Using models to analyze phylogenetic comparative data

Matt Pennell, University of Idaho

A confession

I am not a systematist
and I don't build phylogenetic trees

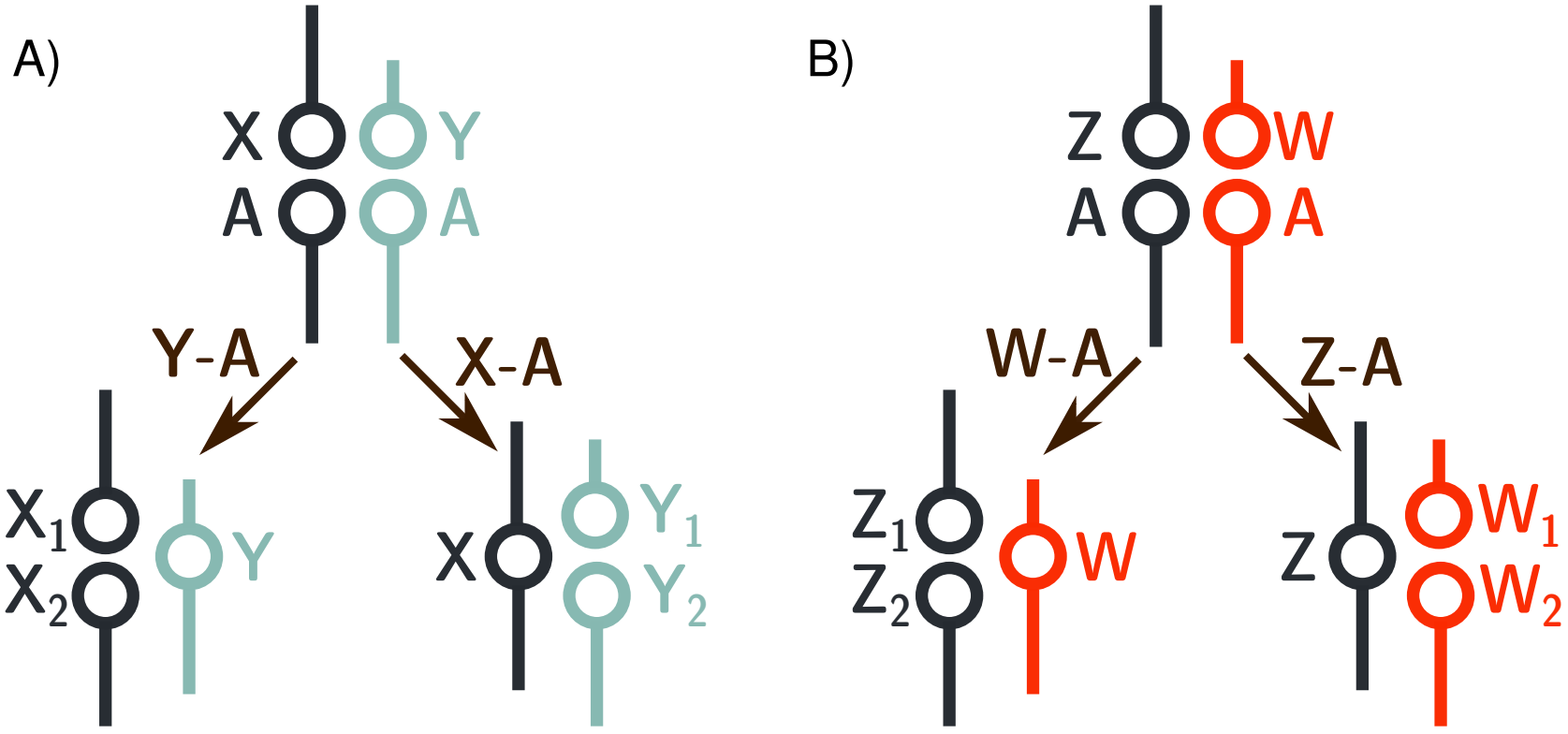
The questions that drive my research

- What are the dynamics of biodiversity through deep time?

- What evolutionary processes have driven these dynamics?

How can we use phylogenetic trees to learn about macroevolution?

Example: Evolution of sex chromosme fusions

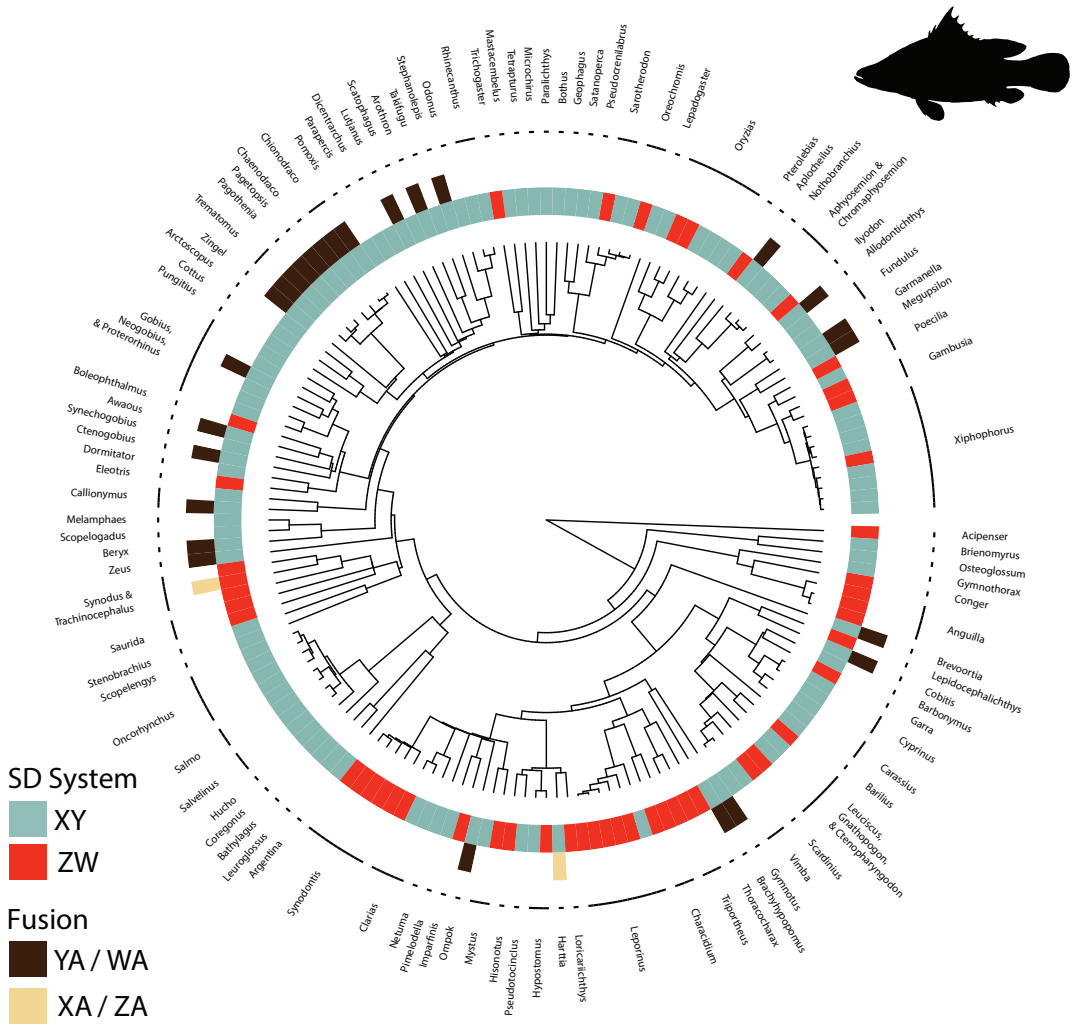


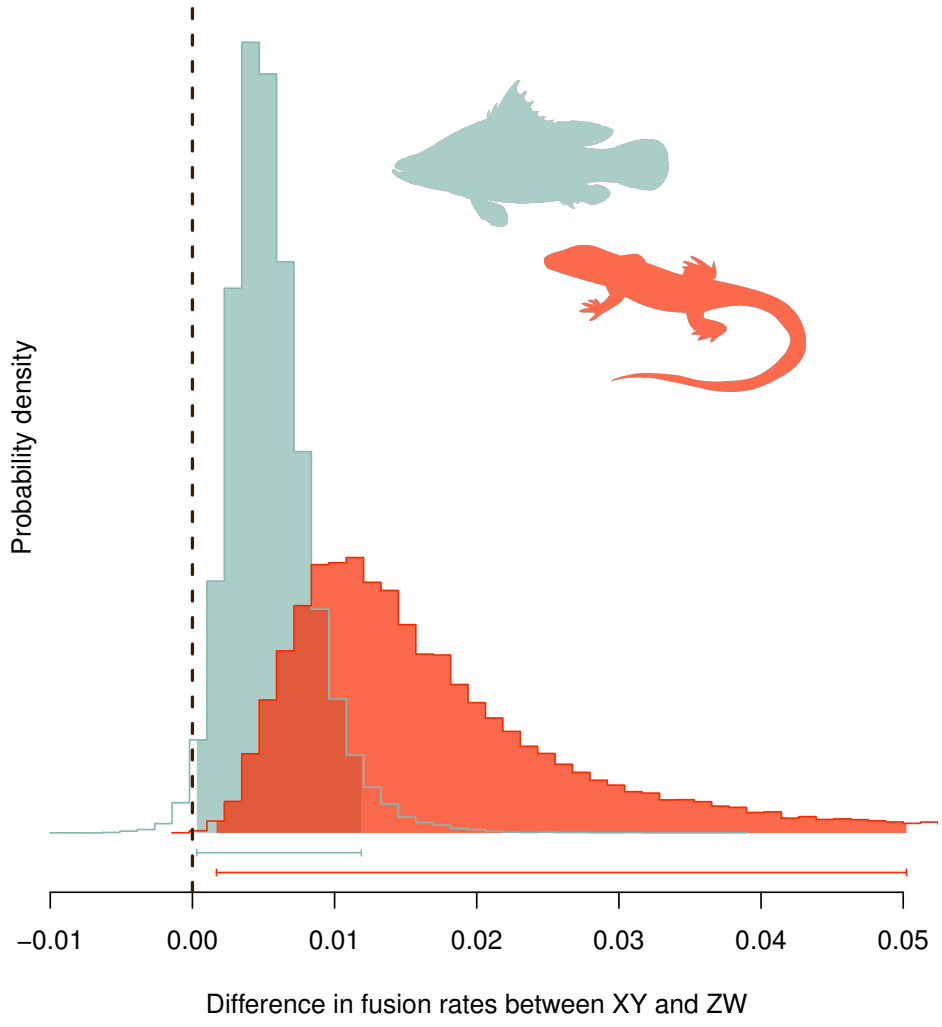
Example: Evolution of sex chromosome fusions

Do Y,X,W, and Z chromosomes fuse at different rates?

