THE UNIVERSITY
of EDINBURGH

The Genomic and Transcriptomic Landscape of the Indian Water Buffalo (*Bubalus bubalis*)

Prasun Dutta
(Student ID: s0928794)

A Thesis submitted for the degree of
Doctor of Philosophy
The University of Edinburgh
2019

Declaration of Authorship

I declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, for any other degree or professional qualification except as specified. Except where it is stated otherwise by reference or acknowledgment, the work presented is entirely my own.

Prasun Dutta

August, 2019

Abstract

The water buffalo (*Bubalus bubalis*) is one of the most important domesticated species in India providing milk, meat, hide and draft power. At over 100 million animals, India has the highest number of water buffalo in the world, however, the species is found across the globe, including Europe where the Mediterranean subspecies is farmed. Despite the importance of this domesticated bovid, there are limited high-resolution genomic and transcriptomic analyses across these animals. The aim of this thesis was to use whole genome and RNA sequencing data to characterise regulatory variation and genome evolution in the water buffalo. Specifically, I explored the presence of regulatory variation in macrophages of water buffalo in the form of allele-specific expression (ASE) and investigated signatures of selection and breed divergence across water buffalo breeds.

Water buffalo are exposed to a range of important pathogens, many of which that are zoonotic in nature. Differences in regulatory variation between animals have been shown to underlie some of the diversity in response to these pathogens. Macrophages are among the first cells of the innate immune system to act against a pathogen through its recognition, phagocytosis and destruction playing an important role in host disease susceptibility. Regulatory variants acting in macrophages are thus important candidates for explaining differences in disease susceptibility to infectious diseases among water buffalo. To detect the presence of regulatory variation, I used whole genome sequencing and RNA-seq data in 4 Mediterranean water buffalo to identify ASE in macrophage expressed genes. The analysis revealed that regulatory variation does exist in macrophage expressed genes which could be reliably detected as ASE signature.

To understand the impact of domestication and how water buffalo have evolved I used whole genome sequencing data from 81 animals spanning seven distinct breeds. I identified the population structure of these breeds and explored how gene flow has shaped their genomes. I also characterised the signatures of putative selection between breeds. Sites identified included genes linked to milk

production, coat colour and body size and interestingly a number of these overlapped those found to be under selection in other domesticated species suggesting some extent of convergent domestication.

In this thesis, I consequently undertook one of the first high-resolution evolutionary and regulatory variation analyses of an important domesticated species, *Bubalus bubalis*. The results from this study are likely to be invaluable to inform future studies of how regulatory variants may confer tolerance to water buffalo pathogens as well as the impact of domestication on its genome.

Lay summary

The water buffalo is one of the most important domesticated animals in India providing milk, meat, hide and draft power. Their dung is used as fertilizer and fuel. India has the highest number of water buffalo in the world (over 100 million) which represents around 57% of the global buffalo population. In India, buffalo are economically important livestock and diseases affecting them directly impact the livelihood of Indian farmers and disturb the economic market of the country. Many of the diseases are zoonotic in nature which means that they affect species beyond the water buffalo, such as humans and livestock that may be in physical proximity.

Many organisms, including the water buffalo, are capable of fighting against disease causing pathogens. There are cells in the body known as 'macrophages' that act as a defence system against a pathogen. Macrophages contain genes which produce proteins that play a central role in this defence. The difference in regulation of protein production in macrophages makes the defence system of some buffalo strong against the pathogen, while some buffalo produce less protein, making their defence system weak. In this thesis, I have studied the possible existence of this regulation in the water buffalo, which will allow us to comment on the differences between various water buffalo defence systems, and help in the production of healthier buffalo in the future.

I have also worked with genomic data from different breeds of water buffalo from India and Europe, and have tried to understand how domestication has brought about differences in their DNA. I have tried to understand if those differences are responsible for the diversity in physical features (such as body size and coat colour, milk production and nutritive content) among buffalo and if there are any advantageous genomic differences that allow some water buffalo breeds to adapt better to their environment and survive among disease causing pathogens, as opposed to others.

Thus, in this thesis, I have tried to evaluate if regulation of certain genes exists in cells which confer tolerance to the water buffalo against pathogens, and how domestication has brought differences in different breeds of water buffalo. The

knowledge produced in this thesis will be important for future work on pin-pointing the reason of the gene regulation, thereby paving a pathway to understand disease tolerance in water buffalo in future research. It will also be useful in understanding how domestication has led to the evolution of its genome which will allow researchers to concentrate on their characteristics of interest such as reproduction, milk production and defence system in the water buffalo.

Acknowledgements

This PhD project would not have been possible without the help and support of many people at various stages of the research.

Firstly, I would like to express my gratitude towards my principal supervisor, Prof David Hume for introducing me to the world of macrophages and his constant inputs and guidance. His constant motivation and mentorship kept me afloat and on my toes through the whole process. I would like to thank my additional supervisors, Dr James Prendergast and Prof Eileen Wall who made sure I meet my deadlines. I would like to especially thank James, who was always there to discuss my queries on population genetics and bioinformatics and was patient even when I barged into his office without prior notification.

My PhD project was made possible with the immense help from my senior colleagues in 'buffalo team'- Dr Rachel Young, Dr Stephen Bush (Steve) and Lucas Lefevre. All the bioinformatics analysis work would not have been possible without the work done by Rachel, Lucas and everyone else who was involved in sample collection, wet-lab experiments and sequencing. Specifically, Rachel and Lucas led the larger buffalo atlas project that provided the data for my PhD. Steve established and ran the bioinformatics pipeline that built the complete buffalo gene expression atlas. An extended thanks to Steve, for guiding me with various bioinformatics techniques/methods during the start of my PhD. I am also thankful to Dr Andrea Talenti, Dr Mazdak Salavati and Anirudh Patir who were not directly involved in my PhD project but were always there for enriching discussions about bioinformatics/statistical methodologies and population genetics concepts.

The research process would have been very difficult without the support of my family and friends. I am grateful to my parents for believing in me when I started this journey. Kudos to my wife Shweta, who dealt with all my crankiness and mood swings during the last phase of my PhD, boosted my confidence about my abilities and made me understand that taking breaks is important. Last but not the least, a big thank you to all my friends in Edinburgh whose company made this journey extremely fruitful and enjoyable.

Table of Contents

List of Abbreviations

AGRI-IS	Animal Genetic Resources of India
AI	Allelic Imbalance
AM	Alveolar Macrophage
APC	Antigen-Presenting Cells
ASE	Allele-Specific Expression
BAM	Binary Alignment Map
BEF	Bovine Ephemeral Fever
BEFV	Bovine Ephemeral Fever Virus
BH	Benjamini-Hochberg
BMDM	Bone Marrow Derived Macrophage
BQ	Base Quality
BVDV	Bovine Viral Diarrhoea Virus
BWA	Burrows-Wheeler Aligner
CDS	Coding Sequence
Cis-eQTL	*Cis*-Expression Quantitative Trait Locus
cM	Centimorgan
CNV	Copy Number Variation
CV	Cross-Validation
DAVID	Database for Annotation, Visualization and Integrated Discovery
DC	Dendritic Cell
DNA	Deoxyribonucleic Acid
DNA-seq	Deoxyribonucleic Acid Sequencing
EHH	Extended Haplotype Homozygosity
eQTL	Expression Quantitative Trait Locus
FAOSTAT	Food and Agriculture Organization Corporate Statistical Database
FDR	False Discovery Rate
FS	FisherStrand
GATK	The Genome Analysis Toolkit
Gb	Gigabase
GFF	Generic Feature Format
GO	Gene Ontology
GOI	Government Of India
GQ	Genotype Quality
GTF	Gene Transfer Format
GVCF	Genomic Variant Call Format
HBV	Hepatitis B Virus
iHS	Integrated Haplotype Score
IMF	Intramuscular Fat

INDELS or Indels	Insertions And Deletions
KEGG	Kyoto Encyclopaedia of Genes and Genomes
LD	Linkage Disequilibrium
LPS	Lipopolysaccharide
LRR	Leucine Rich Repeats
MAE	Monoallelic Expression
MAF	Major Allele Frequency
MAP	Mycobacterium avium subspecies paratuberculosis
MAPQ	Mapping Quality
MBASED	Meta-analysis Based Allele-Specific Expression Detection
Mbp	Mega Base Pair
MDM	Monocyte Derived Macrophage
MDP	Macrophage-DC Progenitor
MEGs	Macrophage-Expressed Genes
MHC	Major Histocompatibility Complex
MPS	Mononuclear Phagocytic System
MQ	Root Mean Square Mapping Quality
mRNA	Messenger Ribonucleic Acid
MT	Mitochondria
NBAGR	National Bureau of Animal Genetic Resources
NCBI	National Center for Biotechnology Information
NDDB	National Dairy Development Board
NET	Neutrophil Extracellular Trap
NGS	Next-Generation Sequencing
NK	Natural Killer
NKR	Natural Killer Cell Activation Receptor
NLR	Nod-like Receptor
NO	Nitric Oxide
PAMP	Pathogen-associated Molecular Pattern
PAR	Pseudoautosomal Regions
PBMC	Peripheral Blood Mononuclear Cell
PC	Principal Component
PCA	Principal Component Analysis
PRRs	Pattern-recognition Receptors
QC	Quality Control
QD	Qualbydepth
RMAE	Random Monoallelic Expression
RNA	Ribonucleic Acid
RNA-seq	Ribonucleic Acid Sequencing
SAM	Sequence Alignment Map

SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
SOR	StrandOddsRatio
TB	Tuberculosis
TF	Transcription Factors
Ti/Tv	Transition vs Transversion
TLR	Toll-like Receptor
TPM	Transcripts Per Million
Trans-eQTL	*Trans*-Expression Quantitative Trait Loci
UTI	Urinary Tract Infection
UTR	Untranslated Region
VCF	Variant Call Format
VQSR	Variant Quality Score Recalibration
WBC	White Blood Cells
WGS	Whole Genome Sequencing
XCI	X Chromosome Inactivation
XP-EHH	Cross-Population Extended Haplotype Homozygosity

List of Figures

List of Tables

Chapter 1. General Introduction

1.1 The Water buffalo

The water buffalo or domestic water buffalo *Bubalus bubalis* (Linnaeus, 1758) belongs to the Bovidae family (antelopes, cattle, gazelles, goats, sheep, and relatives) and bovinae subfamily (bison, African wild buffalo (*Syncerus* genus), Asian wild buffalo (*Bubalus* genus), domestic cattle (*Bos* genus), yaks and relatives). In the Pleistocene period (glacial period that began 2,588,000 years ago and lasted until 11,700 years ago), the genus *Bubalus* was widely distributed in Europe and southern Asia. It became restricted to the Asian continent as the climate became drier. Three distinct wild buffalo types- the anoa from Indonesia (*B. depressicornis*), the tamaraw from Philippines (*B. mindorensis*) and the Indian wild buffalo (*B. arnee)* emerged in Asia. Only the Indian wild buffalo has been domesticated (Mason, 1974). Most likely, the domestic water buffalo is the descendent of *B. arnee* which is currently an endangered species (Lau, et al., 1998; Mason, 1974; McGowan, et al., 2019).

Based on morphological and behavioural criteria, the domestic water buffalo has two subspecies: the swamp type buffalo (*B. bubalis carabanesis*) and the river type buffalo (*B. bubalis bubalis*) (Macgregor, 1939). They differ in chromosome number (swamp type: 2n = 48, river type: 2n = 50) (Harisah, et al., 1989). Despite this difference, the two species can apparently interbreed and generate fertile offspring (Mishra, et al., 2015). The swamp buffalo's natural habitat is the swamp or the marshland whereas the usual habitat of the river buffalo is river valley and they prefer the clean water of rivers (Macgregor, 1939). The swamp type buffalo is found mainly in Southeast Asia and covers the area between the north of China and the state of Assam in the northeast of India. Hence, the area covers countries such as India (northeast), China, Vietnam, Thailand, Indonesia, Philippines and Malaysia. In contrast, the river type buffalo is mainly found in rest of India, and Pakistan. However, they are distributed in southwest Asia (Turkey, Iraq, Iran, Egypt, countries of the Arabian Peninsula, etc.) and southeast Europe (Albania, Bulgaria, Greece, Italy, etc.) (Mason, 1974). The

swamp type of buffalo is primarily used as a work animal in rice-growing countries and for draught purposes whereas the river type buffalo have been mainly selected for developing high performance milk producing breeds (Mason, 1974). This thesis focuses on the different breeds of the river water buffalo and the Mediterranean water buffalo breed of Europe.

Domestication of the water buffalo occurred very early, but the exact time and place is unknown due to lack of clear archaeological evidence. Tamed and hence domesticated representations of buffalo have been found on seals from the third millennium B.C. in Mohenjo-Daro in the Indus valley (now Pakistan) and Mesopotamia (Ur) in Iraq. They have also been mentioned to be present in China from second millennium B.C. (Mason, 1974). Cockrill proposed that the buffalo were domesticated at around 2500 B.C. in various riverine civilisations such as the ones present near the rivers Euphrates and Tigris, the Indus and the Yangtze (Cockrill, 1981). The river water buffalo was probably domesticated approx. 6,300 years ago in the Indian subcontinent (Kumar, et al., 2007; Nagarajan, et al., 2015) whilst the swamp water buffalo originated in the border of south China and north Indochina (Zhang, et al., 2016). Both types have undergone independent domestication events (Yindee, et al., 2010).

1.2 Uses of the water buffalo

The domestic water buffalo is considered to be the animal of the small farmer in countries where it is present in high numbers (Cockrill, 1974). From an economic point of view, water buffalo play a crucial role in the agricultural economy of many countries across the world. A larger proportion of the world population depends on domestic water buffalo than any other livestock species (FAO & UNEP, 2000). The water buffalo is a very good draught animal. It is used to plough and till land, pull carts, raise water from the wells and do other activities requiring draught power. Paddy field cultivation is mainly done using water buffalo that possesses more weight and strength than cattle. It is also referred to as a ‘living tractor’. Due to its low maintenance cost and long living life, it is an asset to the low income small farmer (Cockrill, 1981).

The water buffalo is also the second largest source of milk in the world after the cow. While the river type buffalo is mainly used for milk production, the swamp type buffalo is mainly raised for draught purposes (Borghese, 2005). Buffalo milk has twice the level of butter fat compared to cow milk. It is the main source of *ghee* or clarified butter used in many South Asian households, especially in India and Pakistan. The Mediterranean water buffalo milk from Italy is also used for making the famous Mozzarella cheese (Cockrill, 1981).

Water buffalo are used for meat only when they are done with their primary goals of providing milk and draught power (Deb, et al., 2016). Meat is effectively a by-product and buffalo have not commonly been genetically selected for this purpose (Naveena and Kiran, 2014). One exception to this is the breed called 'buffalypso' in Trinidad, which is reared specifically for meat (P. Bennet, et al., 2010). Buffalo meat is known to be lower in fat and cholesterol and hence considered to be healthy (Infascelli, et al., 2003).

Water buffalo skin or hide is an important source to the leather industry in many countries. It is thicker and tougher than cow hide and is used to produce shoes, bags and garments. The horns have an ornamental value and are also used to create spoons, combs, buttons and knife handles. The dung of the water buffalo is used as a fuel and fertiliser. Buffalo hair is used for making brushes due to their strength, flexibility and thickness (Cockrill, 1977).

1.3 Water buffalo in India

Figure 1 shows the world distribution of water buffalo in the world. According to Food and Agriculture Organization Corporate Statistical Database (FAOSTAT) 2017 data, out of the total water buffalo population in the world i.e. approx. 200 million, Asia has the highest number of water buffalo (approx. 195 million). 'India' has the highest number of buffalo amongst all countries where water buffalo are present. Having 113 million water buffalo in the country, India has about 56.5% of water buffalo in the world. The FAOSTAT data is based on live head counts of water buffalo in 2017 and can be seen in Supplementary Table S 1.

Based on the 19th Livestock Census 2012 (released in June 2014), an official all India report published by Department of Animal Husbandry and Dairy, Government of India (GOI), the water buffalo population in India was recorded to be 108.7 million. Water buffalo contribute to around 21% of the total Indian livestock population after cattle, which contribute to the highest (around 37%). Rest of the livestock includes sheep, goat, pigs and others (horses, ponies, mules, donkeys, mithun or gayal, yaks, camels and poultry such as fowl, ducks, chicken, quails and turkey). The 20th Livestock Census is yet to be released. FAOSTAT and GOI data do not distinguish between river or swamp type of water buffalo.

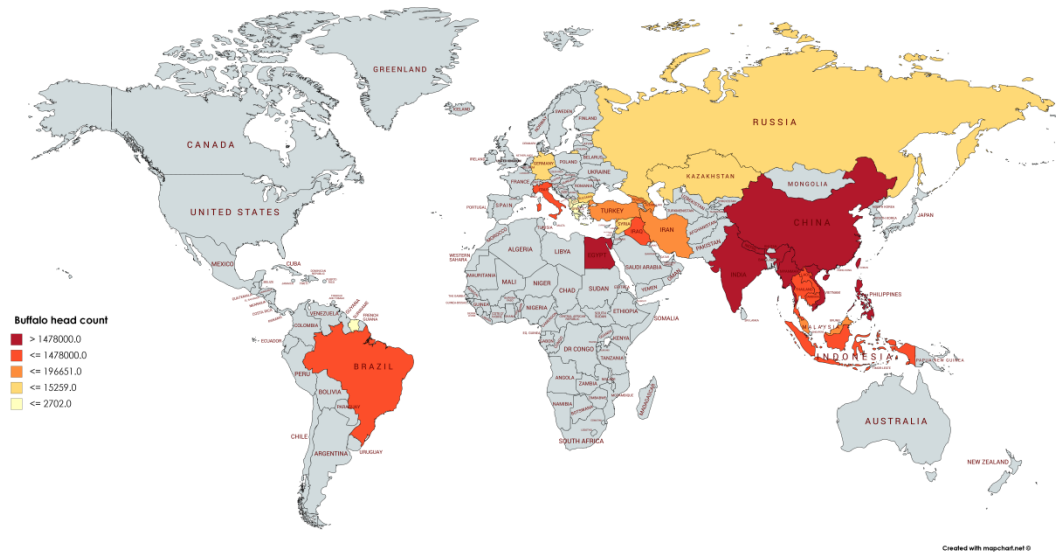

Figure 1: World distribution of buffalo as of 2017. The map has been reproduced from a similar map present in FAOSTAT website (http://www.fao.org/faostat/en/) which is based on the number of live buffalo head count in the year 2017. Information that distinguishes between the types of buffalo (swamp/river) was not provided in FAOSTAT.

The domestic water buffalo plays an essential role in India's agrarian economy (Hoffpauir, 1982). The agriculture sector employs more than 50% of the workforce in India (Sunder, 2018). Buffalo milk contributes about 55% of total milk production in India (Dhillod, et al., 2017) and is economically more valuable to farmers due to its high nutritional content as compared to cow's milk. In India, buffalo are a financial asset to farmers because they can function as insurance against the risk of crop failure due to natural calamities (Dhanda, 2004). The

socio-religious restrictions of India do not allow cow meat consumption. Buffalo are therefore a significant source of meat, cheaper than either chicken or lamb (Deb, et al., 2016; Hoffpauir, 1982). India is one of the largest buffalo meat exporters in the world that directly adds to the Indian economy (Kant, et al., 2018). Since buffalo are economically important livestock in India, diseases affecting them directly impact people's health if they are zoonotic in nature (for example, tuberculosis or brucellosis). Additionally, they may also affect the livelihood of Indian farmers and disturb the economic market of the country.

1.4 The water buffalo breeds of India

As of 2019, there are 16 registered water buffalo breeds in India. This is based on the data provided by the National Bureau of Animal Genetic Resources (NBAGR) which is a GOI nodal agency for registering newly identified livestock and poultry. The registered breeds are: Banni, Bhadawari, Chilika, Jaffarabadi, Kalahandi, Marathwadi, Mehsana, Murrah, Nagpuri, Nili-Ravi, Pandharpuri, Surti, Toda, Bargur, Chhattisgarhi and Luit (Swamp). These breeds are found all over India with specific breeds present in specific states of the country in addition to other 'desi' or non-descript breeds that are present all over India. In this thesis, I have focussed on 6 major Indian riverine buffalo breeds which have been described ahead- Banni, Bhadawari, Jaffarabadi, Murrah, Pandharpuri and Surti. The details of the breeds have been taken from various books and websites such as (Mason, 1974), (Nivsarkar, et al., 2000), https://www.dairyknowledge.in/ (maintained by National Dairy Development Board (NDDB), Government of India) and http://agritech.tnau.ac.in/animal_husbandry/animhus_buffalo%20breeds.html (maintained by Tamil Nadu state government agriculture university), unless otherwise referenced.

1.4.1 Murrah

This is the most popular breed of water buffalo in India which is found in Hisar, Rohtak, Gurgaon and Jind district of Haryana and Delhi. It is also known as 'Delhi', 'Kundi' or 'Kali'. Indian Murrah (Figure 2) is the most diffuse breed in the

world that is present from Bulgaria to South America and all over Asia (Borghese, 2011). The Murrah bull is used to upgrade non-descript and inferior local buffalo and has been used in many other developing countries such as Bulgaria, Philippines, Brazil, Vietnam and Malaysia (Trivedi, 2000). It has jet black skin with hair and may sometimes have white markings on its face or legs. The horns are short and tightly curved in a spiral shape. The breed has a massive body size along with long head and neck. The females have well-developed udders and broad hips. The average weight of a male Murrah buffalo is 567 kg, whereas, for females it is 516 kg. Murrah has high efficiency in producing milk and butter fat in India; an average of 1752 kg milk per lactation and 7.3% butter fat. The males of this breed can be used as draught animals, but are relatively slow and heat-intolerant. Hence, they need to be under shelter during harsh weather conditions (Das, et al., 1999).

Figure 2: Murrah breed of water buffalo

1.4.2 Banni

This Banni breed of buffalo (Figure 3) is found in the Kutchchh/Kutch district of Gujarat, India and is also known as ‘Kutchi’ and ‘Kundi’. The name ‘Banni’ originates from the region where the breed is found. The breed was recognised in 2010. The Banni buffalo is morphologically similar to Murrah, but genetically distinct (Mishra, et al., 2009). The breed is maintained by a local community known as ‘Maldharis’ (traditional livestock keepers) and breeding is done through natural mating only (no artificial insemination). The breed is well-adapted to harsh weather conditions in areas of water scarcity, frequent droughts and high temperatures (Mishra, et al., 2011). Traditionally, the animals are taken away for grazing to the Banni grassland in the night and then brought back to the villages during the day. Physically, the breed has a large wedge shaped body covered with hair. Banni buffalo have vertical coiled horns and can sometimes have double-coiled ones as well. Their teats are conical with round and pointed tips. Their skin is thin and soft and is generally black or copper in colour. Some animals have also been reported to have white coloured patches on their lower legs, tail and forehead. The Banni buffalo are only used for milk. The average yield of the breed is 2,857.2 kg milk per lactation and 6.65% fat. The milk yield can go up to 6,054 kg per lactation.

Figure 3: Banni breed of water buffalo

1.4.3 Bhadawari

The breed which is also known as ‘Etawah’ is found in Bhind and Morena districts of Madhya Pradesh, and Agra and Etawah districts of Uttar Pradesh. The Bhadawari buffalo (Figure 4) are medium sized with greyish-black or copper coloured skin. Their legs also have wheat-straw like colour. They have two white lines called ‘chevrons’ on the lower side of the neck. The average weight of a male and female animal is 475 and 410 kg respectively. The horns of this breed are of average thickness and not very pointed. Bhadawari buffalo are good for draught purposes (including the female buffalo) and have heat tolerance. They are known for their high efficiency in conversion of low quality coarse feed into butter fat, with fat levels as high as 12.8%. This breed has an average milk yield of 1,294 kg per lactation and average fat percentage of 7.8. The Bhadawari breed is frequently cross-bred with Murrah in order to increase fluid milk. This has led to the gradual decrease of the pure Bhadawari breed buffalo (Arora, et al., 2004; Nivsarkar, et al., 2000).

Figure 4: Bhadawari breed of water buffalo

1.4.4 Surti

The Surti breed (Figure 5) is also known by the names ‘Charotari’, ‘Deccani’, ‘Gujarati’, ‘Nadiadi’ and ‘Talabda’ and the breeding tracts for this buffalo are the districts of Vadodara, Bharuch, Kheda and Surat in Gujarat, India. They are medium sized animals, black or brown with silver grey to rusty brown coloured hair. One distinguishing feature of this breed is the presence of white chevrons around the jaw from ear to ear and around the brisket. Some animals also have white marks on the forehead, legs and tail tip. Their heads are broad and long. They possess sickle shaped horns which first go downwards and backwards, and then turn upwards towards the tip to form a hook. They are lighter than other heavy breeds of India. The average weight of male and female Surti buffalo is 435 kg and 401 kg respectively. They produce an average of 1,667 kg of milk per lactation and the average milk fat content is 7.02%.

Figure 5: Surti breed of water buffalo

1.4.5 Jaffarabadi

Jaffarabadi (Figure 6) is one of the heaviest buffalo breeds in India, with the average weight of male and female buffalos at 700 and 620 kg respectively. They are found in Amreli, Bhavnagar, Jamnagar, Junagadh, Porbandar and Rajkot districts of Gujarat state. The breed is named after the town of Jaffarabad (in Amreli district). They have a long body which is black in colour. The breed has been established in and around Gir forest which is a natural habitat for Gir lions. They have a massive head and neck with a prominent forehead. The breed might have been selected for large head and body size to protect against the lions (Kumar, et al., 2006). They possess heavy and broad horns that sometimes cover their eyes. The horns emerge from either side of the head downwards in an inclined fashion, and then move upwards and finally inwards into an incomplete coil. Due to their heavy nature, they provide draught power in the form of ploughing and carting. Like the Banni breed, Jaffarabadi buffalo are also maintained by the ‘Maldharis’ by natural mating (Nivsarkar, et al., 2000).

They are also milk producing animals that yield an average of 2,239 kg of milk per lactation with an average fat percentage of 7.7.

Figure 6: Jaffarabadi breed of water buffalo

1.4.6 Pandharpuri

The Pandharpuri buffalo (Figure 7) are medium sized buffalo from Solapur, Sangli and Kolhapur districts in southeast Maharashtra, India. The breed is also known as 'Dharwari'. The name of the breed comes from the town Pandharpur in Solapur district. The average weight of a male and female Pandharpuri buffalo is 470 (https://www.roysfarm.com) and 416 kg respectively. The body colour varies from light black to dark black. Sometimes, white marks have been observed on forehead, legs and tails. The peculiar characteristic of this breed is the presence of very long horns extending beyond the shoulder blades. They have a long and narrow face along with a long, narrow and straight nasal bone. They are both milk and draught animals. They are fairly good milkers producing an average 1,790 kg of milk per lactation with an average of 8% fat.

Figure 7: Pandharpuri breed of water buffalo

1.5 The Italian Mediterranean river water buffalo and buffalo in the United Kingdom

The Italian buffalo (Figure 8) is a medium sized river animal with a slightly elongated head and thick hair. This breed of water buffalo has brown horns directed laterally and backwards where the base is oval in females and triangular in males. Their coat colour varies from light brown to burnt brown, sometime completely black. They may have white hair in the front and end of the tail (Gonzalez, 2011). Buffalo were first introduced into Italy from central Europe by barbarian invaders (Cockrill, 1981; Salerno, 1974). All buffalo now present in Europe are known as the 'Mediterranean' type (Borghese, 2005). The actual 'Mediterranean' type of buffalo in Europe originates from India and it was introduced into Europe by the Arabs and other central European conquerors in the 6th or 7th century (Borghese, 2005). The Italian buffalo are maintained and bred exclusively for milk from which Mozzarella cheese is derived - the primary product of Italian buffalo livestock (Borghese, 2005). The average milk yield of

the animal was recorded 1.5-3.5 kg for 271 days with 7.87% fat content. Meat production from male Italian buffalo is rapidly increasing (Borghese, 2013). In the past, they were also used as draught animals (Salerno, 1974). Buffalo in the United Kingdom are of the Mediterranean type and they have been reported to be imported from Romania. They have a poor milk yield of about 1.5 kg per year (Borghese, 2011).

Figure 8: Italian Mediterranean river water buffalo

1.6 Diseases affecting water buffalo

Water buffalo are affected by a range of parasitic, fungal, bacterial, viral and endoparasitic pathogens affecting the individual buffalo as well as the herd (Villanueva, et al., 2018). Water buffalo are resilient and appear to be more resistant to common diseases than cattle (Deb, et al., 2016). Bovine pleuropneumonia is uncommon in buffalo and less damaging, (Cockrill, 1981) and foot rot and foot abscess have also been never observed in water buffalo (Cockrill, 1981). Disease resistance is evident in isolated populations, like in Australia and Trinidad, places which in-general have low level of livestock-

related diseases. However, buffalo may become susceptible if a disease from outside is introduced (Cockrill, 1977).

Few bacterial diseases that do impact water buffalo include Leptospirosis, Brucellosis, Tuberculosis (TB) and Johne's disease (paratuberculosis) caused by *Leptospira spp*, *Brucella abortus*, *Mycobacterium bovis* and *Mycobacterium avium* respectively. Leptospirosis is common. For example, 35% of water buffalo in Mexico were reported to be positive for *Leptospira* infection (Romero-Salas, et al., 2017), 48% in Philippines (Villanueva, et al., 2016) and 25% in Thailand (Chadsuthi, et al., 2017). Clinical signs include anaemia, jaundice, meningitis, hemoglobinuria, pyrexia and finally death (Villanueva, et al., 2018). Brucellosis also affects water buffalo (Mathur, 1964). The disease is basically a reproductive disease occurring in cattle with symptoms including late abortions, retained placentas, epididymitis (pain and discomfort in scrotum) and orchitis (inflammation of testes) (Nicoletti, 2001). TB is an important disease that affects both animals and humans worldwide leading to morbidity, mortality and economic loss (Thoen, et al., 2014). In buffalo, TB is caused by *M. bovis*, a Gram-positive intracellular bacteria, and the disease is zoonotic in nature (Grange, 2001). It also affects other ruminants; sheep, goat, bison and farmed deer (de Lisle, et al., 2001). Johne's disease is a chronic infection that leads to diarrhoea, milk production decline, weight loss despite of a good appetite and death. It is spread through contaminated feed, water and milk (Desio, et al., 2013).

Water buffalo are also infected by many viruses. Bovine ephemeral fever (BEF) caused by arthropod-borne bovine ephemeral fever virus (BEFV), Neonatal diarrhoea caused by Rotavirus, Bovine viral diarrhoea caused by Bovine viral diarrhoea virus (BVDV) are some of the diseases and their respective pathogens affecting the water buffalo. Prominent protozoan infections in the water buffalo include *Trypanosoma evansi*, *Neospora caninum* and *Cryptosporidium spp.* There are also reports of fungal infection such as the Deg Nala disease caused by mycotoxin produced by *Fusarium spp*. Endoparasitic diseases such as Fasciolosis, Schistosomiasis also cause immense harm to the buffalo. These diseases cause a variety of clinical signs and symptoms such as

necrosis, decreased milk production, stunted growth, diarrhoea, abortion, weakness and death in extreme cases. There have been few published studies of parasitic gastrointestinal nematodes in water buffalo, but a recent study in Mexico suggest that they share susceptibility to *Strongyloides sp.*, *Cooperia sp.*, *Haemonchus sp.* (10.4%), *Eimeria sp.*, *Moniezia sp.*, *Trichuris sp.* with other domesticated ruminants (Ojeda-Robertos, et al., 2017).

Many of these diseases are zoonotic in nature (brucellosis, TB, schistosomiasis, leptospirosis and rotavirus infection). Pathogenic diseases ultimately lead to the loss of reproduction performance of the water buffalo resulting to economic losses (Villanueva, et al., 2018).

1.7 The immune system

The immune system provides an organism (host) a state of protection (immunity) against foreign pathogens or substances (antigens). Immunity is mediated by cells and molecules, which can recognise, remove or destroy foreign intruders that have the potential to harm the host (Kindt, et al., 2007). Traditionally, the immune system is further classified into two cellular subsystems: 'innate' and 'adaptive' (also called specific/acquired) immune system. The innate immune system has cellular and molecular components that recognise disease-causing pathogens by the specific molecules produced by/on them. For instance, beta-1, 3-glucan is characteristic of certain fungi, peptidoglycan and lipopolysaccharide (LPS) are characteristics of certain bacteria and phosphoglycan is a characteristic of certain parasites (Kimbrell and Beutler, 2001). It generates an inflammatory response stopping the spread of the pathogen and symptoms such as fever, body aches and sickness behaviour (O'Neill, 2005). Various cell types involved in this immune system are neutrophils, macrophages, natural killer cells and dendritic cells. This system acts instantaneously without having any previous introduction of the pathogen. By contrast, adaptive immunity takes weeks to develop and is controlled by B and T cells (lymphocytes). It is capable of recognising and destroying specific pathogens through their structurally unique receptors as compared to the innate

immune response which confers broad protection against pathogens (Kimbrell and Beutler, 2001).

The adaptive immune response is controlled and assisted by the innate immune response and the innate defence system is always ready for instant activation prior to a pathogen attack (Kindt, et al., 2007). Although, the action of lymphocytes is very important in generating an effective immune response, it takes three to five days to amplify pathogen-specific clones which differentiate into effector cells, by which time the pathogen would have caused enough harm to the host organism. Until then, various mechanisms of innate immunity rapidly control the pathogen's replication. This makes the role of the innate immune system extremely important as it plays a fundamental role in host defence (Medzhitov and Janeway, 2000).

Furthermore, in the adaptive immune system, the B and T cell's structurally unique receptors are produced in a somatic nature during B and T cell development. The pathogen-specific receptors are generated randomly and are not passed on to the next generation. In contrast, the innate immunity receptors are present in germ cells in nature which means that receptor specificity is genetically determined which can be passed on to the next generation. The receptors are specific in nature for pathogens and evolve through natural selection (Medzhitov and Janeway, 2000).

1.8 Innate immunity and the cells involved in it

The innate immune system acts as the first-line host defence system as soon as the host is exposed to a pathogen. Innate immunity focuses on highly conserved structures present on surfaces of a broad range of microorganisms, which are called 'pathogen-associated molecular patterns' (PAMP). These are not present on the host cells and are essential for the survival of the pathogen and hence cannot be altered by them (Akira, et al., 2006). The receptors of the innate immune system that recognise the PAMPs are known as 'pattern-recognition receptors' (PRRs) such as 'Toll-like receptors' (TLRs), RIG-I-like receptors (RLRs), NOD-like receptors (NLRs), and DNA receptors (cytosolic sensors for DNA), (Kumar, et al., 2011).

TLRs play a central role in recognising the pathogens and generating an immune response and are mainly expressed by innate immune cells including macrophages, neutrophils and dendritic cells (Cruvinel Wde, et al., 2010). At present, there are 13 TLRs discovered in mammals which are present on the plasma membrane or on endosomal membranes where they interact with components released from internalised pathogens (O'Neill, et al., 2013). TLR 1, 2, 4, 5 and 6 bind to pathogen cell wall and membrane components (such as LPS and lipoteichoic acid from cell walls, lipoproteins (peptidoglycan) of the cell membrane ans flagellin). TLRs 3, 7, 8 and 9 bind to the DNA and RNA of most pathogens (Janeway and Medzhitov, 2002). *TLR10* has no specific ligands and acts as an inhibitory receptor by weakening the *TLR2* response, suppressing the immune response (Oosting, et al., 2014). *TLR11* has been shown to recognise the profilin-like proteins of *Toxoplasma gondii* (Yarovinsky, et al., 2005) and the flagellins of *Escherichia coli* and *Salmonella typhimurium* (Mathur, et al., 2012). In addition to *TLR11*, *TLR12* also recognises the profilin protein of *T. gondii* and regulates interleukin-12 (*IL12*) production by dendritic cells as a response to the pathogen (Raetz, et al., 2013). It has been observed in mice that *TLR13* participates in *group B Streptococcus* recognition (Signorino, et al., 2014).

1.8.1 Dendritic Cells

Dendritic cells (DC) are specialised antigen presenting cells that contribute to the control of functions of B and T cells and hence act as a bridge between innate and adaptive immunity (Banchereau and Steinman, 1998; Steinman and Hemmi, 2006). They reside in tissues like intestines, skin and liver where they are exposed to pathogen-bearing protein antigens. Upon antigen recognition, the cells migrate to regional lymph nodes. As the DCs migrate, they mature to antigen-presenting cells (APC). Those APCs expressing major histocompatibility complex molecules (MHC, containing peptides from processed antigen) and costimulatory molecules are presented to the T cells, leading to an adaptive immune response (Cruvinel Wde, et al., 2010; Kindt, et al., 2007). The precise relationship between DC and macrophages, which can also present antigen to T cells, especially particulate antigens, has been the subject of some debate

(Hume, 2015). It has been suggested that 'antigen presentation' should be considered an inducible function of any cell, rather than forming the basis for definition of a cell type.

1.8.2 Natural killer cells

Natural killer or NK cells are innate immune cells that recognise and lyse cells that are infected by bacteria, viruses or other microbes and tumour cells. They are widely distributed in lymphoid (spleen, thymus, bone marrow, lymph nodes) and non-lymphoid tissues (skin, tonsils, gut, lung, liver). NK cells have multiple germline-encoded receptors on them which are called NK cell activation receptors (NKRs), for example, CD244, Ly49D, Ly49H, NKG2D, NCR1, NCR2, NCR3 and NKG2C. The activation takes place when the receptors interact with pathogen-derived ligands such as PAMPs or foreign ligands present on infected or diseased cells (Abel, et al., 2018). When NK cells are activated by PAMPs, they produce important cytokines such as IFNγ, TNFα and granulocyte-macrophage colony-stimulating factor (GM-CSF) along with a cytokine and chemokine cascade leading to the activation of T cells, macrophages, neutrophils and DCs that helps in the containment of microbial growth (Abel, et al., 2018; Souza-Fonseca-Guimaraes, et al., 2012). NK cells are especially abundant in cattle and have been divided into subsets based on surface markers (Hamilton, et al., 2017).

1.8.3 Granulocytes and Mast cells

Granulocytes are cells of the innate immune system that are of three types: neutrophils, eosinophils and basophils. Neutrophils are the most abundant blood leukocyte population in most species and are rapidly recruited to sites of inflammation. They provide a first-line innate immune defence by phagocytosing, killing, and digesting bacteria and fungi (Segal, 2005). They also kill pathogens extracellularly when an activated neutrophil (for example by LPS) gives rise to structures known as neutrophil extracellular traps (NETs). The NETs are extracellular structures that have a web-like shape which are made up of granule and nuclear proteins (such as neutrophil elastase, calprotectin,

defensins, etc.) and are capable of trapping and killing a variety of pathogens (Papayannopoulos, 2018).

Eosinophils are involved in protection against parasites and allergens. They are activated by various cytokines, immunoglobulins, and complements, which leads to the production of many cytokines (IL2, IL4, IL5, IL10, IL12, IL13, IL16, and IL18), TGF (transforming growth factor) and chemokines (CCL5/RANTES and CCL11/eotaxin-1) having proinflammatory response. Basophils are another type of leukocytes that participate in the host defence and the pathology of acute, chronic and allergic diseases (Siracusa, et al., 2013). Similar to basophils, mast cells also take part in allergic/inflammatory reactions. Once activated, mediators (such as IL4, TNFα, heparin, etc.) are released from them, leading to allergic signs and symptoms such as migration of neutrophils and macrophages, mucus secretion, increased gastrointestinal motility, and bronchoconstriction (Cruvinel Wde, et al., 2010).

1.8.4 Monocytes and the Mononuclear Phagocytic System

Monocytes are part of the Mononuclear Phagocytic System or MPS (which means having a single nucleus). The mononuclear phagocytes include myeloid immune cells such as tissue macrophages, circulating monocytes, promonocytes and their precursor (progenitor) cells in the bone marrow (van Furth, et al., 1972). Based on commonalities of cell expression with macrophages, DCs have also been included in the MPS (Hume, 2008; Hume, 2008). The monocytes circulate in the blood, spleen and bone marrow. However, they do not proliferate in a steady state (Swirski, et al., 2009). The tissue macrophage populations are established by the time of birth and are maintained in the steady state by a combination of self-renewal and replacement by extravasation of blood monocytes. The relative importance of these two mechanisms may vary between tissues and species (Hume, et al., 2019).

In adult animals, blood monocytes are derived from hematopoietic stem cells of the bone marrow, specifically from a common clonogenic progenitor cell called the macrophage-DC progenitor (MDP) (Fogg, et al., 2006; Varol, et al., 2007).

The production of monocytes from the progenitor cells is controlled primarily by macrophage CSF (colony stimulating factor) or *CSF1* (Hume, 2008), which acts through protein tyrosine kinase receptor CSF1 receptor (*CSF-1R*) (Hume and MacDonald, 2012). Like neutrophils, monocytes are recruited to sites of inflammation in response to microbial stimuli and secrete cytokines and antimicrobial factors. They have been reported to respond against many pathogens such as *T. gondii*, *Mycobacterium tuberculosis, Listeria monocytogenes* and *Cryptococcus neoformans* (Serbina, et al., 2008).

