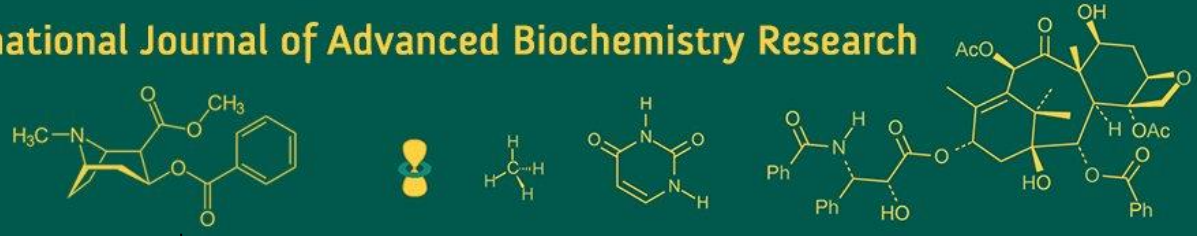
International Journal of Advanced Biochemistry Research



**ISSN Print:** 2617-4693
**ISSN Online:** 2617-4707
IJABR 2024; 8(1): 168-174
www.biochemjournal.com
Received: 15-10-2023
Accepted: 19-11-2023

**Priyanka Swami**
M.V.Sc, Department of Animal Genetics and Breeding, College of Veterinary Science & Animal Husbandry, A.N.D.U.A.T., Kumarganj, Ayodhya, Uttar Pradesh, India

**Jaswant Singh**
Professor, Department of Animal Genetics and Breeding, College of Veterinary Science & Animal Husbandry, A.N.D.U.A.T., Kumarganj, Ayodhya, Uttar Pradesh, India

**Pushkar Sharma**
Assistant Professor, Department of Veterinary Gynaecology and Obstetrics, College of Veterinary Science & Animal Husbandry, N.D.V.S.U., Jabalpur, Madhya Pradesh, India

**Sunil Kumar Meena**
Ph.D. Scholar, Department of Animal Genetics and Breeding, College of Veterinary Science & Animal Husbandry, R.A.J.U.V.A., Udaipur, Rajasthan India

**Corresponding Author:**
**Priyanka Swami**
M.V.Sc, Department of Animal Genetics and Breeding, College of Veterinary Science & Animal Husbandry, A.N.D.U.A.T., Kumarganj, Ayodhya, Uttar Pradesh, India

# Integrative analysis of common SNPS and imputation accuracy across diverse genomic densities in Asian, exotic, and Indian sheep breeds

**Priyanka Swami, Jaswant Singh, Pushkar Sharma and Sunil Kumar Meena**

**DOI:** https://doi.org/10.33545/26174693.2024.v8.i1c.330

**Abstract**
The present study aimed to explore pure SNP densities within Indian, Asian, and exotic sheep breeds using Ovine 50K SNP BeadChip data. This study explored Common Single Nucleotide Polymorphisms (SNPs) in datasets A, B, C, and D, revealing genetic variations among Asian, exotic, and Indian sheep breeds. Venn diagrams identified 13 (0.32%), 201 (1.67%), and 30 (0.15%) common SNPs across all datasets at different genomic densities (1K, 3K, 5K, 10K, and 20K). Using the Frequent item Feature Selection (FIFS) method, unique SNP patterns emphasized genetic differentiation among datasets. Imputation accuracy varied across densities, with dataset A showing the highest average accuracy at 5K (0.6124). Challenges in SNP selection for 10K and 20K densities in datasets A, B, and D indicated difficulties in capturing common SNPs. The study's insights into discriminant SNP loci on specific chromosomes, like ovine chromosomes 17, 6, 4, 15, 19, 12, and 8, offer potential for cost-effective SNP assays in sheep breed assignment. These findings aim to aid in developing low-cost genotyping methods, reducing genotyping expenses in diverse sheep breeds.

**Keywords:** SNP, genotyping, imputation, pruning, frequencies

## Introduction
In the realm of animal genetics and breeding, the application of high-throughput genotyping technologies, such as SNP arrays, has significantly progressed our comprehension of genetic diversity, evolutionary connections, and breeding strategies across diverse livestock populations (Fan, 2010) [5]. Ensuring the reliability, accuracy, and integrity of genotyping data is crucial in genetic analysis and research, involving quality pruning and the exclusion of outlier individuals (Weale, 2010) [17]. Quality pruning necessitates the implementation of rigorous criteria to filter out low-quality or unreliable genotyping data, including the removal of SNPs with high missing rates, those deviating from Hardy-Weinberg equilibrium, and those with low minor allele frequencies (Pavan, 2020) [12]. Similarly, removing outlier individuals eliminates data points potentially affected by experimental or biological factors, ensuring the remaining dataset accurately represents the true genetic composition of the studied population (Motulsky, 2006) [11]. By discarding poor-quality SNPs and outliers, researchers minimize noise and biases that could distort results (Guo, 2013) [8]. By applying quality control measures, researchers can more effectively identify true genetic signals and assess population-specific traits (Fuentes-Pardo & Ruzzante, 2017) [7]. Venn diagrams are employed to depict unique and shared SNPs among various sheep breeds or populations, aiding researchers in understanding genetic diversity and relationships between breeds. For instance, a Venn diagram can illustrate SNPs exclusive to one breed and those shared between two or more breeds (Crispim, 2019) [3].
Genotype imputation is a crucial aspect of genome-wide association studies, allowing precise evaluation of association evidence at ungenotyped markers and consolidating diverse genotyping platform results. This technique utilizes shared haplotypes to accurately estimate effects of ungenotyped variants, relying on identity by descent (IBD) to identify chromosome segments without recombination since a common ancestor (Rabiner, 1989) [13]. The imputation process identifies untyped SNPs with strong association signals, influencing follow-up strategies and aiding in reconstructing missing genotypes in pedigreed data

(Ellinghaus *et al*., 2009) [4]. Genotype imputation relies on haplotype references, often sourced from the International HapMap Project (The International HapMap Consortium, 2003) [15]; (The International HapMap Consortium, 2005) [14]. Various software tools, such as BEAGLE, MINIMAC, MaCH, and FIMPUTE, employ different principles for imputing genotyping data from missing or untyped entries, typically requiring a phasing step.

