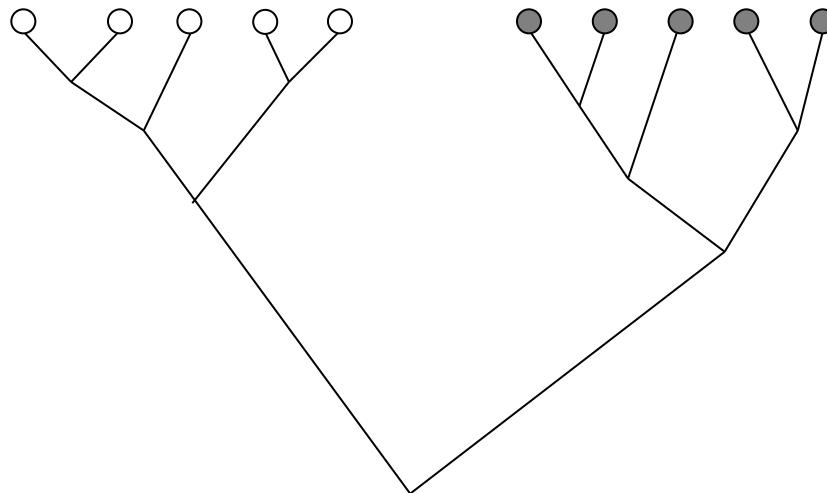
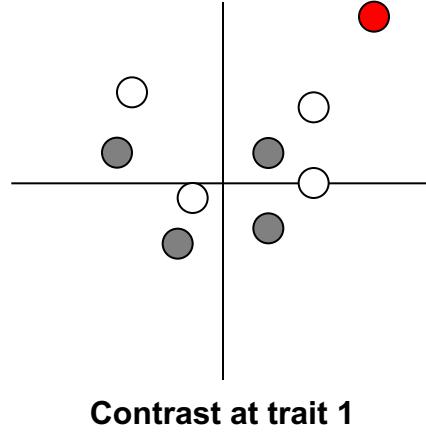


Phylogenies and statistics

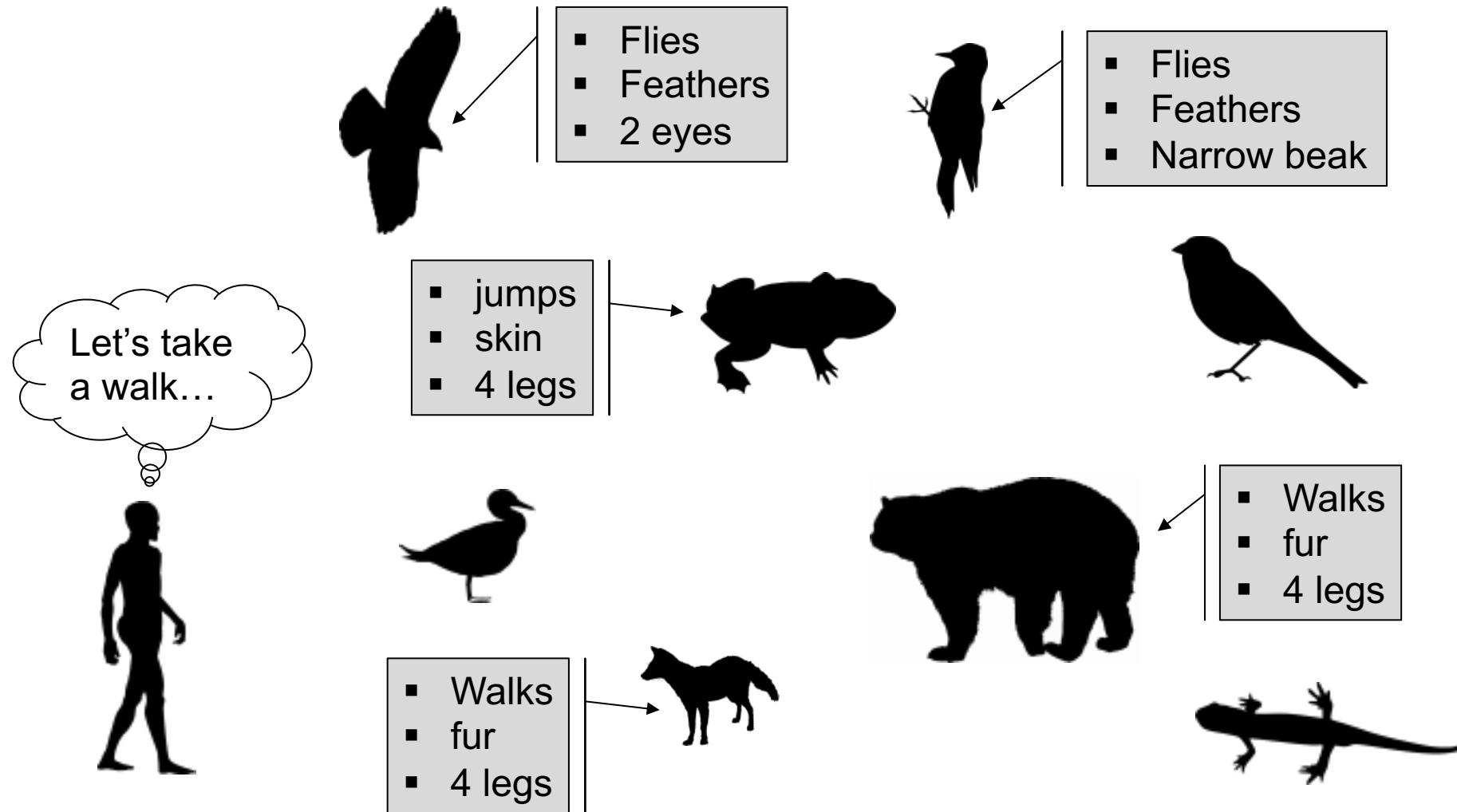


Contrast at trait 2

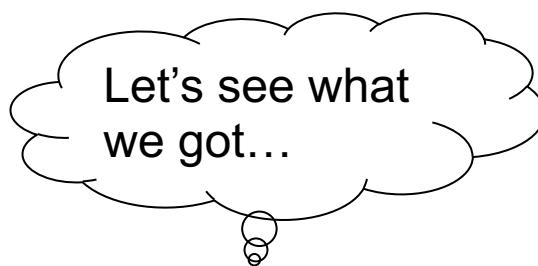


An introduction

An introduction



An introduction



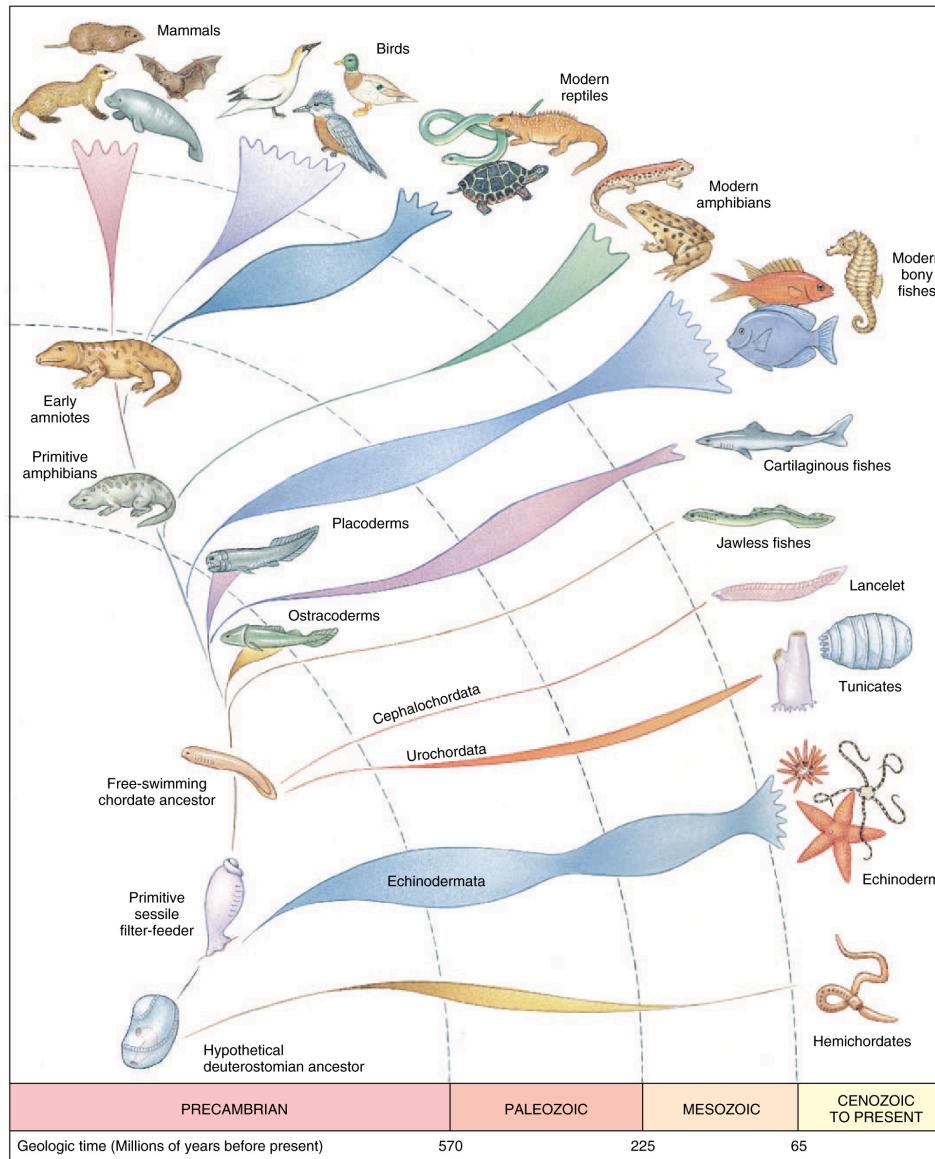
	feathers	other
Animal flies	9	0
Animal doesn't fly	0	7

```
> fisher.test(matrix(c(9,0,0,7),2))
```

Fisher's Exact Test for Count Data

```
data: matrix(c(9, 0, 0, 7), 2)
p-value = 8.741e-05
```

Evolution of flight (and feathers)

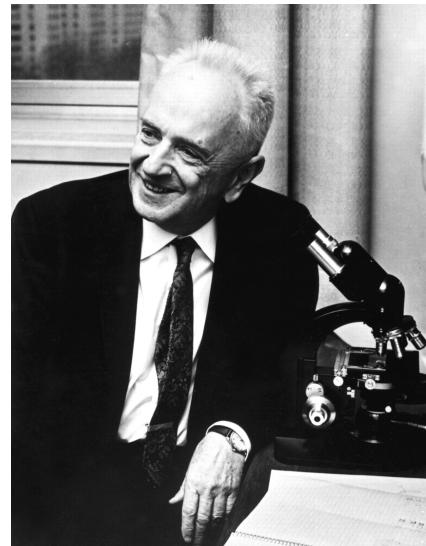


Credit: University Of Kelaniya, Sri Lanka

The comparative method

The comparative method

- Definition
 - It is the name given to a family of methods that aims at correcting for the phylogenetic non-independence of species when comparing them
- Why ?
 - Species cannot be considered without the context of evolution; the observations are not independent

