GENOME-WIDE SNPs DISCOVERY AND ANNOTATION IN GIR CATTLE

THESIS SUBMITTED TO THE
ICAR-NATIONAL DAIRY RESEARCH INSTITUTE, KARNAL
(DEEMED UNIVERSITY)
IN PARTIAL FULFILMENT OF THE REQUIREMENT
FOR THE AWARD OF THE DEGREE OF

MASTER OF VETERINARY SCIENCE
IN
ANIMAL GENETICS AND BREEDING
BY

ANJALI CHOUDHARY
B.V.Sc & A.H.

ANIMAL GENETICS AND BREEDING DIVISION
ICAR - NATIONAL DAIRY RESEARCH INSTITUTE
(DEEMED UNIVERSITY)
KARNAL-132001 (HARYANA), INDIA

2019

Regn. No. 17-M-AG-01

GENOME-WIDE SNPs DISCOVERY AND ANNOTATION IN GIR CATTLE

By

ANJALI CHOUDHARY

THESIS SUBMITTED TO THE

ICAR-NATIONAL DAIRY RESEARCH INSTITUTE, KARNAL

(DEEMED UNIVERSITY)

IN PARTIAL FULFILMENT OF THE REQUIREMENT

FOR THE DEGREE OF

MASTER OF VETERINARY SCIENCE

IN

ANIMAL GENETICS AND BREEDING

Approved By:

	10/07/2019
(EXTERNAL EXAMINER)	**(ARCHANA VERMA)**
	Major Advisor & Chairperson

Advisory Committee Members

1. **Dr. I.D. GUPTA**
   Principal Scientist, AG&B Division
2. **Dr. JAYAKUMAR S.**
   Senior Scientist, NBAGR
3. **Dr. ANUPAMA MUKHERJEE**
   Principal Scientist, AG&B Division
4. **Dr. A. K. MOHANTY**
   Principal Scientist, ABTC
   (Joint Director's Nominee)

ANIMAL GENETICS & BREEDING DIVISION
ICAR-NATIONAL DAIRY RESEARCH INSTITUTE
(DEEMED UNIVERSITY)
KARNAL- 132001 (HARYANA), INDIA

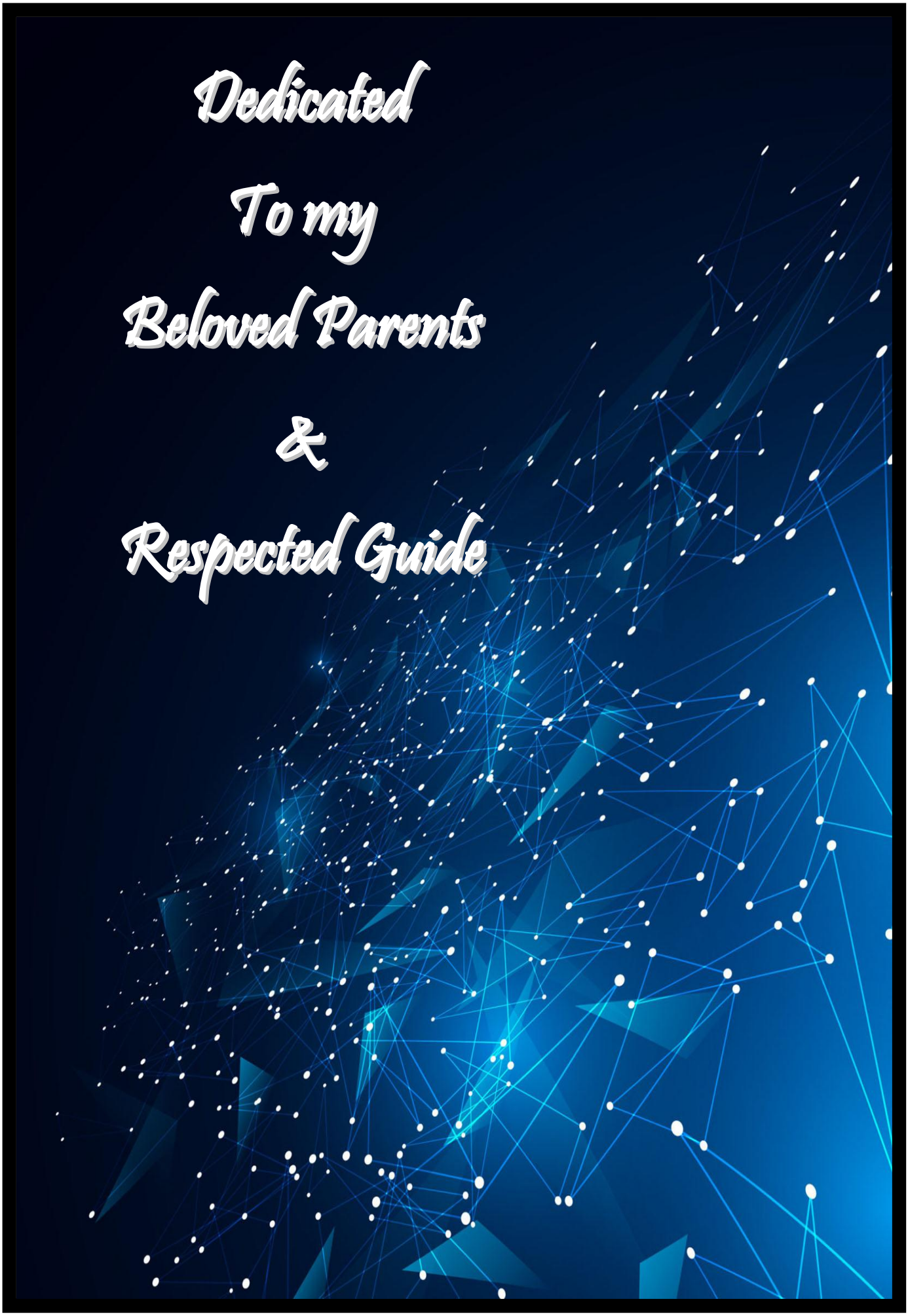
Dr. Archana Verma, Ph.D.
(Principal Scientist)

CERTIFICATE

This is to certify that the thesis entitled, **"GENOME-WIDE SNPs DISCOVERY AND ANNOTATION in GIR CATTLE"** submitted by **Ms. ANJALI CHOUDHARY** towards the partial fulfilment of the award of the degree of **MASTER OF VETERINARY SCIENCE IN ANIMAL GENETICS AND BREEDING** of the **ICAR-National Dairy Research Institute (Deemed University)**, Karnal (Haryana), India, is a bonafide research work carried out by him under my supervision, and no part of the thesis has been submitted for any other degree or diploma.

Dated:

(Archana Verma)
Major Advisor & Chairperson


Dedicated
To my
Beloved Parents
&
Respected Guide

ACKNOWLEDGEMENTS

I feel my foremost duty to express my deepest gratitude to my benevolent advisor Dr. ARCHANA VERMA, Principal Scientist, Animal Genetics and Breeding Division for her enthusiastic supervision, guidance and support. Madam, you will continue to be an eternal source of inspiration to me. Any success that I have or will receive in my journey through science will perpetually be dedicated, in part, to you. Thank you, Madam.

I express my sincere gratitude to members of my advisory committee, Dr. I.D. GUPTA, Principal Scientist, AGB Division; Dr. JAYAKUMAR S. Senior Scientist, AGB Division; Dr ANUPAMA MUKHRJEE, Principal Scientist, AGB Division and Dr. A. K. Mohanty, Principal Scientist, ABTC for their constructive criticism, discussion and constant inspiration during the entire period of research work. My utmost respect to all the teachers who made me sound in the subject, I will remain grateful to each one of them.

My sincere thanks are due to Dr, R. R. B. Singh Director, ICAR-NDRI, Karnal and Dr. S. M. Deb, Head, Animal Genetics and Breeding Division for providing all the necessary facilities to carry out the research. Financial assistance from ICAR in the form of Institute Fellowship is gratefully acknowledged.

I owe profound thanks to my respected seniors Nisha mam, Siddhu sir, Beena mam and Ragini mam, for their immense help, invaluable guidance at every stage of my research and dissertation.

I take this opportunity to thank my seniors Rebecca mam, Vineeth sir, Ravi sir, Kousalya mam, Surya sir, Sushil sir, Tarun singh sir for being with me all through the research work and for the whole hearted help and moral support rendered to me.

I am extremely thankful to my wonderful and ever helping batch mates and friends and juniors Harshit, Joel, Linda, Oshin, Nandhini, Iwan, Namith, Reshma, Ashokan, Eakta, Aruna, Shivam and Abhimanyu. Also, I take the pleasure to thank all my seniors and juniors of AG&B family for their help, care and affection.

Where emotions are involved, words cease to mean. There are no words but only feelings to honorably pay my very regards to Mrs. Kamlesh Choudhary

(Mother), Shri. Gulab Singh Choudhary (Father) and Siblings for their love and care.

Last but not least, I record my sincere thanks to all the well-wishers and I'll ever remain thankful to all those who could not have find separate names but had directly or indirectly helped me.

(Anjali Choudhary)

CONTENT

LIST OF TABLES

LIST OF FIGURES

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	:	Deoxyribo Nucleic 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
MT	:	Metric Tonnes
NCBI	:	National Center for Biotechnology Information
NGS	:	Next Generation Sequencing
OD	:	Optical Density
PCR	:	Polymerase Chain Reaction
QC	:	Quality Control
RAD	:	Restriction Site association DNA sequencing
RD	:	Read Depth
RE	:	Restriction Enzyme
RNA	:	RiboNucleic Acid
RPM	:	Revolution Per Minute
SAM	:	Sequence Alignment Format
SDS	:	Sodium Dodecyl Sulfate
SNP	:	Single Nucleotide Polymorphism
TAE buffer	:	Tris: Acetic acid: EDTA buffer

TE buffer	:	Tris: EDTA buffer
Ts	:	Transition
Tv	:	Transversion
Uv	:	Ultraviolet
VCF	:	Variant Calling Format
WGS	:	Whole Genome Sequencing

ABSTRACT

Molecular genetics approach can increase the genetic gain and selection of traits of economic importance. The drawback of Whole Genome Sequencing (WGS) such as the cost of sequencing and repetitive sequence is overcome by the NGS based RAD sequencing method. In the present study, we identified SNPs by using ddRAD (double Digest Restriction Associated DNA) approach in six samples of Gir cattle. The sequenced data generated by Illumina hiseq 2000 sequencer. A total of 13.1 million raw reads were obtained. After quality control, we retained a total of 12.8 million reads. Good quality processed reads were aligned to the Reference genome of *Bos taurus*, *Bos indicus* and Gir separately which results in 99.76%, 90.36%, and 98.29% respectively. A total of 6.38%, 6.36%, and 6.38% coverage is covered by the generated processed reads with the reference genome of *Bos taurus, Bos indicus*, and Gir respectively. As compared to *Bos taurus* reference genome we identified a total of 193764, 179205 and 160951 SNPs; 12082, 11059 and 9802 INDELs at RD2, RD5 and RD10 respectively. Similarly, when compared to the *Bos indicus* reference genome a total of 117467, 109943 and 99517 SNPs; 12082, 11059 and 9802 INDELs were identified. A total of 174492, 162470 and 146564 SNPs; 10887, 10050 and 9010 INDELs respectively were identified with Gir reference genome. All the SNPs identified in the study were structurally and functionally annotated. Further, the SNPs located in the candidate genes were also annotated. Out of 400 collected genes from the available literature and cgQTL database total 984 SNPs were annotated in 175 genes affecting milk production and composition (milk yield, fat yield, protein yield, milk fat percentage and protein percentage). The SNPs identified in this study may respond as a useful tool in future studies particularly in genome-wide association studies for better understanding of genetic structure revealing the phenotypic difference in cattle.

गिर मवेशियों के पूरे जीनोम में एसऍनपीज़ की खोज तथा एनोटेशन

सारांश

आणविक आनुवांशिकी दृष्टिकोण आनुवंशिक महत्व और आर्थिक महत्व के लक्षणों के चयन को बढ़ा सकता है। अनुक्रमण की लागत और दोहराव अनुक्रम, पूरे जीनोम के अनुक्रमण (WGS) की कमियां हैं, जिन्हे NGS आधारित राड अनुक्रमण विधि से दूर किया जा सकता है। वर्तमान अध्ययन में, हमने गिर मवेशियों के छह नमूनों में ddRAD (डबल डाइजेस्ट प्रतिबंध एसोसिएटेड डीएनए) दृष्टिकोण का उपयोग करके एसएनपी की पहचान की। इलुमिना हाइसेक 2000 सीक्वेंसर द्वारा अनुक्रम डेटा उत्पन्न किया गया। कुल 13.1 मिलियन कच्चे रीड प्राप्त किए गए।गुणवत्ता नियंत्रण के बाद, हमने कुल 12.8 मिलियन रीड्स रखे। अच्छी गुणवत्ता वाली संसाधित रीड्स को क्रमशः बोस टोरस, बोस इंडिकस और गिर के संदर्भ जीनोम से जोड़ दिया गया, जिसके परिणामस्वरूप क्रमशः 99.76%, 90.36% और 98.29% परिणाम आए। संसाधित रीड्स बॉस् टोरस, बॉस् इंडिकस और गिर के संदर्भ जीनोम को क्रमशः कुल 6.38%, 6.36%, और 6.38% कवरेज उत्पन्न कर रहा था। बोस टॉरस संदर्भ जीनोम की तुलना में हमने कुल 193764, 179205 और 160951 एसएनपी तथा 12082, 11059 और 9802 INDELs की पहचान; क्रमशः RD2, RD5 और RD10 में की। गिर संदर्भ जीनोम के खिलाफ हमने कुल 174492, 162470 और 146564 एसएनपी तथा क्रमशः 10887, 10050 और 9010 INDELs की पहचान की। अध्ययन में पहचाने गए सभी एसएनपी संरचनात्मक और कार्यात्मक रूप से एनोटेट किये गए। इसके अलावा, उम्मीदवार जीन में स्थित एसएनपी को भी एनोटेट किया गया था। साहित्य और cgQTL डेटाबेस से 400 एकत्र जीन में से कुल 984 एसएनपी दूध उत्पादन और संरचना (दूध की उपज, वसा की उपज, प्रोटीन की उपज, दूध वसा प्रतिशत और प्रोटीन प्रतिशत) को प्रभावित करने वाले 175 जीनों में एनोटेट किए गए। इस अध्ययन में पहचाने गए एसएनपी भविष्य के अध्ययनों में विशेष रूप से जीनोम-वाइड एसोसिएशन अध्ययनों में एक उपयोगी उपकरण के रूप में प्रतिक्रिया दे सकते हैं ताकि मवेशियों में फेनोटाइपिक अंतर का बेहतर पता लगाया जा सके।

CHAPTER –1

Introduction

INTRODUCTION

Animal husbandry is an integral element of Indian agriculture supporting the livelihood of more than two-third of the rural population. India possesses the largest bovine population in the world and has the status of first in milk production since 1998. The total cattle population of India is 190 million (37.28% of total livestock population) comprising 151.17 million of indigenous cattle population (19th Livestock Census, 2012). Agriculture sector contributes 17.9% of total GDP in the Indian economy, and livestock sector contributes 4.6% of total GDP and 25.6% of total Agriculture GDP (National Accounts Statistics-2016, Central Statistical Organization, Gol). The per capita availability of milk in India which was 130 gram per day during 1950-51, has increased to 374 gram per day in 2017-18 (www.dahd.nic.in). Domesticated cattle are classified into *Bos taurus* and *Bos indicus* and are believed to be descended from the wild species, the aurox (Epstein and Mason, 1984). As compared to taurine cattle which are predominantly found in European countries, indicine cattle possess special attributes like heat tolerance, disease resistance and adaption to tropical climate. At present, there are 43 registered indigenous cattle breeds in India (www.nbagr.res.in) and Gir cattle is one of the important indigenous dairy breeds originated in India. It is also called as Bhodali, Sorthi, Desan, Gujarati, Kathiawari, and Surati (Porter *et al*., 2016) and the native tract of the breed is Gir hills and the surrounding forest of Kathiawar including Junagadh, Gir Somnath, Amreli, and Rajkot district of Gujarat (www.fao.org). Although the breed is native of Gujarat, it is conjointly found in Maharashtra and Rajasthan states in India (www.fao.org). At present, it is found in most States of India and in many countries. The total Gir population in India is 5.1 million and percent share with respect to the total cattle population in 3.38% (Basic Animal Husbandry & Fisheries Statistics 2018). Gir is available in different coat colors like white with dark red or chocolate-brown patches or sometimes black or purely red. The average Milk yield of Gir is 1590 kg per lactation, with a record production of

3182 kg at 4.5% fat in India (Gaur *et al.,* 2003). Gir animals are considered as hardy with a low overall mortality of 3.63 % (www.fao.org). The Scientists of Junagadh University found traces of gold in the urine of Gir cattle (Golakia, 2016) in the form of a water-soluble salt, along with other metals. Gir has been used locally in the improvement of other breeds particularly for the production of rustic crossbred and high milk yield (Reis *et al*., 2010) It was also one of the breeds used in the development of the Brahman breed in North America (Briggs and Briggs, 1980). It is often bred with Holstein cows to make the Girolando breed. The genetic potential of this dairy breed needs more attention for the advancement in its economic characters so as to achieve maximum milk production. The breed improvement programs within the country have targeted traditional breeding techniques to enhance growth, reproduction and production traits (Naqvi, 2007). Low replacement rate, high generation interval, and maintaining low productive animals are the constraints of those breeding programs (Biscarini *et al.,* 2015). The low productivity and continued loss of local breed diversity call for genetic improvement. The molecular genetic approaches in standard selection strategies can increase the genetic gain and choice of the specified traits of economic importance (Mitra *et al.,* 1999). SNPs (Single nucleotide polymorphism) pronounced “snips” the foremost common gene known so far. An SNP is a single base pair change at a specific locus, usually consisting of two alleles (where the allelic frequency is >1. They will act as a biological marker, helping scientist to find genes that are related to some production and reproduction traits. SNPs have some advantages over microsatellites, including greater abundance (Heaton *et al.,* 2005), its genetic stability (Markovtsova *et al*., 2000; Nielsen *et al.,* 2000; Thomson *et al.,* 2000), appropriate for analysis and data interpretation (Wang *et al.,* 1998; Lindblad-toh *et al.,* 2000). Implementation of Marker-Assisted Selection technique has been limited and the increase in genetic gain is small. Genomic selection programs have been found to be costly and the data analysis requires high computing system as well the bioinformatic analysis is difficult. Whole genome sequencing technique reveals the complete DNA makeup of an organism which is sometimes unnecessary. The

repetitive sequence present may cause difficulty in annotation and in bioinformatic result analysis and also very expensive (Ng and Kirkness, 2010). These drawbacks are overcome by using the reduced representation approach which includes restriction digestion of the genome followed by size selection (Van Tassell *et al*., 2008). RAD sequencing is a reduced representative next-generation technique which eliminates the repetitive sequences thereby reduces the cost of sequencing and genotyping. Among all the methods of RAD sequencing like CRoPS (Osrow *et al.,* 2007), RRL (Van-Tassel *et al.,* 2008), GBS (Elshire *et al*., 2011), the ddRAD method (Peterson *et al.,* 2012) having some advantages like there is a use of two restriction enzymes to facilitates multiplexing and size selection which reduces repetitive sequence, DNA loss and increases the accuracy of result and analysis. Genome annotation identifying the location of genes and all of the coding regions in a genome. Although RAD sequencing approaches have been utilized in the livestock sector the ddRAD approach which has numerous advantages over RAD sequencing method has been attempted to a lesser extent. During recent years the cattle genomic studies are greatly increased. One of the most important cattle genomes ever sequenced and annotated was of female Hereford cow by the Bovine genome sequencing and analysis pool (Elsik *et al.,* 2009). The first genome sequence of an Indicine breed “Nellore” has been generated (Canavez *et al*., 2012). Now large numbers of SNPs are known in different cattle breeds, which could be utilized in Genome-wide association studies. This opens up the door for a lot of genetic data from indicine cattle. In *Bos taurus* and *Bos indicus* SNPs occur approximately every 700 bp and 300 bp respectively (The Bovine HapMap Consortium 2009; Seidel, 2009), this information reveals that there is more genetic variation in *Bos indicus* cattle.

So the present study has been planned in Gir cattle with the following objectives:

1. Genome-wide identification of SNPs in Gir cattle using the ddRAD approach
2. Annotation of identified SNPs for the candidate genes affecting milk production and composition traits.

FIGURE 1 - GIR CATTLE

CHAPTER –2

Review of Literature

REVIEW OF LITERATURE

The total cattle population of India is 190 million (37.28% of total livestock population) comprising 151.17 million of indigenous cattle population (19th Livestock Census, 2012). India ranks first among the world's milk-producing nations since 1998 and has the largest bovine population in the World. Milk production in India during the period 1950-51 to 2017-18, has increased from 17 million tonnes to 176.4 million tonnes. At present, there are 43 registered indigenous cattle breeds in India (www.nbagr.res.in). Gir cattle is one of the important indigenous dairy breeds originated in India. It is also called as Bhodali, Desan, Gujarati, Kathiawari, and Surati (Porter *et al*., 2016) and the native tract of the breed is Gir hills and the surrounding forest of Kathiawar including Junagadh, Gir Somnath, Amreli, and Rajkot district of Gujarat (www.fao.org). Although the breed is native of Gujarat, it is conjointly found in Maharashtra and Rajasthan states in India (www.fao.org). This breed is used for the upgradation of local non-descript animals. Traditionally, the main techniques used in breeding and improving the productivity of dairy cattle were 'Selection' and 'mating plan' like the mating of best with best (assortative mating). Economically important traits in cattle selected on the basis of phenotype of the individual or relatives. These breeding techniques take many years to give results and their impact on low heritable or late expressed traits is limited. There are two approaches for identification of markers associated with economic traits, i.e., candidate gene approach and Genome-wide association. Both the methods have their pros and cons. Candidate gene approach was one of the powerful and accurate methods for studying genetic association of complex traits, but this approach mainly focuses on associations between genetic variation within pre-specified genes of interest or depend on previous knowledge of the gene's biological and functional impact on the trait of interest. While genome-wide association studies (GWAS), scan the entire genome for common genetic variation, but the principal disadvantage is cost and high resource requirement (Kwon *et al*., 2000; Zhu *et al*., 2007).The trend of identifying candidate genes affecting QTL

region started since, the 1990s (Hoeschele and Meinert, 1990). Genomic Selection (GS; Meuwissen *et al*, 2001) comes into the light which is based on Genome-wide markers to produce Genomic Estimated Breeding Value (GEBV). This will enhance genetic response by reducing cost and generation interval (Schaeffer, 2006).

2.1 MOLECULAR MARKERS

A molecular marker is a gene or a stretch of a DNA sequence of known function and location. With the help of molecular genetics, remarkable advances have been made over the last decades in the identification of candidate genes associated with economically important traits in livestock production. It is believed that the genes associated with certain trait may show mutation which causes variation in that trait (Hayes *et al.,* 2007). These markers are highly polymorphic, randomly distributed throughout the genome, frequently occurr in the genome and highly reproducible in nature. Molecular markers are significantly important and advantageous than conventional breeding techniques. Molecular markers categorized into hybridization-based, i.e., Restriction Fragment Length Polymorphisms (RFLP) established by Grodzicker *et al*. (1974), It is not widely used now but it was one of the first technique used for DNA analysis. It is widely used in various conservation and breeding programs. Jiang and Gibson (1999) identified 4 new genetic variations in the leptin gene of different pig breeds using RFLP. Polymerase Chain Reaction-based markers i.e, Random amplified length polymorphic DNAs is a PCR based technique developed independently by two different laboratories (Williams *et al*., 1990) and called as RAPD and AP-PCR (Arbitrary primed PCR) respectively. The RAPD technique is based on the principle of polymerase chain reaction. Koh *et al*. (1990) generated specific fingerprinting pattern in 10 different species i.e, wild boar, pig, horse, buffalo, beef, venison, dog, cat, rabbit & kangaroo by using RAPD method. The term "microsatellite" was introduced later,by Litt and Luty (1989). A microsatellite is a tract of repetitive DNA in which certain DNA motifs (ranging in length from 1-6 or more base pairs) are repeated, typically 5–50 time. Microsatellites and Minisatellites, together are classified as VNTR (Variable Number of Tandem Repeats) DNAs. Amplified

fragment length polymorphisms (AFLP) was developed by Zabeau and Vos (1993). It is a combination of the RFLP and PCR techniques. DNA chips and sequencing based DNA markers such as Single nucleotide polymorphisms (SNP) is a substitution of a single nucleotide that occurs at a specific position in the genome. This molecular technology was proposed by Lander (1996). It includes single base transitions, transversions, insertions and deletions (Vignal *et al*., 2002). Among all the mutation of SNPs, transitions are most common about 2/3 (Zhao *et al.,* 2002). SNP belongs to third generation molecular techniques and is widely used as the commonest among all the markers because of their higher abundance in the genome, stability and easy assessment in high throughput automation analysis, because of their remarkable qualities they are highly preferred for the investigation of genetic variation among different species and breeds (Gill *et al*., 2001; Bovine HapMap Consortium, 2009). SNP may fall within the coding sequence of genes and non-coding region of genes or in intergenic regions. SNP in the coding region is of 2 types Synonymous and non-synonymous SNP. Synonymous SNP does not affect the protein sequence while non-synonymous change the amino acid sequence of the protein. Non-synonymous again are of 2 types missense and nonsense. Where a missense change results in a different amino acid while a nonsense change results in a premature stop codon. Consequently, SNPs following biallelic system (only 2 alleles in a population) in contrast to microsatellites which are multiallelic markers resultantly the information content per SNP marker is less. Therefore high throughput technologies needed to scan large numbers of SNPs. Information provided by 5 SNP markers is equal to 1 microsatellite marker (Beuzen *et al*., 2000). SNPs are extremely useful in association studies, gene mapping, and phylogenetic studies.

