Parentage Analysis With Genetic Markers in Natural Populations. I. The Expected Proportion of Offspring With Unambiguous Paternity

Ranajit Chakraborty,* Thomas R. Meagher† and Peter E. Smouse‡

**Center for Demographic and Population Genetics, University of Texas Graduate School of Biomedical Sciences, Houston, Texas 77225, †Department of Botany, Duke University, Durham, North Carolina 27706, and ‡Departments of Human Genetics and Biological Sciences, University of Michigan, Ann Arbor, Michigan 48109*

Manuscript received August 6, 1987
Revised copy accepted November 14, 1987

ABSTRACT

Recent studies indicate that polymorphic genetic markers are potentially helpful in resolving genealogical relationships among individuals in a natural population. Genetic data provide opportunities for paternity exclusion when genotypic incompatibilities are observed among individuals, and the present investigation examines the resolving power of genetic markers in unambiguous positive determination of paternity. Under the assumption that the mother for each offspring in a population is unambiguously known, an analytical expression for the fraction of males excluded from paternity is derived for the case where males and females may be derived from two different gene pools. This theoretical formulation can also be used to predict the fraction of births for each of which all but one male can be excluded from paternity. We show that even when the average probability of exclusion approaches unity, a substantial fraction of births yield equivocal mother-father-offspring determinations. The number of loci needed to increase the frequency of unambiguous determinations to a high level is beyond the scope of current electrophoretic studies in most species. Applications of this theory to electrophoretic data on *Chamaelirium luteum* (L.) shows that in 2255 offspring derived from 273 males and 70 females, only 57 triplets could be unequivocally determined with eight polymorphic protein loci, even though the average combined exclusionary power of these loci was 73%. The distribution of potentially compatible male parents, based on multilocus genotypes, was reasonably well predicted from the allele frequency data available for these loci. We demonstrate that genetic paternity analysis in natural populations cannot be reliably based on exclusionary principles alone. In order to measure the reproductive contributions of individuals in natural populations, more elaborate likelihood principles must be deployed.

THE determination of parentage of individuals from genetic data has become an increasingly interesting activity in population biology, since it is of immediate importance to the study of mating behavior and genetic dispersal (SMITH and ADAMS 1983; ELLSTRAND and MARSHALL 1985; HAMRICK and SCHNABEL 1985; MEAGHER 1986; STANTON 1986) as well as to the management of captive animal populations (MCCRACKEN and BRADBURY 1977; FOLTZ and HOOGLAND 1981; HANKEN and SHERMAN 1981). In many applications, assignment of maternal parentage for each individual is straight-forward (*e.g.*, mother-child pairing can be observed directly without much difficulty), so that parentage analysis reduces to the question of paternity determination. In this sense, the problem is reminiscent of human paternity analysis, whose logic is well understood, though not without controversey (*e.g.*, see WALKER 1983; AICKIN 1984).

When genetic data are available from mother-child-putative father trios, the likelihood of paternity can be determined with inferential procedures of genealogical structure (THOMPSON 1975, 1976a,b; MEAGHER and THOMPSON 1986). This has been standard practice in medicolegal context for nearly 50 yr (ESSEN-MÖLLER 1938), but has recently been criticized on grounds that: (1) it can sometimes result in erroneous paternity assignment (MAJUMDER and NEI 1983), and (2) the resulting probability statement is alleged not to have a sound statistical basis (LI and CHAKRAVARTI 1985). Responses to both criticisms can be found in the literature (*e.g.*, VALENTIN 1984; ELSTON 1986; MICKEY, GJERTSON and TERASAKI 1986; THOMPSON 1986). With genetic data, incompatibility of parent-child marker genotypes is nearly always conclusive proof of nonpaternity. Thus, an alternative procedure for assignment of paternity has evolved, where the likelihood of paternity is derived from the exclusionary data alone, without regard to the actual genotypic configuration of the male parent involved in a paternity dispute (CHAKRAVARTI and LI 1983).

If one adopts genetic exclusion criteria as the primary means of paternity identification, an obvious strategic requirement is to evaluate the efficacy of biochemical markers of varying allele frequencies with reference to their utility in the assignment of paternity. The behavior of genetic exclusion probabilities under varying allele frequencies and number of alleles has been extensively explored in the case where males and females are drawn from the same population and share common allele frequencies (*e.g.*, CHAKRABORTY and SCHULL 1976; CHAKRABORTY and FERRELL 1982; MEAGHER and THOMPSON 1986).

In a well-defined, finite gene pool, as in the case of a captive population, where both males and females are enumerable, it is not unrealistic to assume that the genotypic configurations of *both* parental gene pools are known without error. In other words, it can be assumed that the multi-locus genotypes for all males and females in the population are known, for any particular battery of genetic loci employed. MEAGHER (1986) has discussed situations of this type. It is quite possible that some genetic loci will show allele frequency differences between the two sexes, and this may have a substantial effect on the efficacy of the biochemical markers used. Moreover, such a sex difference has implications for the evaluation of exclusionary power, as well as for the determination of the likelihood of paternity in any particular case.

There are many aspects of paternity evaluation that may be influenced by differing allele frequencies among males and females. In this paper, our objective is to provide analytical solutions for two questions. (1) Given a mother-offspring (M, O) pair, what fraction of the adult males of the population can be excluded from paternity? (2) In what fraction of the births that occur in the population, can all but one male be excluded from paternity, *i.e.*, in what fraction of the offspring can the pedigree relationship of the mother-father-offspring trio be unequivocally established? From a strategic standpoint, the second question can be generalized to consider the probability distribution of the number of non-excluded males for every mother-offspring pair in any given population of finite size.

MEAGHER (1986) discussed the implications of the answers to these questions in the context of ambiguous determination of paternity and their relevance to the extent and spread of genes in a well-defined geographical area. The analytical solutions presented in this paper are based on the logic described in CHAKRABORTY, SHAW and SCHULL 1974; CHAKRABORTY, FERRELL and SCHULL 1979), with the generalization that the allele frequencies are different in the male and female gene pools of the population. Since the number of biochemical markers showing a high degree of polymorphism is rather limited in most practical situations (*e.g.*, MEAGHER 1986), we shall also address a useful strategic question. (3) Which combination of genetic markers will provide optimum information regarding the expected proportion of offspring with unambiguous paternity?

