

Molecular Clocks

Simon Ho

School of Biological Sciences



THE UNIVERSITY OF
SYDNEY

Sydney workshop (July)

- 12 workshops
- Several contributors over the years



MOLECULAR ECOLOGY, EVOLUTION,
& PHYLOGENETICS LABORATORY

Sydney workshop (July)

Day 1	Lecture: Introduction to phylogenetic analysis Practical: Sequence alignment using <i>MEGA</i> Lecture: Evolutionary models Practical: Model selection using <i>MEGA</i> Lecture: Phylogenetic methods Practical: Phylogenetic analysis using <i>MEGA</i>
Day 2	Lecture: Bayesian phylogenetic analysis Lecture: Rates and timescales Practical: Molecular-clock analysis using <i>BEAST</i> Lecture: Analysing populations Practical: Inferring population history using <i>BEAST</i>

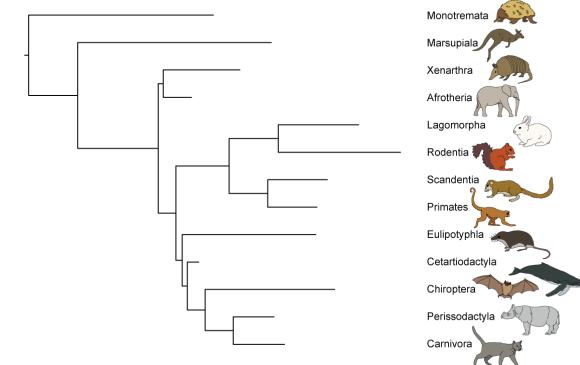
Topics

- Chronograms
- The molecular clock
- Calibrating molecular clocks
- Models of rate variation
- Model selection
- Multiple loci

Chronograms

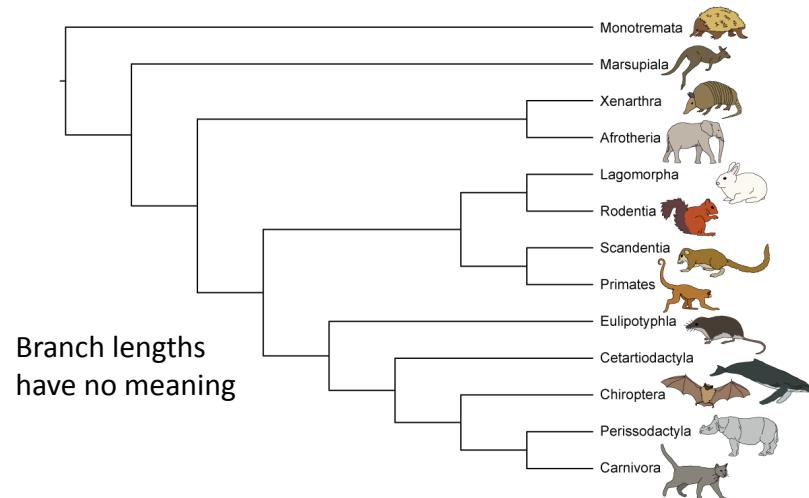
Phylogenetic trees

- A phylogenetic tree has two major components
 - Topology (relationships)
 - Branch lengths (amount of evolutionary change or time)