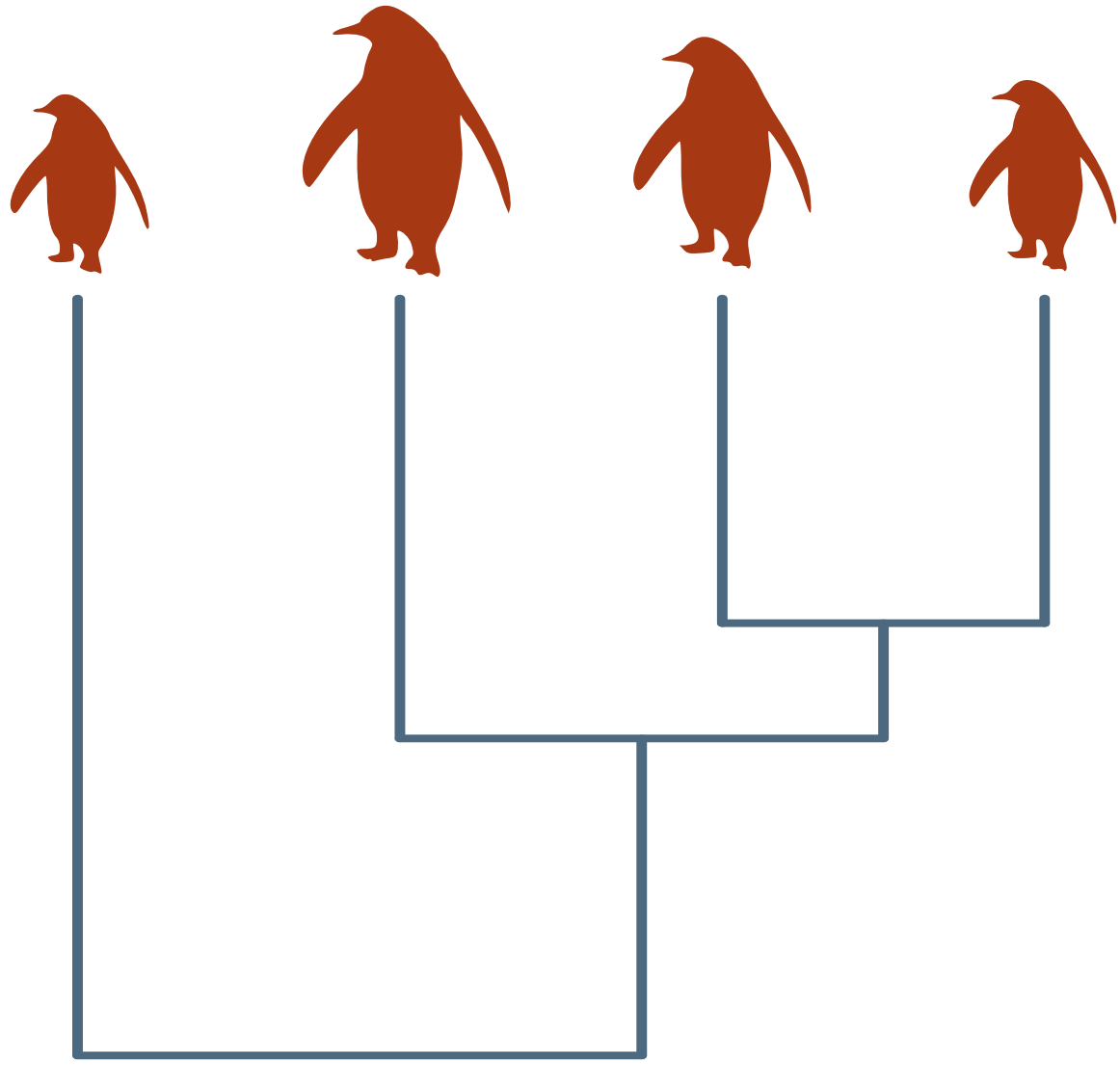
If so, this can provide clues about what processes are driving fusions and their maintenance

Example: Evolution of sex chromosme fusions



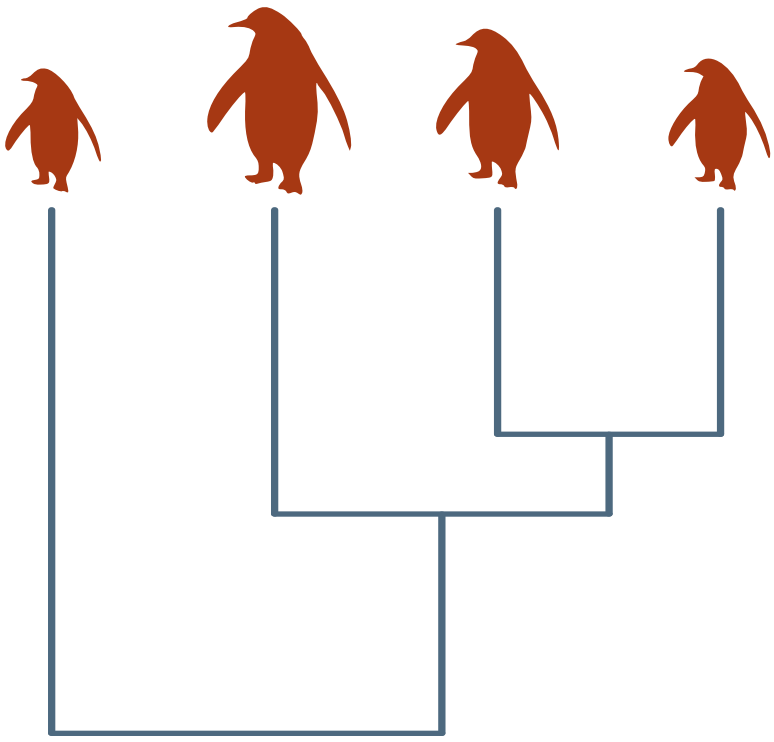

Probability density
−0.01
0.00
0.01
0.02
0.03
0.04
0.05
Difference in fusion rates between XY and ZW

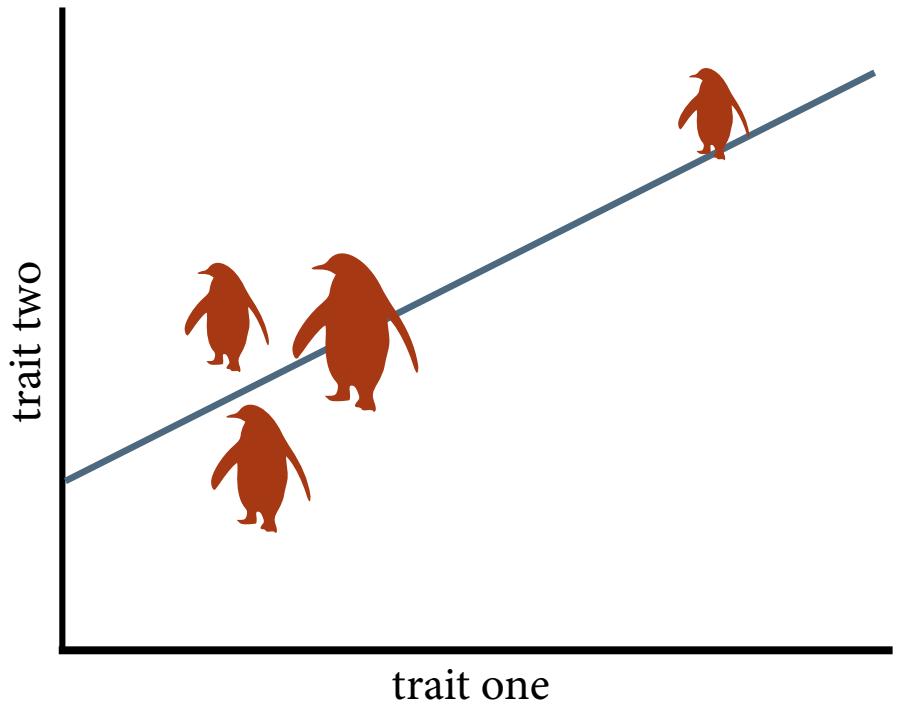
Testing evolutionary hypotheses using phylogenetic trees



Two major types of questions

Are two traits evolutionarily correlated?





Testing adaptive hypotheses

Testing for relationships between traits can provide evidence for adaptation

Especially useful when there is little variation within species

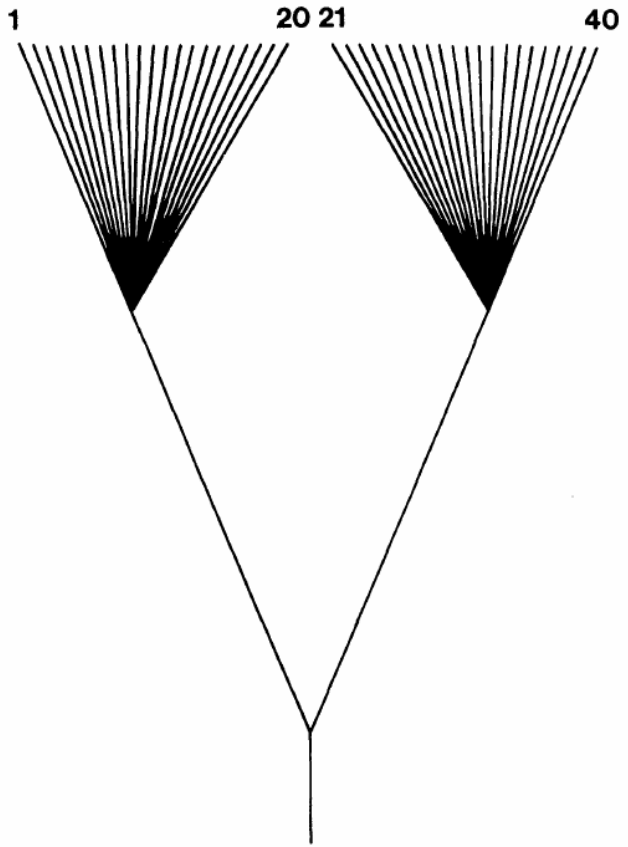
Allows us to assess the generalities of patterns

