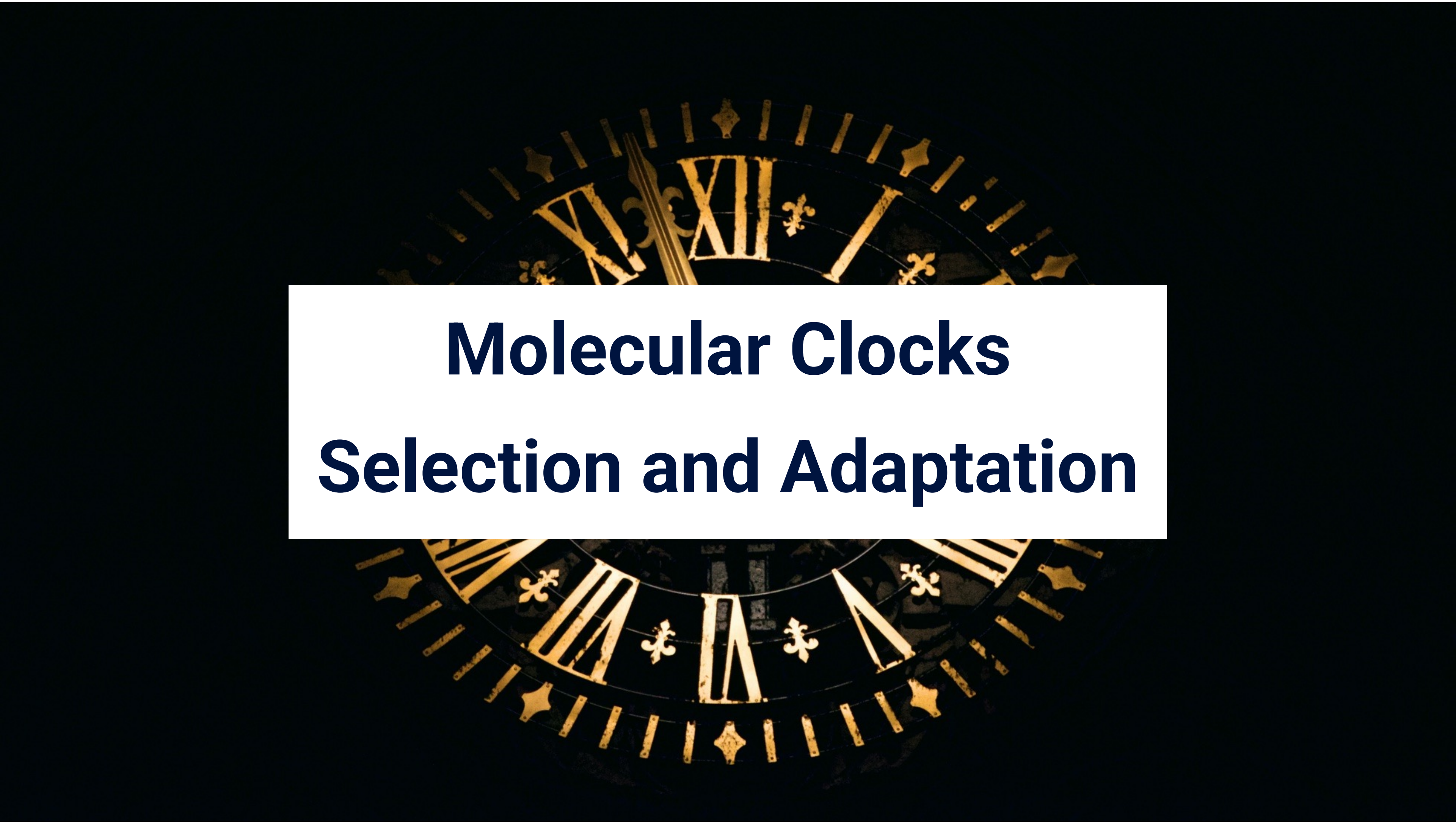
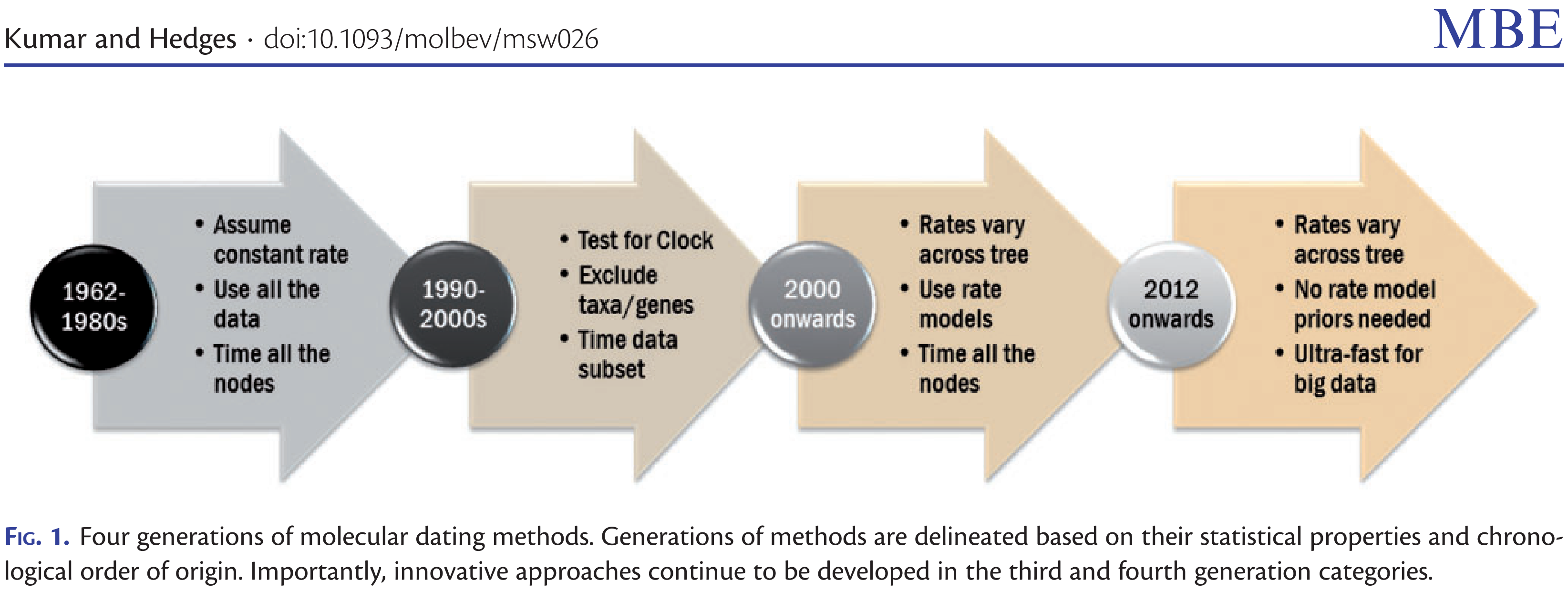


