BRIEF COMMUNICATIONS

Evolution, 59(12), 2005, pp. 2705–2710

THE EFFECT OF INTRASPECIFIC SAMPLE SIZE ON TYPE I AND TYPE II ERROR RATES IN COMPARATIVE STUDIES

LUKE J. HARMON[1] AND JONATHAN B. LOSOS
Department of Biology, Washington University, St. Louis, Missouri 63130

Abstract.—Comparative studies have increased greatly in number in recent years due to advances in statistical and phylogenetic methodologies. For these studies, a trade-off often exists between the number of species that can be included in any given study and the number of individuals examined per species. Here, we describe a simple simulation study examining the effect of intraspecific sample size on statistical error in comparative studies. We find that ignoring measurement error has no effect on type I error of nonphylogenetic analyses, but can lead to increased type I error under some circumstances when using independent contrasts. We suggest using ANOVA to evaluate the relative amounts of within- and between-species variation when considering a phylogenetic comparative study. If within-species variance is particularly large and intraspecific sample sizes small, then either larger sample sizes or comparative methods that account for measurement error are necessary.

Key words.—Independent contrasts, measurement error, phylogenetic comparative method, population variation, statistics.

Received April 22, 2005. Accepted September 29, 2005.

Since Darwin and even before, biologists have compared attributes of groups of organisms to understand how they are related functionally and evolutionarily. The number of such comparative studies has increased greatly in recent years due to advances in statistical and phylogenetic methodologies (reviewed in Felsenstein 2004).

In an ideal world, researchers would conduct comparative studies by collecting data on large numbers of individuals for as many species (or populations, lineages, or other units) as possible. However, in many cases, constraints on the availability of time, effort, or resources may limit the total amount of data that can be collected. As a result, a trade-off will exist between the number of species that can be included in any given study and the number of individuals examined per species.

Because the power of a comparative analysis is a function of the number of species (Freckleton et al. 2002), many workers are tempted to maximize the number of species, even if it means measuring few (at the extreme, one) individuals of each species (e.g., Leal et al. 2002; Patek and Oakley 2003; Ackerly and Nyffeler 2004; Al-kahtani et al. 2004). Measuring few individuals per species increases the probability that species' characteristics will be estimated with error, perhaps even substantial error.

In nonphylogenetic correlational studies, if the individuals to be measured are chosen randomly, measurement error should serve to add noise, and thus should contribute to type II error (failure to detect a real pattern), rather than to type I error (detecting a pattern that does not actually exist). Nonetheless, many workers are skeptical of significant results stemming from analyses with low sample size per species.

In phylogenetic comparative analyses, by contrast, measurement error could artifactually lead to the detection of a relationship that does not actually exist (Martins 1994; Purvis and Webster 1999; Felsenstein 2004). For example, using independent contrasts, measurement error will have the greatest effect on contrasts between closely related species pairs; artificially increasing the value of these contrasts for all variables increases the possibility that a relationship would be detected between the variables when one really does not exist, increasing type I error (Purvis and Webster 1999; Felsenstein 2004). A number of remedies to this problem have been suggested, and some comparative approaches incorporate measurement error into the statistical model when testing for relationships among traits (e.g., Lynch 1991; Martins and Hansen 1997). However, comparative methods that don't account for this error, such as independent contrasts (Felsenstein 1985), are still commonly used, and the magnitude of this potential problem has been little explored.

Because the effect of intragroup sample size on comparative analyses, either phylogenetic or nonphylogenetic, has received little attention, we conducted a simple simulation study to examine the effect of intraspecific sample size on comparative studies.

METHODS

We examined the effect of within-species sample size on type I and type II error rates in between-species comparative studies, both when species are independent (as might occur in a variety of statistical designs when groups do not share similarities with other groups due to historical relationships) and when they are phylogenetically related. We considered three parameters: the extent to which two traits are correlated among species, the extent to which variation is partitioned within-versus among-species, and the number of individuals (ranging from one to 20) per species used to estimate mean values for the two traits.