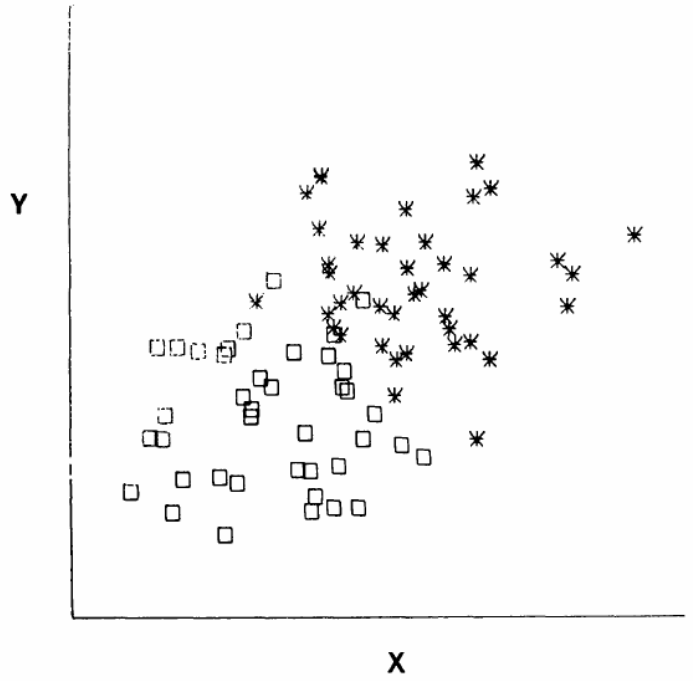
Testing adaptive hypotheses

Species share many traits and trait combinations

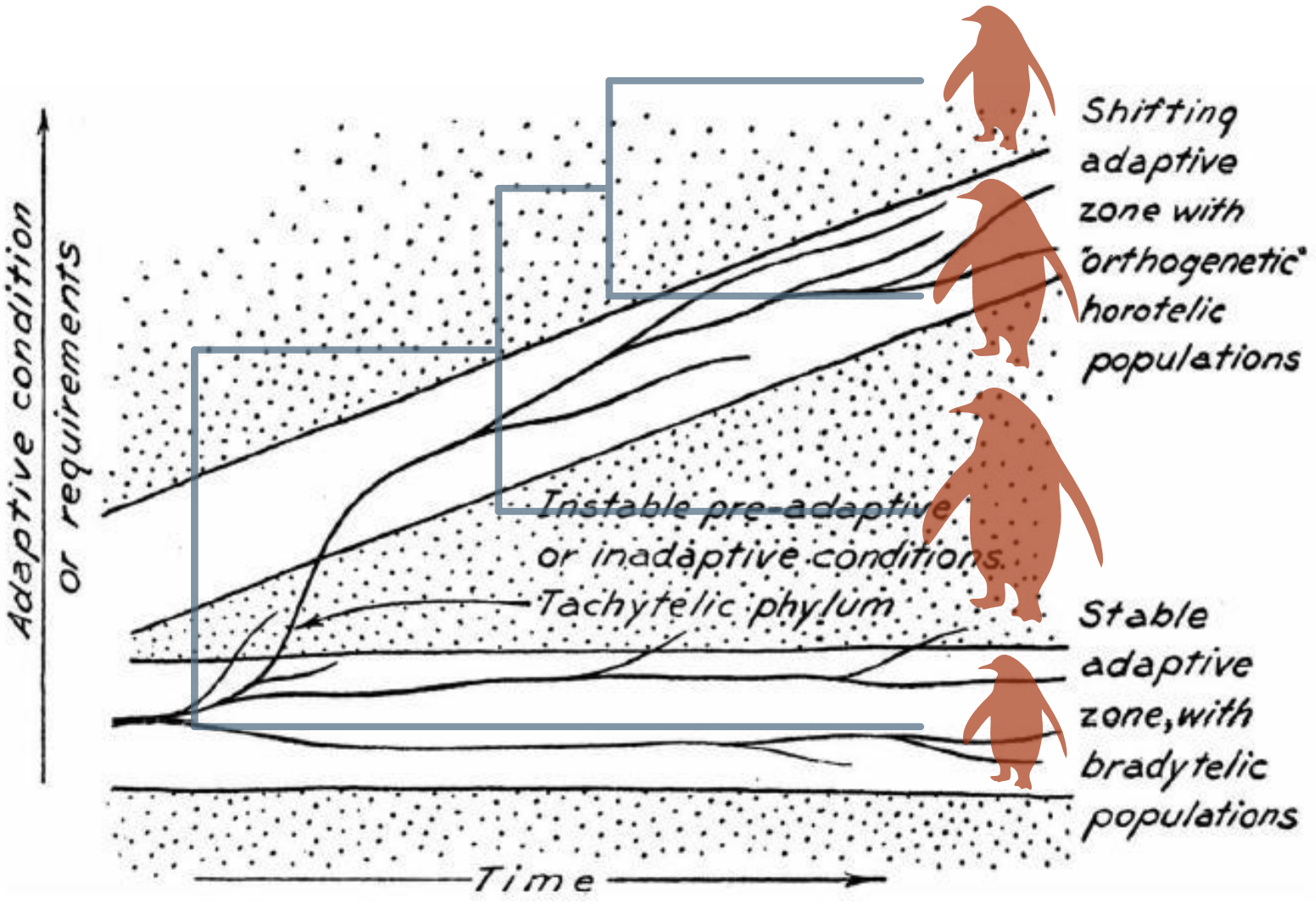
And are therefore not independent data points

Testing adaptive hypotheses





What is the tempo and mode of trait evolution?



Examples of questions we may be interested in:

- What is the general pattern of trait evolution in a group?
- Do different clades show different patterns?
- If so, what is driving these differences?
- What can this tell us the relative importance of different evolutionary processes?

Differences between phylogenetic estimation and comparative studies

Phylogenetic estimation

- Use many traits (DNA)
- Tree (topology and branch lengths) is estimated
- Usually only interested in some of the parameters

Trait evolution

- Use few traits (phenotypes)
- Tree is usually assumed to be known
- Interested in the values of all parameters

In order to test hypotheses we need a statistical model of trait evolution

And branch lengths in units of time!

Continuous traits

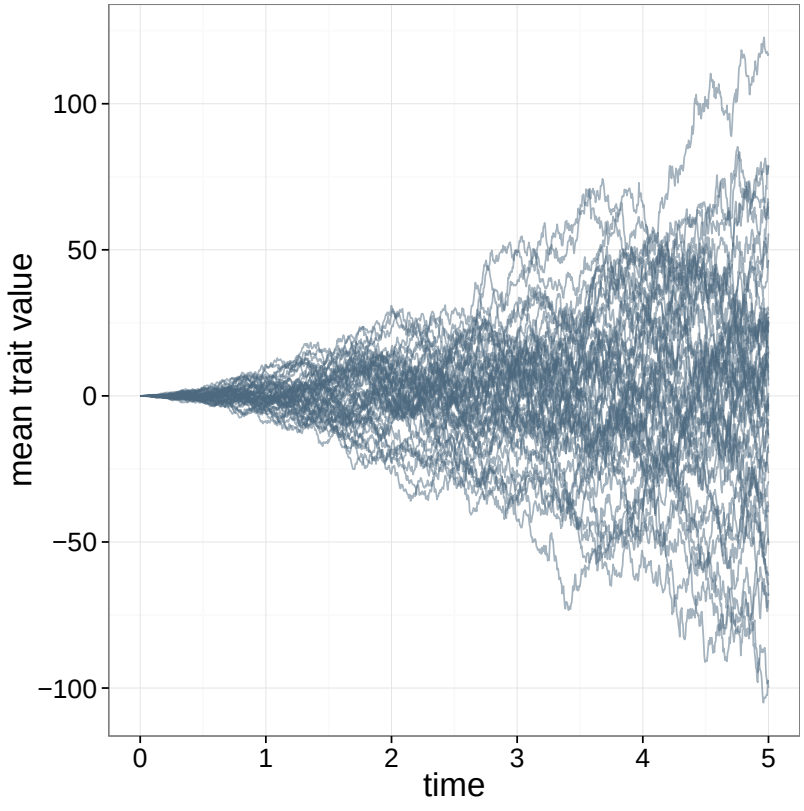
- Phenotypes vary along a continuum (can take any value)
- Usually follow a normal distribution
- E.g.: Body mass of a mammal, height of a plant, etc.

Models for continuous traits: Brownian motion (BM)

Trait evolves via a "random walk" (goes up and down with equal probability)

At the end of the process, on average the trait will be at the starting point

The variance will increase with time

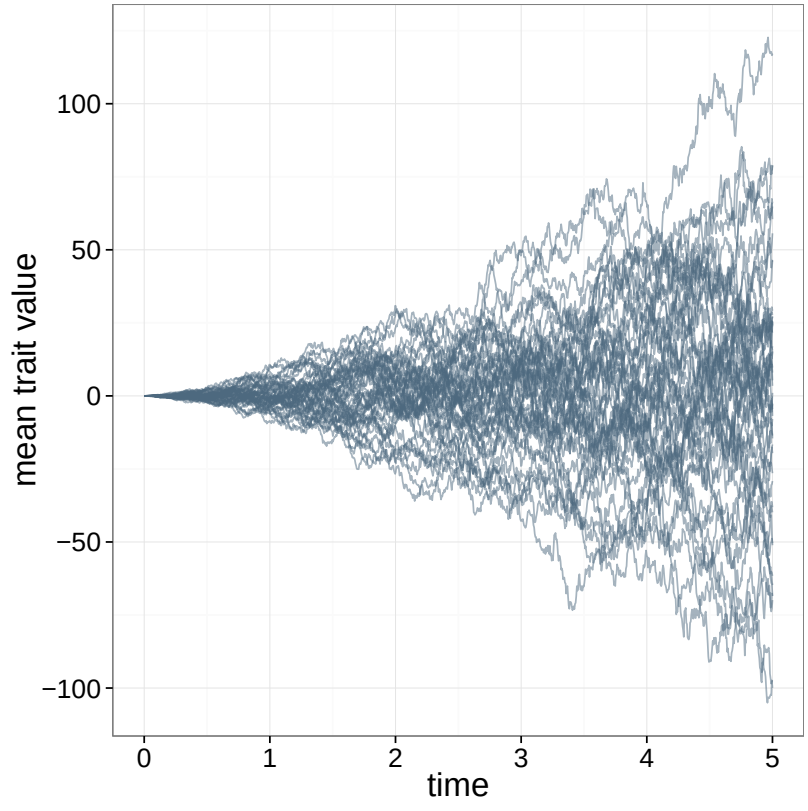


Models for continuous traits: Brownian motion (BM)

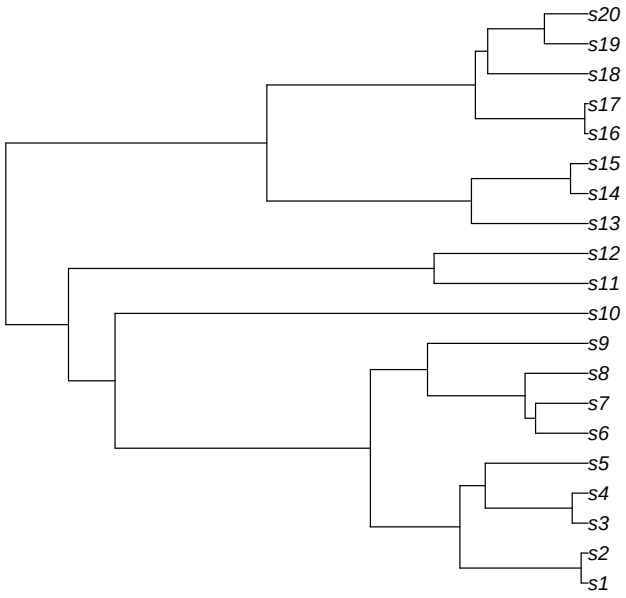
$$\Delta\bar{z} = \sigma dW$$

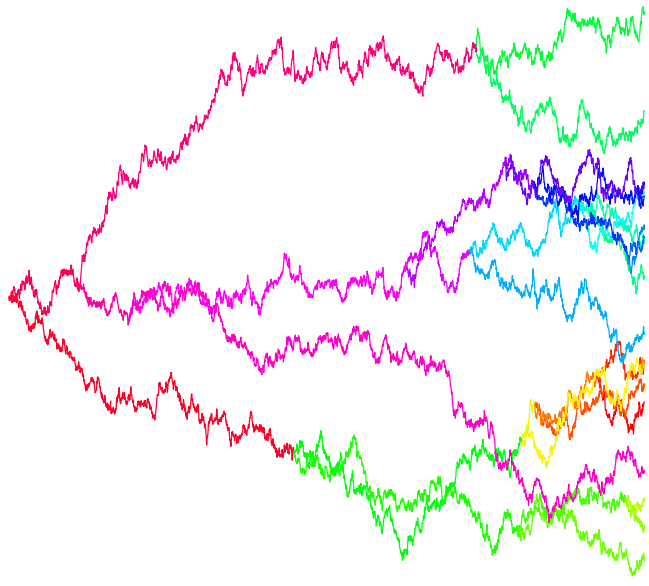
$$\mathrm{E}[\bar{z}(t)] = z(0)$$

$$\mathrm{Var}[\bar{z}(t)] = \sigma^2 t$$



Models for continuous traits: Brownian motion (BM)



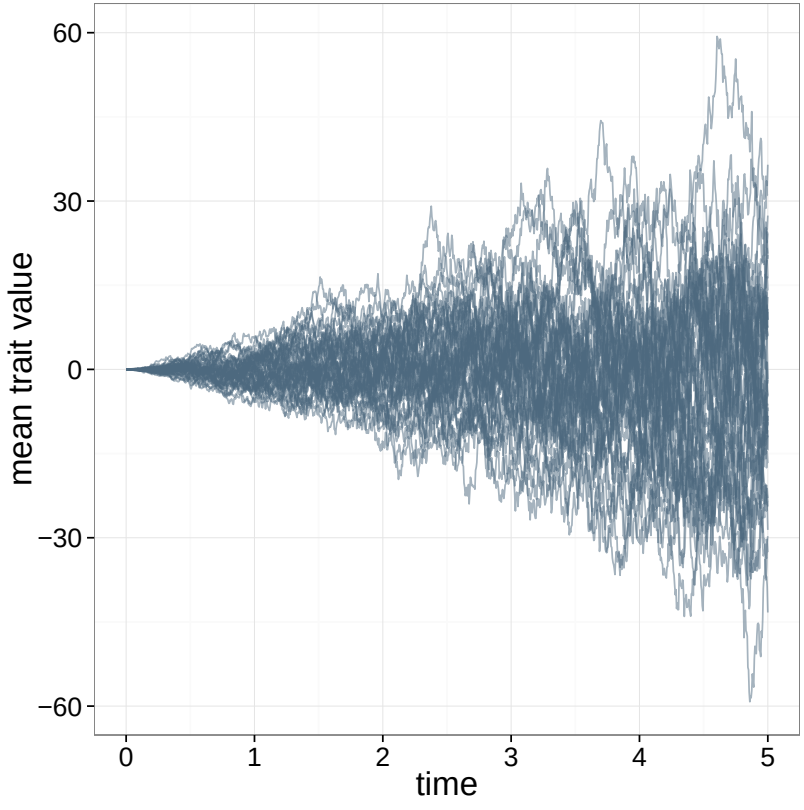


Models for continuous traits: Ornstein-Uhlenbeck (OU)

