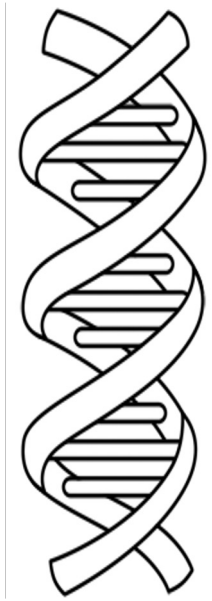
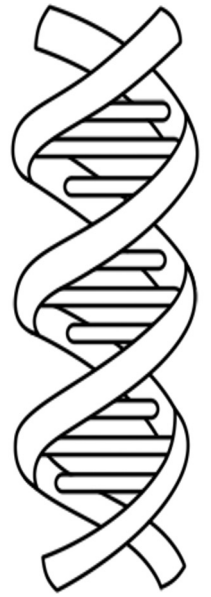


Molecular phylogenetic inference with ancient DNA

Audrey T. Lin, DPhil

American Museum of Natural History



Phylogenetics

The study of evolutionary relationships between taxa

- Species
- Genes
- Individuals

Different (heritable) data can be used:

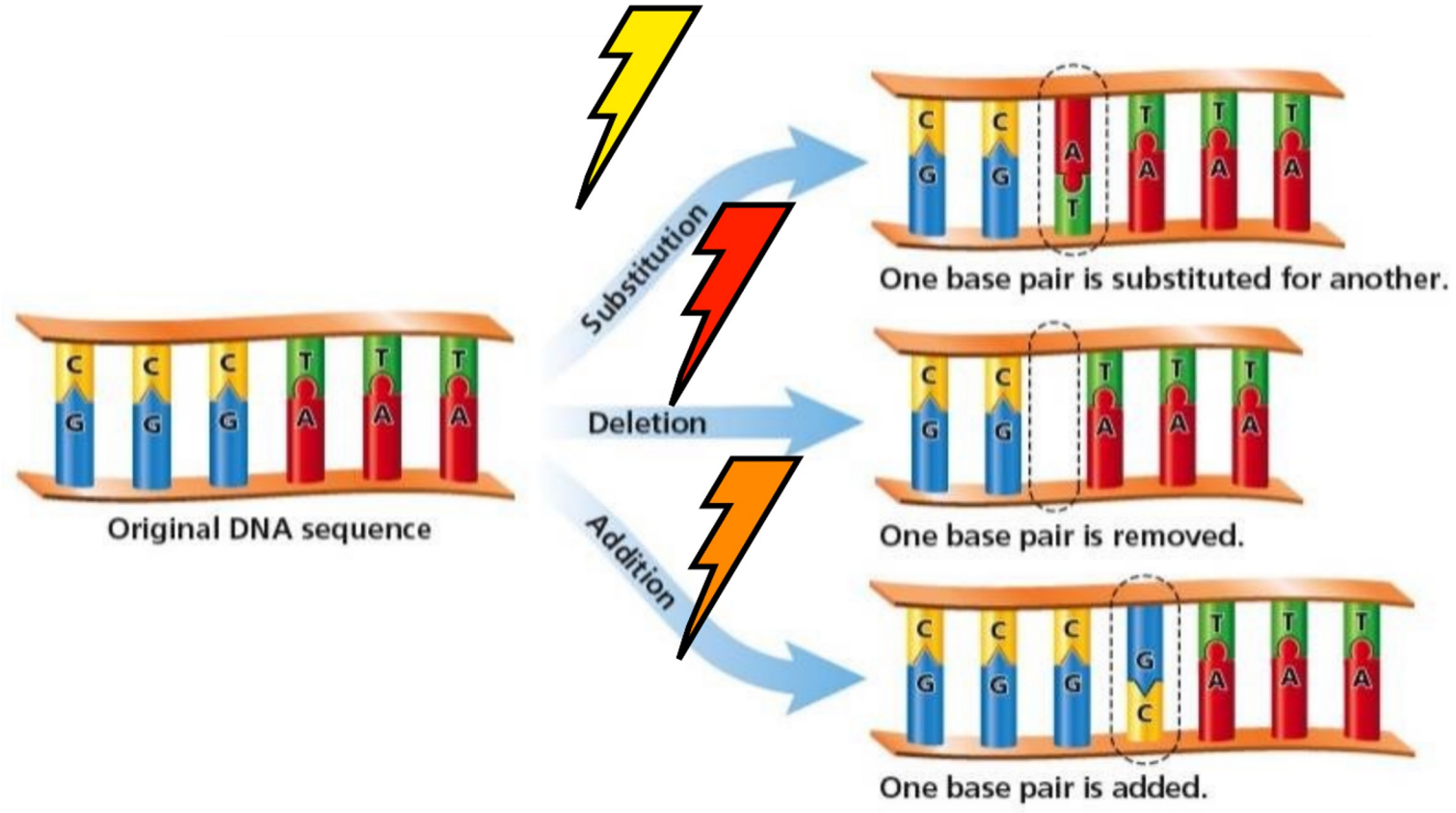
- Morphology
- Protein Sequences
- Genetic Sequences

Computational phylogenetics

- **1966-1981 - the early years**
  - Maximum parsimony introduced
  - Least squares
- **1981-1991 - idealogical warfare and the "Dark Ages" for systematics and molecular evolution**
  - Maximum Parsimony and cladistics peaked
  - Maximum likelihood pruning algorithm introduced (1981)
  - Neighbour-joining introduced (1987)
- **1991- 2001 - the statistical phylogenetics revolution and "a reasonably happy ending"**
  - maximum likelihood matures, parametric bootstrap, KH test et cetera.
  - Bayesian phylogenetics introduced (1996)
- **2001-2010 - Bayesian phylogenetics revolution**
  - MrBayes, BEAST, BayesPhylogeny, PhyloBayes, PHYCAS et cetera
- **2010-2019 - Phylogenomics revolution**
  - Multispecies coalescent, Fossilized birth-death models, methods that go beyond trees are maturing
  - BEAST2, RevBayes
- **2020—now - Integrative phylogenomics? Non-MCMC Bayesian methods? Real-time phylogenetics?**

