Phylogenetics and Molecular Evolution/Filogenética e Evolução Molecular

Octávio S. Paulo
Computational Biology and Population Genomics Group (CoBiG2)

Introdução aos relógios moleculares

Sumário:
Introdução aos relógios moleculares e à sua implementação.

Molecular Clocks

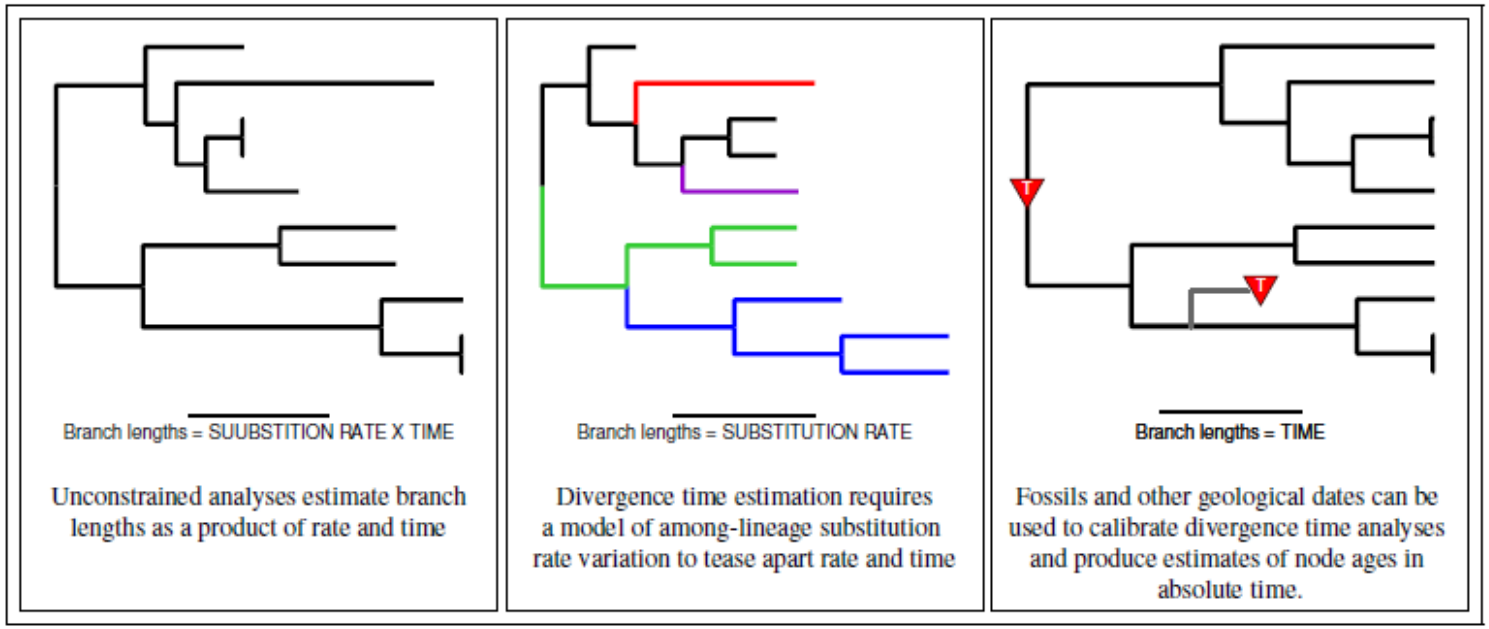


Figure 1: Estimating branch lengths in units of time requires a model of lineage-specific rate variation, a model for describing the distribution of speciation events over time, and external information to calibrate the tree.

Relógios moleculares - Neutral rate

Mutation rate= $2N\mu$

Fixation probability= $P = \frac{1}{2N}$

Substitution rate= $k = P2N\mu$ $k = \mu$

Molecular clock

Box 1 | **The clock and the neutral theory of molecular evolution**

Zuckerkandl and Pauling provided a justification for the molecular clock by suggesting that amino acid changes that accumulate between species are mostly those with little or no effect on the structure and function of the protein, thus reflecting the background mutational process at the DNA level[1]. This hypothesis was formalized by Kimura[106] and by King and Jukes[107] in the neutral theory of molecular evolution, which asserts that most of the genetic variation that we observe (either polymorphisms within species or divergence between species) is due to chance fixation of selectively neutral mutations, rather than due to fixation of advantageous mutations driven by natural selection[6]. Thus, the molecular clock was soon entwined in the controversy surrounding the neutral theory, which was initially proposed to explain the surprising finding of high levels of polymorphism in natural populations[108,109]. If molecular evolution is dominated by neutral mutations, which have little influence on the survival or reproduction of the individual, then an approximately constant rate of evolution is plausible. Indeed, under this theory, the rate of molecular evolution is equal to the neutral mutation rate, which can be assumed to be similar among species with similar life histories.

Most mutations that arise in a generation in a large population are lost by chance within a small number of generations. This is true not only for neutral and deleterious mutations, but also for advantageous mutations unless the advantage is extremely large. For example, if a mutation offers a 1% selective advantage (which is a very large advantage), there is only about 2% chance that the mutation will eventually spread through the whole population[110]. The minority of mutations that are eventually fixed in the population are known as substitutions. Viewed over a very long timescale, this process of new mutations reaching fixation, replacing previous wild-type alleles, is the process of molecular evolution. Suppose the total mutation rate is μ per generation, and a fraction f_0 of the mutations is neutral. The rest of the mutations are deleterious and are removed by natural selection, and do not contribute to the evolutionary process. There are $2N \times \mu f_0$ neutral mutations per generation for a diploid population of size N. The chance that a neutral mutation will eventually reach fixation is $1/(2N)$, because there are $2N$ alleles in the population and each has the same chance of reaching fixation. The molecular substitution rate per generation r (that is, the number of mutations per generation that reach fixation in the population) is thus equal to the number of new neutral mutations produced in each generation multiplied by the probability that they will eventually reach fixation; that is:

$$r = 2N\mu f_0 \times 1/(2N) = \mu f_0 \quad (1)$$

In other words, the substitution rate is equal to the neutral mutation rate (μf_0)[111]. According to this neutral mutation-random drift theory (or the neutral theory), the rate of molecular evolution reflects the neutral mutation rate independently of the population size. Thus, the molecular clock holds if μ and f_0 are approximately constant through time and similar among closely related species.

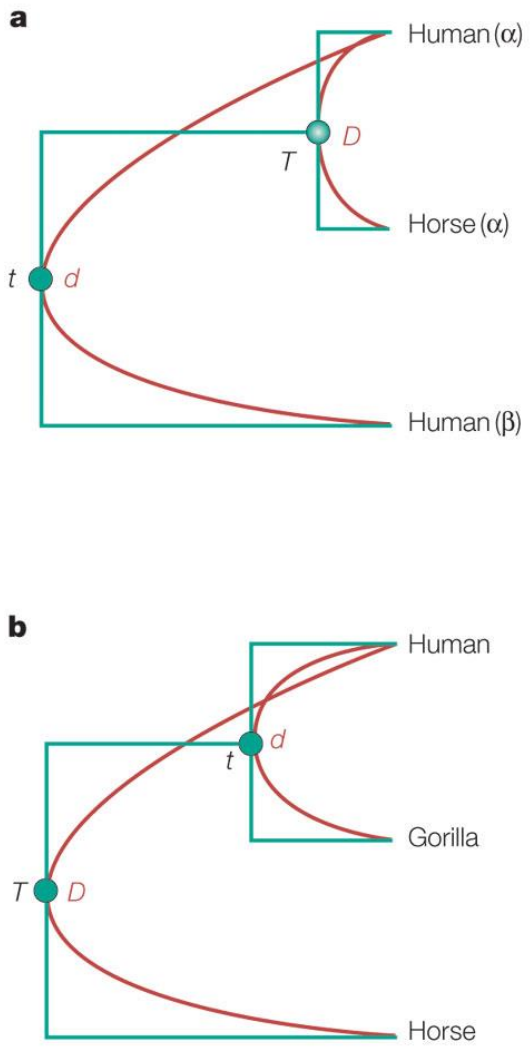
Hence, the neutral theory offers an explanation for the molecular clock, and for a time the clock was considered the most important evidence supporting the neutral theory[6]. Proteins with different functional constraints may have different proportions of neutral mutations (f_0), so that they have different rates of neutral mutation and their clocks tick at different rates. Extensive reviews of the clock-neutral theory controversy are given elsewhere[6,7,112].

Molecular clock

The hypothesis that the rate of molecular evolution is constant over time or among species. Thus, mutations accumulate at a uniform rate after species divergence, keeping time like a timepiece.

Reis et. al 2015 NRG

The earliest uses of the molecular clock



In 1962, Zuckerkandl and Pauling estimated the time of divergence of four members of the haemoglobin gene family (α ,β , γ and δ) by assuming an approximate molecular clock. This was calibrated using the number of observed sequence differences (*D*) between the horse and human α - haemoglobin proteins and the divergence time between the two species (*T*), which is based on the fossil record.

Nature Reviews | Genetics

The earliest uses of the molecular clock

They took a pair-wise approach to estimating divergence times, which is shown schematically for α- and β -haemoglobin in panel **a**. The molecular-clock calibration was carried out by dividing twice the known divergence time by the amount of sequence divergence (2*T*/*D*); the factor of 2 is used here because *D* is equal to the sum of divergence from the common ancestor to the two descendents. This calibration was then used to convert other measurements of protein sequence differences to time. For example, the formula $t = d\ (T/D)$ gives the time when the α - and β -chains diverged, where *d* is the amount of sequence difference between α - and β -chains in humans. The time estimate obtained will have the same units as the time used for clock calibrations (in this case, millions of years).

Zuckerkandl and Pauling also estimated the timing of the human–gorilla divergence using α - and β -chains separately (panel **b**). They calculated the molecular-clock calibration to be 11 to 18 million years (Myr) per amino-acid substitution, based on the observation of 18 differences between human and horse α -haemoglobin proteins and the assumption that these two species diverged 100–160 million years ago (Mya). Using an average calibration of 14.5 Myr per substitution, the human–gorilla divergence was dated to have occurred 14.5 and 7.25 Mya by α - and β -chains, because human and gorilla show two and one differences in these chains, respectively. Therefore, Zuckerkandl and Pauling[3] reported a mean date of 11 Mya for the human–gorilla divergence from an analysis of the two proteins. One year later, Margoliash[5] used the same calibration point to estimate multiple species divergence times. These estimates were based on single, slowly evolving proteins and were therefore not very accurate. In 1965, Zuckerkandl and Pauling[9] predicted that the accuracy of molecular clocks would be improved by using many proteins of different types. Over the past decade, a large number of proteins have been analysed to estimate divergence times among the principal groups of mammals and among animal phyla[35, 39, 60].

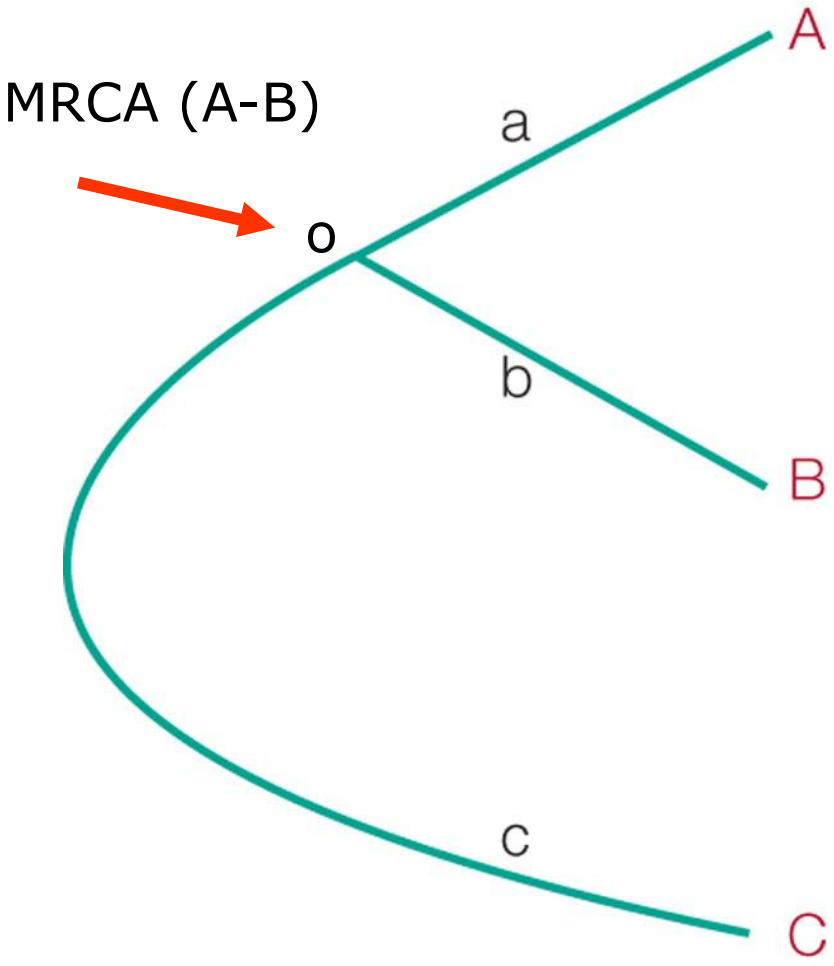
Strick clock

The simplest clock model is called the strict clock (SC) model and assumes a constant rate μ of mutation on all the branches (Zuckerkandl and Pauling 1962).