TABLE 2.1 PROPERTIES OF DIFFERENT MOLECULAR MARKERS

S.No	PARTICULARS	RFLP	RAPD	AFLP	SSRs	SNP
1	Identification method	Hybridization	PCR	PCR	PCR	DNA chips and sequencing based
2	Restriction enzyme	Yes	No	Yes	No	No
3	DNA requires (µg)	10	0.02	0.5-1.0	0.05	0.05
4	DNA quality	High	High	Moderate	Moderate	High
5	Types of polymorphism	Single base change, insertion, deletion	Single base change, insertion, deletion	Single base change, insertion, deletion	Change in repeat length	Single nucleotide change, insertion, deletion
6	Dominance	Co-dominant	Dominant	Dominant/ Co-dominant	Co-dominant	Co-dominant
7	Automation	Low	Moderate	Moderate	High	High
8	Accuracy	Very high	Very low	Medium	High	Very high
9	Cost per analysis	High	Low	Moderate	Low	Low
10	Reproducibility	High	Low	High	High	High
11	Specific primer	No	No	No	Yes	Yes
12	Degree of polymorphism	Low	Medium-high	Medium-high	High	High
13	Radioactive detection	Yes	No	No	No	Yes

(Gous *et al*.,2013)

TABLE 2.2 SNP IDENTIFICATION METHODS

S.NO.	METHODS	REFERENCE
1	Resequencing	Sanger *et al*., 1977
2	Denaturing Gradient Gel Electrophoresis (DGGE)	Fischer and Lerman, 1983
3	Single Strand Conformational Polymorphism analysis (SSCP)	Orita *et al.,* 1989
4	derived/Cleaved Amplified Polymorphic Sequences (dCAPs/CAPs)	Konieczny and Ausubel, 1993
5	Denaturing high-performance liquid chromatography (DHPLC) Wave	Oefner and Underhill, 1995
6	Cleavase Fragment Length Polymorphism (CFLP)	Rossetti *et al*., 1997
7	Pyrosequencing	Ronaghi *et al.,* 1998
8	Taqman assay	Livak *et al.*, 1999
9	Targeting Induced Local Lesions In Genomes (TILLING)	Mccallum *et al.,* 2000
10	Temperature Gradient Capillary Electrophoresis (TGCE)	Hsia *et al.,* 2005
11	DNA chips and microarrays	Gunderson *et al*., 2005
12	SNPlexTM genotyping system	De La Vega *et al.,* 2005

2.2 NEXT GENERATION SEQUENCING:

Previously in absence of whole genome sequence, high-resolution maps and cost-effective technologies for genotyping is difficult to identify polymorphic markers associated with complex and economically important traits.

With the advent of next-generation sequencing (NGS) and the availability of whole genome sequence, discovering, sequencing and genotyping not hundreds but thousands of markers in a single step is possible now.

Sanger (chain termination method) and Maxam-Gilbert (Chemical degradation method) sequencing technologies considered as first generation sequencing technologies developed by Sanger *et al. (*1997) and Maxam *et al.* (1997), for that they got Nobel prize in chemistry by Cambridge University and Harvard University respectively (1980). PhiX174 (5374 bp) and bacteriophage λ (48501 bp) were the first genomes sequenced by the Sanger sequencing (Sanger *et al*.,1980). Though it is a highly accurate method but it is costly and time-consuming. These technologies were dominant for three decades until the emergence of new generation sequencing technologies (2005). NGS technologies are fast, inexpensive and do not need electrophoresis for detecting sequencing output. It includes 3 sequencing platforms i.e, Roche/454 sequencing (2005), Illumina/Solexa sequencing (2006), and ABI/SOLID sequencing (2007).

454 Life Science was founded by Jonathan Rothberg in 2000. 454 Pyrosequencing technology (http://www.454.com/) is based on sequencing by synthesis principle. It works by the detection of pyrophosphate released after each nucleotide incorporation during the synthesis of a new DNA strand. It is capable of sequencing roughly 400-600 megabases of DNA per 10-hour run on the Genome Sequencer FLX with GS FLX Titanium series reagents (Voelkerding *et al*., 2009). Illumina sequencing technology was founded by David Walt (April, 1998), Larry Bock, John Stuelpnagel, Anthony Czarnik, and Mark Chee. Shankar Balasubramanian and David Klenerman of Cambridge University (1990) founded Solexa company which later acquired by Illumina (2007). It is the most used technology in the NGS market based on Sequencing by synthesis approach. In this technology randomly fragmented DNA, ligated to adapters by both the ends. After fixing themselves to the respective complement adapters on the slide they are amplified by PCR bridge amplification results into clusters of identical copies of each sequence. Next, primers and modified nucleotides are added. These nucleotides have reversible 3' blockers that force the polymerase to add on only one nucleotide at a time as well as fluorescent tags. Once the DNA strand has been read, the strand that was just added is washed away. The cluster emits signal

specific lights to each nucleotide which can be detected by coupled-charge device (CCD) these signals translated into nucleotide sequence by camera and computer program. This process continues until the full DNA molecule is sequenced (Meyer *et al.,* 2010).

Choi *et al*. (2014) performed whole-genome analyses of three important cattle breeds in Korea-Hanwoo, Jeju Heugu, and Korean Holstein using the Illumina HiSeq 2000 sequencing platform and achieved 25.5, 29.5, and 29.5 fold coverage, respectively, and identified a total of 10.4 million single nucleotide polymorphisms (SNPs), of which 54.12% were found to be novel and detected 1,063,267 INDELs. Das *et al.* (2015) sequenced four genetically unrelated Danish Holstein cows with a mean coverage of 27X by using Illumina Hiseq 2000 technique and identified 10,796,794 SNPs and 1,295,036 INDELs out of which 482,835 (4.5 %) SNPs and 231,359 (17.9 %) INDELs were novel. Stafuzza *et al.* (2017) sequenced four important cattle breeds in Brazil viz Guzerat (multi-purpose), Gyr, Girolando and Holstein. A total of approximately 4.3 billion reads from an Illumina HiSeq 2000 sequencer generated for each animal with 10.7 to 16.4-fold genome coverage. A total of 27,441,279 SNPs and 3,828,041 insertions/deletions (INDELs) were detected in the samples, of which 2,557,670 SNVs and 883,219 INDELs were novel.

SOLiD sequencing stands for Supported Oligonucleotide Ligation Detection developed by Life Technologies and has been commercially available since 2006. This next generation technology generates hundreds of millions to billions of small sequence reads at one time. SOLiD technology is based on sequencing by ligation approach. In this method, the adapter attached DNA fragments fixed on beads and cloned by PCR amplification. 8-mer fluorescent label ligated DNA fragment placed on a glass slide and the color emitted by the label is subsequently recorded. Single-Molecular real time (SMRT) is the third generation of the sequencing method developed by Pacific Bioscience (PacBio, Menlo Park, CA, USA). A new version of the sequencer called the PacBio RS II was released in April 2013. PacBio sequencing is a method for real-time sequencing.

TABLE 2.3 COMPARISION BETWEEN VARIOUS NGS TECHNOLOGIES

S.NO	PARTICULARS	FIRST GENERATION	SECOND GENERATION				THIRD GENERATION
		Sanger	454/Roche	Ion Torrent	Illumina	Solid/ABI	Pacbio
1	Sequencing mechanism	Di-deoxy chain termination	Pyrosequencing	Semiconductor technique	Sequencing by synthesis	Ligation and 2 base coding	SMRT approach
2	Read length	400 -900 bp	700 bp	400 bp	150 bp	75 bp	~10 kbp
3	Output data/run	0.00069-0.0021kb	0.07Gb	3-5Gb	1.8Tb	320 Gb	7Gb
4	Accuracy	99.999%	99.9%	99.6%	98%	99.94%	99.999%
5	Advantages	High quality, long read length	Read length, fast	Fast, no requirement of optical scanning & fluorescent nucleotides	High throughput	Accuracy	Read length, fast
6	Disadvantages	High cost, low throughput	Error rate with polybase more than 6, high cost, low throughput	High error rates in homopolymers	Short read assembly	Short read assembly	High cost, high overall error rates (~14%), lowest throughput

(Liu *et al*., 2012; Erwin *et al*., 2014; Mehdi *et al*., 2017)

2.3 WHOLE GENOME SEQUENCING TECHNOLOGY

For identifying SNPs, the whole genome sequencing uses short read technologies involving the alignment of millions of reads to a reference genome sequence. Haemophilus influenza was the first organism whose entire genome (1830137 base pairs) was sequenced by Fleischmann *et al*. (1995). The major drawback associated with this technology is that it reveals the complete DNA makeup of an organism resultantly compilation of lots of data which is always not a good thing. The presence of repetitive sequence creates difficulty in annotation and in result analysis and it is expensive too (Hirsch *et al.,* 2014) (Table 2.4).

2.4 REDUCED REPRESENTATION APPROACH

Reduced representation is an approach for sampling and sequencing a small set of genome-wide regions without sequencing the entire genome with the aim of identifying SNP markers in a cost-effective way for mapping several QTL regions (Xiaole *et al*., 2010). Reduced representation approaches include exome capture, genotyping-by-sequencing and transcriptome sequencing (Cory *et al.,* 2014). Exome capture sequencing is a technique which captures only the exonic regions of a genome, increases the depth and coverage by reducing the sequence space of a complex genome (Sarah *et al.,* 2009). Transcriptome sequencing also called RNA sequencing uses NGS technique to identify the quantity of present RNA in a given sample (Wang *et al*., 2009). Genotype by Sequencing (GBS) also called Restriction site Associated DNA method uses restriction enzyme digestion to reduce large repetitive genomes followed by high throughput sequencing. It was developed by Eric Johnson and William Cresko's laboratories at the University of Oregon (2006). It includes digestion of genomic DNA by restriction enzyme followed by adaptors ligation, pooling samples, randomly DNA shearing, size selection, ligation of another adapter, PCR amplification, library preparation and finally sequencing (Baired *et al.,* 2008). This approach facilitates rapid and robust identification of SNPs and small INDELs even with complex genomes (Singh *et al*., 2015).

TABLE 2.4 SNPS IDENTIFIED IN VARIOUS LIVESTOCK 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	Kuchinoshimaushi	1 cow	6303790	93%	15.8X	KwaharaMiki *et al., 2*011
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	Jeju Hengu	1 cow	6484293	98.5%	29.6X	
10	Korean Holstein	1 cow	5814990	98.5%	29.5X	
11	Holstein bull	1 bull	6362988	-	60X	Koks *et al*., 2014
12	Hanwoo & Yanbian cattle	20 (10each)	17976093	98.6%	10.62X	Choi *et al.,* 2015
13	Holstein bull	8 bulls	912302	98.52%	8.1X	Zhang *et al.,* 2016
14	Holstein bull	8 bulls	14821	73.04%	9.6X	Yahui *et al*., 2017
15	Eastern Finncattle, Western Finncattle, and Yakutian cattle	15 (5 each)	17.45 million	98.37%	13.01X	Melak *et al.,* 2018
16	Afrikaner, Drakensberger, and Nguni	3 cow	15442314	-	21.1X	Zwane *et al.*, 2019

RAD sequencing reduces the cost as compared to WGS for SNP discovery (Darvey *et al*., 2010). There are several RAD sequencing methods are available like Complex Reduction of Polymorphism Sequencing (Osrow *et al*., 2007), Reduced Representation Libraries (Van-Tassel *et al.,* 2008). Peterson and his coworker (2012) developed a modified RAD sequencing method which uses two restriction enzymes and DNA size selection step for inexpensive population genotyping. It is used for SNP identification and genotyping (Baired *et al.,* 2008; Willing *et al*., 2011). In ddRAD sequencing, genomic DNA is first digested with a restriction enzyme and then a barcode P1 adaptor is ligated to the DNA fragments, after multiplexing DNA a digested with a second restriction enzyme. Before producing the sequencing library they undergo ligation of P2 adapter and amplification process.

TABLE 2.5 SNPs IDENTIFIED USING RAD SEQUENCING METHODS IN BOVINES

S.NO	SPECIES	NO. OF ANIMALS	METHODS	SNPs	REFERENCE
1	Cattle	47	GBS	52748	De *et al.,* 2013
2	Cattle	1276	GBS	515787	Ibeagha *et al*., 2016
3	Cattle	48	GBS	272103	Jean *et al.,* 2017
4	Buffalo	48	GBS	49607	Imartino *et al*., 2013
5	Buffalo	4	ddRAD	919	Surya, 2018

2.5 BIOINFORMATIC TOOLS FOR SNP IDENTIFICATION

The massive data generated from different NGS techniques were subjected to bioinformatic tools for trimming, quality control, alignment followed by SNP identification (Burt *et al*., 2016). These tools facilitate researchers to retrieve data about SNPs associated with the genes of our interest (Clifford *et al*., 2004). Much of the software working on Linux/ Unix operating system. Next Generation Sequencing technologies combined with fast and reliable bioinformatic softwares to enhance the scope of SNPs identification across different species including complex polyploidy genomes with large repetitive regions (Trick *et al.,* 2009).

TABLE 2.6 List of Bioinformatics Software Tools for Next Generation Sequencing

TASK	TOOLS	REFERENCE
Quality control TOOLS	FastQC	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
	Prinseq	http://prinseq.sourceforge.net/
	FASTX-Toolkit	http://hannonlab.cshl.edu/fastx_toolkit/
	MultiQC	https://multiqc.info/
	QualiMap	http://qualimap.bioinfo.cipf.es/
	Stacks	http://catchenlab.life.illinois.edu/stacks/
Alignment tools	Bowtie2	http://bowtie-bio.sourceforge.net/bowtie2/
	BFAST	http://sourceforge.net/projects/bfast/
	BWA	http://bio-bwa.sourceforge.net/
	SamStat	http://samstat.sourceforge.net/
	Genome Mapper	http://1001genomes.org/
	GMAP	http://www.gene.com/share/gmap/
	MAQ	http://maq.sourceforge.net/
	SOAP	http://soap.genomics.org.cn/
	SWIFT	http://bibiserv.techfak.uni-bielefeld.de/swift
SNP /indel identification	GATK	https://www.broadinstitute.org/gatk/
	SAM Tools	http://samtools.sourceforge.net/
	SOAPsnp	http://soap.genomics.org.cn/soapsnp.html
	VarScan	http://varscan.sourceforge.net/
SNP annotation tools	SNPEff	http://snpeff.sourceforge.net/SnpEff_manual.html
	SnpSift	snpsift.sourceforge.net/
	ANNOVAR	http://annovar.openbioinformatics.org/
	SCAN	http://www.scandb.org/newinterface/about.html
	SNAP	http://www.rostlab.org/services/SNAP
	NGS-SNP	http://stothard.afns.ualberta.ca/downloads/NGS-SNP/
	FAST-SNP	http://fastsnp.ibms.sinica.edu.tw/

2.6 CATTLE GENOME REFERENCE

Recently Hereford taurine genome of 3GB (3 billion base pairs) has been sequenced (Elsik *et al*., 2009), assembled by 2 different groups (Burt 2009) and annotated (Reese *et al*., 2009; Childers *et al*., 2011). However genetic information from indigenous cattle breeds has been limited. The first genome sequence of an indigenous breed (Nellore) generated with 52X coverage by SOLID sequencing platform. The reference genome of cattle (GCA_000003055.5_bos_taurus_3.1.1) is available at NCBI website (www.ncbi.nim.nih.gov/genome?term=bos%20taurus) the bovine reference genome assembly has been revised several times. The newest release of BovineMine (BovineMine v1.5) includes both ARS-UCD1.2 and UMD3.1 genome assemblies.

2.7 ANNOTATION OF THE GENOME

Annotation is a metadatum to predict the effect or the function of an individual SNP using SNP annotation tools. In SNP annotation the biological information is extracted, collected and displayed in a clear form amenable to query.

Liao *et al*. (2013) identified and annotated total 9990733 SNPs and 604308 INDELs in Gir cattle using whole genome sequencing with 4 candidate genes (CAMP, CATHL1, CATHL2, and CATHL3) related to pathogen and parasite-resistance

Jianping *et al.* (2016) identified and annotated 3625 INDELs and 11 promising candidate genes like CENPE, FCGR2B, RETSAT, ACSBG2, NFKB2, TBC1D1, NLK, MAP3K1, SLC30A2, ANGPT1 and UGDH from 8 Holstein bulls responsible for candidate genes affection milk production traits.

Raschia *et al*. (2018) identify candidate genes associated with milk yield in Argentinean dairy cattle with the developed database of candidate genes and SNPs for milk production and composition. Thirty-nine SNPs belonging to 22 candidate genes were genotyped on 1643 animals (Holstein and Holstein x Jersey).

Recently Li *et al*. (2019) studied Genetic association of *DDIT3, RPL23A, SESN2* and *NR4A1* genes with milk yield and composition in Chinese Holstein cattle. A total of 35 SNPs and three insertions/deletions were identified, of which three were found in *DDIT3*, 12 in *RPL23A*, 16 in *SESN2* and seven in *NR4A1*.

For annotation of SNPs available in cattle genome candidate genes responsible for the milk production and composition were collected from available literature.

TABLE 2.7 LIST OF MAJOR CANDIDATE GENES IDENTIFIED THROUGH ASSOCIATION STUDIES FOR MILK PRODUCTION AND COMPOSITION TRAITS IN DAIRY CATTLE

S.NO.	GENE SYMBOL	GENE NAME/DESCRIPTION	CHROMOSOME NO.	REFERENCES
1	CLDN8	Claudin 8	1	Ron *et al*., 2007
2	BTG3	BTG anti-proliferation factor 3	1	Ron *et al*., 2007
3	APOD	Apolipoprotein D	1	Ron *et al*., 2007
4	ST6GAL1	ST6 beta-galactoside alpha-2,6-sialyltransferase 1	1	Ron *et al*., 2007
5	GNB4	G protein subunit beta 4	1	Ron *et al*., 2007
6	SERP1	Stress associated endoplasmic reticulum protein 1	1	Ron *et al*., 2007
7	CP	Ceruloplasmin	1	Ron *et al*., 2007
8	RBP1	Retinol binding protein 1	1	Ron *et al*., 2007
9	COLQ	Collagen like tail subunit of asymmetric acetylcholinesterase	1	Ron *et al*., 2007
10	ESYT3	Extended synaptotagmin 3	1	Ron *et al*., 2007
11	POU1F1	POU class 1 homeobox 1	1	Viale *et al.,* 2017
12	ETS2	ETS proto-oncogene 2, transcription factor	1	Viale *et al.,* 2017
13	CEP63	Centrosomal protein 63	1	Chen *et al*., 2018
14	PDE9A	Phosphodiesterase 9A	1	Jiang *et al.,* 2014
15	DIP2A	Disco interacting protein 2 homolog A	1	Jiang *et al.,* 2014
16	TNFSF10	TNF superfamily member 10	1	Riley *et al*., 2010
17	MIS18A	MIS18 kinetochore protein A	1	Raven *et al.,* 2014

18	FADS1	Fatty acid desaturase 1	1	Bionaz *et al.,* 2008
19	BDH1	3-hydroxybutyrate dehydrogenase 1	1	Bionaz *et al.,* 2008
20	AHSG	Alpha 2-HS glycoprotein	1	D *et al*., 2011
21	SLC37A1	Solute carrier family 37 member 1	1	Raven *et al.,* 2014
22	FABP3	Fatty acid binding protein 3	2	Bionaz *et al.,* 2008
23	DBI/ACBP	Diazepam binding inhibitor, acyl-coa binding protein	2	Ibeagha *et al*., 2016
24	ACTR3	Actin related protein 3	2	Mokhber *et al.,* 2018
25	SOPL	Speckle type BTB/POZ protein like	2	Dong *et al*., 2006
26	HNMT	Histamine N-methyltransferase	2	Sermyagin *et al*., 2018
27	STAT1	Signal transducer and activator of transcription 1	2	Viale *et al.,* 2017
28	SP110	SP110 nuclear body protein	2	Raven *et al.,* 2014
29	SLC40A1	Solute carrier family 40 member 1	2	Fang *et al*., 2014
30	IFIH1	Interferon induced with helicase C domain 1	2	Pimentel *et al*., 2010
31	SDC3	Syndecan 3	2	Raven *et al.,* 2014
32	HSPD1	Heat shock protein family D	2	Ron *et al*., 2007
33	ITGAV	Integrin subunit alpha V	2	Ron *et al*., 2007
34	CYTIP	Cytohesin 1 interacting protein	2	Ron *et al*., 2007
35	CD24	CD24 molecule	2	Ron *et al*., 2007
36	FABP3	Fatty acid binding protein 3	2	Ron *et al*., 2007
37	STMN1	Stathmin 1	2	Ron *et al*., 2007
38	IGFBP2	Insulin like growth factor binding protein 5	2	Nanaei *et al.,* 2019

39	IGFBP5	Insulin like growth factor binding protein 5	2	Ron *et al*., 2007
40	MYL1	Myosin light chain 1	2	Ron *et al*., 2007
41	DGKG	Diacylglycerol kinase gamma	2	Viale *et al.,* 2017
42	SMARCA L1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a like 1	2	Raven *et al*., 2014
43	LEPR	Leptin receptor	3	Viale *et al.,* 2017
44	MUC1	Mucin 1, cell surface associated	3	Raven *et al.,* 2014
45	MACF1	Microtubule actin crosslinking factor 1	3	Sermyagin *et al*., 2018
46	TDRKH	Tudor and KH domain containing	3	Yodklaew *et al*., 2017
47	CERS2/L ASS2	Ceramide synthase 2	3	Rico *et al.,* 2016
48	RHOC	Ras homolog family member C	3	Izumi *et al*., 2016
49	TSHB	Thyroid stimulating hormone beta	3	Yodklaew *et al*., 2017
50	UCK2	Uridine-cytidine kinase 2	3	Ron *et al*., 2007
51	CASQ1	Calsequestrin 1	3	Ron *et al*., 2007
52	KRTCAP2	Keratinocyte associated protein 2	3	Ron *et al*., 2007
53	FDPS	Farnesyl diphosphate synthase	3	Ron *et al*., 2007
54	SSR2	Signal sequence receptor subunit 2	3	Ron *et al*., 2007
55	CRABP2	16090864..16097553	3	Ron *et al*., 2007
56	CTSS	Cathepsin S	3	Ron *et al*., 2007
57	TENT5C	Terminal nucleotidyltransferase 5C	3	Ron *et al*., 2007
58	FRRS1	Ferric chelate reductase 1	3	Ron *et al*., 2007

59	ELOVL1	ELOVL fatty acid elongase 1	3	Ron *et al*., 2007
60	GBA	Glucosylceramidase beta	3	Raven *et al.,* 2014
61	CTTNBP2 NL	CTTNBP2 N-terminal like	3	Raven *et al.,* 2014
62	HSD17B7	Hydroxysteroid 17-beta dehydrogenase 7	3	Cochran *et al*., 2013
63	FAM3C	Family with sequence similarity 3 member C	4	Ron *et al*., 2007
64	AASS	Aminoadipate-semialdehyde synthase	4	Ron *et al*., 2007
65	INSIGI	Insulin induced gene 1	4	Bionaz *et al.,* 2008
66	CACNA2 D1	Calcium voltage-gated channel auxiliary subunit alpha2delta	4	Yuan *et al.,* 2011
67	ETV1	ETS variant 1	4	Raschia *et al*., 2018
68	SNX13	Sorting nexin 13	4	Raschia *et al*., 2018
69	LEP	Leptin	4	Cochran *et al*., 2013
70	DDIT3	DNA damage inducible transcript 3	5	Li *et al.,* 2019
71	OLR1	Oxidized low density lipoprotein receptor 1	5	Khatib *et al*., 2006
72	MKL1/MR TFA	Myocardin related transcription factor A	5	Sun *et al.,* 2006
73	MGST1	Microsomal glutathione S-transferase 1	5	Wang *et al*., 2012
74	HAL	Histidine ammonia-lyase	5	Wang *et al.,* 2014
75	PHLDA1	Pleckstrin homology like domain family A member 1	5	Ron *et al*., 2007
76	RAB3IP	RAB3A interacting protein	5	Ron *et al*., 2007
77	MGP	Matrix Gla protein	5	Ron *et al*., 2007
78	CD9	CD9 molecule	5	Ron *et al*., 2007