For a range of combinations of these three conditions (correlation of traits among species, partitioning of variance, and sample size per species), we conducted both nonphylogenetic and phylogenetic simulations. We then tested for significant

[1] Present address: Biodiversity Research Centre, University of British Columbia, 6270 University Boulevard, Vancouver, British Columbia, V6T 1Z4, Canada; E-mail: harmon@zoology.ubc.ca.

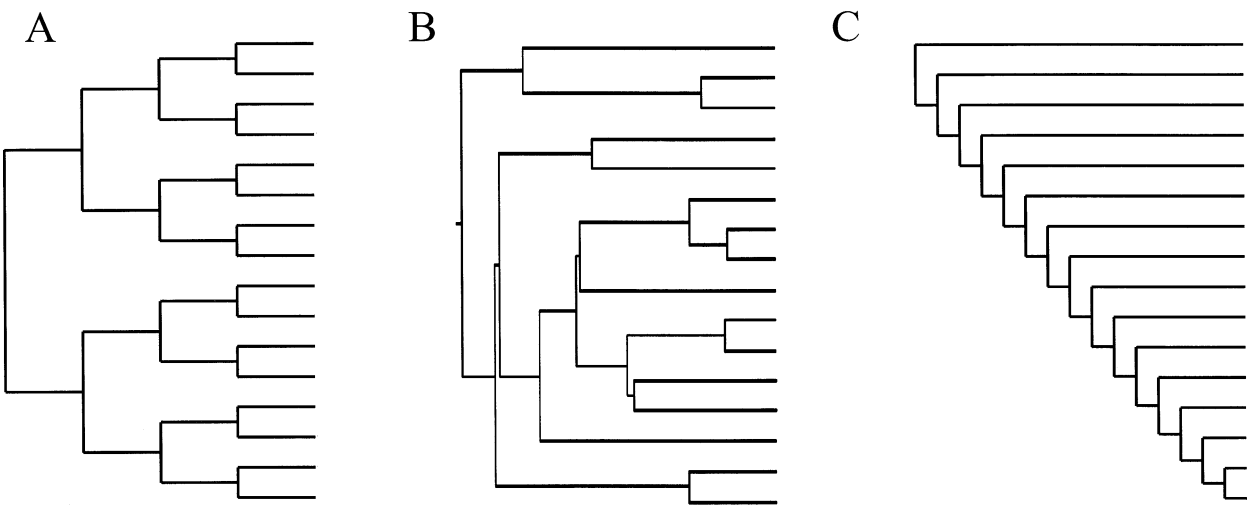

FIG. 1. Trees used for phylogenetic simulations: (A) balanced tree, (B) random pure-birth tree, (C) pectinate tree.

linear correlations among the simulated variables. When the simulated interspecific correlation between traits is zero, then we are examining the rate of type I error (i.e., detecting a relationship that does not exist); conversely, when the correlation is nonzero, we are examining the rate at which real correlations are detected (i.e., the statistical power). The extent to which making incorrect assumptions about evolutionary relationships can affect inferences such as these has been explored extensively in the literature (e.g., Martins and Garland 1991; Ricklefs and Starck 1996; Freckleton et al. 2002; Symonds 2002). We do not address this issue here; instead, we analyze each simulated dataset according to the model under which the data was generated, and focus on the effects of measurement error on these two error rates.

We conducted two types of simulations: nonphylogenetic, where species values were all determined independently of each other (corresponding to many situations—both experimental and observational—when the groups under study have no historical relationship [e.g., Losos et al. 1997; Phelan et al. 2003]), and phylogenetic simulations. We ran phylogenetic simulations on three trees, all including 16 species: one completely balanced, one completely pectinate, and one random tree generated under a pure-birth process by the program Phyl-O-Gen (Rambaut 2002; Fig. 1). To conduct the simulations we simulated measurements of two characters in a dataset of 16 species. To explore the effects of different numbers of species in the analysis, we also simulated data on completely balanced trees including four and 32 species. We did not conduct these additional simulations on pectinate or random trees because the general patterns we describe do not depend strongly on tree shape (see below). The total length (from root to tips) of all phylogenies used was 1.0.