Therefore, a branch of duration li will contain a number of mutations xi which is Poisson distributed with parameter μli.

The SC model (Zuckerkandl and Pauling 1962) has just a single parameter μ and this simplicity is attractive, but it is often too simple because of variations in the mutation rate from one lineage to another.

Strick clock



$$d_{AB} = d_{OA} + d_{OB}$$

$$d_{AC} = d_{OA} + d_{OC}$$

$$d_{BC} = d_{OB} + d_{OC}$$

$$d_{OA} - d_{OB} = d_{AC} - d_{BC}$$

Nature Reviews | Genetics

Test clock

1- Relative ratio test

2- Tajima test

3- Likelihood ratio test

Problemas com o relógio molecular

-Sobre dispersão , processo aleatório, overdispersed Poisson distribution – imprecisão

-Pouco exacto – variação associada provoca enviesamentos nas estimativas

-Tão correcto quanto nossa capacidade de determinar as distancias reais entre as sequencias

- Erro associado aos pontos de calibração geológicos

-Não universal

-Diferentes genes diferentes velocidades

-Efeitos de linhagem

Efeitos de linhagem – fontes de variação

1- Taxas de mutação

a) Eficácia de reparação
b) Taxa metabólica
c) Tempo de geração

2 –Tamanho da população efectiva

3 – Coeficientes Selectivos

1-Taxas de mutação – Eficiencia da reparação (a)

Mutações ocorrem por: erros de replicação e danos não reparados

Associados aos mecanismos de reparação – bateria de enzymas e à eficiencia destes (a)

Roedores menos eficentes que os hominidios

mtDNA menos eficiente que o DNA nuclear

A própria eficiencia de reparação está sujeita a mutação e aos mecanismo de selecção natural e deriva-

nem no máximo da eficácia nem minimo nivel tolerado

1-Taxas de mutação - Taxa metabólica (b)

Taxa metabólica – radicais de oxigénio

Ectotermicos<Endotérmicos

Pequeno tamanho corporal > grande tamanho corporal

mtDNA>Nuclear

1-Taxas de mutação - Tempo de geração (c)

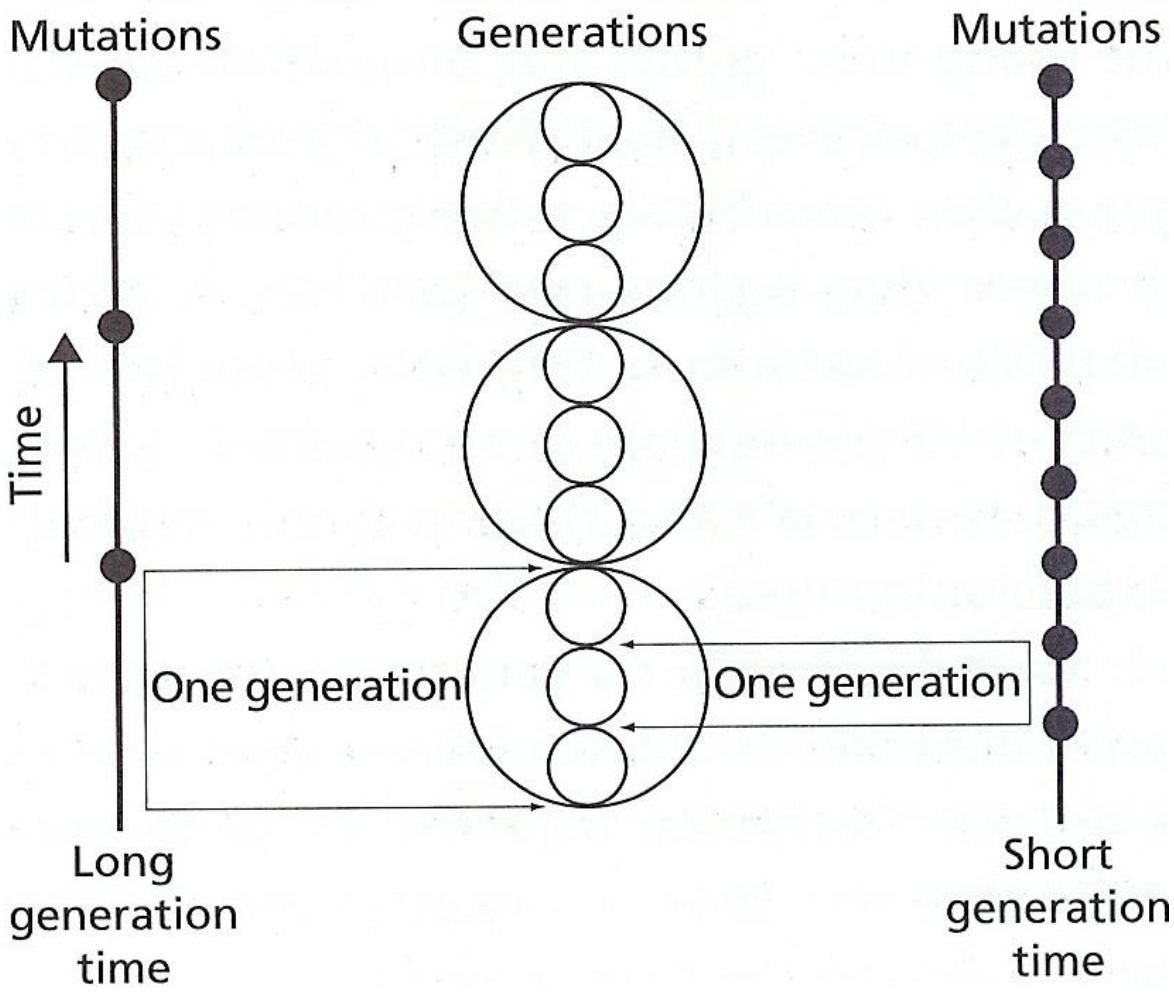
Maior ou menor número de replicações por unidade de tempo

Correlacionado com o tamanho corporal

e relacionado coma taxa de mutação em vertebrados

Gerações curtas>gerações longas

Neutralist model – molecular clock



$$k = \mu / g$$

Tamanho corporal

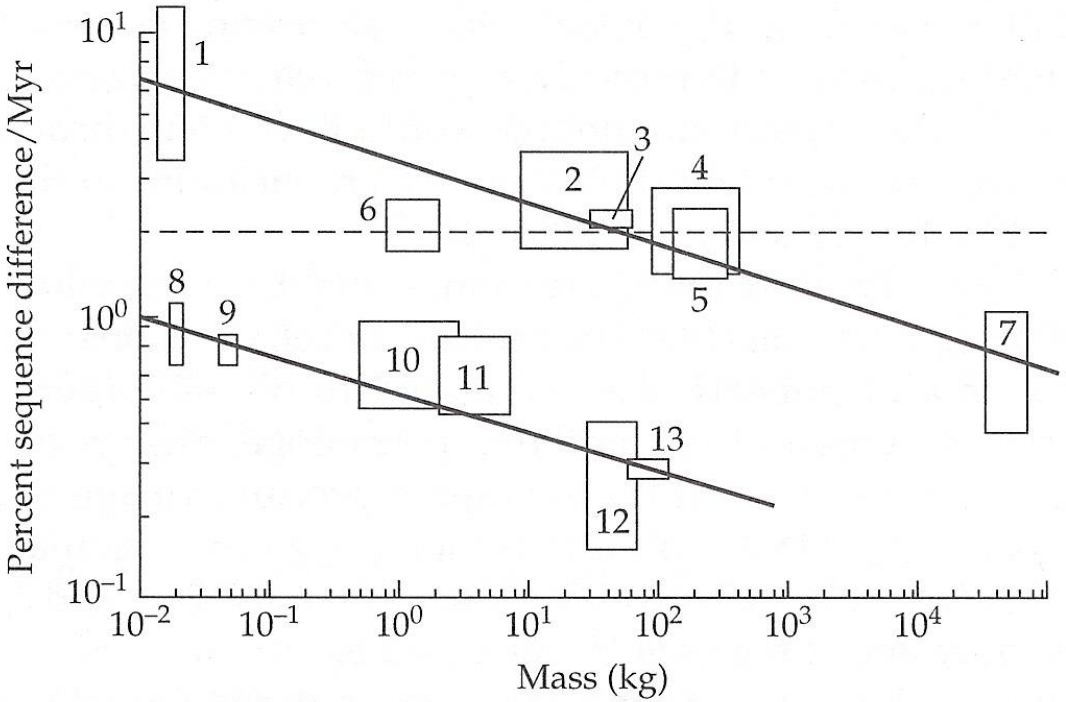
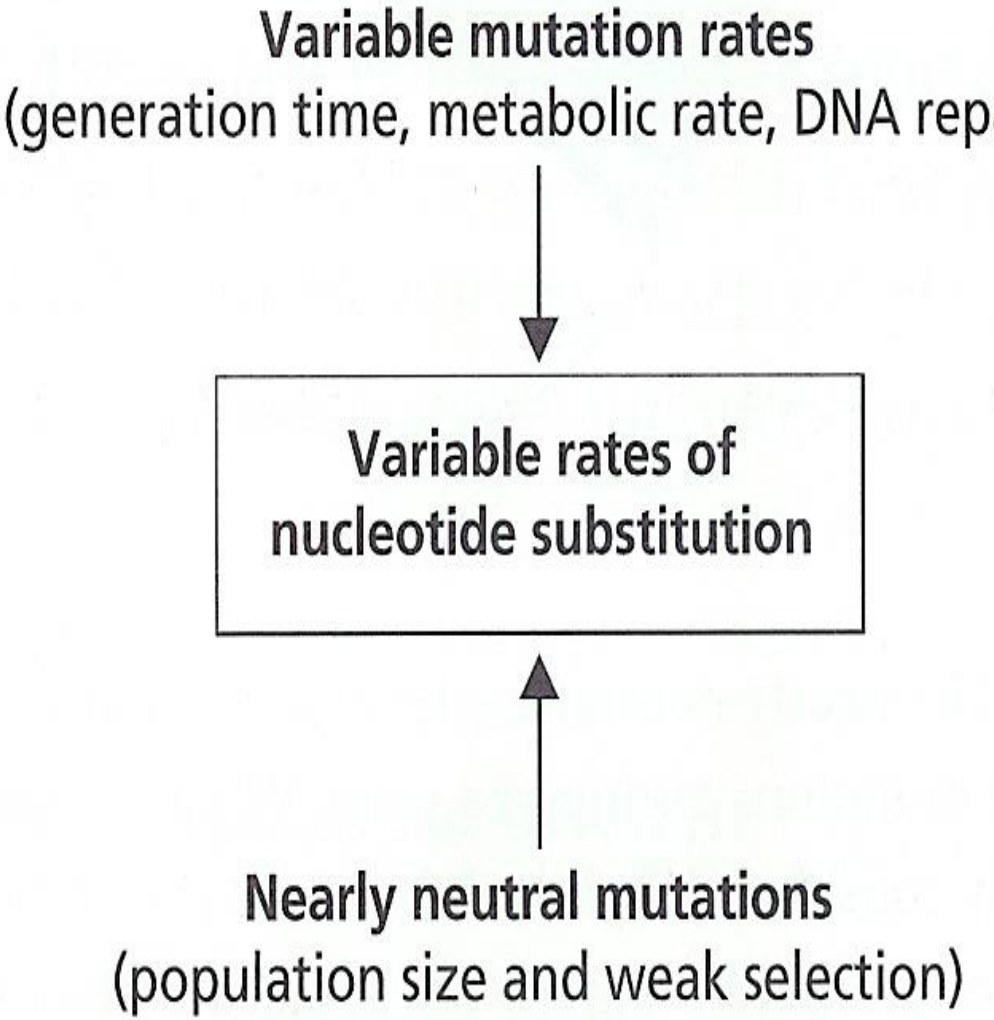
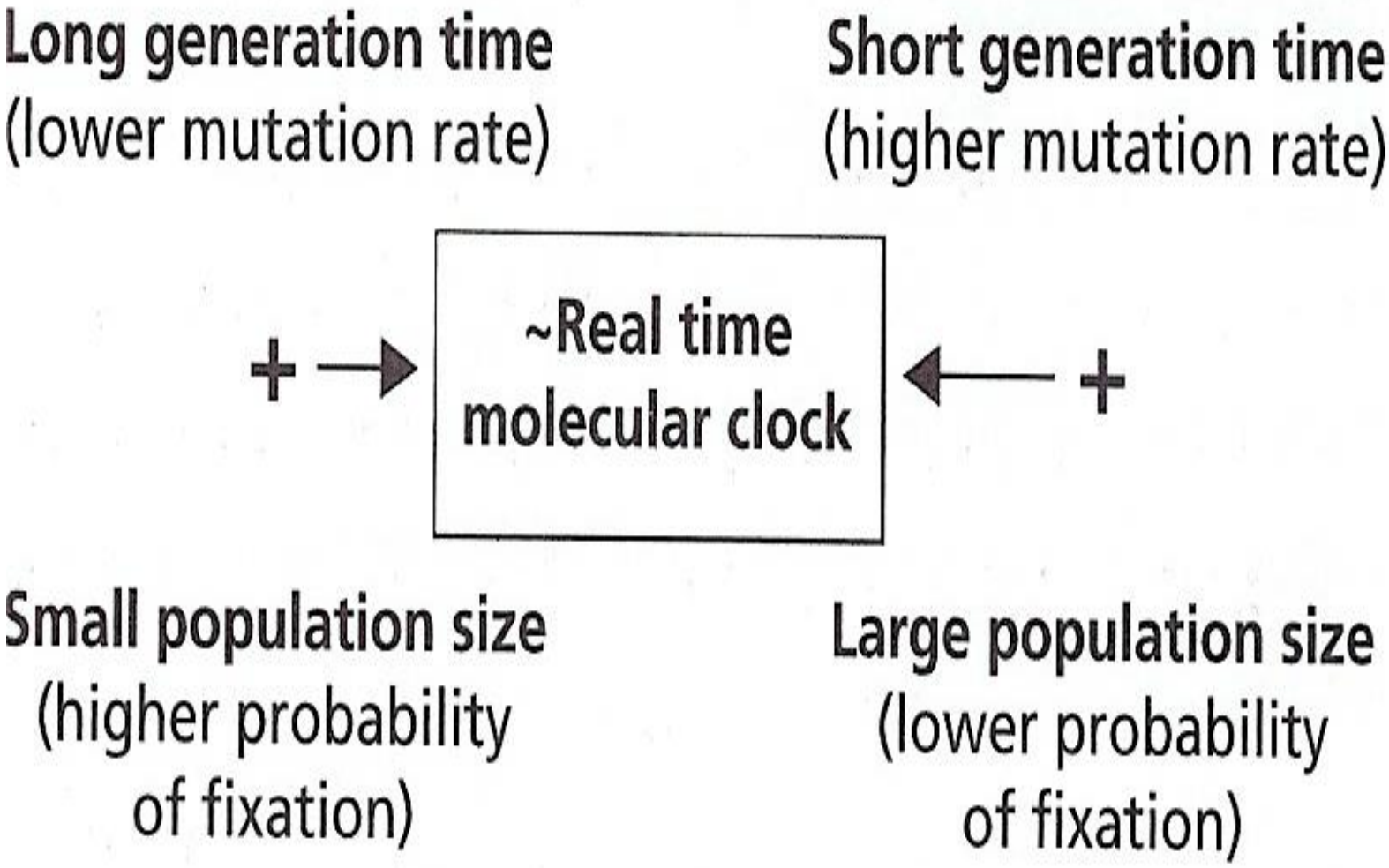