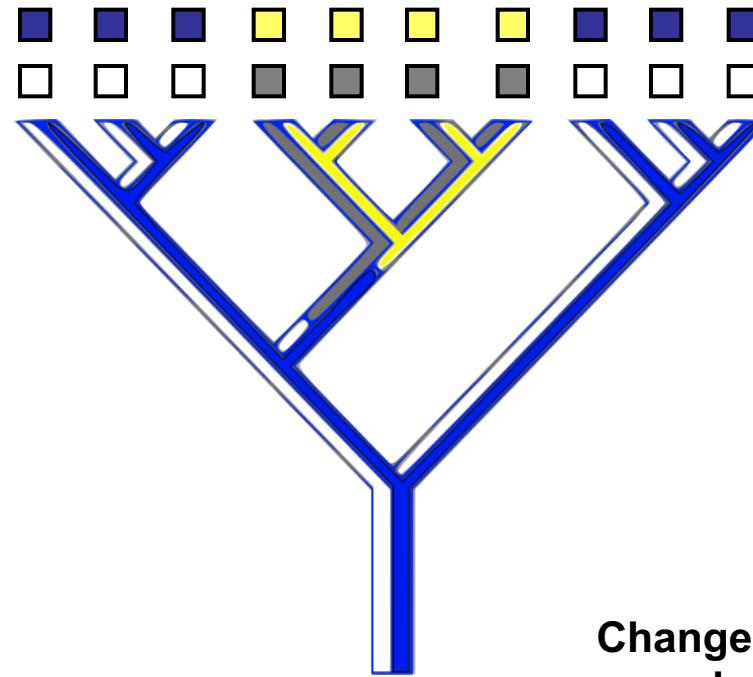


*Nothing in Biology Makes Sense
Except in the Light of Evolution*

Theodious Dobzhansky, 1973

A qualitative example

species	flower	stem
1	blue	spiny
2	blue	spiny
3	blue	spiny
4	blue	spiny
5	yellow	smooth
6	yellow	smooth
7	yellow	smooth
8	yellow	smooth
9	blue	spiny
10	blue	spiny