In the simulations, we first generated means for the two characters for each species. For the nonphylogenetic simulations, these were drawn from a bivariate normal distribution with mean zero, variance of each character equal to one, and an expected covariance between characters that varied from 0 to 0.9. For each simulation run, new species means were drawn. For the phylogenetic simulations, we simulated two continuous characters under a Brownian motion model with rate parameter (σ^2) for each character equal to one, starting value zero, and an expected evolutionary covariance coefficient between characters that varied from 0 to 0.9. This model could correspond to two characters with a given genetic covariance evolving under genetic drift (Lande 1979), or to two genetically independent characters undergoing correlational selection in a constant direction (Lande and Arnold 1983).

We then simulated taking measurements for 1–20 individuals of each of these species by drawing individual values from a bivariate normal distribution with mean equal to the species mean generated above, zero covariance between the two characters, and equal variance for each character. This variance differed between runs from 0.0 (100% of the total variance was among species) to 9.0 (10% of the total variance was among species). We then calculated the mean of each character for all individuals in each species and performed the appropriate correlation analysis. For the phylogenetic simulations, nonindependence among species reduced the expected variance among species means to less than one for each simulated character; to account for this, we calculated the expected variance among species for each of the trees under Brownian motion (Martins and Garland 1991) and scaled the variance due to measurement error accordingly.

For each simulation, we determined the statistical significance at the $P = 0.05$ level for the between-species correlation of the two simulated variables. For the nonphylogenetic simulation, this was done using standard linear regression. For the phylogenetic simulations, we determined statistical significance from the regression of independent contrasts (forced through the origin) for each simulated trait on that phylogeny.

For the simulations, we considered several values of each of the three parameters (interspecific covariance between characters: 0.0, 0.3, 0.6, 0.9; number of individuals per species: 1–20; ratio of between species to total variance: 0.1, 0.3, 0.6, 0.9). For every possible combination of these values, we ran 1000 simulations and counted the number of times that a significant interspecific relationship between characters was found. In cases in which the actual correlation between characters was zero, this represents type I error; in all other

cases these counts indicate the statistical power of the procedure to detect the real correlation. All calculations were carried out using a computer program available from the authors on request.

Results

When there is no correlation among characters (cov = 0), for the nonphylogenetic simulations, the type I error remains near 5% regardless of how many individuals per species are measured and the relative variance within versus between species (Fig. 2A–D). For the phylogenetic simulations, type I error is slightly elevated when variance within species is large relative to interspecific variance and sample sizes are low (1–3 individuals per species; Fig. 3A, B). Although the general pattern was similar among all three model tree shapes (balanced, random, and pectinate), the effect was most pronounced for the pectinate phylogeny, producing, in the worst-case scenario when variance among species is 10% of the total, a type I error at $\alpha = 0.05$ of approximately 0.10 over a range of sample sizes, with a maximum of 0.17 at $n = 3$ (Fig. 3A). In contrast, when most of the variation occurs among species, then type I error is barely inflated, if at all, regardless of tree shape (Fig. 3C–D). Including more species in the simulations did not improve these error rates; in fact, type I error was generally higher in the 32 taxa simulations than in the four- and 16-taxa simulations due to increased power to detect spurious correlations (Fig. 4).

