This is a repository copy of *Comparative analysis of experimental data*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/145191/

Version: Accepted Version

Article:
Freckleton, R.P. orcid.org/0000-0002-8338-864X and Rees, M. orcid.org/0000-0001-8513-9906 (2019) Comparative analysis of experimental data. Methods in Ecology and Evolution, 10 (8). pp. 1308-1321. ISSN 2041-210X

https://doi.org/10.1111/2041-210x.13164

This is the peer reviewed version of the following article: Freckleton, R. P. and Rees, M. (2019), Comparative Analysis of Experimental Data. Methods Ecol Evol., which has been published in final form at https://doi.org/10.1111/2041-210X.13164. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Comparative Analysis of Experimental Data

Robert P. Freckleton

Mark Rees

Department of Animal & Plant Sciences

University of Sheffield

Sheffield S10 2TN

December 11, 2015

Correspondence: r.freckleton@sheffield.ac.uk

Summary

We consider the problem of how to analyse data from experiments conducted on multiple species. This seems to have been largely overlooked in the literature, and we highlight that the use of species as experimental units creates issues for both the design and analysis of experiments. We distinguish fully randomized experiments in which all treatments are applied to all species from those experiments in which the factor of interest varies at the species level, i.e. treatments are not randomly allocated to species. In this latter case, the distribution of the experimental factor across species may be random, phylogenetically structured, or species may be chosen in order to phylogenetically balance the sample (e.g. through sister-species comparisons). We show using simulations that the structure of the experimental factor can affect power and Type I error, and that commonly used approaches (Linear Mixed Models and ANOVAs) may have poor statistical properties when both the predictors and response data show strong phylogenetic signal. We highlight that the true phylogenetic generalized least squares model yield has good statistical properties but show that in some cases the true variance structure may be difficult to identify empirically. Moreover many current comparative tools do not allow such analyses to be easily applied, and we highlight some of those that do.

Introduction

The comparative method is amongst the most widely used approaches for addressing questions in ecology and evolutionary biology ((Harvey and Pagel, 1991); Freckleton & Pagel 2010; Nunn 2012). The rationale of the comparative method is that a group of species will contain more variation than a single species, or is possible to create using experimental manipulation (Maynard Smith, 1978). Consequently, comparative methods can be used to test extremely broad hypotheses. Moreover, comparative methods typically use data collated from the literature and are therefore extremely efficient in terms of time, expense of data collection and reuse.

The potential problems with comparative approaches are well known, and result from the statistical and evolutionary non-independence of species (Harvey, 1996, Harvey and Pagel, 1991). Evolutionary non-independence of species results in similarities within assemblages that are the result of common ancestry rather than independent evolution. Because of this comparative data cannot be safely regarded as being statistically independent, and a suite of approaches have been developed to analyse comparative data whilst incorporating the phylogenetic relationships between the species and these are now routinely used (e.g. (Grafen, 1989);(Martins and Hansen, 1996); (Pagel, 1997, Pagel, 1999);(Hadfield and Nakagawa, 2010)).

Some authors have argued that the problem of phylogeny is not as severe as originally claimed (e.g. Westoby et al. 1995; Starck & Ricklefs 1996), although these claims have been largely dismissed (Harvey, 1996, Harvey et al., 1995). The lessons from simulations are that it is possible for phylogenetic non-independence to generate both incorrect inference and loss of power (Martins and Garland, 1991), and that small amounts of phylogenetic non-independence can create increasingly severe problems as datasets become larger (Freckleton et al. 2011). Although in many ways

the state of the art in comparative methods has for the past decade focused more on the use of evolutionary models to uncover the processes generating current trait diversity (e.g. Pagel 1997; Losos 2008; Freckleton et al. 2011), nevertheless probably the majority of users of the comparative method are primarily aiming to correct statistical analyses for non-independence.

Conceptually the opposite of the comparative method is the experimental approach (Maynard Smith, 1978). Comparative methods are typically observational, and rely on uncovering correlations, with possible confounding effects eliminated statistically. A correlative analysis can never eliminate all confounding effects, nor can causation be distinguished from correlation on purely statistical grounds. On the other hand, experiments are intended to manipulate only the factors of interest, with nuisance and confounding effects eliminated by design and randomization (Mead 1988). The randomized block experiment is, for example, described as the 'gold standard' in testing ecological hypotheses (e.g. Newman et al. 1995).

Experimental and comparative approaches need not be completely divorced, however, and the dichotomy between the two becomes blurred in experiments that use multiple species. Comparative experiments have always been common, and classic examples include mass screening experiments on plants (e.g. (Grime et al., 1990)). In principle the control available in experiments should allow the 'species' effect to be accounted for and treatments randomized across species. Consequently it might be expected that experimental comparative approaches should be less prone to confounding than purely correlative studies.

Unfortunately this is not likely to be a general or safe assumption to make. The reason for this is illustrated in Figure 1, which highlights four designs that are commonly used. The first approach (Figure 1A) is a factorial experiment, in which all

species receive each of the treatments. Species are effectively treated as randomized blocking factors. This design would be used when it is possible to manipulate the factor of interest (e.g. when applying different nutrient regimes to plants in growth experiments).

Frequently, however, the factors analysed in an experiment are properties of the species themselves. For instance, in an example we discuss below plant species may be characterized as possessing one of two photosynthetic pathways (C3 vs C4). Each species is in only of one of two states, and consequently this factor cannot be randomly applied to species. In such situations, the design of the experiment will depend on how species are chosen and/or how traits are distributed across the phylogeny. Figure 1B - D illustrate three idealized cases.

1) Variation in the trait across the phylogeny is random (Figure 1B). Consequently, the levels of treatment factor are distributed randomly through the tree, and there is no phylogenetic structure to the character being analysed.
2) Variation in the factor studied is not distributed randomly across the phylogeny: as shown in Figure 1C. It may be that there is phylogenetic signal and groups of related species share the trait of interest, and are similar in other respects.
3) The sample of species used might be chosen to maximize differences between closely related species, for example as in sister-species comparisons (Figure 1D). In this case, the trait analysed has an over-dispersed distribution.

Several approaches have been employed in the literature for the analysis of data from comparative experiments, although the question of how to analyse phylogenetically referenced experimental data does not appear to have been explicitly addressed. The simplest approach would be to ignore species identity completely, and

treat data as phylogenetically independent: this is potentially defensible if the experiment is appropriately randomised. However, a drawback could be that inter-species variation would add variance to the data and reduce power. Consequently including species identity in the analysis would usually seem preferable. The simplest way to do this would be to introduce 'species' as a factor into the analysis, and test as a fixed or random factor. For a fully orthogonal design, the choice of whether to treat as fixed or random is arbitrary (since either way the effect of species is estimated marginally with respect to the other experimental factors).

In the case where there is phylogenetic structure in the factor of interest, the choice of how to proceed within a conventional statistical framework is not obvious. One approach might be to regard 'species' as the level of replication in the analysis and therefore to stratify the analysis appropriately, e.g. by effectively treating species as a 'split-plot factor' in the parlance of classic ANOVA. The efficacy of this approach would be expected to depend on the precise nature of the phylogenetic structure of the factor, however. Another approach is sister-species comparisons (e.g. see (Weir and Lawson, 2015) for a recent method), however comparisons among sister species may ignore higher level (e.g. above genus level) confounding.

Currently the most widely employed approach to analysing comparative data is based on linear models and generalized least squares (e.g. Felsenstein 1973, 1985; Grafen 1989; Pagel 1997, 1999; Martins & Hansen 1996; Freckleton et al. 2002). This approach is a generalization of classic regression and ANOVA methods to account for phylogenetic non-independence by the explicit inclusion of a variance-covariance matrix that represents species' phylogenetic relationships. Conceptually there is little difference between this and the approaches that one might use in the analysis of comparative or experimental data. For example, fitting a block in an experimental

design simultaneously accounts for uncontrolled variation and reduces non-independence between experimental units. This is akin to the function of a phylogenetic correction (Rees 1995).

