Error tolerant parent identification from a finite set of individuals

MAGALI SANCRISTOBAL* AND CLAUDE CHEVALET

Laboratoire de Génétique Cellulaire, Institut National de la Recherche Agronomique, BP 27, 31326 Castanet Tolosan Cédex, France

(*Received 11 December 1996 and in revised form 24 February and 20 April 1997*)

Summary

We consider using microsatellites for paternity checking and parent identification in different population structures, and allowing for possible typing errors or mutations. Statistical rules derived from the Bayesian and the sampling approaches are discussed in the case involving the choice of the true father–mother pair among a finite set of possible parental pairs. General situations are investigated by means of random simulations, in order to characterize the joint influences of the number and polymorphism of typed loci, the population structure and size, and error rates. Approximate expressions are provided that give the efficiency of a set of markers for identifying the parents in various mating schemes. The importance of a non-zero value for the typing error rate in the likelihood is highlighted.

Introduction

Polymorphic genetic markers have long been used in cases where pedigree information must be ascertained, as in animal breeding selection programmes, or in human paternity analysis (Jamieson, 1965; Chastang, 1973; Hanset, 1975; Elston, 1986). When several potential fathers are proposed for some offspring, each one is checked for its genotype being compatible with the genotypes of the offspring and the mother. A set of loci is characterized by the probability of exclusion, i.e. the probability that one random individual cannot be the father of some proposed individual whose maternal and own genotypes are known. The calculation relies on the frequency of genotypes that are impossible for the true father. The sum of frequencies of all impossible genotypes is the probability of exclusion. It depends only on the frequencies of alleles, and relies on the hypothesis that parents are taken at random in a population with Hardy–Weinberg structure.

Another question is identifying the most probable parents of some individual. Searching for the most likely father happens in certain forensic situations, as well as in animal breeding, e.g. identifying the sire among a set of natural service bulls. Identifying parents is a problem arising in wild-life surveys, analysis of the genetic structure of populations, and in experimental quantitative genetics in fishes (fishes from different sibships are reared together and cannot be identified except through genetic analysis) and trees (retrieving parent trees participating in open pollination).

Large-scale parentage analyses can be divided into three types: (i) identifying one parent when the other is known, (ii) identifying one parent with no information about the other parent, (iii) identifying a parental pair starting with no prior information. We will restrict our analysis here to the first and last cases when the potential parents belong to a finite set of genotyped individuals. Case (ii) involving a parent in an infinite population would need prior knowledge about allele frequencies and a different statistical setting. The potential parents and the offspring are described at a number of unlinked loci, from which data we derive statistical rules in order to identify the parents of the offspring. We consider only single-locus Mendelian systems, and focus on the class of highly polymorphic co-dominant markers provided by microsatellite repeats (Weber & May, 1989; Weber, 1990), which are now available in many species. We assume that genotyping errors may arise, due to the large number of experiments. Mutations that may arise for highly polymorphic markers will also be handled like errors.

We first derive the likelihood of one offspring's phenotype, assuming that parental genotypes are known without error, and then the probability that

* Corresponding author.

https://doi.org/10.1017/S0016672397002851 Published online by Cambridge University Press

any pair of parents has given rise to this offspring. The section is then devoted to the building of the statistical tests needed to answer the question raised. Analysing a simple case allows the main effects of locus polymorphism and error rates to be outlined.

In the second part of the paper we consider various population schemes encountered in animal and plant breeding, and provide numerical results about the actual powers of parent identification tests, according to error rates, the number of loci and the distributions of alleles per locus.

2. Genetic and statistical framework

(i) *Genetic hypotheses*

Let Π_l be the phenotype of some offspring at one locus l $(1 \leqslant l \leqslant L)$, and $\Pi = (\Pi_1, \ldots, \Pi_l, \ldots, \Pi_L)$ the joint phenotypes at the L loci being considered. All calculations are made conditionally on the genotypes of parents taken from a set of potential parental pairs indexed by k $(1 \leqslant k \leqslant N)$. Any discrepancies between observed phenotypes and expected genotypes of offspring are referred to as 'typing errors' and characterized by small numbers, such as the probability that allele (A_k) is seen as (A_i).

(a) *Conditional probability of offspring phenotypes*

At any locus, we first consider the probability of offspring genotypes given the parent genotypes. Let $(A_k A_l)$ and $(A_m A_n)$ be the genotypes of the two parents. In the case of a homozygote genotype $(A_i A_i)$ of offspring, the conditional probability reads

$$\begin{aligned}&\Pr((A_i A_i) \mid (A_k A_l), (A_m A_n))\\ &\qquad = \Pr((A_i) \mid (A_k A_l)) \Pr((A_i) \mid (A_m A_n)),\end{aligned}$$

where $\Pr((A_i) \mid (A_k A_l))$ is the probability that an individual with genotype $(A_k A_l)$ has transmitted an allele (A_i) to its offspring. To allow for error rates or mutations, we write this transmission probability $T(i/kl)$ as

$$\begin{aligned}T(i/kl) &= \Pr((A_i) \mid (A_k A_l))\\ &= \tfrac{1}{2}\epsilon_{ki} + \tfrac{1}{2}\epsilon_{li},\end{aligned}$$

where ϵ_{ki} is the probability that allele (A_k) from a parent yields an allele (A_i) in some offspring: without error or mutation, it is zero if $i \neq k$, or 1 if $i = k$. Allowing for errors, we consider that ϵ_{ki} is small if $i \neq k$, and nearly 1 if $i = k$.

For a heterozygote *phenotype* $[A_i A_j]$ of offspring, we must consider the two alternative *ordered* genotypes $(A_i A_j)$ and $(A_j A_i)$ corresponding to the two possible parental origins of alleles. The probability of *phenotype* $[A_i A_j]$ conditional on parents being $(A_k A_l)$ and $(A_m A_n)$ then reads

$$\begin{aligned}&\Pr([A_i A_j] \mid (A_k A_l), (A_m A_n))\\ &\qquad = T(i/kl)\, T(j/mn) + T(j/kl)\, T(i/mn).\end{aligned}$$

We describe error and mutation rates with a single parameter, a mean overall rate ϵ of incorrect transmission, so that

$$\epsilon_{kk} = 1 - \epsilon$$

and, for any $i \neq k$,

$$\epsilon_{ki} = \frac{\epsilon}{a-1}, \tag{1}$$

where a is the number of alleles (i.e. assuming that misinterpretation of alleles does not depend on allelic types).