The power to detect correlations increases with the strength of the correlation (compare different correlation values within each panel in Figs. 2 and 3), the extent to which overall variance is distributed among, rather than within, species (compare, for any given correlation, the power across Figs. 2 or 3 rows A–D), and with the number of species in the analysis (compare Fig. 3, column 1, with Fig. 4). Finally, at low to intermediate variance ratios (Figs. 2–4 panels B and C), increasing intraspecific sample size increases statistical power. When almost all variation is among species (Figs. 2–4 panel D), then statistical power is maximized even with a sample size of one or very few individuals; conversely, when almost all variation is within species (Figs. 2–4 panel A), then statistical power is extremely low regardless of sample size. These last results do not depend on the type of analysis (phylogenetic or nonphylogenetic) or tree balance.

Discussion

As expected, small intraspecific sample size has no effect on type I error of nonphylogenetic analyses. In contrast, ignoring measurement error when using independent contrasts could lead to increased type I error. However, our simulations have shown that this is not a particularly strong effect as long as either interspecific repeatabilities are high or sample sizes are not extremely small. Maximum type I error was 17% in a case of extremely variable species with low sample sizes on a pectinate phylogeny. As long as interspecific repeatability is 60% or greater, the type I error was never greater than 10%. The implication of these findings is that significant results are generally not called into question unless sample sizes are very small and variation within species relative to that between species is quite large. Even if only one

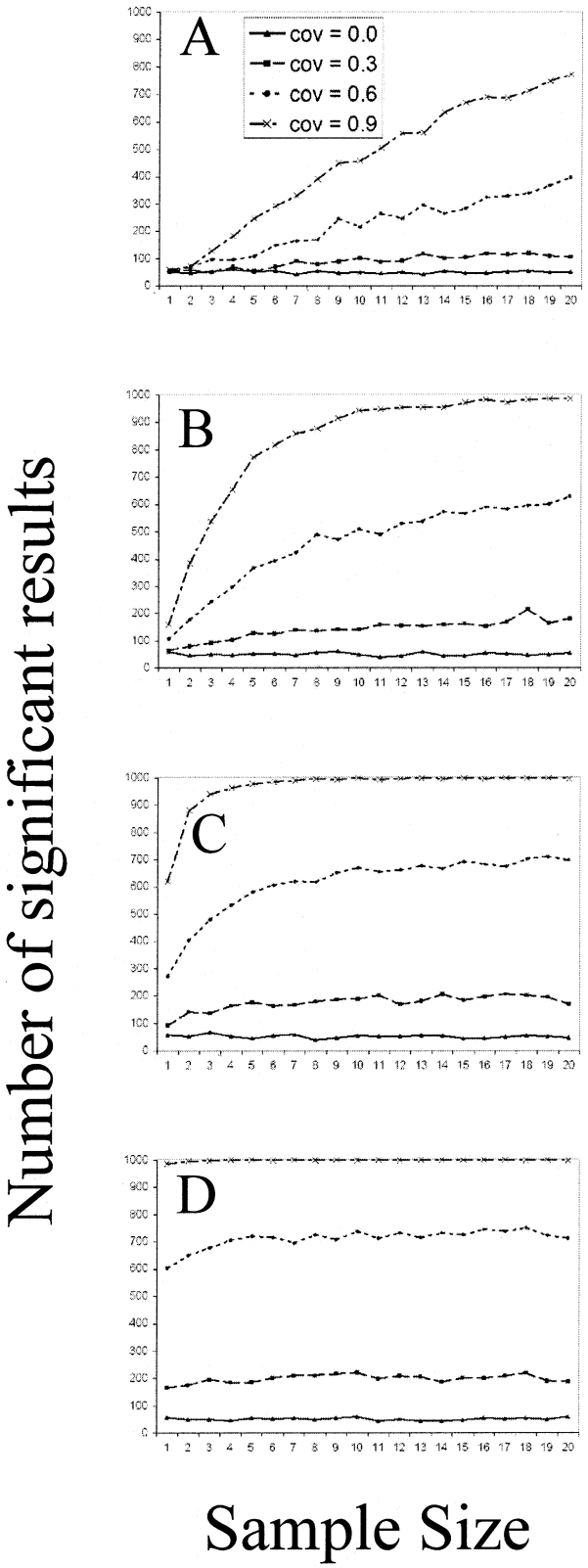

FIG. 2. Relationship between number of individuals measured per species and the number of times a significant correlation was found, out of 1000, for nonphylogenetic simulations. Lines represent levels of covariance between the two characters among species. Individual plots represent levels of intra- versus interspecific variation, such that for each character, interspecific variation as a proportion of the total was (A) 10%, (B) 30%, (C) 60%, and (D) 90%.

