BRIEF COMMUNICATIONS

Evolution, 52(1), 1998, pp. 251–255

DESIGNING EXPERIMENTS TO MAXIMIZE THE POWER OF DETECTING CORRELATIONS

PATRICK C. PHILLIPS
Department of Biology, Box 19498, University of Texas at Arlington, Arlington, Texas 76019-0498
E-mail: pphillips@uta.edu

Abstract.—Studies investigating correlations among traits are increasingly common in evolutionary biology. By providing power calculations, minimum sample sizes, and standard errors of correlation coefficients in a variety of contexts, this note provides guidelines to insure that nonzero correlations will not be erroneously dismissed as nonsignificant in properly designed experiements. Forty individuals is often sufficient for reasonable estimation of correlation coefficients, although 100 or more individuals may be nexessary if a large number of traits are involved or comparisons are to be made among the coefficients.

Key words—Correlation analysis, experimental design, multiple comparisons, power of test.

Received December 17, 1996. Accepted September 29, 1997.

The measurement and interpretation of phenotypic and genetic correlations and covariances has been gaining an increasingly central role in the study of evolutionary processes. The study of developmental and functional integration among traits (Cheverud 1982), the association between fitness and phenotype (Lande and Arnold 1983), the functional analysis of behavior and performance (Arnold 1983), and the discrimination and systematic relationships between populations (Garland et al. 1993; Martins and Garland 1991) all rely to some extent on the analysis of patterns of correlations among traits. These correlations can be very large (> 0.9), for example when functionally related morphological traits are studied, but typically would be expected to be much smaller than this, encompassing the entire spectrum of possible associations between traits. The purpose of this note is to determine how large experiments need to be so that these correlations can be determined accurately and so that potentially meaningful correlations will not be overlooked because of a lack of statistical power.

Rice (1989) has emphasized that when large numbers of correlations are tested simultaneously, there is a likelihood that a number of the correlations will be classified as significant by chance, even if the correlations do not in fact differ from zero. To protect against this possibility, the acceptance criterion or critical *P*-value for each test can be lowered so that the overall experimentwise error is controlled at a defined level (usually 0.05). The dilemma caused by this approach is that setting stringent significance levels makes it more difficult to detect smaller, but nonzero correlations. Thus, while protecting against falsely proclaiming a correlation to be significant (a Type I error), one may be overlooking many real and meaningful correlations (a Type II error). The way around this dilemma is to design experiments that are large enough so that they have sufficient power to detect significant correlations even in the face of the multiple comparisons problem.

Power Calculations

Power is the probability of not making a Type II error; that is, the probability that an experiment will detect a correlation as significant when the correlation does in fact exist. Power depends on three factors: the size of the effect to be detected, the size of the experiment conducted, and the Type I error rate, or α-level (Cohen 1992). The α-level is usually initially set at 0.05, but must be adjusted downward when many comparisons are made. This downward adjustment of α will always reduce power, and, for any given level of correlation, this reduction can only be countered by increasing the size of the experiment.

The *z*-transformation, usually used as an approximation for the standard errors on correlation coefficients, is not very useful for power calculations because this transformation is not very accurate for small sample sizes and often does not adequately reflect the asymmetrical shape of the correlation coefficient distribution. A method of calculating power using the exact distribution (Fisher 1915; Hotelling 1953) is presented in the Appendix. Briefly, the critical value for the correlation coefficient for a given sample size and α-level is calculated. Then the probability of obtaining a correlation coefficient greater than this in a sample of size n and a "true" underlying correlation coefficient of ρ is calculated by integrating the correlation coefficient distribution from the critical value to a value of one. In what follows, only the absolute value of the coefficient will be presented, as the power is the same whether the actual coefficient is positive or negative.

