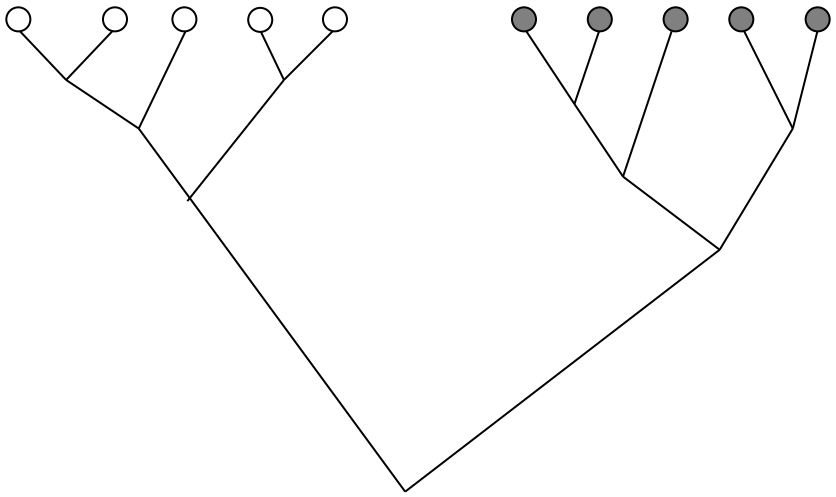
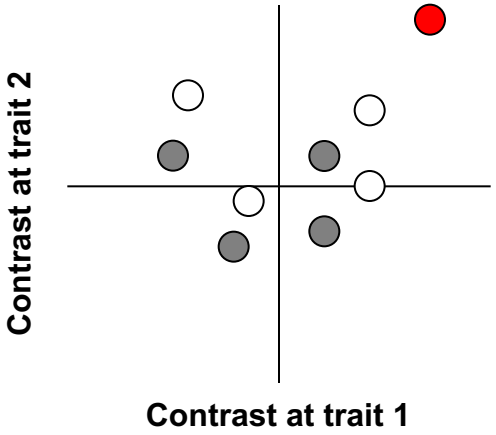
Phylogenies and statistics

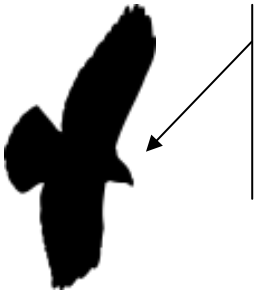




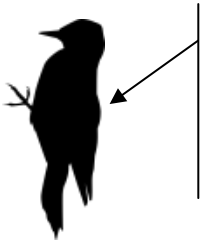
Simon Joly, Ph.D.
Montreal Botanical Garden
QCBS Biodiversity course | November 2017

An introduction

An introduction



- Flies
- Feathers
- 2 eyes



- Flies
- Feathers
- Narrow beak

- jumps
- skin
- 4 legs

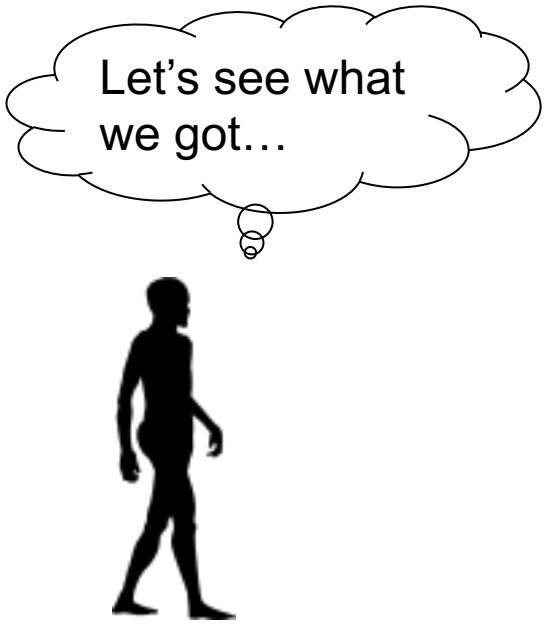
Let's take a walk…

- Walks
- fur
- 4 legs

- Walks
- fur
- 4 legs

Image credits: phylopic.org

An introduction



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

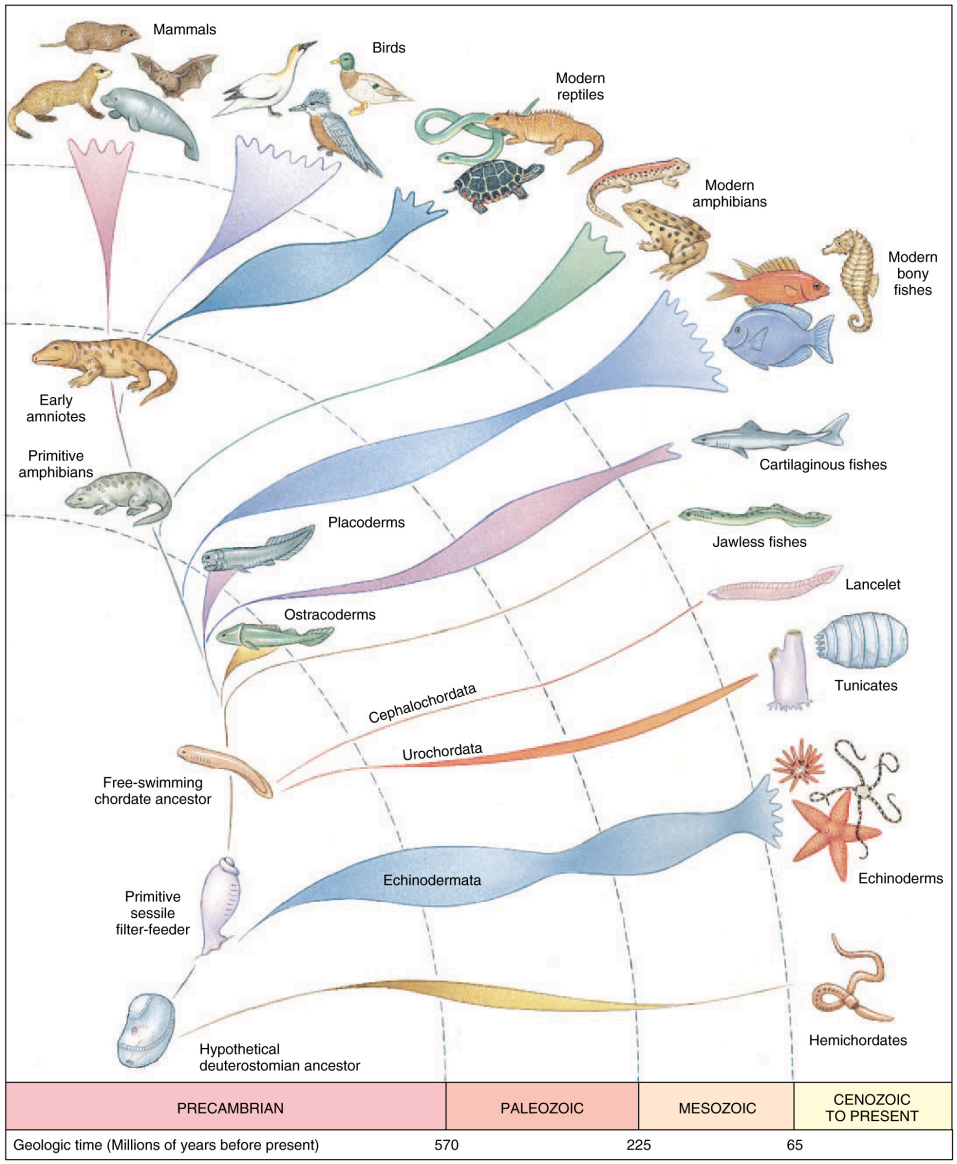
```
> fisher.test(matrix(c(9,0,0,7),2))

Fisher's Exact Test for Count Data

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

Image credits: phylopic.org

Evolution of flight (and feathers)



Credit: University Of Kelaniya, Sri Lanka