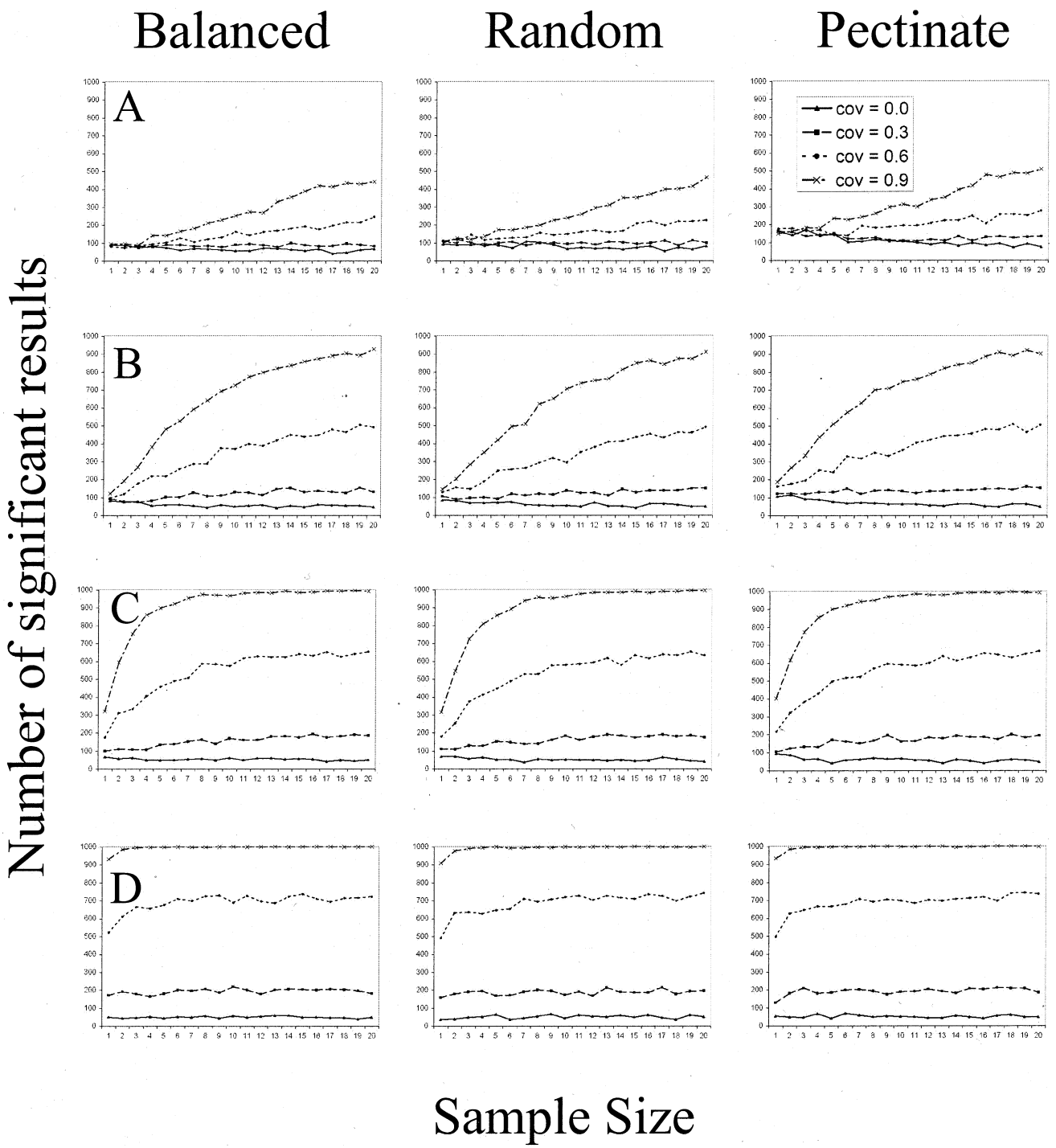

FIG. 3. Relationship between number of individuals measured per species and the number of times a significant correlation was found, out of 1000, for phylogenetic simulations on 16-species trees. Lines represent levels of covariance between the two characters among species. Rows represent levels of intra- versus interspecific variation, such that for each character, variation among species as a proportion of the total was (A) 10%, (B) 30%, (C) 60%, and (D) 90%.

individual is measured per species, as long as variation among species is more than half of the total variation in the dataset, type I error appears to be satisfactory (at least for the parameters and phylogenies we considered). Conversely, nonsignificant results may be the result of low power (type II error) when sample sizes are small or species are extremely variable.

Adding more species to the analysis is not the solution to this problem; in fact, when there is a relatively large amount of measurement error in the dataset, adding more species

