Evolutionary Ecology Research, 2006, **8**: 975–995

A method for calculating means and variances of comparative data for use in a phylogenetic analysis of variance

Patrik Lindenfors*

Department of Zoology, Stockholm University, SE-106 91 Stockholm, Sweden

ABSTRACT

Question: Is there a general method for applying an analysis of variance (ANOVA) to phylogenetic data sets that allows for the incorporation of any assumption of evolutionary model?

Mathematical method: I describe a method that enables the calculation of means and variances of the traits of monophyletically related species. These means and variances are calculated using evolutionary rates derived from estimated nodal values and can thus incorporate any assumption of evolutionary model. Furthermore, calculations of phylogenetically corrected mean squares are described through a worked example to show how these can be implemented into a complete-blocks ANOVA.

Key assumptions: The assumptions of this method are, as for all comparable methods, that the phylogeny is known without error and that estimated nodal values are correct. Since the method can utilize nodal values using any evolutionary model, the latter assumption can to a degree be circumvented.

Conclusions: I compare the presented method to alternative tests conceptually, but also using an idealized example and an example derived from the real world, to show the utility of the approach described herein.

Keywords: analysis of variance, comparative methods, independent contrasts, phylogenies.

INTRODUCTION

One of the most commonly used statistical methods in the biological sciences is the analysis of variance (ANOVA). The aim of this paper is to provide a general method – general in the sense that it is applicable regardless of any assumption of evolutionary model – of how to make phylogenetic adjustments of means and variances so that they can be used in an ANOVA. A typical case where a phylogenetic ANOVA could be applied is where a categorical trait is thought to exert a selective influence on a continuous trait. The hypothesis to be tested would then be that species having one state of the categorical trait should have a differing mean of the continuous trait compared with species having another state of the categorical trait.

* e-mail: Patrik.Lindenfors@zoologi.su.se

Several phylogenetic comparative methods are available, many of which are aimed at making phylogenetically correct regression or correlation analyses between two or more continuous variables (e.g. Felsenstein, 1985; Grafen, 1989; Pagel, 1998), or, conversely, analysing the evolutionary relationships between two discrete characters (e.g. Ridley, 1983; Maddison, 1990; Pagel, 1994). However, quite a few methods have also been proposed to deal with the relationship between one discrete and one or several continuous variables (Stearns, 1983; Grafen, 1989; Møller and Birkhead, 1992; Wickman, 1992; Garland *et al.*, 1993; Purvis and Rambaut, 1995; Hansen and Martins, 1996; Martins and Hansen, 1997; Lindenfors and Tullberg, 1998; Pagel, 1998; Butler *et al.*, 2000; Butler and King, 2004). For example, an early method developed by Stearns (1983) removed variation that could be attributed to a certain level of phylogenetic (or taxonomic) relatedness in preparation for a normal ANOVA. This method, however, assumed statistical independence at all phylogenetic levels above those removed and also removed variation that needed to be explained (Harvey and Pagel, 1991).

Of the other methods available, some pair comparisons by matching clades differing in the categorical trait (Møller and Birkhead, 1992; Wickman, 1992; Purvis and Rambaut, 1995), while others do not pair, but instead group clades sharing the same categorical trait (Grafen, 1989; Garland *et al.*, 1993; Hansen and Martins, 1996; Martins and Hansen, 1997; Lindenfors and Tullberg, 1998; Pagel, 1998; Butler *et al.*, 2000; Butler and King, 2004).

The simplest solution proposed uses matched pairs comparisons (Møller and Birkhead, 1992; Wickman, 1992), where related species, or clades, that differ in the categorical variable are compared with each other in pairs. This method analyses if there is a consistent difference in the continuous variable that can be attributed to some evolutionary process connected with the categorical variable. The method of matched pairs comparisons works fine as long as single species are compared with each other. When whole clades are compared, however, the values of the continuous variables of all species included in a clade are commonly averaged without any regard to phylogenetic history. This introduces an error due to phylogeny, since a clade of species shares a common evolutionary history. Thus, the same variation is used several times when calculating the mean. To solve this problem, novel methods for calculating phylogenetically correct means and variances in a phylogeny are presented here (see also, for example, Garland *et al.*, 1999). There is a useful insight contained in regular matched pairs comparisons, however, in that the method realizes and utilizes the usefulness of *pairing* species differing in the categorical trait being analysed, thus making use of the fact that these species have a common starting point that is not only similar, but identical.

The Brunch-approach implemented in the computer program CAIC (Purvis and Rambaut, 1995) solves the problem of matched pairs comparisons – that the means being compared are calculated without regard to phylogeny – by instead comparing estimated ancestral nodal values arrived at with the method used for carrying out independent contrasts analysis (Felsenstein, 1985). The nodal values of independent contrasts analysis are, however, aimed at removing variation already used, not at producing clade-specific means. Also, when using the Brunch algorithm, the removed variation – that above the utilized nodal values – is not used for hypothesis testing, but only lost.

Several approaches do not pair clades (Grafen, 1989; Garland *et al.*, 1993; Hansen and Martins, 1996; Martins and Hansen, 1997; Lindenfors and Tullberg, 1998; Pagel, 1998; Butler *et al.*, 2000; Butler and King, 2004). The method of Garland *et al.* (1993) produces null distributions of F-statistics via computer simulations of evolution over a 'known' phylogenetic tree (the tree that is used for hypothesis testing). These simulations assume a specific model of evolution, but even so, computer simulations have shown that phylogenetic tests are relatively robust to such assumptions

of evolutionary models (Martins and Garland, 1991; Díaz-Uriarte and Garland, 1996, 1998; Garland and Díaz-Uriarte, 1999). The method presented here does not utilize computer simulations, but instead adjusts the data for phylogeny to enable the use of regular *F*-distributions.

Lindenfors and Tullberg's (1998) 'common origins test' uses parsimonious reconstructions of both the categorical and the continuous character and then tests if changes in the categorical factor have induced changes in the continuous character. The method works by summing changes in the continuous character and then analysing whether such sums of changes differ between groups of clades, the clades having been grouped on the basis of the state of the categorical character. It assumes that branch lengths have no influence on character evolution and that a valid data point is one given by a clade's summed character change after a transition in the categorical variable.

There are also a number of methods based on a generalized least squares (GLS) approach (Grafen, 1989; Hansen and Martins, 1996; Martins and Hansen, 1997; Pagel, 1998; Butler *et al.*, 2000; Blackburn and Duncan, 2001; Butler and King, 2004). The methods differ in the details, but share the quality that they use all of the information contained in the tree for computation, thus also including information not pertaining to the hypothesis being tested. To illustrate the problems this brings with it, analyses of the Hominoidea using the algorithms specified by Grafen (1989) and Pagel (1998) are presented here, testing whether relative testes mass is correlated with the lack of body fur [Fig. 1; data on relative testes weight from Harcourt *et al.* (1981)]. Since humans are the only species in this group that lack body fur, and since the chimpanzees are the only species in the group that have exceptionally large testes, a significant relationship indicating a correlation between the two would be surprising, and the GLS-approaches indicate no such relationship (Grafen's method, $P = 0.374$; Pagel's method, $P = 0.244$).

However, if one adds an outgroup with the typical characteristics of a gibbon (which is a true outgroup), the *P*-value falls. As one keeps adding outgroups with the same

Species	Body fur	Testes/body weight
Homo sapiens	No	0.06
Pan troglodytes	Yes	0.27
Gorilla gorilla	Yes	0.02
Pongo pygmaeus	Yes	0.05
Outgroup 1	Yes	0.1
Outgroup 2	Yes	0.1
etc.	...	...

