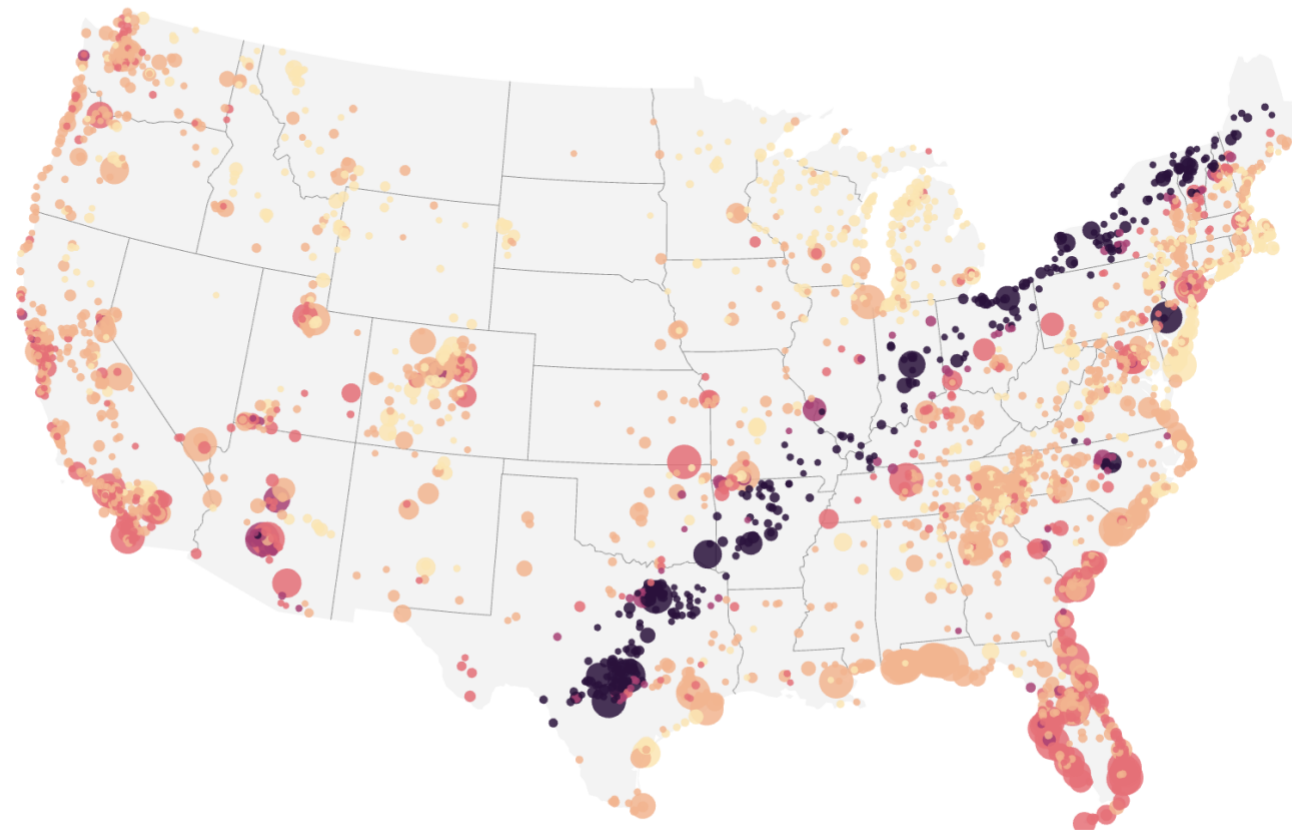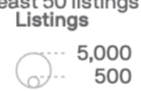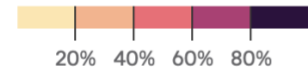# Graph of the week: Airbnb/Vrbo bookings illustrate path of totality

Share of Airbnb and Vrbo listings booked ahead of the April 8 eclipse

Bookings on April 7, 2024, among cities with at least 50 listings

20% 40% 60% 80%

Listings
5,000
500

Data: AirDNA; Map: Erin Davis/Axios Visuals

Species as data points

Outline for today

- The problem with species data

- Phylogenetic signal in ecological traits

- Why phylogeny matters in comparative study

- Phylogenetically independent contrasts (PICs)

- A linear model approach

- A method for categorical data (and issues)

- Many applications

- R: An embarrassment of riches

An example of species data

Mating behaviors in 15 species of water striders (*Gerris*). Males chase females, who flee by skating away. If a male grasps a female, she initiates a series of leaps, rolls, and summersaults that usually toss him off. Males of some species have clasping genitalia that allow them to stay on longer, but females of these species often have spines or other devices that make it difficult for males to grasp her. Mating takes place after a female stops struggling.

Rowe and Arnqvist (2002) measured <u>average duration of female struggles</u> for each species (the periods of evasive action by females in response to lunges or grasps by males); and <u>average mating frequency</u> of females, under controlled lab conditions.
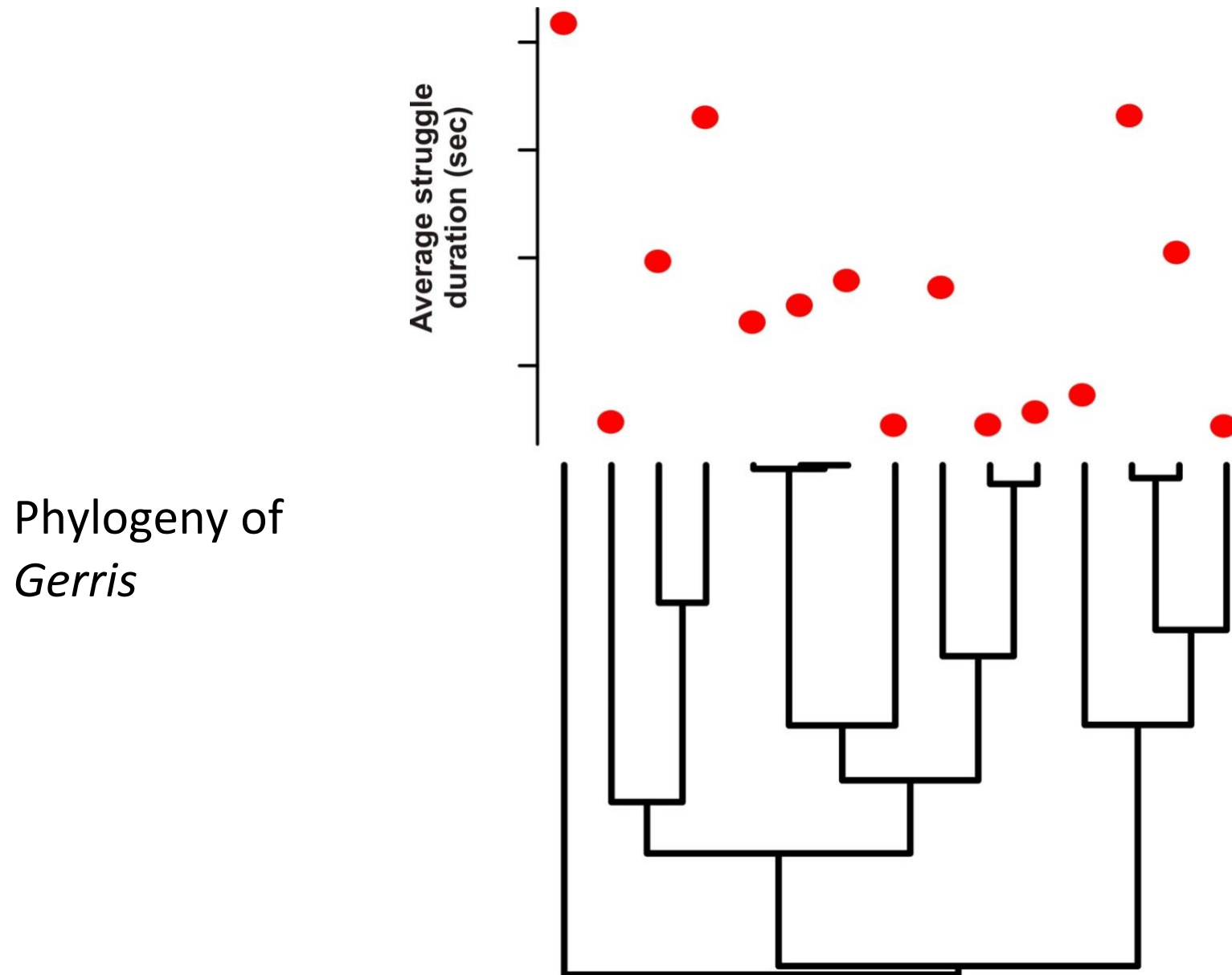
An example of species data

