GENETIC DIVERSITY ANALYSIS USING IDENTIFIED GENOME-WIDE SNPs IN RIVERINE BUFFALOES

Thesis

Submitted to the
DEEMED UNIVERSITY
Indian Veterinary Research Institute
Izatnagar - 243 122 (U.P.), India

Dr. Shiv Kumar Tyagi
Roll No. M-5952

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Veterinary Science
(Animal Genetics and Breeding)

2020

Dedicated To...

My Beloved Family

&

Guide

भा.कृ.अनु.प.–भारतीय पशु चिकित्सा अनुसंधान संस्थान
(सम विश्वविद्यालय)
इज्जतनगर –243122, (उ.प्र.), भारत

DIVISION OF ANIMAL GENETICS
INDIAN VETERINARY RESEARCH INSTITUTE
(Deemed University)
IZATNAGAR - 243 122, U.P., INDIA

Dr. Ashwni Kumar Pandey,
M.Sc., Ph.D.
Principal Scientist

Dated: 14 | 12 | 2020

Certificate

This is to be certified that the research work embodied in this thesis entitled **"Genetic diversity analysis using identified Genome-wide SNPs in Riverine Buffaloes"** *submitted by* **Dr. Shiv Kumar Tyagi, Roll No. M-5952,** *for the award of* **Master of Veterinary Science** *Degree in* **Animal Genetics and Breeding** *at ICAR-Indian Veterinary Research Institute, Izatnagar, is the original work carried out by the candidate himself under my supervision and guidance.*

It is further certified that **Dr. Shiv Kumar Tyagi, Roll No. M-5952,** *has worked for more than 21 months in the Institute and has put in more than 150 days attendance under me from the date of registration for the Master of Veterinary Science Degree in this Deemed University, as required under the relevant ordinance.*

(Ashwni Kumar Pandey)
Chairman
Advisory Committee

Certificate

We the undersigned members of Advisory Committee of **Dr. Shiv Kumar Tyagi, Roll No. M-5952** *a candidate for the degree of* **Master of Veterinary Science** *with the major discipline* **Animal Genetics and Breeding**, *agree that the thesis entitled* **"Genetic diversity analysis using identified Genome-wide SNPs in Riverine Buffaloes"** *may be submitted in partial fulfillment of the requirement for the degree.*

We have gone through the contents of the thesis and are fully satisfied with the work carried out by the candidate, which is being presented for the award of **Master of Veterinary Science** *Degree of this Institute.*

It is further certified that the candidate has completed all the prescribed requirements governing the award of **Master of Veterinary Science** *Degree of the Deemed University, ICAR-Indian Veterinary Research Institute, Izatnagar.*

Signature
Name
External Examiner

Date : .....

(Ashwni Kumar Pandey)
Chairman
Advisory Committee

Date : .....

MEMBERS OF STUDENT'S ADVISORY COMMITTEE

Dr. Amit Kumar, Senior Scientist
Division of Animal Genetics, ICAR-IVRI, Izatnagar .....

Dr. Anuj Chauhan, Senior Scientist
Division of Animal Genetics, ICAR-IVRI, Izatnagar .....

Dr. Arvind Sonwane, Senior Scientist
Division of Animal Genetics, ICAR-IVRI, Izatnagar .....

Dr. A.K.S. Tomar, Principal Scientist
Livestock Production and Management Section, ICAR-IVRI, Izatnagar .....

ACKNOWLEDGEMENTS

"**If we knew what were doing it wouldn't be called research**" - Albert Einstein

I've attempted here to pen down my feelings of gratitude towards all such wonderful people who've rekindled my flame all this time.

First and foremost above every mortal, I thank the ever pervading essence of the universe and the treasure of all knowledge, who kept me alive, flooded me with the energy, hope and allowed me to complete this sojourn.

The process of earning a master degree and writing a dissertation is long and arduous task and it is certainly not done single handedly. I place on record my heartfelt sincere gratitude to my advisor **Dr. Ashwni Kumar Pandey**, P.S., Division of Animal Gentics ICAR-IVRI whose dynamic supervision, keen interest in the work, whole hearted encouragement and philanthropic attitude enabled me to achieve my goals. Without his kind and patient instruction, it would have been impossible for me to finish this thesis. Sir, your encouraging attitude, correlating ideas, understanding the obvious and dedication towards any work you do has helped me to smoothly sail through this long ordeal. "When the going gets tough, the tough get going". I will never find words to tell what I owe to him. From the bottom of my heart, I thank you for everything Sir.

I emphatically owe my sincere thanks to the learned members of my advisory committee **Dr. Amit Kumar**, Senior Scientist, AG Division, **Dr. Anuj Chuahan**, Senior Scientist, AG Division, **Dr. A.K.S. Tomar**, Principal Scientist, LPM Division, and **Dr. Arvind A. Sonawane**, BTY Division for their constant co-operation, help and encouragement during the research work.

I humbly express my deep sense of gratitude and indebtedness to our head (A) of the division **Dr. Bharat Bhushan**, P.S. for his ever encouraging attitude and unflinching support throughout the tenure of my MVSc.

I shall remain grateful to **Dr. Pushpendra Kumar**, Principal Scientist, **Dr. Sanjeev Kumar**, Principal Scientist, **Dr. Ajay Kumar Sharma**, Principal Scientist, **Dr. Gyanendra Kumar Gaur**, Principal Scientist, LPM section, **Dr. Ranvir Singh**, Senior Scientist and **Dr. Manjit Panigrahi**, Scientist (SS) for their generous help, constant support and encouragement in all the spheres of work.

I sincerely thank **Director**, **Joint Director** (Academics) and **Joint Director** (Research), ICAR-IVRI, Izatnagar for providing necessary facilities for this study and IVRI for providing me the financial assistance during this MVSc.

Sincere thanks to **Dr. Jay Prakash Gupta**, Assistant Professor, SDAU And **Dr. Ratan Deep Singh**, Assistant Professor, SDAU, who always had faith in me even when I didn't. Thank you for helping me throughout this journey.

My acknowledgement will never be complete without the special mention of my lovely senior **Dr. Arnav Mehrotra and Dr. Akansha Singh,** for their support. Their all time support in my experiments during MVSc is unforgettable.

From the bottom of my heart, I want to sincerely thank my Seniors **Dr. Babul lal Saini, Dr. Amit Baranwal, Dr. Satish Kumar, Dr. Pruthviraj D, Dr. Ashish Baladhare, Dr. Wagh Shivaji S, Dr. Manoj Kumar, Dr. Amol Talokar, Dr. Mitek Tarang, Dr. Shweta Sachan, Dr. Manjari Pandey, Dr. Sudarshan Mahala, Dr. Chirag Chaudhary, Dr. Ezhilvadhana and Dr. Dhanpal** for their constant support, for their willingness to help whenever I was in need.

I fondly remember the pleasant company of my divisional batch mates **Drs., Mayank Darji, KA Saravanan, Latha Preethi A, Nandhini G & Shahista Sarin Lodhi** and thank them for their constant buoying, fun, in class room.

Thanks to my Juniors Drs, Munish, Chand, Anuradha, Sakshi, Divya, Pranisha and Amit for all the happy moments.

I feel extremely elated to thank my crazy people**, Drs., Vinay, Jitendra, Shashikant, Gyanendra, Alok, Tushar, Pradeep, Anup, Devansh.** My every moments was extremely memorable, joyous and stress free with you all guys.

I would dedicate a sentence to my UG college senior family members **Dr. Shyam Sundar Chaudhary, Dr. Shiv baran Singh, Dr. Dilip Yadav and seniors in IVRI, Dr Rohit Jaiswal, Dr. Sandeep Chaudhary, Dr. Shailesh Patel, Dr. Rohit Kumar singh** who is always by my side, no matter what, kept my spirits high when I felt low and guided to chase my goals always. Thanks Sir

I sincerely and very humbly acknowledge the help and moral support extended by all technical staff – Shri(s) Rajeev ji, Prabhat ji, Harnandan ji, Sarvjeet ji, Ranjeet ji, Saurabh ji, Wasim ji, Horilal ji and Sarafat ji at the Division of Animal Genetics for their support in co-operation in academic affairs.

I owe my thanks to **Darmendra Chachu, D.P.** and **Kuldeep** for their help in presenting this manuscript nicely.

I find no words to express my depth of gratitude to immortal love, blessing and never ending sacrifices of my beloved parents who had toil hard to bring me to stage where I stand now. My father Mr. Dori Lal Tyagi has been a guiding light to me especially when faced with challenges. The sacrifices of my mother Mrs. Manju Devi can never be reciprocated let alone repayment. I have complicated my work with the hope that it would make you proud. I dedicate all my success to you. I express my love and affection to my sweet lil' sister Priyanka and younger brother Vishal.

Date: 14-12-2020
Place: ICAR-IVRI, Izatnagar

(Shiv Kumar Tyagi)

ABBREVIATIONS

AnGR	:	Animal Genetic Resources
BAM	:	Binary Alignment Format
BCF	:	Binary Calling Format
CRoPS	:	Complexity Reduction of Polymorphism Sequencing
ddRAD	:	Double:digest RAD sequencing
DNA	:	DeoxyriboNucleic Acid
EDTA	:	Ethylene Diamine Tetra Acetic acid
GBS	:	Genotyping by Sequencing
GDP	:	Gross Domestic Product
GoI	:	Government of India
GWAS	:	Genome Wide Association Studies
GWSS	:	Genome wide sampling sequencing
INDEL	:	Insertion and Deletion
LD	:	Linkage Disequillibrium
MT	:	Metric Tonnes
NCBI	:	National Center for Biotechnology Information
Ne	:	Effective population size
NGS	:	Next Generation Sequencing
OD	:	Optical Density
PCR	:	Polymerase Chain Reaction
PCA	:	Principal component analysis
QC	:	Quality Control
RAD	:	Restriction Site associated DNA Sequencing
RD	:	Read Depth
RE	:	Restriction Enzyme
ROH	:	Runs of Homozygosity
RNA	:	Ribonucleic Acid
RPM	:	Revolutions Per Minute
RRL	:	Reduced Representation Libraries
SAM	:	Sequence Alignment Format

SDS	:	Sodium Dodecyl Sulfate
SNP	:	Single Nucleotide Polymorphism
TAE buffer	:	Tris : Acetic acid : EDTA buffer
TE buffer	:	Tris : EDTA buffer
T_S	:	Transition
T_V	:	Transversion
UV	:	Ultra violet
VCF	:	Variant Calling Format
WGS	:	Whole Genome Sequencing

LIST OF TABLES

LIST OF FIGURES

CONTENTS

Introduction

1 INTRODUCTION

The livestock sector plays a major role in the economy in many developing countries, including India. It provides food (more precisely animal protein in human diets), employment, job opportunities, draught energy, means of transportation and organic crop production fertilizers. In reality, animals are regarded as farmers ' financial resources because they function as protection against the threat of crop failure due to drought and other unfavorable climatic conditions. Across India, throughout recent decades, the dairy industry has made tremendous development. India is the world's largest producer of milk today, and India contributes about 20% to global milk production. Throughout India, buffalo is the largest dairy commodity and, owing to its adaptability to severe climatic conditions, resistance to tropical diseases and longevity under inadequate feeding and management practices, it holds an important place throughout India's agricultural economy (Thiruvenkadan *et al*., 2013)

Domesticated buffalo species were grouped into river (Bubalus bubalis) and swamp buffalo (Bubalus carabanesis). The current global buffalo population is 193.8 million (http://faostat3.fao.org/browse/Q/QA/E) and has been rising steadily over the past two decades at a rate of 2% per year (Moaeen-ud-Din, 2014). In India the buffalo population has increased from 108.7 million (2012) to 109.85 million (2019) showing a growth of 1.1 % (20th livestock census) that produce 74.70 MT out of the total milk production in India. The contribution of the Agriculture sector in Indian economy is 17.4% of total GDP, of which the Livestock sector contributes 25.07%. India hold the first rank in milk production with a total production of 176.3 MT during 2017 – 2018 (DAHDF, GoI). India is rich in buffalo AnGR, and so far 17 breeds of buffalo have been registered, out of which 8 breeds are famous for their milk

production viz., Murrah, Bhadawari, Jaffarabadi, Surti, Mehsana, Godavari, pandharpuri and Nili-ravi. Haryana-born Murrah buffalo has a relatively higher milk production of 1800 kg/ lactation and its tolerance to India's various environmental conditions. Murrah buffalo are widely used to rate other Indian buffalo populations to increase the production of milk.

The traditional method of genetic improvement of buffalo involves selection, crossbreeding, outbreeding, hybridization, etc. By reducing time, storage, commitment and maintaining long term sustainability, the molecular genetic system will help conventional breeding. Higher productivity can be accomplished by leveraging the genetic variation within and between breeds that can help improve the genetic benefit. Genetic advancement in Indian buffalo breeds includes the discovery and creation of genome-wide markers. The entire water buffalo genome in India was first sequenced in NBAGR, Karnal (Tantia *et al*., 2011). A high-quality water buffalo genome assembly (De-novo), containing 2.76 Gb in scaffolds (Zimin *et al*., 2013), was created and assembled under the International Buffalo Genome Consortium by the University of Maryland, USA, USDA-ARS, USA and CASPUR, Italy. A buffalo genome sequence of 2.77 Gb was reported recently in 2014 by collaborative work between Bangladesh (LalTeer Livestock Limited) and China (Beijing Genomics Institute). Axiom ® Buffalo Genotyping Array was recently developed by Affymetrix (http:/www.affymetrix.com/ catalog/prod740001/AFFY/Axiom percent 26 million 23174 percent 3B-Buffalo-Genotyping-Array#1) utilizing genomic data from several breeds (Italian Mediterranean, Aza-Kheli, Nili-Ravi, Murrah, Egyptiana Kundhi, Jaffarabadi, and Philippine Swamp) which allowed the classification of several types of buffalo breeds. The array of buffalo genotyping includes a total of 90k putative SNPs and has been tested in Italian and Brazilian buffalo populations.

Whole-genome sequencing for genotyping the SNP is complex and the integration of bioinformatics is very difficult. The entire genome has to be sequenced for genome-wide SNP detection, in which there will be a lot of uninformative, repeated series. Also it is more expensive. Compared with entire genome sequencing approaches for SNP exploration, a 35-fold cost reduction by RAD seq methods is feasible. Compared with entire genome sequencing approaches for SNP exploration, a 35-fold cost reduction by RAD seq methods is feasible.

NGS based RAD sequencing methods are an alternative to WGS, wherein simultaneous sequencing, genotyping and multiplexing are facilitated.

SNP exploration was extensively conducted in cow, horse, chicken, donkey, camel, etc. using RAD seq methods. This method having advantage of minimizing the Repetitive sequences, and thereby reducing the cost of sequencing and genotyping. Thus, in a matter of weeks, we can sequence thousands of individuals. There are various RAD sequencing methods viz., Complexity Reduction of Polymorphism Sequencing (CRoPS) (Osrow *et al*., 2007), Reduced Representation Libraries (RRL) (Van-Tassel *et al*., 2008), Restriction Site associated DNA Sequencing (RAD seq) (Baird *et al*., 2008), Genotyping by Sequencing (GBS) (Elshire *et al*., 2011), Double-digest RAD sequencing (dd RAD) (Peterson *et al*., 2012) which have their own benefits and drawbacks.

Genome annotation analyzes a genome's raw sequence and identifies important biological and genomic characteristics such as mutations, mobile elements, repeat elements, replication, and polymorphism, which is a challenging aspect of genome sequencing. The SNPs found by the different approaches to the NGS were not annotated in depth to the genomes.

The presence of genetic diversity within any livestock population is equally important for any type of buffalo population i.e., pure indigenous or graded. As per the latest 20th Livestock census trends, the Buffalo resource base in India is on increasing phase (by 1.04%) and possesses only a little genetic variability that need to be exploited through genetic means. Murrah forms the main milch buffalo that is prevalent in India and is used for selective breeding and upgrading of non-descript buffalo not only in India but also in foreign countries. The genetic diversity exists within and among different populations owing to the processes of evolution, selection and domestication being applied on these populations over the centuries of their rearing. Technically, genetic diversity refers to the occurrence of variation in a population group of various alleles and genotypes and is also reflected in physiological, morphological and behavioural variations within and among different populations (Frankham *et al*., 2002). Apart from the process of evolution, selection and domestication, different other phenomenon also helps in establishing and maintaining the genetic diversity across different populations (Groeneveld *et al*., 2010). Other processes mainly include mutation, migration, adaptation procedures and gentic drift. The presence of genetic diversity within buffalo populations is very important in order to ensure long term adaptation and conservation of efficient Animal

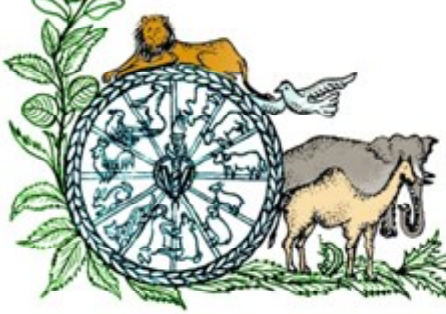
genetic resources (AnGR) (Groeneveld *et al*., 2010; Hanotte *et al*., 2010).

In order to facilitate the processing of a large amount of genomic data generated from raw data, various bioinformatics software has also been developed. In modern times, genome-wide SNP genotypic data for the study and/or prediction of different genetic parameters of an organism or population can be very easily evaluated. With the help of bioinformatics and statistical tools, Runs of homozygosity, linkage disequilibrium, Principal component analysis, admixture and various other genetic predictions have been effectively and accurately made. These tools include a basic interface that can be used by a researcher to make precise predictions based on the variants present at different locations.

In view of the very limited genomic information available in buffalo, it is envisaged to undertake the present work with the following objectives:

1. **To identify genome wide SNPs in riverine buffaloes using ddRAD approach.**
2. **To annotate SNPs to candidate genes for production and reproduction traits.**
3. **To ascertain genetic diversity analysis in the riverine buffaloes.**

Review of Literature



2 REVIEW OF LITERATURE

2.1 ORIGIN AND DOMESTICATION OF BUFFALO

Asian and European buffaloes belong to the Bubalus family, while the Syncerus group belongs to the African buffaloes. The Asian species, also defined as stream or river buffalo, is made up of two forms (river or swamp) that can be differentiated by their size, behavior, use and location. The Indian subcontinent, Egypt and Mediterranean basin of Europe's river buffalo (Bubalus bubalis), mostly for milk production (Cockrill, 1982), but all of them are also dual-purpose animals with good meat characteristics, while their meat capacity remains unexplored and untapped. The swamp buffalo is more or less a perpetual denizen of marshy ground, wallowing in mud and grazing on rough marsh grass. It is found mainly in South East Asia and China and has a very limited or no role in the processing of milk. The buffalo's main contribution in these areas was the draft force for the preparing of grain, rural transportation, threshing, water-raising and oil recovery from oilseeds (National Academy of Science, 1981).

2.2 BREEDS OF RIVER BUFFALOES

In order to devise rational breeding strategies for optimum use and preservation of accessible genetic variability in India, it is essential to understand the genetic structure and relationships between different breeds. Diversity in domestic animal species is generally perceived in terms of differences referred to as breeds. Turton defined breed as a homogeneous, sub-specific group of domestic animals with physical characters definable and observable that allow it to be differentiated from other similarly identified groups within the same genus by

visual evaluation, or a homogeneous category where its distinct identity would generally be accepted by geographical distinction from phenotypically similar groups.' That breed is the result of mutation and genetic drift, as well as independent adaptation and development, with specific climate-imposed selection pressures, infectious pests and diseases, accessible food and human-imposed requirements.

Buffalo breeds of India used in the study are:

2.2.1. Murrah

Murrah is said to have home town in northwest India, but this buffalo breed has established in every part of the country mostly as a popular milk breed, renowned for its jet black body color and tightly coiled horns. These were called "Murrah", meaning' curled. As previously mentioned, this breed has spread to virtually every corner of the country and is either significantly established in pure form or used as an improver breed to bred local buffaloes in India as well as in other countries. Economic characteristics when measured under organized farm conditions were found to be around 2480.38±55.06 of milk yield in large herds in 305 days of lactation with a maximum yield of about 22 kg; 7.32% of fat, 3.62% of protein, 9.52% of SNF (Annual report, NPB 2017/18).

2.2.2. Mehsana

Mehsana a dairy buffalo breed belongs to the northern Gujarat districts of Mehsana, Sabarkantha, and Banaskantha. It is presumed that this breed evolved from the hybrids of Murrah and Surti, so it is similar with these two breeds in many characteristics (Olver, 1938). The legs are smaller but the body is wider than Murrah. Unlike Murrah's bulging eyes, the face is longer and heavier. The horns mimic Murrah, but in the end, they are less rounded than the Murrah type, but with the irregular shape, they are larger. The udder has a good shape. The color of the skin dropped between jet black and gray, and white marks were also noticed on the head, tail tips. The hair found at the bottom of the body and also at the tail switch is usually black. The breed's males aren't being used for draught purpose. Good persistency for milk production is found in this breed. The average milk yield is between 598-3597 Kg with average of 1988 kg per lactation with 6.46±0.17 fat%, 3.87±0.05 protein%, 9.13±0.06 SNF%, 15.59±0.18 TS%. (Mishra *et al*., 2008)

2.2.3. Surti

The Surti buffalo is a water buffalo breed found in Gujarat, India's Charottar region. It is centered between the rivers of the Mahi and the Sabarmati. This breed's strongest livestock are located in the Gujarat districts of Anand, Baroda and Kaira. Many other titles like Surati, Gujarati, Nadiadi, Talabda, Charotar and Deccani even identify it. Surti buffalo is a medium-sized mammal for whom the body color is rusty brown or silver-gray. Such animal's heads are relatively wide and long with convex form in between horns at the edges. The horns are flat and sickle shaped. The horns grow in a downward and backward direction and then form a hook at the tip. The color of their skin is either black or brown. Such animal's back is unique and flat. There are two white collars in some healthy buffaloes. Ecomomic traits when measured under well organized farm conditions found to be around 1617.70 ± 72.01 kg milk production (305 days milk yield) with 9.75 ± 0.24 average peak yield and 8.11% of Fat, 3.67% of Protein, 9.52% of SNF (Annual report, NPB 2017/18).

2.2.4. Bhadawari:

Bhadawari is a dual-type buffalo breed from central and northern India, also known as "Etawah". The breeding tract is part of the former state of Bhadawar, from which the name of the breed originates. Bhadawari buffaloes are present in Uttar Pradesh and Madhya Pradesh in the ravines of the Yamuna, Chambal, and Utangan rivers. The breeding areas are Bah tehsil of Agra; Chakarnagar and Barhpura blocks of Etawah; Ambah and Porsa tehsils of Morena; and Mahangaon tehsil of Bhind district. They are blackish copper to light copper coloured but over the legs, wheat straw like colour is there. On the lower side of the body, there are two white lines, "Chevron," named in local language as "Kanthy". Before running parallel backward near the neck and gradually turning upward, the horns are black curling slightly outward and downward. The animals of this breed are known for their ability to use low-quality coarse fodder available in the region. Although the overall lactation yield is lower, the milk's fat content has been reported to be as high as 12.8%. The Breed's average milk yield is 1294 kg per lactation with an average 7.88% fat (yield varying from 540-1400 kg per lactation and 6 to 12.8% fat) (AGRI-IS, NDDB).

2.2.5. Pandharpuri

Pandharpuri is the native Maharashtra breed. The name of the geographical region is named after them, i.e. Block of Pandharpur in the Maharashtra district of Solapur. The breeding tract includes the Maharashtra districts of Solapur, Sangli and Kolhapur. The buffaloes are concentrated in the Solapur district of Pandharpur, North Solapur, South Solapur, Barshi, Akkalkot, Sangola and Mangalvedha tehsils; the Sangli district of Miraj, Walwa, Jathand Tasgaon tehsils; and the Kolhapur district of Karveer, Shirol, Panhala, Radhanagri, Hatkanangale and Gadhinglaj tehsils. The animals have multiple milk let down capability. In general, Pandharpuri buffaloes are black, but the colour varies from light to deep black. White marks are seen on the forehead; also on legs and tail in few animals. The horns are very long and stretch past the shoulder blade, up to the pin bones occasionally. On average, buffaloes yield 1790 kg of milk per lactation with 8% fat per lactation (AGRI-IS, NDDB).

2.2.6. Toda

The Toda breed is known after its herdsmen, the Toda tribe of the Nilgiris in the State of Tamil Nadu. This is a valuable breed of buffalo that thrives well in high rainfall and high wetland conditions. It is used primarily for function, socio-cultural and religious ceremonies. In general, the calves are fawn coloured at birth and the fawn colour shifts to ash grey at about 2 months of age. The predominant coat colours for adult buffaloes are fawn and ash-grey. The horns are long and variable in form. They are generally set wide apart, emerging slightly downward and upward outward with the points recurving inward, creating a crescent shape or semicircle characteristically. The horns at the base are thick. There are two chevron marks, one just around the jowl and the other anterior to the brisket, a thin band of thick hair spanning the top line from the crest of the neck to the point of origin of the tail. With an average fat of 8.22%, the average lactation milk yield is about 500 kg (AGRI-IS, NDDB).

2.3. REASONS FOR LOSS OF ANIMAL GENETIC RESOURCES

India has rich genetic resources in terms of its breeds of buffalo, but the issue is that no demarcated breeding tract is available for breeding either type. There are a number of factors that may enable the genetic assets of animals to weaken.

Murrah

Mehsana

Surti

Bhadawari

Pandharpuri

Toda

Fig. 2.1: Showing different buffalo breeds of india

The common reasons for defeat, however, are:

- Use of certain breeds through artificial insemination and transfer of embryos, resulting in the loss of genetic assets (FAO, 2007).
- Failed agricultural strategies,
- Changing market demands,
- Natural disaster
- Political unrest and conflict
- Deterioration of habitats,
- Migration of high-producing buffaloes from breeding to metropolitan cities to satisfy milk requirements further contribute to genetic degradation of essential germplasm due to slaughter of these animals after lactation has been completed.

Lack of genetic diversity in a community/breed influences its quality by:

- Reduced population fitness,
- Impaired breeding targets,
- Decreased numbers or even breed extinction,
- Easier spread of infectious diseases,
- Endangered the sustainability of rare domestic breeds,
- Limited options available to satisfy uncertain future needs.

There is an immediate need to identify all Indian buffalo breeds uniformly by using specific breed descriptors, observing their natural climate, management practices, qualitative and quantitative aspects of morphological, physiological and functional features, blood groups and biochemical polymorphisms, cytogenetic criteria, DNA analyzes, usefulness and demographic and geographic distributions. This will contribute to the detection of gene forms and gene variations of different breeds and will also help devise breeding strategies and animal selection systems for preservation, reproduction and enhancement.

With regular and galloping advancement in NGS and related technologies, a huge help is promised to scientists as genome-wide assays have become a practice now. For genome-wide characterization of different species for genetic diversity and population structure studies, this allows cost-effective. A vast amount of data can now be produced to classify the genetic

diversity of the community of different species of animals. Genomic markers have been used to evaluate the depth of genetic variation across multiple species between different animal breeds (Freeman *et al*., 2006; Dadi *et al*., 2008). Widely used genetic markers include microsatellite and SNP variants:

2.3.1 Microsatellites

Microsatellites are simple tandem repetitive DNA sequences varying from 1 to 6 base pairs (bp). The microsatellite markers are often also known as simple sequence repeats (SSR). Mutational rates are immensely high in these types of markers. Microsatellites are part of a strongly polymorphic and insightful class of markers, but their genotyping and scoring were shown to be labor-intensive (Ajmone-Marsan *et al*., 2014). The parameters for genetic diversity are generally high due to high mutation rates. The measurements for heterozygosity easily reach higher than 50% levels.

2.3.2 Single nucleotide polymorphisms

SNPs are single nucleotide substitutes that occur from one base to another in more than one percent of the general population. These are binary markers (biallelic) and have lower variability than multiple allele loci but are the most abundant markers present in animal species. The genome for livestock animals contains several million SNPs (Bovine Genome Sequencing & Analysis Consortium, 2009; Zimin *et al*., 2009).A very small subset of these, though, has been evolved as genetic markers for population genetics, interaction study of disequilibrium, and genome-wide association studies. During the past few decades, SNPs were studied individually or a group of 10-15, in a significant manner of studies to establish some signature associations with an important trait present in different livestock species.

2.4. GENOME WIDE SNPS IDENTIFICATION

Whole genome sequencing was conducted in different species of livestock

Table 2.1: SNP identification in various species by Whole genome sequencing

S. No	Breed	Sample	SNPs	Genome coverage	Depth	Reference
1.	Flekvieh	1 bull	2443637	98%	7.4X	Eck *et al*., 2009
2.	Holstein Friesian	1 bull	6239482	98.3%	14.8X	Zhan *et al*., 2011
3.	Kuchinoshima-Ushi	1 cow	6303790	93%	15.8X	Kwahara-Miki *et al*., 2011
4.	Black Angus	1 bull	3200000	-	21.9X	Stothard *et al*., 2011
5.	Holstein	1 bull	3700000	-	18.6X	
6.	Gir	4 bull	9990733	80-88%	2.8-4.4X	Liao *et al*., 2013
7.	Holstein	1 cow	5923230	-	36.7X	Koks *et al*., 2013
8.	Hanwoo	1 cow	6469804	98.6%	25.5X	Choi *et al*., 2014
9.	JejuHengu	1 cow	6484293	98.5%	29.6X	
10.	Korean Holstein	1 cow	5814990	98.5%	29.5X	
11.	Danish Jutland	1 bull	6812198	98.9%	26.4X	Das *et al*., 2014
12.	Holstein Bull	1 bull	6362988	-	60X	Koks *et al*., 2014
13.	1000 Bull genomes	234 bulls	1600000	-	8.3X	Daetwyler *et al*., 2014

The whole genome has to be sequenced for genome-wide SNP detection, where there will be a lot of uninformative, repeating sequence. It is complex, though, and the assembly of bioinformatics is very challenging.

Whole-genome sequencing for SNP genotyping is technologically impractical, as linkage disequilibrium can be as high as 95%-100% across genetic markers within a gene or genomic area, and the data analysis would pick labeled SNP (Jiang *et al*., 2009). In contrast with the entire genome sequencing techniques of SNP exploration, a 35 fold decrease in costs by RAD- seq approaches is feasible (Davey *et al*., 2011).

2.5. RAD SEQUENCING (RESTRICTION SITE ASSOCIATED DNA SEQUENCING)

RAD sequencing approaches focused on NGS are an option to WGS, allowing concurrent sampling, genotyping and multiplexing. SNP exploration was extensively conducted in cattle, pig, chicken, buffalo, camel, sheep, etc. using RAD seq methods (Table 2.2).

