Poorer resolution of low-density SNP vs. STR markers in reconstructing genetic relationships among seven Italian sheep breeds

E. CIANI1, F. CECCHI2, E. CASTELLANA1, M. D’ANDREA3, C. INCORONATO4, F. D’ANGELO5, M. ALBENZIO5, F. PILLA3, D. MATASSINO4, D. CIANCI1, R. CIAMPOLINI2

1 Dipartimento di Bioscienze, Biotecnologie e Biofarmaceutica, Università degli Studi di Bari, 70126 Bari, Italy
2 Dipartimento di Scienze Veterinarie, Università degli Studi di Pisa, 56124 Pisa, Italy
3 Dipartimento di Scienze Animali Vegetali e dell’Ambiente, Università degli Studi del Molise, 86100 Campobasso, Italy
4 Consorzio per la Sperimentazione, Divulgazione e Applicazione di Biotecniche Innovative (ConSDABI)
Sub National Focal Point Italiano FAO Biodiversità Mediterranea, 82100 Benevento, Italy
5 Dipartimento di Scienze delle Produzioni e dell’Innovazione nei Sistemi Agro-alimentari Mediterranei, Università degli Studi di Foggia, 71100 Foggia, Italy

SUMMARY

Single Nucleotide Polymorphisms (SNPs) are increasingly used in livestock to infer genetic variability and relationship within and among breeds. Since displaying different mutational properties, bi-allelic markers could reconstruct dissimilar scenarios with respect to long-established Short Tandem Repeats (STRs). Here we present the results from the genetic characterization of seven Italian breeds using 111 SNPs, compared to those previously obtained from 19 STR loci typed on the same population sample (Sarda: 97; Comisana: 67; Altamurana: 83; Leccese: 86; Gentile di Puglia: 74; Bagnolese: 33; Laticauda: 30). All the 111 SNP loci resulted to be polymorphic in the total sample, with only six loci being monomorphic in at least one breed. After pruning of loci displaying MAF<0.1, a total of 103 SNPs were left in the overall sample dataset. The overall SNP gene diversity was 0.40 (±0.09). A remarkably low proportion of SNPs (on average 2.2%) displayed significant (P<0.01) deviations from the Hardy–Weinberg equilibrium within the seven breeds. The Leccese breed displayed not only the highest levels of Hardy-Weinberg and gametic disequilibrium (4.9% of SNPs), but also the highest inbreeding coefficient (FIS=0.05) and the highest within-breed allele-sharing distance (D=0.31). These results are consistent with the evidence of sub-structuring within the Leccese breed, as suggested by both the STRUCTURE analysis and the NJ tree based on the inter-individual allele-sharing distance. In addition, the analysis among-breeds highlighted a marked differentiation of Sarda, likely consistent with its insular nature. A more complex scenario was observed when looking at the ability of SNP in reconstructing genetic connections among Italian sheep breeds in comparison to what obtained previously with STR. SNP generally provided poorer resolution in reconstructing the genetic relationships among the seven Italian sheep breeds, likely due to their limited number. Interestingly, the study produced also a first clue of overdominance as shaping the pattern of variability at olfactory receptor genes in the ovine species.

KEY WORDS

Genetic variability, Microsatellite, Single Nucleotide Polymorphisms.

INTRODUCTION

Analysis of genetic variability is essential to preserve and exploit biodiversity. At the molecular level, microsatellite polymorphisms (Short Tandem Repeats, STRs) are by far the most employed loci to investigate biodiversity in livestock, including sheep, but more recently SNPs have also been introduced. However STRs and SNPs have different mutational properties, so, for a more effective characterization of genetic variability, they can be used in a complementary way. Studies carried out by high-throughput SNP genotyping in human datasets previously genotyped at hundreds of STRs highlighted how the fraction of diversity attributable to various hierarchical components of population structure (within populations, between populations within geographic regions, and between geographic regions) differs between STRs and SNPs. By comparing loci from the same genomic regions, Payseur and Jing revealed substantial differences in levels of structure in human populations also at linked STRs and SNPs, thus attributing these differences primarily to variation in the mutational process of the two marker classes. In addition, they highlighted the influence of population divergence time on differences in population structure between STRs and SNPs, suggesting that the informativeness gap between SNPs and STRs increases as divergence times between taxonomic units decrease. In a previous paper, the molecular characterization of seven Italian sheep breeds carried out using STR markers has been discussed. Here we present the results from the genetic characterization of the same breed dataset using 111 SNPs, together with a first evidence of overdominance at sheep olfactory receptor genes.
MATERIAL AND METHODS