79	ARFGAP3	ADP ribosylation factor gtpase activating protein 3	5	Ron *et al*., 2007
80	MGST1	Microsomal glutathione S-transferase 1	5	Ron *et al*., 2007
81	NR4A1	Nuclear receptor subfamily 4 group A member 1	5	Li *et al.,* 2019
82	NCF4	Neutrophil cytosolic factor 4	5	Raven *et al.,* 2014
83	CSF2RB	Colony stimulating factor 2 receptor beta common subunit	5	Raven *et al.,* 2014
84	TXN2	Thioredoxin 2	5	Yodklaew *et al*., 2017
85	LALBA	Lactalbumin alpha	5	Lemay *et al.,* 2009
86	PTHLH	Parathyroid hormone like hormone	5	Ogorevc *et al.,* 2009
87	IGF-1	Insulin like growth factor 1	5	Siadkogska *et al.,* 2006
88	MGP	Matrix Gla protein	5	Geetha *et al*., 2006
89	EPS8	Epidermal growth factor receptor pathway substrate 8	5	Wang *et al*., 2012
90	SOCS2	Suppressor of cytokine signaling 2	5	Riley *et al.,* 2010
91	ATF4	Activating transcription factor 4	5	Raven *et al.,* 2014
92	RPAP3	RNA polymerase II associated protein 3	5	Raven *et al.,* 2014
93	RAP1B	RAP1B, member of RAS oncogene family	5	Niu *et al*., 2016
94	HSP90B1	Heat shock protein 90 beta family member 1	5	Lemay *et al*., 2009
95	VDR	Vitamin D receptor	5	Raven *et al.,* 2014
96	IFNG	Interferon gamma	5	Ogorevc *et al.,* 2009
97	CCDC91	Coiled-coil domain containing 91	5	Pant *et al*., 2010
98	ACSS3	Acyl-coa synthetase short chain family member 3	5	Surya *et al*., 2018
99	GABARAPL1	GABA type A receptor associated protein like 1	5	Pimentel *et al*., 2011

100	ABCG2	ATP binding cassette subfamily G member 2	6	Raschia *et al*., 2018
101	SPPI/OPN	Secreted phosphoprotein 1	6	Raschia *et al*., 2018
102	SLC39A8	Solute carrier family 39 member 8	6	Ron *et al*., 2007
103	PLA2G12A	Phospholipase A2 group XIIA	6	Ron *et al*., 2007
104	SLC34A2	Solute carrier family 34 member 2	6	Ron *et al*., 2007
105	PPARGC1A	PPARG coactivator 1 alpha	6	Raschia *et al*., 2018
106	CSN1S1	Casein alpha s1	6	Raschia *et al*., 2018
107	CXCL8	C-X-C motif chemokine ligand 8	6	Hillreiner *et al*., 2017
108	CSN2	Casein beta	6	Ron *et al*., 2007
109	IGFBP7	Insulin like growth factor binding protein 7	6	Raschia *et al*., 2018
110	CSN3	Casein kappa	6	Raschia *et al*., 2018
111	NAAA/ASAHL	N-acylethanolamine acid amidase	6	Ganguly et al., 2017
112	CENPE	Centromere protein E	6	Jianping *et al.,* 2016
113	TBC1D1	TBC1 domain family member 1	6	Jianping *et al.,* 2016
114	UGDH	UDP-glucose 6-dehydrogenase	6	Jianping *et al.,* 2016
115	EPGN	Epithelial mitogen	6	Raven *et al.,* 2014
116	PKD2	Polycystin 2, transient receptor potential cation channel	6	Tantia *et al.,* 2006
117	BOD1L1	Biorientation of chromosomes in cell division 1 like 1	6	Sermyagin *et al*., 2018
118	IBSP	Integrin binding sialoprotein	6	Cohen *et al.,* 2005
119	PARM1	Prostate androgen-regulated mucin-like protein 1	6	Yodklaew *et al*., 2017

120	EGF	Epidermal growth factor	6	Ogorevec *et al.,* 2009
121	LAP3	Leucine aminopeptidase 3	6	Ogorevec *et al.,* 2009
122	FAM13A	Family with sequence similarity 13 member A	6	Surya *et al*., 2018
123	GRIA1	Glutamate ionotropic receptor AMPA type subunit 1	7	Yodklaew *et al*., 2017
124	RAB3A	RAB2A, member RAS oncogene family	7	D *et al*., 2011
125	CD320	CD320 molecule	7	Ron *et al*., 2007
126	GNA15	G protein subunit alpha 15	7	Ron *et al*., 2007
127	ELL2	Elongation factor for RNA polymerase II 2	7	Ron *et al*., 2007
128	SAR1B	Secretion associated Ras related gtpase 1B	7	Lemay *et al.,* 2009
129	CAST	Calpastatin	7	Yodklaew *et al*., 2017
130	CD14	CD14 molecule	7	Lemay *et al.,* 2009
131	LARP1	La ribonucleoprotein domain family member 1	7	Raven *et al.,* 2014
132	IRF1	Interferon regulatory factor 1	7	Wang *et al.,* 2008
133	TLR4	Toll like receptor 4	8	Viale *et al.,* 2017
134	HPGD	15-hydroxyprostaglandin dehydrogenase	8	Ron *et al*., 2007
135	GLDC	Glycine decarboxylase	8	Ron *et al*., 2007
136	NANS	N-acetylneuraminate synthase	8	Ron *et al*., 2007
137	B4GALT1	Beta-1,4-galactosyltransferase 1	8	Ron *et al*., 2007
138	VLDLR	Very low density lipoprotein receptor	8	Huynh *et al*., 2017
139	SPTLC1	Serine palmitoyltransferase long chain base subunit 1	8	Ganguly *et al*., 2017
140	UGCG	UDP-glucose ceramide glucosyltransferase	8	Ogorevec *et al.,* 2009

141	PLIN2/ADFP	Perilipin 2	8	Mohammad *et al.,* 2013
142	LPL	Lipoprotein lipase	8	Viale *et al.,* 2017
143	NFIB	Nuclear factor I B	8	Robinson *et al*., 2014
144	GNA14	G protein subunit alpha 14	8	Lemay *et al.,* 2009
145	FBP1	Fructose-bisphosphatase 1	8	Ostrowska *et al.,* 2013
146	FBP2	Fructose-bisphosphatase 2	8	Ostrowska *et al.,* 2013
147	TPD52L1	TPD52 like 1	9	Ron *et al*., 2007
148	CITED2	Cbp/p300 interacting transactivator with Glu/Asp rich carboxy-terminal domain 2	9	Ron *et al*., 2007
149	UBE3D	Ubiquitin protein ligase E3D	9	Sermyagin *et al*., 2018
150	TEP1	Telomerase associated protein 1	10	Ibeagha *et al*., 2016
151	SPTLC2	Serine palmitoyltransferase long chain base subunit 2	10	Bionaz *et al*., 2008
152	RAB11A	RAB11A, member RAS oncogene family	10	Lemay *et al.,* 2009
153	DHRS1	Dehydrogenase/reductase 1	10	Ron *et al*., 2007
154	CCNB2	Cyclin B2	10	Ron *et al*., 2007
155	ARG2	Arginase 2	10	Ron *et al*., 2007
156	RAB11B	RAB11B, member RAS oncogene family	10	D *et al*., 2011
157	RORA	RAR related orphan receptor A	10	Nanaei *et al.,* 2019
158	PCK2	Phosphoenolpyruvate carboxykinase 2, mitochondrial	10	Mohammad *et al.,* 2012
159	PAEF/LGB	Progestagen-associated endometrial protein	11	Raschia *et al*., 2018
160	GLT6D1	Glycosyltransferase 6 domain containing 1	11	Raven *et al.,* 2014

161	MCFD2	Multiple coagulation factor deficiency 2	11	Ron *et al*., 2007
162	SLC1A4	Solute carrier family 1 member 4	11	Ron *et al*., 2007
163	RAB1A	RAB1A, member RAS oncogene family	11	Ron *et al*., 2007
164	LCN2	Lipocalin 2	11	Ron *et al*., 2007
165	TOR1B	Torsin family 1 member B	11	Ron *et al*., 2007
166	CEL	Carboxyl ester lipase	11	Ron *et al*., 2007
167	APOB	Apolipoprotein B	11	Venturini *et al.,* 2014
168	RETSAT	Retinol saturase	11	Jianping *et al.,* 2016
169	LPIN1	Lipin 1	11	Viale *et al.,* 2017
170	MAP4K4	Mitogen-activated protein kinase kinase kinase kinase 4	11	Han *et al*., 2019
171	ID2	Inhibitor of DNA binding 2	11	Ron *et al*., 2007
172	MFGE8	Milk fat globule-EGF factor 8 protein	11	Ogorevc *et al*., 2009
173	XDH	Xanthine dehydrogenase	11	Viale *et al.,* 2017
174	NLRP6	NLR family pyrin domain containing 6	11	Raven *et al.,* 2014
175	GFI1B	Growth factor independent 1B transcriptional repressor	11	Raven *et al.,* 2014
176	PRKCE	Protein kinase C epsilon	11	Surya *et al*., 2018
177	NRXN1	Neurexin 1	11	Raven *et al.,* 2014
178	TNFSF11	TNF superfamily member 11	12	Ron *et al*., 2007
179	UFM1	Ubiquitin fold modifier 1	12	Ron *et al*., 2007
180	GJB6	Gap junction protein beta 6	12	Ron *et al*., 2007
181	GJB2	Gap junction protein beta 2	12	Ron *et al*., 2007

182	RAB20	RAB20, member RAS oncogene family	12	Ron *et al*., 2007
183	LCP1	Lymphocyte cytosolic protein 1	12	Lemay *et al.,* 2009
184	RNF219	Ring finger protein 219	12	Raven *et al.,* 2014
185	MATN4	Matrilin 4	13	Wathes *et al*., 2009
186	ACSS1	Acyl-coa synthetase short chain family member 1	13	Bionaz *et al.,* 2008
187	OSBPL2	Oxysterol binding protein like 2	13	Nayeri *et al.,* 2017
188	YWHAB	Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein beta	13	Lemay *et al.,* 2009
189	TFAP2C	Transcription factor AP-2 gamma	13	Ron *et al*., 2007
190	RAB18	RAB18, member RAS oncogene family	13	Ron *et al*., 2007
191	GHRH	Growth hormone releasing hormone	13	Parmentier *et al.,* 1999
192	RAB18	RAB18, member RAS oncogene family	13	Ron *et al*., 2007
193	AVP	Arginine vasopressin	13	Zingg *et al.,* 1988
194	ACSS2	Acyl-coa synthetase short chain family member 2	13	Bionaz *et al.,* 2008
195	KIZ/PLK1S1	Kizuna centrosomal protein	13	Surya *et al*., 2018
196	ANGPT1	Angiopoietin 1	14	Jianping *et al.,* 2016
197	COL22A1	Collagen type XXII alpha 1 chain	14	Choi *et al*., 2009
198	PPP1R16A	Protein phosphatase 1 regulatory subunit 16A	14	Jiang *et al.,* 2014
199	CPSF1	Cleavage and polyadenylation specific factor 1	14	Nayeri *et al*., 2016
200	ARHGAP39	Rho gtpase activating protein 39	14	Nayeri *et al*., 2017
201	MAF1	MAF1 homolog, negative regulator of RNA polymerase III	14	Jiang *et al.,* 2014

202	FOXH1	Forkhead box H1	14	Surya *et al*., 2018
203	NIBP/TRAPPC9	Trafficking protein particle complex 9	14	Surya *et al*., 2018
204	NDRG1	N-myc downstream regulated 1	14	Ron *et al*., 2007
205	TRAM1	Translocation associated membrane protein 1	14	Ron *et al*., 2007
206	AZIN1	Antizyme inhibitor 1	14	Ron *et al*., 2007
207	CA2	Carbonic anhydrase 2	14	Ron *et al*., 2007
208	GRINA	Glutamate ionotropic receptor NMDA type subunit associated protein 1	14	Nayeri *et al*., 2017
209	CYHR1	Cysteine and histidine rich 1	14	Jiang *et al.,* 2014
210	RHPN1	Rhophilin Rho gtpase binding protein 1	14	Jiang *et al.,* 2014
211	PTK2	Protein tyrosine kinase 2	14	Raven *et al.,* 2014
212	PLEC	Plectin	14	Raven *et al.,* 2014
213	ZNF696	Zinc finger protein 696	14	Li *et al*., 2015
214	VPS28	VPS28 subunit of ESCRT-I	14	Jiang *et al.,* 2014
215	ADCK5	Aarf domain containing kinase 5	14	Ibeagha *et al*., 2016
216	TONSL	Tonsoku like, DNA repair protein	14	Ibeagha *et al*., 2016
217	KCNK9	Potassium two pore domain channel subfamily K member 9	14	Jiang *et al.,* 2014
218	TG	Thyroglobulin	14	Anton *et al.,* 2009
219	OPLAH	5-oxoprolinase, ATP-hydrolysing	14	Jiang *et al.,* 2014
220	MAPK15	Mitogen-activated protein kinase 15	14	Jiang *et al.,* 2014
221	EEF1D	Eukaryotic translation elongation factor 1 delta	14	Jiang *et al.,* 2014

222	MYC	MYC proto-oncogene, bhlh transcription factor	14	Raven *et al*., 2014
223	GML	Glycosylphosphatidylinositol anchored molecule like	14	Jiang *et al.,* 2014
224	KCNK9	Potassium two pore domain channel subfamily K member 9	14	Jiang *et al.,* 2014
225	CEBPD	CCAAT enhancer binding protein delta	14	Raven *et al*., 2014
226	FAM83H	Family with sequence similarity 83 member H	14	Jiang *et al.,* 2014
227	CYP11B1	Cytochrome P450, subfamily XI B, polypeptide	14	Jiang *et al.,* 2014
228	LRRC14	Leucine rich repeat containing 14	14	Ibeagha *et al*., 2016
229	SMPD5	Sphingomyelin phosphodiesterase 5	14	Da *et al*., 2019
230	MROH1	Maestro heat like repeat family member 1	14	Tesfayonas *et al*., 2014
231	AGO2	Argonaute RISC catalytic component 2	14	Liu *et al*., 2015
232	KHDRBS3	KH RNA binding domain containing, signal transduction associated 3	14	Surya *et al.,* 2018
233	ZC3H3	Zinc finger CCCH-type containing 3	14	Jiang *et al.,* 2014
234	MAPK15	Mitogen-activated protein kinase 15	14	Jiang *et al.,* 2014
235	OSBP	Oxysterol binding protein	15	Bionaz *et al.,* 2008
236	DEPDC7	DEP domain containing 7	15	Yodklaew *et al.,* 2017
237	PGR	Progesterone receptor	15	Raven *et al*., 2014
238	NEU3	Neuraminidase 3	15	Cochran *et al.,* 2013
239	MMP12	Matrix metallopeptidase 12	15	Ron *et al*., 2007
240	TAGLN	Transgelin	15	Ron *et al*., 2007
241	FXYD2	FXYD domain containing ion transport regulator 2	15	Ron *et al*., 2007
242	NUCB2	Nucleobindin 2	15	Ron *et al*., 2007

243	HBB	Hemoglobin subunit beta	15	Ron *et al*., 2007
244	STARD10	Star related lipid transfer domain containing 10	15	Ron *et al*., 2007
245	COMMD9	COMM domain containing 9	15	Ron *et al*., 2007
246	CD82	CD82 antigen-like	15	Ron *et al*., 2007
247	CD44	CD44 molecule	15	Kolbehdari *et al.,* 2009
248	DHRS3/SDR1	Dehydrogenase/reductase 3	16	Lemay *et al.,* 2009
249	PYCR2	Pyrroline-5-carboxylate reductase 2	16	Ron *et al*., 2007
250	PEX14	Peroxisomal biogenesis factor 14	16	Ron *et al*., 2007
251	EDEM3	ER degradation enhancing alpha-mannosidase like protein 3	16	Ron *et al*., 2007
252	LAMB3	Laminin subunit beta 3	16	Ron *et al*., 2007
253	PIGR	Polymeric immunoglobulin receptor	16	De *et al.,* 2001
254	ATF3	Activating transcription factor 3	16	Han *et al* ., 2017
255	LAX1	Lymphocyte transmembrane adaptor 1	16	Iso *et al*., 2016
256	ACACB	Acetyl-coa carboxylase beta	17	Sun *et al.,* 2009
257	TCN2	Transcobalamin 2	17	Ron *et al*., 2007
258	FGA	Fibrinogen alpha chain	17	Lemay *et al.,* 2009
259	FBRSL1	Fibrosin like 1	17	Chen *et al.,* 2018
260	DTX1	Deltex E3 ubiquitin ligase 1	17	Sermyagin *et al.,* 2018
261	RILPL2	Rab interacting lysosomal protein like 2	17	Ibeagha *et al*., 2016
262	IL15	Interleukin 15	17	Sermyagin *et al*., 2018
263	FGF2	Fibroblast growth factor 2	17	Viale *et al.,* 2017

264	ARHGAP 35/GRLF1	Rho gtpase activating protein 35	18	Viale *et al.,* 2017
265	CARD15/ NOD2	Nucleotide binding oligomerization domain containing 2	18	Viale *et al.,* 2017
266	SPHK2	Sphingosine kinase 2	18	Li *et al*., 2914
267	PGLYRP1	Peptidoglycan recognition protein 1	18	Karthikeyan *et al.,* 2016
268	SLC7A5	Solute carrier family 6 member 2	18	Ron *et al*., 2007
269	DPEP3	Dipeptidase 3	18	Ron *et al*., 2007
270	SLC6A2	Solute carrier family 6 member 2	18	Ron *et al*., 2007
271	RHPN2	Rhophilin Rho gtpase binding protein 2	18	Ron *et al*., 2007
272	LGALS7	Galectin 7	18	Ron *et al*., 2007
273	MIA	MIA SH3 domain containing	18	Ron *et al*., 2007
274	C5AR1	Complement C5a receptor 1	18	Ogorevc *et al.,* 2009
275	MTDH	Metadherin	18	Ron *et al*., 2007
276	HIF3A	Hypoxia inducible factor 3 subunit alpha	18	Ron *et al*., 2007
277	IRX3	Iroquois homeobox 3	18	Ron *et al*., 2007
278	APOE	Apolipoprotein E	18	Lemay *et al.,* 2009
279	NLRP9	NLR family pyrin domain containing 9	18	Yodklaew *et al*., 2017
280	TGFB1	Transforming growth factor beta 1	18	Raven *et al.,* 2014
281	FASN	Fatty acid synthase	19	Raschia *et al*., 2018
282	NLK	Nemo like kinase	19	Legarra *et al*., 2001
283	SREBF1	Sterol regulatory element binding transcription factor 1	19	Angulo *et al*., 2012
284	TP53	Tumor protein p53	19	Ron *et al*., 2007

285	ALDOC	Aldolase, fructose-bisphosphate C	19	Ron *et al*., 2007
286	ENO3	Enolase 3	19	Ron *et al*., 2007
287	DVL2	Dishevelled segment polarity protein 2	19	Ron *et al*., 2007
288	SLC35B1	Solute carrier family 35 member B1	19	Ron *et al*., 2007
289	GRN	Granulin precursor	19	Ron *et al*., 2007
290	NBR1	NBR1 autophagy cargo receptor	19	Ron *et al*., 2007
291	GPS1	G protein pathway suppressor 1	19	Ron *et al*., 2007
292	LLGL2	LLGL scribble cell polarity complex component 2	19	Ron *et al*., 2007
293	ACLY	ATP citrate lyase	19	Ron *et al*., 2007
294	ST6GALNAC2	ST6 N-acetylgalactosaminide alpha-2,6-sialyltransferase 2	19	Ron *et al*., 2007
295	ARHGDIA	Rho GDP dissociation inhibitor alpha	19	Lemay *et al.,* 2009
296	RAB5C	RAB5C, member RAS oncogene family	19	D *et al*., 2011
297	KRT9	Keratin 9	19	Lemay *et al.,* 2009
298	GHDC	GH3 domain containing	19	Raven *et al.,* 2014
299	AP2B1	Adaptor related protein complex 2 subunit beta 1	19	Kolbehdari *et al*., 2009
300	ACLY	ATP citrate lyase	19	Raven *et al*., 2014
301	TP53	Tumor protein p53	19	Ogorevc *et al*., 2009
302	BAIAP2	BAI1 associated protein 2	19	Kolbehdari *et al*., 2009
303	STAT5B	Signal transducer and activator of transcription 5B	19	Khatib *et al*., 2008
304	GHR	Growth hormone receptor	20	Raschia *et al*., 2018; Viale *et al.,* 2017
305	RAB3C	RAB3C, member RAS oncogene family	20	D *et al*., 2011

306	PRLR	Prolactin receptor	20	Raschia *et al*., 2018; Viale *et al.,* 2017
307	MAP3K1	Mitogen-activated protein kinase kinase kinase 1	20	Jianping *et al.,* 2016
308	PELO	Pelota mrna surveillance and ribosome rescue factor	20	Ron *et al*., 2007
309	DAB2	DAB adaptor protein 2	20	Ron *et al*., 2007
310	DAP	Death associated protein	20	Ron *et al*., 2007
311	GDNF	Glial cell derived neurotrophic factor	20	Jiang *et al.,* 2014
312	NIPBL	NIPBL cohesin loading factor	20	Jiang *et al.,* 2014
313	CD180/LY64	CD180 molecule	20	Wang *et al*., 2015
314	RICTOR	RPTOR independent companion of MTOR complex 2	20	Nayeri *et al*., 2016
315	LIFR	LIF receptor alpha	20	Raven *et al.,* 2014
316	OSMR	Oncostatin M receptor	20	Jiang *et al.,* 2014
317	ISL1	ISL LIM homeobox 1	20	Raven *et al.,* 2014
318	EDC3	Enhancer of mrna decapping 3	21	Ryu *et al.,* 2016
319	SERPINA1/PI	Serpin family A member 1	21	Viale *et al.,* 2017
320	PLIN1	Perilipin 1	21	Viale *et al.,* 2017
321	ISG20	Interferon stimulated exonuclease gene 20	21	Holm *et al.,* 2018
322	PDIA3	Protein disulfide isomerase family A member 3	21	Lemay *et al.,* 2009
323	SCAMP2	Secretory carrier membrane protein 2	21	Reinhardt *et al.,* 2006
324	MFGE8	Milk fat globule-EGF factor 8 protein	21	Ron *et al*., 2007
325	CACNA1D	Calcium voltage-gated channel subunit alpha1 D	22	Yodklaew *et al*., 2017

326	ATP2B2	ATPase plasma membrane Ca2+ transporting 2	22	Ogorevc *et al*., 2009
327	PPARG	peroxisome proliferator activated receptor gamma	22	Bionaz *et al.,* 2008
328	CDCP1	CUB domain containing protein 1	22	Izumi *et al.,* 2013
329	CMTM6	CKLF like MARVEL transmembrane domain containing 6	22	Ron *et al*., 2007
330	FEZF2	FEZ family zinc finger 2	22	Ogorevc *et al.,* 2009
331	CMTM8	CKLF like MARVEL transmembrane domain containing 8	22	Ron *et al*., 2007
332	OSBPL10	oxysterol binding protein like 10	22	Ganguly *et al*., 2017
333	RAB7A	RAB7A, member RAS oncogene	22	Lemay *et al.,* 2009
334	LTF	lactotransferrin	22	Raschia *et al*., 2018
335	GOLGA4	golgin A4	22	Yodklaew *et al*., 2017
336	CDKN1A	cyclin dependent kinase inhibitor 1A	23	Han *et al*., 2017
337	BTN1A1	butyrophilin subfamily 1 member A1	23	Ogorevc *et al*., 2009
338	ECI2	enoyl-CoA delta isomerase 2	23	Sermyagin *et al*., 2018
339	GNMT	glycine N-methyltransferase	23	Ron *et al*., 2007
340	ELOVL5	ELOVL fatty acid elongase 5	23	Ron *et al*., 2007
341	TNXB	tenascin XB	23	Ron *et al*., 2007
342	LST1	leukocyte specific transcript 1	23	Ron *et al*., 2007
343	AGPAT1	1-acylglycerol-3-phosphate O-acyltransferase 1	23	Ron *et al*., 2007
344	UBD	Ubiquitin D	23	Ron *et al*., 2007
345	IRF4	Interferon regulatory factor 4	23	Ron *et al*., 2007
346	BPHL	Biphenyl hydrolase like	23	Ron *et al*., 2007

