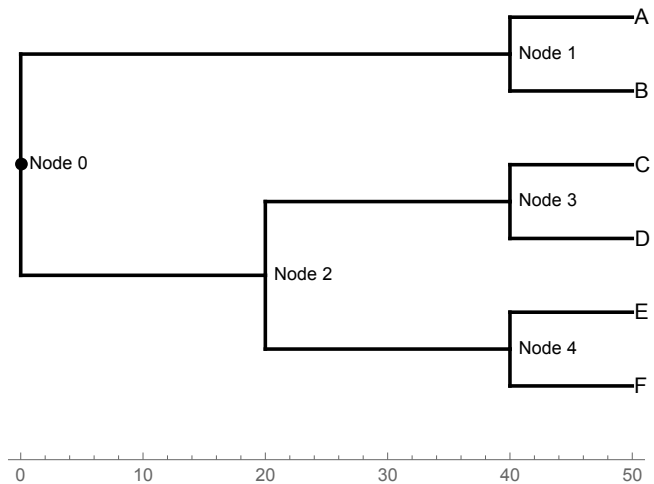
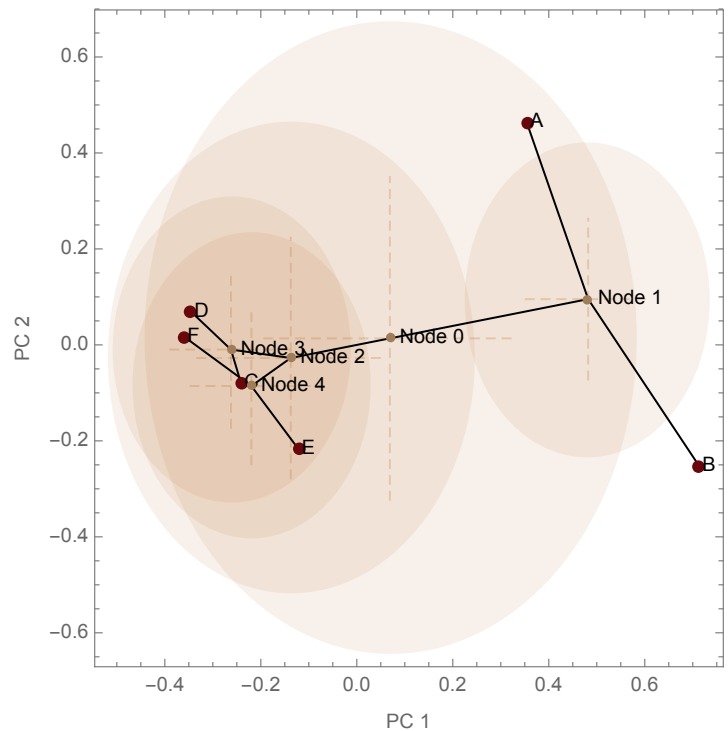
Phylogenetic comparative methods (PCM)

for geometric morphometrics





P. David Polly
Department of Earth and Atmospheric Sciences
Adjunct in Biology and Anthropology
Indiana University
Bloomington, Indiana 47405 USA
pdpolly@indiana.edu

Outline

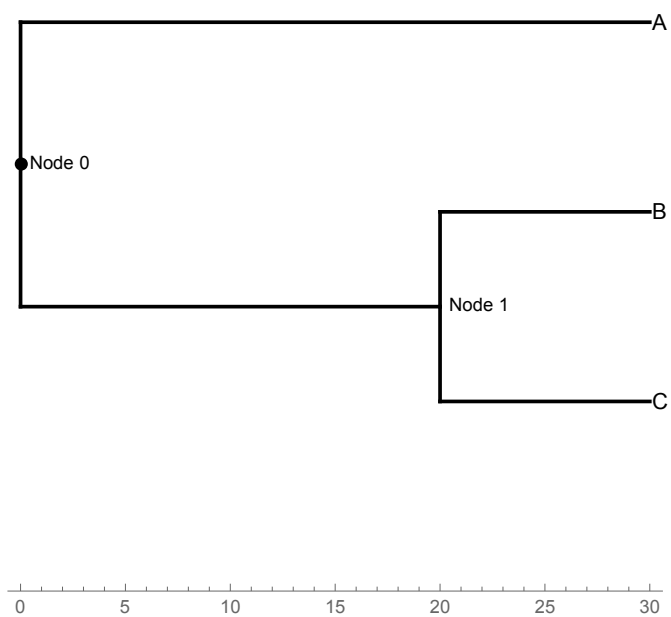
- The reason for phylogenetic comparative methods
- Brownian motion as a null model for comparative methods
- Measuring phylogenetic "signal"
    - Pagel's lambda (λ)
    - Blomberg's kappa (*K*)
- Phylomorphospace: projecting a tree into shape space
    - Ancestor state reconstruction
    - Confidence intervals and probabilities
- Types of PCM for GMM
    - Independent contrasts
    - PGLS "regression"
    - Phylogenetic MANCOVA
- Adams vs. Klingenberg on analyzing multivariate data
- Why visualizing results of PCMs is difficult
    - PCMs measure joint change in shape, not shape per se

The pitfalls of phylogenetic principal components analysis

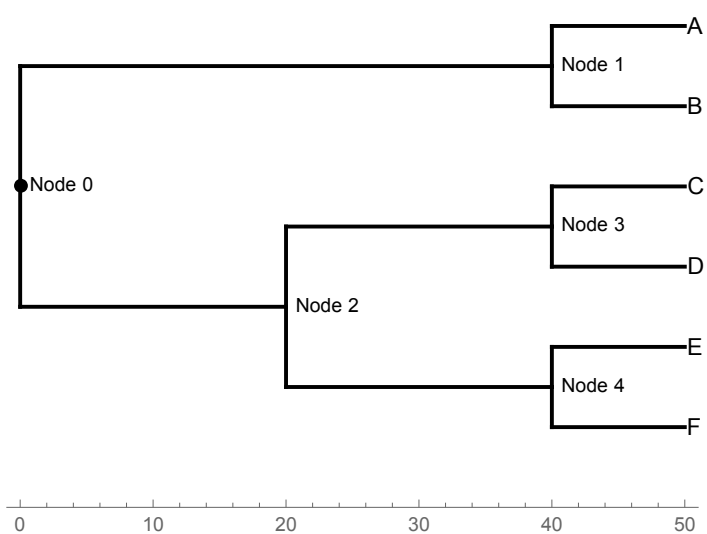
What are phylogenetic comparative methods (PCMs)?

Modified statistical tests that take into account non-independence in data that come from different species.

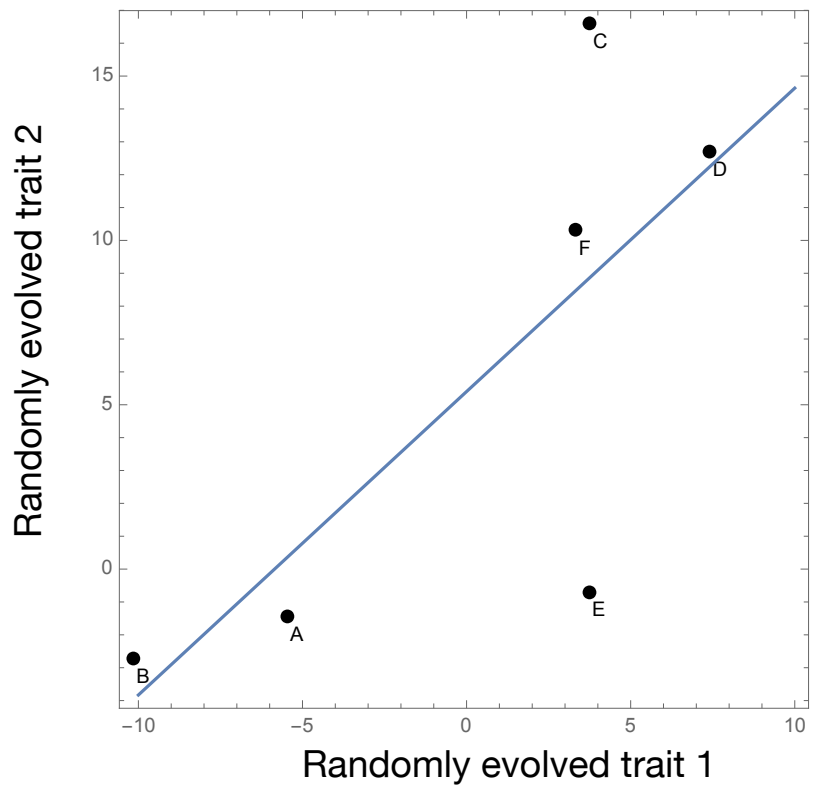
PCMs are applied to regression, MANOVA, and other standard tests, and they can be used to estimate rates of evolution and to reconstruct ancestral trait states on a phylogeny.