Molecular Clocks

Selection and Adaptation

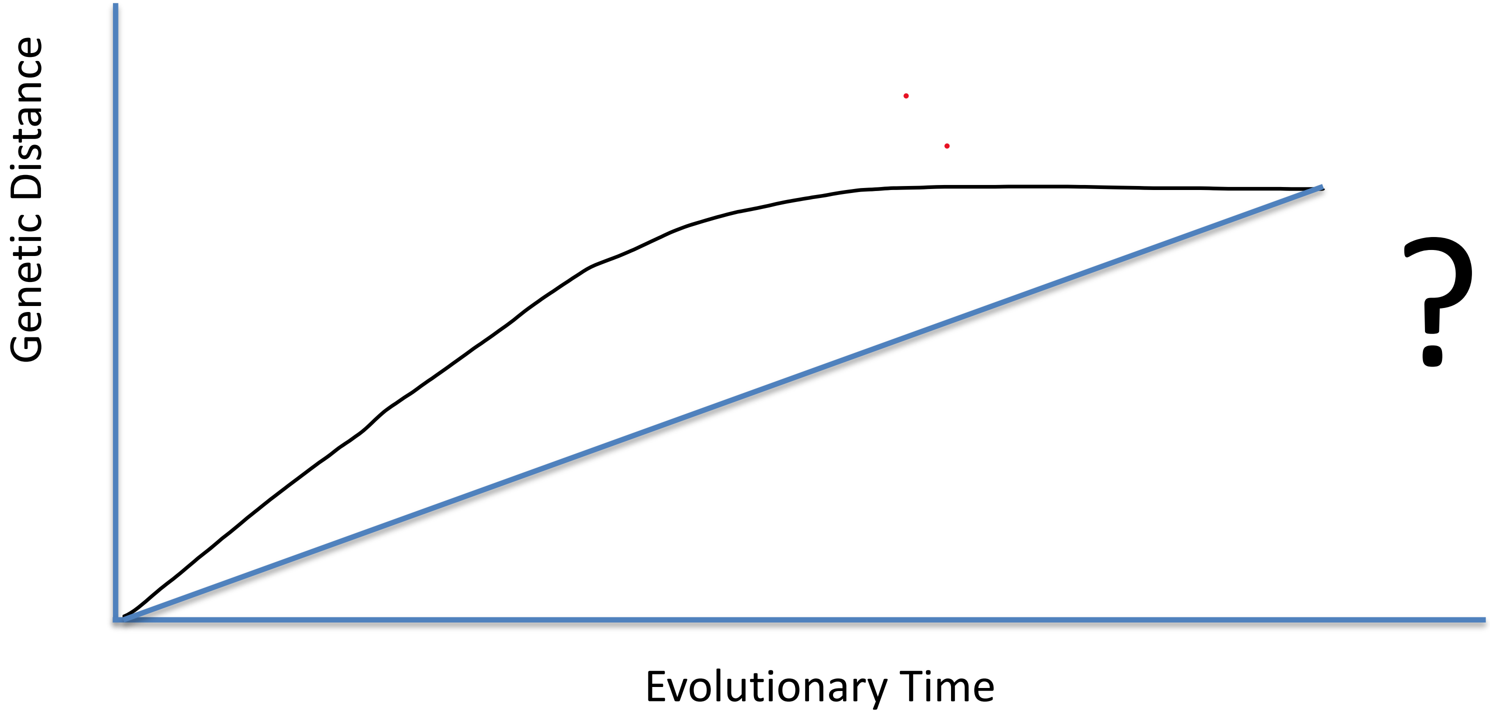
Advances in Molecular Dating Methods



Kumar and Hedges · doi:10.1093/molbev/msw026 MBE

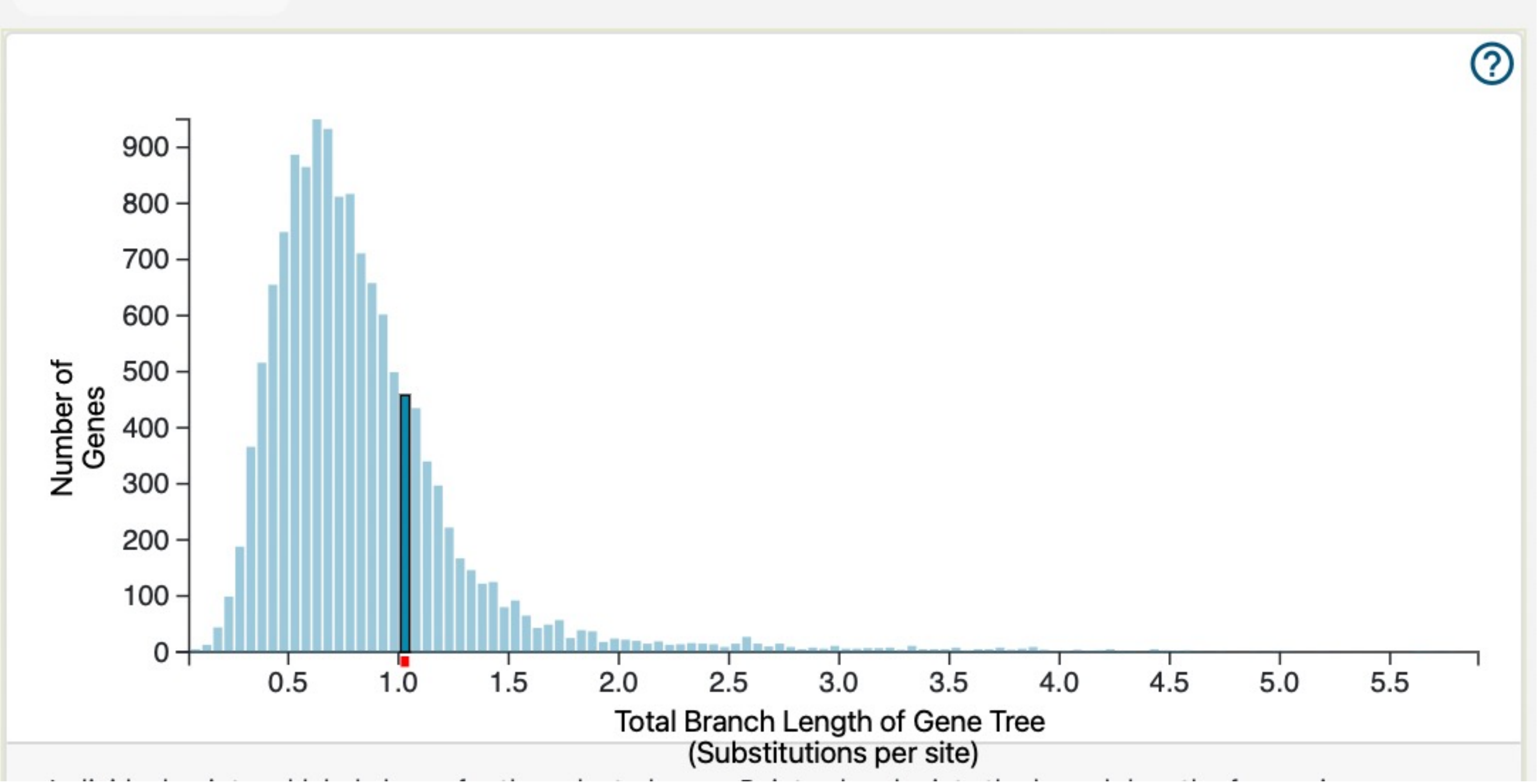
Fig. 1. Four generations of molecular dating methods. Generations of methods are delineated based on their statistical properties and chronological order of origin. Importantly, innovative approaches continue to be developed in the third and fourth generation categories.

Genetic Distance vs. Evolutionary Time



Genetic distance increases with evolutionary time

Evolutionary Rate Distribution: Primate Exome



GDW
Genomics of Disease in Wildlife

Strict 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.

Relaxed Molecular Clock and Bayesian Methods

- Models of evolutionary rate drift over time or across lineages developed to relax the molecular clock hypothesis.
- The parameters are the species divergence times (*t*) and the evolutionary rates (*r*) within a Bayesian model.

Time-Tree: Dating Nodes Within Molecular Genomic Phylogeny

- Testing for Molecular Clocks
  - Strict: substitution rate the same for all branches of topology
  - Relaxed: Heterogeneous substitution rates are permitted with different branches of topology
- Time-Tree estimation with Bayesian framework is computationally prohibitive with large scale genomic sequence data
- Some Work-Arounds with Subsets
    - genes, species or both
    - Codon position (1st and 2nd)
    - Hierarchical lineages within phylogeny
    - Assess sample strategy by testing for consistent rates across lineage
    - Calibration: Multiple fossil record (host) or known date(s) linked with pathogen
    - Iterative subsampling to converge on dates for nodes

