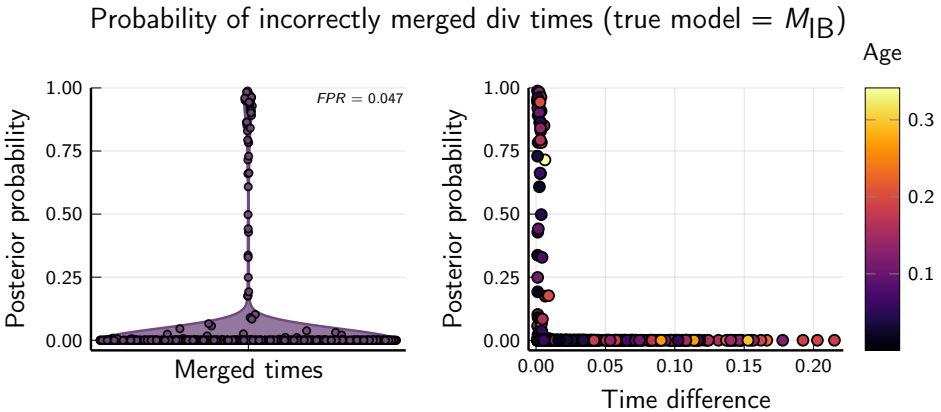
M_G performs well with data simulated on random trees with shared divergences

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

Results: random M_{IB} trees



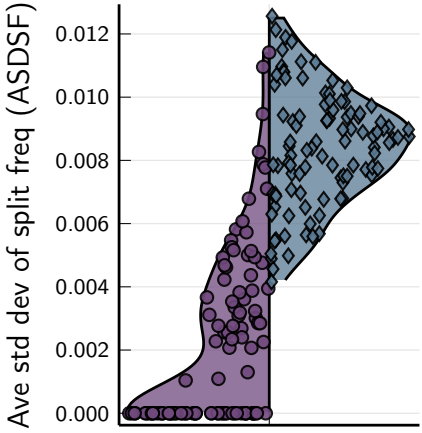
Results: random M_{IB} trees



M_G has low false positive rate

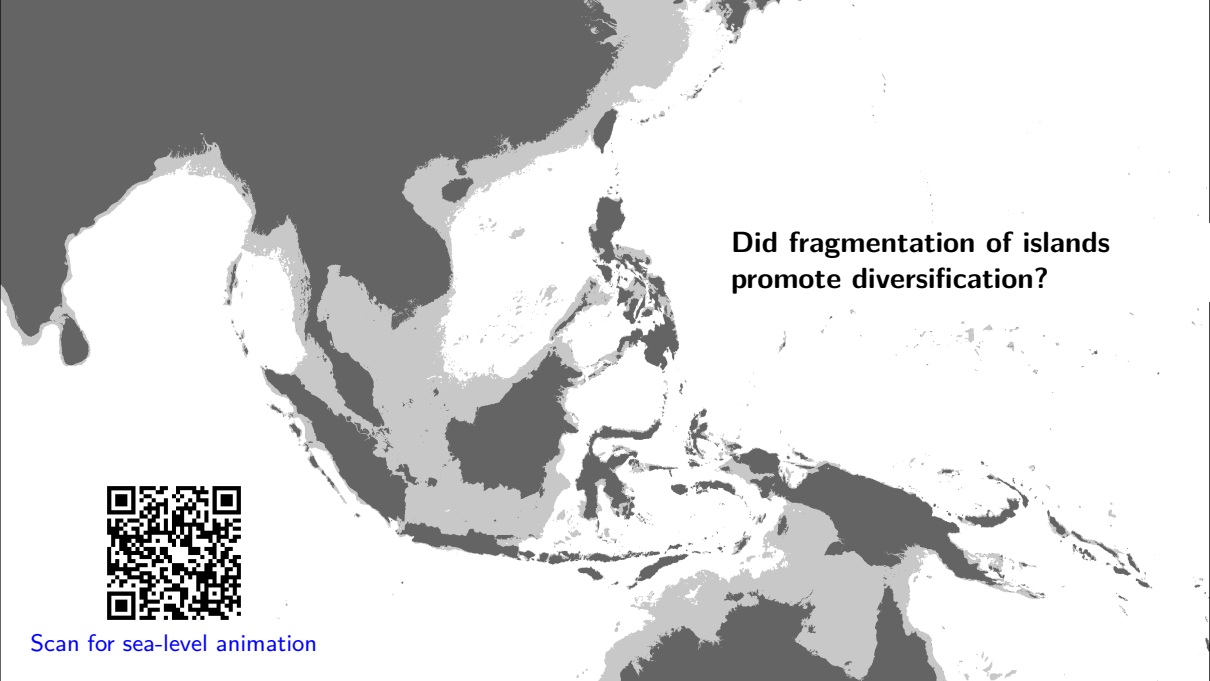
J. R. Oaks et al. (2021). *bioRxiv*

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



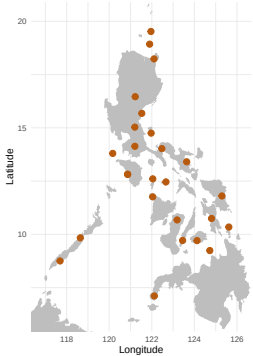
Generalizing tree space improves MCMC convergence and mixing

