Phylogenetic Comparative Methods
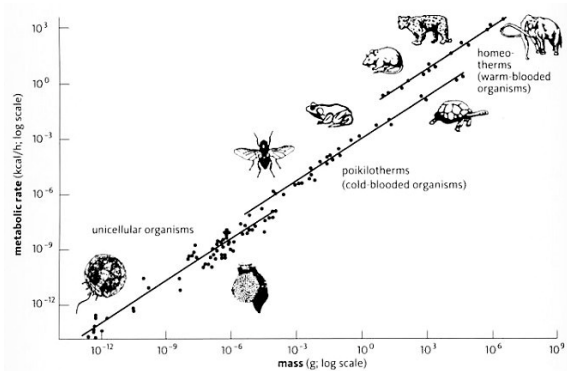
Advanced Biostatistics

Dean Adams
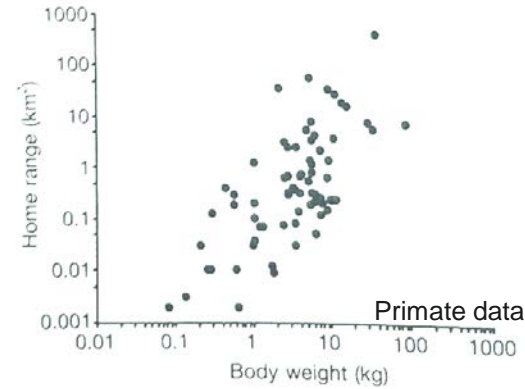
Lecture 12

EEOB 590C

Biologists examine trait correlations to infer adaptation & coevolution
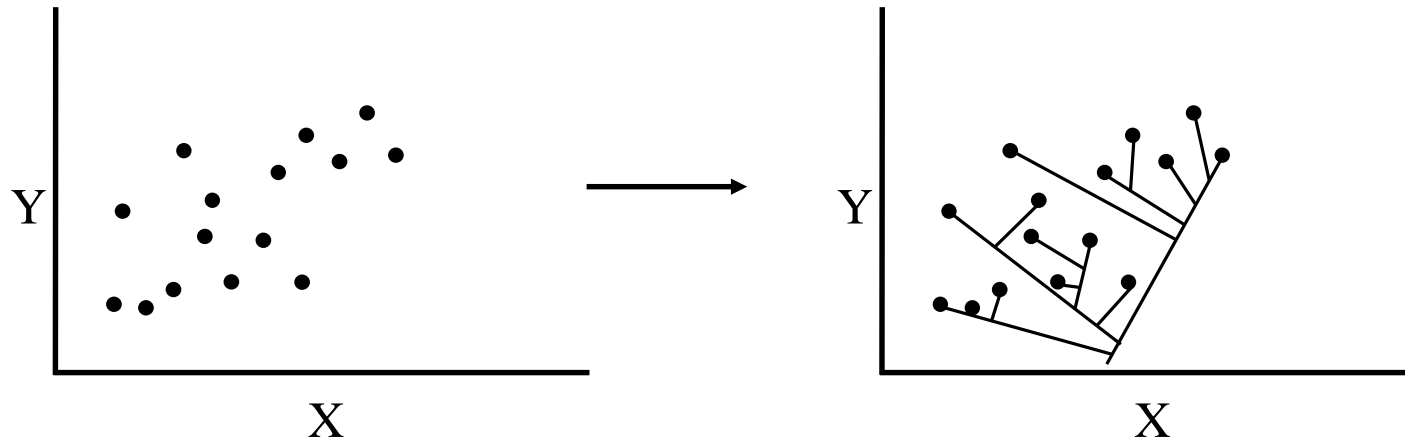

Hemmingsen, 1960

Primate data

Harvey & Clutton-Brock 1981

The problem: taxa are not independent

Phylogenetic comparative methods account for lack of independence

Incorporate phylogeny into statistical analyses

-1970s & 1980s

Early attempts: Nested ANOVA, Phylogenetic autocorrelation

Phylogenetic independent contrasts (PIC) (the breakthrough) **(1985)**

Phylogenetic generalized least squares (PGLS)

-1990s <u>(Niche expansion of H_0)</u>

Phylogenetic signal (λ) (*K* in 2003)

Phylogenetic ANOVA

OU1, ACDC, λ models

Discrete change models

-2000s & 2010s <u>(Maturation of tools)</u>

Synthesis: PIC/PGLS/Phylo-transform

Complex model comparisons (BM1 vs. BMM vs. OU1 vs OUM)

Bayesian approaches

Parameter-shift models (e.g., MEDUSA, BAMM)

Discrete models (Bisse, etc.)

Multivariate PCMs (PPCA, PPLS, $PGLS_{mult}$, phylomorphospace, etc.)

PCMs TODAY: A diverse toolkit for testing evolutionary hypotheses

General linear model expanded to include phylogenetics.

→ Phylogenetic regression/anova/association models

→ Phylogenetic Independent Contrasts
→ Phylogenetic Generalized Least Squares
→ Phylogenetic PLS

→ Phylogenetic ordination

→ Phylogenetic PCA

→ Phylomorphospace

→ Exploring and modelling evolutionary processes

→ Phylogenetic signal
→ Evolutionary rates
→ Evolutionary models

Methods for continuous data

Low power, high type 1 error, biased.

Phylogenetic Mantel test (for critique: Harmon and Glor 2010. *Evol.*)

Phylogenetic Autocorrelation (for critique: Rohlf 2001. *Evol.*)

Phylogenetic Eigenvector Regression (for critique: Adams and Church 2011. Ecography. Freckleton et al. 2011. *Am. Nat.*)

Phylogenetic Independent Contrasts

Phylogenetic Regression — All specific implementations of PGLS

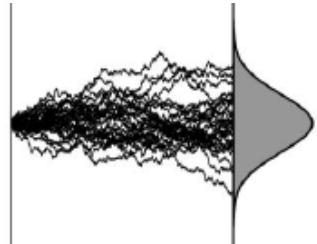Phylogenetic Generalized Least-Squares

The most commonly used approach to take phylogeny into account

Null model of trait change: Brownian motion (BM)

-No change in μ, but $s^2_y \uparrow \propto$ time $dY(t) = \sigma B(t)$

Method uses contrast scores at phylogeny nodes, which are *independent*

-Contrast scores between tips are their difference, standardized for the branch length separating them (μ=0; σ=1)

-Estimation proceeds tips → root, until (n-1) contrasts are obtained (for bifurcating tree)

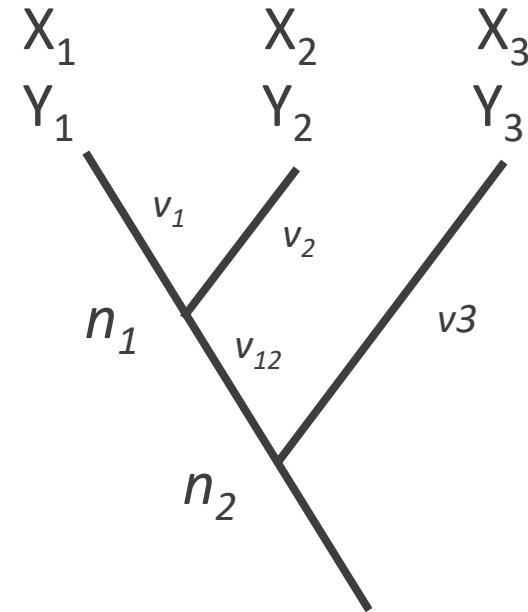Statistical analysis (regression) performed on contrasts rather than tips data

Felsenstein 1985. *Am. Nat.*

Contrast scores

$$Y_{ij} = \frac{Y_i - Y_j}{\sqrt{v_i + v_j}}$$

Internal nodes: weighted average

$$Y_{n1} = \frac{\frac{1}{v_1}Y_1 + \frac{1}{v_2}Y_2}{\frac{1}{v_1} + \frac{1}{v_2}}$$

X_1 X_2 X_3

Y_1 Y_2 Y_3

v_1 v_2

n_1 v_{12} $v3$

n_2

*NOTE: Internal branches adjusted as: $\quad v_{ij}^* = v_{ij} + \left(\dfrac{1}{1/v_i + 1/v_j} \right)$