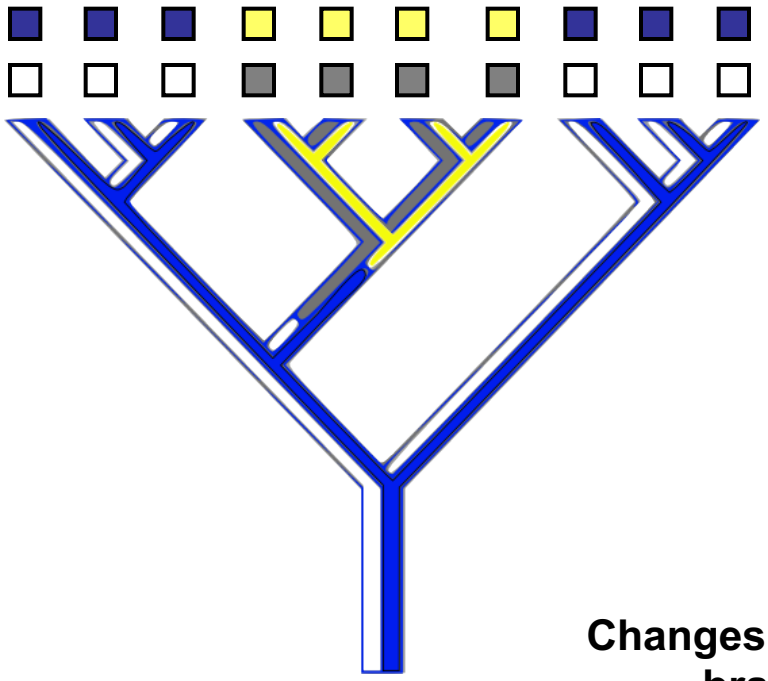
The comparative method

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

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

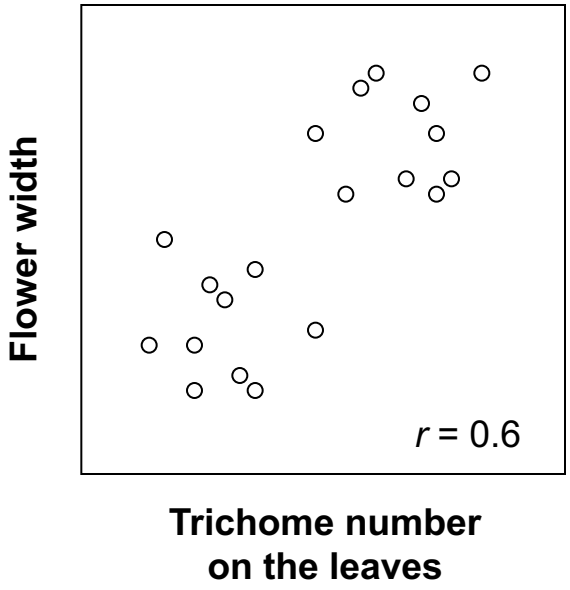
		flower: yes	flower: no
stem	yes	1	0
stem	no	0	17

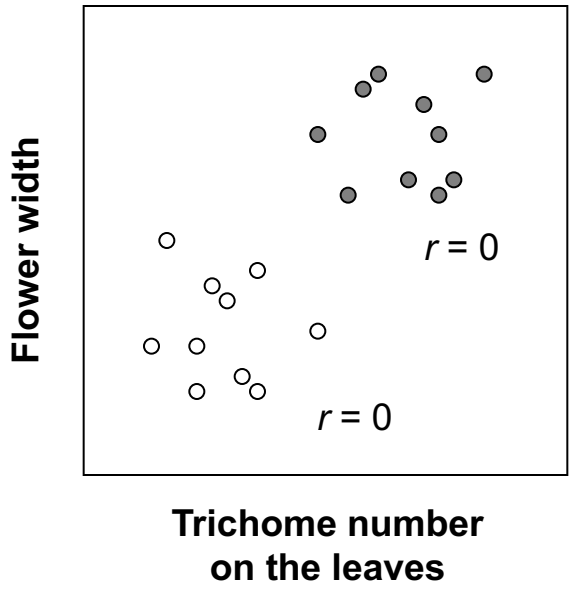
Fisher exact test
$p = 0,056$

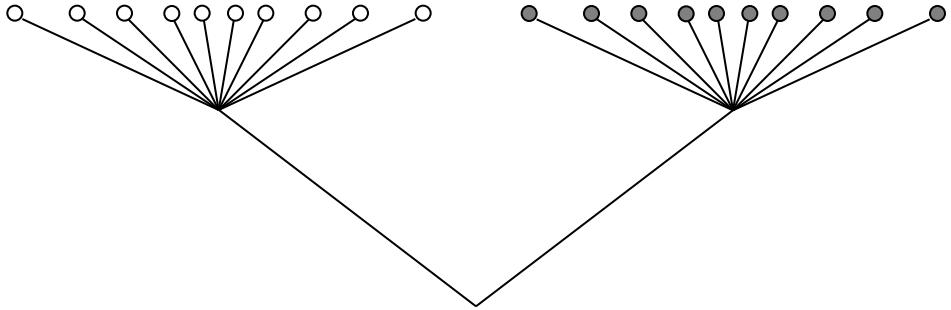
	blue	yellow
spiny	6	0
smooth	0	4

Fisher exact test
$p = 0.00476$

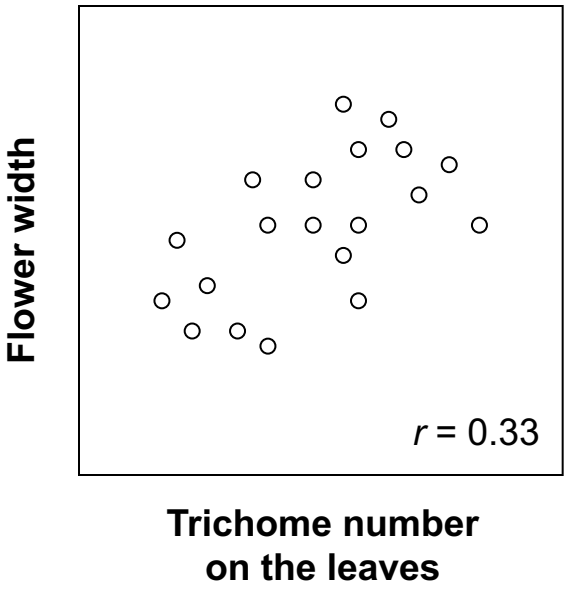
A quantitative example

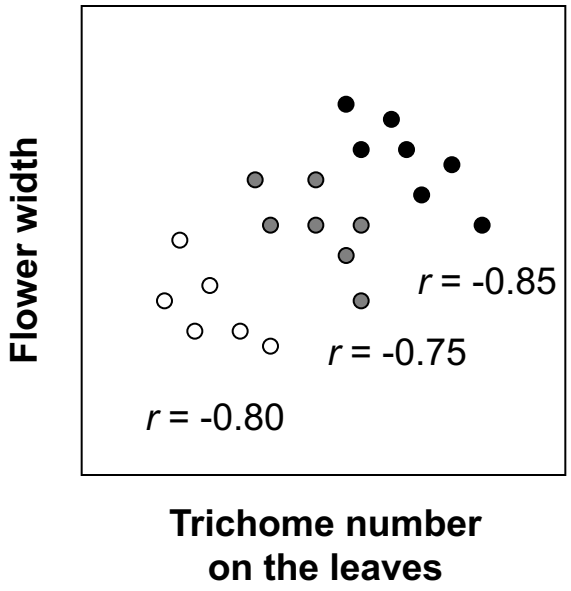


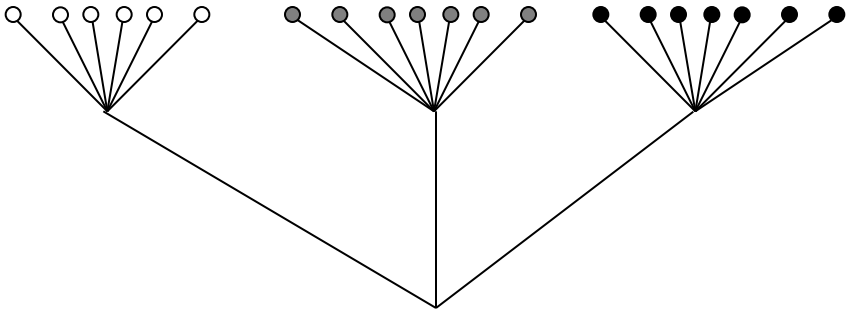




Another quantitative example





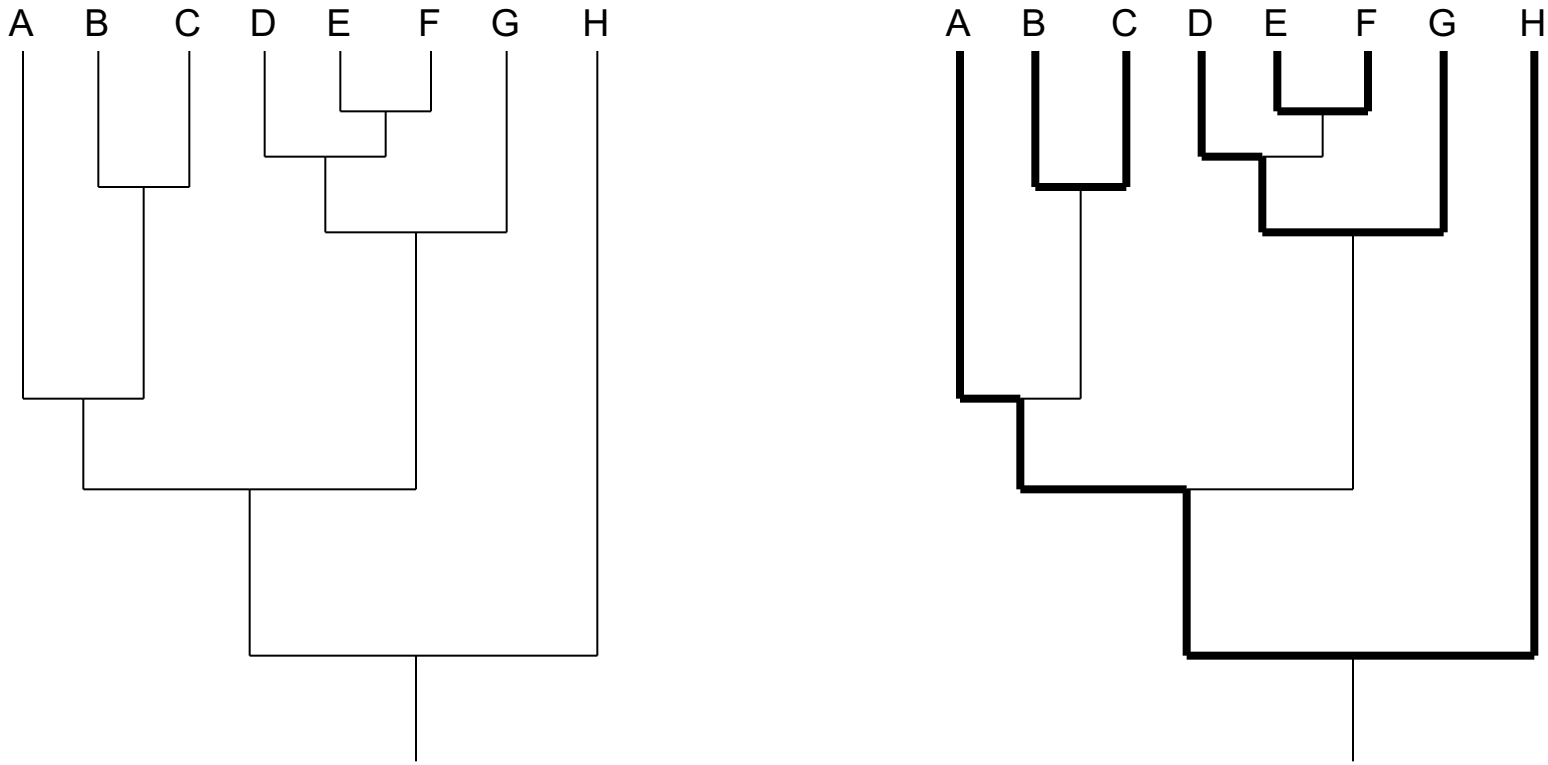


Comparative methods

Different approaches

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

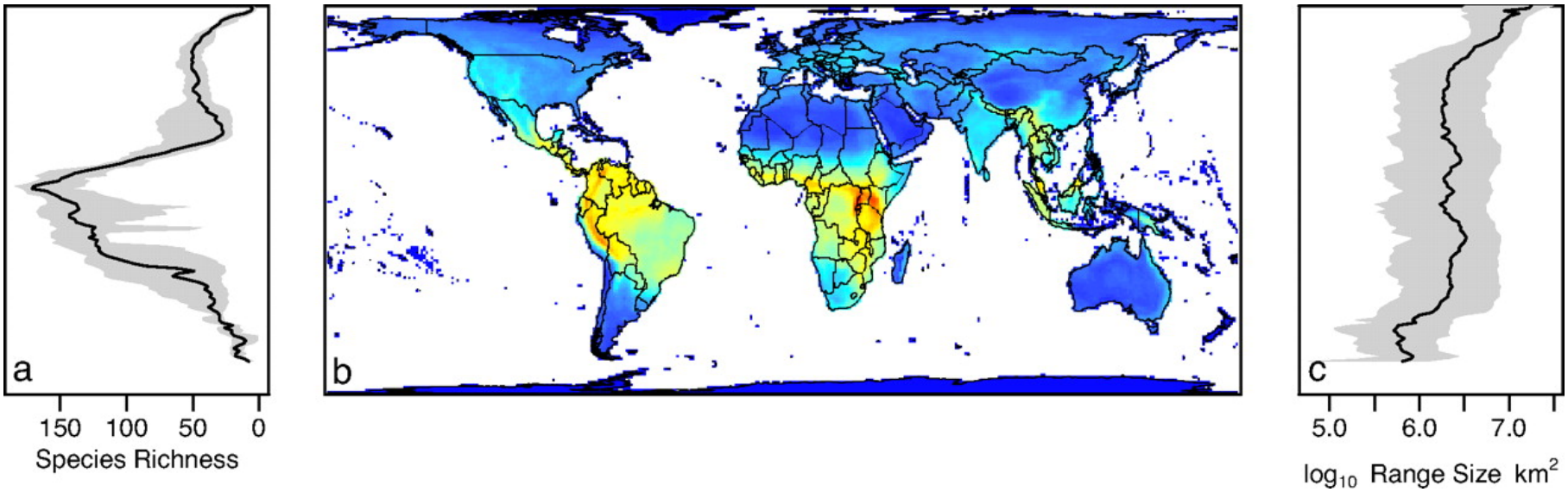
Pairs of species (sometimes sister clades)



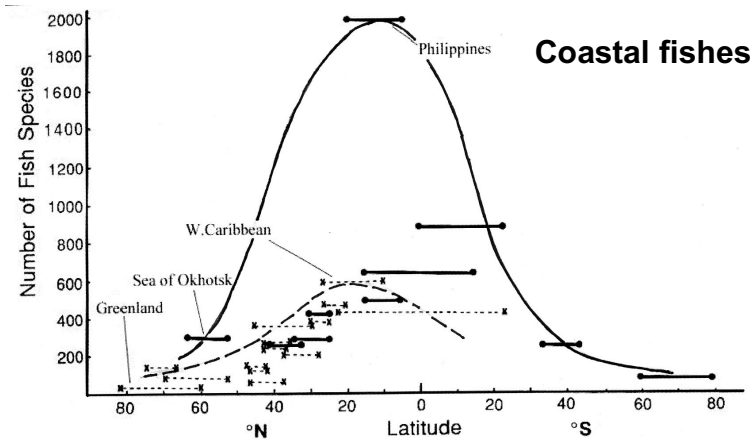
