

# Microsatellite imputation using SNP data for parentage verification in four Italian sheep breeds

Michela Ablondi<sup>1</sup>  | Giorgia Stocco<sup>1</sup>  | Matteo Cortellari<sup>2</sup>  | Antonello Carta<sup>3</sup>  |  
 Andrea Summer<sup>1</sup>  | Alessio Negro<sup>4</sup>  | Silverio Grande<sup>4</sup>  | Paola Crepaldi<sup>2</sup>  |  
 Claudio Cipolat-Gotet<sup>1</sup>  | Stefano Biffani<sup>5</sup> 

<sup>1</sup>Department of Veterinary Science, Università degli studi di Parma, Parma, Italy

<sup>2</sup>Dipartimento di Scienze Agrarie e Ambientali – Produzione, Territorio, Agroenergia, Università degli Studi di Milano, Milan, Italy

<sup>3</sup>Unità di Ricerca di Genetica e Biotecnologie, Agris Sardegna, Sassari, Italy

<sup>4</sup>Ufficio Studi, Associazione Nazionale della Pastorizia, Rome, Italy

<sup>5</sup>Consiglio Nazionale delle Ricerche (CNR), Istituto di Biologia e Biotecnologia Agraria (IBBA), Milan, Italy

## Correspondence

Claudio Cipolat-Gotet, Department of Veterinary Science, Università degli studi di Parma, Parma 43126, Italy.  
 Email: [claudio.cipolatgotet@unipr.it](mailto:claudio.cipolatgotet@unipr.it)

## Abstract

Microsatellite markers (MS) have been widely used for parentage verification in most of the livestock species over the past decades mainly due to their high polymorphic information content. In the genomic era, the spread of genotype information as single-nucleotide polymorphism (SNP) has raised the question to effectively use SNPs also for parentage testing. Despite the clear advantages of SNP panels in terms of cost, accuracy, and automation, the transition from MS to SNP markers for parentage verification is still very slow and, so far, only routinely applied in cattle. A major difficulty during this transition period is the need of SNP data for parents and offspring, which in most cases is not yet feasible due to the genotyping cost. To overcome the unavailability of same genotyping platform during the transition period, in this study we aimed to assess the feasibility of a MS imputation pipeline from SNPs in four native sheep dairy breeds: Comisana ( $N=331$ ), Massese ( $N=210$ ), Delle Langhe ( $N=59$ ) and Sarda ( $N=1003$ ). Those sheep were genotyped for 11 MS and with the Ovine SNP50 Bead Chip. Prior to imputation, a quality control (QC) was performed, and SNPs located within a window of 2 Mb from each MS were selected. The core of the developed pipeline was made up of three steps: (a) storing of both MS and SNP data in a Variant Call Format file, (b) masking MS information in a random sample of individuals (10%), (c) imputing masked MS based on non-missing individuals (90%) using an imputation program. The feasibility of the proposed methodology was assessed also among different training – testing split ratio, population size, number of flanking SNPs as well as within and among breeds. The accuracy of the MS imputation was assessed based on the genotype concordance as well as at parentage verification level in a subset of animals in which assigned parents' MS were available. A total of 8 MS passed the QC, and 505 SNPs were located within the  $\pm 2$  Mb window from each MS, with an average of 63 SNPs per MS. The results were encouraging since when excluding the worst imputed MS (*OARAE129*), and regardless on the analyses performed (within and across breeds) for all breeds, we

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2023 The Authors. *Journal of Animal Breeding and Genetics* published by John Wiley & Sons Ltd.

achieved an overall concordance rate over 94%. In addition, on average, the imputed offspring MS resulted in equivalent parentage outcome in 94% of the cases when compared to verification using original MS, highlighting both the feasibility and the eventual practical advantage of using this imputation pipeline.

#### KEYWORDS

imputation, local breeds, microsatellites, parentage testing, single-nucleotide polymorphism

## 1 | INTRODUCTION

Microsatellite markers (MS), also known as short tandem or simple sequence repeats, consist of motifs from 1 to 6 base pairs (bp) repeated in tandem. Thanks to their high polymorphic information content, MS markers have been widely used for parentage verification in most livestock species over the past decades (Glowatzki-Mullis et al., 2009; Visscher et al., 2002). Their impact on pedigree correctness has been enormous, both ensuring parental verification of registered animals, as well as parental identification when multi-sire candidates in the same herd are present (Davis & DeNise, 1998; Gomez-Raya et al., 2008). Indeed, pedigree correctness is a crucial requirement for the estimation of genetic parameters and breeding values, prediction of genetic gain, and genetic diversity management (Harlizius et al., 2011; Israel & Weller, 2000). However, it has been shown that allele scoring in MS is difficult to be accurately and fully automated, due to the presence of preferential allele amplification, imperfect repeats, null alleles and allelic dropouts (Kelly et al., 2011). Thus, manual MS scoring is still widely performed, although this might cause human errors, due to the high MS polymorphic nature (Baruch & Weller, 2008; Kelly et al., 2011). Concurrently, the spread of genotype information in the form of single-nucleotide polymorphism (SNP) has been massive, raising the question of the possibility to use SNPs for parentage testing instead of MS in the livestock breeding. However, this switch has been very slow and so far, routinely applied only in cattle (International Society for Animal Genetics, 2012). The SNPs are less informative due to their biallelic nature, causing the need of more markers to be typed for parentage verification (generally in the range of 200–700 SNPs) compared to MS (Strucken et al., 2016). Nevertheless, the most tangible benefits in the use of SNPs are the ease of automation and thus the standardization across laboratories, lower error rates as well as the possibility to use these technologies for several other applications, ranging from genomic selection to genetic diversity management. In this scenario, despite the clear advantages of SNP panels in accuracy and automation, a current major drawback in species where genotyping is not yet routinely performed is the need of the same genotyping technology for parents and offspring. Unfortunately, this is

not always possible, especially in small populations, where DNA sources of ancestors might not be any longer available and thus SNP genotyping cannot be performed. Moreover, to avoid extra costs in the case of parents with MS and offspring with SNP data and to overcome the issue of DNA unavailability, a possible solution is the imputation of MS alleles from SNP data. This approach has been successfully proposed in several cattle breeds (McClure et al., 2012, 2013) as well as in the Assaf sheep breed (Marina et al., 2021). In Italy, where over 6 million sheep and almost 100 breeds are reared (BDN, 2022), parentage verification using SNP data is not yet implemented although the number of genotyped animals has been increased in the latest years. The MS imputation based on SNP data might allow parentage verification when different genotype platforms are available and limited economic resources do not allow to test both parents and offspring with genotype data. This latter aspect could be particularly attractive for local breeds, where often the economic resources are restricted and which existence is tightly linked to their sociocultural and economic (e.g., protected designation of origin cheeses; PDO) contribution (Marsoner et al., 2018). This methodology might boost the transition from MS to SNPs in the field of parentage testing by providing the possibility to check parentage when different genotyping platform are available (i.e., parents with MS and offspring with SNP) encouraging the use of SNP in younger animals (without the need to still have MS data for those animals). This possibility might help the transition to the use of SNP data in the coming generations which will in turn allow to verify parentage directly by using SNP data only which has been proven to be the best option (McClure et al., 2018; Tortereau et al., 2017). Therefore, we aimed to evaluate the feasibility of MS imputation pipeline from a medium density SNP panel in four Italian sheep breeds to boost this transition period. Specifically, the objectives were to: (a) test the effect of the training/testing split ratio on the accuracy of MS imputation from SNPs data; (b) assess the accuracy of MS imputation via a within-breed analysis, (c) assess the accuracy of MS imputation via a across-breeds analysis from SNPs data coming from breeds representing the major breeding scenario in the Italian territory and (d) evaluate the concordance in trio parentage verification using original offspring MS and imputed ones.