Changes along
branches

blue yellow

□ spiny

6	0
0	4

Fisher exact test
 $p = 0.00476$

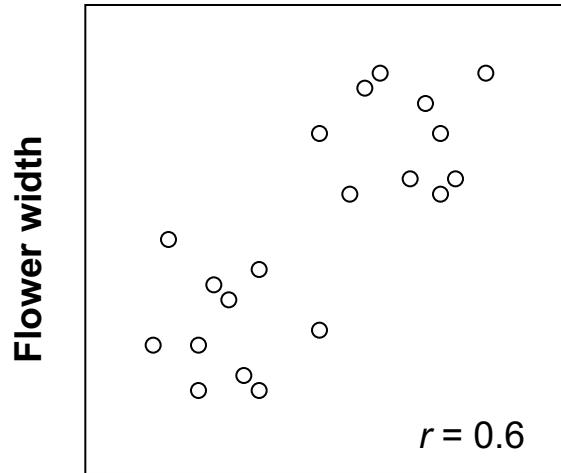
■ smooth

Fisher exact test
 $p = 0.056$

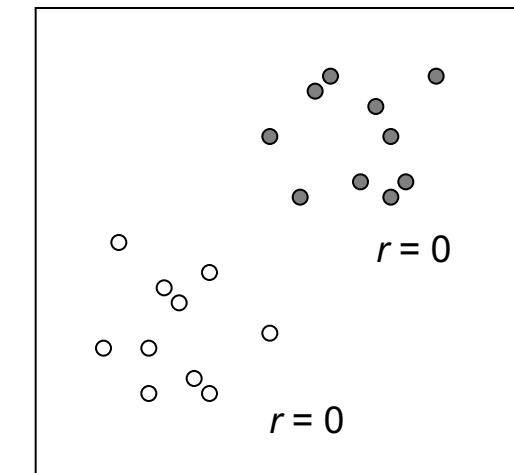
stem

flower		
yes	no	
yes	1	0
no	0	17

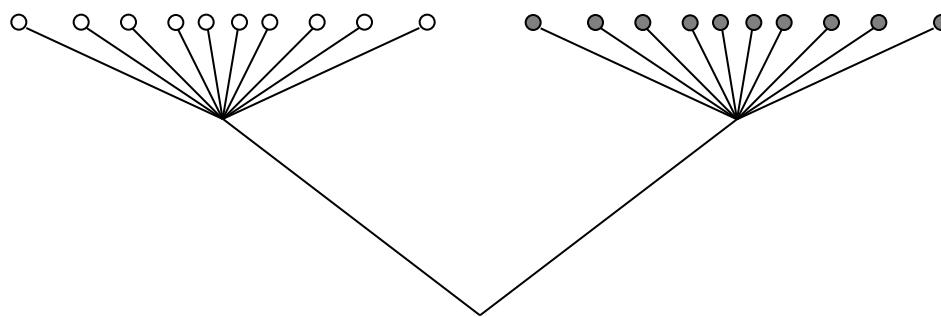
A quantitative example



Trichome number
on the leaves

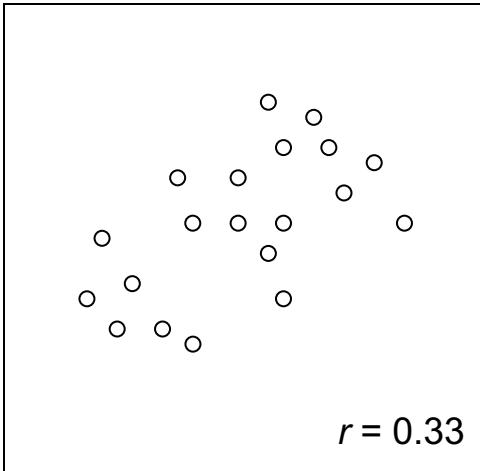


Trichome number
on the leaves



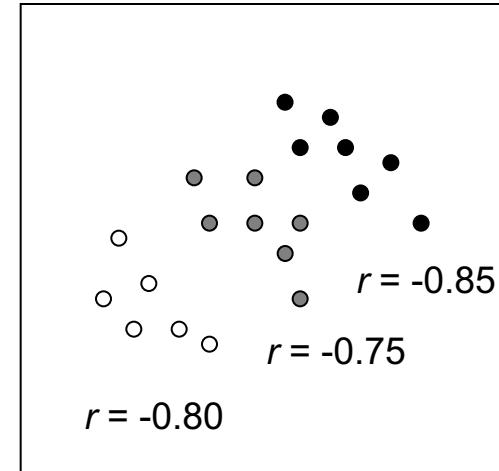
Another quantitative example

Flower width

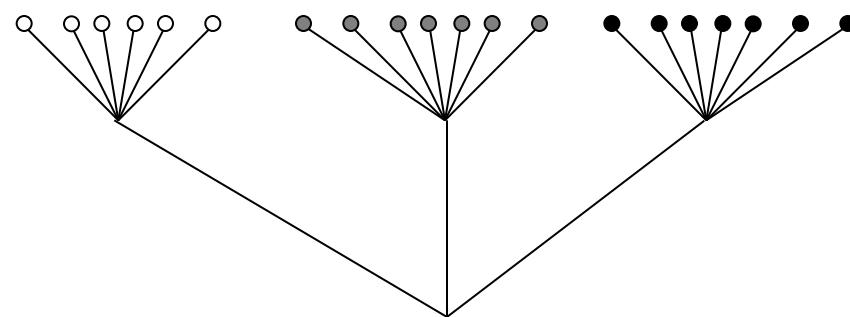


Trichome number
on the leaves

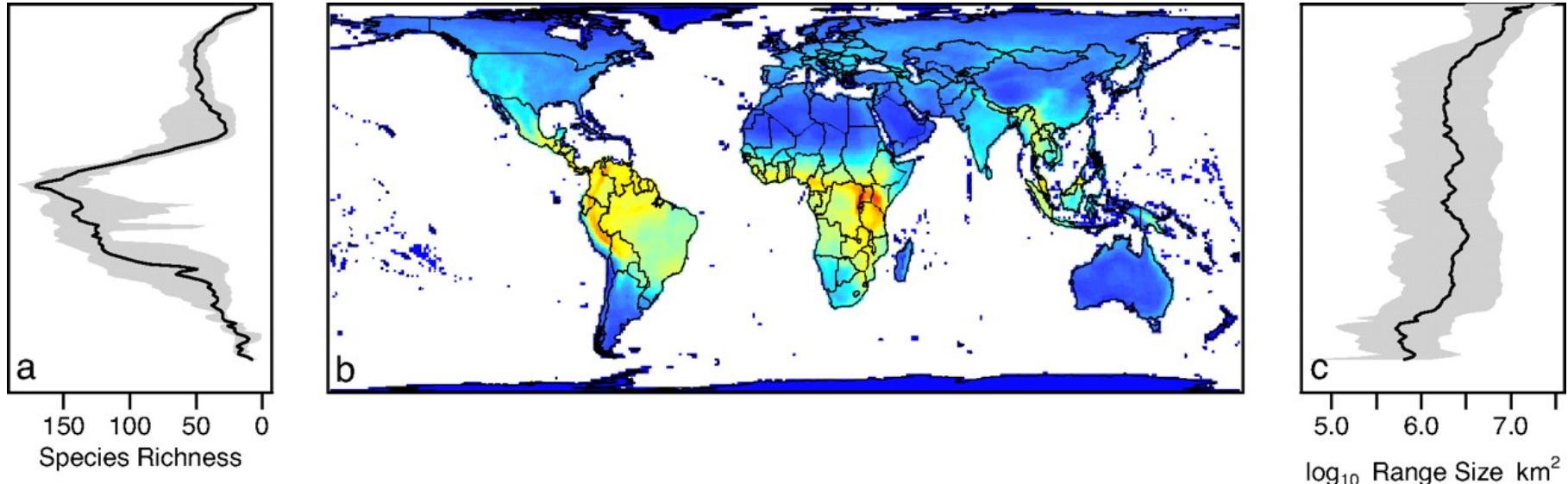
Flower width



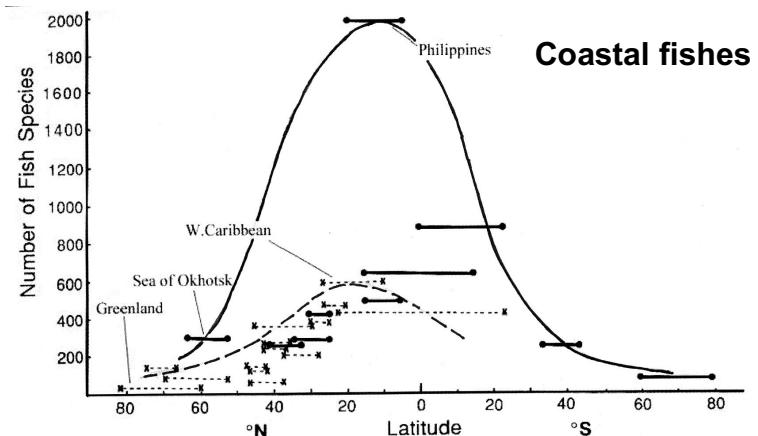
Trichome number
on the leaves



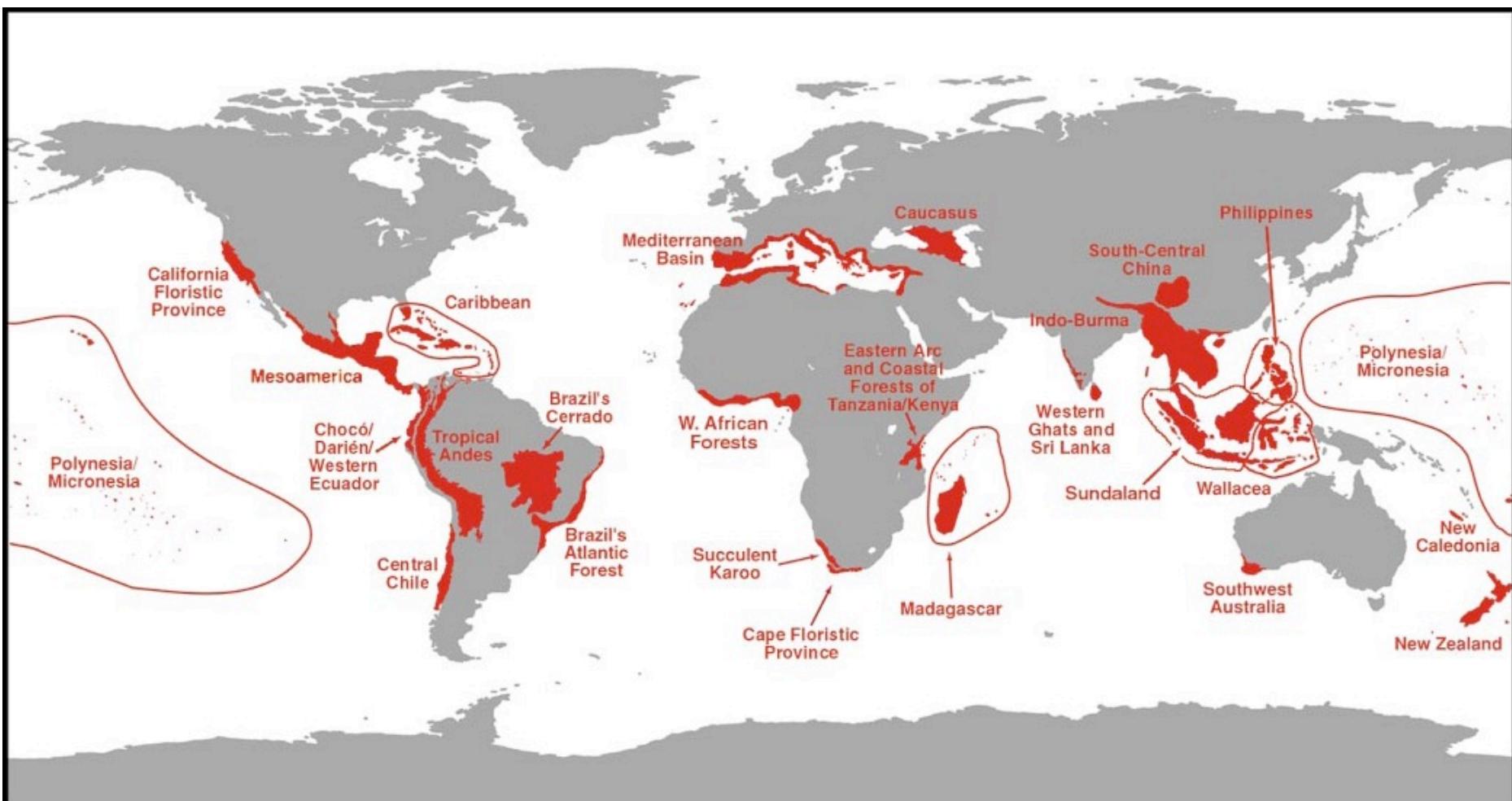
Biodiversity latitudinal gradient



Mammals - Davies T J et al. PNAS 2008;105:11556-11563

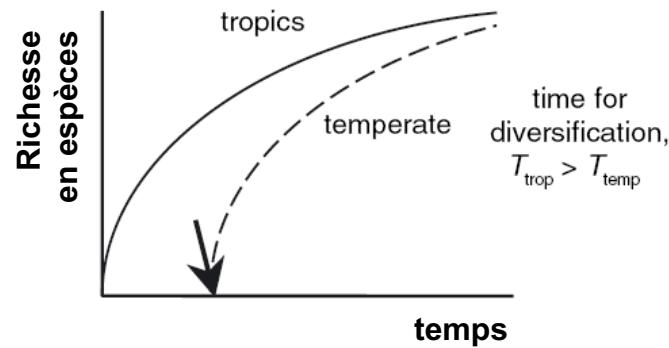


25 biodiversity hotspots

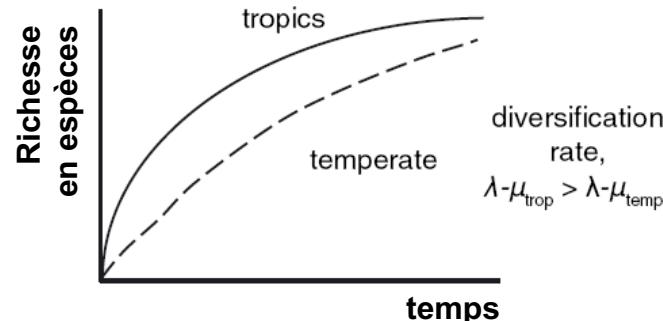


Some evolutionary hypotheses

- More time for speciation in the tropics (older environments)



- Higher diversification rates in the tropics, due to:
 - Higher speciation rates in the tropics (Cradle theory)
 - Lower extinction rates in the tropics (Museum theory)

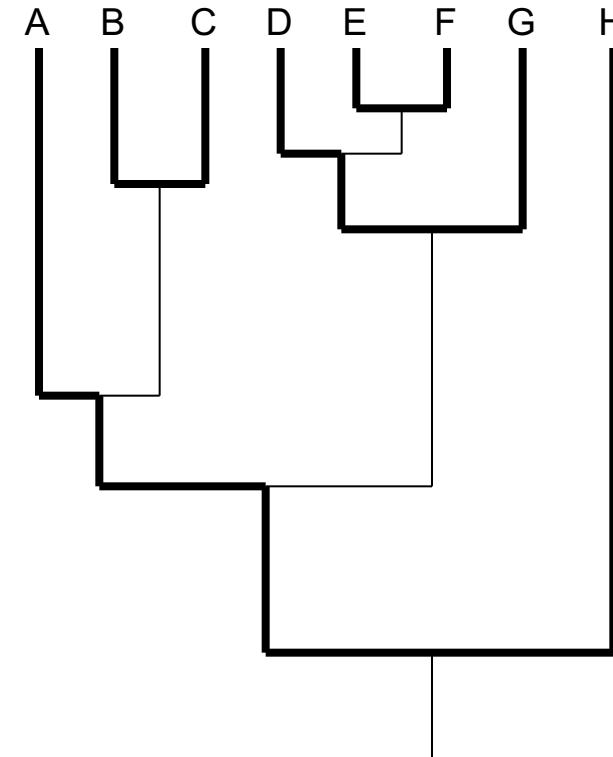
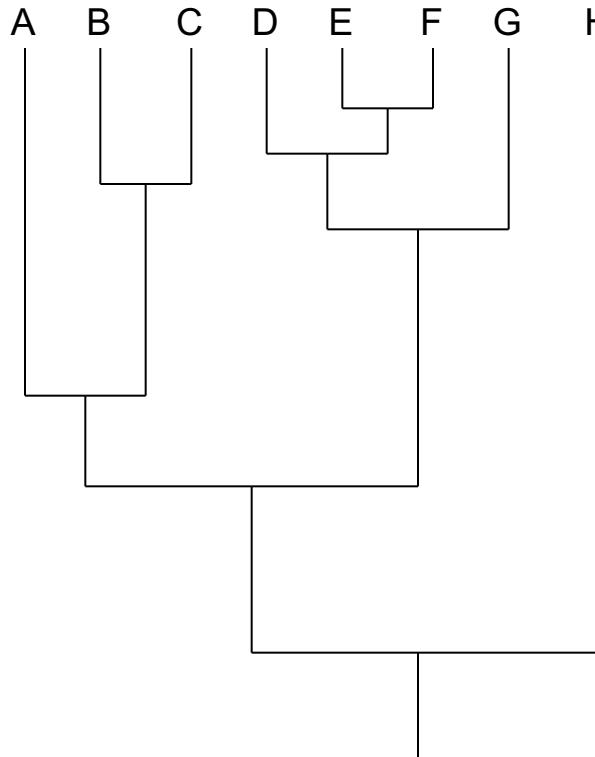


Comparative methods

Different approaches

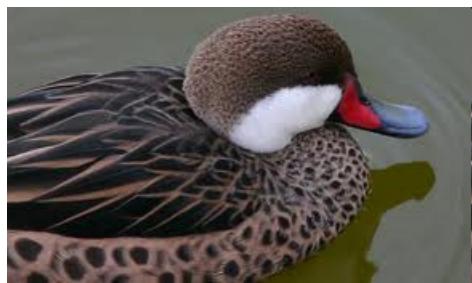
- Studying pairs of species
- Independent contrasts
- Phylogenetic generalized least squares (PGLS)
- Phylogenetic mixed models (PMM)

Pairs of species (sometimes sister clades)

