How to estimate kinship and inbreeding with SNPs

58th Annual Meeting of the EAAP, Session 29, Dublin, Ireland, August 26th-29th 2007.

B. Langlois INRA-SGQA 78 350 Jouy en Josas. France

Abstract

I recently developed (8th wgalp) how to choose genetic markers to infer kinship or inbreeding coefficients. The first draft of the horse genome sequence has now been deposited in public databases and is freely available for use. In addition to sequencing the horse genome a map was produced which should comprise one million of SNPs. We will show in this paper how it could be used to estimate kinship and inbreeding coefficients

Key words
Horse # kinship # inbreeding # SNPs

Introduction

In horse populations, there is a great concern for pedigree. Most stud-books started during the 19th century and some of them even earlier. This administration work was done very carefully as it was done for humans with parish register. This, however, did not exclude some errors (Cunningham et al. 2001, Kavar et al. 2002) 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 the Arab, Thoroughbred and Trotter and commonly used for the others when artificial insemination is used or at random to discourage fraud. The result is a very low percentage of parentage errors in horse breeding. Genetic markers in this breeding are used only 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 forensic reasons. However as shown by Langlois (2005), pedigree information is limited because the total genetic history of an animal or a population cannot be traced from the beginning. Even with very complete and reliable pedigrees (Zechner et al. 2002), there are still events in the past that are not described like bottlenecks or the real number of founders (Mahon and Cunningham 1982, Mac Cluer et al. 1983, Cothran et al. 1984, Moureaux et al. 1996). We recently developed (Langlois 2006) how to choose genetic markers to infer kinship or inbreeding coefficients. We show the advantages of a panel of numerous independents SNPs which alleles frequencies are balanced (i.e. not far from p=0.5) to infer kinship and inbreeding which is quite different as to exclude (Jamieson and Taylor, 1997). In this paper we will develop among many others (Oliehoek et al.

2006) the formulae of estimating the kinship and the inbreeding coefficient in the bi allelic case corresponding to the SNPs which have also the advantage of being co dominant and null allele free. The reason of this paper is that the first draft of the horse genome sequence has now been deposited in public data bases and is freely available for use. The realisation of DNA ships for several hundred of SNPs optimally chosen to trace the parentage is therefore available.

Let us remind the methodology:

We must go back to Malecot (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 homozygous for allele i equals the probability of being IBD defined as the inbreeding coefficient f 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)

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

The probability s_{ij} of being heterozygous for alleles i and j or j and i

$$s_{ij} = (1-f)\ 2p_ip_j \qquad (2)$$

where p_i is the frequency of allele i (resp. j)
Let us add and subtract $p_i(1-p_i)$ to s_{ii}

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

We get after some simplification an expression which looks symmetrical to that of s_{ij} :

$$s_{ii} = (1+f)\ p_i(1-p_i)\ [1 + \frac{(2p_i -1)}{(1-p_i)\ (1+f)}] \qquad (1bis)$$

Indeed, because of bi allelism $p_j = 1-p_i$ and,

$$s_{ij} = (1-f)\ p_i(1-p_i)\ [2] \qquad (2bis)$$

Let us write the likelihood **L** of an individual for m independent loci. It is the product of the likelihood for each locus:

$$\mathbf{L} = \prod_{l=1}^{k} (1+f)\ p_l(1-p_l)\ \{1 + \frac{(2p_l - 1)}{(1-p_l)\ (1+f)}\} \times \prod_{l=1}^{j} (1-f)\ 2p_l(1-p_l) \qquad (3)$$

k loci being homozygous and j loci being heterozygous for the individual with f coefficient of inbreeding.p_l is the frequency of the homozygous allele and $(1-p_l)$ that of the other allele at the locus l.

$$\mathbf{L} = (1+f)^k \times \prod_{l=1}^{k} \Big[p_l(1-p_l) + \frac{p_l(2p_l-1)}{(1+f)}\Big) \times (1-f)^j \prod_{l=1}^{j} 2\, p_l(1-p_l)$$

Let us optimize this likelihood to find f. We first take the natural Logarithm:

$$\text{Log } \mathbf{L} = k \text{ Log } (1+f) + \sum_{l=1}^{k} \text{ Log } \Big[p_l(1-p_l) + \frac{p_l(2p_l-1)}{(1+f)}\Big]$$
$$+ j \text{ Log } (1-f) + \sum_{l=1}^{j} \text{ Log } 2p_l(1-p_l)$$

and derivative of Log $\mathbf{L}$ with respect of f

$$\frac{d \text{ Log } \mathbf{L}}{d f} = k \frac{1}{(1+f)} + \sum_{l=1}^{k} \frac{d \text{ Log } p_l}{d f} + \sum_{l=1}^{k} \frac{d \text{ Log } [p_l + (1-p_l)f]/(1+f)}{d f} - j \frac{1}{(1-f)}$$

