A Brief Overview of Estimating Divergence Times from Molecular Sequence Data

Brian R. Moore
Department of Evolution & Ecology
University of California, Davis

Bodega Phylogenetic Workshop, 2019

Outline

I. Why divergence-time estimates may be helpful

II. The strict molecular-clock model

What it is and why it may be violated

How we can test for violation of the molecular clock

III. Accommodating among-lineage variation in substitution rates

Classification scheme of various approaches

A brief survey of relaxed-clock models

Why Estimate Divergence Times?

Character evolution

time may better reflect the opportunity for character evolution

Biogeographic history

opportunities for dispersal may change over geological time scale

Lineage diversification

branching models exploit the waiting times between speciation events

Coevolution

the ages of associated lineages and timing of their co-diversification is critical

Epidemiology/phylodynamics

the time of origin and timing of spread are central to the study of epidemics

Molecular biology/molecular evolution/genomics

the age of model organisms informs our understanding of the tempo of processes

Etc., Etc.…

a time scale for the Tree of Life can inform countless questions

Outline

I. Why divergence-time estimates may be helpful

II. The strict molecular-clock model

What it is and why it may be violated

How we can test for violation of the molecular clock

III. Accommodating among-lineage variation in substitution rates

Classification scheme of various approaches

A brief survey of relaxed-clock models

Unconstrained *vs.* Clock Phylogenetic Models

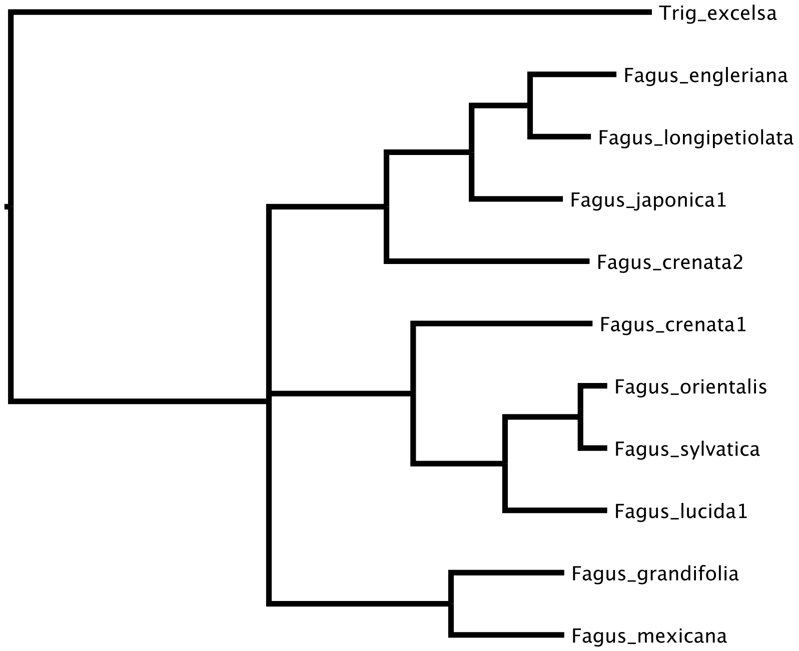
The unconstrained phylogenetic model

Assumes that every branch has an independent substitution rate

Branch lengths are rendered as the expected number of substitutions per site, $v = ut$

Substitution rate, u, and time, t, cannot be estimated independently

To do so, we must impose some assumption about substitution rates



Phylogram

Unconstrained *vs.* Clock Phylogenetic Models

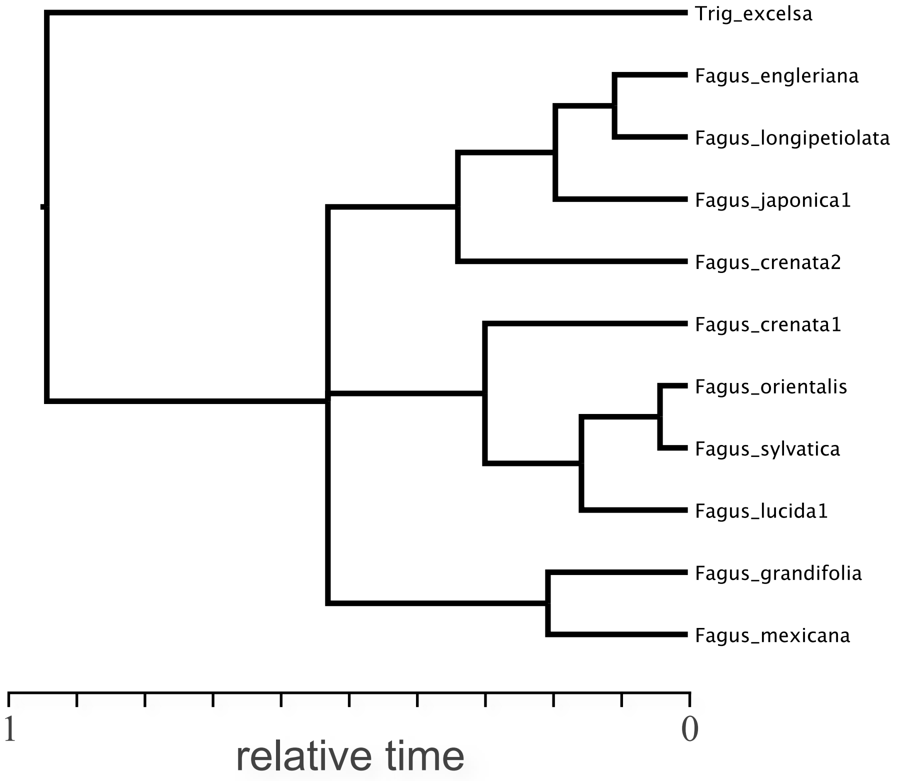
The strict molecular-clock model

Assumes that every branch has the same substitution rate

This allows us to interpret branch lengths as proportional to relative time, $v = ut$

We can also incorporate additional information to calibrate an absolute time scale

- *e.g.*, we may calibrate the tree using estimates of the absolute substitution rate or if we can assign a fossil of known age to one or more internal nodes



Chronogram

Unconstrained *vs.* Clock Phylogenetic Models

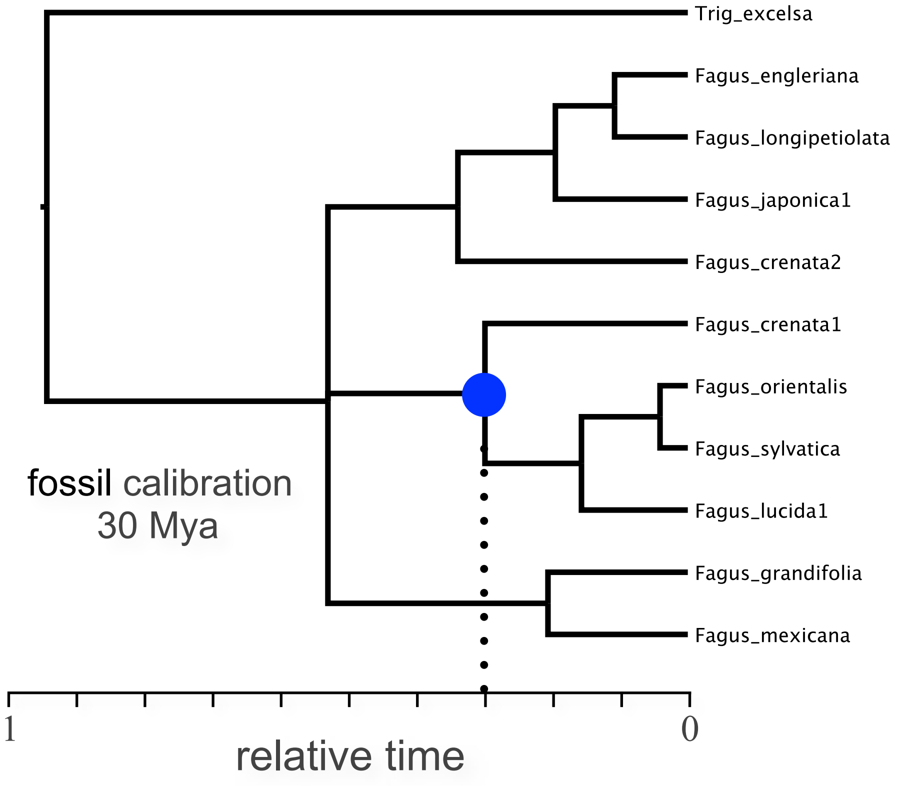
The strict molecular-clock model

Assumes that every branch has the same substitution rate

This allows us to interpret branch lengths as proportional to relative time, $v = ut$