We can thus treat each comparison as independent from each other

from: Felsenstein (2004) Inferring phylogenies

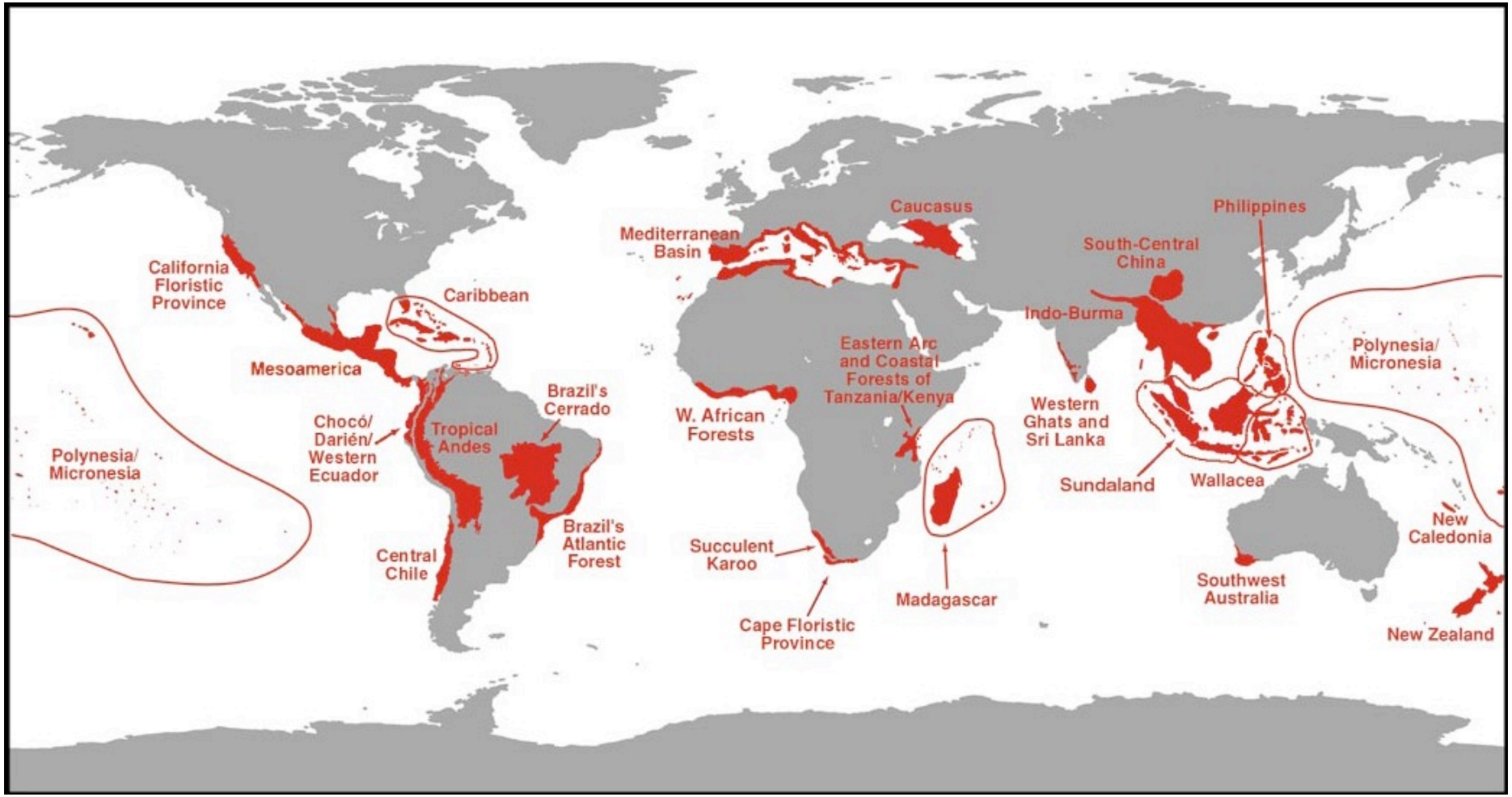
Biodiversity latitudinal gradient



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



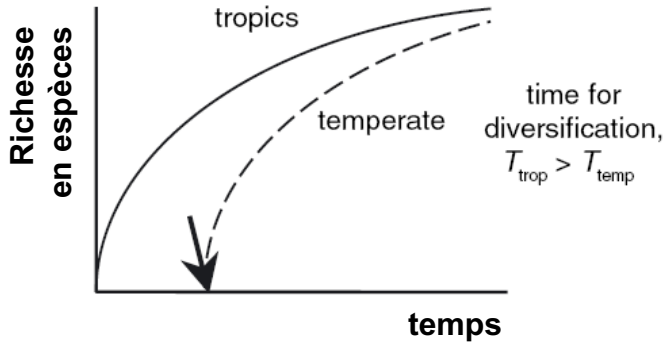
25 biodiversity hotspots



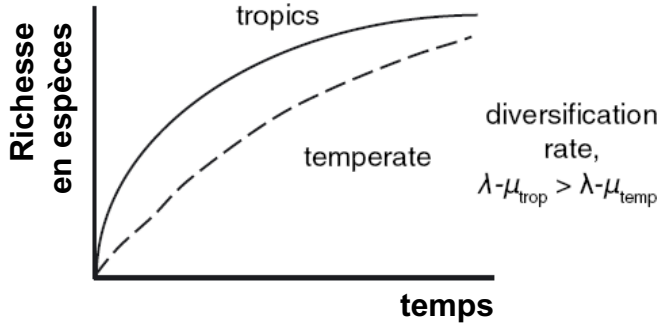
Myers et al. 2000. Nature.

Some evolutionary hypotheses

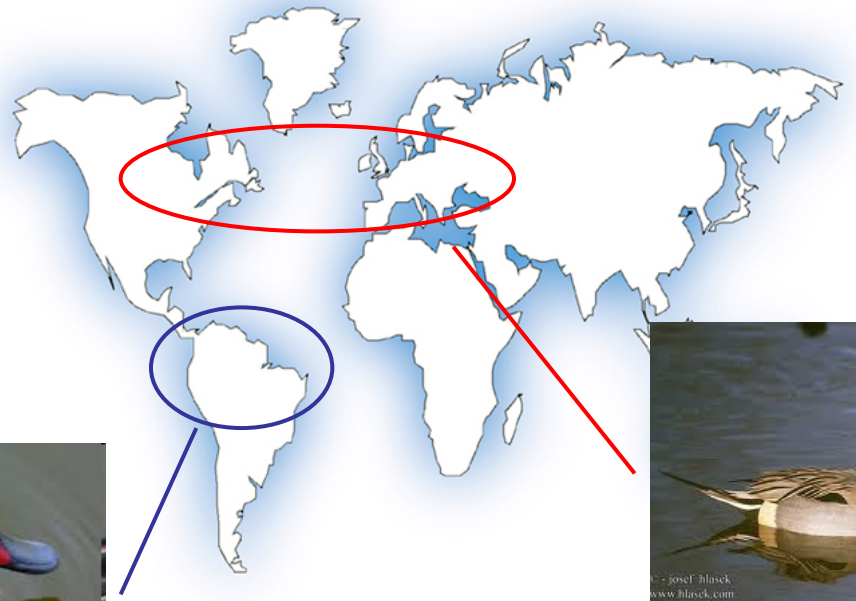
- More time for speciation in the tropics (older environments)

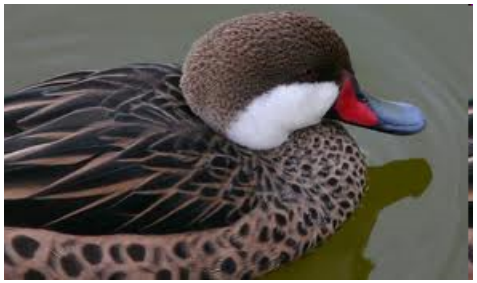


- Higher diversification rates in the tropics, due to:
  - Higher speciation rates in the tropics (Cradle theory)
  - Lower extinction rates in the tropics (Museum theory)



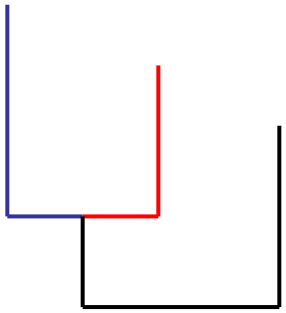
Test for higher diversification rate in the tropics





Anas bahamensis

Anas acuta



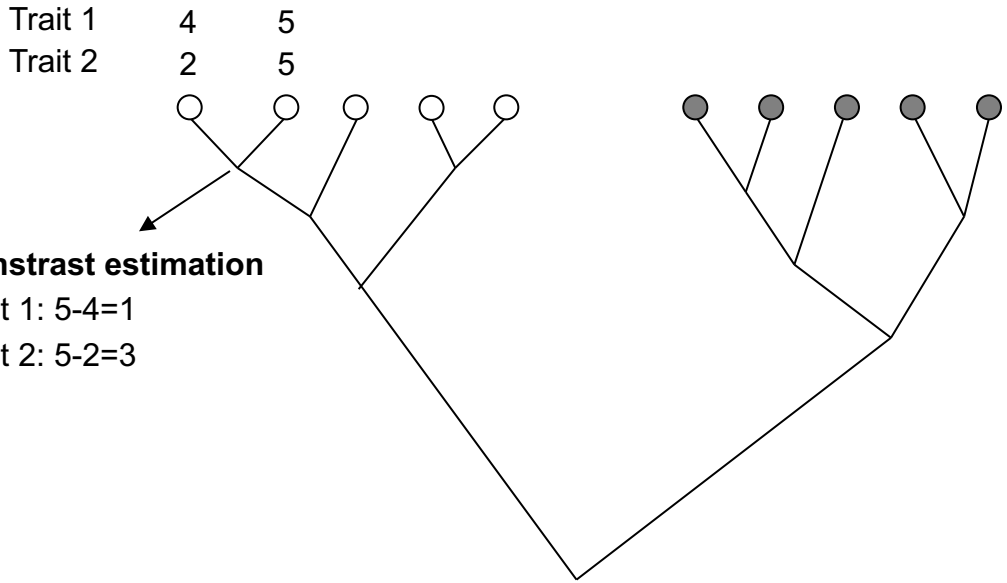
