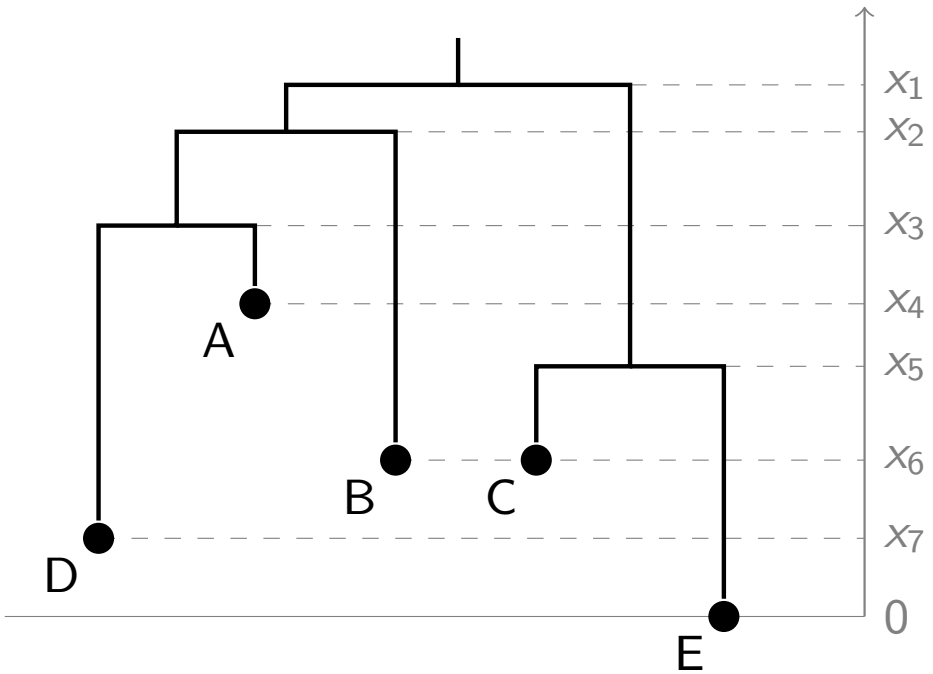
Dating species phylogenies and sampled ancestors

Alexandra Gavryushkina

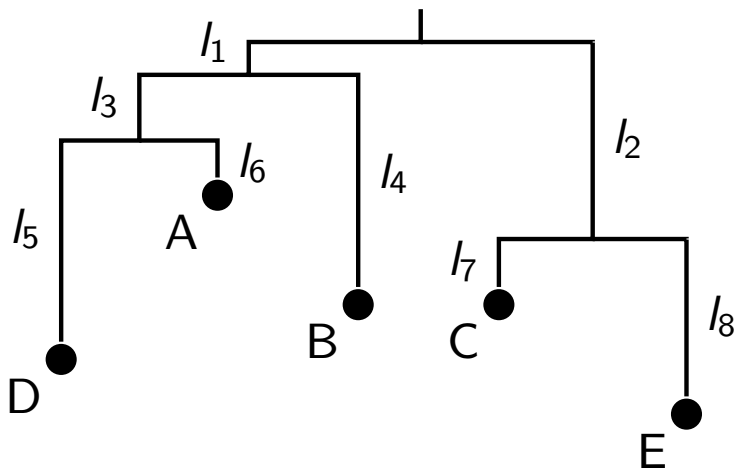
Taming the Beast 2018

Dated phylogenetic tree



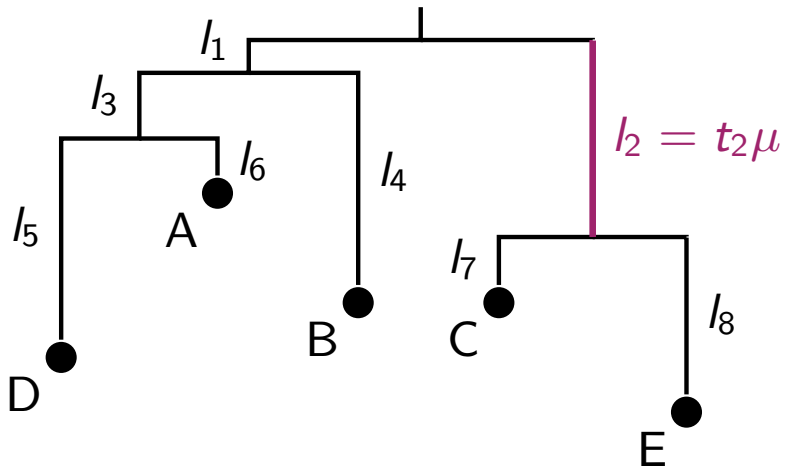
- Each node in a dated phylogenetic tree is assigned a calendar date.

Estimating dated phylogeny (simple case)



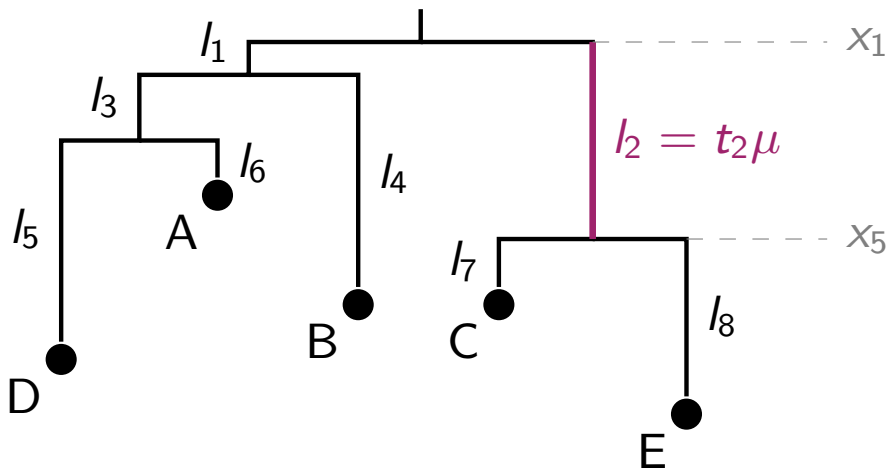
- Having only molecular sequence data we can reconstruct a phylogenetic tree with branch lengths in the units of average number of substitutions per site.

Estimating dated phylogeny (simple case)



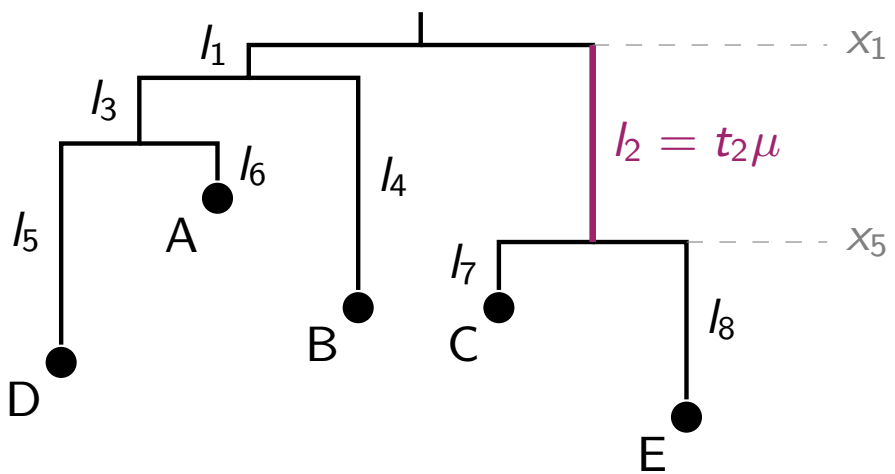
- Having only molecular sequence data we can reconstruct a phylogenetic tree with branch lengths in the units of average number of substitutions per site.
- Each branch length is a product of time t and substitution rate μ.

Estimating dated phylogeny (simple case)



- Having only molecular sequence data we can reconstruct a phylogenetic tree with branch lengths in the units of average number of substitutions per site.
- Each branch length is a product of time t and substitution rate μ.
- If we knew the calendar dates of a few nodes we could reconstruct a dated phylogenetic tree.

Estimating dated phylogeny (simple case)

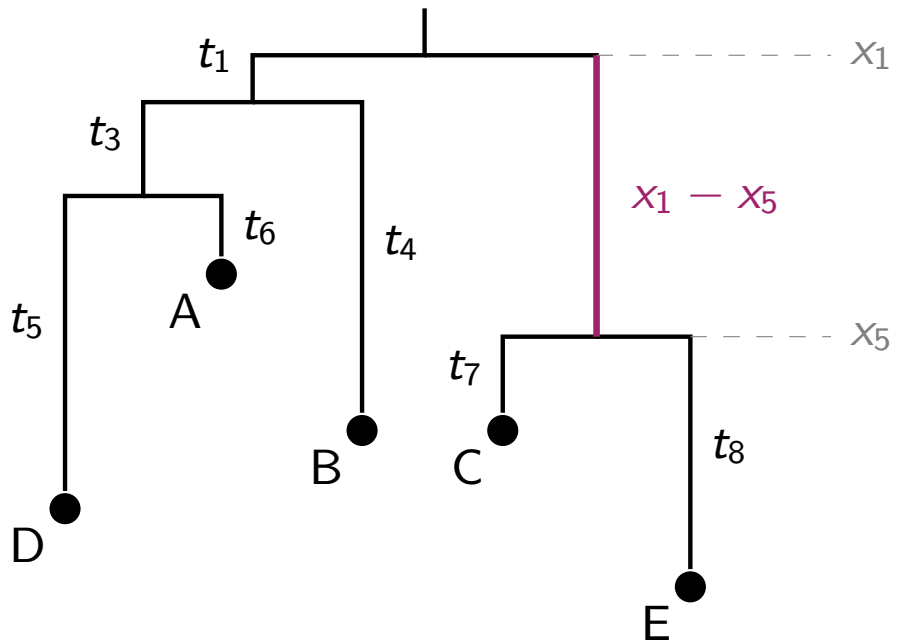


and

$$t_2 = x_1 - x_5$$

$$\mu = \frac{l}{x_1 - x_5}$$

Estimating dated phylogeny (simple case)



$$t_2 = x_1 - x_5$$

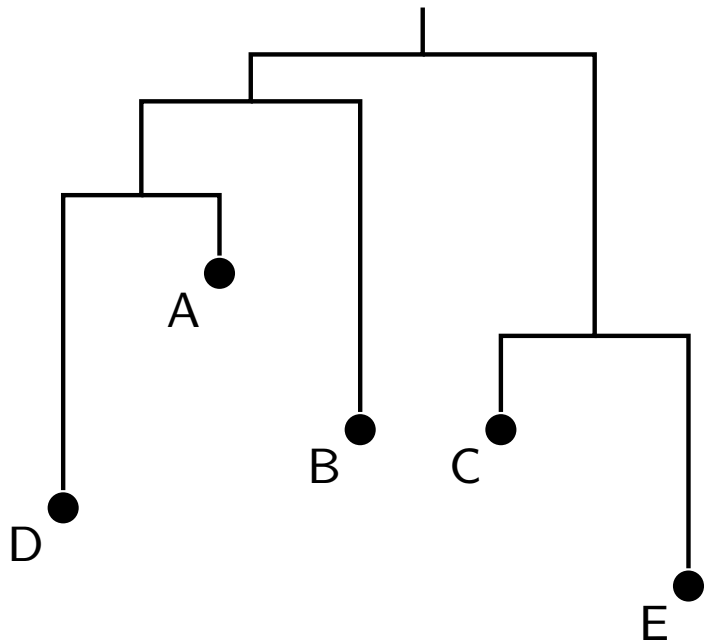
and

$$\mu = \frac{l}{x_1 - x_5}$$

Assuming strict molecular clock:

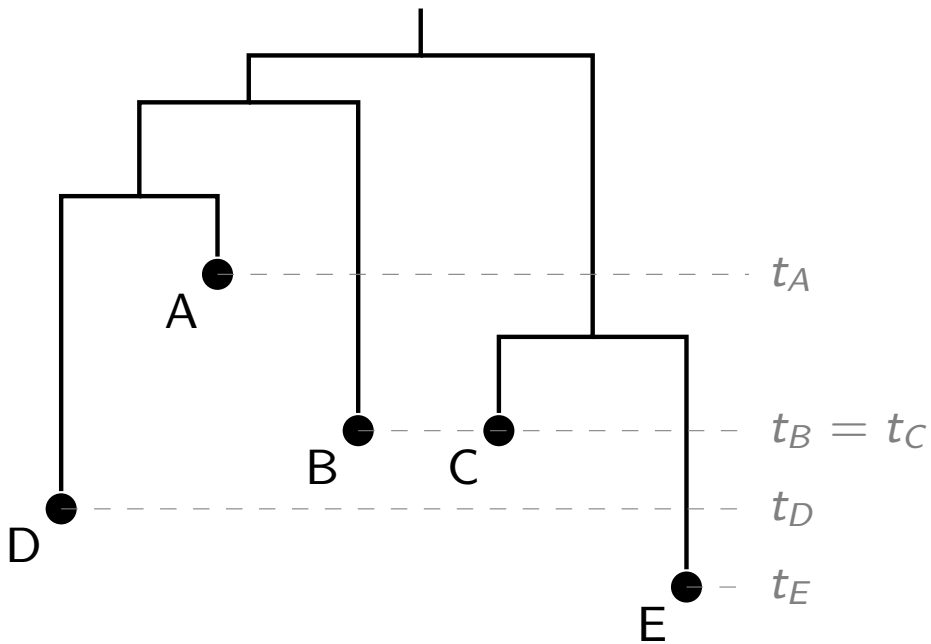
$$t_i = \frac{l_i}{\mu}$$

Estimating dated phylogeny



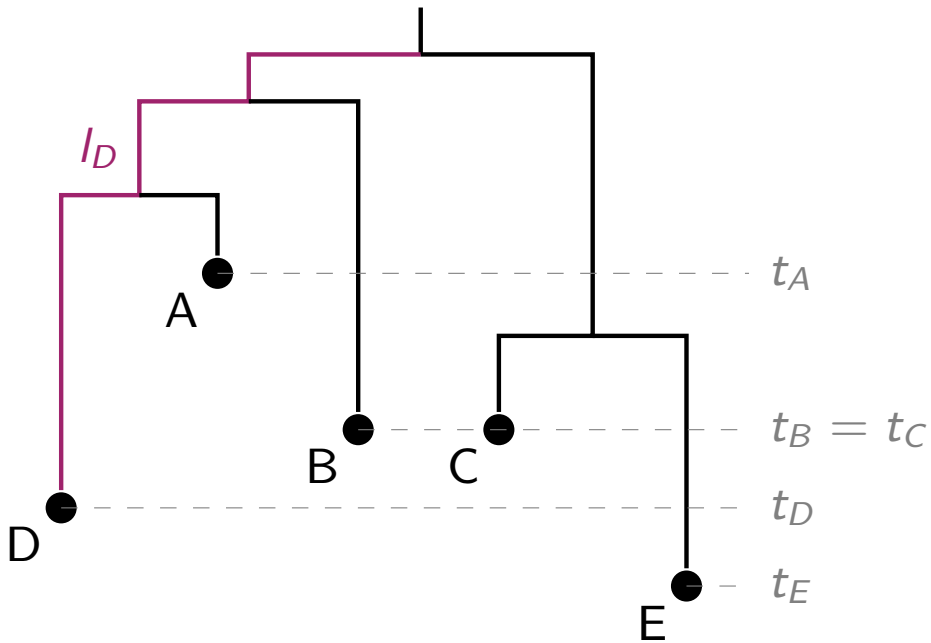
- In measurably evolving organisms the known dates are sampling dates (i.e, tip dates).

Estimating dated phylogeny



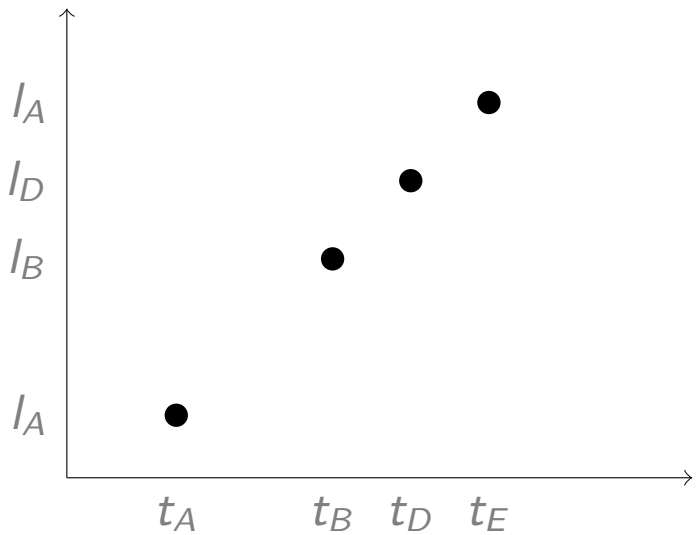
- In measurably evolving organisms the known dates are sampling dates (i.e, tip dates).

Estimating dated phylogeny



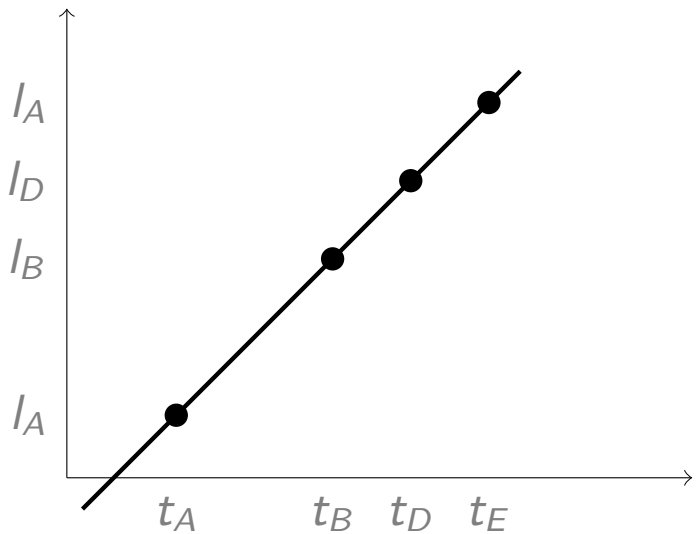
- In measurably evolving organisms the known dates are sampling dates (i.e, tip dates).

Estimating dated phylogeny



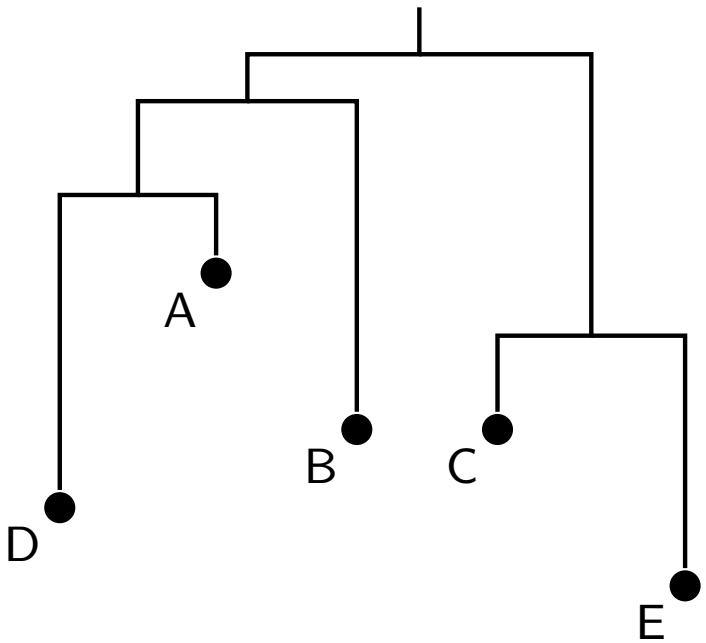
- In measurably evolving organisms the known dates are sampling dates (i.e, tip dates).

Estimating dated phylogeny



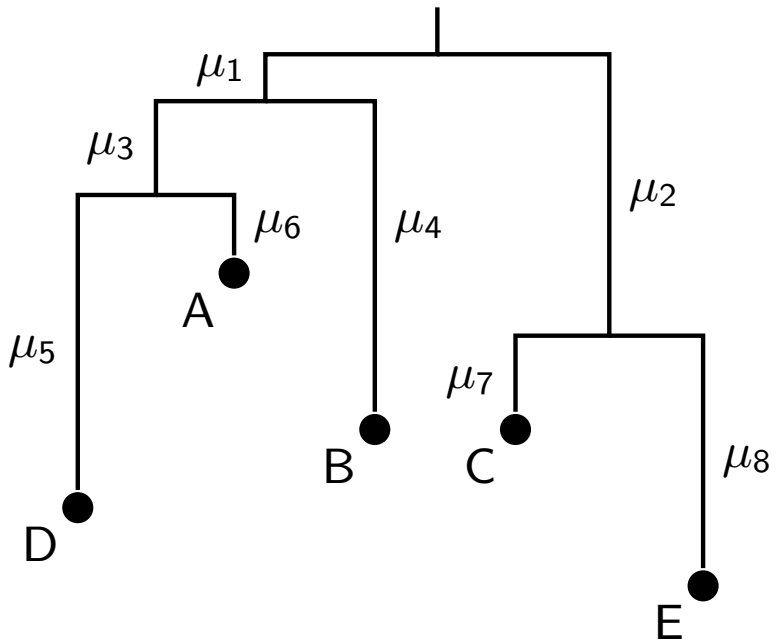
- In measurably evolving organisms the known dates are sampling dates (i.e, tip dates).

Estimating dated phylogeny



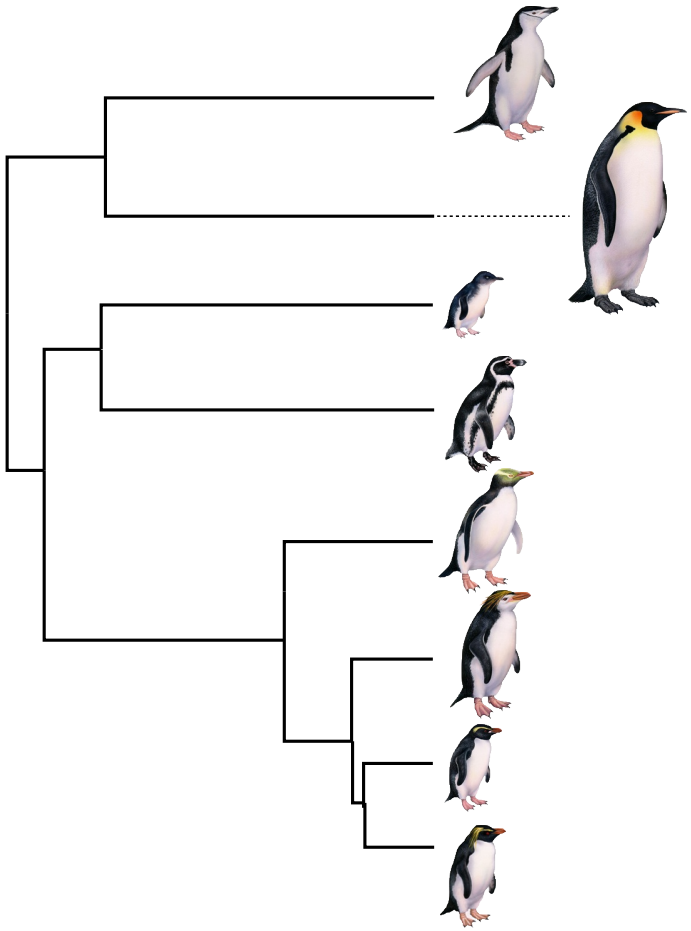
- In measurably evolving organisms the known dates are sampling dates (i.e, tip dates).