Trait evolves by a random walk

But is pulled towards an "optimum" value

The further a trait moves from the optimum, the stronger the pull towards the optimum

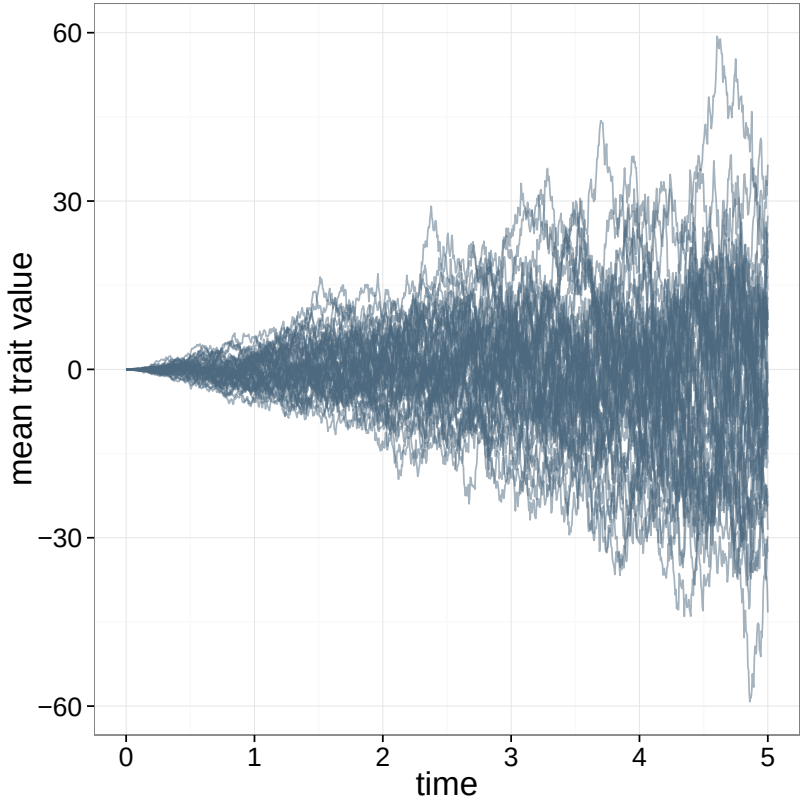


Models for continuous traits: Ornstein-Uhlenbeck (OU)

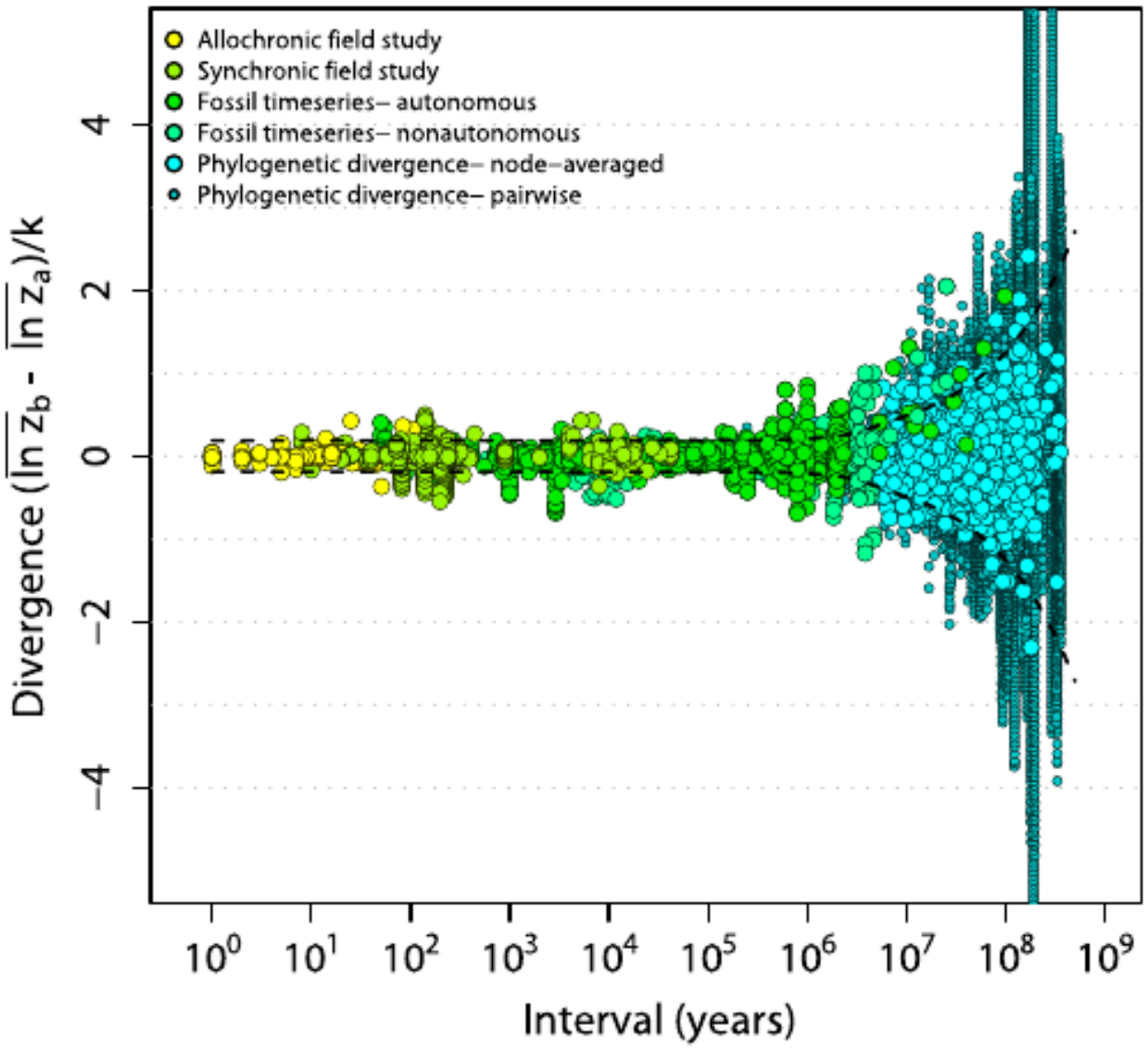
$$\Delta\bar{z} = -\alpha(\bar{z} - \theta) + \sigma dW$$

$$\mathrm{E}[\bar{z}(t)] = z(0)$$

$$\mathrm{Var}[\bar{z}(t)] = \frac{\sigma^2}{2\alpha}(1 - e^{-2\alpha t})$$



Beyond phylogenies



Fit evolutionary model

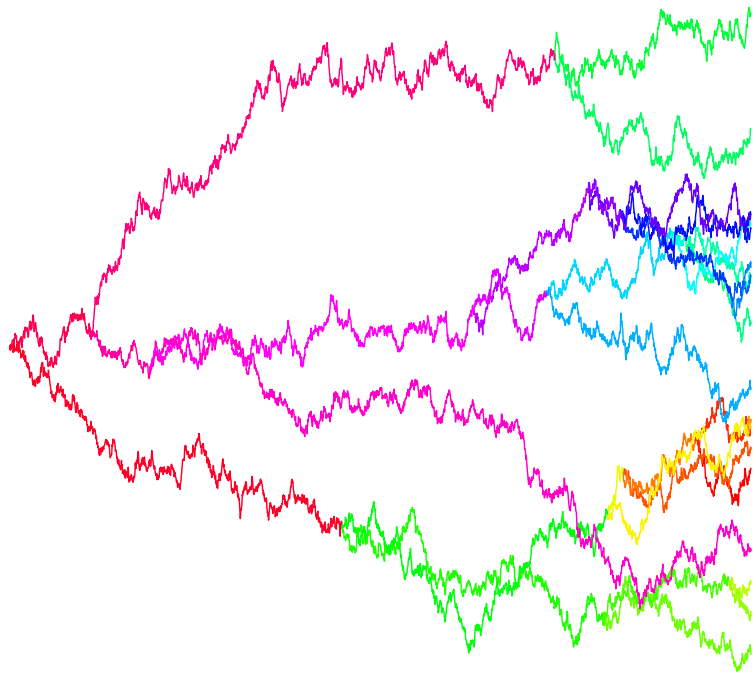
$$\mathcal{L} = -\tfrac{1}{2}\left[n \log(2\pi) + \log|\Sigma| + (\mathbf{Y} - \mu_y \mathbf{X})' \Sigma^{-1} (\mathbf{Y} - \mu_y \mathbf{X})\right]$$

$\mathbf{Y}$ is the observed trait data for the n species

μ_y is the mean of the observed data ($\mathbf{X}$ is just a column of 1's)

Σ is the expected variance–covariance matrix under the model

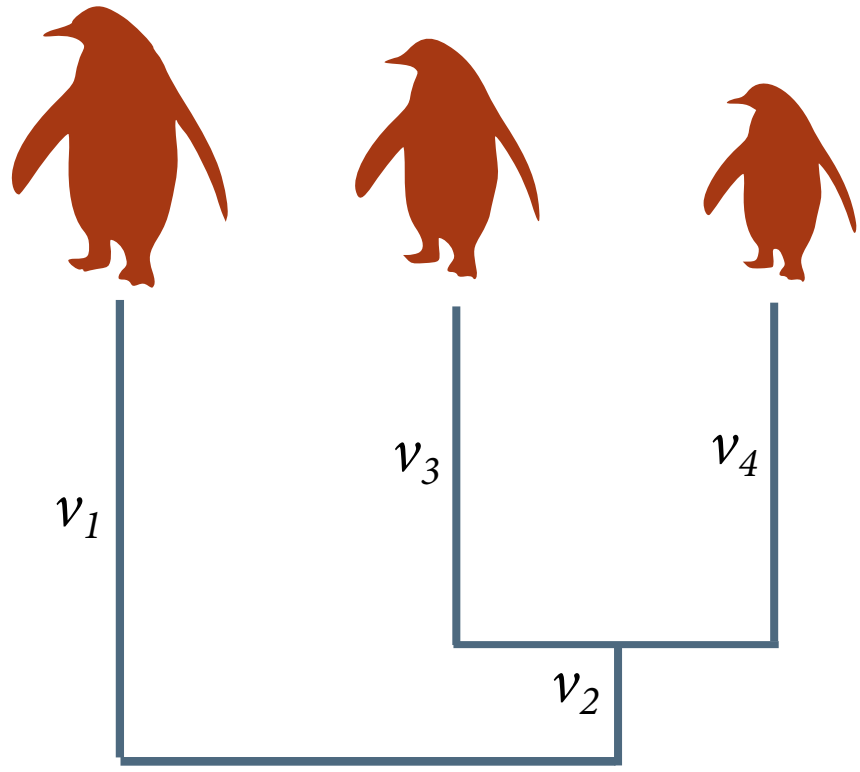
What does multivariate normal mean?