It may be noted that some of these questions had been addressed in the context of human paternity analysis, but most often such analysis either considered equal allele frequencies in the two parental gene pools, or are based on average exclusion probability afforded by a mother-child pair, and thus disregards the variation of exclusion probability over different genotypic combinations of the mother-child pairs observed in a natural population (CHAKRABORTY, FERRELL and SCHULL 1979; SELVIN 1980; CHAKRABORTY and FERRELL, 1982, 1983; SMOUSE and CHAKRABORTY 1986). The following mathematical treatment is proposed to circumvent these limitations, and furthermore it is applicable when the potential parents are drawn from finite populations of known size. Thus, even though the questions addressed here are not novel in parentage analysis, to our knowledge, an analytical treatment of these issues with unequal allele frequencies in finite populations of mates has not been presented before.

MATERIALS AND METHODS

Although the loci employed to examine genetic variation may exhibit dominance relationships among alleles (such as blood groups, or histocompatibility antigens), and although some loci are sex-linked, we shall consider only autosomal codominant loci for the purpose of this paper, since we shall employ only this type of loci in the application to follow. Because of the finiteness of the population, the observed allele frequencies in the two sexes may not be the same (see MEAGHER 1986), and we shall therefore present all computations with unequal allele frequencies for the two sex groups.

Exclusion probability: Let $p_1, p_2, \ldots, p_k$ and $q_1, q_2, \ldots, q_k$ be the allele frequencies of k codominant alleles at an autosomal locus (say, the lth locus) in males and females, respectively. The exclusion probability (PE_l) for such a system will have values given by (CHAKRABORTY and FERRELL 1982):

$$PE_l = \begin{cases} (1-p_i)^2 & \text{with probability } p_i[1-q_i+q_i^2] \\ & \quad \text{for } i = 1, 2, \ldots, k; \\ (1-p_i-p_j)^2 & \text{with probability } q_iq_j(p_i+p_j) \\ & \quad \text{for } j > i = 1, 2, \ldots, k. \end{cases} \tag{1}$$

One can thus obtain $k(k+1)/2$ different values of the exclusion probability for a codominant, k-allele system, depending upon the observed (M, O) genotypic pair (see Table 1 for a three-allele example). The average exclusion probability for such a system is given by

$$\begin{aligned} E(PE_l) &= \sum_{i=1}^{k} p_i \cdot (1-p_i)^2 \cdot (1-q_i+q_i^2) \\ &\quad + \sum_{j>i=1}^{k}\sum q_iq_j \cdot (p_i+p_j) \cdot (1-p_i-p_j)^2 \\ &= 1 - 2a_{20} + a_{30} + 3(a_{11}a_{21} - a_{32}) \\ &\quad - 2(a_{11}^2 - a_{22}), \end{aligned} \tag{2}$$

where $a_{rs} = \sum_{i=1}^{k} p_i^r \cdot q_i^s$, for $r, s = 0, 1, 2, 3$. This

TABLE 1

Expected frequencies of mother-offspring pairs and their probabilities of genetic exclusion of male genotypes for an autosomal locus with three alleles

(A) Expected population frequencies of (M, O) pairs under random mating: 3 alleles

Offspring genotypes	Maternal genotypes A_1A_1	A_1A_2	A_1A_3	A_2A_2	A_2A_3	A_3A_3
A_1A_1	$p_1q_1^2$	$p_1q_1q_2$	$p_1q_1q_3$	0	0	0
A_1A_2	$p_2q_1^2$	$q_1q_2(p_1 + p_2)$	$p_2q_1q_3$	$p_1q_2^2$	$p_1q_2q_3$	0
A_1A_3	$p_3q_1^2$	$p_3q_1q_2$	$q_1q_3(p_1 + p_3)$	0	$p_1q_2q_3$	$p_1q_3^2$
A_2A_2	0	$p_2q_1q_2$	0	$p_2q_2^2$	$p_2q_2q_3$	0
A_2A_3	0	$p_3q_1q_2$	$p_2q_1q_3$	$p_3q_2^2$	$q_2q_3(p_2 + p_3)$	$p_2q_3^2$
A_3A_3	0	0	$p_3q_1q_3$	0	$p_3q_2q_3$	$p_3q_3^2$

(B) Exclusion probabilities for different (M, O) pairs: 3 alleles

Offspring genotypes	Maternal genotypes A_1A_1	A_1A_2	A_1A_3	A_2A_2	A_2A_3	A_3A_3
A_1A_1	$(1 - p_1)^2$	$(1 - p_1)^2$	$(1 - p_1)^2$			
A_1A_2	$(1 - p_2)^2$	$(1 - p_1 - p_2)^2$	$(1 - p_2)^2$	$(1 - p_1)^2$	$(1 - p_1)^2$	
A_1A_3	$(1 - p_3)^2$	$(1 - p_3)^2$	$(1 - p_1 - p_3)^2$		$(1 - p_1)^2$	$(1 - p_1)^2$
A_2A_2		$(1 - p_2)^2$		$(1 - p_2)^2$	$(1 - p_2)^2$	
A_2A_3		$(1 - p_3)^2$	$(1 - p_2)^2$	$(1 - p_3)^2$	$(1 - p_2 - p_3)^2$	$(1 - p_2)^2$
A_3A_3			$(1 - p_3)^2$		$(1 - p_3)^2$	$(1 - p_3)^2$

Note: Exclusion probabilities are not computed for incompatible (M, O) pairs. The value of p_i is the frequency of the *i*th allele of the paternal gene pool, and the value of q_i is the frequency of the *i*th allele in the maternal gene pool.

represents the most general expression of the average exclusion probability for a codominant multiallelic locus, where males and females are from different gene pools, shown explicitly for the first time here. This general formula reduces to the equation II-B of Chakravarti and Li (1983) when $p_i = q_i$, for all i; *i.e.*, when the male and female parents are drawn from the same gene pool. This special case had also been discussed by Selvin (1980), who did not reduce his expression to similar algebraic closed form.

When several such systems are employed for paternity analysis, the combined probability of exclusion $[P_E(C)]$ is given by

$$P_E(C) = 1 - \prod_{l=1}^{L} (1 - PE_l), \tag{3}$$

where PE_l is the exclusionary probability afforded by the lth system (Boyd 1954; Chakraborty, Shaw and Schull 1974).

In Equation 3, the specific value of PE_l can be that of the average of all mother-offspring genotypic pairs or that of a specific (M, O) pair. Furthermore, since PE_l can take different values for different (M, O) genotypic pairs (Table 1, Equation 1), there will be a probability distribution of observed $P_E(C)$ values for all (M, O) pairs with any given battery of genetic markers. Such a distribution can be numerically evaluated using Equations 1 and 3, and can be contrasted with the observed distribution of $P_E(C)$ values in a given situation, as we shall see later. The use of Equation 3 thus provides an expected proportion of excluded males for the specific loci used (giving the estimated effectiveness of the choice of genetic loci).

