Parentage identification using single nucleotide polymorphism genotypes: Application to product tracing

W. G. Hill,*2 B. A. Salisbury,† and A. J. Webb‡

*Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, West Mains Road, Edinburgh, EH9 3JT, United Kingdom; †Clinical Data Inc., Five Science Park, New Haven, CT 06511; and ‡Maple Leaf Foods Inc., 30 St. Clair Avenue West, Toronto, Ontario M4V 3A2, Canada

ABSTRACT: Identification of relatives using SNP markers has many possible applications. One is as a route to tracing a food product such as a cut of meat back to its source of origin by identifying the parents of the animal from which the product was derived. We develop methods for using SNP markers with maximum likelihood, allowing for the possibility of genotyping errors that would cause false exclusions by simpler methods. We use expectations of likelihood ratios to consider how gene frequencies in the parental populations, numbers of loci, and error rates affect accuracy. This is further quantified as the risk, the probability that an incorrect sire is identified from a panel that contains many other putative sires including its relatives, using a breeding structure relevant to pig breeding. This appears to be a straightforward and potentially effective means of product tracing.

Key words: genetic marker, genotyping error, parentage identification, pig, product tracing, single nucleotide polymorphism

©2008 American Society of Animal Science. All rights reserved.

INTRODUCTION

Marker genotypes are used to identify relatives in many applications (Weir et al., 2006). Examples include identification and paternity analysis in livestock (Heaton et al., 2002), identifying family members in nonpedigreed natural populations (Marshall et al., 1998), and in forensic studies to identify disaster victims or find suspects via the genotype of a relative on the database (Brenner and Weir, 2003; Bieber et al., 2006). In tracing food products, genotyping can be used to identify breed of origin and, further, to track a piece of meat with specific desirable or undesirable properties back to the parents of the animals and thus farm of origin (Hayes et al. 2005; Plastow et al., 2007). Tracing could also be used in a breeding program as a way to select on desirable traits, for example meat quality, identified at final product level.

Thus, we analyze methods to identify the source, typically the sire or dam, or both, and through that er how gene frequencies in the parental populations, numbers of loci, and error rates affect accuracy. This is further quantified as the risk, the probability that an incorrect sire is identified from a panel that contains many other putative sires including its relatives, using a breeding structure relevant to pig breeding. This appears to be a straightforward and potentially effective means of product tracing.

MATERIALS AND METHODS

Animal Care and Use Committee approval was not obtained for this study because no animals were used.

Model

Because the populations from which male and female parents are drawn may well be of different breeds
and themselves be crossbred, we allow different gene frequencies between the parent populations and departures from Hardy-Weinberg equilibrium frequencies within them. Consider a locus with alleles A and a, where allele A has frequency \( p_m \) and \( p_f \) in the male and female parental (i.e., sire and dam) populations and \( a \) has frequencies \( q_m = 1 - p_m \) and \( q_f \), respectively. The genotype frequencies in the male population, for example, for AA, Aa, and aa are \( g_{m1}, g_{m2}, \) and \( g_{m3} \), respectively, with \( g_{m1} = p_m^2 \) if they are in Hardy-Weinberg equilibrium.

The expected genotype frequencies in the offspring generation, assuming random mating between the populations, for AA, Aa, and aa, respectively, are described by the following vector:

\[
g_o = (p_m p_f \quad p_m q_f + q_m p_f \quad q_m q_f)\]

For sires (dams) of specified genotype, mating to random dams (sires), and for sire \( \times \) dam genotype pairs, the expected frequencies of offspring genotypes are given in Table 1.

**Genotyping Errors.** Single nucleotide polymorphism genotyping is not devoid of error, and miscalls can arise with automatic or manual systems, so the possibility of genotyping error and indeed of mutation has to be included in the calculations (Sobel et al., 2002). Otherwise, a putative parent with homozygous genotype AA, for example, could be excluded if the offspring sample was incorrectly called as aa (Table 1), regardless of genotypes at other loci.

The distribution of miscalls can be represented by a table or 3 \( \times \) 3 matrix \( E \) in which elements \( e_{ij} \) denote the probability that an individual of real genotype \( i \) is assigned genotype \( j \). The matrices are general, but we assume as a reference point and in most examples, unless specified otherwise, that there is the same probability of miscalling each allele as the other one:

\[
E = \begin{pmatrix}
1 - \delta & \delta^2 & \delta^2 \\
\frac{1}{2} e & 1 - e & \frac{1}{2} e \\
\delta^2 & e & 1 - e - \delta^2
\end{pmatrix}
\]

where \( \delta = \) the allelic error rate and \( e = 28(1 - \delta) \sim 28 \). For simplicity, we shall refer to \( e \) as the genotyping error rate. In sire identification, errors in miscalling heterozygotes as homozygotes cause the most false exclusions, assuming that the probability of calling the wrong homozygote (\( \delta^2 \), e.g., AA as aa) is very small; therefore, we shall also consider a model where the only errors are in miscalling of heterozygotes [i.e., the first and last rows of \( E \) are \((1, 0, 0) \) and \((0, 0, 1) \), respectively]. For generality, and because different procedures may be used for genotyping parents and offspring (for example, the latter may be genotyped in a one-off process using a different assay when a production problem is detected), this matrix may differ among male and female parents and offspring among groups. Hence, we define matrices \( E_m, E_f, \) and \( E_o \), respectively.