Estimating dated phylogeny



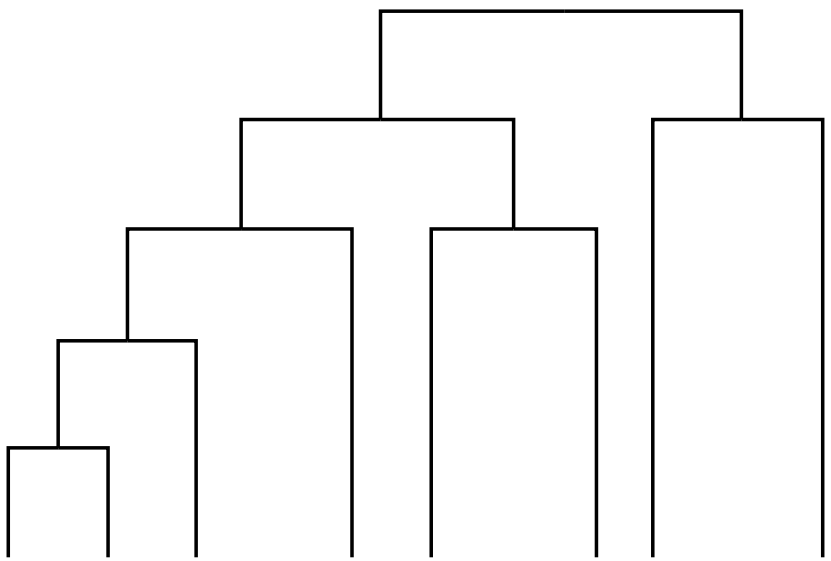
- In measurably evolving organisms the known dates are sampling dates (i.e, tip dates).
- Relaxed molecular clock model assumes:
  - a different rate, μ_i, on each branch and
  - the rates follow some distribution, e.g., $\mu_i \sim LnN(\sigma, \mu)$.

Dating species phylogeny



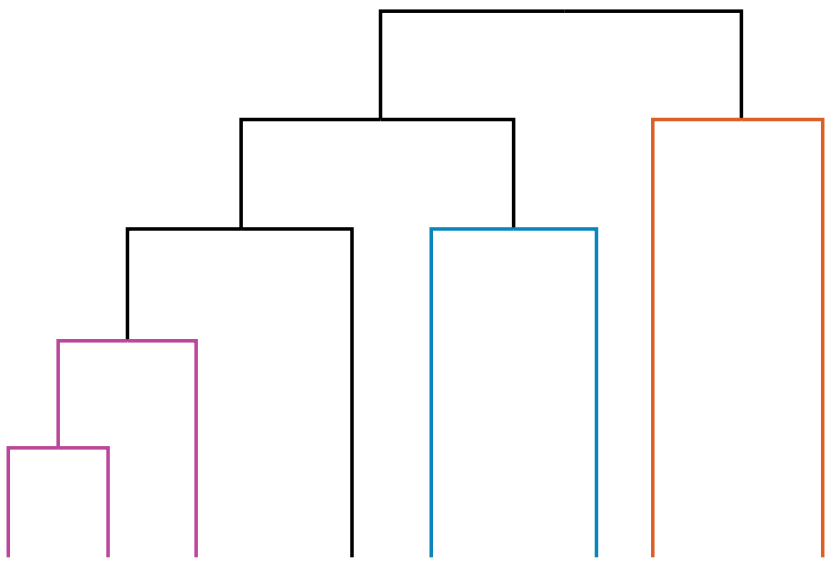
- For macroevolution, the time difference between sampling events is too small.

Calibration approach



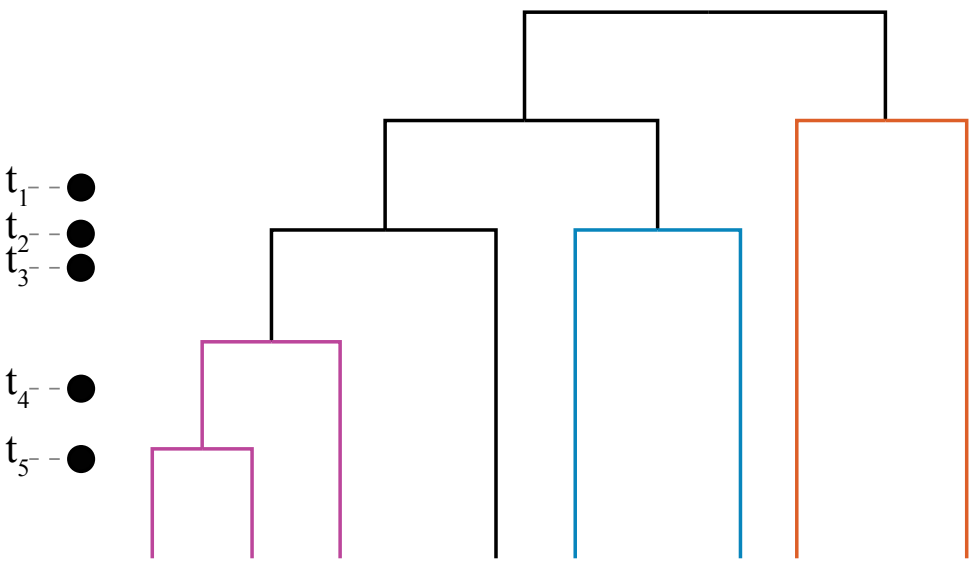
- Estimate undated phylogeny of extant species from molecular data and extract the topology.

Calibration approach



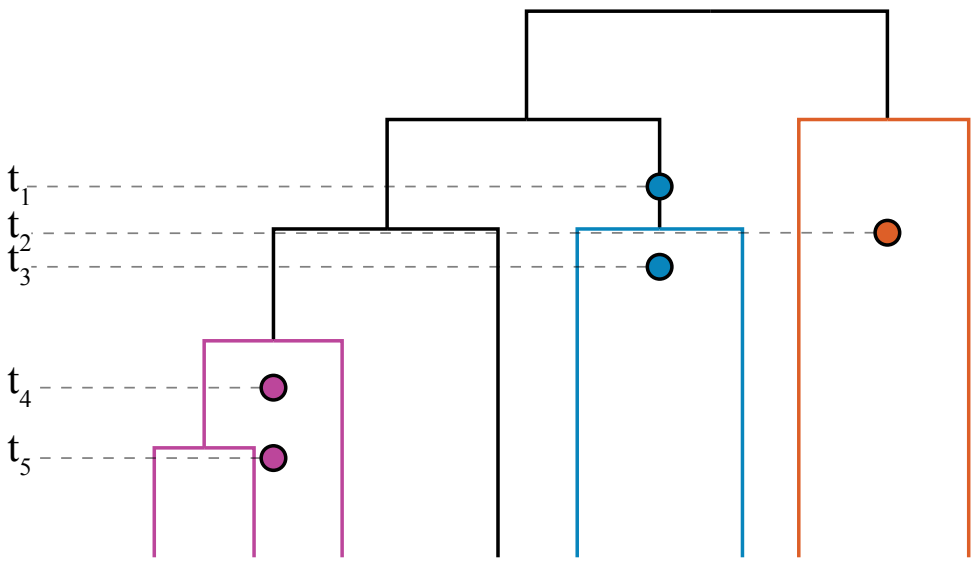
- Estimate undated phylogeny of extant species from molecular data and extract the topology.

Calibration approach



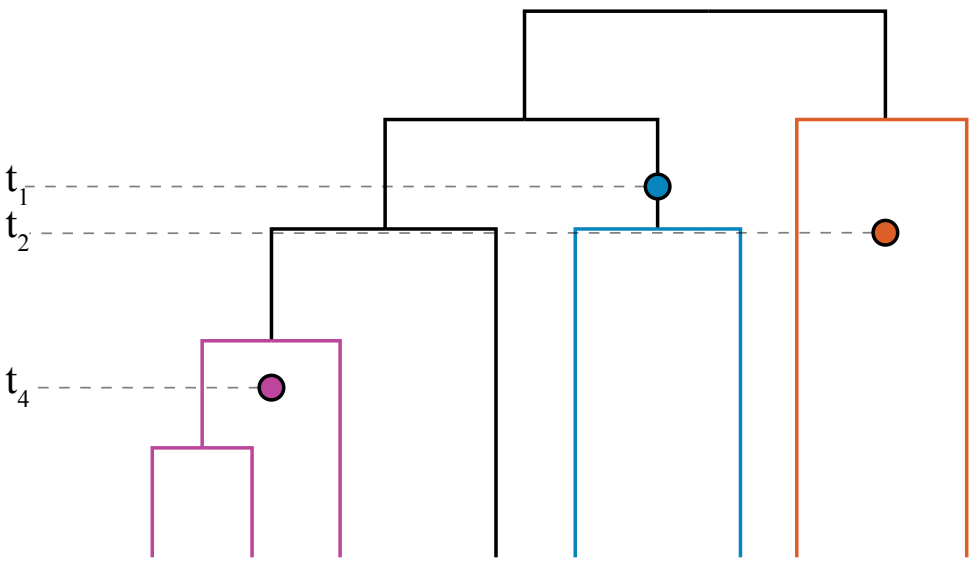
- Estimate undated phylogeny of extant species from molecular data and extract the topology.

Calibration approach



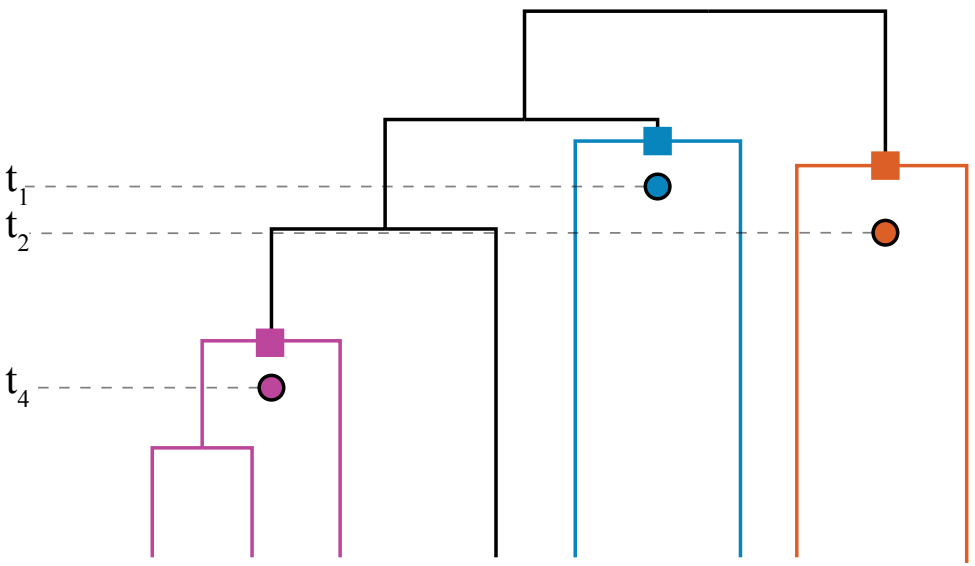
- Estimate undated phylogeny of extant species from molecular data and extract the topology.
- Based on morphological traits assign fossils to clades in the phylogeny.

Calibration approach



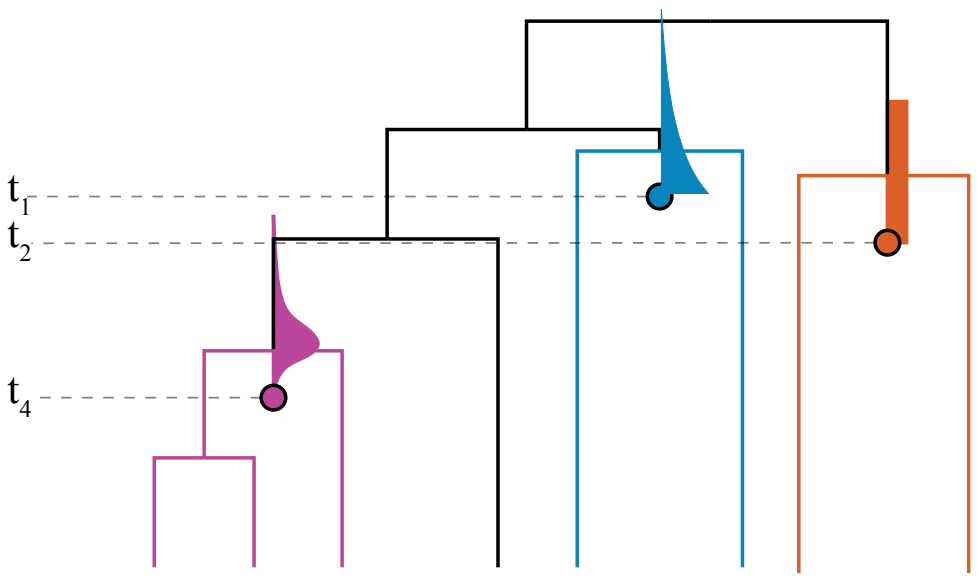
- Estimate undated phylogeny of extant species from molecular data and extract the topology.
- Based on morphological traits assign fossils to clades in the phylogeny.

Calibration approach



- Estimate undated phylogeny of extant species from molecular data and extract the topology.
- Based on morphological traits assign fossils to clades in the phylogeny.

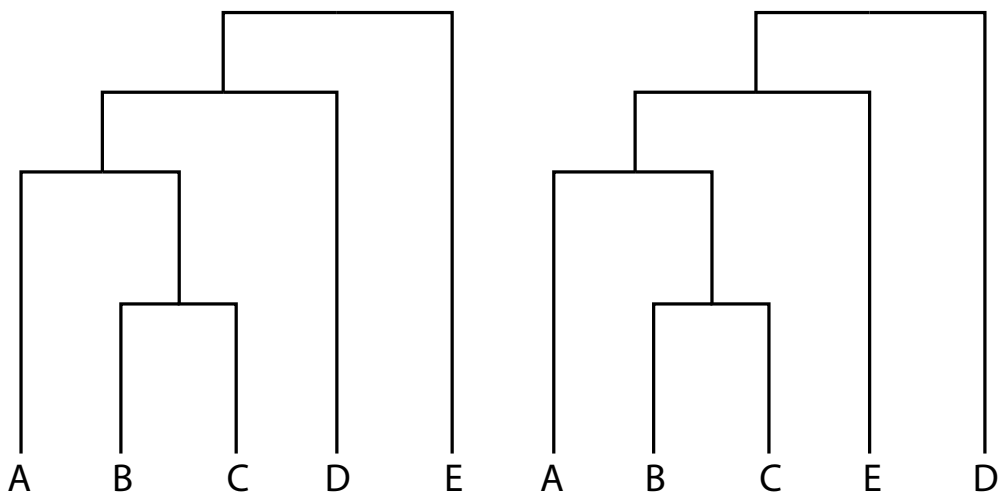
Calibration approach



- Estimate undated phylogeny of extant species from molecular data and extract the topology.
- Based on morphological traits assign fossils to clades in the phylogeny.
- Transform the ages of assigned fossils to calibration densities and use them in a separate analysis of the molecular data to estimate the times of all divergencies.

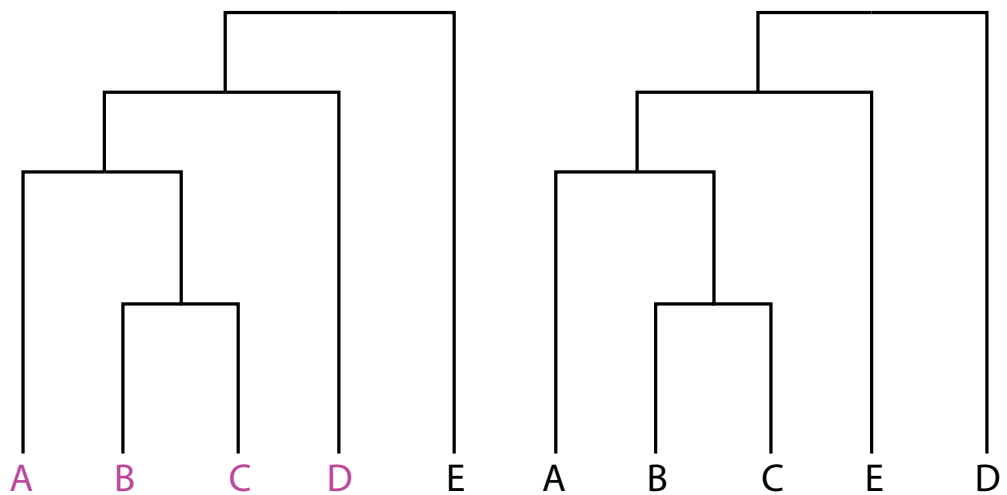
Fixed topology and estimated topology

- It is also possible to co-estimate divergence times and the topology.
- Then either a part of the topology is constraint so that a particular clade is always present
- or the calibration density is applied to the most recent common ancestor of a group of taxa.



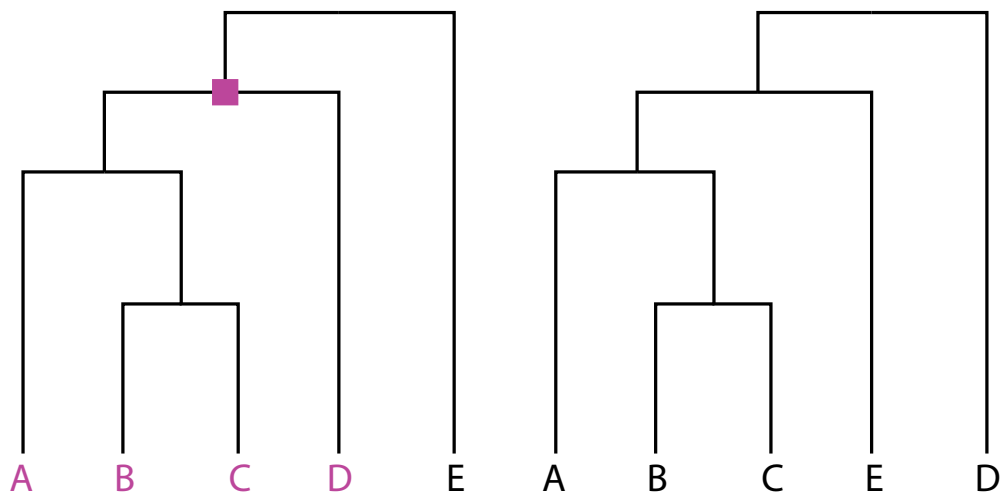
Fixed topology and estimated topology

- It is also possible to co-estimate divergence times and the topology.
- Then either a part of the topology is constraint so that a particular clade is always present
- or the calibration density is applied to the most recent common ancestor of a group of taxa.



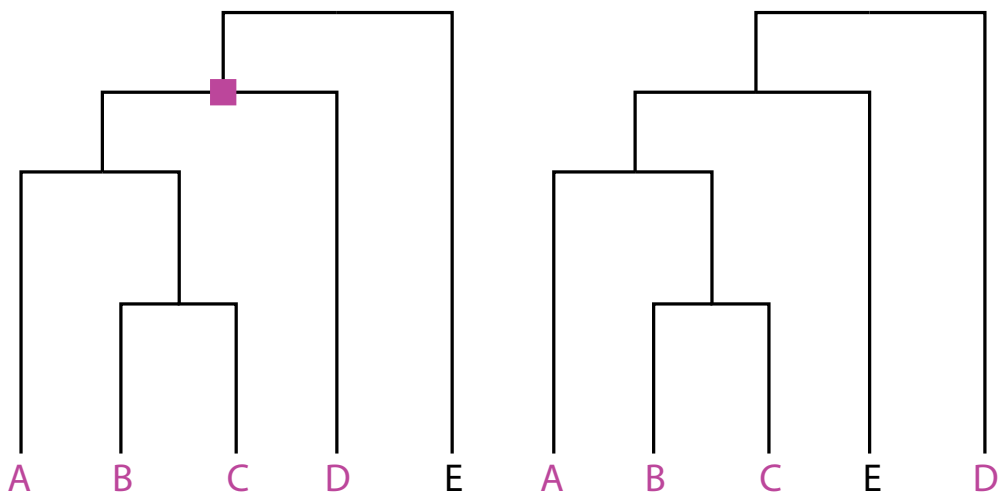
Fixed topology and estimated topology

- It is also possible to co-estimate divergence times and the topology.
- Then either a part of the topology is constraint so that a particular clade is always present
- or the calibration density is applied to the most recent common ancestor of a group of taxa.



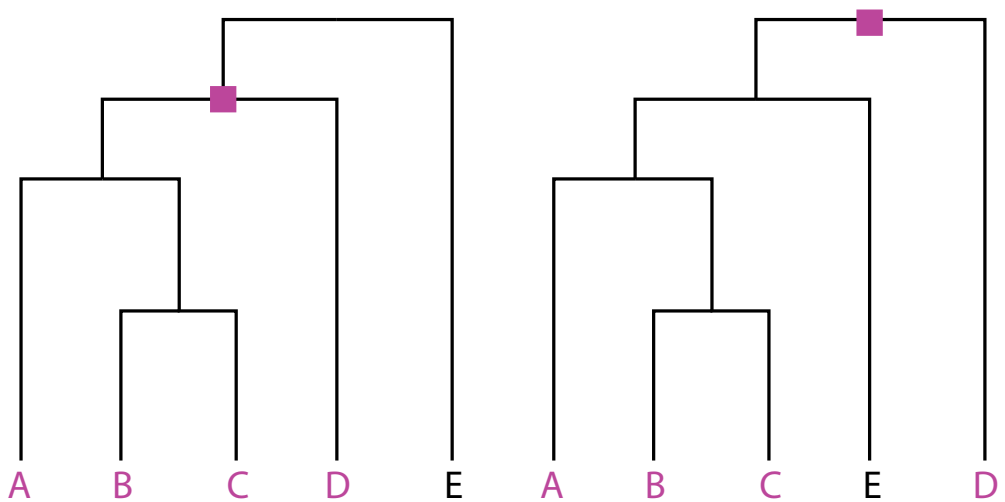
Fixed topology and estimated topology

- It is also possible to co-estimate divergence times and the topology.
- Then either a part of the topology is constraint so that a particular clade is always present
- or the calibration density is applied to the most recent common ancestor of a group of taxa.



