The Impact Of Method On The Estimated Effective Population Size Of A Thoroughbred Population Using Genotype Data

L. J. Corbin[*], S.C. Bishop[*], J. E. Swinburne[†], M. Vaudin[†], S.C. Blott[†] and J. A. Woolliams[*]

Introduction

The release of the Illumina Equine SNP50 BeadChip has increased the potential for in depth genomic analyses of the equine genome. The efficacy of high density SNP arrays is dependent on the extent of linkage disequilibrium (LD) and its rate of decline with distance within populations. The pattern of LD in closed populations is influenced by effective population size (N_e) and estimates of N_e can therefore give an indication as to the likely utility of the genomic tools or be used to assess endangerment. Pedigree data can be used to estimate N_e, but sufficiently lengthy pedigrees are rarely available for horse populations. Alternatively, marker data can be used to predict historical N_e, based on the formula for expected LD derived by Sved (1971) (Hayes et al. (2003); Tenesa et al. (2007); de Roos et al. (2008)). Simulation studies suggest N_e estimates by this method are at least qualitatively correct (Hayes et al. (2003)). Various approaches to this method have been taken for the analysis of human and livestock populations, however the impact of changes made to the original method on subsequent N_e predictions has received little attention. The aim of this study was to investigate the impact of using different approaches, taken from the literature, on N_e prediction.

Material and methods

Data for this study consisted Illumina Equine SNP50 BeadChip genotype data for 817 UK Thoroughbred horses. 34% of the 52,603 autosomal single nucleotide polymorphism (SNP) markers were excluded due to monomorphism, poor genotyping quality (genotyping in <95% of samples), deviation from Hardy-Weinberg equilibrium ($p<0.0001$) and low minor allele frequencies (<0.1) (Hill (1981)). The LD measure r^2 was calculated for all remaining syntenic markers pairs, using an EM algorithm to estimate haplotype frequencies. Under the assumption of a finite population and random mating between individuals, Sved (1971) derived an approximate expression for the expectation of r^2 such that:

$$E(r^2) = \frac{1}{1+4Nc}, \quad (1)$$

where N is effective population size and c is the recombination frequency. This formula can be used as the basis for modelling the decay of LD with distance and for predicting N_e

[*] Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Roslin, EH25 9PS, UK.
[†] Animal Health Trust, Newmarket, CB8 7UU, UK.

(Tenesa et al. (2007); Abasht et al. (2009)). Rearrangement of Eq. 1 allows the prediction of effective population at a given point in time, expressed as generations in the past (Hill (1981); Hayes et al. (2003); Tenesa et al. (2007); de Roos et al. (2008)),

$$N_T = (\frac{1}{4c})(\frac{1}{r^2} - 1), \quad (2)$$

where N_T is the effective population size T generations ago, c is the recombination frequency and $T = 1/(2c)$, assuming linear growth (Hayes et al. (2003)). We used a variety of forms of Eq. 1 and 2 to develop non-linear regression equations to model r^2 and to derive predictions of ancestral N_e, respectively. Details of the equations, along with a brief justification for their inclusion can be found in Table 1. In all cases, a and b are the intercept and regression coefficient for the (non-linear) regression equations, respectively. Chromosome specific megabase to centimorgan conversion rates for Eq. 2 and 9 were based on total physical chromosome length, as stated on the NCBI website, and total chromosome genetic length from the latest complete equine linkage map (Swinburne et al. (2006)).

Results and discussion

The non-linear regression modelling of the decline of LD with distance resulted in both a and b being significantly different from zero in all models. The mean estimates for a and b under the different models are shown in Table 2. Regardless of the model implemented, estimates of a are between 2.2 and 2.3, with a small confidence interval. On fixing a to unity as in Zhao et al. (2005), Abasht et al. (2009) and Toosi et al. (2009), we observed an approximate doubling of b, demonstrating a considerable impact on N_e estimates derived by this method. Based on Eq. 1 in Table 1, the line of predicted r^2 only approximately follows that of the mean observed r^2, with the greatest discrepancy occurring at distances less than 0.03 Mb. This lack of fit is presumably due to the violation of the assumptions of the model, in particular a non-constant population size. In all cases, a significant negative correlation ($p<0.01$) was observed between estimates of b and chromosome length (cM). This contrasts with the findings of Tenesa et al. (2007) who observed a positive relationship, but is in keeping with the observations of Khatkar et al. (2008) and Muir et al. (2008).