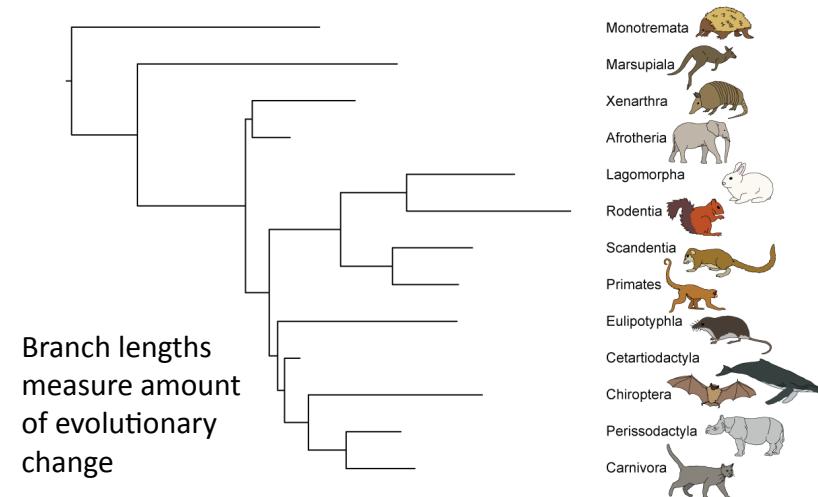
6

Phylogenetic trees: Cladogram



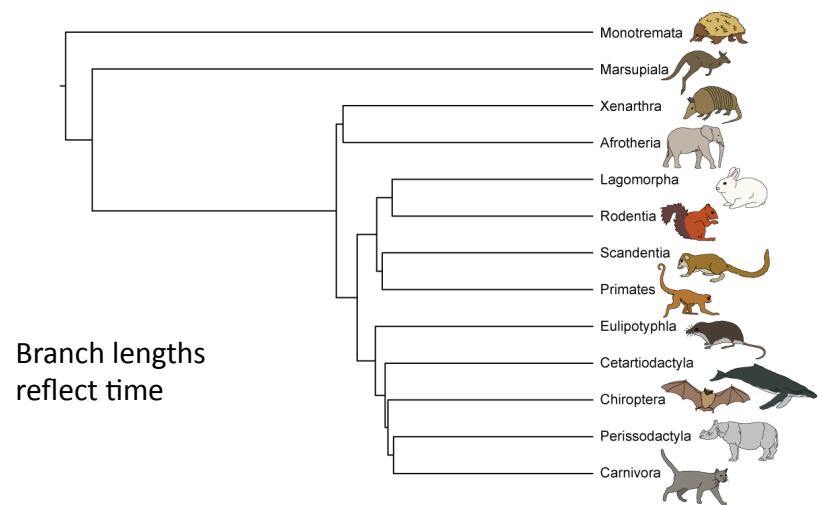
7

Phylogenetic trees: Phylogram

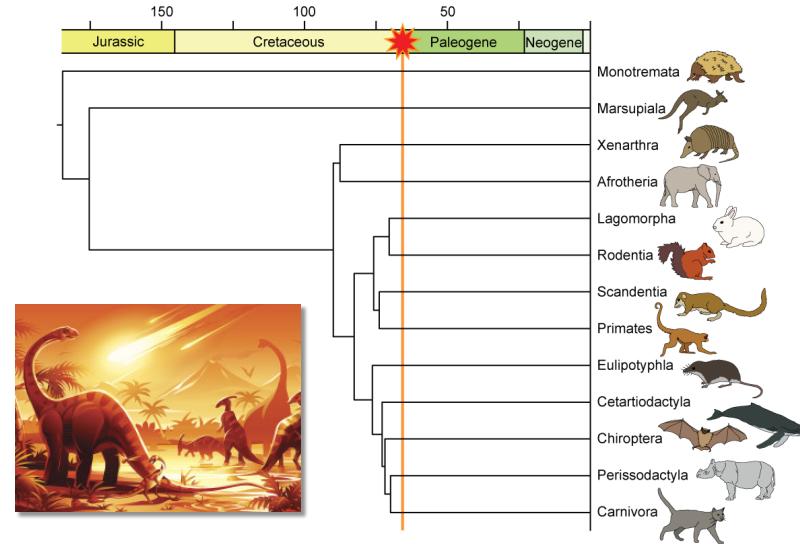


8

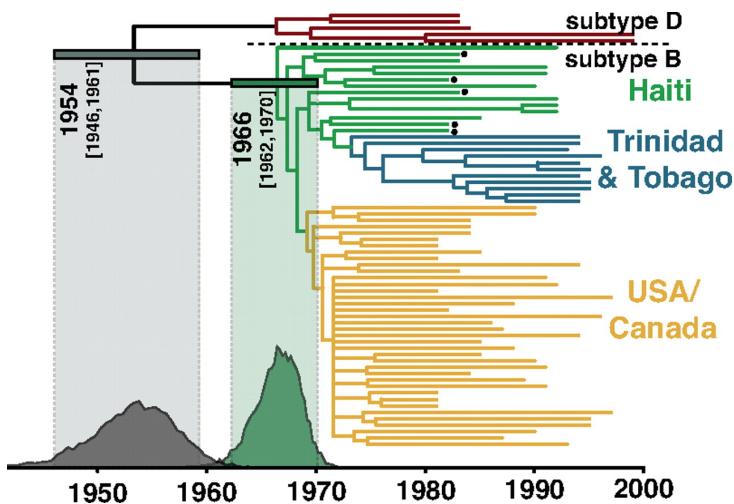
Phylogenetic trees: Chronogram



Ordinal radiation of mammals

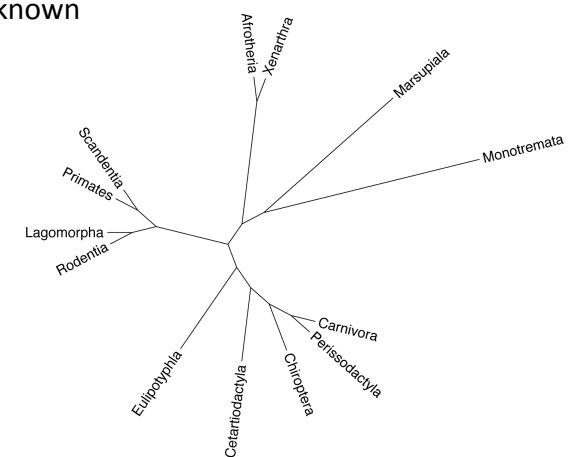


Emergence of HIV/AIDS in the Americas

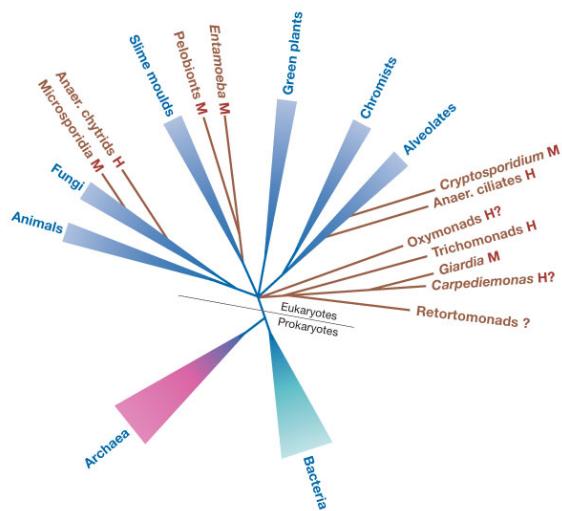


Unrooted trees

- Position of root is unknown
- Usually a phylogram



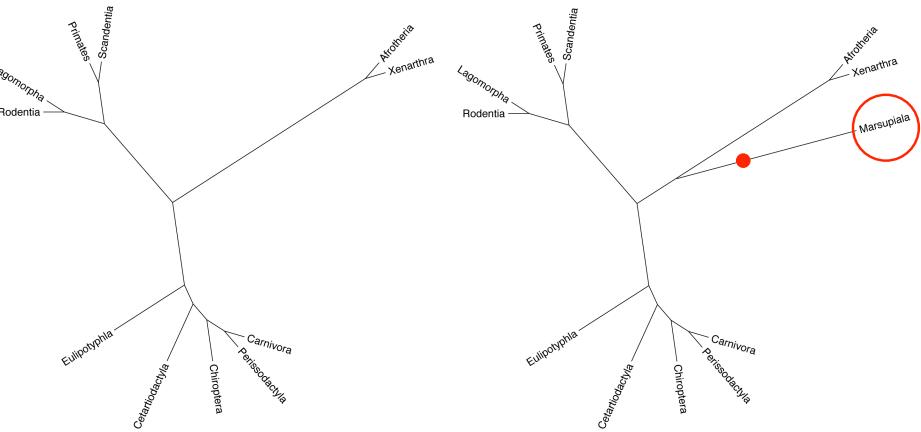
Phylogenetic trees: Unrooted



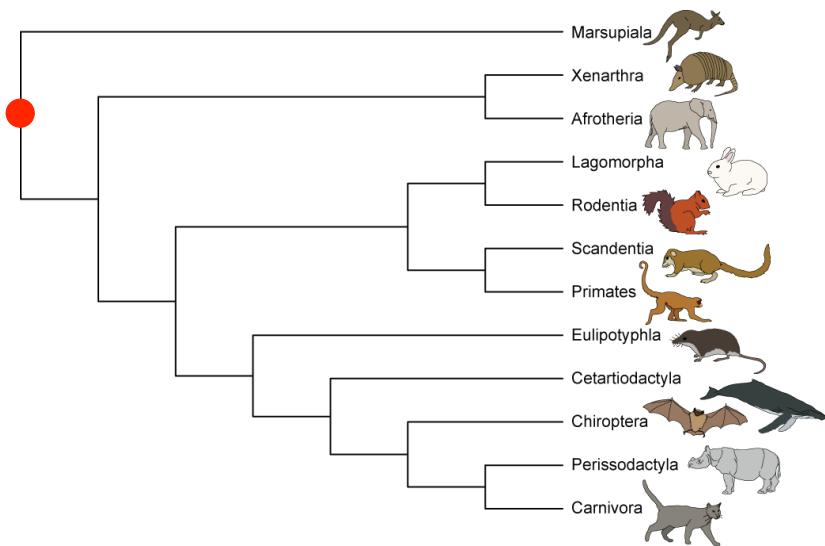
13

Rooting

- Can root a tree by including an outgroup taxon



Rooting



15

Rooting

- Three methods for estimating the root of the tree
- Include an outgroup sequence
- Root at the midpoint of the tree
- Use a molecular clock

