# Accuracy of Imputation of Microsatellite Markers from an SNP50K Chip in Spanish Assaf Sheep

**Hector Marina**
Universidad de León - Campus de Vegazana: Universidad de Leon
**Aroa Suarez-Vega**
Universidad de León - Campus de Vegazana: Universidad de Leon
**Rocio Pelayo**
Universidad de León - Campus de Vegazana: Universidad de Leon
**Beatriz Gutierrez-Gil**
Universidad de León - Campus de Vegazana: Universidad de Leon
**Antonio Reverter**
CSIRO Queensland Bioscience Precinct Agriculture and Food Unit
**Cristina Esteban-Blanco**
Universidad de León - Campus de Vegazana: Universidad de Leon
**Juan-Jose Arranz** ( jjarrs@unileon.es )
Universidad de Leon https://orcid.org/0000-0001-9058-131X

**Research**

**Keywords:** Pedigree verification, sheep, microsatellites, SNPs, marker imputation

**Posted Date:** November 20th, 2020

**DOI:** https://doi.org/10.21203/rs.3.rs-111084/v1

## Abstract

**Background**: Traditional and new genotyping technologies must be combined by applying bridge methodologies that avoid double genotyping costs. This study aims to identify and evaluate a reliable approach to precisely impute microsatellite markers from SNP-chip panels to perform parental verifications in sheep. Moreover, we will assess the optimum number of SNPs necessary to accurately impute microsatellite markers to develop a low-density SNP chip for parentage verification in the Assaf sheep breed.

**Results:** A total of 4,423 animals belonging to the Spanish Assaf sheep breed were genotyped for 19 microsatellites and an ovine custom 49,897 SNP array. The accuracy of microsatellite marker imputation, performed with BEAGLE v5.1 software, was assessed with three metrics, namely, genotype concordance (C), genotype dosage (length $r^2$), and allelic dosage (allelic $r^2$), for all imputation scenarios tested (0.5-10 Mb microsatellite flanking SNP windows). The accuracy of our imputation results for the three metrics analyzed for all haplotype lengths tested was higher than 0.90 (C), 0.80 (length $r^2$), and 0.75 (allelic $r^2$). Considering that the objective of the study was to assess a SNP window length that provides the best accuracy for the microsatellite imputation procedure to design an affordable low-density SNP chip for parentage testing, we considered 2 Mb to be the best SNP haplotype length for further analyses (SNPs/window =74.05, C= 0.970; length $r^2$= 0.952, allelic $r^2$=0.899). We additionally evaluated imputation performance under two null models, naive and random, which showed weak genotype concordance averages in comparison with imputed microsatellites (0.41 and 0.15, respectively).

**Conclusions**: We presented for the first time a precise methodology in dairy sheep to impute multiallelic microsatellite genotypes from biallelic SNP markers. The use of a 2 Mb SNP flanking window for each microsatellite has been shown to achieve high accuracy in the imputation procedure while providing a low-density SNP chip that could be cost-effective. The results from this study will undoubtedly have a significant impact on sheep breeders overcoming the problem of parentage verification when different genotyping platforms have been used across generations.


## Background

Parentage misassignments directly affect genetic gain in traditional breeding programs by biasing heritability estimates of productive traits, genetic parameters, breeding values, and the identification of superior animals for selection (1–3). Therefore, accurate pedigree records are essential for the success of genetic improvement in livestock.

The use of molecular markers, specifically genetic markers, facilitates parentage verification and individual identification by indicating, through different approaches (simple exclusion, genotype reconstruction, or categorical and fractional allocation), the putative relatedness between individuals (4, 5). In this sense, microsatellite variants have become one of the principal molecular markers used in livestock in recent decades for parentage testing. Microsatellites, also known as short tandem repeats (STRs) or simple sequence repeats (SSRs), consist of motifs of 1–6 bp repeated in tandem and represent choice markers for parentage testing in livestock due to their high polymorphic information content with codominant inherited alleles and easy but not fully automatized allele scoring (6).

At present, microsatellite information for parentage verification tests is being gradually replaced by single nucleotide polymorphisms (SNPs). Although SNPs are less informative due to their biallelic nature, which determines the range of markers required for parentage testing (200–700 SNPs compared to 14–20 microsatellites) (7), there is increasing interest in using SNP panels in livestock. The advantages of SNP panels include the more straightforward automation of technology, the lack of need for interlaboratory calibration, lower error rates, the uniform distribution of SNP markers across the genome, and the recently reduced costs in genotyping technology (8–10). Moreover, SNP panels are increasingly used in livestock due to the implementation of genomic selection in breeding schemes (11–13).

In the case of sheep, there are two strategies for parentage testing proposed by the International Society of Animal Sciences (ISAG): a panel of 19 microsatellites (14) and a panel of 163 SNPs with verified qualities to use in diverse sheep breeds (7). Notably, in the Spanish Assaf sheep, most of the animals in the selection scheme are genotyped with microsatellite markers. Therefore, the need for a consistent and reliable pedigree database across generations has made the use of microsatellite information an essential issue. However, since the implementation of genomic selection, with the first genomic evaluation results obtained in 2020, there has been an annual increase in the number of animals genotyped with a 50K SNP panel. Some of these new animals are genotyped with both platforms: SNPs and microsatellites. Parentage verification should be performed using the same technology applied in previous generations, implying additional costs for farmers and breeders' associations. One possible strategy in the migration process from microsatellites to SNP panels to avoid double costs in genotyping is the imputation of microsatellite alleles from SNP haplotypes (15). Therefore, this study aims to identify and evaluate a reliable approach to be used during the transition period to accurately impute microsatellite markers from SNP-chip panels to perform parental verifications in sheep. Moreover, we will evaluate the optimum number of SNPs necessary to accurately impute microsatellite markers to develop a low-density SNP panel for parentage verification.