Phylogenetic correlation in random traits

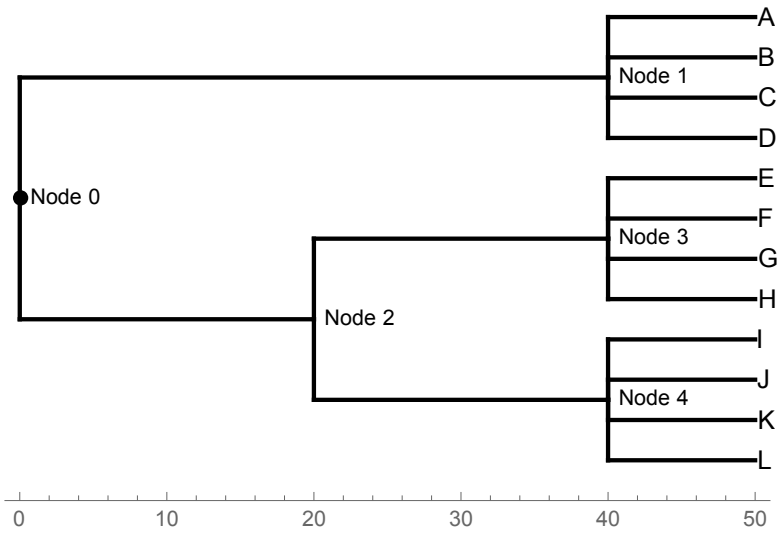


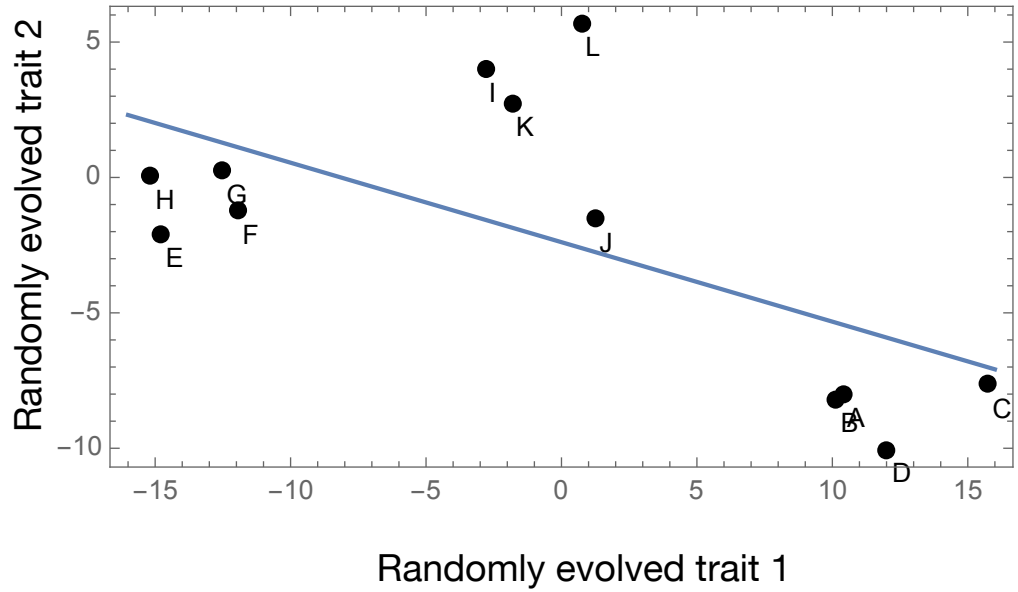
	DF	SS	MS	F-Statistic	P-Value
x	1	192.528	192.528	4.85587	0.0922812
Error	4	158.594	39.6484		
Total	5	351.121			



Additional tips multiplies the effect

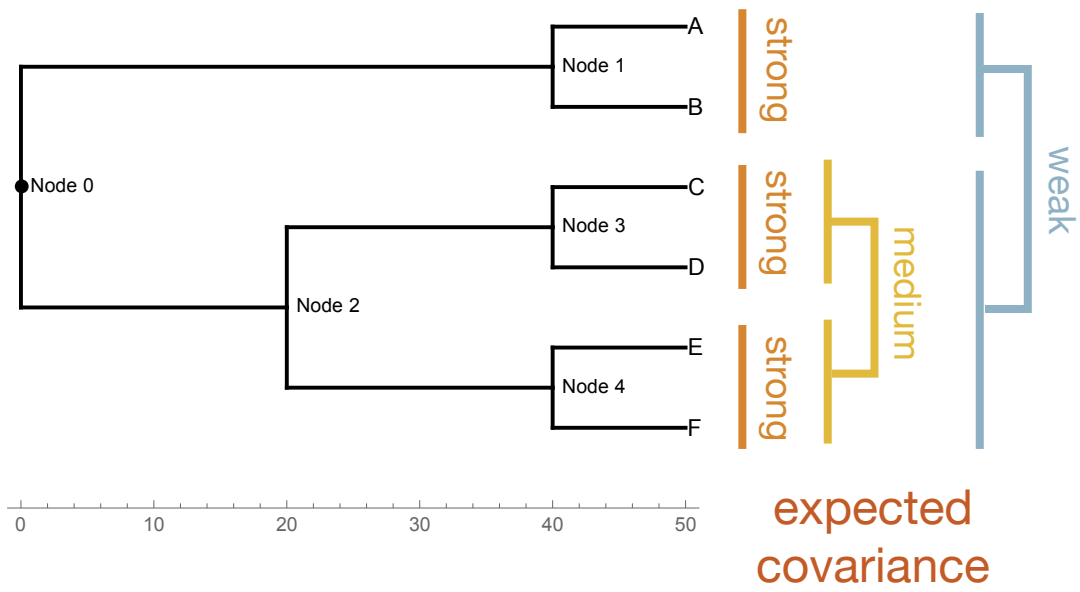
	DF	SS	MS	F-Statistic	P-Value
x	1	116.972	116.972	6.45193	0.0293599
Error	10	181.297	18.1297		
Total	11	298.269			





How do PCMs work?

- PCMs estimate how much covariance a trait should have between taxa (and between traits) is expected from the phylogenetic topology
- The expected phylogenetic covariance is removed, leaving the residual between-trait covariance
- Statistical tests are carried out on the residual component

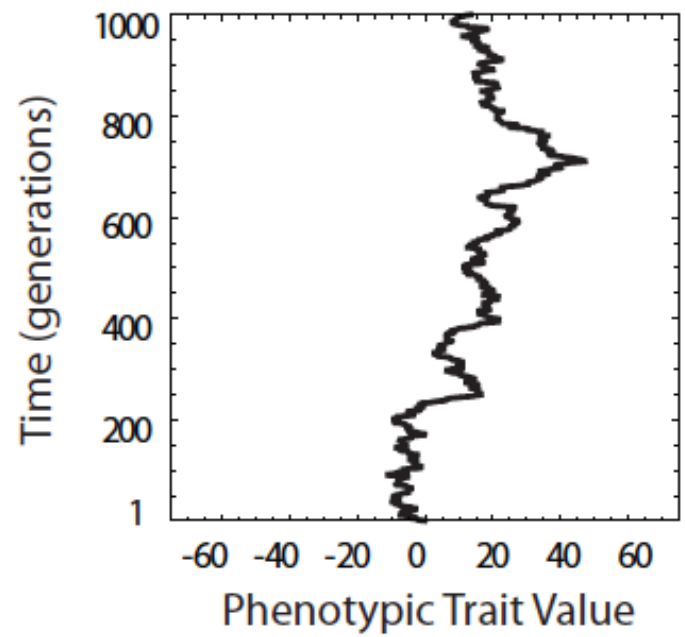


How do we know the expected covariance?

Depends on how the traits evolve: Brownian motion (purely random evolution) is usually used for PCMs.