**Mutations.** Germ-line mutations have a similar influence to genotyping errors on the distribution of genotypes of parent-offspring pairs. Thus, if the mutation rate of \( A \) to \( a \) is \( u_A \) and of \( a \) to \( A \) is \( u_a \), and mutations occur sufficiently rarely that double mutations can be ignored, the total error rate \( E_{o^*} \) recorded in offspring is:

\[
E_{o^*} = E_o + \begin{pmatrix}
-2u_A & 2u_A & 0 \\
u_a & -u_A - u_a & u_A \\
0 & 2u_a & -2u_a
\end{pmatrix}.
\]

The matrices \( E_m \) and \( E_f \) are unchanged.

**Likelihood Calculations.**

We consider as an example the calculation of likelihoods for detection of the sire. That for dams is essentially the same, and the necessary extensions for detection of a sire and dam pair are given subsequently.

**Likelihood when the Putative Sire is Unrelated to the Sire.** The observed frequency (i.e., including possible erroneous genotypes and mutations) of individuals of genotype \( j \) at an arbitrary locus in the offspring population, which is also equal to the expected frequency in the offspring of an unrelated sire, is given by \( g_j = \sum g_o e_{oij} \), an element of the matrix product

<table>
<thead>
<tr>
<th>Sire (frequency)</th>
<th>Random</th>
<th>AA ((g_{m1} = p_{m1}))</th>
<th>Aa ((g_{m2} = 2p_{m1}q_{m1}))</th>
<th>aa ((g_{m3} = q_{m1}^2))</th>
<th>Dam (frequency)</th>
<th>AA</th>
<th>Aa</th>
<th>aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td></td>
<td>AA (p_m p_f)</td>
<td>Aa (p_m q_f + q_m p_f)</td>
<td>aa (q_m q_f)</td>
<td>(1/2p_m)</td>
<td>1/2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AA ((g_{m1} = p_{m1}^2))</td>
<td>(p_f)</td>
<td>0</td>
<td>(1/2)</td>
<td>0</td>
<td>1/2</td>
<td>1/2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Aa ((g_{m2} = 2p_{m1}q_{m1}))</td>
<td>1/2(p_f)</td>
<td>1/2</td>
<td>(1/2)</td>
<td>1</td>
<td>0/2</td>
<td>1/2</td>
<td>1/2</td>
<td>0</td>
</tr>
<tr>
<td>aa ((g_{m3} = q_{m1}^2))</td>
<td>0</td>
<td>(p_f)</td>
<td>(q_f)</td>
<td>0</td>
<td>1/2</td>
<td>1/2</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

\(^1\)The parental genotype frequencies are given in general and under Hardy-Weinberg equilibrium assumptions (e.g., \(g_{m1} = p_{m1}^2\)).
\[ \mathbf{g} = \mathbf{g}_s \mathbf{E}_s. \]

Hence, the probability that the offspring has observed genotype \( j \) and the putative sire has observed genotype \( i \) when it is a random male, unrelated (U) to the sire, does not depend on \( i \) and is

\[ P(i, j | U) = g_j. \]

Combining loci, \( k \), which are assumed to be in linkage equilibrium, and now letting \( g_{kj} \) be the probability that the offspring has genotype \( j \) at locus \( k \), the likelihood of the data given that the putative sire is unrelated to the sire is given by the product

\[ L(U) = \prod \mathbf{h} \mathbf{g}_{kj}. \]

**Likelihood when the Putative Sire is the Sire.** Let the element \( t_{hi} \) of

\[ T = \begin{pmatrix} p_f & q_f & 0 \\ \frac{1}{2} p_f & \frac{1}{2} q_f \\ 0 & p_f & q_f \end{pmatrix} \]

denote the frequency that a sire of true genotype \( h \) has an offspring of true genotype \( j \) (elements of Table 1), and so the \( h, j \) element of the product \( \mathbf{T} \mathbf{E}_s \) denotes the probability that a sire of true genotype \( h \) has an offspring of observed genotype \( j \).

