Genet. Res., Camb. (1999), **74**, *pp.* 255–264. *With* 4 *figures.* Printed in the United Kingdom © 1999 Cambridge University Press

Estimating genetic correlations in natural populations

MICHAEL LYNCH*
Department of Biology, University of Oregon, Eugene, OR 97403, USA

(*Received 14 May 1999 and in revised form 3 August 1999*)

Summary

Information on the genetic correlation between traits provides fundamental insight into the constraints on the evolutionary process. Estimates of such correlations are conventionally obtained by raising individuals of known relatedness in artificial environments. However, many species are not readily amenable to controlled breeding programmes, and considerable uncertainty exists over the extent to which estimates derived under benign laboratory conditions reflect the properties of populations in natural settings. Here, non-invasive methods that allow the estimation of genetic correlations from phenotypic measurements derived from individuals of unknown relatedness are introduced. Like the conventional approach, these methods demand large sample sizes in order to yield reasonably precise estimates, and special precautions need to be taken to eliminate bias from shared environmental effects. Provided the sample consists of at least 20% or so relatives, informative estimates of the genetic correlation are obtainable with sample sizes of several hundred individuals, particularly if supplemental information on relatedness is available from polymorphic molecular markers.

1. Introduction

The field of quantitative genetics has long been concerned with the partitioning of variances and covariances of complex characters into components influenced by various genetic and environmental sources. Literally thousands of studies report estimates of heritabilities for characters of evolutionary and/or economic interest (Falconer & Mackay, 1996; Roff, 1997; Lynch & Walsh, 1998). When sample sizes are adequate, nearly all such studies reveal the existence of additive genetic variation for the traits involved. Such an observation is not surprising since we now know that mutation is a fairly powerful source of genetic variation for most traits (reviewed in Lynch & Walsh, 1998, chapter 12). Thus, for univariate analysis, the only serious question concerns the actual magnitude of the components of variance and the nature of the forces that determine them.

Less straightforward is the genetic correlation between traits, as both the magnitude and the sign of this parameter depend on the general pleiotropic effects of genes and, in some cases, on the pattern of gametic-phase disequilibrium. Genetic correlations between traits are of substantial interest because, depending on their sign, they can either facilitate or impede the joint evolution of the characters involved. A conflict arises when two negatively genetically correlated traits are both selected in the same direction, as the selective advance of each character tends to pull the other character in the opposite direction. A perfect genetic correlation (equal to ± 1) between two traits presents an absolute evolutionary barrier, since no change in either character can occur without a parallel change in the other. Falconer's influential text book greatly elevated our understanding of the evolutionary consequences of genetic correlations in natural and domesticated populations.

Estimation of quantitative-genetic parameters is a demanding enterprise even with nicely balanced designs in the most controlled environments. However, compared with univariate parameters, genetic correlations are particularly difficult to assess because they require accurate estimates of three parameters – the genetic variances of the two traits, and the genetic

* Tel: +1(541) 346 5579. Fax: +1(541) 346 2364. e-mail: mlynch@oregon.uoregon.edu

covariance between them. Because all three estimates are generally obtained from the phenotypic covariances of relatives, they can take on any value. Thus, contrary to the situation with the well-known product–moment correlation, estimates of the genetic correlation can fall outside of the parametric limits (± 1) or can be undefined when one of the genetic-variance estimates is negative (Hill & Thompson, 1978). Sample sizes of a few thousand pairs of relatives are often necessary to achieve estimates that can confidently be interpreted at the level of even single significant digits (Van Vleck & Henderson, 1961; Brown, 1969; Klein, 1974; Visscher, 1998), although a simple knowledge of the sign of the genetic correlation can be achieved with less, but still substantial, effort.

The classical approach to estimating genetic components of variance and covariance relies on the phenotypic resemblance between individuals of known relatedness. Thus, the vast majority of studies in quantitative genetics involve controlled breeding programmes in which individuals are raised in artificial, and often unusually benign, environments. Such treatment raises two significant problems. First, since genetic components of variance and covariance can differ dramatically among environments (Falconer & Mackay, 1996; Roff, 1997; Lynch & Walsh, 1998), the ideal setting for the estimation of quantitative-genetic parameters is the environment of interest; for evolutionary studies, the preferred setting is the natural environment. Secondly, even if the environmental conditions can be made to match those in nature, the conventional known-pedigree approach to quantitative genetics is not an option for species that cannot be raised easily in the laboratory, barnyard or greenhouse in reasonable amounts of time.

The purpose of this paper is to explore the feasibility of two new methods for estimating genetic correlations between characters in natural populations. The first of these methods uses samples of individuals for which the degrees of relationship are completely unknown, whereas the second uses molecular markers to estimate relationship coefficients.

2. Estimation in the absence of pedigree information

Consider two characters, x and y, whose genetic basis is entirely additive, and denote the genetic variances of the two traits as $\sigma_A^2(x)$ and $\sigma_A^2(y)$ and the genetic covariance between the traits as $\sigma_A(x,y)$. Assuming for the time being that shared environmental effects do not contribute to the phenotypic resemblance between relatives, then from well-established results (Falconer & Mackay, 1996; Lynch & Walsh, 1998) the expected phenotypic covariance between individuals is a function of these genetic components of variance and covariance and of the relationship coefficient, r, which equals 0·5 for full-sib and parent–offspring relationships, 0·25 for half-sib and grandparent–grandchild relationships, etc.