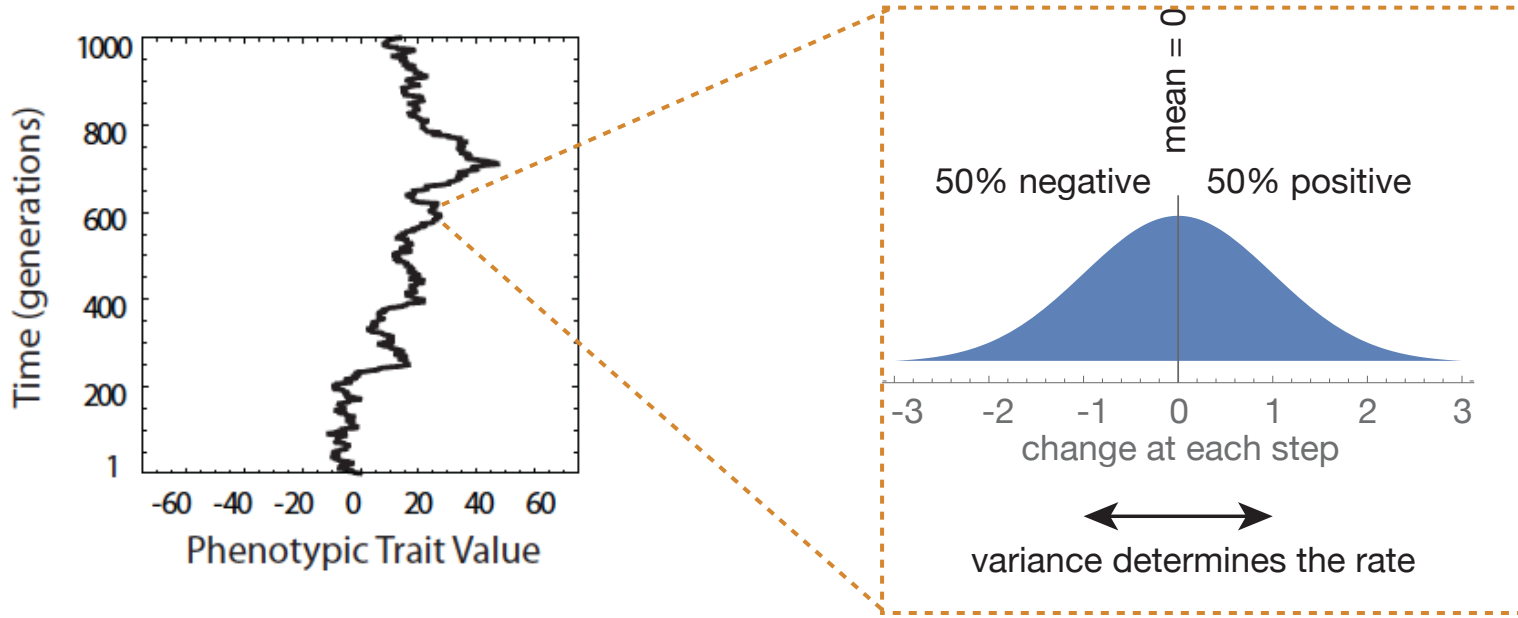
Evolution is change in the mean phenotype (=trait value) from generation to generation...

Evolution = Mean(selection) +
Mean(drift) +
Mean(nongenetic variation)

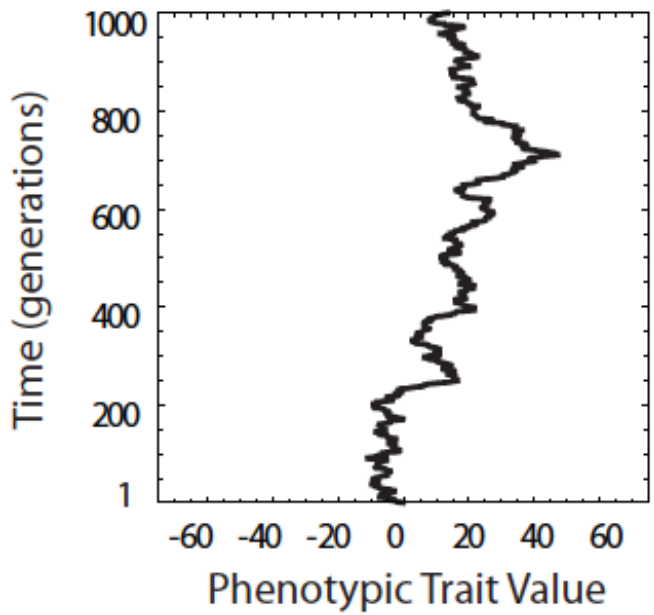


Definition of Brownian motion

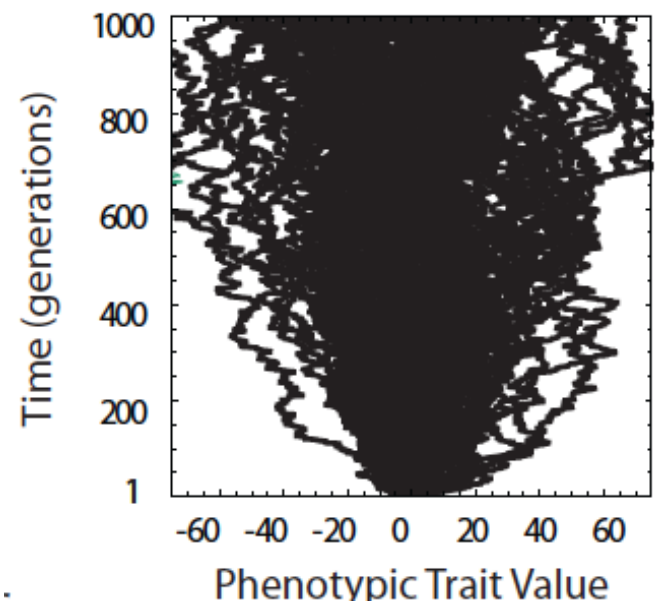
- Brownian motion is equivalent to an unbiased random walk (at type of Markov process)
- change at each step is random with respect to other steps
- change at each step has an equal chance of moving in positive or negative direction
- typical implementation specifies steps as coming from a normal distribution with mean=0 and variance=rate of evolution



Statistical properties of Brownian motion



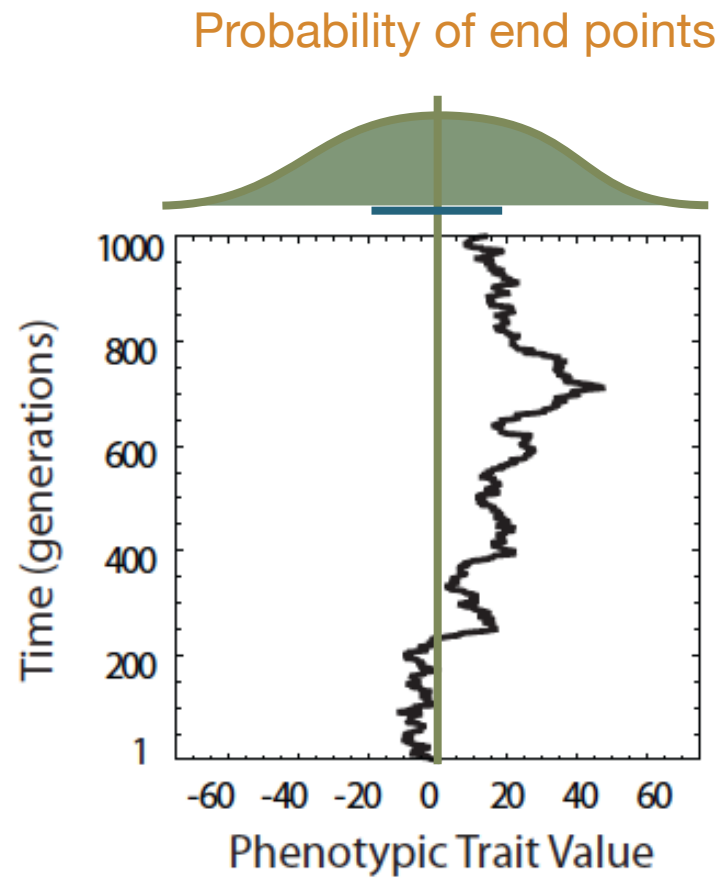
1 random walk



100 random walks

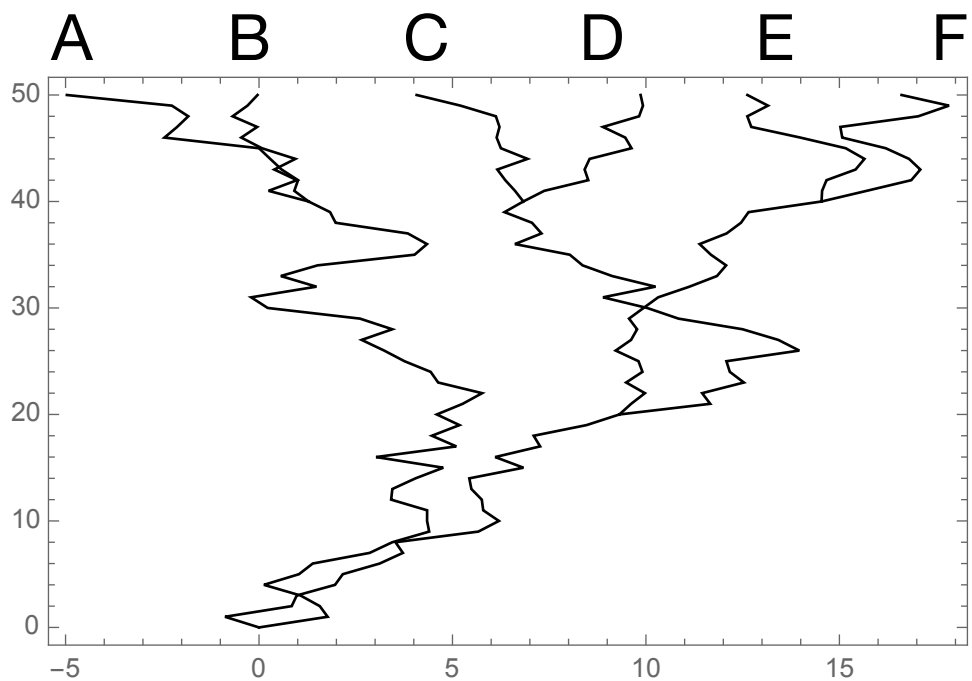
Statistical properties of Brownian motion