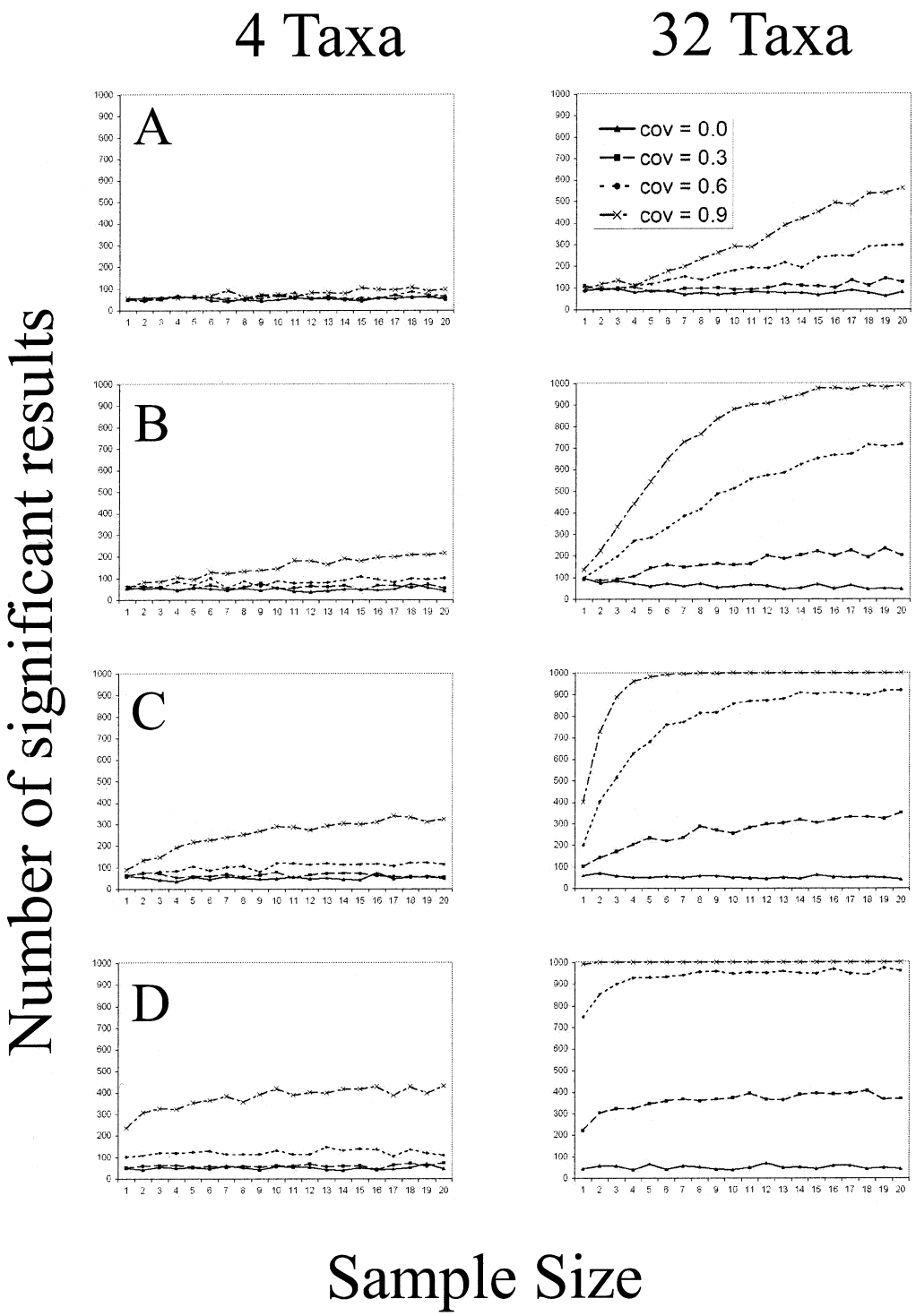

FIG. 4. Relationship between number of individuals measured per species and the number of times a significant correlation was found, out of 1000, for phylogenetic simulations using four-taxon and 32-taxon trees. Lines represent levels of covariance between the two characters among species. Rows represent levels of intra- versus interspecific variation, such that for each character, variation among species as a proportion of the total was (A) 10%, (B) 30%, (C) 60%, and (D) 90%.

increases the type I error rate (Fig. 4, column 2). However, under some circumstances, increasing sample sizes by a few individuals per species can lead to a large increase in statistical power. For example, for a study of 16 species on a balanced phylogenetic tree where interspecific repeatability is 0.3 and the true covariance of the characters is 0.9, measuring one individual per species will detect the correlation about 30% of the time, whereas using five individuals per species will increase the probability of a significant correlation to 90%.

For those planning a comparative study, these simulations suggest some recommendations. First, maximizing intraspecific sample size to the extent possible is generally worthwhile. An important consideration that modifies this recommendation is the extent to which species overlap in trait values. If species in the study tend to be very distinctive such that little overlap exists (i.e., variance distributed among, rather than within, species), then the statistical power of analyses will be great, even with small sample sizes.

Of course, in many cases researchers will not know how variance is distributed a priori. However, to the extent that researchers are aware of great differences among species, then they may be more comfortable with using small intraspecific sample sizes. Similarly, evaluators may place greater credence in nonsignificant results when the variance is shown to reside mostly among species. One approach may be to carry out a preliminary study on a small subset of species that represent a sample of those to