doi:10.1111/j.1365-2052.2008.01800.x

Assessing SNP markers for assigning individuals to cattle populations

R. Negrini*, L. Nicoloso†, P. Crepaldi†, E. Milanesi†, L. Colli*, F. Chegdani*, L. Pariset‡, S. Dunner§, H. Leveziel¶, J. L. Williams and P. Ajmone Marsan***

*Istituto di Zootecnica, Università Cattolica del S. Cuore, Piacenza, Italy. †Sezione di Zootecnica Agraria, Dipartimento di Scienze Animali, Università degli Studi di Milano, Milan, Italy. ‡Dipartimento di Produzioni Animali, Università della Tuscia, Viterbo, Italy. §Departamento de Producción Animal, Facultad de Veterinaria, Madrid, Spain. ¶Unité de Génétique Moléculaire Animale, UMR 1061, INRA/Université de Limoges, Faculté des Sciences et Techniques, Limoges Cedex, France. **Parco Tecnologico Padano, Polo Universitario, Lodi, Italy

Summary

The effectiveness of single nucleotide polymorphisms (SNPs) for the assignment of cattle to their source breeds was investigated by analysing a panel of 90 SNPs assayed on 24 European breeds. Breed assignment was performed by comparing the Bayesian and frequentist methods implemented in the STRUCTURE 2.2 and GENECLASS 2 software programs. The use of SNPs for the reallocation of known individuals to their breeds of origin and the assignment of unknown individuals was tested. In the reallocation tests, the methods implemented in STRUCTURE 2.2 performed better than those in GENECLASS 2, with 96% vs. 85% correct assignments respectively. In contrast, the methods implemented in GENECLASS 2 showed a greater correct assignment rate in allocating animals treated as unknowns to a reference dataset (62% vs. 51% and 80% vs. 65% in field tests 1 and 2 respectively). These results demonstrate that SNPs are suitable for the assignment of individuals to reference breeds. The results also indicate that STRUCTURE 2.2 and GENECLASS 2 can be complementary tools to assess breed integrity and assignment. Our findings also stress the importance of a high-quality reference dataset in allocation studies.

Keywords allocation method, cattle, SNPs, traceability.

Introduction

Allocation or assignment tests use genetic information to establish population membership of individuals, providing the most direct methods to determine the population of origin of unknown individuals. The interest in applying allocation methods has recently come from population genetic investigations, e.g. evaluating the amount of genetic exchange between populations (Cegelski *et al.* 2003), identifying immigrants (Castric & Bernatchez 2004) and detecting hidden population structures (Peter *et al.* 2006). These methods have practical applications, such as parentage analysis and tracing animals and animal products back to their breed of origin (Shackell *et al.* 2001; Dalvit *et al.* 2008a). The latter application has been promoted by the increasing market demand for comprehensive and integrated food safety policies.

The use of DNA molecular markers, especially microsatellites but also single nucleotide polymorphisms (SNPs), for individual identification and paternity testing (Koskinen 2003; Liron *et al.* 2004; Ayres 2005; Heaton *et al.* 2005) has been extensively investigated, but only a few studies have explored their practical applications for tracing meat or meat products at the breed level.

Towards this goal, DNA methods based either on deterministic or probabilistic approaches have already been proposed (Maudet *et al.* 2002; Ciampolini *et al.* 2006; Dalvit *et al.* 2008b). Deterministic approaches are based on the detection of breed-specific alleles, mainly within coat colour genes. However, until now, specific alleles permitted, at most, the discrimination of breed clusters (Maudet *et al.* 2002; Casellas *et al.* 2004). Probabilistic approaches mainly use multiallelic microsatellite markers and are costly because a relatively large number of markers must be

Address for correspondence

R. Negrini, Istituto di Zootecnica, Università Cattolica del S. Cuore, Piacenza, Italy.
E-mail: riccardo.negrini@unicatt.it

Accepted for publication 17 July 2008

analysed to assign an unknown sample correctly to one of a set of poorly differentiated populations such as cattle breeds (Ruzzante *et al.* 2001).

The bovine genome sequencing project and related projects have recently discovered millions of putative SNPs, some of which have already been validated (e.g. cattle SNP database, http://www.animalgenome.org/bioinfo/resources/util/q_bovsnp.html; dbSNP, http://www.ncbi.nlm.nih.gov/projects/SNP/; Werner *et al.* 2004). SNPs are polymorphisms in the DNA sequence occurring at appreciable frequencies in a population and are the most frequent type of mutations in the mammalian genome. Although less informative than multiallelic microsatellites, the diallelic SNPs possess considerable advantages, which include: (i) lower mutation rates, (ii) more robust genotyping and data interpretation (Krawczak 1999), (iii) suitability for standardized representation of genotyping results as digital DNA signatures (Fries & Durstewitz 2001) and (iv) suitability for various genotyping techniques and strong potential for automation (Lindblad-Toh *et al.* 2000). There is growing recognition that large collections of mapped SNPs would provide powerful tools for genetic studies.

In addition, specific statistical methods have been developed for the specific purpose of allocating genetically related individuals to clusters (Corander *et al.* 2003; Baudouin *et al.* 2004). In this study, the efficiency with which SNPs are able to allocate cattle individuals to their source breeds is evaluated. This is addressed by comparing the performance of different Bayesian and frequency-based allocation methods implemented in the software GENECLASS 2 and STRUCTURE 2.2 when applied to data from a panel of 90 SNPs typed on 24 dairy and beef breeds from Italy, France, Spain, Denmark and the UK.