Data on 15 species reveal a positive association between the two variables.
We would like to estimate the strength of the correlation.
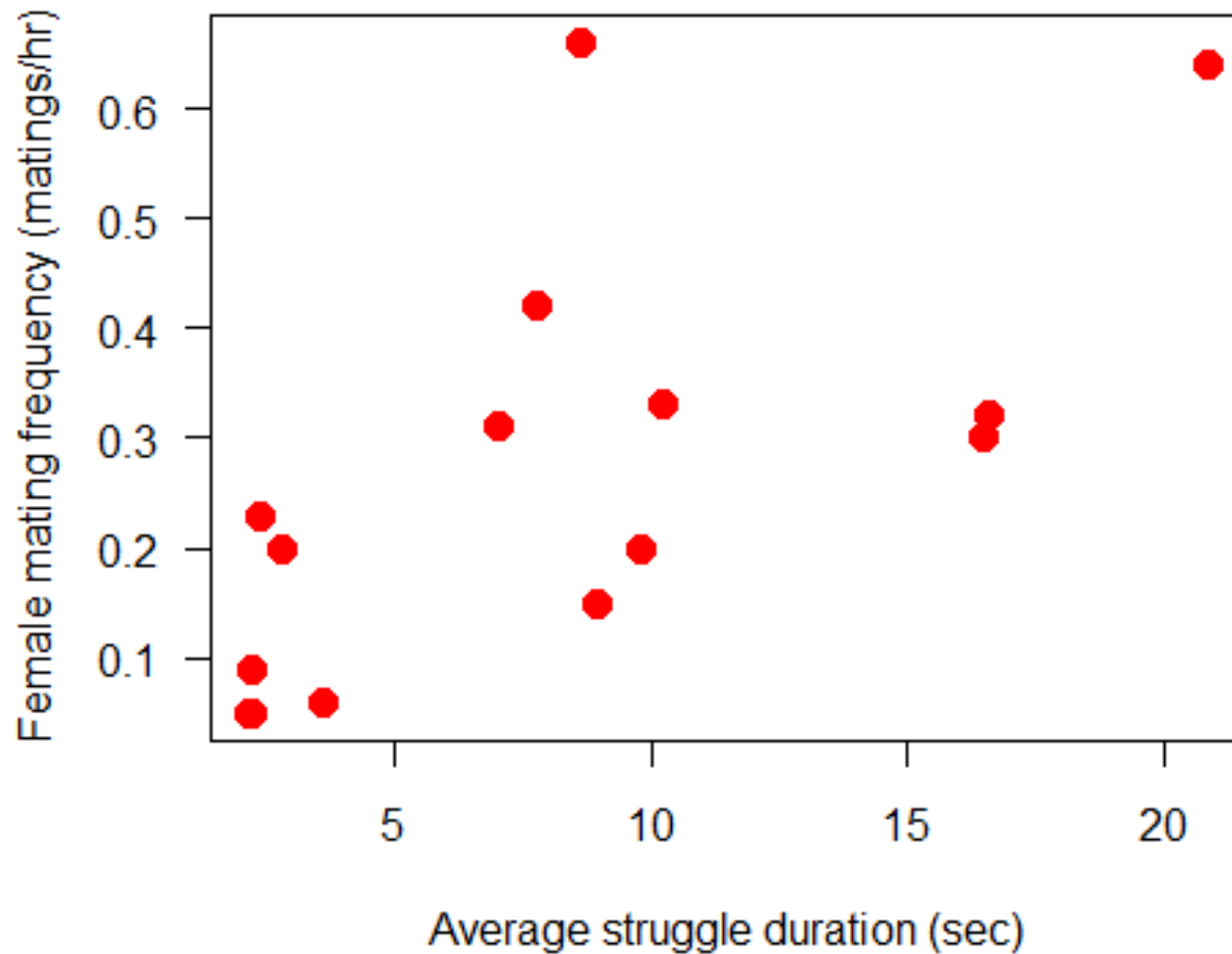
The problem with species data
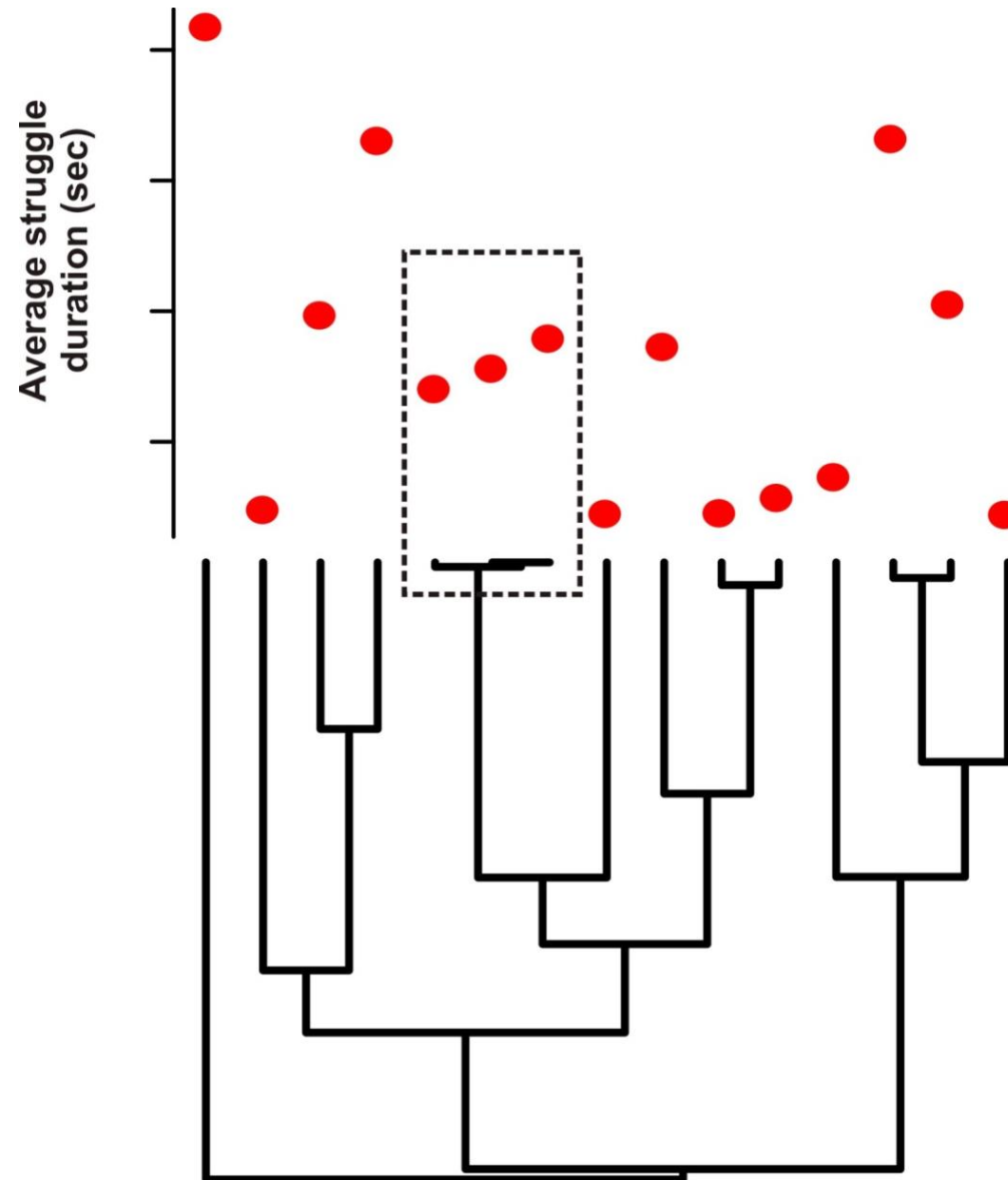
The data points (species) are not independent.
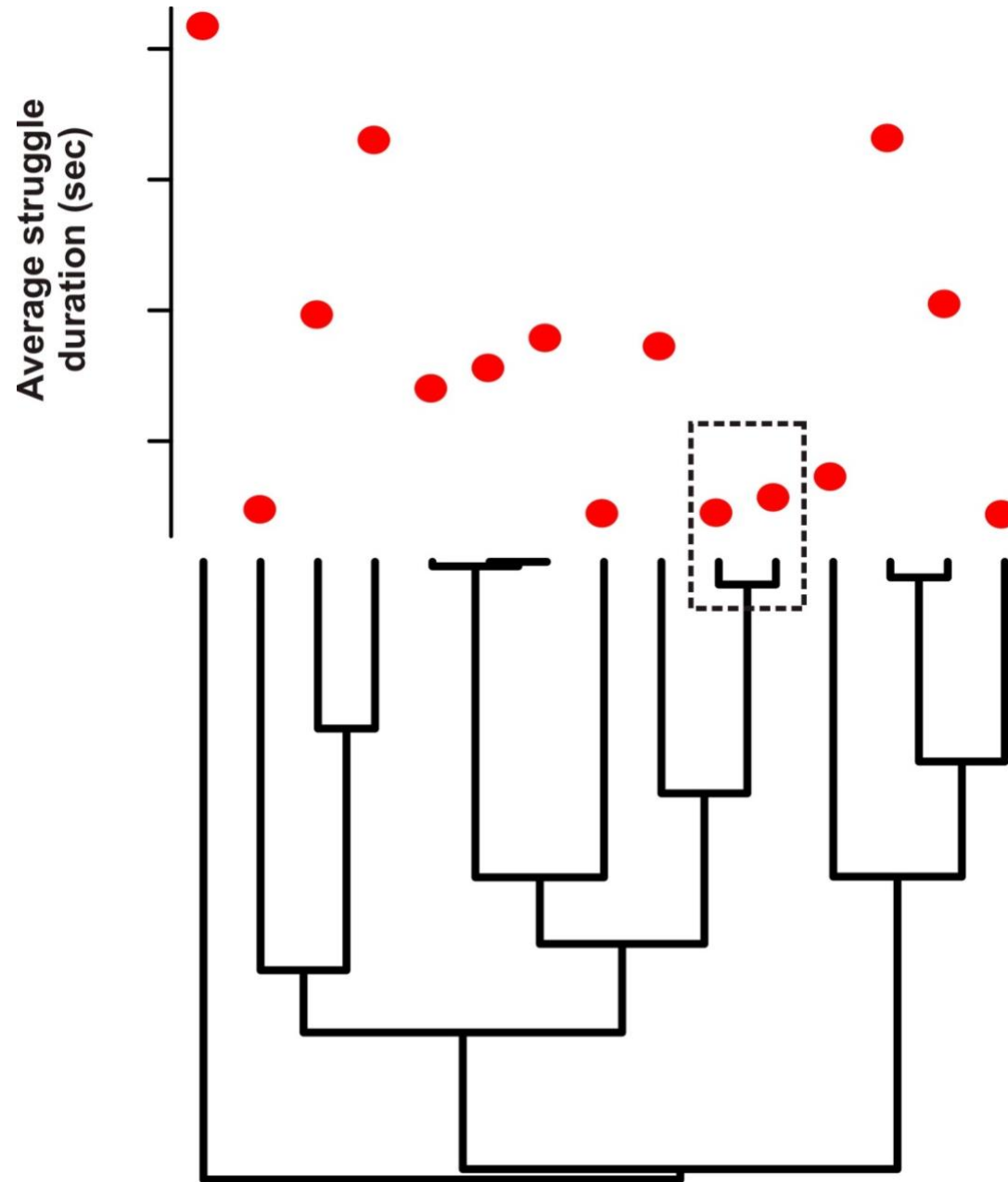
Phylogeny of
Gerris

The problem with species data

Closely related species tend to have similar trait values.

The problem with species data

This tendency is called "phylogenetic signal".

The problem with species data

Non-independence of species data violates a major assumption of conventional statistical methods for data analysis.
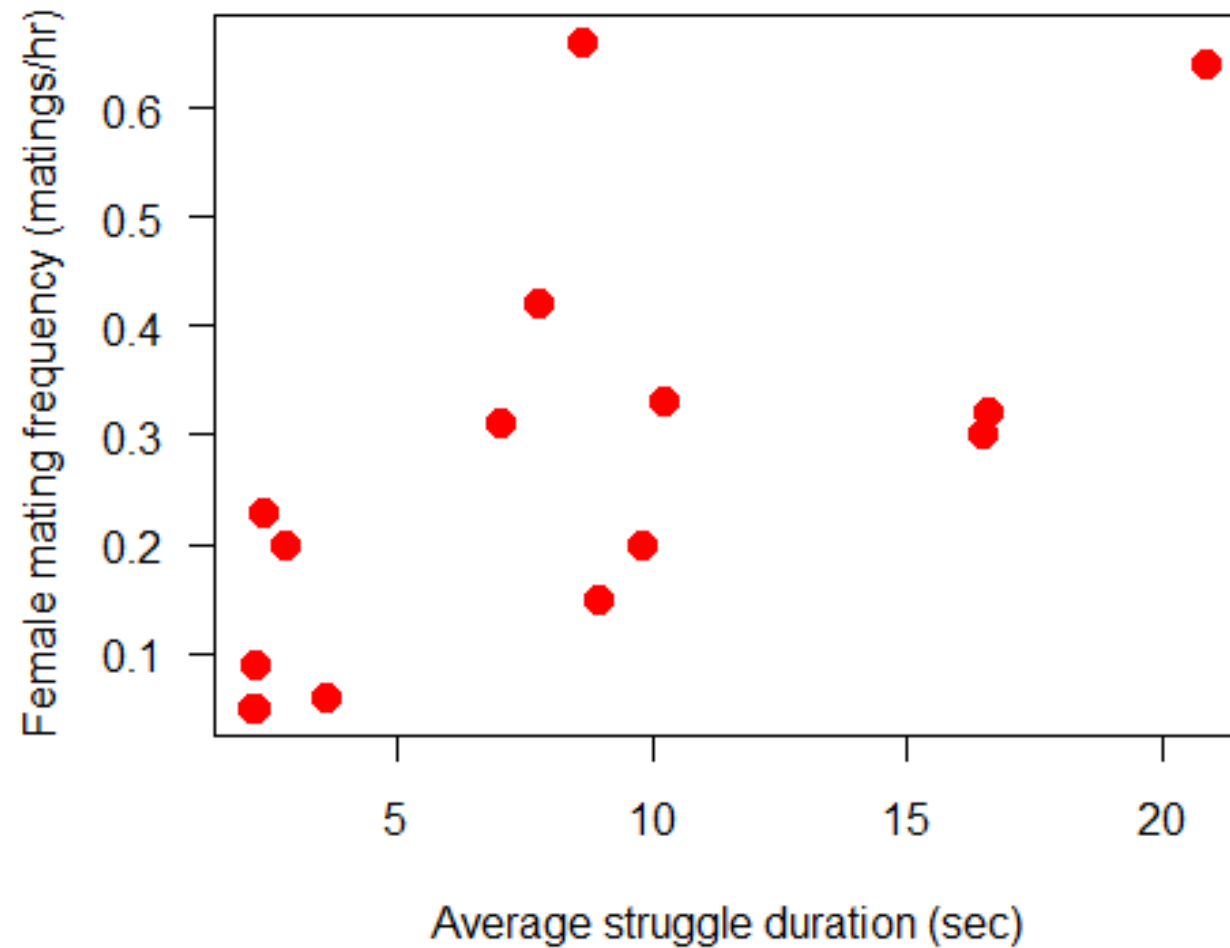
How prevalent is phylogenetic signal in ecologically relevant traits?