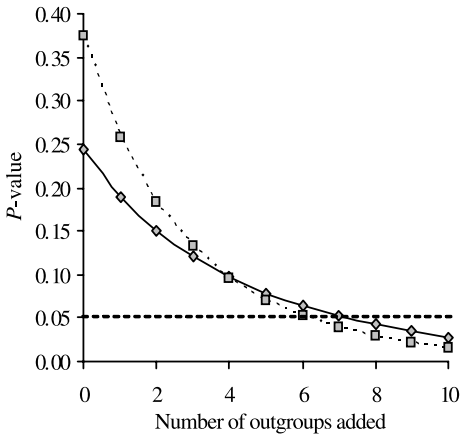

Fig. 1. The phylogeny of the Pongidae with data on body fur and relative testes weight. When the GLS-methods of Grafen (1989) and Pagel (1998) are applied to these data, the *P*-values indicating a relationship between the two variables decreases as outgroups are added. A significant relationship ($P < 0.05$) is observed when seven (Grafen's method – dashed line) or eight (Pagel's method – solid line) outgroups or more are included in the analysis. This is despite the fact that only humans lack body fur and only chimpanzees have exceptionally large testes. Tests utilizing Grafen's (1989) GLS approach were carried out using the MULREG module in NTSYSpc (Rohlf, 2000), while tests using Pagel's GLS approach were carried out in the program Continuous (Pagel, 2000).

characteristics, a significant relationship between body fur and relative testes size is eventually found at seven outgroups or more (Fig. 1). This somewhat disturbing result relating the evolution of one character in one species with another character in another species occurs because this type of approach partitions variation belonging in one section of the tree over the whole tree and also incorporates information not pertaining to the hypothesis into the statistical test. Even though both exceptionally large testes and a lack of body fur are unique to two separate species in the sample phylogeny, GLS and other maximum likelihood-like approaches, in assuming some rate of change – for example, Brownian motion or an Ornstein-Uhlenbeck process – for all characters, obtain results indicating a gradual evolution of these characters from the last common ancestor of chimpanzees and humans down through the whole clade. Note that no assumptions of the GLS-models have been broken in this example.

It should be pointed out that the same type of results can be arrived at for two continuous characters using independent contrasts analysis (Felsenstein, 1985). For example, body size and litter size have been shown to be significantly correlated in haplorhine primates (Lindenfors, 2002), even though all haplorhines have a litter size of one except for most small callitrichids who regularly give birth to twins. Thus, the most probable number of changes in litter size is a single switch from single-birth to twinning in the callitrichid clade (Ah-King and Tullberg, 2000), giving a 'real' sample size of one for analyses of litter size evolution in haplorhine primates. This twinning is, however, 'smeared' down the phylogeny together with the small body size of the callitrichids and thus a significant result is obtained. Hence, one has to be careful not to unintentionally inflate sample sizes when using maximum likelihood-like methods. This is a well-known problem that has previously been described in the context of squared change parsimony reconstructions of ancestral states (e.g. Losos, 1990; Maddison and Maddison, 1992; Butler and Losos, 1997), which is a method similar to those based on maximum likelihood that assume a Brownian motion model of character evolution (Huey and Bennet, 1987).

To carry out phylogenetically correct statistics on comparative materials, one needs to in some way account for phylogenetic relatedness. For this purpose, some methods try to estimate values of internal nodes by using some algorithm that requires, for example, least change per tree (e.g. maximum parsimony methods) or least change per branch length unit (e.g. maximum likelihood methods). These nodal values can then be used in correcting for phylogenetic relatedness. Alternatively, there are methods that claim not to reconstruct ancestral states, but where the algorithm nevertheless either assigns values to internal nodes in the process (Felsenstein, 1985) or where internal node states can be assigned utilizing the same algorithm as that used in the statistical analysis (e.g. Martins and Hansen, 1997; Pagel, 1998). However, whether the nodal estimates are termed 'reconstructed ancestral states' or are only a by-product of a statistical algorithm does not matter – they can still be utilized to calculate phylogenetically correct means of monophyletically related species as described below.

Studies comparing different methods to estimate nodal values reinforce theoretical expectations, simultaneously pointing out that many current methods do not work all that well (e.g. Frumhoff and Reeve, 1994; Butler and Losos, 1997; Cunningham *et al.*, 1998; Oakley and Cunningham, 2000; Polly, 2001; Webster and Purvis, 2002). The method presented here uses values assigned to internal nodes of the phylogeny, but the manner of assigning these nodal values is not central to the functioning of the model. On the contrary, the method aims at being generally useable no matter what algorithm is used to assign nodal values.

Computer simulations have shown that phylogenetic tests are relatively robust to assumptions of evolutionary models (Martins and Garland, 1991; Díaz-Uriarte and Garland, 1996, 1998; Garland and

Díaz-Uriarte, 1999), but depending on the evolutionary model and on the inclusion/exclusion of species exhibiting different character states, one can nevertheless reach very different conclusions (e.g. Schluter *et al.*, 1997). Herein lies one potential utility of a general method enabling the use of alternative evolutionary models. The method of calculating means and variances presented here can be carried out using nodal values estimated in any manner – the method is equally applicable to any model of trait change. To illustrate the method, I use a worked example where nodal values are assigned using maximum parsimony, whereas for a real-world example of the effects of sexual selection on primate size, I use Felsenstein's (1985) independent contrasts to assign nodal values for body size in the primate phylogeny. The nodal values of the categorical variables are assigned using maximum parsimony in both examples. As others have remarked previously, however, the 'choice of estimation method . . . should depend on the available information and the preferences of the individual researcher' (Martins and Hansen, 1997, p. 659).

THE PROBLEM

Consider 26 species, *Species A* to *Species Z*, related according to the phylogeny in Figs. 2 and 3. These species are characterized by two characters: one categorical character, *Colour*, that can take two states, *Black* and *Grey*, as shown in Fig. 2; and one continuous character, *Number*, which varies freely, as shown in Fig. 3. The sample phylogeny also includes information on branch lengths as given in Fig. 2.

The hypothesis to be tested is whether the *Black* and *Grey* species differ consistently in the continuous character *Number*. A non-phylogenetic approach would be simply to pool all

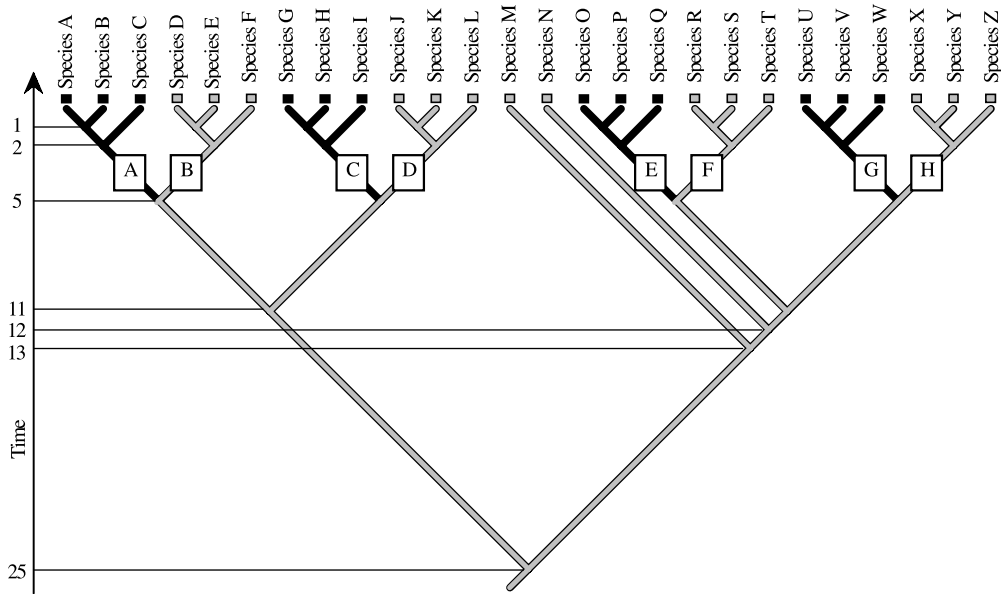

Fig. 2. A sample phylogeny of 26 species (*Species A* to *Species Z*) also showing parsimoniously assigned nodal estimates of the categorical character *Colour* that can take two states: *Black* and *Grey*. The letters *A–H* within the phylogeny denote sub-clades referred to in the text. The ages of representative nodes are indicated.