## Methods

# Animal genotypes and quality control

The genetic profiles of 4,423 animals included in the breeding program of Spanish Assaf dairy sheep were obtained from the Association of Spanish Assaf Sheep Breeders (ASSAFE). These animals were genotyped for the 19 microsatellites recommended by ISAG for paternity control (14), and by an SNP chip with 49,897 markers used in the genomic selection program implemented in the Spanish Assaf sheep breed.

Prior to the imputation process, quality control was applied to both sets of markers. We filtered out microsatellites with call rates below 80% and expected heterozygosity under Hardy-Weinberg equilibrium below 0.095. This value corresponds to a minor allele frequency (MAF) of 5% for biallelic markers. After quality control, microsatellite alleles were recoded to fit the variant call format (VCF) following the VCFtools software specifications (18). The most common allele for each microsatellite was considered the reference allele in the population and recoded as “0”. For the rest of the alleles, a consecutive number was assigned (1,2,3…n) based on the microsatellite allele length. SNP-chip quality control was performed using PLINK (19), and SNPs with call rates under 95% were excluded from the dataset. To maintain haplotype diversity in the population, MAF filtering is not included in the SNP quality control.

# Imputation procedure

The positions of the microsatellite markers in the ovine genome (Oar_v3.1) were obtained from the sheep database from Ensembl v.95 (https://www.ensembl.org/Ovis_aries/ Info/Index), and they were verified through the alignment of the primer sequences to the sheep reference genome using BLAST (16). The genotypes were phased and imputed using the phasing method implemented in the BEAGLE 5.1 software (20) [50 rounds of burn-in and 100 iterations] and the genotype imputation method (21), of the same program. To establish the minimum number of flanking SNPs per microsatellite and the optimal window length to achieve accurate imputation, several SNP window distances on each side of the marker were considered [0.5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 40, 45, 50 Mega bases (Mb)]. In the genotype imputation process, pedigree information and the effective population size (Ne) were also considered.

Given that the animals were genotyped for microsatellite and SNP markers, the imputation performance was estimated by a 10-fold cross-validation approach. For this purpose, we divided the total population into two groups: the training group, which comprises 90% of the total population, and the validation group, which comprises the remaining 10%. The microsatellite information was masked in the validation population, and the genotypes for these markers were imputed by Beagle software using all information (microsatellite and SNP genotypes and pedigree relationships from the reference dataset and the genotypes of SNPs in the validation dataset). The process was repeated for ten rounds, using different animals in the validation dataset in each round following a nonparametric bootstrap of 10% of the total samples using a custom Fortran source code.

# Imputation performance metrics

To assess the accuracy of the microsatellite imputation, we used the metrics genotype concordance, genotype dosage, and allelic dosage, which were previously defined by Saini et al. (17). The genotype concordance ($c_i$) was defined as 0 if neither of the imputed alleles matched a true allele, 0.5 if one of them matched, and 1 if both alleles matched the true alleles. Thus, the genotype concordance for a microsatellite ($C$) was calculated as the average over all the samples of $c_i$ for each microsatellite

$$C = \frac{1}{n}\sum_{i=1}^{n} c_i.$$

The microsatellite genotype dosage [aka length $r^2$] was defined as the Pearson correlation between the sum of the lengths of both alleles at a specific locus in imputed genotypes and the true genotypes. Last, the microsatellite allelic dosage [aka allelic $r^2$] was calculated as the Pearson correlation for each microsatellite allele length α, defined as

$$X_a = \{a_1, a_2, \ldots, a_n\}, \text{ where } a_i = \sum_{j=1}^{2} 1_{(x_{ij}=a)}$$

In addition, Pearson's correlations between the frequencies of the reference microsatellite alleles and the imputed microsatellite alleles were calculated. Furthermore, for each microsatellite, the imputation performance was evaluated by computing the expected value for each metric under two null models: (1) a naive model in which the imputed genotype was selected as the most common allele per microsatellite and (2) a

random model in which the imputed genotype was randomly selected from the available genotypes at each marker, depending indirectly on the allelic frequencies (17).

## Population structure, effective population size, and parental relationships

Finally, we used SNP chip information to evaluate different factors affecting imputation accuracies, such as population structure, effective population size, and parental-progeny pedigree conflicts. To assess the population structure, we estimated the genomic relationship matrix (GRM), and the genomic relationships between the individuals were plotted following the Pedigromics pipeline (22), calculating centrality metrics such as betweenness and closeness coefficients. The effective population size and the parental-progeny conflicts in the pedigree were computed using the BLUPF90 family programs (23).

# Results

## Genotype quality control

All microsatellite markers passed the quality control settings fixed in the analysis. Regarding SNP markers, a total of 3,537 markers showed a call rate lower than 95% and were filtered out. Therefore, a total of 19 microsatellites and 42,665 SNPs were considered for the imputation procedure. The microsatellites were located along the sheep autosomes, and on average, each microsatellite included in this study had 12.73 alleles ranging from 81 to 297 base pairs. The information of the microsatellites considered in this work is summarized in Table 1, and their allele frequencies are presented in Table S1.

