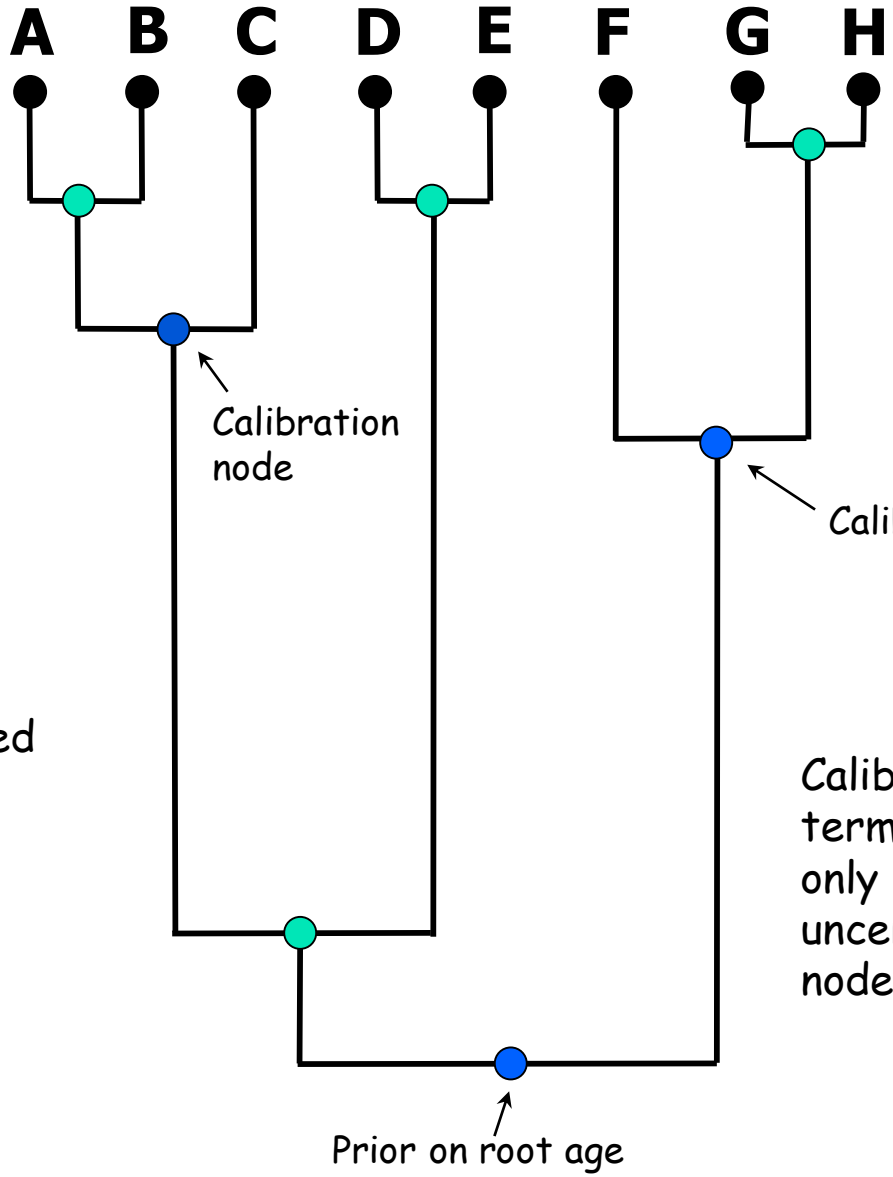
Dating Phylogenies with Fossils

Fredrik Ronquist

Dept. Bioinformatics & Genetics
Swedish Museum of Natural History

Node dating



Molecular analysis of extant taxa using (relaxed) clock model

Node dates
- Fixed
- Drawn from prior
- Unconstrained

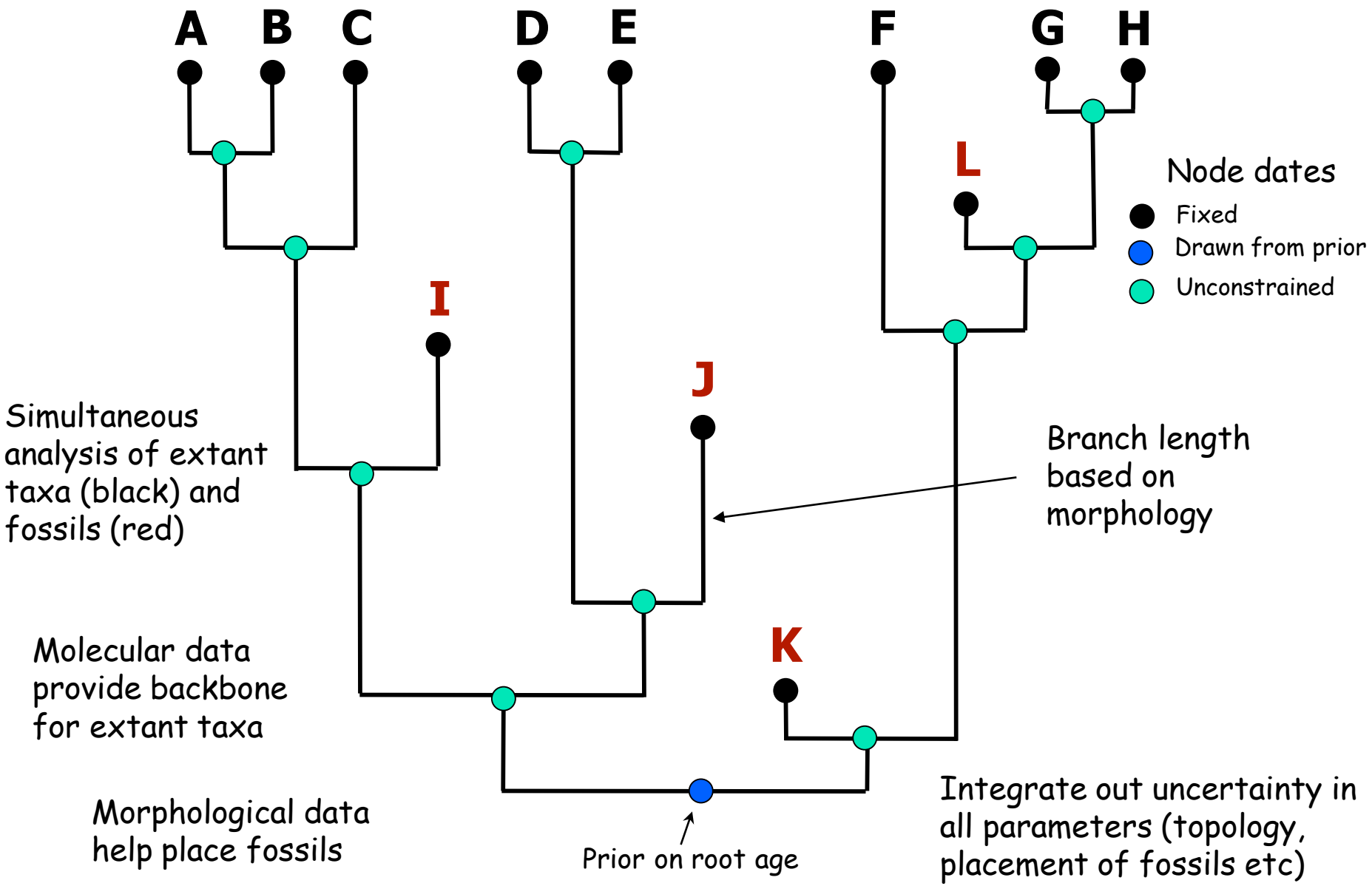
Fossil record is used to derive priors on calibration points

Calibration nodes are fixed in terms of topology -> we can only integrate out topological uncertainty concerning other nodes

Potential Shortcomings

- Phylogeny is not known with certainty but we have to fix clades corresponding to calibration nodes
- Unclear how to derive appropriate calibration distributions
- Fossil placement is often uncertain; unclear if this can be accommodated in the calibration distributions
- Does not incorporate all the data in the analysis
- You have to summarize many fossils in a few calibration points

Total-evidence dating



Total-evidence dating

- Also called tip dating or integrative dating
- Treats fossils and extant taxa in the same analysis
- Fossils placed in the tree according to morphological evidence and assuming a 'morphological clock'
- Relationships among extant taxa usually based on molecular characters
- Using no internal node calibrations derived indirectly from the fossil record
- Fossil ages determined using rock dating methods
- Integrating over the uncertainty in the phylogenetic placement of fossils
- A platform for reconciling evidence from rocks and clocks directly in the same analysis, using probability as the common arbiter

Early radiation of the Hymenoptera

- Documented by a number of incomplete impression fossils that are difficult to place phylogenetically
- 45 fossil and 68 extant taxa
- 343 morphological characters
- Fossil completeness 4 – 20 %
- 5 kb sequence data from 7 markers
- Phylogenetic model:
  - Mk model of morphology
  - Codon-site-partitioned GTR+I+G : SYM+I+G
  - Non-clock, strict clock and relaxed clocks


Abdome
Tórax
Cabeça

Morphological models

- Variable state space
- Arbitrary state labels
- Sampling (ascertainment bias): only variable characters observed
- Different models for ordered and unordered characters

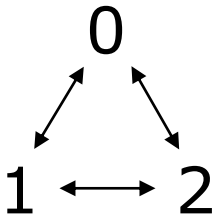
Morphological Models

- Based on Lewis (2002; Syst. Bio.) with several extensions
- Varying state space (k = 2 to k = 10)
- Unordered and ordered characters
- Incomplete coding (coding bias or ascertainment bias)

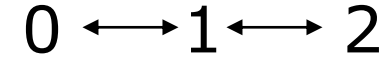
Transformation series

Parsimony Types:

Unordered (Fitch)



Ordered (Wagner)



Probabilistic models

Unordered (M3u)

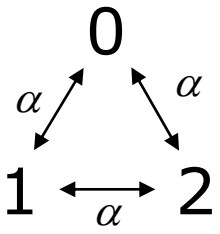
$$Q = \begin{pmatrix} -2\alpha & \alpha & \alpha \\ \alpha & -2\alpha & \alpha \\ \alpha & \alpha & -2\alpha \end{pmatrix}$$

Ordered (M3o)

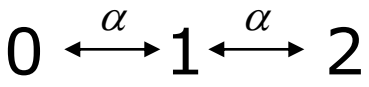
$$Q = \begin{pmatrix} -\alpha & \alpha & 0 \\ \alpha & -2\alpha & \alpha \\ 0 & \alpha & -\alpha \end{pmatrix}$$

Probabilistic models

Unordered (M3u)



Ordered (M3o)



Incomplete coding

A	B	C	D
0	0	1	1
0	0	0	1
0	0	0	0

All

Incomplete coding

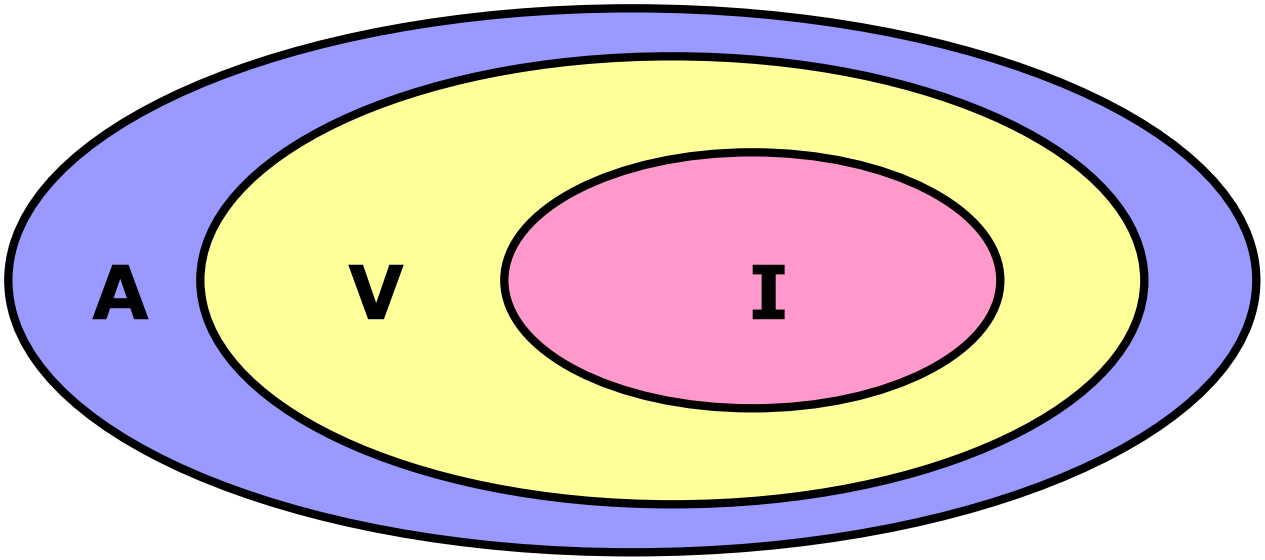
A	B	C	D	
0	0	1	1	**Variable**
0	0	0	1	
0	0	0	0	

Incomplete coding

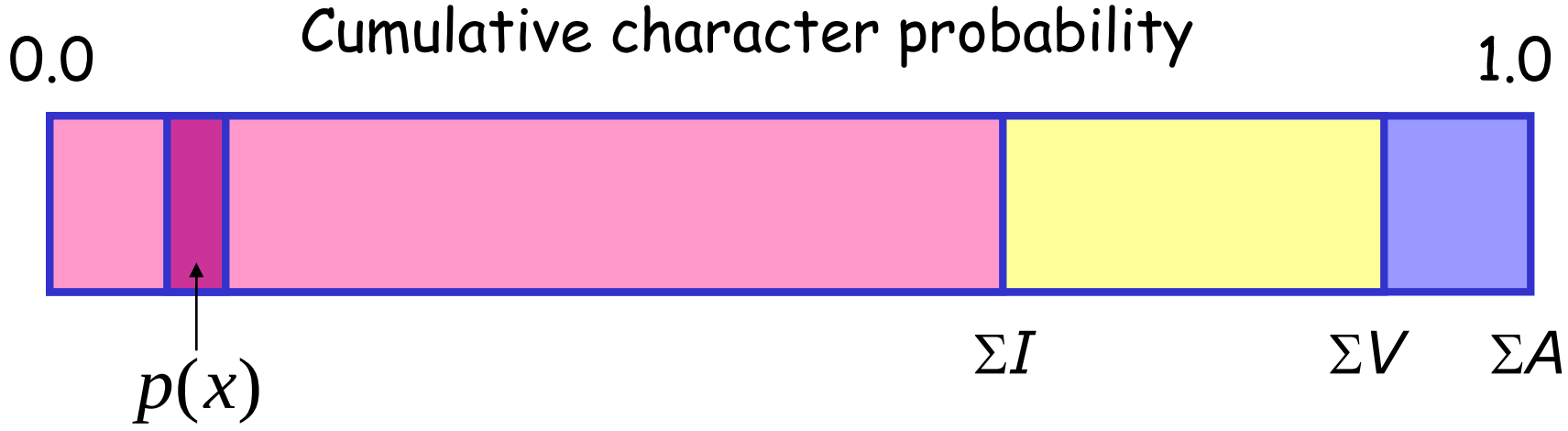
A	B	C	D	
0	0	1	1	**Informative**
0	0	0	1	
0	0	0	0	

Types of characters

A (All), V (Variable), I (Informative)



