

Estimating Divergence Times

Applied Phylogenetics

Jamie R. Oaks¹

¹Kincaid Hall 528 joaks1@gmail.com

February 13, 2013

Acknowledgements

Some of the slides and images are from a divergence-time lecture by Tracy Heath.

Thanks Tracy!

Outline

Why do we want to estimate trees with branches in units of time?

Rates, time & branch lengths

Estimating relative divergence times

Modeling rates of evolution (“clock models”)

Estimating absolute divergence times

Calibrating rates of evolution

Outline

Why do we want to estimate trees with branches in units of time?

Rates, time & branch lengths

Estimating relative divergence times

Modeling rates of evolution (“clock models”)

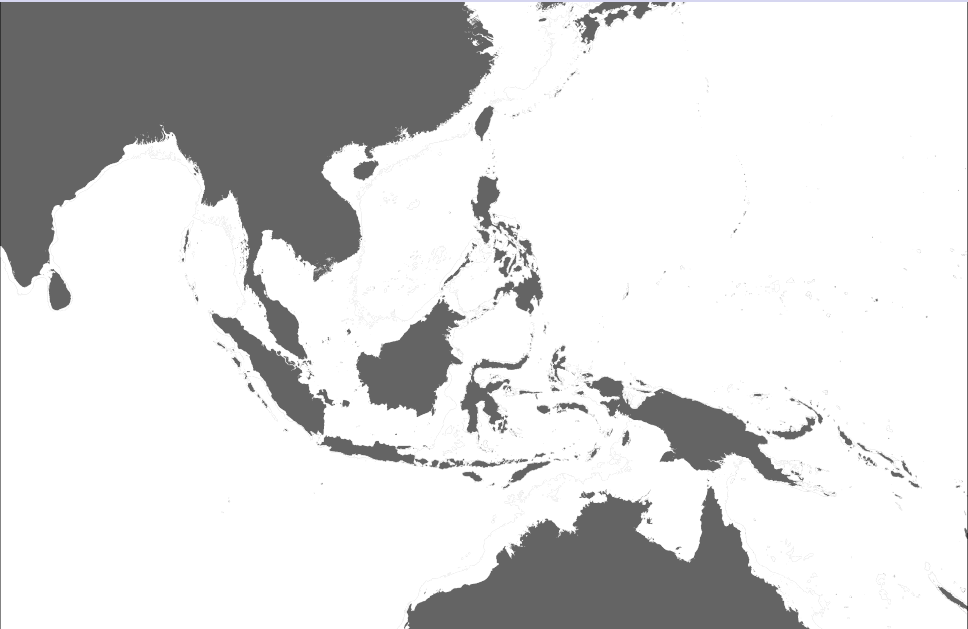
Estimating absolute divergence times

Calibrating rates of evolution

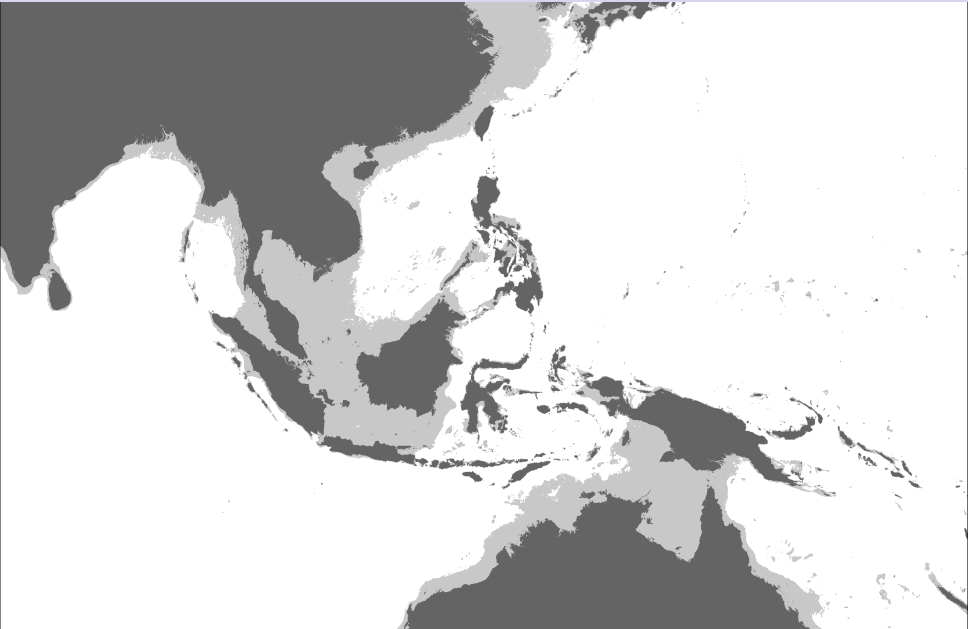
Why estimate divergence times?