Letting the mean population-wide phenotypes of the two traits be μ_x and μ_y, and denoting the phenotypes of individual i as $z_i(x)$ and $z_i(y)$, the expected phenotypic covariance for character x across pairs of individuals (i and j) with relationship r is

$$\sigma_z(x,x\,|\,r) = E[(z_i(x)-\mu_x)(z_j(x)-\mu_x)\,|\,r] = r\sigma_A^2(x), \tag{1a}$$

whereas that for character y is

$$\sigma_z(y,y\,|\,r) = E[(z_i(y)-\mu_y)(z_j(y)-\mu_y)\,|\,r] = r\sigma_A^2(y), \tag{1b}$$

and the expected covariance between the two traits, one in each individual, is

$$\sigma_z(x,y\,|\,r) = E[(z_i(x)-\mu_x)(z_j(y)-\mu_y)\,|\,r] = r\sigma_A(x,y). \tag{1c}$$

The key to estimating the genetic correlation among traits is the fact that the expected values of all three types of phenotypic covariance across individuals are preceded by the relationship coefficient, which cancels out in the function

$$\rho_G(x,y) = \frac{\sigma_z(x,y\,|\,r)}{\sqrt{[\sigma_z(x,x\,|\,r)\,\sigma_z(y,y\,|\,r)]}} = \frac{\sigma_A(x,y)}{\sqrt{[\sigma_A^2(x)\,\sigma_A^2(y)]}}. \tag{2}$$

Thus, the estimate of the genetic correlation, which is obtained by substituting the observed phenotypic covariances for the expectations in this formula, does not directly incorporate the relationship coefficient. Equation (2) embodies the standard approach to estimating genetic correlations when, for example, i and j represent the members of sets of individuals with constant r, e.g. two sibs or parents and offspring.

This same principle applies to a set of individuals with mixed degrees of relatedness. Suppose, for example, that the sample consists of a series of pairs of individuals (with members again being denoted by i, j) with different relationships, such that r_{ij} is the relatedness between the members of a pair. The expected phenotypic covariances involving these pairs of individuals then have expectations

$$\sigma_z(x,x) = \bar{r}\sigma_A^2(x), \tag{3a}$$

$$\sigma_z(y,y) = \bar{r}\sigma_A^2(y), \tag{3b}$$

$$\sigma_z(x,y) = \bar{r}\sigma_A^2(x,y), \tag{3c}$$

where $\bar{r}$ is the mean relationship coefficient between the pairs of assayed individuals. The ratio of phenotypic covariances, $\sigma_z(x,y)/\sqrt{[\sigma_z(x,x)\,\sigma_z(y,y)]}$, again eliminates the unknown parameter $\bar{r}$, thereby reducing to $\sigma_A(x,y)/\sqrt{[\sigma_A(x)\,\sigma_A(y)]}$, the genetic correlation.

Thus, the proposition presented here is that the genetic correlation might be estimable in the absence

https://doi.org/10.1017/S0016672399004243 Published online by Cambridge University Press