1.8.5 Macrophages

Macrophages (from the Greek for 'large eaters') are cells involved in the process of phagocytosis. In vertebrates (including mammals, birds and fish) they are an abundant cell population in all tissues and organ systems, including bone (as osteoclasts), brain (as microglial cells), liver (as Kupffer cells), and connective tissue (as histiocytes), lung (as alveolar macrophages), heart (as cardiac macrophages) and brain (as microglia) (Gordon and Taylor, 2005; Hume, et al., 2019). In addition to their immune function, they are also involved in embryonic development, homeostasis and wound repair (Pollard, 2009). Consequently, macrophages are one of the primary sensors of danger to the host. These danger signals are produced by necrotic cells (cells in stress/attacked by pathogen) and are detected through TLRs, intracellular pattern recognition receptors and interleukin-1-receptor (IL1R) (Mosser and Edwards, 2008).

1.9 Host genetics in disease resistance

Host genes play an important role in conferring natural resistance against pathogens. This means that an organism may have some inherent capacity to be tolerant against disease when exposed to a pathogen without any prior immunization or exposure to the pathogen (Oliver, 1958). The natural defence is controlled by proteins at a molecular level that are encoded by their respective genes. In short, a gene is transcribed to messenger RNA (mRNA) transcripts which are then translated to a protein.

DNA sequence variants of a gene (alleles) may give rise to amino acid changes if they occur within coding regions (exons or splice sites) or to altered levels of the gene product if they impact regulatory regions. If allelic variation of a gene is at a single nucleotide level and the variation is being studied at a population level, it will be called SNP or single nucleotide polymorphism. The term SNP is used to refer specifically to single base variations that have risen to a frequency greater than 1% of the population (Karki, et al., 2015). If the variation alters the structure of the protein, the protein may lose its function. If that protein is responsible for providing defence against a pathogen, the organism that contains the non-functional form of the protein may have a lowered resistance against the pathogen. Fault in a gene that leads to its non-functionality or difference in regulation of a host-pathogen defence mechanism, may change the level of natural resistance of the host against any infectious disease (Adams and Templeton, 1998). It is also possible that a variation can enhance the activity of a protein that may lead to increased resistance of an organism against a pathogen or infectious disease.

Immune-associated genes are subject to strong evolutionary selection as a consequence of the pathogen 'arms race' (Ellegren, 2008). These genes may be involved in functions like phagocytosis, pathogen recognition and immune cell activation. Hence, knowledge about variation amongst those candidate genes, when compared between species or amongst breeds of the same species, is most likely to explain difference in genetic susceptibility to infectious diseases and responses to vaccines as well. Examples of such genes/proteins will be discussed further below.

Previous data and analyses have established the role of host genetics in conferring resistance to various pathogens towards infectious diseases in humans such as HIV (McLaren and Carrington, 2015) and influenza (Arcanjo, et al., 2014). Moreover, studies have reported the association of genetic factors and polymorphisms in different genes for tuberculosis susceptibility in humans (Fernando and Britton, 2006). Cattle have been reported to be naturally resistant to Brucellosis and variation in the *NRAMP1* gene has been associated with difference in susceptibility (Paixão, et al., 2012). Difference in susceptibility

against *Salmonella typhimurium* was observed in chicken (Bumstead and Barrow, 1988). Indigenous cattle breeds of Africa have been reported to be genetically resistant to African trypanosomiasis (Murray, et al., 1984). A recent study revealed that variation in *TICAM1* and *ARHGAP15* involved in important innate immune related pathways were responsible for difference in trypanosomiasis susceptibility (Noyes, et al., 2011).

1.10 Macrophage activation and immune-specific genes involved in it

Previously, macrophage activation has been broadly described via two pathways: classical activation (M1) and alternate activation (M2). Classically activated macrophages respond to microbial pathogens and are associated with IFNγ, proinflammatory cytokines such as TNF, and microbial products such as LPS. Alternatively activated macrophage is stimulated by IL4 or IL13 and has been associated with responses to allergens and parasites. However, this binary classification of macrophage activation was challenged because macrophages can be activated by a spectrum of stimuli (cytokines, chemokines, interleukins, etc.) with unique gene expression changes and the transcriptional profile also changes with time (Hume, 2015).

Within the M1/M2 concept, classical activation centres on IFNγ (originally called macrophage-activating factor), the function of which leads to enhanced antitumor and antimicrobial capacity of the macrophages and the upregulation of antigen processing and presentation pathways (Schroder, et al., 2004). Classical activation involves the induction of TNF transcription by a TLR ligand (such as LPS) which works with IFNγ to activate the macrophage (Mosser and Edwards, 2008). In some cases, the TLR ligand can also activate the Toll/IL1 receptor (TIR) domain containing adaptor protein which in turn induces IFNβ (TRIF)-dependent pathways resulting in IFNβ production. IFNβ can replace IFNγ and also activate the macrophages (Mosser and Edwards, 2008; Yamamoto, et al., 2003). Cytokines such as IL1, IL6 and IL23 are produced by the classically activated macrophages that are helpful in establishing host defence by the development of T helper cells (Mosser and Edwards, 2008). Some other

cytokines secreted by them are CCL15, CCL20, CXCL13, CXCL9, CXCL10, and CXCL11 that coordinate NK and T helper cell recruitment. The microbicidal activity of the macrophages is mediated by phagosome acidification, restriction of iron and other nutrients to the pathogen and release of oxygen free radicals. In rodents, the release of nitric oxide (NO) from L-arginine through NOS2 or the iNOS pathway also contributes to pathogen killing (Bogdan, 2015). Macrophages from most large animals produce little NO, but there is a low level of production in bovid including water buffalo (Young, et al., 2018).

1.11 Polymorphisms associated with receptors for sensing pathogens and other immune related genes in livestock

The most crucial step of any macrophage cell activation is pathogen recognition and pathogen recognition receptors or PRRs such as TLRs and NLRs (Nod-like receptors) play a very important role in such recognition. TLRs are extracellular type I transmembrane glycoproteins playing a very important role in pathogen recognition (Jimenez-Dalmaroni, et al., 2016). As discussed above, 13 mammalian TLRs have been discovered out of which TLR 1-10 are present in humans and TLR 11-13 are present in mice. TLRs can localize on the cell surface (such as TLRs 1, 2, 4, 5, 6, 10) or may have endosomal/intracellular localisation (such as TLRs 3, 7, 8, 9) (Akira, et al., 2006). From a structural point of view, a TLR has an extracellular domain (ectodomain), a single-path transmembrane domain and an intracellular domain (Swirski, et al., 2009). From a sequence point of view, the extracellular N-terminal domain has 16 to 28 Leucine rich repeats (LRR) involved in ligand recognition, and an intracellular C-terminal domain, called the TIR domain, required for interaction and recruitment of adaptor molecules (such as *MyD88* or Myeloid differentiation primary response gene 88) and activating downstream signalling pathways (Medzhitov, 2001). Ectodomain recognition of TLR ligands (such as LPS) initiates dimerisation of the intracellular TIR domain and recruitment of adaptor proteins that drive various signalling pathways producing proinflammatory cytokines and chemokines (Pandey, et al., 2014).

NLRs or NOD-like receptors are intracellular C-type lectin receptors which detect bacterial or viral molecules in the cytoplasm and leads to the secretion of interleukin-1β or IL1β. The family of NLRs contains genes like *NOD1* and *2*, *NOD3*, *NOD9*, *NOD27* NACHT-, LRR- and pyrin-domain containing protein 1 or *NALP1* to *NALP14*, *NLRC4* (NLR family CARD domain-containing protein 4), class II transactivator or *CIITA* and neuronal apoptosis inhibitory protein or *NAIP*. *NOD1* and *2* have been shown to recognise a variety of pathogens such as *Escherichia coli*, *Helicobacter pylori*, *Pseudomonas aeruginosa*, *S. enterica*, *Listeria monocytogenes*, etc. and have been reported to synergise with activated TLR (Diacovich and Gorvel, 2010).

In humans, polymorphisms within TLR genes have been associated with a number of allergic, autoimmune, inflammatory diseases and even cancer (Medvedev, 2013). In livestock, TLR gene polymorphisms may also affect immune-related traits and throw some light on variation in disease resistance (Jann, et al., 2009). A case-control study reported that in a *Mycobacterium bovis* infection in cattle, one SNP within *TLR2* gene was found to be significantly associated with TB resistance (Bhaladhare, et al., 2016). Variation in bovine *TLR2* gene was found to be significantly associated with bovine paratuberculosis infection where the polymorphism was linked to higher macrophage activity (Koets, et al., 2010). A 2014 study on the water buffalo described variants in *TLR2*, *TLR4* and *TLR9* and reported their association with *M. bovis* infection (Alfano, et al., 2014). In another study involving three different breeds of sheep, it was seen that polymorphism within the LRR region of *TLR9* showed significant association with Small Ruminant Lentivirus seropositivity in sheep (*Ovis aries*) (Sarafidou, et al., 2013). Another study reported the association between *TLR4* polymorphism (c.-226G>C SNP) and *Mycobacterium avium subspecies paratuberculosis* (MAP) infection (causing Johne's disease) in Canadian Holsteins cattle, suggesting that G allele supports protection against MAP (Sharma, et al., 2015). Another very recent study in pigs identified genetic variations involving TLR genes impacting the protein sequence of porcine TLR (Clop, et al., 2016). Polymorphisms in *TLR5* and *2* contributed to resistance to salmonellosis within different pig breeds (Shinkai, et al., 2011). In 2012, SNPs in

bovine *TLR1* were reported to be significantly associated with clinical mastitis within a Holstein Friesian herd (Russell, et al., 2012). Another polymorphism in NLR family gene *CARD15* or *NOD2* in Canadian Holstein cattle was proposed to be connected to mastitis resistance (Pant, et al., 2007). *CARD15* gene polymorphism has also been potentially connected to TB susceptibility in Chinese Holstein cow (Wang, et al., 2015).

Several other innate immune genes have also been reported to contribute to disease susceptibility in livestock. *NRAMP1* or *SLC11A1* has been connected to conferring natural resistance against brucellosis in cattle (Paixão, et al., 2012) and Italian water buffalo (Borriello, et al., 2006). Study of polymorphisms in *IFNγ* (along with *TLR4* and *NRAMP1*) in a case-control study revealed their association with Paratuberculosis infection (Johne's disease) in cattle (Pinedo, et al., 2009). SNPs in *IL10RA* have also been connected to the paratuberculosis infection status in dairy cattle (Verschoor, et al., 2010).

The above examples serve as evidence of genetic variation in immune related genes within individuals and amongst breeds, which can cause difference in susceptibility against infectious diseases.

1.12 Regulatory variations and their role in immune-related genes

Regulation of gene expression or the extent to which a gene is transcribed, is affected by various genetic, environmental or epigenetic factors (Stranger and Dermitzakis, 2005). A simple example of regulatory variation involves transcription factor binding sites. Transcription factors (TFs) are proteins that regulate transcription initiation by binding to elements of the promoter of a gene in a sequence-specific manner. They can also act as a repressor where the transcription is switched off. This regulatory variation of genes is important as it allows for certain protein to be expressed in specific tissues or cells that is vital for various cellular functions (Latchman, 1993). Variation in gene expression may also be affected if there is a mutation in the promoter region that may lead to the non-binding of the TF (Buckland, et al., 2004).

Regulatory polymorphisms have been divided into two types: '*cis*' and '*trans*'. The former is present near the locus of a gene, usually in the non-coding part, for example on a promoter site or enhancer site where a TF or an enhancer binds. Alternately, the later refers to any polymorphism that takes place on a locus that is away from the gene. An obvious example is another gene encoding a TF that acts on the target gene. Variation that increases or decreases the activity of the TF would in turn change the expression of the gene it regulates. In this section, we focus on only *cis*-regulation.

Regulatory variations determines an individual's susceptibility to diseases (Knight, 2005). Polymorphisms in the non-coding region of immune-associated genes, for example, the promoter has been associated with disease susceptibility in many published studies in humans. For example, *IL18* gene promoter SNPs have been associated with hepatitis B virus (HBV) infection (Motavaf, et al., 2014). SNPs in *IL10* promoter are associated with severity in leprosy (Moraes, et al., 2004). Two promoter polymorphisms in *CCL2* gene have been associated with disease susceptibility in sepsis (He, et al., 2017). *TLR4* promoter polymorphisms have been reported to influence the innate immunity response against Urinary Tract Infection or UTI (Ragnarsdóttir, et al., 2010). A case-control study revealed the association of polymorphisms in *IL10* promoter with TB susceptibility in Asians and Europeans (Gao, et al., 2015). Regulatory variation arising from promoter polymorphism in the Tumour-necrosis factor-alpha (*TNFα*) gene is associated with malaria in Gambian children (McGuire, et al., 1994). These examples tell us that regulatory variations due to promoter polymorphisms are linked to disease susceptibility.

Regulation of gene expression can be quantified at a population level by looking at the association between the level of expression of a gene and SNPs in its vicinity; so-called expression quantitative trait loci (eQTL) analysis. For example, a population study of human monocyte gene expression revealed heritable variation in expression of the majority of transcripts (Fairfax, et al., 2014). Such regulatory variation can also be quantified by detecting unequal expression of the two alleles in an individual, a phenomenon which is known as allelic imbalance or allele-specific expression (Knight, 2005). In a 2002 study (Yan, et

al., 2002) that involved studying Single nucleotide polymorphisms (SNPs) from 13 genes from 96 individuals, 6 genes showed unequal expression of mRNA transcripts from the two alleles at a heterozygous locus. When the families of 9 individuals who showed allelic imbalance were examined, three families showed allelic imbalance that was consistent with underlying Mendelian inheritance. Based on another study, gene expression variation was found to be less between related individuals than unrelated individuals showing evidence of genetic basis of variation (Cheung, et al., 2003). It can be hypothesised that genes that determine susceptibility to a particular disease, for example, innate immune genes, will also follow this pattern. The presence of regulatory variation in innate immune genes can allow one to prioritise candidate immune genes that can be tested for genetic basis for disease susceptibility at a population level.

1.13 Summary and key points of the thesis

Water buffalo and specifically, the Indian water buffalo are important animals in livestock population. Water buffalo in India are one of the most important livestock in the country and infections lead to economic losses as well as loss of lives in case of zoonotic diseases. Host genetics play an important role in an organism by conferring disease resistance. Variations in innate immune genes are associated with disease susceptibility in humans and livestock. Macrophages are important immune-related cells that are primary sensors of danger in a host organism. TLR and NLR genes are responsible for macrophage activation as they take part in pathogen recognition. Polymorphisms in TLR genes and NLR genes have been associated with disease susceptibility. Macrophage activation leads to the expression of variety of immune-related genes during an infection. Genetic variation in immune related genes within individuals and amongst breeds causes difference in susceptibility against infectious diseases. Regulatory variation in immune related genes determines an individual's susceptibility to diseases which can be studied in the form of unequal expression of the two alleles of a gene. Since regulatory variation is heritable in nature, studying it may help in determining the genetic basis of disease susceptibility. At the outset of this project, there was limited

knowledge of the water buffalo genome and genetic variation amongst individuals and breeds and no published transcriptomic studies that could support analysis of allelic variation.

Chapter 2 describes the examination of the presence of regulatory variation in 4 Italian water buffalo in the form of allele-specific expression (ASE) using RNA-sequencing data. An ASE pipeline was developed *in silico* that was utilised for determining the presence of regulatory variation in immune-related genes.

Chapter 3 describes a genome-wide ASE analysis which involved heterozygous genotypes determined using DNA sequencing data from the same 4 Italian water buffalo and then counting the abundance of RNA-sequencing reads on biallelic heterozygous loci. ASE was observed in innate immune genes and also observed in genes related to cellular homeostasis and genes related to growth and maintenance of cells, which tells us that *cis*-regulation is important in various cellular processes.

Chapter 4 describes the identification of underlying diversity amongst Indian breeds and signatures of positive selection. The results highlight extensive genetic diversity and sights of putative adaptation within Indian water buffalo populations that potentially provide the basis for future genetic selection for improvements in traits such as fertility, productivity and disease resistance.

1.14 Supplementary Material

Domain	Area	Item	Year	Unit	Value	Description
Live Animals	World	Buffalo	2017	Head	200967747	Aggregate, may include official, semi-official, estimated or calculated data
Live Animals	Asia	Buffalo	2017	Head	195772907	Aggregate, may include official, semi-official, estimated or calculated data
Live Animals	Southern Asia	Buffalo	2017	Head	158096534	Aggregate, may include official, semi-official, estimated or calculated data
Live Animals	India	Buffalo	2017	Head	113329671	FAO data based on imputation methodology
Live Animals	Pakistan	Buffalo	2017	Head	37700000	Official data
Live Animals	China	Buffalo	2017	Head	23471754	Aggregate, may include official, semi-official, estimated or calculated data
Live Animals	Eastern Asia	Buffalo	2017	Head	23471754	Aggregate, may include official, semi-official, estimated or calculated data
Live Animals	China, mainland	Buffalo	2017	Head	23469400	FAO estimate
Live Animals	South-Eastern Asia	Buffalo	2017	Head	13602837	Aggregate, may include official, semi-official, estimated or calculated data
Live Animals	Nepal	Buffalo	2017	Head	5177998	Official data
Live Animals	Myanmar	Buffalo	2017	Head	3746870	FAO data based on imputation methodology
Live Animals	Africa	Buffalo	2017	Head	3375752	Aggregate, may include official, semi-official, estimated or calculated data
Live Animals	Egypt	Buffalo	2017	Head	3375727	FAO data based on imputation methodology
Live Animals	Northern Africa	Buffalo	2017	Head	3375727	Aggregate, may include official, semi-official, estimated or calculated data
Live Animals	Philippines	Buffalo	2017	Head	2881894	Official data
Live Animals	Viet Nam	Buffalo	2017	Head	2491662	Official data
Live Animals	Bangladesh	Buffalo	2017	Head	1478000	Official data

Live Animals	Indonesia	Buffalo	2017	Head	1395191	Official data
Live Animals	Americas	Buffalo	2017	Head	1387987	Aggregate, may include official, semi-official, estimated or calculated data
Live Animals	South America	Buffalo	2017	Head	1382130	Aggregate, may include official, semi-official, estimated or calculated data
Live Animals	Brazil	Buffalo	2017	Head	1381395	Official data
Live Animals	Lao People's Democratic Republic	Buffalo	2017	Head	1189000	Official data
Live Animals	Thailand	Buffalo	2017	Head	996307	FAO data based on imputation methodology
Live Animals	Cambodia	Buffalo	2017	Head	655498	FAO data based on imputation methodology
Live Animals	Western Asia	Buffalo	2017	Head	576141	Aggregate, may include official, semi-official, estimated or calculated data
Live Animals	Europe	Buffalo	2017	Head	430836	Aggregate, may include official, semi-official, estimated or calculated data
Live Animals	Southern Europe	Buffalo	2017	Head	404227	Aggregate, may include official, semi-official, estimated or calculated data
Live Animals	Italy	Buffalo	2017	Head	400792	Official data
Live Animals	Sri Lanka	Buffalo	2017	Head	283550	Official data
Live Animals	Iraq	Buffalo	2017	Head	209163	Official data
Live Animals	Azerbaijan	Buffalo	2017	Head	196651	Official data
Live Animals	Turkey	Buffalo	2017	Head	142073	Official data
Live Animals	Iran (Islamic Republic of)	Buffalo	2017	Head	126765	Official data
Live Animals	Timor-Leste	Buffalo	2017	Head	124208	FAO data based on imputation methodology
Live Animals	Malaysia	Buffalo	2017	Head	119264	Official data
Live Animals	Central Asia	Buffalo	2017	Head	25641	Aggregate, may include official, semi-official, estimated or calculated data
Live Animals	Georgia	Buffalo	2017	Head	18358	FAO data based on imputation

						methodology
Live Animals	Eastern Europe	Buffalo	2017	Head	17935	Aggregate, may include official, semi-official, estimated or calculated data
Live Animals	Tajikistan	Buffalo	2017	Head	15259	FAO data based on imputation methodology
Live Animals	Bulgaria	Buffalo	2017	Head	12273	Official data
Live Animals	Kazakhstan	Buffalo	2017	Head	10382	FAO data based on imputation methodology
Live Animals	Syrian Arab Republic	Buffalo	2017	Head	9084	FAO data based on imputation methodology
Live Animals	Germany	Buffalo	2017	Head	8674	Official data
Live Animals	Western Europe	Buffalo	2017	Head	8674	Aggregate, may include official, semi-official, estimated or calculated data
Live Animals	Trinidad and Tobago	Buffalo	2017	Head	5857	FAO data based on imputation methodology
Live Animals	Caribbean	Buffalo	2017	Head	5857	Aggregate, may include official, semi-official, estimated or calculated data
Live Animals	Russian Federation	Buffalo	2017	Head	5662	Official data
Live Animals	Brunei Darussalam	Buffalo	2017	Head	2942	Official data
Live Animals	Greece	Buffalo	2017	Head	2702	FAO data based on imputation methodology
Live Animals	China, Taiwan Province of	Buffalo	2017	Head	2037	Official data
Live Animals	Suriname	Buffalo	2017	Head	735	FAO data based on imputation methodology
Live Animals	Armenia	Buffalo	2017	Head	717	Official data
Live Animals	North Macedonia	Buffalo	2017	Head	646	FAO data based on imputation methodology
Live Animals	Bhutan	Buffalo	2017	Head	550	Official data
Live Animals	China, Hong Kong SAR	Buffalo	2017	Head	317	FAO data based on imputation methodology
Live Animals	Oceania	Buffalo	2017	Head	265	Aggregate, may include official, semi-official, estimated or calculated data

Live Animals	Micronesia	Buffalo	2017	Head	265	Aggregate, may include official, semi-official, estimated or calculated data
Live Animals	Micronesia (Federated States of)	Buffalo	2017	Head	173	FAO data based on imputation methodology
Live Animals	Jordan	Buffalo	2017	Head	95	FAO data based on imputation methodology
Live Animals	Guam	Buffalo	2017	Head	91	FAO data based on imputation methodology
Live Animals	Albania	Buffalo	2017	Head	87	FAO data based on imputation methodology
Live Animals	Mauritius	Buffalo	2017	Head	25	FAO data based on imputation methodology
Live Animals	Eastern Africa	Buffalo	2017	Head	25	Aggregate, may include official, semi-official, estimated or calculated data

Table S 1: FAO data of water buffalo live animals head count according to different areas (world/continent/country) in 2017

Chapter 2. Allele-Specific Expression (ASE) analysis using only RNA-seq data

2.1 Introduction

DNA sequence variation in exons can lead to functional changes in mRNA or a protein encoded by a locus. DNA sequence variants can also alter the level of mRNA produced by a genomic locus (gene expression) leading to discernible changes in phenotype (Albert and Kruglyak, 2015; Wood, et al., 2015). Gene expression is altered by two kinds of genetic variations: *cis*-regulatory, hereon referred to as *cis*-expression quantitative trait loci (*cis*-eQTLs), or *trans*-regulatory (*trans*-eQTLs). *Cis*-eQTLs comprise one or more functional variants that distinguish the two haplotypes at a locus and contribute to differential mRNA output. Functional variants may be single nucleotides or larger gain or loss of a DNA sequence (for example, insertion of mobile genetic elements). These variants may occur in proximal or distal regulatory elements (promoters or enhancers) or could have an effect on the regulation of the mRNA expression (Shi, et al., 2012). *Trans*-eQTLs regulate the transcription of a gene from a distance, for example, they can regulate the transcription of a distal transcription factor which in-turn regulates the expression of the gene (Bryois, et al., 2014; Michaelson, et al., 2009). *Trans*-acting variants may also impact components that transduce environmental signals into gene expression changes (Shi, et al., 2012).

One approach to identify potential regulatory variation in a gene is allele-specific expression (ASE). It is also synonymously known as allelic imbalance (AI). ASE describes expression variation between the two copies of a gene in a diploid organism (Castel, et al., 2015). If an organism is heterozygous for a coding variant, the two alleles or copies of a gene will show unequal levels of expression (Kang, et al., 2016). An extreme case of ASE is monoallelic expression, for example, imprinting, random monoallelic expression and X-inactivation (Metsalu, et al., 2014). An ASE analysis seeks to determine whether

the two alleles of a locus make an equal contribution to the total mRNA output from that locus in heterozygous individuals. We can then seek to link differences in transcript levels produced by each allele to *cis*-acting variation.

RNA-seq as a technique allows the exploration of an organism's transcriptome profile (Wang, et al., 2009). Besides the identification and quantification of transcript isoforms, it enables one to perform differential expression analysis in different conditions (Costa-Silva, et al., 2017). RNA-seq can also capture instances of ASE by digitally quantifying the expression differences of one parental allele over the other (Castel, et al., 2015). Asymmetric expression of the two alleles present in the same cell is potential evidence of *cis*-regulatory differences and a signature of ASE. For detecting ASE, transcribed nucleotide differences between the two gene copies (Reference and Alternate) and sufficient depth of coverage is required (Fontanillas, et al., 2010). In the absence of ASE, the two alleles for a heterozygous genotype at a single nucleotide variant or SNV of any transcript are represented in a ratio of 1:1 of RNA-seq reads. In order to detect ASE, identification of heterozygous SNVs is required, followed by analysis to detect significant divergence from this 50:50 allelic ratio (Harvey, et al., 2014). The accurate quantification of ASE depends on adequate sequencing depth, sufficient heterozygosity and correct alignment (Wood, et al., 2015). Identification of heterozygous sites is ideally done using DNA-seq data, but RNA-seq data can also be used to detect SNVs and the genotypes of loci where a SNV has been detected (Deelen, et al., 2015; Duitama, et al., 2012; Piskol, et al., 2013).

Various tools and platforms have been developed in order to determine ASE. 'AlleleSeq' identifies ASE events by constructing diploid personal genomes (maternal and paternal) using genomic variants from the individual under study and then checks for differences in the number of reads mapped between them (Rozowsky, et al., 2011). 'Allim' requires genomic or RNA-seq information from parents to create a polymorphism aware diploid reference genome or transcriptome respectively. RNA-seq reads from the offspring are then non-ambiguously mapped to the genome or transcriptome to determine ASE (Pandey, et al., 2013). MMSEQ, asSeq and EMASE require phased genotypes

for haplotype construction in order to combine ASE across multiple nearby SNPs/SNVs (Raghupathy, et al., 2018; Sun, 2012; Turro, et al., 2011). In this chapter, I have used MBASED (meta-analysis based allele-specific expression detection) (Mayba, et al., 2014) for quantifying ASE in water buffalo. MBASED was chosen because it can perform gene level ASE analysis using only RNA-seq data and it does not require phased genotype information like other tools/algorithms mentioned above. The MBASED algorithm has been implemented in R (R Core Team, 2018) that measures ASE by combining information across multiple individual heterozygous SNVs within a gene. It can utilise a meta-analytic approach that combines information from multiple studies into a global effect estimate (DerSimonian and Laird, 1986; Mayba, et al., 2014). Although MBASED does not require phase information, it uses a pseudo phasing approach and assigns the 'major' haplotype to the allele with a larger read count at each SNV assuming that ASE is unidirectional along the length of the gene. In this case, 'haplotyping' refers to the grouping of multiple heterozygous SNVs present in a gene into two sets - 'major' and 'minor'. MBASED calculates the extent of deviation from the null hypothesis of equal allele expression (1:1 allelic ratio) for each gene present in a sample. It quantifies the allelic imbalance or ASE effect size in the form of a major allele (haplotype) frequency (MAF) of the gene and also provides a corresponding P-value (known as pValueASE) associated with it (Mayba, et al., 2014).

One major requirement of ASE analysis is getting high confidence heterozygous genotype calls. Single nucleotide polymorphism (SNP) genotyping arrays (Wang, et al., 1998) are a popular means of obtaining genotypes at a particular variant site. In an Illumina based genotyping array, single-stranded fragmented target DNA is hybridised onto arrays that contain unique probe sequences. The probe sequences are designed to bind the target DNA and harbor each of the two alleles of a SNP, say A or B. A SNP can be represented as AA (homozygous) or AB (heterozygous) or BB (homozygous) genotypes. The target DNA that is extracted from source samples is fluorescently labeled, which then binds to the probe sequences and generates a signal in the form of fluorescent

intensity. The fluorescent intensities are then processed, analysed and quality-controlled to infer the SNP genotypes (LaFramboise, 2009; Zhao, et al., 2018).

Another way to attain genotypes is by first sequencing the fragmented target DNA from source samples or individuals using next-generation sequencing methods in a high-throughput manner (Metzker, 2009), followed by the alignment of generated ‘reads’ to a reference genome. After alignment, ‘SNP calling’ or ‘variant calling’ identifies sites that are variable and ‘genotype calling’ infers the genotype of the individuals at those sites (Nielsen, et al., 2011). It should be noted that the usage of ‘SNP’ always refers to a population level, as defined in the last chapter. At an individual level, the word ‘variant’ has been used. Inferring genotypes from sequencing data along with variant calling has been explained in brief further ahead. The word ‘calling’ is synonymous to ‘estimation’.

Whenever a sequencer sequences a nucleotide base from a single strand DNA template, fluorescence intensity is generated and captured by a detector as an image. A base calling algorithm of the sequencer infers the nucleotide from the intensity data and assigns a base quality score as a measure of uncertainty to each base call. In this way, the algorithm incorporates the noise produced during image analysis. The sequencer produces ‘reads’ that consists of the nucleotide bases. The per-base quality score conveys the probability that the base from the read is the actual base that was sequenced (Ewing and Green, 1998). The process of alignment or ‘read mapping’ takes the reads and tries to align them to a reference genome which is crucial to the process of variant calling. Aligners also produce a mapping quality score which is defined as the confidence of the aligner that the read belongs to a particular position in the reference genome (Li, et al., 2008). An important way to increase the alignment accuracy is to use paired-end reads which means that a DNA fragment is sequenced from both forward and reverse ends. The base calls and the quality scores generated during sequencing and alignment are used for genotype and SNP/variant calling. Both base quality and read mapping quality are phred scaled measure of error probability (Ewing and Green, 1998; Ewing, et al., 1998). The larger the phred score, the better the quality of the base call or

mapped read. For example, a phred score of 20 means 99% accuracy or a 1% error rate and a phred score of 30 means 99.9% accuracy or a 0.1% error rate.

As mentioned before, variant calling is a process which aims to determine the positions where bases differ from the reference genome. Genotype calling is the process of determining the genotype of the sample/individual at the location where a variant has been discovered. A simple method for genotype calling is to choose high confidence bases at a particular locus under study (say base quality 20 or 30) and count the number of times an allele has been observed. For example, it was observed that ABI Sanger instrument called a heterozygous genotype when the proportion of the alternate or non-reference allele was between 20% and 80% (Harismendy, et al., 2009). A variant will be called if an individual is homozygous or heterozygous for the alternate allele at the locus (Nielsen, et al., 2011). However, more recently, probabilistic methods are being used while calling genotypes and variants which incorporate uncertainties during sequencing and mapping (phred scaled base and mapping qualities) leading to higher accuracies in genotype and variant calling (Li, 2011).

The effect of regulatory variation in *cis* in the form of ASE has been explored in many cases (Battle, et al., 2014; Larson, et al., 2015; Pickrell, et al., 2010; Zhang, et al., 2018). Whereas eQTL analysis requires access to gene expression profiles of large numbers of individuals to infer linkages between SNVs and the level of mRNA at each locus, ASE can be analysed in individual animals, and population level variation can be inferred from much smaller datasets (Pirinen, et al., 2015). The major constraint is the level of expressed SNVs that distinguish the two alleles at each locus; accordingly ASE is most effective in outbred animals. The motivation of this chapter is to assess the presence of ASE as an indicator of regulatory variation in immune expressed genes in water buffalo using four animals that belong to the Mediterranean breed. The presence of regulatory variation in these genes may determine an individual's susceptibility to diseases.

In this chapter, ASE was determined in 4 types of macrophage samples- bone marrow derived macrophages (BMDMs), BMDMs at 7 hours post LPS or

Lipopolysaccharide stimulation, alveolar macrophages (AMs) and monocyte derived macrophages (MDMs). Sample collection, library preparation and sequencing were performed by methods published in (Young, et al., 2019). As mentioned in the last chapter, macrophages are one of the first lines of host-defense against pathogens and they have a complex transcriptome and many inducible genes. These cells provide the opportunity to check for the presence of regulatory variation in the form of ASE on immune-specific genes. Furthermore, in this chapter, intronic SNVs have also been utilised in ASE calculation along with coding or exonic SNVs. Intronic variants should not be identified using mRNA-seq and any read mapped on the intronic region is potentially a mis-mapping by the aligner (introns are spliced off in mRNA). The intronic variants are generally discarded such as in (Piskol, et al., 2013) and the GATK RNA-seq variant calling pipeline (https://software.broadinstitute.org/gatk/documentation/article.php?id=3891). For this study, the predicted intronic variants were retained, because total RNA-seq provides significant coverage of unspliced nuclear transcripts and SNVs are more prevalent in intronic sequences. In the first instance, an ASE pipeline was developed using only RNA-seq data. The complete pipeline is presented in Supplementary Figure S 2.

2.2 Methods

2.2.1 Raw data description

3 to 4 macrophage samples were sequenced per animal from two male and two female Mediterranean adult buffalo (Paired end-sequencing, using the Illumina HiSeq 2500). As mentioned earlier, these samples were: bone marrow derived macrophages (BMDMs), BMDMs at 7 hours post LPS stimulation, alveolar macrophages (AMs) and monocyte derived macrophages (MDMs). BMDM +/- LPS total RNA was sequenced at a depth of 100 million reads, whereas AM and MDM mRNA was sequenced at a depth of 25 million reads. In total, there were 14 samples. The sample summary has been presented in Table 1. These samples have now been published as part of the gene expression atlas for the water buffalo (Young, et al., 2019).

Animals	Naming in Buffalo expression atlas	Scottish abattoirs where samples were collected from	Samples
Male 1	M2	Grantown	MDM, BMDM +/- LPS
Female 1	F2	Grantown	MDM, BMDM +/- LPS
Male 2	M3	Kirkcaldy	AM, MDM, BMDM +/- LPS
Female 2	F3	Kirkcaldy	AM, MDM, BMDM +/- LPS

Table 1: RNA-seq raw data summary. 14 RNA-seq samples consisting of bone marrow derived macrophages +/- LPS (LPS treated and untreated samples) were sequenced at a depth of 100 million reads. The alveolar and monocyte derived macrophages were sequenced at a depth of 25 million reads.

2.2.2 Read pre-processing and alignment

RNA-seq data processing was done according to the alignment based method described in (Clark, et al., 2017). Briefly, reads from individual samples were screened with FASTQC v0.11.2 (Andrews, 2014) and then cleaned using Trimmomatic v0.36 (Bolger, et al., 2014) with parameters 'TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:100.' Cleaned reads were then aligned to the draft assembly of the water buffalo reference genome (Williams, et al., 2017) present in the NCBI Genome database (NCBI accession GCA_000471725.1) using HISAT2 v2.0.4 (Kim, et al., 2015) with default parameters and -dta (optimise for downstream transcriptome assembly). It should be noted that the reference genome used in this study is not the latest chromosomal assembly currently present in NCBI i.e. assembly UOA_WB_1. Differences between the two assemblies have been mentioned in the next chapter in Table 8.

The reference genome used for alignment was a draft assembly of a female Mediterranean river water buffalo 'UMD_CASPUR_WB_2.0' at 70x coverage and consisting of 366,983 scaffolds. The total number of bases in the genome is 2,836,166,969 (approx. 2.83 Gb) (including the mitochondrion which is 16359 bases long). It was assembled in a collaboration between the University of Maryland (USA), USDA-ARS (USA) and CASPUR (Italy) using MaSuRCA v1.8.3 (Zimin, et al., 2013). As the draft assembly consists of only unplaced scaffolds, an initial analysis of the gene model (Generic feature format or GFF file provided by NCBI) revealed that only 4,338 scaffolds contained genes.

Each of the resulting 14 alignment files (in BAM or Binary Alignment Map format) were filtered with SAMtools v1.4 (Li, et al., 2009), using parameters –F 256 (which removes non-primary alignments) and –F 12 (which removes unmapped reads with unmapped mates). Singleton reads (mapped reads with unmapped mates) were also obtained using parameters –F 4 –f 8. The two subsets were merged using Picard 'MergeSamFiles' (Wysoker, et al., 2013) to create one BAM file of uniquely mapped reads per sample. All the BAM files that belonged to each animal were merged to generate a collective estimate of read counts from all the samples to increase the total depth per animal. Finally, this procedure resulted in 4 BAM files for the 4 water buffalo.

2.2.3 Variant calling and annotation

For each BAM file, variants were called from the RNA-seq data using BCFtools (Gonzalez, 2011) 'mpileup' v1.4 with parameters --max-depth 1000000 --min-MQ 60 (minimum mapping quality (MAPQ) phred score), followed by BCFtools 'call' with parameter -m (allow multiallelic variants) and –v (variant only). 'Mpileup' calculates the genotype likelihood using the coverage of mapped reads on a reference sequence at a single base pair resolution. 'Call' then performs variant calling using this information with the inclusion of a prior (Gonzalez, 2011). BCFtools was selected for calling variants due to its speed in preference to the popular variant calling package - GATK or Genome Analysis toolkit (McKenna, et al., 2010) and for its accuracy which is at par with the GATK variant caller- 'Haplotypecaller' (Poplin, et al., 2018). A very extensive study by (Sahraeian, et al., 2017) demonstrated that when HISAT2 is used as an aligner, both GATK and Samtools (the 'mpileup' function was shifted from Samtools to BCFtools v1.4 onwards) had similar performance. Hence, a more complex approach of GATK can be avoided in RNA-seq variant calling if an accurate aligner is used.

The minimum MAPQ score of 60 was chosen so as to use only uniquely mapped reads for variant calling; HISAT2 assigns a score of 60 to these reads (other aligners use different conventions). Ambiguous mapping reads, arising from the transcription of repetitive genomic regions, are less accurate in

determining allelic imbalance (Chen, et al., 2016). The resulting VCF (variant call format) file (Danecek, et al., 2011) contained both SNVs and indels. Only biallelic heterozygous SNVs were retained using BCFtools 'view' v1.4 with parameters -v snps, -m2 -M2 and -g het.

The resulting SNVs were then annotated using SnpEff v4.3 (Cingolani, et al., 2012) with parameter –onlyProtein (only annotate protein-coding variants). The SnpEff algorithm used a manually added water buffalo reference genome and annotation from NCBI (Williams, et al., 2017) to predict the functional impact of the variants. Each variant may have multiple functional effects. In this scenario, the highest impact was reported first, and for further downstream analyses, each variant was associated with its first annotation only i.e. the gene for which it has the highest effect. When several annotations have the same impact, tie-breaking is done by reporting the canonical transcript annotation before other annotations.

Variants are annotated as intronic/downstream/upstream/untranslated region/intergenic/synonymous/missense/splice region, etc. and the nomenclature is based on the sequence ontology (http://www.sequenceontology.org/). The impacts of each variant is categorised as either as HIGH, MODERATE, LOW or MODIFIER. A HIGH impact variant is anticipated to have a disruptive impact on the translated protein leading to its truncation or loss of function. A MODERATE impact variant is mainly a non-disruptive variant that might change protein effectiveness. A LOW impact variant is likely a harmless variant that does not change a protein's behavior. Finally, a MODIFIER impact variant consists of variants affecting non-coding genes where prediction is difficult or there is no evidence of impact.

2.2.4 Macrophage-expressed genes (MEGs) from water buffalo gene expression atlas

The water buffalo gene expression or transcriptional atlas (Young, et al., 2019) consists of multiple tissues and cell types from different organ systems, including many immune-related tissue and cell populations such as BMDMs +/−

LPS, AM, MDM, or peripheral blood mononuclear cells or PBMCs, spleen, bone marrow cells, etc. Tissues or cell-specific genes will form clusters based on their co-expression patterns. The atlas allowed us to identify MEGs that might be associated with disease resistance traits. MEGs for this ASE analysis were obtained by Dr. Rachel Young based on the water buffalo gene expression atlas, which was primarily developed by Dr. Stephen Bush.

Briefly, the gene expression estimates in the atlas were expressed as transcripts per million (TPM) and averaged across individuals. An .expression file was generated containing the average TPM values of each gene detected across each tissue, and grouped by tissue type or organ system. To visualise the data, the expression file was loaded in Miru (www.kajeka.com/miru), a network graph analysis tool, and a gene-to-gene pairwise Pearson correlation matrix was calculated ($r>0.7$) on gene expression across the samples. A correlation cut off of $r=0.85$ was used to construct a graph containing 15,493 nodes (genes) and 2,846,483 edges (correlations between nodes). Components containing fewer than 5 nodes were removed from the analysis. An expression threshold of TPM >10 was applied to remove noise from the data. The Markov cluster algorithm (van Dongen, 2000) was used with an inflation value of 2.2 to identify clusters of co-expressed genes. Clusters were characterised (including macrophage clusters) by their tissue-specificity or biological process, and in cases where unannotated genes were co-expressed with annotated genes, this information was used to reinforce suggested annotations based on synteny and sequence similarity. Pathway enrichment analysis of clusters was performed using Reactome (http://www.reactome.org/). 527 MEGs were obtained from the macrophage specific clusters by manual inspection. The 527 MEGs are listed in Supplementary Table S 3.