- ▶ Having a history of diversification through time is very powerful
- ▶ Have past environmental changes affected diversification?
- ▶ Epidemiology—Phylodynamics of pathogens

Southeast Asia



Southeast Asia



Outline

Why do we want to estimate trees with branches in units of time?

Rates, time & branch lengths

Estimating relative divergence times

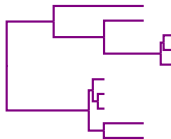
Modeling rates of evolution (“clock models”)

Estimating absolute divergence times

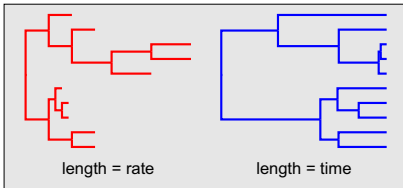
Calibrating rates of evolution

DIVERGENCE TIME ESTIMATION

The **expected # of substitutions/site** occurring along a branch is the product of the **substitution rate** and **time**



length = rate \times time

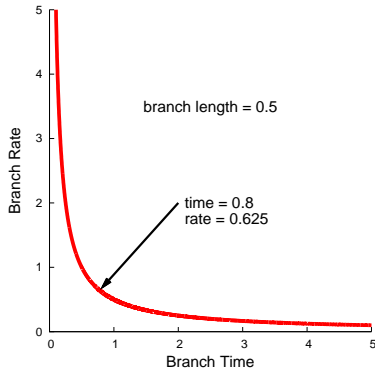


Methods for dating species divergences estimate the **substitution rate** and **time** separately

RATES AND TIMES

The sequence data
provide information
about branch length

for any possible rate,
there's a time that fits
the branch length
perfectly



(based on Thorne & Kishino, 2005)

Rates and Time

- ▶ We need to make some assumptions (i.e., create models) about the rate of evolution.
- ▶ By modeling how the rate of evolution behaves over time, we can tease apart rate and time and estimate branches *proportional* to time
 - ▶ Estimate *relative* divergence times

Outline

Why do we want to estimate trees with branches in units of time?

Rates, time & branch lengths

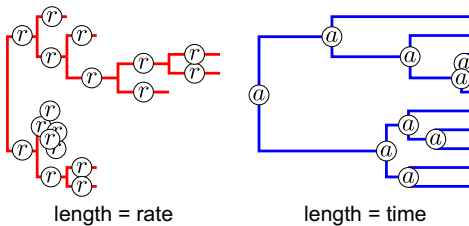
Estimating relative divergence times

Modeling rates of evolution (“clock models”)

Estimating absolute divergence times

Calibrating rates of evolution

BAYESIAN DIVERGENCE TIME ESTIMATION



$$\mathcal{R} = (r_1, r_2, r_3, \dots, r_{2N-2})$$

$$\mathcal{A} = (a_1, a_2, a_3, \dots, a_{N-1})$$

$$N = \text{number of tips}$$

Outline

Why do we want to estimate trees with branches in units of time?

Rates, time & branch lengths

Estimating relative divergence times

Modeling rates of evolution (“clock models”)

Estimating absolute divergence times

Calibrating rates of evolution

MODELING RATE VARIATION

Some models describing lineage-specific substitution rate variation:

- **Global molecular clock** (Zuckerlandl & Pauling, 1962)
- **Local molecular clocks** (Hasegawa, Kishino & Yano 1989; Kishino & Hasegawa 1990; Yoder & Yang 2000; Yang & Yoder 2003, Drummond and Suchard 2010)
- **Compound Poisson process model** (Huelsenbeck, Larget and Swofford 2000)
- **Log-normally distributed autocorrelated rates** (Thorne, Kishino & Painter 1998; Kishino, Thorne & Bruno 2001; Thorne & Kishino 2002)
- **Uncorrelated/independent rates models** (Drummond et al. 2006; Rannala & Yang 2007; Lepage et al. 2007)
- **Mixture models on branch rates** (Heath, Holder, Huelsenbeck 2012)

Global Molecular Clock

- ▶ The simplest model
- ▶ We assume the rate of change is the same on every branch
- ▶ We explain long/short branches due to differences in time
- ▶ But ...