Source: Taming the Beast

DNA changes occur with time

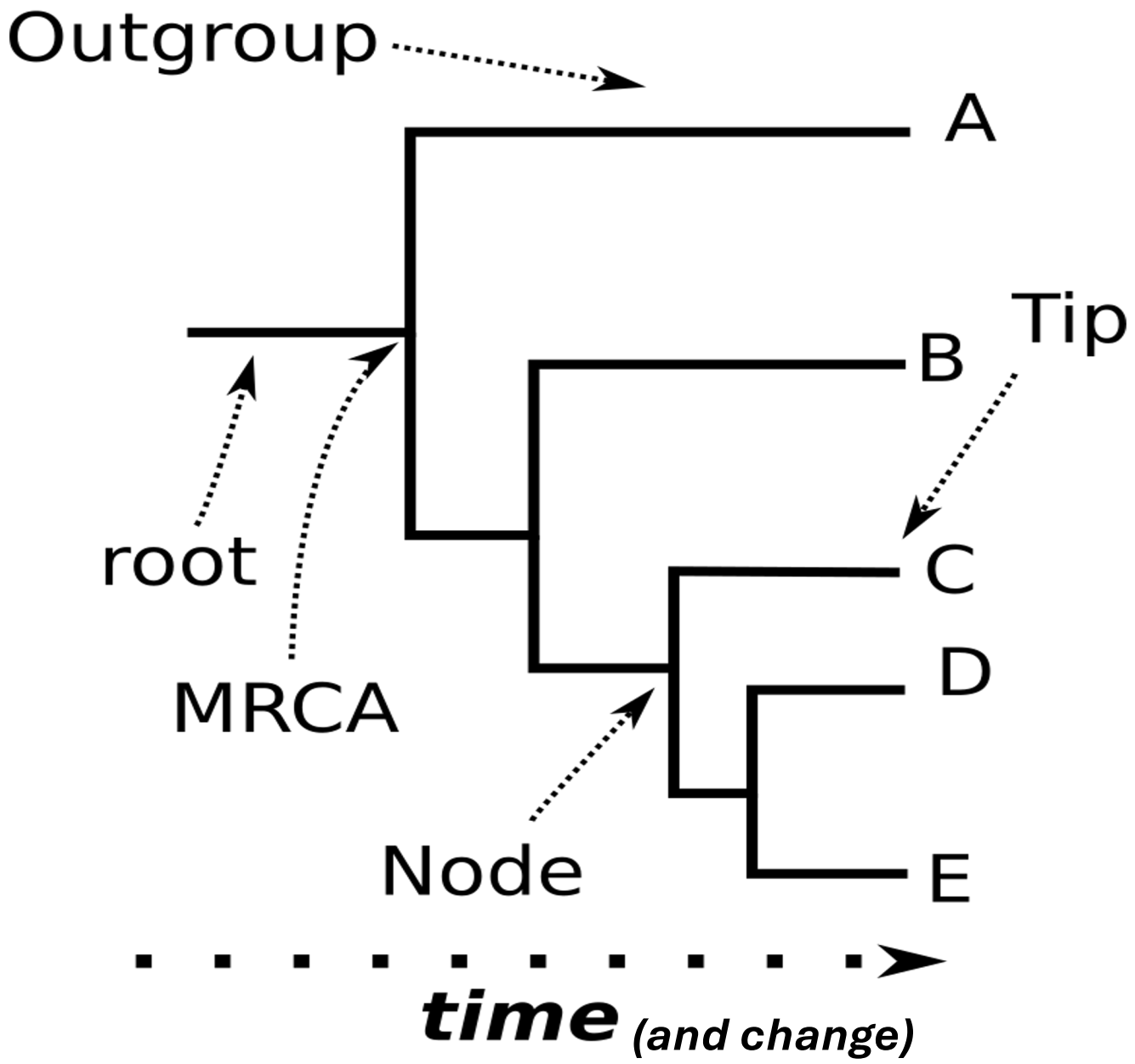


Slides adapted from K Daly

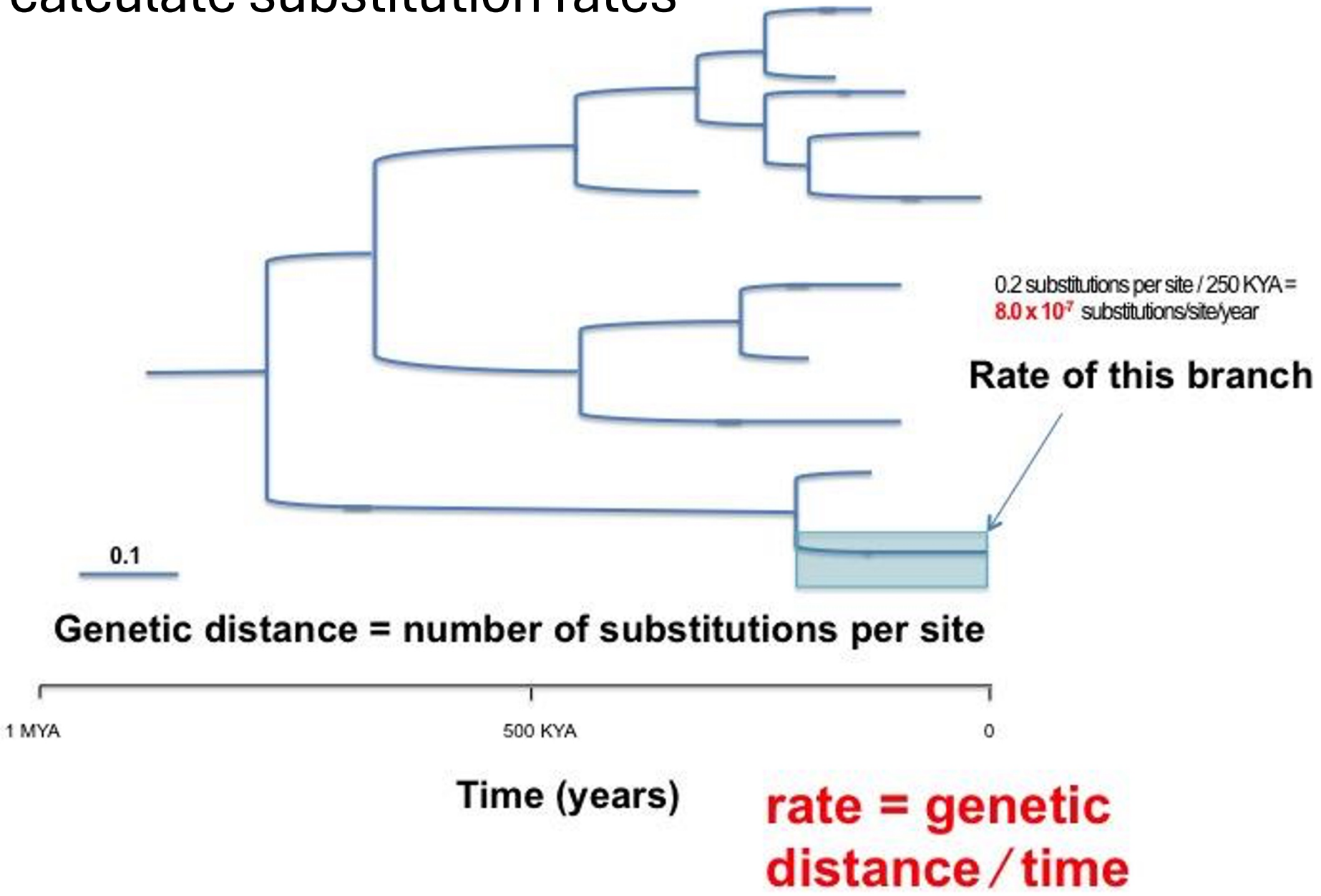
Neutral theory of molecular evolution

- States that most evolutionary changes at the molecular level are **not** caused by selection, but by genetic drift.

evolution.berkeley.edu

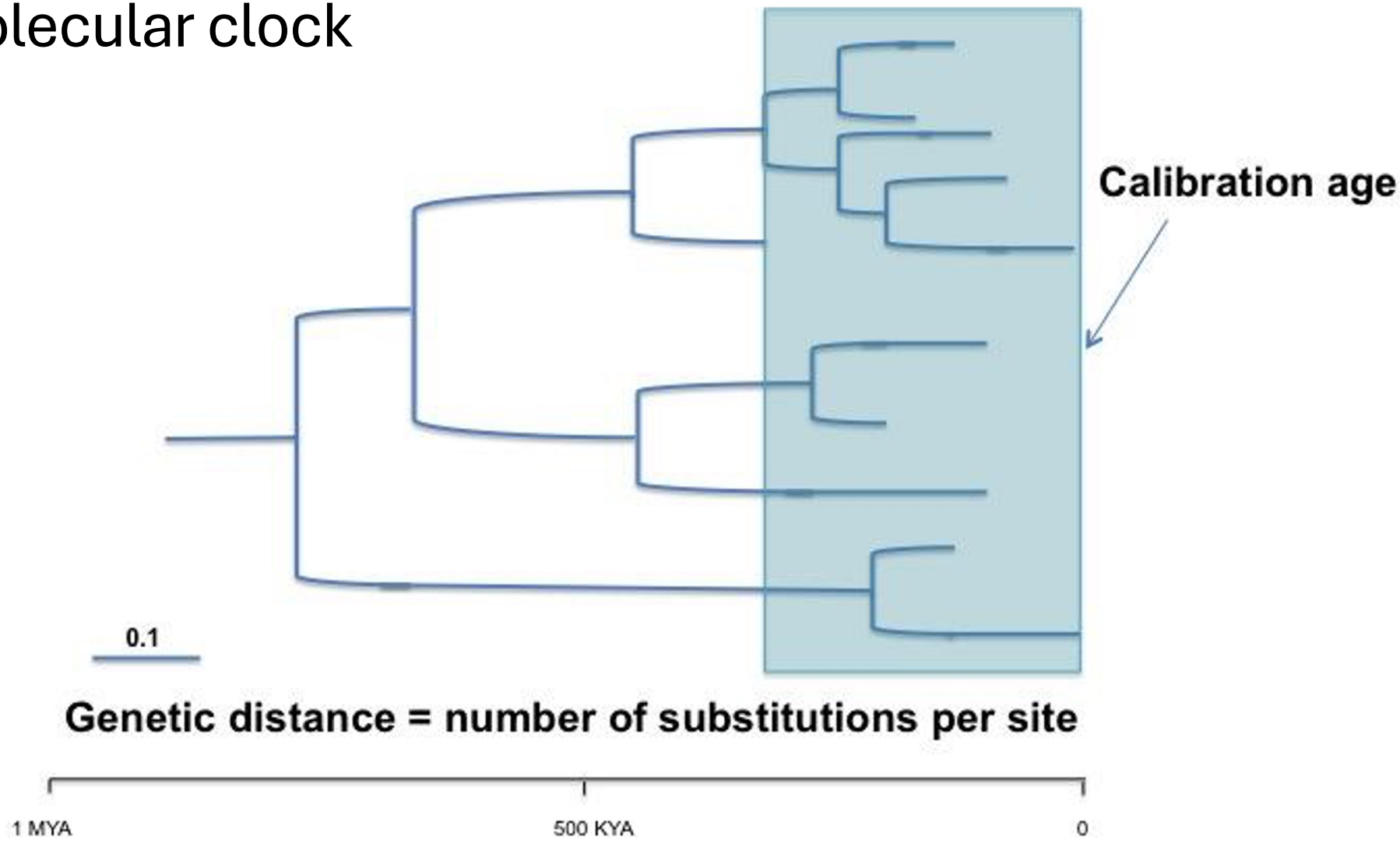

Outgroup
A
B
Tip
root
C
MRCA
D
Node
E
time (and change)