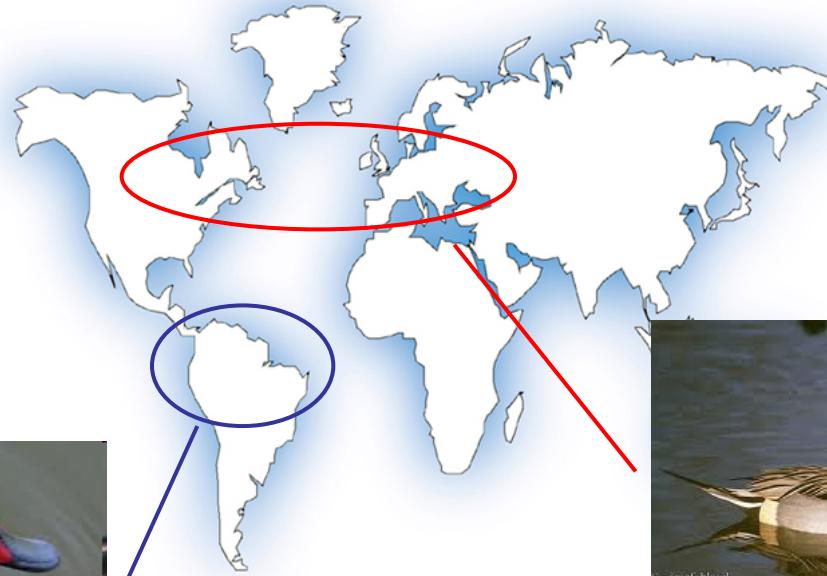


We can thus treat each comparison as independent from each other

Test for higher diversification rate in the tropics



Anas bahamensis



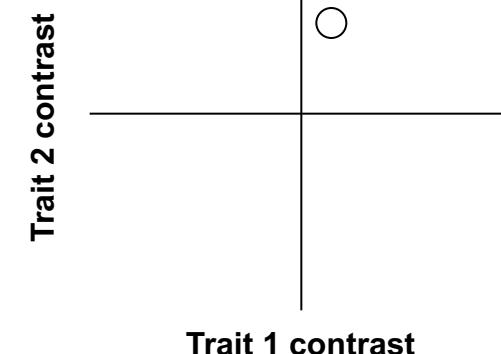
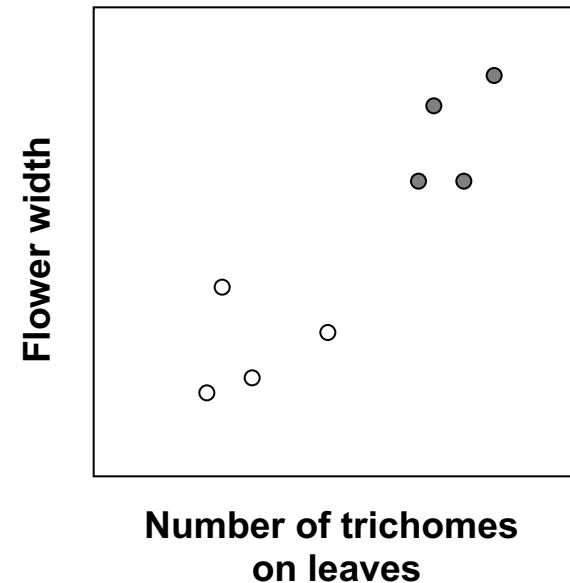
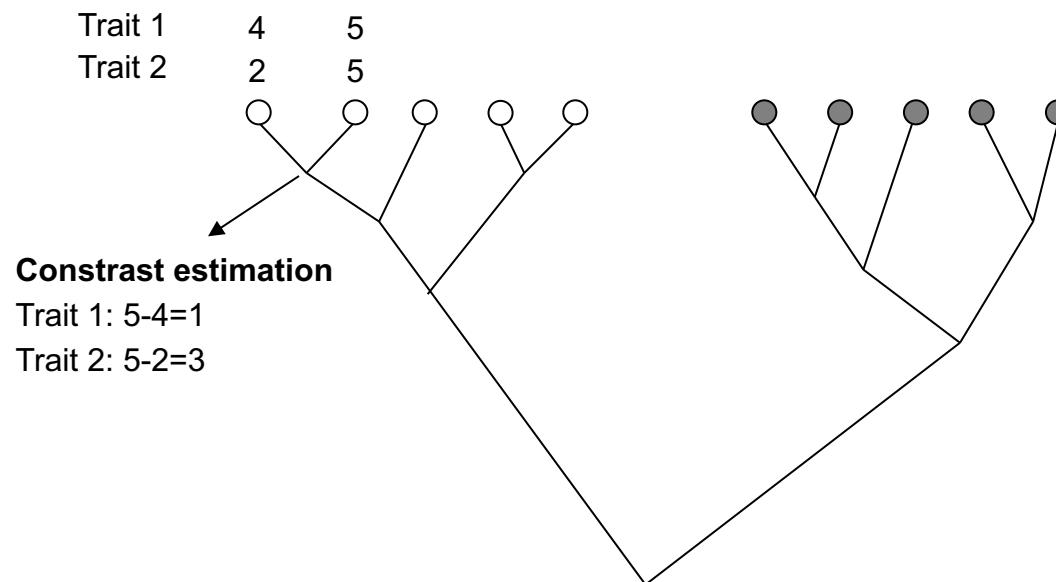
Anas acuta

Number of contrasts	Number of positive contrasts	Sign test	Wilcoxon signed rank test
33	15	0.76	0.36

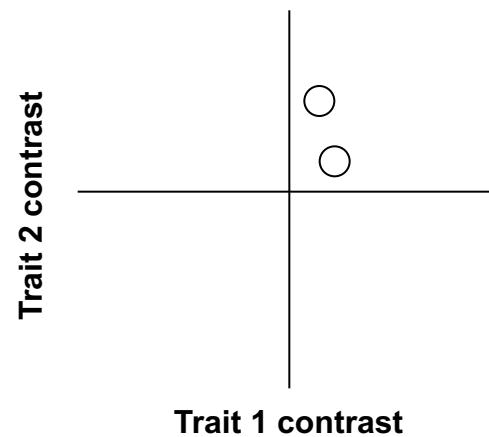
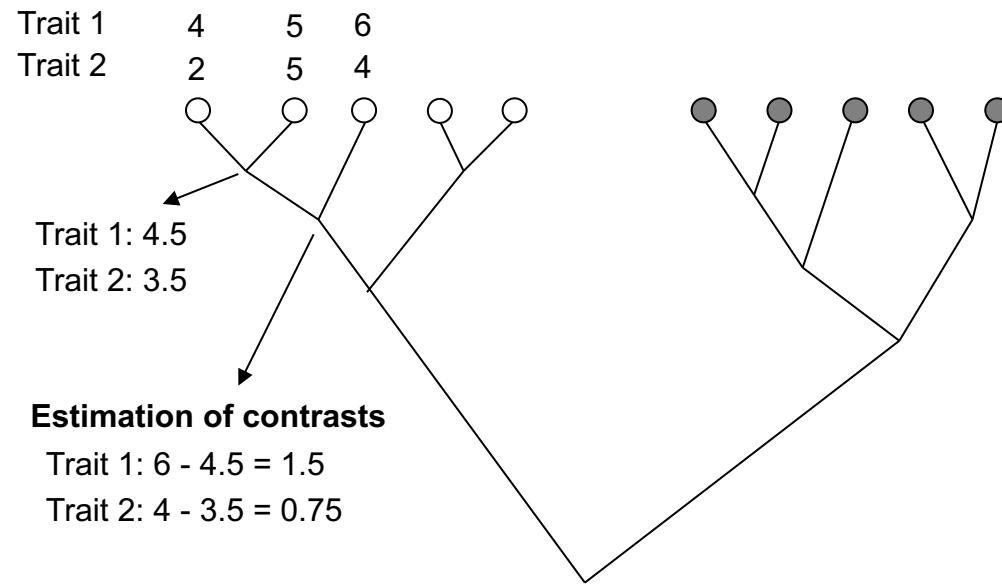
Not significant

Independent contrasts

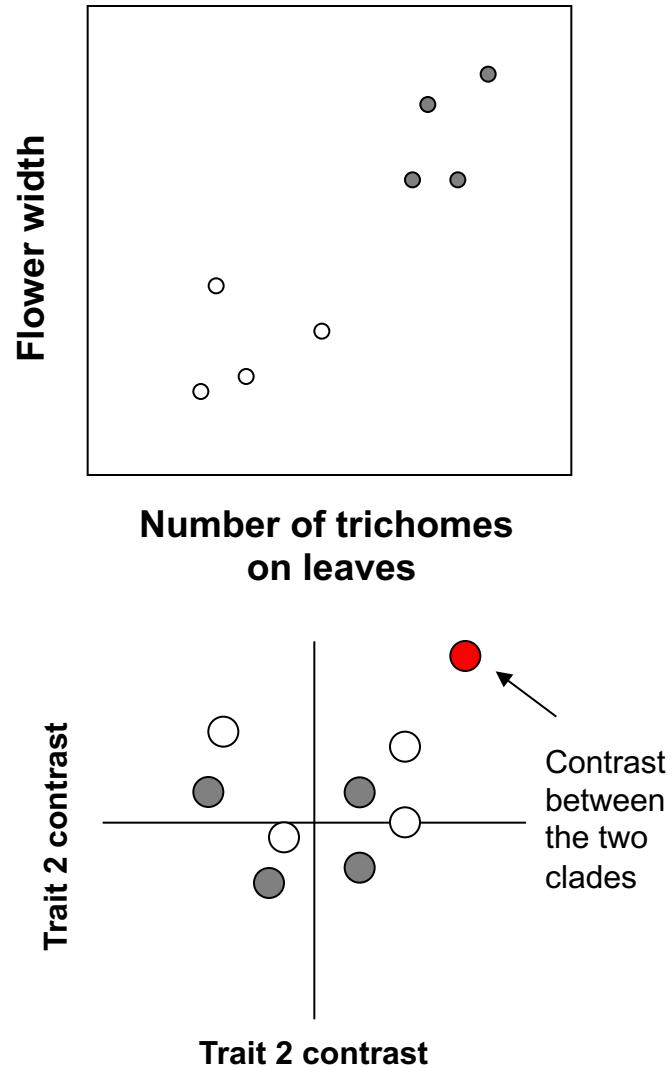
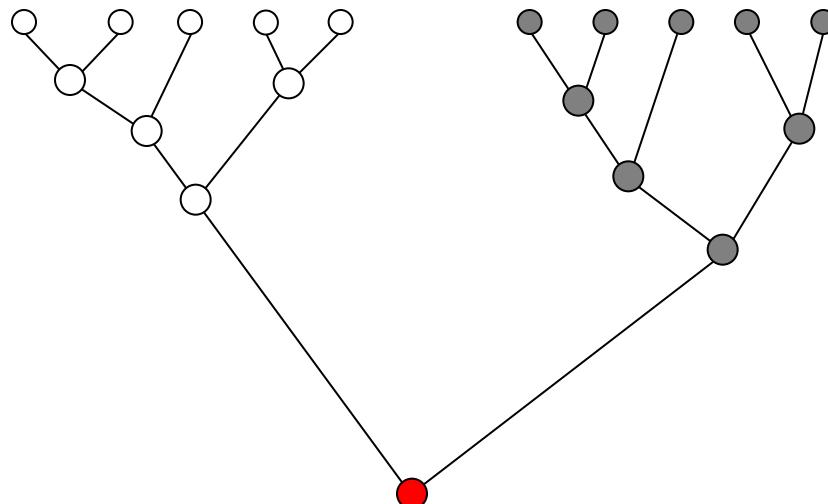
Objective: Test the correlation between characters while taking into account the evolutionary history of species