The joint probability that, for example, a sire has true genotype \( h \) and observed genotype \( i \) is given by \( g_{ih} \mathbf{e}_{sh} \). Hence, the conditional probability that a sire has true genotype \( h \) given observed genotype \( i \) is

\[ c_{sh} = g_{ih} \mathbf{e}_{sh} / \sum g_{ih} \mathbf{e}_{sh}, \]

where \( c_{sh} \) denotes an element of a \( 3 \times 3 \) matrix \( \mathbf{C}_s \).

Hence, \( P(i, j | S) \), the probability that the offspring has observed genotype \( j \) and that the putative sire has observed genotype \( i \) when it is the sire, is obtained by summing over all possible combinations of true genotype of sire and true genotype of offspring. It is given by

\[ x_{ij} = \sum \sum c_{sh} \sum d_{hy} \mathbf{e}_{yd}, \]

where \( x_{ij} \) is an element of

\[ \mathbf{X} = \mathbf{C}_s \mathbf{T} \mathbf{E}_s. \]

Combining all loci under the same assumptions of independence as above, the likelihood of the data given it is the sire is

\[ L(S) = \prod \mathbf{k} x_{kij}. \]

**Likelihood Ratio.** The likelihood ratio \( L(S)/L(U) \) is the relative probability of obtaining the observed genotypes when the putative sire is indeed the sire (S) compared with when it is unrelated (U). The logarithm taken to base 10 of the likelihood ratio (\( lod \)) is determined as follows:

\[ lod_s = \log_{10}[L(S)] - \log_{10}[L(U)]. \]

**Identification of Dams or Both Parents.** For identification of dams with unknown sires, the same analysis as outlined above still applies but with \( \mathbf{C}_s, \mathbf{g}_s \), and \( \mathbf{E}_s \) replaced by \( \mathbf{C}_d, \mathbf{g}_d \), and \( \mathbf{E}_d \), respectively.

For identification of parental pairs, there are now 9 possible real and observed combinations of sire and dam genotypes. The transmission probabilities are given by a \( 9 \times 3 \) matrix \( \mathbf{H} \), with elements given in Table 1, ordered as dam genotype nested within sire genotype, where \( \mathbf{H}_i (3 \times 3) \) defines progeny of sire genotype \( i \):

\[ \mathbf{H} = \begin{pmatrix} \mathbf{H}_1 \\ \mathbf{H}_2 \\ \mathbf{H}_3 \end{pmatrix} \]

with, for example, \( \mathbf{H}_1 = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 1 & 0 \end{pmatrix} \).

For example, the second row of \( \mathbf{H}_1 \) defines the genotype probabilities for offspring of an \( AA \) male mated to an \( Aa \) female. The probability that the sire has true genotype \( h \) given observed genotype \( i \) and that the dam has true genotype \( y \) given observed genotype \( z \) is given by \( c_{sh} c_{dy} \), elements of the \( 9 \times 9 \) matrix \( \mathbf{C}_s \otimes \mathbf{C}_d \), the direct product of \( \mathbf{C}_s \) and \( \mathbf{C}_d \). Hence, the probabilities of each combination of observed offspring genotype and observed sire \( \times \) dam parental genotype pair are given by the elements of the \( 9 \times 3 \) matrix \( \mathbf{X}^s \), where

\[ \mathbf{X}^s = (\mathbf{C}_s \otimes \mathbf{C}_d) \mathbf{H} \mathbf{E}_s. \]

The likelihood, given that these are the true parents, is then the product over loci \( k \) of elements of \( \mathbf{X}^{s*} \), identified by observed genotypes of putative sire \((i)\) and dam \((z)\) and offspring \((j)\):

\[ L(S, D) = \prod \mathbf{k} x_{kij}. \]

If both putative sire and dam are unrelated to the offspring, the likelihood is given by \( L(U) \) as above, and the log-likelihood ratio by \( lod_{SD} = \log_{10}[L(S, D)] - \log_{10}[L(U)] \). Further, if for example the sire is known, \( \log_{10}[L(S, D)] - \log_{10}[L(S)] \) is the conditional log-likelihood ratio for the dam.

**Posterior Probabilities.** If there are a limited number of sires that could possibly be the father, and these have prior probabilities \( \pi_m \), the posterior probability that sire \( M \) is the true sire is given by

\[ \pi_m = \frac{p_m}{\sum p_m}, \]

where

\[ p_m = \pi_m L_M(S)/\sum \pi_m L_m(S). \]

Approximate values for these prior probabilities may be known: for example, a subset of
the sires used in AI may have on average 10 times as many progeny as those used in natural service.