Operating on the Java platform, BEAGLE is platform-independent, with a discrete input genotype format and posterior probability as the output genotype. The Allelic R square serves as the quality measure for BEAGLE's imputation process. Notably, BEAGLE is equipped to handle multi-allelic markers (Browning and Browning, 2007) [2]. MINIMAC is an efficient implementation of the MaCH algorithm for genotype imputation, designed for low memory usage and computational efficiency. The current version of MINIMAC is available in two forms: minimac and minimac-omp. The latter utilizes the OpenMP protocol for multi-threading, resulting in enhanced throughput (Howie *et al*., 2012) [9]; Fuchsberger *et al*., 2014) [6]. Progress in animal genetics, driven by advanced genotyping techniques such as SNP arrays, has greatly enhanced our comprehension of genetic diversity, evolution, and the breeding of livestock. Low-density SNP chips provide an economical way to genotype extensive populations concurrently. Although Venn diagrams serve as a valuable instrument for exploring data, they usually present fixed perspectives of two datasets. These diagrams depict the percentages of common SNPs, highlighting the distinctions between unique and shared SNPs among datasets.

## Materials and Methods

The genotyping data were sourced from publicly available databases, consortia, and/or datasets associated with existing scientific literature. The data specifically pertained to the Ovine50KSNP BeadChip density and were generated for Indian sheep breeds (Changthangi, Tibetan, Deccani, Garole), Asian sheep breeds (Bangladeshi Garole, Bangladeshi East), and Exotic sheep breeds (Rambouillet and Australian Merino) and made datasets A (Indian sheep breeds), B (Indian and Asian sheep breeds), C (Indian and Exotic) and D (Indian, Asian and Exotic sheep breeds) (DataSheet:Agrigenomics;https://www.illumina.com/documents/products/datasheets/datasheet_ovinesnp50.pdf).

FIFS approache and datasets were employed to detect shared Single Nucleotide Polymorphisms (SNPs) among diverse sheep breeds after quality pruning through plink. Genomic datasets A, B, C, and D were subjected to the FIFS method for SNP detection and the identified SNPs were subsequently compared using Venn diagrams generated through web-based tools. Venn diagrams were employed to visualize common SNPs across the diverse datasets. Subsequently, imputation accuracy was assessed by varying densities (1K, 3K, 5K) using the FIFS method in the TRES program. These diagrams aided in distinguishing SNPs that were specific to particular sheep breeds and those that were shared among them. The imputation accuracy of SNP panels with various densities was evaluated using two BEAGLE approaches (Ellinghaus *et al*., 2009) [4]. The original datasets were randomly divided into test and reference datasets at different ratios. Initially, the datasets were separated on a chromosome-wise basis using the PLINK program. Prior to imputation, genotypic data were phased chromosome-wise. The accuracy of imputation was determined by the program's ability to predict accurate genotypes, assessed through the DR2 method (Browning *et al*., 2018). The resulting estimates from the two programs were recorded and evaluated across different low-density panels, with imputation accuracy assessed on a chromosome-wise basis. Additionally, the accuracy of both direct and step-wise imputation methods was also examined.

## Result and Discussion

Common Single Nucleotide Polymorphisms (SNPs) were identified across datasets A, B, C, and D, showcasing distinctions in the genetic makeup of Asian, exotic, and Indian sheep breeds. A Venn diagram was employed to visually represent the shared SNPs among all four datasets. The Venn diagram analysis revealed 13 (0.32%), 201 (1.67%), and 30 (0.15%) common SNPs across all four datasets for each density (1K, 3K, 5K, 10K, and 20K). Further, the study highlighted distinctive patterns of SNP commonality among datasets A, B, C, and D, emphasizing the genetic differentiation among Asian, exotic, and Indian sheep breeds. Utilizing the Frequent item Feature Selection (FIFS) method analysis depicted in Figure (1), 13 (0.32%), 201 (1.67%), and 30 (0.15%) SNPs were identified as common across all four datasets at genomic densities of 1K, 3K, 5K, 10K, and 20K, respectively. Moreover, 987 (24.67%) SNPs were common in datasets A, B, and C, with the additional insight that 987 (24.67%) SNPs were exclusively common in dataset D at densities below 1K. Similarly, 2799 (23.32%) SNPs were shared among datasets A, B, and C, while these SNPs were exclusive to dataset D at densities below 3K. Notably, 415 SNPs were found common in datasets A, B, and C, with an additional 470 SNPs common between datasets A, B, and D at the 5K density

