Evaluation of factors affecting individual assignment precision using microsatellite data from horse breeds and simulated breed crosses

G. Bjørnstad and K. H. Røed

Department of Morphology, Genetics and Aquatic Biology, The Norwegian School of Veterinary Science, Oslo, Norway

Summary

Assignment tests have been utilized to investigate population classification, measure genetic diversity and to solve forensic questions. Using microsatellite data from 26 loci genotyped in eight horse breeds we examined how population differentiation, number of scored loci, number of scored animals per breed and loci variability affected individual assignment precision applying log likelihood methods. We found that both genetic differentiation and number of scored loci were highly important for recognizing the breed of origin. When comparing two and two breeds, a proportion of 95% of the most differentiated breeds ($0.200 \leq F_{ST} \leq 0.259$) could be identified scoring only three loci, while the corresponding number was six for the least differentiated breeds ($0.080 \leq F_{ST} \leq 0.139$). An identical proportion of simulated breed crosses, differentiated from their parental breeds by F_{ST} estimates in the range 0.050–0.069, was identified when scoring 12 loci. This level of source identification was not obtained for the less differentiated breed crosses. The current data further suggested that population sample size and locus variability were not critical for the assignment precision as long as moderately large sample sizes (≥ 20 animals per population) and fairly variable loci were used.

Keywords assignment tests, F_{ST}-statistics, genetic differentiation, horse breeds, hybrid recognition, microsatellites.

Introduction

The availability of multiple polymorphic DNA genetic markers has created the opportunity to use individual genotype information to determine the population origins of individuals (reviewed in Waser & Strobeck 1998; Davies *et al.* 1999). Assignment procedures have been applied for purposes such as detecting population demarcation (e.g. Paetkau *et al.* 1995; MacHugh *et al.* 1998; Primmer *et al.* 1999; Roques *et al.* 1999; Bjørnstad & Røed 2001), comparing sex-biased dispersal rates (Favre *et al.* 1997), investigating migration pathways and wintering habitats of shorebirds from different breeding areas (Haig *et al.* 1997) and in testing the origin of contemporary populations (Nielsen *et al.* 1997). Molecular examination of the population of origin has also been applied in forensics (Primmer *et al.* 2000), and it could be used to examine the origin of livestock products and determine whether meat samples have their origin from endangered or commercially exploited populations.

The potential to discriminate between pure-breds and breed crosses is an emerging challenge for the assignment tests and will be important for securing skilful population management and for the conservation of animal genetic resources. However, several factors will affect the ability to identify the genetic source of both pure-breds and hybrids. The genetic differentiation between populations is likely to influence the capacity to assign individuals to their source population (Cornuet *et al.* 1999) and probably more so, to detect their hybrids. Also the class of genetic marker, number of scored markers and the variability of the loci will affect the discrimination between source populations. Microsatellites are generally more polymorphic, and resolve population of origin better than diallelic markers

Address for correspondence

G. Bjørnstad, International Livestock Research Institute, PO Box 30709, Nairobi, Kenya.
E-mail: g.bjornstad@cgiar.org

Accepted for publication 9 February 2002

(Blott *et al.* 1999) and allozymes, despite the observation that the multilocus F_{ST} estimates computed across populations were similar for the marker categories (Estoup *et al.* 1998). High assignment success can be achieved by scoring even a low number of microsatellites (e.g. Buchanan *et al.* 1994; MacHugh *et al.* 1998; Blott *et al.* 1999). The assignment precision will probably be influenced by within-locus variability, but in which way is somewhat ambiguous (e.g. Bowcock *et al.* 1994; Blott *et al.* 1999). Finally, the number of animals analysed per breed and the choice of assignment methods will affect the success of the population assignment trials (Cornuet *et al.* 1999).