For a set of unlinked loci, the probability of phenotypes Π in offspring, conditional on the genotypes $G^{sd}(k)$ of the kth parental pair, is

$$\Pr(\Pi \mid G^{sd}(k), \epsilon) = \prod_{l=1}^{l=L} \Pr(\Pi_l \mid G_l^{sd}(k), \epsilon), \tag{2}$$

where $\Pr(\Pi_l \mid G_l^{sd}(k), \epsilon)$ is calculated from previous expressions.

Example 1. Suppose two loci are typed. The first is triallelic and the second biallelic. Let the genotype of the parental pair be $(A_1 A_1, A_2 A_3)$ at locus 1 and $(B_1 B_2, B_2 B_2)$ at locus 2, while the phenotype of a given offspring is $[A_1 A_2]$ at locus 1 and $[B_1 B_1]$ at locus 2.

The conditional probabilities of observed phenotypes are, assuming that errors occur independently for different gametes,

$$\begin{aligned}&\Pr([A_1 A_2] \mid (A_1 A_1), (A_2 A_3), \epsilon)\\ &\qquad = T(1/11)\, T(2/23) + T(2/11)\, T(1/23)\\ &\qquad = (1-\epsilon)\left(\tfrac{1}{2}(1-\epsilon) + \tfrac{1}{2}\frac{\epsilon}{2}\right) + \frac{\epsilon}{2}\frac{\epsilon}{2},\\ &\Pr([B_1 B_1] \mid (B_1 B_2), (B_2 B_2), \epsilon) = T(1/12)\, T(1/22)\\ &\qquad = \left(\tfrac{1}{2}(1-\epsilon) + \tfrac{1}{2}\epsilon\right)\epsilon.\end{aligned} \tag{3}$$

(b) *Posterior probability of parental origins*

Consider an individual with phenotype Π, and the genotypes $G^{sd}(k)$, $k = 1, \ldots, N$, of N potential parental pairs. Let θ be the unknown indicator vector of the origin of the offspring among the finite set of these possible parental pairs: its kth component is equal to 1 if and only if the offspring derives from the kth pair, and the other components are 0. This parameter θ can take a finite number of values among the N vectors t_h $(h = 1, \ldots, N)$, where $t_h = (0, \ldots, 0,\ 1,\ 0, \ldots, 0)$ has 0 components except the hth one equal to 1. With these notations, the probability ρ_k that the kth pair has given birth to an offspring with phenotype Π is, for a given ϵ and conditional on the $G^{sd}(h)$ values $(h = 1, \ldots, N)$,

$$\begin{aligned}\rho_k(\epsilon) &= \Pr(\theta = t_k \mid \Pi, \mathbf{G}^{sd}, \epsilon)\\ &= \frac{\Pr(\{\theta = t_k\}, \Pi, \mathbf{G}^{sd}, \epsilon)}{\Pr(\Pi, \mathbf{G}^{sd}, \epsilon)},\end{aligned} \tag{4}$$

https://doi.org/10.1017/S0016672397002851 Published online by Cambridge University Press

where $\mathbf{G}^{sd}$ stands for $\{G^{sd}(h), h = 1, \ldots, N\}$. Writing the numerator as

$$\Pr(\Pi \mid \{\theta = t_k\}, \mathbf{G}^{sd}, \epsilon) \cdot \Pr(\{\theta = t_k\}, \mathbf{G}^{sd}, \epsilon)$$

the first factor depends only on the genotypes $G^{sd}(k)$ of the kth pair, since the condition includes $\{\theta = t_k\}$. Hence, it is equal to the value given by (2). The second factor can be factorized into

$$P_0(\theta = t_k)\, P_0(\epsilon)\, P_0(\mathbf{G}^{sd})$$

with prior probabilities (symbol 'P_0') concerning the distribution of error rates, of the events $\{\theta = t_k\}$, and of parents' genotypes. Some information available from population structure – such as distances between trees in a problem of open pollination, or such as breed structures – might be introduced here.

In (4) the denominator is the sum of terms analogous to the numerator, so that we get

$$\rho_k(\epsilon) = \frac{\Pr(\Pi \mid G^{sd}(k), \epsilon)\, P_0(\theta = t_k)}{\sum_{h=1}^{h=N} \Pr(\Pi \mid G^{sd}(h), \epsilon)\, P_0(\theta = t_h)}. \quad (5)$$

(ii) *Statistical analysis*

Depending on the number and informativity of loci, error rates, and the relationship structure among parents, we have to evaluate the distribution of the ρ_k indices.

(a) *Bayesian approach*

In the present genetic problem, the parameter of prime interest is θ, and the set of ϵ_{ki} values are considered as nuisance parameters. Unless biological or spatial features are known, simple vague prior information on the probability of descent can be translated as equal probabilities, while the distribution of error rates is characterized only by its expectation (assuming independence of causes that may yield errors or mutations).

Given observations (parents' genotypes $\mathbf{G}^{sd}$, and the phenotype Π of the offspring), the posterior distribution of the primary parameter θ is obtained after integrating on ϵ, and is given by the following set of N values (for $k = 1, \ldots, N$):

$$\rho_k = \Pr(\theta = t_k \mid \Pi, \mathbf{G}^{sd}) \propto \Pr(\Pi \mid \{\theta = t_k\}, \mathbf{G}^{sd})\, P_0(\theta = t_k).$$

This means that the posterior distribution of θ is multinominal with parameters 1 and $(\rho_1, \ldots, \rho_k, \ldots, \rho_N)$ equal to:

$$\rho_k = \frac{\Pr(\Pi \mid G^{sd}(k))\, P_0(\theta = t_k)}{\sum_{h=1}^{N} \Pr(\Pi \mid G^{sd}(h))\, P_0(\theta = t_h)}.$$

The difference compared with (5) is that the nuisance parameter ϵ has been integrated out. In fact, under the hypothesis that errors occur independently while reading allelic types, this integral form is obtained from (5) after substituting ϵ values by their prior expectations.

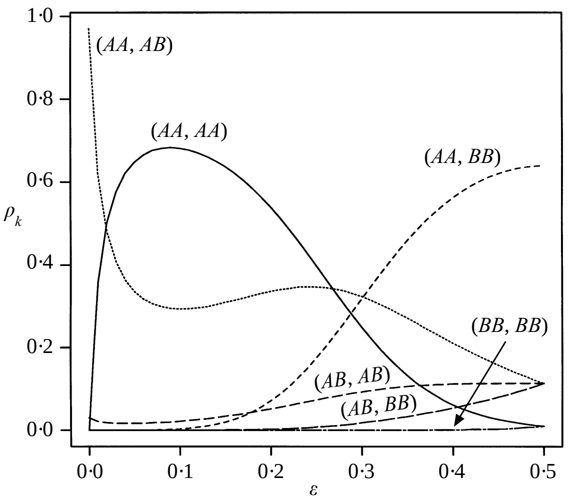

Fig. 1. Estimation of the parental pair by maximum likelihood. Six possible parental pairs are considered, with joint genotypes (the same at six independent loci): (AA, AA), (AA, AB), (AA, BB), (AB, AB), (AB, BB) and (BB, BB), giving six curves ρ_k, $k = 1, \ldots, 6$, depending on the error rate ϵ. The observed offspring phenotype is $[AA]$ at 5 loci and $[AB]$ at 1 locus.

