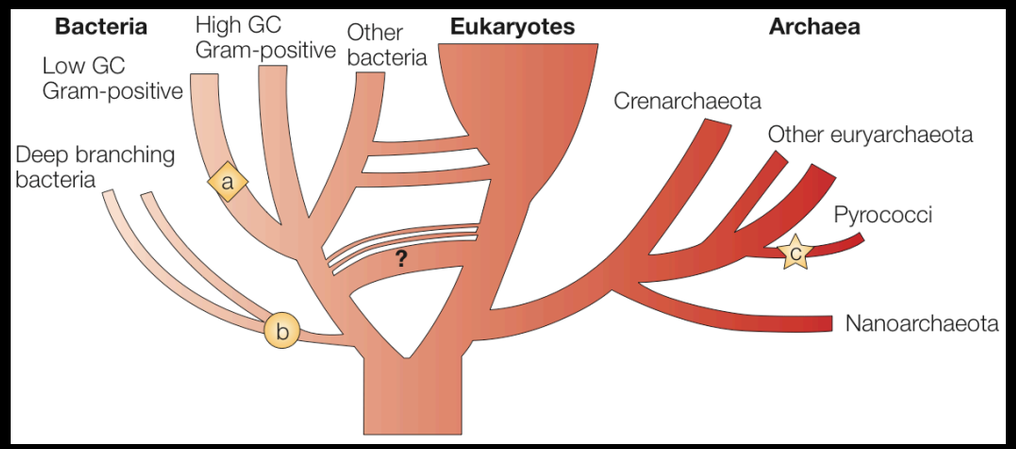
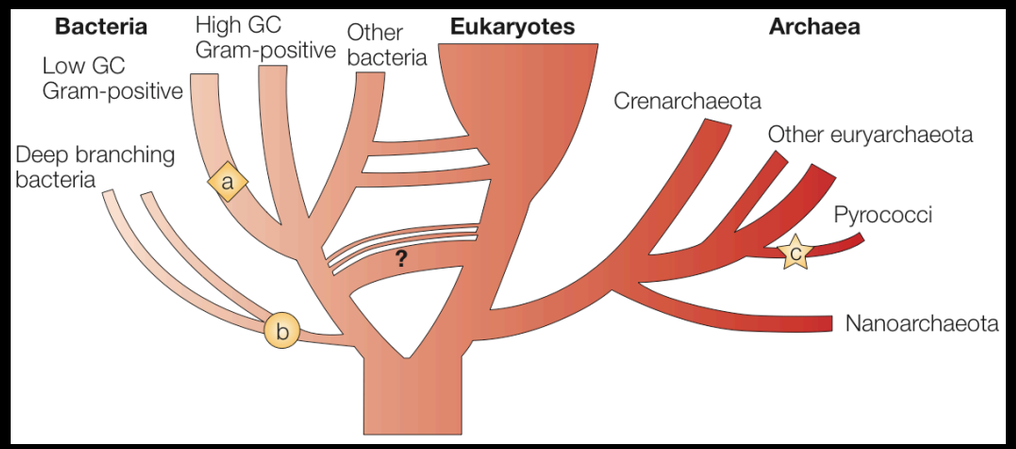
Phylogenomics

Jeffrey P. Townsend
Department of Ecology and Evolutionary Biology
Yale University

How may we improve our inferences?

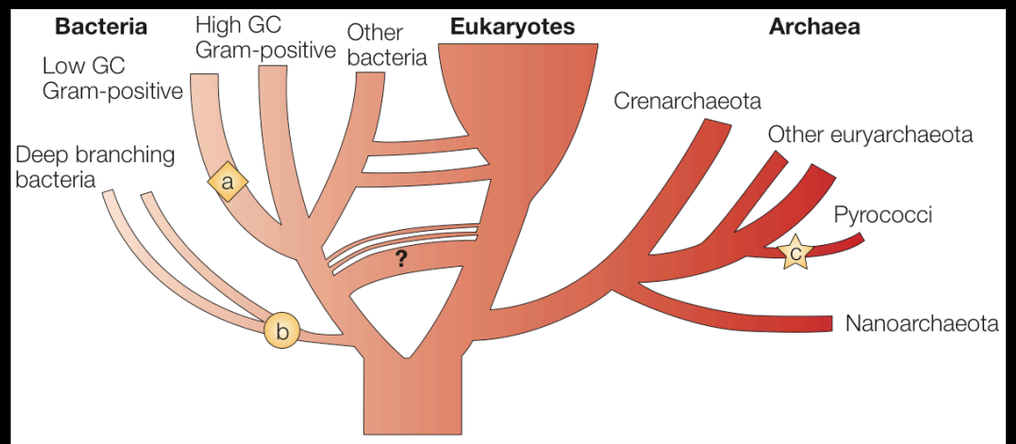


How may we improve our inferences?



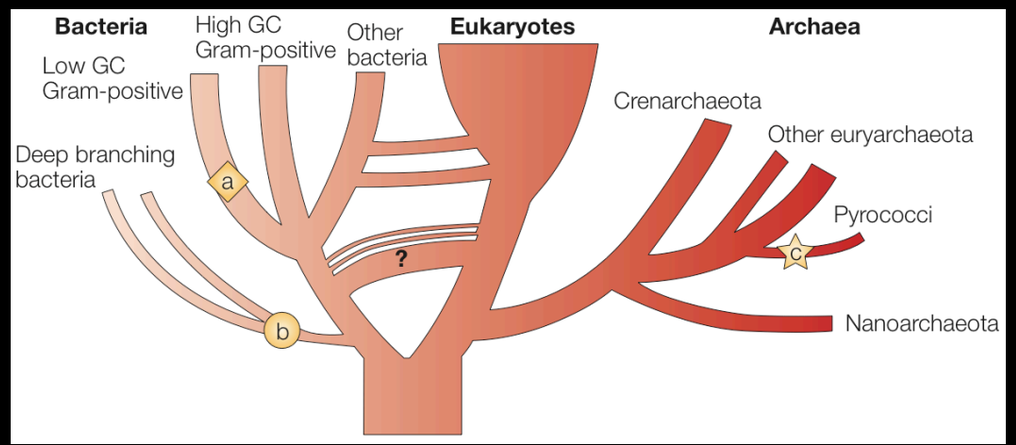
- Inferences I Data

How may we improve our inferences?



- Inferences I Data
- Data I Experimental Design

How may we improve our inferences?



- Inferences I Data
- Data I Experimental Design
- Experimental Design I Prior Knowledge

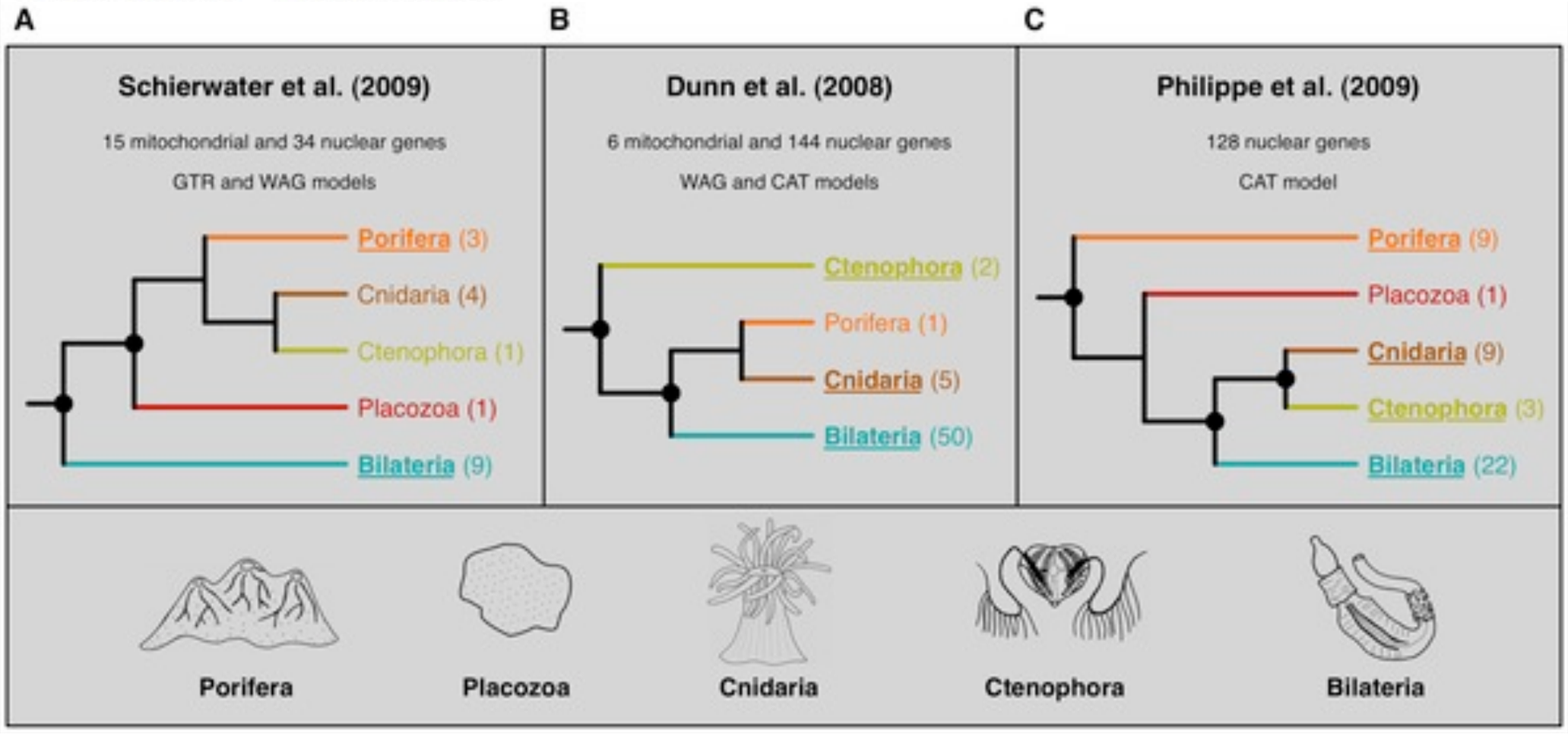
OPEN ACCESS Freely available online

PLoS BIOLOGY

Perspective

Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough

Hervé Philippe[1]*, Henner Brinkmann[1], Dennis V. Lavrov[2], D. Timothy J. Littlewood[3], Michael Manuel[4], Gert Wörheide[5,6], Denis Baurain[7]



Phylogenetic experimental design has a diffuse history

- Which types of characters are most informative? (Collins et al., 2005; Dequeiroz and Wimberger, 1993; Graybeal, 1994; Naylor and Brown, 1997; Rokas and Holland, 2000; Wiens and Servedio, 1997; Yang, 1998; Zwickl and Hillis, 2002)
- Would increased taxonomic or character sampling be more informative? (Graybeal, 1998; Hillis, 1998; Kim, 1996, 1998; Poe, 1998; Pollock et al., 2002; Rannala et al., 1998; Rokas and Carroll, 2005; Rosenberg and Kumar, 2001, 2003; Sullivan et al., 1999)
- Which taxa should be sampled to resolve a given phylogenetic problem? (Goldman, 1998; Huelsenbeck, 1991b; Kim, 1996, 1998; Poe, 2003)

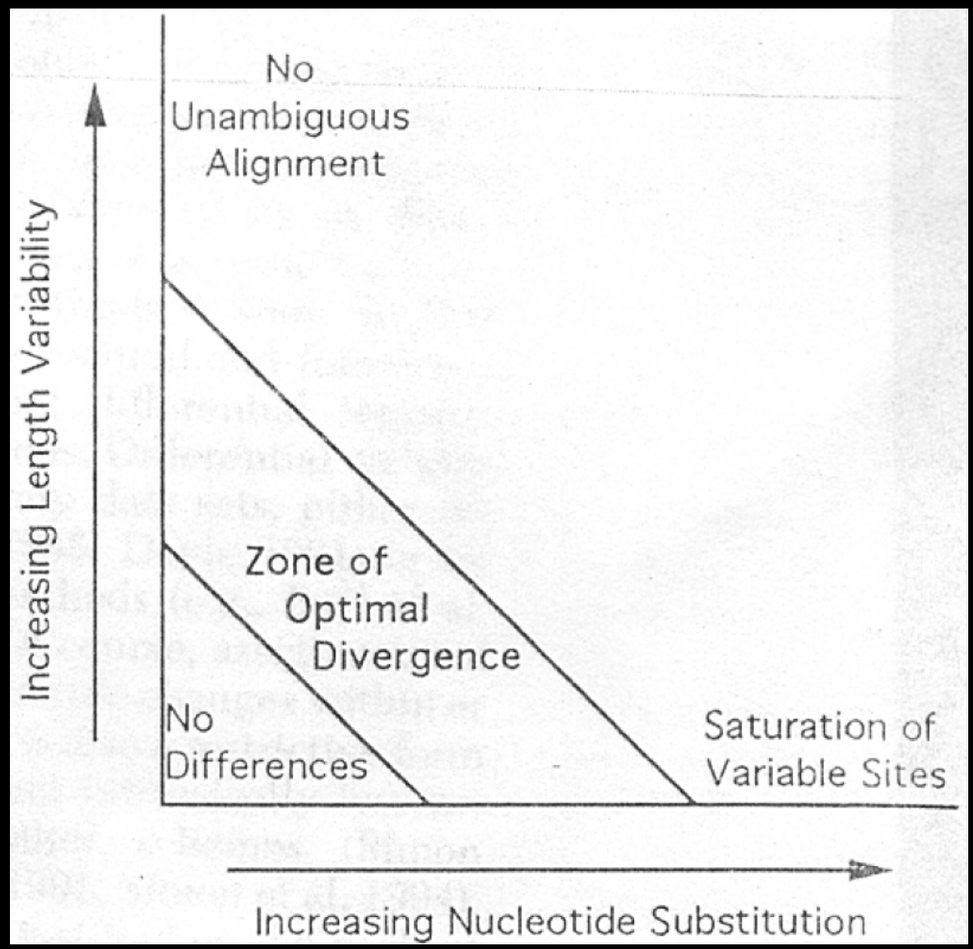
Phylogenetic experimental design has a diffuse history

- **Which types of characters are most informative?** (Collins et al., 2005; Dequeiroz and Wimberger, 1993; Graybeal, 1994; Naylor and Brown, 1997; Rokas and Holland, 2000; Wiens and Servedio, 1997; Yang, 1998; Zwickl and Hillis, 2002)
- Would increased taxonomic or character sampling be more informative? (Graybeal, 1998; Hillis, 1998; Kim, 1996, 1998; Poe, 1998; Pollock et al., 2002; Rannala et al., 1998; Rokas and Carroll, 2005; Rosenberg and Kumar, 2001, 2003; Sullivan et al., 1999)
- Which taxa should be sampled to resolve a given phylogenetic problem? (Goldman, 1998; Huelsenbeck, 1991b; Kim, 1996, 1998; Poe, 2003)

Which types of characters are most informative?

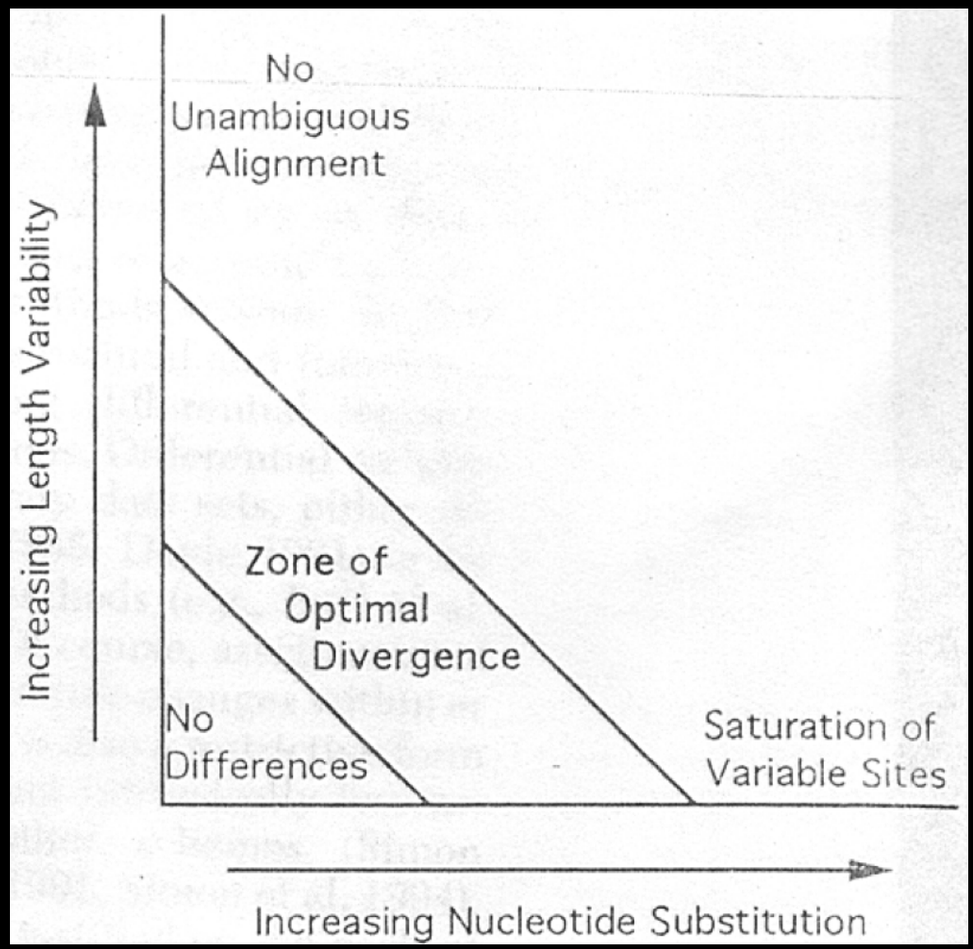
Ancient utility	Recent utility	Reference
Morphological	Behavioral	(Dequieroz & Wimberger, 1993)
Polymorphic	Fixed	(Wiens & Servedio, 1997)
Nonsynonymous	Synonymous	(Graybeal, 1994; Naylor & Brown, 1997)
Amino acids	Nucleotides	(Russo et al. 1996, Gissi et al. 2006)
Rarely changing	characters are best	(many; Rokas & Holland, 2000)

Can we visualize phylogenetic informativeness?



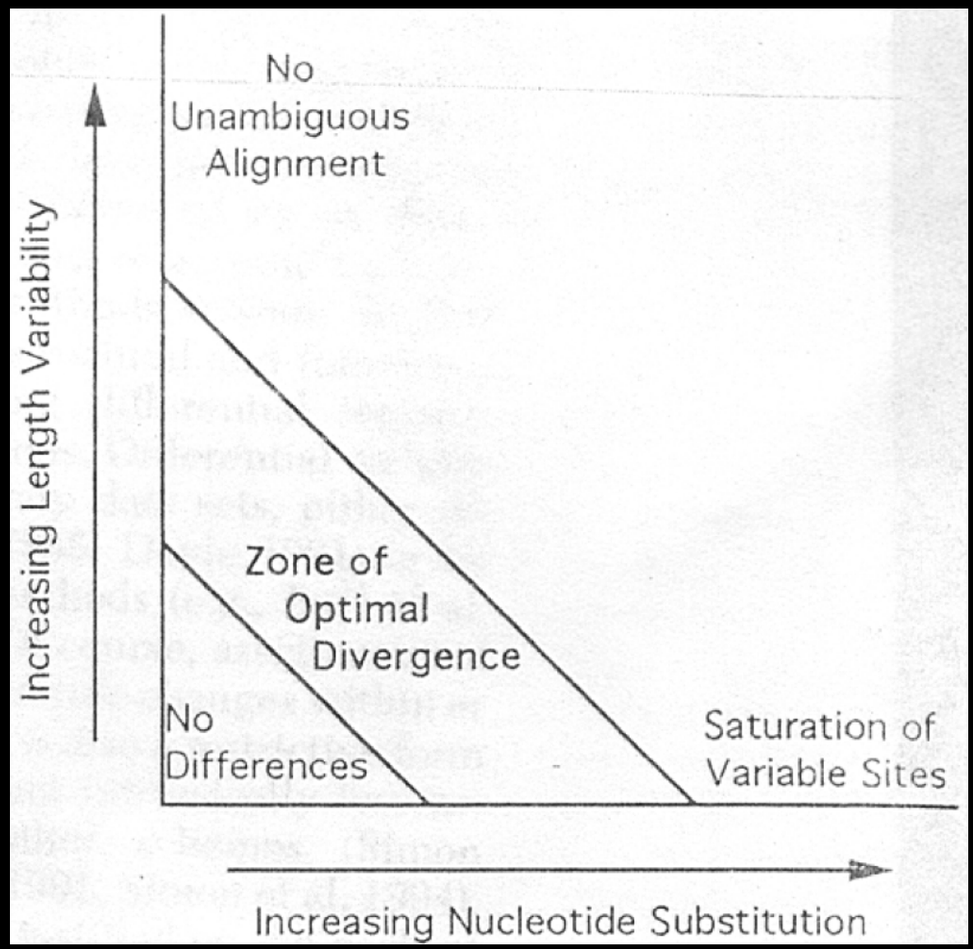
Can we visualize phylogenetic informativeness?

