Comparative Methods

3 February 2017

SESYNC

Comparative Biology

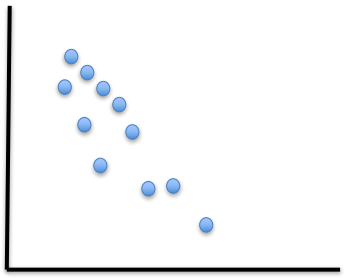
- e.g. taller trees should have bigger seeds

experimental evolution: often not practical

interspecific comparison: test whether traits correlate across species

- problem: related species may share the same traits due to shared ancestry = phylogenetic non-independence
- result is that species cannot be taken as independent data points

Traditional Correlation Tests

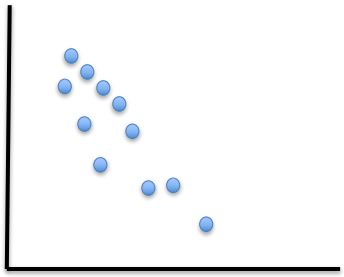


Central Assumptions in Statistics

-Normally distributed data (parametric statistics)

-Independent observations
- Points are independent

Traditional Correlation Tests

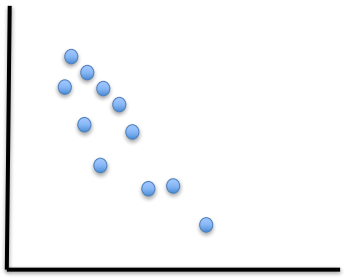


Central Assumptions in Statistics

-Normally distributed data (parametric statistics)

-Independent observations
- Points are independent

Problem with Traditional Correlation Tests



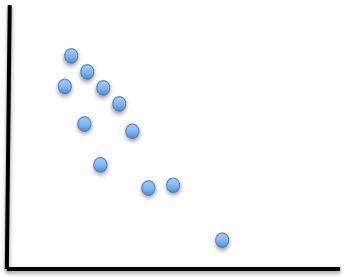
Central Assumptions in Statistics

-Normally distributed data (parametric statistics)

-Independent observations
- Points are independent

Species share evolutionary history and therefore are not independent observations

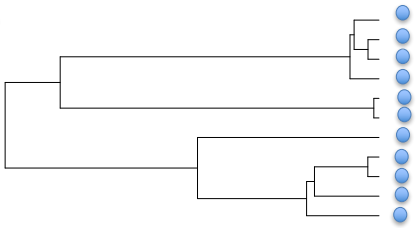
Problem with Traditional Correlation Tests



Central Assumptions in Statistics

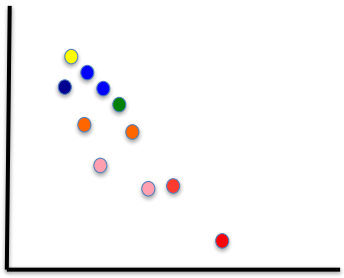
-Normally distributed data (parametric statistics)

-Independent observations
- Points are independent



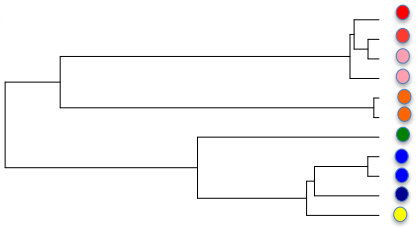
Problem with Traditional Correlation Tests

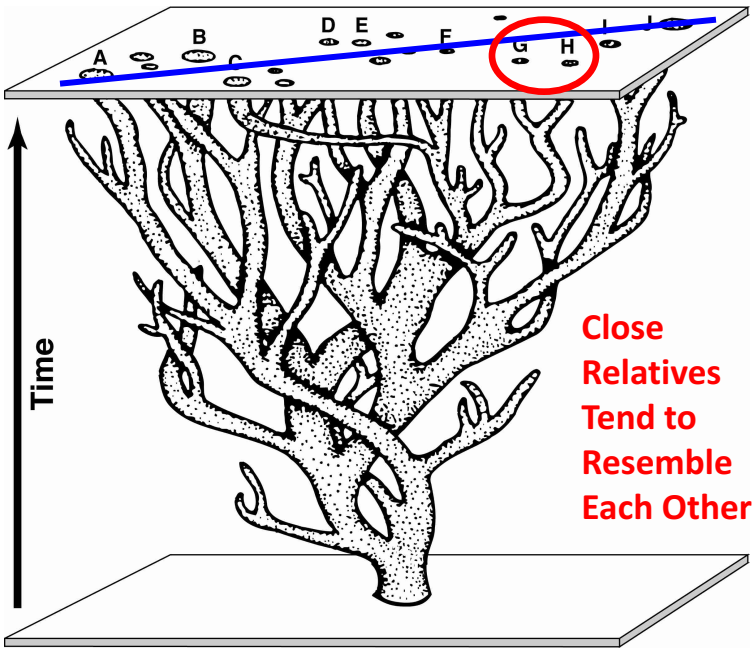
Central Assumptions in Statistics



-Normally distributed data (parametric statistics)

-Independent observations
- Points are independent





Solution

- Because species share common ancestors they cannot be considered independent observations in a correlation study.
- Therefore a central assumption of correlation statistics is violated
- Is there a solution?

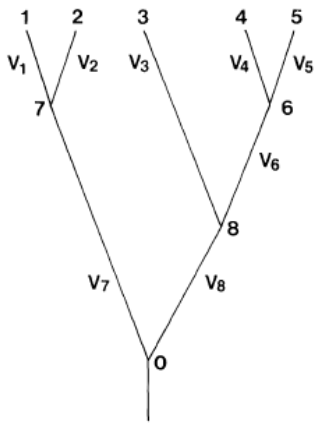
Many Solutions

- Originally using paired species comparisons
- Methods with an underlying model of trait evolution
  - Phylogenetically independent contrasts (PIC)
  - Phylogenetic generalized least squares (pGLS)
- Methods with no underlying model
  - Phylogenetic eigenvectors

Phylogenetically Independent Contrasts

Felsenstein's Solution Step 3

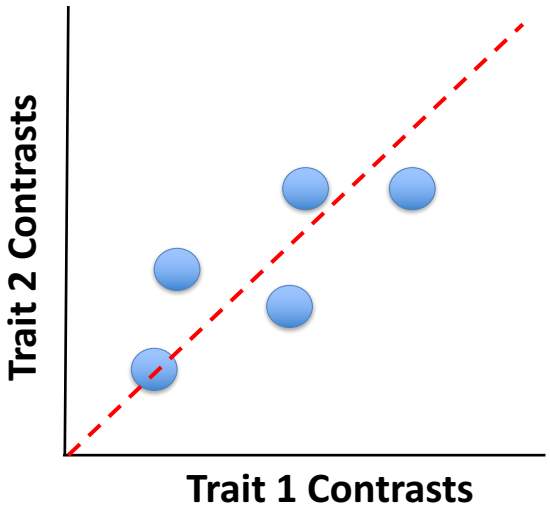
COMPARATIVE METHOD



CONTRAST
$X_1 - X_2$
$X_4 - X_5$
$X_3 - X_6$
$X_7 - X_8$

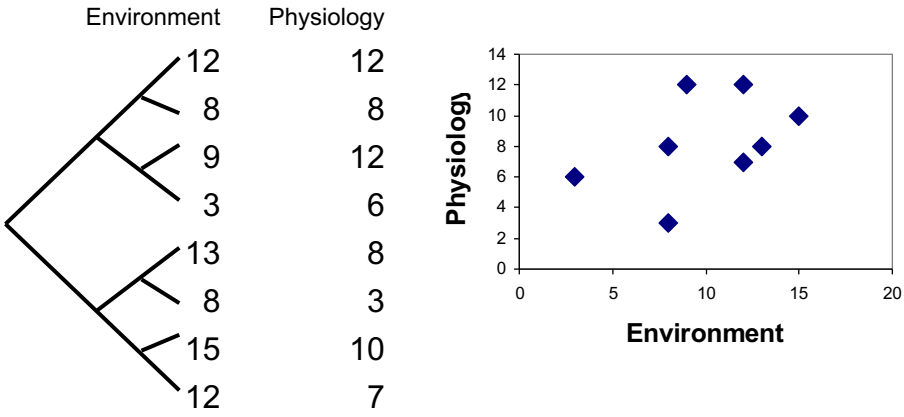
Calculate Regression Through Origin Between Contrasts

