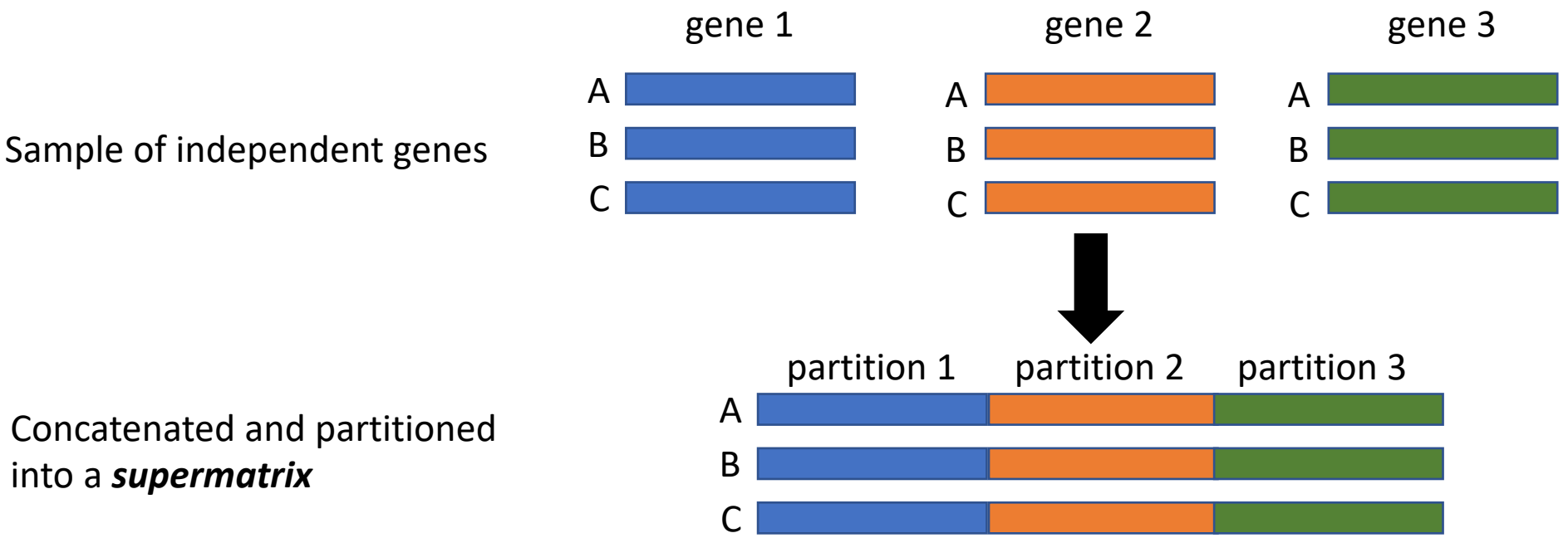
Phylogenetics II

Multi-locus Phylogenetics: Concatenation

Stems from Kluge's "total evidence" philosophy (1989, 2004)



Each partition can be assigned it's own substitution model, and summed over to estimate the likelihood

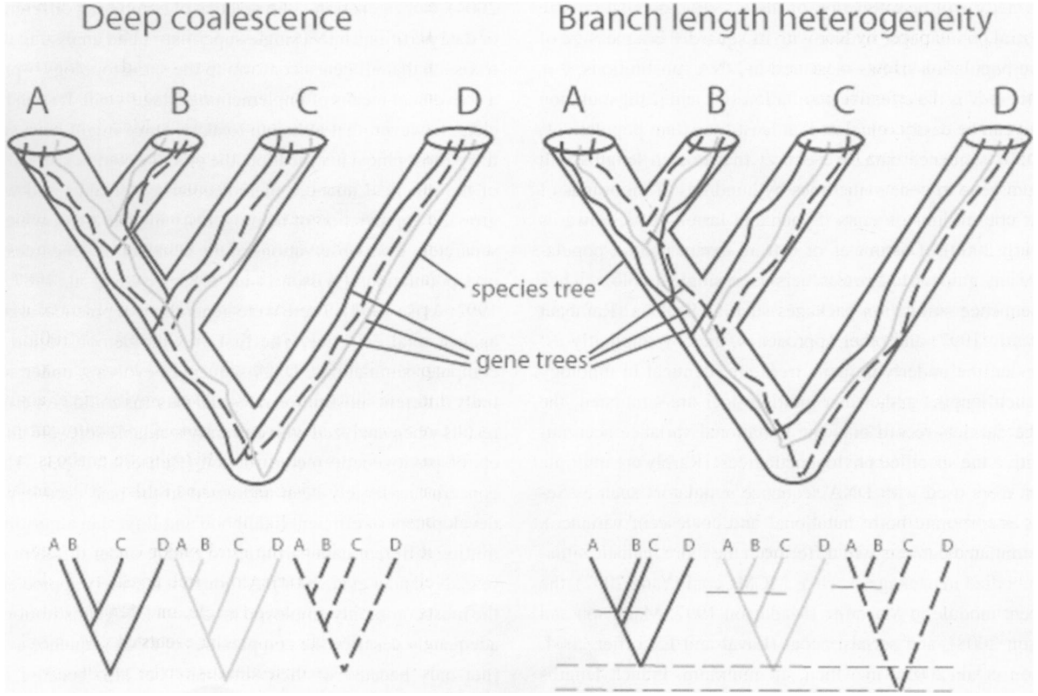
Multi-locus Phylogenetics: problems with concatentation

Deep coalescence = incomplete lineage sorting

- *topological differences between gene trees*

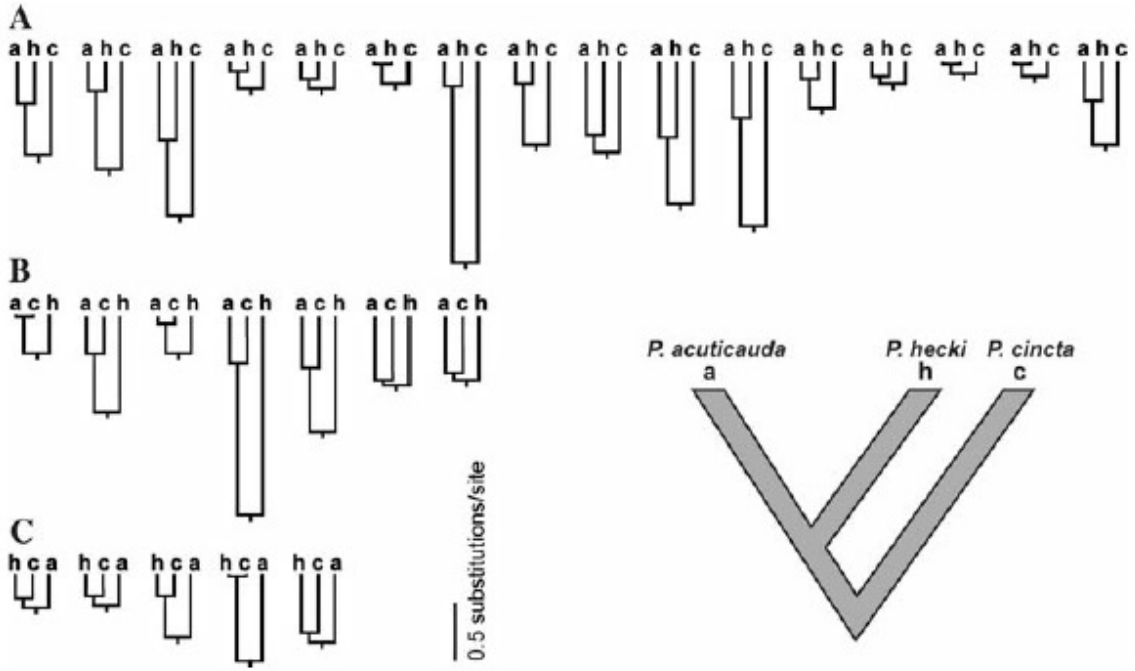
Branch length heterogeneity

- no topological variation but different estimates in rates.



Edwards. 2009. Is a new and general theory of molecular systematics emerging? *Evolution*.

Example of heterogeneity in topology and coalescent times when sampling multiple genes



Brito and Edwards. 2009. Multilocus phylogeography and phylogenetics using sequence-based markers. *Genetica.*

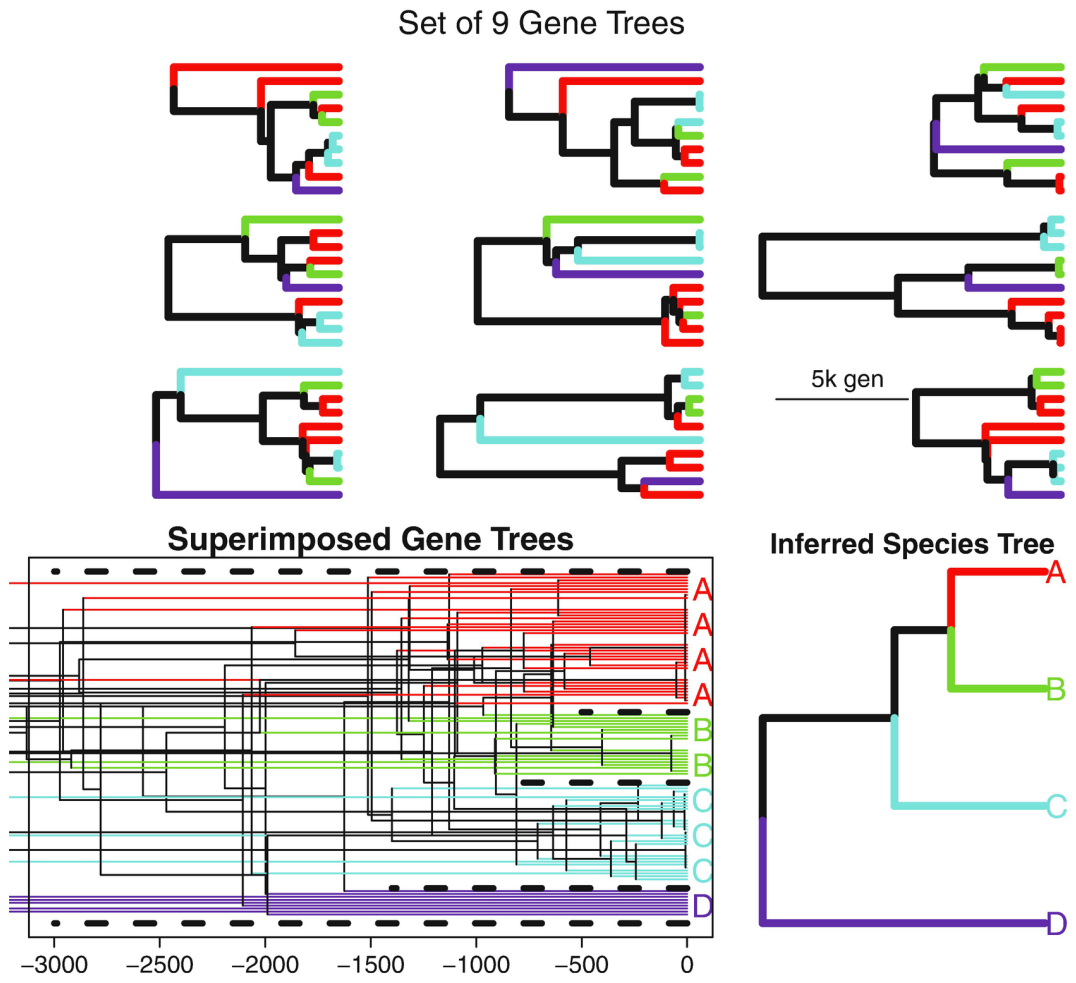
Multispecies Coalescent

Gene tree heterogeneity is the rule, and not the exception.