Following Hoel & Peterson (1949; see Ferguson, 1967, p. 291) an optimal multiple decision rule consists here in choosing k corresponding to the largest ρ_k, over $1, \ldots, N$.

(b) *Sampling approach*

Unknown parameters θ and ϵ can be estimated using the maximum likelihood approach. Since the parameter θ is discrete, the likelihood function $\mathscr{L}(\theta, \epsilon) = \Pr(\Pi \mid \mathbf{G}^{sd}, \theta, \epsilon)$ takes its maximum at a point $(\hat{\theta} = t_{\hat{k}}, \hat{\epsilon})$ corresponding to the greatest conditioned probability of phenotype Π (2) over k ($k \in \{1, \ldots, N\}$) and ϵ values. Numerically, this corresponds to the mode of the joint posterior distribution of θ and ϵ in the Bayesian theory when flat priors are chosen. Fig. 1 gives an illustration of the maximum likelihood approach. For each genotypic state (k) of a parental pair, it shows the likelihood ratio ρ_k as a function of ϵ. As soon as the error rate is not very small, the likelihood of the first type (AA, AA) becomes the largest one, suggesting that it should be preferred to states (AA, AB) and (AB, AB) that are compatible from the strict qualitative (Mendelian) point of view. In other words, unless ϵ is very small, the corresponding joint Mendelian transmission of alleles is less probable than the occurrence of one error.

Assume that ϵ is fixed at some known value. If the maximum likelihood estimator of θ is such that $\hat{\theta} = t_k$, the null hypothesis H_0: $\theta = t_{k'}$, for some $k' \mp k$, can be tested against the alternative H_1: $\theta = t_k$.

The type I error α (the *size* of the test) is the probability that the null hypothesis is wrongly rejected.

https://doi.org/10.1017/S0016672397002851 Published online by Cambridge University Press

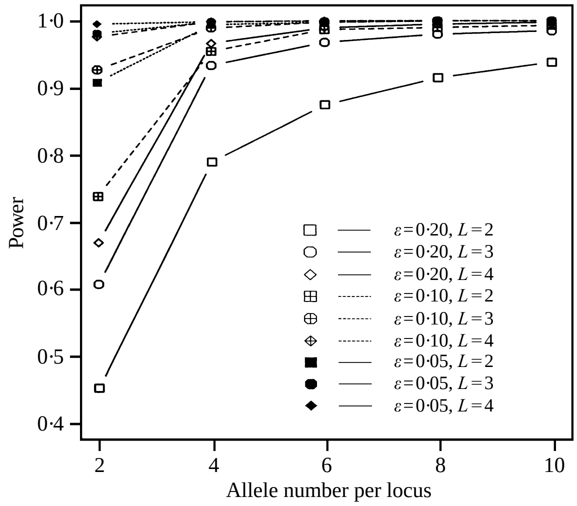

Fig. 2. Power of the test for $H_0: \theta = t_k$ versus $H_1: \theta = t_{k'}$ with parent genotypes $G^{sd}(k) = (A_1 A_1, A_2 A_2)$ and $G^{sd}(k') = (A_1 A_1, A_1 A_1)$, at the 5% size, with varying allele number per locus, number of loci L and typing error rate ϵ (alleles with equal frequencies).

The *power* $1-\beta$, corresponding to a type II error β, is the probability that the null hypothesis is correctly rejected.

According to (2), one can calculate the probabilities of any phenotype Π under both hypotheses H_0 and H_1, and then the distributions of the likelihood ratio

$$LR = \frac{\Pr(\Pi \mid H_1)}{\Pr(\Pi \mid H_0)} = \frac{\Pr(\Pi \mid G^{sd}(k), \epsilon)}{\Pr(\Pi \mid G^{sd}(k'), \epsilon)}$$

under both hypotheses. The Neyman–Pearson lemma allows one to build the most powerful test of given size α. A critical value c_α can be defined such that the decision rule: 'hypothesis H_0 is rejected if the LR value is greater than c_α', has maximal power for the given size α. A procedure to calculate c_α is given in the Appendix, with an example of explicit derivations of size and power.

Example 2. In order to illustrate the effect of alleles number a and of loci number L on the performance of the previous test procedure, we suppose that all the parental pairs are totally inbred, so that any single locus summarizes the whole genome. Moreover, allele frequencies within each locus are assumed equal. Without typing error, the test of

$$H_0: \theta = t_{k'} \quad \text{with} \quad G^{sd}(k') = (A_1 A_1, A_2 A_2)$$

versus

$$H_1: \theta = t_k \quad \text{with} \quad G^{sd}(k) = (A_1 A_1, A_1 A_1)$$

is best of size 0 and has power 1. In that case, one locus is sufficient to decide whether the given individual is an offspring of the k'th pair or the kth pair.

When typing errors occur, the power falls dramatically, as illustrated in Fig. 2 for a size of 5%. The phenomenon is amplified by the increase in typing errors, by a small number of typed loci, and a small number of alleles. Fig. 2 suggests that the biallelic case is the worst, especially when typing errors are numerous, even if several loci are considered. For a constant typing effort (L fixed), multiallelic markers (six alleles or more) provide a very significant improvement.

3. Applications to general situations

(i) *Simulations*

The previous analysis indicates the respective roles of allele number and loci number on the efficiency of paternity retrieval. Although exact calculations are complex (*cf.* Appendix), the dependence of the frequency of correct decisions on markers' informativity, number of loci and error rate can be outlined as follows.