Despite the conceptual links, there is currently a gulf between approaches used to analyse comparative, experimental data and the approaches used in correlative analyses. This is reflected in the software that is currently available (e.g. (Paradis et al., 2004); geiger (Pennell et al., 2014); caper (Orme, 2013)) in which the response variable is assumed to be a single value for each species, and the ability to account for experimental design is not emphasized. For example, in such packages it is not clear how one might include blocking or more sophisticated designs such as error structures that vary between strata (but this may be possible with packages developed in other fields, e.g. see coxme package (Therneau, 2015)). Methods have been developed to account for intra-specific variability ((Felsenstein, 2008); (Ives et al., 2007)), which is similar to doing this. These analyses have shown that there are potentially impacts of accounting for intra-specific structure (e.g. (Silvestro et al., 2015)), and the same would be expected to be true in experimental data.

In this paper we consider the problem of how to analyse data within a comparative, experimental framework. We use the four experimental schemes outlined in Figure 1 as the basis for exploring the performance of different analytical methods. We highlight that the experimental design and phylogenetic structure of the variables analysed have important consequences for the expected performance of methods. Analysis of simulations and real data show that, depending on the structure of the data, application of the incorrect method can lead to either loss of power or Type I error. The results show that, with the correct variance structure, Generalized Least Squares approaches outperform other methods. However, our results also

highlight that comparative experiments with small numbers (<20) of species have limited power to estimate the variance structure, and that in many such cases the correct analysis may remain in doubt.

Methods

Simulations

The four schemes shown in Figure 1 were used as the basis for our simulations. To mimic the type of analysis commonly employed we used a randomized-block experiment. For simplicity we assumed that the treatment of interest was a binary variable (as outlined in the discussion, it is straightforward to generalize our results to more complex treatments).

Phylogenies were generated according to a birth-death process using the package TreeSim ((Stadler, 2015)) in R (R Core (Team, 2015)). Exploratory analysis indicated that the results below are relatively insensitive to the method used to generate the phylogeny, so we held the birth rate of the phylogeny constant at 0.9 and the death rate constant at 0.3. Phylogenies of between 4 and 200 species were generated, representing a typical range that might be encountered in real data.

The basis for the simulations is a general phylogenetic mixed model in which there is a treatment, a blocking structure and residual variance associated with both the individual measurement (e.g. plant-to-plant or pot-to-pot variation in a growth experiment) and with phylogeny.

The model for the vector of observations **y** was:

$$\mathbf{y} = \mathbf{B}\mathbf{b}_b + \mathbf{X}\mathbf{b}_x + \mathbf{e} \qquad (1)$$

In the fully randomized, block design with a binary treatment and b blocks and n species (Figure 1A), **B** is a b x $2n$ design matrix representing the blocking structure.

We assume that the entries of **B** are ordered such that all members of the same block are ordered sequentially. **X** is a *k* x *b* x *n* design matrix representing the experimental factor with *k* levels, with the treatments ordered within blocks. $\mathbf{b}_b$ and $\mathbf{b}_x$ are coefficients that model the differences between blocks and the effects of the treatments, respectively. Finally **e** is a vector of errors.

The errors **e** were assumed to follow a multivariate normal distribution with a variance-covariance matrix **W**, given by:

$$W = \sigma[\lambda(1-\psi)\mathbf{D} \otimes \mathbf{V} + \psi\mathbf{D} \otimes \mathbf{I}_n + (1-\lambda)(1-\psi)\mathbf{I}_{kbn}] \qquad (2)$$

In equation (2) σ is a scale parameter and $\otimes$ denotes the Kronecker product. **V** is an *n* x *n* variance-covariance matrix given by the phylogeny, and scaled so the leading diagonal entries are all one (all trees were ultrametric). **D** is *kb* x *kb* covariance matrix representing the covariance between species across *k* treatments and *b* blocks: this measures the degree to which the phylogenetic variance structure is the same across different treatments and blocks. We saw no reason to make any assumption about how this would vary, so the entries of **D** were all set to unity, i.e. the phylogenetic effect is the same across the whole experiment. **D** $\mathbf{I}_n$ is a *kbn* x *kbn* matrix that codes for species identity (entries corresponding to pairs of experimental units relating to the same species are 1, those corresponding to different species are 0).

There are three components to the variance in equation (2). The first component is the variance among species means that results from phylogenetic dependence (σ_P^2). The second describes variation in the species means that is independent of phylogeny (σ_S^2). The final variance is that between replicate experimental units independent of phylogeny or species identity, i.e. the error variance (σ_e^2).

In equation (2) the parameters, λ and ψ, perform similar but slightly different roles. λ alters the degree of phylogenetic signal in the mean response for each species and thus alters the proportion of variance between species contributed by **V** relative to non-phylogenetic variation (σ_e^2). On the other hand, ψ generates variance in the species means independent of that generated by the phylogeny. This allows for differences between species, unrelated to phylogeny. If ψ is 1 then the phylogenetic component makes no contribution to overall variance. If ψ is 0 then there is no additional species-level variation, i.e. all observations from a species are identical. If $0 < \psi < 1$ and $0 < \lambda < 1$ then both phylogeny and independent species differences play a role.

In order to understand the differences between these different components, it is useful to highlight that the net covariance **W** contains 3 different types of (co)variances. The first is the expected variance of a species *i* in a given block and treatment:

$$w_{i,i} = \sigma\left[\lambda(1-\psi)v_{i,i} + \psi + (1-\lambda)(1-\psi)\right] \quad (3)$$

The second is the expected covariance of species *i* in a given block and treatment with the value measured from species *i* in another block (*b*) and treatment (*k*):

$$w_{i,kbi} = \sigma\left[\lambda(1-\psi)v_{i,i} + \psi\right] \quad (4)$$

Although both (3) and (4) refer to intra-specific (co)variance, the covariance (4) is less than the expected variance for an observation (3) because of the additional observation error in (3). This does not arise in (4) because (4) refers only to covariance, i.e. expected similarity. Finally there is the expected covariance of species *i* with species *j*:

$$w_{i,j} = \sigma\left[\lambda(1-\psi)v_{i,j}\right] \quad (5)$$

Although alternative parameterisations to that used in equation (2) are possible they are not necessarily simpler. One possibility would be to redefine λ as the proportion of the total variance explained by phylogeny (i.e. $\sigma_P^2/\sigma_{Total}^2$). However, this adds complexity to the error component of equation (2), which simple algebra shows to be proportional to $[(1-\lambda)(1-\psi)-\lambda\psi]$. The subtracted variance $\lambda\psi$ in this formulation relates to any variance attributable to both phylogeny and species differences, which is a difficult quantity to understand intuitively. The formulation in equation (2) ensures that λ and ψ are independent, with the components due to phylogeny and species differences clearly separated.

To summarise, the overall variance decomposition is:

Total variance across all observations:

$$\sigma_{Total}^2 = \sigma_P^2 + \sigma_S^2 + \sigma_e^2 \quad (6)$$

Proportion of phylogenetic variance in individual observations relative to phylogenetic variance and error:

$$\lambda = \frac{\sigma_P^2}{\sigma_P^2+\sigma_e^2} \quad (7)$$

Proportion of total variance in observations explained by differences between species means, independent of phylogeny:

$$\psi = \frac{\sigma_S^2}{\sigma_P^2+\sigma_S^2+\sigma_e^2} \quad (8)$$

Proportion of total variance in observations attributable to phylogeny:

$$\lambda' = \lambda(1-\psi) = \frac{\sigma_P^2}{\sigma_P^2+\sigma_S^2+\sigma_e^2} \quad (9)$$

Proportion of total variance that is attributable to error:

$$\varepsilon = (1-\lambda)(1-\psi) = \frac{\sigma_e^2}{\sigma_P^2+\sigma_S^2+\sigma_e^2} \quad (10)$$

As we note below, although the two parameters λ and ψ can play a similar role in determining the relative amounts of variance contributed by phylogenetic and non-phylogenetic sources, they are clearly identifiable.

For the cases in Figure 1B – D, it was assumed that each species was only assigned one treatment, i.e. the dimensions of **X** were *k* x *n*, the dimension of **B** were *b* x *n* and the dimensions of **y** were 1 x *n*. To generate random samples according to the scheme in Figure 1B half of the species were chosen at random and assigned to one treatment, the other half were assigned to another. In order to generate the phylogenetically structured distribution of treatments in Figure 1C we used the phylogeny to create a variance-covariance matrix and then generate a multivariate normal distribution using the mvtnorm package in R ((Genz et al., 2014, Genz and Bretz, 2009)). Species in the bottom 50% quartile of this random variable were assigned to one treatment, the rest were assigned to the other treatment. Finally, in order to generate the scheme shown in Figure 1D, species were ordered according to phylogenetic relatedness and assigned alternating states.