Rates Vary!

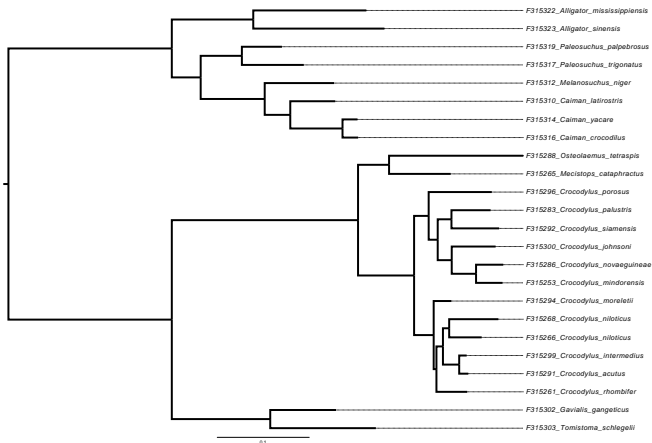
What is a substitution?

- ▶ Mutations are introduced into a populations
- ▶ Some mutations fix to become substitutions

We expect the substitution rate to vary among lineages due to:

- ▶ Differences in underlying mutation rate
 - ▶ Metabolic rates
 - ▶ Generation times
 - ▶ DNA repair mechanisms
- ▶ Differences in selective constraints

Rates Vary!

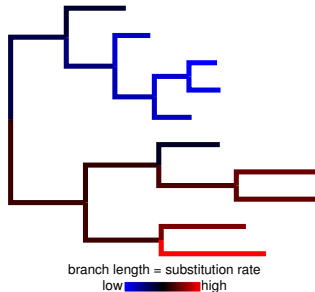


We need “relaxed-clock” models to account for rate variation

AUTOCORRELATED RATES

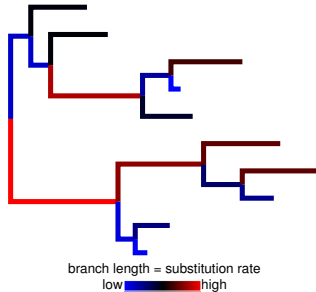
Substitution rates evolve gradually over time – closely related lineages have similar rates

The rate at a node is drawn from a lognormal distribution with a mean equal to the parent rate



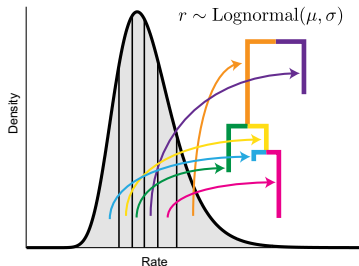
INDEPENDENT/UNCORRELATED RATES

Lineage-specific rates are uncorrelated when the rate assigned to each branch is independently drawn from an underlying distribution



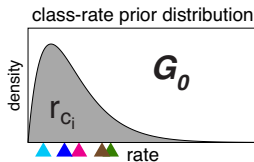
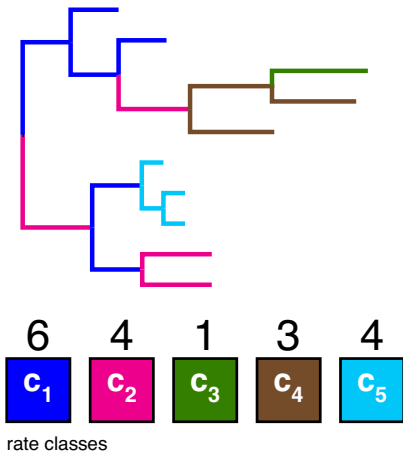
INDEPENDENT/UNCORRELATED RATES

In BEAST, the rates for the branches are drawn from a discretized lognormal distribution