We can also incorporate additional information to calibrate an absolute time scale

- *e.g.*, we may calibrate the tree using estimates of the absolute substitution rate or if we can assign a fossil of known age to one or more internal nodes



Chronogram

Unconstrained *vs.* Clock Phylogenetic Models

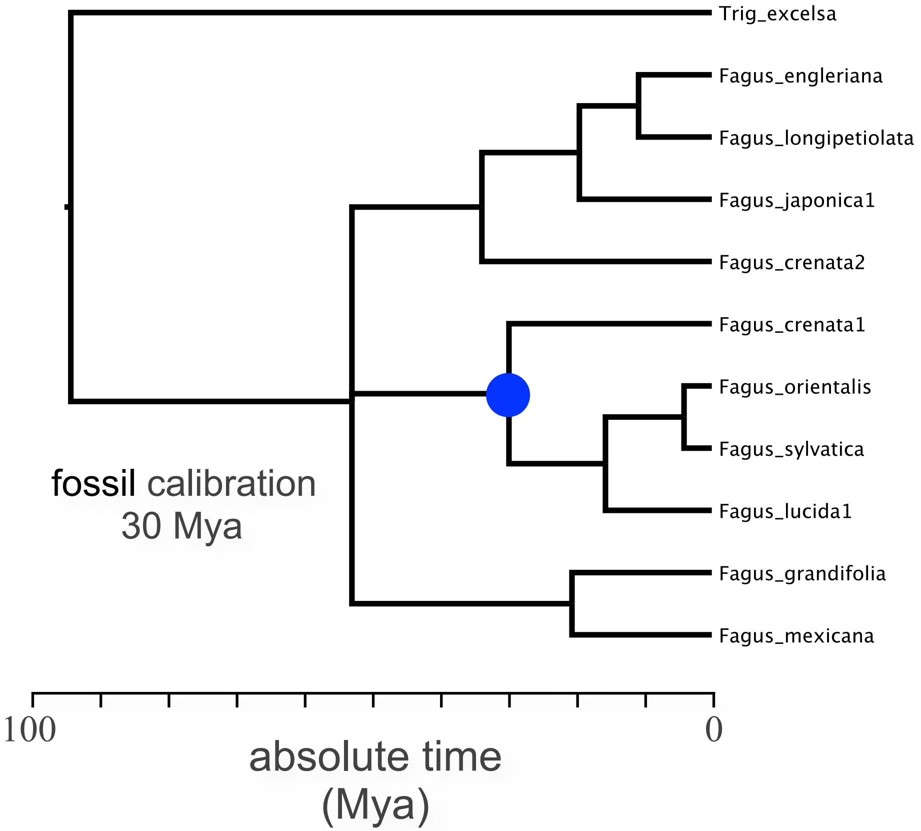
The strict molecular-clock model

Assumes that every branch has the same substitution rate

This allows us to interpret branch lengths as proportional to relative time, $v = ut$

We can also incorporate additional information to calibrate an absolute time scale

- *e.g.*, we may calibrate the tree using estimates of the absolute substitution rate or if we can assign a fossil of known age to one or more internal nodes



Chronogram

Unconstrained *vs.* Clock Phylogenetic Models

The strict molecular-clock model is biologically implausible

Numerous factors may cause substitution rates to vary across lineages:

- variation in generation times across lineages/through time
- variation in selection intensity across lineages/through time
- variation in effective population size across lineages/through time
- functional changes in sequence product across lineages/through time
- evolution of lineage-specific factors (changes in metabolic rates, DNA repair mechanisms, etc.)

Unconstrained *vs.* Clock Phylogenetic Models

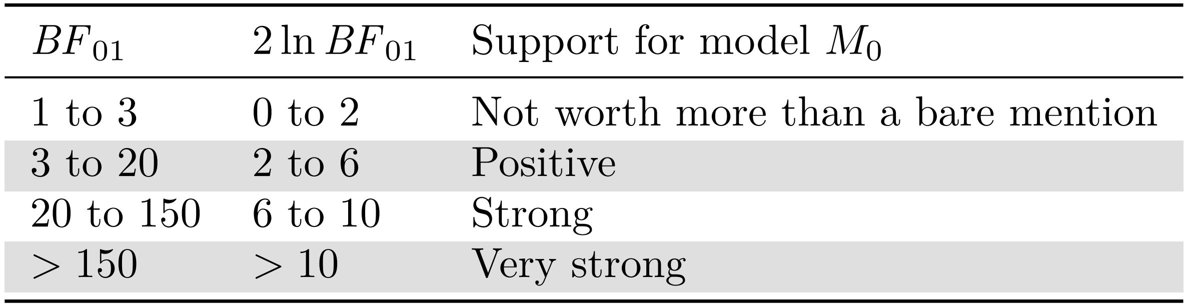
Assessing the fit of the strict molecular-clock model to our data

We can compare the competing models in the usual ways:

- estimate the marginal likelihood for the molecular-clock model, M_0
- estimate the marginal likelihood for the unconstrained model, M_1
- compute the Bayes factor for the two competing models, BF_{01}:

$$2\ln BF_{01} = 2(\ln f(\mathbf{X} \mid M_0) - \ln f(\mathbf{X} \mid M_1))$$

- $BF_{01} > 1$ supports the molecular-clock model, M_0



BF_{01}	$2 \ln BF_{01}$	Support for model M_0
1 to 3	0 to 2	Not worth more than a bare mention
3 to 20	2 to 6	Positive
20 to 150	6 to 10	Strong
> 150	> 10	Very strong

Substitution-rate variation across lineages is a *very* prevalent feature of empirical data

Under simulation, it is known that failure to accommodate substitution-rate variation across lineages will cause divergence-time estimates to be biased