Nine gene trees, no single gene reconstructs the correct species relationships.

Superimposing the gene trees clarifies the relationships.

- ABC diverged from D ~1500 generations ago.
- AB split from C at ~800
- A diverged from B ~600
- The amount of cross breeding within recently diverged taxa implies that C has the smallest effective population size.



Liu et al. 2019. Modern Phylogenomics: Building Phylogenetic Trees Using the Multispecies Coalescent Model.

Multispecies Coalescent

Explains the evolutionary history of multilocus sequences through two levels of biological hierarchy:

1. The gene tree
2. The species tree

Process:

Generate gene trees (ML, Bayesian, whatever)

Incorporate the individual genealogies under a coalescent model

Genes are sampled alleles who have ancestors in ancestral populations

These genealogies “coalesce” in the ancestors

“Averages” over gene trees to infer a species tree

ASTRAL – Accurate Species Tree Algorithm (Zhang et al. 2018 *BMC Bioinformatics*)

Review: Maximum Likelihood Estimation

Asks the question: ***How well does the model fit the data?***

The ***likelihood*** is the probability of the data given the model

- P(D|H)
- Felsenstein (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution.*

Process:

1. Generate a large number of trees using a model of evolution
2. Calculate the likelihood for each tree
3. Choose the tree with the highest likelihood

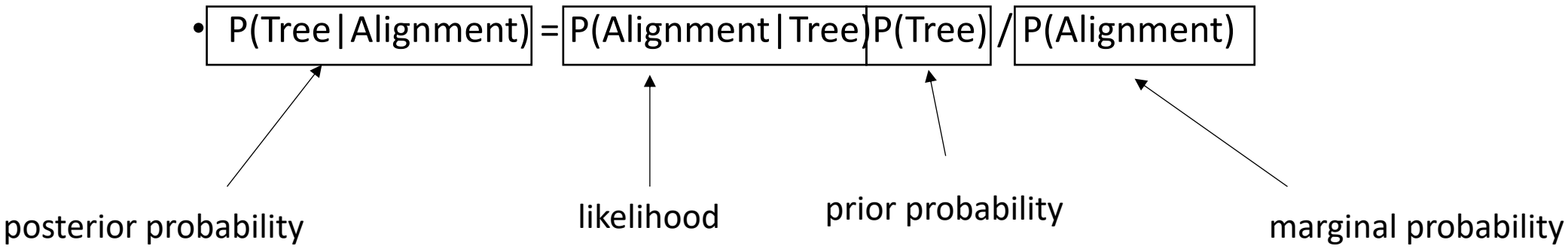
Phylip, RaxML, PhyML, MEGA

Likelihood methods vs Bayesian methods

- Conditional probabilities:
  - P(A|B)P(B) = P(B|A)P(A)
    - the probability of A given B times the probability of B is equal to the probability of B given A times the probability of A
  - P(A|B) = P(B|A)P(A) / P(B)
    - The probability of A given B is equal to the probability of B given A times the probability of A divided by the probability of B
  - In terms of phylogenetics, let's talk about hypotheses and data:
    - P(H|D) = P(D|H)P(H) / P(D)
    - P(Tree|Alignment) = P(Alignment|Tree)P(Tree) / P(Alignment)

Likelihood methods vs Bayesian methods

- Conditional probabilities:
  - P(A|B)P(B) = P(B|A)P(A)
    - the probability of A given B times the probability of B is equal to the probability of B given A times the probability of A
  - P(A|B) = P(B|A)P(A) / P(B)
    - The probability of A given B is equal to the probability of B given A times the probability of A divided by the probability of B
  - In terms of phylogenetics, let's talk about hypotheses and data:
    - P(H|D) = P(D|H)P(H) / P(D)
    - P(Tree|Alignment) = P(Alignment|Tree)P(Tree) / P(Alignment)



Likelihood methods vs Bayesian methods

In a nutshell:

Likelihood:

$P(D|H)$

vs

Posterior probability:

$P(H|D)$

The posterior is proportional to the prior times the likelihood

- Combines information in the prior and in the data

Likelihood methods vs Bayesian methods

Bayesian inference uses probability distributions to describe the uncertainty of all unknowns, including model parameters.

A prior distribution is assigned to model parameters, based on knowledge about them.

In phylogenetics:

- substitution models – parameters we talked about last time

- tree models – depends on clock-like or non-clock model, birth/death processes

Prior distributions each have their own shapes – choose carefully!!!

- uniform, normal, lognormal, exponential, etc…

The Markov chain Monte Carlo (MCMC) is used to simulate samples, which are taken from the prior distribution.

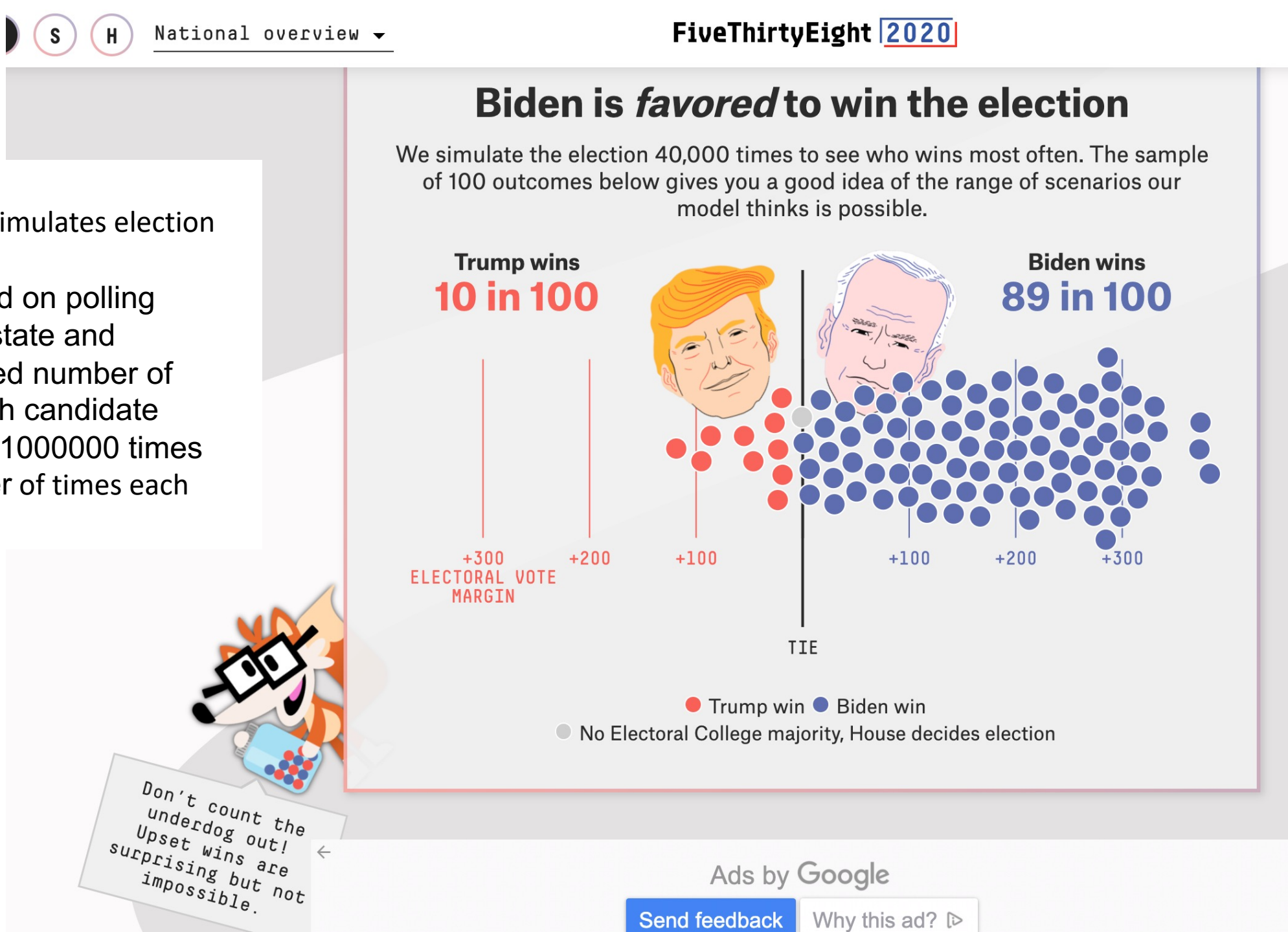
- searches for samples that maximize the <u>POSTERIOR</u> <u>PROBABILITY</u>

- You can then calculate mean, standard deviation, confidence intervals from the posterior.

Election Forecasting:

538 creates a model that simulates election results

- The models are based on polling averages from each state and estimates the expected number of electoral votes to each candidate
- They run the model ~1000000 times and report the number of times each candidate wins



Election Forecasting:

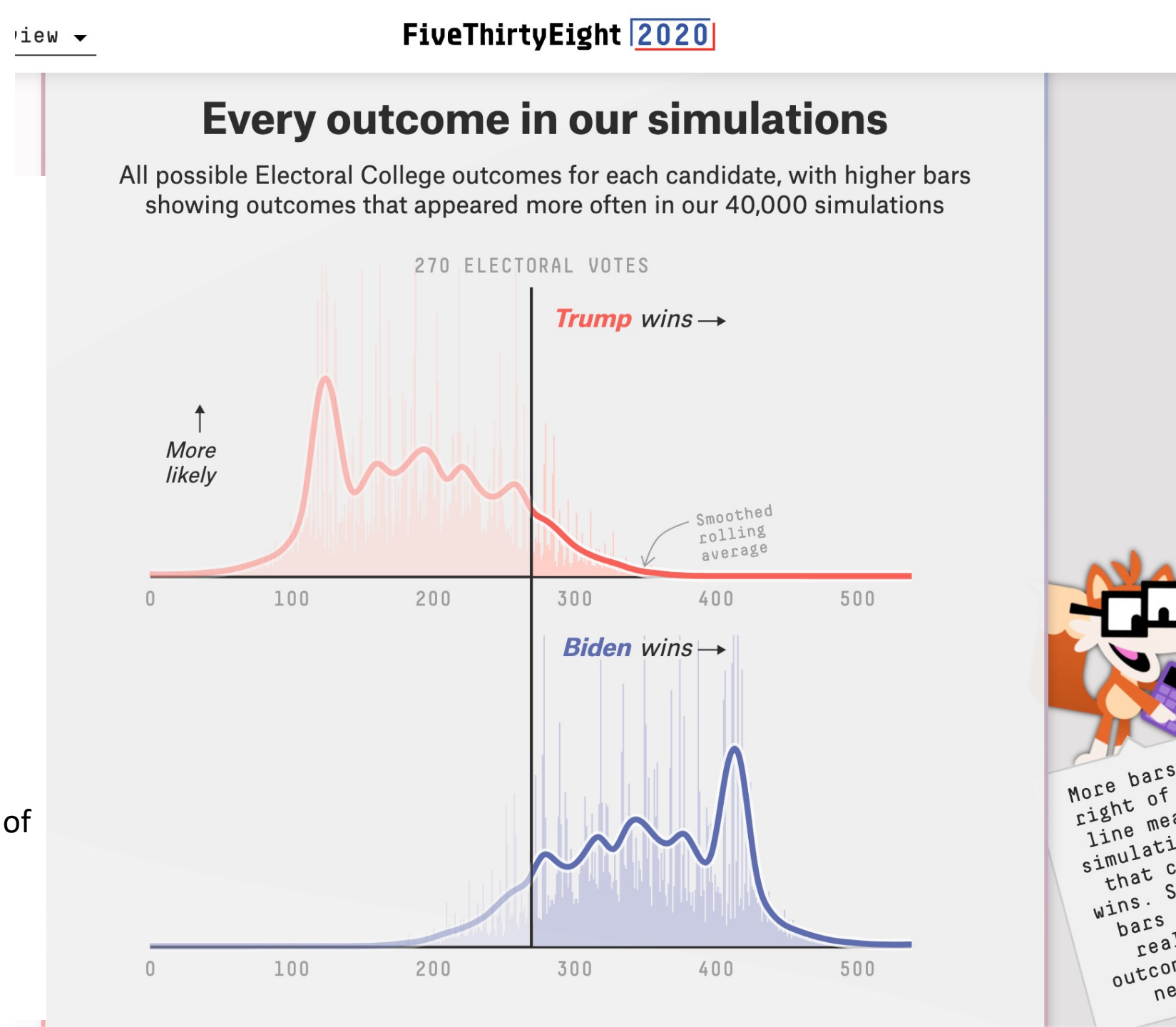
538 creates a model that simulates election results