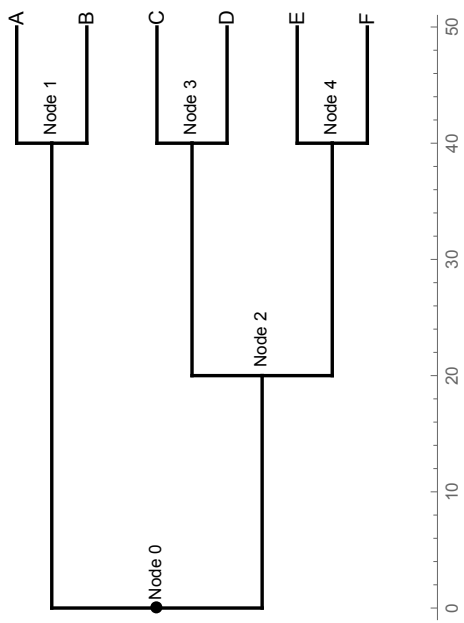
1. Probability of endpoints is a normal distribution (central limit theorem ensures that outcome of series of random events is normally distributed)
2. The most likely endpoint is the starting point
3. The standard deviation of endpoints increases with the square root of time
4. The variance of the endpoints increases linearly with time
5. The variance of the endpoints equals average squared change per step (rate) x number of steps (time)



Expected covariance among tips is proportional to shared branch lengths

Because variance increases linearly with time, covariance between tips is linear with respect to shared branch time

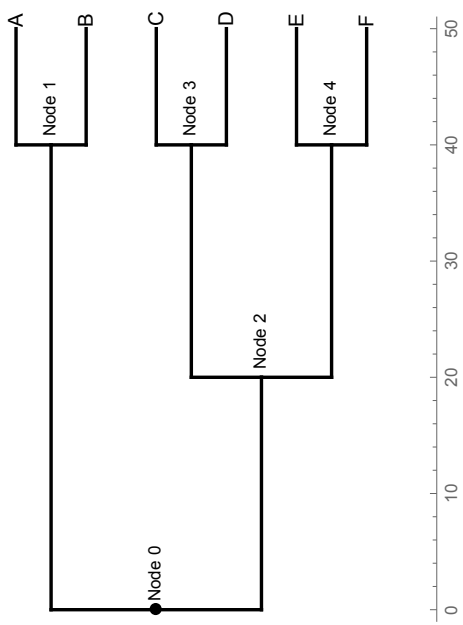




Expected covariance among tips is proportional to shared branch lengths

Phylogenetic covariance matrix ("C" of many authors) has diagonal equal to total branch length between tip and base and off diagonals equal to length of shared branches.

Actual expected variance and covariance is C x rate (as explained above)



Var Y	A	B	C	D	E	F
A	50	40	0	0	0	0
B	40	50	0	0	0	0
C	0	0	50	40	20	20
D	0	0	40	50	20	20
E	0	0	20	20	50	40
F	0	0	20	20	40	50

Expected vs. actual

Covariance expected under Brownian motion

	A	B	C	D	E	F
A	50	40	0	0	0	0
B	40	50	0	0	0	0
C	0	0	50	40	20	20
D	0	0	40	50	20	20
E	0	0	20	20	50	40
F	0	0	20	20	40	50

Covariance of 50 simulated traits

	A	B	C	D	E	F
A	43.	32.	-7.	-3.	-1.	-6.
B	32.	39.	-4.	-1.	-0.	0.
C	-7.	-4.	64.	51.	25.	25.
D	-3.	-1.	51.	52.	26.	25.
E	-1.	-0.	25.	26.	56.	49.
F	-6.	0.	25.	25.	49.	67.

Summary of Brownian motion

- Brownian motion makes a good statistical null because it is a purely random model
- Outcomes of Brownian motion processes are statistically predictable
- Brownian motion can occur in nature through genetic drift or selective drift (selection that changes randomly in direction and magnitude); therefore Brownian motion does not necessarily equate with “neutral evolution”.

Measuring phylogenetic "signal"

The phylogenetic component of data can only be assessed relative to some expectation, such as Brownian motion.

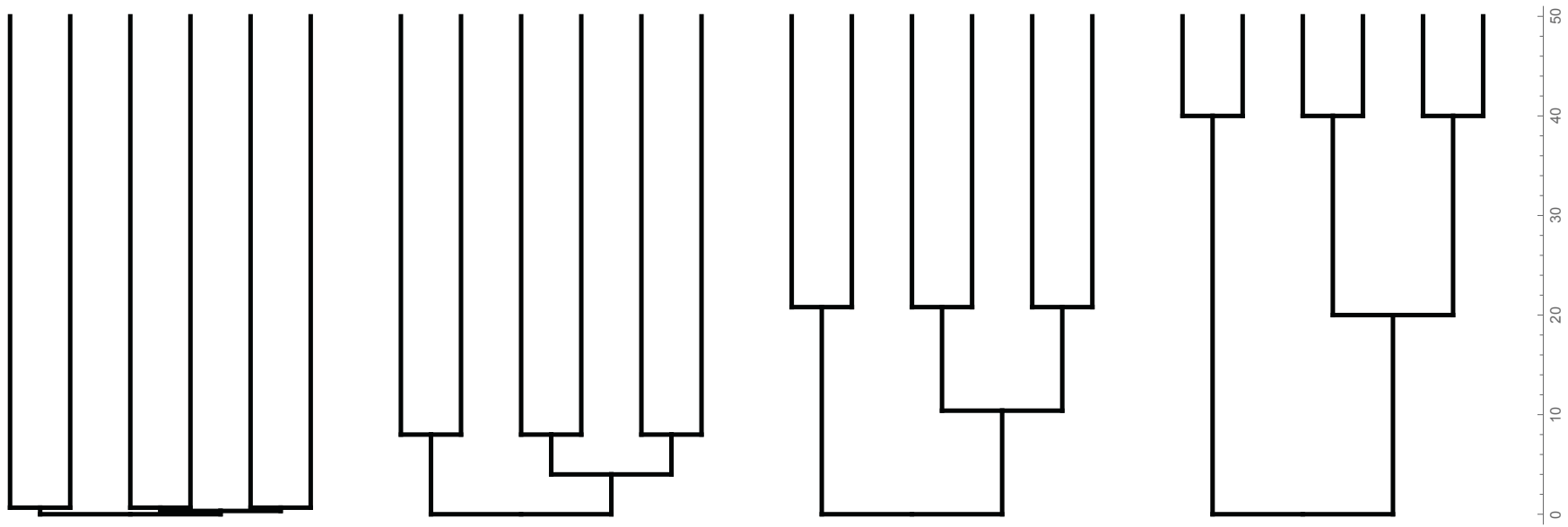
Blomberg's *K* (kappa) = observed cov / expected cov
(Blomberg et al., 2003)
see also Adams' (2014) K_{multi} especially for multivariate data

K can be thought of as the proportion of the covariance that is due to phylogeny. Usually ranges 0 to 1, but can be greater than 1 if phylogenetic covariances are stronger than under BM.

Pagel's *λ* (lambda) = scaling factor so that tree fits BM model
(Pagel, 1999)

λ can be thought of as how you would have to scale the branch lengths so that the data would be obtained under a BM model. Also usually ranges 0 to 1, where 0 is equivalent to no phylogenetic structure and 1 is equal to actual phylogenetic structure. Can also be greater than 1 if phylogenetic covariances are stronger than under BM.

