

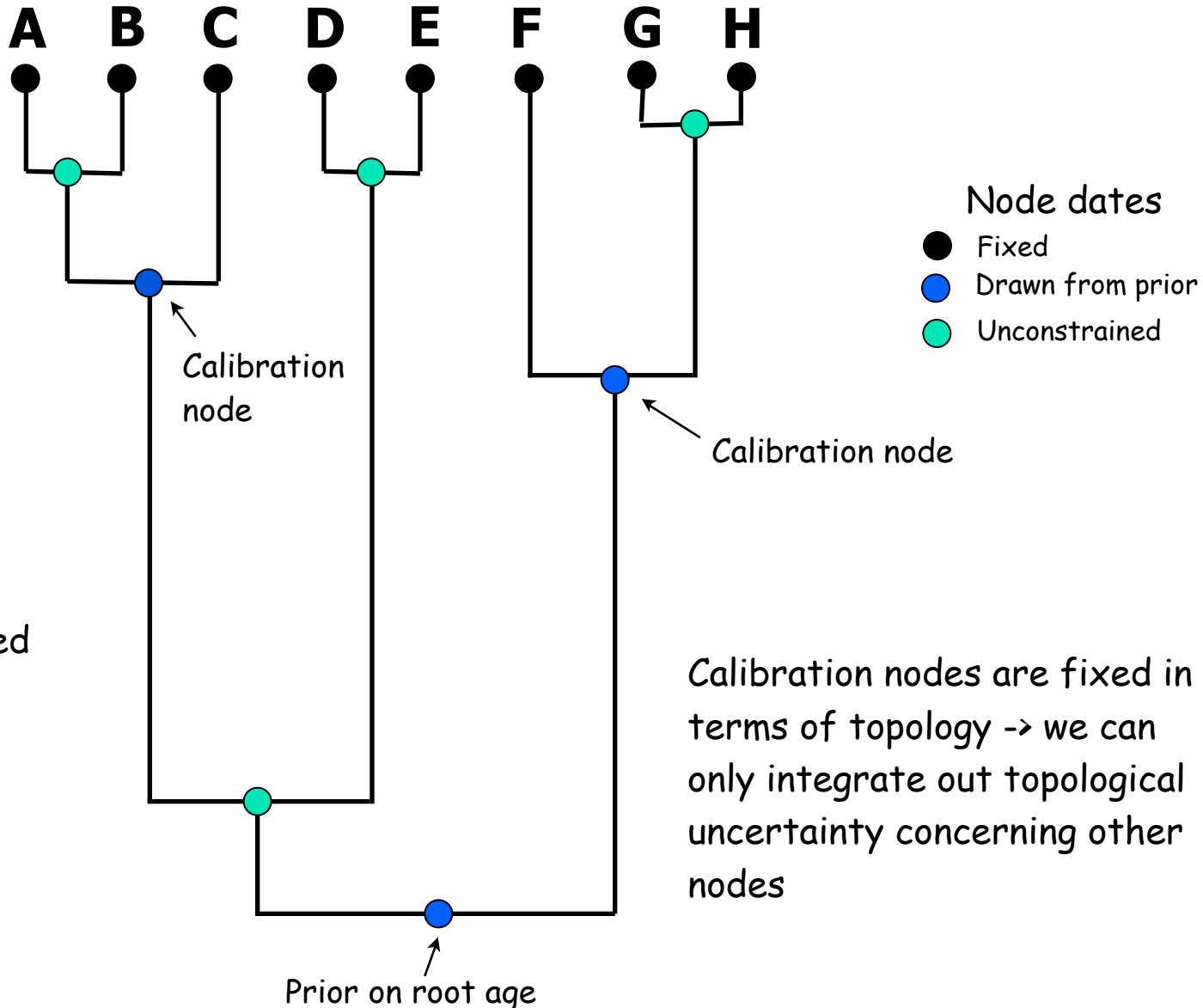
Dating Phylogenies with Fossils

Fredrik Ronquist

Dept. Bioinformatics & Genetics
Swedish Museum of Natural History



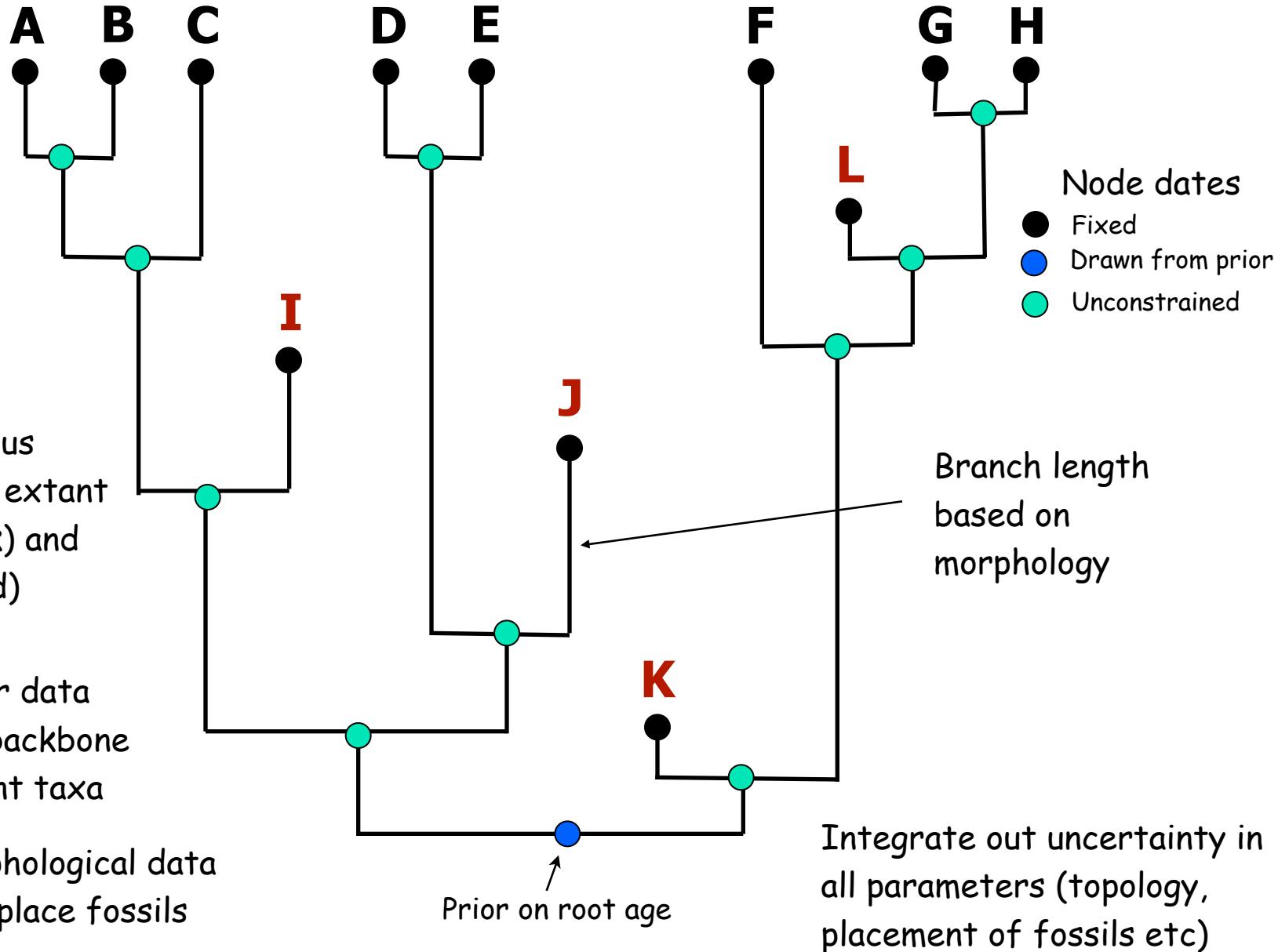
Node dating



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



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

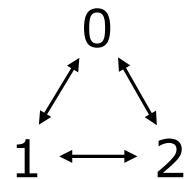


Morphological models

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

Transformation series

Unordered (Fitch)



Ordered (Wagner)

$0 \longleftrightarrow 1 \longleftrightarrow 2$

Probabilistic models

Unordered (M3u)

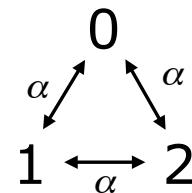
$$Q = \begin{pmatrix} -2\alpha & \alpha & \alpha \\ \alpha & -2\alpha & \alpha \\ \alpha & \alpha & -2\alpha \end{pmatrix}$$

Ordered (M3o)