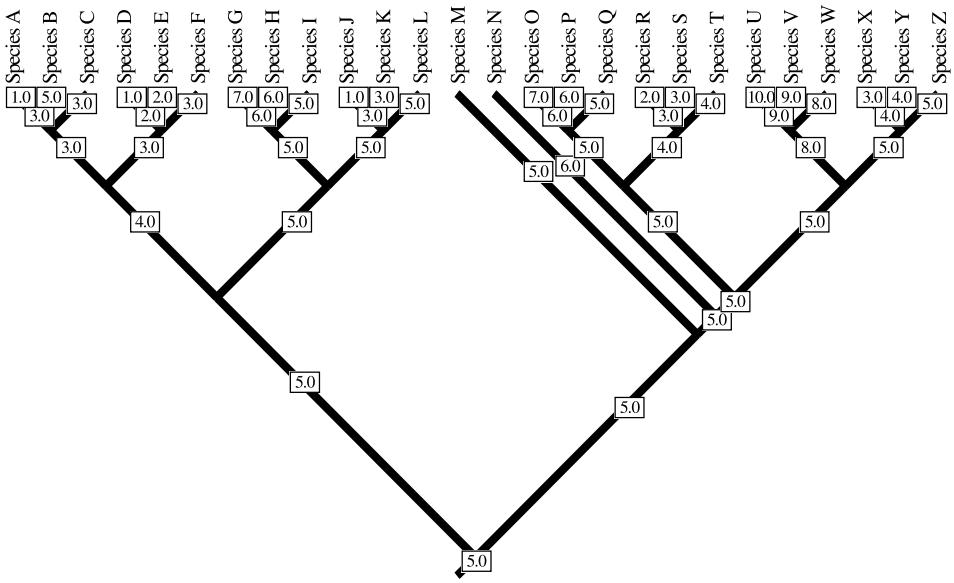

Fig. 3. Least absolute distance parsimonious nodal estimates of the continuous character *Number*. Note that in comparison with Fig. 2, the occurrence of the categorical character state *Black* seems to be correlated with an increase in the continuous character.

Black species together in one group and all *Grey* species together in another and then compare the means of these two groups using an ordinary ANOVA approach. If one looks carefully at the phylogeny, however, it becomes obvious that this is not a phylogenetically correct grouping. The real number of groups to use in the analysis is not two, but eight – if one uses the possibility to pair clades in matched pairs (Møller and Birkhead, 1992; Wickman, 1992) – separated into four comparisons. These are here given by the number of parsimoniously reconstructed evolutionary events in the categorical trait *Colour*. Thus, a paired *t*-test could solve the problem, but the question then becomes what mean values should represent each clade in such a comparison. Also, what if there is more than one state of the categorical variable? What is needed is a general ANOVA-type method that is adjusted for phylogeny.

From the sample phylogeny (Figs. 2 and 3), it seems as if evolution has mimicked an experimental set-up (Table 1). Four groups of species have undergone something akin to a 'treatment', which is a transition from *Grey* to *Black*. Another four groups of species are 'control groups'. That is, they have the same initial conditions as the 'treatment' species, but are not subject to any 'treatment'. Instead, they remain in the *Grey* state. Note that each clade in the phylogeny that has the character state *Black* has a corresponding clade with the character state *Grey* that shares *exactly* the same initial conditions since they were the same species before the 'experiment' started. It is this realization that has prompted the development of matched pairs comparisons (Møller and Birkhead, 1992; Wickman, 1992) and the Brunch algorithm implemented in CAIC (Purvis and Rambaut, 1995). The experimental design analogy makes it clear that the set-up is a blocked two-way mixed-model ANOVA with replication, as shown in Table 2 [a pattern also pointed out by Purvis and Webster (1999)].

Table 1. The experimental set-up as indicated by the phylogenies in Figs. 2 and 3

N		O_1
N	X	O_2

Note: The initial state of the continuous variable for *Grey* and the *Black* species is represented by exactly the same initial value, N, since both groups, in every case, share the same common ancestor. The treatment X is a switch from *Grey* to *Black*, and the observations we can make of extant species' characters are represented by O_1 and O_2.

Table 2. Table of comparisons from the phylogeny in Fig. 2

Blocks	'Control' (*Grey*)		'Treatment' (*Black*)	
1	Species A		Species D	
	Species B	→ Clade A	Species E	→ Clade B
	Species C		Species F	
2	Species G		Species J	
	Species H	→ Clade C	Species K	→ Clade D
	Species I		Species L	
3	Species O		Species R	
	Species P	→ Clade E	Species S	→ Clade F
	Species Q		Species T	
4	Species U		Species X	
	Species V	→ Clade G	Species Y	→ Clade H
	Species W		Species Z	

Note: This is a regular set-up for a blocked two-way mixed-model ANOVA (with or without replication).

From Table 2 it can be deduced that there are three levels of variation available:

1. *Between characters*: The 'experimental' grouping ('treatment' vs. 'control') variation that can be used to test the impact of the character *Colour*.
2. *Between comparisons*: Represents the block level and is random variation concerning the hypothesis, but may be interesting when testing, for example, grade-shifts between different sections of the phylogeny.
3. *Interaction between comparisons and characters*: This level of variation constitutes the error term if only the above two levels of variation are taken into account. It can, however, be used as a regular interaction term if the approach including replicates (see below) is used for the analysis.

These three levels of variation would be used in a typical phylogenetic ANOVA analysis. One more level of variation, however, can be included under exceptional conditions:

4. *Between species within groups*: This level of variation – the replicates – can also be included if this type of data exists (that is, if there are several species of equal numbers in each clade), as is the case in the presented example. For the purpose of showing how to include this level in the analyses, the example given is highly simplified and balanced. Note, however, that this is absolutely *not* a prerequisite for the general model presented here.

Phylogenetic comparisons conducted in this manner can *always* be represented as a complete-blocks design.

PHYLOGENETICALLY CORRECT CALCULATIONS OF MEANS AND VARIANCES

Having data derived through a phylogeny introduces two statistical problems that do not occur in an ordinary ANOVA:

1. How does one compute a phylogenetically correct mean?
2. How does one compute a phylogenetically correct variance?

To demonstrate such a method, let us examine a selected section of the sample phylogeny (Fig. 4). This section consists of the part of the phylogeny illustrating the relatedness of *Species O* to *T* – that is, *Clades E* and *F* (Table 2). A phylogeny such as this can be represented in a time ⇔ continuous character scatter plot as in Fig. 4, which illustrates how the continuous character has evolved over time.

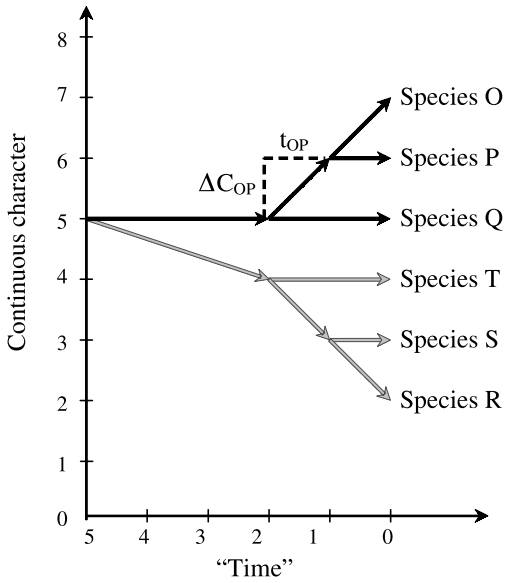

Fig. 4. The phylogenetic relationship between *Species O* to *T* (*Clades E* and *F*) plotted in a scatter plot where the y-axis denotes the continuous character and the x-axis denotes branch length. Time 0 indicates today. A rate of change, r_{OP}, is easily calculated by dividing the change in the continuous character, ΔC_{OP}, with the branch length, t_{OP}. The rates for all branches can be calculated using the same method.