- The models are based on polling averages from each state and estimates the expected number of electoral votes to each candidate
- They run the model ~1000000 times and report the number of times each candidate wins

The EV tally in each outcome is recorded

- Many outcomes include Biden >400 EV
- But also a lot around Biden ~300 EV

In reality, polling errors meant that because of some very close states, there were more plausible scenarios where Trump could win, and the election was closer (306-232)



Election Forecasting:

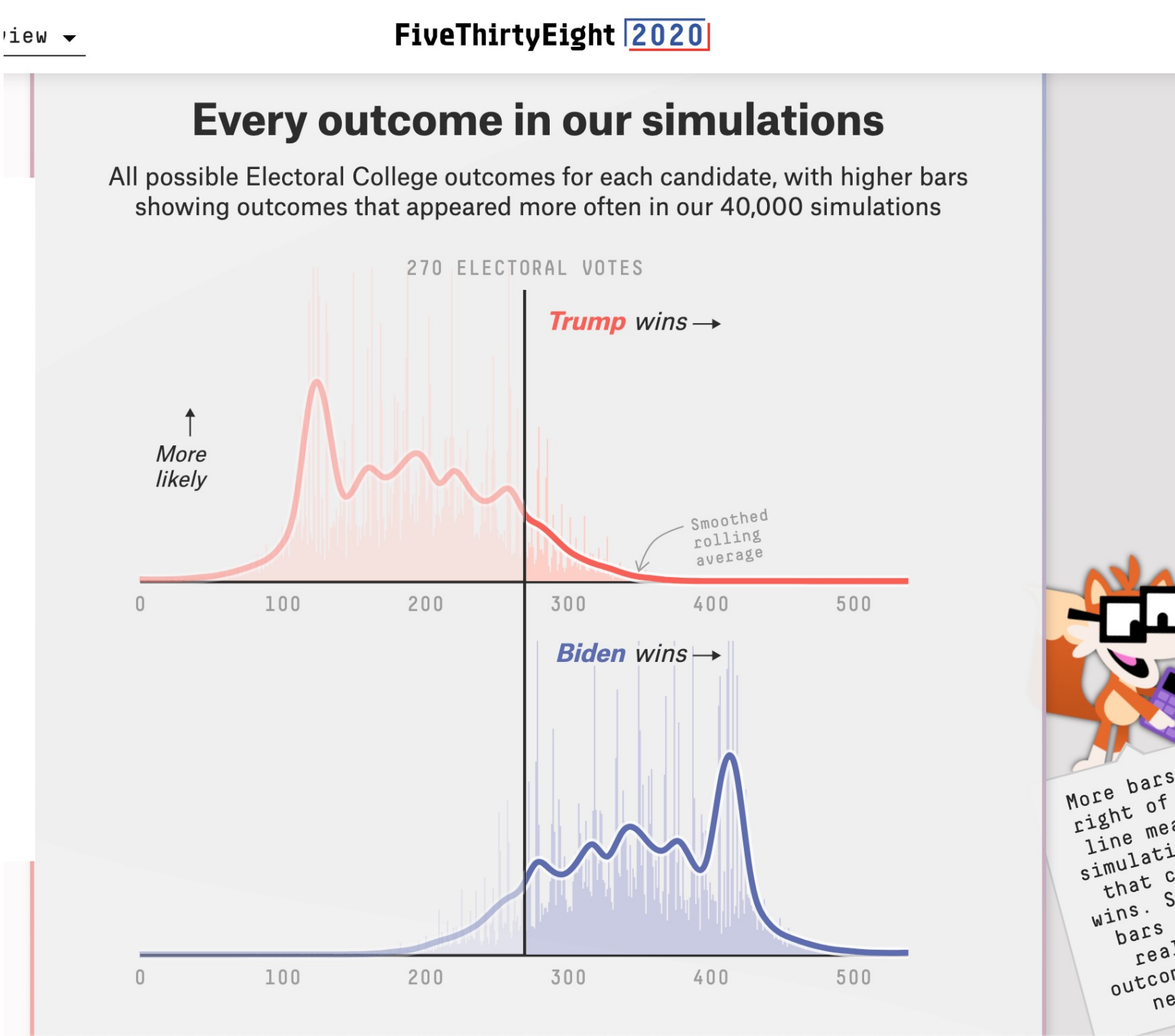
IMO, election forecasts are a product being sold to media and advertisers

They are useful because they help predict an outcome

They are generally correct – serious statisticians and programmers work on them

They are predictions based on probabilities and suffer from the sampling biases underlying all polling techniques

They are not fortunetellers



What Goes Into a Bayesian Phylogenetic Model?

Data

- One or more alignments
- Sampled at one or many timepoints
- Timescale of days to millions of years
- Realisation of a stochastic process

Model

- Description of the process that generated the data (substitution models, trees)
- Parameters are random variables
- We may be interested in only some parameters
- We still need a prior for all the model parameters!

Examples of MCMC sampling of posterior distributions

Parameters d and k

a and b represent ‘good mixing’

a’ and b’ represent ‘bad mixing’

In ‘bad mixing’, the MCMC is not finding ideal values and it gets ‘stuck’ in parameter space.

To get better mixing, you can extend the length of the chain (no. of simulations), or extend the number of steps in between samples.

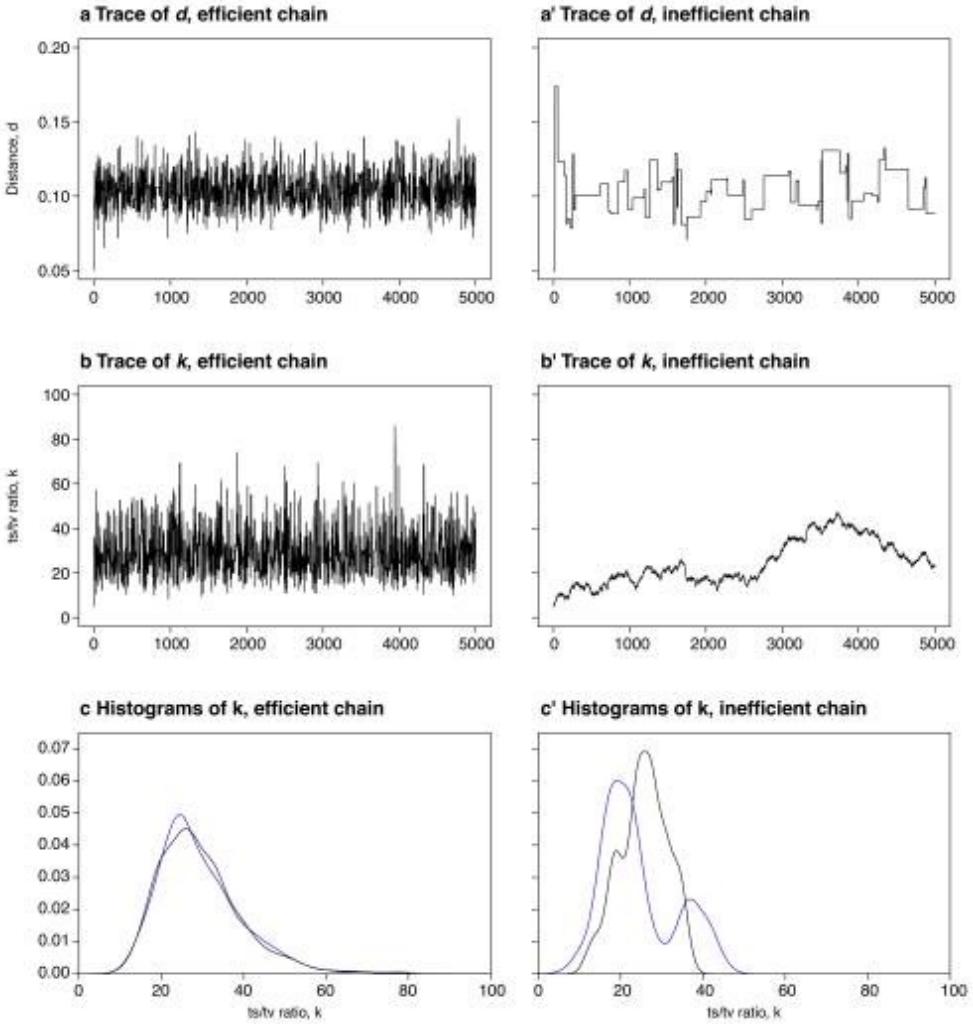
- In phylogenetics, typical MCMC chains are millions of steps.

c and c’ show histograms of k for two runs each of efficient and and inefficient chains.

- The efficient chains both ‘converge’ on the posterior.

“Burn-in” - discard the first 10-25% of the MCMC sampling

- Allows the MCMC to obtain the optimum.