Table 1
Characteristics of microsatellite markers used in the present study.

| Microsatellite ID | CHR | Position (bp) | Nº of Alleles | Range (bp) |
|---|---|---|---|---|
| *INRA006* | 1 | 109478015 | 13 | 104–134 |
| *INRA049* | 1 | 1952560108 | 9 | 134–166 |
| *INRA023* | 1 | 86986507 | 14 | 194–220 |
| *FCB20* | 2 | 153680836 | 14 | 87–115 |
| *AE129* | 5 | 78045895 | 6 | 135–161 |
| *SPS113* | 7 | 23419543 | 11 | 126–152 |
| *ILSTS005* | 7 | 92854099 | 12 | 190–214 |
| *ILSTS011* | 9 | 25256863 | 8 | 268–282 |
| *ILSTS008* | 9 | 45990219 | 2 | 168–170 |
| *McM042* | 9 | 51865313 | 8 | 81–107 |
| *CSRD247* | 14 | 15564041 | 19 | 205–257 |
| *INRA063* | 14 | 39826970 | 18 | 167–207 |
| *SPS115* | 15 | 23269440 | 12 | 237–255 |
| *MAF65* | 15 | 30901387 | 9 | 119–137 |
| *MAF214* | 16 | 33667802 | 16 | 183–269 |
| *CP49* | 17 | 14434435 | 25 | 76–136 |
| *HSC* | 20 | 25764806 | 17 | 263–297 |
| *INRA132* | 20 | 4668849 | 17 | 146–180 |
| *INRA172* | 22 | 20603037 | 12 | 126–172 |

For each SSR marker, the ID, genome position (Oar_v3.1), number of alleles per marker, and allele length range expressed in base pairs are shown in the table.

## Imputation procedure results

To impute the whole population considered by this study, we performed a 10-fold cross-validation approach, as explained above. Therefore, ten imputation procedures were necessary to estimate the microsatellite information in the whole population. Moreover, we assessed the accuracy (concordance, genotype dosage, and allelic dosage) of the imputed microsatellite markers in the proposed imputation scenarios (window lengths from 0.5 Mb to 50 Mb) to determine the best haplotype length for the imputation procedure (Fig. 1).

There was a noticeable increase in the imputation accuracy of the microsatellite markers for the average accuracy metrics when 0.5 Mb (SNPs/window = 19.11, C = 0.922; length $r^2$ = 0.890, allelic $r^2$ = 0.788), 1 Mb (SNPs/window = 38.05, C = 0.962; length $r^2$ = 0.941, allelic $r^2$ = 0.878) and 2 Mb (SNPs/window = 74.05, C = 0.970; length $r^2$ = 0.952, allelic $r^2$ = 0.899) window lengths were compared (Fig. 1; Table S2). Considering a 3 Mb window length, the addition of new information provided by the SNPs localized in the surrounding windows (> 100 SNPs) slightly improved the imputation accuracy metrics (C = 0.972; length $r^2$ = 0.957, allelic $r^2$ = 0.901), while a stabilization in the imputation accuracy was observed for wider windows (Fig. 1). The number of flanking SNPs used in the imputation process for the tested window distances is summarized in Table S3. Considering that the objective of our work was to assess a SNP window length that provides optimum accuracy for the microsatellite imputation procedure to design an affordable low-density SNP chip to be used for parentage testing by breeders, we considered 2 Mb to be the best haplotype for further analyses. The imputation metrics (concordance, genotype dosage, and allelic dosage) for the 2 Mb scenario are summarized in Table 2. The distribution of the allelic $r^2$ values is represented in Figure S1. Using a 2 Mb SNP haplotype, the Pearson correlation between the real microsatellite allele frequency in the population and the frequency of the imputed alleles in these markers was 1.00. These frequencies are represented in Figure S2.

To validate our imputation results, we analyzed the imputation performance under two null imputation models: naive imputation, in which imputed genotypes showed an average concordance of 0.41 (ranging from 0.26 to 0.67) with observed genotypes, and random imputation, which had an average concordance of 0.15 (ranging from 0.06 to 0.60). Both validation procedures revealed considerably fewer concordance values using the two null imputation models than the imputation method proposed in this study, which validates our approach.

Table 2
Imputation performance metrics summary for the 19 metrics microsatellites considered in this study using a 2 Mb window, together with the concordance obtained using the naive and random models.

| CHR | Position | Microsatellite | Concordance | Genotype dosage[1] | Allelic dosage[2] | Minimum allelic dosage[2] | Maximum allelic dosage | Naive concordance | Random concordance |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 86986507 | INRA023 | 0.98 | 0.97 | 0.97 | 0.93 | 0.99 | 0.28 | 0.13 |
| 1 | 109478015 | INRA006 | 0.93 | 0.87 | 0.84 | 0.60 | 1.00 | 0.48 | 0.11 |
| 1 | 195256010 | INRA049 | 0.97 | 0.97 | 0.88 | 0.32 | 0.98 | 0.44 | 0.16 |
| 2 | 153680836 | FCB20 | 0.96 | 0.94 | 0.89 | 0.52 | 1.00 | 0.26 | 0.10 |
| 5 | 78045895 | AE129 | 0.96 | 0.96 | 0.88 | 0.13 | 1.00 | 0.47 | 0.20 |
| 7 | 23419543 | SPS113 | 0.95 | 0.93 | 0.79 | 0.16 | 0.94 | 0.34 | 0.16 |
| 7 | 92854099 | ILSTS005 | 0.99 | 0.97 | 0.97 | 0.97 | 0.97 | 0.41 | 0.11 |
| 9 | 25256863 | ILSTS011 | 0.98 | 0.96 | 0.92 | 0.83 | 0.97 | 0.50 | 0.18 |
| 9 | 45990219 | ILSTS008 | 0.97 | 0.86 | 0.80 | 0.37 | 0.97 | 0.67 | 0.61 |
| 9 | 51865313 | McM042 | 0.97 | 0.97 | 0.93 | 0.70 | 0.98 | 0.48 | 0.17 |
| 14 | 15564041 | CSRD247 | 0.99 | 0.97 | 0.97 | 0.94 | 1.00 | 0.34 | 0.07 |
| 14 | 39826970 | INRA063 | 0.97 | 0.95 | 0.82 | 0.47 | 0.98 | 0.33 | 0.09 |
| 15 | 23269440 | SPS115 | 0.96 | 0.95 | 0.90 | 0.67 | 1.00 | 0.33 | 0.14 |
| 15 | 30901387 | MAF65 | 0.98 | 0.97 | 0.92 | 0.75 | 1.00 | 0.36 | 0.18 |
| 16 | 33667802 | MAF214 | 0.98 | 0.98 | 0.86 | 0.32 | 1.00 | 0.54 | 0.09 |
| 17 | 14434435 | CP49 | 0.98 | 0.97 | 0.92 | 0.84 | 0.98 | 0.39 | 0.06 |
| 20 | 4668849 | INRA132 | 0.98 | 0.97 | 0.95 | 0.84 | 0.97 | 0.29 | 0.11 |
| 20 | 25764806 | HSC | 0.98 | 0.98 | 0.95 | 0.77 | 0.99 | 0.54 | 0.09 |
| 22 | 20603037 | INRA172 | 0.96 | 0.95 | 0.91 | 0.72 | 1.00 | 0.35 | 0.12 |

