RESEARCH ARTICLE

Enhancing Genetic Insights in Animal Breeding: Optimizing Low-Density SNP Panels for Distinguishing Common SNPs in Diverse Sheep Breeds

Priyanka Swami[1], Jaswant Singh[1]*, Pushkar Sharma[2], Sunil Kumar Meena[3]

Abstract

Advancements in animal genetics, propelled by high-throughput genotyping methods like SNP arrays, have significantly expanded our understanding of genetic diversity, evolution, and livestock breeding. Low-density SNP chips offer a cost-effective means of genotyping large populations simultaneously. While Venn diagrams are a valuable tool for data exploration, they typically provide static views of up to datasets. Venn diagrams illustrated the proportions of common SNPs, distinguishing unique and shared SNPs across datasets. In this study, we aimed to develop low-density SNP panels of varying densities using Ovine 50K SNP Bead Chip data from Indian, Asian, and exotic sheep breeds, (a) Select unique and breed specific common SNP via Quality pruning and (b) select 20k panel Using Venn diagram through four methods. Genotyping data were sourced from publicly available databases, consortiums, and datasets referenced in scientific literature. To facilitate our analysis, we merged three sets of sheep breeds into four combinations using appropriate merger commands within the PLINK software. These four datasets underwent quality pruning based on various parameters and thresholds. We generated informative SNP panels for each dataset using the TRES approach, employing the delta, FST, info and combine method to rank markers that distinguish between the underlying breeds. Our findings, obtained through the all four methods, indicated that the 20K SNP panel outperformed the 50K panel in distinguishing common SNPs between Asian, Indian, and exotic sheep breeds. The incorporation of these practices elevates the validity and applicability of genetic insights, fostering informed decision-making and propelling advancements in animal genetics and breeding.

Key words: Genetic diversity, Quality pruning, Ovine 50K SNP BeadChip, SNP arrays, Venn diagram

Ind J Vet Sci and Biotech (2024): 10.48165/ijvsbt.20.3.16

Introduction

In the field of animal genetics and breeding, the utilization of high-throughput genotyping technologies, such as SNP arrays, has significantly advanced our understanding of genetic diversity, evolutionary relationships, and breeding strategies within different livestock populations (Fan *et al.,* 2010).

Quality pruning of genotyping data and the removal of outlier individuals are critical steps in genetic analysis and research (Weale, 2010). These processes ensure the reliability, accuracy, and integrity of the data, which in turn have a profound impact on the validity and meaningfulness of subsequent analyses. Quality pruning involves applying stringent criteria to filter out low-quality or unreliable genotyping data and this includes removing SNPs with high missing rates, those deviating from Hardy-Weinberg equilibrium, and those with low minor allele frequencies (Pavan *et al.,* 2020). Similarly, the removal of outlier individuals eliminates data points that may be erroneous due to experimental or biological factors (Motulsky and Brown, 2006). By discarding such data, researchers ensure that the remaining dataset accurately reflects the true genetic makeup of the population under study (Gress *et al.,* 2018). Pruning poor-quality SNPs and outliers reduces the noise and biases that can distort results (Guo *et al.,* 2013). This, in turn, enhances the power of statistical tests, increases the precision of estimates, and reduces the risk of false positives or false negatives. Quality pruning helps mitigate biases introduced by population structure and admixture, by applying quality control measures, researchers can better identify true

[1]Department of Animal Genetics and Breeding, College of Veterinary Science & Animal Husbandry, ANDUAT, Kumarganj, Ayodhya-224 229, Uttar Pradesh, India

[2]Department of Veterinary Gynaecology and Obstetrics, College of Veterinary Science & Animal Husbandry, NDVSU, Mhow-453 446, Madhya Pradesh, India

[3]Department of Animal Genetics and Breeding, College of Veterinary and Animal Science, Navania, Udaipur RAJUVAS -313 601, Rajasthan, India

Corresponding Author: Jaswant Singh, Department of Animal Genetics and Breeding, College of Veterinary Science & Animal Husbandry, ANDUAT, Kumarganj, Ayodhya-224 229, Uttar Pradesh, India. e-mail: dr.jaswant75@gmail.com

How to cite: Swami, P., Singh, J., Sharma, P., & Meena, S. K. (2024). Enhancing genetic insights in animal breeding: Optimizing low-density SNP panels for distinguishing common SNPs in diverse sheep breeds. *Ind J Vet Sci and Biotech*, 20(3), 82-86.