Proportion of offspring with unambiguous paternity: In a finite, closed population with N males, one of the N males is obviously the actual biological father of each offspring. Paternity can be unambiguously established if all males but one are excluded on the basis of genotypic information. Chakraborty, Ferrell and Schull (1979) provided a theoretical solution for this probability in the process of solving a more general problem. Their treatment, however, depends on equality of allele frequencies in males and females, and used the average exclusion probability. We shall relax these two conditions in the following way. Let $P_E(C)$ denote the combined probability of exclusion (Equation 3) for a given (M, O) pair and a choice of genetic loci. Since one of the N males is the true biological father, the probability that all $(N - 1)$ non-fathers will be excluded from paternity is given by $[P_E(C)]^{N-1}$, using Equation 2 of Chakraborty, Ferrell and Schull (1979). This formulation, however, is dependent on the probability of exclusion afforded by the particular (M, O) pair, which will vary from pair to pair. Therefore, we need to evaluate the expectation of $[P_E(C)]^{N-1}$ for each genotypic pair in the population.

Let N be the number of males (possible fathers) in the population, as before. Let R be the total number of births with which we are concerned. Denote the (M, O) pair for the ith birth ($i = 1, 2, \ldots, R$) by $(M, O)_i$, so that the probability of exclusion for the ith birth may be denoted by $PE(i)$ (evaluated by Equation 2 and combined for all systems by Equation 3). For the ith birth, the probability that only one male will remain nonexcluded (unambiguous paternity) is given by

$$\pi_i = [PE(i)]^{N-1}. \tag{4}$$

Now, let us introduce a sequence of indicator variables $\{X_i;\ i = 1, \ldots, R\}$ defined by

$$X_i = \begin{cases} 1, \text{ if the } i\text{th birth has only one male nonexcluded,} \\ \quad i.e., \text{ the father is unambiguously determined} \\ 0, \text{ otherwise} \end{cases} \tag{5}$$

The variable $X = \sum_{i=1}^{R} X_i$, will indicate the total number of births for which the paternity determinations are unequivocal.

The probability distribution of X can then be worked out with a probability generating function approach (FELLER 1968); *i.e.*,

Prob. $(X = r)$

$$= \text{Coefficient of } t^r \text{ in } \prod_{i=1}^{R} [(1 - \pi_i) + \pi_i t], \quad (6)$$

for some arbitrary variable t $(-1 < t < 1)$. The expectation and variance of X are given by

$$E(X) = \sum_{i=1}^{R} \pi_i \quad (7a)$$

and

$$V(X) = \sum_{i=1}^{R} \pi_i(1 - \pi_i), \quad (7b)$$

respectively. Because of the asymmetry of the distribution of X, these two expressions are, however, not very useful for any inferential purposes in practice. Nevertheless, for a completeness of the analysis of this distribution they are noteworthy.

The probability function of X, represented by Equation 6, is not computationally attractive. However, it is easy to evaluate such functions by the approach suggested by CHAKRABORTY and SCHULL (1976), sequentially adding each birth ($i = 1, 2, \ldots, R$).

This formulation also yields an analytically closed expression for the expected proportion of births, for each of which the paternity determination is conclusive (see APPENDIX for a proof). This is precisely the $(N - 1)$-th moment of $P_E(C)$, which can be obtained from the fact that the distributions of PE_l for $l = 1, 2, \ldots, L$ (Equation 2) are mutually independent, and is given by

$$E[P_E(C)]^{N-1} = \sum_{r=0}^{N-1} \binom{N-1}{r} (-1)^r \prod_{l=1}^{L} E[\{1 - PE_l\}^r], \quad (8)$$

where the rth moment of $\{1 - PE_l\}$ can be evaluated from Equation 2 as

$$E[\{1 - PE_l\}^r] = \sum_{t=0}^{r} \binom{r}{t} (-1)^t \sum_{v=0}^{2t} \binom{2t}{v} (-1)^v \times \Bigg[a_{v+1,0}(l) - a_{v+1,1}(l) - (2^v - 1) a_{v+1,2}(l) \quad (9)$$
$$+ (1/2) \sum_{w=0}^{v+1} \binom{v+1}{w} a_{v+1-w,1}(l)\, a_{w,1}(l) \Bigg],$$

where $a_{\alpha,\beta}(l) = \sum p_i^{\alpha} q_i^{\beta}$, in which p_i's and q_i's are the allele frequencies at the lth locus in males and females, respectively.

Thus, to determine the expected proportions of births in the population for which the paternity can be determined unambiguously, it is not necessary to enumerate all extant (M, O) pairs for the multilocus genotypic combinations. The allele frequencies at each locus provide this information, as long as we assume that the male and female gametes unite at random to form the offspring genotypes, and that the different loci employed are independent.

In field surveys of natural populations, it is possible to enumerate the distribution of the number of genetically possible male parents for all extant (M, O) pairs (see *e.g.*, MEAGHER 1986), which may be contrasted with the above expectation for such observations. For example, if we survey the genotypes of all N male parents of the population, the probability that m $(0 < m < N)$ males (one of whom is the true father) will not be excluded is given by

$$\binom{N-1}{m} [1 - P_E(C)]^{m-1} [P_E(C)]^{N-m},$$
$$\text{for } m = 1, 2, \ldots, N - 1. \quad (10)$$

Since $P_E(C)$ varies for different $(M, O)_i$ pairs, the expected distribution can be obtained by taking the expectation of (10) over all (M, O) pairs. Thus, the probability of m genotypically possible male parents is given by

$$\binom{N-1}{m} E\{[1 - P_E(C)]^{m-1} [P_E(C)]^{N-m}\}$$
$$= \binom{N-1}{m} \sum_{s=0}^{m-1} \binom{m-1}{s} (-1)^s E\{[P_E(C)]^{N-m+s}\}, \quad (11)$$

where the last term can be obtained using Equations 8 and 9 (see APPENDIX for proof).

APPLICATION

Biochemical markers in *Chamaelirium luteum* (L.): MEAGHER (1986) assayed biochemical polymorphism at eight genetic loci (PGI, PGM, GOT_2, GOT_3, TPI_2, TPI_3, GDH and MPI) from 70 females and 273 males from a natural population of *C. luteum* (L.) from Orange County, North Carolina. These polymorphisms were used in an analysis of statistical likelihoods of paternity (THOMPSON 1976a,b) for 2255 offspring seed sampled from known maternal genotypes.

