A review on the methods of parentage and inbreeding analysis with molecular markers

Horse commission, Session 3, invited paper 3

EAAP – BLED – Slovenia, 5- 8 September 2004

Bertrand Langlois

INRA-CRJ-SGQA, 78352 Jouy-en-Josas – France
E-mail : bertrand.langlois@dga.jouy.inra.fr

Abstract:

In horse populations there is a great concern for pedigree. Genetic markers are commonly used for exclusion procedures to assess the right sire and dam of the foal. However pedigree information is limited because the total genetic history of an animal or a population can not be traced from the beginning. In this paper we try to review how genetic markers can help us to overcome these difficulties. Formulae in the literature for estimating F from the state of markers consider the two causes that make sorting two genes alike. They are either identical by descent or alike in state. All authors agree that estimators for pairwise relatedness or individual inbreeding coefficients need a lot of independent co-dominant marker loci where alleles are balanced in frequencies in order to reach a minimum accuracy in estimations. In this perspective the development of a kit of SNP satisfying these conditions would be a tool of great interest to address the problems connected with parentage inbreeding and genetic diversity in horse populations where the good management of pedigree information appears insufficient to do it properly.

Key words: Genetic markers, parentage, inbreeding, Horse

Introduction

In horse populations there is a great concern for pedigree. Most of stud-books started during the 19th century and some of them even earlier. This administrative work was done very carefully as it was done for humans with parish register. However this did not exclude some errors which justified the use of genetic markers in routine procedure as early as the 1970s. Now genetic markers are systematically used for breeds such as arab, thoroughbred and trotter and commonly used for the others in the case of artificial insemination or at random to discourage fraud. The result is a very low percentage of parentage errors in horse breeding. Genetic markers in horse breeding are only used for exclusion procedures to assess the right sire and dam of the foal. Categorical allocation to select the most likely parent from a foal of non-excluded parents is not practised for legal reasons. However pedigree information is limited because the total genetic history of an animal or a population can not be traced from the beginning. Even with very complete and reliable pedigrees, there are still events in the past which are not described like bottlenecks or real number of unrelated founders. In this paper we try to review how genetic markers can help us to overcome these difficulties.

Exclusion

The earliest conceptually simplest technique of parentage analysis is exclusion. This technique based on Mendelian rules of inheritance uses incompatibilities between parents and offspring to reject particular parent-offspring hypotheses. It was used a low scale for a long time in horse breeding: two chestnut (ee) parents are expected to have only chestnut foals (ee). One grey foal (G-)

is expected to have at least one grey parent (G-)$^{-}$ which can be generalised for each recessive allele (black E-/aa) or dominant one (bai E-/A-).

More generally a foal can not receive an allele not present in his parents. One sees immediately that co-dominant loci where genotype of foal and parents will appear will be more efficient that dominant/recessive ones for that technique.

Exclusion is an appealing approach because exclusions of all but one parent pair from a complete sample of all possible parents for each offspring in a population could be considered the paragon of parentage analysis. However it is limited by the occurrence of typing errors and at a lower rate of mutations. The list of markers used for exclusion also plays a major role.Jamieson and Taylor (1997), Dodds et al.(1996) made a thorough analysis of these questions. It is concluded that the exclusion probability increases with the number of loci that can be used as genetic markers, with the number of alleles at each locus and with the evenness of the allele frequency at each locus (Chakraborty et al.1974, Selvin (1980),Ryman and Chakraborty (1982), Smouse and Chakraborty (1986). However the relationship exhibits a diminishing marginal return because an additional marker applies its power of exclusion only on non-excluded parents before its application.

This can be generalised to other types of exclusion. For an autosomal marker first, it is always easier to detect incorrect offspring assignments (i.e. when mating pairs are known) than other types of exclusion and second, paternity (or maternity) exclusion is greater with the other parent known than unknown as expected.

For daughters the X- linked markers always has a greater exclusions probability no matter what situation is being tested except for maternity testing without knowledge of the sire where the autosomal and X- linked markers have the same exclusion probability). For sons the autosomal marker has higher exclusion probability except for maternity testing in which case, the X linked marker would be better.

The importance of exclusion probability to paternity assignment is that an increase in the exclusion probability increases the probability of paternity among the set of non-excluded parents. Clearly the likelihood of choosing the correct non-excluded parent increase. In the extreme case, as the exclusion probability approaches 1, most progeny can be assigned exclusively to a single male or female parent in the population.

It may also be desirable to require exclusion at more than one locus to reduce the effect of possible genotyping errors, mutation or of unknown null alleles.

Table 1 shows the exclusion probability of eleven microsatellites markers routinely used for parentage control in horse breeding in France. The exclusion probability for thoroughbred and Arab breeds is now reaching near one value (Amigues et al. 2000). With older systems (9, haemolytic, 24 aglutination and 10 electrophoretic) it was only of 0.952 for thoroughbred and 0.954 for arab. These now very high exclusion probabilities were recently confirmed by Cho and Cho (2004) for Korean native horses.

As it can be inferred from the studies in horses 10 to 20 polymorphic loci allow probabilities of exclusion close to 1. The paragon of parentage analysis is therefore reached for horse populations. For each foal the right sire and dam can be assigned.

Table 1 – Exclusion probabilities in two breeds for routine microsatellites markers used in France. (source: Amigues et al. 2000)

Microsatellites			Thoroughbred		Arab	
Name	Origin	Reference	Alleles number	Exclusion Probability	Alleles number	Exclusion probability
AHT4	U.K.	Binns et al. 1995	6	0.49	7	0.57
AHT5	U.K.	Binns et al. 1995	6	0.51	6	0.45
ASB2	Australia	Breen et al. 1997	8	0.68	8	0.37
HMS1	France	Guérin et al. 1994	4	0.35	6	0.36
HMS3	France	Guérin et al. 1994	6	0.35	6	0.46
HMS6	France	Guérin et al. 1994	7	0.32	6	0.46
HMS7	France	Guérin et al. 1994	5	0.58	7	0.53
HTG4	Sweden	Ellegren et al. 1992	5	0.25	6	0.41
HTG6	Sweden	Ellegren et al. 1992	7	0.33	7	0.37
HTG10	Sweden	Marklund et al. 1994	7	0.54	8	0.53
VHL20	The Netherlands	Van Haeringen et al. 1994	7	0.50	10	0.62
	Total Probability of identity		68	0.9989	77	0.9991
			$4.6\ 10^{-10}$		$1.8\ 10^{-10}$	

Allocation or assignment

If complete exclusion is not possible it is often not sufficient to derive accurate population statistics on mating patterns. Consequently techniques were developed that assigned progeny to non-excluded parents based on likelihood scores derived from their genotypes. According to Jones and Ardren (2003) these techniques assign offspring either categorically or fractionally.