A number of studies have presented tables of significance levels and confidence regions on correlation coefficients (Odeh 1982; Paul 1988; Subrahmaniam and Subrahmaniam 1983; Jeyaratnam 1992), but it is difficult to design experiments based on the results of these tables (but see Cohen 1988). A common approach is to look at a table of critical values of the correlation coefficient (e.g., Rohlf and Sokal 1981), and then find the sample size that would give a significant result for a given correlation coefficient. In practice, however, if this were the actual parametric value of the correlation coefficient, then we would expect roughly half of the estimates at this sample size to be nonsignificant, thereby yield a power of approximately 0.5. A more robust approach is to actually calculate the power curves and then design the experiment based upon this information.

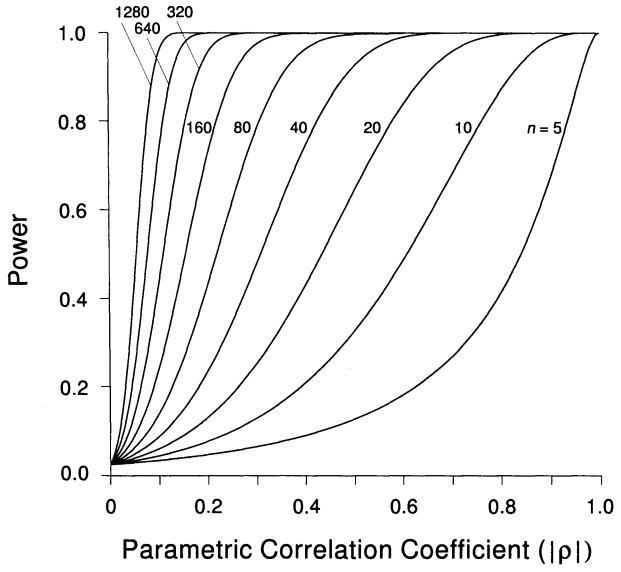

FIG. 1. Effect of the magnitude of the correlation coefficient and sample size on power. Each line gives the power of detecting a given correlation coefficient for an experiment with sample size, *n*. Lines represent successive doublings of sample size. The x-axis is the "true" correlation in the population, and the y-axis is the probability of detecting this correlation with a given sample. Curves for positive and negative correlations of the same magnitude are identical. See the Appendix for details on the calculations.

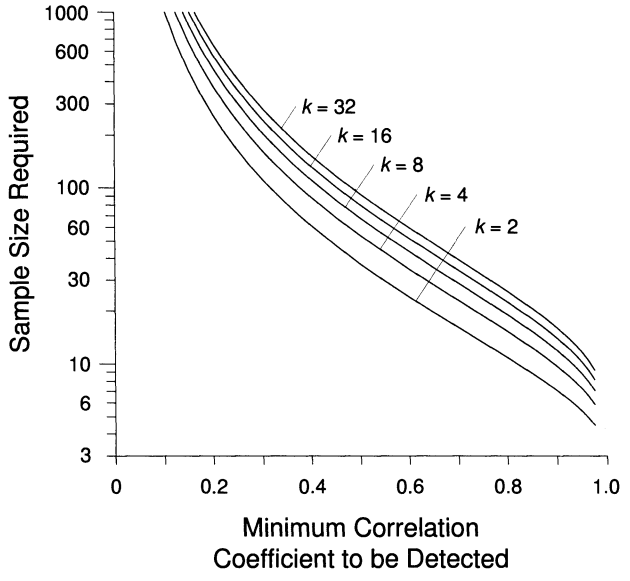

FIG. 2. Sample sizes required to detect correlations of a given size. Lines are isoclines of equal power = 0.9 for experiments involving multiple traits (*k*), with significance levels adjusted by the Bonferroni method (initial $\alpha = 0.05$). Increasingly larger samples are necessary to detect smaller correlations and/or correlations from experiments involving large numbers of traits. Note the logarithmic scale.