How to calculate substitution rates

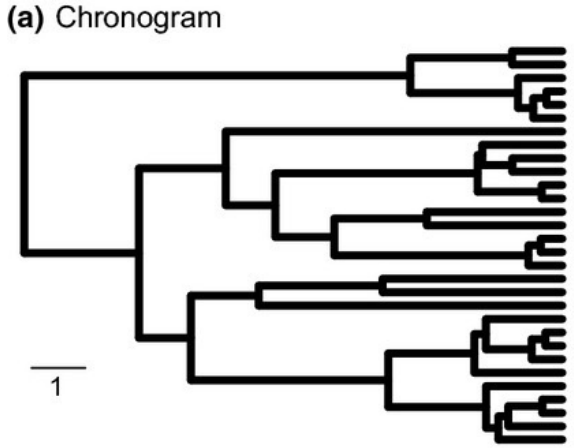


rate = genetic distance / time

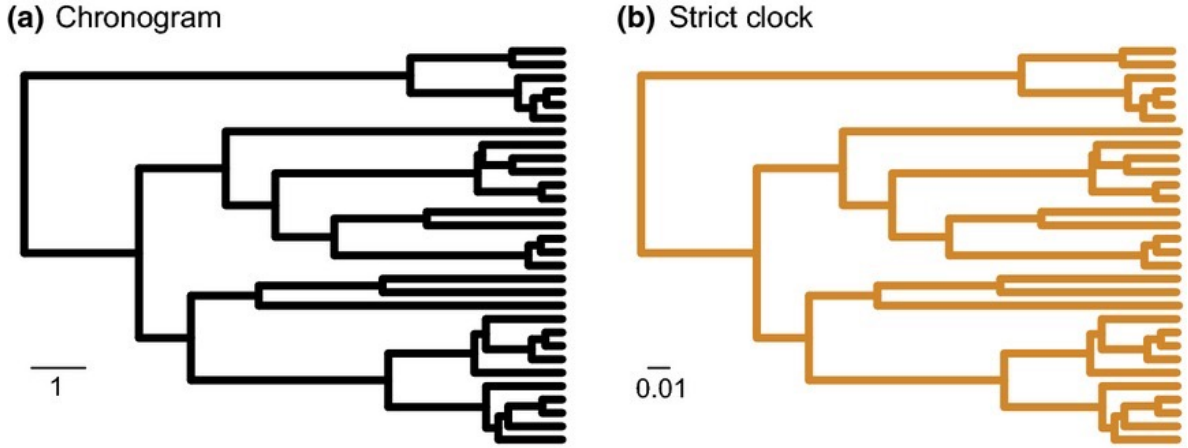
The molecular clock



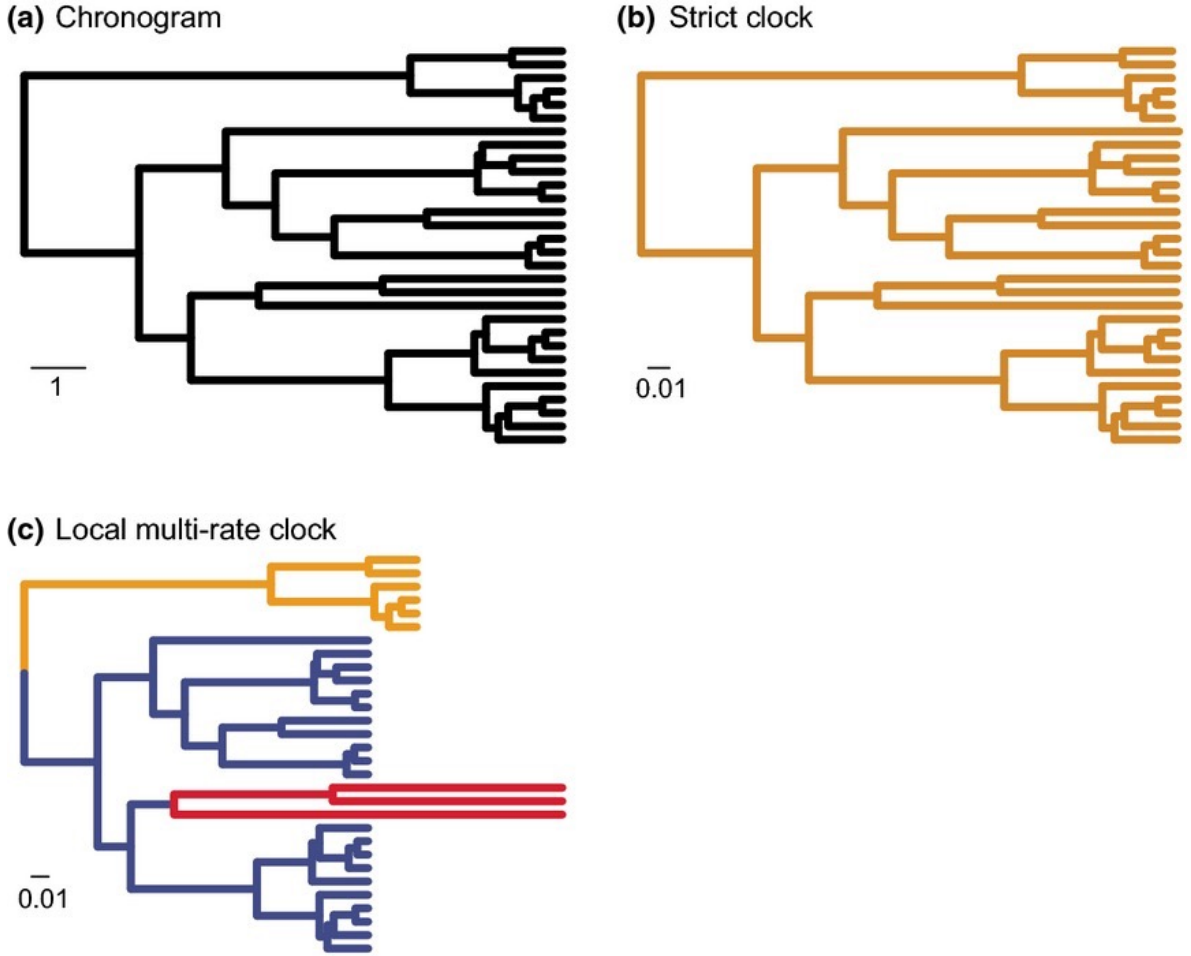
rate = genetic distance / time



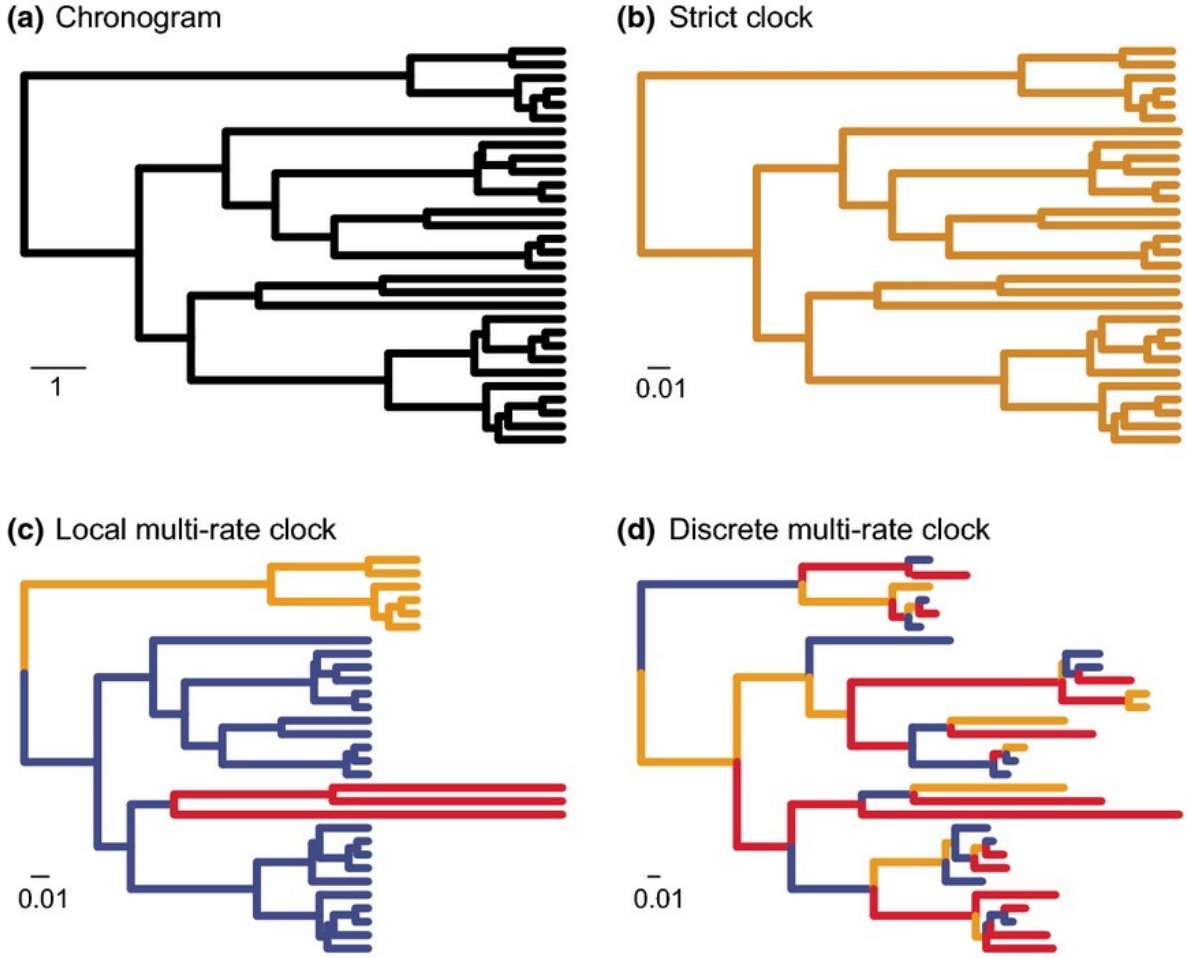
Ho and Duchene (2014)



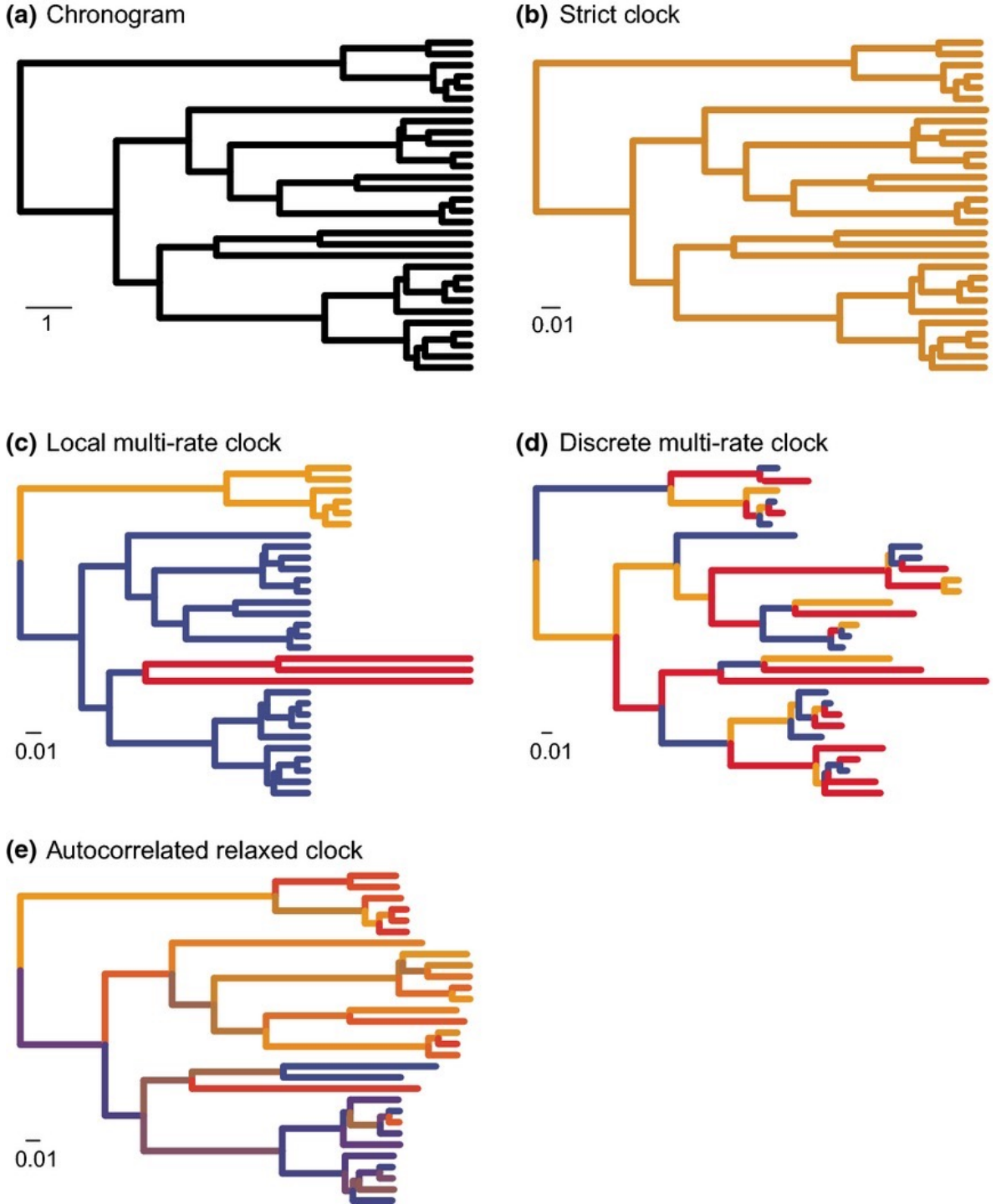
Ho and Duchene (2014)



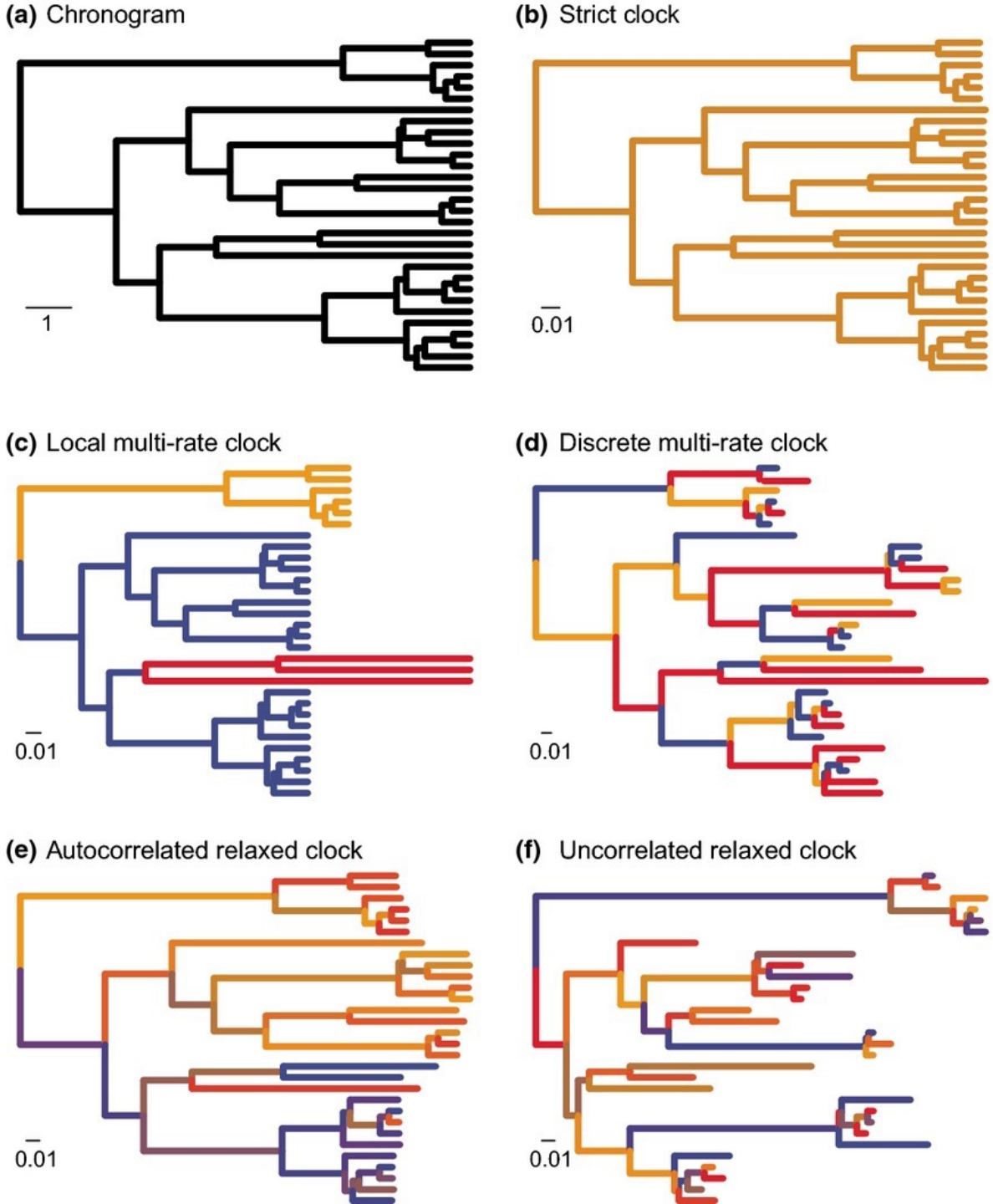
Ho and Duchene (2014)