Categorical allocation uses likelihood-based approaches (Meagher and Thompson, 1987) to select the most likely parent from a pool of non-excluded parents. This method involves calculating the logarithm of the likelihood ratio (LOD score) by dividing the likelihood of an individual (or pair of individuals being the parent (or parents) of a given offspring by the likelihood of these individuals being unrelated. After an exhaustive evaluation of all possible parents, the offspring are assigned to the parent (or parental pair) with the highest LOD score. When all-parent offspring relationships show zero likelihood, offspring are unassigned. Parentage remains also ambiguous when multiple parent-offspring relationships obtain high no zero likelihood. Contrary to strict exclusion methods, likelihood-based allocation method, because it is based on the evaluation of a probability, allows for some degree of transmission errors due to misreading or mutation (SanCristobal and Chevalet 1997).

It is also for this reason that allocation techniques are not acepted in horse breeding. For forensic purpose you need to establish true facts and not only their probabilities. However limiting yourself to true facts limits the amount of information used. Allocations techniques remain therefore appealing because they allow a better use of the available information in statistical terms.

Returning to Meagher's and Thompson's (1987) original proposition for categorical allocation. In all cases we examine genotypes , g_A g_B and g_O at a single autosomal locus for three individuals (O, B and A). Assuming unlinked loci, information from multiple loci can be combined by summing the LOD scores over all loci. Transition probabilities (T) for use of the following equations can be found in Marshall et al. (1998)for co-dominant markers and in Gerber et al. (2000) for dominant markers. Three main cases have to be examined:

a) Identifying one parent when the other is known. Letting B represent the known parent and A the alleged parent, the LOD score for A being the parent of O is:

$$\text{LOD score (A parent of O)} = \text{Log}_e \frac{T(g_O \mid g_B, g_A)}{T(g_O \mid g_B)}$$

Where $T(g_O \mid g_B, g_A)$ is the transition probability of g_O given g_B and g_A and $T(g_O \mid g_B)$ is the transition probability of g_O given g_B.

b) Identifying one parent with no information about the other parent. In this case, no information is available concerning parentage of O. The single parent LOD score for B being the parent of O is:

$$\text{LOD score (B parent of O)} = \text{Log}_e \frac{T(g_O \mid g_B)}{P(g_O)}$$

Where $P(g_O)$ is the frequency of the offspring's genotype in the population.

c) Identifying a parental pair starting with no prior information. Parental pair allocation is an approach for identifying parent-offspring relationships by constructing genotypic triplets consisting of a proposed offspring and proposed maternal and paternal parents. This procedure involves calculating a breeding likelihood, which is defined as the likelihood of a parental pair producing the multi locus genotype found in the offspring being examined. The breeding likelihood of a given offspring on the basis of a single locus is:

$$\text{LOD score (A, B parent of O)} = \text{Log}_e \frac{T(g_O \mid g_A, g_B)}{P(g_O)}$$

The fractional allocation method assigns some function, between 0 and 1, for each offspring to all non-excluded candidate parents. The proportion of an offspring allocated to a particular candidate parent is proportional to its likelihood of parenting the offspring compared to all other non-excluded candidate parents. Single parent and parent pair likelihoods are calculated in the same way as in the categorical allocation method (Devlin et al. 1988). Because the fractional technique splits an offspring among all compatible males it is guaranteed to be incorrect from a biological standpoint, an offspring having only one father and one mother. However for the study of particular problems connected with reproductive success in natural populations this method proved his statistical superiority. Indeed, the categorical allocation by the most likely method as formulated above embodies some bias in that the most likely parent will always be that individual in the population that has the highest number of loci homozygous for the necessary paternal gamete contribution that

complements the maternal ones. It was also emphasised by Thompson and Meagher (1987) that bilateral relatives such as full sibs may be more likely parents than the true parent individuals.

We will also see further that maximum likelihood techniques are asymptotically optimal but can prove to be very inaccurate for a low number of markers.

First conclusion on exclusion and allocation

For horse breeding we are now in a situation where the exclusion probability of the microsatellites markers routinely in use is close to one for all breeds. Therefore the allocation techniques decrease in interest at least to identify the first generation parents (Sire and Dam) to certify the pedigree. To ascertain sire and dam of an offspring when done, over several generations, makes the information of pedigree very reliable.

However this is not sufficient to ascertain exact genetic relationships between individuals of a population when some errors in the past (Kavar et al. 2002) and when the assumption of unrelated founders (ancestors without known parents) can not be accepted as it is mainly the case in horse populations (Mahon and Cunningham 1982, MacCluer et al. 1983, Cothran et al. 1984, Moureaux et al. 1996, Cunningham et al. 2001, Zechner et al. 2002.).

We can also remark that a sire or a dam transmit half of his alleles to his offspring with certainty, this is only the case in probability for other relationships as shown Table 2. One can want to check the realisation of this probability with genetic markers particularly when panmixia is not realised in the case of inbreeding selection and homogamy.

Table 2 – Cotterham's K values for some standard genealogical relationships, in the absence of inbreeding.
Source: Thompson (1975)

Relationship of A to B	K_0	K_1	K_2
Unrelated	1	0	0
Offspring, parent	0	1	0
Sib	1/4	1/2	1/4
Identical twin	0	0	1
Niece, nephew, uncle, aunt Grandparent, grand-child Half-sib	1/2	1/2	0
First-cousin Parent's half-sib, half-sib's child	3/4	1/4	0
Double first cousin	9/16	6/16	1/16
Half-sibs whose non-identical parents are:			
1- sibs or parent-offspring	3/8	1/2	1/8
2- half-sibs	7/16	1/2	1/16

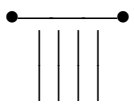
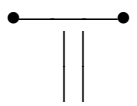
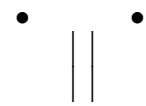
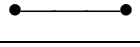
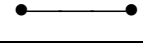
K_0, K_1, K_2 being the probability of 0, 1 and 2 genes in common

Other relationships have $K_1 \leq 1/2$ and $K_2 \leq 1/16$

Short history of the description of genetic pairwise relationships.

Cotterman 1940 first introduced the k coefficients probability that two non-inbred individuals have 0, 1 or 2 genes in common. These are sufficient specification of the relationship between any two non-inbred individual. Malécot 1948 extended this work introducing the parentage coefficient between two individuals. This is the probability of drawing two genes identical by descent in each individual. Consequently the inbreeding coefficient of an individual, the probability that the two genes of a same locus are identical by descent is the parentage coefficient of his parents. This allowed a more thorough use of pedigree information. Wright (1943) defined the kinship coefficient r, as the correlation between uniting gametes. That is two times Malécot's parentage coefficient The last step of description of pairwise relationships was given by Jacquard (1972) for two individuals at one locus nine situations of identity were distinguished according to Figure 1.