Figure 8.3 Relationship between rate of mtDNA sequence divergence (% change per million years) and body size (in kg) for various vertebrates. Data points: 1, mice; 2, dogs; 3, human-chimpanzee; 4, horses; 5, bears; 6, geese; 7, whales; 8, newts; 9, frogs; 10, tortoise; 11, salmon; 12, sea turtles; 13, sharks. Boxes represent the range of rates and body sizes for a given taxon. Solid lines are drawn to pass through the boxes. Dashed line represents the hypothesis of rate constancy. From Martin and Palumbi (1993).

2-Tamanho da população efectiva



2-Tamanho da população efectiva

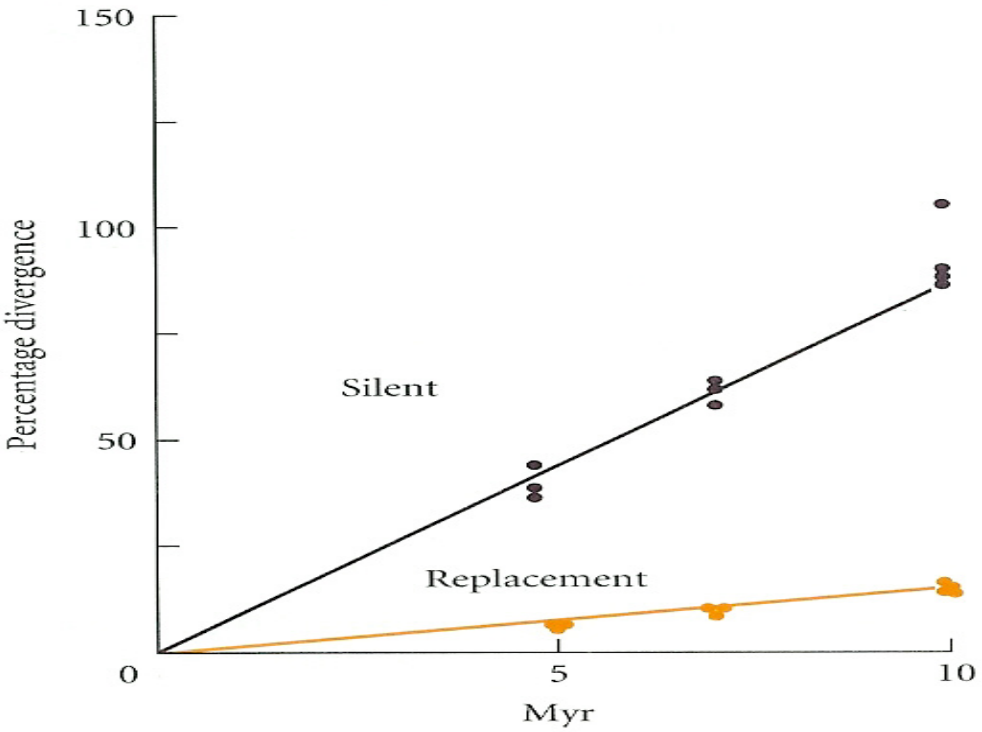


3-Constrangimentos selectivos

Constrangimentos – variação temporal nos constrangimentos (ex: pseudogenes)

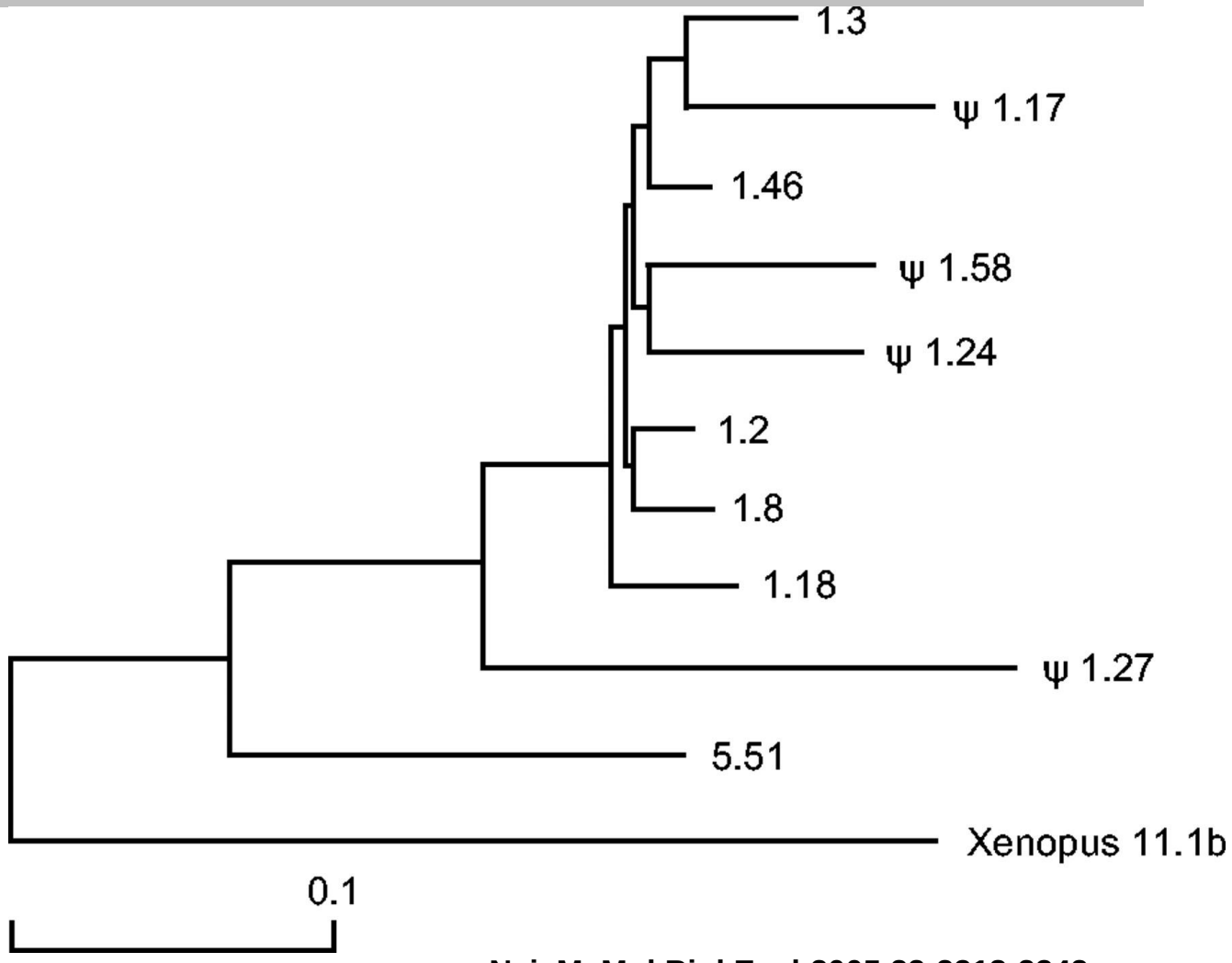
Coeficientes selectivos

3-Constrangimentos selectivos



Immunoglobulin (IG) heavy chain variable region genes (Vh) - Phylogenetic tree of 10 group A human VH genes

All sequences except for *Xenopus* 11.1b were taken from Shin et al. (1991) and Matsuda et al. (1993). The *Xenopus* gene used here is the one of the closest out-group genes. = pseudogene. The branch lengths are measured in terms of the number of nucleotide substitutions with the scale given below the tree.