Conditional character probability

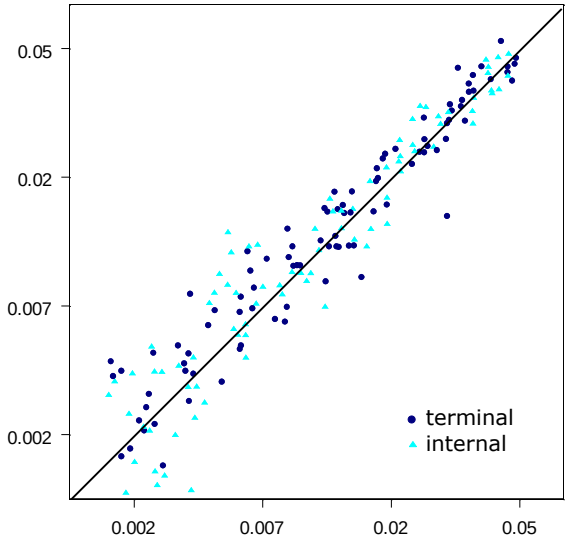


Conditional probability of one character x_i given that only informative characters are coded:

$$p(x_i \mid x_i \in I) = \frac{p(x_i)}{\sum_j p(x_j) : x_j \in I}$$

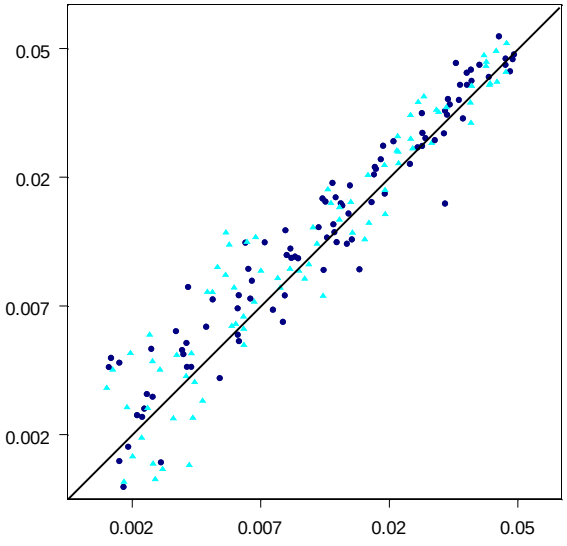
Branch length estimates

All characters
Assuming all



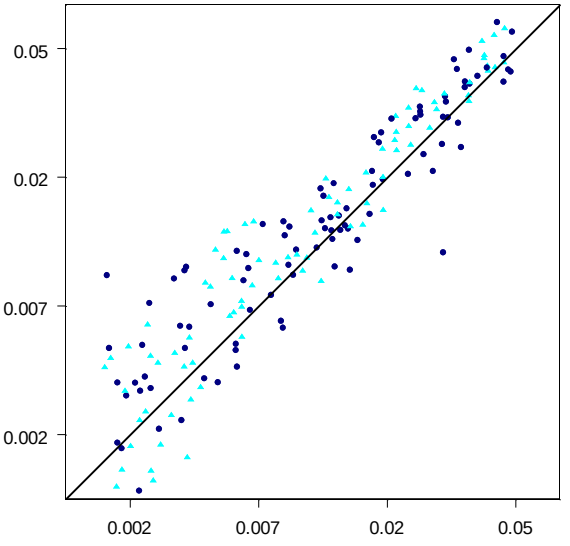
2000 chars

Variable chars
Assuming variable



1176 chars

Informative chars
Assuming informative

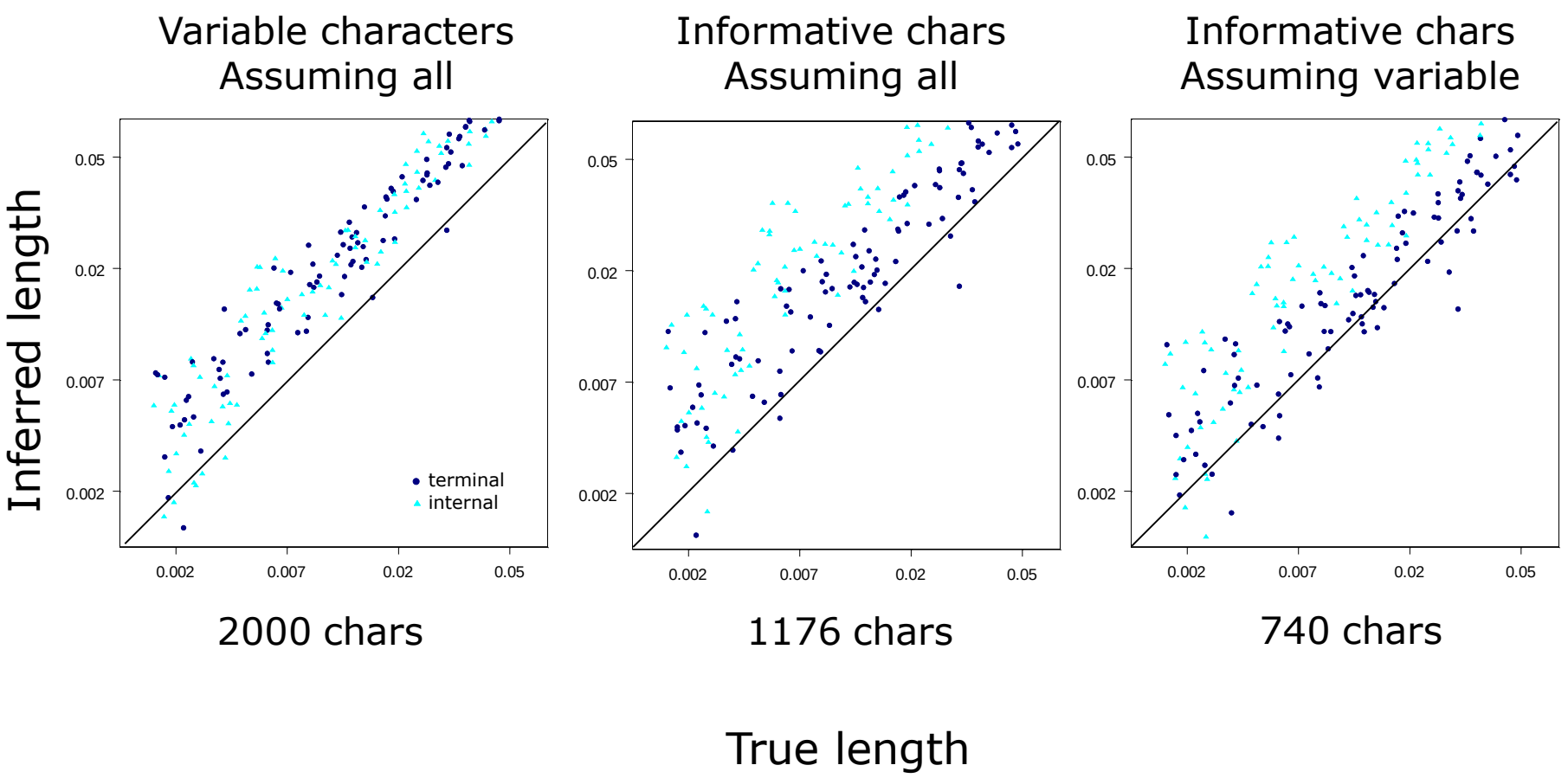


740 chars

Inferred length

True length

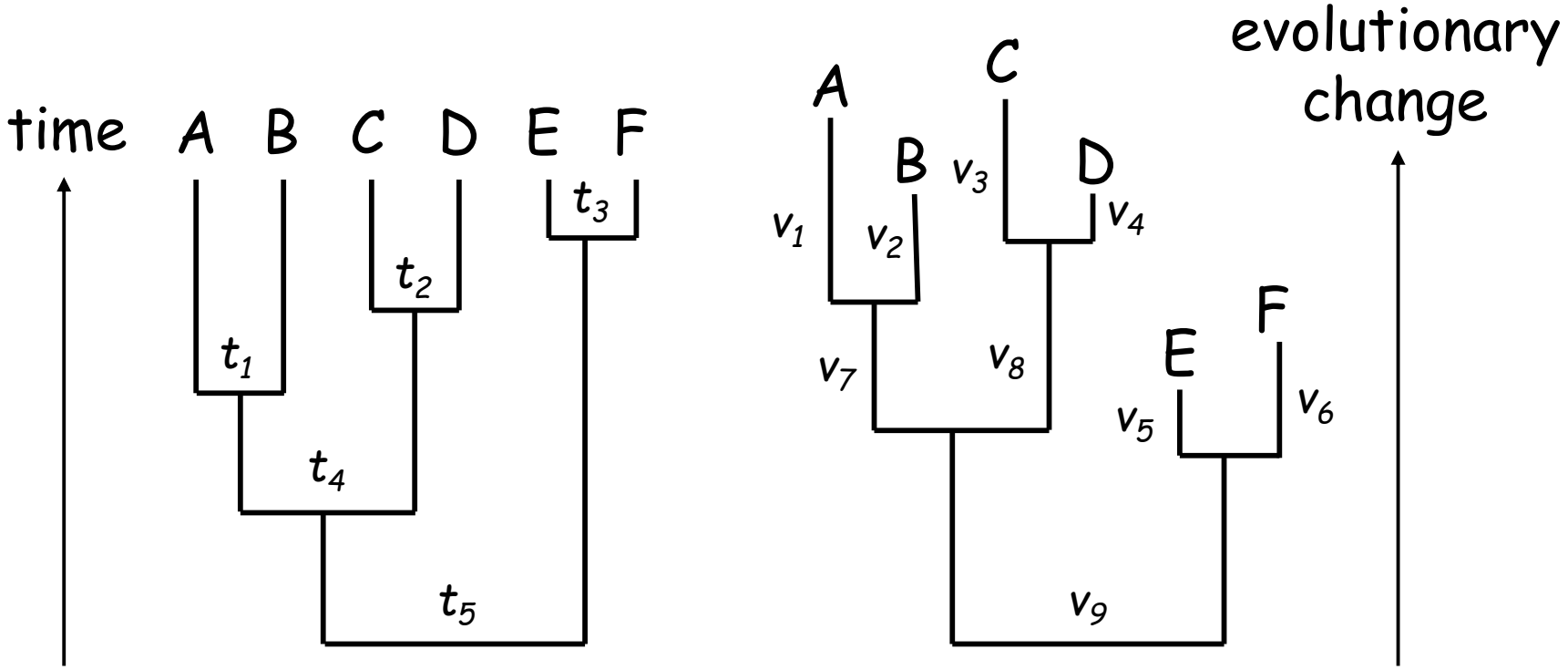
Branch length estimates



Relaxed clock models

- Thorne-Kishino 2002 (TK02) model: continuous autocorrelated model
- Compound Poisson process (CPP) model: discrete autocorrelated model
- Independent gamma rates (IGR) model: uncorrelated continuous model

Clock and Non-clock Trees



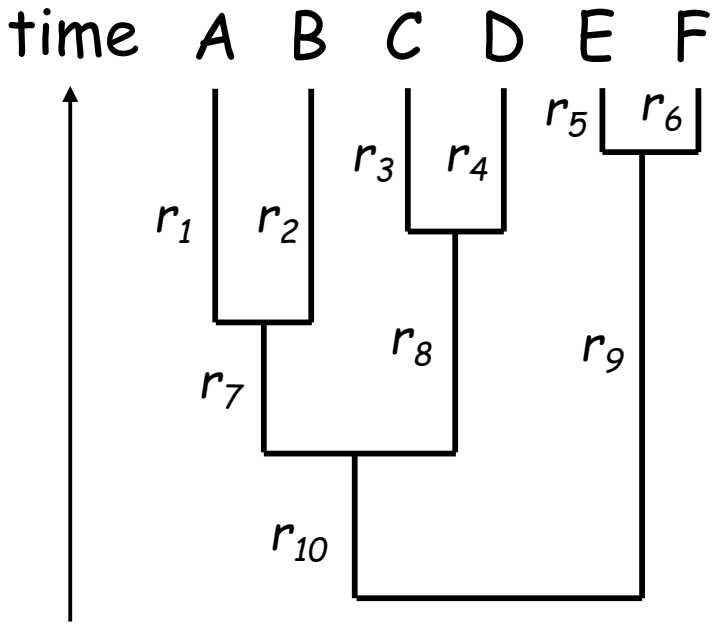
Clock tree

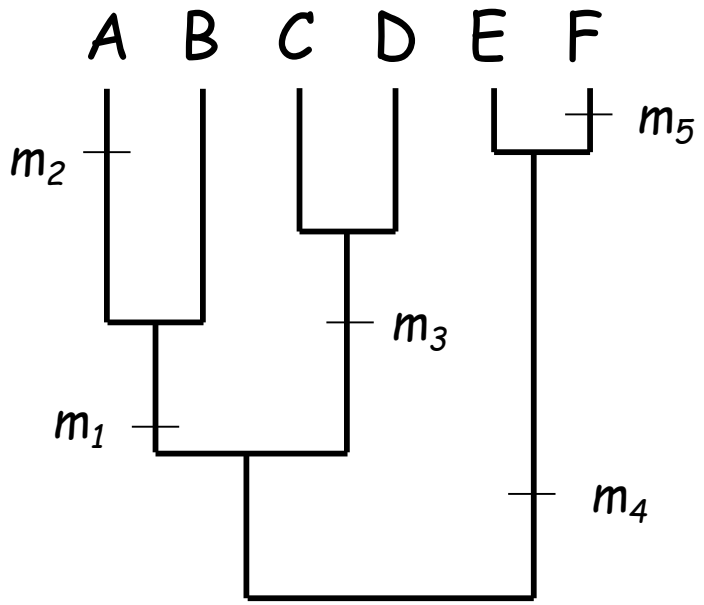
n – 1 node times

Non-clock tree

2n – 3 branch lengths

Relaxed clocks and dating





Branch rate models:
r_i follow Brownian motion
r_i drawn iid
both cases one variance param.

Compound Poisson Process (CPP):
Rate multipliers m drawn iid and generated according to a Poisson process; variance and rate parameters

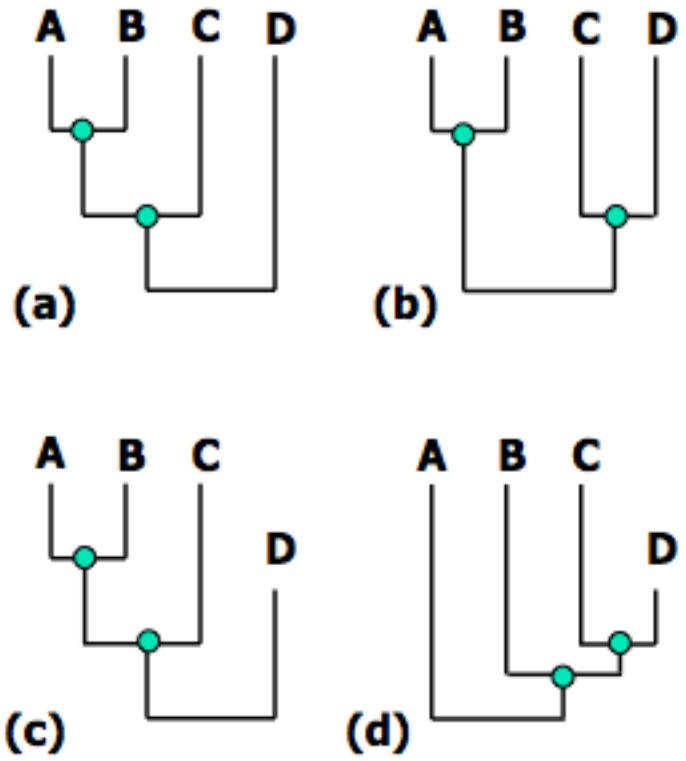
Relaxed clocks and dating

- MrBayes implements three relaxed clock models:
  - The Compound Poisson Process (CPP) relaxed clock (discrete autocorrelated model)
  - The Thorne-Kishino 2002 (TK02) model (continuous autocorrelated model)
  - The Independent Gamma Rates (IGR) model (continuous truly uncorrelated model)
- Date using tip and/or node calibrations
- Dates can be fixed or associated with uncertainty
- Rich summaries from sumt, including effective branch lengths, rates and ages
- Summary trees guaranteed to be clock trees and have positive branch lengths