Figure 1 – Scheme of the nine situations of identity according to Jacquard 1972

N°	1	3	5	7	9
Individual A					
Individual B					

N°	2	4	6	8
Individual A				
Individual B				

According to these nine situations a probability Δ_i is given according to the pedigree and we have the following relations with inbreeding coefficient f and the parentage coefficient φ.

$$f_A = \Delta_1 + \Delta_2 + \Delta_3 + \Delta_4$$
$$f_B = \Delta_1 + \Delta_2 + \Delta_5 + \Delta_6$$
$$\varphi_{AB} = \Delta_1 + \frac{1}{2}(\Delta_3 + \Delta_5 + \Delta_7) + \frac{1}{4}\Delta_8$$

The correspondence with Cotterman's coefficients is:
k_0 in situations 2 4 6 and 9 (0 gene in common for A and B)
k_1 in situations 8 3 5 (1 gene in common)
k_2 in situations 1 and 7 (2 genes in common)

How to infer the estimation of f or φ .from the situation of genetic markers identity.

We must go back to Malécot (1948) to define the two concepts that makes two alleles at the same locus alike. They are either "identical by descent" (IBD) or "alike in state" (AIS). He wrote

therefore the probability s_{ii} of being homozygote for allele i equals the probability of being IBD defined as the inbreeding coefficient multiplied by the probability of drawing the i allele, plus the probability (1-f) of not being IBD multiplied by the probability of drawing at random twice the same allele (probability of being AIS):

$$s_{ii} = fp_i + (1-f)p_i^2 = fp_i + (1-p_i) + p_i^2$$

The probability s_{ij} of being heterozygote for alleles i and j equals the probability of not being IBD multiplied by the probability of drawing at random i and j or j and i.

$$S_{ij} = (1\text{-}f)2p_ip_j$$

Where p_i is the frequency of allele i. The probability of being homozygous at the locus is derived:

$$\sum_i s_{ii} = f + (1-f)\sum_i p_i^2 \qquad (1)$$

or $$\sum_i s_{ii} = f(1-\sum_i p_i^2) + \sum_i p_i^2$$

And the probability of being heterozygous:

$$\sum_{i\neq j} s_{ij} = 1 - \sum_i s_{ii}$$

$$= 1 - \sum_i p_i^2 - f(1-\sum_i p_i^2)$$

$$\left(1-\sum_i s_{ii}\right) = (1-f)\left(1-\sum_i p_i^2\right)$$

From the knowledge of p_i and the observed s_{ii} or s_{ij} f can be estimated.

This formula is presented many times in the literature under different forms considering f or parentage coefficients φ of the parents, multi- or bi-allelism. Let us cite Wright (1978)analogous formulae for subdivided population discussed by Malécot (1969):

$$(1-F_{IT}) = (1-F_{IS})(1-F_{ST})$$

where according to Eding (2002) or Robertson and Hill (1984) F_{IT} is defined as the total kinship between two individuals within the whole subdivided population. F_{IS} is the kinship between two individuals within a subpopulation and can be extracted from the (limited) pedigree information, $F_{IS} = f$. F_{ST} is the correlation between random gametes from the same sub population relative to the whole population: $F_{ST} = 1 - H_s / H_t$

H_s is the heterozygosity intra subpopulation
H_t is the heterozygosity for the whole population

.

Let us also cite Lynch (1988) formula:

$$f = \frac{\sum_i s_{ii} - \sum_i p_i^2}{1 - \sum_i p_i^2}$$

(see equation (1) probability of being homozygous)

All techniques of estimating pairwise relationships from the state of molecular markers derive from this approach. They can be applied at the individual level or at the intra- and between- population levels. However, one can immediately anticipate how low will be the accuracy of the one locus approach. We are therefore inclined to propose a multi-locus approach.

Multi-locus approach

For the choice of an adequate and efficient set of markers the weighting of information of each locus plays a major role for determining best statistical estimators (Ritland 1996) and also for the choice of efficient and adequate markers. Indeed, one can easily understand that fixed loci will not give any information on parentage and that at the opposite loci with more balanced frequencies will do it.

From a statistical standpoint (Ritland 1996) one problem in estimating relatedness or inbreeding for individual is statistical bias caused by small samples. In considering sample size, there are two dimensions the number of individuals, and the number of marker loci. The number of individuals if considered alone is at a bare minimum of one for inbreeding, two for relatedness, magnifying the bias due to small samples, even when a large number of marker loci are used. This can be a significant problem when using maximum likelihood estimators which are often recognized to show bias with small sample sizes. However individuals or mating pairs are not isolated they belong to a population or a sub-population. Indeed, the parentage coefficient of two individuals and therefore the inbreeding of an individual have no meaning per se. The concept takes sense only when it is related to a population or a subpopulation constituting the gene pool. This question about the number of individuals can therefore be partly translated on the problem of estimating allele frequencies in these populations. It will give the basis for the genotypic probabilities of randomly chosen animals. Deviations from this basis will serve for parentage analysis. Without reference there is no measurement possible.

Method of Moment Estimator (MME)

Generality

Its primary advantage is the reduction of bias with individual level estimates and a lack of distributional assumptions.

To describe the data we denote S_i as the observed proportion of pairs similar for marker allele i. It can be regarded as an indicator variable of relationship. For the case of inbreeding coefficient f, then $S_i = 1$ if the two alleles at a locus are allele i ; otherwise $S_i = 0$. For the case of two-individual relatedness φ, there are four equally probable ways of sampling two alleles, two for each two relatives. Si is the average over the four ways that a given pair of alleles can be sampled.

For individual A: $S_i = 1$ for the situation 1, 2, 3 and 4.of Jacquard (1972) $S_i = 0$ for the others.

For individual B: $S_i = 1$ for the situations 1, 2, 5 and 6. $S_i = 0$ for the others.

For the pair of individual A and B we have defined $S_i = \frac{1}{4}(I_{11} + I_{12} + I_{21} + I_{22})$. Therefore:

$S_i = 1$ for the situation 1,

$S_i = 1/2$ for the situations 3, 5 and 7

$S_i = 1/4$ for the situation 8,

$S_i = 0$ for the situations 2, 4, 6 and 9.

The expectation of S_i (denoted s_i) conditioned upon relationship, is as we have seen:

$$s_i = \rho p_i + (1-\rho)p_i^2, \qquad (1bis)$$

Where ρ is the two-gene relationship, which equals either f or φ. This expectation assumes the population gene frequencies equal the pedigree gene frequencies (the gene pool from which alleles were randomly drawn during the formation of the pedigree. The probabilities over several independent loci are the product of these single-locus probabilities.

