

# Temporal calibration of molecular phylogenies

- Chris Yesson -



Institute of Zoology

LIVING CONSERVATION

# Molecular Clocks: Background

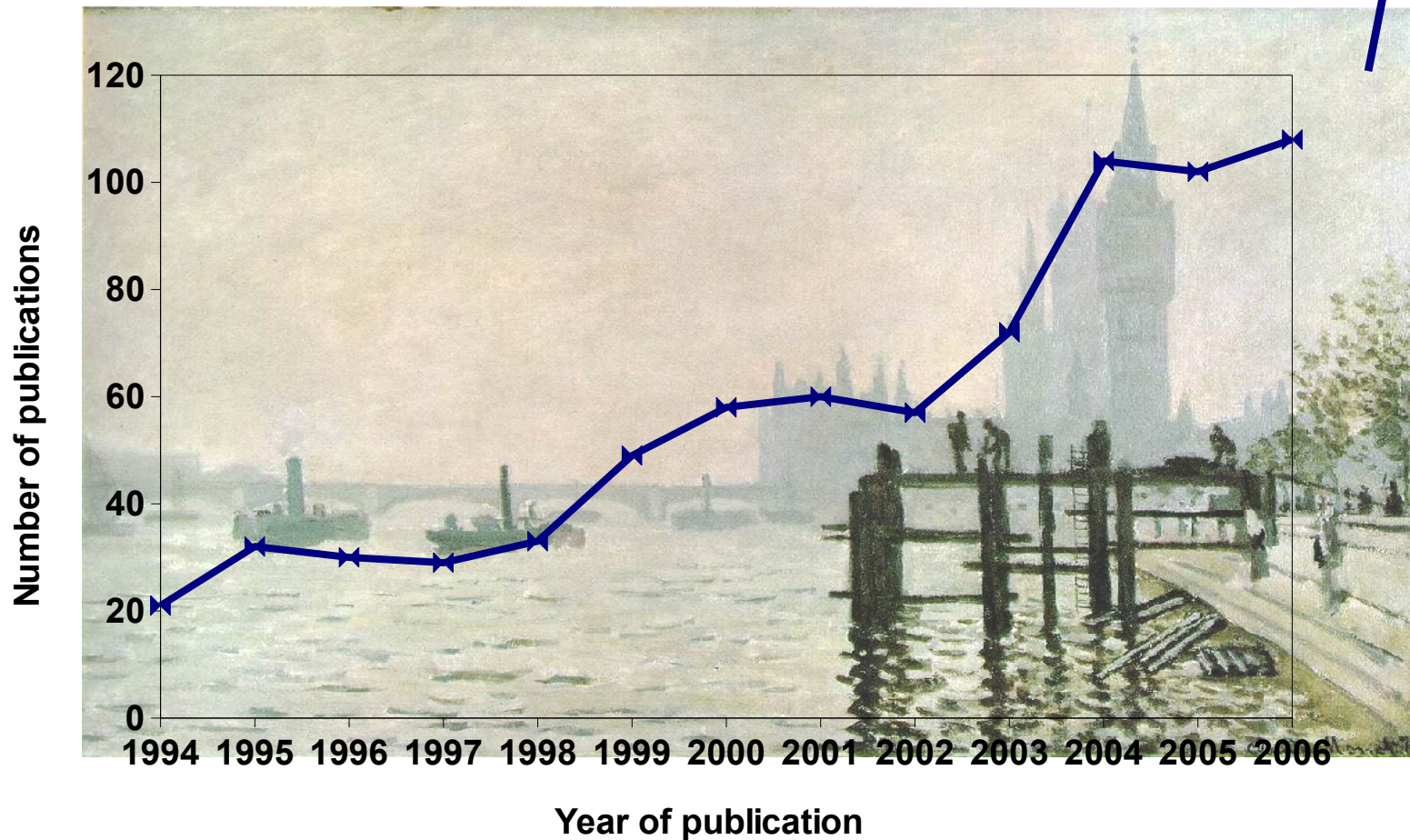


- Molecular clock Zuckerkandl and Pauling 1962, 1965
- Kimura's neutral theory (1983)
  - Molecular substitutions are “random”
  - Constant accumulation of mutations
  - Constant among lineages
- Controversial... But popular

# The dating of phylogenies is popular

- searching web of science for  
"phylogeny and (molecular clock or dating)"