because the derivative of the sum is the sum of the derivatives.
Let us remark that $d \text{ Log } p_l / d f = 0$
And because

$$d \text{ Log } \frac{p_l + (1-p_l)f}{(1+f)} / d f$$

Is quite heavy to develop with these notations, let us for simplification write more scholarly:

$$y = \text{Log } \frac{a + (1-a)x}{(1+x)}$$

$$y' = \frac{u'}{u} \qquad \text{with } u = \frac{a + (1-a)x}{1 + x}$$

$$u' = \frac{(1-a)(1+x) - [a + (1-a)x]}{(1+x)^2}$$

$$u' = \frac{(1-a)(1 + x - x) - a}{(1+x)^2}$$

$$u' = \frac{(1 - 2a)}{(1+x)^2}$$

$$y' = \Big[\frac{(1-2a)}{(1+x)^2}\Big] \Big\{ \frac{(1 + x)}{[a + (1-a)x]} \Big\}$$

$$y' = \left[\frac{(1-2a)}{(1+x)\,[a + (1-a)x]} \right]$$

Which leads with a = p_l and x =f to the following equation:

$$\frac{d \operatorname{Log} \mathbf{L}}{d f} = k \frac{1}{(1+f)} + \sum_{l=1}^{k} \frac{(1-2p_l)}{(1+f)\,[p_l + (1-p_l)f]} - j \frac{1}{(1-f)} \qquad (4)$$

which is zero when:

$$(1-f) \quad \{ \sum_{l=1}^{k} (1 - \frac{(2p_l-1)}{p_l + (1-p_l)f}) \} = j\,(1+f) \qquad (5)$$

Defining S_l = 1 for homozygous and S_l = 0 for heterozygous for the m = k + j loci, we have:

$$(1 - f)\,\{ \sum_{l=1}^{m} S_l \,(1 - \frac{(2p_l -1)}{pl + (1-pl)f})\} = (1+f) \sum_{l=1}^{m} (1-S_l) \qquad (5bis)$$

By definition of C and D,
$(1-f)C = (1+f)D$
and

$$f = \frac{C-D}{C+D}$$

D = j in any case and C is the sum of the weights w_l of the homozygous loci.

$$w_l = [1 - \frac{2p_l - 1}{p_l + (1-p_l)f}]$$

This weight is one when p_l = 0.5. Therefore when p_l = 0.5 for every $_l$,

f = (k-j)/m

which is a very simple estimator, which also does not need any prior assumptions on f and no iterative resolution.
In this problem, this kind of bi allelic markers with p_l # 0.5 are particularly worth full.

Let us now consider the estimation of Φ, the relationship coefficient between two individuals. It is the inbreeding coefficient of their virtual offspring. Four main situations have to be considered according to the probability of producing a homozygous offspring:

-1- The two individuals are homozygous for the same allele which frequency is p_l. They will produce with probability 1 a homozygous offspring for this allele: $S_1 = 1$ for alleles p or $q=(1-p)$ and $(1-S_1) = 0$.
-2- The two individuals are homozygous for two different alleles: they will never produce a homozygous: $S_2 = 0$ and $(1-S_2) = 1$.
-3- One individual is homozygous for the allele with frequency p_l, the other is heterozygous. They should produce a half homozygous and a half heterozygous. $S_3 = ½$ for alleles p or q.
-4- The two individuals are heterozygous, they will produce a half of homozygous $S_4 = ½$ from which $S_{4a} = ¼$ for allele p and an other $S_{4b} = ¼$ for the other allele q.
Replacing f by Φ and S_l by the corresponding values for each loci (i.e. $S_1 = 1$; $S_2 = 0$; $S_3 = ½$; $S_4 = S_{4a} + S_{4b}$ with $S_{4a} = S_{4b} = ¼$, according to the situation of identity of the two individuals in equation (5bis) when using adequate alleles frequencies will lead to the estimation of Φ instead of f. That is:

$$\Phi = \frac{C - D}{C + D} \quad \text{with} \quad C = \sum_{l=1}^{m} S_l \left(1 - \frac{(2p_l - 1)}{p_l + (1-p_l)\Phi}\right) \quad \text{and } D = \sum_{l=1}^{m} (1 - S_l)$$

Considering that from the m loci m_1 are in situation 1(two homozygous individuals for the same allele), m_2 are in situation 2 (two homozygous individuals for different alleles), m_3 are in situation 3 (one individual homozygous, the other heterozygous), m_4 are in situation 4 (two heterozygous individuals).