Fixed topology and estimated topology

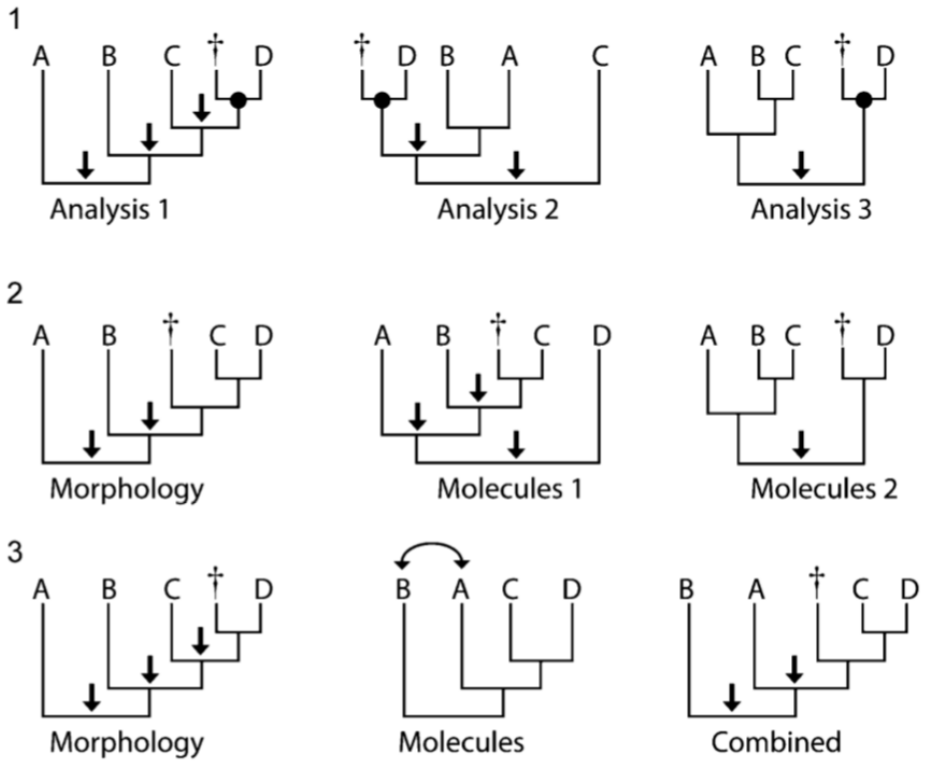
- It is also possible to co-estimate divergence times and the topology.
- Then either a part of the topology is constraint so that a particular clade is always present
- or the calibration density is applied to the most recent common ancestor of a group of taxa.



Problems with calibration approach

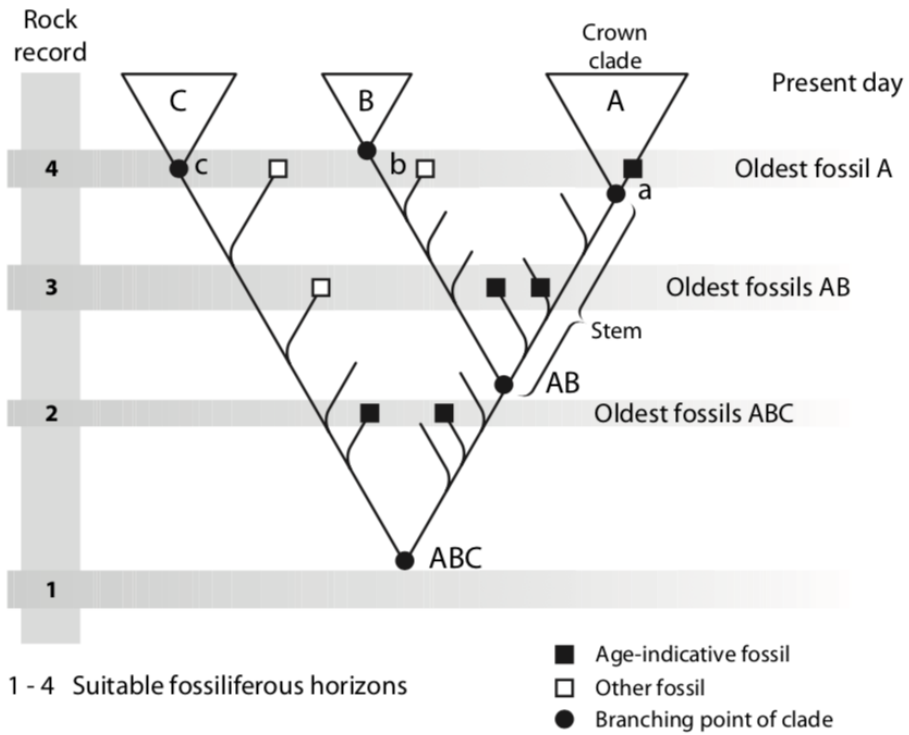
Assigning fossils to clades is problematic

- Fossils are placed on the phylogeny using parsimony methods.
- Sometimes no phylogenetic methods are used but the assignment is based on apomorphies — traits that are unique for a taxon.
- Inference from molecular and morphological data may support different topologies. Then fossil assignment becomes ambiguous.



Parham *et al.* (2012)

Transformation of fossils to calibration densities is *ad hoc*



Benton and Donoghue (2006)

- The oldest fossil is used to specify minimum age.
- Stem fossils and fossils from other clades are used to specify the maximum age (or upper tail of the calibration density).

Many available fossils are ignored

Calibration approach only uses:

- The oldest fossils in each clade. The ages of these fossils are directly used to specify the lower bound of a calibration density.
- A few other fossils are indirectly used to specify the upper tail of a calibration density.

Bayesian inference of divergence times

D	molecular sequence data
T	fixed topology
$t = (t_1, t_2)$	divergence dates
t_1	unknown divergence dates
t_2	ages of calibration nodes
Θ	tree prior hyper-parameters
Ω	molecular evolution model parameters
E	molecular clock model parameters

$$P(t, \Theta, E, \Omega \mid D) = \frac{P(D \mid T, t, \Theta, E, \Omega)P(T, t, \Theta, E, \Omega)}{P(D)} =$$

$$\frac{P(D \mid T, t, E, \Omega)\ P(T, t \mid \Theta)P(t_2)\ P(\Theta)P(E)P(\Omega)}{P(D)}$$

Bayesian inference of divergence times

D	molecular sequence data
T	fixed topology
$t = (t_1, t_2)$	divergence dates
t_1	unknown divergence dates
t_2	ages of calibration nodes
Θ	tree prior hyper-parameters
Ω	molecular evolution model parameters
E	molecular clock model parameters

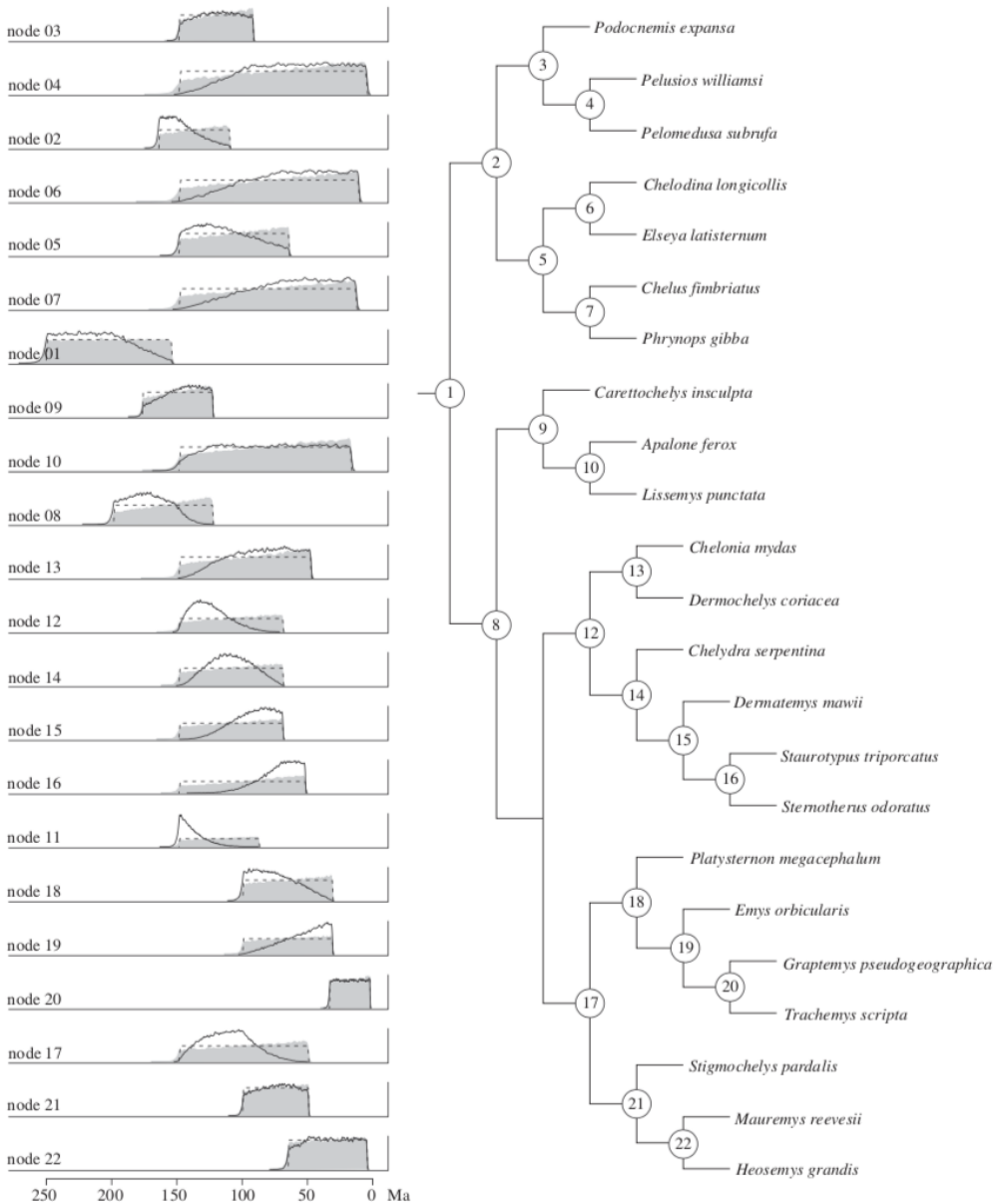
$$P(t, \Theta, E, \Omega \mid D) = \frac{P(D \mid T, t, \Theta, E, \Omega) P(T, t, \Theta, E, \Omega)}{P(D)} =$$

$$\frac{P(D \mid T, t, E, \Omega) \boxed{P(T, t \mid \Theta) P(t_2)} P(\Theta) P(E) P(\Omega)}{P(D)}$$

Problems with the inference

- Several calibration densities interact with each other and with the tree prior leading to altered prior distributions for the ages of calibration nodes.

Warnock *et al.* (2014):



Problems with the inference

$t = (t_1, t_2)$ divergence dates
t_1 unknown divergence dates
t_2 ages of calibration nodes

$$\frac{P(D \mid T, t, E, \Omega)\boxed{P(T, t \mid \Theta)P(t_2)}P(\Theta)P(E)P(\Omega)}{P(D)}$$

$$\frac{P(D \mid T, t, E, \Omega)\boxed{P(T, t_1 \mid t_2, \Theta)P(t_2)}P(\Theta)P(E)P(\Omega)}{P(D)}$$

Problems with the inference

- Correctly incorporating calibration densities into an analysis when topology is co-estimated is challenging.

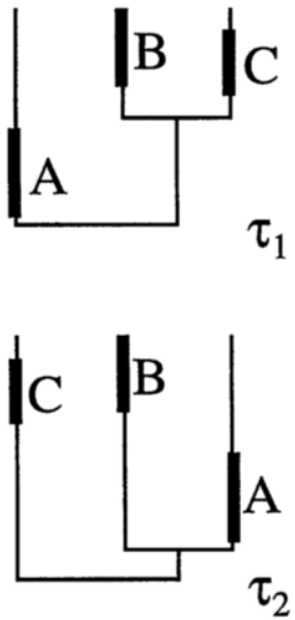
T	topology (or constraint topology)
$t = (t_1, t_2)$	divergence dates
t_1	unknown divergence dates
t_2	ages of calibration nodes

$$P(t, T, \Theta, E, \Omega \mid D) = \frac{P(D \mid T, t, \Theta, E, \Omega)P(T, t, \Theta, E, \Omega)}{P(D)} =$$

$$\frac{P(D \mid T, t, E, \Omega)\boxed{P(T, t_1 \mid t_2, \Theta)P(t_1)}P(\Theta)P(E)P(\Omega)}{P(D)}$$

Incoherency

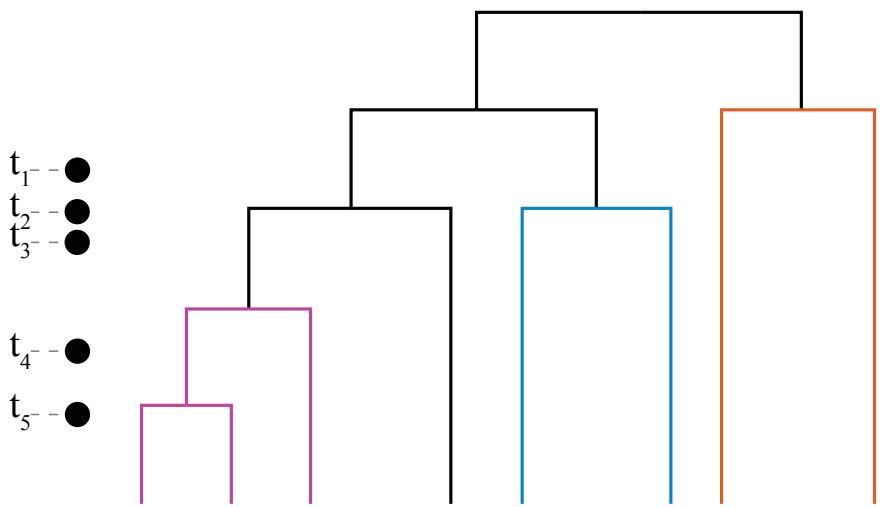
- Fossils come from the same process that generates the molecular phylogeny. However fossil data and molecular data are treated independently.
- The topology and divergence times are often estimated separately but they are influenced by each other.



Huelsenbeck and Rannala (1997)

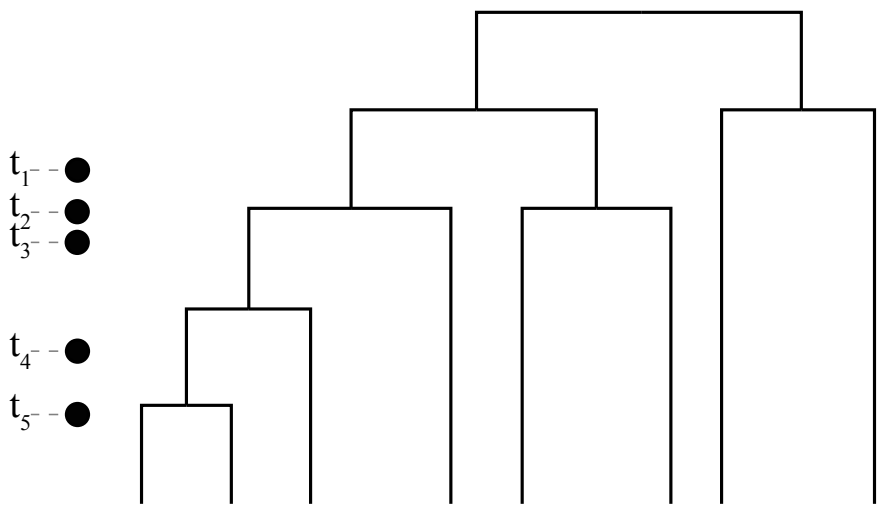
Joint/total-evidence inference

Joint inference



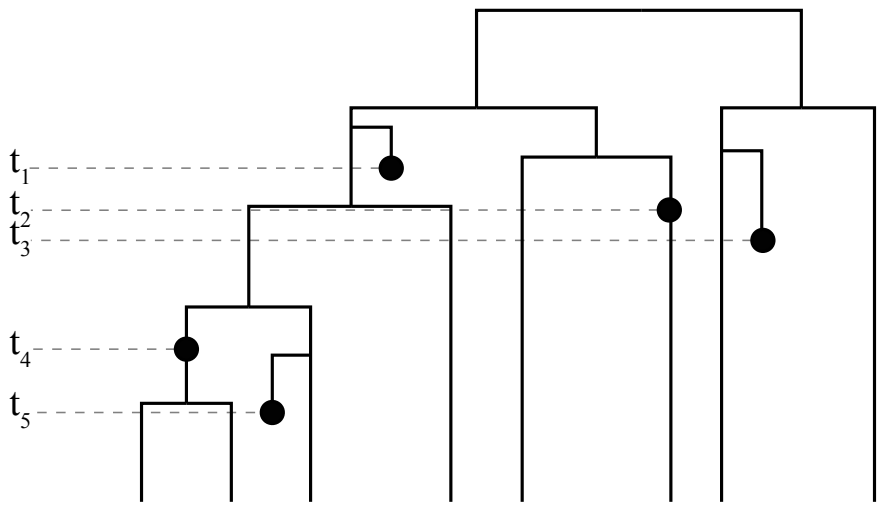
- Extinct species are a part of the phylogeny

Joint inference



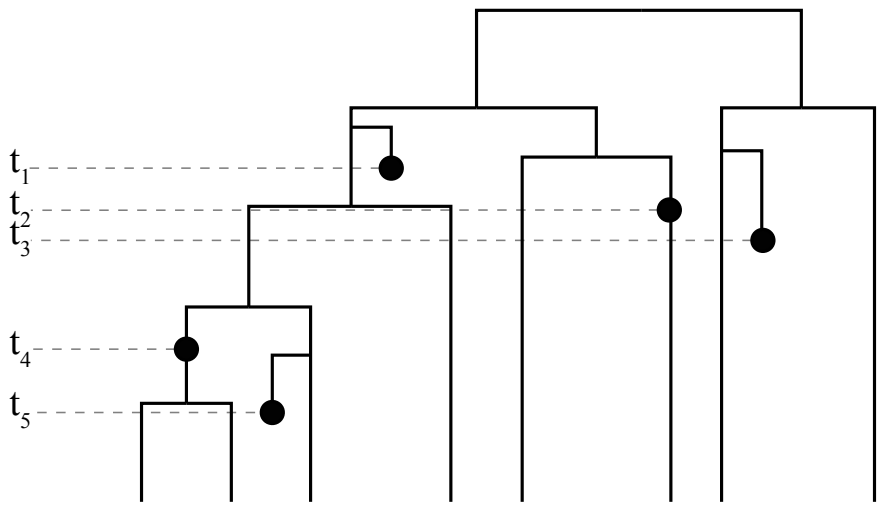
- Extinct species are a part of the phylogeny

Joint inference



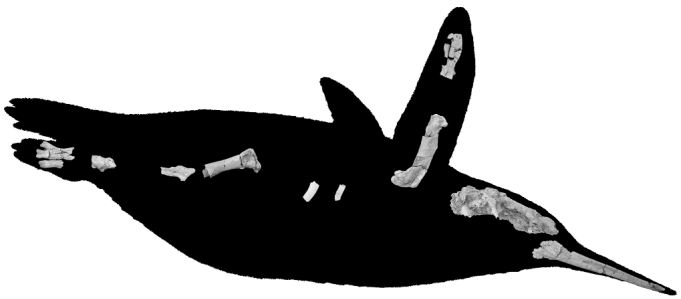
- Extinct species are a part of the phylogeny
- Fossil ages can be used to estimate dated phylogeny in a similar way as sampling dates of measurably evolving organisms.

Joint inference



- Extinct species are a part of the phylogeny
- Fossil ages can be used to estimate dated phylogeny in a similar way as sampling dates of measurably evolving organisms.
- Molecular data of fossils is not available (except for rear ancient DNA).

Morphological data

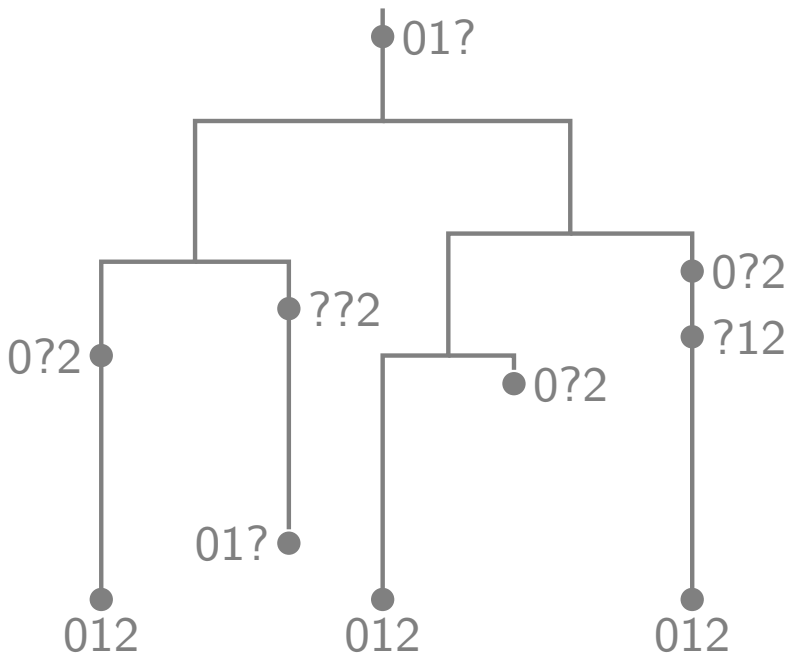


002113???3?0014?70210?...

Ksepka and Clarke (2010)

- Extant species morphology can help in estimating dated phylogenetic trees.
- We treat traits as discrete characters.