Nascimento, dos Reis and Yang. 2017. ***A biologist’s guide to Bayesian phylogenetic analysis.*** *Nature Ecol Evol.*

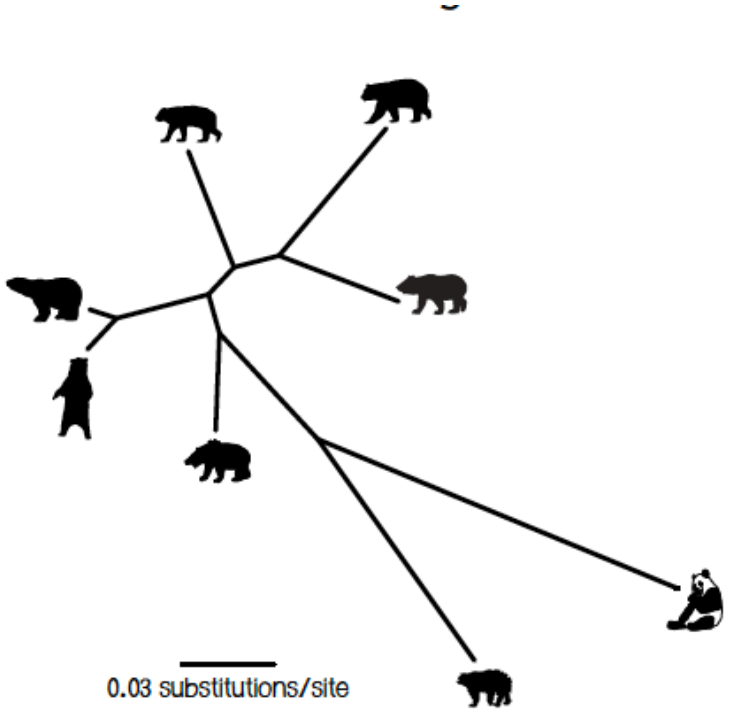
List of Bayesian Programs

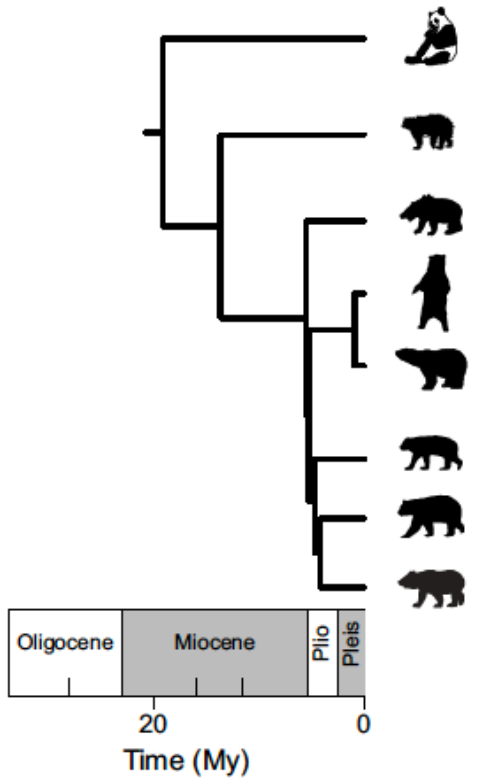
Program	Description	Reference
BEAST	Joint estimation of tree topology, divergence times, and more	Bouckaert R, et al. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput Biol. 2014;10:e1003537
MrBayes	Estimates species phylogenies and divergence times	Ronquist F, et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol.* 2012;61:539–542
RevBayes	Hierarchical Bayesian models	Höhna S, et al. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Syst Biol.* 2016;65:726–736.
MCMCTree	Estimates divergence times on a fixed phylogeny	Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 2007;24:1586–1591.
BPP	Species delimitation using a coalescent model and multi-locus data	Yang Z. The BPP program for species tree estimation and species delimitation. *Curr Zoo.* 2015;61:854–865.
Tracer	A program for MCMC diagnostics	Rambaut A, Suchard MA, Xie D, Drummond AJ. Tracer v1.6. 2014

Time Scale for Macroevolution

An unrooted tree with branch lengths in terms of ***substitutions per site*** only give us a relative view of evolution.

Phylogenies with branch lengths ***proportional to time*** tell us more information about evolutionary history.

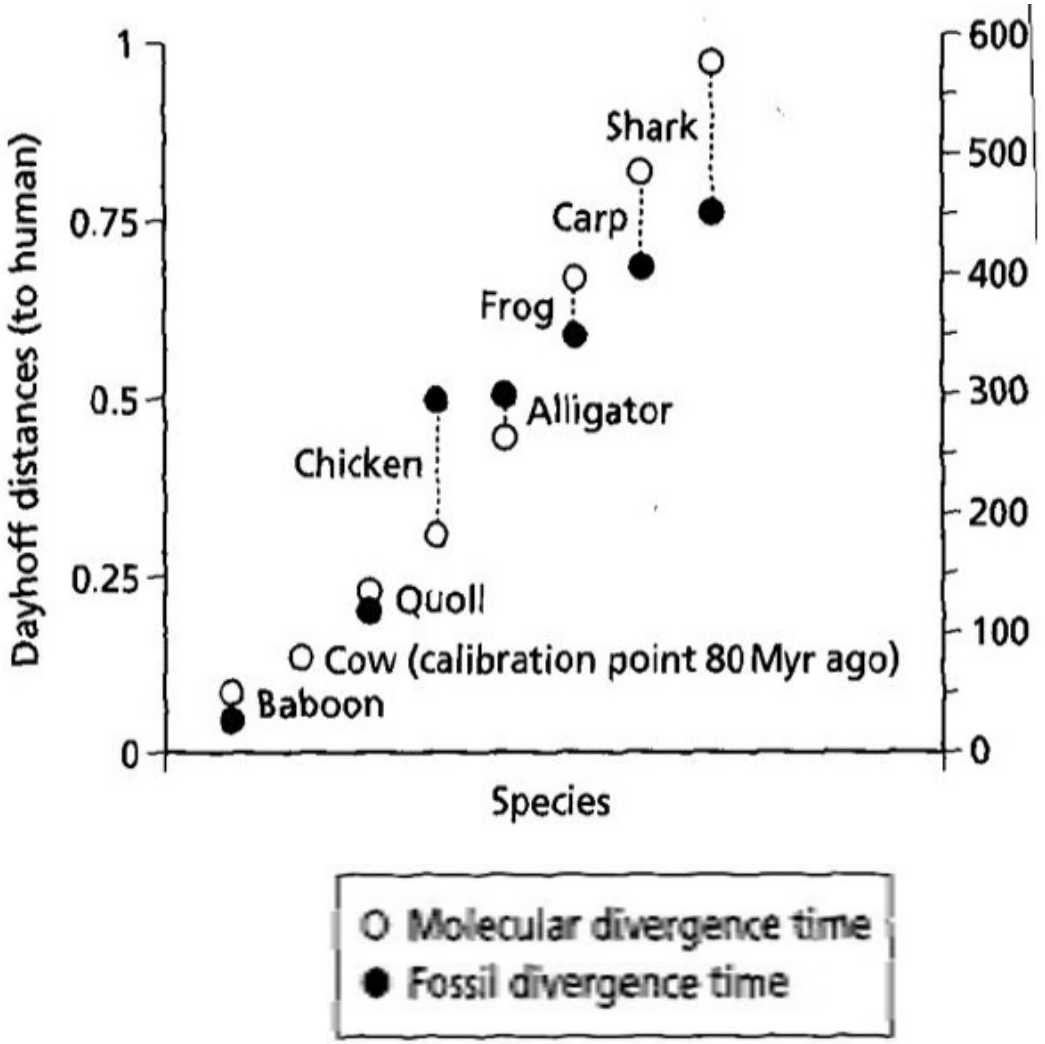




Tracy Heath

The Molecular Clock

Zuckerlandi and Pauling (1965) suggested the rate of evolution at the molecular level is ***constant through time and among species***.



Zuckerlandi and Pauling 1965

The Molecular Clock

Zuckerlandi and Pauling (1965) suggested the rate of evolution at the molecular level is ***constant through time and among species***.

Kimura (1968) suggested most mutations do not affect fitness (neutral theory), will become fixed through genetic drift.

- The rate at which neutral mutations become fixed is the ***substitution rate***.
- If mutation rates are similar among species, then substitution rates should be constant.

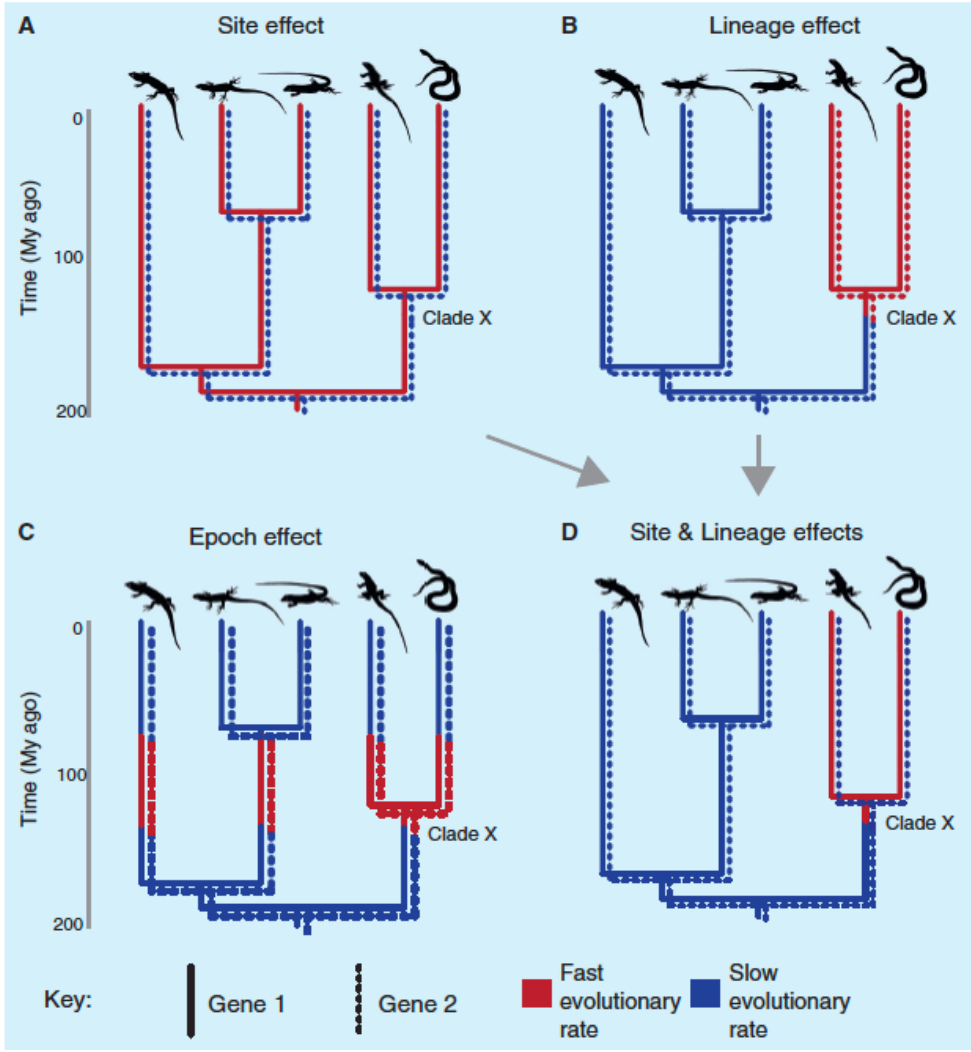
Motoo Kimura. 1983.
The Neutral Theory of Molecular Evolution

Evolutionary Rate Variation

Kimura's formulation turned out to be overly simplistic, especially as DNA sequencing took off in the 1990s-2000s.

Molecular clocks are now a suite of different models that try to explain this variation and take its uncertainties into account.

Kimura's clock is considered a 'strict' clock, versus 'relaxed' clocks.