In this study we addressed how assignment precision was influenced by interbreed genetic differentiation, number of scored loci, locus variability and breed sample size. Published characterizations based on empirical data of eight horse breeds using 26 microsatellites were used as the data source for the present study (Bjørnstad *et al.* 2000a; Bjørnstad & Røed 2001). A clear demarcation among the current breeds has been confirmed (Bjørnstad & Røed 2001). We therefore wanted to evaluate the ability to distinguish breed crosses from their source breeds. The relatively high number of breeds, the different levels of differentiation among the breeds, and the high number of analysed loci, make the horse a suitable model for analysing the factors affecting the potential to distinguish both among pure-bred animals and hybrids.

Materials and methods

Genotype data

Genotyping data from 26 microsatellites in eight horse breeds was the basis for the present study. The panel of breeds included the four native Norwegian breeds Fjord Horse ($n = 40$), Nordland/Lyngen Horse ($n = 30$), Døle Horse ($n = 40$) and Coldblooded Trotter ($n = 44$). In addition, Icelandic Horse ($n = 37$), Shetland Pony ($n = 34$), Standardbred Trotter ($n = 41$) and Thoroughbred ($n = 44$) were included. Laboratory protocols, including the procedure for microsatellite genotyping, were described in Bjørnstad *et al.* (2000a,b).

The genetic differentiation between loci and breeds, measured by F_{ST} (θ, Weir & Cockerham 1984) across breeds and loci, respectively, were estimated using FSTAT version 1.2 (Goudet 1995). The differentiation of loci ranged between 0.09 and 0.38 (Table 1), while the differentiation values between breed pairs were distributed in the range from 0.080 to 0.258 (Table 2). Locus variability was also

Table 1 Locus characteristics including number of alleles, average heterozygosity, F_{ST} and error rate, defined as fails in the assignment trials, estimated across eight horse breeds. The loci are sorted according to locus error rate.

Locus	Number of alleles	Heterozygosity	F_{ST}	Error rate
NVHEQ82	7	0.66	0.12	0.48
AHT5	9	0.67	0.21	0.49
NVHEQ11	8	0.62	0.20	0.50
NVHEQ100	11	0.66	0.20	0.50
ASB17	15	0.76	0.13	0.51
HMS2	13	0.65	0.19	0.52
NVHEQ18	19	0.75	0.14	0.53
VHL20	11	0.72	0.13	0.53
HTG4	6	0.60	0.19	0.54
UCDEQ425	10	0.74	0.11	0.55
ASB2	14	0.79	0.09	0.56
NVHEQ79	10	0.62	0.24	0.56
NVHEQ43	10	0.72	0.14	0.57
AHT4	13	0.76	0.09	0.59
NVHEQ29	11	0.69	0.12	0.59
HTG7	6	0.54	0.24	0.59
HTG6	8	0.41	0.38	0.61
LEX20	11	0.66	0.12	0.62
HMS7	10	0.61	0.10	0.62
NVHEQ21	5	0.56	0.20	0.64
NVHEQ70	9	0.78	0.10	0.65
HMS6	8	0.73	0.10	0.67
NVHEQ40	7	0.66	0.14	0.69
HTG14	7	0.69	0.12	0.70
NVHEQ54	3	0.31	0.09	0.77
NVHEQ5	4	0.47	0.10	0.80

Table 2 The differentiation index F_{ST} (θ, Weir & Cockerham 1984) estimated between eight horse breeds (below diagonal) and their hybrids (above diagonal). The estimates were based on 26 microsatellite loci.

	1	2	3	4	5	6	7	8
1 Shetland Pony		0.023	0.043	0.037	0.037	0.058	0.045	0.059
2 Icelandic Horse	0.100		0.024	0.019	0.018	0.040	0.034	0.050
3 Nordland/Lyngen	0.171	0.105		0.030	0.033	0.057	0.037	0.050
4 Fjord Horse	0.152	0.087	0.126		0.020	0.037	0.032	0.041
5 Coldblooded Trotter	0.150	0.083	0.140	0.093		0.018	0.034	0.049
6 Døle Horse	0.221	0.162	0.218	0.152	0.080		0.056	0.069
7 Standardbred	0.178	0.137	0.148	0.133	0.137	0.212		0.031
8 Thoroughbred	0.226	0.195	0.200	0.165	0.191	0.258	0.129	

measured as average heterozygosity and number of detected alleles across the breeds. The number of alleles within a locus varied between 3 and 19, while locus heterozygosity ranged between 0.31 and 0.79 (Table 1).