Table 2.2: SNP detection using RAD seq techniques

Species	Method	SNPs	Reference
Cattle	GBS	52748	Donato *et al*., 2013
Buffalo	GBS	49607	Imartino *et al*., 2013
Buffalo	ddRAD	130688	Surya *et al*., 2018
Chicken	RAD	75587	Zhai *et al*., 2014
Pig	GBS	185206	Yang *et al*., 2017
Sheep	GBS	300000	Clarke *et al*., 2016
Camel	GBS	310311	Holl *et al*., 2016

This approach has the benefit of minimizing the repetitive sequences and thus reducing the sequencing and genotyping costs. So in a matter of weeks, we will sequence thousands of samples. There are different methods of sequencing RAD (Table 2.3) that have their own upsides and downsides (Table 2.4).

Table 2.3: RAD sequencing methods

S no.	Sequencing methods	References
1.	Complexity Reduction of Polymorphism Sequencing (CRoPS)	Osrow *et al*., 2007
2.	Reduced Representation Libraries (RRL)	Van-Tassel *et al*., 2008
3	Restriction Site associated DNA Sequencing(RAD seq)	Baird *et al*., 2008
4.	Genotyping by Sequencing (GBS)	Elshire *et al*., 2011
5.	Double-digest RAD sequencing (dd RAD)	Peterson *et al*., 2012

Table 2.4: Comparison of RAD sequencing methods

Methods	RAD	GBS	RRL	DdRAD
Digestion by RE	Yes	Yes	Yes	Yes
Ligation	Yes	Yes	Yes	Yes
Pooling	Yes	Yes	Yes	Yes
Shearing	Yes	No	No	No
Size selection	Yes	No	Yes	Yes
Multiplexing	Yes	No	No	Yes
Remarks	DNA loss	No Size Selection	No Multiplexing	Multiplexing with Size Selection

Donato *et al*. (2013) reported that Genotyping by Sequencing (GBS) method was a novel, efficient, cost-effective method which allowed simultaneous SNP identification and genotyping in multiple individuals. A total of 40,725 SNPs were identified in that study using Restriction digestion based methods with call rate of $\geq$ 90%. He used the GBS approach for genotyping 47 animals containing 7 taurine species and showing U.S. and African cattle breeds. Barcodes can be used in a single lane for multiplexing DNA samples up to 384. The repetitive sequence were discarded using methylation-sensitive restriction enzymes, which in effect reduces the cost of sampling, improves 2-3 fold higher efficiency and also simplifies more downstream data analysis by minimizing the data size of the series. A maximum of 49,607 SNP with value > 5 and 10,335 SNP with quality > 10 was reported by GBS (Reduced Representation Approach) using the water buffalo genome comparison in the analysis of Imartino *et al*. (2013).

2.6. REFERENCE GENOME

For understanding the diversity within a species and for harnessing this diversity to wards the genetic improvement of livestock, a reference assembly and a quality gene annotation are important. A significantly enhanced water buffalo (Bubalus bubalis) assembly, UOA-WB-1 (GCF-003121395.1), has been made available in the past year, replacing the fragmented assembly, UMD-CASPUR-WB-2.0 (GCF-000471725.1). Gene models projected to be more complete on the new assembly which are better supported by experimental evidence and have higher coverage hits than models annotated on the old assembly on UniProtKB/ SwissProt proteins. Furthermore, compared to 1.6 billion for UMD-CASPUR-WB-2.0, the 15 billion RNA-Seq reads used in the UOA-WB-1 annotation provide a richer annotation by increasing the number of genes with alternative splice variants by 40%.

2.7. GENOME ANNOTATION

Genome annotation analyzes a genome's raw sequence and identifies important biological and genomic properties such as mutations, mobile elements, repeat elements, replication, and polymorphisms, which is a challenging aspect of genome sequence.

Patel *et al*. (2017) annotated the SNPs in Indian buffalo, which were collected using the cattle reference genome using a directed sequencing process. A total of 447996 SNPs are

annotated from combined dairy production and fertility trait sets, 540414 SNPs from fertility trait analysis sub-set and 383550 SNPs from milk production information. The number of SNPs in the intronic region was comparatively higher in all the data set (44-55%).

The candidate genes responsible for the production trait, reproduction traits and other Major genes were collected from various literature for annotation of SNPs present in the buffalo genome.

2.8. GENETIC DIVERSITY

In any culture, genetic diversity promotes the mechanism of adjustment to various changes throughout environmental conditions and encourages genetic development (Notter, 1999). Genetic diversity legislation results in decreased inbreeding quality (Barczak *et al*., 2009). Maintaining genetic diversity and knowing the species structure for a particular animal community is important for enhancing breeding strategy(ies). Crossbreeding and selection can have a significant impact on the genetic diversity and demographic composition of various animal species (Visser *et al*., 2016; ligda *et al*., 2009). Studies involving human genetics and epidemiology include researching genetic diversity and social composition of different populations. (Novembre and peter, 2016; Elhaik *et al*., 2014; Brye *et al*., 2010; Shtir *et al*., 2009; Tishkoff and Campbell, 2008; Wang *et al*., 2007). The SNPs as possible markers are commonly used for genetic diversity research in various species of livestock (Periasamy *et al*., 2014; Moradi *et al*., 2012; Kijas *et al*., 2012). Some of the important parameters which are critically studied include Heterozygosity (number of heterozygotes), Runs of homozygosity (ROH), linkage disequilibrium (LD), ancestral effective population size (Ne), estimate of haplotype blocks, principal component analysis, admixture analysis, phylogenetic analysis along with neighbour joining tree, etc.

2.8.1. Runs of homozygosity

A ROH is known as a contiguous length of genotypes that are homozygous. A appropriate length of ROH suggests that the two copies of the chromosome are identical-by-descent (IBD) in this region (Purfield *et al*., 2012). The length and frequency of ROH is helpful in gaining insights into an individual's and the population's history of inbreeding. The

detection of ROH may also assist in identifying the footprints of genetic selection on the genome (Marras *et al*., 2015; Peripolli *et al*., 2017; Purfield *et al*., 2012). ROH, however, is indicative, yet not definitive, of natural or artificially selected genomic regions since several factors other than ROH, such as recombination rate, population structure, mutation rate and inbreeding rate, affect the frequency, degree and distribution of ROH throughout the genome (Peripolli *et al*., 2017).

2.8.2. Linkage disequilibrium

Genetic linkage is the propensity during the meiosis process of sexual reproduction to pass DNA molecules on a chromosome together. T.H. Morgan gave the concept in 1911, though in the early 1900s, Bateson and Punnett obtained 1st research on pea plants. The purpose of the linkage analysis is to extract all available inheritance information from pedigrees and to test the co-heritance of chromosome regions with a trait. Analysis of linkage was the primary genetic mapping tool. Linkage disequilibrium is the measure of non-random allele association at different loci as a function of recombining physical positions in the genome (Barbato *et al*., 2015). Admixture and genetic drift (Wright, 1943; Wang, 2005) or systematic sweeping (Smith *et al*., 1974) or background filtering (Charlesworth *et al*., 1997) by 'Hitchhiking' may also derive LD signatures." LD study leads us to find effective population size (Barbato *et al*., 2015), LD maps and haplotype block structure characterization (Mokry *et al*., 2014; villa-angulo *et al*., 2009) and many others.

2.8.3. Effective population size (N_e)

This is the amount of people who would give rise to the estimated variation in sampling or inbreeding frequency if they were raised in the idealized population fashion (Falconer and Mackay, 1996). An enhanced genetic drift caused by bottlenecks and founding effects is of particular interest when deciphering the potential causes of speciation and diversity (Barton and Charlesworth, 1984). In order to conserve livestock populations and ensure safe production, early detection of population declines is important. Antao *et al*. (2011) tested the output of two N E estimators to predict reductions in the population: the temporal two-sample model and one Linkage Disequillibrium (LD) based sample process. The LD model outperformed

the temporal approach by requiring less extreme population losses to be observed sooner. Methods for estimating Ne from LD was developed-40 years ago, however, due to the lack of significant amounts of information on genetic markers, the study was not quite impactful. With the help of SNP marker knowledge, Ne study can help to identify the underlying cause of genetic variation in this era of next-generation sequencing (Barbato *et al*., 2015). The LD model is reliable and generally applicable to small populations utilizing genome-wide data, both in overlapping and non-overlapping generations (Saura *et al*., 2015). It encourages the use of the tool to predict Ne where pedigree information is available to track and control populations efficiently and to identify early decreases in the population.

2.8.4. Haplotype Block structure

Haplotype block applies to a set of alleles that are inherited from a common ancestor along a chromosome (Mokry *et al*., 2014). For buffalo, there are many reports of haplotype block parameters that differ in many ways (breed of interest, marker shapes, marker density, and chromosome regions), producing average haplotype block scales from a few kb in size, i.e. 5.7 kb taking into account two or more SNPs (Villa-Angulo *et al*., 2009), 26.2 kb taking into account four or more SNPs to hundreds of kb in size (Kim *et al*., 2009). Large LD amounts up to ~100 kb and haplotype block range around 30 bases and 75 kb with an average of 10.3 kb was observed in a cattle genome analysis (Villa-Angulo *et al*., 2009). From the bovine HapMap Consortium's SNP data, Villa-Angulo found new results like haplotype block structure on the scale of 1-100 kb, revealing the similarity between cattle and the human genome, as well as similarities between dairy and beef breeds They often grouped various breeds based on the common haplotype proportion.

2.8.5. Inbreeding coefficient

Inbreeding coefficient (F) is the likelihood that the pair of alleles borne by the gametes that generated it was equivalent by descent (Falconer and Mackay, 1996). Inbreeding in a group contributes to the production of an increased number of homozygotes at the cost of heterozygotes. Because of inbreeding, the deleterious alleles come with homozygous conditions and thus, through the phenomenon of inbreeding depression, affect the characteristics of fitness

and viability. Reducing the characteristics of reproduction and lower inbred animal evolutionary ability may have serious effects on community development (Iacolina *et al*., 2016). To devise successful breeding strategies for a species, it is necessary to determine the rate of inbreeding in a community. Conventionally, the inbreeding coefficient of individuals belonging to a specific population is determined on the basis of pedigree data, using wright formulae, assuming the inheritance is halving in nature. However, with the advent of genome-wide SNP and the compilation of large genotypic results, which can sometimes be unreliable due to incomplete or defective pedigree records, the genomic inbreeding coefficient can be estimated instead of depending on pedigree (F_{PED}). Some notable methods are used, such as run of homozygosity (ROH)-based inbreeding coefficients (F_{ROH}), excess homozygosity-dependent inbreeding coefficients (F_{HOM}), method-based inbreeding coefficient (F_{MOM}), maximum likelihood estimate-based inbreeding coefficients (F_{MLE}), inbreeding matrix-based inbreeding coefficients (F_{GRM}) and comparison of uniting gametes (F_{UNI}).

2.8.5.1. F_{HOM}

Inbreeding dependent on excess homozygosity was estimated on the basis of the calculation of excess homozygosity shown in the Wright (1948) model of genotypical SNP results as; $F_{HOM} = [O_{(HOM)} - E_{(HOM)}]/1 - E_{(HOM)}$

Where O(HOM) and E(HOM) are the amount of homozygous genotypes observed and expected in the study, respectively.

2.8.5.2. $F_{ROH:}$

Homozygosity run (ROH) applies to homozygous genotypes contiguous sizes. Large chunks of homozygous genotypes can be detected with the inclusion of genotypic evidence from the SNP range (Zhang *et al*., 2015). The Runs of Homozygosity (ROH) is believed to be a reliable approximation of genomic autozygosity and to differentiate regions that are identical by descent (MeQuillan *et al*., 2008). ROH-based homozygosity measurements from genomic data are defined as the total length of the genome covered by ROH divided by the total length of the genome covered by SNPs;

$$F_{ROH} = L_{ROH} / L_{AUTO}$$

Where the number of ROH lengths is L_{ROH} and the maximum amount of autosomes provided by reads is L_{AUTO}. ROH-based estimates are not affected by allele frequencies and explicitly reflect homozygosity levels, but it is important to accurately classify short ROH (Zhang *et al*., 2015).

2.8.6 Genetic admixture

When individuals from two or more previously isolated populations interbreed and produce a new viable hybrid animal population, genetic admixture occurs. An admixture of different breeds has been prevalent throughout the history of animal breeding to exploit the phenomenon of complementarity and heterosis of distinct traits. In a population, genetic admixture analysis allows geneticists to classify individuals from a population into separate groups based on certain genome-wide markers that are eventually linked to certain biological entities.

These levels of admixture can be studied at different levels from individual to population, i.e. population, individual or chromosome-length specific regions (locus level). We can determine the genetic composition of different breeds with the use of genomic tools and analyze if they are purebred or graded or crossbred (Anderson, 2008). The advent of these breeds can also be historically and geographically identified by recognizing their recent or distant admixture (Larmer *et al*., 2014). Divergence and mixing can also be evaluated through genomic studies there.

2.8.6.1 Approaches to the estimation of admixtures by using bioinformatics and statistical methods

The goal of global ancestry is to estimate the ancestral contribution contributed by each constituent population in crossbred populations. The study of the existence of marker variants across the entire genome is based on these estimates. There are two major approaches for the detection of admixture levels in any population, i.e. model based and non-parametric method based approach (Shriner, 2013).

2.8.6.2 Model-based approach

The approaches based on the model, detect the chunks of chromosomes and segments that remain unbroken from the base population along the ancestry after their origin. Quantification of the proportions of different breeds in the present population is possible with the aid of the recognition of these chunks (Falush *et al*., 2003). Based on various underlying models, these models aim to estimate the ancestry of admixed populations. STRUCTURE (Pritchard *et al*., 2000) and ADMIXTURE (Alexander *et al*., 2009) are the most widely used bioinformatics methods. Both of these programmes operate on the model using the proportions of ancestry and population allele frequencies from genotypic data assuming the presence of Hardy-Weinberg Equilibrium and linkage disequilibrium among these loci (Liu *et al*., 2013; Skotte *et al*., 2013).

2.8.6.2.1 Admixture

ADMIXTURE is a model-based ancestry estimation programme from large autosomal SNP genotype datasets in a model-based manner, where individuals are unrelated (for example, individuals in a case-control association study). ADMIXTURE's input is binary PLINK (.bed), ordinary PLINK (.ped), or EIGENSTRAT (.geno) formatted files and its output is simple space-delimited files containing estimates of parameters.

You need an input file and an idea of K or your assumption about the number of ancestral populations, to use ADMIXTURE.

2.8.6.3 Non-parametric approaches/ tests

Non-parametric tests do not need the modality of the data for inference of population structure. Different elements of multivariate statistical analysis are used by non-parametric experiments. The cluster analysis and the Principal component analysis (PCA) primarily include these methods. These methods aim to stratify the population on the basis of the linearity of multi-dimensional variability found in the genotypic data. These methods help to directly infer the presence of subsets in the genotypical data comprising various populations/breeds.

2.8.6.3.1 Cluster analysis

One of the key parts of non-parametric tests performed on genome-wide SNP data is the clustering of individuals from the population into different clusters based on their respective

allele frequencies at different SNP variant loci. The ultimate objective of cluster analysis is to find data subsets and thus populations in the dataset representing various population groups (Bouaziz *et al*., 2011).

2.8.6.3.2 Principal component analysis

Principal Component Analysis (PCA) refers to a methodology that seeks to reduce complex datasets in a linear dimension. For the original vector or associated variables, this transformation of dimensionality is carried out to a vector of uncorrelated main components. The major part of the variation between populations and between individuals is determined by these Principal components (Menozzi *et al*., 1978). PCA helps to evaluate various key components in population genetic studies for a set of markers in different populations of livestock. PCA also helps to differentiate populations with various components that clarify substantial parts of the existing variation. In population studies, PCA is a commonly used statistical method for genetic structure analysis. PCA is now routinely applied to the processed data from different animal species with the advent of high density genome wide SNP marker panels. This is done specifically in order to detect and measure the genetic makeup of livestock populations.

Most water buffalo populations have shown a steady reduction in population sizes in recent decades (Borghese, 2011), which is generally correlated with biodiversity loss. In recent years, the use of standardised single nucleotide polymorphism (SNP) marker panels for major livestock species has been particularly useful in studying the genomic diversity of farm animals both globally (Kijas *et al*., 2012; Decker *et al*., 2014) and locally (Ciani *et al*., 2014; Nicoloso *et al*., 2015), allowing post-domestication evolution research to be carried out.

Materials and Methods

3 MATERIALS AND METHODS

3.1 Blood collection

About 10 ml of blood was collected from total 96 animals of 6 different breeds of buffalo (30 Murrah, 15 Pandharpuri, 15 Bhadawari, 15 Mehsana, 15 Surti and 6 Toda buffaloes) aseptically in a sterile vial containing 0.5% EDTA from Jugular vein puncture. The blood samples were stored in the freezer at -20°C before the DNA isolation.

3.2 Extraction of genomic DNA from blood samples

The fundamental success of all molecular biology experiments is the quality and quantity of DNA being used. The DNA collected for subsequent use must therefore be devoid of proteins and other inhibitors. DNA may be isolated from fresh or frozen whole blood, blood stains, sperm cells, hair tissue, bone and other biological specimens. As per the method defined by Sambrook *et al*. (1989), genomic DNA was isolated with minor modifications. The phenol-chloroform extraction accompanied by ethanol precipitation for DNA isolation is most widely followed. Almost all DNA isolation protocols from mammalian tissues or any other material, such as blood, hair, etc., involve four main steps:

- Cell lysis using SDS detergent.
- Digestion of cell lysis-released proteins with proteinase k
- DNA extraction with phenol.
- DNA sodium acetate precipitation with alcohol.

3.2.1 DNA isolation Procedure

- DNA blood samples were thawed before performing the isolation by storing them in an ice-containing box and taking them to room temperature.
- 2 volumes of lysis buffer were added to 5 ml of blood and blended well after that these tubes were kept for 8 to 10 minutes in ice for incubation.
- The samples were centrifuged for 10 min @ 12000 rpm at the temperature of 40°C, the supernatant was discarded and the cell pellet was suspended in a lysis buffer and incubated for 8 to 10 minutes in ice.
- Again, the samples at 12000 rpm were centrifuged at 4^0C for 10 minutes. The supernatant had been discarded.
- In 5 ml of extraction buffer, 25 µl of Proteinase-k and 125 µl of 0.5% SDS, the pellet was dissolved and the tubes were incubated overnight at 56^0C.
- After overnight incubation equal volume of phenol (pH 8.0) (normal pH of phenol-4.0, Tris Cl is added to adjust pH 8.0) was added to the samples and mixed gently.
- These tubes were centrifuged at 12,000 rpm for 10 min at 25^0C.
- After centrifugation, with the aid of the Pasteur pipette, the upper aqueous layer was collected without disrupting the organic layer and the aqueous solution was transferred to a fresh tube.
- Phenol: chloroform: isoamyl alcohol (25:24:1) was added to this aqueous layer and the solution was gently blended. These samples were centrifuged at 12,000 rpm at 25^0C for 10 minutes.
- Without upsetting the organic layer, the upper aqueous layer was collected and then chloroform: isoamyl alcohol (24:1) had been added in equal quantity to the aqueous layer and blended gently.
- These tubes were centrifuged at 12,000 rpm at 25°C for 10 min, collecting the upper aqueous layer.

3.2.2 DNA precipitation:

- 500µl of 3 M sodium acetate and 2 volumes of chilled ethanol were added to this aqueous layer. Tubes have been inverted 3 to 4 times. At this stage, DNA will be precipitated.

- DNA with 70% ethanol was spooled and washed twice..
- The tubes were dried at room temperature until the alcohol evaporated.
- A suitable amount of TE buffer (500 µl) was applied to this DNA precipitate..
- The DNA pellet was dissolved for nuclease inactivation and at 65^0c/30 min, the tubes were incubated.
- Isolated DNA samples on agarose gels were checked.

3.2.3 Quality of DNA

The quality of the genomic DNA was checked by electrophoresis of the agarose gel. For this purpose, 0.8% agarose (0.24 gm) was mixed in a 1X TAE buffer of 30 ml and heated until the agarose was completely melted and dissolved, to obtain a transparent solution. Afterwards, the solution was cooled to room temperature. At the rate of 1.7 µl per 30 ml of gel, ethidium bromide (10 mg / ml stock solution) was added and mixed thoroughly.

The gel was then poured into the casting tray and allowed to solidify for at least 30 minutes before the gel became pure milky white in colour. After proper solidification, the gel was removed and submerged in the buffer tank, which also contained 1X TAE buffer.

2 µl of DNA was mixed with 2 µl of 6X loading dye for loading the samples, and placed into the wells. At 50 volts for one hour, electrophoresis was performed. Under the UV transilluminator, the gel was then visualised. Good quality DNA samples were devoid of shearing and smearing, which suggested low quality.

3.2.4 Quantity of DNA

Using the factor 50 LP 0.2 mm lid, the quantity of DNA was determined by the digital nanophotometer. In each sample, optical density (OD) was also calculated as the ratio of OD260 to OD280.

For ddRAD library preparation, samples meeting the necessary QC parameters were considered.

3.3 Library preparation

Using appropriate restriction enzymes (Sph I & MluC I), double digestion of Genomic DNA (1 microgram) was performed. The digested product was cleaned-up using Ampure beads. Using T4 DNA ligase, ligation of P1 (Barcoded) and P2 adaptors was performed, followed by pooling and clean-up of the ligated product. Size selection of the product was done after 2% agarose gel electrophoresis. PCR amplification was carried out to enrich and add specific adapters and flow cell annealing sequences to the Illumina. Final pooling and sequencing in 96 plex was carried out after QC test on bio-analyzer.

3.4 OBJECTIVE 1

3.4.1 De-multiplexing:

The samples will be demultiplexed first to obtain reads for each sample. Up to one mismatch will be allowed to demultiplex the sample data.

3.4.2 Adapter trimming:

The Illumina 5' and 3' adapter sequences will be removed from the reads for final analysis.

3.4.3 Quality control of reads:

NGS data has several sequence artifacts including poor quality reads, primer/adaptor contamination, and base calling errors are quite common which will significantly affect the final result interpretation. Hence, the sequence data generated from these technologies should be of good quality through various quality control (QC) process to proceed further for the downstream analysis.

The reads were quality checked using FastQC. Trimming of Illumina universal adapters and quality filtering was performed by the process_radtags function of the STACKS v2 software. Reads were examined using a sliding window spanning 15% of the read length and the reads which had an average phred score of < 15 were discarded. The first 15 bp of the reads were trimmed using Cutadapt 2.10 as they contained barcoding sequences used for demultiplexing the samples.

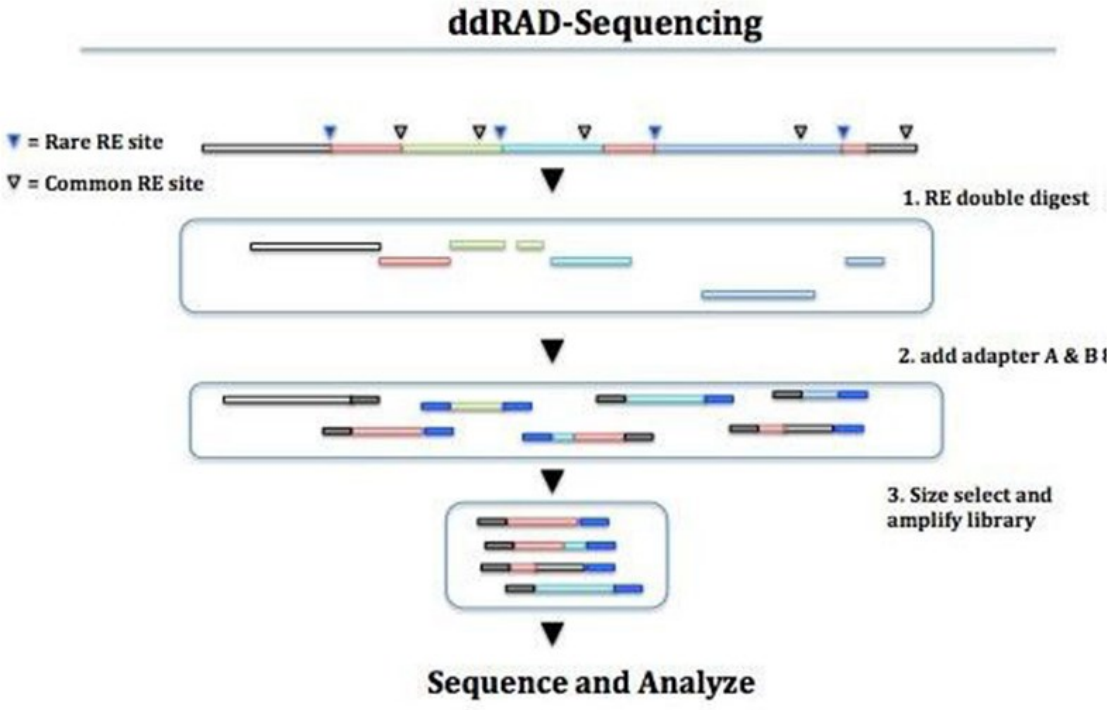


Fig. 3.1: Library Preparation Workflow

3.4.4 Alignment

The paired reads were aligned to the *Bubalus bubalis* assembly UOA_WB_1 using the Burrows-Wheeler Aligner's (BWA) memfunction with default settings.

3.4.5 Variant calling:

The resulted SAM file from the sequence alignment were then converted to BAM file using samtools view followed by Sorting, Indexing and Merging of the reads using Samtools program (Samtools version 1.6) for further downstream analysis.

Samtools program (Samtools version 1.6) was used to call variants from the BAM file which resulted in the output of VCF file.

Variant calling was performed through the bcftools mpileup utility of the Samtools suite in a multi-sample mode. The minimum base (-m) and mapping quality (-M) were set to 30 as recommended by Wright *et al.* (2019). The mapping quality of the reads containing excessive mismatches was downgraded by a coefficient of 50 (-C) which is recommended for bwa alignmemnts. The flag -E was used to recalculate the Base Alignment Quality score to reduce false SNP calls.

The output of the mpileup utility was piped to bcftools call function. Multi-allelic caller was used through the flag -m and only the variant sites were called using the flag -v. The raw variant calls were output as a bcf file:

bcftools mpileup -f <reference genome> -b <bam files> -q 30 -Q 30 -C 50 -Ou -E | bcftools call -mv -Ob -o raw.bcf

The raw variants with quality score greater than 30 and a read depth of 10 were retained using the function bcftools filter and the output VCF file was obtained:

bcftools filter -i '%QUAL>30 && DP > 10'raw.bcf -o filtered.vcf.gz

Bioinformatics workflow for ddRAD sequence analysis

De-multiplexing
↓
Enzyme cut site overhang trimming
↓
Quality and base distribution trimming
↓
Adapter sequence trimming
↓
Denovo sequence assembly
↓
Sequence alignment
↓
Variant calling

3.5 OBJECTIVE 2

3.5.1 Annotation

SnpEff v4.3T was used for structural and functional annotation of the discovered variants. First, a reference database was built for annotation using the UOA_WB_1 genome sequence and its corresponding gene annotation file (.GTF) from NCBI. This database was used to annotate the combined and breed-wise SNP sets using default settings.

Annotation workflow

Collection of genes from literature
↓
Finding the location of gene from NCBI

For assigning the variants to each gene, the genes involved in different traits were downloaded from CattleQTLdb. The Animal Quantitative Trait Loci (QTL) Database (Animal QTLdb) strives to collect all publicly available trait mapping data, *i.e.* QTL (phenotype/ expression, eQTL), candidate gene and association data (GWAS), and copy number variations (CNV) mapped to livestock animal genomes, in order to facilitate locating and comparing discoveries within and between species. New data and database tools are continually developed to align various trait mapping data to map-based genome features such as annotated genes.

3.6. OBJECTIVE 3 - GENETIC DIVERSITY ANALYSIS

3.6.1. Data preparation

The annotated variants were subjected to further filters before diversity analysis. Non-autosomal and unmapped SNPs were removed. SNPs missing in more than 25% individuals and below the MAF threshold of 0.01 were also filtered out. Indels were removed.

(Plink command: —maf 0.01 —geno 0.25 —snps-only —chr-set 24 —autosome)

Genotype imputation has been suggeted for improving the genotype quality of low coverage data. Beagle 4.1 was used for imputation of missing genotypes by applying the genotype likelihood (-gl) input flag, resulting in a dataset containing 237,762 common SNPs across all the breeds.

3.6.1.1 Population genetics analysis.

The parameters estimated to study the genomic diversity in buffalo populations included estimates of heterozygosity and detection of runs of homozygosity (ROH).

The observed (Ho) and expected (He) heterozygosities, for buffalo populations were estimated using PLINK1.9 sofware (Purcell *et al*., 2007). The F_{ST} was estimated using PLINK sofware, which uses method introduced by Weir and Cockerham (1984).

3.6.2. Detection of autozygosity in genomic region

To assess the degree of autozygosity in the buffalo populations, homozygosity (ROH) runs were observed. A ROH is known as a contiguous length of genotypes of homozygotes (Gibson *et al*., 2006). A appropriate length of ROH suggests that the two copies of the chromosome are identical-by-descent (IBD) in this region (Purfield *et al*., 2012). The length and frequency of ROH is helpful in gaining insight into the history of an individual's inbreeding and the population.

The combined dataset of the 96 animals was used for estimation of ROH using the –homozyg function of PLINK v1.9. The best-practices guidelines for ROH estimation in PLINK as specified by Meyermans *et al.* (2020) were followed. No further MAF or LD filtering was

applied. The minimum ROH length (L) was calculated using the method proposed by Lencz *et al.* (2007).

$$L = \frac{\log_e \frac{\alpha}{n_s n_i}}{\log_e (1 - het)}$$

where α is the false positive threshold (0.05), n_s is the number of markers, n_i is the number of individuals, and het is the average heterozygosity over all loci. To be considered as ROH, a genome segment had to consist of minimum 70 SNP and be 1000 kb in length. The SNP density was 30 kb/SNP. 5 missing and 3 heterozygous genotypes were allowed in the scanning window as suggested by Ceballos *et al.* (2018) for low coverage WGS data. All other settings were set to default. A completely homozygous individual was simulated to ensure that the above settings covered the entire genome length.

plink —bfile plink.imputed —homozyg —cow —homozyg-window-het 3 —homozyg-kb 1000 —homozyg-snp 70 —homozyg-density 30 —homozyg-window-missing 5

In the present study, genomic autozygosity, F_{ROH} of each individual was estimated as the sum of length of autosomal ROH divided by the total length of the autosomes covered by markers (McQuillan *et al*., 2008).

$$F_{ROH} = \frac{\sum_{j-1}^{n} L_{ROHi}}{L_{total}}$$

where L_{ROHj} is the length of ROH_j, and L_{total} is the total size of the autosomes covered by markers.