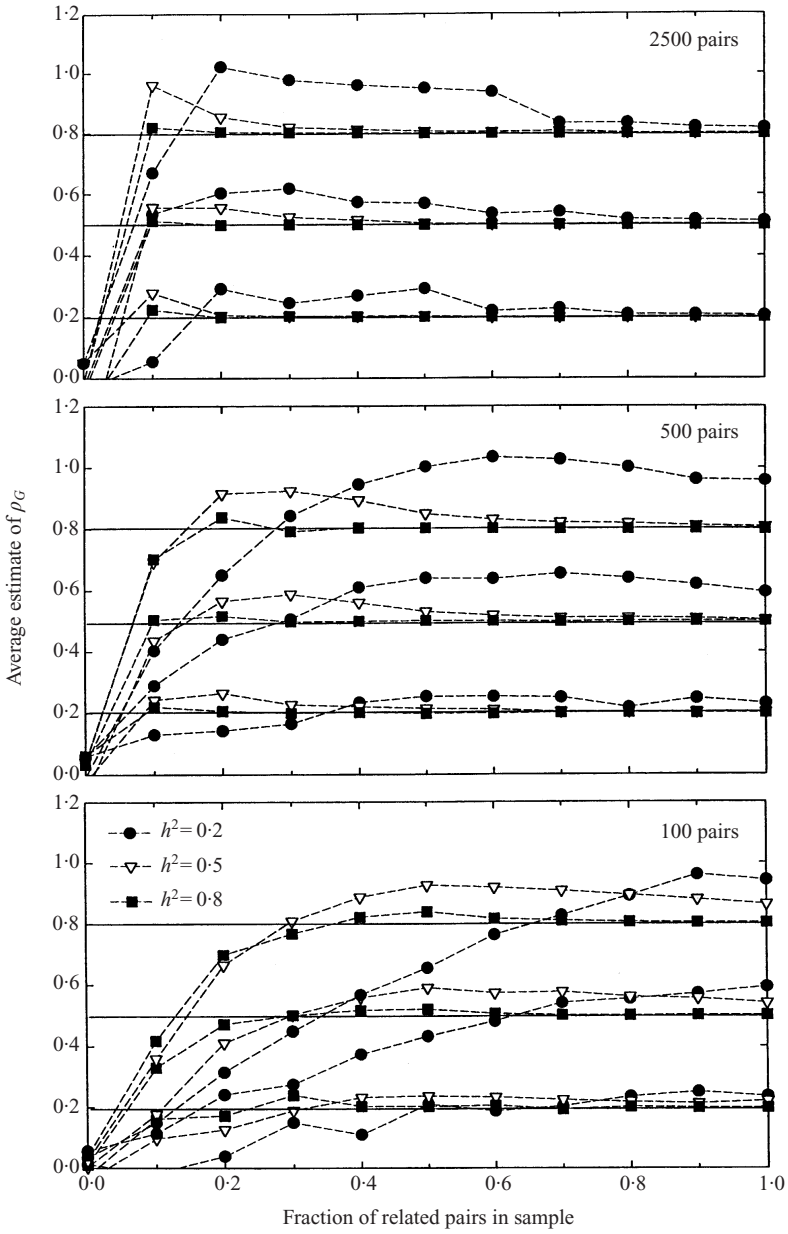

Fig. 1. Mean estimates of the genetic correlation as a function of the fraction of pairs of individuals that are full sibs (the remaining pairs consisting of nonrelatives). Results are given for three genetic correlations (0·2, 0·5 and 0·8, denoted by the continuous horizontal lines), three heritabilities (denoted by the different symbols, and three sample sizes (2500, 500 and 100 pairs of individuals, shown in different panels). The environmental correlation within individuals is equal to zero, and the environmental and genetic values of the two traits are bivariate normally distributed with zero genotype–environment covariance.

of known relationships by compiling a list of pairs of individuals and computing the three phenotypic covariances involving characters x and y. Provided that some of the pairs consist of relatives, then the expected values of all three covariances will be nonzero assuming there is genetic variance (covariance) for the traits, and all three will be proportional to $\bar{r}$. The unknown parameter $\bar{r}$ is then eliminated by dividing the phenotypic covariance across traits by the square root of the product of the phenotypic covariances within traits. Letting Cov denote an observed covariance across the members of pairs of individuals, the proposed estimate for the genetic correlation becomes

$$\hat{\rho}_G(x, y) = \frac{\mathrm{Cov}[z_i(x), z_j(y)]}{\sqrt{\{\mathrm{Cov}[z_i(x), z_j(x)] \,.\, \mathrm{Cov}[z_i(y), z_j(y)]\}}} \qquad (4)$$

Statistical properties

The general utility of the proposed technique will depend on a number of issues. First, as noted above,

https://doi.org/10.1017/S0016672399004243 Published online by Cambridge University Press

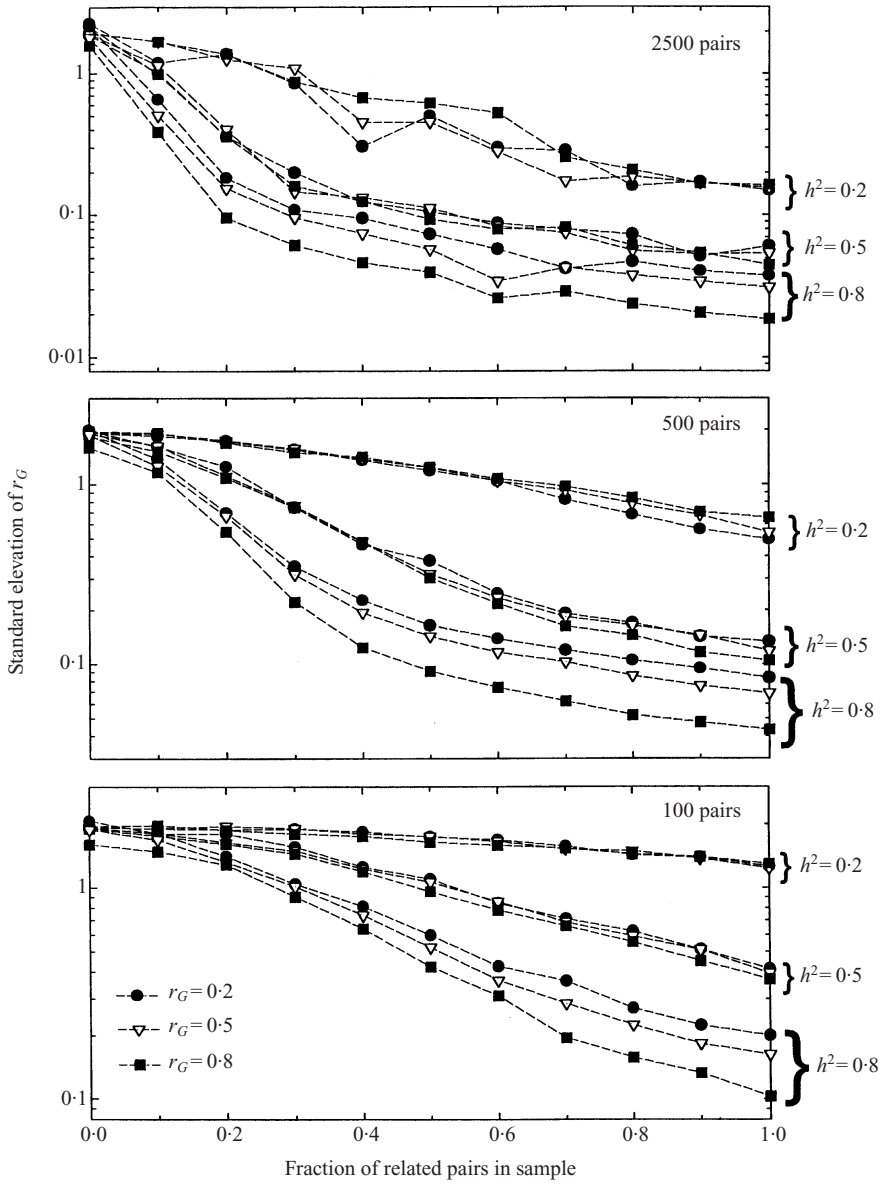

Fig. 2. Sample standard deviations of the genetic correlation as a function of the fraction of pairs of individuals that are full sibs (the remaining pairs consisting of non-relatives). Results are given for three genetic correlations (denoted by the different symbols), three heritabilities (0·2, 0·5 and 0·8) and three sample sizes (2500, 500 and 100 pairs). The environmental correlation is equal to zero, and the environmental and genetic values of the two traits are bivariate normally distributed with zero genotype–environment covariance.

even with a balanced experimental design with known relatives, the sample size requirements for accurate estimates of genetic correlations are substantial. They must be even greater for samples of individuals of uncertain and distant relationships. Secondly, the sample composition is expected to be an important determinant of statistical power, with the latter increasing with the fraction of relatives and the affinity of their relationships. Thirdly, even if samples with adequate numbers of relatives are achievable, it is unclear whether the preceding estimator will yield unbiased estimates when the degree of relatedness varies. Fourthly, characters with higher heritabilities are expected to yield more accurate estimates of genetic correlations because the phenotype more accurately reflects the underlying genetic values.