16

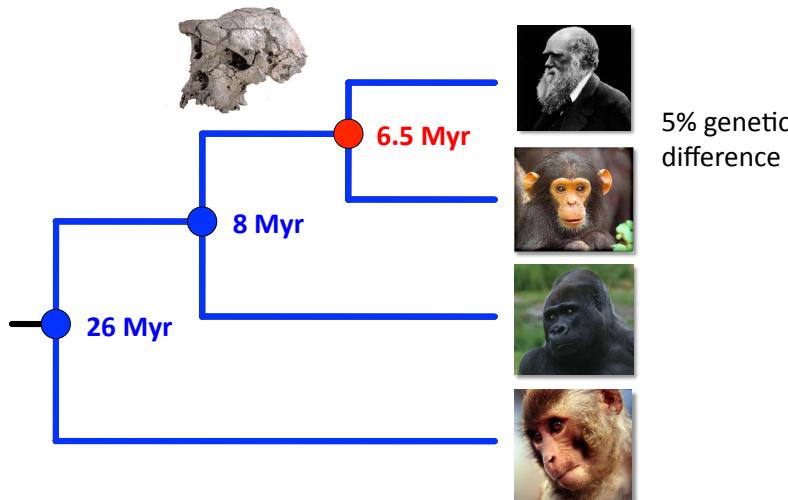
Newick format

- Without branch lengths (cladogram):
 - (Monotremata,(Marsupiala,((Afrotheria,Xenarthra),((Rodentia,Lagomorpha),(Primates,Scandentia)),(Eulipotyphla,(Cetartiodactyla,(Chiroptera,(Carnivora,Perissodactyla)))))));
- With branch lengths (phylogram/chronogram):
 - (Monotremata:12.0,(Marsupiala:11.0,((Afrotheria:1.0,Xenarthra:1.0):9.0,(((Rodentia:1.0,Lagomorpha:1.0):2.0,(Primates:1.0,Scandentia:1.0):2.0):5.0,(Eulipotyphla:4.0,(Cetartiodactyla:3.0,(Chiroptera:2.0,(Carnivora:1.0,Perissodactyla:1.0):1.0):1.0):4.0):2.0):1.0):1.0);

17

The Molecular Clock

The molecular clock



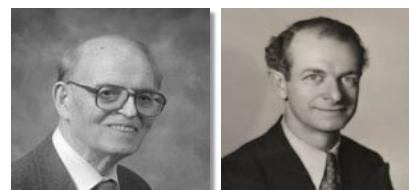
19

The molecular clock

- Zuckerkandl & Pauling (1962)
- Margoliash (1963)
- Doolittle & Blomback (1964)
- Zuckerkandl & Pauling (1965)**

Assumed constant rate among species to estimate timing of globin gene duplications

Proportional relationship between genetic distance and time since divergence



Examined correlation between time and genetic divergence in mammalian fibrinopeptides

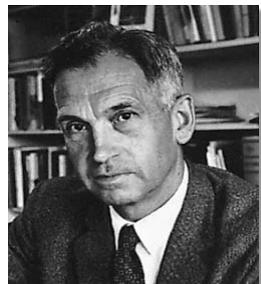
Introduced the term 'molecular evolutionary clock'

20

Immediate criticism

Ernst Mayr (1965)

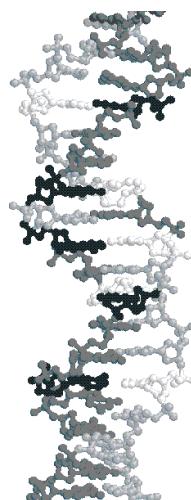
Evolution is too complex and too variable a process, connected with too many factors, for the time dependence of the evolutionary process at the molecular level to be a simple function



21

Mutations

- Mutations in the genome
 - Replication/recombination errors
 - Damage incorrectly repaired
 - Mutagens (e.g., chemicals, radiation)
- Source of variation
 - Occur stochastically
 - Raw material for natural selection



23

Nature of the molecular clock

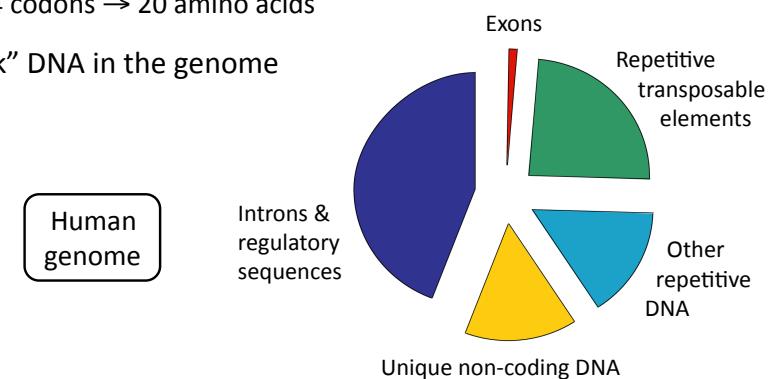
- Metronomic clock (deterministic)
 - Constant **frequency** of mutation events
 - Mutations occur at regular intervals
- Stochastic clock (probabilistic)
 - Constant **probability** of mutation events
 - Variance about the mean
 - Poisson process ($\text{var}/\text{mean} = 1$)
- But why should we expect rates to be constant among species?



22

Effects of mutations

- Redundancy in genetic code
 - 64 codons → 20 amino acids
- “Junk” DNA in the genome

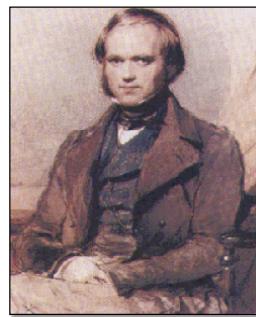


24

Charles Darwin

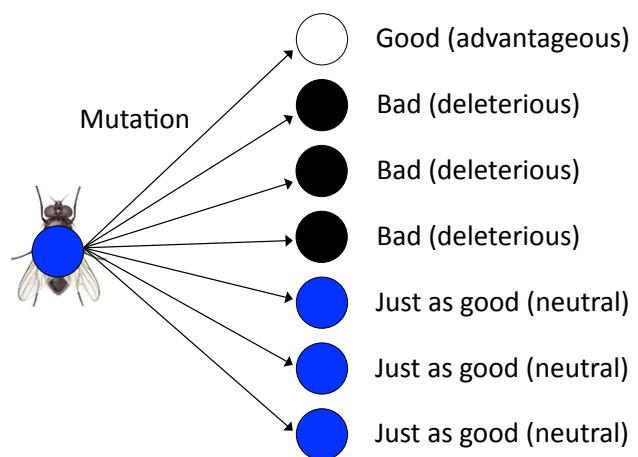
Variation neither useful nor injurious would not be affected by natural selection, and would be left either a fluctuating element, as perhaps we see in certain polymorphic species, or would ultimately become fixed.