Number of contrasts	Number of positive contrasts	Sign test	Wilcoxon signed rank test
33	15	0.76	0.36

Not significant

Bromham & Cardillo (2003)

Independent contrasts

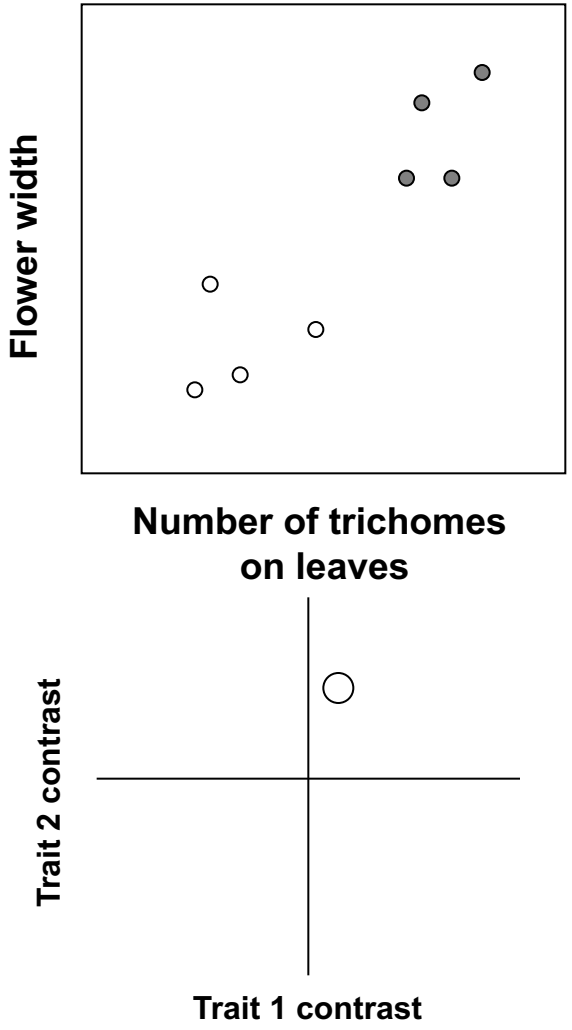
Objective: Test the correlation between characters while taking into account the evolutionary history of species



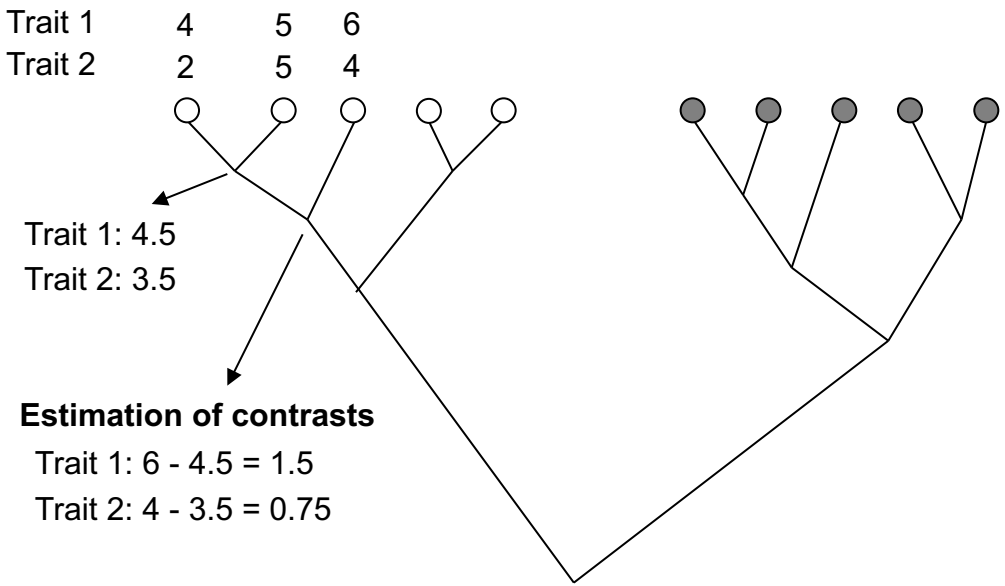
Constrast estimation

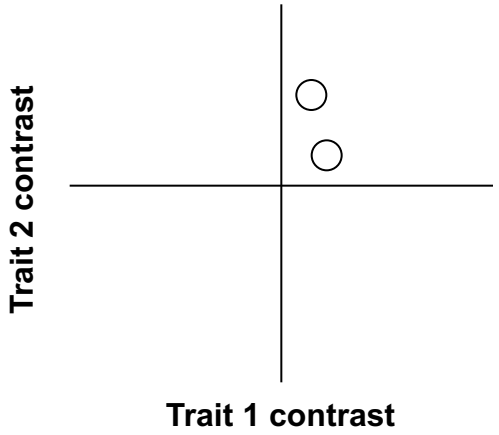
Trait 1: 5-4=1

Trait 2: 5-2=3

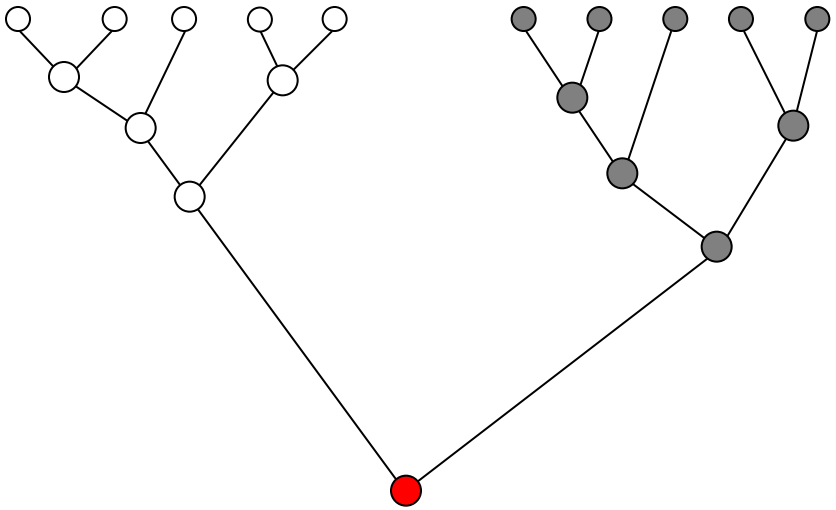


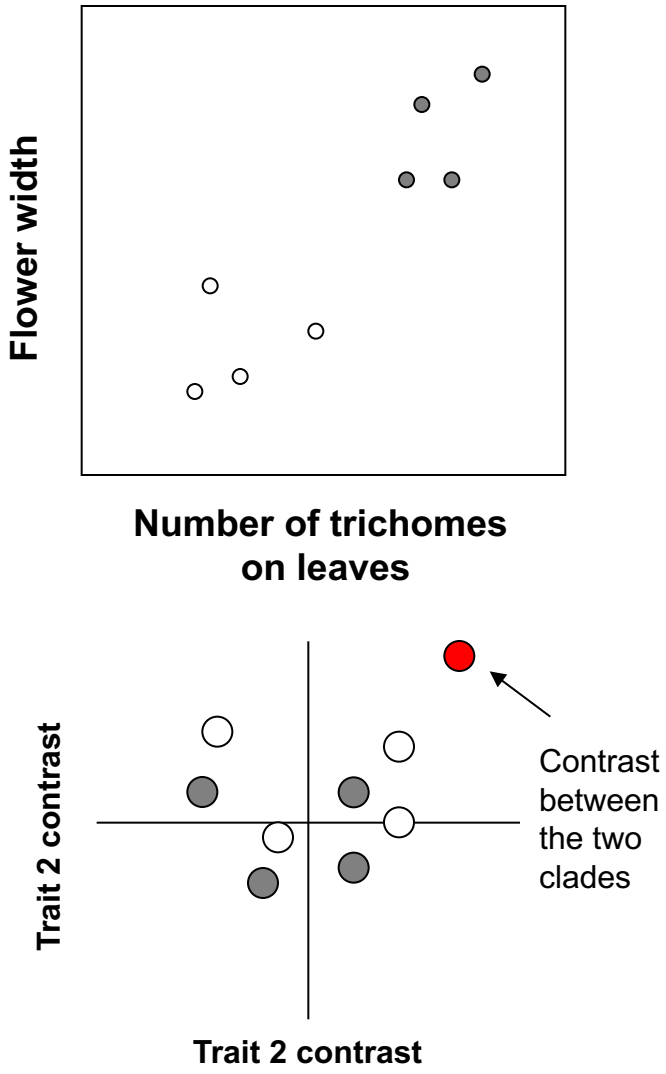
Independent contrasts





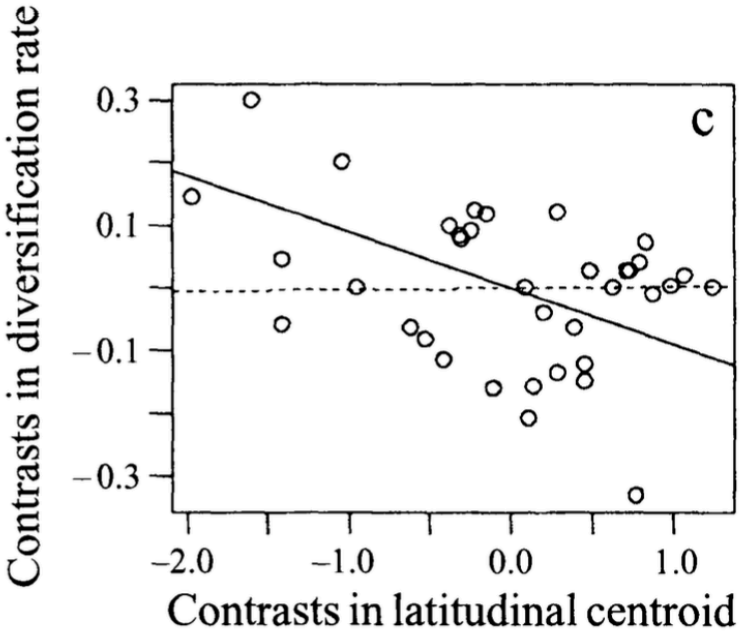
Independent contrasts





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

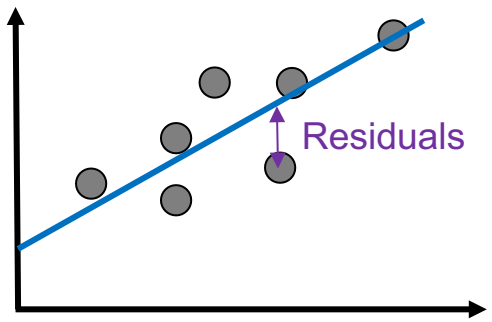
Higher diversification rates in the tropics

Cardillo et al. (2005)

Phylogenetic Generalized Least Squares

Linear model (ordinary least squares – OLS)

$$Y_i = \alpha + \beta_i X_i + \epsilon_i$$