347	TRIM26	Tripartite motif containing 26	23	Ron *et al*., 2007
349	HSPA1A	Heat shock protein family A (Hsp70) member 1A	23	Gurskly *et al.,* 2016
350	JARID2	Jumonji and AT-rich interaction domain containing 2	23	Fang *et al.,* 2014
351	ATP5A1	ATP synthase F1 subunit alpha	24	Lemay *et al.,* 2009
352	FHOD3	Formin homology 2 domain containing 3	24	Ron *et al*., 2007
353	CIDEA	Cell death inducing DFFA like effector a	24	Ron *et al*., 2007
354	DSC2	Desmocollin 2	24	Yodklaew *et al*., 2017
355	DTX2	Deltex E3 ubiquitin ligase 2	25	Yodklaew *et al*., 2017
356	ACTB	Actin beta	25	Yadav et al., 2012
357	IGFALS	Insulin like growth factor binding protein acid labile subunit	25	Ron *et al*., 2007
358	TNP2	Transition protein 2	25	Ron *et al*., 2007
359	EEF2K	Eukaryotic elongation factor 2 kinase	25	Ron *et al*., 2007
360	KIF22	Kinesin family member 22	25	Ron *et al*., 2007
361	TUFM	Tu translation elongation factor, mitochondrial	25	Ron *et al*., 2007
362	CLDN3	Claudin 3	25	Ron *et al*., 2007
363	FAM20C	FAM20C golgi associated secretory pathway kinase	25	Ron *et al*., 2007
364	BAIAP2L1	BAI1 associated protein 2 like 1	25	Ron *et al*., 2007
365	PMM2	Phosphomannomutase 2	25	Cochran et al., 2013
366	PAM16	Presequence translocase associated motor 16	25	Nayeri *et al*., 2016
367	GNB2	G protein subunit beta 2	25	Lemay *et al.,* 2009
368	GLIS2	GLIS family zinc finger 2	25	Li *et al.,* 2019

369	CLEC16A	C-type lectin domain containing 16A	25	Iso *et al.,* 2016
370	NFKB2	Nuclear factor kappa B subunit 2	26	Jianping *et al.,* 2016
371	ACTA2	Actin alpha 2, smooth muscle	26	Ron *et al*., 2007
372	KIF11	Kinesin family member 11	26	Ron *et al*., 2007
373	ALDH18A1	Aldehyde dehydrogenase 18 family member A1	26	Ron *et al*., 2007
374	LIPA	Lipase A, lysosomal acid type	26	Ron *et al*., 2007
375	CTBP2	C-terminal binding protein 2	26	Ron *et al*., 2007
376	MKI67	Marker of proliferation Ki-67	26	Ron *et al*., 2007
377	GPAM	Glycerol-3-phosphate acyltransferase, mitochondrial	26	Bionaz *et al.,* 2008
378	DMBT1	Deleted in malignant brain tumors 1	26	D *et al.,* 2016
379	BTRC	Beta-transducin repeat containing E3 ubiquitin protein ligase	26	Raven *et al.,* 2014
380	PRKG1	Protein kinase cgmp-dependent 1	26	Li *et al.,* 2014
381	SUFU	SUFU negative regulator of hedgehog signaling	26	Raven *et al.,* 2014
382	NEURL1	Neuralized E3 ubiquitin protein ligase 1	26	Iso *et al.,* 2016
383	DKK1	Dickkopf WNT signaling pathway inhibitor 1	26	Raven *et al.,* 2014
384	ACSL1	Acyl-coa synthetase long chain family member 1	27	Bionaz *et al.,* 2008
385	GPAT4	Glycerol-3-phosphate acyltransferase 4	27	Wang *et al*., 2012
386	MFHAS1	Malignant fibrous histiocytoma amplified sequence 1	27	Ron *et al*., 2007
387	GINS4	GINS complex subunit 4	27	Raven *et al.,* 2014
388	SGPL1	Sphingosine-1-phosphate lyase 1	28	Bionaz *et al*., 2008
389	SAR1A	Secretion associated Ras related gtpase 1A	28	Lemay *et al.,* 2009

390	KCNK1	Potassium two pore domain channel subfamily K member 1	28	Lemay *et al.,* 2009
391	MS4A8/MS4A8B	Membrane spanning 4-domains A8	29	Velmala *et al*., 1995
392	FADS1	Fatty acid desaturase 1	29	Ibeagha *et al*., 2016
393	THRSP	Thyroid hormone responsive	29	Bionaz *et al*., 2008
394	FADS2	Fatty acid desaturase 2	29	Ibeagha *et al*., 2016
395	CTSC	Cathepsin C	29	Ron *et al*., 2007
396	SPTBN2	Spectrin beta, non-erythrocytic 2	29	Ron *et al*., 2007
397	CAPN6	Calpain 6	X	Ron *et al*., 2007
398	GJB1	Gap junction protein beta 1	X	Ron *et al*., 2007
399	TIMP1	TIMP metallopeptidase inhibitor 1	X	Ron *et al*., 2007
400	PRDX4	Peroxiredoxin 4	X	Ron *et al*., 2007

CHAPTER –3

Materials & Methods

MATERIALS AND METHODS

3.1 LOCATION OF THE STUDY

The study was conducted in Animal Genetics and Breeding division (AGB Division), the Livestock research center (LRC); NDRI-ICAR and bioinformatics analysis at ICAR NBAGR.

3.2 S OURCE OF SAMPLES

Samples were collected aseptically in LRC; ICAR-NDRI farm from 6 Gir cattle.

3.3 BLOOD SAMPLE COLLECTION

Blood samples were collected from the 6 Gir cows. About 10ml of venous blood was collected from the jugular vein aseptically in a 15ml polypropylene centrifuge tubes under sterile condition using 0.5 ml of EDTA as an anticoagulant. Collected blood samples were immediately transported to the laboratory and stored in the refrigerator at -20 degree Celcius until further processing.

3.4 DNA EXTRACTION

DNA isolation from the blood samples was done by the Phenol: Chloroform method as described by Sambrook *et al. (*2001) with slight modifications.

Phenol-chloroform extraction is a liquid-liquid extraction technique in molecular biology used to separate nucleic acids from proteins and lipids.

The protocol for isolation of genomic DNA is as follows:

- The blood samples stored at - 20◦C were thawed to room temperature.
- The tubes were filled with chilled RBC lysis buffer (1x), mixed end to end, incubated in ice for 10 min and centrifuged at 12000 rpm for 8-10 min at room temperature.
- The reddish tinged supernatant containing plasma and lysed RBC were discarded by pipetting.

- Steps 2-3 were repeated, till the WBC pellet appeared nearly white in color.
- About 5 ml of DNA extraction buffer was added in the tubes and incubated at 56℃ for overnight.
- For overnight incubation equal volume of phenol (pH 8.0) was added to the samples and mixed gently.
- These tubes were centrifuged at 12000 rpm for 10 min at 25℃.
- After centrifugation, the upper aqueous layer was collected without disturbing the organic layer with the help of Pasteur pipette and the aqueous solution was transferred to a fresh tube.
- To this aqueous layer phenol: chloroform: isoamyl alcohol (25:24:1) was added and the solution was mixed gently. These samples were centrifuged at 12000 rpm for 10 min at 25℃.
- The upper aqueous layer was collected without disturbing the organic layer and chloroform: isoamyl alcohol (24:1) was added to the aqueous layer in equal volume and mixed gently.
- These tubes were centrifuged at 12000 rpm for 10 min at 25℃, the upper aqueous layer was collected.
- 500µl of 3M sodium acetate and 2 volumes chilled ethanol was added to this aqueous layer. Tubes were inverted for 3-4 times. DNA got precipitated at this step.
- DNA was spooled and washed twice with 70% ethanol.
- The tubes were air dried at room temperature till the alcohol evaporates.
- To this DNA precipitate, an appropriate amount of TE buffer (500 µl) was added.
- The DNA pellet was dissolved for inactivation of nuclease; the tubes were incubated at 65℃/30 min.
- The isolated DNA samples were checked on agarose gels.

3.5 QUALITY AND QUANTITY CHECKING OF GENOMIC DNA

Agarose gel electrophoresis was carried out for checking the quality of DNA. 2 µl DNA mixed with 2 µl of 6X gel-loading dye was loaded in 0.8% agarose gel [0.8gm agarose + 10 ml TBE (10X) + 90 ml double distilled water] in horizontal mini electrophoresis unit using 1X TBE as running buffer at 100 volts for 45 minutes. The gel was stained with 1% ethidium bromide solution and a photograph was taken by a gel documentation system (Bio-Rad). Sharp and intact bands indicated good quality DNA. Quality and quantity of DNA were also estimated by nanodrop spectrophotometer. DNA Samples with OD_{260}/OD_{280} ranging between 1.7 and 1.9 were of good quality, while the quantity of DNA was calculated using the following formula.

{Quantity of DNA (µg/ml) = OD_{260} x dilution factor x 50}

3.6 GENOMIC LIBRARY PREPARATION

About 1 microgram of genomic DNA was first digested with a suitable restriction enzyme, SphI (CATG) and MluCI (AATT). After clean-up using Ampure beads, barcoded P1 adapter was ligated to the fragments using T4 DNA ligase. The adapter-ligated fragments from different samples are combined if samples are multiplexed, and the DNA is digested by a second restriction enzyme. The fragments were size-selected using 2% agarose gel electrophoresis and purified. The P2 adapter-primers are ligated, and the fragments are amplified by PCR amplification to enrich and add the Illumina specific adapters and flowcell annealing sequences. The library prepared were then pooled and sent for sequencing using Illumina HiSeq 2000 sequencer.

3.7 BIOINFORMATIC PIPELINE

3.7.1 FASTQC

Fastqc (http;//www.bioinformatics.babraham.c.uk/projects/fastqc/) is a quality control application that facilitates numerous quality control checks on raw sequence data generated by high throughput sequencing pipelines such as Illumina and ABI SOLiD platforms in fastq format. It generates as an output comprehensive multi-

page report on the composition and quality of reads in HTML format, with one page for each of the reads (e.g. Single End, Paired End: forward, Paired End: reverse)**.** The HTML report includes results from multiple modules that were run by fastqc, and provides a quick assessment of the quality of the results labeled as normal (green checkmark), slightly abnormal (orange triangle), and very unusual (red cross) reads.

The modules included in the report are as follows:

- **BASIC STATISTICS:** provides introductory compositional statistics such as filename, file type, encoding, total sequences, number of sequences flagged as poor quality, sequence length and %GC for the analyzed read.
- **PER BASE SEQUENCE QUALITY:** displays an overview of the range of quality values across all bases at each position in the fastq file. The graph is vertically partitioned into three quality ranges: good (green), reasonable (orange), poor (red).
- **PER SEQUENCE QUALITY SCORES:** displays the quality score distribution over all the sequences, which allows users to see if a subset of the sequences has universally low-quality values.
- **PER BASE SEQUENCE CONTENT:** displays the proportion of each DNA base (A, T, C, G) called at a given position in all the sequences.
- **PER SEQUENCE GC CONTENT:** displays the GC distribution over all the sequences across the whole length and compares it to a modeled normal distribution of GC content.
- **PER BASE N CONTENT:** displays the percentage of base calls at each position for which an N was called. N at a given position indicates the inability to make a normal base call with sufficient confidence.
- **SEQUENCE LENGTH DISTRIBUTION:** displays the distribution of fragment sizes across all sequences and is highly dependent on the sequencing platform.
- **SEQUENCE DUPLICATION LEVELS:** counts the degree of duplication for every sequence in a library and creates a plot showing the proportion of