- Characters with no differences across taxa are useless



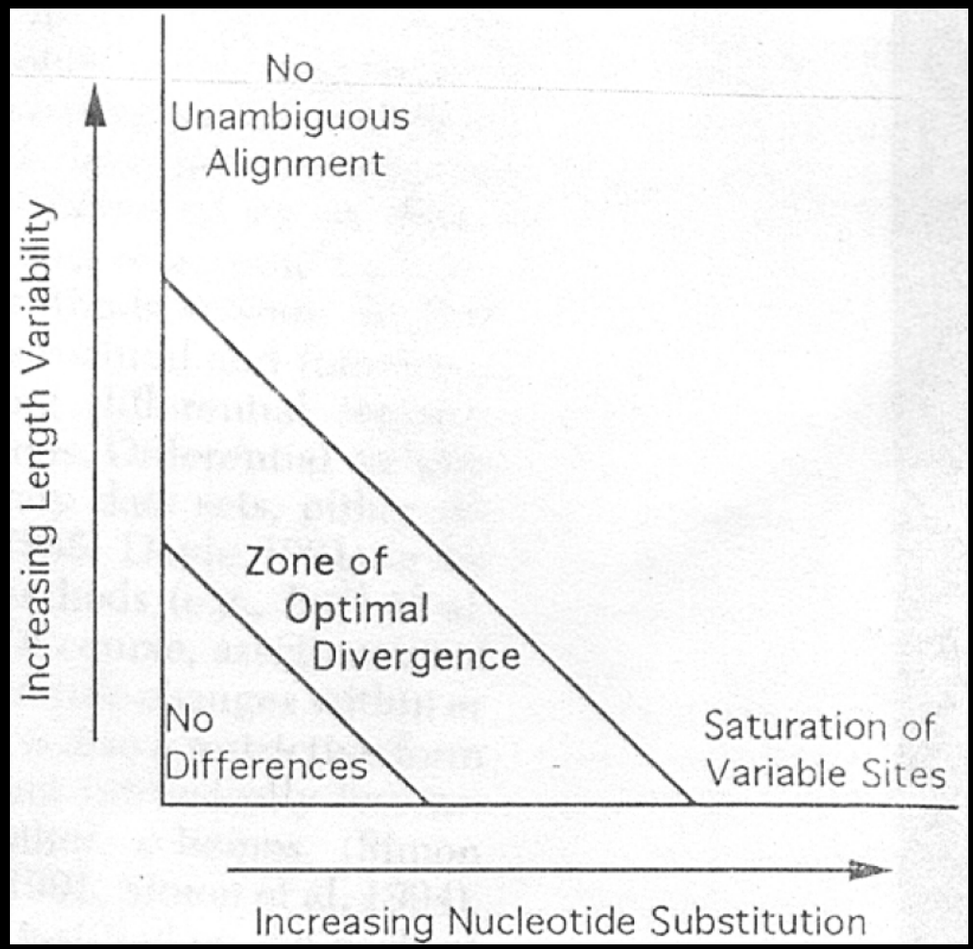
Can we visualize phylogenetic informativeness?

- Characters with no differences across taxa are useless
- Characters that are saturated with mutations across taxa are misleading



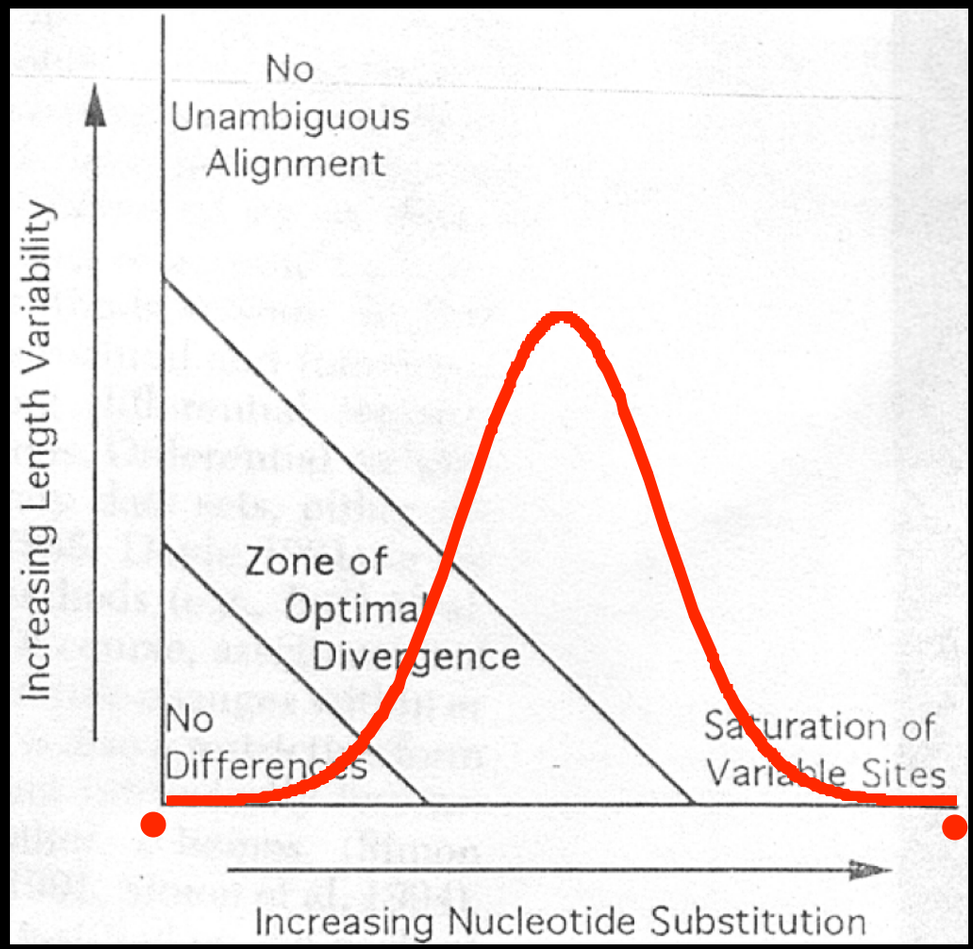
Can we visualize phylogenetic informativeness?

- Characters with no differences across taxa are useless
- Characters that are saturated with mutations across taxa are misleading
- There exists a happy medium



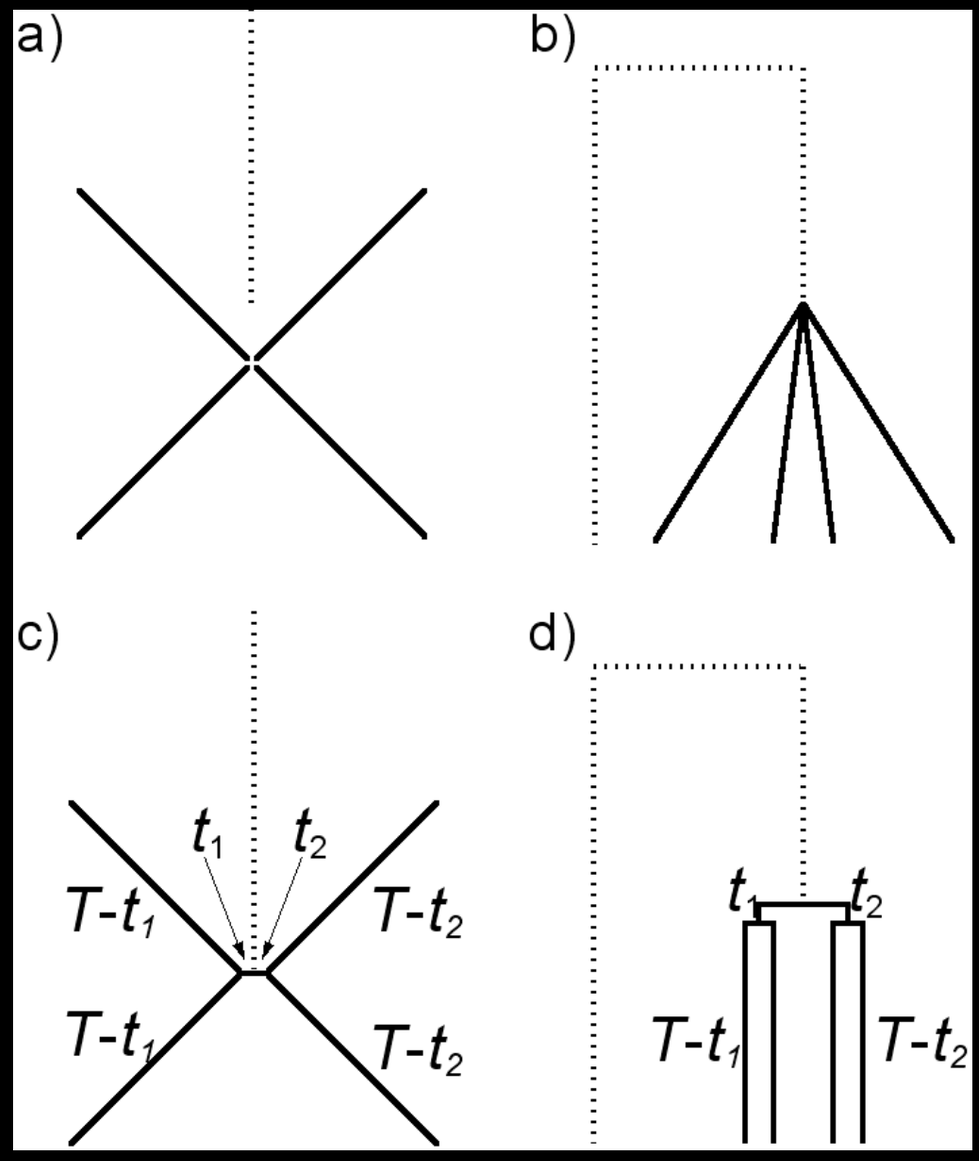
Can we visualize phylogenetic informativeness?

- Characters with no differences across taxa are useless
- Characters that are saturated with mutations across taxa are misleading
- There exists a happy medium



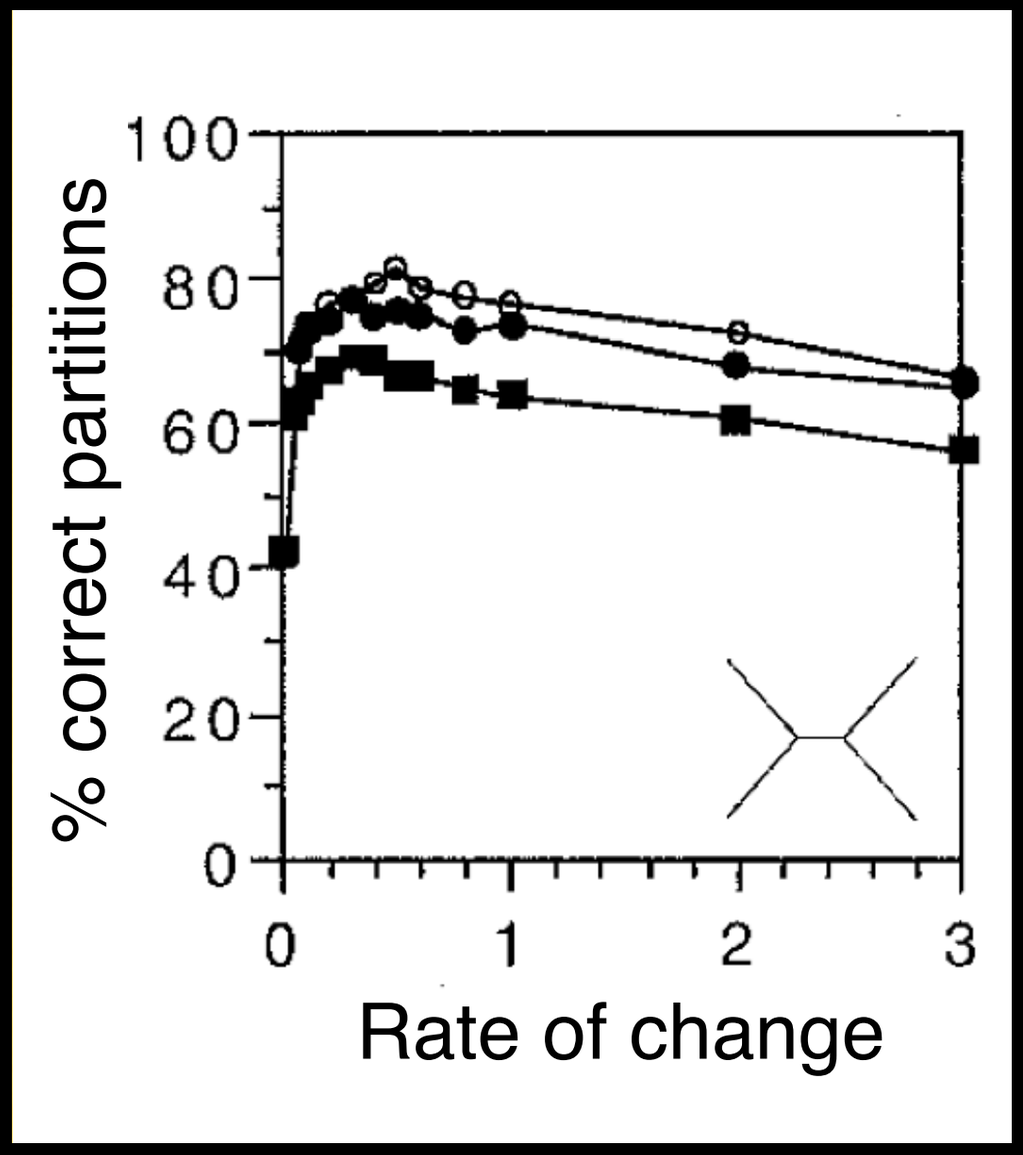
Deriving phylogenetic informativeness

- Calculate the probability of a site being informative, given a rate of evolution, λ , and letting $t_1 + t_2 \rightarrow 0$.
- $\hat{\lambda} = \frac{1}{4}T$
-



Townsend, 2007, *Systematic Biology* 56:222-231

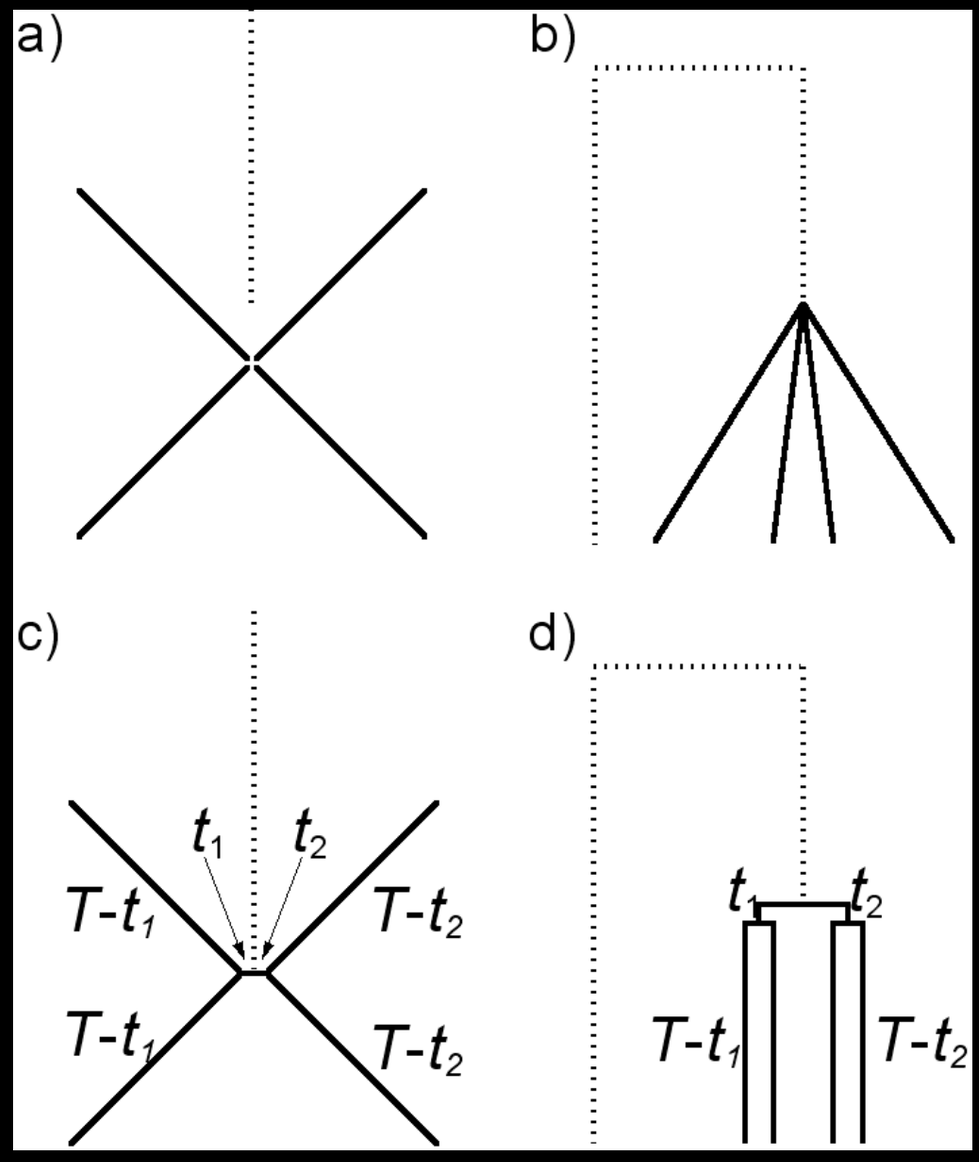
Simulations demonstrate a 1/4*T* optimal rate of character change



Yang, *Systematic Biology* (1998)

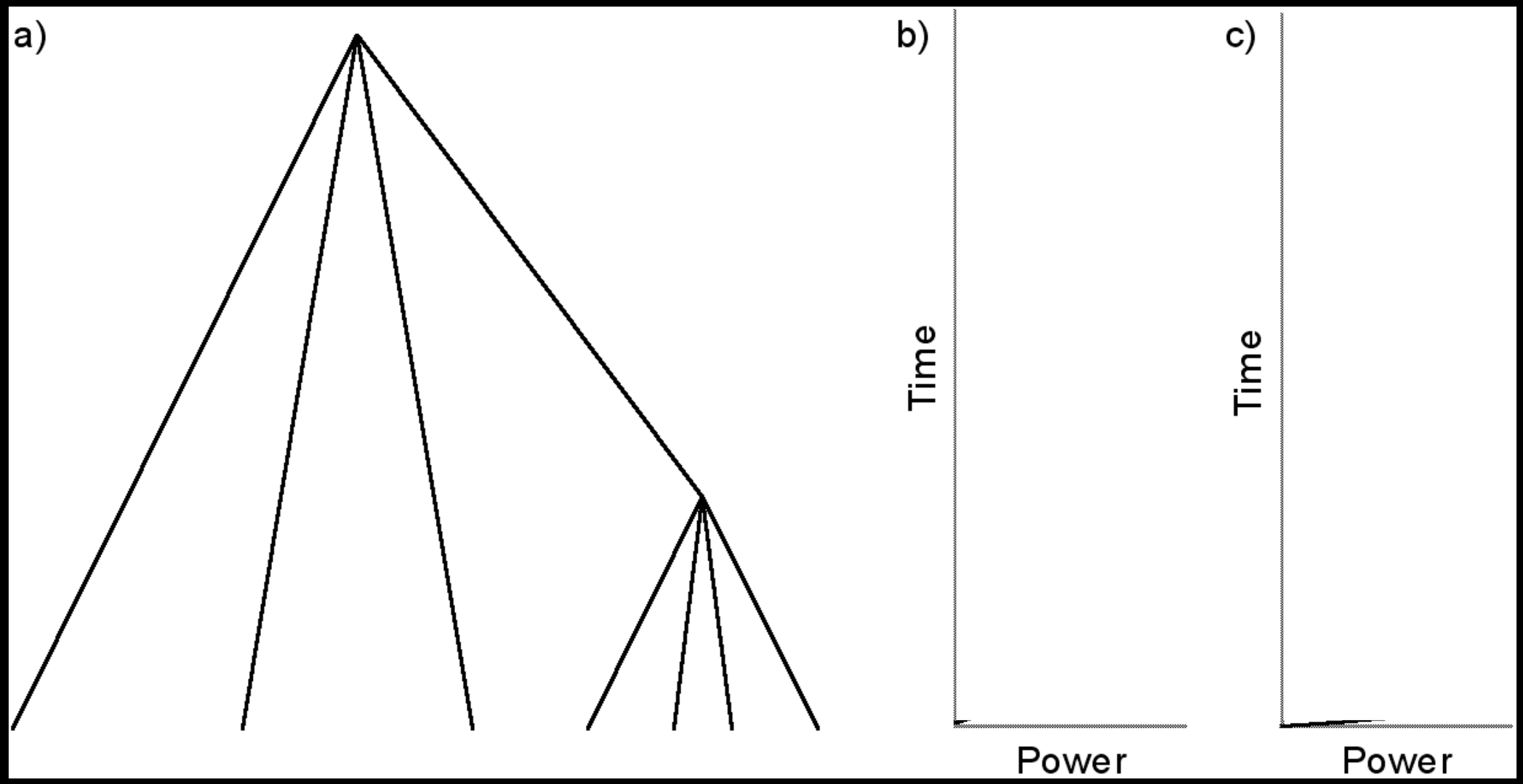
Deriving phylogenetic informativeness

- Calculate the probability of a site being informative, given a rate of evolution, λ , and letting $t_1 + t_2 \rightarrow 0$.
- $\hat{\lambda} = \frac{1}{4}T$
- $\rho\left(T;\lambda_1,\ldots,\lambda_n\right) = \sum_{i=1}^{n} 16\lambda_i{}^2 T e^{-4\lambda_i T}$



