Acta Scientific Veterinary Sciences (ISSN: 2582-3183)

Volume 4 Issue 9 September 2022

Research Article

Identification of Microsatellite-Based Markers and Breed-Specific Single Nucleotide Polymorphism Panels for Parentage Assignment in Bovines

Kirtypal Singh[1], CS Mukhopadhyay[2]*, Simarjeet Kaur[3] and JS Arora[4]

[1]Ph.D. Scholar, College of Animal Biotechnology, Guru Angad Dev Veterinary and Animal Sciences University, Ludhiana

[2]Senior Scientist, College of Animal Biotechnology, Guru Angad Dev Veterinary and Animal Sciences University, Ludhiana

[3]Simarjeet Kaur, Head and Senior Geneticist, Department of Animal Genetics and Breeding, Guru Angad Dev Veterinary and Animal Sciences University, Ludhiana

[4]Scientist, College of Animal Biotechnology, Guru Angad Dev Veterinary and Animal Sciences University, Ludhiana

***Corresponding Author:** CS Mukhopadhyay, Senior Scientist, College of Animal Biotechnology, Guru Angad Dev Veterinary and Animal Sciences University Ludhiana.

DOI: 10.31080/ASVS.2022.04.0505

Received: August 19, 2022

Published: August 29, 2022

Abstract

Single nucleotide polymorphism (SNP) markers are nowadays used as potential markers for parentage determination in bovines. Five trios of each of the species, namely, Bubalus bubalis (Indian water buffalo) and *Bos taurus* (Holstein Friesian crosses) were selected for the study. Parentage conformity of trios (Probability of Paternity = 99.99%) with six fluorescently labelled microsatellite loci was done. Single nucleotide polymorphism-based parentage detection included identification of informative single nucleotide polymorphism (based on trio-wise allele comparison) followed by exclusion of Mendelian erroneous (MDE) SNPs for minimizing the errors. About 100 to 100 nucleotide long stretches of DNA, harbouring haplotypes (consisting of at least 5 SNPs) were identified PCR amplification followed by Sanger sequencing for validation of SNP-variations in new sets of trios. The principal component analysis (PCA) based on the Minor allele frequency (MAF) distribution was analysed. Principal component analysis have narrowed down the SNP-data to figure out the most informative SNPs. We could identify 51 and 1857 most informative SNPs for buffalo and cattle, respectively, which could explain cumulative variance of up to 95.4% and 95.43% of the components, respectively. However, the validation results were not much appreciable as some of the single nucleotide polymorphism could not be detected in those amplicons. Hence, the accuracy of parentage assignment using the SNP-based approach was quite less efficient. In conclusion, it can be stated that the microsatellite-based approach for parentage determination is well standardized and efficient in accuracy of parentage assignment has compared to the SNP-based approach for parentage assignment.

Keywords: Bovines; Microsatellite; Parentage; PCA, SNPs

Abbreviations

BLAST: Basic Local Alignment Search Tool; dd-RAD: Double Digest Restriction-Site Associated DNA; DNA: Deoxyribonucleic acid; GBS: Genotyping by Sequencing; He: Heterozygotes; Ho: Homozygotes; IAEC: Institutional Animal Ethics Committee; LOD: Log of Odds; MAF: Minor Allele Frequency; MDE: Mendelian Erroneous; MDE: Mendelian error; PCA: Principal Component Analysis; PCR: Polymerase Chain Reaction; PIC: Polymorphism Information Content; PP: Probability of Paternity; SNP: Single Nucleotide Polymorphism

Introduction

Parentage testing for bovines is important to keep a check on reproduction and estimating the genetic merits of the germ pool. Multi-sire bovine herds are often found to be impracticable due to incorrect parentage assignment, which affects the genetic gains [13]. Since many decades, parentage testing has been performed using different DNA-based techniques (such as DNA Fingerprinting, PCR) and before that, parentage testing was performed using blood groups [21]. Later microsatellite markers have dominated the labs [3], and now transitional change has happened from the recent decade, as the single nucleotide polymorphism (SNP) based approach is playing a crucial role in parentage testing [5]. Adoption of new technology depends upon the initial cost and the availability of technological support and their trade off with the efficacy of new technique to reduce parentage identification errors [1,23]. SNP-based marker system usually relies on the polymorphism, which varies between the population and species [9]. SNP-based markers are quite attractive because they are abundant in eukaryotic genome [4], genetically stable in mammals [12,14,22] and they are amenable to high-throughput automated analysis [11,24]. Many high-throughput genotyping technologies have been developed [10] to add to the score of a growing number of published SNPs [15]. Many of these techniques are quite suitable for their use in bovine SNPs identification.

Genotyping by sequencing (GBS) is a novel approach of next-generation sequencing protocols for discovering and genotyping the SNPs in populations. GBS approach is based on the digestion of the genomic DNA with restriction enzymes (rare cutter and frequent cutter), followed by the ligation of the barcode adapter, and finally, streamlining of sequenced data using a bioinformatics pipeline [6]. While developing the efficient SNP-based marker systems for the parentage identification in bovines, it is critical to consider that SNP information may vary significantly between the populations and species [9]. For developing a bovine panel of SNPs for animal identification and paternity analysis, one must remember that in the description of a minimal set of SNPs should have sufficient power to uniquely identify animals and their parents in a variety of popular breeds and crossbred populations. In this study, we have selected 6 sets of microsatellite loci, each for cattle and buffalo parentage determination with high accuracy and also have identified the panel of highly informative SNPs for animal identification and breed-specific parentage assignment.

Methodology

Animal selection

The experiment was conducted at the College of Animal Biotechnology of Guru Angad Dev Veterinary and Animal Sciences University, Punjab, India. In this study, five trios (two parents and one offspring) of each species (HF crossbred cattle and Murrah buffalo) were selected from the dairy farm of the Department of Animal Genetics and Breeding, Guru Angad Dev Veterinary and Animal Sciences University, Ludhiana and also from farmer's herds. The pedigree of each trio (collected from University farm) was recorded for at least four generations to ensure that the trios were genetically unrelated. Experimental procedures conducted in this study were approved by the Institutional Animal Ethics Committee (IAEC) (vide Memo No.: IAEC/2018/1001-1022 Dated: 11.04.2018) constituted as per article number 13 of the CPCSEA rules followed by the Government of India.