Materials and methods

Selection of SNPs

From a panel of 701 SNPs identified in candidate genes for meat quality in cattle (GeMQual project QLRT-1999-30147), a subset of 97 SNPs in 73 different genes were selected (accession numbers ss77831721–ss77831810). Of these SNPs, 40 (41.2%) were located in introns, 41 (42.3%) in exons, 14 (14.3%) in 3′ or 5′ UTRs and 2 (0.02%) in promoter regions. Eighty-five mutations (87.6%) were transitions and 12 (12.3%) were transversions. Nine of the 41 SNPs in exons were non-synonymous.

Collection of biological samples, DNA extraction and SNP genotyping

Biological samples (whole blood and semen) from 249 minimally related animals belonging to 13 cattle breeds from Italy (Chianina = 19; Romagnola = 19; Marchigiana = 17; Piedmontese = 18; Maremmana = 22; Italian Red Pied = 23; Italian Brown = 21; Italian Holstein = 19) and France (Blonde d'Aquitaine = 19; Maine Anjou = 19; Limousin = 19; Salers = 20; Parthenais = 14) were collected. Genomic DNA was extracted using a commercial kit (GenElute™ Mammalian Genomic DNA kit; Sigma-Aldrich) following the manufacturer's instructions. The DNA was tested for quality and concentration by electrophoresis on 0.8% agarose gel, stained with ethidium bromide and compared to a commercial standard. Large-scale genotyping of all animals with the panel of 97 SNPs was performed by out-sourcing to a commercial genotyping company (http://www.kbioscience.co.uk).

Additional samples

To complete the dataset, we included genotype data of 558 individuals belonging to 11 other breeds (GeMQual project QLRT-1999-30147). These were from France (Charolais = 82), the UK (Jersey = 46; Highland = 46; South Devon = 35; Aberdeen-Angus = 38), Spain (Pirenaica = 71; Asturiana de la Montaña = 55; Avilena = 53; Asturiana de los Valles = 56) and Denmark (Simmental = 19; Red Cattle = 57). An additional cohort of 240 individuals from four Italian breeds (Italian Holstein-Gq = 58; Piedmontese-Gq = 67; Marchigiana-Gq = 38; Limousin-Gq = 77) were included for data quality control. The final dataset comprised a total of 1047 animals belonging to 24 breeds.

Statistical analysis

Summary statistics were calculated using POWERMARKER ver. 3.25 (http://www.powermarker.net). The unbiased estimator of gene diversity, often referred to as expected heterozygosity, was calculated according to the method of Weir (1996). All the possible allele-pair combinations were tested for linkage disequilibrium using the Lewontin D' method (Lewontin 1988), considering a threshold value of 0.7 (Khatkar *et al.* 2007).

Population subdivision was measured using F-statistics (Wright 1965) by calculating the amount of inbreeding-like effects within subpopulations (F_{IS} or f), among subpopulations (F_{ST} or θ) and within the entire dataset (F_{IT} or F). Hardy–Weinberg (HW) equilibrium across the complete data set was tested using the exact test (Guo & Thompson 1992). Deviation from HW equilibrium was considered significant only if detected in more than 50% of the populations. Reynolds genetic distances (Reynolds *et al.* 1983) between all pairs of breeds were calculated considering the double-sampled breeds as independent. The contribution of sampling size to breed allocation was estimated by bootstrapping using the software GENEDIST (A. Valentini, unpublished data). Average Jaccard similarities and coefficients of variation (C.V.) were calculated on 1000 bootstrapping replicates at each increasing step of five individuals. The minimum number of individuals that maintained the C.V. below 10% was selected as threshold.

Allocation test

The allocation tests were performed using the frequency-based method of Paetkau *et al.* (1995) and the Bayesian-based methods of Rannala & Mountain (1997; hereafter referred to as R&M), Baudouin & Lebrun (2001; hereafter referred to as B&L) and Pritchard *et al.* (2000). The first three algorithms were implemented in GENECLASS 2 and the latter was implemented in STRUCTURE 2.2. These software programs are freely available at http://www.montpellier.inra.fr/URLB/geneclass/geneclass.html and at http://pritch.bsd.uchicago.edu respectively. For GENECLASS 2, the probability of assignment was performed either by using the likelihood method without associated probabilities or by simulating 1000 individuals with a Markov chain (MC) re-sampling procedure, setting the type I errors to 0.05 (assignment threshold of score = 0.05; Piry *et al.* 2004). Five independent runs were compared. For STRUCTURE 2.2, a 20 000 initial burn-in was used to minimize the effect of the starting configurations, followed by 100 000 MC iterations, as recommended by Falush *et al.* (2007). Depending on the purpose, prior information on populations was (i) not included when STRUCTURE 2.2 was used as the unsupervised method, i.e. considering the genotyping data only, (ii) included when running STRUCTURE 2.2 as the supervised method and (iii) considered only for the individuals of reference populations when running the field tests. In all cases, five repeated runs were performed setting the parameter K equal to the number of reference populations.