On the Origin of Species, 1859



25

Neutral mutations



27

Birth of the neutral theory

Evolutionary Rate at the Molecular Level

1968

by
MOTOO KIMURA
National Institute of Genetics,
Mishima, Japan

Calculating the rate of evolution in terms of nucleotide substitutions seems to give a value so high that many of the mutations involved must be neutral ones.



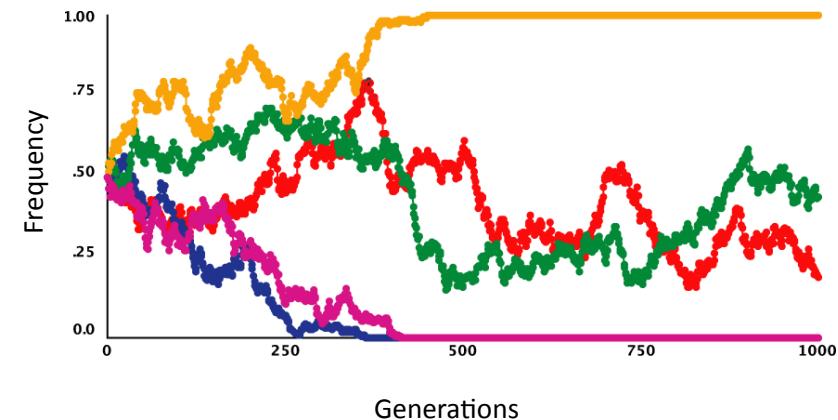
Non-Darwinian Evolution

Most evolutionary change in proteins may be due to neutral mutations and genetic drift.

Jack Lester King and Thomas H. Jukes
1969

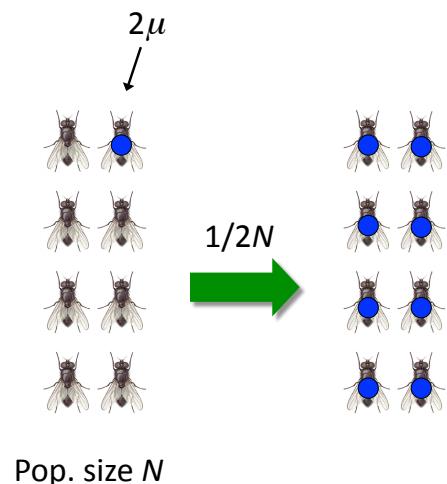
26

Genetic drift



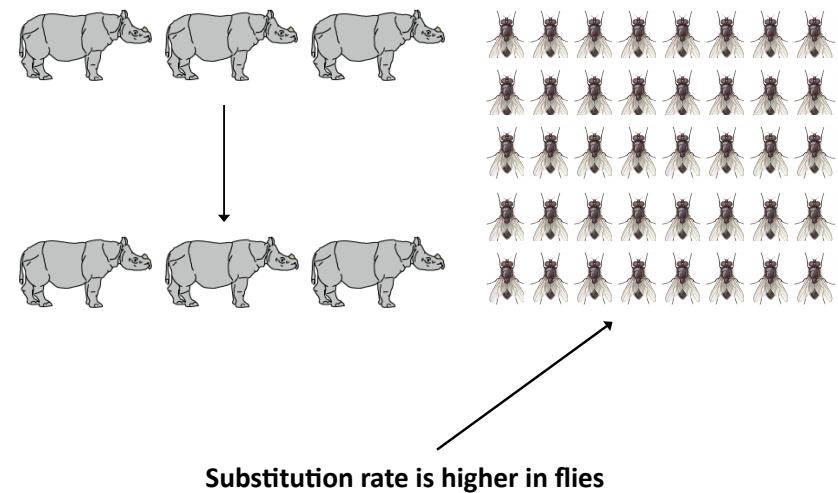
28

Neutral rate



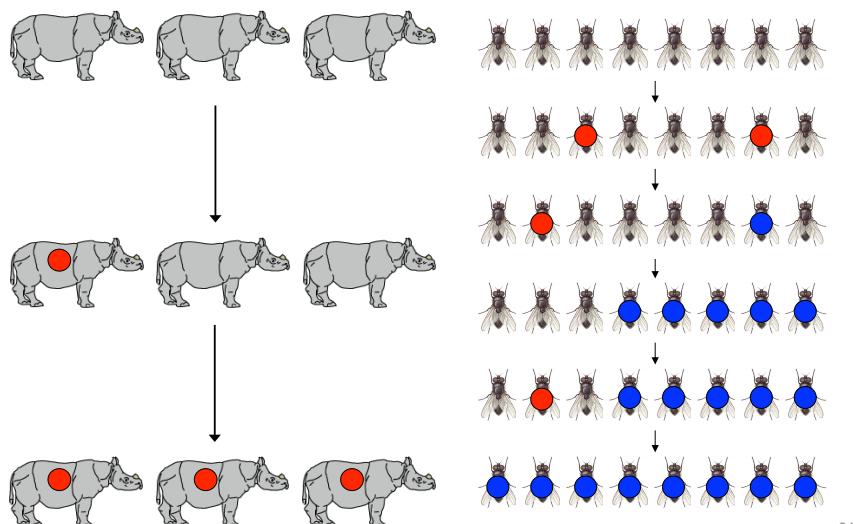
$$\begin{aligned}\text{Substitution rate } (k) &= \text{mutation rate} \times \\ &\quad \text{Pr(fixation)} \\ &= 2\mu N \times (1/2N) \\ &= \mu \\ &= \text{mutation rate}\end{aligned}$$

Generation-time effect



30

Population-size dependence

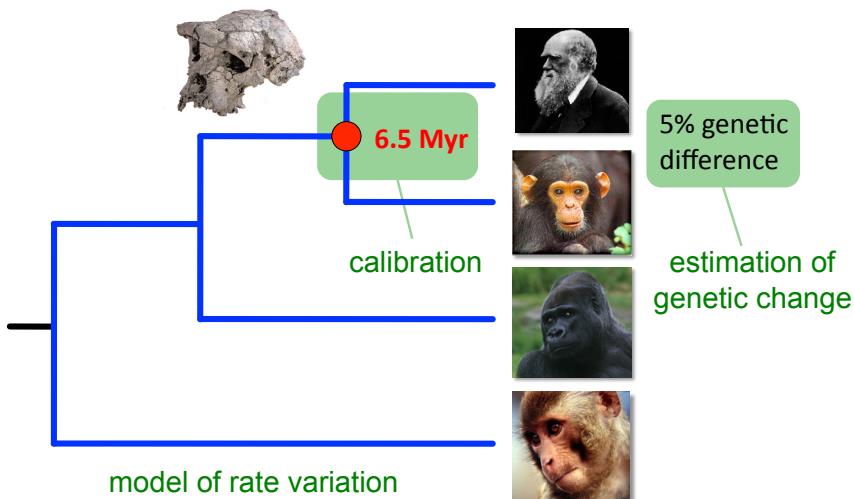


Population-size dependence

- | | |
|--|--|
| <p>Few mutations</p> <ul style="list-style-type: none"> ▪ Few individuals ▪ Long generations <p>High probability of fixation</p> <ul style="list-style-type: none"> ▪ Genetic drift dominates | <p>Many mutations</p> <ul style="list-style-type: none"> ▪ Many individuals ▪ Short generations <p>Low probability of fixation</p> <ul style="list-style-type: none"> ▪ Natural selection dominates |
|--|--|
- Effects cancel out → substitution rate independent of population size and generation time**

