Parentage exclusion of close relatives in haplodiploid species

Jinliang Wang[1]*, Andrew F. G. Bourke[2]

[1]*Institute of Zoology, Zoological Society of London, London NW1 4RY, United Kingdom*

[2]*School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, NR4 7TJ, United Kingdom*

Left running head: J Wang

Right running head: Parentage exclusion of relatives

Key words: Parentage exclusion, Markers, Close relatives, Full siblings, Maternity

**Corresponding author:*

Jinliang Wang

Institute of Zoology

Regent's Park

London NW1 4RY

United Kingdom

Tel: 0044 20 74496620

Fax: 0044 20 75862870

Email: jinliang.wang@ioz.ac.uk

ABSTRACT

Parentage exclusion probability is usually calculated to evaluate the informativeness of a set of markers for, and the statistical power of, a parentage analysis. Equations for parentage exclusion probability have been derived in various scenarios such as paternity exclusion when maternity is known or unknown or when candidate males are unrelated or loosely related (being from the same subpopulation) to the father. All previous work assumes a diploid species. Although marker-based parentage analyses have been conducted in haploidiploid species (such as ants, bees and wasps) for diploid offspring at the individual level or haploid offspring at the class level, rigorously derived formulations of parentage exclusion probability for haploid offspring at the individual level are lacking, which prevents the precise evaluation of the informativeness for and the statistical power of a parentage analysis. In this study we derive equations for the exclusion probability of maternity of a haploid male when multiple mother candidates (workers or queens) are unrelated or fullsibs to the mother. The usefulness of the equations is exemplified by numerical examples, and the results are discussed in the context of the study of worker reproductivity in eusocial haplodiploid species. The results are especially valuable for an optimal experimental design in determining sampling intensities (e.g. number of markers and number of individuals) to achieve satisfactory statistical power of a parentage analysis in investigating workers' reproductivity in eusocial haplodiploid species.

1. Introduction

Genetic marker-based parentage analysis has been widely applied in human and wildlife forensics (Ogden et al., 2009), in studies of social behaviour, social organization, reproductive success, mating systems, dispersal and spatial genetic structure in natural populations of wild species (Hughes, 1998; Coltman et al., 1999; Garant et al., 2001; Avise et al., 2002; Robledo-Arnuncio and Gil, 2005; Bretman and Tregenza, 2005), in the conservation management of endangered species in captivity and in the wild (Moran et al., 2021), and in the selective breeding of domestic animals and crops (Heaton et al., 2014). Both exclusion and likelihood approaches have been developed to assign the parentage of an offspring to a candidate using the genotype data of the individuals at some marker loci (Flanagan and Jones, 2019).

In the experimental design stage of a parentage study in determining, among other things, the appropriate sampling intensities of markers and individuals, a statistic called parentage exclusion probability (P_E) is usually calculated to evaluate the marker informativeness for, and the statistical power of, a parentage analysis (e.g. Dodds et al., 1996; Jamieson and Taylor, 1997). P_E is usually defined as the average probability that a randomly selected individual is excluded from the parentage of a randomly selected offspring based on their genotypes at a set of marker loci. The individual is excluded from the parentage of the offspring if they have genotypes that mismatch at one or more marker loci, and is unexcluded if they have completely matched genotypes. A low P_E means an individual who is unrelated to an offspring is excludable from the parentage of the offspring at a low probability, signifying that the set of markers used in calculating P_E is not informative and the parentage analysis using the markers is powerless. In contrast, a high P_E means an individual who is unrelated to an offspring is excludable from the parentage of the offspring at a high probability, signifying that the set of markers has sufficient information and the parentage analysis using the markers is powerful enough to yield accurate parentage assignments. Although P_E is based on exclusion, it is relevant to a parentage analysis regardless of the methods, exclusion or likelihood.

Formulas for P_E have been derived in the literature for diallelic (e.g. Wiener et al., 1930) and multiallelic (e.g. Jamieson, 1965; Ohno et al., 1982; Dodds et al., 1996) markers, for excluding an individual who is unrelated to (e.g. Jamieson and Taylor, 1997) or is loosely related to (being from the same subpopulation, e.g. Ayres, 2002) the sampled individuals involved in a parentage analysis, and for excluding a close relative of the true parent (MacCluer and Schull, 1963; Salmon and Brocteur, 1978; Thompson and Meagher, 1987; Double et al., 1997; Fung et al., 2002; Hu et al., 2005). Among many insights gleaned from these formulations, it was shown that the exclusionary capability of a set of markers is much reduced by genetic relatedness between the alleged and true parents.

In a parentage analysis, usually many individuals are candidates for parentage assignment of an offspring, and all but one must be excluded before the unexcluded one can be confidently assigned parentage. When these candidates are unrelated to the true parent and unrelated among themselves, the probability of multiple candidate exclusions of an offspring can be easily calculated from that of single exclusion (Chakraborty et al., 1988), P_E; The probability of excluding n such unrelated candidates is simply $P_{E(n)} = (P_E)^n$. When multiple candidates are close relatives, say fullsibs, to the true parent, however, the probability of excluding all of them can no longer be calculated from the probability of excluding a single candidate. Until now there have been no algebraic derivations of the probability of excluding multiple (n) close relatives to the true parent in a parentage analysis, $P_{E(n)}$. In the absence of formulas, Double et al. (1997) used simulations instead to evaluate $P_{E(n)}$ for excluding multiple relatives to the true parent in diploid species.

Previous studies calculating P_E, as well as empirical applications, assume diploid species (e.g. Dodds et al., 1996; Double et al., 1997; Jamieson and Taylor, 1997). However, it is also desirable to conduct parentage analysis in eusocial insects including the ants, bees and wasps (Hymenoptera), all of which are haplodiploid. In some studies of eusocial Hymenoptera, individual-level parentage analyses have been conducted assigning diploid offspring to potential mothers, e.g. worker offspring to coexisting mother queens (Hammond et al., 2006). In many species of eusocial Hymenoptera, workers are capable, through haplodiploidy, of producing unfertilized eggs that develop into males, which may occur in the queen's presence (queenright conditions) or more frequently in colonies consisting of workers remaining after the mother queen has died (queenless conditions) (Bourke, 1988; Ratnieks et al., 2006; Friend and Bourke, 2014). Nearly all studies of male parentage in eusocial Hymenoptera have assigned males to queens or workers as a class (e.g. Foster et al., 2001; Hammond et al., 2003; Alaux et al., 2004). A single pioneering study of the ant *Pachycondyla villosa* aimed to assign males to individual worker parents (Trunzer at al., 1999), and this used experimentally-established groups of unrelated workers although the study species, as is almost universal in eusocial Hymenoptera, lives in colonies of related workers. The lack of studies aiming to exclude related reproductive workers as parents of worker-produced males at the individual level is attributable to the inherent difficulty of performing such exclusions when potential worker mothers may be related to one another by values as high as 0.75 (i.e. full sister relatedness in workers produced by a single, once-mated queen). In fact, no analytical expressions have previously been derived to perform such exclusions, despite potential applications for understanding the distribution of direct fitness (i.e. individual production of sons) among workers in the same eusocial colony.

In this study, we derive equations for the probability of excluding (for parentage) an arbitrary number of N (>0) workers who are either unrelated or related as full sibs to the true mother of a male using codominant marker data. In the latter case, all candidate mothers as well as the true mother of a male are full sib workers who are all daughters of

one, singly-mated queen, and therefore a candidate mother is an aunt of the male. The validity of the equations is verified by simulations and the power of excluding multiple aunts as the mother of a male is investigated using some numerical examples. The results should be useful for the experimental design and execution of studies of worker reproduction by a parentage analysis of marker data at the individual worker level in eusocial Hymenoptera, and should be valuable in assessing the informativeness of markers for, and the power of parentage analysis in, haplodiploid species in general.

2. Derivation of exclusion probability

We consider the exclusion of maternity of a haploid male in haplodiploid species first when the candidate diploid females are unrelated among themselves and are unrelated to the true mother, and then when the candidate diploid females are full siblings to the true mother (i.e. the aunts of the male). The case of unrelated candidate diploid females is rare in eusocial Hymenoptera, which typically consist of colonies of related queens and workers (Ross, 2001; Rubenstein and Abbot, 2017), but is nonetheless included in this study for comparison with the focal case of full-sib candidate females. This focal case arises when worker offspring of one, singly-mated queen produce male offspring in queenless conditions, these males being grandsons of the departed queen, as occurs relatively frequently in eusocial Hymenoptera (see Introduction).