GDW
Genomics of Disease in Wildlife

Bayesian Programs to Estimate Molecular Clock

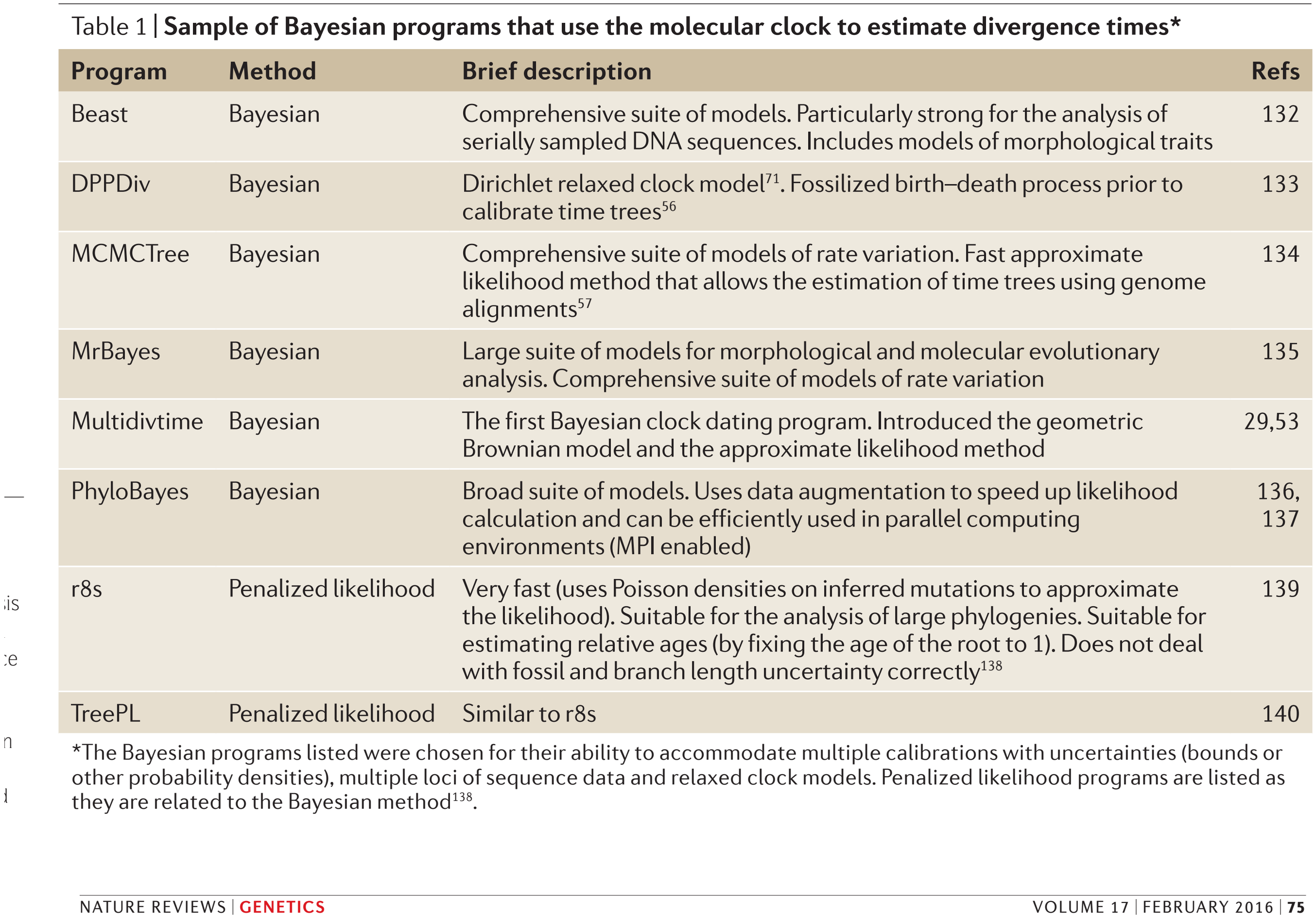
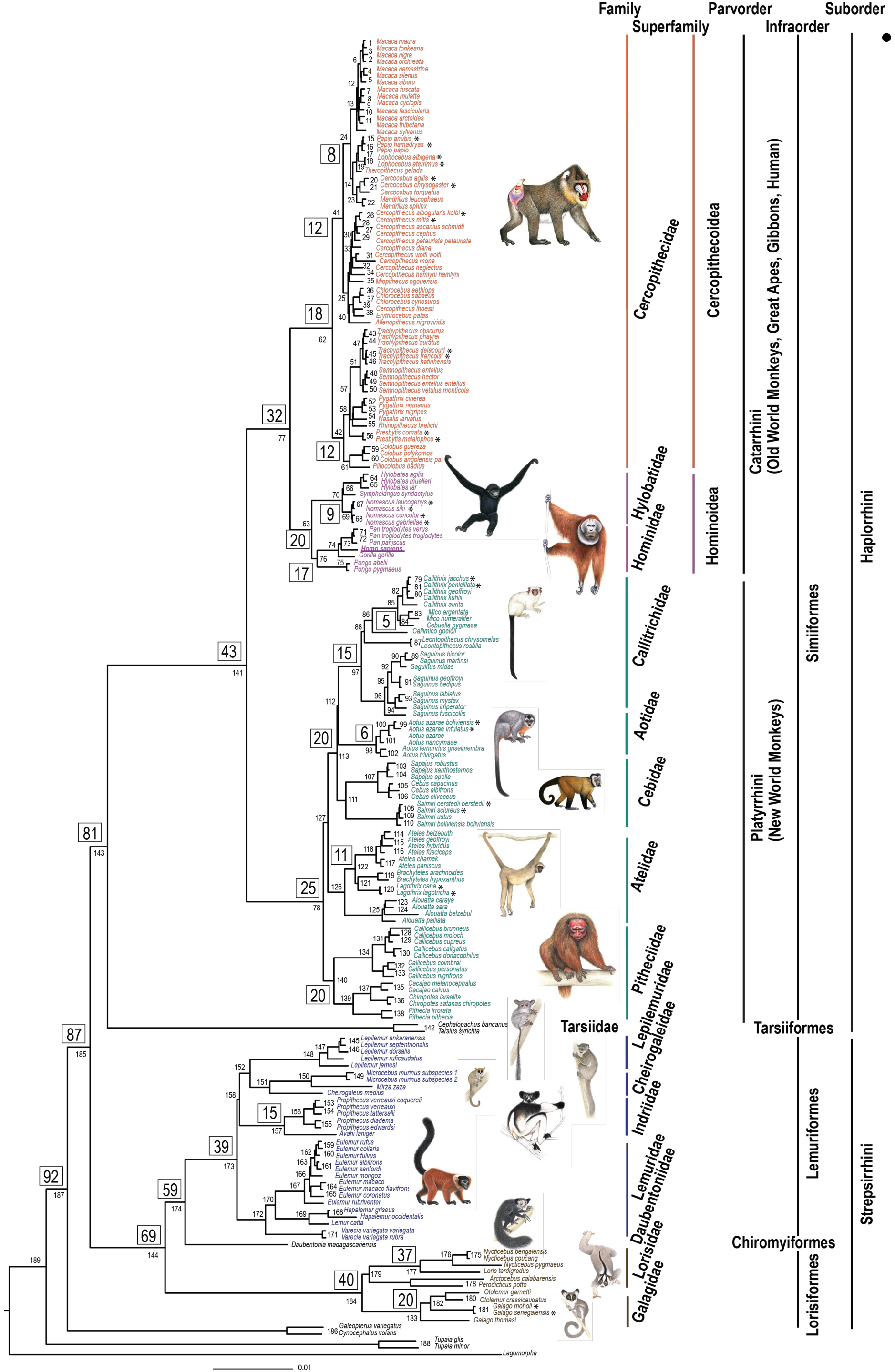