The breed specific allele frequencies were used to simulate hybrid genotypes. From each of the 28 different breed combinations 100 hybrids were generated, by drawing one allele per marker from each of the parental gene pools. The average differentiation values between hybrids and their parental breeds were distributed in the range from 0.018 to 0.069 (Table 2), and cover the F_{ST} range below the lowest differentiation index observed between pure breeds.

Both the pure breeds and the simulated hybrid populations were tested for deviations from Hardy–Weinberg equilibrium and linkage equilibrium using GENEPOP, version 1.2 (Raymond & Rousset 1995). The populations did not show significant deficiency of heterozygotes (i.e. no Wahlund effect for the breed crosses) or more linkage disequilibrium than expected by chance.

Population assignments

Assignment methods are based on the likelihood that the multilocus genotype of the individual to be assigned occurs in each of two or more candidate populations. The likelihood computations rely on the assumptions that loci are in Hardy–Weinberg equilibrium and linkage equilibrium. We used the frequency method for assigning individuals to populations, first presented by Paetkau *et al.* (1995). The method involves the following steps: first, the allele frequency distributions of all candidate populations are calculated in the absence of the individual in question, then the likelihoods of the individual's genotype occurring in each population are calculated, and finally, the individual is assigned to the population in which the individual's genotype is most likely to occur. The calculations were performed using WhichRun 3.2 (Banks & Eichert 2000). The confidence of each assignment was resolved by utilizing the log of the odds (lod) ratio for the two most likely source populations. The assignment pattern was investigated as a function of the estimated differentiation between breed pairs. Consequently only two breeds were regarded as a possible source for each test. The assignment pattern was also investigated as a function of the number of scored loci, using a random order of loci.

The assignment error rate of single microsatellites and the effect of breed sample size were investigated across the eight pure breeds using Geneclass 1.0.02 (Cornuet *et al.* 1999), with the default settings of the frequency likelihood method, i.e. 'leave one out' (individuals were excluded from their population when testing their origin) and in case of absence of alleles in possible source populations, such null allele frequencies were set to a value of 0.01. A Bayesian method (Rannala & Mountain 1997; Cornuet *et al.* 1999) was also evaluated for these computations, but only the results of the frequency method are reported because the approaches gave approximately the same results. To test the effect of breed sample sizes, five to 30 animals per breed were analysed, using 10 sets of loci, each consisting of 10 randomly selected microsatellites. For testing the error rate of single microsatellites, breed sample sizes were made uniform by randomly selecting 30 animals per breed to avoid the bias of overall error rate estimates caused by different assignment precision of the various breed comparisons.

Results

Effects of number of scored loci and breed differentiation

More than 95% of the pure bred animals were correctly identified with high certainty (lod2) when scoring at least 13 microsatellites (Fig. 1). Far lower number of loci was required to obtain this precision when using less strict assignment thresholds, i.e. only six random loci at lod0 level and nine loci at lod2 threshold (Fig. 1). Using different orders in which the loci were scored gave approximately the same assignment results.

Different levels of breed differentiation had relatively little effect on the assignment precision when using the relaxed lod0 criterion (Fig. 1), but the way in which the breed differentiation influenced the assignment precision was more apparent when applying the stricter assignment criteria. For instance, the lod2 threshold required analysis of six loci to identify 95% of the most differentiated breeds, while this number of loci only identified roughly 75% of the least differentiated breeds (Fig. 1). When more than 10–12 loci were scored, the degree of breed differentiation was less important, because the assignment precision was high (more than 90%) for all breed combinations.