For each animal F_{ROH} (F_{ROH}–2 Mb, F_{ROH} 2–4 Mb, F_{ROH} 4–8 Mb, F_{ROH} 8–16 Mb, and F_{ROH} > 16 Mb) was calculated based on ROH distribution of five minimum different lengths (ROHj): 1–2, 2–4, 4–8, 8–16, and >16 Mb, respectively.

Inbreeding based on the observed versus expected number of homozygous genotypes (F_{HOM}) was calculated using the –het flag on the same dataset in PLINK v1.90 by computing observed and expected autosomal homozygous genotypes counts for each sample, as follows:

$$F_{HOM} = \frac{\text{Observed hom count} - \text{Expected count}}{\text{Total observation} - \text{Expected count}}$$

Pearson correlation coefficient was used to compare the two methods.

3.6.3. Extent of Linkage Disequilibrium

For each breed, the pairwise LD for all pairs of SNP within a 500Kb sliding window **(-ld-window-kb 500)** on each autosomal chromosome was calculated using the method specified by Hill and Robertson (1968), as implemented in PLINK 1.9.

The LD between two SNPs was evaluated using the statistics r^2 and the absolute D-value (|D'|) which were calculated as follows:

$$r^2 = \frac{D^2}{f(A)f(a)f(B)f(b)}$$

Where,

$$D = \text{freq AB} - \text{freq A} * \text{freq B}$$

And,

$$D' = \begin{cases} \dfrac{D}{\min(\text{freq A} * \text{freq b}, \text{freq B} * \text{freq a})} & \text{if } D > 0 \\ \dfrac{D}{\min(\text{freq A} * \text{freq B} * \text{freq a} * \text{freq b})} & \text{if } D < 0 \end{cases}$$

Where SNP pairs had alleles *A* and *a* at the first locus and *B* and *b* at the second locus, *freq A*, *freq a*, *freq B and freq.b* denote frequencies of alleles *A*, *a*, *B*, and *b*, respectively, and *freq AB* denote frequency of haplotype *AB* in the population. The r^2 and |D'| were calculated between adjacent markers and SNP pairs with physical distances from 0 to 15 Mb for each population (Sargolzaei M, University of Guelph, Canada).

The interpretation of r^2 in terms of the power to detect an association leads to the concept of useful LD (Kruglyak, 1999). Sample size is usually limiting in association studies and large large increases in sample size to compensate for weak LD between a marker and the susceptibility locus are impractical.

The min R2 (-ld-window-r2) setting was set to 0. Alleles below a MAF of< 0.01 and significantly ($p < 0.0001$) deviating from the Hardy Weinberg Equilibrium were filtered out. The 500 Kb estimation window was divided into 10 kb bins and average r^2 for each bin was used to plot LD decay.

3.6.4. Ancestral effective population size

The size of a hypothetical ideal population that will result in the same amount of genetic drift as in the (actual) population (Wright, 1931) can be defined as the effective population size (Ne) of a real population X. It is an important population parameter that enables us to understand the evolution of populations (Falconer and Mackay, 1996) and can be used to enhance the understanding and modeling of the architecture underlying complex genetic traits (Hayes *et al*., 2003).

The effective population size (Ne) for each generation was estimated on the basis of the average linkage disequilibrium (r^2) used to estimate patterns of ancestral effective population size using SNeP 1.1 for each group. By using the formula defined by Sved,

$$E(r^2) = \frac{1}{1 + 4N_eC}$$

Where, c is the distance in Morgan between the SNPs and T is equal to 1/2c which represents the age of Ne.

In the absence of a linkage map for Buffalo genome, the genetic distance was approximated as 1 cM ~ 1 Mb and (Sved and Feldman, 1973) recombination rate modifier was applied. The Ne was estimated for different generations using the average of c and r^2 at every 0.10 Mbp for distance between 0.05 Mbp and 10 Mbp & 0.5 Mbp for distance between 10 and 20 Mbp. The plots were generated using the R package ggplot2.

3.6.5. Identification of Haplotype blocks and tag SNPs

Haplotype block refers to a combination of alleles linked along a chromosome, which are inherited together from a common ancestor (Hapmap, 2005).

The dataset of individual breeds were phased using Beagle 4.1. Haplotype blocks were identified through the PLINK function –blocks, which is based on Gabriel *et al*. (2002) haplotype block definition, as implemented in Haploview. Since no Haplotype blocks could be identified in Toda at default settings, the confidence interval thresholds were relaxed as mentioned in Seefried, (2016) and Pritchard *et al.* (2018) using the following settings for all the breeds.

-blocks no-pheno-req no-small-max-span —blocks-max-kb 5000 —blocks-recomb-highci 0.8 —blocks-strong-highci 0.905 —blocks-strong-lowci 0.505

For each breed, the Tag SNPs were identified chromosome-wise using the Tagger algorithm in Haploview (Bakker *et al.* 2005). Pairwise-tagging was performed to identify a minimum panel of SNPs which were in strong LD (r^2>= 0.8) with all other SNPs on the chromosome. The results for Haploblock identification and tagging were summarized and plotted using R.

For PCA, Treemix and Admixture analysis, additional filtering for markers in linkage disequilibrium was performed using PLINK. Each chromosome was scanned through a 50 kb sliding window with a step size of 5 variants. SNPs above r^2> 0.2 were removed, resulting in a dataset of 67,798 SNPs.

(Plink command –indep-pairwise 50 5 0.2).

3.6.6. PCA

The population structure was categorized by: (1) phylogenetic distance-based analysis, (2) model-based clustering, and (3) principal component analysis (PCA) that is assumption-free. The -indep command in PLINK was used to prune the SNPs that passed the initial filtering stage in order to ensure that studies would not be skewed by the existence of SNPs in a strong linkage disequilibrium (LD).

First, a genomic relationship matrix was prepared (-make-grm) by GCTA v1.93 using the LD pruned dataset. 10 principal components were extracted from the matrix using –pca 10, and the eigenvectors of the first 3 PC were plotted in R and two dimensional graphical representations of the dataset were visualized in graphical output. The graphical output was designated by corresponding breed/population names after confirming with the corresponding coordinates of the plot.

3.6.7 Admixture analysis

A model-based population structure estimation with the software ADMIXTURE ver. 1.3 (Alexander *et al*., 2009) was obtained via a maximum likelihood method. Unlike the

EigenStrat format that we used for Principal Component Analysis, ADMIXTURE includes PLINK bed format input data. Fortunately, using Eigensoft's tool convertf, we can simply convert Eigenstrat to bed. Now we can run Admixture. The basic syntax is dead-easy: admixture $INPUT.bed $K. Here, $K is a number that indicates the number of clusters you want to infer by. ADMIXTURE 1.3 was run on the LD pruned dataset for K values ranging from 2 – 6. Each run was repeated 10 times with a random seed to gain accurate estimates. The results were visualized using PONG.

3.6.7.1 Choosing the correct value for K

For it, use ADMIXTURE's cross validation procedures. Compared with other K values, a successful value of K should exhibit a low cross-validation error. By simply adding the —cv flag to the ADMIXTURE command line, cross-validation is enabled. The cross-validation method performs 5-fold CV- in this default setting. So can get 10-fold CV for example, using —cv=10. The cross-validation error is documented in the output.

3.6.8 Phylogenetic analysis

The branching diagrams which showing the evolutionary relationships between organisms are the phylogenetic trees. Typically, such trees are built on the basis of the sequence similarity between the highly conserved 16S rRNA genes or a collection of several organism's housekeeping genes. This constraint on a small set of input sequences can be problematic.

Since the phylogeny of single genes does not necessarily reflect the phylogeny of the complete organisms. Therefore, it is highly beneficial to use all core genome genes as an input for tree computation, which improves its reliability significantly (Gontcharov *et al*., 2004).

The Identity by State matrix (IBS) prepared using PLINK's –distance flag was used to construct the phylogenetic tree including each individual using the UPGMA algorithm in Phylip v3.697. Visuation was done by FigTree v1.1.

PHYLIP is a comprehensive collection of software tools that apply various algorithm to create the Phylogenetic trees. Four of the most notable algorithms are:

- UPGMA
- Neighbour-joining (NJ)

- Maximum parsimony (MP)
- Maximum-likelihood (ML)

UPGMA: UPGMA is the simplest method of the distance matrix which uses sequential methods. Clustering to create a rooted phylogenetic tree. To measure the distance matrix, firstly all sequences are compared by pairwise alignment. The two sequences with the least distance are identified and clustered as a single pair using this matrix. Then, to construct a new matrix, the distance between this pair and all other sequences is recalculated. Using this new matrix, the sequence that is nearest to the first pair is marked and clustered. This process is repeated until all the sequences in the cluster have been incorporated.

Using TreeMix software version 1.12, the frequency of migration events was evaluated (Pickrell and Pritchard, 2012). TreeMix estimates the relationships between the studied populations, models a user-defined number of migrations (mi) within the tree-like graph by relying on a drift-based evolutionary model, and estimates the proportion of admixture exhibited by the receiving groups. So the Treemix software was used to make a phylogram by maximum-likelihood method to infer the ancestral relationships and migration patterns between the breeds.

Results

4 RESULTS

4.1 Identification of SNPs

4.1.1 Raw reads:

Total 397.8 million paired end reads of 150 bp length were generated by the sequencer, averaging 2.07 million reads per sample. The libraries were constructed based on restriction digestion methods using Sph I & MluC I restriction enzymes.

4.1.2 Quality control of raw reads:

After the quality control steps, a total of 367.2 million reads (92.3% of total reads) of 135 bp length were used for downstream processing (Table 4.1) and the remaining poor quality reads were discarded.

4.1.3 Alignment:

The paired reads were aligned to the *Bubalus bubalis* assembly UOA_WB_1 using the Burrows-Wheeler Aligner's memfunction (BWA-MEM) with default settings. 99.82% of the reads aligned to the reference genome.

4.1.4 Variant calling:

The variants identified using SAMTOOLS software program were quality filtered at read depth of 10 along with phred-like consensus quality score of > 15 and the reads which had an average phred score of < 15 were discarded. In this study, Total 569,535 variants were discovered, out of which 502,476 were SNP and 67,059 were indels. 551,458 variants were present on autosomes, 15315 on the X chromosome and 12 on the mtDNA

(NC_006295.1) and 2750 variants were located on unmapped contigs. A variant was discovered for every 4,637 bp of the genome length.

Table 4.1 Variants discovered per chromosome (all breeds)

S. no.	Chromosome_code	Chr. Name	Length (bp)	Discovered variants	Bp_per_ variant
1.	NC_006295.1	Mitochondria	16359	12	1363
2.	NC_037545.1	1	202105980	45310	4460
3.	NC_037546.1	2	188952477	39315	4806
4.	NC_037547.1	3	175630833	38378	4576
5.	NC_037548.1	4	165345888	34549	4785
6.	NC_037549.1	5	127681980	28003	4559
7.	NC_037550.1	6	120641659	26282	4590
8.	NC_037551.1	7	117219835	26192	4475
9.	NC_037552.1	8	119769348	25544	4688
10.	NC_037553.1	9	110236936	23679	4655
11.	NC_037554.1	10	104521508	22648	4615
12.	NC_037555.1	11	102258935	21968	4656
13.	NC_037556.1	12	106433551	22999	4627
14.	NC_037557.1	13	90503546	21296	4249
15.	NC_037558.1	14	83509883	18118	4609
16.	NC_037559.1	15	82162863	19975	4113
17.	NC_037560.1	16	84651441	19740	4288
18.	NC_037561.1	17	73313739	16455	4455
19.	NC_037562.1	18	65927732	14885	4429
20.	NC_037563.1	19	71842292	18040	3982
21.	NC_037564.1	20	68863396	15933	4322
22.	NC_037565.1	21	60877646	13838	4399
23.	NC_037566.1	22	62062344	15114	4106
24.	NC_037567.1	23	51760919	12621	4101
25.	NC_037568.1	24	42448106	10576	4013
26.	NC_037569.1	X	143533794	15315	9372

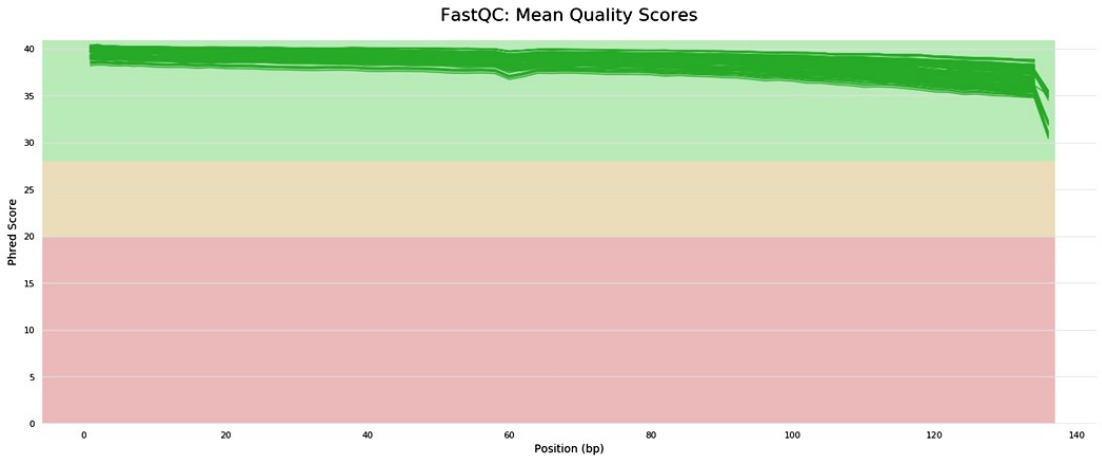
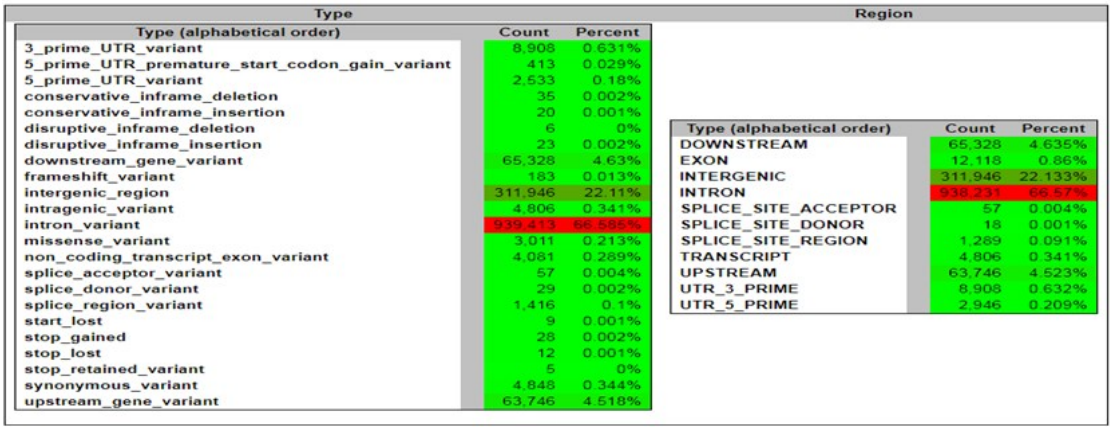


Fig. 4.1: Fastqc report showing mean quality scores for all the samples



Type

Type (alphabetical order)	Count	Percent
3_prime_UTR_variant	8,908	0.631%
5_prime_UTR_premature_start_codon_gain_variant	413	0.029%
5_prime_UTR_variant	2,533	0.18%
conservative_inframe_deletion	35	0.002%
conservative_inframe_insertion	20	0.001%
disruptive_inframe_deletion	6	0%
disruptive_inframe_insertion	23	0.002%
downstream_gene_variant	65,328	4.63%
frameshift_variant	183	0.013%
intergenic_region	311,946	22.11%
intragenic_variant	4,806	0.341%
intron_variant	939,413	66.585%
missense_variant	3,011	0.213%
non_coding_transcript_exon_variant	4,081	0.289%
splice_acceptor_variant	57	0.004%
splice_donor_variant	29	0.002%
splice_region_variant	1,416	0.1%
start_lost	9	0.001%
stop_gained	28	0.002%
stop_lost	12	0.001%
stop_retained_variant	5	0%
synonymous_variant	4,848	0.344%
upstream_gene_variant	63,746	4.518%

Region

Type (alphabetical order)	Count	Percent
DOWNSTREAM	65,328	4.635%
EXON	12,118	0.86%
INTERGENIC	311,946	22.133%
INTRON	938,231	66.57%
SPLICE_SITE_ACCEPTOR	57	0.004%
SPLICE_SITE_DONOR	18	0.001%
SPLICE_SITE_REGION	1,289	0.091%
TRANSCRIPT	4,806	0.341%
UPSTREAM	63,746	4.523%
UTR_3_PRIME	8,908	0.632%
UTR_5_PRIME	2,946	0.209%

Fig. 4.2: Number of effects by type and region

Type (alphabetical order)	Count	Percent
HIGH	302	0.021%
LOW	6,462	0.458%
MODERATE	3,095	0.22%
MODIFIER	1,399,534	99.3%

Table 4.3: Number of effects by impact

Type (alphabetical order)	Count	Percent
MISSENSE	3,031	38.333%
NONSENSE	24	0.304%
SILENT	4,852	61.363%

Table 4.4: Number of effects by functional class
Missense / Silent ratio: 0.6247

	A	C	G	T
A	0	18,320	80,245	13,585
C	22,239	0	19,858	96,873
G	96,833	19,812	0	22,423
T	13,415	80,471	18,402	0

Table 4.5: Base changes (SNPs)

Transitions	11,967,236
Transversions	4,678,732
Ts/Tv ratio	2.5578

Table 4.6: Ts/Tv (transitions / transversions)

Ts/Tv is calculated by using SNPs only. The value of this ratio is more in targeted sequencing methods and less in whole genome sequencing method

	*	-	?	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
*	5				1								1					9	1				
-			64	16			1	3	20			15	2			1		3	19				1
?																							
A		53		610		51	6		5							2		2	39	88	85		
C	5	2			166			9	2									20	11				16
D		5				196	9		21	12					31						9		15
E	2	17		3		50	95		16			53					9				2		
F		4			8			127			1		35						3		1		
G		21		19	2	29	18		337									9	34		29	2	
H		1				1				136					9	18	12	42					35
I		7						5			133	2	5	8	5				6	36	74		
K	1	11					42					108			19		8	21		19			
L		8						34		2	14		418	10		47	5	4	12		30	7	
M		1								6	27	1	5					1		36	22		
N						24				4		8			228				49	2			1
P		50		27						13			31			573	42	5	70	8			
Q	6	11					17			6		5	18			14	126	52					
R	3	1			63				36	94		38	6	1		12	89	215	35	7		37	
S	7	86		59	16			51	29		5		22		32	37		34	554	14			6
T		13		130							46	6		60	1	7		17	24	536			
V		31		48		1	3	10	2		105		21	41				1			168		
W	2	1			2				1				3					11					
Y	2	1			5	1		3		7					2				2				186

Table. 4.7: Changes in Amino acid due to SNPs

How to read this table:

- Rows are Here the reference amino acids are presented in rows and the changed amino acids are in column. E.g. Row ‘E’ column ‘A’ indicates how many ‘E’ amino acids have been replaced by ‘A’ amino acids.
- Grey background color in the table indicate diagonal.

Table 4.2 Breed-wise SNPs, Insertions and Deletions at Read Depth = 10

	Murrah	Bhadawari	Mehsana	Pandharpuri	Surti	Toda
SNPs	484449	473909	489738	469311	470603	448714
INS	26518	25815	26714	25565	25783	24452
DEL	32510	31843	32594	31444	31104	29407

In this study, Highest number of SNPs was found for Mehsana followed by Murrah, while least number of SNPs for Toda breed of Buffalo.

4.2 Annotation

4.2.1 Genome wide annotation of SNPs:

The high quality SNPs identified against Water buffalo reference genome at read depth > 10 and quality > 30 were annotated using the SNPEFF software. Based on the sequence ontology terms, 66.57% SNPs present in intron region, 22.133% SNPs present in intergenic region, and 0.341% SNPs present in transcript region. The highest number of identified SNPs were present in the intron region and the lowest number of SNPs were present in the SPLICE_SITE_DONOR region. Identification of SNPs in the non-coding region may also help to find significant markers (Guo *et al*., 2012).

4.2.2 Gene wise annotation of Variants:

The genes involved in different traits were downloaded from CattleQTLdb. SNPeff results were used to assign the variants to each gene.

4.2.2.1 Gene wise annotation of Variants for the candidate genes responsible for Calving interval.

In this study, total 9 genes are annotated for calving interval.

S. No.	Gene symbol	Gene name Description	Gene Location	Variants impact High	Variants impact Low	Variants impact moderate	Variants impact modifier
1.	GHR	Growth hormone receptor	Chromosome 19 - NC_037563.1	0	2	1	59
2.	GNRHR	Gonadotropin releasing hormone receptor	Chromosome 7 - NC_037551.1	0	0	0	9
3.	LENG8	Leukocyte receptor cluster member 8	Chromosome 18 - NC_037562.1	0	0	0	4
4.	LEP	Leptin	Chromosome 8- NC_037552.1	0	0	0	2
5.	LRWD1	leucine rich repeats and WD repeat domain containing 1	Chromosome 24 - NC_037568.1	0	0	0	2
6.	LTF	Lactotransferrin	Chromosome 21 - NC_037565.1	0	0	0	1
7.	RECQL5	RecQ like helicase 5	Chromosome 3 - NC_037547.1	0	0	0	5
8.	SELP	Selectin P	Chromosome 5- NC_0375549.1	0	0	0	6
9.	TLR4	Toll like receptor 4	Chromosome 3- NC_037547.1	0	0	0	9

4.2.2.2 Gene wise annotation of Variants for the candidate genes responsible for 305 DMY (305 DAILY MILK YIELD).

In this study, total 13 genes are annotated for 305 DMY.

S. No.	Gene symbol	Gene name Description	Gene Location	Variants impact High	Variants impact Low	Variants impact moderate	Variants impact modifier
1.	ABCG2	ATP binding cassette subfamily G member 2 (Junior blood group)	Chromosome 7- NC_037551.1	0	0	0	16
2.	ACACB	acetyl-CoA carboxylase beta	Chromosome 17 - NC_037561.1	0	2	0	36

3.	ARL4A	ADP ribosylation factor like GTPase 4A	Chromosome 8 - NC_037552.1	0	0	0	1
4.	DGAT1	diacylglycerol O-acyltransferase 1	Chromosome 15- NC_037559.1	0	0	0	3
5.	IGF1	insulin like growth factor 1	Chromosome 4 - NC_037548.1	0	0	0	6
6.	LEP	Leptin	Chromosome 8- NC_037552.1	0	0	0	2
7.	LPIN1	lipin 1	Chromosome 12 - NC_037556.1	0	0	0	51
8.	LTF	Lactotransferrin	Chromosome 21 - NC_037565.1	0	0	0	1
9.	MAP4K4	mitogen-activated protein kinase kinase kinase kinase 4	Chromosome 12 - NC_037556.1	0	0	0	59
10.	NPY	neuropeptide Y	Chromosome 8 - NC_037552.1	0	0	0	14
11.	PRLR	prolactin receptor	Chromosome 19 - NC_037563.1	0	0	0	54
12.	SELE	selectin E	Chromosome 5 - NC_037549.1	0	0	0	1
13.	SRC	SRC proto-oncogene, non-receptor tyrosine kinase	Chromosome 14 - NC_037558.1	0	0	0	11

4.2.2.3 Gene wise annotation of Variants for the candidate genes responsible for Age at first calving (AFC).

In this study, total 6 genes are annotated for Age at First calving.

S. No.	Gene symbol	Gene name Description	Gene Location	Variants impact High	Variants impact Low	Variants impact moderate	Variants impact modifier
1.	APP	amyloid beta precursor protein	Chromosome 1 - NC_037545.1	0	0	0	97
2.	ARHGEF3	Rho guanine nucleotide exchange factor 3	Chromosome 21 - NC_037565.1	0	0	0	30

3.	EXOC4	exocyst complex component 4	Chromosome 8 - NC_037552.1	0	0	0	157
4.	LEP	Leptin	Chromosome 8- NC_037552.1	0	0	0	2
5.	SELP	Selectin P	Chromosome 5- NC_0375549.1	0	0	0	6
6.	SETD3	SET domain containing 3, actin histidine methyltransferase	Chromosome 20 - NC_037564.1	0	0	0	4

4.2.2.4 Gene wise annotation of Variants for the candidate genes responsible for Conception Rate.

In this study, total 190 genes are annotated for Conception rate.

S. No.	Gene symbol	Gene name Description	Gene Location	Variants impact High	Variants impact Low	Variants impact moderate	Variants impact modifier
1.	ABCC9	ATP binding cassette subfamily C member 9	Chromosome 4 - NC_037548.1	0	0	0	46
2.	ACAP2	ArfGAP with coiled-coil, ankyrin repeat and PH domains 2	Chromosome 1 - NC_037545.1	0	0	0	22
3.	ACAT2	acetyl-CoA acetyltransferase 2	Chromosome 10 - NC_037554.1	0	0	0	7
4.	ACCSL	1-aminocyclopropane-1-carboxylate synthase homolog (inactive) like	Chromosome 16 - NC_037560.1	0	0	1	28
5.	ACER1	alkaline ceramidase 1	Chromosome 9 - NC_037553.1	0	0	0	18
6.	ADGRE3	adhesion G protein-coupled receptor E3	Chromosome 9 - NC_037553.1	0	0	0	27
7.	AFF1	AF4/FMR2 family member 1	Chromosome 7 - NC_037551.1	0	0	0	57
8.	AGPAT4	1-acylglycerol-3-phosphate O-acyltransferase 4	Chromosome 10 - NC_037554.1	0	0	0	21

9.	AKT2	AKT serine/threonine kinase 2	Chromosome 18 - NC_037562.1	0	0	0	7
10.	AKTIP	AKT interacting protein	Chromosome 18 - NC_037562.1	0	0	0	1
11.	AMN1	antagonist of mitotic exit network 1 homolog	Chromosome 4 - NC_037548.1	0	0	0	14
12	ANKRD17	ankyrin repeat domain 17	Chromosome 7 - NC_037551.1	0	0	0	7
13.	AP3B1	adaptor related protein complex 3 subunit beta 1	Chromosome 11 - NC_037555.1	0	0	0	39
14.	ASB18	ankyrin repeat and SOCS box containing 18	Chromosome 6 - NC_037550.1	0	2	2	61
15.	ATF6	activating transcription factor 6	Chromosome 6 - NC_037550.1	0	0	0	40
16.	AXL	AXL receptor tyrosine kinase	Chromosome 18 - NC_037562.1	0	0	0	5
17.	BCAS1	brain enriched myelin associated protein 1	Chromosome 14 - NC_037558.1	0	0	0	31
18.	BCKDHA	branched chain keto acid dehydrogenase E1 subunit alpha	Chromosome 18 - NC_037562.1	0	0	0	3
19.	BRF1	BRF1 RNA polymerase III transcription initiation factor subunit	Chromosome 6 - NC_037550.1	0	1	0	4
20.	BRINP3	BMP/retinoic acid inducible neural specific 3	Chromosome 5 - NC_037549.1	0	1	0	124
21.	C1QB	complement C1q B chain	Chromosome 2 - NC_037546.1	0	0	0	2
22.	C2CD5	C2 calcium dependent domain containing 5	Chromosome 4 - NC_037548.1	0	0	0	10
23.	CACNA1D	calcium voltage-gated channel subunit alpha1 D	Chromosome 21 - NC_037565.1	0	0	0	62
24.	CACNA1H	calcium voltage-gated channel subunit alpha1 H	Chromosome 24 - NC_037568.1	0	0	0	13

25.	CAMK2D	calcium/calmodulin dependent protein kinase II delta	Chromosome 7 - NC_037551.1	0	0	0	46
26.	CAST	Calpastatin	Chromosome 9 - NC_037553.1	0	0	0	38
27.	CCDC171	coiled-coil domain containing 171	Chromosome 3 - NC_037547.1	0	0	0	39
28.	CCDC86	coiled-coil domain containing 86	Chromosome 5 - NC_037549.1	0	0	0	5
29.	CCT8	chaperonin containing TCP1 subunit 8	Chromosome 1 - NC_037545.1	0	0	0	7
30.	CD109	CD109 molecule	Chromosome 10 - NC_037554.1	0	0	1	31
31.	CD40	CD40 molecule	Chromosome 14 - NC_037558.1	0	0	0	13
32.	CDH23	cadherin related 23	Chromosome 4 - NC_037548.1	0	1	0	95
33.	CENPC	centromere protein C	Chromosome 7 - NC_037551.1	0	0	0	11
34.	CEP128	centrosomal protein 128	Chromosome 11 - NC_037555.1	0	0	0	60
35.	CEP152	centrosomal protein 152	Chromosome 11 - NC_037555.1	0	0	0	17
36.	CHD9	chromodomain helicase DNA binding protein	Chromosome 18 - NC_037562.1	0	0	0	16
37.	CHI3L2	chitinase 3 like 2	Chromosome 6 - NC_037550.1	0	0	0	2
38.	CHST8	carbohydrate sulfotransferase 8	Chromosome 18 - NC_037562.1	0	0	0	45
39.	CIT	citron rho-interacting serine/threonine kinase	Chromosome 17 - NC_037561.1	0	0	0	29
40.	CKAP5	cytoskeleton associated protein 5	Chromosome 16 - NC_037560.1	0	1	0	14
41.	CLASP2	cytoplasmic linker associated protein 2	Chromosome 21 - NC_037565.1	0	0	0	15
42.	COQ9	coenzyme Q9	Chromosome 18 - NC_037562.1	0	0	0	1