When comparing the correct hypothesis H_1 (the true father is the kth one) against a wrong hypothesis H_0 (testing the k'th father), a wrong decision is taken if the likelihood of H_0 is larger than that of H_1. Roughly speaking, likelihood is large if offspring and parent genotypes are compatible, and it is small if they are not, because an error must be assumed to make parenthood admissible so that the error rate ϵ must be introduced as a factor in the likelihood. Allowing for only one error in the case of a true relationship, two situations may give rise to likelihoods taking similar values under both hypotheses: if the false father is compatible (probability equal to $1-E_g$, where E_g is the exclusion probability), or if the true father is compatible, but for one locus (with probability equal to $LE_1(1-E_1)^{L-1}$, where E_1 is the one-locus probability of exclusion – assuming homogeneous and independent loci). In the latter case, the likelihood for the right father involves the factor ϵ if an error has occurred for one allele transmission at the same locus, an event of probability $\epsilon(1-\epsilon)^{2L-1}$. Hence, the order of magnitude of the frequency of a wrong decision for any pair (k, k') of true and false fathers can be written as

$$1-p_1 \sim (1-E_g) + LE_1(1-E_1)^{L-1}(1-\epsilon)^{2L-1}\epsilon$$

or

$$1-p_1 \sim (1-E_g)(1+K\epsilon),$$

where p_1 stands for the frequency of a correct decision involving one (k, k') comparison, and where K depends on E_g and on ϵ (only slightly for small values). When the true father k is opposed to a finite set of s false fathers, the frequency of joint correct decision is $p_s = p_1^s$ since all the alternatives $k' \neq k$ must be rejected. For large E_g and small ϵ values, one can write

$$\log\left(\frac{p_s}{1-p_s}\right) \sim -\log(s) - \log(1-E_g) - K\epsilon. \qquad (6)$$

Note that this relationship is only qualitative, and that the linear dependence on ϵ is only asymptotic for small values since K is also a function of ϵ and E_g. Although

https://doi.org/10.1017/S0016672397002851 Published online by Cambridge University Press

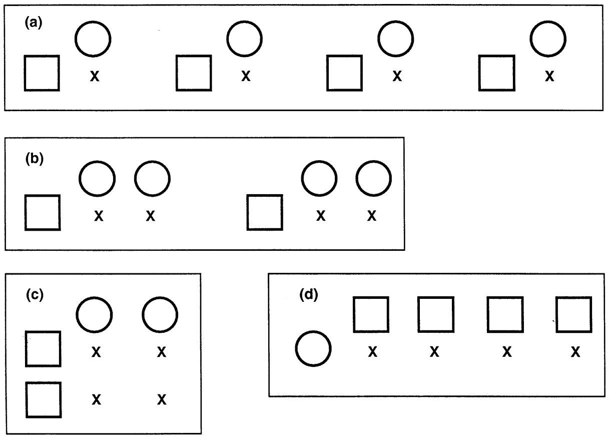

Fig. 3. The four population structures envisioned in the simulations: (*a*) independent pairs, (*b*) hierarchical scheme, (*c*) crossed scheme, (*d*) paternity analysis. Each scheme involves a number of male (squares) and female (circles) parents and their matings (crosses). According to the scheme, each individual can be mated to one or several individuals of the other sex.

obtained in a simple case, these qualitative considerations suggest combinations of parameters which are likely to make quantitative predictions possible in realistic cases, by means of simulations and statistical analysis.

Simulations performed involve various population structures and variations of parameters (number of loci, allelic frequencies, error rates).

(a) *Population structures*

Four structures of population were considered (see Fig. 3):

(i) N independent pairs of $s = N$ unrelated sires and $d = N$ unrelated dams. This could correspond to some natural bird populations.

(ii) s sires are mated to d dams each, so the number of parental pairs is $N = s \cdot d$. This hierarchical scheme is appropriate for animal designs in swine or poultry.

(iii) s males are mated to d females in a crossed scheme, leading to $N = s \cdot d$ parental pairs. Fishes and trees are mated in this way.

(iv) $s = N$ potential fathers when the couple (mother, child) is known, i.e. paternity testing – a special case of the previous two schemes.

Simulations were run with variable numbers of parents (from 5 to 1000).

(b) *Typing error rates and decision rule*

Several typing error rates from 0 to 10% were considered in the simulations, in two instances. A first value ϵ_T is used to generate data with errors occurring at this rate: at each transmission event from a parent to the offspring, a random variable U is drawn in [0, 1] and compared with the error rate. If U is smaller than the error rate, an allele name is drawn at random and transmitted to the offspring. A second given error rate ϵ_C is used to build the likelihood and to make the decision (the ϵ values included in (2)).

In each run, the most probable parental pair is chosen according to the maximum likelihood approach using a fixed value for the error rate ϵ_C. For each set of parameter values, and for each run, the most probable pair was checked against the true parents. The criterion used is the frequency of good and unique decisions: only cases involving no uncertainty are considered to estimate the parental pair. Cases with several pairs sharing the same likelihood are recorded as erroneous decisions.

(c) *Allele frequencies and polymorphism measures*

Actual allele frequencies in populations are not equal as we assumed in the analytical derivations. To propose more realistic results, polymorphism was generated at two levels: the number of loci and the number of alleles per locus. This should allow the present results to be used in a large variety of situations, since available polymorphism in populations is not controllable.

Analytical results suggest that 5 loci with 5 alleles per locus are minimal requirements, while 8 loci with 8 equiprobable alleles at each locus are very informative. We therefore simulated random unequal allele frequencies following a uniform distribution, in the range of 5 to 8 alleles per locus, while the number of loci was drawn at random between 5 and 8 (inclusive). For each population structure and choice of error rates, ten systems were drawn from this distribution, and their polymorphism characterized by the global probability of exclusion (E_g): if E_l is the probability of exclusion attached to the lth locus, the global probability of exclusion for a system of L loci is

$$E_g = 1 - \prod_{l=1}^{l=L} (1 - E_l).$$

E_l is calculated after the expressions given by, e.g. Hanset (1975) or Smouse & Chakraborty (1986), from the first five moments μ_1 to μ_5 of the distribution of allele frequencies, assuming Hardy–Weinberg genotype frequencies in the population:

$$E_l = 1 - 2\mu_2 + \mu_3 + 3(\mu_2 \mu_3 - \mu_5) - 2(\mu_2^2 - \mu_4).$$

Global heterozygosity and global polymorphic information content (PIC) were defined in the same way as the global exclusion probability. They were also considered as potential predictors of polymorphism in the present context, although they are highly correlated (the complementary global exclusion probability $1 - E_g$ and the complementary global PIC are linearly dependent on a logarithm scale, at fixed number of loci (not shown)).

https://doi.org/10.1017/S0016672397002851 Published online by Cambridge University Press

Table 1. *Coefficients of the linear predictor η (standard errors) for the generalized linear model with binomial error and logit link*