For the most recent ten generations, our results suggest an increase in N_e from generation ten to generation two, at which point N_e levels off (Fig. 1). This trend remains largely intact, regardless of changes to the formulae used and can be rationalised by what is known about the history of the Thoroughbred breed. However, N_e values for the most recent ten generations are dependent on the method used and range from 150 to 350 at generation one. Such differences may be critical if, for example, predictions are used for assessing endangerment. It is also not clear what impact population structure in the sample, such as overlapping generations, might have on the most recent N_e predictions and whether the suggested dip in N_e at generation one is merely an artefact of the method itself. Using recombination rate in place of linkage distance had the biggest impact on predictions of recent N_e (Eq. 12). However, using recombination rate with Sved's (1971) formula prior to approximation (Eq. 13), gives N_e estimates closer to those computed using linkage distance (Eq. 8). As predicted by Zhao et al. (2005), Eq. 11 affects primarily those predictions of N_e many generations in the past. Predictions from the other equations at this point are more consistent, predicting a decrease in N_e from the distant past to approximately 20 generations ago (data not shown).

Table 1: Non-linear regression models and N_T equations tested

Eq. Pair	Eq. No.	Non-linear regression equation[1]	Eq. No.	Eq. for prediction of ancestral effective population size[2]
A	1[a]	$y_i = \frac{1}{(a + 4bd_i)} + e_i$	8[a]	$N_T = (\frac{1}{4d})(\frac{1}{r^2} - 1)$
B	2[b]	$y_i = \frac{1}{(a + 4bd_i)} + e_i$	9[b]	$N_T = (\frac{1}{4d})(\frac{1}{r^2} - 1)$
C	3[c]	$adj.\ y_i = \frac{1}{(a + 4bd_i)} + e_i$	10[c]	$N_T = (\frac{1}{4d})(\frac{1}{adj.\ r^2} - 1)$
D	4[d]	$y_i = \frac{1}{(1 + 4bd_i)} + e_i$	8[a]	$N_T = (\frac{1}{4d})(\frac{1}{r^2} - 1)$
E	5[e]	$y_i = \frac{1}{(2 + 4bd_i)} + e_i$	11[e]	$N_T = (\frac{1}{4d})(\frac{1}{r^2} - 2)$
F	6[f]	$y_i = \frac{1}{(a + 4bc_i)} + e_i$	12[f]	$N_T = (\frac{1}{4c})(\frac{1}{r^2} - 1)$
G	7[g]	$y_i = \frac{1}{(a + 4bc_i[\frac{(1-c_i/2)}{(1-c_i)^2}])} + e_i$	13[g]	$N_T = (\frac{1}{4c[\frac{(1-c/2)}{(1-c)^2}]})(\frac{1}{r^2} - 1)$

[1]Where, y_i is r^2 for SNP pair i, at linkage distance d_i in Morgans or recombination rate c_i assuming that 1 cM ≈ 1 Mb (except for Eq. 2).

[2]Where, N_T is the effective population size T generations ago, d is the marker distance in Morgans or c is the recombination rate assuming that 1 cM ≈ 1 Mb (except for Eq. 9) and $T = 1/(2c)$, under the assumption of linear growth (Hayes et al. (2003)).

[a]Previously implemented by Tenesa et al. (2007) (Eq. 1/8) and de Roos et al. (2008) (Eq. 8)

[b]Using chromosome specific megabase to centimorgan conversion rates, as suggested by Abasht et al. (2009) and applied by Qanbari et al. (2009).

[c]r^2 (y_i) replaced with *adj.* r^2 (y_i) = r^2 - (1/2n), where n is the sample size. Suggested to adjust for sampling effect (Hill(1981); Tenesa et al. (2007)).