FIGURE 2: PER BASE SEQUENCE QUALITY

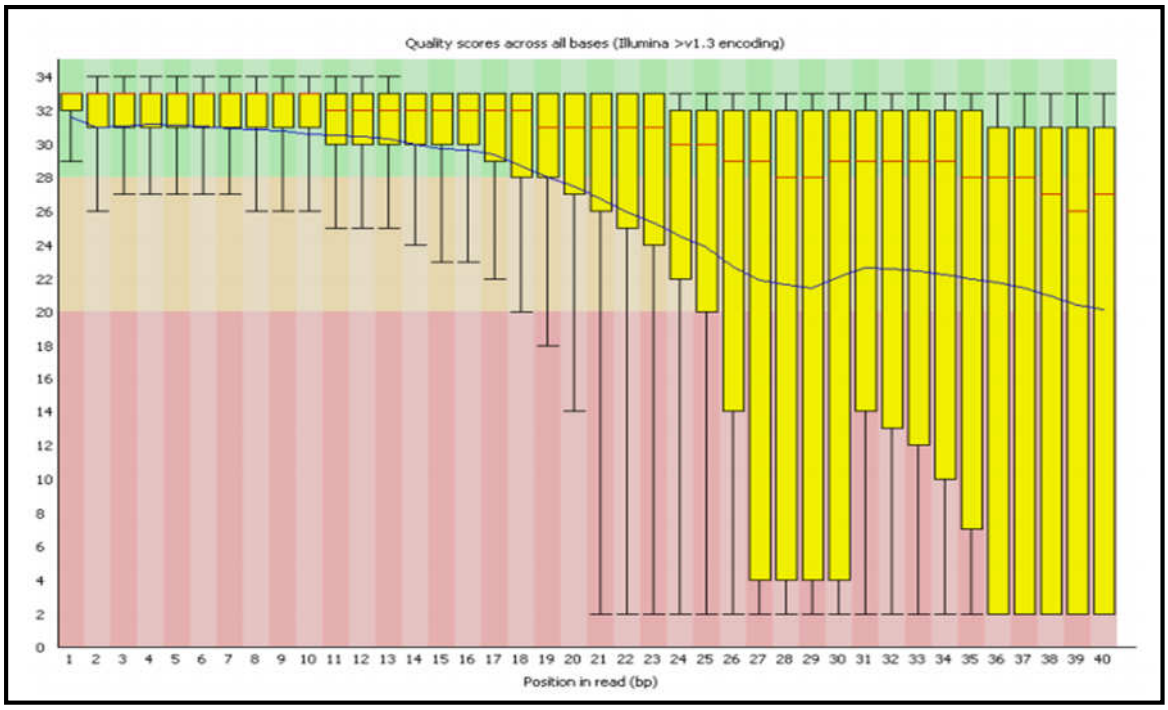


FIGURE 3: PER SEQUENCE QUALITY SCORE

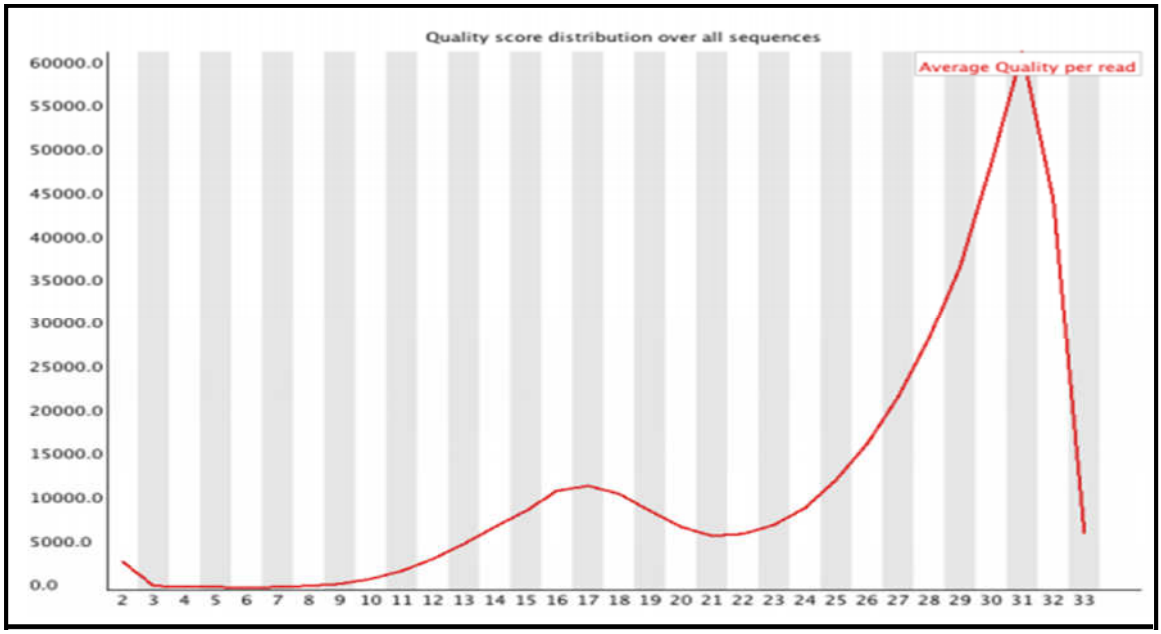


FIGURE 4 : PER BASE SEQUENCE CONTENT

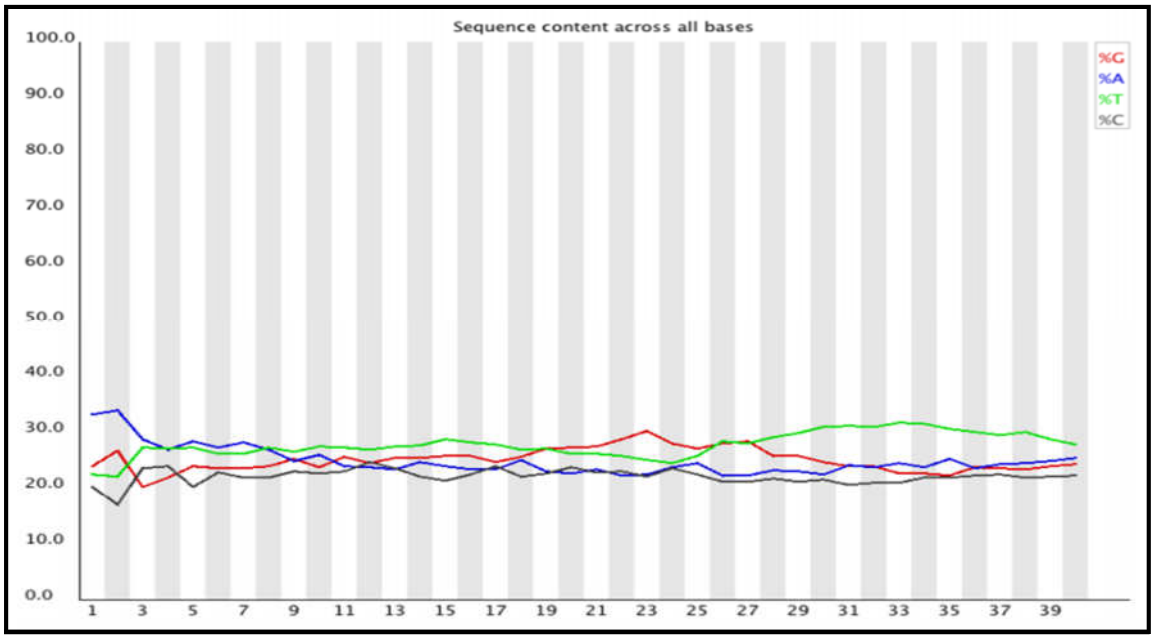


FIGURE 5 : PER SEQUENCE GC CONTENT

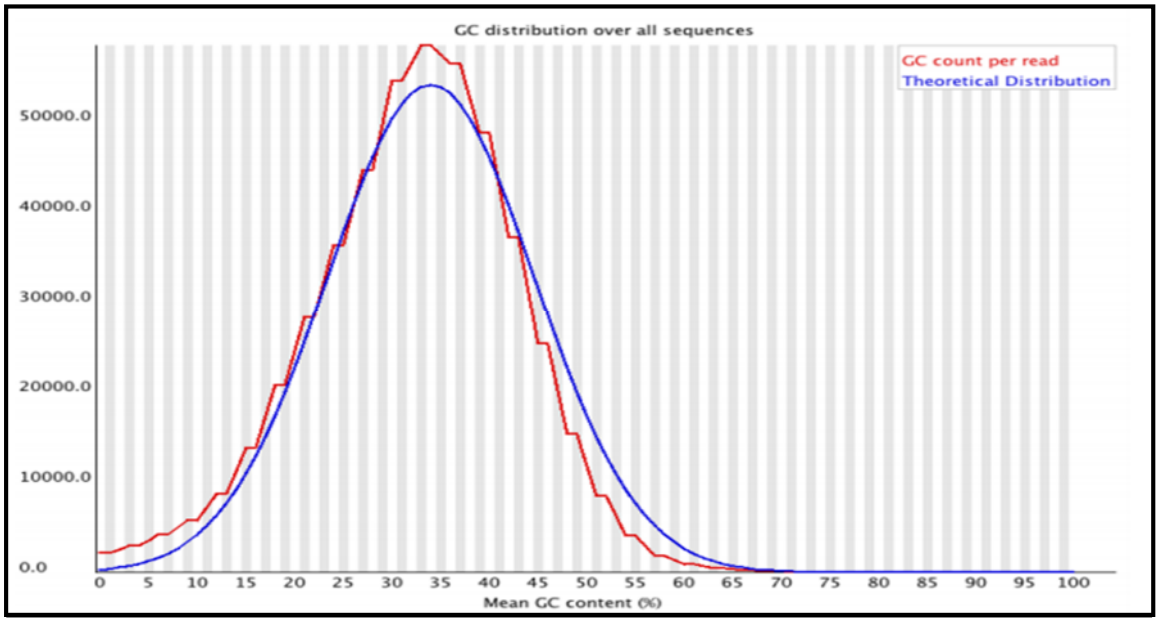


FIGURE 6: PER BASE N CONTENT

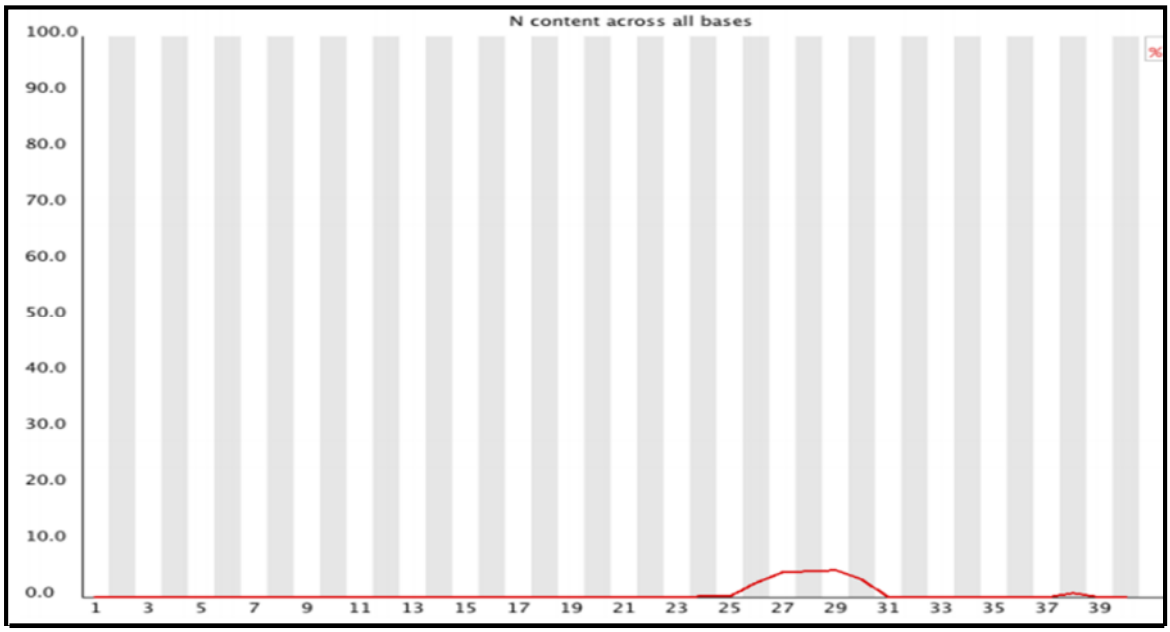


FIGURE 7: SEQUENCE LENGTH DISTRIBUTION

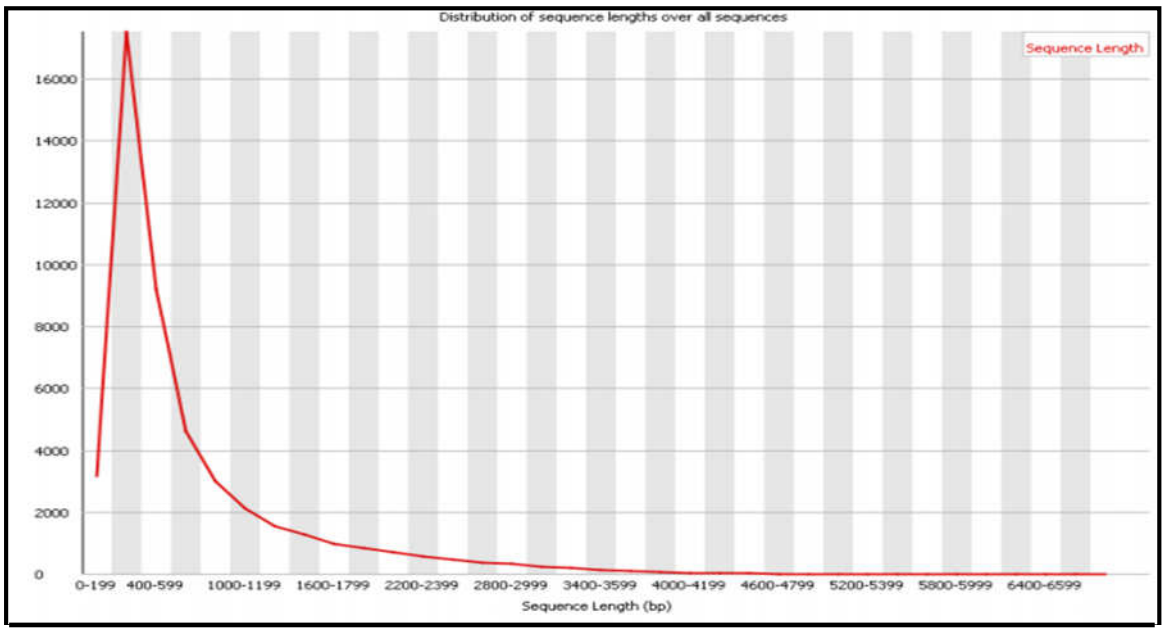


FIGURE 8: SEQUENCE DUPLICATE LEVELS

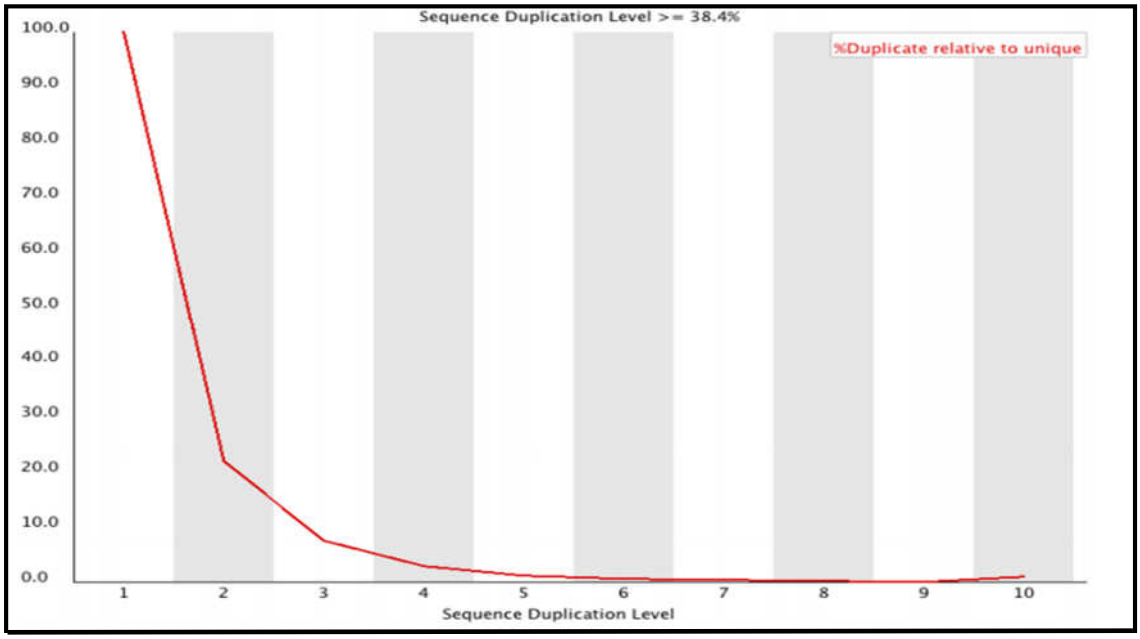
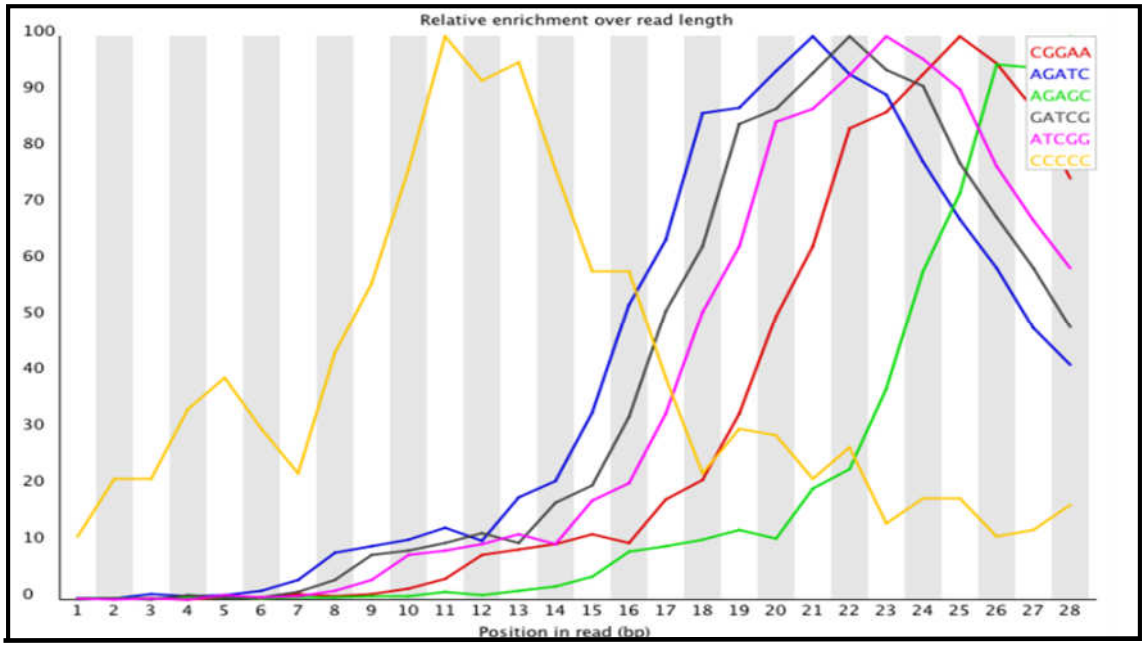


FIGURE 9: OVERPRESENTED KMERS



sequences with different degrees of duplication (in blue) and de-duplicated sequences (in red).

- **OVERREPRESENTED SEQUENCES:** lists all of the sequence which makes up more than 0.1% of the total. An overrepresented sequence implies that either it is highly biologically significant, the library is contaminated, or is not that diverse.
- **ADAPTER CONTENT:** displays the proportion of sequences which have an adapter sequence at a given position. This is informative in deciding whether there is a significant amount of adapter present in the sequences and can be subjected to trimming.
- **K-mer CONTENT:** Displays relative K-mer (k=7) enrichment over the read length for the top six K-mers. More specifically, it measures the number of each 7-mer at each position in the library and then uses a binomial test to look for significant deviations from an even coverage at all positions and reports the 7-mers with positionally biased enrichment.

3.7.2 ADAPTER TRIMMING

The data received from the Illumina sequencer as fastq files. The first requirement is to demultiplex, the raw data to recover the individual samples in the Illumina library. While doing this, we used the scores provided in the fastq files to discard sequencing reads of low quality. These tasks are accomplished by the software prinseq and stacks. Prinseq software (http://prinseq.sourcefore.net/) was used to filter, reformat, or trim genomic and metagenomic sequence data. It generates summary statistics of sequences in graphical and tabular format. We used data in fasta (and qual) format or fastq format as input. The files were compressed with the zip, gzip, or bzip2 algorithm to reduce upload time. It provided summary statistics for data including reading length, GC content, sequence complexity, and quality score distributions, the number of read duplicates, the occurrence of Ns and poly-A/T tails, assembly quality measures, tag sequences, and more.

3.7.3 QUALITY CONTROL

The poor quality reads, primer/adaptor contamination and base calling error which affects the final result interpretation was done by the stacks software**.** Stacks was created by Julian Catchen *et al.,* 2011 at the University of Oregon. This program specializes in identifying and genotyping loci from next-generation sequencing data, where large numbers of short reads collected throughout the genome using restriction enzyme cut sites. Stacks uses a sliding-window quality filtering algorithm that allows for isolated low-quality base calls while discarding reads that show a degenerating quality level across the read length. The program will also demultiplex data according to a set of barcodes and check for the presence of a restriction enzyme cut site for both single and double-digested data. The process rad-tags program can also filter adapter sequence from raw reads. It removed the reads which do not have cut site for restriction enzymes and having a quality score of less than 15 (phred score).

3.7.4 ALIGNMENT

The paired-end alignment was done by bowtie2. Bowtie sequence aligner was originally developed by Langmead *et al. (*2009) at the University of Maryland (2009). On 16 October 2011, the developers released a beta fork of the project called Bowtie2. It is an ultrafast and memory-efficient tool for aligning sequencing reads to long references sequences Bowtie2 takes a bowtie2 index and a set of sequencing read files and outputs a set of alignments in SAM format.

3.7.5 VARIANT CALLING

Samtools (http://github.com/samtools/samtools/releases) provide various utilities for manipulating alignments in the SAM (Sequence Alignment/Map) format, including sorting, merging, indexing and generating alignments in a BAM (Binary alignment format) format. The mapped reads were used for SNP identification using samtools software program considering phred quality score ≥ 30 and read depth of 2, 5 and 10. Samtools uses a reference genome and a file with aligned reads (BAM file) to call variants. We need to first convert the SAM to its binary counterpart, BAM

format. The binary format is much easier for computer programs to work with. To convert SAM to BAM, we must specify that our input is in SAM format using the -S option and the output in the BAM format with the -b option. Samtools convert SAM file to BAM file for sorting and call variants from the sorted BAM file which resulted as the output of VCF file. The variant call format (VCF) is a generic format used for storing DNA polymorphism data such as SNPs, insertions, deletions, and structural variants, together with rich annotations. VCF is usually stored in a compressed manner and can be indexed for fast data retrieval of variants from a range of positions on the reference genome. The aim of vcftools (http://snpeff.sourceforge.net) is to provide easily accessible methods for working with complex genetic variation data in the form of VCF files.

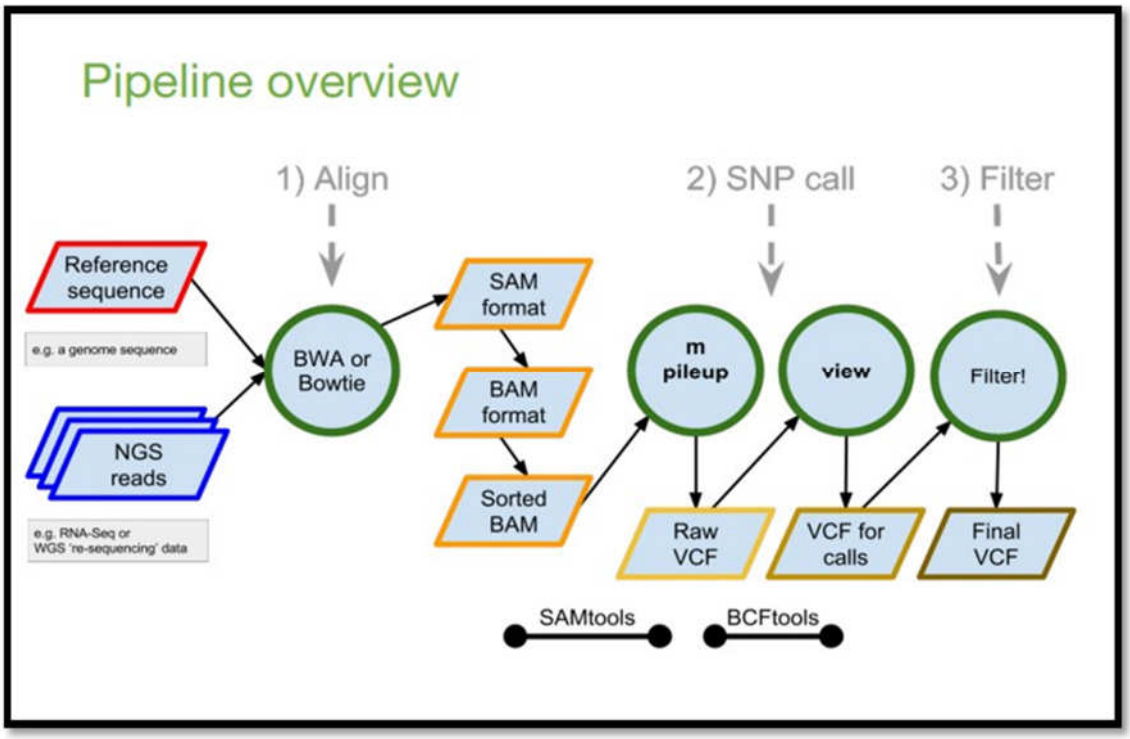
This toolset can be used to perform the following operations on VCF files:-

- ❖ Filter out specific variants
- ❖ Compare files
- ❖ Summarize variants
- ❖ Validate and merge files
- ❖ Create intersections and subsets of variants.

3.7.6 ANNOTATION

Annotation was done by the software SnpEff (http://snpeff.sourceforge.net) and SnpSift (http://snpsift.sourceforge.net). SnpEff software is used to predict the effect or function of an individual SNP. It annotates and predicts the effects of genetic variants on genes and proteins (such as amino acid changes). Genes will be collected from the literature responsible for the Milk production and composition traits. Information regarding gene ID, gene name and gene description will collect from the NCBI website (http://www.ncbi.nlm.nih.gov/). The SNPs falling in those genes were annotated for their position in the gene, chromosome number, reference, and alternate allele. Gene-wise SNP information was done by the software SnpSift.

FIGURE 10 - SNP CALLING PIPELINE



CHAPTER –4

Results and Discussion

RESULTS AND DISCUSSION

In the present study, we identified SNPs by using ddRAD (double Digest

Restriction Associated DNA) sequencing approach in the six samples of Gir cattle. The sequenced data generated by Illumina hiseq 2000 sequencer were processed to obtain genome-wide SNPs.

4.1 ddRAD LIBRARY PREPARATION

Quality of extracted DNA was checked by the agarose gel electrophoresis and quantity was checked by Nanodrop spectrophotometer (1.8 OD). High-quality DNA was then digested with a restriction enzyme (Sphl and MluCl). The digested DNA was then separated in agarose gel electrophoresis along with the ladder. Size selection was carried out with the fragment size between 100-200 bp. The isolated and purified DNA samples were then ligated with specific barcodes, followed by PCR amplification, attachment of Illumina specific adapters and flowcell annealing sequencing. The library from each sample was then pooled and sequenced using Illumina HiSeq 2000.

4.2 RAW READ CHARACTERISTICS

From the sequencing of 6 samples of Gir cattle, a total of 13133430 (13.1 million) raw reads were obtained in fastq format. An average of 2.18 million reads was obtained per sample.

4.3 BIOINFORMATIC ANALYSIS

The obtained raw reads were then analyzed to identify SNPs among different samples of Gir cattle. For this, a workflow has been designed which includes demultiplexing, adapter trimming, filtration, sequence alignment, SNP identification, and annotation.

4.3.1 QUALITY CONTROL OF RAW DATA

The Raw data obtained after ddRAD sequencing platform are subjected to quality control step to remove sequence artifacts such as base calling errors,

INDELs, poor quality reads and adaptor contamination. After demultiplexing, adapter trimming and filtering the ddRAD sequences, we retained a total of 12892759 (12.8 Million) reads across all the samples, with an average of approximately 2 million reads per sample. The initial data analyses have revealed that the read quality (Phred score) within 6 samples is greater than 30 indicating that the data produced is of good quality and the remaining poor quality reads were discarded.

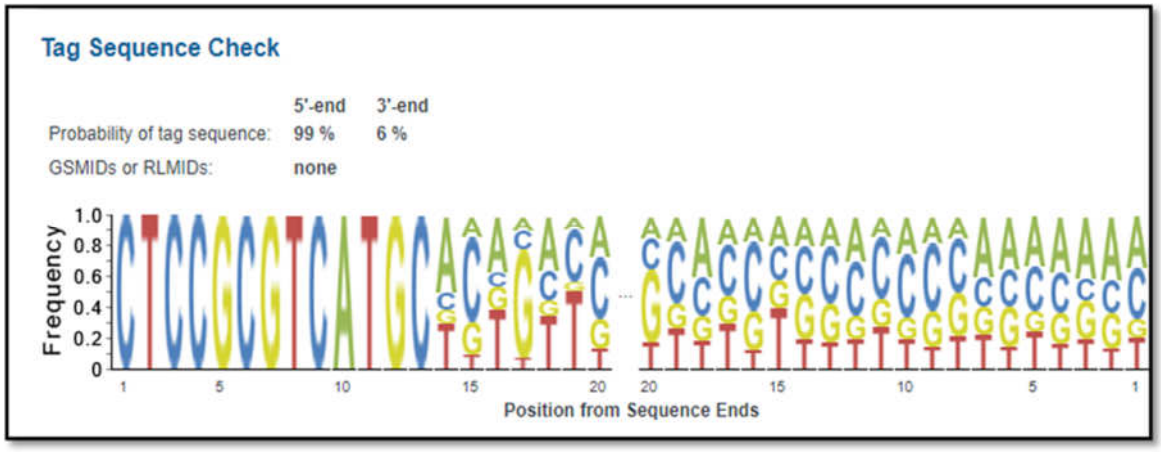
TABLE 4.1 SUMMARY OF THE SEQUENCING READS BEFORE AND AFTER QUALITY CONTROL OF SEQUENCE READS

SAMPLE NUMBER	RAW READS	PROCESSED READS
1	765970	754342
2	3310806	3223269
3	3305652	3249045
4	1283522	1264553
5	840808	829157
6	3626672	3572393
Total	13133430 (13.1 M)	12892759 (12.8 M)
QC % OF PROCESSED GOOD QUALITY READS		98.16%

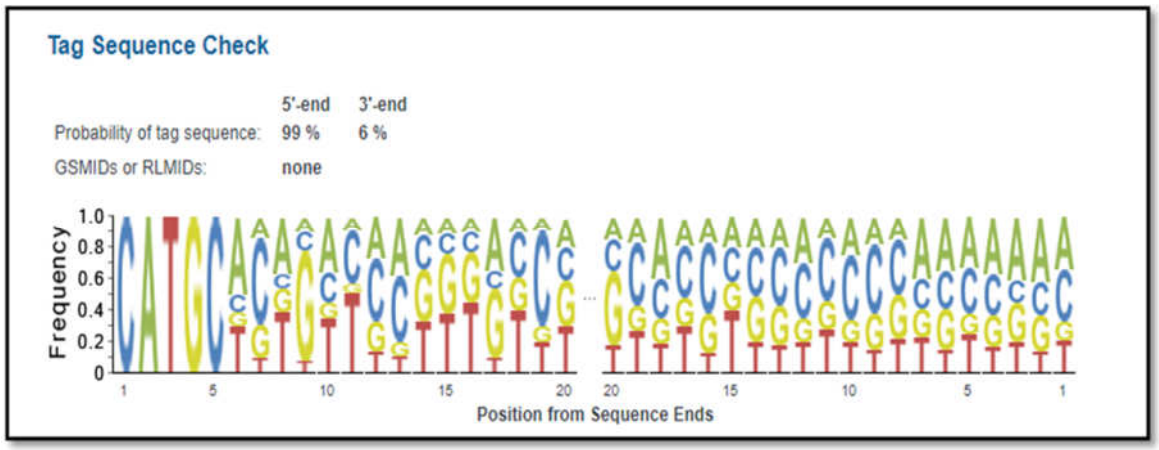
4.3.2 MAPPING AND ALIGNMENT OF READS

After quality control step by using software bowtie2, good quality processed reads were aligned to the reference genomes of *Bos taurus, Bos indicus*, and Gir separately which is available at NCBI website (www.ncbi.nim.nih.gov/ genome?term=bos%20taurus). The percentages of overall aligned reads with reference to *Bos taurus, Bos indicus,* and Gir genomes was 99.76%, 90.36%, and 98.29% respectively. A total of 6.38%, 6.36%, and 6.38% coverage was covered by the generated processed reads with the reference genome of *Bos Taurus, Bos*

FIGURE 11. SCREENSHOT OF THE SEQUENCE READ BEFORE (A) AND AFTER (B) ADAPTER TRIMMING



(A)



(B)

indicus and Gir respectively. Approximately 21.31%, 26.91% and 21.36% of the high quality read in the samples were uniquely mapped to the *Bos taurus, Bos indicus*, and Gir reference genome respectively.

4.3.3 VARIANT CALLING

In this study with reference to *Bos taurus* genome, we identified a total of 193764, 179205 and 160951 SNPs; 12082, 11059 and 9802 INDELs at RD2, RD5 and RD10 respectively. Similarly, when compared to the *Bos indicus* reference genome, we identified a total 117467, 109943 and 99517 SNPs; 12082, 11059 and 9802 INDELs at RD2, RD5 and RD10 respectively. Against Gir reference genome we identified a total of 174492, 162470 and 146564 SNPs; 10887, 10050 and 9010 INDELs at RD2, RD5, and RD10 respectively. The Ts/Tv ratio was found to be 2.62.

4.4 SNP ANNOTATION

4.4.1 FUNCTIONAL CHARACTERIZATION OF SNPs

Annotation of SNPs using SnpEff, revealed that the highest number of SNPs were present in the transcript region (30.99%), followed by intron region (30.58%), intergenic region (25.90%), upstream region (5.79%) and downstream region (5.70%) and the lowest number of SNPs were present in the UTR_5_PRIME region (0.11%).

TABLE 4.2 SUMMARY OF NUMBER OF EFFECT BY IMPACT AND FUNCTIONAL CLASS

(A) NUMBER OF EFFECT BY IMPACT

TYPE	COUNT	PERCENT
High	18	0.007%
Low	734	0.302%
Moderate	385	0.159%
Modifier	241,582	99.532%

(B) NUMBER OF EFFECT BY FUNCTIONAL CLASS

TYPE	COUNT	PERCENT
Missense	387	39.051%
Nonsense	6	0.605%
Silent	598	60.343%
Missense / Silent ratio		0.6472

TABLE 4.3 SUMMARY OF BASE CHANGES AND Ts/Tv RATIO

(A) BASE CHANGES SNPs

	A	C	G	T
A	0	3,662	15,708	2,196
C	3,859	0	4,291	20,458
G	20,643	4,285	0	3,831
T	2,049	15,447	3,386	0

(B) Ts/Tv RATIO (TRANSITION/TRANSVERSION)

Transitions	271,828
Transversions	103,483
Ts/Tv ratio	2.6268

4.4.2 CHROMOSOME-WISE VARIANT IDENTIFICATION

Chromosome-wise variant identification reveals a total of 99815 variants along with the total variant rate of 26789.

SNPEFF VARIANT ANALYSIS:-

FIGURE 12. NUMBER OF EFFECTS BY TYPE AND REGION

Number of effects by type and region

Type

Type (alphabetical order)	Count	Percent
3_prime_UTR_variant	732	0.301%
5_prime_UTR_premature_start_codon_gain_variant	47	0.019%
5_prime_UTR_variant	236	0.097%
downstream_gene_variant	13,852	5.704%
intergenic_region	62,875	25.893%
intragenic_variant	1,364	0.562%
intron_variant	74,340	30.614%
missense_variant	385	0.159%
non_coding_transcript_exon_variant	360	0.148%
non_coding_transcript_variant	73,864	30.418%
splice_acceptor_variant	9	0.004%
splice_donor_variant	1	0%
splice_region_variant	99	0.041%
start_lost	2	0.001%
stop_gained	6	0.002%
synonymous_variant	598	0.246%
upstream_gene_variant	14,058	5.789%

Region

Type (alphabetical order)	Count	Percent
DOWNSTREAM	13,852	5.707%
EXON	1,339	0.552%
INTERGENIC	62,875	25.904%
INTRON	74,245	30.589%
SPLICE_SITE_ACCEPTOR	9	0.004%
SPLICE_SITE_DONOR	1	0%
SPLICE_SITE_REGION	97	0.04%
TRANSCRIPT	75,228	30.994%
UPSTREAM	14,058	5.792%
UTR_3_PRIME	732	0.302%
UTR_5_PRIME	283	0.117%

FIGURE 13. CHANGES IN AMINO ACID DUE TO SNPs

	*	?	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
*																						
?		19																				
A			76																19	12		
C				8														1				1
D					20	1		1						19						4		2
E	2				1	11		2			6					1						
F							19					3						2				2
G	1		3	10		5		37									3	3		1		
H									23							6	4					3
I							2			12									3	4		
K	1					6				4	17						8					
L							1			2		47			11	4	2	1		5		
M												1					1		7	7		
N										18				33				2				
P			1									12			79	1		12	9			
Q									4		1				1	24	5					
R	1			18				4	4		13	1				8	26	1			5	
S			1	1			2	9				1		2	2			61	4			
T			19							6	1		4		2			4	46			
V			8					1		6		3	4							29		
W																						
Y	1								3													11

TABLE 4.4 CHROMOSOME-WISE VARIANTS AND VARIANT RATE AT RD10

CHROMOSOME	VARIANTS	VARIANT RATE
NC_032650.1	5141	31,337
NC_032651.1	4471	31,465
NC_032652.1	4389	29,133
NC_032653.1	4407	28,234
NC_032654.1	3854	32,600
NC_032655.1	4087	30,008
NC_032656.1	3835	29,191
NC_032657.1	3770	31,018
NC_032658.1	3641	29,689
NC_032659.1	3750	28,349
NC_032660.1	4332	25,451
NC_032661.1	3653	23,389
NC_032662.1	3562	23,703
NC_032663.1	3309	24,602
NC_032664.1	3365	25,200
NC_032665.1	3398	22,927
NC_032666.1	3072	24,908
NC_032667.1	3148	20,949
NC_032668.1	2778	23,512
NC_032670.1	3069	22,583
NC_032671.1	2679	23,102
NC_032672.1	2501	21,323
NC_032673.1	2903	22,190
NC_032674.1	2405	18,313
NC_032675.1	2145	24,177
NC_032676.1	2251	21,656
NC_032677.1	2302	20,028
NC_032678.1	2246	23,210
NC_032679.1	2076	42,638
NC_032680.1	48	821,272
TOTAL	99,815	26,798

4.4.3 GENE WISE VARIANT ANNOTATION

For gene-wise SNP annotation, we collected candidate genes responsible for the milk production and composition traits from the available literature and cgQTL database (http://cowry.agri.huji.ac.il/QTLMAP/qtlmap.htm). The information regarding Gene symbol, Gene name, Gene location is taken from the NCBI website (http://www.ncbi.nlm.nih.gov/). SNPs present in those genes were annotated for their position, reference allele, and alternate allele by using the SNPsift software. Out of 400 collected genes, a total of 984 SNPs were annotated in 175 genes affecting milk production and composition traits (Milk yield, Fat yield, Protein yield, Milk fat percentage, and protein percentage).