Felsenstein's Solution Step 3

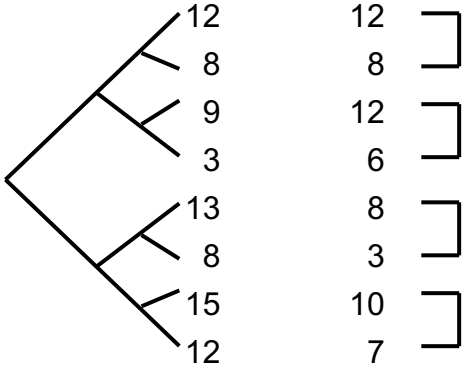


Example of analysis that takes phylogeny into account

- Phylogenetically independent contrasts



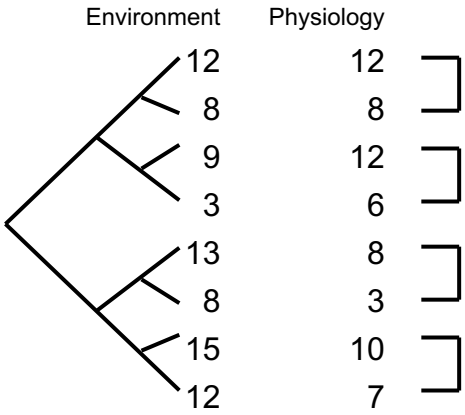
Phylogenetically independent contrasts



Contrasts

Environment	Physiology
12 - 8 = 4	12 – 8 = 4
9 – 3 = 6	12 – 6 = 6
13 – 8 = 5	8 – 3 = 5
15 – 12 = 3	10 – 7 = 3

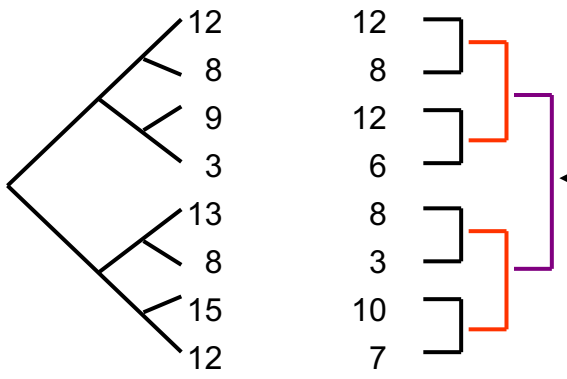
Phylogenetically independent contrasts



Contrasts

Environment	Physiology
12 - 8 = 4	12 – 8 = 4
10 – 6 = 4	10 – 9 = 1
9 – 3 = 6	12 – 6 = 6
13 – 8 = 5	8 – 3 = 5
10.5–13.5=-3	5.5-8.5 = -3
15 – 12 = 3	10 – 7 = 3

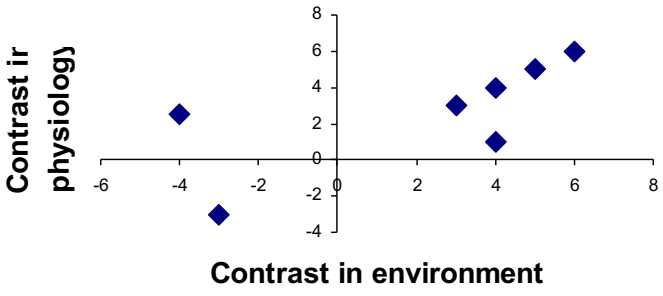
Phylogenetically independent contrasts



Contrasts

Environment	Physiology
12 - 8 = 4	12 – 8 = 4
10 – 6 = 4	10 – 9 = 1
9 – 3 = 6	12 – 6 = 6
8 -12 = -4	9.5 – 7 = 2.5
13 – 8 = 5	8 – 3 = 5
10.5–13.5=-3	5.5-8.5 = -3
15 – 12 = 3	10 – 7 = 3

Phylogenetically independent contrasts



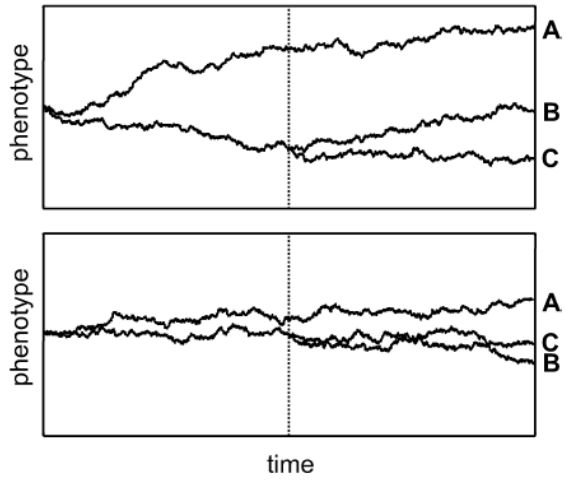
Contrasts

Environment	Physiology
4	4
4	1
6	6
-4	2.5
5	5
-3	-3
3	3

PIC ASSUMPTIONS

- The phylogenetic topology is 'correct'. IE we know the correct phylogeny
  - Phylogenies as hypotheses
- Branch lengths are measured in the units expected to correlate with character evolution
- We know the species trait values
- Traits evolve under a Brownian motion model
  - Increasingly, people are developing methods to rely on other potential models or that fit models given the data

Brownian Motion



Limits to phylogenetically based analyses

- What if a particular adaptation only occurred once?
  - Don't want to ignore a "neat story" that is supported by other kinds of evidence
- Trait relationships may be non-linear
  - Must transform data to satisfy BM assumption
  - Reduces statistical power

How might this influence my statistical inference?