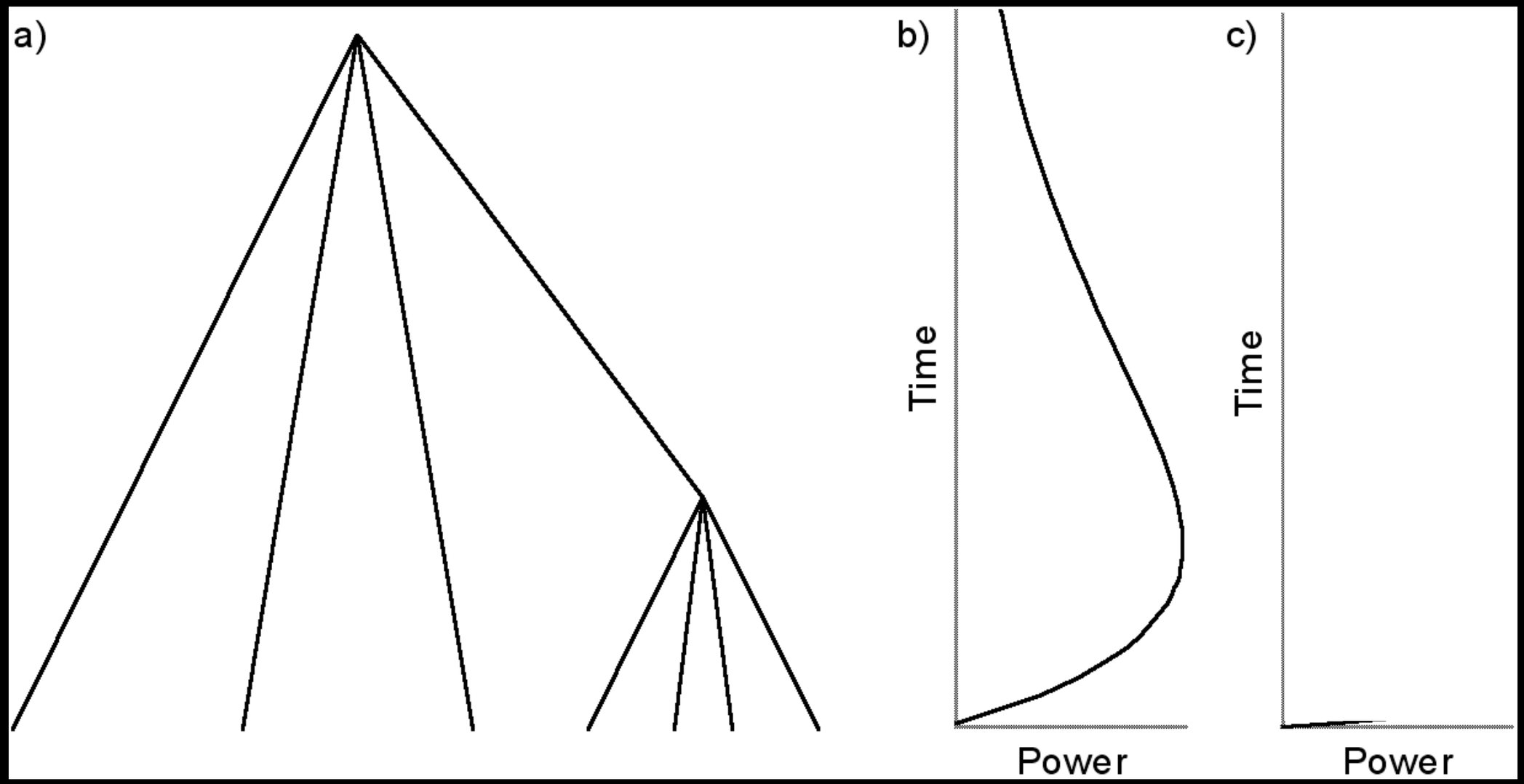
Townsend, 2007, *Systematic Biology* 56:222-231

Informativeness is not obvious



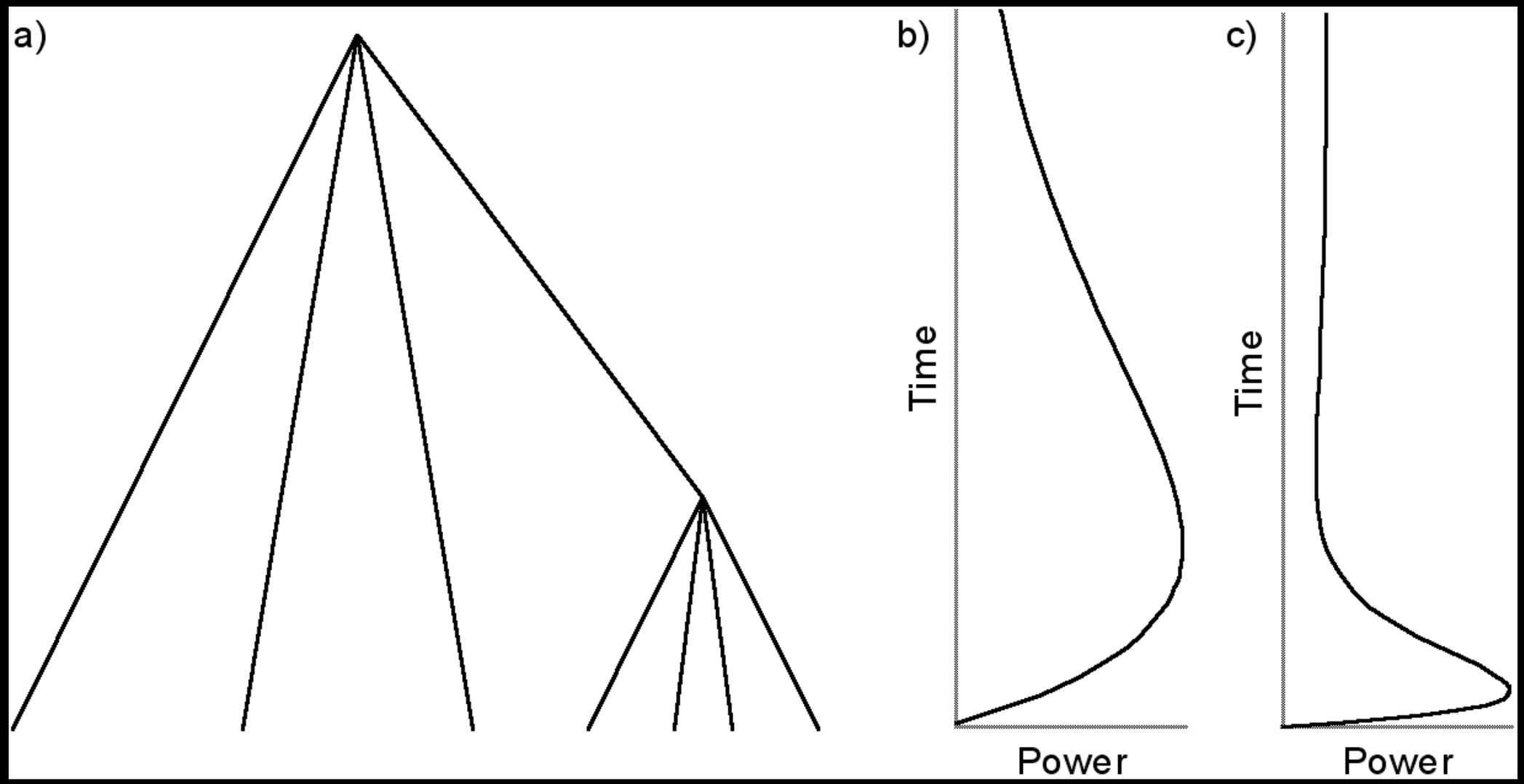
Townsend, 2007, *Systematic Biology* 56:222-231

Informativeness is not obvious



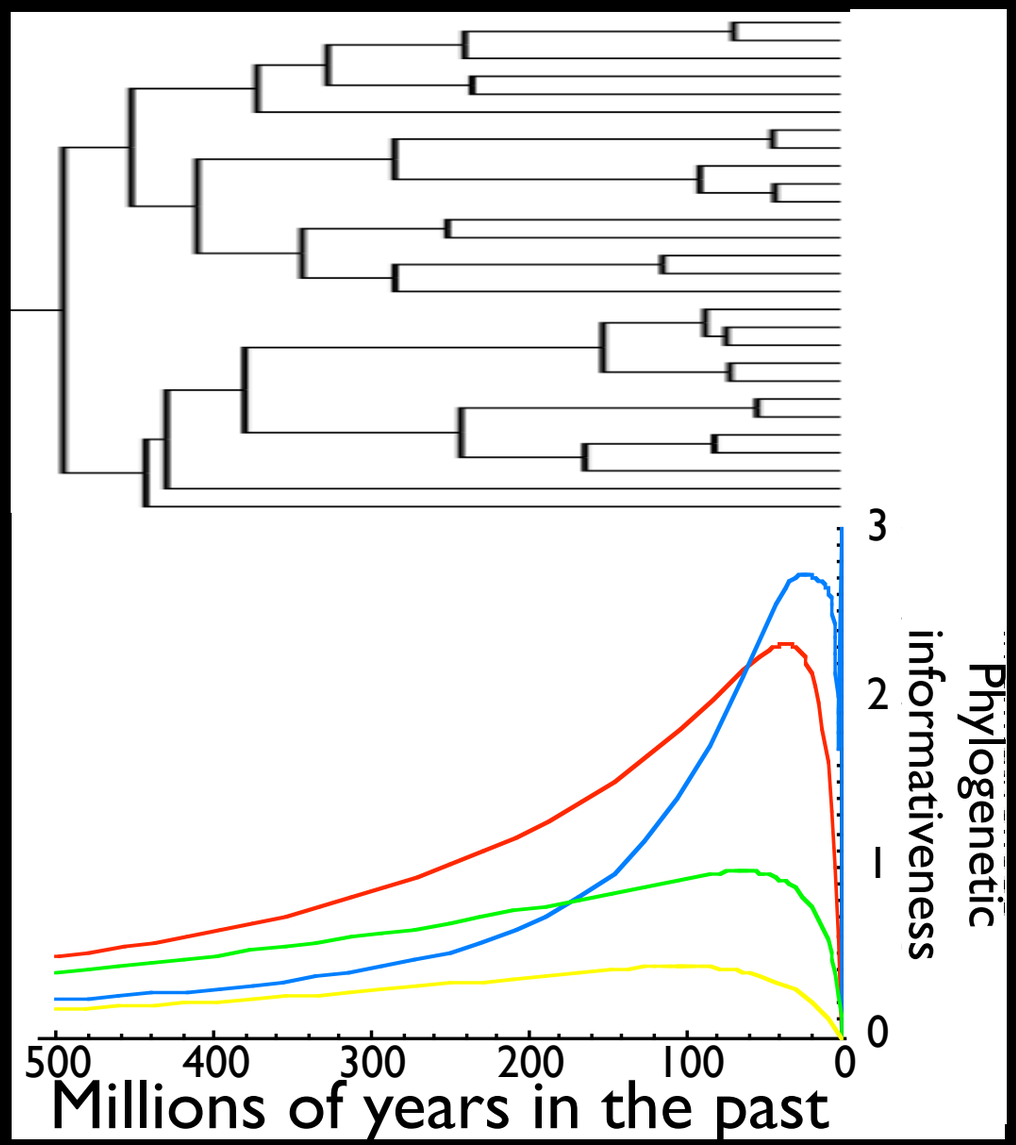
Townsend, 2007, *Systematic Biology* 56:222-231

Informativeness is not obvious



Townsend, 2007, *Systematic Biology* 56:222-231

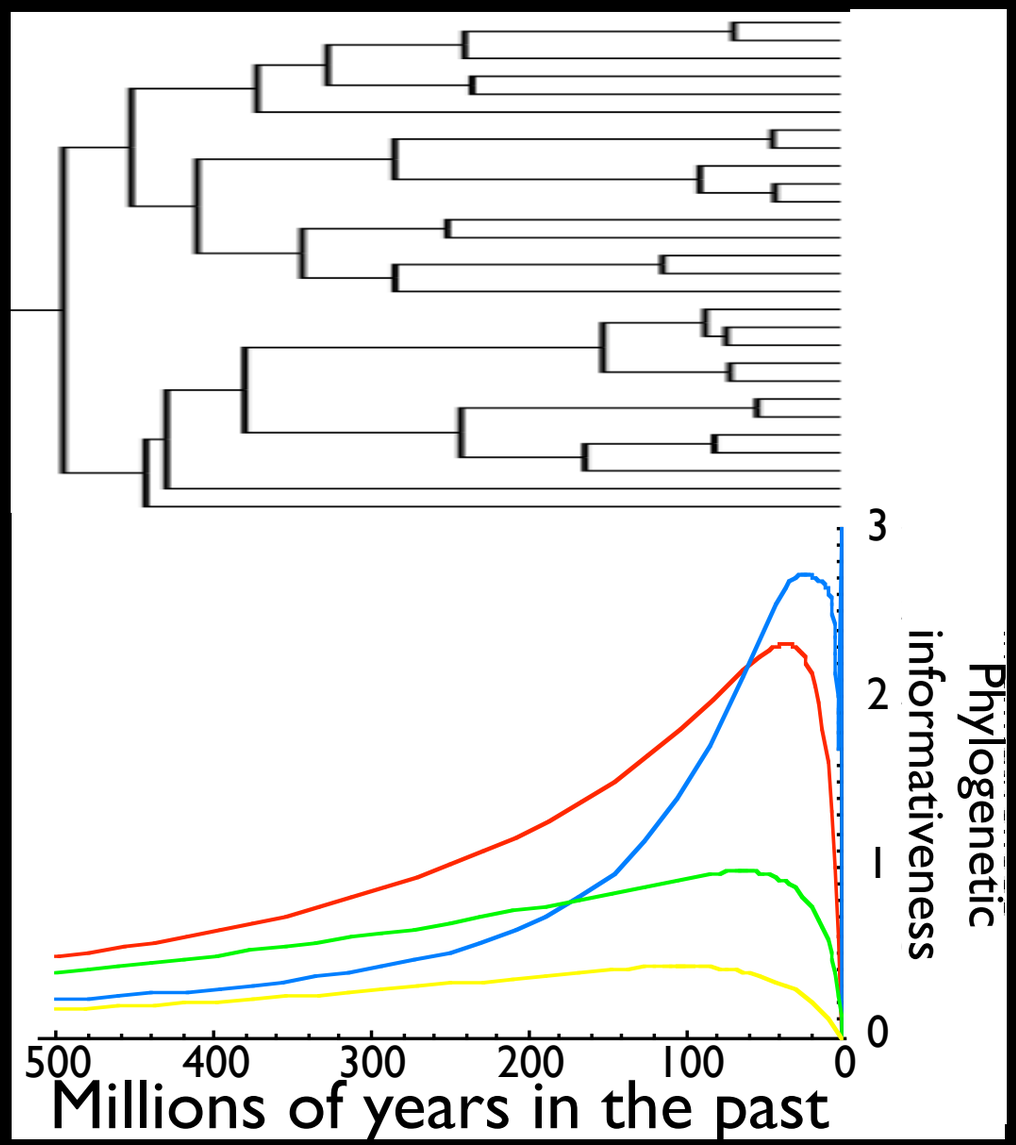
Phylogenetic informativeness yields predictions of relative performance



- Generally, higher phylogenetic informativeness of a gene indicates a higher ability to resolve the corresponding node.

Townsend et al., 2008, *JME* 67: 222-231

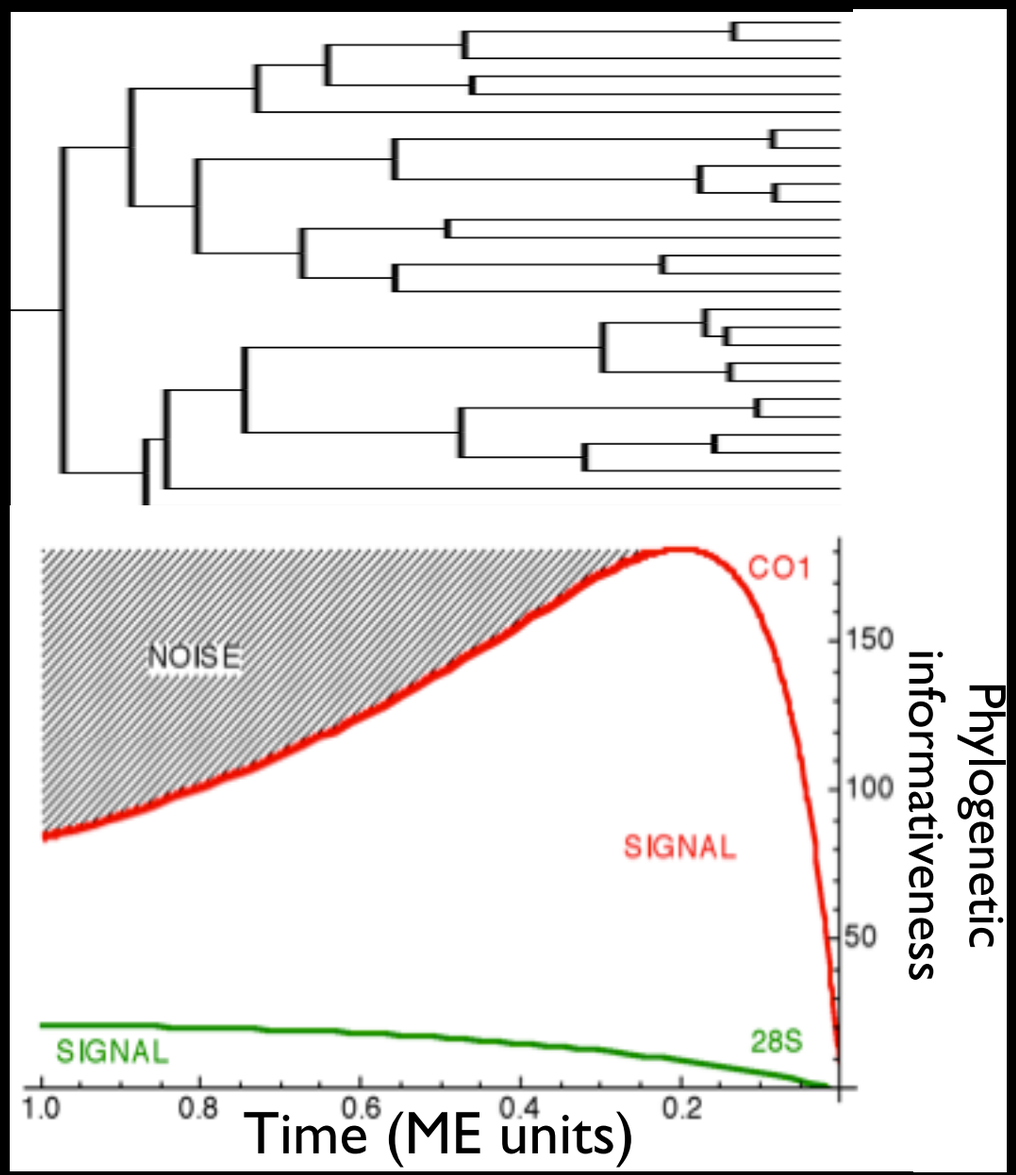
Phylogenetic informativeness must be considered with caveats



- Informativeness profiles provide no clear expectation of performance for specific nodes
- No reason to expect an x-fold difference in informativeness would result in an X-fold difference in the number of nodes resolved
- Profiles of informativeness predict signal but not the misleading effect of noise (convergence or parallelism)

Townsend et al., 2008, *JME* 67: 222-231

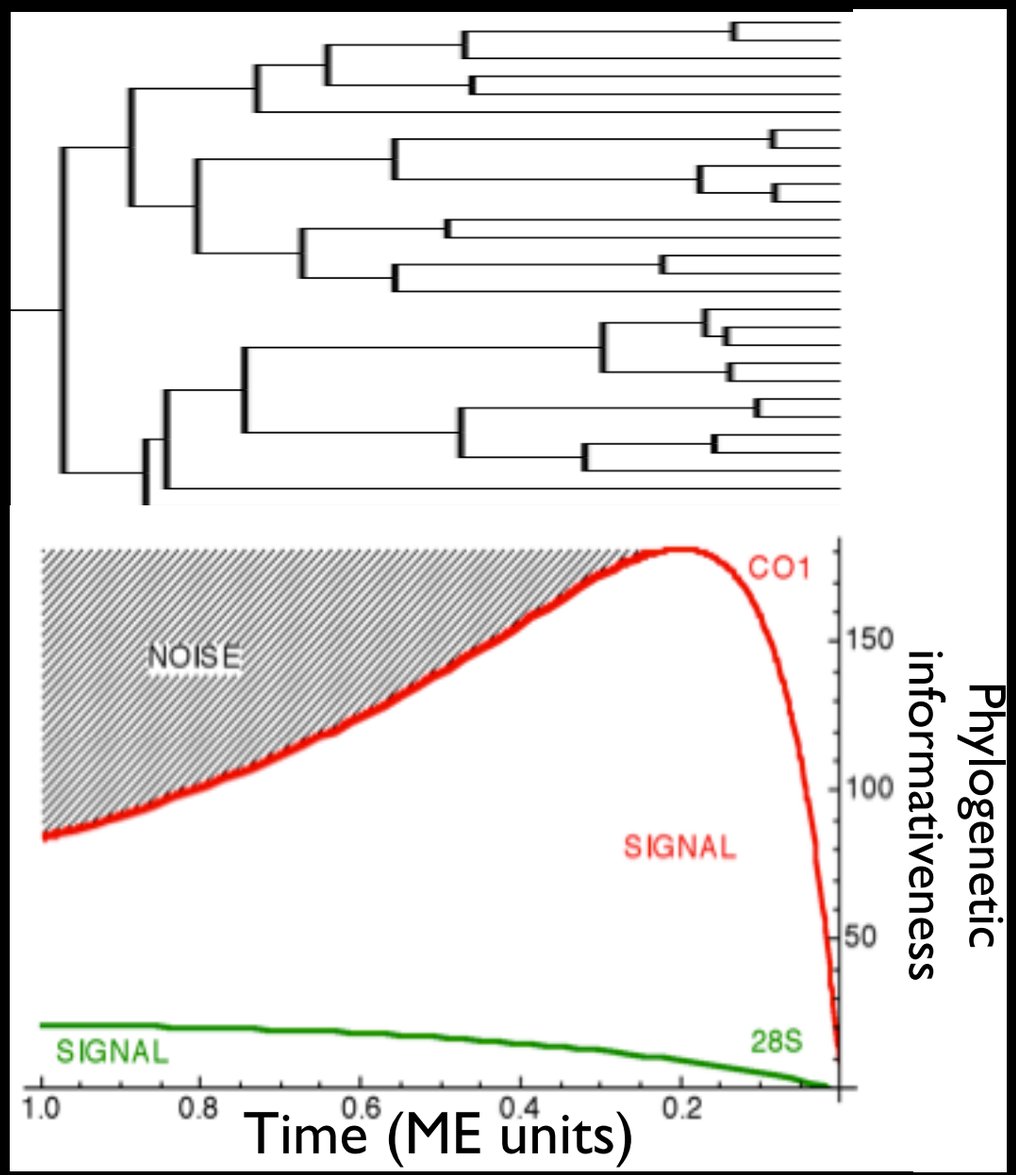
With profiles alone, the effect of noise can be depicted, but not quantified



- Informativeness profiles do not account for misleading effects of noise (convergence or parallelism)
- No quantitative rule, but as a rule of thumb, selecting genes that peak deeper than the time interval of interest will minimize the influence of noise

Townsend et al., 2008, *JME* 67: 222-231

Large data sets lead to both signal and noise



- Gathering large data sets may enable signal to outweigh noise...
- ...but signal *and* noise are contributed by larger datasets.
- Thus, large datasets can lead to spurious resolutions of deep polytomies.
- Deep time and short internodes exacerbate this issue.

