GENOTYPIC CORRELATION AND REGRESSION IN SOCIAL GROUPS: MULTIPLE ALLELES, MULTIPLE LOCI AND SUBDIVIDED POPULATIONS

PEKKA PAMILO[1]

School of Zoology, University of New South Wales, Kensington, New South Wales 2033, Australia

Manuscript received September 20, 1983
Revised copy accepted January 6, 1984

ABSTRACT

Genotypic correlations and regressions can be estimated from multiallelic data sets either by weighting the allelic effects additively or by specifically weighting the genotypic interactions. Both methods can be extended to multiple loci, but they do not fully take into account the joint segregation patterns at the loci. These genotypic statistics have a great importance in sociobiological contexts, as they can be used for genetic descriptions of social structures. In this paper I examine the two estimation methods and show how the genotypic correlation and regression coefficients from genotypic interactions are connected to other statistics of standard population genetics; special emphasis is given to the sample-size correction when intracolony correlations from small samples were estimated. I also show how genotypic correlation and regression can be estimated in subdivided populations, both in continuous populations with isolation by distance and in populations divided into separate subpopulations. The latter analysis is an example of a more general hierarchic correlation analysis.

GENEALOGICAL relationships lead into gene identities, and differing relationships give rise to varying probabilities of sharing identical genes. Such a probability can be calculated from known pedigrees, or it can be estimated as a correlation coefficient from frequency data. The gametic correlation between a mating pair is interpretable as an inbreeding coefficient of the offspring, and genotypic correlation between two individuals is known as the coefficient of relationship (WRIGHT 1922). The coefficient of relationship can be interpreted in terms of the probability of the individuals sharing identical genes in common. In recent sociobiological literature, it has proved important to estimate this probability unidirectionally, because we are interested in the following problem: If one individual helps another, what is the probability that the individual being helped will transmit to its offspring genes carried by the helper? (HAMILTON 1964). This unidirectional relationship can be measured by genotypic regression (HAMILTON 1972; MICHOD and HAMILTON 1980; UYENOYAMA and FELDMAN 1981). The importance of genotypic regression is

[1] Permanent address: Department of Genetics, University of Helsinki, P. Rautatiekatu 13, SF-00100 Helsinki 10, Finland.

that it can be used to predict the gene transmission and the change in gene frequencies in connection to the helping behavior (see MICHOD 1982 for a review).

There are two important aspects I want to point out before proceeding further. First, the correlation and regression coefficients convert into general pedigree estimates only in the absence of selection. When social evolution is studied, the locus affecting social behavior is under selection and leads to a different regression value than would a neutral marker locus (UYENOYAMA and FELDMAN 1981). Second, the coefficients are calculated as group estimates, and, hence, they depend on the group used as a reference. This has been worked out for gametic correlation, which can be estimated in subdivided populations using WRIGHT's (1943) well-known formula

$$1 - F_{IT} = (1 - F_{ST})(1 - F_{IS}). \quad (1)$$

In the present paper I examine how *genotypic* correlation and regression coefficients are estimated on the basis of neutral marker loci. HAMILTON (1972) discussed such measures and showed that genotypic correlation among group members in a subdivided population is

$$r = 2F_{ST}/(1 + F_{IT}) \quad (2)$$

(see also MICHOD and HAMILTON 1980; UYENOYAMA and FELDMAN 1981). It is clear from (2) that the choice of an appropriate measure to describe the population structure depends on whether one wants to estimate genetic differentiation and population-breeding structure separately (gametic correlation) or wants to include the effects of the breeding structure in the correlation estimate (genotypic correlation). When studying a population subdivided in separate subpopulations, we are interested in factors (such as drift and migration) affecting the allele frequency differentiation, and it seems best to describe the population structure using WRIGHT's F statistics. When a Mendelian population is subdivided in social groups, we are not primarily interested in allele frequency differentiation but in the genotypic basis of that subdivision. The genotypic estimates are appropriate for that case. Much of the current theory concerning social evolution refers to the population structure defined in genotypic terms, and although the correlation estimates at neutral loci differ from those at loci under selection, they provide useful information for considering thresholds of weak selection in the population with a given genotypic structure.

The theoretical background for calculating the coefficients of genotypic relationships is well developed for pedigrees (see *e.g.*, CANNINGS and THOMPSON 1982), but their estimation from genotype frequencies observed in natural populations has appeared problematic. There are two basic approaches for estimating genotypic relationships: the estimate can be based on weighting either the allelic effects or the genotypic interactions. My associates and I have earlier introduced a sample-size-corrected regression method for a biallelic locus based on weighting the allelic effects (PAMILO and CROZIER 1982), and STANTON (1960) has derived a correlation coefficient for multiple alleles by weighting genotypic interactions. Here, I show how STANTON's estimate is

connected to other statistics of standard population genetics and how it is affected by sample-size correction. Weighting of the allelic effects can also be done with multiple alleles, and both methods are extended to multiple loci. I also examine the estimation of genotypic correlation in subdivided populations in a way analogous to WRIGHT's (1951) F statistics (1).

