Am. J. Hum. Genet. 62:698–702, 1998

On the Probabilities of Identity States in Permutable Populations

C. Cannings

School of Mathematics and Statistics, University of Sheffield, Sheffield, United Kingdom

Summary

Génin and Clerget-Darpoux recently discussed the derivation of the probabilities of identity states for populations in which there was some degree of kinship, primarily to allow the extension of the classical affected-sib-pair method to such populations. It is argued here that their derivation makes certain assumptions that are valid only for some very restricted population models and that are not needed for an appropriate treatment. Here the probabilities of the identity states of two individuals with a given genealogical relationship are specified in terms of the kinship parameters of the underlying population, from which the founders of the individuals' genealogy have been randomly selected. It is argued that an appropriate representation for a permutable population, one in which gene identity does not depend on the pattern of genes across individuals, requires three parameters. This representation is related to that of Génin and Clerget-Darpoux and to that of Weir.

Introduction

In a recent paper, Génin and Clerget-Darpoux (1996) have sought to extend the sib-pair method of linkage analysis to the case in which the underlying population is consanguineous. In doing so, they require specification of the probabilities of the nine possible identity states ($S_1, \ldots, S_9$ in their notation—which I have adopted, albeit with some reordering) of two individuals drawn at random from the population. It is argued here that Génin and Clerget-Darpoux's set of probabilities are incorrect, in the sense that they do not correspond to any fully specified population model (except in certain very restricted and uninteresting cases). Additionally, it is demonstrated how one can derive, in a compact way, the probabilities of the gene-identity states for any pair of individuals and, in particular, for a sib pair whose parents have some specific relationship within a genealogy whose founders are drawn from a population with nonzero kinship. Earlier, Bishop and Williamson (1990) had extended the affected-sib-pair method to more-distant relationships, using Cotterman's (1940) k coefficients, but here we require the full identity states.

Received January 15, 1997; accepted December 30, 1997; electronically published March 13, 1998.

Address for correspondence and reprints: Dr. C. Cannings, School of Mathematics and Statistics, University of Sheffield, Sheffield S3 7RH, United Kingdom. E-mail: c.cannings@sheffield.ac.uk

0002-9297/98/6203-0025$02.00

Permutable Populations

A first point, although not the main one, of this article concerns the assumption made by Génin and Clerget-Darpoux—that, if a population is stable in the sense that the genetic structure does not change, then the coefficient of inbreeding, denoted by "α," is equal to the coefficient of kinship, denoted by "ϕ." This is not the case in general. For example, suppose that a population of infinite size has genetic makeup $\frac{1}{6}AA + \frac{2}{3}AB + \frac{1}{6}BB$ (here the letters denote alleles that are identical by descent), with the mating rule that each AA mates with a BB and that each AB mates with an AB. In the absence of fertility or viability differences, this population is stable, with $\alpha = \frac{1}{3}$, whereas the coefficient of kinship is $\phi = \frac{1}{4}$. On the other hand, if an infinite population mates at random, then it will be in Hardy-Weinberg equilibrium with respect to the frequency of alleles identical by descent, and $\alpha = \phi$. It might seem paradoxical to assert that a population in Hardy-Weinberg equilibrium has any inbreeding at all, but, if such a population (or an approximation of it) is created from a founder population that then expands rapidly, it makes sense to identify in that final population the distinct founder genes and to calculate appropriate kinship coefficients.

In a finite population, stability of coefficients of inbreeding and kinship, other than unity at fixation, will be achieved if there is mutation that introduces new "founder" genes. The infinite-alleles model is such a model: mutation occurs at a rate, μ, that is independent of the "target" gene, each new mutant is considered as a new founder gene, and any identity by descent requires that ancestry be traced back only to such an ancestral mutant. Weir (1994) discusses this model and gives the stationary coefficients of kinship for as many as four genes—that is, the state reached by such a population

Table 1

Probabilities of Identity States under the Génin and Clerget-Darpoux Model and under the Permutable Model