Imputation accuracy analysis demonstrated varying results across densities for each dataset. The imputation accuracy of the FIFS method for dataset A varied between 0.1300 and 0.4293, 0.4917 to 0.5902, and 0.5614 to 0.6564 for densities of 1000, 3000, and 5000, respectively. The average imputation accuracy was highest for the 5K density at 0.6124. For 1K and 3K densities, the average imputation accuracy was 0.3182 and 0.5476, respectively. In the 1K density, chromosome number 17 had the highest imputation accuracy, while chromosome number 26 had the lowest. In the 3K density, chromosome number 6 had the highest imputation accuracy, while chromosome number 21 had the lowest. In the 5K density, chromosome number 4 had the highest imputation accuracy, while chromosome number 26 had the lowest. The selection of SNPs for 10000 and 20000 densities in dataset A using FIFS failed, possibly due to the inability to choose common SNPs frequent in all breeds. Dataset A exhibited the highest average imputation accuracy at the 5K density (0.6124), with notable variations in imputation accuracy across chromosomes. Similar trends were observed in datasets B, C, and D, where the 5K density consistently yielded the highest average imputation accuracy. Notably, the failure of SNP selection for 10K and 20K densities in datasets A, B, and D using the FIFS method pointed to potential challenges in capturing common

SNPs across diverse populations. The imputation accuracy for various densities of dataset A using the Frequent item Feature Selection (FIFS) method in the TRES program is presented in Table 1, with the corresponding graphical representation depicted in Figure (2).

For dataset B, the imputation accuracy using the FIFS method ranged from 0.1067 to 0.3835, 0.4102 to 0.5699, and 0.4958 to 0.6346 for densities of 1000, 3000, and 5000, respectively. The highest average imputation accuracy was observed for the 5K density, reaching a value of 0.5855. For 1K and 3K densities, the average imputation accuracy was 0.2690 and 0.5120, respectively. The median imputation accuracy peaked at the 5K dataset with an estimate of 0.5859, while it was 0.2826 and 0.5123 for 1K and 3K densities. Notably, chromosome 17 exhibited the highest imputation accuracy within the 1K density, while chromosome 6 and 4 showed the highest accuracy in the 3K and 5K densities, respectively. However, the selection of SNPs for 10,000 and 20,000 densities in dataset B using the FIFS method failed, likely due to the inability to choose common SNPs prevalent in all breeds. Table No.2 illustrates the imputation accuracy for different densities of dataset B while employing the FIFS method in the TRES program, with the graphical representation presented in Figure (3).

Similarly, for dataset C, the imputation accuracy using the FIFS method ranged from 0.1051 to 0.4416, 0.4231 to 0.6645, and 0.4231 to 0.6645 for densities of 1000, 3000, and 5000, respectively. The highest average imputation accuracy was observed for the 5K density, with a value of 0.6126. The average imputation accuracy for 1K and 3K densities was 0.2842 and 0.6132, respectively. The median imputation accuracy peaked at the 5K dataset with an estimate of 0.6173, while it was 0.2936 and 0.6140 for 1K and 3K densities. Chromosome 17 exhibited the highest imputation accuracy within the 1K density, while chromosome 4 showed the highest accuracy in the 3K and 5K densities. Similar to dataset B, the selection of SNPs for 10,000 and 20,000 densities in dataset C using the FIFS method failed. Table No. 3 presents the imputation accuracy for different densities of dataset C while employing the FIFS method in the TRES program, with the graphical representation depicted in Figure (4).

For dataset D, the imputation accuracy using the FIFS method varied from 0.1300 to 0.1067 to 0.3835, 0.4102 to 0.5699, and 0.4958 to 0.6346 for densities of 1000, 3000, and 5000, respectively. The highest average imputation accuracy was observed for the 5K density, reaching a value of 0.5855. The average imputation accuracy for 1K and 3K densities was 0.2690 and 0.5120, respectively. The median imputation accuracy peaked at the 5K dataset with an estimate of 0.5859, while it was 0.2826 and 0.5123 for 1K and 3K densities. Chromosome 17 exhibited the highest imputation accuracy within the 1K density, while chromosome 6 showed the highest accuracy in the 3K density, and chromosome 4 showed the highest accuracy in the 5K density. Similar to datasets B and C, the selection of SNPs for 10,000 and 20,000 densities in dataset D using the FIFS method failed. Table No.4 illustrates the imputation accuracy for different densities of dataset D while employing the FIFS method in the TRES program, with the graphical representation presented in Figure (5).

Across all datasets, the highest average imputation accuracy was consistently observed for the 5K density using the FIFS method. Notably, the discriminant SNP loci on ovine chromosomes 17, 6, 4, 15, 19, 12, and 8 exhibited superior imputation accuracy for various marker panels, providing valuable insights for cost-effective SNP assays in sheep breed assignment. The study's findings are expected to contribute to the development of low-cost genotyping methods for accurately assigning unknown animals to their true population of origin in diverse sheep breeds, with the potential to commercially reduce genotyping costs through the targeted genotyping of discriminant SNP loci. Vergara *et al.* (2014) [16] emphasise and explain the issues that call for feature selection techniques and reported feature selection methods in order to provide a state-of-the-art of feature selection methods with an implementation of mutual information feature selection framework.

Kavakiotis *et al.* (2017) [10] used a dataset of 446 individuals divided into 14 sub-populations, genotyped at 59,436 SNPs in pig breed types found in the United Kingdom, and concluded that FIFS can surpass the assignment accuracy threshold of 95% while using half the number of SNPs, gives better results than other approaches, and can aid biologists in selecting the most informative markers with maximum discrimination power for cost-effect optimization.