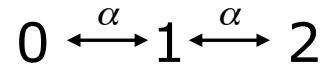
$$Q = \begin{pmatrix} -\alpha & \alpha & 0 \\ \alpha & -2\alpha & \alpha \\ 0 & \alpha & -\alpha \end{pmatrix}$$

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
0	0	0	1
0	0	0	0

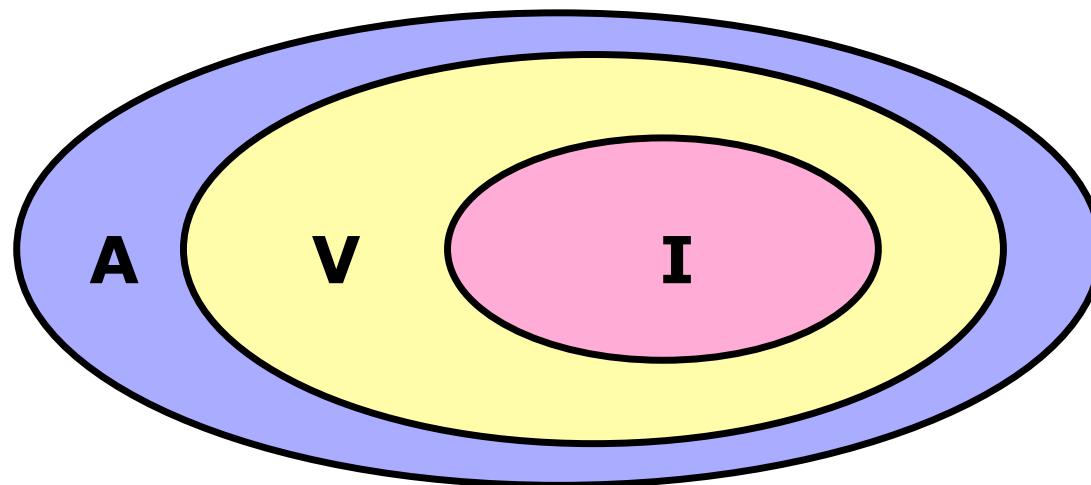
Variable

Incomplete coding

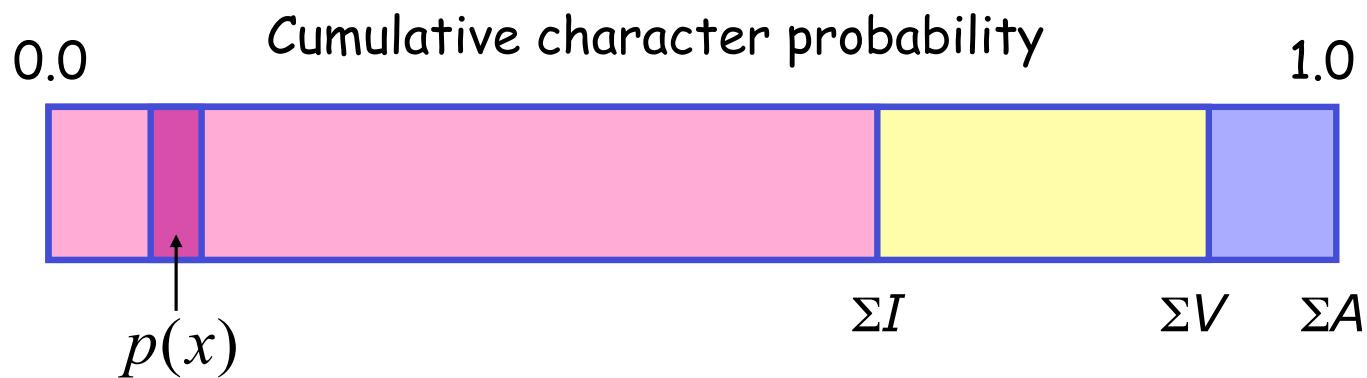
A	B	C	D	
0	0	1	1	Informative
0	0	0	1	
0	0	0	0	