Methods of analysis

In the comparison of methods we outline six different approaches for analysing the data generated according to these models. (1), (2) and (4)-(6) may be used to fit to the experimental scheme shown in Figure 1A, whilst, (1) and (3) – (6) can be applied to data generated according to the schemes shown in Figure 1B-D. To make the models fully explicit we include R formulae for the models where appropriate. The full simulation code is available as a supplement.

In the models below, $y_{i,j,k}$ is the observation. $\beta_{0,i}$ is the intercept term with separate intercepts for the $i = 1...b$ blocks. $\beta_{X,j}$ is the effect of treatment j, where j

=1...*t* representing the *t* treatments. We assume that both block and treatment are fitted as fixed effects. Typically there are either too few blocks in an experiment or no within-block replication, so that it is not possible to efficiently or reliably treat blocks as random factors. Subscript *k* refers to species *k* = 1...*n*. The approaches to modelling the data differ with respect to how the species-specific effects are included. All models incorporate an error term $e_{i,j,k}$ which is a random error for each experimental observation, which is assumed to be normally distributed $e_{i,j,k} \sim N(0, \sigma_e^2)$.

Model (1) OLS ANOVA.

This is undoubtedly an inappropriate model whenever data contain phylogenetic signal. However, this approach is still commonly employed in the literature. We show the results obtained using this approach to demonstrate unequivocally the importance of accounting for phylogeny in such analyses. We simply fit 'block' and 'treatment' as fixed factors and species identity was ignored. The model fitted is:

$$y_{i,j,k} = \beta_{0,i} + \beta_{X,j} + e_{i,j,k} \quad (11)$$

In [R] format, the model is:

```
model1 <- lm( y ~ block + treatment )
```

Model (2) OLS ANOVA with species as a 'fixed' factor.

This is the simplest approach to including phylogenetic effects, by adding 'species' as an additional fixed factor, i.e.

$$y_{i,j,k} = \beta_{0,i} + \beta_k + \beta_{X,j} + e_{i,j,k} \quad (12)$$

In this model, the additional term β_k models species-specific variation. This is fitted in [R] using the model:

```
model2 <- lm( y ~ block + species + treatment )
```

Model (3) OLS ANOVA stratified by species.

For traits which species are assigned one treatment only (i.e. schemes B – D in Figure 1), model 2 will not fit because species and treatment are confounded. In this case the appropriate error for the model is the species level. This is the same as treating species as a split plot factor. In this case the fitted model is:

$$y_{i,j,k} = \beta_{0,i} + \beta_{X,k} + e_k + e_{i,j,k} \qquad (13)$$

The term e_k is a species-specific error term, $e_k \sim N(0, \sigma_S^2)$. The treatment is also species-specific, i.e. $\beta_{X,k}$ does not include the j subscript and consequently, the effective level of replication is the species. Hence the appropriate error term for testing the treatment effect is the variance at the species level. In [R] this is done by fitting the following:

```
model3 <- aov( y ~ block + treatment + Error( species ) )
```

Model (4) Linear Mixed Model with species as a 'random' factor.

In this model we treat 'species' as a random factor. For an orthogonal balanced design (e.g. scheme A in Figure 1), this offers no advantage over Model 2, and would yield identical results in terms of parameter errors and variances for the fixed effects. This is because the fixed effects in Model 4 are calculated marginally with respect to the random effects. The complication with Model 4 is the calculation of statistical significance for the main effects as the calculation of residual degrees of freedom is not straightforward (e.g. (Pinheiro and Bates, 2000)). We therefore fitted model 4 by maximum likelihood (rather than REML) and used likelihood ratio tests to test the fixed effect 'treatment'. The model fitted is:

$$y_{i,j,k} = \beta_{0,i} + b_k + \beta_{X,j} + e_{i,j,k} \qquad (14)$$

$b_k \sim N(0, \sigma_S^2)$

In this model, the random effects term b_k is assumed to be normally distributed with mean 0 and variance σ_S^2. The R command for fitting this model is:

```
model4 <- lmer( y ~ block + treatment + (1 | species) )
```

Model (5) PGLS model with known variance components

This model directly fits the 'true' variance structure with λ and ψ set to their true values. The reason for fitting this model is to demonstrate how the 'ideal' model for the system behaves relative to other approaches. We would expect this approach to behave well but we show that in some cases the other methods do approximate this model. As we outline below (model 6) it is possible to estimate the data-generating model.

We fitted model (1) directly to the data, using a PGLS approach (e.g.(Pagel, 1997, Pagel, 1999); (Hansen and Martins, 1996); (Freckleton et al., 2002)). The additional element here is that we account for the blocking and treatment structure inherent in an experiment of this design. In contrast the majority of previous analyses of comparative and cross-species experimental analyses focus on the analysis of data in which each species is represented by only a single measurement. Models for the analysis of data in which each species is measured more than once are described by (Felsenstein, 2008), (Ives et al., 2007) and (Hadfield and Nakagawa, 2010). Assuming that λ is known, we fitted a pgls model with block and treatment as fixed effects using maximum likelihood.

Model (6) Phylogenetic generalized least squares with estimated variance components.

At risk of giving away the punchline, our simulations shows that Model 5 is the *only* method that behaves well in *all* circumstances. Model 5 assumes that the parameter λ and ψ are known. However in real applications these parameters are unknown. Using the technique of 'Estimated' Generalised Least Squares (EGLS) (e.g. mentioned in a phylogenetic context by (Freckleton et al., 2002) and (Ives and Zhu, 2006)), the variance structure may be estimated from the data. We used the function lmekin() in the R package coxme ((Therneau, 2015)) to fit equation (1) to the data and estimate the variance components in equation (2).

Models (1) and (4) can be expressed as special cases of (5), specifically:

$\lambda = 0$ and $\psi = 1$: this is model (1)

$\lambda = 0$ and $0 < \psi < 1$: this is model (4)

$0 < \lambda < 1$ and $0 < \psi < 1$: this is model (5)

The middle model requires some explanation: in this model the phylogenetic variance structure is reduced to a diagonal matrix representing the model for the mean species responses. This implies no covariance among species resulting from phylogeny, but does represent a variance structure for within species variance. In the models considered here, this means that replicates from the same species are more similar to each other than to other species because $\sigma_S^2 > 0$. The difference between models (4) and (5) is that model (4) allows for species differences unrelated to phylogeny, whereas (5) models these differences as a function of phylogenetic distance.

Simulation details

Phylogenies of between 4 and 200 species were generated. The value of λ was set at 1 (high phylogenetic signal) generating a situation where species mean responses show

strong phylogenetic signal or set to zero, i.e. no phylogenetic signal in the mean response.

The effect of between species variation unrelated to phylogeny was assumed to be low ($\psi = 0$) or high ($\psi = 0.8$). Block effects were included throughout, although initial analysis showed that the assumptions about these were unimportant so long as all species x treatment combinations are present in each block. The treatment was assumed to be binary, and we either assumed no treatment effect (i.e. testing Type I error), or that there was a low to moderate treatment effect (i.e. testing power; $b_0 = 0$, $b_1 = 0.05, 0.1$).

All of the methods described above would be expected to yield unbiased estimates of the model coefficients ($\mathbf{b}_b$ and $\mathbf{b}_X$) in equation (1) ((McCullagh and Nelder, 1989)), so we did not study the behaviour of the parameter estimates further.

For each set of parameters we calculated the P-value for the inclusion of the treatment variable in the model as the main output of the simulation. This focus on P-values might seem a bit retro given the modern emphasis on effect sizes and model selection. However, in this context, in which we are studying the effect of including a binary factor in the model, the test effectively reduces to a t-test on the inclusion of an additional parameter, i.e. $t = (b_{obs} - b_{test}) / se(b)$ with $n - k$ degrees of freedom, where k is the number of estimated parameters. Given that we we know from theory that ($b_{obs} - b_{test}$) is unbiased (McCullagh & Nelder 1989), the P-value effectively summarises the information on both the variance and degrees of freedom ((Murtaugh, 2014)). In the simulations we found large effects of the method employed on the P-value, so this seems to be an informative metric. Moreover, in an experimental context, the primary aim of the analysis is usually to test the significance of the treatment variable, so this is a practically relevant measure.