Tree model for total-evidence dating

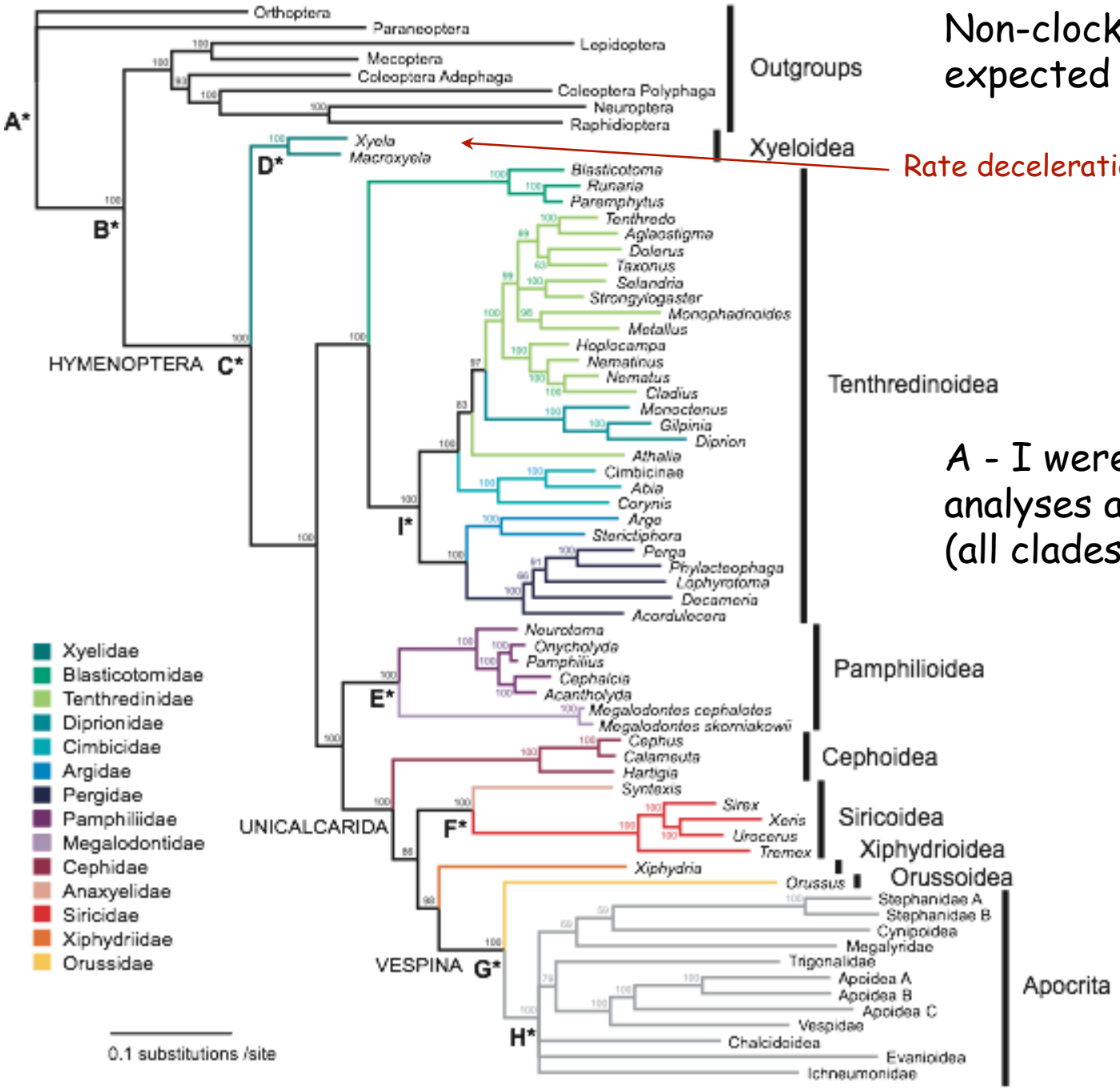
- Coalescent model: not relevant model for higher-level phylogenies
- Birth-death model: problem of modeling speciation, extinction, sampling and fossilization
- Uniform model: can be extended to serially sampled trees

Uniform prior on serially sampled trees



Two approaches to dating

- Node dating
  - 68 extant taxa
  - Seven Hymenoptera calibration points derived from 45 fossils (C-I)
  - Two outgroup calibration points (A-B)
  - Offset exponential priors, mean being min of the next more inclusive calibration point
  - Calibrations set together with the leading paleontological and morphological experts on the Hymenoptera
- Total-evidence dating
  - 68 extant + 45 fossil taxa in simultaneous analysis
  - Position of fossil taxa determined by morphological characters (343 characters in total, 4 – 20% coded for fossils)
  - Extant phylogeny mostly determined by molecular characters (5 kb sequence data from 7 markers)
  - All calibration constraints removed except the two outgroup calibrations

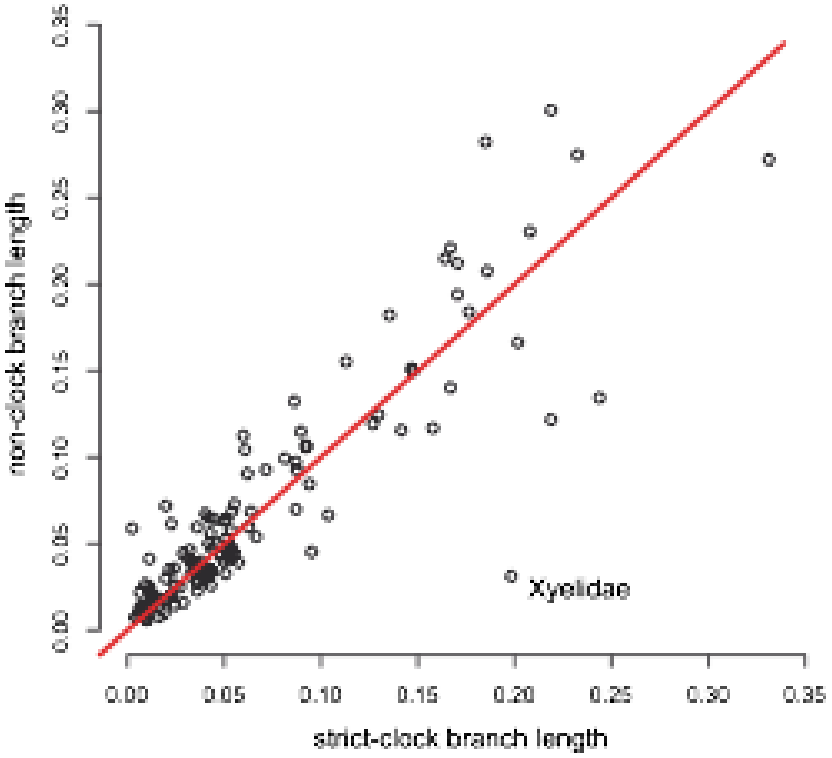


Non-clock tree retrieves expected relationships

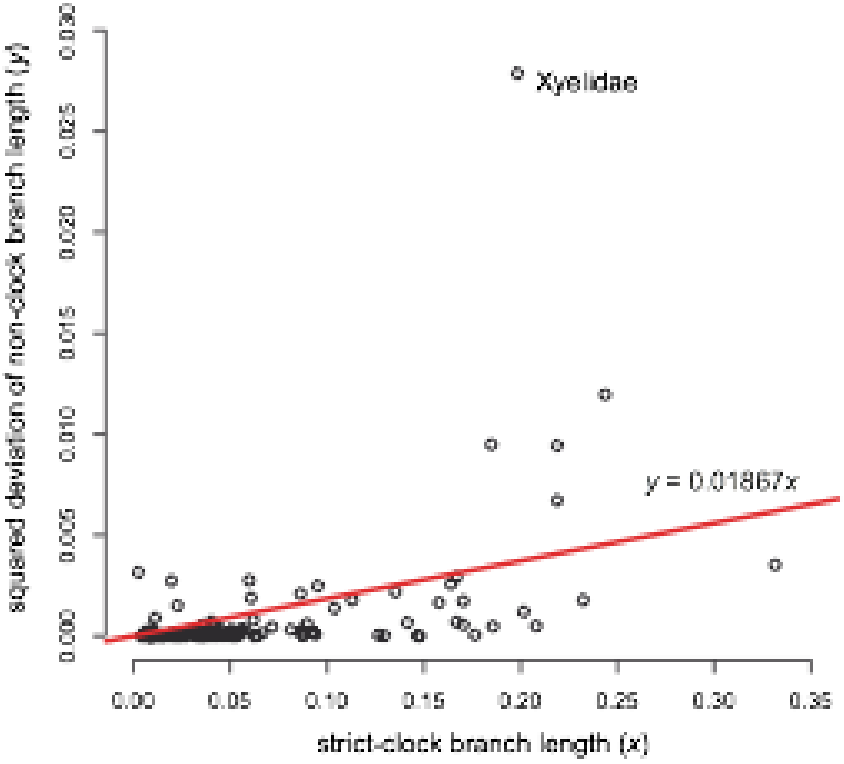
A - I were used in clock analyses as calibration points (all clades well supported)

Comparing strict clock to non-clock branch lengths sampled from the non-clock topology

Correlation

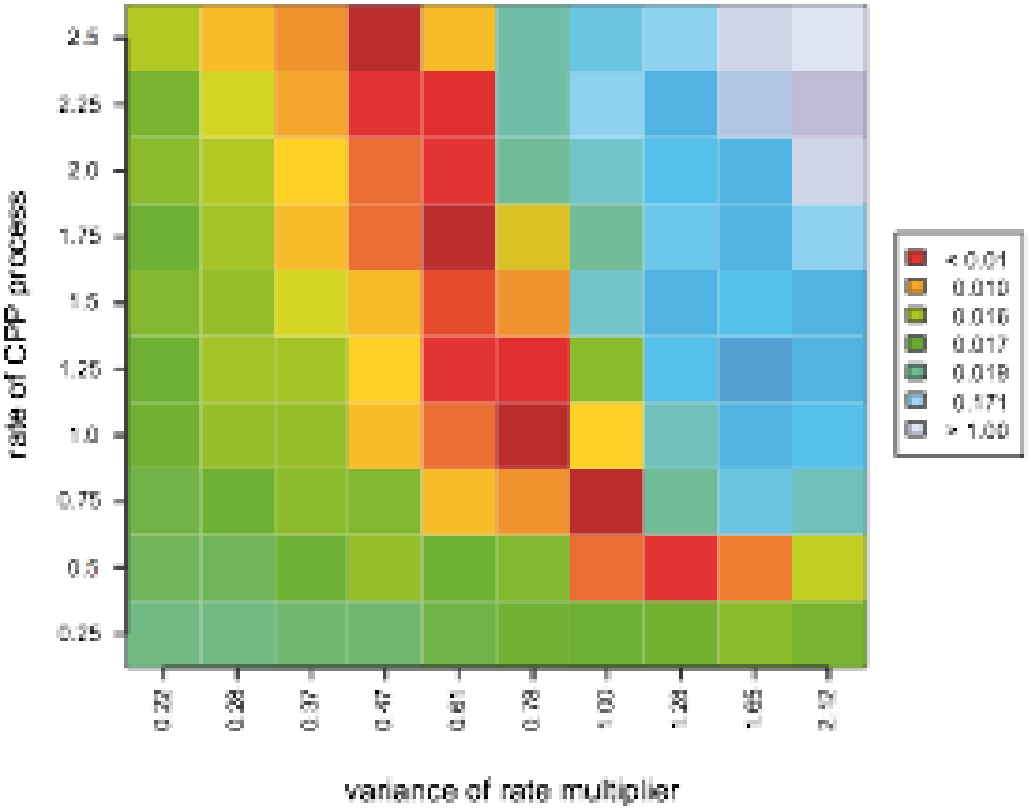


Squared deviation



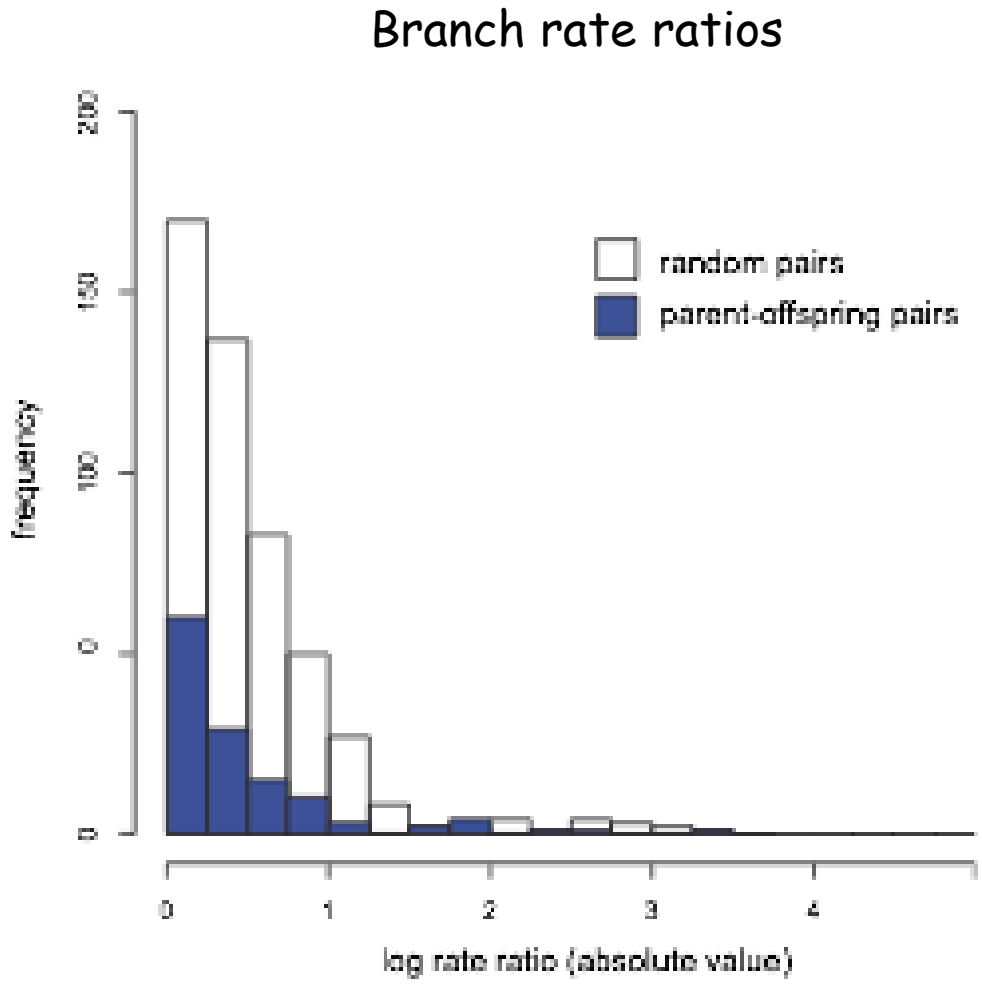
Finding suitable priors for the CPP model

Poisson rate – multiplier variance space

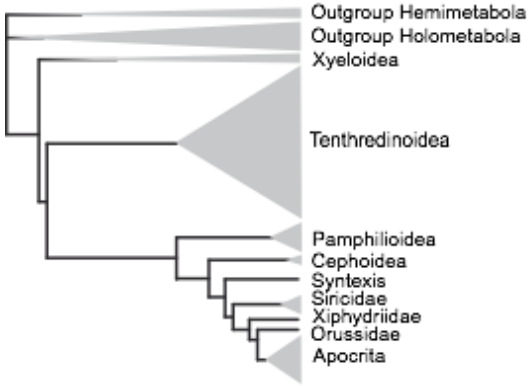


Deviation from observed variance

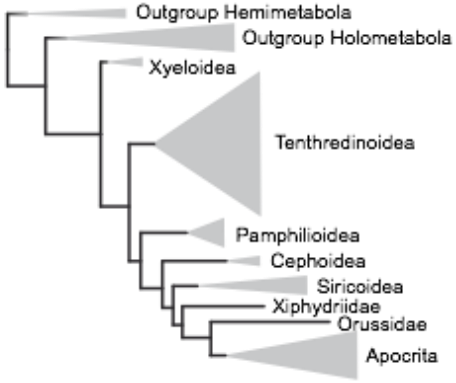
Slight but significant rate autocorrelation