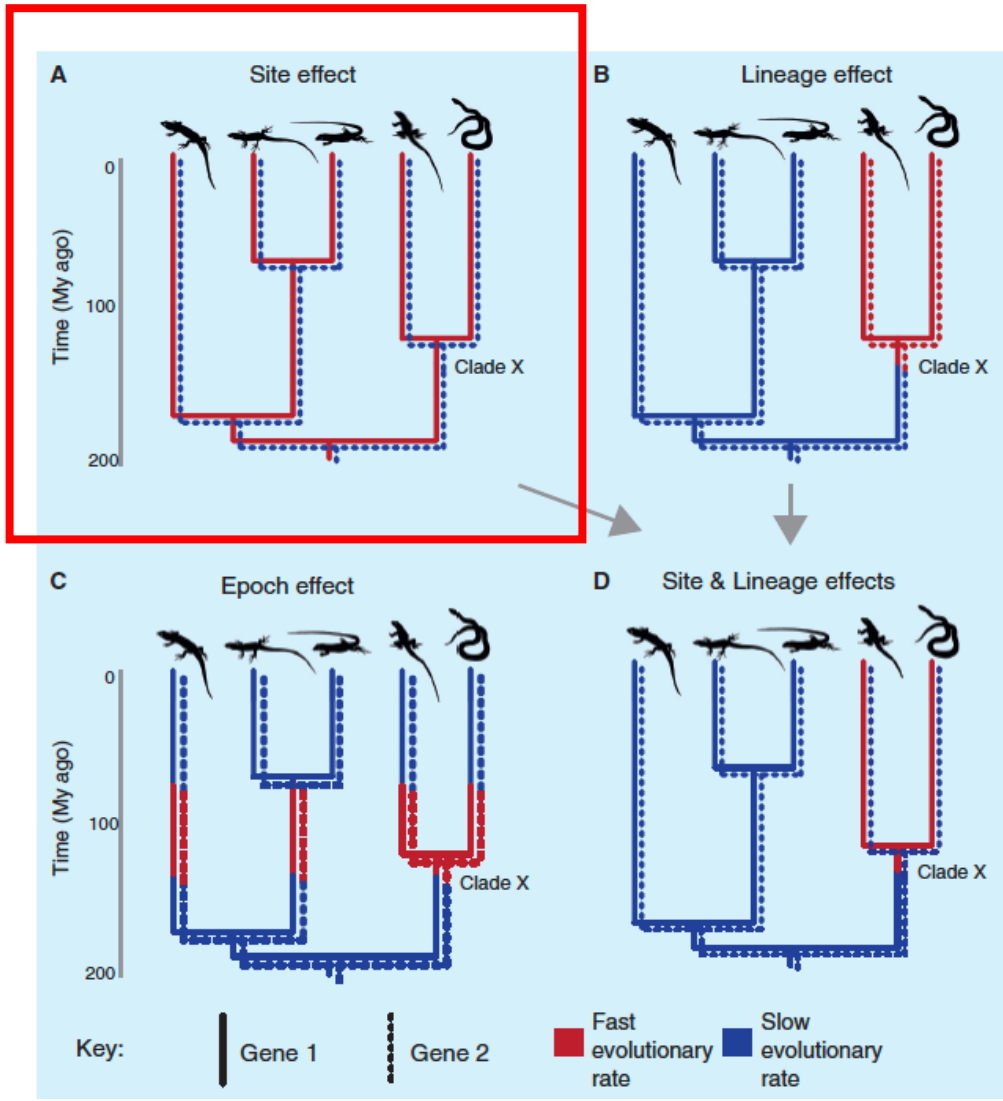


Lee and Ho. 2016. Molecular Clocks. *Current Biology*

Evolutionary Rate Variation

Site effects

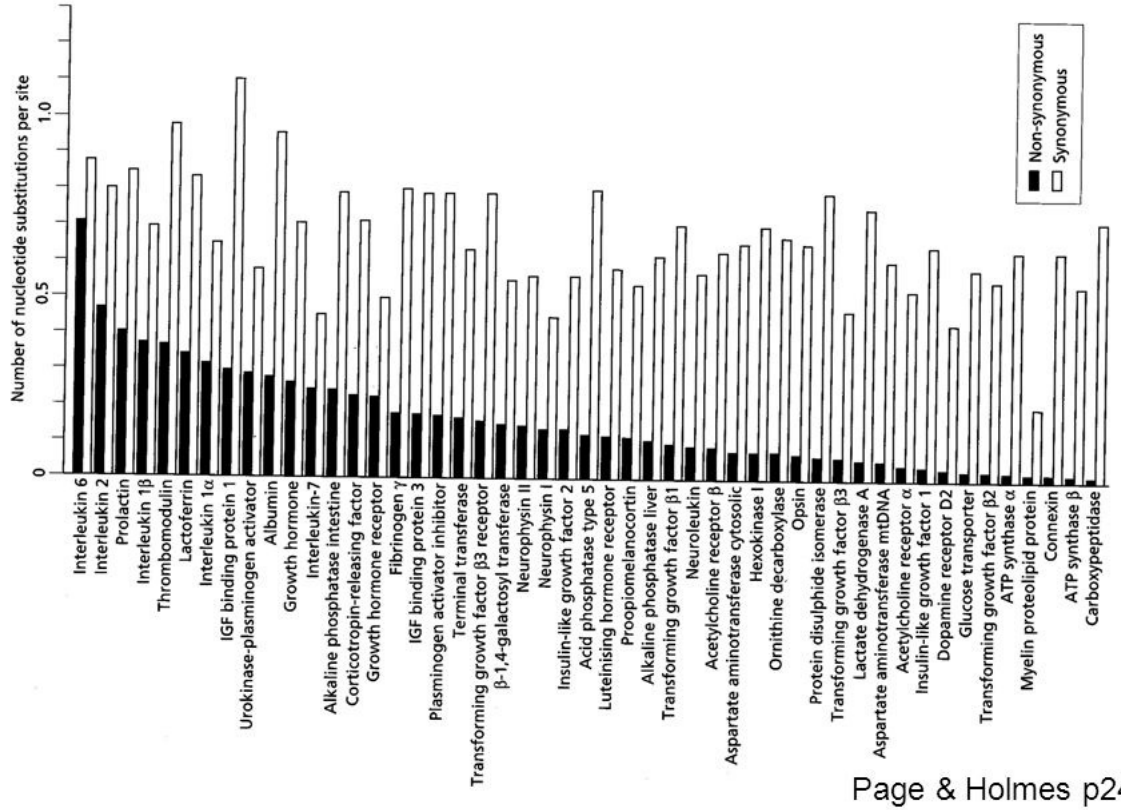
- Different parts of the genome evolve at distinct rates.
- For instance, ***protein-coding genes*** have higher rates at 3rd codons positions than at 1st and 2nd positions.
- There is a difference in the rates of **nonsynonymous** and **synonymous** substitutions.



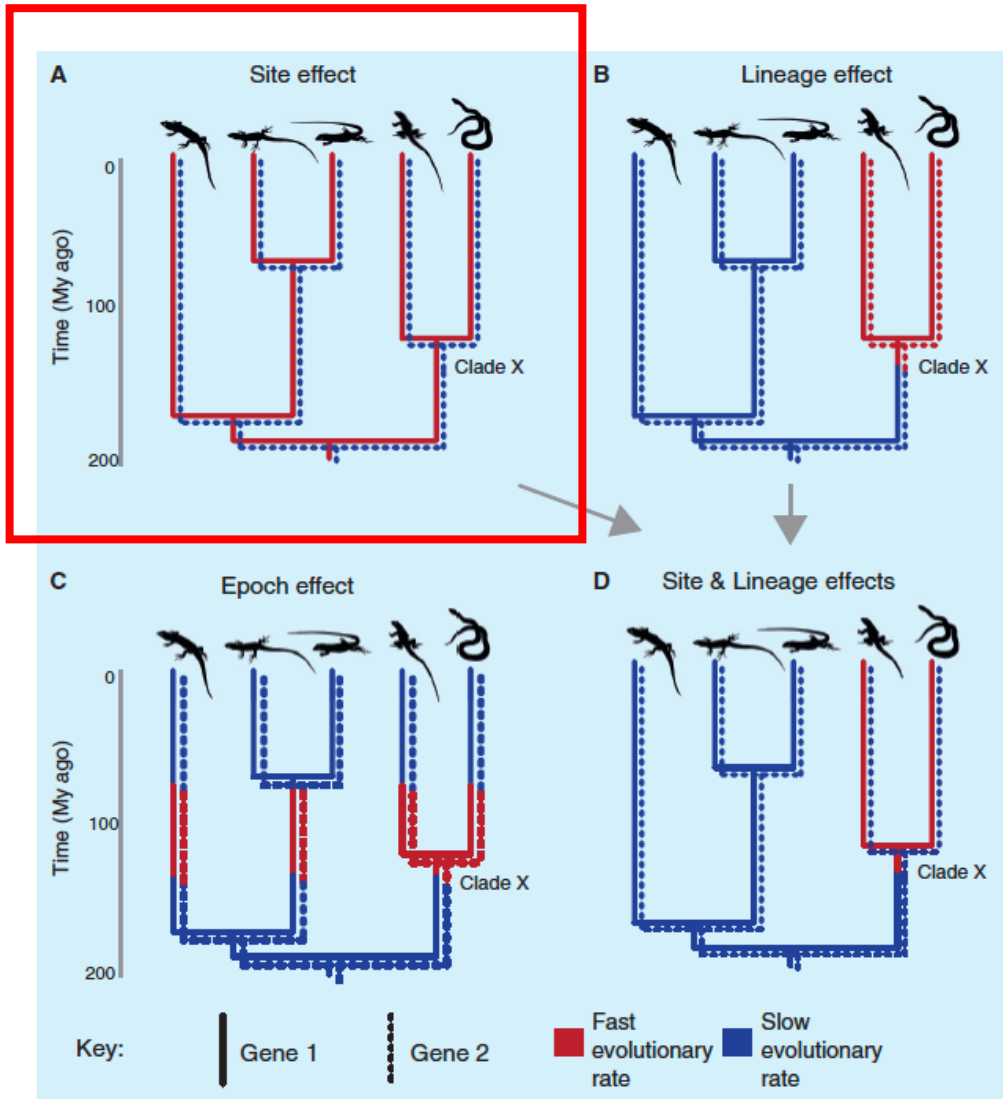
Lee and Ho. 2016. Molecular Clocks. *Current Biology*