Case study

In order to demonstrate the consequences of phylogenetic structure and method of analysis for the conclusions drawn from experimental data, we present an analysis of data taken from (Taylor et al., 2011)). We chose this dataset because the fully blocked experiment included treatments that were both fully randomized and phylogenetically structured. This dataset thus encompasses a broad range of scenarios considered in the simulations.

The data are taken from 13 species of grasses, 7 of which have C_4 photosynthesis and 6 of which have C_3 photosynthesis. Five experimental blocks were set up in which plants were either exposed to drought or were watered. One plant was assigned to each treatment within each block. There were 5 (blocks) x 2 (treatments) x 13 (species) = 130 experimental units at the start of the experiment. Full details are given in (Taylor et al., 2011). We report the analysis of measurements of photosynthetic rate (log transformed).

The two factors of interest in this experiment are the watering treatment and the photosynthetic pathway. The watering treatment was assigned in a fully randomized manner, whilst photosynthetic pathway was not. Photosynthetic pathway is a property of the species, and cannot be allocated at random to species. The experimental design involved some degree of phylogenetic balancing, so this treatment is somewhere between case (C) and (D) (see Figure 4).

We used this approach to analyse the effects of watering and photosynthetic type. The effect of watering treatment (watering versus drought) was first analysed using models (1) – (3). The effect of photosynthetic type (C_3 versus C_4) was analysed

using models (1), (2) and (4). The results are summarized in Table 1. To analyse the data from this experiment we used Model (6).

Results

Simulation results

The simulations show that the results of analyses of experimental data on multiple species are dependent on both the structure of the data and the method of analysis employed (Table 1 & 2). The only situation where results are robust to varying the analysis method employed is a fully randomized experiment in which all treatments are applied to each species (scenario A in Figure 1). Apart from Model 1, which ignores species identity completely, all methods have acceptable Type I error (Table 1A) and the same power (Table 2A & E). Model 1 ignores species identity and consequently is very conservative and has low power. This is because the variance resulting from species differences is not accounted for.

When each species does not receive every treatment, the results of the analyses are sensitive to the experimental design, method of analysis and degree of phylogenetic dependence. When phylogenetic dependence is strong, the difference in Type I error between phylogenetically balanced versus randomly distributed treatments in only evident for the non-phylogenetic analysis (Table 1B, D). Otherwise in these cases the performance of the phylogenetically corrected approaches is unaffected by the distribution of the traits. These methods are thus reasonably robust in terms of Type I errors.

On the other hand, phylogenetic structure in the treatment variable leads to high Type I errors for all methods apart from PGLS when phylogenetic signal is strong (Table 1C). Notably, the Type I error rate increases with the number of species

in the analysis, reflecting the misattribution of variance. Even for methods such as mixed models and ANOVA that 'control' for species differences, the rates of Type I error can exceed 50% for larger phylogenies.

In terms of power PGLS yields the highest or equal-highest power for all parameter combinations (Table 2). This emphasises the importance of accurately identifying the correct variance structure even in experimental analyses. Table 2 measures effective power, i.e. the proportion of times a statistically significant result is recorded, minus the proportion of times a Type I error is expected. Thus the power of tests that have high Type I errors in Table 1 is naturally low in Table 2.

In terms of design, fully randomised experiments have consistently high power (Table 2). Phylogenetically random and balanced designs combined with PGLS methods yield moderate to high power, with phylogenetically structured treatments yielding experiments with considerably lower effective power. When there is power over the allocation of treatments in an experiment, therefore, choice of design is important.

There is a qualitative difference between the effect of varying λ and Ψ. When λ is 1 and ψ is 0, PGLS (Model 5) has high power (Table 2), appropriate Type I error (Table 1) and other methods perform poorly. However, when ψ is high (0.8), then irrespective of the value of λ, the power of all techniques is low. The reason is that when ψ is high, there are large unknown differences between species, and these differences have to be estimated from the data. The effect of setting ψ greater than zero is akin to including phylogenetic branch lengths that are unknown resulting in unaccounted differences between species. In contrast when λ is high and the phylogeny is known, the expected covariance among species is known and can be potentially corrected for. The need to estimate species effects from the data when ψ is

high therefore reduces the power. Note, that the same is true for models (2)-(4) when applied to data in which ψ is 0, but λ is high. In this case these methods are attempting to deal with phylogenetic independence without an estimate of the covariance structure among the species and yield similarly low power (Table 2).

In summary, each of models (1) – (5) can yield statistically acceptable or equivalent results under some circumstances. However, PGLS is the only method to perform well under all tested scenarios. The problem, of course is that our simulations with PGLS assume that the true variance structure is known. In reality the true variance structure can never be known, with one course of action being that we estimate this from the data. The EGLS simulations in which the variance structure is estimated from the data indicate that the outcome of this can be mixed.

EGLS performs well in terms of Type I errors (Table 1) or power (Table II) when phylogenetic signal is low ($\lambda = 0$). In this case the method is accurate at identifying a lack of phylogenetic signal (e.g. see simulations in Freckleton et al. 2002) and the results are almost identical to those from the PGLS in which ψ is fixed to its true value.

When phylogenetic signal is high ($\lambda = 1$), then there are two circumstances under which EGLS performs below PGLS in terms of power: (i) when the number of species is small and experiments are not fully randomised; (ii) when the treatments are phylogenetically structured. Particularly in the latter case the rates of Type I error can be considerable, although not as high as those of methods that do not appropriately account for phylogeny. Unfortunately when phylogenetic signal exists in the treatment variable, our results indicate that it may be difficult to reliably fit models to experimental data.

Analysis of real data

As shown in Figure 2, the structure of the data mirrors the situations envisaged in Figure 1. The watering treatment was fully randomized as in Figure 1A. The photosynthetic type varies with a combination of phylogenetic balancing by design (Figure 1D) and phylogenetic structure (Figure 1C). The analysis of the effect of the watering treatment indicated that the conclusions drawn were relatively invariant to the choice of method of analysis. In this case, the results obtained from ANOVA, random effects model and the PGLS were very similar indeed (Table 3a). The probability value for the OLS model ignoring phylogeny altogether was larger reflecting the lower power of this method observed in the simulations. Note that in this dataset the maximum likelihood value of λ is zero, which means that the PGLS model and the REML mixed model produce equivalent parameter estimates (to within a trivial numerical difference).

The analysis of effect of photosynthetic type however indicated that the choice of method was important. Depending on the approach chosen, the result was statistically highly significant (OLS), clearly non-significant (ANOVA), marginally non-significant (random effects model) or marginally significant (EGLS) (Table 1b). This reinforces the conclusion from the modelling, namely that when phylogenetic signal is strong in both the data and the test variables, the results obtained may be very sensitive to the choice of model.

Discussion

The strongest message from the results presented above is that the power of comparative experiments is maximized when the correct variance structure is

incorporated into the analysis. Depending on the design of the experiment, there may also be increased Type I error rates when the correct variance structure is not included. This occurs particularly if the treatments are naturally varying ones that show phylogenetic structure, and in this case the more commonly used methods for experimental data analysis can perform poorly. The Type I error rate for EGLS is frequently high, which is a concern: the acceptable rates for PGLS rely on knowing the correct variance structure, however this can only ever be estimated empirically.

The only experimental design that is completely robust is the fully randomized design, with all treatments applied to each species. For many factors of interest, however, this is not possible and consequently the method of analysis is important. To illustrate this, we presented a simple case study in which it is possible to get a range of answers depending on the method employed.

The analysis of experimental data seems to have been overlooked in the comparative literature. In the development of new methods it is typically assumed that phylogenetic comparative methods are applied in a correlative manner, frequently to literature-derived observational data. On the other hand, many studies report experiments performed on multiple species and our results highlight that these should carefully consider the choice of analysis.