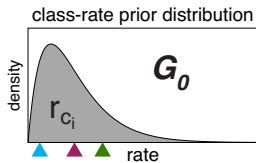
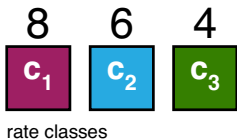
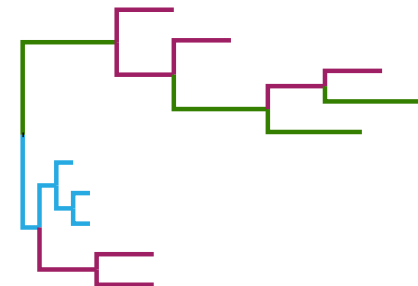
THE DIRICHLET PROCESS PRIOR (DPP)

branch length = substitution rate



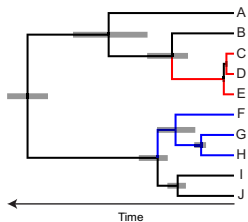
THE DIRICHLET PROCESS PRIOR (DPP)

branch length = substitution rate



BAYESIAN INFERENCE UNDER THE DPP

Current implementation: DPPDiv



Availability:

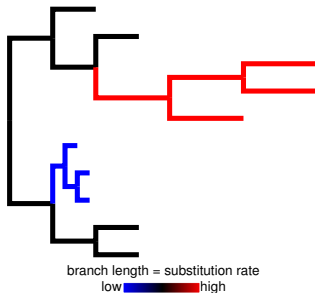
<http://phylo.bio.ku.edu/content/tracy-heath-dppdiv>

*with optimized and paralleized versions by Diego Darriba, Tomáš Flouri, & Alexis Stamatakis

LOCAL MOLECULAR CLOCKS

Rate shifts occur
infrequently over the tree

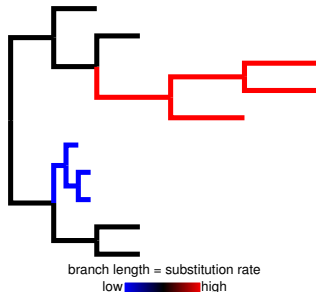
Closely related lineages
have equivalent rates
(clustered by sub-clades)



LOCAL MOLECULAR CLOCKS

Most methods for estimating local clocks required specifying the number and locations of rate changes *a priori*

Drummond and Suchard (2010) introduced a Bayesian method that samples over a broad range of possible *random local clocks*



Relaxed-clock models

- ▶ There are a lot of models for rate variation over time
- ▶ By modeling rate variation, these relaxed-clock methods allow us to estimate trees with branches proportional to time
- ▶ What about absolute time?

Outline

Why do we want to estimate trees with branches in units of time?

Rates, time & branch lengths

Estimating relative divergence times

Modeling rates of evolution (“clock models”)

Estimating absolute divergence times

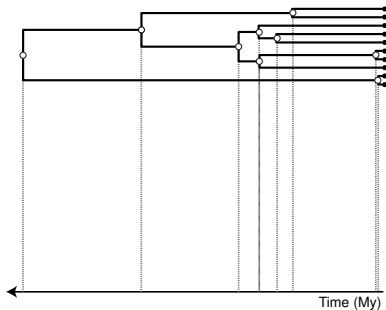
Calibrating rates of evolution

Calibrating rates

- ▶ Sequence data are only informative on *relative* divergences
- ▶ We need external information (like fossils) to calibrate (scale) the tree to absolute time
- ▶ A Bayesian statistical framework allows us to incorporate such information while accounting for uncertainty (priors!)

FOSSIL CALIBRATION

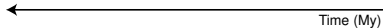
Fossil and geological data can be used to estimate the absolute ages of ancient divergences



FOSSIL CALIBRATION



The ages of extant taxa
are known



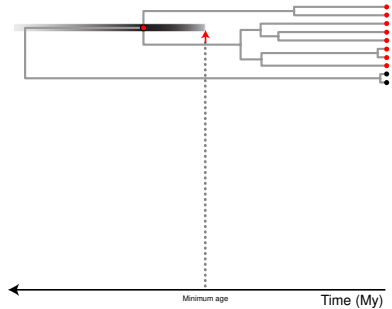
FOSSIL CALIBRATION

Fossil taxa are assigned to monophyletic clades



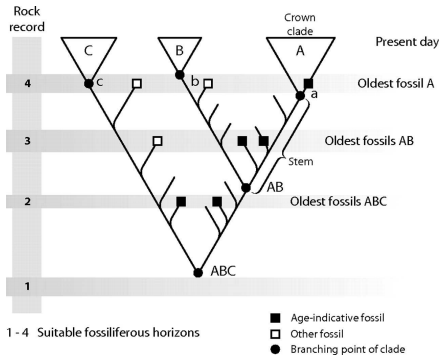
FOSSIL CALIBRATION

Fossil taxa are assigned to monophyletic clades and constrain the age of the MRCA