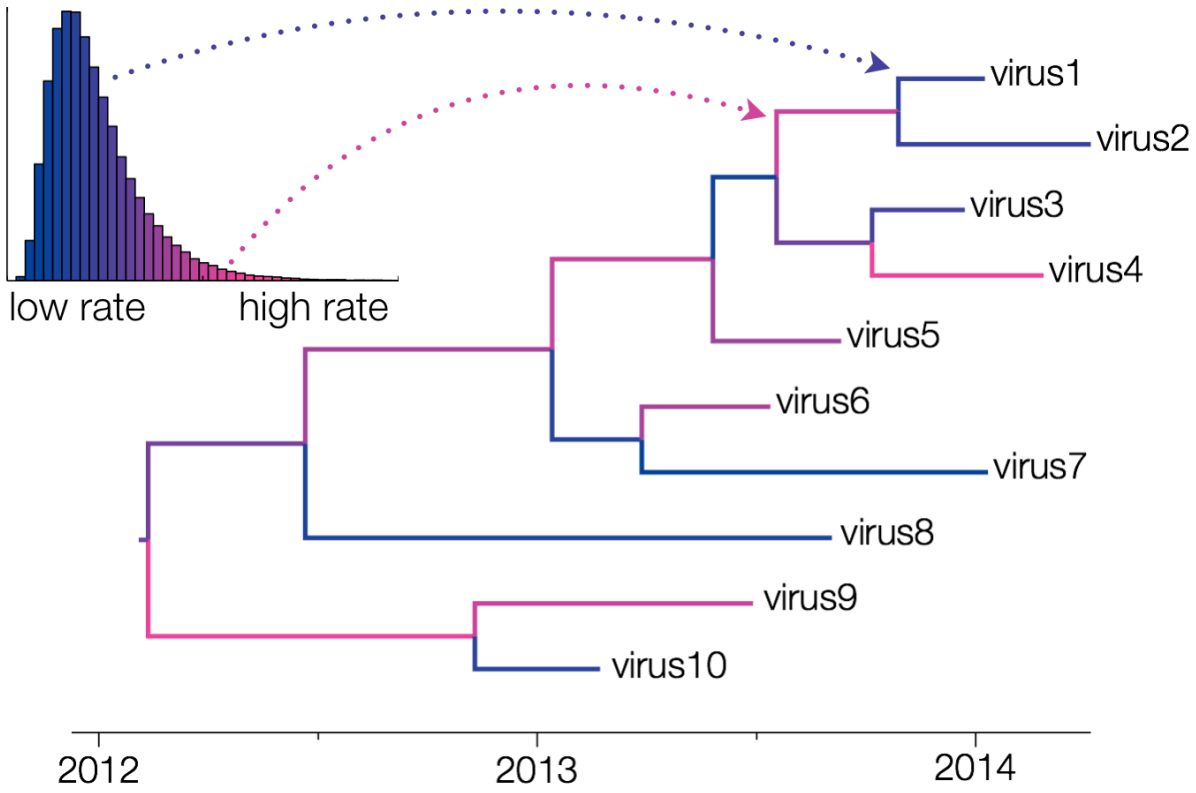
Ho and Duchene (2014)



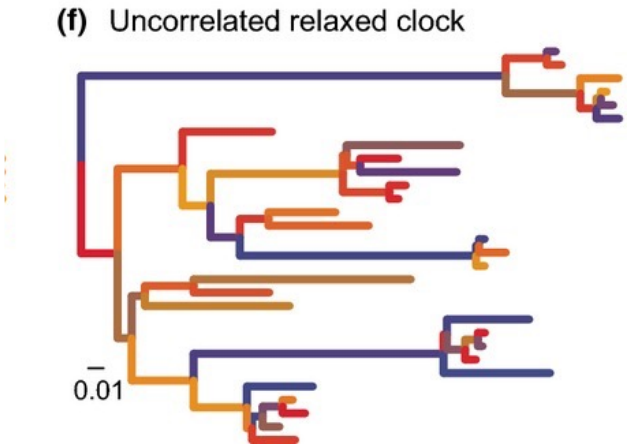
Ho and Duchene (2014)



Ho and Duchene (2014)



https://beast.community/clocks



Ho and Duchene (2014)

Questions to ask:

- What data should we use?

Questions to ask:

- What data should we use?
- Which method should we use?

Questions to ask:

- What data should we use?
- Which method should we use?
- Which evolutionary model should we use?

What data to use?

- Great care must be taken in the data used to reconstruct a tree that accurately reflects the evolutionary history

What data to use?

- Great care must be taken in the data used to reconstruct a tree that accurately reflects the evolutionary history
- Ideally the choice is a genomic region that appears exactly once in every species and whose evolutionary history is “identical” to that of the species

What data to use?

- Great care must be taken in the data used to reconstruct a tree that accurately reflects the evolutionary history
- Ideally the choice is a genomic region that appears exactly once in every species and whose evolutionary history is “identical” to that of the species
- The region should have very little if any traces of HGT

What data to use?

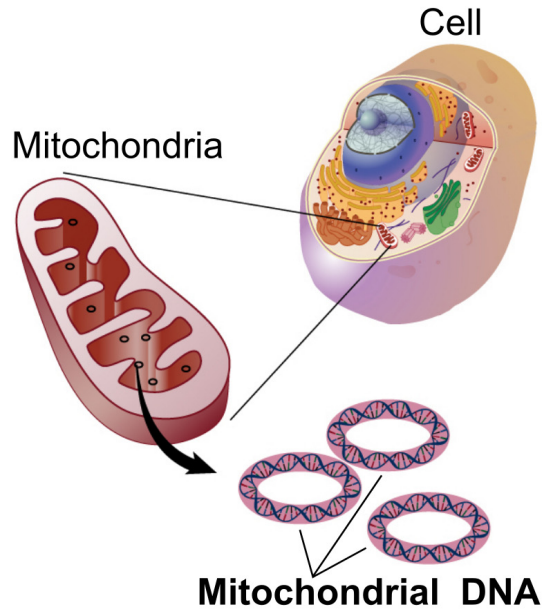
- Great care must be taken in the data used to reconstruct a tree that accurately reflects the evolutionary history
- Ideally the choice is a genomic region that appears exactly once in every species and whose evolutionary history is “identical” to that of the species
- The region should have very little if any traces of HGT
- The rate of changes in the region should be fast enough to distinguish between closely related species, but not so fast that regions from very distantly related species cannot be reliably aligned

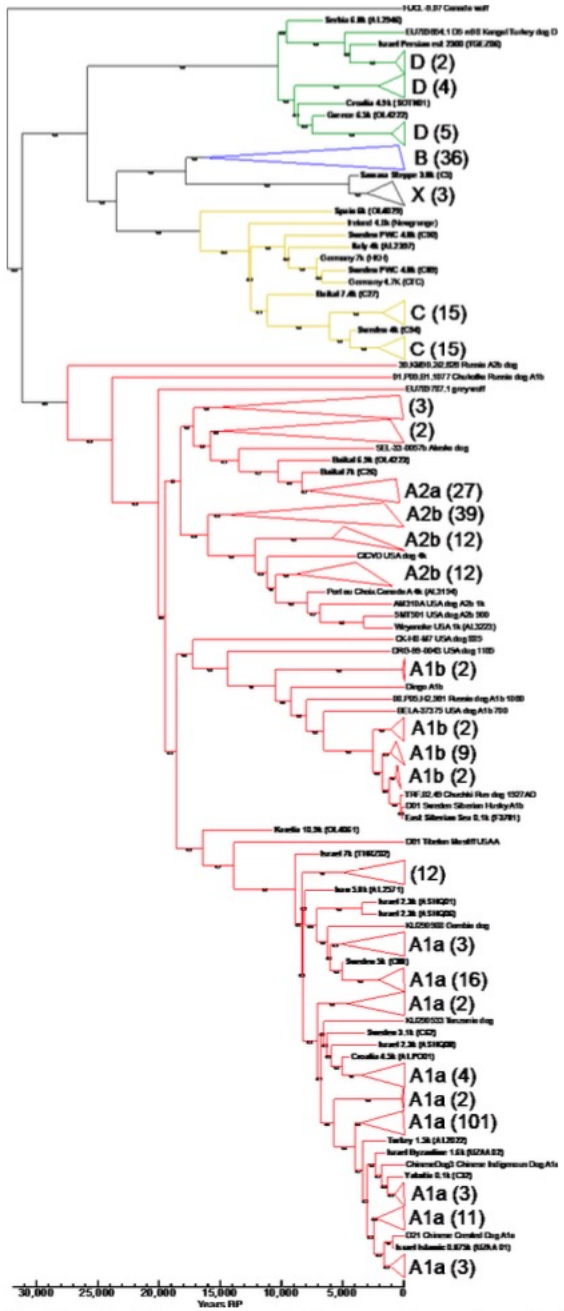
What data to use?

- Great care must be taken in the data used to reconstruct a tree that accurately reflects the evolutionary history
- Ideally the choice is a genomic region that appears exactly once in every species and whose evolutionary history is “identical” to that of the species
- The region should have very little if any traces of HGT
- The rate of changes in the region should be fast enough to distinguish between closely related species, but not so fast that regions from very distantly related species cannot be reliably aligned
    - Small ribosomal subunit rRNA
    - mtDNA (d-loop/control region, etc)

mtDNA

- Shorter sequences
- High copy number
- Can be used for DNA barcoding (~700 bp) to quickly identify species
- Relatively fewer differences between species
- No introns
- Whole mtDNA has enough markers to differentiate between populations
- Can’t really make inferences on admixture/gene flow/recombination/rate heterogeneity/incomplete lineage sorting





D (mostly Near East farmers)

B (Eurasia)

C (Mesolithic Eurasian)

A2a (Arctic)

A2b (Indigenous N Am)

A1b (Asia/Arctic)

A1a (Eurasia/most common dog haplotype in the world)

A

Dog mtDNA haplotypes

Bergstrom et al. 2020

Choice of method

- Many computational methods exist for phylogeny reconstruction
- 2 categories – distance-based methods & sequence-based methods
- **Distance-based methods** first compute pairwise distances from the sequences, and then use these distances to obtain the tree
  - UPGMA (unweighted pair-group method using arithmetic averages)
  - NJ (neighbor-joining)

Choice of method

- **Sequence-based methods** use the sequence alignment directly, and usually search the tree space using an optimality criterion that is defined on the columns of the alignment
  - Maximum parsimony (MP)
  - Maximum likelihood (ML)
  - Bayesian inference

Choice of model of evolution

- **p-distance** uses the proportion (**p**) of nucleotide sites at which two sequences compared are different.
  - Divide the # of nucleotide differences (n_d) by the total number of nucleotides compared (n).
  - $\boldsymbol{p = n_d/n}$
- Usually an underestimate of true evolutionary distance
  - Correction is therefore necessary
  - Models of evolution can be used to derive the distance corrections