Correlation method

To obtain an efficient method of moments estimator (MME) for two-gene relationship, one first obtains estimates for each marker allele i, for i=1 to n (the number of alleles at the locus), based upon the observation of whether the alleles are both of type i or not. Although there are n(n+1)/2 combinations of alleles each of which can give an estimate of relationship, these estimates are not independent, and only the set of n estimates corresponding to the sharing of allele i, i=1, n, are sufficient to capture all information in the data (Robertson and Hill, 1984). The variance-covariance matrix of these n estimates is then used to optimally combine the n estimates in a linear fashion into a single estimate.

By equating observed quantities to their expectations in (1bis), we obtain an estimator for each allele i at an n-allele locus as:

$$\hat{\rho}_i = \frac{S_i - P_i^2}{P_i Q_i}, i = 1,....,n \qquad (2)$$

where $P_i = 1 - Q_i$ is the estimate of gene frequency p_i (capital letters are used to denote estimated quantities), and the hat denotes the estimate. For simplicity, gene frequency can be estimated by collecting alleles in the entire sampled population (this assumes low mean relationship).

The total estimate of relationship (relatedness or inbreeding is then the weighted average:

$$\hat{\rho} = \sum_i w_i \hat{\rho}_i \qquad (3)$$

Where the weights w_i sum to unity.

To obtain the optimal weights, note that the n estimates of relationship (2) have variances and covariances:

$$\text{var}(\hat{\rho}_i) = \frac{s_i(1-s_i)}{cp_i^2q_i^2}$$

$$\text{Cov}(\hat{\rho}_i\,\hat{\rho}_j) = \frac{-s_is_j}{cp_ip_jq_iq_j},\ i,\ j = 1,\ 2,\ldots,\ n$$

These are obtained by noting that the S_i are multinomially distributed with variances $s_i(1-s_i)$ and covariances - s_is_j, and that $\text{Var}(aX+b) = a^2\text{Var}(x)$ for a and b constant. The constant c=1 for f while $C \le 4$ for φ; its exact value is irrelevant because it cancels when computing weights.

The optimal weights are then found via a standard procedure of weighting correlated estimates.

Briefly these weights minimize $\text{Var}(\hat{\rho}) = w^T V w$ where w is an n element column vector of weights and V is the variance – covariance matrix of allele- specific estimates.

Unless one assumes a prior of $\rho = 0$ or $\rho = 1$ the expression of w must be solved numerically.

Then multi-locus estimates of relatedness involve a second stage of weighting. After a weighted estimate is found for each locus, a "grand" weighted estimate is found by weighting estimates across loci. If loci are unlinked and in linkage equilibrium, estimates from different loci will be independent and the weighting used for a given locus is simply proportional to the inverse of its variance as computed by the above weighting procedure.

A simple simplified MME estimator can be obtained by assuming $\rho = 0$ in the weights. The procedure for obtaining optimal weights gives for allele i $w_i = q_i/(n-1)$ for n number of alleles at the locus. This gives an estimator for a single locus, which combines information along alleles, as

$$\hat{\rho} = \sum_i \frac{S_i - P_i^2}{(n-1)P_i}$$

To combine estimates among loci, we use the fact that at zero true relationship and known gene frequency, the variance o f single locus weighted MME is proportional to $1/(n-1)$, regardless of the frequency distribution of alleles. The inverse of this quantity serves as the weight. This gives a simplified multi locus estimator of relationship, based upon a prior ρ of zero as:

$$\hat{\rho} = \sum_{i,l} \frac{S_{il} - P_{il}^2}{P_{il}} \Big/ \sum_l (n_l - 1)$$

where l denotes the locus. This estimator was first described by Li and Horvitz (1953).

A second simple method of moment estimator for ρ can be obtained by assuming $\rho=1$ in the weights. The weights then become $p_i q_{i/(1-J)}$, for J the expected homozygosity. Over m independent loci, this estimator equals:

$$\hat{\rho} = \frac{S - J}{1 - J}$$

for $J = \frac{1}{m}\sum_{i,l} P_{il}^2$ the mean expected homozygosity over the m loci and $S = \frac{1}{m}\sum_{i,l} S_{il}$ the arithmetic average of allele similarity between the two individuals across loci.

Simulation results showed the variance of the MME to be approximately a function of 1/m for m the number of loci. However for relation ships spanning a wide range and for many different distributions of gene frequency a systematic bias on the order of 1/N was observed, for N the number of individuals used to estimate gene frequency. Greater efficiency is obtained by using loci with even gene frequencies. The estimation of MME almost plateaus by 40-60 individuals where it nearly equals the predicted asymptotic variance $(1/\lfloor 4(n-1)m \rfloor$ for n alleles at each of m loci.

Regression method

Lynch and Ritland (1999) pursueing their search for optimal estimators for common situations when the number of loci are under 50, changed the name MME in that of correlation method and proposed a new one on the same principles but based on a regression approach. They also refined their analyses in proposing estimators for "higher-order" coefficients. The relatedness (kinship) coefficient for two individuals (x and y), two times their coefficient of coancestry (or parentage), can be written:

$$r_{xy} = \frac{\Phi_{xy}}{2} + V_{xy}$$

Where Φ_{xy} .is the probability that a single gene in x is identical by descent with one in y, and V_{xy} is the probability that each of the two genes in x is identical by descent with one in y. For parents and offspring, Φ =1 and ν=0; for full sibs, Φ=0.5 and ν=0.25; and for half sibs, Φ =0.25 and ν=0.
Consider a single locus with n alleles and let x be the reference individual (with alleles a and b) and y be the proband individual (with alleles c and d). The conditional probabilities for the n(n+1)/2 possible genotypes in y can be expressed as a function of Φxy, νxy and the known allele frequencies:

$$\begin{aligned} P(y = cd \mid x = ab) = {} & P_0(cd).(1 - \Phi_{xy} - V_{xy}) \\ & + P_1(cd \mid ab).\Phi_{xy} + P_2(cd \mid ab).V_{xy} \end{aligned}$$

Where $P_0(cd)$ is the Hardy-Weinberg probability of genotype cd, and $P_1(cd \mid ab)$ and $P2(cd \mid ab)$ denote the probabilities of genotype cd in y given genotype ab in x, the first being conditional on the two individuals having one gene identical by descent and the second being conditional on two genes being identical by descent.
Considering first x being homozygous for allele i and letting pi be the frequency of the ith allele, the preceding equation can be written:

$$P(ii\,|\,ii) = p_i^2 + p_i(1-p_i)\Phi_{xy} + (1-p_i^2)V_{xy}$$
$$P(i.\,|\,ii) = 2p_i\,(1-p_i) + (1-p_i)(1-2p_i)\Phi_{xy} - 2p_i(1-p_i)V_{xy}$$

which can be rearranged to yield the following estimators:

$$\hat{\Phi}_{xy} = \frac{(1+p_i)\hat{P}(i.\,|\,ii) + 2p_i\hat{P}(ii\,|\,ii) - 2p_i}{(1-p_i)^2}$$
$$V_{xy} = \frac{p_i^2 - p_i\hat{P}(i.\,|\,ii) + (1-2p_i)\,\hat{P}(ii\,|\,ii)}{(1-p_i)^2}$$