The performance of the allocation algorithms was compared in terms of (i) sensitivity, calculated as the number of correct individuals allocated to breed j divided by the number of animals sampled from breed j, (ii) overall average assignment probability, calculated per breed as the average of the probability of any correct assignment and (iii) specificity, calculated as the number of correct assignments to breed j divided by the total (correct + incorrect) assignments to breed j. For the simulated field trial, the double-sampled breeds (Marchigiana, Piedmontese, Limousin and Italian Friesian) were split according to the sampling date and individuals from the two sampling periods were used alternatively as unknowns or reference populations.

Results

Summary statistics

The Lewontin D' measure of linkage disequilibrium was significant for seven SNP pairs ($D' > 0.7$, chi-squared probability > 0.9). One SNP of each pair was then removed from the dataset to meet the assumption of independence between loci. The final panel of SNPs assayed for breed traceability therefore comprised 90 independent polymorphisms in 72 genes. About 43% of these SNPs were found to be polymorphic in each of the 24 breeds investigated and about 76% were polymorphic in at least 20 breeds. Only one out of the 90 markers was polymorphic in four breeds only (Fig. 1).

The chi-squared goodness-of-fit test performed within populations indicated that 39 SNPs (43.3%) were not in HW equilibrium in one or more populations. However, none of the SNPs was out of HW equilibrium for more than five (20.8%) populations and hence all SNPs were retained in the dataset. The number of polymorphic SNPs per breed ranged from 69 in Highland to 88 in Marchigiana. Observed heterozygosity ranged from 0.22 in Highland to 0.31 in Marchigiana, Romagnola, Parthenaise and Red Cattle. All genetic variability parameters are reported in Table 1. No significant differences were detected between observed and expected heterozygosity.

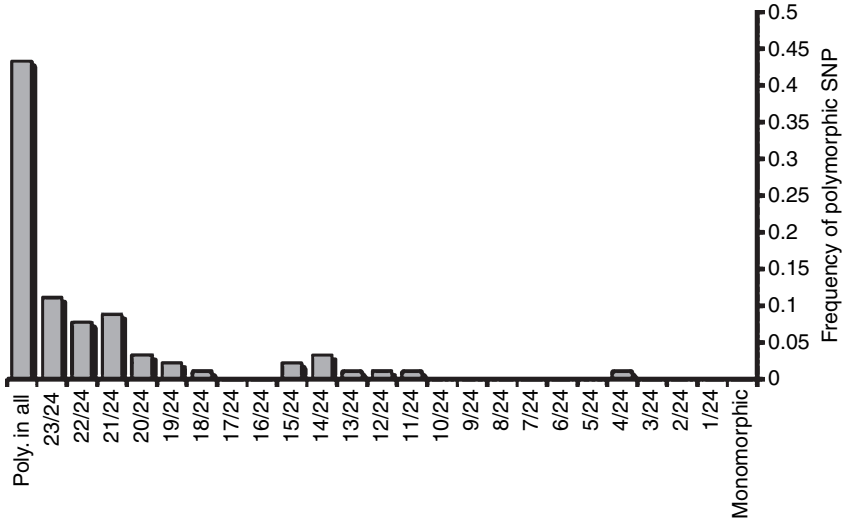

Figure 1 Distribution of polymorphic SNPs among breeds. Sixty-nine out of 90 SNPs were polymorphic in at least four breeds. One SNP was fixed in 20 breeds and polymorphic in four weakly related breeds: Marchigiana from Italy, Highland from the UK, Red Cattle from Denmark and Asturiana from Spain.

Genetic diversity

All F_{ST} values for pair-wise comparisons between different breeds were significant and ranged from 0.05 (Italian Red Pied vs. Simmental) to 0.26 (Jersey vs. Highland). Conversely, pair-wise genetic distances between double samplings of the same breed were very small and not significant (Limousin vs. Limousin-Gq = 0.022; Piedmontese vs. Piedmontese-Gq = 0.023; Italian Holstein vs. Italian Holstein-Gq = 0.024 and Marchigiana vs. Marchigiana-Gq = 0.028). Among the 24 breeds, Charolais, Piedmontese and Limousin had the lowest average genetic distance (0.09), whereas Highland, Jersey and Italian Holstein were the most distant breeds on average (0.19, 0.17 and 0.16 respectively). The F_{ST} value of 0.11 indicates that almost 90% of the total variability can be ascribed to the within-breed component, and that 11% separates the breeds.

Assignment test

The performance of STRUCTURE 2.2, when used as the 'unsupervised method' on the complete dataset, was very poor. In fact, on the basis of allele frequencies only, STRUCTURE 2.2 was able to allocate only 79 of 1047 individuals (7.5%) to their source breed with an individual probability of assignment higher than 50%. Those individuals belonged to the Highland (83% correct assignment), Jersey (46% correct assignment) and Aberdeen-Angus (29% correct assignment) breeds.

The assignment of individuals to their breed of origin using the supervised methods available in GENECLASS 2 and STRUCTURE 2.2 with prior information on breeds to assist clustering is reported in Table 2. Using prior information, the performance of STRUCTURE 2.2 increased significantly, with 96.1% (1006 of 1047) correct assignments at $P \geq 50\%$ and 90.4% (946 of 1047) unambiguous correct allocations at $P \geq 90\%$. The overall specificity was 99% and the average probability of assignment was 96%. About 20% (8 of 41) of the unassigned individuals were assigned to an incorrect breed with a probability higher than 50%. The other individuals remained unallocated.