	Independent pairs	Hierarchical scheme	Crossed scheme	Paternity search
Intercept	2·22 (0·11)	1·37 (0·09)	1·44 (0·05)	1·56 (0·05)
ϵ	−1·68 (1·24)	1·98 (0·87)	1·29 (0·25)	−1·82 (0·37)
$\log(1-E_g)$	−1·17 (0·02)	−0·96 (0·01)	−0·97 (0·01)	−0·83 (0·01)
$\log s$	−0·78 (0·03)	−0·56 (0·03)	−0·87 (0·02)	−0·96 (0·01)
$\log d$	—	−0·83 (0·02)	−0·89 (0·01)	—
$\epsilon \cdot \log(1-E_g)$	3·79 (0·24)	4·09 (0·04)	4·27 (0·04)	3·59 (0·05)
$\epsilon \cdot \log s$	−0·99 (0·30)	−2·04 (0·36)	—	1·92 (0·05)
$\epsilon \cdot \log d$	—	1·44 (0·23)	1·79 (0·04)	—
$\log(1-E_g) \cdot \log s$	—	−0·05 (0·003)	−0·03 (0·003)	−0·01 (0·001)
$\log(1-E_g) \cdot \log d$	—	0·02 (0·002)	—	—
$\log s \cdot \log d$	—	0·02 (0·007)	0·11 (0·003)	—
$\epsilon \cdot \log s \cdot \log d$	—	0·41 (0·10)	—	—

The response is the frequency of good and unique decisions. Mating schemes are independent pairs (the number of observations is 115), hierarchical scheme (440 observations), crossed scheme (479 observations) and paternity search (200 observations). The variable ranges are: $\epsilon \in [0, 0{\cdot}1]$, $1-E_g \in [8 \times 10^{-5}, 5{\cdot}7 \times 10^{-2}]$, $s \in [5, 100]$, $d \in [5, 100]$ with $s \leqslant d$. The natural logarithm is used.

(d) *Program*

For any set of polymorphic markers (number of loci, and for each locus the number and frequencies of alleles), the programme calculates the global polymorphism measures, generates N parental pairs' genotypes according to the population scheme, draws at random one pair and generates one offspring from this pair (allowing for errors). For each drawing, the N values ρ_h (5) are calculated and sorted. The largest one identifies the likeliest pair, which is compared with the true one. The frequency of correct decisions is then calculated over a large number of independent runs.

(ii) *Results*

(a) *Allowing for errors is necessary*

Since data were generated according to various typing error rates ($\epsilon_T = 0$, 1%, 5%, 10%) and likelihood calculated using a single ϵ_C, we first investigated the effect of ϵ_C on the efficiency of the decision rule by comparing the results obtained from the same simulated data. For a single value ϵ_T, the frequencies of correct final decisions were of the same order, no matter whether the right typing error rate was used or not for calculating likelihoods, as long as the zero value ($\epsilon_C = 0$) is not used. For this purpose, the Wilcoxon paired rank test was used to test the null hypothesis of zero median difference, at the 1% level. For example, in the crossed mating scheme, 120 runs were done with a true error rate equal to $\epsilon_T = 5\%$. The Wilcoxon statistics are equal to $-9{\cdot}4974$ (P value $< 0{\cdot}001$) and $-0{\cdot}4321$ (P value $= 0{\cdot}67$) comparing the frequencies of correct decisions with $\epsilon_C = 0$ and $\epsilon_C = 5\%$ on one hand, and $\epsilon_C = 1\%$ and $\epsilon_C = 5\%$ on the other hand.

(b) *Approximate expression of efficiency*

The simulation program allows the probability of a correct decision to be related to number of males, females, level of polymorphism and the error rate under the four population structures defined earlier. A total of $n = 10000$ runs are done for any combined values of the error rate ϵ_T, the numbers of alleles at the locus set, the number of males and females in a mating scheme. In order to link the true probability p (unknown) of correct decisions with the previous covariates, a Generalized Linear Model (McCullagh & Nelder, 1989) was used with binomial error and logit link, using the Splus package (Becker *et al.*, 1988):

$$\operatorname{logit}(p) = \log \frac{p}{1-p} = \eta, \tag{7}$$

where η is a linear predictor involving the available explanatory variables. The three measures of polymorphism were tested for their predictive value. The global probability of exclusion gave the best fits in terms of deviance, compared with the global PIC value and the global heterozygosity. All analyses were performed using $\log(1-E_g)$ as the measure of markers' informativity. Best predictors for the other parameters were confirmed to be ϵ_T, $\log s$, and $\log d$, as suggested by (6). All data were generated using a single ϵ_C value set to 1%.

For each mating scheme, a saturated model with interactions was fitted, and a strategy of backward elimination of variables was applied, via χ^2 deviance tests. Table 1 gives the final models. Observed frequencies of correct decisions were compared with fitted values $\hat{p} = \operatorname{logit}^{-1}(\hat{\eta})$ for each scheme, giving satisfactory results. Residuals were verified to be independent (not shown).

4. Discussion and conclusion

We have shown how it is possible to cope with errors, or mutations, in order to identify the parents of some offspring among a finite set of individuals, from the observations of genotypes at a number of loci. More specifically, we showed that taking account of possible errors is really necessary as soon as errors are expected in large-scale surveys, and we provided practical rules to choose markers, according to their polymorphic informativity (measured by the global probability of exclusion) and to some required efficiency (defined as the probability of identifying the correct parental pair).

Qualitatively, the most important result is that an efficient decision rule is obtained as soon as a non-zero error rate is allowed for, even if no precise data are available concerning this rate. In fact, a rule based on the hypothesis of no error (i.e. assuming $\epsilon_{\mathrm{C}} = 0$) yields quite high proportions of wrong decisions, even if true errors occur at a low rate, and all the more as more markers are used, i.e. in cases when a correct decision can be made if a small error rate is allowed for. For example in the search for the true parental pair among 50 independent ones, with a true error rate of 2 % and no error allowed in the likelihood calculation (i.e. $\epsilon_{\mathrm{C}} = 0$), the proportion of correct decisions is about 88 % using a system of 5 loci with 5 alleles, and 83 % using a system of 8 markers with 5 alleles. Low rates of correct decisions are mainly due to the fact that, when allowing for no error, no decision can be made as soon as no candidate parental pair is compatible with offspring genotype. The frequency of this situation increases with the exclusion probability and with the error rate, hence with the number of loci, and decreases with the number of candidates. Using a small non-zero error rate in the likelihood calculation (using ϵ_{C} as low as 10^{-3}) allows the proportion of correct decisions to increase to nearly 100 %, and it can be seen in the same examples that using such a low ϵ_{C} value for likelihood calculations is not harmful if there is no error in the typing process ($\epsilon_{\mathrm{T}} = 0$) (Table 2).