Source of support: Nil

Conflict of interest: None

Submitted 18/01/2024 **Accepted** 21/02/2024 **Published** 10/05/2024

genetic signals and accurately assess population-specific traits (Fuentes-Pardo and Ruzzante, 2017). Venn diagrams are used to illustrate the unique and shared SNPs among different sheep breeds or populations. This helps researchers understand the genetic diversity and relatedness between breeds. For example, a Venn diagram can show which SNPs are exclusive to one breed and which are common between two or more breeds (Crispim, 2019).

This study aimed to investigate the pure SNP densities within Indian, Asian, and exotic sheep breeds using the Ovine 50K SNP BeadChip data.

Materials and Methods

The genotyping data utilized in this study were sourced from publicly available databases, consortiums, and datasets present in the scientific literature (DataSheet:Agrigenomics;https://www.illumina.com/documents/products/datasheets/datasheet_ovinesnp50.pdf). The Ovine 50K SNP BeadChip, which encompasses a total of 54,241 markers, was employed to generate array data for the different sheep breeds. Four distinct datasets were created based on breed composition:

1. Dataset A: Indian sheep breeds only
2. Dataset B: Indian and Asian sheep breeds
3. Dataset C: Indian and exotic (non-Asian) sheep breeds
4. Dataset D: Indian, Asian, and exotic sheep breeds

Quality pruning of the genotyping data was performed to ensure the reliability and accuracy of downstream analyses. The following thresholds were applied for quality control using the plink software (Whole genome association analysis toolset, 2023): Autosomal coordinates and missing rate per person (--mind 0.1); Minor allele frequency (--maf< 0.05); Missing rate per SNP (--geno 0.1); Hardy-Weinberg equilibrium (--hwe 0.001); Mendel error rate (--me 0.05 or 0.1)

The genotyping data underwent processing with TRES software to create rankings of markers that distinguish between the various breeds in separate datasets. For each dataset, lists containing the top 1000, 3000, 5000, 10000, and 20000 SNP markers were generated according to the criteria mentioned above.

Multiple techniques (including combine, info, fst, and delta) and datasets were utilized to identify common Single Nucleotide Polymorphisms (SNPs) across diverse sheep breeds. These selected SNPs were then compared using Venn diagrams, which were created using web-based tools. These Venn diagrams facilitated the recognition of SNPs that were either unique to specific sheep breeds or shared among them.

Results and Discussion

In this study, the genotypic data of 8 sheep breeds were taken from the public repository and classified into four different groups including dataset A includes Indian sheep breeds, dataset B includes Indian and Asian sheep breeds, dataset C includes Indian and exotic (other than Asian) sheep breeds and dataset D includes Indian, Asian and exotic sheep breeds, respectively. Quality pruning was done on each dataset based on thresholds as autosomal coordinates and missing rate per person (--mind 0.1), minor allele frequency (--maf <0.05), missing rate per SNP (--geno 0.1), Hardy-Weinberg equilibrium (--hwe 0.001), Mendel error rate (--me 0.05 or 0.1) with plink software (Table 1).

A total of 8,764 SNP markers were removed from dataset A on using above thresholds. --autosome (3,565 SNP markers were removed), --mind 0.1 (no removal of markers), --maf 0.05 (3,650 SNP markers were removed), --geno 0.1; (1.00 SNP markers were removed), --hwe 0.001; (1,548 SNP markers were removed), --me 0.05 0r 0.1; (no removal of markers). Genotyping rate of data set A was 0.999622 or 99.96%, A total of 41620 markers were left in Dataset A after quality pruning for downstream analysis.

A total of 11,201 SNP markers were removed from dataset B on using above thresholds. --autosome (3,565 SNP markers were remove), --mind 0.1 (no removal of markers), --maf 0.05 (3,845 SNP marker were remove), --geno 0.1 (0.00 SNP marker were removed), --hwe 0.001 (3,791 SNP marker were removed), --me 0.05 0r 0.1 (no markers removed). Genotyping rate of data set B was 99.97%. A total of 39183 markers were left in dataset B after quality pruning for downstream analysis.