Enabling the computations of probabilities associated with the assignment, the two Bayesian methods and the frequency-based methods available in GENECLASS 2 performed equally, with an overall sensitivity of 85% (887 out of 1047, considering an individual correctly assigned if its

Table 1 Descriptive statistics calculated at the breed level.

Breed	Country	Sample size	No. of loci[1]	No. of loci in HWd[2]	Availability[3]	H_E	H_O	F_{IS}[4]
Aberdeen-Angus	UK	38	78	2	0.94	0.29	0.29	−0.0015
Asturiana de los Valles	ES	56	87	5	0.97	0.30	0.29	0.0212
Avilena	ES	53	83	2	0.97	0.29	0.28	0.0262
Blonde d'Aquitaine	FR	19	78	3	0.98	0.27	0.28	0.0111
Italian Brown	IT	21	77	1	0.99	0.26	0.27	−0.0184
Asturiana de la Montaña	ES	55	84	2	0.96	0.28	0.28	−0.0024
Charolais	FR	82	86	2	0.96	0.30	0.30	0.0149
Chianina	IT	19	71	1	0.98	0.23	0.24	0.0204
Highland	UK	46	69	2	0.95	0.22	0.22	0.0282
Italian Friesian	IT	77	84	4	0.98	0.28	0.28	0.0001
Jersey	UK	46	80	4	0.94	0.27	0.28	0.0072
Limousin	FR	96	85	2	0.98	0.29	0.29	0.0088
Maine Anjou	FR	19	77	0	0.99	0.26	0.27	−0.0191
Marchigiana	IT	55	88	0	0.97	0.31	0.31	0.0191
Maremmana	IT	22	77	4	0.98	0.28	0.29	−0.0152
Parthenaise	FR	14	81	0	0.98	0.29	0.31	−0.0318
Italian Red Pied	IT	23	78	0	0.99	0.27	0.29	−0.0382
Piedmontese	IT	85	87	3	0.98	0.28	0.27	0.0272
Pirenaica	ES	71	84	2	0.97	0.29	0.29	−0.0101
Red Cattle	DK	57	84	3	0.97	0.31	0.31	0.0037
Romagnola	IT	19	82	0	0.98	0.29	0.31	−0.0420
Salers	FR	20	73	1	0.99	0.25	0.26	−0.0125
Simmental	DK	19	72	1	0.97	0.27	0.29	−0.0275
South Devon	UK	35	78	5	0.95	0.28	0.29	−0.0125

[1]Number of polymorphic loci within the breed.

[2]Number of loci in Hardy–Weinberg disequilibrium within the breed.

[3]Defined as 1 − *Obs*/*n*, where *Obs* is the number of observations and *n* is the number of individuals sampled.

[4]Inbreeding-like effects within the population.

Table 2 Number of animals sampled per breed and number of animals not correctly assigned, as well as sensitivity, specificity and average probability values calculated for each breed using different assignment methods.