GENOTYPE INTERACTIONS

Intracolony correlation: STANTON (1960) formulated genotypic correlation between two groups, Z and U, with the help of an interaction matrix in which the genotypes of the Z group form the rows, genotypes of the U group form the columns and the elements of the matrix consist of the frequencies of pairwise interactions between the genotypes in the two groups. The correlation coefficient is

$$r = \frac{\Sigma zu - \Sigma z \Sigma u}{\sqrt{[\Sigma z^2 - (\Sigma z)^2][\Sigma u^2 - (\Sigma u)^2]}} \tag{3}$$

where the symbols refer to allele, genotype and genotype-interaction frequencies, and the sums of z and z^2 are over the columns, the sums of u and u^2 over the rows and the sum of zu's is calculated over all the elements of the interaction matrix (STANTON 1960; CROZIER, PAMILO and CROZIER 1984). The sums are calculated by weighting the genotypic interactions with a special weighting scheme that depends on the number of alleles at the locus. If we modify STANTON's (1960) method to make an intragroup correlation coefficient, the matrix consists of interactions of an individual with the other group members and is symmetric. This means that intragroup correlation equals intragroup regression. I will next analyze the nature of these estimators and the effects of sample-size correction on them. My main interest is in estimating genetic relatedness among and between individuals living in social groups. I will call such a group a *colony* and, accordingly, use the term intracolony correlation.

If a sample from a given colony m has N_m individuals, there are $N_m(N_m - 1)/2$ interactions between them. Inserting these in the interaction matrix of a locus with s alleles ($s(s + 1)/2$ genotypes) and weighting each colony equally, we get after considerable algebra

$$(\Sigma z)^2 = (\Sigma u)^2 = c(1 - (s/(s - 1))H_{\text{exp}})$$

$$\Sigma z^2 = \Sigma u^2 = c(1 - (s/(s - 1))H_{\text{obs}}/2)$$

$$\Sigma zu = c - \frac{s}{s - 1} \Sigma \frac{1}{N_m - 1} \left(N_m h_{\text{exp},m} - \frac{1}{2} h_{\text{obs},m} \right)$$

where $h_{\text{exp},m} = 1 - \Sigma x_{i,m}^2$ and $h_{\text{obs},m} = 1 - \Sigma X_{ii,m}$ are the expected and observed heterozygosities within the colony m and H_{exp} and H_{obs} in the whole population. Note that the mean allele ($\bar{x}_i$) and genotype ($\bar{X}_{ij}$) frequencies are calculated by weighting the colonies equally. We can now insert these sums into (3) and

obtain

$$r = \frac{H_{exp} - \frac{1}{c}\Sigma h_{exp,m} - \frac{1}{c}\Sigma \frac{1}{N_m - 1}\left(h_{exp,m} - \frac{1}{2}h_{obs,m}\right)}{H_{exp} - \frac{1}{2}H_{obs}} \quad (4)$$

We see from (4) that r can be expressed in terms of expected and observed heterozygosities within the colonies and in the whole population. Similarly, CHAKRABORTY (1980) earlier noticed that the result does not depend on the number of alleles used for constructing the original weighting scheme introduced by STANTON (1960). The same formula (4) can also be used in the case of genetic dominance by considering all of the dominant phenotypes as homozygotes for the dominant allele, but one has to remember that the resulting correlation coefficient differs from that with intermediate heterozygotes. I should also remark that (4) does not give the genotypic correlation when the locus is affected by selection.

It is of interest to know how (4) is related to our earlier biallelic estimate of relatedness, which is given as a linear regression coefficient (PAMILO and CROZIER 1982) as

$$b = \frac{SS_{xy}}{SS_x} = \frac{\Sigma xy - \frac{1}{c}\Sigma x \Sigma y}{\Sigma x^2 - \frac{1}{c}(\Sigma x)^2} \quad (5)$$

where x refers to the allele frequency of a given individual and y to the allele frequency in the rest of the sample from that colony. The sums are calculated first over all individuals within a colony and then over the colonies weighting these equally. Using the same notations as (5) (except that we need no subscript to indicate the allele), we find that

$$\Sigma x = \Sigma y = c\bar{x}$$

$$\Sigma x^2 = c\left(\bar{x} - \frac{1}{4}H_{obs}\right)$$

$$\Sigma xy = \Sigma \frac{1}{N_m - 1}\left(N_m x_m^2 - x_m + \frac{1}{4}h_{obs,m}\right)$$

and inserting these in (5), we see that it equals (4) for two alleles.