Felsenstein 1985. *Am. Nat.*

$$c_1 = \frac{4 - 5}{\sqrt{1 + 1}}$$
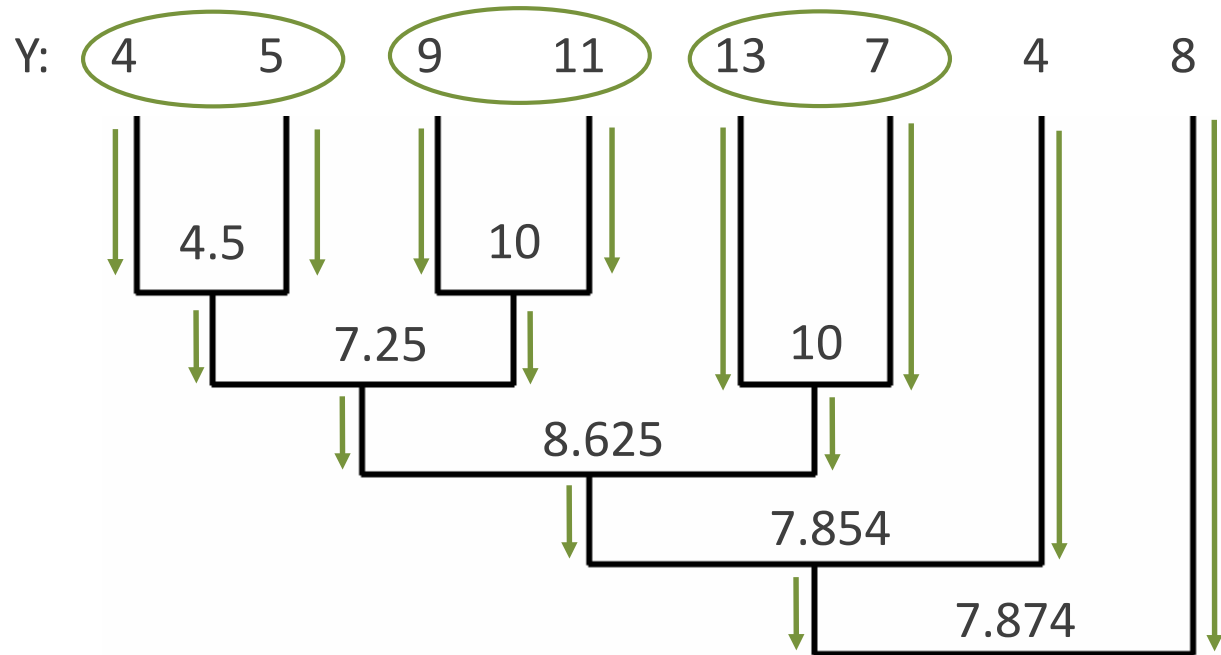
$$c_2 = \frac{9 - 11}{\sqrt{1 + 1}}$$

$$c_3 = \frac{13 - 7}{\sqrt{1.5 + 1.5}}$$

$$c_4 = \frac{4.5 - 10}{\sqrt{0.5 + 0.5}}$$

$$c_5 = \frac{7.25 - 10}{\sqrt{0.5 + 0.5}}$$

$$c_6 = \frac{8.625 - 4}{\sqrt{0.5 + 2.5}}$$

$$c_7 = \frac{7.854 - 8}{\sqrt{0.5 + 3}}$$

Y: 4 5   9 11   13 7   4   8
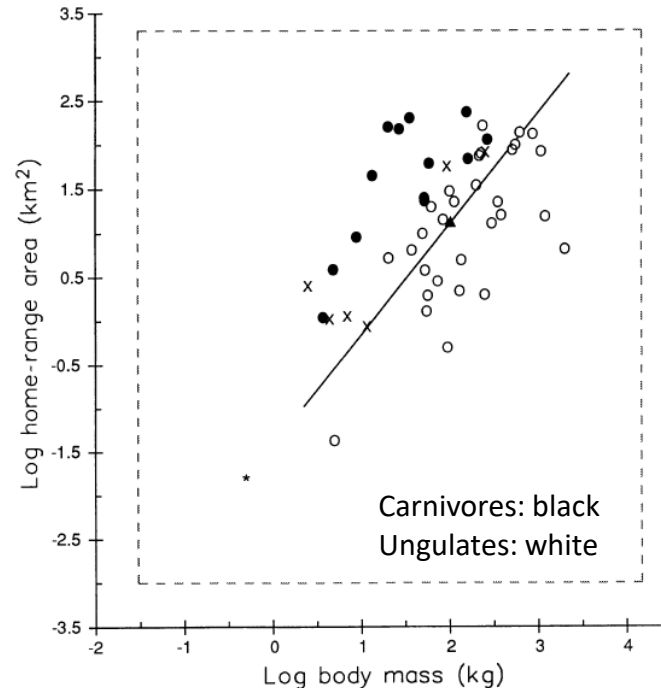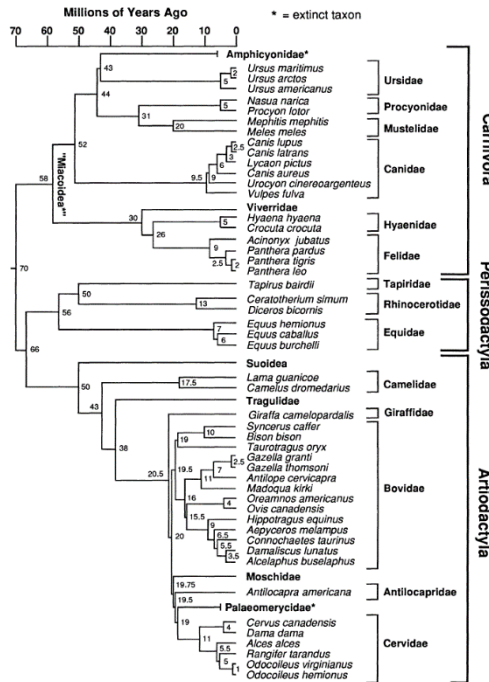
4.5   10   10

7.25

8.625

7.854

7.874

Contrast scores

$$Y_{ij} = \frac{Y_i - Y_j}{\sqrt{v_i + v_j}}$$

Internal nodes: weighted average

$$Y_{n1} = \frac{\frac{1}{v_1} Y_1 + \frac{1}{v_2} Y_2}{\frac{1}{v_1} + \frac{1}{v_2}}$$

Felsenstein 1985. *Am. Nat.*

Is there an evolutionary relationship between body size and range in mammals?

Carnivores: black
Ungulates: white

ANCOVA: Significant BM-HR slope
carnivore HR > ungulate HR
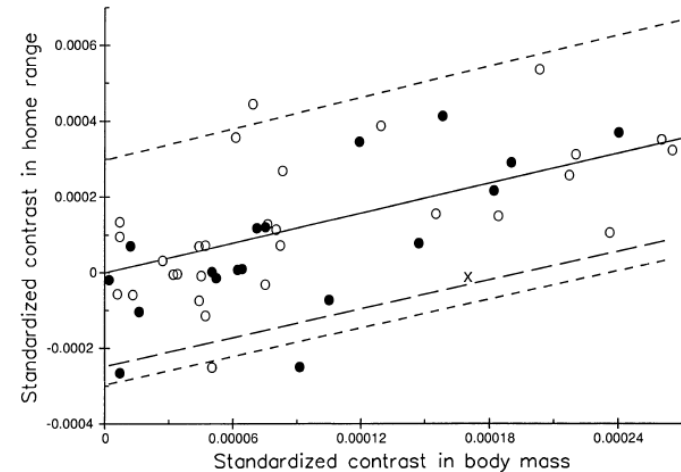
ANCOVA: Significant BM-HR regression
no HR differences

Garland et al. 1993, Syst.Biol.

Intuitive approach that follows a reasonable null model

The method is an algorithm with logical steps

Three or more branches from the same node.

Phylogenies with polytomies can be analysed by adding zero-length branches

 *note: anova-like designs can also be examined, but not simple to implement

(must perform contrasts on columns of the design matrix X)

PIC is a special case of Phylogenetic Generalized Least-Squares; a more general (and flexible) statistical model

Generalized Least-Squares (GLS) is the general form of GLM

$$\mathbf{Y} = \mathbf{X\beta} + \mathbf{\varepsilon}$$

In GLM, the error is iid: $\varepsilon \sim N(0,1)$

-GLS more flexible and can utilize other error structures $\varepsilon \sim N(0,\mathbf{V})$

   -Here, $\mathbf{V}$ describes the *expected* covariance among objects due to some model

For PGLS: $\mathbf{V}$ describes phylogenetic non-independence among species

PGLS is analogous to a 'weighted' GLM, where weights are inverse of structured error

Grafen, 1989. *Phil. Trans. Roy. Soc.*
Martins and Hansen, 1997. Am. Nat.
Rohlf 2001. Evol.

Under BM, trait variation $\propto$ to the rate of change σ^2 relative to the time species have evolved

-Thus, the expected covariance between species ε = **V**=σ^2**C**

- **C** is obtained from the branch lengths of the phylogeny
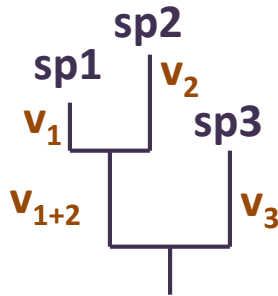
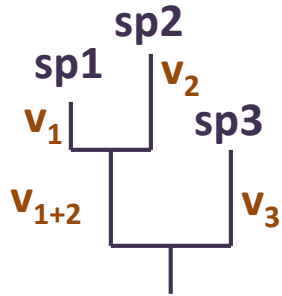$$C = \begin{pmatrix} v_1 + v_{1+2} & v_{1+2} & 0 \\ v_{1+2} & v_2 + v_{1+2} & 0 \\ 0 & 0 & v_3 \end{pmatrix} \begin{matrix} \text{sp1} \\ \text{sp2} \\ \text{sp3} \end{matrix}$$

sp1   sp2   sp3

C is the phylogenetic covariance matrix

1. Diagonal elements: species variance
(distance from the root in branch length)

2. Off-diagonal elements: covariance
(shared evolutionary path, from root to MRCA)

NOTE: **V** is different for other
evolutionary models

$$C = \begin{pmatrix} v_1 + v_{1+2} & v_{1+2} & 0 \\ v_{1+2} & v_2 + v_{1+2} & 0 \\ 0 & 0 & v_3 \end{pmatrix}$$

PGLS model:

GLS $\quad \mathbf{Y} = \mathbf{X\beta} + \mathbf{\varepsilon}$

Error $\quad \mathbf{\varepsilon} = \mathbf{V} = \sigma^2 \mathbf{C}$ $\quad$ σ^2: evolutionary rate following Brownian Motion (see later on)