**Table 1:** Imputation accuracy of dataset A while employing FIFS method across different densities

| S. No. | FIFS_A_1k | FIFS_A_3k | FIFS_A_5k |
|---|---|---|---|
| OAR-1 | 0.3303 | 0.5565 | 0.6266 |
| OAR-2 | 0.3462 | 0.5769 | 0.6233 |
| OAR-3 | 0.3760 | 0.5642 | 0.6364 |
| OAR-4 | 0.3322 | 0.5746 | 0.6564 |
| OAR-5 | 0.3469 | 0.5421 | 0.5969 |
| OAR-6 | 0.3446 | 0.5902 | 0.6488 |
| OAR-7 | 0.3025 | 0.5247 | 0.5994 |
| OAR-8 | 0.3796 | 0.5437 | 0.6159 |
| OAR-9 | 0.3678 | 0.5371 | 0.6190 |
| OAR-10 | 0.3375 | 0.5490 | 0.6088 |
| OAR-11 | 0.3724 | 0.5875 | 0.6430 |
| OAR-12 | 0.3551 | 0.5682 | 0.6322 |
| OAR-13 | 0.3012 | 0.5734 | 0.6447 |
| OAR-14 | 0.4205 | 0.5450 | 0.5836 |
| OAR-15 | 0.2935 | 0.5384 | 0.5927 |
| OAR-16 | 0.3006 | 0.5621 | 0.6100 |
| OAR-17 | 0.4293 | 0.5783 | 0.6397 |
| OAR-18 | 0.2606 | 0.5325 | 0.6012 |
| OAR-19 | 0.3071 | 0.5527 | 0.6138 |
| OAR-20 | 0.2768 | 0.5433 | 0.6079 |
| OAR-21 | 0.2102 | 0.4917 | 0.5723 |
| OAR-22 | 0.2203 | 0.5645 | 0.6092 |
| OAR-23 | 0.3102 | 0.5016 | 0.6024 |
| OAR-24 | 0.3297 | 0.5101 | 0.5678 |
| OAR-25 | 0.2938 | 0.5109 | 0.6091 |
| OAR-26 | 0.1301 | 0.5207 | 0.5615 |

**Table 2:** Imputation accuracy of dataset B while employing FIFS method across different densities

| S. No. | FIFS_B_1k | FIFS_B_3k | FIFS_B_5k |
|---|---|---|---|
| OAR-1 | 0.2651 | 0.5339 | 0.6119 |
| OAR-2 | 0.2961 | 0.5527 | 0.6170 |
| OAR-3 | 0.3481 | 0.5405 | 0.6300 |
| OAR-4 | 0.2861 | 0.5394 | 0.6347 |
| OAR-5 | 0.2942 | 0.5122 | 0.5896 |
| OAR-6 | 0.3399 | 0.5699 | 0.6202 |
| OAR-7 | 0.2413 | 0.5117 | 0.5820 |
| OAR-8 | 0.3577 | 0.5185 | 0.5848 |
| OAR-9 | 0.2367 | 0.5060 | 0.5847 |
| OAR-10 | 0.3047 | 0.5110 | 0.5925 |
| OAR-11 | 0.2826 | 0.5349 | 0.5891 |
| OAR-12 | 0.2826 | 0.4981 | 0.5878 |
| OAR-13 | 0.2883 | 0.5175 | 0.6120 |
| OAR-14 | 0.3729 | 0.5039 | 0.5788 |
| OAR-15 | 0.2426 | 0.5029 | 0.5608 |
| OAR-16 | 0.2563 | 0.5208 | 0.5802 |
| OAR-17 | 0.3835 | 0.5427 | 0.6186 |
| OAR-18 | 0.2115 | 0.5076 | 0.5840 |
| OAR-19 | 0.2065 | 0.5159 | 0.5997 |
| OAR-20 | 0.2597 | 0.5125 | 0.5765 |
| OAR-21 | 0.1335 | 0.4102 | 0.5307 |
| OAR-22 | 0.1678 | 0.5327 | 0.5840 |
| OAR-23 | 0.3485 | 0.4966 | 0.5871 |
| OAR-24 | 0.3097 | 0.4651 | 0.5143 |
| OAR-25 | 0.1731 | 0.4768 | 0.5777 |
| OAR-26 | 0.1067 | 0.4792 | 0.4959 |

**Table 3:** Imputation accuracy of dataset C while employing FIFS method across different densities

| S. No. | FIFS_C_1k | FIFS_C_3k | FIFS_C_5k |
|---|---|---|---|
| OAR-1 | 0.3275 | 0.4231 | 0.4231 |
| OAR-2 | 0.2994 | 0.6551 | 0.6551 |
| OAR-3 | 0.3337 | 0.6606 | 0.6606 |
| OAR-4 | 0.3427 | 0.6646 | 0.6646 |
| OAR-5 | 0.2923 | 0.6112 | 0.6112 |
| OAR-6 | 0.3770 | 0.6510 | 0.6472 |
| OAR-7 | 0.2697 | 0.6402 | 0.6434 |
| OAR-8 | 0.3832 | 0.6113 | 0.6098 |
| OAR-9 | 0.3057 | 0.6264 | 0.6285 |
| OAR-10 | 0.2551 | 0.6358 | 0.6330 |
| OAR-11 | 0.2760 | 0.5958 | 0.5979 |
| OAR-12 | 0.2925 | 0.6187 | 0.6182 |
| OAR-13 | 0.2965 | 0.6309 | 0.6305 |
| OAR-14 | 0.3218 | 0.6064 | 0.6088 |
| OAR-15 | 0.2524 | 0.6600 | 0.6569 |
| OAR-16 | 0.2974 | 0.6494 | 0.6522 |
| OAR-17 | 0.4416 | 0.6155 | 0.6164 |
| OAR-18 | 0.2430 | 0.5931 | 0.5930 |
| OAR-19 | 0.2605 | 0.6285 | 0.6269 |
| OAR-20 | 0.2623 | 0.6126 | 0.6109 |
| OAR-21 | 0.1897 | 0.5985 | 0.6004 |
| OAR-22 | 0.1863 | 0.5917 | 0.5974 |
| OAR-23 | 0.3189 | 0.5920 | 0.5889 |
| OAR-24 | 0.2949 | 0.5668 | 0.5299 |
| OAR-25 | 0.1660 | 0.6021 | 0.6212 |
| OAR-26 | 0.1051 | 0.6025 | 0.6017 |