Independent contrasts

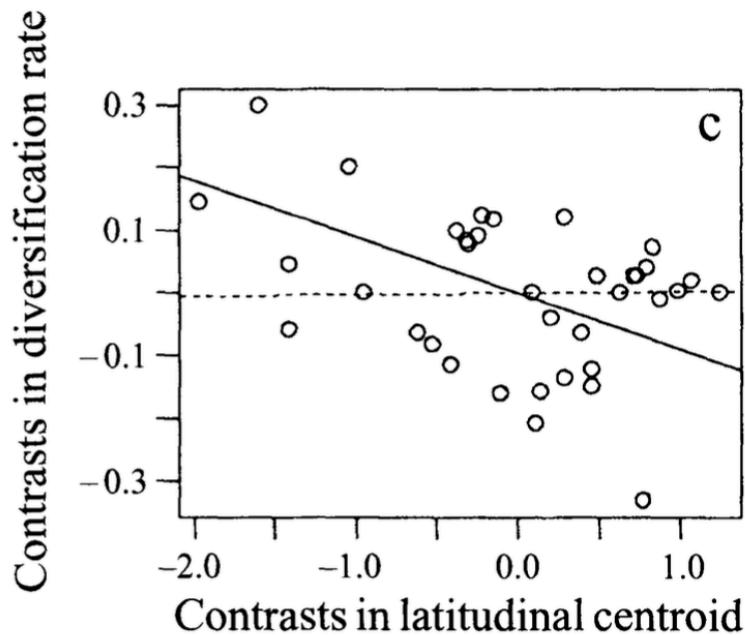


Independent contrasts



Test for higher diversification rate in the tropics

Test: Using a bird genera phylogeny, Cardillo et al. (2005) estimated for each genera their diversification rate and their mean latitude. They then estimated the PIC to test if there is a correlation between latitude and the diversification rate.



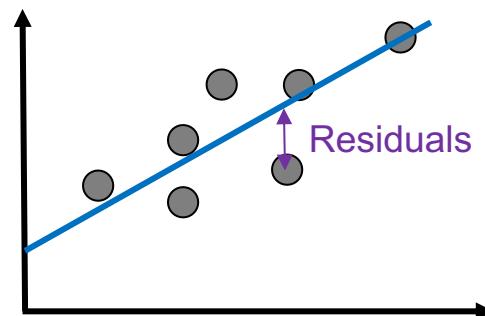
P-value = 0.023

**Higher diversification rates
in the tropics**

Phylogenetic Generalized Least Squares

Linear model (ordinary least squares – OLS)

$$\mathbf{y} = \alpha + \beta \mathbf{x} + \mathbf{e}$$



Linear model (ordinary least squares – OLS)

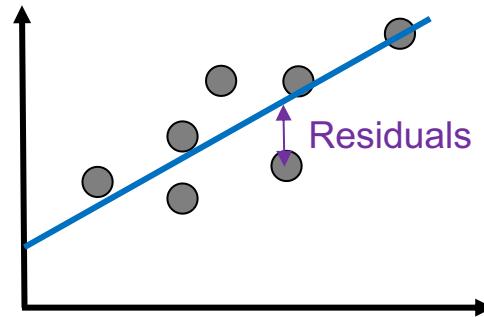
$$\mathbf{y} = \alpha + \beta \mathbf{x} + \mathbf{e}$$

Many assumptions

- Normality, homogeneity, fixed X , independence, and correct model specification
- Independence:
 - The value of Y_i at X_i should not be influenced by other values of X_i

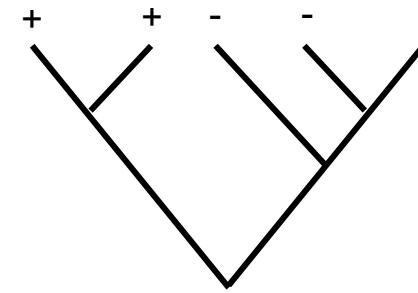
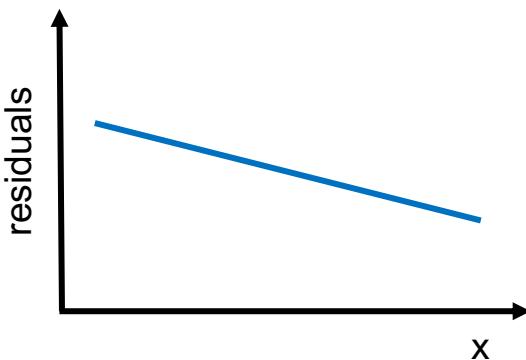
$$\mathbf{e} \sim \mathcal{N}(0, \sigma^2)$$

Examples of violations



Residuals correlated with the explanatory variable

Residuals correlated with the phylogenetic relationships



Practicals in R



www.github.com/simjoly/ComparativeMethods-HalfDayWorkshop

Up to Challenge #1

Phylogenetic generalized least squares

Phylogenetic generalized least squares (PGLS)

- A special case of generalized least squares (GLS)
 - Allow the residuals of the model to be correlated in a specific way
 - Used for spatial correlation, time series, phylogenetic relationships, etc.

Phylogenetic generalized least squares (PGLS)

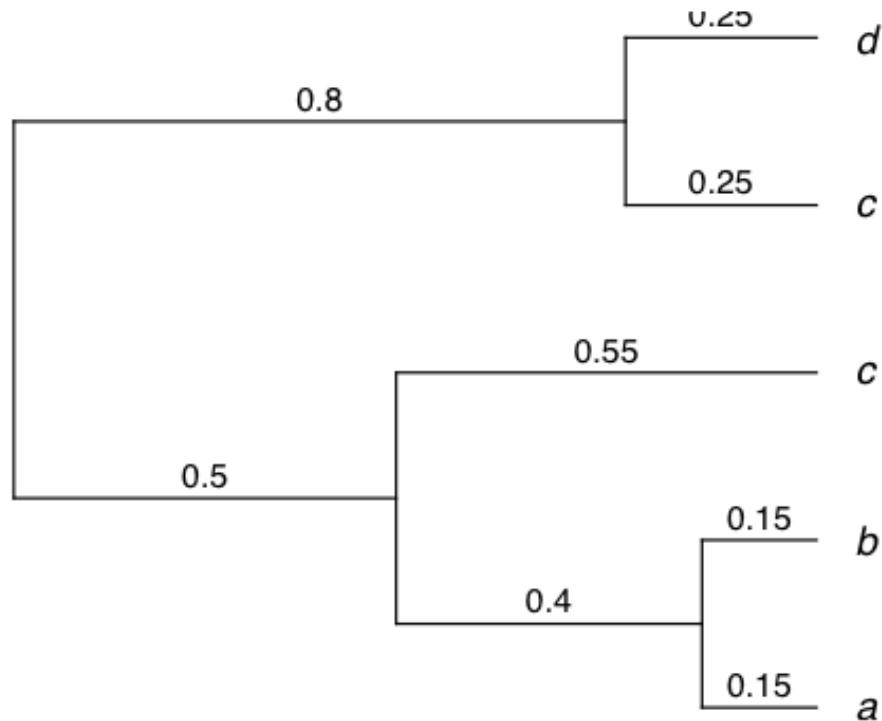
$$\mathbf{y} = \alpha + \beta \mathbf{x} + \mathbf{e}$$

$$\mathbf{e} \sim \mathcal{N}(0, \sigma^2 \mathbf{C})$$

$$\sigma^2 \mathbf{C} = \sigma^2 \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \dots & \rho_{1i} \\ \rho_{21} & 1 & \rho_{23} & \dots & \rho_{2i} \\ \rho_{31} & \rho_{32} & 1 & \dots & \rho_{3i} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{i1} & \rho_{i2} & \rho_{i3} & \dots & 1 \end{pmatrix}$$

correlation matrix

From a tree to a variance-covariance structure



	a	b	c	c	d
a	1.05	0.90	0.50	0.00	0.00
b	0.90	1.05	0.50	0.00	0.00
c	0.50	0.50	1.05	0.00	0.00
c	0.00	0.00	0.00	1.05	0.80
d	0.00	0.00	0.00	0.80	1.05

... to a correlation matrix

	a	b	c	c	d
a	1.05	0.90	0.50	0.00	0.00
b	0.90	1.05	0.50	0.00	0.00
c	0.50	0.50	1.05	0.00	0.00
c	0.00	0.00	0.00	1.05	0.80
d	0.00	0.00	0.00	0.80	1.05

VCV matrix

	a	b	c	c	d
a	1.000	0.857	0.476	0.000	0.000
b	0.857	1.000	0.476	0.000	0.000
c	0.476	0.476	1.000	0.000	0.000
c	0.000	0.000	0.000	1.000	0.762
d	0.000	0.000	0.000	0.762	1.000

correlation matrix

Back to PGLS

$$\mathbf{e} \sim \mathcal{N}(0, \sigma^2 \mathbf{C})$$

$$\sigma^2 \mathbf{C} = \sigma^2 \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \dots & \rho_{1i} \\ \rho_{21} & 1 & \rho_{23} & \dots & \rho_{2i} \\ \rho_{31} & \rho_{32} & 1 & \dots & \rho_{3i} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{i1} & \rho_{i2} & \rho_{i3} & \dots & 1 \end{pmatrix}$$