Types of characters

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



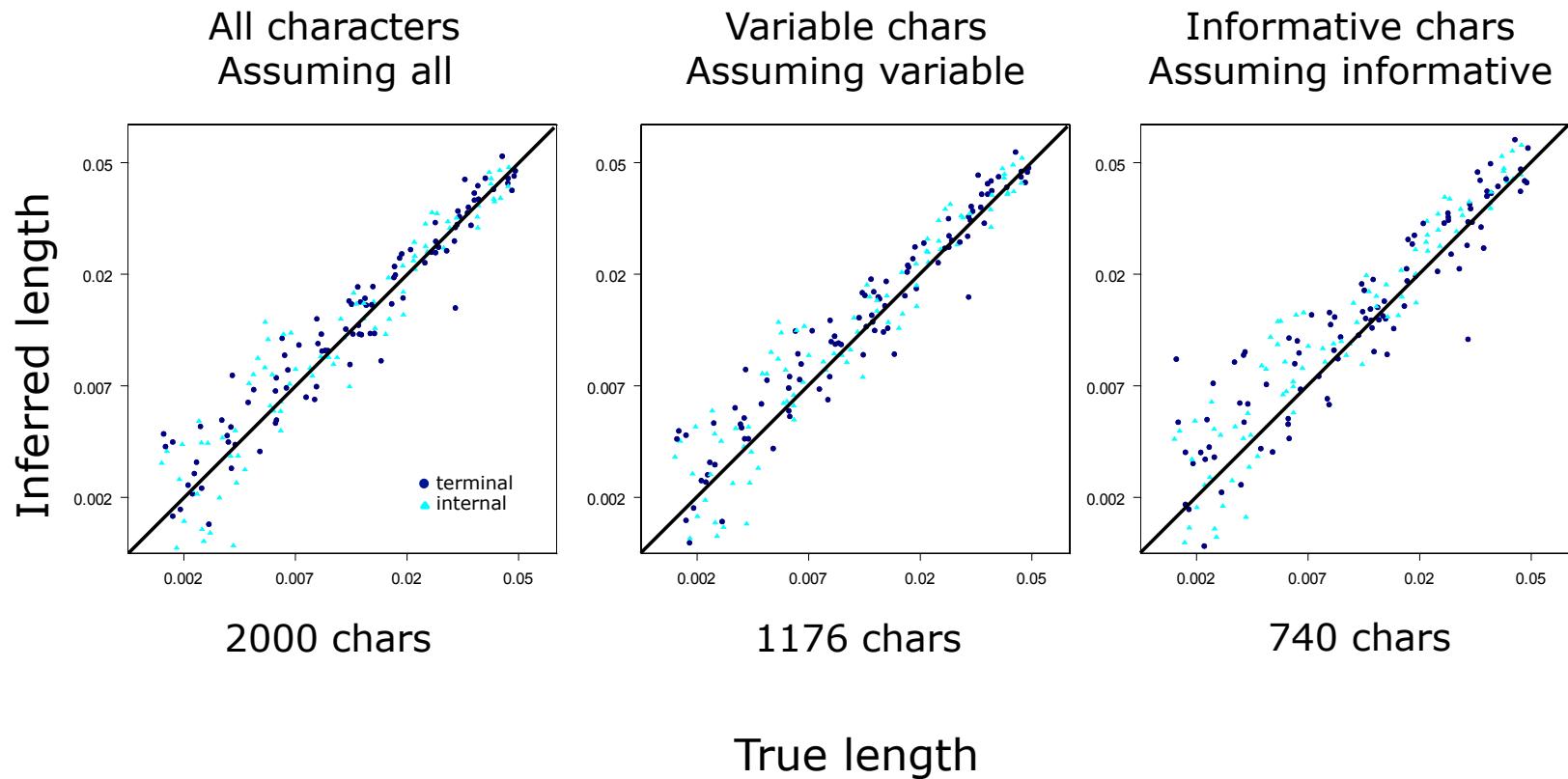
Conditional character probability



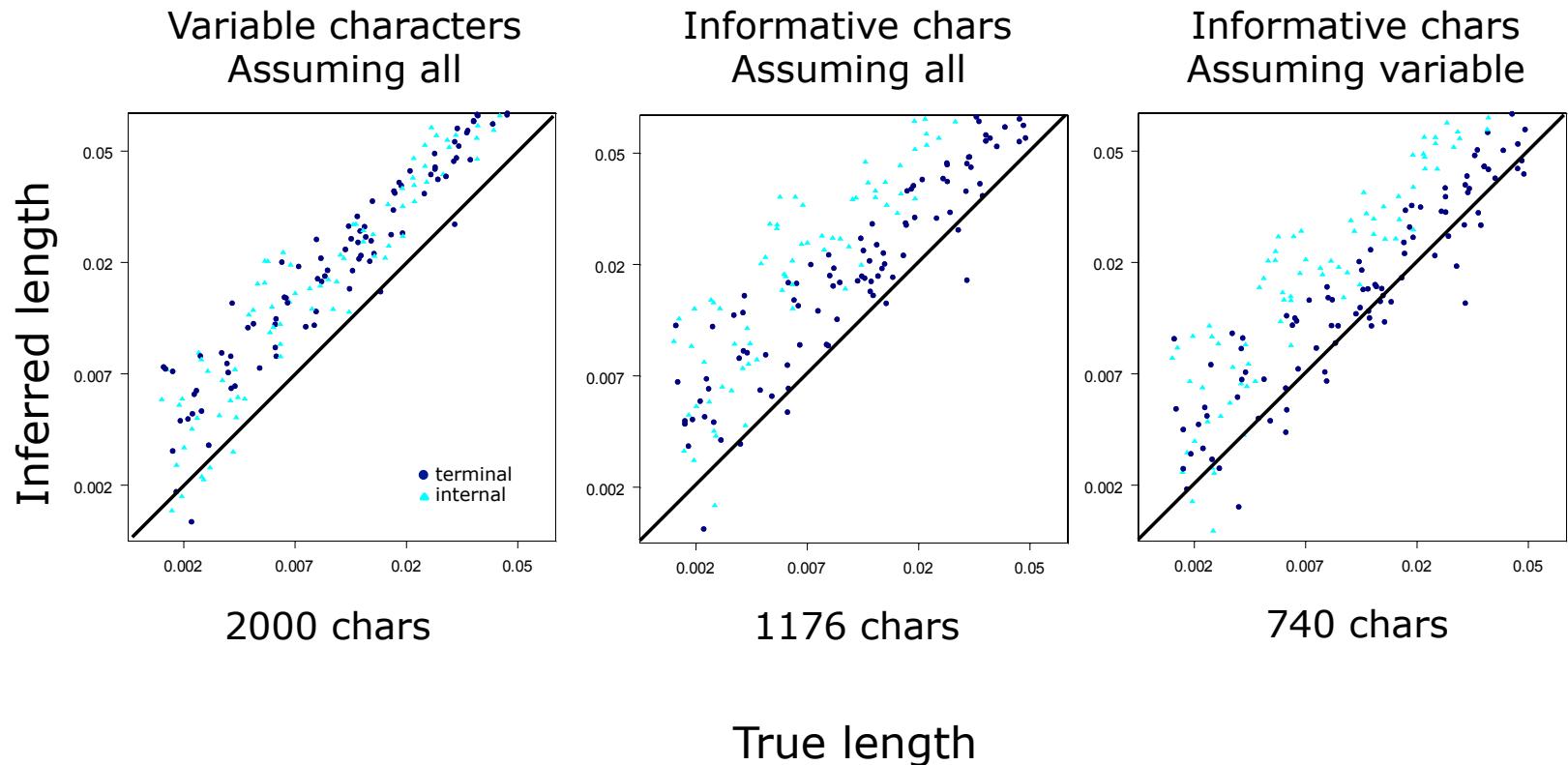
Conditional probability of one character x_i given that only informative characters are coded:

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

Branch length estimates



Branch length estimates



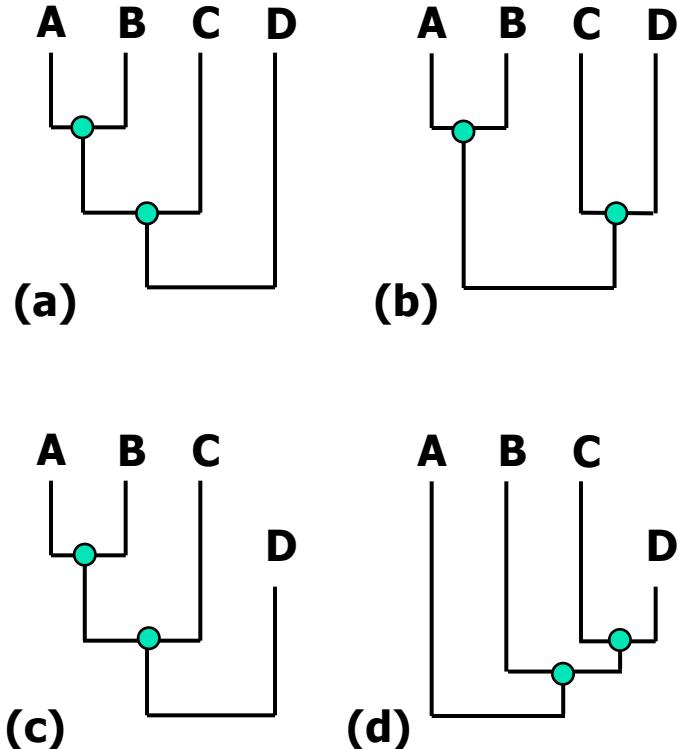
Relaxed clock models

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

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

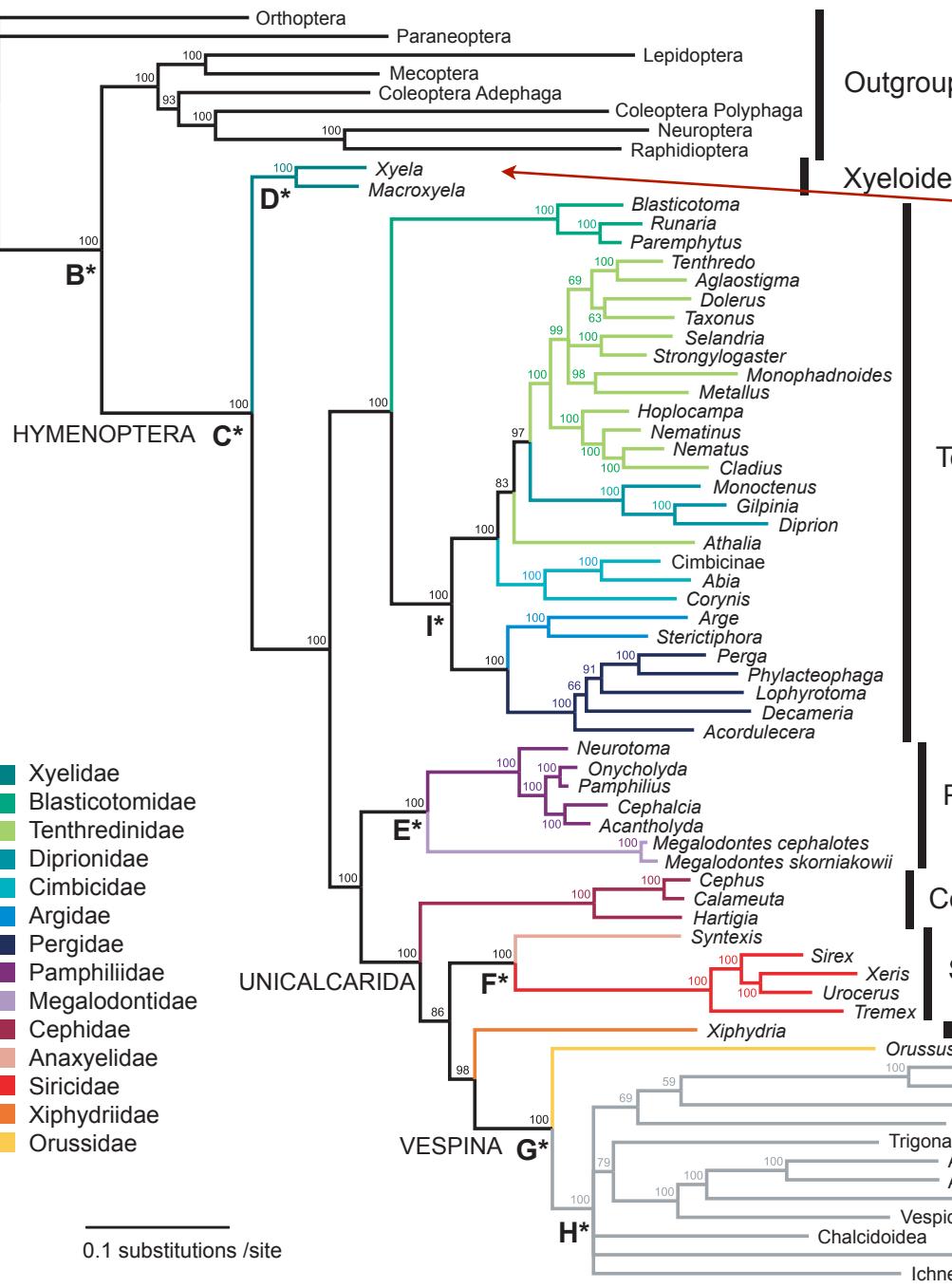
Uniform prior on serially sampled trees



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

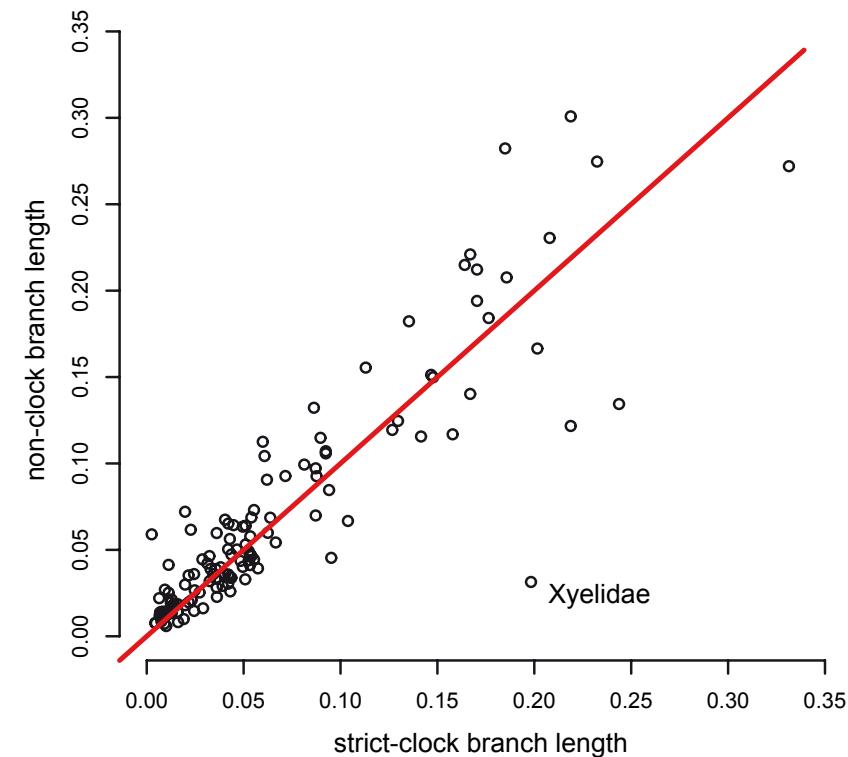


Rate deceleration in Xyelidae!

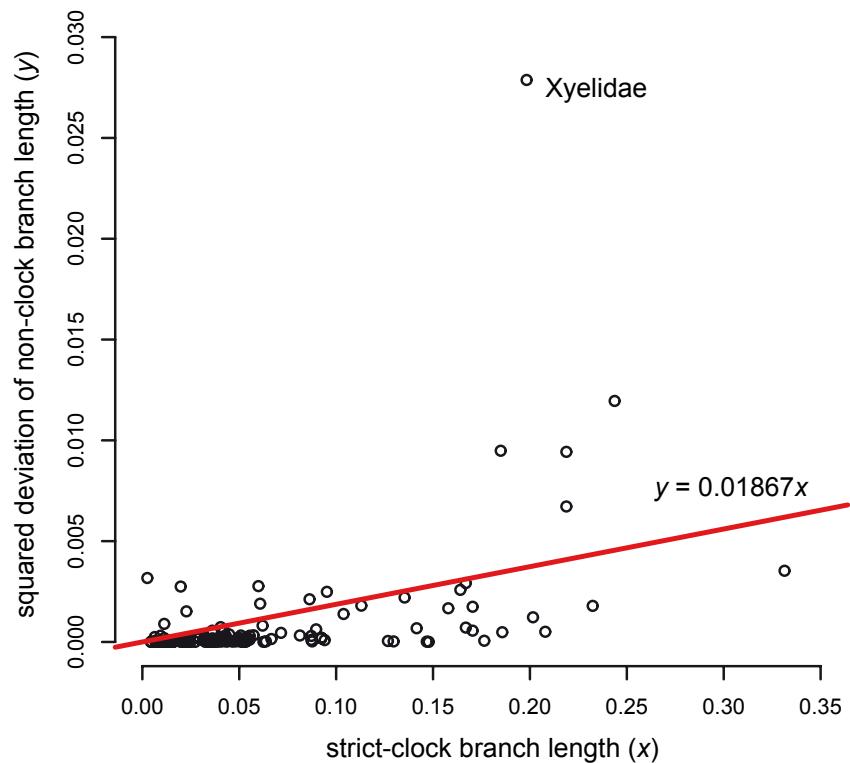
A - I were used in clock analyses as calibration points (all clades well supported)