- This depends on the degree of phylogenetic signal in data
- If there is no phylogenetic signal in trait data, then your result is similar compared to traditional correlation
- If there is high phylogenetic signal in trait data, then your inference can be very different compared to traditional correlation

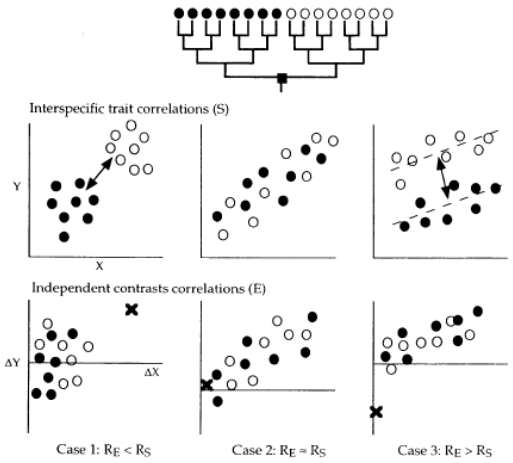


Figure 20.3 Hypothetical examples of correlated traits to illustrate how non-phylogenetic interspecific correlations may differ from phylogenetically structured comparisons based on independent contrasts. Upper panels: scatterplots of species trait values, with closed and open symbols for the two lineages in the phylogeny above; lower panels: scatterplots of corresponding independent contrasts, showing the contrasts within clades, and the contrast between clades with an 'X'. Case (1): if the correlation across species is primarily a result of a discrete difference between two major groups, then the correlation of independent contrasts will be weaker or non-existent. Case (2): if correlated evolution occurs within both clades, and there is little divergence between them, the correlation of independent contrasts is similar to the interspecific correlations. Case (3): a displacement between two groups, coupled with correlated evolution within groups, results in significant correlations using interspecific contrasts, but no correlation when species data are viewed as a whole.

What kind of data can (or should) be analyze using Phylogenetically Independent Contrasts?

- Where species are units of observation
  - Though, a handful of methods are now available that utilize variation or probability distributions w/i species
- Comparative Questions
  - Do species in drier soils have smaller leaves (data are soil moisture preference and leaf size for many species)
  - Is maximum height of species correlated with their seed size? (data are maximum heights and seed size of many species)

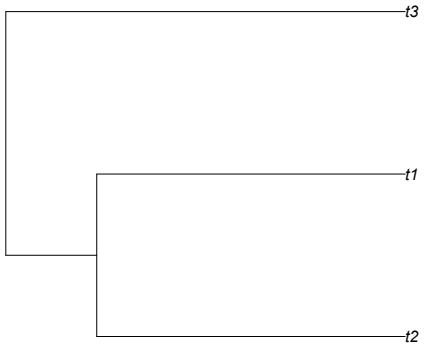
Phylogenetic GLS

Phylogenetic GLS

- In the simplest case equivalent to independent contrast analysis (Grafen 1989; Martins & Hansen 1997) but various extensions.
- One important extension is fitting a model of trait evolution to the residuals

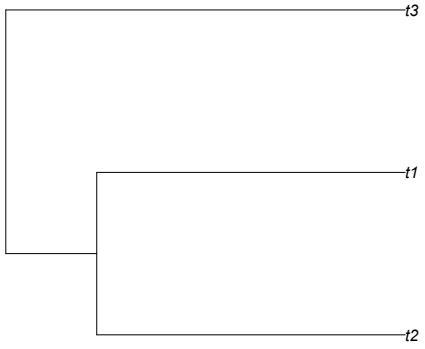
Phylogenetic GLS

- We start with a phylogeny
- Must convert it to a phylogenetic variance-covariance (VCV) matrix



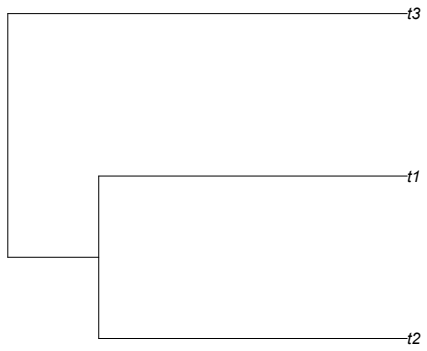
What is a phylogenetic VCV?

- A phylo VCV is constructed assuming a model of trait evolution.
  - Usually Brownian motion
  - Trait variance and co-variance scales with branch length



What is a phylogenetic VCV?

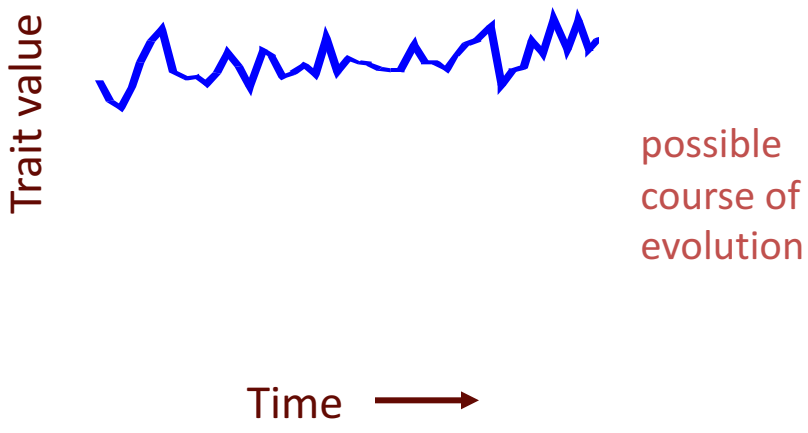
```
          t2        t1        t3
t2 0.6174106 0.1405757 0.0000000
t1 0.1405757 0.6174106 0.0000000
t3 0.0000000 0.0000000 0.6174106
```

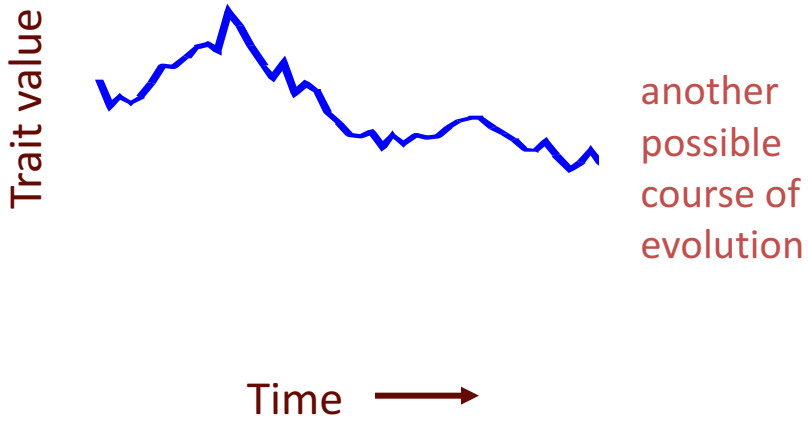


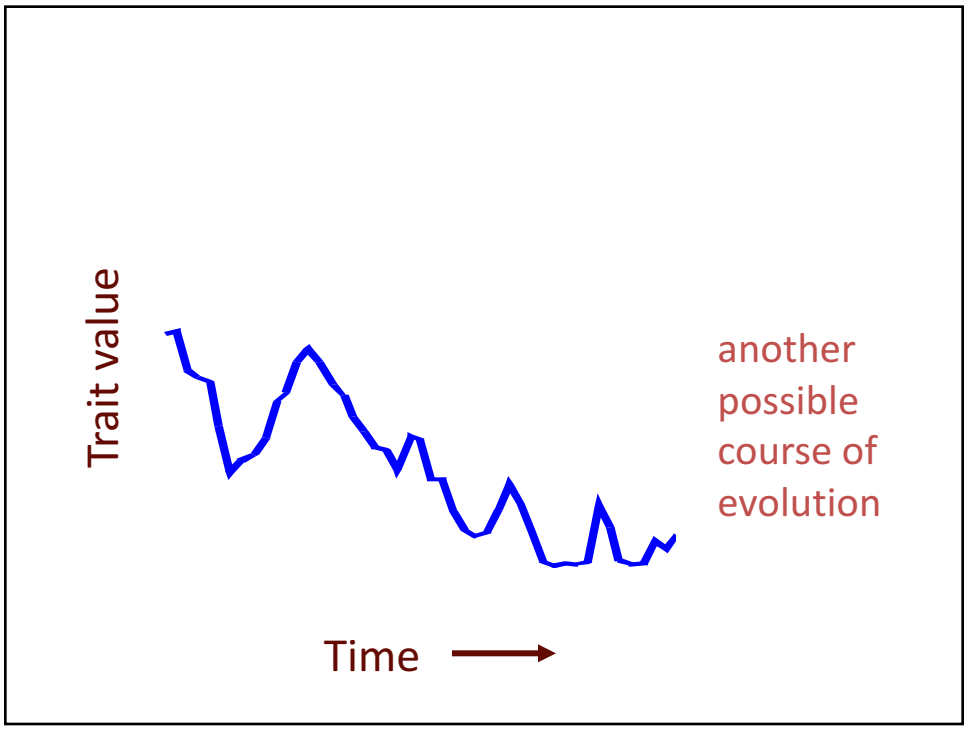
What is a phylogenetic VCV?