Blood and semen sample collection

Peripheral blood (2 ml) was aseptically collected from the jugular vein of the animals using a 16G needle in a 5 ml sterile tube with an anticoagulant (0.5 M EDTA). Frozen semen straws were procured from the GADVASU semen centre and private facility. Genomic DNA was isolated from whole blood using Phenol, Chloroform, Iso-amyl alcohol method [18] and stored at -20°. Whereas, genomic DNA from semen samples were isolated using NaCl and iso-amyl alcohol method [2].

Microsatellite genotyping

Fluorescent dye labelled (F'-5') six microsatellite- primer-pairs were selected from reported literature [8,20]. The selection of loci was based on the reported high polymorphism information content (PIC), high heterozygosity, and homogenous repeat motifs (Table 1 and 2). The DNA samples and the primer-sets were sent to Xcelris Lab Private Limited, Ahmedabad, and Gujarat for sequencing, length polymorphism analysis and typing.

Genotyped data were analysed bio-computationally using GENEPOP 32 [16,17] to find out the allelic frequencies. Cervus 3.0 [7] was used for analyzing the parameters viz. parentage analysis by calculating the log of odds (LOD) scores given by the natural logarithm of the overall likelihood ratio, homozygotes (Ho), heterozygotes (He), polymorphism information content (PIC), the average non-exclusion probability for one candidate parent (NE-1P), the average non-exclusion probability for one candidate parent given the genotype of a known parent of the opposite sex (NE-2P); the average non-exclusion probability for a candidate parent pair (NE-PP), estimated null allele frequency F (Null), the average non-exclusion probability for the identity of two unrelated individuals (NE-I), the average non-exclusion probability for the identity of two siblings (NE-SI) and Probability of Paternity (PP).

SN	Microsatellite Marker	Primer Sequence (5′ to 3′)	Primer Length	5′ Forward Dye Label	Tm (°C)	Expected Range	Reference
1	ILSTS089	F:AATTCCGTGGACTGAGGAGC R:AAGGAACTTTCAACCTGAGG	20 20	6-FAM	55	112-120	Kathiravan., *et al.* 2010 and Singh., *et al.* 2018
2	ILSTS095	F:GAAAGATGTTGCTAGTGGGG R:ATTCTCCTGTGAACCTCTCC	20 20	HEX	55	187-215	Singh., *et al.* 2018
3	CSSM019	F:TTGTCAGCAACTTCTTGTATCTTT R:TGTTTTAAGCCACCCAATTATTTG	24 24	TAMRA	55	127-161	Kathiravan., *et al.* 2010 and Singh., *et al.* 2018
4	ILSTS025	F: GTTACCTTTATATAAGACTCCC R: AATTTCTGGCTGACTTGGACC	22 21	6-FAM	55	110-126	Kathiravan., *et al.* 2010 and Singh., et al. 2018
5	ILSTS060	F:TAGGCAAAAGTCGGCAGC R:TTAAGGGGACACCAGCCC	18 18	HEX	55	162-198	Kathiravan., *et al.* 2010 and Singh., *et al.* 2018
6	ILSTS058	F: GCCTTACTACCATTTCCAGC R: CATCCTGACTTTGGCTGTGG	20 20	TAMRA	56	120-152	Kathiravan., *et al.* 2010 and Singh., *et al.* 2018

Table 1: Fluorescent labelled microsatellite markers for *Bubalus bubalis* population.

SN	Microsatellite Marker	Primer Sequence (5′ to 3′)	Primer Length	5′ Forward Dye Label	T_a (°C)	Expected Range	Reference
1	INRA035	F: ATCCTTTGCAGCCTCCACATTG R: TTGTGCTTTATGACACTATCCG	22 20	6-FAM	55	104-170	Jei Pei., *et al.* 2018
2	ILSTS034	F:AAGGGTCTAAGTCCACTGGC R:GACCTGGTTTAGCAGAGAGC	20 20	HEX	56	137-199	Sodhi., *et al.* 2007
3	ILSTS011	F: GCTTGCTACATGGAAAGTGC R: CTAAAATGCAGAGCCCTACC	20 20	TAMRA	57	249-169	Sodhi., *et al.* 2007
4	HEL09	F:CCCATTCAGTCTTCAGAGGT R: CACATCCATGTTCTCACCAC	20 20	6-FAM	55	137-169	Sodhi., *et al.* 2007
5	CSSM08	F:CTTGGTGTTACTAGCCCTGGG R:GATATATTTGCCAGAGATTCTGCA	21 24	HEX	56	180-202	Sodhi., *et al.* 2007
6	CSSM033	F:CACTGTGAATGCATGTGTGTGAGC R:CCCATGATAAGAGTGCAGATGACT	24 24	TAMRA	57	309-317	Pei., *et al.* 2018

Table 2: Fluorescent labelled microsatellite markers for *Bos taurus* population.

Genotyping by sequencing (dd-RAD)

Genotyping by sequencing (GBS) method is the reductional method of genome representation and widely used for identifying the SNP's within the genome using restriction enzymes. The double digest restriction-site associated DNA (dd-RAD) approach of GBS was used in this study and outsourced from Agri genome Private Labs Ltd, Kerala, India. Genomic DNA (> 50 ng/ul) was delivered to the facility.

Bio computational analysis of SNPs

Raw SNPs files were arranged in a spreadsheet, for each chromosome number, position, accession number, reference allele, MAF. Firstly, Mendelian error (MDE) SNPs were excluded from the genotype data-set using an in-house designed bioinformatics pipeline, to minimize the errors. X-Chromosome and scaffold data were excluded from the analysis. In the final selection, SNPs with at least 0.05 MAF (minor allele frequency) were retained for further analysis.