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

Correlation

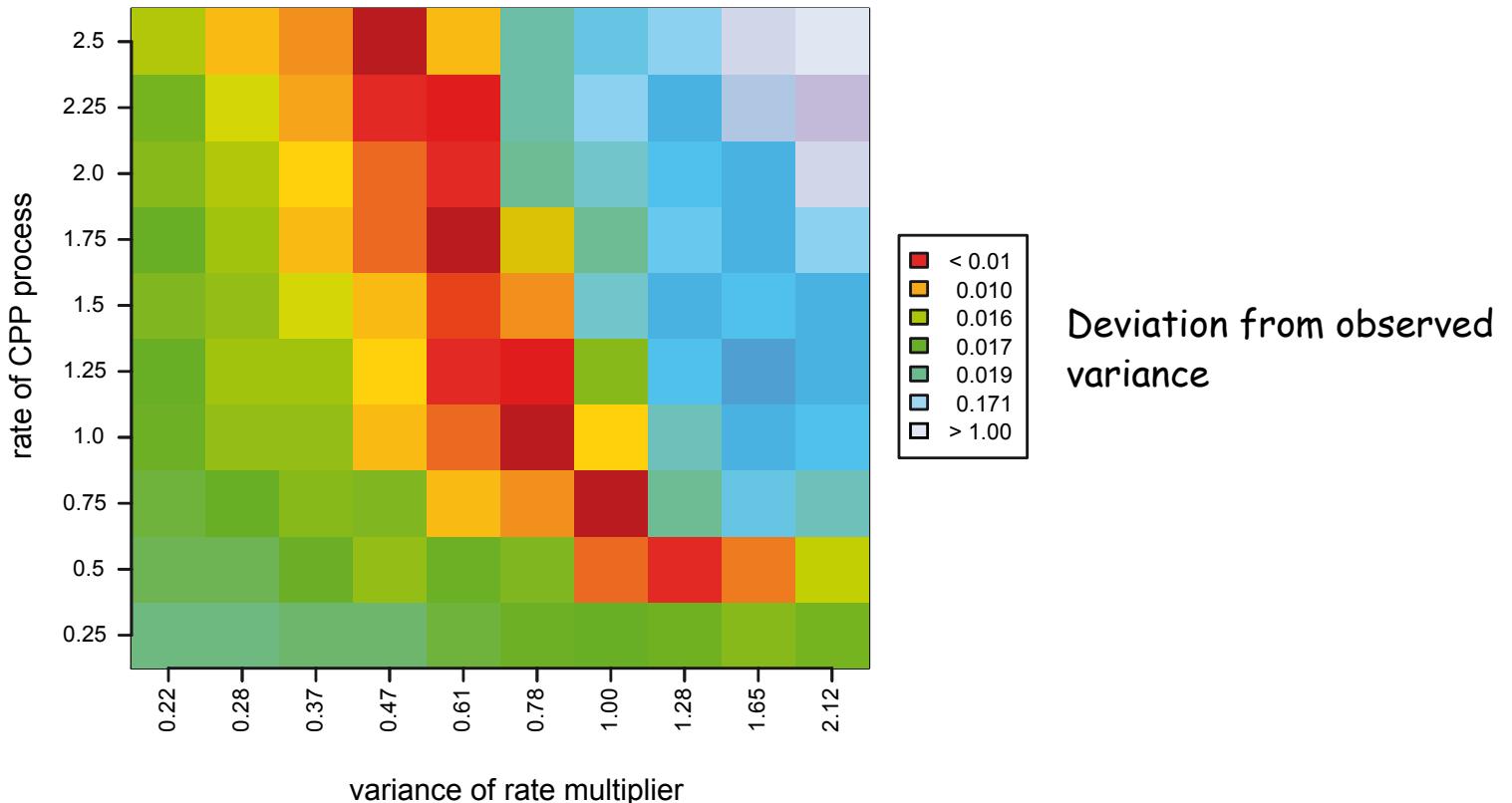


Squared deviation

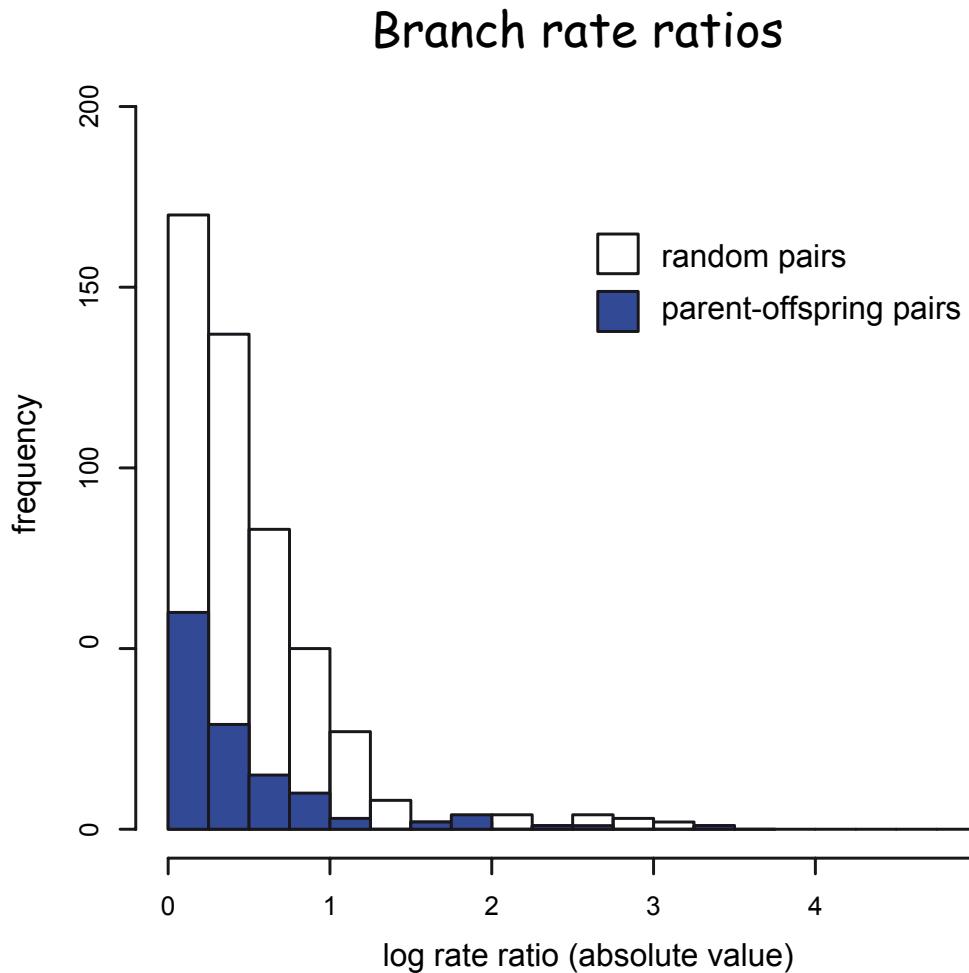


Finding suitable priors for the CPP model

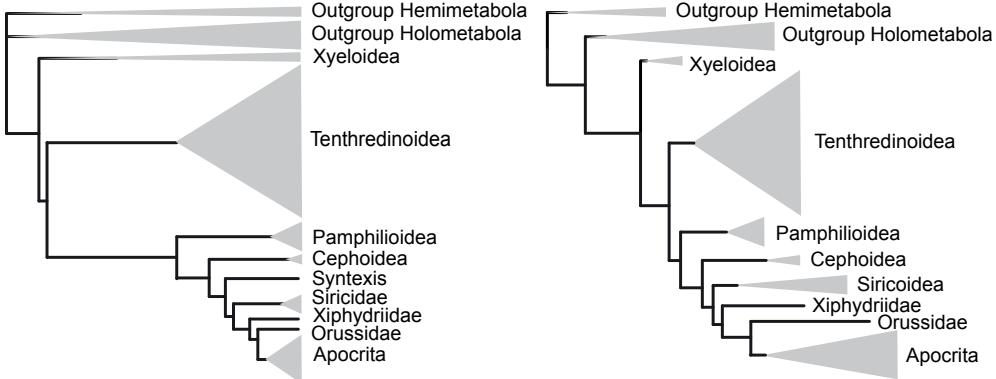
Poisson rate - multiplier variance space



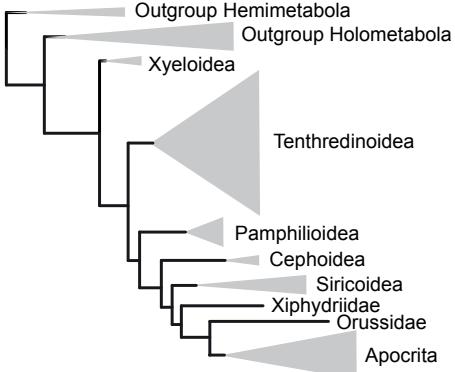
Slight but significant rate autocorrelation