Table 2. *Proportion of correct decisions obtained with various true error rates* (ϵ_{T}), *supposed error rates* (ϵ_{C}), *numbers of parental pairs* (N), *and number of loci* (L), *with 5 equiprobable alleles per locus*

ϵ_{T}	ϵ_{C}	N	$L = 5$	$L = 8$
0	0	10	0·998	1·00
		50	0·989	1·00
	0·001	10	0·998	1·00
		50	0·989	1·00
0·02	0	10	0·889	0·818
		50	0·884	0·826
	0·001	10	0·996	0·999
		50	0·973	0·999

The quantitative results obtained by fitting the probabilities of a good decision to a statistical model, allow one to answer practical questions in the fields of applied genetics. For example, the effectiveness of two sets of markers in the turbot and trout species are quantified in Estoup *et al.* (1997) under different mating schemes. Another question arising in fish quantitative genetics is: given a typing error rate and a certain number of males and females in a crossed scheme, how can one choose a set of markers in order to obtain an average rate of good parent–offspring relationships equal to some value p_0? The previous fitted GLM can be used in this purpose, using relation (7) and Table 1.

A simple numerical illustration in a search for paternity follows: Assume a reliability of $p_0 = 0{\cdot}9$ is required, for identifying the father among $s_0 = 10$ possible fathers. Using a genotyping method with an error level of 2 % ($\epsilon_{\mathrm{T0}} = 0{\cdot}02$), the question is to determine the required informativity E_g of the set of loci to be used. From relation (7) and Table 1, we can write down the following expression:

$$\log \frac{p_0}{1-p_0} = 1{\cdot}56 - 1{\cdot}82\,\epsilon_{\mathrm{T0}} - 0{\cdot}96 \log (s_0) + 1{\cdot}92\,\epsilon_{\mathrm{T0}} \log (s_0) + [-0{\cdot}83 + 3{\cdot}59\,\epsilon_{\mathrm{T0}} - 0{\cdot}01 \log (s_0)] \log (1 - E_g).$$

Replacing known values p_0, s_0 and ϵ_{T0} by their numerical values allows the required value of informativity E_g to be derived: $E_g \simeq 0{\cdot}972$, which needs for example 2 loci with 12 equiprobable alleles.

A confidence interval can be calculated using, for example, the method of Carroll *et al.* (1988). However, it is an approximation even if the covariables are precisely known, so that this interval is to be considered with care since only an estimate of the typing error rate $\epsilon_{\mathrm{T,0}}$ can be used here. Nevertheless, this problem of calibration can be roughly answered: at the 5 % level, $E_g \in (0{\cdot}948,\ 0{\cdot}985)$.

A more general answer can be given with a plot as in Fig. 4.

In the previous derivations, we assumed a simple model for typing errors (1). A more realistic modelling of errors for microsatellite markers should assume that the rates of substitution between alleles depend on their differences in repeat number, since alleles corresponding to quite different repeat numbers are less likely to be confounded, while there is higher chance of confusion between alleles differing only by one repeat. Moreover, mis-sizing of one allele could tend to be associated with mis-sizing of the other allele, meaning a non-independence of errors between alleles within loci and individuals. Such indications might be included in the calculations of likelihoods, possibly with locus-dependent values. However, efficiency of parent identification is better when error rates are lower, so that our results can be considered

https://doi.org/10.1017/S0016672397002851 Published online by Cambridge University Press

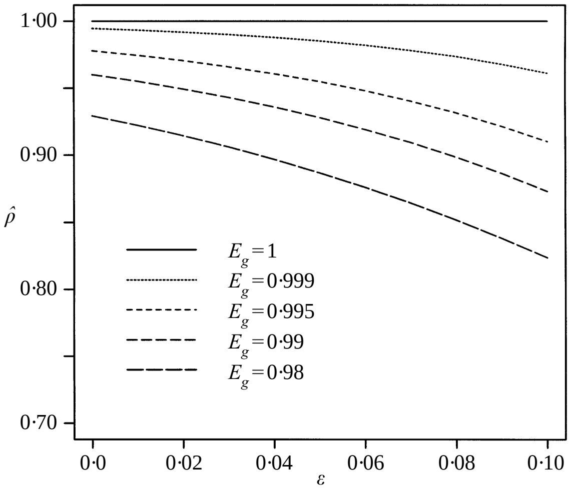

Fig. 4. Relationship between the probability of a correct decision and rate of typing error, for five sets of markers of differing exclusionary power, in a paternity analysis with 10 potential sires. Fitted values are equal to $\hat{p} = \mathrm{logit}^{-1}\hat{\eta} = \exp\hat{\eta}/(1+\exp\hat{\eta})$ where $\hat{\eta}$ is the estimated linear predictor (Table 1).

to be conservative if one considers the error rate used as the highest error rate over all possible allelic misinterpretations. Actual error rates are quite low. Lathrop *et al.* (1983) estimated the overall pedigree error and the overall laboratory error for a human population of a South Pacific island. Errors on father–child (or mother–child) relationships represent 4 % (or 0), while the laboratory error rate was 1 %. In a large-scale pig experiment in which typing made use of automated DNA analysers, selected markers showed an overall error rate of 0·5–1 % (D. Milan and N. Woloszyn, personal communication). These low error rates are minimal values, since they are estimated from segregation analysis performed in situations for which pedigree information is available. Although small, such values lead to quite a high proportion of individuals exhibiting incorrect multilocus genotypes, and hence the requirement that errors be systematically taken into account: for example, setting ϵ_T to 2 % means that the frequency of individuals with exact genotypes of L loci, equal to $(1-\epsilon_T)^{2L}$, is only about 82 % for 5 loci, and 72 % for 8 loci.

Another limitation of the model may be not taking account of null alleles, if present. Non-amplification of some microsatellite alleles occurs quite often and may be important in the present situation without family information. Then any *phenotype* $[A_i A_i]$ (denoted with two allele symbols as a homozygote) may reflect either a true homozygote with two (A_i) alleles or the genotype $(A_i A_0)$ with a null allele (A_0). In such cases, likelihoods involving parent or offspring homozygote phenotypes should be modified. In particular, non-compatible genotypes (implying an error or a mutation in a co-dominant setting) may become admissible. For example the conditional probability of the second locus genotype in Example 1 (3), should allow for the possibility that the offspring be of genotype $(B_1 B_0)$ and the second parent be of genotype $(B_2 B_0)$. Calculation of likelihoods then needs knowledge of allele frequencies: denoting by p_0 and p_2 the frequencies of alleles B_0 and B_2, (3) would be replaced by a value approximately equal to

$$\frac{1}{2}\frac{p_0}{p_2+2p_0}$$

instead of

$$\tfrac{1}{2}\epsilon.$$

Hence, taking account of such a situation in the framework of our model requires that high error rates are allowed for, up to the order of magnitude of a null allele frequency. It should also be stressed that exclusion probabilities should be computed in a different way, since null alleles make them lower (Chakravarti & Li, 1983). Adapting the calculation of likelihood to the occurrence of null alleles is possible, and some trials have indicated that it is a way to recover high efficiency of paternity identification. This may be useful when working with a specific set of data for which the choice of markers is limited, with evidence of a single null allele. However, the model chosen here for handling null alleles is not general. We consider a null allele as corresponding to a mutation in the primer sequence, allowing no identifiable amplification. Other circumstances may lead to a complex pattern of dominance and recessivity, with weak alleles being observable if associated with an allele of quite different length, but not observable when amplified together with a strong allele of nearly equal length. It does not seem worth developing such marker-dependent models for microsatellites, as for special loci (e.g. blood group systems), since searching for new primers or choosing another microsatellite marker are relatively straightforward tasks.