One approach to analyzing the data from multiple species in an experiment is to generate a single mean for the response for each species, and use this in conventional phylogenetic analysis (e.g. PGLS). In the example considered in the simulations above, this would involve averaging values across the blocks. In the case of the fully randomized design (Figure 1A) this would generate two values per species; in the case of the others (Figure 1B – D) this would yield one value per species. This analysis is effectively the same as model (3) if no further phylogenetic

correction is applied. This is justifiable, but with three limitations: (i) the design of the experiment must permit simple averaging of species traits (e.g. ideally completely balanced). (ii) Consequently there can be no missing data as the means are taken across experimental replicates. (iii) The estimates of phylogenetic signal and experimental error cannot be disentangled, i.e. λ and □□□ cannot be separately estimated. If the experimental design is not simple (e.g. split plot or repeated measures are included) then the model could not be fit using this approach without carefully and appropriately averaging across clusters, and we are aware of no examples of studies that have analysed such designs in a phylogenetic comparative context.

If these three conditions hold, then the implementation of the phylogenetic approach described above (Model 6) can be simplified. Recall the net variance matrix for all observations is (equation 2):

$$\mathbf{W} = \sigma[\lambda(1-\psi)\mathbf{D} \otimes \mathbf{V} + \psi\mathbf{S} + (1-\lambda)(1-\psi)\mathbf{1}]$$

This can be simplified to the following model for just the species means:

$$\mathbf{W} = \sigma[\lambda'\mathbf{D} \otimes \mathbf{V} + \mathbf{\Sigma}] \quad (17)$$

In equation (17) $\mathbf{\Sigma}$ is a vector of variances, comprising the individual variance for each species and the variance about the mean resulting from error. In constructing $\mathbf{\Sigma}$ it would be tempting to calculate the variance across experimental units for each species and use these values as species-specific estimates of variance. In the above simulations, for example, this would be the variance in measurements for each species across the blocks (in each treatment where relevant). This approach has been used to model measurement errors in comparative analyses (e.g. (Silvestro et al., 2015)). However, this would be inadvisable in the analysis of experimental data, as the individual species-level values will individually be poor estimates of the residual

variance. This is particularly so if the number of blocks is small (typical values being in the range 3 – 10 for most experiments). It is advised that measurement error cannot be accurately calculated from fewer than ~12 observations (Hansen and Bartoszek, 2012). The better approach would be to use a pooled estimate of variance across all of the replicates, as in conventional ANOVA.

It should also be noted that $\mathbf{\Sigma}$ in equation (17) includes the variance attributed to species mean differences generated by non-zero values of ψ. Adding a vector of expected intra-specific variances will not account for such differences. Empirically, however, this should be compensated through the estimate of λ: as pointed out above the effects of varying λ and ψ are not separable when data are summarised to species means.

Models for intra-specific variation in comparative data have been described, and these are closely related to the approaches for analysing experimental data described here (e.g. (Felsenstein, 2008)). Such models allow for intra-specific variation and are specifically designed to consider the variation within measured units, which is typically assumed to have a deterministic basis. These methods are essentially the same as the pgls approach: for example, through including additional variables experimental designs such as blocking could be included.

Linear mixed models are frequently used to analyse data from experiments, with species specified as a random factor (e.g. (Jamil et al., 2013, Brown et al., 2014)). The problems with using LMMs are two-fold. First, although they account for differences between species, they do not allow for higher level phylogenetic dependence. They assume that $\lambda = 0$, although the inclusion of higher level taxonomic grouping factors could improve the performance of this approach providing the taxonomic relationships reflect the underlying phylogeny. The second problem with

LMMs is that it is not straightforward to calculate degrees of freedom relating to the random components of the model. If the random components of the models are not varied then this is not an issue for examining the fixed effects. It is worth noting that in the situations simulated above the parameter estimates and error variances from the LMM and the fixed effect models were identical. This strongly suggests that in the LMMs the number of grouping factors (i.e. # of species) is a number of estimated parameters.

Our analyses of real data showed that in one case the results obtained were relatively insensitive to the choice of method. But in another, the results in terms of the fixed factor of interest varied widely. Depending on the choice of method, the treatment effect was highly significant, non-significant, marginally non-significant or marginally significant.

The comparative method is typically thought of as an observational approach (Harvey & Pagel 1990). In practice, however, there are many studies that integrate comparative and experimental methods (Weber & Agrawal 2012). Arnold & Nunn (2010) considered one aspect of this integration, namely how one chooses species in order to maximise the power of analyses. They highlighted that appropriate phylogenetic targeting could considerably enhance the power of tests and that even for the same number of species, power could vary considerably depending on how they are distributed phylogenetically. Our results similarly show that the phylogenetic distribution of traits play an important role in determining both the Type I error and power of experimental studies.

Recommendations

In designing experiments our results highlight some clear recommendations. Most importantly if the effects of naturally varying traits are examined, then the phylogenetic distribution of these is extremely important. Our simulation analysis assumed that this was generated according to a discretized Brownian process. This model generates phylogenetic structure (e.g. see also) but more extreme cases are possible. For example, "grade shifts" occur when the values of entire clades are identical. In the Brownian model the outcome is not as extreme and there is usually variation within clades (e.g. See case (C) in Figure 1). Examples could include sexual determinations systems, which are XY for mammals and WZ for birds (). Comparative analysis on such traits would be more complex: Model 3 would be relevant with the stratum (Error(…) in the model function) defined by clades rather than species in such cases.

We recommend that the phylogenetic distribution of experimental factors is checked prior to undertaking an experiment. For example, the phylogenetic signal of a binary predictor can be assessed using the D statistic (Fritz and Purvis, 2010). For a multi-level predictor, the phylogenetic distribution of each level could be assessed using the same approach. Our results clearly indicate that all empirically applicable methods, including EGLS, may be severely compromised if experimental factors show strong phylogenetic dependence, and that approaches such as phylogenetic balancing can help to ensure that power and Type I error are improved.

The technique of phylogenetic balancing is closely related to the sister-species approach that has been used a great deal over several decades (e.g. see (Weir and Lawson, 2015) for a recent application). The sister species approach is based on choosing closely related pairs of species that differ in some key characteristic. Differences between other traits of the sister species are then tested. This is akin to

blocking in a conventional experimental design. The ANOVA and LMM approaches considered here are related, but lack power when there is considerable phylogenetic signal (Table 2D & H). When phylogenetic signal is considerable, both PGLS and EGLS are able to both deal yield high power and appropriate Type I error. Sister species comparisons would ignore higher level confounding (being based on comparisons at the species-pair level only) and also reduce sample sizes relative to fully phylogenetically explicit methods (being based on $n/2$ comparisons for n species).

We would also recommend the use of simulations to test the statistical performance of experimental designs prior to analyses. Such simulations are relatively straightforward to implement and would give insights into the likely pitfalls of any proposed experimental schemes, or the relative performance of alternative designs.

The results with varying ψ indicated that high values of ψ led to low power for all tests apart from fully randomized experiments, irrespective of sample size. As noted above, high values of ψ yield large species specific mean differences that detract from the power to compare species in terms of other traits. Fully randomised experiments perform better because these differences are controlled through design and, indeed, the results are thus insensitive to the method of analysis. It cannot be predicted in advance whether species differ in this way, and hence a risk of any experiment that relies on natural variation is that species-specific non-phylogenetic variation might obscure treatment effects.

Concluding remarks

Despite the popularity of comparative methods, the bottom line in our analysis is that the fully randomized block experiment remains the gold standard for experimental

cross-species studies. Our results highlight several pitfalls, notably when natural variation among species is used to define experimental treatments and this variation is phylogenetically structured, as well as when species show large non-phylogenetic mean differences. Techniques such as EGLS can under many circumstances approximate the ideal PGLS solution, however our simulations reveal that there are conditions under which methods perform poorly. Our overall recommendation is that such eventualities are considered and to the greatest extent possible dealt with at the design rather than analysis stage.