2.2.5 Variant filtration, variant statistics generation and read counting

The master VCF files were filtered to create new VCF files for each water buffalo individual containing only those annotated variants related to the MEGs using SnpSift (Cingolani, et al., 2012). The gene name based filtered VCFs were

further filtered on the basis of variant quality. The BCFtools ‘filter’ program was used with the -i option to include only those variants with the parameters 'QUAL >= 20, MQ = 60, DP >= 2, FORMAT/GQ >= 20 and FORMAT/SP <= 60'. These filtering metrics are intentionally conservative to get high quality biallelic heterozygous SNVs.

The hard-filtering thresholds were chosen based on a small analysis explained ahead wherein the percentage frequency distribution graph of the four parameters - QUAL (phred-scaled quality score for the assertion made in ALT), MQ (RMS Mapping quality), DP (raw read depth), FORMAT/GQ (phred-scaled quality score of the assigned genotype) and FORMAT/SP (phred-scaled strand bias P-value) were plotted whose values were obtained from the VCF files of all the animals.

Figure 9D shows that most of the variants had QUAL values between 0 and 10 i.e. a 10% error rate. Figure 9B shows that, most of the variants had a raw depth of 1. This means that the variant caller was not sure about those variant sites. Accordingly, a minimum raw depth of 2 was applied as a threshold to enable the retention of intronic variants in our analyses. These have relatively low depth as compared to exons or coding sequence (CDS). Since variant calling is probabilistic in nature and probability is multiplied for independent events, the second read increases the probability that a variant exists at that locus (the variant caller also includes sequencing and mapping errors into consideration while calculating the probability). As shown in Figure 9C, most of the sites in all samples had GQ in the range of 120-130. Keeping a GQ threshold of 20 seemed robust (1% error rate). The phred scaled strand bias p-value was added in case any strand bias occurs due to mapping errors and its distribution can be visualised in Figure 9A. The strand bias test checks if the proportion of reference bases on the forward and reverse strand is different from the proportion of alternate bases. If most reference bases are on one strand and most alternate bases are on the same strand, then there is no bias (https://sourceforge.net/p/lofreq/discussion/general/thread/ee151ab0/). 99% of variant sites had very low strand bias, as expected, since the gene expression atlas project employed a stranded RNA-seq protocol (Young, et al., 2019).

BCFtools ‘stats’ v1.6 was used to calculate the transition/transversion (Ti/Tv) ratio (number of transition SNVs divided by number of transversion SNVs) for the master VCF files, which have both SNVs and indels, and also for filtered VCF files consisting of high quality heterozygous biallelic SNVs belonging to MEGs. Ti/Tv is often used as a quality indicator of variation data produced from NGS experiments and a higher Ti/Tv ratio means better quality SNVs. However, the Ti/Tv ratio is genomic region dependent and is different for exons, introns, intergenic, lncRNA and miRNA (Wang, et al., 2015). Variant counts were obtained using the BCFtools ‘view’ tool. It lists the total number of variants discovered per animal (-H <VCF_FILE>), the number of SNVs present in the master VCF file (-H -v snps <VCF_FILE>), the number of indels present in the master VCF file (-H -v indels <VCF_FILE>), number of biallelic heterozygous SNVs in the master VCF file (-H -v snps -m2 -M2 -g het <VCF_FILE>), number of biallelic heterozygous SNVs shared between the MEGs (-H -v snps -m2 -M2 -g het <MEG_VCF_FILE_unfiltered>) and number of biallelic heterozygous SNVs shared between the MEGs which consists of filtered high quality SNVs (-H -v snps -m2 -M2 -g het <MEG_VCF_FILE_filtered>).

Read counts over the reference and alternate alleles were obtained using a python based read counter program called allelecounter v0.6 (https://github.com/secastel/allelecounter) with parameters --min_cov 4, --min_baseq 20 and --min_mapq 60 and --max_depth 10000. The python program works by counting the number of reads present on the locus having the SNV and uses the mpileup file (locus based read coverage data) generated by Samtools ‘mpileup’ program. Samtools counts overlapping reads (due to smaller fragment size in paired end sequencing) only once which takes care of the double counting problem, a common issue in ASE studies (Castel, et al., 2015). In a double counting problem, a read counter may count overlapping mate pairs from a paired end sequence set twice leading to a technical error during ASE quantification.

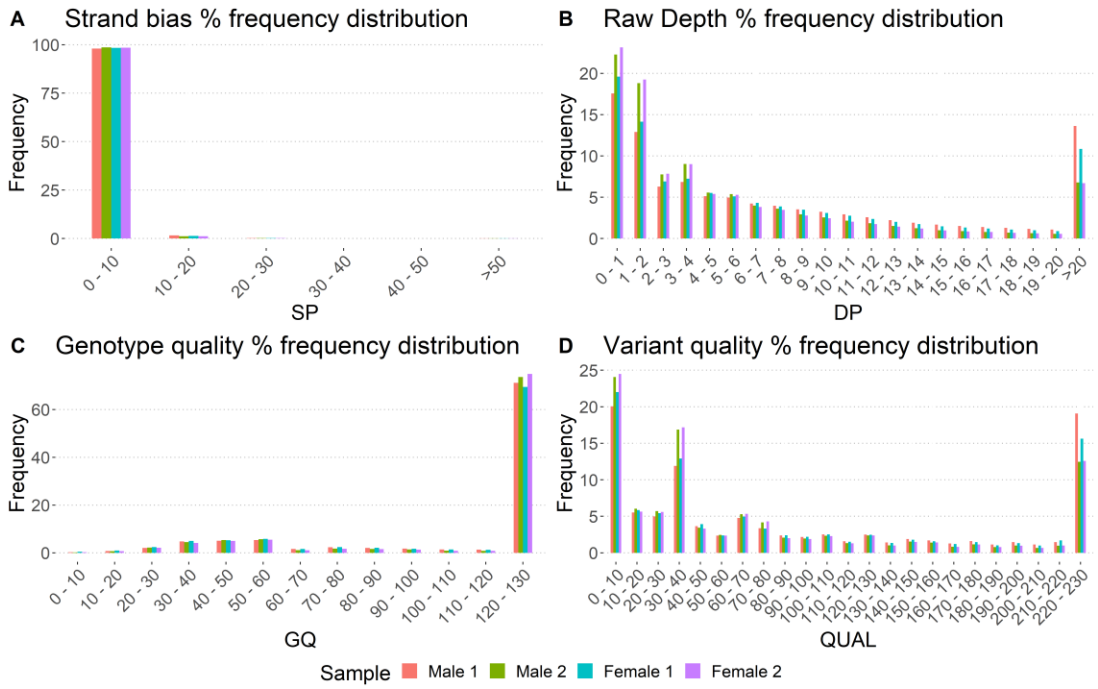

Figure 9: Percentage frequency distribution graphs of various chosen hard-filtering parameters. A- Phred scaled strand bias p-value or SP, B- raw depth at a variant locus or DP, C-Phred scaled Genotype quality or GQ and D- Phred scaled variant quality or QUAL

2.2.6 Quantification of ASE using MBASED

The python program 'allelecounter' counted and reported the number of reads that map to either the reference or alternate allele at each heterozygous SNV of MEGs. Each SNV was linked to the gene on which it has an effect, based on the SnpEff annotation using a custom R Script. As mentioned above, SnpEff associates a variant to the gene within which it has the highest impact. The longest gene transcript gets associated with the variant if multiple effects with the same impact are reported. In this study, this annotation was used to associate a variant to a gene. A caveat to this approach is that this could have led to some expressed variants present outside the gene boundaries to be associated with the gene. MBASED (Mayba, et al., 2014) was used to get gene-based ASE estimates. The MBASED program requires the input in a particular format in which each row corresponds to a single SNV containing column wise information on locus ID (in the form of scaffoldID_geneName), location, reference allele, alternate allele, reference read counts and alternate read counts. This specific input file was prepared using a custom python script. Only

those SNVs that had at least 10 reads on either reference or alternate allele were retained. Finally, the resulting output file was given as an input to the MBASED package with parameters isPhased=FALSE, numSim=10^6, BPPARAM=SerialParam() to get gene based estimates of ASE and corresponding p values (pValueASE). The resulting pValueASE was again FDR adjusted, and only those genes where the Benjamini-Hochberg (Benjamini and Hochberg, 1995) adjusted P-value was <=0.05 and majorAlleleFrequency was >= 0.7 were retained. The MBASED output was further filtered to contain only those genes with more than 100 reads (reads over reference allele and alternate allele) and the ratio between reference and alternate reads was >=1.5.

2.3 Results and Discussion

2.3.1 Alignment statistics

The merged BAM files statistics from four water buffalo are shown in Table 2. The statistics have been generated using bamtools 'stats' (Barnett, et al., 2011). There are no unmapped reads (100% mapped reads) as they were already removed in the post-processing step. No duplicates have been reported because duplicate reads were not marked in the BAM files. Alignment statistics tools identify duplicate reads with a BAM FLAG value of 1024. Since measurement of ASE depends on counting absolute read counts over each allele, marking duplicates will discount reads from highly-expressed regions that are saturated with reads (Williams, et al., 2014). Computational tools such as Picard 'MarkDuplicates' and Samtools 'rmdup' cannot distinguish whether a read is an actual PCR-duplicate or if it belongs to a highly expressed gene. An average of 85% of reads were 'proper-pairs' i.e. reads present on the same chromosome, oriented towards each other with a reasonable insert size. A very low number of reads were singletons i.e. an average of 1.44%.

Reported statistics for each individual	Male 1	Male 2	Female 1	Female 2
Total reads	508,862,462	308,570,302	446,449,471	321,822,267
Mapped reads	508,862,462 (100%)	308,570,302 (100%)	446,449,471 (100%)	321,822,267 (100%)
Failed QC	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Paired-end reads	482,204,121 (94.76%)	295,597,940 (95.80%)	426,733,284 (95.58%)	308,000,352 (95.71%)
'Proper-pairs'	411,808,940 (85.40%)	247,955,500 (83.88%)	375,443,656 (87.98%)	253,275,134 (82.23%)
Singletons	6,939,381 (1.44%)	4,604,224 (1.56%)	4,797,914 (1.12%)	5,153,922 (1.67%)

Table 2: Alignment statistics of four water buffalo BAM files calculated using bamtools stats.

2.3.2 Variant calling and variant filtration statistics

The complete variant statistics per animal are shown in Table 3. The genome-wide Ti/Tv ratio for Male 1, Male 2, Female 1 and Female 2 were 2.37, 2.38, 2.35 and 2.39 respectively, which shows more transitions (C↔T and A↔G) than transversions (A↔C, A↔T, C↔G and G↔T). These values were calculated using the unfiltered variants from master VCF files. It can be seen that the average genome-wide Ti/Tv ratio of the water buffalo is 2.37. This is similar to the genome-wide Ti/Tv ratio of 2.46 calculated for the water buffalo in another study (Surya, et al., 2018) and is slightly higher than humans. The global or genome-wide estimate of Ti/Tv ratio for human has been reported to be between 2.0 and 2.2 (DePristo, et al., 2011). The Ti/Tv ratio for filtered biallelic heterozygous SNVs in MEGs for Male 1, Male 2, Female 1 and Female 2 were 2.74, 2.71, 2.65 and 2.76 respectively. The complete variation statistics of all four animals are present in Table 3. Male 2 and Female 2 have fewer variants than Male 1 and Female 1 which is likely due to low number of mapped reads in these two animals leading to fewer SNVs for analysis (Table 2).

Variant statistics	Male 1 (n=3)	Male 2 (n=4)	Female 1 (n=3)	Female 2 (n=4)
Total variants	2,130,375	1,175,144	1,967,515	1,133,361
Number of SNVs	2,072,840	1,140,181	1,914,668	1,099,445
Number of Indels	57,535	34,963	52,847	33,916
Number of biallelic heterozygous SNVs	948,931	412,386	856,450	382,995
Number of biallelic heterozygous SNVs in MEGs (unfiltered)	37,233	28,291	38,148	28,420
Number of biallelic heterozygous SNVs in MEGs (filtered)	33,094	22,850	33,105	23,100

Table 3: Variant statistics of four water buffalo samples calculated using BCFtools stats. 'n' denotes the number of samples merged for each animal.

2.3.3 Variant annotation (functional annotation) statistics

The variant annotations (effects by type) of filtered biallelic heterozygous SNVs from each animal performed by SnpEff are shown in Table 4. SnpEff predicted that the largest number of variants lie in introns, as previously observed in sheep (Suárez-Vega, et al., 2017). The merged BAM file contains samples from both mRNA and total RNA. Total RNA samples contain immature transcripts that have not been spliced. The immature transcripts can further comprise pre-mRNA or nascent transcripts on which RNA polymerase is still attached and has not reached the 3' end (Ameur, et al., 2011). Furthermore, variants annotated as 'intergenic_region' or downstream and upstream gene variants could also arise from unannotated exons or from novel non-coding mRNA transcripts as reported in the sheep transcriptome (Suárez-Vega, et al., 2017). Because of the high depth of sequencing, aggregating data from multiple libraries, these variants are reliably detected and can be included in the ASE analysis.

Effect by type	Male 1	Male 2	Female 1	Female 2
intron_variant	614,170	245,020	542,016	219,553
intergenic_region	145,619	44,931	120,717	42,135
downstream_gene_variant	62,319	37,539	62,137	36,540
upstream_gene_variant	55,429	28,996	53,929	27,857
3_prime_UTR_variant	19,737	17,507	23,875	18,937
intragenic_variant	19,071	8,693	16,623	7,816
synonymous_variant	13,640	12,542	16,155	13,243
missense_variant	8,239	7,344	9,449	7,355
5_prime_UTR_variant	4,875	3,734	5,357	3,857
splice_region_variant&intron_variant	2,929	3,172	3,019	2,853
5_prime_UTR_premature_start_codon_gain_variant	967	721	1,023	783
splice_donor_variant&intron_variant	815	1,016	883	959
splice_acceptor_variant&intron_variant	351	487	392	458
splice_region_variant&synonymous_variant	304	300	350	262
missense_variant&splice_region_variant	184	164	200	151
stop_gained	111	82	122	101
splice_region_variant	99	79	111	76
stop_lost	24	17	33	19
stop_retained_variant	22	13	23	14
start_lost	13	16	24	17
initiator_codon_variant	6	5	5	3
stop_lost&splice_region_variant	4	3	5	3
stop_gained&splice_region_variant	1	5	1	3
start_lost&splice_region_variant	1	0	1	0
splice_region_variant&stop_retained_variant	1	0	0	0
Grand Total	948,931	412,386	856,450	382,995

Table 4: Summary statistics of the functional annotation of biallelic heterozygous SNVs in the four animals predicted by SnpEff

2.3.4 Allelic imbalance in MEGs

Based on the filtered output generated by MBASED, Male 1, Male 2, Female 1, and Female 2 showed significant ASE in 91, 71, 111 and 62 MEGs respectively. The results are summarised in Figure 10. It can be seen that in Male 1, Male 2, Female 1 and Female 2, 21%, 18%, 25% and 15% of MEGs tested show ASE out of the total genes that could be tested for ASE.

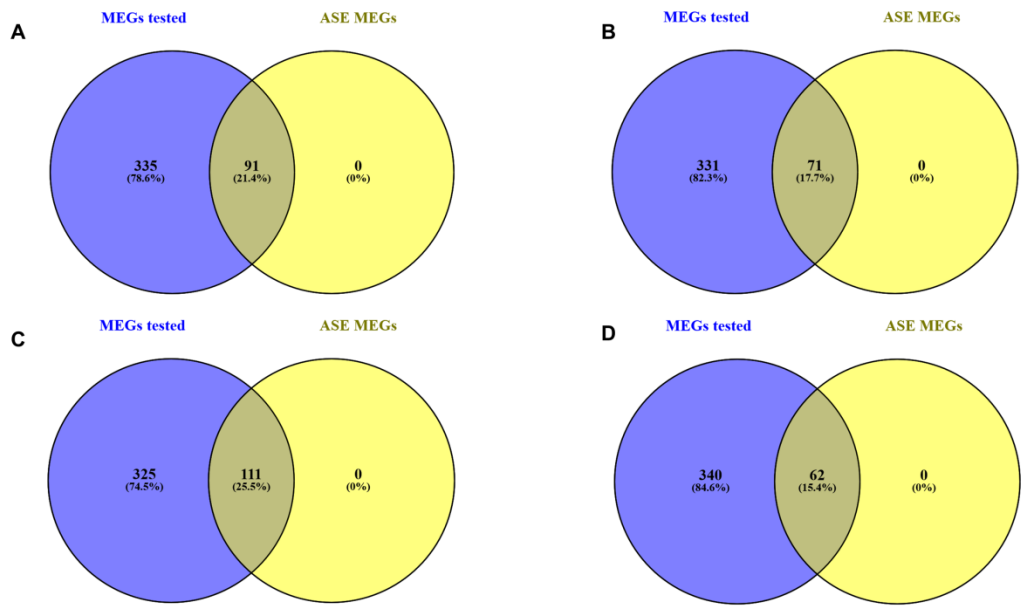

Figure 10: MBASED Results. The figure shows the number of MEGs that could be tested for ASE and the number of MEGs that showed significant ASE in (A) Male 1, (B) Male 2, (C) Female 1 and (D) Female 2.

A total of 201 unique genes showed significant ASE in at least one animal from the set of 527 MEGs. Other genes could not be tested for ASE due to the lack of heterozygous biallelic SNVs. Figure 11 shows an UpSet plot (Lex, et al., 2014) made using the R package UpSetR (Conway, et al., 2017). An UpSet plot is used to perform an intersectional analysis, for example, MEGs showing ASE amongst all 4 water buffalo samples. The distribution of the 201 genes showing significant ASE in various intersections can be observed from the figure. The ASE profile of MEGs is highly individual-specific: 20 in Male 1, 33 in Male 2, 34 in Female 1 and 22 in Female 2. The dominance of private ASE profiles and lower numbers of shared ASE genes amongst individuals was also observed in a sheep BMDM +/- LPS samples (Salavati, et al., 2019). The prevalence of individual-specific ASE MEGs supports the premise that there is substantial variation in the immune system of individuals (Brodin and Davis, 2017). In part, it is also a reflection of the presence or absence of informative SNVs. For example, the *EMP3* gene showed ASE in only Female 1 but none of the other animals had any heterozygous SNVs for that gene that could have been used for quantifying ASE. Similarly, *LOC102409115* showed ASE only in Male 2

which was the only animal with informative heterozygous SNVs. So, although these genes may also show differences in their expression between chromosomal copies in the other animals, it is not possible to test this, i.e. they are inaccessible to this approach.

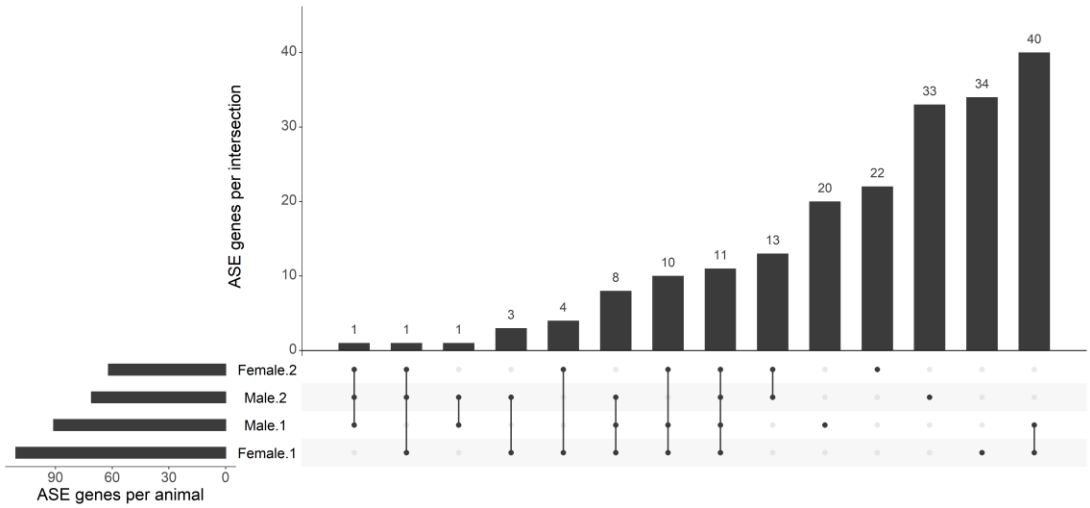

Figure 11: Intersection plot (UpSet plot) of MEGs showing ASE. The goal of this plot is to see the intersections between MEGs showing ASE amongst 4 water buffalo samples under study. The vertical black bars represent the number of elements (ASE genes) an intersection contains. The numbers on the top of the vertical bars represent the cardinality (number of ASE genes in each intersection). The dots represent the sets that contribute to an intersection. The horizontal bars to the left represent the total number of ASE genes in each sample.

40 genes showed ASE in both Male 1 and Female 1 whereas only 13 ASE genes were common between Male 2 and Female 2. Very few sex-specific ASE genes were observed in this study i.e. 1 gene in males and 4 genes in females. This should be dealt with caution due to the usage of a small sample size (n=2). The main aim of this study was to determine whether regulatory variation exists in the form of ASE in macrophage-expressed genes and to establish that it can be reliably detected. This aim has clearly been realised.

Gene symbol	Gene name	Function
ARL11	ADP Ribosylation Factor Like GTPase 11	Tumour suppressor
GALK2	Galactokinase 2	Galactose metabolism
GK	Glycerol Kinase	Glycerol uptake and metabolism
HEBP1	Heme Binding Protein 1	Calcium metabolism and chemotaxis in monocytes and dendritic cells
LOC102393437	protein FAM111B-like	Involved in congenital defects
LOC102396311	galectin-3-like	Neutrophil activation and adhesion, chemo-attraction of monocytes/macrophages
LOC102398746	nuclear autoantigen Sp-100-like	Tumour suppressor
LOC102402934	NACHT, LRR and PYD domains-containing protein 2-like	Involved in the innate immune response releasing proinflammatory cytokines upon bacterial infection
LOC102407627	2'-5'-oligoadenylate synthase 1-like	Activates latent RNase L, which results in viral RNA degradation and the inhibition of viral replication
PGK1	Phosphoglycerate Kinase 1	Glycolysis in macrophages
VSIG4	V-Set And Immunoglobulin Domain Containing 4	Negative regulator of T-cell proliferation and IL2 production

Table 5: 11 MEGs showing significant ASE in all four water buffalo

Table 5 summarises the MEGs that showed evidence of ASE in all 4 animals. Out of the 11 genes, *ARL11* had two missense variants that were shared across all 4 animals. This gene is required for regulating macrophage activation during LPS stimulation or pathogen encounter (Arya, et al., 2018). *LOC102402934* or NACHT, LRR and PYD domains-containing protein 2-like (*NLRP2-like*) is a pathogen recognition receptor that has essential roles in immune protection against various pathogens and is part of the inflammasome which are innate immune system receptors and sensors regulating inflammation in response to infectious diseases (Guo, et al., 2015; Yang, et al., 2016). This gene also had many missense variants associated with it that were common amongst all four animals.

Most of the genes mentioned encode receptor molecules that either recognize the pathogen or trigger an immune response against it. Since these genes were induced with LPS (mimicking a pathogen), we have quantitatively shown the presence of regulatory variation in macrophage-expressed genes through ASE and have made a catalogue of genes that can be used further for case-control studies to identify genes associated with disease susceptibility.

2.4 Conclusions

This chapter characterises regulatory variation in the form of ASE at a gene-level in macrophage expressed genes in water buffalo using only RNA-seq data. The presence of regulatory variation in immune-related genes may determine an individual's susceptibility to diseases. Macrophage-specific samples were chosen as they harbor many innate immune genes necessary for starting an innate immune response which is the first line of host defense against any pathogen attack.

ASE studies have been done in various livestock such as cattle (Chamberlain, et al., 2015; Guillocheau, et al., 2019), sheep (Salavati, et al., 2019), goats (Cao, et al., 2019), pigs (Ahn, et al., 2019; Stachowiak, et al., 2018) and chicken (Zhuo, et al., 2017). This is the first ASE study that has been performed in water buffalo.

Merging of multiple macrophage samples per animal into a single BAM file provided an advantage of greater depth that allowed reliable identification of SNVs and ASE quantification. However, this method was confounding to the identification of sample-specific ASE which is a caveat to this study.

The ASE analysis revealed the presence of regulatory variation in some important PRR genes including *NLRP2-like, TLR7*, *NOD1*, *NLRC4* (NLR family CARD domain-containing protein 4) which could contribute to variation in disease resistance in water buffalo which is an economically important trait. ASE has been associated with disease resistance or regulation of immune response in humans (Lappalainen, et al., 2013), chicken (Maceachern, et al., 2011) and goats (Cao, et al., 2019).

This study also suggested that ASE is very individual-specific. In some cases, this is simply a function of the available informative SNVs in the mRNA to enable detection. It is possible that more variation could be uncovered with even greater depth of sequencing and lower stringency. In a larger study with more animals, it may also be possible to identify and compare the level of expression

in macrophages from animals that are homozygous for different haplotypes to confirm the existence of high and low expression alleles.

A limitation of this this study is that copy number variants or CNVs were not taken into account. CNVs are large DNA segments of length >= 1kb which are present in variable copy number when compared to a reference genome (Feuk, et al., 2006). CNV regions harbor many protein-coding genes and as a consequence lead to gene dosage imbalance, disruption of protein coding regions and affecting overall gene regulation (Zhou, et al., 2011). Any gene present in a CNV region showing ASE would be an artefact. A CNV analysis should be performed and any protein-coding genes present in those regions should be discarded from further analysis as done in (Guillocheau, et al., 2019).

Another limitation of this study was not considering the presence of overlapping genes. If an SNV is present in the overlapping segment of two genes, it is very difficult to assign a gene name to it. Although the variant annotation tool assigns multiple annotations to a particular variant, we have considered only the first annotation and assigned the gene name of that annotation to the variant.

We were not able to explore/identify monoallelic expression, which is an extreme form of ASE. In the current study, variant calling was performed using RNA-seq data. Hence, true monoallelic sites will have a homozygous genotype. An extensive whole-genome ASE analysis on the water buffalo was carried out which has been discussed in the next chapter. The issue of reference bias, which is an important caveat in this chapter, has also been explored.

2.5 Supplementary Material

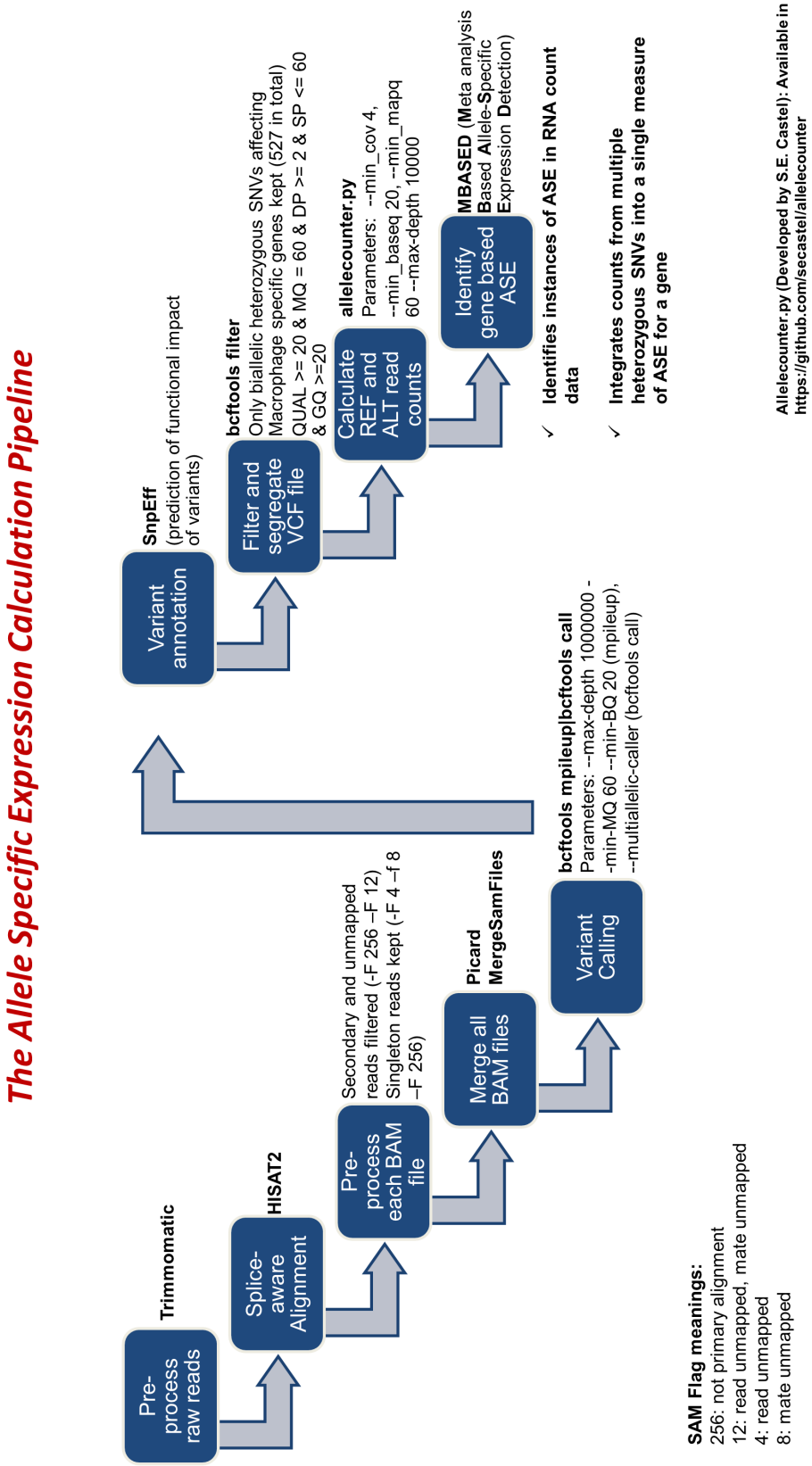

Figure S 2: Allele-specific expression calculation pipeline

APBB3	GCLC	LOC102398746	LOC102414334	PPBP	TFEC	CTSH	LOC102390560	NOD1	XM_006044187	M6PR	BIRC3	ATP6V1E1	MRC1
ARG2	GCNT2	LOC102398779	LOC102414937	PRKX	TIRAP	CTSK	LOC102391055	NRP2	XM_006045970	MKNK1	CASP8	C7ORF60	NAAA
ARID5B	GDNF	LOC102398793	LOC102415147	PTGS2	TLR2	CTSS	LOC102392421	OCSTAMP	XM_006051370	MMP19	CCL5	CLTC	PPT1
ARL11	GK	LOC102399573	LOC102415899	PTPN12	TLR4	CTSZ	LOC102393233	OSBPL11	XM_006078282	MR1	CD40	GLS	PRCP
BATF3	GPR110	LOC102400283	LOC102415907	PTPRE	TMEM170B	CX021	LOC102396311	OSCAR	ZMYND15	MSN	CD80	HTRA1	SELPLG
BCL2A1	HERC6	LOC102400303	LOC102416225	PTPRJ	TMEM2	CYP4V2	LOC102396916	PKIB	AMDHD2	MVP	CMPK2	LOC102393257	TLR8
CASP13	IER3	LOC102400350	LRRC18	PTX3	TNFRSF8	DCLRE1B	LOC102397367	PLAU	ARL4C	NFE2L2	CXCL16	LOC102403104	TNFAIP8L2
CCL2	IFIT1	LOC102401129	LRRC8C	PVR	TNFSF13B	DFNA5	LOC102398694	PLIN2	ARPC1B	OSTF1	EIF4E	LOC102412417	
CCL24	IL17A	LOC102401256	LVRN	RAPH1	TNFSF15	DNMT3L	LOC102398755	PLXNC1	ARPC5	OSTM1	EMILIN2	LOC102415810	
CCL3	IL17F	LOC102401348	MAFG	RIPPLY3	TRAF1	DRAM1	LOC102399881	PSAP	CD53	P4HA1	FMNL3	MFSD11	
CCL4	IL17REL	LOC102402934	MAP3K8	RNF13	TRPV3	EMB	LOC102401690	PTAFR	CTSA	PDCD6IP	GBP1	MFSD12	
CD44	IL1A	LOC102402993	MBOAT1	RSAD2	TXN	EMR3	LOC102402134	RAB31	CTSD	PICALM	HERC5	NIPA1	
CD55	IL1B	LOC102403078	MCOLN3	RSPRY1	TXNRD1	FAM20A	LOC102402272	RAB7B	DNASE2	PLBD2	IFI44	NPC1	
CLEC4D	IL1RAP	LOC102403552	METTL1	S100A12	UNC93A	FGD4	LOC102402326	RBPJ	DOK2	PLEKHO2	IFI44L	PAQR8	
CLEC4E	IL27	LOC102403680	MICALCL	S100A8	UPP1	FLVCR2	LOC102402675	RGL1	DRAM2	PQLC2	IFIH1	PGAM1	
CSF2	IL36G	LOC102403711	MOB1B	S100A9	USP12	FTH1	LOC102405926	S100A4	DUSP6	PRDX1	IFIT3	PGK1	
CSF3	IL6	LOC102403850	MT2A	SAMD9	WFS1	GALNS	LOC102406324	SCIMP	EFHD2	PTPLAD2	IL12B	PTPN11	
CXCL6	IL7R	LOC102403864	MX1	SCNN1D	ZBTB20	GAS2L3	LOC102408791	SCIN	EVI2A	PWWP2B	IL15RA	RCBTB1	
CYBB	IL8	LOC102404177	MX2	SDC4	ZC3H12C	GBA	LOC102409415	SCPEP1	FAM49A	RAB32	IL1RN	SLC36A4	
CYTIP	INHBA	LOC102404503	NAV3	SDS	ZDHHC18	GGCT	LOC102409999	SEMA3A	FTL	RAB8B	IRF7	SLC48A1	
CYTX	IQGAP1	LOC102405513	NEK6	SERPINB2	ZNFX1	GLA	LOC102410120	SIGLEC15	FXYD5	RALB	ISG20	SNX10	
DCSTAMP	IRAK2	LOC102405577	NEU1	SERPINB8	ADAM28	GLMP	LOC102411087	SKAP2	GALK2	RFESD	JAK2	STAM	
DDX58	ISG15	LOC102405722	NFAT5	SH2D1B	ADAP2	GNS	LOC102411162	SLAMF8	GLIPR1	RGS10	LOC102389096	STX6	
DNAH17	ITGAV	LOC102405755	NINJ1	SLAMF7	BLVRA	GRN	LOC102412304	SLC2A9	GLIPR2	RNASEL	LOC102395641	TBC1D2	
DNAJC13	ITGB8	LOC102406056	NLRC4	SLC11A1	C17ORF96	HEBP1	LOC102413747	SLC35F6	GNB4	S100A11	LOC102402487	AOAH	
EDN1	LOC102389031	LOC102406566	NLRP3	SLC11A2	C5AR1	HK2	LOC102414053	SLC43A2	GNG2	SARS	LOC102407126	C3AR1	
EMP3	LOC102389399	LOC102407099	LOC102401026	SLC13A5	CAPG	HNMT	LRRC25	SLC6A6	GPX1	SDCBP	LOC102409538	COLGALT1	

EMR1	*LOC102389879*	*LOC102407404*	*NUDT9*	*SLC2A6*	*CCR1*	*IFI30*	*LXN*	*SPP1*	*HEXA*	*SLC29A3*	*LOC102412516*	*CSF1R*	
ENTPD7	*LOC102390393*	*LOC102407627*	*NUMB*	*SLC31A1*	*CCR5*	*IGSF6*	*LY9*	*TGFBI*	*IFNAR1*	*SMIM3*	*MB21D1*	*FOLR2*	
ETV3L	*LOC102390729*	*LOC102408026*	*NXT1*	*SLC31A2*	*CD14*	*ITGAM*	*MAFB*	*THBS1*	*IL7*	*SPI1*	*NFKB2*	*GCNT1*	
FAM102B	*LOC102390987*	*LOC102408072*	*PARP3*	*SLC38A6*	*CD300E*	*ITGAX*	*MERTK*	*TLR7*	*IRF5*	*SPPL2A*	*PARP14*	*GDA*	
FBXO30	*LOC102393437*	*LOC102408349*	*PDPN*	*SLC6A12*	*CD68*	*ITGB2*	*MMP1*	*TM4SF19*	*LAMP1*	*TAGLN2*	*SLC25A19*	*GPR34*	
FCAR	*LOC102393611*	*LOC102408483*	*PDXK*	*SLC7A11*	*CLEC5A*	*LDHA*	*MMP12*	*TMEM150B*	*LGALS3BP*	*TALDO1*	*STAT1*	*KCTD12*	
FCER1G	*LOC102395010*	*LOC102408763*	*PF4*	*SOD2*	*CLEC6A*	*LGMN*	*MMP14*	*TMEM251*	*LOC102395533*	*TMEM104*	*TDRD7*	*KLHL5*	
FGR	*LOC102395851*	*LOC102409115*	*PFKFB3*	*SQSTM1*	*CLEC7A*	*LHFPL2*	*MPEG1*	*TMEM26*	*LOC102401307*	*TMEM243*	*TNIP1*	*LOC102394636*	
FKBP15	*LOC102396130*	*LOC102409205*	*PIK3AP1*	*SRXN1*	*CPM*	*LIMS1*	*MSR1*	*TREM2*	*LOC102405541*	*TOM1*	*TRIM25*	*LOC102404615*	
FLT1	*LOC102396163*	*LOC102409828*	*PIK3R5*	*SUCNR1*	*CREG1*	*LIPA*	*MYOF*	*TRPV2*	*LOC102410916*	*TPP1*	*USP18*	*LOC102406660*	
FNIP2	*LOC102397333*	*LOC102409866*	*PLA2G4F*	*SUSD1*	*CSTB*	*LOC102389759*	*NAGLU*	*TYROBP*	*LOC102413162*	*ZCWPW1*	*ATP6V0D1*	*LOC102410536*	
FOSL1	*LOC102397971*	*LOC102410193*	*PLEKHM3*	*TBC1D9*	*CTSB*	*LOC102389993*	*NAIP*	*VAT1*	*LOC102414960*	*ZNRF2*	*ATP6V1A*	*LOC102415243*	
G0S2	*LOC102398505*	*LOC102410742*	*PLXNA1*	*TCIRG1*	*CTSC*	*LOC102390451*	*NCF2*	*VSIG4*	*LOC102416018*	*ACSL5*	*ATP6V1B2*	*MARCH1*	

Table S 3: 527 macrophage expressed genes (MEGs) obtained from the macrophage specific clusters from the water buffalo gene expression atlas

Chapter 3. ASE using DNA-seq and RNA-seq data

3.1 Introduction

One of the major steps in determining allele-specific expression (ASE) requires the identification of heterozygous biallelic variant sites. From a next-generation sequencing (NGS) perspective, when reads generated by a sequencer are aligned to a reference genome, 'variant calling' is performed to identify the biallelic variant sites and 'genotype calling' determines the precise genotype. As mentioned in the previous chapter, RNA-seq variant calling can be applied directly to identification/quantification of transcript isoforms provided sufficient sequencing depth is available. RNA-seq variant calling is constrained by alignment errors due to splicing, increased error rate due to reverse transcription, failure to call variants and infer genotypes in lowly expressed genes, incorrect inference of RNA editing sites as genomic variation and finally the inability to call complete ASE i.e. monoallelic expression (Xu, 2018). These points have been briefly described ahead.

In the previous chapter, RNA-seq reads were aligned to a reference genome to identify genomic variants. RNA-seq short reads are derived from mature-RNA (in the case of mRNA-seq) which do not have introns in their sequence. A read generated from an mRNA-seq experiment may span multiple exons, while the reference genome will have an exon, followed by an intron leading to the splitting of a single read (known as 'split-reads' formed due to a splice junction). This would need an aligner to deal with issues such as mapping ambiguity due to the short length of reads and sequencing errors leading to misaligned split reads. 'Splice-aware' aligners follow different strategies to identify the splice junctions and alternative splicing that enables a gene to code for multiple proteins through the inclusion or exclusion of exons (Engström, et al., 2013; Mapleson, et al., 2018; Williams, et al., 2014). Errors introduced during the processes of reverse transcription and sequencing cannot be distinguished from

genuine variants (Carey, 2015). RNA-seq read distribution on a reference genome (when aligned to one) is non-uniform in nature which means that reads will be concentrated on highly expressed genes relative to genes which have low or no expression (Muzzey, et al., 2015). By contrast, reads are distributed more uniformly across the whole genome in DNA-seq (except repetitive regions) (Rozowsky, et al., 2011) which increases the confidence of variant calling. RNA editing refers to a post-transcriptional modification in which a mature-RNA sequence differs from its template DNA sequence because of base substitution leading to RNA mutations and altering gene regulation (Brennicke, et al., 1999; Ouyang, et al., 2018). When RNA-seq data alone is used to detect variation by alignment to the reference genome, it is impossible to distinguish edits from SNVs. Having the genomic sequence of the same organism under study enables identification and correct interpretation of such events (Bahn, et al., 2012; Bakhtiarizadeh, et al., 2018; Eisenberg, et al., 2005; Park, et al., 2017).