Linear model (ordinary least squares – OLS)

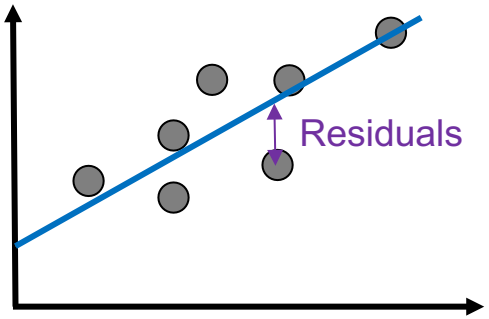
$$Y_i = \alpha + \beta_i X_i + \epsilon_i$$

Many assumptions

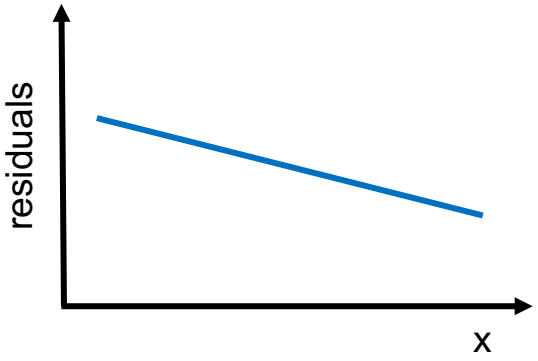
- Normality, homogeneity, fixed *X*, independence, and correct model specification
- Independence:
  - The value of Y_i at X_i should not be influenced by other values of X_i

$$\epsilon_i \sim N(0, \sigma^2)$$

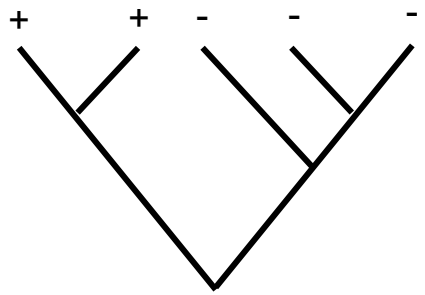
Examples of violations



Residuals correlated with the explanatory variable



Residuals correlated with the phylogenetic relationships



Practicals in R

www.github.com/simjoly/QCBS_workshop_2017

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)

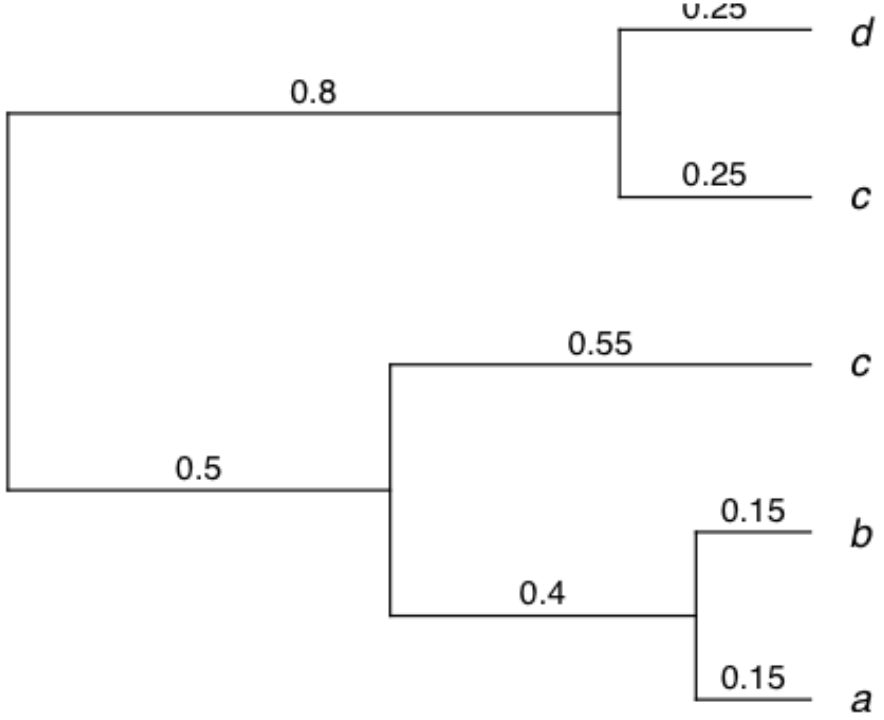
$$Y_i = \alpha + \beta_i X_i + \epsilon_i$$

$$\epsilon_i^{GLS} \sim MVN(0, \sigma^2 \mathbf{C})$$

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