RESULTS

Properties of Likelihood Ratios

To assess the efficacy of the likelihood ratio comparisons to find the correct parent and to judge the numbers of loci to be used as a function of their frequency, we consider the distribution of the likelihood ratios under the 2 hypotheses that the putative parent is or is not the real parent. If we assume loci are unlinked and in linkage equilibrium, it is sufficient to evaluate these quantities for single loci. Thus, we compute the expectations for sire identification,

\[
E(lod | S) = E[L(S) | S - L(U) | S] \quad \text{and} \quad E(lod | U) = E[L(S) | U - L(U) | U],
\]

and the variances of each of these quantities. The expectations were computed by summing the likelihoods over all possible genotypic combinations, weighted by their frequencies according to whether the putative sire was the sire, to obtain \( E(lod | S) \), or was unrelated, to obtain \( E(lod | U) \). From these we compute:

\[
E(lod_S) = E(lod | S) - E(lod | U),
\]

the expected difference in lod score between the sire and a putative sire unrelated to it. The actual parameters used in computing their expectation to weight the likelihoods and those assumed in computing the likelihoods do not have to be the same; therefore, the effects of wrong assumptions can be checked.

Results are given in Table 2 for cases in which the gene frequency is the same in the male and female populations, \( p_m = p_l = p \), and only for \( p \leq 0.5 \), because those for gene frequency \( p \) and \( 1 - p \) are the same. In each case, there is assumed to be a genotyping error rate \( e \) of 0.005 (based on Kuruvilla et al., 2006). As an example, consider results for sire identification with \( p = 0.5 \) for a single locus when \( E(lod | S) = 0.0718 \) and \( E(lod | U) = -0.2126 \), with the difference \( E(lod) = 0.2845 \) (Table 2). With \( n \) loci, these expected values increase \( n \)-fold; for example, with 100 loci the expectations are 7.18, -21.26, and 28.45, respectively. The standard deviation of the lod score given that the putative sire is indeed the sire, \( \sigma(lod | S) = \sqrt{E[((lod | S) - E(lod | S))^2]} \), is 0.148, and correspondingly, if it is not the sire, it is 0.683 (Table 2). The standard deviation is much greater for the nonparent, because exclusions occur with binomial frequencies, and each such exclusion generates a large negative lod (with magnitude depending on the assumed genotyping error rate). There is no meaningful standard deviation of the difference of the lod under the 2 hypotheses, because the calculations are conditional on the hypothesis. The standard deviation \( \sigma \) increases in proportion to \( \sqrt{n} \) for \( n \) loci, and thus, \( E(lod | S) / \sigma(lod | S) \) increases in proportion to \( \sqrt{n} \). For example, with 100 loci of frequency 0.5, the mean and standard deviation of the lod, if the putative sire is the true sire, is 7.18 ± 1.48; therefore, a negative lod score would be very improbable. Loci at intermediate frequency show the greatest values of \( E(lod | S) \), presumably because these have the greatest heterozygosity and also the greatest chance of showing an exclusion.

The likelihood ratios depend on the frequencies in both parental populations, even when only 1 parent is to be identified. This is illustrated in Figure 1, which shows \( E(lod_2) \) as curves for a series of gene frequencies \( p_2 \) plotted against frequency in the dam population \( p_1 \). In general, the differences in lod are maximized when the gene frequency in the dam population is near

| \( p_m = p_l = p \) | \( E(lod | S) \) | \( \sigma(lod | S) \) | \( E(lod | U) \) | \( \sigma(lod | U) \) | \( E(lod_2) \) |
|---|---|---|---|---|---|
| 0.1 | 0.0430 | 0.192 | -0.0519 | 0.256 | 0.0950 |
| 0.2 | 0.0567 | 0.184 | -0.1053 | 0.432 | 0.1620 |
| 0.3 | 0.0651 | 0.168 | -0.1587 | 0.569 | 0.2237 |
| 0.4 | 0.0701 | 0.154 | -0.1981 | 0.654 | 0.2682 |
| 0.5 | 0.0718 | 0.148 | -0.2126 | 0.683 | 0.2845 |

\( E(lod_2) = E(lod | S) - E(lod | U) \),

\( \sigma(lod | S) = \sqrt{E[((lod | S) - E(lod | S))^2]} \),

\( E(lod | S) \) and \( E(lod | U) \) are maximized when the gene frequency in the dam population is near 0.5.
0 or 1. This is, presumably, because there is no ambiguity as to the source of an A gene, say, in a heterozygous offspring if the dam line is fixed for aa. (Indeed, this is just an example of the utility of inbreds in test crossing.)

The expected lod scores between a parental pair of true sire and dam and an unrelated pair of putative parents differ substantially more than that between just sire and unrelated individuals (Table 2), because the information content is much greater. For example, an AA sire and aa dam can have homozygous progeny only as a result of genotyping errors or mutation.