When the sample size N_m from each colony is equal (*e.g.*, N), (5) reduces to (PAMILO 1982a)

$$b = \frac{N}{N-1}\frac{2F_{ST}}{1+F} - \frac{1}{N-1} \quad (6)$$

where

$$F_{ST} = \frac{s_x^2}{\bar{x}(1-\bar{x})} = \frac{\frac{1}{c} SS_x}{\bar{x}(1-\bar{x})}$$

$$F = 1 - H_{obs}/H_{exp}.$$

When N is very large, (6) approaches the limit given by (2). We can also show that, with equal sample size N, (4) reduces to (6) where F_{ST} is replaced by NEI's (1973) multiallelic estimator of F_{ST} (also called G_{ST})

$$F_{ST} = 1 - \frac{1}{c} \Sigma h_{exp,m}/H_{exp}. \tag{7}$$

We can thus conclude that our earlier biallelic estimate is a special case of STANTON's intracolony correlation and that the estimates approach the expected value when the sample size increases.

It can also be noted that combining (4) and (2) yields a sample-size-corrected estimator for multiallelic F_{ST}, denoted here by F_{ST}^*

$$F_{ST}^* = 1 - \frac{1}{H_{exp}} \frac{1}{c} \Sigma \frac{1}{N_m - 1} \left(N_m h_{exp,m} - \frac{1}{2} h_{obs,m} \right). \tag{8}$$

This can be compared with the estimator derived by NEI and CHESSER (1983). This part of (8) corresponding to $\frac{1}{c} \Sigma h_{exp,m}$ of (7) and generally denoted by H_S is identical with NEI and CHESSER's formulation when the samples are of equal size, but there is a slight difference between the two when N_m varies from colony to colony. NEI and CHESSER also derived an unbiased estimator for H_{exp}, but as they note, the ratio of two unbiased estimates does not necessarily lead to an unbiased estimate of the ratio.

The equation (8) can be derived in an alternative way as

$$F_{ST}^* = 1 - (1 - P_1)/(1 - P_2) \tag{9}$$

where P_1 is the probability of sampling an identical allele from two individuals in the same colony, and P_2 is the same probability for the whole population. To produce (8) we have to assume that P_2 approaches Σx_i^2, which is not exactly true in small samples (see NEI and CHESSER 1983), and

$$1 - P_1 = 1 - \frac{1}{c} \Sigma \frac{1}{N_m - 1} \left[\sum_{i<j} X_{ii} \left(x_i - \frac{1}{N_m} \right) + \frac{1}{2} X_{ij} \left(x_i + x_j - \frac{1}{N_m} \right) \right]$$

$$= \frac{1}{c} \Sigma \frac{1}{N_m - 1} \left(N_m h_{exp,m} - \frac{1}{2} h_{obs,m} \right)$$

and we see that (9) equals (8).

Intergroup correlation and regression: In intracolony calculations STANTON's interaction matrix is symmetric, and correlation is the same as regression. In

intergroup comparisons this is no longer true, because we now recognize two parties (*e.g.*, *X* and *Y*) which may, or may not, be from the same colony. Let us take an example in which genotypic regression is calculated between pairs *X* and *Y*, and let $x_{i,m}$ and $y_{i,m}$ denote the frequencies of allele *i* in the *m*th pair of *X* and *Y*. When we construct an interaction matrix and apply (3) to calculate regression of *Y* on *X*, the resulting regression coefficient appears to take a form similar to the intracolony coefficient (4), if we replace Σx_i^2 by $\Sigma x_i y_i$ and $\Sigma\Sigma x_{i,m}^2$ by $\Sigma\Sigma x_{i,m} y_{i,m}$, the summations taking place over the alleles and (in the latter) over the *X*, *Y* pairs. There is no need for the sample-size correction term that appeared in (4), and the regression coefficient of *Y* on *X* is

$$b_{YX} = \frac{\frac{1}{c}\Sigma\Sigma x_{i,m} y_{i,m} - \Sigma \bar{x}_i \bar{y}_i}{H_{\text{exp},X} - \frac{1}{2} H_{\text{obs},X}}. \tag{10}$$

In our earlier studies we have calculated intergroup regression from biallelic data as a linear regression between the allele frequencies in *X* and *Y* (PAMILO and VARVIO-AHO 1979). For that coefficient (when calculated only for one allele) $SS_{xy} = \Sigma x_m y_m - \frac{1}{c}\Sigma x_m \Sigma y_m$, $SS_x = \frac{c}{2}(H_{\text{exp},X} - H_{\text{obs},X}/2)$, and we see immediately that the resulting regression coefficient equals that of (10).

Genotypic correlation between *X* and *Y* is obtained as the geometric mean of the regression coefficients b_{YX} and b_{XY}.

Multiple loci: If there are several polymorphic loci, it would be of interest to use their joint segregation pattern in estimating genotypic correlations. Unfortunately, the preceding method does not allow that, and the multilocus estimate has to be computed using the loci as independent units. We can calculate the means of F_{ST} (or its equivalents in equations 4 and 10) and *F* over the loci and then apply (2). This procedure assumes that the loci are expected to give identical information, *i.e.*, the segregation patterns are not disturbed by different selection pressures, geographic differentiation, etc. Note that the multilocus data for this averaging method need not even come from the same individuals.