		Baudouin & Lebrun (2001)				Rannala & Mountain (1997)				Paetkau *et al.* (1995)				Pritchard *et al.* (2000)			
Breed	No. of samples	Not correctly assigned	Sensitivity[1]	Specificity	Average probability	Not correctly assigned	Sensitivity	Specificity	Average probability	Not correctly assigned	Sensitivity	Specificity	Average Probability	Not correctly assigned	Sensitivity	Specificity	Average probability
Aberdeen-Angus	38	6	0.84	0.94	92.8	6	0.84	0.89	99.5	6	0.84	0.94	95.8	2	0.95	1.00	0.92
Asturiana de los Valles	56	18	0.68	0.67	84.9	18	0.68	0.70	88.1	18	0.68	0.63	87.2	3	0.95	1.00	0.91
Avilena	53	5	0.91	0.89	95.7	4	0.92	0.91	95.8	5	0.91	0.89	96.4	0	1.00	1.00	0.95
Blonde	19	6	0.68	0.50	84.5	6	0.68	0.52	86.1	6	0.68	0.52	87.8	2	0.89	1.00	0.9
Italian Brown	21	2	0.90	0.76	92.9	2	0.90	0.76	92.8	2	0.90	0.79	90.7	0	1.00	0.91	0.89
Asturiana de la Montaña	55	6	0.89	0.92	96.8	6	0.89	0.94	94.4	5	0.91	0.94	96.4	4	0.93	0.99	0.93
Charolais	82	12	0.85	0.84	95.1	14	0.83	0.82	95.5	12	0.85	0.83	95.6	5	0.94	0.99	0.99
Chianina	19	1	0.95	0.90	99.7	1	0.95	0.95	99.8	1	0.95	0.95	99.8	1	0.95	1.00	0.98
Italian Friesian	77	3	0.96	0.94	99.7	3	0.96	0.94	99.7	3	0.96	0.94	99.8	1	0.99	1.00	0.97
Highland	46	0	1.00	0.98	99.9	0	1.00	0.98	99.9	0	1.00	0.98	99.9	0	1.00	1.00	0.98
Jersey	46	0	1.00	1.00	99.1	0	1.00	1.00	99.0	0	1.00	1.00	98.9	1	0.98	1.00	0.98
Limousin	96	22	0.77	0.85	83.4	22	0.77	0.84	86.1	22	0.77	0.85	88.8	1	0.99	0.99	0.95
Maine Anjou	19	2	0.89	1.00	91.8	3	0.84	0.89	94.4	4	0.79	0.94	96.3	2	0.89	1.00	0.93
Marchigiana	55	7	0.87	0.91	95.8	6	0.89	0.92	96.6	6	0.89	0.92	97.5	2	0.96	0.95	0.94
Maremmana	22	1	0.95	0.78	98.8	1	0.95	0.78	99.5	1	0.95	0.78	97.4	1	0.95	1.00	0.91
Italian Red Pied	23	6	0.74	0.61	91.4	6	0.74	0.61	91.8	7	0.70	0.59	93.8	1	0.96	1.00	0.94
Parthenaise	14	6	0.57	0.73	84.1	6	0.57	0.57	86.4	6	0.57	0.62	87.8	9	0.36	1.00	0.87
Piedmontese	85	18	0.79	0.86	92.7	16	0.81	0.86	89.6	17	0.80	0.85	91.1	0	1.00	1.00	0.93
Pirenaica	71	13	0.82	0.84	91.2	14	0.80	0.83	88.0	14	0.80	0.84	91.6	3	0.96	1.00	0.96
Red Cattle	57	13	0.77	0.90	92.3	13	0.77	0.90	93.2	13	0.77	0.88	93.7	1	0.98	0.98	0.93
Romagnola	19	2	0.89	0.77	95.2	2	0.89	0.74	94.4	3	0.84	0.76	95.6	0	1.00	1.00	0.89
Salers	20	3	0.85	0.74	89.4	3	0.85	0.74	91.0	3	0.85	0.77	90	0	1.00	1.00	0.91
Simmental	19	6	0.68	0.54	89.2	6	0.68	0.62	86.9	6	0.68	0.59	88.4	2	0.89	1.00	0.93
South Devon	35	1	0.97	0.97	99.1	1	0.97	1.00	99.0	1	0.97	0.94	98.7	0	1.00	1.00	0.96
Overall	104	159	0.85	0.83	93.15	159	0.85	0.82	93.6	161	0.85	0.82	94.125	41	0.96	0.99	0.93

probability was > 50%) and an average probability of assignment > 93%. The assignment computations without associated probabilities increased the sensitivity by < 1%.

The average specificity index (83%, 82% and 82% for B&L, R&M and Paetkau methods respectively) was also comparable.

A large proportion of the samples (86%, or 763 of 887) were correctly assigned with a probability greater than 90% and can therefore be considered as unambiguous. About 14% (124 of 887) were ambiguous, but always with a probability of correct assignment ≥ 50% and with a difference with the second-best assignment > 20%. Twenty-six of the 160 incorrect allocations (16%) were truly unassigned ($P < 50\%$ in each breed) and the remaining unambiguous allocations (134 individuals) were assigned by the three methods to a same incorrect breed.

The lowest sensitivity was observed in Parthenaise (57% and 35% with GENECLASS 2 and STRUCTURE 2.2 respectively) followed by Asturiana de los valles, Blonde d'Aquitaine and Simmental. The Jersey and Highland breeds were invariably assigned correctly by all methods. Specificity was the lowest for Blonde d'Aquitaine, Asturiana de los valles and Simmental and the highest (>98%) for Jersey and Highland. Sensitivity was highly correlated with both average Reynolds genetic distances and specificity ($r = 0.70$, $P < 0.01$; $r = 0.82$, $P < 0.01$ respectively).

Field trials were carried out by splitting the four double-sampled breeds (Marchigiana, Piedmontese, Limousin and Italian Holstein) according to the sampling date. The two batches of different sizes (71 and 241 individuals respectively) were then used alternately as reference populations and unknown samples to compare the performance of the R&M Bayesian approach of GENECLASS 2 with that of STRUCTURE 2.2. The results obtained by applying R&M and STRUTCURE 2.2 approaches are shown in Table 3.

Considering the complete dataset, ranking the 90 SNPs for their F_{ST} values and using only those with a value ≥0.1 (35 SNPs) allowed the correct assignment of 56.8% of the individuals; this is in comparison with the 85% reached with all 90 SNPs (R&M method).

Discussion

This study assessed the potential of SNP markers in combination with Bayesian and frequentist statistics to cluster groups of breeds and assign individuals to breed clusters using a dataset comprising more than a thousand purebred individuals from 24 European breeds genotyped for 90 independent SNPs. Only individuals registered in herd books were sampled and hence these should represent the most genetically typical examples of the breeds, although incorrect allocations caused by technical errors (e.g. in labelling or in pedigree records) cannot be ruled out. To avoid an upward bias in assignment success, we used the 'leave one out' procedure (Smouse & Chevillon 1988), whereby each individual to be reassigned was removed from the source population when estimating the allele frequencies. The results of such a reallocation procedure, together with two field tests, were used to estimate the efficiency of different assignment algorithms.