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

$$\epsilon_i^{GLS} \sim MVN(0, \sigma^2 \mathbf{C})$$

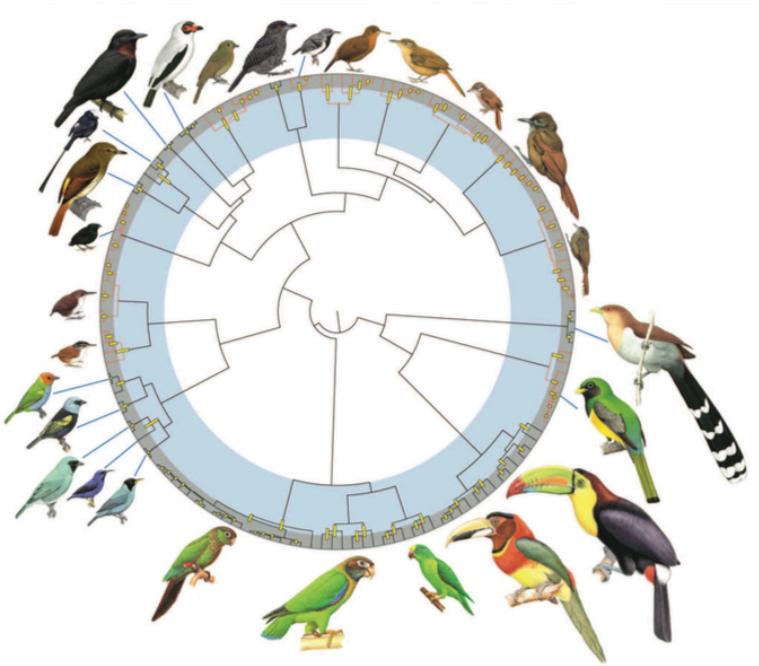
$$\sigma^2 \mathbf{C} = \sigma^2 \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1i} \\ \rho_{21} & 1 & \rho_{23} & \cdots & \rho_{2i} \\ \rho_{31} & \rho_{32} & 1 & \cdots & \rho_{3i} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{i1} & \rho_{i2} & \rho_{i3} & \cdots & 1 \end{pmatrix}$$

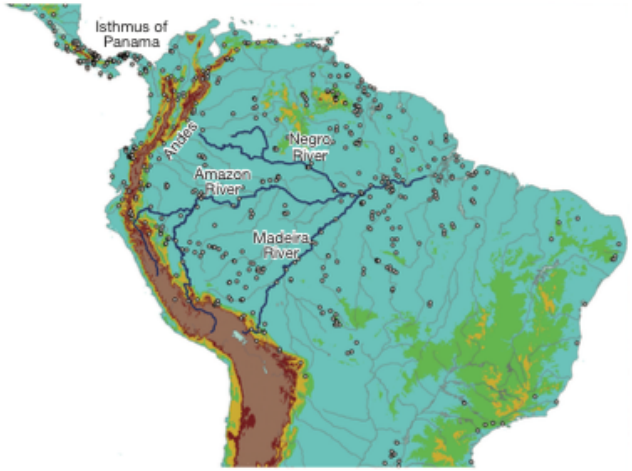
correlation matrix

Practicals in R

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





Smith et al. 2014. *Nature*

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*	ΔAICc
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

Smith et al. 2014. *Nature*

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

The solution

- What happens is there are other sources of errors in the data than the phylogenetic correlations?