To evaluate the power of the proposed technique, computer simulations were used to generate the joint distributions of two characters in pairs of individuals with known degrees of relatedness. The environmental variances of the characters were scaled to $\sigma_E^2(x) = \sigma_E^2(y) = 1$ throughout, the environmental correlation between traits within the same individual was assumed to be equal to zero, and the expected character means were set equal to zero. The genetic correlation between traits, $\rho_G(x, y)$, was treated as a free parameter, as were the genetic variances for the two traits. The distributions of genotypic and environmental values were

https://doi.org/10.1017/S0016672399004243 Published online by Cambridge University Press

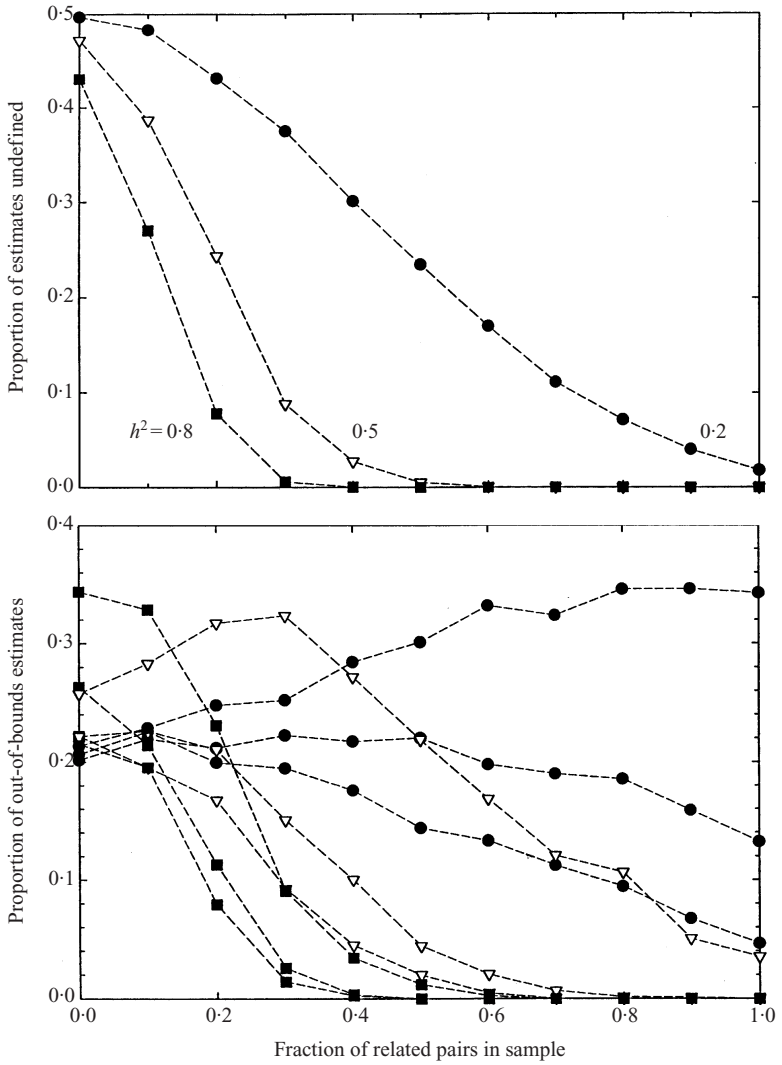

Fig. 3. Upper panel: Proportion of genetic correlation estimates that are undefined as a consequence of a negative variance-component estimate. Results are given as a function of the heritabilities of the two traits and of the proportion of sampled pairs of individuals that are full sibs. Each set of conditions involves samples of 500 total pairs and assumes an environmental correlation equal to zero. Lower panel: Proportion of genetic correlation estimates that exceed the parametric bounds of ± 1. Results are given for the conditions in the upper panel, using the same symbols for the different heritabilities. Within each set of heritabilities, there are three sets of points; the lines with the lowest, intermediate and highest values denote the results for genetic correlations equal to 0·20, 0·50 and 0·80 respectively.

always multivariate normal, and the subroutine for generating the four phenotypic values in pairs of individuals was tested extensively to ensure that it was generating the expected levels of genetic and environmental variances and covariances within and between individuals. All the following analyses involve sets of mixtures of various proportions of full sibs and non-relatives.

For each set of parameter values explored, 10000 random data sets were generated and assayed for the genetic correlation, as estimated by (4). Thus, for each simulated data set there was a list of pairs of individuals (i and j), each with phenotypic values of the two characters, x and y. The phenotypic means for each character were estimated for each column of individuals, and then the four covariances of characters ($z_i(x)$ vs $z_j(x)$, $z_i(x)$ vs $z_j(y)$, $z_i(y)$ vs $z_j(x)$, and $z_i(y)$ vs $z_j(y)$) were estimated. The covariance across characters, in the numerator of (4), was estimated as the average of the reciprocal covariances involving $z_i(x)$ versus $z_j(y)$ and $z_i(y)$ versus $z_j(x)$. The sample standard deviation of $\hat{\rho}_G(x, y)$, i.e. the standard deviation among estimates derived from independent samples of the same population, was derived from the 10000 replicate estimates.