ALLELIC WEIGHTING

An alternative way to estimate genotypic correlation and regression is the approach used in quantitative genetics (FISHER 1918; FISHER 1958, pp. 30–37; FALCONER 1960, pp. 112–128). The alleles are given specific values (a_i for allele *i*), and the genotypic value is taken as the sum of the allelic values (haploid males at sex-linked loci can be treated as homozygotes). Correlation and regression can now be calculated on the basis of these genotypic values. Note that setting $a_i = a_j$ means pooling the alleles *i* and *j* together, and if this is not wanted we must have $a_i \neq a_j$. For a biallelic locus this method is equivalent to the weighting scheme used by PAMILO and CROZIER (1982).

Intergroup calculations can be done using the mean genotypic values of the *X* and *Y* groups in each colony, and the intracolony correlation (generally called

intraclass correlation) can be estimated by several alternative methods (DONNER and KOVAL 1980; KARLIN, CAMERON and WILLIAMS 1981). I will next discuss two commonly used methods, the ANOVA method and the sib-pair method.

If the sample size from each of c colonies is the same and equal to N, the intracolony correlation coefficient from the ANOVA method is

$$r = \frac{s_A^2 - s_W^2}{(N-1)s_W^2 + s_A^2} \tag{11}$$

where s_A^2 and s_W^2 are the estimates of the among-colony and within-colony variances. For a biallelic locus, (11) can be written (PAMILO 1982a)

$$r = \frac{N}{N-1}\frac{c}{c-1}\frac{2F_{ST}}{1+F+2F_{ST}/(c-1)} - \frac{1}{N-1} \tag{12}$$

which results in slightly different values than (4) and (6), especially when c is small and F_{ST} is high. With an increasing number of colonies, (12) will approach the value obtained from (4) and hence, with large values of N, the expected value based on (2).

The sib-pair method calculates the correlation coefficient using all pairwise comparisons between colony members,

$$r = \Sigma\Sigma\Sigma(x_{mi} - \bar{x})(x_{mj} - \bar{x})/(cN(N-1)s_x^2) \tag{13}$$

where x_{mi} is the genotypic value of the individual i in colony m, and $\bar{x}$ and s_x^2 are the sample mean and variance. When the sample size is the same, N, from each colony, the correlation coefficient calculated from (13) is the maximum likelihood estimator (ROSNER, DONNER and HENNEKENS 1977). It is easy to show that we get the same answer if, instead of calculating the correlation between pairs x_{mi} and x_{mj}, we use pairs x_{mi} and $(Nx_m - x_{mi})/(N-1)$, where x_m is the mean value in colony m. This is the approach used by PAMILO and CROZIER (1982).

The problem of allelic weighting with multiple alleles and multiple loci is that the result depends on the weighting scheme, although the estimates from different weighting schemes should converge when large sample sizes are used. It might be good to rotate the weights, so that the alleles are given the same set of weights in different orders, and then use the mean of the obtained correlation estimates. As we are dealing with the Mendelian segregation of genotypes, we have no epistatic interactions, and such a model can assume simple additive effects of the loci (ANDERSON and KEMPTHORNE 1954). If the alleles have values a_i at locus A and b_i at locus B, the genotypic value of an individual $A_iA_jB_kB_l$ is $(a_i + a_j + b_k + b_l)/2$. Calculating correlations and regressions for multilocus values is otherwise identical with the single-locus procedure.

SUBDIVIDED POPULATIONS

The genotypic correlation and regression estimates depend on the allele frequency variance among the colonies, and this variance increases when there is allele frequency differentiation within the study area, *i.e.*, when the colonies

do not come from the same gene pool. The effect of population subdivision has been worked out for gametic correlation (1) (WRIGHT 1943, 1951; COCKERHAM 1973), and in this section I will examine the effects of subdivision on intracolony genotypic correlation.

Let us first examine a continuous population with isolation by distance. If we assume that there is no local inbreeding, the observed inbreeding coefficient F measures the areal allele frequency differentiation, and the standardized allele frequency variance among the colonies, F_{CT}, is a function of F. If we denote by F_{CS} the estimate that we would have without gene frequency differentiation, we can partition the F_{CT} estimate (see WRIGHT 1943; JACQUARD 1975; HARPENDING 1979) by dividing it into two components following WAHLUND's (1928) principle: that due to the subdivision of the population into colonies and that due to geographic allele frequency differentiation

$$F_{CT} = (1 - F)F_{CS} + F. \tag{14}$$

If we assume that there is no inbreeding (no positive assortative mating between relatives), the genotypic correlation without gene frequency differentiation would be, from (2), $b_{CS} = 2F_{CS}$. This can be now written on the basis of (14)

$$\begin{aligned} b_{CS} = 2F_{CS} &= 2(F_{CT} - F)/(1 - F) \\ &= (b_{CT} - 2F/(1 + F))/(1 - 2F/(1 + F)). \end{aligned} \tag{15}$$

This result can be applied in continuous populations assuming that the social structure remains the same in different parts of the population. The equation (15) allows a comparison with WRIGHT's (1943) classic formula for subdivided populations, my equation (1). When WRIGHT's formula and (15) are compared, we find that $1 - F_{CS} = 1/(1 - F_{IC})$ or $F_{CS} = -F_{IC}/(1 - F_{IC})$, where F_{IC} is the mean gametic correlation within the colonies.