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

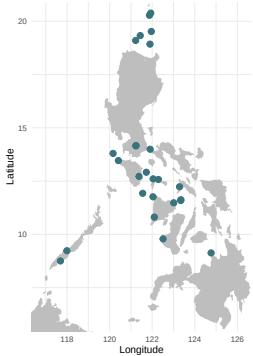

Did fragmentation of islands promote diversification?
Scan for sea-level animation

Cyrtodactylus

©Rafe M. Brown



Gekko

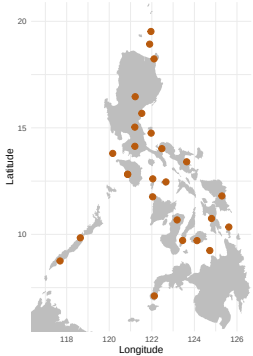


©Rafe M. Brown

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

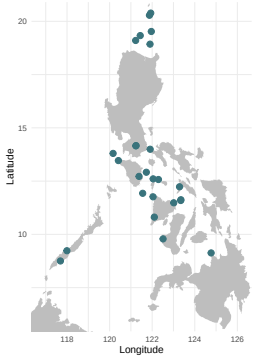
Cyrtodactylus

©Rafe M. Brown



1702 loci
155,887 sites

Gekko

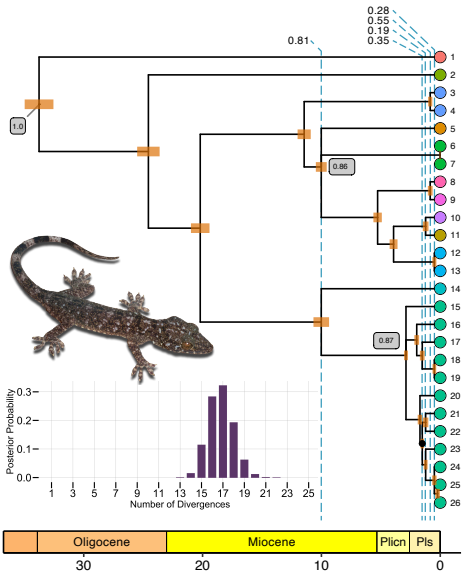


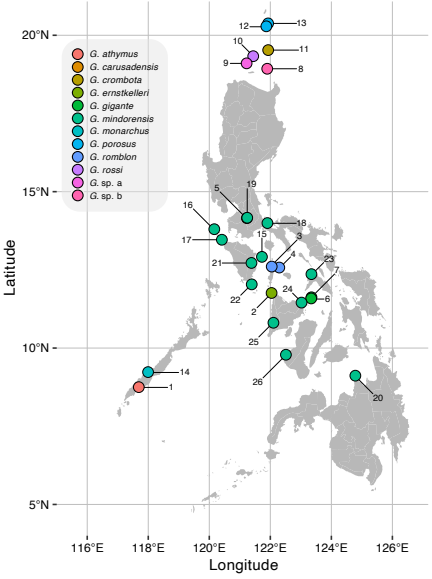
©Rafe M. Brown

1033 loci
94,813 sites