Pagel's λ measures the extent to which closely related species are similar in their trait values (phylogenetic signal). Here is a survey of λ-values from many studies and traits by Freckleton et al (2002):
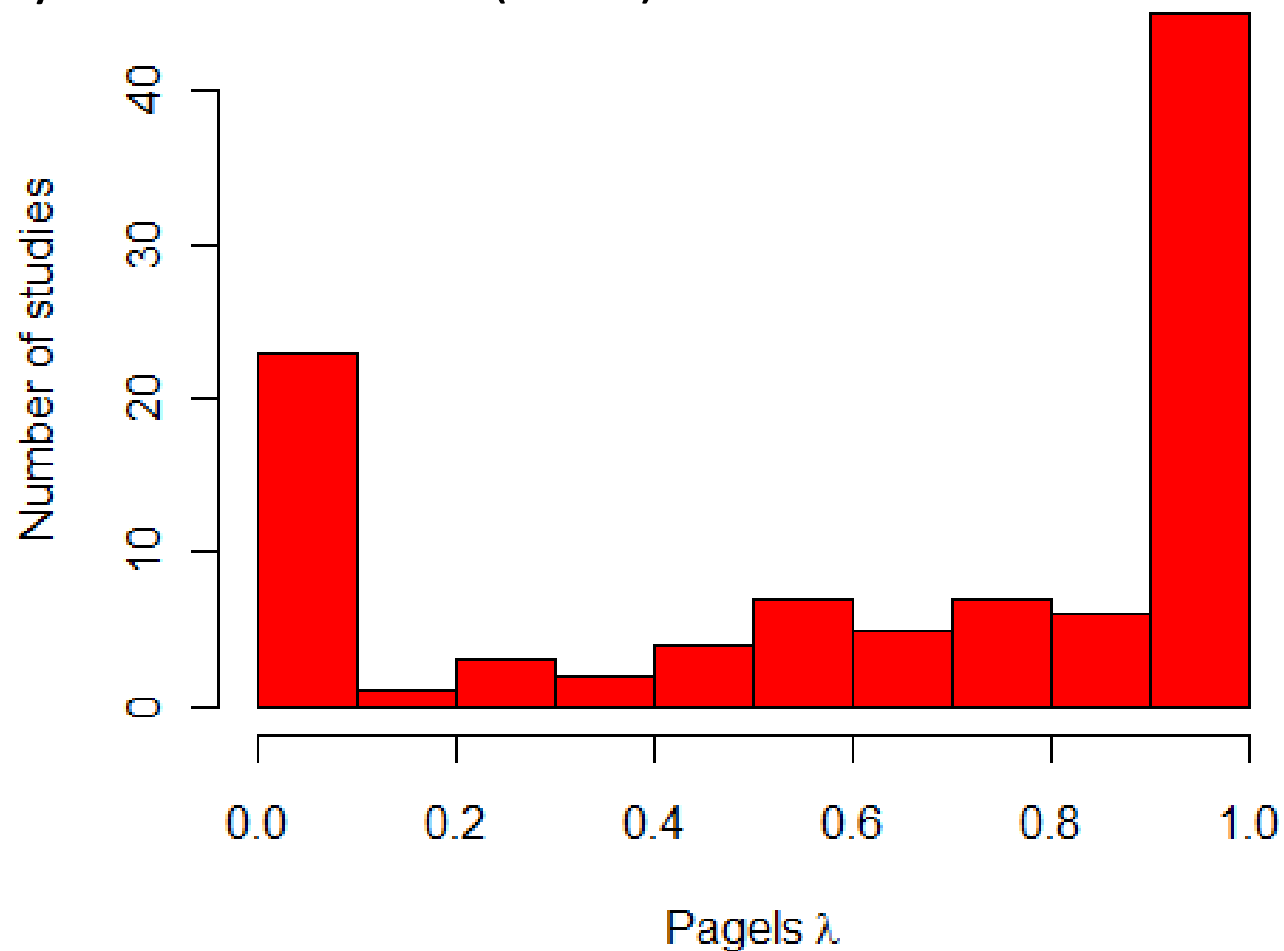
Why is phylogenetic signal a problem?

Non-independence leads to wrong calculations of precision (standard errors, confidence intervals). It leads to wrong Type 1 error rates in null hypothesis significance testing.

Example scenario:
Data on two traits
for 40 species

Looks like a strong
correlation between
variables Y and X

Felsenstein (1985) *Am Nat*
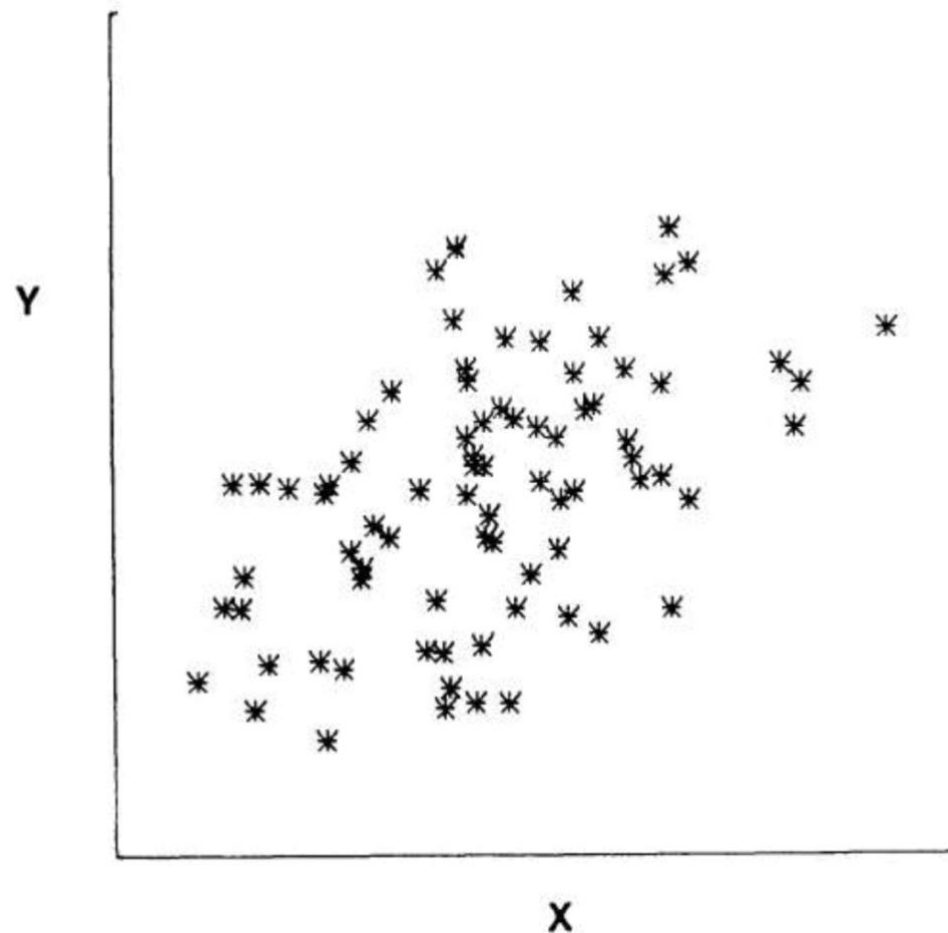
Why is phylogenetic signal a problem?

Felsenstein's "worst case scenario" for the phylogeny of the 40 species.
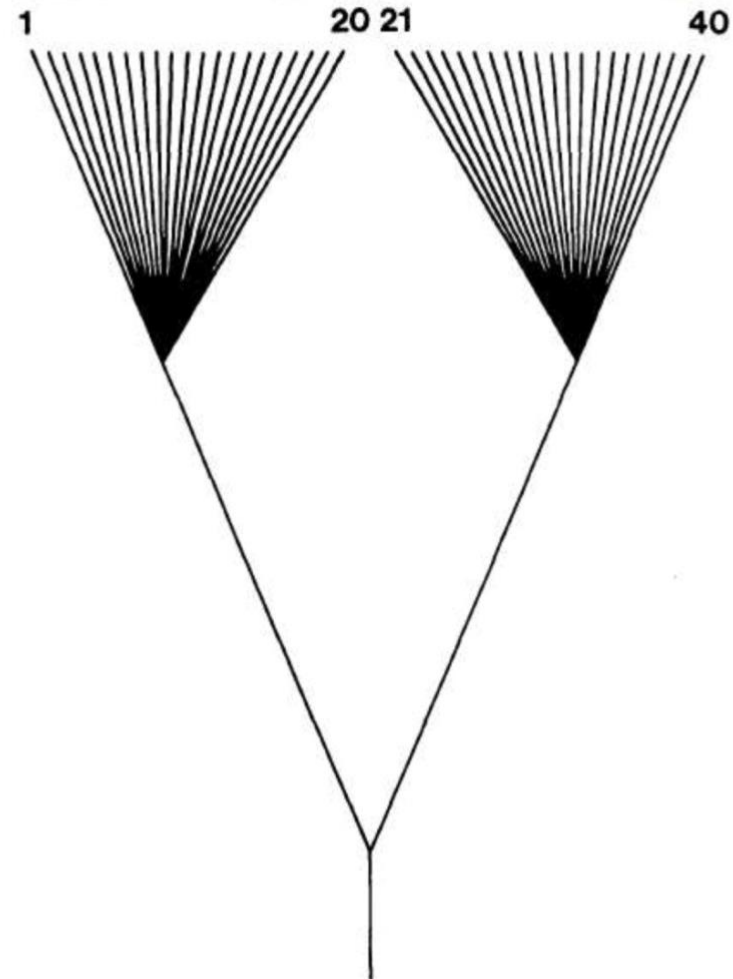
Felsenstein (1985) *Am Nat*

FIG. 5.—A "worst case" phylogeny for 40 species, in which there prove to be 2 groups each of 20 close relatives.

Why is phylogenetic signal a problem?