Table 1 | **Sample of Bayesian programs that use the molecular clock to estimate divergence times***

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].

NATURE REVIEWS | GENETICS VOLUME 17 | FEBRUARY 2016 | 75



- **DATING NODES (BEAST)**

Calibrated by fossil record: Strong support from consensus of paleoanthropologist with extensive citation record per dates

A Galigidae-Lorisidae split 38-42 MYA

B Simiiformes 36-50 MYA

C Catarrhini 20-38 MYA

D Platyrrhini 20-27 MYA

E Papionini 6-8 MYA

F Theropithecus 3.5-4.5 MYA

G Hominidae 13-18 MYA

H Homo-Pan split 6-7 MYA

5 Different Sets Of Species (BEAST):

1) Genus-level data set including 61 Primate genera, two Dermoptera genera and one Scandentia genus rooted by Lagomorpha,
2) Catarrhini species with outgroups,
3) Platyrrhini species with outgroups,
4) Strepsirrhini species with outgroups and
5) Genus-level analysis with a partitioned data set allowing for rate heterogeneity and different substitution models for autosome, X-chromosome, and Y-chromosome sequences.

Selection and Adaptation

Testing Coding Regions for Signals of Adaptive Evolution

- Is there evidence of selection operating on a gene?

- Where does selection occur within the gene? What regions, motifs, amino acids are under selection?

- Mapping the selection event on the phylogenetic tree. Is there a specific species or lineage that is experiencing selection?

- Assessing the form of selection (i.e. negative or positive) and identifying the codon (s), and *a priori* testing for statistical rigor.

- Are other genes exhibiting compensatory changes coincident with selection?

Testing for Selection in Aligned Sequences (MSA)

- Aligned codon sequences must be in frame with no stop codons.

- Remove recombination motifs and/or analyze partitions of MSA that are confirmed identical by descent.
  - Permits unbiased estimates of parameters

- A resolved phylogenetic tree of the multiple sequence file.
  - Pre-existing species tree established from other analyses
  - A poorly resolved tree will be unable to adequately test for selection and result in spurious results.

Molecular Selection In Coding Sequences

dN – nonsynonymous (missense) substitution

dS – synonymous substitutions

ω = dN/dS

ω = 1 (Neutral), no functional effect

ω < 1 (Purifying selection), selected to maintain function

ω > 1 (Diversifying selection), selected for adaptation to change in function

Genomics of Disease in Wildlife

Categories of Codon Selection Models

- Among sites:
  - Ho: Variable selection pressure possible among sites within YGOI (your-gene-of-interest) but no sites exhibit positive selection
- Among branches:
  - Ho: Average dN/dS is the same among all branches in the phylogeny for YGOI.
- Among clades:
  - Ho: Average dN/dS for YGOI is the same for each lineage within the phylogeny
- Branch-site:
  - Ho: Variable selection pressure is possible among sites with YGOI and no sites exhibit positive selection in any particular lineage relative to the rest of the phylogeny.