RESULTS AND DISCUSSION

Power for Single Correlations.—Both sample size and the magnitude of the correlation coefficient to be detected have a large influence on power (Fig. 1). Experiments involving 10 or fewer individuals are unlikely to detect any significant correlations. Experiments intended to detect moderate to large correlations should probably involve at least 40 individuals, whereas accurate detection of smaller coefficients will require samples of perhaps several hundred individuals. Much like the setting of an α-level, an acceptable level of power is something of a subjective decision. Certainly, one would not like to conduct an experiment that has little hope of producing interpretable results. The standard used here is to take a power value of 0.9 as a design cutoff, although others (Cohen 1992) have recommended 0.8. Because of the shape of the power curves (Fig. 1), choice of the cutoff actually has a small effect on sample size in this region of power. Choosing a smaller value for power will result in smaller experiments, but also increases the probability of a Type II error. A power of 0.9 means that if one were to conduct the same experiment 10 times, then one would expect to obtain a significant result for a given correlation coefficient nine of those times.

Power for Multiple Correlations.—When large sets of traits are analyzed simultaneously then the α-level must be adjusted for multiple comparisons. Using the traditional Bonferroni method, all of the significant correlations must have *P*-values that exceed $2\alpha/k(k - 1)$, where k is the number of traits. However, in the sequential Bonferroni method at least one of the correlation *P*-values must exceed this value, with subsequent coefficients being tested against sequentially larger α-levels (Holm, 1979; Rice, 1989). It may appear at first that the Bonferroni adjustment is too harsh in the case of correlations, since the errors associated with correlations are likely to be correlated if the traits themselves are correlated. However, the experimentwise error that the Bonferroni method controls is based on the null hypothesis of no correlation among any of the traits (i.e., the probability that any one of the correlations that we observe would be significant by chance alone). The Bonferroni method therefore actually provides the exact correction necessary in this case (simulation results not shown). It may be that if some subset of correlations is initially found to be significant, then even the sequential Bonferroni method will be excessively conservative for subsequent tests. This depends on the actual (parametric) values of the underlying correlations, however. Since measuring these correlations is the point of the experiment in the first place, biologists are unlikely to want to build these assumptions into their tests, and instead are more apt to test against the assumption of no correlation structure. More work on this is necessary, but for now it appears that the sequential Bonferroni method is the best multiple comparison correction available for tables of correlations (Rice 1989). From the standpoint of power, the implication of the multiple comparison problem is, then, that experiments involving large numbers of traits will have lower power than a simple two-trait correlation because of the more stringent significance test.

As shown in Figure 2, testing many traits simultaneously means that many more individuals will need to be sampled to achieve the same level of power. For example, it takes 61 individuals to estimate a correlation of 0.4 with 90% power if only two traits are tested, but 111 individuals if eight traits are tested and 152 individuals if 32 traits are tested. Although the required sample size can increase dramatically, the effect becomes somewhat attenuated as more and more traits are

added. The sample size required beyond the two-trait case increases roughly as the quartic root of the number of traits. It should be noted that Figure 2 shows only the minimum correlation to be detected. Using the simple Bonferroni method, the given sample sizes will give the correct power for all correlations in a matrix, but this size will be correct for only the largest correlation if the sequential Bonferroni method is used. In the latter method, subsequent correlations would be tested at slightly higher powers. In practice, the gain will probably not be very large as the function $2\alpha/k(k-1)$ is fairly flat when k is larger than five or six. Nevertheless, experiments involving 100 individuals should be of sufficient size to detect moderate to large correlations even if the number of traits being studied is large (Fig. 2). Several hundred or a thousand individuals are needed to adequately test small correlations on the order of 0.25 or less.