Animals and breeds
A total of 470 animals, representative of seven breeds from southern and insular Italy (Sarda, SAR: 97; Comisana, COM: 67; Altamurana, ALT: 83; Leccese, LEC: 86; Gentile di Puglia, GEN: 74; Bagnolesi, BAG: 33; Laticauda, LAT: 30) were sampled from different flocks trying to avoid closely related individuals. In particular, from 3 (Altamurana) to 6 (Leccese and Gentile di Puglia) farms were considered for the small local sheep populations while samples from the more represented Sarda and Comisana were supplied by national breeding stations to guarantee a wider territorial coverage.

Sarda is a high milk-yielding breed indigenous of the Sardinian island, thought to derive from the Sardinian mouflon. After a long-lasting period of very strict isolation, half a century ago it began to spread out over the Italian peninsula replacing low-producing local sheep breeds; it counts at present more than 3 million heads, mainly located in Sardinia, Tuscany and Latium.

Comisana originated at the beginning of the XX century in southeast Sicily by crossbreeding local ewes with rams from Malta and North Africa; it is nowadays spread throughout the peninsula though being most concentrated in Southern Italy where the breed is preferred to Sarda for its rusticity and excellent adaptation to harsh and semi-arid climate conditions, associated to a valuable milk productivity also when reared under extensive or semi-extensive systems.

Out of the three autochthonous Apulian breeds, two (Altamurana and Leccese) are considered as belonging to the Zackel group of sheep and display a prevailing attitude toward milk production while Gentile di Puglia, also known as Apulian Merino, is a fine-wool sheep whose true origin is still object of debate. All these breeds have suffered a severe numerical reduction in the last fifty years, more pronounced for Altamurana, whose population size dropped from 140,000 heads in 1960s to very few hundred heads at the present days. Laticauda and Bagnolesi, both autochthonous of Campania, are usually reared by small family farms under semi-extensive systems and are thought to derive by crossbreeding local sheep from the Apennines with fat-tailed North African sheep, likely imported under the Bourbons dynasty in the XVIII century. Laticauda has a prevailing attitude toward meat production, associated to a good prolificacy, while Bagnolesi is a dual-purpose breed, particularly valued for the excellent quality of dairy products derived from its milk.

SNP analysis
Genomic DNA was extracted from peripheral blood samples following standard protocols. Animals were genotyped using 111 ovine SNPs selected among 1.5K markers from the ISGC pilot sheep SNP array (http://www.sheephapmap.org/). The full list of the selected 111 SNPs is available upon request to the authors.

Minor Allele Frequency (MAF) and pair-wise gametic disequilibrium were evaluated using PowerMarker v. 3.259. Deviations from the Hardy-Weinberg equilibrium (HWE) were evaluated using the ARLEQUIN package v. 3.1.10. The genetic distance (D) between all pair-wise combinations of individuals was calculated, using PLINK v 1.01, as one minus the average proportion of alleles shared identical by state and then used to construct a NJ tree via MEGA v.4.11.

In order to take into account more complex evolutionary relationships (reticulated evolution) among the studied breeds, network analysis using Reynolds genetic distance was also carried out by the software Neighbor-net. Breed differentiation was also investigated using the Bayesian clustering algorithm implemented in the STRUCTURE software v. 2.2. The ad hoc model, without providing a priori information on population membership, with 5 independent runs was performed for each K value from 2 to 10 (where K is the number of clusters to be tested), adopting a burn-in period of 100,000 generations, followed by 100,000 iterations. The algorithm of Evanno et al. was adopted in order to evaluate the most probable value of K. Only loci displaying MAF ≥ 0.1 were used in all the above analyses.