In this case the non-independence is severe, and creates an apparent association between X and Y where there is none.
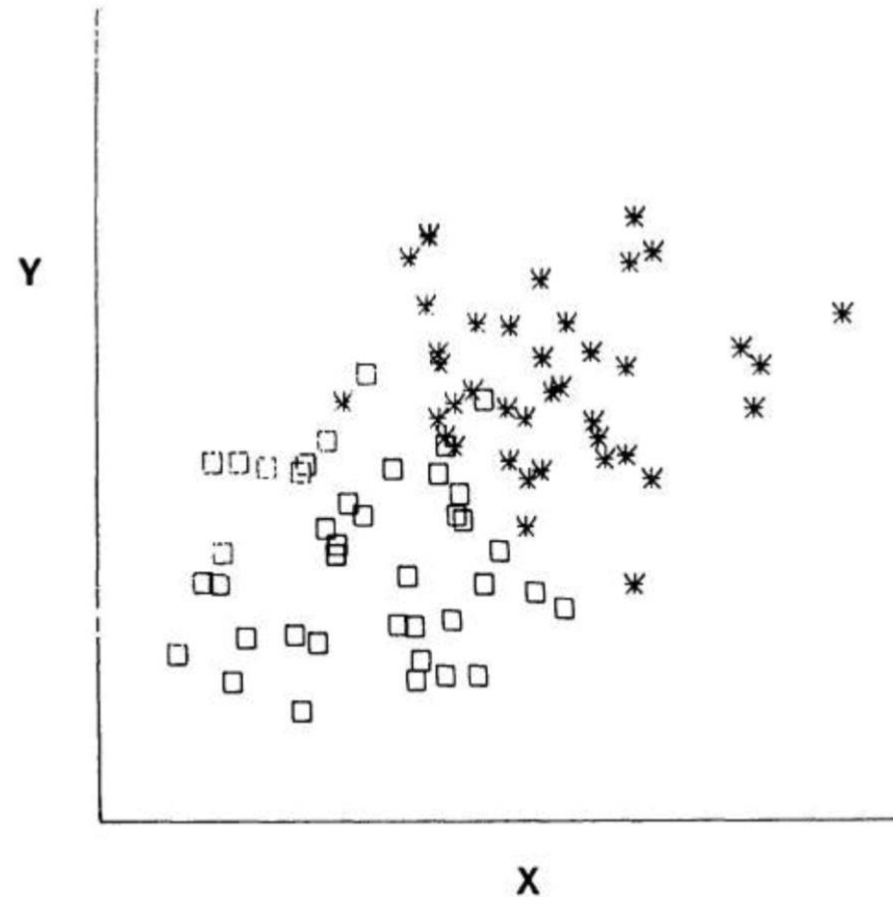
Felsenstein (1985) *Am Nat*

FIG. 7.—The same data set, with the points distinguished to show the members of the 2 monophyletic taxa. It can immediately be seen that the apparently significant relationship of fig. 6 is illusory.

What we are really assuming when we ignore phylogeny

That the species are related as in a "star" phylogeny, which leads to no phylogenetic signal.
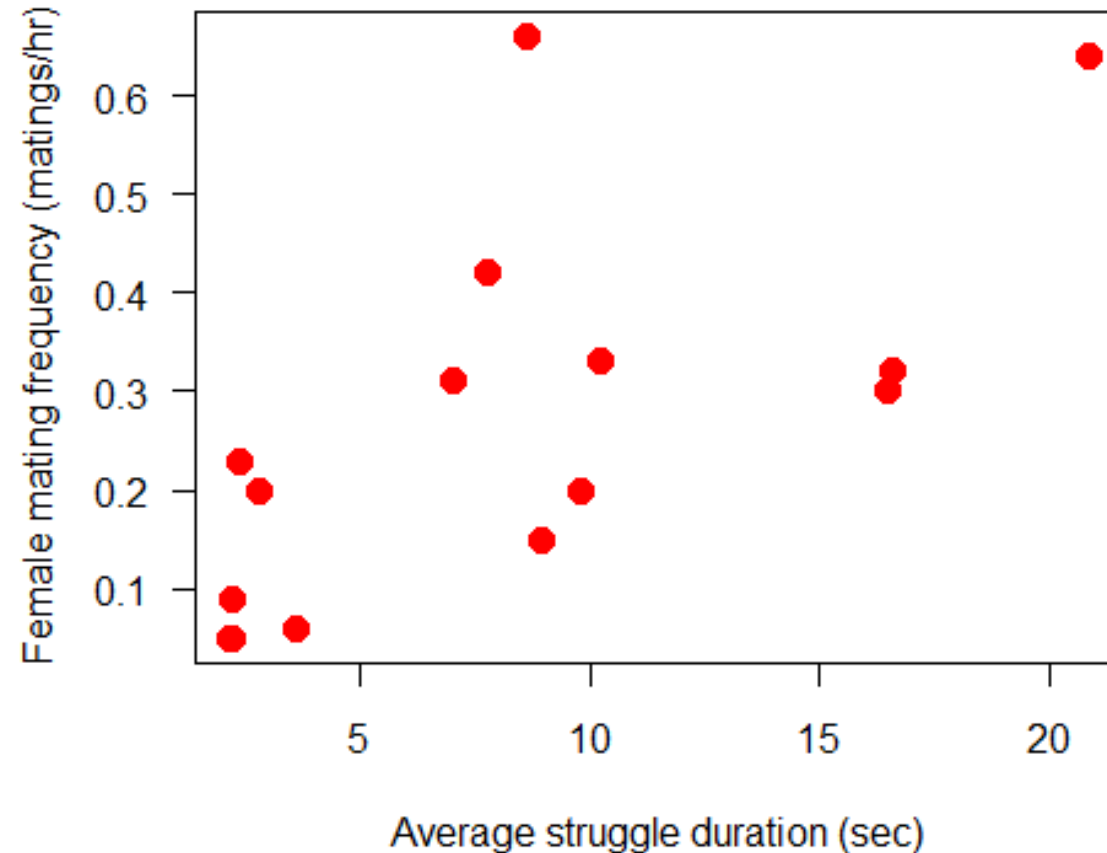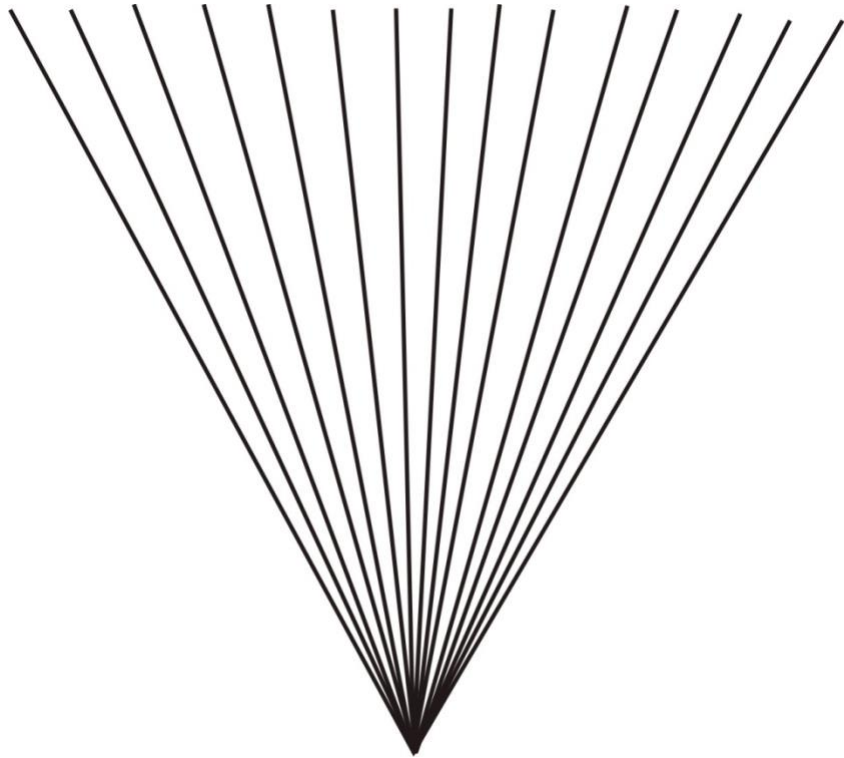
Felsenstein's (1985) solution

Method assumes that the evolution of traits is mimicked by a continuous random walk (Brownian motion).

Under Brownian motion, the difference between any two species in a trait has a normal probability distribution with mean 0 and variance proportional to the time since their common ancestor.

Felsenstein (1985) *Am Nat*
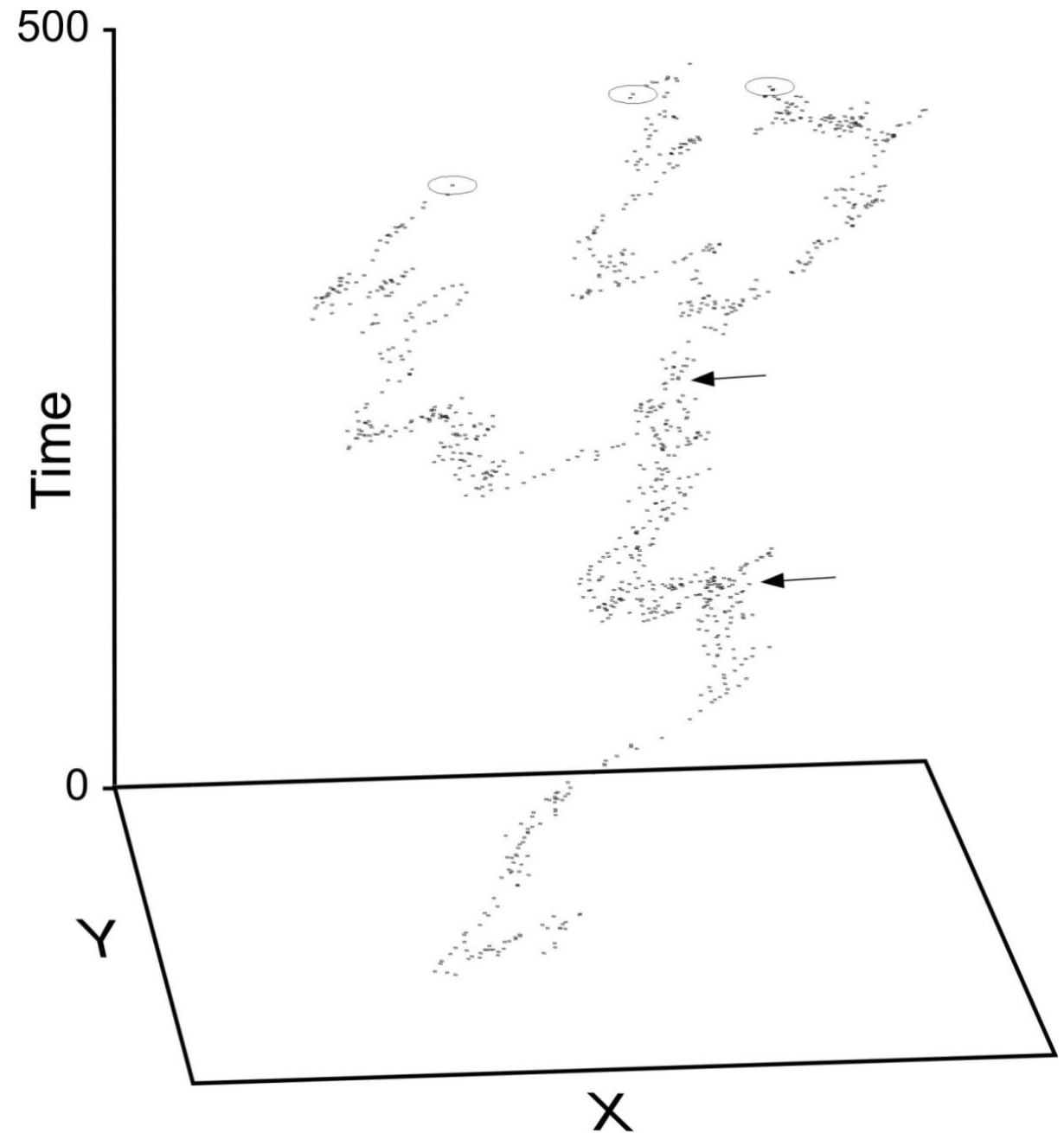
Felsenstein's method of phylogenetically independent contrasts