For illustrative purposes – and since the method is explained graphically – the terminology used here is that mainly appropriate for reconstructed ancestral states. This is for ease of comprehension only, however, and should not be taken as indicating that the method is only appropriate for methods claiming to reconstruct ancestral states. On the contrary, the method can be used with nodal values estimated using any assumption of evolutionary model.

As can be seen in Fig. 4, the continuous character is estimated to have increased over time in the *Black* section of the phylogeny, while it has decreased over time in the *Grey* section. If one were to calculate the mean of *Species R*, *S*, and *T* (*Clade F*) as one usually does, however, it would include variation from one common event three times: the first decrease, and on two occasions the variation from the time before *Species R* and *S* had speciated. The same problem, of course, applies to *Species O*, *P*, and *Q* (*Clade E*). This type of problem is the rationale behind all phylogenetic methods, and the same reasoning should also hold true for mean and variance estimates as calculated from species' values.

The phylogeny plotted in Fig. 4 represents a set of trajectories of a continuous character through time. There is a certain rate of change that makes it possible to make a statement such as 'the continuous character *Number* has increased over time'. Evolutionary rates can also be calculated with, for example, maximum likelihood (e.g. Lynch, 1991) or GLS (Martins and Hansen, 1997; Pagel, 1998) methods, but it is also possible to calculate such rates using parsimoniously estimated nodal values – or indeed any nodal values assigned using any available method – as shown below.

In Fig. 4, the information required to calculate the rate is illustrated for the branch leading to the last common ancestor of *Species O* and *P* in the sample. By dividing the branch-specific change for the continuous character, ΔC_{OP}, with the branch length, t_{OP}, one arrives at a rate of change, r_{OP}, for this specific branch. Or, more generally:

$$r_i = \frac{\Delta C_i}{t_i}$$

Using the same reasoning, a clade-typical rate of change, *R*, can be calculated by dividing the sum of changes of the continuous variable along the *q* branches with the sum of branch lengths in the clade one is examining:

$$R = \frac{\sum_{i=1}^{q} \Delta C_i}{\sum_{i=1}^{q} t_i}$$

Note that this formula assumes that all branches contain equal amounts of information on the evolutionary rate in the clade. Thus, for example, rapid speciation after an initial change in a character, where the new species do not change, will 'dilute' the rate. To circumvent this problem, a solution implemented in a comparable method to the one presented here – the common origins test (Lindenfors and Tullberg, 1998) – can be used. With this method, to arrive at a clade-specific rate the rates of all branches in the clade are instead summed. Since the evolution of a character not under selection can be assumed to proceed by Brownian motion (Felsenstein, 1985) – assuming rates of change that on average will sum to zero – rapid

speciation will then not 'unduly' influence a sum. In general, however, it is a better assumption that all branches contribute equally to the clade-specific rate than to ignore information from the onset.

The information contained in the branches is now summed together as vectors (Fig. 5). This sum of vectors is the sum of all evolutionary trajectories of the continuous variable *Number* in the subset of the sample phylogeny. If one were to attach this summed trend line to the starting point of the sub-clade, s_{clade}, the phylogenetically correct mean is where the trend line intersects with the present time, t_{clade}. This intersection is the point where a hypothetical species with the clade-typical rate of change would end up, on average, given the starting value of the continuous variable *Number* in the sub-clade examined. Observe that the starting point, s_{clade}, always will be the same value for two sub-clades being compared, since this value represents their last common ancestor:

$$\overline{Y} = R \cdot t_{clade} + s_{clade}$$

A clade average as described here will thus depend on four things: (1) the phylogenetic topology, (2) the clade-typical rate of change, (3) the age of the clade, and (4) the estimated value of the most basal node of the clade. Thus, clade averages will be different between clades consisting of species with identical values of the continuous character if their relatedness, evolutionary rates of change, ages, or starting points differ. Hence, a clade-specific average will not be a 'true' average in the sense that it is an observed average of all observed species, but it will be an average including history in the equations – which is necessary if one wants to make a historical analysis.

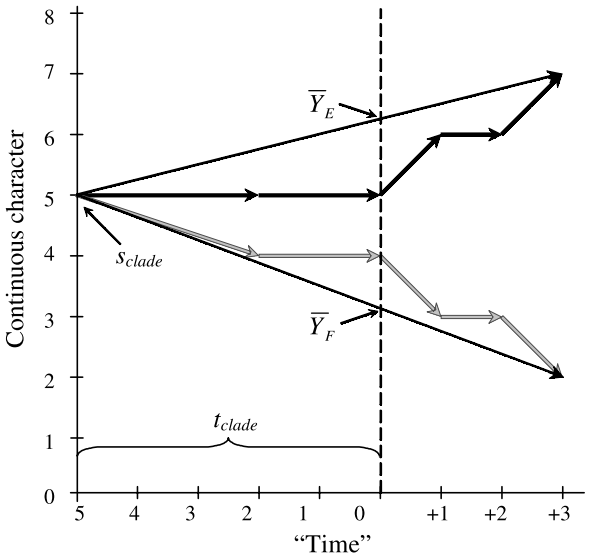

Fig. 5. Using the sample part of the phylogeny representing *Clades E* and *F* (*Species O* to *T*), this graph shows the trajectories represented by all individual branches, as indicated by the *Black* (*Clade E*) and *Grey* (*Clade F*) arrows. Time 0 indicates today. The long thin black arrows illustrate the clade-typical trajectories, *R*. The true means of the sub-phylogenies ($\overline{Y}_E$ and $\overline{Y}_F$), where each branch is included in the calculations only once, is indicated by the intersections of the clade-typical trajectories (starting at s_{clade}) and t_{clade}, which is the age of the clades – that is, the time from the last speciation common to all species and Time 0. See the text on how to calculate these entities.

Using a similar approach as when computing the mean, one can arrive at a clade-typical variance within a clade. It is not possible, however, to only subtract each species' value of the character *Number* from the clade-typical mean, as is the common procedure when calculating the variance statistic, because this would include the same variation several times in the calculations. Instead, the vector approach can be used again to determine how much each individual branch's end-point, O_i, differs from an end-point, E_i, as expected from the clade-typical rate of change. R is in each case the clade-typical rate of change for the species investigated:

$$E_i - O_i = R \cdot t_i - \Delta C_i$$

To arrive at the proper variance one cannot, however, simply take the sum of these deviations squared, divided by the number of deviations minus one, as is normally done; this is because the degrees of freedom are limited by the original number of data points, which in this case is the number of species in the investigated sub-clade. The proper variance is instead the sum of the squared deviations divided by the number of species, n, minus one:

$$v = \frac{\sum_{i=1}^{q} (E_i - O_i)^2}{n - 1}$$

Note that the variance arrived at using this formula is independent of the influence of t_{clade}, meaning that the method for calculating variances presented here is useable even when the matched pairs are of unequal ages (e.g. if the age of *Clades A* and *B* are different from the age of *Clades C* and *D*). Using the formulas given above, one can arrive at means and variances for all clades in the example. Note that t_{clade} is 5 in all cases and that *Species N* and *Species M* are not included in the analysis as they lack matching species (or clades) to be paired with. This is an intentional effect of using pairs of species, or clades, since *Species N* and *Species M* are not part of the inferred 'experimental set-up'. To use them would be to deviate from the powerful utility of all pairs having the exact same starting point.