Table 3 Analysis of real experiment data using different approaches. The data are taken from Taylor et al. (2011). The response variable is log(photosynthetic rate). The predictors in the model are watering treatment (assigned in a randomized, fully factorial manner) and photosynthetic type (C3 or C4, with each species being one or the other). The data were analysed in 4 ways: (i) OLS: the predictors were fitted singly, not accounting for phylogeny. (ii) ANOVA: in the case of watering treatment, species was fitted as an additional fixed factor; photosynthetic type varies at the species level, so a split-plot ANOVA was used; (iii) Random effects: species identity was fitted as a random effect. This was fitted by Maximum Likelihood (ML) and the effect of the predictor tested using a likelihood ratio test, relative to a simpler model. (iv) PGLS: as described in the text, a phylogenetic variance matrix was included in the model, and the parameters λ and ψ were estimated to measure the effects of phylogenetic structure and species-specific variation, respectively.

Model	(a) Watering Treatment		(b) Photosynthetic type	
OLS	□	-0.078	0.520	
	se	0.131	0.123	
	t	-0.598	4.237	
	P	0.551	0.000	
ANOVA	□	-0.083	-	
	se	0.094	-	
	t	-0.884	F = 1.081	
	P	0.379	0.370	
Random		ML	REML	ML	REML
Effects	□	-0.083	-0.083	0.499	0.498

	se	0.092	0.094	0.258	0.281
	t	-0.900	-0.880	1.934	1.776
	P^*	0.369		0.070	
PGLS	□	-0.083	$\lambda = 0$	0.493	$\lambda = 0.483$
	se	0.092	$\psi = 0.487$	0.234	$\psi = 0$
	t	-0.900		2.100	
	P	0.370		0.036	

* - tested by likelihood ratio test.

Figure Legends

Figure 1 Possible experimental designs by which two treatments are assigned to multiple species. (A) A fully randomized design, with both treatments assigned to each species. In (B), (C) and (D), treatments are a property of the species and are not fully randomized. (B) Treatments are assigned at random to species. (C) Treatments are phylogenetically non-randomly distributed, such that closely related species are more similar to each other than to distantly related ones. (D) Phylogenetically balanced treatments, i.e. species are chosen so that sister species differ in the factor studied.

Figure 2 Experimental design in an example dataset. There are 13 species related by the phylogeny shown. There were two factors considered. Watering treatment was a binary variable, with each treatment applied to each species. Photosynthetic type is a binary variable, but each species was either C_3 or NADP.

References

Brown, A. M., Warton, D. I., Andrew, N. R., Binns, M., Cassis, G. & Gibb, H. (2014) The fourth-corner solution - using predictive models to understand how species traits interact with the environment. *Methods in Ecology and Evolution,* **5,** 344-352.

Felsenstein, J. (2008) Comparative methods with sampling error and within-species variation: contrasts revisited and revised. *Amercian Naturalist,* **171,** 713-725.

Freckleton, R. P., Harvey, P. H. & Pagel, M. (2002) Phylogenetic analysis and comparative data: A test and review of evidence. *American Naturalist,* **160,** 712-726.

Fritz, S. A. & Purvis, A. (2010) Selectivity in mammalian extinction risk and threat types: a new measure of phylogenetic signal strength in binary traits. *Conservation Biology,* **24,** 1042-1051.

Genz, A. & Bretz, F. (2009) *Computation of multivariate normal and t probabilities.* Springer-Verlag, Heidelberg.

Genz, A., Bretz, F., Miwa, T., Mi, X., Liesch, F., Scheipl, F. & Hothorn, T. (2014) mvtnorm: Multivariate normal and t distributions. R package 1.0-2.

Grafen, A. (1989) The phylogenetic regression. *Philosophical Transactions of the Royal Society of London Series B,* **326,** 119-157.

Grime, J. P., Hodgson, J. G. & Hunt, R. (1990) *Comparative Plant Ecology.* Chapman & Hall, London.

Hadfield, J. D. & Nakagawa, S. (2010) General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for

continuous and categorical characters. *Journal of Evolutionary Biology,* **23,** 494-508.

Hansen, T. F. & Bartoszek, K. (2012) Interprerting the evolutionary regression: the interplay between observational and biological errors in phylogenetic comparative studies. *Systematic Biology,* **61,** 413-425.

Hansen, T. F. & Martins, E. P. (1996) Translating between microevolutionary process and macroevolutionary patterns: the correlation structure of interspecific data. *Evolution,* **50,** 1404-1417.

Harvey, P. H. (1996) Phylogenies for ecologists. *Journal of Animal Ecology,* **65,** 255-263.

Harvey, P. H. & Pagel, M. D. (1991) *The comparative method in evolutionary biology.* Oxford University Press, Oxford.

Harvey, P. H., Read, A. F. & Nee, S. (1995) Why ecologists need to be phylogenetically challenged. *Journal of Ecology,* **83,** 535-536.

Ives, A. R., Midford, P. E. & Garland, T. (2007) Within-species variation and measurement error in phylogenetic comparative methods. *Systematic Biology,* **56,** 252-270.

Ives, A. R. & Zhu, J. (2006) Statistics for correlated data: phylogenies, space and time. *Ecological Applications,* **16,** 20-32.

Jamil, T., Ozinga, W. A., Kleyer, M. & ter Braak, C. J. F. (2013) Selecting traits that explain species-environment relationships: a generalized linear mixed model approach. *Journal of Vegetation Science,* **24,** 988-1000.

Martins, E. P. & Garland, T. (1991) Phylogenetic analyses of the correlated evolution of continuous characters: a simulation study. *Evolution,* **45,** 534-557.

Martins, E. P. & Hansen, T. F. (1996) The statistical analysis of interspecific data: a review and evaluation. *Phylogenies and the comparative method in animal behaviour* (ed E. P. Martins), pp. 22-75. Oxford University Press, Oxford.

Maynard Smith, J. (1978) Optimization theory in evolution. *Annual Review of Ecology and Systematics,* **9,** 31-56.

McCullagh, P. & Nelder, J. A. (1989) *Generalisd linear models.* Chapman & Hall, London.

Murtaugh, P. A. (2014) In defense of P values. *Ecology,* **95,** 611-617.

Orme, D. (2013) The caper package: comparative analysis of phylogenetics and evolution in R.

Pagel, M. (1997) Inferring evolutionary processes from phylogenies. *Zoologica Scripta,* **26,** 331-348.

Pagel, M. (1999) Inferring the historical patterns of biological evolution. *Nature,* **401,** 877-884.

Paradis, E., Claude, J. & Strimmer, K. (2004) APE: analysis of phylogenetics and evolution in R language. *Bioinformatics,* **20,** 289-290.

Pennell, M. W., Eastman, J. M., Slater, G. J., Brown, J. W., Uyeda, J. C., FitzJohn, R. G., Alfaro, M. E. & Harmon, L. J. (2014) geiger 2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics,* **30,** 2216-2218.

Pinheiro, J. C. & Bates, D. M. (2000) Springer, New York.

Silvestro, D., Kosktikova, A., Litsios, G., Pearman, P. B. & Salamin, N. (2015) Measurment errors should always be incorporated in phylogenetic comparative analysis. *Methods in Ecology and Evolution,* **6,** 340-346.

Stadler, T. (2015) TreeSim: simulating phylogenetic Trees. R package version 2.2.

Taylor, S. H., Ripley, B. S., Woodward, F. I. & Osborne, C. P. (2011) Drought limitation of photosynthesis differs betwee C3 and C4 grass species in a comparative experiment. *PLant , Cell and Environment,* **34,** 65-75.

Team, R. C. (2015) R: a language and environment for statistical computing. R Foundation for Statistical Computing Vienna, Austria.

Therneau, T. M. (2015) coxme: mixed effects cox models. R package version 2.2-5.

Weir, J. T. & Lawson, A. (2015) Evoluitonary rates across gradients. *Methods in Ecology and Evolution,* **6**.

Table 1 Type I error rates for different methods of analysis of experimental data generated according to a range of different designs and assuming a 5% threshold for statistical significance. Type I errors proportions of 0.1 and higher (i.e. twice the nominal rate and higher) are highlighted in red.