Outline

I. Why divergence-time estimates may be helpful

II. The strict molecular-clock model

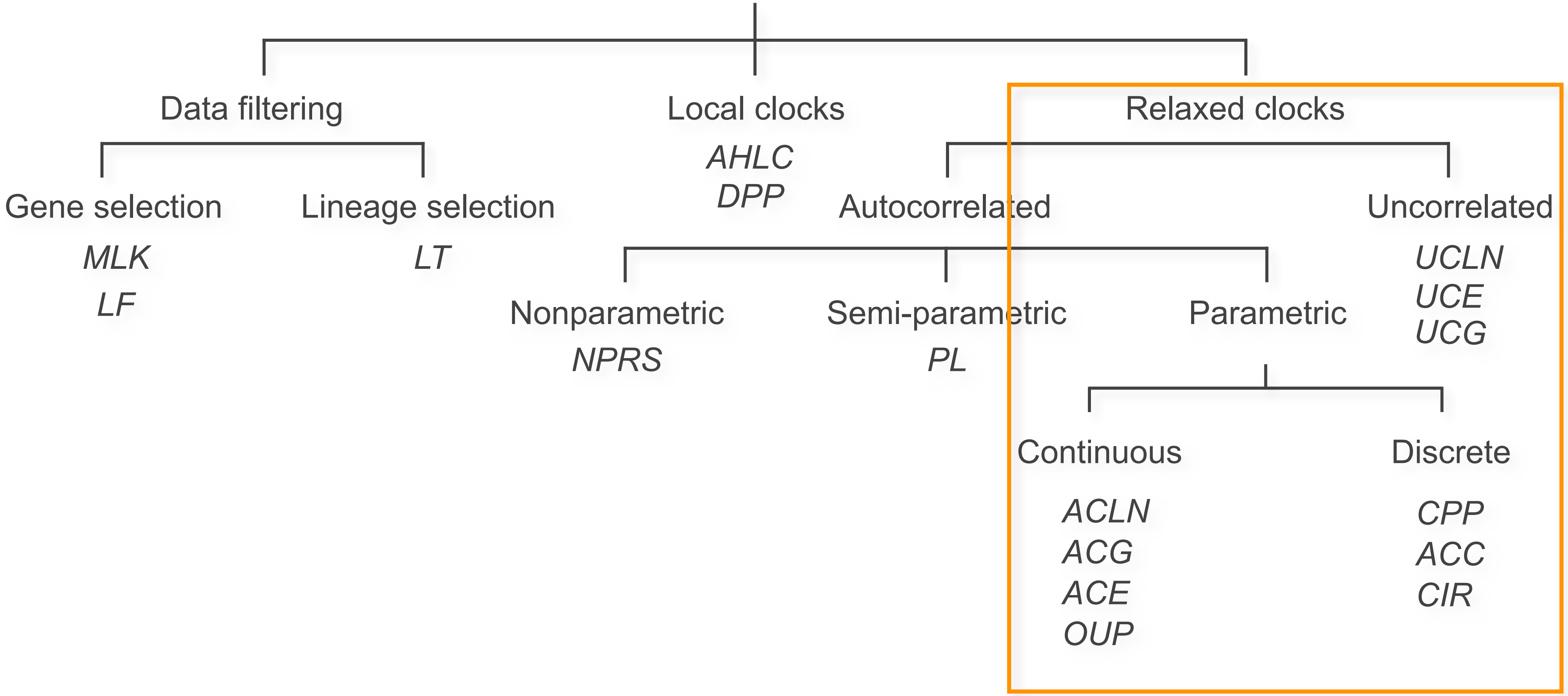
- What it is and why it may be violated
- How we can test for violation of the molecular clock

III. Accommodating among-lineage variation in substitution rates

- Classification scheme of various approaches
- A brief survey of relaxed-clock models

Accommodating Substitution-Rate Variation

Divergence-time estimates methods that accommodate rate variation



Data filtering

Gene selection

MLK

LF

Lineage selection

LT

Local clocks

AHLC

DPP

Relaxed clocks

Autocorrelated

Nonparametric

NPRS

Semi-parametric

PL

Parametric

Continuous

ACLN

ACG

ACE

OUP

Discrete

CPP

ACC

CIR

Uncorrelated

UCLN

UCE

UCG

Outline

I. Why divergence-time estimates may be helpful

II. The strict molecular-clock model

What it is and why it may be violated

How we can text for violation of the molecular clock

III. Accommodating among-lineage variation in substitution rates

Classification scheme of various approaches

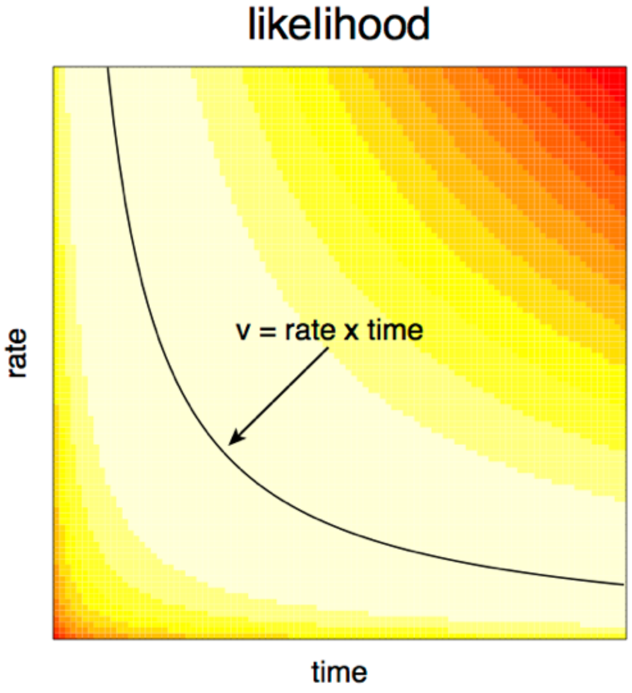
A brief survey of relaxed-clock models

Bayesian Relaxed-Clock Models

Rate and time are non-identifiable

Branch lengths are rendered as the expected number of substitutions per site, $v = ut$

Substitution rate, u, and time, t, cannot be estimated independently



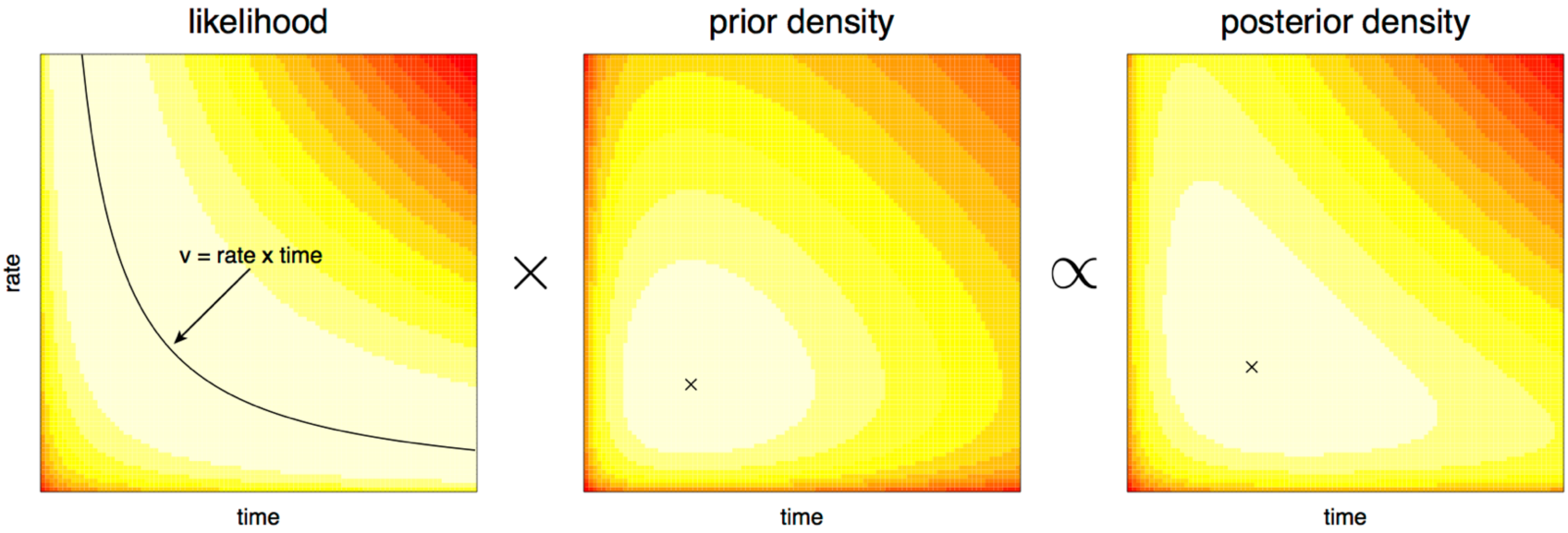
Bayesian Relaxed-Clock Models

Rate and time are non-identifiable

Branch lengths are rendered as the expected number of substitutions per site, $v = ut$

Substitution rate, u, and time, t, cannot be estimated independently

To do so, we must impose some assumption about substitution rates



Bayesian Relaxed-Clock Models

Biology motivates the extension of models

If substitution-rate variation is prevalent in empirical data, let’s model it!