be included in a comparative study. By measuring more than one individual for these species, and then carrying out an ANOVA with species as factors, the R^2 is an estimate of the proportion of total variation that is between species. The plots included here can then be used to get a general idea of the sample sizes needed to have acceptable levels of type I error and statistical power. Because these results do not seem to depend strongly on tree balance, these plots may be valid for a wide range of phylogenetic trees. If the investigator finds that the relative variation within species is high, and sample sizes cannot be made large enough to avoid problems with type I error, we suggest using comparative methods that account for measurement error, such as the GLM framework (Martins and Hansen 1997).

We think that these simulations identify, in general, the conditions under which comparative results are suspect. These recommendations must be treated conservatively, however, because the simulations presented here represent a first attempt to quantify the effect of measurement error on comparative statistical studies. Our simulations represent only a small subset of possibilities in terms of the distribution of variation in a group of species. Furthermore, we have used only three model phylogenetic trees for our simulations. Real trees include a much wider range of topologies and branch lengths than analyzed here. One factor that is of particular importance for statistical error in comparative analyses is the length of terminal branches on the phylogenetic tree. If the tree includes two or more very short tip branches, then these results should be used with caution, because contrasts involving such branches can be greatly inflated by measurement error (Purvis and Webster 1999, Felsenstein 2004). If these contrasts are not identified as outliers, they can lead to greatly inflated type I error (results not presented). This emphasizes the need for diagnostic tests of independent contrasts (e.g., Garland et al. 1992) to identify such outliers prior to carrying out any statistical tests.

