More intuition on phylogenetic tree estimation

Working with real data

- Tips in the phylogeny: genes, species, individuals, or some other evolving entity, are referred to as taxa.

Add taxa

- Adding taxa breaks up branches in the phylogeny. This often improves the performance of maximum likelihood
- Improve model evaluation and model parameter estimation
- Adding more outgroup taxa often improves rooting
- Adding taxa can considerably increase the computational demand of analyses

Types of real data

- Full genome: Largely restricted to organisms with small genomes. Expensive and difficult to build phylogenies with large genome sequences, though this is rapidly changing.

- Transcriptome: RNA gives a snapshot of an enriched subset of the genome mRNA is then copied to complimentary DNA (cDNA) and sequenced. Less stable than DNA.

- Targeted enrichment: We design short bait sequences that are similar to conserved genome regions. Data are difficult to combine across different studies that used different baits.

- RAD-seq: Sequence data from specific regions scattered across the genome. It uses intrinsic properties of the genome for enrichment, rath,er than user-designed baits

Building a phylogenetic tree

01

Assume a substitution model (Jukes-Cantor, HKY, Kimura)

02

Assume a clock (constant or variable molecular clock)

03

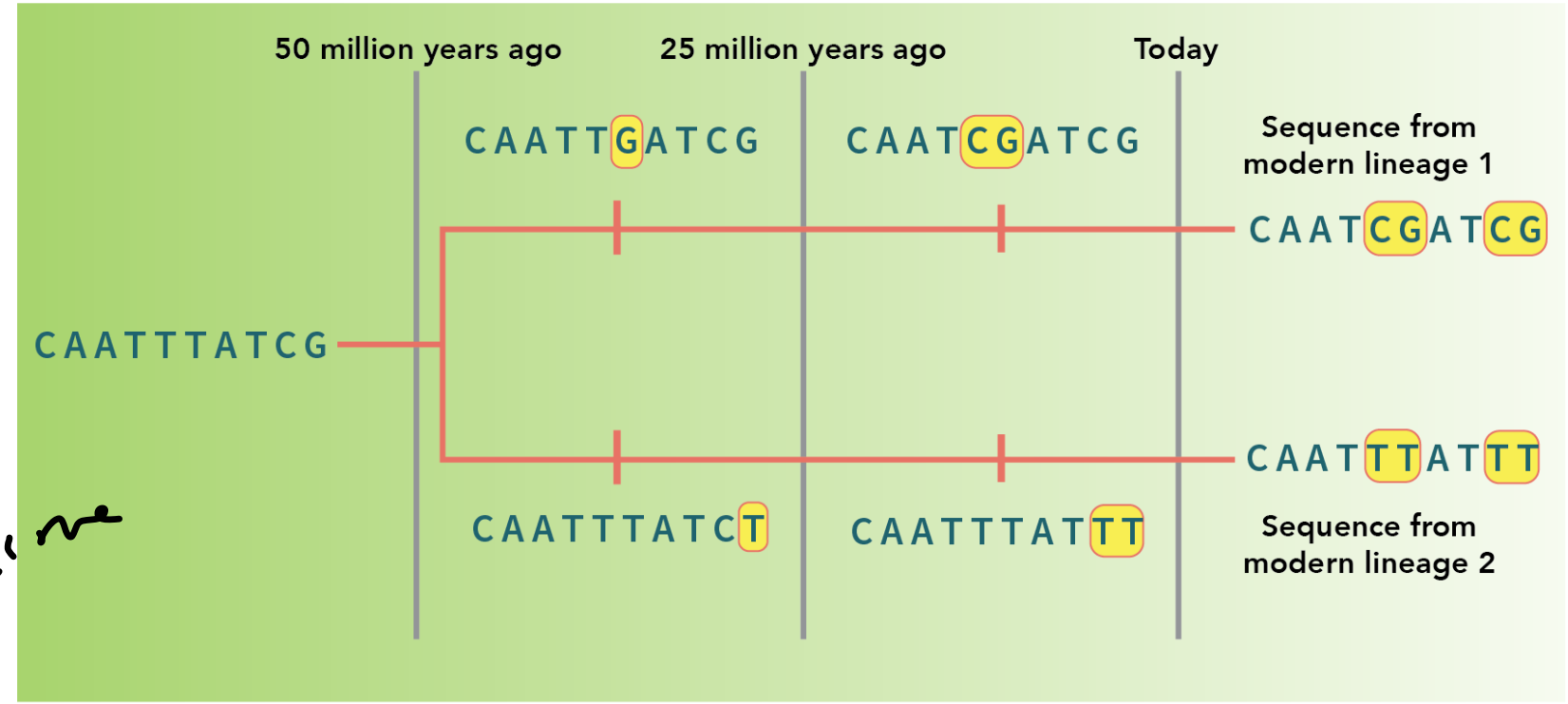
Give an algorithm to propose new tree shapes (cladograms)