Table 1: Quality pruning results and filtered SNP markers for different datasets

Dataset	Original number SNP marker	SNPs with known autosomal coordinates	Missing rate per person of 10% and below	Minor allele frequency of 0.05 and below	Missing rate per SNP of 10% and below	HWE with p value of 0.001	Mendel error rate	Filtered SNPs
Dataset A	50384	46819	46819	43169	43168	41620	41620	41620
Dataset B	50384	46819	46819	42974	42974	39183	39183	39183
Dataset C	50384	46819	46819	45330	45330	38791	38791	38791
Dataset D	50384	46819	46819	45323	45323	34640	34640	34640

SNPs filtering parameters: Unmapped, X, Y, Mt SNPs, SNP CR (< 0.001) Mt: Mitochondria, MAF: Minor Allele Frequency, HWE: Hardy Weinberg Equilibrium

A total of 11,593 SNP markers were removed from dataset C on using above thresholds. --autosome (3,565 SNP markers were removed), --mind 0.1 (no removal of markers), --maf 0.05, (1,489 SNP markers were removed), --geno 0.1 (0.00 SNP markers were removed), --hwe 0.001 (6,539 SNP markers were removed), --me 0.05 0r 0.1 (no markers were removed). Genotype rate of data set C was 99.98%. A total of 38791 markers were left in Dataset C after quality pruning for downstream analysis.

A total of 15,744 SNP markers were removed from dataset D on using above thresholds. --autosome (3,565 SNP markers were removed),--mind 0.1 (no markers were removed), --maf 0.05 (1,496 SNP markers were removed), --geno 0.1 (no markers were removed), --hwe 0.001 (10,683 SNP markers removed), --me 0.05 0r 0.1 (no markers were removed). Genotyping rate of data set D was 98.98%. A total of 34640 markers were left in dataset D after quality pruning for downstream analysis.Quality pruning of each data set was done according to Ahmad e*t al.* (2021). After quality control, a total of 41620, 39,183, 38791, and 34640 SNPs were retained for dataset A, dataset B, dataset C, and dataset D, respectively.

In this study, the removal of SNP markers based on established thresholds, such as missing rates, minor allele frequency, and Hardy-Weinberg equilibrium, led to the retention of high-quality SNP markers for each dataset. The creation of distinct datasets, encompassing different combinations of Indian, Asian, and exotic sheep breeds, allowed for the investigation of evolutionary relationships and genetic diversity within and across these populations. The findings from this study contribute to a deeper understanding of the genetic makeup of Indian sheep breeds and their relationships with both regional and exotic counterparts as also found by Ahmad *et al.* (2021). Furthermore, the identification of pure SNPs densities and genetic variations within these sheep breeds has significant implications for India's small ruminant breeding policies. These insights can inform decisions related to breed improvement, conservation efforts, and the development of targeted breeding strategies to enhance desirable traits and overall livestock productivity.

Common SNPs were identified across datasets A, B, C, and D, with variations in these SNPs distinguishing between Asian, exotic, and Indian sheep breeds. Utilizing different analytical methods, the following percentages of common SNPs were observed in all four datasets, including 1K, 3K, 5K, 10K, and 20K using Venn diagram. Venn diagram was plotted to depict the common SNPs belonging to all four datasets under the study.

The Combine method revealed 79 (1.97%), 463 (8.61%), 1034 (5.17%), 3,067 (7.66%), and 9,164 (11.45%) SNPs were common between datasets A, B, C and D at 1k, 3k, 5k, 10k, 20k panels orderly (Table 2). The Delta method indicated 66 (1.65%), 432 (3.6%), 970 (4.85%), 2,998 (7.49%), and 9,433 (11.79%) SNPs were common between datasets A, B, C and D at 1k, 3k, 5k, 10k, 20k panels orderly (Table 2). The F_{ST} method yielded 19 (0.47%), 153 (1.27%), 482 (2.41%), 1,989 (4.97%), and 7,790 (9.73%) SNPs were common between datasets A, B, C and D at 1k, 3k, 5k, 10k, 20k panels orderly (Table 2). The Information method analysis showed 73 (1.8%), 416 (3.46%), 935 (4.67%), 2,838 (7.09%), and 8,731 (10.91%) SNPs were common between datasets A, B, C and D at 1k, 3k, 5k, 10k, 20k panels orderly (Table 2). In all datasets 20 k panels have more unique and common breed specific SNP showing at figure 1, 2, 3 and 4. Venn diagram represents the overlap of genetic data at the same genomic position. Czech *et al*. (2018) also reported the number of common SNPs among various breed combinations. Specifically, 23,113 filtered SNPs were common to each breed, indicating a polymorphic characteristic in the Domino lineage but monomorphic in other breeds.