In an ideal world, researchers would always have large intraspecific sample sizes. In the real world, however, the results of these simulations can provide guidance concerning when sample size is likely to affect the interpretation of comparative analyses.

Acknowledgments

We thank K. Nicholson, M. Johnson, J. J. Kolbe, L. Revell, L. Mahler, J. Vonesh, R. B. Langerhans, M. Bjorklund, and one anonymous reviewer for comments. This research was supported by the National Science Foundation grant DEB 9982736.

Literature Cited

Ackerly, D. D., and R. Nyffeler. 2004. Evolutionary diversification of continuous traits: phylogenetic tests and application to seed size in the California flora. Evol. Ecol. 18:249–272.

Al-kahtani, M. A., C. Zuleta, E. Caviedes-Vidal, and T. Garland, Jr. 2004. Kidney mass and relative medullary thickness of rodents in relation to habitat, body size, and phylogeny. Physiol. Biochem. Zool. 77:346–365.

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

———. 2004. Inferring phylogenies. Sinauer Associates, Sunderland, MA.

Freckleton, R. P., P. H. Harvey, and M. Pagel. 2002. Phylogenetic analysis and comparative data: a test and review of evidence. Am. Nat. 160:712–726.

Garland, T., Jr., P. H. Harvey, and A. R. Ives. 1992. Procedures for the analysis of comparative data using phylogenetically independent contrasts. Syst. Biol. 41:18–32.

Lande, R. 1979. Quantitative genetic analysis of multivariate evolution, applied to brain: body size allometry. Evolution 33: 402–416.

Lande, R., and S. J. Arnold. 1983. The measurement of selection on correlated characters. Evolution 37:1210–1226.

Leal, M., A. K. Knox, and J. B. Losos. 2002. Lack of convergence in aquatic *Anolis* lizards. Evolution 56:785–791.

Losos, J. B., K. I. Warheit, and T. W. Schoener. 1997. Adaptive differentiation following experimental island colonization in *Anolis* lizards. Nature 387:70–73.

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

Martins, E. P. 1994. Estimating the rate of phenotypic evolution from comparative data. Am. Nat. 144:193–209.

Martins, E. P., and T. Garland, Jr. 1991. Phylogenetic analyses of the correlated evolution of continuous characters: a simulation study. Evolution 45:534–557.

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

Patek, S. N., and T. H. Oakley. 2003. Comparative tests of evolutionary trade-offs in a palinurid lobster acoustic system. Evolution 57:2082–2100.

Phelan, J. P., M. A. Archer, K. A. Beckman, A. K. Chippindale, T. J. Nusbaum, and M. R. Rose. 2003. Breakdown in correlations during laboratory evolution. I. Comparative analyses of *Drosophila* populations. Evolution 57:527–535.

Purvis, A., and A. J. Webster. 1999. Phylogenetically independent comparisons and primate phylogeny. Pp. 44–70 *in* P. C. Lee, ed. Comparative primate socioecology. Cambridge Univ. Press, Cambridge, U.K.

Rambaut, A. 2002. Phyl-O-Gen: phylogenetic tree simulator package, Vers. 1.1. University of Oxford, Oxford, U.K.

Ricklefs, R. E., and J. M. Starck. 1996. Applications of phylogenetically independent contrasts: a mixed progress report. Oikos 77:167–172.

Symonds, M. R. E. 2002. The effects of topological inaccuracy in evolutionary trees on the phylogenetic comparative method of independent contrasts. Syst. Biol. 51:541–553.

Corresponding Editor: M. Björklund