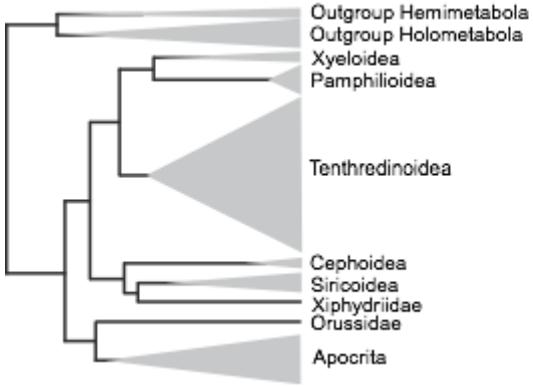
Morphology



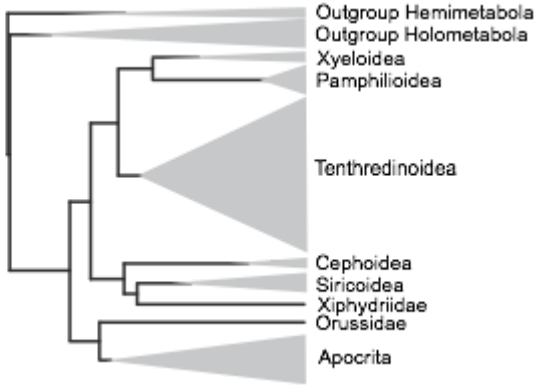
Non-clock



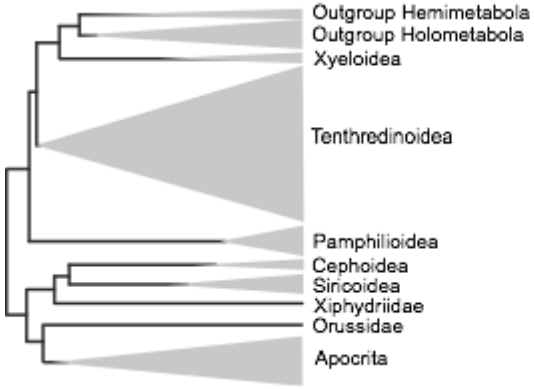
Strict clock



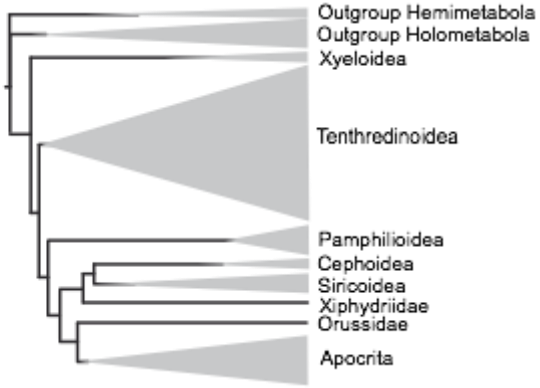
Strict clock with rooting constraint



Relaxed clock



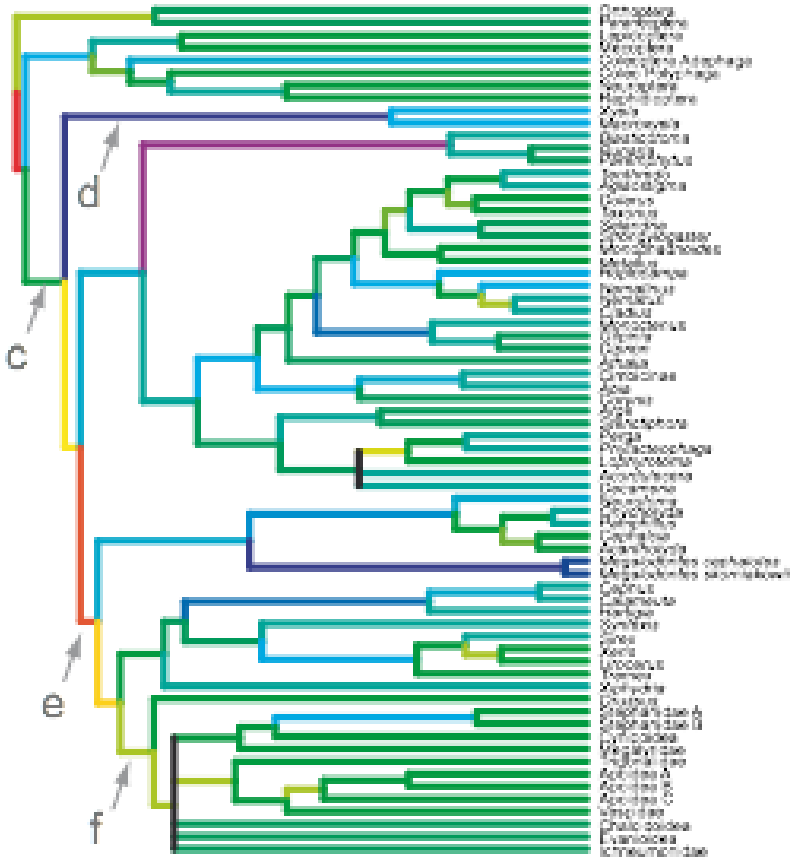
Relaxed clock with rooting constraint



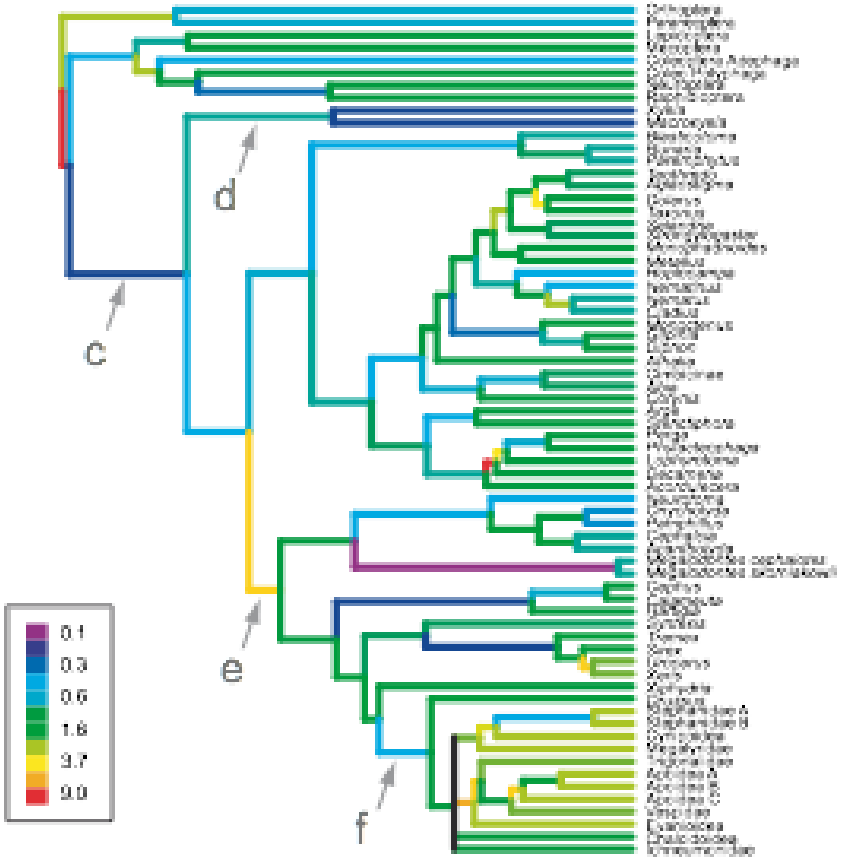
Relaxed clock models may need rooting constraint