**Table 4:** Imputation accuracy of dataset D while employing FIFS method across different densities

| S. No. | FIFS_D_1k | FIFS_D_3k | FIFS_D_5k |
|---|---|---|---|
| OAR-1 | 0.2986 | 0.5332 | 0.6290 |
| OAR-2 | 0.2998 | 0.5555 | 0.6233 |
| OAR-3 | 0.3187 | 0.5218 | 0.6115 |
| OAR-4 | 0.3694 | 0.5621 | 0.6332 |
| OAR-5 | 0.2638 | 0.5353 | 0.6378 |
| OAR-6 | 0.3215 | 0.5064 | 0.5923 |
| OAR-7 | 0.2699 | 0.5145 | 0.5984 |
| OAR-8 | 0.3360 | 0.5797 | 0.6483 |
| OAR-9 | 0.2246 | 0.5069 | 0.6117 |
| OAR-10 | 0.3263 | 0.5303 | 0.6252 |
| OAR-11 | 0.2857 | 0.4611 | 0.5777 |
| OAR-12 | 0.3761 | 0.5372 | 0.6115 |
| OAR-13 | 0.2595 | 0.4341 | 0.5454 |
| OAR-14 | 0.2437 | 0.4970 | 0.5705 |
| OAR-15 | 0.3665 | 0.5338 | 0.5969 |
| OAR-16 | 0.1591 | 0.5018 | 0.6168 |
| OAR-17 | 0.2721 | 0.5446 | 0.6047 |
| OAR-18 | 0.2827 | 0.5039 | 0.6235 |
| OAR-19 | 0.3917 | 0.5308 | 0.6156 |
| OAR-20 | 0.3550 | 0.5179 | 0.5706 |
| OAR-21 | 0.2431 | 0.4576 | 0.5521 |
| OAR-22 | 0.1791 | 0.4290 | 0.5514 |
| OAR-23 | 0.2383 | 0.5023 | 0.5868 |
| OAR-24 | 0.3038 | 0.4887 | 0.5699 |
| OAR-25 | 0.4299 | 0.5469 | 0.6108 |
| OAR-26 | 0.2328 | 0.4502 | 0.5432 |