Examples of substitution models

- **Jukes-Cantor (JC)** – equal base frequencies, all substitutions equally likely
- **Felsenstein 1981 (F81)** – variable base frequencies, all substitutions equally likely
- **Kimura 2-parameter (K80)** – equal base frequencies, one transition rate one transversion rate
- **Hasegawa-Kishino-Yano (HKY)** – variable base frequencies, one transition rate, one transversion rate
- **General time reversible (GTR)** – variable base frequencies, different rates for transitions and transversions

Models describing rate variation in a sequence

- **Gamma distribution (G)** – gamma distributed rate variation among sites
- **Proportion of invariable sites (I)** – extent of static, unchanging sites in a dataset

JC69

- All rates are equal
- 1-parameter

	A	T	C	G
A	-	λ	λ	λ
T	λ	-	λ	λ
C	λ	λ	-	λ
G	λ	λ	λ	-

HKY85

- Ts/Tv rate bias
- Base composition bias
- 5-parameters

	A	T	C	G
A	-	$\beta\pi_T$	$\beta\pi_C$	$\alpha\pi_G$
T	$\beta\pi_A$	-	$\alpha\pi_C$	$\beta\pi_G$
C	$\beta\pi_A$	$\alpha\pi_T$	-	$\beta\pi_G$
G	$\alpha\pi_A$	$\beta\pi_T$	$\beta\pi_C$	-

GTR (REV)

- Time reversible
- Different rates
- 9-parameters

	A	T	C	G
A	-	$a\pi_T$	$b\pi_C$	$d\pi_G$
T	$a\pi_A$	-	$c\pi_C$	$e\pi_G$
C	$b\pi_A$	$d\pi_T$	-	$f\pi_G$
G	$c\pi_A$	$e\pi_T$	$f\pi_C$	-

Nonhomogeneous and nonstationary

- Bayesian
- Fixed number of sets of base frequencies
- Assignment of the branches to the sets varies by MCMC

+I proportion of invariant sites

+Γ rate variation across sites

1969 — 1980 — 1985 — 1989 — 1993 — 1994 — 1995 — 2004

Observations

- Count the number of different bases
- p-distance

K80

- Ts/Tv rate bias
- 2-parameters

	A	T	C	G
A	-	β	β	α
T	β	-	α	β
C	β	α	-	β
G	α	β	β	-

TN93

- Purine/Pyramidine rates
- Ts/Tv rate bias
- 6-parameters

	A	T	C	G
A	-	$\beta\pi_T$	$\beta\pi_C$	$\alpha_1\pi_G$
T	$\beta\pi_A$	-	$\alpha_2\pi_C$	$\beta\pi_G$
C	$\beta\pi_A$	$\alpha_2\pi_T$	-	$\beta\pi_G$
G	$\alpha_1\pi_A$	$\beta\pi_T$	$\beta\pi_C$	-

UNREST

- All different rates
- Not time reversible
- 12-parameters

	A	T	C	G
A	-	a	b	c
T	d	-	e	f
C	g	h	-	i
G	j	k	l	-

Nonhomogeneous and nonstationary

- Maximum likelihood
- Different base frequency parameters in the rate matrix for different branches

Credit: Jose Barba Montoya

Model selection tools

- jModelTest2
- PartitionFinder
- IQTREE
- BModelTest (implemented in BEAST2)

Exercises!?

Do we even need a model selection step???

- Abadi et al. 2019, Nat Comm: “Model selection may not be a mandatory step for phylogeny reconstruction”
- Fabreti and Höhna, 2023: “Nucleotide Substitution Model Selection Is Not Necessary for Bayesian Inference of Phylogeny With Well-Behaved Priors”
    - Suggests skipping model selection and instead using the most parameter-rich model, GTR+I+G leads to similar inference

Tip-dating: molecular dating of phylogenetic trees

- Using sequence data to co-estimate the timing of evolutionary events and rates of molecular evolution

Tip-dating: molecular dating of phylogenetic trees

- Using sequence data to co-estimate the timing of evolutionary events and rates of molecular evolution
  - Fast-dating methods: LSD, Treetime, Treedater, Treepl, Reltime
  - These methods need a starting tree (e.g. ML tree from RAxML)

Molecular clock phylogenies with Bayesian methods

- BEAST
- BEAST2
- MrBayes
- RevBayes

- MCMCTree (cannot directly infer tree from alignment, need to provide starting tree)

- Probably more…

Upcoming workshops

Coming soon...

Past workshops

Squamish BEAST Tamers

- August 14, 2023-August 18, 2023
- Squamish, British Columbia
- Programme
- Lecture slides
- Flyer

Taming the BEAST Online

- June 07, 2021-June 11, 2021
- Online!
- Programme
- Lecture slides
- Flyer

Taming the BEAST 2020

- **Cancelled**
- Oberägeri, Switzerland
- Programme
- Flyer

Taming the BEAST Eh!

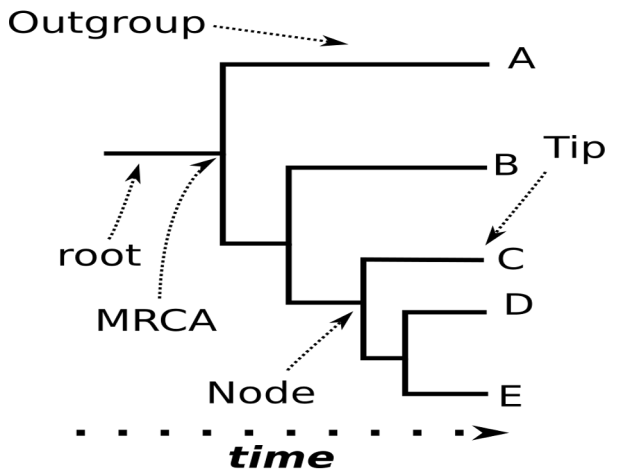
- August 12, 2019-August 16, 2019
- Squamish, British Columbia
- Programme
- Lecture slides
- Flyer

Tip-dating: molecular dating of phylogenetic trees

- Using sequence data to co-estimate the timing of evolutionary events and rates of molecular evolution
- Requires converting genetic divergence between sequences into absolute time

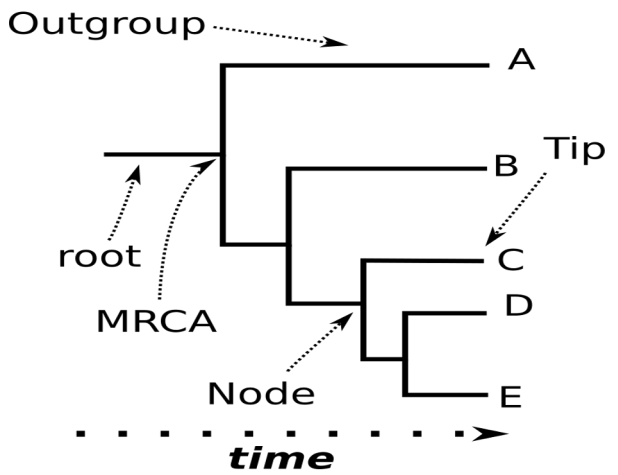
Tip-dating: molecular dating of phylogenetic trees

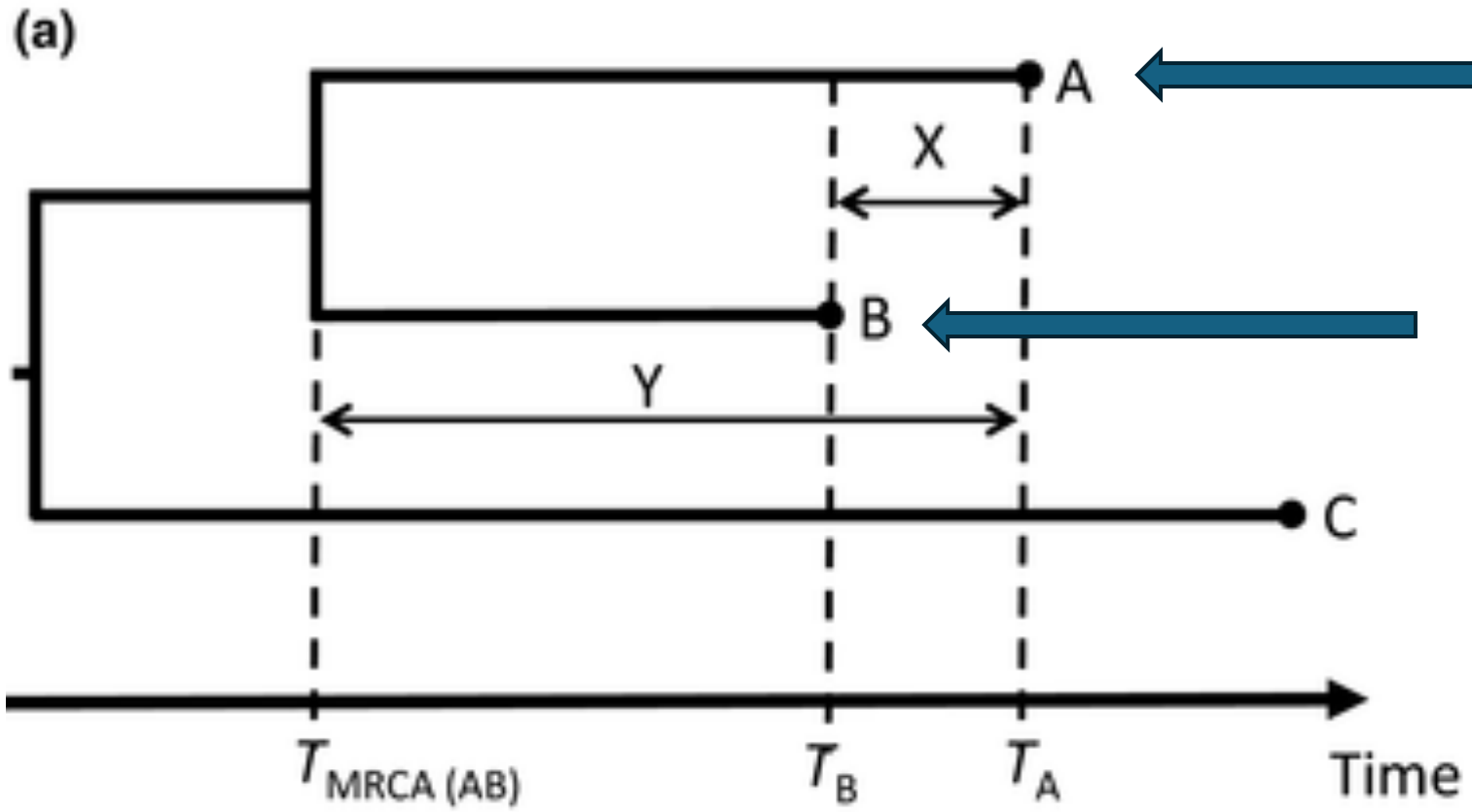
- Using sequence data to co-estimate the timing of evolutionary events and rates of molecular evolution
- Requires converting genetic divergence between sequences into absolute time
- Tip-dating origins in study of fast-evolving genomes (e.g. viruses, some bacteria)