As noted above, genetic correlation estimates are undefined if by chance the estimate of one of the entries in the denominator of (4) is negative. Thus, for each set of parameter values, the frequency of undefined correlation estimates was estimated. In the computation of the means and standard deviations of $\hat{\rho}_G(x, y)$, undefined estimates were ignored, but those

https://doi.org/10.1017/S0016672399004243 Published online by Cambridge University Press

outside the ± 1 bounds were employed except in the rare event that they exceeded ± 10, the latter treatment being primarily a precaution to prevent bias from rare outliers.

The average estimated genetic correlation does not always strictly coincide with the expected value when the preceding method is applied, although sets of conditions do exist in which the bias is negligible compared with the substantial sampling variance of the estimates (Fig. 1). In general, estimates of ρ_G tend to be biased downwards when the number of informative pairs of individuals in the sample is low (either because of a small fraction of pairs of relatives or because of a small total sample size), and biased upwards when the fraction of shared relatives is high but the heritabilities of the trait are low. For example, when the sample consists of a mixture of full sibs and non-relatives and a total sample of 500 pairs of individuals, estimates of ρ_G are nearly unbiased provided at least 20% of the pairs of individuals are sibs and the heritabilities of the two traits are on the order of 0·5 or greater (Fig. 1). With very low heritabilities and a large fraction of relatives, the average estimate of ρ_G can be as high as 1·00 when the true value is 0·80 or as high as 0·65 when the true value is 0·50. It should be noted, however, that this upward bias in estimates of ρ_G also exists when the sample size consists entirely of pairs of relatives (in which case the estimate is equivalent to that obtained by conventional quantitative-genetic analysis).

The standard deviations of ρ_G estimates are very high and increase with decreasing genetic information content in the sample, i.e. with decreasing frequency of pairs of relatives and with decreasing trait heritabilities (Fig. 2). For example, when 500 pairs of individuals are assayed, regardless of ρ_G, approximately 60% of the sample must consist of close relatives (full sibs in the example provided in Fig. 2) before the standard deviation drops below one if $h^2 = 0{\cdot}20$; and for $h^2 = 0{\cdot}50$ and 0·80, respectively, the critical fractions of full sibs are on the order of 25% and 15%. With 2500 pairs, standard errors of approximately 0·5 can be achieved with fractions of full sibs of 50% when $h^2 = 0{\cdot}02$, 20% *when* $h^2 = 0{\cdot}5$ and 10% when $h^2 = 0{\cdot}8$.

As noted above, undefined estimates of ρ_G arise when sampling error results in a negative estimate of the phenotypic covariance of the same trait among pair members. Because this problem is a function of univariate analysis, its incidence does not depend on the magnitude of the genetic correlation, but it does depend on the frequency of related pair members (upper panel, Fig. 3). With 500 pairs of measured individuals, the problem is negligible when the fraction of relatives exceeds 30% when the traits have heritabilities equal to 0·80, but is non-negligible even when all pairs consist of relatives when heritabilities equal 0·20. In addition, the frequency of out-of-bounds estimates ($\hat{\rho}_G^2 > 1{\cdot}00$) can be considerable unless heritabilities are high, even when all pairs contain related individuals (lower panel, Fig. 3).

3. Marker-assisted estimates

The method introduced in the previous section assumes zero knowledge about the degree of relatedness of pairs of individuals. However, even when it is not possible to ascertain relationships with certainty from direct observation, it is often feasible to estimate relatedness by using polymorphic molecular markers. In principle, use of this additional information should increase the precision of estimates of the genetic correlation.

A regression method, modified from Ritland (1996), provides a marker-assisted means for estimating the genetic variances and covariances of two traits. Again letting $z_i(x)$ be the phenotypic value of character x in individual i, and μ_x be the population mean phenotype for the trait, an index of the phenotypic covariance for trait x in two individuals is

$$C_{ij}(x,x) = [z_i(x) - \mu_x][z_j(x) - \mu_x].$$

Likewise, indices of phenotypic covariance can be defined for character y,

$$C_{ij}(y,y) = [z_i(y) - \mu_y][z_j(y) - \mu_y],$$

and for character x in individual i and character y in individual j,

$$C_{ij}(x,y) = [z_i(x) - \mu_x][z_j(y) - \mu_y].$$

The expected values of these quantities are equal to the products of the relationship coefficient for the two individuals and the respective genetic variances or covariance:

$$E[C_{ij}(x,x)] = r_{ij}\,\sigma_A^2(x), \tag{5a}$$

$$E[C_{ij}(y,y)] = r_{ij}\,\sigma_A^2(y), \tag{5b}$$

$$E[C_{ij}(x,y)] = r_{ij}\,\sigma_A(x,y). \tag{5c}$$

Letting $\hat{}$ denote an estimate, the preceding expressions suggest that linear regressions of $\hat{C}_{ij}(x,x)$, $\hat{C}_{ij}(y,y)$, and $\hat{C}_{ij}(x,y)$ on $\hat{r}_{ij}$ fit through the origin will provide least-squares estimates of the genetic variances and covariance:

$$\mathrm{Var}_A(x) = \frac{\sum \hat{r}_{ij}\,\hat{C}_{ij}(x,x)}{\sum \hat{r}_{ij}^2}, \tag{6a}$$

$$\mathrm{Var}_A(y) = \frac{\sum \hat{r}_{ij}\,\hat{C}_{ij}(y,y)}{\sum \hat{r}_{ij}^2}, \tag{6b}$$

$$\mathrm{Cov}_A(x,y) = \frac{\sum \hat{r}_{ij}\,\hat{C}_{ij}(x,y)}{\sum \hat{r}_{ij}^2}, \tag{6c}$$

where the summations are over pairs of individuals. Although these three estimates are all downwardly

https://doi.org/10.1017/S0016672399004243 Published online by Cambridge University Press