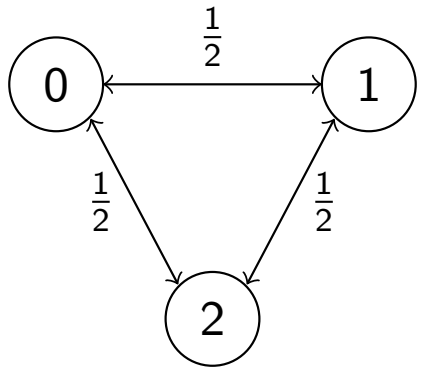
Morphological data



- We consider phylogenetic trees that relate both: extant and fossil species.
- Sequences of discrete characters of fossil and extant species are treated in a similar way as molecular sequences.

Lewis Mk model

Continuous-time Markov chain process with **k** states: $0, 1, 2, \ldots, k-1$.



The instantaneous transition rates from any state to any other state are equal:

$$Q_k = \frac{1}{(k-1)} \begin{bmatrix} -(k-1) & 1 & 1 & \ldots & 1 \\ 1 & -(k-1) & 1 & \ldots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \ldots & -(k-1) \end{bmatrix}$$

The frequencies of the characters at equilibrium are all equal:

$$[\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k}]$$

Characters with different numbers of states

0000010010110100 02
000041011001011002
000120100110020002
100030?00110000001
??1?????101 1?0110
????1????1?00?0110

- Characters are partitioned in groups with the same number of states (greater than one).

Characters with different numbers of states

0000010010110100 02
0000410110010110 02
0001201001100200 02
1000307?0011000000 1
??1??????1011?0110
???1????1?00?0110

- Characters are partitioned in groups with the same number of states (greater than one).

Characters with different numbers of states

0000010010110100 02
0000410110010110 02
0001201001100200 02
1000303?0011000000 1
??1??????1011?0110
??????1????1?00?0110

- Characters are partitioned in groups with the same number of states (greater than one).

Characters with different numbers of states

```
0000010010110100 02
0000410110010110 02
0001201001100200 02
1000300?0011000 0001
??1??????1011?0110
????1????1?00?0110
```

- Characters are partitioned in groups with the same number of states (greater than one).

Characters with different numbers of states

```
0000010010110100002
0000410110010110002
0001201001100200002
1000300?00110000001
??1??????1011?0110
????1????1?00?0110
```

- Characters are partitioned in groups with the same number of states (greater than one).

Characters with different numbers of states

	2 states	3 states	5 states
0000010010110100002	00010010110000	12	0
0000410110010110002	00010110010100	12	4
0001201001100200002	00101001100000	22	2
100030?00110000001	1000?001100000	01	3
??1??????1011?0110	?1?????1011011	?0	?
????1????1?00?0110	???????1?00011	?0	1

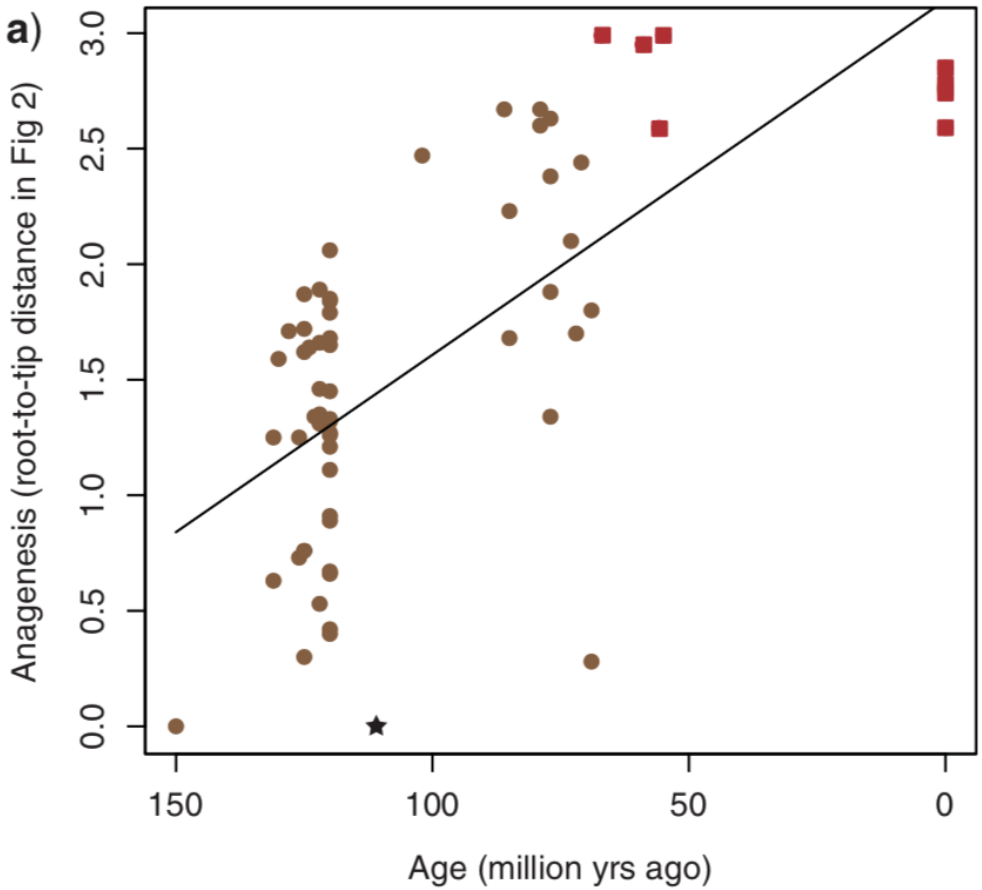
- Characters are partitioned in groups with the same number of states (greater than one).
- Models with different k are used for each group.

Characters with different numbers of states

2 states	3 states	5 states	constant
0000010010110100 02	00010010110000	12	0	0
0000410110010110 02	00010110010100	12	4	0
0001201001100200 02	00101001100000	22	2	0
100030?00110000 01	1000?001100000	01	3	0
??1??????1011?0110	?1?????1011011	?0	?	?
????1????1?00?0110	???????1?00011	?0	1	?

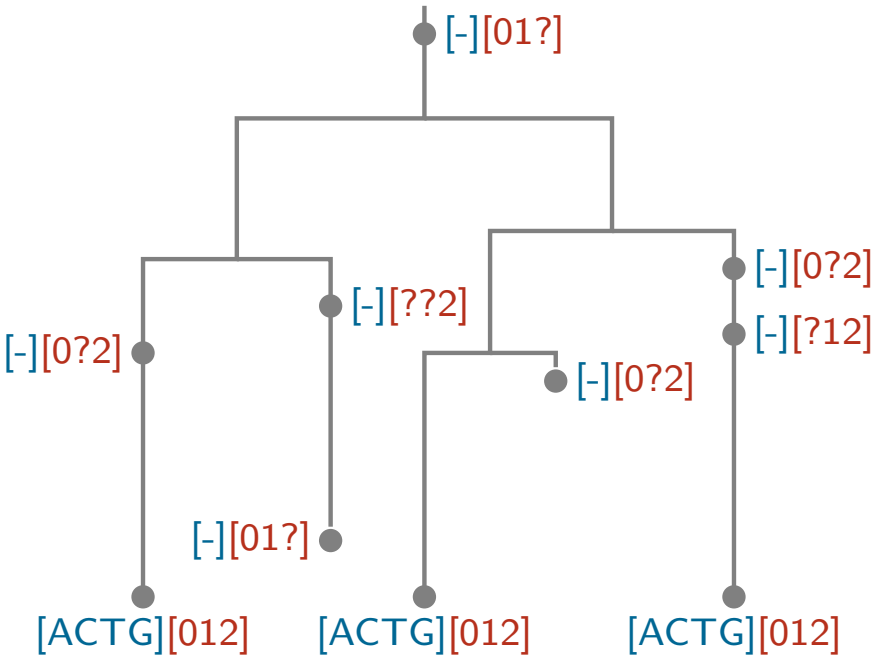
- Characters are partitioned in groups with the same number of states (greater than one).
- Models with different k are used for each group.
- If constant characters are removed from the matrix then Mkv model should be used.

Morphological clock



Lee *et al.* (2014)

Total-evidence analysis of molecular and morphological data



D = molecular data M=morphological data

Other settings for morphological evolution models

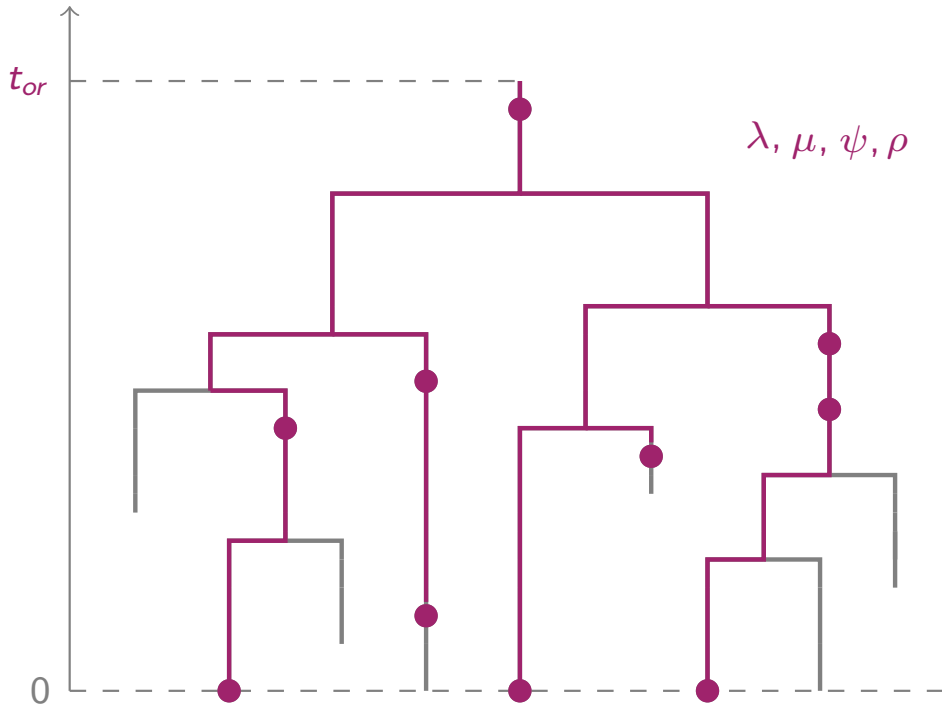
Similarly to modelling evolution of molecular data we can model:

- different rates at different branches: relaxed clock model
- or varying rates for different characters: Gamma variation.

But!

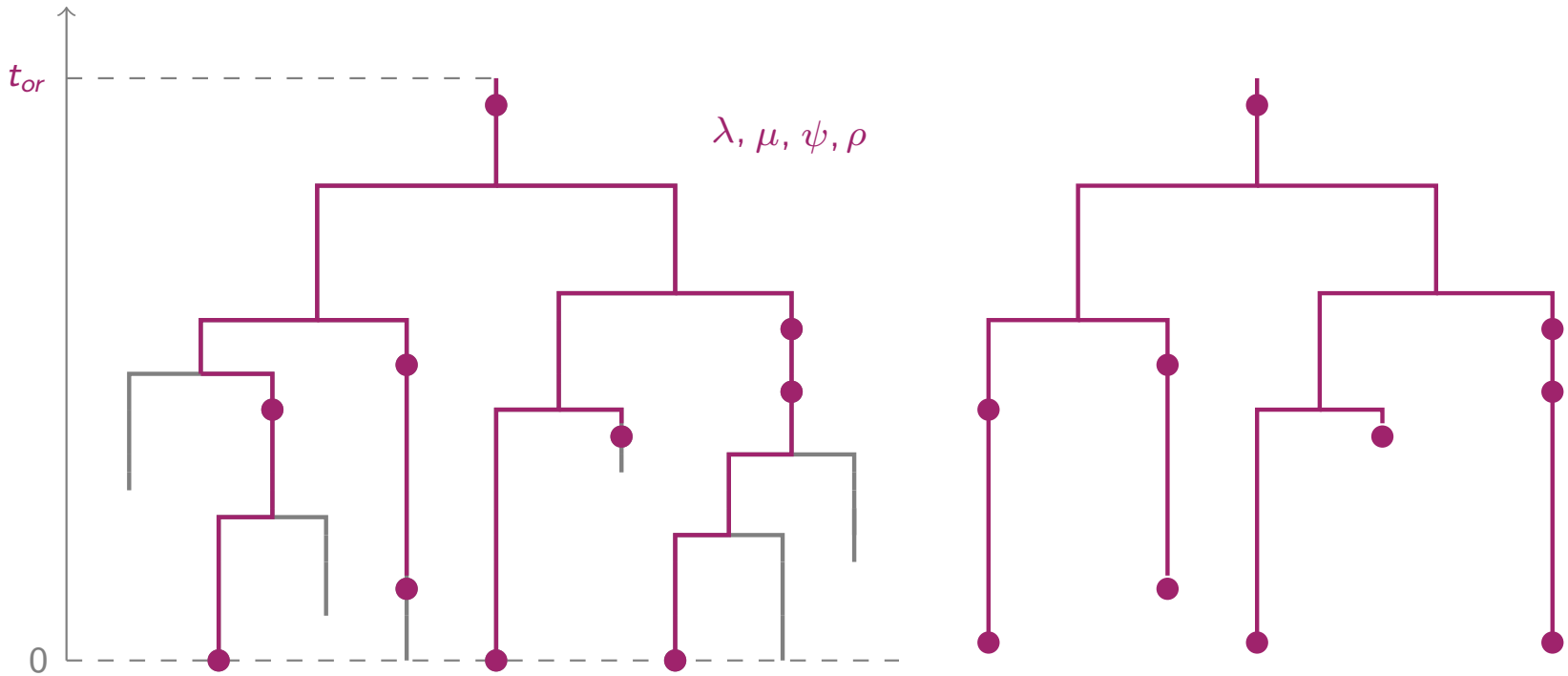
- While heterogeneously sampled molecular data can be enough to estimate a dated phylogeny, morphological data is too sparse.
- We should employ fossil temporal data, which is also informative about the dated phylogeny.

Speciation fossilisation model



Complete phylogeny

Speciation fossilisation model



Complete phylogeny

Sampled phylogeny ($\mathcal{T}$)

Fossilised birth-death model (FBD)

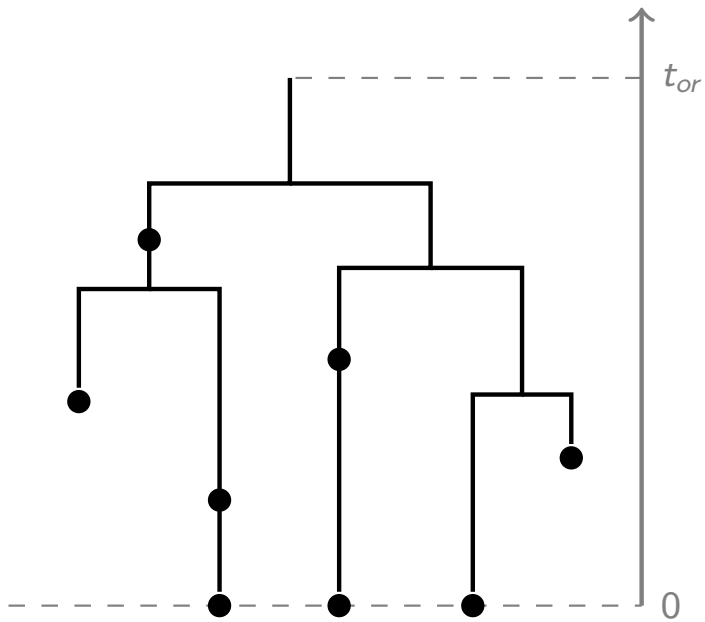
Stadler 2010, Heath *et al.* 2014.

The process starts at time $t_{or} > 0$ and ends at time zero (present time).

- birth rate λ
- death rate μ
- sampling rate ψ
- sampling at present probability ρ

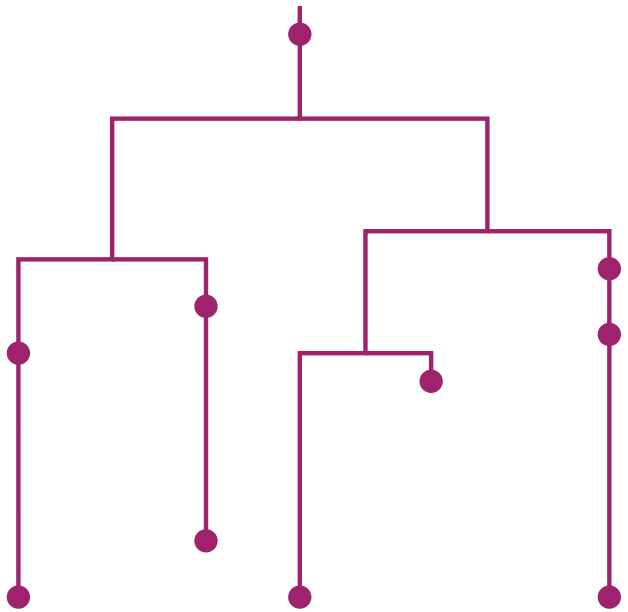
Model parameters: $\eta = (t_{or}, \lambda, \mu, \psi, \rho)$.

All the parameters are identifiable.



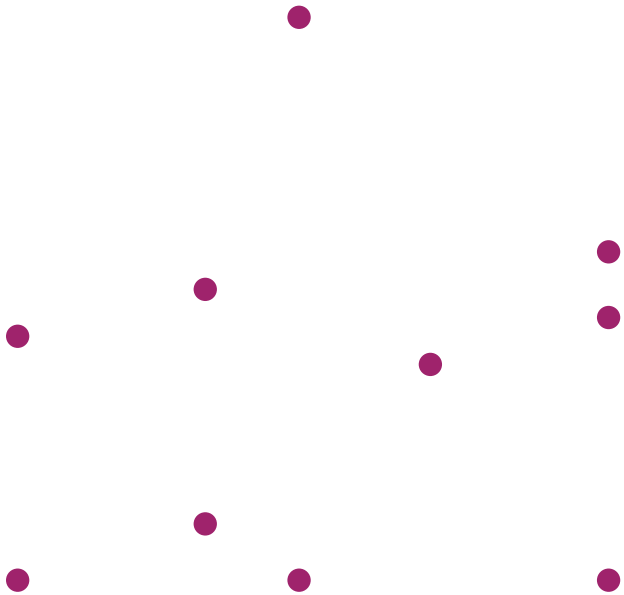
Sampled tree

Data generated by speciation fossilisation model

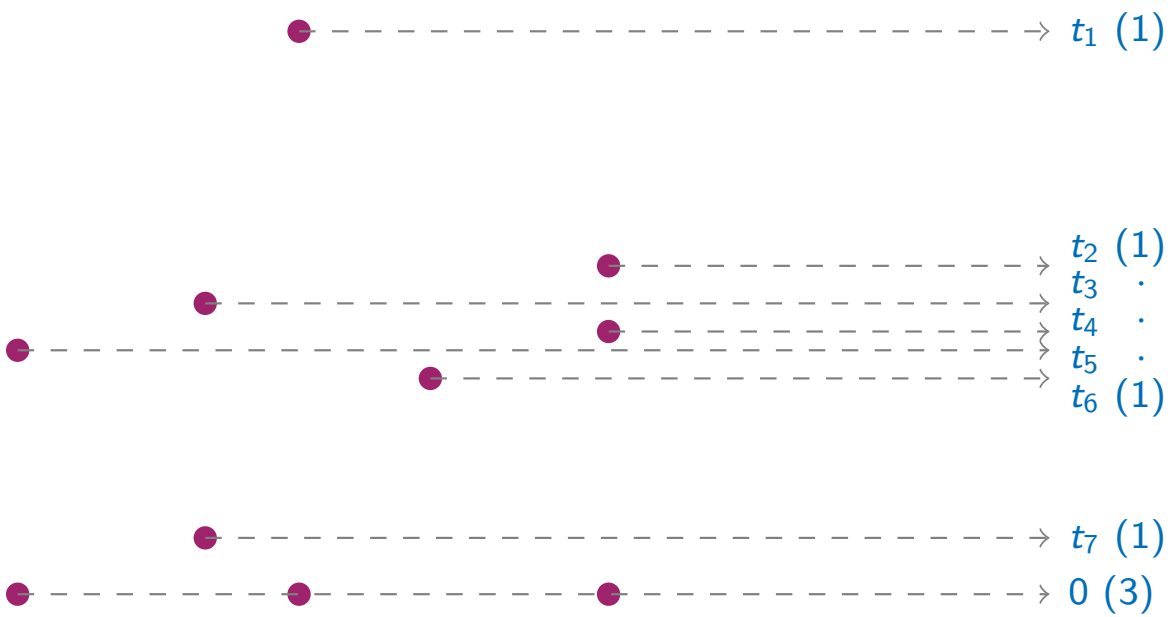


Sampled phylogeny ($\mathcal{T}$)

Data generated by speciation fossilisation model

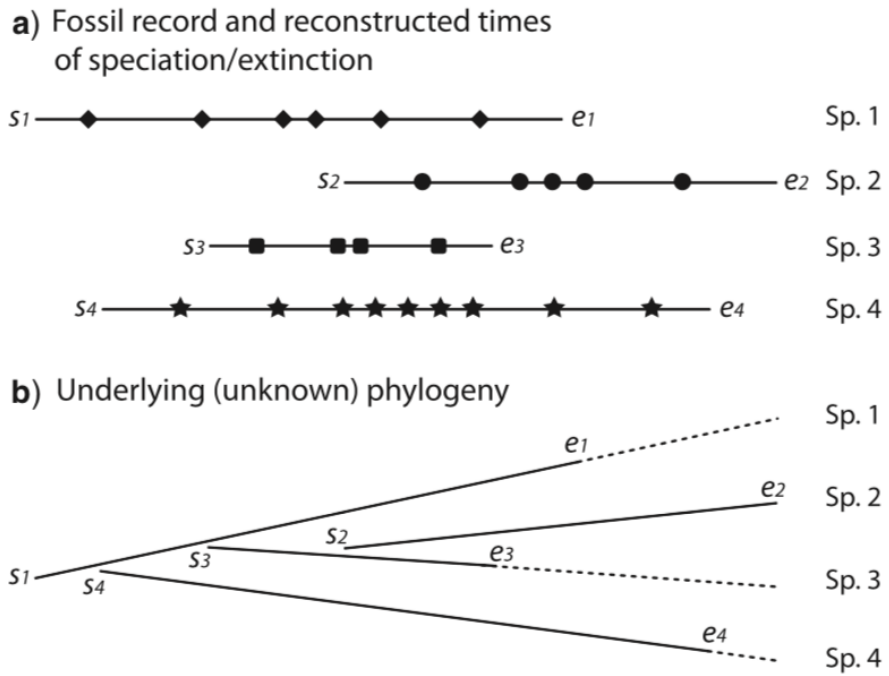


Data generated by speciation fossilisation model



τ = dates with # of samples

What these data tells us?