Parameters of GLS may be found as:

$$\mathbf{\beta} = (\mathbf{X^t C^{-1} X})^{-1} \mathbf{X^t C^{-1} Y}$$

COMMENTS:
-Independent contrasts are a special case of PGLS

-Method is general:
  -Polytomies easily accommodated
  -Regression, ANOVA, and other scenarios possible
  -Other evolutionary models (e.g., OU) implemented with change to **V**

OLS comparative model:

$$\beta = \left(X^t X\right)^{-1} X^t Y$$

OLS is an unweighted model:

$$\beta = \left(X^t C^{-1} X\right)^{-1} X^t C^{-1} Y$$

$$C = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

PGLS is a weighted model:

$$\beta = \left(X^t C^{-1} X\right)^{-1} X^t C^{-1} Y$$
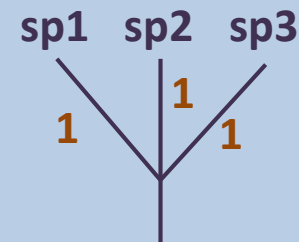
$$C = \begin{pmatrix} v_1 + v_{1+2} & v_{1+2} & 0 \\ v_{1+2} & v_2 + v_{1+2} & 0 \\ 0 & 0 & v_3 \end{pmatrix}$$

In PGLS, the weights are the phylogenetic distances, which describe the phylogenetic non-independence

Attention! Not taking phylogeny into account, corresponds to assuming a star phylogeny

$$C = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Several implementations of PGLS are common

1: ML: Fit data to model, find β using ML optimization
  -e.g. 'optim' in R

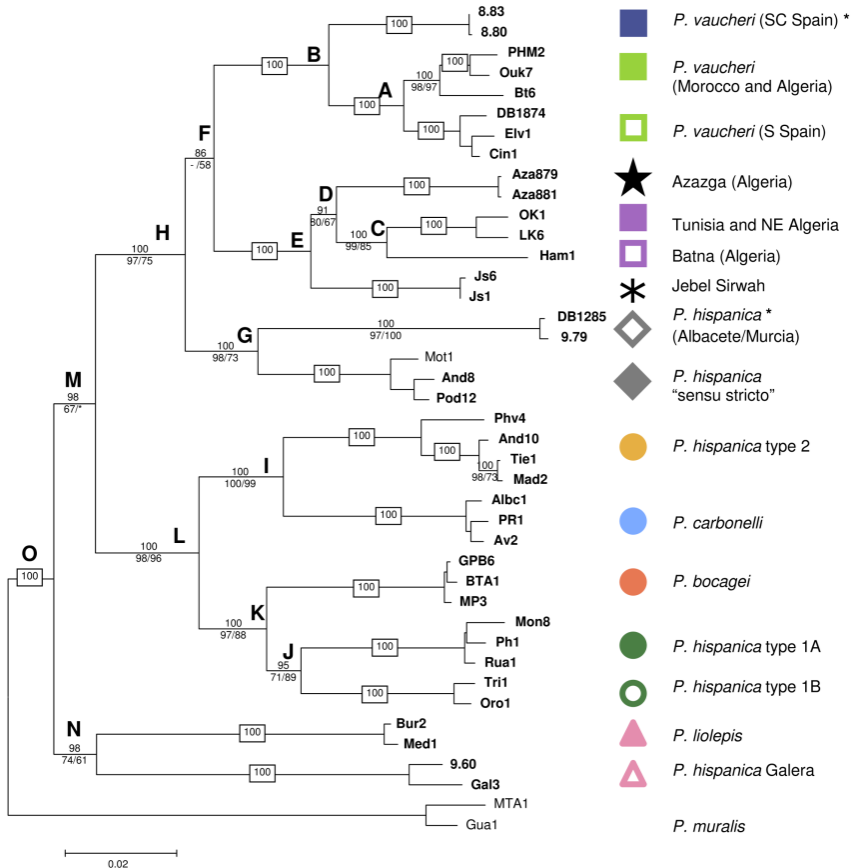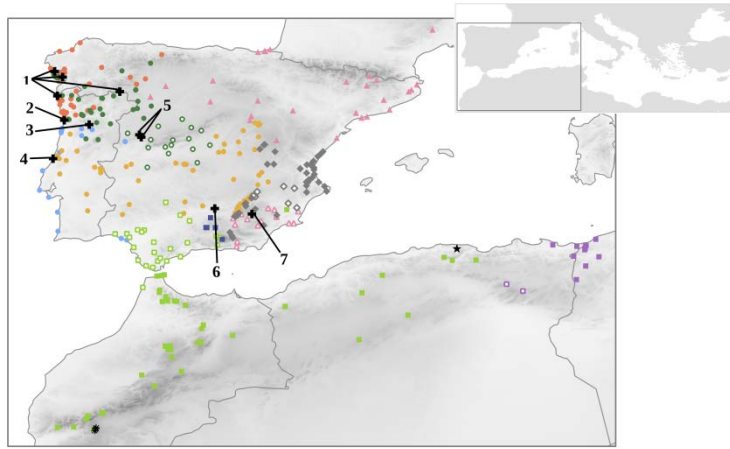2: Algebraic I GLS: Fit model and find β using weighted GLM

$$\boldsymbol{\beta} = \left(\mathbf{X^t C^{-1} X} \right)^{-1} \mathbf{X^t C^{-1} Y}$$

3: Algebraic II: Phylogenetic transform: Transform X & Y by the phylogeny, find β using OLS

$$\mathbf{C} = \mathbf{UWU}^{-1}$$

$$\mathbf{P} = (\mathbf{UW^{1/2}U^T})^{-1}$$

$$\boldsymbol{\beta} = \left((\mathbf{PX})^t \, \mathbf{PX} \right)^{-1} (\mathbf{PX})^t \, \mathbf{PY}$$

Head height varies extensively

Different species live in different structural habitats

Flatter body plan is advantageous for saxicolous species

Does habitat use explain variation in head height?

Kaliontzopoulou et al. 2015, *J. Evol. Biol.*

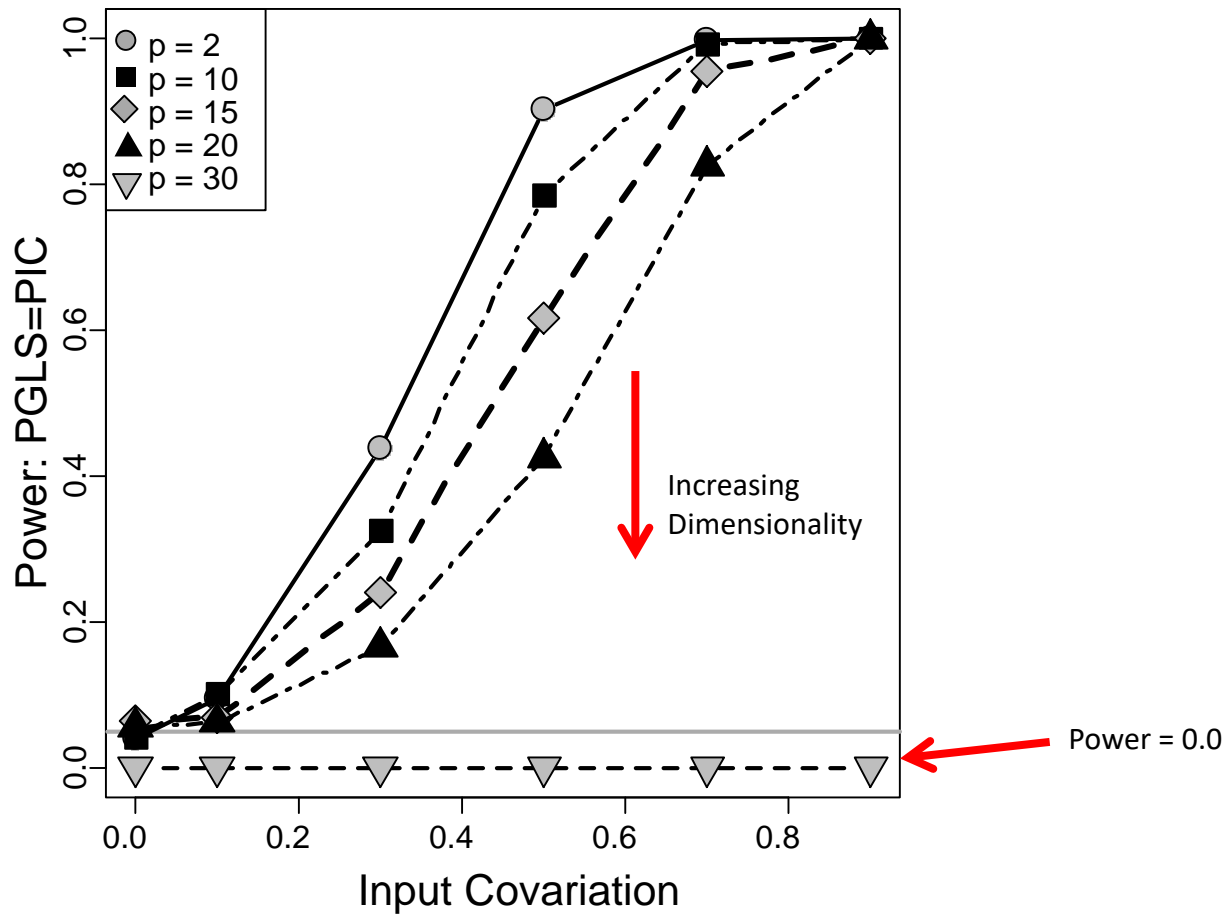1) Date the tree and make it ultrametric

2) Each species is classified depending on habitat use

3) Calculate expected covariance matrix (C)