And,

$$\hat{r}_{xy} = \frac{\hat{P}(i.\,|\,ii) + 2\hat{P}(ii\,|\,ii) - 2p_i}{2(1-p_i)}$$

P(i./ii) and P(ii/ii) are estimated as 0/1 variables. Both probabilities are 0 if the proband y has no alleles in common with the reference x. Thus for example when individual y contains 2, 1 and 0 i alleles the estimates of r_{xy} are 1, (1-2pi)/2(1-pi) and –pi/(1-pi) respectively.
When x is heterozygous and the locus multiallelic there are six classes of conditional probabilities. Then the number of observed 0/1 variables exceeds the number of unknows (Φ and Δ). To deal with this situation a weighted least-square approximation is provided.
A general one locus estimator which cover all the cases is best described by introducing "indicator variables" for the sharing of pairs of alleles:
As before let the reference individual x have the alleles a and b and the proband individual y alleles c and d. If the reference individual is homozygous, S_{ab}=1 while if it is heterozygous S_{ab}=0. Likewise if allele a from the reference individual is the same as allele c from the proband S_{ac}=1, while S_{ac}=0 if it is different. In total, there are six S's corresponding to the six ways of choosing two objects without replacement from a pool of four objects. Letting p_a, p_b be the frequencies of alleles a and b in the population, the fully general expressions for the two locus-specific coefficients of primary interest are:

$$\hat{r}_{xy} = \frac{p_a(S_{bc}+S_{bd}) + p_b(S_{ac}+S_{ad}) - 4p_ap_b}{(1+S_{ab})(p_a+p_b) - 4p_ap_b}$$

$$V_{xy} = \frac{2p_ap_b - p_a(S_{bc}+S_{bd}) - p_b(S_{ac}+S_{ad}) + (S_{ac}S_{bd}) + (S_{ad}S_{bc})}{(1+S_{ab})(1-p_a-p_b) + 2p_ap_b}$$

There is no particular reason to use one member of a pair of individuals as the reference and the other as proband. Thus the reciprocal estimates *xy* and r*yx* can be arithmetically averaged to further refine the pairwise relationship estimates.

Multilocus estimates

As shown before, with statistically independent marker loci the locus-specific weights that minimize the sampling variance of the overall estimates are simply the inverse of the sampling variance of the locus-specific estimates. Approximations can be obtained by assuming x and y unrelated and general expressions for the weights w(*l*) are given by:

$$w_{r,x}(l) = \frac{1}{Var\left[\hat{r}_{xy}(l)\right]} = \frac{(1+S_{ab})(p_a+p_b)-4p_ap_b}{2p_ap_b}$$
$$w_{V,x}(l) = \frac{1}{Var\left[V_{xy}(l)\right]} = \frac{(1+S_{ab})(1-p_a-p_b)+2p_ap_b}{2p_ap_b}$$

Other methods

Queller and Goodnight (1989) presented also a regression based estimator for two-gene relatedness. Their one locus estimator was designed to estimate relatedness within groups of individuals but it can be adapted for estimating pair wise relationships. However their estimator

$$\hat{r}_{xy} = \frac{0.5\left(S_{ac}+S_{ad}+S_{bc}+S_{bd}\right)-p_a-p_b}{1+S_{ab}-p_a-p_b}$$

has limited utility with diallelic loci. Indeed, if x is heterozygous then Sab=0 and the equation is undefined because $p_a+p_b=1$

Eding and Meuwissen (2001) with similar approach and starting from Lynch's (1988) formula estimating f from the observed similarity S_l at a locus l and $h_l = \sum_{i=1}^{nl} p_{i,l}^2$ the probability of alleles of locus l being AIS (alike in state), are writing:

$$\begin{aligned} E(S_l) &= Pl \\ &= f + (1-f)h_l \\ &= h_l + (1-h_l)f \end{aligned} \qquad (1 \text{ ter})$$

This leads to the variance of $\hat{f}$

$$var(\hat{f}) = \frac{1}{(1-h_l)^2} var(S_l) \qquad (4)$$

Since S is the probability that two random alleles drawn from two individuals are alike, the distribution of S is binomial. The variance of S_l for a locus l is given as:

$$var(S_l) = P_l(1-P_l) \qquad (5)$$

Filling (1 ter) in (5) yields

$$\begin{aligned} Var(S_l) &= f(1-h_l)+h_l-\left[f^2(1-h_l)^2+2fh_l+h_l^2\right] \\ &= f(1-h_l)(1-2h_l)+h_l(1-h_l)-f^2(1-h_l)^2 \end{aligned} \qquad (6)$$

Substitution of 6 in 4 gives:

$$\begin{aligned} Var(\hat{f}) &= \frac{f(1-h_l)(1-2h_l)+h_l(1-h_l)-f^2(1-h_l)^2}{(1-h_l)^2} \\ &= \frac{h_l+f(1-2h_l)-f^2(1-h_l)}{(1-h_l)} \end{aligned} \qquad (7)$$

An over all loci estimation of f can be obtained through averaging over m analysed loci. We may use the inverse of the variance of the estimates of f for each independent locus as weights. We obtain the following estimation:

$$(8) \qquad f = \frac{\sum_{l=1}^{m} \hat{f}_l \left[\frac{(1-h_l)}{h_l + f(1-2h_l) - f^2(1-h_l)} \right]}{\sum_{l=1}^{m} \left[\frac{(1-h_l)}{h_l + f(1-2h_l) - f^2(1-h_l)} \right]}$$

Maximum Likelihood Estimator (MLE)

The maximum likelihood procedure was extensively investigated by Thompson (1975, 1976) for inferring pairwise relationship. She discussed the power of likelihood to distinguish among major types of relationships (parent-offspring, full sibs, half sibs, etc…) and unrelated. She found that due to large errors of inference it is difficult, even with 20 highly polymorphic loci, to distinguish among the major classes of relatives. However MLE gives asymptotically efficient estimates when the number of loci exceeds 50. This allows test of hypothesis via likelihood ratios and a better analytical analysis of the problem as we will try to demonstrate now.