			Probability of Identity State	
Label	Identity State	Probability Designation	Génin and Clerget-Darpoux Model	Permutable Model
S_1	(1,1,1,1)	π_1	α^3	$\alpha^2\beta$
S_2	(1,1,2,2)	π_2	$\alpha^2(1-\alpha)$	$\alpha^2(1-\beta)$
S_3	(1,2,1,2)	π_3	$\alpha^2(1-\alpha)^2$	$2\alpha^2(1-\beta)$
S_4	(1,1,1,2)	π_4	$\alpha^2(1-\alpha)$	$\alpha(1-\alpha)\gamma$
S_5	(1,2,2,2)	π_5	$\alpha^2(1-\alpha)$	$\alpha(1-\alpha)\gamma$
S_6	(1,1,2,3)	π_6	$\alpha(1-\alpha)^2$	$\alpha(1-\alpha)(1-\gamma)$
S_7	(1,2,3,3)	π_7	$\alpha(1-\alpha)^2$	$\alpha(1-\alpha)(1-\gamma)$
S_8	(1,2,1,3)	π_8	$2\alpha(1-\alpha)^3$	$4\alpha(1-\alpha)(1-\gamma)$
S_9	(1,2,3,4)	π_9	$(1-\alpha)^4$	$1-6\alpha+3\alpha^2+4\alpha(1-\alpha)\gamma+2\alpha^2\beta$

after a long time. However, it is not true that, at equilibrium, $\alpha = \phi$ in a diploid population. In fact, $\alpha = (1-\mu)^2\phi$, since the two genes at a locus within a diploid individual are identical by descent only if those of the two randomly selected parents are identical and neither mutates during transmission. This fact is not noted by Weir, since he uses the haploid version of the model, in which there is no meaning to α. Of course, there is approximate equality if μ is small, as is usually the case.

Thus it is not true that $\alpha = \phi$ in general, and to assume so is to make some statement about the structure of the population. I have argued above that, for certain infinite-population models and, approximately, in some finite-population models, the assumption is reasonable. It is this class of models that I discuss here, since these accord with Génin and Clerget-Darpoux's assumptions. I consider a class that I term "permutable" and make the assumption that, if a set of genes is selected from the population and the probabilities of the various possible identity states are calculated, then these probabilities will not be dependent on how the genes were drawn with respect to diploid individuals; for example, for four genes, it is not important whether they were drawn from two, three, or four individuals.

The Probabilities of Identity States

I turn now to the derivation of probabilities of identity states of a random pair of individuals in a permutable population. Table 1 specifies the set of identity states, their labeling (the S_i), and the notation for their probabilities (π_i). Génin and Clerget-Darpoux (1996, Appendix A) do not give a derivation of their probabilities but would appear to have used a conditional argument, first taking the appropriate probabilities for the identity within individuals and then, conditional on these states, calculating the probability of the between-individual states. Thus, to calculate the probability of $S_1 = (1,1,1,1)$, one has a probability α^2 that the individuals have the correct within-individual identity (thus restricting them to $S_1 = (1,1,1,1)$ or $S_2 = (1,1,2,2)$) and then has the probability of $S_1 = \alpha^2\beta$, where β is the probability that the individuals are in state S_1, given that both of the individuals are inbred. Génin and Clerget-Darpoux give a final probability of α^3 for S_1, thus requiring that $\beta = \alpha$. In general, there is no reason to assume that $\beta = \alpha$. Furthermore, we have, summing appropriately over identity states, $\alpha = (\pi_1 + \pi_2 + \pi_4 + \pi_6)$, whereas

$$\phi = \pi_1 + \frac{1}{2}(\pi_3 + \pi_4 + \pi_5) + \frac{1}{4}\pi_8 = \alpha \ .$$

Génin and Clerget-Darpoux's expressions for the π's satisfy the first expression but not the second. There are other restrictions on the π_i ($\Sigma_i\pi_i = 1$, $\pi_4 = \pi_5$, $\pi_6 = \pi_7$, and $\pi_1 + \pi_2 = \alpha^2$), which are satisfied by the Génin and Clerget-Darpoux's expressions. On the other hand, permutability implies that $\pi_3 = 2\pi_2$ and that $\pi_8 = 4\pi_6$, which are not true for Génin and Clerget-Darpoux's expressions. These equalities follow from consideration of the arrangement of the four alleles among the pairs of individuals. Thus, if the four alleles are two 1's, one 2, and one 3, then the probability that the two 1's are assigned to the first individual, giving S_6, is $\frac{1}{6}$, whereas the probability that they are assigned one to each individual, giving S_8, is $\frac{2}{3}$ (hence, $\pi_8 = 4\pi_6$).