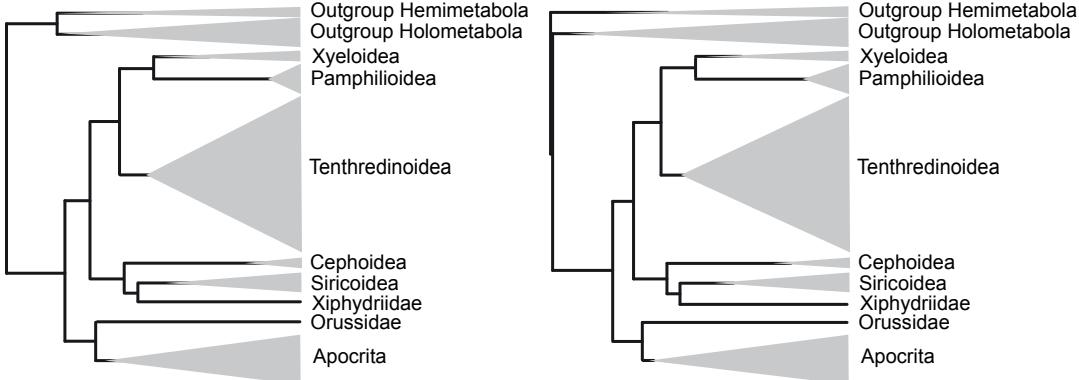
Morphology



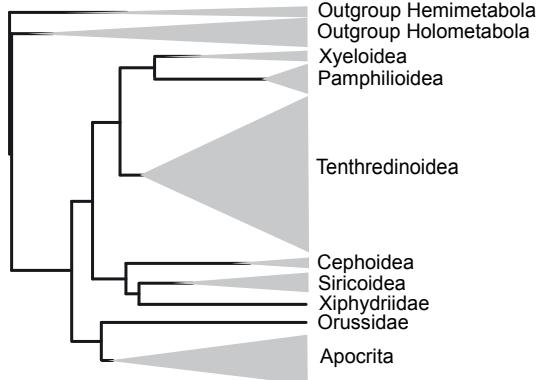
Non-clock



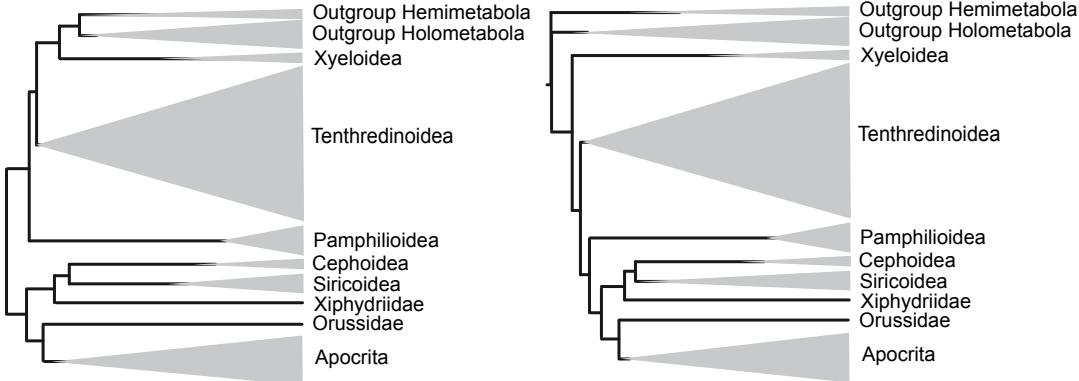
Strict clock



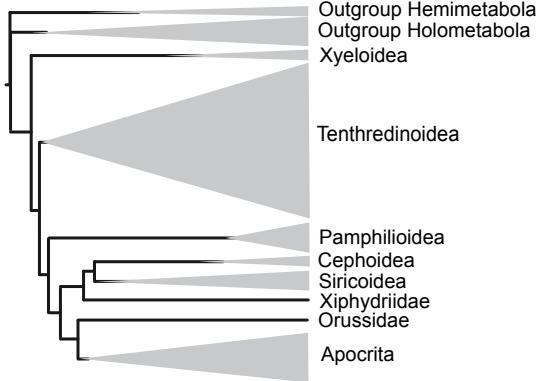
Strict clock with rooting constraint



Relaxed clock



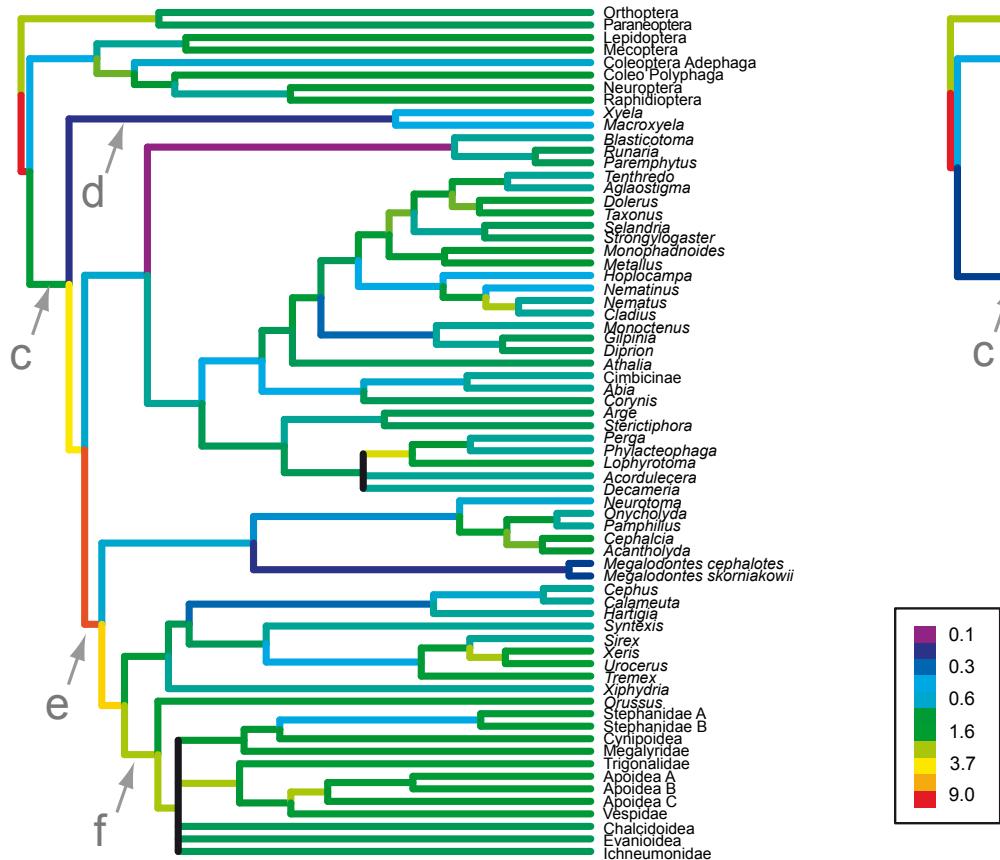
Relaxed clock with rooting constraint



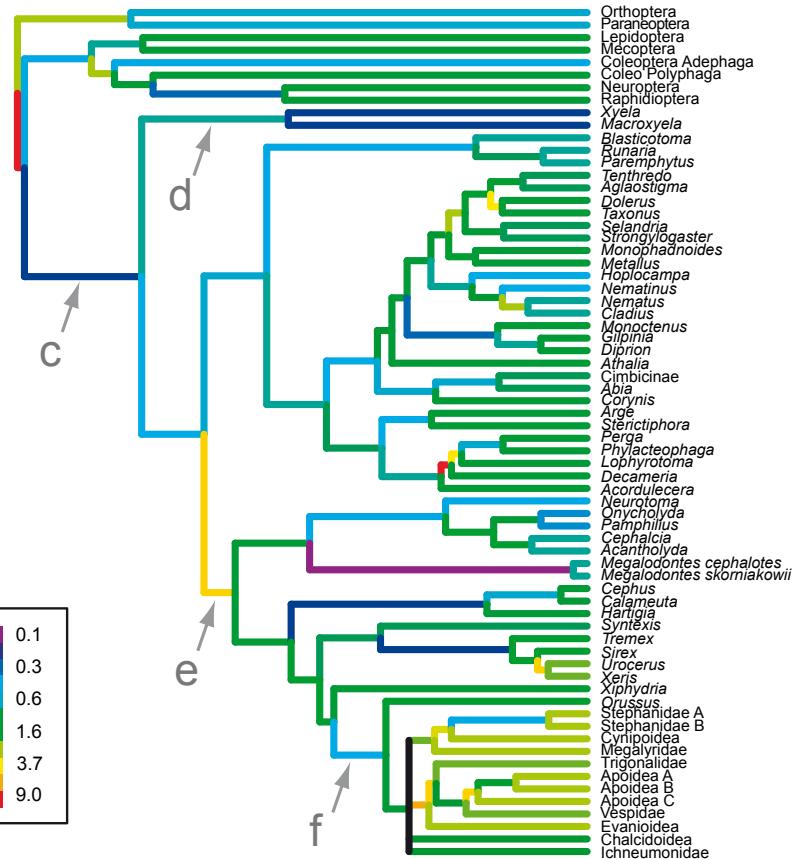
Relaxed clock
models may need
rooting constraint