The likelihood Y of the genotype of individual A for m independent loci is the product of the likelihood for each locus:

$$Y = \prod_{l=1}^{k} \left[h_l + (1-h_l) f_A \right] \times \prod_{l=1}^{j} (1-h_l)(1-f_A)$$

k loci being homozygous and j loci being heterozygous for individual A with f_A coefficient of inbreeding. h_l being the probability of being homozygous for the locus l in panmixia (equals the two alleles being AIS), $(1-h_l)$ being the probability of being heterozygous:

$$Y = (1+f_A)^k \prod_{l=1}^{k} \left[h_l \frac{(1-f_A)}{(1+f_A)} + \frac{f_A}{(1+f_A)} \right] \times (1-f_A)^j \prod_{l=1}^{j} (1-h_l)$$

Taking the natural logarithm:

$$\mathrm{Log_e}\, Y = k\, \mathrm{Log_e}\,(1+f_A) + \sum_{l=1}^{k} \mathrm{Log_e} \left[h_l \frac{(1-f_A)}{(1+f_A)} + \frac{f_A}{(1+f_A)} \right] + j\, \mathrm{Log_e}\,(1-f_A) + \sum_{l=1}^{j} \mathrm{Log_e}\,(1-h_l)$$

Derivative of $\mathrm{Log_e}Y$ with respect to f_A:

$$\frac{\partial \mathrm{Log_e}\, Y}{\partial f_A} = k \frac{1}{(1+f_A)} + \sum_{l=1}^{k} \frac{1-2h_l}{(1+f_A)\left[h_l + (1-h_l) f_A\right]} - j \frac{1}{1-f_A}$$

Which is zero for:

$$(1-f_A)\left\{\sum_{l=1}^{k}(1-\frac{2h_l-1}{[h_l+(1-h_l)f_A]})\right\}=j(1+f_A)$$

Defining $S_l=1$ for homozygotes and $S_l=0$ for heterozygotes we have for the $m=k+j$ loci

$$(1-f_A)\left\{\sum_{l=1}^{m}S_l(1-\frac{2h_l-1}{h_l+(1-h_l)f_A})\right\}=(1+f_A)\sum_{l=1}^{m}(1-S_l)$$

Considering S*l* taking the values 0 ¼ ½ 1, according to the situation of identity (see p7: generality) this formula allows the estimation of φ the parentage coefficient instead of f_A.
By definition of A and B

$$(1-f_A)\,A=(1+f_A)\,B$$

$$f_A=\frac{A-B}{A+B}$$

One can note that for

$$h_l=\sum_{i=1}^{m}p_i^2\ \#0.5\ \forall 1$$

$$f_A=\frac{k}{m}-\frac{j}{m}$$

A very simple estimator. This estimator is also independent from f_A and need therefore no prior assumptions.
The same argument at the allele level (not as before at the locus level) starting from the formulae just before (1) leads to similar results. In this case the weights r*i* take in account p*i* the allele frequency changing h*l* in p*i*.
Let us study the weight $r_l=\left(1-\frac{2h_l-1}{h_l+(1-h_l)f_A}\right)$ of a homozygous locus according to $h_l=\sum_{i-1}^{n}p_i^2$ and the prior on f_A .These weights are indeed functions of the parameters that we are trying to estimate. Their estimation needs therefore prior assumptions or iterative resolution. We can also remark that h_l is the inverse of the effective number of alleles A_e at the locus (i.e the equivalent number of alleles when even frequencies; A_e=2 for two equiprobable alleles, A_e=n for n equiprobable alleles). The weight r_l can easily be expressed in terms of A_e:

$$r_l=\left\{1-\frac{(2-A_e)}{[1+(A_e-1)]f_A}\right\}$$

Table 3 and figure 2 shows that this weight increases as h_l tends to zero and as f_A tends also to zero. This increase tends to be very drastic for h_l being under 0.10 and f_A smaller than 0.05.

Table 3 – r_l weight of a homozygote locus in estimation of f by MLE according to h_l probability of AIS and the prior on f

$$r_l = \left(1 - \frac{2h_l - 1}{h_l + (1 - h_l)\, f_A}\right) \qquad h_l = \sum_{i=1}^{n_l} p_i^2$$

h_l \ f_A	0	0.01	0.05	0.125	0.25	0.50	1
0.05	**19.000**	**16.126**	**10.231**	**6.333**	4.130	2.714	1.9
0.10	**9.000**	**8.339**	**6.517**	4.765	3.462	2.455	1.8
0.20	4.000	3.885	3.500	3.000	2.500	2.000	1.6
0.30	2.333	2.303	2.194	2.032	1.842	1.615	1.4
0.40	1.500	1.493	1.784	1.421	1.364	1.286	1.2
0.50	**1**	**1**	**1**	**1**	**1**	**1**	**1**
0.60	0.667	0.669	0.677	0.692	0.714	0.750	0.8
0.70	0.429	0.431	0.441	0.458	0.484	0.529	0.6
0.80	0.250	0.252	0.259	0.273	0.294	0.158	0.2
0.90	0.111	0.112	0.116	0.123	0.135	0.158	0.2
1	0	0	0	0	0	0	0

This over weighting of some homozygous loci favouring very polymorphic loci (or rare alleles) will make the estimation of f_A too much dependent of the situation of identity observed at few loci and the exact determination of alleles frequencies at such loci will be more difficult due to the low expected values. Contrary to general agreement I would therefore not recommend to use such loci. Biallelic loci with the value of h_l not so far from 0.5 would in my opinion allow more precise estimations because they are independent of the prior on f_A and are allowing more precise estimations of alleles frequencies. Table 4 shows the variations in h_l according to n the number of alleles and a frequency disequilibrium supposing a constant decrease of allele frequency from the most to the least frequent one. It can be observed that h_l is minimum for evenness and is decreasing with the number of alleles. Near 0.5 values for h_l are more easily obtained for bi allelic loci, multi allelic have to respect a constant decrease in alleles frequencies near 0.30 to satisfy to the condition. This observation is leading us to propose a kit of single nucleotide polymorphism (SNP) to study parentage and connected problems in horse populations.

Figure 2 – Illustration of Table 3
Weight r according to h and f.