An issue arising from mapping reads to the reference genome is the reference bias due to preferential mapping of the reference allele (Pai, et al., 2009). Various methods have been previously described in order to address this issue such as converting known variant sites in the reference genome with a third base that is neither the reference allele nor the alternate allele (SNP or Single-nucleotide polymorphism masking) (Pai, et al., 2009), masking known heterozygous variant sites with an ambiguous nucleobase 'N' (also known as 'N-masking') (Krueger and Andrews, 2016), creating personalised diploid genome (Rozowsky, et al., 2011) and creating 'pseudogenomes' or parental genomes from a standard reference genome by incorporating variant information (Huang, et al., 2014). These methods do not completely eliminate reference bias and some methods require pedigree information. For example, AlleleSeq incorporates phased variation data such as SNPs, indels and structural variations from the parents and the child for creating a personalised diploid genome to identify ASE (Rozowsky, et al., 2011). Pedigree information is unavailable in most cases. Mapping to pseudogenomes can lead reads containing either the reference or alternate allele to not map uniquely or map to an incorrect position in the genome (Krueger and Andrews, 2016; van de Geijn,

et al., 2015). Genome masking methods can remove the differences in the reference and alternate versions of the genome which can allow the reference and alternate allele containing reads to map to the masked genome equally (Krueger and Andrews, 2016; Pai, et al., 2009). However, an aligner can consider that one base mismatch to be present anywhere in the genome and the reads may incorrectly map to some other homologous region (Pai, et al., 2009). MBASED can incorporate information about pre-existing reference or allelic bias into ASE quantification (Mayba, et al., 2014). The details have been explained further ahead in the chapter.

This chapter also explores the extent of monoallelic expression or MAE in water buffalo. In mammals, MAE has been mainly divided into three types. The first type is genomic-imprinting or parent-of-origin imprinting in which one of the alleles (either maternal or paternal) becomes transcriptionally silent due to marks (for example, methylation) present on the promoter region of an allele during gametogenesis. All the cells where a gene is imprinted will express the same allele of the gene according to its parent of origin. The remaining two types are random monoallelic expression (RMAE) (including X chromosome inactivation or XCI) and autosomal RMAE. These two types are independent of parent-of-origin. XCI is random silencing of one of the X chromosomes in female diploid cells. In autosomal RMAE, autosomal genes of some cells express the maternal allele and some express the paternal allele. This feature is very random in nature and independent of parental origin. This usually happens in order to generate diversity among cells and their clonal descendants leading to the generation of unique cell identity among individual cells. XCI is an RMAE event which is restricted only to the X chromosome. RMAE events are generally widespread among autosomal chromosomes and they are present in genes that show cell-type specific expression and encode cellular surface protein and diverse signalling molecules such as immunoglobulins, T-cell receptors, natural killer cell receptors, interleukins and genes from odorant and pheromone receptor genes (Chess, 2016; Gimelbrant, et al., 2007). RMAE events can provide functional mosaicism among similar cells of a tissue and affect processes that have an important role in the interface between cells and the

environment (Nag, et al., 2015). Whereas all imprinting events are MAE events, all MAE events (such as XCI or autosomal RMAE) may or may not be imprinting events.

Since RMAE events are random and heterogeneous at a cellular level, RMAE events are deciphered using single-cells or clonal cell lines (Eckersley-Maslin and Spector, 2014). This is because RMAE is not observed in non-clonal cell populations or whole tissue unlike imprinted genes. Even in clonal cell lines, the same gene can show biallelic expression in one cell line and show monoallelic expression in another (Zwemer, et al., 2012). A genome-wide study in humans using B-lymphoblastoid clonal cell lines discovered that ~5 to 10% of 4,000 autosomal genes under study showed RMAE (Gimelbrant, et al., 2007). A similar range of RMAE events was observed in mouse lymphoblast clonal cell lines where over 10% of the 1,300 autosomal genes studied showed RMAE (Zwemer, et al., 2012). Single-cell RNA-seq experiments are beneficial in exploring such RMAE events (Deng, et al., 2014).

ASE analysis on the X chromosome has been avoided in this study. ASE involving sex chromosomes is a complex phenomenon. Females have two copies of the X chromosome, whereas males have a single copy of it along with a Y chromosome. As mentioned above, XCI is a biological process in which transcription of genes in one of the X chromosome copy in female cells is silenced in order to balance the expression dosage in males and females cells (Tukiainen, et al., 2017). The pseudoautosomal regions (PAR1 and 2) on the sex chromosomes have genes whose expression is biallelic like autosomal genes. Genes present in the non-pseudo-autosomal region (X-linked genes) will show MAE. However, some genes outside the pseudo-autosomal region escape XCI in order to express both gene copies (Balaton and Brown, 2016; Carrel and Willard, 2005).

In a 2018 study encompassing 29 tissues from 554 donors, a catalogue of genes that show incomplete XCI along with genes that escape XCI was observed to be variable between cells, tissues and individuals (Tukiainen, et al., 2017). Using fibroblast and lymphoblastoid cell lines from 5 individuals, this kind

of XCI heterogeneity was also observed in human females in which the ability of the genes escaping XCI varied among different individuals, cells and genes (Garieri, et al., 2018). Tissue and cell-line specific incomplete XCI was also observed in mice (Berletch, et al., 2015). X chromosome gene expression patterns including genes escaping XCI have been extensively studied in cattle involving various tissues by Min *et al* (Min, et al., 2017) and Duan *et al* (Duan, et al., 2018). Furthermore, the pseudo-autosomal region has also been identified in cattle (Das, et al., 2009). Such studies documenting incomplete XCI and genes escaping XCI are absent for water buffalo. The pseudo-autosomal region has not been characterised for the water buffalo. Keeping in mind the aforementioned complexities involving the sex chromosomes, the X chromosome has been removed from ASE quantification. Additionally, Y chromosome is not a part of this study because the reference genome belongs to a female water buffalo.

MAE also occurs where one of the two alleles contains a mutation that ablates detectable expression. Such variation may arise from the loss or interruption (for example by repeat element insertion) of regulatory elements (Feschotte, 2008). It may also arise from protein-coding mutations that lead to premature termination and nonsense-mediated decay (Rivas, et al., 2015). This kind of MAE is invisible to RNA-seq, which will detect only homozygous SNVs, but may be functionally important indicating recessive loss-of-function alleles that could be deleterious as homozygotes.

This chapter deals with the determination of genome-wide ASE by identifying and genotyping variants using DNA-seq data. The variant calling workflow has been 'adapted from' the GATK (The Genome Analysis Toolkit) (McKenna, et al., 2010) best practices workflow for germline single nucleotide polymorphisms (SNPs) and Indels (Insertions or deletions) for our non-human organism as mentioned in https://software.broadinstitute.org/gatk/best-practices/workflow?id=11145. The results present important observations of presence of regulatory variation in the form of ASE, not only in immune-related genes, but also in genes involved in important biological processes such as growth and development.

3.2 Methods

3.2.1 Whole Genome Sequencing (WGS) raw data description

As shown in Table 6, the raw data consisted of WGS data from 81 water buffalo from 6 Indian breeds plus the Mediterranean breed from which the reference sequence was derived. Paired-end sequencing of the animals was done across two sequencing centres: Edinburgh Genomics in the UK and SciGenom in India. The mean sequencing depths at each sequencing centre were 30x and 10x respectively. Further sample information is provided in Table 6. Edinburgh Genomics sequenced the animals using the Illumina HiSeq X platform (read length: 150 bp) whereas SciGenom used the Illumina HiSeq 2500 platform (read length: 250 bp). The animals from the Mediterranean breed consist of two animals from Lodi (Italy) and four animals from the same farm in Kirkcaldy, Fife, on the East coast of Scotland. The Indian breed samples belong to their respective breeding tracts as mentioned in Table 7.

Breed	Number of animals	Sequencing Centre	Depth
Italian (Mediterranean)	6	Edinburgh Genomics	30x
Jaffrabadi	13	SciGenom (India) and Edinburgh Genomics	6 animals-10x 7 animals-30x
Pandharpuri	13	SciGenom (India) and Edinburgh Genomics	6 animals-10x 7 animals-30x
Banni	12	SciGenom (India) and Edinburgh Genomics	6 animals-10x 6 animals-30x
Surti	12	SciGenom (India) and Edinburgh Genomics	6 animals-10x 6 animals-30x
Murrah	12	SciGenom (India) and Edinburgh Genomics	6 animals-10x 6 animals-30x
Bhadawari	13	SciGenom (India) and Edinburgh Genomics	6 animals-10x 7 animals-30x
7 BREEDS	**81 ANIMALS**	**2 SEQUENCING CENTRES**	**2 DEPTHS**

Table 6: WGS raw data summary showing sequencing information of 81 animals from 7 breeds

Breed	Breeding Tract in India	Link
Banni	Banni area of Kutch district of Gujarat	http://dairyknowledge.in/article/banni
Bhadawari	Bhind and Morena districts of Madhya Pradesh and Agra and Etawah districts of Uttar Pradesh	http://dairyknowledge.in/article/bhadawari
Jaffrabadi	Amreli, Bhavnagar, Jamnagar, Junagadh, Porbandar and Rajkot districts of Gujarat state	http://dairyknowledge.in/article/jaffarabadi
Murrah	Hisar, Rohtak, Gurgaon and Jind district of Haryana and Delhi	http://dairyknowledge.in/article/murrah
Pandharpuri	Solapur, Sangli and Kolhapur districts of Maharashtra	http://dairyknowledge.in/article/pandharpuri
Surti	Vadodara, Bharuch, Kheda and Surat districts of Gujarat	http://dairyknowledge.in/article/surti

Table 7: Breeding tract information of six Indian water buffalo breeds. Courtesy: Information System on Animal Genetic Resources of India (AGRI-IS) - developed at National Bureau of Animal Genetic Resources, Karnal, Haryana, India

3.2.2 Read alignment to the new water buffalo reference genome and pre-processing before variant calling

Raw paired-end reads from all 81 samples were aligned using BWA-MEM v0.1.17 (Li, 2013) to a newly assembled water buffalo reference genome (Low, et al., 2019) released on the NCBI website on 14th May, 2018. The genomic sequence was provided by Dr. Lloyd Low from University of Adelaide prior to its publication. NCBI completed and released its gene annotation for this genome on 25th June 2018 (ftp://ftp.ncbi.nih.gov/genomes/Bubalus_bubalis/README_CURRENT_RELEASE). All analysis has been done using the unofficial copy of the currently available reference genome. The unofficial version of the assembly is extremely similar to the one that is currently present in NCBI named 'UOA_WB_1' and can be found in the following link- ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/vertebrate_mammalian/Bubalus_bubalis/latest_assembly_versions/GCF_003121395.1_UOA_WB_1.

Figure 12 depicts this point in a dotplot made by performing a pairwise alignment using minimap2 (Li, 2018) and then using the 'minidot' program from miniasm (Li, 2016).

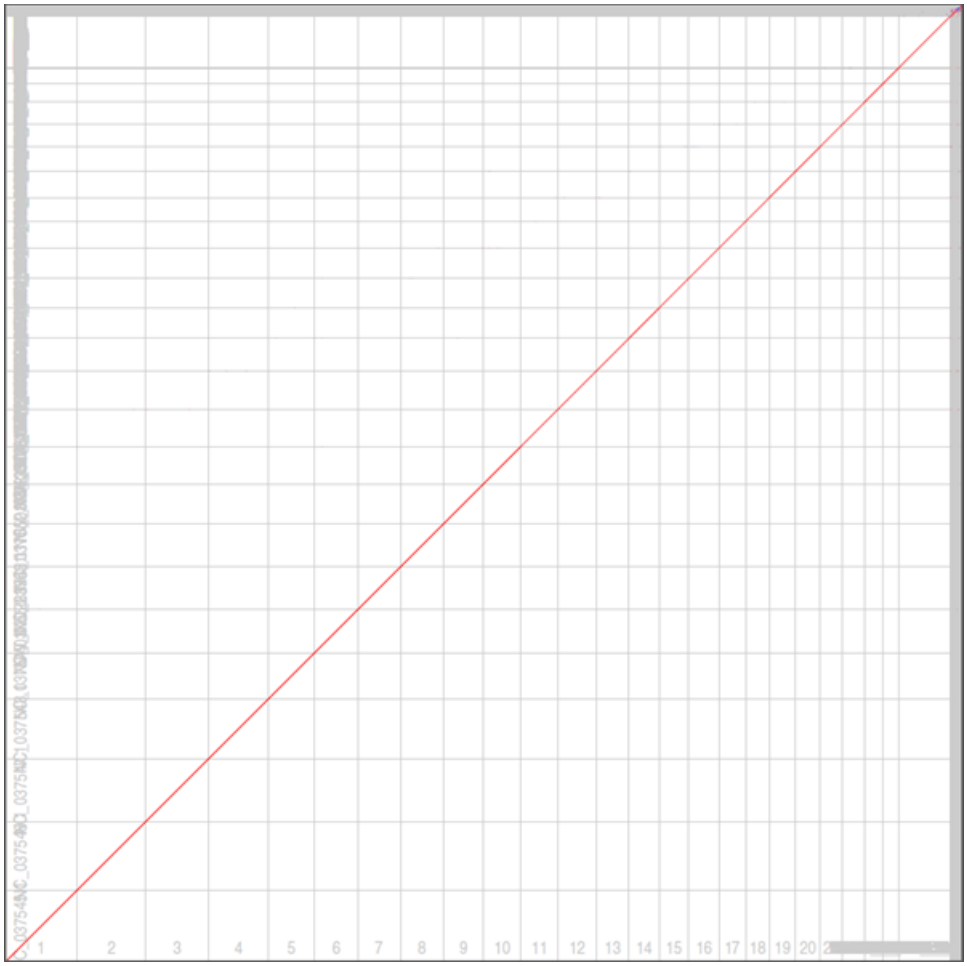

Figure 12: A dotplot between the unofficial (x-axis) and NCBI RefSeq version (y-axis) of the water buffalo genome assembly created using minimap2 and miniasm. The straight line signifies that there are no differences in any of the large contig after pairwise alignment between the two assemblies with only slight differences in the top- right corner

The plot contains a straight line throughout with only few differences. The differences are due to the following reasons:

a) hic5004_arrow (unplaced scaffold) was removed by RefSeq as a contaminant.

b) A mitochondrial genome (MT) was added by RefSeq (RefSeq-Accn: NC_006295.1) that belongs to a swamp buffalo from China (https://www.ncbi.nlm.nih.gov/nuccore/52220982).

c) NCBI annotation assigned a unique RefSeq accession id to all of the chromosomes. Since the original assembly was used in our analysis

before annotation, the chromosomes have been assigned numbers: 1,2,3,4 and so on. An already developed mapping file can be easily used to update the chromosome IDs to RefSeq assigned IDs if needed.

Statistic	Old water buffalo genome	New water buffalo genome
Assembly name	UMD_CASPUR_WB_2.0	UOA_WB_1
Submitter	University of Maryland	University of Adelaide
Date	9th Sep 2013	14th May 2018
Assembly level	Scaffold	Chromosome
Assembly method	MaSuRCA v. 1.8.3	Falcon-Unzip v. 1.8.7
Genome coverage	70.0x	69.0x
Sequencing technology	Illumina GAIIx; Illumina HiSeq; 454	PacBio
Total scaffolds	366,982	509
Contig-N50	21,938	22,441,509
Total-length	2,836,150,610	2,655,780,776
Total-gap length	74,388,041	373,500

Table 8: Difference between old water buffalo draft assembly and improved water buffalo new assembly (Data from NCBI RefSeq database)

The new assembly was generated using PacBio long read sequencing technology that gave it better contiguity than the old draft assembly. Both the assemblies are from the same animal, i.e., a female Mediterranean water buffalo named 'Olympia'. Table 8 describes the differences between the two assemblies. The contig-N50 (length of the shortest contig at the 50% assembly length) is clearly improved in the new assembly and the total-length (2.65 Gb) is reduced (2.83 Gb) through the resolution of the many gaps (total-gap length=74388041). Furthermore, the difference in length can also be attributed to the reason mentioned above i.e. removal of contaminants and addition of mitochondrial genome.

The new assembly has a total of 509 scaffolds, 24 autosomes and 1 sex chromosome (X) and 484 unplaced scaffolds; already superior to cattle GCF_002263795.1_ARS-UCD1.2 (2,211 scaffolds) and human GCF_000001405.38_GRCh38.p12 assemblies (874 scaffolds).

The new reference genome was first indexed using the BWA index command (v0.1.17) followed by the actual alignment process using BWA-MEM v0.1.17

using the paired-end reads of each sample. For each sample, the BWA-MEM generated Sequence Alignment Map (SAM) output was converted to Binary Alignment Map (BAM) output using Samtools view v1.6 with -b parameter. On completion of the alignment process, each of the 81 BAM files from individual animals was coordinate sorted using Samtools sort v1.6 with default parameters. Duplicates were marked using Picard MarkDuplicates v2.14.0 (Wysoker, et al., 2013). In brief, the tool locates and marks/tags duplicates in a BAM file (marks duplicate reads with the hexadecimal value of 0x0400, which corresponds to a decimal value of 1024), which allows downstream GATK tools to exclude duplicates from analysis based upon the probability of being non-independent measurements from the exact same template DNA. Then, each de-duplicated BAM file was subjected to Picard 'AddOrReplaceReadGroups' v2.14.0 with parameters RGLB=library RGPL=illumina RGPU=barcode RGSM=<sample_name> CREATE_INDEX=true for adding 'read groups'. This refers to a set of reads that was generated from a single run of a sequencing instrument and is necessary for all downstream GATK tools. As seen from the parameters, the BAM index was generated on the fly during the tool execution. After performing these steps, 81 pre-processed BAM files were used for variant calling.

3.2.3 Variant calling

For calling variants, the GATK tool called HaplotypeCaller v4.0.4.0 (Poplin, et al., 2018) was run per sample in GVCF (Genomic VCF/Variant Call Format) mode with the -ERC GVCF parameter. HaplotypeCaller utilizes reads from BAM/SAM files and performs variant calling per sample to produce unfiltered genotype likelihoods (Poplin, et al., 2018).

HaplotypeCaller produced 81 GVCFs in total for our cohort. Before joint genotyping, GenomicsDBImport v4.0.4.0 was used to aggregate the GVCF files from all 81 samples. It takes one or more single-sample GVCFs and imports data over a single interval, and outputs a directory containing a GenomicsDB datastore with combined multi-sample data. In v4.0.4.0 of GenomicsDBImport, there was a caveat that it could only be run on a single genomic interval (i.e. a

maximum of one contig) at a time (this caveat being removed in v4.0.6.0). GenotypeGVCFs v4.0.4.0 with the -new-qual parameter was used to read from the created GenomicsDBs directly and output the final multi-sample VCF file per scaffold. GenotypeGVCFs performs joint genotype calling, assigns a quality score (QUAL) to each variant and removes low quality variants (Poplin, et al., 2018). Finally, Picard GatherVcfs v2.14.0 was used to concatenate variants called per chromosome/unplaced scaffold to get the final multisample VCF file for all scaffolds together.

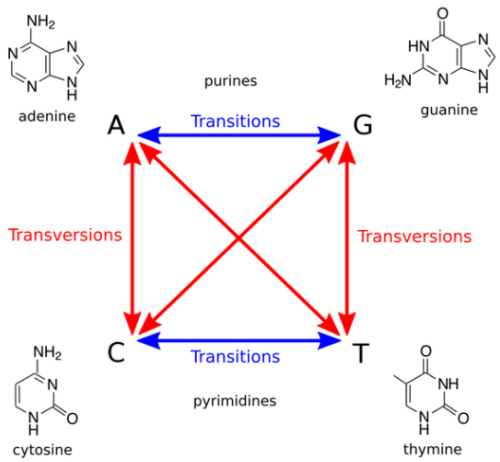

Figure 13: The transition vs transversion (Ti/Tv ratio) is calculated by dividing the number of transition SNPs by the number of Transversions SNPs. Image courtesy: By Petulda [CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0)], from Wikimedia Commons

3.2.4 Variant Filtration

Several errors in NGS technologies can lead to false- positives in the variant and genotype calling (Ribeiro, et al., 2015). These may arise due to base calling errors or may creep in during read alignment (Nielsen, et al., 2011). As per the GATK 'best practices' workflow for SNPs and Indels, the raw variant callset should pass through a 'soft-filtering' pipeline, performing variant quality score recalibration (VQSR) as described in the following link- https://software.broadinstitute.org/gatk/documentation/article?id=11084. Based on a 'truth-set' of known and validated good quality, true-positive variants (from sources such as HapMap (Gibbs, et al., 2003) , dbSNP (Sherry, et al., 2001),

etc.), the pipeline analyses/studies variant annotations (properties or statistics that describe each variant) in the raw variant callset. Based on the analysis, it develops a machine learning model to distinguish the good from the bad variants. The rules are then applied to the whole variant callset and a single score is assigned to each variant which defines the variant to be true or false under the model.

The VQSR pipeline requires the 'truth-set' which is clearly not available for *Bubalus bubalis*. Hence, a 'hard-filtering' approach was adopted. In the hard-filtering approach, the distribution of each variant annotation is checked individually across the variant data and a threshold is chosen. Variants are filtered out above or below this threshold (based on the type of variant annotation or metric). This approach can produce significant false-positive and false negative rates, which can only be calibrated empirically (https://software.broadinstitute.org/gatk/documentation/article?id=11069).

The transition vs transversion (Ti/Tv) ratio is often used as a quality indicator of variation data produced from NGS experiments and higher Ti/Tv ratio is an indicator of good quality SNVs (Wang, et al., 2015). As shown in Figure 13, transitions are the interchange within purines and pyrimidines whereas transversions are changes from a purine to a pyrimidine or vice versa. Transition mutations are more common than transversion mutations because in the latter, the interchange between one-ring and two-ring systems require more energy than the former (Wang, et al., 2015). Also, transitions are less likely to produce a difference in the amino acid sequence leading to a non-synonymous change compared to transversions (Guo, et al., 2017). Another reason why transitions occur more often is cytosine methylation, the most common form of modification in eukaryotic DNA (Cooper and Gerber-Huber, 1985). 5-methylcytosine often spontaneously deaminates turning it into a T (Coulondre, et al., 1978). The Ti/Tv ratio differs between genomic regions and it can also be used to distinguish between exonic and non-exonic regions. For example, Table 9 compiles Ti/Tv ratios calculated for five major genomic regional categories in humans (Wang, et al., 2015). The higher values of Ti/Tv in exons is due to the presence of a higher number of methylated cytosines within exons (Hodges, et al., 2009). Previous

studies have also found that for WGS data in humans, the genome-wide Ti/Tv ratio should be around approx. 2.0 - 2.1. A ratio which is lower than this can indicate poor quality data (DePristo, et al., 2011). The Ti/Tv ratio for WGS data of *Bos taurus* (cattle) was also found to be in the same range as those in humans (Baes, et al., 2014).

Genomic Region	Ti/Tv ratio
Exon	3
Intron	2.2
Intergenic	2.06
lncRNA	2.06
miRNA	2.59-2.95

Table 9: Ti/Tv ratios showing different values for different genomic regions in humans

In the variant filtration analysis step, six GATK recommended annotations were focused on that seem to be highly informative and robust (https://software.broadinstitute.org/gatk/documentation/article?id=11069). They are: QualByDepth (QD), FisherStrand (FS), StrandOddsRatio (SOR), RMSMappingQuality (MQ), MappingQualityRankSumTest (MQRankSum) and ReadPosRankSumTest (ReadPosRankSum). These annotations are briefly explained below:

1) QD is the QUAL score (phred-scaled quality score or the variant confidence score) divided by the allele depth of non-homozygous reference samples. This is used instead of the QUAL score because a locus with deep coverage will have an inflated QUAL score.

2) FS is the phred scaled probability that a locus has strand bias. In paired-end sequencing, a real variant should be present in both forward and reverse reads. Strand bias is a type of sequencing bias in which one DNA strand is favored over the other. The higher the FS value, the more likely there is to be a bias.

3) MQ is the root mean square mapping quality over all the reads at the site.

4) SOR is another metric to evaluate strand bias in the data but is considered as an updated form of FS which performs better in high coverage situations. A higher value indicates increased strand bias.

5) ReadPosRankSum is a score from a rank sum test which is calculated by comparing the positions of reference or alternate alleles within the reads. If an allele is generally only present at the end of the read, it is more likely to be an error. A negative value indicates the excess presence of alternate alleles at the end of the reads.

6) The final metric is MQRankSum which is a score that is calculated from a rank sum test by comparing the mapping qualities of the reads supporting the reference allele with those supporting the alternate allele. A positive value suggests that the alternate allele read qualities are greater than those of reference allele reads ones whereas a negative value suggests that the mapping qualities of the reference allele reads are greater than of the alternate alleles.

As we are only interested in biallelic sites for ASE studies, the final multisample VCF file was filtered using BCFtools (Li, 2011) 'view' v1.6 with the parameters -v snps, -m2 and -M2. The biallelic SNVs were then filtered to remove possible false positives. In order to get optimal threshold values for the above mentioned metrics, plots were made between the Ti/Tv ratio and various changing values of the six metrics. The principle is that there will be a point in the graph where the Ti/Tv ratio will not change as the value of the metric changes as false positives have largely been removed. That point is the threshold for the particular metric that was chosen for variant filtration. Based on the selected thresholds for each metric, BCFtools 'filter' v1.6 was used with parameters -i to include variants where QD >= 15, FS <= 60, SOR <= 2, MQ >= 50, MQRankSum >= -2.5 and ReadPosRankSum >= -2.5 to get the filtered biallelic multisample VCF file.

3.2.5 RNA-seq raw data description

3 to 4 macrophage samples were sequenced per animal from two male and two female Mediterranean adult buffalo (Paired end-sequencing, using the Illumina HiSeq 2500): bone marrow derived macrophages (BMDMs), BMDMs at 7 hours post LPS stimulation, alveolar macrophages and monocyte derived macrophages (MDMs). BMDM +/- LPS total RNA was sequenced at a depth of

100 million reads, whereas AM and MDM mRNA was sequenced at a depth of 25 million reads. In total, there were 14 samples. The sample summary is shown in Table 10.

Animals	Naming in Buffalo expression atlas	Scottish abattoirs where samples were collected from	Samples
Male 1	M2	Grantown	MDM, BMDM +/- LPS
Female 1	F2	Grantown	MDM, BMDM +/- LPS
Male 2	M3	Kirkcaldy	AM, MDM, BMDM +/- LPS
Female 2	F3	Kirkcaldy	AM, MDM, BMDM +/- LPS

Table 10: RNA-seq raw data summary. 14 RNA-seq samples consisting of bone marrow derived macrophages +/- LPS (LPS treated and untreated samples) were sequenced at a depth of 100 million reads. The alveolar and monocyte derived macrophages were sequenced at a depth of 25 million reads.

3.2.6 RNA-seq read alignment

The alignment was performed by the method present in (Clark, et al., 2017). Briefly, raw reads from 14 RNA-seq samples were screened with FASTQC v0.11.2 (Andrews, 2014) and then cleaned using Trimmomatic v0.36 (Bolger, et al., 2014) with parameters 'TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:100.' This will remove low quality bases (below quality 20) from the trailing end of the reads. Furthermore, it will scan the read with a 4-base wide sliding window, and cut the read when the average quality per base drops below 20. It will also drop reads if the read length goes below 100 after trimming. Cleaned reads were then aligned to the new water buffalo reference genome (Low, et al., 2019) using HISAT2 v2.0.4 (Kim, et al., 2015) with its default parameters plus -dta (optimised for downstream transcriptome assembly). The resulting BAM files were filtered with SAMtools v1.4 (Li, et al., 2009), using parameters –F 256 (which removes non-primary alignments) and –F 12 (which removes pairs of unmapped reads) to create a subset. Singleton reads (mapped reads with unmapped mates) were also filtered using the parameters –F 4 –f 8 to create another subset from the main BAM files. The two subsets were cleaned using Picard 'CleanSam' v2.5.0, which soft-clips beyond-end-of-reference alignments and sets mapping quality to 0 for unmapped reads. These were then merged using Picard 'MergeSamFiles' v2.5.0 to create one BAM file per sample.

3.2.7 Obtaining RNA-seq reads corresponding to biallelic heterozygous sites obtained from genomic (DNA-seq) data

As RNA-seq data was only present for 4 Mediterranean water buffalo, GATK SelectVariants v4.0.4.0 was used to extract the 4 samples corresponding to the 4 animals from the biallelic filtered multisample VCF file obtained from calling variants using the DNA-seq data and to create 4 single-sample VCF files. BCFtools view v1.6 was used to keep only biallelic heterozygous sites which had a genotype quality greater than or equal to a Phred score of 40 using the parameters -g het and GQ >= 40. Since some of the samples belonged to mRNA and others total RNA, only exonic variants were retained in the ASE study. In this study, intronic variants were not used from the total RNA-seq samples (as done in the previous chapter) to keep the study more uniform across all samples. The exonic coordinates of all the genes present in the water buffalo genome gene annotation file were extracted using GTF_extract (https://github.com/fls-bioinformatics-core/GFFUtils/blob/master/docs/GTF_extract.rst). The resulting file was coordinate corrected using a custom R script to make it a 0-based coordinate system bed file. BEDtools 'intersect' v2.26.0 with the -u parameter was used to retain only high quality heterozygous exonic variants. The 'intersect' tool tries to find overlapping coordinates between two files (in this case a VCF file and the bed file consisting of only exon information) and the -u parameter ensures that no variant entry is written more than once in the output VCF file because overlapping exons of different transcripts of the same gene would otherwise lead to the reporting of the same variant multiple times.

After obtaining high quality heterozygous exonic variants from the genomic data, RNA-seq read counts over the reference and alternate alleles were obtained using a python-based read counter program called allelecounter v0.6 (https://github.com/secastel/allelecounter) (Castel, et al., 2015) with parameters --min_cov 4, --min_baseq 20 and --min_mapq 60 and --max_depth 100000. The tool's method of working has been described in the previous chapter. The ASE

analysis program ‘MBASED’ (Mayba, et al., 2014) required variants to be associated with gene names in the input file along with RNA-seq read counts of the reference and alternate alleles. To obtain the gene name associated with a variant, the VCF files were annotated using the program SnpEff v4.3 (Cingolani, et al., 2012) with parameters -no-intron, -no-downstream, -no-intergenic and -no-upstream so that the tool only reports the exon-related putative effects (for example, missense or synonymous or 5’ and 3’ UTR variant, etc.). The gene name was extracted from the first annotation in case of multiple effects/consequences for a particular variant. Custom scripts were then used to generate the input files for MBASED.

3.2.8 Reference bias estimation and duplicate sample sequencing discovery

When sequencing reads are aligned to a linear haploid reference genome, reads carrying alternate (non-reference) alleles will have more mismatches in the alignment than reads which have the reference allele and consequently, will be less likely to be mapped correctly. This is known as a ‘reference bias’ or ‘read mapping bias’. In an ASE analysis, reference bias leads to an over estimation of the relative proportion of reads carrying the reference allele (Pai, et al., 2009; Satya, et al., 2012).

MBASED can incorporate pre-existing reference bias into its ASE estimates as a global reference bias assuming that the bias is constant across all SNVs within a sample (Mayba, et al., 2014). To estimate the extent of reference bias in the RNA-seq samples, the reference allele ratio (proportion of reads containing the reference allele) was calculated (i.e. ref reads/ref reads + alt reads) per exonic heterozygous SNV site per sample based on the ‘allelecounter’ program’s output. To visualize the extent of the reference bias, a density plot of the reference allele ratio was calculated for each sample and the median of the ratio per sample was noted (Figure 15). The sample specific reference bias medians were considered as the extent of reference bias present in each sample respectively. This was incorporated into the MBASED custom script to determine sample specific ASE.

The expected distribution of the proportion of reads containing the reference allele is broadly normal with skewness towards the right in case of the presence of reference bias (Pai, et al., 2009). Two of the samples (Female 1 BMDM and Female 1 BMDM7hrLPS) out of fourteen showed bimodal distributions with very high numbers of sites which had either only reference allele reads or alternate allele reads. To determine the underlying reason, variant calling was performed using only RNA-seq data using the two RNA-seq BAM files using BCFtools 'mpileup' v1.4 with parameters --max-depth 1000000, --min-MQ 60 (minimum mapping quality (MAPQ) Phred score), --min-BQ (minimum base quality Phred score) followed by BCFtools 'call' with parameter -m (allow multiallelic variants) and –v (variant only). The called variants (in the form of a VCF file) were filtered to get only good quality biallelic SNVs, with a minimum depth of 10 reads and genotype quality of 20 (1% error rate) using BCFtools 'view' v1.4 and the parameters -v snps, -m2 -M2, and -i 'DP >= 10 & FORMAT/GQ >= 20'.

The principle of this process is that the genotypes of the variant sites discovered using RNA-seq data should be similar (with some sequencing or genotype calling errors) to the genotypes of the sites discovered using the DNA-seq data of the same animals. Genotypes were extracted from the resulting VCFs using SnpSift's (Cingolani, et al., 2012) 'extract' command. Genotypes were also extracted from the VCFs of the four animals containing filtered biallelic variants discovered using their genomic data. BCFtools 'isec' v1.4 was used to intersect the RNA-seq generated VCF files with the DNA-seq generated VCF files to find common variants for the genotype comparison. A genotype concordance Pearson correlation value was calculated using a custom R script by matching the genotypes of common variants. A high Pearson correlation value of 0.98 was found between the genotypes of the biallelic SNV sites of these Female 1 RNA-seq samples and the genotypes of Male 1 discovered through DNA-seq. This suggested that the Female 1 BMDM and Female 1 BMDM7hrLPS RNA-seq samples are actually derived from Male 1. Accordingly the BAM files from Female 1 BMDM were merged with those of Male 1 BMDM and Female 1 BMDM7hrLPS was merged with Male 1 BMDM7hrLPS to get a greater depth for the Male 1 RNA-seq samples. The procedure of obtaining RNA-seq reads

corresponding to biallelic heterozygous sites obtained from DNA-seq was repeated for the new high depth samples and reference bias was also estimated by the above mentioned procedure. Therefore, the total number of RNA-seq samples reduced to 12.

3.2.9 ASE quantification using MBASED

The reference and alternate allele read counts from the 12 RNA-seq samples calculated using the program 'allelecounter' were processed using the method described in the previous chapter. The same ASE quantification method was used (as mentioned in Chapter 2, section 2.2.6), except for the incorporation of the reference bias value.

3.3 Results and Discussion

3.3.1 RNA-seq samples alignment results

Supplementary Table S 4 shows the alignment summary of the 12 macrophage-specific RNA-seq samples extracted from four Mediterranean water buffalo. The summary statistics were generated using the SAMtools flagstat v1.6 (Li, et al., 2009) program that utilised the pre-processed BAM files generated using HISAT2. On average, around 85% of properly-paired mapped and aligned reads were obtained for each sample.

3.3.2 DNA-seq alignment, variant calling and variant filtration

As shown in Supplementary Table S 5, the mean mapping percentage was 99% and 95% of reads were properly-paired. The mean duplicate percentage was 8% across all 81 samples. The final multisample VCF file contained a total of 43,243,663 variants that contained a mix of SNVs and indels. Out of this total, there were 36,541,726 biallelic SNVs. The remaining variants consisted of Indels and multiallelic sites. The variant calling summary statistics are shown in Table 11.

As mentioned in the methods section, the Ti/Tv ratio of biallelic SNVs was used to derive a value that can be used as an appropriate cut-off for removing false positive variants. Various values of the chosen metrics or variant annotations

were chosen and the biallelic SNVs were filtered based on those values and their Ti/Tv ratio. A graph was then plotted based on each metric as shown in Figure 14. The aim was to maximise the Ti/Tv ratio whilst minimising the number of variants removed.

Statistic	Original base data	Biallelic SNVs	Filtered biallelic SNVs
Number of records	43,243,663	36,541,726	26,247,559
Number of SNVs	37,682,631	36,541,726	26,247,559
Number of indels	5,897,230	0	0
Number of multiallelic sites	2,296,557	0	0
Number of multiallelic SNV sites	785,862	0	0
Ti/Tv	1.93	2.01	2.29

Table 11: Multisample variant calling statistics showing variant information from original multisample VCF file, number of biallelic SNVs and number of filtered biallelic SNVs along with Ti/Tv values

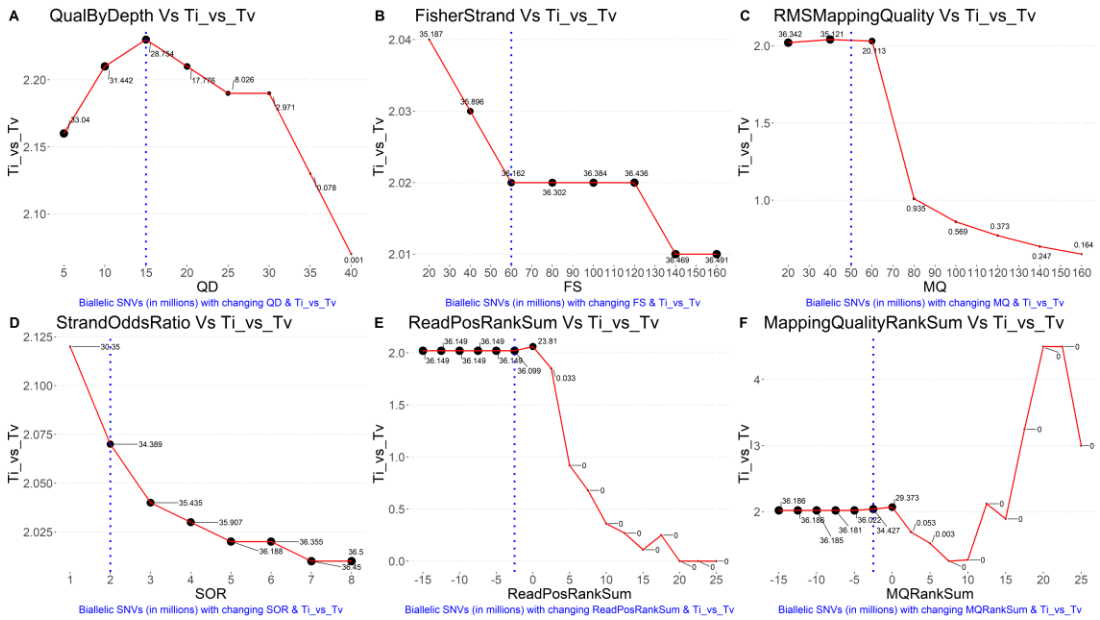

Figure 14: Graphs showing the relationship between changing variant annotations and Ti_vs_Tv ratio. (A) Biallelic SNVs (in millions) with changing QualByDepth and Ti_vs_Tv ratio, (B) Biallelic SNVs (in millions) with changing FisherStrand and Ti_vs_Tv ratio, (C) Biallelic SNVs (in millions) with changing RMSMappingQuality and Ti_vs_Tv ratio, (D) Biallelic SNVs (in millions) with changing StrandOddsRatio and Ti_vs_Tv ratio, (E) Biallelic SNVs (in millions) with changing ReadPosRankSum and Ti_vs_Tv ratio, (F) Biallelic SNVs (in millions) with changing MQRankSum and Ti_vs_Tv ratio. The size of points in each graph denotes the number of SNVs (in millions). The number of biallelic

SNVs have been rounded up to three digits after decimal in the plots which resulted in a few points being rounded off to 0.

The variant filtration parameters and thresholds were chosen keeping in mind the GATK hard-filtering recommendations that were provided in place of VQSR (https://software.broadinstitute.org/gatk/documentation/article?id=11069). They derived on their chosen values based on the overlap of proportion of variants that passed or failed VQSR. Based on our analysis, it is observed that In Figure 14A, QD >=15 was the optimal threshold for the cut-off with the highest Ti/Tv ratio. In Figure 14B, from FS=60 to 120, the Ti/Tv value did not change. As the FS value was lowered, there was an increase of the Ti/Tv. Accordingly, the GATK recommended value, i.e., 60 was chosen. In Figure 14C, At MQ = 60, the Ti/Tv ratio attains a plateau. However, at MQ=40, the Ti/Tv value only slightly increased but there was a large increase in number of SNVs. Hence, MQ >= 50 was chosen as a filtering cut-off in this case. In Figure 14D, SOR <=2 was chosen as an appropriate cut-off value for StrandOddsRatio. From Figure 14E, ReadPosRankSum >= -2.5 was chosen as a threshold. Based upon data in Figure 14F MQRankSum >= -2.5 was chosen as the filtering threshold. Variants with values greater than or equal to -2.5 for ReadPosRankSum and MQRankSum were retained. It should be noted that in some cases, for example, in Figure 14E, since we have put up a threshold value where we retained variants with ReadPosRankSum >= -2.5, we have also retained variants whose Ti/Tv ratio goes as low to 0. It would have been better if an additional filter was applied by choosing an upper limit for ReadPosRankSum value and lose those variants which contribute in lowering the Ti/Tv ratio. It is same with MQRankSum as well. Putting an upper limit would not result to loosing too many variants. The process of choosing an optimal hard-filtering threshold is always arbitrary in nature and here we have tried to logically determine the best cut-off for our data.