The genetic differentiation from their parental breeds had a strong effect on the assignment precision of the breed crosses. This was seen when using the relaxed threshold, and even clearer when using the more strict ones (Fig. 1). For instance at the lod0 threshold, analysis of 12 loci identified more than 95% of the crosses between the most differentiated breeds, while crosses between the least differentiated breeds did not reach this recognition level (Fig. 1). In particular, the assignment precision of crosses between closely related breeds was seriously affected when the assignment stringency was increased (Fig. 1). But all together, more than 90% of the simulated breed crosses could be identified when scoring at least 18 loci and using the most relaxed assignment threshold.

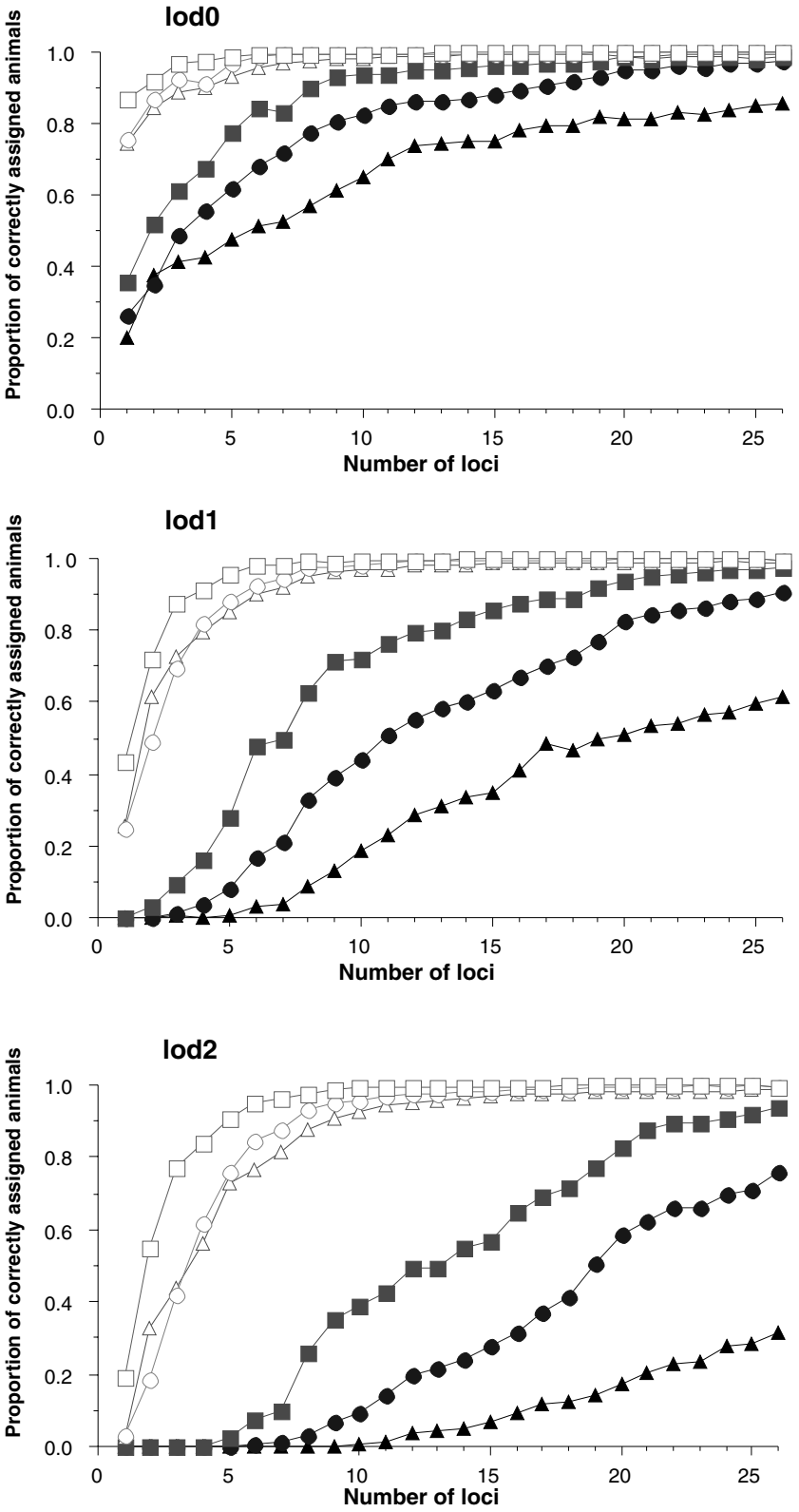

Figure 1 The proportion of correctly assigned animals interpreted as a function of the genetic differentiation between the breeds and the number of scored loci, applying three acceptance thresholds. Breed differentiation was labelled in the intervals: (▲) 0.010–0.029 (6), (●) 0.030–0.049 (15), (■) 0.050–0.069 (7), (△) 0.080–0.139 (11), (○) 0.140–0.199 (11) and (□) 0.200–0.259 (6). The filled symbols refer to simulated hybrids and the open symbols refer to pure breeds. The number of data points used for each differentiation interval is given in parenthesis.

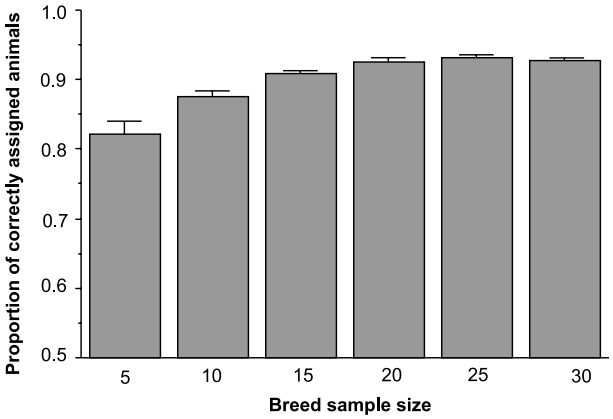

Figure 2 The proportion of correctly assigned individuals as a function of breed sample size, scoring 10 sets of loci each consisting of 10 random microsatellites. Standard errors are given as vertical bars.