Haplotype analysis: The stretches of SNPs posited within 100 to 300 bp (for each chromosomal DNA) was identified through haplotype analysis, using an in-house bioinformatics pipeline. The criteria for haplotype selection were existing high variation among SNPs and presence of a minimum of five SNPs in the single haplo-

type. Primers were designed with an amplicon size of a maximum of 500 bp (minimizing sequencing errors) and specific to the target within the genome, without any spurious amplification (checked with Primer-BLAST (https://www.ncbi.nlm.nih.gov/tools/primer-blast/)). The Sanger sequencing facility was provided by Agrigenome Labs Private Limited, Kerala, India.

Validation of parentage through SNPs: The target regions were amplified for one offspring and three Sires DNA (among which one Sire was biological father of the offspring and the other two sides were unrelated). The SNP positions identified in haplotypes were compared among the groups to identify the pattern of inheritance using the multiple sequence alignment tools (Clustal Omega).

Paternity assignment: Sequoia R-Programming package is generally used for the paternity assignment in missing pedigree data. The requisite input datasets are the pedigree records with animal numbers and date of birth and animal data with their respective SNP positions. In analysis, two sets of data were used for paternity assignment, MDE free SNPs, and raw data without any filtering.

Principal component analysis (PCA): The SNP datasets were subjected to PCA study for three-dimensional representation and streamlining of huge data sets. In this study, *firstly* the sum of the first three PCA was compared with the different MAF values and *secondly,* the most informative SNPs were based on the possible combinations: (1) No copy of reference allele (AA/0) (2) Single copy of reference allele (AB/1) (3) Double copy of reference allele (BB/2).

Now, in this case we have 15 animals (belonging to 5 trios) for each species. The possible distribution of the genotypes among the 15 animals could range as shown in Table 8. Here we have got total 136 combinations. To find out the most informative distribution of genotypes among these 15 animals, the count of genotypes was multiplied to get a value representing the maximum possible score. In Table 11, the unique values of these scores (obtained by multiplication) have been organized and it is evident that there is total 20 such scores ranging from 0 to 125.

We considered the most informative SNP was that which is balanced in Sire, Dam and offspring i.e., present in equal number of genotypes, viz. 5 animals each of AA, AB and BB genotypes.

The buffalo data provided 66825 SNPs which were then classified into the above 20 scores, based on the number of observations (Table 9). That means, there were 51 cases where all the three genotypes (AA, AB, BB) had equal number of animals (i.e., 5 each), hence yielding the score as 125 (Figure 1).

The cumulative sum of thirteen PCA's was compared with obtained twenty groups of genotype combinations (scores obtained above; Table 10) for both of the species. This helped figure out the highest PCA explaining the maximum number of SNPs in each population.

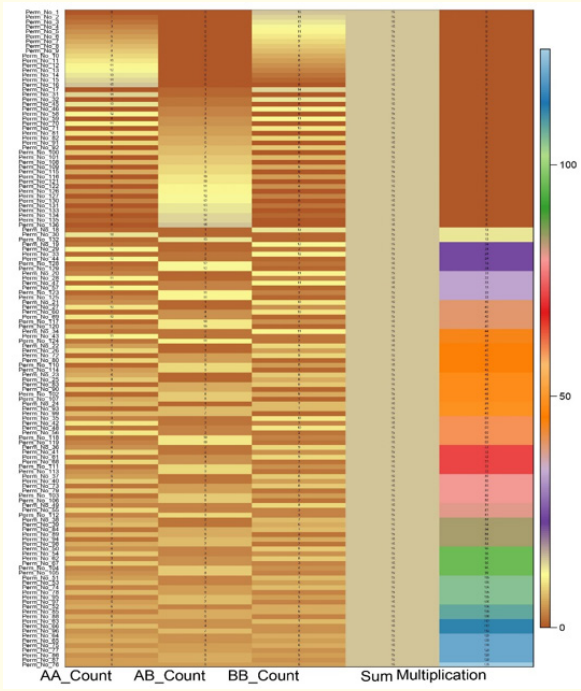

Figure 1: Possible permutation of the number of animals for 3 genotypes vis-a-vis the multiplication values of the number of animals.

Results and Discussion

In this study, parentage conformity between trios was accomplished using microsatellite genotyping. Microsatellite loci were genotyped based on allele length; bio-computational analysis was done separately.

Buffalo and cattle trios yielded observed heterozygosity (H_{obs}) = 1 and Expected heterozygosity (H_{exp}) ranged between 0.959 to

0.977 (Table 3), and 0.972 to 0.979, respectively (Table 4) with average polymorphic information content (PIC) value of 0.900 for each marker. Findings are suggestive of an isolate-breaking effect (the mixing of two previously isolated populations) because heterozygosity is higher than expected as animals in the study are not in-breeding, therefore confirms their unrelatedness. PIC for a sound genetic marker as described by [19] found the average value of PIC = 9.000, in this study is highly significant for considering the genetic markers as a valuable tool.

Locus	K	N	Hobs	Hexp	PIC	NE-1P	NE-2P	NE-PP	NE-I	NE-SI	HW	F(Null)	PP
INRA089	18	15	1.000	0.959	0.922	0.256	0.147	0.035	0.010	0.289	ND	-0.0394	99.99%
ILSTS095	20	15	1.000	0.977	0.942	0.203	0.113	0.022	0.006	0.279	ND	-0.0287	99.99%
CSSM019	19	15	1.000	0.975	0.939	0.211	0.118	0.024	0.006	0.281	ND	-0.0299	99.99%
ILSTS025	18	15	1.000	0.968	0.932	0.232	0.131	0.029	0.008	0.284	ND	-0.0336	99.99%
ILSTS060	19	15	1.000	0.972	0.937	0.218	0.122	0.026	0.007	0.282	ND	-0.0312	99.99%
ILSTS058	18	15	1.000	0.970	0.934	0.225	0.127	0.028	0.007	0.283	ND	-0.0323	99.99%

Table 3: The parameters of fifteen microsatellites in the *Bubalus bubalis*.