43.	CSPP1	centrosome and spindle pole associated protein 1	Chromosome 15 - NC_037559.1	0	0	0	23
44.	CYTH4	cytohesin 4	Chromosome 4 - NC_037548.1	0	0	0	12
45.	DCK	deoxycytidine kinase	Chromosome 7 - NC_037551.1	0	0	0	3
46.	DCP1A	decapping mRNA 1A	Chromosome 21 - NC_037565.1	0	0	0	21
47.	DDX55	DEAD-box helicase 55	Chromosome 17 - NC_037561.1	0	0	0	6
48.	DEPDC5	DEP domain containing 5, GATOR1 subcomplex subunit	Chromosome 17 - NC_037561.1	0	0	0	12
49.	DERA	deoxyribose-phosphate aldolase	Chromosome 4 - NC_037548.1	0	0	0	20
50.	DGAT1	diacylglycerol O-acyltransferase 1	Chromosome 15 - NC_037559.1	0	0	0	3
51.	DHX57	DExH-box helicase 57	Chromosome 12 - NC_037556.1	0	0	0	11
52.	DHX9	DExH-box helicase 9	Chromosome 5 - NC_037549.1	0	0	0	9
53.	DIP2B	disco interacting protein 2 homolog B	Chromosome 4 - NC_037548.1	0	1	0	42
54.	DNMBP	dynamin binding protein	Chromosome 23 - NC_037567.1	0	0	0	20
55.	DYNC1I2	dynein cytoplasmic 1 intermediate chain 2	Chromosome 2 - NC_037546.1	0	0	0	4
56.	DZIP3	DAZ interacting zinc finger protein 3	Chromosome 1 - NC_037545.1	0	0	0	11
57.	EFCAB3	EF-hand calcium binding domain 3	Chromosome 3 - NC_037547.1	0	0	0	17
58.	ELP4	elongator acetyltransferase complex subunit 4	Chromosome 16 - NC_037560.1	0	0	0	22
59.	EPAS1	endothelial PAS domain protein 1	Chromosome 12 - NC_037556.1	0	0	0	5
60.	EPB41L5	erythrocyte membrane protein band 4.1 like 5	Chromosome 2 - NC_037546.1	0	0	0	13
61.	EPHA5	EPH receptor A5	Chromosome 7 - NC_037551.1	0	0	0	63

62.	EXOC2	exocyst complex component 2	Chromosome 2 - NC_037546.1	0	0	0	11
63.	EXOC6B	exocyst complex component 6B	Chromosome 12 - NC_037556.1	0	0	0	52
64.	EYA4	EYA transcriptional coactivator and phosphatase 4	Chromosome 10 - NC_037554.1	0	0	0	115
65.	FAR2	fatty acyl-CoA reductase 2	Chromosome 4 - NC_037548.1	0	0	0	27
66.	FETUB	fetuin B	Chromosome 1 - NC_037545.1	0	0	0	3
67.	FGD4	FYVE, RhoGEF and pH domain containing 4	Chromosome 4 - NC_037548.1	0	0	0	20
68.	FNIP2	folliculin interacting protein 2	Chromosome 17 - NC_037561.1	0	0	0	21
69.	GADD45GIP1	GADD45G interacting protein 1	Chromosome 9 - NC_037553.1	0	0	0	3
70.	GALK2	galactokinase 2	Chromosome 11 - NC_037555.1	0	0	0	17
71.	GBX1	gastrulation brain homeobox 1	Chromosome 8 - NC_037552.1	0	0	0	10
72.	GC	GC vitamin D binding protein	Chromosome 7 - NC_037556.1	0	0	0	41
73.	GCNT3	glucosaminyl (N-acetyl) transferase 3, mucin type	Chromosome 11 - NC_037555.1	0	0	0	6
74.	GHR	Growth hormone receptor	Chromosome 19 - NC_037563.1	0	2	1	59
75.	GMPR	guanosine monophosphate reductase	Chromosome 2 - NC_037546.1	0	0	0	14
76.	GOLGA4	golgin A4	Chromosome 21 - NC_037565.1	0	0	0	18
77.	GPLD1	glycosylphosphatidy-linositol specific phospholipase D1	Chromosome 2 - NC_037546.1	0	0	0	20
78.	GRIA4	glutamate ionotropic receptor AMPA type subunit 4	Chromosome 16 - NC_037560.1	0	0	0	112
79.	HIP1	huntingtin interacting protein 1	Chromosome 24 - NC_037568.1	0	0	0	22

80.	HSD17B7	hydroxysteroid 17-β dehydrogenase 7	Chromosome 6 - NC_037550.1	0	0	1	40
81.	HSD17B12	hydroxysteroid 17-β dehydrogenase 12	Chromosome 16 - NC_037560.1	0	0	0	14
82.	IBSP	integrin binding sialoprotein	Chromosome 7 - NC_037551.1	0	0	0	6
83.	IGF2BP2	insulin like growth factor 2 mRNA binding protein 2	Chromosome 1 - NC_037545.1	0	0	0	16
84.	IL20RA	interleukin 20 receptor subunit alpha	Chromosome 10 - NC_037554.1	0	0	0	11
85.	ILDR1	immunoglobulin like domain containing receptor 1	Chromosome 1 - NC_037545.1	0	0	0	5
86.	IQGAP1	IQ motif containing GTPase activating protein 1	Chromosome 20 - NC_037564.1	0	0	0	17
87.	ITGB5	integrin subunit beta 5	Chromosome 1 - NC_037545.1	0	0	0	9
88.	ITPR2	inositol 1,4,5-tris-phosphate receptor type 2	Chromosome 4 - NC_037548.1	0	3	0	134
89.	KALRN	kalirin RhoGEF kinase	Chromosome 1 - NC_037545.1	0	0	0	76
90.	KAT8	lysine acetyltransferase 8	Chromosome 24 - NC_037568.1	0	0	0	1
91.	KCNIP4	potassium voltage-gated channel interacting protein 4	Chromosome 7 - NC_037551.1	0	0	0	182
92.	KCNMB2	potassium calcium-activated channel subfamily M regulatory beta subunit 2	Chromosome 1 - NC_037545.1	0	0	0	55
93.	KCNN2	potassium calcium-activated channel subfamily N member 2	Chromosome 11 - NC_037555.1	0	0	0	90
94.	KLHL1	kelch like family member 1	Chromosome 13 - NC_037557.1	0	0	0	70
95.	KPNA3	karyopherin subunit alpha 3	Chromosome 13 - NC_037557.1	0	0	0	7

96. LDB3	LIM domain binding 3	Chromosome 4 - NC_037548.1	0	1	0	4
97. LPIN2	lipin 2	Chromosome 22 - NC_037566.1	0	0	0	16
98. LRBA	LPS responsive beige-like anchor protein	Chromosome 17 - NC_037561.1	0	0	0	11
99. LSG1	large 60S subunit nuclear export GTPase 1	Chromosome 1 - NC_037545.1	0	0	0	91
100. MAP1B	microtubule associated protein 1B	Chromosome 19 - NC_037563.1	0	0	0	14
101. KPNA6	karyopherin subunit alpha 6	Chromosome 2 - NC_037546.1	0	0	0	3
102. MAP3K3	mitogen-activated protein kinase kinase kinase 3	Chromosome 3 - NC_037547.1	0	0	0	3
103. MAP6	microtubule associated protein 6	Chromosome 16 - NC_037560.1	0	0	0	20
104. MAPKAP1	MAPK associated protein 1	Chromosome 12 - NC_037556.1	0	0	0	35
105. MCF2L2	MCF.2 cell line derived transforming sequence-like 2	Chromosome 1 - NC_037545.1	0	0	0	59
106. MOG	myelin oligodendrocyte glycoprotein	Chromosome 2 - NC_037546.1	0	0	0	14
107. MON1B	MON1 homolog B, secretory trafficking associated	Chromosome 18 - NC_037562.1	0	0	0	10
108. MOXD1	monooxygenase DBH like 1	Chromosome 10 - NC_037554.1	0	0	0	18
109. MRPL48	mitochondrial ribosomal protein L48	Chromosome 16 - NC_037560.1	0	0	0	19
110. MRPS35	mitochondrial ribosomal protein S35	Chromosome 4 - NC_037548.1	0	0	0	14
111. MS4A8	membrane spanning 4-domains A8	Chromosome 5 - NC_037549.1	0	1	0	14
112. MTMR7	myotubularin related protein 7	Chromosome 1 - NC_037545.1	0	1	0	26
113. MUC19	mucin 19, oligomeric	Chromosome 4 - NC_037548.1	0	0	1	23

114. MYEF2	myelin expression factor 2	Chromosome 11 - NC_037555.1	0	0	0	4
115. MYO10	myosin X	Chromosome 19 - NC_037563.1	0	0	0	80
116. MYO16	myosin XVI	Chromosome 13 - NC_037557.1	0	2	0	165
117. MYO5A	myosin VA	Chromosome 11 - NC_037555.1	0	0	0	14
118. NBAS	NBAS subunit of NRZ tethering complex	Chromosome 12 - NC_037556.1	0	1	2	91
119. NCAM2	neural cell adhesion molecule 2	Chromosome 1 - NC_037545.1	0	0	0	148
120. NEU3	neuraminidase 3	Chromosome 16 - NC_037560.1	0	0	0	6
121. NOL4	nucleolar protein 4	Chromosome 22 - NC_037566.1	0	1	0	60
122. NTRK2	neurotrophic receptor tyrosine kinase 2	Chromosome 3 - NC_037547.1	0	0	0	80
123. OCLN	Occluding	Chromosome 19 - NC_037563.1	0	0	0	13
124. PAFAH1B1	platelet activating factor acetylhydrolase 1b regulatory subunit	Chromosome 13 - NC_037547.1	0	0	0	1
125. PAK2	p21 (RAC1) activated kinase 2	Chromosome 1 - NC_037545.1	0	0	0	15
126. PARM1	prostate androgen-regulated mucin-like protein 1	Chromosome 7 - NC_037551.1	0	0	1	37
127. PARN	poly(A)-specific ribonuclease	Chromosome 24 – NC_037568.1	0	1	0	57
128. PCCB	propionyl-CoA carboxylase subunit beta	Chromosome 1 - NC_037545.1	0	0	0	11
129. PCED1B	PC-esterase domain containing 1B	Chromosome 4 - NC_037548.1	0	0	0	35
130. PDE10A	phosphodiesterase 10A	Chromosome 10 - NC_037554.1	0	0	0	65
131. PELI2	pellino E3 ubiquitin protein ligase family member 2	Chromosome 11 - NC_037555.1	0	0	0	24

132. PGAP1	post-GPI attachment to proteins inositol deacylase 1	Chromosome 2 - NC_037546.1	0	0	0	10
133. PITRM1	pitrilysin metallo peptidase 1	Chromosome 14 - NC_037558.1	0	1	0	15
134. PIWIL3	piwi like RNA-mediated gene silencing 3	Chromosome 17 - NC_037561.1	0	0	0	9
135. PKHD1	PKHD1 ciliary IPT domain containing fibrocystin/polyductin	Chromosome 2 – NC_037546.1	0	0	0	184
136. PKIA	cAMP-dependent protein kinase inhibitor alpha	Chromosome 15 - NC_037559.1	0	0	0	13
137. PLA2G12B	phospholipase A2 group XIIB	Chromosome 4 - NC_037548.1	0	0	0	12
138. PLA2R1	phospholipase A2 receptor 1	Chromosome 2 - NC_037546.1	0	0	0	27
139. PLCB1	phospholipase C beta 1	Chromosome 14 - NC_037558.1	0	0	0	285
140. PLSCR5	phospholipid scramblase family member 5	Chromosome 1 - NC_037545.1	0	0	0	8
141. PMM2	phosphomannomutase 2	Chromosome 24 - NC_037568.1	0	0	0	3
142. POLD3	DNA polymerase delta 3, accessory subunit	Chromosome 16 - NC_037560.1	0	0	0	6
143. PPFIBP1	PPFIA binding protein 1	Chromosome 4 - NC_037548.1	0	0	0	31
144. PRKG1	protein kinase cGMP-dependent 1	Chromosome 4 – NC_037567.1	0	0	0	453
145. PROP1	PROP paired-like homeobox 1	Chromosome 9 - NC_037553.1	0	0	0	15
146. PTPRD	protein tyrosine phosphatase receptor type D	Chromosome 3 - NC_037547.1	0	0	0	451
147. QKI	QKI, KH domain containing RNA binding	Chromosome 10 - NC_037554.1	0	0	0	39
148. RABEP2	rabaptin, RAB GTPase binding effector protein 2	Chromosome 24 - NC_037568.1	0	0	0	3
149. RABGAP1	RAB GTPase activating protein 1	Chromosome 12 - NC_037556.1	0	0	0	10

150. RECQL5	RecQ like helicase 5	Chromosome 3 - NC_037547.1	0	0	0	5
151. RIMS1	regulating synaptic membrane exocytosis 1	Chromosome 10 - NC_037554.1	0	0	0	116
152. RNF17	ring finger protein 17	Chromosome 13 - NC_037557.1	0	0	0	10
153. ROPN1	rhophilin associated tail protein 1	Chromosome 1 - NC_037545.1	0	0	0	10
154. RPE65	Retinoid isomerohydrolase	Chromosome 6 – NC_037550.1	0	0	0	7
155. SEC14L1	SEC14 like lipid binding 1	Chromosome 3 - NC_037547.1	0	0	0	8
156. SGCD	sarcoglycan delta	Chromosome 9 - NC_037553.1	0	0	0	171
157. SIN3B	SIN3 transcription regulator family member B	Chromosome 9 - NC_037553.1	0	0	0	23
158. SIPA1L3	signal induced proliferation associated 1 like 3	Chromosome 18 - NC_037562.1	0	0	0	49
159. SLAIN2	SLAIN motif family member 2	Chromosome 7 - NC_037551.1	0	0	0	12
160. SLC18A2	solute carrier family 18 member A2	Chromosome 23 - NC_037567.1	0	0	0	10
161. SLC25A48	solute carrier family 25 member 48	Chromosome 9 - NC_037553.1	0	0	0	32
162. SLC4A4	solute carrier family 4 member 4	Chromosome 7 - NC_037551.1	0	0	0	65
163. SLC9A9	solute carrier family 9 member A9	Chromosome 1 - NC_037545.1	0	0	0	205
164. SPATA16	spermatogenesis associated 16	Chromosome 1 - NC_037545.1	0	2	0	79
165. SPOCK1	SPARC (osteonectin), cwcv and kazal like domains proteoglycan 1	Chromosome 9 - NC_037553.1	0	1	0	153
166. STAP1	signal transducing adaptor family member 1	Chromosome 7 - NC_037551.1	0	0	0	7
167. SV2C	synaptic vesicle glycoprotein 2C	Chromosome 11 - NC_037555.1	0	1	0	102
168. SYNE1	spectrin repeat containing nuclear envelope protein 1	Chromosome 10 - NC_037554.1	0	11	4	128

169. TANGO2	transport and golgi organization 2 homolog	Chromosome 17 - NC_037561.1	0	0	0	3
170. TBC1D24	TBC1 domain family member 24	Chromosome 24 - NC_037568.1	0	0	0	4
171. TDRKH	tudor and KH domain containing	Chromosome 6 - NC_037550.1	0	0	0	2
172. TESK2	testis associated actin remodelling kinase 2	Chromosome 6 - NC_037550.1	0	0	0	21
173. TIAM1	TIAM Rac1 associated GEF 1	Chromosome 1 – NC_037545.1	0	1	0	69
174. TMTC1	transmembrane O-mannosyltransferase targeting cadherins 1	Chromosome 4 - NC_037548.1	0	2	2	94
175. TOX3	TOX high mobility group box family member 3	Chromosome 18 - NC_037562.1	0	0	0	28
176. TRPM7	transient receptor potential cation channel subfamily M member 7	Chromosome 11 - NC_037555.1	0	0	0	11
177. TSGA10IP	testis specific 10 interacting protein	Chromosome 5 - NC_037549.1	0	0	0	4
178. UBR1	ubiquitin protein ligase E3 component n-recognin 1	Chromosome 11 – NC_037555.1	0	0	0	11
179. UBXN7	UBX domain protein 7	Chromosome 1 - NC_037545.1	0	0	0	4
180. UPK1B	uroplakin 1B	Chromosome 1 - NC_037545.1	0	0	0	3
181. USH2A	usherin	Chromosome 5 - NC_037549.1	0	0	0	72
182. UTRN	utrophin	Chromosome 10 - NC_037554.1	0	0	1	88
183. VTI1A	vesicle transport through interaction with t-SNAREs 1A	Chromosome 23 - NC_037567.1	0	0	0	41

4.2.2.5 Gene wise annotation of Variants for the candidate genes responsible for Milk Fat Percentage.

In this study, total 219 genes are annotated for Milk fat percentage.

S. No.	Gene symbol	Gene name Description	Gene Location	Variants impact High	Variants impact Low	Variants impact moderate	Variants impact modifier
1.	A2ML1	alpha-2-macroglobulin like 1	Chromosome 4 - NC_037548.1	0	0	0	3
2.	AASDHPP	aminoadipate-semialdehyde dehydrogenase-phosphopantetheinyl transferase	Chromosome 16 - NC_037560.1	0	0	0	3
3.	ABCC9	ATP binding cassette subfamily C member 9	Chromosome 4 - NC_037548.1	0	0	0	46
4.	ABCG2	ATP binding cassette subfamily G member 2 (Junior blood group)	Chromosome 7 - NC_037551.1	0	0	0	16
5.	ACACB	acetyl-CoA carboxylase beta	Chromosome 17 - NC_037561.1	0	2	0	36
6.	ACO2	aconitase 2	Chromosome 4 - NC_037548.1	0	0	0	5
7.	ADAMTS1	ADAM metallopeptidase with thrombospondin type 1 motif 1	Chromosome 1 - NC_037545.1	0	0	0	129
8.	ADGRB1	adhesion G protein-coupled receptor B1	Chromosome 15 - NC_037559.1	0	0	0	4
9.	AGO2	argonaute RISC catalytic component 2	Chromosome 15 - NC_037559.1	0	0	0	10
10.	ANKRD17	ankyrin repeat domain 17	Chromosome 7 - NC_037551.1	0	0	0	7
11.	APBA1	amyloid beta precursor protein binding family A member 1	Chromosome 3 - NC_037547.1	0	0	0	51
12.	APBB2	amyloid beta precursor protein binding family	Chromosome 7 - NC_037551.1	0	0	0	60

		B member 2					
13.	ARHGAP39	Rho GTPase activating protein 39	Chromosome 15 - NC_037559.1	0	0	0	17
14.	ARID5B	AT-rich interaction domain 5B	Chromosome 4 - NC_037548.1	0	0	0	22
15.	ARRB1	arrestin beta 1	Chromosome 16 - NC_037560.1	0	0	0	9
16.	BAIAP2	BAR/IMD domain containing adaptor protein 2	Chromosome 3 - NC_037547.1	0	0	0	11
17.	BCAT1	branched chain amino acid transaminase 1	Chromosome 4 - NC_037548.1	0	0	0	21
18.	BCL2L14	BCL2 like 14	Chromosome 4 - NC_037548.1	0	0	0	4
19.	BRINP3	BMP/retinoic acid inducible neural specific 3	Chromosome 5 - NC_037549.1	0	1	0	124
20.	BTN1A1	butyrophilin subfamily 1 member A1	Chromosome 2 - NC_037546.1	0	0	0	8
21.	BYSL	bystin like	Chromosome 2 - NC_037546.1	0	1	0	5
22.	C1R	complement C1r	Chromosome 4 - NC_037548.1	0	0	0	2
23.	C6	Complement C6	Chromosome 19 - NC_037563.1	0	0	0	20
24.	C7	Complement C7	Chromosome 19 - NC_037563.1	0	0	0	25
25.	CACHD1	cache domain containing 1	Chromosome 6 - NC_037550.1	0	1	0	102
26.	CCDC57	coiled-coil domain containing 57	Chromosome 3 - NC_037547.1	0	0	0	8
27.	CCDC91	coiled-coil domain containing 91	Chromosome 4 - NC_037548.1	0	0	0	40
28.	CCL28	C-C motif chemokine ligand 28	Chromosome 19 - NC_037563.1	0	0	0	3
29.	CCND3	cyclin D3	Chromosome 2 - NC_037546.1	0	0	0	10

30.	CD2	CD2 molecule	Chromosome 6 - NC_037550.1	0	0	0	1
31.	CD4	CD4 molecule	Chromosome 4 - NC_037548.1	0	0	0	7
32.	CDC42BPA	CDC42 binding protein kinase alpha	Chromosome 5 - NC_037549.1	0	0	0	44
33.	CDH2	cadherin 2	Chromosome 22 - NC_037566.1	0	0	0	52
34.	CDKN1A	cyclin dependent kinase inhibitor 1A	Chromosome 23 - NC_037546.1	0	0	0	3
35.	CLEC2B	C-type lectin domain family 2 member B	Chromosome 4 - NC_037548.1	0	0	0	5
36.	CLU	clusterin	Chromosome 3 NC_037547.1	0	0	0	2
37.	CNTN5	contactin 5	Chromosome 16 - NC_037560.1	0	0	0	330
38.	COL22A1	collagen type XXII alpha 1 chain	Chromosome 15 - NC_037559.1	0	0	0	68
39.	CPQ	carboxypeptidase Q	Chromosome 15 - NC_037559.1	0	0	0	166
40.	CRACR2A	calcium release activated channel regulator 2A	Chromosome 4 - NC_037548.1	0	0	0	39
41.	CREBL2	cAMP responsive element binding protein like 2	Chromosome 4 - NC_037548.1	0	0	0	8
42.	CSAD	cysteine sulfinic acid decarboxylase	Chromosome 4 - NC_037548.1	0	0	0	3
43.	CSN1S1	casein alpha s1	Chromosome 7 - NC_037551.1	0	0	0	1
44.	CSN2	casein beta	Chromosome 7 - NC_037551.1	0	0	0	4
45.	CSTF3	cleavage stimulation factor subunit 3	Chromosome 16 - NC_037560.1	0	0	0	12
46.	DAB2	DAB adaptor protein 2	Chromosome 19 - NC_037563.1	0	0	0	12

47.	DDIT3	DNA damage inducible transcript 3	Chromosome 4 - NC_037548.1	0	0	0	2
48.	DENND3	DENN domain containing 3	Chromosome 15 - NC_037559.1	0	0	0	8
49.	DERA	deoxyribose-phosphate aldolase	Chromosome 4 - NC_037548.1	0	0	0	20
50.	DGAT1	diacylglycerol O-acyltransferase 1	Chromosome 15 - NC_037559.1	0	0	0	3
51.	DGKG	diacylglycerol kinase gamma	Chromosome 1 - NC_037545.1	0	2	0	25
52.	DOCK5	dedicator of cytokinesis 5	Chromosome 3 - NC_037547.1	0	1	0	92
53.	DPP10	dipeptidyl peptidase like 10	Chromosome 2 - NC_037546.1	0	0	0	206
54.	EEF1D	eukaryotic translation elongation factor 1 delta	Chromosome 15 - NC_037559.1	0	0	0	1
55.	EFNA1	ephrin A1	Chromosome 6 - NC_037550.1	0	0	0	1
56.	EFR3A	EFR3 homolog A	Chromosome 15 - NC_037559.1	0	0	0	17
57.	EGFLAM	EGF like, fibronectin type III and laminin G domains	Chromosome 19 - NC_037563.1	0	0	0	98
58.	ELOVL6	ELOVL fatty acid elongase 6	Chromosome 7 - NC_037551.1	0	0	0	18
59.	EMB	Embigin	Chromosome 19 - NC_037563.1	0	0	0	16
60.	EMP1	epithelial membrane protein 1	Chromosome 4 - NC_037548.1	0	0	0	4
61.	EPS8	epidermal growth factor receptor pathway substrate 8	Chromosome 4 - NC_037548.1	0	0	0	16
62.	ERBB2	erb-b2 receptor tyrosine kinase 2	Chromosome 3 - NC_037547.1	0	0	0	5
63.	ERCC6L2	ERCC excision repair 6 like 2	Chromosome 3 - NC_037547.1	0	0	0	14

64.	ETS2	ETS proto-oncogene 2, transcription factor	Chromosome 1 - NC_037545.1	0	0	0	1
65.	ETV6	ETS variant transcription factor 6	Chromosome 4 - NC_037548.1	0	0	0	59
66.	FABP4	fatty acid binding protein 4	Chromosome 15 - NC_037559.1	0	0	0	4
67.	FAM135B	family with sequence similarity 135 member B	Chromosome 15 - NC_037559.1	0	0	0	114
68.	FAM13A	family with sequence similarity 13 member A	Chromosome 7 - NC_037551.1	0	0	0	41
69.	FBLN5	fibulin 5	Chromosome 20 - NC_037564.1	0	0	0	10
70.	FHIT	fragile histidine triad diadenosine triphosphatase	Chromosome 21 - NC_037565.1	0	1	0	306
71.	GABARAPL1	GABA type A receptor associated protein like 1	Chromosome 4 - NC_037548.1	0	0	0	7
72.	GBA	glucosylceramidase beta	Chromosome 6 - NC_037550.1	0	0	0	3
73.	GDNF	glial cell derived neurotrophic factor	Chromosome 19 - NC_037563.1	0	0	0	2
74.	GHR	growth hormone receptor	Chromosome 19 - NC_037563.1	0	2	1	59
75.	GINS4	GINS complex subunit 4	Chromosome 1 - NC_037545.1	0	0	0	3
76.	GPIHBP1	glycosylphosphatidylinositol anchored high density lipoprotein binding protein 1	Chromosome 15 - NC_037559.1	0	0	0	1
77.	GRHL2	grainyhead like transcription factor 2	Chromosome 15 - NC_037559.1	0	0	0	26
78.	GRHL3	grainyhead like transcription factor 3	Chromosome 2 - NC_037546.1	0	0	0	12
79.	GRIA4	glutamate ionotropic receptor AMPA type subunit 4	Chromosome 16 - NC_037560.1	0	0	0	112

80.	GRIN2B	glutamate ionotropic receptor NMDA type subunit 2B	Chromosome 4 - NC_037548.1	0	0	0	146
81.	GSG1	germ cell associated 1	Chromosome 4 - NC_037548.1	0	0	0	9
82.	GUCY2C	guanylate cyclase 2C	Chromosome 4 - NC_037548.1	0	1	0	31
83.	HCN1	hyperpolarization activated cyclic nucleotide gated potassium channel 1	Chromosome 19 - NC_037563.1	0	0	0	73
84.	HPSE2	heparanase 2 (inactive)	Chromosome 23 - NC_037567.1	0	0	0	159
85.	HSF1	heat shock transcription factor 1	Chromosome 15 - NC_037559.1	0	0	0	2
86.	HSPA8	heat shock protein family A (Hsp70) member 8	Chromosome 16 - NC_037560.1	0	1	0	1
87.	IDH1	isocitrate dehydroge-nase (NADP(+)) 1	Chromosome 2 - NC_037546.1	0	0	0	5
88.	IFIH1	interferon induced with helicase C domain 1	Chromosome 2 - NC_037546.1	0	0	0	2
89.	IGF1	insulin like growth factor 1	Chromosome 4 - NC_037548.1	0	0	0	6
90.	IGF1R	insulin like growth factor 1 receptor	Chromosome 20 - NC_037564.1	0	0	0	35
91.	IGF2	insulin like growth factor 2	Chromosome 5 - NC_037549.1	0	0	0	3
92.	IGFBP2	insulin like growth factor binding protein 2	Chromosome 2 - NC_037546.1	0	0	0	16
93.	IGFBP7	insulin like growth factor binding protein 7	Chromosome 7 - NC_037551.1	0	0	0	5
94.	IL1RAPL2	interleukin 1 receptor accessory protein like 2	Chromosome X - NC_037569.1	0	0	0	127
95.	IQANK1	IQ motif and ankyrin repeat containing 1	Chromosome 15 - NC_037559.1	0	1	1	0

96. IRF9	interferon regulatory factor 9	Chromosome 11 - NC_037555.1	0	0	0	3
97. ITPR2	inositol 1,4,5-tris phosphate receptor type 2	Chromosome 4 - NC_037548.1	0	3	0	134
98. KCNIP4	potassium voltage-gated channel interacting protein 4	Chromosome 7 - NC_037551.1	0	0	0	182
99. KCNK9	potassium two pore domain channel subfamily K member 9	Chromosome 15 - NC_037559.1	0	0	0	21
100. KCNQ3	potassium voltage-gated channel subfamily Q member 3	Chromosome 15 - NC_037559.1	0	0	0	115
101. KHDRBS3	KH RNA binding domain containing, signal transduction associated 3	Chromosome 15 - NC_037559.1	0	0	0	41
102. KLHL41	kelch like family member 41	Chromosome 2 - NC_037546.1	0	0	0	3
103. KSR2	kinase suppressor of ras 2	Chromosome 17 - NC_037561.1	0	0	0	138
104. LAP3	leucine amino peptidase 3	Chromosome 7 - NC_037551.1	0	0	0	14
105. LEP	leptin	Chromosome 8 - NC_037552.1	0	0	0	2
106. LGMN	Legumain	Chromosome 20 - NC_037564.1	0	0	0	6
107. LIPA	lipase A, lysosomal acid type	Chromosome 23 - NC_037567.1	0	0	0	13
108. LPIN1	lipin 1	Chromosome 12 - NC_037556.1	0	0	0	51
109. LPL	lipoprotein lipase	Chromosome 3 - NC_037547.1	0	0	0	12
110. LRP6	LDL receptor related protein 6	Chromosome 4 - NC_037548.1	0	0	0	11

111. LTF	lactotransferrin	Chromosome 21 - NC_037565.1	0	0	0	1
112. MANSC1	MANSC domain containing 1	Chromosome 4 - NC_037548.1	0	0	0	27
113. MAP4K4	mitogen-activated protein kinase kinase kinase kinase 4	Chromosome 12 - NC_037556.1	0	0	0	59
114. MATN2	matrilin 2	Chromosome 15 - NC_037559.1	0	0	0	40
115. MFGE8	milk fat globule EGF and factor V/VIII domain containing	Chromosome 20 - NC_037564.1	0	0	0	2
116. MGST1	microsomal glutathione S-transferase 1	Chromosome 4 - NC_037548.1	0	0	0	12
117. MOCS1	molybdenum cofactor synthesis 1	Chromosome 2 - NC_037546.1	0	0	0	16
118. MROH1	maestro heat like repeat family member 1	Chromosome 15 - NC_037559.1	0	0	0	2
119. MRPL32	mitochondrial ribosomal protein L32	Chromosome 8 - NC_037552.1	0	0	0	3
120. MRPS30	mitochondrial ribosomal protein S30	Chromosome 19 - NC_037563.1	0	0	1	4
121. MS4A8	membrane spanning 4-domains A8	Chromosome 5 - NC_037549.1	0	1	0	14
122. MTMR12	myotubularin related protein 12	Chromosome 19 - NC_037563.1	0	0	0	7
123. MTMR3	myotubularin related protein 3	Chromosome 17 - NC_037561.1	0	0	0	8
124. MYO16	myosin XVI	Chromosome 13 - NC_037557.1	0	2	0	165
125. NALCN	sodium leak channel, non-selective	Chromosome 12 - NC_037557.1	0	0	0	72
126. NBAS	NBAS subunit of NRZ tethering complex	Chromosome 12 - NC_037556.1	0	1	2	91
127. NCKAP5	NCK associated protein 5	Chromosome 2 - NC_037546.1	0	0	0	273
128. NFIB	nuclear factor I B	Chromosome 3 - NC_037547.1	0	0	0	99