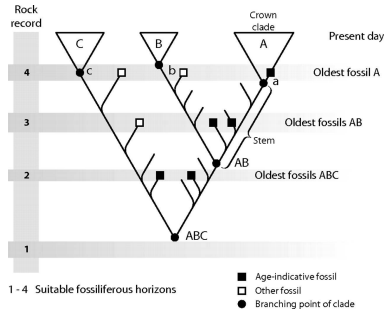
ASSIGNING FOSSILS TO CLADES

Misplaced fossils can affect node age estimates throughout the tree – if the fossil is older than its presumed MRCA



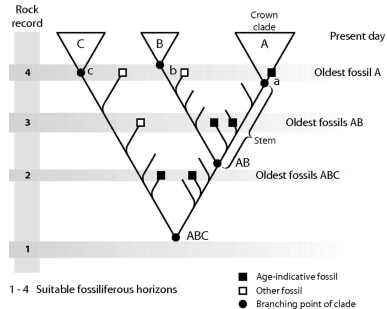
ASSIGNING FOSSILS TO CLADES

Crown clade: all living species and their most-recent common ancestor (MRCA)



ASSIGNING FOSSILS TO CLADES

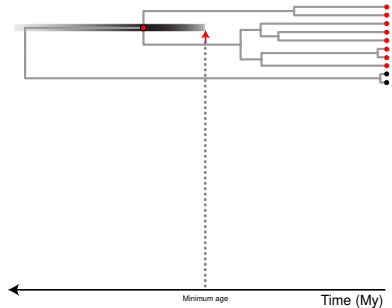
Stem lineages:
purely fossil forms
that are closer to
their descendant
crown clade than
any other crown
clade



FOSSIL CALIBRATION

Age estimates from fossils can provide **minimum** time constraints for internal nodes

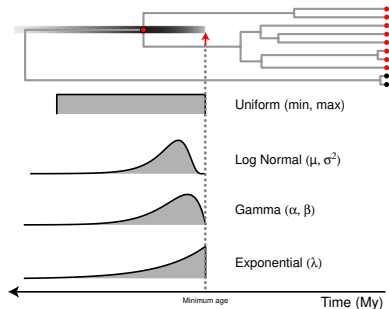
Reliable **maximum** bounds are typically unavailable



PRIOR DENSITIES ON CALIBRATED NODES

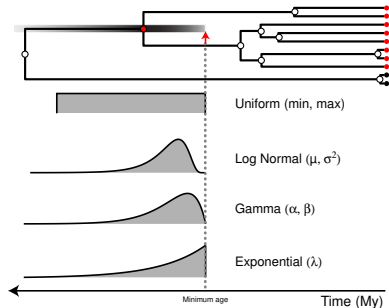
Parametric distributions are typically off-set by the age of the oldest fossil assigned to a clade

These prior densities do not (necessarily) require specification of maximum bounds



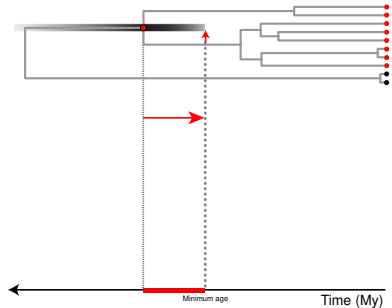
PRIOR DENSITIES ON CALIBRATED NODES

Describe the waiting time between the divergence event and the age of the oldest fossil



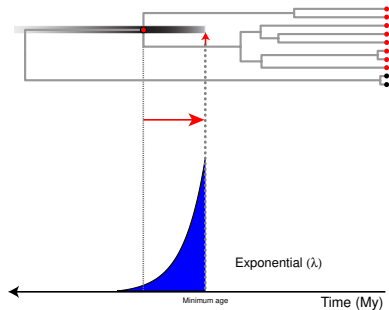
PRIOR DENSITIES ON CALIBRATED NODES

Describe the waiting time between the divergence event and the age of the oldest fossil