We shall use these same data to show that in spite of the appreciable average exclusionary power of such a battery of genetic markers, conclusive determination of paternity cannot be resolved by exclusionary analysis alone. Table 1 of MEAGHER (1986) gives the allele frequencies at each of these eight loci in the parental (male and female) populations. Before considering these data for the present analysis, we note that there are two typographical errors in Table 1 of MEAGHER (1986): (1). Female allele frequencies are based on a sample of 70 individuals, and (2) the fourth allele of the MPI locus has a frequency of 0.007 among females. The eight loci were shown to segregate independently of each other, and for two systems (PGM and TPI_3), the allele frequencies differed significantly in the two sexes. In another two systems (TPI_3 and MPI), the paternal gene pool contains alleles that are absent from the maternal gene pool, and one system (TPI_2) exhibits polymorphism only in the paternal gene pool. These data are ideal to illustrate computations of the exclusion probabilities for the case with unequal allele frequencies in the two parental gene pools. Use of Equations 1 and 2, along with the allele frequency data of MEAGHER (1986; Table 1) gives the locus-specific exclusionary chances presented in Table 2. In this table, we present the average exclusion probability

TABLE 2

Locus-specific paternity exclusion probabilities (*PE*) and their ranges for *C. luteum* (L.), as surveyed by MEAGHER (1986)

		Probability of exclusion $(PE)^b$										
		Average		Minimum			Maximum			Most likely		
Locus[a]	No. of alleles	*PE*	Rank	*PE*	Rank	*f*[c]	*PE*	Rank	*f*	*PE*	Rank	*f*
PGI	4	0.347	(1)	0.040	(1)	0.121	0.880	(5)	0.059	0.176	(2)	0.440
PGM	3	0.215	(3)	0.010	(2)	0.132	0.812	(7)	0.096	0.065	(3)	0.620
GOT_2	2	0.071	(6)	0.0	(7)	0.079	0.839	(6)	0.077	0.007	(6)	0.844
GOT_3	2	0.009	(7)	0.0	(8)	0.007	0.982	(4)	0.009	8.1×10^{-5}	(7)	0.984
TPI_2	2	0.007	(8)	4.9×10^{-5}	(4)	0.993	0.986	(3)	0.007	4.9×10^{-5}	(8)	0.993
TPI_3	3	0.082	(5)	1.6×10^{-5}	(5)	0.014	0.992	(2)	0.004	0.008	(5)	0.897
GDH	2	0.112	(4)	0.0	(6)	0.137	0.716	(8)	0.133	0.024	(4)	0.730
MPI	5	0.287	(2)	0.008	(3)	0.180	0.996	(1)	0.002	0.270	(1)	0.361
Combined[d]		0.728		0.056		2.9×10^{-8}	1.0		2.9×10^{-14}	0.459		0.053

[a] The abbreviations for loci are: phosphoglucose isomerase (PGI), phosphoglucomutase (PGM), glutamate oxaloacetate transaminase (two loci: GOT_2 and GOT_3), triose phosphate isomerase (two loci: TPI_2 and TPI_3), glutamate dehydrogenase (GDH), and mannose-6-phosphate isomerase (MPI).

[b] The *PE* values were computed by using Equation 1 for each possible (M, O) genotypic pair, along with the relative frequencies of such (M, O) pairs in the population (represented as *f*). The average *PE* was computed for each system using Equation 2.

[c] *f* is the population frequency of (M, O) pairs yielding the respective *PE* values. Numbers in parentheses are rankings of loci, based on respective *PE* values.

[d] Combined values of *PE* were computed by using Equation 3, substituting the average, min, max, and most likely *PE* values for each locus. The frequencies for these respective combined *PE* values are obtained by multiplying the locus-specific relative frequencies of the respective (M, O) pairs which yields these values.

(PE_l) for each system, as well as the maximum, minimum, and most likely values of PE_l for particular (M, O) pairs, along with the relative frequencies with which such PE_l values might occur in a random (M, O) pair drawn from the population.

Several observations may be made from this table. *First,* the average exclusionary power depends upon the number of alleles as well as the allele frequencies. A system with more alleles may not always provide better average exclusionary power, as is evident in the case of the MPI locus, which exhibits five allelic variants but which has less average exclusionary power than the PGI system, which has four segregating alleles in the population. *Second,* the ranking of the loci with respect to the average exclusion probability (shown in parentheses in Table 2) is not necessarily the same as that of the most likely, maximum, or minimum values of PE_l. *Third,* since for a two-allelic locus, no male can be excluded when both mother and offspring are heterozygous ($PE_l = 0$), the choice of loci for which the *average* exclusionary power is maximum may lead to a substantial frequency of (M, O) pairs for which such loci will be completely uninformative with respect to an exclusionary event. In fact, the LOD score computations of paternity (GÜRTLER 1956; MEAGHER 1986) will also be uninformative for such loci when M and O are both heterozygous.

Equation 3 may be applied to these PE_l values to examine the exclusionary power of various combinations of biochemical loci. In Figure 1, we perform this in a graphic fashion. The loci are combined in this graph in a sequential manner, picking the best system (that with rank 1, from single-locus analysis), the best pair of loci (ranked 1 and 2), the best triplet (ranked 1, 2 and 3), and so on, to all eight loci. The uppermost *dashed line* of Figure 1 represents the approximate upper 95% values of $P_E(C)$ for the specific choice of combinations of loci, according to the ranking of average $P_E(C)$, while the lowest dashed line is the lower 5% values of $P_E(C)$. The middle solid line is the trajectory of the average $P_E(C)$ values

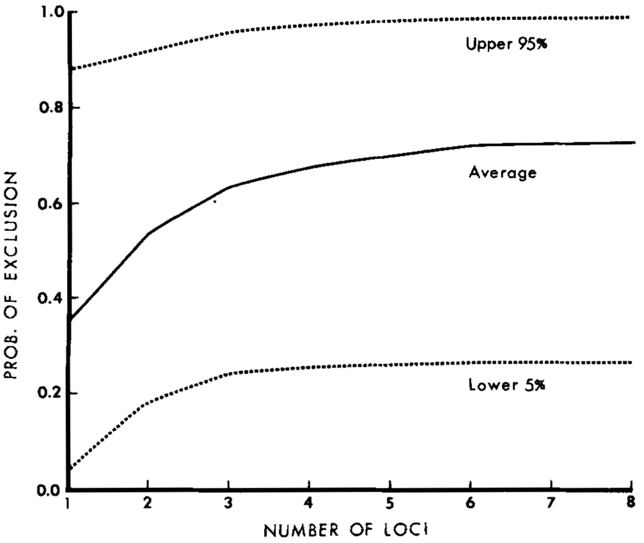

FIGURE 1.—The 90% confidence limits of paternity exclusion probabilities (*dashed lines*) for *C. luteum* (L.), obtained by sequentially adding the 8 loci (according to rankings of their average *PE* values shown in Table 2).