129. NIPBL	NIPBL cohesin loading factor	Chromosome 19 - NC_037563.1	0	0	0	21
130. NNT	nicotinamide nucleotide transhydrogenase	Chromosome 19 - NC_037563.1	0	1	0	24
131. NPC1	NPC intracellular cholesterol transporter 1	Chromosome 22 – NC_037566.1	0	4	1	20
132. NPR3	natriuretic peptide receptor 3	Chromosome 19 – NC_037563.1	0	0	0	44
133. NR4A1	nuclear receptor subfamily 4 group A member 1	Chromosome 4 - NC_037548.1	0	0	0	3
134. NT5DC3	5'-nucleotidase domain containing 3	Chromosome 4 - NC_037548.1	0	0	0	12
135. NT5E	5'-nucleotidase ecto	Chromosome 10 - NC_037554.1	0	0	0	15
136. NUCB2	nucleobindin 2	Chromosome 16 - NC_037560.1	0	0	0	10
137. NUP98	nucleoporin 98 and 96 precursor	Chromosome 16 – NC_037560.1	0	0	0	1
138. OCIAD2	OCIA domain containing 2	Chromosome 7 - NC_037551.1	0	0	0	7
139. OCLN	occludin	Chromosome 19 - NC_037563.1	0	0	0	13
140. OLR1	oxidized low density lipoprotein receptor 1	Chromosome 4 - NC_037548.1	0	0	0	3
141. OPLAH	5-oxoprolinase, ATP-hydrolysing	Chromosome 15 - NC_037559.1	0	0	0	3
142. OSMR	oncostatin M receptor	Chromosome 19 - NC_037563.1	0	0	0	11
143. OXCT1	3-oxoacid CoA-transferase 1	Chromosome 19 - NC_037563.1	0	0	0	50
144. P4HA3	prolyl 4-hydroxylase subunit alpha 3	Chromosome 16 - NC_037560.1	0	0	0	11
145. PAH	phenylalanine hydroxylase	Chromosome 4 – NC_037548.1	0	0	0	16
146. PARD3	par-3 family cell polarity regulator	Chromosome 14 - NC_037558.1	0	1	0	69

147. PARP12	poly(ADP-ribose) polymerase family member 12	Chromosome 8 - NC_037552.1	0	0	0	5
148. PARP8	poly(ADP-ribose) polymerase family member 8	Chromosome 19 - NC_037563.1	0	0	0	31
149. PDE1B	phosphodiesterase 1B	Chromosome 4 - NC_037548.1	0	0	0	3
150. PDGFRB	platelet derived growth factor receptor beta	Chromosome 9 - NC_037553.1	0	0	0	4
151. PHF20L1	PHD finger protein 20 like 1	Chromosome 15 - NC_037559.1	0	0	0	9
152. PIK3C2G	phosphatidylinositol-4-phosphate 3-kinase catalytic subunit type 2 gamma	Chromosome 4 - NC_037548.1	0	0	0	118
153. PKD2	polycystin 2, transient receptor potential cation channel	Chromosome 7 - NC_037551.1	0	0	0	16
154. PLBD1	phospholipase B domain containing 1	Chromosome 4 - NC_037548.1	0	0	0	12
155. PLCE1	phospholipase C epsilon 1	Chromosome 23 - NC_037567.1	0	0	0	131
156. PLCXD3	phosphatidylinositol specific phospholipase C X domain containing 3	Chromosome 19 - NC_037563.1	0	0	0	40
157. PLEC	plectin	Chromosome 15 – NC_037559.1	0	0	0	3
158. PLEKHA5	pleckstrin homology domain containing A5	Chromosome 4 - NC_037548.1	0	0	0	33
159. PPARGC1A	PPARG coactivator 1 alpha	Chromosome 7 - NC_037551.1	0	0	0	19
160. PRKDC	protein kinase, DNA-activated, catalytic subunit	Chromosome 15 - NC_037559.1	0	2	2	22
161. PRLR	prolactin receptor	Chromosome 19 - NC_037563.1	0	0	0	54

162. PRNP	prion protein	Chromosome 14 - NC_037558.1	0	0	0	6
163. PTK2	protein tyrosine kinase 2	Chromosome 15 - NC_037559.1	0	0	0	15
164. PTPRO	protein tyrosine phosphatase receptor type O	Chromosome 4 - NC_037548.1	0	0	0	85
165. RAI1	retinoic acid induced 1	Chromosome 3 – NC_037547.1	0	0	0	8
166. RAI14	retinoic acid induced 14	Chromosome 19 - NC_037563.1	0	1	0	26
167. RERG	RAS like estrogen regulated growth inhibitor	Chromosome 4 – NC_037548.1	0	0	0	49
168. RFX4	regulatory factor X4	Chromosome 4 - NC_037548.1	0	0	0	36
169. RGS22	regulator of G protein signaling 22	Chromosome 15 - NC_037559.1	2	1	1	32
170. RICTOR	RPTOR independent companion of MTOR complex 2	Chromosome 19 - NC_037563.1	0	0	0	11
171. RNF19A	ring finger protein 19A, RBR E3 ubiquitin protein ligase	Chromosome 15 – NC_037559.1	0	0	0	4
172. RPL23A	ribosomal protein L23a	Chromosome 3 - NC_037547.1	0	0	0	1
173. RUNX2	RUNX family transcription factor 2	Chromosome 2 - NC_037546.1	0	0	0	50
174. SCARB1	scavenger receptor class B member 1	Chromosome 17 - NC_037561.1	0	0	0	34
175. SCRN1	secernin 1	Chromosome 8 - NC_037552.1	0	0	0	7
176. SDC2	syndecan 2	Chromosome 15 - NC_037559.1	0	0	0	18
177. SDR16C5	short chain dehydrogenase/ reductase family 16C member 5	Chromosome 15 - NC_037559.1	0	0	0	10

178. SERPINA1	serpin family A member 1	Chromosome 20 - NC_037564.1	0	0	0	13
179. SESN2	sestrin 2	Chromosome 2 - NC_037546.1	0	0	0	3
180. SLC15A5	solute carrier family 15 member 5	Chromosome 4 - NC_037548.1	0	0	0	59
181. SLC1A3	solute carrier family 1 member 3	Chromosome 19 - NC_037563.1	0	0	0	17
182. SLC35F6	solute carrier family 35 member F6	Chromosome 12 - NC_037556.1	0	0	0	12
183. SLC45A4	solute carrier family 45 member 4	Chromosome 15 - NC_037559.1	0	0	0	6
184. SLC8A1	solute carrier family 8 member A1	Chromosome 12 - NC_037556.1	0	0	1	104
185. SLC8A3	solute carrier family 8 member A3	Chromosome 11 - NC_037555.1	0	1	1	39
186. SMCO3	single-pass membrane protein with coiled-coil domains 3	Chromosome 4 - NC_037548.1	0	0	0	17
187. SMPD5	sphingomyelin phosphodiesterase 5	Chromosome 15 - NC_037559.1	0	0	0	5
188. SOX5	SRY-box transcription factor 5	Chromosome 4 - NC_037548.1	0	0	0	167
189. SPEF2	sperm flagellar 2	Chromosome 19 - NC_037563.1	0	0	0	34
190. SPOCK1	SPARC (osteonectin), cwcv and kazal like domains proteoglycan 1	Chromosome 9 - NC_037553.1	0	1	0	153
191. ST3GAL1	ST3 beta-galactoside alpha-2,3-sialyltransferase 1	Chromosome 15 - NC_037559.1	0	0	0	32
192. ST8SIA1	ST8 alpha-N-acetyl-neuraminide alpha-2, 8-sialyltransferase 1	Chromosome 4 - NC_037548.1	0	0	0	49
193. STAT1	signal transducer and activator of transcription 1	Chromosome 2 - NC_037546.1	0	0	0	5

194. SUPT3H	signal transducer and activator of transcription 1	Chromosome 2 - NC_037546.1	0	0	0	22
195. SWT1	SWT1 RNA endoribonuclease homolog	Chromosome 5 - NC_037549.1	0	0	0	14
197. SYN3	synapsin III	Chromosome 4 - NC_037548.1	0	0	0	122
198. TBC1D1	TBC1 domain family member 1	Chromosome 7 – NC_037551.1	0	0	0	34
199. TBCD	tubulin folding cofactor D	Chromosome 3 - NC_037547.1	0	0	0	19
200. TDRKH	tudor and KH domain containing	Chromosome 6 - NC_037550.1	0	0	0	2
201. TG	thyroglobulin	Chromosome 15 - NC_037559.1	0	2	3	58
202. TGM2	transglutaminase 2	Chromosome 14 - NC_037558.1	0	0	0	1
203. TLR4	toll like receptor 4	Chromosome 3 – NC_037547.1	0	0	0	9
204. TNFSF10	TNF superfamily member 10	Chromosome 1 - NC_037545.1	0	1	0	12
205. TPCN1	two pore segment channel 1	Chromosome 17 - NC_037561.1	0	0	0	14
206. TRAPPC9	trafficking protein particle complex 9	Chromosome 15 - NC_037559.1	0	0	0	95
207. TSHB	thyroid stimulating hormone subunit beta	Chromosome 6 - NC_037550.1	0	0	0	2
208. TSNARE1	t-SNARE domain containing 1	Chromosome 15 - NC_037559.1	0	0	0	3
209. TSPAN32	tetraspanin 32	Chromosome 5 - NC_037549.1	0	1	1	2
210. TTC33	tetratricopeptide repeat domain 33	Chromosome 19 - NC_037563.1	0	0	0	6
211. UGDH	UDP-glucose 6-dehydrogenase	Chromosome 7 - NC_037551.1	0	0	0	4

212. VIPR2	vasoactive intestinal peptide receptor 2	Chromosome 8 - NC_037552.1	0	0	0	27
213. VPS13B	vacuolar protein sorting 13 homolog B	Chromosome 15 - NC_037559.1	0	2	0	74
214. VPS28	VPS28 subunit of ESCRT-I	Chromosome 15 - NC_037559.1	0	1	0	1
215. ZC3H3	zinc finger CCCH-type containing 3	Chromosome 15 - NC_037559.1	0	0	0	4
216. ZDHHC17	zinc finger DHHC-type palmitoyltransferase 17	Chromosome 4 - NC_037548.1	0	0	0	16
217. ZMYND11	zinc finger MYND-type containing 11	Chromosome 14 – NC_037558.1	0	0	0	4
218. ZNF484	zinc finger protein 484	Chromosome 3 - NC_037547.1	0	0	0	9
219. ZNF696	zinc finger protein 696	Chromosome 15 - NC_037559.1	0	1	0	1

4.3 Genetic diversity

4.3.1 Population diversity parameters

Table 4.8 Observed (Ho) and expected (He) Heterozygosity estimated in the Indian buffalo population

Breeds	Observed. Het	Expected. Het
Bhadawari	0.2343	0.2366
Mehsana	0.2314	0.2239
Murrah	0.2372	0.2462
Pandharpuri	0.2366	0.2390
Surti	0.2361	0.2255
Toda	0.2150	0.2111

The observed heterozygosity (Ho) and expected heterozygosity (He) was found highest in Murrah breed (0.2372 and 0.2462) followed by Pandharpuri breed (0.2366 and 0.2390), while lowest was observed in Toda breed (0.2150 and 0.2111) respectively.

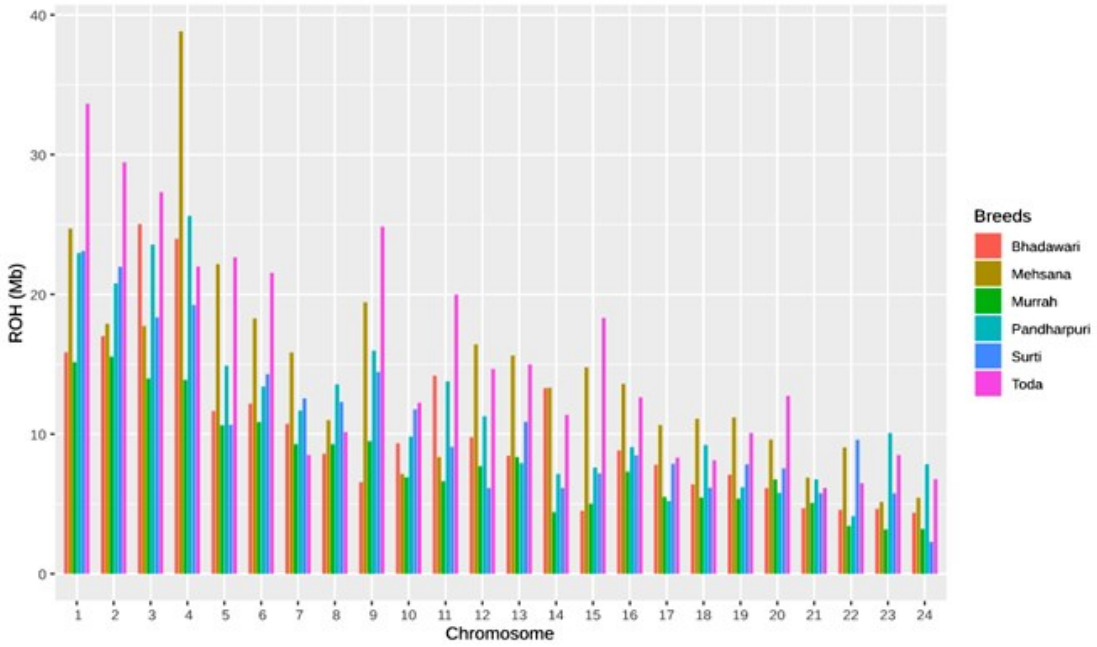


Fig. 4.3: Average ROH length per Chromosome

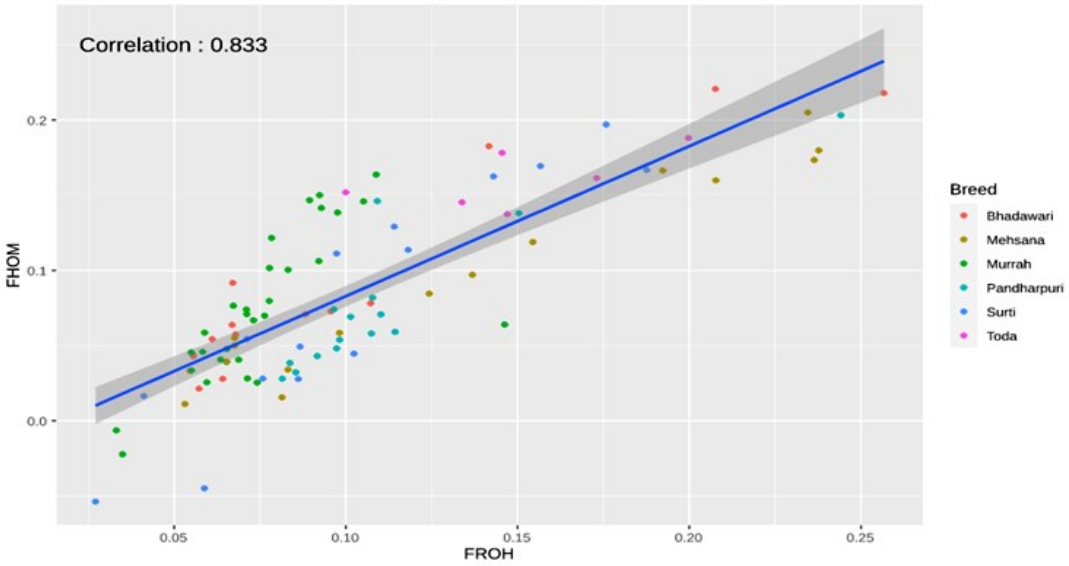


Fig. 4.4: Correlation between FHOM & FROH

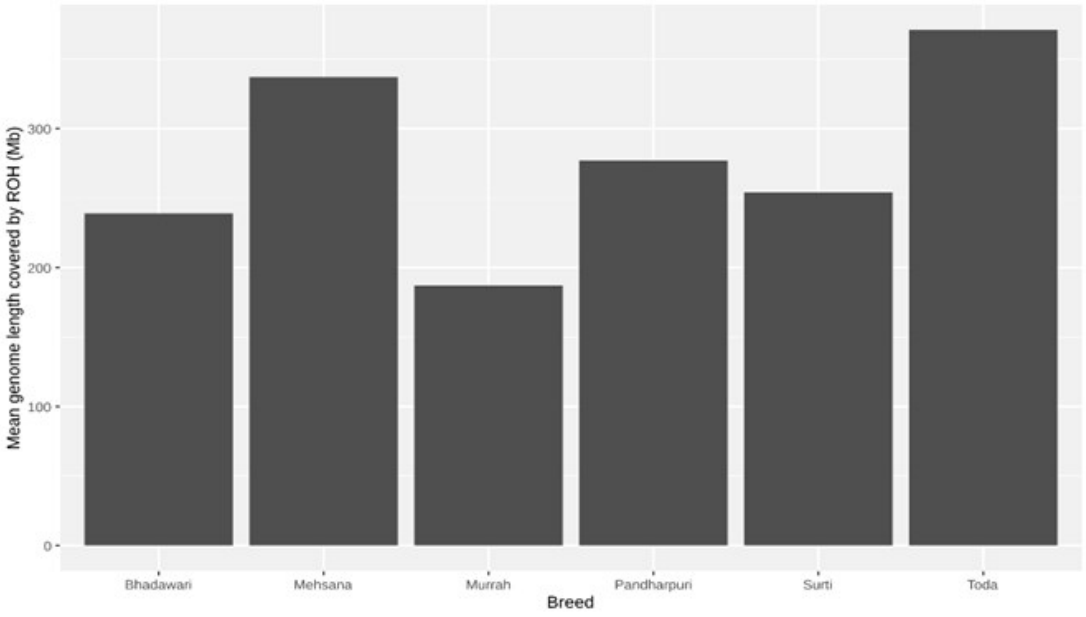


Fig. 4.5: Mean Genome length covered by ROH per Breed

4.3.2 F-statistics (FST)

Table 4.9 F_{ST} in buffalo breeds

Bhadawari	Mehsana	Murrah	Pandharpuri	Surti	Toda
Bhadawari					
Mehsana	0.155				
Murrah	0.111	0.167			
Pandharpuri	0.136	0.199	0.154		
Surti	0.072	0.13	0.085	0.117	
Toda	0.121	0.176	0.134	0.166	0.095

F_{ST} values showed lowest genetic distance between Bhadawari and Surti (0.072) followed by Murrah and Surti (0.085) while highest genetic distance was found in Mehsana and Pandharpuri (0.199) followed by Mehsana and Toda (0.176).

4.3.3 Runs of homozygosity

Runs of homozygosity (ROH) in the autosomes of 96 buffalo animals were determined using PLINK1.9 and consisted of 237,762 SNPs across all the breeds after quality control. The average total length of ROH per animal were 2, 3, 1, 2, 2 & 2 mb and Average numbers of ROH per animal were 107, 111, 114, 118, 124 & 180 in the Bhadawari, Mehsana, Murrah, Pandharpuri, Surti and Toda breeds of Buffalo, respectively. The average total length of ROH per animal was 239, 337, 187, 277, 254, 371 Mb respectively in Bhadawari, Mehsana, Murrah, Pandharpuri, Surti and Toda. The average ROH length per chromosome was shown in figure 4.3. Individuals with an almost equal portion of the genome covered by ROH had different numbers and lengths of ROH, which could be an indication of different combinations of recent and distant inbreeding events in the samples.

Among F_{ROH} estimates, it can be observed an increase in variation with ROH lengths, being evidenced by the coefficient of variation (CV). In this study F_{HOM} is highly correlated with F_{ROH} with the value of 0.833 (Fig. 4.4). Mean Genome length covered by ROH is more in Toda and less in Murrah (fig. 4.5).

Table 4.10 Breed-wise summary statistics of ROH observed

	Bhadawari	Mehsana	Murrah	Pandharpuri	Surti	Toda
Total No. of ROH	1610	1662	3411	1769	1865	1078
Avg. No. of ROH	107	111	114	118	124	180
Std.Deviation	20.64	13.34	27.53	16.15	44.09	21.47
Min. No.ROH	86	82	54	104	45	152
Max. No. ROH	157	134	162	158	180	206
Total ROH Length (Mb)	239	337	187	277	254	371
Avg. ROH Length (Mb)	2	3	1	2	2	2
F_{ROH}	0.0965	0.1360	0.0754	0.1119	0.1028	0.1498
Std.deviation	0.06	0.07	0.02	0.04	0.05	0.16
F_{HOM}	0.0853	0.0966	0.0740	0.0763	0.0781	0.1603
Std.deviation	0.07	0.06	0.05	0.05	0.08	0.02
Correlation_F_{ROH}—F_{HOM}	0.9394	0.9745	0.7060	0.8775	0.9268	0.6174

4.3.4 Linkage Disequillibrium

Genome-wide average LD (r^2) decreased with increasing genomic distance for all breeds. The maximum average r2 of 0.32, 0.35, 0.25, 0.32, 0.33, 0.46 obtained for Bhadawari, Mehsana, Murrah, Pandharpuri, Surti, Toda, respectively at a distance of <10 kb while the average minimum r^2 values obtained in this study for Bhadawari and Mehsana after 280 Kb, Murrah and Pandharpuri after 310 Kb, Surti after 330 Kb and Toda after 390 Kb (Table 4.11). In this study, the LD values are generally larger in Toda and making major differences with other breeds. We note that Pandharpuri, Surti and Bhadawari had a roughly similar LD decay pattern (fig 4.6). Murrah had lower LD which rapidly decayed with increasing distance between markers compared to Toda.

Table 4.11 Genome-wide average Linkage Disequillibrium (r^2) in Buffalo breeds

S. no.	Distance (Kb)	Bhadawari	Mehsana	Murrah	Pandharpuri	Surti	Toda
1.	0-10	0.32	0.35	0.25	0.32	0.33	0.46
2.	10-20	0.22	0.25	0.16	0.22	0.23	0.36
3.	20-30	0.19	0.23	0.13	0.20	0.20	0.33
4.	30-40	0.18	0.21	0.12	0.18	0.19	0.32
5.	40-50	0.17	0.20	0.11	0.17	0.18	0.30
6.	50-60	0.16	0.19	0.11	0.17	0.17	0.30
7.	60-70	0.16	0.19	0.10	0.16	0.16	0.29
8.	70-80	0.15	0.18	0.09	0.15	0.16	0.29
9.	80-90	0.15	0.18	0.09	0.15	0.16	0.29
10.	90-100	0.14	0.17	0.09	0.15	0.15	0.28
11.	100-110	0.14	0.17	0.08	0.14	0.15	0.28
12.	110-120	0.14	0.17	0.08	0.14	0.14	0.27
13.	120-130	0.13	0.17	0.08	0.14	0.14	0.27
14.	130-140	0.13	0.16	0.07	0.13	0.14	0.26
15.	140-150	0.13	0.16	0.07	0.13	0.14	0.26
16.	150-160	0.13	0.16	0.07	0.13	0.13	0.26
17.	160-170	0.13	0.16	0.07	0.13	0.13	0.26
18.	170-180	0.13	0.16	0.07	0.13	0.13	0.26
19.	180-190	0.12	0.16	0.07	0.13	0.13	0.25
20.	190-200	0.12	0.15	0.07	0.13	0.13	0.25
21.	200-210	0.12	0.15	0.06	0.12	0.13	0.25
22.	210-220	0.12	0.15	0.06	0.12	0.12	0.25
23.	220-230	0.12	0.15	0.06	0.12	0.12	0.25
24.	230-240	0.12	0.15	0.06	0.12	0.12	0.25
25.	240-250	0.12	0.15	0.06	0.12	0.12	0.25
26.	250-260	0.12	0.15	0.06	0.12	0.12	0.25
27.	260-270	0.12	0.15	0.06	0.12	0.12	0.25
28.	270-280	0.11	0.15	0.06	0.12	0.12	0.25
29.	280-290	0.11	0.14	0.06	0.12	0.12	0.24
30.	290-300	0.11	0.14	0.05	0.12	0.12	0.24
31.	300-310	0.11	0.14	0.06	0.11	0.12	0.24
32.	310-320	0.11	0.14	0.05	0.11	0.12	0.24
33.	320-330	0.11	0.14	0.05	0.11	0.12	0.24
34.	330-340	0.11	0.14	0.05	0.11	0.11	0.24

35.	340-350	0.11	0.14	0.05	0.11	0.11	0.24
36.	350-360	0.11	0.14	0.05	0.11	0.11	0.24
37.	360-370	0.11	0.14	0.05	0.11	0.11	0.24
38.	370-380	0.11	0.14	0.05	0.11	0.11	0.24
39.	380-390	0.11	0.14	0.05	0.11	0.11	0.24
40.	390-400	0.11	0.14	0.05	0.11	0.11	0.23
41.	400-410	0.11	0.14	0.05	0.11	0.11	0.23
42.	410-420	0.11	0.14	0.05	0.11	0.11	0.23
43.	420-430	0.11	0.14	0.05	0.11	0.11	0.23
44.	430-440	0.11	0.14	0.05	0.11	0.11	0.23
45.	440-450	0.11	0.14	0.05	0.11	0.11	0.23
46.	450-460	0.11	0.14	0.05	0.11	0.11	0.23
47.	460-470	0.11	0.13	0.05	0.11	0.11	0.23
48.	470-480	0.11	0.13	0.05	0.11	0.11	0.23
49.	480-490	0.11	0.13	0.05	0.11	0.11	0.23
50.	490-500	0.11	0.14	0.05	0.11	0.11	0.23

4.3.5 Effective population size (Ne)

Past and recent effective population sizes were estimated from the average r^2 for markers separated by various genomic distances. The extent of LD over greater recombinational distances indicated more recent Ne while that over shorter distances provided ancestral Ne (Hayes *et al.*, 2003). The Ne values for the different breeds of buffalo are listed in Table-4.12 with their generations ago. In figure 4.8 ancestral Ne i.e., upto 1000 generations presents while more recent Ne presents in figure 4.9. In general, Ne declined over time from larger to smaller Ne across the breeds (Makina *et al.,* 2015), we can see it in this study (Table 4.12). Result shows that Murrah has had the greatest genetic diversity over the generations among all the sampled populations, while Toda had the least. When we see the ancient plot of Ne, it can easily observed that there is some deviation in the curve of all different breeds at 250 generations ago except Toda which follows a constant curve in Ne throughout (fig. 4.8). The purebred breed had higher estimates of Ne than the crossbred breed, and vice versa. Bhadawari and Surti follows roughly similar effective population size (Ne) after 65 generations ago.

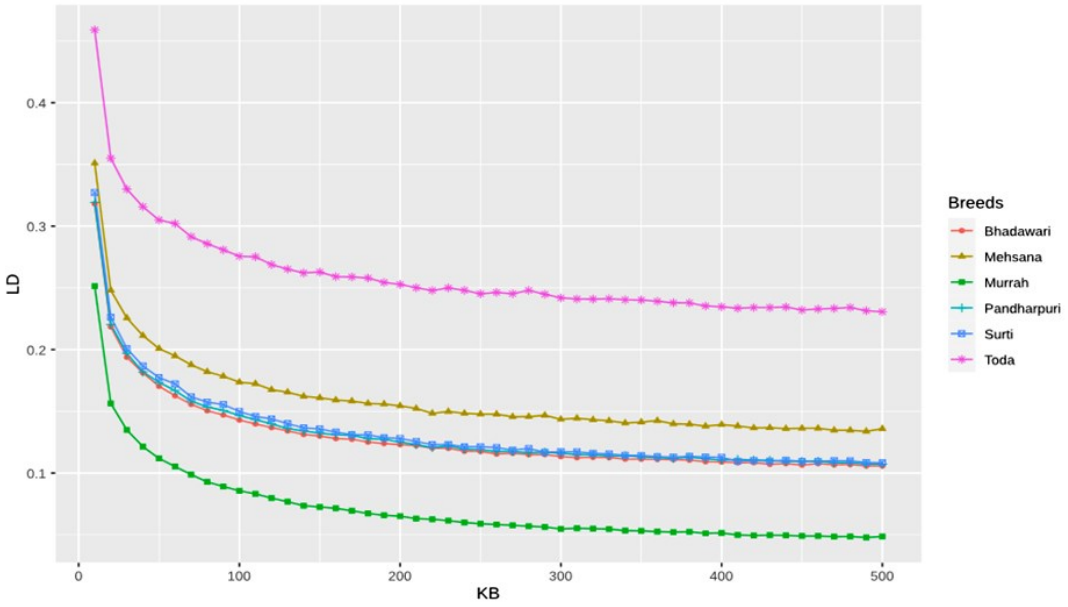


Fig. 4.6: Linkage disequilibrium (r^2) decay pattern graph in buffalo breeds

Table 4.12 Effective population size (Ne) for different breed of buffaloes

Generation ago	Bhadawari	Murrah	Pandharpuri	Surti	Toda
5	26	68	30	26	11
6	30	77	33	29	12
7	34	87	38	33	14
8	38	100	43	37	16
9	44	115	48	43	19
11	51	134	56	49	22
13	59	156	64	57	25
15	69	183	75	67	30
18	82	216	88	79	36
22	98	257	103	94	43
27	117	309	123	114	52
33	143	372	147	138	63
41	175	448	177	169	76
52	216	548	216	208	95
65	267	666	268	258	118
84	332	809	333	320	149
108	418	975	411	401	189
142	524	1154	515	498	241
187	660	1380	650	623	312
248	820	1605	808	780	405
329	1031	1876	1004	964	522
433	1266	2168	1238	1178	671
556	1535	2510	1492	1408	842
687	1792	2845	1785	1682	1016
809	2051	3141	2000	1907	1172
903	2216	3325	2200	2036	1269
960	2331	3411	2260	2117	1376
988	2336	3567	2209	2178	1435
998	2322	3351	2334	2163	1449

4.3.5 Haplotype Blocks and Tag SNPs

The quality control filtering for the haplotype block study was performed using the PLINK software (Purcell et al., 2007). Haplotype blocks formed by only two SNPs were discarded to avoid spurious block formation.

In this study, the total number of blocks identified was 7615, 8379, 11642, 7792, 7747, 1363 in Bhadawari, Mehsana, Murrah, Pandharpuri, Surti and Toda, respectively (Table 4.13). The highest number of blocks (11642) in Murrah and lowest number of blocks (1363) in Toda were found. Total number of SNPs (63444) in blocks was found highest in Mehsana and lowest (11967) in Toda. Chromosome 1 showed the most number of SNPs (5190) in Mehsana Buffalo and most number of blocks (938) in Murrah Buffalo, while Chromosome 19 showed the least number of SNPs (181) and Chromosome 24 showed the least number of blocks (27) in Toda Buffalo.