The Σ matrix

We can represent any phylogeny as a matrix **C**

$$\mathbf{C} = \begin{pmatrix} v_1 & 0 & 0 \\ 0 & v_2 + v_3 & v_2 \\ 0 & v_2 & v_2 + v_4 \end{pmatrix}$$



The Σ matrix

Models are transformations of the **C** matrix

Brownain motion:

$$\Sigma_{i,j} = \sigma^2 C_{i,j}$$

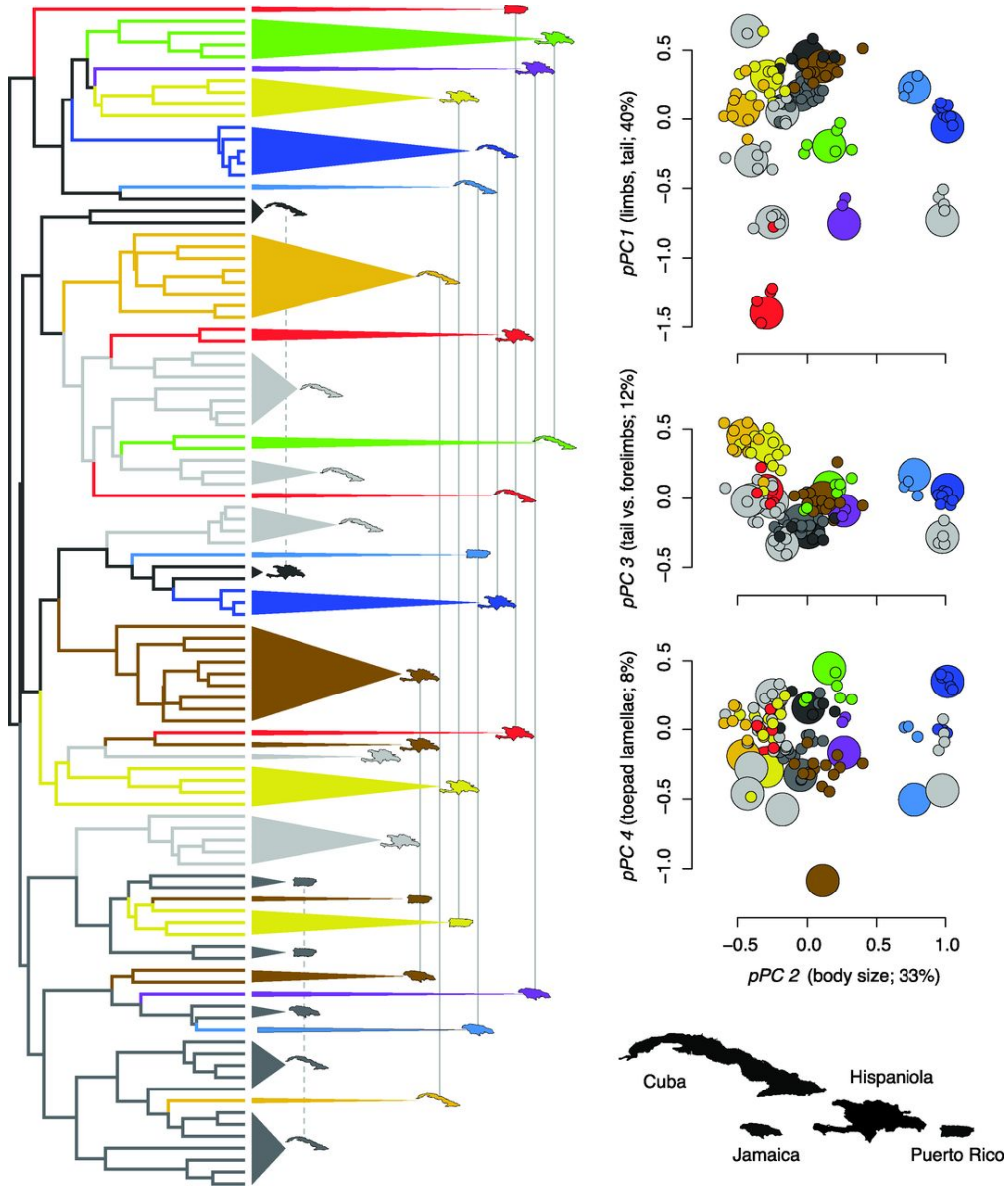
Ornstein-Uhlenbeck:

$$\Sigma_{i,j} = \frac{\sigma^2}{2\alpha}\left(1 - e^{-2\alpha t}\right)e^{-\alpha C_{i,j}}$$

Early Burst:

$$\Sigma_{i,j} = \sigma_0^2\left(\frac{e^{-rC_{i,j}} - 1}{r}\right)$$

Many extensions to these type of models

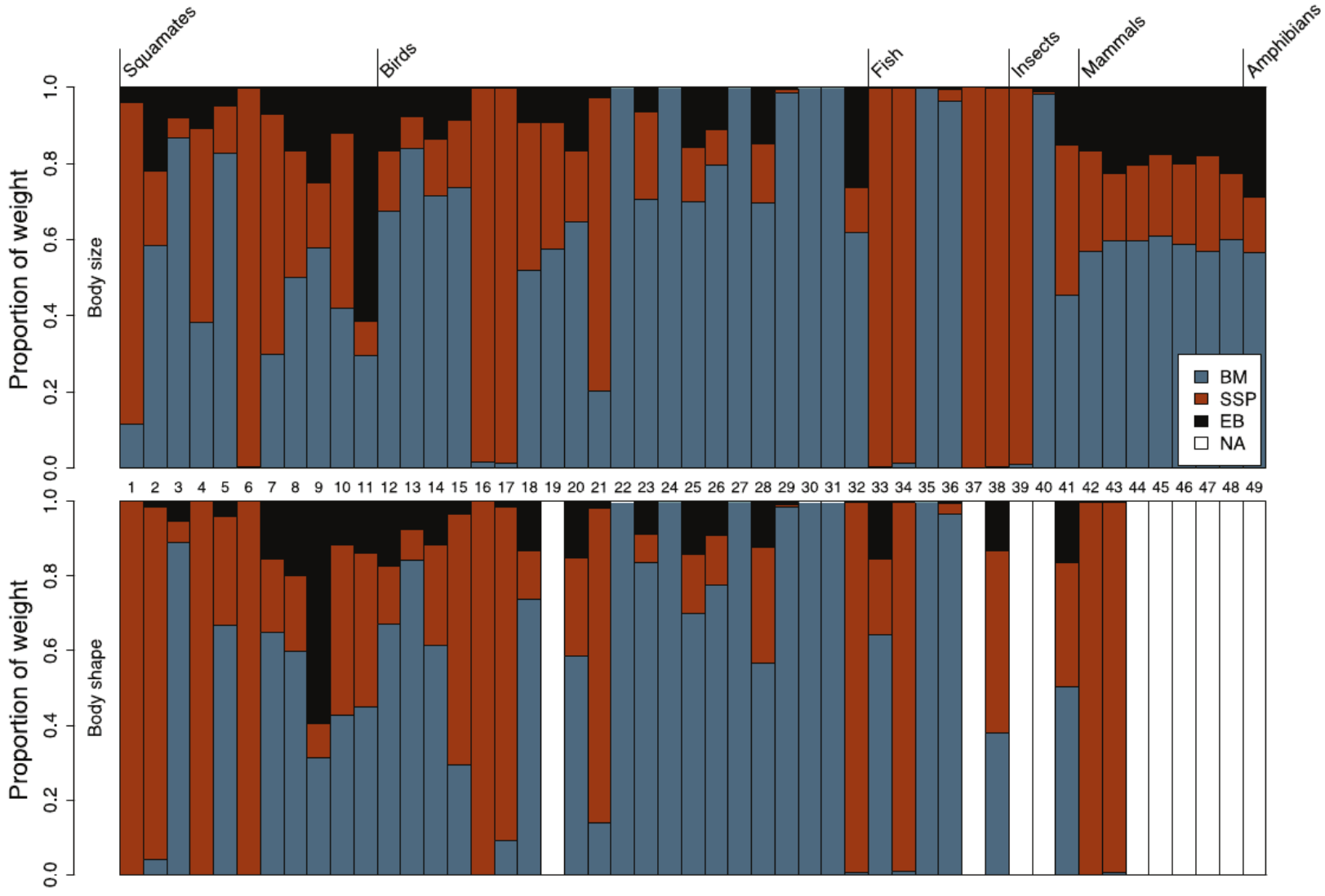


Mahler et al. 2013

Many extensions to these type of models

- Multi-rate BM
- Multi-optimum OU
- Mixed models
- Variation in rates through time
- Change concentrated at speciation events
- Combinations of the above

Can compare the relative fit of models



Harmon et al. 2010

The more challenging question is:
how should we interpret models of trait evolution

The Quantitative Genetics view

$$\Delta \bar{z} = \mathbf{G}\beta$$

- Phenotypes are controlled by an effectively infinite number of alleles of small effect
- Phenotypes are normally distributed

The Quantitative Genetics view

Macroevolutionary models literally represent microevolutionary hypotheses

E.g., Brownian motion

$$\begin{aligned}\Delta\bar{z} &= \sigma^2 t \\ &= \frac{V_A}{N_e} t \\ &= 2V_M t\end{aligned}$$

V_A Additive genetic variance

N_e Effective population size

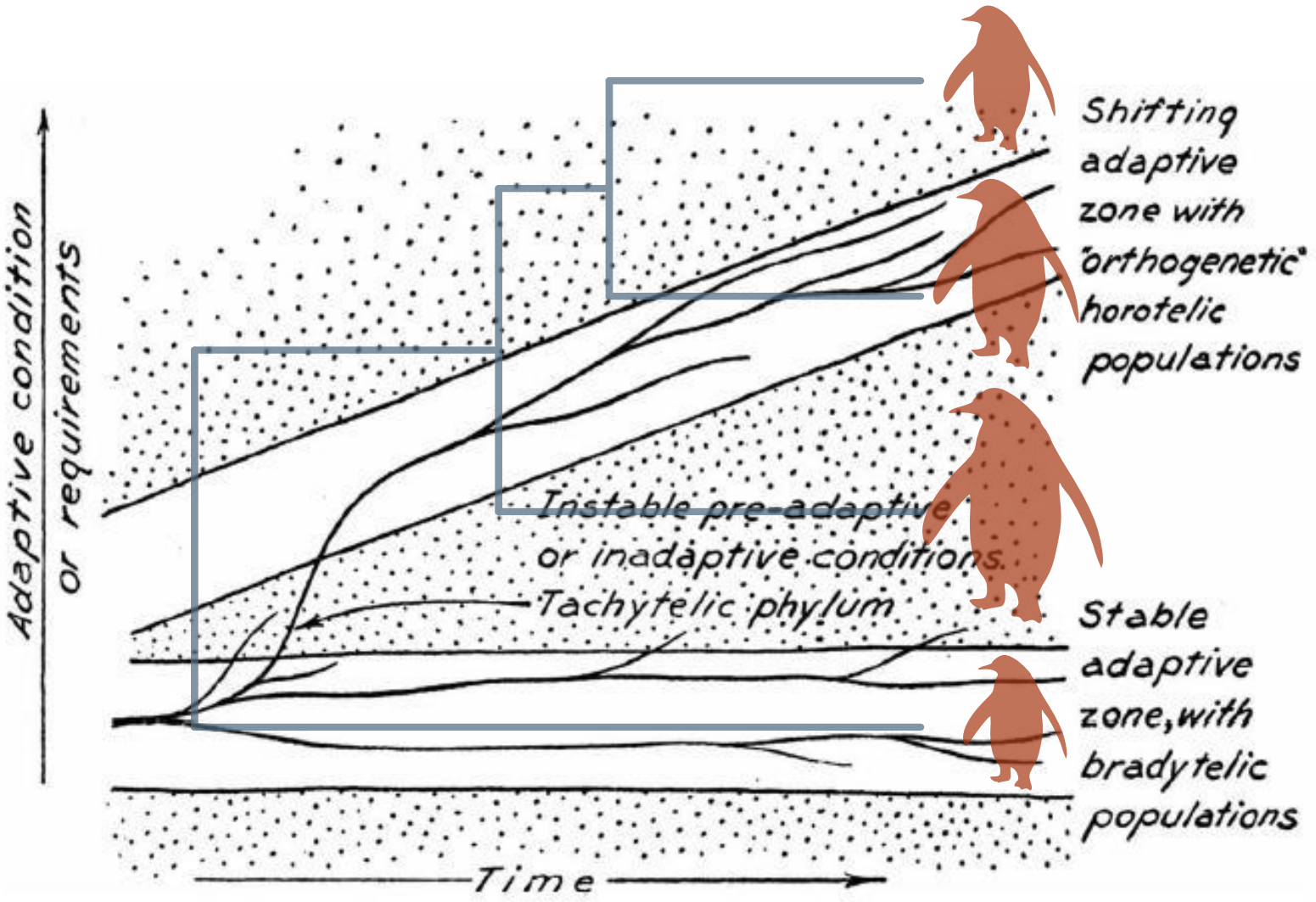
V_M Mutational variance

The purely statistical view

The models we use are just phenomenological constructs that describe general patterns

Like models of sequence evolution???