Now that there is no phylogenetic influence on the means and the variances, the phylogenetically corrected values can be represented graphically, as shown in Fig. 6.

THE ANOVA CONTINUED

The data of the sample phylogeny plus the calculated means are given in Table 3. The notation in the section that follows is that of Sokal and Rohlf (1995). The phylogenetic ANOVA design is a complete-blocks ANOVA, carried out here both with and without replication, where calculations of means are performed as follows:

- The subgroup means, $\overline{Y}$, are calculated as described above and are thus phylogenetically adjusted.
- The 'treatment' (column) means, $\overline{C}$, are calculated by averaging the subgroup means ($\overline{Y}$) within each 'treatment'.
- The comparison (row, block) means, $\overline{R}$, are calculated by averaging the subgroup means ($\overline{Y}$) within each comparison.

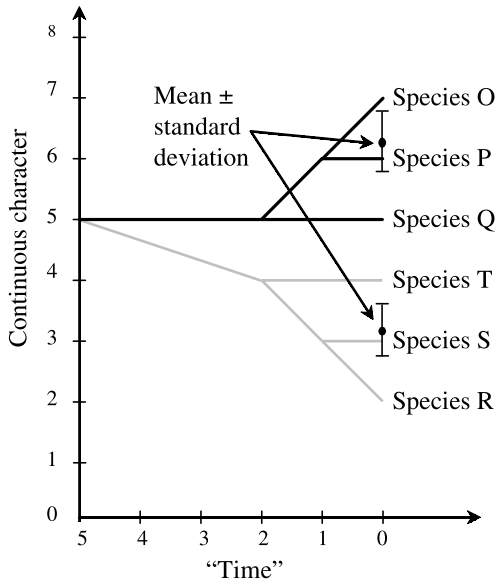

Fig. 6. Phylogenetically correct means and standard deviations as calculated for *Clade E* (*Species O* to *Q*) and *Clade F* (*Species R* to *T*).

- The grand mean, $\overline{\overline{Y}}$, is calculated by averaging all subgroup means ($\overline{Y}$). Note that this is not the grand mean of the phylogeny itself, but the grand mean of the sub-clades used in the ANOVA. A grand mean of the total phylogeny can instead be calculated using the method of calculating clade-means described above, but applied to the whole phylogeny. Such a grand mean could be of utility when, for example, placing a phylogenetically correct regression line back in the original data-space.

The next step is to calculate the four entities below, where *r* is the number of blocks (in this case four because there are four independent comparisons of *Black* and *Grey* groups), *c* is the number of discrete characters (in this case two: *Black* and *Grey*), and *n* is the number of data-points (in this case three, as there are three species in each group). Note that the formula given at point 4 is only valid if an equal number of species (*n*) are included in the comparisons in all cases. This will not be true for most analyses, in which case this entity can be left out completely or calculated with formulas adjusted for unbalanced designs. If this level of variation is dropped, *n* will be equal to one and the design will reduce to a complete-blocks mixed-model ANOVA without replication, which would also make calculations of phylogenetic variances redundant. The ANOVA is mixed-model, since the effect of the 'treatments' has an expected direction, while the effect of the 'blocks' level is expected to be random.

1. $SS_{columns} = \sum^{c} rn(\overline{C} - \overline{\overline{Y}})^2 = 58.594$

2. $SS_{rows} = \sum^{r} cn(\overline{R} - \overline{\overline{Y}})^2 = 39.211$

Table 3. Data table with phylogenetically adjusted means given

Comparisons (r) (block = row)	'Treatment' (c) (column)		Comparison means $\overline{R}$
	Black	*Grey*	
1	1	1	
	5	2	
	3	3	
Subgroup means $\overline{Y}$	*3.000*	*1.750*	*2.375*
2	7	1	
	6	3	
	5	5	
Subgroup means $\overline{Y}$	*6.250*	*2.500*	*4.375*
3	7	2	
	6	3	
	5	4	
Subgroup means $\overline{Y}$	*6.250*	*3.125*	*4.688*
4	10	3	
	9	4	
	8	5	
Subgroup means $\overline{Y}$	*8.125*	*3.750*	*5.938*
'Treatment' means $\overline{C}$	*5.906*	*2.781*	
Grand mean $\overline{\overline{Y}}$	*4.344*		

Note: Explanations of how to calculate these means are given in the text.

3. $SS_{interaction} = \sum^{r} \sum^{c} n(\overline{Y} - \overline{R} - \overline{C} + \overline{\overline{Y}})^2 = 8.203$

4. $SS_{within} = \sum^{rc} \sum^{n} (E - O)^2 = 29.000$

For an unbalanced design, which is what most researchers will encounter, no agreement exists on how to calculate the error mean square, but three disparate views exist on how to proceed (Quinn and Keough, 2002). In such a case, however, note that the same result is arrived at for the 'treatment' term if a full complete-blocks mixed-model ANOVA with replication is carried out, as if the same calculations are done without replication (Tables 4 and 6, respectively). This is because the denominator for the F-ratio in both cases is the interaction

Table 4. Results table for the phylogenetic ANOVA incorporating subgroup variation

Source of variation	d.f.	SS	MS	F_s		P
$\bar{C} - \bar{\bar{Y}}$ (columns)	$c - 1 = 1$	58.594	58.594	$\frac{MS_{columns}}{MS_{interaction}}$	$= 21.429$	0.019
$\bar{R} - \bar{\bar{Y}}$ (rows)	$r - 1 = 3$	39.211	13.070	$\frac{MS_{rows}}{MS_{error}}$	$= 7.211$	0.003
$\bar{Y} - \bar{R} - \bar{C} + \bar{\bar{Y}}$ (interaction)	$(r - 1)(c - 1) = 3$	8.203	2.734	$\frac{MS_{interaction}}{MS_{error}}$	$= 1.509$	0.250
$Y - \bar{Y}$ (error)	$rc(n - 1) = 16$	29.000	1.813			

Note: The difference between the colours ('*Black*' and '*Grey*') in the given example is significant, as is the difference between the 'blocks' (rows), in a real analysis possibly indicating a grade-shift. The *F*-ratio for the 'treatment' term (*Colour*) is calculated using the interaction term in the denominator because the blocking factor is a random factor (Quinn and Keough, 2002).

Table 5. Results table for the phylogenetic ANOVA incorporating subgroup variation and pooling the interaction term with the error term

Source of variation	d.f.	SS	MS	F_s		P
$\bar{C} - \bar{\bar{Y}}$ (columns)	$c - 1 = 1$	58.594	58.594	$\frac{MS_{columns}}{MS_{error} + MS_{interaction}}$	$= 29.924$	0.00003
$\bar{R} - \bar{\bar{Y}}$ (rows)	$r - 1 = 3$	39.211	13.070	$\frac{MS_{rows}}{MS_{error} + MS_{interaction}}$	$= 6.675$	0.003
$\bar{Y} - \bar{R} - \bar{C} + \bar{\bar{Y}} + Y - \bar{Y}$ (interaction + error)	$(r - 1)(c - 1) + rc(n - 1) = 19$	37.203	1.958			

Note: The difference between the colours ('*Black*' and '*Grey*') is significant, as is the difference between the 'blocks' (rows), in a real analysis possibly indicating a grade-shift.

mean squares, and both the enumerator and denominator for the *F*-ratio reduce by the same factor, *n*, being the number of replicates. Thus, in the highly likely case that the design is unbalanced, the replicates can be ignored. If the replicate level is dropped, however, the possibility to test for an interaction effect is lost, as is power when testing for a block effect.