2.1 Excluding candidates unrelated to the true mother of a male

We assume N candidate females (workers or queens) compete for maternity assignment to a male, and these females are unrelated among themselves and unrelated to the true mother. Females are diploid, while males are haploid and have developed from unfertilized eggs laid by a female (queen or worker). All individuals are genotyped at L codominant marker loci. In the absence of mutations and genotyping errors, the allele of a male at each locus should be found in its mother genotype. In other words, the true mother will always have a genotype compatible with that of its male offspring at all loci. Specifically, at a given locus, a female of genotype A_iA_j will have sons of genotypes A_i or A_j at frequencies of ½ and ½, respectively. If a candidate female is not the mother of a male, then it possibly has genotypes incompatible or mismatched with those of the male at one or more loci. When such an event occurs, the candidate is excluded from the maternity of the male. Otherwise, it is not excluded. When all but one of the N candidate mothers are excluded of the maternity of a male and the markers used in the analysis is sufficiently informative, then the male's maternity is assigned confidently to the unexcluded candidate. Given a set of markers with known allele frequencies, we calculate the average probability that a randomly selected female (worker or queen) who is unrelated to the true mother is excluded from the maternity of a randomly selected male, P_{E1}. The value of P_{E1} (from 0 to 1) signifies the information content of the set of markers for, and measures the power of, a parentage analysis.

Consider a locus l with k_l codominant alleles, A_i, for i=1, 2, …, k_l. The frequency of allele A_i in the population is denoted by p_{li}, with $p_{li} > 0$ and $\sum_{i=1}^{k_l} p_{li} = 1$. Therefore, a

male taken at random from the population will have allele A_i with probability p_{li}. Given the male genotype, a random candidate can be excluded as the mother of the male if its genotype does not contain allele A_i, which will occur with probability $(1-p_{li})^2$ in a population with random mating. The overall exclusion probability considering all possible genotypes of the male is $\sum_{i=1}^{k_l} p_{li}(1-p_{li})^2$.

Now consider a number of L loci. Exclusion occurs when the male and the candidate mother have incompatible genotypes at 1 or more loci. The exclusion probability considering L loci is thus

$$P_{E1} = 1 - \prod_{l=1}^{L}\left(1 - \sum_{i=1}^{k_l} p_{li}(1-p_{li})^2\right). \quad (1)$$

This formula gives the probability that a female (worker or queen) is excluded from the maternity of a male based on their genotypes at L loci with known allele frequencies p_{li} (for l=1, 2, ...L; i=1, 2, ..., k_l). (1) can be further simplified to

$$P_{E1} = 1 - \prod_{l=1}^{L}(2a_{l2} - a_{l3}), \quad (2)$$

where a_{lb} is the sum of powers of allele frequencies at locus l, with $a_{lb} = \sum_{i=1}^{k_l} p_{li}^b$ for $b = 2, 3$.

Now consider the probability of excluding N random workers who are not the mother and are unrelated to the mother of a male. It is simply calculated from (2) as

$$P_{E1(N)} = (P_{E1})^N = (1 - \prod_{l=1}^{L}(2a_{l2} - a_{l3}))^N. \quad (3)$$

$P_{E1(N)}$ depends on the allele frequencies at each of the L loci. It is maximized when each locus l (l=1, 2, ..., L) has k_l equal-frequency alleles (i.e. $p_{li} = 1/k_l$ for i=1, 2, ..., k_l). In such a situation, $2a_{l2} - a_{l3}$ in (3) is minimized to $(2k_l - 1)/k_l^2$ (Appendix 1), and $P_{E1(N)}$ is maximized to

$$P_{E1(N)} = \left(1 - \prod_{l=1}^{L}\frac{2k_l - 1}{k_l^2}\right)^N. \quad (4a)$$

When all L loci have the same number of alleles, k, and the same equal allele frequency of $1/k$, the exclusion probability is further simplified to

$$P_{E1(N)} = \left(1 - \left(1 - \left(1 - \frac{1}{k}\right)^2\right)^L\right)^N. \quad (4b)$$

2.2 Excluding candidates who are full siblings to the true mother of a male

Excluding aunts (i.e. full-sibling to the mother) from being assigned as the mother of a haploid male is much more difficult, as they have genotypes similar to that of the true mother and thus inclined to be compatible with those of the male. Specifically, at a given locus, workers that are daughters of a mother queen of genotype A_iA_j, who has mated singly with a male of genotype A_m, will be of genotypes A_iA_m or A_jA_m at frequencies of ½

and ½, respectively, and their male offspring will be of genotypes A_i, A_j or A_m at frequencies of ¼, ¼ and ½, respectively. Most often a male may have a genotype (allele) at a locus that produces no maternity exclusion of its aunts. Occasionally, however, it may have a genotype (allele) that is absent from the genotypes of some aunts who are then excluded from the maternity of the male. A schematic illustration of a pedigree in which some males may and others may not allow maternity exclusion of their aunts is shown in Figure 1. In general, the exclusion power of a single locus is rather poor, and many loci are needed to confidently exclude assigning maternity to the aunts of males.

It is more difficult to derive the equation for excluding multiple aunts from the maternity of a male, because the aunts are highly related (relatedness 0.75) among themselves, and thus cannot be considered independently as in the previous case involving multiple unrelated workers taken at random from the population. Many more markers are therefore required to provide sufficient information for excluding multiple aunts from the maternity of a male.

Consider a locus l having k_l codominant alleles A_i with frequencies p_{li} for i=1, 2, …, k_l. A male taken at random from the population will have allele A_i with probability p_{li}. It could come from a mother produced from four possible grandparent mating types (Table 1). Only two of the four grandma-grandpa mating types (i.e. the mating type of [i] the queen producing the workers and [ii] the queen's mate) could produce a grandson that allows the exclusion of its aunts being assigned as the mother. The two mating types, together with the pooled type of matings that do not allow maternity exclusion, are detailed below.

2.2.1. Grandma-grandpa mating type 1: $A_iA_j \times A_i$

This produces two types of workers, A_iA_j and A_iA_i, at an equal frequency of ½, as depicted in Figure 1. A male from the workers of this mating type has a genotype A_j with a probability of ¼. This is the only male type that allows exclusion of its aunts when they display the genotype A_iA_i (i.e. having no male allele, A_j, in their genotype).

The overall frequency of this mating type is

$$q_1 = \sum_{i=1}^{k_l} 2p_{li}^2(1 - p_{li}) = 2(a_{l2} - a_{l3}),$$

where a_{lb} is the sum of powers of allele frequencies at locus l as shown above. For equal allele frequency $p_{li} = \frac{1}{k}$ at a locus with k alleles, q_1 reduces to

$$q_1 = \frac{2(k-1)}{k^2}.$$

2.2.2. Grandma-grandpa mating type 2: $A_iA_j \times A_m$

In this mating type, the grandpa has an allele, A_m, different from any of the two alleles of the grandma's heterozygous genotype, A_iA_j. This mating type occurs only when a locus

has more than two alleles. It produces two types of workers, A_iA_m and A_jA_m, at an equal frequency of ½. A male from a A_iA_m worker has a genotype A_i with a probability of ½, and this male allows the exclusion of an aunt when she has the genotype A_jA_m. Similarly, a male from a A_jA_m worker has a genotype A_j with a probability of ½, and this male allows the exclusion of an aunt when she has the genotype A_iA_m.

The overall frequency of this mating type is

$$q_2 = \sum_{i=1}^{k_l} \sum_{j=i+1}^{k_l} 2p_{li}p_{lj}\left(1 - p_{li} - p_{lj}\right) = 1 - 3a_{l2} + 2a_{l3}.$$

For a locus with k equal-frequency alleles, $p_{li} = \frac{1}{k}$, the overall frequency of this mating type reduces to

$$q_2 = 1 - \frac{3}{k} + \frac{2}{k^2}.$$

2.2.3. Grandma-grandpa mating type 3: All others

The rest of the mating types are pooled to form mating type 3, which does not allow any exclusion. Grandmas from this pooled mating type are always homozygotes and thus all females (i.e. the mother and aunts of a male) produced from the mating type are of the same genotype (Table 1). The frequency of this pooled mating type is

$$q_3 = 1 - q_1 - q_2 = 1 - 2(a_{l2} - a_{l3}) - (1 - 3a_{l2} + 2a_{l3}) = a_{l2}.$$

For a locus with k equal-frequency alleles, $p_{li} = \frac{1}{k}$, it reduces to

$$q_3 = \frac{1}{k}.$$

2.2.4. Summing the 3 mating types

The three mating type frequencies sum to 1, $q_1 + q_2 + q_3 \equiv 1$, as expected. The relative frequencies of mating types 1, 2 and 3 depend on the number and frequencies of alleles at a locus. For diallelic markers, we have $q_2 \equiv 0$ and $q_3 \geq q_1$ with $q_3 - q_1 = (1 - 2p_{l1})^2$. When $p_{l1} = p_{l2} = 0.5$, $q_3 = q_1 = 0.5$. Otherwise, $q_3 > q_1$, and the difference increases with an increasing departure from the equifrequency $p_{l1} = p_{l2} = 0.5$. With more than 2 alleles at a locus, $q_2 > 0$ and the sum of frequencies of exclusion-permitting mating types, $q_1 + q_2 = 1 - a_{l2}$, is always larger than $q_3 = a_{l2}$. The higher is the polymorphism (with a larger k and a more even allele frequency distribution) of a marker, the greater are the frequencies of exclusion-permitting mating types and thus the higher is the information content of the marker for parentage analysis. This is an intuitive conclusion that is partially verified by numerical analysis (Figures 2 and 3 in Results below).