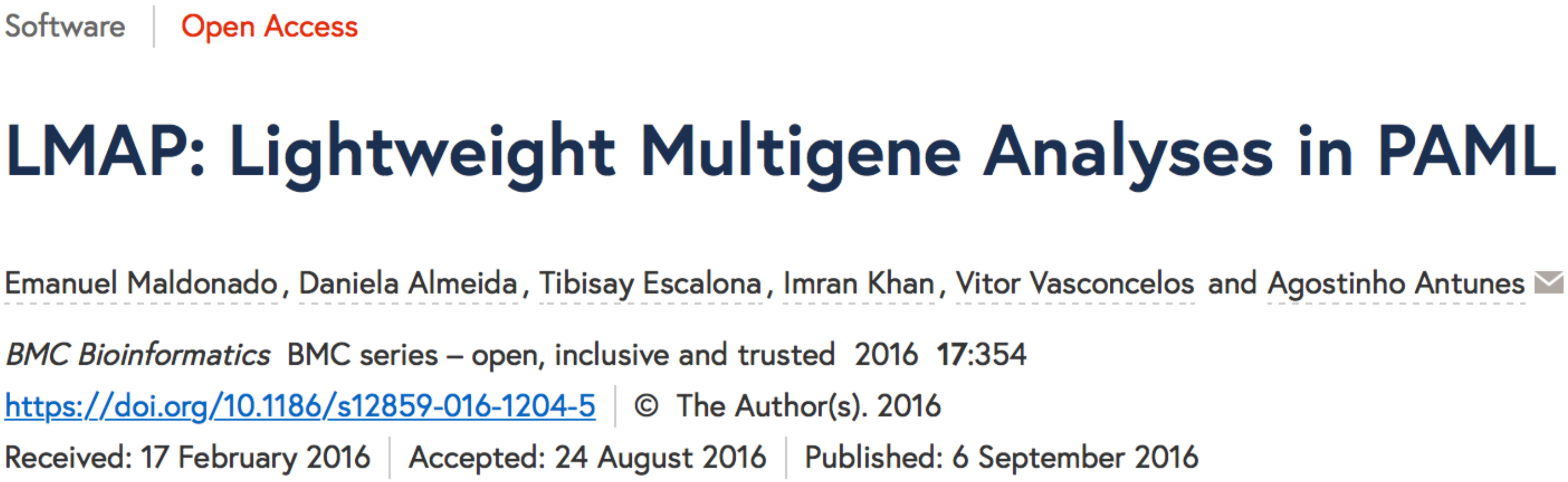
Why Use PAML?

- PAML offers a rigorous maximum likelihood algorithm to detect selection in MSA

- Conservative: Reduces the risk of a false positive result.

- Codon models incorporate empirical sequence characteristics
  - Instantaneous rate matrix
  - Transition: transversion bias
  - Codon frequency bias (due to redundancy in genetic code)

Scalable Version of PAML



Software | Open Access

LMAP: Lightweight Multigene Analyses in PAML

Emanuel Maldonado, Daniela Almeida, Tibisay Escalona, Imran Khan, Vitor Vasconcelos and Agostinho Antunes ✉

BMC Bioinformatics BMC series – open, inclusive and trusted 2016 **17**:354

https://doi.org/10.1186/s12859-016-1204-5 | © The Author(s). 2016

Received: 17 February 2016 | Accepted: 24 August 2016 | Published: 6 September 2016

Genomics of Disease in Wildlife

Files and Format Needed for CodeML Analyses

- Data file:
  - Nucleotide data
  - No stop codons
  - PHYLIP format
- Treefile
  - Topology only, no branch lengths (suggested)
  - Tree must reflect true phylogeny
  - Unresolved nodes will decrease CodeML utility
  - PHYLIP (newick) format
- Control file
  - Text file containing specific settings for Model analysis

GDW
Genomics of Disease in Wildlife

Basic Steps to CodeML

Create 3 files: data, tree and control

Load files into GUI interface

Select Parameters for appropriate test

> Depending on numbers of sequences, genetic information, pattern of mutation, length of sequence.....
>
> CodeML can take minutes or hours to run