Table 4.13 Summarized Table of Haplotype in different breeds

	Total No. of Blocks	Total SNPs in Blocks	Highest No. of SNPs & Blocks	Lowest No. of SNPs & Blocks	Total Block length coverage (Mb)	Avg. Chromosome coverage by block (%)
Bhadawari	7615	44359	Chr-1 (4086 SNPs) (617 Blocks)	Chr-24 (989 SNPs) (150 Blocks)	400.26	15.82
Mehsana	8379	63444	Chr-1 (5190 SNPs) (668 Blocks)	Chr-23 (1196 SNPs) Chr-24 (150 Blocks)	672.17	26.84
Murrah	11642	62773	Chr-1 (5163 SNPs) (938 Blocks)	Chr-24 (1252 SNPs) Chr-18 (296 Blocks)	392.20	15.40
Pandharpuri	7792	51934	Chr-1 (4243 SNPs) (611 Blocks)	Chr-24 (945 SNPs) (157 Blocks)	429.33	16.93
Surti	7747	52526	Chr-1 (4456 SNPs) (625 Blocks)	Chr-24 (952 SNPs) (144 Blocks)	473.93	18.65
Toda	1363	11967	Chr-1 (1269 SNPs) (126 Blocks)	Chr-19 (181 SNPs) Chr-24 (27 Blocks)	125.07	4.95

4.3.6 Principal Component Analysis

The first and second PCs explain 3.4% and 2.86% of the total variance, respectively (Fig. 4.15). The PCA plot separated the individuals of the six breeds. PC1 clearly separated Mehsana from the rest of the breeds. Murrah and Bhadawari were found to be closer in

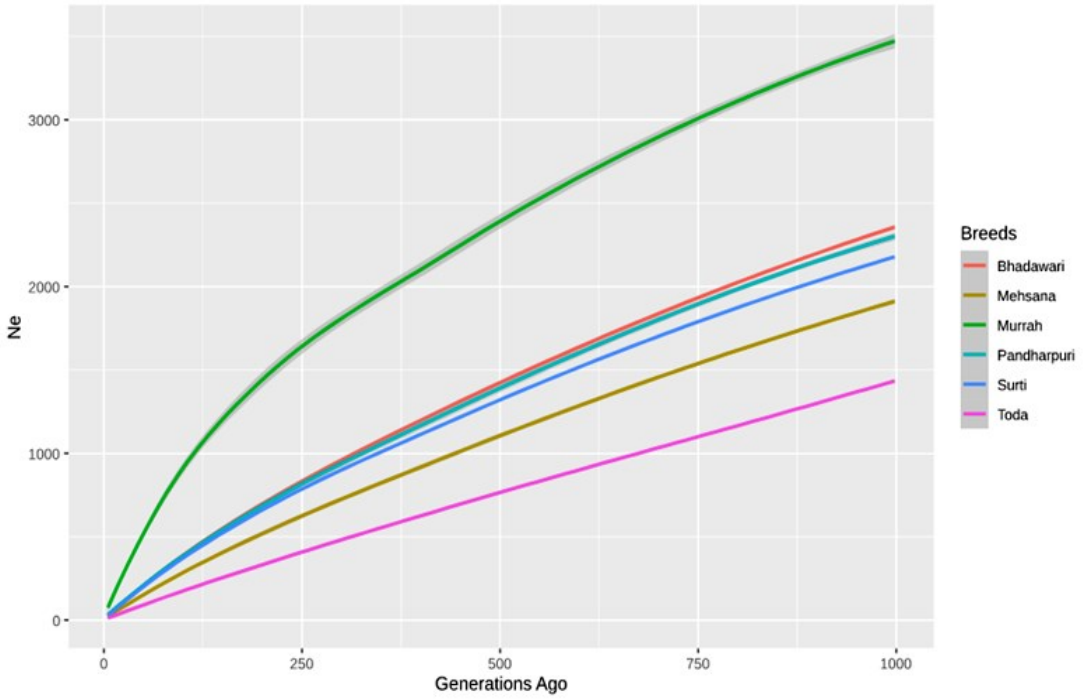


Fig. 4.7: Effective population size (Ne) before 1000 years ago in different breeds

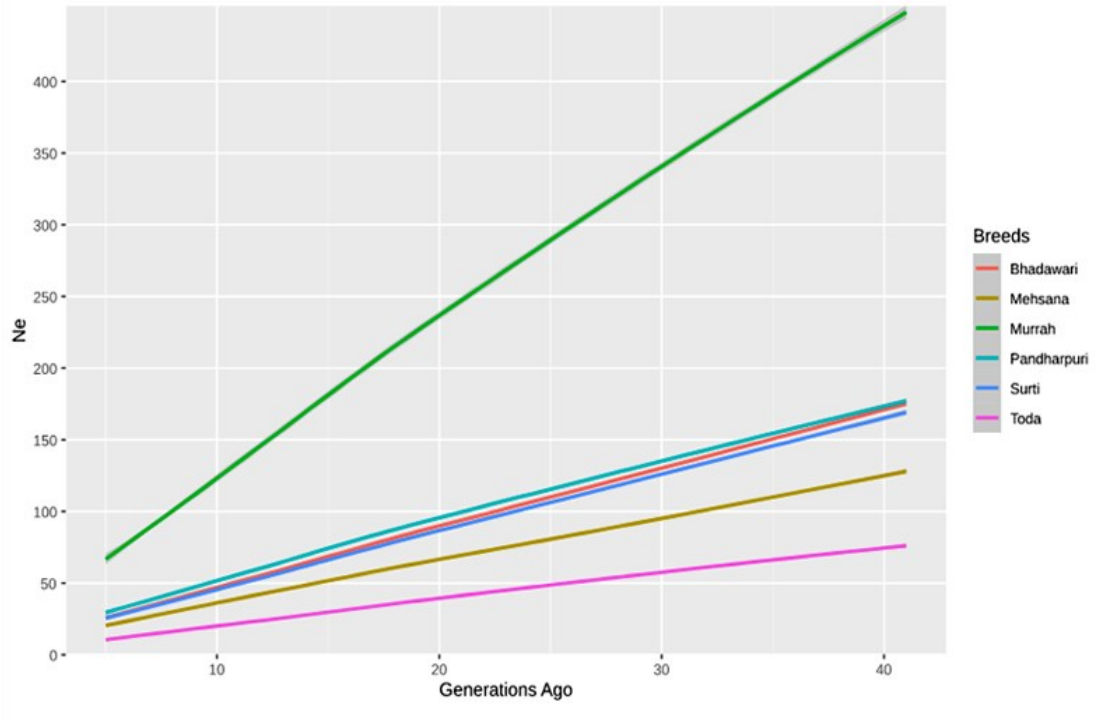


Fig. 4.8: Recent Effective population size (Ne) in different Buffalo breeds

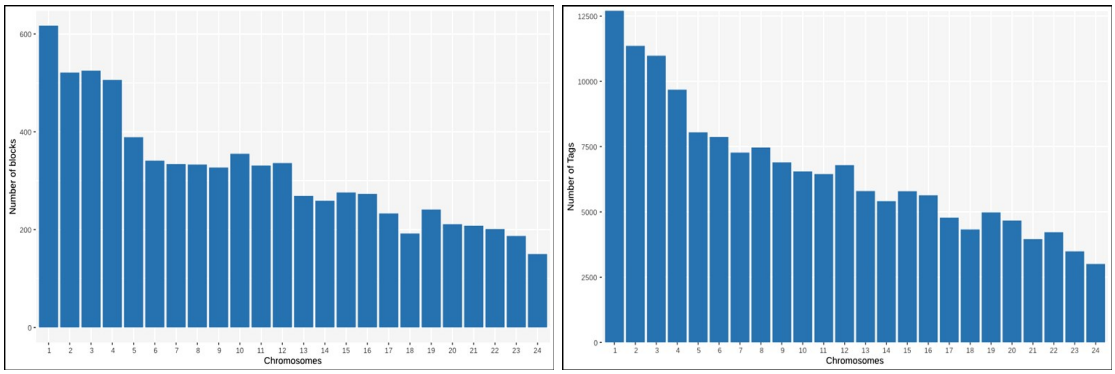


Fig. 4.9: Showing haplotype blocks and Tagged SNPs in Bhadawari Buffalo

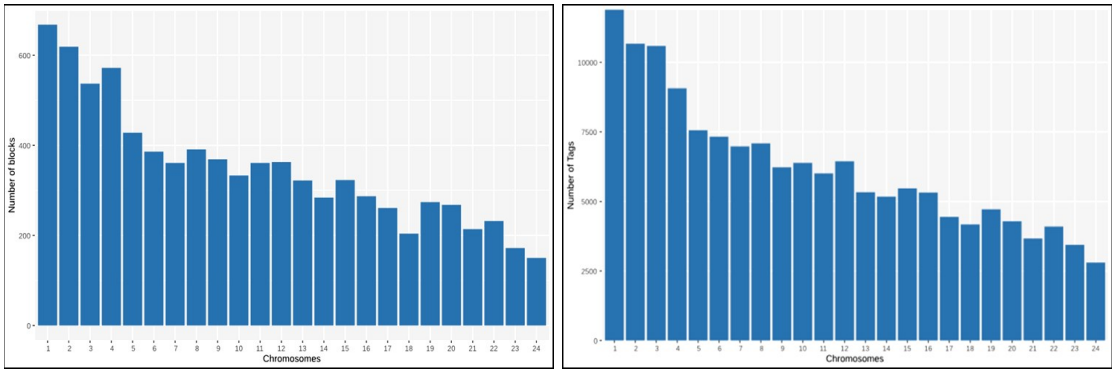


Fig. 4.10: Showing haplotype blocks and Tagged SNPs in Mehsana Buffalo

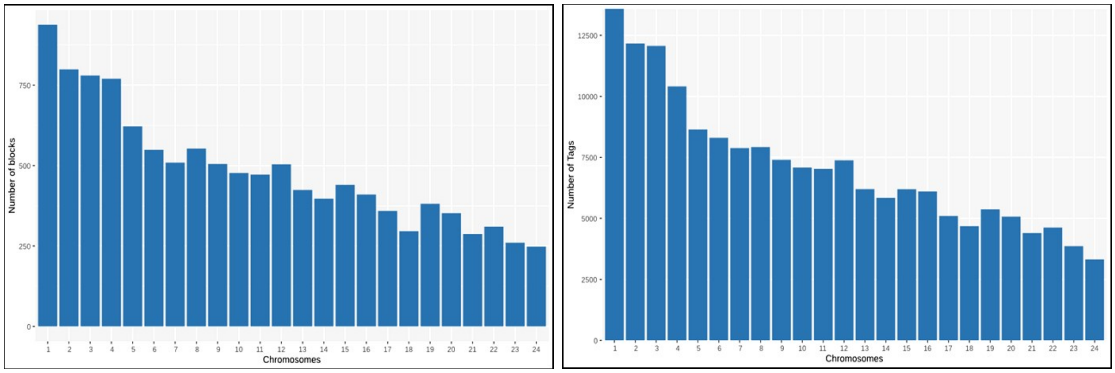


Fig. 4.11: Showing haplotype blocks and Tagged SNPs in Murrah Buffalo

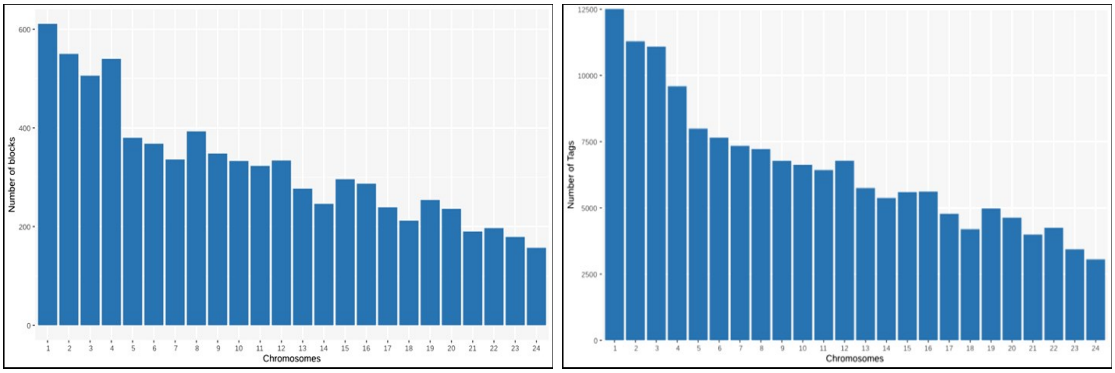


Fig. 4.12: Showing haplotype blocks and Tagged SNPs in Pandharpuri Buffalo

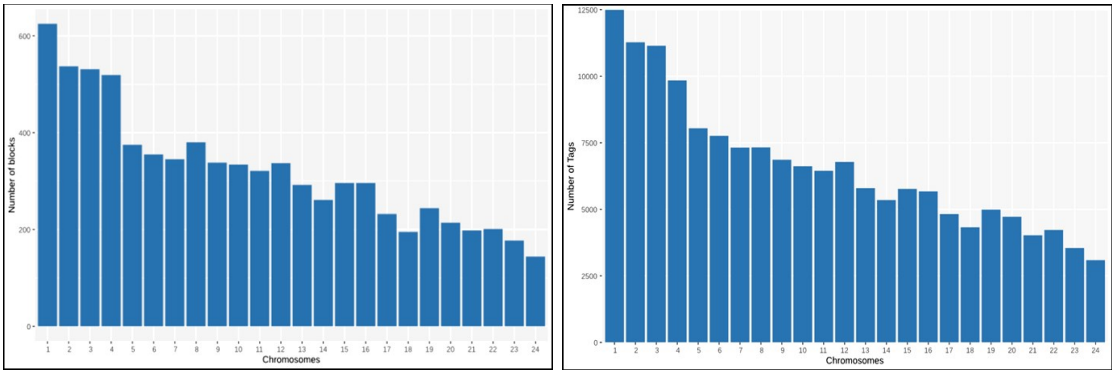


Fig. 4.13: Showing haplotype blocks and Tagged SNPs in Surti Buffalo

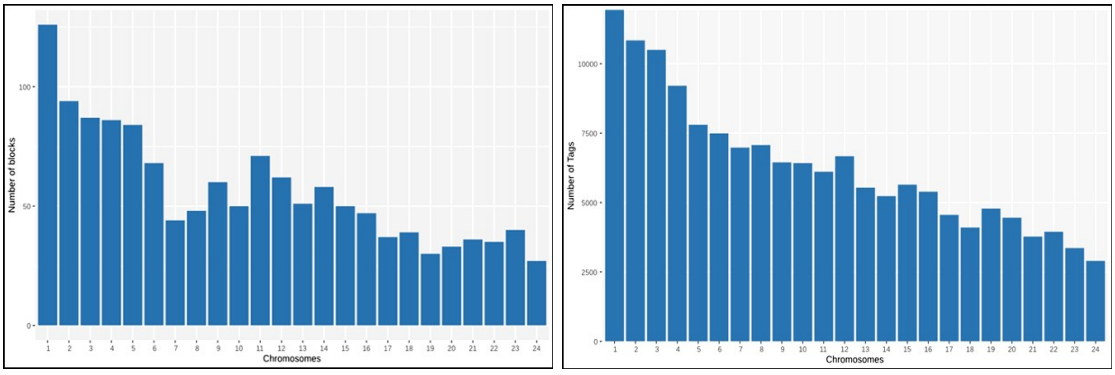


Fig. 4.14: Showing haplotype blocks and Tagged SNPs in Toda Buffalo

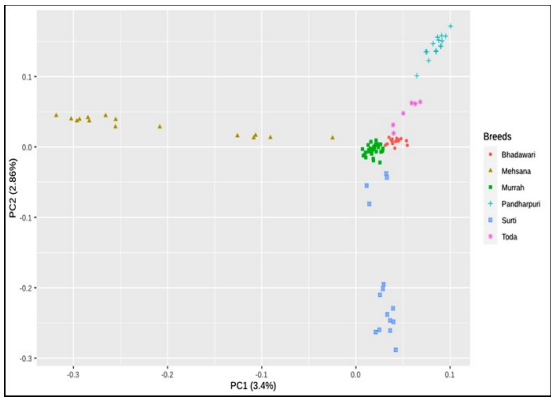


Fig. 4.15: PC1 & PC2

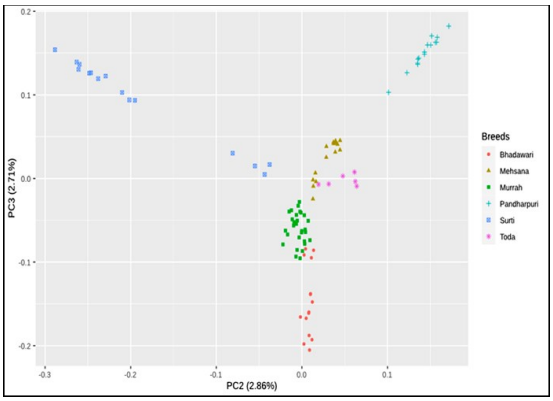


Fig. 4.16: PC2 & PC3

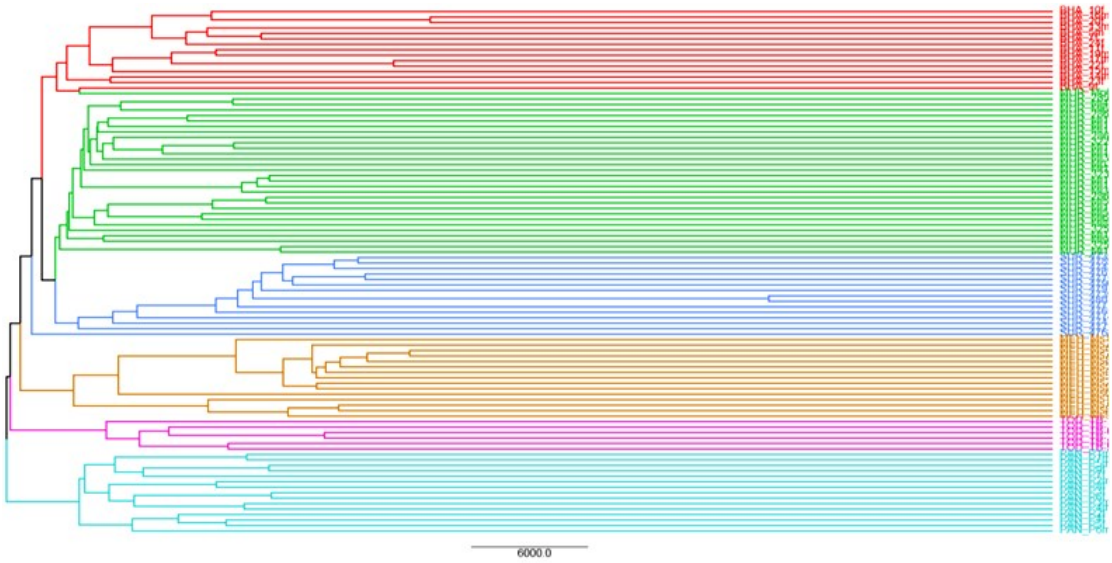


Fig. 4.17: Phylogenetic tree prepared from the IBS distance matrix

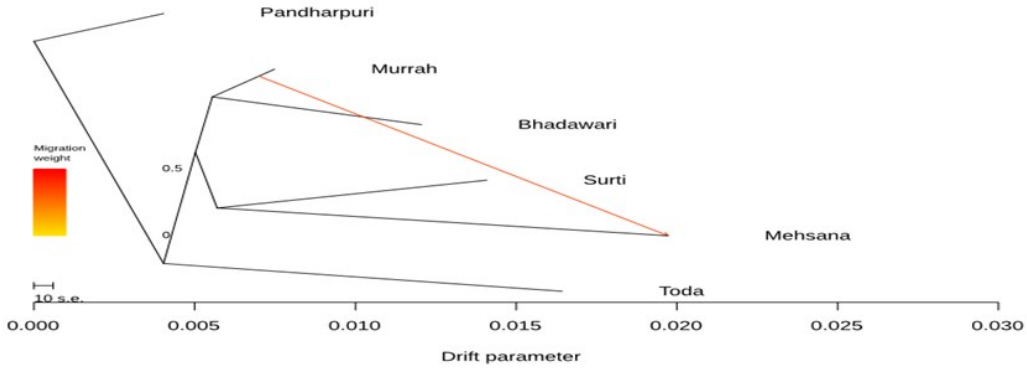


Fig. 4.18: Tree-mix phylogram

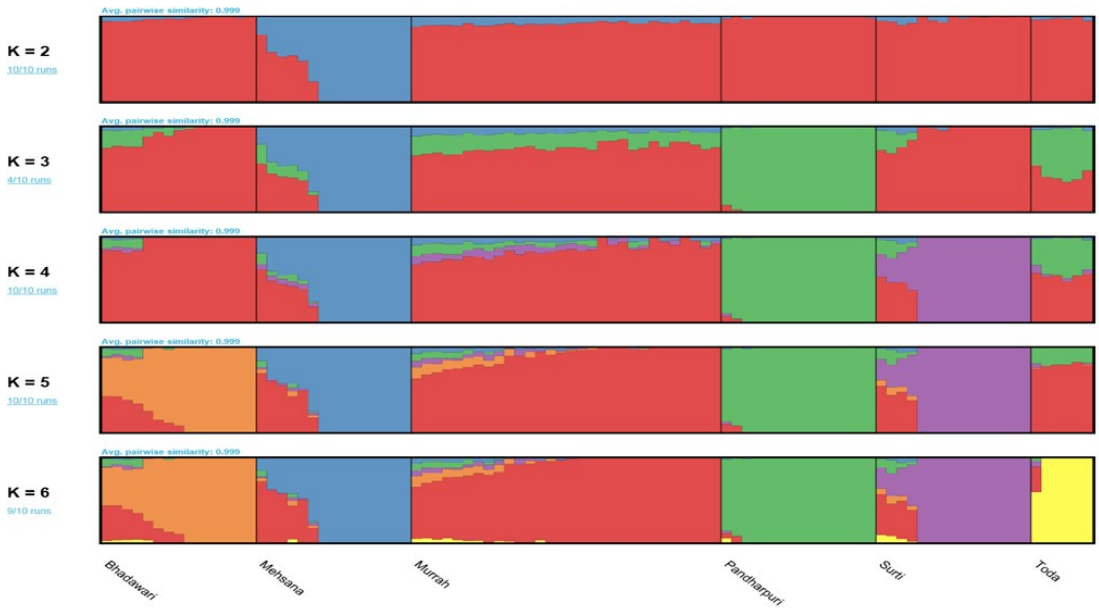


Fig. 4.19: Showing the Admixture analysis for K values of 2 to 6

cluster. Some of the animals from Surti buffalo were found closer to Murrah and distant to Mehsana buffalo while PC2 separated Mehsana, Pandharpuri, Toda and Bhadawari from Murrah, and Surti. PC3 explains the 2.71% of the total variance, there is a clearer separation between Murrah and Surti unlike the PC2 (Fig. 4.16).

4.3.7 Phylogenetic Analysis

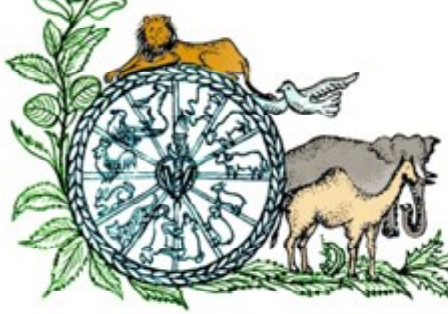
The phylogenetic tree constructed using the IBS matrix showed that the animals clustered among their respective breed groups (Fig. 4.17). Pandharpuri formed a separate lineage to the rest of the breeds.

The maximum likelihood phylogram constructed with Treemix also displayed similar tree topology (Fig. 4.18). The addition of one migration path in Treemix revealed the introgression of Murrah inheritance in Mehsana buffaloes. This tree explained 99.6% of the covariance observed between populations, whereas the tree without any migration events explained only 98.3% of the covariance.

4.3.8 Admixture Analysis

According to ADMIXTURE software analysis, the first subdivision highlighted at K = 2 (Fig. 4.19). As with PC1, Mehsana was separated from the rest of the breeds at K = 2. It became evident from the Admixture plot that Mehsana animals had varying levels of shared ancestry with the Murrah cluster, which also explains their spread along the PC2 axis in the PCA plot. K = 3 clearly separated Pandharpuri, which gives credence to the results of the phylogenetic analyses. Toda showed a mixture of Pandharpuri and Murrah inheritance. At K = 6, all the breeds were assigned to their own clusters, with varying levels of Murrah ancestry appearing in other breeds which may be due to use of Murrah buffalo in every region of India.

Discussion



5 DISCUSSION

5.1 SNP Genotyping

The phenotypic and genotypic characterization of livestock forms an integral part of any breeding programme. The phenotypic characterization, on the one hand, includes only the phenotypic features of a breed/population; while, on the other hand, the genotypic characterization strives to analyze the population's patterns of genetic uniformity, admixture, subdivisions, inbreeding and introgression. In addition, genetic characterization seeks to provide useful insights into the development of breeds/strains that may be equally applicable in crossbred populations. SNP markers are third generation markers that have a cutting edge over other markers, including mitochondrial DNA or other single locus markers, although there are different markers that have commonly been used. They have much greater resolution of population histories and coalescent nature of animal populations (McTavish *et al*., 2013). In recent studies on admixture, population composition and tracing discrete domestication and divergence trends, these types of markers have been recommended for use.

The ddRAD-seq method, even in the absence of an available genome sequence, is a powerful and cost-effective means of SNP discovery, producing thousands of high-quality SNPs (Peterson *et al*., 2012). Through this technique, 397.8 million paired end reads (pre-QC) *and total 367.2 million reads after QC were generated* in 6 different Buffalo breeds having 96 samples selected from 6 different farms located in different states in India.

5.2 Identification of SNPs

5.2.1 Raw reads

In this study, we found total 397.8 million paired end reads for 96 Buffalo samples averaging 2.07 million reads per sample.

In the study by Imartino *et al*. (2013), an average of 2.12 million reads per sample for 48 water buffalo were obtained by GBS using the reference of water buffalo genome which is similar to the present study. Furthermore, in the current analysis, the double digestion approach was used instead of GBS. The study done by Surya *et al*. (2018) revealed total of 21.2 million raw reads from 4 pooled female Murrah buffalo samples using ddRAD approach. In the study by Mishra *et al*. (2020), A total of 683580 putative SNPs were obtained from 85 animals based on variations in genomic regions with respect to four buffalo traits.

5.2.3 Quality control of raw reads:

After QC, total 367.2 million reads (92.3% of total reads) of 135 bp length were used for downstream processing. The raw reads were filtered using the PRINSEQ software program previous studies (Upadhyay *et al*., 2015; Patel *et al*., 2017) for SNP identification in buffalo via targeted sequencing, which resulted in further loss of reads at filtration stage. STACKS software program was used in the current analysis, which is comparatively better than PRINSEQ for Reduced Representation approach since STACKS will check the mean quality score at the window level (default window size is 15 bp), while PRINSEQ software takes into account the mean quality score of the reads that may result in more reads being lost. Besides providing the basic QC functions, the Restriction Enzyme (RE) sites and rad-tags are also tested by STACKS.

5.2.4 Alignment:

99.82% of the reads aligned to the reference genome (UOA_WB_1). The mapping rate was comparatively higher than targeted sequencing method used in previos studies (Upadhyay *et al*., 2015; Menon *et al*., 2016; Patel *et al*., 2017). This may have been due to the ddRAD sequencing and the newest version of Buffalo reference assembly used in this study.

5.2.5 Variant calling:

A greater number of high quality SNPs were identified through reduced representation approach in the present study compared to the study of Immartino et al., (2013), where an average of 2.12 million reads per sample for 48 water buffalo were used to identify the SNPs at minimum base quality score of 30. A total of 49,607 SNP with quality $\geq$ 5 and 10,335 SNP with quality $\geq$ 10 through GBS using the water buffalo reference genome.

In Indian buffalo, Upadhyay *et al*. (2015) studied 23 buffalo through targeted sequencing method which were separated into three groups viz., Fertility trait group (8 animals), General Health trait group (8 animals), Production trait group (7 animals) for identifying SNVs. A total of 2838 SNVs from general health trait group, 36743 SNVs from the fertility trait group and 56996 SNVs from the production trait group were identified using *Bos taurus* reference genome.

Menon *et al*. (2016), studied a total of 12 water buffalo from three breeds viz., Banni, Mehsani, and Jaffarabadi through targeted sequencing method for identification and annotation of SNPs responsible for high and low milk producing buffalo groups. They reported 1.203 million SNPs in high milk yield group and 1.315 million SNPs in low milk yield group using *Bos taurus* reference genome at read depth $\geq$ 5 and quality score $\geq$ 30.

Patel *et al*., (2017), reported a total of 477996 SNPs at read depth $\geq$ 5 and quality $\geq$ 25 in 24 samples from three breeds viz., Banni, Mehsani and Jaffarabadi through targeted genome sequencing method using the reference genome of *Bos taurus* genome build 4.6.1. The authors reported that high number of SNPs identified in regulatory regions, which may lead to conformational changes in transcription factor-binding sites, which play crucial role in gene expression in low milk producing animals.

The reduction in the number of SNPs in the present study compared to Patel *et al*. (2017) may be due to the stringent minimum read depth (>10) used in this study.

The T_S/T_V ratio was applied to check the potential false identification of the SNPs which was found to be 2.5578 in the present study. The T_S/T_V ratio for whole genome sequencing

ranging from 2.0 - 2.1 (1000 genome project) and for exome target regions ranges from 2.6 - 3.5 (http://www.1000genomes.org). When the genome targeted methods were applied, the T_S/T_V ratio was more than the whole genome sequencing method (Upadhyay *et al*., 2015; Menon *et al*., 2016; Patel *et al*., 2017). In the present study the genome was reduced using restriction enzyme digestion method to exclude the repetitive sequences which resulted in increase of T_S/T_V ratio.

From the above comparisons, using ddRAD method was found to be more efficient resulting in more number of high quality SNPs. Since buffalo and cattle were closely related before 5-10 million years ago (Upadhyay *et al*., 2015) they might have developed high difference at genome level during the course of time.

5.3 Annotation

In this study, the genes involved in different traits were downloaded from CattleQTLdb. SNPeff results were used to assign the variants to each gene. Total number of genes was 9, 13, 6, 190 and 213 that were annotated for calving interval, 305 days daily milk yield, age at first calving, conception rate and milk fat percentage, respectively. Using ddRAD, several variants with various impact like high, low, moderate and modifier were identified in genes involved in economic traits, which may be used in association studies with phenotypes or to discover selection signatures in future.