Anatomy of a relaxed-clock model

Site model is used to estimate branch lengths (in the usual way)

Branch-rate prior model describes the distribution of substitution rates across branches

Node-age prior model describes the distribution of topologies and speciation times

likelihood
(substitution model)

prior probability on
rates and times

$$f(u, t \mid \mathbf{X}) = \frac{\boxed{f(\mathbf{X} \mid u, t)}\boxed{f(u, t)}}{f(\mathbf{X})}$$

branch-rate
prior model

$$f(u, t) = \boxed{f(u)}\boxed{f(t)}$$

node-age
prior model

The prior models allow us to tease apart rate and time from the branch-length estimates

Bayesian Relaxed-Clock Models

Node-age prior models

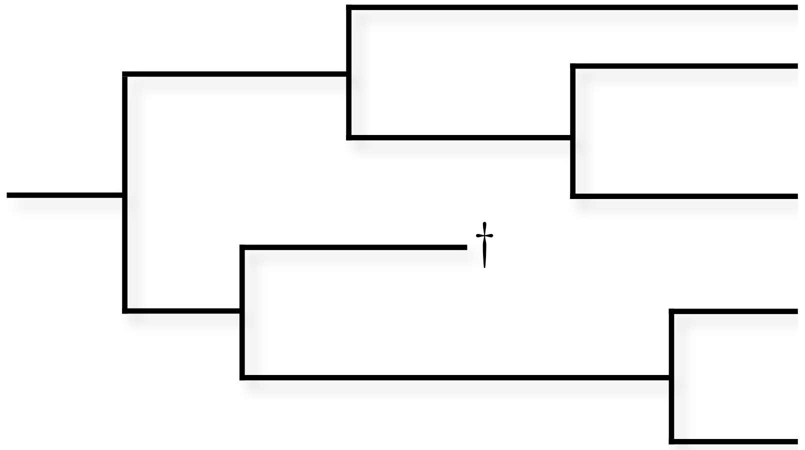
Specify a prior probability distribution on tree topologies and node ages

Types of node-age prior models

Stochastic-branching process models:

- constant-rate Yule (pure-birth) branching process
- constant-rate birth–death branching process
- sampled constant-rate birth–death branching process

λ	instantaneous speciation rate	$\dfrac{\lambda}{(\lambda + \mu)}$	relative speciation probability
μ	instantaneous extinction rate		
$(\lambda + \mu)$	exponential waiting time	$\dfrac{\mu}{(\lambda + \mu)}$	relative extinction probability



Bayesian Relaxed-Clock Models

Node-age prior models

Specify a prior probability distribution on tree topologies and node ages

Types of node-age prior models

Stochastic-branching process models:

- constant-rate Yule (pure-birth) branching process
- constant-rate birth–death branching process
- sampled constant-rate birth–death branching process

Population-level process models:

- coalescent
- multi-species coalescent

Phenomenological models:

- uniform
- Dirichlet

You can (and *should*) ask your data which probability distribution best reflects the process of substitution rates variation by they were generated

Bayesian Relaxed-Clock Models

Branch-rate prior models

Describe the prior distribution of substitution rates across branches

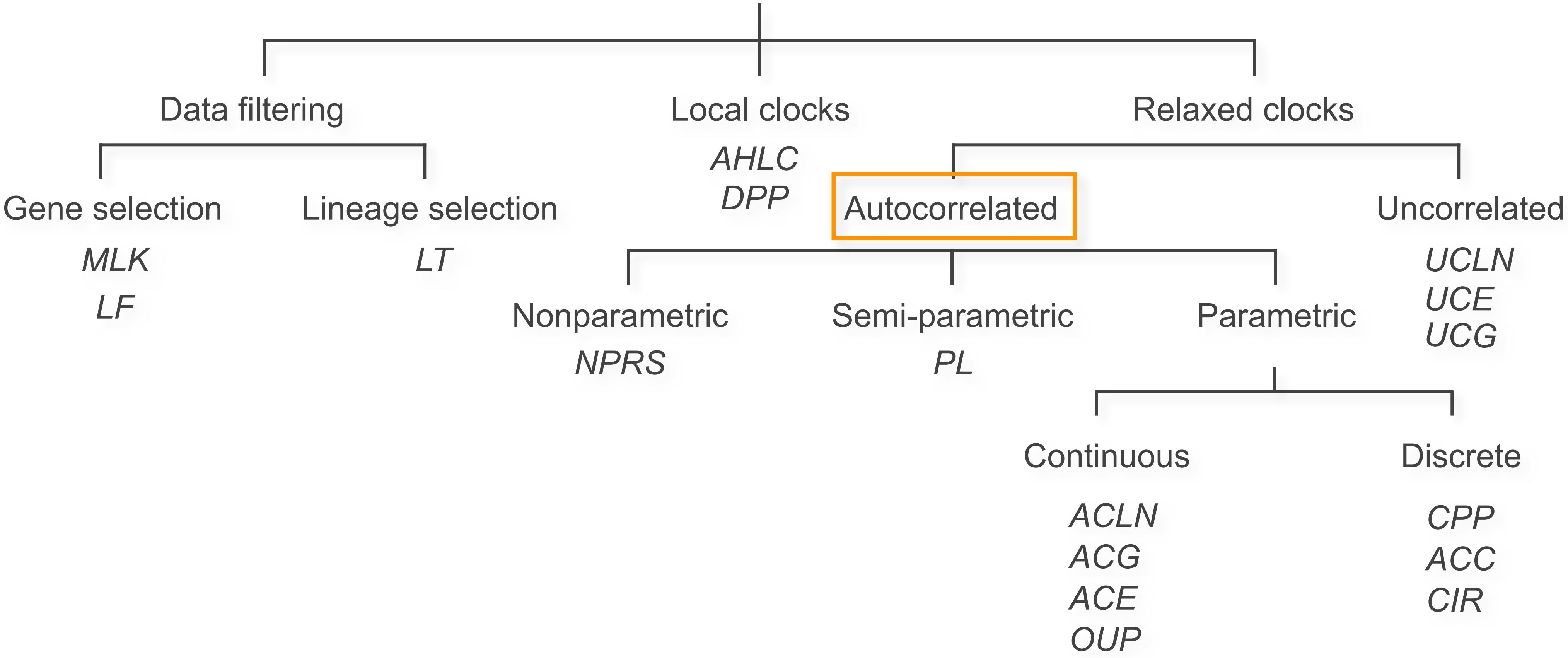
Types of branch-rate prior models

Autocorrelated models assume that the substitution-rate variation is heritable

Uncorrelated models assume that the substitution-rate variation is not heritable

Accommodating Substitution-Rate Variation

Divergence-time estimates methods that accommodate rate variation



Data filtering

Gene selection

MLK

LF

Lineage selection

LT

Local clocks

AHLC

DPP

Relaxed clocks

Autocorrelated

Nonparametric

NPRS

Semi-parametric

PL

Parametric

Continuous

ACLN

ACG

ACE

OUP

Discrete

CPP

ACC

CIR

Uncorrelated

UCLN

UCE

UCG

Autocorrelated Relaxed-Clock Models

Substitution rates may vary across lineages, but are heritable

We relax the assumption that descendant lineage inherit *identical* substitution rates with the assumption that they inherit *similar* substitution rates

These models are motivated by the following biological considerations:

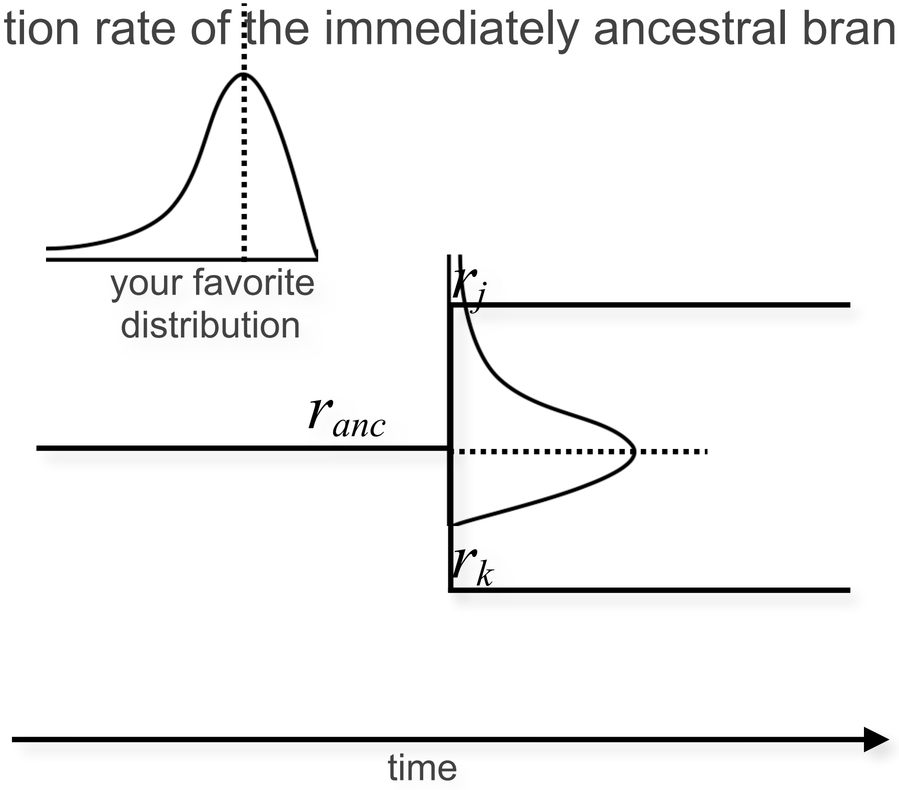
- many of the factors that impact rates of mutation are heritable
- assuming that mutation rate and substitution rate are tightly correlated, the largest component of substitution-rate variation should also be heritable
- the substitution rate of a branch should therefore be *similar* (but not necessarily *identical*) to that of its immediate ancestor

Autocorrelated Relaxed-Clock Models

Substitution rates may vary across lineages, but are heritable

We explicitly model the change in substitution rate along ancestor-descendent lineages by means of a probability distribution

The rates for descendant branches are drawn from a distribution that is centered on the substitution rate of the immediately ancestral branch



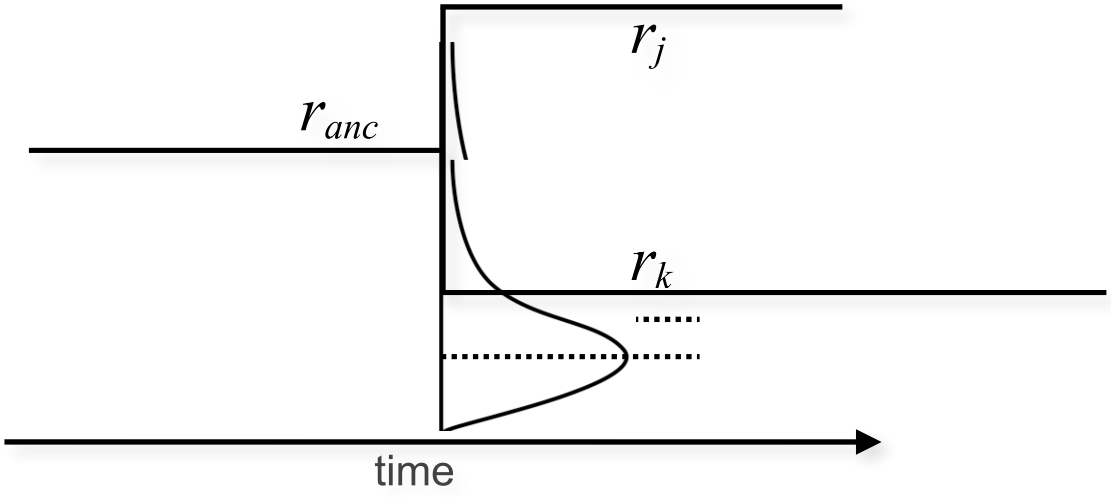
Autocorrelated Relaxed-Clock Models

Substitution rates may vary across lineages, but are heritable

We explicitly model the change in substitution rate along ancestor-descendent lineages by means of a probability distribution

The rates for descendant branches are drawn from a distribution that is centered on the substitution rate of the immediately ancestral branch

Variance in substitution rate (typically) scales with the duration of the branch



Autocorrelated Relaxed-Clock Models

Different probability distributions can be used to model autocorrelation

You can select different probability distributions to reflect your prior beliefs about how substitution rates change in an autocorrelated manner

Continuous-autocorrelated rate variation

- autocorrelated lognormal branch-rate prior model (*ACLN*)
- autocorrelated gamma branch-rate prior model (*ACG*)
- autocorrelated exponential branch-rate prior model (*ACE*)
- autocorrelated Ornstein–Uhlenbeck branch-rate prior model (*ACOUP*)

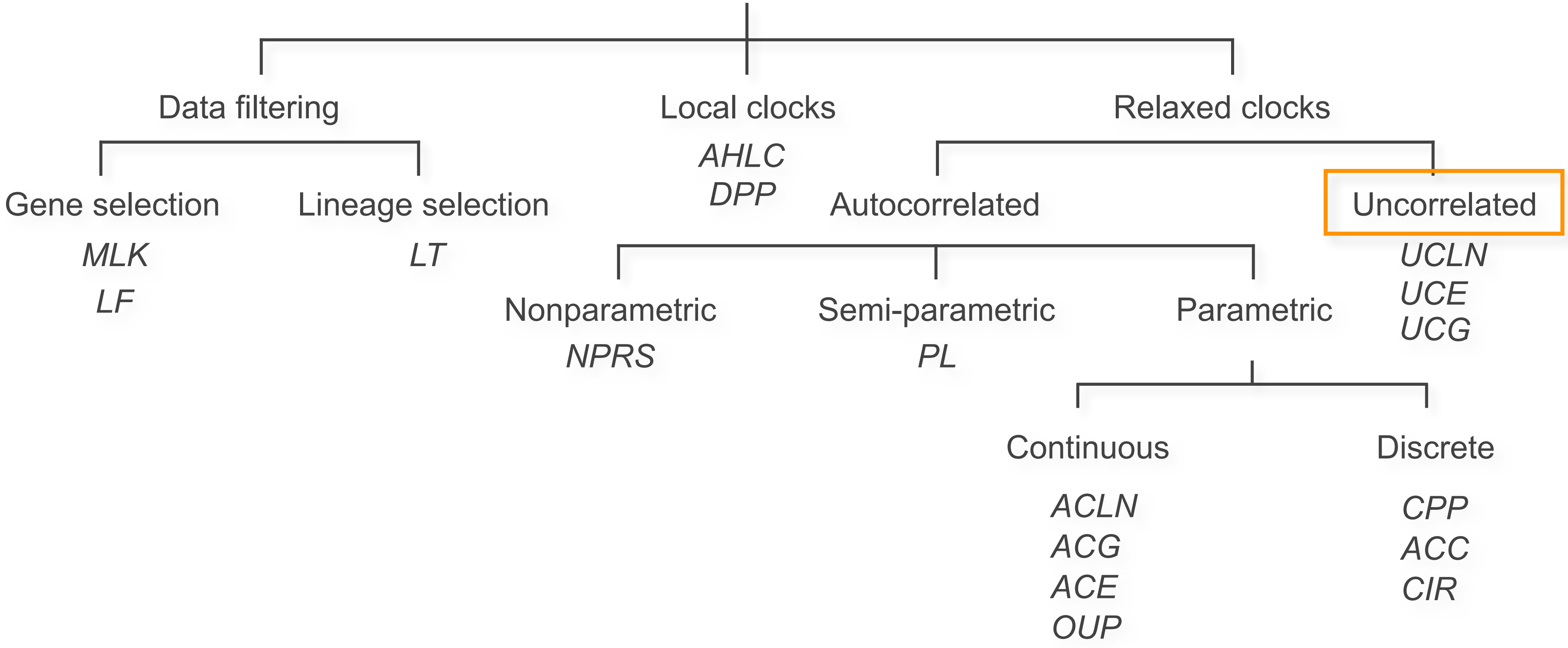
Stepwise-autocorrelated rate variation

- autocorrelated compound Poisson process branch-rate prior model (*ACPP*)
- autocorrelated Cox branch-rate prior model (*ACG*)
- autocorrelated Cox–Ingersoll–Ross process branch-rate prior model (*CIR*)