Effect of breed sample size

The assignment precision depended on how many individuals that were regarded to represent the source population ($\chi^2 = 38.3$, d.f. = 5, $P < 0.0001$), with lower assignment precision for the smallest sample sizes. But still, a relatively high assignment precision, 88–92%, was achieved when sampling only 10–15 individuals from each population and scoring 10 loci (Fig. 2). The assignment precision was not different when analysing 20 animals per breed compared to 25 and 30 animals per breed ($\chi^2 = 0.39$, d.f. = 2, $P = 0.82$).

Effect of locus variability

The assignment error rate of single microsatellites varied in the range from 0.48 to 0.80 (Table 1). The locus error rate was negatively correlated both to the number of alleles within the locus (Fig. 3, $r_{sp} = -0.47$, $P = 0.01$), and to locus F_{ST} ($r_{sp} = -0.42$, $P = 0.03$). The significance of these correlations disappeared when the two microsatellites with the highest error rates (*NVHEQ*5 and *NVHEQ*54) were removed from the analysis, but a similar tendency persisted [$r_{sp} = -0.33$, $P = 0.12$ (number of alleles), $r_{sp} = -0.30$, $P = 0.15$ (F_{ST})]. No significant correlation was detected between locus heterozygosity and locus error rate ($r_{sp} = -0.23$, $P = 0.26$).

Discussion

The horse breeds included in this study show distinct breed demarcation (Bjørnstad & Røed 2001), and this suggested that also breed crosses could be genetically distinguishable. But identifying hybrids can be complicated, because of their