5.4 Genetic Diversity Analysis

Among the different purebred and crossbred buffalo populations, genetic diversity and population structure are two essential aspects. These assist in genetic improvement by optimizing existing diversity-based breeding plans; enhancing local environmental adaptation of these populations; maximizing and optimizing production performance; and conservation of these breeds (Groenebeld *et al*., 2010). Initially, the genetic diversity studies performed in India for buffalo have depended mainly on the use of microsatellite markers (Kataria *et al*., 2009; Tantia *et al*., 2006). Recently, high-density genotype markers have successfully been used to reveal the genetic architecture and diversity of Indian cattle and buffalo (Dash *et al*. 2017; Surya *et al*., 2018).

5.4.1 Population Diversity parameters

The observed (Ho) and expected heterozygosity (He) was found highest in Murrah breed (0.2372 and 0.2462), respectively (Table 4.2). These results were as Murrah is the best breed of buffalo for milk production in India and present in large numbers in the field. In Toda breed (Ho; 0.2150 and He; 0.2111), the lowest value of this parameter was observed, suggesting that inbreeding in conjunction with a small population size resulted in loss of variation within the breed. The study done by Shokrollahi et al. (2009) using microsatellites, reported that the expected heterozygosity in the population of Iranian buffalo varied from 0.69 (Guilani) to 0.78 (Khuzestani). Colli et al. (2018) recently did one of the most comprehensive studies on genomic diversity in buffalo by integrating the buffalo populations around the world. In their analysis, the average heterozygosity found for different buffalo populations ranged from 0.302 (swamp buffalo) to 0.481 (Aza-Kheli), high heterozygosity (0.386) and low to medium inbreeding (0.045) for the Khuzestani river buffalo.

5.4.2 Runs of Homozygosity (ROH)

The lengths of homozygous genotypes that were > 1000 kb and contained only one heterozygous genotype were classified as ROH. In view of the strong linkage imbalance (LD) between SNPs with a distance of up to 100 Kb (Mokhber *et al*., 2019), short homozygous haplotypes in the buffalo genome are expected to be prevalent. Thus, to be considered as ROH, a genome segment had to consist of minimum 70 SNP and be 1000 kb in length (as described in materials and methods) to avoid detecting small and prevalent haplotypes as ROH. In comparison to human populations, livestock species typically have higher autozygosity levels and longer ROHs (Peripolli *et al*., 2018, McQuillan *et al*., 2008, Mastrangelo *et al*., 2016). Genotyping errors, however, can still affect the quality of ROH calling (Ceballos *et al*., 2018). Therefore, we allowed one heterozygous SNP in ROH to avoid losing particularly long ROH because of a single genotyping error (Szmatola *et al*., 2016; Ghozeishifar *et al*., 2019).

As presented in Table 4.5, The average ROH length detected in Bhadawari, Pandharpuri, Surti and Toda was 2 Mb whereas 1 Mb in Murrah and 3 Mb in Mehsana. The proportion of different ROH lengths may be used as an measure of the amount of past generations

in which inbreeding has occurred, since the chromosomes may be rearranged through recombination events and the ROH length can be decreased. Thus, recent inbreeding results in longer ROH because of long IBD stretches. In contrast, as a result of ancient inbreeding, short ROHs occur because the long IBD segments are broken down in meiosis through generations. We detected ROH with an average length from 1 to 3 Mb from all of the samples, which might indicate that some inbreeding events occurred about 20-25 generations ago (Howrigan *et al*., 2011). However, because of the smaller number of samples from Toda, our findings should be viewed with caution.

Due to their varying distances from the last common ancestor, animals with the same cumulative length of ROH presented different numbers of ROH with distinct lengths (Mastrangelo *et al*., 2017). The total genomic length (Mb) covered by ROH per individual was roughly proportional to the average number of ROH per individual; the total number of ROHs increased synchronously with an increase in the total length of individual ROHs (Except in Toda, the total number of ROH decreases with an increase in the total length of individual ROH, may be due to the very small sample size as compared to the sample size of other breeds. So, we took average number of ROH for comparison) (fig. 4.3). To some degree, these findings could reflect the level of inbreeding or the variations between different populations in population history, where the higher the level of inbreeding, the greater the number of ROH in the genome, and the longer the total length of ROH. Some extreme individuals with ROH lengths exceeding 500 Mb were identified among the Bhadawari, Mehsana and Pandharpuri with the longest total ROH length in an individual was 621.34 Mb (Bhadawari) (Table 4.4). This outcome represented the lack of effective management of inbreeding in the population. Murrah and Mehsana are dairy breeds that have been subjected to systematic breeding; Toda Buffaloes are a local breed that has not been subjected to systematic breeding. Generally, the F_{ROH} of Toda should be smaller than those of Murrah and Mehsana but the results of this study indicated that the F_{ROH} of Toda was high (0.15) to that of other breeds (Table 4.5). This outcome may be due to the small effective population size, such that recent generations may have resulted in a high degree of inbreeding from the small effective population (Curik *et al*., 2014). The inbreeding coefficient of the few animals measured as F_{HOM} was negative, which

may have been due to the limited number of animal samples in our research, and random sampling errors may also have resulted in a negative result (Purcell *et al*., 2007) (Fig. 4.2).

5.4.3 Linkage Disequilibrium (LD) and Effective population size (Ne)

The decision to use r^2 instead of 'D' to test LD measurements was due to the fact that in a finite population size, it is less influenced by allele frequencies compared to 'D', which appears to overestimate LD in small samples and low frequency alleles (Lu *et al*., 2012, Hedrick., 1987, Ardlie *et al*., 2002). Mean r^2 values above 0.30 can be considered as a strong LD and useful for QTL mapping, according to the literature (Ardlie *et al*., 2002), while mean r^2 values above 0.20 are considered adequate for genomic breeding value (GBV) estimation to achieve an accuracy of 0.85 (Meuwissen *et al*., 2001). Mean values of $r^2 = 0.30$ and above were found in all the breeds taken in our study at a genomic distance of 10 Kb except Murrah.

Ne is one of the most important criteria for the management of genetic resource conservation, the performance of breeding programmes and the improvement of artificial selection design. The minimum level recommended by FAO (2007) for the effective population size of at least 50 animals is sufficient to retain genetic variation and to prevent inbreeding depression in different populations. In addition, in order to preserve initial evolutionary potential in perpetuity, the Ne should be over 1000. It is assumed that the breed Mehsana was developed from Murrah and Surti buffalo a few centuries ago (less than 100 generations may have finished). So, we did not indicate the Ne value for the Mehsana buffalo in the table 4.7. Therefore, the findings should be interpreted in the context of theoretical assumptions.

A downward trend for Ne over the generations was observed for all the breeds in this study, where Murrah (68) had the highest, and Toda had the least recent (11) Ne. The majority of research on Ne estimates in livestock have concentrated on cattle, in particular dairy cattle. Biegelmeyer *et al*. (2016) found that in the Herford and Braford breeds, the Ne was 153 and 220, respectively. Zhu *et al*. (2013) demonstrated that in the Chinese Simmental breed, the Ne was 73. Qanbari *et al*. (2010) recorded that four generations ago, in the German Holstein breed, the Ne was close to 103. Karimi *et al*. (2016) stated that this value ranged from 13

(Sarabi) to 107 (Mazandarani) in order to estimate the Ne in indigenous cattle breeds of Iran. Deng *et al*. (2019) recorded the decreased Ne trend across the purebred and crossbred buffalo population from 1,000 to 100 generations ago, demonstrating the impact of historical domestication, breed development and artificial selection process. In comparison, Fallahi *et al*. (2019) showed Ne of 422 in the Azerbaijani river buffalo population, suggesting adequate population diversity. In the study completed on the Khuzestani River buffalo, Davoudi et al. (2020) found a Ne value of 240. Using pedigree data, Santana *et al*. (2011) studied Ne on milk buffalo in Brazil (Murrah buffalo) and reported a very low amount of Ne (n = 40). Despite this small amount in the Ne in the Murrah buffalo, Malhado *et al*. (2013) stated that in the other Brazilian buffalo breeds (Jafarabadi buffalo), the Ne is 10. In order to avoid high inbreeding in these populations, it appears that a specific mating system should be strictly considered.

The disparity in Ne describes the incidents that have happened in the history of the population. Therefore, the Ne provides valuable conservation information, especially for indigenous breeds. Although the Ne in this study was found higher for Murrah (n = 68), it is the only breed in our study which had more number of population than prescribed by the FAO (2007). Rest of the breeds in our study had a value lower than the FAO recommended measure. Due to the demand for high yield of milk and meat production in recent years. The Buffalo population has increasingly being brought under selective breeding. Therefore, Ne has shown a more serious decline in recent years. For this purpose, various strategies, such as designing suitable mating systems and developing gene pools, need to be considered.

5.4.4 Haplotype blocks

The magnitude of haplotype blocks between different autosomes differed significantly. In this study, the average Chromosome coverage by block (%) in the Bhadawari, Mehsana, Murrah, Pandharpuri, Surti was 15.82, 26.84, 15.40, 16.93, 18.65 respectively, which was higher than the Khuzestani river buffalo (8.2/ %) studied by Davoudi *et al*. (2020) except Toda which had a coverage of 4.95%. In this research, the average block coverage of the genome was higher than that of many cattle studies, which corresponds to the lower genetic

diversity in the Buffalo breeds under study. For example, Qanbari *et al*. (2010) found that haplotype blocks in the German Holstein cattle breed covered 4.7 % of the genome; Salem *et al*. (2018) recorded that this rate in the Portuguese Holstein cattle breed was 6.2 %.

The total number of blocks identified in this study for Bhadawari, Mehsana, Murrah, Pandharpuri, Surti, Toda was 7615, 8379, 11642, 7792, 7747, 1363 respectively, which was higher than in Khuzestani river buffalo (n = 1726) (Doudani *et al*., 2020), the Azerbaijani buffalo (n = 1693) (Fallahi *et al*., 2019), the Portuguese Holstein (n/ =/ 969) (Salem *et al*., 2018), and German Holstein (n/ =/ 712) (Qanbari *et al*., 2010).

There are many factors influencing the characteristics of the haplotype blocks, including breed, marker types, marker density, chromosome region and the haplotype block definition process, according to previous findings. Since the haplotypes have a stronger LD with Quantitative Trait Loci (QTL) compared to individual SNPs, the use of haplotype blocks in genomic and GWAS studies has always been recognised as useful techniques (Sun *et al*., 2016). By understanding the structure of haplotypes, the GWAS and GS studies can be successfully planned and interpreted in livestock populations. The GWAS studies showed that the haplotype association test rather than the single SNPs analysis model could increase the accuracy of detecting candidate genes. The prediction accuracy in the GS could be increased using suitable covariates for haplotype alleles compared to the SNPs, according to a report by Hess *et al*. (2017), which can be result in improved genetic advantage by changing the ranking of selection candidates.

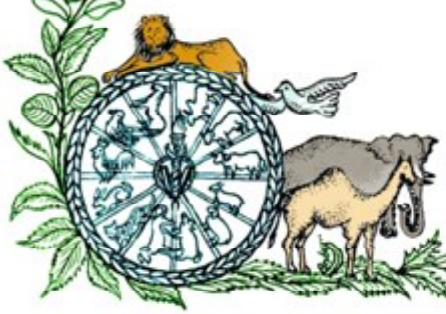
5.4.5 PCA, Phylogenetic and Admixture Analysis:

The results of the PCA (PC1 & PC2) revealed the higher amount of genetic similarities among Murrah and Bhadawari forming a separate cluster with few samples of Surti and Toda buffalo near this cluster, while Surti, Pandharpuri and Mehsana showed greater genetic differentiations with three distinct clusters (fig. 4.15 & fig. 4.16). PC3 separated Murrah and Bhadawari individuals. Also, the crossbred Mehsana was clustered between its parent breeds Murrah and Surti along PC3. In this study, Mehsana, Surti and Pandharpuri grouped in separate clusters, likewise the study done by Thakor *et al*. (2018) also kept Surti and Pandharpuri in different clusters. However, it was shown in single cluster by Kumar *et al*. (2006).

In the Admixture analysis, at K = 2, Mehsana was separated from the rest of the breeds and showed introgression of Murrah inheritance. Mehsana also showed small amount of admixture with Surti when the K value was increased beyond 3. Pandharpuri clearly separated at K = 3 and showed very little admixture of Murrah and Surti in further K values. It is also supported by the Principal component analysis and Tree mix analysis where Pandharpuri formed a separate lineage. At the value of K = 6, each breed separated into their own clusters with the varying level of Murrah ancestry in these breeds which may be due to use of Murrah buffalo in every region of India. Same result had been found by the study of Thakor *et al.* (2018) in which they found that Surti and Pandharpuri were forming a separate cluster while Murrah and Mehsana were showing admixture. The reason behind is that these two breeds have been most popular amongst the buffalo breeds in terms of high milk yield. Murrah semen has been extensively and indiscriminately used for artificial insemination (AI) across the country while Toda, Bhadawari, Surti and Pandharpuri are less in number and been less or not utilized for insemination throughout the country, which has led to a steady decline in the genetic diversity present in the less characterized populations.

There is a hypothesis that Mehsana breed has been developed using Murrah bulls on local Surti buffaloes (Pundhir *et al*., 2000). Although the Admixture software could not separate the Murrah and Surti inheritance in Mehsana, the PCA (along PC3) results and the Treemix analysis where Mehsana is in the same branch as Surti, and shows Murrah introgression, supports this hypothesis.

Summary and Conclusions



6 SUMMARY AND CONCLUSIONS

A total of 397800000 raw reads of 150 bp length were generated by the sequencer were obtained from the 96 samples of the 6 different breeds in 96 plex libraries obtained from restriction digestion based methods using Sph I & MluC I restriction enzymes. After the QC, a total of 367200000 reads were obtained (92.3% of total raw reads). 99.82% of the reads aligned to the reference genome (Bubalus bubalis assembly UOA_WB_1). Total 569,535 variants were discovered, out of which 502,476 were SNP and 67,059 were indels. 551,458 variants were present on autosomes, 15315 on the X chromosome and 12 on the mtDNA (NC_006295.1). 2750 variants were located on unmapped contigs. A variant was discovered for every 4,637 bp of the genome length. When compared to the reference genome, a total of 484449, 473909, 489738, 469311, 470603 and 448714 SNPs; 59028, 57658, 59308, 57009, 56887 and 53859 INDELs were identified at RD10 in Murrah, Bhadawari, Mehsana, Pandharpuri, Surti and Toda, respectively.

The genome wide SNPs identified by Reduced Representation Approach in Buffalo breeds using Bubalus bubalis assembly UOA_WB_1. The high-quality SNPs identified against Water buffalo reference genome at read depth $\geq$ 10 and quality $\geq$ 30 were annotated using the SNPEFF software. The variants of different genes affecting the different traits were identified across the genome of Water buffalo. The Ts/Tv ratio was 2.5578 in this study. The genes involved in different traits were downloaded from CattleQTLdb and SNPeff results were used to assign the variants to each gene. Gene wise annotation of Variants for the candidate genes was done in which total 9 genes for Calving interval, 13 genes for 305 days milk yield, 6 genes

for Age at first calving, 190 genes for Conception rate and 219 genes for Milk fat (%) was annotated. The SNPs found in this study was further use in the genetic diversity study and for the phylogeny analysis.

The annotated variants were subjected to further filters before diversity analysis. Non-autosomal and unmapped SNPs were removed. SNPs missing in more than 25% individuals and below the MAF threshold of 0.01 were also filtered out. Indels were removed. Runs of homozygosity (ROH) in the autosomes of 96 buffalo animals were determined using PLINK1.9 and consisted of *237,762* SNPs across all the breeds after quality control. The average total length of ROH per animal were 2, 3, 1, 2, 2 & 2 Mb and Average numbers of ROH per animal were 107, 111, 114, 118, 124 & 180 in the Bhadawari, Mehsana, Murrah, Pandharpuri, Surti and Toda breeds of Buffalo, respectively. The results of distribution of ROH revealed that ancient and recent inbreeding have had an influence on the genome of the buffalo breeds taken in our study. In this study, the LD values are generally larger in Toda and making major differences with other breeds. We found higher Ne for Murrah at 5 generation ago (Ne = 68) which was more than the recommended value of FAO (Ne = 50). We note that Pandharpuri, Surti and Bhadawari had a roughly similar LD decay pattern. Result shows that Murrah has had the greatest genetic diversity over the generations among all the sampled populations, while Toda had the least. This study has reported the characterization of haplotype blocks and tagged SNPs.

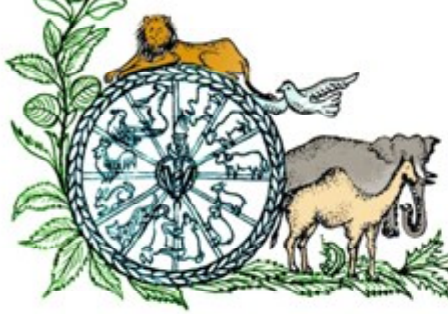
For PCA, Treemix and Admixture analysis, additional filtering for markers in linkage disequilibrium was performed using PLINK. Each chromosome was scanned through a 50 kb sliding window with a step size of 5 variants. SNPs above $r^2 > 0.2$ were removed, resulting in a dataset of 67,798 SNPs. In PCA, Mehsana is separated from the Murrah and Surti. In PC2, few samples of Mehsana are nearby to the cluster of Murrah. In Admixture analysis, at K = 3, Pandharpuri forms a separate cluster showing some part of the introgression of Murrah and no introgression of Surti was shown in Mehsana at any K value. The formation of separate cluster was also supported by Tree mix analysis in which Pandharpuri forming a separate lineage from the other breeds taken in the study.

Mini Abstract

7 MINI ABSTRACT

Next Generation Sequencing (NGS) based Restriction-site Associated DNA sequencing (RAD Seq) methods are an alternative to Whole Genome Sequencing (WGS), wherein simultaneous sequencing, genotyping and multiplexing are facilitated. WGS for SNP genotyping is more expensive and a reduction of about 35 fold in cost is possible through RAD seq methods as compared to WGS method. The present study was carried out to identify the genome wide SNPs and INDELs, to annotate the SNPs related to Production and Reproduction traits and genetic diversity analysis in Genome of Riverine buffaloes (Murrah, Mehsana, Bhadawari, Pandharpuri, Surti and Toda). A total of 397.8 million raw reads from 96 buffalo samples were obtained using restriction enzyme digestion and sequencing with Illumina Hiseq 2000. The raw reads were quality filtered which yielded a total of 367.2 million good quality reads (92.3%) and were aligned with Bubalus bubalis assembly (UOA_WB_1) which resulted in 99.82% alignment. The candidate genes affecting Milk production (305 Days Milk yield, Milk Fat Percentage) and Reproduction (Age at first calving, Conception Rate, Calving interval) were annotated. A total of 9 genes affecting Calving interval; 13 genes affecting 305 DMY; 6 genes affecting AFC, 123 genes affecting conception rate and 129 genes affecting Milk fat percentage were annotated. The observed (Ho) and expected heterozygosity (He) was found highest in Murrah breed (0.2372 and 0.2462). The results of distribution of ROH revealed that ancient and recent inbreeding have had an influence on the genome of the buffalo breeds taken in our study. All the breeds except Murrah showing the strong LD at a distance of <10 Kb. We found higher Ne for Murrah at 5 generation ago (Ne = 68) which was more than the recommended value of FAO (Ne = 50). This study has reported the characterization of haplotype blocks and tagged SNPs. According to PCA and Admixture analysis, Mehsana showing the introgression of Murrah not of Surti as per the hypothesis. Pandharpuri forming a separate lineage in phylogenetic analysis which was strongly supported by PCA and Admixture analysis. The SNPs identified and annotated in the present study may be used for genotyping in larger number of samples for further association studies.

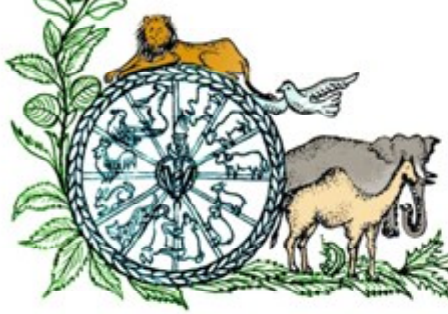
लघु सारांश



8 लघु सारांश

अगली पीढ़ी की अनुक्रमण (NGS) आधारित प्रतिबंध–साइट एसोसिएटेड डीएनए अनुक्रमण (RAD Seq) विधियाँ पूरे जीनोम अनुक्रमण (WGS) का एक विकल्प हैं, जिसमें एक साथ अनुक्रमण, जीनोटाइपिंग और बहुसंकेतन की सुविधा होती है। एसएनपी जीनोटाइपिंग के लिए WGS अधिक महंगा है और WGS विधि की तुलना में RAD seq विधियों के माध्यम से लागत में लगभग 35 गुना सस्ता है। वर्तमान अध्ययन जीनोम वाइड एसएनपी और INDELs की पहचान करने के लिए किया गया था, जेनोम में रिवराइन भैसों (मुर्रा, मेहसाणा, भसावरी, पंढरपुरी, सुरती और टोडा) के उत्पादन और प्रजनन विविधता से संबंधित एसएनपी की व्याख्या करने के लिए किया गया था। 96 भैंस के नमूनों से कुल 397.8 मिलियन कच्चे रीड्स इल्लुमिना हिसक 2000 के साथ प्रतिबंध एंजाइम पाचन और अनुक्रमण का उपयोग करके प्राप्त किए गए थे। कच्ची रीड गुणवत्ता वाले फिल्टर किए गए थे, जिनकी कुल 367.2 मिलियन अच्छी गुणवत्ता रीड (92.3%) आयी थी और उन्हें बुबलस बुबलिस असेंबली (UOA_WB_1) के साथ जोड़ा गया था, जिसके परिणामस्वरूप 99.82% संशोधन हुआ था। दुग्ध उत्पादन (305 दिन की दूध की उपज, मिल्क फैट का प्रतिशत) और प्रजनन (पहले ब्यांत होने पर आयु, गर्भाधान दर, बछड़ा अंतराल) को प्रभावित करने वाले उम्मीदवार जीन को एनोटेट किया गया था। बछड़ा अंतराल को प्रभावित करने वाले कुल 9 जीन, 305 डीएमवाई को प्रभावित करने वाले 13 जीन, एएफसी को प्रभावित करने वाले 6 जीन, गर्भाधान दर को प्रभावित करने वाले 123 जीन और दूध के वसा प्रतिशत को प्रभावित करने वाले 129 जीन को एनोटेट किया गया था। देखे गए (Ho) और अपेक्षित विषमयुग्मजी (He) मुर्राह नस्ल (0.2372 और 0.2462) में सबसे अधिक पाया गया। आरओएच के वितरण के परिणामों से पता चला है कि प्राचीन और हाल ही में इनब्रीडिंग का हमारे अध्ययन में ली गई भैंस नस्लों के जीनोम पर प्रभाव पड़ा है। मुर्राह को छोड़कर सभी नस्लों <10 Kb की दूरी पर मजबूत एलडी दिखा। हमने 5 पीढ़ी पहले (Ne = 68) पर मुर्रा के लिए उच्च Ne पाया, जो FAO (Ne = 50) के अनुशंसित मूल्य से अधिक था। इस अध्ययन में हैप्लोटाइप ब्लॉक के लक्षण वर्णन और एसएनपी को टैग किया गया है। PCA और Admixture के विश्लेषण के अनुसार, मेहसाणा ने परिकल्पना केक अनुसार सुरति का नहीं मुर्रा का अंतर्ज्ञान दिखाया। पंढरपुरी ने फाइलोजेनेटिक विश्लेषण में एक अलग वंश का गठन किया जो पीसीए और एडमिक्सचर विश्लेषण द्वारा दृढ़ता से समर्थित था। वर्तमान अध्ययन में एसएनपी की जो पहचान और व्याख्या की गयी है, उसका उपयोग आगे के अध्ययन के लिए बड़ी संख्या में नमूनों में जीनोटाइपिंग के लिए किया जा सकता है।

References



9 REFERENCES

Ajmone-Marsan, P., Colli, L., Han, J.L., Achilli, A., Lancioni, H., Joost, S., Crepaldi, P., Pilla, F., Stella, A., Taberlet, P. and Boettcher, P., 2014. The characterization of goat genetic diversity: Towards a genomic approach. Small Ruminant Res, **121**(1), pp.58-72.

Alexander, D.H., Novembre, J. and Lange, K., 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome research, **19**(9), pp.1655-1664.

Anderson, E.C., 2008. Bayesian inference of species hybrids using multilocus dominant genetic markers. Philosophical Transactions of the Royal Society B: Biol. Sci, **363**(1505), pp.2841-2850.

Antao, T., Pérez Figueroa, A. and Luikart, G., 2011. Early detection of population declines: high power of genetic monitoring using effective population size estimators. Evolutionary Applications, **4**(1), pp.144-154.

Ardlie KG, Kruglyak L, Seielstad M: Patterns of linkage disequilibrium in the human genome. Nat Rev Genet 2002, 3:299-309.

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A. and Johnson, E.A., 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PloS one, **3**(10), p.e3376.

De Bakker, P.I., McVean, G., Sabeti, P.C., Miretti, M.M., Green, T., Marchini, J., Ke, X., Monsuur, A.J., Whittaker, P., Delgado, M. and Morrison, J., 2006. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nat. Genet, **38**(10), pp.1166-1172.

Barbato, M., Orozco-terWengel, P., Tapio, M. and Bruford, M.W., 2015. SNeP: a tool to estimate trends in recent effective population size trajectories using genome-wide SNP data. Front genet, 6, p.109.

Barczak, E., Wolc, A., Wójtowski, J., Slósarz, P. and Szwaczkowski, T., 2009. Inbreeding and inbreeding depression on body weight in sheep. J. Anim. Feed Sci, **18**(1), pp.42-50.

Barton, N.H. and Charlesworth, B., 1984. Genetic revolutions, founder effects, and speciation. Annu Rev Ecol Evol S, **15**(1), pp.133-164.

Biegelmeyer, P., Gulias-Gomes, C.C., Caetano, A.R., Steibel, J.P. and Cardoso, F.F., 2016. Linkage disequilibrium, persistence of phase and effective population size estimates in Hereford and Braford cattle. BMC Genet, **17**(1), p.32.

Borghese, A., 2013. Buffalo livestock and products in Europe. Buffalo Bulletin, 32 (Special Issue 1), pp.50-74.

Bouaziz, M., Ambroise, C. and Guedj, M., 2011. Accounting for population stratification in practice: a comparison of the main strategies dedicated to genome-wide association studies. PloS one, **6**(12), p.e28845.

Bryc, K., Auton, A., Nelson, M.R., Oksenberg, J.R., Hauser, S.L., Williams, S., Froment, A., Bodo, J.M., Wambebe, C., Tishkoff, S.A. and Bustamante, C.D., 2010. Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proc. Natl. Acad. Sci, **107**(2), pp.786-791.

Ceballos, F.C., Joshi, P.K., Clark, D.W., Ramsay, M. and Wilson, J.F., 2018. Runs of homozygosity: windows into population history and trait architecture. Nat. Rev. Genet, **19**(4), p.220.

Charlesworth, B., 2013. Background selection 20 years on: the Wilhelmine E. Key 2012 invitational lecture. J Hered, **104(**2), pp.161-171.

Choi, J.W., Liao, X., Stothard, P., Chung, W.H., Jeon, H.J., Miller, S.P., Choi, S.Y., Lee, J.K., Yang, B., Lee, K.T. and Han, K.J., 2014. Whole-genome analyses of Korean native and Holstein cattle breeds by massively parallel sequencing. PloS one, **9**(7), p.e101127.

Ciani, E., Lasagna, E., D'andrea, M., Alloggio, I., Marroni, F., Ceccobelli, S., Bermejo, J.V.D., Sarti, F.M., Kijas, J., Lenstra, J.A. and Pilla, F., 2015. Merino and Merino-derived sheep breeds: a genome-wide intercontinental study. Genet. Sel. Evol, 47(1), pp.1-12.

Clark, E.L., Bush, S.J., McCulloch, M.E., Farquhar, I.L., Young, R., Lefevre, L., Pridans, C., Tsang, H., Wu, C., Afrasiabi, C. and Watson, M., 2017. A high resolution atlas of

gene expression in the domestic sheep (Ovis aries). PLoS genet, **13**(9), p.e1006997.

Cockrill, W.R., 1981. The water buffalo: a review. Br. Vet., **137**(1), pp.8-16.

Colli, L., Milanesi, M., Vajana, E., Iamartino, D., Bomba, L., Puglisi, F., Del Corvo, M., Nicolazzi, E.L., Ahmed, S.S., Herrera, J.R. and Cruz, L., 2018. New insights on water buffalo genomic diversity and post-domestication migration routes from medium density SNP chip data. Front genet, 9, p.53.

Curik, I., Ferenèakoviæ, M. and Sölkner, J., 2014. Inbreeding and runs of homozygosity: a possible solution to an old problem. Livest Sci, 166, pp.26-34.

Dadi, H., Tibbo, M., Takahashi, Y., Nomura, K., Hanada, H. and Amano, T., 2008. Microsatellite analysis reveals high genetic diversity but low genetic structure in Ethiopian indigenous cattle populations. Anim Genet, **39**(4), pp.425-431.

Daetwyler, H.D., Capitan, A., Pausch, H., Stothard, P., Van Binsbergen, R., Brøndum, R.F., Liao, X., Djari, A., Rodriguez, S.C., Grohs, C. and Esquerré, D., 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet, **46**(8), p.858.

Das, A., Panitz, F. and Holm, L., 2014, August. Identification and annotation of genetic variants (SNP/Indel) in Danish Jutland cattle. In Vancouver: 10th World Congress on Genetics Applied to Livestock Production (pp. 17-22).

Davey, J.W., Hohenlohe, P.A., Etter, P.D., Boone, J.Q., Catchen, J.M. and Blaxter, M.L., 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet, **12**(7), p.499.

Davoudi, P., Moradi-Shahrbabak, H., Mehrabani-Yeganeh, H., Ghoreishifar, S.M., Gholami, S. and Abdollahi-Arpanahi, R., 2020. Exploring the structure of haplotype blocks, runs of homozygosity and effective population size in khuzestani river buffalo. Slovak J. Anim. Sci. **53(**02), pp.67-77.

Decker, J.E., McKay, S.D., Rolf, M.M., Kim, J., Alcalá, A.M., Sonstegard, T.S., Hanotte, O., Götherström, A., Seabury, C.M., Praharani, L. and Babar, M.E., 2014. Worldwide patterns of ancestry, divergence, and admixture in domesticated cattle. PLoS Genet, **10**(3), p.e1004254.