Constant molecular clock

- 1 substitution rate per 25 million lyears

$$P(t) = e^{Qt}$$

Q ← rate
t ← time



The shared ancestor of the two modern lineages lived 50 million years ago.	After 25 million years, the two descendent lineages have diverged. Each has had a single base mutation, so that the lineages now differ by two bases.	After 50 million years, the two descendent lineages have diverged further by another base mutation.	The lineages now differ by four bases.

Rejecting the Global Molecular Clock

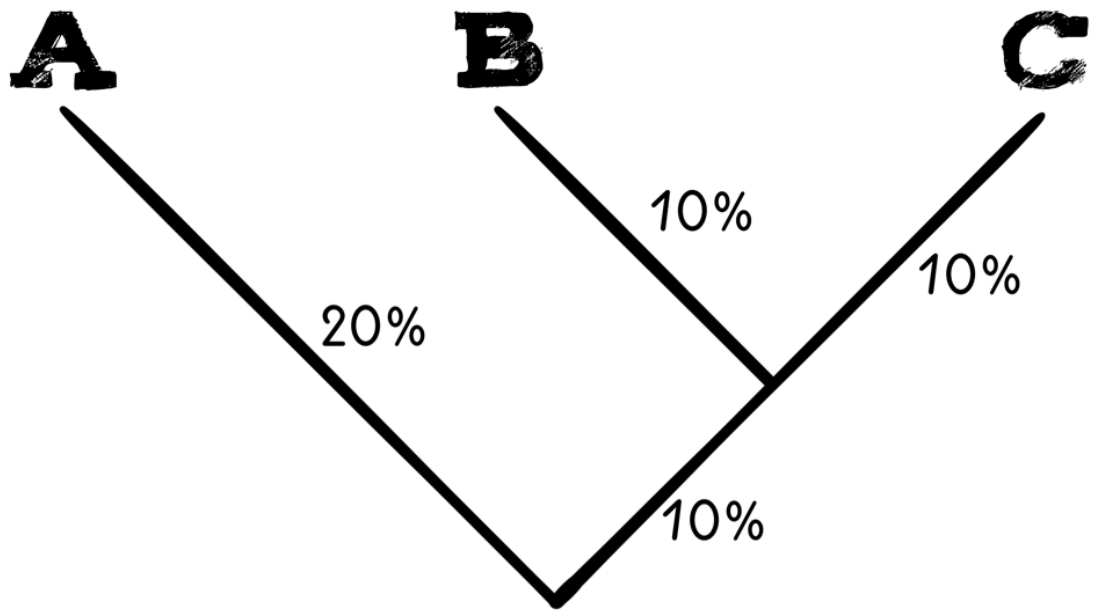
Rates of evolution vary across lineages and over time