Now consider a male genotype at L loci, each having k alleles of the same frequencies $\{p_1, p_2, \ldots, p_k\}$ (the subscript l for locus is thus dropped out hereafter). The probability that, among the L alleles in a male genotype, n_1, n_2 and $n_3 = L - n_1 - n_2$

come from grandma-grandpa mating types 1, 2, and 3 follows the multinomial distribution

$$P[n_1, n_2, n_3] = \frac{L!}{n_1! n_2! n_3!} q_1^{n_1} q_2^{n_2} q_3^{n_3}.$$

Suppose a male has a genotype with n_1 alleles (loci) coming from grandma-grandpa mating type 1. The probability that, among these n_1 loci, each of n_{11} (=0,1,..., n_1) loci has an allele permitting exclusion follows a binomial distribution, and can be derived from Table 1 as

$$R_1[n_{11}, n_1 - n_{11}] = \frac{n_1!}{n_{11}!(n_1 - n_{11})!} \left(\frac{1}{4}\right)^{n_{11}} \left(\frac{3}{4}\right)^{n_1 - n_{11}}.$$

Similarly, the probability that, among the n_2 loci of mating type 2, each of n_{21} (=0,1,..., n_2) loci has an allele permitting exclusion is

$$R_2[n_{21}, n_2 - n_{21}] = \frac{n_2!}{n_{21}!(n_2 - n_{21})!} \left(\frac{1}{2}\right)^{n_{21}} \left(\frac{1}{2}\right)^{n_2 - n_{21}} = \frac{n_2!}{n_{21}!(n_2 - n_{21})!} \left(\frac{1}{2}\right)^{n_2}.$$

Given a male with n_{11} and n_{21} loci displaying exclusionary alleles coming from mating type 1 and 2 respectively, the probability that its N aunts are excluded is

$$Q[n_{11}, n_{21}] = \left(1 - \left(\frac{1}{2}\right)^{n_{11} + n_{21}}\right)^N.$$

The overall exclusion probability considering all possible male genotypes from all possible grandma-grandpa mating types is

$$P_{E2(N)} = \sum_{n_1=0}^{L} \sum_{n_2=0}^{L - n_1} P[n_1, n_2, L - n_1 - n_2] \times$$

$$\sum_{n_{11}=0}^{n_1} R_1[n_{11}, n_1 - n_{11}] \sum_{n_{21}=0}^{n_2} R_2[n_{21}, n_2 - n_{21}]\, Q[n_{11}, n_{21}]$$

$$= \sum_{n_1=0}^{L} \sum_{n_2=0}^{L - n_1} \frac{L!}{n_1!\, n_2!\, (L - n_1 - n_2)!} q_1^{n_1} q_2^{n_2} q_3^{L - n_1 - n_2} \sum_{n_{11}=0}^{n_1} \frac{n_1!}{n_{11}!\, (n_1 - n_{11})!} \left(\frac{1}{4}\right)^{n_{11}} \left(\frac{3}{4}\right)^{n_1 - n_{11}}$$

$$\sum_{n_{21}=0}^{n_2} \frac{n_2!}{n_{21}!(n_2 - n_{21})!} \left(\frac{1}{2}\right)^{n_2} \left(1 - \left(\frac{1}{2}\right)^{n_{11} + n_{21}}\right)^N. \quad (5)$$

In equation (5), a male genotype has n_1, n_2 and $L - n_1 - n_2$ alleles (loci) coming from grandma-grandpa mating type 1, 2 and 3, respectively. Therefore, n_1 varies between 0 and L (number of loci), while n_2 varies between 0 and $L - n_1$. Among these n_1 loci from mating type 1, n_{11} (=0,1,..., n_1) loci have alleles that make it possible to exclude maternity. Similarly, among these n_2 loci from mating type 2, n_{21} (=0,1,..., n_2) loci have alleles that make it possible to exclude maternity. q_1, q_2 and q_3 are the frequencies of

grandpa with grandma mating types 1, 2 and 3 as described in sections 2.2.1, 2.2.2 and 2.2.3. The computational load of (5) increases rapidly with L, and becomes substantial even when L is as small as 10. To facilitate its application, (5) is implemented in software (see below).

For the diallelic marker case, $q_2 \equiv 0$ and $n_2 \equiv 0$ (such that the sums over n_2 and over n_{21} are both empty), and (5) reduces to

$$P_{E2(N)} = \sum_{n_1=0}^{L} \frac{L!}{n_1!(L-n_1)!} q_1^{n_1} q_3^{L-n_1} \sum_{n_{11}=0}^{n_1} \frac{n_1!}{n_{11}!(n_1-n_{11})!} \left(\tfrac{1}{4}\right)^{n_{11}} \left(\tfrac{3}{4}\right)^{n_1-n_{11}} \left(1-\left(\tfrac{1}{2}\right)^{n_{11}}\right)^{N}. \quad (6)$$

3. Simulations

To check the validity of the formula, we conducted some Monte Carlo simulations. For these numerical examples, the formula and simulations agree very well (Table 2). For a given allele frequency distribution (equal frequency, or frequencies in a triangular distribution where the frequency of allele j, p_j, is proportional to j for j = 1, 2, ..., k at a locus with k alleles), exclusion probabilities increase rapidly with: a decreasing number of potential candidate parents (aunts); an increasing number of loci; and an increasing number of alleles per locus. At the same number of alleles at a locus, the same number of loci and the same number of candidates, equal allele frequency distribution leads to a substantially higher exclusion probability than a triangular distribution of allele frequencies.

Simulations were also conducted to investigate the impact of genome size (linkage) on the exclusion probabilities of a set of markers. In deriving equations (3) and (5), we assumed no linkage among the markers. This is a very good approximation when the number of markers (L) is small. However, with a large L, some markers might be physically linked (located on the same chromosome), and (3) and (5) may overestimate the exclusionary power of the L markers. To understand the impact of linkage, we simulated a genome of various map lengths (m, from 1 to 32 Morgans) and assumed the L markers are equally spaced in the genome. The number of crossovers in generating a gamete from a diploid female was drawn from a Poisson distribution with parameter m, and the locations where crossovers occurred were randomly chosen without interference between different crossover events. The results in Table 2 show that linkage can decrease the exclusionary power of a set of L markers substantially when roughly $3m < L$, which means that when the genome size m is small or the number of markers L is large such that on average 3 or more markers are located on 1 Morgan of the genome. For the range of L (10-80) considered in the simulation, a genome with $m \geq 16$ is hardly affected by linkage in determining exclusion probabilities.

4. Results

4.1 Exclusion of candidates unrelated to the true mother

Confirming the proof in Appendix 1, the maximal exclusion probability is attained when all alleles at a locus *l* have the same frequency of $\frac{1}{k_l}$. For the case of a diallelic marker ($k_l = 2$), for example, P_{E1} as a function of allele frequencies is shown in Figure 2. A locus with rare alleles contributes little to maternity exclusion. For L diallelic loci with allele frequencies $\{p, 1-p\}$, P_{E1} reduces to $P_{E1} = 1 - (1 - p + p^2)^L$. The number ($L$) of loci with a rare allele frequency p required to attain the same exclusion probability as a single diallelic locus with equal allele frequency (0.5) can be solved from the equation $1 - (1 - p + p^2)^L = 1 - (1 - 0.5 + 0.5^2)^1$. Figure 3 plots L as a function of p, where L is solved from the previous equation. L increases loglinearly with a decreasing p. Loci with rare alleles have little exclusionary power. For example, about 6 diallelic loci with allele frequencies (0.05, 0.95) or 29 diallelic loci with allele frequencies (0.01, 0.99) have roughly the same exclusion power, $P_{E1} = 0.25$, as a single diallelic locus with equifrequent (0.5, 0.5) alleles.

Some numerical examples of (4) are shown in Figure 4. For markers with equifrequent alleles, the number of alleles of a locus has a large impact on the exclusion probability. A diallelic locus affords the smallest amount of information for maternity exclusions. It can be shown using (4) that a locus with 10 equifrequent alleles has the same exclusionary capability as 2.8 triallelic loci with equifrequency or 5.8 diallelic loci with equifrequency. $P_{E1(N)}$ decreases rapidly with an increasing N and a decreasing L.