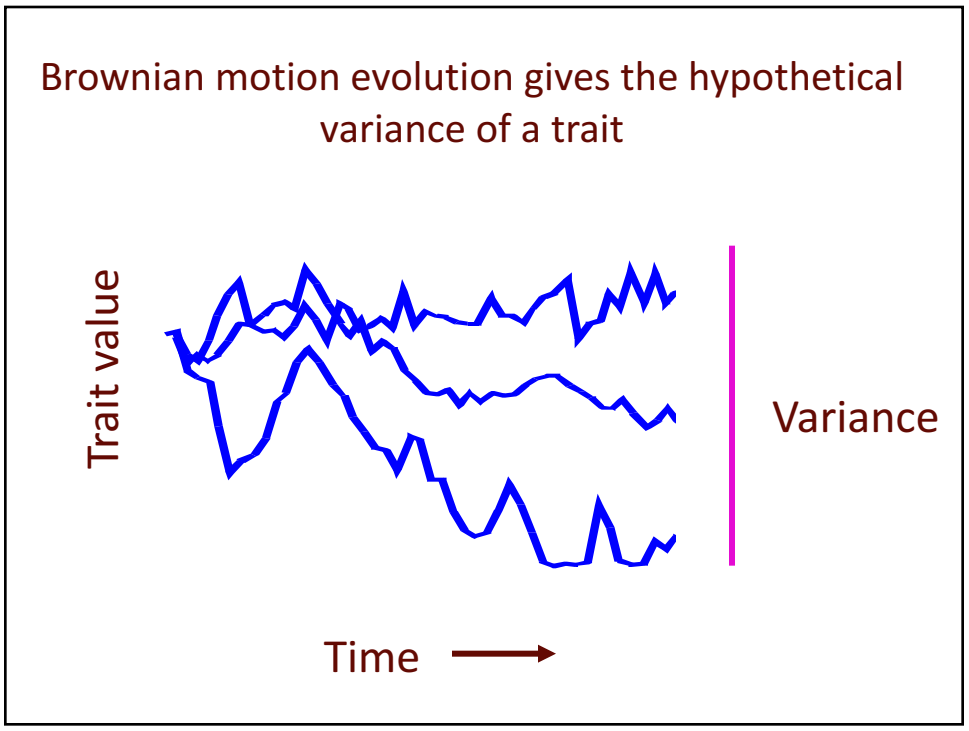
```
          t3        t1        t2
t3 0.5795273 0.5625709 0.0000000
t1 0.5625709 0.5795273 0.0000000
t2 0.0000000 0.0000000 0.5795273
```