## 2 | MATERIALS AND METHODS

### 2.1 | Dairy sheep breeds

Data for the present study were provided by the (Asso.Na.Pa.) and included both MS and SNP information from 1603 animals belonging to four native dairy breeds: Comisana ( $N=331$ ), Massese ( $N=210$ ), Delle Langhe ( $N=59$ ) and Sarda ( $N=1003$ ). Those animals were chosen to represent as much variability as possible within breed to reduce the potential bias due to highly related animals, with an average of 2.78 offspring per sire (Table 1).

The Sarda is an autochthonous breed of the Sardinia Island (Italy, Western Mediterranean Sea) and belong to the Mediterranean group. This breed accounts for more than 50% of whole sheep population in Italy (BDN, 2022) with over three million animals. Data were available from flocks both located in the Sardinia Island and from specialized dairy sheep farms from other Italian regions. The Sarda is a large size, polled breed with white wool. Most of the milk from Sarda ewes is used to produce three of the most important ovine PDO cheeses (Pazzola et al., 2014). The Comisana sheep breed belongs to the Mediterranean type and originated from the Maltese and Sicilian breeds in the late 19th and early 20th century in the Southeast region of Sicily Island (Italy, Central Mediterranean Sea). Also known as 'red head' due to its characteristic red face, the Comisana is a medium-large size, polled breed, with white wool. This breed is well adapted to the semi-arid Mediterranean environment representing an important resource for the marginal areas of Southern Italy (Selvaggi et al., 2017) with more than 60,000 animals (BDN, 2022). The Massese is an indigenous breed of Tuscany (Central Italy), it belongs to the Alpine group, and counts roughly 25,000 animals (BDN, 2022). Nowadays is mainly reared in the Tuscany and Emilia Romagna regions (Northern-Central Italy). This is a medium-small size, horned breed, with black wool. For the Comisana and Massese, data were available from the Asso.Na.Pa. nucleus farm, which manages the breeding and provides semen and rams for genetic improvement of these two breeds in commercial flocks. The Delle Langhe is an autochthonous breed of the Piedmont region (Northern Italy), native to the province of Cuneo, in the Delle Langhe hilly area. This breed accounts for about 6000 animals (BDN, 2022), whose milk is mostly used to produce typical cheeses and PDO products.

**TABLE 1** Population structure of the 1603 animals used in the study divided by breed: Comisana (COM), Massese (MAS) Delle Langhe (ODL), and Sarda (SAR).

	COM	MAS	ODL	SAR
Sire	187	151	14	269
Dam	169	252	33	887
Average offspring per sire	1.77	1.39	4.21	3.73
Average offspring per dam	1.96	0.83	1.79	1.13

The Delle Langhe is reared in the province of Cuneo, Asti and Savona (Northern Italy). It is a medium size polled breed, with white wool.

### 2.2 | Microsatellite and data preparation

Specifically, MS were sequenced and stored by the Associazione Nazionale della Pastorizia Asso.Na.Pa. as part of the official parentage verification routine. In 2015, SNP genotyping and storing for the Italian sheep breeds was established in the framework of two national projects, CHEESR and SHEEP&GOAT, both funded by the European Agricultural Fund for Rural Development (Measure 10.2: Support for conservation and sustainable use and development of genetic resources in agriculture). The sheep used in this study were genotyped with 11 MS routinely used for parentage testing and with the Ovine SNP50 Bead Chip. The position of each MS in the ovine genome (OAR\_v4.0) was obtained through the alignment of the primer sequences to the sheep reference genome using BLAT (Kent, 2002). Prior to imputation, a quality control (QC) was performed and MS with a call rate below 80% were removed. After the QC, each MS was recoded to fit the variant call format (VCF) required by the VCFtools software specifications (Danecek et al., 2011). The most common allele in each MS was considered as the reference and recoded as "0," whereas consecutive numbers (1, 2, 3, ...,  $n$ ) were assigned to the remaining ones based on the MS allele length. To keep haplotype diversity in the sample, a MAF filter was not performed in the case of SNP data and only SNPs and animals with a call rate lower than 90% were removed using PLINK v1.9 (Purcell et al., 2007). Plink v1.9 was also used to extract SNPs located within a window of  $\pm 2$ Mb (as suggested by Marina et al. (2021)) from each MS and to recode the SNP data in the VCF format (Purcell et al., 2007). The two VCF files were then merged using MergeVcfs in GATK version 4.1 (Auwera et al., 2013).

### 2.3 | Imputation procedure and accuracy evaluation

The imputation of MS based on SNP data was performed following the procedure described in Marina et al. (2021). The core of this methodology was made up

of three main steps: (a) storing of both MS and SNP data in a Variant Call Format (VCF) text file, (b) masking the MS information in a random sample of individuals, and (c) imputing masked MS based on SNP and MS data from individuals with non-masked information using an imputation program. In the [Figure S1](#), a thorough description of all the steps for the imputation procedure is available.

To test the feasibility of the proposed methodology with different sample size as well as within and across breeds the following approach was used. Specifically, the accuracy of MS imputation were assessed by three different Cross-Validation (CV) procedures:

- (i) To evaluate the effect of the training/testing split ratio, only the data available for the Sarda breed were used. We tested different split ratio from 10% training and 90% testing till 90% training and 10% testing of the total records, increasing each run of 10% the training set. To account for individual variability, all the procedures were repeated 20 times per each of the nine different split ratio scenarios randomly selecting the test and reference populations each time. Furthermore, the nine split ratio scenarios were run while increasing the population size (randomly sampling 50, 100, 500, 1000 animals – 20 replicates each) to evaluate the combined effect of split ratio and population size.
- (ii) To evaluate the imputation accuracy within-breed, the available records within breed were splitted into a training and a testing dataset that contained 90% (training) and 10% (testing) of the total records, respectively. To account for individual variability, the procedure was repeated 100 times.
- (iii) To evaluate the imputation accuracy among breeds, all the records were splitted into a training and a testing dataset that contained 90% (training) and 10% (testing) of the total records, respectively. In this CV, a proportional random sampling (10% within breed) was used in order to ensure that, in each training and testing sets, a balanced number of samples per each of the four breeds was included. To account for individual variability, the procedure was repeated 100 times.

For all the aforementioned CV procedures, the MS information was masked in the testing set, and consequently phased and imputed in BEAGLE 5.4 software (Browning et al., 2018) using the training VCF file that contained both MS and SNP data. Phasing and imputation were implemented after correcting for small effective population size (Ciani et al., 2014).

The accuracy of the MS imputation was assessed based on the genotype concordance rate, which was defined as 0 if none of the imputed alleles matched the true allele, 0.5 if only one of them matched, and 1 if both alleles matched the true alleles. The imputation accuracy was established for each MS separately and calculated as the average over all the MS within-breed as well as in the across-breeds analysis. In addition, in the case of Comisana and Massese, for over 60% of the animals, assigned parents' MS were available. Thus, a further accuracy evaluation was tested via the comparison between parentage verification performed using original offspring MS and imputed ones. The parentage verification was applied in Cervus, which uses maximum-likelihood methods to predict parent-offspring relationships (Kalinowski et al., 2007). A concordance rate was established for trios given that if parentage was either verified or discarded based on both methods for that animal a value of one was assigned, otherwise a zero. We scored the parentage verification at trio level and not as performed at allele level (0, 0.5 and 1), since this is what is going to happen in practice. Finally, since the number of SNPs located within a window of  $\pm 2$ Mb from the MS was significantly lower for one MS (*OARAE129*–10 SNPs instead of over 50 in the remaining ones) we randomly selected 10 SNPs within the  $\pm 2$ Mb window per each MS to assess the effect of SNPs number on the concordance rate. The latter analysis was performed in the Sarda breed as it is the breed with the highest sample size.