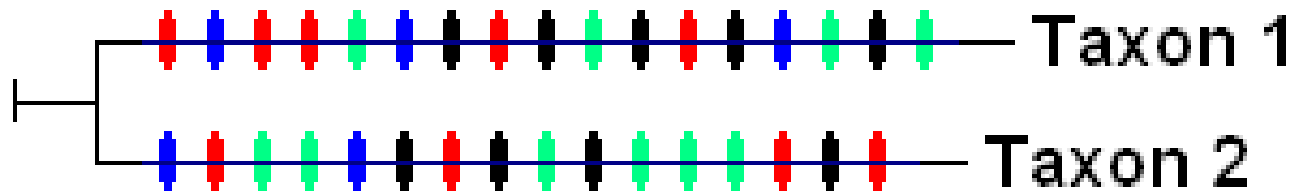
ZSL



# Molecular Clocks



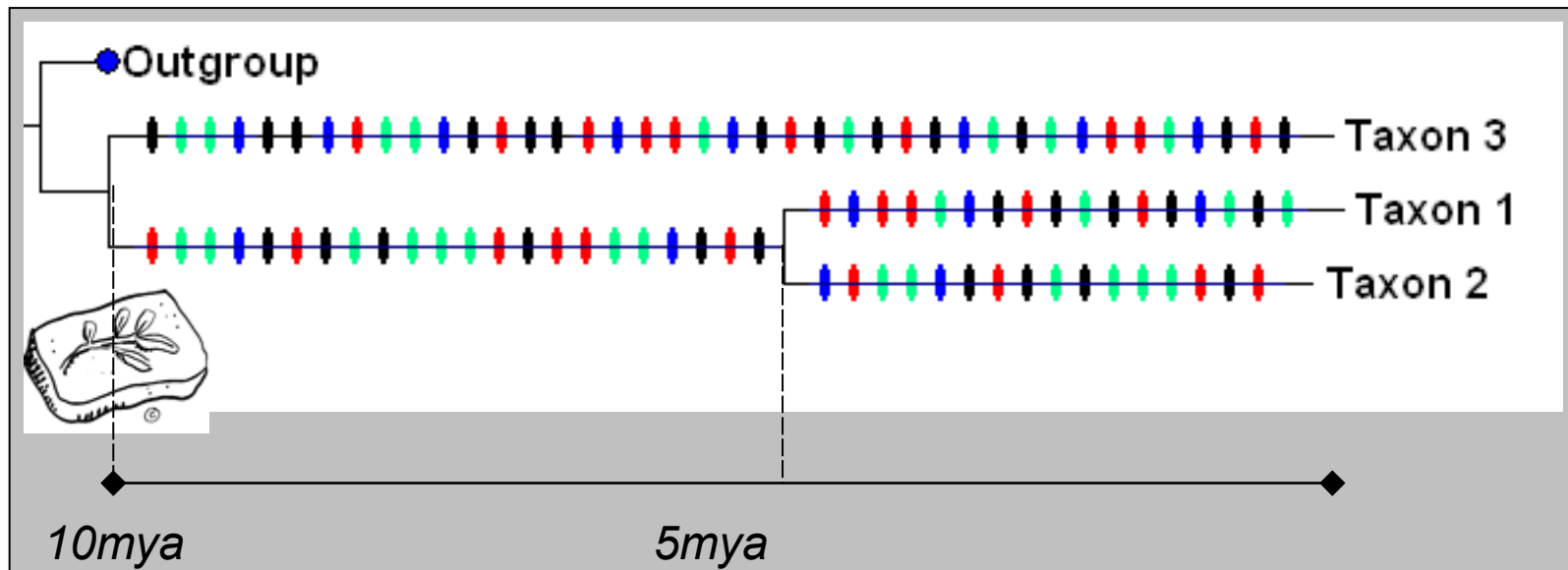
- Random mutations/substitutions implies “same” number of mutations for sister taxa



# Molecular Clocks



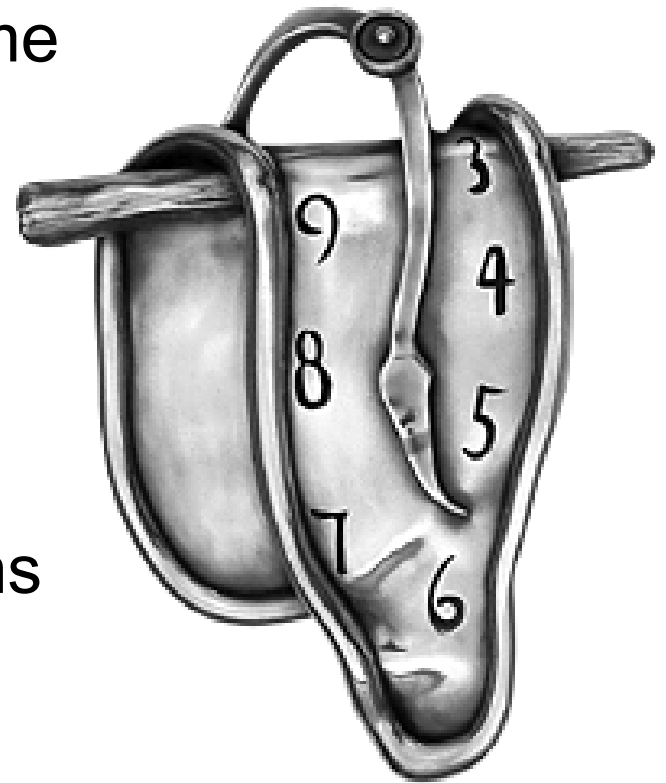
- Fix one age on the tree = fix all ages on the tree



# Problems with the molecular clock...?



- Inconsistency of evolutionary rates
  - Some genes are clocklike, some are not
- Long-term stability vs. short-term rate fluctuations
- Test your data to see if it conforms to the molecular clock!



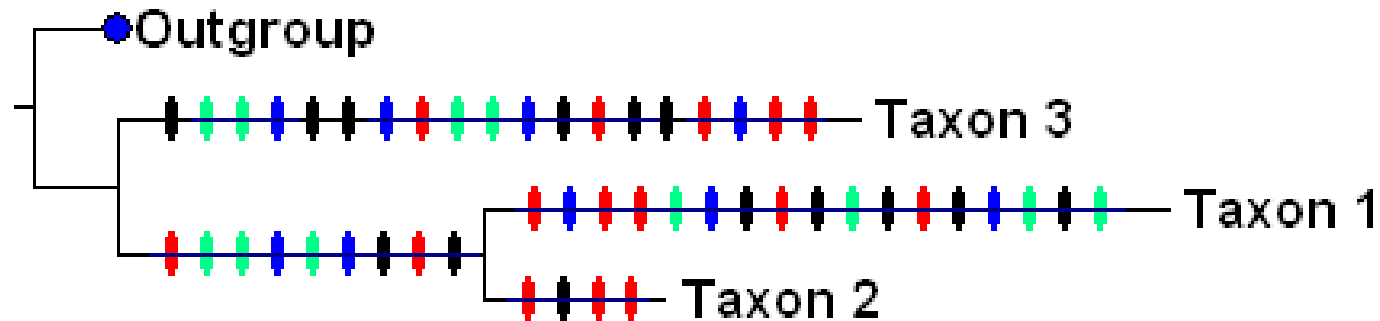
## Testing for a molecular clock using a Likelihood Ratio Test\*



- Choose a simple likelihood model (i.e. HKY85)
- Find a tree close to the 'best' topology (NJ tree)
- Calculate the difference between likelihood scores with and without a molecular clock enforced (LR)
- Find  $\Lambda = |2 \times \text{LR}|$
- Compare  $\Lambda$  to Chi-squared distribution (n-2 df, where n = number of taxa in tree)
- Reject the clock if  $p < 0.05$