Moura *et al*. (2019) conducted a study involving two groups of goats to identify common and unique SNPs in the Moreta and Anglo Nubian goat breeds. Our work is in alignment with their research, as we also aimed to investigate common and unique SNPs within sheep breeds. Kranis *et al.* (2013) demonstrated the presence of overlapping SNPs between broiler, layer, and inbred lines, as well as between broiler, WEL, and BEL. Approximately 23% of the 10 million SNPs were identified as shared among broiler, layer, and inbred lines, while 31% of the SNPs were found to be common between broiler, BEL, and WEL. The size of the circles in Venn diagrams also visually reflects the relative proportions of SNPs contributed by different groups within the 10 million SNP dataset.

Employing quality pruning techniques, we carefully selected SNPs to mitigate bias. Subsequently, we applied the Delta, F_{ST}, Combine, and Information (Info) methods, visualizing the results through Venn diagrams, to uncover common and unique SNPs. These SNPs provide valuable insights into breed purity, population structure, and disease resistance within the studied populations.

Table 2: Venn diagram of SNPs detected by different methods in different densities of datasets

Panel of SNP	Common SNPs through Combine method	Common SNPs through Delta method	Common SNPs through F_{ST} method	Common SNPs through Info method
1000(1K)	79	66	19	73
3000(3K)	463	432	153	416
5000(5K)	1034	970	482	935
10,000(10K)	3067	2998	1989	2838
20,000(20K)	9164	9433	7790	8731

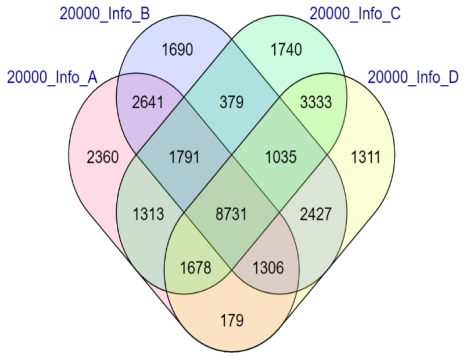

Fig. 1: Venn diagrams showing the common SNPs via info method in different datasets

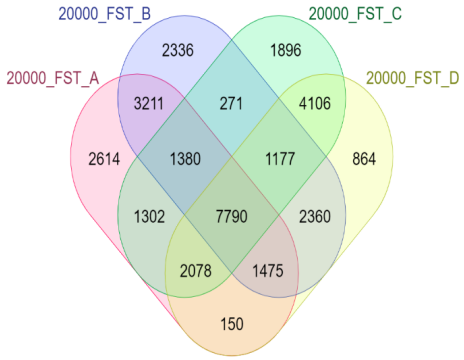

Fig. 2: Venn diagrams showing the common SNPs via F_{ST} method in different datasets

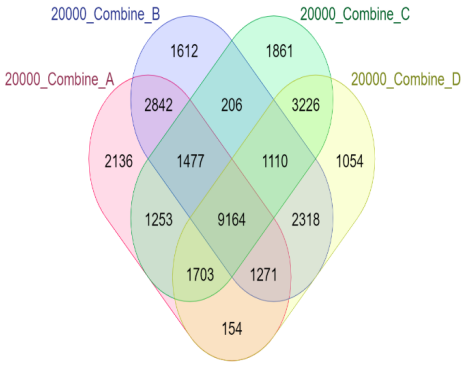

Fig. 3: Venn diagrams showing the common SNPs via Combine method in different datasets

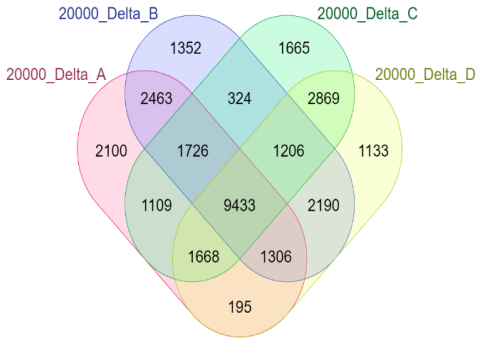

Fig. 4: Venn diagrams showing the common SNPs via Delta method in different datasets