4.2 Exclusion of candidates who are full sibs to the true mother

The results from (5) for some parameter combinations are shown in Figure 5. Similar to the case of excluding candidate females who are unrelated to the mother, the probability of excluding aunts as mother increases rapidly with both the number of loci (L) and the number of alleles per locus (k). However, excluding aunts is much more difficult than excluding females unrelated to the mother. To attain the same exclusion power, many more loci are necessary when the candidate females are aunts rather than unrelated individuals. For markers with each having 5 or more equifrequent alleles, about 10 loci are required to yield a probability 0.99 for excluding 100 unrelated females. However, when the candidate females are aunts, about 50 such loci are required to yield the same exclusion power. Diallelic markers, such as SNPs, have a much reduced exclusion power than multiallelic markers. For the above example, about 150 diallelic loci with equifrequent alleles are required to exclude 100 aunts as the mother at a probability of 0.99. This is because, when k=2, only one mating type ($A_iA_j \times A_i$) instead of two (when $k > 2$) can generate males that may allow exclusions.

Similar to the case of unrelated candidates, the maximal exclusion probability is attained when all alleles at a locus *l* have the same frequency of $\frac{1}{k_l}$. Figure 2 shows the probability of excluding N = 10 aunts from the maternity of a male using L=10, 20, 40 and 160 diallelic markers ($k_l = 2$) as a function of allele frequencies. Again, a locus with a rare allele contributes little to maternity exclusion. For example, to exclude 10 aunts

from the maternity of a male at a probability of 0.996 would require 120, 162, 340 and 650 diallelic markers with each having an allele frequency equal to 0.5, 0.25, 0.1 and 0.05, respectively. Compared with the exclusion of unrelated candidate workers (Figure 2), excluding aunts from maternity of a male is much more difficult. At the same allele frequency distribution, roughly L and $4L$ diallelic markers afford the same probability of excluding N unrelated candidates and N aunts of a male, respectively.

5. Discussion

Although marker-based parentage analyses have been conducted in haplodiploid species (for diploid offspring at the individual level and haploid offspring at the class level), previously there has been no study in haplodiploids on the average probability of marker-based parentage exclusion for haploid offspring at the individual level. All studies on parentage exclusion at the individual level in the literature assume a diploid species (see Introduction). Herein we investigated the exclusion of maternity of a haploid male in haplodiploid species when multiple candidate mothers to be excluded are unrelated or are full siblings to the mother. These equations are especially useful for the study of worker reproductivity, where many fullsib females (e.g. workers produced by one, singly-mated queen) may compete for maternity of a male.

As shown by the numerical examples (Figures 2 and 3), a marker with equifrequent alleles affords the maximal exclusionary power. A marker with rare alleles (i.e. allele frequencies close to zero) holds little exclusion capability. This is understandable from an inspection of Table 1, which shows that a male from a homozygous grandma or a homozygous mother does not allow maternity exclusion of any aunts. Homozygosity is expected to increase with allele frequencies departing increasingly from an equifrequent distribution or with the frequency of one allele approaching 1 and the frequencies of the other alleles approaching 0.

As is the case for parentage exclusion in diploid species (Salmon and Brocteur, 1978; Double et al., 1997), maternity exclusion of a male in haplodiploid species becomes much more difficult when the female candidates are fullsibs of the mother rather than random individuals unrelated to the mother. This is because, being from the same pair of parents, aunts and the true mother share similar genotypes. Hence aunts are more likely to have genotypes compatible with those of their nephew than unrelated females drawn at random from a population. Therefore, hundreds of diallelic loci are required to exclude 100 aunts as the mother of a male at a probability of 0.99. However, nowadays SNPs from next generation sequencing can easily provide hundreds of diallelic loci for parentage and similar analyses (Helyar et al., 2011), and as studies characterising SNPs in eusocial Hymenoptera grow in number (e.g. Theodorou et al., 2018; Southon et al., 2019), parentage analyses involving haploid offspring using hundreds of markers will become increasingly feasible.

Following previous work we assume marker data are perfect in maternity exclusion analysis. Unfortunately, in reality, genotyping errors and mutations are rules

rather than exceptions. Regardless of genotyping methods (e.g. by PCR for microstellites or by sequencing for SNPs), typing errors are ubiquitous (Pompanon et al., 2005). False maternity exclusion might occur because the mother's genotype and the male's genotype may mismatch at one or more loci due to genotyping abnormality or mutations. To reduce false exclusion, a common convention is to exclude putative mothers only when they have genotypes that mismatch with male genotypes at two or more loci. By making this mismatch allowance, the exclusionary power of a set of markers could be substantially reduced (Double et al., 1997). To obtain a given probability (say, 0.99) of excluding all false mothers, therefore, a few more markers than that determined from (3) or (5) for the case of perfect markers would be required.

The formula for exclusion probabilities, (3) and (5), are derived by assuming the absence of linkage among markers. However, in the case of many markers (L) in a small genome (m Morgans in genetic map length), some of the markers are inevitably located on the same chomosome and are thus physically linked. As shown by our simulations, linked markers could have a substantially reduced exclusionary power compared to unlinked markers (Table 2). When roughly $L \geq 3m$, the predictions by (3) and (5) should be taken as an upper limit of the exclusionary power of the L markers. For maternity exclusion of unrelated females, eqn (3) should be largely valid for any species because L is generally small. However for maternity exclusion of many aunts, eqn (5) might be too optimistic because L can be larger than $3m$ in some species. Species of ants, for example, show huge variation in the number of chromosomes, from only one chromosome, as in the males of the Australian bulldog ant *Myrmecia croslandi*, to as many as 60 chromosomes, as in the males of the giant Neotropical ant *Dinoponera lucida* (Cardoso and Cristiano, 2021). With 60 chromosomes, (5) should be accurate except when L is extremely high, say $L > 180$. With 1 chromosome, on the contrary, (5) may always overestimate the power of a set of L markers because they are likely to be closely linked.

It should be emphasized that (3) or (5) give the *average* probability of excluding a randomly drawn sample of candidates (who are unrelated or fullsibs to the mother) as the mother of a male drawn at random from a population. The actual exclusion probability varies depending on the genotypes of the males and the genotypes of the candidate females, as shown for diploid species (Chakraborty et al., 1988). For a male with genotypes coming from mothers homozygous at an exceptionally high proportion of loci, it is difficult to exclude false maternity because the loci at which the maternal genotypes are heterozygous could be too few to allow maternity exclusion. For this reason and others (such as the presence of population genetic structure, e.g. subdivision), the average exclusion probability calculated by (3) or (5) could be too liberal and a few more markers than those determined by the equations are required to yield accurate parentage analysis results.

Maternity exclusion is different from maternity assignment. The probability of exclusion depends on the genetic structure of a population, calculable from the allele frequencies of the markers in the population. It can be determined before genotype data

are acquired, and therefore is valuable in experimental design in optimizing the sampling intensities of markers and individuals. The probability of maternity assignment depends on the genotypes of candidate females, males and their relatedness as well as allele frequencies in the population. Two approaches can be used to make maternity assignments. One is based on exclusion. When all candidate females except for one are excluded from the maternity of a male, then the maternity can be assigned to the unexcluded female (Jones et al., 2010). The confidence of the assignment is determined by the quantity and quality of marker data. However, implementing exclusion-based parentage assignment can be tricky due to complexities such as the presence of genotyping errors and mutations, and the approach is rarely powerful enough to assign parentage unambiguously in reality. Quite often more than one candidate may remain unexcluded from the parentage of an offspring based on their genotype data.

A more powerful and flexible approach to parentage assignment is based on likelihood. It can optimally use allele frequency and genotyping error rate information, in addition to genotype data as used by the exclusion approach, in calculating the probability of each candidate being the parent of an offspring (Marshall et al., 1998; Wang and Santure, 2009). For example, the sharing of rare alleles between a candidate female and an offspring is strong evidence that they are a mother-offspring dyad in the likelihood approach, but this allele frequency information is wasted in the exclusion approach. Frequently parentage assignment can be determined with confidence by the likelihood approach in situations where parentage assignment is inconclusive by the exclusion approach. Although exclusion probability described above and in the literature is based on exclusion or genotype mismatches, it is informative for likelihood parentage analysis in helping determine the sufficiency of marker information for, and the power of, a parentage analysis.

As the formula calculating the probability of excluding an arbitrary number of aunts from the maternity of a haploid male, (5), is complicated, it is implemented in software AuntExclusion available for free download from https://www.zsl.org/science/software/auntexclusion. It has a Windows GUI for data and parameter input and for analysis results visualization. The software also includes a simulation module which can be used to simulate the probability of excluding multiple aunts from maternity of a male in haplodiploid species and in diploid species.

Data availability

No raw data are generated in this study. The software from this study is posted on https://www.zsl.org/science/software/auntexclusion.

Acknowledgments

We thank the editor N. Rosenberg and two anonymous referees who provided us excellent reviews which have helped in improving this article significantly. We also thank Zarif Ahsan for providing us the proof that $2a_{l2} - a_{l3}$ is minimized at an equal allele frequency for a locus with an arbitrary number of k (≥2) alleles shown in Appendix 1.