[1] Genotype dosage: length $r^2$

[2] Allelic dosage: allelic $r^2$

# Population structure and effective population size

Figure 2 presents the population structure of the 4,423 animals included in the study using the GRM created with the 42,665 SNPs remaining after quality control filtering. Individuals are represented as nodes in the network, and two animals are connected by an edge when a pre-defined genomic kinship exists, e.g., parent-offspring. Those animals not related to the main population were filtered in the representation. Genomic relationship higher than 0.2 and 0.5 were represented in Fig. 2 and Figure S3, respectively. The Pedigromics approach of the Assaf breed showed low average values of the betweenness centrality coefficient (0.003) and closeness coefficient (0.237), both ranging between 0 and 1. Centrality coefficients reflect the influence of each vertex over the graph structure. In this case, closeness centrality is based on the average length of the shortest paths from a given node to other reachable nodes in the network (24), given how genomic information is spread in the population (25). The betweenness centrality coefficient reflects the amount of control that a node exerts over the interactions with other nodes in the network. Animals with high betweenness centrality in a pedigree graph could have a role in connecting disconnected groups (25). The low average values of the betweenness centrality and closeness coefficient suggest a low relationship among the samples included in the population studied. However, 21% of the animals had a closeness coefficient higher than the third quartile of the value distribution (0.24), which is represented on a green to a red color scale in Fig. 2. These samples are distributed in eight related family groups (as shown in Fig. 2). The low degree of relationship between these groups and the rest of the animals suggests that the population is neither highly related nor structured. Moreover, we estimated the effective population size of the studied Assaf population, which was 214 animals.

# Parental relationships

The pedigree records available for the Assaf population under study integrated 1,450 parental-progeny relationships that could be confirmed with the SNP information to detect parental-progeny conflicts. A total of 24 misassignments were found in the pedigree, representing a total of 1.66% of all the parental relationships analyzed in the pedigree.

## Discussion

This research presented for the first time a precise methodology in sheep to impute multiallelic genotypes from biallelic information. Traditional and new genotyping technologies must be joined by applying bridge methodologies, which allow breeders to avoid additional costs of regenotyping historical data. Our study combines microsatellite and SNP markers in an efficient approach to impute microsatellite markers through SNP haplotypes, achieving high concordance rates. Therefore, the imputation procedure developed represents a useful and inexpensive approach to perform parentage verification when different genotyping platforms have been used across generations. The results from this study will undoubtedly have a great impact on Assaf sheep breeders, allowing them to perform a transition from microsatellite maker kinship verification to the use of SNP panels (26). In addition to constituting a clear advantage for sheep producers, the imputation methodologies developed can provide advantages in genomic studies by combining both types of data, such as in genome-wide association analyses (GWAS). In this approach, microsatellite information could improve the detection of new associations, provide complementary information, and explain part of the missing heritability for the trait under study (17).

In general, as shown in Fig. 1, the accuracy of our imputation results for the three metrics analyzed (C, length $r^2$, and allelic $r^2$) in the different scenarios tested (SNP windows ranged between 0.5 and 10 Mb) was higher than 0.90 (C), 0.80 (length $r^2$), and 0.75 (allelic $r^2$) for all haplotype lengths. The accuracy results presented in this study were higher than those found in a previous study performed in cattle by Sharma et al. (26), which reached a concordance of 0.40 and a correlation between the real and imputed microsatellites of 0.31. In addition, we have explored not only the viability of performing microsatellite imputation but also the optimum number of SNPs necessary to perform accurate imputation of microsatellite information. According to Strucken et al. (7) 700 SNP markers are required to reduce false-positive results in parentage testing, which in our approach correspond to an SNP haplotype length of 1 Mb, covering 38.05 SNPs per microsatellite with adequate imputation accuracy rates (C = 0.962; length $r^2$=,0.941, allelic $r^2$ = 0.878). However, the imputation performance reached high accuracy values at a SNP haplotype length of 2 Mb: 0.97 (C) 0.95 (length $r^2$) and 0.90 (allelic $r^2$), with all accuracy metrics higher than 0.90. The SNPs located in the 2 Mb window distance used in the imputation procedure have been summarized in Table S4. These results were slightly higher than those obtained by Saini et al. (17), who achieved a genotype concordance of 0.97, a genotypic dosage of 0.91, and an allelic dosage of 0.86. In our study, accuracy metrics were obtained using a 50 k SNP chip in sheep compared to the SNP data from whole-genome sequencing (27,185,239 SNPs) with a SNP window of 100 Kb used by Saini et al. (17) in humans. Moreover, concordance rates of the null models obtained by Saini et al. [naive (0.72), and random (0.61)] are higher than those obtained in the present study [naive (0.41) and random (0.15)]. This highlights the genetic diversity of the microsatellite markers in sheep and the high efficiency of the imputation procedure presented in this work.