These restrictions reduce the df for the case of four alleles, to 3. One possible parameterization is to use α and β, as defined above, and $\gamma = P(S_4 \mid \text{individual 1 inbred} \cap \text{individual 2 not inbred})$. It is then easy to obtain the expressions given in the final column of table 1. It should be noted that Génin and Clerget-Darpoux used $\beta = \alpha$. It is difficult to think of conditions under which

Table 2

Identity States: Probabilities for Two and Three Genes

Label	Identity State	Probability Designation	Probability under Permutable Model
T_1	(1,1)	ρ_1	α_2
T_2	(1,2)	ρ_2	$(1 - \alpha_2)$
U_1	(1,1,1)	τ_1	α_3
U_2	(1,1,2)	τ_2	$(\alpha_2 - \alpha_3)$
U_3	(1,2,1)	τ_3	$(\alpha_2 - \alpha_3)$
U_4	(2,1,1)	τ_4	$(\alpha_2 - \alpha_3)$
U_5	(1,2,3)	τ_5	$(1 - 3\alpha_2 + 2\alpha_3)$

Table 3

Identity States: Probabilities for Four Genes

Label	Identity State	Probability Designation	Probability under Permutable Model
S_1	(1,1,1,1)	π_1	α_4
S_2	(1,1,2,2)	π_2	$(\alpha_2^2 - \alpha_4)$
S_3	(1,2,1,2)	π_3	$2(\alpha_2^2 - \alpha_4)$
S_4	(1,1,1,2)	π_4	$2(\alpha_3 - \alpha_4)$
S_5	(1,2,2,2)	π_5	$2(\alpha_3 - \alpha_4)$
S_6	(1,1,2,3)	π_6	$\alpha_2(1 - \alpha_2) - 2(\alpha_3 - \alpha_4)$
S_7	(1,2,3,3)	π_7	$\alpha_2(1 - \alpha_2) - 2(\alpha_3 - \alpha_4)$
S_8	(1,2,1,3)	π_8	$4[\alpha_2(1 - \alpha_2) - 2(\alpha_3 - \alpha_4)]$
S_9	(1,2,3,4)	π_9	$1 - 6\alpha_4 + 8\alpha_3 - 6\alpha_2 + 3\alpha_2^2$

this could reasonably be expected to hold. For example, if one has an infinite, random-mating population with alleles A_1,A_2, ... (in terms of identity—i.e., with reference back to the founders), having frequencies p_i, where $\sum p_i = 1$, then we have $\alpha = \sum p_i^2$ and $\beta\alpha^2 = \sum p_i^4$. Note that all sums are over unequal indices. It can be proved (see the Appendix, below) that $\beta \geqslant \alpha$, with equality if—and only if—in each case $p_i = \frac{1}{n}$. This latter condition implies that $\gamma = 2\alpha$. It should be noted that there are difficulties with each of the other π's in Génin and Clerget-Darpoux's model. This can easily be seen, since, if there were only two alleles (with respect to identity), then $\pi_6 = \pi_7 = \pi_8 = \pi_9 = 0$, which is not possible in Génin and Clerget-Darpoux's model; nor can $\pi_9 = 0$ in their model when there are only three alleles (with respect to identity).

An Alternate Parameterization

The expressions discussed above for the probabilities of the various identity states are particularly appropriate to the discussion of the Génin and Clerget-Darpoux formulas, since these latter appear to have been derived by a conditioning argument. However, a more natural parameterization for this model is to use a set of $\alpha_i = P(i$ randomly selected genes are identical by descent). This is the approach adopted by Weir (1994), although he also requires an additional parameter (his "Δ") in dealing with four genes, since he does not assume a permutable population, and so must consider two pairs of genes as distinct from four genes. We can now express, in a straightforward manner, the probability of identity states involving two, three, and four genes, exploiting the invariance under permutations. For example, for three genes and with the notation defined in table 2, we see, on examining the first two genes of the triplet, that the probability that they are identical by descent is $\alpha_2 = \alpha$, so that $\tau_1 + \tau_2 = \alpha_2$, implying that $\tau_2 = \alpha_2 - \alpha_3$. Table 3 gives similarly derived expressions for four genes.

Deriving Sib-Pair Probabilities

Having derived the vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_9)$ for a pair of individuals randomly selected from a permutable population, we can obtain the vector π appropriate for a pair of sibs whose parents are such a randomly selected pair, using the standard formula $\pi_{\text{sibs}} = \pi_{\text{parents}} M$, where

$$16\mathbf{M} = \begin{bmatrix} 16 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 16 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 2 & 4 & 4 & 4 & 0 & 0 & 0 & 0 \\ 4 & 0 & 4 & 4 & 4 & 0 & 0 & 0 & 0 \\ 4 & 0 & 4 & 4 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 8 & 0 & 0 & 0 & 0 & 8 & 0 \\ 0 & 0 & 8 & 0 & 0 & 0 & 0 & 8 & 0 \\ 1 & 0 & 3 & 2 & 2 & 1 & 1 & 6 & 0 \\ 0 & 0 & 4 & 0 & 0 & 0 & 0 & 8 & 4 \end{bmatrix}.$$