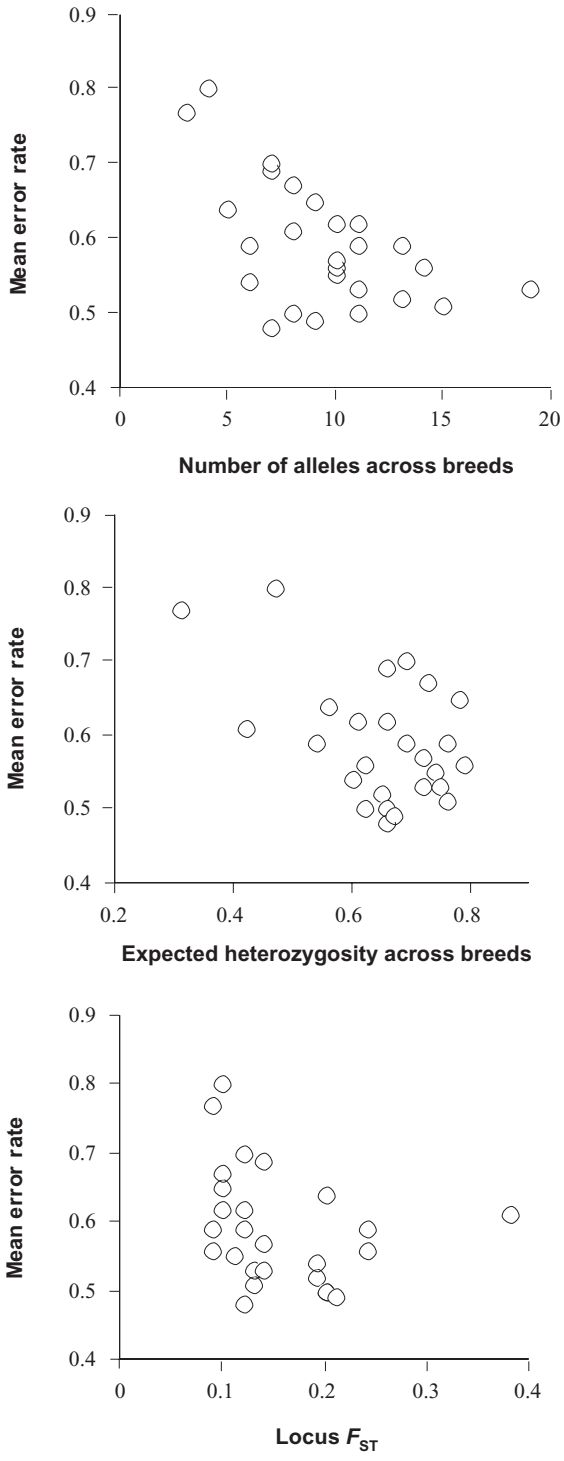

Figure 3 The mean error rate, defined as fails in the assignment trials, for the 26 microsatellites plotted against the number of alleles within the locus, heterozygosity and F_{ST} across breeds.

close relationship to both their parental breed. Using the empirical data from Bjørnstad & Røed (2001), we examined the different factors affecting the assignment precision particularly for pure breeds, but also for simulated breed crosses.

Quantifying the assignment precision across species suggests that correct breed designations can be inferred with accuracies of 90–95% using less than 10 microsatellites (e.g. sheep: Buchanan *et al.* 1994; cattle: MacHugh *et al.* 1998; Blott *et al.* 1999; horse: Bjørnstad & Røed 2001). A general pattern is a rapid increase in assignment precision when scoring the first loci. Then the assignment curves flatten out, and only a marginally higher proportion of the breeds can be assigned after a certain number of loci are scored. Why not all individuals can be correctly assigned when analysing a high number of loci seem to rely on some individuals being genetically atypical of their breeds, rather than analysis of too few loci, or limited polymorphic content within the loci. The genetic differentiation between the breeds affected the assignment precision, but this was most notable when scoring just a few loci for pure breeds, as also described in a simulation study (Cornuet *et al.* 1999). For breed crosses, the assignment precision increased at a much slower rate than compared with pure breeds, and the differentiation to their parental breeds was critical even when scoring a high number of loci. Thus, the misassignments of breed crosses are strongly related to their minor differentiation from their parental breeds.