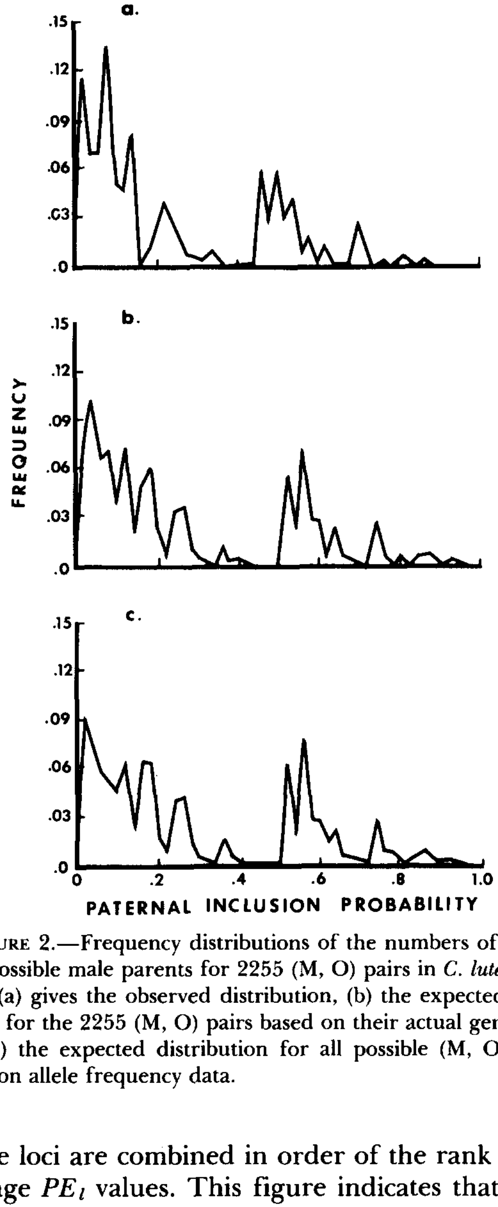

FIGURE 2.—Frequency distributions of the numbers of genetically possible male parents for 2255 (M, O) pairs in *C. luteum* (L). Panel (a) gives the observed distribution, (b) the expected distribution for the 2255 (M, O) pairs based on their actual genotypes, and (c) the expected distribution for all possible (M, O) pairs, based on allele frequency data.

as the loci are combined in order of the rank of the average PE_l values. This figure indicates that there is a wide range of possible exclusion probabilities for (M, O) pairs in the population, even when the genetic loci are optimally chosen for paternity analysis. Furthermore, it shows that even for a battery of genetic loci that has a high exclusionary power on average, there remain a considerable proportion of (M, O) pairs for which very few males can be excluded.

As these computations are based on the theoretical expectations of $P_E(C)$ values for specific (M, O) pairs, it is worthwhile to compare how these expectations tally with observations. Figure 2a performs this task, where the observed proportions of excluded males for 2255 (M, O) genotypic pairs assayed by MEAGHER (1986) are plotted; these can be compared with the distribution of $P_E(C)$ computed for the actual genotypes of these same (M, O) pairs (Figure 2, b and c). This graph shows that when each mother-offspring-putative father trio is scored for each of the eight genetic loci considered here, there are 57% (M, O) pairs for which less than 10% of the males can be excluded from paternity (the expected proportion of such (M, O) pairs based on allele frequency data is 63%). This clearly demonstrates that exclusionary criteria alone cannot fully resolve the problem of paternity assignment. We might add that this feature is not unique to the present data, and a high frequency of nonresolving cases will remain even if the battery of genetic markers were extended to enhance the average probability of exclusion to a value approaching unity, as suggested by the confidence belt of Figure 1.

Unambiguous determination of paternity and distribution of number of excluded males in population of finite size: MEAGHER (1986) earlier evaluated the empirical distribution of the number of genetically possible male parents for 575 seeds, collected from this natural population of *C. luteum*, where the seeds and their maternal parents could either be assigned to particular males with high probability by LOD score analysis or by excluding all but one male in the population. For the remaining 1680 seeds, paternity assignments was not possible, because the LOD score values are not very discriminatory. Equations 10 and 11 enable us to compare the emprical distribution of potential male parents for all 2255 seeds with their expectations, based on allele frequency counts in this population.

In addition to the observed distribution of the numbers of genetically possible male parents for all 2255 mother-offspring pairs (Figure 2a), in the other two panels of Figure 2 we plotted the expected distribution for these same (M, O) pairs (panel b), and finally the expected distribution for all possible (M, O) combinations, given the reported allele frequencies (panel c). It is clear that even though the average exclusion probability in this case is quite high (73%), in a large majority of cases (2198/2255 = 97.5%), paternity determination by exclusion criteria alone is ambiguous. Therefore, this figure also establishes that unambiguous assignment of paternity in a natural population is not generally feasible, based solely on exclusionary events.

DISCUSSION AND CONCLUSION

In terms of the basic objectives of this paper, the above considerations clearly show that the number of genetically possible male parents for any given (M, O) pair represents a substantial fraction of the head-count of available males in the population, even when the average exclusion probability afforded by the (M, O) pairs is high. As a consequence of this observation, we may also conclude that if we rely completely on exclusionary events, paternity can be unambiguously assigned for only a small proportion of births.

Two issues may be raised by way of objection,

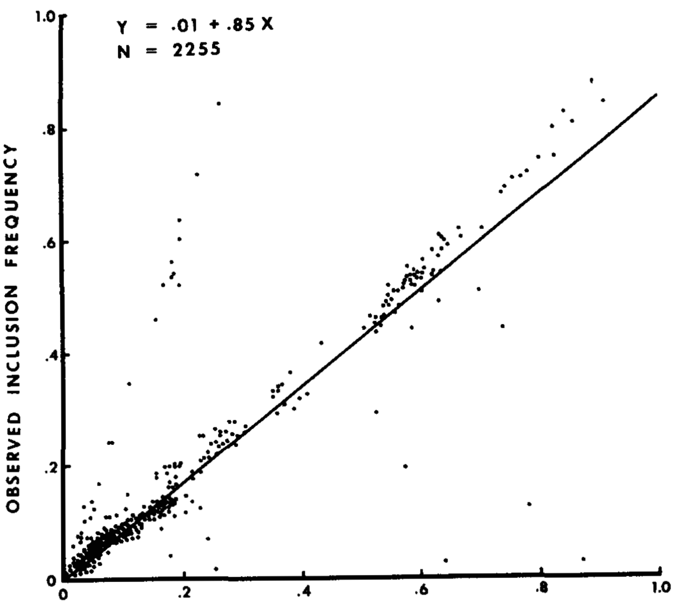

FIGURE 3.—Regression analysis of observed and expected proportions of genetically possible male parents in *C. luteum* (L). The data used are the same presented in panels (a) and (b) of Figure 2.

relative to the use of the data for these analyses. *First*, the theory discussed here depends heavily on the assumptions that the offspring genotypes are formed by random union of male and female gametes with respect to all loci employed, and that the males are all unrelated, so that the probabilities for each male can be assumed to be independent and identically distributed. *Second*, the eight-locus analysis examined here does not reflect the situation where the average probability of exclusion can be much higher than 73%. As we shall see below, these points do not change any of our conclusions in a qualitative way.