## 3 | RESULTS

### 3.1 | Quality control (QC)

The MS assessed in the present study are presented in [Table 2](#). Specifically, eight out of the 11 MS routinely used for parentage verification passed the QC and were further used in the imputation process. The number of alleles per MS ranged between 12 in the case of the *OARAE129* to 28 for *OARCP49*. The variability of the MS allele size weakly differed across breeds. The least variable MS, although presenting 18 allelic size in the whole population, was the *FCB304* in the Comisana and in the Sarda with the most common allele size being the 170 found in 46.4% and 71.0% of the sheep, respectively. In the Massese, the least variable MS was the *MAF214* with 62.0% of the sheep carrying the 189 allelic sizes, while in the whole population 14 allelic sizes were observed. In contrast, in the Delle Langhe breed, the least variable MS was the *OARAE129* with 70.0% of the animals carrying the 147 allelic size ([Figure 1](#)). Although few breed-specific MS allele size were found in the dataset, they had low allele frequencies. A total of 506 SNPs

**TABLE 2** Microsatellite (MS) used in the current study. For each MS, the following information are shown: the genome position in the Oar\_4.0, the allele length expressed in base pairs, the number of alleles per MS and the number of SNPs located within a window size of  $\pm 2$  Mb from the start position of each MS.

Microsatellite	Chr.	Position (bp)	Range	No. alleles	No. SNPs
<i>FCB11</i>	2	248,883,054	118–144	15	50
<i>OARAE129</i>	5	78,045,895	135–165	12	10
<i>CSRD247</i>	14	15,515,671	211–257	22	69
<i>INRA063</i>	14	41,508,190	167–207	19	76
<i>MAF214</i>	16	33,667,891	181–265	14	75
<i>OARCP49</i>	17	14,426,572	76–134	28	75
<i>FCB304</i>	19	49,669,589	146–188	18	75
<i>HSC</i>	20	25,684,012	263–297	18	76

were located within a window size of  $\pm 2$  Mb from each MS, with an average of 63 SNPs per MS (Table 2). In the case of the *OARAE129* MS, only 10 SNPs were located in the  $\pm 2$  Mb window size, while 76 SNPs overlapped the *INRA063* and *HSC* MS  $\pm 2$  Mb region. Thus, since some level of breed similarity at MS level was visible, testing an imputation of all breeds together seemed feasible and further tested in the across-breeds analysis.

### 3.2 | Training – testing Split ratio evaluation

Prior to testing the within and among breed procedures, the effect of data dimension in the developing an MS-SNP imputation model has been quantified in the Sarda breed (Figure 2). The average of the concordance rate across split ratio scenario varied considerably with a non-linear increase between the lowest in the case of split ratio 10:90 (mean = 75%) and the highest in the 90:10 scenario (mean = 91%). The variability of replicates ( $N = 20$ ) across the tested split ratio scenarios assumed erratic values. The pattern observed in Figure 2 shows the importance of carefully consider the proportion of animals to include in the training and in the testing sets for monitoring the concordance rate. Specifically, the 90:10 scenario resulted in the highest concordance rate while the largest variation across replicates was found in the case of 50:50 scenario.

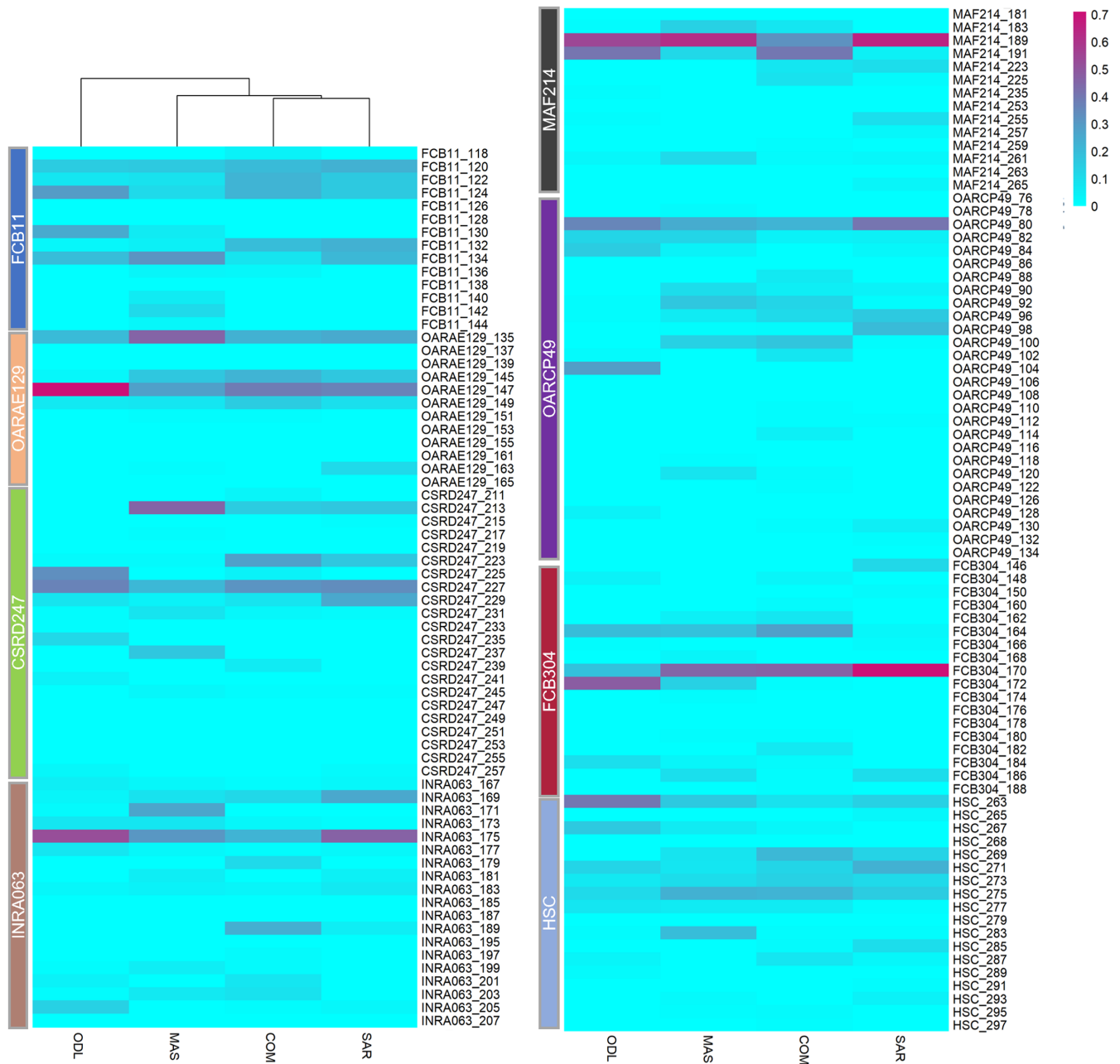
In Figure 3, the concordance rate per each split ratio scenario across MS is presented. Although the average values of concordance rates were quite different, the patterns from the 10:90 to 90:10 split ratio scenarios were similar across MS, except in the case of *OARAE129*. Most of the MS seemed to reach a concordance rate plateau from 40:60 split ratio scenario onwards (Figure 3). The highest average concordance rate across split ratio scenarios was achieved by the *FCB304* with a mean of 96% (SD = 0.02) followed by *CSRD247* (mean = 94%, SD = 0.03). The *OARAE129* performed always worse than the others having an average concordance rate of 59% (SD = 0.16) across

split ratio scenarios and presenting a cubic pattern across split ratio scenarios. The evaluation of the combined effect of split-ratio and population size showed that when population size was over 500 animals, already with 50% of the animals in the training set – which translates in 250 animals in practice – the imputation accuracy was over 85% and it reached a plateau at this point. Similar pattern was shown for the 1000 animals' scenario already at the 30:70 split ratio scenario which corroborates the hypothesis that the accuracy was roughly 85% even if a limited number of animals were included in the training set. Likewise for the 50 and 100 animals' scenario, a plateau was visible although more variability was present across validation sets (Figure 4). In Figure 5 is shown the concordance rate for the 8 MS when only 10 SNPs within the  $\pm 2$  Mb window were retrieved to assess the effect of SNPs number in the concordance rate. The highest concordance rate was still found for the *FCB304*, which reached a mean of 93.7%, and in contrast there were 4 MS that performed worse than the *OARAE129* (Figure 5).