Hard filtering is clearly empirical, but the chosen thresholds are conservative and more likely to remove true positives than to include false positives. After applying the above filters, 26,247,559 biallelic filtered SNVs were left for further downstream analysis (Table 11).

3.3.3 Pre-processing of SNVs for ASE analysis

The ASE analysis utilises the reads obtained from RNA-seq samples belonging to four Mediterranean water buffalo that mapped to reference or alternate allele from heterozygous biallelic exonic sites. In order to identify those sites from the four water buffalo samples, sample-specific SNVs were obtained from the base filtered multisample VCF file obtained above. Each sample-specific VCF file was pre-processed to obtain high quality biallelic heterozygous exonic SNVs. Since the samples were extracted from a multisample VCF file, some variant positions will have a homozygous reference genotype with respect to the reference genome. This is because there may be other samples with genotypes with an alternate allele because of which the particular locus was called as a variant with respect to the reference genome. Hence, the counts of number of biallelic SNVs per sample was obtained using BCFtools view v1.6 with parameters -i GT="alt" which means obtain sites with at least one alternate allele. Table 12 shows sample-specific number of biallelic SNVs, number of heterozygous biallelic SNVs that have a phred scaled genotype quality or GQ>=40 (99.99% accuracy or 0.01% error rate), number of heterozygous biallelic SNVs present in exonic regions and number of genes associated with those exonic SNVs. The high quality heterozygous biallelic exonic sites were used to obtain reads from RNA-seq samples that mapped onto those sites. Table 13 shows the number of genes associated with the SNVs onto which the RNA-seq reads could map. The RNA-seq sample-specific genes in Table 13 are less than the number of genes with heterozygous biallelic exonic SNVs at GQ >=40 in Table 12 because SNVs were only selected for ASE analysis where the number of RNA-seq reads was at least 10 carrying either the reference or alternate allele.

Sample	Number of Biallelic SNVs	Number of Heterozygous Biallelic SNVs at GQ >=40	Number of Heterozygous Biallelic exonic SNVs at GQ >=40	Number of genes with Heterozygous Biallelic exonic SNVs at GQ >=40
Male 1	8,723,607	5,750,866	133,050	16,974
Male 2	8,373,498	5,590,632	125,959	16,283
Female 1	9,468,361.00	6,513,025	144,762	18,957
Female 2	8,842,271	5,881,652	132,249	17,261

Table 12: Sample-wise variant statistics of 4 Mediterranean water buffalo samples calculated after extracting their data from the base filtered multisample VCF file.

Sample	MDM	AM	BMDM	BMDM7hrLPS
Male 1	4,209	-	4,519	4,200
Male 2	4,183	3,700	5,602	5,131
Female 1	5,352	-	-	-
Female 2	4,663	4,130	5,634	5,656

Table 13: Number of genes with heterozygous biallelic exonic SNVs at GQ >=40 onto which RNA-seq reads could map. Four cells are empty due to absence of those RNA-seq samples. BMDM is bone marrow derived macrophages, BMDM7hrLPS is BMDM sample at 7 hrs post LPS treatment, AM is alveolar macrophages and MDM is monocyte derived macrophages

3.3.4 Reference bias estimation

All of the 12 RNA-seq samples showed some evidence of reference bias (Figure 15). The reference allele ratio or the proportion of reads containing the reference allele was calculated (i.e. ref reads/ref reads + alt reads) per exonic heterozygous SNV site for all samples. In MBASED, the pre-existing allelic bias is incorporated during ASE calculation as a global reference bias (which is the ratio of all reference reads to total reads in the sample across all heterozygous loci) assuming that the reference bias is constant across all the SNVs in a given sample (Mayba, et al., 2014). In the current study, the sample-specific global reference bias was assumed to be the median of the reference ratio. In MBASED, loci with the top 5% of read counts were removed while calculating the global reference bias (Mayba, et al., 2014). However, the median of the reference ratio calculated per site for each sample was used, to retain all sites but minimise the impact of outliers.

The plots also show the existence of monoallelic expression or MAE in the tails of the read distributions where the reference ratio is 1 or 0. In Figure 15, a reference ratio of 1 means that at a heterozygous SNV site, there is only expression of reads containing the reference allele and 0 means that there is only expression of reads carrying the alternate allele. The relative enrichment of sites only expressing the reference allele is potentially partly due to the presence of genotyping errors which may have remained even though a strict genotype quality filter (GQ>=40) and a strict SNV filtering criteria was used. It should be also noted that since males have only one X chromosome, all loci outside the pseudoautosomal regions (PAR) will be monoallelically expressed on the male X chromosome. As shown in Figure 15, the male RNA-seq samples have more SNVs with a reference ratio of 1, as expected. Since, genomic variant calling was done with ploidy=2 (diploid) for all samples (irrespective of male or female sex), many heterozygous calls on male X chromosome are incorrect.

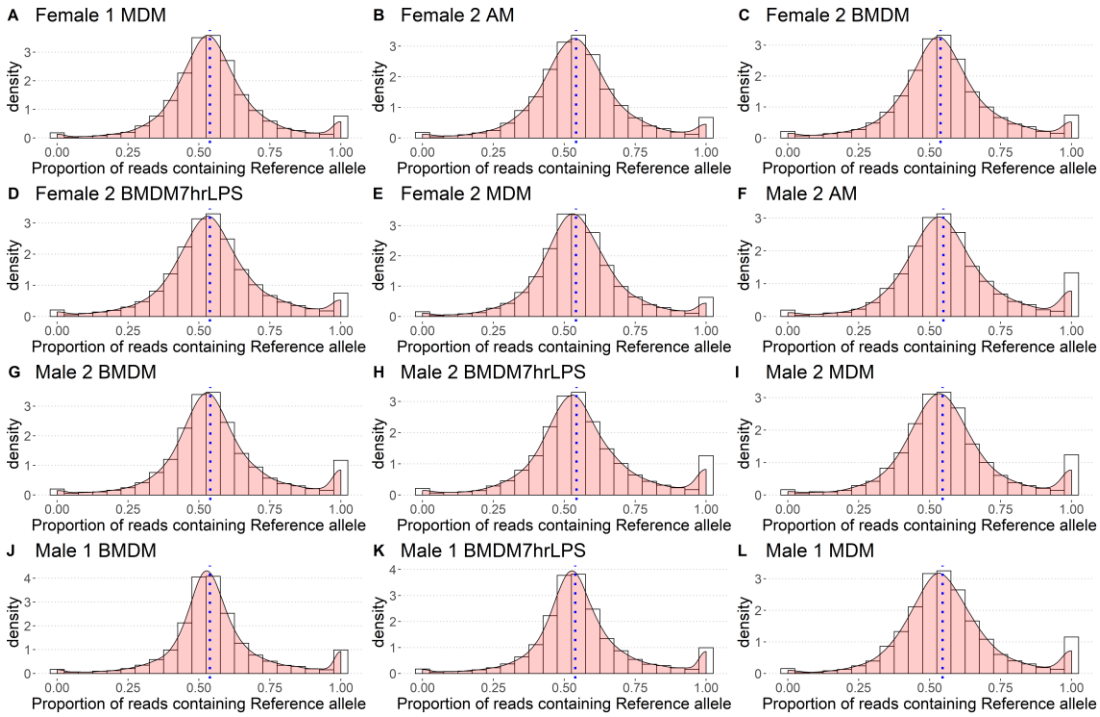

Figure 15: Genome-wide reference bias plots. The plots (A to L) show the distribution across heterozygous exonic SNVs of the proportion of RNA-seq reads containing the reference allele in 12 RNA-seq samples. The reference allele ratio (proportion of reads containing the reference allele) was calculated i.e. ref reads/ (ref reads + alt reads) per

exonic heterozygous SNV site for each of the samples. The plots show that there is a slight skew towards higher numbers of reads carrying the reference allele in all the samples revealing the presence of reference/read mapping bias. The blue dashed line represents the median value of the distribution. The sample-wise median values of the reference ratio calculated per site are- (A) 0.54 (B) 0.54 (C) 0.54 (D) 0.54 (E) 0.54 (F) 0.55 (G) 0.54 (H) 0.54 (I) 0.55 (J) 0.54 (K) 0.54 (L) 0.55. The samples J and K represent the merged samples wherein the Female 1 samples were merged with Male 1 samples as it was found that there was a library preparation error wherein Male 1 sample was re-sequenced but labelled as Female 1. BMDM is bone marrow derived macrophages, BMDM7hrLPS is BMDM sample at 7 hrs post LPS treatment, AM is alveolar macrophages and MDM is monocyte derived macrophages.

The level of MAE seen here was also observed in a tissue specific ASE analysis in cattle (Chamberlain, et al., 2015). In this study, MAE was defined to be present in SNVs when the frequency of the major allele was >0.9. It was also found that SNVs with reference ratio of 1 were more prevalent than SNVs with reference ratio of 0. MAE events that are related to parent-of-origin rather than *cis*-acting variation can only be confirmed with access to parental genomes and multiple independent crosses (Baran, et al., 2015). In a recent work on tissue based ASE profiling in sheep (Salavati, et al., 2019), loci with strict MAE were removed from their analysis. In the ASE results produced by MBASED, genes which have an estimated major allele frequency or MAF>0.9 were considered to be having MAE (Mayba, et al., 2014). Here, the analysis is based on the threshold followed by MBASED authors. The reasons are explained below. The analysis is limited to autosomal genes due to the complexity involved in dealing with sex chromosomes, as explained in the introduction.

3.3.5 ASE analysis on autosomal genes

Table 14 shows the number of autosomal genes that were considered for the ASE analysis.

Sample	MDM	AM	BMDM	BMDM7hrLPS
Male 1	4,187	-	4,495	4,179
Male 2	4,159	3,678	5,571	5,105
Female 1	5,239	-	-	-
Female 2	4,588	4,072	5,534	5,561

Table 14: Number of autosomal genes used for ASE analysis with heterozygous biallelic exonic SNVs at GQ >=40 onto which RNA-seq reads could map. Four cells are empty due to absence of those RNA-seq samples.

MBASED calculates the extent of deviation from the null hypothesis of equal allele expression (1:1 allelic ratio) for each gene present in a sample. It quantifies the allelic imbalance or ASE effect size in the form of major allele (haplotype) frequency (MAF) of the gene and also provides a corresponding P-value (known as pValueASE) associated with it. The pValueASE generated by MBASED is not FDR (false discovery rate) corrected. Hence, for each sample, the pValueASE have been Benjamini-Hochberg (Benjamini and Hochberg, 1995) or BH adjusted and genes with BH adjusted P-value <=0.05 (FDR of 5%) along with MAF>=0.7 are declared to be showing evidence of ASE. During software testing using real world data, the authors of MBASED encountered many genes which had high RNA-seq coverage and showed significant ASE. However, their allelic ratio was not very distinct from 1:1. Hence, MBASED takes a MAF threshold of 0.7 along with a BH P-value cut-off for assigning ASE status, which is roughly equivalent to a 2:1 ratio of the two alleles/haplotypes.

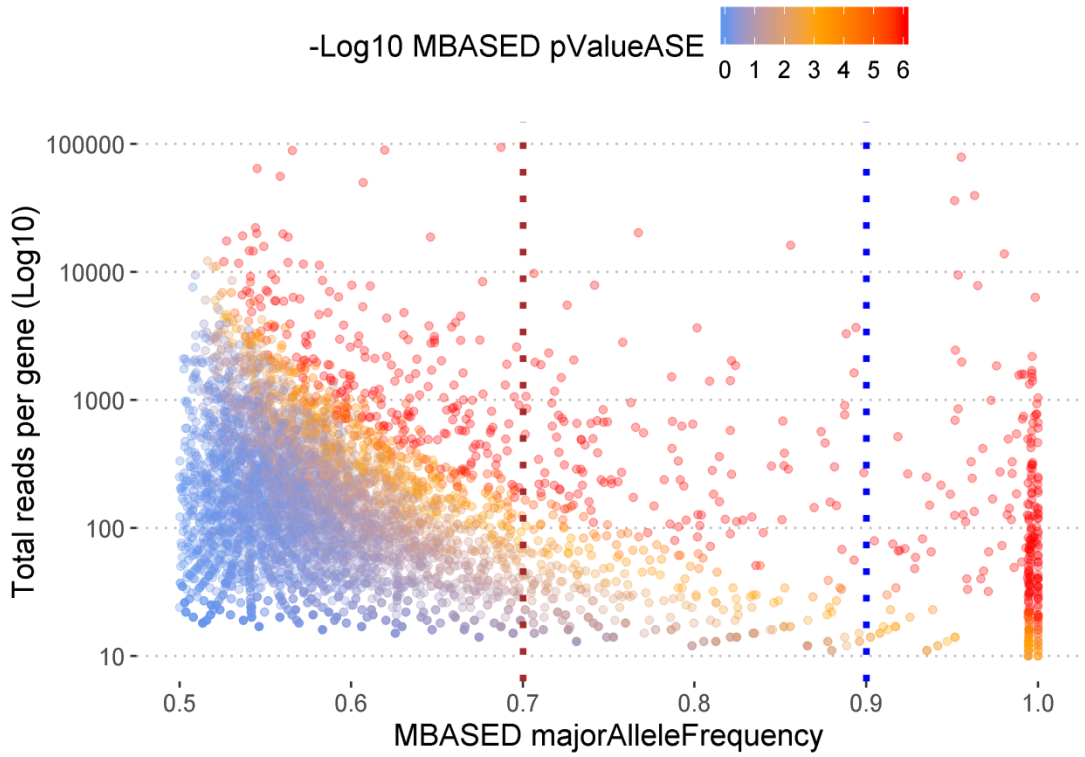

Figure 16: Relationship between MBASED calculated majorAlleleFrequency or MAF (ASE Effect size) and total number of reads per gene for Female 1 MDM sample. The y-axis is in Log10 scale to spread the data vertically for easy visualisation. Each point represents a gene and are coloured differently according to their MBASED pValueASE (FDR uncorrected). There are 5,239 genes in this plot. The pValueASE has been transformed into -Log10 scale for easy visualisation. The brown dashed line represents the MAF threshold of 0.7 above which genes have been considered to show ASE. The blue dashed line represents the MAF threshold of 0.9 above which genes have been categorised to show MAE.

Figure 16 shows the relationship between MAF and total reads per gene for one sample 'Female 1 MDM', which consisted of 5,239 genes that underwent MBASED analysis. The plot was made to see the pValueASE (FDR uncorrected) distribution among those genes. As MAF corresponds to the frequency of the common haplotype its value will always be >=0.5. Most of the genes have a high value of pValueASE (blue in colour). As mentioned before, there are many genes that have significantly low pValueASE and high depth as well, but the MAF is between 0.5 and 0.7. This can be observed from the yellow and red coloured dots in the plot whose MAF is less than 0.7. This phenomenon

was also observed by the MBASED authors (personal communication; plot not shown in the paper). The brown dashed line indicates the MAF cut-off (MAF>=0.7) applied to characterise genes potentially showing ASE. The blue dashed line indicates the cut-off (MAF >=0.9) considered to characterise genes to be showing MAE (the extreme form of ASE). Genes showing MAE are good candidates that may be tested for imprinting or the presence of candidate null expression alleles. Similar plots describing the relationship between MAF and total reads per gene for all 12 RNA-seq samples are present in Figure 17.

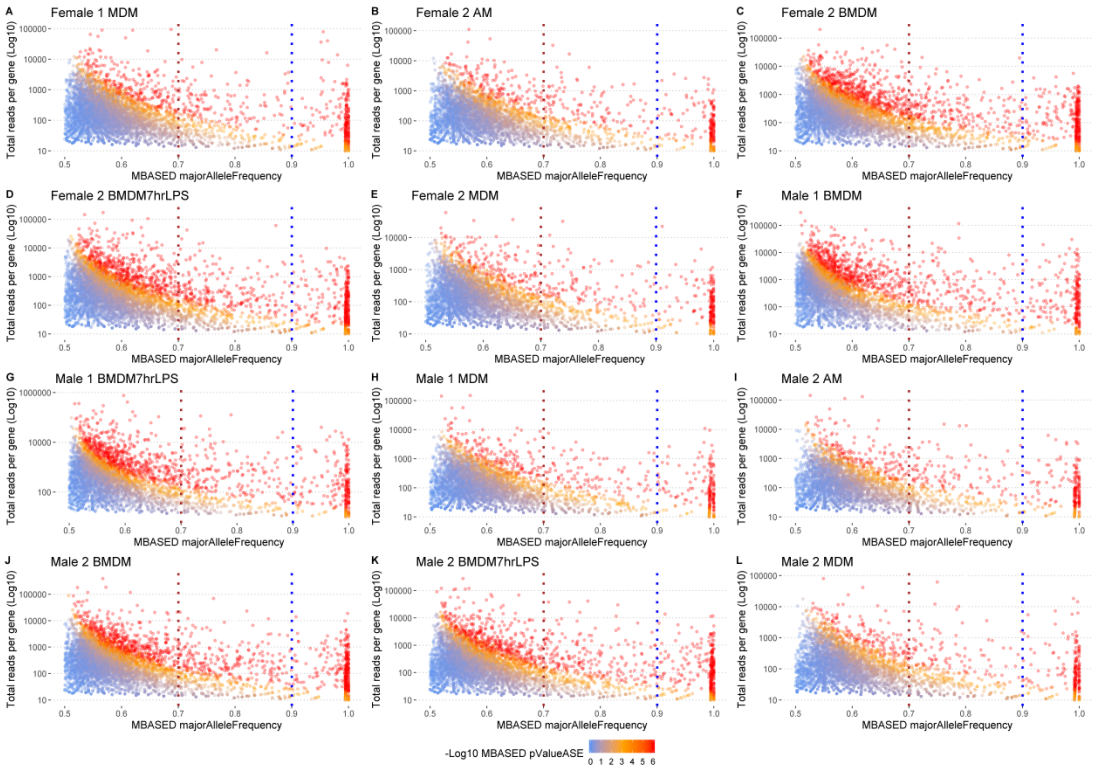

Figure 17: Relationship between MBASED calculated majorAlleleFrequency or MAF (ASE Effect size) and total number of reads per gene for all 12 RNA-seq samples under study

Sample	Genes tested	ASE genes (MAF>=0.7 and BH adj pValueASE<=0.05)	ASE genes %	MAE genes (MAF>=0.9)	MAE genes %
Female 1 MDM	5,239	563	11%	291	6%
Female 2 AM	4,072	472	12%	215	5%
Female 2 BMDM	5,534	848	15%	378	7%
Female 2 BMDM7hrLPS	5,561	847	15%	376	7%
Female 2 MDM	4,588	474	10%	222	5%
Male 1 BMDM	4,495	561	12%	255	6%
Male 1 BMDM7hrLPS	4,179	548	13%	240	6%
Male 1 MDM	4,187	446	11%	213	5%
Male 2 AM	3,678	454	12%	228	6%
Male 2 BMDM	5,571	787	14%	361	6%
Male 2 BMDM7hrLPS	5,105	762	15%	357	7%
Male 2 MDM	4,159	499	12%	245	6%
Average	4,697	605	13%	282	6%
Total	10,624	3,278	31%	1,205	11%

Table 15: MBASED results of the number of autosomal genes showing ASE and MAE. It shows number of genes tested, number of genes showing ASE and its percentage, number of genes showing MAE and its percentage averaged across all samples. It also shows the total number of genes tested for ASE that were present in at least one sample, genes showing ASE in at least one sample and its percentage and genes showing MAE in at least one sample and its percentage.

Table 15 shows the cumulative result of the MBASED analysis across 12 RNA-seq samples under study. A total of 10,624 genes were tested for ASE that was present at least in one sample. 3,278 genes (31% of all genes tested) showed ASE in at least one sample. 1,205 genes (11% of all genes tested) showed MAE in at least one sample. In the cattle ASE validation study involving 20 samples from each of white blood cells (WBC) and liver, an average of 13–29% of 3,531 genes tested in WBC showed significant ASE (Chamberlain, et al., 2015). This broadly matches the ASE pattern in water buffalo in Table 15.

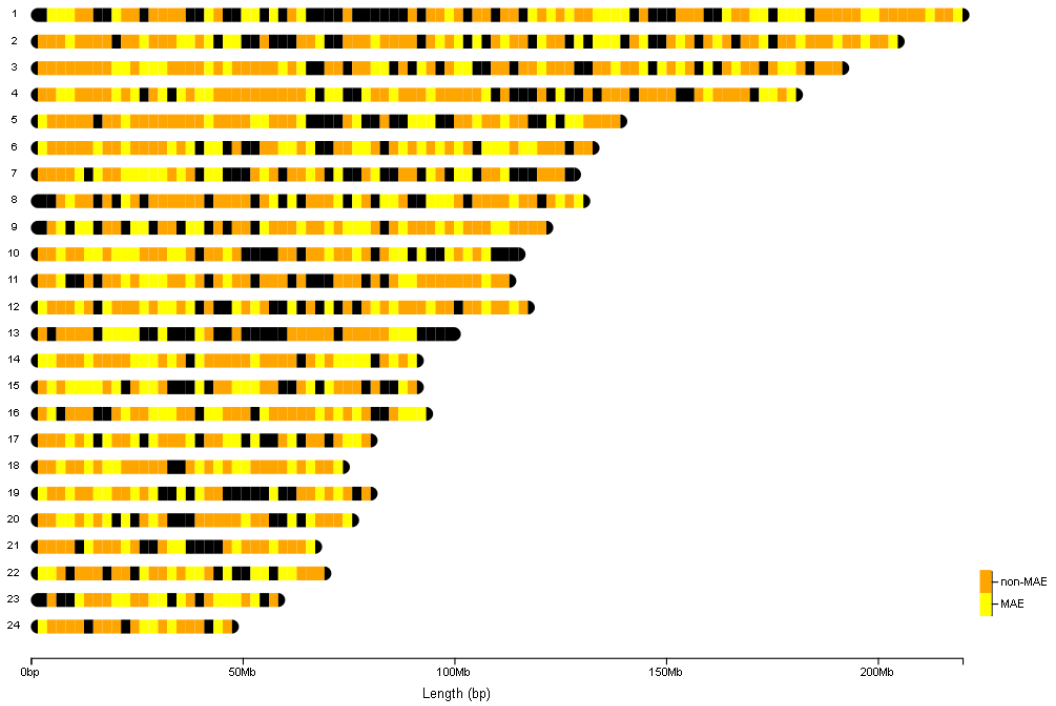

Figure 18: Widespread ASE on water buffalo autosomes. The plot shows the distribution of ASE genes along the 24 autosomes. The ASE genes have been divided into MAE and non-MAE genes. The black bands on the autosomes denote areas where no genes show ASE. The yellow bands denote areas where genes showing MAE exist and the orange bands show areas where non-MAE genes are present. The genes represented in this plot show ASE in at least one sample.

In order to visualize ASE hotspots (genomic regions with a high density of ASE genes), all ASE genes (including MAE genes) were plotted along water buffalo autosomal chromosomes using the R package chromoMap (Anand, 2019) (Figure 18). There is no obvious concentration of genes showing MAE. A similar pattern was observed on human autosomes (Gimelbrant, et al., 2007).The scattered presence of SNPs displaying allele-specific gene expression across all human chromosomes has been observed previously as well (Palacios, et al., 2009; Schroder, et al., 2004).

3.3.6 Intersectional analysis between RNA-seq samples

An intersectional analysis plot gives a snapshot of the number of ASE genes that are common to all RNA-seq samples, unique to a single sample and shared between different samples. Absence of genes showing ASE in a sample may be

due to the absence of informative SNVs (biallelic heterozygous SNVs) for that gene in the sample, making it inaccessible for ASE analysis.

Figure 19 depicts an UpSet plot (Lex, et al., 2014) of the ASE genes (with a MAF>=0.7 and BH adj pValueASE <=0.05) present in six BMDM samples. UpSet plots were made using the UpSetR package (Conway, et al., 2017) in R. Within BMDM samples, there are 2,391 unique genes showing 4,353 instances of ASE. The numbers above the vertical bars represent the distribution of the 2,391 genes in different intersections.

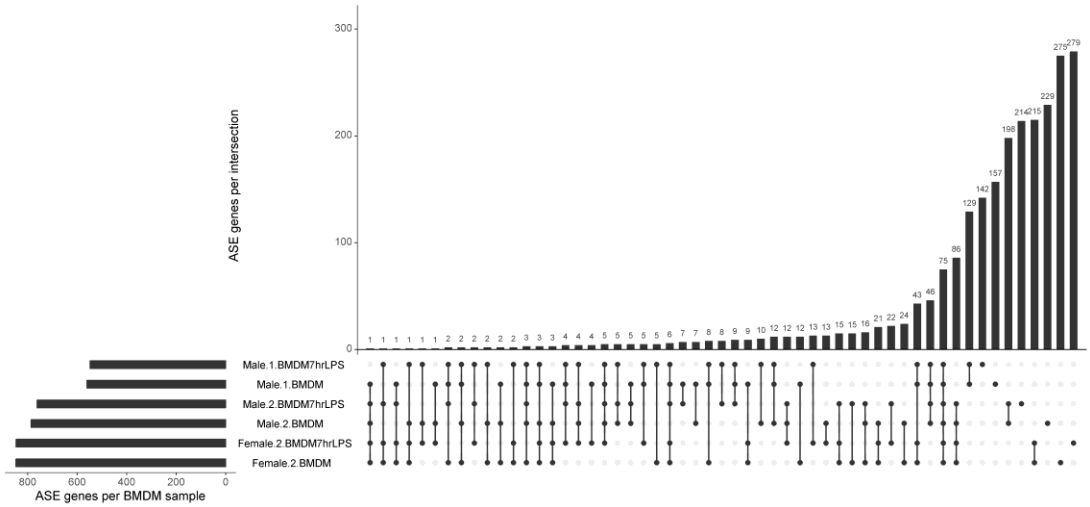

Figure 19: UpSet plot showing intersection among BMDM samples. The goal of the UpSet plot is to see the intersections between the BMDM samples of genes showing ASE. The vertical black bars represent the number of elements (ASE genes) an intersection contains. The numbers on the top of the vertical bars represent the cardinality (number of ASE genes in each intersection). The dots represent the sets that contribute to an intersection. The horizontal bars to the left represent the total number of ASE genes in each sample.

Most of the ASE genes are sample-specific (extreme right) and require replication for validation. Individual-specific ASE can also be observed from the plot. 129 genes in Male 1, 198 genes in Male 2 and 215 genes in Female 2 show individual-specific ASE. In these sets, the expression bias is replicated in separate RNA-seq libraries (for example Male 1 BMDM and Male 1 BMDM7hrLPS) but it is absent in the other animal sets. 75 genes (present in the six-way intersection) show ASE in all the samples. Out of these 75 genes, only

20 genes had an official gene symbol. A functional enrichment analysis using Database for Annotation, Visualization and Integrated Discovery (DAVID) did not reveal any significant biological process, molecular function or cellular compartment that was enriched among this small subset of genes. However, it helped to associate each gene to its function. One gene was *STEAP3*. Analysis of mouse BMDM +/- LPS samples revealed the importance of the *STEAP3* gene in regulating iron homeostasis and the *TLR4*-mediated inflammatory response in macrophages (Zhang, et al., 2012). The extent of ASE for the 20 genes which showed ASE in all BMDM +/- LPS samples is summarised in Supplementary Table S 6. The MAF values have been plotted in Figure 20. Most of the genes showed similar values of MAF in all samples. However, *RRP12*, *CHN1* and *CCNI* showed extreme fluctuations. For example, *RRP12* has MAF near 0.7 in both Female BMDM samples. However, in all Male BMDM samples, the MAF is 1. A similar trend was observed for the gene *CCNI* wherein MAF is exactly 1 for both Male 1 BMDM +/- LPS samples, but lower in remaining four BMDM samples. *CHN1* gene's MAF value is 0.74 for Female 1 BMDM which increases to 0.93 for Female 1 BMDM7hrLPS. These examples suggest that ASE profile differs between individuals and may be condition dependent. Condition dependent ASE can be correctly measured using a different protocol where differential ASE is detected between paired samples from the same individual and is currently, beyond the scope of this chapter. They have been developed in MBASED itself and in GeneiASE (Edsgärd, et al., 2016).

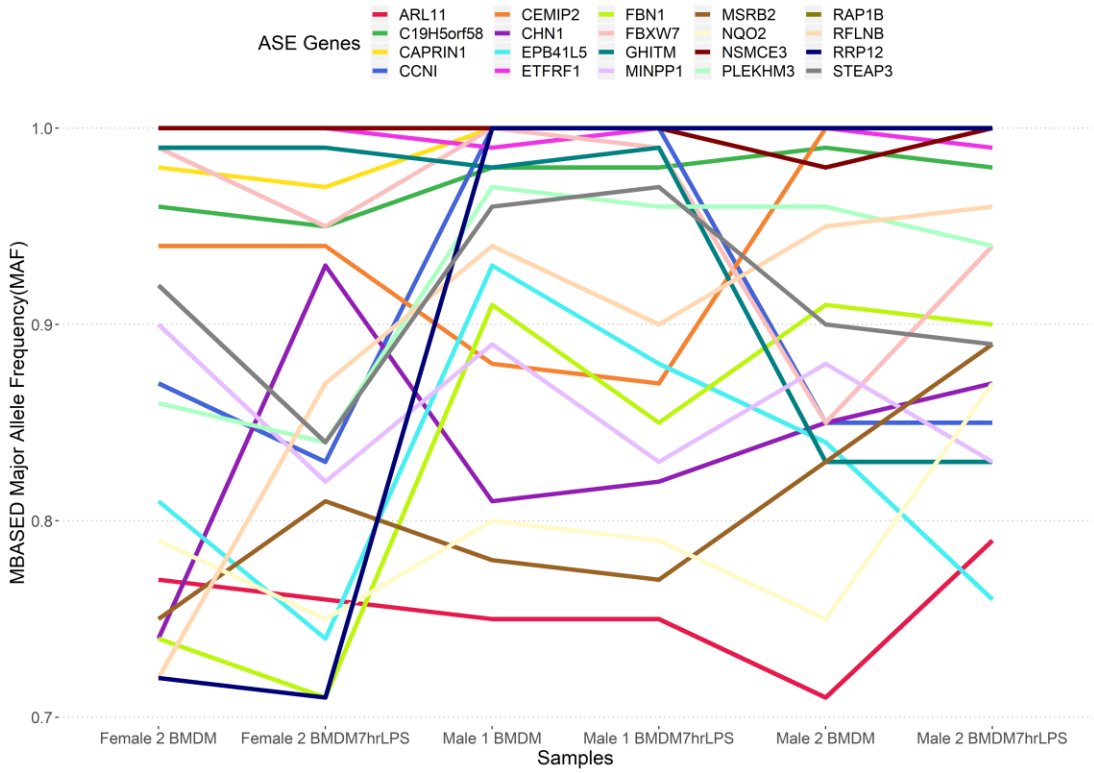

Figure 20: Graphical representation of MBASED Major Allele Frequency (MAF) values for 20 genes with official gene symbol across BMDM +/- LPS samples. The plot shows that the genes have almost similar MAF values across different samples (with slight fluctuations) except few exceptions like *RRP12*, *CHN1* and *CCNI*.

There were two genes that showed ASE across BMDM control samples (Male 1 BMDM, Male 2 BMDM and Female 2 BMDM) - *IL17B* and *LOC102415270*. *IL17B* induces cytokine production (TNF-α and IL-1β) and also takes part in neutrophil recruitment (Bie, et al., 2017). The other gene is unannotated. Two genes - *LOC112585628* and *LOC102401561* - showed ASE in all three BMDM LPS 7hrs samples. *LOC112585628* is a guanylate-binding protein 7-like. Guanylate-binding proteins are cytokine (IFN-γ, TNF-α and IL-1β) induced proteins from the family of GTPases (guanosine triphosphatase) (Tripal, et al., 2007). They are involved in the anti-microbial host defence system. However, the exact mechanism is unknown (Vestal and Jeyaratnam, 2011). *LOC102401561* encodes a zinc finger protein 790-like (or *ZNF790*-like). According to the KEGG database (Kanehisa and Goto, 2000), the gene is involved in the Herpes simplex virus 1 infection pathway.

The plot also highlights genes that showed ASE specifically when BMDM samples were induced by LPS (Male 2 BMDM7hrLPS, Male 1 BMDM7hrLPS, Female 2 BMDM7hrLPS). This set includes inducible genes that likely achieve sufficient sequence depth for inclusion in ASE analysis only in the induced state.

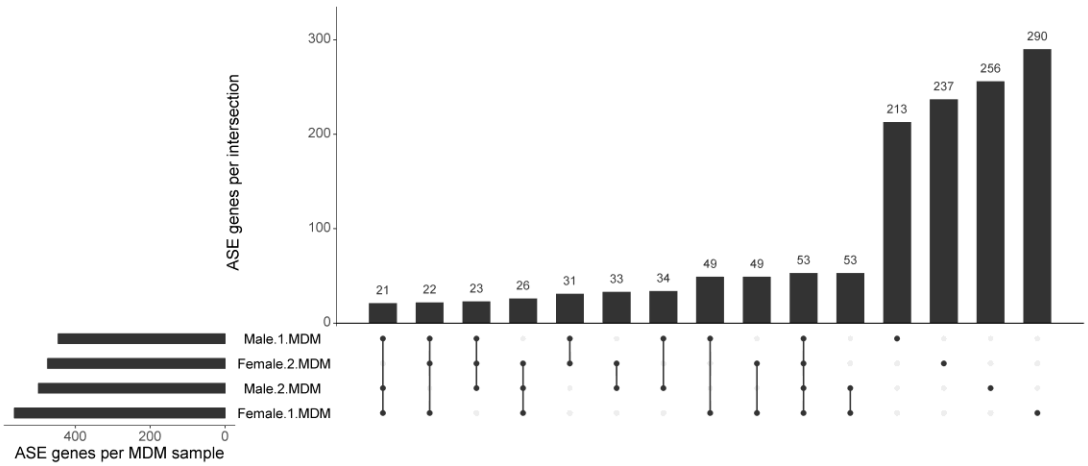

Figure 21: UpSet plot among MDM samples

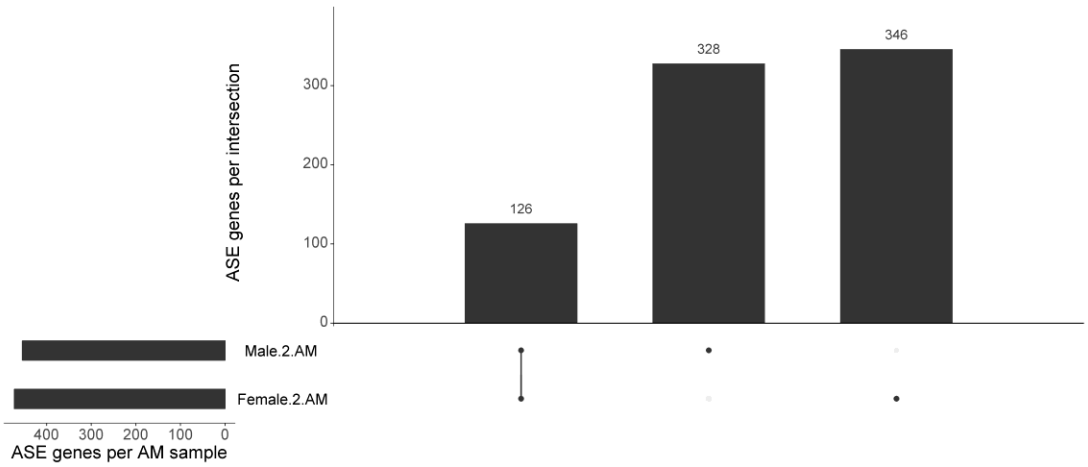

Figure 22: UpSet plot among AM samples

As seen in Figure 21, within the four MDM samples, 1,390 unique genes showed 1,982 instances of ASE. 53 genes were common to all individuals. In this case, a large number of genes showed individual or sample specific ASE. Amongst the 53 genes that show ASE in all MDM samples, 10 genes had an official gene symbol- *STEAP3*, *PLEKHM3*, *CEMIP2*, *MGST1*, *ETFRF1*, *GHITM*, *NASP*, *CCNI*, *ARL11* and *CAPRIN1*.

Within the two AM samples (Figure 22), 800 unique genes showed 926 instances of ASE. 126 genes showed ASE in both AM samples, of which 45 had an official gene symbol.

There were 33 genes that showed significant ASE among all 12 samples .i.e. all BMDM +/- LPS, MDM and AM samples, 6 with an official gene symbol: *ARL11*, *CAPRIN1*, *CCNI*, *CEMIP2*, *GHITM* and *PLEKHM3*. These genes are neither condition-specific nor individual specific. *ARL11* has been previously reported to be endogenously expressed in macrophages and helps in its activation and cytokine production upon any pathogenic infection (Arya, et al., 2018). *CAPRIN1* (Cell Cycle Associated Protein 1) is a gene that is involved in the cell cycle process and is essential for cellular proliferation (Wang, et al., 2005). *CCNI* or cyclin-1 is involved in cell-cycle progression (Nagano, et al., 2013). *GHITM* or *MICS1* is an important protein for cellular survival as it has been reported to interfere with the apoptotic process and promote cell growth (Oka, et al., 2008). *PLEKHM3* or *DAPR* (differentiation-associated protein) takes part in the phosphatidylinositol 3-kinase (PI3K) signalling pathway that is required in tissue development and cell cycle progression (Virtanen, et al., 2009). *CEMIP2* or *TMEM2* encodes a cell surface protein called hyaluronidase that is needed for the degradation of hyaluronic acid, an important extracellular high molecular weight glycosaminoglycan that permits easy movement of immune cells to the site of tissue repair. During an inflammatory immune response, high level of hyaluronic acid has been reported to slow down the response and its degradation is necessary in order to bring back homeostasis. Alveolar macrophages have been reported to carry out this process in an inflamed lung with the help of *TMEM2* (Johnson, et al., 2018).

3.3.7 Monoallelic expression of autosomal genes

Based on our MBASED analysis, MAE (genes whose MAF>=0.9) was observed in 1,205 autosomal genes out of 10,624 genes tested (i.e.11.3%) in at least one of the 12 samples under study. Based on an ASE study on cattle involving multiple tissues, 4,298 of 7,985 genes tested (54%) showed MAE (MAF>0.9 in this case) in at least one tissue (Chamberlain, et al., 2015). The authors

considered true MAE to be present in 28% of the genes when they only considered genes that had more than one heterozygous SNP with ASE. The remaining may have been present due to errors in the cattle genome sequence, but the true estimate may be around 28-54%. However, a direct comparison of the extent of MAE with our study is difficult because the cattle study involved 18 tissues from each animal compared to 3 samples from 4 water buffalo in our case. Furthermore, they had also included X chromosome in their study.

Out of 1,205 autosomal genes that showed MAE, 518 genes had an annotated gene name. A DAVID analysis of those 518 genes did not reveal any pathway that was significantly enriched in the dataset (BH FDR p-value <=0.05). Interleukins such as *IL34*, *IL36B* and *IL36G* were present along with *TLR10*, *CD84* (Signaling Lymphocytic Activation Molecule 5), *NLRP1* (NLR family pyrin domain containing 1), *NLRC4* (NLR family CARD domain containing 4), etc. Under nominal significance, many of the genes encoded for membrane bound proteins nominally enriched with the cellular component GO term 'integral component of membrane' and 'plasma membrane'. In terms of molecular function, some genes were involved in 'carbohydrate binding', 'glutathione peroxidase activity', etc. such as *KLRG2* (killer cell lectin like receptor G2), *LY75* (lymphocyte antigen) or *CD205*, and *MRC2* (mannose receptor C type 2). CD205 is a receptor for ligands expressed during apoptosis and hence, takes part in the phagocytosis of apoptotic cells (Shrimpton, et al., 2009). MRC2 is a pattern recognition receptor (PRR) present on macrophages, and belongs to the mannose receptor family. Genes from the mannose receptor family are involved in macrophage activation and also mediate endocytosis and phagocytosis when they recognise and bind to pathogen-associated molecular patterns or PAMPs (for example, glycoprotein such as mannose) (Stahl and Ezekowitz, 1998). There are also genes that encode proteins containing leucine-rich repeats or LRR (Ng and Xavier, 2011) such as *FLRT2* (F-box and leucine rich repeat protein 20), *LRRC2* (leucine rich repeat containing 2), *LRRC8A* (leucine rich repeat containing 8 family member A), *FBXL20* (F-box and leucine rich repeat protein 20), *NLRP1*, *NLRC4*, *NRROS* (negative regulator of reactive oxygen species) or *LRRC33* and *TLR10*. LRR proteins, through their LRR receptor,

sense various pathogens such as bacteria, fungi, parasites and viruses. They recognise PAMPs present on the microbes (for example, lipopolysaccharides or LPS on bacteria) in order to generate an immune response. This analysis indicates many of the genes showing MAE are transmembrane proteins or genes encoding cell surface markers, an observation noticed in humans as well (Gimelbrant, et al., 2007).