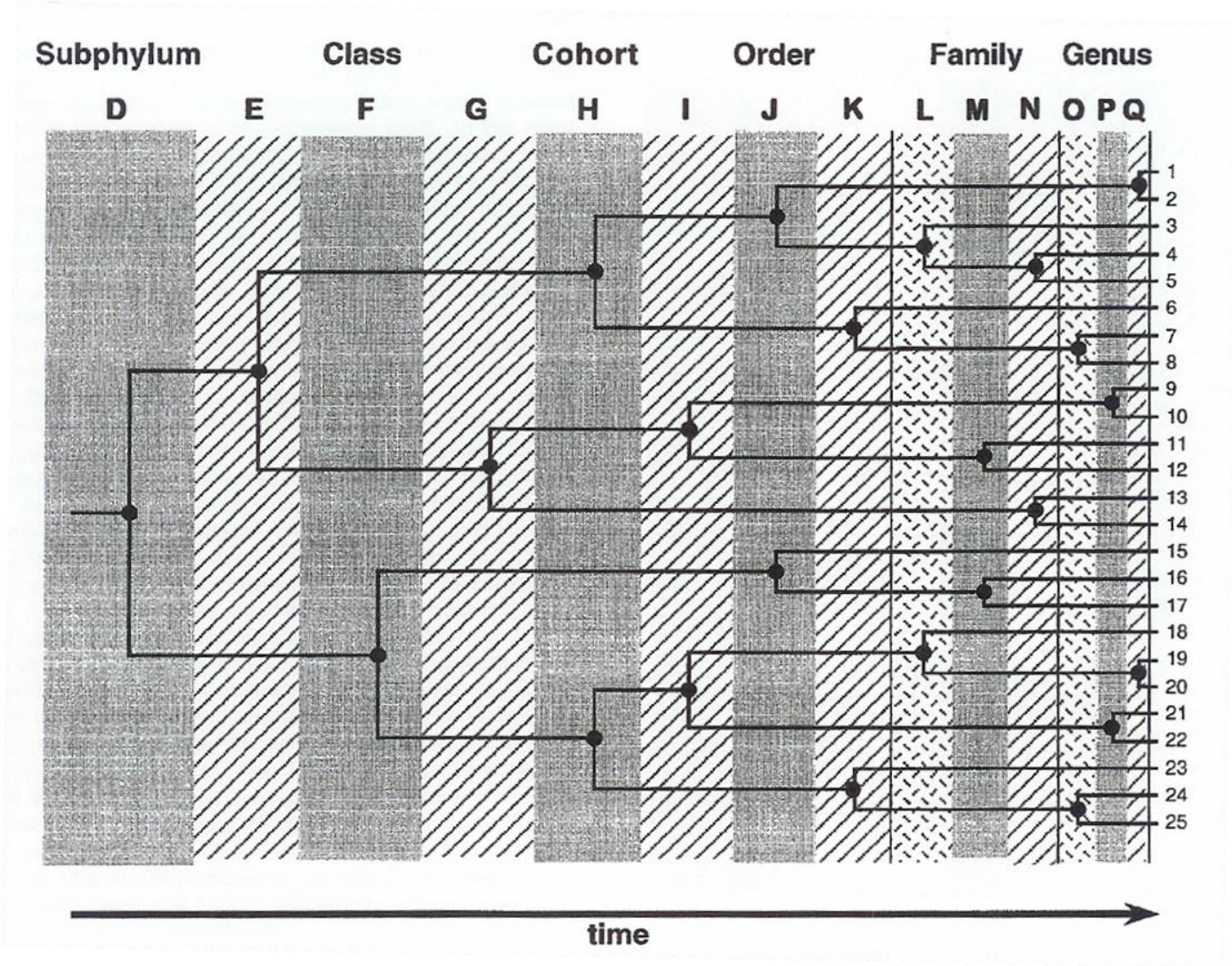


Nei, M. Mol Biol Evol 2005 22:2318-2342

Rates of nucleotide substitution per site per year x10-9 for mammalian globin pseudogenes and their functional homologues

		Functional gene		
Gene	Pseudogene	Position 1	Position 2	Position 3
Mouse Ψα3	5.0	0.75	0.68	2.65
Human Ψα1	5.1	0.75	0.68	2.65
Rabbit Ψα2	4.1	0.94	0.71	2.02
Goat Ψβ Ψ	4.4	0.94	0.71	2.02
average	4.7	0.85	0.7	2.34

Temporal framework?



Fossil calibrations

Fossil-age calibrations
Constraints on the timing of lineage divergence in molecular clock dating. They are established through fossil-based minimum and maximum constraints on clade ages (node calibrations) or through the inclusion of dated fossil species in the analysis (tip calibrations).

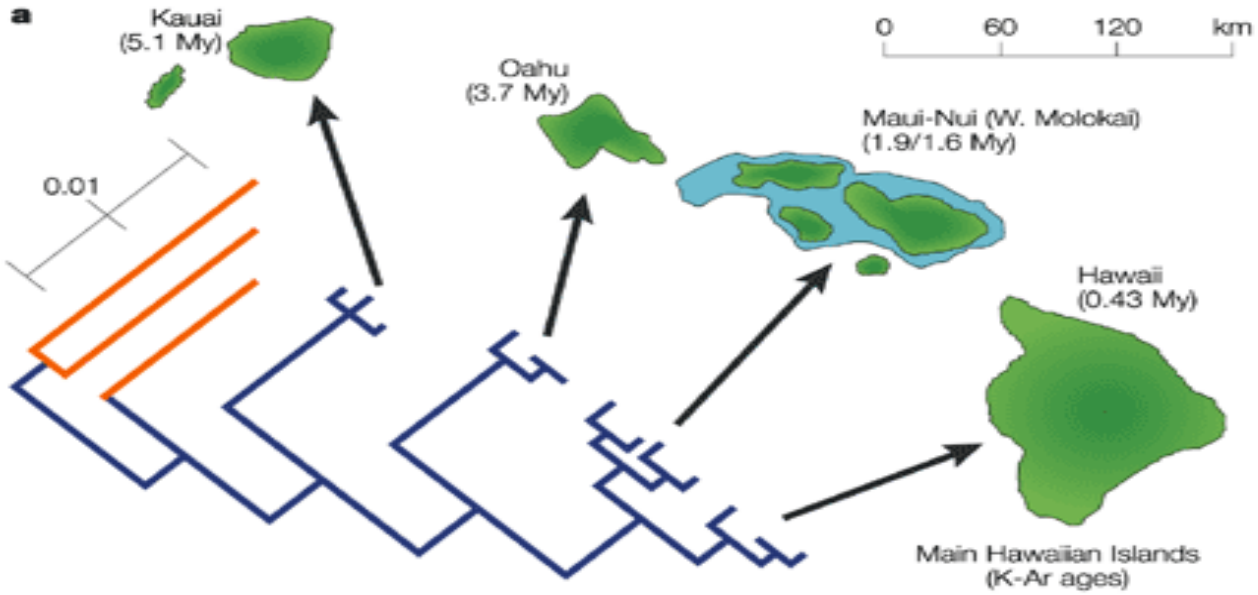
Geological and Biogeographic calibrations

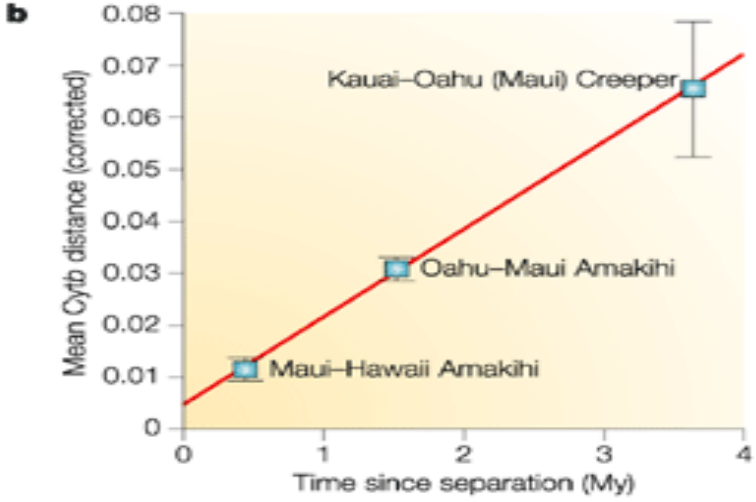
Geological and biogeographic calibrations

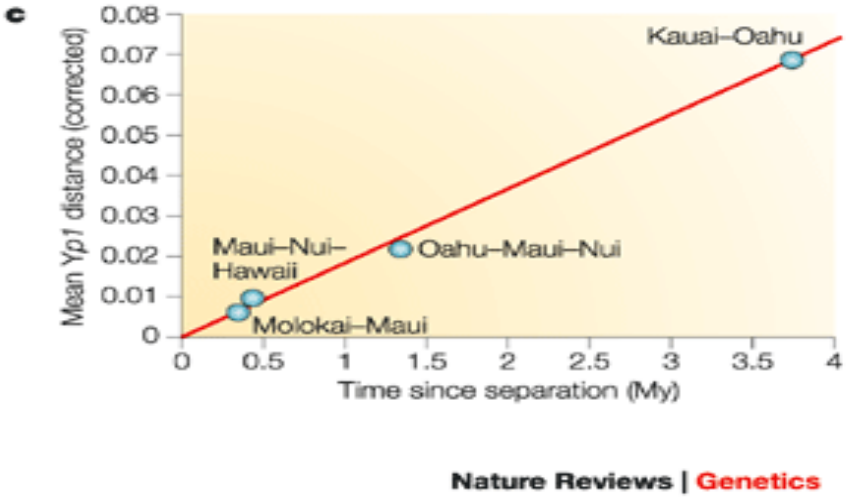
If the phylogenetic tree contains lineage divergences that are thought to be associated with particular geological events, the estimated age of each geological event can be used to calibrate the molecular clock. These can include any geological events that promote reproductive isolation or produce a founder effect, ranging from ancient continental drift to recent island colonizations (Kodandaramaiah 2011). However, biogeographic calibrations are often controversial because of the strong assumption that genetic divergence is tied to geological events. This practice is susceptible to a range of confounding factors, including errors in the estimation of geological ages, the degree of association between geological events and genetic divergences and the impacts of taxon sampling and lineage extinction (Heads 2011; Kodandaramaiah 2011).

Newer approaches to geological calibrations include tying them to demographic events, rather than to genetic divergences. This reduces the impact of the discrepancy between population and genetic divergence, which can be a considerable source of estimation bias (Edwards & Beerli 2000; Peterson & Masel 2009; Ho *et al.* 2011). In a study of marine invertebrates, Crandall *et al.* (2012) assumed that the timing of sea level rise following the last glacial maximum was tied to population expansion, rather than to the most recent common ancestor of the sampled individuals. Similar analyses using demography-based calibrations will provide further insights into the utility and effectiveness of this approach.

Relógios molecular para o Hawai (dados das aves *Hemignathus vires e H. wilsoni em a e b e Drosophila em c)*







Nature Reviews | Genetics

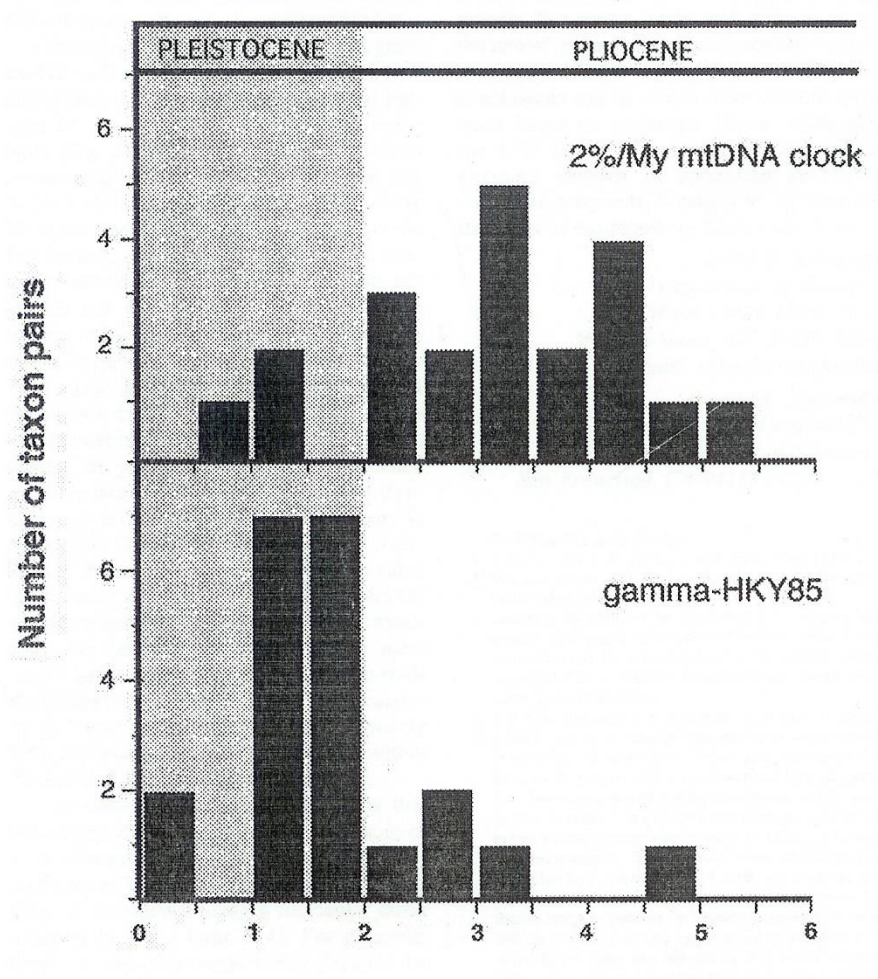
Calibrations

Heterochronous sequence data

In data sets comprising nucleotide or amino acid sequences from ancient specimens or viruses, the samples often have distinct ages. If the temporal span of the samples is sufficient to allow the accumulation of genetic variation, the ages of the samples can be used for calibration (Rambaut 2000; Drummond *et al.* 2003). In studies of viruses, sample ages are often known from collection dates, patient records or other forms of medical and historical documentation. In contrast, ancient DNA sequences are often obtained from samples of unknown age. These samples can be dated using radiometric methods, from the dates of associated remains or from stratigraphic analysis. Uncertainty in sample ages, such as the error in radiocarbon dating, can be incorporated into the analysis (Molak *et al.* 2014). However, accounting for uncertainty in sample ages often has only a small impact on the resulting estimates of rates and divergence times (Molak *et al.* 2013). If the ages of the heterochronous sequences are unknown, they can be estimated as part of the molecular-clock analysis (Shapiro *et al.* 2011).

relógios moleculares provisórios

Cytochrome b 2% por MY



uncorrelated relaxed clock (RC) model (Drummond et al. 2006)

Each branch has its own mutation rate μi,

these per-branch rates are independent of one another.