Evolutionary Rate Variation

Site effects



Page & Holmes p240

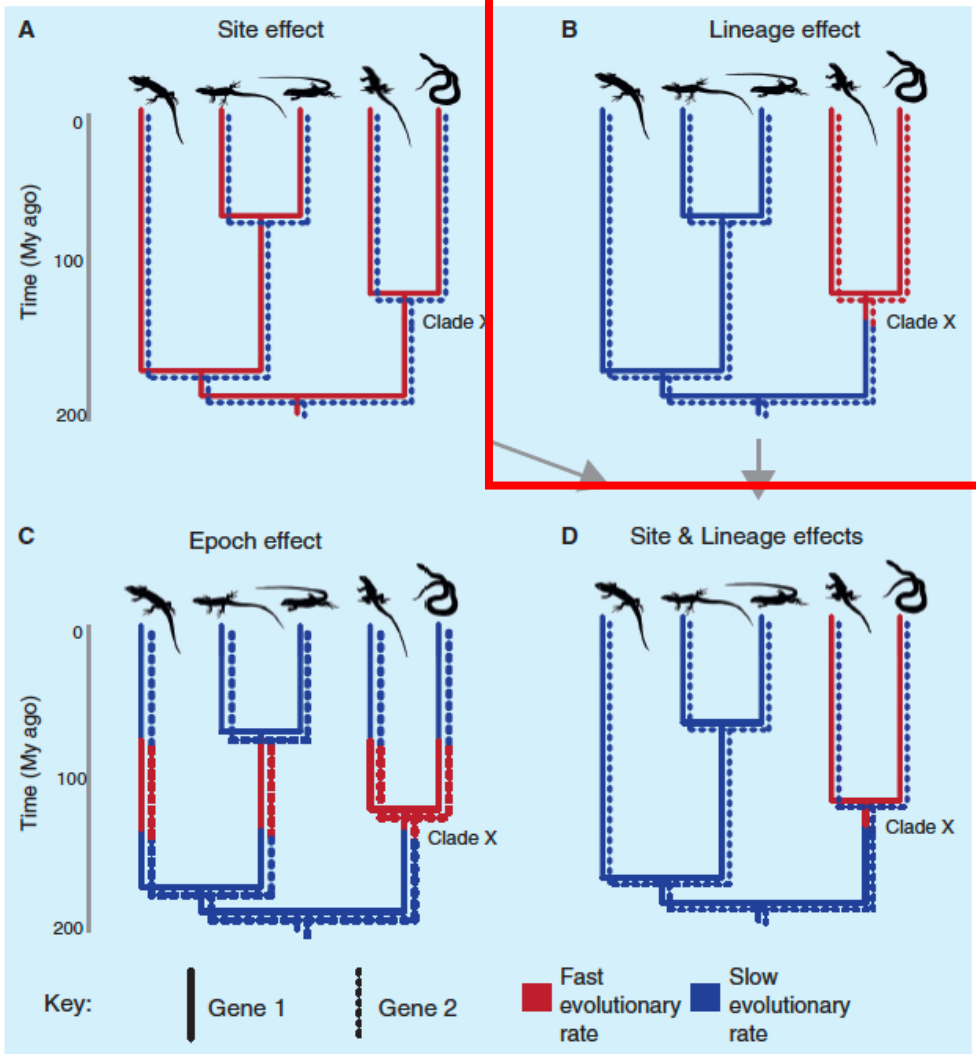


Lee and Ho. 2016. Molecular Clocks. *Current Biology*

Evolutionary Rate Variation

Lineage effects

- Different taxa evolve at distinct rates.
- This led to the development of ‘relaxed clock’ approaches.
- These clock models allow rate variation among branches in the phylogeny.
- Can estimate divergence times even when rates vary across lineages.

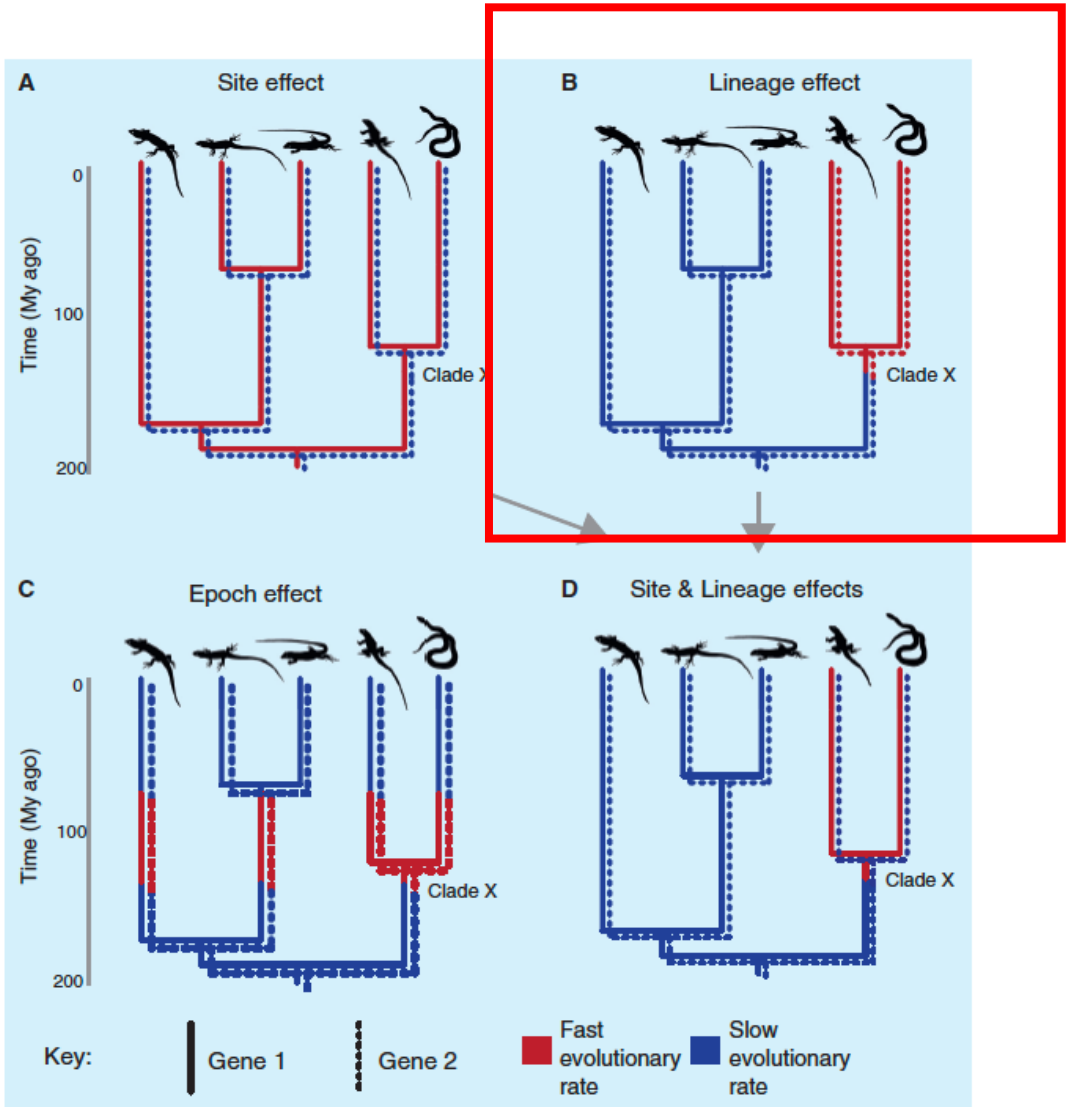


Lee and Ho. 2016. Molecular Clocks. *Current Biology*

Evolutionary Rate Variation

Lineage effects

Cause	Reason
DNA repair mechanisms	Error-prone polymerases
Metabolic rate	Production of free radicals
Generation time	Differences in the rate at which DNA is copied
Population size	Effect the substitution rate (fixation of alleles through drift)

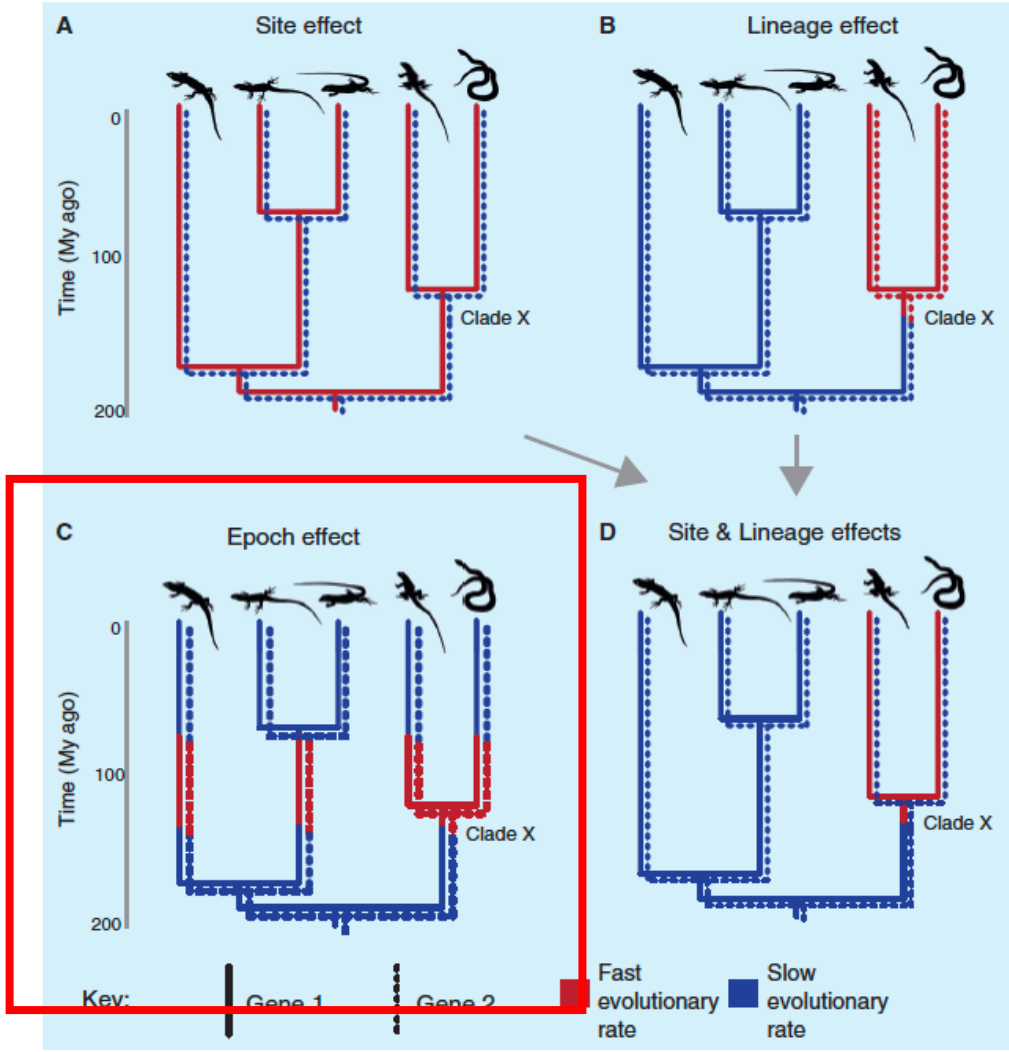


Lee and Ho. 2016. Molecular Clocks. *Current Biology*

Evolutionary Rate Variation

Epoch effects

- Rates of evolution differ across different time slices.
- Often the case in viral evolution.