Random union of gametes and independence of male genotypes: The theory developed here assumes random union of gametes from the male and female gene pools to form the offspring genotypes. As we considered samples of seeds collected from this population, this may not be the case in reality, if there are any selective factors involved in the polymorphisms of the loci used. Furthermore, the assumption of independence of male genotypes may not hold if there is a genealogical structure in the population. To see the effect of departures from these two assumptions, we performed a regression analysis of the number of genotypically possible male parents for all of the 2255 seeds (observed versus expected). If the two assumptions mentioned above are correct, such a regression is expected to be linear through the origin with a slope of unity.

In the present regression analysis (Figure 3), the overall fit was quite good ($F_{1,2253}$ = 16702; $P <$ 0.0001), suggesting a very tight linear relationship between observed and expected inclusion probabilities ($1 - P_E(C)$). The amount of variance explained by the linear regression is 89%. The intercept (a = 0.013 ± 0.002), although significantly different from zero, was nevertheless very small and is probably more an artifact of the large sample, and holds little biological significance. The slope of the regression (b = 0.851 ± 0.007) was significantly and convincingly less than unity, indicating that expected exclusion probabilities are an *optimistic* view of the degree of resolution afforded to a given battery of genetic loci under natural conditions.

A slope of less than unity is probably a reflection of the underlying genealogical structure of the population. It has been shown elsewhere that if there are genealogical relationships among the potential mates in the population, exclusion probabilities will be smaller than that used in our calculations (MacCluer and Schull 1963; Salmon and Brocteur 1978). The reliability of exclusionary events for unambiguous assignment of paternity will be even smaller than that predicted by our theory, which should thus be viewed as a "best case" result.

There are some clear "outliers" in Figure 3, that can be easily explained. The outliers above the regression line all come from 5 maternal sibships, which provide (M, O) genotypic pairs having exclusion power much worse than the expected. The scattered outliers below the line are from two maternal sibships whose exclusionary powers are better than the expected probability of exclusion.

Number of loci needed to attain reasonable value for exclusion of all-but-one male from paternity: The analysis considered here uses a battery of eight loci yielding an average probability of exclusion of 73%. One generally aims at developing a battery of loci with much higher average exclusion probability. Consider the probability of excluding all-but-one male in a situation where each of L loci provides the same exclusion probability, PE. In a population of N potential fathers, the probability that all-but-one male will be excluded can be extracted from Equation (3), and is seen to be

$$PE_{N-1} = [1 - (1 - PE)^L[^{N-1}. \qquad (13)$$

Figure 4 plots this function, where each of the L loci are taken to be biallelic, codominant, and with allele frequencies p and $1 - p$. Note that for maximally informative two-allele loci (p = 0.5), we need at least 50 such ideal loci to exclude all-but-one male from paternity with probability larger than 0.99, given N = 273. For less efficient loci ($p \neq 0.5$), the number of loci needed is much larger.

To draw analogy with our illustration with *C. luteum*, we can take multiples of the same eight-locus systems to evaluate PE_{N-1}, to determine when this probability exceeds 0.99. We would have to at least

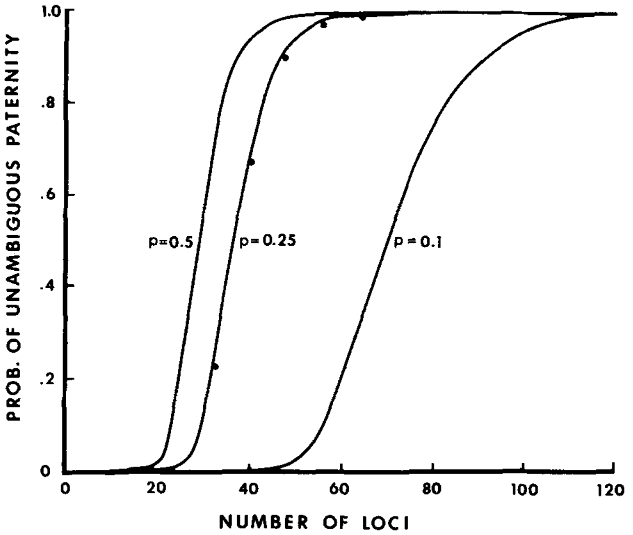

FIGURE 4.—Probability of unequivocal assignment of paternity, as a function of number of equivalent genetic markers. The *solid circles* represent the replicates of the set of eight loci surveyed for *C. luteum* (L). The *solid lines* are for sets of bi-allelic loci with allele frequencies p and $(1 - p)$, where p values are given in the figure.

triple the number of available markers to make this strategy work (as shown by the *dark circles* in Figure 4). It is not feasible with traditional electrophoresis to detect much additional allozymic variation, since the eight polymorphic loci used here were chosen as a result of an intensive survey involving more than 25 enzyme systems.

Use of the restriction fragment length polymorphisms (RFLPs) may provide additional power in this regard, since such polymorphic DNA markers are relatively more abundant than the electrophoretic ones (COOPER and SCHMIDTKE 1984). One may argue that with enough RFLPs, we could achieve any level of genetic resolution desired (LI and CHAKRAVARTI 1985). We should reiterate here that even when the average probability of exclusion is high, there will be a substantial fraction of (M, O) pairs that will have a low power of exclusion. With a large number of DNA markers, we will also have loci that are not independently segregating, and will require a more elaborate haplotype analysis, the problems with which are not fully resolved (SMOUSE and CHAKRABORTY 1986). In addition, the exclusionary power of RFLPs is also limited because of the low heterozygosity of individual site-specific polymorphisms and the necessity of construction and hybridization of new probes [see *e.g.*, QUINN *et al.* (1987) and comments by HILL (1987)]. Some of the difficulties attendant to the RFLP technology can, however, be circumvented by the use of the hypervariable minisattelite probes developed by JEFFREYS, WILSON and THEIN (1985) and used successfully in human, mice, cats, dogs, and birds for establishing and/or disproving genealogical relationships (JEFFREYS et al. 1987; JEFFREYS and MORTON 1987; BURKE and BRUFORD 1987; WETTON *et al.* 1987). The recently developed variable number of tandem repeat (VNTR) markers are also a promising technology in this regard (NAKAMURA *et al.* 1987).