In principle, in an imprinting event, one parental copy of a gene is silenced due to methylation (Baran, et al., 2015; Tycko, 1994). Putative imprinting events can be identified using genetic/sequencing data, wherein a heterozygous SNV identified through genomic data will have a RNA-seq reads containing only reference or the alternate allele at the locus. 53 experimentally-validated MAE events were reported to be based on a global allele-specific gene expression involving 4 *Bos taurus indicus* x *Bos taurus taurus* F1 conceptuses (foetus and placenta) (Chen, et al., 2016). RNA-seq data from multiple tissues (brain, muscle, placenta, liver and kidney) was utilised and the 53 genes showed MAE in at least one tissue. Out of 53, 35 genes were reported to be imprinted in this cattle study of which 8 genes were novel and were experimentally validated. A summary of the 53 genes is presented in Table 16. 3 of the 1205 MAE genes (*NAP1L5*, *PLAGL1* and *SGCE*) discovered in water buffalo are contained with this set, but this would not represent a significant enrichment relative to the full set of protein-coding genes. Most imprinted genes are not expressed by macrophages. Many play an important role in prenatal stages (foetal growth and development) and many genes remain imprinted only in that stage i.e. they are placental specific (Bartolomei and Tilghman, 1997; Monk, et al., 2006). However, some imprinted genes are also important in normal postnatal growth and development, for example, neuronal development (Lozano-Urena, et al., 2017). The cattle imprinted gene *NAP1L5* (Nucleosome Assembly Protein 1 Like 5) has been reported to be expressed in the brain and adrenal gland of adult mice (Smith, et al., 2003). It was also found to have possible roles in foetal growth and development in cattle (Zaitoun and Khatib, 2006). No such observation was found for *PLAGL1* and *SGCE*.

Gene symbol	Imprinted	Comment
BEGAIN	yes	had been reported before/experimentally validated
CDKN1C	yes	had been reported before/experimentally validated
DIRAS3	yes	had been reported before/experimentally validated
DLK1	yes	had been reported before/experimentally validated
GNAS	yes	had been reported before/experimentally validated
H19	yes	had been reported before/experimentally validated
IGF2	yes	had been reported before/experimentally validated
IGF2R	yes	had been reported before/experimentally validated
INPP5F	yes	had been reported before/experimentally validated
MAGEL2	yes	had been reported before/experimentally validated
MEG3	yes	had been reported before/experimentally validated
NAP1L5	yes	had been reported before/experimentally validated
NNAT	yes	had been reported before/experimentally validated
PEG10	yes	had been reported before/experimentally validated
PEG3	yes	had been reported before/experimentally validated
PHLDA2	yes	had been reported before/experimentally validated
PLAGL1	yes	had been reported before/experimentally validated
RTL1	yes	had been reported before/experimentally validated
SGCE	yes	had been reported before/experimentally validated
SNRPN	yes	had been reported before/experimentally validated
MEG9	yes	had been reported before/experimentally validated
MIMT1	yes	had been reported before/experimentally validated
USP29	yes	had been reported before/experimentally validated
MGC157368	ND	lack of informative SNPs
LOC508098	ND	lack of informative SNPs
LOC100848941	ND	lack of informative SNPs
LOC100848941 brain	ND	lack of informative SNPs

isoform		
LOC101907203	ND	lack of informative SNPs
XLOC_045114	ND	lack of informative SNPs
AOX1	ND	lack of informative SNPs
APCS	ND	lack of informative SNPs
AS3MT	ND	lack of informative SNPs
KRT7	ND	lack of informative SNPs
OOEP	ND	lack of informative SNPs
XLOC_009410	ND	lack of informative SNPs
GNASXL	yes	novel/experimentally validated
COPG2IT1	yes	novel/experimentally validated
MEG8	yes	novel/experimentally validated
LOC100298176	yes	novel/experimentally validated
LOC104974975	yes	novel/experimentally validated
LOC101907679	yes	novel/experimentally validated
LOC100849023	yes	novel/experimentally validated
XLOC_052524	yes	novel/experimentally validated
C1R	no	only monoallelically expressed
C1S	no	only monoallelically expressed
CDA	no	only monoallelically expressed
LOC101905472	no	only monoallelically expressed
RDH16	no	only monoallelically expressed
XLOC_012439	no	only monoallelically expressed
PON3	yes	previously reported to be imprinted in human and/or mouse
PPP1R9A	yes	previously reported to be imprinted in human and/or mouse
MKRN3	yes	previously reported to be imprinted in human and/or mouse
ZIM2	yes	previously reported to be imprinted in human and/or mouse

Table 16: Gene list of 53 genes showing MAE along with their imprinted status and comments. 'ND' means not determined.

3.4 Conclusions

This chapter dealt with analysing the presence of regulatory variation in the form of ASE in macrophage cells. This was mainly done in order to determine whether regulatory variation existed in immune genes. Since macrophages are one of the first lines of host-defense against pathogens (Weiss and Schaible, 2015) and helpful in triggering an immune response, these cells provide the opportunity to perform this study on immune specific genes. In order to perform the ASE analysis, multiple samples were utilised to call high quality

heterozygous variants, and accounted for various technical errors, such as low quality RNA-seq reads, double-counting of RNA-seq reads and reference mapping bias. The limitation of genes showing artefactual ASE in CNV regions discussed in the last chapter remains a limitation in this chapter too. An extensive CNV analysis needs to be performed in genes present in CNV regions should be discarded from ASE analysis in this chapter as well. The limitation of considering the presence of overlapping genes in ASE analysis also remains in this chapter.

The ASE analysis revealed the existence of high and low expressed alleles in many macrophage-expressed genes. The prevalence is broadly consistent with ASE studies done in cattle. Similar ASE is inferred from eQTL study of human monocytes, where >80% of transcripts showed evidence of heritable variation in the level of expression (Fairfax, et al., 2014). ASE was often sample specific and individual specific, as intersectional analysis revealed that most of the genes were private to each sample or each individual with some genes that were common to all samples/individuals. The reason may be differences in accessibility such as unavailability of biallelic heterozygous SNVs or informative SNVs. This observation was also seen in an ASE profile study involving sheep (Salavati, et al., 2019). Since the DNA sequence is available for each animal, it would be possible to identify candidate *cis*-acting variation in each locus that might explain the allelic imbalance. Alternatively an eQTL analysis might reveal whether SNVs associated with ASE loci can also be linked to substantial variation in expression, especially when present as homozygotes. There is also the potential to extend this study using the methods optimised herein to the much larger dataset generated in the water buffalo expression atlas (Young, et al., 2019).

3.5 Supplementary Material

Sample	total reads	mapped reads	mapped %	paired end	paired end %	properly paired	properly paired %	with itself and mate mapped	with itself and mate mapped %	singletons	Singletons %
Female 1 MDM	42,061,248	42,061,248	100	40,839,664	97.10	36,755,870	90	40,445,092	99.03	394,572	0.97
Female 2 AM	33,202,689	33,202,689	100	31,816,096	95.82	27,597,386	86.74	31,380,912	98.63	435,184	1.37
Female 2 BMDM	123,262,612	123,262,612	100	117,976,442	95.71	95,342,416	80.81	114,893,396	97.39	3,083,046	2.61
Female 2 BMDM7hrLPS	130,294,180	130,294,180	100	124,618,295	95.64	100,909,100	80.97	121,505,090	97.50	3,113,205	2.5
Female 2 MDM	34,429,374	34,429,374	100	32,694,095	94.96	27,765,614	84.93	32,130,878	98.28	563,217	1.72
Male 1 BMDM7hrLPS_Female 1 BMDM7hrLPS	461,242,648	461,242,648	100	436,548,039	94.65	376,215,302	86.18	428,657,872	98.19	7,890,167	1.81
Male 1 BMDM_Female 1 BMDM	423,651,801	423,651,801	100	403,350,837	95.21	347,906,442	86.25	395,956,630	98.17	7,394,207	1.83
Male 1MDM	28,611,390	28,611,390	100	27,764,562	97.04	24,513,384	88.29	27,489,400	99.01	275,162	0.99
Male 2 AM	33,997,375	33,997,375	100	32,545,308	95.73	27,940,610	85.85	32,083,162	98.58	462,146	1.42
Male 2 BMDM	134,561,118	134,561,118	100	129,011,249	95.88	107,054,912	82.98	126,143,866	97.78	2,867,383	2.22
Male 2 BMDM7hrLPS	109,228,135	109,228,135	100	104,373,953	95.56	87,002,030	83.36	102,214,134	97.93	2,159,819	2.07
Male 2 MDM	30,550,624	30,550,624	100	29,242,284	95.72	25,138,940	85.97	28,793,936	98.47	448,348	1.53

Table S 4: RNA-seq alignment summary of 12 samples from four animals. Alignment summary generated using Samtools 'flagstat' program.

Sample	total reads	mapped reads	mapped %	paired end	paired end %	properly paired	properly paired %	with itself and mate mapped	with itself and mate mapped %	singletons	singletons %	Duplicates	Duplicates %
Banni_1	916,989,584	913,422,674	99.61	913,569,572	99.63	896,185,068	98.1	907,158,754	99.30	2,843,908	0.31	89,145,339	9.72
Banni_2	920,660,181	917,089,155	99.61	916,136,528	99.51	897,462,764	97.96	909,713,182	99.30	2,852,320	0.31	97,043,326	10.54
Banni_3	933,964,778	930,663,893	99.65	930,088,628	99.58	913,000,890	98.16	924,131,436	99.36	2,656,307	0.29	87,381,511	9.36
Banni_4	903,886,821	899,983,286	99.57	900,121,974	99.58	881,881,448	97.97	893,029,500	99.21	3,188,939	0.35	100,851,415	11.16
Banni_5	936,346,329	932,253,071	99.56	932,520,860	99.59	913,100,804	97.92	925,049,942	99.20	3,377,660	0.36	98,142,279	10.48
Banni_6	925,733,806	922,069,526	99.6	922,277,066	99.63	905,296,562	98.16	915,803,804	99.30	2,808,982	0.3	151,259,350	16.34
Bhadawari_1	932,928,305	929,794,237	99.66	930,036,612	99.69	914,525,818	98.33	924,576,612	99.41	2,325,932	0.25	98,876,451	10.60
Bhadawari_2	926,248,832	922,788,775	99.63	923,074,804	99.66	906,678,290	98.22	916,999,586	99.34	2,615,161	0.28	103,116,943	11.13
Bhadawari_3	910,376,110	905,146,523	99.43	906,745,602	99.60	884,785,168	97.58	897,264,440	98.95	4,251,575	0.47	85,989,092	9.45
Bhadawari_4	936,669,884	933,620,921	99.67	933,866,454	99.70	918,693,824	98.38	928,598,500	99.44	2,218,991	0.24	96,468,698	10.30
Bhadawari_5	939,179,082	936,099,569	99.67	935,682,580	99.63	919,902,218	98.31	930,312,002	99.43	2,291,065	0.24	117,590,829	12.52
Bhadawari_6	933,426,672	929,713,736	99.6	929,238,310	99.55	910,721,886	98.01	922,668,802	99.29	2,856,572	0.31	147,240,294	15.77
Bhadawari_male	958,406,227	948,627,441	98.98	954,427,252	99.58	923,850,678	96.8	939,974,208	98.49	4,674,258	0.49	189,373,892	19.76
Bhadhwari-9-369	133,708,359	132,874,443	99.38	131,397,748	98.27	112,875,790	85.9	130,275,338	99.15	288,494	0.22	2,250,947	1.68
Bhadhwari-B167	132,175,709	131,707,799	99.65	128,799,632	97.45	122,293,854	94.95	128,059,944	99.43	271,778	0.21	1,358,134	1.03
Bhadhwari-B254	125,543,118	124,868,139	99.46	123,486,794	98.36	103,738,642	84.01	122,591,682	99.28	220,133	0.18	1,962,717	1.56
Bhadhwari-B277	126,874,407	126,180,207	99.45	124,638,800	98.24	117,159,742	94	123,622,232	99.18	322,368	0.26	1,390,858	1.10
Bhadhwari-B278	126,016,659	123,664,236	98.13	124,130,474	98.50	110,448,996	88.98	121,362,868	97.77	415,183	0.33	2,048,428	1.63
Bhadhwari-B284	127,570,304	126,982,438	99.54	125,720,034	98.55	112,383,424	89.39	124,988,838	99.42	143,330	0.11	1,666,896	1.31
Bu.B.2	144,821,753	144,346,535	99.67	143,204,712	98.88	139,284,626	97.26	142,451,566	99.47	277,928	0.19	2,238,288	1.55
Bu.B.4	149,439,517	148,845,157	99.6	147,275,440	98.55	142,592,820	96.82	146,343,518	99.37	337,562	0.23	2,999,353	2.01
Bunny-0903	127,052,005	126,369,851	99.46	124,321,644	97.85	116,065,456	93.36	123,329,000	99.20	310,490	0.25	1,558,806	1.23
Bunny-0914	127,581,642	126,898,977	99.46	125,721,316	98.54	110,772,704	88.11	124,831,660	99.29	206,991	0.16	1,947,877	1.53

Bunny-0982	125,375,793	124,609,681	99.39	123,484,576	98.49	108,481,610	87.85	122,376,538	99.10	341,926	0.28	1,689,409	1.35
Bunny-960	123,708,564	122,944,963	99.38	121,279,090	98.04	108,458,854	89.43	120,132,940	99.05	382,549	0.32	1,688,462	1.36
Female_Grant own_buffalo	1,006,208,525	1,002,779,063	99.66	1,002,570,974	99.64	984,844,656	98.23	996,764,314	99.42	2,377,198	0.24	186,711,144	18.56
Jaffarabadi_1	869,214,856	864,515,459	99.46	865,650,520	99.59	846,432,958	97.78	857,749,926	99.09	3,201,197	0.37	110,182,340	12.68
Jaffarabadi_2	895,473,643	888,234,687	99.19	891,259,678	99.53	866,835,688	97.26	880,916,104	98.84	3,104,618	0.35	88,613,074	9.90
Jaffarabadi_3	909,679,525	905,001,474	99.49	906,303,128	99.63	887,499,628	97.93	898,192,600	99.11	3,432,477	0.38	101,816,019	11.19
Jaffarabadi_4	894,796,482	890,344,459	99.5	891,289,420	99.61	873,157,188	97.97	883,816,636	99.16	3,020,761	0.34	80,630,714	9.01
Jaffarabadi_5	887,714,709	884,634,834	99.65	884,665,954	99.66	869,485,376	98.28	879,562,956	99.42	2,023,123	0.23	97,220,000	10.95
Jaffarabadi_6	915,659,031	910,520,505	99.44	912,186,938	99.62	892,121,442	97.8	903,074,796	99.00	3,973,616	0.44	124,916,290	13.64
Jaffarabadi_ male	901,521,907	894,127,305	99.18	897,222,974	99.52	870,667,306	97.04	885,573,776	98.70	4,254,596	0.47	81,194,494	9.01
Jaffrabadi-0685	128,024,055	127,381,670	99.5	126,216,728	98.59	113,958,808	90.29	125,394,398	99.35	179,945	0.14	1,604,988	1.25
Jaffrabadi-0845	119,855,051	119,060,116	99.34	117,927,646	98.39	104,971,466	89.01	116,872,840	99.11	259,871	0.22	1,496,544	1.25
Jaffrabadi-2304430	128,865,451	128,055,598	99.37	126,625,760	98.26	106,304,834	83.95	125,532,202	99.14	283,705	0.22	1,887,823	1.46
Jaffrabadi-230964	125,817,115	125,269,060	99.56	123,841,432	98.43	120,196,498	97.06	122,989,476	99.31	303,901	0.25	2,253,016	1.79
Jaffrabadi-548	132,662,332	131,817,349	99.36	130,234,470	98.17	118,778,854	91.2	128,958,908	99.02	430,579	0.33	2,154,048	1.62
Jaffrabadi-971	127,603,393	126,969,023	99.5	125,239,258	98.15	120,045,376	95.85	124,307,732	99.26	297,156	0.24	1,828,633	1.43
Kirkcaldy_ female	978,757,402	973,430,844	99.46	974,939,548	99.61	951,206,282	97.57	965,920,286	99.07	3,692,704	0.38	249,130,660	25.45
Kirkcaldy_ male	999,866,533	992,483,668	99.26	994,956,774	99.51	966,521,768	97.14	983,369,874	98.84	4,204,035	0.42	271,549,195	27.16
Lodi_female	919,161,311	911,055,864	99.12	915,908,656	99.65	890,427,482	97.22	903,959,526	98.70	3,843,683	0.42	171,721,929	18.68
Lodi_male	946,845,386	940,871,128	99.37	943,938,912	99.69	922,490,264	97.73	933,754,016	98.92	4,210,638	0.45	140,862,472	14.88
Male_Granto wn_buffalo	1,006,889,509	1,000,747,093	99.39	1,002,827,242	99.60	979,966,826	97.72	993,913,578	99.11	2,771,248	0.28	199,103,382	19.77
Murrah_1	871,228,199	867,222,467	99.54	867,979,470	99.63	850,229,184	97.95	860,787,692	99.17	3,186,046	0.37	77,128,769	8.85
Murrah_2	913,536,190	909,563,387	99.57	909,982,646	99.61	891,885,310	98.01	902,809,058	99.21	3,200,785	0.35	88,113,486	9.65
Murrah_3	952,711,445	949,142,812	99.63	949,014,758	99.61	931,164,400	98.12	942,734,088	99.34	2,712,037	0.29	131,607,106	13.81
Murrah_4	952,715,728	947,830,412	99.49	948,903,714	99.60	927,428,958	97.74	939,903,522	99.05	4,114,876	0.43	148,834,738	15.62
Murrah_5	939,958,479	936,699,276	99.65	936,910,064	99.68	921,057,212	98.31	931,174,872	99.39	2,475,989	0.26	109,454,357	11.64

Murrah_6	912,809,640	909,535,003	99.64	909,465,922	99.63	893,628,242	98.26	903,707,268	99.37	2,484,017	0.27	114,384,410	12.53
Murraha-2984	122,658,657	121,928,635	99.4	120,799,018	98.48	108,038,672	89.44	119,864,142	99.23	204,854	0.17	1,688,332	1.38
Murraha-RASM-28	122,134,222	121,490,538	99.47	120,098,700	98.33	105,937,892	88.21	119,212,400	99.26	242,616	0.2	1,109,791	0.91
Murraha-RASM2	125,689,459	125,011,818	99.46	123,662,588	98.39	109,375,374	88.45	122,737,428	99.25	247,519	0.2	1,108,339	0.88
Murraha-RRP277	130,866,580	130,321,018	99.58	128,307,216	98.04	123,053,644	95.91	127,515,688	99.38	245,966	0.19	1,453,123	1.11
Murraha-RSM-36	133,861,430	133,138,140	99.46	131,769,560	98.44	113,543,406	86.17	130,808,198	99.27	238,072	0.18	1,678,448	1.25
Murraha-RSM-3B	123,457,572	122,660,089	99.35	121,669,356	98.55	107,285,976	88.18	120,623,670	99.14	248,203	0.2	1,954,523	1.58
Pandharpuri-62222	138,748,158	137,159,216	98.85	136,927,462	98.69	124,039,774	90.59	134,944,314	98.55	394,206	0.29	3,436,859	2.48
Pandharpuri-M240	126,425,978	125,759,579	99.47	124,438,762	98.43	109,933,458	88.34	123,605,468	99.33	166,895	0.13	1,800,828	1.42
Pandharpuri-M256	180,079,868	179,144,552	99.48	176,804,474	98.18	169,639,984	95.95	175,483,694	99.25	385,464	0.22	3,156,407	1.75
Pandharpuri-M257	122,578,506	121,718,617	99.3	120,887,734	98.62	105,912,880	87.61	119,845,198	99.14	182,647	0.15	1,653,625	1.35
Pandharpuri-M73	126,315,139	125,552,275	99.4	124,419,262	98.50	120,004,086	96.45	123,339,882	99.13	316,516	0.25	2,093,457	1.66
Pandharpuri-M7	137,045,272	136,159,518	99.35	134,621,918	98.23	128,144,674	95.19	133,291,088	99.01	445,076	0.33	2,919,673	2.13
Pandharpuri_1	947,468,938	940,067,985	99.22	943,699,698	99.60	918,500,832	97.33	932,313,228	98.79	3,985,517	0.42	132,952,926	14.03
Pandharpuri_2	951,910,472	944,097,047	99.18	947,942,882	99.58	920,841,036	97.14	935,788,192	98.72	4,341,265	0.46	134,099,202	14.09
Pandharpuri_3	959,306,945	952,113,630	99.25	955,339,294	99.59	930,109,128	97.36	944,356,444	98.85	3,789,535	0.4	112,267,629	11.70
Pandharpuri_4	980,263,040	972,994,451	99.26	976,071,578	99.57	949,912,568	97.32	964,947,592	98.86	3,855,397	0.39	119,985,897	12.24
Pandharpuri_5	964,337,365	957,436,053	99.28	960,394,016	99.59	935,699,246	97.43	950,123,092	98.93	3,369,612	0.35	118,312,862	12.27
Pandharpuri_6	921,630,597	915,386,326	99.32	917,921,894	99.60	895,454,012	97.55	908,660,948	98.99	3,016,675	0.33	100,607,546	10.92
Pandharpuri_female	876,243,757	812,363,254	92.71	873,190,342	99.65	792,052,478	90.71	804,799,278	92.17	4,510,561	0.52	109,870,354	12.54
Su.B.1	148,201,913	147,537,293	99.55	146,686,218	98.98	142,766,468	97.33	145,645,166	99.29	376,432	0.26	1,873,477	1.26
Su.B.3	134,321,048	133,621,831	99.48	132,268,170	98.47	127,922,188	96.71	131,374,720	99.32	194,233	0.15	1,925,707	1.43
Surti-078	123,315,669	122,606,282	99.42	121,225,106	98.30	107,031,610	88.29	120,278,138	99.22	237,581	0.2	1,555,593	1.26
Surti-214	127,858,410	127,121,207	99.42	125,643,190	98.27	111,247,158	88.54	124,669,516	99.23	236,471	0.19	1,770,931	1.39
Surti-251	134,161,592	133,683,900	99.64	131,999,840	98.39	127,976,906	96.95	131,309,522	99.48	212,626	0.16	1,937,222	1.44

Surti-B367	127,621,242	127,018,755	99.53	125,546,906	98.37	106,681,554	84.97	124,769,350	99.38	175,069	0.14	1,630,557	1.28
Surti_1	916,870,577	913,031,824	99.58	913,279,678	99.61	895,432,092	98.05	906,466,670	99.25	2,974,255	0.33	147,543,880	16.09
Surti_2	923,559,763	920,231,955	99.64	920,440,550	99.66	904,255,722	98.24	914,519,308	99.36	2,593,434	0.28	107,823,859	11.67
Surti_3	822,971,146	819,964,833	99.63	820,015,546	99.64	805,988,074	98.29	814,785,332	99.36	2,223,901	0.27	95,980,742	11.66
Surti_4	825,481,964	822,431,847	99.63	822,767,688	99.67	808,510,422	98.27	817,470,560	99.36	2,247,011	0.27	59,554,171	7.21
Surti_5	861,745,281	858,368,235	99.61	858,447,216	99.62	842,610,764	98.16	852,538,988	99.31	2,531,182	0.29	108,115,809	12.55
Surti_6	892,310,079	889,537,461	99.69	889,397,234	99.67	875,009,454	98.38	884,601,620	99.46	2,022,996	0.23	94,183,571	10.56

Table S 5: DNA-seq alignment summary from 81 water buffalo. Alignment summary generated using from Samtools flagstat program.

Gene	Female 2 BMDM	Female 2 BMDM7hr LPS	Male 1 BMDM	Male 1 BMDM7hr LPS	Male 2 BMDM	Male 2 BMDM7hrLPS
ARL11	0.77	0.76	0.75	0.75	0.71	0.79
C19H5orf58	0.96	0.95	0.98	0.98	0.99	0.98
CAPRIN1	0.98	0.97	1.00	1.00	1.00	1.00
CCNI	0.87	0.83	1.00	1.00	0.85	0.85
CEMIP2	0.94	0.94	0.88	0.87	1.00	1.00
CHN1	0.74	0.93	0.81	0.82	0.85	0.87
EPB41L5	0.81	0.74	0.93	0.88	0.84	0.76
ETFRF1	1.00	1.00	0.99	1.00	1.00	0.99
FBN1	0.74	0.71	0.91	0.85	0.91	0.90
FBXW7	0.99	0.95	1.00	0.99	0.85	0.94
GHITM	0.99	0.99	0.98	0.99	0.83	0.83
MINPP1	0.90	0.82	0.89	0.83	0.88	0.83
MSRB2	0.75	0.81	0.78	0.77	0.83	0.89
NQO2	0.79	0.75	0.80	0.79	0.75	0.87
NSMCE3	1.00	1.00	1.00	1.00	0.98	1.00
PLEKHM3	0.86	0.84	0.97	0.96	0.96	0.94
RAP1B	0.72	0.87	0.94	0.90	0.95	0.96
RFLNB	0.72	0.87	0.94	0.90	0.95	0.96
RRP12	0.72	0.71	1.00	1.00	1.00	1.00
STEAP3	0.92	0.84	0.96	0.97	0.90	0.89

Table S 6: MBASED Major Allele Frequency (MAF) values for genes showing ASE in all BMDM +/- LPS samples. These were the only genes that had official gene symbols out of 75 genes that showed ASE in all BMDM +/- LPS samples

Chapter 4. Genetic diversity analysis

4.1 Introduction

The water buffalo (*Bubalus bubalis*) is one of the most important species in the Indian subcontinent providing milk, meat, hide and draught power (Cockrill, 1977). According to, FAOSTAT (http://www.fao.org/faostat/), as of 2017, India has the highest number of water buffalo in the world (approx. 113 million). In India, different breeds of river buffalo (subspecies of the water buffalo, the other being the swamp buffalo) have been developed through selection into many high performance dairy animals (Cockrill, 1977). An early study analysed the maternal lineages of eight distinct breeds based upon mitochondrial D-loop region sequences (Kumar, et al., 2007). These studies indicated that breed differentiation and haplotype expansion was relatively ancient, following domestication around 6,300 years ago. Through migration and importation, water buffalo spread worldwide, and have become an important commercial source of milk production in Middle Eastern and Mediterranean countries (Cockrill, 1981). The 90K SNP genotyping array was developed based upon SNP discovery in several river buffalo breeds (Iamartino, et al., 2017). An analysis of 31 water buffalo populations across the world, using this array, supported the view that the westward spread of river buffalo from India occurred in multiple independent waves (Colli, et al., 2018). This water buffalo SNP array was developed based upon a relatively fragmented draft genome (Williams, et al., 2017) and with limited representation of the diverse breeds from the Indian subcontinent. The rapid decrease in the cost of genomic DNA sequencing has produced a revolution in the analysis of genotype-phenotype relationships in livestock and in the identification of selective sweeps associated with performance traits and adaptation. This chapter deals with identifying underlying diversity amongst Indian breeds and signatures of positive selection.

Biological evolution and diversity is mainly categorised into two forms. The first form is macroevolution, which deals with evolution above the species level (i.e. differences between species or even genera or phyla) and happens over a

larger period of time (Callahan, 2002). The second form is microevolution, which can be defined as change or variation within the population of a species (Callahan, 2002). It explains how individuals are different from other individuals in the same population. For example, one of the two alleles or two forms of a gene in a population may become more common or rare as the population evolves from one generation to another. Microevolution is basically this change of allele frequency (how common an allele of a gene is) in a population which occurs over a comparatively short period of time. There are four basic mechanisms by which microevolution takes place, which are - natural selection, mutations, genetic drift and gene flow (Hendry and Kinnison, 2001). This chapter focusses on understanding the impact of natural selection.

Mutations can be defined as the mechanism that results in a heritable change in the DNA of a gene (Anthony J.F. Griffiths, et al., 2015). At a phenotypic level, point mutations (substitution of one base) affect protein functions if they occur in the protein coding regions, such as exons. Synonymous mutations change the codon of an amino acid into another that codes for the same amino acid. Missense or non-synonymous mutations change the codon for one amino acid into a codon that codes for another amino acid. Nonsense mutations take place when one amino acid changes into a stop codon, terminating the translation of the coded protein. Base insertions and deletions (indels) result in the addition or deletion of base pairs. SNV or single nucleotide variant is a type of point mutation that happens at the single base level. SNP or single nucleotide polymorphism is a term that is used at the population level often to refer specifically to single base mutations that have risen to a frequency greater than 1% of the population (Karki, et al., 2015). This means that the frequency of the least abundant allele or minor allele frequency will be 1% or more for a SNP (Brookes, 1999). The vast majority of DNA sequence variants in any individual do not reach the level to be regarded as SNPs, and are private or rare. This basically means they have arisen relatively recently within a family or pedigree. For example, in humans, 100 to 200 new DNA sequence variants arise in each generation (Xue, et al., 2009). In this study, since we have applied a minor allele

frequency threshold of 0.05 (or 5%) in most of our downstream analysis, the SNVs left after applying the threshold are referred to as SNPs.

Based on how mutation effects an individual, it can be good/beneficial if it increases the organism's fitness (ability to reproduce in an environment), it can be bad/deleterious if it decreases the fitness or it can be indifferent/neutral if it has little effect (Loewe and Hill, 2010). Beneficial traits arising from that gene or mutation can be those that allow an individual to adapt to its external environment, some of which may be changes in coat colour and body size, the ability to use a new food source or resistance to a pathogen. It is the force that pushes the growth of such advantageous traits and plays a central role in the development of a species (Sabeti, et al., 2006).

Genetic drift can be defined as allele frequency change due to sampling (Anthony J.F. Griffiths, et al., 2015). In other words, it is the change in the frequency of an allele of a gene in the next generation of a population as a result of random mating in the current generation. The size of the population plays an important role in genetic drift. A beneficial allele of a gene can be eliminated due to random mating amongst the individuals of a population or a deleterious variant can achieve fixation (reach 100% frequency) by chance. But these changes can be achieved more quickly if the population is small whereas if it is large, genetic drift may take a very long time to achieve such a result (Anthony J.F. Griffiths, et al., 2015). Genetic drift can be amplified by a 'bottleneck effect'. A population achieves a bottleneck if its size contracts generation after generation. Environmental fluctuations or natural disasters such as floods, earthquakes, fires, etc. or even human interventions such as hunting may cause reduction in the population size, leading to a potentially increased impact of genetic drift within that population (Anthony J.F. Griffiths, et al., 2015). This is because such drastic events may lead to the loss of certain alleles completely leading to loss of genetic variation/diversity.

Migration or gene flow is a phenomenon that takes place due to the movement between populations leading to change in allele frequencies in either group (Anthony J.F. Griffiths, et al., 2015).

Natural selection can be defined as the process by which individuals who have certain heritable features have a greater chance to reproduce and survive than individuals that do not have those features (Anthony J.F. Griffiths, et al., 2015). 'Positive' selection is the phenomenon in which a beneficial allele of a gene or a beneficial mutation is spread across the population and increases in frequency. When a deleterious mutation is removed from the population, it is called 'purifying' or 'negative' selection. It prevents any adaptive features to be removed from the population (Anthony J.F. Griffiths, et al., 2015). Another form of selection other than natural selection which allows the promotion of advantageous traits is 'artificial selection'. For example, humans have selected a favourable trait in animals and plants while domesticating them (Anthony J.F. Griffiths, et al., 2015).

Detecting recent positive selection in organisms can give an idea about the timescales of when genetic adaptation and disease resistance arose (Sabeti, et al., 2002). Positive selection is the primary mechanism by which adaptation takes place (development of phenotypes important for specific environment) (Vitti, et al., 2013). When a beneficial allelic variant is selected, variants on the same chromosome, which share identity by descent but may confer no selective advantage, are co-selected. This phenomenon is referred to as 'hitchhiking effect' (Smith and Haigh, 1974). As a consequence, the degree of variation around the selected allele decreases in the selected population, leading to a phenomenon known as a 'selective sweep' (Nielsen, et al., 2005). Selective sweep is considered to be a signature of positive selection and can be detected using different methods that have been extensively described by Sabeti *et al* (Vitti, et al., 2013). We have used a method based on Linkage disequilibrium or LD. Many LD based methods have been developed to detect selection signature indicating positive selection, including EHH (extended haplotype homozygosity) (Sabeti, et al., 2002), iHS (integrated haplotype score) (Voight, et al., 2006) and XP-EHH (Cross Population Extended Haplotype Homozygosity) (Sabeti, et al., 2007).

Two variants are said to be in LD when certain combinations of their alleles are inherited together, more often than would be expected by chance. During a

selective sweep, the causal variant or the beneficial allele remains in strong LD or physical linkage with its nearby hitchhiking variants, thus reducing genetic diversity. This results in the formation of a long haplotype (combination of alleles that are inherited together), since the hitchhiking takes place at a very high rate and recombination does not have enough time to break the haplotype. Extended regions of such strong LD (and hence, long haplotypes) may be a signature of positive selection. The EHH method finds areas of selection based on the association between an allele's frequency and the magnitude of LD around it, specifically measuring haplotype homozygosity between one locus and other loci at multiple distances away (Sabeti, et al., 2002). The iHS is a statistic for detecting the evidence of positive selection at a locus. It is based on differential levels of LD around a positively selected allele compared to a background allele at the same locus. This test requires the knowledge of the ancestral and derived allele. XP-EHH is a variant of the EHH method which compares haplotype lengths between populations. It detects selective sweeps in which the causal allele has reached fixation in one population but is polymorphic in another. The XP-EHH score is directional in nature i.e. a positive score indicates that selection has likely happened in the first population present in the comparison, whereas a negative value indicates the same for the other population (Sabeti, et al., 2007).

Positive selection resulting from domestication has been identified in a variety of livestock species. An exploration of various cattle breeds from South Korea and West Africa revealed that many common and breed specific genes are related to biological and economic important traits such as milk production, reproduction and meat (Taye, et al., 2017). Genes related to immunity (*IRAK3*), thermotolerance (*HSF5*), coat colour (*PMEL*), reproduction and fertility (*SRD5A3 and AFP*) were found to be positively selected in Butana and Kenana dairy zebu cattle of Africa (Bahbahani, et al., 2018). Based on approx. 70,000 SNPs from seven cattle breeds, 17 genes were identified to be under selection and associated with body size, milk production, reproduction, coat colour and muscle development (Zhao, et al., 2015). In another recent study, selection signatures were identified that included genes and pathways related to trypanosome

tolerance, coat colour, horn development, heat tolerance and tick resistance across African cattle breeds (Kim, et al., 2017). Multiple genes associated with features such as milk production, growth, immunity and apoptosis were identified under selection in Iranian water buffalo breeds (Mokhber, et al., 2018). *IGF2* gene, which is associated with growth and leanness in meat, was under selection in three pig breeds (Ojeda, et al., 2008). A selective sweep analysis in European domestic pigs revealed *MC4R* gene to be under selection that is associated with growth and feed intake (Rubin, et al., 2012). European pig breeds were examined to reveal selection signatures in genes related to coat colour, ear morphology and pork production traits that are used to define breed standards (Wilkinson, et al., 2013). A study involving 74 sheep breeds from all over the world identified genes related to coat colour, morphological and muscle development and the ability to survive in cold environments, which were under selection (Fariello, et al., 2014). 86 candidate genes were identified in Brazilian local sheep that were associated with immunity, nervous system development, sensory perception and reproduction (de Simoni Gouveia, et al., 2017). Regions with selective sweeps were identified in the 'Charolais de Cuba' breed of cattle that contained genes related to heat tolerance, immunity, muscle development and meat quality (Rodriguez-Valera, et al., 2018). Signatures of selection were also observed in European taurine cattle breeds in loci containing many genes related to coat colour, milk composition, metabolism, reproduction, etc. with causal variants present in the regulatory regions of the genes (Boitard, et al., 2016).

In the current study we have sequenced the genomes of 81 riverine buffalo across 7 breeds, out of which 6 breeds are from different locations in India in addition to the Mediterranean breed obtained from Italy and Scotland. Figure 23 shows the breeding tracts (area where a particular breed is found and reproduces) in India to which the 6 breeds belong. The map was created using ggmap package (Kahle, et al., 2019) in R (R Core Team, 2018) based on breeding tract information provided in the Dairy knowledge portal website maintained by the National Dairy Development Board (NDDB), Government of India (information present in Table 18). The results highlight extensive genetic

diversity and sights of putative adaptation within Indian water buffalo populations, which can potentially provide the basis for future genetic selection to improve traits such as fertility, productivity and disease resistance.

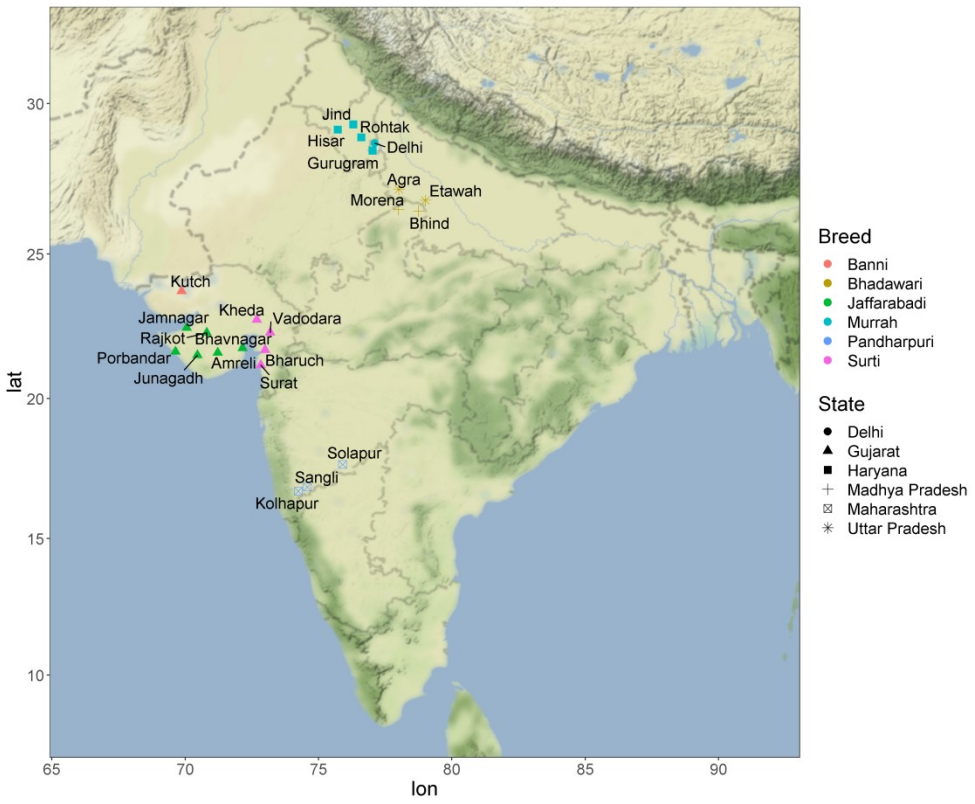

Figure 23: Breeding tracts in India to which the 6 Indian breeds under study belong. The 6 breeds are represented by different colours whereas the Indian states to which the breed belongs have been given different shapes. The name of the areas/district/cities to which the breed belongs has been named in the plot based on their latitudinal and longitudinal coordinates.

4.2 Methods

4.2.1 Sample information and sequencing

As shown in Table 17, the raw data consisted of Whole Genome Sequencing (WGS) data from 81 water buffalo across 7 breeds. The 7 breeds consisted of 6 Indian breeds and one Mediterranean breed, from which the reference sequence was derived. Paired-end sequencing of the animals was done across two sequencing centres: Edinburgh Genomics in the UK and SciGenom in India. The mean sequencing depths at each sequencing centre were 30x and 10x

respectively. Further sample information is provided in Table 17. Edinburgh Genomics sequenced the animals using the Illumina HiSeq X platform (read length: 150 bp) whereas SciGenom used the Illumina HiSeq 2500 platform (read length: 250 bp). The animals from the Mediterranean breed consisted of two animals from Lodi (Italy) and four animals from the same farm in Kirkcaldy, Fife, on the East coast of Scotland. The Indian breed samples belong to their respective breeding tracts as mentioned in Table 18.

4.2.2 Alignment, Variant calling and Variant filtration

Raw paired-end reads from all 81 samples were aligned to the newly assembled water buffalo reference genome released by NCBI on 14th May 2018 (Low, et al., 2019). NCBI completed and released accompanying gene annotation on 25th June 2018 . The paired-end reads were aligned to the reference genome using BWA-MEM v0.1.17 (Li, 2013). For each sample, BWA-MEM generated Sequence Alignment Map (SAM) output that was converted to Binary Alignment Map (BAM) output using Samtools v1.6. Each BAM file was coordinate sorted using Samtools 'sort' v1.6 with default parameters. Duplicates were marked using Picard 'MarkDuplicates' v2.14.0. Read groups (set of reads that were generated from a single sequencing run) were added to each BAM file using Picard 'AddOrReplaceReadGroups' v2.14.0.