Mutation rate

- metabolic rate
- generation time
- DNA repair

Fixation rate

- strengths/targes of selection
- population size

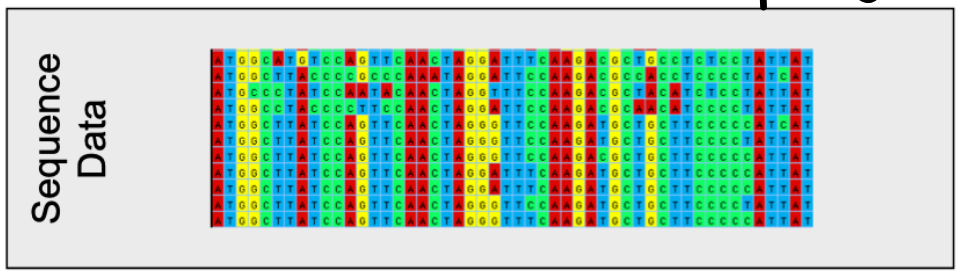


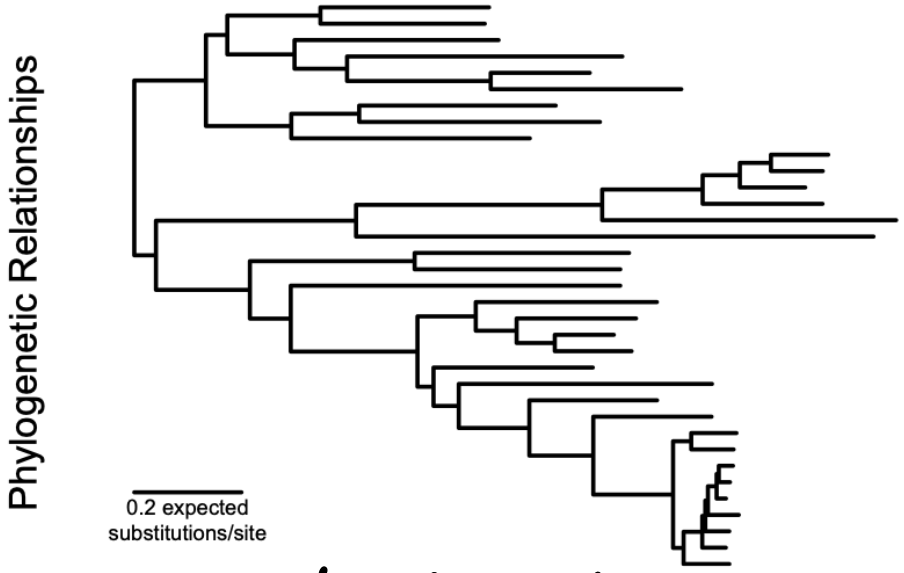
Unconstrained Analysis

Sequence data provide information about **branch lengths**

In units of the **expected # of substitutions per site**

branch length = rate × time





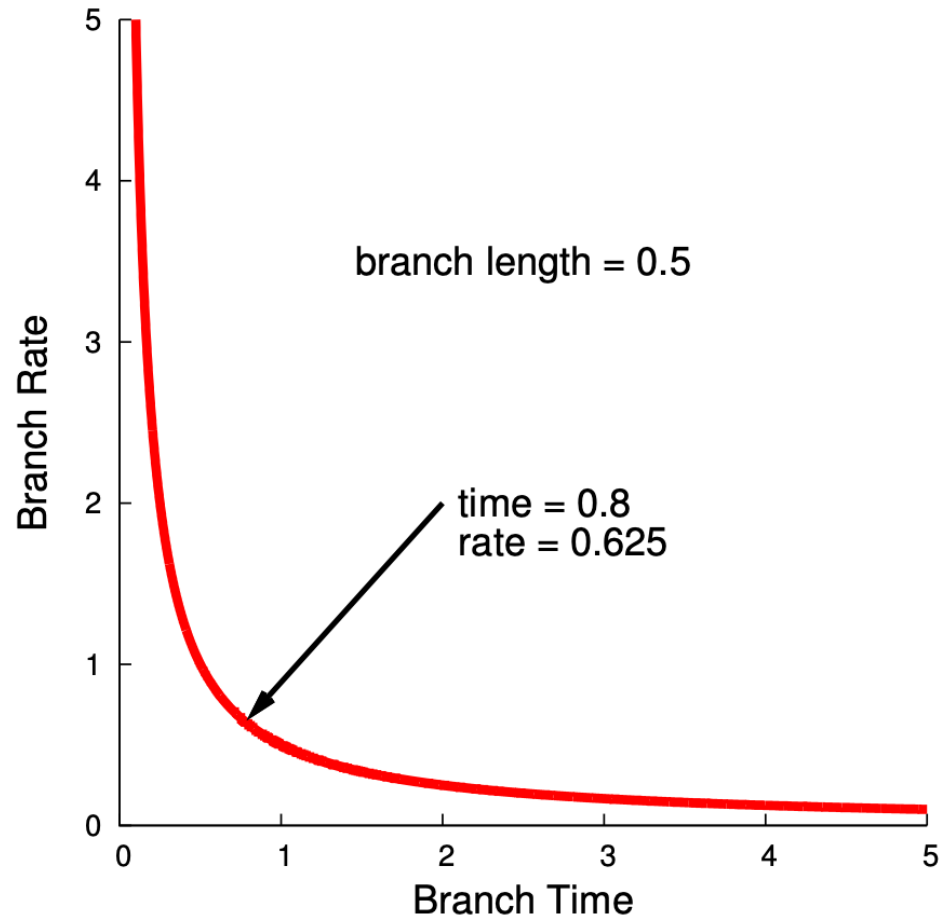
Tree dating
Model Model
absolute times x building your tree