[d]Previously implemented by Zhao et al. (2005), Abasht et al. (2009) and Toosi et al. (2010).

[e]To take account of mutation (Hill(1975); Tenesa et al. (2007)).

[f]Using recombination rate instead of linkage distance as in Sved's (1971) derivation.

[g]Version of formula prior to approximation, which assumed small c (Sved (1971)).

Table 2: Parameter estimates for Models 1-7[a]

	Parameter Estimate			
	a		*b*	
Model	Mean	95% C.I.	Mean	95% C.I.
1	2.25	[2.18; 2.33]	127.74	[119.56; 135.91]
2	2.25	[2.18; 2.33]	103.06	[95.84; 110.29]
3	2.25	[2.17; 2.32]	129.79	[121.45; 138.13]
4	NA	NA	249.73	[230.06; 269.39]
5	NA	NA	137.38	[127.55; 147.20]
6	2.20	[2.12; 2.28]	134.91	[126.69; 143.14]
7	2.27	[2.20; 2.35]	125.08	[116.99; 133.16]

[a]Mean across autosomes, as calculated by meta-analysis (an inverse variance method and a random effects model)

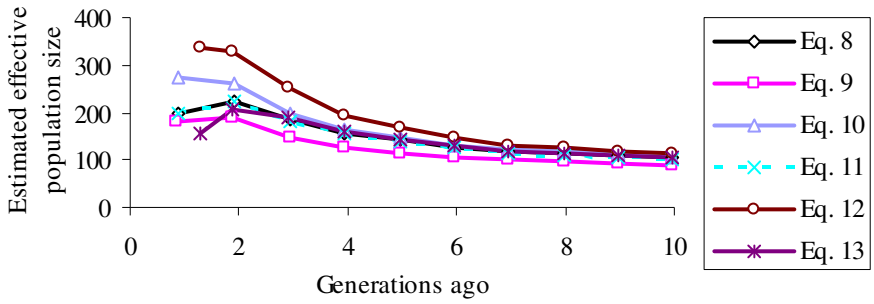

Figure 1: Effective population size as a function of generations in the past, as estimated by Eq. 8-13 (truncated at ten generations)

Conclusion

Estimates of N_e obtained by non-linear regression are dependent on assumptions made with respect to the intercept. In situations where population size has not been constant over time, this model may provide a poor fit to the data. Whilst the trend of ancestral N_e remains consistent despite changes to the prediction equation, large discrepancies between recent N_e estimates are of concern and lead us to question the ability of these methods to produce reliable estimates of more recent N_e. Pedigree analyses of the UK Thoroughbred population may enable us to determine the most accurate method for predicting N_e in this population.

References

Abasht, B., Sandford, E., Arango, J. et al. (2009). *BMC Genomics*, 10(Suppl. 2):S2.

de Roos, A., Hayes, B., Spelman, R. et al. (2008). *Genetics*, 179:1503-1512.

Hayes, B., Visscher. P., McPartlan, H. et al. (2003). *Genome Res*., 13:635-643.

Hill, W. (1975). *Theor. Pop. Biol*., 8:117-126.

Hill, W. (1981). *Genet. Res*., 38:209-216.

Khatkar, M., Nicholas, F., Collins, A. et al. (2008). *BMC Genomics*, 9:187.

Muir, W., Wong, G., Zhang, Y. et al. (2008). *World's Poultry Sci. J.*, 64:219-226.

Qanbari, S., Pimentel, E., Tetens, J. et al. (2009). *Anim. Genet*., 10.1111/j.1365-2052.2009.02011.x

Sved, J. (1971). *Theor. Pop. Biol*., 2:125-141.

Swinburne, J., Boursnell, M., Hill, G. et al. (2006). *Genomics*, 87:1-29.

Tenesa, A., Navarro, P., Hayes, B. et al. (2007). *Genome Res*., 17:520-526.

Toosi, A., Fernando, R., and Dekkers, J. (2010). *J. Anim Sci*., 88(1):32-46.

Zhao, H., Nettleton, D., Soller, M. et al. (2005). *Genet. Res*., 86:77-87.