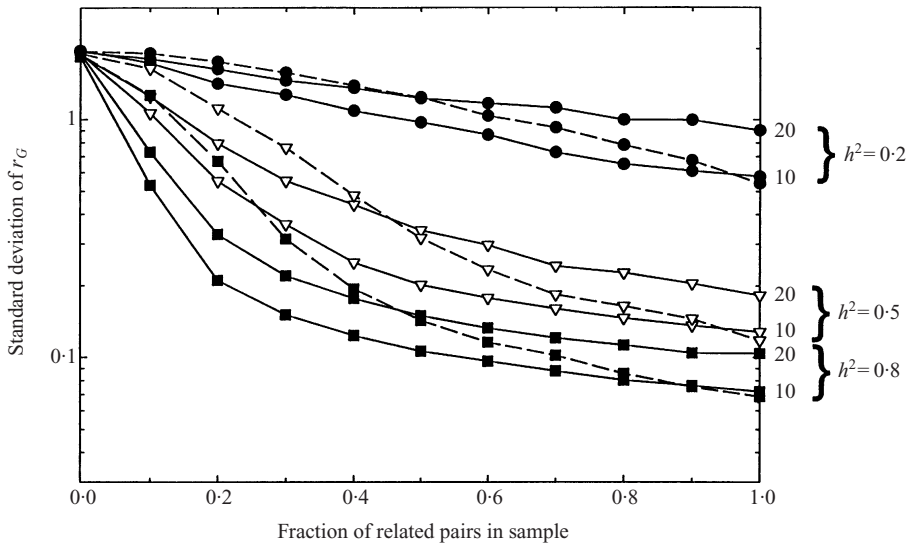

Fig. 4. Sampling standard deviation of the genetic correlation in marker-assisted analysis for two conditions involving equal effort in genotyping – 20 marker loci and 250 pairs of individuals, and 10 marker loci and 500 pairs of individuals (continuous lines). Five equally frequent alleles/loci are assumed to be in Hardy–Weinberg and gametic-phase equilibrium. Results are given for three different heritabilities (assumed to be the same for both traits) as a function of the fraction of related pairs of individuals (full sibs) in the total sample. The genetic correlation is equal to 0·50 in all cases, and the environmental correlation is equal to 0·00. The dashed lines give the results for analysis using 500 pairs of individuals and no markers, and comparison of them with the lines denoted by 10 provides insight into the reduction in the standard deviation that is expected with the addition of 10 molecular markers to the analysis.

biased by sampling variance in the estimates of r (Ritland, 1996), the magnitude of the bias is identical in all three cases. Thus, a potentially unbiased estimate of the genetic correlation is given by

$$\hat{\rho}_G = \frac{\sum \hat{r}_{ij}\,\hat{C}_{ij}(x,y)}{\sqrt{[\sum \hat{r}_{ij}\,\hat{C}_{ij}(x,x).\sum \hat{r}_{ij}\,\hat{C}_{ij}(y,y)]}}, \qquad (7)$$

which is equivalent to the estimator presented by Ritland (1996).

Statistical properties

To evaluate the potential utility of marker-assisted analysis in the estimation of genetic correlations, sets of computer simulations identical in all respects to those described above were run, with the addition of molecular markers in the surveyed individuals. In all cases reported on here, the marker loci were assumed to have five co-dominant alleles in equal frequencies (i.e. 0·20) and to be autosomal, unlinked, and in Hardy–Weinberg and gametic-phase equilibrium. The pairs of sampled individuals were again assumed to consist of either non-relatives or full sibs, and upon drawing the multilocus marker genotype of one individual, the markers in the other member of the pair were drawn conditional on the relationship. A number of methods exist for the estimation of relatedness with co-dominant markers. The following analyses utilize the estimator of Lynch & Ritland (1999), which is computationally simple and, relative to other estimators, has near minimal sampling variance.

The properties of bias with marker-assisted estimation are very similar to those illustrated above for analysis in the absence of markers. That is, there is downward bias if the incidence of pairs of related individuals is less than 20% or so, and upward bias in other cases, with the latter being generally very minor unless the heritabilities of the traits are very low (data not shown). Of greater interest is the behaviour of the sampling error.

Results for two sets of combinations of numbers of marker loci and numbers of pairs of individuals, both involving the same total amount of effort in genotyping, are given in Fig. 4 (continuous lines) – 20 marker loci and 250 pairs, and 10 marker loci and 500 pairs. It is clear that a higher degree of precision is achieved by maximizing the number of pairs at the expense of markers. A comparison of the results for 10 marker loci with the average results for zero marker loci, with 500 pairs being assayed in both cases (the dashed lines in Fig. 4, taken from Fig. 2), clarifies the relative advantages of a marker-based approach. If the trait heritabilities are moderately high and the fraction of pairs of related individuals in the sample is fairly low, then dramatic reductions in the standard deviation of $\hat{\rho}_G$ arise when the phenotypic information is supplemented with marker information. However, as the fraction of pairs of related individuals increases, the gain in precision from the use of markers becomes diminishingly small. When the fraction of related

https://doi.org/10.1017/S0016672399004243 Published online by Cambridge University Press

individuals exceeds a threshold (generally $> 90\%$), the use of markers actually induces a small increase in the sampling standard deviation.