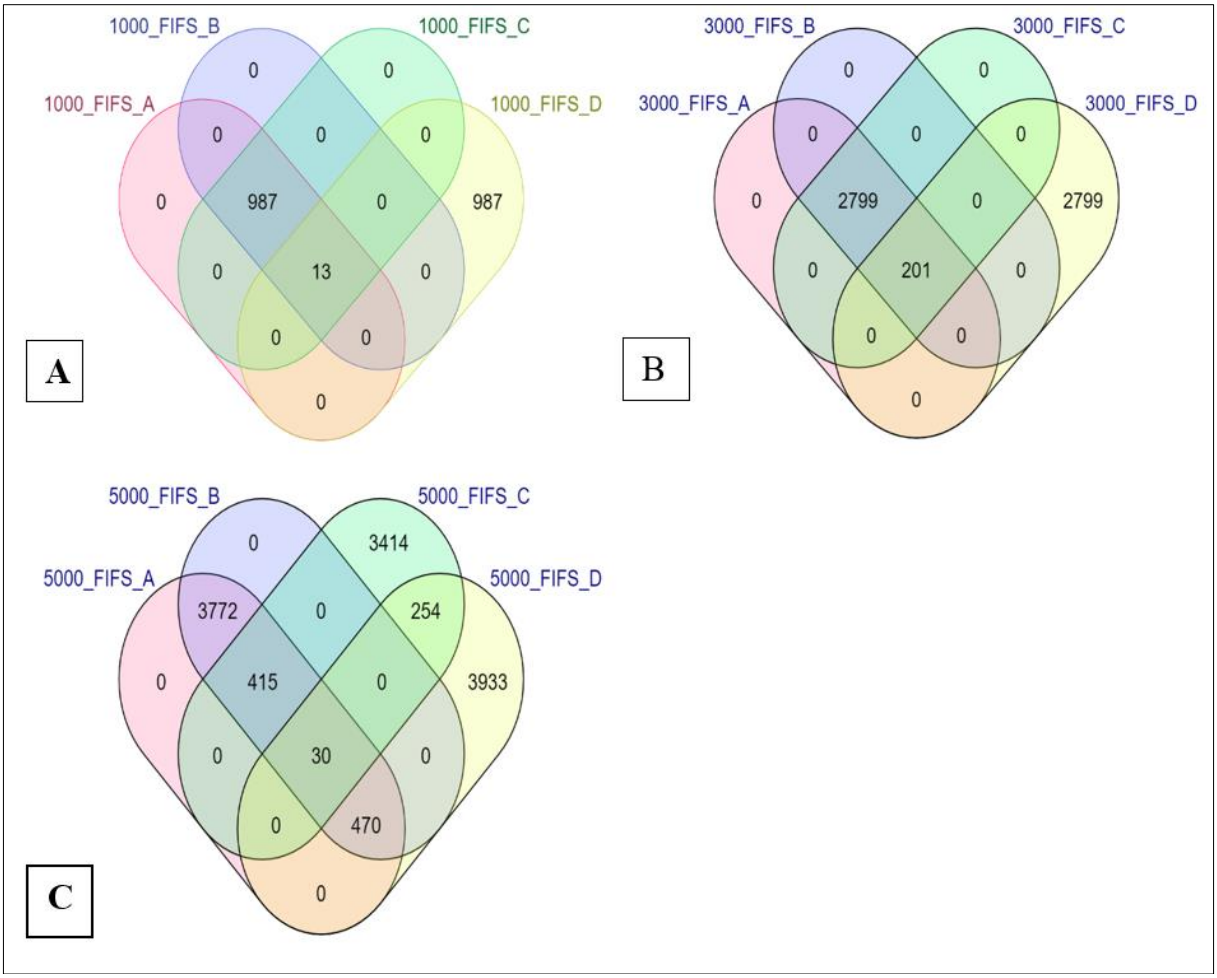



**Fig 1:** The following Venn diagram represents the overlap of genetic data at the same genomic position. Venn diagram showing the total number of putative SNPs significantly associated with FIFS method. The total number of SNPs is reported in each panel. The 13, 201 and 30 SNPs were common between datasets A, B, C and D.