References

Alaux, C., Savarit, F., Jaisson, P., Hefetz, A., 2004. Does the queen win it all? Queen-worker conflict over male production in the bumblebee, *Bombus terrestris*. Naturwissenschaften 91, 400–403.

Ayres, K.L., 2002. Paternal exclusion in the presence of substructure. *Forensic Sci. Int.* 129, 142-144.

Avise, J.C., Jones, A.G., Walker, D., DeWoody, J.A., 2002. Genetic mating systems and reproductive natural histories of fishes: lessons for ecology and evolution. Annu. Rev. Genet. 36, 19–45.

Bourke, A.F.G., 1988. Worker reproduction in the higher eusocial Hymenoptera. Q. Rev. Biol. 63, 291–311.

Bretman, A., Tregenza, T., 2005. Measuring polyandry in wild populations: a case study using promiscuous crickets. Mol. Ecol. 14, 2169–2179.

Cardoso, D.C., Cristiano, M.P., 2021. Karyotype diversity, mode, and tempo of the chromosomal evolution of Attina (Formicidae: Myrmicinae: Attini): is there an upper limit to chromosome number?. *Insects*, *12*(12), 1084.

Chakraborty, R., Meagher, T.R., Smouse, P.E., 1988. Parentage analysis with genetic markers in natural populations. I. The expected proportion of offspring with unambiguous paternity. Genetics 118, 527-536.

Coltman, D.W., Bancroft, D.R., Robertson, A., Smith, J.A., Clutton-Brock, T.H., Pemberton, J.M., 1999. Male reproductive success in a promiscuous mammal: behavioural estimates compared with genetic paternity. Mol. Ecol. 8, 1199–1209.

Dodds, K.G., Tate, M.L., McEwan, J.C., Crawford, A.M., 1996. Exclusion probabilities for pedigree testing farm animals. Theor. Appl. Genet. 92, 966-975.

Double, M.C., Cockburn, A., Barry, S.C., Smouse, P.E., 1997. Exclusion probabilities for single-locus paternity analysis when related males compete for matings. Mol. Ecol. 6,1155-1166.

Flanagan, S.P., Jones, A.G., 2019. The future of parentage analysis: From microsatellites to SNPs and beyond. Mol. Ecol. 28, 544-567.

Foster, K.R., Ratnieks, F.L.W., Gyllenstrand, N., Thorén, P.A., 2001. Colony kin structure and male production in *Dolichovespula* wasps. Mol. Ecol. 10, 1003–1010.

Friend, L.A., Bourke, A.F.G., 2014. Workers respond to unequal likelihood of future reproductive opportunities in an ant. Anim. Behav. 97, 165-176.

Fung, W.K., Chung, Y.K., Wong, D.M., 2002. Power of exclusion revisited: probability of excluding relatives of the true father from paternity. Int. J. Legal Med. 116, 64–67.

Garant, D., Dodson, J.J., Bernatchez, L., 2001. A genetic evaluation of mating system and determinants of individual reproductive success in Atlantic salmon (*Salmo salar L.*). J. Hered. 92, 137–145.

Giehr, J., Senninger, L., Ruhland, K., Heinze, J., 2020. Ant workers produce males in queenless parts of multi-nest colonies. Sci. Rep. 10, 1-8.

Hammond, R.L., Bruford, M.W., Bourke, A.F.G., 2003. Male parentage does not vary with colony kin structure in a multiple-queen ant. J. Evol. Biol. 16, 446-455.

Hammond, R.L., Bruford, M.W., Bourke, A.F.G., 2006. A test of reproductive skew models in a field population of a multiple-queen ant. Behav. Ecol. Sociobiol. 61, 265-275.

Heaton, M.P., Leymaster, K.A., Kalbfleisch, T.S., Kijas, J.W., Clarke, S.M., McEwan, J., Maddox, J.F., Basnayake, V., Petrik, D.T., Simpson, B., Smith, T.P., 2014. SNPs for parentage testing and traceability in globally diverse breeds of sheep. PloS one *9*(4), p.e94851.

Helyar, S.J., Hemmer-Hansen, J., Bekkevold, D., Taylor, M.I., Ogden, R., Limborg, M.T., Cariani, A., Maes, G.E., Diopere, E., Carvalho, G.R., Nielsen, E.E., 2011. Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. Mol. Ecol. Res. 11 (Suppl. 1), 123–136.

Hu, Y.Q., Fung, W.K., Hu, Y.Q., 2005. Power of excluding an elder brother of a child from paternity. Forensic Sci. Int. 152, 321–322.

Hughes, C.R., 1998. Integrating molecular techniques with field methods in studies of social behavior: a revolution results. Ecology 79, 383–399.

Jamieson, A., 1965. The genetics of transferrin in cattle. Heredity 20, 419–441.

Jamieson, A., Taylor, S.C.S., 1997. Comparisons of three probability formulae for parentage exclusion. Anim. Genet. 28, 397–400.

Jones, A.G., Small, C.M., Paczolt, K.A., Ratterman, N.L., 2010. A practical guide to methods of parentage analysis. Mol. Ecol. Res. 10, 6-30.

MacCluer, J.W., Schull, W.J., 1963. On the estimation of the frequency of nonpaternity. Am. J. Hum. Genet. 15, 191-202.

Marshall, T.C., Slate, J., Kruuk, L.E.B., Pemberton, J.M., 1998. Statistical confidence for likelihood-based paternity inference in natural populations. Mol. Ecol. 7, 639–655.

559 Moran, B.M., Thomas, S.M., Judson, J.M., Navarro, A., Davis, H., Sidak-Loftis, L., Korody, M.,
560 Mace, M., Ralls, K., Callicrate, T., and Ryder, O.A., 2021. Correcting parentage
561 relationships in the endangered California Condor: Improving mean kinship
562 estimates for conservation management. The Condor 123, p.duab017.

563 Ogden, R., Dawnay, N., McEwing, R., 2009. Wildlife DNA forensics—bridging the gap
564 between conservation genetics and law enforcement. Endanger. Species Res. 9,179-
565 195.

566 Ohno, Y., Sebetan, I.M., Akaishi, S., 1982. a simple method for calculating the probability
567 of excluding paternity with any number of codominant alleles. Forensic Sci. Int. 19,
568 93 – 98.

569 Pompanon, F., Bonin, A., Bellemain, E., Taberlet, P., 2005. Genotyping errors: causes,
570 consequences and solutions. Nat. Rev. Genet. 6, 847-859.

571 Ratnieks, F.L.W., Foster, K.R., Wenseleers, T., 2006. Conflict resolution in insect societies.
572 Annu. Rev. Entomol. 51, 581-608.

573 Robledo-Arnuncio, J.J., Gil, L., 2005. Patterns of pollen dispersal in a small population of
574 *Pinus sylvestris* L. revealed by total exclusion paternity analysis. Heredity 94, 13–22.

575 Ross, K.G. 2001., Molecular ecology of social behaviour: analyses of breeding systems
576 and genetic structure. Mol. Ecol. 10, 265-284.

577 Rubenstein DR, Abbot P., 2017. Comparative Social Evolution. Cambridge University
578 Press, Cambridge.

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

581 Southon, R.J., Bell, E.F., Graystock, P., Wyatt, C.D.R., Radford, A.N., Sumner, S., 2019. High
582 indirect fitness benefits for helpers across the nesting cycle in the tropical paper
583 wasp *Polistes canadensis*. Mol. Ecol. 28, 3271-3284.

584 Theodorou, P., Radzevičiūtė, R., Kahnt, B., Soro, A., Grosse, I., Paxton, R.J., 2018. Genome-
585 wide single nucleotide polymorphism scan suggests adaptation to urbanization in
586 an important pollinator, the red-tailed bumblebee (*Bombus lapidarius* L.). Proc. R.
587 Soc. B 285, 20172806.

588 Thompson, E.A., Meagher, T.R., 1987. Parental and sib likelihoods in genealogy
589 reconstruction. Biometrics 43, 585–600.

590 Trunzer, B., Heinze, J., Hölldobler, B., 1999. Social status and reproductive success in
591 queenless ant colonies. Behaviour 136, 1093-1105.

592 Wang, J., Santure, A., 2009. Parentage and sibship inference from multilocus genotype
593 data under polygamy. Genetics 181,1579-1594.

Wiener, A.S., Lederer, M., Polayes, S.H., 1930. Studies in isohemagglutination. IV. On the chances of proving non-paternity; with special reference to the blood groups. J. Immunol. 19, 259-282.