Townsend et al., 2008, *JME* 67: 222-231

Matrix methods quantify signal and noise

Townsend et al., 2012, *Systematic Biology*

Matrix methods quantify signal and noise

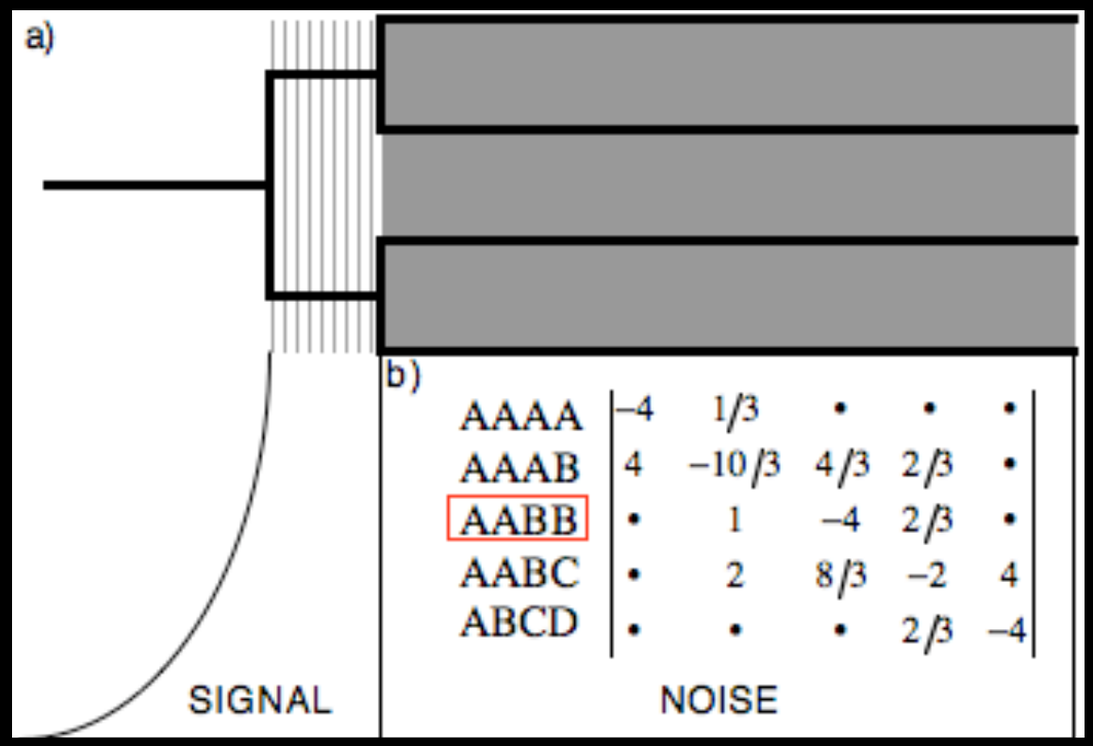
Townsend et al., 2012, *Systematic Biology*

Matrix methods quantify signal and noise

- Begin with a four taxon tree, branches to time T, internode t_0

Townsend et al., 2012, *Systematic Biology*

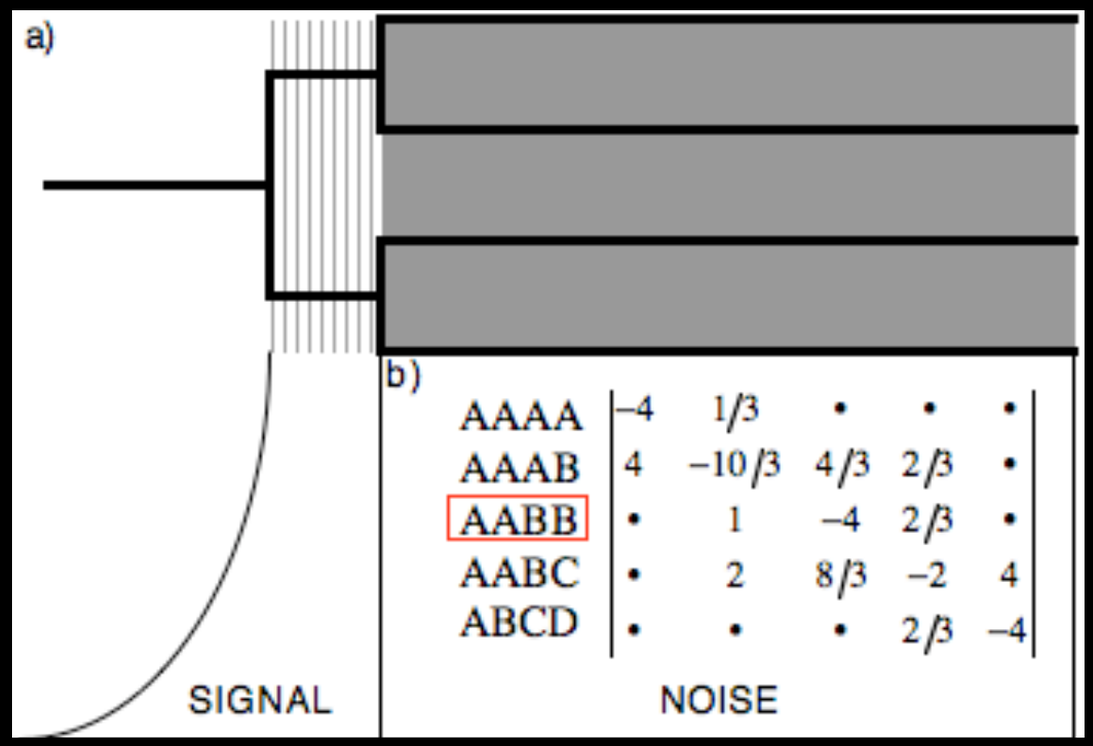
Matrix methods quantify signal and noise



- Begin with a four taxon tree, branches to time T, internode t_0

Townsend et al., 2012, *Systematic Biology*

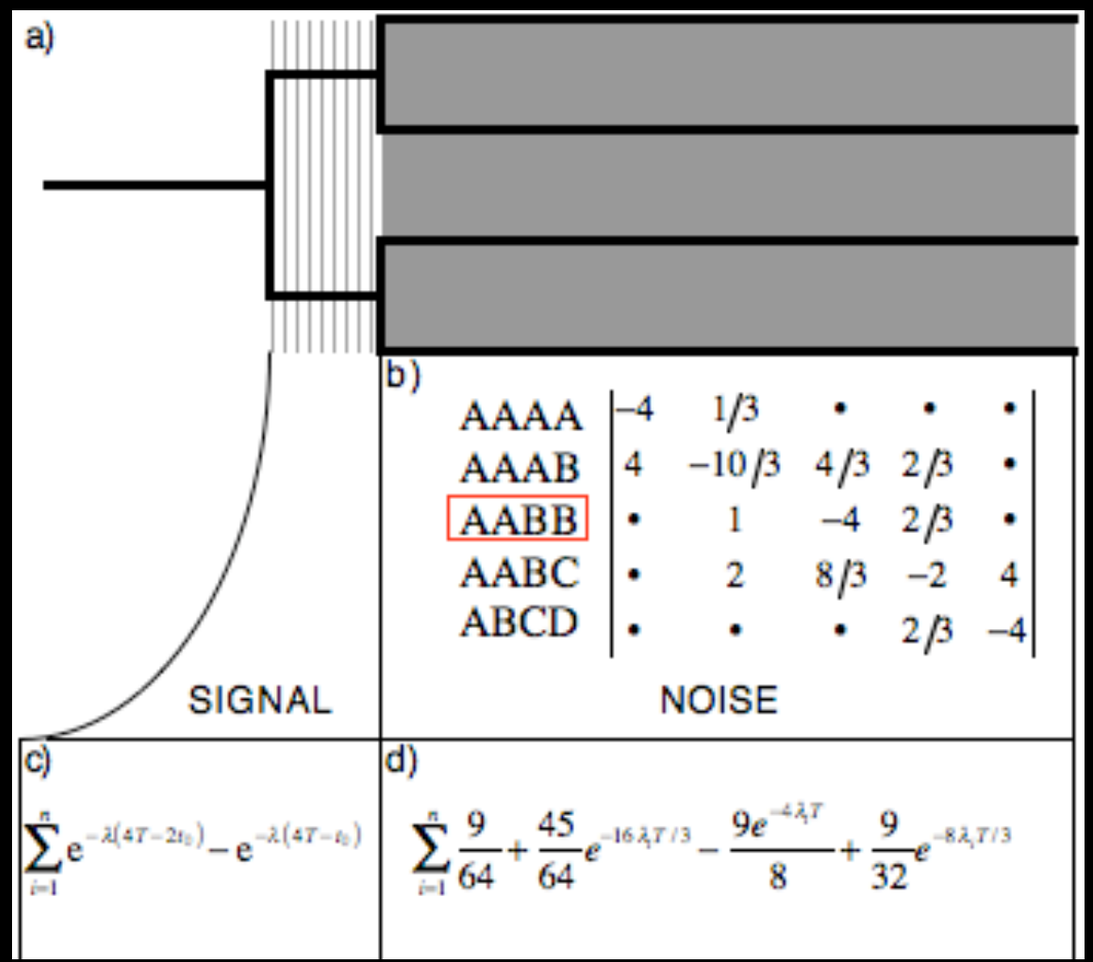
Matrix methods quantify signal and noise



- Begin with a four taxon tree, branches to time T, internode t_0
- Derive transition matrix for the site patterns for the four species: AAAA, AAAB, AABB, AABC, ABCD

Townsend et al., 2012, *Systematic Biology*

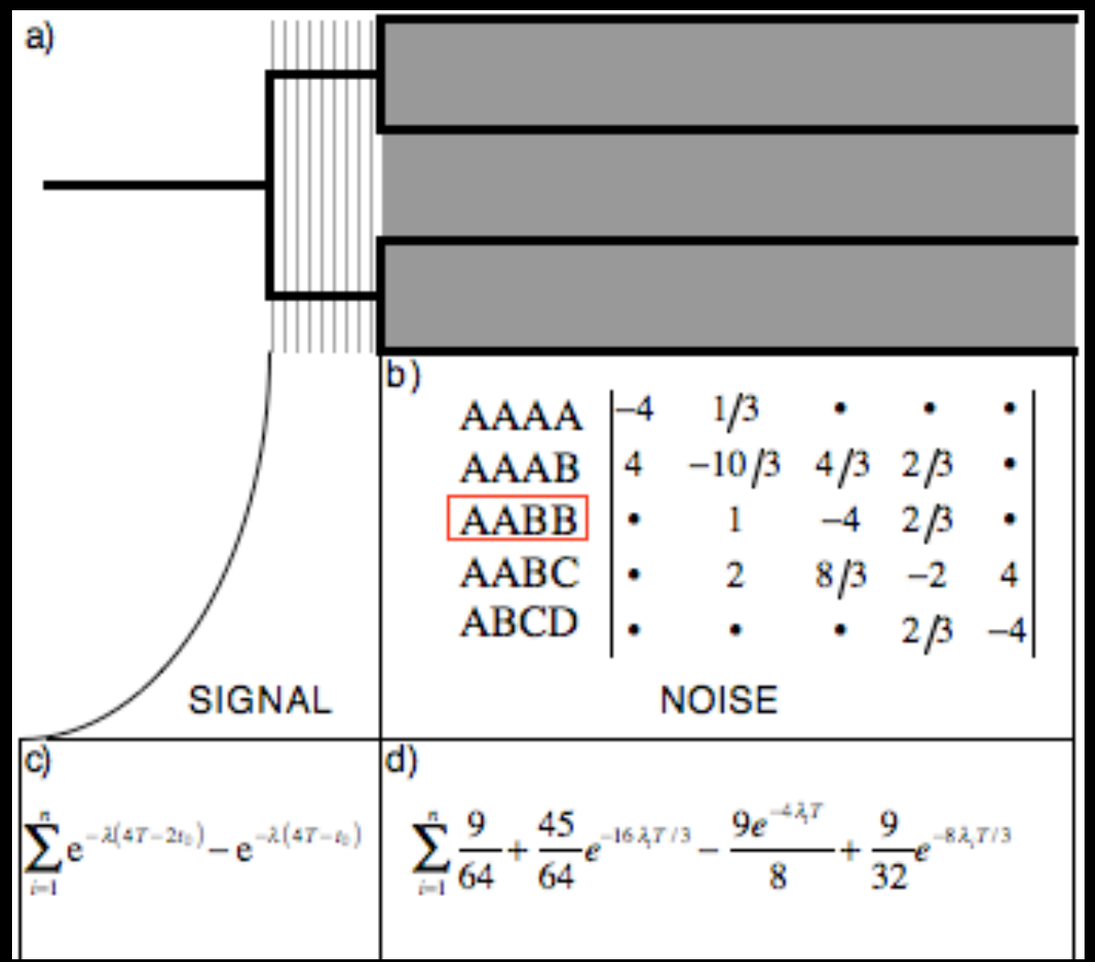
Matrix methods quantify signal and noise



- Begin with a four taxon tree, branches to time T, internode t_0
- Derive transition matrix for the site patterns for the four species: AAAA, AAAB, AABB, AABC, ABCD

Townsend et al., 2012, *Systematic Biology*

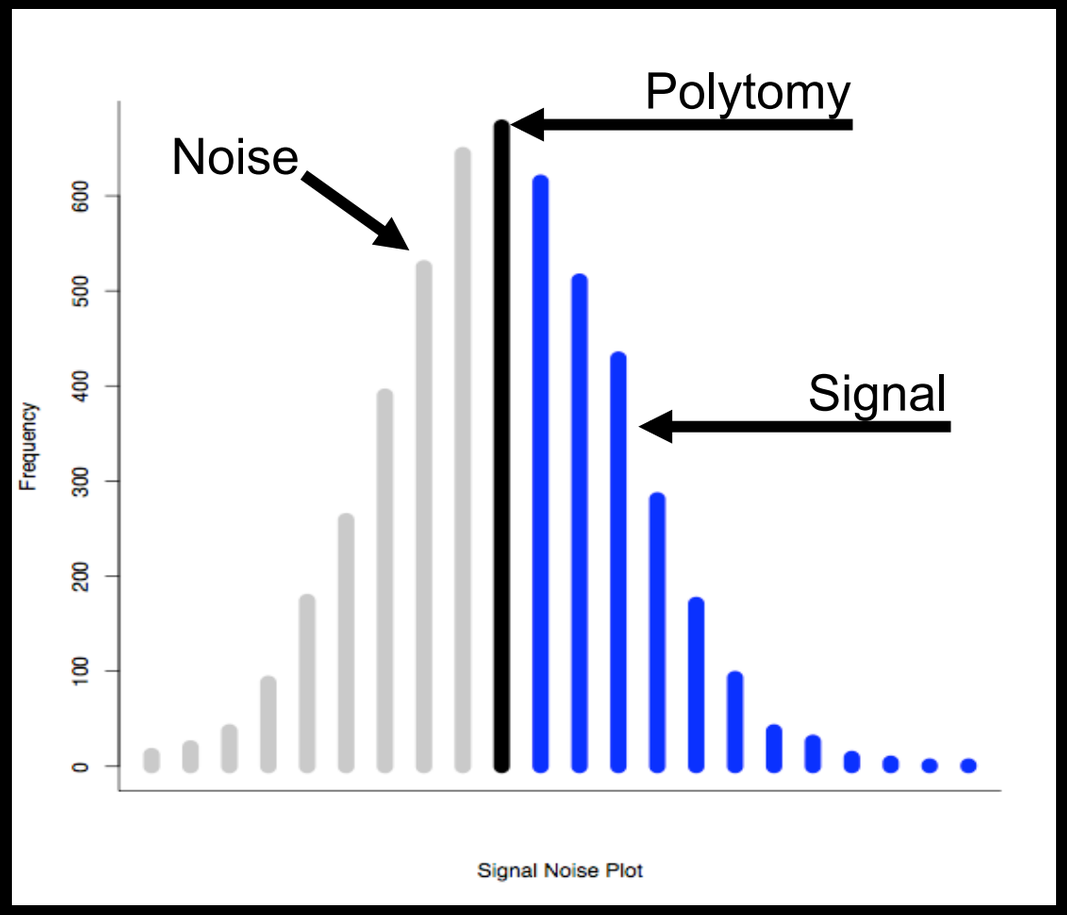
Matrix methods quantify signal and noise



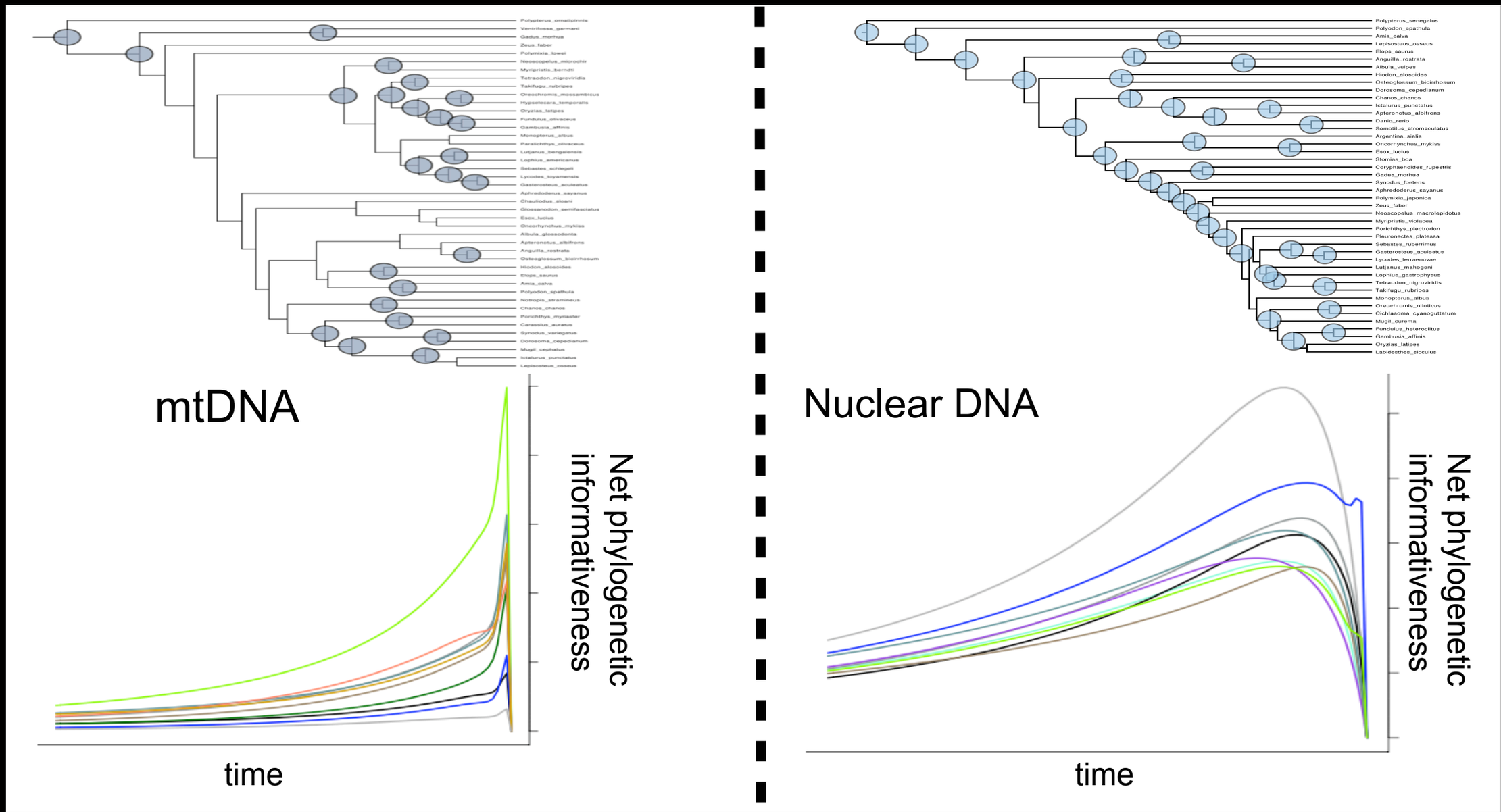
- Begin with a four taxon tree, branches to time T, internode t_0
- Derive transition matrix for the site patterns for the four species: AAAA, AAAB, AABB, AABC, ABCD
- Power to resolve a node can then be reduced to a biased random walk problem, with each site representing a "step."