In current implementations of the uncorrelated RC model,
the rates μi are drawn independently and identically
from a well-defined rate distribution, for example:
a lognormal distribution (Drummond et al. 2006),
exponential distribution (Drummond et al. 2006; To et al. 2016),
a normal distribution (Sagulenko et al. 2018),
gamma distribution (Volz and Frost 2017; Didelot et al. 2018).

Bayesian clock

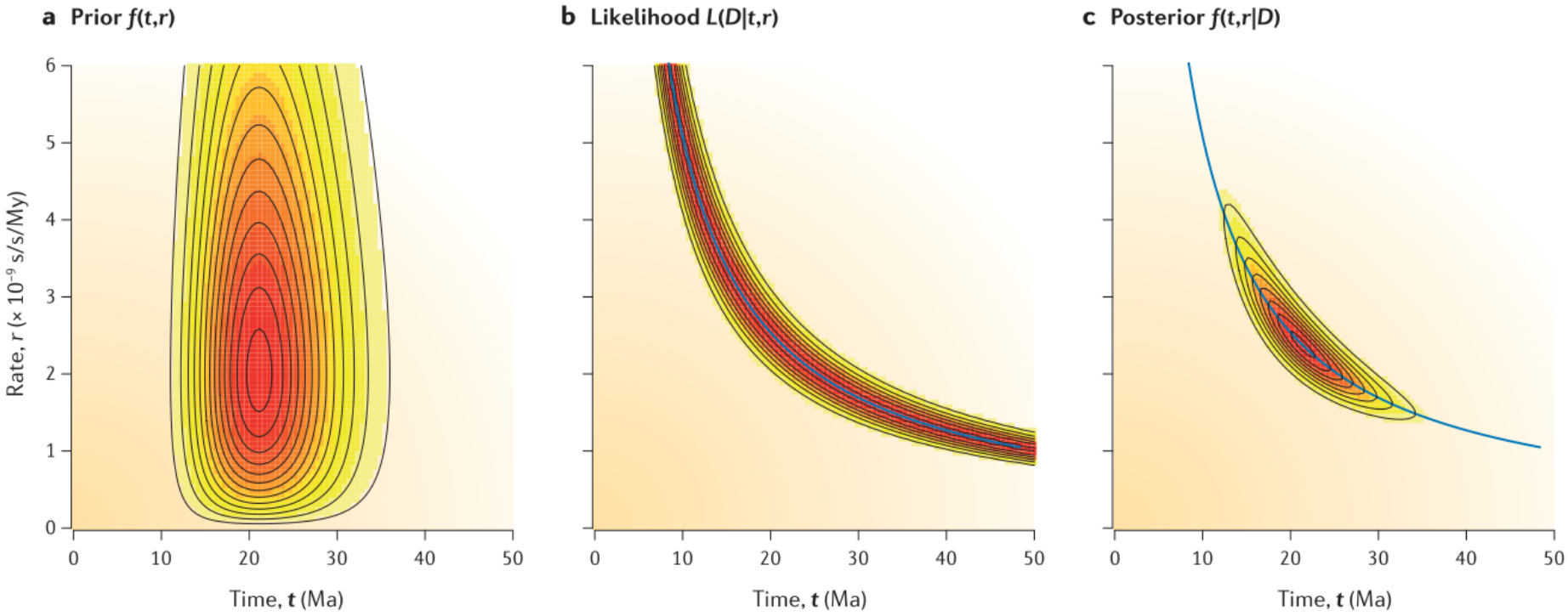


Figure 1 | **Bayesian molecular clock dating.** We estimate the posterior distribution of divergence time (t) and rate (r) in a two-species case to illustrate Bayesian molecular clock dating. The data are an alignment of the 12S RNA gene sequences from humans and orang-utans, with 90 differences at 948 nucleotides sites. The joint prior (part **a**) is composed of two gamma densities (reflecting our prior information on the molecular rate and on the geological divergence time of human–orang-utan), and the likelihood (part **b**) is calculated under the Jukes–Cantor model. The posterior surface (part **c**) is the result of multiplying the prior and the likelihood. The data are informative about the molecular distance, $d = tr$, but not about t and r separately. The posterior is thus very sensitive to the prior. The blue line indicates the maximum likelihood estimate of t and r, and the molecular distance d, with $\hat{t}\hat{r} = \hat{d}$. When the number of sites is infinite, the likelihood collapses onto the blue line, and the posterior becomes one-dimensional[62].

Reis et. al 2015 NRG

Bayesian clock

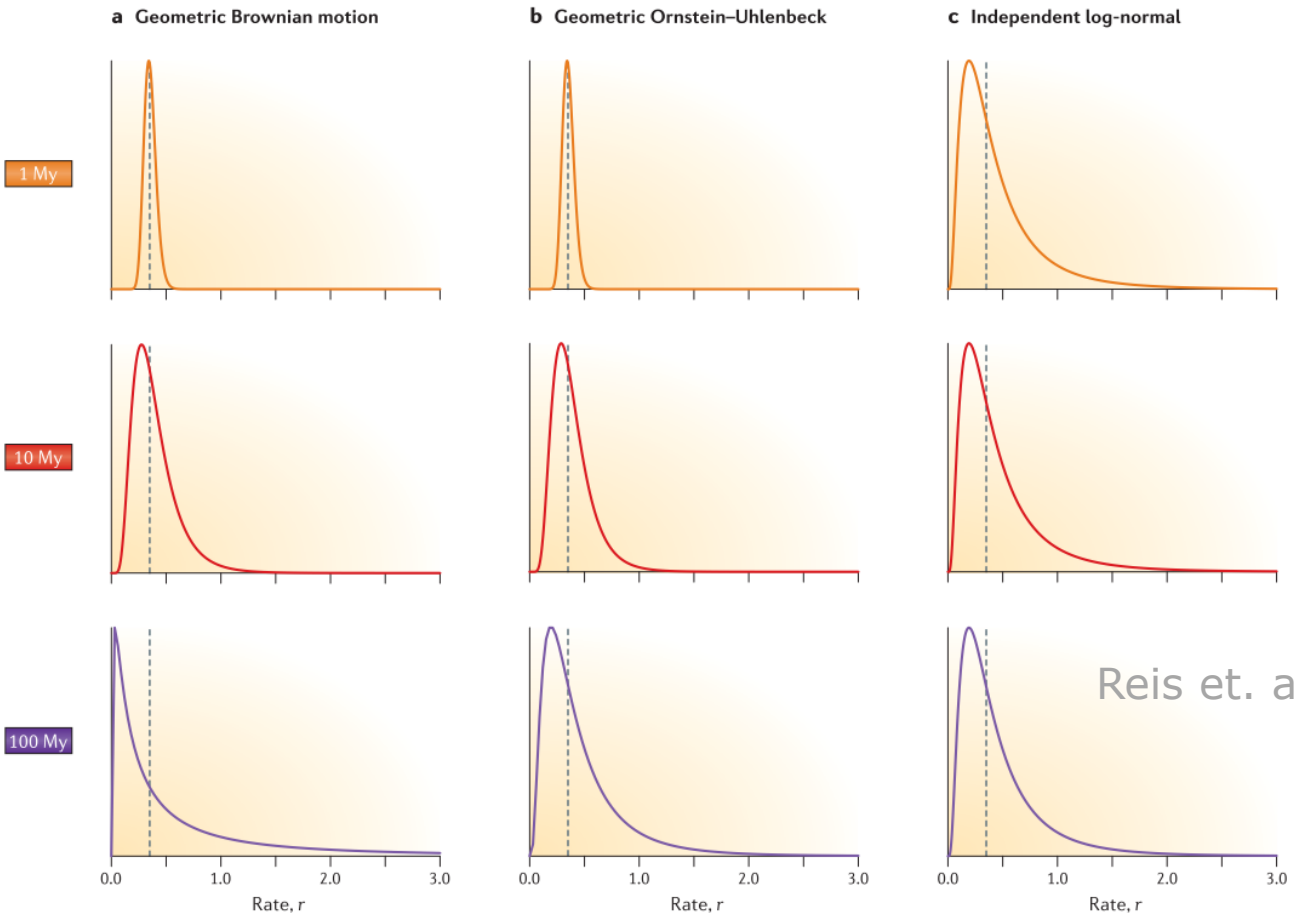
posterior probability distribution. In molecular clock dating, the parameters are the species divergence times (t) and the evolutionary rates (r). Given the sequence data (D), the posterior of times and rates is given by the Bayes theorem as follows:

$$f(t, r|D) = \frac{1}{z} f(t)\, f(r|t)\, L(D|t, r) \qquad (2)$$

Here, $f(t)$ is the prior on divergence times, which is often specified using a model of cladogenesis (of speciation and extinction[54,56], and so on) and incorporates the fossil calibration information[52,54]; $f(r|t)$ is the prior on the rates of branches on the tree, which is specified using a model of evolutionary rate drift[29–31]; and $L(D|t, r)$ is the likelihood or the probability of the sequence data, which is calculated using standard algorithms[11]. FIGURE 1

Reis et. al 2015 NRG

Bayesian clock



Reis et. al 2015 NRG

Figure 3 | **Three relaxed clock models of rate drift.** The rate of molecular evolution among lineages (species) is described by a time-dependent probability distribution (plotted here for three time points: 1 My, 10 My and 100 My) since the lineages diverged from a common ancestral rate ($r_0 = 0.35$ substitutions per site per 100 My (represented by the dashed line)). **a** | The geometric Brownian process[29,31,52] (here with drift parameter $v = 2.4$ per 100 My). This model has the undesirable property that the variance increases with time and without bound, and at large times the mode of the distribution is pushed towards zero. **b** | The geometric Ornstein–Uhlenbeck process (here with $v = 2.4$ per 100 My and dampening force $f = 2$ per 100 My) converges to a stationary distribution with constant variance when time is large. **c** | The independent log-normal distribution[30,31] is a stationary process, and the variance of rate among lineages remains constant through time (here with log-variance $\sigma^2 = 0.6$, the same as the long-term log-variance of the Ornstein–Ulhenbeck process above). The branch length (the amount of evolution along the branch) under the rate-drift models of parts **a** and **b** is usually approximated in Bayesian dating software[31,52]; methods for exact calculation have recently been developed[55].