\* Felsenstein, J. (1988) *Annual Review of Genetics*. **22**:521-565

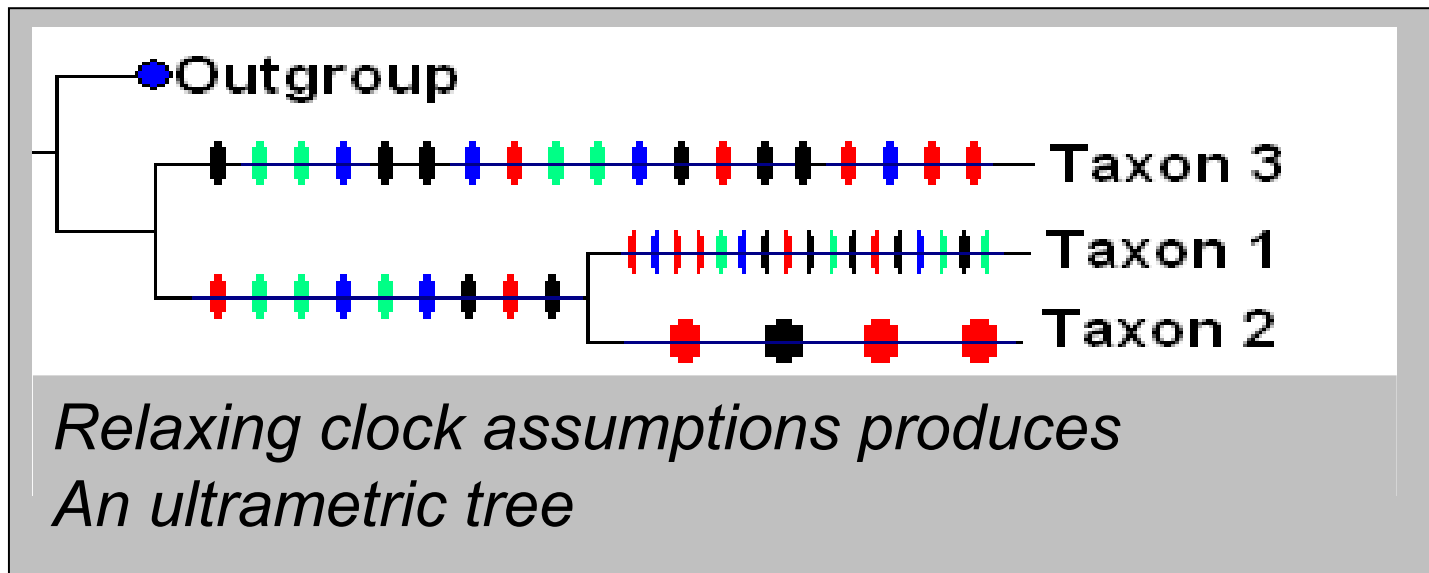
Your data may not conform to the molecular clock



*Data fails the clock test*



Your data may not conform to the molecular clock



## Methods that don't assume a clock



- Local Clock Methods
- Smoothing methods
  - NPRS
  - Penalized Likelihood
- Bayesian Methods

Simple



Complex

*Reviews:*

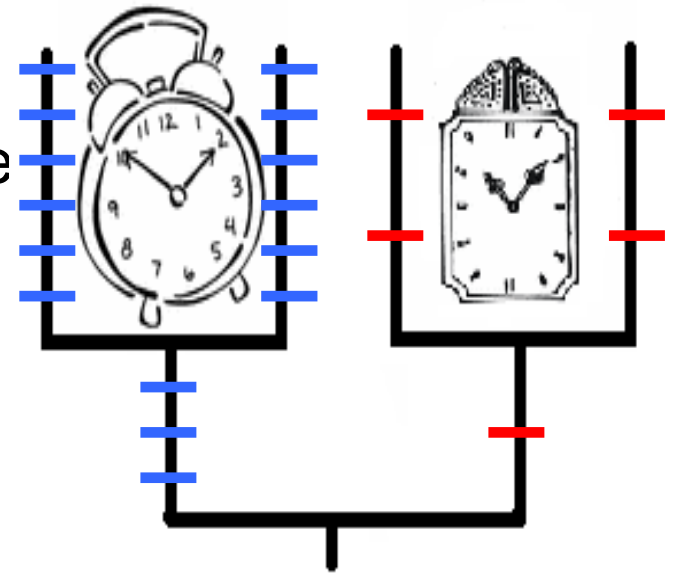
*Sanderson et al. (2004) American Journal of Botany. 91, 1656-1665*

*Rutschmann, F. 2006 Diversity and Distributions 12:35-48.*

# Local clock methods



- Assumes that parts of tree conform to different clocks
- Difficult choosing how to partition the tree
- No dedicated software?