Tip-dating: molecular dating of phylogenetic trees

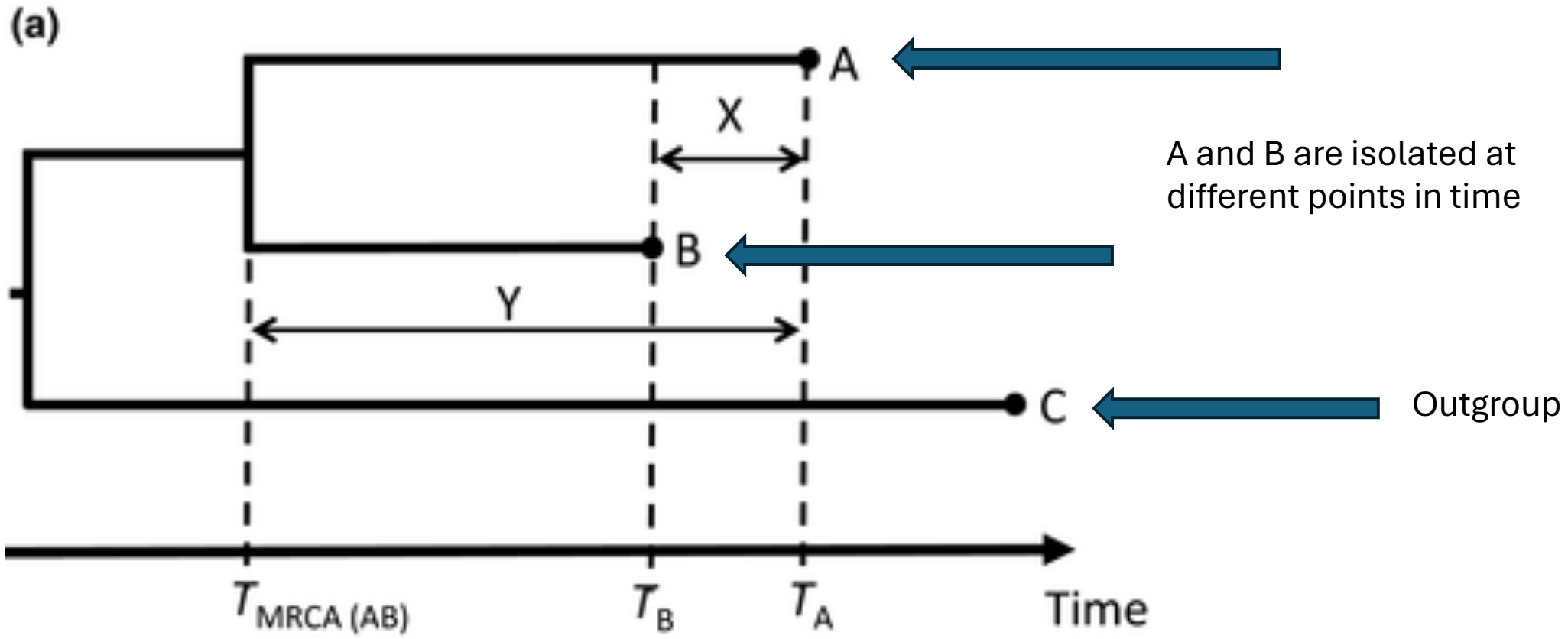
- Using sequence data to co-estimate the timing of evolutionary events and rates of molecular evolution
- Requires converting genetic divergence between sequences into absolute time
- Tip-dating origins in study of fast-evolving genomes (e.g. viruses, some bacteria)
- Genomic data of ancient DNA extends tip-based calibration to variety of other taxa





Tips of the tree: A and B are isolated at different points in time

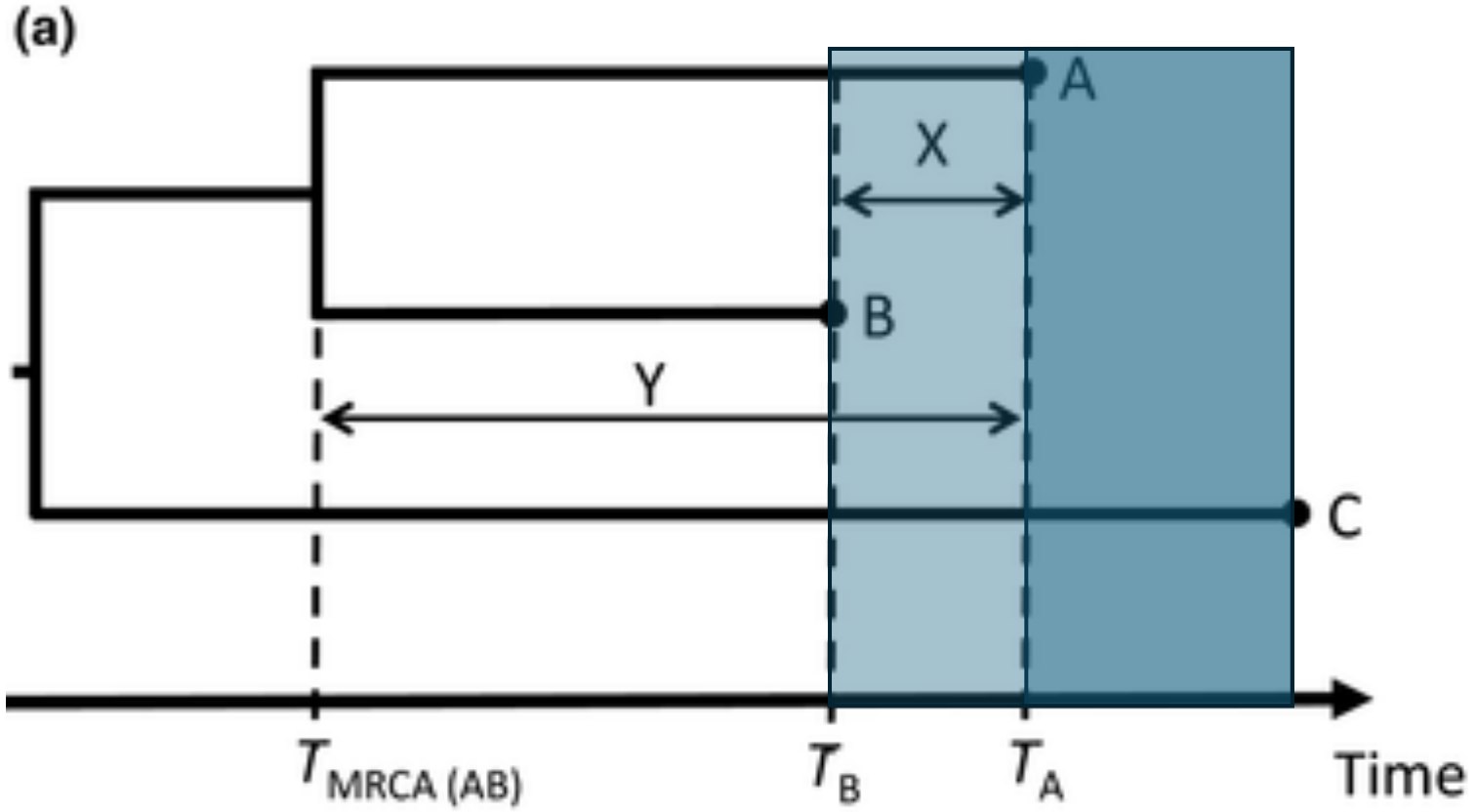
Rieux and Balloux (2016)



A and B are isolated at different points in time

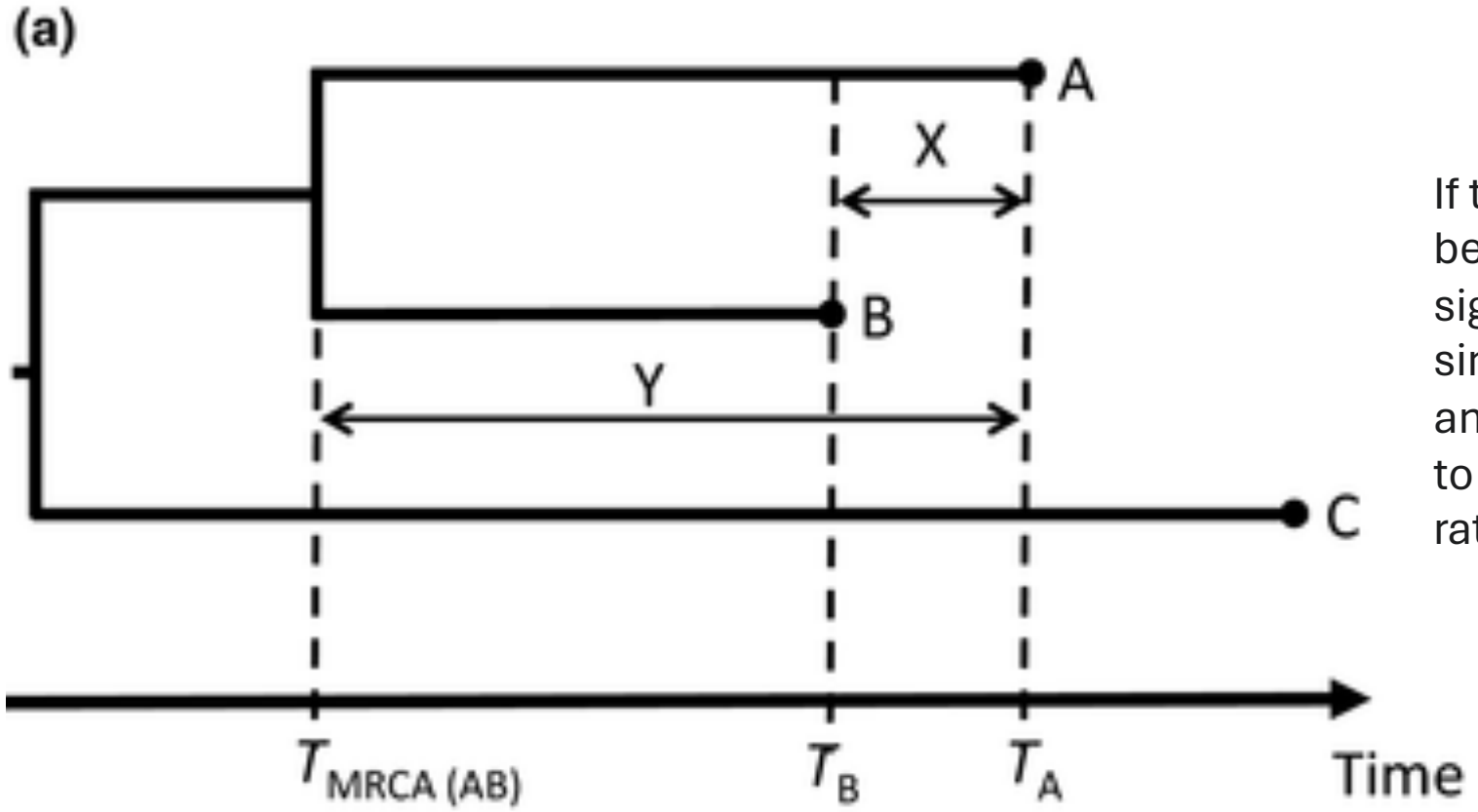
Outgroup

Rieux and Balloux (2016)



If the rate of evolution is the same in lineages A and B, then the amount of molecular evolution expected to occur between T_A and T_B is <u>equal</u> to $d_{AC} - d_{BC}$