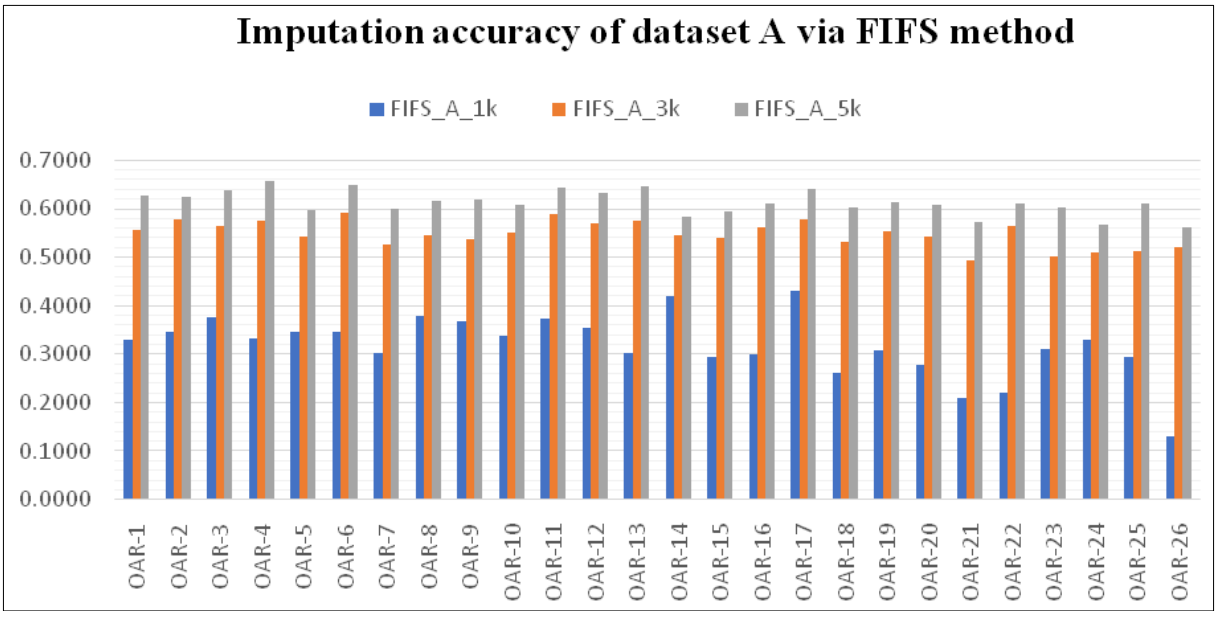



**Fig 2:** Imputation accuracy of dataset A via FIFS method

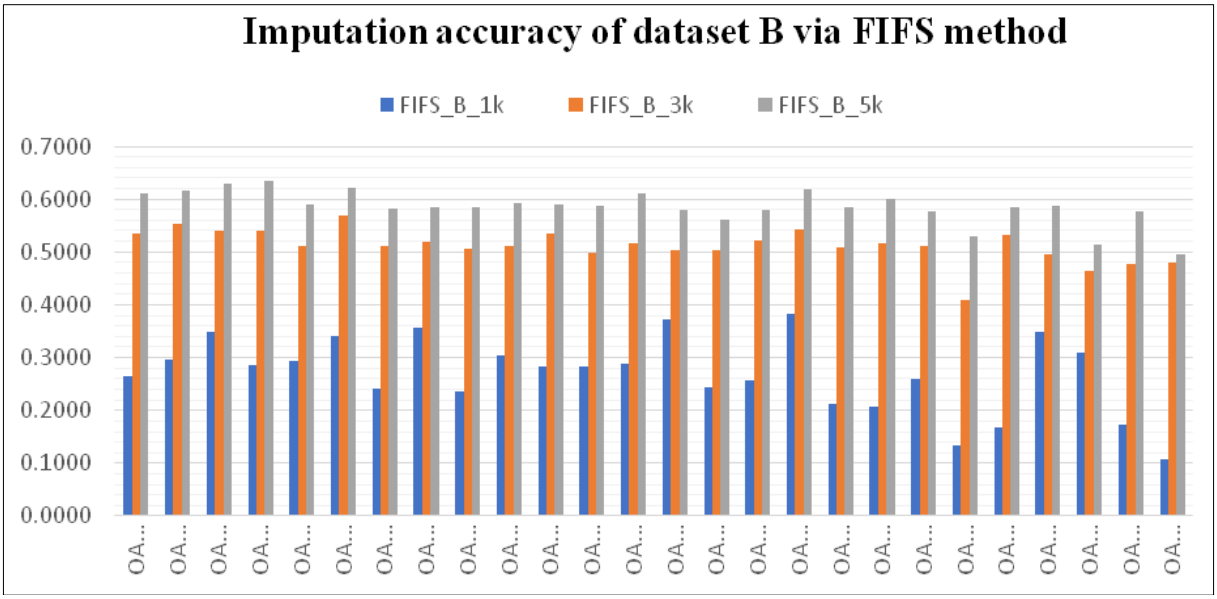



**Fig 3:** Imputation accuracy of dataset B via FIFS method

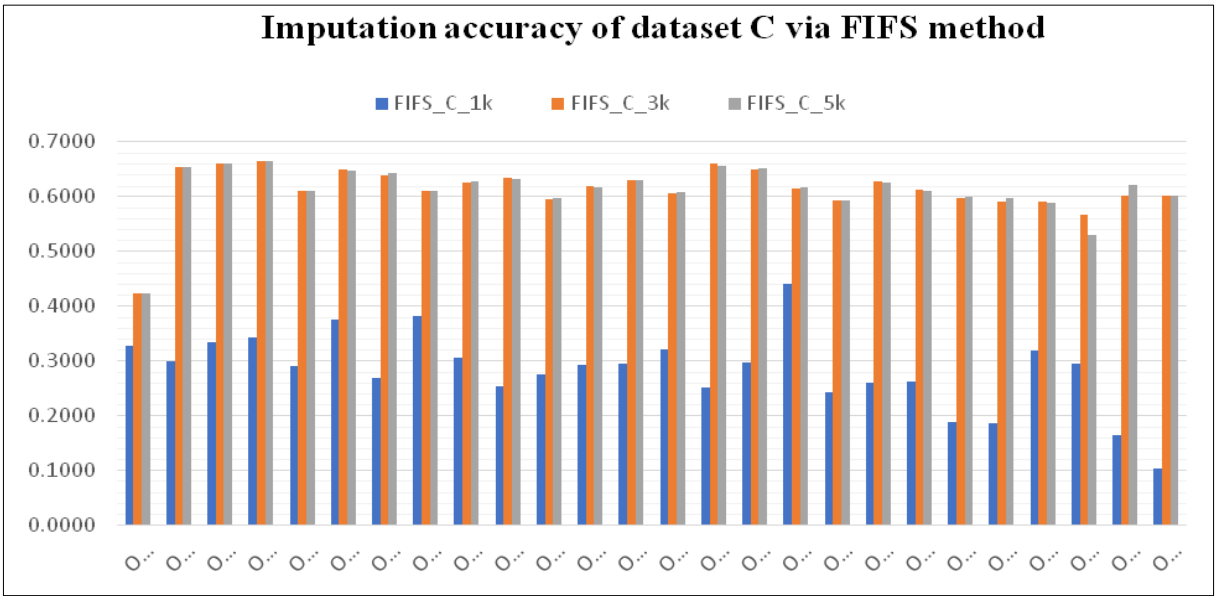



**Fig 4:** Imputation accuracy of dataset C via FIFS method

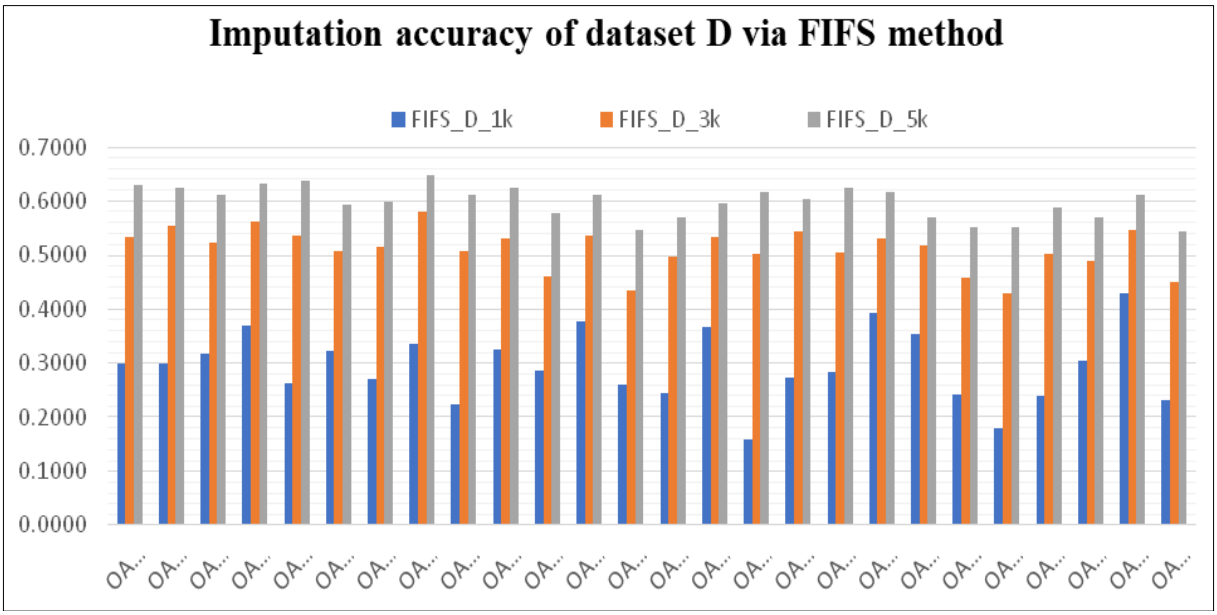



**Fig 5:** Imputation accuracy of dataset D via FIFS method

## Conclusion

This research provides a comprehensive understanding of SNP commonalities and imputation accuracy in different genomic densities for Asian, exotic, and Indian sheep breeds, and highlights the significance of the FIFS method in identifying common SNPs and underscores the challenges in imputing SNPs from low to high density. Insights into chromosomal variations affecting imputation accuracy contribute to refining genomic analyses, offering practical applications in breed-specific genomic research and breeding programs. The findings lay the foundation for the development of cost-effective SNP assays, facilitating accurate population assignments in sheep breeds. summary, this study explored Common Single Nucleotide Polymorphisms (SNPs) in datasets A, B, C, and D, revealing genetic variations among Asian, exotic, and Indian sheep breeds. Venn diagrams highlighted unique sets of common SNPs across different genomic densities, emphasizing distinctive genetic features within each dataset. The Frequent item Feature Selection (FIFS) method efficiently identified informative SNP patterns, emphasizing genetic differentiation among the populations.

Imputation accuracy varied across densities, with Dataset A (Indian breeds) showing the highest average accuracy at 5K (0.6124). Challenges in SNP selection for 10K and 20K densities suggested difficulties capturing common SNPs, a trend observed in datasets B, C, and D. Chromosome-specific analyses pinpointed discriminant SNP loci, offering opportunities for cost-effective SNP assays in sheep breed assignment. The result of our study can provide valuable information for developing a method in designing a low cost SNP assay for assigning unknown animals to their true population of origins in other sheep breeds. It is expected that the genotyping of 5K discriminant SNP loci selected from ~50000 SNPs, available on Illumina Ovine 50KSNPBeadChip, will commercially decrease the cost of the genotyping.

These findings provide valuable insights for developing low-cost genotyping methods, aiming to accurately assign unknown animals to their true population in diverse sheep breeds. The focus on discriminant SNP loci suggests a targeted approach for genotyping, potentially reducing commercial genotyping costs. The focus on discriminant SNP loci suggests a targeted approach for genotyping, potentially reducing commercial genotyping costs, research lays the groundwork for further investigations into optimizing SNP assays and advancing genetic studies in sheep breeding and population genetics.