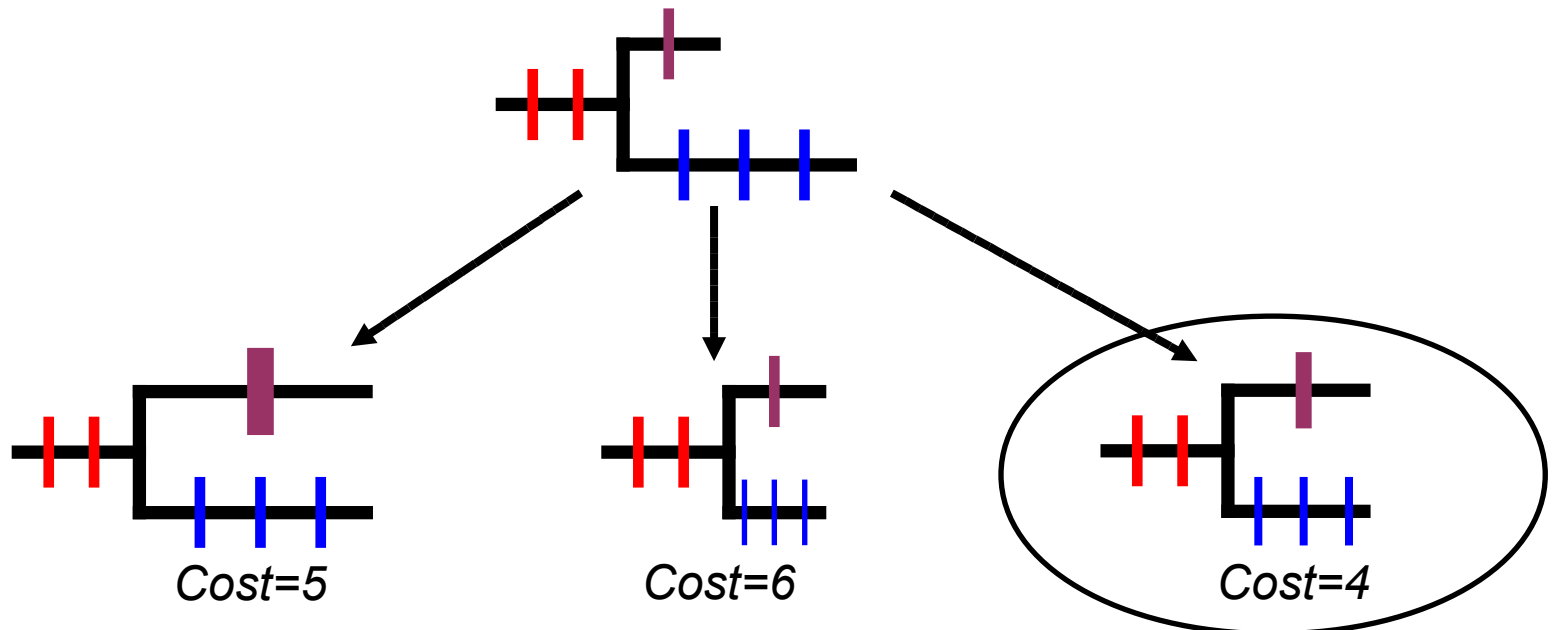
Yoder, A. D., & Yang, Z. 2000. *Molecular Biology and Evolution* 17:1081–1090.

Hasegawa, M., Kishino, H. & Yano, T. (1989). *Journal of Human Evolution* 18: 461–476.

# Smoothing Methods



- Each branch has different rate
- Changes in rate carry a cost
- Optimal solution minimises rate-change costs



- Least squares smoothing of substitution rates
- Popular
  - “Easy” to use
  - No parameters
- Widely criticised
  - Calibration at tips is bad

- A semi-parametric smoothing method
- NPRS + Likelihood
- “Tuned” to allow greater or lesser rate variation
- Zero-length terminal branches can cause problems

Implemented in Sanderson's r8s <http://ginger.ucdavis.edu/r8s/index.html>

*Sanderson, M. J. (2002) Molecular Biology and Evolution. 19, 101-109*

- Adopt an explicit probabilistic model of how evolutionary rates change over time
- Parameter rich
- Possible to incorporate models into initial tree searching
- Difficult to select/justify parameter values
- Computationally expensive

*Implemented in Thorne's multidivtime <http://statgen.ncsu.edu/thorne/multidivtime.html>*

*Thorne, J. L., et. al. (1998) Molecular Biology and Evolution. **15**, 1647-1657*

*Huelsenbeck, J. P., et. al. (2000) Genetics. **154**, 1879-1892*

*Kishino, H., et. al. (2001) Molecular Biology and Evolution. **18**, 352-361.*

# Combining tree searching and rate smoothing



- Incorporating rate smoothing into the evolutionary model used in tree searching is a step forward
- Multiple trees are dated and can be used as replications to provide confidence intervals for dates
- Replicates have different topologies and so incorporate topological uncertainty into the dates
- Trees found incorporating rate smoothing models can be different topologies to those found without

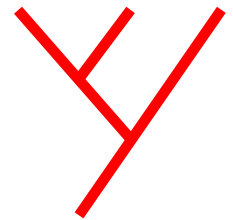


- Incorrect Phylogeny
  - Inadequate sampling?
  - Is the rooting correct?
- Rate heterogeneity
  - A major source of error... but poorly understood
  - Generation times / selection pressures?
- Orthology/Paralogy
  - Gene duplications can pre-date speciation leading to a downward bias in age estimates

# Calibration



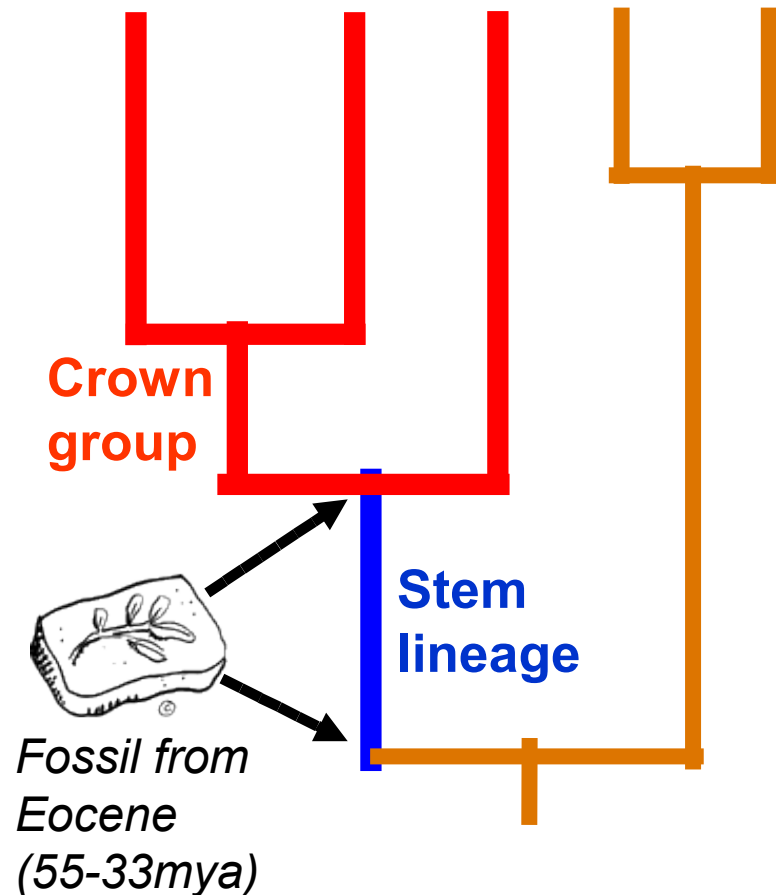
- Fossils provide minimum ages
- Geography can provide maximum ages
- Secondary calibrations are also possible