Under Brownian motion, a, b, and c are not independent, but the <u>difference</u> ("contrast") between a and b is independent of the <u>difference</u> between c and (a+b)/2.
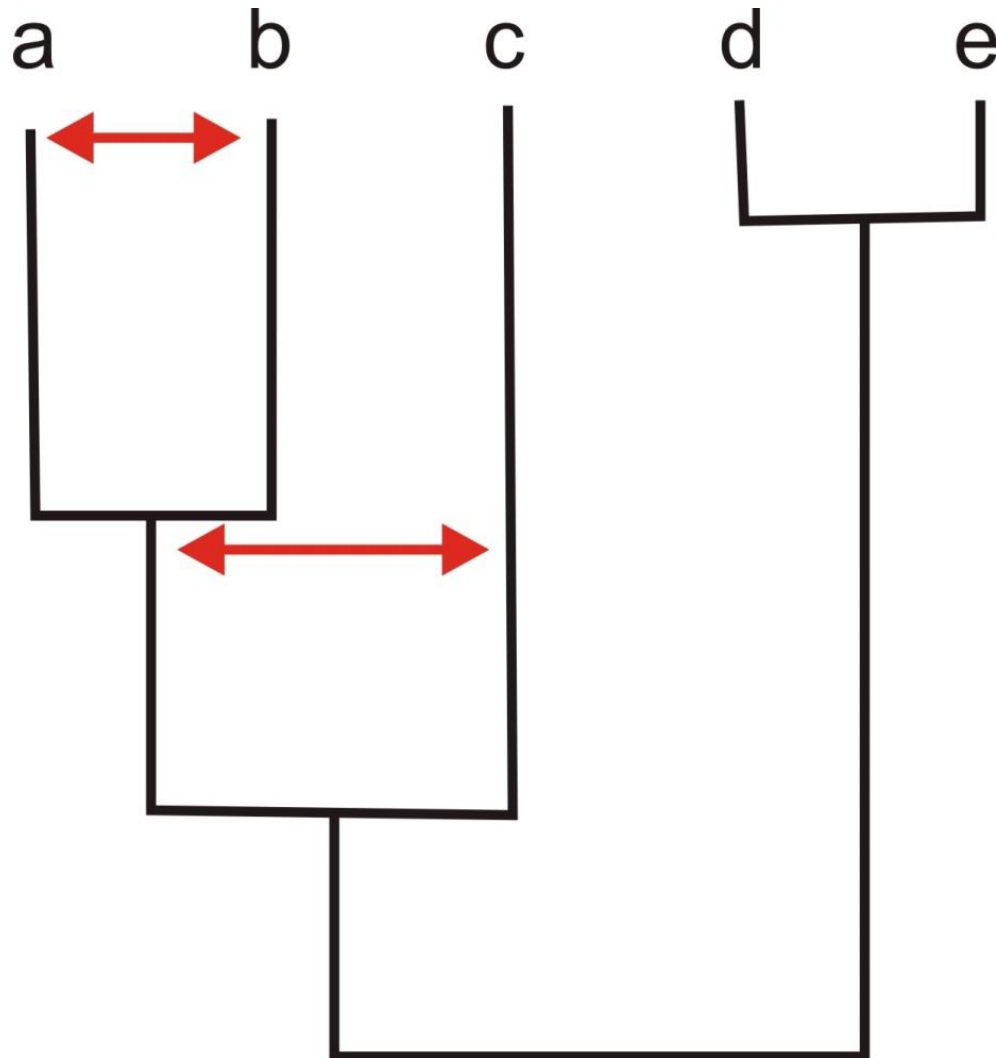
Phylogenetically independent contrasts

There are $n - 1$ independent contrasts for n species.
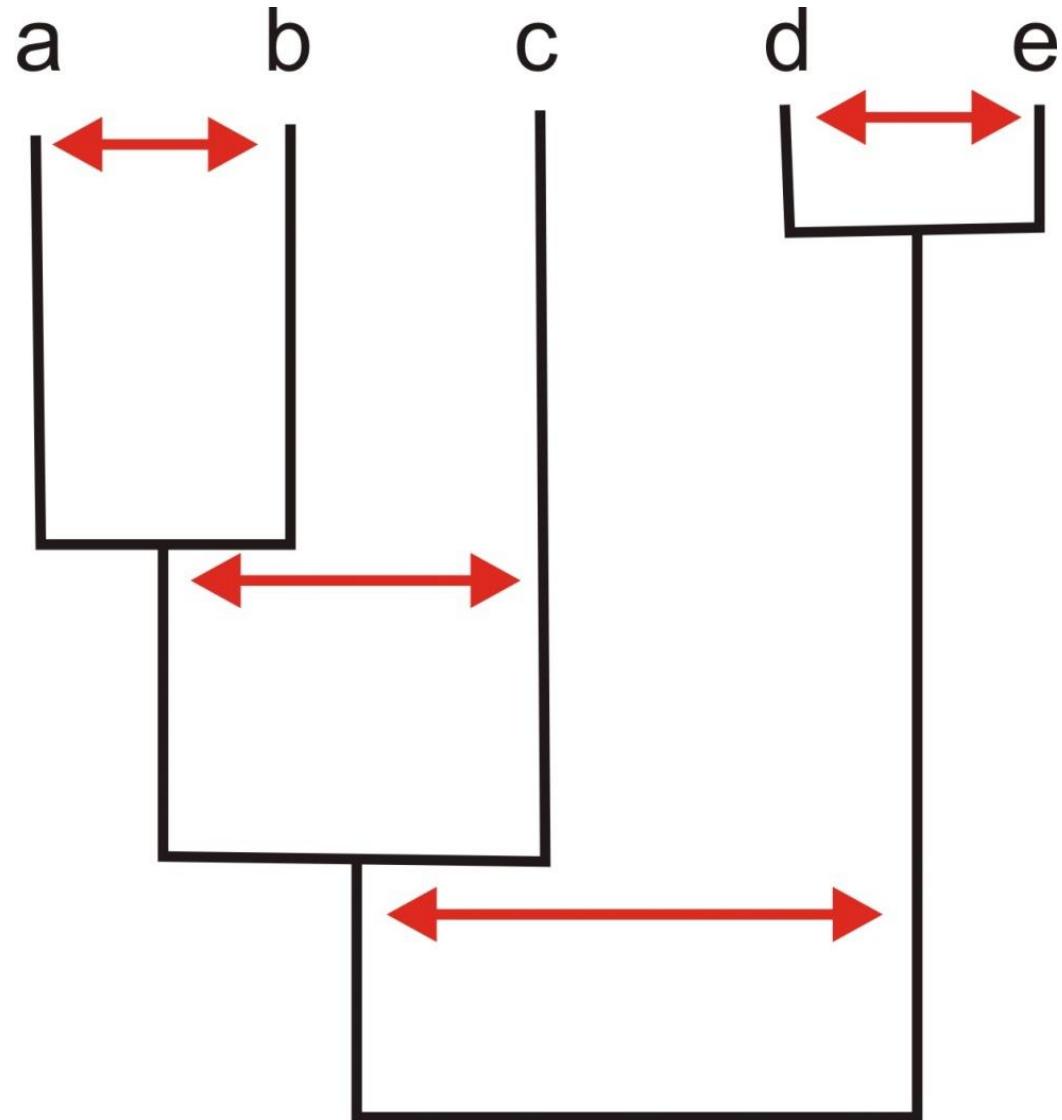
Phylogenetically independent contrasts

Calculation details. Usually, contrasts are standardized by the square root of the expected variance, which is proportional to branch length.
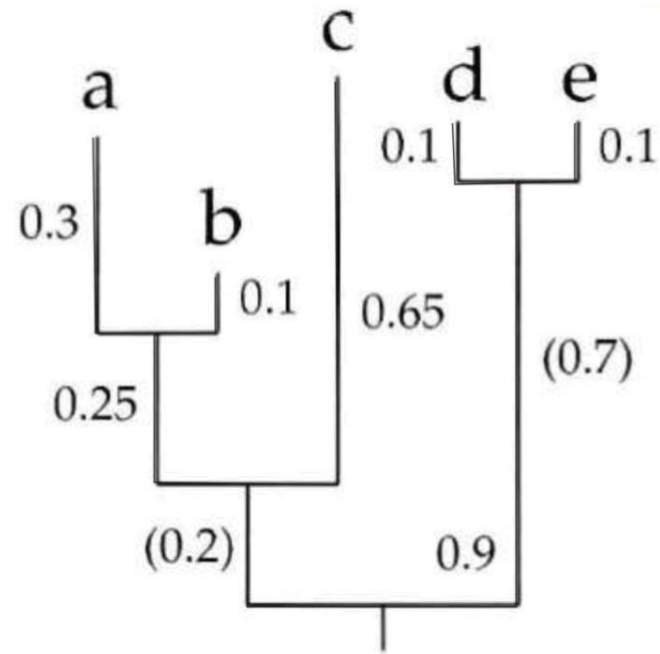
	Contrast						Variance proportional to
y_1	$=$	$x_a - x_b$					0.4
y_2	$=$	$\frac{1}{4} x_a + \frac{3}{4} x_b - x_c$					0.975
y_3	$=$				$x_d - x_e$		0.2
y_4	$=$	$\frac{1}{6} x_a + \frac{1}{2} x_b + \frac{1}{3} x_c - \frac{1}{2} x_d - \frac{1}{2} x_e$					1.11666