correlation matrix

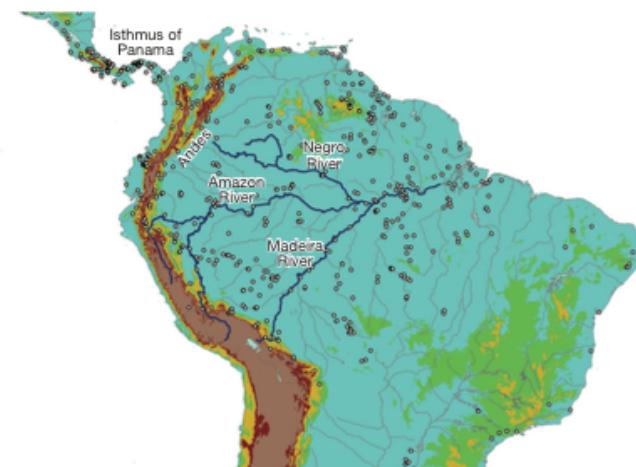
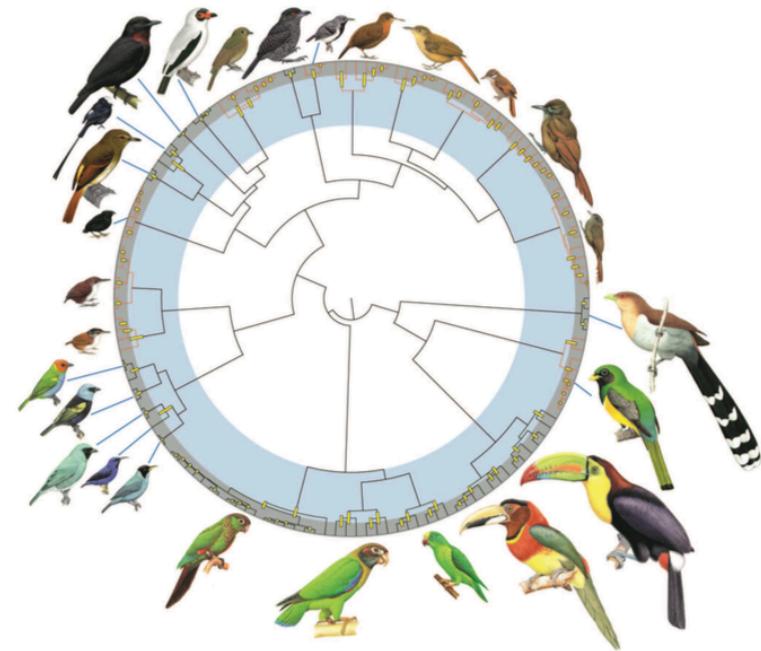
Practicals in R



Up to Challenge #3
(you can also look at the next section on Phylogenetic Independent Contrasts)

Example

- What drove bird diversification in the neotropics?
 - Lineage age
 - Foraging behaviour (stratum: understorey or canopy)
 - Ancestral area
 - Niche breadth
- Studied the diversity of 27 lineages of birds



Example

- What drove bird diversification in the neotropics?

Table 1 | Phylogenetic generalized least-squares regression showing the effects of historical and ecological variables on species diversity

Effect	Estimate	Standard error	t value	P	ΔAIC_c
Lineage age	0.1187	0.0283	4.1907	0.0004	6.9586
Foraging stratum	0.5188	0.2025	2.5623	0.0178	4.0122
Ancestral origin	-0.1921	0.2023	-0.9495	0.3527	-1.9546
Niche breadth	1.0097	1.0658	0.9473	0.3538	-1.9595

Relax the assumption that the residuals
need to be all phylogenetically correlated

The solution

- What happens if there are other sources of errors in the data than the phylogenetic correlations?
- Suppose for instance that there are measurement errors that are expected to be the same for all species.

$$\mathbf{y} = \alpha + \beta \mathbf{x} + \mathbf{e}$$

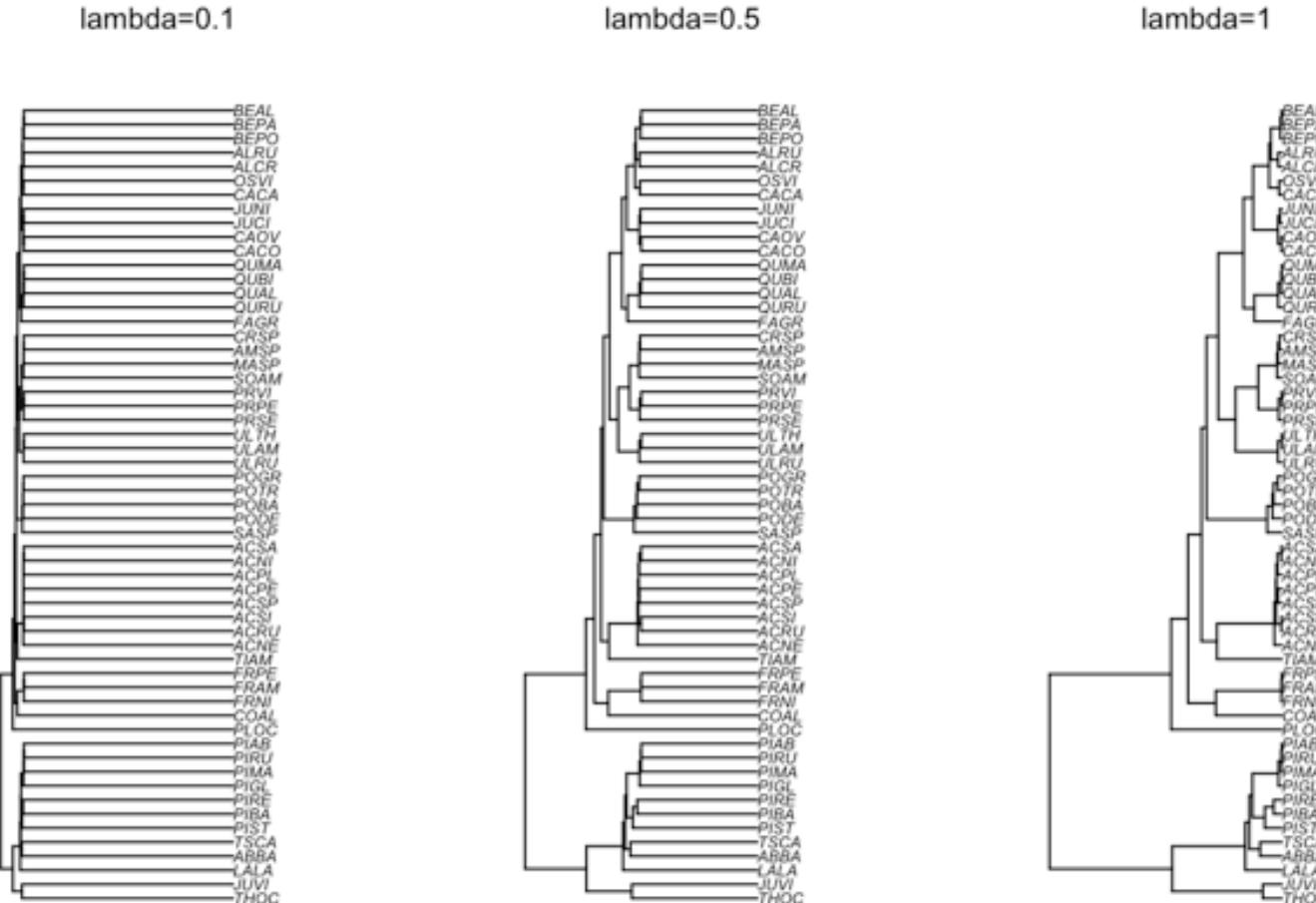
$$\mathbf{e} \sim \mathcal{N}(0, \sigma^2 \mathbf{C})$$

The solution: use the λ model

- Estimate a parameter λ in the GLS that determines the strength of the phylogenetic signal in the residuals.
 - $\lambda = 0$: no phylogenetic signal (equivalent to OLS)
 - $\lambda = 1$: perfect phylogenetic signal

$$\sigma^2 \mathbf{C} = \sigma^2 \begin{pmatrix} 1 & \lambda\rho_{12} & \lambda\rho_{13} & \dots & \lambda\rho_{1i} \\ \lambda\rho_{21} & 1 & \lambda\rho_{23} & \dots & \lambda\rho_{2i} \\ \lambda\rho_{31} & \lambda\rho_{32} & 1 & \dots & \lambda\rho_{3i} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \lambda\rho_{i1} & \lambda\rho_{i2} & \lambda\rho_{i3} & \dots & 1 \end{pmatrix}$$

Effect of parameter λ on the phylogeny



Practicals in R



Up to Challenge #4

Optional: look at the next section on phylogenetic ANOVA

Model Selection or Testing

Likelihood

$$L(D|M)$$

Concept of model choice

The choice of a model is generally based on the concept of parsimony, that stipulates that one should choose a more complex model only if it is necessary

A more complex model will always have a better fit (better likelihood), but its parameters will have greater variance (and thus greater uncertainty).

How then should we make the cutoff?

Likelihood Ratio Test

$$D = -2 \ln \frac{L(D|M_0)}{L(D|M_1)}$$

$$D = -2(\ln L(D|M_0) - \ln L(D|M_1))$$

The distribution of D can be approximated by a χ^2 distribution with x degree of freedom, where the degree of freedom correspond to the difference in parameters between the models.

M_0 is the null (simpler) model and M_1 is the more complex model.

Aikake Information Criterion (AIC)

$$AIC = -2\ln L(D|M) + 2k$$

where k is the number of parameters in the model.

The best model is the one with the smaller AIC.