		$\lambda = 1$ / $\psi = 0.8$				$\lambda = 1$ / $\psi = 0$				$\lambda = 0$ / $\psi = 0.8$			
		Number of species				Number of species				Number of species			
Experiment Design	Analysis Method	10spp	20spp	50spp	100spp	10spp	20spp	50spp	100spp	10spp	20spp	50spp	100spp
Effect Size = 0													
(A) Fully	(1) OLS	0.04	0.04	0.04	0.04	0.00	0.00	0.00	0.00	0.03	0.03	0.03	0.03
randomized	(2) Anova	0.05	0.05	0.06	0.06	0.05	0.05	0.05	0.04	0.05	0.06	0.05	0.05
	(3) Anova	-	-	-	-	-	-	-	-	-	-	-	-
	(4) LMM	0.05	0.05	0.06	0.06	0.05	0.06	0.05	0.04	0.06	0.06	0.05	0.05
	(5) PGLS	0.05	0.04	0.05	0.05	0.03	0.02	0.01	0.00	0.06	0.06	0.05	0.05
	(6) EGLS	0.07	0.06	0.06	0.06	0.04	0.03	0.03	0.04	0.06	0.06	0.06	0.04
(B) Phylo-													
genetically	(1) OLS	0.11	0.13	0.10	0.12	0.47	0.43	0.41	0.40	0.20	0.16	0.15	0.16
random	(2) Anova	-	-	-	-	-	-	-	-	-	-	-	-

	(3) Anova	0.05	0.05	0.05	0.06	0.05	0.05	0.06	0.04	0.07	0.05	0.05	0.06
	(4) LMM	0.07	0.06	0.05	0.06	0.09	0.07	0.06	0.05	0.10	0.06	0.05	0.06
	(5) PGLS	0.05	0.05	0.05	0.05	0.04	0.04	0.06	0.05	0.07	0.05	0.05	0.06
	(6) EGLS	0.08	0.05	0.04	0.03	0.22	0.12	0.09	0.07	0.09	0.05	0.06	0.05
(C) Phylo-	(1) OLS	0.21	0.29	0.46	0.60	0.67	0.75	0.81	0.86	0.17	0.15	0.18	0.16
genetically	(2) Anova	-	-	-	-	-	-	-	-	-	-	-	-
structured	(3) Anova	0.11	0.21	0.36	0.51	0.23	0.37	0.56	0.67	0.05	0.05	0.05	0.05
	(4) LMM	0.16	0.23	0.37	0.52	0.30	0.40	0.57	0.67	0.08	0.06	0.06	0.05
	(5) PGLS	0.04	0.05	0.06	0.05	0.06	0.06	0.07	0.06	0.05	0.04	0.05	0.04
	(6) EGLS	0.13	0.16	0.16	0.11	0.36	0.33	0.29	0.26	0.05	0.04	0.05	0.04
(D) Balanced	(1) OLS	0.07	0.06	0.04	0.05	0.24	0.15	0.04	0.02	0.17	0.15	0.15	0.15
Design	(2) Anova	-	-	-	-	-	-	-	-	-	-	-	-
	(3) Anova	0.02	0.02	0.02	0.02	0.00	0.00	0.00	0.00	0.05	0.05	0.04	0.05
	(4) LMM	0.05	0.03	0.02	0.02	0.02	0.00	0.00	0.00	0.09	0.06	0.05	0.05
	(5) PGLS	0.06	0.06	0.05	0.06	0.07	0.04	0.05	0.06	0.05	0.05	0.06	0.05
	(6) EGLS	0.05	0.05	0.05	0.05	0.11	0.05	0.02	0.01	0.06	0.07	0.06	0.04

Table 2 Power of different methods of analysis of experimental data generated according to a range of different designs. In this Table, power is measured as the proportion of times that a statistically significant result is recorded, minus the proportion of times a Type I error is expected (Table 1).

		$\lambda = 1 / \psi = 0.8$				$\lambda = 1 / \psi = 0$				$\lambda = 0 / \psi = 0.8$			
		Number of species				Number of species				Number of species			
Experiment Design	Analysis Method	10spp	20spp	50spp	100spp	10spp	20spp	50spp	100spp	10spp	20spp	50spp	100spp
Effect Size = 0.05													
(A) Fully	(1) OLS	0.01	0.02	0.06	0.08	0.00	0.01	0.04	0.16	0.02	0.02	0.04	0.09
randomized	(2) Anova	0.01	0.02	0.07	0.10	0.95	0.95	0.96	0.96	0.02	0.02	0.06	0.12
	(3) Anova	-	-	-	-	-	-	-	-	-	-	-	-
	(4) LMM	0.02	0.03	0.07	0.10	0.95	0.94	0.95	0.96	0.02	0.02	0.06	0.13
	(5) PGLS	0.02	0.03	0.07	0.12	0.96	0.98	1.00	1.00	0.01	0.02	0.06	0.12
	(6) EGLS	0.01	0.03	0.06	0.12	0.96	0.97	0.97	0.96	0.00	0.02	0.04	0.11
(B) Phylo-	(1) OLS	0.01	0.01	0.03	0.05	0.02	0.02	0.04	0.08	0.00	0.01	0.04	0.03
genetically	(2) Anova	-	-	-	-	-	-	-	-	-	-	-	-
random	(3) Anova	0.01	0.01	0.00	0.04	0.00	0.00	0.00	0.03	0.00	0.02	0.02	0.00

	(4) LMM	0.01	0.01	0.01	0.03	0.00	0.00	0.00	0.04	0.00	0.01	0.02	0.00
	(5) PGLS	0.00	0.02	0.01	0.07	0.04	0.12	0.55	0.89	0.00	0.00	0.01	0.00
	(6) EGLS	0.00	0.01	0.02	0.02	0.00	0.04	0.14	0.42	0.00	0.03	0.00	0.01
(C) Phylo-	(1) OLS	0.02	0.02	0.00	0.00	0.00	0.00	0.02	0.02	0.01	0.03	0.01	0.05
genetically	(2) Anova	-	-	-	-	-	-	-	-	-	-	-	-
structured	(3) Anova	0.00	0.01	0.00	0.00	0.00	0.01	0.04	0.02	0.01	0.00	0.02	0.03
	(4) LMM	0.00	0.01	0.00	0.00	0.00	0.01	0.05	0.02	0.02	0.01	0.02	0.03
	(5) PGLS	0.01	0.01	0.00	0.01	0.00	0.02	0.06	0.20	0.01	0.02	0.02	0.04
	(6) EGLS	0.01	0.00	0.00	0.02	0.00	0.00	0.04	0.10	0.00	0.00	0.03	0.02
(D) Balanced	(1) OLS	0.00	0.02	0.02	0.04	0.01	0.02	0.07	0.11	0.00	0.01	0.02	0.07
	(2) Anova	-	-	-	-	-	-	-	-	-	-	-	-
	(3) Anova	0.01	0.00	0.01	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03
	(4) LMM	0.00	0.01	0.01	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.04
	(5) PGLS	0.01	0.01	0.02	0.04	0.05	0.20	0.73	0.94	0.00	0.00	0.01	0.03
	(6) EGLS	0.00	0.02	0.02	0.01	0.02	0.06	0.21	0.56	0.02	0.03	0.00	0.03

Effect size = 0.1

(E) Fully	(1) OLS	0.01	0.08	0.22	0.41	0.06	0.13	0.52	0.79	0.04	0.07	0.17	0.37
randomized	(2) Anova	0.02	0.10	0.24	0.43	0.95	0.95	0.96	0.96	0.05	0.08	0.21	0.45
	(3) Anova	-	-	-	-	-	-	-	-	-	-	-	-
	(4) LMM	0.02	0.11	0.25	0.43	0.95	0.94	0.95	0.96	0.05	0.08	0.21	0.45
	(5) PGLS	0.02	0.10	0.24	0.43	0.97	0.98	1.00	1.00	0.04	0.08	0.22	0.45
	(6) EGLS	0.06	0.09	0.23	0.46	0.96	0.97	0.97	0.96	0.05	0.09	0.23	0.48
(F) Phylo-	(1) OLS	0.03	0.04	0.13	0.17	0.01	0.03	0.09	0.17	0.00	0.04	0.11	0.16
genetically	(2) Anova	-	-	-	-	-	-	-	-	-	-	-	-
random	(3) Anova	0.01	0.02	0.04	0.09	0.00	0.01	0.03	0.10	0.00	0.02	0.06	0.08
	(4) LMM	0.01	0.02	0.04	0.09	0.00	0.01	0.03	0.10	0.00	0.02	0.06	0.08
	(5) PGLS	0.01	0.02	0.06	0.10	0.11	0.38	0.88	0.95	0.00	0.01	0.05	0.08
	(6) EGLS	0.02	0.02	0.04	0.09	0.05	0.17	0.42	0.83	0.00	0.03	0.03	0.09
(G) Phylo-	(1) OLS	0.01	0.05	0.01	0.00	0.00	0.00	0.01	0.01	0.00	0.05	0.06	0.17
genetically	(2) Anova	-	-	-	-	-	-	-	-	-	-	-	-
structured	(3) Anova	0.01	0.03	0.03	0.10	0.01	0.03	0.00	0.03	0.00	0.03	0.03	0.09
	(4) LMM	0.01	0.05	0.02	0.10	0.01	0.04	0.00	0.03	0.00	0.04	0.03	0.10