Lod0, simply meaning that one population is a more likely source population that another, is the most frequently used assignment criterion (e.g. Buchanan *et al.* 1994; Paetkau *et al.* 1995; MacHugh *et al.* 1998), but using log-likelihood procedures the strength of the assignments could be measured (reviewed in Davies *et al.* 1999). In this study, application of strict assignment thresholds identified the majority of the pure-breds, even when scoring a relatively low number of loci. The ability to identify hybrids has implications for efficient conservation and management of populations. Using the relaxed criterion lod0, a high proportion of the present breed crosses could be identified, and for several crosses all the simulated individuals were recognized. But the hybrid assignments had considerably lower confidence than the pure-breds. Thus, strict thresholds, such as lod2 or larger, can be too rigid for practical implementation of assignment testing, because high assignment confidence will be achieved at the expense of the capacity to identify breed crosses. However, in situations where correct assignment is critical, such as in forensics, strict acceptance thresholds could be valuable (Shriver *et al.* 1997). It should be emphasized that the inclusion of many potential source breeds reduces the capacity to detect the breed of origin. Hence, the potential to recognize hybrids will primarily be based on suspicion of crossbreeding, with a limited number of possible source breeds.

Genetic markers varying greatly between populations but relatively little within them is expected to be best suited for assigning individuals. We found that the assignment error rates correlated negatively with locus F_{ST}. There was also a negative correlation between error rate and the number of alleles within a locus, but not with the heterozygosity. The heterozygosity for most of the loci included in this study was in the range between 0.6 and 0.8, and the heterozygosity may thus represent a limited expression of variability. Highly variable loci, in terms of high number of alleles and high heterozygosity, have also previously been reported to be most efficient in assignment testing (Blott *et al.* 1999) as well as in examining population structure using a related method to assignment testing based on extensive simulation without *a priori* population knowledge (Rosenberg *et al.* 2001). Hence, loci with high numbers of low frequency alleles do not increase the noise rather than the power, as has been suggested (Bowcock *et al.* 1994). It can be summarized that loci containing an intermediate to high amount of variability within and across populations would yield higher assignment precision than less polymorphic loci.

The question of how many animals have to be analysed per breed to achieve satisfactory assignment precision is of high importance, as both access to samples and the analysis costs can be limiting factors for population analyses. In this study, only very small sample sizes had a negative effect on the assignment precision. Even when sampling 10–15 individuals from a population an assignment precision of approximately 90% can be accomplished using 10 loci. The current data, consisting of fairly differentiated breeds [F_{ST}: 0.08–0.26], suggests that breed sample size is not a critical factor for the assignment precision as long as moderately large sample sizes ($\geq$ 20 animals per population) are used.

In summary, the present study showed that breed differentiation and number of scored loci were the most critical factors for assignment precision, and these factors were more critical for simulated breed crosses than for pure breeds. Hence, with known differentiation level between populations, the number of loci required for identifying a satisfactory proportion for both pure breeds and breed crosses could be predicted. Population sample size and locus variability were relatively less important for the assignment performance, as long as the parameters maintained a certain standard.

Acknowledgements

Alan Wilton, Gunnar Eie and two anonymous reviewers are thanked for comments on the manuscript. The study was

funded by the Norwegian Research Council and the Norwegian Trotting Association.

References

Banks M.A. & Eichert W. (2000) WHICHRUN (version 3.2): a computer program for population assignment of individuals based on multilocus genotype data. *Journal of Heredity* **91**, 87–9.

Bjørnstad G. & Røed K.H. (2001) Breed demarcation and potential for breed allocation of horses assessed by microsatellite markers. *Animal Genetics* **32**, 59–65.

Bjørnstad G., Gunby E. & Røed K.H. (2000a) Genetic structure of Norwegian horse breeds. *Journal of Animal Breeding and Genetics* **117**, 307–17.

Bjørnstad G., Midthjell L. & Røed K.H. (2000b) Characterization of ten equine dinucleotide microsatellite loci: *NVHEQ21*, *NVHEQ54*, *NVHEQ67*, *NVHEQ70*, *NVHEQ75*, *NVHEQ77*, *NVHEQ79*, *NVHEQ81*, *NVHEQ82* and *NVHEQ83*. *Animal Genetics* **31**, 78–9.

Blott S.C., Williams J.L. & Haley C.S. (1999) Discriminating among cattle breeds using genetic markers. *Heredity* **82**, 613–9.