Comparisons of Correlations.—Comparisons of correlation coefficients from two or more populations require even more individuals, because in this case two random variables are being compared (the coefficients) rather than a single random variable being tested against a fixed value (zero). In this case, it may be easier to examine the expected confidence intervals on the range of correlation coefficients to be compared. Figure 3 presents confidence intervals for correlation coefficients calculated from the distribution given in the Appendix. These graphs can be read in two ways. Given a parametric coefficient, the expected range of observed coefficients can be found by moving across the graph horizontally to the point at which the parametric value intersects the curves for the sample size of interest. After the experiment is conducted, however, the confidence interval of the observed coefficient can be found by moving up and down on the graph vertically in an analogous fashion. It is impossible to make a general statement about the size of an experiment necessary to compare two correlations because, unlike something like a *t*-test, the comparison depends on the actual values being compared rather than just difference between the values. (For example, a case of no difference because the correlations were both 0.9 has different sampling characteristics than a case of no difference because the correlations were both 0.2.) As the sample size becomes larger, however, this discrepancy become less prominent. In general, a small sample on the order of 40 individuals will only be able to distinguish between correlations that differ by about 0.7, while 100 or more individuals will be necessary to detect differences between correlations on the order of 0.3 (Fig. 3). An easy way to figure this is to find the upper confidence interval for the smallest correlation that you are interested in (or zero if you want to be the most conservative), and then move directly up from there to find the lower confidence interval for the difference that you would like to be able to detect.

Nonzero Correlations.—When a large number of traits is being studied, it is likely that no single correlation is of particular interest, but instead it is the overall pattern of correlation that is important. For example, Cheverud et al. (1989) have detailed an approach in which correlations among traits are compared against functional hypotheses represented by pattern matrices. A similar approach was used by Kingsolver and Wiernasz (1987) to compare correlations among butterfly wing characters and a functional map relating melanin pattern

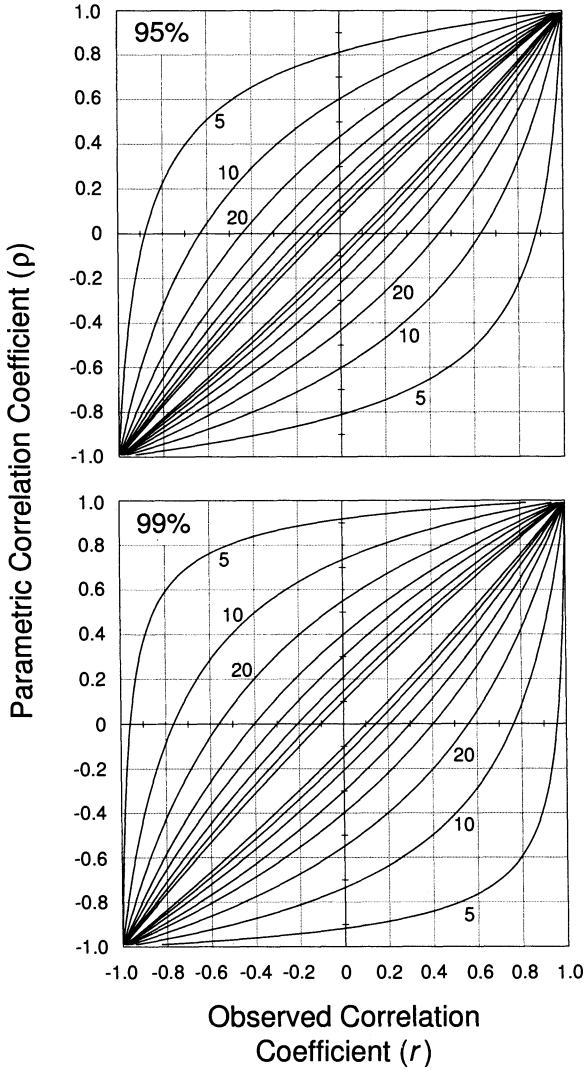

FIG. 3. Confidence intervals on correlation coefficients. The graphs give the 95% and 99% confidence intervals for correlation coefficients drawn from normally distributed populations of different sample sizes. Each curve corresponds to either the upper or lower confidence interval for a given sample size, which increase following the doubling pattern n = 5, 10, 20, 40, 80, 160, 320, 640 (labels for the higher numbers are not shown for the sake of clarity). The graph can be read in two ways. First, for a given parametric correlation coefficient, ρ, the graph can be followed horizontally from that point on the y-axis across to where a horizontal line would intersect the lower and upper confidence intervals for that value and sample size. This provides the region in which 95% or 99% of the actual observations would be expected to fall. Second, for a given observed correlation coefficient, r, the graph can be read vertically from that point on the x-axis to where a vertical line would intersect the lower and upper confidence intervals for that value and sample size. This provides the interval in which the parametric value would be expected to fall 95% or 99% of the time.