RESULTS

Whitin-breed diversity
All the 111 SNP loci resulted to be polymorphic in the total sample, with only six loci being monomorphic in at least one breed (DU179443, DU216500, DU222192, DU355225, DU465451, DU485358; data not shown). After pruning of loci displaying MAF<0.1, a total of 103 SNPs were left in the overall sample dataset (data not shown).

The overall SNP gene diversity (0.40±0.09; data not shown) was appreciably lower than that estimated previously at STR loci (0.79±0.08) by Ciani et al. (2013) on the same samples, likely due to the bi-allelic nature of SNPs. Within-breed gene diversity ranged from 0.37 (ALT, BAG, LAT) to 0.40 (GEN) and the average observed heterozygosity ranged from 0.36 (LEC) to 0.41 (LAT) (Table 1).

A remarkably low proportion of SNPs (on average 2.2%) displayed significant (P<0.01) deviations from the Hardy-Weinberg equilibrium within the seven breeds (data not shown), with LEC exhibiting the highest number of loci (4.9% of SNPs) in significant departure from HWE. Interestingly, all the seven breeds displayed a significant (P<0.01) departure from HWE due to excess of heterozygous genotypes at locus DU526865 (OAR15) (Table 2).

The squared allele-frequency correlation, r², as a measure of gametic disequilibrium, was assessed for the 5,253 pair-wise combinations possible with 103 bi-allelic loci. SNP r² values in the total sample were generally higher than STR values, with a maximum value of 0.64 (data not shown). Within-breed percentages of locus pairs with significant P values were generally consistent with those observed at STR loci, highlighting LEC as the breed displaying the highest values and BAG and LAT as the breeds displaying the lowest values (data not shown).

When estimating inbreeding, lower FIS values were observed compared to STR loci, the highest being observed in LEC (0.05, Table 1). The allele-sharing distance (D) between all possible individual pairs within each breed highlighted LAT as the less differentiated breed (0.26), while the most differentiated breeds resulted to be LEC and GEN (0.31, Table 1).

Genetic relationships among breeds
The neighbour-net networks based on the Reynolds distance (Figure 1) highlighted a poor structure resolution (star-like topology with internal reticulations), with a slightly more
Table 1 - Within-breed genetic diversity parameters.

<table>
<thead>
<tr>
<th>Breed</th>
<th>N</th>
<th>Region of origin</th>
<th>$H_o^*$</th>
<th>$H_e^*$</th>
<th>$F_{IS}^*$</th>
<th>D*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Altamura 83</td>
<td>Apulia</td>
<td>0.38</td>
<td>0.37</td>
<td>-0.01</td>
<td>0.29</td>
<td></td>
</tr>
<tr>
<td>Bagnoiese 33</td>
<td>Campania</td>
<td>0.37</td>
<td>0.37</td>
<td>-0.01</td>
<td>0.28</td>
<td></td>
</tr>
<tr>
<td>Comisana 67</td>
<td>Sicily</td>
<td>0.38</td>
<td>0.38</td>
<td>0.01</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>Gentile di Puglia 74</td>
<td>Apulia</td>
<td>0.39</td>
<td>0.40</td>
<td>0.02</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>Laticauda 30</td>
<td>Campania</td>
<td>0.41</td>
<td>0.37</td>
<td>-0.11</td>
<td>0.26</td>
<td></td>
</tr>
<tr>
<td>Leccese 86</td>
<td></td>
<td>0.36</td>
<td>0.38</td>
<td>0.05</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>Sarda 97</td>
<td>Sardinia</td>
<td>0.38</td>
<td>0.38</td>
<td>0.00</td>
<td>0.29</td>
<td></td>
</tr>
</tbody>
</table>

* $H_o$, observed heterozygosity; $H_e$, expected heterozygosity; $F_{IS}$, inbreeding coefficient; D, allele sharing distance.