4) Run PGLS

Kaliontzopoulou et al. 2015, *J. Evol. Biol.*
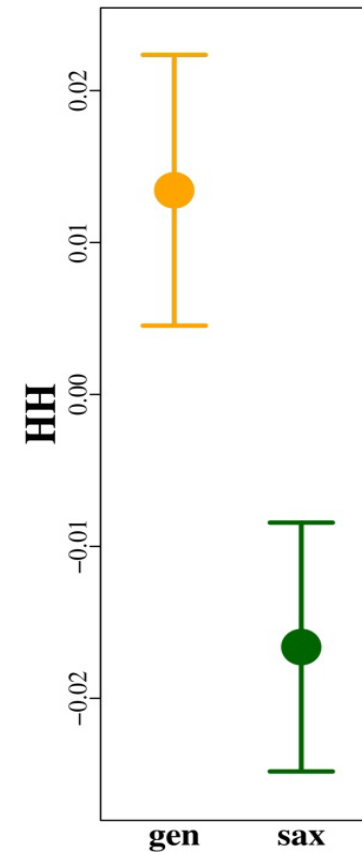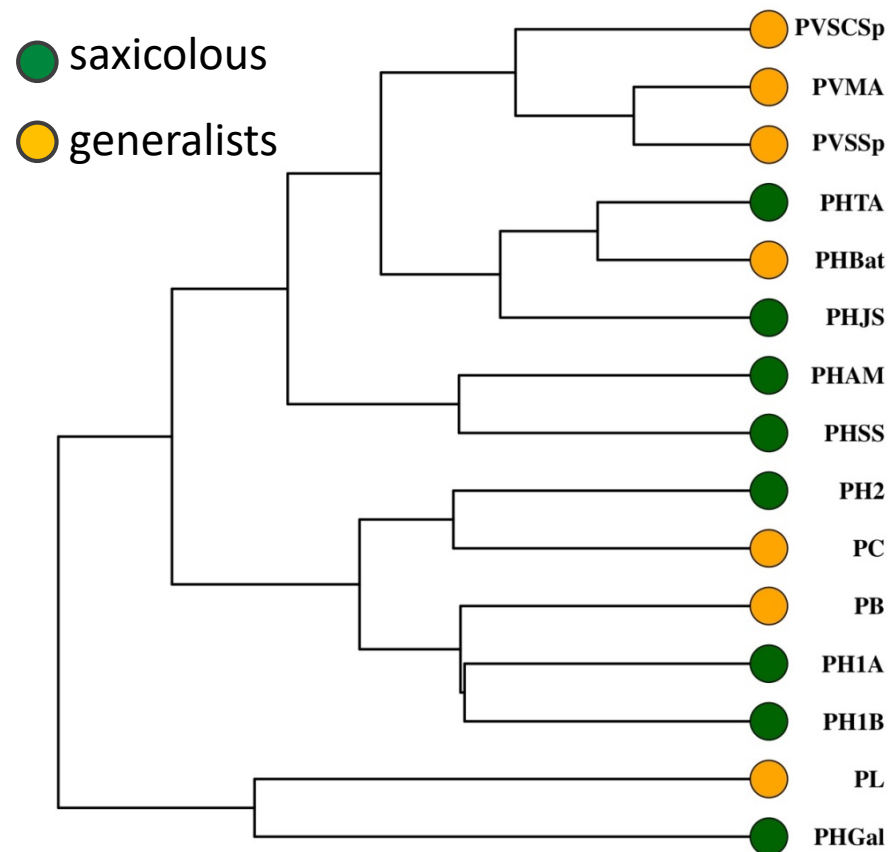
Method directly extendable to multivariate

However parametric PIC and PGLS suffer from Rao's paradox with high-D data

For multivariate data, an alternative is required

Adams 2014, *Evolution*.

Phylogenetic transformation + perm-manova generalizes PGLS

Method projects data into a transformed space where data are independent of phylogeny

D-PGLS steps:

**NOTE: Identical to *PGLS* for univariate

-Decompose phy. Cov. Mat: $\mathbf{C} = \mathbf{U}\mathbf{W}\mathbf{U}^{-1}$

-Obtain transformation matrix: $\mathbf{P} = (\mathbf{U}\mathbf{W}^{1/2}\mathbf{U}^{t})^{-1}$

-Transform data: $\mathbf{Y}_{phy} = \mathbf{P}\mathbf{Y}$ $\qquad \mathbf{X}_{phy} = \mathbf{P}\mathbf{X}$

-Fit model: $\widehat{\mathbf{Y}}_X = predict(Y_{phy} \sim X_{phy})$

-E.g,. SS$_X$: $tr\left(\left(\widehat{\mathbf{Y}}_X - \widehat{\mathbf{Y}}_1\right)\left(\widehat{\mathbf{Y}}_X - \widehat{\mathbf{Y}}_1\right)^t\right)$ $\qquad$ Note: tr(=$\mathbf{Y}^t\mathbf{Y}$) = tr($\mathbf{Y}\mathbf{Y}^t$)
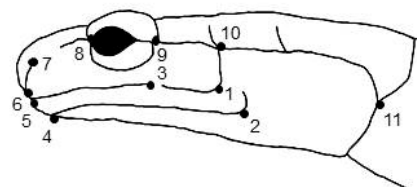
-Obtain SS, MS, R^2 from model

-Permute Y, repeat steps

**NOTE: multivariate PIC also proposed, where PIC are obtained and PIC$_y$ permuted (Klingenberg and Marugán-Lobón 2013. *Syst. Biol*)

Adams 2014, *Evolution*.

Which implementation to use: PIC_rand or D-PGLS?

Y~X|phy

vs. SVL

• Identical SS, MS, F, R^2

D-PGLS	df	SS	MS	F	R^2	P_{Yrand}	
SVL	1	0.0006586	0.0006586	3.0288	0.07039	0.221 NS	
Residual	40	0.0086976	0.0086976				
Total	41	0.0093562	0.00021744				

PIC	df	SS	MS	F	R^2	$P_{PICrand}$	P_{Yrand}
SVL	1	0.0006586	0.0006586	3.0288	0.07039	0.026	0.221 NS
Residual	40	0.0086976	0.0086976				
Total	41	0.0093562	0.00021744				

Why difference in significance levels?

Adams and Collyer 2015. *Evol.*

Must shuffle the exchangeable units under H_0
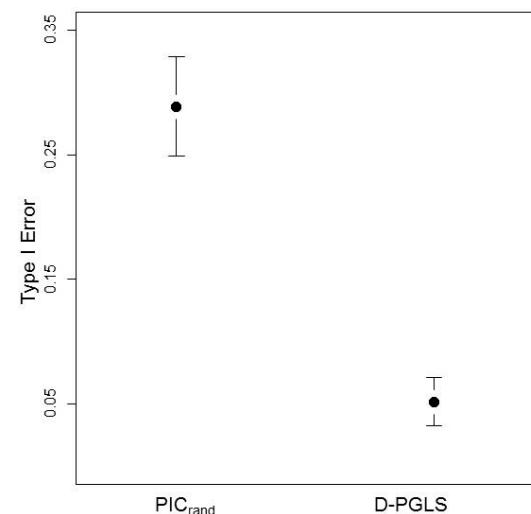D-PGLS shuffles data, PIC_{rand} shuffles PIC_Y (permutes phylogeny in Y vs X)

Result: PIC_{rand} displays high type I error rates

D-PGLS	df	SS	MS	F	R^2	P_{Yrand}
SVL	1	0.0006586	0.0006586	3.0288	0.07039	0.221 NS
Residual	40	0.0086976	0.0086976			
Total	41	0.0093562	0.00021744			

PIC	df	SS	MS	F	R^2	$P_{PICrand}$	P_{Yrand}
SVL	1	0.0006586	0.0006586	3.0288	0.07039	0.026	0.221 NS
Residual	40	0.0086976	0.0086976				
Total	41	0.0093562	0.00021744				

What you shuffle matters!