Locus: Marker name; k: Number of alleles at the locus; N: Number of individuals typed at the locus; Hobs: Observed heterozygosity; Hexp: Expected heterozygosity; PIC: Polymorphic information content; NE-1P: Average non-exclusion probability for one candidate parent; NE-2P: Average non-exclusion probability for one candidate parent given the genotype of a known parent of the opposite sex; NE-PP: Average non-exclusion probability for a candidate parent pair; F(Null): Estimated null allele frequency; PP: Probability of Paternity

Locus	K	N	Hobs	Hexp	PIC	NE-1P	NE-2P	NE-PP	NE-I	NE-SI	HW	F(Null)	PE
INRA035	20	15	1.000	0.977	0.942	0.203	0.113	0.022	0.006	0.279	ND	-0.0287	99.99%
ILSTS034	19	15	1.000	0.975	0.975	0.211	0.118	0.024	0.006	0.281	ND	-0.0299	99.99%
ILSTS011	19	15	1.000	0.972	0.972	0.218	0.122	0.026	0.007	0.282	ND	-0.0312	99.99%
HEL9	20	15	1.000	0.977	0.977	0.203	0.113	0.022	0.006	0.279	ND	0.0287	99.99%
CSSM08	20	15	1.000	0.977	0.977	0.203	0.113	0.022	0.006	0.279	ND	-0.0287	99.99%
CSSM033	21	15	1.000	0.979	0.979	0.196	0.108	0.021	0.005	0.278	ND	0.0276	99.99%

Table 4: The parameters of fifteen microsatellites in the cattle (*Bos taurus*).

Locus: Marker name; k: Number of alleles at the locus; N: Number of individuals typed at the locus; Hobs: Observed heterozygosity; Hexp: Expected heterozygosity; PIC: Polymorphic information content; NE-1P: Average non-exclusion probability for one candidate parent; NE-2P: Average non-exclusion probability for one candidate parent given the genotype of a known parent of the opposite sex; NE-PP: Average non-exclusion probability for a candidate parent pair; F(Null): Estimated null allele frequency; PP: Probability of Paternity

Assignment of parentage using microsatellite markers: LOD score for paternity confirmation of buffalo and cattle trios ranged from 7.04 to 8.18 and 2.62 to 8.53, respectively. It demonstrated that the offspring are correctly assigned to their respective sires. The findings were consistent with the pedigree information for both species. Further, the trio confidence was above 99% which has been indicated with "*" for all offspring (Table 5 and Table 6).

Probability of paternity (PP) was calculated as given by Stephenson 2016 for all trios (Table 3 and 4), which has been strongly indicative (PP = 99.99%) of correct assignments of all offspring to their respective parents and suggestive the results conform to the trio record with allele length-based microsatellite analysis.

Offspring	Assigned Sire	Dam	Pair Loci Number	Pair Loci Mismatching	Pair LOD Score	Pair Delta	Trio Confidence
Offspring-3	1	2	6	0	7.84E + 00	7.84E + 00	*
Offspring-6	4	5	6	0	7.04E + 00	7.04E + 00	*
Offspring-9	7	8	6	0	7.26E + 00	7.26E + 00	*
Offspring-12	10	11	6	0	8.18E + 00	8.18E + 00	*
Offspring-15	13	14	6	0	7.44E + 00	7.44E + 00	*

Table 5: Paternity assignment in *Bubalus bubalis.*

Offspring	Assigned Sire	Dam	Pair Loci Number	Pair Loci Mismatching	Pair LOD Score	Pair Delta	Trio Confidence
Offspring-103	101	102	6	0	8.35E + 00	8.35E + 00	*
Offspring-106	104	105	6	0	7.84E + 00	7.84E + 00	*
Offspring-109	107	108	6	0	8.53E + 00	8.53E + 00	*
Offspring-112	111	110	6	0	7.84E + 00	7.84E + 00	*
Offspring-115	114	113	6	0	2.62E + 00	2.62E + 00	*

Table 6: Paternity assignment in *Bos taurus.*

SNP-based approach

This study focused on the identification of breed-specific SNPs using haplotype analysis for identification of stretches of SNPs as mentioned above: R-programming based in-house designed pipeline used for analysis. Closely associated, SNPs positions less than < 100 bp, identified in each set of the chromosome of both of the species. The closely associated SNP positions were grouped, according to nucleotide start and end position within the chromosome. Position with high variability, filtered for primer designing using BLAST PRIMER *(In-silico),* targets specific positions, non-spurious amplification within the genome of the population (Table 7).

SN	Primer	Sequence	Length	SNP's covered	Tm	Amplicon *Bubalus*	Target Chr. No	Amplicon *Bos taurus*	Target Chr. No
1	PB36	CCAGGACTAAACGAAACGACCT	22	9	60	427	2	446	23
		CTACAGTTGCCATGCAGACG	20	9	60	427	2	446	23
2	PB51	GACCTCATAGTAAAACAAGACCGGA	25	9	60	379	12	400	11
		TCAGCATGCAGTACATCTCCC	21	9	60	379	12	400	11
3	PB68	ACGGCCTCCAAGCAACTTTAT	21	8	60	380	11	400	10
		ACCCTCACCTATTCCGTGTT	20	8	60	380	11	400	10
4	PB80	TCTCTTGCATTTCTGCTTTTAGCC	24	8	60	378	4	400	5
		CATGTCCTCCTGTGTATGGTGG	22	8	60	378	4	400	5
5	PB100	GTGGTGTGCGTTATTCAGCTAC	22	4	60	379	21	410	22
		CTGAACCCTAGTTGCTCCCAG	21	4	60	379	21	410	22
6	PC44	ATACTCCTGAAGTGTGATAGCTC	23	7	60	482	21	199	20
		CTCCAGTGTTGTCAACGGGA	20	7	60	482	21	199	20
7	PC56	TTGTATGGGAGAAAGTAAAAGCATC	25	6	58	400	3	401	6
		TCTTGCTTGGGTTCCTACTGA	21	6	58	400	3	401	6

8	PC91	CAAGGGGTTGTTGCAGTGAAT	21	6	56	413	24	413	22
		GAGCCCTCCCCATTCCTAAATC	22	6	56	413	24	413	22
9	PC94	GAGCTGTAGTCAAATAAAGAGCAGA	25	6	58	400	27	399	1
		GCTGCCTCGTTTGAGTCCTTT	21	6	58	400	27	399	1
10	PC42	CTCTGTGACAAGATGCCTTAAAGTT	25	6	60	400	20	399	19
		ACTTTGTCCATTTCCCTGTGGTG	23	6	60	400	20	399	19

Table 7: Unique Haplotype primers designed In-silico for *Bos taurus* and *Bubalus bubalis* population.