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

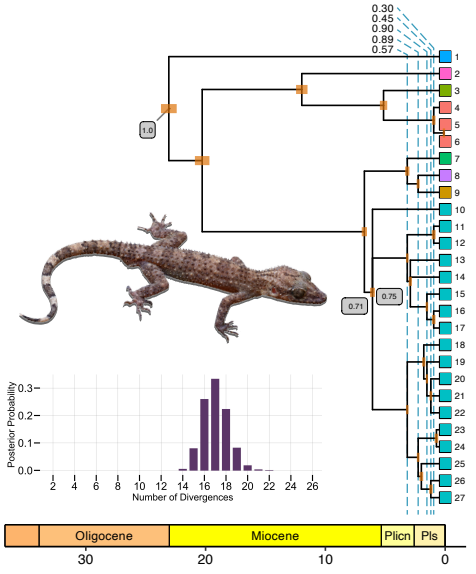
Gekko

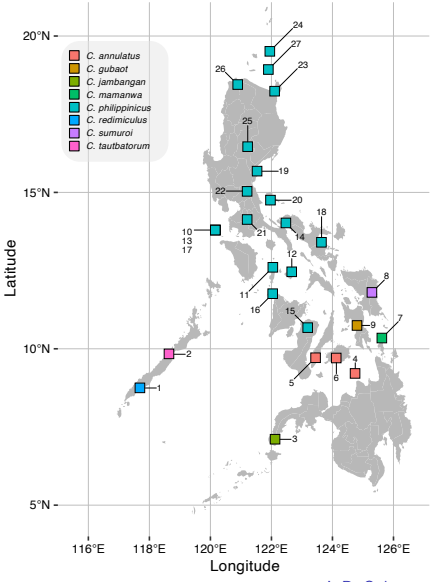




J. R. Oaks et al. (2021). *bioRxiv*

Cyrtodactylus





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

Take-home points

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

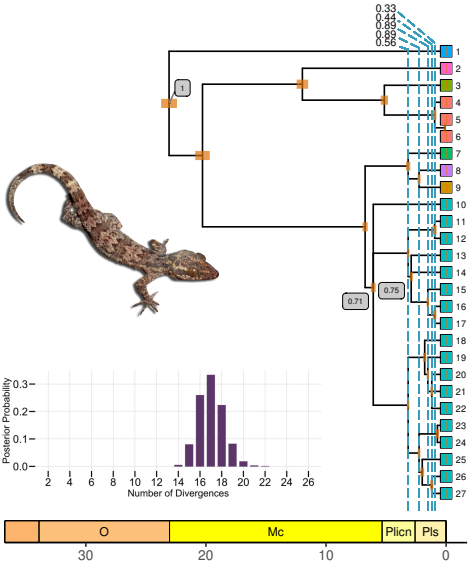
Take-home points

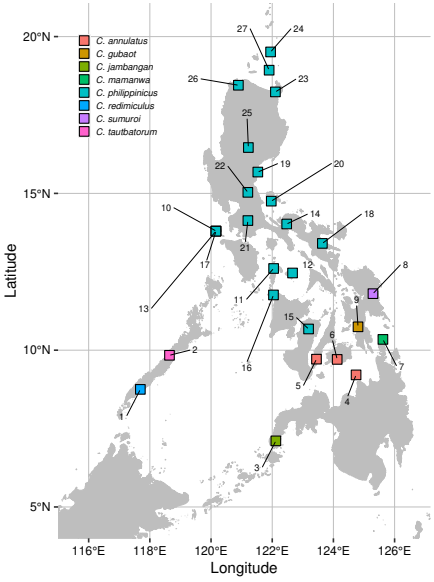
- We can accurately infer phylogenies with shared divergences with moderately sized data sets
- Generalizing tree space avoids spurious support and improves MCMC mixing