32

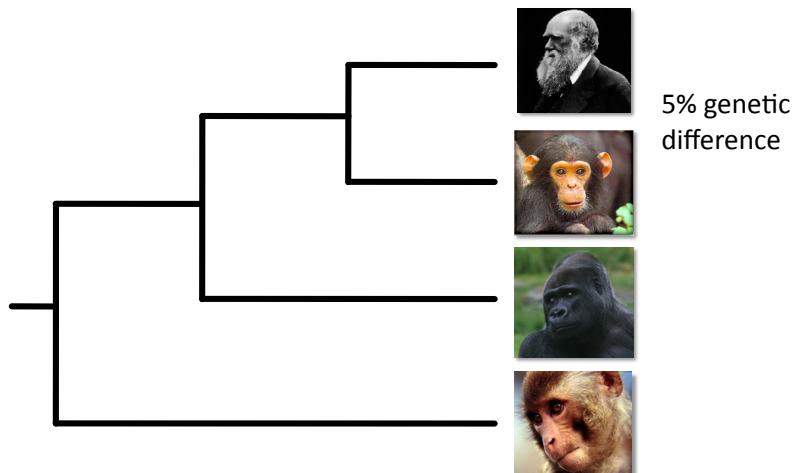
Sources of error



Calibrating Molecular Clocks

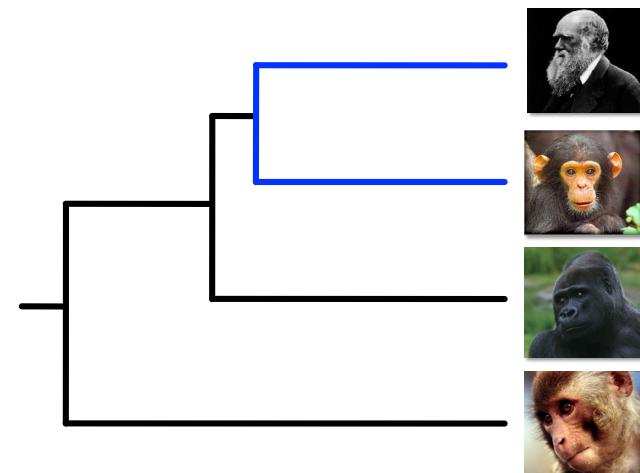
33

Calibrating the molecular clock



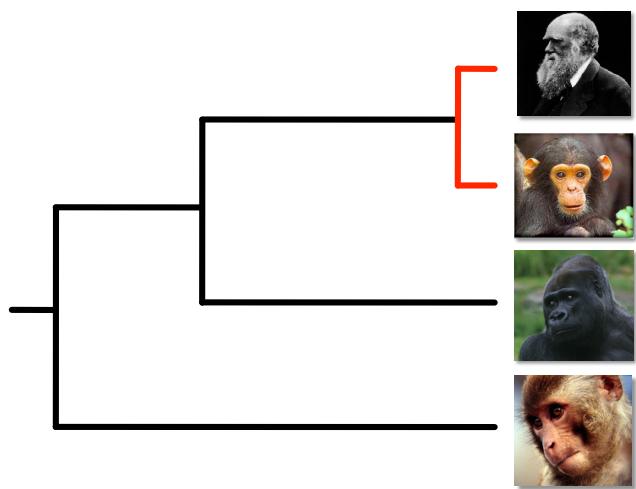
35

Calibrating the molecular clock



36

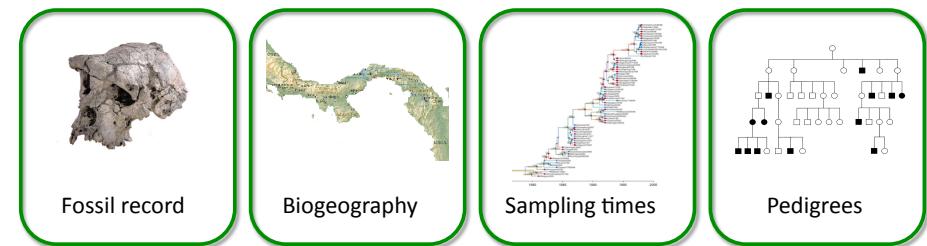
Calibrating the molecular clock



37

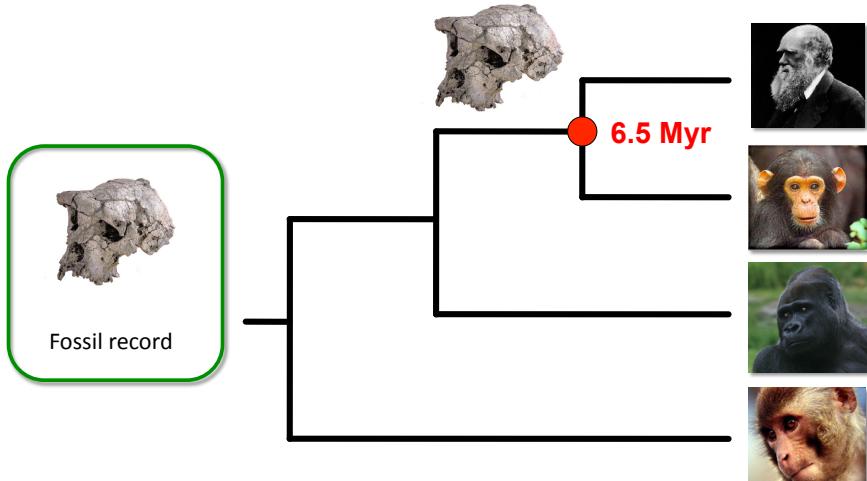
Calibrating information

- Information about the rate
 - Substitution rate obtained from an independent study
- Information about time