The equations (14) and (15) do not hold if there is local inbreeding, because in that case F reflects not only local differentiation but also the assortative mating between the relatives. This effect can be taken into account if we can distinguish between the two components in F. This can be done with the help of (1) if we have a population divided into a number of separate subpopulations and each subpopulation has a number of colonies. Let us denote F_{IT} as the total inbreeding coefficient, F_{IS} as the mean inbreeding coefficient within the subpopulations, F_{ST} as the standardized allele frequency variance among the subpopulations, F_{CT} as the standardized allele frequency variance among all of the colonies in the population and F_{CS} as the standardized allele frequency variance among the colonies in the same subpopulation. From (14)

$$F_{CT} = (1 - F_{ST})F_{CS} + F_{ST}.$$

and

$$\begin{aligned} r_{CS} = \frac{2F_{CS}}{1 + F_{IS}} = \frac{2(F_{CT} - F_{ST})}{1 + F_{IT} - 2F_{ST}} &= \frac{\dfrac{2F_{CT}}{1 + F_{IT}} - \dfrac{2F_{ST}}{1 + F_{IT}}}{1 - \dfrac{2F_{ST}}{1 + F_{IT}}} \\ &= (r_{CT} - r_{ST})/(1 - r_{ST}) \end{aligned} \tag{16}$$

where r_{CT} is the measured intracolony correlation in the whole population, r_{ST} the correlation among the subpopulations and r_{CS} the estimate of the intracolony correlation within subpopulations. Because F_{CT} and F_{IT}, when calculated from the pooled data, are weighted by the number of colonies sampled from each subpopulation, F_{ST} should also be calculated by weighting the subpopulations according to the number of colonies. This means that r_{ST} will become similarly weighted. Equation (16) can also be written as

$$1 - r_{CT} = (1 - r_{ST})(1 - r_{CS}) \tag{17}$$

which is of the same form as that for gametic correlation (1). It should be noted that here r_{CS} includes the effects of inbreeding within the subpopulations. Equation (17) is an example of a general property of intraclass correlation analysis in a hierearchically organized material as easily demonstrated from the expected values of the intraclass correlation coefficient at various levels of the hierarchy (*e.g.*, KEMPTHORNE 1957, pp. 243–244).

APPLICATION TO DATA

Unequal sample sizes: I have discussed only the cases in which each colony is weighted equally. It may sometimes be of interest to weight the colonies according to the colony size. Assume that we are studying an ant population with a varying number of queens in the colonies and we estimate the genotypic correlation among the coexisting queens. When weighting each colony equally, we get an estimate of average relatedness per nest, and when weighting according to the queen number, we get an estimate of average relatedness per queen in that population.

One problem in equal weighting of colonies with unequal sample sizes is that small samples can bias the results because of large sampling errors associated to them. On the other hand, if each colony is weighted by the sample size, large samples get much weight in determining the sample means and variances and can bias the results. Thus, it seems desirable to use equal sample sizes or to take such samples that the smallest one is large enough to eliminate strong sampling errors and weight all the samples equally. Another possibility would be to set an upper limit, *e.g.*, the median sample size, and weight all of the samples greater than that by this limit value and the smaller samples according to their sample size. The problems associated to unequal sample sizes are also discussed by DONNER and KOVAL (1980) and KARLIN, CAMERON and WILLIAMS (1981), and the present methods are easily modified to allow unequal weighting. It may be useful to calculate both weighted and unweighted estimates and base the conclusions on both of these and on the difference between them.

Variance estimates: The variances of the correlation and regression estimates can be obtained by subsampling techniques (see CROZIER, PAMILO and CROZIER 1984). I have examined two subsampling techniques, jackknife and bootstrap (EFRON 1981), by computer simulations similar to those used by PAMILO and CROZIER (1982) and CROZIER, PAMILO and CROZIER (1984). The jackknife technique forms c subsamples from the original data of c colonies by leaving out one colony at a time, and the bootstrap technique can be used to create any number of subsamples drawn from an imaginary population having a distribution

identical with that observed. According to EFRON (1981), bootstrapping gives better results than jackknifing in the case of parametric linear correlation, but my Monte Carlo simulations of STANTON's (1960) interaction method showed that jackknifing was superior to bootstrapping (with c to $4c$ subsamples). The superiority of jackknifing is based on two observations: (1) the ratio of the mean standard error estimate (s_1) from 50 simulations to the standard deviation of the point estimate (s_2) from the same 50 simulations was generally closer to 1.0, and (2) the coefficient of variation (CV) of the standard error estimate was smaller than in bootstrapping. Hence, I prefer the jackknife technique for estimating variances.