### 3.3 | Within and among breeds analysis

Figure 6 shows the concordance rate from the within breed analysis per each different MS and breed. The within breed average concordance rate ranged from 91.0% (SD = 0.04) in the Delle Langhe to 95.0% (SD = 0.03) in the Sarda breed across MS. The concordance rate varied among MS as well, with the least performing one being the *OARAE129* in all within breed analysis. In contrast, the highest concordance rate was found for the *FCB304* for all the breeds (mean = 98.3%) except for the Comisana where the *MAF214* reached a concordance rate of 98.6% (Figure 6).

The average concordance rate, calculated from all MS and regardless breed type, in the among breeds analyses was equal to 94% (SD = 0.04). In particular, the highest concordance rate was 98% (SD = 0.007) for the *FCB304*, whereas the lowest was 84% (SD = 0.023) for the



**FIGURE 1** Allele frequency per each microsatellite divided by the breeds used in this study: Delle Langhe (ODL), Massese (MAS), Comisana (COM) and Sarda (SAR), respectively. The name and length of the MS is expressed as name plus length divided by \_ on the right of the figure and the frequency is shown as sequential palette with lighter blue being the lowest frequent allele and the dark pink being the most frequent one. Hierarchical clustering of MS is shown as dendrogram at the top of the figure.

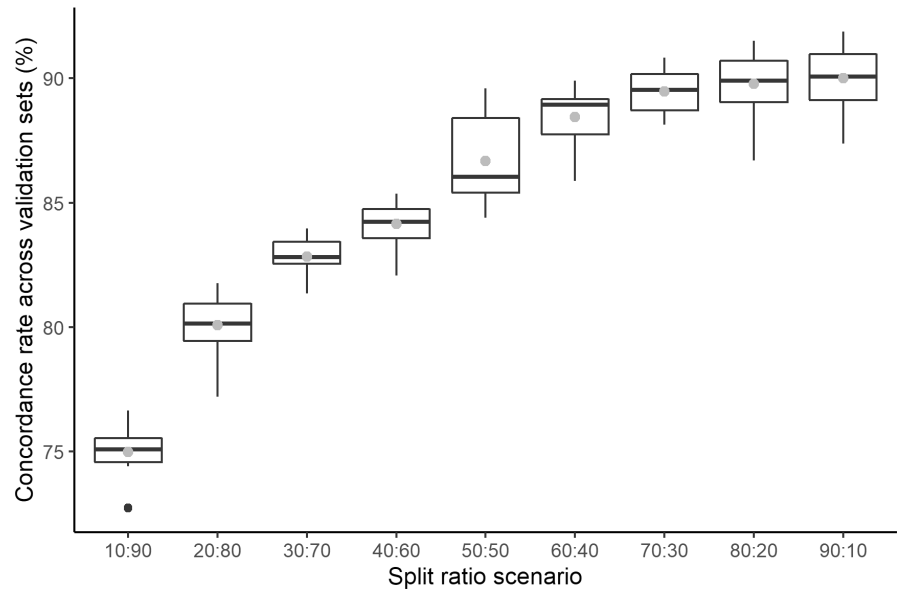
*OARAE129*. Similar trends were found when evaluating the concordance rate per MS within breed in the among breeds analysis. The lowest concordance rate was found for the *OARAE129* for all breeds ranging between 68% for the Delle Langhe to 86% for the Sarda. The *FCB304* was the best performing one in all breeds except for the Comisana where the best concordance rate was found for the *MAF214* (98%; Figure 7). The breed with the highest concordance rate and lowest standard deviation was the Sarda ( $94\% \pm 0.3$ ), whereas the lowest concordance was

found in the Delle Langhe with an average rate of 89% and the largest variation across testing sets (Figure 7).

### 3.4 | Concordance rate at trio parentage verification

For 60% of Comisana and Massese, the assigned parents' MS were available and thus, the parentage verification was performed using both original and imputed offspring

**FIGURE 2** Boxplot of concordance rate (%) across validation sets per each split ratio scenario in the Sarda breed. The mean is shown as grey dot, the median as straight line and outliers as black dots.



MS based on the within breed imputation pipeline. A total of 2985 imputed MS-animal combinations based on the 100 times imputation procedure was available and used to assess parentage verification concordance rate. The comparison with the parentage verification using original offspring MS resulted in an overall accuracy of 95.6% in the case of Comisana and of 93.1% in the Massese breed, respectively.