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

Cyrtodactylus





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

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

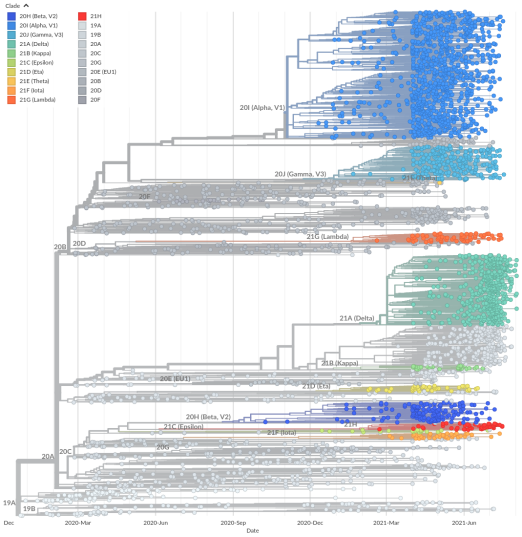
Ongoing work

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



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

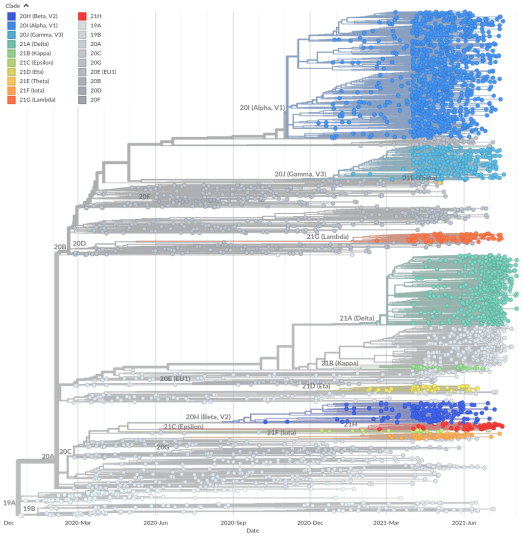
Ongoing work

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



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

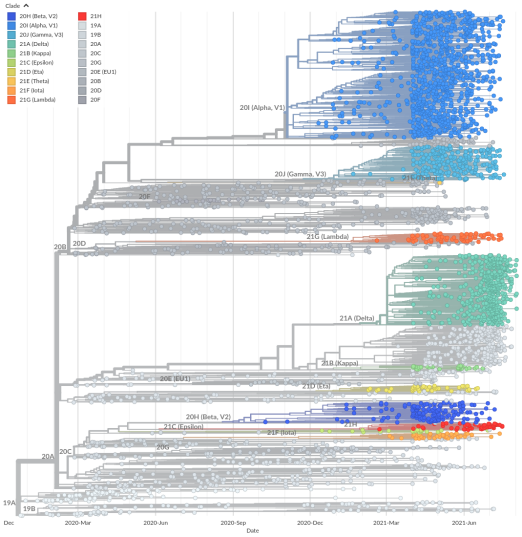
Ongoing work

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



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

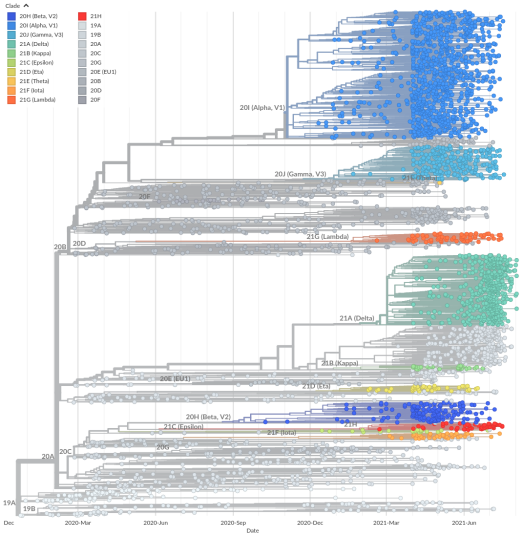
Ongoing work

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



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

Acknowledgments

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

Computation:

- Alabama Supercomputer Authority
- Auburn University Hopper Cluster

Funding:

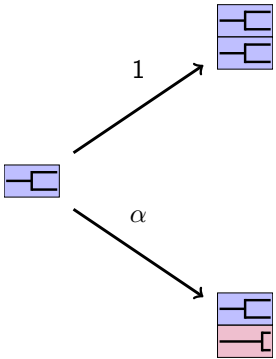
Photo credits:

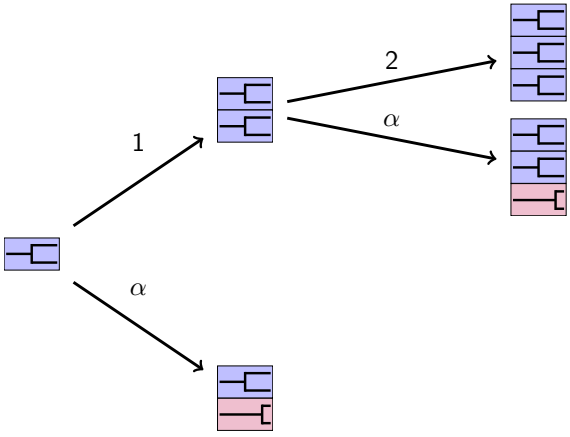
- Rafe Brown
- Perry Wood, Jr.
- PhyloPic

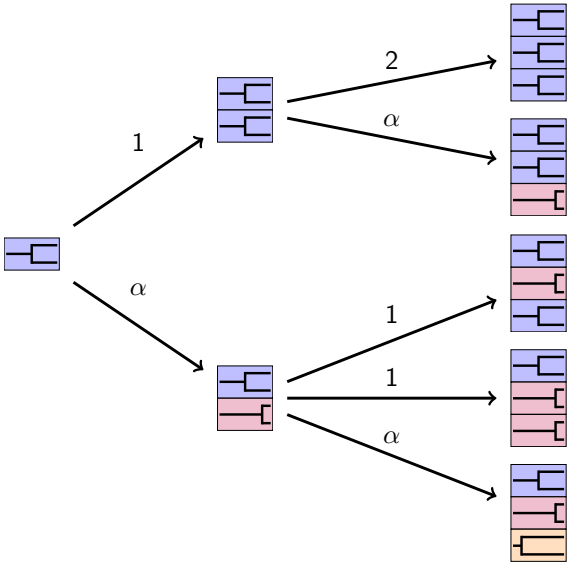
Questions?