Hypothetical trait for a single species under Brownian motion evolution

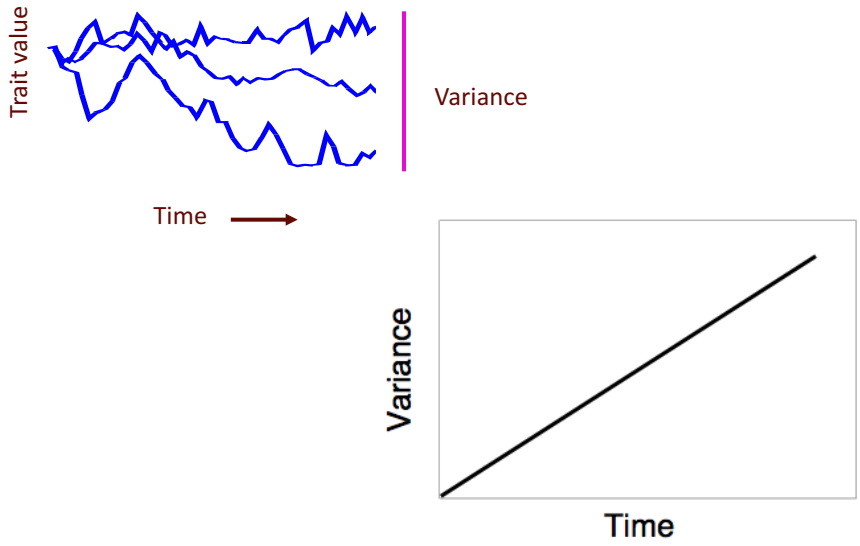




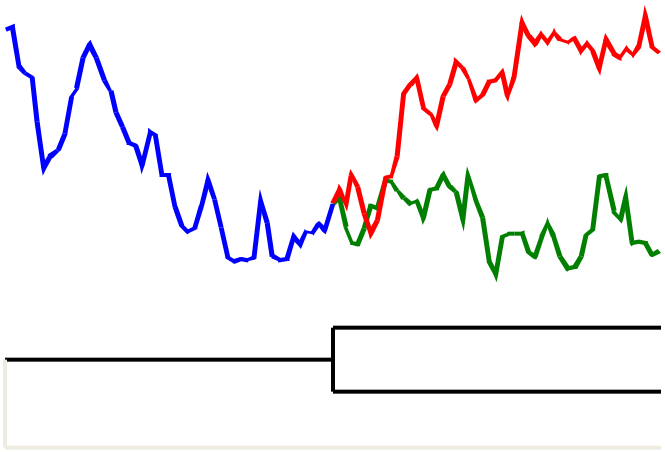

Trait value
another
possible
course of
evolution
Time

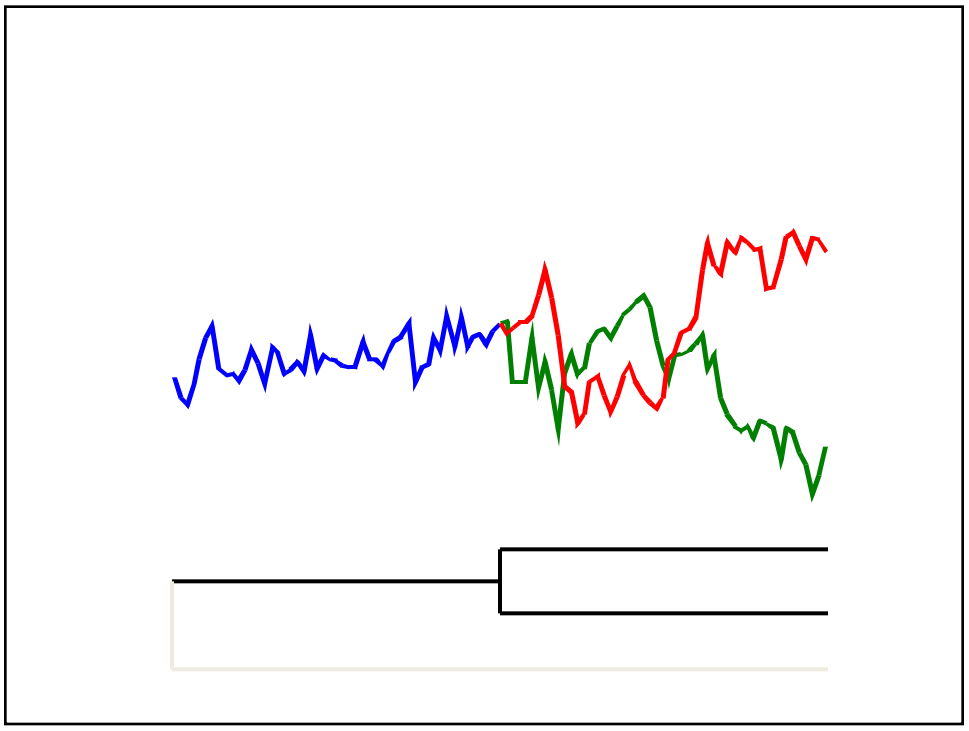

Brownian motion evolution gives the hypothetical
variance of a trait
Trait value
Variance
Time

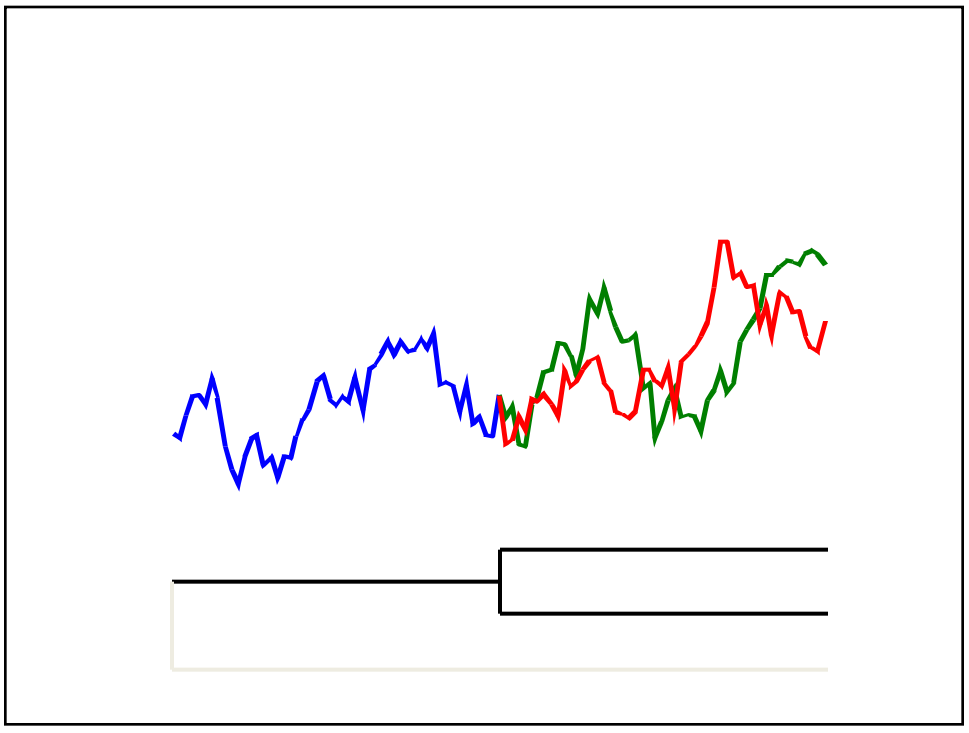
Brownian motion evolution

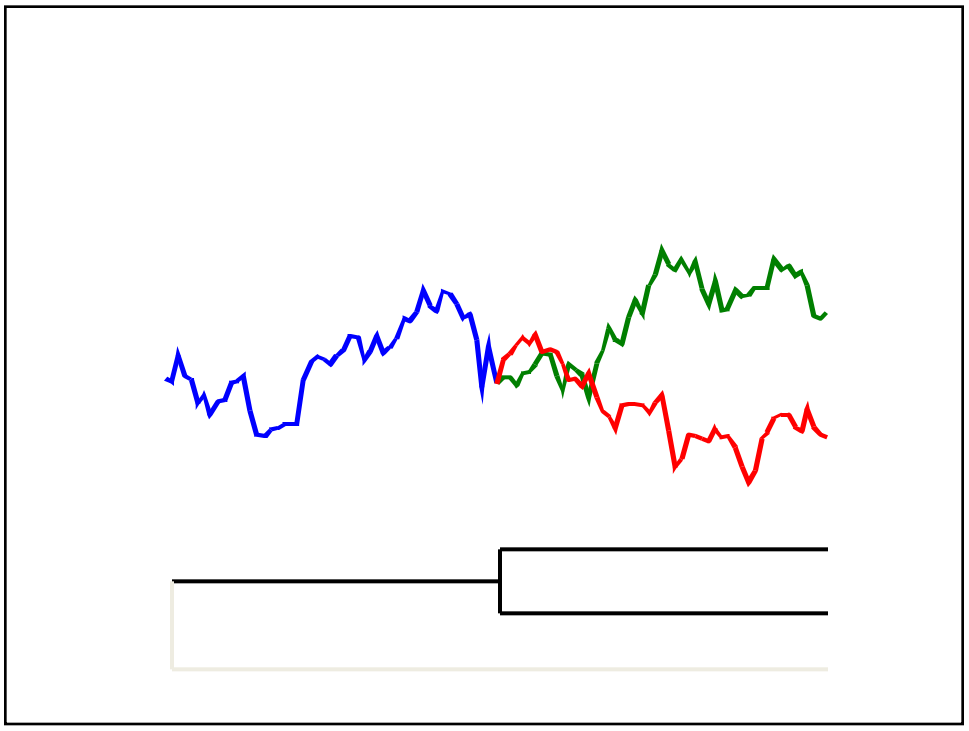


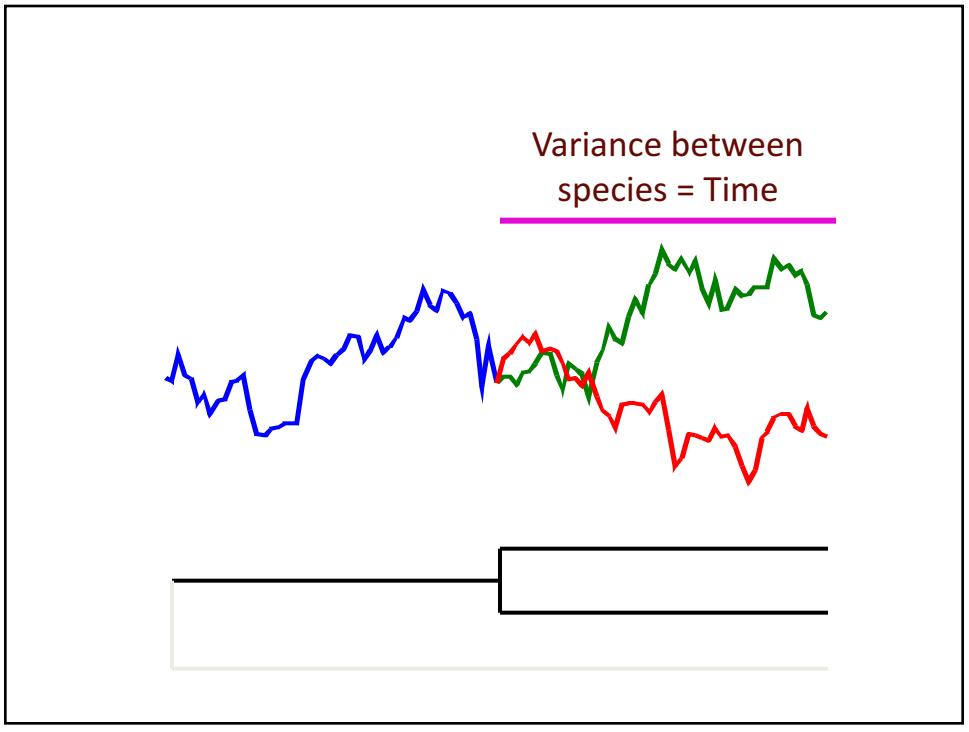
Brownian motion evolution of a hypothetical trait during speciation