Sequoia R-Programming based analysis

Sequoia R-programming package was used for the assignment of paternity of the given samples using MDE free SNPs. The allelic frequencies, allele missingness for minor allele frequency, Likelihood ratios for dam, sire and parent pair and parent assignment was computed for two types of data set i.e., *(a) raw data* and *(b) filtered data* (excluding MDE error and 0.5 MAF), but the results were not so promising to assign the paternity as comparative to the paternity assigned through microsatellite markers. It could be due to the fewer sample size for the study.

Tapered SNP analysis

It is an approach, based on the streamlined process of narrowing down the huge data set of SNPs, to identify the most informative SNPs for the parentage assignment. PCA is the sole criteria to explain this approach. Firstly, we have demonstrated the use of the first three directions of principal component analysis. In order, if we expand further then it includes the further hurdles while maximizing the explanation.

The graph shows the maximum number of sums of three PCA is 0.440 (44%) and the least is 0.430 (43%). While reading the graph from right to left, with 0.5 MAF to 0.3 MAF the number of SNP's are increasing but the sum of the first three PCA is reducing down. Further, in the graph with MAF 0.22 to 0.1, the sum of the first three PCA start increasing but it doesn't reach up to the maximum sum (i.e., 0.440). This observation is it is not following any specific trend, to explain the PCA, three values were selected MAF > = 0.1 (Not maximum but near to maximum), MAF > = 0.22 (Least) and MAF = 0.499 (Maximum) (Figure 2).

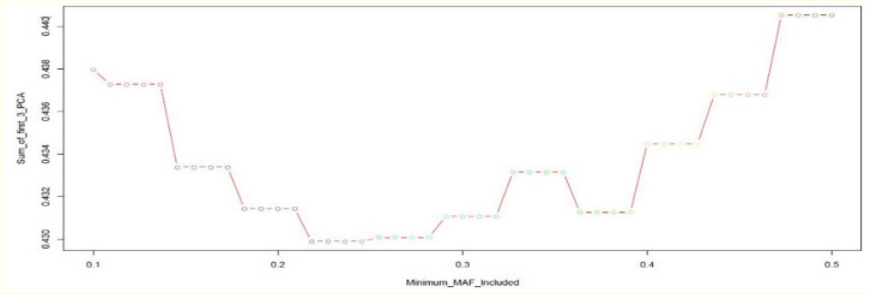

Figure 2: The sum of first three principal components (PCs) against different levels of MAF (minor allele frequency).

In another, the approach of strategy the informative SNPs were selected based on the reference allele of each position for each animal. In general, there could be three possible combinations (1) Where, no copy of reference allele (AA/0) is present (2) Where, a single copy of reference allele (AB/1) is present (3) and where the double copy of reference allele (BB/2). Based on the above three criteria of selecting the number of SNPs were calculated for an individual column of each SNP position. The total number of SNPs for all of the three criteria be equal to the number of the animals in each population i.e., 15. While multiplying the SNPs in three criteria for each position will give permutation numbers. Frequent repeating numbers are selected. It came out to be a total of "20"

different sets of genotypes in each, *Bos taurus* and *Bubalus bubalis* population. The variability among the individuals will be maximum for any SNP position if the count of each type of genotype is near to 5. i.e., the heterozygosity in the population will be balanced with homozygosity. It will help identify the most informative SNPs (e.g., 5*5*5 = 125, 6*5*4 = 120,6*6*3 = 108, 7*5*3 = 105). The sum of all PCA was calculated and the Cumulative sum of each genotype was calculated. The heatmap of the cumulative sum of PCA was drawn using R-programming to figure out the final selection of the SNPs in both of the populations.

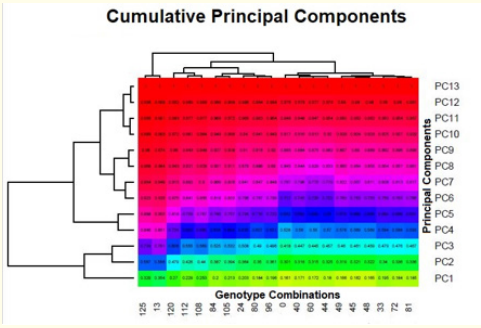

Figure 3: Heatmap of Cumulative sum of Principal Components (PCs) of buffalo trios.

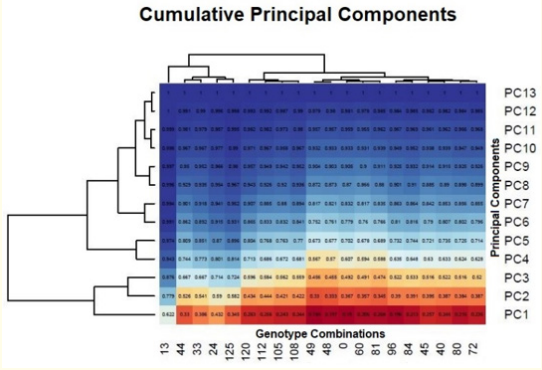

Figure 4: Heatmap of cumulative sum of principal components (PCs) of cattle trios.