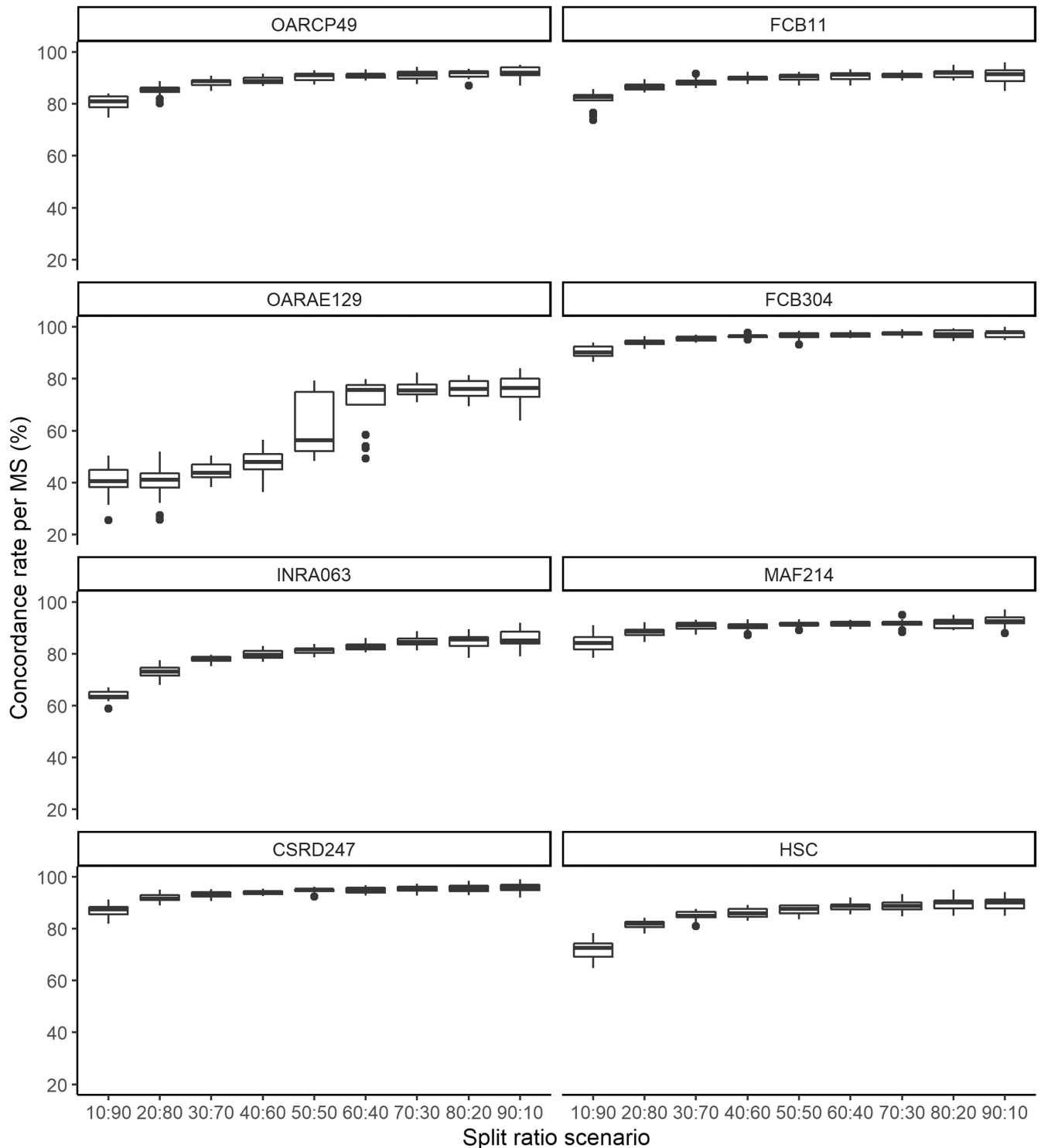
## 4 | DISCUSSION

In this study, we tackled the need to evaluate a method which might bridge traditional and new technologies in the field of parentage testing in sheep breeds. Thus, this study comprised a total of 1603 animals from the most important four Italian sheep breeds that depict different farming system scenarios from the Pulina et al. (2018). The experimental design employed followed the criteria on optimal window distance suggested in a previous study in sheep, where SNPs located within a window size of  $\pm 2$  Mb from each MS were used in the imputation strategies (Marina et al., 2021). This approach allows to keep the number of SNPs used in the imputation process to a limited number ( $N=506$ ), which in turn results to a cost-effective solution which might be used in practice.

Another key aspect to consider prior to potentially use this methodology in practice, it is to assess how much variability to include in the training test to ensure high accuracy in the imputation. Therefore, in this study, we tested different proportions of the data to include in the training set to dig into how the performance of the model changes as the amount of training data changes. For this first objective, we limited the analysis to the Sarda breed due to the highest number of observations available in this

breed ( $N=1003$ ) and since it represents the most spread local breed in the Italian territory. As expected, the 90:10 scenario resulted in the highest concordance rate (above 90%), which highlights the importance to include in the training set as much variability as possible to achieve accurate results. Nevertheless, it is important to also consider that by excluding the worst performing MS (*OARAE129*), the average concordance rate increased considerably (from  $\sim 40\%$  to 88.8%) regardless of the split ratio scenario. In addition, several MS seem to reach a concordance plateau from 40:60 split ratio scenario onwards except for the *FCB304* which reached at scenario 10:90 already a concordance rate of 91%. Conservatively and to be aligned with previous studies, we opted to use the 90:10 CV scenario in the within-breed and across-breeds analyses (Marina et al., 2021; McClure et al., 2013; Nolte et al., 2022). It is key to realize that the objective of the cross-validation using different split ratio was to test the effectiveness of the methodology and its accuracy while including different percentage of the whole variability in the data. At the field level, all individuals with both SNPs and MS have to be used in the imputation. Nevertheless, the analysis combining different split ratios and population sizes indicated that the methodology is robust also when a limited number of animals have both SNPs and MS.

The within breed analyses showed a concordance rate over 91% in all the tested breeds highlighting that MS genotypes can be accurately imputed based on SNP genotypes. However, similarly to what presented above, differences among MS imputation performances were shown. Yet, the worst imputed MS was the *OARAE129*, and this result was found also in the among breeds analyses. Out of all the MS tested, the *OARAE129* clearly represented the greatest challenge and in all analyses had the lowest concordance rate. While this marker

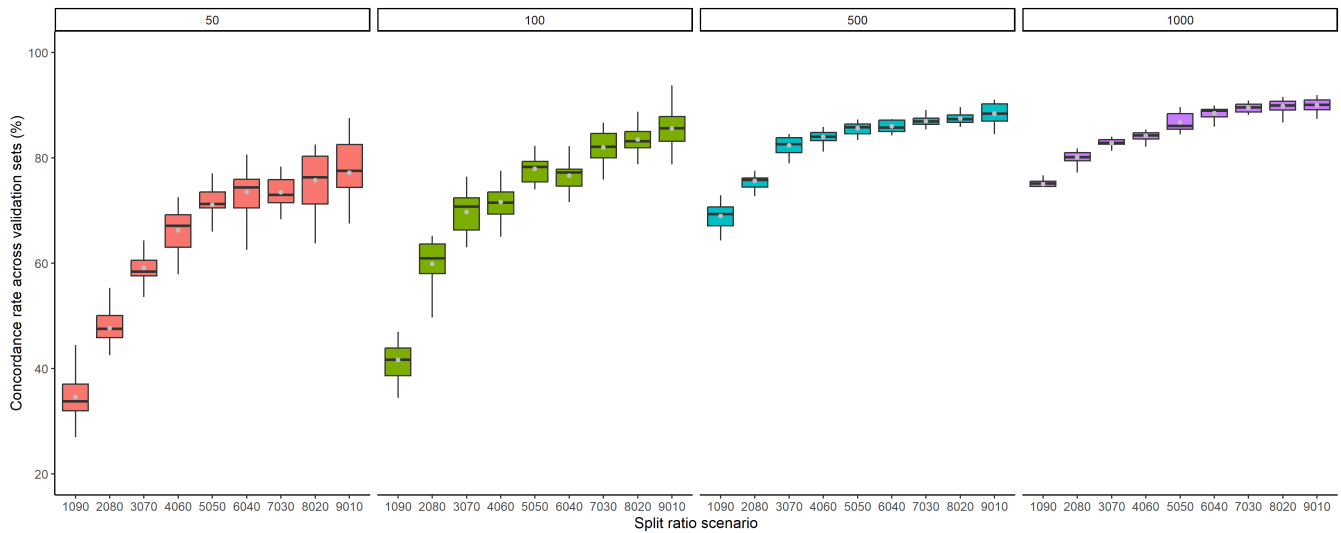


**FIGURE 3** Boxplot of concordance rate (%) across testing sets per each split ratio scenario per MS in the Sarda breed. The median is shown as straight line and outliers as black dots.