Pagel's lambda illustrated



λ = 0(ish)

λ = 0.2

λ = 0.5

λ = 1.0

no phylogenetic "signal"

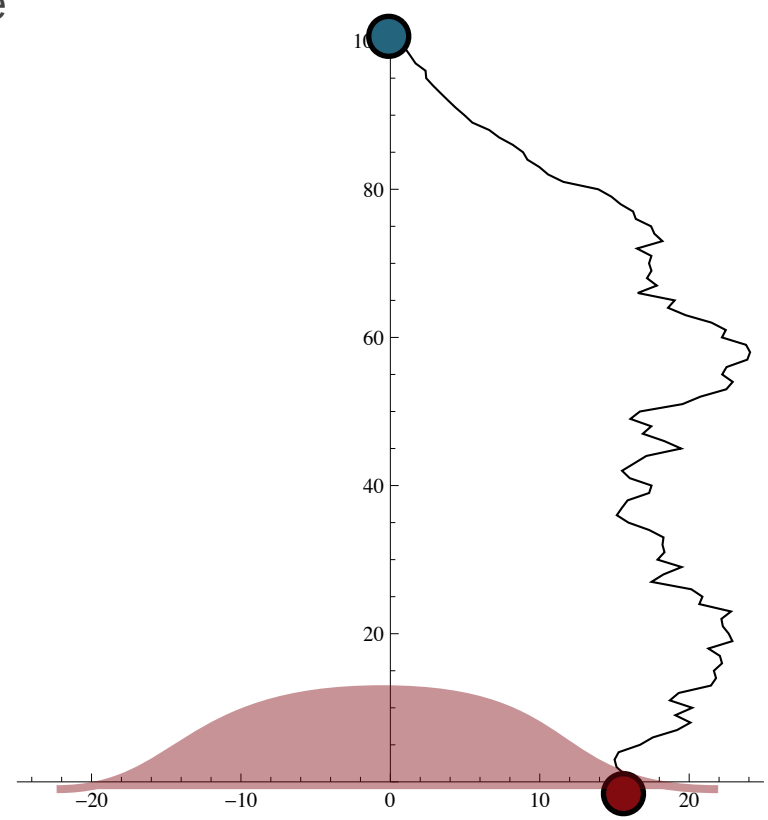
full phylogenetic "signal" (BM)

Reconstructing evolution of shape

Brownian motion in reverse

Most likely ancestral phenotype is same as descendant, variance in likelihood is proportional time since the ancestor lived

Descendant

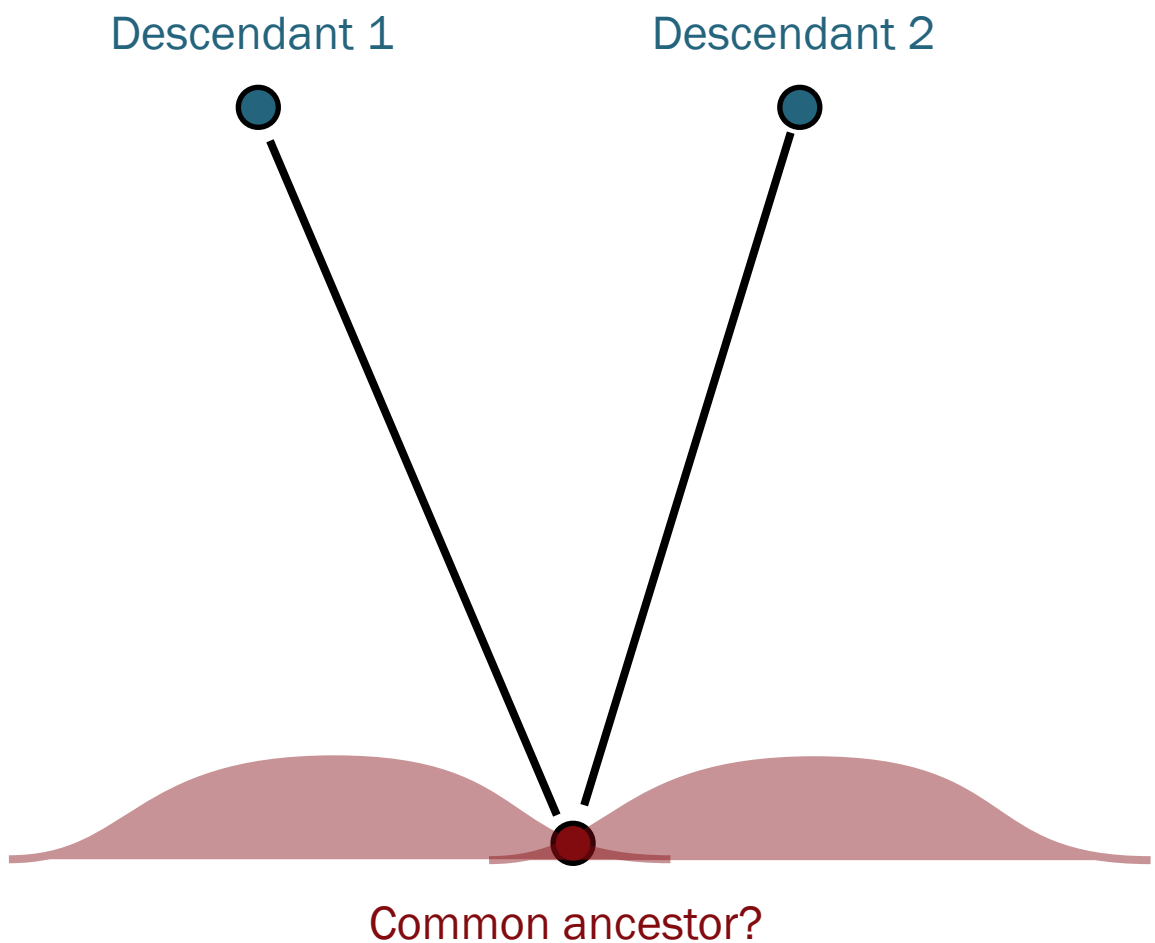


Ancestor?

Ancestor of two branches on phylogenetic tree

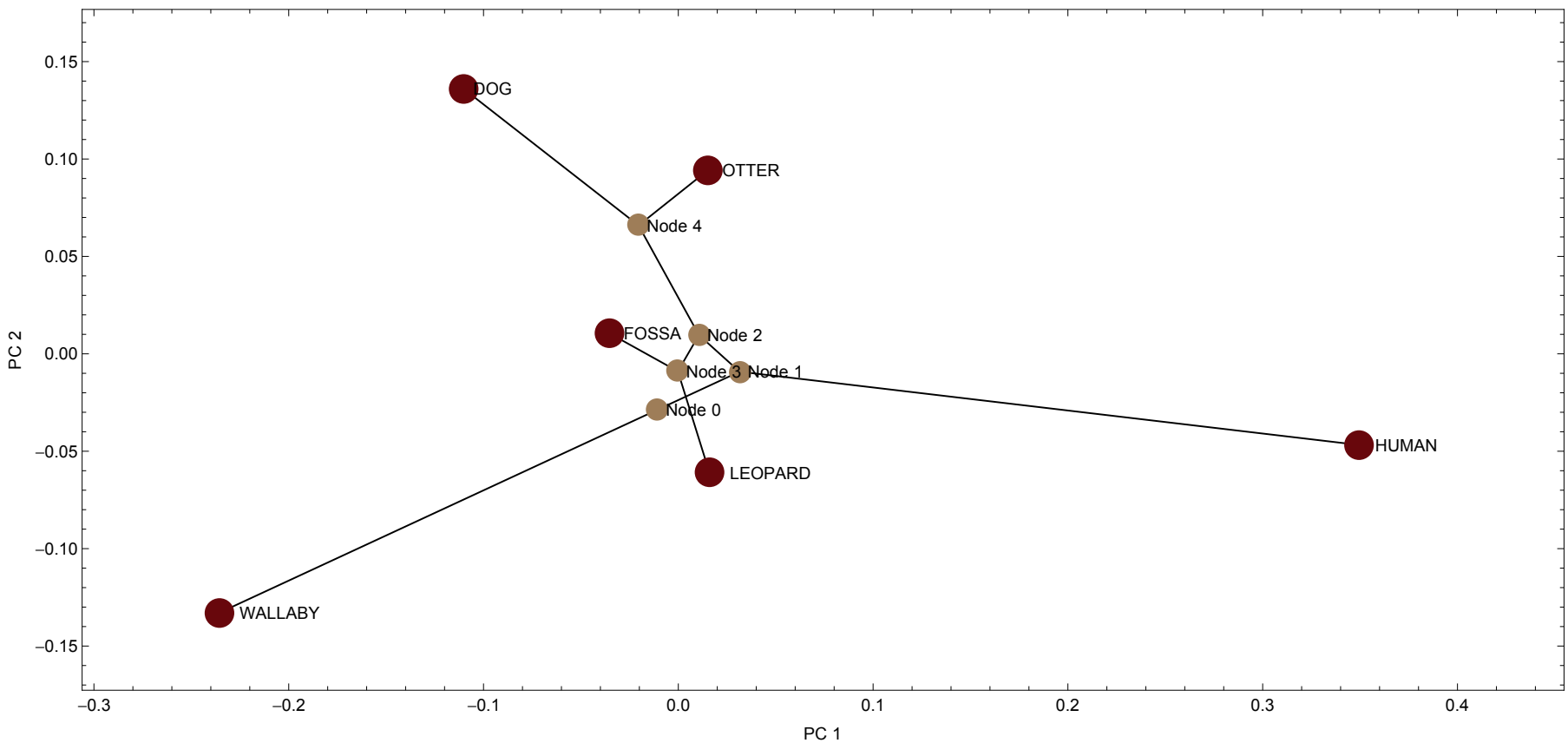
If likelihood of ancestor of one descendant is normal distribution with variance proportional to time, then likelihood of two ancestors is the product of their probabilities.