The number of haplotypes per microsatellite and the frequency of these haplotypes did not significantly impact the allele dosage, with correlations of 0.33 and 0.18, respectively. Therefore, as the number of alleles and their frequency increases, the concordance tends to rise. However, the naive and random models' concordance rate decreased as the number of alleles increased because they depended on the number of haplotypes of each microsatellite (correlations were − 0.45 and − 0.70).

The imputation accuracies obtained might be overestimated due to (i) a highly structured and related population (27) or due to (ii) a low effective population size (28). On the one hand, the population included in the present work, represented using the Pedigromics pipeline (Fig. 2), achieved low rates of centrality coefficients (betweenness coefficient = 0.003 and closeness coefficient = 0.237), which suggests that the population is nonstructured or highly related. In addition, the selection of the reference and test populations during cross-validation by a nonparametric bootstrap approach avoids the overestimation of the imputation metrics by avoiding the selection of immediate relative samples in the different groups. On the other hand, the effective population size was 214, higher than in highly selected cattle breeds (26), but in the wide range of effective population sizes described in sheep breeds from values of 78 in Romney, 100 in Wiltshire breed, 128 in Churra breed, to 1,317 in Qezel (29–31). Lower values can lead to an overestimation of imputation accuracy metrics; however, if we compared our concordance rates with the concordance rates obtained in the microsatellite imputation in cattle carried out by Sharma et al. (26), we achieved more than double the concordance (0.90 vs. 0.40). Small population sizes reduce the genetic diversity in the population (32) and would influence the naive and random models' concordance rates, increasing their accuracy parameters. Nevertheless, the average of the naive and random concordance rates for these two models (0.41 and 0.15, respectively) was far lower than those obtained in humans by Saini et al. (17), [0.72 and 0.61, respectively]. This difference between the imputation accuracy and the accuracy of the null models could be because the effective population size and the genetic diversity of the Assaf population analyzed are large enough to perform an accurate imputation of the microsatellite information. In particular, high genetic diversity in the reference population would help achieve high squared

correlations in the imputation process (10, 27, 33) and reduce the probability of accurate imputations in the naive and random models. Therefore, the large number of samples included in this study, and as a consequence, the large number of individuals genotyped in the reference population, could influence the high accuracy rates achieved because it is necessary to impute the odd haplotypes (28) accurately and could also reduce the concordance rates obtained in the null models. Therefore, this finding explained the higher concordance rates obtained than those in previous studies on microsatellite imputation from SNP data conducted with lower sample sizes in humans (17) [1,916 samples] and cattle (26) [1,482 samples].

Last, the development of a low-density SNP panel with the 1,407 SNPs (2 Mb SNP window) proposed in this approach (Table S4) would also help to reduce the number of kinship errors in the pedigree due to its lower error rates compared with microsatellite markers and the lack of need for interlaboratory calibration and easier automation (8–10).

## Conclusion

This study presents an effective methodology to overcome the problem presented in the transition from multiallelic (microsatellites) to biallelic markers (SNPs) for pedigree verification analyses in sheep. The use of a flanking 2 Mb SNP window for each microsatellite has been shown to achieve high accuracy in the imputation procedure while providing a cost-effective, low-density SNP chip for breeders. The microsatellite imputed information could be used for individual identification and parentage verification in sheep, postulating a useful approach in the sheep industry to avoid double genotyping.

## Abbreviations

*GD: Genotype dosage; GRM: Genomic relationship matrix; GWAS: Genome-wide association studies. ISAG: International Society of Animal Sciences; SNP: Single nucleotide polymorphism; STR: Short tandem repeats; SSR: Simple sequence repeats; MAF: Minor frequency allele; Ne: Effective population size.*

## Declarations

# Ethics approval and consent to participate

As the data were obtained from the Spanish Assaf breeders association (ASSAFE) database, no direct experimentation on animals was performed in this work. According to the Research Ethics Committee of the University of León, formal ethical approval was not necessary for this case.

# Consent for publication

Not applicable.

# Availability of data and materials

Supplementary material related to this article can be found in the online version. The genotype datasets that support the results of this study belong to the Spanish Assaf breeders association (ASSAFE) and should be requested from that association.

# Competing interests

The authors declare that they have no competing interests regarding the publication of this article.

# Funding

This research work was supported by the RTI2018-093535-B-I00 project of the Spanish Ministry of Economy and Science and Innovation (Madrid, Spain), co-funded by the European Regional Development Fund. H. Marina is funded by an FPU contract from the Ministry of Science, Innovation and Universities (MICIU, Ref. FPU16/01161).

# Authors' contributions

Conceived and designed the experiments: ASV, BGG, HM and JJA. Analyzed the data: HM and JJA. Wrote the paper: ASV, BGG, CEB, HM, RP, AR and JJA. All authors read and approved the final manuscript.

# Acknowledgments

The authors would like to thank the National Association of Sheep Breeders of Assaf Breed (ASSAFE) (http://assafe.es/) for allowing us to access the database of genotypes of the animals in the selection nucleus.

# Authors' information

Departamento de Producción Animal, Facultad de Veterinaria, Universidad de León, Campus de Vegazana s/n, León 24071, Spain.