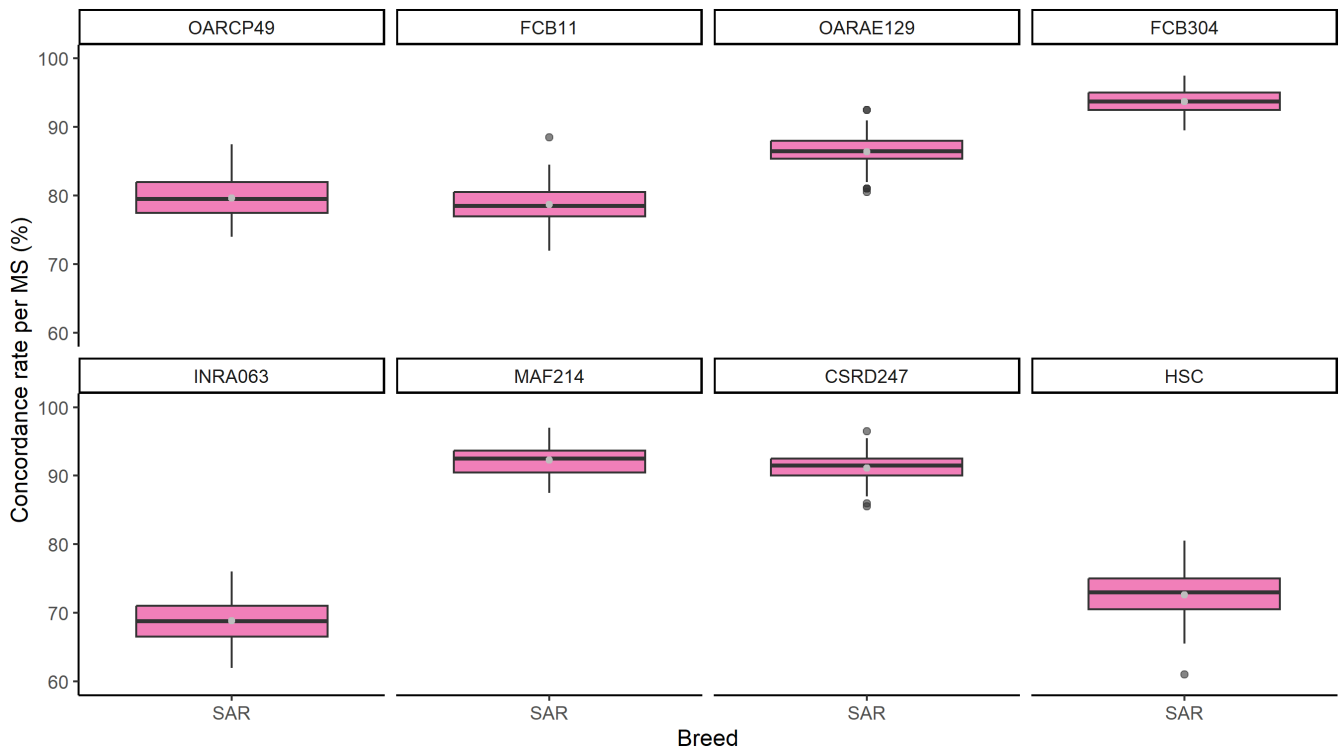
performed slightly better in the within breed analysis (mean = 83.36%) compared to the among breeds analysis (mean = 79.64%), both analyses resulted into an insufficient concordance rate. A possible explanation might be related to the position of this MS, which is located on chromosome 5:78045895 and where a limited

number of SNPs overlapped the  $\pm 2$ Mb window from *OARAE129*. Indeed, only 10 SNPs downstream of this MS were used for imputation which might prejudice the concordance rate. This potential explanation is in line with what previously found in horses where a MS located on a telomeric region end, which overlapped





**FIGURE 4** Boxplot of concordance rate (%) across validation sets per each split ratio scenario in the Sarda breed randomly sampling 50, 100, 500 and 1000 animals. The mean is shown as grey dot and the median as straight line.

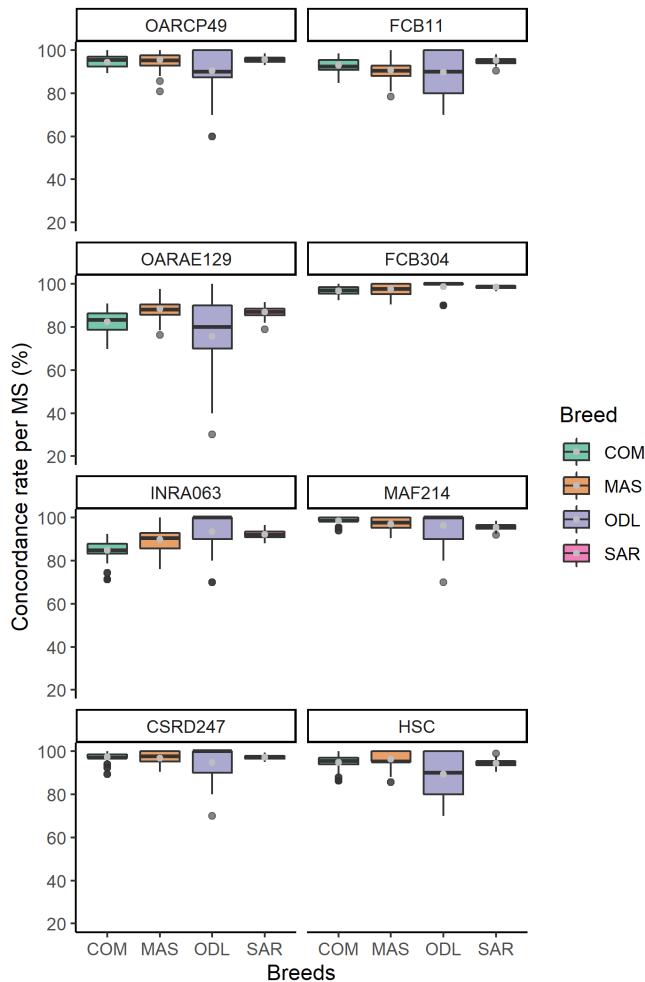


**FIGURE 5** Boxplot of concordance rate (%) per MS in the Sarda breed when only 10 SNPs within the  $\pm 2$ Mb window per each MS were retrieved to assess the effect of SNPs number in the concordance rate. The mean is shown as grey dot and the median as straight line.

with eight SNPs only, yielded to worst accuracy results (Nolte et al., 2022). This hypothesis was corroborated by performing the analysis with 10 flanking SNPs only from each MS where all MS showed concordance rate lower than what found including all SNPs available in the  $\pm 2$ Mb window. On top of that, we found MS that performed even worse than the *OARAE129*, which indicated the importance of the combined effect of the

number of flanking SNPs and their informative value on the imputation accuracy.

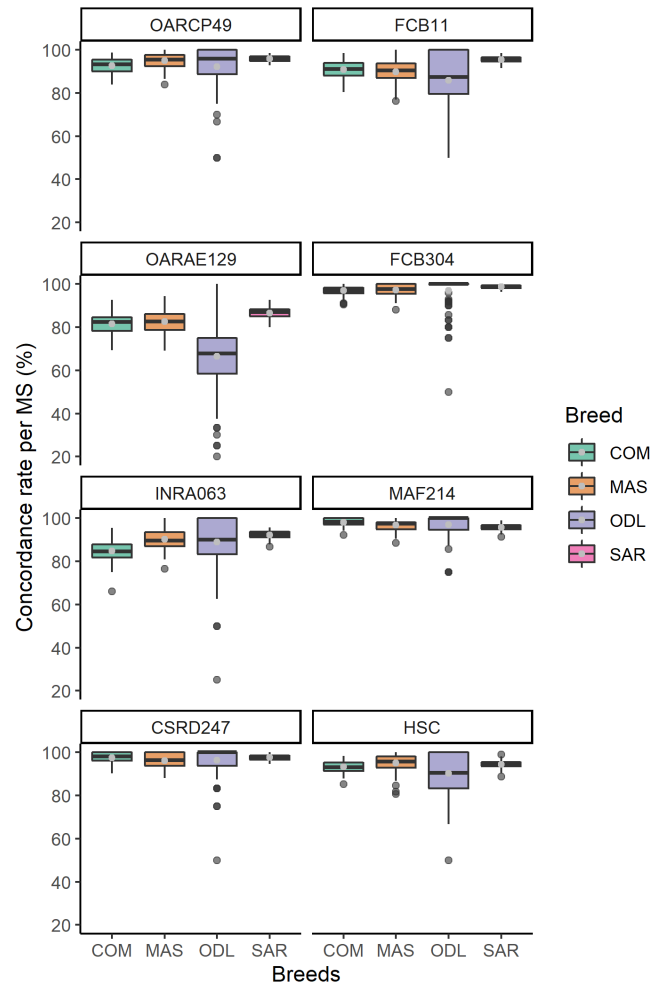
The breed with the lowest concordance rate both in the within and among breeds analyses was the Delle Langhe, which reached an overall imputation accuracy of over 90% only in the within-breed analysis. Unfortunately, this might be due to the very limited number of animals available for this breed, which might



**FIGURE 6** Boxplot of concordance rate (%) from the within breed analysis per each different microsatellite and breed. The mean is shown as grey dot, the median as straight line and outliers as black dots. Comisana (COM), Massese (MAS), Delle Langhe (ODL) and Sarda (SAR).

not be sufficient to underline the genetic variability present in the Delle Langhe breed. Moreover, we cannot either rule out the hypothesis that a peculiar genetic structure of this breed might determine those results. Still, we decided to keep this breed in the model since it is quite remarkable that although the reference panel is so limited in numbers, we could nonetheless reach appreciable level of concordance rate, especially when we did not consider the *OARAE129* (within breed and among breeds average were 93.3% and 92.6%, respectively). We decided to keep the Delle Langhe breed since it resembles several small autochthonous populations reared in the Italian context.