This is the maximum likelihood method for estimating phylogeny, and for reconstructing ancestral phenotypes. (Felsenstein,

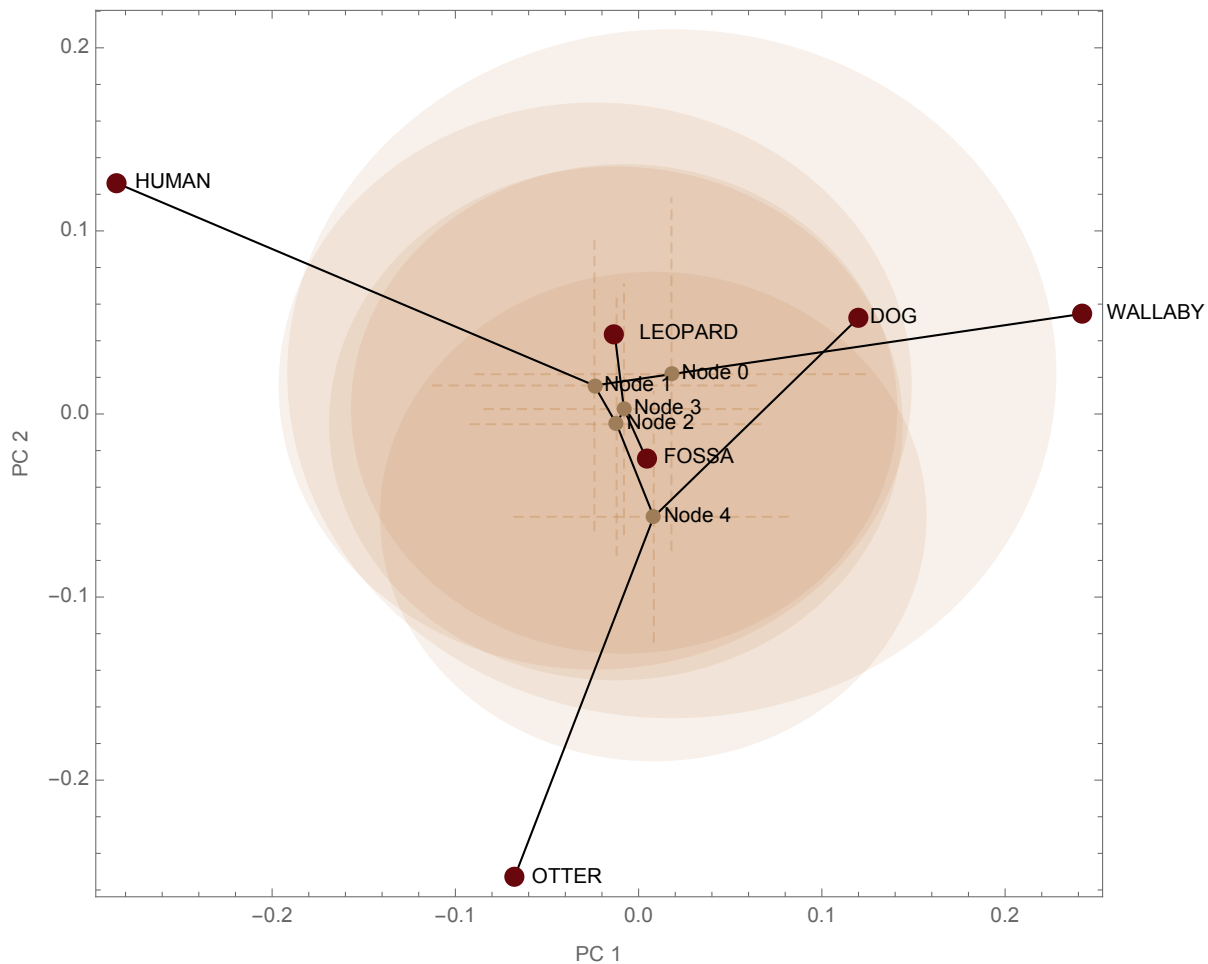


Phylogenetic tree projected into morphospace

- Ancestral shape scores reconstructed assuming Brownian motion
- Ancestors plotted in morphospace
- Tree branches drawn to connect ancestors and nodes

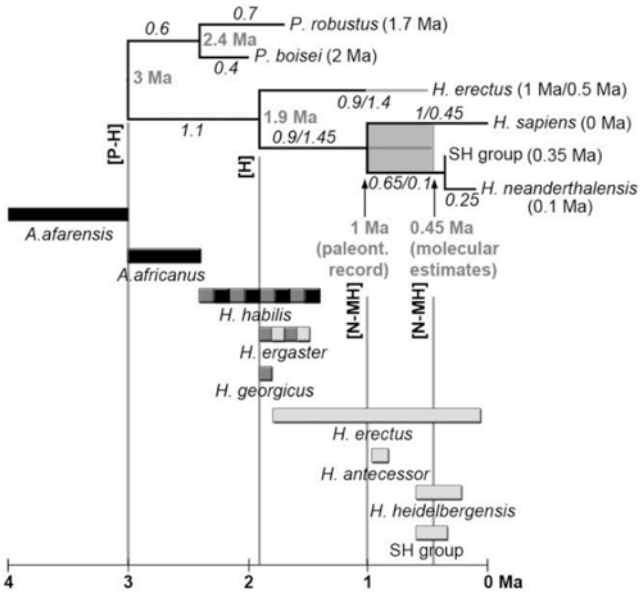


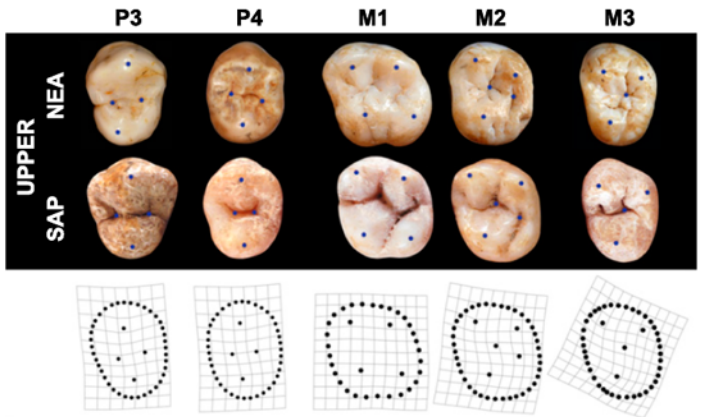
Phylogenetic tree with 95% confidence intervals

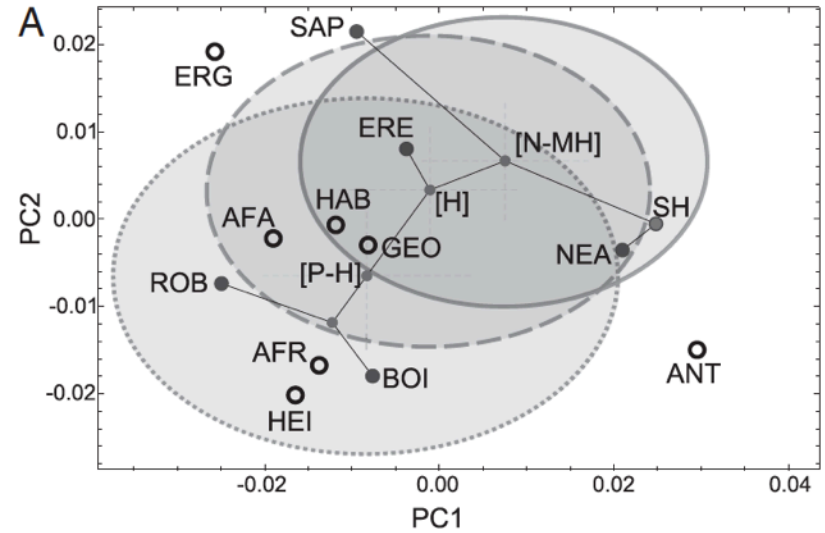


No known hominin species matches the expected dental morphology of the last common ancestor of Neanderthals and modern humans

Aida Gómez-Robles[a,b,1], José María Bermúdez de Castro[c], Juan-Luis Arsuaga[d], Eudald Carbonell[e], and P. David Polly[f]