Bayesian clock

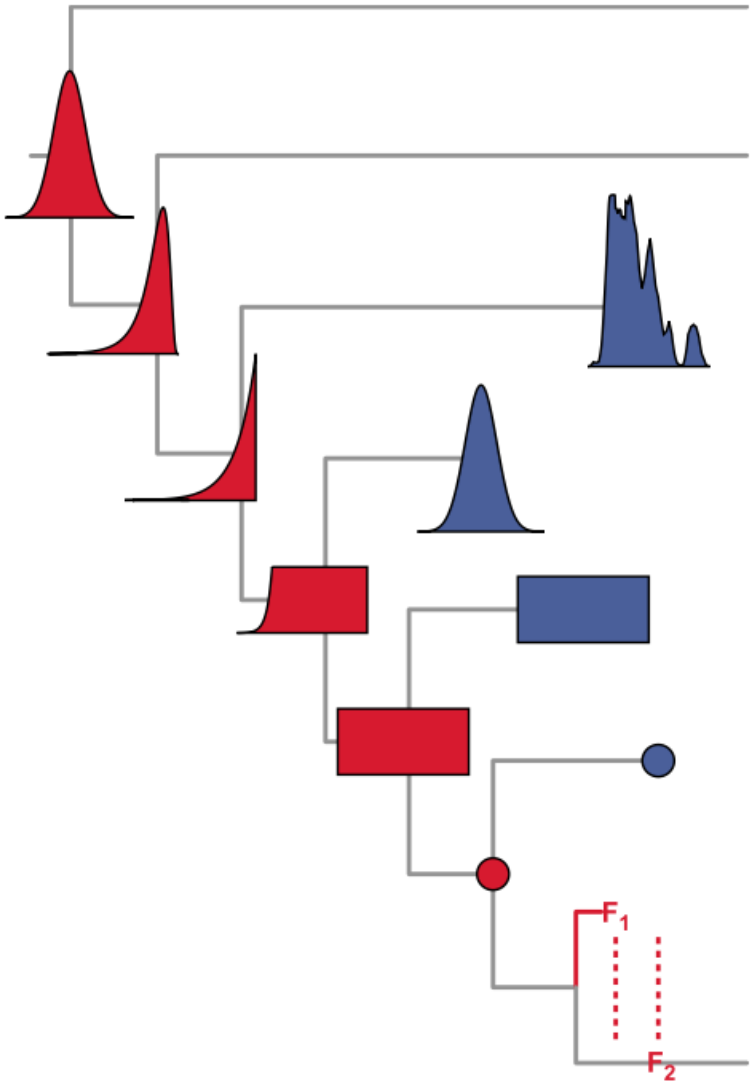


Fig. 1 Different approaches for representing uncertainty in calibrations in a phylogenetic tree. Calibrations at internal nodes (red), from top to bottom: normal distribution, lognormal distribution, exponential distribution (Drummond *et al.* 2006), uniform distribution with hard minimum and soft maximum bounds (Yang & Rannala 2006), uniform distribution with hard minimum and maximum bounds, point value and fossilized birth–death model (Heath *et al.* 2014). F_1 and F_2 represent fossil occurrences of known age. Calibrations at terminal nodes (blue), from top to bottom: empirical calibrated radiocarbon sampler (Molak *et al.* 2014), normal distribution, uniform distribution with hard minimum and maximum bounds, and point value.

Ho et. al 2014 MolEco

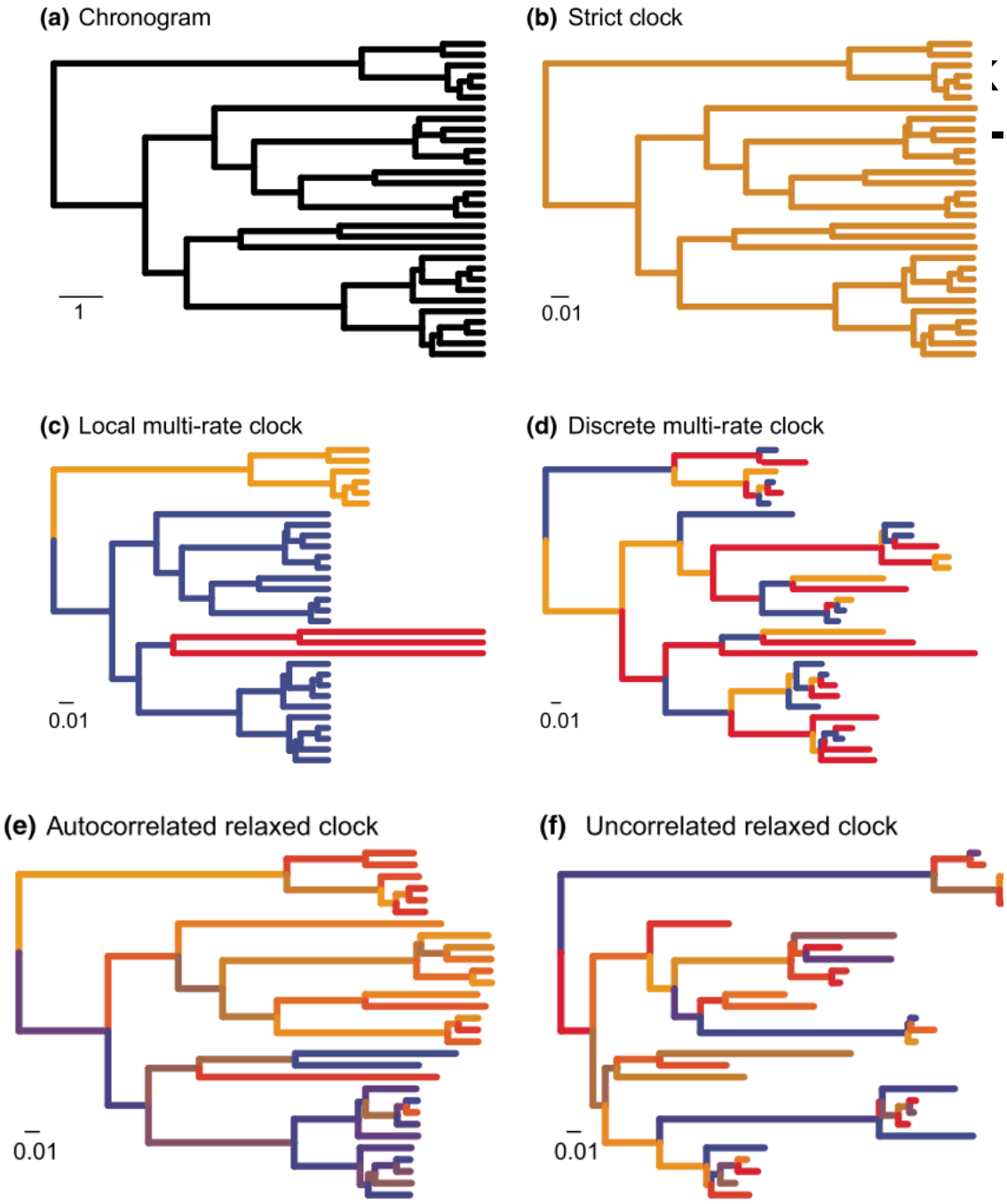


Fig. 2 Phylogenetic trees showing differences among models of rate variation. Panel (a) shows a chronogram with branch lengths measured in time units. The scale bar indicates 1 time unit. The remaining panels show phylograms with branch lengths generated under a range of different clock models: (b) strict clock, with a constant rate among branches; (c) local multi-rate clock, with a distinct rate in each of three groups of branches; (d) discrete multi-rate clock, with a small number of branch-specific rates distributed throughout the tree; (e) autocorrelated relaxed clock, with a distinct rate along each branch that is correlated with the rate along its parent branch; and (f) uncorrelated relaxed clock, with a distinct rate along each branch drawn from a chosen probability distribution. In panels (b) to (f), scale bars represent 0.01 substitutions per site.

Ho et. al 2014 MolEco

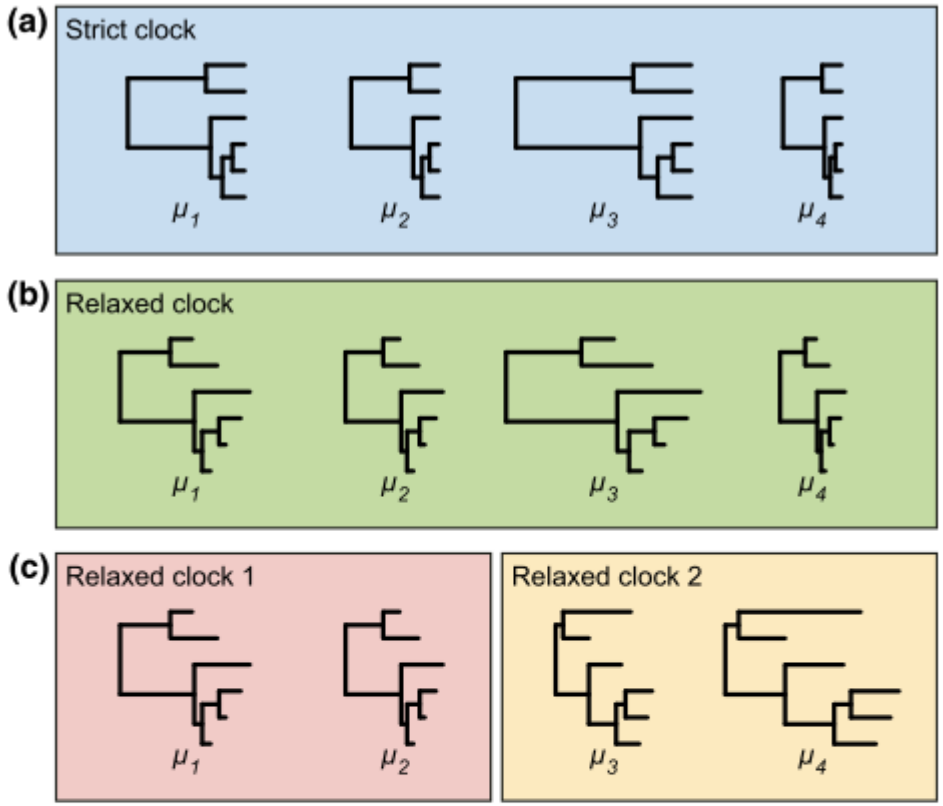


Fig. 3 Clock-model schemes for dealing with different patterns of rate variation in a four-gene data set. (a) When only gene effects are present, a strict-clock model can be used, with a different rate (μ_i) for each gene. (b) When gene and lineage effects are both present, a relaxed-clock model can be used, with the branch-specific rates being scaled differently for each gene. The ratio of rates along different branches is identical across genes. (c) When gene-by-lineage (residual) effects are present, separate relaxed-clock models should be applied to clusters of genes that share the same pattern of among-lineage rate variation. For each relaxed clock, the branch-specific rates are scaled differently for each gene.

Ho et. al 2014 MolEco

Bayesian clock

Table 1 | **Sample of Bayesian programs that use the molecular clock to estimate divergence times***

Program	Method	Brief description	Refs
Beast	Bayesian	Comprehensive suite of models. Particularly strong for the analysis of serially sampled DNA sequences. Includes models of morphological traits	132
DPPDiv	Bayesian	Dirichlet relaxed clock model[71]. Fossilized birth–death process prior to calibrate time trees[56]	133
MCMCTree	Bayesian	Comprehensive suite of models of rate variation. Fast approximate likelihood method that allows the estimation of time trees using genome alignments[57]	134
MrBayes	Bayesian	Large suite of models for morphological and molecular evolutionary analysis. Comprehensive suite of models of rate variation	135
Multidivtime	Bayesian	The first Bayesian clock dating program. Introduced the geometric Brownian model and the approximate likelihood method	29,53
PhyloBayes	Bayesian	Broad suite of models. Uses data augmentation to speed up likelihood calculation and can be efficiently used in parallel computing environments (MPI enabled)	136, 137
r8s	Penalized likelihood	Very fast (uses Poisson densities on inferred mutations to approximate the likelihood). Suitable for the analysis of large phylogenies. Suitable for estimating relative ages (by fixing the age of the root to 1). Does not deal with fossil and branch length uncertainty correctly[138]	139
TreePL	Penalized likelihood	Similar to r8s	140

*The Bayesian programs listed were chosen for their ability to accommodate multiple calibrations with uncertainties (bounds or other probability densities), multiple loci of sequence data and relaxed clock models. Penalized likelihood programs are listed as they are related to the Bayesian method[138].