The unsupervised method of the STRUCTURE 2.2 software (i.e. considering the genetic data only), which is widely used to solve clustering problems (Lecis *et al.* 2006; Linz *et al.* 2007), fails to allocate most individuals to their breed of origin. As the likelihood of clustering individuals into single breeds is highly dependent on the between-breed

Table 3 Results obtained in the simulated field tests 1 and 2.

	Field test 1					Field test 2				
Breed	No. of samples	Not correctly assigned	Sensitivity[1]	Specificity	Average probability	No. of samples	Not correctly assigned	Sensitivity[1]	Specificity	Average probability
Rannala & Mountain (1997) method										
Marchigiana	38	13	0.66	0.96	0.97	18	3	0.83	1.00	0.93
Italian Friesian	58	4	0.93	1.00	0.97	19	1	0.95	1.00	0.99
Piedmontese	68	39	0.43	0.88	0.84	18	6	0.67	0.86	0.86
Limousin	77	42	0.45	0.90	0.96	19	5	0.74	0.93	0.92
Average	241[2]	98[2]	0.62	0.93	0.94	74[2]	15[2]	0.80	0.95	0.93
STRUCTURE 2.2 method										
Marchigiana	38	21	0.45	1.00	0.70	17	5	0.71	1.00	0.75
Italian Friesian	58	2	0.97	1.00	0.76	19	1	0.95	1.00	0.84
Piedmontese	67	46	0.31	1.00	0.73	18	11	0.39	1.00	0.64
Limousin	77	49	0.36	1.00	0.67	19	9	0.53	1.00	0.72
Average	240[2]	118[2]	0.51	1.00	0.72	73[2]	26[2]	0.65	1.00	0.74

[1]Proportion correctly assigned.

[2]Value obtained by summing the column.

component of genetic diversity, the animals that were correctly assigned belonged to the most genetically divergent breeds. Using STRUCTURE 2.2 as a supervised method (i.e. providing the putative breed of origin of each samples as prior), only 4% of the breed allocations were incorrect.

STRUCTURE 2.2 performed better than allocation algorithms available in GENECLASS 2, in terms of both sensitivity (96% vs. 85%) and specificity (99% vs. 83%). Interestingly, using both approaches together, the sensitivity increased to 97.5%. The average probability of correct allocation was, conversely, similar (93%) for all methods. Direct comparison of the methods should, however, account for the differences in the outputs of the two programs: STRUCTURE 2.2 provides the posterior probability that each individual belongs to each population, whereas GENECLASS 2 gives the proportions of the simulated samples (representative of the same population) the probability of which is less than or equal to that of the reference individual. Furthermore, GENECLASS 2 takes the reference allele frequencies as they are, whereas STRUCTURE 2.2 reconstructs ancestral allele frequencies on the basis of both reference and test samples.

In this study, the sensitivity values were higher than those in previous reports, where 80–90% of individuals belonging to five to 10 cattle or sheep breeds were correctly assigned using allele frequencies of 21 microsatellites (MacHugh *et al.* 1998; Diez-Tascon *et al.* 2000), and were comparable to values obtained by Maudet *et al.* (2002) and Ciampolini *et al.* (2006) in assignment studies on six French and four Italian breeds respectively. A panel of 90 SNPs, therefore, seems to be as efficient as 19–23 microsatellites markers. However, considering the larger number of breeds included in the dataset presented here, it is likely that discrimination among populations in a complex dataset will benefit from high-throughput SNP genotyping technology.

The French Parthenaise breed showed the lowest assignment rate across all methods, although it had an above-average mean pair-wise genetic distance (Parthenaise = 0.10; mean = 0.12). This may have been caused by the reduced sample size (14 individuals), which was inadequate to estimate precisely the allele frequencies necessary to define the breed. To obtain good breed definition, the minimum number of individual samples from each breed ranged from 20 to 30 for all breeds (data not shown). The STRUCTURE 2.2 software showed lower sensitivity, as it failed to assign more than 65% of individuals of Parthenaise; this is in comparison to the 40% of failed assignments by GENECLASS 2 algorithms.

In addition to Parthenaise, Blonde d'Aquitaine, Asturiana de los valles and Simmental were poorly assigned to a breed cluster, probably because of their lower-than-average mean pair-wise genetic distance (Parthenaise = 0.10; Blonde d'Aquitaine = 0.11; Simmental = 0.11 and Asturiana de los valles = 0.08; average = 0.125). Indeed, the sensitivity results were highly correlated with average genetic diversity ($r = 0.70$; $P < 0.01$) in both the frequency and Bayesian methods, as already observed by Negrini *et al.* (2007) using biallelic dominant AFLP markers.

As expected, sensitivity and specificity also depend on the individual assignment probability thresholds. Increasing the threshold results not only in increasing specificity but also in decreasing sensitivity (with R&M algorithm in GENECLASS 2, probability threshold 70%: sensitivity 80% and specificity 88%; probability threshold 85%: sensitivity 74% and specificity 93%).