The ANOVA method outlined here is readily extended to situations where the categorical character can take more states than two, or if an analysis of covariance is the desired analysis. The formulas to be used are available in any standard statistics text (e.g. Sokal and Rohlf,

Table 6. Results table for the phylogenetic ANOVA not incorporating subgroup variation

Source of variation	d.f.	SS	MS	F_s	P
$\overline{C} - \overline{\overline{Y}}$ (columns)	$c - 1 = 1$	19.531	19.531	$\frac{MS_{columns}}{MS_{interaction}} = 21.429$	0.019
$\overline{R} - \overline{\overline{Y}}$ (rows)	$r - 1 = 3$	13.070	4.357	$\frac{MS_{rows}}{MS_{interaction}} = 4.780$	0.116
$\overline{Y} - \overline{R} - \overline{C} + \overline{\overline{Y}}$ (interaction)	$(r - 1)(c - 1) = 3$	2.734	0.911		

Note: The difference between the colours ('*Black*' and '*Grey*') is significant. In comparison with Table 5, the inclusion/exclusion of subgroup variation does not influence the results for the 'treatment' term as both the enumerator and the denominator for the *F*-ratio reduce by the same term, *n*. For the blocking term, however, statistical power is reduced.

1995; Quinn and Keough, 2002), while the phylogenetic adjustments are those as described in this paper.

As can be seen in the sample ANOVA tables (Tables 4, 5, and 6), the conclusion regarding the 'treatment' effect is the same whether the replicates are included in the calculations or not. When the ANOVA is carried out with replication but the interaction term is non-significant, however, the interaction mean squares can be pooled with the error mean squares for the testing of significance and hence increase the power of the test (Table 5).

COMPARISONS WITH OTHER TESTS

As reviewed in the Introduction, a number of other tests already exist to handle the same set-up as the phylogenetic ANOVA is designed to analyse. This section briefly compares and discusses similarities and differences in the results that some of these alternative tests provide. As pointed out in the Introduction, these alternative tests differ among each other on one central point, which is whether they use the possibility to pair sub-phylogenies and thus construct a blocked ANOVA design, or if they instead just construct a regular one-way ANOVA design. The phylogenetically adjusted means and variances presented here can be used for both purposes. When the data come sorted in phylogenetic comparisons, however, this makes the possibility for a paired grouping obvious. It is a waste of statistical power not to use it.

For the sake of comparison, however, the comparisons with the alternative tests will be made both using the possibility to construct matched pairs, as described in the example above, and without using this possibility (Table 7). Tests using the pairing option include matched pairs comparisons (Møller and Birkhead, 1992; Wickman, 1992) and the Brunch-option in CAIC (Purvis and Rambaut, 1995), while tests not using this option include the GLS tests (Grafen, 1989; Pagel, 1998) and Garland and colleagues' (1993) simulation approach. These tests are compared to the phylogenetic ANOVA presented here, as well as a regular (species-level) ANOVA.

As can be seen in Table 7, all tests agree on the effect of the 'treatment' variable in the given example, with the exception of CAIC which gives a result that is nearly significant.

Table 7. Comparisons of the results of different alternative tests on the given example

Test		Factor	*F*-ratio	*P*
(1)	Phylogenetic ANOVA (with replication)	*Colour*	21.429	0.019
		Block	7.211	0.003
		Interaction	1.509	0.250
	Blocked mixed-model ANOVA (with replication, but without phylogenetic corrections)	*Colour*	13.500	0.035
		Block	9.143	0.001
		Interaction	2.286	0.118
(2)	Phylogenetic ANOVA (with replication and pooled error term)	*Colour*	29.924	0.00003
		Block	4.780	0.003
	Blocked mixed-model ANOVA (with replication and pooled error term, but without phylogenetic corrections)	*Colour*	13.500	0.00007
		Block	4.000	0.002
(3)	Phylogenetic ANOVA (without replication)	*Colour*	21.429	0.019
		Block	4.780	0.116
	Blocked mixed-model ANOVA (without replication and phylogenetic corrections)	*Colour*	13.500	0.035
		Block	4.000	0.142
	Matched pairs comparisons (Møller and Birkhead, 1992; Wickman, 1992)	*Colour*	13.500	0.035
	Brunch in CAIC (Purvis and Rambaut, 1995)	*Colour*	8.643	0.061
(4)	Phylogenetic ANOVA (with replication, but without blocking)	?	11.124	0.00005
	One-way ANOVA (with replication, but without blocking)	?	11.511	0.00004
	Generalized least squares (GLS) (Grafen, 1989)	*Colour*	4.029	0.056
	Generalized least squares (GLS) (Pagel, 1998)	*Colour*	(2.018)*	0.045
	Simulation approach (Garland *et al.*, 1993)	?	11.511	0.040†

* Ln likelihood ratio that is compared with a χ^2-distribution.
† The relationship between the *F*-ratio and *P*-value is adjusted via simulations for this test.

Note: The tests are grouped into four groups depending on how the data are utilized: (1) with replicates included, (2) with replicates included and a pooled error term, (3) with replicates excluded, and (4) without using the option to pair clades differing in the categorical variable (see text for further explanations). For the tests with a question mark in the factor column, the significant result does not indicate an influence of the variable *Colour*, but a significant difference between the groups. It can thus be the case that only one of the *Black* clades differs from all other groups.

The phylogenetic ANOVA reports a more significant effect of the 'treatment', but a less significant difference between the blocks, than a regular ANOVA. This is to be expected given the construction of the example (Figs. 3 and 4). If the replicates are dropped, the power to test for an effect of the blocking factor decreases and the possibility to test for an interaction effect is lost. Matched pairs comparisons (Møller and Birkhead, 1992; Wickman, 1992) and the Brunch-option in CAIC (Purvis and Rambaut, 1995) do not facilitate the testing of differences between blocks.

Without pairing, the data are arranged in one large group having the character state *Grey* to be compared with four clades having the character state *Black* (*Clades A*, *C*, *E*, and *G*). This is how Grafen's (1989) and Pagel's (1998) GLS methods, Lindenfors and Tullberg's (1998) common origins test, and Garland and colleagues' (1993) simulation approach use the data.

The method of Garland *et al.* allows specification of alternative groupings, however, and in principle even allows the pairing of clades, although this is not implemented in the PDAP package (Garland *et al.*, 1993). If the phylogenetic data are analysed without blocking, however, Garland and colleagues' test requires a further *post-hoc* test to determine which groups differ from each other. A significant result can, for example, indicate that only one of the *Black* clades differs from the others.

The method of Garland *et al.* (1993) works by producing null distributions of *F*-statistics via computer simulations of evolution over a 'known' phylogenetic tree (the tree that is used for hypothesis testing). For this example, the simulation approach is equivalent to lowering the α-level where a difference is deemed significant from $P = 0.05$ to $P = 0.00007$, decreasing statistical power proportionally [though such a comparison only makes real sense when comparing tests having the same Type 1 error rate (Martins and Garland, 1991; T. Garland, personal communication)].

The common origins test (Lindenfors and Tullberg, 1998) is not applicable for the given example. This is because no reconstructed origin of the character state *Grey* exists in the phylogeny. As the common origins test defines its groups from such origins, the group defined by the character state *Grey* is 'forbidden'.

In summary, the test presented here is to be preferred over matched pairs analyses (Møller and Birkhead, 1992; Wickman, 1992) because it corrects for phylogeny throughout the test and not only when determining what species to compare. It also adds to the possibilities provided by the Brunch-option in CAIC (Purvis and Rambaut, 1995), as it facilitates testing for grade shifts and – in rare cases – the inclusion of replicates. Furthermore, it performs better than Grafen's (1989) and Pagel's (1998) GLS methods because it increases statistical power through the use of blocking and because it has none of the problems of the GLS approaches that accrue from using information outside the comparisons for hypothesis testing as outlined in the Introduction. The phylogenetic ANOVA is a computational alternative to Garland and colleagues' (1993) simulation approach, but is indicated to have higher power to detect evolutionary relationships – although this has to be validated through proper comparisons on simulated data.