Given that a strictly exclusionary solution to the problem of parentage assessment is not attainable in most natural populations, there is need for an alternative procedure that optimally apportions total paternity in the population among the candidate males, without the necessity of assigning a definite father to any particular mother-offspring pair. We will show in a subsequent paper (P. E. SMOUSE, R. CHAKRABORTY and T. R. MEAGHER, unpublished data) that instead of an exclusionary solution, the standard paternity analysis approach can be adapted for this purpose. Similar procedures have been explored in the context of constructing genealogical relationships from genetic data (see *e.g.*, THOMPSON, 1976a,b; MEAGHER and THOMPSON 1986). We will show that such an approach offers the opportunity to model reproductive success both as a function of size and social dominance of a candidate male and as a function of the phenotypic, kinship, and spatial relationships between mating partners.

The authors would like to thank JOSEPH TWOREK for help with the programming and computations. This work is supported by U.S. Public Health Service Research grants NIH-GM-20293 (R.C.), NSF BSR-83-14887 (T.R.M.), DE-FG02-ER60533 and NIH-R01-GM-32589 (P.E.S.).

LITERATURE CITED

AICKIN, M., 1984 Some fallacies in the computation of paternity probabilities. Am. J. Hum. Genet. **36:** 904–915.

BOYD, W. C., 1954 Tables and nomograms for calculating chances of excluding paternity. Am. J. Hum. Genet. **6:** 426–433.

BURKE, T., and M. W. BRUFORD, 1987 DNA fingerprinting in birds. Nature **327:** 149–152.

CHAKRABORTY, R., and R. E. FERRELL, 1982 Efficient choice of genetic markers for paternity testing and correlation of paternity index and exclusion probability. Forensic Sci. Intl. **19:** 113–124.

CHAKRABORTY, R., and R. E. FERRELL, 1983 Paternity index and its use in parentage diagnosis. pp. 115–122. In: *Inclusion Probabilities in Parentage Testing*, Edited by R. H. WALKER. American Association of Blood Banks, Arlington, Va.

CHAKRAVARTI, A., and C. C. LI, 1983 The effect of linkage on paternity calculations. pp. 411–420. In: *Inclusion Probabilities in Parentage Testing*, Edited by R. H. WALKER. American Association of Blood Banks, Arlington, Va.

CHAKRABORTY, R., and W. J. SCHULL, 1976 A note on the distribution of the number of exclusions to be expected in paternity testing. Am. J. Hum. Genet. **28:** 615–618.

CHAKRABORTY, R., R. E. FARRELL and W. J. SCHULL, 1979 Paternity exclusion in primates: two strategies. Am. J. Phys. Anthropol. **50:** 367–372.

CHAKRABORTY, R., M. W. SHAW and W. J. SCHULL, 1974 Exclusion of paternity: The current state of the art. Am. J. Hum. Genet. **28:** 477–488.

COOPER, D. N., and J. SCHMIDTKE, 1984 DNA restriction fragment length polymorphisms and heterogeneity in the human genome. Hum. Genet. **66:** 1–16.

ELLSTRAND, N. C., and D. L. MARSHALL, 1985 Interpopulation gene flow by pollen in wild radish, *Raphanus sativus*. Am. Nat. **126:** 606–616.

Elston, R. C., 1986 Probability and paternity testing. Am. J. Hum. Genet. **39:** 112–122.

EssenMöller, E., 1938 Die Beweiskraft der Ähnlichkeit im Vaterschaftsnachweis. Theoretische Grundlagen. Mitt. Anthropol. Ges. Wien **68:** 9–53.

Feller, W., 1968 *An Introduction to Probability Theory and Its Applications.* Wiley, New York.

Foltz, D. W., and J. L. Hoogland, 1981 Analysis of mating system in the black-tailed prairie dog (*Cynomys ludovicianus*) by likelihood of paternity. J. Mammal. **62:** 706–712.

Gürtler, H., 1956 Principles of blood group statistical evaluation of paternity cases at the University Institute of Forensic Medicine, Copenhagen. Acta Med. Leg. Soc. **9:** 83–93.

Hamrick, J. L., and A. Schnabel, 1985 Understanding the genetic structure of plant populations: some old problems and a new approach. pp. 50–70. In: *Population Genetics in Forestry,* Edited by H.-R. Gregorius. Springer-Verlag, New York.

Hanken, J., and P. W. Sherman, 1981 Multiple paternity in Belding's ground squirrel litters. Science **212:** 351–353.

Hill, W. G., 1987 DNA fingerprints applied to animal and bird populations. Nature **327:** 98–99.

Jeffreys, A. J., and D. B. Morton, 1987 DNA fingerprints of dogs and cats. Anim. Genet. **18:** 1–15.

Jeffreys, A. J., V. Wilson and S. L. Thein, 1985 Individual-specific 'fingerprints' of human DNA. Nature **316:** 76–79.

Jeffreys, A. J., V. Wilson, R. Kelly, B. A. Taylor and G. Bulfield, 1987 Mouse DNA 'fingerprints': analysis of chromosome localization and germ line stability of hypervariable loci in recombinant inbred strains. Nucleic Acid Res. **15:** 2823–2836.

Li, C. C., and A. Chakravarti, 1985 Basic fallacies in the formulation of the paternity index. Am. J. Hum. Genet. **37:** 809–818.

MacCluer, J., and W. J. Schull, 1963 On the estimation of the frequency of nonpaternity. Am. J. Hum. Genet. **15:** 191–202.

Majumder, P. P., and M. Nei, 1983 A note on positive identification of paternity by using genetic markers. Hum. Hered. **33:** 29–35.

McCracken, G. F., and J. W. Bradbury, 1977 Paternity and genetic heterogeneity in the polygynous bat *Phyllostomus hastatus*. Science **198:** 303–306.

Meagher, T. R., 1986 Analysis of paternity within a natural population of *Charmaelirium lutem*. I. Identification of most-likely parents. Am. Nat. **128:** 199–215.