to body temperature (Kingsolver 1987). In these cases, getting the sign of the correlation right might be as important as its absolute magnitude. The probability of correctly estimating the sign of a correlation is given in Figure 4 (see the Appendix for the methodology). These results support the notion that it is indeed easier to get the sign of the correlation

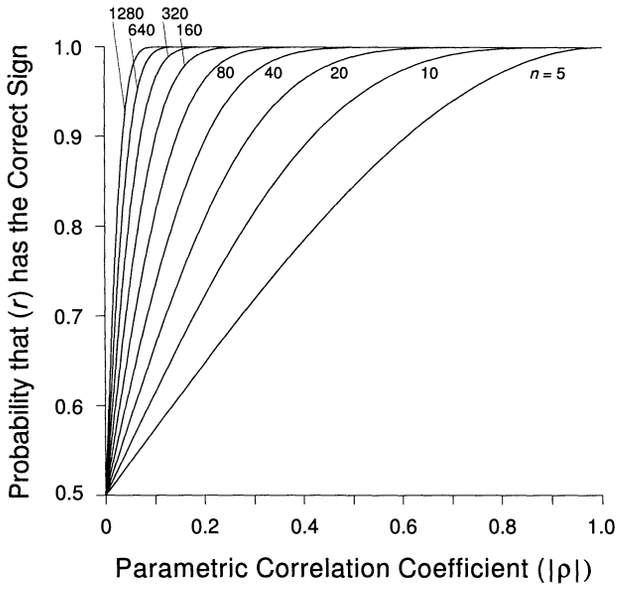

FIG. 4. Probability that the sample correlation coefficient (r) has the correct size given a particular parametric correlation coefficient. Lines give this probability for samples of size n.

correct than to tell whether the correlation is significant. However, a number of individuals must still be sampled (say, > 40) for reasonable performance on most correlations. In addition, if the pattern of a large correlation matrix is to be tested, then the fact that each individual correlation could have the wrong sign becomes important. In the worst case, if γ is the probability of obtaining a correct sign on the correlation, and there are m entries in the correlation matrix, then $(1 - \gamma^m)$ is the probability of making an error somewhere in the matrix if all entries are treated as independent. This number approaches one rapidly as m increases unless γ is also close to one. Therefore, as was the case for the influence of multiple comparisons on significance, more individuals will need to be sampled for the pattern of correlation signs to be correct in large matrices.

In the end, the analysis and comparison of patterns of correlation are probably best studied by composite methods that extract patterning information directly from the correlation matrix (e.g., Zelditch 1987; Flurry 1988; Phillips and Arnold, unpubl.). The power of these composite methods remains to be seen, and may be higher or lower depending on the underlying pattern of the correlations. However, any method of correlational analysis depends on accurate correlation estimates, and the results presented here should help serve as a general guide.

Types of Correlations.—The results presented above are applicable to any dataset in which the observations to be correlated are taken on every individual. These would include phenotypic correlations and some forms of genetic correlations, such as those based on family means. These results cannot be used for more general quantitative genetic approaches, as correlations in this context are based on estimating covariance components derived from the relationships among relatives (Falconer and Mackay 1996). Sample sizes required for equivalent power in quantitative genetic experiments are likely to be substantially larger, and here power relies on a balance between the number of families and the total number of individuals (Klein 1974).