Silvestro *et al.* (2014)

- We can estimate parameters: λ, μ, ψ, ρ from only fossil occurrence data.
- Parameters of the birth-death model define the distribution of the branch lengths.
- Fossil sampling dates on their own are informative about the times of the events on the phylogeny.

Hierarchical model

- A typical model in BEAST2 is hierarchical, that is, parameters of one model are the outcomes of another model.
- The tree is an outcome of the speciation fossilisation model.
- The same tree is a parameter for the substitution and clock models.

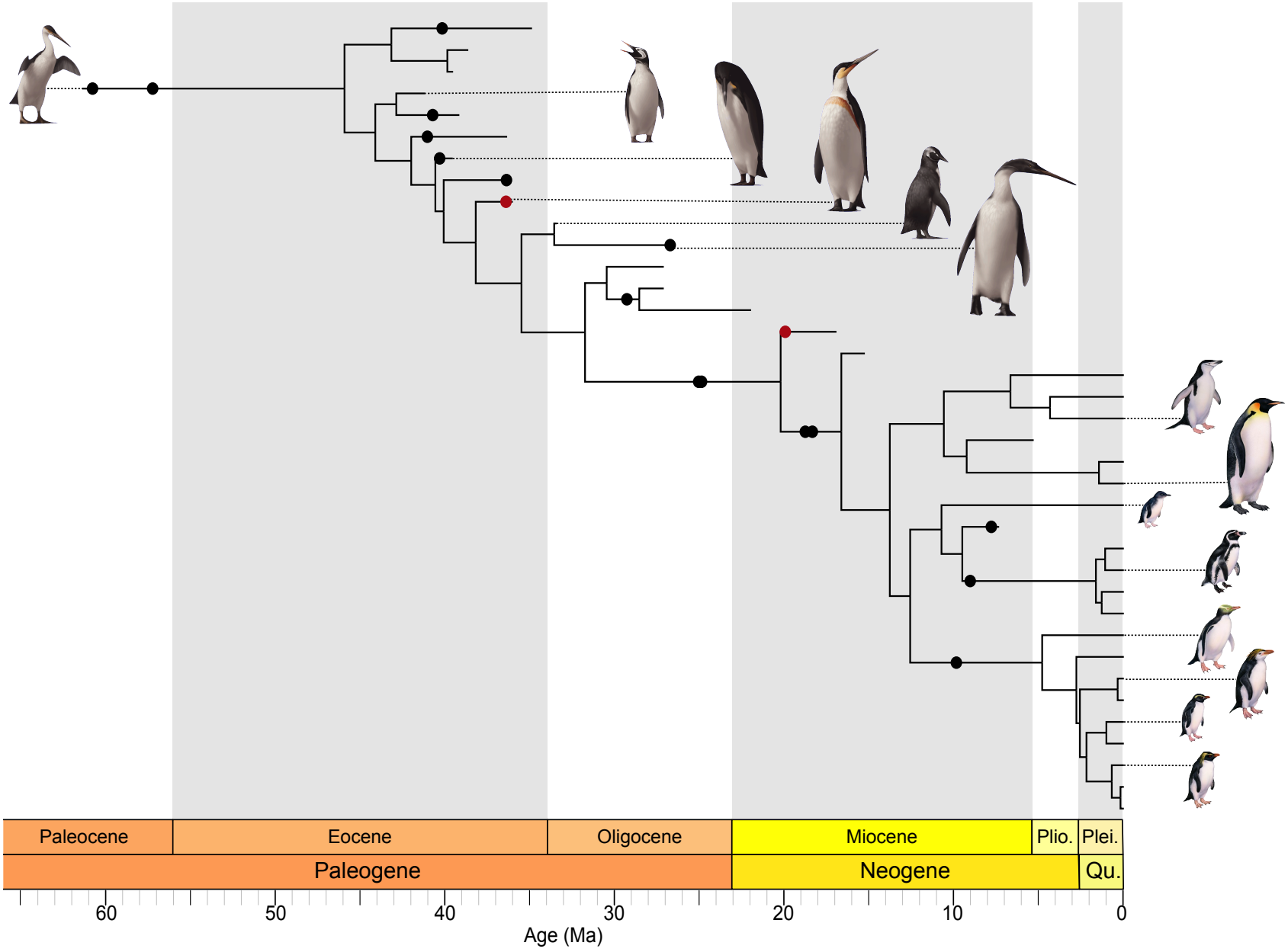
Hierarchical model

D	molecular sequence data
M	morphological data matrix
τ	sampling dates
T	dated phylogeny without sampling dates
$\mathcal{T} = (T, \tau)$	dated phylogeny
Θ	tree (FBD) model parameters
$\Omega = (\Omega_D, \Omega_M)$	molecular and morphological evolution model parameters
$E = (E_D, E_M)$	molecular and morphological clock model parameters

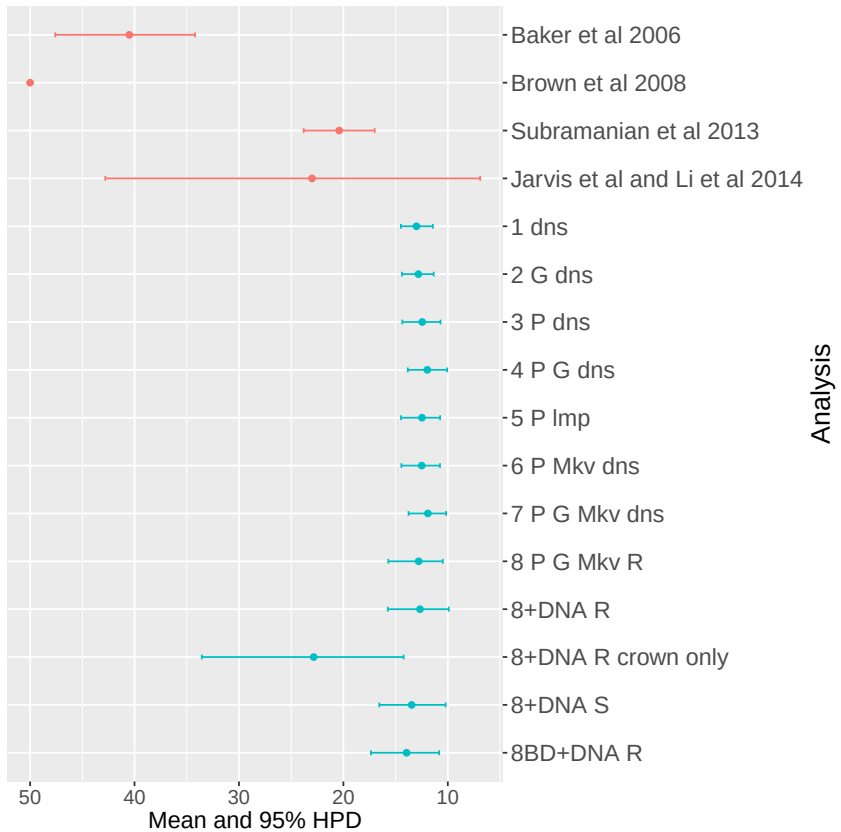
$$P(T, \Theta, E, \Omega \mid D, M, \tau) = \frac{P(D, M, \tau \mid T, \Theta, E, \Omega) P(T, \Theta, E, \Omega)}{P(D, M, \tau)} =$$

$$\frac{P(D, M \mid T, \tau, E, \Omega) P(T, \tau \mid \Theta) P(\Theta) P(E) P(\Omega)}{P(D, M, \tau)}$$

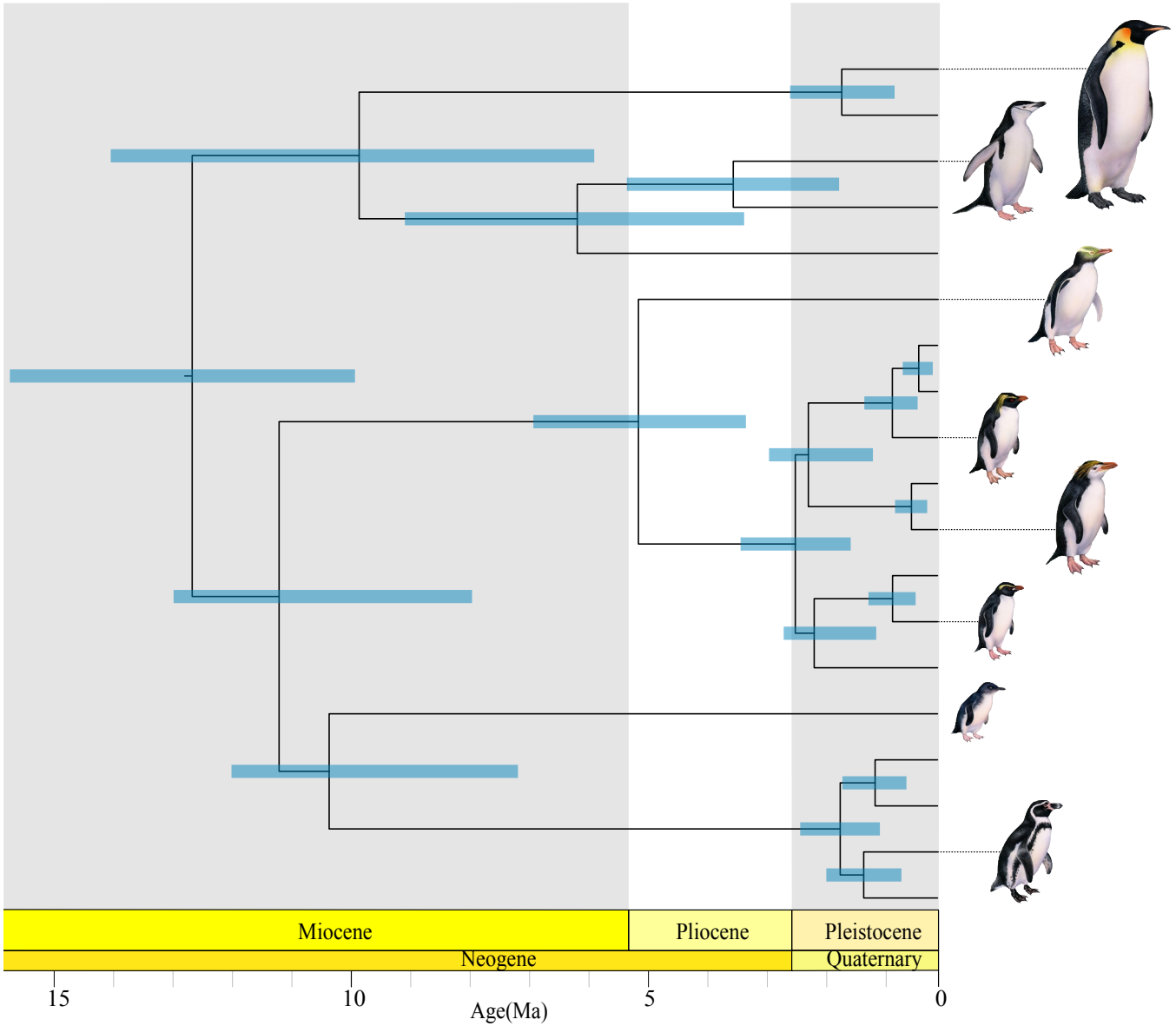
Total evidence analysis of penguins



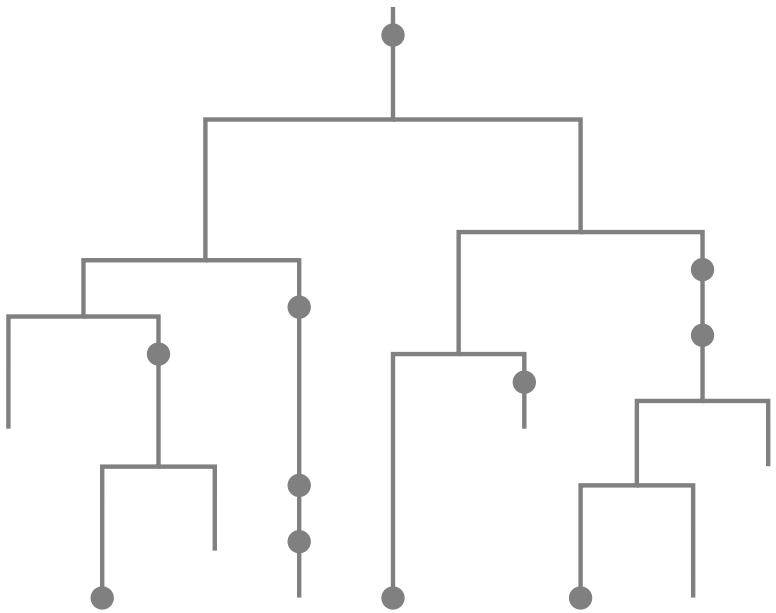
The age of the MRCA of extant penguins



Dated extant penguin phylogeny

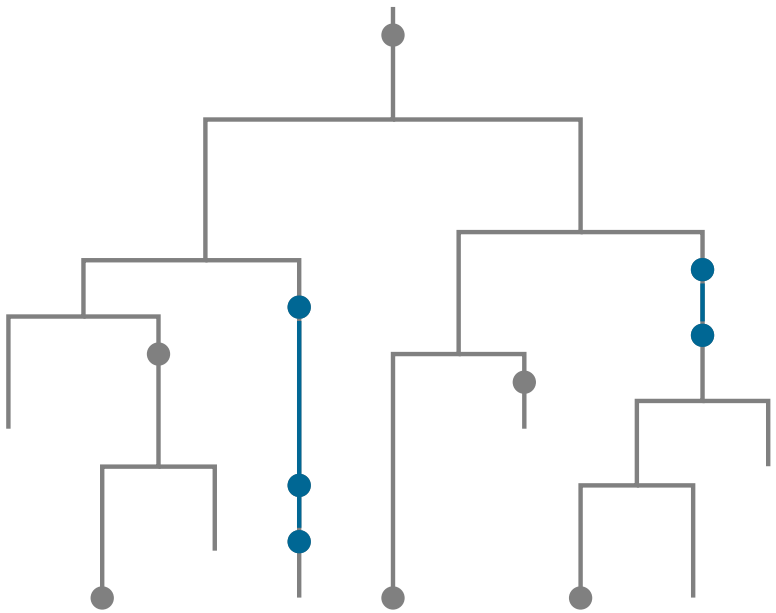


Fossil ages



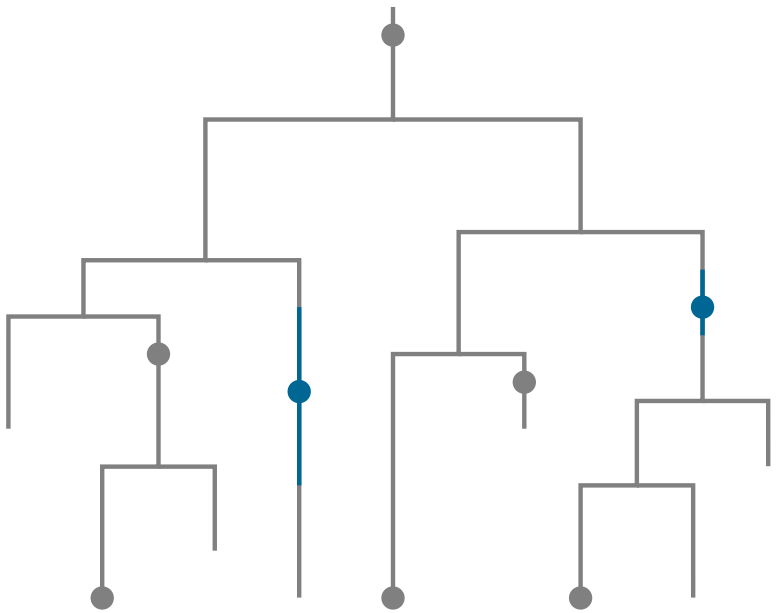
- Several fossil specimens are replaced with a single specimen
- The morphological data from multiple specimens is merged to a single record.
- Stratigraphic ranges package which allows assigning multiple fossils to the same species is on the way.

Fossil ages



- Several fossil specimens are replaced with a single specimen
- The morphological data from multiple specimens is merged to a single record.
- Stratigraphic ranges package which allows assigning multiple fossils to the same species is on the way.

Fossil ages



- Several fossil specimens are replaced with a single specimen
- The morphological data from multiple specimens is merged to a single record.
- Stratigraphic ranges package which allows assigning multiple fossils to the same species is on the way.

Advantages of the joint inference

- All available fossil data can be used.
- This approach does not require an out-group species.
- It accounts for uncertain fossil placement.
- It is model based: we directly model fossil sampling process (e.g., with FBD).

Problems with the joint inference

- Morphological data is very sparse: therefore the fossil sampling model will have a great impact on the results.
- Models of morphological evolution are very primitive (yet are better than parsimony).

Speciation-fossilisation models

Speciation-fossilisation models

Only a few models have been implemented for the joint inference to date. Most of the models are variants of the birth-death model with or without sampling.

1. Yule model (pure birth without sampling)
2. Uniform model (not a birth-death model)
3. Birth-death model (no sampling)
4. FBD (birth-death with sampling)
5. Skyline FBD
6. Diversified skyline FBD

Speciation-fossilisation models

Only a few models have been implemented for the joint inference to date. Most of the models are variants of the birth-death model with or without sampling.

1. Yule model (pure birth without sampling)
2. Uniform model (not a birth-death model)
3. Birth-death model (no sampling)
4. FBD (birth-death with sampling)
5. Skyline FBD
6. Diversified skyline FBD

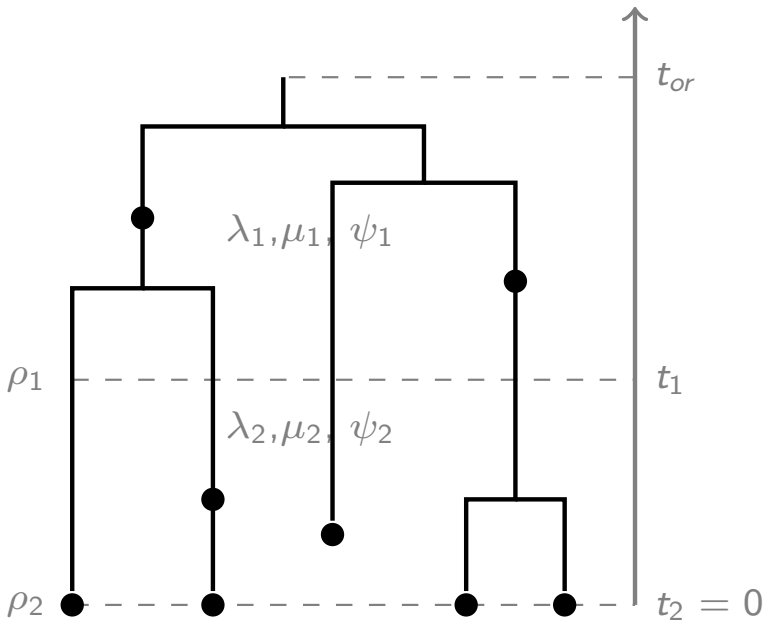
Skyline FBD

Stadler *et al.* (2012), Gavryushkina *et al.* (2014)

There are k time intervals and parameters remain constants within the intervals but may vary from one interval to another

- birth rates $\lambda_1, \ldots, \lambda_k$
- death rates $\mu_1, \ldots, \mu_k$
- sampling rates $\psi_1, \ldots, \psi_k$
- sampling at interval end points probabilities $\rho_1, \ldots, \rho_k$

Model parameters: $\eta = (t_{or}, \bar{\lambda}, \bar{\mu}, \bar{\psi}, \bar{\rho})$



Sampled tree

Skyline FBD

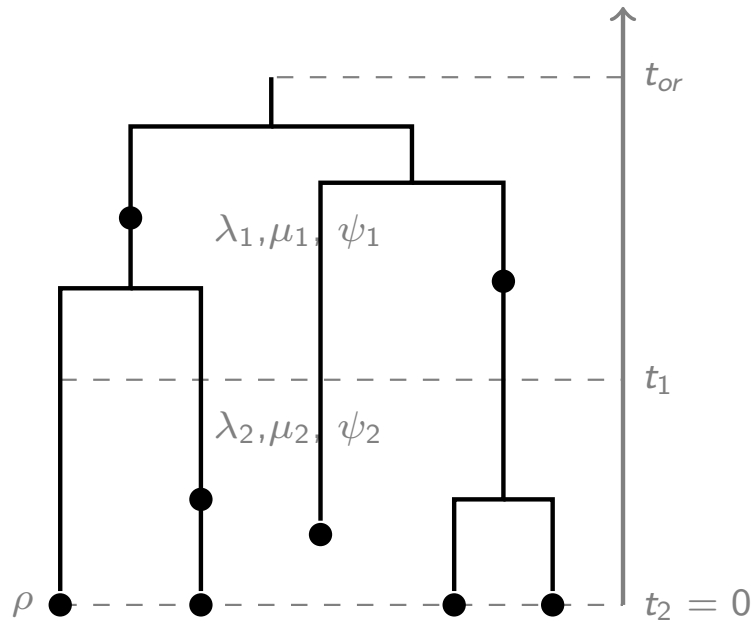
Stadler *et al.* (2012), Gavryushkina *et al.* (2014)

There are k time intervals and parameters remain constants within the intervals but may vary from one interval to another

- birth rates $\lambda_1, \ldots, \lambda_k$
- death rates $\mu_1, \ldots, \mu_k$
- sampling rates $\psi_1, \ldots, \psi_k$
- sampling at interval end points probabilities $\rho_1, \ldots, \rho_k$

Model parameters: $\eta = (t_{or}, \bar{\lambda}, \bar{\mu}, \bar{\psi}, \bar{\rho})$

Often $\rho_1 = \ldots = \rho_{k-1} = 0$



Sampled tree

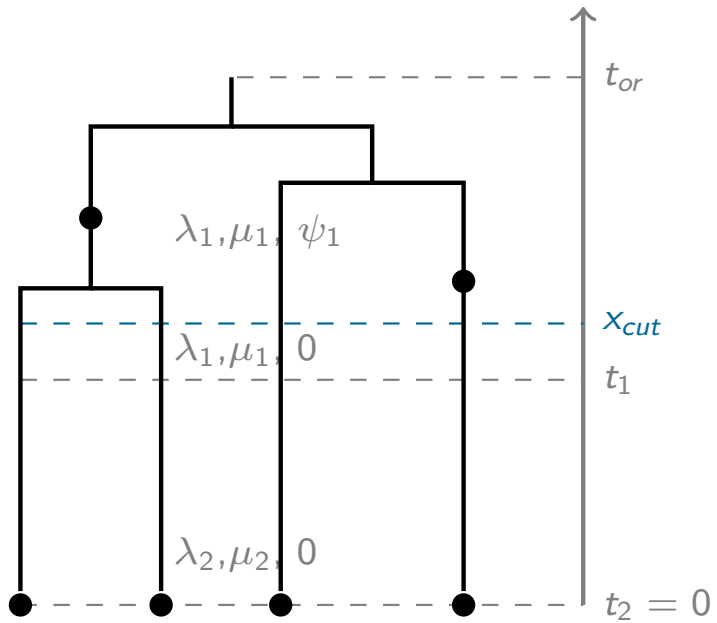
Diversified skyline FBD

Hönna *et al.* (2011) and Zhang *et al.* (2016)

There is a cut-off time x_{cut}. There are no fossil samples after x_{cut} and a single descendant of every branch existing at time x_{cut} is sampled at present.