Practicals in R

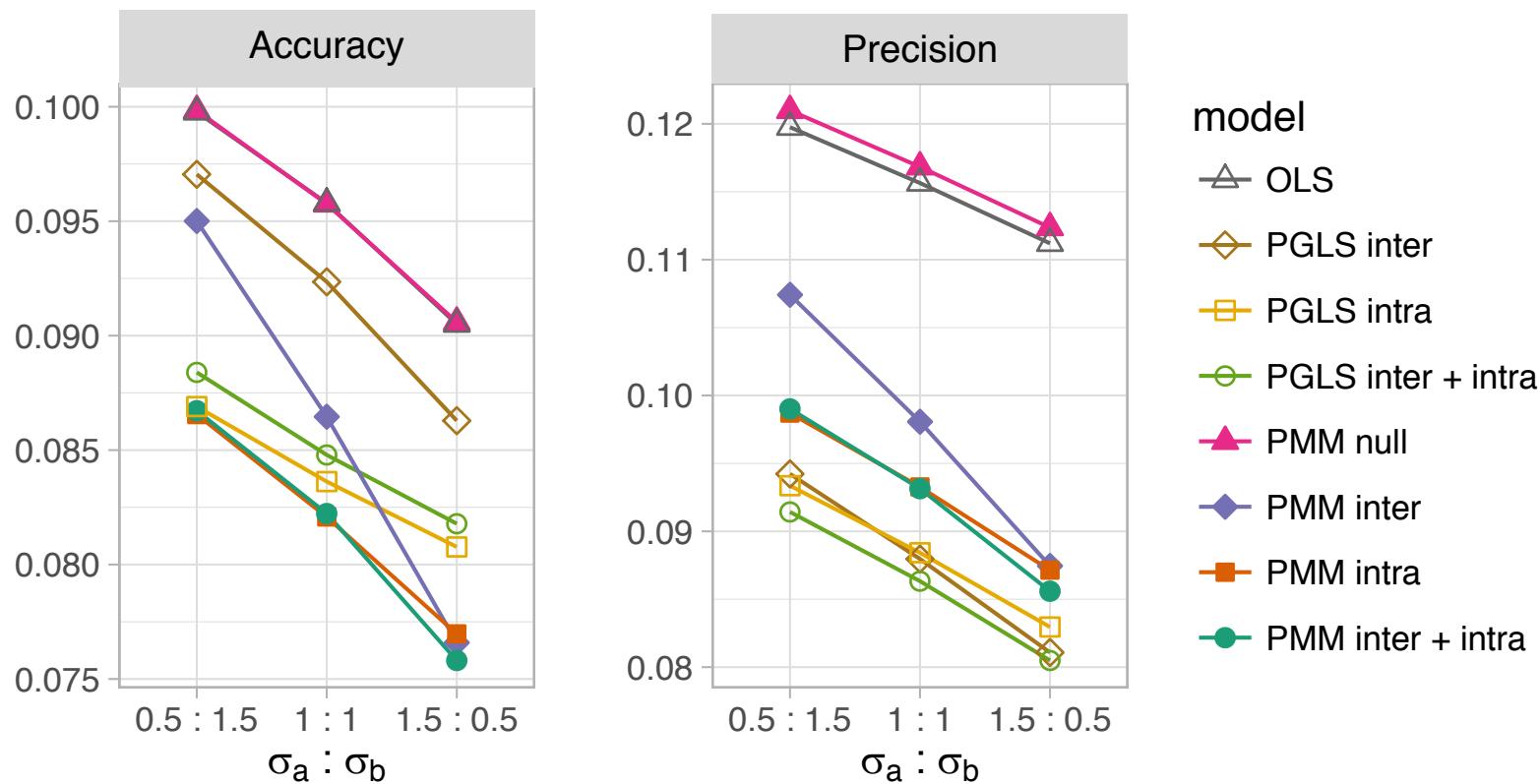


Look at the practicals on model testing

Why use comparative methods?

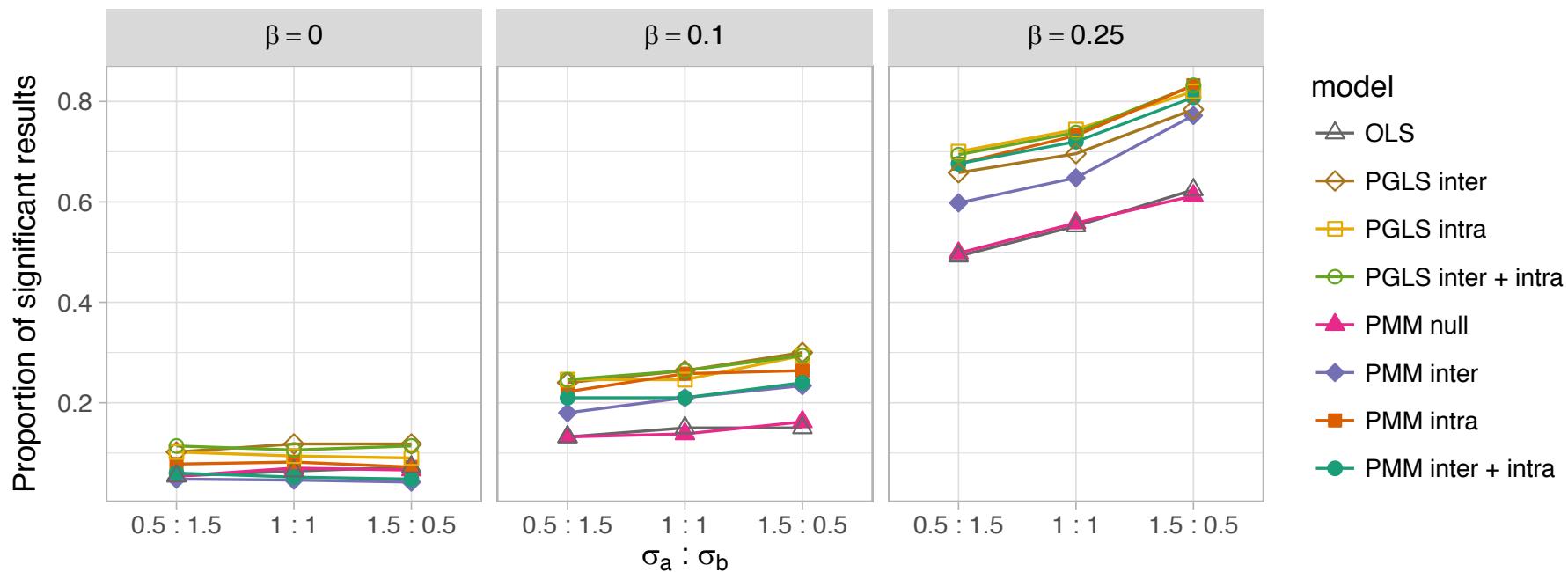
Why use comparative methods?

- Better accuracy and precision for fixed effects



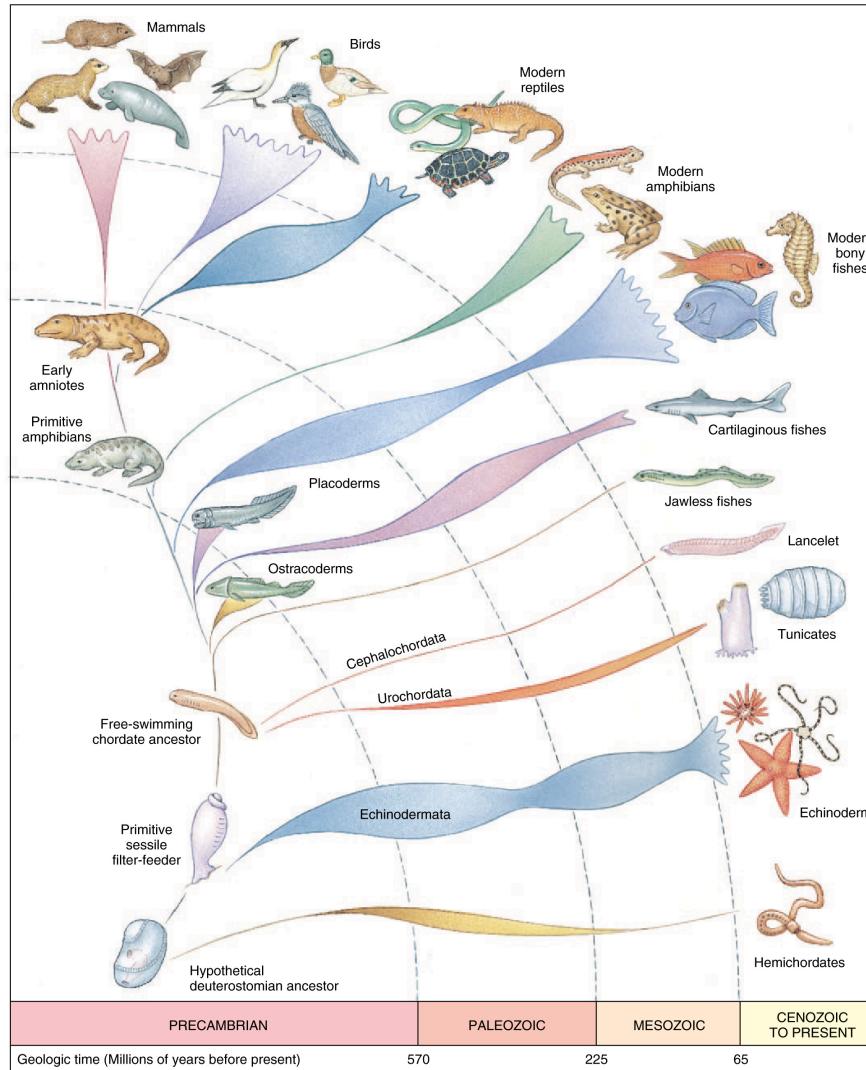
Why use comparative methods?

- Lower type I error and higher power

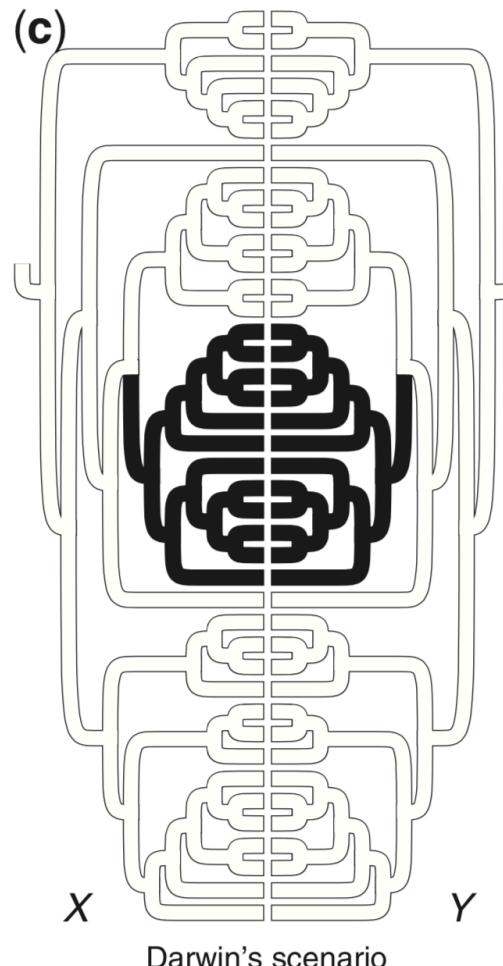


The problem of replication

Replication in evolutionary studies



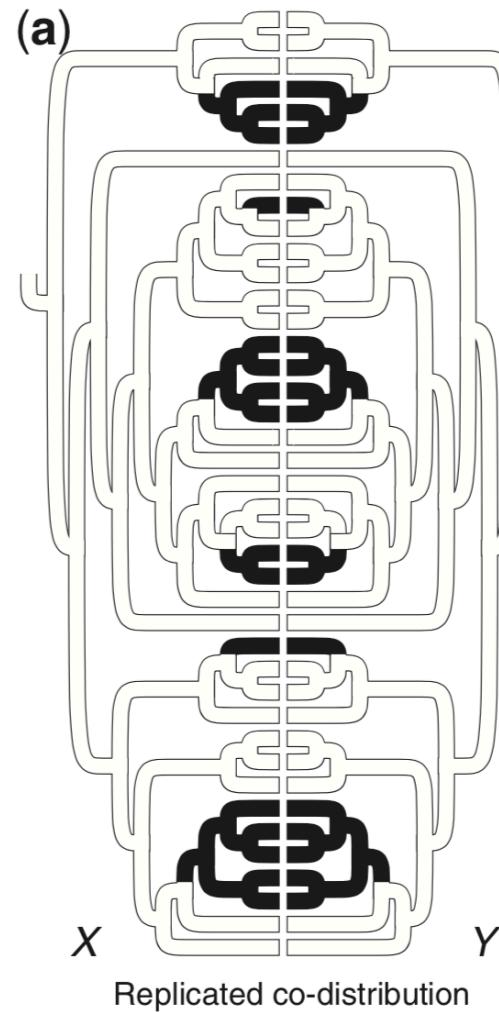
No replication (Darwin's scenario)