## References

1. Dodds KG, Tate ML, Sise JA. Genetic evaluation using parentage information from genetic markers1. J Anim Sci [Internet]. 2005 Oct 1 [cited 2020 Sep 21];83(10):2271–9. Available from: https://academic.oup.com/jas/article/83/10/2271/4790459
2. Geldermann H, Pieper U, Weber WE. Effect of Misidentification on the Estimation of Breeding Value and Heritability in Cattle1. J Anim Sci [Internet]. 1986 Dec 1 [cited 2020 Sep 21];63(6):1759–68. Available from: https://academic.oup.com/jas/article/63/6/1759-1768/4662094
3. Heaton MP, Leymaster KA, Kalbfleisch TS, Kijas JW, Clarke SM, McEwan J, et al. SNPs for Parentage Testing and Traceability in Globally Diverse Breeds of Sheep. Wade C, editor. PLoS One [Internet]. 2014 Apr 16 [cited 2020 Sep 21];9(4):e94851. Available from: https://dx.plos.org/10.1371/journal.pone.0094851
4. Jones AG, Ardren WR. Methods of parentage analysis in natural populations [Internet]. Vol. 12, Molecular Ecology. Mol Ecol; 2003 [cited 2020 Sep 21]. p. 2511–23. Available from: https://pubmed.ncbi.nlm.nih.gov/12969458/
5. Jones AG, Small CM, Paczolt KA, Ratterman NL. A practical guide to methods of parentage analysis [Internet]. Vol. 10, Molecular Ecology Resources. Mol Ecol Resour; 2010 [cited 2020 Sep 21]. p. 6–30. Available from: https://pubmed.ncbi.nlm.nih.gov/21564987/
6. Chambers GK, MacAvoy ES. Microsatellites: Consensus and controversy. Vol. 126, Comparative Biochemistry and Physiology - B Biochemistry and Molecular Biology. Elsevier Inc.; 2000. p. 455–76.
7. Strucken EM, Lee SH, Lee HK, Song KD, Gibson JP, Gondro C. How many markers are enough? Factors influencing parentage testing in different livestock populations. J Anim Breed Genet [Internet]. 2016 Feb 1 [cited 2020 Sep 21];133(1):13–23. Available from: https://pubmed.ncbi.nlm.nih.gov/26234440/
8. Glover KA, Hansen MM, Lien S, Als TD, Høyheim B, Skaala Ø. A comparison of SNP and STR loci for delineating population structure and performing individual genetic assignment. BMC Genet [Internet]. 2010 Jan 6 [cited 2020 Apr 16];11(1):2. Available from: https://bmcgenet.biomedcentral.com/articles/10.1186/1471-2156-11-2
9. Carta A, Casu S, Salaris S. Current state of genetic improvement in dairy sheep. Vol. 92, Journal of Dairy Science. American Dairy Science Association; 2009. p. 5814–33.
10. Zhang P, Zhan X, Rosenberg NA, Zöllner S. Genotype imputation reference panel selection using maximal phylogenetic diversity. Genetics. 2013 Oct;195(2):319–30.
11. Cesarani A, Gaspa G, Correddu F, Cellesi M, Dimauro C, Macciotta NPP. Genomic selection of milk fatty acid composition in Sarda dairy sheep: Effect of different phenotypes and relationship matrices on heritability and breeding value accuracy. J Dairy Sci. 2019 Apr 1;102(4):3189–203.
12. Lillehammer M, Sonesson AK, Klemetsdal G, Blichfeldt T, Meuwissen THE. Genomic selection strategies to improve maternal traits in Norwegian White Sheep. J Anim Breed Genet [Internet]. 2020 Jul 31 [cited 2020 Sep 21];137(4):384–94. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/jbg.12475
13. Brito LF, Clarke SM, McEwan JC, Miller SP, Pickering NK, Bain WE, et al. Prediction of genomic breeding values for growth, carcass and meat quality traits in a multi-breed sheep population using a HD SNP chip. BMC Genet [Internet]. 2017 Dec 26 [cited 2020 Apr 29];18(1):7. Available from: http://bmcgenet.biomedcentral.com/articles/10.1186/s12863-017-0476-8
14. Di Stasio L. ISAG panels of markers for parentage verification [Internet]. 2002 [cited 2020 Sep 22]. Available from: http://www.isag.us/Docs/consignmentforms/02_PVpanels_LPCGH.doc
15. McClure M, Sonstegard T, Wiggans G, Van Tassell CP. Imputation of Microsatellite Alleles from Dense SNP Genotypes for Parental Verification. Front Genet [Internet]. 2012 Aug 14 [cited 2020 Sep 15];3(AUG):140. Available from:

http://journal.frontiersin.org/article/10.3389/fgene.2012.00140/abstract
16. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403–10.
17. Saini S, Mitra I, Mousavi N, Fotsing SF, Gymrek M. A reference haplotype panel for genome-wide imputation of short tandem repeats. Nat Commun. 2018 Dec 1;9(1).
18. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics [Internet]. 2011 Aug 1 [cited 2018 Nov 9];27(15):2156–8. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr330
19. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet [Internet]. 2007 Sep [cited 2018 Oct 11];81(3):559–75. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17701901
20. Browning BL, Zhou Y, Browning SR. A One-Penny Imputed Genome from Next-Generation Reference Panels. Am J Hum Genet. 2018 Sep 6;103(3):338–48.
21. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007;81(5):1084–97.
22. Reverter A, Dominik S, Ferraz JBS, Corrigan L, Porto-Neto LR. Pedigromics: a network-inspired approach to visualise and analyse pedigree structures. Proc. Assoc. Advmt. Anim. Breed. Genet. 2019;23:540-543. Available from: http://www.aaabg.org/aaabghome/AAABG23papers/133Reverter23540.pdf
23. Misztal I, Tsuruta S, Lourenco D, Aguilar I, Legarra A, Vitezica Z. Manual for BLUPF90 family of programs. [Internet]. 2019 [cited 2019 Oct 15]. Available from: http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90_all2.pdf.
24. Doncheva NT, Assenov Y, Domingues FS, Albrecht M. Topological analysis and interactive visualization of biological networks and protein structures. Nat Protoc [Internet]. 2012 Apr 15 [cited 2020 Oct 2];7(4):670–85. Available from: https://www.nature.com/articles/nprot.2012.004
25. Da Costa Perez B. Strategies to improve results from genomic analyzes in small dairy cattle populations. Universidade de São Paulo; 2019.
26. Sharma A, Park J-E, Park B, Park M-N, Roh S-H, Jung W-Y, et al. Accuracy of Imputation of Microsatellite Markers from BovineSNP50 and BovineHD BeadChip in Hanwoo Population of Korea. Genomics Inform [Internet]. 2018 Jan 1 [cited 2020 Jun 25];16(1):10–3. Available from: https://pubmed.ncbi.nlm.nih.gov/29618182/
27. Bolormaa S, Gore K, van der Werf JHJ, Hayes BJ, Daetwyler HD. Design of a low-density SNP chip for the main Australian sheep breeds and its effect on imputation and genomic prediction accuracy. Anim Genet [Internet]. 2015 Oct [cited 2020 Jan 19];46(5):544–56. Available from: http://doi.wiley.com/10.1111/age.12340
28. Druet T, Macleod IM, Hayes BJ. Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. Heredity (Edinb) [Internet]. 2014 Jan 3 [cited 2019 Jun 10];112(1):39–47. Available from: http://www.nature.com/articles/hdy201313
29. Kijas JW, Lenstra JA, Hayes B, Boitard S, Neto LR, Cristobal MS, et al. Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. PLoS Biol. 2012;10(2).
30. García-Gámez E, Sahana G, Gutiérrez-Gil B, Arranz J-J. Linkage disequilibrium and inbreeding estimation in Spanish Churra sheep. BMC Genet [Internet]. 2012 [cited 2020 Jan 13];13(1):43. Available from: http://bmcgenet.biomedcentral.com/articles/10.1186/1471-2156-13-43
31. Prieur V, Clarke SM, Brito LF, McEwan JC, Lee MA, Brauning R, et al. Estimation of linkage disequilibrium and effective population size in New Zealand sheep using three different methods to create genetic maps. BMC Genet [Internet]. 2017 Dec 21 [cited 2020 Jan 12];18(1):68. Available from: http://bmcgenet.biomedcentral.com/articles/10.1186/s12863-017-0534-2
32. Frankham R. Relationship of Genetic Variation to Population Size in Wildlife. Conserv Biol [Internet]. 1996 Dec 1 [cited 2020 Oct 4];10(6):1500–8. Available from: https://conbio.onlinelibrary.wiley.com/doi/full/10.1046/j.1523-1739.1996.10061500.x
33. Boichard D, Chung H, Dassonneville R, David X, Eggen A, Fritz S, et al. Design of a Bovine Low-Density SNP Array Optimized for Imputation. Liu Z, editor. PLoS One [Internet]. 2012 Mar 28 [cited 2020 Apr 6];7(3):e34130. Available from: https://dx.plos.org/10.1371/journal.pone.0034130