- birth rates $\lambda_1, \ldots, \lambda_k$
- death rates $\mu_1, \ldots, \mu_k$
- sampling rates $\psi_1, \ldots, \psi_m, 0, \ldots, 0$

Model parameters:
$\eta = (t_{or}, \bar{\lambda}, \bar{\mu}, \psi_1, \ldots, \psi_m)$



Sampled tree

Influence of the speciation-fossilisation model

Matzke and Wright (2016) analysis of fossil Canidae:

	Canidae	crown Caninae	crown *Canis*
Uniform	49 Ma	38.9 Ma	27.5 Ma
FBD	36.3 Ma	9.8 Ma	2.8 Ma

Zhang *et al.* (2016) analysis of Hymenoptera + outgroups:

	Hymenoptera
Uniform	306 Ma
Skyline FBD	346.6 Ma
Diversified Skyline FBD	251.7 Ma

Using FBD model

For more accurate results:

- use informative prior distributions for parameters where possible,
- allow for rate variation (Skyline FBD),
- where appropriate use FBD diversified model,
- If the data violates the assumptions of the available FBD models then further work is needed! (*e.g.*, theoretical extension of FBD, simulation studies).

Modelling morphological evolution

Problems with models of morphological evolution:

- Lewis Mk model is too simple (other models are developed: ordered characters, inertia model, *etc*).
- Continuous traits are presented as discrete characters (models of continuous character evolution are available).
- Correlation of characters is ignored.
- Morphological characters are merged from different specimens.
- Morphological clock models need to be developed.

When to use joint or total-evidence analysis

Joint or total-evidence inference requiers:

- dated fossils from several species belonging to the clade (more than five),
- enough morphological data from both fossil and extant species.

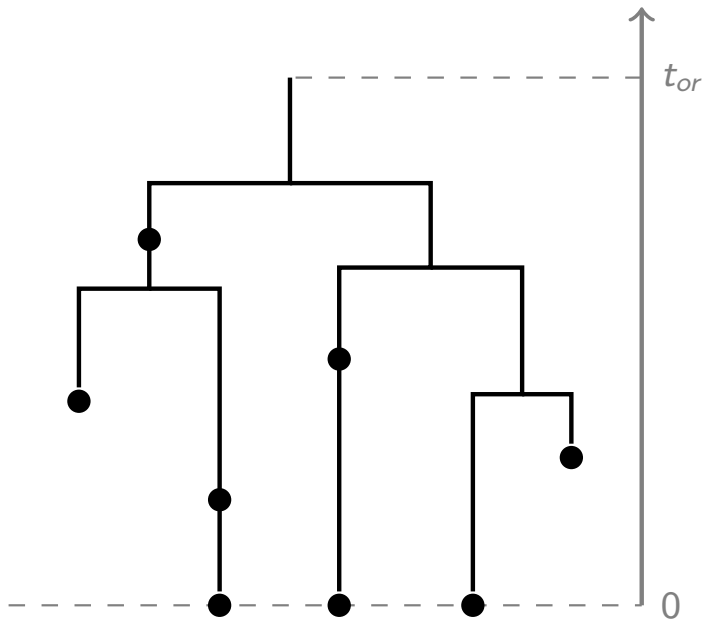
Sampled ancestors and sampled ancestor trees

Birth-death-sampling model

Stadler 2010, Stadler *et al.* 2011

The process starts at time $t_{or} > 0$ and ends at time zero (present time).

- birth rate λ
- death rate μ
- sampling rate ψ
- removal probability r
- sampling at present probability ρ



Sampled tree

- FBD is a birth-death sampling model with $r = 0$.
- When $\rho = 0$ only three out of four parameters (λ, μ, ψ, r) are identifiable.

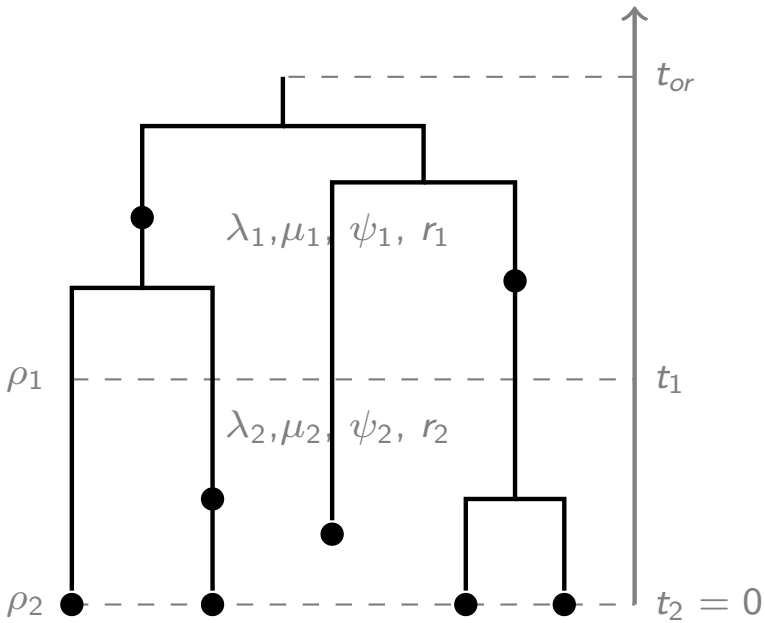
Skyline birth-death-sampling model

Stadler *et al.* (2012), Gavryushkina *et al.* (2014)

There are k time intervals with parameters:

- birth rates $\lambda_1, \ldots, \lambda_k$
- death rates $\mu_1, \ldots, \mu_k$
- sampling rates $\psi_1, \ldots, \psi_k$
- removal probabilities $r_1, \ldots, r_k$
- sampling at interval end points $\rho_1, \ldots, \rho_k$

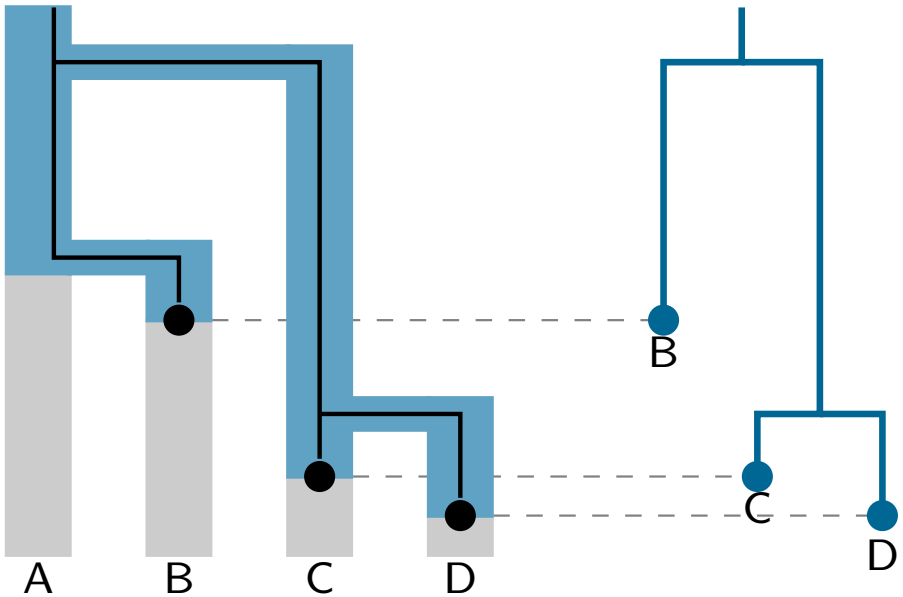
Model parameters: $\eta = (t_{or}, \bar{\lambda}, \bar{\mu}, \bar{\psi}, \bar{r}, \bar{\rho})$



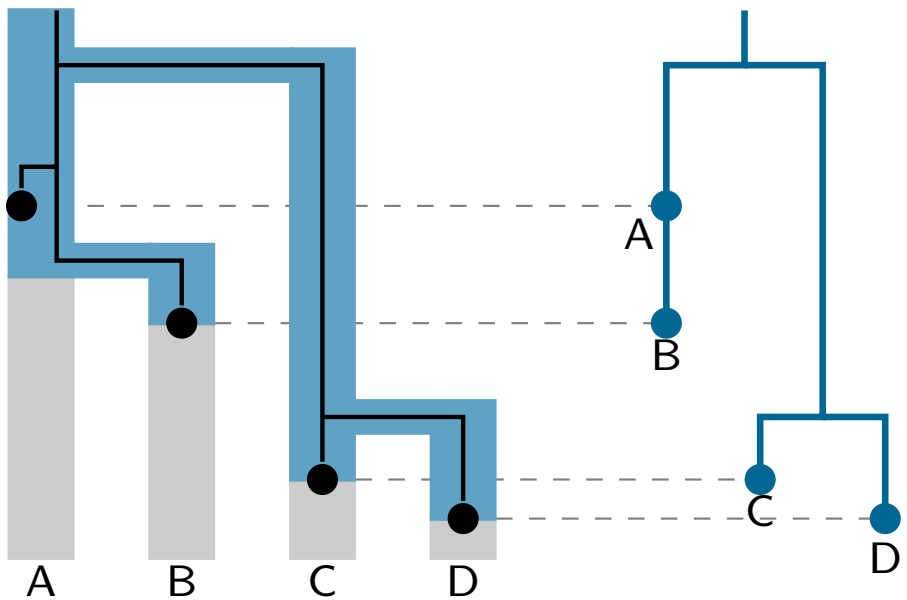
Sampled tree

- When $\rho_1 = \ldots = \rho_{k-1} = 0$ only $4k$ out of $4k + 1$ parameters $(\bar{\lambda}, \bar{\mu}, \bar{\psi}, \bar{r}, \rho)$ are identifiable.

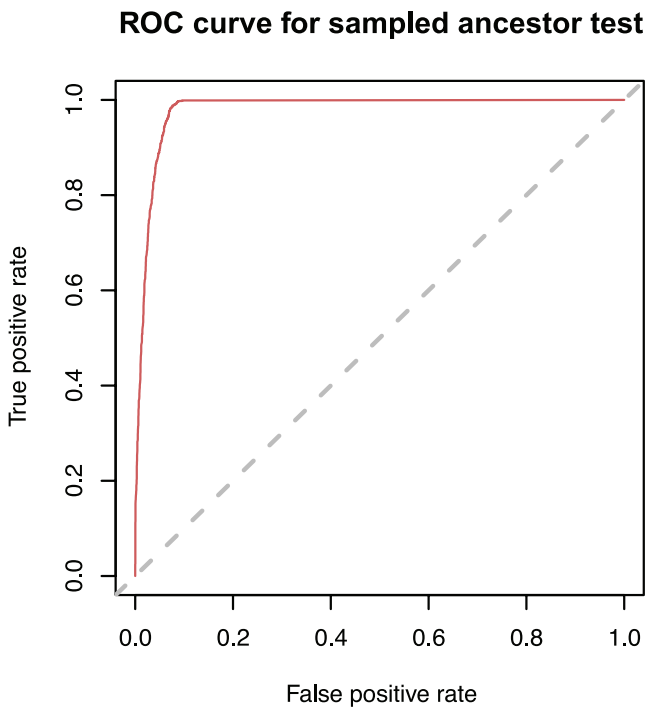
Transmission trees



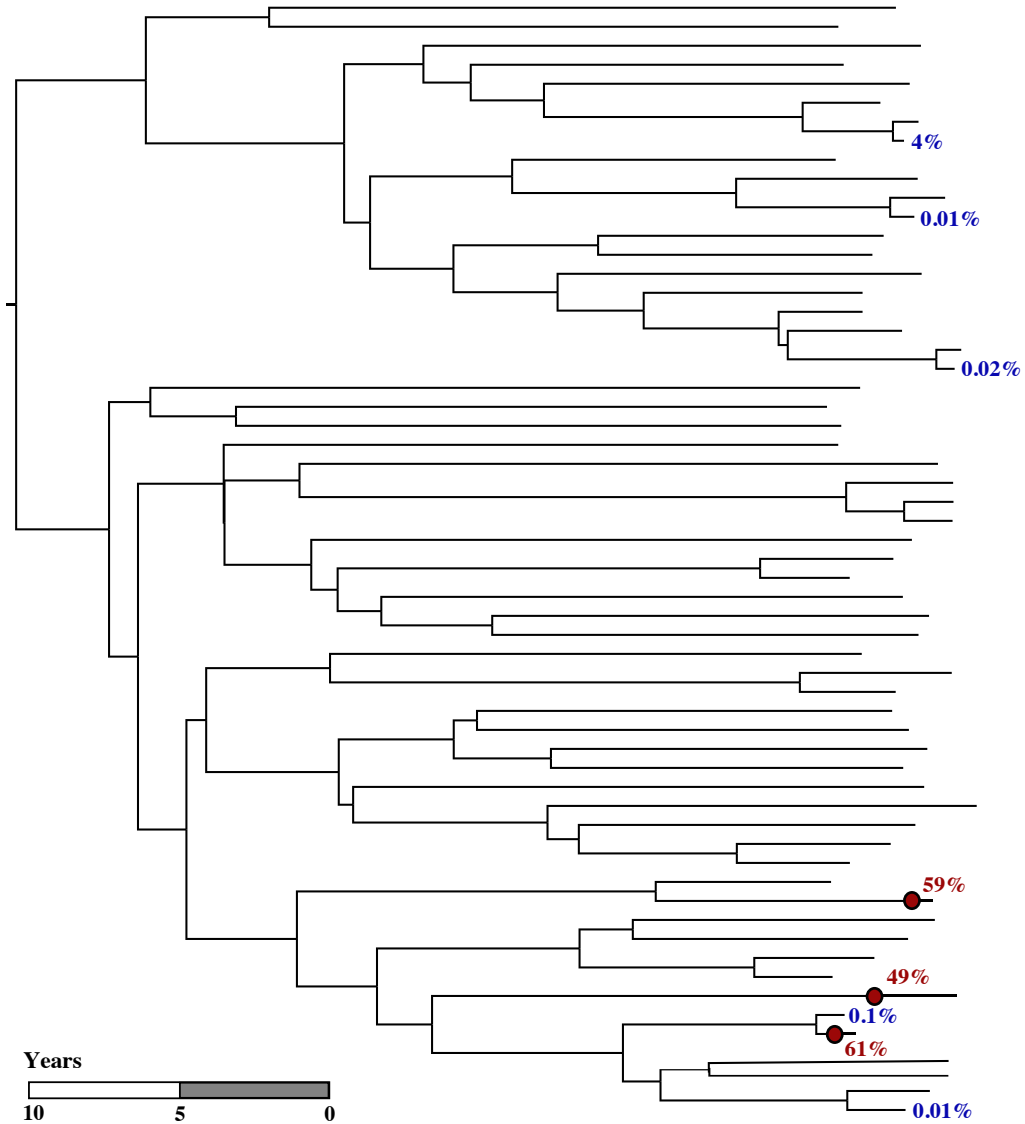
Sampled ancestors in transmission trees



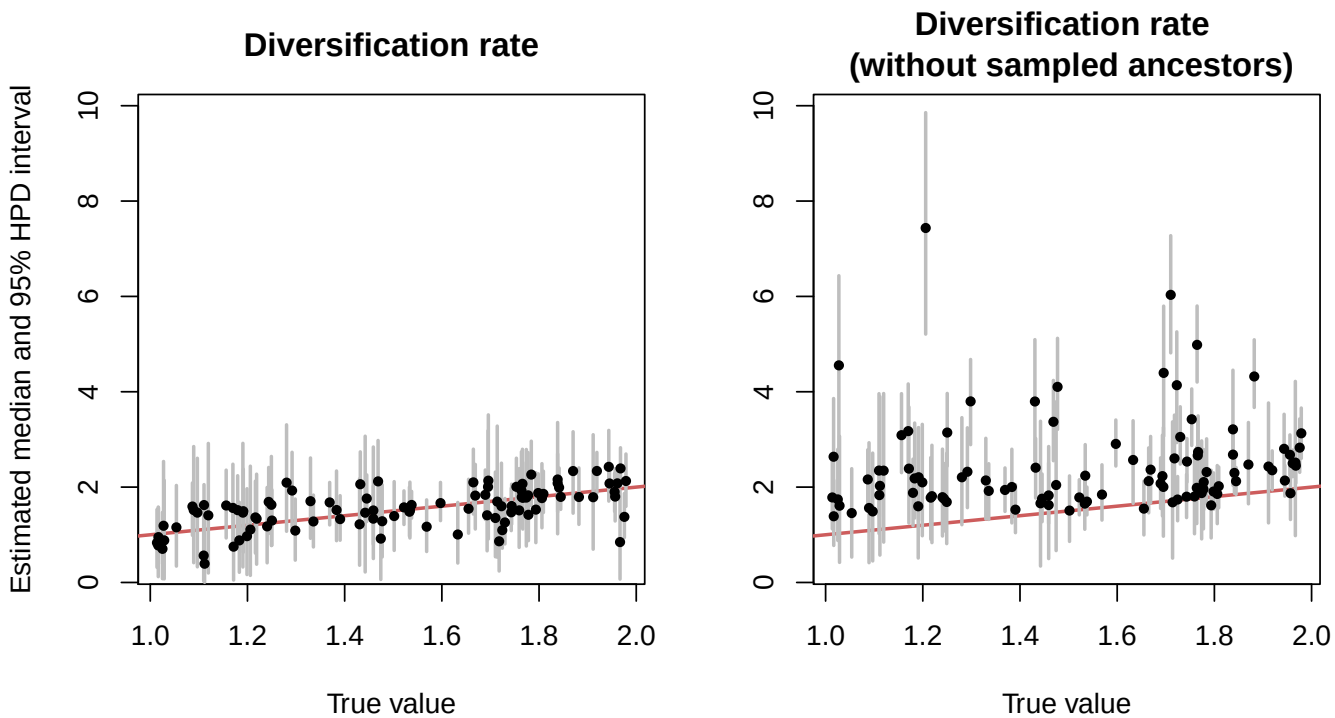
Sampled ancestors



HIV phylogeny

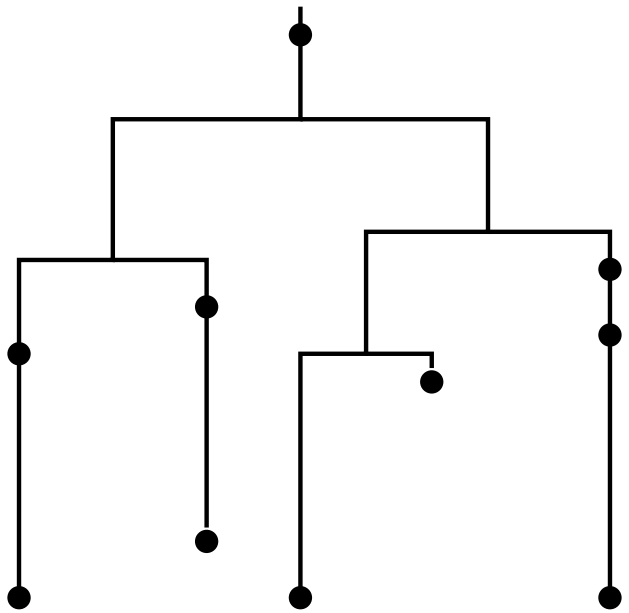


Not accounting for sampled ancestors leads to biased parameter estimates



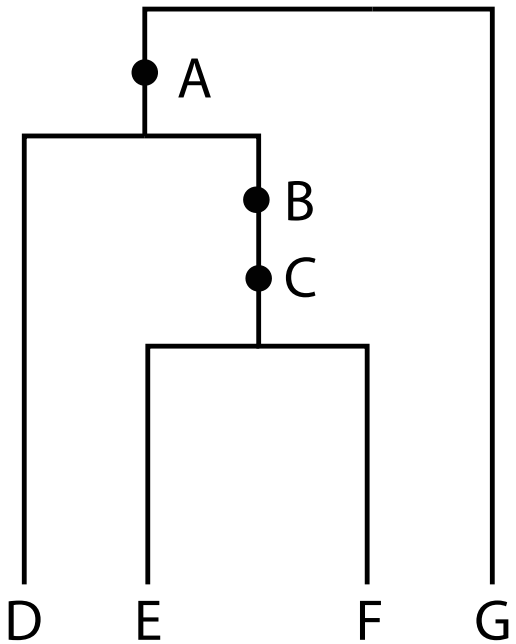
Diversification rate $d = \lambda - \mu$

Summarising sampled ancestor (SA) trees



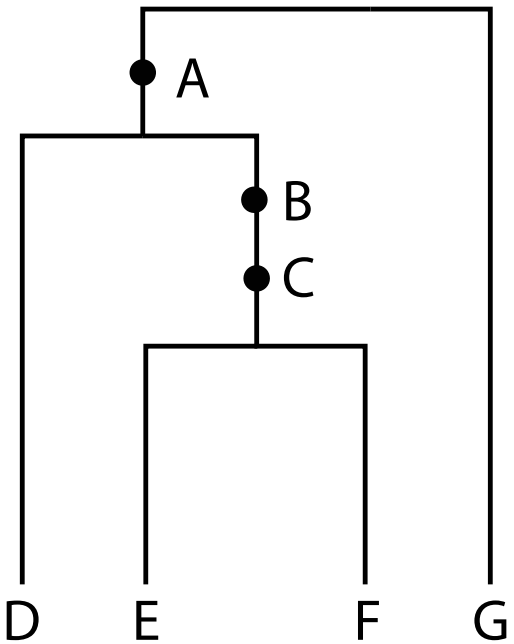
Sampled ancestor tree

Sampled ancestor (SA) clade



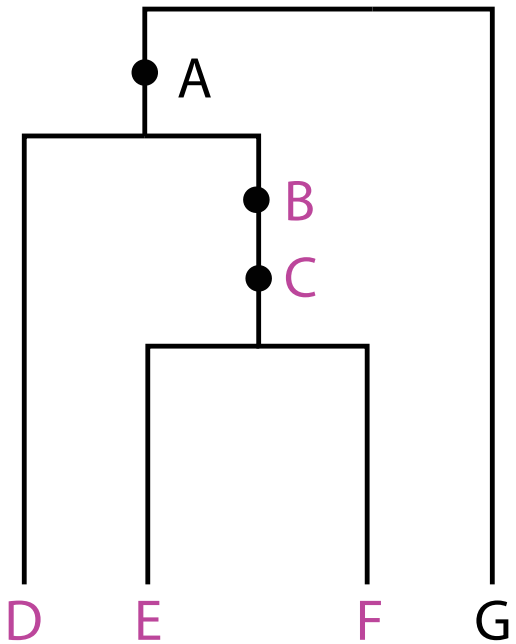
- A sampled ancestor (SA) clade is either a set of taxa $\{A_1, \ldots, A_n\}$ or a pair of taxon B and a set of taxa $\{A_1, \ldots, A_n\}$.