4. Discussion

The goal of this paper has been to explore the feasibility of some new methods for estimating the genetic correlations between characters expressed in individuals in completely undisturbed natural populations. Previous attempts to estimate quantitative-genetic parameters in natural populations include cross-fostering procedures (reviewed on pp. 696–700 in Lynch & Walsh, 1998) and regression of phenotypes of laboratory reared progeny on wild parents (Coyne & Beecham, 1987; Riska *et al*., 1989). However, there are few organisms other than nest-box inhabiting birds for which the first method can be applied, and the second method can yield extremely biased results in the presence of genotype × environment interaction. In addition, both methods require accurate information on parentage for the assayed individuals, which can be difficult to near impossible to acquire for many organisms. For many species, such as long-lived trees, there are added problems concerning the time-scale necessary for conventional quantitative-genetic investigation.

The procedures outlined above provide a potential means for estimating genetic correlations in the absence of any direct observations on relatedness and without requiring any manipulations of individuals (other than those necessary for obtaining phenotypic measurements, and for obtaining molecular-marker profiles in the case of marker-assisted analysis). Provided certain conditions are fulfilled and large sample sizes are available, it appears that genetic correlations can be estimated successfully with collections of paired individuals taken from natural environments. Two key issues are the spatial scale of sampling of individuals and the optimal allocation of effort to sampling markers as opposed to individuals.

Both proposed methods require samples containing a moderate number of relatives. Generally, related individuals tend to be more closely associated geographically than random members of a population. Thus, a logical sampling scheme would involve the procurement of data on pairs of individuals that are not so distant from each other as to eliminate any possibility of relatedness. On the other hand, a key assumption of the proposed methods is that individuals exhibit phenotypic similarity only because of shared genes. Shared environmental effects can be important in well-designed laboratory experiments, due for example to shared maternal environment, but they may be especially important in samples drawn from natural populations where individuals are non-randomly distributed across the landscape. Such effects can bias estimates of genetic variances and/or covariances by causing phenotypes of relatives to be more similar than expected on the basis of genes alone.

There are a number of ways in which the influence of shared environmental effects can be ascertained and minimized in the analysis of natural populations. For example, as noted above, such effects would be expected to result in a dependence of phenotypic covariance on geographic distance separating pairs of individuals. Quantification of the magnitude of such spatially dependent effects would be difficult in situations in which there is no information on relatedness, because geographic distance will typically be correlated with both relatedness (shared genes) and shared environment, thereby confounding the two. However, in a marker-assisted study, it should be possible to isolate the two effects by performing a joint regression of pairwise phenotypic covariance on the estimated degree of relatedness and geographic distance. Consider, for example, the regression of $C_{ij}(x, x)$ (the estimated phenotypic covariance of character x in pairs of individuals) on estimated relatedness (r_{ij}) and physical distance (d_{ij}):

$$C_{ij}(x,x) = a + b_r r_{ij} + b_d d_{ij} + e_{ij}$$

(cf. Ritland, 1996). The estimated regression coefficient b_r is proportional to the genetic variance for character x and is unbiased by shared environmental effects (under the assumption that such effects decline linearly with distance), whereas the coefficient b_d quantifies the degree of spatially dependent phenotypic similarity. In principle, other indicator variables (differences in temperature, light intensity, etc.) would be added to such a regression, and the independent variables could be transformed to allow for non-linear responses. Alternatively, with certain types of organisms, it may be possible artificially to transplant pairs of individuals of fixed relatedness (e.g. full sibs or clonemates) into sites with a range of geographic distances in order directly to quantify the effect of spatial location on pairwise phenotypic covariance.

A second assumption in the preceding analyses is that all members of the population are equivalent with respect to the ability to express the characters under consideration. In reality, differences in mean phenotypes may exist among members of the different sexes, among individuals inhabiting different microhabitats within the total sampling area, and for populations with overlapping age distributions, among individuals in different cohorts. In principle, all such differences may be eliminated by the computation of fixed effects associated with sex, spatial location and time. Once the individual phenotypic measures are standardized with respect to these fixed effects, the pairwise comparisons outlined above can then be computed and subjected to the proposed methods of analysis.

Assuming that problems of shared environmental effects and fixed effects can be dealt with adequately, one is still confronted with the fact that estimates of the genetic correlation in natural populations can have much greater sample size requirements than those derived from controlled experiments with sets of known relatives. For many situations (e.g. most vertebrates and long-lived plants), it is actually much easier to procure large collections of data on individual phenotypes in the field than to perform large numbers of controlled matings, but a further limitation of methods for estimating the genetic correlation without known pedigrees is the need for a minimum fraction of pairs of relatives in the total sample in order to avoid substantial downward bias. The results noted above suggest the need for on the order of 20% for samples of 500 pairs consisting of either non-relatives or full sibs, and limited simulations suggest a similar threshold value when the sample consists of non-relatives and half sibs. Such levels may be difficult to achieve when samples are random over the entire range of a population. However, when the biology of the study population is reasonably well understood, it will often be possible to enrich the sample with pairs of relatives. For example, exploratory molecular analysis can provide a very useful means for determining the distribution of relatedness in the field, and hence for identifying the geographic scale beyond which pairs of individuals are unlikely to be close relatives (Lynch & Ritland, 1999).