Phylogenetically independent contrasts

The idea is to convert the data on <u>both</u> traits to their independent contrasts using the phylogeny of the species. Then calculate the correlation between the independent contrasts of the two traits.
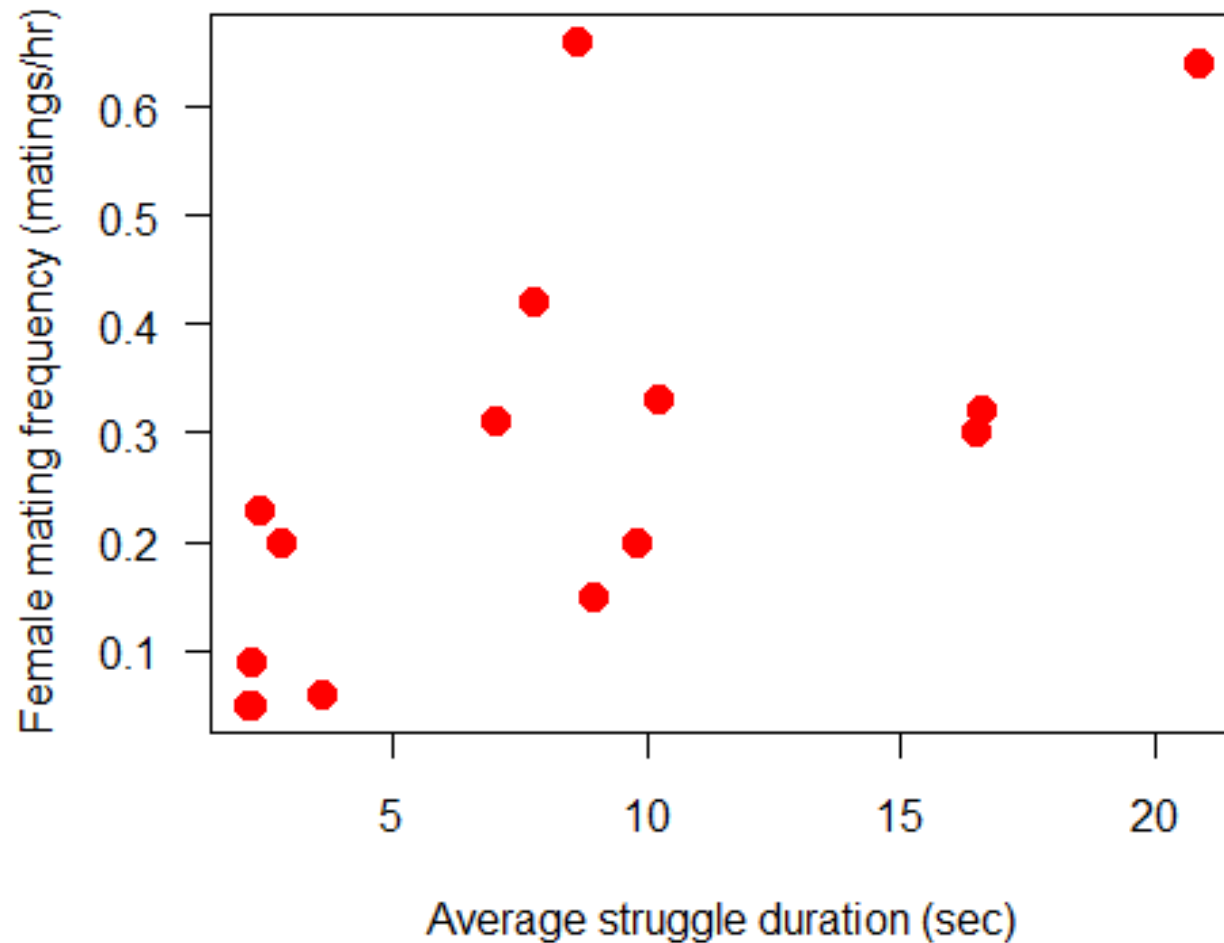
Phylogenetically independent contrasts

A cutaway of the independent contrasts for the water strider mating behavior data. The direction of each contrast is arbitrary, but the contrast direction must be the same for both variables.

	Struggle duration	Mating rate
sphagnetorum	3.63	0.06
incognitus	2.19	0.05
pingreensis	10.24	0.33
gillettei	16.58	0.32
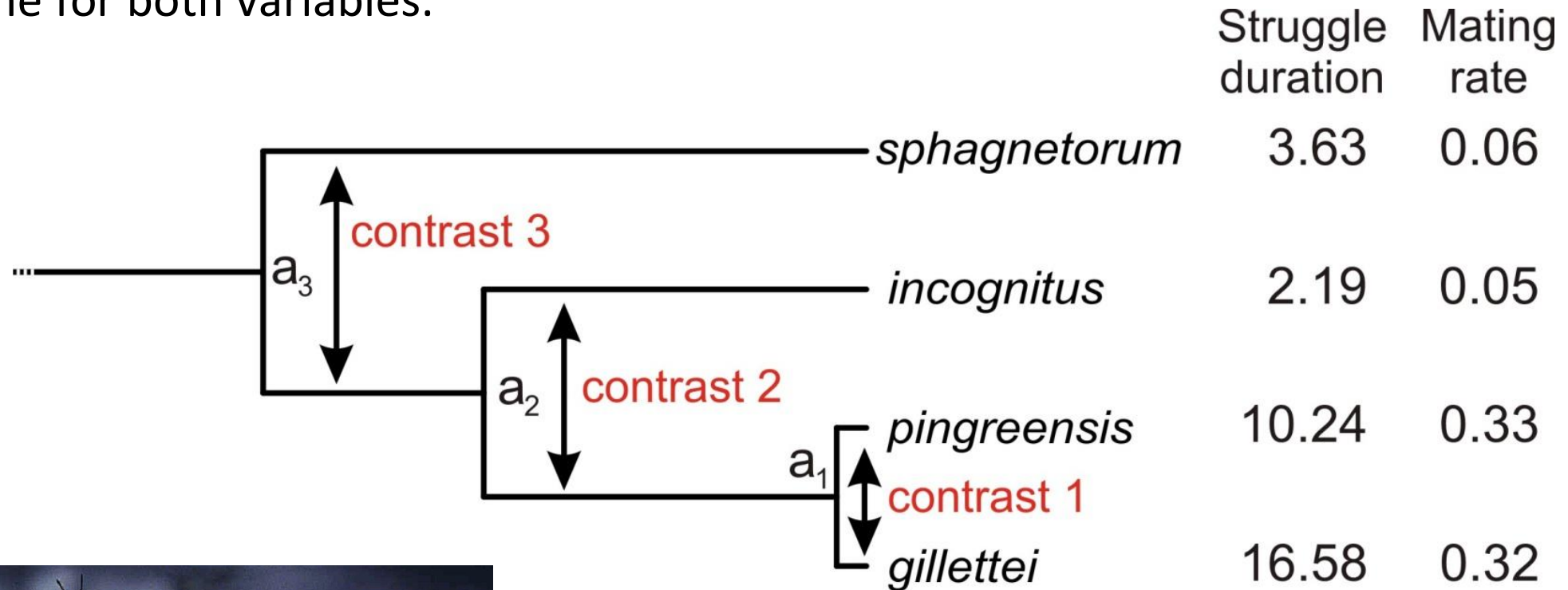
Phylogenetically independent contrasts

Because the direction of the contrast is arbitrary, the correlation or regression using independent contrasts is fitted through the origin (0,0).

The `ape` package in R implements phylogenetically independent contrasts.

Positive correlation confirmed!
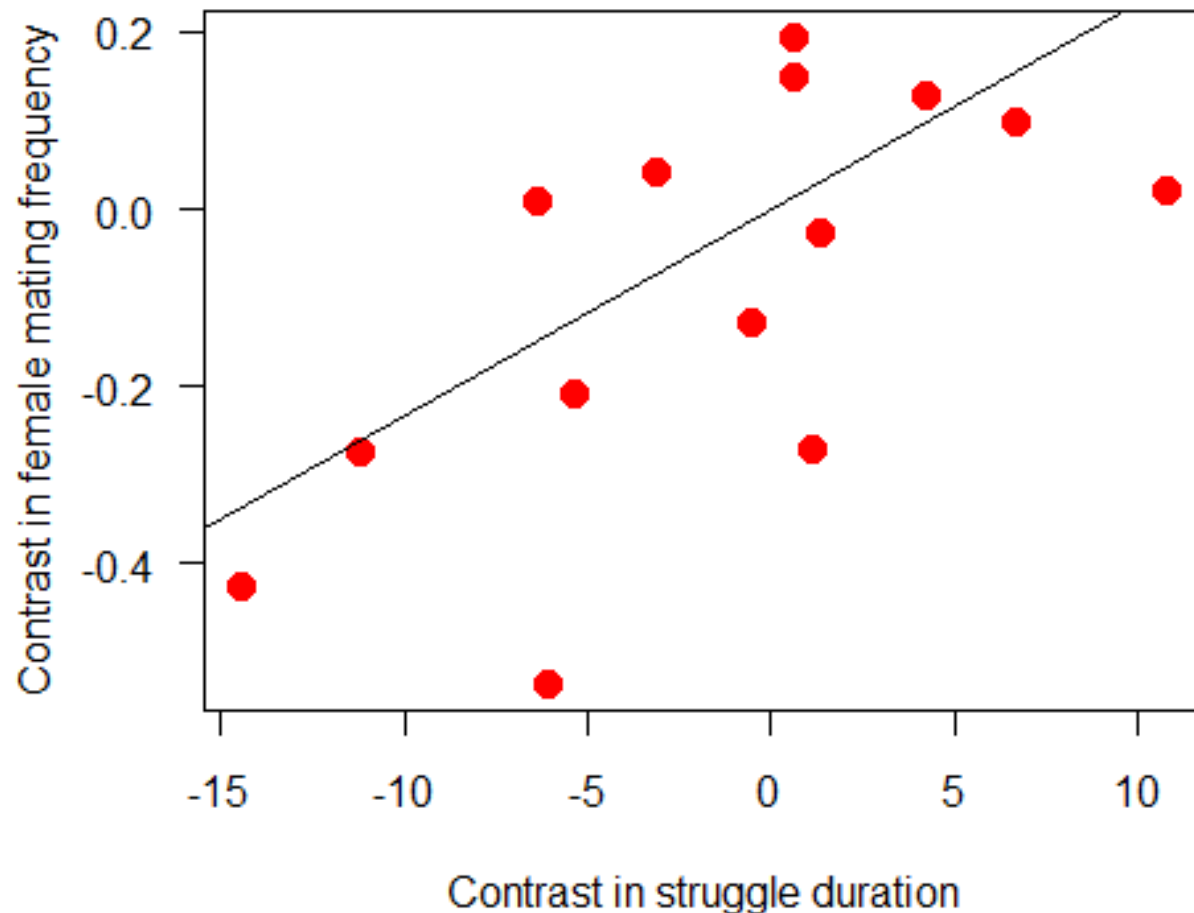
A linear model approach

General least squares (GLS) is a linear model technique mathematically equivalent to phylogenetically independent contrasts.

GLS allows the residuals to be correlated and have unequal variances. The method incorporates them using a "weight" matrix of expected covariances between species traits.

Using GLS gives access to all the tools of linear models, including model selection methods (AIC, etc).

The function `gls()` in the `nlme` package can be used to fit phylogenetic linear models.

Specifying the covariance matrix between data points

	Homo	Pongo	Macaca	Ateles	Galago
Homo	1.00	0.79	0.51	0.38	0
Pongo	0.79	1.00	0.51	0.38	0
Macaca	0.51	0.51	1.00	0.38	0
Ateles	0.38	0.38	0.38	1.00	0
Galago	0.00	0.00	0.00	0.00	1

To analyze, we must know what the variances and correlations are between species. Under Brownian motion, the expected covariance between two species is the proportion of total history, from root to tip, that they share.

Specifying the covariance matrix between data points

	Homo	Pongo	Macaca	Ateles	Galago
Homo	1.00	0.79	0.51	0.38	0
Pongo	0.79	1.00	0.51	0.38	0
Macaca	0.51	0.51	1.00	0.38	0
Ateles	0.38	0.38	0.38	1.00	0
Galago	0.00	0.00	0.00	0.00	1

These expected covariances between pairs of data points (species) are used as "weights" in the linear model fitting. A pair of data points (species) that share most of their phylogenetic history end up being down-weighted in the analysis. In effect, each of them is counted as only a fraction of a data point.
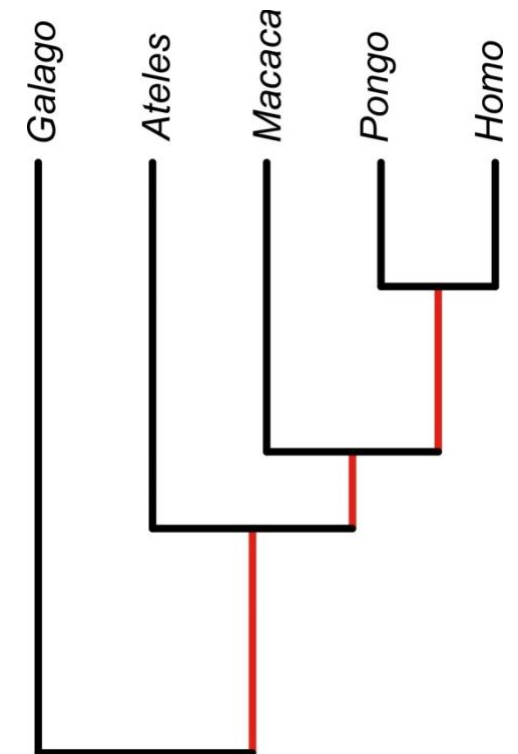
Assumptions of the method

- Evolution in each trait mimics a continuous random walk in time (Brownian motion).