Table 2 - Observed ($H_o$) and expected ($H_e$) heterozygosity for the locus DU528685 (OAR15).

<table>
<thead>
<tr>
<th>Breed</th>
<th>$H_o$</th>
<th>$H_e$</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Altamura 0.52</td>
<td>0.39</td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>Bagnoiese 0.55</td>
<td>0.40</td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>Comisana 0.75</td>
<td>0.47</td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>Gentile di Puglia 0.59</td>
<td>0.42</td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>Laticauda 0.58</td>
<td>0.42</td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>Leccese 0.70</td>
<td>0.46</td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>Sarda 0.61</td>
<td>0.43</td>
<td></td>
<td>*</td>
</tr>
</tbody>
</table>

*Hardy-Weinberg exact test P-value <0.01.

Figure 1 - Neighbor-Net network constructed using Reynolds distance.

pronounced differentiation of the Sarda breed. A rough breed resolution was also observed when plotting the NJ dendrogram based on the inter-individual allele-sharing distance (D) calculated by PLINK (data not shown), with the exception of the Sarda animals, well discriminated from the other subjects, and the Leccese animals, grouped into two different sub-clusters.

Breed differentiation was also investigated adopting a Bayesian clustering analysis carried out using the software STRUCTURE v. 2.2.15. The plot of the results is presented in Figure 2, where the number of sub-populations assumed in the total sample is indicated with K followed by the corresponding integer. In general, bi-allelic markers provided more noisy STRUCTURE patterns than STR markers. Notwithstanding, they confirmed the higher differentiation of Sarda, as highlighted both in Figure 2 (with Sarda clustering in a separate group already at K = 2) and in Table 3, where the proportions of genomic ancestry (Q) obtained at each K value for each breed are reported. As can be observed, at almost all the tested K values, Sarda was the best discriminated breed, though lower Q values were observed with SNPs compared to STRs. The second best discriminated breed was Altamura that displayed, at least at lower K values (≤ 4), Q values higher or nearly approaching those of Sarda. In addition, the STRUCTURE analysis (Figure 2) confirmed the presence of genetic sub-structuring within the Leccese breed sample previously observed by STR markers.

Table 3 - Average proportion of membership to assumed STRUCTURE clusters for each breed.

<table>
<thead>
<tr>
<th>Breed</th>
<th>Average Q values*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>K = 2</td>
</tr>
<tr>
<td>Altamura 0.95</td>
<td>0.87</td>
</tr>
<tr>
<td>Bagnoiese 0.89</td>
<td>0.44</td>
</tr>
<tr>
<td>Comisana 0.88</td>
<td>0.86</td>
</tr>
<tr>
<td>Gentile di Puglia 0.86</td>
<td>0.78</td>
</tr>
<tr>
<td>Laticauda 0.87</td>
<td>0.74</td>
</tr>
<tr>
<td>Leccese 0.91</td>
<td>0.43</td>
</tr>
<tr>
<td>Sarda 0.92</td>
<td>0.88</td>
</tr>
</tbody>
</table>

*For each K value, only data for the cluster showing the highest average Q value is reported.
DISCUSSION

Neither STR nor SNP used here were specifically selected for information content on the Italian populations under study. An SNP/STR ratio of about six was adopted, based on the main literature contrasting the two marker types\(^4, 17, 19, 20, 21\). Rosenberg et al.\(^22\) and Liu et al.\(^21\) showed that, in general, randomly chosen set of microsatellites provided greater informativeness to discriminate between human populations (respectively, 2.8 to 4.3 and 2.5 to 6.3 times) than a random chosen set of SNPs. Narum et al.\(^19\) found mean ratios of random SNPs to microsatellites in Chinook salmon (from 3.9 to 4.1) to be within the range of those previously shown in humans. Adopting a SNP to microsatellite ratio of 5.5, we observed a better discrimination ability of microsatellite markers versus SNP loci. The selection of highly diagnostic SNPs from larger panels has been proven to dramatically increase population differentiation, as demonstrated by Glover et al.\(^18\) in Atlantic salmon and by Lao et al.\(^23\) in a large panel from humans where the best 10 loci allowed to clearly assign all the individuals to the correct continental region of sampling. A similar scenario may be assumed for sheep, where the current availability of medium- to high-density SNP chips will allow to identify the most discriminating SNPs.