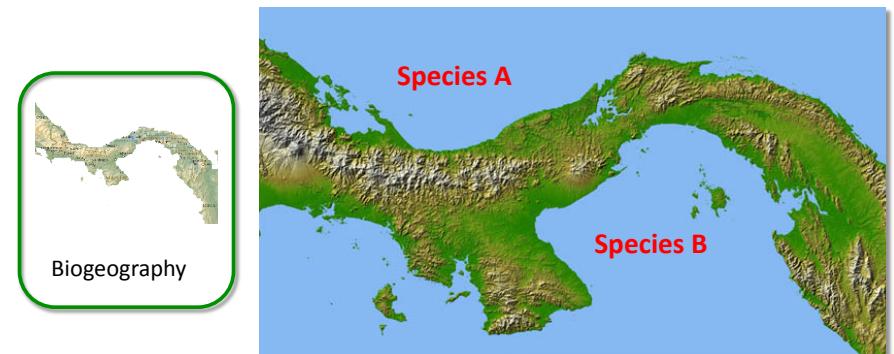
38

Calibration: Fossil record



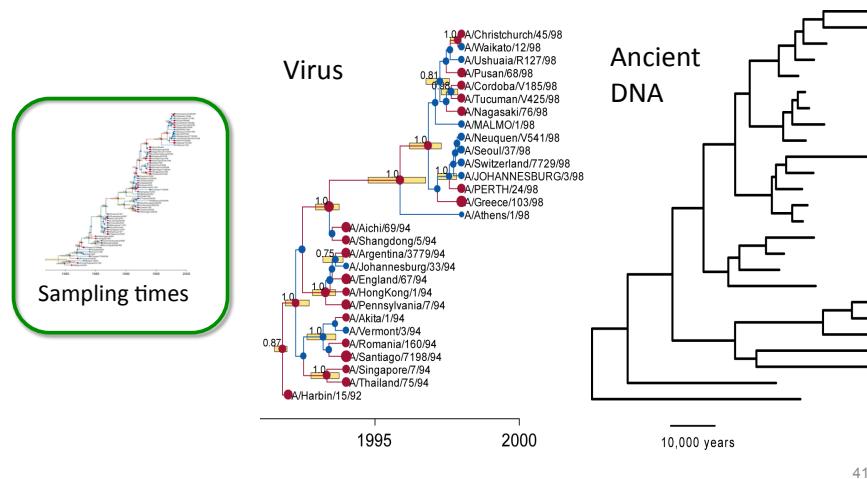
39

Calibration: Biogeography



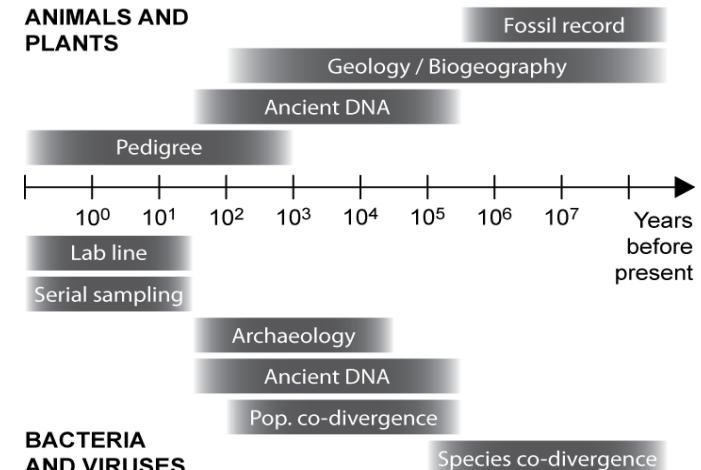
40

Calibration: Sampling times



41

Calibrations

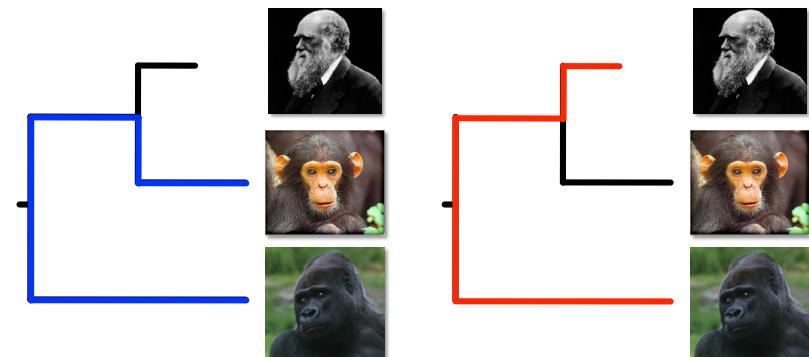


42

Models of Rate Variation

Testing for clocklike evolution

- Relative-rates test (Fitch, 1976)



44

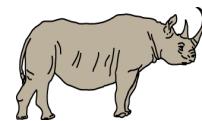
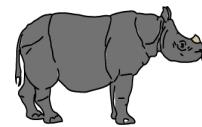
Testing for clocklike evolution

- Likelihood-ratio test
 - Strict clock vs unconstrained model
- Bayes factors
 - Comparing strict clock against other models
 - Comparing numbers of rate categories in DPP model
 - Comparing numbers of distinct rates in random local clock

45

Departures from the clock

- Rates vary among lineages
 - Differences in mutation rates
 - Differences in strength and direction of selection
 - Differences in population size
- Predictors of rates:
 - Longevity (mammals)
 - Height (flowering plants)



46

Why keep the molecular clock?

- The behaviour of most real sequences does not satisfy the assumption of a strict molecular clock

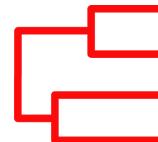
Bromham & Penny (2003):

The molecular clock is an irreplaceable source of information in evolutionary biology and it would be foolish to abandon it altogether

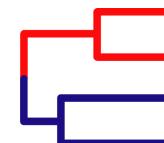
47

Molecular-clock models

- Strict or 'global' molecular clock



- Multi-rate clocks



- Relaxed clocks



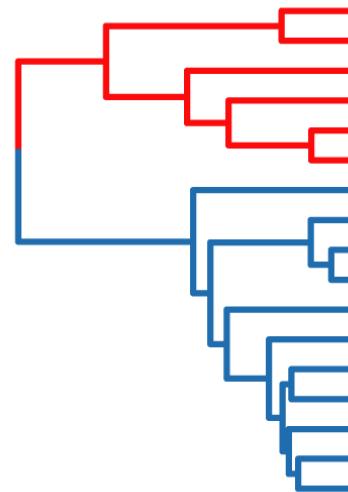
48

Multi-rate clocks

- Small number of rates
 - More than 1 rate (i.e., not a strict clock)
 - Fewer than number of branches (i.e., not a relaxed clock)
- Local clock
 - Same rate shared by neighbouring branches
- Discrete clock
 - Small number of branch rates, distributed across tree

49

Local clocks



50

Local clocks

- User-defined local clock
 - Fixed tree topology
- Random local clock
 - Each branch has a probability of inheriting rate from ancestor
 - Tree estimated

51

Discrete clocks

- User-defined discrete clocks
 - Fixed tree topology
- Dirichlet process prior (DPP) model
 - Branch rates drawn from gamma distribution
 - Concentration parameter (α) governs number of rate categories and number of branches assigned to each rate category