Using the jackknife technique, I next compare the interaction and the sib-pair allelic weighting methods of estimating genotypic correlation. The allelic weighting is done by giving the allele a_i weight i, etc. The simulation results with multiple alleles show that STANTON's interaction method gives generally better results than allelic weighting (Table 1; the s_1/s_2 ratio closer to 1.0, smaller CV, and the point estimate generally closer to the expected value). The simulation results also show that the multilocus estimate based on the interaction method gives narrower confidence limits for the correlation coefficient than any of the single-locus estimates. The properties of the interaction method and jackknife technique in estimating genetic relatedness in colonies of social insects are examined more closely by CROZIER, PAMILO and CROZIER (1984).

We have applied the present methods in studying population structures in social insects (CROZIER, PAMILO and CROZIER, 1984). Another application of the genotypic correlation could be in estimating polyandry (see WILSON 1981; GRIFFITHS, MCKECHNIE and MCKENZIE 1982 for biallelic methods). In diploid organisms, the offspring of a single female, fathered by n unrelated males (each male contributing equally), have an expected genotypic correlation among them

$$r = 1/(2n) + (n - 1)/(4n). \tag{18}$$

SASSAMAN (1978) published mother-offspring genotype data (one locus with six alleles) from 20 broods of *Porcellio scaber*. From this material we get, using (4), genotypic correlation ($r \pm$ SE) among the offspring 0.382 ± 0.055, which corresponds to 1.90 effective inseminations per female. The point estimate obtained by WILSON (1981) using analysis of variance for the same data classified in biallelic form, 2.26, is somewhat greater than my present estimate but well within the range given by mean ± SE, 1.34–3.25.

Examples of subdivided populations: I will next apply the present methods in reanalyzing genetic data from subdivided populations of Formica ants. The data are based on electrophoretic variation at the *MDH-2* locus in four species: in island populations of *F. exsecta* and *F. fusca* and in continuous populations of *F. transkaucasica* and *F. sanguinea* (PAMILO 1983). The *sanguinea* population had five alleles; all of the others were biallelic. The calculations here are done using STANTON's interaction method.

In the island populations the genotypic correlation among workers in a single

TABLE 1

Results from simulated sampling of the model populations

Locuw and x_i's	One queen per nest								Five queens per nest							
	Stanton's method				Sib-pair method				Stanton's method				Sib-pair method			
	r	s_2	s_1/s_2	CV	r	s_2	s_1/s_2	CV	r	s_2	s_1/s_2	CV	r	s_2	s_1/s_2	CV
a) $N = 10$																
A 0.5 0.5	0.74	0.06	0.90	0.13	0.74	0.06	0.90	0.13	0.15	0.06	0.84	0.22	0.15	0.06	0.84	0.22
B 0.33 0.33 0.33	0.75	0.04	0.97	0.11	0.74	0.06	0.97	0.17	0.15	0.03	1.06	0.17	0.17	0.06	0.89	0.18
C 0.25 0.25 0.25 0.25	0.76	0.03	0.92	0.12	0.76	0.06	0.91	0.19	0.15	0.03	0.86	0.17	0.15	0.06	0.84	0.21
All three loci	0.75	0.03	0.86	0.12	0.75	0.05	1.03	0.16	0.15	0.02	0.93	0.14	0.16	0.06	0.79	0.21
A 0.8 0.2	0.75	0.06	1.19	0.25	0.75	0.06	1.19	0.25	0.16	0.05	1.01	0.30	0.16	0.05	1.01	0.30
B 0.8 0.1 0.1	0.75	0.06	1.10	0.23	0.74	0.07	1.13	0.24	0.15	0.04	1.11	0.27	0.16	0.05	1.10	0.32
C 0.8 0.067 0.067 0.067	0.76	0.05	1.14	0.18	0.76	0.08	1.01	0.28	0.15	0.04	0.88	0.32	0.15	0.06	0.82	0.34
All three loci	0.75	0.03	1.19	0.16	0.76	0.05	1.19	0.23	0.15	0.03	0.99	0.26	0.14	0.05	1.01	0.22
b) $N = 2$																
A 0.5 0.5	0.75	0.08	1.02	0.19	0.75	0.08	1.02	0.19	0.16	0.16	1.03	0.13	0.16	0.16	1.03	0.13
B 0.33 0.33 0.33	0.75	0.06	1.04	0.13	0.74	0.08	1.04	0.23	0.15	0.13	0.95	0.11	0.14	0.16	1.06	0.12
C 0.25 0.25 0.25 0.25	0.74	0.06	0.91	0.10	0.74	0.08	1.06	0.24	0.15	0.10	1.01	0.08	0.16	0.16	1.06	0.14
All three loci	0.75	0.04	0.97	0.12	0.77	0.06	1.19	0.30	0.15	0.08	1.08	0.10	0.16	0.16	1.01	0.13
A 0.8 0.2	0.73	0.12	0.86	0.31	0.73	0.12	0.86	0.31	0.13	0.18	0.97	0.19	0.13	0.18	0.97	0.19
B 0.8 0.1 0.1	0.76	0.10	0.94	0.27	0.75	0.12	0.94	0.35	0.16	0.14	0.99	0.21	0.13	0.19	0.92	0.26
C 0.8 0.067 0.067 0.067	0.76	0.09	0.99	0.24	0.75	0.13	0.97	0.35	0.14	0.14	1.00	0.24	0.11	0.20	0.86	0.27
All three loci	0.75	0.06	0.90	0.18	0.76	0.11	0.84	0.40	0.15	0.09	1.10	0.15	0.10	0.19	0.88	0.22