# Problems with Fossils



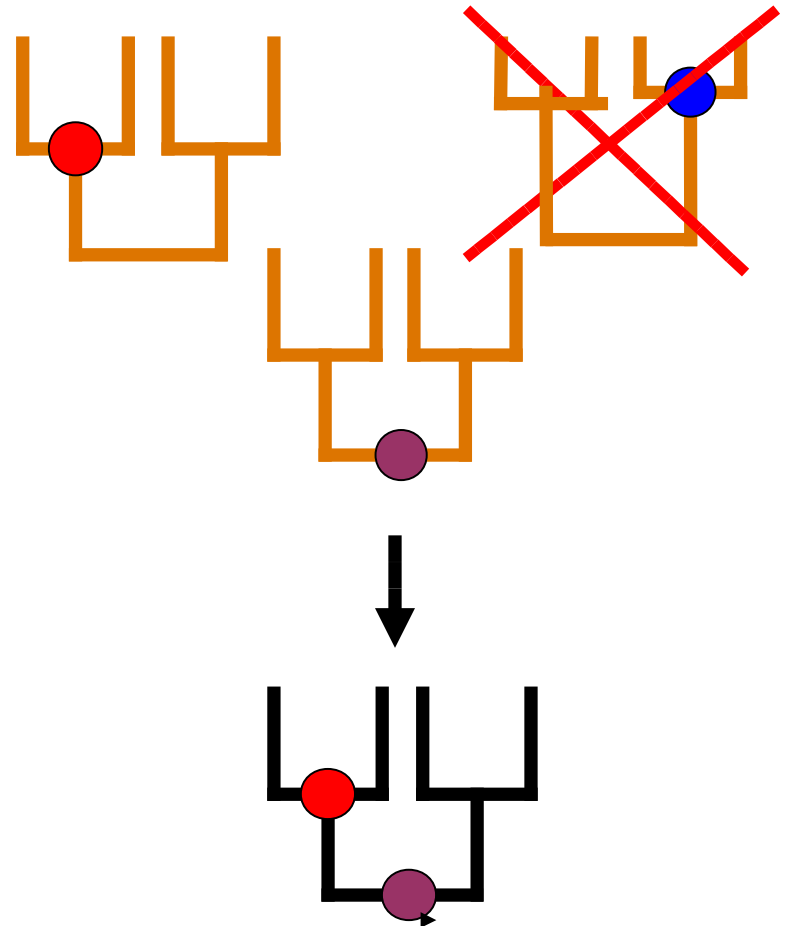
- Incompleteness of fossil record
- Identification
  - Species / Genus / Family?
- Position
  - Stem or Crown?
- Which date?
  - Min/Mid/Max of Epoch?



# Solving Problems with Fossils



- Use many calibration points
- Check if different calibrations agree
- Discard discordant fossils



# Confidence Intervals



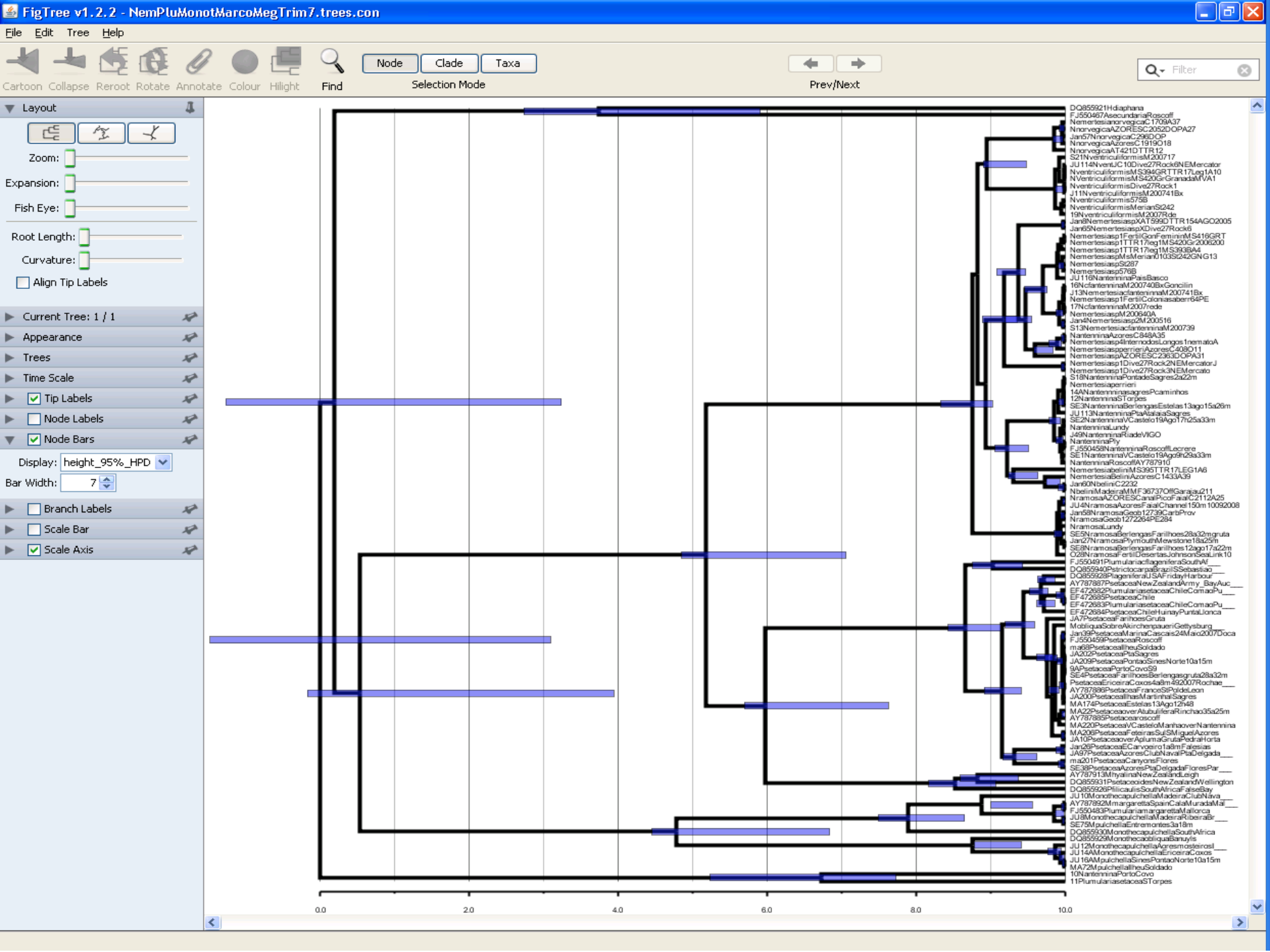
- Bootstrap re-sampling can provide confidence intervals (CI) for dates
  - Create bootstrap replicates for your chosen topology
  - Date each replicate with your chosen method
  - Age range = 95% CIs for the mean age of node
- Other methods exist
  - Bayesian dating methods produce ranges without bootstrapping
  - Can use Bayesian replicates from phylogenetic analysis instead of bootstrap replicates