You can (and *should*) ask your data which probability distribution best reflects the process of substitution rates variation by they were generated

Accommodating Substitution-Rate Variation

Divergence-time estimates methods that accommodate rate variation



Data filtering

Gene selection

MLK

LF

Lineage selection

LT

Local clocks

AHLC

DPP

Relaxed clocks

Autocorrelated

Nonparametric

NPRS

Semi-parametric

PL

Parametric

Continuous

ACLN

ACG

ACE

OUP

Discrete

CPP

ACC

CIR

Uncorrelated

UCLN

UCE

UCG

Uncorrelated Relaxed-Clock Models

Substitution rates may vary across lineages, and are not heritable

We relax the assumption that descendant lineage inherit *identical* substitution rates with the assumption that they are independently sampled from a shared distribution

These models are motivated by the following biological considerations:

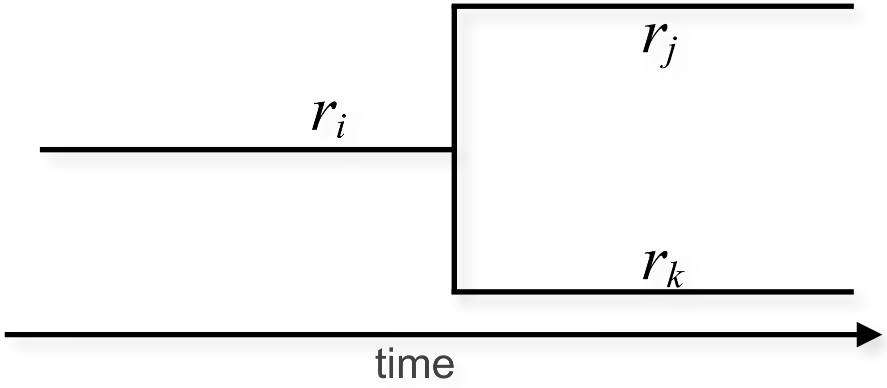
- rate autocorrelation will occur when the largest component is due to heritable factors
- at very *small* time scales "[autocorrelation is so strong that very little of the variation can be attributed to inherited factors" (Drummond *et al*., 2006)
- conversely, at very *large* time scales "there may be so much variation in inherited factors that autocorrelation along branches may break down" (Drummond *et al*., 2006)

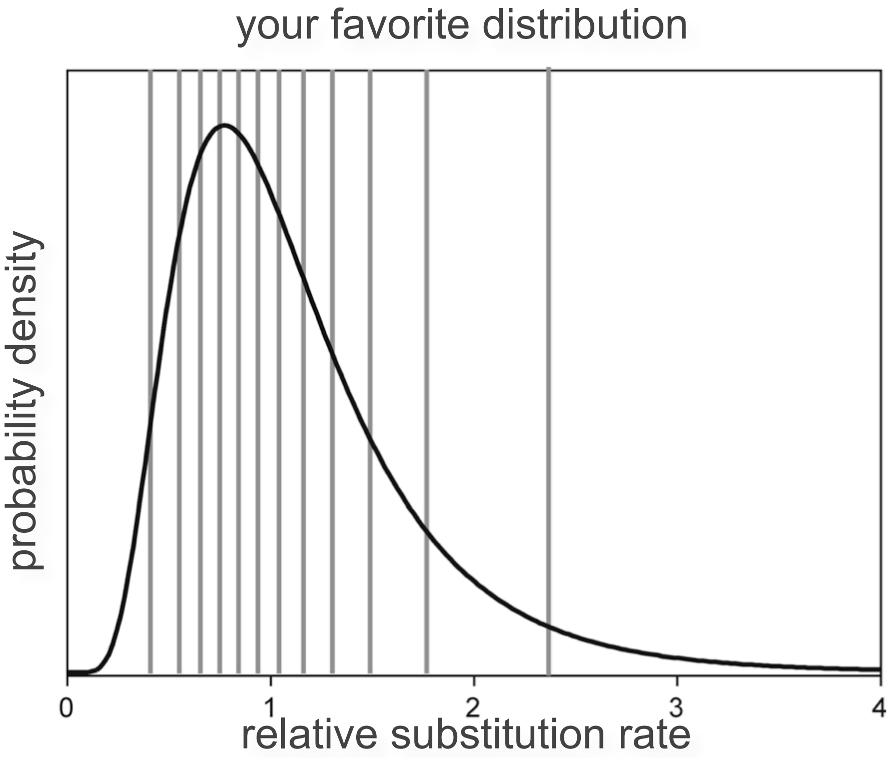
Uncorrelated Relaxed-Clock Models

Substitution rates may vary across lineages, and are not heritable

We explicitly model the change in substitution rate across ancestor-descendent lineages by means of a shared probability distribution

The rates for each branch are independently drawn from a shared distribution with parameters that are estimated from the data





Uncorrelated Relaxed-Clock Models

Different probability distributions can be used to model rate variation

You can select different probability distributions to reflect your prior beliefs about how substitution rates change in an uncorrelated manner

There are many probability distributions that may be used:

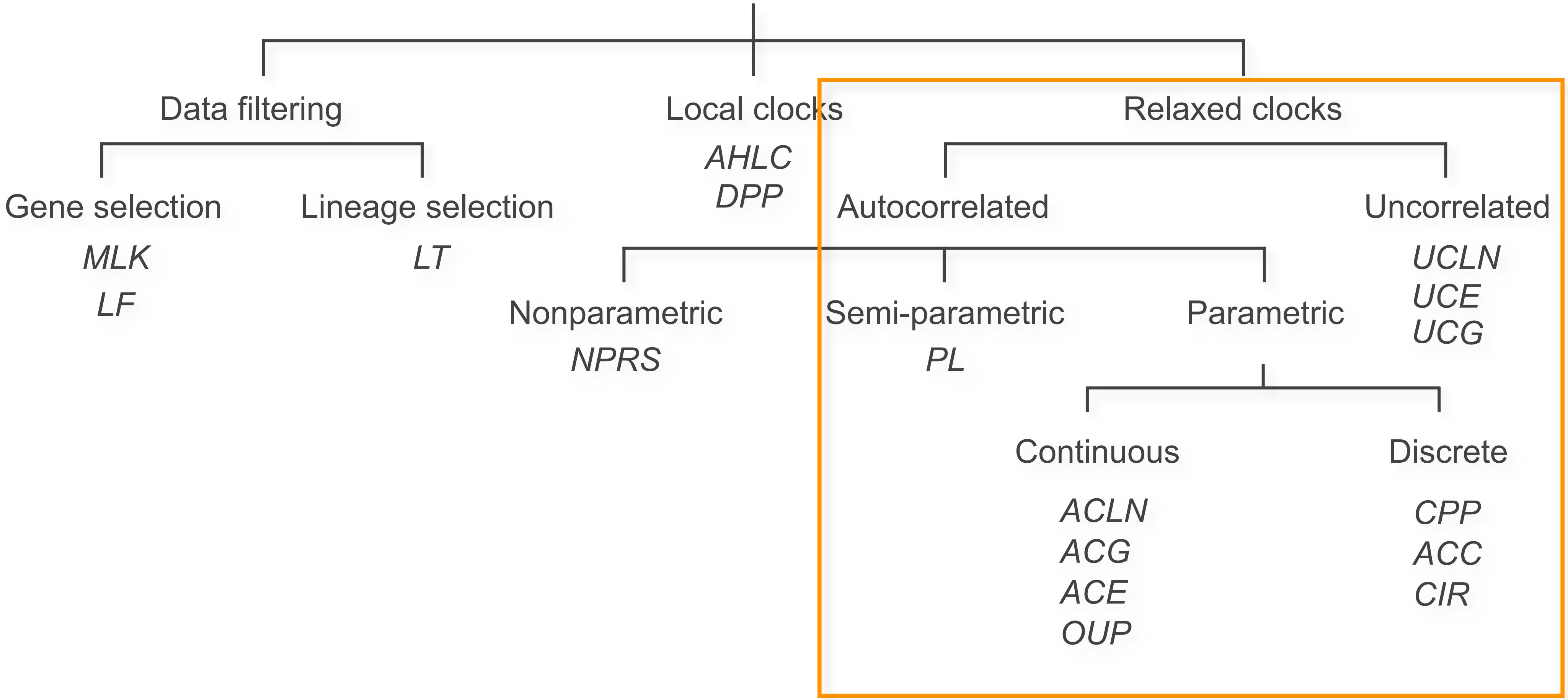
- uncorrelated lognormal branch-rate prior model (*UCLN*)
- uncorrelated gamma branch-rate prior model (*UCG*)
- uncorrelated exponential branch-rate prior model (*UCE*)
- independent gamma branch-rate prior model (*IGR*)