Rate variation across the tree

IGR model

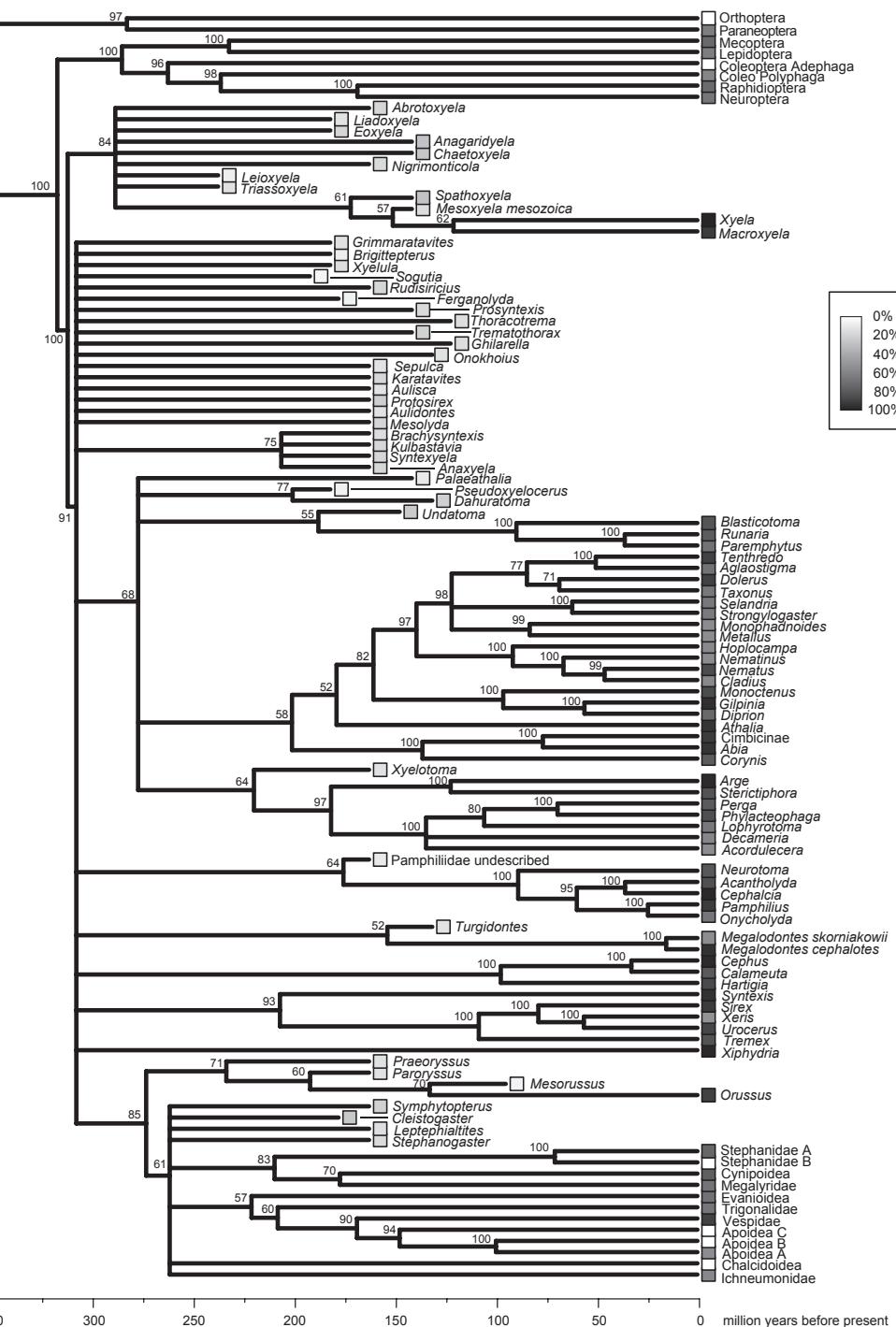


CPP model



Majority rule consensus with fossils

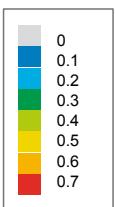
Completeness of
morphology scores



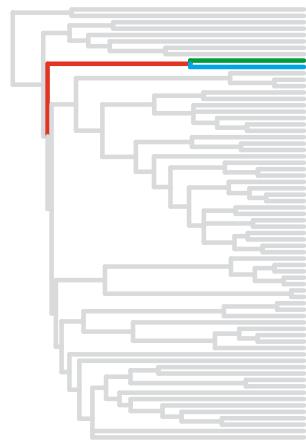
Uncertainty in the phylogenetic position varies across fossils