- The rate of evolution is constant through time and along all branches of the phylogeny.

- Speciation and extinction are unrelated to trait values.

These assumptions are difficult to verify.

Branch lengths of phylogenies can be transformed to improve agreement with Brownian motion assumption.

If the assumptions are not met, then in extreme cases using independent contrasts might be worse than simply treating the species data as though they were independent (Harvey and Rambaut 2000).

Assumptions of the method
Diagnostic plots can help

Phylogenetic trees

Read tree from file

Plot tree

Write trees to file

Tree formats

Trait data

Species names

Row names

Row order

Plot continuous traits on trees

Phylogenetic signal

Independent contrasts

PICs

Correlation

Zero branch lengths

General least squares method

Phylogenetic correlation matrix

Using GLS

Diagnostic plots for GLS

Alternative evolutionary models

The Ornstein–Uhlenbeck process

gls.fit() function

Diagnostic plots for GLS

The GLS method essentially transforms the variables in your linear model to a new scale where all the usual assumptions of linear models – independent residuals having equal variance — are met (assuming that your model of evolution is the correct one). GLS then fits an ordinary linear model to these transformed variables.

You can evaluate linear model assumptions by making scatter plots and residual plots of these transformed variables from a GLS analysis using the `lm.gls()` function at the end of this page. Cut and paste the code for the function into your R command console. You'll need to load the `visreg` package too.

I illustrate using data from Rolland et al (2020) "Vulnerability to fishing and life history traits correlate with the load of deleterious mutations in teleosts", Molecular Biology and Evolution 37: 2192–2196. The linear model will fit an estimate of deleterious mutation accumulation in fish species to a measure of fish species vulnerability in the face of human exploitation.

```
# read the tree
fishtree <- read.tree(url("https://www.zoology.ubc.ca/~schluter/R/csv/fishtree.tre"))
fishtree
```

```
##
## Phylogenetic tree with 65 tips and 64 internal nodes.
##
## Tip labels:
##   Astyanax_mexicanus, Danio_rerio, Gasterosteus_aculeatus, Myoxocephalus_scorpius, Sebastes_
norvegicus, Chaenocephalus_aceratus, ...
##
## Rooted; includes branch lengths.
```

```
# read the data
fishdat <- read.csv(url("https://www.zoology.ubc.ca/~schluter/R/csv/fishdat.csv"),
                    row.names = 1)
head(fishdat)
```

```
##                      dNdS ss.Vulnerability
## Anabas testudineus   0.08792156         12.47
```

Assumptions of the method
Diagnostic plots can help

```
## 1 0.8287927   10.15330   0.077005090   0.05926235   0.017742739
## 2 0.8287927   11.08866   0.055717328   0.05991080  −0.004193474
## 3 0.1766355    2.66602   0.064105281   0.01297831   0.051126967
## 4 0.1766355  −39.79760  −0.046142947  −0.01646017  −0.029682779
## 5 0.1785635   36.81465   0.053106918   0.03677380   0.016333115
## 6 0.1785635  −11.79493  −0.008038477   0.00307455  −0.011113028
```

Use `visreg` to show a scatter plot of the transformed variables.

```
visreg(z$lm.fit, "x", ylab = "Corrected dN/dS", xlab = "Corrected Vulnerability")
```

Phylogenetic trees
Read tree from file
Plot tree
Write trees to file
Tree formats

Trait data
Species names
Row names
Row order
Plot continuous traits on trees
Phylogenetic signal

Independent contrasts
PICs
Correlation
Zero branch lengths

General least squares method
Phylogenetic correlation matrix
Using GLS
Diagnostic plots for GLS
Alternative evolutionary models
The Ornstein–Uhlenbeck process
gls.fit() function

A residual plot is obtained by comparing `resid` with `yhat` .

Assumptions of the method

The GLS linear model approach makes it easy to transform branch lengths of the tree to better meet the assumption of Brownian motion.

Under Brownian motion, Pagel's phylogenetic signal λ = 1.

If phylogenetic signal λ is less than one, each of the non-diagonal elements of the phylogenetic matrix can be multiplied by the estimated λ. This allows us to fit a model in which phylogenetic signal in the data is weaker than expected under simple Brownian motion.

The `ape` package in R can finds the "best" estimate of λ for a given data set using maximum likelihood. We'll try this in the workshop.
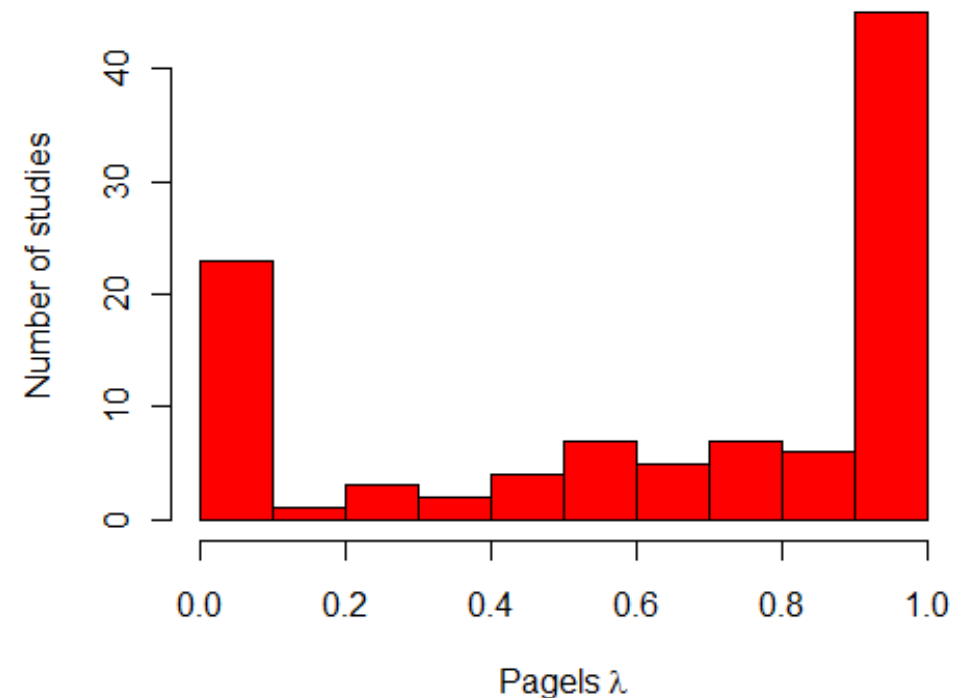
Categorical species data

Patterson and Givnish (2002) found that lily species flowering in the low light environment of the forest understory, such as the blue bead lily (*Clintonia borealis*), tend to have small and inconspicuous flowers whitish or greenish in color.

Lilies that live in sunny, open habitats, or that live in deciduous woods but flower before the tree leaves come out, such as the Turk's-cap lily (*Lilium superbum*), tend to have large, showy flowers.

Categorical species data

Data from 17 lily species indicated an almost perfect association between habitat and flower type. All ten species flowering in open habitats had large and showy flowers. Six of the seven species flowering in shaded habitats had relatively small and inconspicuous flowers. This seemed like a strong association.

	Open habitat	Shaded habitat
Showy flowers	10	0
Inconspicuous flowers	1	6

Categorical species data

But the phylogeny of the group reveals the same problem as in the water strider example: closely related species tend to be similar.

Even though there are 17 species, there might have been as few as three transitions between habitats in the past, leaving fewer effective data points than first assumed.
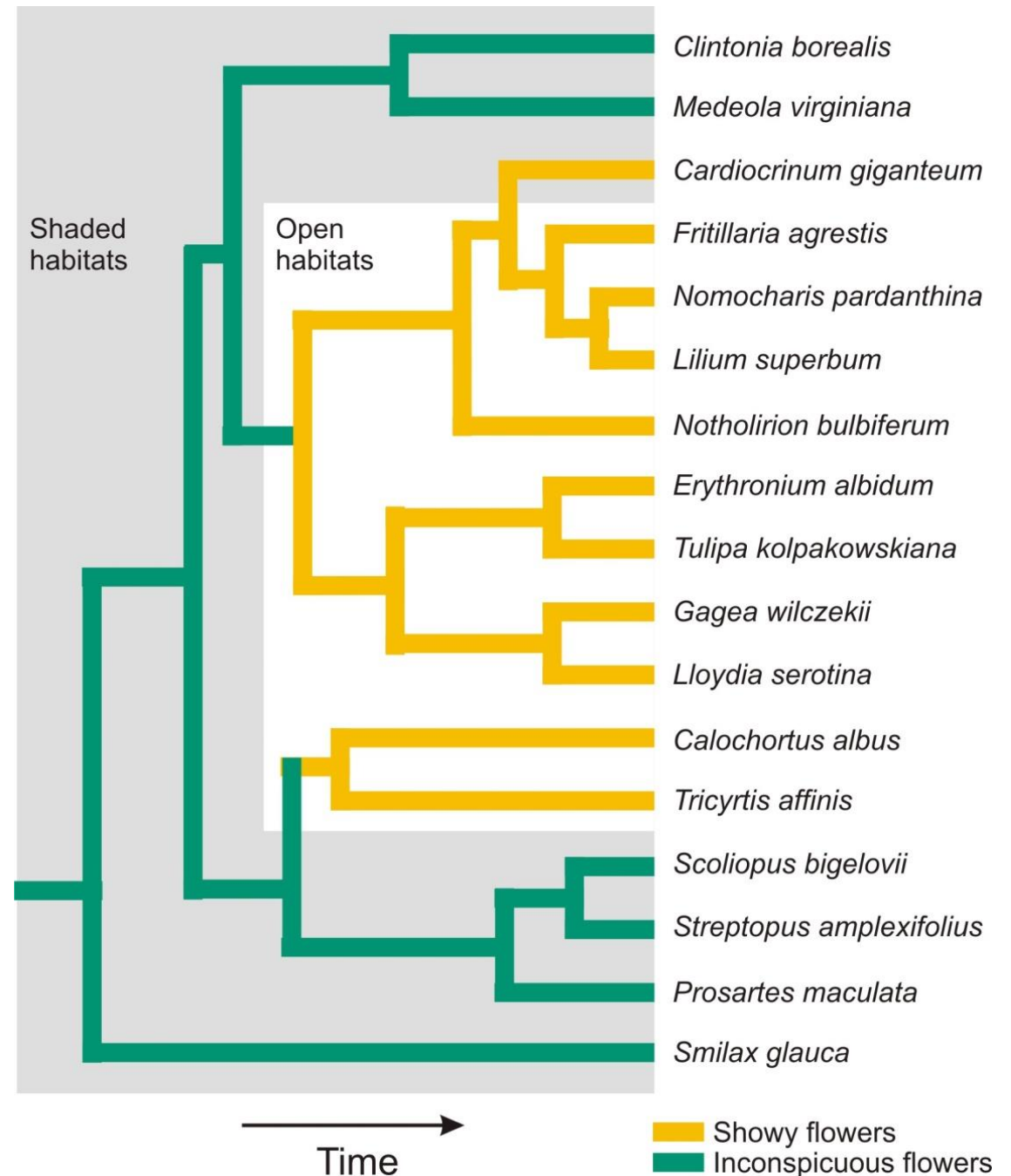
Categorical species data

Pagel (1994) developed a maximum likelihood method for analyzing discrete characters. The method assumes that evolution in each trait mimics a discrete random walk in time (Markov process).