Adams and Collyer 2015. _Evol._

Statistically account for phylogenetic non-independence in error $\varepsilon \sim N(0,\mathbf{V})$

   -Is a weighted GLM

   -Flexible: ANOVA, regression

   -BM or other models can be evaluated (by altering $\mathbf{V}$)

-PIC (the most common comparative analysis) is a special case

Assess covariation between blocks of variables while accounting for phylogeny

-PLS of evolutionary covariance matrix

$$\mathbf{R}_{12} = \mathbf{U}_{\mathbf{R}1}\mathbf{D}\mathbf{V}_{\mathbf{R}2}^{t}$$

$$\mathbf{R} = \frac{(\mathbf{Y} - E(\mathbf{Y}))^{t}\,\mathbf{C}^{-1}(\mathbf{Y} - E(\mathbf{Y}))}{N-1}$$
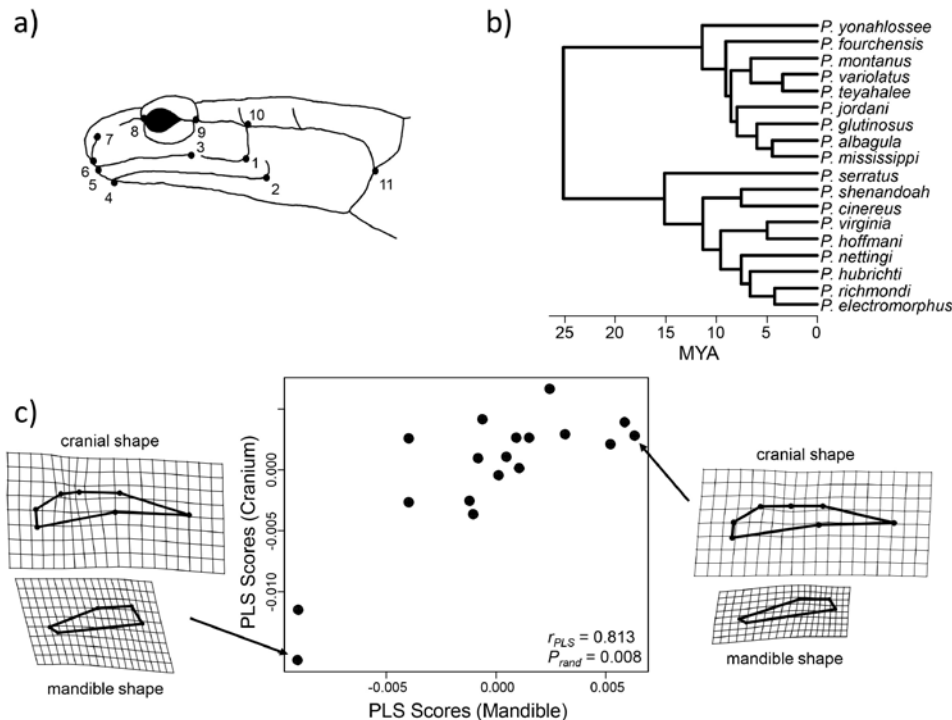
Adams and Felice 2014. *PloS One.*

-PLS of covariance matrix from PIC is mathematically equivalent

Klingenberg and Marugán-Lobón 2013. *Syst. Biol.*

Phylo-canonical correlation also possible:
Revell & Harrison. 2008. Bioinformatics.

Adams and Felice 2014. *PloS One.*

Ordinations that incorporate phylogenetic relatedness

Two approaches:

1: Phylogenetic PCA (pPCA): account for phylogeny in PCA computations:

SVD of evolutionary rate matrix (the covariance matrix 'standardized' by phylogeny)

$$R = \frac{(\mathbf{Y} - E(\mathbf{Y}))^{\mathbf{T}}\mathbf{C}^{-1}(\mathbf{Y} - E(\mathbf{Y}))}{N - 1}$$
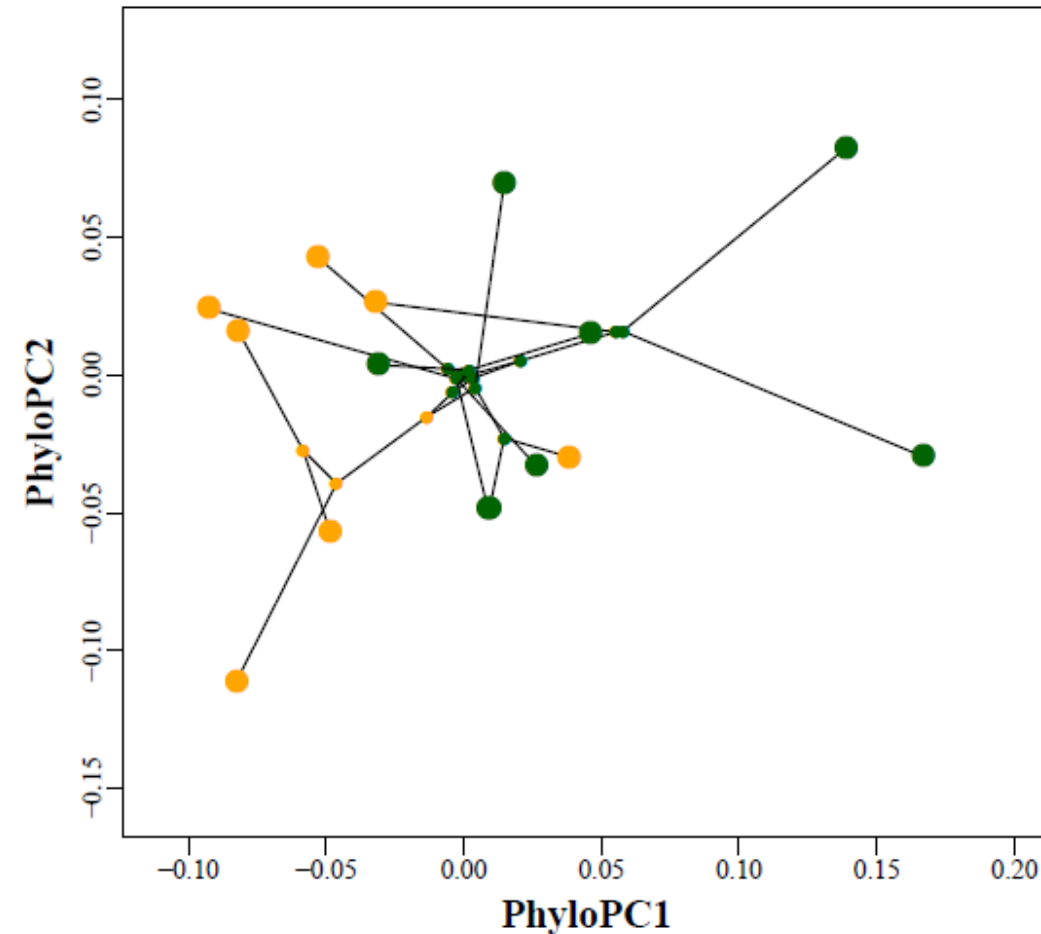
Revell 2009. *Evol.*

Produces phylogenetically corrected eigenvectors

Preserves specimen distances, but pPCA scores correlated

2: Phylomorphspace: project phylogeny into PCA space (using ancestral states)

Can provide important visual insights into macroevolutionary patterns

```
> library(phytools)
> phyl.pca(tree, morpho)

Phylogenetic pca
Starndard deviations:
        PC1          PC2
0.09321575 0.07317544

Loads:
              PC1          PC2
TRL   0.64293015   0.3673396
HL   -0.40347305  -0.2379650
HW   -0.09881369  -0.6795454
HH    0.51833347  -0.8385141
FFL  -0.95357849  -0.2108364
HFL  -0.91580552  -0.3113380
```
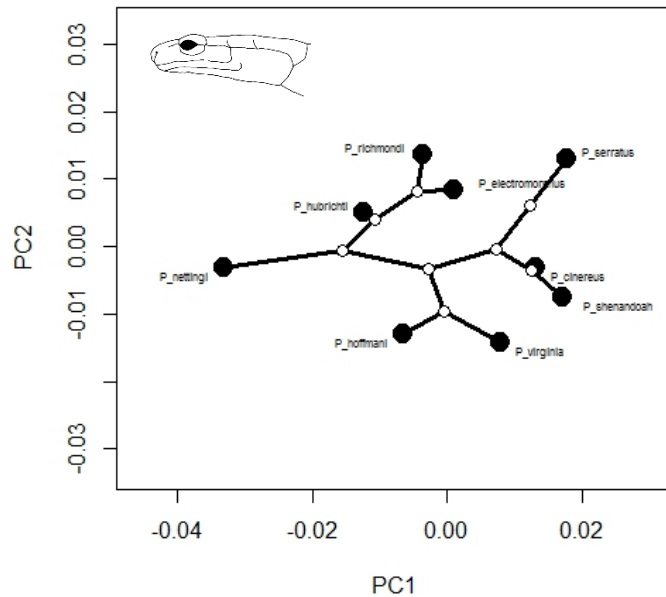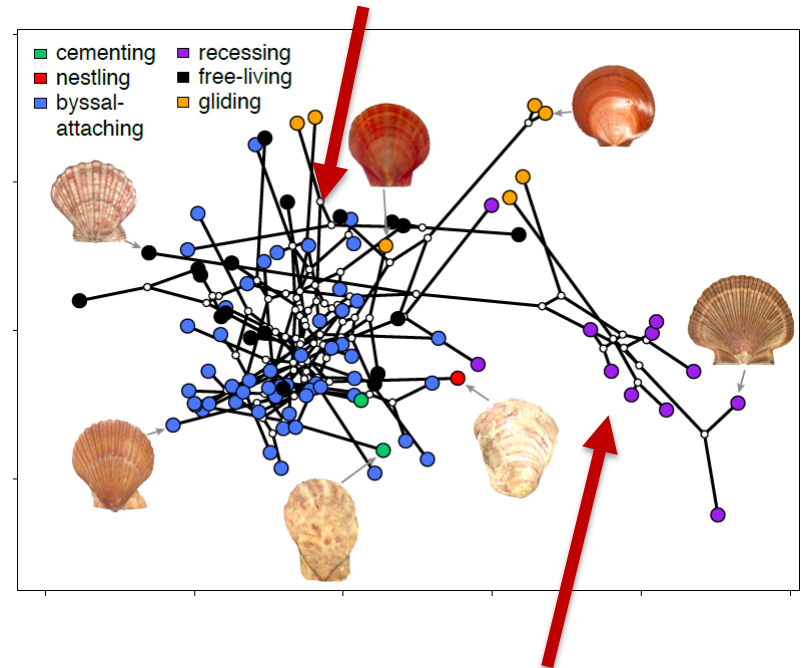
Species emanating out from ancestors: shows shape divergence over macroevolutionary time