The above heatmap shows that the cumulative sum of 51 SNPs explained 95.4% components in buffalo population and comparatively, in cattle cumulative sum of 1857 SNPs explained 95.43% of the components. Therefore, the set of 51 SNPs in buffalo and 1857 SNPs in cattle are highly informative for breed-specific parentage assignment studies in bovines.

SN	AA/0	AB/1	BB/2	Total	Multiplication
1	0	0	15	15	0
2	1	0	14	15	0
3	2	0	13	15	0
4	3	0	12	15	0
5	4	0	11	15	0
6	5	0	10	15	0
7	6	0	9	15	0
8	7	0	8	15	0
9	8	0	7	15	0
10	9	0	6	15	0
11	10	0	5	15	0
12	11	0	4	15	0
13	12	0	3	15	0
14	13	0	2	15	0
15	14	0	1	15	0
16	15	0	0	15	0
17	0	1	14	15	0
18	1	1	13	15	13
19	2	1	12	15	24
20	3	1	11	15	33
21	4	1	10	15	40
22	5	1	9	15	45
23	6	1	8	15	48
24	7	1	7	15	49
25	8	1	6	15	48
26	9	1	5	15	45
27	10	1	4	15	40
28	11	1	3	15	33
29	12	1	2	15	24
30	13	1	1	15	13
31	14	1	0	15	0
32	0	2	13	15	0
33	1	2	12	15	24
34	2	2	11	15	44
35	3	2	10	15	60
36	4	2	9	15	72
37	5	2	8	15	80
38	6	2	7	15	84
39	7	2	6	15	84

40	8	2	5	15	80
41	9	2	4	15	72
42	10	2	3	15	60
43	11	2	2	15	44
44	12	2	1	15	24
45	13	2	0	15	0
46	0	3	12	15	0
47	1	3	11	15	33
48	2	3	10	15	60
49	3	3	9	15	81
50	4	3	8	15	96
51	5	3	7	15	105
52	6	3	6	15	108
53	7	3	5	15	105
54	8	3	4	15	96
55	9	3	3	15	81
56	10	3	2	15	60
57	11	3	1	15	33
58	12	3	0	15	0
59	0	4	11	15	0
60	1	4	10	15	40
61	2	4	9	15	72
62	3	4	8	15	96
63	4	4	7	15	112
64	5	4	6	15	120
65	6	4	5	15	120
66	7	4	4	15	112
67	8	4	3	15	96
68	9	4	2	15	72
69	10	4	1	15	40
70	11	4	0	15	0
71	0	5	10	15	0
72	1	5	9	15	45
73	2	5	8	15	80
74	3	5	7	15	105
75	4	5	6	15	120
76	5	5	5	15	125
77	6	5	4	15	120
78	7	5	3	15	105
79	8	5	2	15	80
80	9	5	1	15	45
81	10	5	0	15	0
82	0	6	9	15	0
83	1	6	8	15	48
84	2	6	7	15	84
85	3	6	6	15	108
86	4	6	5	15	120
87	5	6	4	15	120
88	6	6	3	15	108
89	7	6	2	15	84
90	8	6	1	15	48
91	9	6	0	15	0
92	0	7	8	15	0
93	1	7	7	15	49
94	2	7	6	15	84
95	3	7	5	15	105
96	4	7	4	15	112
97	5	7	3	15	105
98	6	7	2	15	84
99	7	7	1	15	49
100	8	7	0	15	0
101	0	8	7	15	0
102	1	8	6	15	48
103	2	8	5	15	80
104	3	8	4	15	96
105	4	8	3	15	96
106	5	8	2	15	80
107	6	8	1	15	48
108	7	8	0	15	0
109	0	9	6	15	0
110	1	9	5	15	45
111	2	9	4	15	72
112	3	9	3	15	81
113	4	9	2	15	72
114	5	9	1	15	45
115	6	9	0	15	0
116	0	10	5	15	0
117	1	10	4	15	40
118	2	10	3	15	60
119	3	10	2	15	60
120	4	10	1	15	40
121	5	10	0	15	0

122	0	11	4	15	0
123	1	11	3	15	33
124	2	11	2	15	44
125	3	11	1	15	33
126	4	11	0	15	0
127	0	12	3	15	0
128	1	12	2	15	24
129	2	12	1	15	24
130	3	12	0	15	0
131	0	13	2	15	0
132	1	13	1	15	13
133	2	13	0	15	0
134	0	14	1	15	0
135	1	14	0	15	0
136	0	15	0	15	0

Table 8: All Permutation numbers obtained using possible genotypes.

SN	AA	AB	BB	Total	Multiplication
1	5	5	5	15	125
2	5	4	6	15	120
3	4	4	7	15	112
4	6	3	6	15	108
5	5	3	7	15	105
6	4	3	8	15	96
7	6	2	7	15	84
8	3	3	9	15	81
9	5	2	8	15	80
10	4	2	9	15	72
11	3	2	10	15	60
12	7	1	7	15	49
13	6	1	8	15	48
14	5	1	9	15	45
15	2	2	11	15	44
16	4	1	10	15	40
17	3	1	11	15	33
18	2	1	12	15	24
18	1	1	13	15	13

Table 9: Unique Haplotype combinations were selected.

SN	Genotype Combinations	Genotypes Included
1	125	34
2	120	231
3	112	158
4	108	204
5	105	350
6	96	445
7	84	721
8	81	216
9	80	673
10	72	652
11	60	533
12	49	645
13	48	1324
14	45	1441
15	44	283
16	40	1351
17	33	1740
18	24	1483
19	13	1857
20	0	38963

Table 10: Genotype Combinations with respective total genotypes.