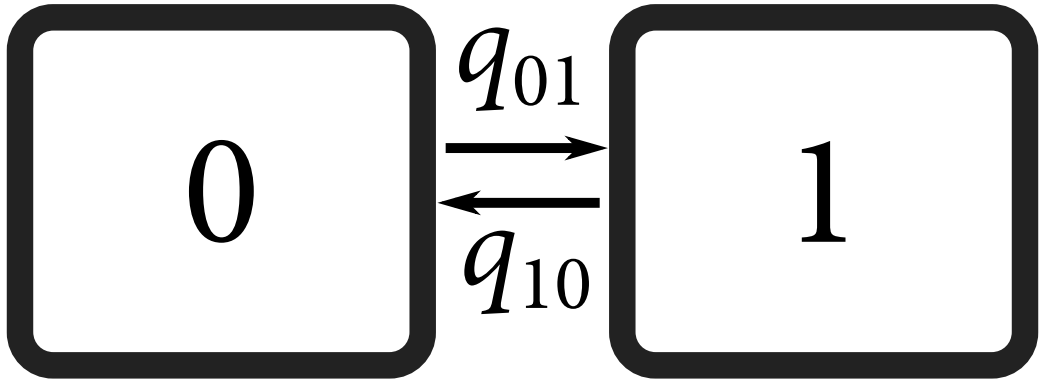
The macroevolutionary landscapes view



Discrete traits

- Only a fixed number of states are possible
- E.g.: floral aymmetry/symmetry, presence/absence of woody tissue, DNA

Discrete traits



Discrete traits

- Exactly the same likelihood calculations used for sequence data
- We can also use arbitrary **Q** matrices depending on the biology of the system
- **Remember:** When optimizing transition rates between states, we are assuming that we have the correct tree

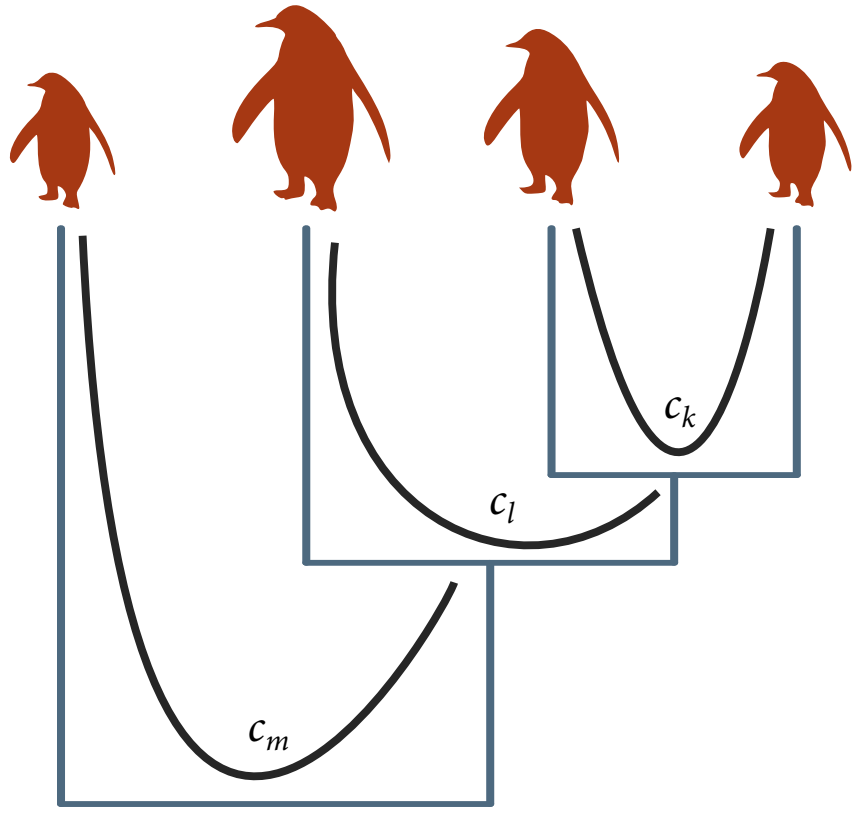
Independent contrasts

Instead of using species as data, consider changes (contrasts)

Assume a Brownian motion of evolution for the traits

Independent contrasts

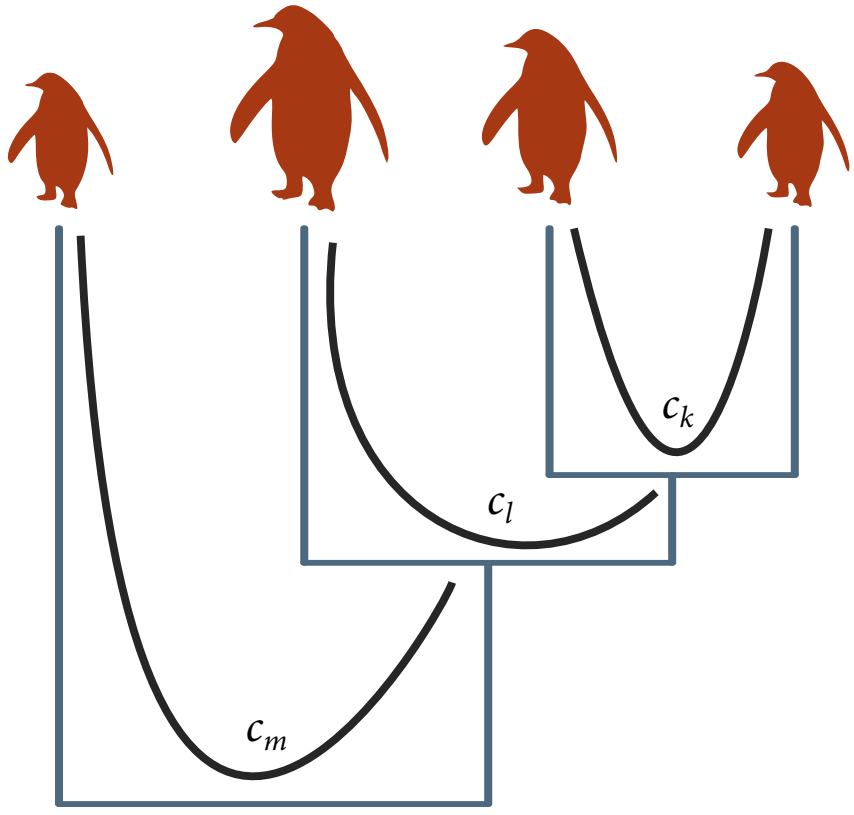
- Start from the tips
- For each node, calculate the average of the daughter species
- Standardize by the branch lengths; the longer the branch length, the more evolution we expect



Independent contrasts

$n - 1$ contrasts for n tips

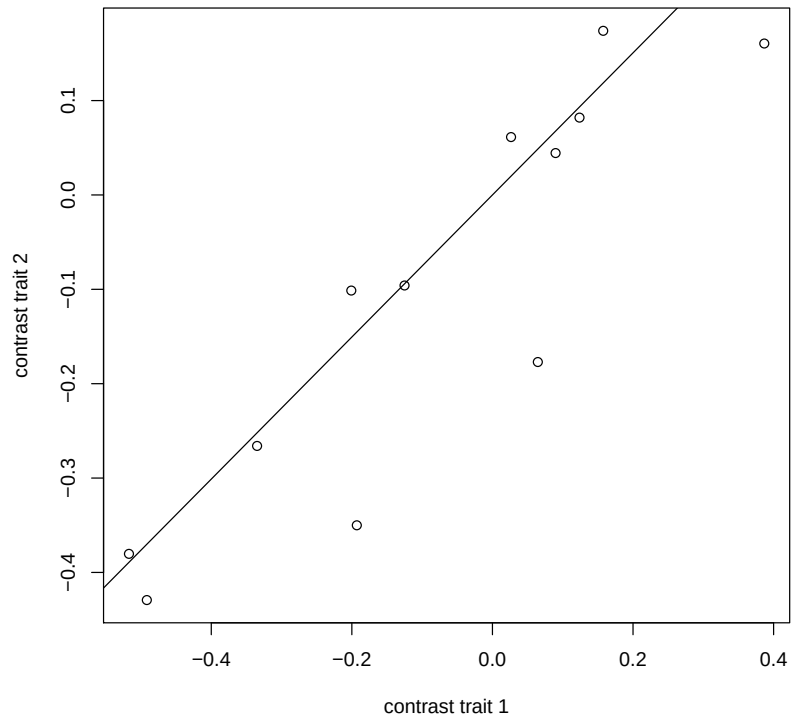
Under BM, contrasts will be Independent and Identically Distributed (I.I.D.) ~ $\mathcal{N}(0, \sigma)$



Independent contrasts

Compute contrasts at every node for *both* traits

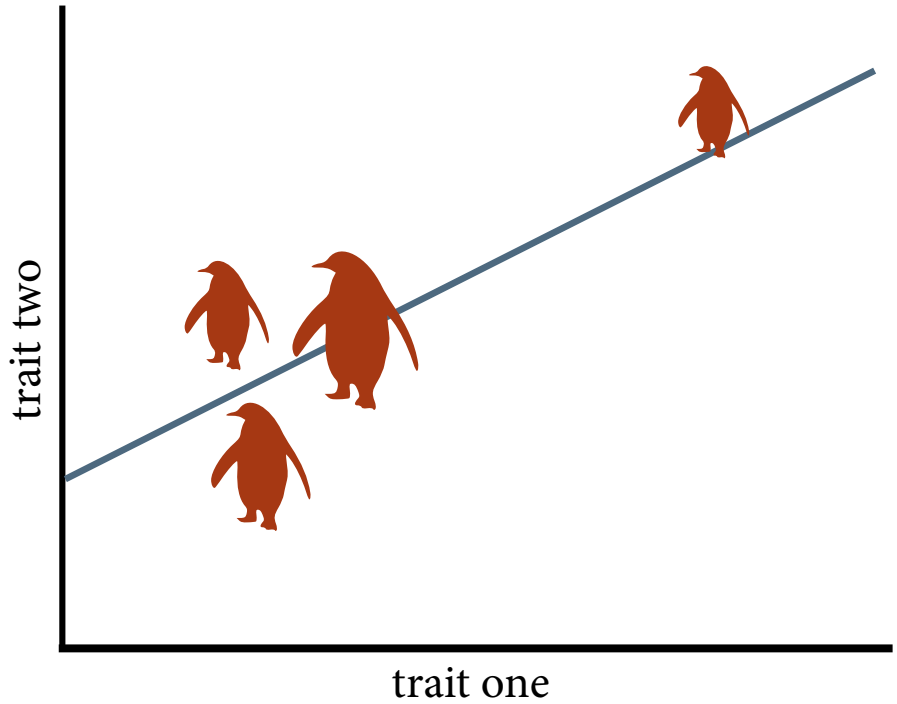
Fit a linear model between the contrasts, forcing the model through the origin (0,0)