Species trajectories 'cross-cross' one another (convergent/repeated evolution/homoplasy)
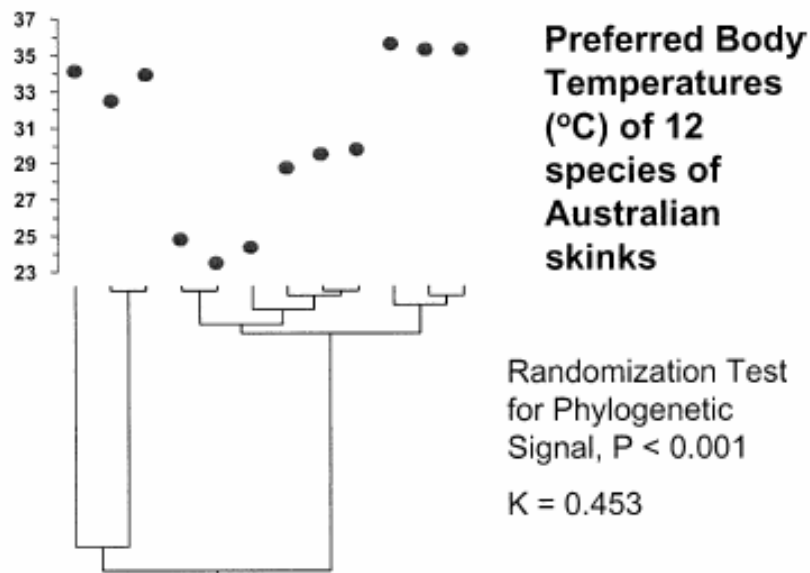

Adams 2014a. *Syst. Biol.*

Potential directional trend

Sherratt et al. *Evol.* 2016.

The degree to which phenotypic similarity associates with phylogenetic relatedness

Blomberg´s K measure:

$$K = \frac{\left(\mathbf{Y} - E(\mathbf{Y})\right)^{t}\left(\mathbf{Y} - E(\mathbf{Y})\right)}{\left(\mathbf{Y} - E(\mathbf{Y})\right)^{t}\mathbf{C}^{-1}\left(\mathbf{Y} - E(\mathbf{Y})\right)} \Bigg/ \frac{tr(\mathbf{C}) - N(\mathbf{1}^{t}\mathbf{C}^{-1}\mathbf{1})^{-1}}{N-1}$$

Preferred Body Temperatures (°C) of 12 species of Australian skinks

Randomization Test for Phylogenetic Signal, P < 0.001

K = 0.453

$K \in (0, \infty)$

$E(K) = 1$ under Brownian Motion
$K < 1$: related taxa less similar than that expected under BM
$K > 1$: related taxa more similar than that expected under BM

Blomberg et al. 2003. *Evolution* 57: 717-745.

Extended to multivariate using the equivalency between variances and distances

(remember PCA and PCoA)

Distances to multivariate mean

$$K_{mult} = \frac{\mathbf{D}_{Y,\hat{a}}^{t}\mathbf{D}_{Y,\hat{a}}}{\mathbf{PD}_{U,0}^{t}\mathbf{PD}_{U,0}} \bigg/ \frac{tr\mathbf{C} - N(\mathbf{1}^{1}\mathbf{C}^{-1}\mathbf{1})^{-1}}{N-1}$$

Distances to origin after phylogenetic transformation

Phylogenetic transformation

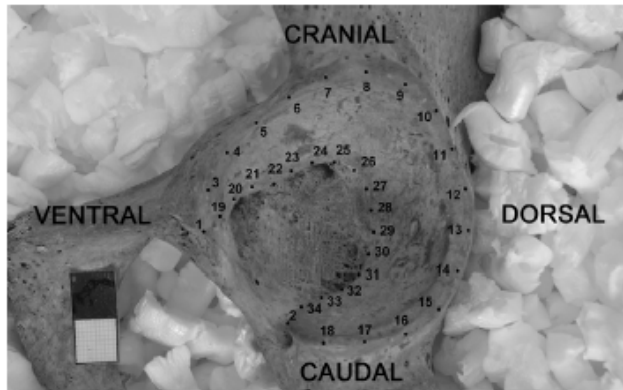1. Calculate eigenvectors ($\mathbf{U}$) and eigenvalues ($\mathbf{W}$) of $\mathbf{C}$

$$\mathbf{C} = \mathbf{UWU}^{-1}$$

2. Calculate phylogenetic transform matrix $\mathbf{P} = (\mathbf{UW}^{1/2}\mathbf{U}^{\mathbf{T}})^{-1}$
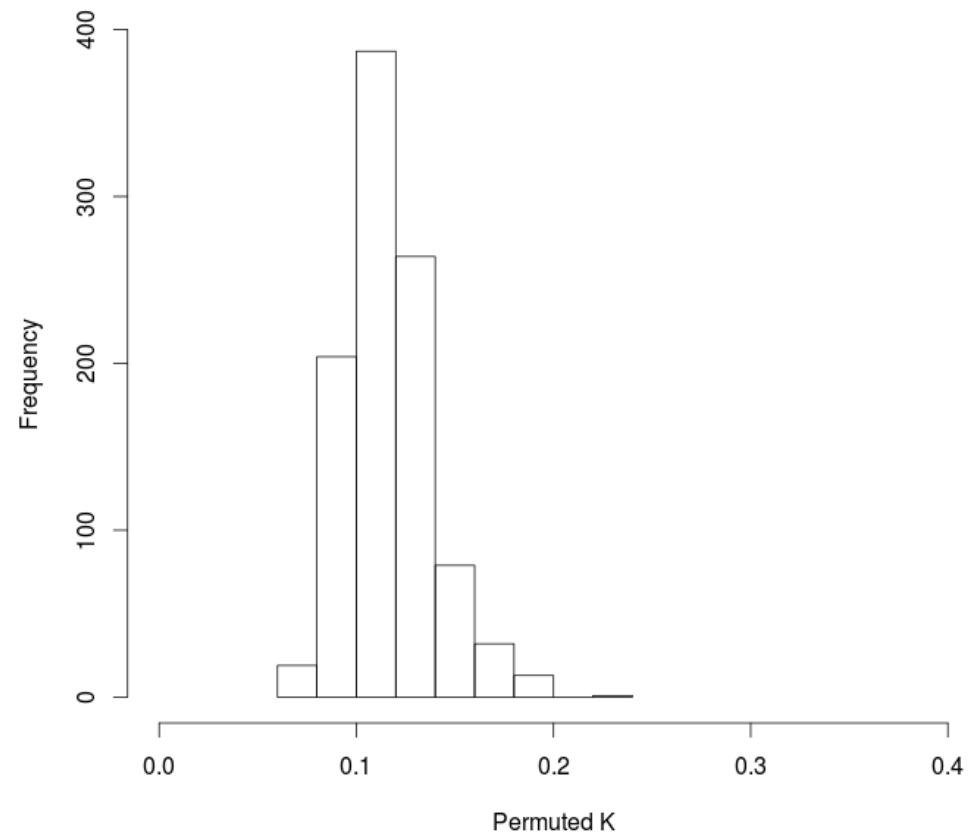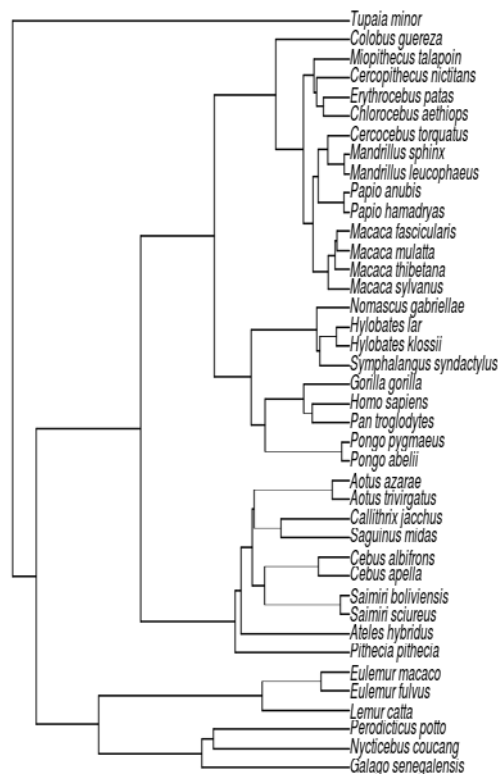
3. Transform phenotypic data for phylogeny $\mathbf{U_X} = \mathbf{P}(\mathbf{X} - E(\mathbf{X}))$