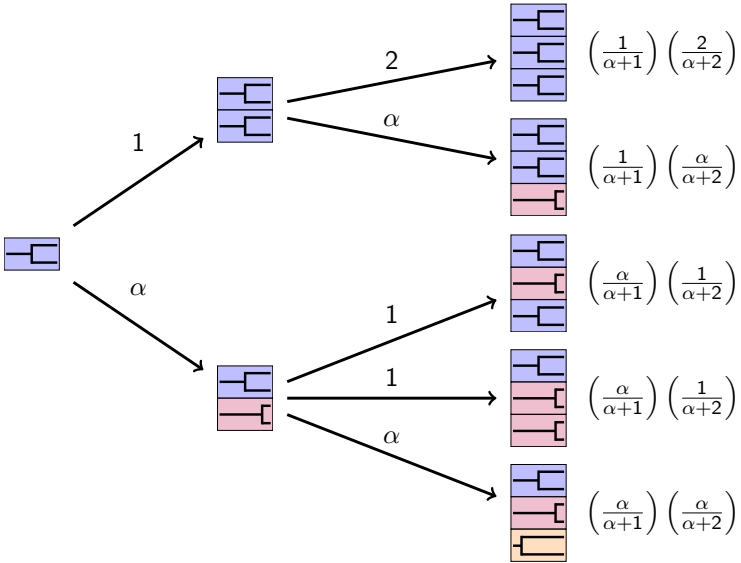
joaks@auburn.edu

phyletica.org

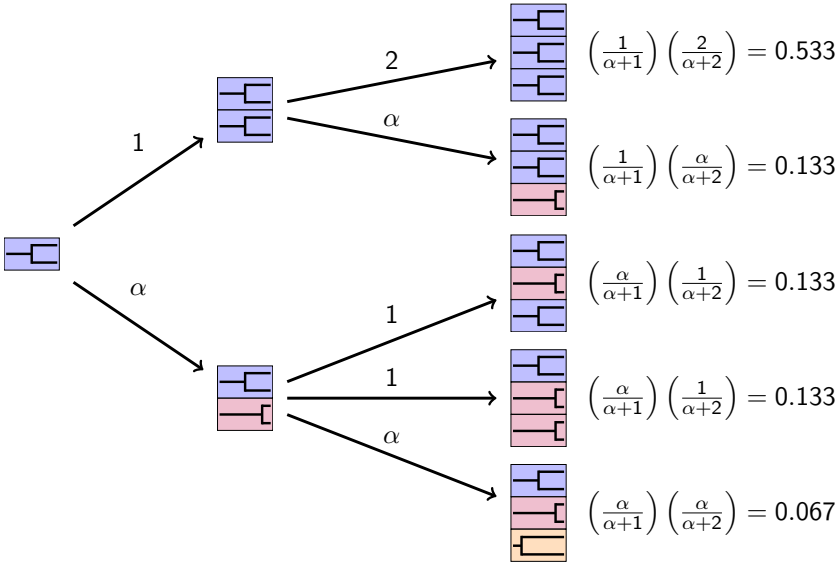

1
α


1
2
α
α

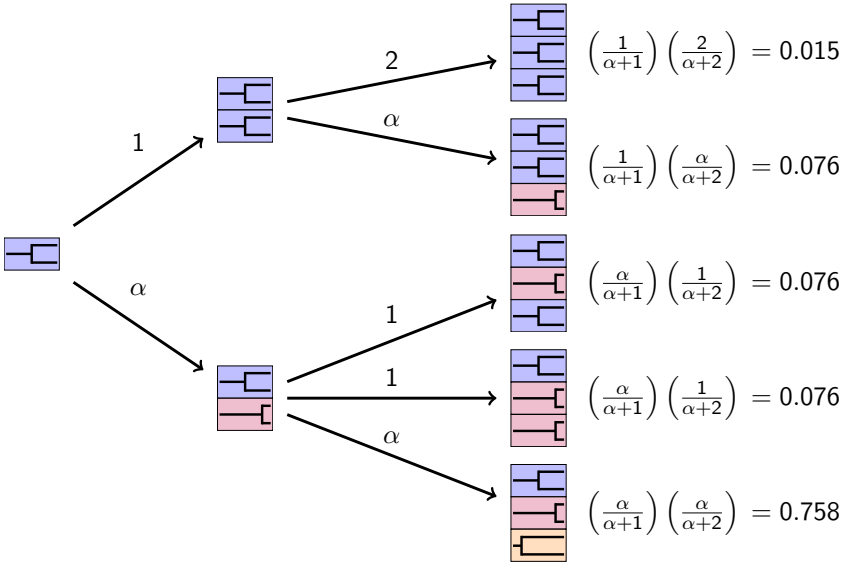

1
α
2
α
α
1
1
α


1
α
2
α
1
1
α
$\left(\frac{1}{\alpha+1}\right)\left(\frac{2}{\alpha+2}\right)$
$\left(\frac{1}{\alpha+1}\right)\left(\frac{\alpha}{\alpha+2}\right)$
$\left(\frac{\alpha}{\alpha+1}\right)\left(\frac{1}{\alpha+2}\right)$
$\left(\frac{\alpha}{\alpha+1}\right)\left(\frac{1}{\alpha+2}\right)$
$\left(\frac{\alpha}{\alpha+1}\right)\left(\frac{\alpha}{\alpha+2}\right)$