$$C = \sum_{l=1}^{m_1} \left(1 - \frac{2p_l - 1}{p_l + (1-p_l)\Phi}\right) + \sum_{l=1}^{m_3} 1/2 \left(1 - \frac{2p_l - 1}{p_l + (1-p_l)\Phi}\right.$$
$$+ \sum_{l=1}^{m_4} \left\{ 1/4 \left(1 - \frac{2p_l - 1}{p_l = (1-p_l)\Phi}\right) + 1/4 \left(1 - \frac{2q_l - 1}{q_l + (1-q_l)\Phi}\right)\right\}$$

and what ever p_l, q_l and Φ,

$D = m_1 + ½ (m_3 + m_4)$

If as recommended p_l # q_l # 0.5

$C = m_1 + ½ (m_3 + m_4)$

Gives:

$\Phi = (m_1 - m_2)/m$

a very simple estimator of Φ

CONCLUSION

From the above studies, it can be concluded that parentage analysis needs a lot of markers to reach a reasonably good precision in practice. We propose the realisation of an SNP kit. This kind of marker has the advantage of being easily revealed by DNA chips, being bi-allelic, co-dominant and null allele free. In addition, it is thought that an SNP can be found in mammals every 500 to 1000 pairs of bases. Micro-satellites are expected only every 25 to 100 kilo-bases. The screening of the horse genome would therefore be much more precise with SNP than with micro-satellites. It is also known that the mammal genome is approximately constituted of 60 segments of 50 centimorgans. Sixty independent markers at a bare minimum can therefore be expected and 120 at the maximum when independence is strictly respected. Some slight dependence when accepted (Slate *et al.* 2004) could easily lead to a panel of more than 300 markers that could be efficient in parentage analysis (Fernandez et al. 2005). Due to their potential great number and their revelation facilities (positive or negative responses on DNA chips) allowing to squeeze sequencing for routine analysis, SNP markers allow the consideration of the tracing of parentage. The realisation of a kit of several hundreds of SNP would allow precise estimation of allele frequencies and a choice of 100-120 independent loci in order to trace the parentage as seen before. This could be a goal leading to a better mastering of the real parentage between individuals at the end. For small populations the question of the evolution of inbreeding should also be better faced than currently by only taking pedigrees into account. The realisation of such a tool is only a problem of engineering and financing and not a question of know-how. In my opinion, the future of genomics in horse breeding will depend on the solution of this political problem. I treated here only one part of the whole problem. But this part appears sufficient to justify the approach.

REFERENCES:

Cothran E.G., Mac Cluer J.W., Weitkamp L.R., Pfenning D.W., Boyce A.J. (1984) *J. Heredity* **75:** 220-224.

Cunningham E.P., Dooley J.J., Splan R.K., Bradley D.G. (2001) *Animal Genetics* **32:** 360-364.

Fernandez J., Villanueva B., Pong-Wong R., Toro M.A. (2005) *Genetics,* **170:** 1313-1321.

Kavar T., Brem G., Habe F., Sölkner J., Dovc P. (2002) *Genet. Sel Evol.* **34:** 635-648.

Langlois B. (2005) EAAP publication **n°116:** 35-54

Langlois B. (2006) 8th wgalp session 8 4pp

Li C.C., Horvitz D.G. (1953) *Am. J. Hum. Genet.* **5:** 107-117.

Li C.C., Weeks D.E., Chakravati A. (1993) *Hum. Hered.* **43:** 45-52.

Lynch M. (1988) *Mol. Biol. Evol.* **5:** 584-599.
Lynch M., Ritland K. (1999) *Genetics* **152:** 1753-1766.
Mac Cluer J.W., Boyce A.J., Dyke B., Weitkamp L.R., Pfenning D.W., Parsons C.J. (1983) *J. Heredity*, **74:** 394-399.
Mahon G.A.T., Cunningham E.P. (1982) *Livest. Prod. Sci.* **9:** 743-754.
Malécot G. (1948) Les mathématiques de l'hérédité. Paris, Masson et Cie.64p.
Moureaux S., Verrier E., Ricard A., Mériaux J.C. (1996) *Genet. Sel. Evol.* **28:** 83-102.
Oliehoek P.A.,Winding J.J., van Arendonk J.A.M., Bijma P. 2006. Genetics, 173, 483-496.
Ritland K. (1996) *Genet. Res*, **67 :** 175-186.
Slate J., David P., Dodds K.G. Veenvliet B.A., Glass B.C., Broad T.E., McEwan J.C. (2004) Heredity, **93 :** 255-265.
Thompson E.A. (1975) *Ann. Hum. Genet. Lond,* **39:** 173-188.
Thompson E.A. (1976). *Social Sciences Information* **15:** 477-526.
Zechner P., Sölkner J., Bodo I., Druml T., Baumung R., Achmann R., Marti E., Habe F., Brem G. (2002). *Livest. Prod. Sci.* **77:** 137-146.