*Torsten Eriksson r8s-bootkit is a useful tool for doing bootstrapping*  
[http://www.bergianska.se/index\\_forskning\\_soft.html](http://www.bergianska.se/index_forskning_soft.html)

# Confidence Intervals



- Software written by the writers of BEAST is very useful for generating confidence intervals for dates
- Treeanotator runs through tree replicates in a file and works out confidence intervals for all nodes on a given tree (<http://beast.bio.ed.ac.uk/>)
- Figtree displays the tree with confidence intervals as bars on the phylogeny (<http://beast.bio.ed.ac.uk/FigTree>)



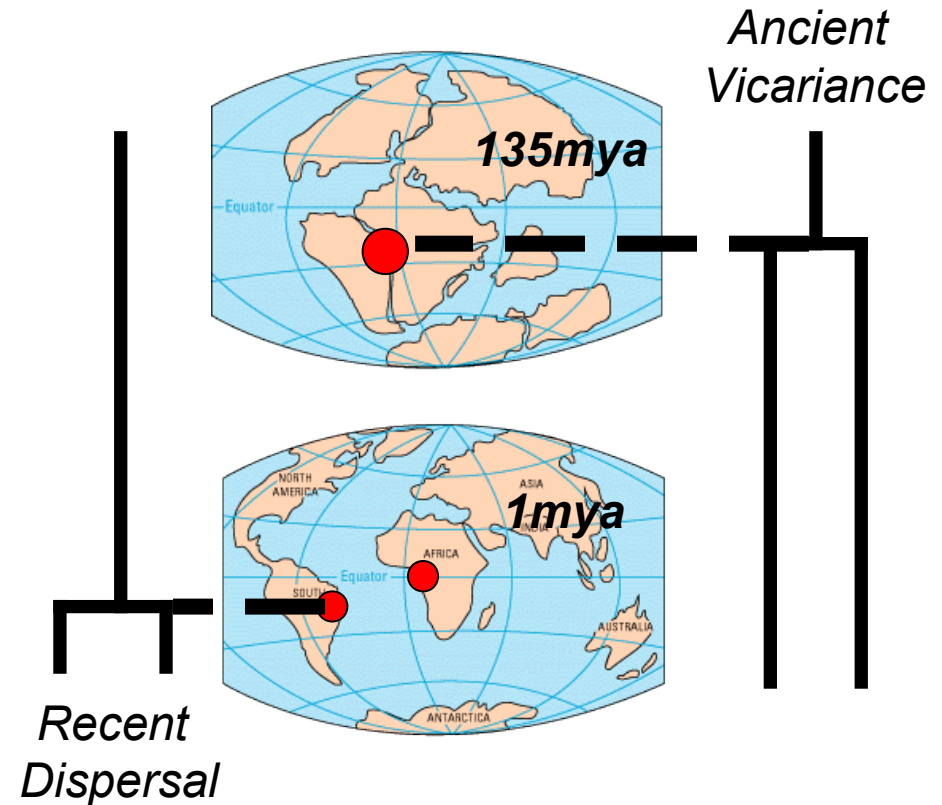
- Many trees reject the molecular clock
- Popular NPRS is widely criticised
- Finding & positioning fossils is difficult
- Bootstrapping is a tedious process
- Running an entire phylogenetic analysis in beast can take time and often is done alongside a traditional analysis
- Much of the software is not user friendly and prone to error
  - Sanderson on r8s: *“Frankly, I’m always surprised when it works...”*



# So Why Bother?



- Deeper understanding of evolutionary processes
- Biogeography
  - Dispersal or vicariance?
- Impact of historic climate change