Dated phylogenies

1- Dated phylogenies can be built directly from the genetic data using Bayesian phylogenetic methods implemented in BEAST

2 - Two-step approach:

The first step (standard phylogenetics): RAxML, PhyML, FastTree, IQ-Tree

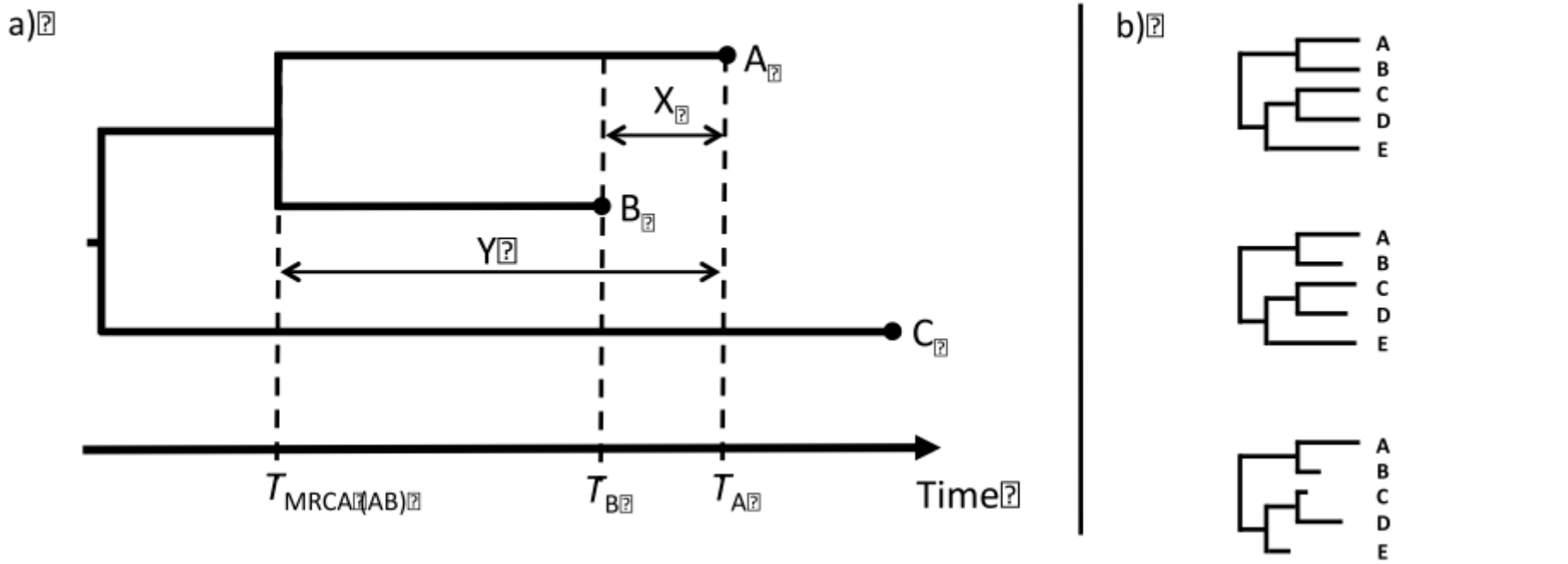
The second step (phylogeny dating) can be performed, for example, using:
LSD (To et al. 2016),
node.dating (Jones and Poon 2017),
treedater (Volz and Frost 2017),
TreeTime (Sagulenko et al. 2018),
BactDating (Didelot et al. 2018).

Didelot et. al 2021 MBE

Tip-dating principle

Figures

Rieux & Balloux 2016 MolEco



Panel a) In this simplified theoretical situation adapted from Rambaut (2000), sequences A and B were isolated at different points in time (T_A and T_B respectively) and C is an outgroup sequence. If we assume the rate of evolution to be the same in lineage A and B, then the amount of molecular evolution expected to have occurred between T_A and T_B is equal to $d_{AC} - d_{BC}$ (d_{AC} and d_{BC} being the genetic distance between A&C and B&C, respectively). If the time X between T_A and T_B represents a significant proportion of the time Y since A and B last shared a common ancestor, then one can use tip dates to conjointly estimate the rate of evolution $\mu = (AC - BC) / (T_A - T_B)$ and extrapolate the age of $T_{MRCA(AB)}$

Panel b) Top: tree with modern samples only for which no divergence time estimate is possible without calibrations on internal nodes or a strong prior on the rate of molecular clock. Middle: tree where tip dates may not be widely spread enough for accurate inferences. Bottom: tree where tip date width should be sufficiently broad to allow divergence-time and rate of evolution estimates with a good degree of certainty, since the sample dates cover a relatively large fraction of the total age of the tree.

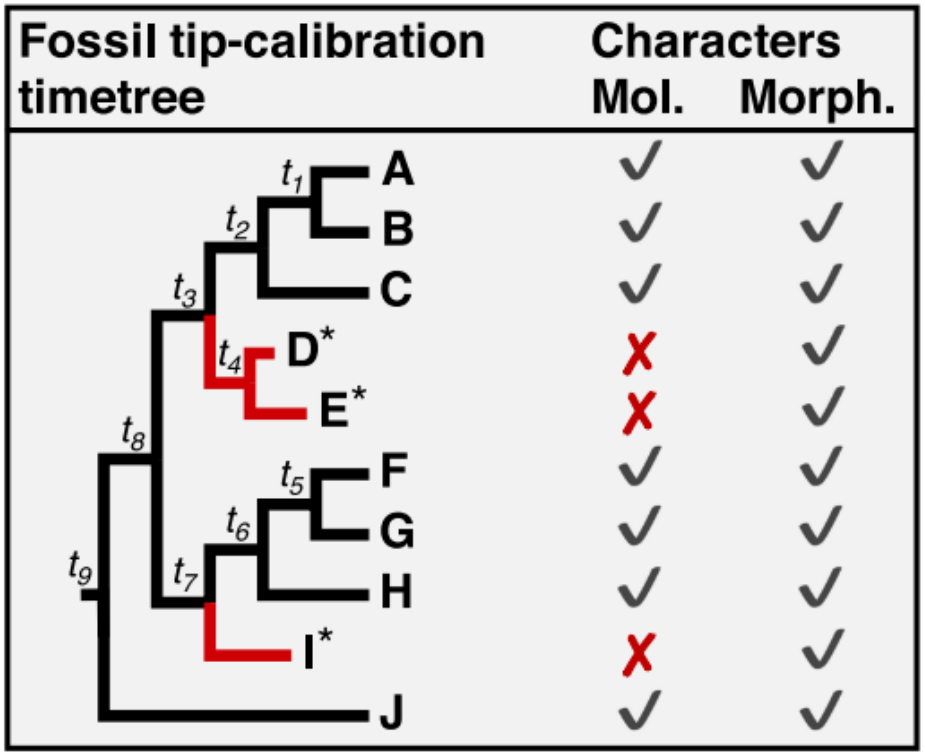


Fig. 1 Integration of morphological characters from living and extinct species in a combined analysis with molecular data. The combined matrices contain morphological characters from both fossil (*) and living species, but no molecular data from fossil species. Red branch lengths are created due to the inclusion of extinct taxa and are estimated using non-molecular data. Intuitively, we expect t_3, t_4, and t_7 to be estimated well if the non-molecular data has a strong time structure. The inclusion of non-molecular data will only benefit if the time structure is concordant between molecular and non-molecular datasets. However, we emphasize the importance of using morphological data to infer the phylogenetic position of extinct taxa in the tree to determine which internal nodes are to be calibrated

Barbara-Montoya et. al 2021 BMC EE

LETTER

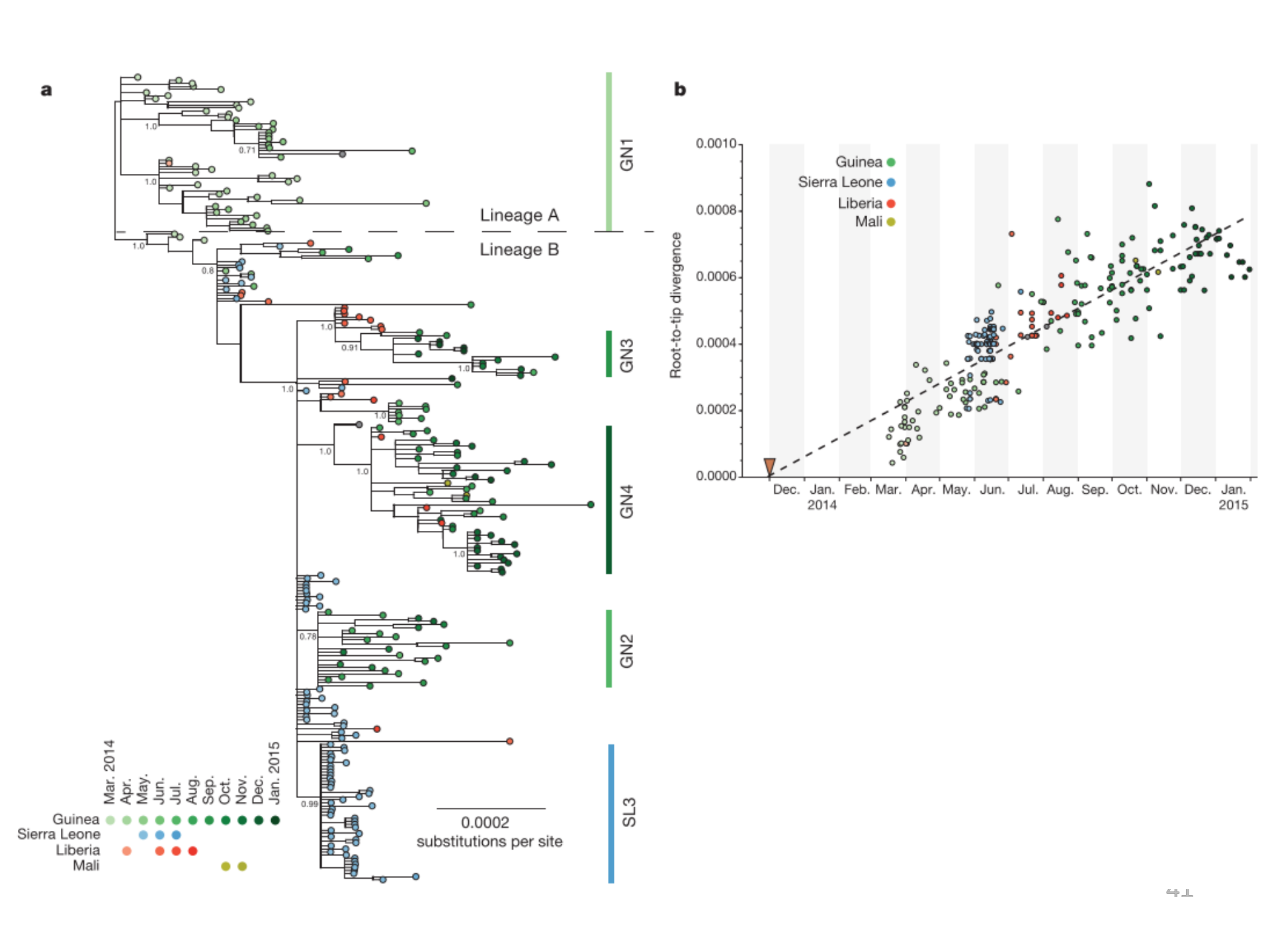
OPEN

doi:10.1038/nature14594

Temporal and spatial analysis of the 2014–2015 Ebola virus outbreak in West Africa

Miles W. Carroll[1,2,3], David A. Matthews[4]*, Julian A. Hiscox[5]*, Michael J. Elmore[1]*, Georgios Pollakis[5]*, Andrew Rambaut[6,7,8]*,

Figure 2 | Phylogenetic relatedness and nucleotide sequence divergence of EBOV isolates from the 2013–2015 outbreak. **a**, Phylogenetic relatedness of EBOV isolates. Phylogenetic tree inferred using MrBayes[11] for full-length EBOV genomes sequenced from 179 patient samples obtained between March 2014 and January 2015. Displayed is the majority consensus of 10,000 trees sampled from the posterior distribution with mean branch lengths. Posterior support is shown for selected key nodes. Twenty-two samples originated in Liberia and were collected between March and August 2014 and six samples from Sierra Leone were obtained in June and July 2014. In our analysis we also included published sequences, including the three early Guinean sequences[2] and 78 sequences described by Gire *et al.*[6]. A number of lineages predominantly circulating in Guinea are denoted as GN1–4 along with a uniquely Sierra Leone lineage (SL3) recognised in Gire *et al.*[6]. **b**, EBOV nucleotide sequence divergence from root of the phylogeny in Fig. 2a plotted against time of collection of each virus. The date of the first documented case near Meliandou in eastern Guinea is indicated by the red triangle.


a
Lineage A
Lineage B
GN1
GN3
GN4
GN2
SL3
1.0
0.71
1.0
1.0
0.8
1.0
0.91
1.0
1.0
1.0
1.0
1.0
0.78
0.99
Mar. 2014
Apr.
May.
Jun.
Jul.
Aug.
Sep.
Oct.
Nov.
Dec.
Jan. 2015
Guinea
Sierra Leone
Liberia
Mali
0.0002
substitutions per site
b
Root-to-tip divergence
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
Guinea
Sierra Leone
Liberia
Mali
Dec.
2014
Jan.
Feb.
Mar.
Apr.
May.
Jun.
Jul.
Aug.
Sep.
Oct.
Nov.
Dec.
Jan.
2015

LETTER

OPEN

doi:10.1038/nature14594

Temporal and spatial analysis of the 2014–2015 Ebola virus outbreak in West Africa