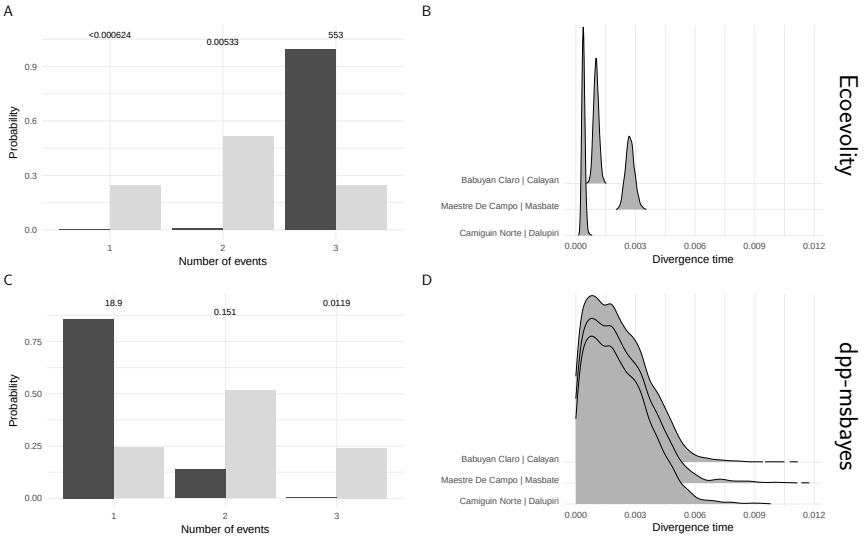
$\alpha = 0.5$


1
α
2
α
1
1
α
$\left(\frac{1}{\alpha+1}\right)\left(\frac{2}{\alpha+2}\right) = 0.533$
$\left(\frac{1}{\alpha+1}\right)\left(\frac{\alpha}{\alpha+2}\right) = 0.133$
$\left(\frac{\alpha}{\alpha+1}\right)\left(\frac{1}{\alpha+2}\right) = 0.133$
$\left(\frac{\alpha}{\alpha+1}\right)\left(\frac{1}{\alpha+2}\right) = 0.133$
$\left(\frac{\alpha}{\alpha+1}\right)\left(\frac{\alpha}{\alpha+2}\right) = 0.067$

$\alpha = 10.0$


1
α
2
α
1
1
α
$\left(\frac{1}{\alpha+1}\right)\left(\frac{2}{\alpha+2}\right) = 0.015$
$\left(\frac{1}{\alpha+1}\right)\left(\frac{\alpha}{\alpha+2}\right) = 0.076$
$\left(\frac{\alpha}{\alpha+1}\right)\left(\frac{1}{\alpha+2}\right) = 0.076$
$\left(\frac{\alpha}{\alpha+1}\right)\left(\frac{1}{\alpha+2}\right) = 0.076$
$\left(\frac{\alpha}{\alpha+1}\right)\left(\frac{\alpha}{\alpha+2}\right) = 0.758$

ABC vs ecoevolity results



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

Why so different?



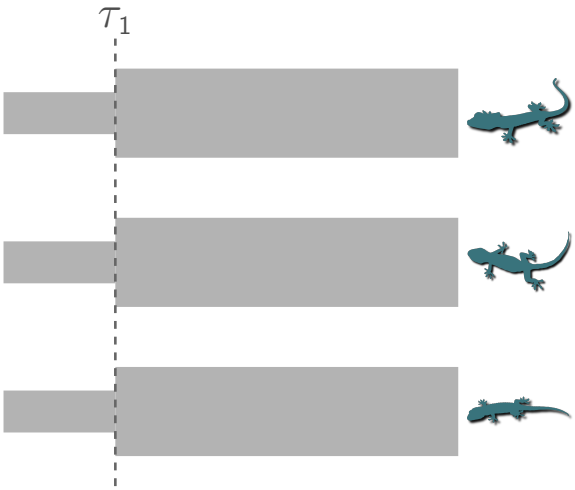
J. R. Oaks et al. (2019). *Systematic Biology* 68: 681–697