**Consequences of Errors in Assumptions**

The results in Table 2 and Figure 1 were obtained under the assumption that the gene frequencies used in the calculations were indeed those in the parental populations, but in practice and certainly early in use of a parent assignment scheme, these frequencies may not be known accurately. Examples are given in Table 3 for frequencies in error by 0.1 or so, more than would be expected for all but the smallest initial genotype samples (under 50 or so individuals). It is seen that, although the expected likelihood ratios under both the unrelated and sire assumptions are influenced by the assumed gene frequency, the difference in $E(lod_s)$ is little affected over these examples. It seems possible for $E(lod_s)$ to be greater with incorrect assumptions on frequency, basically because similarity of genotype for a rare allele of father and son is expected to be rarer than it actually is. In general, however, the methodology appears to be robust.

Similarly, genotyping error rates are unlikely to be estimated accurately, certainly for each of the loci used. Some examples are given in Table 4 of the consequences on mean and standard deviation of lod scores of different actual and assumed rates of errors in genotyping (Kuruvilla et al., 2006), but assuming that the same rates apply in sire, dam, and offspring popula-

---

### Table 3. Influence of errors in estimates of gene frequency ($p_m, p_f$) on expectations and standard deviations of lod scores ($=\log_{10}(\text{likelihood ratio})$) for sire identification

| $p_m = p_f$ | $lod | S$ | $lod | U$ |
|-------------|-------|-------|
| Actual      | Assumed | $E(lod | S)$ | $\sigma(lod | S)$ | $E(lod | U)$ | $\sigma(lod | U)$ | $E(lod_s)$ |
| 0.4         | 0.4    | 0.0701 | 0.154 | $-0.1981$ | 0.654 | 0.2682 |
| 0.4         | 0.5    | 0.0808 | 0.152 | $-0.1843$ | 0.664 | 0.2651 |
| 0.4         | 0.3    | 0.0821 | 0.191 | $-0.1875$ | 0.646 | 0.2896 |
| 0.3         | 0.3    | 0.0651 | 0.168 | $-0.1583$ | 0.569 | 0.2237 |
| 0.3         | 0.4    | 0.0762 | 0.149 | $-0.1414$ | 0.587 | 0.2176 |
| 0.3         | 0.5    | 0.1077 | 0.159 | $-0.1033$ | 0.605 | 0.2110 |

$^1$Genotyping error rate $e = 0.005$. S = sire; U = unrelated.

---

**Figure 1.** Expected lod scores $=[\log_{10}(\text{likelihood ratio})]$ for sire identification, $E(lod_s)$, plotted for a range of gene frequencies ($p_m$) in the sire population (lines) and gene frequencies in the dam population ($p_f$). Genotyping error rate ($e$) is 0.005.
tions. In Table 4, errors in genotyping both homozygotes and heterozygotes are included, as in Tables 2 and 3. It is seen (Table 4) that, for the sire, \(E(\text{lod}|S)\) is not very sensitive to either the real or assumed error rate, although tending to fall with increased real error rate and rise with increased assumed error rate. Its standard deviation changes more and in the opposite direction, thus assuming a low error rate leads to an increasing probability of a negative value. For an unrelated individual, the actual rate has little effect on \(E(\text{lod}|U)\) or \(\sigma(\text{lod}|U)\), but a low assumed rate increases the magnitude of negative values. For the sire, the greater the penalty attached to an apparent exclusion, the distribution of \(\text{lod}|U\) is skewed downwards by these tiny probabilities, because the apparent exclusion completely outweighs all of the data coming from other loci which have nonexcluded genotype combinations. If the genotyping error rate is assumed to be zero, when actually it is not, then \(E(\text{lod}|U)\) becomes infinite.

The relative effect of genotyping errors of homozygotes and heterozygotes is analyzed in Table 4. The heterozygote rate was kept constant at 0.005, and the real and assumed rate in homozygotes varied. The means and standard deviations of the \(\text{lod}\) scores for sire evaluation are little affected by genotyping errors of homozygotes, basically because they do not lead to exclusions, except in the very rare case that the wrong homozygote is assigned. The effect is greater when assigning 2 parents (data not shown), in which falsely reading as \(Aa\) an \(AA\) offspring of an \(AA \times AA\) mating leads to an apparent exclusion.