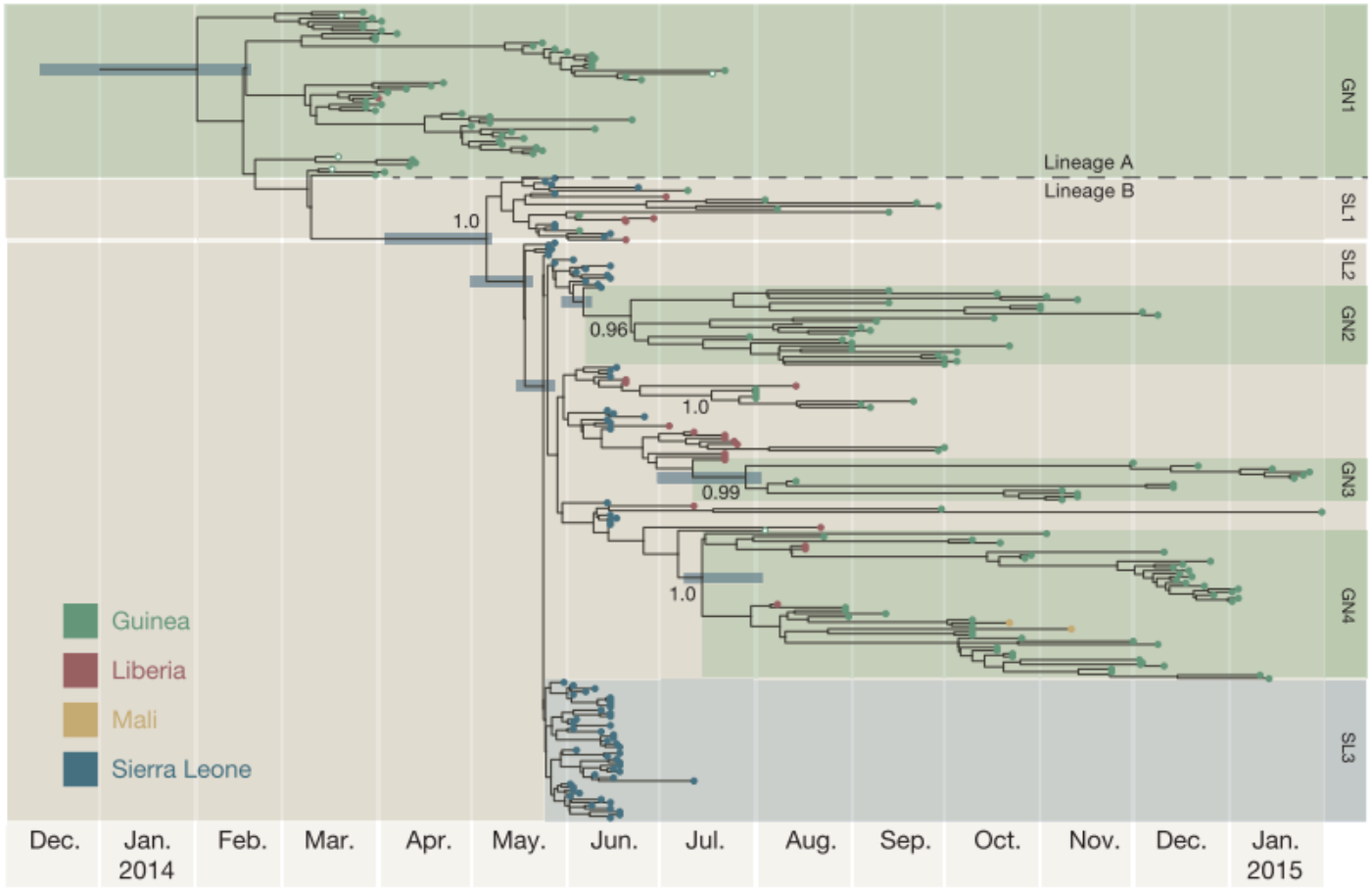


Figure 3 | A time-scaled phylogenetic tree of 262 EBOV genomes from Guinea, Sierra Leone, Liberia and Mali. Shown is a maximum clade credibility tree constructed from 10,000 trees sampled from the posterior distribution with mean node ages. Clades described in Gire *et al.*[6] are identified here (SL1, SL2 and SL3) as well as a number of lineages predominantly circulating in Guinea and posterior probability support is given for these. For certain key node ages, 95% credible intervals are shown by horizontal bars.

LETTER

OPEN

doi:10.1038/nature14594

Temporal and spatial analysis of the 2014–2015 Ebola virus outbreak in West Africa

Miles W. Carroll[1,2,3], David A. Matthews[4]*, Julian A. Hiscox[5]*, Michael J. Elmore[1]*, Georgios Pollakis[5]*, Andrew Rambaut[6,7,8]*,

Phylogenetic analysis. Phylogenetic analysis comprised the 179 EBOV genomes from this study, 78 genomes from Sierra Leone[6], three sequences from Guinea[2] and two sampled from Mali[15]. The genomes were partitioned into four sets of sites—1st, 2nd and 3rd codon positions of the protein-coding regions and the non-coding intergenic regions—with each partition being assigned a generalized time reversible substitution model[16], gamma distributed rate heterogeneity[17] and a relative rate of evolution. This model was used to construct a Bayesian nucleotide divergence tree (Fig. 2) using MrBayes[11] and a time-scaled phylogenetic analysis (Fig. 3) using BEAST[18] with a log-normal distributed relaxed molecular clock[19], and the 'Skygrid' non-parametric coalescent tree prior[20]. The alignments and control files for both analyses are available in Supplementary Data Files 2 and 3 and provide documentation of all model parameters.

Species origin

https://doi.org/10.1038/d41586-023-03487-4

News & views

Evolution

Genes often uninformative for dating species' origins

Matt Pennell

The time frame of a species' origins provides context for evolutionary questions. However, dates from fossils are often inconsistent with estimates from genetic data. Emerging evidence points to a new explanation for this discrepancy.

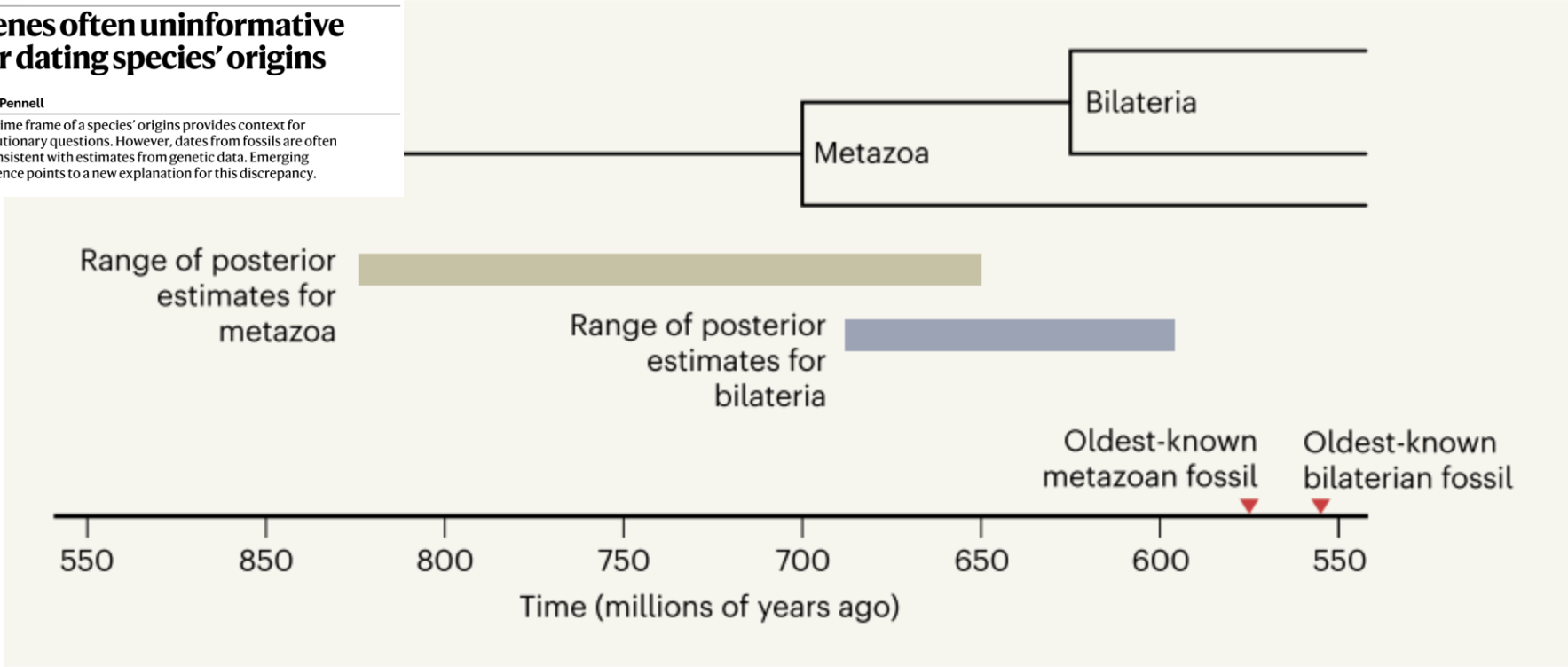
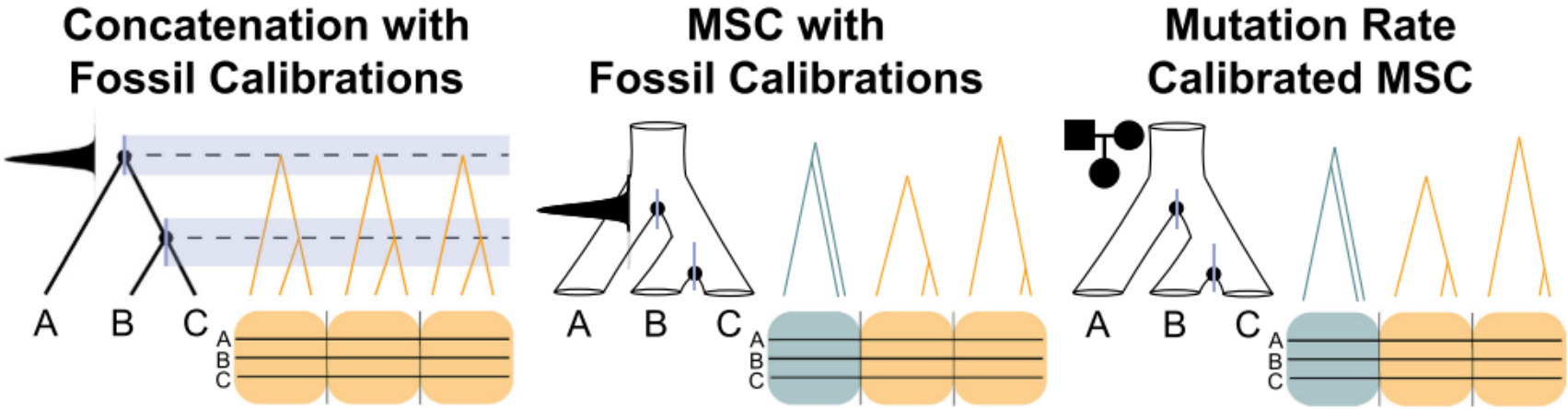


Figure 1 | Discrepancies between genetic and fossil data underpinning evolutionary trees. It can be difficult to assemble the branch points of a tree corresponding to the ancient origins of key groups, such as multicellular animals, termed metazoans, and the origins of a type of metazoa called bilateria (which includes humans). Assembly of such trees typically involves statistical analysis of genetic data to generate what are termed posterior estimates of probable origin dates. However, these dates are often substantially earlier than the oldest known fossils of the groups of interest. Budd and Mann[3] provide an explanation for why these discrepancies occur. Data shown are from refs 1 and 3.



Strengths

Computationally efficient for large numbers of tips and loci

Considers discordance between gene trees and species trees

Does not require calibrations on nodes from external information such as fossils

Weaknesses

May produce biased estimates when ILS is high or when gene sequence divergence is far from species divergence

Increased computational complexity from averaging over gene trees to estimate species tree parameters

Requires external mutation rate estimates from sequenced pedigrees and potentially not appropriate for distant taxa

Common Programs

BEAST2 [96]
MCMCTREE [97]
MrBayes [98]
PhyloBayes [99]

BPP [5,71]
StarBEAST2 [6]

BPP [5,71]
StarBEAST2 [6]

Tiley et. al 2020 TiG

Trends in Genetics

Figure 5. Differences between Bayesian Methods for Divergence Time Estimation and Programs for Implementing Them. A number of methods that estimate divergence times with concatenated data [96–99] or the MSC [5,6,71] are available with some variations in prior distributions and relaxed-clock models. The choice of concatenation or MSC methods, and whether divergence times are calibrated with fossils or mutation rates, is dependent on the data set size, prevalence of ILS among species, and appropriateness of a single germline mutation rate.