Rate variation across the tree

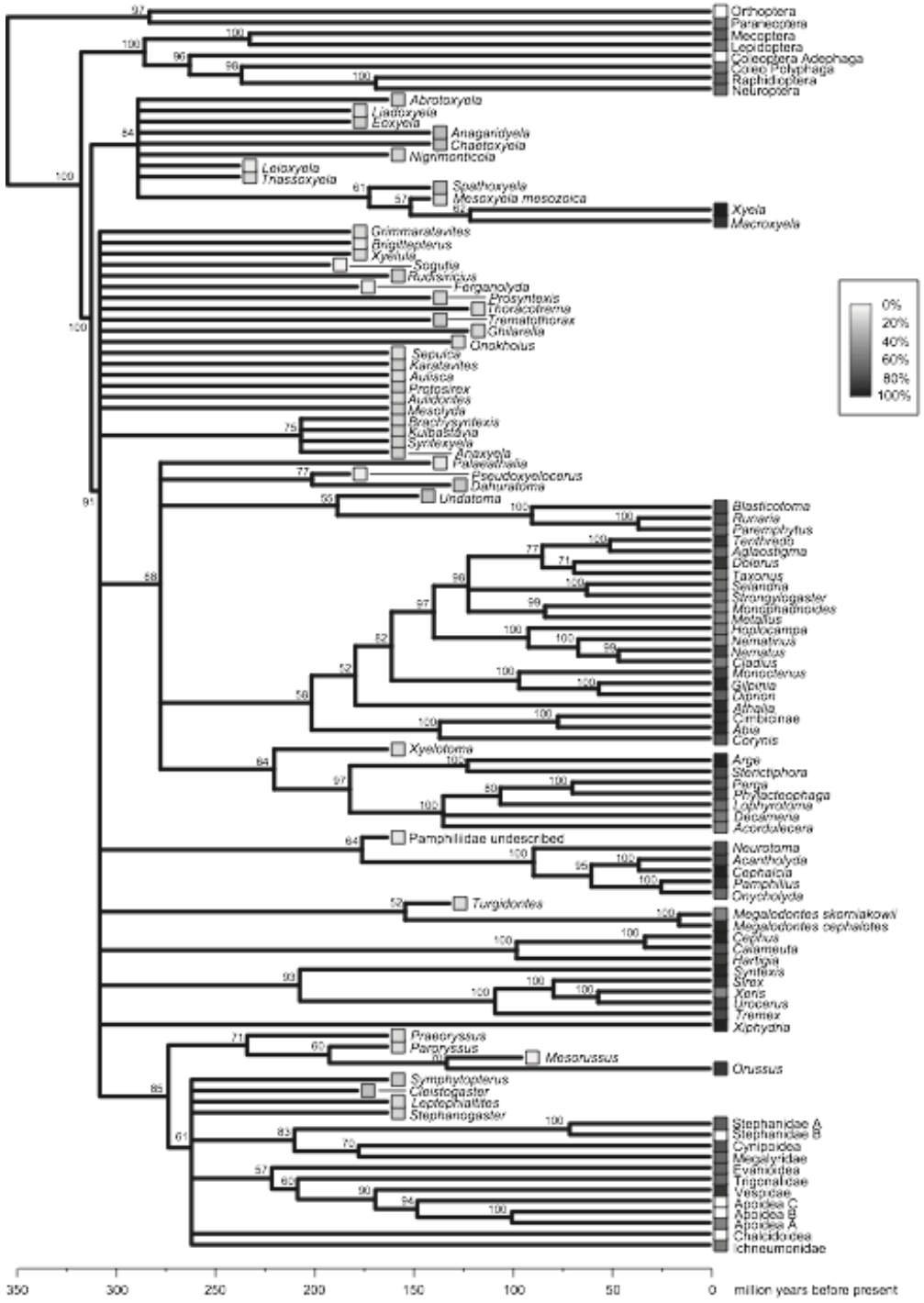
IGR model



CPP model

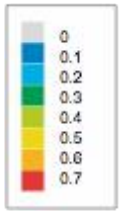


Majority rule consensus with fossils

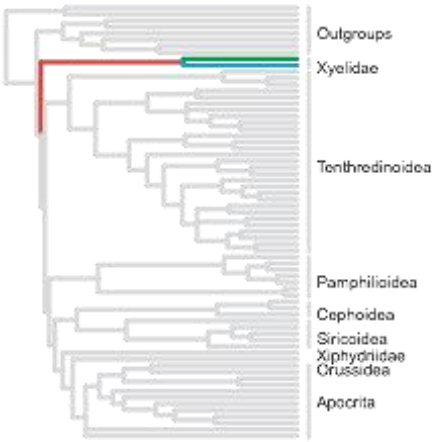


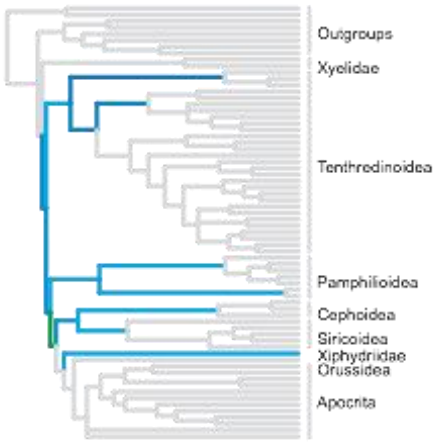
Completeness of morphology scores

Mesoxyela mesozoica

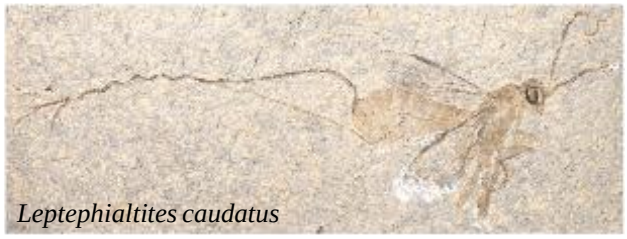


Sogutia liassica

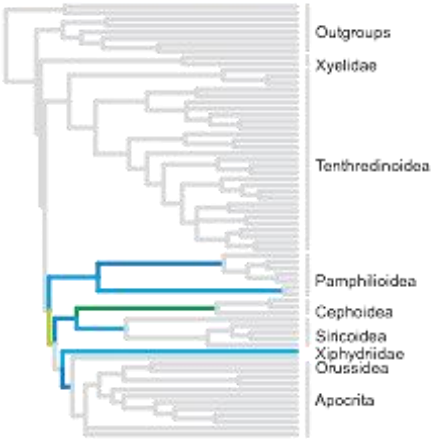


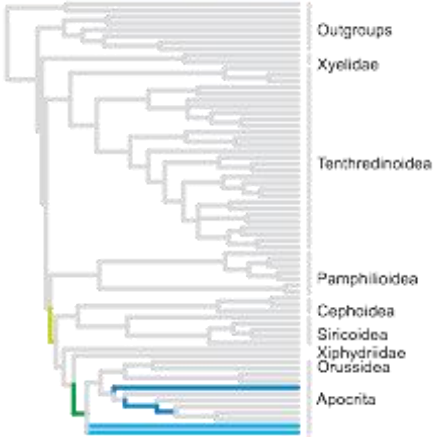


Aulisca odontura



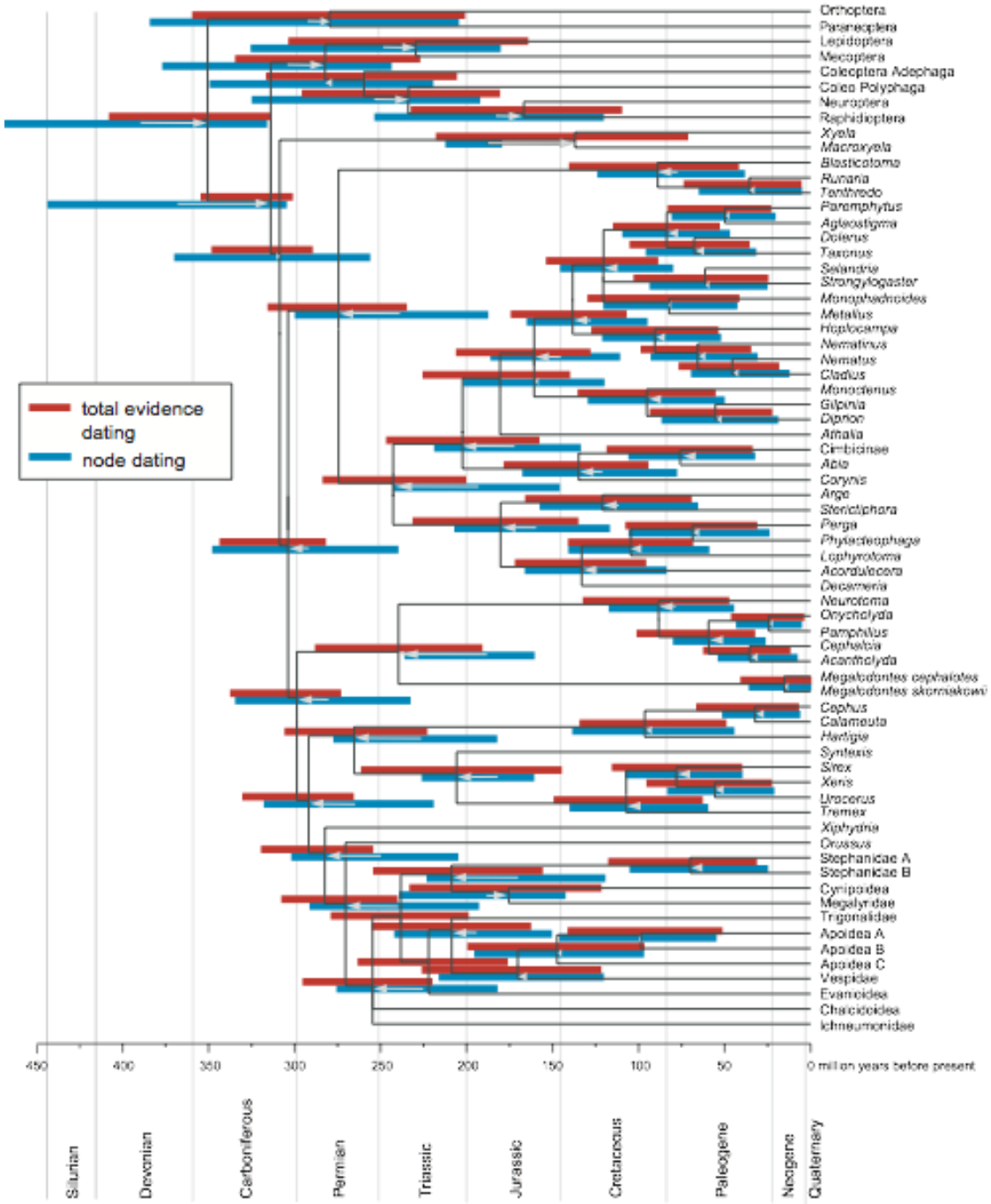
Leptephialtites caudatus





Uncertainty in the phylogenetic position varies across fossils

Estimated divergence times



Posteriors on node ages

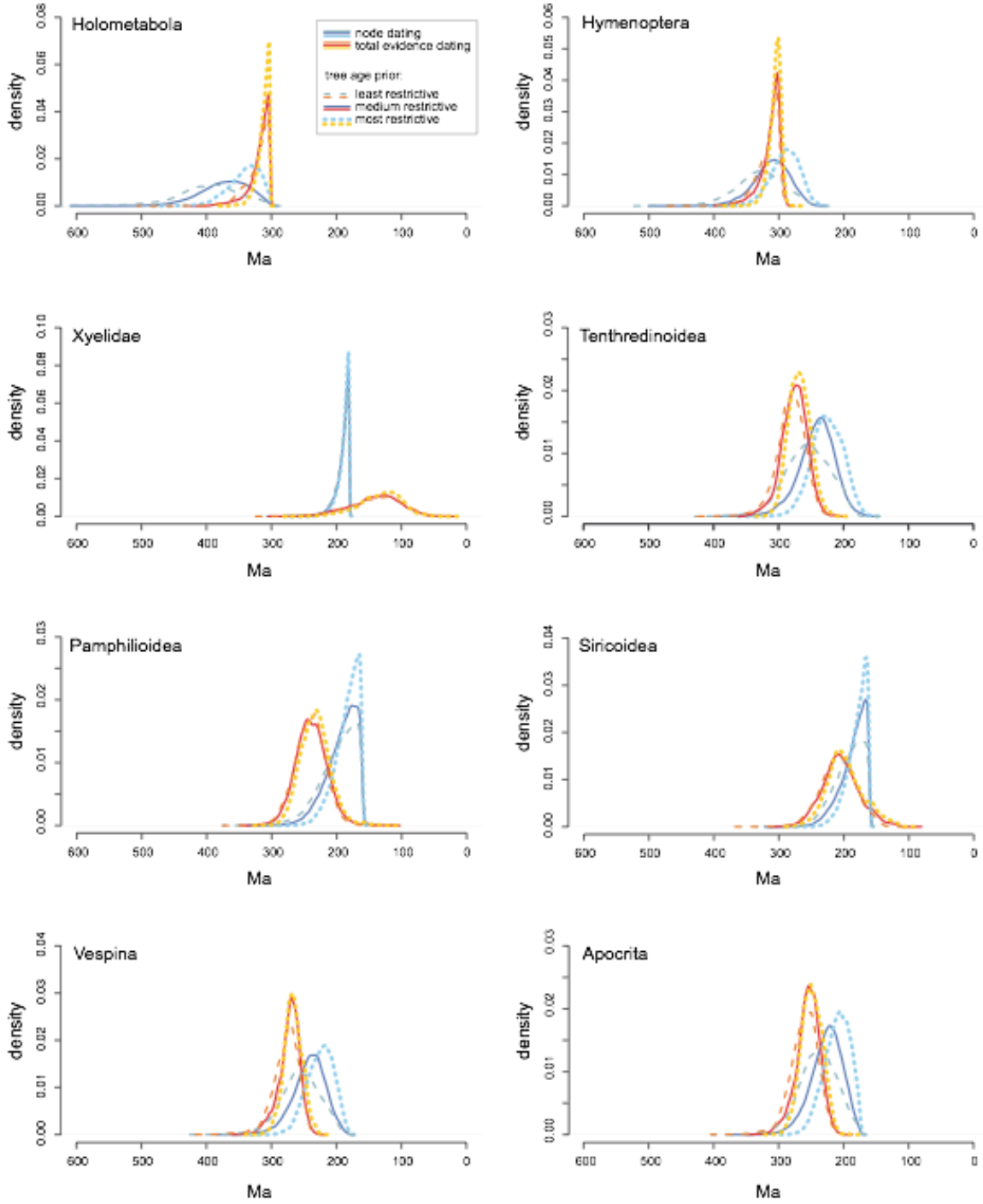


Table 2. Fossils used in the node-dating analyses and calibration point prior settings. The minimal and mean age for the offset-exponential prior are given, along with the corresponding fossils and references.

Calibration point	Prior on age (Ma)	Fossil(s)	Reference	PP correct[1]
A. Neoptera	min: 315	*Katerinka* (oldest Neoptera)	Prokop & Nel 2007	
	mean: 396	*Rhyniognatha* (oldest insect)	Engel & Grimaldi 2004	
B. Holometabola	min: 302	insect gall (oldest Holometabola)	Labandeira & Philips 1996	
	mean: 396	*Rhyniognatha*	Engel & Grimaldi 2004	
C. Hymenoptera	min: 235	*Triassoxyela, Asioxyela*	Rasnitsyn & Quicke 2002	96%
	mean: 302	insect gall	Labandeira & Philips 1996	
D. Xyelidae[2]	min: 180	*Eoxyela*	Rasnitsyn 1983	0%
E. Pamphilioidea[2]	min: 161	*Aulidontes, Pamphilidae undescribed*	Rasnitsyn & Zhang 2004	48%
F. Siricoidea[2]	min: 161	*Aulisca, Anaxyela, Syntexyela, Kulbastavia, Brachysyntexis*	Zhang & Rasnitsyn 2006	0%
G. Vespina[2]	min: 180	*Brigittepteris*	Rasnitsyn et al. 2003	7%
H. Apocrita[2]	min: 176	*Cleistogaster*	Rasnitsyn 1975	34%
I. Tentrhedinoidea s.str.[2,3]	min: 140	*Palaeathalia*	Zhang 1985	100%

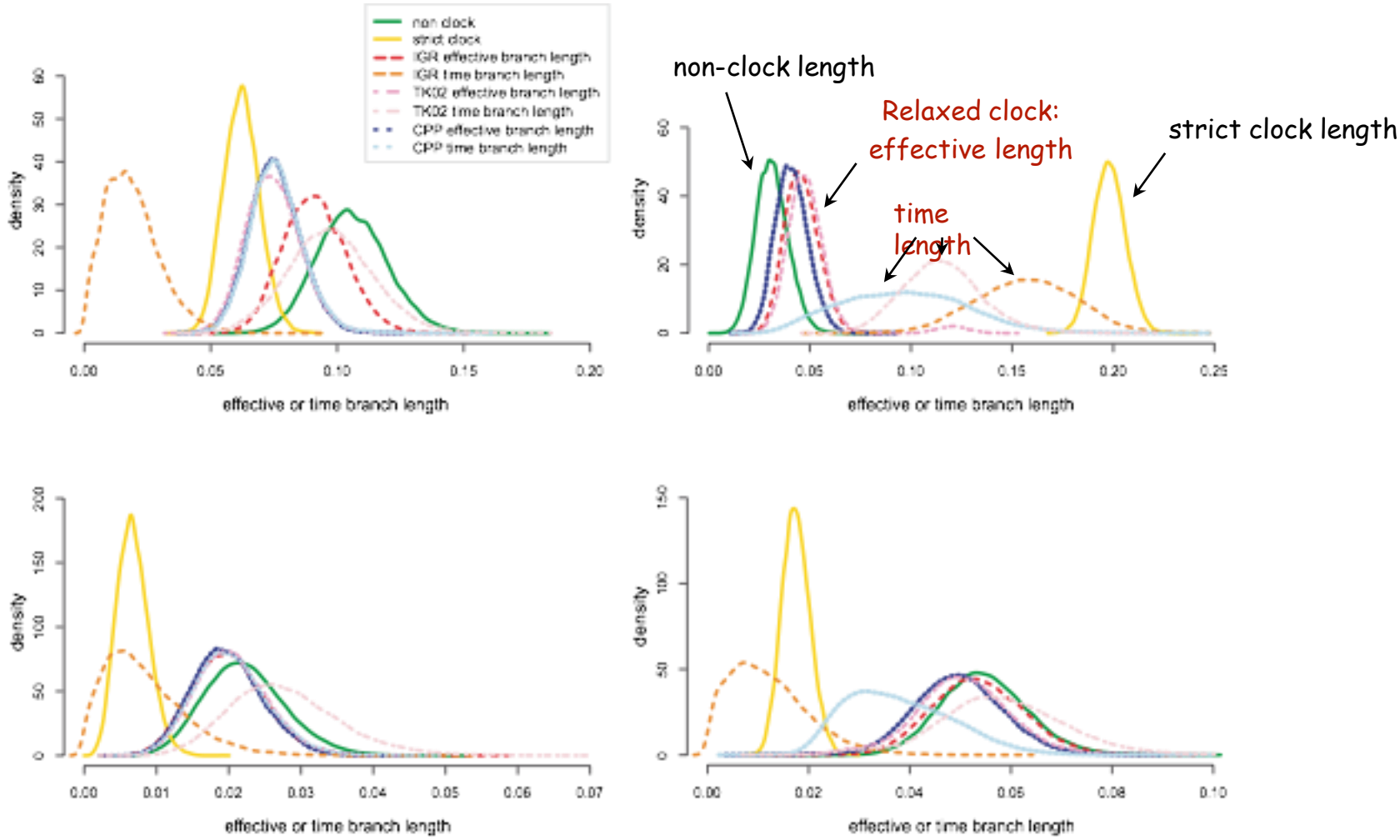
[1]Posterior probability (PP) from the total-evidence analysis that the fossil attaches at the position assumed in the node-dating analysis. Note that these posterior probabilities take both the morphological data and the ages of the fossils into account.

[2]The mean age for all intra-hymenopteran calibration points was assumed to be the minimal age of Hymenoptera, i.e. 235 Ma (*Triassoxyela, Asioxyela*).

[3]Tenthredinoidea excluding Blasticotomidae.

Error in divergence time estimation is not influenced to a large extent by molecular character data

Branch length posteriors for different models on four example branches



Conclusions 1(2)

- Total-evidence dating is preferable because it:
  - explicitly incorporates fossil evidence
  - allows powerful analysis of the available data
  - results in divergence times that are
    - more precise
    - less sensitive to prior assumptions
    - probably more accurate
  - provides better platform for future development, such as explicit modeling of fossilization, speciation, extinction, and sampling

Conclusions 2(2)

- There is a limit to how much molecular characters can help reduce the errors in divergence time estimates
- Most significant improvements will come from
  - more intense study of the fossil record
  - better understanding of morphological evolution
  - better models of rate variation across sites and lineages
  - better modeling of speciation, extinction, fossilization and sampling of fossil and extant taxa
- Challenges with total-evidence dating under birth-death prior with fossilization:
  - Dealing with trees where fossils are ancestors (sit on branches)
  - Sampling probabilities and biases, both for fossils and extant taxa
  - Uniform fossilization or "slice sampling"
  - Priors for speciation and extinction rates

Birth-death model in phylogenetics

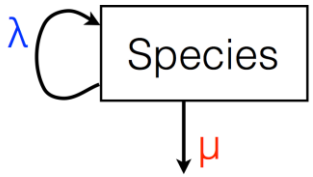
Parameters

λ Speciation rate

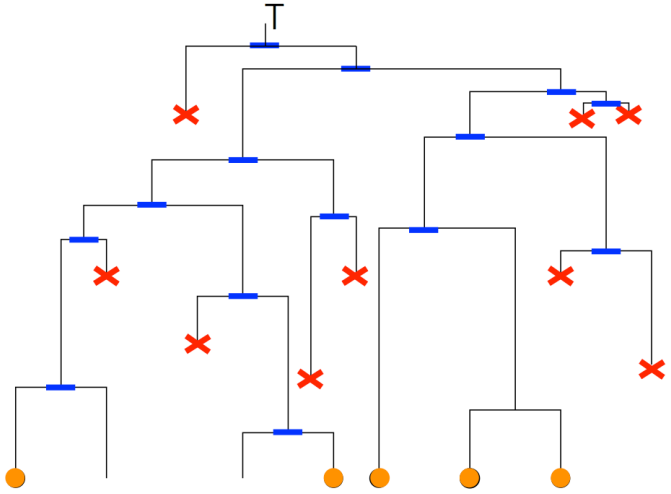
μ Extinction rate

r Sampling probability

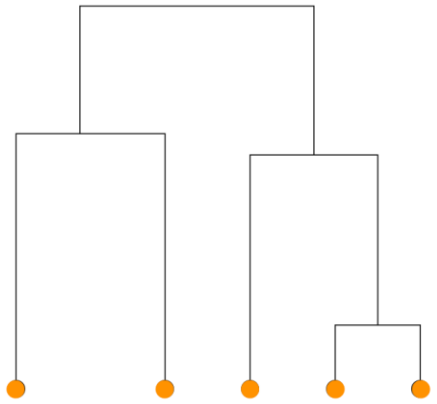
T Time of origin



State machine representation

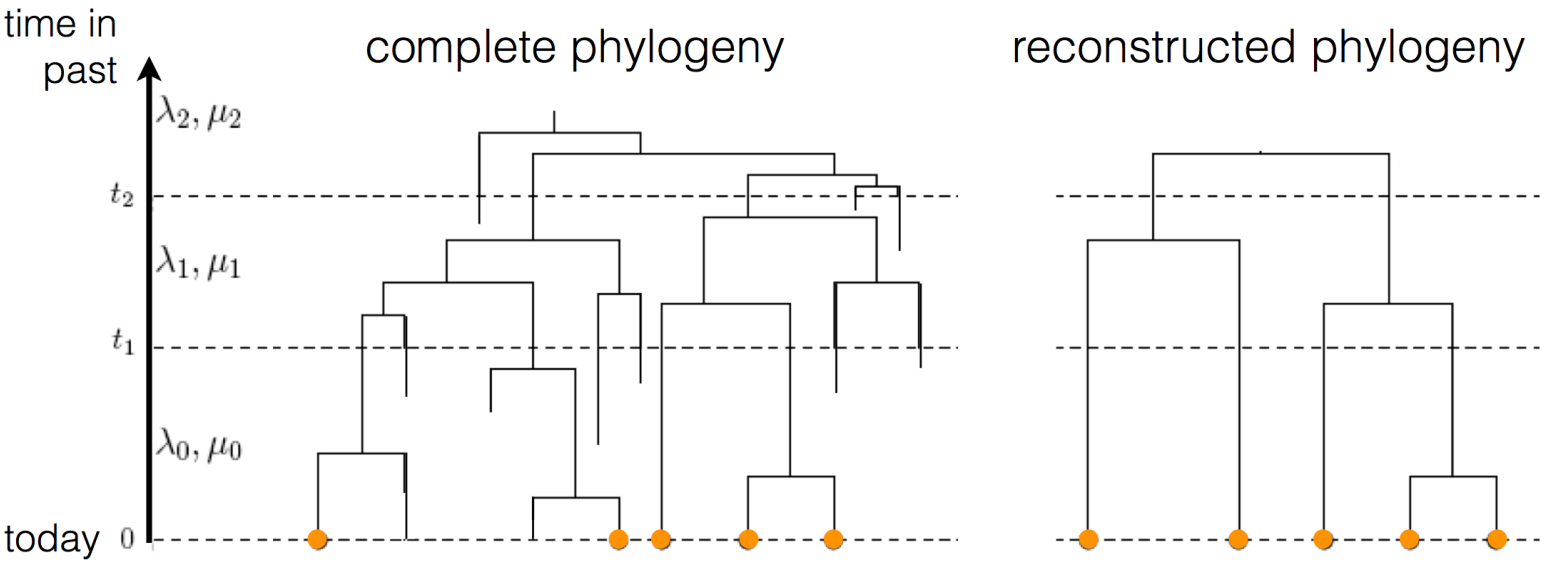


Complete tree



Sampled tree

The piece-wise constant birth-death model



Probability of the reconstructed tree is an integral over all complete trees. It can be calculated efficiently using recursion and by solving differential equations.