## Figures

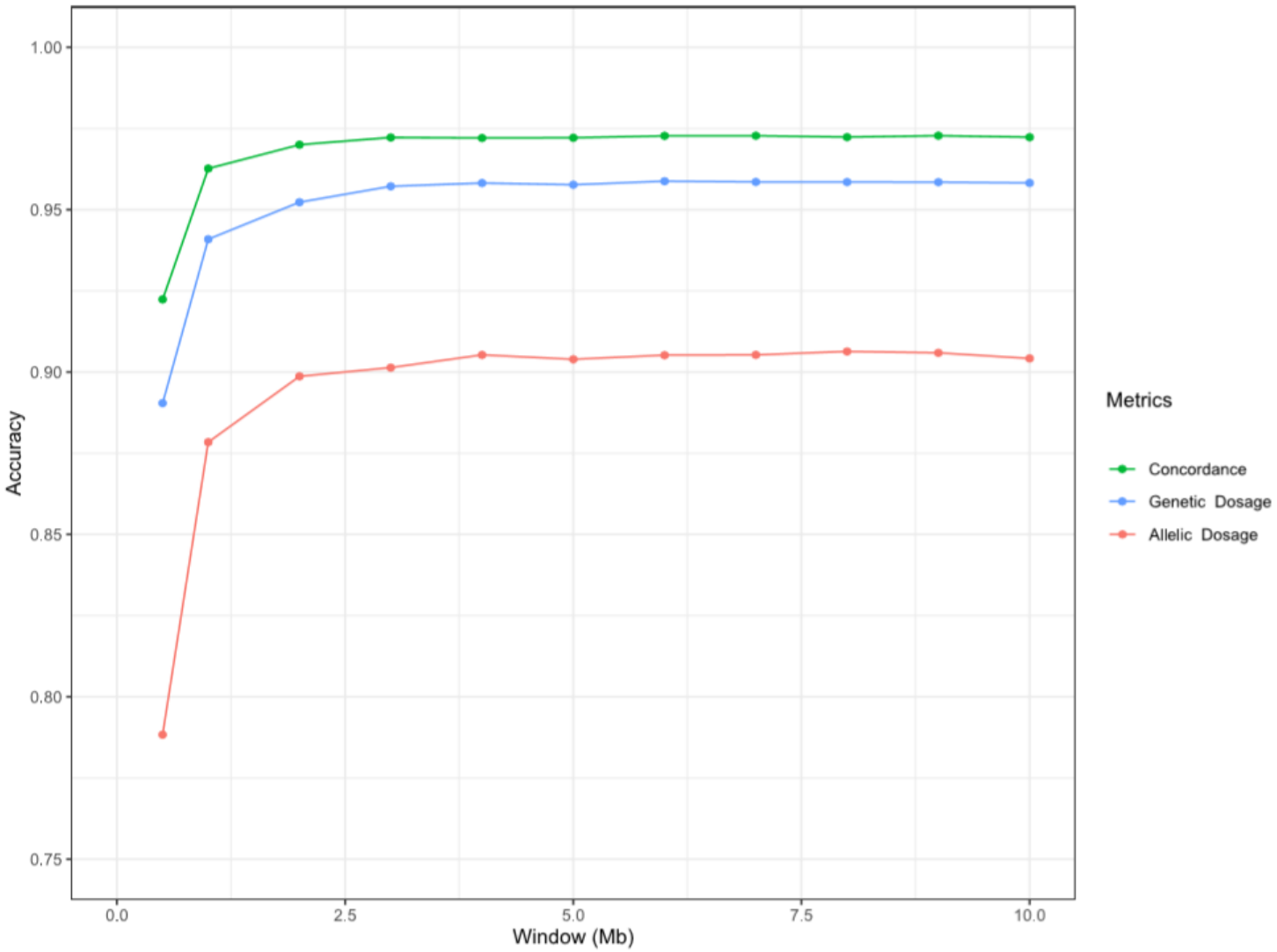


**Figure 1**

Graphical representation of the accuracy of the microsatellite imputation considering different window lengths in the imputation process. The x-axis represents the window sizes (in bp) considered in the imputation process. The y-axis represents the average of the imputation accuracy parameters (concordance [green], genotype dosage [blue] and allelic dosage [orange]) of the 19 microsatellites included in this study.

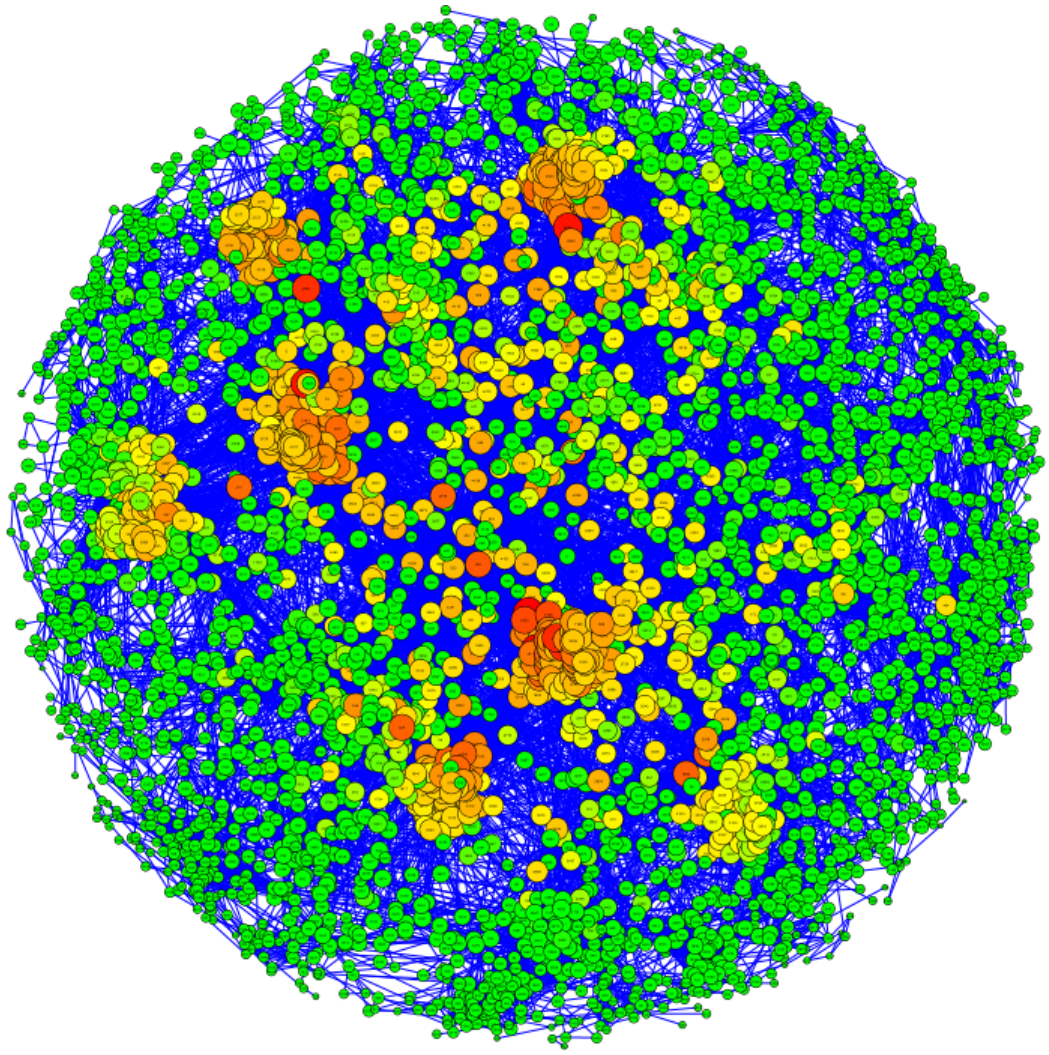

**Figure 2**

Population structure using the Pedigromics approach. Genomic relationships (>0.2) among the individuals are displayed. Each node represents one animal from the population. The color and the size of the nodes are based on the closeness coefficient, on a green to a red color scale, with the higher values represented by a large size and red color.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- FigureS1.pdf
- FigureS2.pdf
- FigureS3.pdf
- TableS1.xlsx
- TableS2.xlsx
- TableS3.xlsx
- TableS4.xlsx