It has been assumed that the parental genotypes are known exactly. This should be the case in experimental designs, where parental genotypes can be ascertained from a sample of their progeny, but not necessarily true in population surveys or wild-life studies. A similar probability analysis can be carried out to account for such errors. The difference is that the whole progeny of a mistyped parent is expected to be difficult to recognize as descending from one of the proposed parental pairs, since one error must be systematically assumed. To overcome the problem, e.g. to identify the mistyped parent, it is necessary that the likelihood for this parent be larger than the likelihood for a random parent. The first one, involving the estimated error rate ϵ_C for one allele, is of the order of

$$\epsilon_C\,(\tfrac{1}{2})^{2L-1}$$

while the second one is of the order of

$$(1-E_g)\,(\tfrac{1}{2})^{2L},$$

https://doi.org/10.1017/S0016672397002851 Published online by Cambridge University Press

where E_g is the global probability of exclusion. Therefore, using an ϵ_C value one order of magnitude larger than $(1-E_g)$ can account for one error among parent genotypes; noting, however, that further errors in offspring typing would probably lead to an erroneous decision.

The main qualitative conclusion of this work is that typing errors *must* be considered in the model, even if they rarely occur. If typing errors are ignored, wrong conclusions are often drawn concerning the detection of the true parental pair. This situation can only be avoided by taking a non-zero error rate into account in the model. More multiallelic loci need to be typed to overcome the lack of information due to typing errors, and biallelic markers must be avoided. Typing errors are also worth considering in other fields. In linkage analysis, numerous markers are used to type numerous individuals, and it is usual that typing errors, when they occur, do perturb genetic mapping (Simianer & Wild, 1995). Large-scale genotyping is also undertaken in surveys of genetic diversity of populations (Deka *et al.*, 1995; van Zeveren *et al.*, 1995; Moazami-Goudarzi *et al.*, 1997), and it might be interesting to check the incidence of typing errors on classifications and on estimations of genetic distances.

Appendix. Explicit derivation of the size and power of the Likelihood Ratio test

The distributions of the likelihood ratio

$$LR = \frac{P(\Pi\,|\,H_1)}{P(\Pi\,|\,H_0)}$$

can be written down explicitly, under both hypotheses, $H_0: \theta = t_{k'}$ and $H_1: \theta = t_k$, according to (2). The Neyman–Pearson lemma is applied, and allows one to build the most powerful test of given size α. One defines the test, ϕ, as:

$$\phi(\Pi) = \begin{cases} 1 \quad \text{if} \quad \dfrac{P(\Pi\,|\,H_1)}{P(\Pi\,|\,H_0)} > c_\alpha \\[2ex] \gamma(\Pi) \quad \text{if} \quad \dfrac{P(\Pi\,|\,H_1)}{P(\Pi\,|\,H_0)} = c_\alpha \\[2ex] 0 \quad \text{if} \quad \dfrac{P(\Pi\,|\,H_1)}{P(\Pi\,|\,H_0)} < c_\alpha, \end{cases}$$

where $c_\alpha > 0, 0 \leqslant \gamma(\Pi) \leqslant 1$.

The first type error can be explicitly calculated:

$$\begin{aligned} \alpha &= E_{H_0}(\phi) \\ &= P_{H_0}\left(\frac{P(\Pi\,|\,H_1)}{P(\Pi\,|\,H_0)} > c_\alpha\right) + \gamma P_{H_0}\left(\frac{P(\Pi\,|\,H_1)}{P(\Pi\,|\,H_{0)}} = c_\alpha\right) \\ &= \sum_\Pi P(\Pi\,|\,H_0)\left[I\left\{\frac{P(\Pi\,|\,H_1)}{P(\Pi\,|\,H_0)} > c_\alpha\right\}\right. \\ &\quad \left. + \gamma I\left\{\frac{P(\Pi\,|\,H_1)}{P(\Pi\,|\,H_0)} = c_\alpha\right\}\right], \end{aligned} \tag{A 1}$$

where $I(E)$ is an indicator taking values 1 or 0 depending on the statement E being true or false, and $\gamma \in [0,1]$. The power is

$$\begin{aligned} 1-\beta &= E_{H_1}(\phi) \\ &= \sum_\Pi P(\Pi\,|\,H_1)\left[I\left\{\frac{P(\Pi\,|\,H_1)}{P(\Pi\,|\,H_0)} > c_\alpha\right\}\right. \\ &\quad \left. + \gamma I\left\{\frac{P(\Pi\,|\,H_1)}{P(\Pi\,|\,H_0)} = c_\alpha\right\}\right]. \end{aligned} \tag{A 2}$$

The values c_α and γ are chosen with the following procedure:

(i) if there exists c such that

$$P_{H_0}\left(\frac{P(\Pi\,|\,H_1)}{P(\Pi\,|\,H_0)} > c\right) = \alpha \text{ then } c_\alpha = c \text{ and } \gamma = 0;$$

(ii) if there exists c such that

$$P_{H_0}\left(\frac{P(\Pi\,|\,H_1)}{P(\Pi\,|\,H_0)} > c\right) < \alpha \leqslant P_{H_0}\left(\frac{P(\Pi\,|\,H_1)}{P(\Pi\,|\,H_0)} \geqslant c\right)$$

then $c_\alpha = c$ and

$$\gamma = \frac{\alpha - P_{H_0}\left(\dfrac{P(\Pi\,|\,H_1)}{P(\Pi\,|\,H_0)} > c_\alpha\right)}{P_{H_0}\left(\dfrac{P(\Pi\,|\,H_1)}{P(\Pi\,|\,H_0)} = c_\alpha\right)}.$$

Example. Explicit derivation of the size and power are derived for the simple example assuming that parental pairs derive from homozygous inbred lines. The special case of one locus with a alleles is developed here. Let ϵ_1 denote $\epsilon/(a-1)$. The hypotheses are:

$H_0: \theta = t_{k'}$ with $G^{sd}(k') = (A_1A_1, A_2A_2)$ denoted 11, 22

$H_1: \theta = t_k$ with $G^{sd}(k) = (A_1A_1, A_1A_1)$ denoted 11, 11

Table A 1 gives the probabilities that offspring x is of phenotype Π under the null H_0 and the alternative H_1.