Bowcock A.M., Ruiz-Linares A., Tomfohrde J., Minch E., Kidd J.R. & Cavalli-Sforza L.L. (1994) High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **368**, 455–7.

Buchanan F.C., Adams L.J., Littlejohn R.P., Maddox J.F. & Crawford A.M. (1994) Determination of evolutionary relationships among sheep breeds using microsatellites. *Genomics* **22**, 397–403.

Cornuet J.-M., Piry S., Luikart G., Estoup A. & Solignac M. (1999) New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics* **153**, 1989–2000.

Davies N., Villablanca F.X. & Roderick G.K. (1999) Determining the source of individuals: multilocus genotyping in nonequilibrium population genetics. *Trends in Ecology and Evolution* **14**, 17–21.

Estoup A., Rousset F., Michalakis Y., Cornuet J.-M., Adriamanga M. & Guyomard R. (1998) Comparative analysis of microsatellite and allozyme markers: a case study investigating microgeographic differentiation in brown trout (*Salmo trutta*). *Molecular Ecology* **7**, 339–53.

Favre L., Balloux F., Goudet J. & Perrin N. (1997) Female-based dispersial in the monogamous mammal *Crocidura russula*: evidence from field data and microsatellite patterns. *Proceedings of Royal Society of London B* **264**, 127–32.

Goudet J. (1995) FSTAT (version 1.2): a computer program to calculate *F*-statistics. *Journal of Heredity* **86**, 485–6.

Haig S.M., Gratto-Trevor C.L., Mullins T.D. & Colwell M.A. (1997) Population identification of western hemisphere shorebirds throughout the annual cycle. *Molecular Ecology* **6**, 413–27.

MacHugh D.E., Loftus R.T., Cunningham P. & Bradley D.G. (1998) Genetic structure of seven European cattle breeds assessed using 20 microsatellite markers. *Animal Genetics* **29**, 333–40.

Nielsen E.E., Hansen M.M. & Loeschcke V. (1997) Analysis of microsatellite DNA from old scale samples of Atlantic salmon *Salmo salar*: a comparison of genetic composition over 60 years. *Molecular Ecology* **6**, 487–92.

Paetkau D., Calvert W., Stirling I. & Strobeck C. (1995) Microsatellite analysis of population structure in Canadian polar bears. *Molecular Ecology* **4**, 347–54.

Primmer C.R., Aho T., Piironen J., Estoup A., Cornuet J.-M. & Ranta E. (1999) Microsatellite analysis of hatchery stocks and natural populations of Arctic charr, *Salvelinus alpinus*, from the Nordic region: implications for conservation. *Hereditas* **130**, 277–89.

Primmer C.R., Koskinen M.T. & Piironen J. (2000) The one that did not get away: individual assignment using microsatellite data detects a case of fishing competition fraud. *Proceedings of Royal Society of London B* **267**, 1699–704.

Rannala B. & Mountain J.L. (1997) Detecting immigration by using multilocus genotypes. *Proceedings of National Academy of Science USA* **94**, 9197–201.

Raymond M. & Rousset F. (1995) GENEPOP (version 1.2): population genetic software for exact tests and ecumenicism. *Journal of Heredity* **86**, 248–9.

Roques S., Duchesne P. & Bernatchez L. (1999) Potential of microsatellites for individual assignment: the North Atlantic redfish (genus *Sebastes*) species complex as a case study. *Molecular Ecology* **8**, 1703–17.

Rosenberg N.A., Burke T., Elo K. *et al.* (2001) Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. *Genetics* **159**, 699–713.

Shriver M.D., Smith M.W., Jin L., Marcini A., Akey J.M., Deka R. & Ferrell R.E. (1997) Ethnic-affiliation estimation by use of population-specific DNA markers. *American Journal of Human Genetics* **60**, 957–64.

Waser P.M. & Strobeck C. (1998) Genetic signatures of interpopulation dispersal. *Trends in Ecology and Evolution* **13**, 43–4.

Weir B.S. & Cockerham C.C. (1984) Estimating F-statistics for the analysis of population structure. *Evolution* **38**, 1358–70.