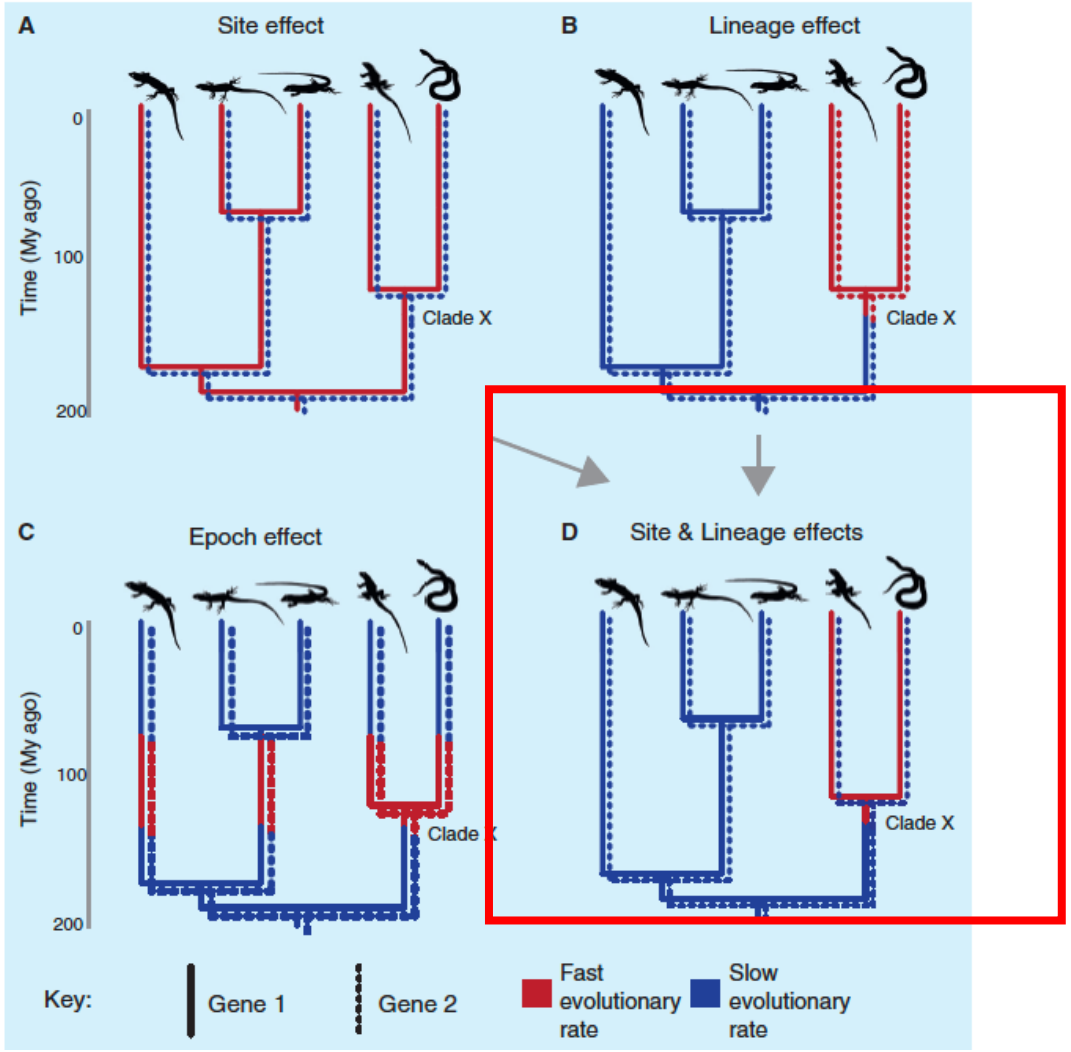


Lee and Ho. 2016. Molecular Clocks. *Current Biology*

Evolutionary Rate Variation

Interaction between sources of heterogeneity

- For instance, site and lineage effects can interact when different genes have different patterns across taxa.
- Selection may be relaxed on particular genes in particular taxa.
- For example, sloths and anteaters lack tooth enamel. Since selection is relaxed on genes coding for tooth enamel in these lineages, they may evolve ‘faster’.

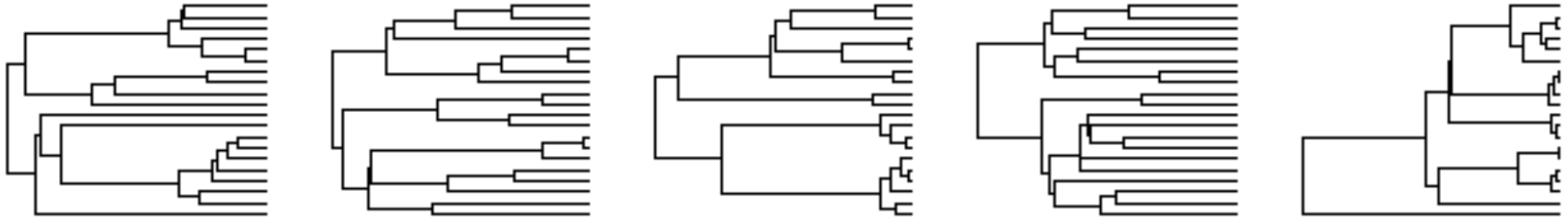


Lee and Ho. 2016. Molecular Clocks. *Current Biology*

Priors on Node Times

Sequence data alone is only informative for relative rates and times

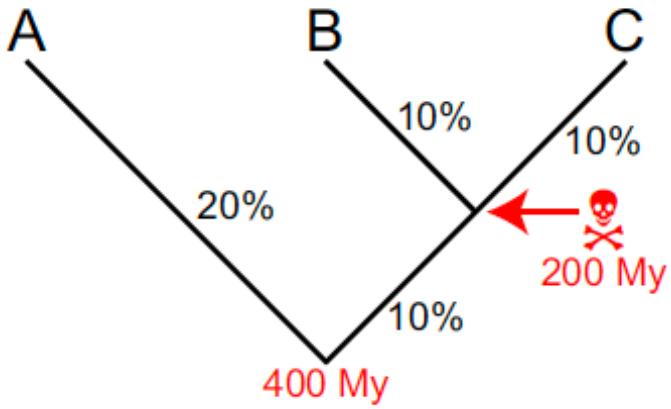
Ideally, we can have absolute estimates in real time units.



External evidence such as fossils allow us to calibrate (or scale) the tree to obtain absolute times.

Calibrating Divergence Times

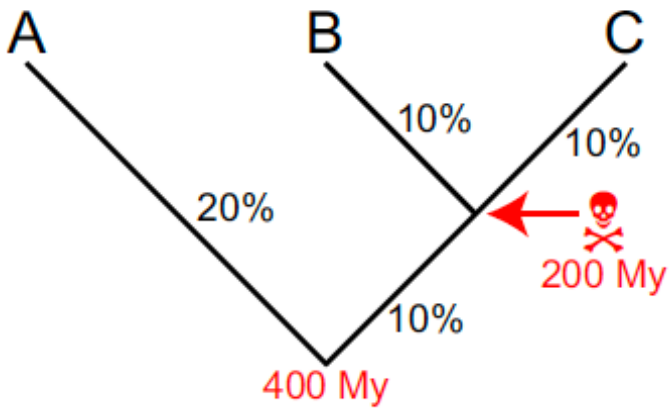
Fossils (or other data) are needed to estimate node ages.



However, there is uncertainty in the placement of fossils.

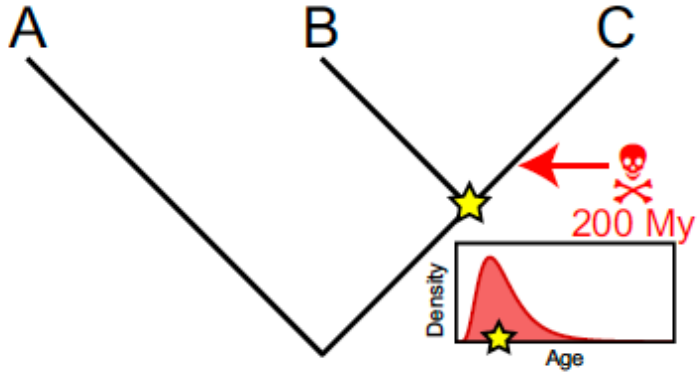
Calibrating Divergence Times

Fossils (or other data) are needed to estimate node ages.



However, there is uncertainty in the placement of fossils.

Bayesian approaches are well-suited to deal with this uncertainty.



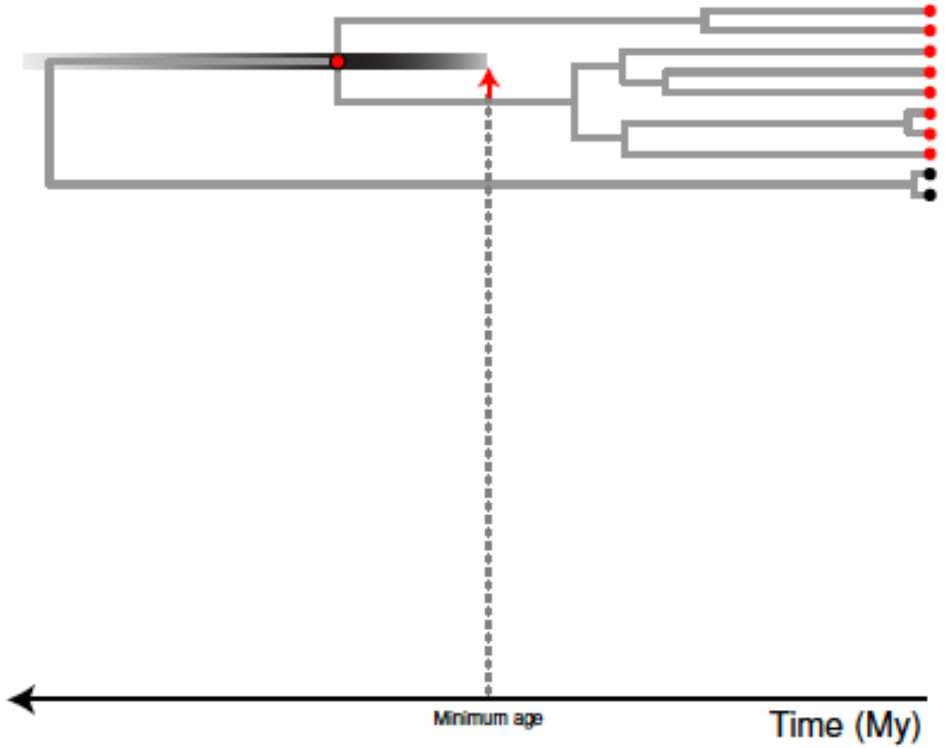
The divergence time of a clade can be treated as a parameter with a prior probability distribution.

We can use MCMC to sample from the posterior distribution.

Fossil Calibration

Age estimates from fossils can provide ***minimum*** constraints for internal nodes

Reliable ***maximum*** bounds are very hard to come by!!!!!

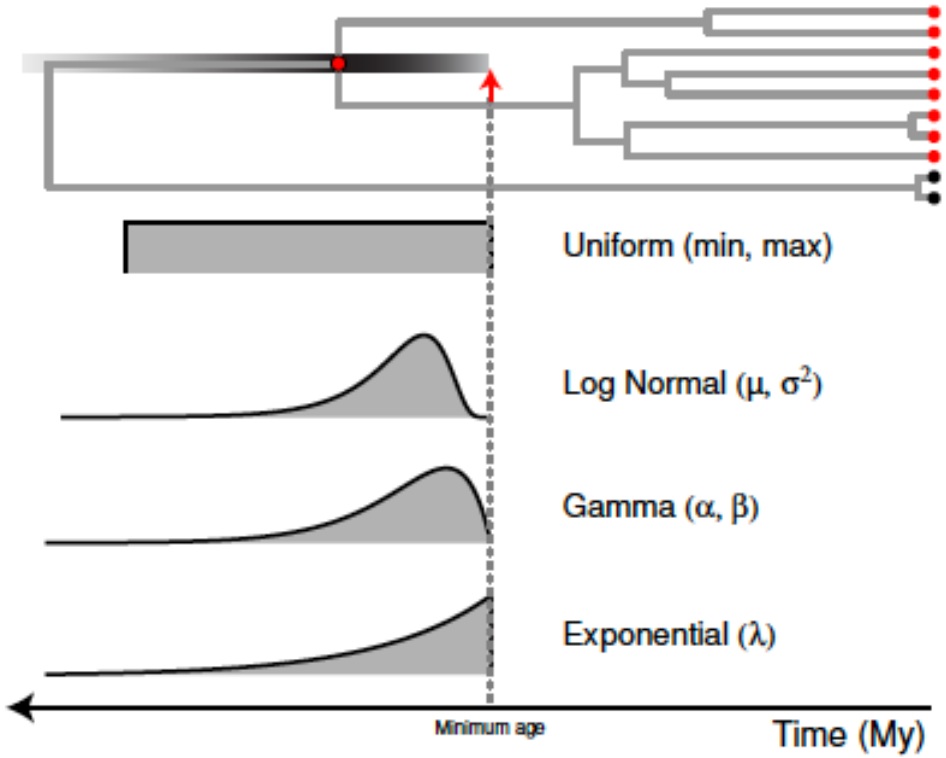


Tracy Heath

Prior Densities on Calibrated Nodes

Parametric distributions are typically off-set by the age of the ***oldest fossil assigned to a clade***.