Conclusion

This study utilized Ovine 50K SNP BeadChip data to investigate pure SNP densities within Indian, Asian, and exotic sheep breeds. The application of quality control measures resulted in the identification of high-quality SNP markers for downstream analysis. The findings shed light on the evolutionary relationships between different sheep populations and hold important implications for India's breeding policies concerning small ruminants. Using the Combine, Info, FST and Delta method and the Venn diagram, display notable and frequent SNPs in distinct panels. These frequent SNPs distinguish between Indian, exotic, and Asian sheep breeds. Higher the common SNPs, the easier it is to discern breeds. Because it selected more common SNPs from a 20K panel, it is the best panel.

In summary, this study employed rigorous quality control measures to filter and retain high-quality SNP markers in different datasets of sheep breeds. The research provided critical insights into genetic relationships and diversity within these breeds, with implications for breed improvement and conservation efforts. Additionally, the identification of common and unique SNP markers contributes to our understanding of breed purity, population structure, and disease resistance. The findings build upon previous research and enhance our knowledge of the genetic characteristics of these livestock populations.

Acknowledgment

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

References

Ahmad, S.F., Mehrotra, A., Charles, S., & Ganai, N.A. (2021). Analysis of selection signatures reveals important insights into the adaptability of high-altitude Indian sheep breed Changthangi. *Gene, 799,* 145809.

Crispim, B.A. (2019). Genetic diversity in Brazilian sheep breeds. *Small Ruminant Research, 178,* 70-77.

Czech, B., Frąszczak, M., Mielczarek, M., & Szyda, J. (2018) Identification and annotation of breed-specific single nucleotide polymorphisms in *Bos taurus* genomes. *Plos One, 13*(6), 1-9.

Fan, B., Du,Z.Q., Gorbach, D.M., & Rothschild, M.F. (2010). Development and application of high-density SNP arrays in genomic studies of domestic animals. *Asian-Australasian Journal of Animal Sciences, 23*(7), 833-847.

Fuentes-Pardo, A.P., & Ruzzante, D.E. (2017). Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. *Molecular Ecology, 26*(20), 5369-5406.

Gress, T.W., Denvir, J., & Shapiro, J.I. (2018). Effect of removing outliers on statistical inference: Implications to interpretation of experimental data in medical research. *Marshall Journal of Medicine, 4*(2), 1-19.

Guo, Y., Ye, F., Sheng, Q., Clark, T., & Samuels, D.C. (2013). Three-stage quality control strategies for DNA re-sequencing data. *Briefings in Bioinformatics, 15*(6), 879–889.

Kranis, A., Gheyas, A.A., Boschiero, C., Turner, F.,Yu, L.,Smith, S., Talbot, R., Pirani, A., Brew, F., Kais, P., Hocking, P.M., Fife, M., Salmon, N., Fulton, J., Strom, T.M., Haberer, G., Weigend, S., Preisinger, R., Gholami, M., & Burt, D.W. (2013). Development of a high density 600K SNP genotyping array for chicken. *BMC Genomics, 14*(59), 1-13.

Motulsky, H.J., & Brown, R.E. (2006). Detecting outliers when fitting data with nonlinear regression – a new method based on robust nonlinear regression and the false discovery rate. *BMC Bioinformatics, 7*(1), 1-20.

Moura, J.deO, Campelo, J.E.G., Baja, M.M., Sarmento, J.L.R., & Araújo, A.M.de. (2019). Population genomics and gene introgression in goat herds naturally adapted to Brazil. *Revista Ciencia Agronomica, 50*(3), 476-483.

Pavan, S., Delvento, C., Ricciardi, L., Lotti, C., Ciani, E., & D'Agostino, N. (2020). Recommendations for Choosing the Genotyping Method and Best Practices for Quality Control in Crop Genome-Wide Association Studies. *Frontiers in Genetics, 11,* 1-13.

Weale, M.E. (2010). Quality Control for Genome-Wide Association Studies. In: Barnes, M., Breen, G. (eds) Genetic Variation. *Methods in Molecular Biology,* vol 628. Humana Press, Totowa, NJ.

Whole genome association analysis toolset. (2023) PLINK: Whole genome data analysis toolset. (n.d.). https://zzz.bwh.harvard.edu/plink/tutorial.shtml