Breed	Sample Size	Sequencing Centre	Depth
Italian (Mediterranean)	6	Edinburgh Genomics	30x
Jaffarabadi	13	SciGenom (India) and Edinburgh Genomics	6 animals-10x 7 animals-30x
Pandharpuri	13	SciGenom (India) and Edinburgh Genomics	6 animals-10x 7 animals-30x
Banni	12	SciGenom (India) and Edinburgh Genomics	6 animals-10x 6 animals-30x
Surti	12	SciGenom (India) and Edinburgh Genomics	6 animals-10x 6 animals-30x
Murrah	12	SciGenom (India) and Edinburgh Genomics	6 animals-10x 6 animals-30x
Bhadawari	13	SciGenom (India) and Edinburgh Genomics	6 animals-10x 7 animals-30x
7 BREEDS	**81 ANIMALS**	**2 SEQUENCING CENTRES**	**2 DEPTHS**

Table 17: Whole genome sequencing data summary of 81 samples

Breed	Breeding Tract in India	Link
Banni	Banni area of Kutch district of Gujarat	http://dairyknowledge.in/article/banni
Bhadawari	Bhind and Morena districts of Madhya Pradesh and Agra and Etawah districts of Uttar Pradesh	http://dairyknowledge.in/article/bhadawari
Jaffarabadi	Amreli, Bhavnagar, Jamnagar, Junagadh, Porbandar and Rajkot districts of Gujarat state	http://dairyknowledge.in/article/jaffarabadi
Murrah	Hisar, Rohtak, Gurgaon and Jind district of Haryana and Delhi	http://dairyknowledge.in/article/murrah
Pandharpuri	Solapur, Sangli and Kolhapur districts of Maharashtra	http://dairyknowledge.in/article/pandharpuri
Surti	Vadodara, Bharuch, Kheda and Surat districts of Gujarat	http://dairyknowledge.in/article/surti

Table 18: Breeding tract information of six Indian water buffalo breeds. Courtesy: Information System on Animal Genetic Resources of India (AGRI-IS) - developed at National Bureau of Animal Genetic Resources, Karnal, Haryana, India

For calling variants, the variant calling workflow was adapted from the GATK (The Genome Analysis Toolkit) (McKenna, et al., 2010) best practices workflow for Germline Single Nucleotide Polymorphisms (SNPs) and Indels (Insertions or Deletions) for non-human organism as mentioned in https://software.broadinstitute.org/gatk/best-practices/workflow?id=11145.

GATK HaplotypeCaller v4.0.4.0 (Poplin, et al., 2018) was run per sample in GVCF (Genomic Variant Call Format) file mode using the parameter -ERC GVCF. HaplotypeCaller utilises reads from BAM/SAM files and performs variant calling per sample to produce unfiltered genotype likelihoods. GenomicsDBImport v4.0.4.0 was then used to aggregate all 81 GVCF files per chromosome/scaffold. The tool takes one or more single-sample GVCFs and imports data over a single interval, and outputs a directory containing a GenomicsDB datastore with combined multi-sample data. Then, GenotypeGVCFs v4.0.4.0 with the -new-qual parameter was used to read from the created GenomicsDBs directly and output the final multi-sample VCF file per scaffold. GenotypeGVCFs performs joint genotype calling, assigns a quality score (QUAL) to each variant and removes low quality variants (Poplin, et al., 2018). Finally, Picard 'GatherVcfs' v2.14.0 was used to concatenate variants called per chromosome/unplaced scaffold to get the final multisample VCF file containing variants for all scaffolds. Only biallelic SNVs, which were obtained using BCFtools 'filter' v1.6 with the -v snps and -m2 -M2 parameters, were retained.

Sources of errors in Next-Generation Sequencing (NGS) technologies can lead to false-positives in variant and genotype calling (Ribeiro, et al., 2015). False-positives may arise due to base calling errors or may appear during read alignment (Nielsen, et al., 2011). In the variant filtration analysis step, six GATK recommended annotations were utilised that can be highly informative and robust in identifying errors . They are: QualByDepth (QD), FisherStrand (FS), StrandOddsRatio (SOR), RMSMappingQuality (MQ), MappingQualityRankSumTest (MQRankSum) and ReadPosRankSumTest (ReadPosRankSum). Only those variants were kept whose QD was >= 15, FS <= 60, SOR <= 2, MQ >= 50, MQRankSum >= -2.5, ReadPosRankSum >= -2.5. The process followed to get to these optimised thresholds has already been described in the previous chapter. The complete variant calling pipeline with all tools involved and their respective parameters used to generate the final multisample VCF is described in detail in Table 19.

Job	Tools with version	Parameter
Alignment	BWA-MEM 0.7.17, SAMtools view 1.6 (for conversion from SAM to BAM)	Defaults for bwa index (indexing the reference genome) Defaults for bwa mem (alignment) –b (for SAMtools view)
BAM file sorting	Samtools sort 1.6	Defaults
Marking Duplicates	Picard 2.14.0 MarkDuplicates	ASSUME_SORTED=true VALIDATION_STRINGENCY=SILENT MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=1024 TMP_DIR=<temp_directory> METRICS_FILE=<metrics_filename>
Adding read groups	Picard 2.14.0 AddOrReplaceReadGroups	RGLB=library RGPL=illumina RGPU=barcode RGSM=<sample_name> CREATE_INDEX=true
Call variants per sample in GVCF mode	GATK 4.0.4.0 Haplotypecaller	-ERC GVCF
Consolidate GVCFs	GATK 4.0.4.0 GenomicsDBImport	--TMP_DIR <temp_directory> --sample-name-map <sample_name_mapping_file> --reader-threads 2 --genomicsdb-workspace-path <chromosomes_and unplaced_scaffolds_name> -L <chromosomes_and unplaced_scaffolds_name>
Joint call cohort	GATK 4.0.4.0 GenotypeGVCF	-new-qual -V gendb:// <chromosomes_and unplaced_scaffolds_name>
Concatenate variants called per chromosome/unplaced_scaffold into a final multi-sample VCF file	Picard 2.14.0 GatherVcfs	Defaults
Index the final VCF file	Tabix	Defaults
Get only biallelic SNVs	BCFtools view 1.6	-v snps -m2 -M2
Hard filter biallelic variants	BCFtools filter 1.6	-i 'QD >= 15 & FS <= 60 & SOR <= 2 & MQ >= 50 & MQRankSum >= -2.5 & ReadPosRankSum >= -2.5'

Table 19: Variant calling pipeline used to perform joint variant calling across the 81 water buffalo samples

4.2.3 Population differentiation and structure

To understand the population structure within our samples, we limited ourselves to the 24 autosomes of the water buffalo. For Principal Component Analysis (PCA), PLINK v1.90b4 64-bit (Purcell, et al., 2007) was used to generate the principal components (PCs) from the filtered biallelic SNVs which were then plotted using ggplot2 package (Wickham, 2016) in R to get an idea of the population structure. The PCs were generated using the following parameters in PLINK: --allow-extra-chr (because the base multisample VCF file contained

unplaced scaffolds), --chr 1-24 (only consider autosomal chromosomes from 1 to 24), --chr-set 24, --geno 0 (only include SNVs genotyped in every sample), --maf 0.05 (minor allele frequency threshold was kept at 5% to minimise sequencing errors and keep common variants as it is more difficult to distinguish between a rare variant and a sequencing error), --pca 30 (produce 30 PCs), --vcf-min-gq <10,20,30,40,50>. Various values of minimum genotype qualities were tried in order to determine an appropriate threshold and a genotype quality of 20 (1% error rate) was chosen for all downstream analyses. A relatedness study was undertaken using VCFtools v0.1.13 with the –relatedness2 parameter (Manichaikul, et al., 2010) that outputs a relatedness statistic that was plotted using the R package Complexheatmap (Gu, et al., 2016) in order to infer sample relationships and as a quality control procedure to check if any samples had been sequenced twice. Admixture v1.3.0 (Alexander, et al., 2009) was used to estimate the ancestry in unrelated individuals for K=2 to 7 where K is the probable number of ancestral populations. For selecting variants for the admixture analysis, the autosomal biallelic filtered variants were processed using PLINK (with the same parameters used to process the variants for the PCA analysis), except two duplicate samples that were dropped using the --remove option bringing down the total number of animals to 79. In brief, genetic admixture takes place when a distantly related population (that has been comparatively reproductively isolated) interbreeds, results in the mixture of DNA from both lineages (Yang and Fu, 2018).

To estimate the genetic distance between populations (measurement of level of interbreeding) and to understand population genetic differentiation, an F_{st} (Wright, 1949) or Fixation index analysis was done using VCFtools v0.1.13 wherein Weir and Cockerham (Weir and Cockerham, 1984) weighted F_{st} values were calculated for each pair of populations. For this analysis, the biallelic filtered variants for 79 samples were processed using PLINK with these parameters: --allow-extra-chr, --allow-no-sex, --chr 1-24, --chr-set 24, --double-id, --geno 0, --nonfounders, --pca 30, var-wts, header, tabs, --recode vcf and --vcf-min-gq 20. In short, genetic distance can be defined as the degree of genetic or genomic difference observed between populations or species that can be

measured by a numerical method (F_{st} analysis in this case) (Nei, 2001). F_{st} values range between 0 and 1, with 0 indicating a shared population and 1 indicating completely isolated populations or no interbreeding. 0.00 to 0.05 signifies low genetic differentiation, 0.05 to 0.15 signifies moderate genetic differentiation and >0.15 signifies high genetic differentiation (Hartl and Clark, 1997; Sethuraman).

4.2.4 Inferring population splits and migration events

The software TreeMix v1.13 (Pickrell and Pritchard, 2012) was used to determine population splits and migration events amongst the populations involved in this study. The software utilises allele frequency data to model the course of evolution through a maximum likelihood approach. It allows one to visualise historical relationships between populations in the form of a phylogeny tree, whose branches split when the populations diverge because of genetic drift and unite when an event of admixture happens (two populations hybridise). It may also be used to visualise gene flow events between diverged populations (Pickrell and Pritchard, 2012).

The filtered biallelic SNVs were processed using PLINK v1.90b4 64-bit with these parameters: --allow-extra-chr, --allow-no-sex, --chr 1-24, --chr-set 24, --double-id, --geno 0, --maf 0.05, --make-bed, --nonfounders, --remove (to remove individuals that were sequenced twice), --update-ids (for updating family IDs) and --vcf-min-gq 20. The resulting PLINK binary file was again processed using PLINK v1.90b4 64-bit for calculating family-wise allele frequencies using these parameters: --chr 1-24, --chr-set 24, --family and –freq. Finally, a breed-stratified allele frequency report was obtained for 79 individuals. This file was converted to an input file suitable for the TreeMix program using the python script ‘plink2treemix.py’ (https://bitbucket.org/nygcresearch/treemix/downloads/). The TreeMix algorithm was run for 7 migration events using the –m parameter. The maximum likelihood tree was constructed using blocks of 1,000 SNVs to account for variants that are non-independent because they lie close to each other (-k 1000). The root was defined using the –root parameter considering the Mediterranean breed as ‘outgroup’. The Mediterranean water buffalo breed was

chosen as the 'outgroup' as it is distantly related to other Indian breeds of water buffalo. The outputs were plotted using the R function 'plot_tree' which was included with TreeMix software.

4.2.5 Identification of selective sweeps between populations using pairwise XP-EHH analysis

XP-EHH is a cross-population comparison test that detects selective sweeps in which one allele has achieved fixation in one population (for example a breed in our case) but remains polymorphic in another population of the same species (Sabeti, et al., 2007). It detects genomic signature of long haplotypes to identify regions of recent positive selection that may have relevance to important traits such as adaptability and productivity. 21 pairwise XP-EHH tests were performed between all pairs of breeds (7 breeds in total, 21 unique combinations) using hapbin's 'XP-EHH' program v1.3.0 (Maclean, et al., 2015) that calculates XP-EHH scores based on the method published in (Sabeti, et al., 2007). For this, the filtered biallelic SNVs from 79 samples were processed through PLINK v1.90b4 64-bit using the following parameters: --allow-extra-chr, --allow-no-sex, --chr 1-24, --chr-set 24, --double-id, --geno 0.2 (remove variants that have a genotype call rate less than 80%), --maf 0.05, --nonfounders and --vcf-min-gq 20. Missing genotypes were self-imputed (to infer missing genotypes) using Beagle v5.0 (Browning and Browning, 2007) with default parameters, as XP-EHH does not tolerate missing data. Phasing was done using Beagle v5.0 using default parameters. The recombination maps required for XP-EHH calculation were created per chromosome using a custom R script. The recombination rate was assumed to be 1 cM (centiMorgan) per Mbp for the water buffalo genome. Breed wise VCF files (each VCF file contained individuals from only one breed) were produced from the phased and imputed base VCF file using BCFtools v1.6, which were then converted to IMPUTE hap format (suitable for hapbin) using VCFtools v0.1.13. The hap files along with the map files were given as an input to hapbin (hap files given as input in pairs). The hapbin 'XP-EHH' program outputs unstandardised XP-EHH scores, which were then converted to standardised Z-scores for simplifying interpretation and visualisation. The Z-

scores resemble a normal distribution with a mean of 0 and standard deviation 1. This was done using the 'scale' function in R with parameters center = T and scale = T. For visualising the XP-EHH score distribution across the whole genome, Manhattan plots were created using the qqman package (Turner, 2014) in R. Benjamini-Hochberg (BH) (Benjamini and Hochberg, 1995) FDR corrected p-values was calculated from the Z-scores using the 'pnorm' function (to convert Z-scores to p-values) and 'p.adjust' function (for applying BH FDR correction). Clumping was performed using PLINK v1.90b4 64-bit on candidate SNVs to keep only one representative variant per region of linkage disequilibrium using parameters --allow-no-sex, --chr 1-24, --chr-set 24, --clump, --double-id and --nonfounders . To select candidate SNVs for clumping, the SNV corresponding to the lowest p-value (not FDR corrected) across all 21 combinations was kept and this was done across the whole genome. Absolute Z-scores corresponding to the index SNVs were obtained through a custom R script from all combinations and were then plotted as a heatmap using ComplexHeatmap (Gu, et al., 2016). Index SNVs present in the heatmap breed specific clusters were obtained and a bed file was generated (containing the coordinates of the region covering the index SNV and its associated clumped SNVs) using a custom R script. The cluster SNVs bed file was sorted and merged (for merging overlapping coordinates) using BEDTools v2.27.1 (Quinlan and Hall, 2010). In order to identify peaks with extreme XP-EHH scores, a custom R script was written that would extract SNVs with FDR corrected P values <= 0.01. Another custom R script was written that would extract the SNV with the highest absolute Z-score and create BED files adding 500 kbp upstream and downstream. The resulting BED files were then intersected with another bed file containing all the gene coordinates from the water buffalo gene annotation file in GTF using BEDTools intersect v2.27.1 to get a list of candidate genes that may be under selection.

4.2.6 Gene set enrichment analysis

The gene set enrichment analysis was done using DAVID v6.8 (Huang da, et al., 2009a; Huang da, et al., 2009b) to functionally annotate and cluster our

genes of interest. In an enrichment analysis, genes which are over-represented with a particular pathway, disease or phenotype annotation are identified. Through this analysis, we have tried to associate our genes to gene ontology terms. Unfortunately, the biological annotation of *Bubalus bubalis* is incomplete in DAVID's knowledgebase. Since the genes present in the human genome are well annotated as compared to water buffalo, the annotations of the orthologous human genes were used for this analysis. Hence, all the human genes along with their Ensembl IDs were downloaded from Ensembl genome browser (Frankish, et al., 2017) using Biomart (Smedley, et al., 2009). A list was made of all the genes with matching official gene names in the gene annotation file of *Bubalus bubalis* genome assembly. Using a custom R script, both the lists were mapped to each other to obtain human Ensembl IDs for *Bubalus bubalis* gene names. A total of 15938 out of 31995 genes from *Bubalus bubalis* had a corresponding human gene Ensembl ID, providing the background reference list for DAVID enrichment analyses.

4.3 Results and Discussion

4.3.1 Alignment, Variant calling and Variant filtration

In total, 81 water buffalo were sequenced that contained animals from 7 different breeds. BWA-MEM managed to map the sequences to the water buffalo reference genome (Low, et al., 2019) at an average alignment rate of 99.37% out of which an average of 94.78% reads were properly paired. The final multisample VCF file contained a total of 43,243,663 variants. Out of this total, there were 36,541,726 biallelic SNVs. The remaining variants consisted of Indels and multiallelic sites (Table 20). After variant filtration to keep only high quality biallelic SNVs, this number reduced to 26,247,559. The complete variant calling results are presented in Table 20.

Statistic	Original base data	Biallelic SNVs	Filtered biallelic SNVs
Number of records	43,243,663	36,541,726	26,247,559
Number of SNVs	37,682,631	36,541,726	26,247,559
Number of indels	5,897,230	0	0
Number of multiallelic sites	2,296,557	0	0
Number of multiallelic SNV sites	785,862	0	0
Ti/Tv	1.93	2.01	2.29

Table 20: Multisample variant calling results

Total autosomal variants: 25,513,085				
GQ	**Total genotyping rate**	**Variants removed because genotype was not present in all samples**	**Variants removed due to minor allele threshold (0.05)**	**Variants passing filters and QC**
10	0.955805	22,091,661	820,877	2,600,547
20	0.886553	25,442,450	5,681	64,954
30	0.782666	25,506,247	207	6,631
40	0.680179	25,508,230	97	4,758
50	0.646665	25,509,272	90	3,723

Table 21: PLINK statistics generated to decide on a genotype quality (GQ) threshold

4.3.2 Population differentiation and structure

PCA was performed on the 25,513,085 autosomal SNVs. A variety of genotype qualities were tested (from 10 to 50) in order to choose the optimal threshold to select variants per breed. From Table 21, it is seen that at GQ 10 (expected 10% error rate in genotype calling), only 2,600,547 variants were left after all the filter and quality control (QC) thresholds were applied. The variant number gets reduced to just 64,954 at GQ 20 (1% error rate), which drops to 6,631 at GQ 30 (0.1% error rate). Hence, GQ 20 seemed to be an optimal score in terms of maintaining enough variants with good quality genotypes across all animals. These 64,954 variants were therefore used to understand the genetic structure amongst the different breeds using PCA. The PCA plot attained from the first two principal components, which explains most of the variance in the data, is presented in Figure 24.

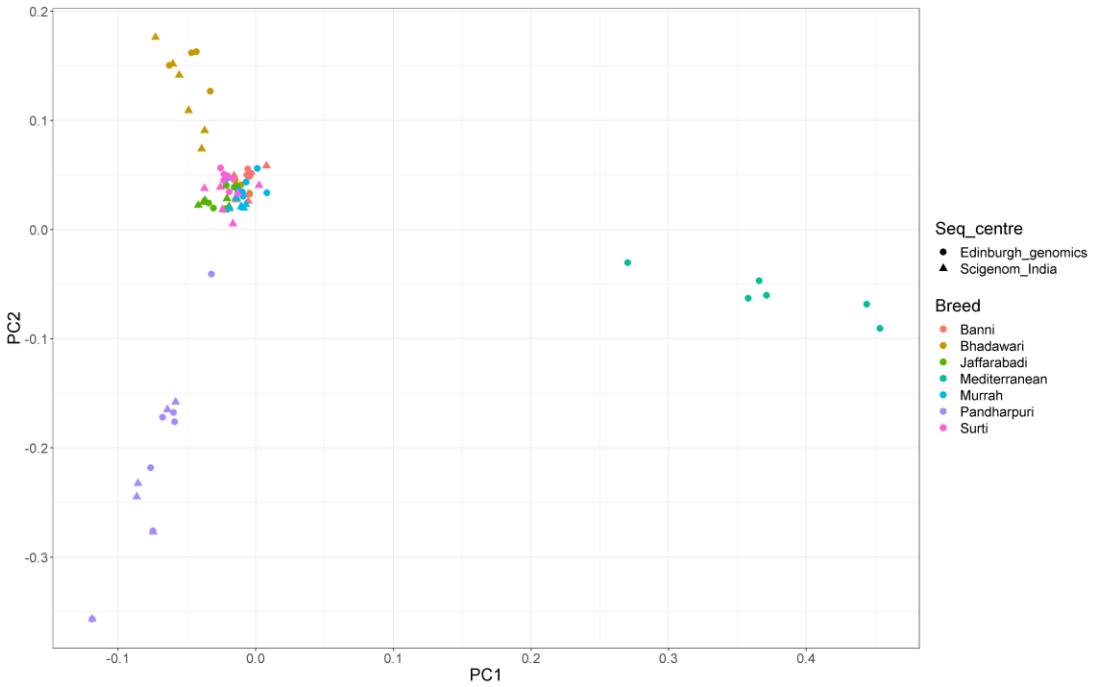

Figure 24: Principal Component (PC) analysis, PC1 versus PC2. The plot explains the breed differences based on 64,954 variants from 81 animals from 7 breeds. The points in the plot represent each animal, with different colours denoting their respective breed and the shape of the points denoting their respective sequencing centre.

From Figure 24, it can be seen that the first principal component separates the Mediterranean breed from the Indian breeds. On the second principal component, the Pandharpuri and Bhadawari breeds separate, whereas the remaining breeds cluster together. This suggests that they are either very closely related or the animals were collected from near borders of Indian states where these breeds are prominent, and crossbreeding may have occurred. The plot also suggests that Indian breeds are, as expected, genetically very different from the Mediterranean breed. Sequencing centres have been indicated in Figure 24. There is no evidence of a sequencing centre batch effect.

To get a clearer view of the structure among Indian breeds, the Mediterranean breed animals were removed, PCs recalculated and the PCA graph replotted (Figure 25). The Bhadawari and Pandharpuri breeds clearly form separate clusters, whereas the Jaffarabadi, Surti, Banni and Murrah breeds are more closely-related to each other. Murrah is a very important milk producing breed from north India and is the recommended breed for grading-up (breeding system

used to improve unproductive breeds) of other local breeds in Punjab, Haryana and Uttar Pradesh (Sreenivas, 2013).

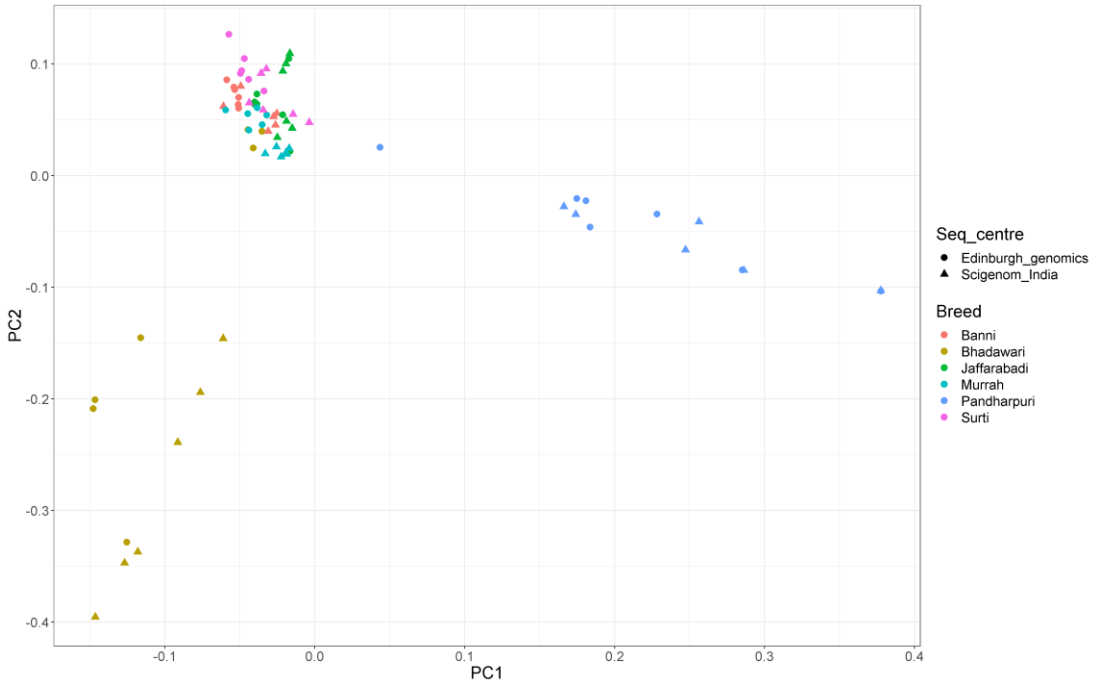

Figure 25: PC analysis after excluding the Mediterranean animals.

From Figure 25, it is seen that some Bhadawari animals lie close to the Murrah cluster. This may imply cases of crossbreeding between Bhadawari and Murrah for grading-up Bhadawari, a common practice in India (Das AK, 2008). Shortage of breeding bulls and low milk production was the main reason for this crossbreeding and grading up using the Murrah breed, a policy undertaken by the State Government of Uttar Pradesh, India (Kushwaha, et al., 2007). In Figure 25, some Bhadawari animals form a distinct cluster, a similar trend seen before in the maternal lineage of Bhadawari animals in a pairwise F_{st} analysis of eight Indian river buffalo breeds (Kumar, et al., 2007). The Pandharpuri animals form a distinct cluster in both Figure 24 and Figure 25 suggesting they are genetically different from other breeds. The Pandharpuri breed has also been identified to have its own lineage in a genetic variation study among eight Indian water buffalo breeds using 27 microsatellite loci (Kumar, et al., 2006).

Figure 26 shows the violin plots (combination of box plot and density plot) made to test for the presence of batch effects. The figure shows the relationship of PC

to sequencing centre. To understand if any of our principal components (1 to 15 in this case) correlate to the sequencing centre, violin plots were plotted from PC1 to PC15. No PC was correlated to sequencing centre.

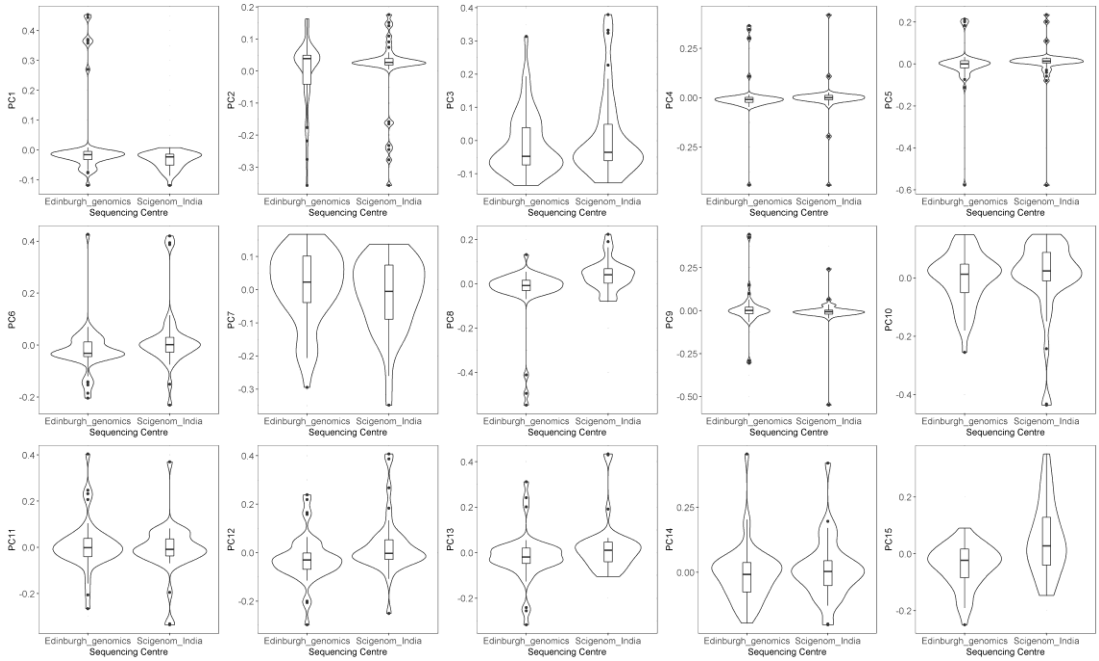

Figure 26: Violin plots showing that none of the PCs (from 1 to 15) was observed to be strongly correlated to sequencing centre as the median of both the sequencing centres of all the 15 cases are approximately the same

On closer examination, it was found that Pandharpuri-M240_10x overlapped with Pandharpuri_2_30x, and Pandharpuri-M256_10x overlapped with Pandharpuri_6_30x. This suggested that two animals from the Pandharpuri breed may have been sequenced twice. To confirm this, a relatedness study was done Using VCFtools v0.1.13. The scoring criteria used to assign relationship amongst pair of individuals has been mentioned in Table 22 and is based on the method by (Manichaikul, et al., 2010). The relatedness score is based on the 'kinship coefficient', which can be defined as the probability that two alleles - where one allele has come from one individual and the other allele has come from another individual - are identical by descent (Gillespie, 1998). Figure 27 is the heatmap that was plotted showing autosomal relatedness amongst all 81 animals. Different breeds have been represented with different colours. This analysis also supports the view that two animals have likely been

sequenced twice. Hence, from this step onwards, two samples were discarded- Pandharpuri-M240_10x and Pandharpuri-M256_10x. The related samples with higher depth of 30x were kept. This reduced our total sample number from 81 to 79.

Relationship	Relatedness score
Duplicate samples/Monozygotic twins	> 0.354
1st degree relatives	0.177–0.354
2nd degree relative	0.0884–0.177
3rd degree relatives	0.0442–0.0884
Unrelated samples	< 0.0442

Table 22: Scoring criteria for relatedness amongst pair of individuals

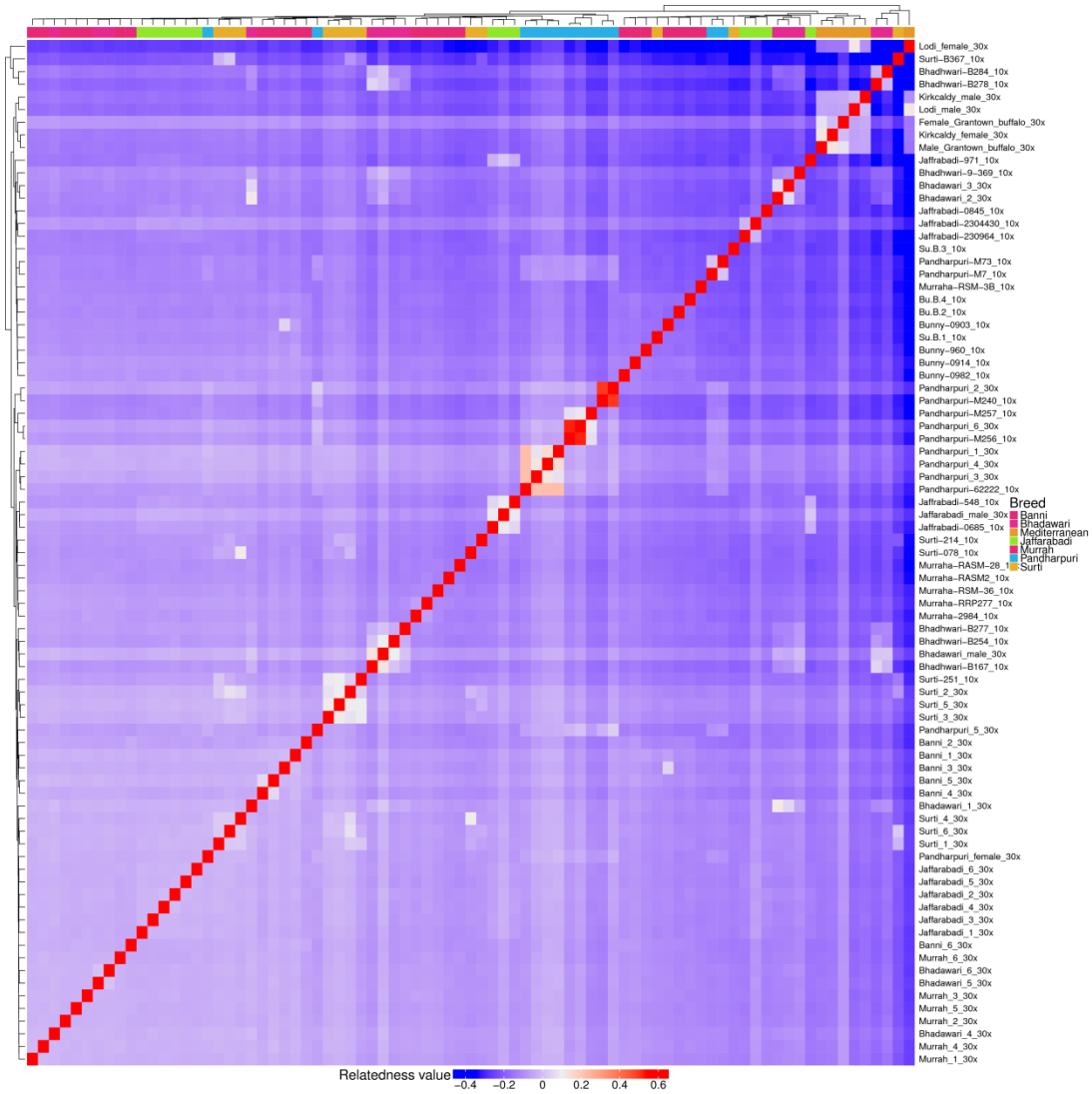

Figure 27: Heatmap showing autosomal relatedness amongst the 81 water buffalo. This heatmap mainly shows that two animals of one breed have probably been sequenced twice (two big red squares)

To understand the genetic admixture amongst the breeds, Admixture v1.3.0 was used on approx. 74,007 filtered biallelic SNPs from 79 animals that were obtained using PLINK. K was tested in the range from 2 to 7, where K is defined as the assumed number of ancestral populations (Figure 29). To determine the most appropriate value of K, CV (cross-validation) error was used (Alexander, et al., 2009). For this analysis, the CV error continued to increase as the value of K was increased as shown in Figure 28.

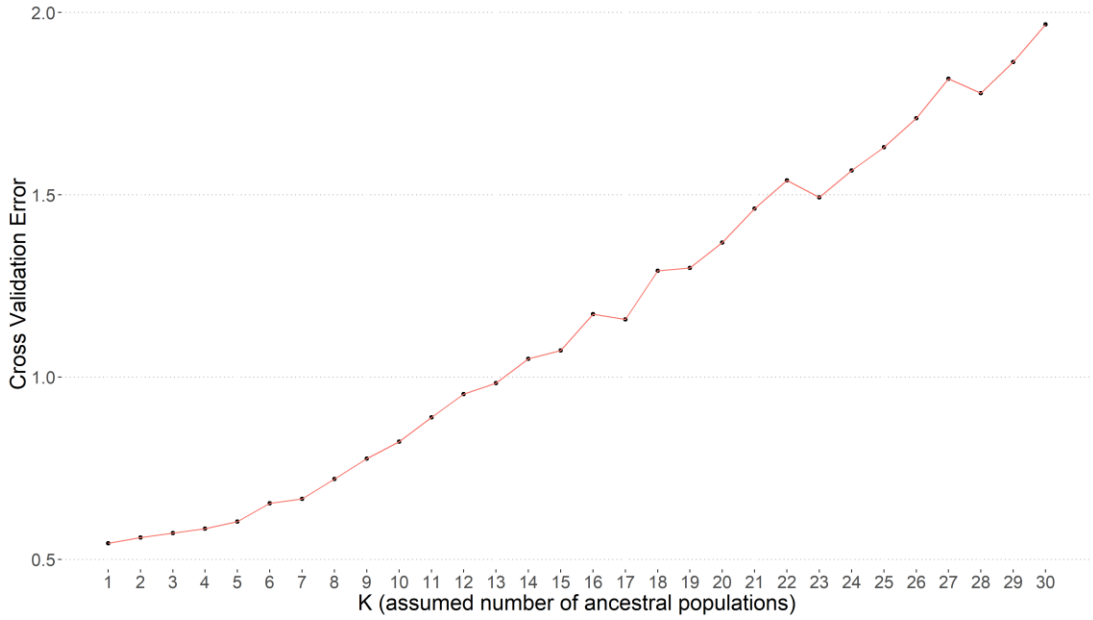

Figure 28: Cross-Validation Error values for different admixture models (K =1 to 30)

The lowest CV error was found at K=2. The analysis suggested that at K=2, there were two genetically distinct groups within our samples, consistent with the PCA analysis. In Figure 29, the mixture of colours per row in the plots suggests that an individual has ancestry from more than one subpopulation. This indicates that some amount of gene flow or introgression has happened between the Mediterranean and Indian breeds. This also reflects the fact that the Indian breeds have separated from the Mediterranean breed many years ago and show limited inter-breeding. The introgression is consistent with previous analysis (Kumar, et al., 2007) indicating the common origins of Mediterranean and Indian domestic river buffalo. A similar geographical analysis of the water buffalo was done very recently to determine the post-domestication migration route. This analysis also suggested that after domestication took place in the Indian subcontinent, the river water buffalo populations spread to south-west Asia and then entered Europe through a probable migration wave (Colli, et al., 2018). At K=3, Indian breed heterogeneity is revealed. The Bhadawari breed and Pandharpuri breed also appears separate at the ancestral level showing genetic heterogeneity with other Indian breeds. Murrah, Bhadawari, Banni and Surti share genomic features (show genetic admixture). In another analysis involving eight river buffalo breeds from India, genetic admixture between

Bhadawari, Surti and Murrah was observed (Kumar, et al., 2006). Banni was not a part of that study. Separation of Bhadawari and Pandharpuri breeds, but clustering of few Bhadawari animals was observed in the PCA in Figure 25 as well. At K=4, Banni, Jaffarabadi, Murrah and Surti separate off from the other breeds and show a greater level of admixture amongst them. Bhadawari and Pandharpuri also separate along with Mediterranean. This was also seen in the PCA analysis from Figure 24 where Pandharpuri, Bhadawari and Mediterranean separate from the rest of the breeds that cluster together, showing a greater level of admixture. As K is increased further, more heterogeneity is revealed within the Indian breeds.

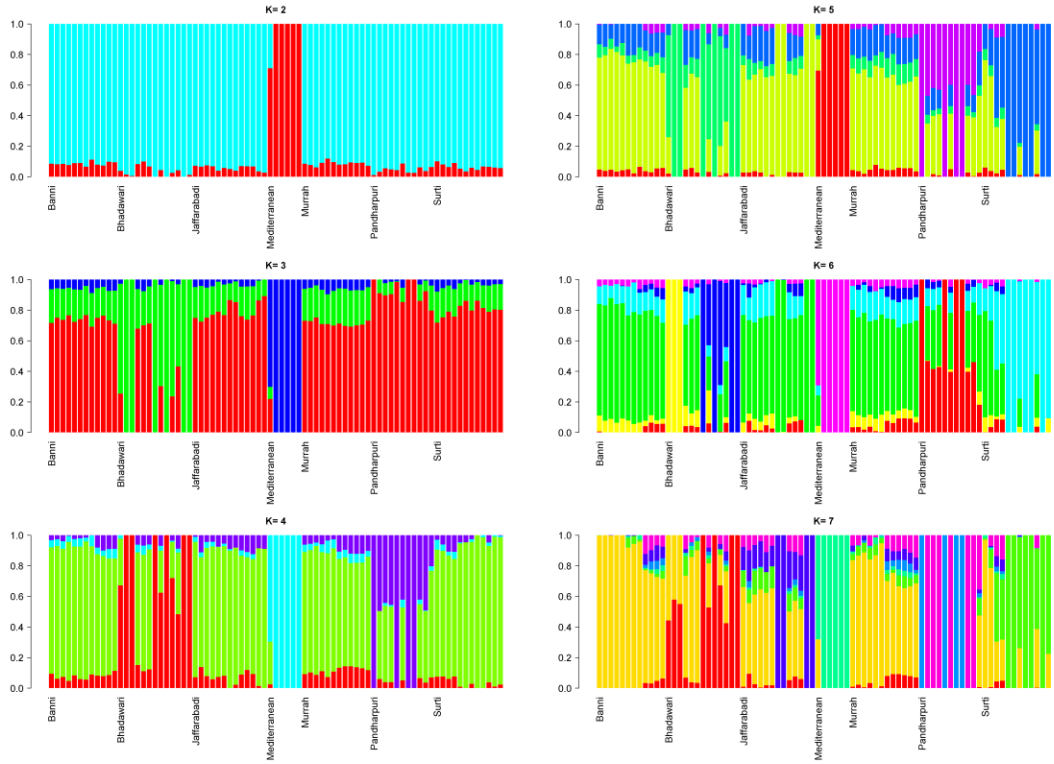

Figure 29: Ancestral proportion of each breed assuming different number of ancestral populations (K= 2 to 7). Each vertical line represents each individual's genome from the corresponding population, the white lines separating each individual. The colours in each vertical line represents the ancestry proportion i.e. the percentage of an individual's genomic data that was inherited from one of the seven ancestral populations present in the complete genomic dataset. The admixture analysis is based on 74,007 biallelic SNPs from 79 animals.