DISCUSSION

In the genome-wide SNPs identification study using ddRAD approach, an average of 1.4 million reads was obtained, using GBS method in 47 samples of cattle (De *et al*., 2013). The average number of raw reads by GBS method was 0.87 million reads per sample in cattle (De *et al*., 2013). In the present study average reads obtained per sample was comparatively higher than the above-mentioned study. Because, in the present study, the ddRAD method was used instead of the GBS method, along with two restriction enzymes in library preparation in order to cover the maximum number of loci across the genome.

When using Prinseq alone number of raw reads was lost in the quality control processing in buffalo (Upadhyay *et al*., 2015; Patel *et al.,* 2017). So, in the present study, raw reads were screened for assessing their quality prior to quality control in the fastqc software program and in Prinseq program, adapter and barcodes were removed. The reads with low quality were discarded with minimum loss of sequences as possible in Stacks software.

In this study for alignment, Bowtie2 was used due to its ultrafast and memory-efficient in aligning paired-end sequences to the reference genome of complex species (Langmead *et al*., 2012; Langdon, 2013). In the present study, the overall alignment for mapped reads was less compared to other genome-wide

studies. Because in the ddRAD approach based upon which type of restriction enzyme used in library preparation it cover the genome as well as the number of loci in that genome. So, it may be the reason in obtaining a lesser number of alignment rate compared to other studies.

A total of 52,748 SNPs were identified in 47 animals using GBS method (De *et al.,* 2013). But, it is comparatively less than SNPs identified in the present study; the reason may be ddRAD approach compared to other RAD approaches covers the maximum region of the genome with the minimum number of samples (Peterson *et al*., 2012). A total of 107488 SNPs were identified by GBS approach in 24 animals belonging to seven Indian cattle breeds (Malik *et al.,* 2018) and 8065 high-quality SNPs in 48 individuals of cattle (Gurgul *et al.,* 2019). A total of 238725 high-quality SNPs was reported in indigenous cattle breeds of China using RAD sequencing approach (Wang *et al.,* 2018) and 10058 SNPs in 40 dairy cows using a modified RAD sequencing method (Yang *et al.,* 2018). When double enzyme GBS method was used, 272103 SNPs were identified in 48 dairy cows (Brouard *et al.,* 2017).

Ts/Tv ratio ranging for was 2.0-2.2 (Choi *et al.,* 2014; Kawahara *et al*., 2011; Stothard *et al.,* 2011). In whole-genome sequencing studies for variant identification in cattle, the Ts/Tv ratios were 2.2 in *Bos taurus, Bos indicus* and their crossbreds (Stothard *et al.,* 2011; Choi *et al*., 2014; Stafuzza *et al.,* 2017). However, in the present study, the Ts/Tv ratio obtained was higher than the ratio of whole-genome sequencing. Because in this study whole genome-reduced representation method used rather than whole genome sequencing may be the reason for getting a higher ratio than the above studies. Higher Ts/Tv ratios have also been reported in targeted sequencing approaches as well in the other reduced representation approaches for genome-wide SNP discovery (Kraus *et al*., 2011; Le and Durbin, 2011; Ba *et al*., 2017).

Annotation of the genes associated with Milk production and composition traits in the present study was done by the SnpEff which shows the highest number of SNPs present in the transcript region followed by the intron and intergenic region.

The maximum number of candidate genes responsible for Milk production and composition traits were found on chromosome number 14. Among all the collected candidate genes like PRKG1, FHOD3, TG, GRIA1, OSBPL10, ATP2B2, ACACB, COL22A1, PRKCE revealed a number of SNPs affecting Milk production and composition trait. PRKG1 having a maximum of 67 SNPs mainly responsible for the protein percentage. Nanaei *et al*. (2019) showed that several candidate genes such as CSN3, IGFBP-2, RORA, ABCG2, B4GALT1, and GHR are positively selected for milk production traits in Kenana cattle. Recently Lung *et al*. (2019) studied genome-wide association study for Milk production trait in Brazilian Holstein population. He showed that candidate genes like MGST1, DGAT1, PAEP, COL18A1, LTTC19, SLC3JA1, LTBB1, MFSD4A were associated with milk yield, Somatic cell score, fat percentage, fatty acid composition. Han *et al*. (2019) detected the polymorphisms of LPIN1 and verified their genetic effects on milk yield and composition in a Chinese Holstein cow population. Li *et al*. (2019) genetic effects of HSPA8 and ERBB2 on milk protein concentration in a large Chinese Holstein population and to evaluate the genetic effects of both genes on other milk production traits. There were 2 single nucleotide polymorphisms (SNPs) identified for HSPA8 and 11 SNPs for ERBB2 by sequencing 17 unrelated Chinese Holstein sires. Raschia *et al*. (2018) identified Single nucleotide polymorphisms in 22 candidate genes like *ETV1, LEP, ABCG2, OPN, PPARGC1A, CSN1S1, CSN2, CSN3, LGB, DGAT1, GH, GHR, PRLR, LTF,* and *PRL* ,*ARL4A, SNX13, ORL1, STAT5A, FASN, UTMP* and *SCD1* are associated with milk yield in Argentinean Holstein and Holstein x Jersey cows. Viale *et al*. (2017) studied association of candidate genes like CCL3, AGPAT6, DGKG, PPARGC1A, CSN1S1, GHR, TLR4, POU1F1 and CSN2 with milk production traits, yield, composition and somatic cell score in Italian Holstein-Friesian sires. Grisart *et al*. (2002) studied a missense mutation in the DGAT1 gene on chromosome 14 that has been identified to have major effects on milk composition and fat content in dairy cattle.

The SNPs identified in this study may respond as a useful tool in future studies particularly in genome-wide association studies for better understanding of genetic structure revealing the phenotypic difference in cattle.

TABLE4.5: GENE-WISE ANNOTATION OF SNPs FOR THE CANDIDATE GENES RESPONSIBLE FOR MILK PRODUCTION AND COMPOSITION TRAITS IN DAIRY CATTLE

S.NO	GENE SYMBOL	CHROMOSOME NO.	GENE NAME	SNP POSITION	REFERENCE ALLELE	ALTERNATE ALLELE
1	ETS2	1	ETS proto-oncogene 2, transcription factor	154600344	G	C
	DGKG	1	Diacylglycerol Kinase gamma	82870399	C	G
2	CEP63	1	Centrosomal Protein 63	137148341	G	A
3	PDE9A	1	Phosphodiesterase 9A	145866522	T	C
				145866634	T	A
				145866677	C	T
				145867984	A	G
				145868201	A	G
				145912157	T	G
				145912325	C	T
				145919132	G	A
				145919165	C	T
				145942570	G	A
				145942617	G	A
				145942633	G	A
				145942636	G	A
				145942639	G	A
				145942674	T	C
4	DIP2A	1	Disco interacting protein 2 homolog A	149209084	G	A
5	BDH1	1	3-hydroxybutyrate dehydrogenase 1	73131951	T	C
				73132000	T	C

6	SLC37A1	1	Solute carrier family 37 member 1	145789863	T	C
7	COLQ	1	Collagen like tail subunit of asymmetric acetylcholinesterase	156005593	A	G
				156009862	G	A
				156009871	A	C
				156009895	G	A
				156009898	A	G
				156037847	G	T
				156037859	G	A
				156037889	G	A
				156043048	T	C
8	APOD	1	Apolipoprotein D	73215625	G	T
				73215737	C	T
				73215739	G	C
				73215799	A	G
9	ST6GAL1	1	ST6 beta-galactoside alpha-2,6-sialyltransferase 1	82070124	G	A
				82164825	T	C
				82165225	C	T
				82165235	T	C
				82165242	A	C
				82165243	A	T
				82165262	T	A
				82183228	A	G
				82183268	A	G
				82183294	T	C
10	GNB4	1	G protein subunit beta 4	89773512	T	C
				89831057	C	T
11	CP	1	Ceruloplasmin	120893934	T	C

				120893936	G	A
12	RBP1	1	Retinol binding protein 1	131812060	G	C
				131812080	C	T
				131835746	T	G
				131860169	G	T
14	HNMT	2	Histamine N-methyltransferase	62020705	A	G
				62020797	A	G
				62040740	G	T
				62040755	A	G
15	SMARCAL1	1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a like 1	108592394	T	C
				108592460	C	T
				108592496	G	A
16	IGFBP2	2	Insulin like growth factor binding protein 2	108800257	G	A
				108800305	A	G
				108800317	A	G
				108800321	C	T
				108800323	G	A
				108800324	T	C
				108800395	C	T
17	ITGAV	2	Integrin subunit alpha V	10270018	A	G
18	CYTIP	2	Cytohesin 1 interacting protein	40262777	T	C
				40262813	G	A
				40262918	G	A
				40262938	T	C
				40311972	G	C
				40312014	C	A
				40320129	C	T

19	IGFBP5	2	Insulin like growth factor binding protein 5	108823833	C	T
				108823876	G	A
				108827155	C	A
				108827581	A	G
20	LEPR	3	Leptin receptor	85659344	T	C
21	MACF1	3	Microtubule actin crosslinking factor 1	113640762	C	T
				113644803	C	G
				113667741	A	G
				113667827	C	A
				113690551	G	A
				113725299	T	C
				113725377	T	C
				113786642	A	G
				113786649	G	A
				113786671	A	C
				113786696	G	A
				113786722	C	T
				113793826	C	T
				113793843	G	A
				113793910	G	A
				113847682	T	A
				113847689	G	T
				113847771	G	C
				113847808	G	A
				113847872	G	C
22	TDRKH	3	Tudor and KH domain containing	20522383	C	G
				20522394	A	G

				20522504	T	C
23	RHOC	3	Ras homolog family member C	33141042	G	T
24	UCK2	3	Uridine-cytidine kinase 2	3707827	G	C
				3707859	C	A
				3707863	T	C
				3732925	C	T
				3733028	A	G
				3740715	T	C
				3740762	G	A
				3740799	C	T
				3750234	T	C
				3750272	A	G
				3750319	C	T
				3763031	C	T
				3763093	A	G
				3763247	T	C
				3763254	A	G
				3763268	A	G
				3763329	G	T
				3763334	T	C
				3763335	G	A
25	KRTCAP2	3	Keratinocyte associated protein 2	16824123	A	G
26	CTSS	3	Cathepsin S	21519963	T	C
27	TENT5C	3		27889716	C	T
				27889758	G	C
				27889769	A	G
			Terminal nucleotidyltransferase 5C	27889795	A	G

				27895174	A	G
28	FRRS1	3		46451163	C	T
			Ferric chelate reductase 1	46451274	C	T
29	ELOVL1	3	ELOVL fatty acid elongase 1	109569820	C	T
30	CACNA2D1	4	CTTNBP2 N-terminal like	39814037	T	G
				39814182	G	A
				39814189	T	C
				39814190	G	A
				39814197	C	G
				39909781	C	A
				39909881	C	A
				39909961	A	G
				39959414	G	A
				39959444	A	G
31	LEP	4	Leptin	95663586	T	C
				95663647	C	T
				95663754	T	C
				95663931	T	C
				95671426	G	T
				95671439	T	C
				95671453	G	A
				95671482	A	G
				95671493	A	G
32	ETS1	4	ETS variant 1	23090579	A	G
33	SNX13	4	Sorting nexin 13	27562488	T	C
				27617000	A	C
				27617069	G	A

				27617084	A	G
34	AASS	4	Aminoadipate-semialdehyde synthase	89538970	A	G
				89539054	C	T
				89542760	G	A
				89542794	C	T
				89542881	A	G
				89542883	C	T
				89543323	T	G
				89543400	T	C
				89543448	G	C
				89543471	A	G
				89544872	C	T
35	MKL1/MRTFA	5	Myocardin related transcription factor A	118742041	T	G
				118742042	A	G
				118742043	T	A
				118742044	A	G
				118742145	C	A
				118742200	G	A
				118742218	G	A
				118756794	A	T
				118776341	T	G
				118790268	G	A
				118790326	C	A
36	MGST1	5	Microsomal glutathione S-transferase 1	100159790	G	C
				100159834	A	T
				100159842	A	G

37	EPS8	5	Epidermal growth factor receptor pathway substrate 8	100814065	G	A
				100814068	A	G
				100951755	T	A
38	VDR	5	Vitamin D receptor	35571110	C	G
				35581713	C	T
				35621647	T	G
39	CCDC91	5	Coiled-coil domain containing 91	87207729	A	C
				87207736	C	G
				87207976	T	C
				87396973	A	G
				87399972	G	A
				87400070	G	A
40	ACSS3	5	Acyl-coa synthetase short chain family member 3	12815539	C	T
				12857830	T	C
				12857838	A	G
				12857850	T	G
				12858716	G	A
				12928985	C	A
				12929160	G	C
41	RAB3IP	5	RAB3A interacting protein	47199321	C	T
42	CD9	5	CD9 molecule	110786705	G	A
				110815376	G	C
43	ARFGAP3	5	ADP ribosylation factor gtpase activating protein 3	120594125	C	T
				120608455	A	G
				120608663	T	G
				120608723	A	G

44	ABCG2	6	ATP binding cassette subfamily G member 2	37341559	A	G
				37341644	C	G
45	PPARGC1A	6	PPARG coactivator 1 alpha	44889243	G	A
46	IGFBP7	6	Insulin like growth factor binding protein 7	75171520	T	C
47	NAAA	6	N-acylethanolamine acid amidase	94106003	C	T
48	CENPE	6	Centromere protein E	23279534	T	G
				23279542	T	C
				23279590	T	C
				23279637	G	A
				23279688	A	G
				23335199	G	A
				23335244	G	C
				23335298	C	T
49	TBC1D1	6	TBC1 domain family member 1	59765007	T	C
				59765025	T	C
				59765103	G	T
50	UGDH	6	UDP-glucose 6-dehydrogenase	61063338	T	C
				61063367	T	C
				61063397	T	A
				61063412	G	A
				61063475	C	T
51	PKD2	6	Polycystin 2, transient receptor potential cation channel	37440111	C	G

52	PARM1	6	Prostate androgen-regulated mucin-like protein 1	93131062	T	C
				93136609	G	C
				93136636	C	T
				93136673	A	T
				93163651	A	C
				93163800	C	A
53	EGF	6	Epidermal growth factor	16849365	T	C
54	LAP3	6	Leucine aminopeptidase 3	37980915	C	T
				37980947	C	G
				37980968	T	C
				37980977	G	A
				37980998	G	A
				37981008	G	A
				37981022	T	C
55	FAM13A	6	Family with sequence similarity 13 member A	36816604	A	G
				36816677	T	G
				36816706	T	G
56	GRIA1	7	Glutamate ionotropic receptor AMPA type subunit 1	64456033	G	T
				64456471	C	G
				64456472	G	A
				64510313	T	C
				64585962	C	G
				64585964	A	G
				64585966	G	A
				64586749	G	A
				64586908	A	G

				64586973	A	G
				64599967	G	C
				64600034	C	T
				64600058	G	C
				64600072	G	A
				64638191	T	C
				64638192	G	A
				64667402	T	C
				64667450	G	A
				64687729	G	A
				64699965	A	G
				64699974	G	T
				64706234	T	C
				64730309	A	G
				64730326	T	C
				64730353	G	T
				64730369	T	C
				64730406	G	A
				64730516	C	T
				64730521	G	A
				64730565	C	A
				64730587	C	A
				64730594	C	G
				64730629	C	A
				64730631	C	A
				64766212	C	A
				64766213	T	G

				64766214	T	G
				64766221	A	C
				64766314	C	A
57	CAST	7	Calpastatin	97316657	C	T
				97367239	C	T
				97367265	A	G
				97367330	C	T
58	LARP1	7	La ribonucleoprotein domain family member 1	65651091	G	T
				65651191	C	T
59	ELL2	7	Elongation factor for RNA polymerase II 2	96403225	A	C
60	SPTLC1	8	Serine palmitoyltransferase long chain base subunit 1	90100295	C	T
				90100354	C	G
				90100405	C	T
61	NFIB	8	Nuclear factor I B	31673531	G	T
				31673656	G	A
				31698178	A	G
				31705469	T	G
				31711290	G	A
				31733893	T	C
				31829082	G	C
				31829094	A	G
				31870232	T	A
				31870387	C	A
62	GNA14	8	G protein subunit alpha 14	55907644	C	G
				55967439	G	A
				55967440	A	G

				55990833	A	G
63	FBP1	8	Fructose-bisphosphatase 1	85419243	C	T
				85419301	A	G
				85419311	T	C
64	FBP2	8	Fructose-bisphosphatase 2	85356826	C	A
				85356877	C	T
65	NANS	8	N-acetylneuraminate synthase	66003964	C	T
				66003980	T	C
66	B4GALT1	8	Beta-1,4-galactosyltransferase 1	78997582	T	G
67	UBE3D	9	Ubiquitin protein ligase E3D	23488777	C	G
				23488865	G	A
				23488884	T	G
				23495602	T	C
				23495757	C	G
				23488777	C	G
68	TPD52L1	9	TPD52 like 1	27038183	A	G
				27038195	A	G
				27056911	G	C
				27056933	A	T
				27056945	T	C
				27056965	T	C
69	SPTLC2	10	Serine palmitoyltransferase long chain base subunit 2	91271065	G	A
				91271232	T	C
				91271250	A	C
				91271260	T	G
				91271351	G	A

70	RAB11A	10	RAB11A, member RAS oncogene family	12585475	A	G
71	RORA	10	RAR related orphan receptor A	50037457	A	C
73	PCK2	10	Phosphoenolpyruvate carboxykinase 2, mitochondrial	21321275	A	G
74	GLT6D1	11	Glycosyltransferase 6 domain containing 1	107273273	T	G
75	LPIN1	11	Lipin 1	88813274	A	G
				88822559	C	T
				88822673	G	A
				88822678	G	A
				88822743	A	G
				88822913	C	T
				88867020	T	C
				88867124	T	C
				88888200	T	C
				88888218	A	G
				88888415	T	G
				88888444	T	C
				88888523	G	A
				88888743	T	G
76	NRXN1	11	Neurexin 1	33606307	T	C
				33606322	G	A
				33618998	T	A
				33619015	A	G
				33619130	G	C
				33735092	C	G
				33735094	C	T

				33825277	C	T
				33825418	T	C
				33825447	G	A
				33909224	T	C
				34157999	G	A
				34158062	A	G
				34199854	T	C
				34199911	C	T
				34199914	C	T
77	MAP4K4	11	Mitogen-activated protein kinase kinase kinase kinase 4	6642830	G	A
78	PRKCE	11	Protein kinase C epsilon	29216891	C	A
				29216984	A	C
				29216998	C	T
				29216999	T	G
				29334141	A	G
				29334169	A	G
				29334174	T	C
				29334176	C	T
				29334179	G	T
				29334242	A	G
				29461103	G	C
				29461991	G	A
				29462031	A	G
				29485975	G	A
				29509796	T	C
				29509848	T	G

				29509859	C	T
				29510035	C	T
				29510056	C	T
				29525954	A	G
				29620356	C	T
				29620419	T	C
				29699008	C	T
				29702944	G	A
79	MCFD2	11	Multiple coagulation factor deficiency 2	30502606	G	C
80	LCP1	12	Lymphocyte cytosolic protein 1	15240072	G	T
				15240125	C	T
				15240138	C	A
				15240220	A	C
				15240259	C	A
				15240269	A	T
				15296484	T	C
				15313635	C	T
				15313640	C	T
				15313643	T	C
				15313651	G	T
				15313680	A	G
				15313698	C	T
				15313711	G	A
				15313794	T	C
				15313808	C	G
				15313841	G	A

81	TNFSF11	12	TNF superfamily member 11	11463522	T	C
				11463581	A	G
				11463697	A	G
				11467236	A	G
				11467351	A	G
				11467352	C	G
				11467381	T	G
82	GJB6	12	Gap junction protein beta 6	36363141	G	A
83	MATN4	13	Matrilin 4	74418993	A	G
84	ACSS1	13	Acyl-coa synthetase short chain family member 1	42783203	G	A
				42783275	G	A
85	OSBPL2	13	Oxysterol binding protein like 2	55677934	A	G
86	YWHAB	13	Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein beta	74096450	A	C
87	KIZ/PLK1S1	13	Kizuna centrosomal protein	40622579	G	A
				40654769	T	C
				40654816	C	T
				40656648	G	A
				40656693	A	G
				40671589	G	A
				40671598	C	T
88	ANGPT1	14	Angiopoietin 1	54853319	T	C
				54853346	C	T
				55004423	G	A
				55004480	A	C
				55004564	C	T

				55004571	A	C
				55077125	C	T
				55077161	G	A
				55105829	T	G
				55106040	C	T
89	COL22A1	14	Collagen type XXII alpha 1 chain	3502324	G	A
				3507352	G	A
				3507385	A	G
				3524527	T	G
				3534501	C	G
				3581977	C	T
				3582040	C	A
				3582086	A	G
				3582088	T	C
				3613770	C	T
				3613849	G	T
				3613901	G	T
				3613938	C	T
				3626111	G	A
				3626167	T	C
				3626168	G	C
				3626282	T	C
				3662581	T	C
				3675487	C	T
				3675488	A	G
				3686209	A	C
				3705812	G	A

				3705826	A	G
90	ARHGAP39	14	Rho gtpase activating protein 39	174404	A	G
91	NIBP/TRAP PC9	14	Trafficking protein particle complex 9	2762649	A	G
				2851646	C	T
92	PTK2	14	Protein tyrosine kinase 2	2132311	G	T
				2132327	T	A
				2171281	A	C
				2173413	G	A
				2173541	G	T
				2173625	T	C
				2173706	G	C
				2173792	T	G
				2257142	G	A
				2257163	C	T
93	TG	14	Thyroglobulin	7660940	T	C
				7660951	A	G
				7661025	G	A
				7669819	C	G
				7672649	G	T
				7687241	C	G
				7687330	T	C
				7687780	C	T
				7687813	T	C
				7688008	C	T
				7688165	G	A
				7688175	T	C
				7732382	C	G

				7732461	C	G
				7733972	G	C
				7778530	A	G
				7778535	C	T
				7838260	A	G
				7838262	C	T
				7838438	C	T
				7838516	G	A
				7838570	C	T
				7838596	T	C
				7838745	G	A
				7838826	T	C
				7838852	T	G
				7838868	C	T
				7838918	A	G
				7838922	A	G
				7843862	C	T
				7843925	T	C
				7844037	G	A
				7881108	A	G
94	FAM83H	14	Family with sequence similarity 83 member H	845746	A	G
95	AGO2	14	Argonaute RISC catalytic component 2	2339981	T	C
				2340063	C	T
				2343697	A	C
				2343761	T	G
				2343764	T	C