Statistical testing through phylogenetic permutation (permutation of tip values)
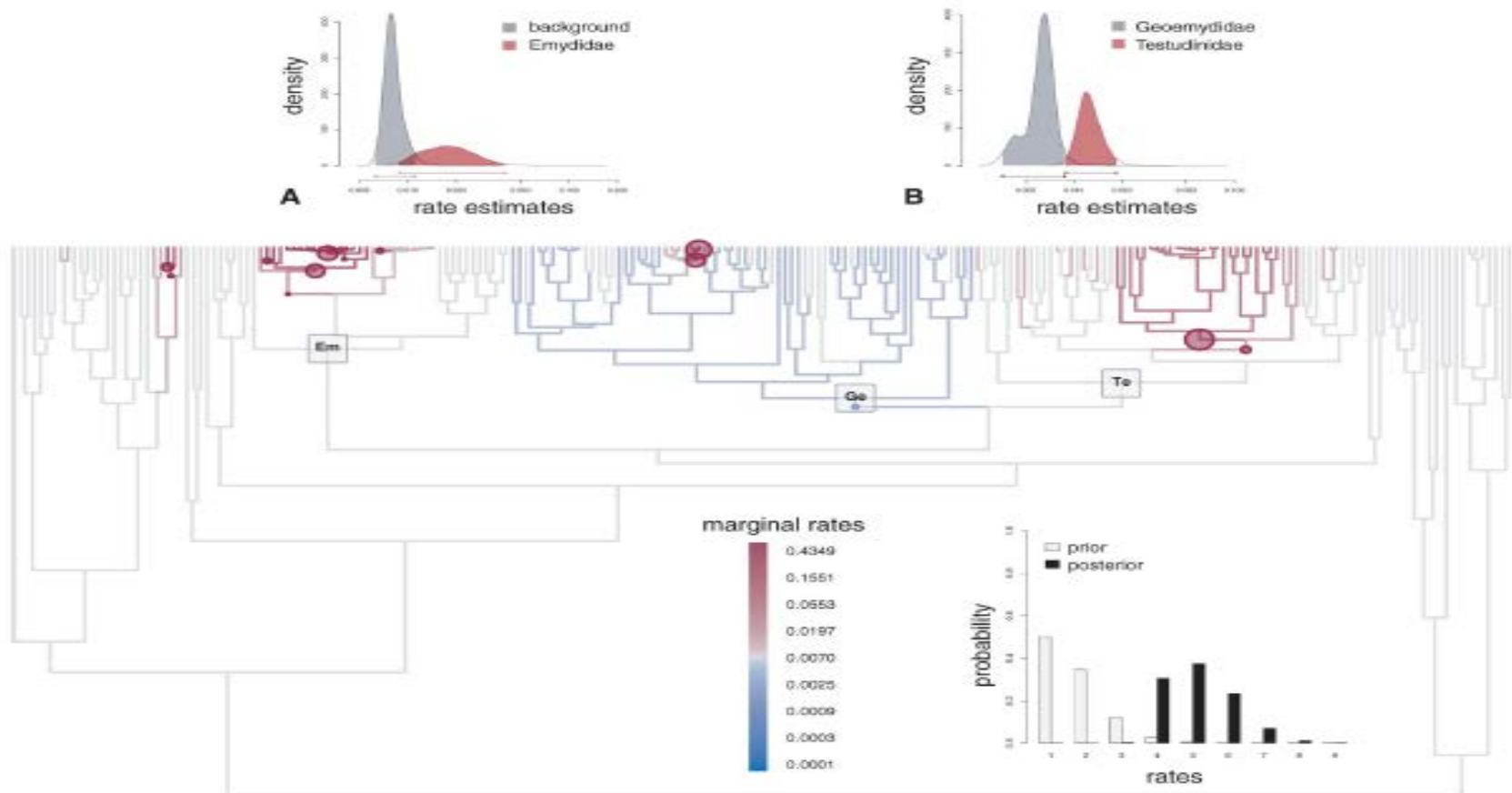
Is acetabular shape related to phylogeny in primates?

Kmult = 0.347
p = 0.001

San Millán et al. 2015, *J. Hum. Evol.*

"How fast, as a matter of fact, do animals evolve in nature?" (Simpson, 1944)

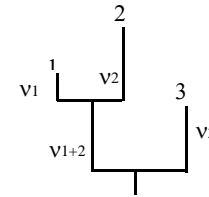

Eastman et al. 2011

Evolutionary rates = σ^2 (a phylogenetically 'standardized' variance)

$$\sigma^2 = \frac{\left(\mathbf{Y} - E(\mathbf{Y})\right)^t \mathbf{C}^{-1}\left(\mathbf{Y} - E(\mathbf{Y})\right)}{N}$$

Felsenstein 1973. *Am J. Hum. Gen.*

$$\mathbf{C} = \begin{pmatrix} v_1 + v_{1+2} & v_{1+2} & 0 \\ v_{1+2} & v_2 + v_{1+2} & 0 \\ 0 & 0 & v_3 \end{pmatrix}$$
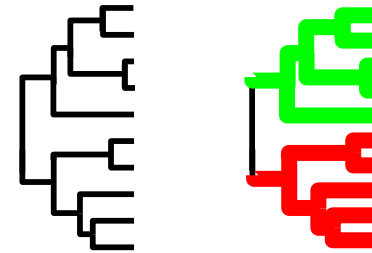
For multiple traits, σ^2 is a matrix **(R)**

$$\mathbf{R} = \begin{bmatrix} \sigma_1^2 & & \\ \sigma_{21} & \sigma_2^2 & \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{bmatrix}$$

Revell & Harmon 2008. *Ev. Ecol. Res.*

1: Use logL_{mult} to compare alternative rate-based models
   -Define 'regimes' for models (BM1, BMM, etc.)

   -Estimate **R** and log$L$

$$\log\left(L_{(\mathbf{R})}\right) = \log\left[\frac{\exp\left(-\frac{1}{2}\left((\mathbf{Y} - E(\mathbf{Y}))^t (\mathbf{R} \otimes \mathbf{C})^{-1}(\mathbf{Y} - E(\mathbf{Y}))\right)\right)}{\sqrt{(2\pi)^{Np} \times \det(\mathbf{R} \otimes \mathbf{C})}} \right]$$

-Compare log$L$ (LRT tests, AIC, phylogenetic simulation, etc).

Methods can be adapted to compare rates among clades (does one trait evolve faster than another)?

Obtain R_o and logL:
$$\mathbf{R} = \frac{\left(\mathbf{Y} - E(\mathbf{Y})\right)^t \mathbf{C}^{-1} \left(\mathbf{Y} - E(\mathbf{Y})\right)}{N}$$
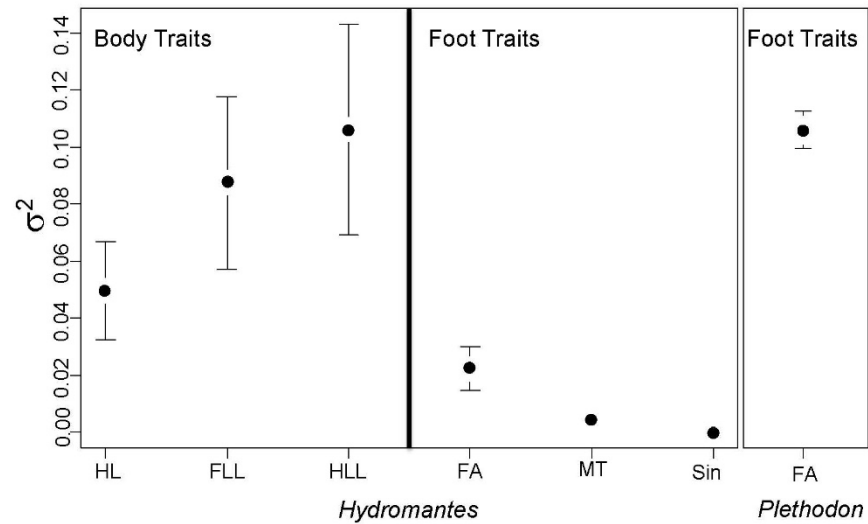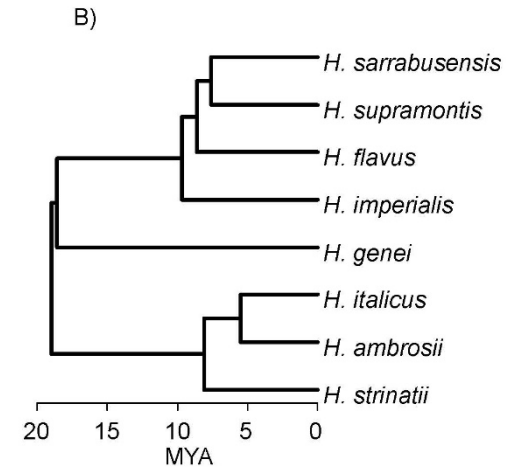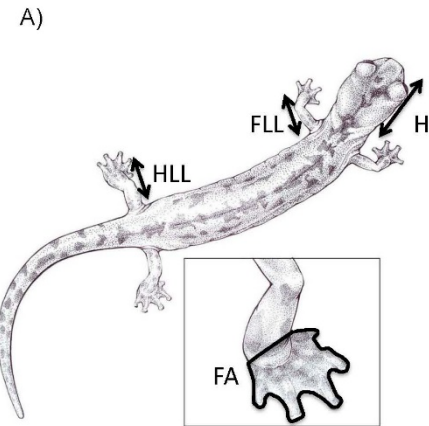
Estimate R_c & logL, where rates are constrained to be the same

$$\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_p^2 \qquad \mathbf{R}_C = \begin{bmatrix} \sigma_1^2 & & \\ \sigma_{21} & \sigma_2^2 & \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{bmatrix}$$

Compare the two models with LRT

Adams. (2013). *Syst. Biol.*
Extended to multivariate: Denton and Adams, 2015. *Evol.*

Compare morphological rates in cave-dwelling *Hydromantes*

Climbing traits evolve more slowly (consistent with evolutionary constraint)

Adams et al. 2017. *Am. Nat.*

For multiple (univariate) traits, the evolutionary rate becomes a matrix

$$\mathbf{R} = \frac{\left(\mathbf{Y} - E(\mathbf{Y})\right)^t \mathbf{C}^{-1}\left(\mathbf{Y} - E(\mathbf{Y})\right)}{N} = \begin{bmatrix} \sigma_1^2 & & \\ \sigma_{21} & \sigma_2^2 & \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{bmatrix}$$