r is the observed mean correlation, the expected values of which are 0.75 for one queen and 0.15 for five queens; s_1 is the mean standard error from the simulations; s_2 is the standard deviation of the simulated point estimates, CV is the coefficient of variation of the standard errors in the simulations and N is the sample size.

nest ($r_{CT} \pm$ SE) from pooled material is 0.69 ± 0.04 in *exsecta* (121 nests) and 0.62 ± 0.06 in *fusca* (232 nests). The estimates of r_{ST} are 0.14 and 0.15, respectively, and the corrected correlations, according to (16), are $r_{CS} = 0.64$ in *exsecta* and $r_{CS} = 0.56$ in *fusca*. The estimate in *fusca* agrees well with those obtained from single islands (0.57 and 0.56, PAMILO 1983) and indicates either multiple mating or partial polygyny (*i.e.*, multiple queens) of the nests (the expected correlation for one single-mated queen per nest is 0.75). Single-island estimates in two island populations of *exsecta* were 0.62 and 0.78, and the genotype distributions in the nests agreed with the assumption of them having a single queen (PAMILO and ROSENGREN 1983). The obtained correlation estimates in this species are probably boosted by inbreeding within the islands (PAMILO and ROSENGREN 1983), and the results thus suggest either slight multiple mating or shared egg laying by two to several queens in some nests.

In *transkaucasica* and *sanguinea* I examined continuous populations but sampled them in a discontinuous way using a grid pattern (see PAMILO 1983), and it is possible to calculate the standardized allele frequency variance among the sampling plots. This F_{ST} was 0.11 in *transkaucasica* and 0.10 in *sanguinea*. The genotypic correlation among worker nest mates in the pooled material was $r_{CT} = 0.47 \pm 0.04$ in *transkaucasica* (161 nests) and 0.42 ± 0.03 in *sanguinea* (137 nests). Using (16) we get $r_{CS} = 0.34$ and 0.27, respectively, in the two species. In *transkaucasica*, I earlier estimated the genotypic correlation as 0.33 from a sample of 55 nests in a small area (PAMILO 1982b); the present result agrees well with this estimate. As there seems to be no local inbreeding, we can pool all of the 216 nests and get $r_{CT} = 0.45 \pm 0.03$, $F = 0.09 \pm 0.03$, and using (15) $r_{CS} = 0.33 \pm 0.04$, also in good agreement with the earlier result. In *sanguinea*, I also examined 137 nests in a core area of the same population from which the grid samples were taken. Within this core area $b = 0.42 \pm 0.03$, but this is probably boosted by local allele frequency differences (PAMILO 1983). However, the allele frequency differentiation within the core area did not yield a positive inbreeding coefficient, but there is an excess of heterozygotes. Thus, it seems probable that the nests have several queens sharing the egg laying; the level of multiple insemination detected in this species does not explain the observed level of relatedness (PAMILO 1982c). As there was no indication of increased allele frequency differentiation with distances greater than those within the core area, we might expect that the standardized allele frequency variance 0.10 also describes allele frequency differentiation in the core population. Using this value as our estimate of F and pooling all of the 266 nests together we get the estimates $r_{CT} = 0.42 \pm 0.02$ and, applying (16), $r_{CS} = 0.29$, very close to the value obtained from the grid samples alone.

The methods described here are probably best suited for social insects but have a great potential use in estimating average genetic relatedness in other organisms living in social groups. But, one has to remember that the average relatedness is a population estimate, and the actual relatedness is likely to vary among single colonies. Computer programs for STANTON's interaction method are available by writing to the author.

I thank R. H. CROZIER for introducing STANTON's article to me, and B-O. BENGTSSON, R. H.

CROZIER, M. NEI and the reviewers for critically reading earlier drafts of the manuscript. The work has been supported by a grant from the Australian Research Grants Committee to R. H. CROZIER.

LITERATURE CITED

ANDERSON, V. L. and O. KEMPTHORNE, 1954 A model for the study of quantitative inheritance. Genetics **39:** 883–898.

CANNINGS, C. and E. A. THOMPSON, 1982 *Genealogical and Genetic Structure.* University Press, Cambridge, Massachusetts.