Townsend et al., 2012, *Systematic Biology*

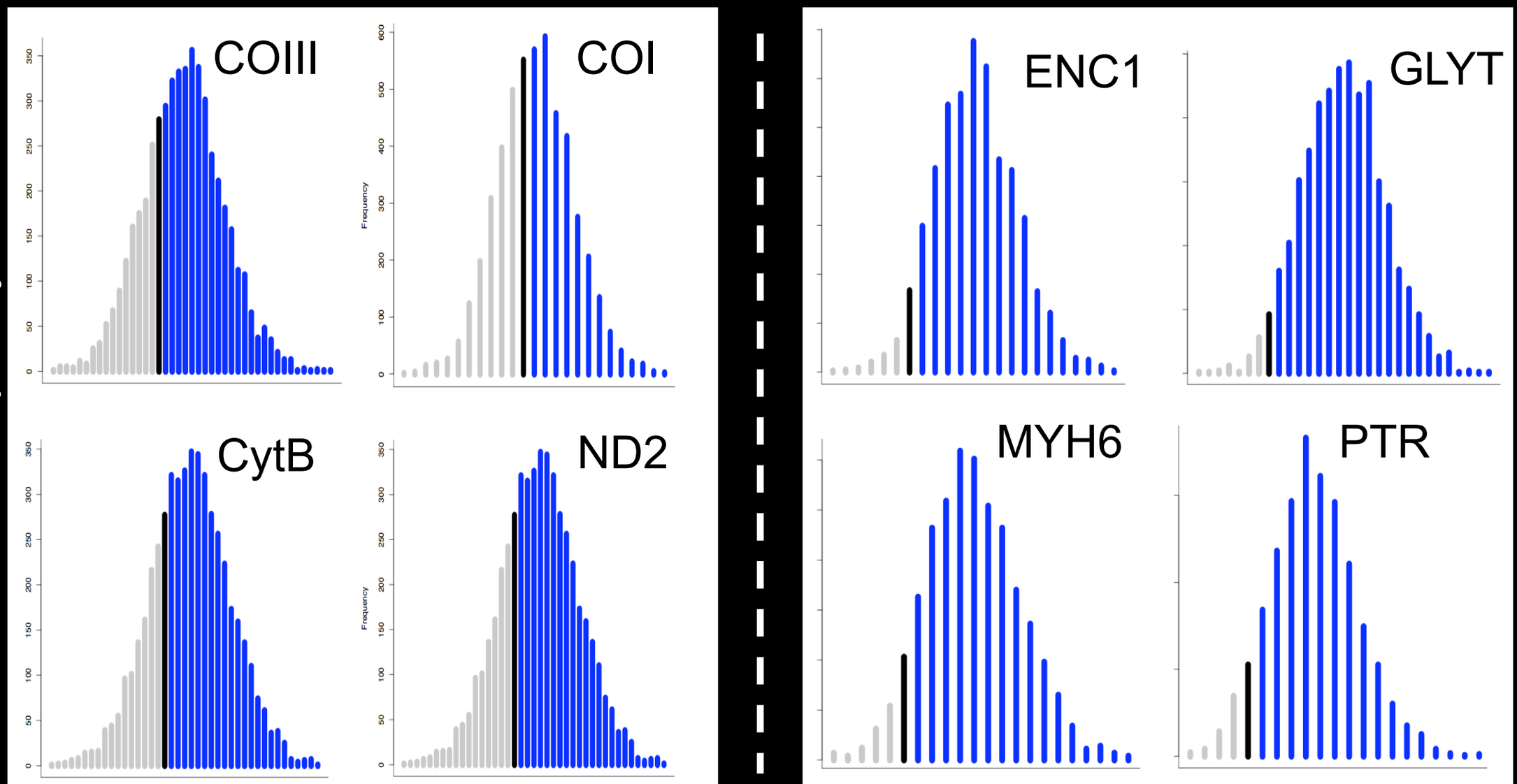
Each site contributes to the probability of correct resolution, incorrect resolution, or polytomy



- Probability signal 0.43
- Probability polytomy 0.14
- Probability noise 0.43



Dornburg et al., in revision

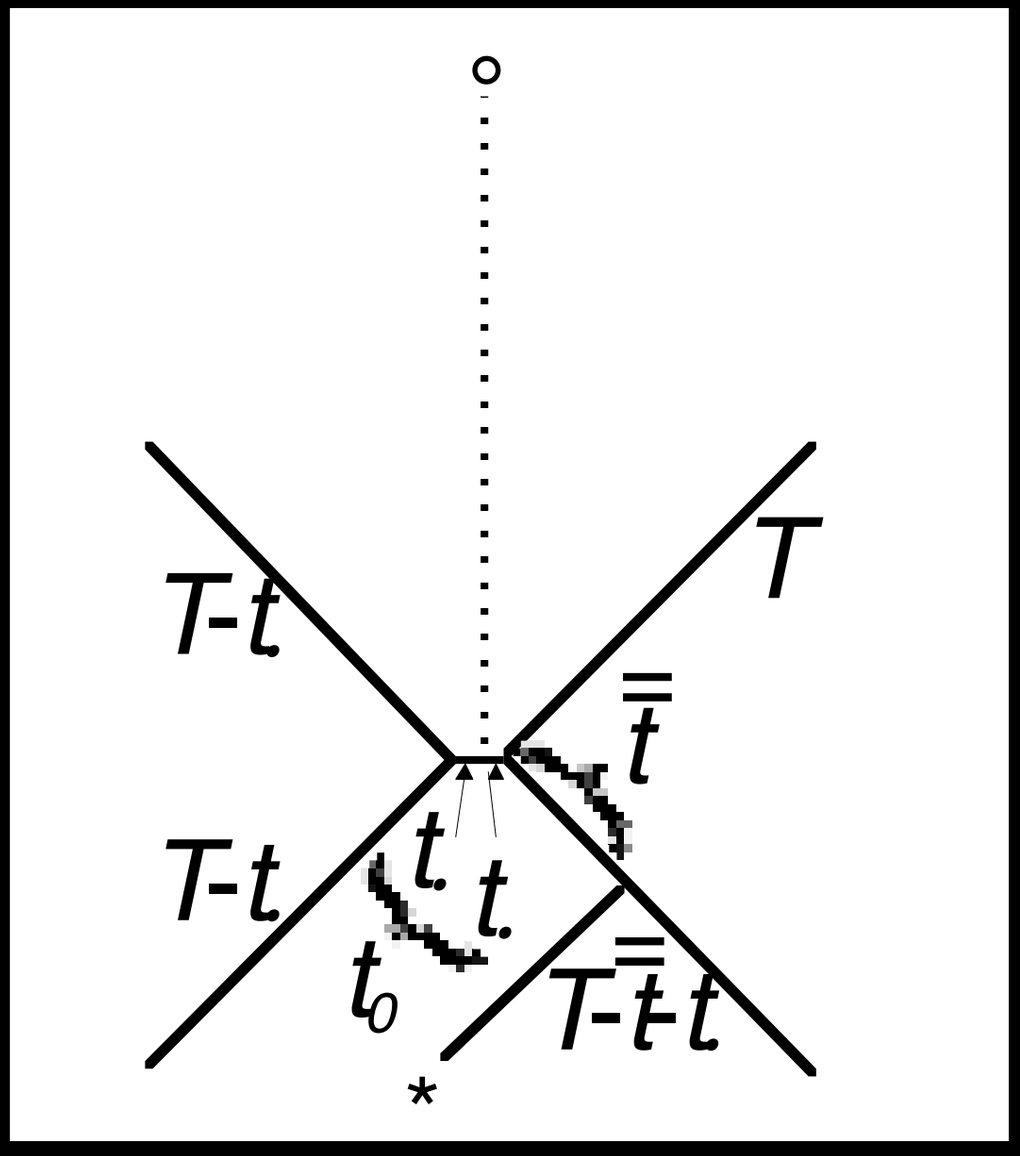


Dornburg et al., in revision

Phylogenetic experimental design has a diffuse history

- Which types of characters are most informative? (Collins et al., 2005; Dequeiroz and Wimberger, 1993; Graybeal, 1994; Naylor and Brown, 1997; Rokas and Holland, 2000; Wiens and Servedio, 1997; Yang, 1998; Zwickl and Hillis, 2002)
- **Would increased taxonomic or character sampling be more informative?** (Graybeal, 1998; Hillis, 1998; Kim, 1996, 1998; Poe, 1998; Pollock et al., 2002; Rannala et al., 1998; Rokas and Carroll, 2005; Rosenberg and Kumar, 2001, 2003; Sullivan et al., 1999)
- Which taxa should be sampled to resolve a given phylogenetic problem? (Goldman, 1998; Huelsenbeck, 1991b; Kim, 1996, 1998; Poe, 2003)

Phylogenetic informativeness can be modified to optimize taxon sampling

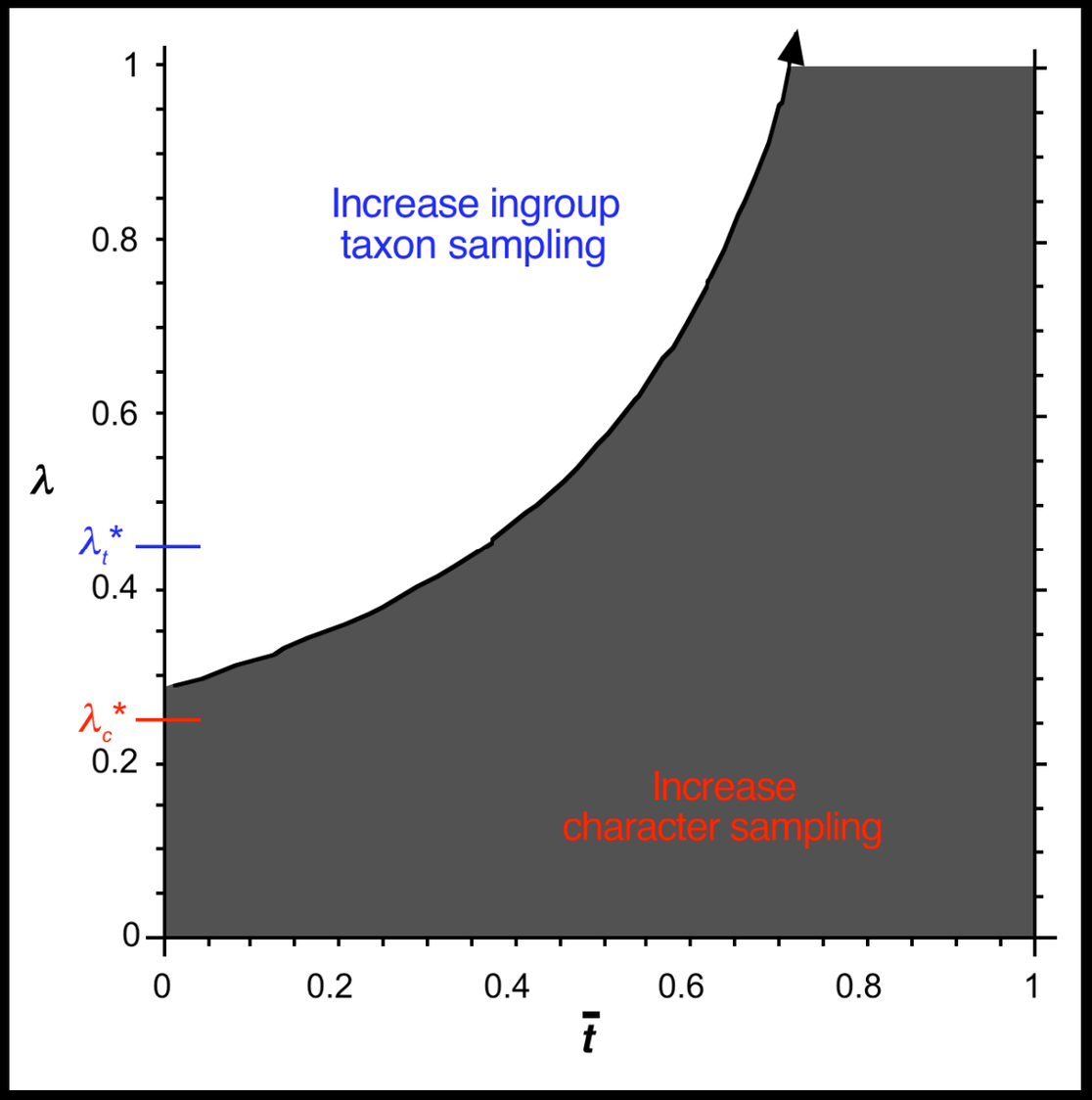


- Pr{inf}= $e^{-4T\lambda}\left(1-e^{-t_0\lambda}\right)\left(1-e^{-\left(T-\bar{\bar{t}}\right)\lambda}\right)$

$$\hat{\lambda}_t \approx \frac{1}{(2.2)T}$$

Townsend & Lopez-Giraldez, *Systematic Biology* 59: 446-457

Increase taxon sampling with fast-evolving characters, increase character sampling with slow-evolving characters



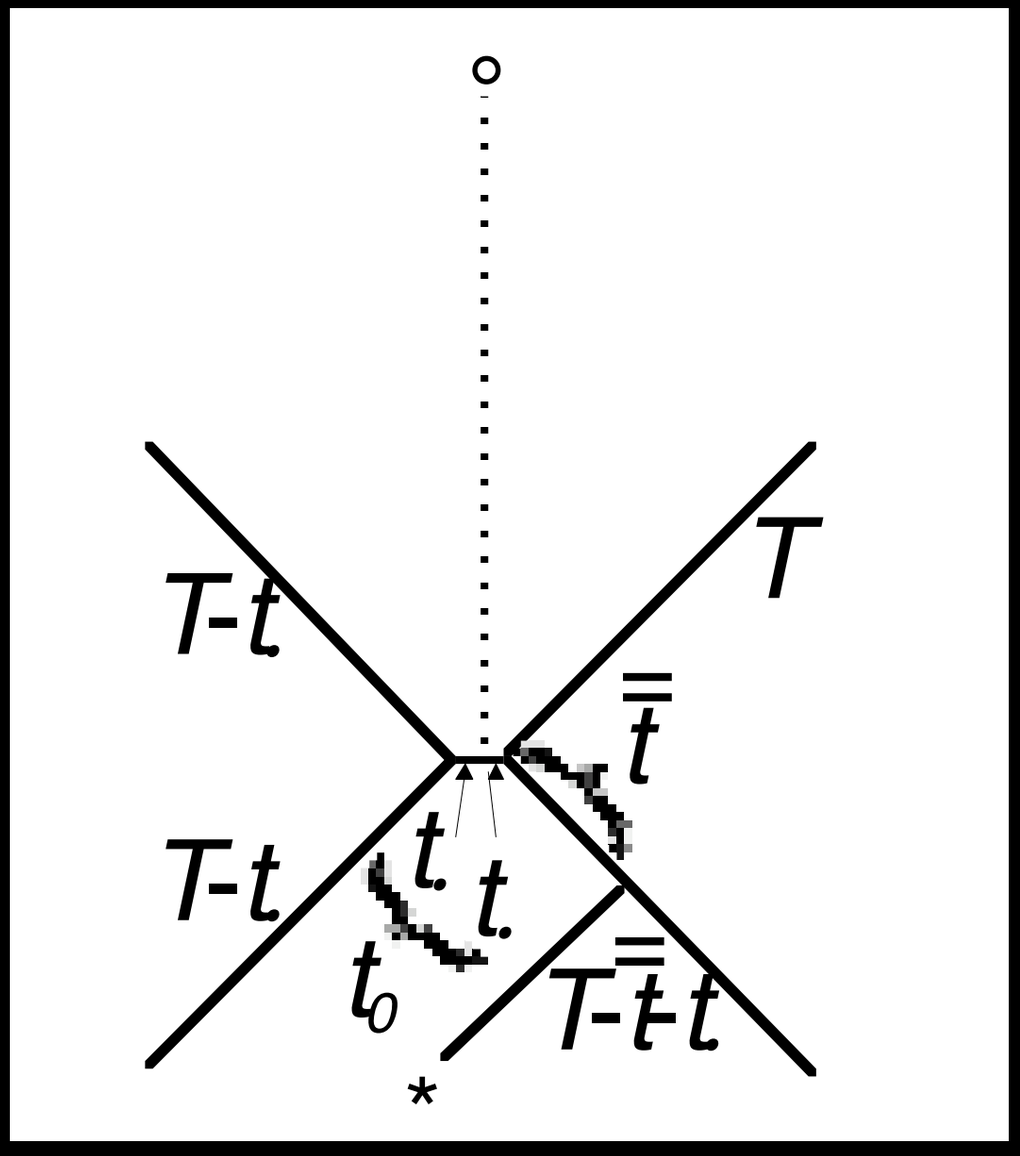
- Optimal choice of taxon and gene depends on rate of character evolution and ingroup divergence time

Townsend & Lopez-Giraldez, *Systematic Biology* 59: 446-457

Phylogenetic experimental design has a diffuse history

- Which types of characters are most informative? (Collins et al., 2005; Dequeiroz and Wimberger, 1993; Graybeal, 1994; Naylor and Brown, 1997; Rokas and Holland, 2000; Wiens and Servedio, 1997; Yang, 1998; Zwickl and Hillis, 2002)
- Would increased taxonomic or character sampling be more informative? (Graybeal, 1998; Hillis, 1998; Kim, 1996, 1998; Poe, 1998; Pollock et al., 2002; Rannala et al., 1998; Rokas and Carroll, 2005; Rosenberg and Kumar, 2001, 2003; Sullivan et al., 1999)
- **Which taxa should be sampled to resolve a given phylogenetic problem?** (Goldman, 1998; Huelsenbeck, 1991b; Kim, 1996, 1998; Poe, 2003)