52

Relaxed clocks

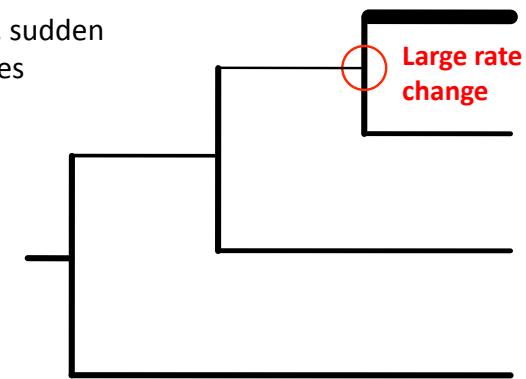
- We know that life-history characteristics:
 - Have effects on rates of molecular evolution
 - Are usually heritable to some degree
- Treat molecular rate as a heritable trait
- Relaxed clocks generally assume that closely related species share similar rates



53

Likelihood-based relaxed clocks

- Allow a different rate in each branch
- Penalise large, sudden changes in rates



$$PL = \text{Likelihood} - \lambda * (\text{roughness penalty})$$

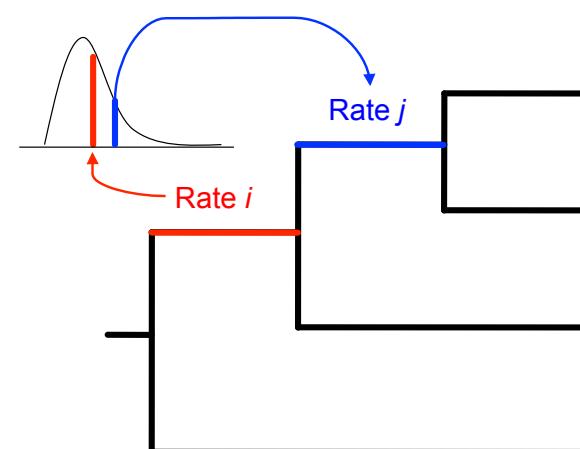
54

Bayesian relaxed clocks

- Allow a different rate in each branch
- Statistical models of rates among branches
- Rates can be autocorrelated or uncorrelated
 - **Autocorrelated:** rates in neighbouring branches are related
 - **Uncorrelated:** rates identically and independently distributed among branches

55

Autocorrelated relaxed clock



56

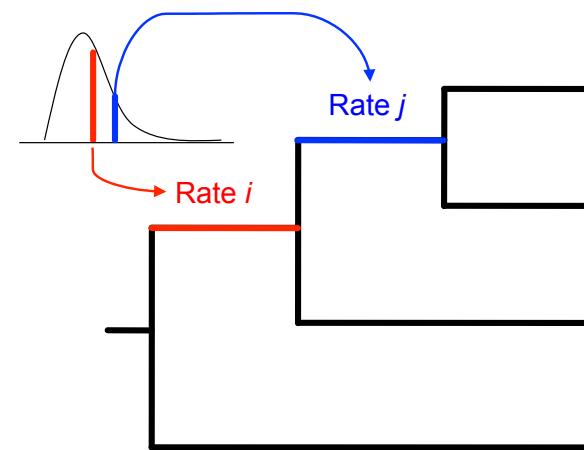
Autocorrelated relaxed clocks

- Continuous
 - $r_j \sim \text{Lognormal}(\mu_i, \sigma^2 t_i)$
 - $r_j \sim \text{Gamma}(\alpha, \lambda)$
 - Cox-Ingersoll-Ross
 - Ornstein-Uhlenbeck
- Episodic
 - $r_j \sim \text{Exponential}(1/r_i)$

Variance of distribution depends
on time duration of branch

57

Uncorrelated relaxed clock



58

Uncorrelated relaxed clocks

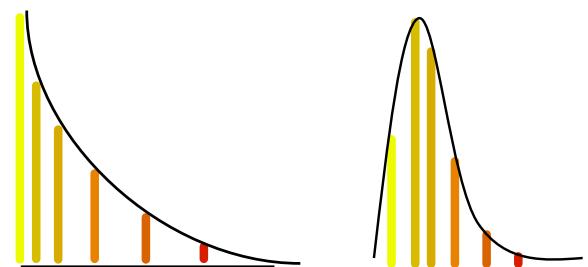
- Lognormal
- Gamma
- Exponential
- White noise / IGR

Branch rates
increasingly
independent

59

Uncorrelated relaxed clock

- Models available in BEAST
 - **Exponential distribution**
Most rates are quite low
 - **Lognormal distribution**
Most rates cluster around the mean



60

Lognormal uncorrelated relaxed clock

- In the uncorrelated lognormal relaxed clock, two statistics can be obtained:
 1. **Coefficient of variation of rates**
Measures the rate variation among branches
A value of 0 indicates clocklike evolution
 2. **Covariance of rates**
Measures autocorrelation of rates between adjacent branches

61

Model Selection

Selecting a clock model

- Bayes factor
 - Harmonic-mean estimator (deprecated) – BEAST/Tracer
 - Stepping-stone and path sampling – BEAST, MrBayes, RevBayes
- Treat model as a variable
 - Model averaging (Li & Drummond 2012) – BEAST

63

Multiple Loci

Multigene data sets

- Analysis of large sequence alignments is now routine
- Multiple genes from multiple genomes
(nuclear, mitochondrial, chloroplast)

	Mitochondrial		Nuclear		
	Gene A	Gene B	Gene C	Gene D	Gene E
Species 1	[]	[]	[]	[]	[]
Species 2	[]	[]	[]	[]	[]
Species 3	[]	[]	[]	[]	[]
Species 4	[]	[]	[]	[]	[]
Species 5	[]	[]	[]	[]	[]

65

Partitioning scheme

- Separate substitution model for each genome

	Mitochondrial		Nuclear		
	Gene A	Gene B	Gene C	Gene D	Gene E
Species 1	[orange]	[orange]	[green]	[green]	[green]
Species 2	[orange]	[orange]	[green]	[green]	[green]
Species 3	[orange]	[orange]	[green]	[green]	[green]
Species 4	[orange]	[orange]	[green]	[green]	[green]
Species 5	[orange]	[orange]	[green]	[green]	[green]

66

Partitioning scheme

- Separate substitution model for each genome
- Separate substitution model for each gene

	Mitochondrial		Nuclear		
	Gene A	Gene B	Gene C	Gene D	Gene E
Species 1	[orange]	[red]	[green]	[olive]	[blue]
Species 2	[orange]	[red]	[green]	[olive]	[blue]
Species 3	[orange]	[red]	[green]	[olive]	[blue]
Species 4	[orange]	[red]	[green]	[olive]	[blue]
Species 5	[orange]	[red]	[green]	[olive]	[blue]