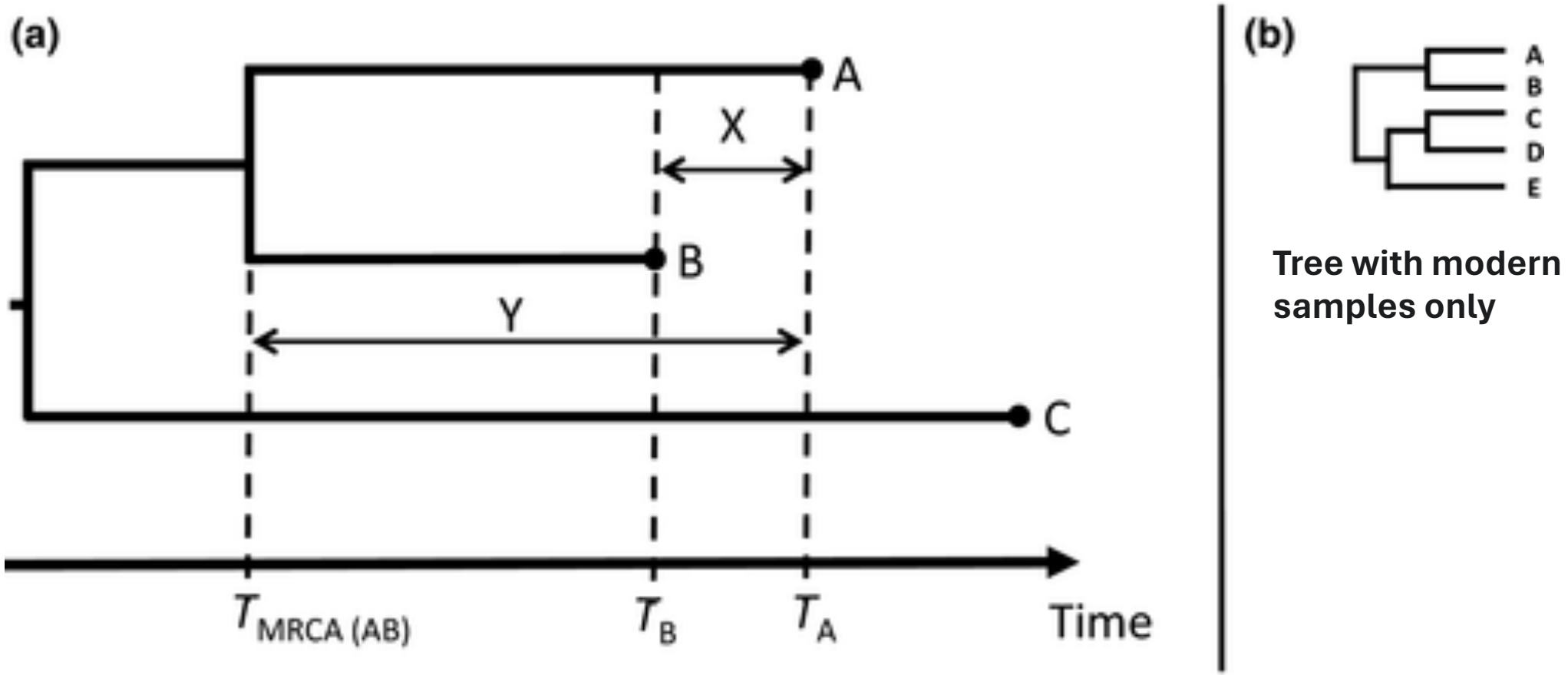
Rieux and Balloux (2016)



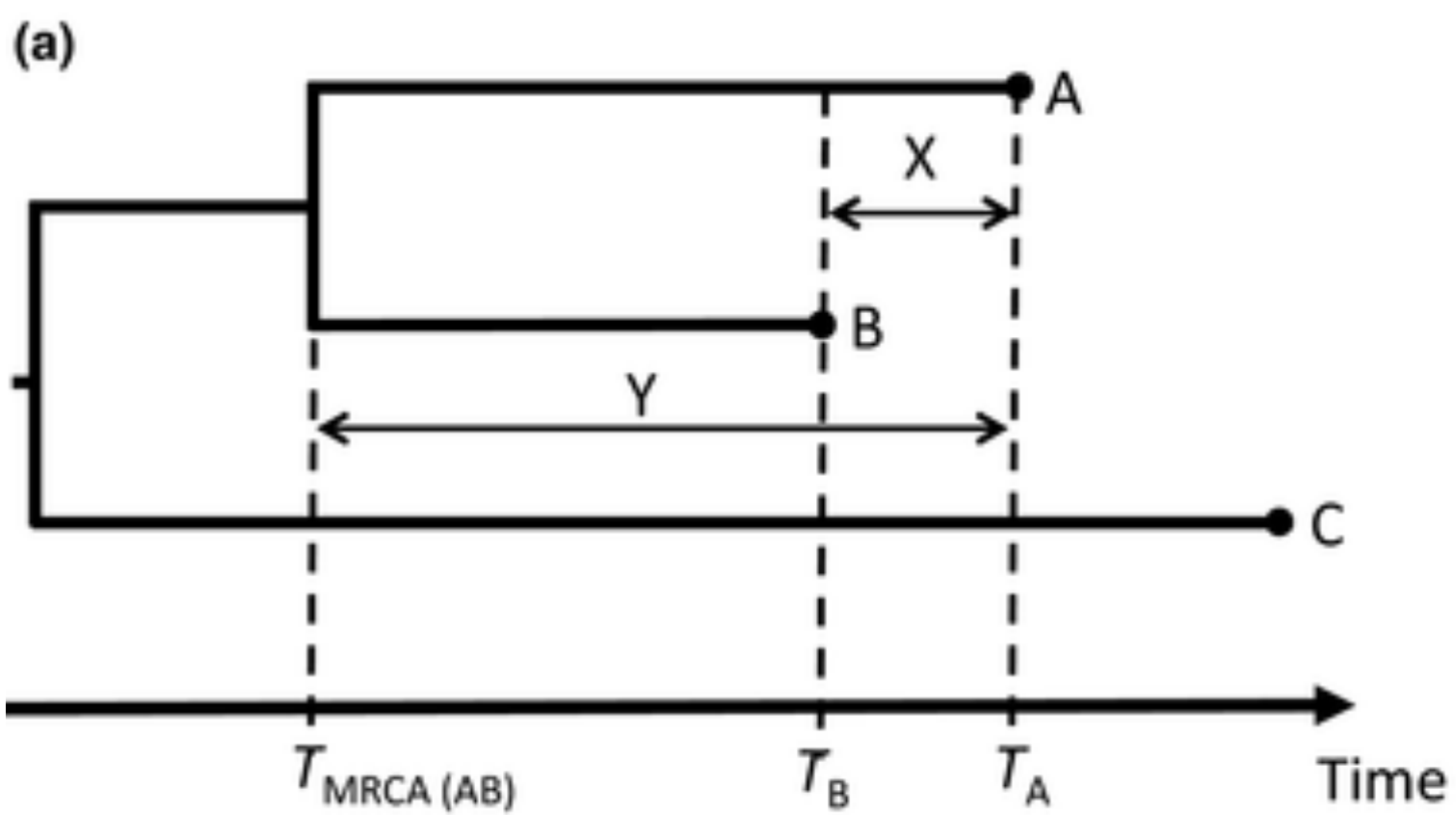
If the time X between T_A and T_B represents a significant proportion of the time Y since A and B last shared a common ancestor, then one can use tip dates to estimate conjointly estimate the rate of evolution.

$$\mu = (AC - BC)/(T_A - T_B)$$

Rieux and Balloux (2016)

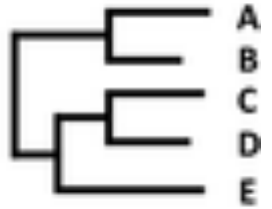


Rieux and Balloux (2016)

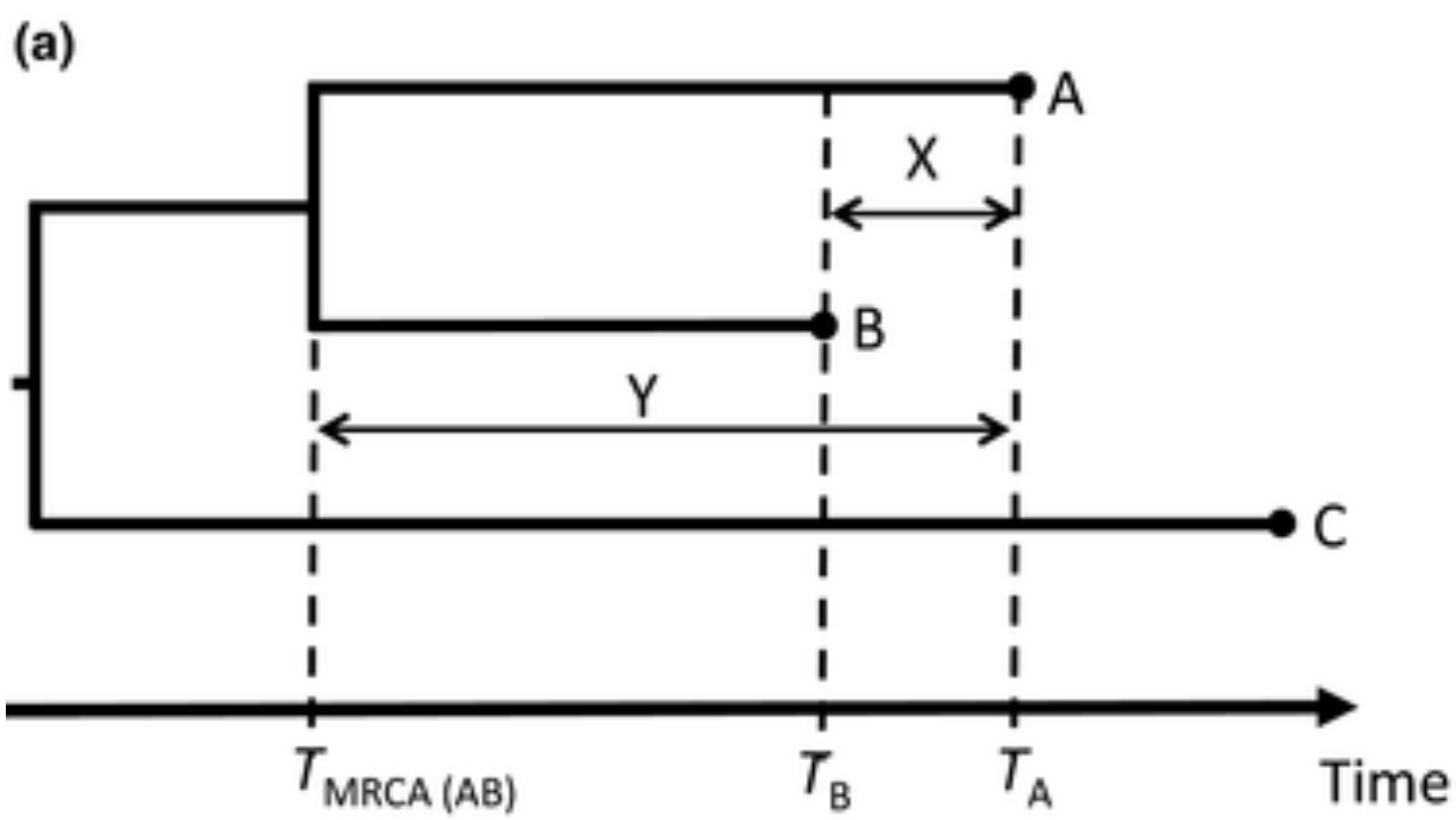


(b)

Tree where tip dates are not widely spread enough for accurate inferences

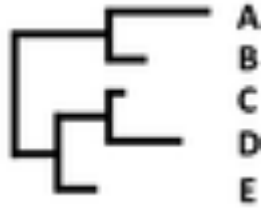


Rieux and Balloux (2016)



(b)

Tree where tip date width is broad enough to allow divergence time and evolutionary estimates with good certainty



Rieux and Balloux (2016)

Important note

Calibration by tip-dating can only be used if there is sufficient temporal signal in your dataset.