Record Ln likelihood value

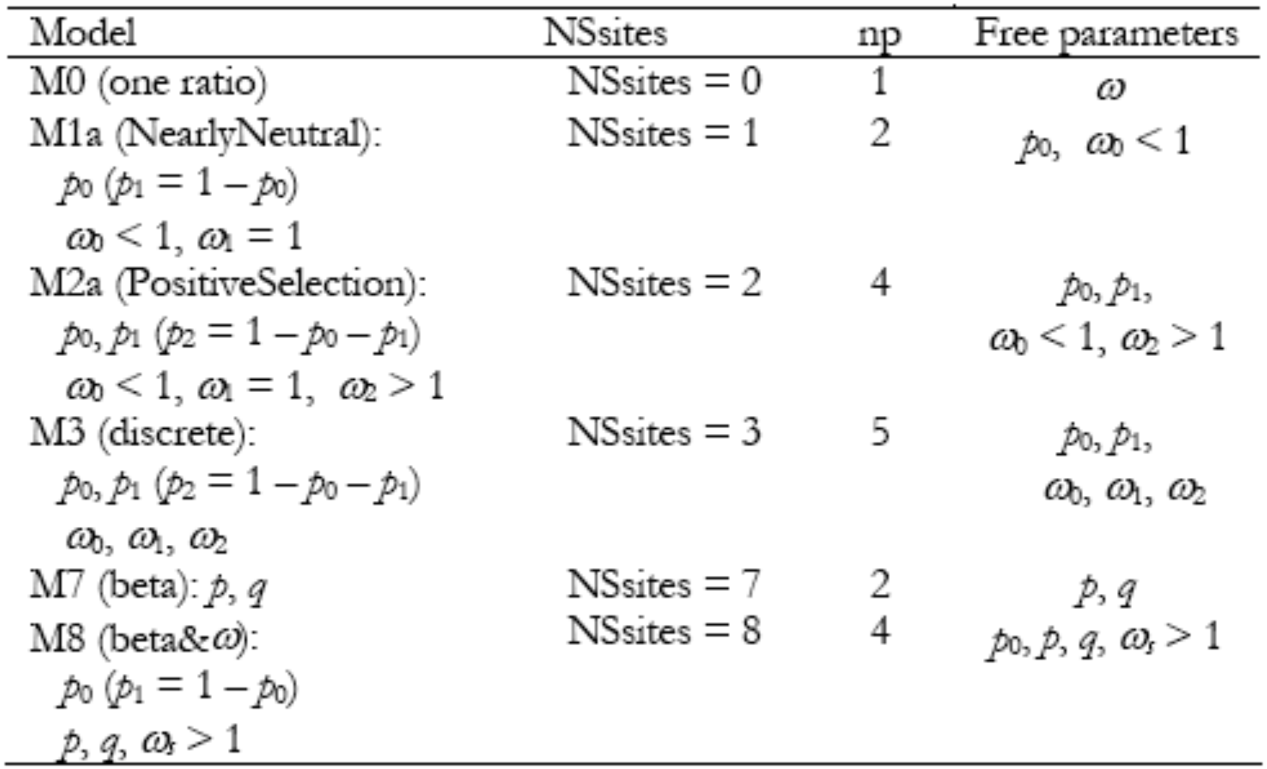
> Compare with ln likelihood of null model.
>
> Determine significance by the log-likelihood ratio model (LRT).
>
> $\Delta\lambda = 2\ (l_1 - l_0)$ chi-square 2 d.f.

For Branch-Site Models

> $\Delta\lambda = 2\ (l_1 - l_0)$ chi-square 2 d.f. P-value/2

GDW
Genomics of Disease in Wildlife

Site Models

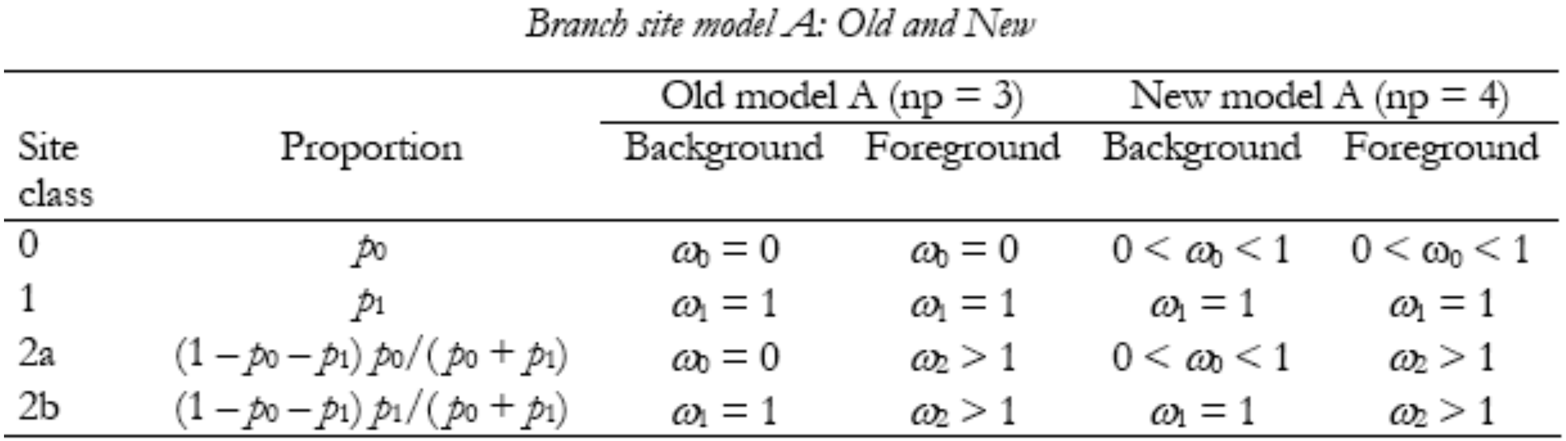


Model	NSsites	np	Free parameters
M0 (one ratio)	NSsites = 0	1	ω
M1a (NearlyNeutral): p_0 ($p_1 = 1 - p_0$) $\omega_0 < 1$, $\omega_1 = 1$	NSsites = 1	2	p_0, $\omega_0 < 1$
M2a (PositiveSelection): p_0, p_1 ($p_2 = 1 - p_0 - p_1$) $\omega_0 < 1$, $\omega_1 = 1$, $\omega_2 > 1$	NSsites = 2	4	p_0, p_1, $\omega_0 < 1$, $\omega_2 > 1$
M3 (discrete): p_0, p_1 ($p_2 = 1 - p_0 - p_1$) ω_0, ω_1, ω_2	NSsites = 3	5	p_0, p_1, ω_0, ω_1, ω_2
M7 (beta): p, q	NSsites = 7	2	p, q
M8 (beta&ω): p_0 ($p_1 = 1 - p_0$) p, q, $\omega_s > 1$	NSsites = 8	4	p_0, p, q, $\omega_s > 1$