598 **Table 1: Maternity exclusion of aunts of a haploid male from different**
599 **grandparent mating types**. (Freq. = frequency)

Grandparent Mating		Sibling Worker from Mating		Male Produced by Worker		Excluded Genotype	Exclusion Probability
Type	Freq.	Type	Freq.	Type	Freq.		
$A_iA_i \times A_i$	p_{li}^3	A_iA_i	1	A_i	1	-	0
$A_iA_i \times A_j$ $(j \neq i)$	$p_{li}^2(1-p_{li})$	A_iA_j	1	A_i	½	-	0
				A_j	½	-	0
$A_iA_j \times A_i$ $(j \neq i)$	$2p_{li}^2(1-p_{li})$	A_iA_i	½	A_i	1	-	0
		A_iA_j	½	A_i	½	-	0
				A_j	½	A_iA_i	½
$A_iA_j \times A_m$ $(j \neq i, m \neq i, m \neq j)$	$p_{li}p_{lj}(1-p_{li}-p_{lj})$	A_iA_m	½	A_i	½	A_jA_m	½
				A_m	½	-	0
		A_jA_m	½	A_j	½	A_iA_m	½
				A_m	½	-	0

600

601

602 **Table 2: Check of equation (5) by simulations with and without linkage**

Allele Frequency	***N***	***k***	***L***	**Simulated exc. prob. for a genome in size (Morgan)**							**Eqn (5)**
				1	**2**	**4**	**8**	**16**	**32**	**∞**	
Equal	10	5	10	0.1915	0.2712	0.3297	0.3401	0.3408	0.3411	0.3409	0.3408
	10	5	20	0.2770	0.4398	0.6277	0.7650	0.7999	0.7965	0.7962	0.7964
	100	5	20	0.1028	0.1698	0.2619	0.3424	0.3526	0.3546	0.3554	0.3547
	100	5	40	0.1774	0.3034	0.5124	0.7630	0.9118	0.9398	0.9361	0.9362
	10	2	10	0.0545	0.0580	0.0550	0.0533	0.0536	0.0535	0.0536	0.0536
	10	2	20	0.1343	0.1798	0.2134	0.2220	0.2226	0.2228	0.2228	0.2228
	100	2	40	0.0697	0.1045	0.1450	0.1766	0.1796	0.1808	0.1808	0.1809
	100	2	80	0.1384	0.2321	0.3816	0.5711	0.7199	0.7711	0.7689	0.7688
Triangular	10	5	10	0.1729	0.2380	0.2792	0.2841	0.2847	0.2847	0.2847	0.2848
	10	5	20	0.2641	0.4144	0.5856	0.7071	0.7346	0.7317	0.7317	0.7317
	100	5	20	0.0908	0.1459	0.2154	0.2650	0.2649	0.2672	0.2675	0.2675
	100	5	40	0.1651	0.2823	0.4760	0.7127	0.8633	0.8897	0.8860	0.8860
	10	2	10	0.0451	0.0461	0.0427	0.0416	0.0418	0.0419	0.0418	0.0418
	10	2	20	0.1185	0.1535	0.1757	0.1785	0.1792	0.1791	0.1791	0.1792
	100	2	40	0.0595	0.0860	0.1122	0.1268	0.1224	0.1243	0.1244	0.1243
	100	2	80	0.1263	0.2099	0.3407	0.5050	0.6314	0.6670	0.6651	0.6652

Note each of the L loci is assumed to have k alleles with either equal frequencies (=1/k) or frequencies in a triangular distribution (i.e. frequency of allele j is proportional to j for j = 1, 2, …, k). The average probability of excluding N aunts from maternity of a male taken at random from the population is calculated by equation (5) and simulations. For simulations, the genome size is assumed to be 1, 2, 4, … 32 Morgans in genetic map length or to be ∞ for free recombination. The average in simulation is taken over 10000000 replicates.

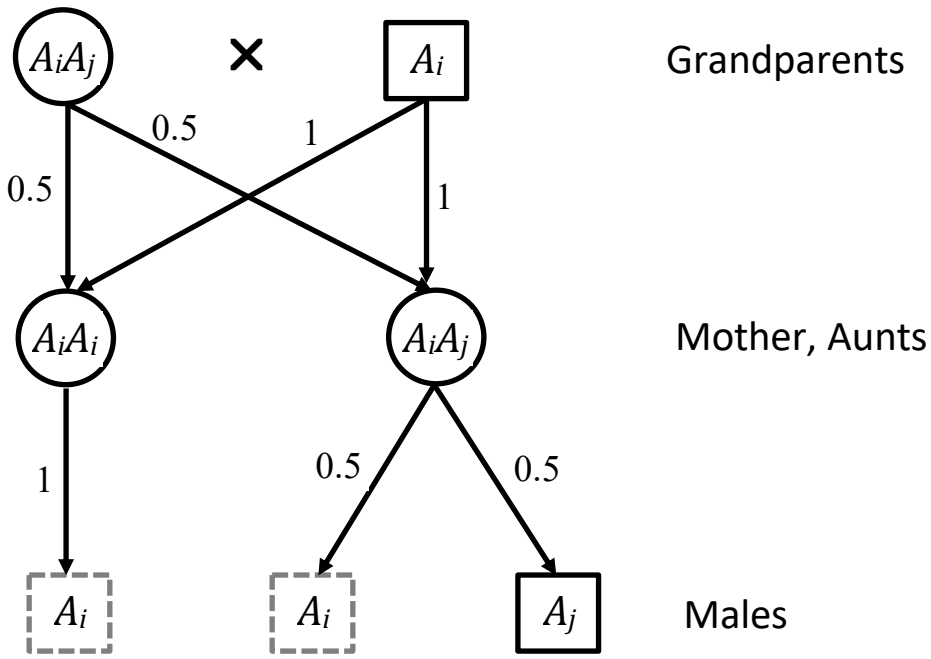

Figure 1: A schematic illustration of a pedigree in which a male has possible genotypes that may or may not allow the maternity exclusion of its aunts. A male who has a genotype (A_j) that allows the exclusion of some aunts (with genotype A_iA_i) as its mother is depicted in a black solid-lined box, while a male who has a genotype (A_i) that does not allow the exclusion of any aunts as its mother is depicted in a grey dashed-lined box. The figures beside the arrowed lines are the corresponding transmission probabilities from a parental to an offspring genotype.

603

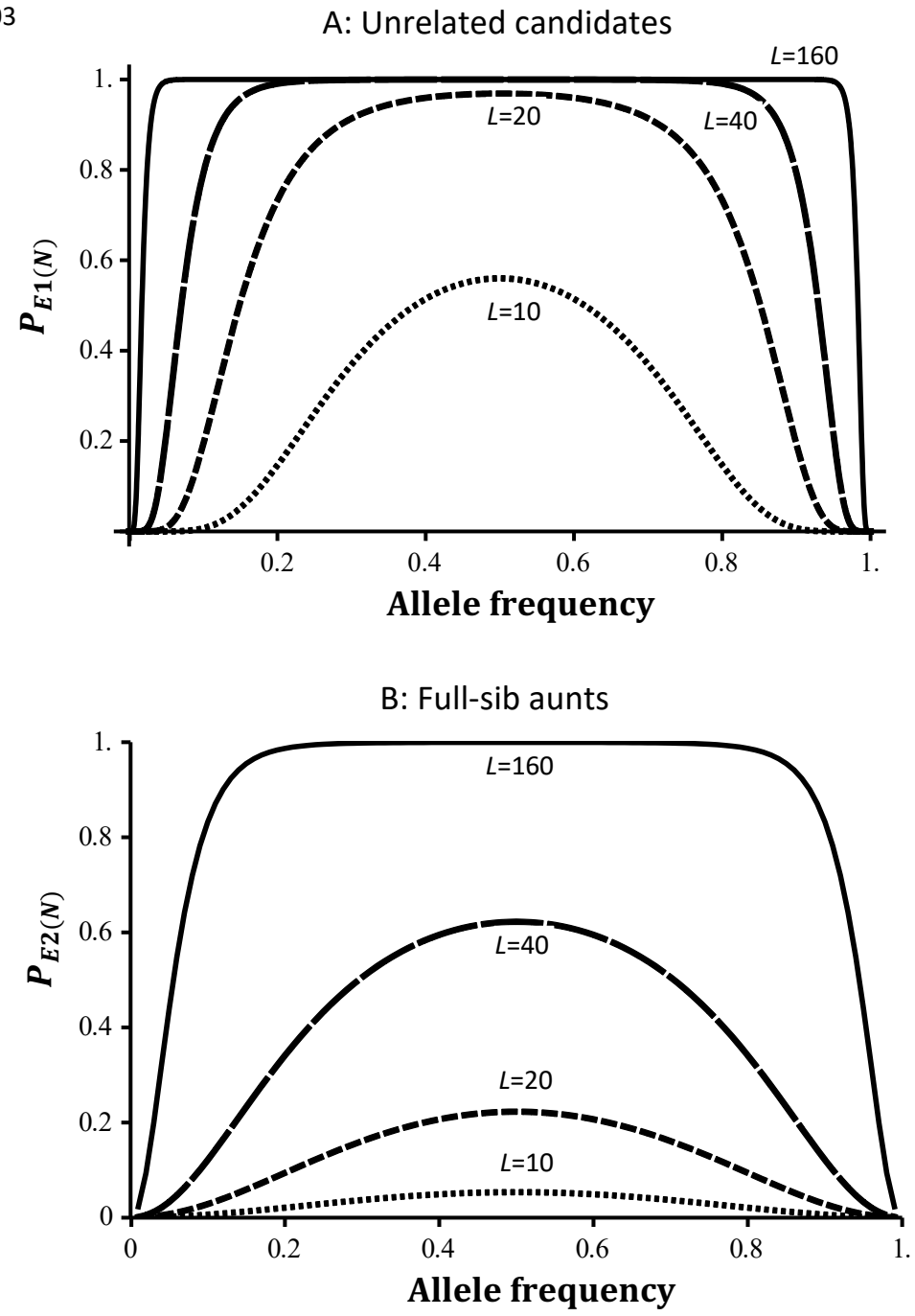

Figure 2: Average maternity exclusion probability of males as a function of allele frequencies at diallelic loci. (A) N=10 candidate females unrelated to the true mother of a male, and (B) N=10 candidates who are full siblings to the true mother of a male, are to be excluded as the mother of the male using L = 10, 20, 40, 160 loci with each locus having K=2 codominant alleles of frequencies shown on the x axis. Eqns (3) and (6) are used in calculating the average maternity exclusion probabilities in cases (A) and (B) respectively.