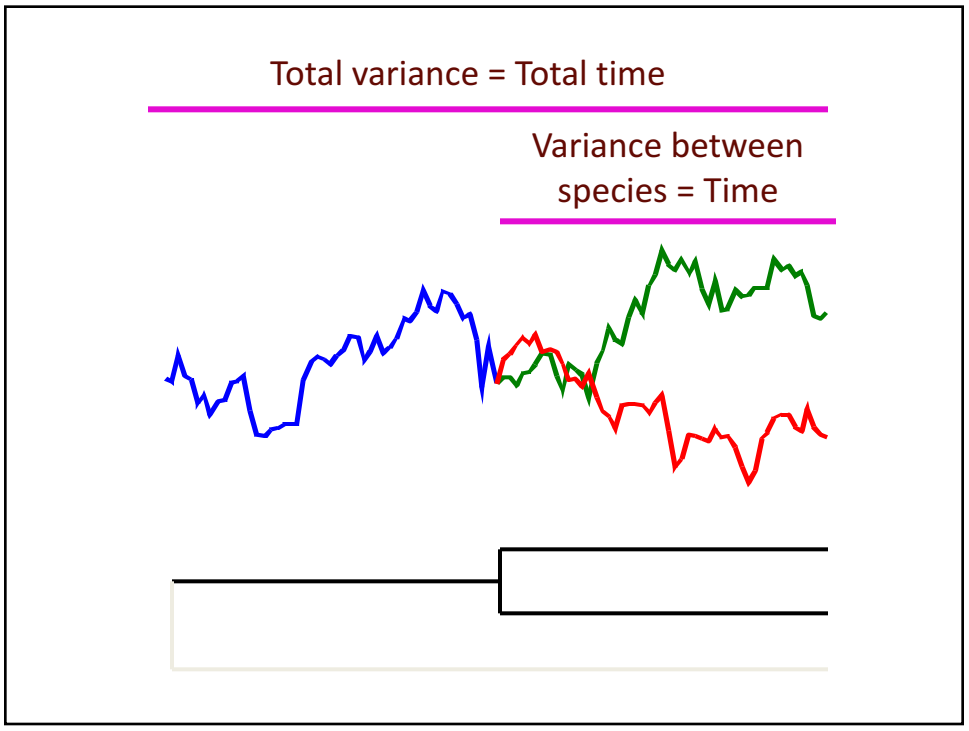


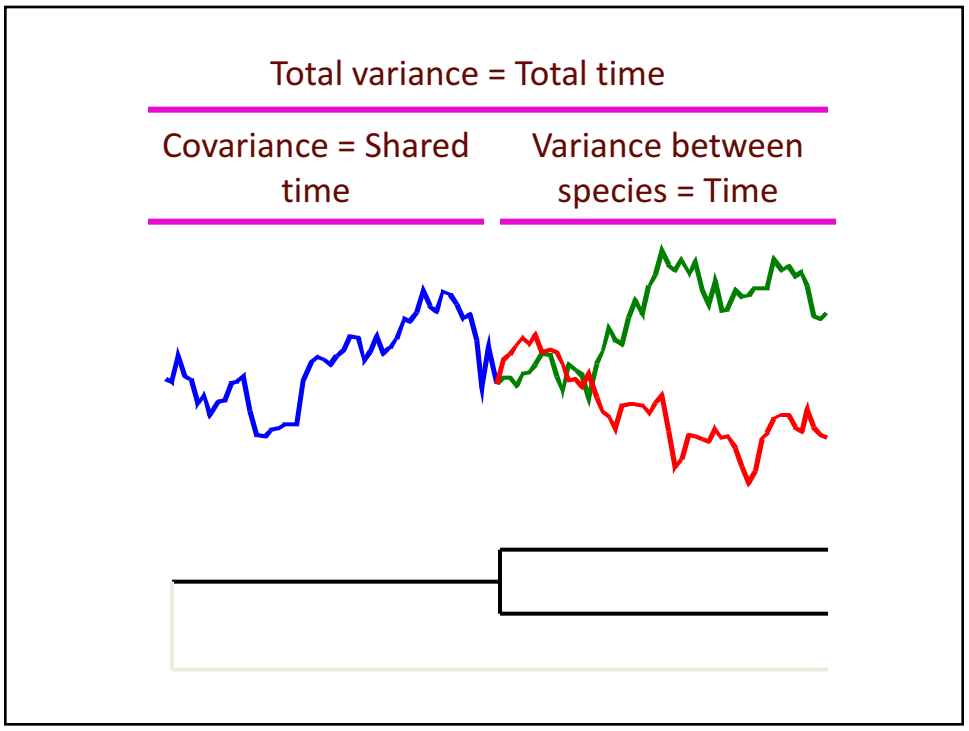


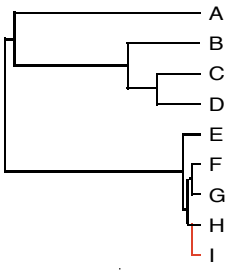





Variance between
species = Time


Total variance = Total time
Variance between species = Time


Total variance = Total time
Covariance = Shared time
Variance between species = Time