$$Y_i = \alpha + \beta_i X_i + \epsilon_i$$

$$\epsilon_i^{GLS} \sim MVN(0, \sigma^2 \mathbf{C})$$

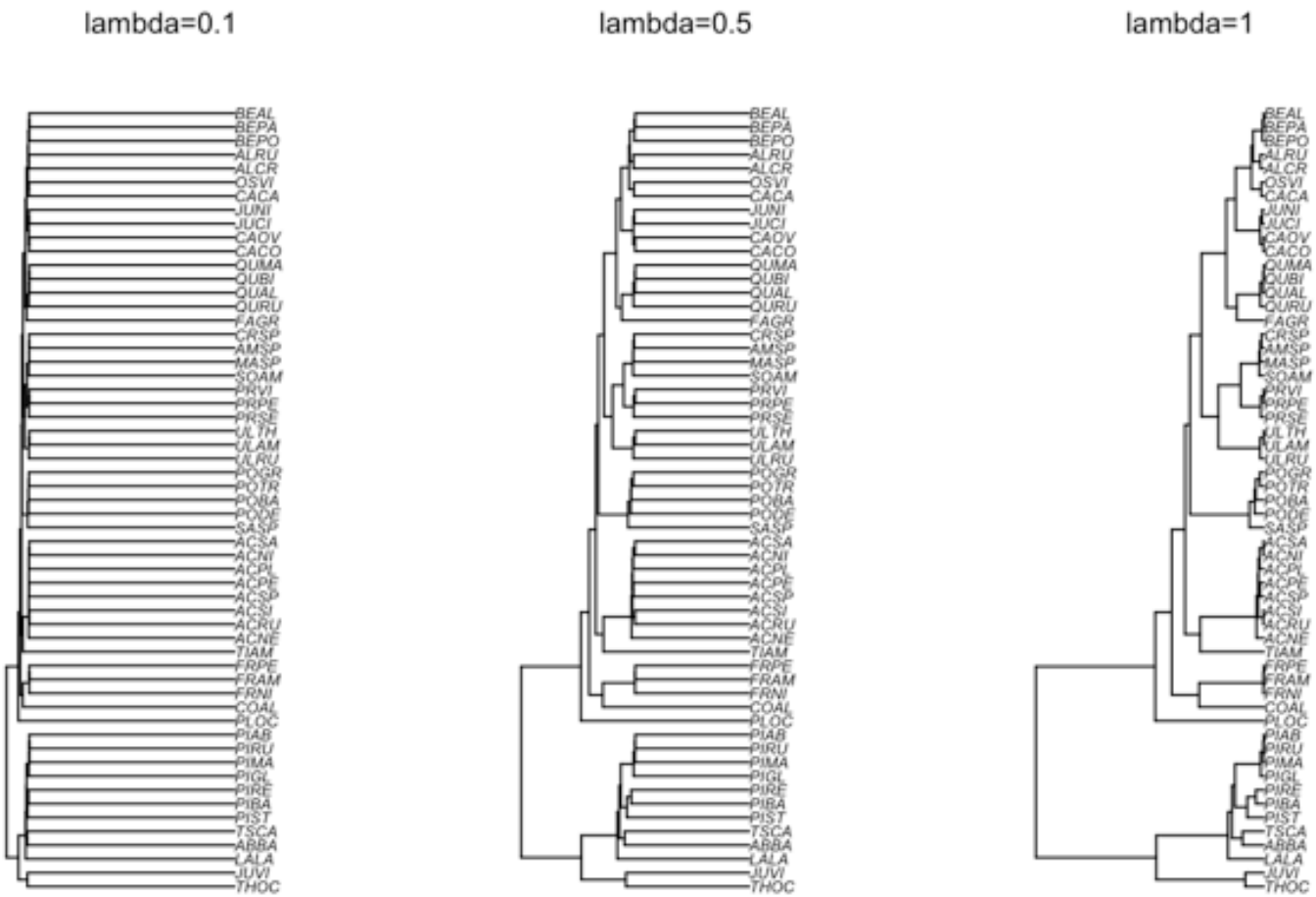
The solution: use the λ model

- 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

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

Pagel 1999 *Nature*

Effect of parameter λ on the phylogeny



Practicals in R

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

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

$$D = -2(lnL(D|M_0) - lnL(D|M_1))$$

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.

Aikake Information Criterion (AIC)

$$AIC = -2lnL(D|M) + 2k$$

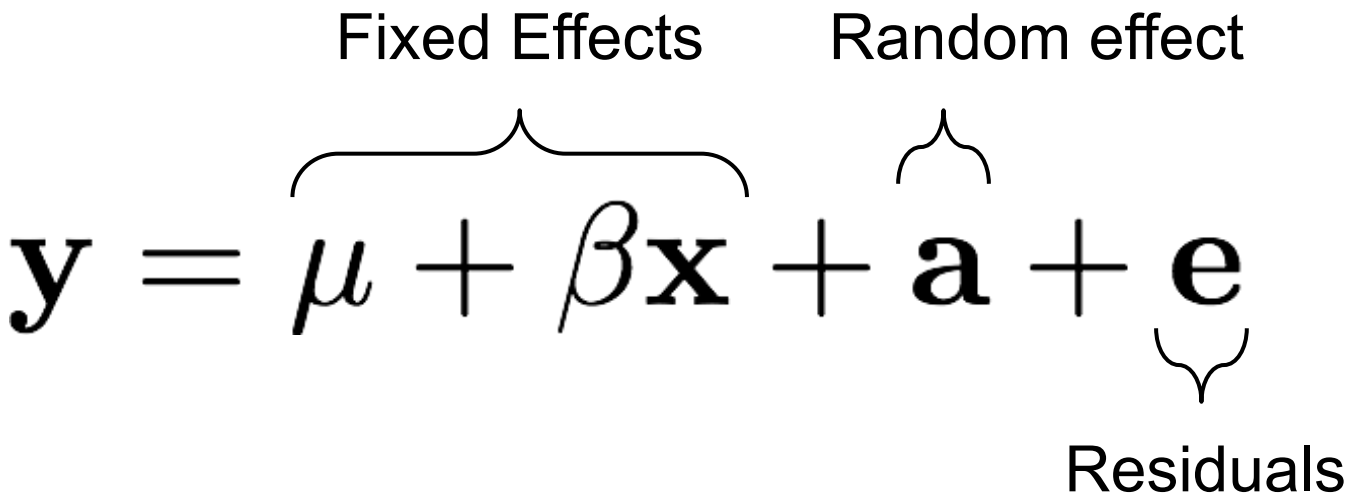
where *k* is the number of parameters in the model.

The best model is the one with the smaller AIC.

Practicals in R

The Phylogenetic Mixed Model

Mixed Model


Fixed Effects
Random effect
$\mathbf{y} = \mu + \beta\mathbf{x} + \mathbf{a} + \mathbf{e}$
Residuals

Fixed vs. random effects

- **Fixed effects**
  - The parameter you are interested in.
    - For instance, if you want to estimate the impact of temperature on generation times of frogs, then the temperature is a fixed effect.

- **Random effects**
  - Parameter that have been measured and that can affect you model. However, they are not of primary importance in your study.
    - e.g., Blocks in experimental settings, phylogenies!