67

PartitionFinder

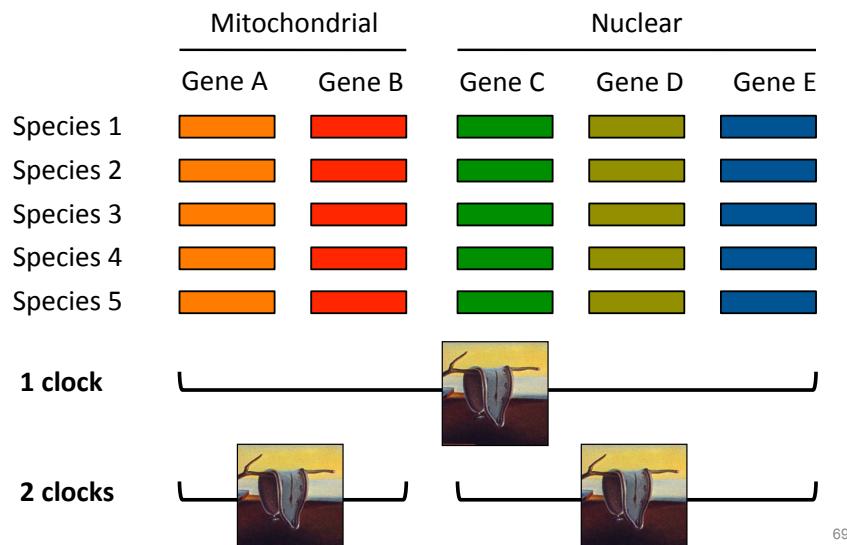
- Too many possible partitioning schemes
 - 15 schemes for 4 genes
 - 52 schemes for 5 genes
 - 203 schemes for 6 genes

PartitionFinder: Combined Selection of Partitioning Schemes and Substitution Models for Phylogenetic Analyses

Robert Lanfear,^{*1} Brett Calcott,^{1,2} Simon Y. W. Ho,³ and Stephane Guindon⁴

68

Multiple clock models



69

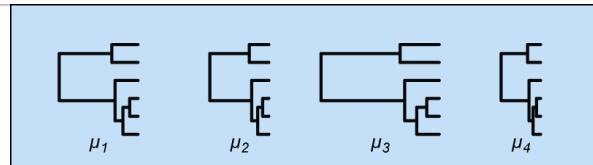
Multiple clock models

- Simple approach
 - Single relaxed-clock model
 - Scaling factor for each gene
- Multiple relaxed-clock models
 - Separate relaxed-clock model for each data partition
 - Potentially many parameters

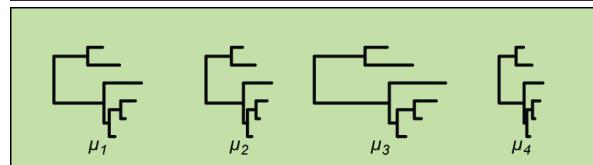
70

Multiple clock models

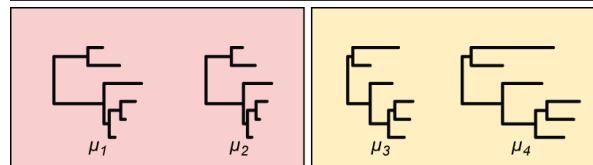
- Among loci
(gene effects)



- Among lineages
(lineage effects)

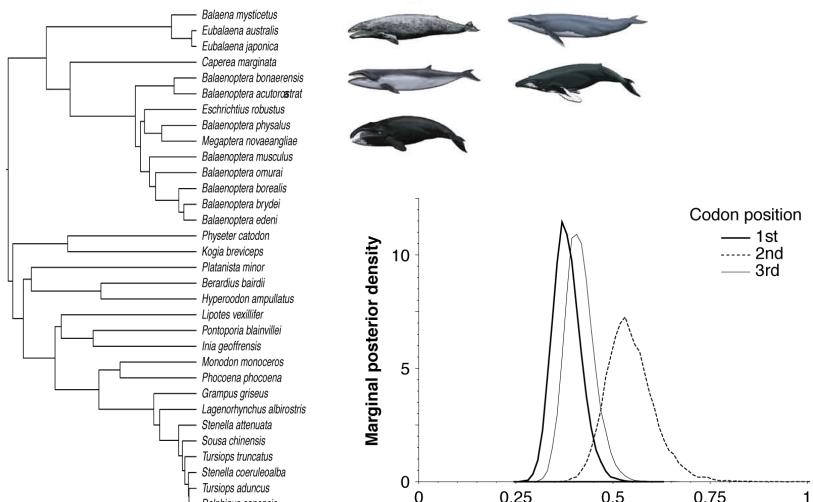


- Gene by lineage interaction
(residual effects)



71

Multiple clock models



72

Selecting the number of clocks

- Assign clock models according to genome, gene, codon position, etc.
- Many possible combinations
- Not feasible to use Bayes factors for model selection

BIOINFORMATICS APPLICATIONS NOTE

Vol. 30 no. 7 2014, pages 1017–1019
doi:10.1093/bioinformatics/btt665

Phylogenetics

Advance Access publication November 14, 2013

ClockstaR: choosing the number of relaxed-clock models in molecular phylogenetic analysis

Sebastián Duchêne^{*,†}, Martyna Molak and Simon Y. W. Ho[†]

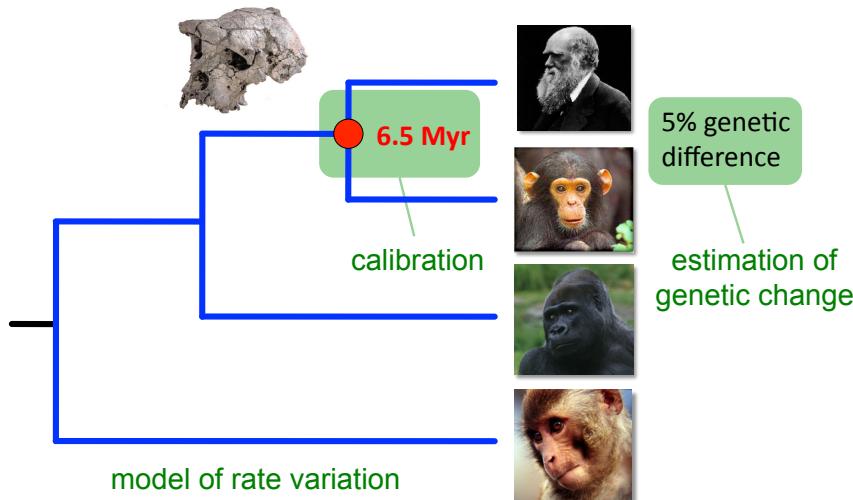
73

74

ClockstaR

- Fixed tree topology
- Estimate branch lengths for each gene tree
- Identify clusters of gene trees that have similar branch lengths
- Assign a separate relaxed-clock model to each cluster of gene trees

Sources of error



75

Useful references

- Molecular-clock methods for estimating evolutionary rates and timescales
Ho & Duchene (2014). *Molecular Ecology*, in press.
- The changing face of the molecular evolutionary clock
Ho (2014) *Trends in Ecology and Evolution*, 29: 496–503.
- Molecular Evolution: A Statistical Approach
Yang (2014) Oxford University Press, Oxford.
- Bayesian inference of species divergence times
Heath & Moore (2014) In: *Bayesian Phylogenetics: Methods Algorithms, and Applications* (eds Chen, Kuo, & Lewis), pp. 277–318. CRC Press, Boca Raton, Florida.

76