Brownian motion $\Rightarrow$ **V**

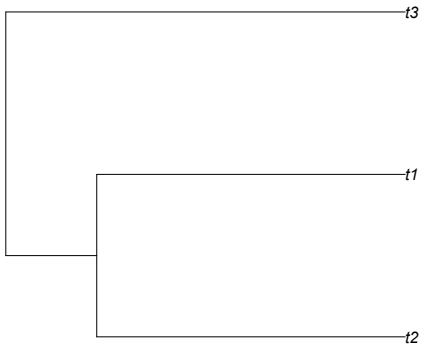
V Covariance matrix giving phylogenetic covariances among species

v_{ii} diagonal elements give the total variance for species i

v_{ij} off-diagonal elements give covariances between species i and species j

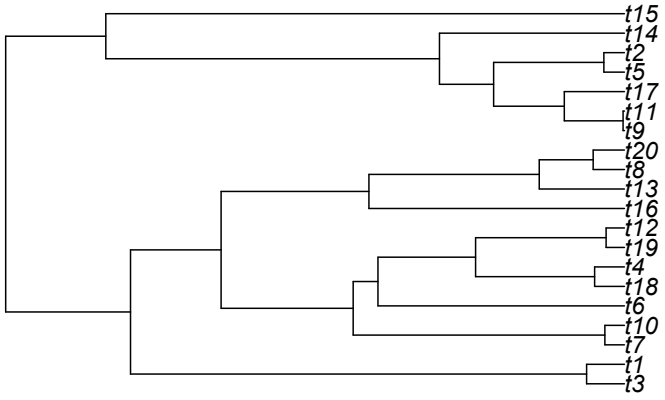
Phylo GLS

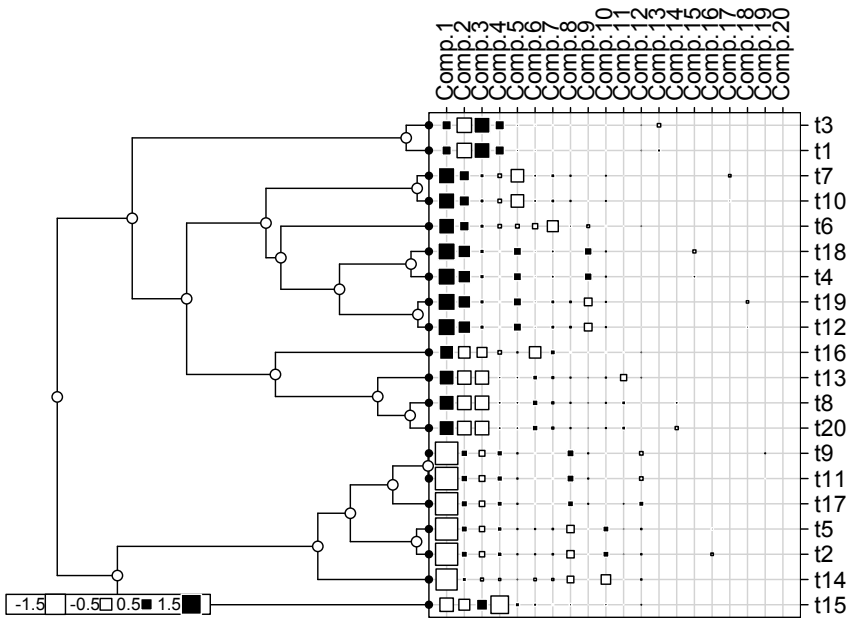
- The phylo VCV is then used in a generalized linear model
  - Can use the r command pGLS()



Phylogenetic Eigenvectors

- Inputs a phylogenetic distance matrix into a PCA to decompose the phylo into many eigenvectors





Phylogenetic Eigenvectors

- Phylogenetic eigenvectors are then used in a linear model similar
  - E.g. as independent variables in the model
- Some well discussed weaknesses
  - No explicit model of trait evolution
  - How do we pick which eigenvectors to use?

Phylogenetic ANOVAs/t-tests/PCAs/etc

- Phylogenetic ANOVAs and MANOVAS can be conducted (or t-tests for that matter)
- A standard way is to calculate a regular (M)ANOVA
- Next trait values are simulated on the phylogeny under a given model of trait evolution
  - This generates a null distribution

Intra-specific variation

- Error/variance
- Re-sampling from probability distribution

Our old friend the vcv

```
          t3        t1        t2
t3 0.5795273 0.5625709 0.0000000
t1 0.5625709 0.5795273 0.0000000
t2 0.0000000 0.0000000 0.5795273
```

Ives et al. 2007 sys biol

ANALYSES

Univariate Analyses and Phylogenetic Signal

The problem of finding the best estimator of the expectation of a random variable when there is phylogenetic correlation and measurement error is given by the statistical model

$$\mathbf{X}^* = a + \varepsilon \qquad (1)$$
$$\mathbf{X} = \mathbf{X}^* + \eta$$

where $\mathbf{X}^*$ is a $N \times 1$ dimensional vector containing the true values of a trait in a sample of N species (tips), a is a scalar giving the expected value of the trait, ε is a $N \times 1$ vector of zero-mean error terms depicting the evolutionary variance of the trait among species, $\mathbf{X}$ is a $N \times 1$ vector containing the observed values of the trait, and η is the $N \times 1$ vector of errors associated with measurement. Note that for notational convenience we have written $\mathbf{X}^* = a + \varepsilon$ as the sum of a scalar and a vector to represent $\mathbf{X}^* = a\mathbf{1} + \varepsilon$ where $\mathbf{1}$ is the $N \times 1$ vector of ones.

Because closely related species will likely have similar values of trait $\mathbf{x}$, values of ε will be correlated among species. Thus, we assume the covariance matrix for ε is given by $\mathrm{E}\{\varepsilon\varepsilon'\} = \sigma^2\mathbf{C}$, where σ^2 scales the overall phylogenetically inherited variance (sometimes referred to as the rate of evolution; Garland et al., 1999; Garland and Ives, 2000), and $\mathbf{C}$ gives the correlation structure created by phylogenetic relatedness. The most common assumption in phylogenetic analyses is that evolution proceeds like a "Brownian motion" process; through time, the value of a trait changes in small increments in random directions, like a random walk in continuous time (Felsenstein, 1985). Under this assumption, ε has a multivariate normal distribution in which the element c_{ij} of $\mathbf{C}$ is proportional to the length of the shared

Although we do not consider correlated measurement errors in detail, nonzero off-diagonal elements of $\mathbf{M}$ can be used in all of the methods we derive. Finally, although we will typically assume that ε and η have multivariate normal distributions, for some of the statistical procedures described below, ε and η need not be restricted to being normally distributed.

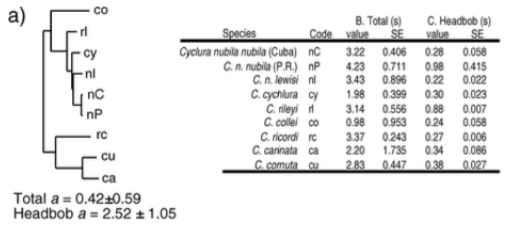
Consider first the case of no measurement error. Equation 1 can be reformulated as a phylogenetic regression problem in which the error terms are correlated, and hence can be analyzed using either independent contrasts or, as we will do here, generalized least squares (Hansen and Martins, 1996; Garland and Ives, 2000; Rohlf, 2001). Because $\mathbf{C}$ is a covariance matrix (and hence real, symmetric, and nonsingular), there exists another matrix $\mathbf{D}$ such that $\mathbf{DCD}' = \mathbf{I}$, where the apostrophe denotes transpose and $\mathbf{I}$ is the $N \times N$ identity matrix. Matrix $\mathbf{D}$ can be used to transform values of trait $\mathbf{x}$ by letting $\mathbf{Z} = \mathbf{DX}$, $\mathbf{U} = \mathbf{D1}$ (the $N \times 1$ vector of 1's), and $\alpha = \mathbf{D}\varepsilon$. From Equation 1 (with $\eta = 0$), this gives

$$\mathbf{Z} = \mathbf{U}a + \alpha \qquad (2)$$

The covariance matrix of α is $\mathrm{E}\{\alpha\alpha'\} = \mathrm{E}\{\mathbf{D}\varepsilon\ (\mathbf{D}\varepsilon)'\} = \mathrm{E}\{\mathbf{D}\varepsilon\varepsilon'\mathbf{D}'\} = \mathbf{D}\mathrm{E}\{\varepsilon\varepsilon'\}\mathbf{D}' = \mathbf{D}(\sigma^2\mathbf{C})\,\mathbf{D}' = \sigma^2\mathbf{I}$. Thus, no covariance terms appear in the covariance matrix of α, so the error terms α_i are uncorrelated. Equation 2 can, therefore, be analyzed as a standard least-squares regression problem with independent errors. Specifically, the generalized least-squares (GLS) estimator of a is

$$\hat{a} = \frac{\mathbf{U}'\mathbf{Z}}{\mathbf{U}'\mathbf{U}} = \frac{(\mathbf{D1})'(\mathbf{DX})}{(\mathbf{D1})'(\mathbf{D1})} = \frac{\mathbf{1}'\mathbf{D}'\mathbf{DY}}{\mathbf{1}'\mathbf{D}'\mathbf{D1}} = \frac{\mathbf{1}'\mathbf{C}^{-1}\mathbf{X}}{\mathbf{1}'\mathbf{C}^{-1}\mathbf{1}} \qquad (3)$$