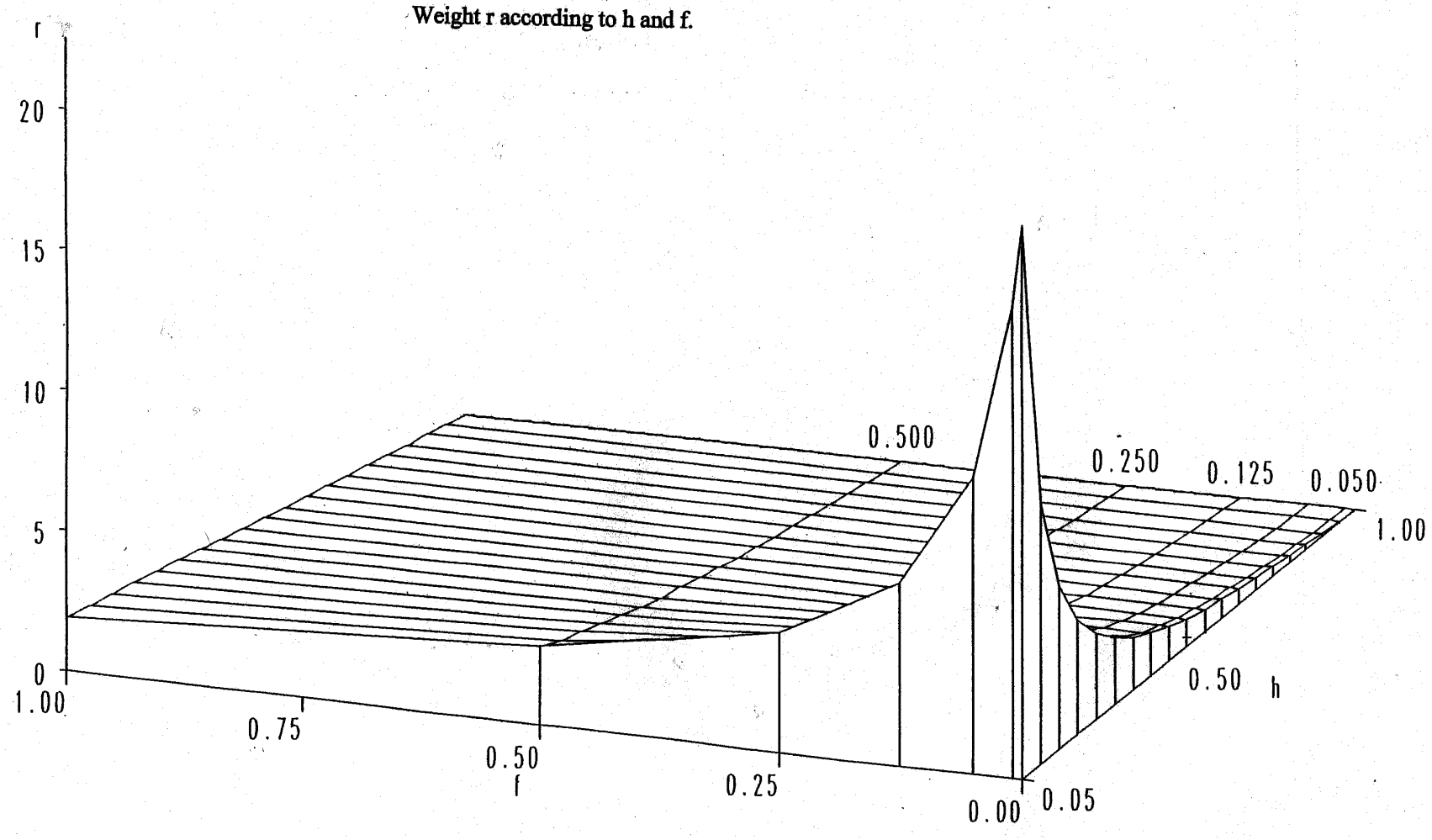


Table 4 - h_1 probability of being alike in state according to the number of alleles n and a model supposing constant decrease of allele frequency from the most to the least frequent one. $0 < a < 1$ is the constant percentage of decrease. a=1 represent the even frequencies case. $h_1 = \sum_{i=1}^{n_1} p_i^2$.

a \ n	2	3	4	5	10	20
0.1	0.835	0.820	0.818	0.818	0.818	0.818
0.2	0.722	0.677	0.669	0.667	0.667	0.667
0.3	0.645	**0.568**	**0.547**	**0.541**	**0.538**	**0.538**
0.4	0.592	0.487	0.451	0.437	0.429	0.429
0.5	**0.556**	0.429	0.378	0.355	0.334	0.333
0.6	**0.531**	0.388	0.324	0.292	0.253	0.250
0.7	**0.516**	0.361	0.288	0.248	0.187	0.177
0.8	**0.506**	0.344	0.265	0.219	0.138	0.114
0.9	**0.501**	0.336	0.253	0.204	0.109	0.067
1.0	**0.500**	0.333	0.250	0.200	0.100	0.050

Promoting the realisation of a kit of SNP

From the above studies it can be concluded that parentage analysis need a lot of markers to reach a reasonably good precision in practice. The problem to ascertain sire and dam of a foal is not so complicated and we have shown (Table1) that 11 polymorphic microsatellites markers are sufficient to solve it properly. However the problem of remote parentage remains open and pedigree information is often not available to solve it. To help for the resolution of this dilemma we propose the realisation of a kit of SNP.

This kind of markers has the advantage of being easily revealed by DNA chips, being bi-allelic, co-dominant and null alleles free. This greatly simplify their management in terms of population genetics. Although not as discriminant as polymorphic loci, 5 to 10 SNP are considered equivalent from this standpoint to one microsatellite.

It is thought in addition that a SNP can be found in mammals every 500 to 1000 pairs of bases. Microsatellites are expected only every 25 to 100 kilo-bases. The screening of horse genome would be therefore much more precise with SNP than with microsatellites.

It is also known that mammal's genome is approximately constituted by 60 segments of 50 centimorgans. 60 independent markers at a bare minimum can therefore be expected and 120 at the maximum.

Conclusion

Due to their potential great number and their revelation facilities (positive or negative responses on DNA chips) allowing to squize sequencing for routine analysis, SNP markers allow to consider the tracing of parentage.

The realisation of a kit of several hundreds of SNPs would allow precise estimation of allele frequencies and a choice of 100-120 independent loci to trace the parentage as seen before. This could be a goal for at the end a better mastering the real parentage between individuals. For small populations the question of the evolution of inbreeding should also be better faced than actually by only taking pedigrees in account.

This new kit would also facilitate the comparisons of horse populations according to more precise genetic distances.

The realisation of such a tool is only a problem of engineering and financing not a question of know how. In my opinion from the solution of this political problem will depend the future of genomic in horse breeding. I treated here only one part of the whole problem. But this part appears sufficient to justify the approach.

Acknowledgements

We thank Louis Ollivier for helpful comments

References

Amigues Y., Mériaux J.C., Boscher M.Y.,2000. Utilisation de marqueurs génétiques en sélection les activités de Labogéna. Inra Prod. Anim.,n° hors série "Génétique moléculaire : principes et applications aux populations animales",203-210.

Ayman N., Chakraborty R., 1982. Evaluation of paternity-testing data from the joint distribution of paternity index and the rate of exclusion. Hereditas, 96, 49-54.

Binns M.M., Holmes N.G., Holliman A., Scott A.M., 1995. The identification of polymorphic microsatellite loci in the horse and their use in thoroughbred parentage testing British Veterinary Journal, 151, 9-15.