Branch lenghts are # of expected substitutions

Estimating Rate & Time

The sequence data provide information about branch length

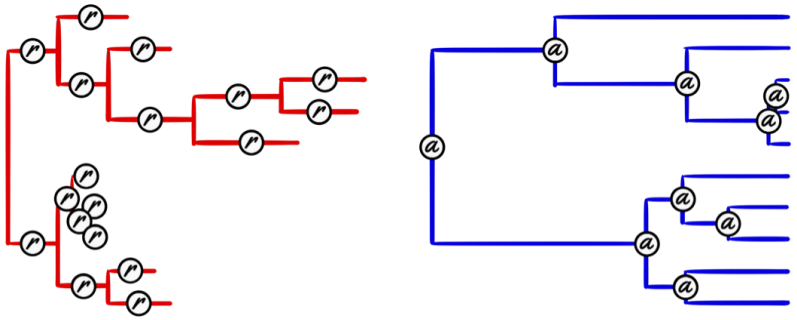
for any possible rate, there's a time that fits the branch length perfectly



(figure based on Thorne & Kishino 2005)

Phylogenetic tree dating

Bayesian Divergence Time Estimation

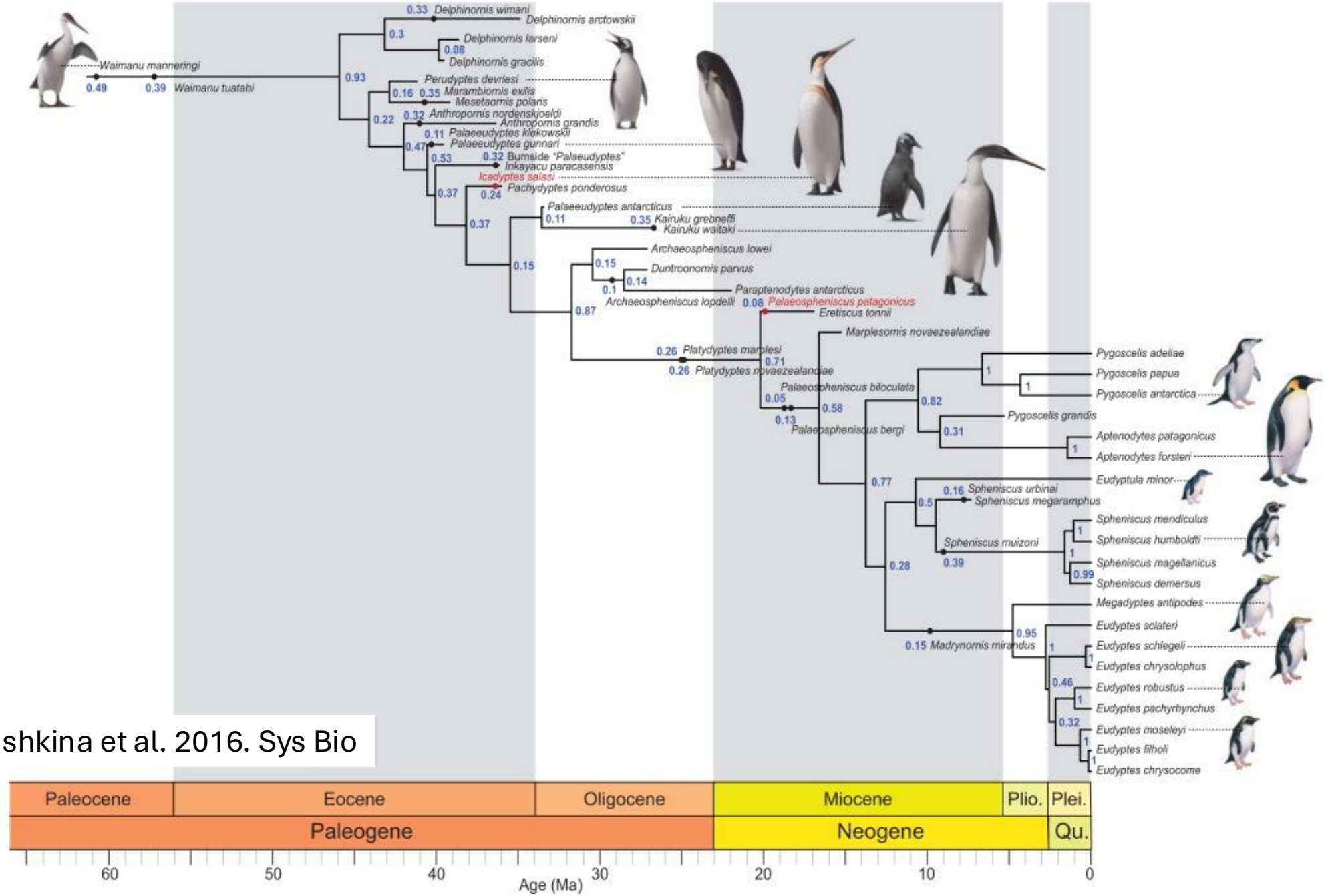


For each branch we estimate a rate r_i and for each node we estimate an age a_i

$$\boldsymbol{R} = (r_1, r_2, r_3, \ldots, r_{2N-2})$$

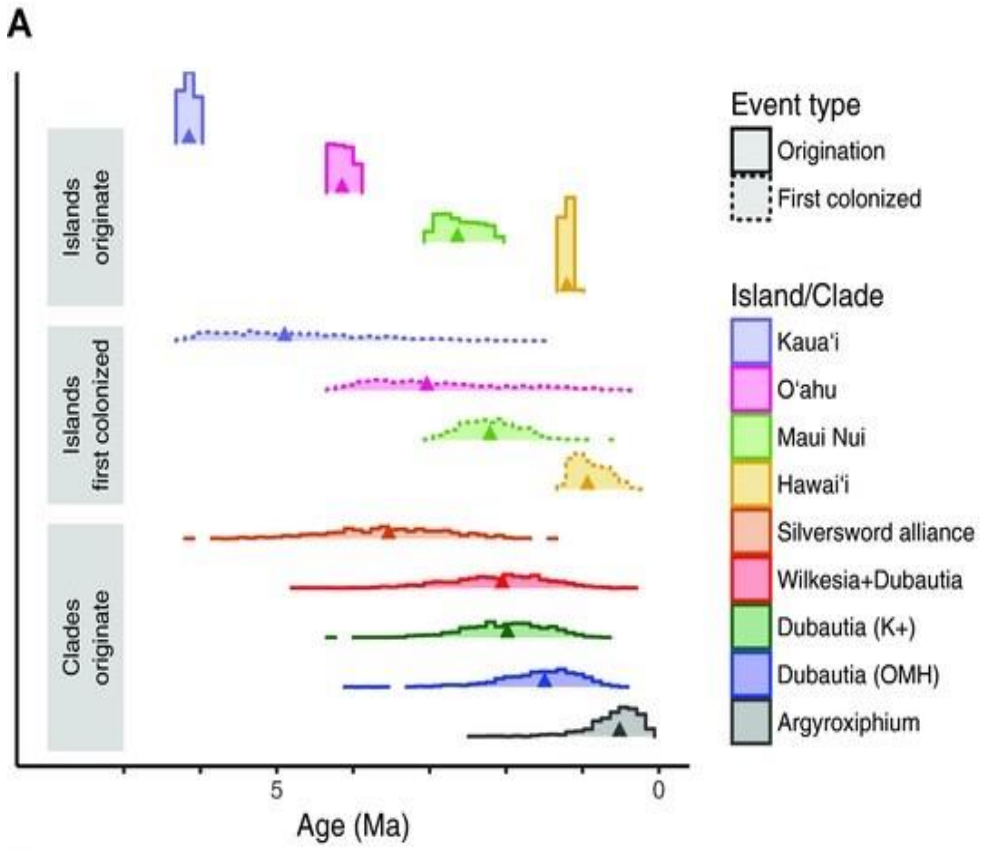
$$\boldsymbol{A} = (a_1, a_2, a_3, \ldots, a_{N-1})$$