Interactive coin-flipping demo

tl;dr:

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

An aside for a related problem with population demography

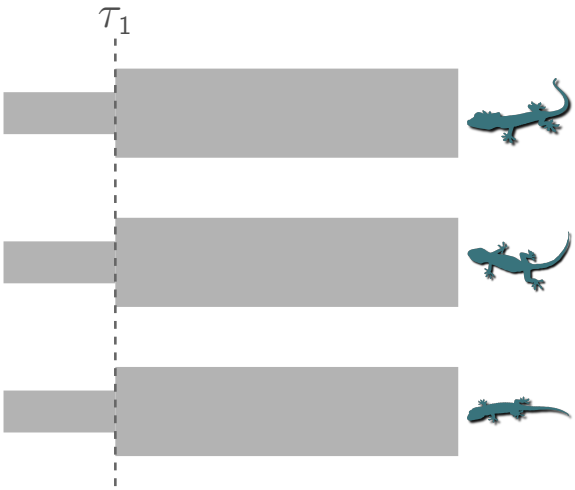


- Recent interest in testing shared demographic changes

[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

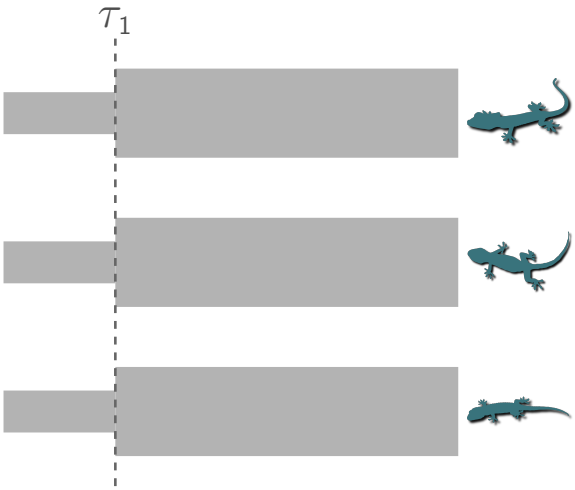


- Recent interest in testing shared demographic changes
- Several nice ABC approaches[1,2,3]
  - Inferred simultaneous expansion of 5 Alaskan populations of sticklebacks (posterior probability = 0.99)[2]

[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

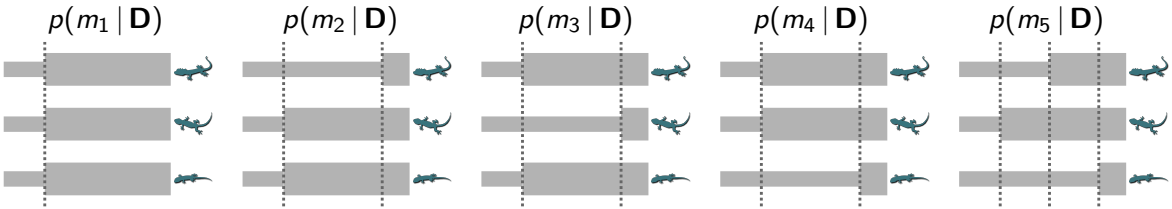


- Recent interest in testing shared demographic changes
- Several nice ABC approaches[1,2,3]
  - Inferred simultaneous expansion of 5 Alaskan populations of sticklebacks (posterior probability $= 0.99$)[2]
- It's a tricky inference problem
  - Change in population size can become unidentifiable in 3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

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

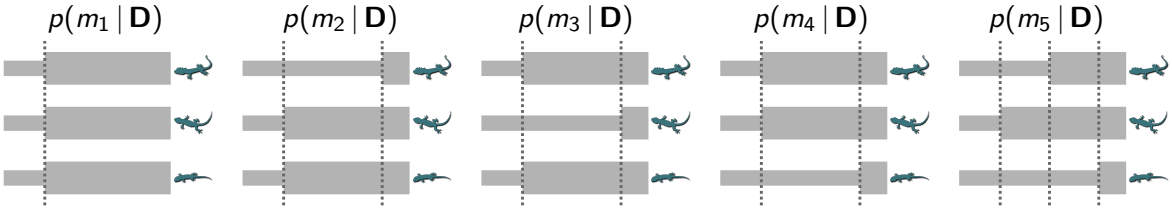
[2] A. T. Xue and M. J. Hickerson (2015). *Molecular Ecology* 24: 6223–6240

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



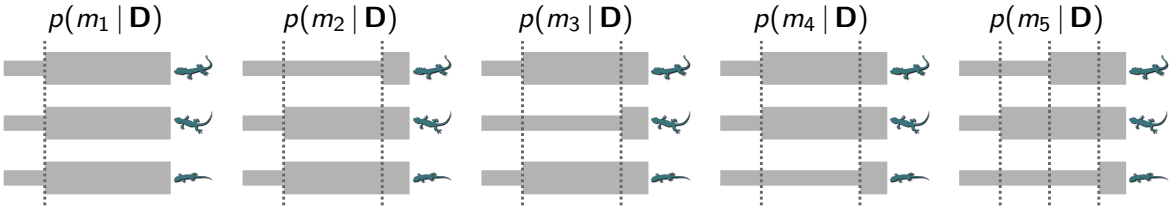
Given genomic data, can we infer the correct model and the timing of the demographic events?