CONCLUSIONS

If only large correlations are of concern, then a few tens of individuals should provide sufficient power, whereas 100 or so individuals will be adequate for most uses in evolutionary studies (Fig. 2). Studies concentrating on small correlations, however, will require several hundred individuals to have sufficient power to make conducting the experiment worthwhile. This is especially problematic if large numbers of traits are studied simultaneously (Fig. 2). If it is important that only the sign of the correlation be correct then far smaller samples are necessary, with a few dozen individuals being sufficient in many cases (Fig. 4). Accurate comparisons between correlations are very difficult to achieve, and certainly several hundred individuals will be required if reasonable comparisons are to be made (Fig. 3). In general, then, conducting studies looking at patterns of correlation are not trivial undertakings if performed correctly. The consequence of not designing experiments with sufficient power, however, is that one can be left at the end of a project not really knowing its outcome, one way or another.

ACKNOWLEDGMENTS

A Mathematica (Wolfram 1996) notebook for carrying out these power calculations is available via http://www.uta.edu/biology/phillips or from the author on request. I wish to thank two anonymous reviewers for many helpful comments on the manuscript. This work was funded in part by National Science Foundation BIR-9612469.

LITERATURE CITED

ARNOLD, S. J. 1983. Morphology, performance and fitness. Am. Zool. 23:347–361.

CHEVERUD, J. M. 1982. Phenotypic, genetic, and environmental morphological integration in the cranium. Evolution 36:499–516.

CHEVERUD, J. M., G. P. WAGNER, AND M. M. DOW. 1989. Methods of the comparative analysis of variation patterns. Syst. Zool. 38: 201–213.

COHEN, J. 1988. Statistical power analysis for the behavioral sciences. 2d ed. Erlbaum, Hilsdale, NJ.

———. 1992. A power primer. Psych. Bull. 112:155–159.

FALCONER, D. S., AND T. F. C. MACKAY. 1996. Introduction to quantitative genetics. 4th ed. Longman, Essex, U.K.

FISHER, R. A. 1915. Frequency-distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika 10:507.

FLURRY, B. 1988. Common principal components and related multivariate models. Wiley, New York.

GARLAND, T., JR., A. W. DICKERMAN, C. M. JANIS, AND J. A. JONES. 1993. Phylogenetic analysis of covariance by computer simulation. Syst. Biol. 42:265–292.

HOLM, S. 1979. A simple sequentially rejective multiple test procedure. Scand. J. Statist. 6:65–70.

HOTELLING, H. 1953. New light on the correlation coefficient and its transformations. J. R. Stat. Soc. B 15:193–232.

JEYARATNAM, S. 1992. Confidence intervals for the correlation coefficient. Stat. Prob. Lett. 15:389–393.

KINGSOLVER, J. G. 1987. Evolution and coadaptation of thermoregulatory behavior and wing pigmentation pattern in pierid butterflies. Evolution 41:472–490.

KINGSOLVER, J. G., AND D. C. WIERNASZ. 1987. Dissecting correlated characters: adaptive aspects of phenotypic covariation in melanization pattern of *Pieris* butterflies. Evolution 41:491–503.
KLEIN, T. W. 1974. Heritability and genetic correlation: statistical power, population comparisons, and sample size. Behav. Genet. 4:171–189.
LANDE, R., AND S. J. ARNOLD. 1983. The measurement of selection on correlated characters. Evolution 37:1210–1226.
MARTINS, E. P., AND T. GARLAND JR. 1991. Phylogenetic analyses of the correlated evolution of continuous characters: a simulation study. Evolution 45:534–557.
ODEH, R. E. 1982. Critical values of the sample product-moment correlation coefficient in the bivariate normal distribution. Commun. Statist.-Simula. Computa. 11:1–26.
PAUL, S. R. 1988. Estimation of and testing significance for a common correlation coefficient. Commun. Statist.-Theory Meth. 17:39–53.
RICE, W. R. 1989. Analyzing tables of statistical tests. Evolution 43:223–225.
ROHLF, F. J., AND R. R. SOKAL. 1981. Statistical tables. 2d ed. Freeman, New York.
STUART, A., AND J. K. ORD. 1987. Kendall's advanced theory of statistics. Vol. 1. Distribution theory. 5th ed. Oxford Univ. Press, New York.
SUBRAHMANIAM, K., AND K. SUBRAHMANIAM. 1983. Some extensions to Miss F. N. David's tables of the sample correlation coefficient: distribution function and percentiles. Sanknya B 45:75–147.
WOLFRAM, S. 1996. The mathematica book. 3d ed. Wolfram Media, Champaign, IL.
ZELDITCH, M. L. 1987. Evaluating models of developmental integration in the laboratory rat using confirmatory factor analysis. Syst. Zool. 36:368–380.