# Examples



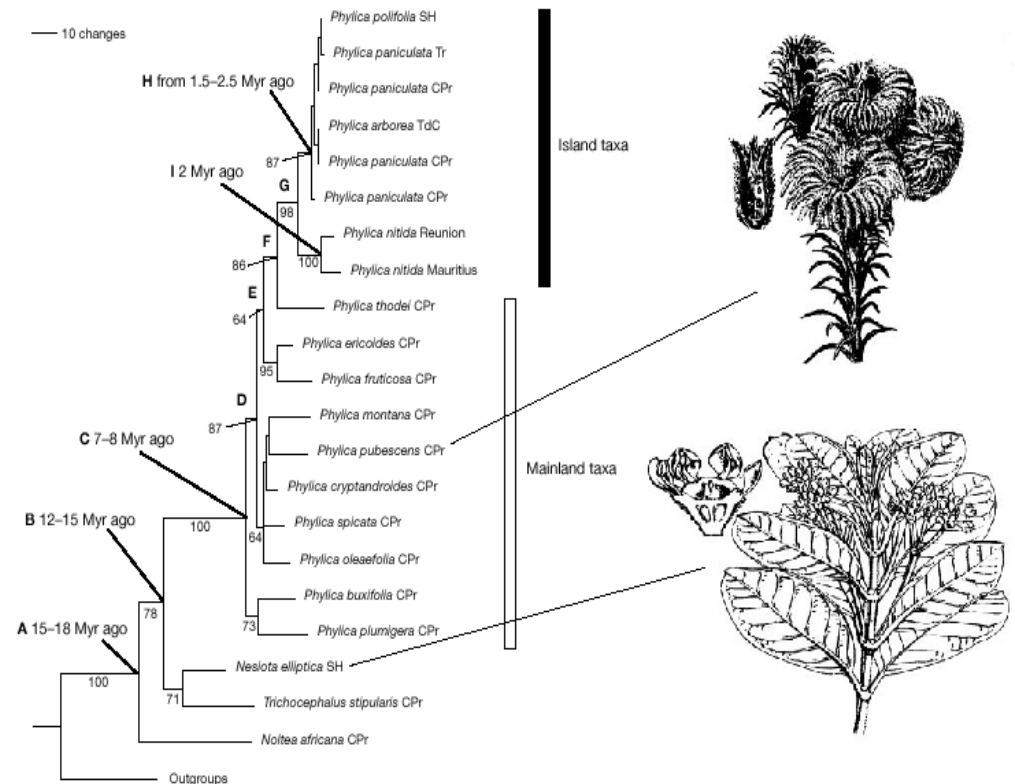
## letters to nature

### Rapid and recent origin of species richness in the Cape flora of South Africa

James E. Richardson<sup>\*,†</sup>, Frans M. Weitz<sup>‡,§</sup>, Michael F. Fay<sup>\*</sup>, Quentin C. B. Cronk<sup>||,¶</sup>, H. Peter Linder<sup>†,§</sup>, G. Reeves<sup>\*</sup> & Mark W. Chase<sup>\*</sup>

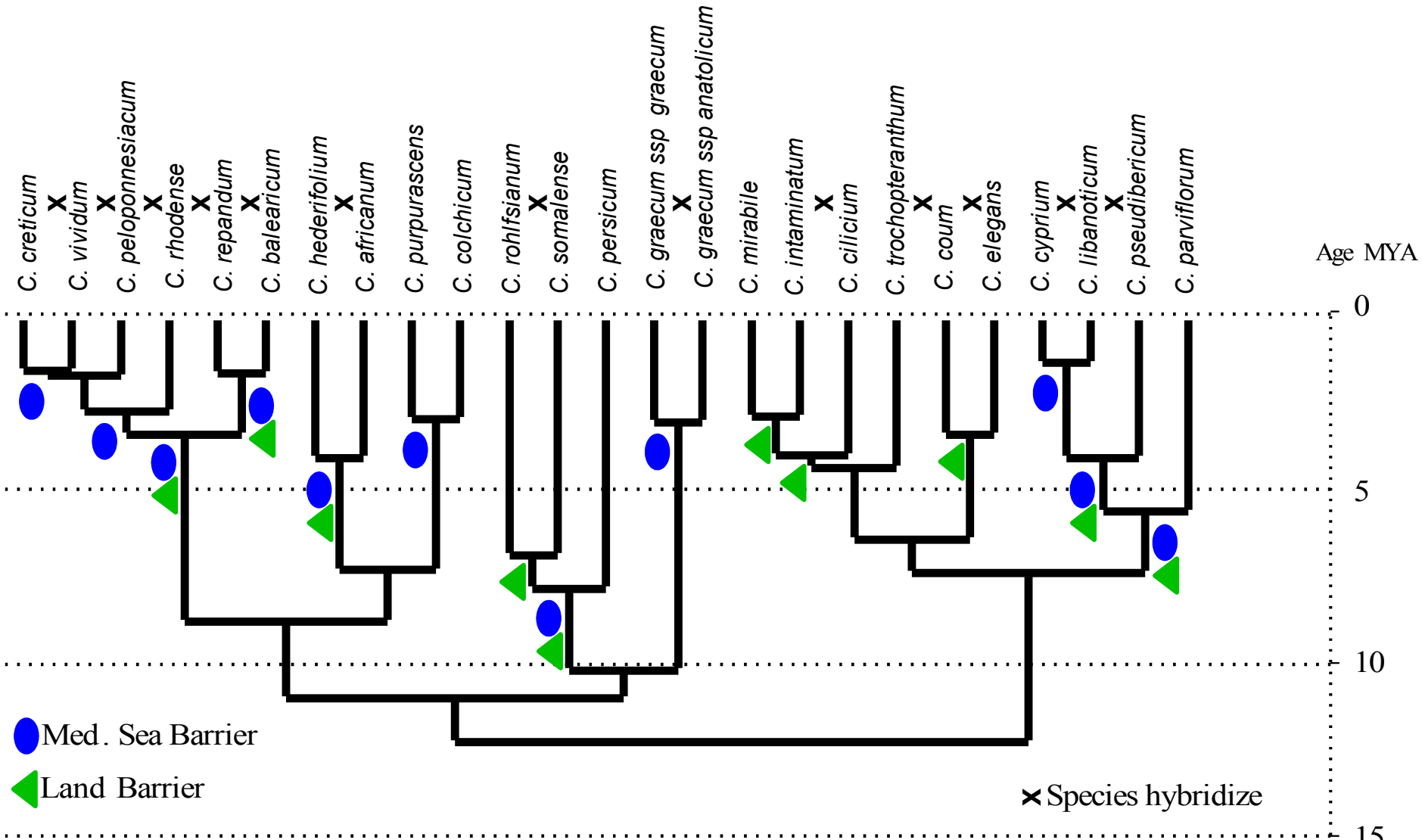


Fig. 1. Distribution of *Phyllica*. Tristan da Cunha indicated here represents Tristan da Cunha, Nightingale and Inaccessible islands.