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

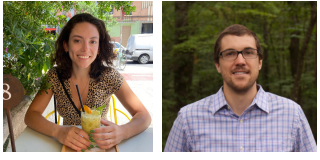


Given genomic data, can we infer the correct model and the timing of the demographic events?

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



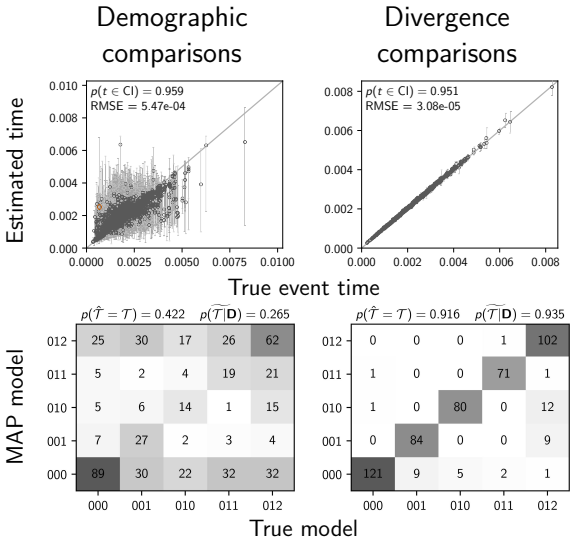
Given genomic data, can we infer the correct model and the timing of the demographic events?



- 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

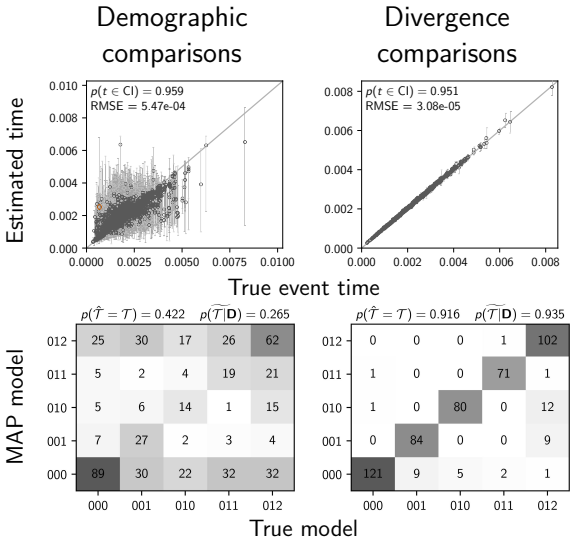
Simulation results



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

Simulation results

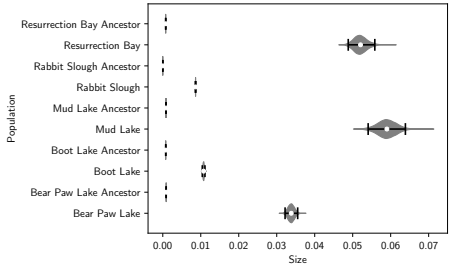
Simulations confirm theoretical expectations that changes in population size are difficult to estimate



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

Stickleback results

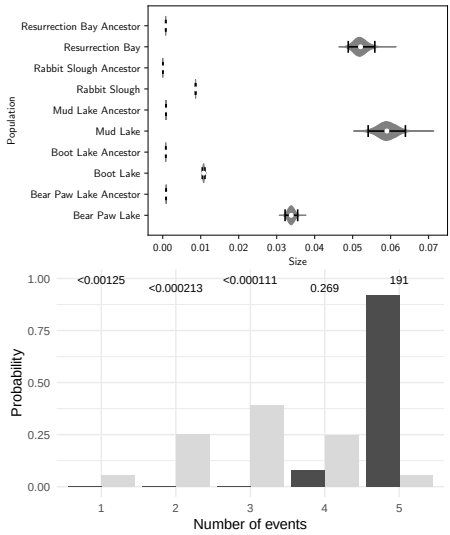
- Strong support for expansions in all 5 populations; consistent with ABC results



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

Stickleback results

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