Meagher, T. R., and E. Thompson, 1986 The relationship between single parent and parent pair genetic likelihoods in genealogy reconstruction. Theor. Popul. Biol. **29:** 87–106.

Mickey, M. R., D. W. Gjertson and P. I. Terasaki, 1986 Empirical validation of the Essen-Moller probability of paternity. Am. J. Hum. Genet. **39:** 123–132.

Nakamura, Y., M. Leppert, P. O'Connell, R. Wolfe, T. Holm, M. Culver, C. Martin, E. Fujimoto, M. Hoff, E. Kumlin and R. White, 1987 Variable number of tandem repeat (VNTR) markers for human gene mapping. Science **235:** 1616–1622.

Quinn, T. W., J. S. Quinn, F. Cooke and B. N. White, 1987 DNA marker analysis detects multiple maternity and paternity in single broods of the lesser snow goose. Nature **326:** 392–394.

Salmon, D. B., and J. Brocteur, 1978 Probability of paternity exclusion when relatives are involved. Am. J. Hum. Genet. **30:** 65–75.

Selvin, S., 1980 Probability of nonpaternity determined by multiple allele codominant systems. Am. J. Hum. Genet. **32:** 276––278.

Smith, D. B., and W. T. Adams, 1983 Measuring pollen contamination in clonal seed orchards with the aid of genetic markers. pp. 69–77. In: *Proceedings of the 17th Southern Forest Tree Improvement Conference.* University of Georgia, Athens.

Smouse, P. E., and R. Chakraborty, 1986 The use of restriction fragment length polymorphisms in paternity analysis. Am. J. Hum. Genet. **38:** 918–939.

Stanton, M., 1986 Unveiling the mystery of plant paternity. Trends Ecol. Evol. **1:** 116–117.

Thompson, E., 1975 The estimation of pairwise relationship. Ann. Hum. Genet. **39:** 173–188.

Thompson, E., 1976a Inference of genealogical structure. II. Quantifying genetic information. Soc. Sci. Inform. **15:** 491–506.

Thompson, E., 1976b Inference of genealogical structure. III. The reconstruction of genealogies. Soc. Sci. Inform. **15:** 507–526.

Thompson, E., 1986 Likelihood inference of paternity. Am. J. Hum. Genet. **39:** 285–287.

Valentin, J., 1984 Paternity index and attribution of paternity. Hum. Hered. **34:** 255–257.

Walker, R. H., 1983 *Inclusion Probabilities in Parentage Testing.* American Association of Blood Banks, Arlington, Va.

Wetton, J. H., R. E. Carter, D. T. Parkin and D. Walters, 1987 Demographic study of a wild house sparrow population by DNA fingerprinting. Nature **327:** 147–149.

Communicating editor: M. T. Clegg

APPENDIX

Derivations of Equations 8 and 9: Note that from Equation 3, we have

$$
\begin{aligned}
[P_E(C)]^{N-1} &= \left[1 - \prod_{l=1}^{L} (1 - PE_l)\right]^{N-1} \\
&= \sum_{r=0}^{N-1} \binom{N-1}{r} (-1)^r \left[\prod_{l=1}^{L} (1 - PE_l)\right]^r \\
&= \sum_{r=0}^{N-1} \binom{N-1}{r} (-1)^r \prod_{l=1}^{L} (1 - PE_l)^r \qquad \text{(A1)}
\end{aligned}
$$

Taking expectations of both sides of Equation (A1), and using the fact that the L loci are independent, we have Equation 8 of the text.

To derive Equation 9, note that for the lth locus,

$$(1 - PE_l)^r = \sum_{t=0}^{r} \binom{r}{t} (-1)^t \{PE_l\}^t,$$

in which PE_l has a distribution given by Equation 1, where p_i and q_i values represent the allele frequencies in males and females for the lth locus (the suffix l is suppressed by simplicity of notation).

Therefore,

$$
\begin{aligned}
E[\{PE_l\}^t] &= \sum_{i=1}^{k} (1 - p_i)^{2t} p_i (1 - q_i + q_i^2) \\
&\quad + \sum_{j>i=1}^{k}\sum (1 - p_i - p_j)^{2t} q_i q_j (p_i + p_j).
\end{aligned}
$$

Using the binomial expansions for $(1 - p_i)^{2t}$ and $(1 - p_i - p_j)^{2t}$, we have

$$
\begin{aligned}
(1 - p_i)^{2t} &= \sum_{v=0}^{2t} \binom{2t}{v} (-1)^v p_i^v \\
(1 - p_i - p_j)^{2t} &= \sum_{v=0}^{2t} \binom{2t}{v} (-1)^v (p_i + p_j)^v.
\end{aligned}
$$

This, in turn, yields

$$E[\{PE_l\}^t] = \sum_{i=1}^{k} \sum_{v=0}^{2t} \binom{2t}{v} (-1)^v p_i^{v+1}(1 - q_i + q_i^2) + \sum\sum_{j>i=1}^{k} \sum_{v=0}^{2t} \binom{2t}{v} (-1)^v (p_i + p_j)^{v+1} q_i q_j. \tag{A2}$$

Again, if we use the binomial expansion

$$(p_i + p_j)^{v+1} = \sum_{w=0}^{v+1} \binom{v+1}{w} p_i^w p_j^{v+1-w},$$

and note that

$$\sum_{w=0}^{v+1} \binom{v+1}{x} = 2^{v+1},$$

algebraic simplification of (A2) leads to Equation 9 with the notation

$$a_{r,s} = \sum_{i=1}^{k} p_i^r \cdot q_i^s.$$

Derivations of Equations 10 and 11: In a population of N adult males, we assume that for each birth, the biological father is one of these males. From the genotypic combinations of mother-offspring (M, O) pairs, however, not all fathers can be unambiguously determined. If $P_E(C)$ represents the probability of exclusion obtained for a specific (M, O) pair, combining information on all L loci, the probability that exactly m, $(0 < m < N)$, males are not excluded is given by the binomial expression

$$\binom{N-1}{m} [1 - P_E(C)]^{m-1} [P_E(C)]^{N-m},$$

$$\text{for } m = 0, 1, 2, \ldots, N - 1;$$

since one of the N males is the true father.

However, since $P_E(C)$ varies over the different (M, O) pairs, the probability of m genotypically possible male parents for a large array of offspring in the population is given by

$$\binom{N-1}{m} E\{[1 - P_E(C)]^{m-1} [P_E(C)]^{N-m}\},$$

which reduces to Equation 11, since

$$[1 - P_E(C)]^{m-1} = \sum_{s=0}^{m-1} \binom{m-1}{s} (-1)^s [P_E(C)]^s.$$