In field trials, as expected, all methods produced the best results when the smallest batch size of animals was treated as unknown and assigned to the reference population (the largest batch), which was evident in terms of sensitivity (80% vs. 62% with GENECLASS 2; 65% vs. 51% with STRUCTURE 2.2), but not specificity. It is likely that widening the genetic base of the reference populations (i.e. increasing the number of individuals) favours the assignment of samples with atypical genotype profiles. Interestingly, in the simulated field test, the GENECLASS 2 algorithm was more efficient than STRUCTURE 2.2, giving about 15% more correct assignments, even if the performance of this software in the reallocation procedures was significantly lower (85% vs. 96% of correct assignment respectively). The performance of the two methods in the reallocation tests was probably affected by the different prior assumed in modelling the allele frequencies: STRUCTURE 2.2 assumed the same prior for all genotypes, whereas the R&M algorithms assumed higher prior probability values for the homozygote genotypes (Baudouin *et al.* 2004). Apparently, the higher stringency in selecting reference individuals employed by GENECLASS 2 during the reallocation phase produced a better reference dataset and resulted in a higher allocation rate of field samples. Moreover, the R&M approach may benefit by allowing for departures from HW equilibrium; in contrast, STRUCTURE 2.2 relies on a theoretical genetic model that includes HW equilibrium among its assumed parameters. The combinations of the two methods increased the percentage of correct assignment in field test 1 (using the small batch size as reference and big batch as unknown) on average by 2%.

The effect of reducing the number of markers without affecting the assignment power was also tested. The SNPs were ranked using the F_{ST} index and considering only 35 markers, which had values > 0.1. The results obtained here indicate that the selection of SNPs with the highest F_{ST} values allowed the use of a significantly smaller number of loci, with an acceptable loss in assignment rate.

In conclusion, the results presented here demonstrate the effectiveness of SNP markers for identifying the source breed of individuals of unknown origin. Although tested in a complex dataset and in situations of poor genetic differentiation, the number of correct allocations using a set of 90 SNPs was rather large. The methods allowed the

precise allocation of more than 90% of individuals on average. Selecting markers with large F_{ST} values in the populations analysed provided the greatest discriminating power.

Comparing different analysis methods, the supervised option of the STRUCTURE 2.2 algorithm performed better than the Bayesian and frequentist approaches implemented in GENECLASS 2 in reallocating individuals to their source breeds. However, the latter methods showed greater sensitivity in allocating unknown samples to a predefined reference dataset. In our view, the two methods can be considered complementary: the methods of R&M, B&L and Paetkau *et al.* (1995) require a shorter calculation time, are easier to use and are based on true population data. This indicates that they are most suitable for routine applications. In contrast, in complex situations where the presence of hybrids is suspected, more elaborate calculation techniques based on the Markov chain Monte Carlo approach can be useful. Using both STRUCTURE 2.2 and GENECLASS 2, the percentage of correct assignment increased to 97.5% and 67.5% in the reallocation assays and field test 1 respectively, indicating that improved statistical tools would increase the power of this approach.

The decreasing cost of SNP analysis and the recent availability of many thousands of validated SNPs in cattle will permit the selection of a large informative set of markers and may increase their applicability to molecular tracing in the field. Finally, our results stress the importance of using a reference dataset composed of a sufficient number of individuals possibly sampled in their region of origin and accurately selected, considering all information available (e.g. pedigree and agreement with breed standards).

Acknowledgements

The authors acknowledge the Eu-TRACE project (contract no. 006942; http://www.trace.eu.org) and the Fondazione CARIPLO (no. 2003.0721/11.8094) for financial support, as well as the Eu-GeMQual project QLRT-1999-30147 for providing some of the data. The authors also acknowledge the Italian Breeders Association for the provision of samples and pedigree information. The content of this article does not necessarily reflect the view of the European Commission or its services.

References

Ayres K.L. (2005) The expected performance of single nucleotide polymorphism loci in paternity testing. *Forensic Science International* **154,** 167–72.

Baudouin L. & Lebrun P. (2001) An operational Bayesian approach for the identification of sexually reproduced cross-fertilized populations using molecular markers. *Acta Horticulturae* **546,** 81–94.

Baudouin L., Piry S. & Cornuet J.M. (2004) Analytical Bayesian approach for assigning individuals to populations. *Journal of Heredity* **95,** 217–24.

Casellas J., Jimenez N., Fina M., Tarres J., Sanchez A. & Piedrafita J. (2004) Genetic diversity measures of the bovine Alberes breed using microsatellites, variability among herds and types of coat colour. *Journal of Animal Breeding and Genetics* **121,** 101–10.

Castric V. & Bernatchez L. (2004) Individual assignment test reveals differential restriction to dispersal between two salmonids despite no increase of genetic differences with distance. *Molecular Ecology* **13,** 1299–312.

Cegelski C.C., Waits L.P. & Anderson N.J. (2003) Assessing population structure and gene flow in Montana wolverines (*Gulo gulo*) using assignment-based approaches. *Molecular Ecology* **12,** 2907–18.

Ciampolini R., Cetica V., Ciani E. *et al.* (2006) Statistical analysis of individual assignment tests among four cattle breeds using fifteen STR loci. *Journal of Animal Science* **84,** 11–19.

Corander J., Waldmann P. & Sillanpaa M.J. (2003) Bayesian analysis of genetic differentiation between populations. *Genetics* **163,** 367–74.

Dalvit C., De Marchi M., Dal Zotto R., Gervaso M., Meuwissen T. & Cassandro M. (2008a) Breed assignment test in four Italian cattle breeds. *Meat Science* **80,** 389–95.