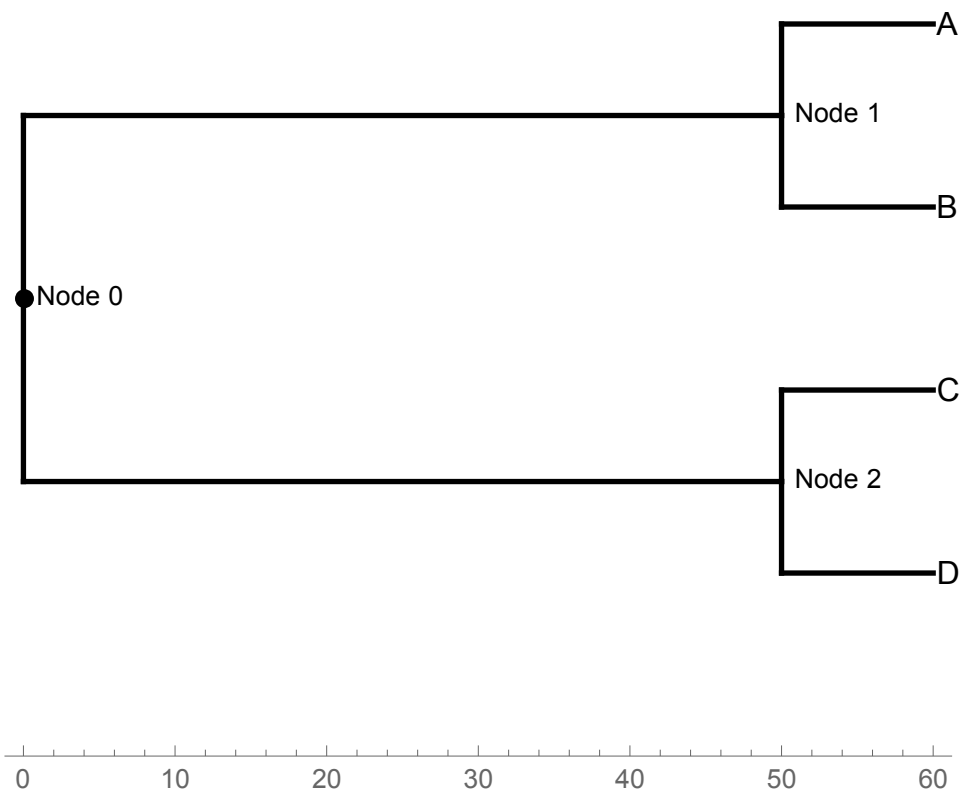
Example: *ProbabilitiesOfShapesAsAncestors[proc, labels, tree]*

	Node 0	Node 1	Node 2	Node 3	Node 4	Node 5
A.africanus	**0.8920**	**0.7500**	0.0249	0.0130	0.0002	0.0000
H.habilis	**0.6460**	**0.9110**	**0.2520**	**0.6490**	**0.2460**	0.0009
H.ergaster	0.0392	0.0022	0.0000	0.0002	0.0001	0.0000
H.georgicus	**0.3120**	0.0099	0.0000	0.0001	0.0000	0.0000
H.antecessor	0.0109	0.0008	0.0000	0.0001	0.0000	0.0000
H.heidelbergensis	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000

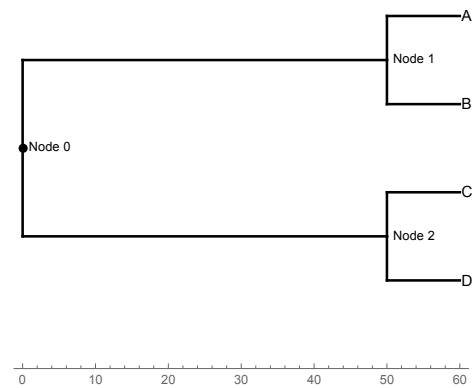
Gómez-Robles et a.. 2013. *PNAS,* 110: 18196-18201.

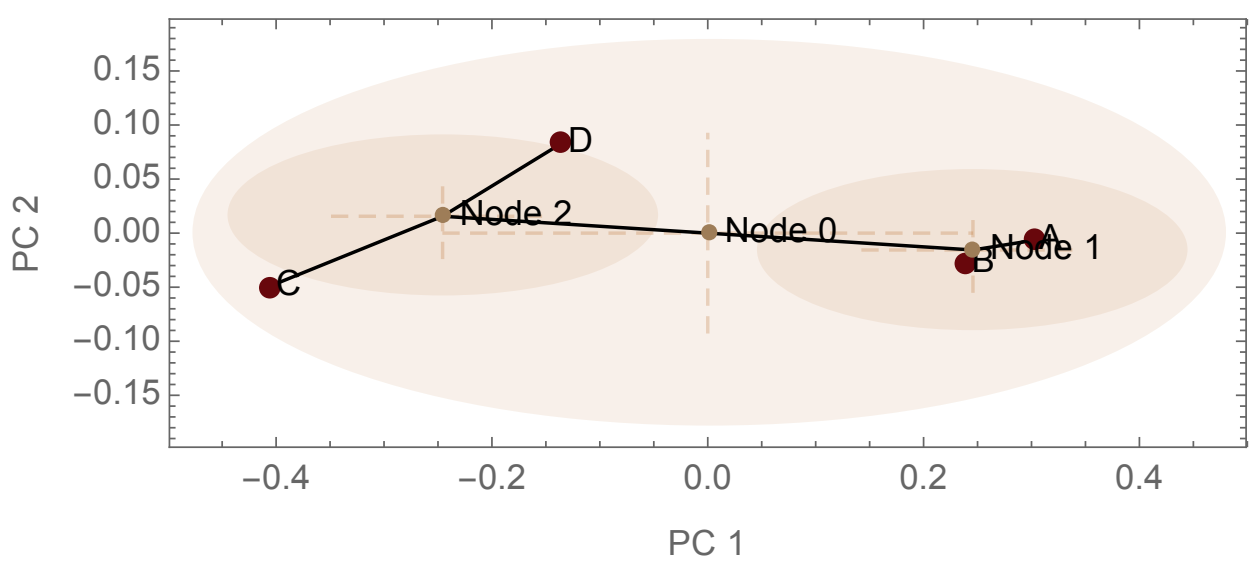
What relationship do we expect between phylogeny and PCA space?

how will these taxa be arranged in a PCA?