LRT

Model M1a and Model M2a, 2 df
Model M7 and Model M8, 2 df

Branch Site Models



Branch site model A: Old and New

Site class	Proportion	Old model A (np = 3)		New model A (np = 4)	
		Background	Foreground	Background	Foreground
0	p_0	$\omega_0 = 0$	$\omega_0 = 0$	$0 < \omega_0 < 1$	$0 < \omega_0 < 1$
1	p_1	$\omega_1 = 1$	$\omega_1 = 1$	$\omega_1 = 1$	$\omega_1 = 1$
2a	$(1 - p_0 - p_1)\, p_0/(p_0 + p_1)$	$\omega_0 = 0$	$\omega_2 > 1$	$0 < \omega_0 < 1$	$\omega_2 > 1$
2b	$(1 - p_0 - p_1)\, p_1/(p_0 + p_1)$	$\omega_1 = 1$	$\omega_2 > 1$	$\omega_1 = 1$	$\omega_2 > 1$

- Branch-site
  - Model A recommended
    - Model = 2
    - NSsites=2

Compare LRT with Null Model A

Model=2

NSsites=2

But, ω=1, fixed.

Gene Trees and Species Trees

- **Gene tree**-genealogical history of the gene or gene family over time.
- **Species tree**-genealogical relationships of species from which genes were sequenced
- **Concatenation methods**-A single sequence per taxon that is a composite of multiple gene regions selected across the genome.
- **Incomplete lineage sorting**-evolutionary history of genes are discordant with patterns of speciation.
- **Multi-Species Coalescent methods**-individual gene trees are simultaneously or separately estimated and then a summation method to ascertain phylogenetic relationships.

Incomplete Lineage Sorting (ILS)

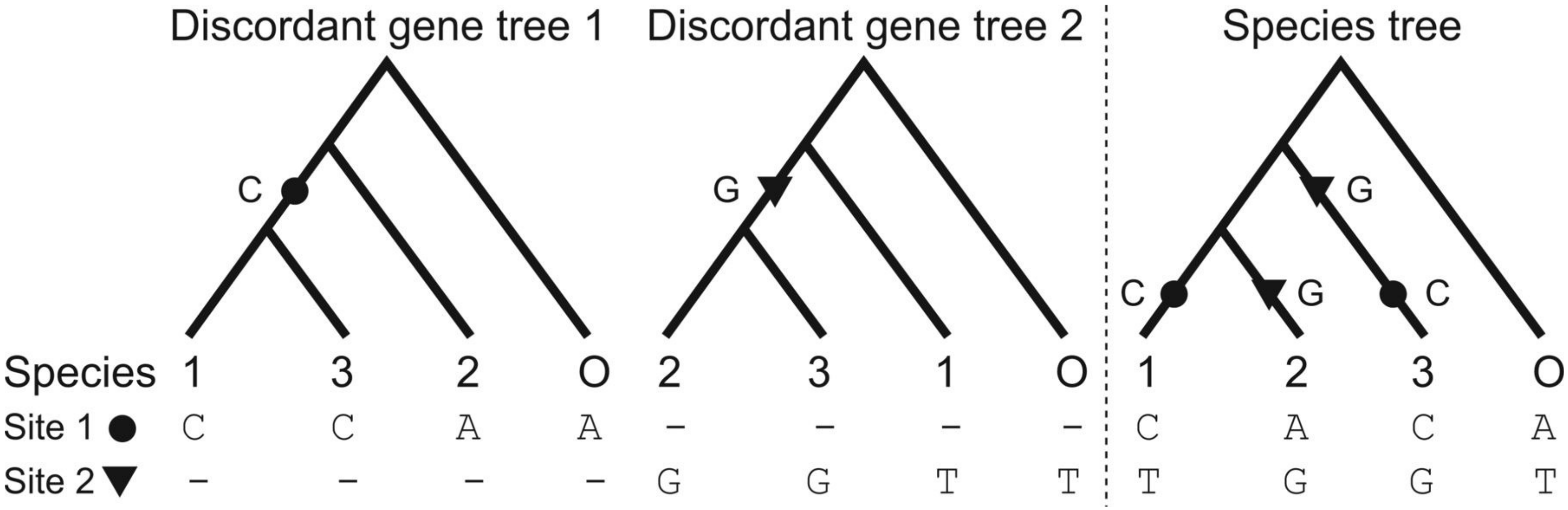
- Incomplete lineage sorting occurs when gene trees do not reflect true species phylogeny (reciprocal monophyly). The genes, due to biological processes such as inter-species hybridization, introgression, reticulate evolution, or molecular processes (i.e mutation rate, recombination, or genome function), provide alternate topologies.

- Characteristic in radiations characterized by short internal internode distances correlated with rapid diversification.

- Usually nodes are weakly supported (<70%) with bootstrap values or Bayesian proportions (<0.9)

- Concatenation studies tend to ignore these genes alternate topologies as it generates too much ‘noise’ or homoplasy (convergent, reversal, multiple hits)
- **But they are actually incredibly interesting**!

Two discordant gene trees & their species tree.



Species	1	3	2	O	2	3	1	O	1	2	3	O
Site 1 ●	C	C	A	A	–	–	–	–	C	A	C	A
Site 2 ▼	–	–	–	–	G	G	T	T	T	G	G	T

Mol Biol Evol, Volume 33, Issue 12, December 2016, Pages 3299–3307, https://doi.org/10.1093/molbev/msw197