To date, only a very restricted number of studies have reported levels of genetic diversity in sheep using biallelic markers. Hoda et al.\(^2\) investigated three Albanian sheep breeds using 36 SNP markers originally developed by Pariset et al.\(^24\) in known sheep genes; Kijas et al.\(^1\) performed a genome-wide survey of SNP variation in sheep breeds sampled at a global scale using a 1.5K pilot SNP chip and Kijas et al.\(^2\) investigated the history of domestic sheep through a medium density SNP chip. In our study, on average, 99% of the 103 SNPs were polymorphic in all the seven breeds, with Sarda and Altamura displaying a fairly lower proportion of polymorphic loci (98 and 97%, respectively, data not shown). On the contrary, Sarda was among the breeds displaying the highest proportion of polymorphic SNPs (90%) in the work of Kijas et al.\(^1\) where contrasted with world-wide sheep breeds. The observed difference (90% vs. 98%) may be explained considering that we used a sub-set of SNPs selected from the 1.5K SNP chip adopting a criterion of heterozygosity maximization, therefore removing from our panel loci less likely to be polymorphic in our breeds. In fact, a higher proportion of loci (56%) displayed a high degree of polymorphism (MAF >0.30 in the total sample) compared to the proportion (45%) observed by Kijas et al. (2009)\(^1\). For a similar reason, when the 50K SNP chip was adopted on a set of 2,819 sheep, belonging to 74 breeds sampled from around the world (Kijas et al., 2012)\(^3\) the within-breed proportion of polymorphic loci never exceeded 96%, with generally higher values in Southern and Western European breeds.

The average (1-IBS) distances between all possible individual pairs within each of the seven breeds (data not shown) were higher than observed by Kijas et al.\(^1\) in 22 sheep populations typed at a 1.5K SNP chip, thus suggesting a more pronounced allelic variability within Italian breeds. However, average distance values at our 103 SNP panel resulted to be higher than those observed at the 1.5K SNP panel, also when considering a breed (Sarda) represented in both studies (0.294 vs. 0.267, respectively). This may be due to a markedly larger sample size of the breeds considered here compared to population samples of far less than 40 animals per breed adopted in the study of Kijas et al.\(^1\). Moreover, values observed in our survey may have been inflated by the SNP selection process, as already specified above.
Analysis of Hardy-Weinberg proportions confirmed SAR and GEN, respectively, as one of the less and the most variable breed. Interestingly, a significant excess of heterozygous genotypes was shared among the seven breeds at the locus DU526865 (OAR15). Heterozygote excess may be due to several factors, like balancing selection (overdominance), negative assortative mating, population mixing, CNVs, segmental duplications, pseudogenes. Out of all the loci displaying a significant (P<0.01) departure from HWE in at least one breed (n = 8), only a minor proportion (23%; n = 2) displayed heterozygote excess, therefore suggesting that this may not be due to factors acting on a genome-scale like negative assortative mating or population mixing. It must be pointed out that the SNP DU526865 is harboured in a region homologous to the human OR5W2 gene, known to be a functional olfactory receptor interestingly. By investigating the contribution of overdominance to the maintenance of polymorphism in the human genome using the HapMap genotypic data, Alonso et al.25 found that olfactory receptor activity is a molecular function enriched in genes with higher heterozygosity than expected, also when accounting for pseudogenes or functional genes not significantly expressed in olfactory epithelium. The maintenance of high variability in olfactory receptors has been observed so far also in primates, mouse and pig and would be related to the ability to discriminate among closely related structural odorants. Considering all the above evidences, the results presented in this work may be interpreted as a first clue to a role of overdominance in shaping the pattern of variability at olfactory receptor genes in the ovine species. Unfortunately, literature on inference of genetic relationships among sheep breeds by SNP markers is still very poor. Kijas et al.1 observed FST values comparable to those reported in the present study among breeds of European origin such as in the comparisons Sarda vs. Merino (FST = 0.053), vs. German Mountain Brown (FST = 0.070) and vs. Comisana (FST = 0.078) while obviously observing much higher FST values (>0.25) when comparing breed pairs sampled from different continents.