Stadler 2011, PNAS

The fossilized birth-death (FBD) model

Parameters

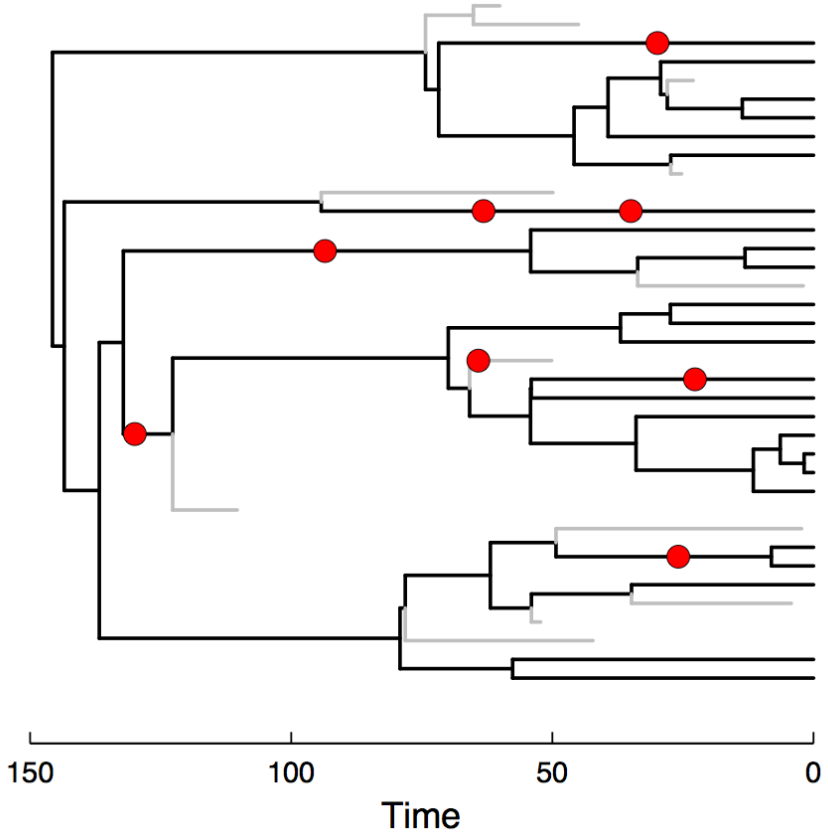
λ Speciation rate

μ Extinction rate

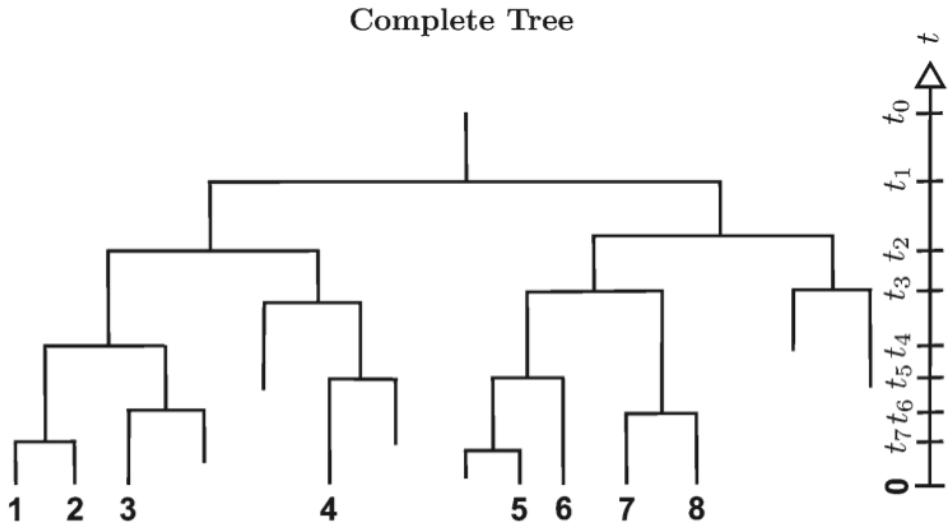
ψ Fossilization rate

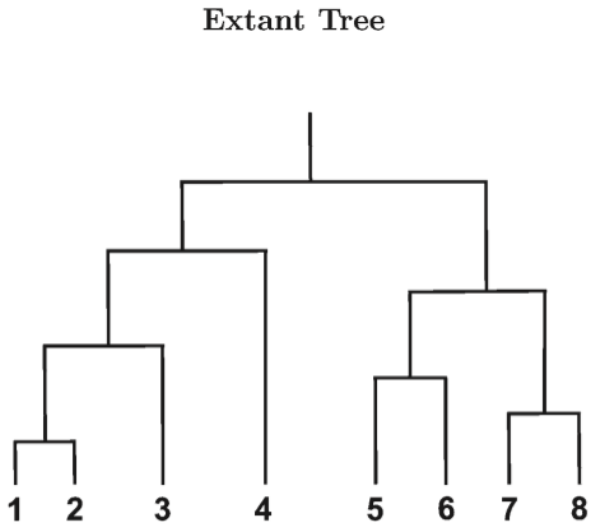
r Sampling probability

T Time of origin

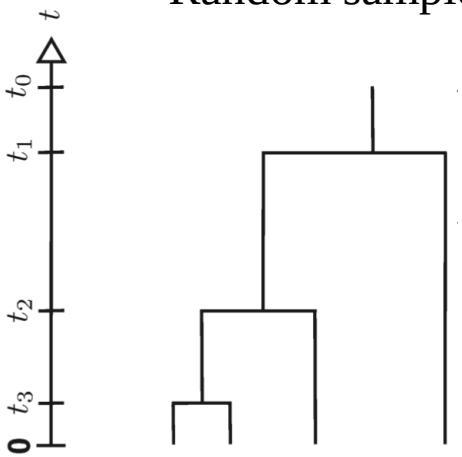


Sampling of extant taxa

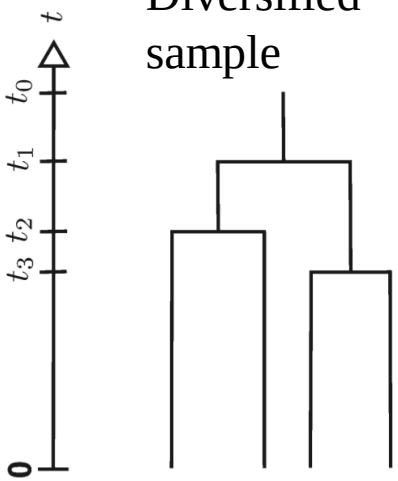




Random sample



Diversified sample

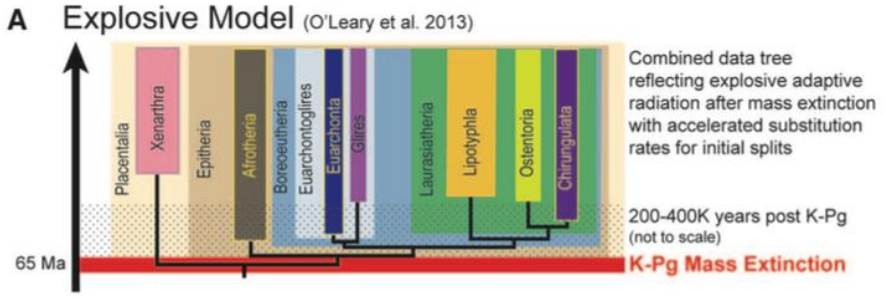


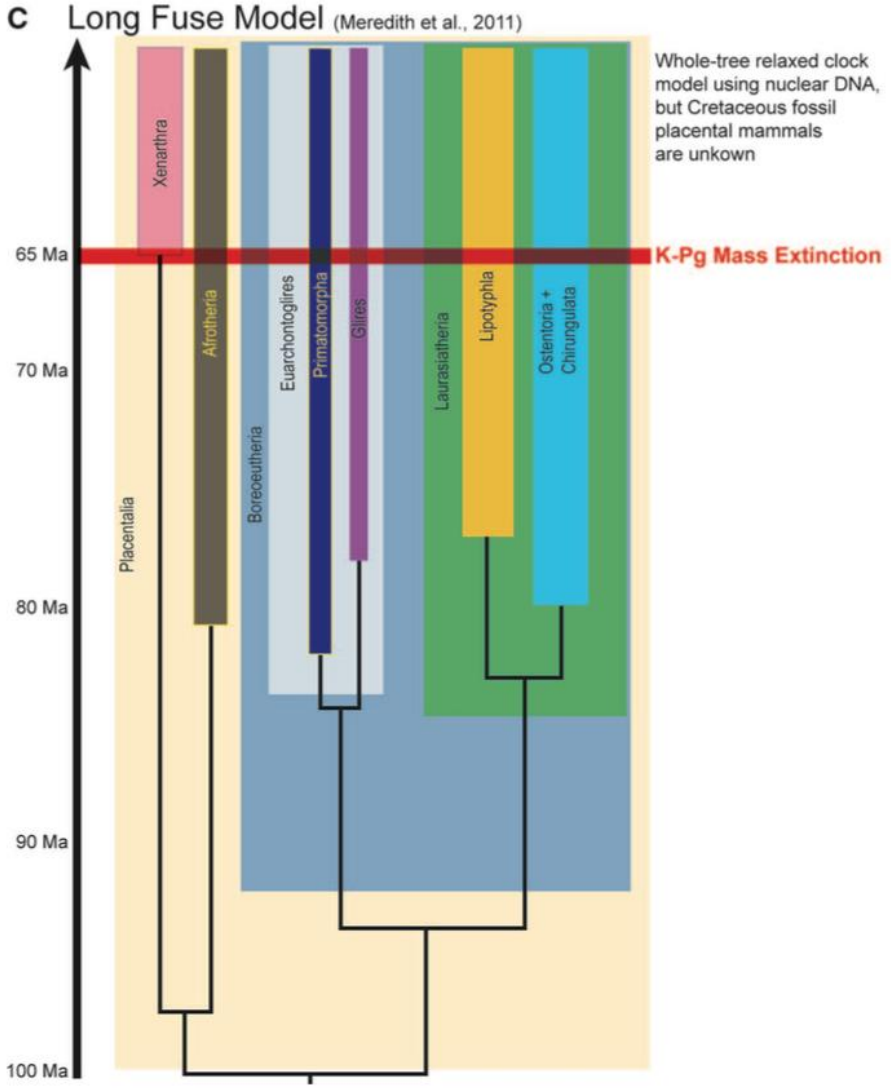
Random or
complete sampling

Diversified sampling

Placental radiation

- Controversial dating problem:
  - Bininda-Emonds et al. 2007, Nature (supertree analysis): 99 (93-108) Ma
  - Meredith et al. 2011, Science (calibrated molecular clock [supermatrix]): 101 (92, 117) Ma
  - dos Reis et al. 2012, PRS (genomics, multiple soft calibrations): (88, 92) Ma
  - O'Leary et al. 2013, Science (ghost lineage analysis): 65 Ma
  - Beck & Lee 2014, PRS (total-evidence dating, without internal node calibrations): 165 (150, 180) Ma
- Why does total-evidence dating widen and not close the gap between rocks and clocks?
  [see also review by O'Reilly et al. 2015, Trends in Genetics]



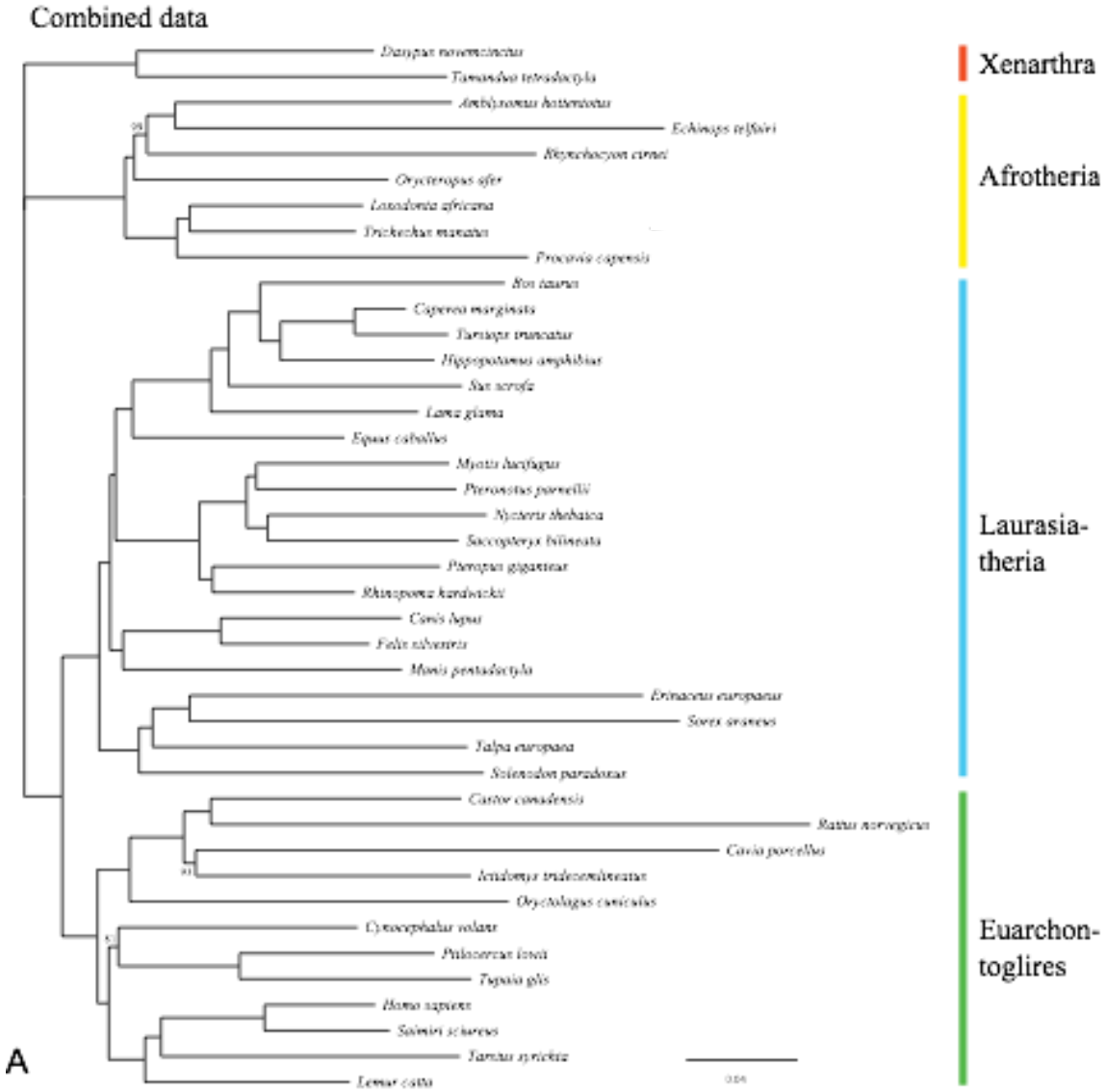


[O'Leary et al. 2013, Science, in reply to comment by Springer et al.]