Direct Probabilities from Any π

As an alternative to deriving π_{sibs} from π_{parents}, we can proceed directly from π^*_{sibs}, where the asterisk (*) indicates that the expressions take into account kinship only within the observed genealogy and that they do not take into consideration the possible kinship of the founders of that genealogy, using the theory developed earlier for two, three, and four genes. Moreover, we can allow any relationship between the parents to be incorporated, so that sib-pair analysis for families with parents who are first cousins, for example, can easily be performed. In fact, for any pair of individuals with the $\pi^*_{\text{individuals}}$ calculated for them back to their founders within their defining genealogy, we can readily find $\pi_{\text{individuals}}$, provided that the population is of the permutable kind.

I have $\pi_{\text{individuals}} = \pi^*_{\text{individuals}}\mathbf{W}$, where

$$\mathbf{W} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \alpha_2 & (1-\alpha_2) & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \alpha_2 & 0 & (1-\alpha_2) & 0 & 0 & 0 & 0 & 0 & 0 \\ \alpha_2 & 0 & 0 & (1-\alpha_2) & 0 & 0 & 0 & 0 & 0 \\ \alpha_2 & 0 & 0 & 0 & (1-\alpha_2) & 0 & 0 & 0 & 0 \\ \alpha_3 & (\alpha_2-\alpha_3) & 0 & 2(\alpha_2-\alpha_3) & 0 & [1] & 0 & 0 & 0 \\ \alpha_3 & (\alpha_2-\alpha_3) & 0 & 0 & 2(\alpha_2-\alpha_3) & 0 & [1] & 0 & 0 \\ \alpha_3 & 0 & (\alpha_2-\alpha_3) & (\alpha_2-\alpha_3) & (\alpha_2-\alpha_3) & 0 & 0 & [1] & 0 \\ \alpha_4 & 2(\alpha_2-\alpha_3) & 2(\alpha_2^2-\alpha_4) & 2(\alpha_3-\alpha_4) & 2(\alpha_3-\alpha_4) & [2] & [2] & 4[2] & [3] \end{bmatrix}$$

with $[1] = (1 - 3\alpha_2 + 2\alpha_3)$, $[2] = \alpha_2(1 - \alpha_2) - 2(\alpha_3 - \alpha_4)$, and $[3] = 1 - 6\alpha_4 + 8\alpha_3 - 6\alpha_2\alpha + 3\alpha^2$.

The derivation of $\mathbf{W}$ is fairly straightforward from tables 2 and 3. For example, consider the fourth column of $\mathbf{W}$—ω_4, say. $\pi^*_{\text{individuals}}\omega_4$ equals the fourth element of $\pi_{\text{individuals}}$—that is, the probability of state S_4, or (1,1,1,2). We work through the states for the individuals, imposing the additional kinship of the population. Thus we see that, when population kinship is incorporated, (1,1,1,2) can only arise from the states (1,1,1,2), (1,1,2,3), (1,2,1,3), or (1,2,3,4) in the pair of individuals. If the pair have identity state (1,1,1,2) within their genealogy, then we require that the two alleles 1 and 2 remain distinct, which, on the basis of the data in table 2, has probability $(1 - \alpha_2)$. In a similar manner, if their within-genealogy state is (1,1,2,3), then we require that the three distinct alleles 1, 2, and 3 are, in fact, in state (1,1,2) or state (1,2,1) with respect to the population, which has probability $2(\alpha_2 - \alpha_3)$. For (1,2,1,3), the alleles must be in state (1,1,2), with probability $(\alpha_2 - \alpha_3)$, and, finally, (1,2,3,4) must be precisely (1,1,1,2), with probability $2(\alpha_3 - \alpha_4)$.