				2343799	G	A
				2343803	C	T
96	KHDRBS3	14	KH RNA binding domain containing, signal transduction associated 3	5809834	C	T
97	SPAG1	14	Sperm associated antigen 1	62369336	C	T
				62369449	T	C
98	NDRG1	14	N-myc downstream regulated 1	7543542	C	A
				7543588	T	C
				7543589	A	G
99	TRAM1	14	Translocation associated membrane protein 1	34634613	G	A
				34643587	C	T
100	CA2	14	Carbonic anhydrase 2	75977704	T	G
101	DEPDC7	15	DEP domain containing 7	63408043	A	G
102	PGR	15		6534502	T	C
				6534517	T	A
				6534594	G	A
				6534624	C	T
				6534640	A	G
			Progesterone receptor	6579366	T	C
103	CD44	15	CD44 molecule	65550772	T	C
				65550804	A	G
				65571377	C	A
				65571461	C	T
				65594979	C	T
104	FXYD2	15	FXYD domain containing ion transport regulator 2	27033823	T	C
105	STARD10	15	Star related lipid transfer domain	51984452	C	G

			containing 10	51991866	A	G
106	COMMD9	15	COMM domain containing 9	66661920	T	C
107	DHRS3/SDR 1	16	Dehydrogenase/reductase 3	37847573	C	T
				37873978	T	A
				37874148	A	C
108	PIGR	16	Polymeric immunoglobulin receptor	3735214	T	C
				3735263	C	T
109	COQ8A	16	Coenzyme Q8A	26980174	C	T
				26980190	T	C
				26980214	G	A
110	LAX1	16	Lymphocyte transmembrane adaptor 1	503787	C	T
111	PEX14	16	Peroxisomal biogenesis factor 14	39948435	G	A
				39948529	A	G
				39950963	G	A
				39950969	A	G
				39990067	G	A
				39990120	C	T
				39990201	G	A
112	LAMB3	16	Laminin subunit beta 3	71856616	G	A
113	ACACB	17	Acetyl-coa carboxylase beta	67163745	C	G
				67163769	A	G
				67182104	A	C
				67182200	G	A
				67190909	G	C
				67190975	G	A
				67191165	G	A
				67195319	G	A

				67195508	T	G
				67195547	G	A
				67208946	C	T
				67209014	T	C
				67209059	C	T
				67209060	G	A
				67209764	A	G
				67209769	C	T
				67209771	C	G
114	FBRSL1	17	Fibrosin like 1	47151751	T	G
				47154198	G	A
115	RILPL2	17	Rab interacting lysosomal protein like 2	54848974	G	A
116	IL15	17	Interleukin 15	17180578	C	T
				17180610	C	G
				17182718	C	T
117	EDNRA	17	Endothelin receptor type A	11505989	A	C
				11506019	C	T
				11506020	A	G
118	ARHGAP35 /GRLF1	18	Rho gtpase activating protein 35	53839560	G	A
				53839563	C	T
				53839580	G	A
				53839591	G	A
119	CARD15/NOD2	18	Nucleotide binding oligomerization domain containing 2	18145975	C	A
				18145978	C	T
				18155560	G	C
				18162437	C	T
				18162438	G	A

				18162457	T	G
120	SLC7A5	18	Solute carrier family 6 member 2	12572523	T	C
				12572578	T	C
				12587075	C	A
				12590250	G	A
121	SLC6A2	18	Solute carrier family 6 member 2	23500254	C	T
				23500258	G	A
				23507415	T	C
				23507517	G	A
				23514432	G	A
				23514437	G	A
				23514443	C	T
122	RHPN2	18	Rhophilin Rho gtpase binding protein 2	42911004	A	G
				42941120	G	C
123	HIF3A	18	Hypoxia inducible factor 3 subunit alpha	53374679	G	A
124	RAB5C	19	RAB5C, member RAS oncogene family	43595108	T	C
				43613507	T	C
125	BAIAP2	19	BAI1 associated protein 2	53096057	T	C
				53108584	G	A
				53108593	T	C
				53111646	G	C
				53111681	A	G
				53111751	G	A
				53111790	C	G
126	LLGL2	19	LLGL scribble cell polarity complex component 2	57713457	A	T
				57730759	C	T

127	GHR	20	Growth hormone receptor	33928184	G	A
				33928198	A	G
				33928297	C	T
				33928303	A	G
				33928333	A	G
				33934384	C	T
				33934425	T	C
				34019161	C	T
				34019187	T	C
				34134848	T	A
				34140190	T	A
				34149632	G	A
				34164679	G	A
				34189415	T	G
128	RAB3C	20	RAB3C, member RAS oncogene family	21965895	G	A
				22061826	A	G
				22061830	T	G
				22061838	T	G
				22061895	C	T
				22101868	G	C
				22101893	C	T
				22134151	C	G
				22134198	C	T
				22134253	T	C
				22250203	T	C
				22270316	A	G
129	PRLR	20	Prolactin receptor	41417700	G	A

130	MAP3K1	20	Mitogen-activated protein kinase kinase kinase 1	23965088	C	G
131	NIPBL	20	NIPBL cohesin loading factor	39454723	C	T
				39580256	C	G
132	RICTOR	20	RPTOR independent companion of MTOR complex 2	37664290	G	A
				37664405	T	C
133	LIFR	20	LIF receptor alpha	38185461	A	G
				38185465	G	A
				38185498	A	G
134	OSMR	20	Oncostatin M receptor	37758990	C	A
				37759093	A	G
135	DAB2	20	DAB adaptor protein 2	37286857	T	C
				37289722	A	C
				37289731	C	T
				37289739	G	A
136	DAP	20	Death associated protein	66235660	G	A
				66235713	A	C
				66235738	G	T
				66235761	C	T,G
				66263685	A	G
				66268895	C	T
				66276056	G	A
				66276062	C	T
137	EDC3	21	Enhancer of mrna decapping 3	34097147	G	A
				34097161	T	C
				34097220	C	T
				34097274	A	G

138	SERPINA1/ PI	21	Serpin family A member 1	59304946	T	C
139	PDIA3	21	Protein disulfide isomerase family A member 3	55841645	G	A
				55841803	G	C
140	SCAMP2	21	Secretory carrier membrane protein 2	33889872	G	A
141	ISG20	21	Interferon stimulated exonuclease gene 20	19247175	T	C
				19247327	G	A
				19247362	T	C
142	CACNA1D	22	Calcium voltage-gated channel subunit alpha1 D	48276675	A	C
				48285942	C	G
				48287082	A	G
				48348709	T	G
				48377433	T	C
				48377434	G	A
				48399900	G	A
143	ATP2B2	22	Atpase plasma membrane Ca2+ transporting 2	55805125	G	A
				55807606	T	C
				55832192	A	G
				55832336	T	C
				55832411	C	T
				55835094	C	T
				55843229	T	C
				55843348	T	C
				55894614	T	C
				55894724	C	T
				55895872	C	T

				55897065	T	C
				55902945	C	T
				55906567	G	T
				55907874	G	A
				55908062	A	G
				55908070	C	T
				55908122	G	A
				55908126	G	A
				55908169	A	G
				55908252	C	T
				55908266	T	A,G
				55977508	G	A
				55977596	T	G
144	OSBPL10	22	Oxysterol binding protein like 10	6154447	G	A
				6154529	C	T
				6174247	C	T
				6174369	C	T
				6174444	G	A
				6189637	C	A
				6189643	A	G
				6189784	T	C
				6189789	T	G
				6190283	T	G
				6190284	A	T
				6190329	G	A
				6211218	T	C
				6211254	T	C

				6211341	A	G
				6219892	A	G
				6219926	G	A
				6220004	G	A
				6220022	A	G
				6220032	G	C
				6221151	G	A
				6221219	G	A
				6222013	C	G
				6222027	A	C
				6222065	C	T
				6235839	G	A
				6235880	G	A
				6235892	A	G
145	RAB7A	22	RAB7A, member RAS oncogene	61046292	T	G
				61046400	G	C
				61046424	T	C
				61046426	C	T
				61062670	C	G
				61062712	T	C
				61071421	C	T
				61071494	C	T
146	CDCP1	22	CUB domain containing protein 1	55526015	G	A
				55562243	C	T
				55562256	T	C
				55562266	C	G
				55569788	C	T

147	CMTM6	22	CKLF like MARVEL transmembrane domain containing 8	6956891	C	T
				6956916	C	G
				6956996	G	T
148	ECI2	22	Enoyl-coa delta isomerase 2	50987662	T	A
				50987668	G	A
149	JARID2	23	Jumonji and AT-rich interaction domain containing 2	42063325	G	T
				42063467	A	C
				42084802	A	G
				42084807	G	C
				42084857	C	G
				42084884	T	C
				42084891	T	C
				42084938	A	C
				42130928	G	A
				42130984	C	T
				42131043	C	T
				42156524	T	C
				42213942	G	A
150	ELOVL5	23	ELOVL fatty acid elongase 5	26008832	C	T
				26009074	T	A
151	DSC2	24	Desmocollin 2	27110933	T	C
				27119403	G	A
152	FHOD3	24	Formin homology 2 domain containing 3	21447685	A	G
				21459126	C	T
				21529767	C	T
				21558194	G	A
				21573104	G	A

				21573112	C	T
				21581781	G	C
				21581782	A	T
				21581796	G	A
				21583680	T	C
				21583738	C	T
				21583818	A	G
				21583819	C	T
				21618968	G	C
				21619013	G	A
				21619078	A	G
				21619081	A	G
				21619155	G	A
				21619202	A	G
				21619285	G	A
				21643623	T	A
				21643633	T	A
				21643635	T	C
				21692857	T	C
				21692940	G	A
				21692948	G	A
				21734481	C	T
				21778063	T	C
				21778073	A	C
				21778085	C	A
				21778093	G	A
				21778098	G	A

				21829730	G	A
				21872281	C	T
				21872348	G	A
				21923554	C	T
				21927383	G	A
				21927397	A	C
				21927399	C	G
				21927437	G	C
				21927510	T	C
				21927652	C	T
				21927655	A	G
				21941581	T	C
				21941713	G	A
				21941718	C	A
				21941761	G	A
				21941790	G	C
				21941817	G	A
				21941825	C	T
153	DTX2	25	Deltex E3 ubiquitin ligase 2	36603766	C	T
154	PMM2	25	Phosphomannomutase 2	8555543	G	A
				8555591	G	A
155	CLEC16A	25	C-type lectin domain containing 16A	10719312	G	C
				10730924	G	A
				10739286	A	G
				10842907	G	A
				10845091	A	C
156	EEF2K	25	Eukaryotic elongation factor 2 kinase	21135019	C	G

157	GPAM	26	Glycerol-3-phosphate acyltransferase, mitochondrial	33209741	C	T
158	DMBT1	26		43116429	A	G
				43116525	T	C
				43167127	G	A
				43167154	A	G
				43167195	A	G
				43167212	G	A
				43167265	G	A
				43167268	C	T
			Deleted in malignant brain tumors 1	43167282	G	C
159	BTRC	26	Beta-transducin repeat containing E3 ubiquitin protein ligase	22633058	G	T
160	PRKG1	26	Protein kinase cgmp-dependent 1	7229666	T	A
				7229679	A	G
				7249309	G	T
				7249343	C	T
				7249414	C	T
				7333491	G	T
				7333498	C	G
				7333557	G	A
				7333559	A	T
				7333566	C	T
				7420395	C	T
				7420453	C	A
				7420502	G	C
				7420513	T	C

				7420546	G	T
				7420598	T	C
				7436424	A	G
				7436534	G	A
				7523456	G	A
				7523510	C	A
				7598271	T	A
				7667288	C	T
				7667332	T	C
				7667333	G	A
				7676215	C	A
				7697548	C	T
				7697567	T	C
				7784058	A	C
				7824384	C	T
				7824397	T	A
				7824413	A	G
				7824480	G	A
				7858660	T	G
				7915475	G	A
				8033738	G	C
				8049201	G	A
				8111393	T	C
				8111443	G	A
				8111513	C	T
				8218634	C	T
				8254562	C	T

				8254604	A	G
				8254605	A	G
				8254718	A	G
				8254794	A	C
				8254801	G	A
				8254813	T	C
				8260116	G	A
				8260129	C	T
				8260155	C	T
				8260221	C	T
				8260226	G	A
				8280747	A	G
				8280829	G	C
				8280889	A	G
				8484302	A	G
				8484318	T	C
				8653973	T	C
				8666661	T	C
				8666746	T	C
				8666759	G	C
				8666790	C	T
				8666800	T	C
				8674124	G	A
				8674138	C	A

161	SUFU	26	SUFU negative regulator of hedgehog signaling	23705651	C	T
				23705685	G	A
				23705769	C	T
				23705770	C	G
				23747730	C	G
162	NEURL1	26	Neuralized E3 ubiquitin protein ligase 1	24741443	A	G
163	ALDH18A1	26	Aldehyde dehydrogenase 18 family member A1	17586257	A	G
				17586274	G	C
				17598824	C	T
				17598882	G	A
				17598954	C	T
				17602647	A	G
				17602707	G	A
				17602739	G	A
				17602750	C	G
				17602778	C	G
				17607437	A	G
				17607456	T	C
				17607480	A	G
				17607485	A	G
164	LIPA	26	Lipase A, lysosomal acid type	11563087	C	T
				11586056	G	C
165	CTBP2	26	C-terminal binding protein 2	45520428	T	C
				45567223	A	G
				45567226	A	G
				45567244	C	T
166	GPAT4	27	Glycerol-3-phosphate acyltransferase 4	38939989	T	C
167	ACSL1	27	Acyl-coa synthetase long chain family	16519377	C	T

			member 1	16519474	A	C
168	GINS4	27	GINS complex subunit 4	38873845	G	A
169	MFHAS1	27	Malignant fibrous histiocytoma amplified sequence 1	26868432	C	A
				26868520	C	T
				26912549	T	C
				26941607	C	T
				26950340	A	G
				26950344	C	T
				26950349	A	C
				26950357	C	T
				26950513	C	T
				26962894	A	G
				26963012	C	T
				26963098	A	G
				26963151	T	C
170	KCNK1	28	Potassium two pore domain channel subfamily K member 1	4817023	C	T
				4817129	G	T
				4817132	G	T
				4817142	G	A
				4817144	T	C
				4817177	C	T
				4820229	T	C
				4820345	A	G
				4820427	C	T
171	FADS1	29	Fatty acid desaturase 1	41938678	C	T
				41938694	C	T
				41966748	C	T

				41966854	C	T
172	THRSP	29	Thyroid hormone responsive	18949067	A	T
				18949077	C	T
173	FADS2	29	Fatty acid desaturase 2	42003574	G	A
				42003741	G	T
				42028878	G	A
174	CTSC	29	Cathepsin C	7509979	T	A
				7510282	G	A
				7510354	T	C
				7510485	C	A
				7510552	C	T
				7510556	T	G
				7539590	A	G
				7539603	C	T
				7539684	T	C
				7557784	T	C
175	SPTBN2	29	Spectrin beta, non-erythrocytic 2	46764984	C	T
				46765043	G	T

CHAPTER –5

Summary and Conclusions

SUMMARY AND CONCLUSION

The present study was undertaken to identify the genome-wide variations in Gir cattle using ddRAD approach. The blood samples collected from Gir cattle was further used for DNA isolation using the phenol-chloroform method and was checked for its quality and quantity. The DNA samples were custom sequenced using the ddRAD methodology in Illumina Hiseq 2000. The raw reads obtained were checked for its quality and the poor quality reads were removed. The processed reads were then further used for the identification of the variants and were annotated. Moreover, annotation of the SNPs in the candidate genes for milk production was also carried out.

- A total of 13.1 million raw reads were obtained by Illumina Hiseq 2000 sequencer in fastq format and the initial quality control of raw reads were done by FASTQC. Adaptor trimming and quality control of raw reads were done by the software Prinseq and Stacks.
- A total of 12.8 good quality reads were generated after the initial quality control procedure.
- The alignment was carried out using bowtie2 and the average number of aligned reads was 99.76%, 90.36% and 98.29% for *Bos taurus*, *Bos indicus*, and Gir reference genomes respectively.
- A total of 6.38%, 6.36%, and 6.38% coverage was obtained from the processed reads with the reference genome of *Bos taurus, Bos indicus*, and Gir respectively.
- Approximately 21.31%, 26.91% and 21.36% of the high-quality reads were uniquely mapped to the *Bos taurus, Bos indicus*, and Gir reference genome respectively.
- Variant calling was done by samtools and a total of 193764, 179205 and 160951 SNPs; 12082, 11059 and 9802 INDELs at RD2, RD5 and RD10 respectively were identified in the present study as compared to *Bos taurus* reference genome. Similarly, when compared to the *Bos indicus* reference

genome, a total 117467, 109943 and 99517 SNPs; 12082, 11059 and 9802 INDELs at RD2, RD5, and RD10 respectively were identified. Against Gir reference genome we identified a total of 174492, 162470 and 146564 SNPs; 10887, 10050 and 9010 INDELs at RD2, RD5, and RD10 respectively.

- ❖ The Ts/Tv ratio is found to be 2.62.
- ❖ Annotation of the SNPs using SnpEff revealed that the highest number of SNPs were present in the transcript region (30.99%), followed by intron region (30.58%) and intergenic region (25.90%).
- ❖ Out of 400 collected genes from the available literature and the cgQTL database (http://cowry.agri.huji.ac.il/QTLMAP/qtlmap.htm), a total of 984 SNPs was annotated in Gir cattle among the 175 genes affecting milk production and composition traits. In summary, ddRAD sequencing method provides a cost-effective and robust way to identify SNPs associated with the candidate genes responsible for milk production and its composition traits.

CONCLUSION:

- ❖ The genome-wide SNPs identified by the ddRAD approach in Gir cattle using *Bos taurus*, *Bos indicus*, and Gir as reference genomes.
- ❖ SNPs identified were annotated structurally and functionally across the *Bos indicus* genome.
- ❖ The SNPs present in the candidate genes responsible for milk production and composition traits were annotated.

From results obtained in the present study, may conclude that the high-quality SNPs identified using *Bos indicus* genome as a reference, are highly specific to the *Bos indicus* compare to that of SNPs identified using *Bos taurus* reference genome.

RECOMMENDATION:

The SNPs identified in this study may serve as a useful tool in future studies of quantitative trait loci and linkage mapping, as well as genome-wide association studies in cattle.

Bibliography

BIBLIOGRAPHY

Angulo, J., Mahecha, L., Nuernberg, K., Nuernberg, G., Dannenberger, D., Olivera, M., Boutinaud, M., Leroux, C., Albrecht, E. and Bernard, L. (2012). Effects of polyunsaturated fatty acids from plant oils and algae on milk fat yield and composition are associated with mammary lipogenic and SREBF1 gene expression. *Animal. 6*(12): 1961-1972.

Anton, I., Kovács, K., Holló, G., Farkas, V., Szabó, F., Egerszegi, I., Rátky, J., Zsolnai, A. and Brüssow, K.P. (2012). Effect of DGAT1, leptin and TG gene polymorphisms on some milk production traits in different dairy cattle breeds in Hungary. *Archives Animal Breeding*. *55*(4): 307-314.

Ba, H., Jia, B., Wang, G., Yang, Y., Kedem, G. and Li, C. (2017). Genome-wide SNP discovery and analysis of genetic diversity in farmed sika deer (Cervus nippon) in northeast China using double-digest restriction site-associated DNA sequencing. *G3: Genes, Genomes, Genetics*. *7*(9): 3169-3176.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., and Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS one*. *3*(10): 3376.

Belew, A.K., Tesfaye, K., Belay, G. and Assefa, G. (2016). The State of Conservation of Animal Genetic Resources in Developing Countries: A Review. *International Journal of Pharma Medicine and Biological Sciences*. *5*(1): 58.

Beuzen, N.D., Stear, M.J. and Chang, K.C. (2000). Molecular markers and their use in animal breeding. *The Veterinary Journal*. *160*(1): 42-52.

Bionaz, M. and Loor, J.J. (2008). Gene networks driving bovine milk fat synthesis during the lactation cycle. *BMC Genomics*. *9*(1): 366.

Biscarini, F., Nicolazzi, E.L., Stella, A., Boettcher, P.J. and Gandini, G. (2015). Challenges and opportunities in genetic improvement of local livestock breeds. *Frontiers in Genetics*. *6*: 33.

Bovine HapMap Consortium. (2009). Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science*. *324*(5926): 528-532.

Brouard, J.S., Boyle, B., Ibeagha-Awemu, E.M. and Bissonnette, N. (2017). Low-depth genotyping-by-sequencing (GBS) in a bovine population: strategies to maximize the selection of high quality genotypes and the accuracy of imputation. *BMC Genetics. 18*(1): 32.

Canavez, F.C., Luche, D.D., Stothard, P., Leite, K.R., Sousa-Canavez, J.M., Plastow, G., Meidanis, J., Souza, M.A., Feijao, P., Moore, S.S. and Camara-Lopes, L.H. (2012). Genome sequence and assembly of Bos indicus. *Journal of Heredity*. *103*(3): 342-348.

Catchen, J.M., Amores, A., Hohenlohe, P., Cresko, W. and Postlethwait, J.H. (2011). Stacks: building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics*. *1*(3): 171-182.

Chen, Z., Yao, Y., Ma, P., Wang, Q. and Pan, Y. (2018). Haplotype-based genome-wide association study identifies loci and candidate genes for milk yield in Holsteins. *PloS One*. *13*(2): 0192695.

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): 101127.

Cochran, S.D., Cole, J.B., Null, D.J. and Hansen, P.J. (2013). Discovery of single nucleotide polymorphisms in candidate genes associated with fertility and production traits in Holstein cattle. *BMC Genetics*. *14*(1): 49.

Cohen-Zinder, M., Seroussi, E., Larkin, D.M., Loor, J.J., Everts-Van Der Wind, A., Lee, J.H., Drackley, J.K., Band, M.R., Hernandez, A.G., Shani, M. and Lewin, H.A. (2005). Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. *Genome Research*. *15*(7): 936-944.

Da, Y., Jiang, J., Prakapenka, D., Ma, L., Vanraden, P. and Cole, J.B. (2019). A large-scale genome-wide association study in US Holstein cattle. *Frontiers in Genetics*. *10*: 412.

D'Alessandro, A., Zolla, L. and Scaloni, A. (2011). The bovine milk proteome: cherishing, nourishing and fostering molecular complexity. An interactomics and functional overview. *Molecular BioSystems*. *7*(3): 579-597.

Das, A., Panitz, F., Gregersen, V.R., Bendixen, C. and Holm, L.E. (2015). Deep sequencing of Danish Holstein dairy cattle for variant detection and insight into potential loss-of-function variants in protein coding genes. *BMC Genomics*. *16*(1): 1043.

Davey, J.W. and Blaxter, M.L. (2010). RADSeq: next-generation population genetics. *Briefings in Functional Genomics*. *9*(5-6): 416-423.

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): 62137.

de Groot, N., van Kuik-Romeijn, P., Lee, S.H. and de Boer, H.A. (2001). Over-expression of the murine pIgR gene in the mammary gland of transgenic mice influences the milk composition and reduces its nutritional value. *Transgenic Research*. *10*(4): 285-291.

DongTshan, Y.A.N.G., ChenTguang, D.U., Fei, G.A.O. and Shorgan, B.O.U. (2006). In vitro culture of bovine fetal fibroblast cells using tissue explant attachment and gene transfection through electroporation. *Zoological Research*. *27*(1): 41-47.

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 Biology*. *10*(8): R82.

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): 19379.

Elsik, C.G., Tellam, R.L. and Worley, K.C. (2009). The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science*. *324*(5926): 522-528.

Fang, M., Fu, W., Jiang, D., Zhang, Q., Sun, D., Ding, X. and Liu, J. (2014). A multiple-SNP approach for genome-wide association study of milk production traits in Chinese Holstein cattle. *PLoS One*. *9*(8): 99544.

Fischer, S.G. and Lerman, L.S. (1983). DNA fragments differing by single base-pair substitutions are separated in denaturing gradient gels: correspondence with melting theory. *Proceedings of the National Academy of Sciences*. *80*(6): 1579-1583.

Ganguly, B., Ambwani, T.K. and Rastogi, S.K. (2017). Electronic Northern Analysis of Genes and Modeling of Gene Networks Underlying Bovine Milk Fat Production. *Genetics Research International*, *2017*.

Gaur, G.K., Kaushik, S.N. and Garg, R.C. (2003). The Gir cattle breed of India characteristics and present status. *AnimalGenetic Resources/Resources Génétiques Animales/Recursos Genéticos Animales*. *33*: 21-29.

Geetha, E., Chakravarty, A.K. and Kumar, K.V. (2006). Genetic persistency of first lactation milk yield estimated using random regression model for Indian

Murrah buffaloes. *Asian-Australasian Journal of Animal Sciences*. *19*(12): 1696-1701.

Gill, P. (2001). An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes. *International journal of legal medicine*, *114*(4-5): 204-210.

Grisart, B., Coppieters, W., Farnir, F., Karim, L., Ford, C., Berzi, P., Cambisano, N., Mni, M., Reid, S., Simon, P. and Spelman, R. (2002). Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. *Genome Research*. *12*(2): 222-231.

Grodzicker, T., Williams, J., Sharp, P. and Sambrook, J. (1974), January. Physical mapping of temperature-sensitive mutations of adenoviruses. In *Cold Spring Harbor Symposia on Quantitative Biology*.39: 439-446.