Mesoxyela mesozoica



Sogutia liassica



Outgroups

Xyelidae

Tenthredinoidea

Pamphilioidea

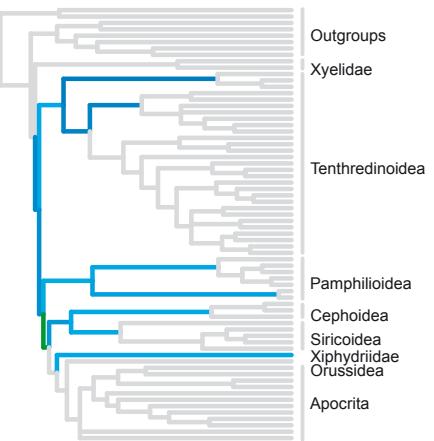
Cephoidea

Siricoidea

Xiphydriidae

Orussidae

Apocrita



Outgroups

Xyelidae

Tenthredinoidea

Pamphilioidea

Cephoidea

Siricoidea

Xiphydriidae

Orussidae

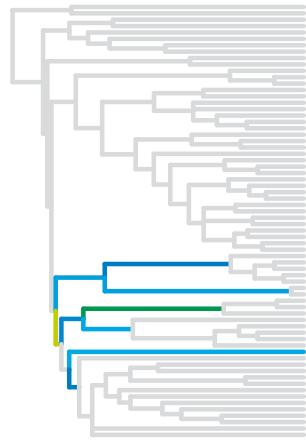
Apocrita



Aulisca odontura



Leptephialtites caudatus



Outgroups

Xyelidae

Tenthredinoidea

Pamphilioidea

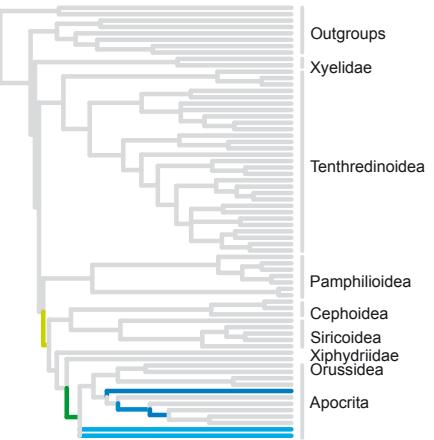
Cephoidea

Siricoidea

Xiphydriidae

Orussidae

Apocrita



Outgroups

Xyelidae

Tenthredinoidea

Pamphilioidea

Cephoidea

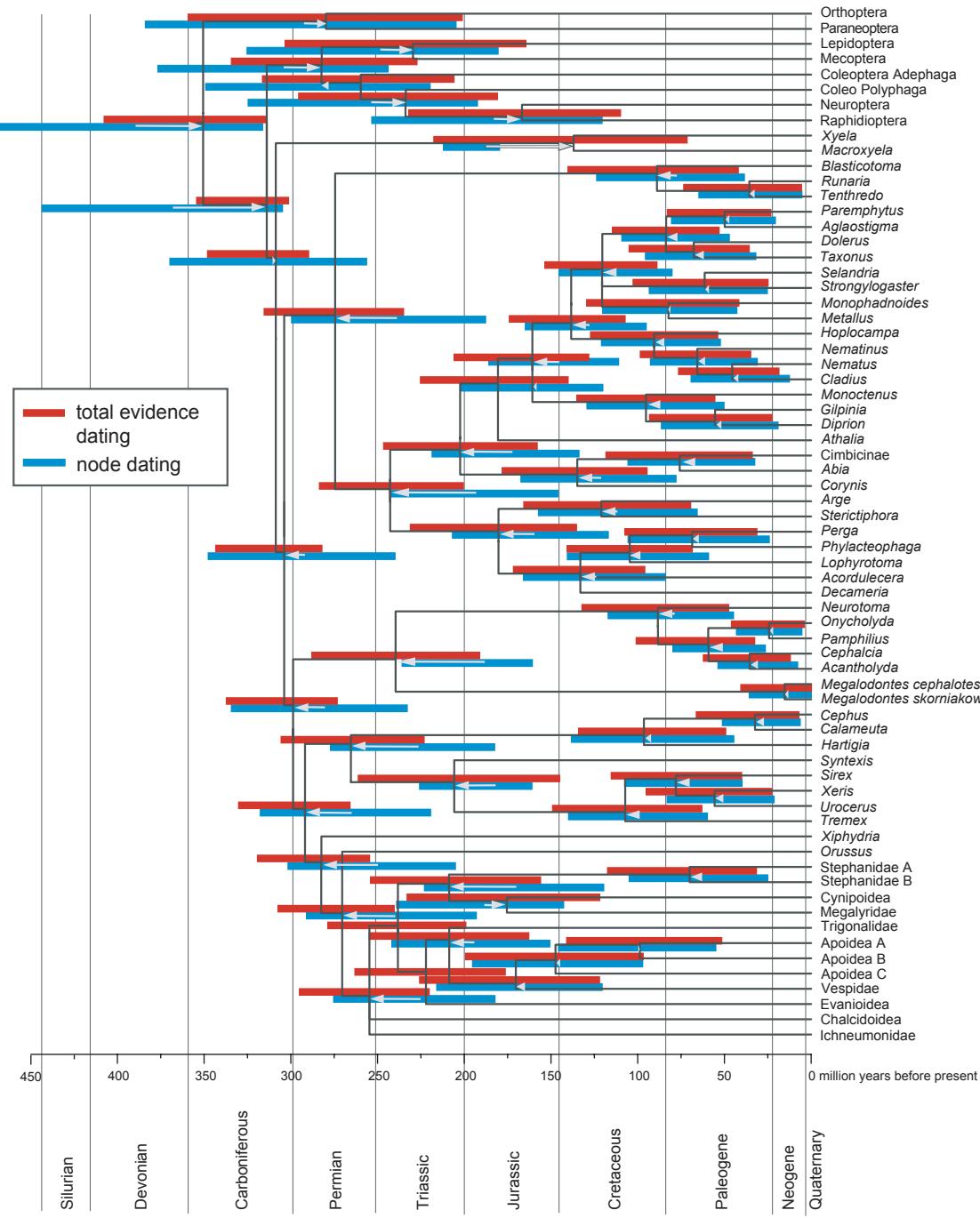
Siricoidea

Xiphydriidae

Orussidae

Apocrita

Estimated divergence times



Posteriors on node ages

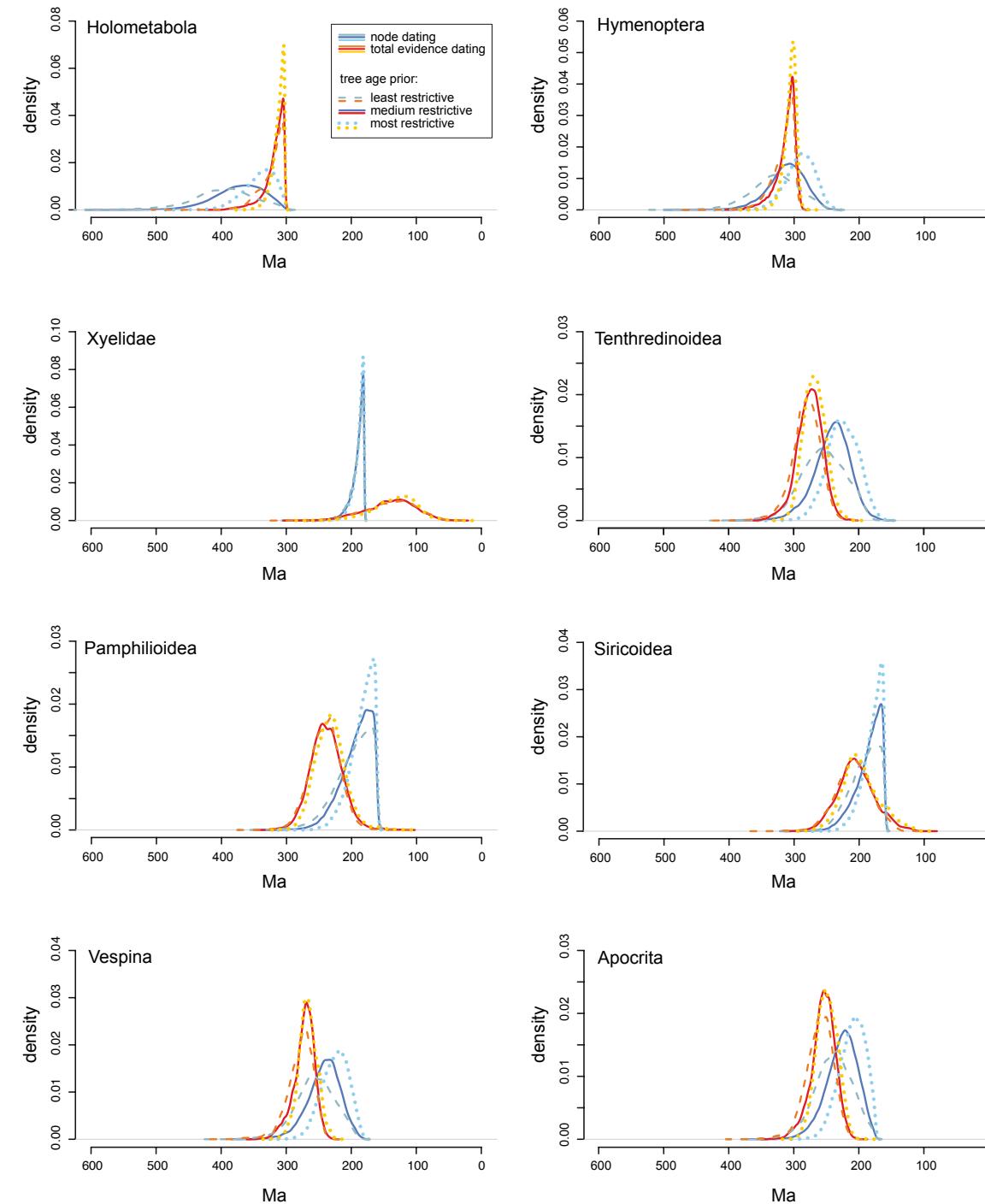


Table 2. Fossils used in the node-dating analyses and calibration point prior settings. The minimal and mean age for the offset-exponential prior are given, along with the corresponding fossils and references.

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

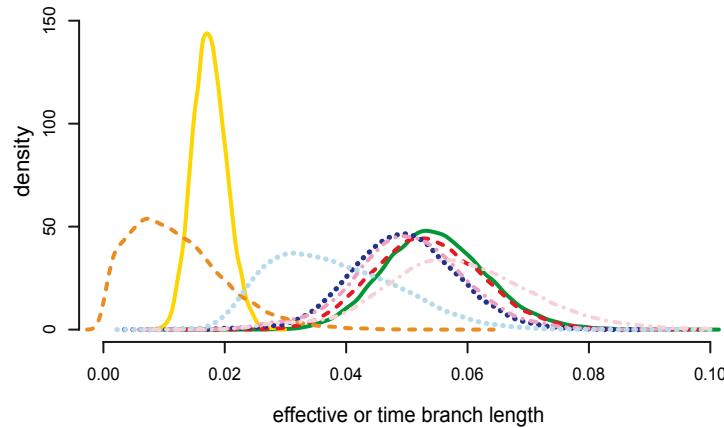
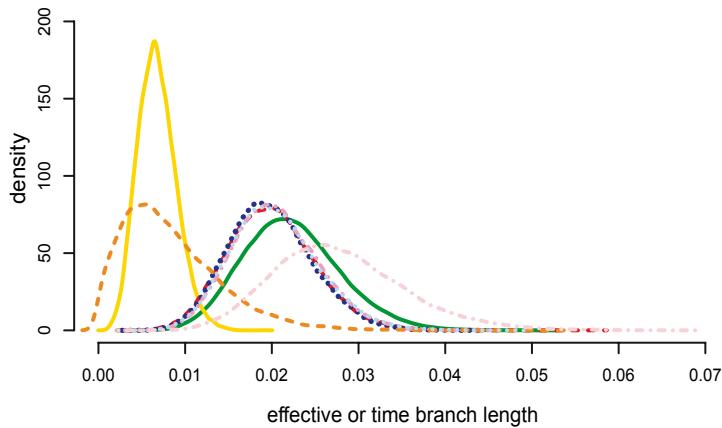
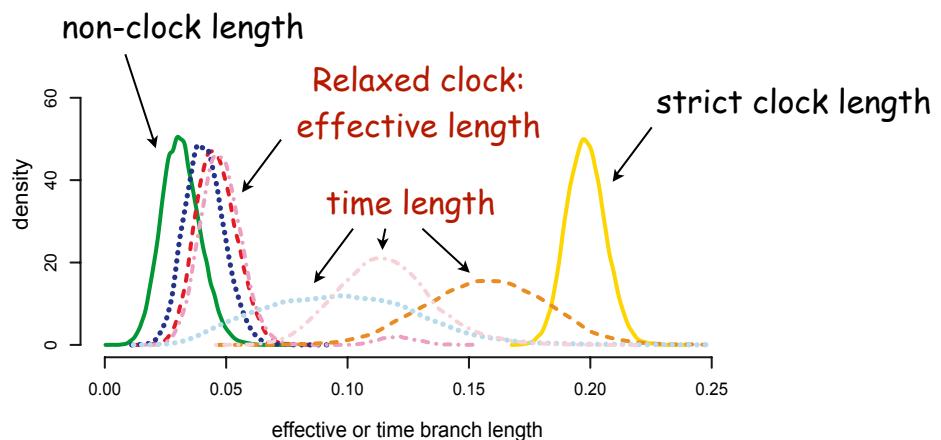
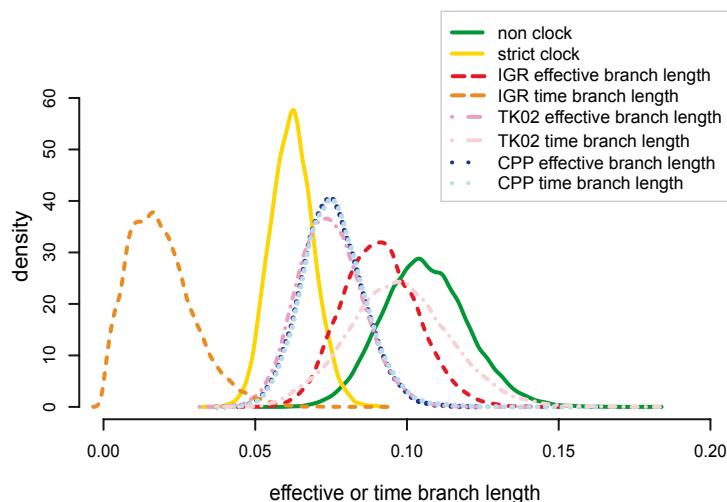
¹Posterior probability (PP) from the total-evidence analysis that the fossil attaches at the position assumed in the node-dating analysis. Note that these posterior probabilities take both the morphological data and the ages of the fossils into account.

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

³Tenthredinoidea 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

Acknowledgements

- Alex Rasnitsyn
- Lars Vilhelmsen
- Susanne Schulmeister
- Debra Murray
- Seraina Klopfstein
- Chi Zhang
- Tracy Heath
- Tanja Stadler
- Nicolas Lartillot
- Funding: US National Science Foundation (HymAToL), Swedish Research Council, Swiss National Science Foundation