Thus we can easily derive the appropriate expressions for any sib pair whose parents have a specific relationship in a permutable population with known values of α_2, α_3, and α_4 . For example, for sibs with unrelated parents, $\pi^*_{\text{sibs}} = (0, 0, 1, 0, 0, 0, 0, 2, 1)/4$; for sibs whose parents are first cousins, $\pi^*_{\text{sofc}} = (1, 0, 15, 2, 2, 1, 1, 30, 12)/64$; and, for sibs whose parents are double first cousins, $\pi^*_{\text{sodfc}} = (4, 1, 29, 8, 8, 3, 3, 54, 18)/128$. These expressions are most easily derived by use of the earlier formula, $\pi_{\text{sibs}} = \pi_{\text{parents}}M$, which is equally valid for π^*_{sibs} and π^*_{parents}. For these particular relationships, π comes directly from the k coefficients ($\pi_3 = k_2$, $\pi_8 = k_1$, $\pi_9 = k_0$, and all other $\pi_i = 0$). The k coefficients are, respectively, (1,0,0), $(\frac{3}{4},\frac{1}{4},0)$, and $(\frac{9}{16},\frac{6}{16},\frac{1}{16})$. Of course, one can proceed directly from the π^*_{parents}, which are commonly given in the literature, in a single step, via the formula $\pi_{\text{individuals}} = \pi^*_{\text{parents}}\mathbf{MW}$, by combining the two matrix multiplications.

Discussion

It has been emphasized throughout the foregoing remarks that expressions for appropriate π vectors require a carefully specified model of the population structure. Here that specification is via a permutable population, which is precise for a Hardy-Weinberg population and is approximately true for a finite infinite-alleles model. As usual, one needs to be careful to distinguish between identity-by-descent and identity-by-state alleles. Here reference is always to identity-by-descent alleles, so that, even though reference is made to a Hardy-Weinberg population, it is supposed that this has arisen from a finite population at some stage in the past and so alleles can be identified with founder genes. Thus, in this case, the coefficients of identity, which are usually regarded as zero in a Hardy-Weinberg population, are nonzero and measure real kinship.

In order to utilize the aforementioned formulas, one needs good estimates of the various coefficients. This issue has recently been discussed in some detail by Morton and Teague (1996). It is worth noting that one of the assumptions of Génin and Clerget-Darpoux and of the present article—that is, that the coefficients of inbreeding and kinship are equal—is empirically supported by Morton's work in a variety of populations (see table 5.1 of Morton and Teague [1996], and references therein).

Appendix

We have $\alpha = (\sum p_i^2)$ and $\beta\alpha^2 = \sum p_i^4$. The sign of $\beta - \alpha$ is the same as that of $d = \beta\alpha^2 - \alpha^3$. Now,

$$\begin{aligned} d &= (\sum p_i^4) - (\sum p_i^2)^3 \\ &= (\sum p_i)^2(\sum p_i^4) - (\sum p_i^2)^3, \text{ since } \sum p_i = 1 \\ &= 2\sum p_ip_j(p_i - p_j)^2(p_i^2 + p_ip_j + p_j^2) \\ &\quad + \frac{1}{3}\sum p_ip_jp_k(p_i^3 + p_j^3 + p_k^3 - 3p_ip_jp_k) \end{aligned}$$

after tedious but straightforward rearrangement, where each Σ is over indices that are distinct; that is, $i \neq j$, $i \neq k$, and $j \neq k$. The first term of the final expression is greater than or equal to zero, with equality only if p_i is constant for all i or if $p_i = 1$ for some i, since each component is nonnegative. The second term is greater than or equal to zero, since, for three nonnegative quantities—x_1, x_2, and x_3—we have $\frac{1}{3}(x_1 + x_2 + x_3) \geqslant \sqrt[3]{x_1 x_2 x_3}$ (i.e., see Hardy et al. 1967), with equality if, and only if, x_i is constant. Taking $x_i = p_i^3$, we have $(p_1^3 + p_2^3 + p_3^3)$, and so the second term is greater than or equal to zero, with equality if, and only if, p_i is constant or if, at most, two p_i's are nonzero. Combining results and taking n to be the number of alleles with nonzero frequency, we see that $d \geqslant 0$, and thus $\beta \geqslant \alpha$, with equality if, and only if, p_i is constant for all i.

References

Bishop DT, Williamson JA (1990) The power of identity-by-state methods for linkage analysis. Am J Hum Genet 46: 254–265

Cotterman CW (1940) A calculus for statistico-genetics. PhD thesis, Ohio State University, Columbus

Génin E, Clerget-Darpoux F (1996) Consanguinity and the sib-pair method: an approach using identity by descent between and within individuals. Am J Hum Genet 59: 1149–1162

Hardy GH, Littlewood JE, Pólya G (1967) Inequalities. Cambridge University Press, Cambridge

Morton NE, Teague JW (1996) Kinship, inbreeding and matching probabilities. In: Boyce AH, Mascie-Taylor CGN (eds) Molecular biology and human diversity. Cambridge University Press, Cambridge, pp 51–62

Weir BS (1994) The effect of inbreeding on forensic calculations. Annu Rev Genet 28:597–621