### Table 4. Influence on expectations and standard deviations of lod scores (=\(\log_{10}\)likelihood ratio) for sire identification of actual, assumed genotyping error rate (\(e\)) of heterozygotes and homozygotes, and actual and assumed genotyping error rates in homozygotes with the actual and assumed error rate for heterozygotes of 0.005

| Error rate (\(e\)) | Actual  | Assumed | \(E(\text{lod}|S)\) | \(\sigma(\text{lod}|S)\) | \(E(\text{lod}|U)\) | \(\sigma(\text{lod}|U)\) | \(E(\text{lod})\) |
|-------------------|---------|---------|---------------------|---------------------|---------------------|---------------------|---------------------|
| 0.000             | 0.005   | 0.0729  | 0.137              | −0.1980             | 0.654               | 0.2709              |                     |
| 0.001             | 0.005   | 0.0724  | 0.140              | −0.1980             | 0.654               | 0.2704              |                     |
| 0.005             | 0.005   | 0.0701  | 0.154              | −0.1981             | 0.654               | 0.2682              |                     |
| 0.010             | 0.005   | 0.0673  | 0.170              | −0.1983             | 0.654               | 0.2656              |                     |
| 0.005             | 0.0001  | 0.0686  | 0.190              | −0.3939             | 1.192               | 0.4624              |                     |
| 0.005             | 0.001   | 0.0697  | 0.167              | −0.2786             | 0.875               | 0.3483              |                     |
| 0.005             | 0.005   | 0.0701  | 0.154              | −0.1981             | 0.654               | 0.2682              |                     |
| 0.005             | 0.01    | 0.0700  | 0.148              | −0.1636             | 0.559               | 0.2335              |                     |
| 0.000             | 0.000   | 0.0709  | 0.155              | −0.2006             | 0.658               | 0.2714              |                     |
| 0.000             | 0.005   | 0.0709  | 0.154              | −0.2001             | 0.657               | 0.2710              |                     |
| 0.005             | 0.000   | 0.0701  | 0.154              | −0.1986             | 0.655               | 0.2687              |                     |
| 0.005             | 0.005   | 0.0701  | 0.154              | −0.1981             | 0.654               | 0.2682              |                     |

\(^1\)Gene frequencies \(p_m = p_f = 0.4\). \(S\) = sire; \(U\) = unrelated.

### Distinguishing Parents from Relatives of the Parents

Consider a relative of the sire, with relationship \(R\) to it (strictly the numerator relationship, or twice the coancestry). Because the offspring shares 1 gene with the true parent, the probability is \(R\) that it shares 1 gene with the relative of the parent. Thus, \(R = 0.5\) for a full sib or parent of the sire (i.e., an uncle or grandparent of the progeny), and \(R = 0.25\) for a half-sib of the sire (i.e., a half-uncle of the progeny).

The expectation of the lod for a putative sire related to the true sire is therefore the weighted average of the lod for a sire and for an unrelated individual:

\[ E(\text{lod}|R) = R E(\text{lod}|S) + (1 - R) E(\text{lod}|U); \]

and therefore,

\[ E(\text{lod}|R) - E(\text{lod}|U) = R E(\text{lod}) \]

or

\[ E(\text{lod}|S) - E(\text{lod}|R) = (1 - R) E(\text{lod}) \]

Similarly, the standard deviation of the lod is given by that for the mix of 2 groups:

\[ \sigma(\text{lod}|R) = \sqrt{R \sigma^2(\text{lod}|S) + (1 - R) \sigma^2(\text{lod}|U)} \]

\[ + R(1 - R)[E(\text{lod}|S) - E(\text{lod}|U)]^2 1/2. \]

Results given in Table 2 and Figure 1 can be used accordingly. Basically, it is harder to distinguish a sire from a full uncle or grandsire than from a half-uncle than, in turn, an unrelated individual.
The probability of incorrect assignment is small, an incorrect sire assignment would usually be to a full sib in these examples. On this basis, further simulations were run with many more replicates (1,000,000, comprising 1,000 sires each with 1,000 offspring) and more loci, but considering only 4 full sibs in addition to the true sire, such that more accurate estimates of the risk could be obtained when it was low (Table 6). With 250 loci and minor allele frequencies of 0.25 or more, the probability of wrong identification is less than 10^-5. Some values from Tables 5 and 6 are plotted together in Figure 2, which shows that the risk of wrong assignment falls approximately linearly as \(-\log(\text{number of loci})\).

Poor estimates of genotyping error rates or of gene frequencies do not greatly increase the risk of incorrect assignment of the sire (Table 7), as anticipated from the analysis of expected lod (Table 4). Indeed, as might be hoped, the real values have a greater influence than do the assumed values of these parameters.