The simulation results presented above indicate that although estimates of the genetic correlation can be biased downwardly when the incidence of relatives is low, upward bias occurs when the incidence of relatives is high. However, this upward bias, which is most pronounced when heritabilities are low and genetic correlations are high, is not unique to the methods introduced herein, as it even occurs in the ideal case in which all pairs consist of related individuals. Although such bias has been noticed (Van Vleck & Henderson, 1961; Brown, 1969), it has received little attention in previous studies of the statistical behaviour of the genetic correlation, and given its magnitude relative to the sampling variance of estimates, however, it will not generally be an overwhelming concern.

Despite the fact that marker-based estimates of pairwise relatedness are notoriously noisy (Lynch & Ritland, 1999), the supplementation of a field study on the genetic correlation with information on molecular markers can lead to a substantial gain in precision of estimates, particularly when a large segment of the sample consists of non-relatives. For example, when the incidence of close relatives in a sample is on the order of 30%, the standard deviation of estimates can be reduced by as much as 50% when only 10 markers are used to infer relatedness. More markers will further increase the degree of precision. However, when there is a tradeoff between the number of markers and the number of individuals that can be assayed, it appears that greater precision is achieved if more effort is put into sampling individuals and less into sampling loci. Multilocus DNA profiles involving co-dominant markers, such as those generated by DNA-fingerprinting, may be useful in this regard, as they generate large amounts of data that, when applied to the formula of Lynch (1988), can yield estimates of r that can be equally or more precise than those generated with locus-specific probes (Lynch & Ritland, 1999). In addition, a substantial amount of efficiency can be gained by utilizing loci with large numbers of alleles. For distant relatives, the sampling variance of relatedness estimates derived with the method of Lynch & Ritland (1999) is approximately $1/[n(m-1)]$, where n is the number of loci and m the number of alleles per locus, so a doubling in the number of alleles per locus reduces the sampling variance of r by nearly 50%.

It should be noted that the gain in precision from the incorporation of molecular markers can often be minor. In extreme cases, when most pairs of individuals consist of relatives, the use of markers can actually be counterproductive. This latter behaviour is a consequence of the crudeness of estimates of relatedness unless very large numbers of multiallelic loci are assayed. When most pairs of individuals are related, the error in inference of relatedness can overwhelm the small gain in precision that would occur with perfect knowledge of pedigrees.

To provide insight into the principles underlying the estimation of genetic correlations in natural populations, the examples presented above contained many simplifications including the assumptions that all pairs of individuals contain either non-relatives or full sibs, that the molecular markers all have equal allele frequencies, that the environmental correlation between-individuals is equal to zero, etc. In reality, samples from natural populations will generally present a range of relationships, marker loci will exhibit a diversity of uneven allele-frequency distributions, and shared environmental effects will be non-zero. Although such conditions are likely to result in higher sampling error of the genetic correlation than noted above and therefore call for even larger sample sizes, they do not alter the utility of the basic methodology.

Finally, it should be noted that the sampling standard deviations of the genetic correlation obtained by computer simulation in this study denote the expected standard deviation of estimates derived from replicate samples taken from the same population. Empirical studies almost always rely on single samples. Although analytical expressions exist for the sampling variance of the genetic correlation under standard

balanced designs (Lynch & Walsh, 1998, chapter 21), for the most part these are only rough approximations, and they are not easily extended to natural populations with heterogeneous mixtures of relationships.

For the current purpose, bootstrap analysis would appear to be a reasonable approach for constructing confidence limits – either sampling over n individuals and randomly constructing pairs when the total sample has been acquired randomly, or sampling over $n/2$ fixed pairs when pairs have been intentionally selected based on suspected relationships. Each resampling will generate different sets of $n/2$ pairs of relatives, so that averaging estimates over the full resampling procedure should fully utilize the information contained within the entire sample, while avoiding non-independence problems that would arise from a single analysis involving all $n(n-1)/2$ possible pairs of individuals.

Helpful comments were provided by W. Bradshaw, W. Hill, D. Roff, and two anonymous reviewers. This work was supported by NSF grant DEB-9629775.

References

Brown, G. H. (1969). An empirical study of the distribution of the sample genetic correlation coefficient. *Biometrics* **22**, 63–72.

Coyne, J. A. & Beecham, E. (1987). Heritability of two morphological characters within and among natural populations of *Drosophila melanogaster*. *Genetics* **117**, 727–737.

Falconer, D. S. & Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics*. 4th edn. Harlow, Essex: Longman.

Hill, W. G. & Thompson, R. (1978). Probabilities of non-positive definite between-group or genetic covariance matrices. *Biometrics* **34**, 429–439.

Klein, T. W. (1974). Heritability and genetic correlation: statistical power, population comparisons, and sample size. *Behavioral Genetics* **3**, 355–364.

Lynch, M. (1988). Estimates of relatedness by DNA fingerprinting. *Molecular Biology and Evolution* **5**, 584–599.

Lynch, M. & Ritland, K. (1999). Estimation of pairwise relatedness with molecular markers. *Genetics* **152**, 1753–1766.

Lynch, M. & Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sunderland, MA: Sinauer.

Riska, B., Prout, T. & Turelli, M. (1989). Laboratory estimates of heritabilities and genetic correlations in nature. *Genetics* **123**, 865–871.

Ritland, K. (1996). A marker-based method for inferences about quantitative inheritance in natural populations. *Evolution* **50**, 1062–1073.

Roff, D. A. (1997). *Evolutionary Quantitative Genetics*. New York: Chapman and Hall.

Van Vleck, L. D. & Henderson, C. R. (1961). Empirical sampling estimates of genetic correlations. *Biometrics* **17**, 359–371.

Visscher, P. M. (1998). On the sampling variance of intraclass correlations and genetic correlations. *Genetics* **149**, 1605–1614.