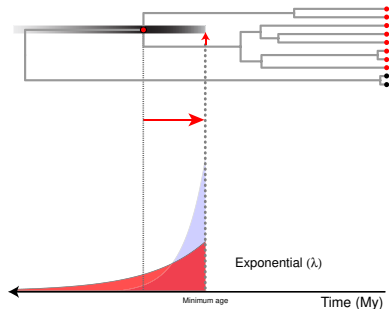
PRIOR DENSITIES ON CALIBRATED NODES

Overly **informative** priors
can bias node age
estimates to be too young

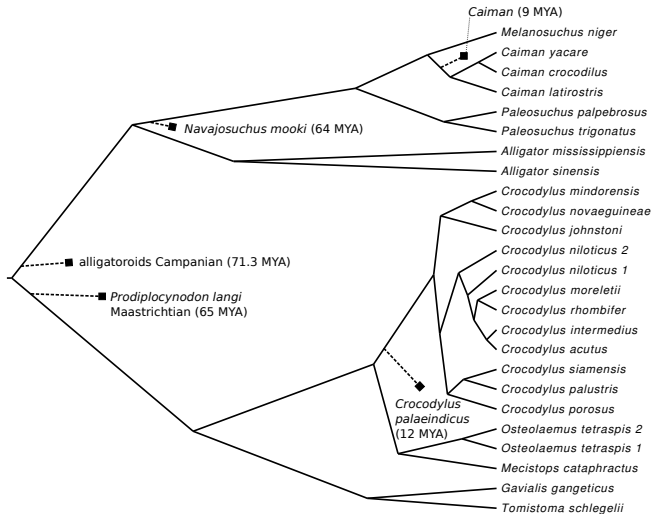


PRIOR DENSITIES ON CALIBRATED NODES

Uncertainty in the age of the MRCA of the clade relative to the age of the fossil may be better captured by **vague** prior densities



Fossil calibrations



Fossil Calibrations—an issue

- ▶ We are mixing priors on node ages!
- ▶ There is a prior on the tree and its branch lengths (e.g., birth-death process)
- ▶ We are adding additional priors on node ages based on fossils
- ▶ The resulting *marginal* prior might not be what we expect!
- ▶ This is currently a “hot” area of research

Fossil Calibrations—an issue

Potential solutions

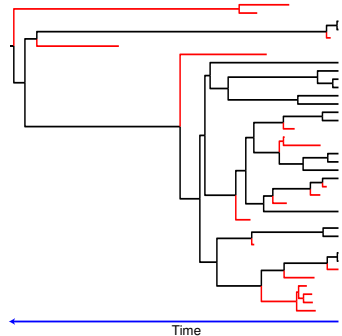
- ▶ Incorporating fossil information directly into the prior on the tree
 - ▶ Heled and Drummond (2013) arXiv
 - ▶ Heath et al. (2013) arXiv
- ▶ Fossils as tips!

FOSSIL TIP DATING

Ideally, we would like to include all of the available data

Account for uncertainty in the placement of fossil lineages

Keep all fossil data, not just the oldest descendant for a given node

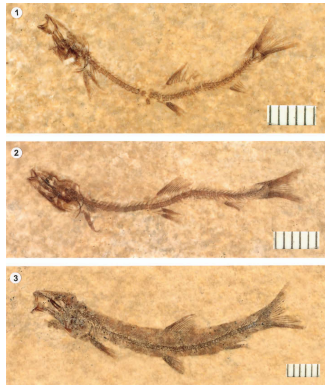


FOSSIL TIP DATING

In groups with rich fossil records, tip dating is an ideal approach

Allows for dating trees with more of the available fossils

Investigate questions (i.e. historical biogeography, character evolution) with extinct lineages

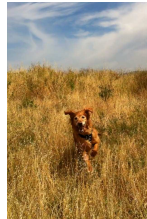


Notogoneus osculus — early growth series
illustrating the ontogeny of the scale covering

(Grande & Grande *J. Paleont.* 2008)

FOSSIL TIP DATING

Fossil tip-dating methods are available in MrBayes and BEAST, though our understanding of how well these methods work is still incomplete



Some Criticisms of Divergence-time Estimation

- ▶ Priors are too informative
- ▶ Results are often very sensitive to the priors used for fossil calibrations, the tree, and relaxed-clock parameters
- ▶ We have to make some simplifying assumptions to partition genetic distances into rates and time.

Some Criticisms of Divergence-time Estimation

What to do?

- ▶ Assess prior sensitivity
- ▶ Incorporate uncertainty from all aspects of the model (Bayesian joint inference)
- ▶ Sample from the prior only
- ▶ Present uncertainty in the results

Questions?