The dataset

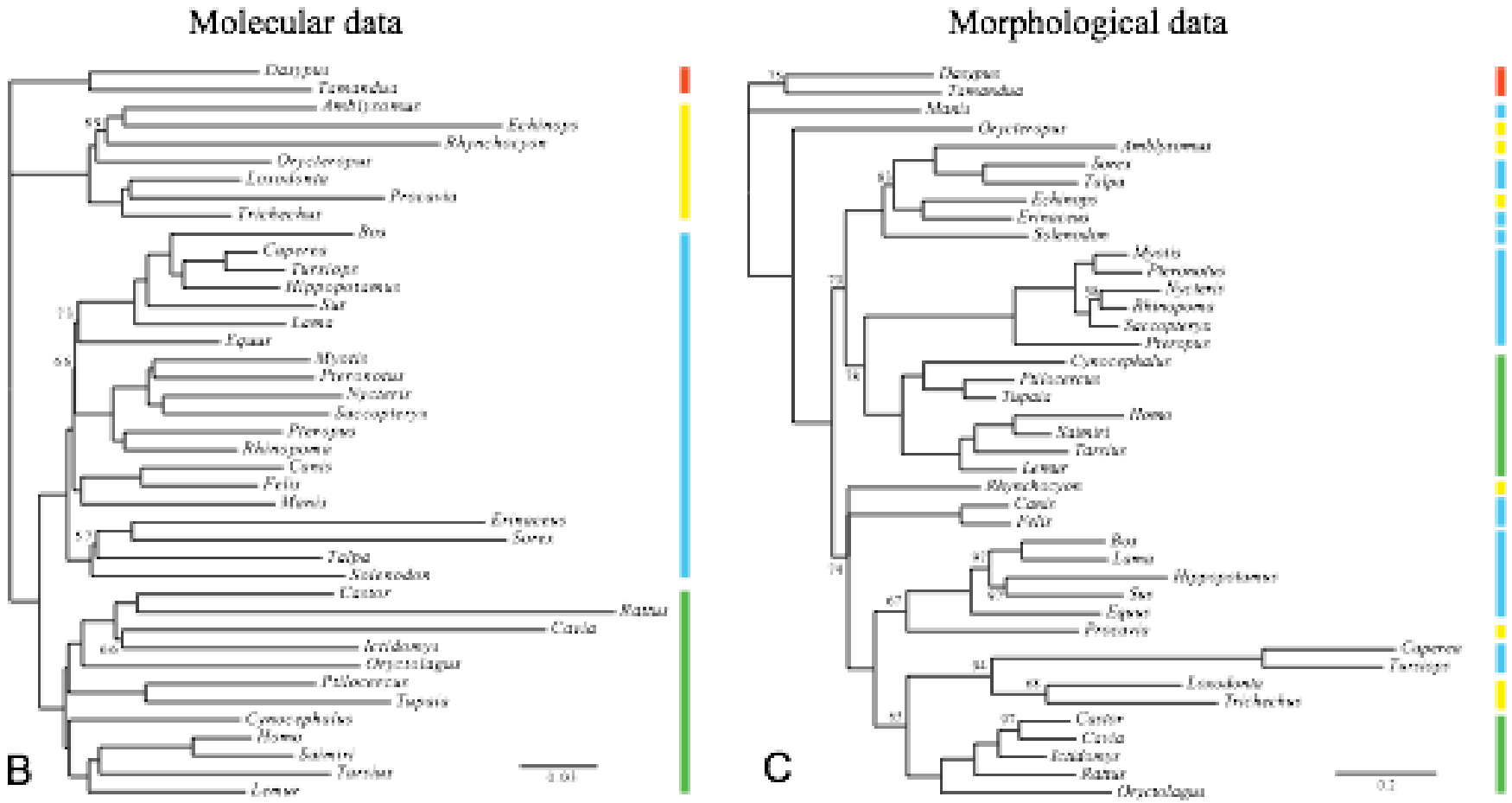
- From O'Leary et al. (2013), removing a recent fossil, the most recent speciation event, and non-eutherian taxa (from 86 to 74 taxa)
- Unprecedented morphological (phenomics) dataset: 4.5 k characters (1,284 cranial, 1,451 dental, 925 postcranial, 881 soft)
- Rich molecular dataset: 36.9 kb from 38 nuclear protein-coding genes
- 33 fossils and 41 recent taxa
- Total-evidence dating under different (fossilized birth-death) models to explore the reasons for discordance between estimated dates and the fossil record
- Vague priors on tree age and clock rate; independent gamma rates (white noise) relaxed clock model
- No internal node calibrations

[Non-clock analysis]

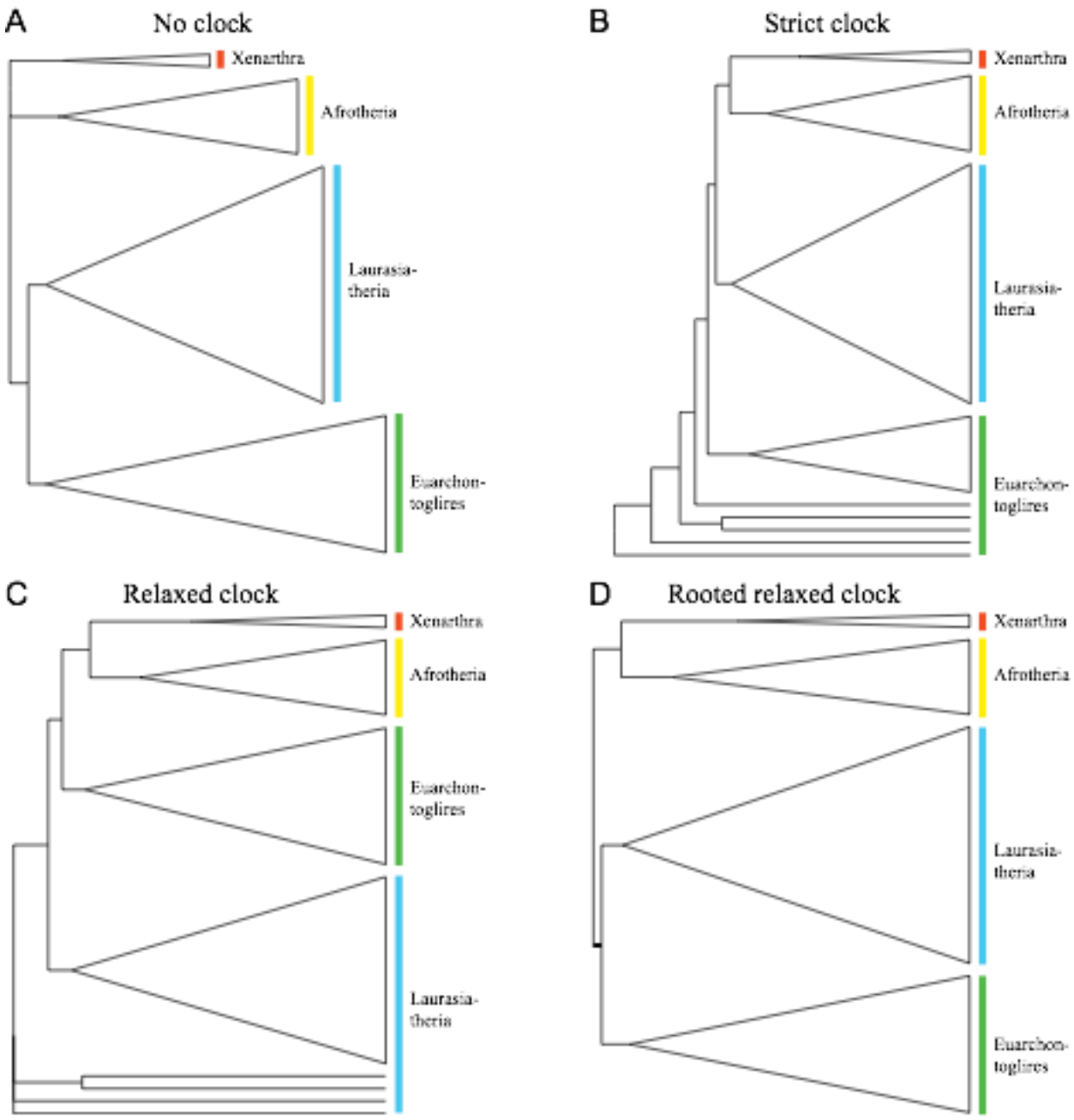


Combined tree retrieves expected relationships

[Non-clock analysis]

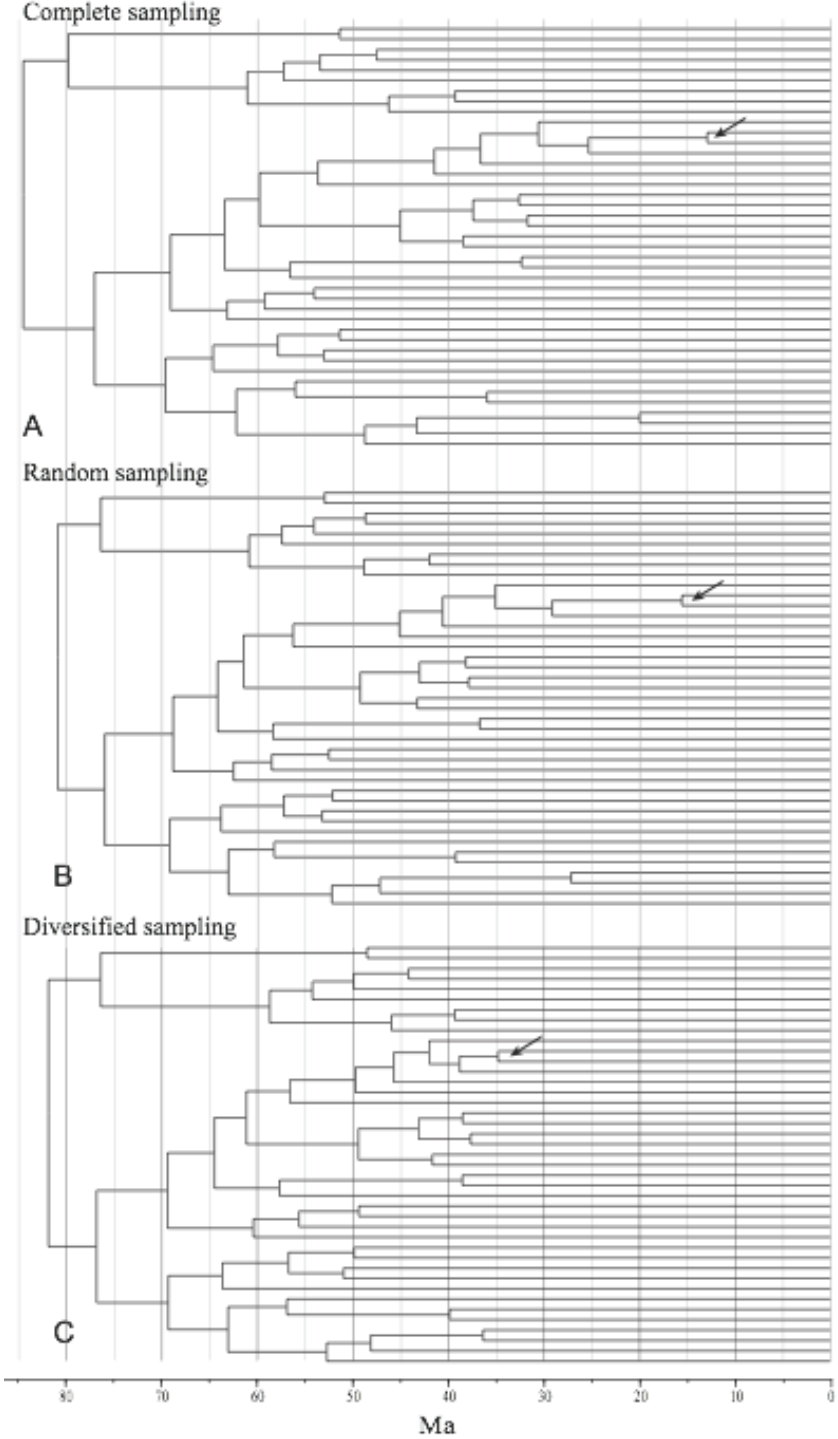


Combined tree largely reflects molecular data; morphological tree retrieves conflicting relationships but signal is weak



Relaxed clock analyses have difficulties finding the root

[Birth-death model without fossils]



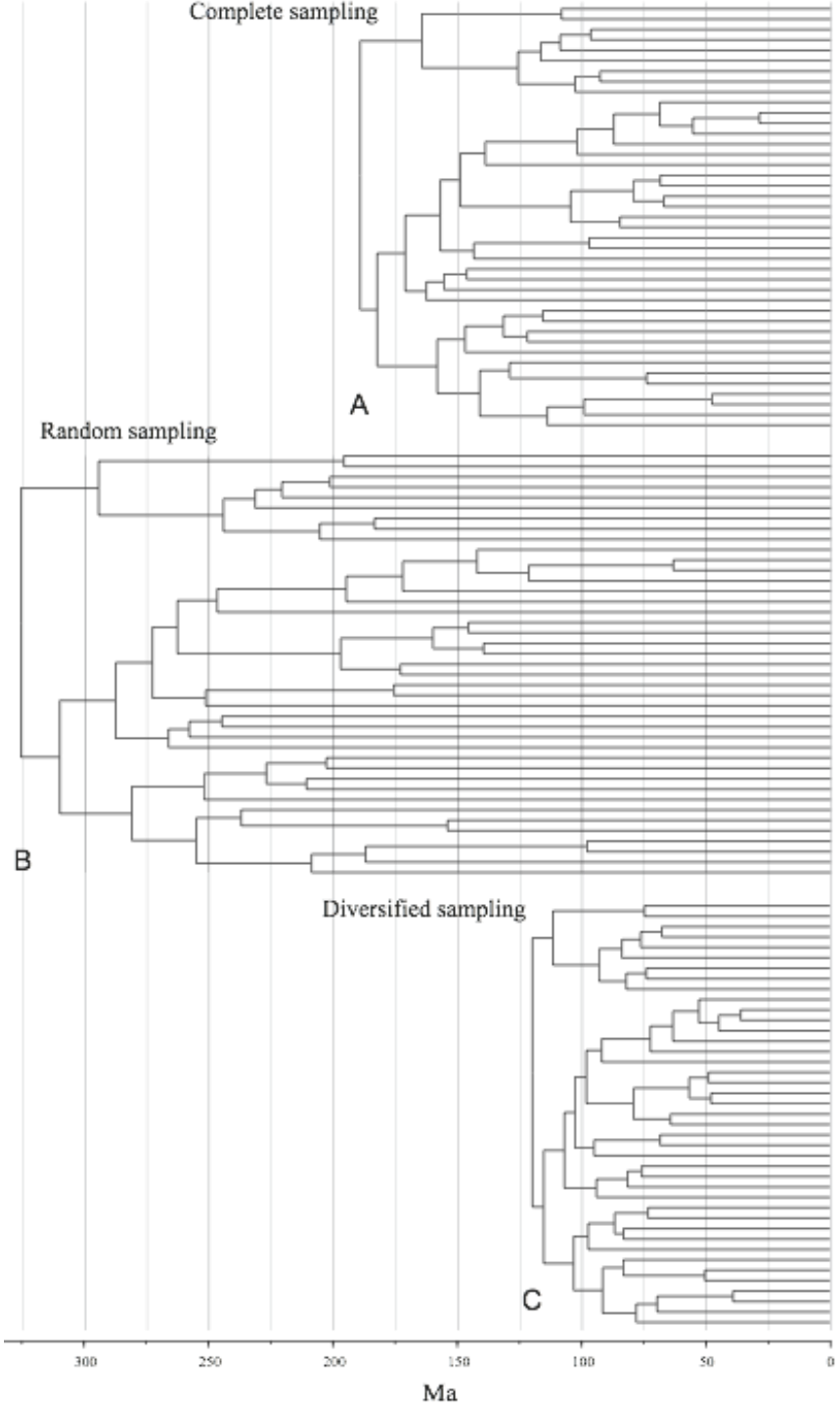
whale–dolphin split 13 Ma

whale–dolphin split 16 Ma

whale–dolphin split 35 Ma

Accounting for the tip sampling procedure is important in birth-death (speciation-extinction) models

[Fossilized birth-death model with fossils]



Under total-evidence dating, erroneous tip sampling assumptions have dramatic effects

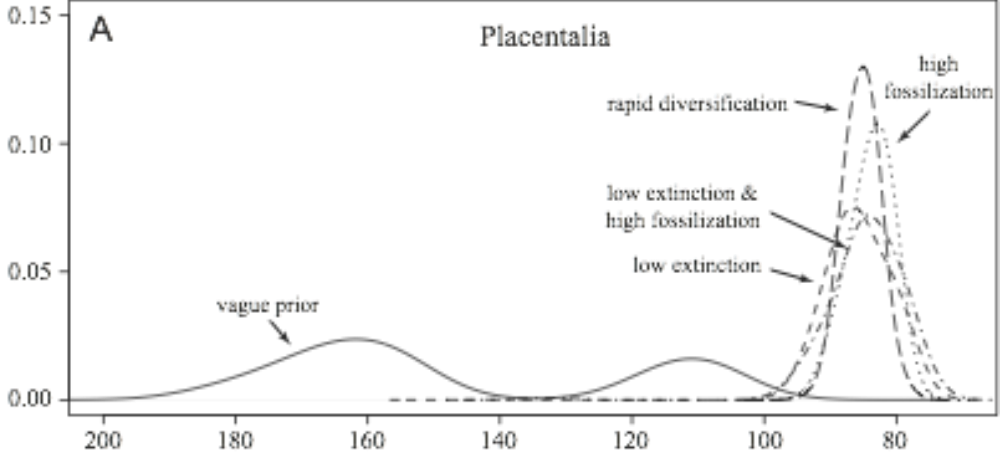
We call this phenomenon Deep Root Attraction (DRA)