**DISCUSSION**

**Exclusions**

A simple alternative to the use of ML to identify the parent(s) is to eliminate others by simple exclusion of genotype combinations, for example an AA sire can-

---

### Table 5. Frequency from 50,000 replicates in which the lod score (=\(\log_{10}(\text{likelihood ratio})\)) for one or more of its 4 full sibs, 20 half-sibs, or 976 unrelated males exceeded the lod of the true sire for different actual and assumed gene frequencies (\(p = p_m = p_0\))

<table>
<thead>
<tr>
<th>Loci</th>
<th>50</th>
<th>75</th>
<th>100</th>
<th>125</th>
<th>150</th>
<th>175</th>
<th>200</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>42.4</td>
<td>14.36</td>
<td>39.9 (1.43)</td>
<td>0.886</td>
<td>0.314 (0.218)</td>
<td>0.076</td>
<td>0.040 (0.034)</td>
</tr>
<tr>
<td>0.25</td>
<td>34.4</td>
<td>8.18</td>
<td>1.68 (0.73)</td>
<td>0.380</td>
<td>0.088 (0.074)</td>
<td>0.004</td>
<td>0.004 (0.004)</td>
</tr>
<tr>
<td>0.3</td>
<td>29.8</td>
<td>5.55</td>
<td>0.98 (0.49)</td>
<td>0.176</td>
<td>0.038 (0.036)</td>
<td>0.008</td>
<td>0.000 (0.000)</td>
</tr>
<tr>
<td>0.35</td>
<td>25.6</td>
<td>3.92</td>
<td>0.41 (0.24)</td>
<td>0.076</td>
<td>0.008 (0.008)</td>
<td>0.006</td>
<td>0.000 (0.000)</td>
</tr>
<tr>
<td>0.4</td>
<td>20.7</td>
<td>2.38</td>
<td>0.27 (0.18)</td>
<td>0.038</td>
<td>0.010 (0.010)</td>
<td>0.002</td>
<td>0.000 (0.000)</td>
</tr>
<tr>
<td>0.45</td>
<td>19.8</td>
<td>2.16</td>
<td>0.24 (0.17)</td>
<td>0.042</td>
<td>0.010 (0.010)</td>
<td>0.002</td>
<td>0.000 (0.000)</td>
</tr>
<tr>
<td>0.5</td>
<td>21.0</td>
<td>2.12</td>
<td>0.23 (0.16)</td>
<td>0.026</td>
<td>0.002 (0.002)</td>
<td>0.000</td>
<td>0.000 (0.000)</td>
</tr>
</tbody>
</table>

1Values indicate the percentage assigned to an incorrect sire. In parentheses are examples of the frequency in which the lod for the sire was exceeded by that for 1 or more of its full sibs. The genotyping error rate \(e = 0.005\).
not produce an aa offspring. If there are no genotyping errors or mutations and there is random mating, the probability of exclusion of a randomly sampled unrelated individual is the frequency of different homozygote pairs, \( P_X = p_m^2 (1 - p_o)^2 + (1 - p_m p_o)^2 \). If \( p_m > 0.5 \), \( P_X \) is maximized when \( p_o = 0 \), otherwise when \( p_m = 1 \). With equal frequencies in the lines \( P_X = 2p^2q^2 \), which has a maximum of 0.125 at \( p = 0.5 \), is 0.1152 if \( p = 0.4 \), but is only 0.0162 if \( p = 0.1 \). The probability of exclusion for a related sire, for example, is simply a factor \( 1 - R \) of that for an unrelated individual, assuming there are no genotyping errors.

To enable comparison with results from the likelihood analysis, consider the example in which the sire has 4 full brothers, 20 half-brothers, and 976 unrelated males in the data set. The probability that at least 1 putative sire is, by chance, not excluded if 100 loci each with frequency 0.4 are genotyped without error is given by:

\[
1 - \left[ 1 - (1 - (0.1152)^{100})^{976} \right] [1 - (1 - 0.75) \\
\times 0.1152]^{100} [1 - (1 - 0.5 \times 0.1152)^{100}]^4 \approx 1.8%.
\]

In comparison, there is only a 0.27% chance that the \( lod \) of the sire will be exceeded by that of any of the other putative sires, allowing for a genotyping error rate of 0.0005 (Table 5).

An important complication that we have incorporated quantitatively within the ML analysis is the effect of errors in genotyping or mutation, which can lead to false exclusions and thus potential wrong decisions if genotypic exclusion is the sole criterion. We consider just the most important class, calling heterozygotes as homozygotes, simply assuming \( Aa \) is equally likely to be called as \( aa \) or \( AA \), and ignoring second order terms. If the error rate is \( e_m \) in sires and \( e_o \) in offspring, the exclusion rate for a sire, using Table 1, comprises terms for \( AA \) sires with aa offspring \((1/2p_m q_m p_m e_m)\), AA sires with aa offspring \((1/2p_m q_m p_m e_m)\), etc., totaling \(1/2p_m q_m (p_m e_m + q_m e_m) + 1/2q_m p_m (q_m e_o + p_m e_m)\). With equal gene frequencies in males and females, and error rates in parents and offspring, this reduces to \( pq e \), or, for example, to \( 0.00125 \) for \( p = 0.5 \) and \( q = 0.005 \), the error rate assumed in the ML analysis (Table 2). There is thus a probability of false exclusion based on a single SNP of about 12% with 100 SNP tested. For gene frequencies closer to 0 or 1, the probability of a false exclusion of a sire increases relative to that of the correct exclusion of a nonsire \((e/2pq \text{ approximately})\).