A REAL-WORLD EXAMPLE: SEXUAL SELECTION ON PRIMATE SIZE

To show the method's utility using an example taken from the real world, I have chosen to analyse the effects of sexual selection on haplorhine primate size. This is a well-researched problem (e.g. Clutton-Brock and Harvey, 1977; Alexander *et al.*, 1979; Gaulin and Sailer, 1984; Harvey and Harcourt, 1984; Mitani *et al.*, 1996; Lindenfors and Tullberg, 1998) and the emerging result is clear: sexual selection has resulted in an increased body size dimorphism in haplorhine primates [the claim of causality is based on a result in Lindenfors and Tullberg (1998) establishing temporal order – body size changes *after* transitions in mating system]. Phylogenetic analyses have further revealed the increase in dimorphism to be the result of sexual selection causing an increase in male body size, even though female size also increases, though to a lesser extent (Lindenfors and Tullberg, 1998).

To analyse the issue of haplorhine dimorphism and sexual selection with the phylogenetic ANOVA, the haplorhine section of the primate phylogeny by Purvis (1995) was used. This phylogeny was made using a 'super-tree' technique, combining a large number of source phylogenies based on both molecular and morphological data. Data on mating systems were taken from Lindenfors and Tullberg (1998), while body weights were taken from Smith

and Jungers (1997). Male and female body weights, as well as the ratio male/female weight, were $\log_{10}$-transformed before the analyses.

Mating system (uni-male, multi-male, and monogamous), used here as an indication of the strength of sexual selection, was considered a three-state unordered character and nodal values parsimoniously estimated (Swofford and Maddison, 1987) using MacClade (Maddison and Maddison, 1992). Internal node values for body mass were estimated using Felsenstein's (1985) independent contrasts method, with branch lengths as given in Purvis (1995) and polytomies set as being branches of zero length.

The trichotomous variable mating system introduces a special problem in that the multi-male mating system indicates a higher degree of sexual selection than monogamy, but a lower degree of sexual selection than the uni-male mating system. Thus, a group of multi-male species used in a comparison is expected to be larger and more dimorphic than a group of monogamous species, but smaller and less dimorphic than a group of uni-male species. For this reason, all comparisons were recoded as being between two groups: 'more' or 'less' sexually selected (Lindenfors and Tullberg, 1998; Lindenfors *et al.*, 2003), where multi-male species therefore ended up in one of these groups depending on whether they were used in comparisons with monogamous or uni-male species, respectively.

A total of 15 comparisons could be made in the phylogeny, as given in Appendix 2 in Lindenfors and Tullberg (1998). Nested comparisons were handled as in any comparison-type method; that is, once the comparison closest to the tips of the phylogeny has been made, the included species are "removed" from the phylogeny to advance another comparison towards the tips of the phylogeny, and so on (Purvis and Rambaut, 1995).

For two reasons the replicates level of the analysis was dropped: (1) the independent factor interesting for hypothesis testing is mating system, while the blocking factor is of lesser interest, and (2) the design is highly unbalanced. Note again, however, that because the *F*-ratio for the 'treatment' term (mating system) is calculated using the interaction term in the denominator – whether replicates are included or not [because the blocking factor is a random factor (Quinn and Keough, 2002)] – the results for the effects of mating systems are the same as if one were to include the replicates (see Table 8).

Table 8. Comparisons of results of different tests on the effects of sexual selection on haplorhine primate body size and sexual size dimorphism

Test	Factor	*Male weight*	*Female weight*	*Dimorphism*
Phylogenetic ANOVA (without replication)	*Mating system*	0.012	0.018	0.011
	Comparison	<0.001	<0.001	<0.001
Blocked mixed-model ANOVA (without replication and phylogenetic corrections)	*Mating system*	0.011	0.016	0.009
	Comparison	<0.001	<0.001	0.012
Matched pairs comparisons (Møller and Birkhead, 1992; Wickman, 1992)	*Mating system*	0.011	0.016	0.009
Brunch in CAIC (Purvis and Rambaut, 1995)	*Mating system*	0.046	0.046	0.090

Note: The tests are mostly consistent regarding the results on the effects of mating system, differing somewhat only in significance. A strong effect of the comparison level is also indicated (see text for discussion). Note that the results for the matched pairs comparisons on the effects of differences in mating system are identical to those observed using a species-level ANOVA.

The results are in general consistent between the phylogenetic ANOVA and the three comparable tests, but differ somewhat in significance, in that dimorphism, as well as male and female body size, is larger in more sexually selected species. Note that the species level ANOVA and the matched pairs analyses give the same results for the effect of sexual selection. This is because these two tests utilize the data in the same way – that is, they simply average species values belonging to each block in the comparisons. The two tests that facilitate testing for grade-shifts – the phylogenetic ANOVA and a regular species-level ANOVA – also show that there is a highly significant effect of the comparison level. This is not surprising regarding body size evolution *per se* – primates vary in body size for a vast number of other reasons than sexual selection.

It is more surprising, however, that sexual size dimorphism also shows a significant effect of blocks. One possible conclusion based on this result is that dimorphism may vary for reasons other than sexual selection. It has to be remembered, however, that comparisons in this test differed in the starting point of sexual selection, in that some tests compared groups of species where the ancestral reconstruction gave a multi-male mating system, whereas others were made where the reconstructions instead indicated an ancestral monogamous mating system. Thus the blocks level also contains some aspect of sexual selection. Also, for some comparisons the 'treatment' was a switch to a mating system indicating more sexual selection, while in others it was a switch towards less sexual selection.

Adding to this is the fact that even though mating system is a good indicator of the strength of sexual selection, it is not a perfect measurement of it. The classification of species according to differences in mating systems surely does not include everything encompassed under the labels 'more' and 'less' sexual selection. Thus, even though mating systems enable us to make confident statements on the effects of sexual selection on haplorhine primates, it goes without saying that a more fine-grained measurement would be preferred [e.g. operational sex-ratio (Mitani *et al.*, 1996), harem size (Lindenfors *et al.*, 2002)].

ACKNOWLEDGEMENTS

I wish to thank T. Garland, Jr., B.S. Tullberg, J.L. Gittleman, S. Heard, O. Leimar, S. Nylin, D. Faith, and F. Lutzoni for comments on previous drafts of the manuscript. This work was supported by the Swedish Research Council.

REFERENCES

Ah-King, M. and Tullberg, B.S. 2000. Phylogenetic analysis of twinning in Callitrichinae. *Am. J. Primatol.*, **51**: 135–146.

Alexander, R.D., Hoogland, J.L., Howard, R.D., Noonan K.M. and Sherman, P.W. 1979. Sexual dimorphism and breeding systems in pinnipeds, ungulates, primates and humans. In *Evolutionary Biology and Human Social Behaviour: An Anthropological Perspective*. (N. Chagnon and W. Irons, eds.), pp. 402–435. North Scituate, MA: Duxbury.

Blackburn, T.M. and Duncan, R.P. 2001. Determinants of establishment success in introduced birds. *Nature*, **414**: 195–197.

Butler, M.A. and King, A.A. 2004. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *Am. Nat.*, **164**: 683–695.

Butler, M.A. and Losos, J.B. 1997. Testing for unequal amounts of evolution in a continuous character on different branches of a phylogenetic tree using linear and squared-change parsimony: an example using Lesser Antillean *Anolis* lizards. *Evolution*, **51**: 1623–1635.

Butler, M.A., Schoener, T.W. and Losos, J.B. 2000. The relationship between habitat type and sexual size dimorphism in Greater Antillean Anolis lizards. *Evolution*, **54**: 259–272.