Validation of haplotype-based primers for paternity assignment

Validation of three primer sets (PB51, PB68, PB100) for buffalo (Table 11-13) and three primer sets (PC44, PC91, PC42) for cattle (Table 14-16) were used in laboratory conditions for PCR amplicons. PCR amplification for aforesaid primer sets were done using high fidelity Taq Polymerase (PROMEGA). Amplicons were sent for the sanger sequencing at Agri genome Private Limited Labs, Kerela, India. The validation of primers is based on one set of biological offspring (named OS 1) sequence was aligned with reference genomic sequence, biological sire sequence and other two putative sire sequences, multiple sequence alignment tool (Clustal Omega). The positions of SNP's were compared throughout using reference SNP positions, none of them shown any significant difference within SNPs against query sequences. Primers sets have not shown any consistency in identifying the polymorphic SNPs for paternity

S. No.	Details	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7
1	Ref Seq Pos Orig Seq	47670593	47670595	47670598	47670611	47670614	47670617	47670622
2	Ref Seq Pos Base position	197	199	202	215	218	221	226
3	Ref SNP	T/Y	A/W	T/Y	G/K	C/M	A/M	A/M
4	Offspring	T	A	C	G	G	C	C
5	Sire 1	T	A	C	G	G	C	C
6	Sire 2	T	A	C	G	G	C	C
7	Sire 3	T	A	C	G	G	C	C
8	Original Pos	47670593	47670595	47670598	47670611	47670614	47670617	47670622
9	Seq Start Pos	47670397	47670397	47670397	47670397	47670397	47670397	47670397

Table 11: SNP positions, start and end positions, SNP details of PB51 primer validation.

S. No.	Details	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7
1	Ref Seq Pos Orig Seq	20028548	20028550	20028583	20028651	20028667	20028673	20028686
2	Ref Seq Pos Base position	11	13	46	114	130	136	149
3	Ref SNP	A/R	C/Y	T/Y	G/R	A/W	G/K	T/Y
4	Offspring	NA	NA	NA	NA	A	G	T
5	Sire 1	NA	NA	NA	NA	A	G	T
6	Sire 2	NA	NA	NA	NA	A	T	T
7	Sire 3	NA	NA	NA	NA	A	T	T
8	Original Pos	20028548	20028550	20028583	20028651	20028667	20028673	20028686
9	Seq Start Pos	20028538	20028538	20028538	20028538	20028538	20028538	20028538
10	New Pos	11	13	46	114	130	136	149

Table 12: SNP positions, start and end positions, SNP details of PB68 primer validation.

S. No.	Details	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7
1	Ref Seq Pos Orig Seq	41969649	41969689	41969777	41969793	41969833	41969912	41969918
2	Ref Seq Pos Base position	168	208	296	312	352	431	437
3	Ref SNP	C/Y	T/Y	C/S	G/S	T/Y/C	G/K	C/Y
4	Offspring	C	T	C	G	NA	NA	NA
5	Sire 1	C	T	C	G	NA	NA	NA
6	Sire 2	C	T	C	NA	NA	NA	NA
7	Sire 3	C	T	NA	G	NA	NA	NA
8	Original Pos	41969649	41969689	41969777	41969793	41969833	41969912	41969918
9	Seq Start Pos	41969482	41969482	41969482	41969482	41969482	41969482	41969482
10	New Pos	168	208	296	312	352	431	437

Table 13: SNP positions, start and end positions, SNP details of PB100 primer validation.

S. No.	Details	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7
1	Ref Seq Pos Orig Seq	14793900	14793908	14793919	14793965	14793966	14793979	14793998
2	Ref Seq Pos Base position	195	203	214	260	261	274	293
3	Ref SNP	A/W	C/Y	A/R/G	C/Y	G/R	T/Y	A/R
4	Offspring	A	C	G	C	A	T	A
5	Sire 1	A	C	A	C	G	T	A
6	Sire 2	A	T	A	C	G	T	A
7	Sire 3	A	C	A	C	G	T	A
8	Original Pos	14793900	14793908	14793919	14793965	14793966	14793979	14793998
9	Seq Start Pos	14793706	14793706	14793706	14793706	14793706	14793706	14793706
10	New Pos	195	203	214	260	261	274	293

Table 14: SNP positions, start and end positions, SNP details of PC44 primer validation.

S. No.	Details	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6
1	Ref Seq Pos Orig Seq	33955161	33955185	33955247	33955250	33955268	33955277
2	Ref Seq Pos Base position	181	205	267	270	288	297
3	Ref SNP	A/G/R	T/C/Y	C/A/M	A/R	G/R	C/T/R
4	Offspring	A	T	C	A	G	G
5	Sire 1	A	T	C	A	G	G
6	Sire 2	A	T	C	A	G	G
7	Sire 3	G	C	A	A	G	G
8	Original Pos	33955161	33955185	33955247	33955250	33955268	33955277
9	Seq Start Pos	33954981	33954981	33954981	33954981	33954981	33954981
10	New Pos	181	205	267	270	288	297

Table 15: SNP positions, start and end positions, SNP details of PC91 primer validation.

S. No.	Details	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7
1	Ref Seq Pos Orig Seq	69171419	69171449	69171471	69171506	69171516	69171529	69171538
2	Ref Seq Pos Base position	120	150	172	207	217	230	239
3	Ref SNP	T/Y	A/R	T/Y	C/S	C/Y	T/Y	C/Y
4	Offspring	T	G	C	C	T	C	T
5	Sire 1	T	A	T	C	C	T	C
6	Sire 2	T	A	T	C	C	T	C
7	Sire 3	T	A	T	C	C	T	C
8	Original Pos	69171419	69171449	69171471	69171506	69171516	69171529	69171538
9	Seq Start Pos	69171300	69171300	69171300	69171300	69171300	69171300	69171300
10	New Pos	120	150	172	207	217	230	239

Table 16: SNP positions, start and end positions, SNP details of PC42 primer validation.

assignment even in known biological parent-offspring pairs, failure could be due to the sequencing error and large number of identified (Table 7) primer sets further need to be validated.