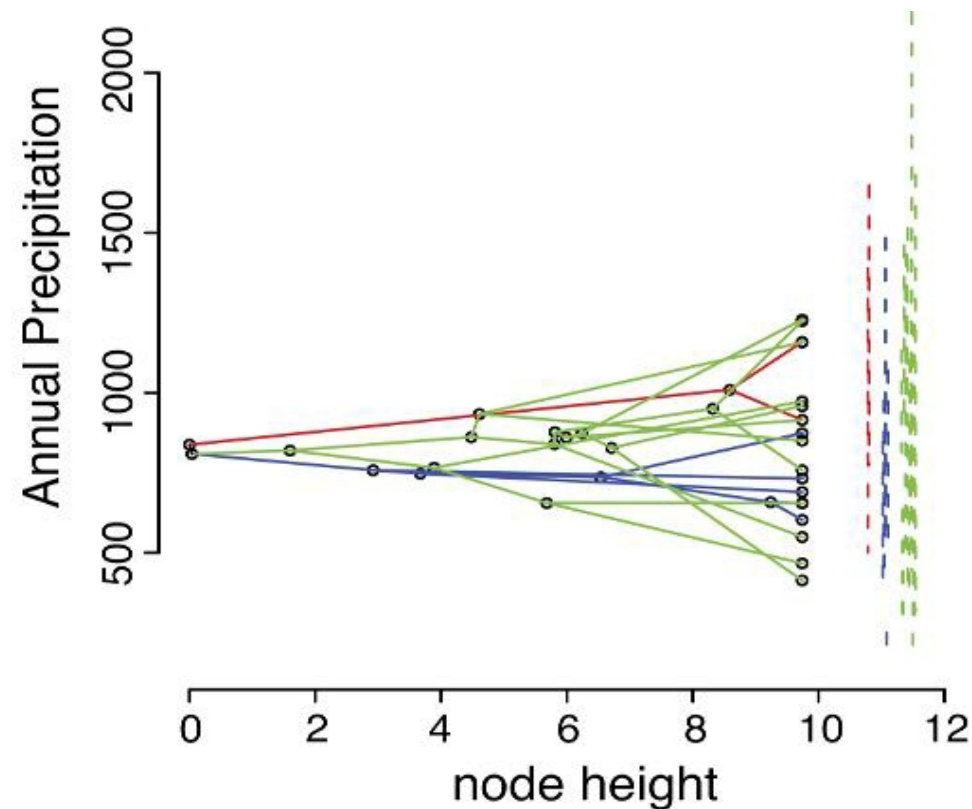
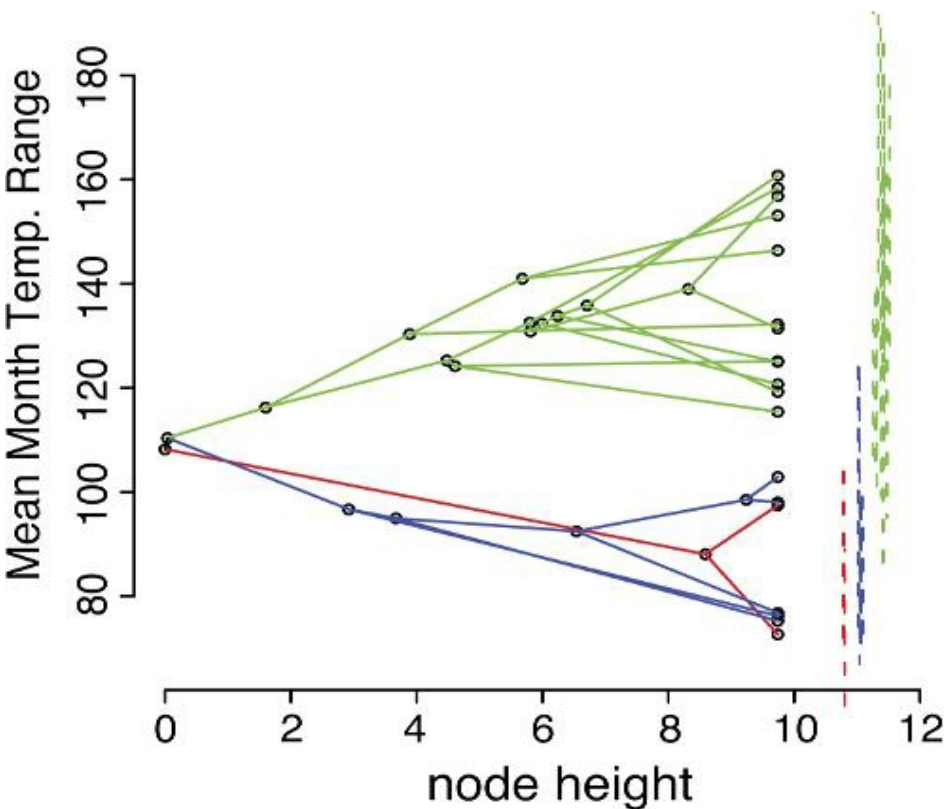
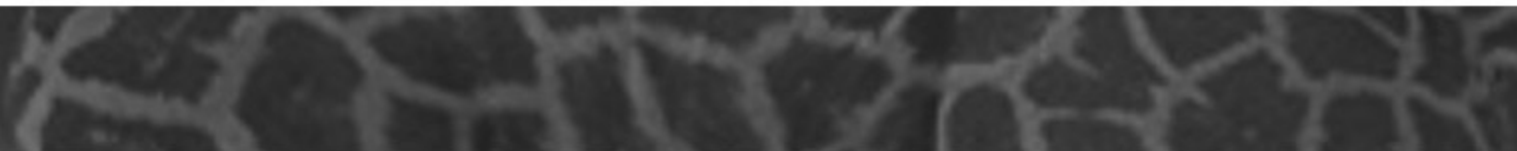


Richardson *et al.* (2001) *Nature* **412**:181

# Example



Yesson, C.; Toomey, N. H. & Culham, A. Cyclamen: time, sea and speciation biogeography using a temporally calibrated phylogeny *Journal of Biogeography*, 2009, 36, 1234-1252



Smith, S. & Donoghue, M. Combining Historical Biogeography with Niche Modeling in the Caprifolium Clade of Lonicera (Caprifoliaceae, Dipsacales) Systematic biology, 2010, 59, 322-341