	(5) PGLS	0.02	0.04	0.05	0.10	0.03	0.08	0.25	0.52	0.01	0.03	0.03	0.10
	(6) EGLS	0.01	0.02	0.05	0.10	0.04	0.01	0.13	0.28	0.01	0.02	0.06	0.07
(H) Balanced	(1) OLS	0.00	0.00	0.08	0.14	0.06	0.10	0.28	0.47	0.02	0.04	0.10	0.18
	(2) Anova	-	-	-	-	-	-	-	-	-	-	-	-
	(3) Anova	0.00	0.02	0.04	0.09	0.01	0.00	0.01	0.02	0.01	0.01	0.06	0.10
	(4) LMM	0.00	0.03	0.06	0.10	0.00	0.01	0.01	0.02	0.01	0.02	0.06	0.11
	(5) PGLS	0.00	0.02	0.04	0.09	0.17	0.57	0.94	0.94	0.01	0.01	0.05	0.09
	(6) EGLS	0.00	0.02	0.04	0.09	0.10	0.20	0.60	0.95	0.02	0.02	0.04	0.11

Figure 1

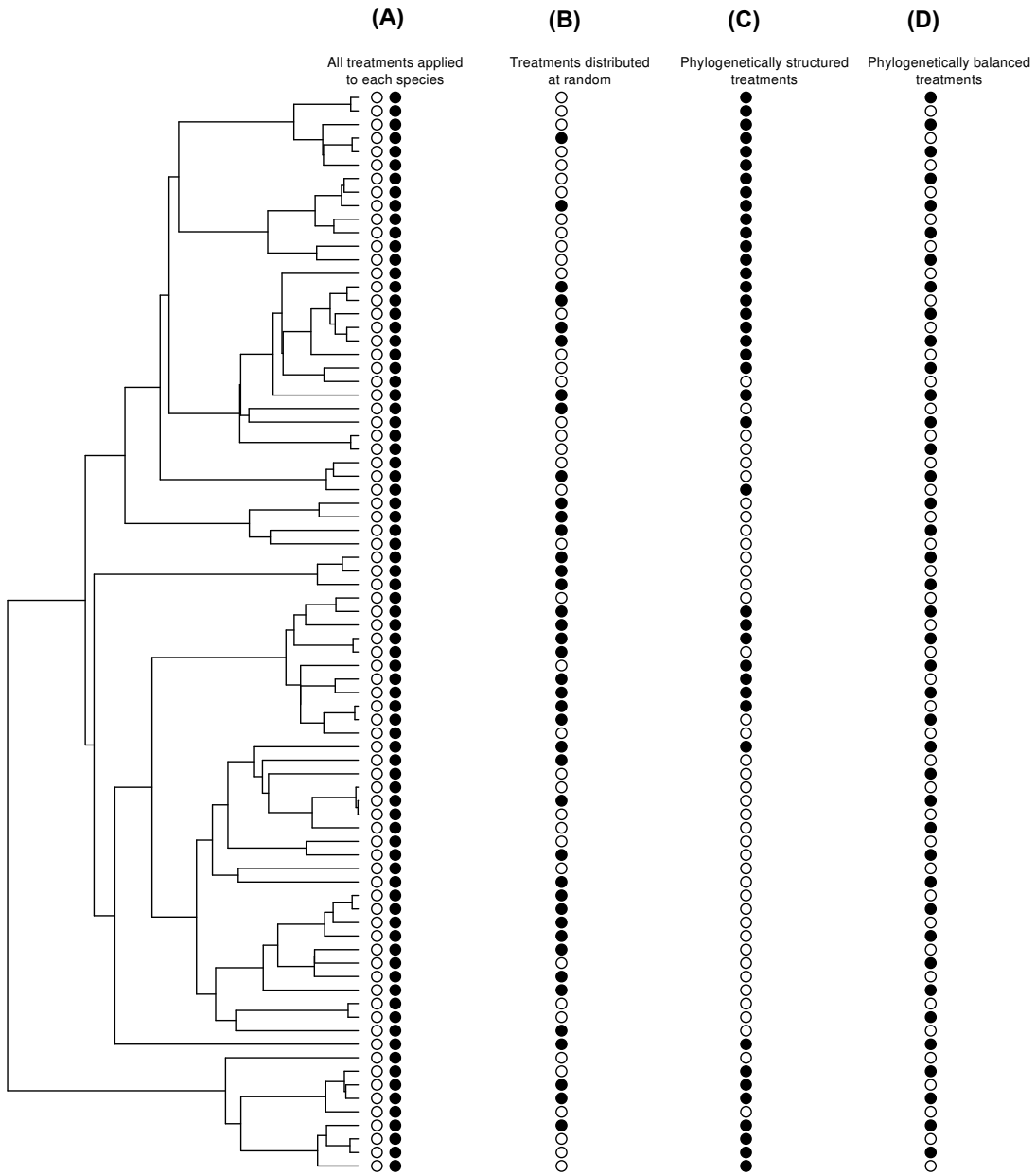
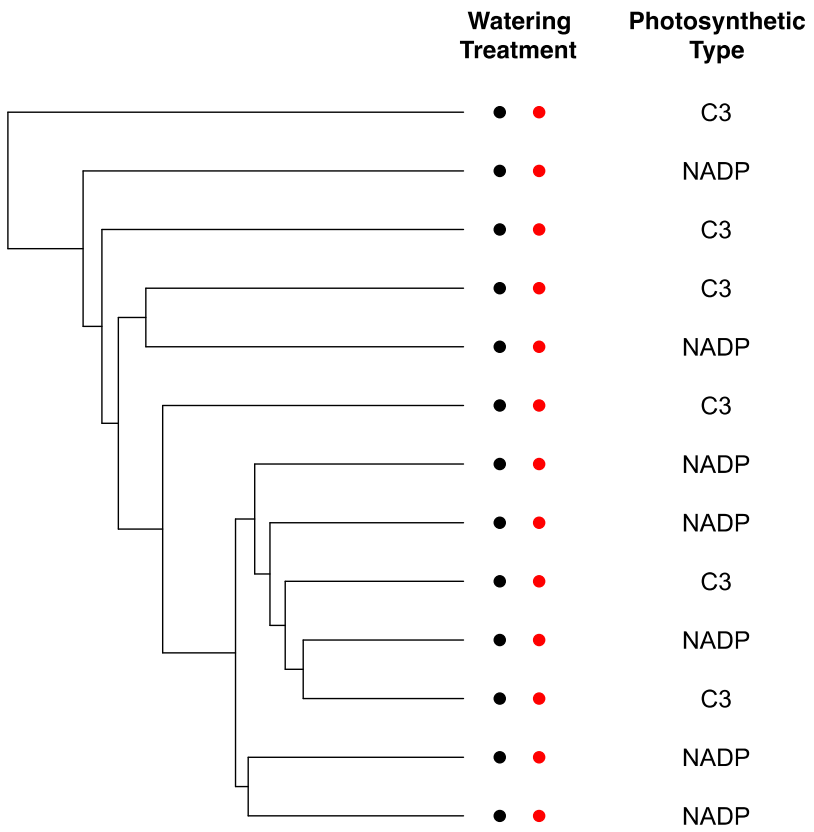

(A)
All treatments applied to each species
(B)
Treatments distributed at random
(C)
Phylogenetically structured treatments
(D)
Phylogenetically balanced treatments

Figure 2


Watering Treatment
Photosynthetic Type
C3
NADP
C3
C3
NADP
C3
NADP
NADP
C3
NADP
C3
NADP
NADP