Sampled ancestor (SA) clade



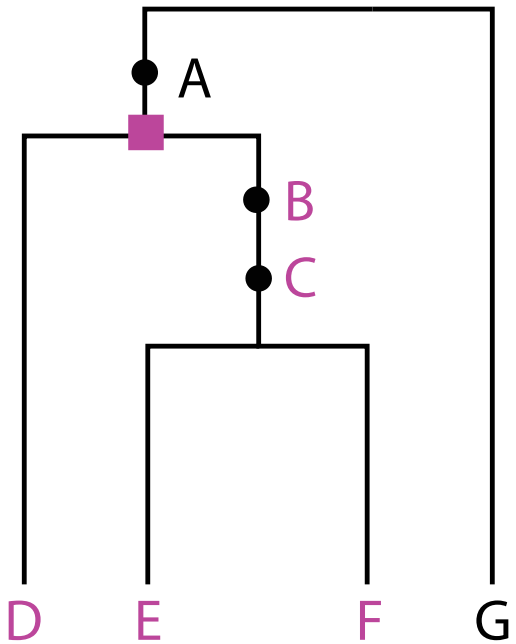
- A sampled ancestor (SA) clade is either a set of taxa $\{A_1, \ldots, A_n\}$ or a pair of taxon B and a set of taxa $\{A_1, \ldots, A_n\}$.
- A SA tree has a SA clade $\{A_1, \ldots, A_n\}$ if $A_1, \ldots, A_n$ are monophyletic and the most recent common ancestor of $A_1, \ldots, A_n$ is a bifurcation node.

Sampled ancestor (SA) clade



- A sampled ancestor (SA) clade is either a set of taxa $\{A_1, \ldots, A_n\}$ or a pair of taxon B and a set of taxa $\{A_1, \ldots, A_n\}$.
- A SA tree has a SA clade $\{A_1, \ldots, A_n\}$ if $A_1, \ldots, A_n$ are monophyletic and the most recent common ancestor of $A_1, \ldots, A_n$ is a bifurcation node.

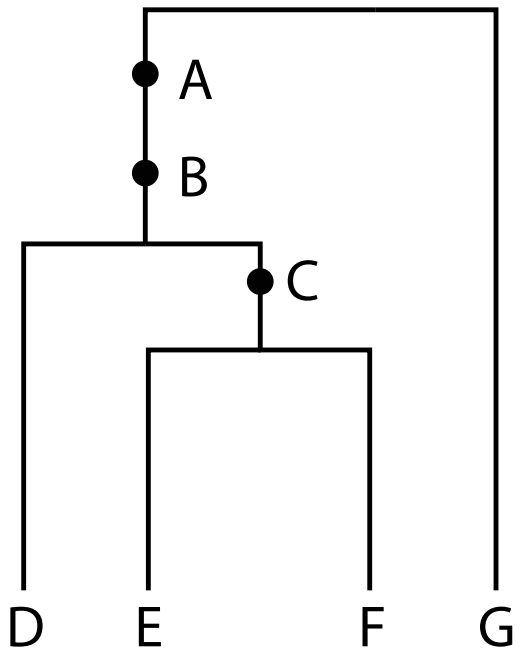
Sampled ancestor (SA) clade



$\{B, C, D, E, F\}$

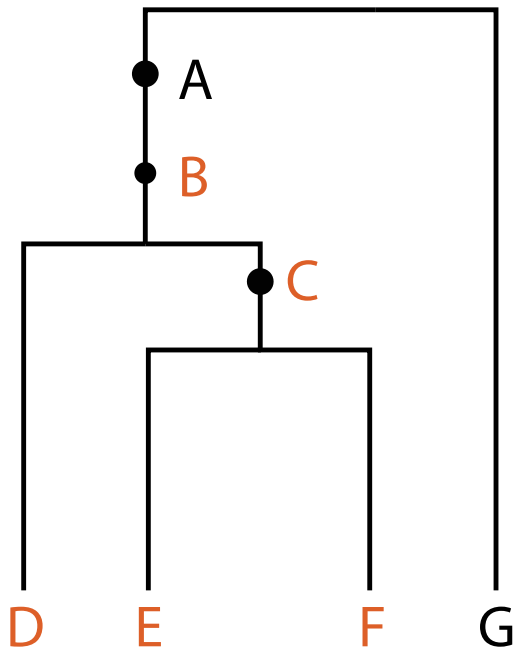
- A sampled ancestor (SA) clade is either a set of taxa $\{A_1, \ldots, A_n\}$ or a pair of taxon B and a set of taxa $\{A_1, \ldots, A_n\}$.
- A SA tree has a SA clade $\{A_1, \ldots, A_n\}$ if $A_1, \ldots, A_n$ are monophyletic and the most recent common ancestor of $A_1, \ldots, A_n$ is a bifurcation node.

Sampled ancestor (SA) clade



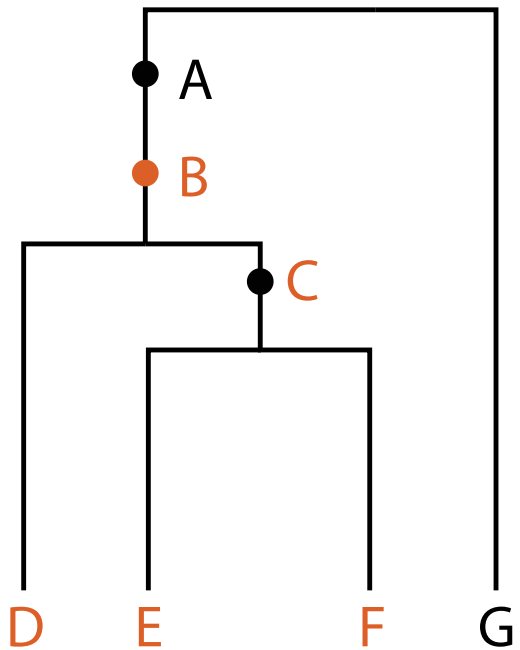
- A sampled ancestor (SA) clade is either a set of taxa $\{A_1, \ldots, A_n\}$ or a pair of taxon B and a set of taxa $\{A_1, \ldots, A_n\}$.
- A SA tree has a SA clade $\{A_1, \ldots, A_n\}$ if $A_1, \ldots, A_n$ are monophyletic and the most recent common ancestor of $A_1, \ldots, A_n$ is a bifurcation node.
- A SA tree has a SA clade $(B, \{A_1, \ldots, A_n\})$ if $B, A_1, \ldots, A_n$ are monophyletic and the most recent common ancestor of $B, A_1, \ldots, A_n$ is a sampled ancestor B.

Sampled ancestor (SA) clade



- A sampled ancestor (SA) clade is either a set of taxa $\{A_1, \ldots, A_n\}$ or a pair of taxon B and a set of taxa $\{A_1, \ldots, A_n\}$.
- A SA tree has a SA clade $\{A_1, \ldots, A_n\}$ if $A_1, \ldots, A_n$ are monophyletic and the most recent common ancestor of $A_1, \ldots, A_n$ is a bifurcation node.
- A SA tree has a SA clade $(B, \{A_1, \ldots, A_n\})$ if $B, A_1, \ldots, A_n$ are monophyletic and the most recent common ancestor of $B, A_1, \ldots, A_n$ is a sampled ancestor B.

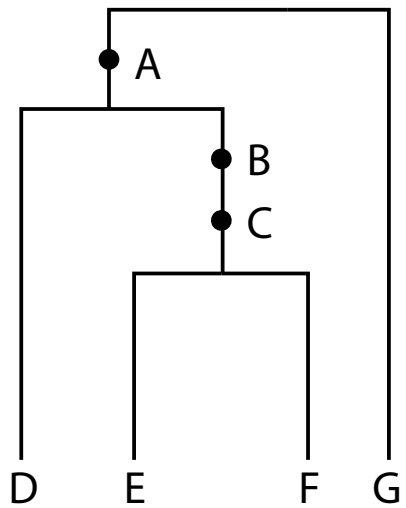
Sampled ancestor (SA) clade



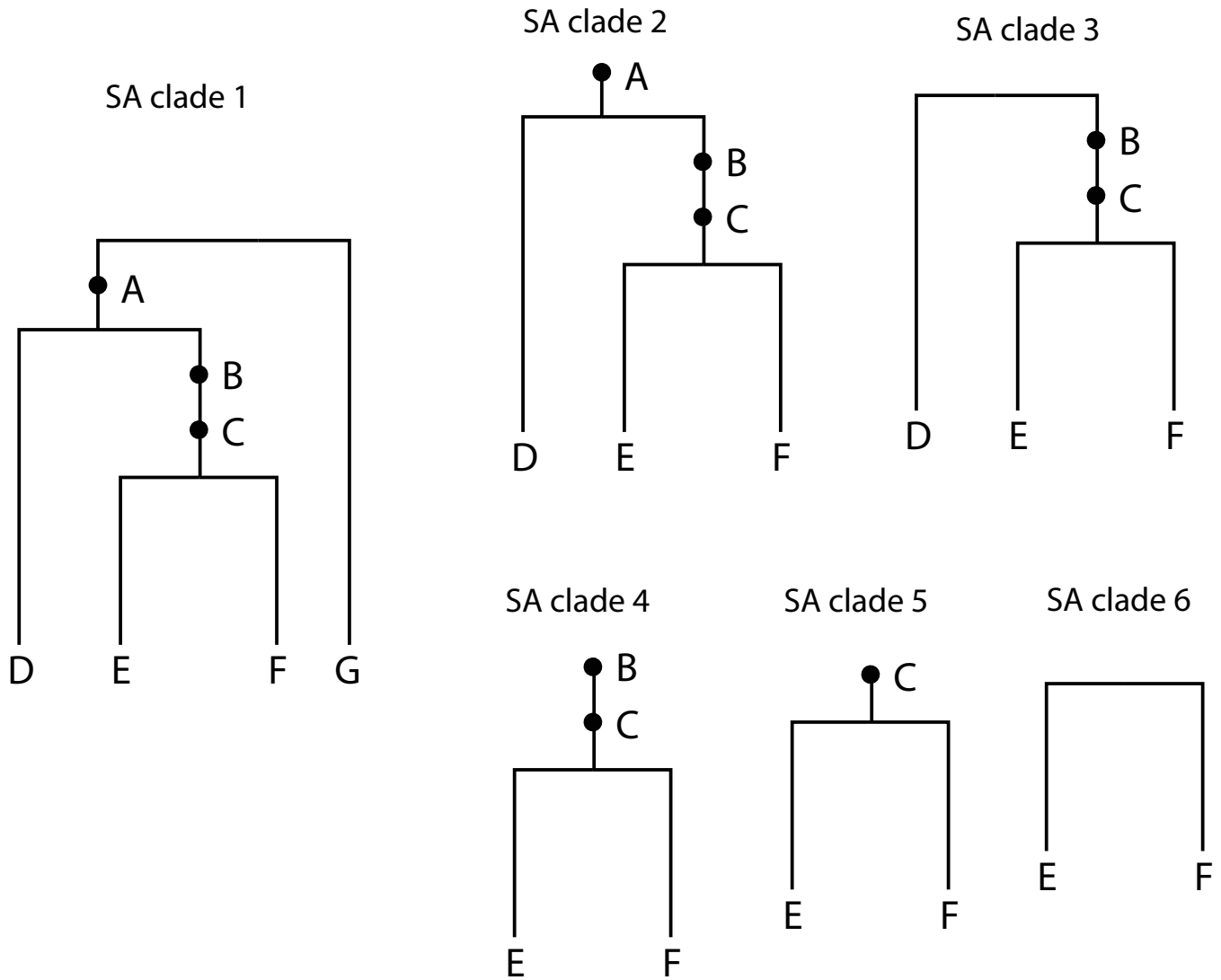
$(B, \{C, D, E, F\})$

- A sampled ancestor (SA) clade is either a set of taxa $\{A_1, \ldots, A_n\}$ or a pair of taxon B and a set of taxa $\{A_1, \ldots, A_n\}$.
- A SA tree has a SA clade $\{A_1, \ldots, A_n\}$ if $A_1, \ldots, A_n$ are monophyletic and the most recent common ancestor of $A_1, \ldots, A_n$ is a bifurcation node.
- A SA tree has a SA clade $(B, \{A_1, \ldots, A_n\})$ if $B, A_1, \ldots, A_n$ are monophyletic and the most recent common ancestor of $B, A_1, \ldots, A_n$ is a sampled ancestor B.

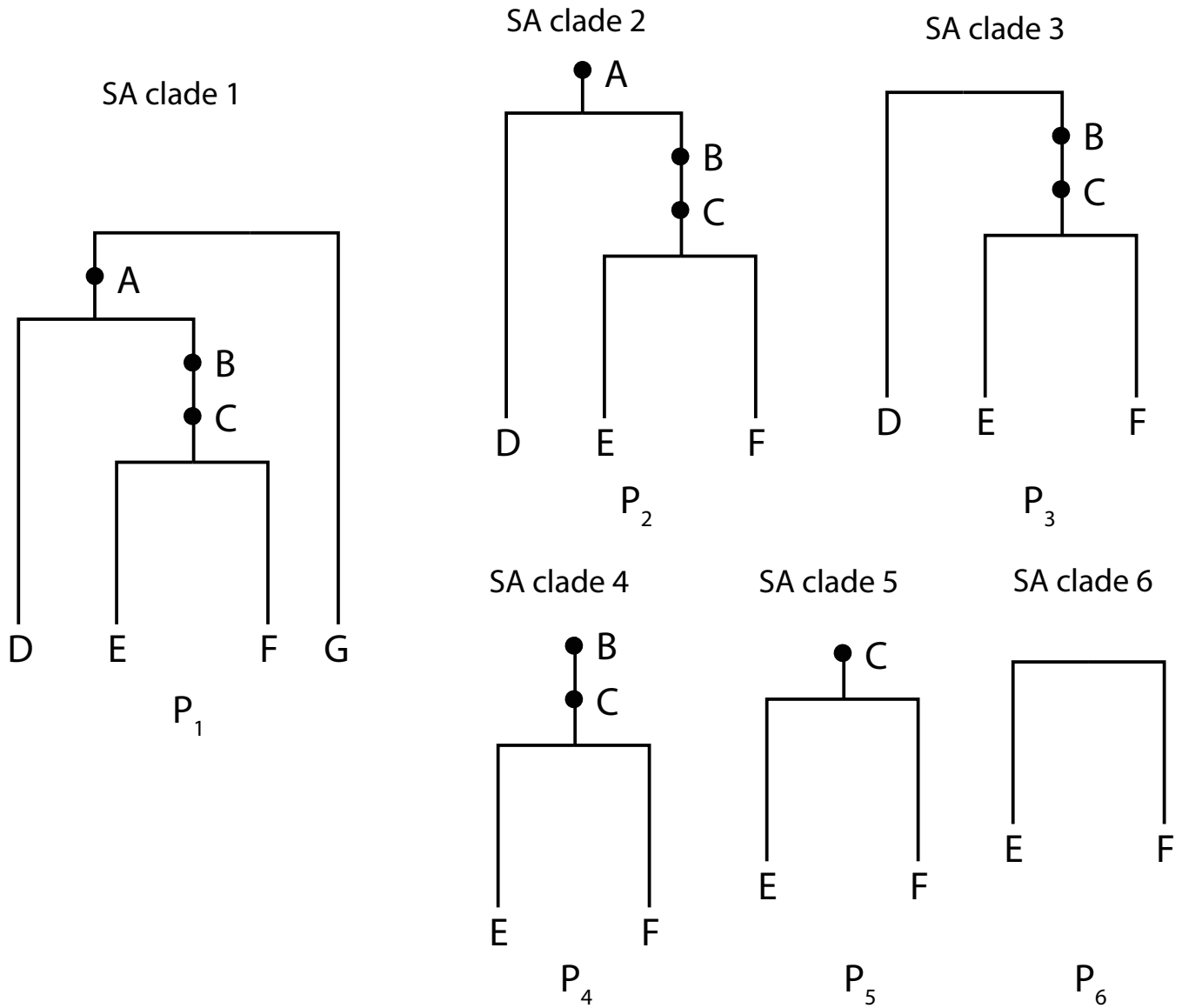
Maximum sampled ancestor clade credibility (MSACC) tree



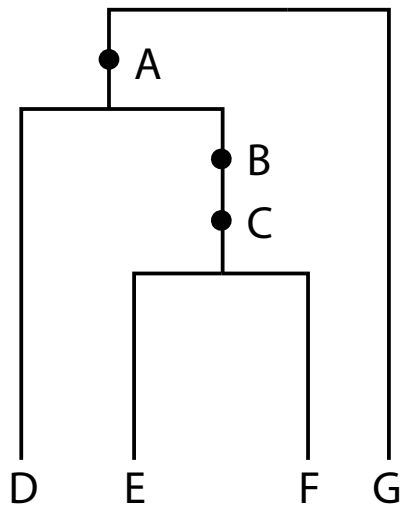
Maximum sampled ancestor clade credibility (MSACC) tree



Maximum sampled ancestor clade credibility (MSACC) tree

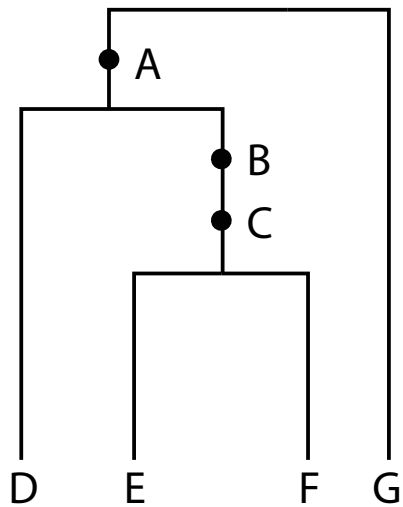


Maximum sampled ancestor clade credibility (MSACC) tree



$$P_1 \times P_2 \times P_3 \times P_4 \times P_5 \times P_6$$

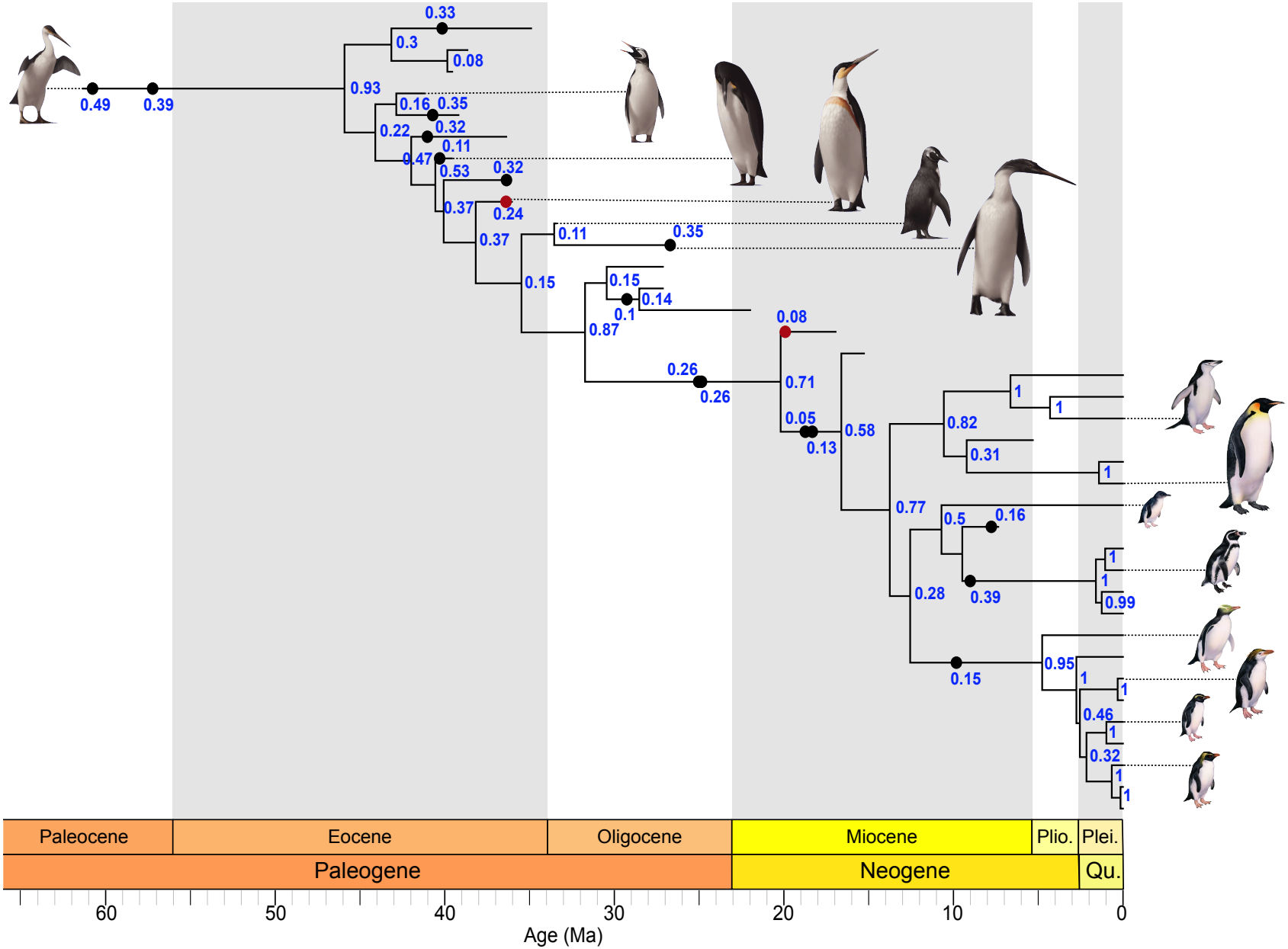
Maximum sampled ancestor clade credibility (MSACC) tree



$$P_1 \times P_2 \times P_3 \times P_4 \times P_5 \times P_6$$

A MSACC tree is a tree from the posterior distribution which maximises the product of the posterior probabilities of its SA clades.