These prior densities do not require specification of maximum bounds

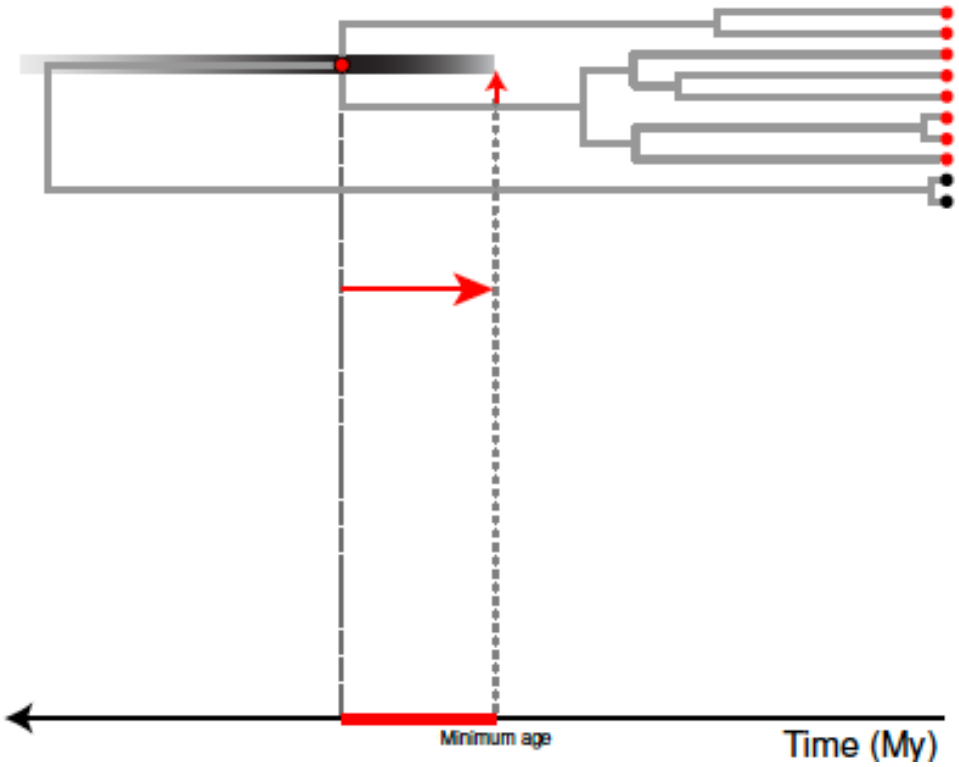


Tracy Heath

Prior Densities on Calibrated Nodes

The ***divergence event*** is estimated by the molecules.

There is a "***waiting time***" in between the divergence and the age of the oldest fossil.



Tracy Heath

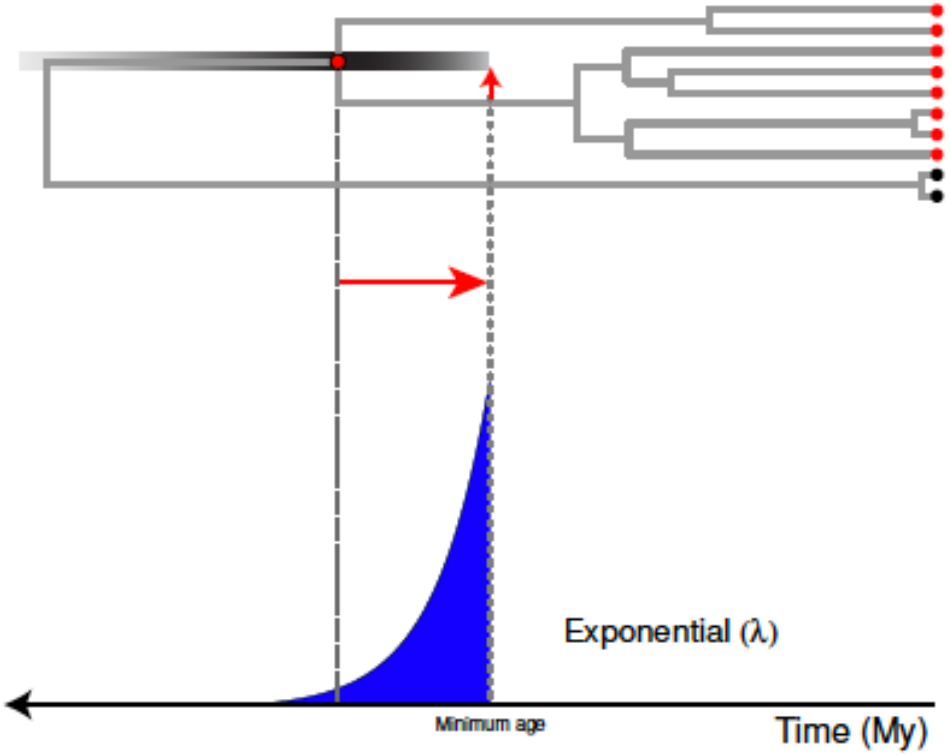
Prior Densities on Calibrated Nodes

It is possible to bias the age of this node with the ***wrong prior***.

In this case, the prior is not diffuse enough.

Too much of the prior distribution is centered close to the age of the fossil.

This ***overly informative*** prior will bias the estimate to be ***too young***.

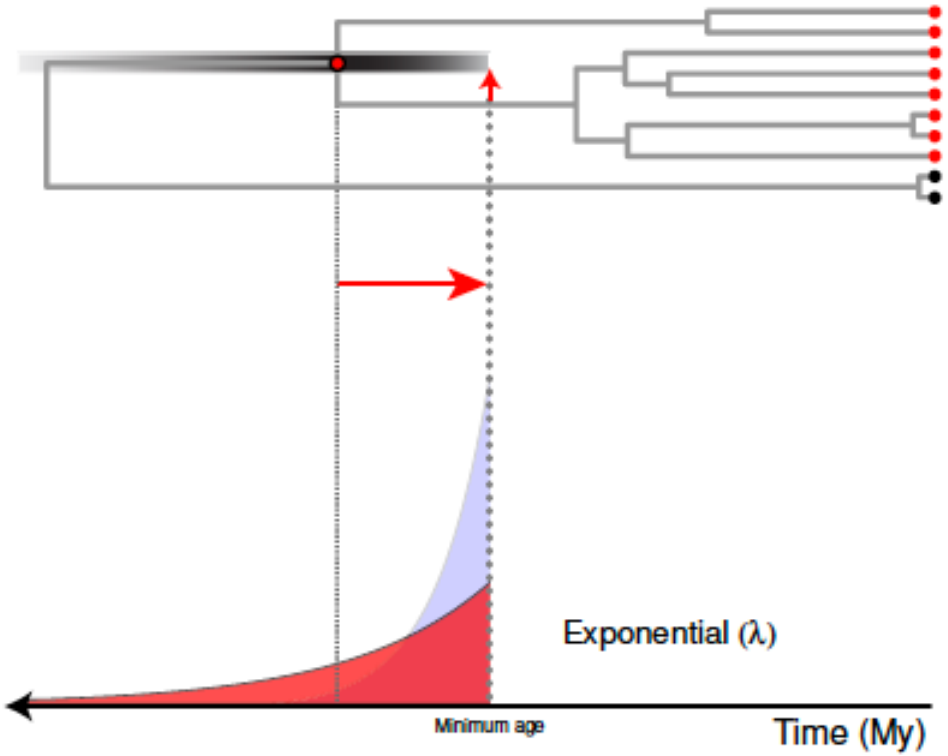


Tracy Heath

Prior Densities on Calibrated Nodes

A more ***diffuse prior*** would have more of it's distribution extending further back in time.

This better captures the ***uncertainty*** in the age of the most recent common ancestor.



Tracy Heath

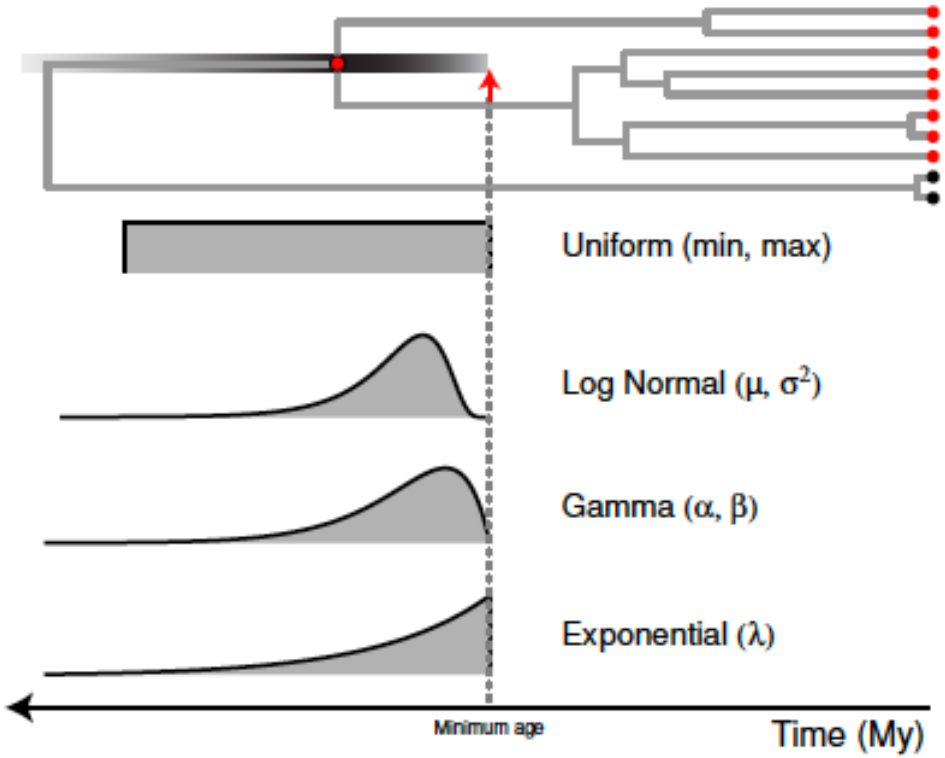
Prior Densities on Calibrated Nodes

Estimates of absolute node ages are driven by the calibration density.

Specifying appropriate priors is a challenge for most molecular biologists.

Uniform priors do not contain enough information, and may lead to skewed posterior estimates.

Often, a ***Log Normal prior*** distribution captures the uncertainty best.



Tracy Heath

Next

Track B does not include a Proposal Assignment, so those will not be graded.

Track B Deliverable 3, I will grade according to the criteria on BBLearn.

Track B Deliverable 4 annotated bibliographies are due 10/23, see rubric for expectations

I will create the peer review matrix necessary for Deliverable 5 and distribute after I receive the Deliverable 4 assignments.