The breeds analysed in this study represent different scenario in terms of breeding management although all are officially milk-recorded and routinely genetically evaluated for milk yields. The Sarda is managed using a pyramidal scheme (Carta et al., 2009), while the selection



**FIGURE 7** Boxplot of concordance rate (%) from the among breeds analysis per each different microsatellite and breed. The mean is shown as grey dot, the median as straight line and outliers as black dots. Comisana (COM), Massese (MAS), Delle Langhe (ODL) and Sarda (SAR).

scheme for Comisana and Massese is based on two different closed nucleus flocks reared at the Genetic Center of Asciano (Siena, Italy). The selection in Delle Langhe is mainly performed using an estimated breeding value to improve milk production. In addition, breeders can use a software to optimize genetic breeding values and manage mating to reduce inbreeding in the coming generations. Therefore, the inclusion of all those different types of breeding scheme paves the way to ensure that this imputation approach can be efficiently used in the Italian sheep scenario which is quite heterogeneous in terms of breeding strategies. We believe that by including more animals we will be able to account for most of the variability present in the Italian sheep population which in turn will lead to higher accuracy in the imputation process. Indeed, although the total number of animals tested in this study was lower than what found in Marina et al., the concordance rates were on the same scale (Marina et al., 2021).

It is known that in local breeds, the imputation accuracies might be overestimated mostly due to either highly structured and related population (Bolormaa et al., 2015) or due to a low effective population size (Druet et al., 2014). Although all tested breeds were limited in size, there are recent evidence that their inbreeding level is modest since all breeds except Sarda showed average inbreeding between 1% and 6% based on Runs of Homozygosity estimate ( $F_{ROH}$ ) (Cortellari et al., 2022). The Sarda breed showed an average  $F_{ROH}$  of 13.5%, which might partially explain, together with the highest number of animals considered in the study, why the highest concordance rate was found in this breed. Finally, thanks to the availability of MS data of the assigned parents in the case of Comisana and Massese breeds, we could bring forward the evaluation of the accuracy of the imputation pipeline. The concordance between parentage verification performed using original offspring MS and imputed ones was on average over 94%, which encourages the implementation of this imputation pipeline in practice.

## 5 | CONCLUSIONS

This study represents the first attempt to evaluate a multi-breed imputation approach in several Italian breeds to help the transition from MS to SNPs data in the field of parentage testing. The results are encouraging since when excluding the worst imputed MS (*OARAE129*) and regardless on the analyses performed (within and among breeds) for all breeds we achieved an overall concordance rate over 94%. In addition, at the parentage verification level, the imputed offspring MS resulted in equivalent diagnoses for about 94% of the cases in the Comisana and Massese breeds, where assigned parents' MS were available. By combining different split ratios and population sizes we showed a concordance rate plateau of over 85% when at least 250 animals were included in the training set. This latter aspect indicated that the imputation can be used also when a limited number of animals have both MS and SNP information. At the field level, all animals with both information must be included in the imputation pipeline to ensure higher accuracy rate. Moreover, this approach should be tested in other species and scenarios. A technical note, when performing this type of study, is key to consider which set of MS was used and with which reference genome was performed the alignment of MS position.

In the future, the use of this approach in practice might lead to a substantial reduction in cost for breeders and breeding associations since they might not have to double genotype animals to test parentage. This in turn might encourage the use of SNP data only in young animals, which will allow to a full transition of SNP data for parentage verification in the future.

## FUNDING INFORMATION

There are no funders.

## CONFLICT OF INTEREST STATEMENT

None.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.


## ORCID

Michela Ablondi  <https://orcid.org/0000-0003-3700-1042>

Giorgia Stocco  <https://orcid.org/0000-0002-6786-9806>

Matteo Cortellari  <https://orcid.org/0000-0002-5161-0648>

Antonello Carta  <https://orcid.org/0000-0003-0534-4876>

Andrea Summer  <https://orcid.org/0000-0002-4833-657X>

Alessio Negro  <https://orcid.org/0000-0002-2433-0210>

Silverio Grande  <https://orcid.org/0000-0001-8685-6202>


Paola Crepaldi  <https://orcid.org/0000-0002-6526-2162>

Claudio Cipolat-Gotet  <https://orcid.org/0000-0002-2318-4231>

Stefano Biffani  <https://orcid.org/0000-0001-5559-3630>

Stefano Biffani  <https://orcid.org/0000-0001-5559-3630>

Stefano Biffani  <https://orcid.org/0000-0001-5559-3630>

Stefano Biffani  <https://orcid.org/0000-0001-5559-3630>

Stefano Biffani  <https://orcid.org/0000-0001-5559-3630>

## REFERENCES

- Auwers, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altschuler, D., Gabriel, S., & DePristo, M. A. (2013). From FastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, 43, 11.10.1–11.10.33. <https://doi.org/10.1002/0471250953.bi1110s43>
- Baruch, E., & Weller, J. I. (2008). Estimation of the number of SNP genetic markers required for parentage verification. *Animal Genetics*, 39, 474–479. <https://doi.org/10.1111/j.1365-2052.2008.01754.x>
- BDN, S.I.V. (2022). *Consistenza Capi Ovini e Caprini Identificati Singolarmente per Razza*. [https://www.vetinfo.it/j6\\_statistiche/#/report-pbi/89](https://www.vetinfo.it/j6_statistiche/#/report-pbi/89)
- Bolormaa, S., Gore, K., van der Werf, J. H. J., Hayes, B. J., & Daetwyler, H. D. (2015). Design of a low-density SNP chip for the main Australian sheep breeds and its effect on imputation and genomic prediction accuracy. *Animal Genetics*, 46, 544–556. <https://doi.org/10.1111/age.12340>
- Browning, B. L., Zhou, Y., & Browning, S. R. (2018). A one-penny imputed genome from next-generation reference panels. *American Journal of Human Genetics*, 103, 338–348. <https://doi.org/10.1016/j.ajhg.2018.07.015>
- Carta, A., Casu, S., & Salaris, S. (2009). Invited review: Current state of genetic improvement in dairy sheep. *Journal of Dairy Science*, 92, 5814–5833. <https://doi.org/10.3168/jds.2009-2479>