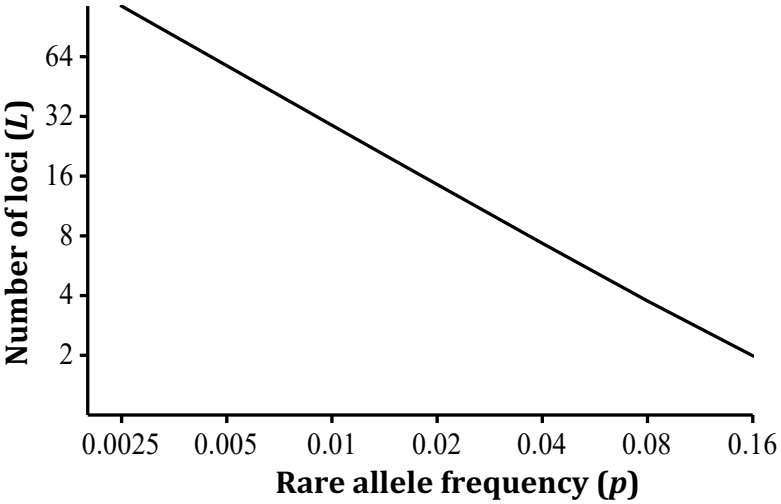

Figure 3: Number of loci L (y axis) with rare allele frequency p (x axis) required to attain the same exclusion probability as a single diallelic locus with equifrequent (0.5, 0.5) alleles. Note that the axes have logarithmic scales. Eqn (3) is used in the calculations.

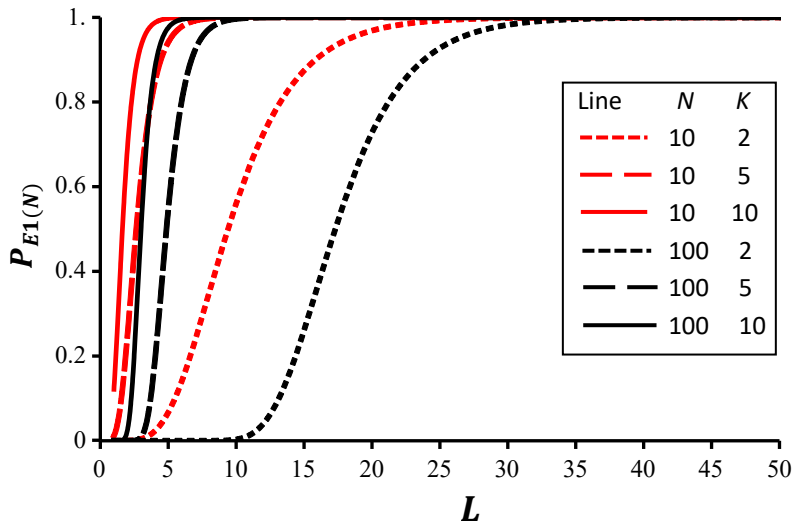

Figure 4: The probability of excluding *N* (10 or 100) candidate females unrelated to a male as the maternity of the male. The exclusion probability is calculated by eqn (4b) using *L* loci (on *x* axis), each having *K*=2, 5 or 10 equifrequent codominant alleles.

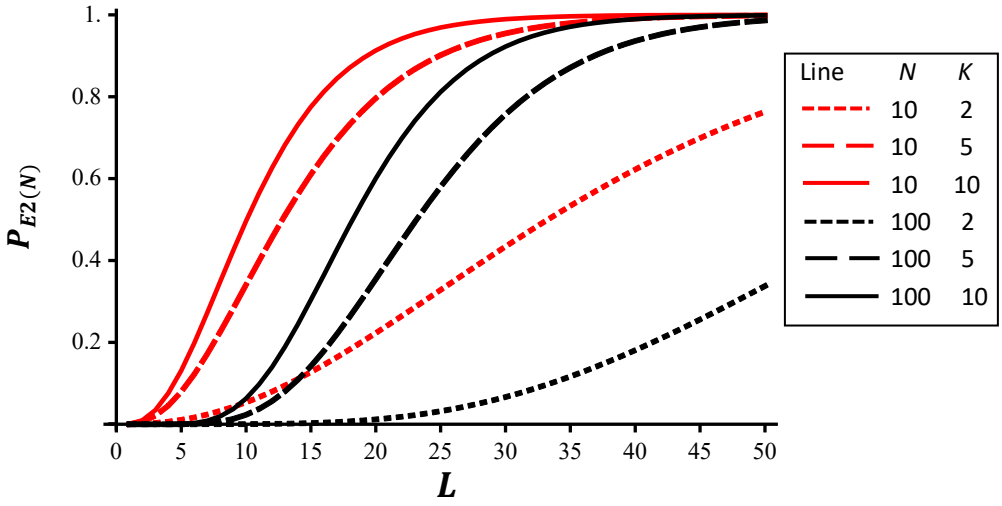

Figure 5: The probability of excluding N (10 or 100) candidate females who are full siblings to the true mother of a male as maternity of the male. The exclusion probability is calculated by eqn (5) using L loci (on x axis), each having K=2, 5 or 10 equifrequent codominant alleles.

604

Appendix 1: Proof that $2a_{l2} - a_{l3}$ is minimized at an equal allele frequency

We provide a proof that $2a_{l2} - a_{l3}$ in equation (3) is minimized to $\frac{2k_l-1}{k_l^2}$ and therefore the exclusion probability is maximized to $\left(1 - \prod_{l=1}^{L} \frac{2k_l-1}{k_l^2}\right)^N$ of equation (4a) when locus *l* has an equal allele frequency of $1/k_l$. For clarity, we drop the subscript *l* and consider a locus with $k \geq 2$ alleles of frequencies p_i for *i*=1, 2, …, *k*, where p_i is apparently subject to the constraints $0 < p_i < 1$ and $\sum_{i=1}^{k} p_i \equiv 1$.

When p_k is replaced by $1 - q_{k-1}$, where $q_{k-1} = \sum_{i=1}^{k-1} p_i$, the quantity $2a_2 - a_3$, denoted by X_k, is reduced to

$$X_k = 2a_2 - a_3 = 2\sum_{i=1}^{k-1} p_i^2 + 2(1 - q_{k-1})^2 - \sum_{i=1}^{k-1} p_i^3 - (1 - q_{k-1})^3.$$

To derive the minimum value of X_k and the corresponding values of p_i, we first obtain the critical points of X_k by setting its first derivatives to zero and solving the resultant equations. We then examine these points and choose the points that satisfy the constraints $0 < p_i < 1$ and $\sum_{i=1}^{k} p_i \equiv 1$. The chosen valid critical points are then used in the second derivative test to determine whether function X_k attains a minimum, a maximum or otherwise at the critical points. We first consider the simplest cases of a diallelic (*k*=2) and triallelic (*k*=3) locus, and then the general case of any number of alleles ($k \geq 2$) at a locus.

1. Two alleles, $k = 2$

In the simplest case of a diallelic locus with *k*=2 alleles, function X_k reduces to

$$X_k = 2p_1^2 + 2(1 - p_1)^2 - p_1^3 - (1 - p_1)^3.$$

By setting the first derivative to zero,

$$\frac{\partial X_k}{\partial p_1} = 2p_1 - 1 = 0,$$

we obtain the sole critical point p_1=1/2. Apparently, the point p_1=1/2 (and thus $p_2 = 1 - p_1$=1/2) is valid, satisfying the constraints $0 < p_i < 1$ and $\sum_{i=1}^{k} p_i \equiv 1$. The second derivative of X_k is 2, which is a positive value and signifies that X_k attains a minimum value at the critical point {p_1=1/2, $p_2 = 1/2$}. The minimal value of X_k at point $p_i = 1/2$ (*i*=1, 2) is $\frac{2k-1}{k^2}$ =¾.