Phylogenetic regression

$$Y = \beta_0 + \beta_1 X + \epsilon$$

$$\epsilon \sim \mathcal{N}(0, \Sigma)$$



Phylogenetic regression

$$Y = \beta_0 + \beta_1 X + \epsilon$$

$$\epsilon \sim \mathcal{N}(0, \Sigma)$$

Residuals are structured according to the phylogenetic model

Phylogenetic regression

$$Y = \beta_0 + \beta_1 X + \epsilon$$

$$\epsilon \sim \mathcal{N}(0, \Sigma)$$

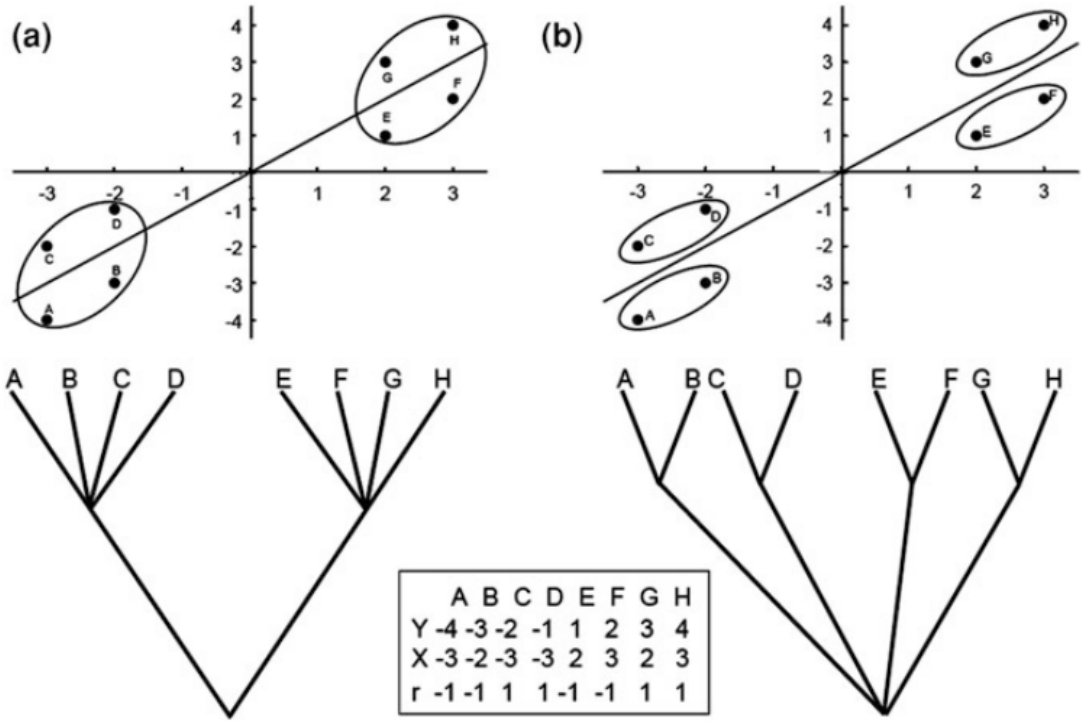
When Σ generated from BM model, phylogenetic regression is equivalent to Independent Contrasts method

Phylogenetic regression

The advantages of using phylogenetic regression over independent contrasts:

- Σ can be constructed from any model (not just Brownian motion)
- Because it is formulated as a standard regression model, we can use all the standard linear model tricks

Phylogenetic regression



Hansen 2014

Phylogenetic regression

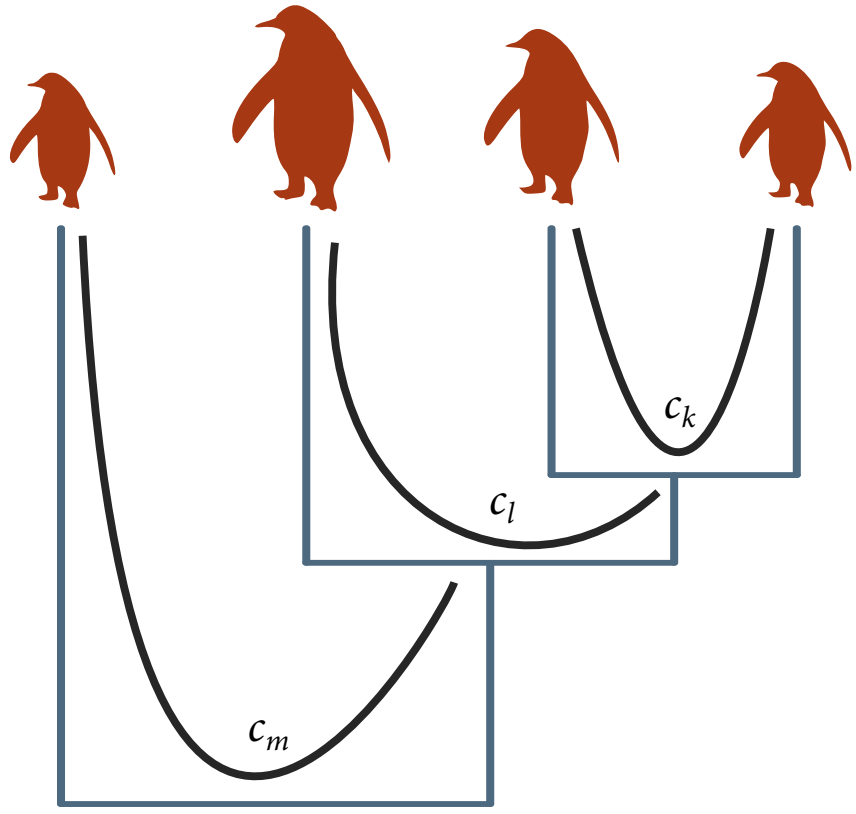
Common misconceptions:

- The traits X and Y must show phylogenetic signal
- The traits X and Y are assumed to be multivariate normal
- Phylogenetic regression removes the "effect of phylogeny"
- Phylogenetic regression partitions the variation into "phylogenetic" from "ecological" components

Independent contrasts

$n - 1$ contrasts for n tips

under BM, contrasts will be independent and identically distributed (i.i.d.) $\sim \mathcal{N}(0, \sigma)$



Independent contrasts

If: we assume that the residuals are distributed according to a Brownian motion model of trait evolution

Then: $Y_{PIC} = \beta_1 X_{PIC} + \epsilon_\sigma$ will be equivalent to the PGLS estimator $Y = \beta_0 + \beta_1 X + \epsilon_\Sigma$ (Blomberg et al. 2012 Sys Bio)

How do traits influence diversification?

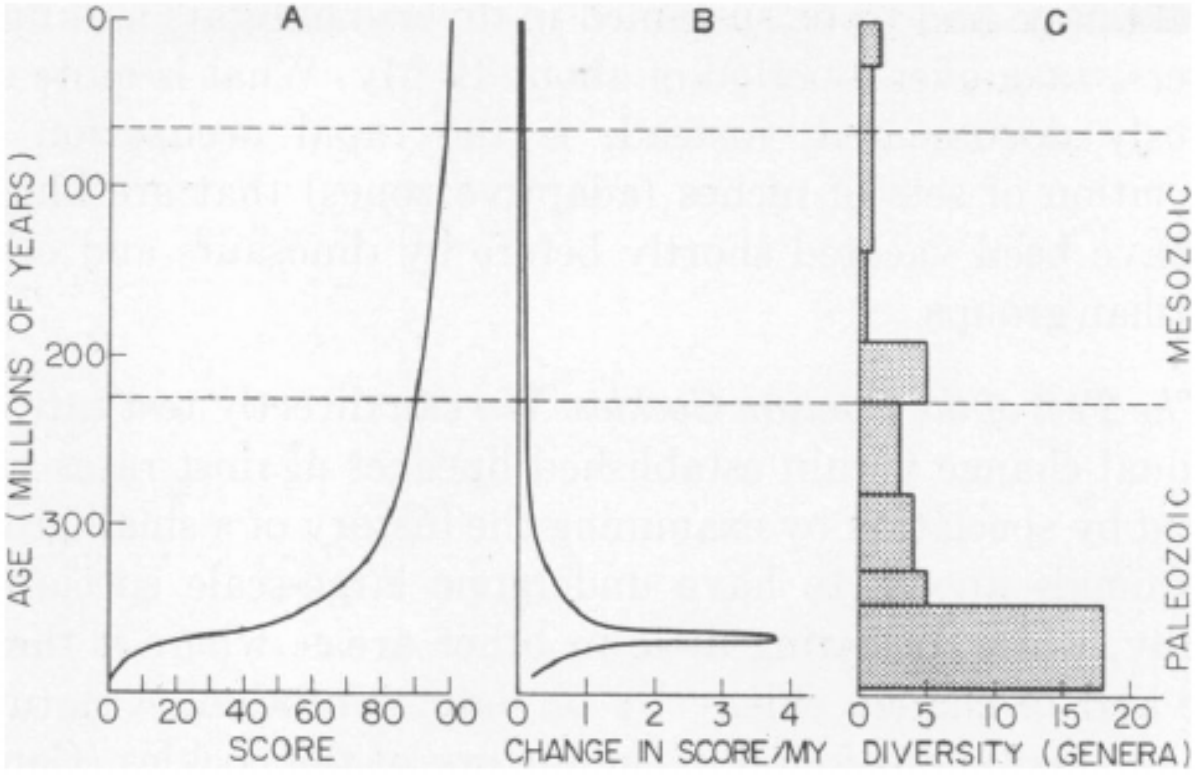
The ingredients of natural selection

- Variation with population
- Heritability
- Differential fitness

In principle, this can also apply to species

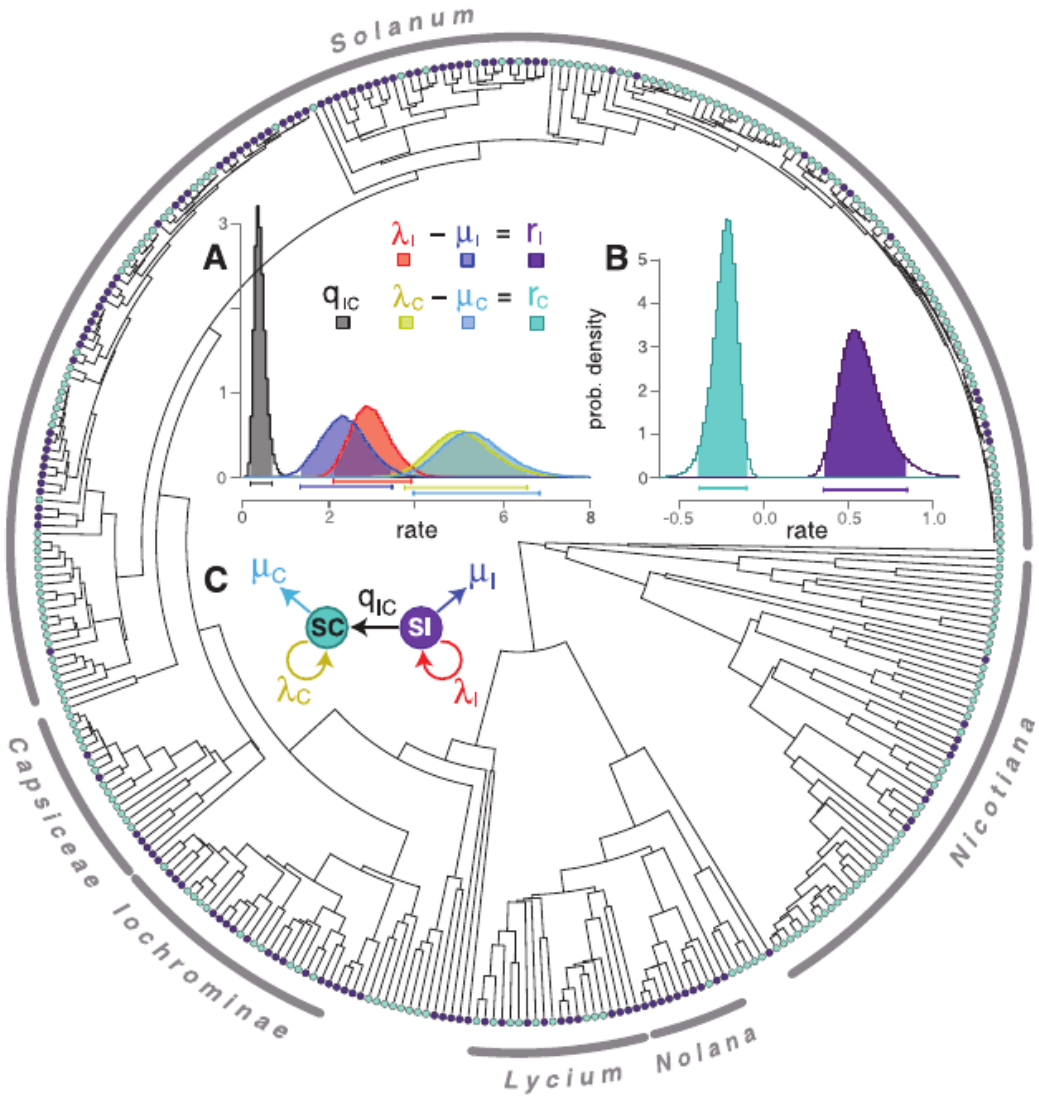
- Variation in traits among species
- New species tend to resemble ancestors
- Traits promote speciation