CHAKRABORTY, R., 1980 Relationship between single- and multi-locus measures of gene diversity in a subdivided population. Ann. Hum. Genet. **43:** 423–428.

COCKERHAM, C. C., 1973 Analysis of gene frequencies. Genetics **74:** 679–700.

CROZIER, R. H., P. PAMILO and Y. C. CROZIER, 1984 Relatedness and microgeographic genetic variation in *Rhytidoponera mayri*, an Australian arid-zone ant. Behav. Ecol. Sociobiol. In press.

DONNER, A. and J. J. KOVAL, 1980 The estimation of intraclass correlation in the analysis of family data. Biometrics **36:** 19–25.

FALCONER, D. S., 1960 *Introduction to Quantitative Genetics.* Ronald Press, New York.

FISHER, R. A. 1918 The correlation between relatives on the supposition of Mendelian inheritance. Trans. R. Soc. Edinb. **52:** 399–433.

FISHER, R. A., 1958 *The Genetical Theory of Natural Selection*, Ed. 2. Dover, New York.

GRIFFITHS, R. C., S. W. MCKECHNIE and J. A. MCKENZIE, 1982 Multiple mating and sperm displacement in a natural population of *Drosophila melanogaster*. Theor. Appl. Genet. **62:** 89–96.

HAMILTON, W. D., 1964 The genetical evolution of social behaviour I. J. Theor. Biol. **7:** 1–16.

HAMILTON, W. D., 1972 Altruism and related phenomena, mainly in social insects. Annu. Rev. Ecol. Syst. **3:** 193–232.

HARPENDING, H. R., 1979 The population genetics of interaction. Am. Nat. **113:** 622–630.

JACQUARD, A., 1975 Inbreeding: one word, several meanings. Theor. Pop. Biol. **7:** 338–363.

KARLIN, S., E. C. CAMERON and P. T. WILLIAMS, 1981 Sibling and parent-offspring correlation estimation with variable family size. Proc. Natl. Acad. Sci. USA **78:** 2664–2668.

KEMPTHORNE, O., 1957 *An Introduction to Genetic Statistics.* Iowa State University Press, Ames.

MICHOD, R. E., 1982 The theory of kin selection. Annu. Rev. Ecol. Syst. **13:** 23–55.

MICHOD, R. E. and W. D. HAMILTON, 1980 Coefficients of relatedness in sociobiology. Nature **288:** 694–697.

NEI, M., 1973 Analysis of gene diversities in subdivided populations. Proc. Natl. Acad. Sci. USA **70:** 3321–3323.

NEI, M. and R. K. CHESSER, 1983 Estimation of fixation indices and gene diversities. Ann. Hum. Genet. **47:** 253–259.

PAMILO, P., 1982a Genetic structure of *Formica* populations. Ph.D. dissertation, University of Helsinki, Helsinki, Finland.

PAMILO, P., 1982b Genetic population structure in polygynous *Formica* ants. Heredity **48:** 95–106.

PAMILO, P., 1982c Multiple mating in *Formica* ants. Hereditas **97:** 37–45.

PAMILO, P., 1983 Genetic differentiation within subdivided populations of *Formica* ants. Evolution **37:** 1010–1022.

PAMILO, P. and R. H. CROZIER, 1982 Measuring genetic relatedness in natural populations: methodology. Theor. Pop. Biol. **21:** 171–193.

PAMILO, P. and R. ROSENGREN, 1983 Evolution of nesting strategies of ants: genetic evidence from different population types of *Formica* ants. Biol. J. Linn. Soc. In press.

PAMILO, P. and S. VARVIO-AHO, 1979 Genetic structure of nests in the ant *Formica sanguinea*. Behav. Ecol. Sociobiol. **6:** 91–98.

ROSNER, B., A. DONNER and C. H. HENNEKENS, 1977 Estimation of interclass correlation from familial data. Appl. Statistics **26:** 179–187.

SASSAMAN, C., 1978 Mating systems in porcellionid isopods: multiple paternity and sperm mixing in *Porcellio scaber* Latr. Heredity **41:** 385–397.

STANTON, R. G., 1960 Genetic correlations with multiple alleles. Biometrics **16:** 235–244.

UYENOYAMA, M. and M. W. FELDMAN, 1981 On relatedness and adaptive topography in kin selection. Theor. Pop. Biol. **19:** 87–123.

WAHLUND, S., 1928 Zusammensetzung von Population und Korrelationserscheinungen vom Standpunkt der Vererbungslehre aus betrachtet. Hereditas **11:** 65–106.

WILSON, J., 1981 Estimating the degree of polyandry in natural populations. Evolution **35:** 664–673.

WRIGHT, S., 1922 Coefficients of inbreeding and relationship. Am. Nat. **56:** 330–338.

WRIGHT, S., 1943 Isolation by distance. Genetics **28:** 114–138.

WRIGHT, S., 1951 The genetical structure of populations. Ann. Eugen. (Lond.) **15:** 323–354.

Corresponding editor: M. NEI