You can (and *should*) ask your data which probability distribution best reflects the process of substitution rates variation by they were generated

Application of Relaxed-Clock Models

Inference under relaxed-clock models is based on *the* model

So many models to choose from!!



Data filtering

Gene selection

MLK

LF

Lineage selection

LT

Local clocks

AHLC

DPP

Relaxed clocks

Autocorrelated

Nonparametric

NPRS

Semi-parametric

PL

Parametric

Continuous

ACLN

ACG

ACE

OUP

Discrete

CPP

ACC

CIR

Uncorrelated

UCLN

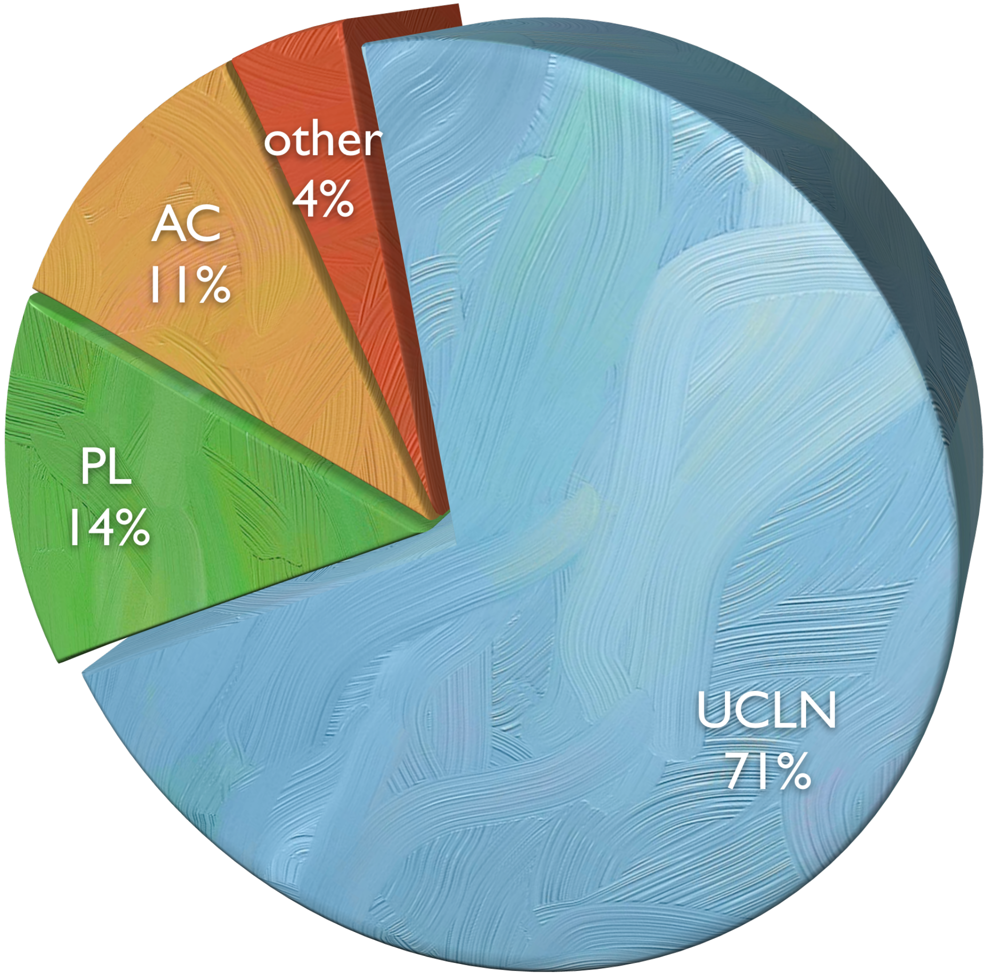
UCE

UCG

Application of Relaxed-Clock Models

Inference under relaxed-clock models is based on *a* model

Methods used in empirical divergence-time studies between 2008–2014:

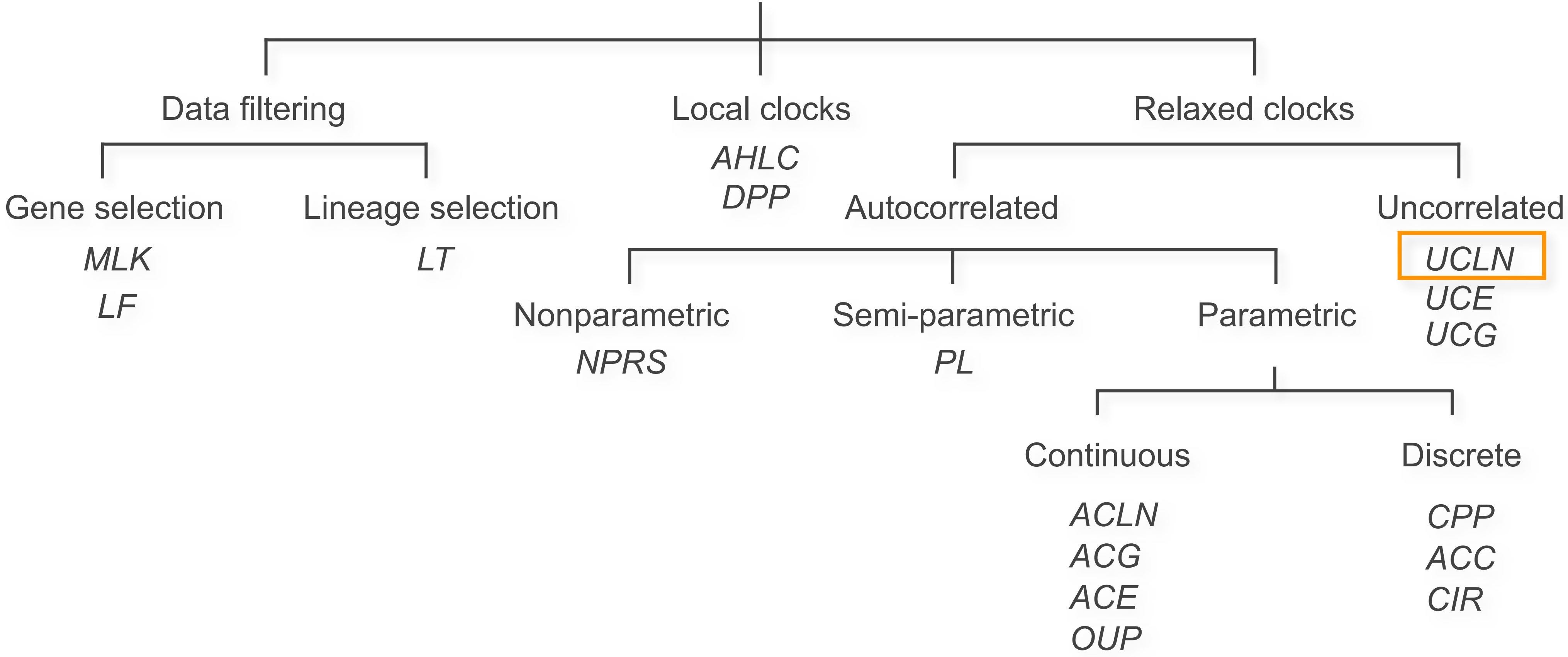


(studies published in *Systematic Biology*, *MPE*, and *Systematic Botany*)

Application of Relaxed-Clock Models

Inference under relaxed-clock models is based on *a* model

But everyone uses the same model! Maybe it doesn’t matter???