Phylogenetic informativeness can be modified to optimize taxon sampling

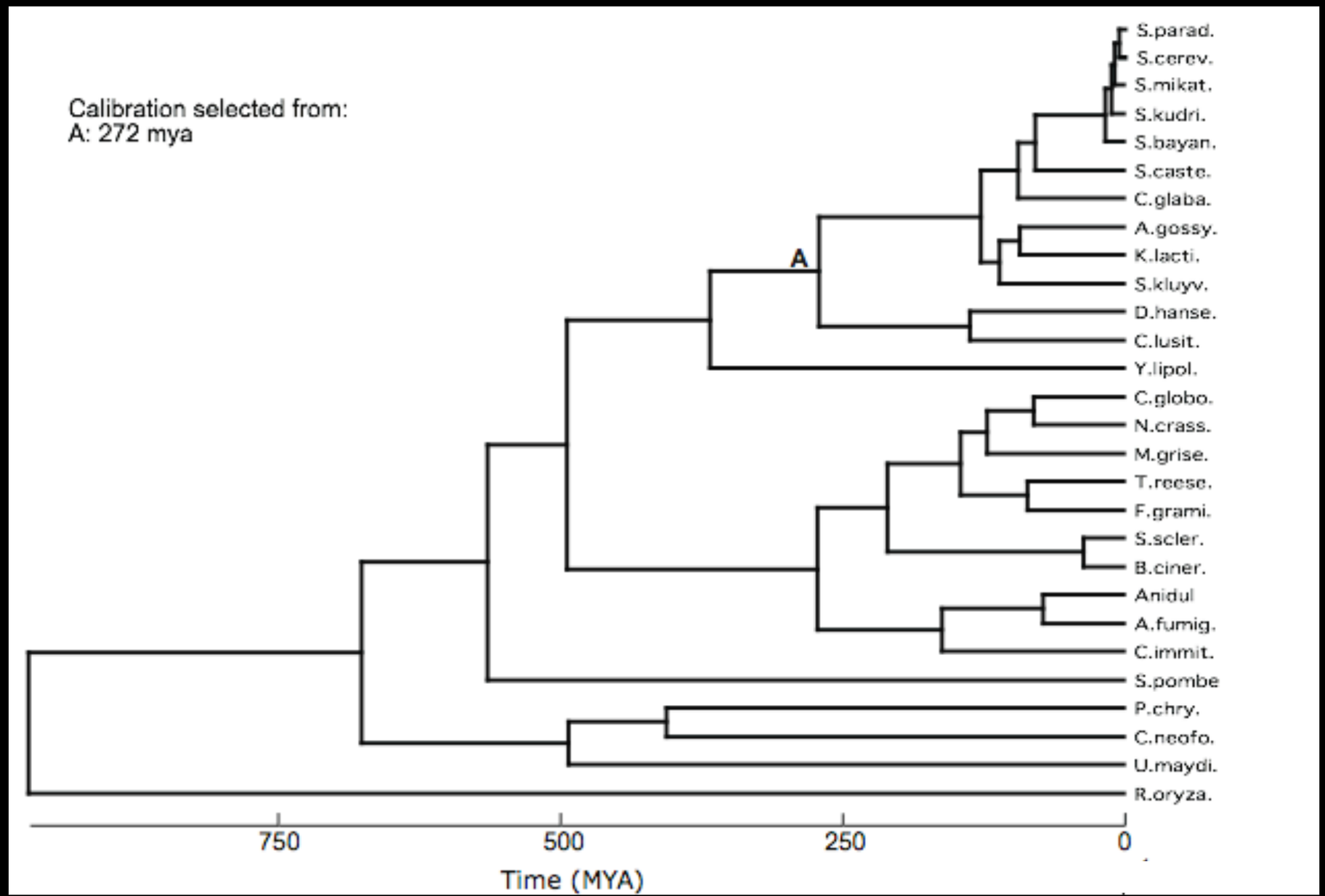


- Pr{inf}= $e^{-4T\lambda}\left(1-e^{-t_0\lambda}\right)\left(1-e^{-\left(T-\bar{\bar{t}}\right)\lambda}\right)$

$$\hat{\lambda}_t \approx \frac{1}{(2.2)T}$$

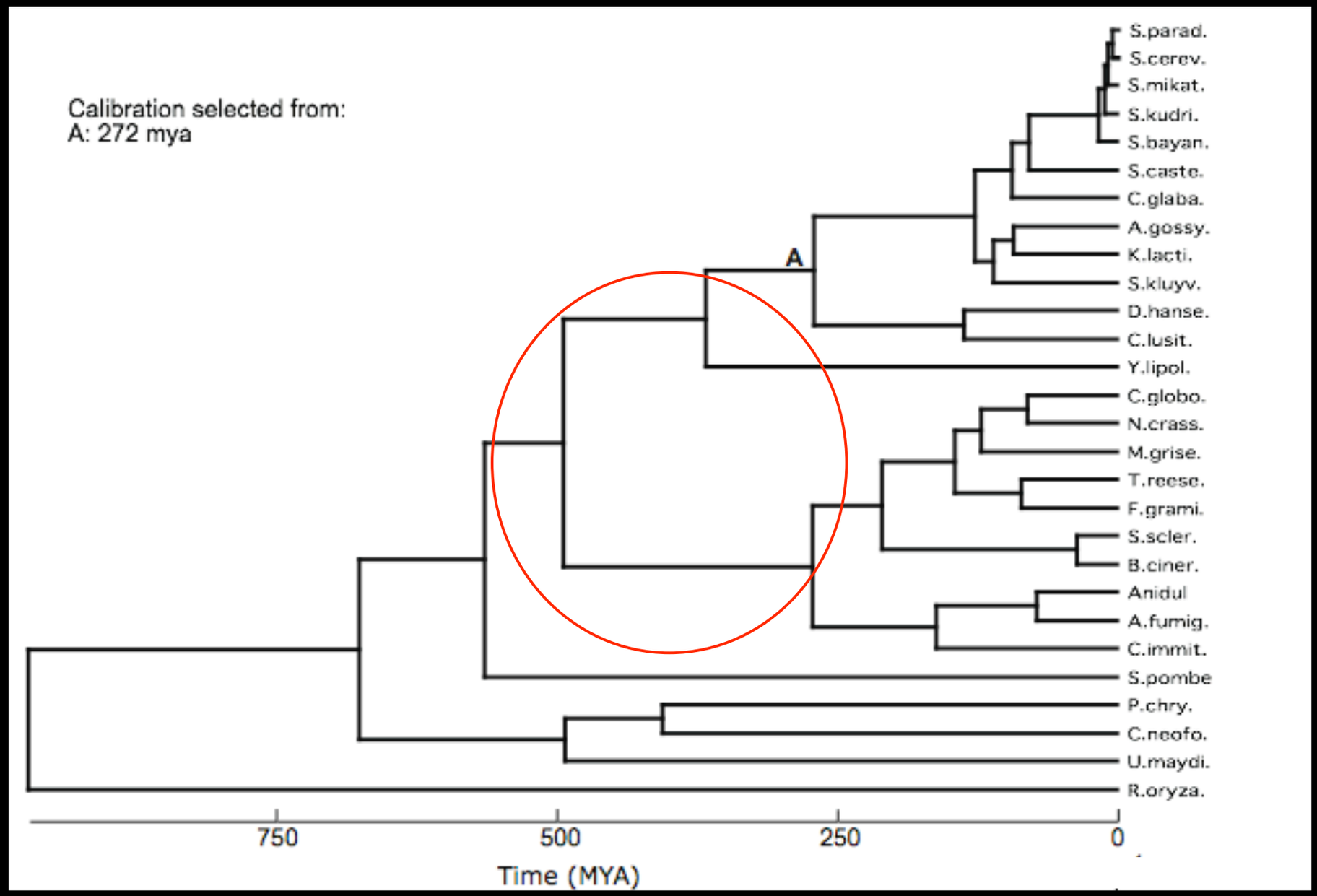
Townsend & Lopez-Giraldez, *Systematic Biology* 59: 446-457

Four taxa and their ingroups were used to evaluate PITA



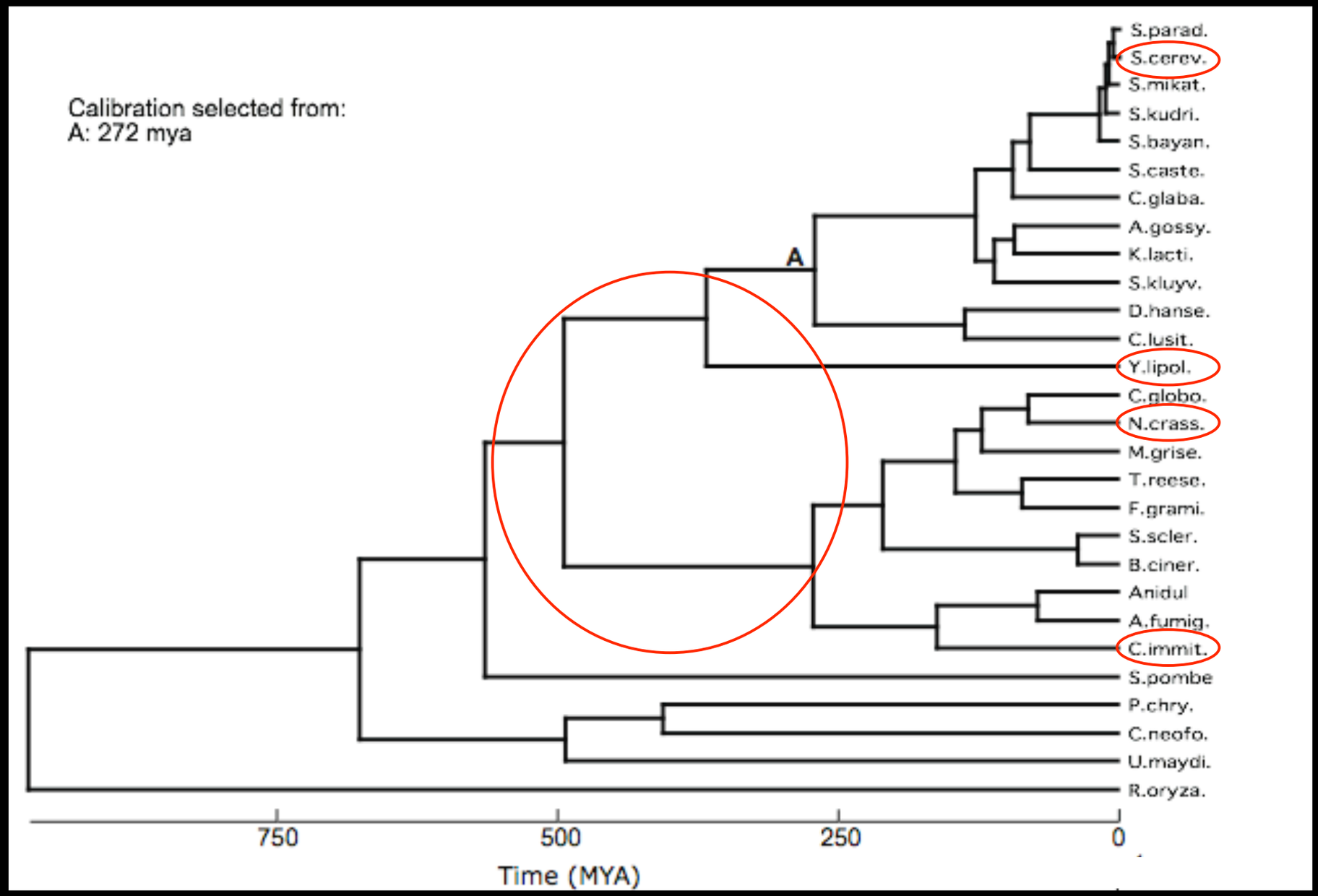
Townsend & Lopez-Giraldez, *Systematic Biology* 59: 446-457

Four taxa and their ingroups were used to evaluate PITA



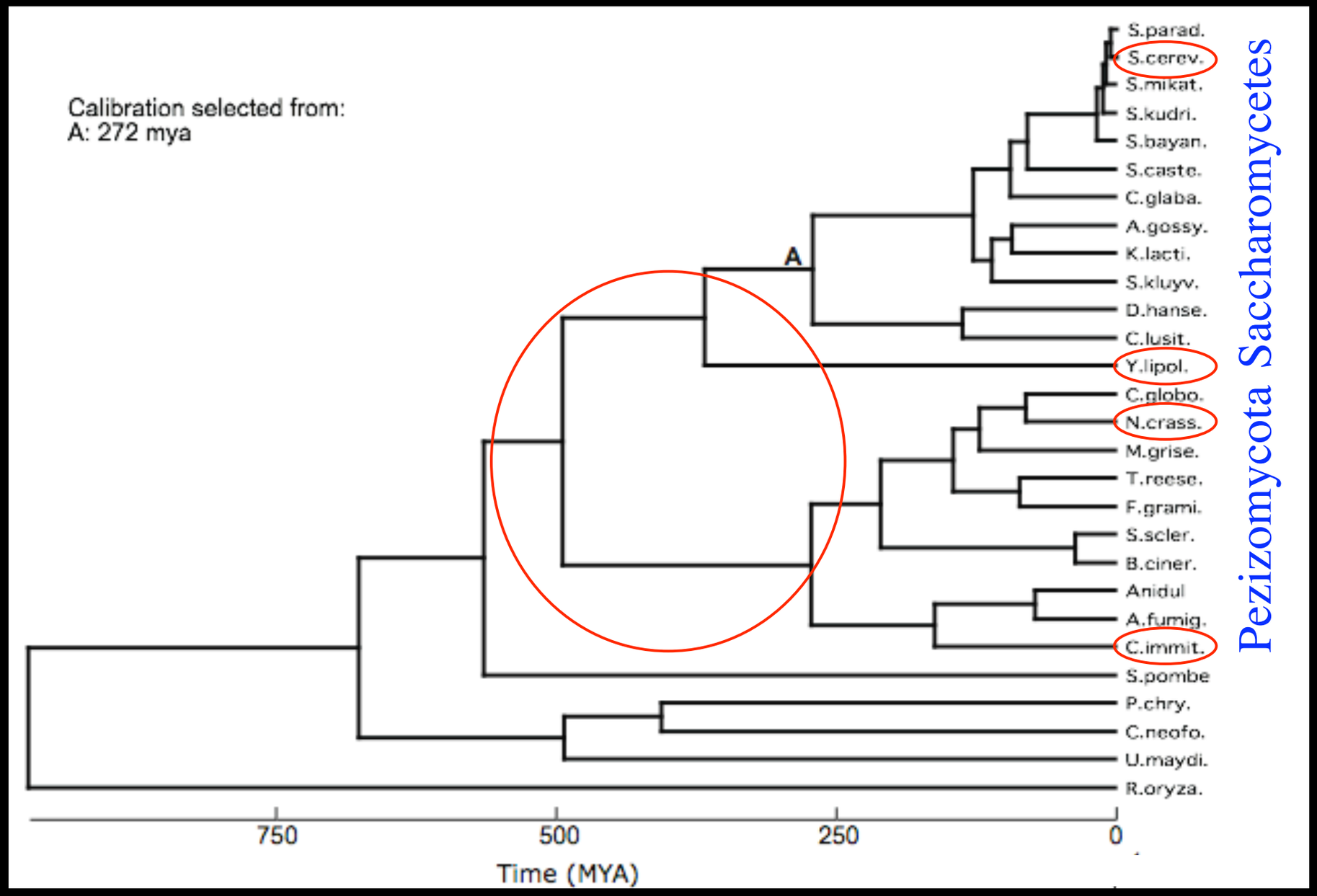
Townsend & Lopez-Giraldez, *Systematic Biology* 59: 446-457

Four taxa and their ingroups were used to evaluate PITA



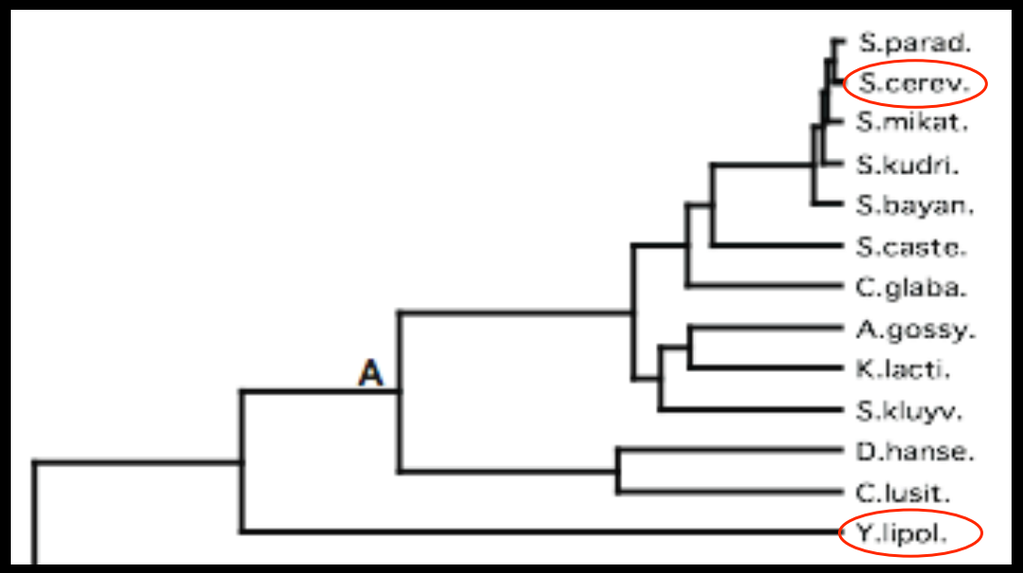
Townsend & Lopez-Giraldez, *Systematic Biology* 59: 446-457

Four taxa and their ingroups were used to evaluate PITA



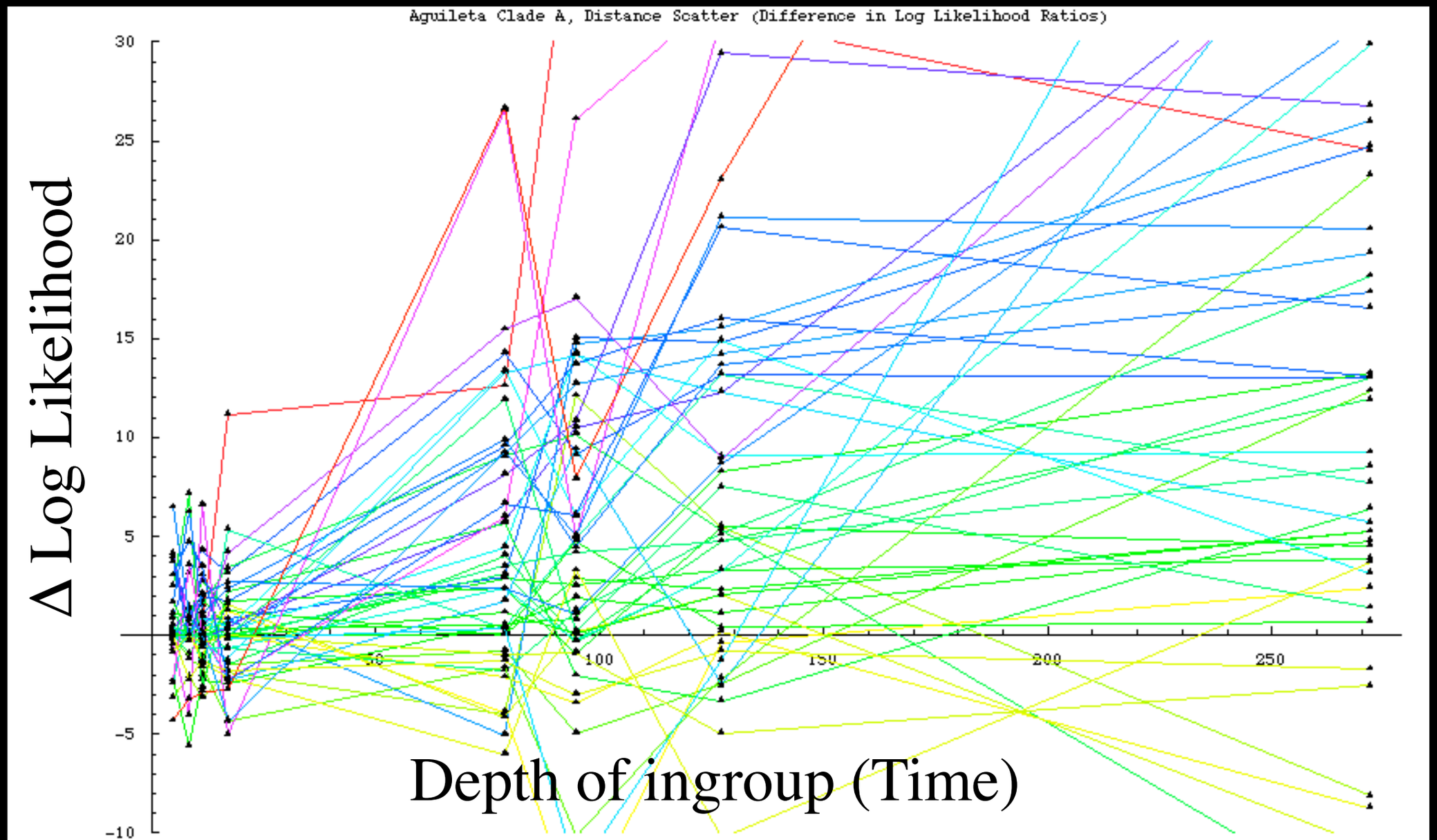
Townsend & Lopez-Giraldez, *Systematic Biology* 59: 446-457

The Saccharomycetes chronogram features



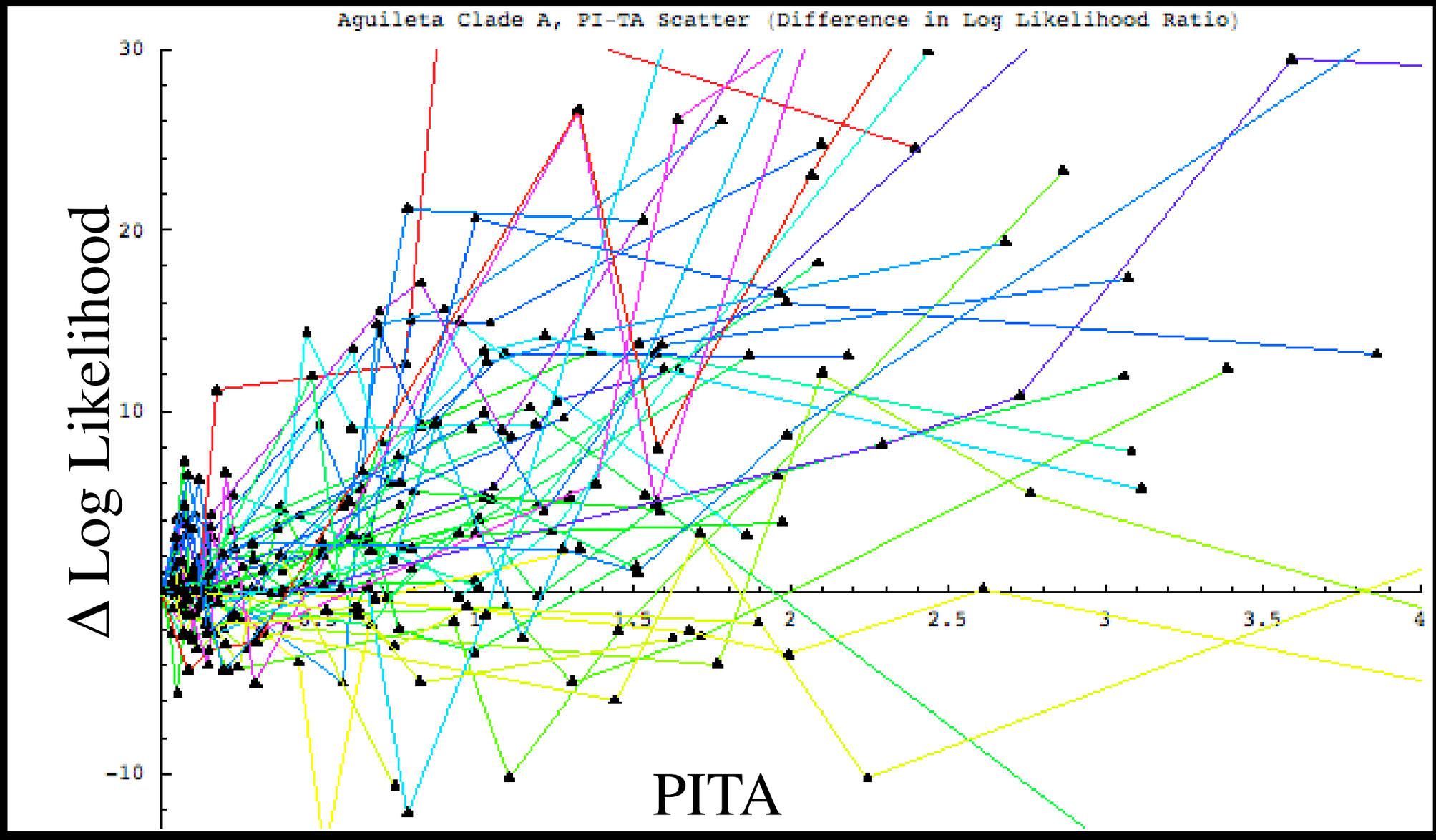
- four tip lineages in the genus Saccharomyces, closely related to *S. cerevisiae*, and
- a range of ingroup taxa at other depths