Table A 1

Π	11	$1y(y \neq 1,2)$	$yy(y \neq 1,2)$	$yz(y \neq 1,2;\ z \neq 1,2;\ y \neq z)$	12	$2y(y \neq 1,2)$	22
H_0	$\epsilon_1(1-\epsilon)$	$\epsilon_1(1-\epsilon)+\epsilon_1^2$	ϵ_1^2	$2\epsilon_1^2$	$(1-\epsilon)^2+\epsilon_1^2$	$\epsilon_1(1-\epsilon)+\epsilon_1^2$	$\epsilon_1(1-\epsilon)$
H_1	$(1-\epsilon)^2$	$2\epsilon_1(1-\epsilon)$	ϵ_1^2	$2\epsilon_1^2$	$2\epsilon_1(1-\epsilon)$	$2\epsilon_1^2$	ϵ_1^2

https://doi.org/10.1017/S0016672397002851 Published online by Cambridge University Press

Let $J(\Pi)$ be defined as

$$J(\Pi) = I\left\{\frac{P(\Pi \mid H_1)}{P(\Pi \mid H_0)} > c\right\} + \gamma I\left\{\frac{P(\Pi \mid H_1)}{P(\Pi \mid H_0)} = c\right\}.$$

For example,

$$J(\Pi = 11) = I\left\{\frac{1-\epsilon}{\epsilon_1} > c\right\} + \gamma I\left\{\frac{1-\epsilon}{\epsilon_1} = c\right\}.$$

Equation (8) leads to:

$$\begin{aligned}\alpha = {} & \epsilon_1(1-\epsilon)\,J(\Pi = 11) + (a-2)\,\epsilon_1(1-\epsilon+\epsilon_1)\,J(\Pi = 1y)\\ & + (a-2)\,\epsilon_1^2\,J(\Pi = yy) + (a-2)\,(a-3)\,\epsilon_1^2\,J(\Pi = yz)\\ & + ((1-\epsilon)^2 + \epsilon_1^2)\,J(\Pi = 12)\\ & + (a-2)\,\epsilon_1(1-\epsilon+\epsilon_1)\,J(\Pi = 2y)\\ & + \epsilon_1(1-\epsilon)\,J(\Pi = 22)\end{aligned}$$

and (9) to:

$$\begin{aligned}1-\beta = {} & (1-\epsilon)^2 J(\Pi = 11) + 2(a-2)\,\epsilon_1(1-\epsilon)\,J(\Pi = 1y)\\ & + (a-2)\,\epsilon_1^2\,J(\Pi = yy)\\ & + (a-2)\,(a-3)\,\epsilon_1\,J(\Pi = yz)\\ & + 2\epsilon_1(1-\epsilon)\,J(\Pi = 12) + 2(a-2)\,\epsilon_1^2\,J(\Pi = 2y)\\ & + \epsilon_1^2\,J(\Pi = 22).\end{aligned}$$

The Bayesian test corresponds to $c = 1$.

When L loci are considered, all the combinations of the possible phenotypes $\Pi_1, \ldots, \Pi_l, \ldots, \Pi_L$ have to be written down. Only the first terms of the development in ϵ can be explicitly calculated. Straightforward numerical programming was used to construct Fig. 2.

Thanks are due to René Guyomard, Arnaud Estoup and Karim Gharbi for stimulating discussions and access to actual data, and to Philippe Mulsant for reading the manuscript.

References

Becker, R. A., Chambers, J. M. & Wilks, A. R. (1988). *The New S Language*. Pacific Grove, California: Wadsworth and Brooks/Cole.

Carroll, R. J., Spiegelman, C. H. & Sacks, J. (1988). A quick and easy multiple-use calibration curve procedure. *Technometrics* **30**, 137–141.

Chakravarti, A. & Li, C. C. (1983). The probability of exclusion based on the HLA locos. *American Journal of Human Genetics* **35**, 1048–1052.

Chastang, C. (1973). *Contribution à l'aide au diagnostic: analyse d'un problème de parenté*. Thèse pour le doctorat en médecine, University of Clermont, France.

Deka, R., Jin, L., Shriver, M. D., Yu, L. M., DeCroo, S., Hundrieser, J., Bunker, C. H., Ferrell, R. E. & Chakraborty, R. (1995). Population genetics of dinucleotide (dC-dA)n.(dG-dT)n polymorphisms in world populations. *American Journal of Human Genetics* **56**, 461–474.

Elston, R. C. (1986). Probability and paternity testing. *American Journal of Human Genetics* **39**, 112–122.

Estoup, A., Gharbi, K., SanCristobal, M., Chevalet, C., Haffray, P. & Guyomard, R. (1997). Parentage assignment using microsatellites in turbot (*Scrophtalmus maximus*) and rainbow trout (*Oncorhynchus mykiss*) hatchery populations. *Canadian Journal of Fisheries and Aquatic Sciences* (in the Press).

Ferguson, T. S. (1967). *Mathematical Statistics: A Decision Theoretic Approach*. New York: Academic Press.

Hanset, R. (1975). Probabilité d'exclusion de paternité et de monozygotie, probabilité de similitude. Généralisation à N alleles co-dominants. *Annales de Médecine Vétérinaire* **119**, 71–80.

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

Lathrop, G. M., Hooper, A. B., Huntsman, J. W. & Ward, R. H. (1983). Evaluating pedigree data. I. The estimation of pedigree error in the presence of marker mistyping. *American Journal of Human Genetics* **35**, 241–262.

Moazami-Goudarzi, K., Laloë, D., Furet, J. P. & Grosclaude, F. (1997). A study of the impact of additional microsatellite data on analysis of genetic relationships between 10 cattle breeds. *Animal Genetics* (in the Press).

McCullagh, P. & Nelder, J. A. (1989). *Generalised Linear Models*, 2nd edn. London: Chapman and Hall.

Simianer, H. & Wild, V. (1995). Effect and treatment of misclassified samples in linkage analysis. *Journal of Animal Breeding and Genetics* **112**, 243–254.

Smouse, P. E. & Chakraborty, R. (1986). The use of restricted fragment length polymorphisms in paternity analysis. *American Journal of Human Genetics* **38**, 918–939.

Weber, J. (1990). Informativeness of human (dC-dA)n(dG-dT)n polymorphisms. *Genomics* **7**, 524–530.

Weber, J. & May, P. (1989). Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *American Journal of Human Genetics* **44**, 388–396.

van Zeveren, A., Peelman, L., Van de Weghe, A. & Bouquet, Y. (1995). A genetic study of four Belgian pig populations by means of seven microsatellite loci. *Journal of Animal Breeding and Genetics* **112**, 191–204.