It estimates the transition rates q between states through time on a phylogeny.

It uses likelihood to estimate and test how transitions between states in one trait (e.g., flower conspicuousness) depend on the character states of a second trait (e.g., habitat).

The method is implemented in the corHMM package in R.

0   0   0   0   1

q01=.08
q10=.32

time
0
1
2
3
4
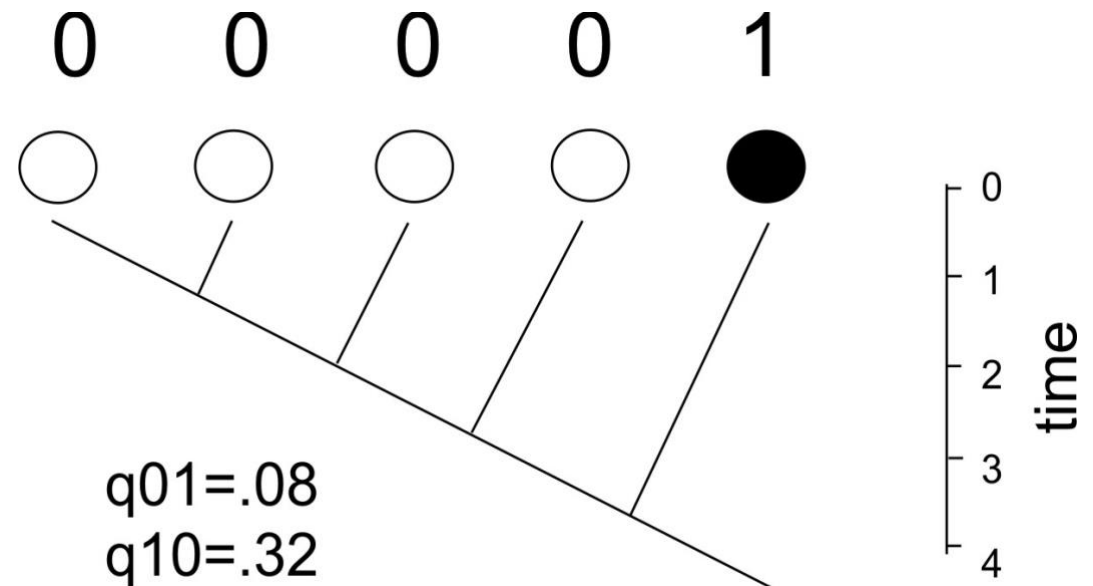
Categorical species data

Maddison, W. and R. Fitzjohn. 2015. *The unsolved challenge to phylogenetic correlation tests for categorical characters*. Syst. Biol. 64:127–136.

"… Pagel's test is susceptible to yielding significant results from the effects of a single change in one of the characters, …. Other tests suffer the same problem, which we will call "within-clade pseudoreplication".

Possible solution:

Beaulieu, J. M., & B. C. O'Meara. 2014. *Hidden Markov models for studying the evolution of binary morphological characters*. In Modern phylogenetic comparative methods and their application in evolutionary biology (pp. 395-408). Springer, Berlin, Heidelberg.

Is phylogenetically independent contrasts/GLS also susceptible?

Uyeda, J. C., R. Zenil-Ferguson, and M. W. Pennell. 2018. *Rethinking phylogenetic comparative methods*. Syst. Biol 67: 1091-1109.

"...phylogenetically independent contrasts can be misled by a single extraordinary event..."

Method development continues apace.

Phylogenetic methods have many applications

Article

Revealing uncertainty in the status of biodiversity change

T. F. Johnson[1,✉], A. P. Beckerman[1], D. Z. Childs[1], T. J. Webb[1], K. L. Evans[1], C. A. Griffiths[1,10], P. Capdevila[2,3,4], C. F. Clements[2], M. Besson[2,11], R. D. Gregory[5,6], G. H. Thomas[1], E. Delmas[1,7,8] & R. P. Freckleton[1,9]

Biodiversity faces unprecedented threats from rapid global change[1]. Signals of biodiversity change come from time-series abundance datasets for thousands of species over large geographic and temporal scales. Analyses of these biodiversity datasets have pointed to varied trends in abundance, including increases and decreases. However, these analyses have not fully accounted for spatial, temporal and phylogenetic structures in the data. Here, using a new statistical framework, we show across ten high-profile biodiversity datasets[2-11] that increases and decreases under existing approaches vanish once spatial, temporal and phylogenetic structures are accounted for. This is a consequence of existing approaches severely underestimating trend uncertainty and sometimes misestimating the trend direction. Under our revised average abundance trends that appropriately recognize uncertainty, we failed to observe a single increasing or decreasing trend at 95% credible intervals in our ten datasets. This emphasizes how little is known about biodiversity change across vast spatial and taxonomic scales. Despite this uncertainty at vast scales, we reveal improved local-scale prediction accuracy by accounting for spatial, temporal and phylogenetic structures. Improved prediction offers hope of estimating biodiversity change at policy-relevant scales, guiding adaptive conservation responses.

Accelerating rates of species extinction are driving global changes in biodiversity, threatening ecosystems and the services they provide[1]. In an attempt to reverse biodiversity declines, world leaders, policymakers and academics have called for action[12]. Evidence-based actions require long-term datasets and rigorous modelling to reliably detect and attribute biodiversity change through time[13,14]. At present, some of the most influential estimates of biodiversity change are calculated using datasets such as BioTIME[2], the Living Planet[15] or the North American Breeding Bird Survey[3]. Inferences from these abundance datasets have shaped policy[16] and are considered by some to be a key pillar of global biodiversity monitoring[17].

Biodiversity datasets are complex and typically subject to one or more sources of non-independence across the axes of time, space and evolution. This presents a challenge for analysis, as omission of even one of these sources of non-independence from a statistical model can lead to underestimation of uncertainty, incorrect trends and poorly resolved prediction, and ultimately undermines current interpretation of wildlife abundance trends[18-20]. A unifying feature of previous studies is that they are characterized by the consistent omission of one or more of these dependencies from their analysis. This imposes a risk that past estimates of abundance change—pointing to declines[15,21], no net change[18,22,23] and recovery[24]—may be unreliable.

Non-independence can be classified in a variety of ways, which we split into two core types: hierarchical, for which observations are pseudoreplicated or nested (for example, multiple trends for a given species, site or region in time); and correlative, for which observations become increasingly correlated (sometimes termed autocorrelation) when close in time[25], space[26] or phylogeny[27]. Under correlative non-independence, we may expect sequential abundance values in a time series to be more similar, and trends should be similar when near in space or in closely related species (Fig. 1). Although studies commonly account for hierarchical non-independence using features such as random effects in mixed models, a literature review covering hundreds of papers published in high-impact journals since 2010 revealed that studies rarely account for correlative non-independence across space (accounted for in 7% of studies), phylogeny (14%) or time (32%; Supplementary Table 1). Further, no biodiversity model has yet been formalized to account for all three sources of correlative non-independence at the same time.

Here we show that ignoring non-independence has serious consequences for inference of biodiversity trends. We introduce the correlated effect model, which incorporates hierarchical non-independence and all three sources of correlative non-independence, and apply it to ten high-profile, multi-species datasets that have been used to infer

[1]School of Biosciences, Ecology and Evolutionary Biology, University of Sheffield, Sheffield, UK. [2]School of Biological Sciences, Biosciences, University of Bristol, Bristol, UK. [3]Departament de Biologia Evolutiva, Ecologia I Ciències Ambientals, Universitat de Barcelona (UB), Barcelona, Spain. [4]Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona (UB), Barcelona, Spain. [5]RSPB Centre for Conservation Science, The Lodge, Sandy, UK. [6]Centre for Biodiversity & Environment Research, Department of Genetics, Evolution and Environment, University College London, London, UK. [7]Habitat, Montreal, Quebec, Canada. [8]Institut des Sciences de la Forêt Tempérée, Université du Québec en Outaouais, Ripon, Quebec, Canada. [9]Debrecen Biodiversity Centre, University of Debrecen, Debrecen, Hungary. [10]Present address: Swedish University of Agricultural Sciences, Department of Aquatic Resources, Institute of Marine Research, Lysekil, Sweden. [11]Present address: Sorbonne Université, CNRS, Biologie Intégrative des Organismes Marins, BIOM, Banyuls-sur-Mer, France. ✉e-mail: t.f.johnson@sheffield.ac.uk

Phylogenetic methods have many applications

Conserving avian evolutionary history can effectively safeguard future benefits for people

Rikki Gumbs[1,2,3]*, Claudia L. Gray[1,3], Michael Hoffmann[1], Rafael Molina-Venegas[4], Nisha R. Owen[3,5], Laura J. Pollock[3,6]

Phylogenetic diversity (PD)—the evolutionary history of a set of species—is conceptually linked to the maintenance of yet-to-be-discovered benefits from biodiversity or "option value." We used global phylogenetic and utilization data for birds to test the PD option value link, under the assumption that the performance of sets of PD-maximizing species at capturing known benefits is analogous to selecting the same species at a point in human history before these benefits were realized. PD performed better than random at capturing utilized bird species across 60% of tests, with performance linked to the phylogenetic dispersion and prevalence of each utilization category. Prioritizing threatened species for conservation by the PD they encapsulate performs comparably to prioritizing by their functional distinctiveness. However, species selected by each metric show low overlap, indicating that we should conserve both components of biodiversity to effectively conserve a variety of uses. Our findings provide empirical support for the link between evolutionary history and benefits for future generations.

INTRODUCTION

Biodiversity contributes a wide variety of benefits and services to humanity including food, fuel, medicine, materials, and a myriad other economic and cultural values (*1*, *2*). Unfortunately, humanity's reliance on biodiversity is now a major driver of the unprecedented declines across species and ecosystems globally (*3*, *4*). Accordingly, the goal of maintaining the benefits contributed by biodiversity for current and future generations, through conservation and sustainable use, now sits at the heart of global biodiversity policy (*5*), including as part of the recently adopted Kunming-Montreal Global Biodiversity Framework under the Convention on Biological Diversity (*6*).

There are many ways to value biodiversity and nature in general (*7*), the most prominent of which is through ecosystem services (*8*).

phylogenetic branches that connect them (*12*)—has been proposed to fulfill this role, under the assumption that maintaining a greater amount of PD will conserve distinct features and consequently a wider variety of potential benefits (*5*).

Although it is not possible to predict the precise nature of future benefits it is reasonable to assert that known benefits today were, at some point in the human history, unknown future options for humanity. For example, most biodiversity benefits today could be seen as option value for the future generations of the first humans that appeared in Africa roughly 200,000 years ago. Thus, work has been done to assess the performance of PD at capturing known benefits from plants when applied naively (i.e., selecting species for conservation based on PD with no knowledge of the distribution of benefits). Forest *et al.* (*13*) found that selecting sets of plant genera to

Workshop on phylogenetic comparative methods

This Thursday!

R: an embarrassment of riches

cran.r-project.org/web/views/Phylogenetics.html

Use R!

This course was an introduction to more advanced methods in data analysis in ecology and evolution, how they work, and how you can avoid *some* of the most common misinterpretations and perils.

These methods will likely be useful to your future work. Hopefully you have a basis to go further as needed.

The R tips web site and the workshops will remain online and available for the foreseeable future. I'll do my best to keep it up to date. Revisit and refresh your memories as needed.

Lots of people use R for data analysis here, so there is help all around. Start a data analysis group!

Bye!