Sampling the deepest ingroups typically yields the greatest resolution



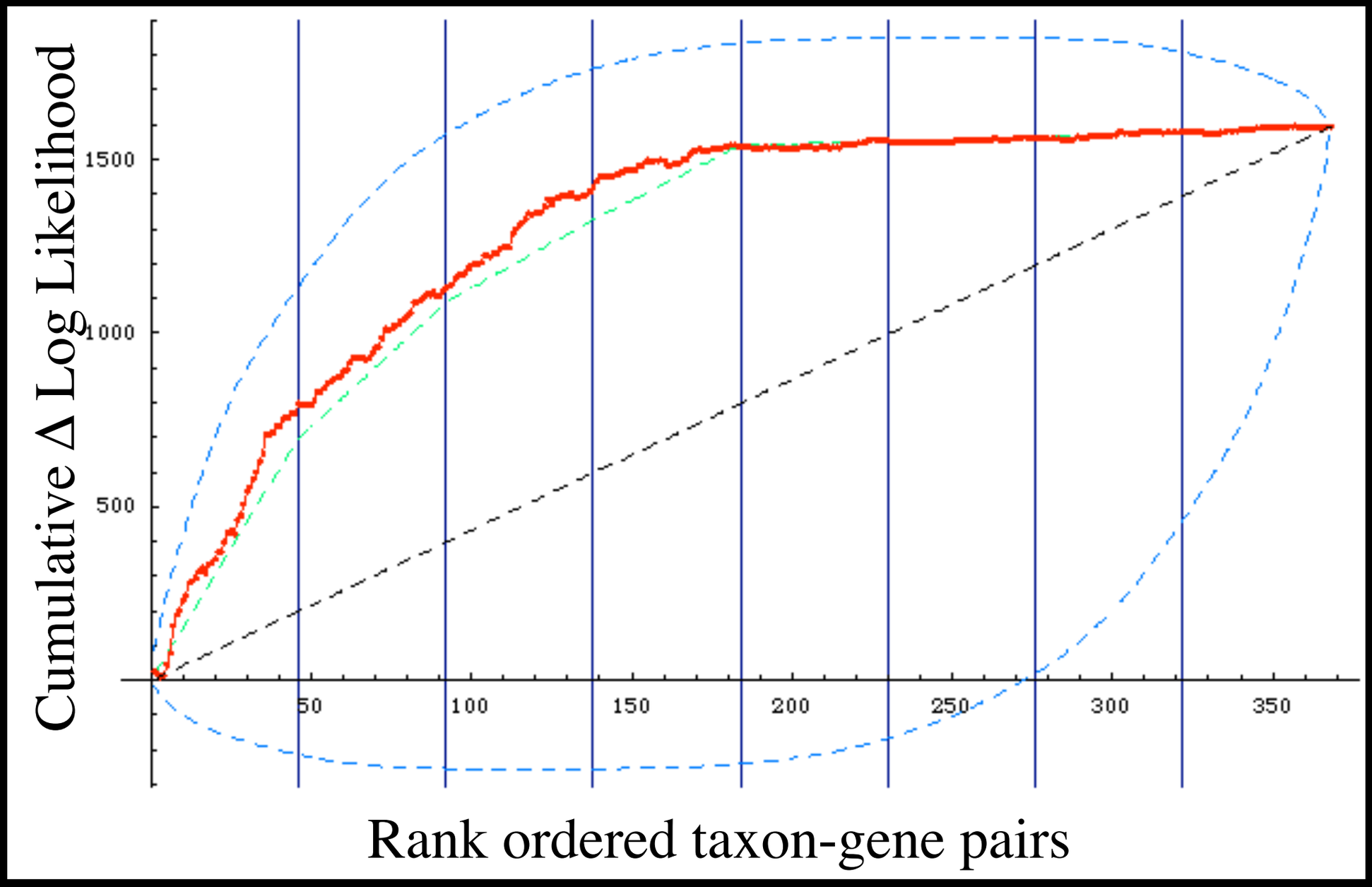
Townsend & Lopez-Giraldez, *Systematic Biology* 59: 446-457

Sampling using PITA is correlated with support



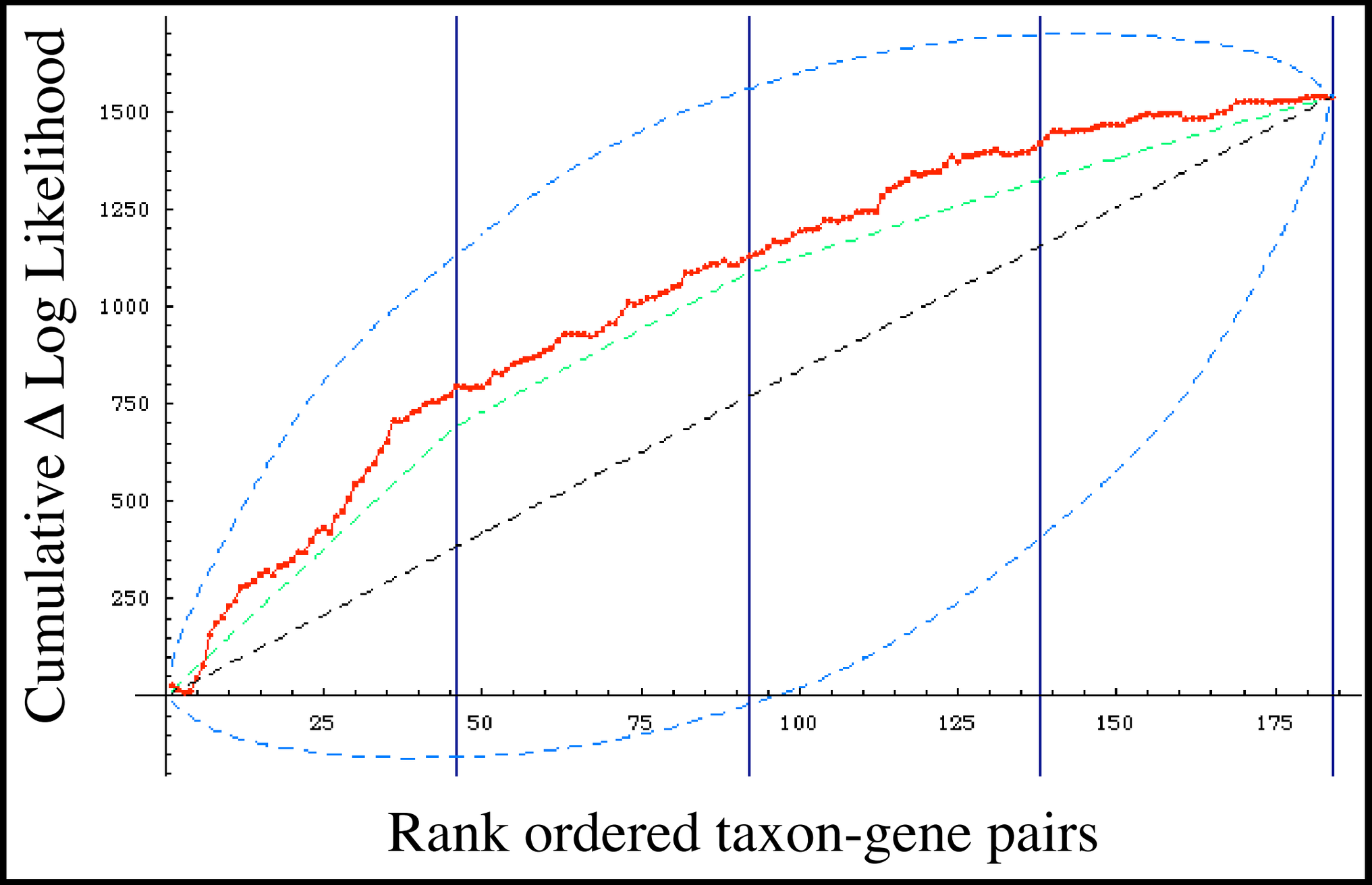
Townsend & Lopez-Giraldez, *Systematic Biology* 59: 446-457

Saccharomycete clade: PITA performs best



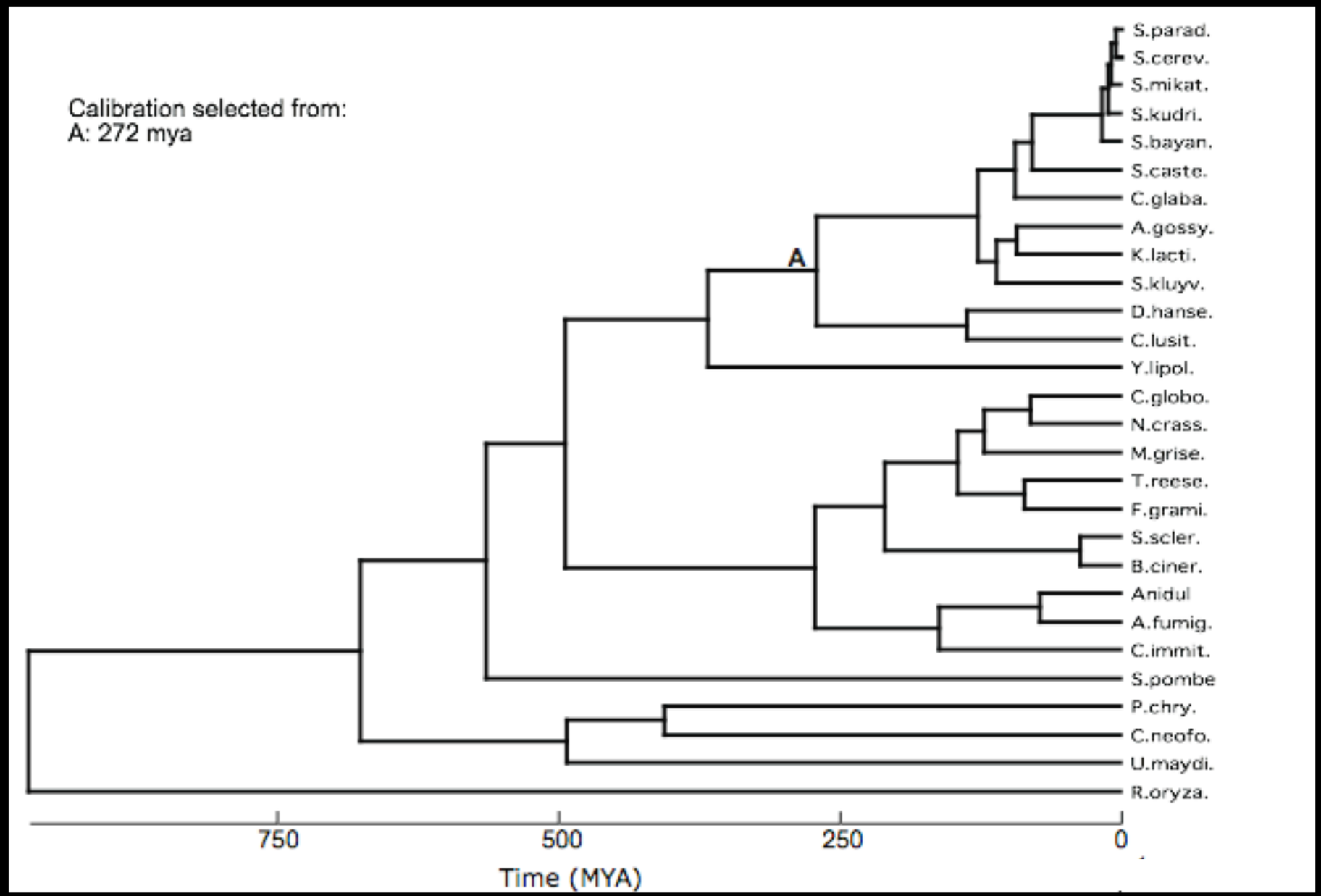
Townsend & Lopez-Giraldez, *Systematic Biology* 59: 446-457

Deep Saccharomycete clade: PITA performs best

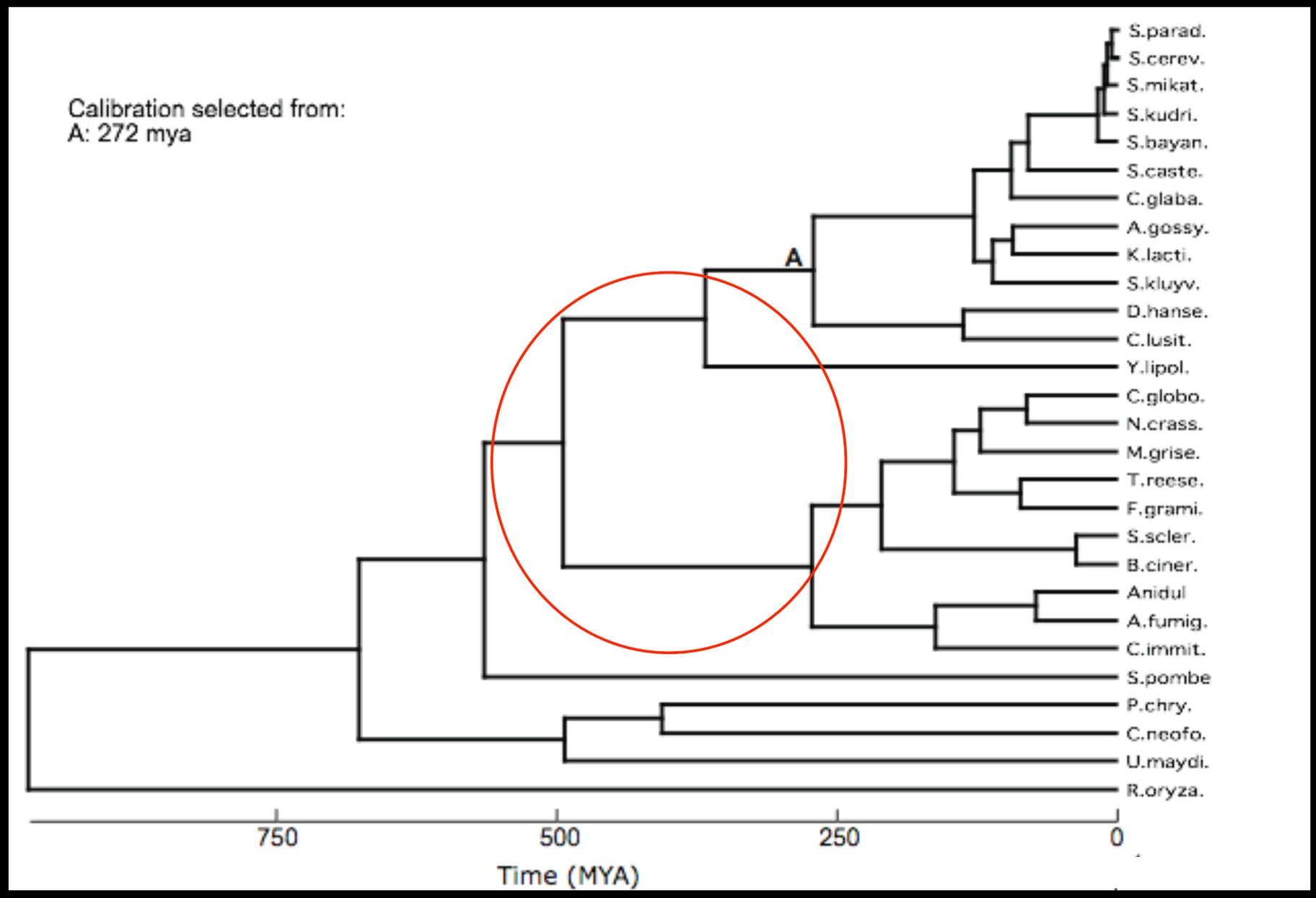


Townsend & Lopez-Giraldez, *Systematic Biology* 59: 446-457

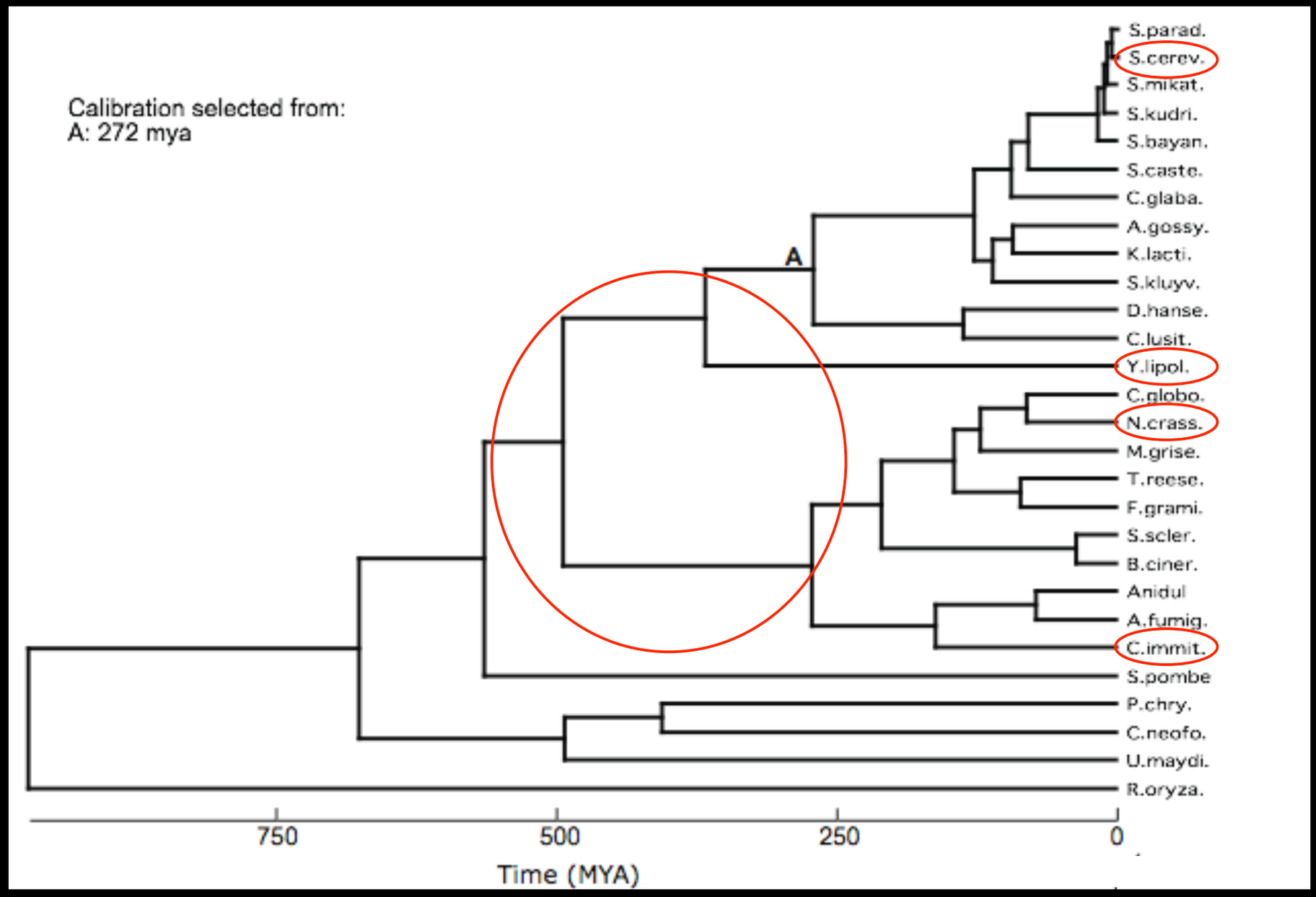
Does the other ingroup clade show the same patterns?