Breen M., Lindgren G., Binns M.M., Norman J., Irvin Z., Bell K., Sandberg K., Ellegren H., 1997. Genetical and physical assignments of equine microsatellites – First integration of anchored markers in horse genome mapping. Mamm. Genome, 8, 267-273.

Chakraborty R., Shaw M., Schull W.J., 1974. Exclusion of paternity, the current state of the art. Am. J. Hum. Genet., 26, 477-488.

Cho G.J., Cho B. W. 2004. Microsatellite DNA typing using 16 markers for parentage verification of the Korean native horse. Asian-Aust. J. Anim. Sci. 17(6) 750-754.

Cothran E.G., Mac Cluer J.W., Weitkamp L.R., Pfenning D.W., Boyce A.J., 1984. Inbreeding and reproductive performance in standardbred horses. J. Heredity, 75, 220-224.

Cotterman C.W., 1940. A calculus for statistico-genetics. PhD Thesis, University, Columbus, Ohio.

Cunningham E.P., Dooley J.J., Splan R.K., Bradley D.G. 2001. Microsatellite diversity, pedigree relatedness and the contributions of founder lineages to thoroughbred horses. Animal Genetics,32, 360-364.

Devlin B., Roeder K., Ellstrand N.C., 1988. Fractional paternity assignment: theoretical development and comparison to other methods. Theor. Appl. Genet., 76, 369-380.

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

Eding H., 2002. Conservation of genetic diversity. PhD. Thesis 120p.

Eding H., Meuwissen T.H.E., 2001. Marker based estimates of between and within population kinships for the conservation of genetic diversity. J. of Anim. Breed. Genet., 118, 141-159.
Ellegren H., Johansson M., Sandberg K., Anderson L., 1992. Cloning of highly polymorpic microsatellites in the horse. Animal Genetics, 23, 133-142.

Gerber S., Mariette S., Streiff R., Bodénès C., Kremer A., 2000. Comparison of microsatellites and amplified fragment lenght polymorphysm markers for parentage analysis Mol. Ecol., 9, 1037-1048.

Guérin G., Bertaud M., Amigues Y., 1994. Characterization of seven new horse microsatellites: HMS1, HMS2, HMS3, HMS5, HMS6, HMS7 and HMS8. Animal Genetics, 25, 62.

Jacquard, A., 1972. Genetic information given by a relative. Biometrics, 28, 1101-1114.

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

Jones A., Ardren W.R., 2003. Methods of parentage analysis in natural populations. Molecular Ecology, 12, 2511-2523.

Kavar T., Brem G., Habe F., Sölkner J., Dovc P. 2002. History of Lipizzan horse maternal lines revealed by mtDNA analysis. Genet. Sel. Evol. 34, 635-648.

Li C.C., Horvitz D.G., 1953. Some methods of estimating the inbreeding coefficient. Am. J. Hum. Genet., 5, 107-117.

Li C.C., Weeks D.E., Chakravati A., 1993.Similarity of DNA fingerprints due to chance and relatedness. Hum. Hered. 43, 45-52.

Lynch M. Ritland K. 1999. Estimation of pairwise relatedness with molecular markers. Genetics, 152, 1753-1766.

Lynch M., 1988. Estimation of relatedness by DNA fingerprinting. Mol. Biol. Evol., 5, 584-599.

Mac Cluer J.W., Boyce A.J., Dyke B., Weitkamp L.R., Pfenning D.W., Parsons C.J., 1983. Inbreeding and pedigree structure in Standard bred horses. J. Heredity, 74, 394-399.

Mahon G.A.T., Cunningham E.P., 1982. Inbreeding and the inheritance of fertility in Thoroughbred mare. Livest. Prod. Sci., 9, 743-754.

Malécot G., 1948. Les mathématiques de l'hérédité. Paris, Masson et Cie.64p.

Malécot G. 1969 Consanguinité panmictique et consanguinité systématique (coefficients de Wright et de Malécot). Ann. Génét. Sél. Anim. 1, 237-242.

Marklund S., Ellegren H., Eriksson S., Sandberg K., Anderson L., 1994. Parentage testing and linkage analysis in the horse using a set of highly polymorphic microsatellites. Anim. Genet., 25, 19-23.

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.

Meagher T., 1986. Analysis of paternity within a natural population of chromaelirum luteus I Identification of the most likely parents. Am. Nat., 128, 199-215.

Meagher T.R., Thompson E.A., 1986. The relationship between single and parent pair genetic likelihoods in genealogy reconstruction. Theoretical Population Biology, 29, 87-107.

Meagher T.R., Thompson E.A., 1987. Analysis of parentage for naturally established seedlings within a population of Chamaelirium luteum (liliaceae). Ecology, 68, 803-812.

Moureaux S., Verrier E., Ricard A., Mériaux J.C., 1996. Genetic varaibility within French race and riding horse breeds from genealogical data and blood marker polymorphisms. Genet. Sel. Evol., 28, 83-102.

Queller, D.C., Goodnight K.F., 1989. Estimating relatedness using genetic markers. Evolution 43,258-275.

Ritland K., 1996. Estimators for pairwise relatedness and individual inbreeding coefficients. Genet. Res., 67, 175-186.

Robertson A., Hill W.G., 1984. Deviation from Hardy Weinberg proportions: sampling variances and use in estimation of inbreeding coefficients 1984. Genetics, 107, 703-718.

San Cristobal M., Chevalet C., 1997. Error tolerant parent identification from a finite set of individuals Genet. Res., 70, 53-62.

Selvin S., 1980. Probability of non paternity determined by multiple allele co-dominant systems. Am. J. Hum. Genet., 32, 276-278.

Smouse P.E., Chakraborty R., 1986. The use of restriction fragment length polymorphism in paternity analysis. Am. J. Hum. Genet., 38, 918-939.

Thompson E.A. 1975. The estimation of pairwise relationships. Ann. Hum. Genet. Lond., 39, 173-188.

Thompson E.A. 1976. Inference of genealogical structure. Social Sciences Information 15,477-526.

Thompson E.A., Meagher T.R., 1987. Parental and sib likelihoods in genealogy reconstitution. Biometrics, 43, 585-600.

Van Haeringen H., Bowling A.T., Stott M.L., Lenstra J.A., Zwaagstra K.A., 1994. A highly polymorphic horse microsatellite locus: VHL20. Anim. Genet., 25, 207.

Wright S., 1978. Evolution and the Genetics of Populations, Vol 4, Variability within and among Natural Populations. Univ. of Chicago press, Chicago, USA.

Zechner P., Sölkner J., Bodo I., Druml T., Baumung R., Achmann R., Marti E., Habe F., Brem G. 2002. Analysis of diversity and population structure in the lipizzan horse breed based on pedigree information. Livest. Prod. Sci. 77, 137-146.