## Acknowledgment

The author would like to express gratitude to the Dean, College of Veterinary Science and Animal Husbandry, A.N.D.U.A. & T., Kumarganj, Ayodhya (Uttar Pradesh) for providing funds and necessary support for the research.

## Conflict of interest statement

The authors declare there is no conflict of interest on this article.

## References

1. Browning BL, Zhou Y, Browning SR. A one-penny imputed genome from next-generation reference panels. Am J Hum Genet. 2018;103:338-348. DOI: 10.1016/j.ajhg.2018.07.015
2. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing data inference for whole genome association duties by use of localized haplotype. Am J Hum Genet. 2007;81:1054-1097.
3. Crispim BA. Genetic diversity in Brazilian sheep breeds. Small Ruminant Res. 2019;178:70-77.
4. Ellinghaus E, Schreiber S, Franke A, Nothnagel M. Current software for genotype imputation. Hum Genomics. 2009;3(4):371-380.
5. Fan B, Du ZQ, Gorbach DM, Rothschild MF. Development and application of high-density SNP arrays in genomic studies of domestic animals. Asian-Australas J Anim Sci. 2010;23(7):833-847.
6. Fuchsberger C, Abecasis GR, Hinds DA. minimac2: faster genotype imputation. Bioinformatics. 2014;31(5):782-4. DOI: 10.1093/bioinformatics/btu704.
7. Fuentes-Pardo AP, Ruzzante DE. Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. Mol Ecol. 2017;26(20):5369-5406.
8. Guo Y, Ye F, Sheng Q, Clark T, Samuels DC. Three-stage quality control strategies for DNA re-sequencing data. Brief Bioinform. 2013;15(6):879-889.
9. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet. 2012;44(8):955-9.
10. Kavakiotis I, Samaras P, Triantafyllidis A, Vlahavas I. FIFS: A data mining method for informative marker selection in high dimensional population genomic data. Control Biol Med. 2017;90:146-154.
11. Motulsky HJ, Brown RE. Detecting outliers when fitting data with nonlinear regression – a new method based on robust nonlinear regression and the false discovery rate. BMC Bioinformatics. 2006;7(1):1-20.
12. Pavan S, Delvento C, Ricciardi L, Lotti C, Ciani E, D'Agostino N. Recommendations for Choosing the Genotyping Method and Best Practices for Quality Control in Crop Genome-Wide Association Studies. Front Genet. 2020;11:1-13.
13. Rabiner LR. A tutorial on hidden Markov-models and selected applications in speech recognition. Proc IEEE. 1989;77:257-286.
14. The International HapMap Consortium. A haplotype map of the human genome. Nature. 2005;437:1299-1320.
15. The International HapMap Consortium. The International HapMap Project. Nature. 2003;426:789-796.
16. Vergara JR, Estévez PA. A review of feature selection methods based on mutual information. Neural Comput Appl. 2014;24(1):175.
17. Weale ME. Quality Control for Genome-Wide Association Studies. In: Barnes M, Breen G (eds) Genetic Variation. Methods Mol Biol. Humana Press, Totowa, NJ; c2010. p. 628.