The corresponding estimate of σ^2 is the mean squared



Species	Code	B. Total (s) value	B. Total (s) SE	C. Headbob (s) value	C. Headbob (s) SE
Cyclura nubila nubila (Cuba)	nC	3.22	0.406	0.28	0.058
C. n. nubila (P.R.)	nP	4.23	0.711	0.98	0.415
C. n. lewisi	nl	3.43	0.896	0.22	0.022
C. cychlura	cy	1.98	0.399	0.30	0.023
C. rileyi	rl	3.14	0.556	0.88	0.007
C. collei	co	0.98	0.953	0.24	0.058
C. ricordi	rc	3.37	0.243	0.27	0.006
C. carinata	ca	2.20	1.735	0.34	0.086
C. comuta	cu	2.83	0.447	0.38	0.027

FIGURE 1. For the univariate case using data from Martins and Lamont (1998), the effects of measurement error can be visualized by constructing a tree that corresponds to the covariance structure of the data combining both phylogenetic covariance and measurement error variance (see text). The measurement error variance lengthens the terminal branch segments of the tree, with the length of the tip extension giving the measurement error variance relative to the variance of the evolutionary process. (a) The phylogenetic tree from which the covariance matrix $\sigma^2\mathbf{C}$ is calculated. (b) The phylogenetic tree with the variance associated with measurement error for total display duration graphed onto the tips of the tree, thereby giving a graphical representation of the covariance matrix $\sigma^2\mathbf{C} + \sigma^2_z\mathbf{M}$. For comparison, (c) is like (b) but with measure error for another trait, headbob duration. By increasing the expected within-species variances without changing among-species covariances, measurement error decreases the among-species correlations in the observed data. The table gives trait values and standard errors of the measurements for both traits and estimates of the phylogenetic mean a are given for each tree.

Evans et al. 2009 AmNat

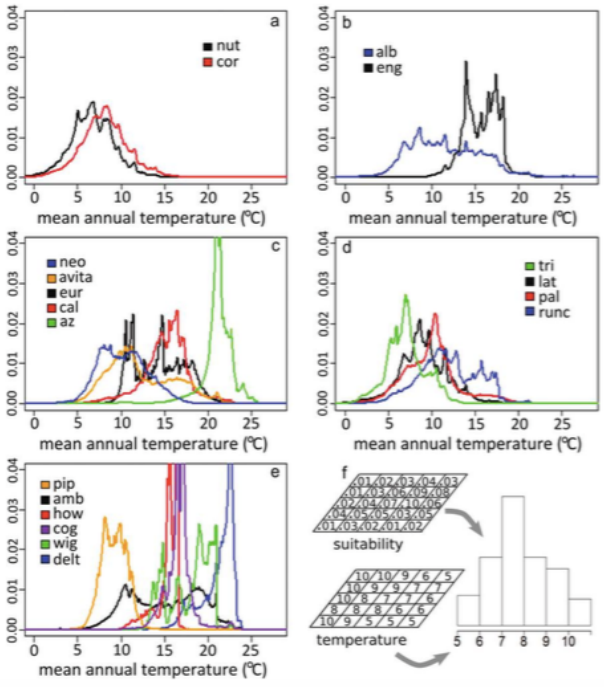
VOL. 173, NO. 2 THE AMERICAN NATURALIST FEBRUARY 2009

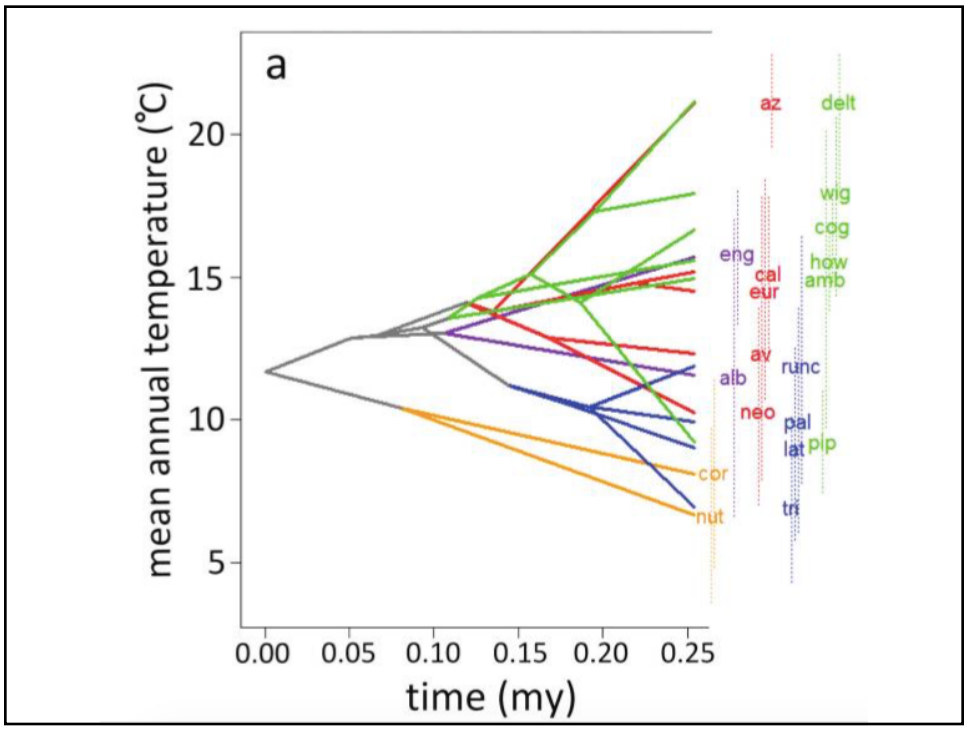
Climate, Niche Evolution, and Diversification of the "Bird-Cage" Evening Primroses (*Oenothera*, Sections *Anogra* and *Kleinia*)

Margaret E. K. Evans,[1,2,*] Stephen A. Smith,[1,†] Rachel S. Flynn,[1,‡] and Michael J. Donoghue[1,3,§]

Evans et al. 2009 AmNat

- Obtain a probability distribution for a trait
- Re-sample that distribution and estimate ancestral states each time




a
mean annual temperature (°C)
20
15
10
5
0.00
0.05
0.10
0.15
0.20
0.25
time (my)
az
delt
wig
cog
how
amb
eng
cal
eur
av
runc
alb
neo
pal
lat
pip
cor
nut
tri