MSACC tree of penguins



Node ages in the summary tree

For each SA clade presented in the tree:

- consider all the ages of a given SA clade that occur in the posterior distribution,
- use the mean of the ages as the age of the SA clade in the phylogeny or
- the median of the ages as the age of the SA clade in the phylogeny

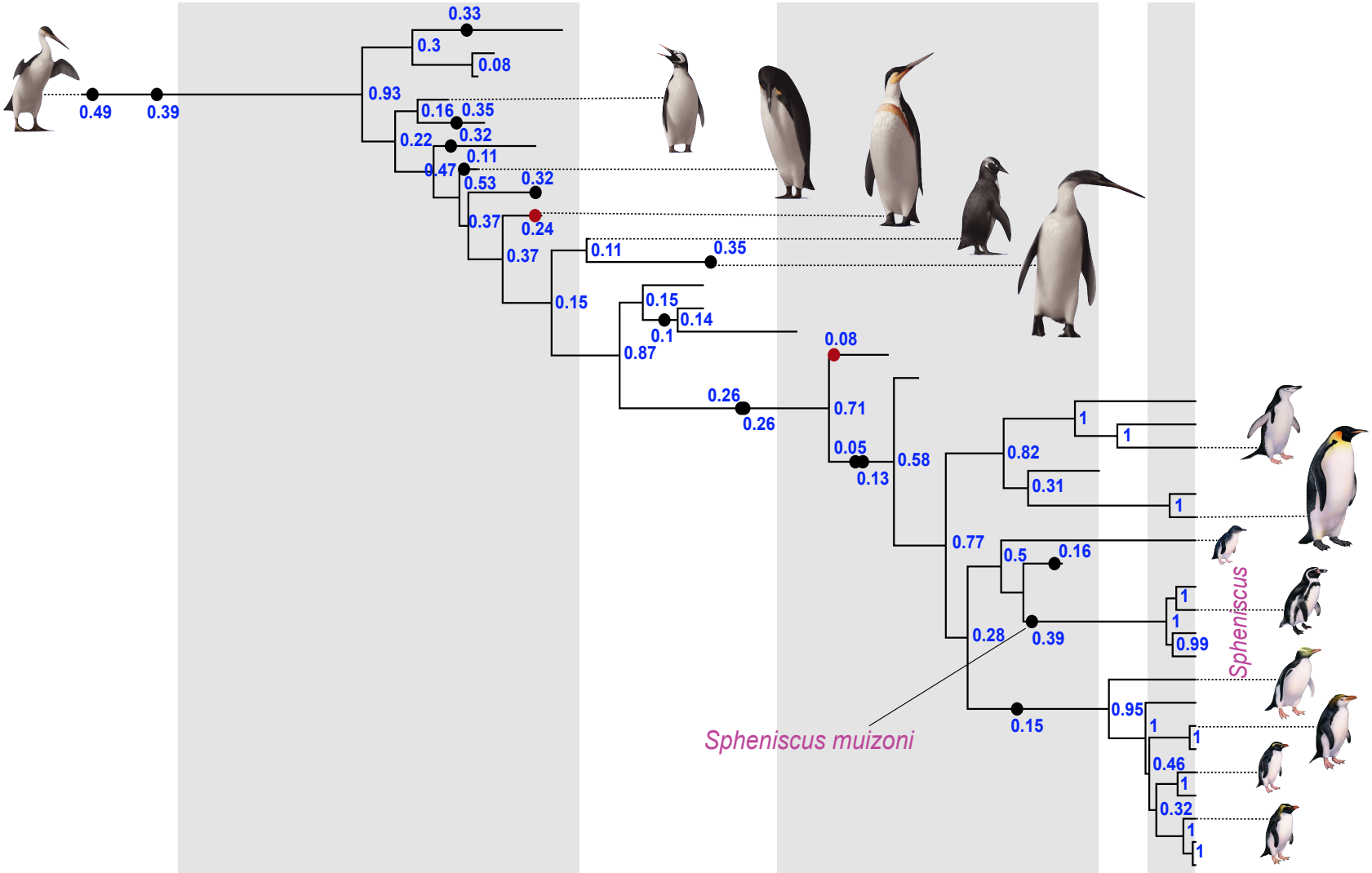
Problems:

- This approach can result in negative branch lengths.
- 'Common ancestor age' is not available for SA trees.
- One can keep the ages of the MSACC tree.

Summarising the posterior distribution of SA trees

- A summary tree does not fully represent the posterior distribution.
- Given that the morphological data of fossil species is limited, there is a high degree of uncertainty in the positions of fossils in the phylogeny.
- Further tree processing is required to interpret the results of a total-evidence analysis.

MSACC tree of penguins



- *Spheniscus muizoni* lies on the branch leading to the extant *Spheniscus* clade in 61% of the posterior trees.

Inferring sampled ancestors

- We can estimate the posterior probability $P(H_1|D, \bar{\tau}, M)$ of an individual to be a sampled ancestor as the proportion of trees in which this individual is a sampled ancestor.
- A prior probability $P(H_1|M)$ that an individual sampled at time x before present is a sampled ancestor is

$$\int (1-r)(1-p_0(x \mid \eta))f(\eta)d\eta$$

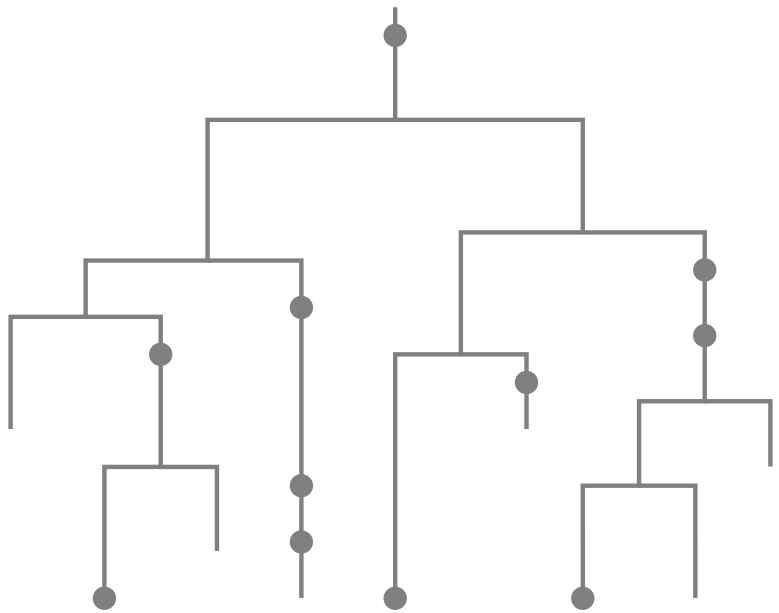
where $\eta = (\lambda, \mu, \psi, r, \rho)$ and $p_0(x \mid \eta)$ is the probability that an individual alive at time x before present leaves no sampled descendants.

$$BF = \frac{P(D, \bar{\tau}|H_1, M)}{P(D, \bar{\tau}|H_2, M)} = \frac{P(H_1|D, \bar{\tau}, M)P(H_2|M)}{P(H_2|D, \bar{\tau}, M)P(H_1|M)} =$$

$$= \frac{P(H_1|D, \bar{\tau}, M)(1 - P(H_1|M))}{(1 - P(H_1|D, \bar{\tau}, M))P(H_1|M)}$$

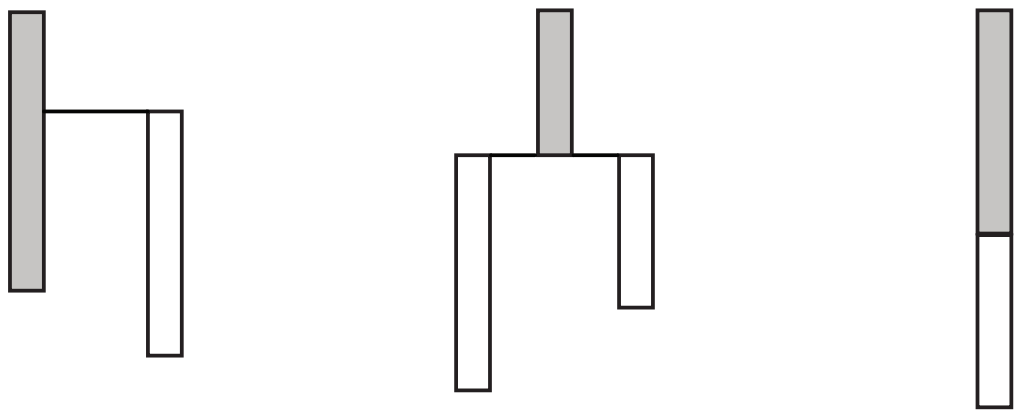
Stratigraphic range model

Multiple fossils of the same species



Three types of speciation[1]

(i) asymmetric speciation (ii) symmetric speciation (iii) anagenetic speciation

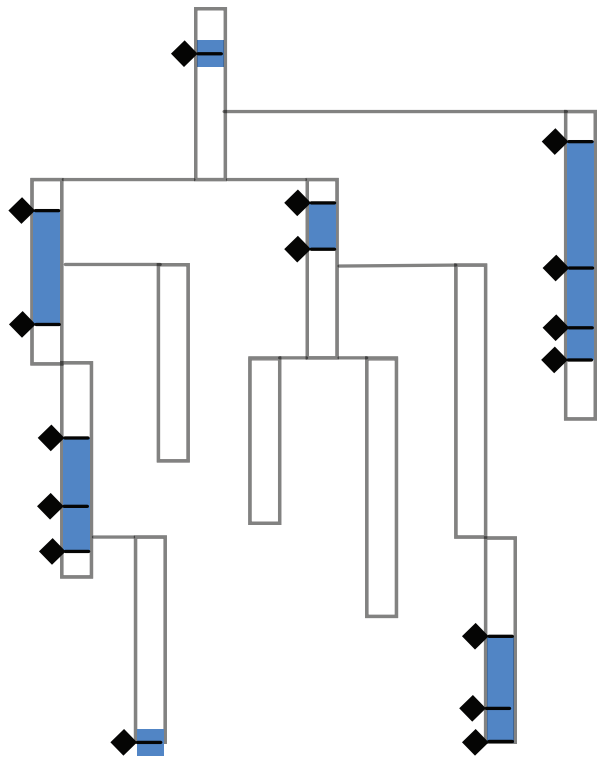


[1] Foote (1996)

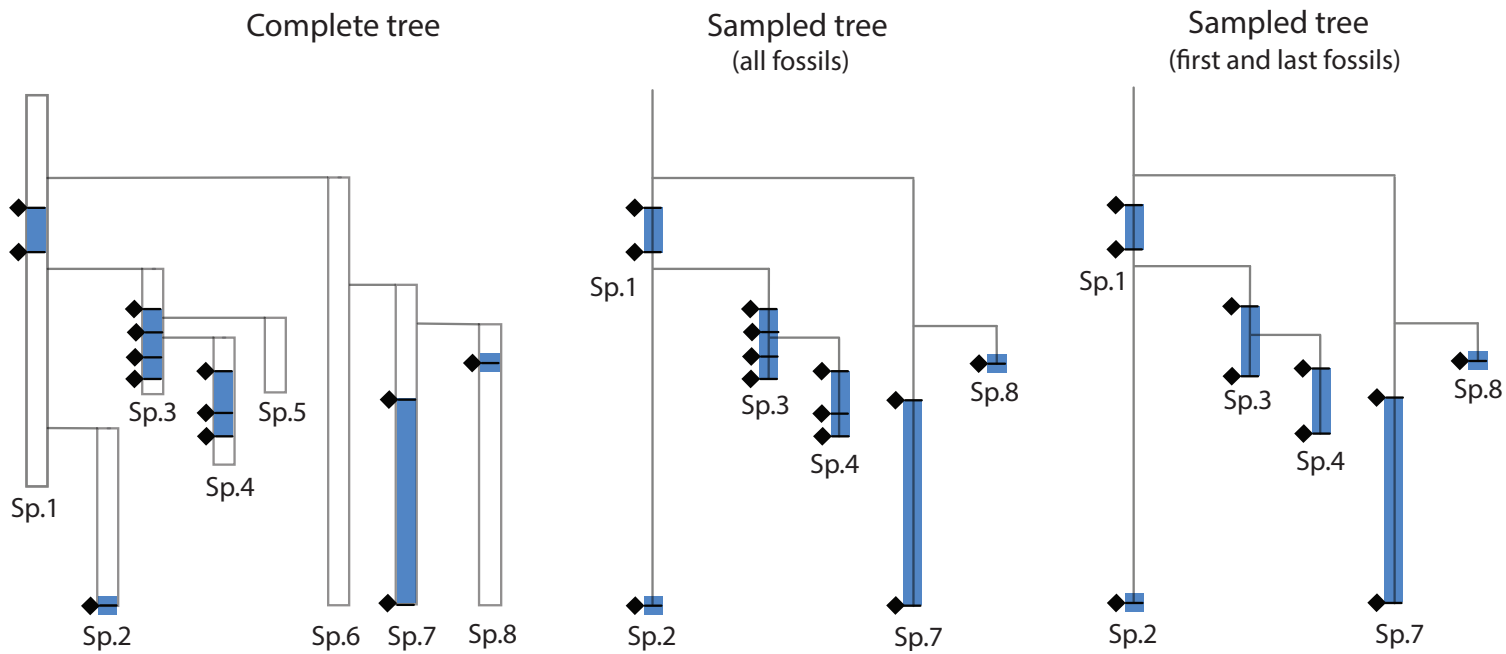
Stratigraphic range (SR) birth-death process

Stadler *et al.* (2018)

- branching rate, λ,
- death rate, μ,
- fossil sampling rate, ψ,
- probability of a symmetric branching event, β,
- anagenetic speciation rate, λ_a, and
- probability of sampling at present, ρ.



Pure budding process $\beta = \lambda_a = 0$



Probability density of an oriented sampled tree with and without intermediate fossils

$\mathcal{T}_s^o$ an oriented sampled tree.
$\mathcal{T}_{s,r}^o$ an oriented sampled tree with κ intermediate fossils removed.
κ the total number of *intermediate* fossils that neither represent the first nor the last occurrences.
L_s the sum of the lengths of all stratigraphic ranges.

$$f[\mathcal{T}_s^o \mid \lambda, \mu, \psi, \rho, x_0] =$$

$$\frac{\psi^k \rho^l \lambda^{n-j-1}}{1-p(x_0)} \prod_{i=1}^{3n-j-1} \hat{q}_{asym}(B_i) \prod_{i=1}^{n-j-l} p(y_i) \prod_{i \in I} \left(1 - \frac{q(o_i)}{\widetilde{q}_{asym}(o_i)} \frac{\widetilde{q}_{asym}(y_{a(i)})}{q(y_{a(i)})}\right)$$

$$f[\mathcal{T}_{s,r}^o \mid \lambda, \mu, \psi, \rho, x_0] = \psi^{-\kappa} f[\mathcal{T}_s^o \mid \lambda, \mu, \psi, \rho, x_0] e^{\psi L_s}$$

Estimating oriented sampled trees

Stratigraphic-ranges addon for BEAST2 implements operators to sample oriented sampled trees with stratigraphic ranges represented by only the first and the last samples (or single sample).

Analysis of North American canids

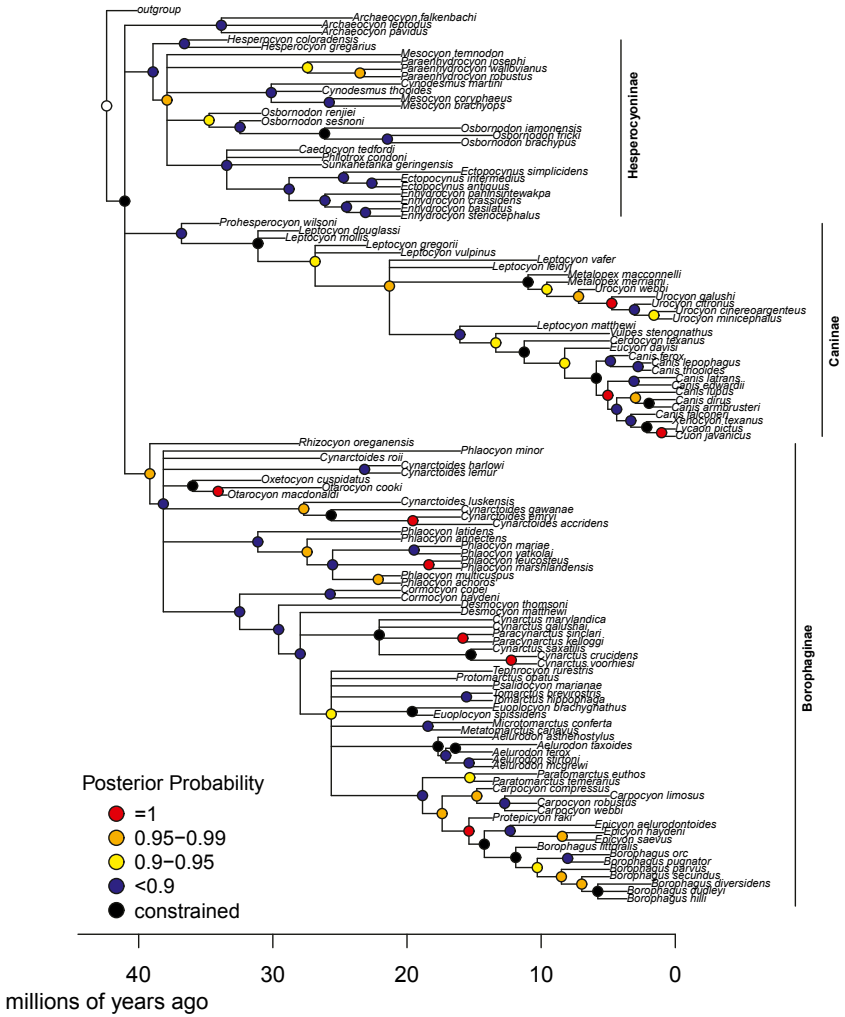
Data from Slater (2015):

- Morphological data of 120 living and fossil North American canid species.
- The age range of the first occurrence of every taxon.
- The age range of the last occurrence (if any) of every taxon.
- The last occurrence for extant species is fixed to zero.

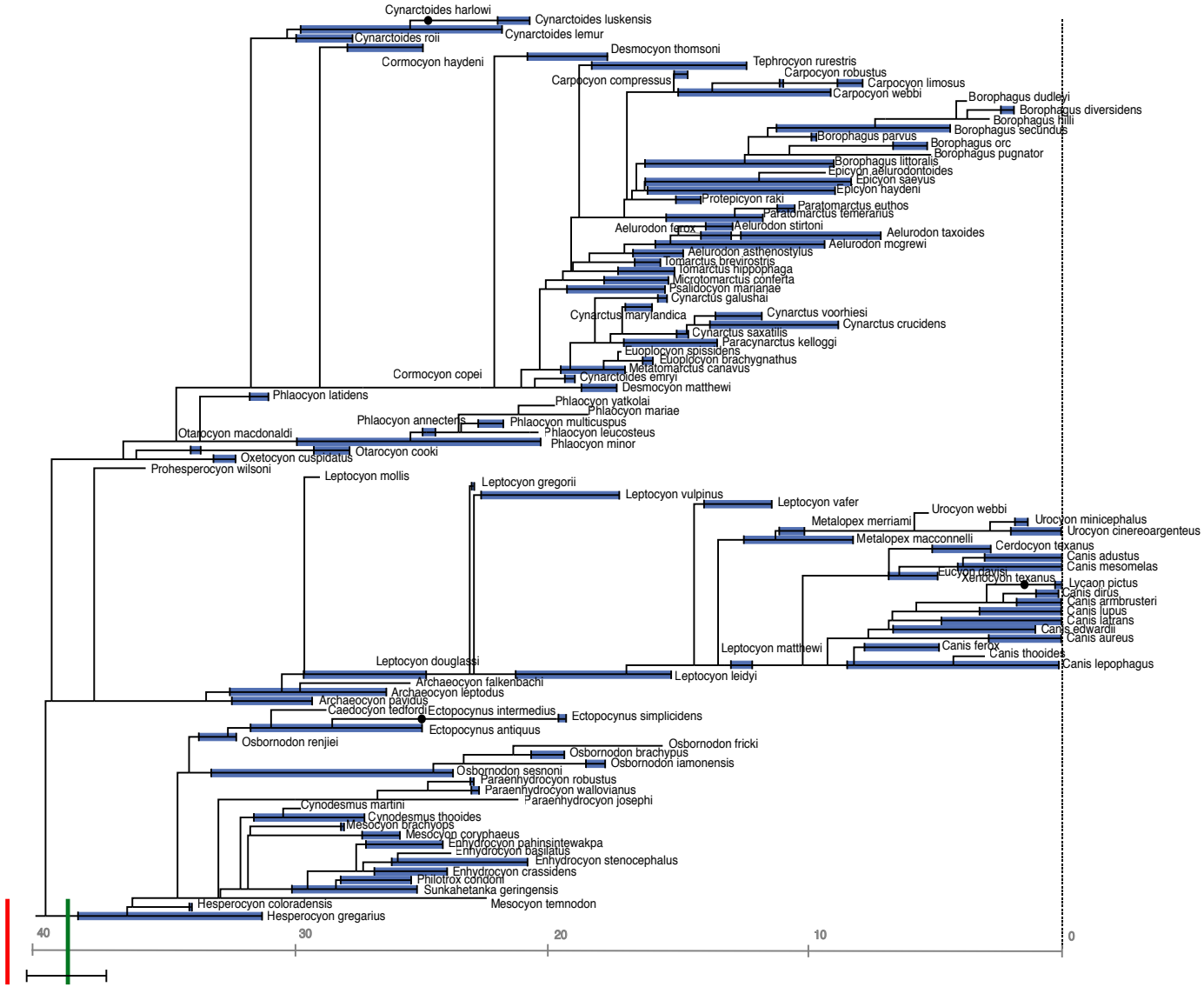
Analysis:

- The first and last occurrences are sampled within the ranges in MCMC.
- The same morphological coding is assigned to the first and the last occurrences.

Analysis by Slater (MCC tree)



Random tree from the posterior distribution



Thank you!