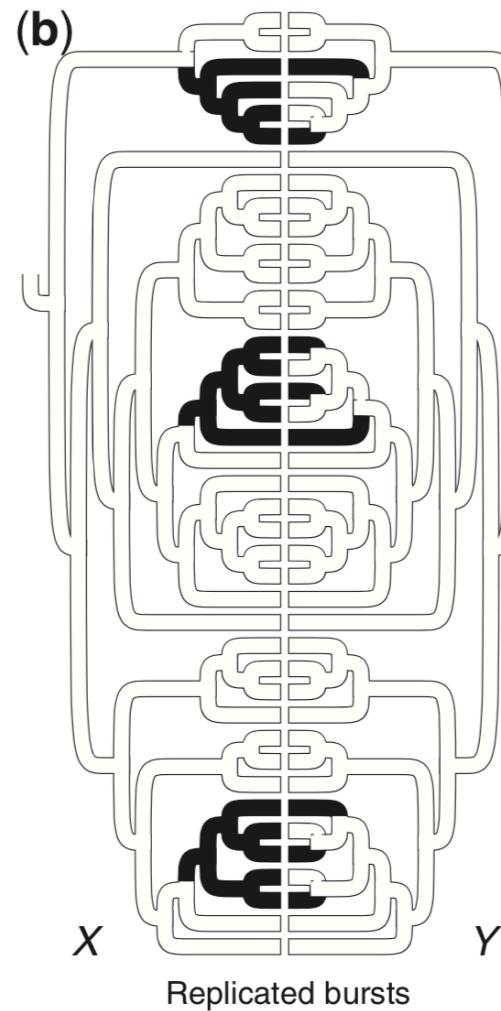
“We may often falsely attribute to correlated variation structures which are common to whole groups of species, and which in truth are simply due to inheritance; for an ancient progenitor may have acquired through natural selection some one modification in structure, and, after thousands of generations, some other and independent modification; and these two modifications, having been transmitted to a whole group of descendants with diverse habits, would naturally be thought to be in some necessary manner correlated.”

Darwin (1872)

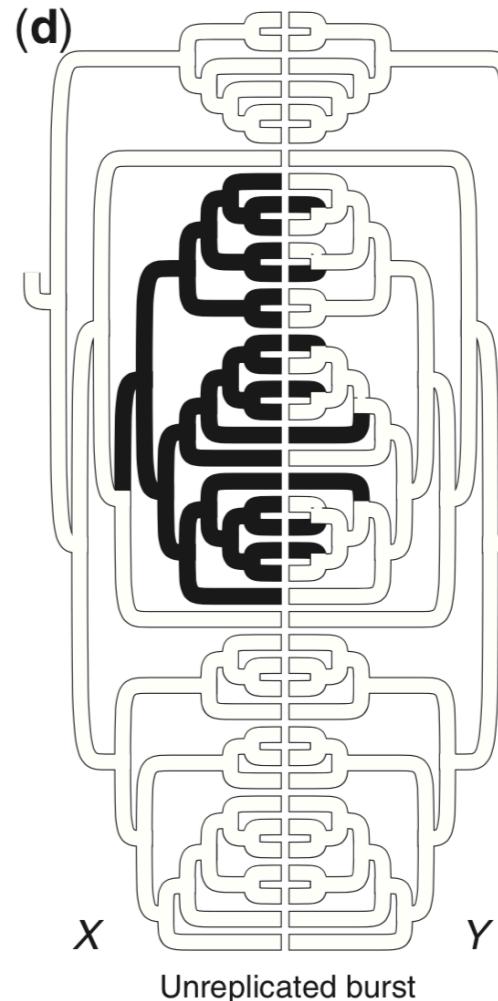
True replication



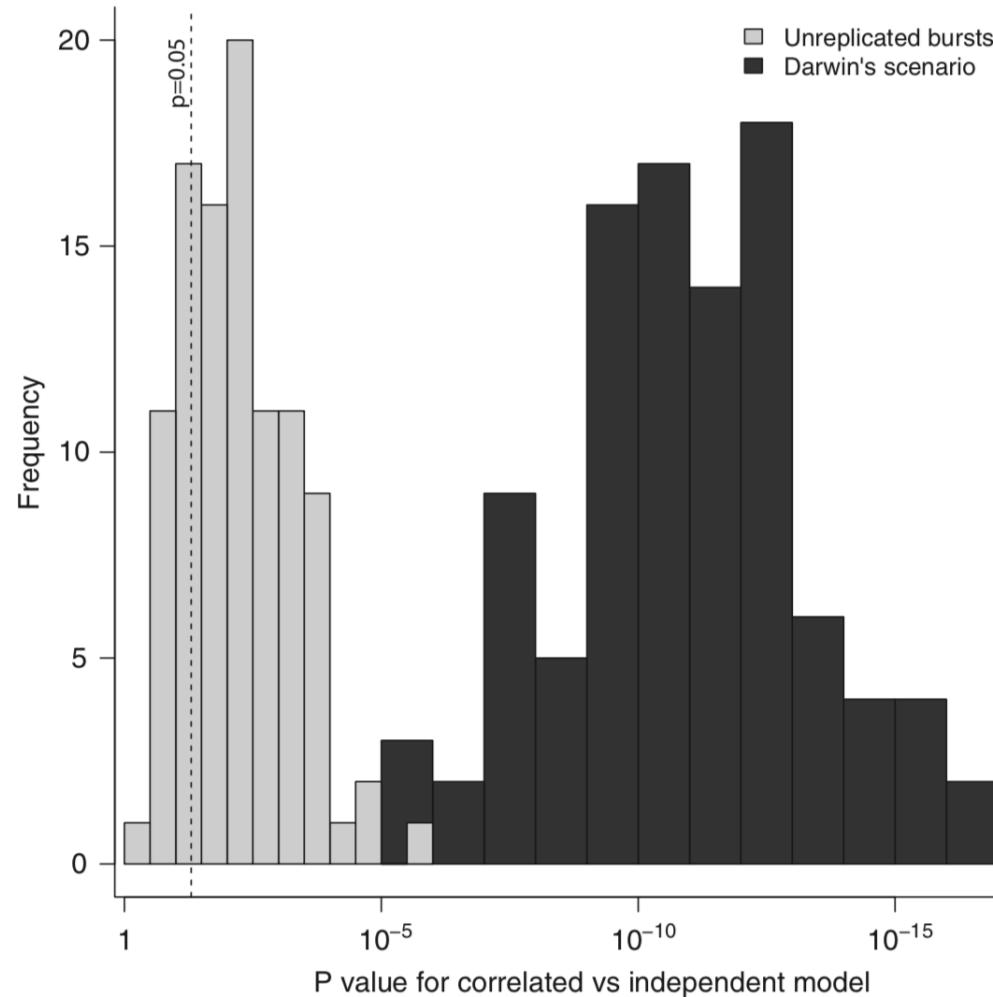
Replicated burst



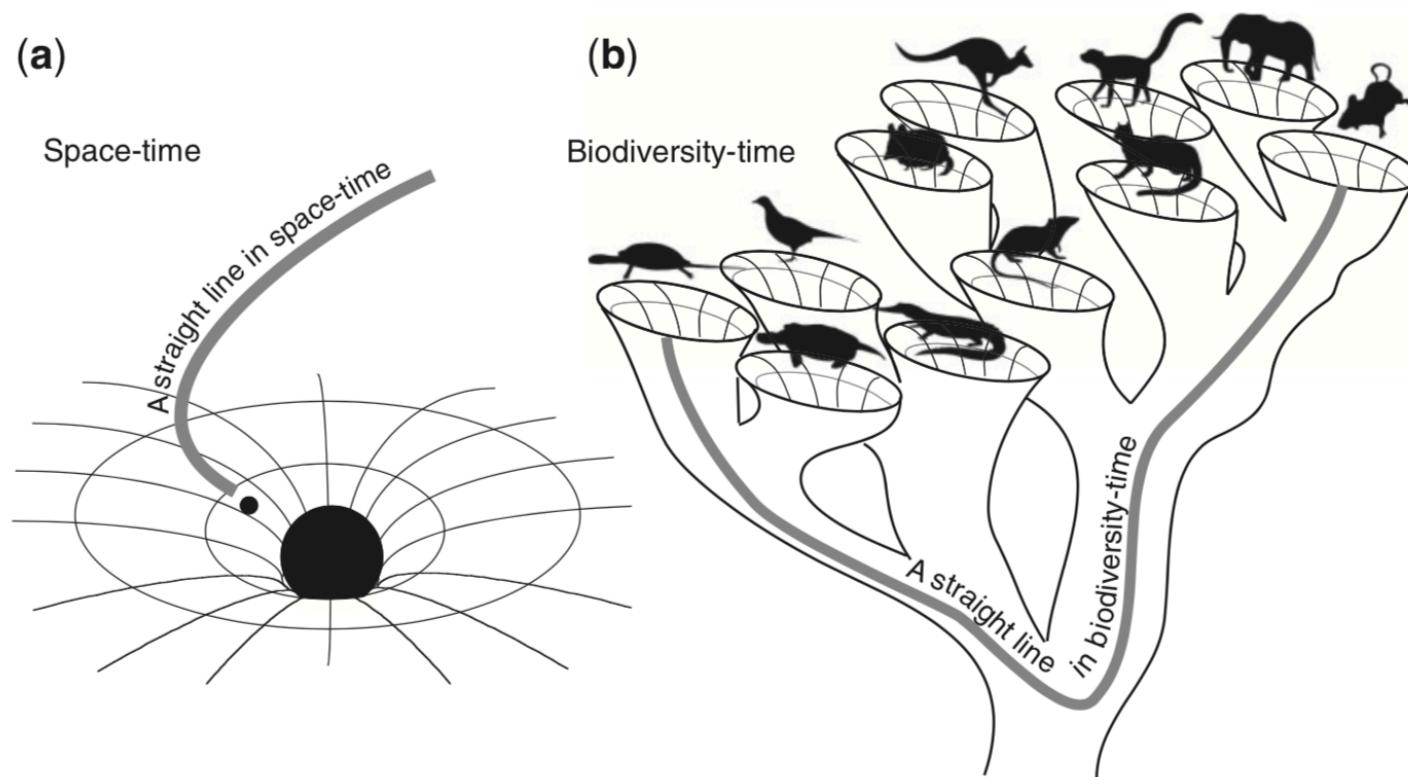
Un-replicated burst



Phylogenetic methods could be misleading !



We need to change our perspectives



Beware ...

Many actual PCM do not follow the biodiversity-time line correctly.

Before new methods are developed, users need to check whether their results are realistic and contain sufficient replication for being significant!