2. Three alleles, *k*=3

For a triallelic locus with *k*=3 alleles, function X_k reduces to

$$X_k = 2\sum_{i=1}^{2} p_i^2 + 2(1 - q_2)^2 - \sum_{i=1}^{2} p_i^3 - (1 - q_2)^3,$$

where $q_2 = \sum_{i=1}^{2} p_i$. By setting the first derivatives to zero,

$$\frac{\partial X_k}{\partial p_1} = (p_1 - 1 + q_2)(1 - 3p_1 + 3q_2) = 0,$$

$$\frac{\partial X_k}{\partial p_2} = (p_2 - 1 + q_2)(1 - 3p_2 + 3q_2) = 0,$$

we obtain a set of two equations. Solving the equations, we obtain 4 critical points of $\{p_1, p_2\}$, which are {1/3, 1/3}, {-1/3, 5/3}, {5/3, -1/3}, {-1/3, -1/3}. Except for the first point, all other points are invalid because they contain negative values which are infeasible for allele frequencies. The first point is the sole valid one that satisfies the constraints $0 < p_i < 1$ and $\sum_{i=1}^{k} p_i \equiv 1$. From the constraints on allele frequencies, we obtain $p_3 = 1/3$ at the critical point.

The Hessian matrix is

$$H(p_1, p_2) = \begin{bmatrix} 2 + 6q_2 - 6p_1 & -2 + 6q_2 \\ -2 + 6q_2 & 2 + 6q_2 - 6p_2 \end{bmatrix},$$

which becomes

$$H(p_1, p_2) = \begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix}$$

at the critical point $\{p_1, p_2\}$={1/3, 1/3}. The eigenvalues of $H(p_1, p_2)$ at the critical point are 6 and 2, which are both positive, signifying that function X_k attains a minimum value at the critical point $\{p_1, p_2\}$={1/3, 1/3}. The minimum value of X_k at point $p_i = 1/3$ (i=1, 2, 3) is $\frac{2k-1}{k^2} = 5/9$.

3. Any number of alleles, $k \geq 2$

For a locus with an arbitrary number of $k \geq 2$ alleles, the function X_k is

$$X_k = 2\sum_{i=1}^{k-1} p_i^2 + 2(1 - q_{k-1})^2 - \sum_{i=1}^{k-1} p_i^3 - (1 - q_{k-1})^3,$$

where $q_{k-1} = \sum_{i=1}^{k-1} p_i$. The partial derivatives of X_k with respect to p_j are

$$\frac{\partial X_k}{\partial p_j} = (p_j - 1 + q_{k-1})(1 - 3p_j + 3q_{k-1}),$$

for j=1, 2, …, k-1. To obtain the critical points of function X_k, we set these partial derivatives to zero,

$$\frac{\partial X_k}{\partial p_j} = (p_j - 1 + q_{k-1})(1 - 3p_j + 3q_{k-1}) = 0.$$

Note that, in the above equation, the factor $1 - 3p_j + 3q_{k-1} = 1 + 3\sum_{i=1}^{j-1} p_i + 3\sum_{i=j+1}^{k-1} p_i > 1$ as $p_i > 0$ for i=1, 2, …, k-1. Therefore, we have $p_j - 1 + q_{k-1} = 0$ and thus $p_j = 1 - q_{k-1}$ for j=1, 2, …, k-1. Hence, function X_k has only one valid critical point within the permissible parameter space defined by the constraints $0 < p_i < 1$ and $\sum_{i=1}^{k} p_i \equiv 1$, which is $p_1 = p_2 = \cdots = p_k = 1/k$.

We now show that function X_k reaches a minimum at this critical point of $p_i = 1/k$ for i=1, 2, …, k. The Hessian matrix for function X_k is

$$H(p_1, p_2, \ldots, p_{k-1}) = \begin{bmatrix} 2 + 6q_{k-1} - 6p_1 & -2 + 6q_{k-1} & \ldots & -2 + 6q_{k-1} \\ -2 + 6q_{k-1} & 2 + 6q_{k-1} - 6p_2 & \ldots & -2 + 6q_{k-1} \\ \ldots & \ldots & \ldots & \ldots \\ -2 + 6q_{k-1} & -2 + 6q_{k-1} & \ldots & 2 + 6q_{k-1} - 6p_{k-1} \end{bmatrix},$$

At the critical point $p_i = 1/k$, the matrix becomes

$$H(p_1, p_2, \ldots, p_{k-1}) = \begin{pmatrix} 2n & n & \ldots & n \\ n & 2n & \ldots & n \\ \ldots & \ldots & \ddots & \ldots \\ n & n & \ldots & 2n \end{pmatrix},$$

where $n = (4k - 6)/k$. This matrix is symmetrical, with an identical diagonal element $2n = \frac{8k-12}{k}$ and an identical nondiagonal element $n = \frac{4k-6}{k}$. We show, using Sylvester's criterion (Roger et al., 1990, p.439), that this matrix is positive definite, as each of the leading $m \times m$ minors are positive, i.e. the upper left $m \times m$ corner has positive determinant for each $m = 1, \ldots, k - 1$. We show this by induction.

Let a_m be the determinant of the upper left $m \times m$ corner. For the base case of $m = 1$, we have, for $k \geqslant 2$, that

$$a_1 = \det(2n) = 2n = \frac{8k-12}{k} > 0.$$

For the inductive stage, suppose $a_m > 0$. The upper $(m + 1) \times (m + 1)$ corner is given by

$$\begin{pmatrix} 2n & n & \ldots & n \\ n & 2n & \ldots & n \\ n & n & \ldots & n \\ \ldots & \ldots & \ddots & \ldots \\ n & n & \ldots & 2n \end{pmatrix}.$$

Note that the determinant of a matrix is unchanged if we subtract a column by another (Roger et al., 1990, p.10). Subtracting the second column from the first, we have

$$a_{m+1} = \det \begin{pmatrix} 2n & n & \ldots & n \\ n & 2n & \ldots & n \\ n & n & \ldots & n \\ \ldots & \ldots & \ddots & \ldots \\ n & n & \ldots & 2n \end{pmatrix} = \det \begin{pmatrix} n & n & \ldots & n \\ -n & 2n & \ldots & n \\ 0 & n & \ldots & n \\ \ldots & \ldots & \ddots & \ldots \\ 0 & n & \ldots & 2n \end{pmatrix}.$$

Via cofactor expansion (Roger et al., 1990, p.8) along the first column, we find the determinant of this matrix as

$$a_{m+1} = n \times \det \begin{pmatrix} 2n & n & \ldots & n \\ n & 2n & \ldots & n \\ \ldots & \ldots & \ddots & \ldots \\ n & n & \ldots & 2n \end{pmatrix} - (-n) \times \det \begin{pmatrix} n & n & \ldots & n \\ n & 2n & \ldots & n \\ \ldots & \ldots & \ddots & \ldots \\ n & n & \ldots & 2n \end{pmatrix}.$$

The left matrix is the same as the upper left $m \times m$ corner, so

$$a_{m+1} = na_m + n \times \det \begin{pmatrix} n & n & \ldots & n \\ n & 2n & \ldots & n \\ \ldots & \ldots & \ddots & \ldots \\ n & n & \ldots & 2n \end{pmatrix}.$$

By subtracting the top row from each of the other rows in the remaining matrix (which also preserves the determinant), we obtain

$$a_{m+1} = na_m + n \times \det\begin{pmatrix} n & n & \dots & n \\ 0 & n & \dots & 0 \\ \dots & \dots & \ddots & \dots \\ 0 & 0 & \dots & n \end{pmatrix}.$$

Finally, using cofactor expansion along the first column, this becomes

$$a_{m+1} = na_m + n^2 \times \det\begin{pmatrix} n & 0 & \dots & 0 \\ 0 & n & \dots & 0 \\ \dots & \dots & \ddots & \dots \\ 0 & 0 & \dots & n \end{pmatrix} = na_m + n^{m+1}.$$

We can see $a_{m+1} > 0$ since, for $k \geq 2$, $n^{m+1} = ((4k-6)/k)^{m+1} > 0$ and $na_m = (4k-6)/k)a_m > 0$ by our inductive hypothesis. This completes our induction, so $a_m > 0$ and each upper left $m \times m$ corner of $H(\frac{1}{k},\dots,\frac{1}{k})$ has a positive determinant for $m = 1, \dots, k-1$.

Therefore, by Sylvester's criterion, $H(\frac{1}{k},\dots,\frac{1}{k})$ is positive definite, so function X_k has an absolute minimum when $p_1 = \cdots = p_k = 1/k$ where $k \geqslant 2$.

References

Roger A. Horn and Charles R. Johnson. Matrix Analysis. Cambridge University Press, 1990.