Clutton-Brock, T.H. and Harvey, P.H. 1977. Primate ecology and social organization. *J. Zool., Lond.*, **183**: 1–39.

Cunningham, C.W., Omland, K.E. and Oakley, T.H. 1998. Reconstructing ancestral character states: a critical reappraisal. *Trends Ecol. Evol.*, **13**: 361–366.

Díaz-Uriarte, R. and Garland, T., Jr. 1996. Testing hypotheses of correlated evolution using phylogenetically independent contrasts: sensitivity to deviations from Brownian motion. *Syst. Biol.*, **45**: 27–47.

Díaz-Uriarte, R. and Garland, T., Jr. 1998. Effects of branch length errors on the performance of phylogenetically independent contrasts. *Syst. Biol.*, **47**: 654–672.

Felsenstein, J. 1985. Phylogenies and the comparative method. *Am. Nat.*, **125**: 1–15.

Frumhoff, P.C. and Reeve, H.K. 1994. Using phylogenies to test hypotheses of adaptation: a critique of current proposals. *Evolution*, **48**: 172–180.

Garland, T., Jr. and Díaz-Uriarte, R. 1999. Polytomies and phylogenetically independent contrasts: an examination of the bounded degrees of freedom approach. *Syst. Biol.*, **48**: 547–558.

Garland, T., Jr., Dickerman, A.W., Janis, C.M. and Jones, J.A. 1993. Phylogenetic analysis of covariance by computer simulation. *Syst. Biol.*, **42**: 265–292.

Garland, T., Jr., Midford, P.E. and Ives, A.R. 1999. An introduction to phylogenetically based statistical methods, with a new method for confidence intervals on ancestral values. *Am. Zool.*, **39**: 374–388.

Gaulin, S.J.C. and Sailer, L.D. 1984. Sexual dimorphism in weight among the primates: the relative impact of allometry and sexual selection. *Int. J. Primatol.*, **5**: 515–535.

Grafen, A. 1989. The phylogenetic regression. *Phil. Trans. R. Soc. Lond. B*, **326**: 119–156.

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

Harcourt, A.H., Harvey, P.H., Larson, S.G. and Short, R.V. 1981. Testis weight, body weight and breeding system in primates. *Nature*, **293**: 55–57.

Harvey, P.H. and Harcourt, A.H. 1984. Sperm competition, testes size, and breeding system in primates. In *Sperm Competition and the Evolution of Animal Mating Systems* (R.L. Smith, ed.), pp. 589–659. London: Academic Press.

Harvey, P.H. and Pagel, M.D. 1991. *The Comparative Method in Evolutionary Biology*. Oxford: Oxford University Press.

Huey, R.B. and Bennet, A.F. 1987. Phylogenetic studies of coadaption: preferred temperatures versus optimal performance temperatures of lizards. *Evolution*, **41**: 1098–1115.

Lindenfors, P. 2002. Sexually antagonistic selection on primate size. *J. Evol. Biol.*, **15**: 595–607.

Lindenfors, P. and Tullberg, B.S. 1998. Phylogenetic analyses of primate size evolution: the consequences of sexual selection. *Biol. J. Linn. Soc.*, **64**: 413–447.

Lindenfors, P., Tullberg, B.S. and Biuw, M. 2002. Phylogenetic analyses of sexual selection and sexual size dimorphism in pinnipeds. *Behav. Ecol. Sociobiol.*, **52**: 188–193.

Lindenfors, P., Székely, T. and Reynolds, J.D. 2003. Directional changes in sexual size dimorphism in shorebirds, gulls and alcids. *J. Evol. Biol.*, **16**: 930–938.

Losos, J.B. 1990. Ecomorphology, performance capability, and scaling of West Indian *Anolis* lizards: an evolutionary analysis. *Ecol. Monogr.*, **60**: 369–388.

Lynch, M. 1991. Methods for the analysis of comparative data in evolutionary biology. *Evolution*, **45**: 1065–1080.

Maddison, W.P. 1990. A method for testing the correlated evolution of two binary characters: are gains or losses concentrated on certain branches of a phylogenetic tree? *Evolution*, **44**: 539–557.

Maddison, W.P. and Maddison, D.R. 1992. *MacClade*. Sunderland, MA: Sinauer Associates.

Martins, E.P. and Garland, T., Jr. 1991. Phylogenetic analyses of the correlated evolution of continuous characters: a simulation study. *Evolution*, **41**: 534–557.

Martins, E.P. and Hansen, T.F. 1997. Phylogenies and the comparative method: a general approach in incorporating phylogenetic information into the analysis of interspecific data. *Am. Nat.*, **149**: 646–667.

Mitani, J.C., Gros-Louis, J. and Richards, A.F. 1996. Sexual dimorphism, the operational sex ratio, and the intensity of male competition in polygynous primates. *Am. Nat.*, **147**: 966–980.

Møller, A.P. and Birkhead, T.R. 1992. A pairwise comparative method as illustrated by copulation frequency in birds. *Am. Nat.*, **139**: 644–656.

Oakley, T.H. and Cunningham, C.W. 2000. Independent contrasts succeed where ancestor reconstruction fails in a known bacteriophage phylogeny. *Evolution*, **54**: 397–405.

Pagel, M.D. 1994. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proc. R. Soc. Lond. B*, **255**: 37–45.

Pagel, M.D. 1998. Inferring evolutionary processes from phylogenies. *Zoologica Scripta*, **26**: 331–348.

Pagel, M.D. 2000. *Continuous*. Reading, UK: Department of Animal and Microbial Sciences, University of Reading.

Polly, P.D. 2001. Paleontology and the comparative method: ancestral node reconstructions versus observed node values. *Am. Nat.*, **157**: 596–609.

Purvis, A. 1995. A composite estimate of primate phylogeny. *Phil. Trans. R. Soc. Lond. B*, **348**: 405–421.

Purvis, A. and Rambaut, A. 1995. Comparative analysis by independent contrasts (CAIC): an Apple Macintosh application for analyzing comparative data. *Comput. Appl. Biosci.*, **11**: 247–251.

Purvis, A. and Webster, A.J. 1999. Phylogenetically independent comparisons and primate phylogeny. In *Comparative Primate Socioecology* (P.C. Lee, ed.), pp. 44–68. Cambridge: Cambridge University Press.

Quinn, G.P. and Keough, M.J. 2002. *Experimental Design and Data Analysis for Biologists*. Cambridge: Cambridge University Press.

Ridley, M. 1983. *The Explanation of Organic Diversity: The Comparative Method and Adaptions for Mating*. Oxford: Clarendon Press.

Rohlf, F.J. 2000. *NTSYS-pc v2.1: Numerical Taxonomy and Multivariate Analysis System*. Setauket, NY: Exeter Software.

Schluter, D., Price, T., Mooers, A.Ø. and Ludwig, D. 1997. Likelihood of ancestor states in adaptive radiation. *Evolution*, **51**: 1699–1711.

Smith, R.J. and Jungers, W.L. 1997. Body mass in comparative primatology. *J. Human Evol.*, **32**: 523–559.

Sokal, R.R. and Rohlf, F.J. 1995. *Biometry*, 3rd edn. New York: W.H. Freeman.

Stearns, S.C. 1983. The influence of size and phylogeny on patterns of covariation among life-history traits in the mammals. *Oikos*, **41**: 173–187.

Swofford, D.L. and Maddison, W.P. 1987. Reconstructing ancestral character states under Wagner parsimony. *Math. Biosci.*, **87**: 199–229.

Webster, A.J. and Purvis, A. 2002. Testing the accuracy of methods for reconstructing ancestral states of continuous characters. *Proc. R. Soc. Lond. B*, **269**: 143–149.

Wickman, P.-O. 1992. Sexual selection and butterfly design – a comparative study. *Evolution*, **46**: 1525–1536.