Deng, T., Liang, A., Liu, J., Hua, G., Ye, T., Liu, S., Campanile, G., Plastow, G., Zhang, C., Wang, Z. and Salzano, A., 2019. Genome-Wide SNP data revealed the extent of

linkage disequilibrium, persistence of phase and effective population size in purebred and crossbred buffalo populations. Front genet, 9, p.688.

De Donato, M., Peters, S.O., Mitchell, S.E., Hussain, T. and Imumorin, I.G., 2013. Genotyping-by-sequencing (GBS): a novel, efficient and cost-effective genotyping method for cattle using next-generation sequencing. PloS one, **8(**5), p.e62137.

Eck, S.H., Benet-Pagès, A., Flisikowski, K., Meitinger, T., Fries, R. and Strom, T.M., 2009. Whole genome sequencing of a single Bos taurus animal for single nucleotide polymorphism discovery. Genome Biol, **10**(8), p.R82.

Elhaik, E., Tatarinova, T., Chebotarev, D., Piras, I.S., Calò, C.M., De Montis, A., Atzori, M., Marini, M., Tofanelli, S., Francalacci, P. and Pagani, L., 2014. Geographic population structure analysis of worldwide human populations infers their biogeographical origins. Nat. Commun, **5**(1), pp.1-13.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S. and Mitchell, S.E., 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PloS one, **6**(5), p.e19379.

Falconer, D.S. and Mackay, T.F.C., 1996. Introduction to quantitative genetics. 1996. Harlow, Essex, UK: Longmans Green, **3**.

Fallahi, M.H., Shahrbabak, H.M., Shahrbabak, M.M., Arpanahi, R.A. and Gholami, S., 2019. Detection of Haplotypic structure for genome of Azerbaijani Buffalo using high density SNP markers. Russian J Genet, **55**(8), pp.1000-1007.

Falush, D., Stephens, M. and Pritchard, J.K., 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genet, **164**(4), pp.1567-1587.

Frankham, R., Ballou, J.D. and Briscoe, D.A., 2002. Conservation genetics. Cambridge.

Freeman, A.R., Bradley, D.G., Nagda, S., Gibson, J.P. and Hanotte, O., 2006. Combination of multiple microsatellite data sets to investigate genetic diversity and admixture of domestic cattle. Anim Genet, **37**(1), pp.1-9.

Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M. and Liu-Cordero, S.N., 2002. The structure of haplotype blocks in the human genome. Science, **296**(5576), pp.2225-2229.

Ghoreishifar, S.M., Moradi-Shahrbabak, H., Parna, N., Davoudi, P. and Khansefid, M., 2019. Linkage disequilibrium and within-breed genetic diversity in Iranian Zandi sheep. Arch. Anim. Breed, **62**(1), p.143.

Gibson, J., Morton, N.E. and Collins, A., 2006. Extended tracts of homozygosity in outbred human populations. Hum mol genet, **15**(5), pp.789-795.

Gontcharov, A.A., Marin, B. and Melkonian, M., 2004. Are combined analyses better than single gene phylogenies? A case study using SSU rDNA and rbc L sequence comparisons in the Zygnematophyceae (Streptophyta). Mol. Boil. Evol, **21**(3), pp.612-624.

Groeneveld, L.F., Lenstra, J.A., Eding, H., Toro, M.A., Scherf, B., Pilling, D., Negrini, R., Finlay, E.K., Jianlin, H., Groeneveld, E.J.A.G. and Weigend, S., 2010. Genetic diversity in farm animals–a review. Anim Genet, 41, pp.6-31.

Guo, Y., Long, J., He, J., Li, C.I., Cai, Q., Shu, X.O., Zheng, W. and Li, C., 2012. Exome sequencing generates high quality data in non-target regions. BMC genomics, **13**(1), p.194.

Hanotte, O., Dessie, T. and Kemp, S., 2010. Time to tap Africa's livestock genomes. Sci, **328(**5986), pp.1640-1641.

Hayes, B.J., Visscher, P.M., McPartlan, H.C. and Goddard, M.E., 2003. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. Genome res, **13**(4), pp.635-643.

Hess, G.T., Tycko, J., Yao, D. and Bassik, M.C., 2017. Methods and applications of CRISPR-mediated base editing in eukaryotic genomes. Mol. cell, **68**(1), pp.26-43.

Hill, W.G. and Robertson, A., 1968. Linkage disequilibrium in finite populations. Theor appl genet, **38**(6), pp.226-231.

Howrigan, D.P., Simonson, M.A. and Keller, M.C., 2011. Detecting autozygosity through runs of homozygosity: a comparison of three autozygosity detection algorithms. BMC genom, **12**(1), p.460.

Iacolina, L., Stronen, A.V., Pertoldi, C., Tokarska, M., Nørgaard, L.S., Muñoz, J., Kjærsgaard, A., Ruiz-Gonzalez, A., Kamiński, S. and Purfield, D.C., 2016. Novel graphical analyses of runs of homozygosity among species and livestock breeds. Int J genomics, 2016.

Iamartino, D., Strozzi, F., Nicolazzi, E.L., Capoferri, R., Ramelli, P., Pasquariello, R., Stella, A. and Williams, J.L., www. tecnoparco. org.

Jiang, L., Liu, J., Sun, D., Ma, P., Ding, X., Yu, Y. and Zhang, Q., 2010. Genome wide association studies for milk production traits in Chinese Holstein population. PloS one, **5**(10), p.e13661.

Jiang, Z., Wang, H., Michal, J.J., Zhou, X., Liu, B., Woods, L.C.S. and Fuchs, R.A., 2016. Genome wide sampling sequencing for SNP genotyping: methods, challenges and future development. Int J Biol Sci, **12**(1), p.100.

Karimi, K., Koshkoiyeh, A.E., Fozi, M.A., Porto-Neto, L.R. and Gondro, C., 2016. Prioritization for conservation of Iranian native cattle breeds based on genome-wide SNP data. Conserv. Genet, **17**(1), pp.77-89.

Kataria, R.S., Sunder, S., Malik, G., Mukesh, M., Kathiravan, P. and Mishra, B.P., 2009. Genetic diversity and bottleneck analysis of Nagpuri buffalo breed of India based on microsatellite data. Russ. J. Genet, **45**(7), p.826.

Kijas, J.W., Lenstra, J.A., Hayes, B., Boitard, S., Neto, L.R.P., San Cristobal, M., Servin, B., McCulloch, R., Whan, V., Gietzen, K. and Paiva, S., 2012. Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. PLoS biology, **10**(2), p.e1001258.

Kim, E.S. and Kirkpatrick, B.W., 2009. Linkage disequilibrium in the North American Holstein population. Anim Genet, **40**(3), pp.279-288.

Kõks, S., Lilleoja, R., Reimann, E., Salumets, A., Reemann, P. and Jaakma, Ü., 2013. Sequencing and annotated analysis of the Holstein cow genome. Mammalian genome, **24**(7-8), pp.309-321.

Kõks, S., Reimann, E., Lilleoja, R., Lättekivi, F., Salumets, A., Reemann, P. and Jaakma, Ü., 2014. Sequencing and annotated analysis of full genome of Holstein breed bull. Mammalian genome, **25**(7-8), pp.363-373.

Kruglyak, L., 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat. Genet, **22**(2), pp.139-144.

Kumar, S., Gupta, J., Kumar, N., Dikshit, K., Navani, N., Jain, P. and Nagarajan, M., 2006. Genetic variation and relationships among eight Indian riverine buffalo breeds. Mol. Ecol, **15**(3), pp.593-600.

Kumar, P., Freeman, A.R., Loftus, R.T., Gaillard, C., Fuller, D.Q. and Bradley, D.G., 2003. Admixture analysis of South Asian cattle. Heredity, **91**(1), p.43.

Larmer, S., Ventura, R., Buzanskas, M.E., Sargolzaei, M. and Schenkel, F.S., 2014. Assessing admixture by quantifying breed composition to gain historical perspective on dairy cattle in Canada. In Vancouver, Canada: 10th World Congress on Genetics Applied to Livestock Production.

Lencz, T., Lambert, C., DeRosse, P., Burdick, K.E., Morgan, T.V., Kane, J.M., Kucherlapati, R. and Malhotra, A.K., 2007. Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. PNA S, **104**(50), pp.19942-19947.

Liao, X., Peng, F., Forni, S., McLaren, D., Plastow, G. and Stothard, P., 2013. Whole genome sequencing of Gir cattle for identifying polymorphisms and loci under selection. Genome, **56**(10), pp.592-598.

Ligda, C.H., Altarayrah, J., Georgoudis, A. and Econogene Consortium, 2009. Genetic analysis of Greek sheep breeds using microsatellite markers for setting conservation priorities. Small Ruminant Res, **83**(1-3), pp.42-48.

Litt, M., Luty, J.A., 1989. A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am J Hum Genet., 44, pp. 397–401.

Liu, J., Wang, K., Ma, S. and Huang, J., 2013. Accounting for linkage disequilibrium in genome-wide association studies: A penalized regression method. Statistics and its interface, **6**(1), p.99.

Lu, D., Sargolzaei, M., Kelly, M., Li, C., Vander Voort, G., Wang, Z., Plastow, G., Moore, S. and Miller, S., 2012. Linkage disequilibrium in Angus, Charolais, and Crossbred beef cattle. Front. Genet, 3, p.152.

Makina, S.O., Taylor, J.F., van Marle-Köster, E., Muchadeyi, F.C., Makgahlela, M.L., MacNeil, M.D. and Maiwashe, A., 2015. Extent of linkage disequilibrium and effective population size in four South African Sanga cattle breeds. Front. genet, 6, p.337.

Malhado, C.H.M., Ferraz, P.C., Ramos, A.A., Carneir, P., Aragao, E.S., Barbosa, A.C.B. and Carrillo, J.A., 2013. Inbreeding, Average Relatedness Coefficient and Effective Population Size In Jaffarabadi Buffaloes Raised In Brazil. EB, p.641.

Marras, G., Gaspa, G., Sorbolini, S., Dimauro, C., Ajmone Marsan, P., Valentini, A., Williams, J.L. and Macciotta, N.P., 2015. Analysis of runs of homozygosity and their relationship with inbreeding in five cattle breeds farmed in Italy. Animal genet, **46**(2), pp.110-121.

Mastrangelo, S., Tolone, M., Di Gerlando, R., Fontanesi, L., Sardina, M.T. and Portolano, B., 2016. Genomic inbreeding estimation in small populations: evaluation of runs of homozygosity in three local dairy cattle breeds. Animal, **10**(5), pp.746-754.

Mastrangelo, S., Portolano, B., Di Gerlando, R., Ciampolini, R., Tolone, M. and Sardina, M.T., 2017. Genome-wide analysis in endangered populations: a case study in Barbaresca sheep. Animal, **11**(7), p.1107.

McQuillan, R., Leutenegger, A.L., Abdel-Rahman, R., Franklin, C.S., Pericic, M., Barac-Lauc, L., Smolej-Narancic, N., Janicijevic, B., Polasek, O., Tenesa, A. and MacLeod, A.K., 2008. Runs of homozygosity in European populations. Am J Hum Genet, **83**(3), pp.359-372.

McTavish, E.J., Decker, J.E., Schnabel, R.D., Taylor, J.F. and Hillis, D.M., 2013. New World cattle show ancestry from multiple independent domestication events. PNA S, **110(**15), pp.E1398-E1406.

Menon, R., Patel, A. B., & Joshi, C. (2016). Comparative analysis of SNP candidates in disparate milk yielding river buffaloes using targeted sequencing. PeerJ, 4, e2147.

Menozzi, P. and Krimbas, C.B., 1992. The inversion polymorphism of D. subobscura revisited: synthetic maps of gene arrangement frequencies and their interpretation. J. Evol. Biol, **5**(4), pp.625-641.

Meuwissen TH, Hayes BJ, Goddard ME: Prediction of total genetic value using genome-wide dense marker maps. Genetics 2001, 157:1819-1829.

Meyermans, R., Gorssen, W., Buys, N. and Janssens, S., 2020. How to study runs of homozygosity using PLINK? A guide for analyzing medium density SNP data in livestock and pet species. BMC Genom, **21**(1), pp.1-14.

Mishra, D.C., Sikka, P., Yadav, S., Bhati, J., Paul, S.S., Jerome, A., Singh, I., Nath, A., Budhlakoti, N., Rao, A.R. and Rai, A., 2020. Identification and characterization of trait-specific SNPs using ddRAD sequencing in water buffalo. Genomics.

Moaeen-ud-Din, M. (2014). Buffalo genome research-a review. Anim. Sci. Pap. Rep, **32**(3), 187-199.

Mokhber, M., Shahrbabak, M.M., Sadeghi, M., Shahrbabak, H.M., Stella, A., Nicolzzi, E. and Williams, J.L., 2019. Study of whole genome linkage disequilibrium patterns of Iranian water buffalo breeds using the Axiom Buffalo Genotyping 90K Array. Plos one, **14**(5), p.e0217687.

Mokry, F.B., Buzanskas, M.E., de Alvarenga Mudadu, M., do Amaral Grossi, D., Higa, R.H., Ventura, R.V., de Lima, A.O., Sargolzaei, M., Meirelles, S.L.C., Schenkel, F.S. and da Silva, M.V.G.B., 2014. Linkage disequilibrium and haplotype block structure in a composite beef cattle breed. BMC Genom, **15**(S7), p.S6.

Moradi, M.H., Nejati-Javaremi, A., Moradi-Shahrbabak, M., Dodds, K.G. and McEwan, J.C., 2012. Genomic scan of selective sweeps in thin and fat tail sheep breeds for identifying of candidate regions associated with fat deposition. BMC genet, **13**(1), p.10.

Notter, D.R., 1999. The importance of genetic diversity in livestock populations of the future. Sci. J. Anim. Sci, **77**(1), pp.61-69.

Novembre, J. and Peter, B.M., 2016. Recent advances in the study of fine-scale population structure in humans. Curr Opin Genetics Dev, **41**, pp.98-105.

Olver, A., 1938. A brief survey of some of the important breeds of cattle in India. A brief survey of some of the important breeds of cattle in India., (17).

Patel, A. B., Subramanian, R. B., Padh, H., Shah, T. M., Mohapatra, A., Reddy, B., ... & Joshi, C. G. (2017). Identification of single nucleotide polymorphism from Indian Bubalus bubalis through targeted sequence capture. CURR. SCI, **112**(6), 1230.

Periasamy, K., Pichler, R., Poli, M., Cristel, S., Cetra, B., Medus, D., Basar, M., Thiruvenkadan, A.K., Ramasamy, S., Ellahi, M.B. and Mohammed, F., 2014. Candidate gene approach for parasite resistance in sheep–variation in immune pathway genes and association with fecal egg count. PLoS One, **9**(2), p.e88337.

Peripolli, E., Munari, D.P., Silva, M.V.G.B., Lima, A.L.F., Irgang, R. and Baldi, F., 2017. Runs of homozygosity: current knowledge and applications in livestock. Animal genet, **48**(3), pp.255-271.

Peripolli, E., Stafuzza, N.B., Munari, D.P., Lima, A.L.F., Irgang, R., Machado, M.A., do Carmo Panetto, J.C., Ventura, R.V., Baldi, F. and da Silva, M.V.G.B., 2018. Assessment of runs of homozygosity islands and estimates of genomic inbreeding in Gyr (Bos indicus) dairy cattle. BMC genomics, **19**(1), pp.1-13.

Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S. and Hoekstra, H.E., 2012. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PloS one, **7**(5), p.e37135..

Pickrell, J. and Pritchard, J., 2012. Inference of population splits and mixtures from genome-wide allele frequency data. Nature Precedings, pp.1-1.

Pritchard, V.L., Mäkinen, H., Vähä, J.P., Erkinaro, J., Orell, P. and Primmer, C.R., 2018. Genomic signatures of fine scale local selection in Atlantic salmon suggest involvement of sexual maturation, energy homeostasis and immune defence related genes. Mol. Ecol, **27**(11), pp.2560-2575.

Pritchard, J.K., Stephens, M. and Donnelly, P., 2000. Inference of population structure using multilocus genotype data. Genetics, **155**(2), pp.945-959.

Pundir R, Sahana G, Navani N, Jain P, Singh D, Kumar S, et al. (2000) Characterization of Mehsana buffaloes in India. Anim. Genet. Resour. 28: 53-62.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., Daly, M.J. and Sham, P.C., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet, **81**(3), pp.559-575.

Purfield, D.C., Berry, D.P., McParland, S. and Bradley, D.G., 2012. Runs of homozygosity and population history in cattle. BMC genet, **13**(1), p.70.

Qanbari, S., Pimentel, E.C.G., Tetens, J., Thaller, G., Lichtner, P., Sharifi, A.R. and Simianer, H., 2010. The pattern of linkage disequilibrium in German Holstein cattle. Anim Genet, **41**(4), pp.346-356.

Sambrook, J., Fritsch, E.F. and Maniatis, T., 1989. Molecular cloning: a laboratory manual (No. Ed. 2). Cold spring harbor laboratory press.

Santana Jr, M.L., Aspilcueta-Borquis, R.R., Bignardi, A.B., Albuquerque, L.G.D. and Tonhati, H., 2011. Population structure and effects of inbreeding on milk yield and quality of Murrah buffaloes. J. Dairy Sci **94**(10), pp.5204-5211.

Saura, M., Fernández, A., Varona, L., Fernández, A.I., de Cara, M.Á.R., Barragán, C. and Villanueva, B., 2015. Detecting inbreeding depression for reproductive traits in Iberian pigs using genome-wide data. Genet. Sel. Evol, **47**(1), p.1.

Seefried, F., 2016. Genomic prediction using haplotypes in Brown Swiss. Interbull Bulletin, (50).

Shokrollahi, Borhan, Cyrus Amirinia, Navid Dinparast Djadid, Noor Amirmozaffari, and Mohammad Ali Kamali. "Development of polymorphic microsatellite loci for Iranian river buffalo (Bubalus bubalis)." African Journal of Biotechnology 8, no. 24 (2009).

Shriner, D., 2013. Overview of admixture mapping. Current protocols in human genetics, **76**(1), pp.1-23.

Shtir, C.J., Marjoram, P., Azen, S., Conti, D.V., Le Marchand, L., Haiman, C.A. and Varma, R., 2009. Variation in genetic admixture and population structure among Latinos: the Los Angeles Latino eye study (LALES). BMC genetics, **10**(1), p.71.

Skotte, L., Korneliussen, T.S. and Albrechtsen, A., 2013. Estimating individual admixture proportions from next generation sequencing data. Genetics, **195**(3), pp.693-702.

Smith, W.L. and Wheeler, W.C., 2006. Venom evolution widespread in fishes: a phylogenetic road map for the bioprospecting of piscine venoms. J Hered, **97**(3), pp.206-217.

Stothard, P., Choi, J.W., Basu, U., Sumner-Thomson, J.M., Meng, Y., Liao, X. and Moore, S.S., 2011. Whole genome resequencing of black Angus and Holstein cattle for SNP and CNV discovery. BMC genomics, **12**(1), p.559.

Sun, C., Wang, B., Yan, L., Hu, K., Liu, S., Zhou, Y., Guan, C., Zhang, Z., Li, J., Zhang, J. and Chen, S., 2016. Genome-wide association study provides insight into the genetic control of plant height in rapeseed (Brassica napus L.). Front. Plant Sci, 7, p.1102.

Surya, T., Vineeth, M.R., Sivalingam, J., Tantia, M.S., Dixit, S.P., Niranjan, S.K. and Gupta, I.D., 2019. Genomewide identification and annotation of SNPs in Bubalus bubalis. Genomics, **111**(6), pp.1695-1698.

Sved, J.A., Cameron, E.C. and Gilchrist, A.S., 2013. Estimating effective population size from linkage disequilibrium between unlinked loci: theory and application to fruit fly outbreak populations. PLoS One, **8**(7), p.e69078.

Sved, J.A. and Feldman, M.W., 1973. Correlation and probability methods for one and two loci. Theor. Popul. Biol, **4**(1), pp.129-132.

Szmato³a, T., Gurgul, A., Ropka-Molik, K., Jasielczuk, I., Z¹bek, T. and Bugno-Poniewierska, M., 2016. Characteristics of runs of homozygosity in selected cattle breeds maintained in Poland. Livest Sci, 188, pp.72-80.

Tantia MS, Vijh RK, Bhasin V, Sikka P, Vij PK, Kataria RS, Mishra BP, Yadav SP, Pandey AK, Sethi RK, Joshi BK, Gupta SC, Pathak KML: Whole-genome sequence

assembly of the water buffalo (Bubalus bubalis). Indian J Anim Sci. 2011, **81** (5): 38-46.

Tantia, M. S., Vijh, R. K., Mishra, B. P., Mishra, B., Kumar, S. B., & Sodhi, M. (2006). DGAT1 and ABCG2 polymorphism in Indian cattle (Bos indicus) and buffalo (Bubalus bubalis) breeds. BMC Vet Res, **2**(1), 32.

Thakor, P.B., Hinsu, A.T., Bhatiya, D.R., Shah, T.M., Nayee, N., Sudhakar, A. and Joshi, C.G., 2018. High-throughput genotype based population structure analysis of selected buffalo breeds. bioRxiv, p.395681.

Thiruvenkadan, A.K., Rajendran, R. and Muralidharan, J., 2013. Buffalo genetic resources of India and their conservation. Buffalo Bull, 32, pp.227-235.

Yang, T., 2017. Application of High-throughput Genomic Data in the Genetic Analysis of Pigs.

Campbell, M.C. and Tishkoff, S.A., 2008. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. Annu. Rev. Genomics Hum. Genet., 9, pp.403-433.

Turton, J.D., 1974, October. The collection, storage and dissemination of information on breeds of livestock. In Proceedings of the 1st World Congress on Genetics Applied to Livestock Production, II (pp. 61-74).

Upadhyay, M. R., Patel, A. B., Subramanian, R. B., Shah, T. M., Jakhesara, S. J., Bhatt, V. D., ... & Joshi, C. G. (2015). Single nucleotide variant detection in Jaffrabadi buffalo (Bubalus bubalis) using high-throughput targeted sequencing. Front Life Sci, **8**(2), 192-199.

Van Orsouw, N.J., Hogers, R.C., Janssen, A., Yalcin, F., Snoeijers, S., Verstege, E., Schneiders, H., van der Poel, H., Van Oeveren, J., Verstegen, H. and Van Eijk, M.J., 2007. Complexity reduction of polymorphic sequences (CRoPS™): a novel approach for large-scale polymorphism discovery in complex genomes. PloS one, **2**(11), p.e1172.

Van Tassell, C.P., Smith, T.P., Matukumalli, L.K., Taylor, J.F., Schnabel, R.D., Lawley, C.T., Haudenschild, C.D., Moore, S.S., Warren, W.C. and Sonstegard, T.S., 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat. methods, **5**(3), p.247.

Villa-Angulo, R., Matukumalli, L.K., Gill, C.A., Choi, J., Van Tassell, C.P. and Grefenstette, J.J., 2009. High-resolution haplotype block structure in the cattle genome. BMC Genet, **10**(1), p.19.

Visser, C., Lashmar, S.F., Van Marle-Köster, E., Poli, M.A. and Allain, D., 2016. Genetic diversity and population structure in South African, French and Argentinian Angora goats from genome-wide SNP data. PLoS One, **11**(5), p.e0154353.

Wang, Y.H., Reverter, A., Mannen, H., Taniguchi, M., Harper, G.S., Oyama, K., Byrne, K.A., Oka, A., Tsuji, S. and Lehnert, S.A., 2005. Transcriptional profiling of muscle tissue in growing Japanese Black cattle to identify genes involved with the development of intramuscular fat. AUST J EXP AGR, **45**(8), pp.809-820.

Wang, S., Lewis Jr, C.M., Jakobsson, M., Ramachandran, S., Ray, N., Bedoya, G., Rojas, W., Parra, M.V., Molina, J.A., Gallo, C. and Mazzotti, G., 2007. Genetic variation and population structure in Native Americans. PLoS genetics, **3**(11), p.e185.

Weir, B.S. and Cockerham, C.C., 1984. Estimating F-statistics for the analysis of population structure. evolution, pp.1358-1370.

Wright, S., 1948. On the roles of directed and random changes in gene frequency in the genetics of populations. Evolution, pp.279-294.

Wright, S., 1931. Evolution in Mendelian populations. Genetics, **16**(2), p.97.

Wright, S., 1943. Isolation by distance. Genetics, **28**(2), p.114.

Zhan, B., Fadista, J., Thomsen, B., Hedegaard, J., Panitz, F. and Bendixen, C., 2011. Global assessment of genomic variation in cattle by genome resequencing and high-throughput genotyping. BMC genomics, **12**(1), p.557.

Zhang, Q., Calus, M.P., Guldbrandtsen, B., Lund, M.S. and Sahana, G., 2015. Estimation of inbreeding using pedigree, 50k SNP chip genotypes and full sequence data in three cattle breeds. BMC genet, **16**(1), p.88.

Zhu, Bo, Hong Niu, Wengang Zhang, Zezhao Wang, Yonghu Liang, Long Guan, Peng Guo et al. "Genome wide association study and genomic prediction for fatty acid composition in Chinese Simmental beef cattle using high density SNP array." BMC genomics 18, no. **1** (2017): 464.

Zimin, A.V., Delcher, A.L., Florea, L., Kelley, D.R., Schatz, M.C., Puiu, D., Hanrahan, F., Pertea, G., Van Tassell, C.P., Sonstegard, T.S. and Marçais, G., 2009. A whole-genome assembly of the domestic cow, Bos taurus. Genome biol, **10**(4), p.R42.

Zimin, A.V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S.L. and Yorke, J.A., 2013. The MaSuRCA genome assembler. Bioinformatics, **29**(21), pp.2669-2677.

Websites and others

1. http://faostat3.fao.org/browse/Q/QA/E
2. http://www.affymetrix.com/catalog/prod740001/AFFY/Axiom%26%23174%3B-Buffalo-Genotyping-Array#1
3. AGRI-IS, NDDB
4. Annual report, NPB 2017/18
5. Bovine Genome Sequencing& Analysis Consortium, 2009
6. Department of Animal Husbandry, Dairying & Fisheries, Ministry of Agriculture, GoI.
7. *Hapmap, 2005*
8. National Academy of Science, 1981
9. Sargolzaei M, University of Guelph, Canada
10. 19th livestock census.

Appendix

APPENDIX

Preparation of Reagents

Tris (1 M) pH 8.0

Dissolve 121.10 gm of Tris base in 800 ml of double distilled water and adjust the pH to 8.0 by adding concentrated HCl (around 40 ml). Allow the solution to cool to room temperature before making final volume to 1 litre and store at 4 ^{0}C.

Sodium chloride (5 M)

Dissolve 29.40 gm of Sodium chloride in 80 ml of double distilled water and make the final volume up to 100 ml. Sterilize by autoclaving and stored at 4 ^{0}C.

Sodium acetate (3 M)

Dissolve 24.60 gm of Sodium acetate (anhydrous) in 80 ml of double distilled water. Adjust the pH to 5.2 with glacial acetic acid and make up the final to 100 ml. Sterilize by autoclaving and store at room temperature.

EDTA (0.5 M)

Dissolve 18.60 gm of Disodium salt of EDTA in 80 ml of double distilled water by adding NaOH pellets. Adjust the pH to 8.0 and make up the final volume to 100 ml. Sterilize by autoclaving and store at 4 ^{0}C.

Ammonium chloride (1 M)

Dissolve 5.35 gm ammonium chloride in 80 ml of double distilled water and make the final volume up to 100 ml. Sterilize by autoclaving and Store at 4 ^{0}C.

Potassium bicarbonate (1 M)

Dissolve 10.00 gm of sodium bicarbonate in 80 ml of double distilled water and mak up to a final volume of 100 ml. Sterilize by autoclaving.

Sodium dodecyl sulphate (SDS) 10 %

Dissolve 10.00 gm of Sodium dodecyl sulphate in 80 ml of autoclaved double distilled water and make up to a final volume of 100 ml. No autoclaving needed.

Proteinase-K

Dissolve Proteinase-k (20 mg) in 1ml of double distilled water. Store at -20 ^{0}C.

Ethanol 70 %

Ethanol 99.9 %	70 ml
Distilled water	30 ml

RBC Lysis buffer (1 X prepared freshly every time**)**

NH_4Cl (155 mM)	155.0 ml
$KHCO_3$ (10 mM)	10.0 ml
EDTA (0.1 mM)	2.0 ml

Make up to 1000 ml by double distilled water. Sterilize by autoclaving and store at 4 ^{0}C.

Phenol equilibration (Tris saturation)

Liquify Phenol crystals (500 gm) stored at –20 ^{0}C by keeping in a water bath maintained at 65 ^{0}C for 1 hour. Add 8-hydroxyquinoline to liquefied phenol at a final concentration of 0.1 %. Add equal volume (500 ml) of 0.5 M Tris (pH 8.0) and stir for 4 hours on magnetic stirrer, acheck pH repeatedly till it reached 8.0. Finally, 0.1 M Tris was added to an equilibrated phenol and stirred well and stored in amber colored bottles at (4 ^{0}C).

TE buffer (10 mM)

Tris (1M, pH 8.0)	1.00 ml
EDTA (0.5M, pH 8.0)	200 µl
Final volume made up to	100 ml

Sterilize by autoclaving and store the buffer at room temperature.

Double distilled water	90 ml

TAE buffer (50X)

Tris base	242.0 gm
Glacial acetic acid	57.1 ml
EDTA (0.5 M, pH 8.0)	100 ml

Add double distilled water to make the final volume 1000ml, filter and autoclave.

VITAE

Name	:	**Dr. Shiv Kumar Tyagi**
Father's Name	:	Mr. Dori Lal Tyagi
Mother's Name	:	Mrs. Manju Tyagi
Date of Birth	:	09/01/1994
Permanent Address	:	32- Shivani Dham Colony, 100 ft Road, Kalindi Vihar, Agra, Uttar Pradesh (282006)
E-mail	:	shivtyagi632@gmail.com
Mobile No.	:	8218319348
Language Known	:	Hindi and English
Nationality	:	Indian
Major Field of Specialization	:	Animal Genetics & Breeding

Academic Qualifications :

Degree	Board/University	Year of passing	Percentage
B.V.Sc. & A.H.	SDAU, Gujarat	2018	81.25%
M.V.Sc.	ICAR-IVRI, Izatnagar	2020	85.57%

Awards and Fellowship

- ICAR-IVRI Institute Fellowship for M.V.Sc programme
- UGC-Net Qualified (2019)