Conclusion

In this study, we have conducted the paternity assignment study using fluorescent labelled microsatellite markers have assigned the paternity with a confidence interval of 98% and probability of paternity of putative sires recorded as 99.99%, both of the studied population has high genetic variability along with PIC > 0.05 for markers. Confirmed trios were used for the SNP-based paternity assignment through R-programming, due to less sample size is found to be inadequate. The sum of the first three PCA compared (maximum 0.440 and 0.430 least) within the range of 0.1-0.5 MAF. Selected, three values SNPs with MAF> = 0.1 (*Bos:* 16620 SNPs, *Bubalus*: 14965 SNPs), MAF> = 0.22 (*Bos:* 1796 SNPs, *Bubalus:* 4207 SNPs) and MAF = 0.499 (*Bos:* 1671 SNPs, *Bubalus:* 5202 SNPs) shows no specific trend. The cumulative proportion of thirteen PCA's was compared with twenty Genotype combinations in buffalo and cattle. The cumulative sum of 51 SNPs explained 95.4% of components. Whereas, in *Bos taurus* Cumulative sum of 1857 SNPs explains 95.43% of the components. A set of 51 SNP's in Bubalus bubalis and 1857 SNPs are highly informative for breed-specific parentage assignment studies in bovines. In this study, we could identify a minimum number of most informative SNPs for the two species were 51 for buffalo (explaining 72.4% variation for the first 3 PCA) and 34 SNPs for crossbred cattle (explaining 72.4% variation for the first 3 PCA). The in silico identified SNPs warrants validation in unrelated trios. It can be concluded that the identified SNPs are different in cattle and buffalo and the SNPs also vary with the MAF-threshold being considered for the selection of SNPs. Laboratory based evaluation for the accuracy of haplotype-based SNP primers is further warranted in large sets.

Conflict of Interest

We certify that there is no conflict of interest with any financial organization regarding the material discussed in the manuscript.

Acknowledgement

The authors thankfully acknowledge the financial support provided by "*Rashtriya Krishi Vikash Yojna*" (National Agricultural Development Scheme) (RKVY) project "Identification of the species-specific molecular markers for parentage determination in livestock" (RKVY-12: Enhancement of production and productivity potential of livestock, poultry, and fisheries sector for the socio-economic upliftment of the farmers of Punjab).

Bibliography

1. Banos Georgios., *et al.* "Impact of paternity errors in cow identification on genetic evaluations and international comparisons". *Journal of Dairy Science* 84.11 (2001): 2523-2529.

2. Darbandi Mahsa., *et al.* "Reactive oxygen species-induced alterations in H19-Igf2 methylation patterns, seminal plasma metabolites, and semen quality". *Journal of Assisted Reproduction and Genetics* 36.2 (2019): 241-253.

3. Davis GP and SK DeNise. "The impact of genetic markers on selection". *Journal of Animal Science* 76.9 (1998): 2331-2339.

4. Heaton Michael P., *et al.* "Interleukin-8 haplotype structure from nucleotide sequence variation in commercial populations of US beef cattle". *Mammalian Genome* 12.3 (2001): 219-226.

5. Heaton Michael P., *et al.* "Selection and use of SNP markers for animal identification and paternity analysis in US beef cattle". *Mammalian Genome* 13.5 (2002): 272-281.

6. He Jiangfeng., *et al.* "Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding". *Frontiers in Plant Science* 5 (2014): 484.

7. Kalinowski, Steven T., *et al.* "Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment". *Molecular Ecology* 16.5 (2007): 1099-1106.

8. Kathiravan Mathur Nadarajan., *et al.* "Sonoassisted microbial reduction of chromium". *Applied Biochemistry and Biotechnology* 160.7 (2010): 2000-2013.

9. Krawczak Michael. "Informativity assessment for biallelic single nucleotide polymorphisms". *ELECTROPHORESIS: An International Journal* 20.8 (1999): 1676-1681.

10. Kwok Pui-Yan. "Methods for genotyping single nucleotide polymorphisms". *Annual Review of Genomics and Human Genetics* 2.1 (2001): 235-258.

11. Lindblad-Toh Kerstin., *et al.* "Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse". *Nature Genetics* 24.4 (2000): 381-386.

12. Markovtsova Lada., *et al.* "The age of a unique event polymorphism". *Genetics* 156.1 (2000): 401-409.

13. Munoz Patricio R., *et al.* "Genomic relationship matrix for correcting pedigree errors in breeding populations: impact on genetic parameters and genomic selection accuracy". *Crop Science* 54.3 (2014): 1115-1123.

14. Nielsen Rasmus. "Estimation of population parameters and recombination rates from single nucleotide polymorphisms". *Genetics* 154.2 (2000): 931-942.

15. Ravi Sachidanandam., *et al.* "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms". *Nature* 409.6822 (2001): 928-933.

16. Raymond Michel and François Rousset. "An exact test for population differentiation". *Evolution* (1995): 1280-1283.

17. Rousset Francois. "genepop'007: a complete re-implementation of the genepop software for Windows and Linux". *Molecular Ecology Resources* 8.1 (2008): 103-106.

18. Sambrook JF and Russell DW. "Molecular Cloning: A Laboratory Manual (3-Volume Set) Edition". 3Publisher: Cold Spring Harbor Laboratory Press.

19. Serrote Caetano Miguel Lemos., *et al.* "Determining the Polymorphism Information Content of a molecular marker". *Gene* 726 (2020): 144175.

20. Singh Simrat Pal., *et al.* "Evolutionary divergence of the rye Pm17 and Pm8 resistance genes reveals ancient diversity". *Plant Molecular Biology* 98.3 (2018): 249-260.

21. Stormont Clyde. "Contribution of blood typing to dairy science progress". *Journal of Dairy Science* 50.2 (1967): 253-260.

22. Thomson Russell., *et al.* "Recent common ancestry of human Y chromosomes: evidence from DNA sequence data". *Proceedings of the National Academy of Sciences* 97.13 (2000): 7360-7365.

23. Visscher PM., *et al.* "Estimation of pedigree errors in the UK dairy population using microsatellite markers and the impact on selection". *Journal of Dairy Science* 85.9 (2002): 2368-2375.

24. Wang David G., *et al.* "Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome". *Science* 280.5366 (1998): 1077-1082.