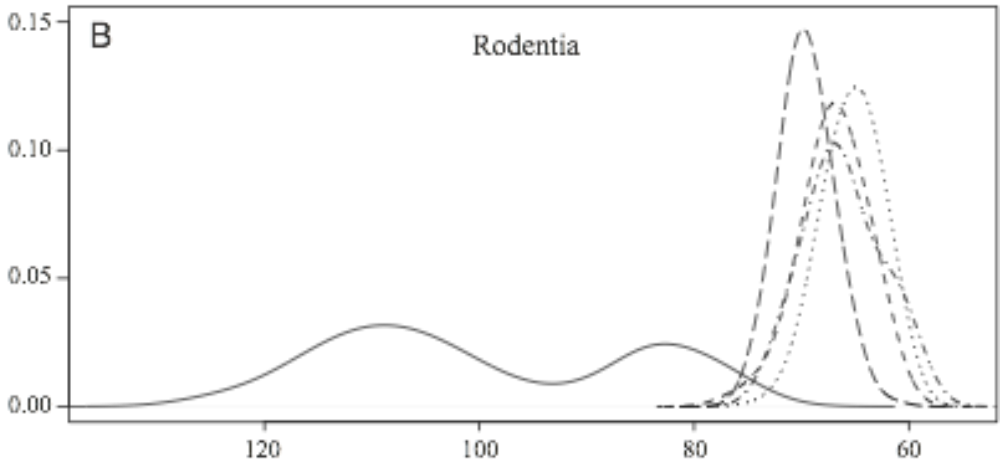
Deep root attraction (DRA)

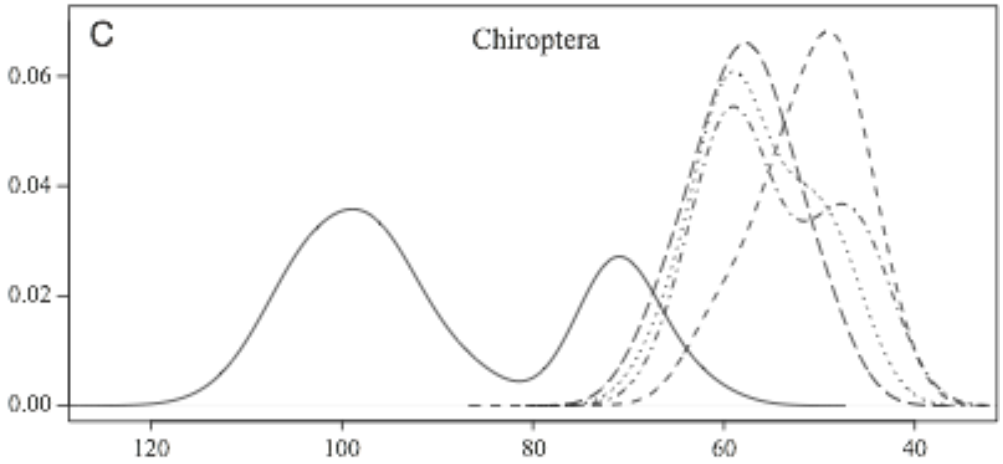
- Occurs under vague priors or erroneous models
- Occurs when long ghost lineages that are unobserved in the fossil record carry little cost
- Occurs when there is low net diversification (speciation and extinction rates are approximately balanced, so that we expect many lineages in the past)
- Occurs when there is a high extinction rate (high turnover) and a low fossil sampling probability
- Occurs when background information allowing us to conclude that long unobserved ghost lineages are unlikely is not accounted for in the analysis (e.g., very few of the available fossils included in the analysis)
- Aggravated by model inadequacies and conflicts between data partitions

Addressing DRA

- Using informative priors assuming:
  - … a high diversification rate
  - … a low extinction rate
  - … a high fossil sampling probability
  - … a combination of low extinction rate and high fossil sampling probability

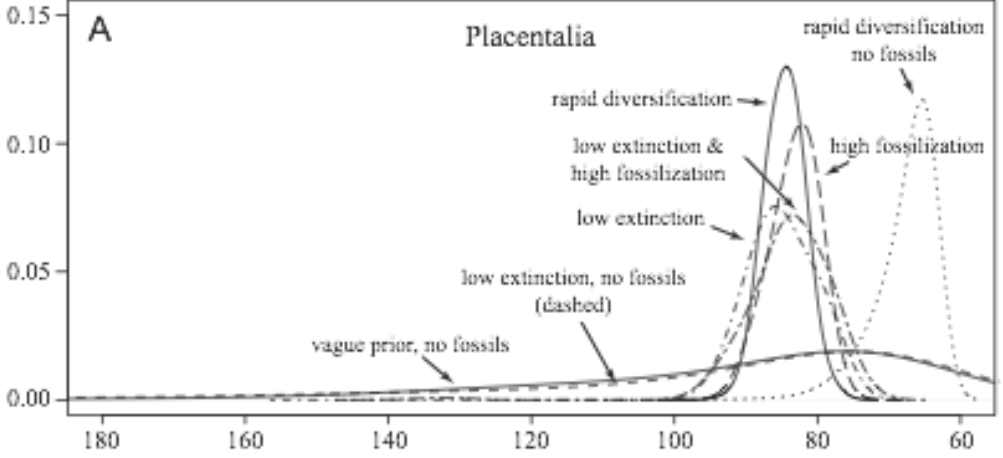


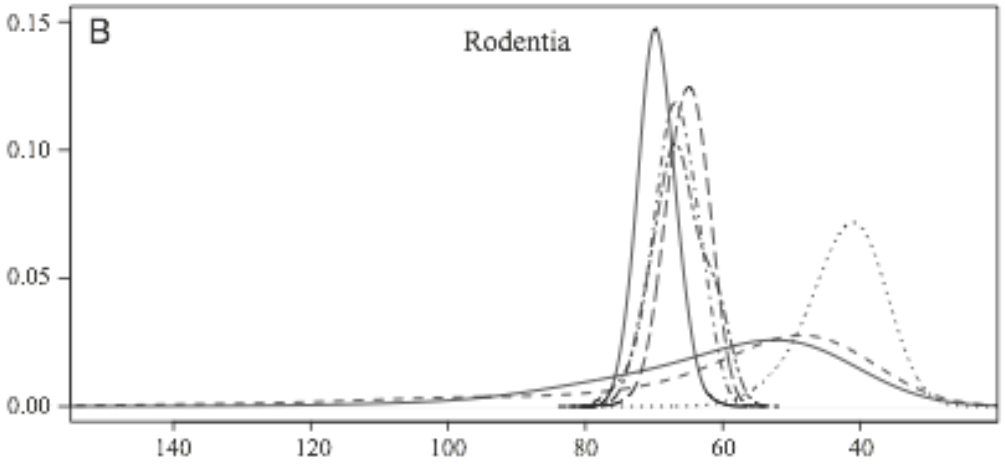


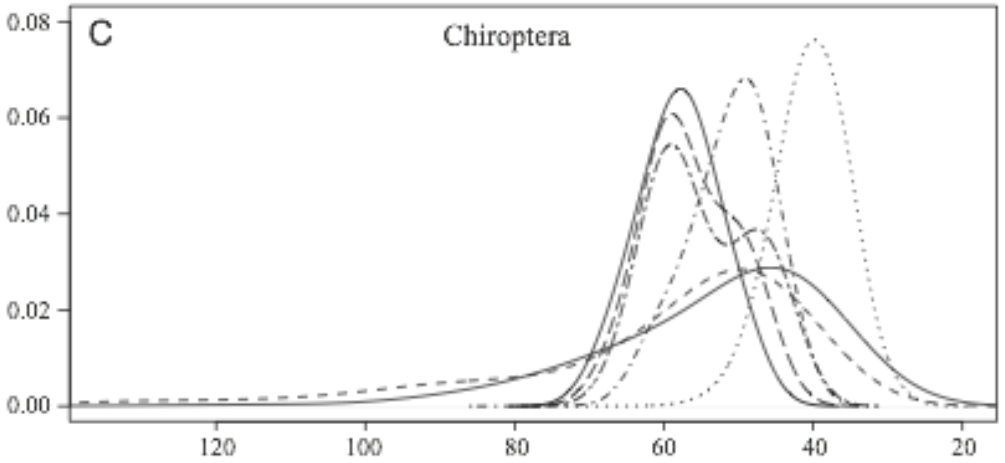


[Vague and informative priors, fossilized-birth death with fossils]

Introducing a modest penalty for ghost lineages corrects DRA and stabilizes divergence time estimates



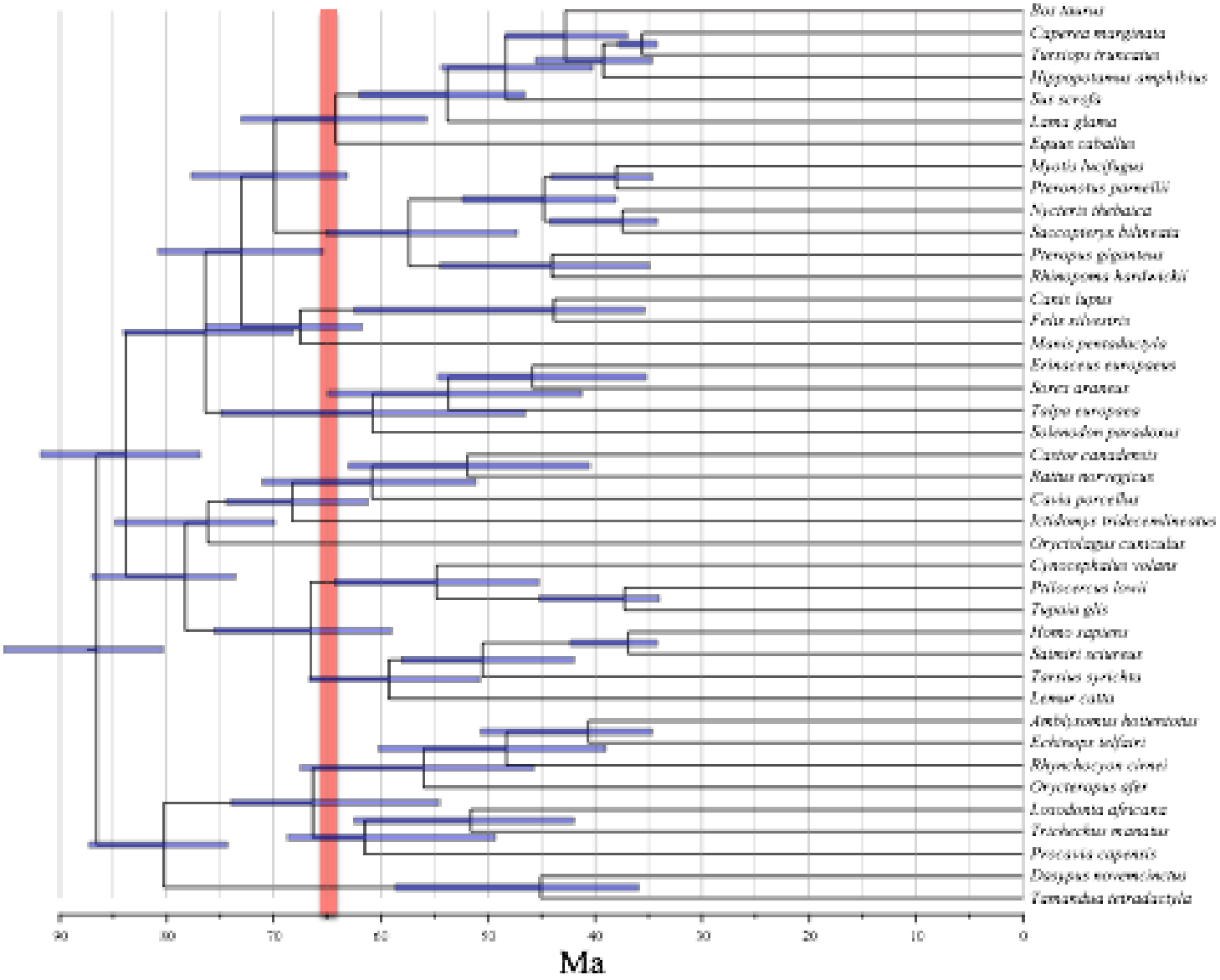




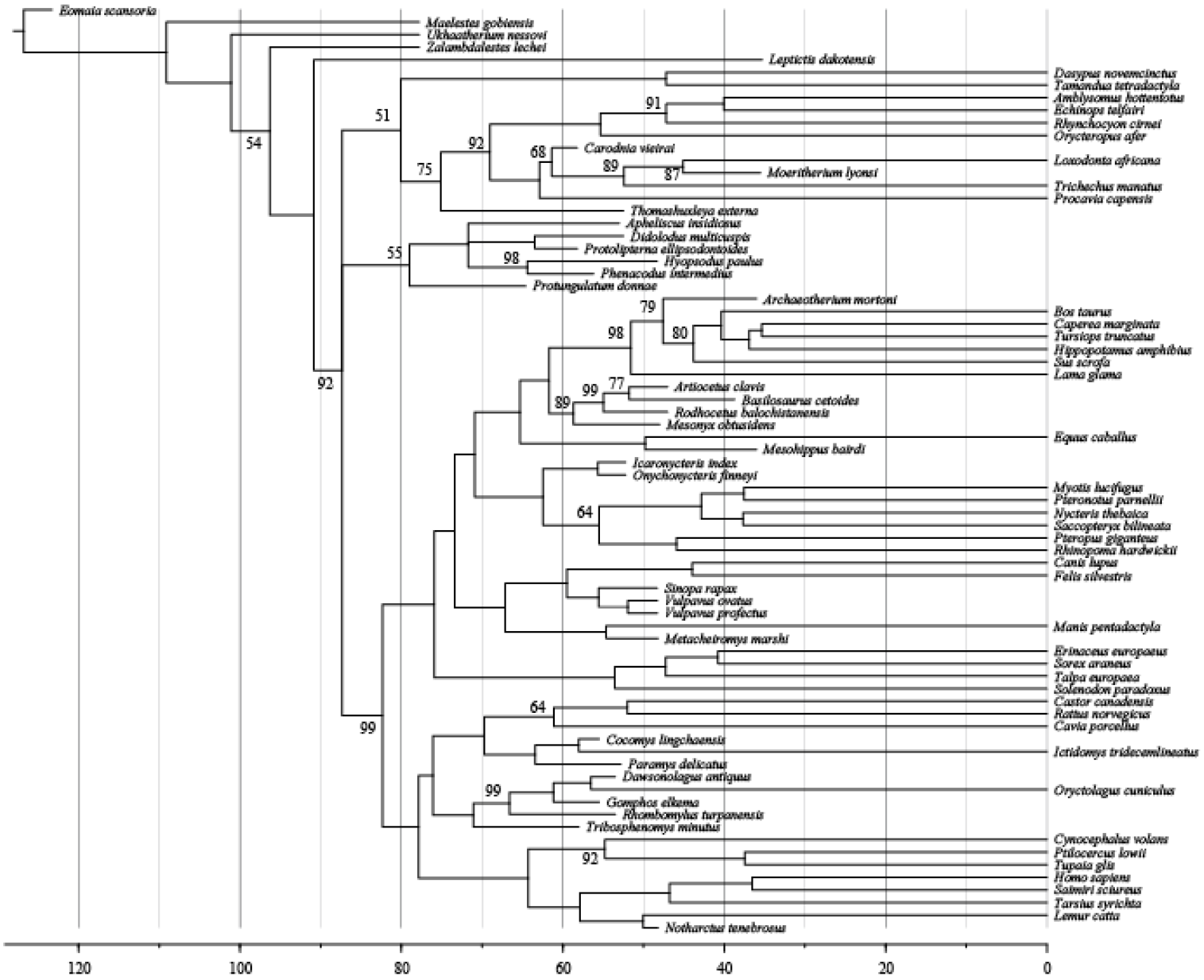
[Vague and informative priors, fossilized-birth death with fossils and birth-death without fossils]

Fossils stabilize divergence time estimates and increase the precision of those estimates

Total-evidence dating assuming rapid (initial) diversification



Total-evidence dating placement of fossils



Improving placental TE dating

- Modeling fossil preservation probabilities and biogeographically dependent sampling probabilities
- More sophisticated diversification models, e.g., skyline or logistic growth models
- Better understanding of rate variation across characters and across lineages in morphological (and molecular) characters
- Relaxing the assumption of coupled rate variation across lineages in molecular and morphological clocks
- Better understanding of morphological evolution
  - Directional evolution [Klopstein et al. Syst Biol 2015]
  - Modeling character dependencies to address convergence in large correlated character suites driven by functional adaptation