Gurgul, A., Miksza-Cybulska, A., Szmatoła, T., Jasielczuk, I., Piestrzyńska-Kajtoch, A., Fornal, A., Semik-Gurgul, E. and Bugno-Poniewierska, M. (2019). Genotyping-by-sequencing performance in selected livestock species. *Genomics*. *111*(2): 186-195.

Han, B., Yuan, Y., Liang, R., Li, Y., Liu, L. and Sun, D. (2019). Genetic Effects of LPIN1 Polymorphisms on Milk Production Traits in Dairy Cattle. *Genes*. *10*(4): 265.

Hayes, B.J., Chamberlain, A.J., McPartlan, H., Macleod, I., Sethuraman, L. and Goddard, M.E. (2007). Accuracy of marker-assisted selection with single markers and marker haplotypes in cattle. *Genetics Research*. *89*(4): 215-220.

Heaton, M.P., Keen, J.E., Clawson, M.L., Harhay, G.P., Bauer, N., Shultz, C., Green, B.T., Durso, L., Chitko-McKown, C.G. and Laegreid, W.W. (2005). Use of bovine single nucleotide polymorphism markers to verify sample

tracking in beef processing. *Journal of the American Veterinary Medical Association*. *226*(8): 1311-1314.

Hillreiner, M., Schmautz, C., Ballweg, I., Korenkova, V., Pfaffl, M.W. and Kliem, H. (2017). Gene expression profiling in pbMEC–in search of molecular biomarkers to predict immunoglobulin production in bovine milk. *BMC Veterinary Research*. *13*(1): 369.

Hilton Marshall Briggs, Dinus M. Briggs (1980). *Modern Breeds of Livestock*. London; New York: Macmillan. Also cited in: Breeds of Livestock - Brahman Cattle. Department of Animal and Food Sciences, Oklahoma State University.

Hirsch, C.D., Evans, J., Buell, C.R. and Hirsch, C.N. (2014). Reduced representation approaches to interrogate genome diversity in large repetitive plant genomes. *Briefings in Functional Genomics*. *13*(4): 257-267.

Hoeschele, I. and Meinert, T.R. (1990). Association of genetic defects with yield and type traits: the weaver locus effect on yield. *Journal of Dairy Science*. *73*(9): 2503-2515.

Hsia, A.P., Wen, T.J., Chen, H.D., Liu, Z., Yandeau-Nelson, M.D., Wei, Y., Guo, L. and Schnable, P.S. (2005). Temperature gradient capillary electrophoresis (TGCE)–a tool for the high-throughput discovery and mapping of SNPs and IDPs. *Theoretical and Applied Genetics*. *111*(2): 218-225.

Huynh, H., Wei, W. and Wan, Y., (2017). mTOR inhibition subdues milk disorder caused by maternal VLDLR loss. *Cell Reports*, *19*(10): 2014-2025.

Ibeagha-Awemu, E.M., Li, R., Ammah, A.A., Dudemaine, P.L., Bissonnette, N., Benchaar, C. and Zhao, X. (2016). Transcriptome adaptation of the bovine mammary gland to diets rich in unsaturated fatty acids shows

greater impact of linseed oil over safflower oil on gene expression and metabolic pathways. *BMC Genomics*. *17*(1): 104.

Iso-Touru, T., Sahana, G.,Guldbrandtsen, B., Lund, M.S. and Vilkki, J. (2016). Genome-wide association analysis of milk yield traits in Nordic Red Cattle using imputed whole genome sequence variants. *BMC Genetics*, *17*(1): 55.

Iung, L.H.S., Petrini, J., Ramírez-Díaz, J., Salvian, M., Rovadoscki, G.A., Pilonetto, F., Dauria, B.D., Machado, P.F., Coutinho, L.L., Wiggans, G.R. and Mourão, G.B. (2019). Genome-wide association study for milk production traits in a Brazilian Holstein population. *Journal of Dairy Science*.

Izumi, H., Shimizu, T. and Sekine, K., Morinaga Milk Industry Co Ltd. (2016). *Method for Screening for Diet Providing Production of Milk having Immunoregulatory Action*. U.S. Patent Application 15/211,338.

Jiang, L., Liu, X., Yang, J., Wang, H., Jiang, J., Liu, L., He, S., Ding, X., Liu, J. and Zhang, Q. (2014). Targeted resequencing of GWAS loci reveals novel genetic variants for milk production traits. *BMC Genomics*. *15*(1): 1105.

Jiang, Z.H. and Gibson, J.P. (1999). Genetic polymorphisms in the leptin gene and their association with fatness in four pig breeds. *Mammalian Genome*. *10*(2): 191-193.

Kawahara-Miki, R., Tsuda, K., Shiwa, Y., Arai-Kichise, Y., Matsumoto, T., Kanesaki, Y., Oda, S.I., Ebihara, S., Yajima, S., Yoshikawa, H. and Kono, T.(2011). Whole-genome resequencing shows numerous genes with nonsynonymous SNPs in the Japanese native cattle Kuchinoshima-Ushi. *BMC Genomics*. *12*(1): 103.

Khatib, H., Leonard, S.D., Schutzkus, V., Luo, W. and Chang, Y.M. (2006). Association of the OLR1 gene with milk composition in Holstein dairy cattle. *Journal of Dairy Science*. *89*(5): 1753-1760.

Koh, M.C., Lim, C.H., Chua, S.B., Chew, S.T. and Phang, S.T.W. (1998). Random amplified polymorphic DNA (RAPD) fingerprints for identification of red meat animal species. *Meat Science*. *48*(3-4): 275-285.

Kolbehdari, D., Wang, Z., Grant, J.R., Murdoch, B., Prasad, A., Xiu, Z., Marques, E., Stothard, P. and Moore, S.S. (2009). A whole genome scan to map QTL for milk production traits and somatic cell score in Canadian Holstein bulls. *Journal of Animal Breeding and Genetics*. *126*(3): 216-227.

Konieczny, A. and Ausubel, F.M. (1993). A procedure for mapping Arabidopsis mutations using co-dominant ecotype-specific PCR-based markers. *The Plant Journal*. *4*(2): 403-410.

Kraus, R.H., Kerstens, H.H., Van Hooft, P., Crooijmans, R.P., Van Der Poel, J.J., Elmberg, J., Vignal, A., Huang, Y., Li, N., Prins, H.H. and Groenen, M.A. (2011). Genome wide SNP discovery, analysis and evaluation in mallard (Anas platyrhynchos). *BMC Genomics*. *12*(1): 150.

Kwon, J.M. and Goate, A.M. (2000). The candidate gene approach. *Alcohol Research and Health*. *24*(3):164-168.

Lander, E.S. (1996). The new genomics: global views of biology. *Science*. *274*(5287): 536-539.

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*. *10*(3): R25.

Langdon, W.B. (2013). Which is faster: Bowtie2GP> Bowtie> Bowtie2> BWA. *arXiv preprint arXiv:1301.5187*.

Le, S.Q. and Durbin, R. (2011). SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Research*. *21*(6):952-960.

Legarra, A. and Ugarte, E. (2001). Genetic parameters of milk traits in Latxa dairy sheep. *Animal Science*. *73*(3): 407-412.

Lemay, D.G., Lynn, D.J., Martin, W.F., Neville, M.C., Casey, T.M., Rincon, G., Kriventseva, E.V., Barris, W.C., Hinrichs, A.S., Molenaar, A.J. and Pollard, K.S. (2009). The bovine lactation genome: insights into the evolution of mammalian milk. *Genome Biology*. *10*(4): R43.

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): 592-598.

Li, C., Wang, M., Cai, W., Liu, S., Zhou, C., Yin, H., Sun, D. and Zhang, S. (2019). Genetic Analyses Confirm SNPs in HSPA8 and ERBB2 are Associated with Milk Protein Concentration in Chinese Holstein Cattle. *Genes*, *10*(2): 104.

Li, Y., Han, B., Liu, L., Zhao, F., Liang, W., Jiang, J., Yang, Y., Ma, Z. and Sun, D. (2019). Genetic association of DDIT 3, RPL 23A, SESN 2 and NR 4A1 genes with milk yield and composition in dairy cattle. *Animal Genetics*.

Lindblad-Toh, K., Winchester, E., Daly, M.J., Wang, D.G., Hirschhorn, J.N., Laviolette, J.P., Ardlie, K., Reich, D.E., Robinson, E., Sklar, P. and Shah, N. (2000). Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nature Genetics*. *24*(4): 381.

Litt, M. and Luty, J.A. (1989). A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *American Journal of Human Genetics*. *44*(3):397.

Livak, K.J. (1999). Allelic discrimination using fluorogenic probes and the 5′ nuclease assay. *Genetic analysis: Biomolecular Engineering*. *14*(5-6): 143-149.

Malik, A.A., Sharma, R., Ahlawat, S., Deb, R., Negi, M.S. and Tripathi, S.B. (2018). Analysis of genetic relatedness among Indian cattle (Bos indicus) using genotyping-by-sequencing markers. *Animal Genetics*. *49*(3): 242-245.

Markovtsova, L., Marjoram, P. and Tavaré, S. (2000). The age of a unique event polymorphism. *Genetics*. *156*(1): 401-409.

Matukumalli, L.K., Lawley, C.T., Schnabel, R.D., Taylor, J.F., Allan, M.F., Heaton, M.P., O'Connell, J., Moore, S.S., Smith, T.P., Sonstegard, T.S. and Van Tassell, C.P. (2009). Development and characterization of a high density SNP genotyping assay for cattle. *PloS One*. *4*(4): 5350.

McCallum, C.M., Comai, L., Greene, E.A. and Henikoff, S. (2000). Targeted screening for induced mutations. *Nature Biotechnology*. *18*(4): 455.

Meyer, M. and Kircher, M., (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*. *2010*(6): 5448.

Miah, G., Rafii, M., Ismail, M., Puteh, A., Rahim, H., Islam, K. and Latif, M. (2013). A review of microsatellite markers and their applications in rice breeding programs to improve blast disease resistance. *International Journal of Molecular Sciences*. *14*(11): 22499-22528.

Mitra, A., Yadav, B.R., Ganai, N.A. and Balakrishnan, C.R. (1999). Molecular markers and their applications in livestock improvement. *Current Science*. *77*(8): 1045-1053.

Mokhber, M., Moradi-Shahrbabak, M., Sadeghi, M., Moradi-Shahrbabak, H., Stella, A., Nicolzzi, E., Rahmaninia, J. and Williams, J.L. (2018). A genome-wide scan for signatures of selection in Azeri and Khuzestani buffalo breeds. *BMC Genomics*. *19*(1): 449.

Nanaei, H.A., Qanatqestani, M.D. and Esmailizadeh, A. (2019). Whole-genome resequencing reveals selection signatures associated with milk production traits in African Kenana dairy zebu cattle. *Genomics*.

Naqvi, A.N. (2007). Application of molecular genetic technologies in livestock production: potentials for developing countries. *Advances in Biological Research*, *1*(3-4): 72-84.

Nayeri, S., Sargolzaei, M., Abo-Ismail, M.K., Miller, S., Schenkel, F., Moore, S.S. and Stothard, P. (2017). Genome-wide association study for lactation persistency, female fertility, longevity, and lifetime profit index traits in Holstein dairy cattle. *Journal of Dairy Science*. *100*(2): 1246-1258.

Ng, P.C. and Kirkness, E.F. (2010). Whole genome sequencing. In *Genetic variation* (pp. 215-226). Humana Press, Totowa, NJ.

Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E. and Bamshad, M. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*. *461*(7261): 272.

Nielsen, R. (2000). Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*. *154*(2): 931-942.

Niu, Q., Bonsergent, C., Rogniaux, H., Guan, G., Malandrin, L. and Moreau, E. (2016). RAP-1a is the main rhoptry-associated-protein-1 (RAP-1) recognized during infection with Babesia sp. BQ1 (Lintan)(B. motasi-like phylogenetic group), a pathogen of sheep in China. *Veterinary Parasitology*, *232*, pp.48-57.

Oefner, P.J. (1995). Comparative DNA sequencing by denaturing high performance liquid chromatography (DHPLC). *Am. J. Hum. Genet. 57*: 226.

Ogorevc, J., Kunej, T., Razpet, A. and Dovc, P. (2009). Database of cattle candidate genes and genetic markers for milk production and mastitis. *Animal Genetics*. *40*(6): 832-851.

Orita, M., Iwahana, H., Kanazawa, H., Hayashi, K. and Sekiya, T. (1989). Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. *Proceedings of the National Academy of Sciences*. *86*(8): 2766-2770.

Pant, S.D., Schenkel, F.S., Verschoor, C.P., You, Q., Kelton, D.F., Moore, S.S. and Karrow, N.A. (2010). A principal component regression based genome

wide analysis approach reveals the presence of a novel QTL on BTA7 for MAP resistance in holstein cattle. *Genomics*. *95*(3): 176-182.

Pareek, C.S., Smoczynski, R. and Tretyn, A. (2011). Sequencing technologies and genome sequencing. *Journal of Applied Genetics*. *52*(4): 413-435.

Parmentier, I., Portetelle, D., Gengler, N., Prandi, A., Bertozzi, C., Vleurick, L., Gilson, R. and Renaville, R. (1999). Candidate gene markers associated with somatotropic axis and milk selection☆. *Domestic Animal Endocrinology*. *17*(2-3): 139-148.

Patel, A.B., Subramanian, R.B., Padh, H., Shah, T.M., Mohapatra, A., Reddy, B., Jakhesara, S.J., Koringa, P.G., Dash, D. and Joshi, C.G. (2017). Identification of single nucleotide polymorphism from Indian Bubalus bubalis through targeted sequence capture. *CURRENT SCIENCE*. *112*(6): 1230.

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): 37135.

Pimentel, E.C.G., Tietze, M., Simianer, H., Reinhardt, F., Tetens, J., Thaller, G., Bauersachs, S., Wolf, E. and Köning, S. (2010), August. Study of Relationships between Production and Fertility traits in dairy cattle using genomic data. In *9th World Congress on Genetic applied to Livestock Production. Leipzig, Germany*.

Porter, V., Alderson, L., Hall, S.J.G. and Sponenberg, D.P. (2016). Mason's world encyclopedia of livestock breeds and breeding. Volume 1 and Volume 2. *Mason's World Encyclopedia of Livestock Breeds and Breeding. Volume 1 and Volume 2.*

Raschia, M.A., Nani, J.P., Maizon, D.O., Beribe, M.J., Amadio, A.F. and Poli, M.A. (2018). Single nucleotide polymorphisms in candidate genes associated

with milk yield in Argentinean Holstein and Holstein x Jersey cows. *Journal of Animal Science and Technology*. *60*(1): 31.

Raven, L.A., Cocks, B.G. and Hayes, B.J. (2014). Multibreed genome wide association can improve precision of mapping causative variants underlying milk production in dairy cattle. *BMC Genomics*. *15*(1): 62.

Reinhardt, T.A. and Lippolis, J.D. (2006). Bovine milk fat globule membrane proteome. *Journal of Dairy Research*. *73*(4): 406-416.

Reis Filho, J.C., Lopes, P.S., Verneque, R.D.S., Torres, R.D.A., Teodoro, R.L. and Carneiro, P.L.S. (2010). Population structure of Brazilian Gyr dairy cattle. *Revista Brasileira de Zootecnia*. *39*(12): 2640-2645.

Richard, G.F., Kerrest, A. and Dujon, B. (2008). Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiol. Mol. Biol. Rev. 72*(4): 686-727.

Rico, J.E., Mathews, A.T., Lovett, J., Haughey, N.J. and McFadden, J.W. (2016). Palmitic acid feeding increases ceramide supply in association with increased milk yield, circulating nonesterified fatty acids, and adipose tissue responsiveness to a glucose challenge. *Journal of Dairy Science*. *99*(11): 8817-8830.

Riley, L.G., Gardiner-Garden, M., Thomson, P.C., Wynn, P.C., Williamson, P., Raadsma, H.W. and Sheehy, P.A. (2010). The influence of extracellular matrix and prolactin on global gene expression profiles of primary bovine mammary epithelial cells in vitro. *Animal Genetics*. *41*(1): 55-63.

Robinson, G.W., Kang, K., Yoo, K.H., Tang, Y., Zhu, B.M., Yamaji, D., Colditz, V., Jang, S.J., Gronostajski, R.M. and Hennighausen, L. (2014). Coregulation of genetic programs by the transcription factors NFIB and STAT5. *Molecular Endocrinology*. *28*(5): 758-767.

Ron, M., Israeli, G., Seroussi, E., Weller, J.I., Gregg, J.P., Shani, M. and Medrano, J.F. (2007). Combining mouse mammary gland gene expression and

comparative mapping for the identification of candidate genes for QTL of milk production traits in cattle. *BMC Genomics*. *8*(1): 183.

Ronaghi, M., Uhlén, M. and Nyrén, P. (1998). A sequencing method based on real-time pyrophosphate. *Science*. *281*(5375): 363-365.

Rossetti, S., Englisch, S., Bresin, E., Pignatti, P.F. and Turco, A.E. (1997). Detection of mutations in human genes by a new rapid method: cleavage fragment length polymorphism analysis (CFLPA). *Molecular and Cellular Probes*. *11*(2): 155-160.

Sambrook, J. and Russell, D.W. (2001). Molecular cloning: a laboratory manual

3rd edition, (Coldspring-Harbour Laboratory Press, UK)

Schaeffer, L.R. (2006). Strategy for applying genome-wide selection in dairy cattle. *Journal of Animal Breeding and Genetics*. *123*(4): 218-223.

Sermyagin, A.A., Gladyr, E.A., Plemyashov, K.V., Kudinov, A.A., Dotsev, A.V., Deniskova, T.E. and Zinovieva, N.A. (2018). Genome-wide association studies for milk production traits in Russian population of Holstein and black-and-white cattle. In *Proceedings of the Scientific-Practical Conference" Research and Development-2016"* (pp. 591-599). Springer, Cham.

Seidel, G.E. (2009). Brief introduction to whole-genome selection in cattle using single nucleotide polymorphisms. *Reproduction, Fertility and Development*. *22*(1): 138-144.

Stafuzza, N.B., Zerlotini, A., Lobo, F.P., Yamagishi, M.E.B., Chud, T.C.S., Caetano, A.R., Munari, D.P., Garrick, D.J., Cole, J.B., Machado, M.A. and Martins, M.F. (2017). 164 Single nucleotide variants and indels identified from whole-genome resequencing of Gyr, Girolando, and Holstein cattle breeds. *Journal of Animal Science*. *95*(suppl_4): 80-81.

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): 559.

Sun, Y., Boyd, K., Xu, W., Ma, J., Jackson, C.W., Fu, A., Shillingford, J.M., Robinson, G.W., Hennighausen, L., Hitzler, J.K. and Ma, Z. (2006). Acute myeloid leukemia-associated Mkl1 (Mrtf-a) is a key regulator of mammary gland function. *Molecular and Cellular Biology*. *26*(15): 5809-5826.

Surya, T., Vineeth, M.R., Sivalingam, J., Tantia, M.S., Dixit, S.P., Niranjan, S.K. and Gupta, I.D. (2018). Genomewide identification and annotation of SNPs in Bubalus bubalis. *Genomics*.

Tesfayonas, Y.G. (2014). Genome wide association study of milk composition traits in Swedish Red cows.

Thomson, R., Pritchard, J.K., Shen, P., Oefner, P.J. and Feldman, M.W. (2000). Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. *Proceedings of the National Academy of Sciences*. *97*(13): 7360-7365.

Tobler, A.R., Short, S., Andersen, M.R., Paner, T.M., Briggs, J.C., Lambert, S.M., Wu, P.P., Wang, Y., Spoonde, A.Y., Koehler, R.T. and Peyret, N. (2005). The SNPlex genotyping system: a flexible and scalable platform for SNP genotyping. *Journal of Biomolecular Techniques: JBT*. *16*(4): 398.

Upadhyay, M.R., Patel, A.B., Subramanian, R.B., Shah, T.M., Jakhesara, S.J., Bhatt, V.D., Koringa, P.G., Rank, D.N. and Joshi, C.G. (2015). Single nucleotide variant detection in Jaffrabadi buffalo (Bubalus bubalis) using high-throughput targeted sequencing. *Frontiers in Life Science*. *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): 1172.

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. *Nature Methods*. *5*(3): 247.

Velmala, R., Vilkki, J., Elo, K. and Mäki-Tanila, A.(1995). Casein haplotypes and their association with milk production traits in the Finnish Ayrshire cattle. *Animal Genetics*. *26*(6): 419-425.

Venturini, G.C., Cardoso, D.F., Baldi, F., Freitas, A.C., Aspilcueta-Borquis, R.R., Santos, D.J.A., Camargo, G.M.F., Stafuzza, N.B., Albuquerque, L.G. and Tonhati, H. (2014). Association between single-nucleotide polymorphisms and milk production traits in buffalo. *Genet Mol Res*. *13*(4): 10256-68.

Viale, E., Tiezzi, F., Maretto, F., De Marchi, M., Penasa, M. and Cassandro, M. (2017). Association of candidate gene polymorphisms with milk technological traits, yield, composition, and somatic cell score in Italian Holstein-Friesian sires. *Journal of Dairy Science*. *100*(9): 7271-7281.

Vignal, A., Milan, D., SanCristobal, M. and Eggen, A., (2002). A review on SNP and other types of molecular markers and their use in animal genetics. *Genetics Selection Evolution*. *34*(3): 275.

Wang, D.G., Fan, J.B., Siao, C.J., Berno, A., Young, P., Sapolsky, R.Ghandour, G., Perkins, N., Winchester, E., Spencer, J. and Kruglyak, L., (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*. *280*(5366): 1077-1082.

Wang, H., Jiang, L., Wang, W., Zhang, S., Yin, Z., Zhang, Q. and Liu, J.F. (2014). Associations between variants of the HAL gene and milk production traits in Chinese Holstein cows. *BMC Genetics*. *15*(1): 125.

Wang, Z., Gerstein, M. and Snyder, M.(2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*. *10*(1): 57.

Wathes, D.C., Cheng, Z., Chowdhury, W., Fenwick, M.A., Fitzpatrick, R., Morris, D.G., Patton, J. and Murphy, J.J. (2009). Negative energy balance alters global gene expression and immune responses in the uterus of postpartum dairy cows. *Physiological Genomics*. *39*(1): 1-13.

Weldenegodguad, M., Popov, R., Pokharel, K., Ammosov, I., Yao, M., Ivanova, Z., and Kantanen, J. (2018). Whole-genome sequencing of three native cattle breeds originating from the northernmost cattle farming regions. *Frontiers in Genetics*,. *9*: 728.

Wu, Z., Wang, B., Chen, X., Wu, J., King, G.J., Xiao, Y. and Liu, K. (2016). Evaluation of linkage disequilibrium pattern and association study on seed oil content in Brassica napus using ddRAD sequencing. *PloS One*. *11*(1): 0146383.

Yadav, P., Singh, D.D., Mukesh, M., Kataria, R.S., Yadav, A., Mohanty, A.K. and Mishra, B.P.(2012). Identification of suitable housekeeping genes for expression analysis in mammary epithelial cells of buffalo (Bubalus bubalis) during lactation cycle. *Livestock Science*, *147*(1-3): 72-76.

Yang, F., Chen, F., Li, L., Yan, L., Badri, T., Lv, C., Yu, D., Chen, J., Xing, C., Li, J. and Wang, G. (2018). GWAS using 2b-RAD sequencing identified three mastitis important SNPs via two-stage association analysis in Chinese Holstein cows. *BioRxiv*, p.434340.

Yodklaew, P., Koonawootrittriron, S., Elzo, M.A., Suwanasopee, T. and Laodim, T. (2017). Agriculture and Natural Resources.

Yuan, Z.R., Li, J., Liu, L., Zhang, L.P., Zhang, L.M., Chen, C., Chen, X.J., Gao, X., Li, J.Y., Chen, J.B. and Gao, H.J. (2011). Single nucleotide polymorphism of CACNA2D1 gene and its association with milk somatic cell score in cattle. *Molecular Biology Reports*. *38*(8): 5179-5183.

Zabeau, M.(1993). Selective restriction fragment amplification: a general method for DNA fingerprinting. European. *Patent Application 924026 29. 7*.

Zhao, Z. and Boerwinkle, E. (2002). Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome. *Genome Research*. *12*(11): 1679-1686.

Zhu, M. and Zhao, S. (2007). Candidate gene identification approach: progress and challenges. *International Journal of Biological Sciences*. *3*(7):420.

Zingg, H.H. and Lefebvre, D.L. (1988). Oxytocin and vasopressin gene expression during gestation and lactation. *Molecular Brain Research*. *4*(1): 1-6.

Zwane, A.A., Schnabel, R.D., Hoff, J., Choudhury, A., Makgahlela, M.L., Maiwashe, A., Marle-Koster, V. and Taylor, J.F. (2019). Genome-wide SNP discovery in indigenous cattle breeds of South Africa. *Frontiers in Genetics*. *10*: 273.