Interestingly, although SNPs generally depicted less defined genetic relationships among breeds than microsatellites6, they constituted the majority among the most informative markers ranked on the basis of FST (data not shown). This result is in agreement with those observed by Rosenberg et al.22, who found some SNP loci outperforming the average SNPs by a large margin. We hypothesize that this phenomenon may be due to the fact that the stronger “selection intensity” usually applied on the choice of SNP markers than microsatellites due to the availability of far larger panels of bi-allelic markers would facilitate detection of highly informative SNPs showing breed-specific allele frequency distributions. In addition, homoplasy may contribute to keep informativeness of the best-discriminating STR loci at lower levels than that of the best SNPs.

As anticipated, SNP produced poor network resolution compared to STR3, with patterns generally inconsistent with any previous knowledge about breed inter-relationships. These results do not confirm those of Smith et al.24, Ryynänen et al.23, Narum et al.21 and Glover et al.20 who observed generally comparable topologies adopting the two classes of markers in various salmon populations. On the contrary, patterns of genetic structure inferred by using the Bayesian approach implemented in the popular software package STRUCTURE resulted to be roughly consistent between the two classes of markers, with STR showing only slightly better discrimination ability than SNP loci.

The branching pattern of the NJ dendrogram based on the inter-individual allele-sharing distance (D, data not shown), well differentiated Sarda animals from the others and highlighted sub-clustering of the Lecceesi animals in two groups, grossly corresponding to animals belonging, respectively, to the old-type and to the improved Lecceesi breed. These results would support the oral historical knowledge that modern-day improved Lecceesi animals were obtained during the second half of the last century by crossing local Lecceesi ewes with rams from different breeds, mainly Bergamasca, a sheep breed from Northern Italy (also known as the Giant of Bergamo, due to its body size) widely used in the past as crossing breed for several Alpine and Apenninic breeds. Breed sub-structuring within the Lecceesi population was also evident from the Bayesian assignment test performed on the whole population sample. Moreover, when the analysis was repeated uniquely on the Lecceesi sample an incredible fine and precise dissection of genetic variability was obtained, with animals from different flocks falling each in a corresponding cluster (data not shown). This result cannot be explained only by the above-cited crossbreeding events (unless assuming that rams from different breeds were used in each flock) and seems more likely due to a very close reproductive isolation among flocks, that would also justify the high levels of inbreeding and gametic disequilibrium observed in this study.

CONCLUSION

This paper presents the results from the genetic characterization of seven Italian sheep breeds through a low-density SNP panel. Results are compared to those previously observed at STR loci for the same population data-set. The two class of markers generally displayed low consistency both in measuring within-breed genetic variability and in reconstructing genetic relationship among breeds, likely due to a poor informativeness of bi-allelic loci as a consequence of the limited number. However, STR and SNP loci coherently highlighted some striking evidences, notably the greater differentiation of Sarda from all the other breeds and the presence of genetic sub-structuring within the Lecceesi population sample. Advancing fast and cheap whole genome sequencing and genome-wide SNP genotyping techniques are now providing very large panels of bi-allelic markers, validated over many world-wide sheep breeds, that will facilitate the detection of highly informative SNPs displaying breed-specific allele frequency distributions.

AKNOWLEDGEMENTS

Authors would like to thank the Ministero delle Politiche Agricole, Alimentari e Forestali (MIPAAF) for funding the research, Associazione Nazionale della Pastorizia, Istituto Zooprofilattico della Sicilia, Azienda Sperimentale Cavone di Spinazzola and farmers for assistance with sample collection.
References