- Ciani, E., Crepaldi, P., Nicoloso, L., Lasagna, E., Sarti, F. M., Moiola, B., Napolitano, F., Carta, A., Usai, G., D'Andrea, M., Marletta, D., Ciampolini, R., Riggio, V., Occidente, M., Matassino, D., Kompan, D., Modesto, P., Macciotta, N., Ajmone-Marsan, P., & Pilla, F. (2014). Genome-wide analysis of Italian sheep diversity reveals a strong geographic pattern and cryptic relationships between breeds. *Animal Genetics*, *45*, 256–266. <https://doi.org/10.1111/age.12106>
- Cortellari, M., Negro, A., Bionda, A., Grande, S., Cesarani, A., Carta, A., Macciotta, N., Biffani, S., & Crepaldi, P. (2022). Using pedigree and genomic data toward better Management of Inbreeding in Italian dairy sheep and Goat breeds. *Animals*, *12*, 2828. <https://doi.org/10.3390/ani12202828>
- Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*, 2156–2158. <https://doi.org/10.1093/bioinformatics/btr330>
- Davis, G. P., & DeNise, S. K. (1998). The impact of genetic markers on selection. *Journal of Animal Science*, *76*, 2331. <https://doi.org/10.2527/1998.7692331x>
- Druet, T., Macleod, I. M., & Hayes, B. J. (2014). Toward genomic prediction from whole-genome sequence data: Impact of sequencing design on genotype imputation and accuracy of predictions. *Heredity*, *112*, 39–47. <https://doi.org/10.1038/hdy.2013.13>
- Glowatzki-Mullis, M.-L., Gaillard, C., Wigger, G., & Fries, R. (2009). Microsatellite-based parentage control in cattle. *Animal Genetics*, *26*, 7–12. <https://doi.org/10.1111/j.1365-2052.1995.tb02612.x>
- Gomez-Raya, L., Priest, K., Rauw, W. M., Okomo-Adhiambo, M., Thain, D., Bruce, B., Rink, A., Torell, R., Grellman, L., Narayanan, R., & Beattie, C. W. (2008). The value of DNA paternity identification in beef cattle: Examples from Nevada's free-range ranches. *Journal of Animal Science*, *86*, 17–24. <https://doi.org/10.2527/jas.2007-0068>
- Harlizius, B., Lopes, M. S., Duijvesteijn, N., van de Goor, L. H. P., van Haeringen, W. A., Panneman, H., Guimarães, S. E. F., Merks, J. W. M., & Knol, E. F. (2011). A single nucleotide polymorphism set for paternal identification to reduce the costs of trait recording in commercial pig breeding. *Journal of Animal Science*, *89*, 1661–1668. <https://doi.org/10.2527/jas.2010-3347>
- International Society for Animal Genetics. (2012). *Guidelines for cattle parentage verification based on SNP markers*. Cairns.
- Israel, C., & Weller, J. I. (2000). Effect of misidentification on genetic gain and estimation of breeding value in dairy cattle populations. *Journal of Dairy Science*, *83*, 181–187. [https://doi.org/10.3168/jds.S0022-0302\(00\)74869-7](https://doi.org/10.3168/jds.S0022-0302(00)74869-7)
- Kalinowski, S. T., Taper, M. L., & Marshall, T. C. (2007). Revising how the computer program cervus accommodates genotyping error increases success in paternity assignment. *Molecular Ecology*, *16*, 1099–1106. <https://doi.org/10.1111/j.1365-294X.2007.03089.x>
- Kelly, A. C., Mateus-Pinilla, N. E., Douglas, M., Douglas, M., Shelton, P., & Novakofski, J. (2011). Microsatellites behaving badly: Empirical evaluation of genotyping errors and subsequent impacts on population studies. *Genetics and Molecular Research*, *10*, 2534–2553. <https://doi.org/10.4238/2011.October.19.1>
- Kent, W. J. (2002). BLAT—The BLAST-like alignment tool. *Genome Research*, *12*, 656–664. <https://doi.org/10.1101/gr.229202>
- Marina, H., Suarez-Vega, A., Pelayo, R., Gutiérrez-Gil, B., Reverter, A., Esteban-Blanco, C., & Arranz, J. J. (2021). Accuracy of imputation of microsatellite markers from a 50K SNP Chip in Spanish Assaf sheep. *Animals*, *11*, 86. <https://doi.org/10.3390/ani11010086>
- Marsoner, T., Egarter Vigl, L., Manck, F., Jaritz, G., Tappeiner, U., & Tasser, E. (2018). Indigenous livestock breeds as indicators for cultural ecosystem services: A spatial analysis within the alpine space. *Ecological Indicators*, *94*, 55–63. <https://doi.org/10.1016/j.ecolind.2017.06.046>
- McClure, M., Sonstegard, T., Wiggans, G., & Van Tassell, C. P. (2012). Imputation of microsatellite alleles from dense SNP genotypes for parental verification. *Frontiers in Genetics*, *3*, 140. <https://doi.org/10.3389/fgene.2012.00140>
- McClure, M. C., McCarthy, J., Flynn, P., McClure, J. C., Dair, E., O'Connell, D. K., & Kearney, J. F. (2018). SNP data quality control in a National Beef and dairy cattle system and highly accurate SNP based parentage verification and identification. *Frontiers in Genetics*, *9*, 84. <https://doi.org/10.3389/fgene.2018.00084>
- McClure, M. C., Sonstegard, T. S., Wiggans, G. R., Van Eenennaam, A. L., Weber, K. L., Penedo, C. T., Berry, D. P., Flynn, J., Garcia, J. F., Carmo, A. S., Regitano, L. C. A., Albuquerque, M., Silva, M. V. G. B., Machado, M. A., Coffey, M., Moore, K., Boscher, M.-Y., Genestout, L., Mazza, R., ... Van Tassell, C. P. (2013). Imputation of microsatellite alleles from dense SNP genotypes for parentage verification across multiple *Bos taurus* and *Bos indicus* breeds. *Frontiers in Genetics*, *4*, 1–11. <https://doi.org/10.3389/fgene.2013.00176>
- Nolte, W., Alkhoder, H., Wobbe, M., Stock, K. F., Kalm, E., Vosgerau, S., Krattenmacher, N., Thaller, G., Tetens, J., & Kühn, C. (2022). Replacement of microsatellite markers by imputed medium-density SNP arrays for parentage control in German warm-blood horses. *Journal of Applied Genetics*, *63*, 783–792. <https://doi.org/10.1007/s13353-022-00725-9>
- Pazzola, M., Dettori, M. L., Cipolat-Gotet, C., Cecchinato, A., Bittante, G., & Vacca, G. M. (2014). Phenotypic factors affecting coagulation properties of milk from Sarda ewes. *Journal of Dairy Science*, *97*, 7247–7257. <https://doi.org/10.3168/jds.2014-8138>
- Pulina, P. (2018). Evaluation of rural development policies in the post 2020 perspective. *Agriregionieuropa*, *14*.
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*, 559–575. <https://doi.org/10.1086/519795>
- Selvaggi, M., D'Alessandro, A. G., & Dario, C. (2017). Environmental and genetic factors affecting milk yield and quality in three Italian sheep breeds. *The Journal of Dairy Research*, *84*, 27–31. <https://doi.org/10.1017/S0022029916000765>
- Strucken, E. M., Lee, S. H., Lee, H. K., Song, K. D., Gibson, J. P., & Gondro, C. (2016). How many markers are enough? Factors influencing parentage testing in different livestock populations. *Journal of Animal Breeding and Genetics*, *133*, 13–23. <https://doi.org/10.1111/jbg.12179>
- Tortereau, F., Moreno, C. R., Tosser-Klopp, G., Servin, B., & Raoul, J. (2017). Development of a SNP panel dedicated to parentage



assignment in French sheep populations. *BMC Genetics*, 18, 50.

<https://doi.org/10.1186/s12863-017-0518-2>

Visscher, P. M., Woolliams, J. A., Smith, D., & Williams, J. L. (2002).

Estimation of pedigree errors in the UK dairy population using microsatellite markers and the impact on selection. *Journal of Dairy Science*, 85, 2368–2375. [https://doi.org/10.3168/jds.S0022-0302\(02\)74317-8](https://doi.org/10.3168/jds.S0022-0302(02)74317-8)

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Ablondi, M., Stocco, G., Cortellari, M., Carta, A., Summer, A., Negro, A., Grande, S., Crepaldi, P., Cipolat-Gotet, C., & Biffani, S. (2023). Microsatellite imputation using SNP data for parentage verification in four Italian sheep breeds. *Journal of Animal Breeding and Genetics*, 00, 1–13. <https://doi.org/10.1111/jbg.12839>