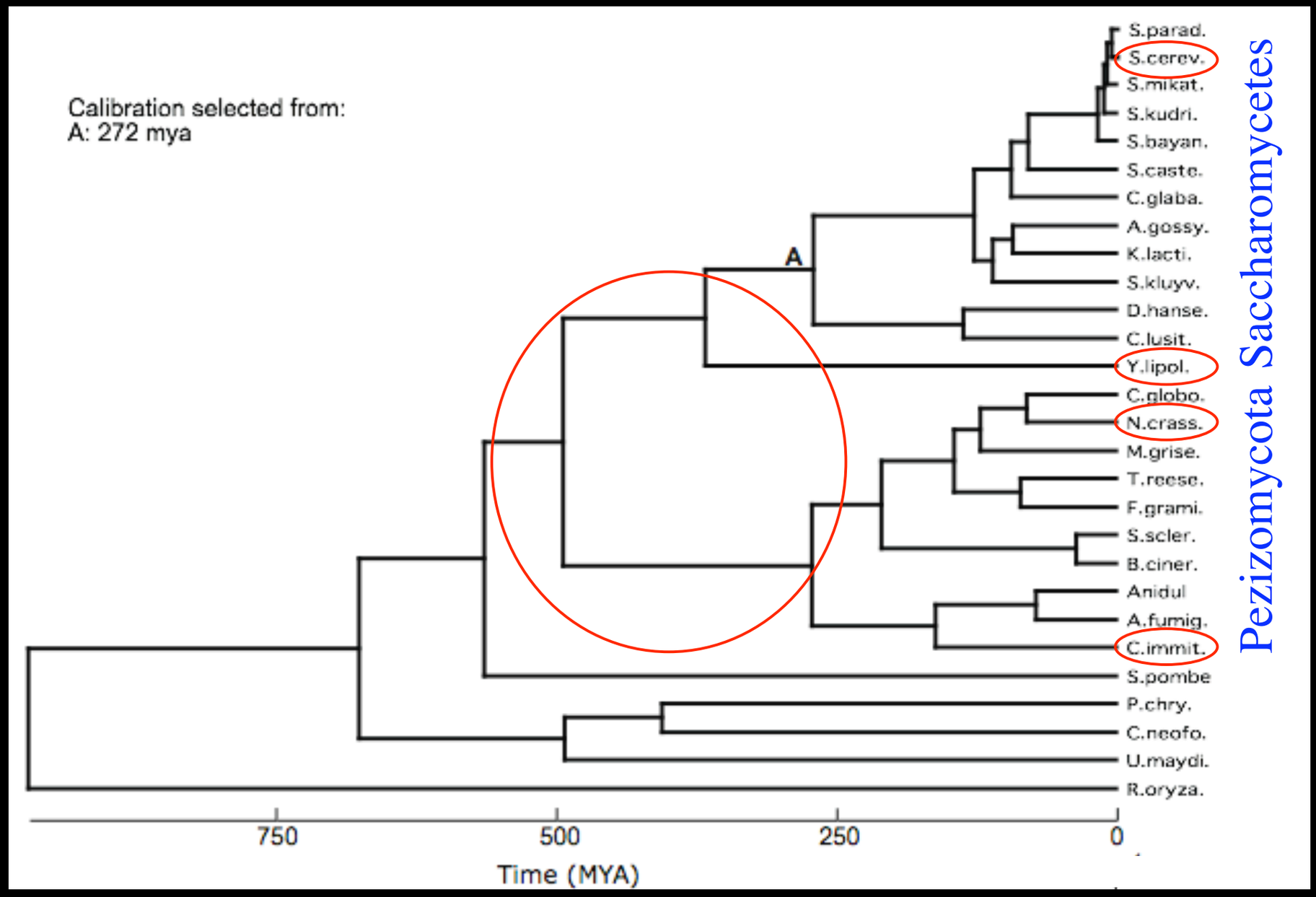
Does the other ingroup clade show the same patterns?



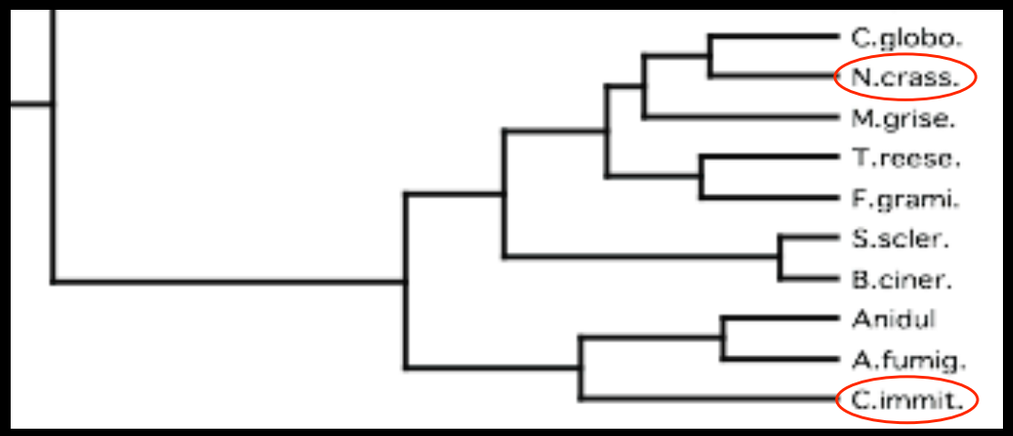
Does the other ingroup clade show the same patterns?



Does the other ingroup clade show the same patterns?

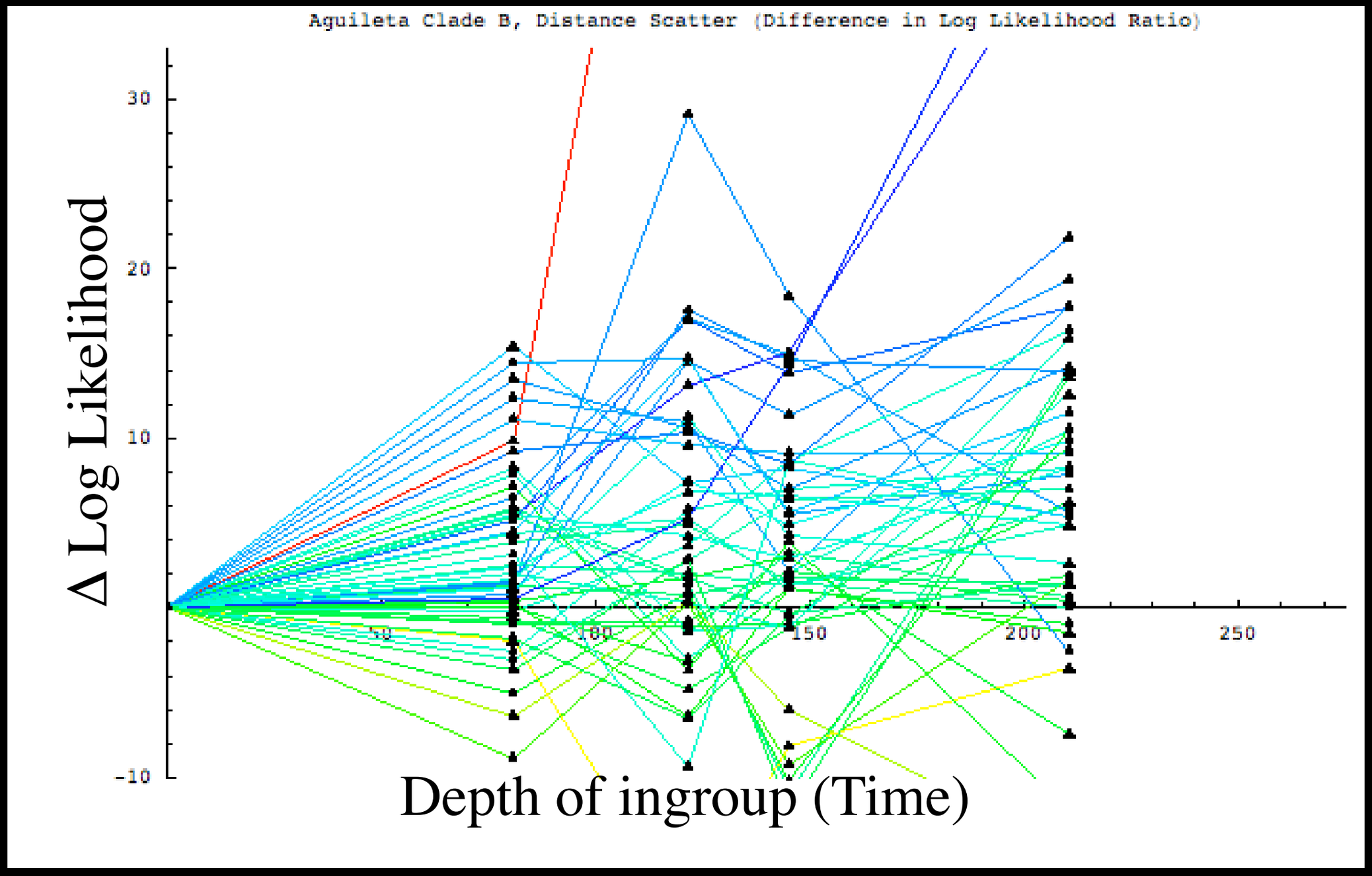


The Pezizomycetes chronogram features



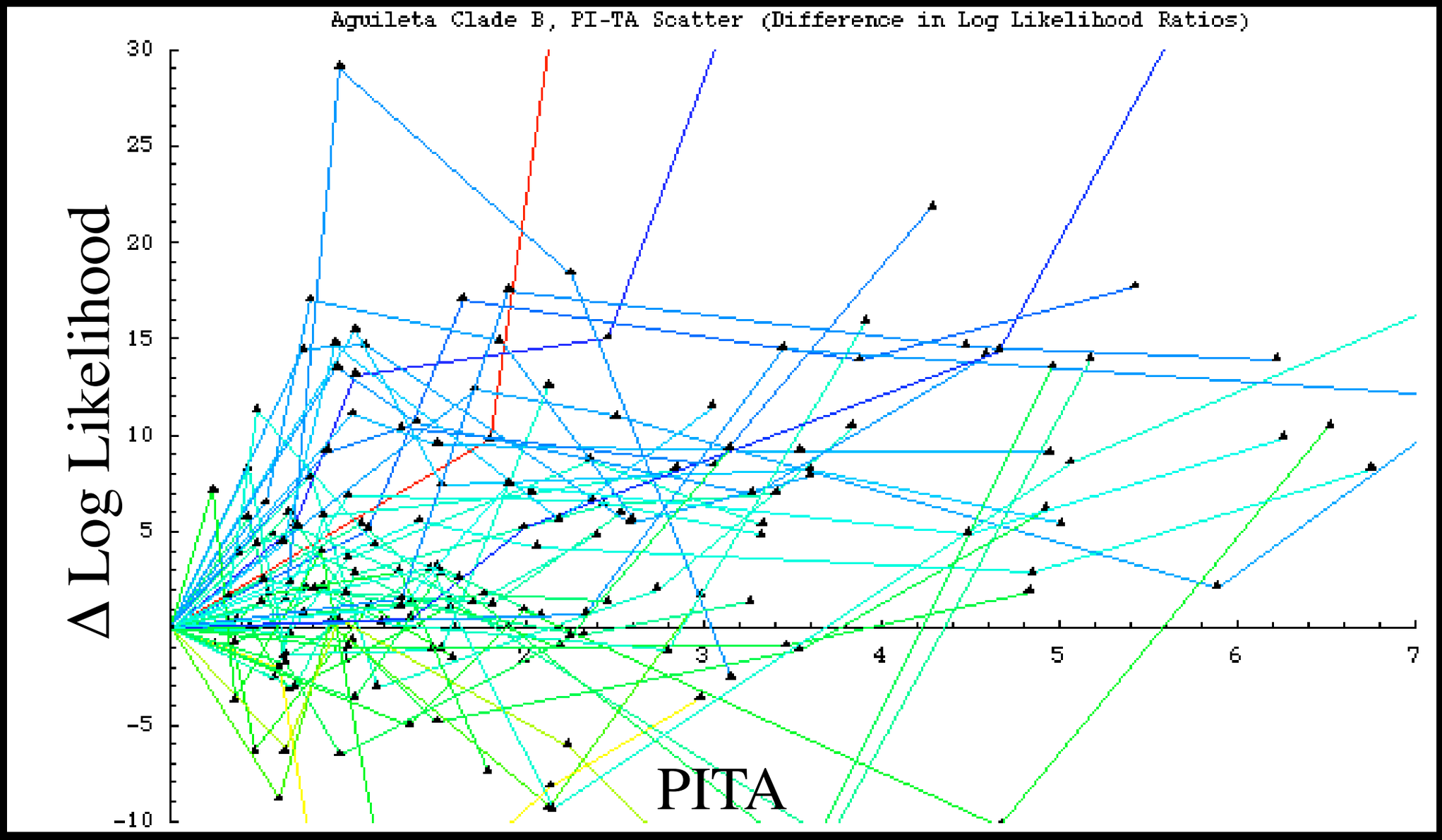
- no lineages closely related to *N. crassa* or *C. immitis*
- a smaller range of depths of ingroup taxa than in the Saccharomycetes

Sampling the deepest ingroups yields the greatest resolution



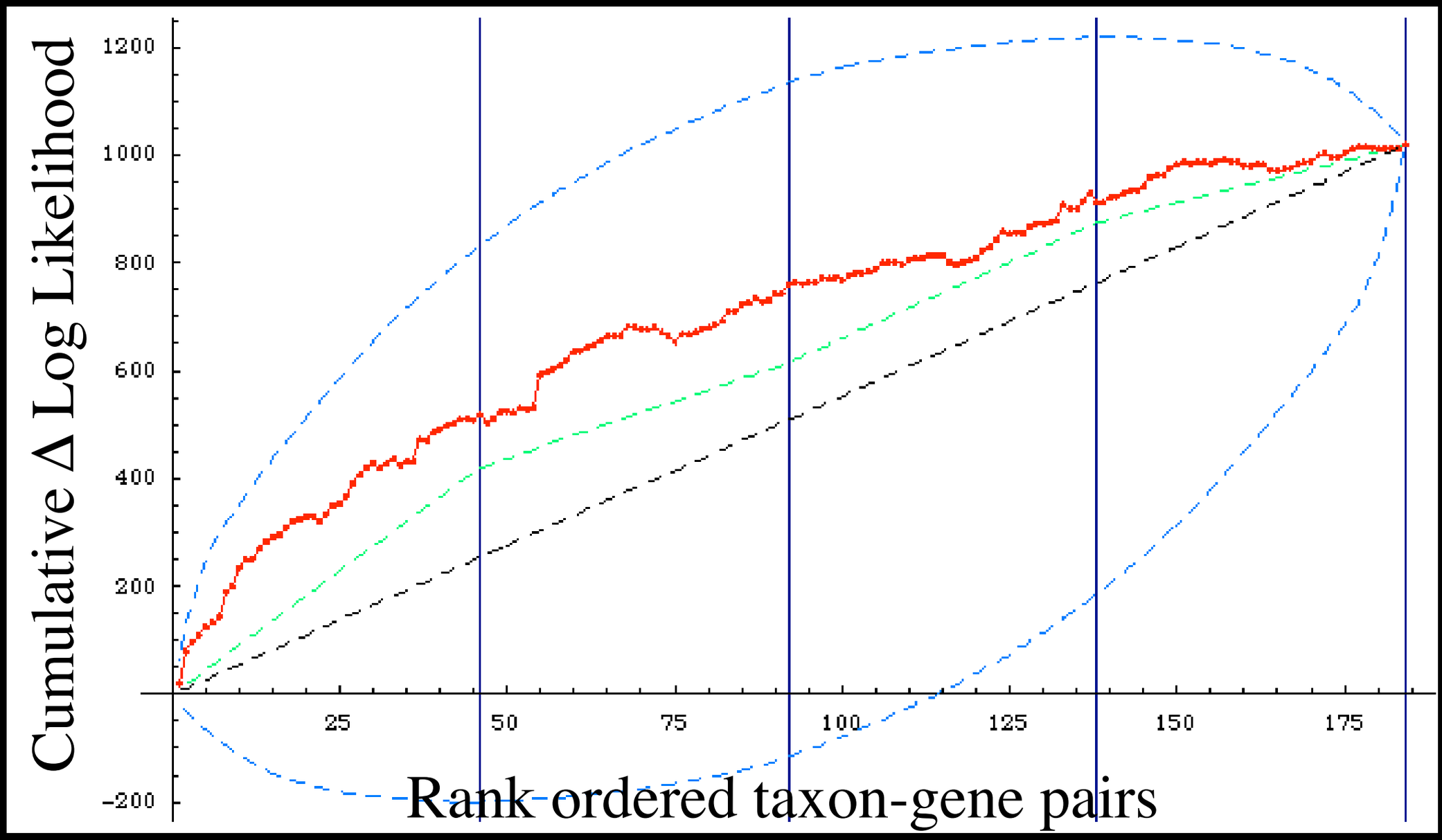
Townsend & Lopez-Giraldez, *Systematic Biology* 59: 446-457

PITA is correlated with support



Townsend & Lopez-Giraldez, *Systematic Biology* 59: 446-457

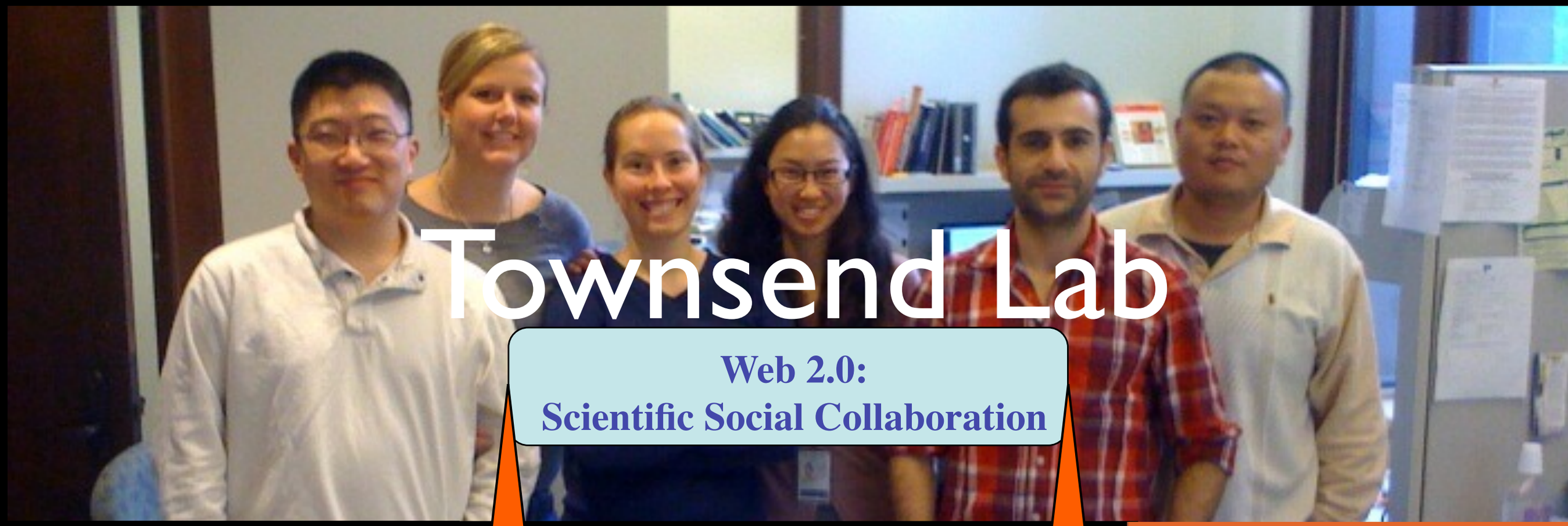
Pezizomycete clade: PITA performs best



Townsend & Lopez-Giraldez, *Systematic Biology* 59: 446-457

Conclusions

- Large data sets contribute both signal and noise.
- Phylogenetic informativeness profiles can help you prioritize loci - but remember the effect of noise.
- With fast-evolving characters, increase taxa.
- With slow-evolving characters, increase characters.
- The most informative taxa to sample to resolve a node are the deepest ingroups to the node.

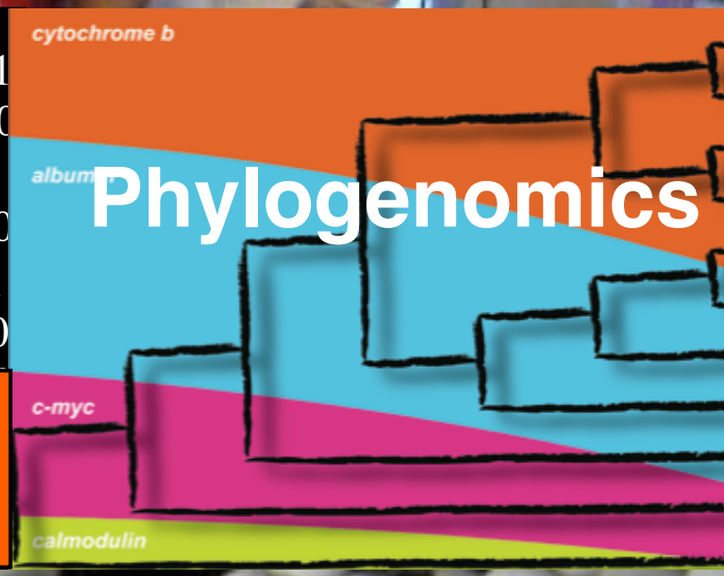


Townsend Lab

Web 2.0:
Scientific Social Collaboration

Transcriptomics

Computational and statistical analysis of large datasets



Phylogenomics

Fungal Biology and Phylogeny