Corresponding Editor: L. Leamy

APPENDIX

The density function of the correlation coefficient used in this analysis is based on the hypergeometric function expression of Hotelling (1953), and is given by

$$d\varphi = \frac{(n-2)dr}{(n-1)\sqrt{2}B(1/2, n-1/2)}(1-\rho^2)^{1/2(n-1)} \times (1-r^2)^{1/2(n-4)}(1-\rho r)^{3/2-n}F\left[\frac{1}{2},\frac{1}{2},n-\frac{1}{2},\frac{1}{2}(1+\rho r)\right], \quad \text{(A1)}$$

where ϕ is the density function, n is the sample size, ρ is the parametric value of the correlation coefficient, r is the estimate of the correlation coefficient, $B()$ is the beta function, and $F()$ is the hypergeometric function (Stuart and Ord 1987, eq. 16.66). Power was calculated as the integral of this density with limits of integration from the value of r yielding the appropriate $\alpha/2$ (two-tailed test) on the distribution with $\rho = 0$ to positive infinity. Numerical solutions to the hypergeometric function tended to become unstable as n became large, so $F()$ was assumed to be 1.0 when $n > 40$, as is the asymptotic result.

The test of coefficient sign (Fig. 4) was performed using the relationship,

$$P(r \geq 0) = P\left\{t_{n-1} \geq -\left[\frac{(n-1)\rho^2}{1-\rho^2}\right]^{1/2}\right\}, \quad \text{(A2)}$$

where t_{n-1} is a t-value for the given α (Stuart and Ord 1987, eq. 16.68). All calculations were performed using Mathematica (Wolfram 1996).

Evolution, 52(1), 1998, pp. 255–260

THE GENETIC STRUCTURE OF A GYNODIOECIOUS PLANT: NUCLEAR AND CYTOPLASMIC GENES

DAVID E. MCCAULEY
Department of Biology, Vanderbilt University, Nashville, Tennessee 37235
E-mail: mccaulde@ctrvax.vanderbilt.edu

Abstract.—Sex expression in gynodioecious plants is often determined by an interaction between biparentally and maternally inherited genes. Their relative rates of gene flow should be considered when modeling the evolution of the sex ratio in structured populations. In order to understand patterns of gene flow in *Silene vulgaris*, a gynodioecious plant, genetic structure was estimated from biparentally inherited genetic markers (allozymes) and a maternally inherited marker (chloroplast DNA) using Wright's F_{st}. Based on data from 16 local populations, chloroplast DNA showed considerably more genetic structure than did allozymes (F_{st} values of 0.62 and 0.22, respectively). This suggests that the rate of gene flow is about three times greater for nuclear genes.

Key words.—cpDNA, gene flow, gynodioecy, population structure, *Silene*.

Received May 22, 1997. Accepted September 12, 1997.

Gynodioecious species are those in which individuals can be classified as being either hermaphroditic or functionally female. In gynodioecious plants, sex expression is usually a cytonuclear phenomenon (Saumitou-Laprade et al. 1994). That is, factors in the cytoplasm are capable of conferring cytoplasmic male sterility (CMS) unless the phenotypic effects of these factors are masked by restorer alleles, which are associated with one or more nuclear loci. Thus, sex expression could be considered a multilocus phenomenon with strong epistatic interactions between loci. If the sex ratio of a gynodioecious population is defined by the relative proportions of female and hermaphroditic individuals, then it should be a function of the allele frequencies at the loci that determine sex expression.

Studies of natural populations of gynodioecious species have revealed that both CMS and restorer loci often display