Dalvit C., De Marchi M., Targhetta C., Gervaso M. & Cassandro M. (2008b) Genetic traceability of meat using microsatellite markers. *Food Research International* **41,** 301–7.

Diez-Tascon C., Littlejohn R.P., Almeida P.A. & Crawford A.M. (2000) Genetic variation within the Merino sheep breed, analysis of closely related populations using microsatellites. *Animal Genetics* **31,** 243–51.

Falush D., Stephens M.W. & Pritchard J.K. (2007) Inference of population structure using multilocus genotype data, dominant markers and null alleles. *Molecular Ecology Notes* **7,** 574–8.

Fries R. & Durstewitz G. (2001) Digital DNA signatures for animal tagging. *Nature Biotechnology* **19,** 508.

Guo S.W. & Thompson E.A. (1992) Performing the exact test of Hardy–Weinberg proportion for multiple alleles. *Biometrics* **48,** 361–72.

Heaton M.P., Keen J.E., Clawson M.L., Harhay G.P., Bauer N., Shultz C., Green B.T., Durso L., Chitko-McKown C.G. & Laegreid W.W. (2005) Use of bovine single nucleotide polymorphism markers to verify sample tracking in beef processing. *Journal of the American Veterinary Medical Association* **226,** 1311–4.

Khatkar M.S., Zenger K.R., Hobbs M. *et al.* (2007) A primary assembly of a bovine haplotype block map based on a 15k SNP panel genotyped in Holstein–Friesian cattle. *Genetics* **176,** 763–72.

Koskinen M.T. (2003) Individual assignment using microsatellite DNA reveals unambiguous breed identification in the domestic dog. *Animal Genetics* **34,** 297–301.

Krawczak M. (1999) Informativity assessment for biallelic single nucleotide polymorphisms. *Electrophoresis* **20,** 1676–81.

Lecis R., Pierpaoli M., Biro Z.S., Szemethy L., Ragni B., Vercillo F. & Randi E. (2006) Bayesian analyses of admixture in wild and domestic cats (*Felis silvestris*) using linked microsatellite loci. *Molecular Ecology* **15,** 119–31.

Lewontin R.C. (1988) On measures of gametic disequilibrium. *Genetics* **120,** 849–52.

Lindblad-Toh K., Winchester E., Daly M.J. *et al.* (2000) Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nature Genetics* **24,** 381–6.

Linz B., Balloux F., Moodley Y. *et al.* (2007) An African origin for the intimate association between humans and *Helicobacter pylori*. *Nature* **445,** 915–8.

Liron J.P., Ripoli M.V., Garcia P.P. & Giovambattista G. (2004) Assignment of paternity in a judicial dispute between two neighbour Holstein dairy farmers. *Journal of Forensic Science* **49,** 96–8.

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.

Maudet C., Luikart G. & Taberlet P. (2002) Genetic diversity and assignment tests among seven French cattle breeds based on microsatellite DNA analysis. *Journal of Animal Science* **80,** 942–50.

Negrini R., Milanesi E., Colli L., Pellecchia M., Nicoloso L., Crepaldi P., Lenstra J.A. & Ajmone-Marsan P. (2007) Breed assignment of Italian cattle using biallelic AFLP markers. *Animal Genetics* **38,** 147–53.

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

Peter C., Bruford M., Peretz T., Dalamitra S., Hewitt G. & Erhardt G. (2006) The ECONOGENE Consortium Genetic diversity and subdivision of 57 European and Middle-Eastern sheep breeds. *Animal Genetics* **38,** 37–44.

Piry S., Alapetite A., Cornuet J.M., Paetkau D., Baudouin L. & Estoup A. (2004) GENECLASS 2, a software for genetic assignment and first-generation migrant detection. *Journal of Heredity* **95,** 536–9.

Pritchard J.K., Stephens M. & Donnelly P. (2000) Inference of population structure using multilocus genotype data. *Genetics* **155,** 945–59.

Rannala B. & Mountain J.L. (1997) Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of Sciences of the United States of America* **94,** 9197–221.

Reynolds J., Weir B.S. & Cockerham C.C. (1983) Estimation of the Coancestry coefficient, basic for a short-term genetic distance. *Genetics* **105,** 767–79.

Ruzzante D.E., Hansen M.M. & Meldrup D. (2001) Distribution of individual inbreeding coefficients, relatedness and influence of stocking on native anadromous brown trout (*Salmo trutta*) population structure. *Molecular Ecology* **10,** 2107–28.

Shackell G.H., Tate M.L. & Anderson R.M. (2001) Installing a DNA-based traceability system in the meat industry. *Proceedings of the Association for the Advancement of Animal Breeding and Genetics* **14,** 533–6.

Smouse P.E. & Chevillon C. (1988) Analytical aspects of population-specific DNA fingerprinting for individuals. *Journal of Heredity* **89,** 143–50.

Weir B.S. (1996) *Genetic Data Analysis II*. Sinauer Associates, Inc., Sunderland, MA.

Werner F.A., Durstewitz G., Habermann F.A. *et al.* (2004) Detection and characterization of SNPs useful for identity control and parentage testing in major European dairy breeds. *Animal Genetics* **35,** 44–9.

Wright S. (1965) The interpretation of population structure by *F*-statistics with special regard to systems of mating. *Evolution* **19,** 395–420.