No temporal signal = you can't calibrate the molecular clock = no confidence in your result

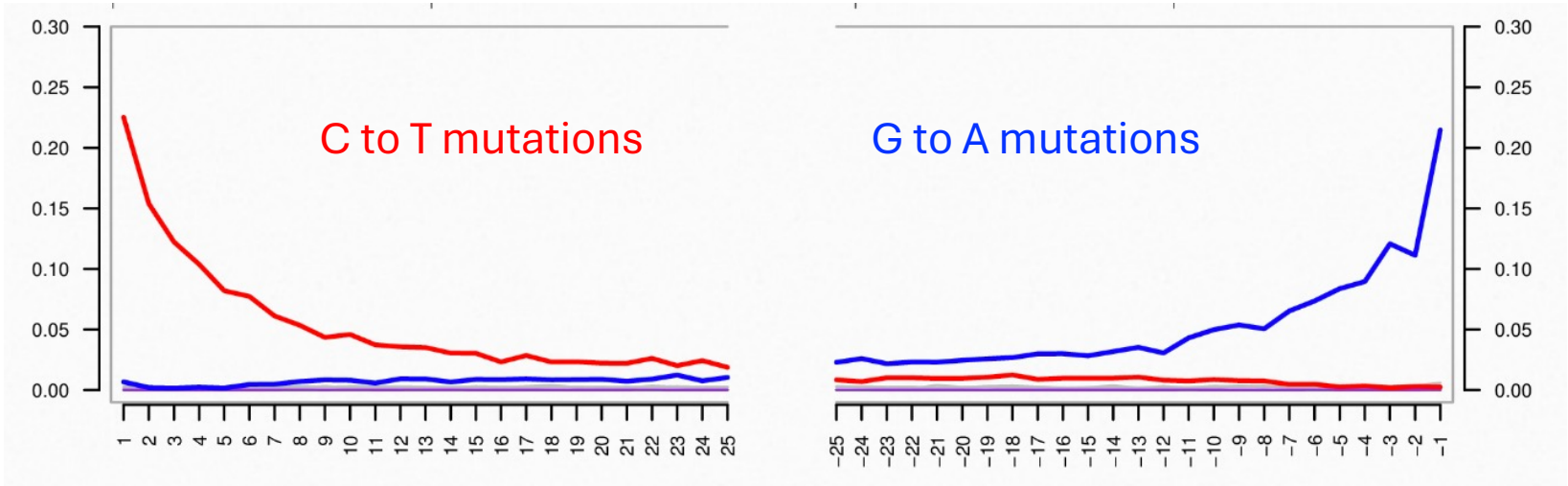
"Garbage in, garbage out"

Temporal signal will be absent if the sampling period too short, if you have too few ancient timepoints, or if evolutionary rates vary a lot between lineages

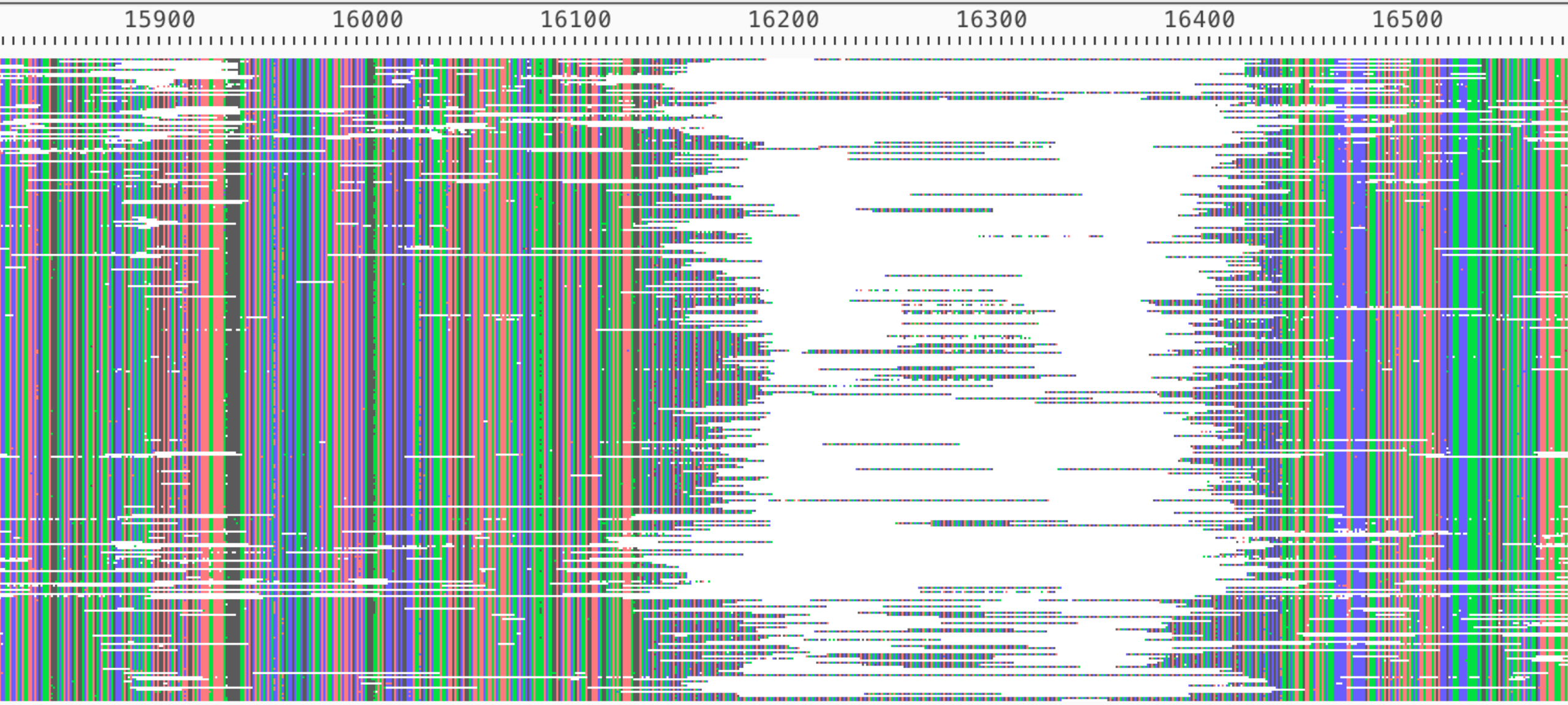
Many methods to check for temporal signal in your dataset

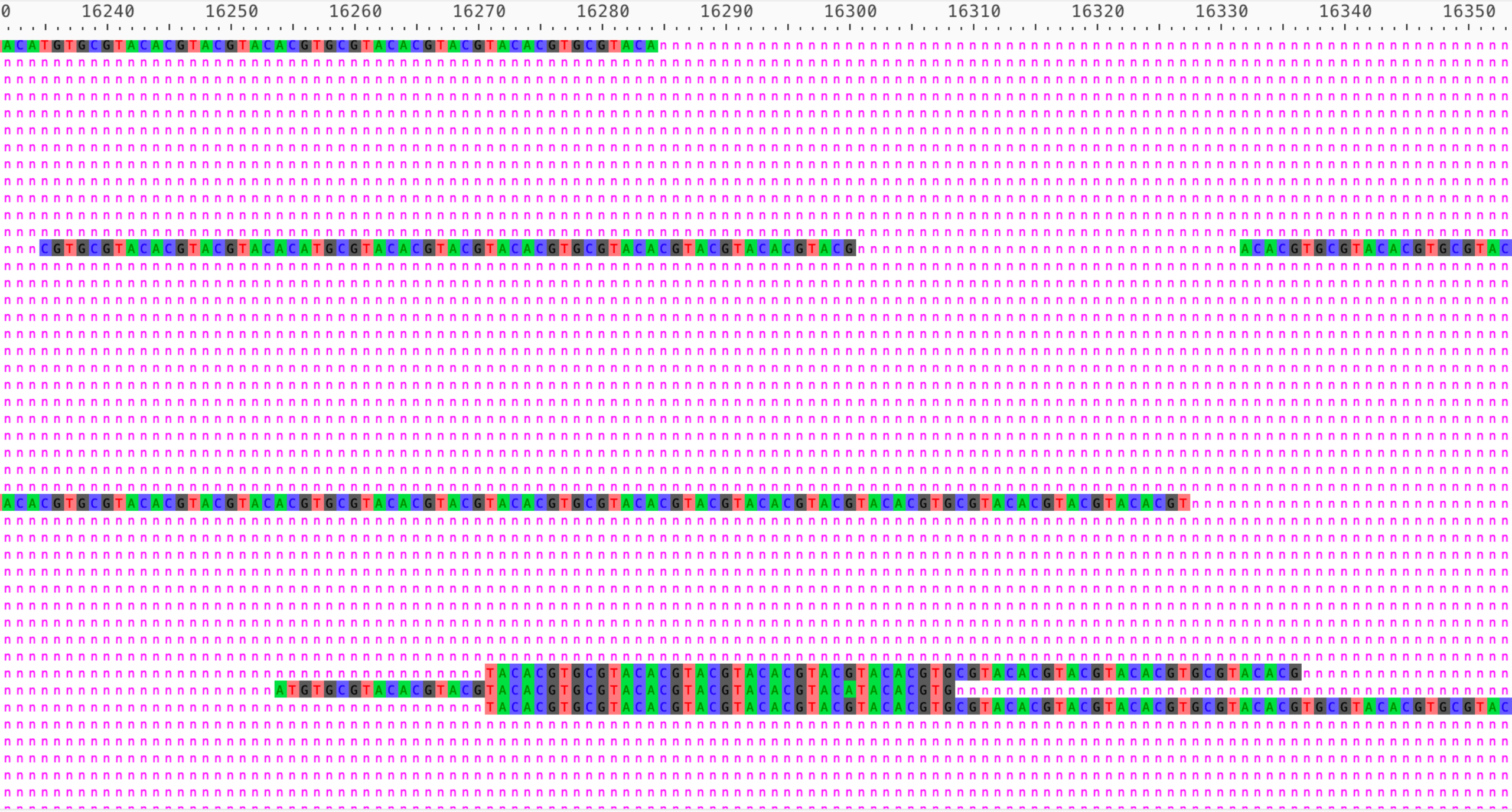
- Date randomization tests (e.g. Ramsden et al. 2008, *MBE*; Duchene et al. 2015, *MBE*; Murray et al. 2016, *Methods Ecol. Evol.*)
- Model selection/comparison (e.g. Murray et al. 2016, Methods Ecol. Evol.)
- BETS (Bayesian evaluation of temporal signal) (Duchene et al. 2020, *MBE*)
- Nested Sampling (NS package in Beast2)

Figure 6.13 RAxML tree of chicken dataset with red jungle fowl as outgroup


C to T mutations
G to A mutations
0.30
0.25
0.20
0.15
0.10
0.05
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
−25 −24 −23 −22 −21 −20 −19 −18 −17 −16 −15 −14 −13 −12 −11 −10 −9 −8 −7 −6 −5 −4 −3 −2 −1

AliView - 1x.fasta




0
16240
16250
16260
16270
16280
16290
16300
16310
16320
16330
16340
16350

Modeling postmortem decay for phylogenies?

Modeling postmortem decay for phylogenies?

- Ho SYW, Heupink TH, Rambaut A, Shapiro B. Bayesian estimation of sequence damage in ancient DNA. Mol Biol Evol. 2007;24:1416–22. doi:10.1093/molbev/msm062.
  - Models postmortem damage as independent of sample age
  - Implemented in BEAST

Modeling postmortem decay for phylogenies?

- Rambaut A, Ho SYW, Drummond AJ, Shapiro B. Accommodating the effect of ancient DNA damage on inferences of demographic histories. Mol Biol Evol. 2009;26:245–8. doi:10.1093/molbev/msn256.
  - Models postmortem damage as dependent on sample age
  - Implemented in BEAST

Modeling postmortem decay for phylogenies?

- Final consensus aDNA sequences typically have very low error rate due to damage

Modeling postmortem decay for phylogenies?

- Final consensus aDNA sequences typically have very low error rate due to damage
- When the two models of damage were directly compared, very low rates of damage found.

Exercises!?