Data filtering

Gene selection

MLK

LF

Lineage selection

LT

Local clocks

AHLC

DPP

Relaxed clocks

Autocorrelated

Nonparametric

NPRS

Semi-parametric

PL

Parametric

Continuous

ACLN

ACG

ACE

OUP

Discrete

CPP

ACC

CIR

Uncorrelated

UCLN

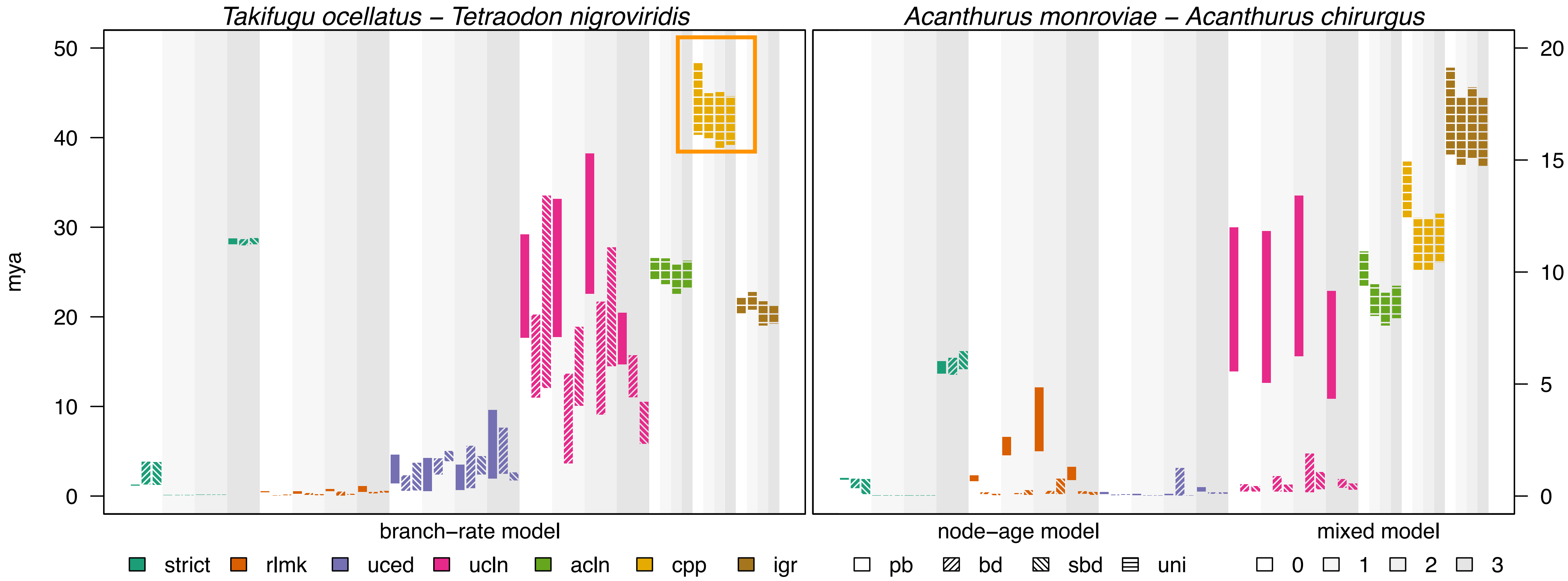
UCE

UCG

Application of Relaxed-Clock Models

Inference under relaxed-clock models is based on the model

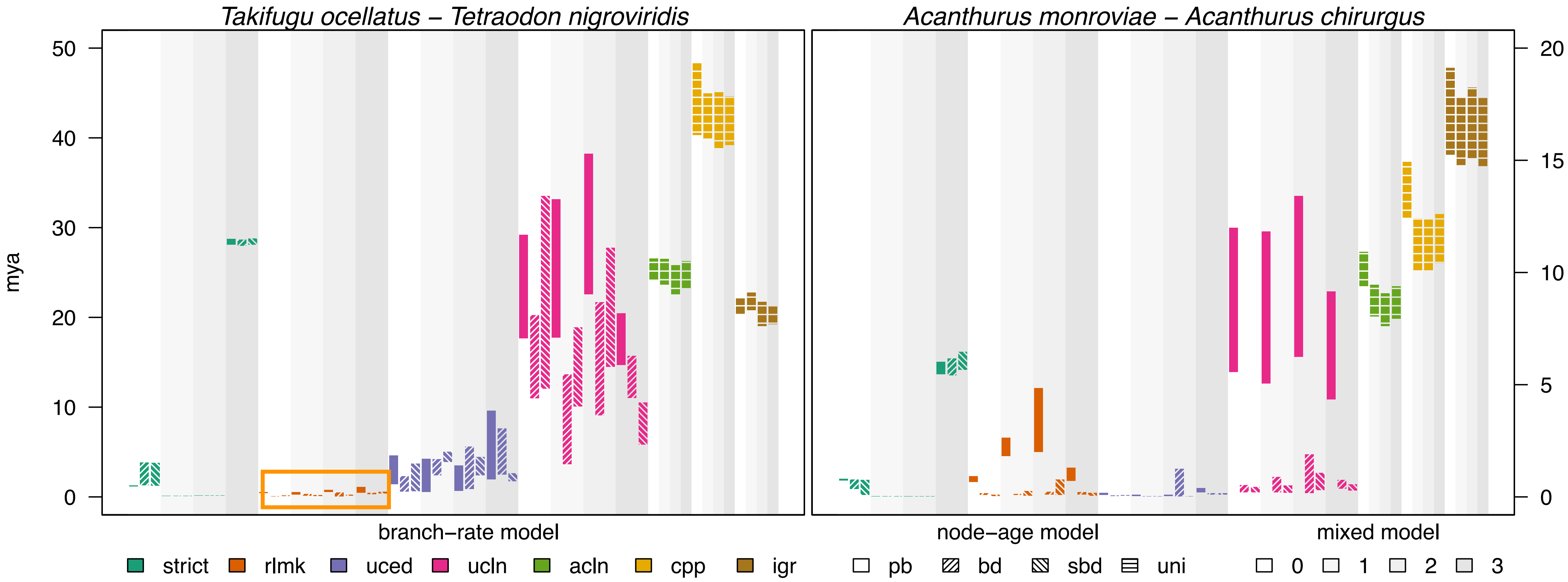
You shouldn't be surprised to learn that the model matters!!
Depending on the model, these species either diverged ~45 Mya



Application of Relaxed-Clock Models

Inference under relaxed-clock models is based on the model

You shouldn’t be surprised to learn that the model matters!!
Depending on the model, these species either diverged ~45 Mya, or ~1Mya!!



To Calibrate or Not To Calibrate?

Depending on your interests, you may not need calibration

For many inference problems, estimates of *relative* divergence times/substitution rates may be adequate :

- character evolution
- lineage diversification
- rates and patterns of molecular evolution

For other inference problems, estimates of *absolute* divergence times/substitution rates may be necessary:

- biogeography
- co-evolution
- epidemiology
- events in Earth history