This is an old idea



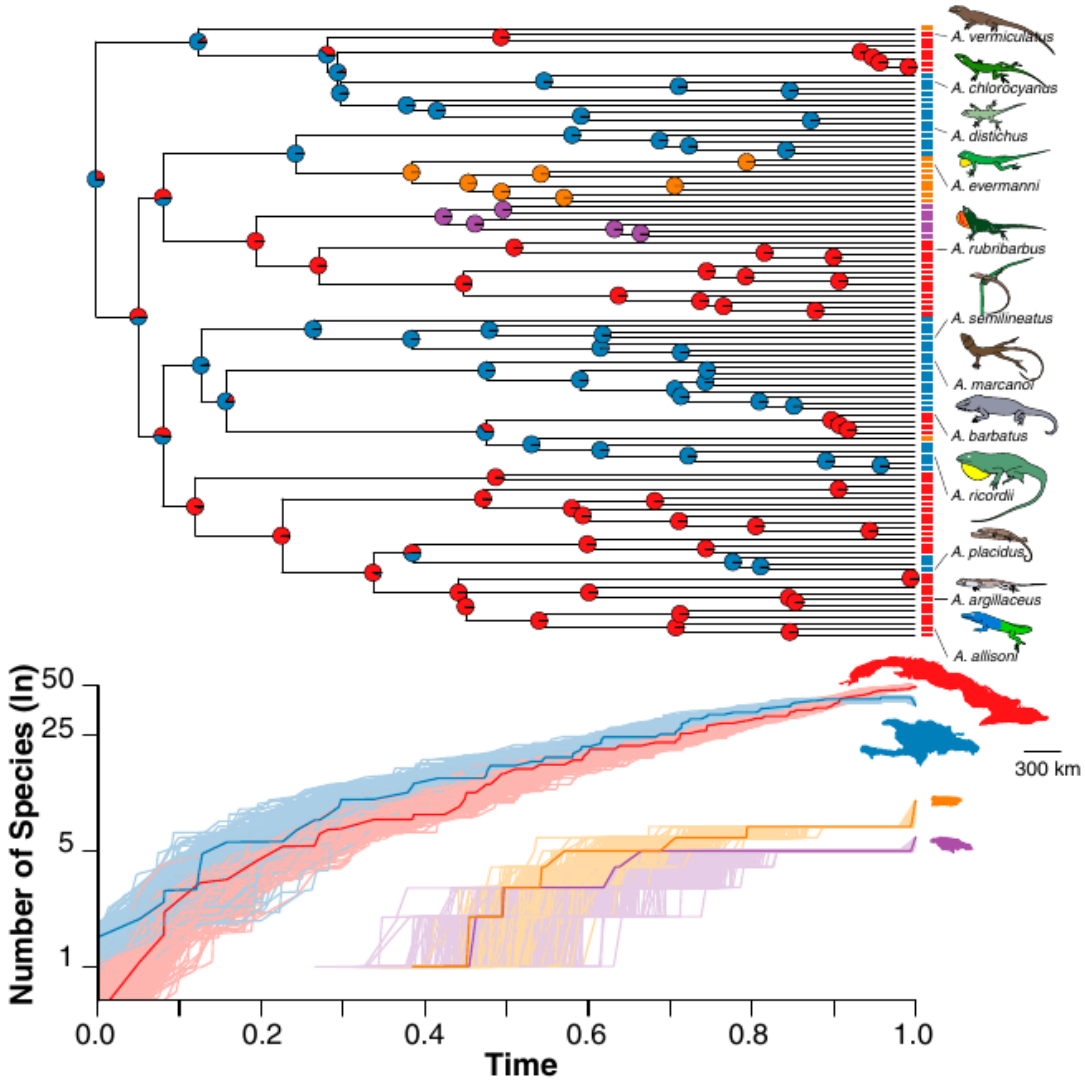
Stanley 1975

But is a huge topic again



Goldberg et al. 2010

Different types of traits may influence diversification

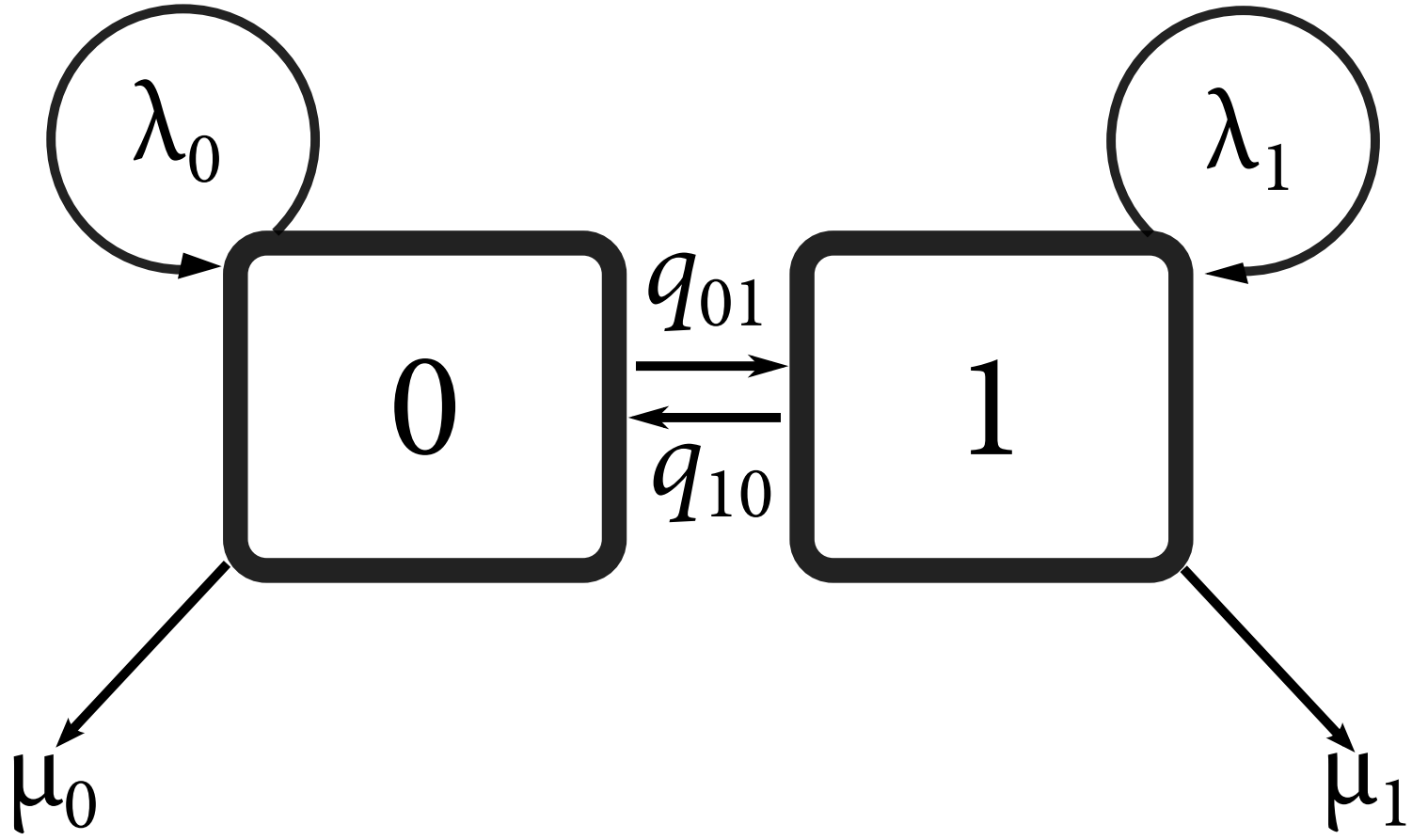


Rabosky and Glor 2010

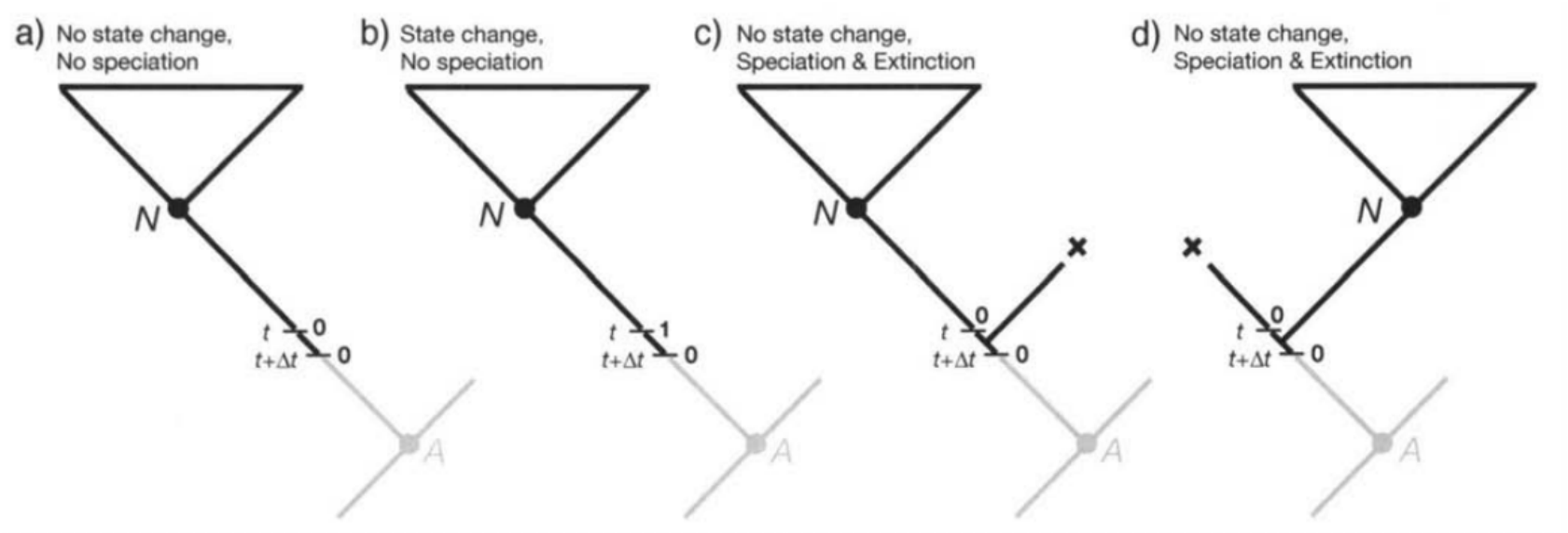
Binary State Speciation Extinction Model

- Transitions between two states occur at rates q_{01} and q_{10}
- Assume that diversification occurs via a "birth-death" model
- Lineages in state 0 speciate at rate λ_0 and go extinct at rate μ_0
- Lineages in state 1 speciate at rate λ_1 and go extinct at rate μ_0
- Simultaneously model speciation, extinction and transitions between states

Binary State Speciation Extinction Model



Binary State Speciation Extinction Model



Maddison et al. 2007

The xxSSE class of models

- Multistate traits (MuSSE, FitzJohn 2012)
- Geography (GeoSSE, Goldberg et al. 2011)
- Continuous traits (QuaSSE, FitzJohn 2010)
- Trait change at speciation (BiSSE-ness, Magnuson-Ford and Otto 2012; ClaSSE, Goldberg and Igić 2012)