The phylogenetic Mixed Model (PMM)

Fixed Effects Phylogenetic effect

$$\mathbf{y} = \overbrace{\mu + \beta\mathbf{x}} + \overbrace{\mathbf{a}} + \mathbf{e}$$

$$\mathbf{a} \sim \mathcal{N}(0, \sigma_a^2 \mathbf{A})$$

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

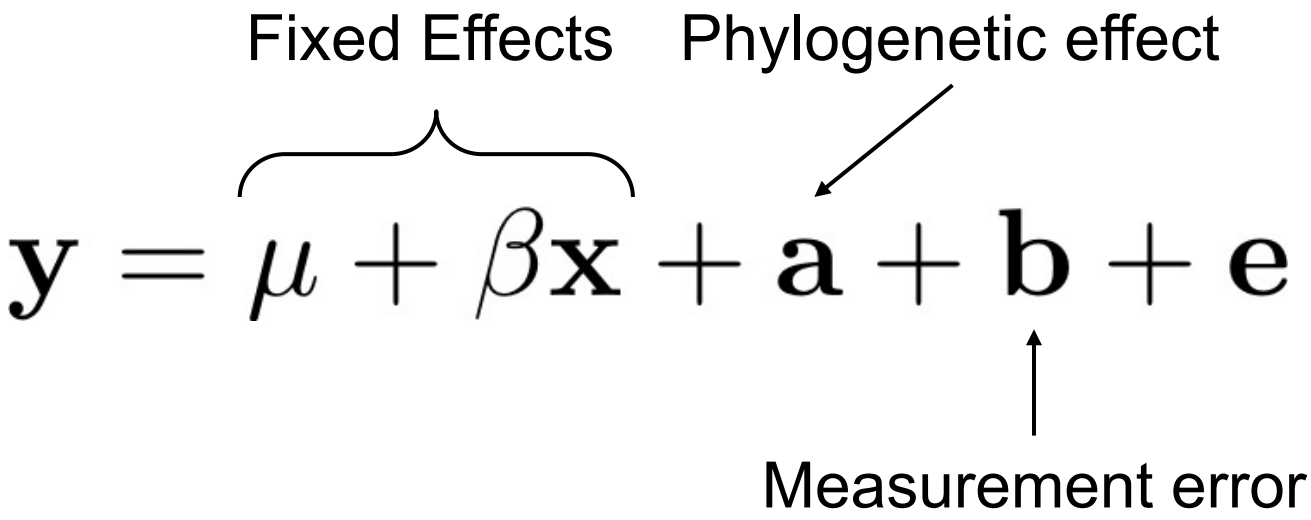
Heredity (or phylogenetic signal)

Variance structure

$$\mathbf{y} = \mu + \beta\mathbf{x} + \overbrace{\mathbf{a} + \mathbf{e}}$$

$$h^2 = \frac{\sigma_a^2}{(\sigma_a^2 + \sigma_e^2)}$$

Flexibility of the PMM



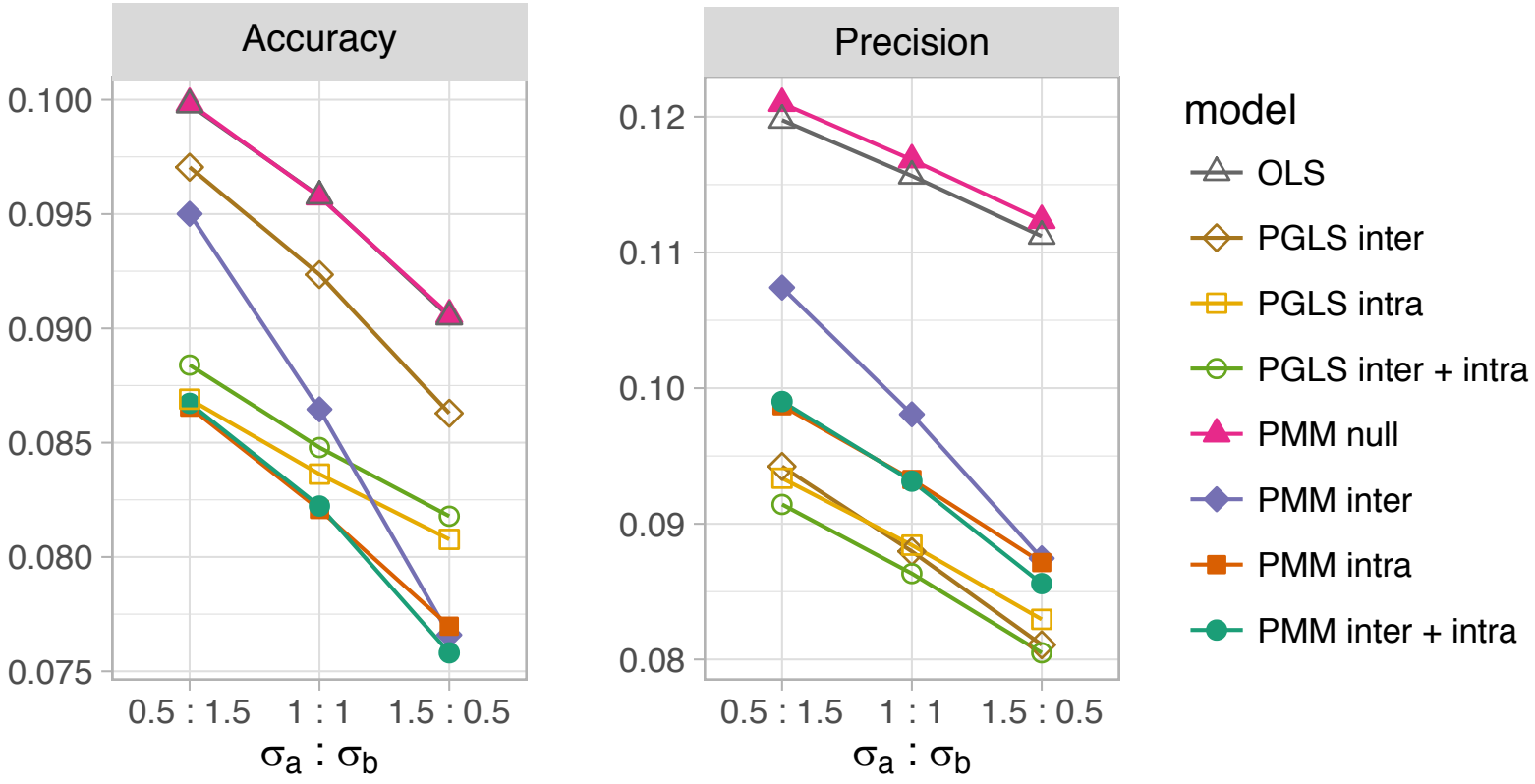
You can easily add more random effects. This is not possible with PGLS

Practicals in R

Why use comparative methods?

Why use comparative methods?

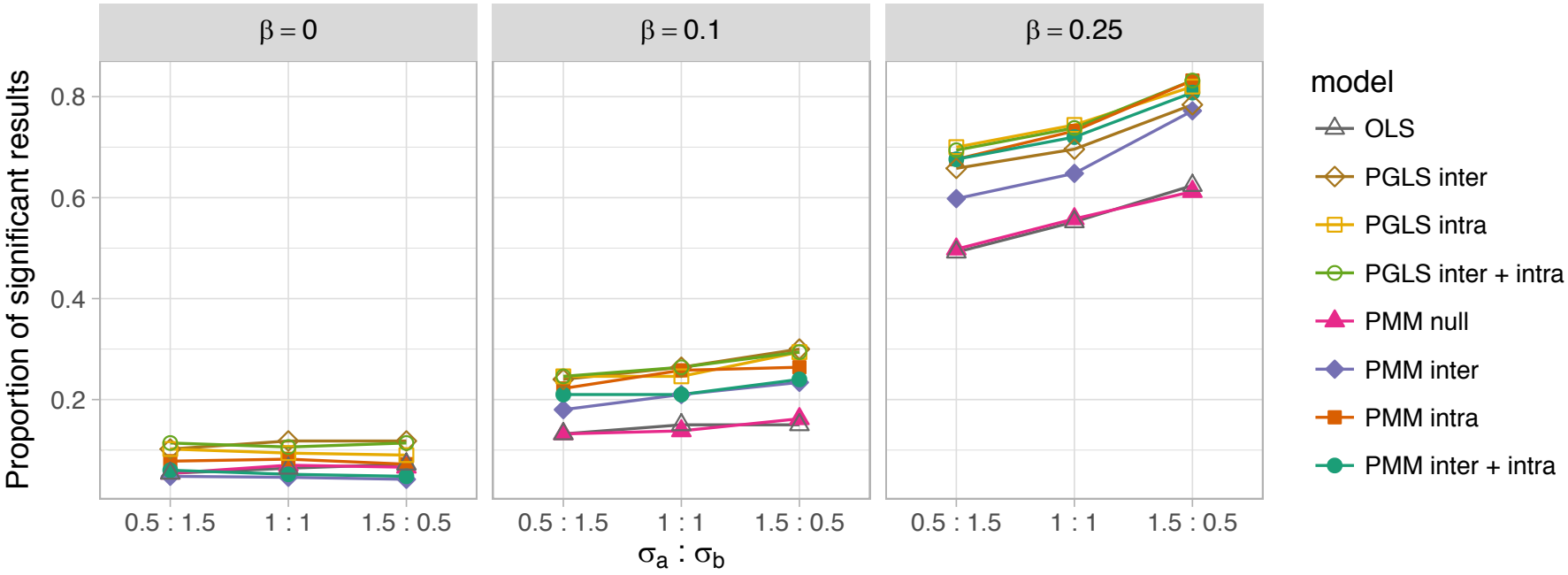
- Better accuracy and precision for fixed effects



Joly and Wolkovich submitted

Why use comparative methods?

- Lower type I error and higher power



Joly and Wolkovich submitted