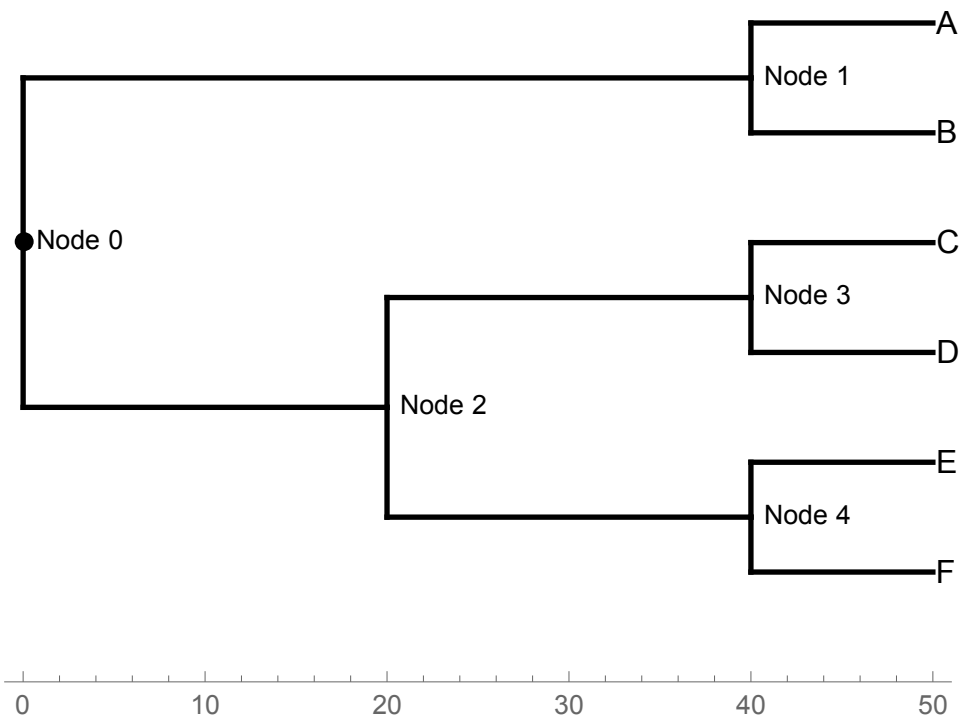


Two clades separated by PC 1

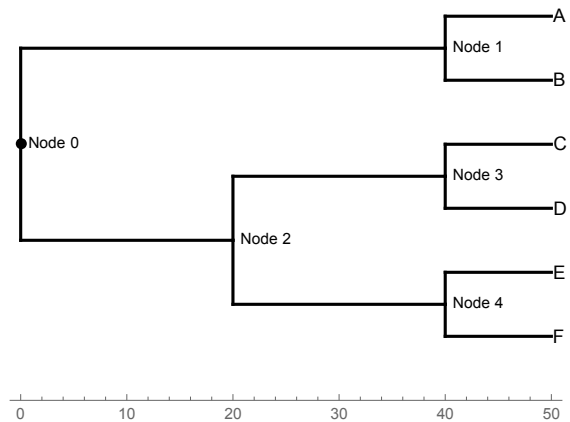


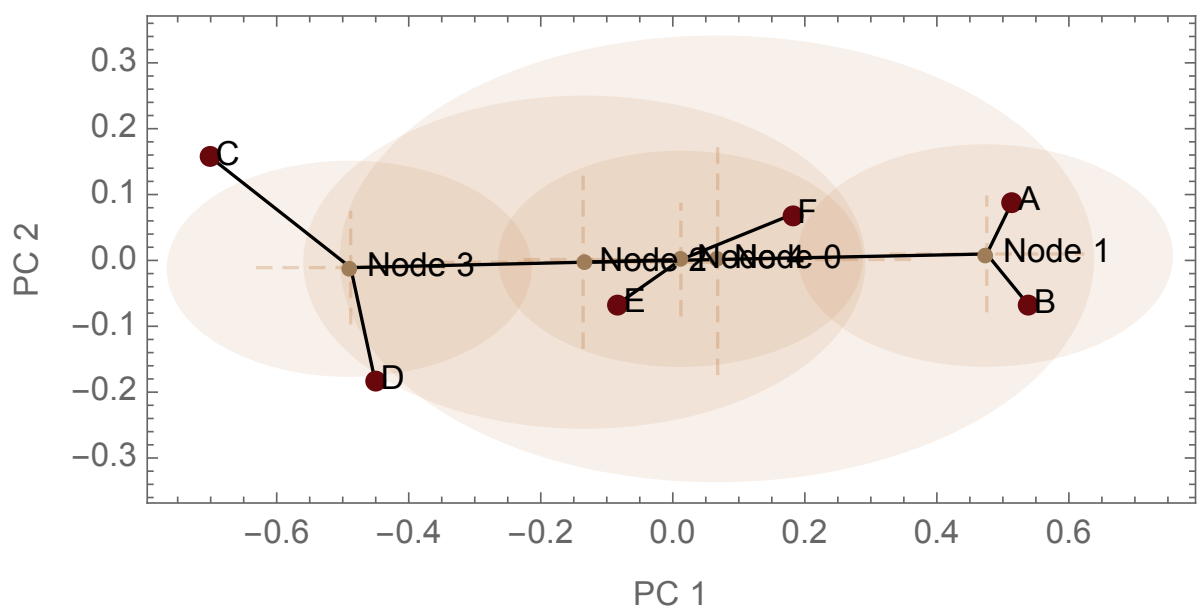


These?

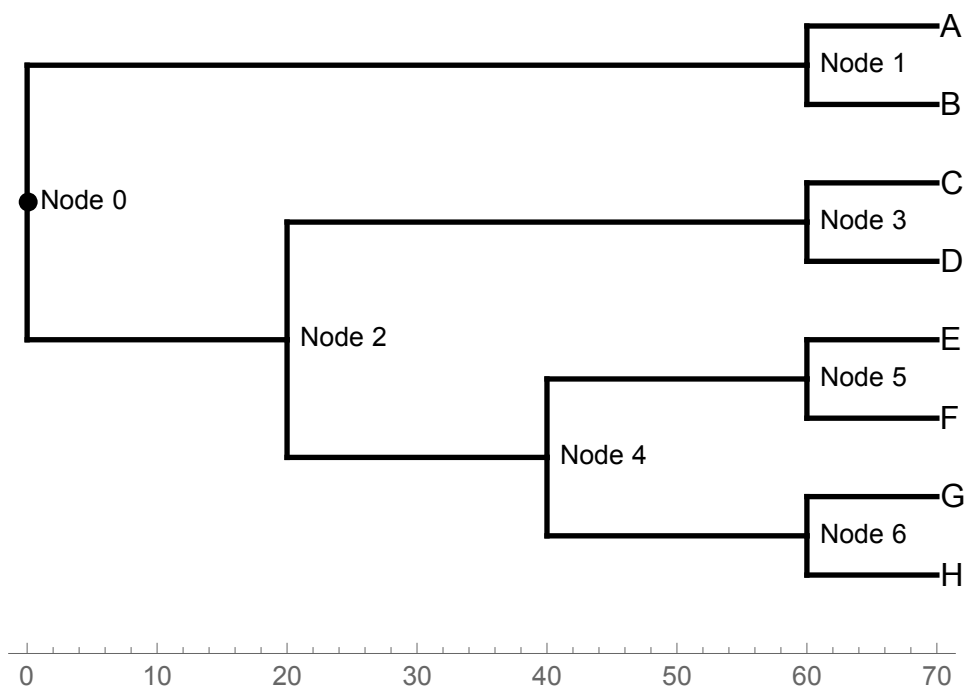


PC 1 = deep node, PC 3 = shallow node

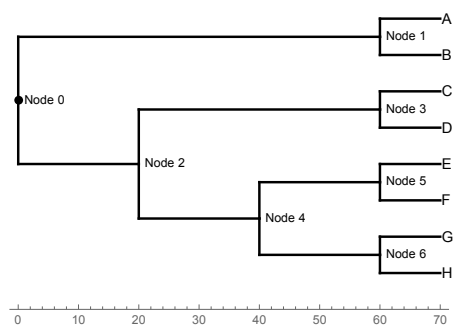


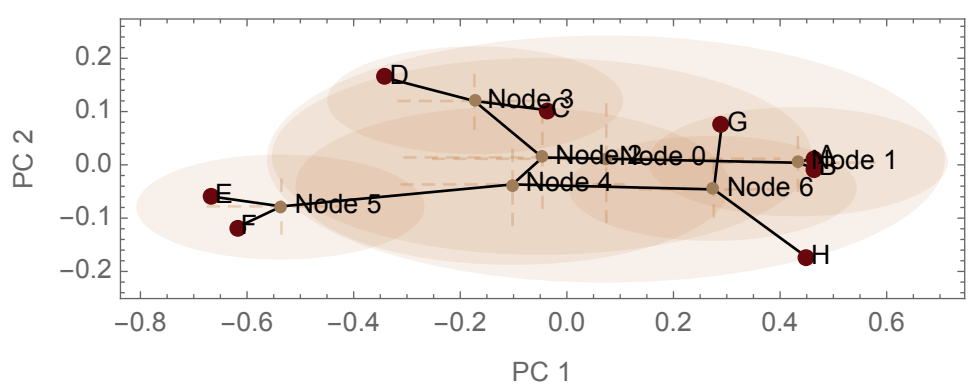


This?



PC 1 = A,B, PC 2 = C,D





PCMs useful in GMM

Phylogenetic Independent Contrasts (PIC or IC)

- Flexible method in which data are transformed to change along non-overlapping segments of a tree (hence "contrasts")
- Analogous to "first differencing" in time series analysis
- Contrasts are calculated for each trait of interest, then treated as phylogeny-free variables in ordinary statistical tests (regression, MANOVA, etc.)
- Developed by Felsenstein (1985). Multivariate implementation for GMM by Klingenberg (1996)

Phylogenetic Generalized Least Squares (PGLS)

- Uses C matrix to correct covariance structure in a least-squares regression
- Analogous to regressing out phylogenetic covariance as part of a trait-on-trait regression
- Results are same as with PIC or ML methods under Brownian motion
- Developed by Martins & Hansen (1997)

Phylogenetic MANOVA

- Essentially a MANOVA on independent contrasts for testing
- Developed by Revell et al. (2007)

Phylogenetic MANCOVA

- Extension of PGLS for assessing group differences holding a continuous covariate constants
- Developed by Smaers and Rohlf (2016)

Adams vs. Klingenberg approaches

Phylogenetic Independent Contrasts (PIC) approach of Klingenberg

- Widely used approach, including the implementation in Chris Klingenberg's *MorphoJ* system
- Method is simple:
  - Procrustes coordinates transformed to PICs prior to analysis (assumes BM model)
  - Regressions, MANOVA, etc. performed on the PIC matrix and summed across variables (e.g., fully multivariate)

Phylogenetic transformation approach of Adams

- Newer approach by Dean Adams implemented in the *geomorph* package, designed as high-dimensional alternative to PIC and PGLS
- Method is more abstract:
  - Data are rotated into a "phylogenetic space" based on the C matrix
  - Distance matrices are then calculated. Distances circumvent problems due to high numbers of variables in GMM data. (capitalizes on the Q-mode / R-mode equivalence)
  - Regression,MANOVA and other model fitting are based on the distance matrices to produce P, R, R^2, and other test statistics

Visualization of results is difficult with PCMs

Transformations for phylogenetic correction (PICs, C matrix, etc.) distort the shape space (indeed, even the variables on which it is based are changed with PICs)

Visualizing results as landmark deformations therefore requires a convoluted process and is sometimes impossible

When visualization is possible, it shows the pure relationship between independent variable and shape (e.g., between diet and mandible shape), without differences related to phylogeny

If relationship to independent factor is also strongly phylogenetic (e.g.. if each clade has a different and unique dietary specialization), PCMs may “over correct” and remove the variation that is actually associated with the factor

Pitfalls of Phylogenetic Principal Components Analysis (pPCA)

pPCA is a PCA-like analysis based on a covariance matrix that has been adjusted with the C matrix

developed by Revell (2009) and available in *R*

pPCA has many undesirable properties:

- pPCA does not remove effects of phylogeny
- pPCA axes are correlated
- pPCA eigenvalues do not describe the variance on the pPC axes
- pPC 1 does not describe the greatest variance in the data
- pPCA does not affect the outcome of statistical tests

Recommendation: do not use pPCA. Use normal PCM methods instead.

For more details see:
Polly et al. 2013. Phylogenetic principal components analysis and geometric morphometrics. *Hystrix*, 24: 1-9.