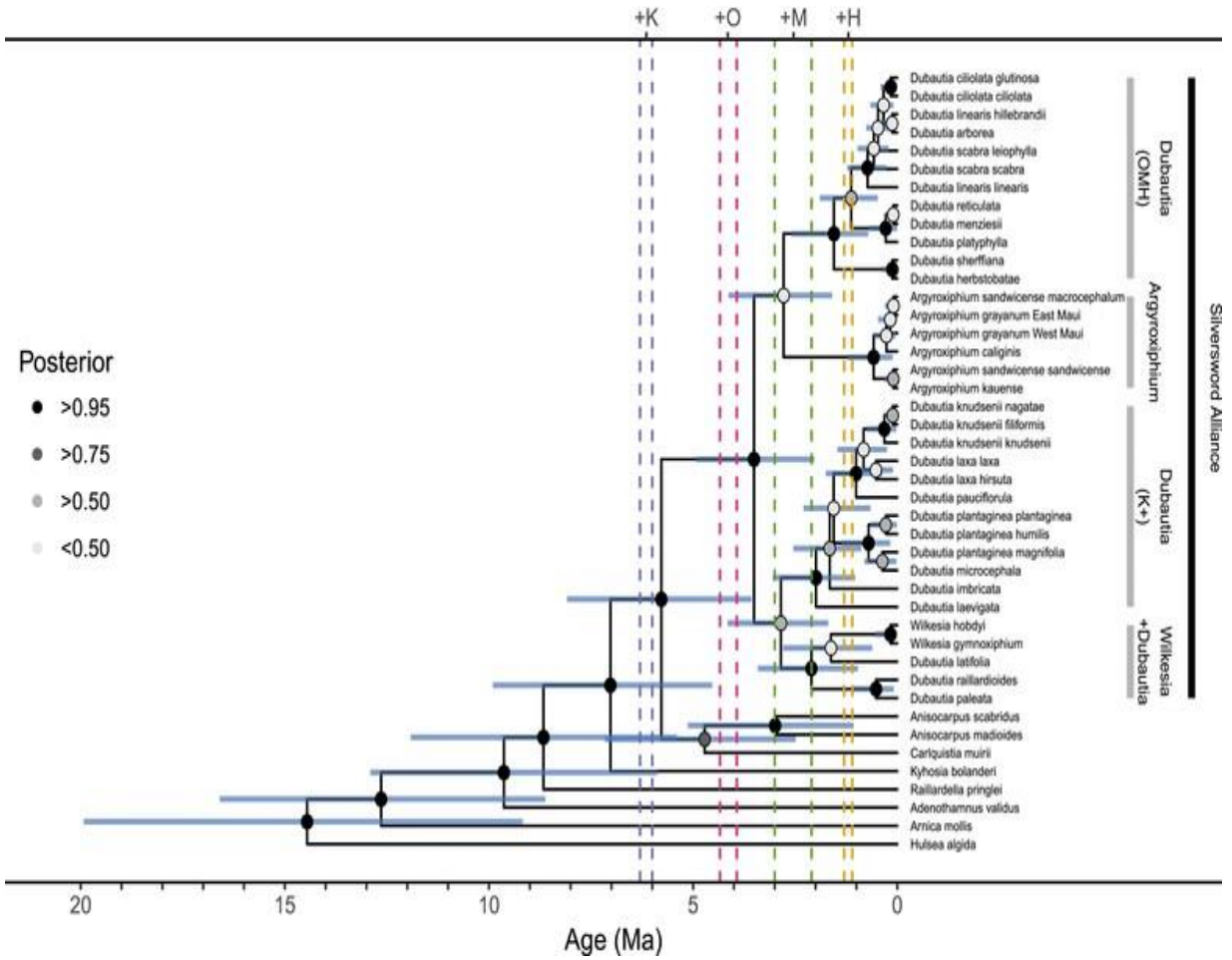
- Proposes as new model called birth-and-death model that has speciation rates, extinction rates but also absolute times for some nodes using fossils, or biogeography.



Gavryushkina et al. 2016. Sys Bio

- Island orogenesis / absolute time





Landis et al. 2018. *Sys Bio*