### Practical Issues

An alternative to using SNP would be to adopt highly polymorphic markers such as microsatellites or variable number tandem repeats. Although many more SNP are needed to achieve the same level of precision in identifying relatives (Brenner, 1999; Gill, 2001; Vignal et al., 2002), many costs are the same, such as collection of blood-tissue samples and DNA extraction. The developments in technology are such that relative costs and accuracies of high-throughput genotyping are changing such that SNP are becoming the method of choice. An added advantage is the binary nature of SNP, with 2 alleles, such that error rates are relatively low. Thus, SNP have been proposed for use in paternity testing (Heaton et al., 2002) and product tracing (Plastow et al., 2007) in livestock.

Unsurprisingly, the analysis shows that SNP at intermediate gene frequencies in the sire population are most informative and become increasingly so the nearer are loci in the female line to homozygosity. The fall in efficiency (exemplified by Table 2 for \( E(lod) \) or Table 5 for risk) is small, however, over the range 0.3 < \( p < 0.7 \). The numbers of loci required to be almost certain to identify the correct parent depend on the number of putative competitors and particularly on how many are closely related to the sire (e.g., full sibs or father). Even so, with large numbers of competitors including relatives, 100 to 150 loci seem sufficient, and simulations of data sets of populations having small numbers of breeding individuals so that there are many relationships among the sires illustrate that correct identification occurs with under 150 loci.
The analysis has not considered relationships among putative sires other than full or half-sibs. In practice, there are almost certain to be additional relationships, for example cousins and second cousins, and various multiple relationships. Hence, the calculations of both the exclusion probabilities (Dodds et al., 1996; Double et al., 1997) and the risk of wrong assignment are conservative and so may be greater in practice. Because the risk of wrong assignment is greatest for the closest relatives, full sibs (Table 5), particularly when many loci are used, it is unlikely that further relatives among putative sires will greatly affect the probabilities of misidentification of the sire unless these are more highly related than full sibs. A clone of the sire could not, of course, be distinguished from the sire by genotyping. In the risk calculations, we have not, however, assumed an equal prior probability for all sires, whereas in practice, this is unlikely to be the case. For example, some sires may have been used in AI, and others not, and the pattern of use over time of sires may be known.

In our analysis of risk, we have assumed the genotype of the actual sire is present in the data set and designated as sire the putative sire with the greatest lod score, regardless of the actual and relative sizes of the lod scores of the winner and runners-up. In practice, we advocate presenting lod scores for at least the top few candidates for perusal and decision as in CERVUS (Marshall et al., 1998; Kalinowski et al., 2007; http://www.fieldgenetics.com/pages/aboutCervus_Functions.jsp). In the context of product tracing, if sufficiently important, possible ties or marginally convincing assignments could be resolved by further genotyping using an additional set of SNP (or a search for the absent fathers).

Simulations can be undertaken, as in CERVUS for example, on the actual data structure to assess the confidence of the parentage assignment. In the more limited context of product tracing, in which only a single relationship is to be identified and the population is known, the information on the standard deviations of lod scores would seem to provide a guide in most cases.

We use the results on the distributions of lod scores, which can readily be computed for any population. As a simple example, assume all gene frequencies are 0.4 and taking results from Table 2, for a single locus, the mean and standard deviation of lod 1S are 0.0701 ± 0.154. Thus, for 100 such markers, lod |S ~ 7.01 ± 1.54; similarly, for unrelated individuals, lod |U ~ 19.81 ± 6.54; and for full brothers, lod |R ~ 6.40 ± 4.94. As noted previously, the distribution of the lod scores, even with large numbers of loci, departs from normality, in particular being skewed downwards, but even so, the normal distribution provides a guide. Assuming no relative could be closer than a full sib, then an observed lod in excess of 5 is strong evidence that the sire has been identified, whereas a negative lod is strong evidence that the sire has not been found.

Although genotyping errors should, of course, be minimized, the analysis shows that the ML method is reasonably robust, providing the error rate is not assumed to be too small (when false exclusions generate very high negative lod scores). The actual error rate to be used will depend on the genotyping platform and assignment rate in use. Therefore, the base figure of 0.005 mainly used should be regarded as no more than a typical figure.

The application considered here to product tracing is simpler than other potential applications, because the kind of relationship to be identified is limited solely to parents. In this case, there is clearly more information to be gained from a limited number of loci by identifying both parents, even though the actual matings made are not recorded (e.g., artificial insemination using mixed semen), because the likelihood ratios are much larger (Table 2). If sire tracing is sufficient, it may be much more economical, because fewer animals need to be genotyped.

**LITERATURE CITED**