Some traits are multi-dimensional. Here, a *net rate* of evolution is found as:

$$\sigma_{mult}^2 = \frac{\mathbf{PD}_{\mathbf{U},0}^t \mathbf{PD}_{\mathbf{U},0}}{N}$$

Found from distances among objects in phylogenetic transformed space

Rates may be compared using permutation
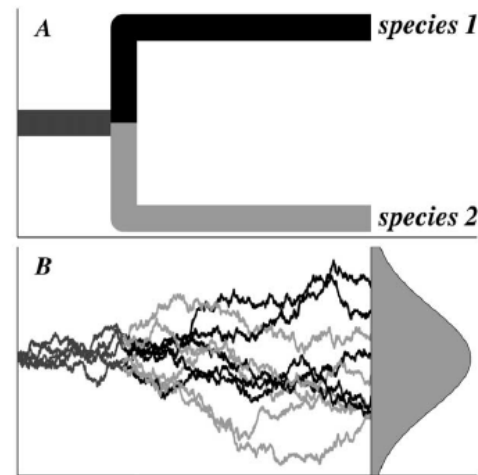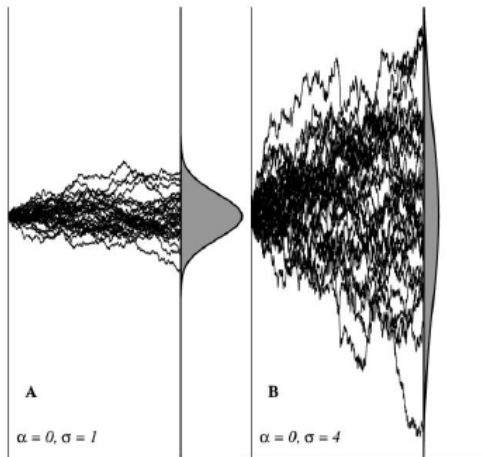
Adams 2014b. *Syst. Biol*

In what manne3r do traits evolve?

Brownian Motion: neutral change under drift (no selection)

Variance increases with time, but the mean value remains constant

$$dY(t) = \sigma B(t)$$

Character change   Evolutionary rate   Small random perturbations

Images from Butler and King 2004. Am. Nat. 164: 683-695.

OU models include drift and selection

$$dX(t) = a(\beta - X(t)) + \sigma B(t)$$
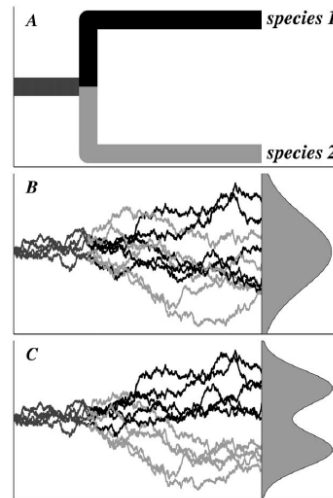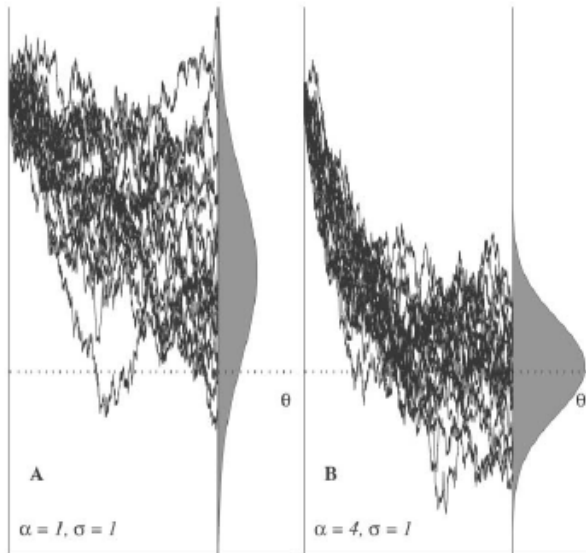
Strength of selection

Distance from optima

Brownian motion portion

Trait values are "pulled" towards one or more optima
   Single optimum value: stabilizing selection
   Two or more optima: diversifying selection

Hansen and Martins 1996. Evolution
Martins and Hansen 1997. Am. Nat.
Butler and King, 2004. Am. Nat.

Evolutionary rate changes along phylogeny (accelerates/decelerates)

   -Contains σ^2 and additional parameter 'g' which scales rate of trait change

Can model early burst of phenotypic evolution (as expected in adaptive radiations)

Blomberg et al. 2003. *Am. Nat.*

Evolutionary models describe 'expected' trait variation (described by **V**)

Fit models: BM1, BMM, OU1, OUM, ACDC, etc.

   -Obtain parameters, and logL

Compare evolutionary models using LRT and AIC

How did Anolis body size groups (small, medium, large) evolve?

-5 models: BM, OU_1, OU_3 OU_4 (3 group+anc), OU_{LP} (3 gp + history of colonization)

-OU_{LP} (3 gp + col. hist.) best explains body size evolution
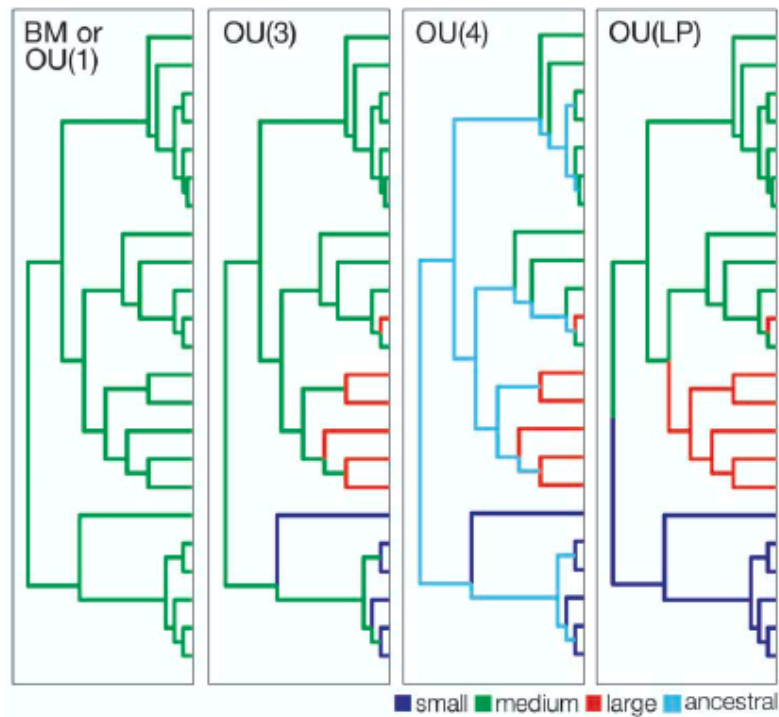

■ small ■ medium ■ large ■ ancestral

Table 1: Performance of alternative models for body size evolution in the character displacement study

|	BM	OU(1)	OU(3)	OU(4)	OU(LP)
$-2 \log \mathcal{L}$	−34.66	−34.66	−40.21	−47.22	−49.69
AIC	−30.66	−26.66	−28.21	−33.22	−37.69
SIC	−28.39	−22.12	−21.40	−25.27	-30.88
LR		0	5.55	12.56	15.03
P value		1	.24	.028	.0046

Butler and King. 2004. *Am. Nat.*

Some work on multivariate phyologenetic evolutionary models, but

-Multivariate rate matrix models (**R**) have high type I error as p ↑
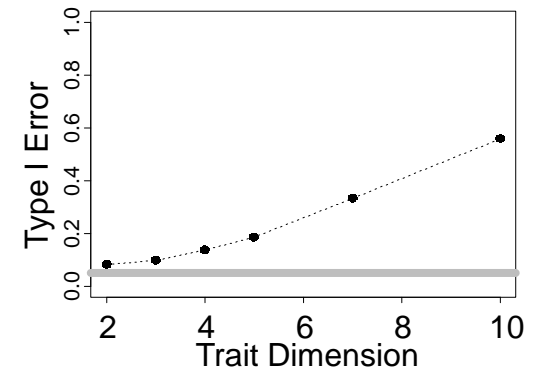
Adams 2014b. *Syst. Biol.*

-Multivariate OU comparisons also misspecify models

$N = 32$

$p = 4$ $p = 8$

% Misspecification

BM$_1$ vs. OU$_1$
(mvMORPH)
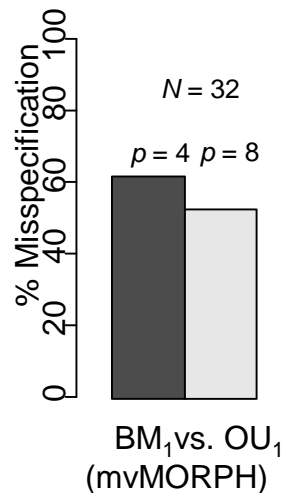
Adams and Collyer 2017. *In Review.*

More work is needed here

Adams 2014b. *Syst. Biol.*
Adams and Collyer (In Review)

Testing evolutionary hypotheses requires phylogenetic perspective

Analytical methods developed for evaluating:
- Evolutionary correlations (PGLS)
- Compare evolutionary models (BM, OU, etc.)
- Compare evolutionary rates
- Evaluation phylogenetic signal

Multivariate analogs for most methods exist

      PGLS

      PPLS

      Phylomorphospace

      K_{mult}

      σ^2_{mult}

Robust multivariate OU models still in need of development