The F_{st} analysis measures the degree of inter-population diversity, like the PCA and admixture analysis, also indicates that the Mediterranean breed is most distinct from the Indian breeds. Figure 30 shows the heatmap constructed using weighted F_{st} values calculated using VCFtools v1.3.0 on 79 samples. High F_{st} values for the Mediterranean breed indicate comparative isolation from the Indian breeds and little inter-breeding. Amongst the Indian breeds, the Pandharpuri breed and Bhadawari breed seem to be relatively distinct from other Indian breeds with moderately high F_{st} values. Murrah, Banni, Jaffarabadi and Surti appear genetically nearer to each other compared to other breeds.

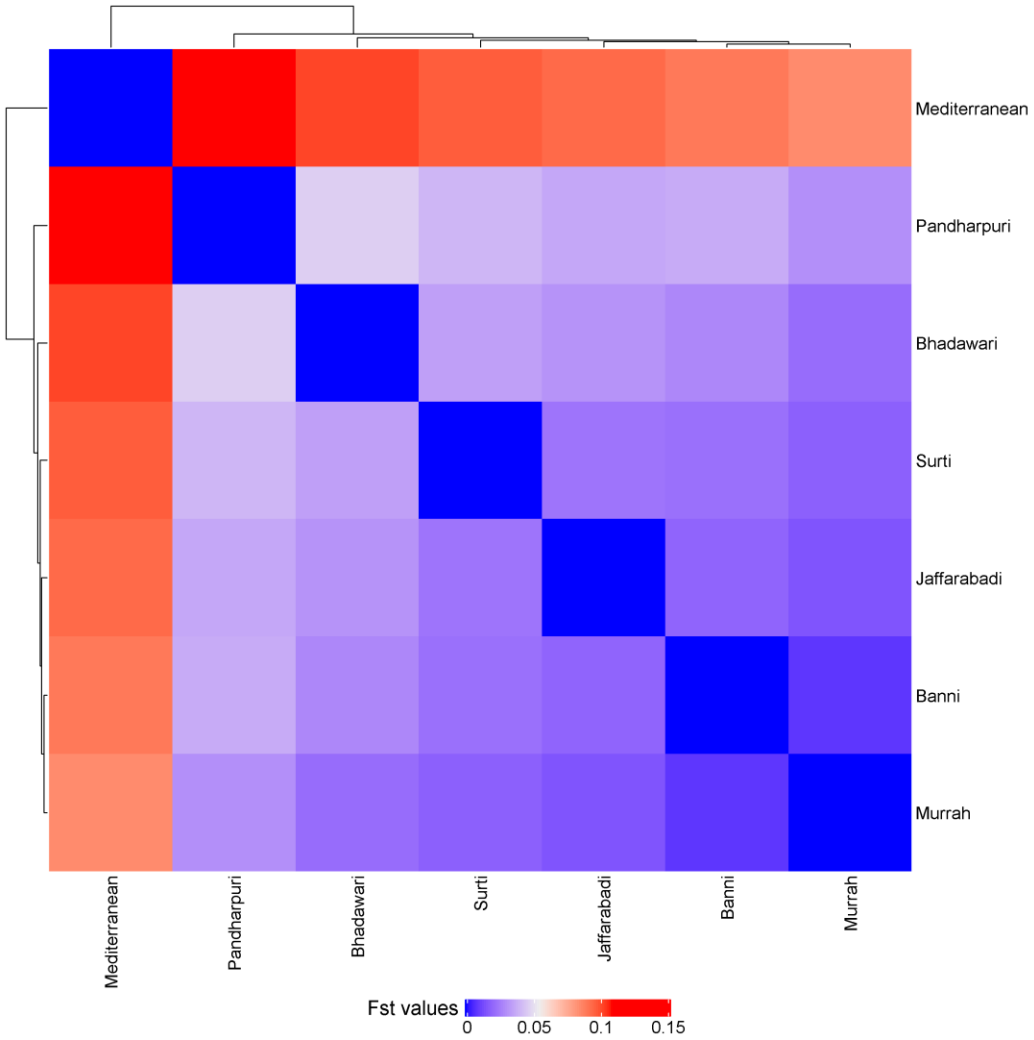

Figure 30: Heatmap showing the extent of genetic distance amongst different breeds of water buffalo. F_{st} values are shown where lower values (towards 0 and the colour blue) indicate high levels of inter-breeding whereas higher values (towards 1 and towards red) indicate more isolated populations.

4.3.3 Inferring population split and migration events

PLINK processing of filtered biallelic SNVs produced 74,007 SNPs from 79 individuals, which were used in the TreeMix v1.13 software (Pickrell and Pritchard, 2012). The software relies on a drift based evolutionary model and utilises allele frequencies to study relationships amongst populations under study and constructs maximum likelihood trees to reveal any existing population history and admixture events. This analysis does not model population history explicitly, but should be considered as a *post hoc* analysis (statistical analysis done after knowing about the data) to see if the tree represents actual population history and admixture events (Pickrell and Pritchard, 2012). All seven breeds were used in this study. Figure 31 shows the maximum likelihood tree without any migration events.

All the analysis above is consistent with evidence that the Mediterranean and Indian breeds have a common lineage, but have been largely isolated (Kumar, et al., 2007). Keeping the Mediterranean breed as the 'outgroup', all Indian breeds separate away from the Mediterranean breed. The algorithm made two major clusters, keeping Mediterranean breeds in one and other Indian breeds grouped into another. The no migration model explained 99.26% of variance. The horizontal branch length is proportional to the amount of genetic drift experienced in the population. Figure 31 shows that the Mediterranean population has experienced more genetic drift compared to Indian breeds. The Italian Mediterranean buffalo population has been inferred to have been largely isolated from when it was introduced in Italy from Africa or Europe in the sixth or seventh century (Cockett and Kole, 2008), except for some instances. For example, Italian Mediterranean buffalo were taken to Egypt to improve the milk production capacity of Egyptian buffalo (Nasr, 2017).

Banni and Murrah seem to have diverged from a common lineage. The two breeds also have similarities in morphology and production potential (Mishra, et al., 2009). However, F_{st} analysis involving only Banni and Murrah breeds (Mishra, et al., 2009) estimated a divergence time of around 7,286 years. The F_{st} analysis and TreeMix analysis here is consistent with this divergence. The

two breeds are nevertheless more genetically similar to each other than to the other breeds.

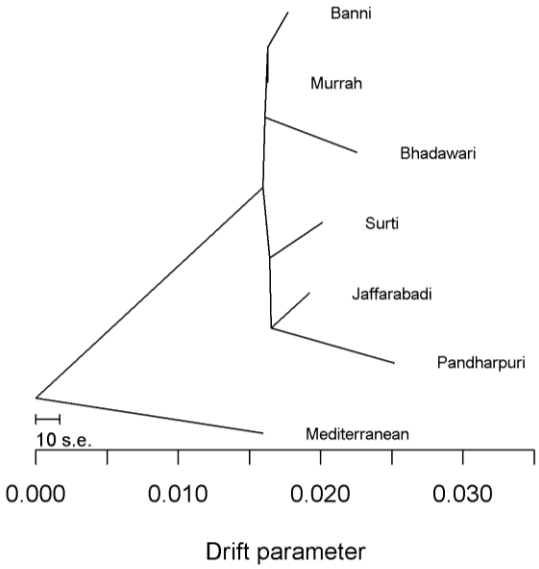

Figure 31: TreeMix maximum likelihood tree inferred from 7 breeds without any migration events. The length of the branch is proportional to the drift of each population. The scale bar indicates 10 times the average standard error of the relatedness among populations.

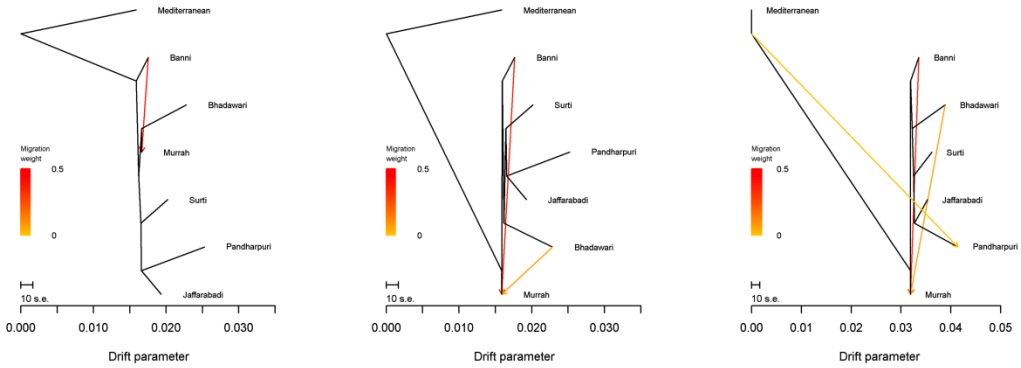

Figure 32: TreeMix maximum likelihood trees for three migration scenarios where m or migration event is 1, 2 and 3 (number of migrations equal to the number of arrows in the figure) and migration arrows are coloured according to their weights. The migration weight represents the fraction of ancestry derived from the migration edge. The direction of the arrows represents the direction of the gene flow from the migrant population to the recipient breed and the colour denotes the amount of mixture percentage.

To model the gene flow events between populations, various levels of migration were added to the model (Figure 32). At the first migration event, it was seen that there has been gene flow from Banni to Murrah suggesting interbreeding in

the past. At the second migration event, introgression was seen to take place from Banni to Murrah and Bhadawari to Murrah. This is possible as crossbreeding between Bhadawari and Murrah is a common practice as a process of grading-up (Kushwaha, et al., 2007). However, the gene flow amount is low and a lower migration weight is assigned to the gene flow arrow. As we advanced to the third migration event, gene flow was modelled between the common ancestor of Mediterranean and Indian breed group and Pandharpuri breed. The variance explained for the three migration events was 99.5%, 99.5% and 99.7% respectively.

4.3.4 Identification of signatures of putative selective sweeps between populations using pairwise XP-EHH analysis

The goal of this analysis was to identify evidence for signatures of positive selection between water buffalo breeds and to identify sites which may have possible importance in adaptation and productivity. PLINK processing resulted in 18,358,331 biallelic variants from 79 samples which were then imputed and phased. This VCF file was then separated into 7 new VCF files consisting of individuals from the same breed, which were then used in pairs (21 unique pairs in total) to seek evidence of selection signatures based on the pairwise XP-EHH test. The XP-EHH absolute Z-scores obtained from 21 breed pairs have been plotted in the form of a heatmap (Figure 33).

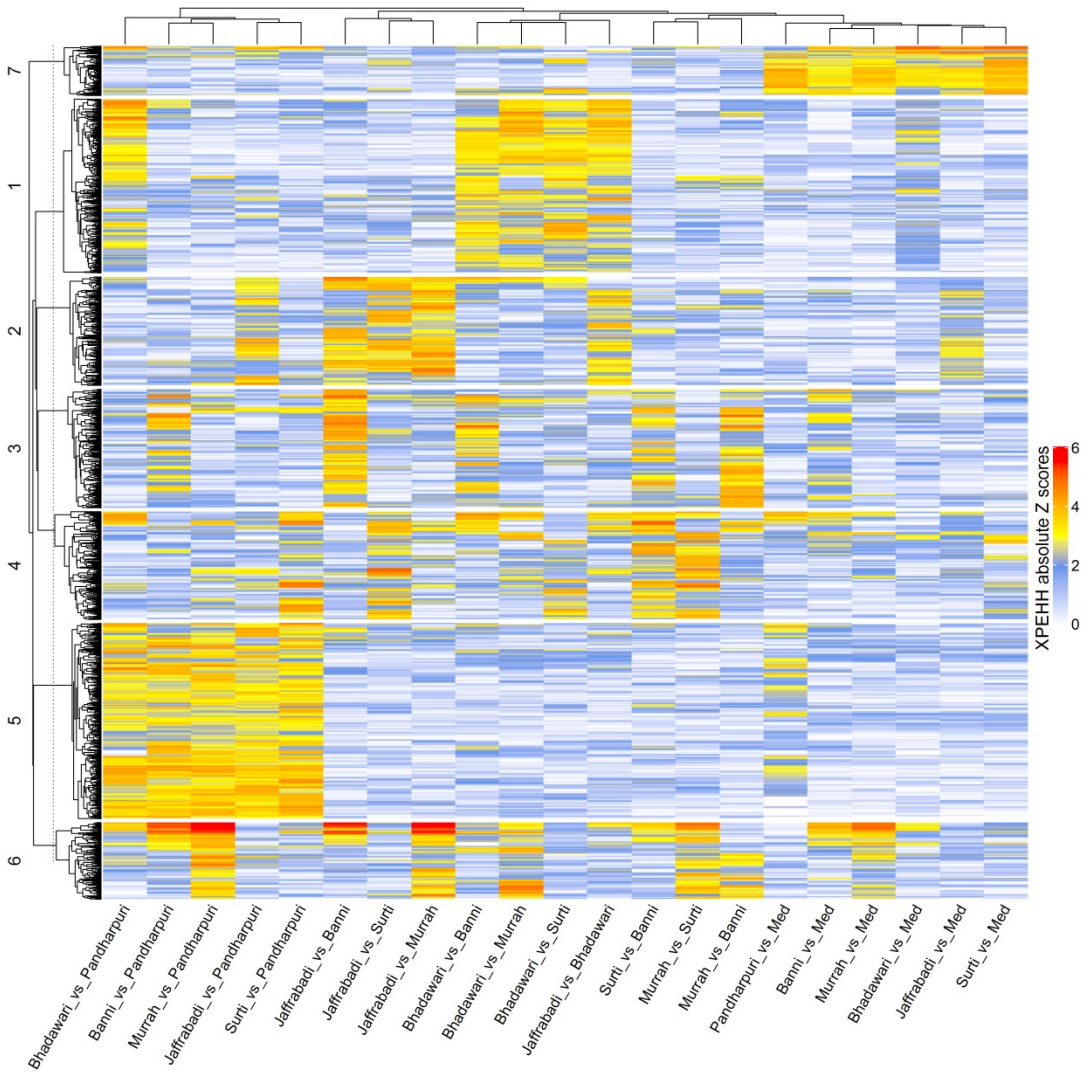

Figure 33: Heatmap showing XP-EHH absolute Z-score distribution amongst 21 breed pairs. Rows correspond to genomic loci. The score ranges from 0 to 6. Higher the score (towards red), greater the evidence of lineage specific adaptation between the breeds at the locus. The k-means clusters from 1 to 7 are shown wherein each cluster is largely a breed specific cluster. Cluster 1 corresponds to Bhadawari, 2 - Jaffarabadi, 3 - Banni, 4 - Surti, 5 - Pandharpuri, 6 - Murrah and 7 - Mediterranean

In Figure 33 , the clusters formed in the heatmap largely reflected breed-specific groups of sites of putative adaptation. The primary breed associated with each cluster is indicated in Table 23. The most interesting regions in the heatmap are the SNVs that have very high absolute XP-EHH Z-scores (indicated in red). Those SNVs are a part of the region in the genome that shows signatures of selection between breeds. Out of 21 pairwise XP-EHH combinations, 15

combinations had significantly high peaks with respect to other regions in the chromosome (FDR corrected P values <= 0.01). The peaks signify strong signatures of selection, where the focal locus (SNV with the highest XP-EHH absolute Z-score) and the surrounding region is linked to a putative selective sweep. Table 24 summarises the 1 Mbp region surrounding the focal SNV for the 15 candidate regions. An example of a Manhattan plot along with its zoomed in image of the peak has been shown in Figure 34 and Figure 35 respectively. In combination, the 15 regions containing apparent signatures of breed-specific selection contained 179 genes. When restricting to just those with official gene names, 105 genes remained.

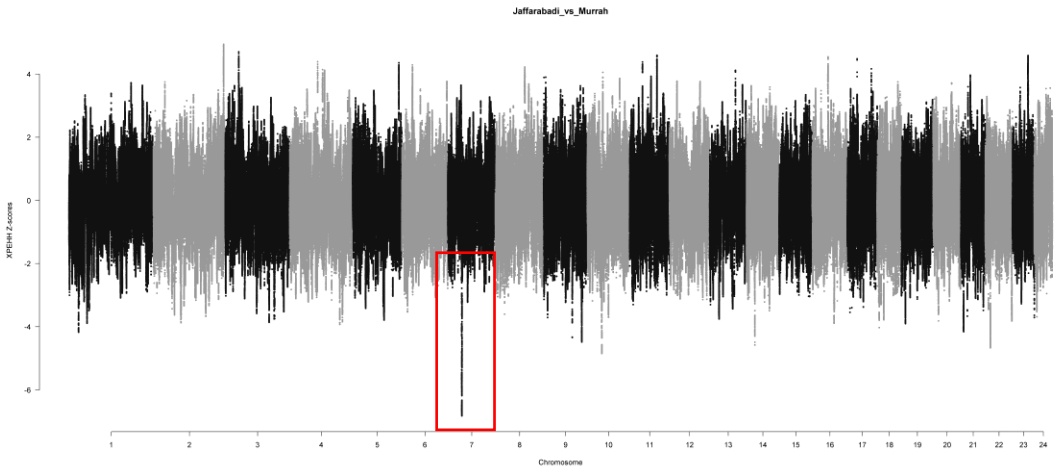

Figure 34: Genome wide Manhattan plot of standardised XP-EHH Z-scores calculated between Jaffarabadi and Murrah breeds showing region of high XP-EHH Z-score (inside red rectangle)

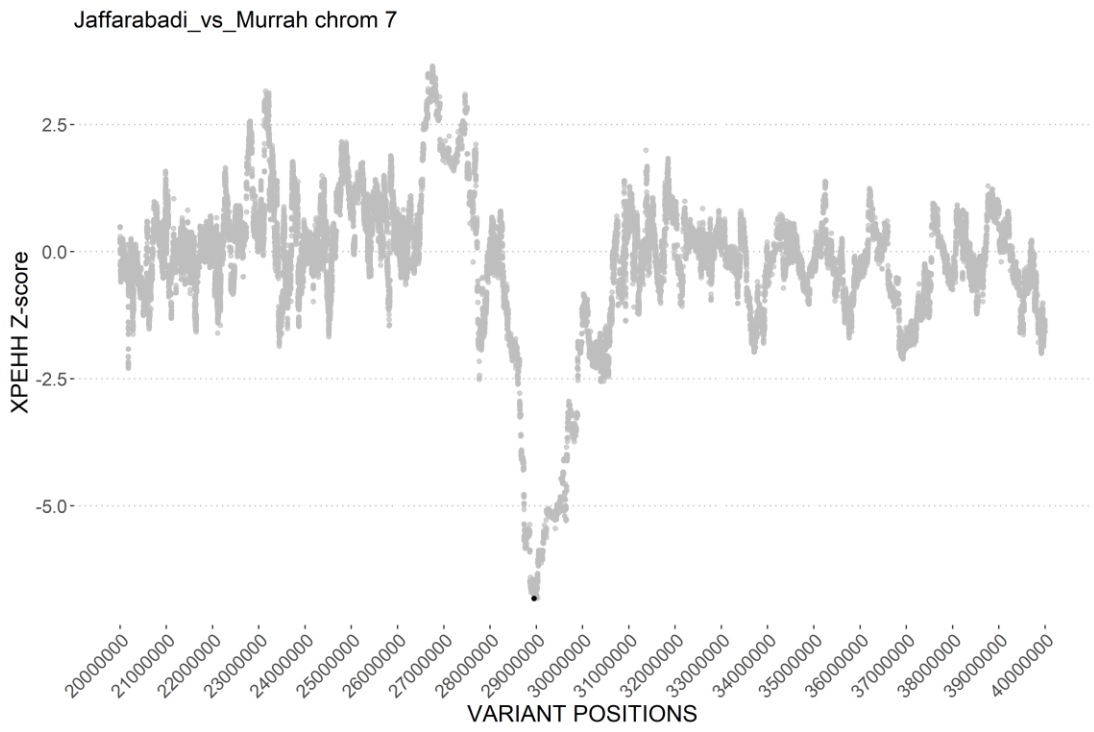

Figure 35: A zoomed-in image of Figure 34 showing regions that have undergone putative selective sweeps in the Murrah breed (inside red rectangle). The black dot at the tip of the cone pointing downwards is the SNV that has the highest absolute XP-EHH Z-score and is the focal SNV.

Cluster number	Associated breed
1	Bhadawari
2	Jaffarabadi
3	Banni
4	Surti
5	Pandharpuri
6	Murrah
7	Mediterranean

Table 23: Breed specific clusters found after k-means clustering performed during heatmap generation of XP-EHH absolute Z-scores from 21 breed pairs amongst which XP-EHH was calculated

4.3.5 Gene set enrichment analysis and functional annotation of genes present in the selective sweep region

The 105 genes obtained from the XP-EHH test were subject to a gene set enrichment analysis through DAVID v6.8. For this, corresponding human gene Ensembl IDs was obtained for 105 genes. Four transfer-RNA genes did not

have their corresponding IDs- *TRNAK-CUU*, *TRNAC-GCA*, *TRNAG-CCC* and *TRNAE-CUC*. *TRNAG-CCC* appeared twice. So, the final list that was given as an input to DAVID was 99 Ensembl IDs corresponding to the 99 official gene names. The human Ensembl IDs corresponding to the 15,938 genes with orthologues in *Bubalus bubalis* were used as the background list. The functional annotation of the genes was summarised on the basis of four annotation categories: molecular function, biological process and cellular component and KEGG pathways. The functional analysis chart (Figure 36) provides the output of the overrepresentation analysis done by DAVID identifying the most relevant biological terms (Gene ontology or GO) associated with our gene list (Huang da, et al., 2009b). The p-value associated with each GO terms is BH adjusted (Benjamini and Hochberg, 1995) controlling for the false discovery rate. Only 'CXCR chemokine receptor binding' was observed to be under 0.05 FDR with three genes in it. No one pathway was strongly linked across these sites.

Category	Term	RT	Genes	Count	%	P-Value	Benjamini
GOTERM_MF_DIRECT	CXCR chemokine receptor binding	RT		3	3.0	2.4E-4	4.5E-2
GOTERM_BP_DIRECT	male gonad development	RT		5	5.1	2.1E-3	7.3E-1
GOTERM_MF_DIRECT	chemokine activity	RT		3	3.0	1.2E-2	6.8E-1
GOTERM_BP_DIRECT	negative regulation of neuron death	RT		3	3.0	2.5E-2	1.0E0
GOTERM_BP_DIRECT	chemokine-mediated signaling pathway	RT		3	3.0	3.4E-2	1.0E0
GOTERM_MF_DIRECT	GPI anchor binding	RT		2	2.0	3.8E-2	9.1E-1
KEGG_PATHWAY	Cytokine-cytokine receptor interaction	RT		5	5.1	4.1E-2	9.9E-1
GOTERM_BP_DIRECT	DNA repair	RT		5	5.1	4.7E-2	1.0E0
GOTERM_MF_DIRECT	poly(A) RNA binding	RT		12	12.1	5.1E-2	9.2E-1
GOTERM_CC_DIRECT	ESCRT III complex	RT		2	2.0	6.6E-2	1.0E0
GOTERM_BP_DIRECT	defense response to bacterium	RT		3	3.0	7.5E-2	1.0E0
GOTERM_CC_DIRECT	DNA replication factor A complex	RT		2	2.0	7.7E-2	1.0E0
GOTERM_MF_DIRECT	metallopeptidase activity	RT		3	3.0	7.9E-2	9.6E-1
GOTERM_BP_DIRECT	positive regulation of leukocyte chemotaxis	RT		2	2.0	7.9E-2	1.0E0
GOTERM_CC_DIRECT	nucleus	RT		35	35.4	8.2E-2	9.9E-1
GOTERM_BP_DIRECT	regulation of mitotic spindle assembly	RT		2	2.0	8.5E-2	1.0E0
GOTERM_BP_DIRECT	vacuolar transport	RT		2	2.0	8.5E-2	1.0E0
GOTERM_BP_DIRECT	mitotic G2 DNA damage checkpoint	RT		2	2.0	9.0E-2	1.0E0
GOTERM_BP_DIRECT	chondrocyte development	RT		2	2.0	9.0E-2	1.0E0
GOTERM_CC_DIRECT	catalytic step 2 spliceosome	RT		3	3.0	9.5E-2	9.8E-1
GOTERM_BP_DIRECT	2-oxoglutarate metabolic process	RT		2	2.0	9.6E-2	1.0E0
KEGG_PATHWAY	Chemokine signaling pathway	RT		4	4.0	9.9E-2	1.0E0
GOTERM_BP_DIRECT	negative regulation of apoptotic process	RT		6	6.1	9.9E-2	1.0E0

Figure 36: Functional annotation chart from DAVID v6.8 of 99 genes based on background gene list of 15,938 genes

4.3.6 Identifying candidate regions in breeds under putative positive selection from pairwise XP-EHH analysis and their biological relevance

During a selective sweep, the ancestral allele is generally replaced by the new derived or beneficial allele so that its prevalence rises in the population. The set of genes with candidate regions is presented in Table 24. There is most likely a single variant in each region that has been positively-selected with remaining genes hitchhiking on the selective sweep. Out of the 15 regions, 3 plots have been generated for an in-depth visualisation of regions under selection. They have been specifically chosen as each of them contain a gene that has been reported to be under selection in previous studies. The plots are present as supplementary figures - Figure S 7, Figure S 8 and Figure S 9.

Aside from selection traits, it is very likely that the breeds were adapted to different environments and developed resistance to pathogens. Indian water buffalo in general live in tropical and sub-tropical weather in harsh environmental conditions and are exposed to a variety of infections and diseases such as mastitis (Dhakal, 2006), Brucellosis (Jain, et al., 2011), bovine tuberculosis (Srinivasan and Easterling, 2018), foot and mouth disease (Maddur, et al., 2009), Johne's disease (Singh, et al., 2008), *Pasteurella multocida* infection (Sethi, et al., 2011), *Peste des petits ruminants* virus infection (Dhanasekaran, et al., 2014), and tick infection (Miranpuri, 1988). Innate immune system has an important role to play in combating such infections. This section discusses plausible candidate genes in each selected region that could have driven the selection. The candidate genes present in the candidate regions have been mentioned in Table 24.

Pairwise Breed Comparison	Chr	Start	End	Genes with trait
Murrah and Pandharpuri	2	136,435,459	137,435,460	***DNAH7* (Fertility)**, ***SLC39A10* (Fertility)**
Jaffarabadi and Mediterranean, Bhadawari and Mediterranean, Surti and Mediterranean	4	72,245,983	73,246,027	***GRIP1* (Fertility)**, ***HELB* (Fertility)**, ***HMGA2* (Body size)**, ***IRAK3* (Immunity)**, ***LLPH* (Growth)**, *LOC102396915*, *LOC102398844*, *LOC112584844*, ***TMBIM4* (Growth)**, *TRNAK-CUU*
Surti and Banni, Surti and Pandharpuri	4	100,861,550	102,060,235	***DUSP6* (Growth)**, ***KITLG* (coat colour)**, *LOC102391483*, *LOC102393435*, *LOC102393752*, *LOC102395105*, *LOC112584637*, *LOC112584638*, *LOC112584639*, *LOC112584640*, *LOC112584641*, *LOC112584793*, *POC1B*, *TRNAG-CCC*
Jaffarabadi and Banni	6	28,142,486	29,142,487	***AMPD1* (Growth/Meat)**, *BCAS2*, *CSDE1*, *DENND2C*, *LOC102401117*, *LOC102401745*, *LOC112585585*, *LOC112585586*, *LOC112585841*, ***NRAS* (Growth/Meat)**, ***SIKE1* (Immunity)**, *SYCP1* , ***SYT6* (fertility)**, ***TRIM33* (Immunity)**
Banni and Pandharpuri, Jaffarabadi and Banni, Jaffarabadi and Murrah, Murrah and Mediterranean, Murrah and Pandharpuri, Murrah and Surti	7	28,452,839	29,820,124	***AFM* (Dairy)**, ***AFP* (Fertility)**, ***ALB* (climatic adaptation)**, ***ANKRD17* (Immunity)**, ***CXCL6* (Immunity)**, ***CXCL8* (Immunity)**, *LOC102395897*, *LOC102396414*, *LOC102411166*, *LOC102413257*, *LOC102414909*, *LOC102415247*, *LOC102415563*, *LOC102415899*, *LOC102416225*, *LOC112586113*, *LOC112586168*, *LOC112586170*, *LOC112586293*, ***MTHFD2L* (Growth)**, ***PPBP* (Immunity)**, ***RASSF6* (Growth)**
Bhadawari and Banni	7	105,036,356	106,036,357	***BMPR1B* (Fertility)**, *HPGDS*, *LOC102408403*, *LOC102415909*, ***PDLIM5* (Meat)**, *SMARCAD1*
Jaffarabadi and Murrah	10	35,381,663	36,381,664	*ARHGAP18*, *L3MBTL3*, *LOC102408911*, *LOC102412861*, *SAMD3*, *TMEM200A*, *TMEM244*
Bhadawari and Murrah	10	55,178,713	56,178,714	*ASCC3*, ***GRIK2* (tameness)**, *LOC112587467*, *LOC112587469*, *LOC112587470*, *LOC112587668*
Banni and Pandharpuri, Murrah and Banni	11	1,370,186	2,385,225	*EFCAB11*, *FOXN3*, *KCNK13*, *LOC102396186*, *LOC112587755*, *NRDE2*, *PSMC1*, *TDP1*, *TRNAC-GCA*
Bhadawari and Murrah	14	19,409,983	20,409,984	***ACSS2* (Dairy)**, ***AHCY* (Coat colour)**, ***ASIP* (Coat colour)**, *CHMP4B*, *DYNLRB1*, *EIF2S2*,

				GGT7, ***ITCH* (Coat colour)**, *LOC102392477*, *LOC102392793*, *LOC102410343*, *LOC102411041*, *LOC102412315*, *LOC112578730*, *LOC112578732*, *LOC112578733*, *LOC112578981*, *LOC112579002*, *MAP1LC3A*, ***NCOA6* (Dairy)**, *PIGU*, ***RALY* (Coat colour)**, *TP53INP2*, *TRNAG-CCC*, *ZNF341*
Bhadawari and Banni, Bhadawari and Murrah, Murrah and Surti, Surti and Banni	18	13,908,598	14,908,828	***ANKRD11* (Growth)**, *C18H16orf87*, *CDK10*, *CENPBD1*, *CHMP1A*, *CPNE7*, *DBNDD1*, *DEF8*, *DPEP1*, *FANCA*, ***GAS8* (Meat/Fertility)**, ***GPT2* (Growth)**, *LOC102394579*, *LOC102395986*, *LOC102400019*, *LOC102402068*, *LOC102403160*, *LOC112580223*, *LOC112580225*, *LOC112580226*, *LOC112580227*, *LOC112580478*, *LOC112580495*, ***MYLK3* (Meat)**, *ORC6*, *RPL13*, ***SHCBP1* (Meat)**, *SPATA2L*, ***SPATA33* (Fertility)**, *SPG7*, *SPIRE2*, *TCF25*, ***VPS35* (Growth)**, *VPS9D1*, *ZNF276*
Murrah and Surti, Surti and Pandharpuri	19	69,321,400	70,456,448	***IRX2* (Dairy)**, *LOC102392162*, *LOC112580653*, *LOC112580664*
Jaffarabadi and Surti	19	31,705,724	32,705,725	*C19H5orf51*, *FBXO4*, ***GHR* (Dairy)**, ***OXCT1* (Dairy)**
Jaffarabadi and Banni	19	6,078,842	7,078,843	*ENC1*, ***FAM169A* (Dairy)**, *GFM2*, *LOC102398516*, *LOC102398732*, *LOC102407744*, ***MSX2* (Dairy)**, *NSA2*
Bhadawari and Pandharpuri	23	44,355,173	45,355,174	***ADAM12* (Growth/Meat)**, *BCCIP*, ***DHX32* (Immunity)**, ***EDRF1* (Dairy)**, ***FANK1* (fertility)**, *LOC102410615*, *LOC102413592*, *LOC112581580*, *LOC112581600*, *LOC112581646*, *LOC112581676*, ***TEX36* (fertility)**, *TRNAE-CUC*, *UROS*

Table 24: Candidate genes present in the putative selective sweep candidate regions found through pairwise XP-EHH analysis between breeds. One of these genes is under selection in the region that caused a selective sweep. Candidate genes that could be connected to a production/disease resistance/phenotypic trait have been given a different colour than the rest of the genes. The traits have been written in brackets next to the gene name itself and their functions have been described in the main text.

4.3.6.1 2:136435459-137435460

Murrah and Pandharpuri breed pair showed putative selection in a candidate region on chromosome 2 which had two fertility/reproduction related genes-

DNAH7 and *SLC39A10*. The protein encoded by *DNAH7* is involved in the formation of the sperm flagellum during spermatogenesis (Horowitz, et al., 2005). Protein encoded by *SLC39A10* is a solute carrier that transports zinc that is important for oocyte development (Hester and Diaz, 2018; Lisle, et al., 2013).

4.3.6.2 4:72245983-73246027

Putative selection was observed in a candidate region on chromosome 4 between three breed pairs - Jaffarabadi and Mediterranean, Bhadawari and Mediterranean, and Surti and Mediterranean - and it seems that the selection may have taken place in Mediterranean specific region. The SNV having the highest absolute Z-score is located within the *HELB* locus. *HELB* encodes a DNA helicase that is involved in cell cycle progression and helps to mitigate replication stress (Guler, et al., 2012). The gene has also been associated with inhibin regulation affecting Sertoli cell (testicular cells) proliferation and testicular size in domestic cattle (Fortes, et al., 2013). The adjacent *GRIP1* locus encodes an estrogen receptor cofactor that has been reported to have reproductive functions, as *Grip1* knockout mice had impaired fertility in both sexes (Gehin, et al., 2002). The gene product is also used as an oestrus detection marker for predicting the ovulation time in cattle (Lee, et al., 2017). In India, selective breeding is employed in non-descript local buffalo breeds using bulls of high milk producing breeds such as Murrah, Surti, Bhadawari and Jaffarabadi (Sreenivas, 2013). As artificial insemination is extensively used in India for selective breeding, care is taken to get the best quality semen (Singh and Balhara, 2016). In a study involving Murrah, a significant correlation was also found between scrotal circumference with semen volume and sperm concentration wherein a breeding bull can be selected by taking into reference the size of scrotal circumference (Dabas, et al., 2010). Vaginal fluid is used to detect estrous cycle in cows, a practice also followed in buffalo in India (Selvam and Archunan, 2017). However, detecting estrous behaviour is difficult in water buffalo as they exhibit silent estrous (Awasthi, et al., 2007).

IRAK3, also known as *IRAK-M* (Wesche, et al., 1999), is an immune-specific gene that is involved in regulating TLR and IL-1 signalling pathways. It has been

shown that *IRAK3* negatively regulates TLR signalling when stimulated by TLR to prevent extra inflammation and tissue damage (Kobayashi, et al., 2002). Some other genes that came up were genes involved in development such as *LLPH* and *TMBIM4* and *HMGA2*. *LLPH* is a gene that is involved in neuronal growth and synaptic transmission (Yu, et al., 2016). It was previously found to be positively selected in domestic cattle (Xu, et al., 2015). It was also detected to be one of the genes present in candidate regions for 'stayability' (probability of a cow giving birth at a certain age) in Nellore cattle (Barreto Amaral Teixeira, et al., 2017). *TMBIM4* regulates Ca^{2+} levels and fluxes conferring resistance to apoptotic stimuli (Carrara, et al., 2017). *HMGA2* is another regulatory protein related to cell growth and has been associated with bone development (Kuipers, et al., 2009) and height (Yang, et al., 2010) in humans and has also been found to be positively selected in Russian cattle breeds (Yurchenko, et al., 2018).

4.3.6.3 4:100861550-102060235

Two breed pairs - Surti and Banni, and Surti and Pandharpuri - showed selection in another putative candidate region in chromosome 4. The candidate region contained *KITLG* (*KIT*-ligand or mast cell growth factor) and *DUSP6* along with other genes which are not annotated. *KITLG* is a gene which is linked to coat colour and is responsible for roan coat colour in goats (white hairs intermixed with pigmented hairs) (Talenti, et al., 2017) and cattle (Seitz, et al., 1999) and six-white-point colour in pigs (white coat on the four feet, the head, and the end of tail but with the remaining coat being black) (Lü, et al., 2016). *KITLG*, along with its receptor c-KIT triggers the MAPK signalling pathway that plays an important role in controlling skin pigmentation (Picardo and Cardinali, 2011). The gene along with its receptor regulates melanocyte development and melanin synthesis (Picardo and Cardinali, 2011) and variations related to the gene lead to variable coat colour phenotypes, as in the case of pigs mentioned above. Surti seems to have peculiar white marking on its coat, one around the jaw and another on the brisket, whereas its coat colour is black (Kumar Yadav, 2017). *DUSP6* on the other hand has an important role in muscle growth and has been reported to be a candidate gene regulator for body mass as it

negatively regulates the MAPK pathway which is the central pathway for cell growth and differentiation and its inhibition leads to increase in myoblast proliferation leading to muscle growth (Vo, et al., 2019).

4.3.6.4 6:28142486-29142487

Between Jaffarabadi and Banni, chromosome 6 had a candidate region that seemed to be under putative selection and the SNV with the highest absolute Z-score was present in an intergenic region with no gene associated to it. *TRIM33* is a gene in the candidate region that has been reported to regulate the proinflammatory function of T helper 17 cells by inducing IL-17 production. This in turn activates a signal cascade recruiting many chemokines which recruit monocytes and neutrophils to the actual site of infection (Tanaka, et al., 2018). *SIKE1* or *SIKE* is another gene in the candidate region that has a proposed immune related regulatory role. In the mammalian system, when a virus enters a cell through the process of endocytosis, it is recognised by Toll-like receptor 3 (*TLR3*) that activates Interferon regulatory factor 3 (*IRF-3)* and *NF-κB*. This in-turn leads to the production of type 1 interferons (IFNs) which activate the immune cells that finally upregulate the host defence against the virus. Activation of *IRF-3* requires its phosphorylation by two serine-threonine kinases i.e., TBK1 and IKKi, and *SIKE1* is an inhibitor of their interaction with *IRF-3* leading to the disruption of TLR3-mediated induction of type I IFNs (Huang, et al., 2005). Two other genes in the candidate region were *AMPD1* and *NRAS* which have been present as signatures of selection in goats and associated with meat production traits due to their link with skeletal muscles (Bertolini, et al., 2018). *SYT6* or Synaptotagmin VI is a fertility related gene that participates in sperm's acrosomal exocytosis, which is the secretion of sperm content or male chromatin from the sperm's head (acrosome) into the egg's cytoplasm for the fertilisation process to occur (Michaut, et al., 2001).

4.3.6.5 7:28452839-29820124

A candidate region in chromosome 7 showed putative selection in six breed pairs: Banni and Pandharpuri; Jaffarabadi and Banni; Jaffarabadi and Murrah;

Murrah and Mediterranean; Murrah and Pandharpuri; Murrah and Surti. In this scenario, since directionality cannot be assigned, it is possible that this region is specific to the Murrah and Banni breeds between which we observed evidence for gene flow in the Treemix and Admixture analyses before. We would predict *a priori* that genes related to defence mechanism or immune related genes will be under selective pressure, as breeds were selected in different environments with different pathogen challenges (Nielsen, et al., 2007). There is some evidence to support this inference within breed-specific candidate regions. This candidate region contains many immune and development genes. *ANKRD17* is an important gene related to innate immunity due to its involvement in the pattern recognition receptors NOD-1 and NOD-2 mediated responses in humans (Menning and Kufer, 2013). The gene product has been reported to act against viral response where it interacts with virus-induced signalling adaptor protein which in-turn upregulates the RLR or RIG-I-like receptor-mediated immune signalling pathway and finally enhances the transcription of the cytokine IFN-β (Wang, et al., 2012). The SNV related to *ANKRD17* also had the highest XP-EHH absolute Z-score in the pairwise XP-EHH combination of Murrah and Mediterranean breed.

CXCL6, *CXCL8* (*IL-8*) and *PPBP* or *CXCL7* are chemokines (signalling proteins) that are pro-inflammatory in nature and recruit immune cells to the site of infection by the process of chemotaxis (Rollins, 1997). *CXCL6* is expressed by macrophages during an infection and is important in activating neutrophils that provide a first-line innate immune defence by phagocytosing, killing, and digesting bacteria and fungi (Linge, et al., 2008; Segal, 2005). *CXCL8* is another chemokine produced by macrophages that produces an acute inflammatory response by recruiting and activating neutrophils (Harada, et al., 1994). Macrophages are likely to be the first cells to release *CXCL8* upon an infection (Arango Duque and Descoteaux, 2014). *CXCL8* has been associated with a number of diseases in humans such as tuberculosis (Selvaraj, et al., 2006), gastric cancer (Gonzalez-Hormazabal, et al., 2018) and upper respiratory tract infection (Zehsaz, 2015). *CXCL8* has been seen to possibly enhance innate immunity response in human TB patients by directly binding to tubercle bacilli

leading to augmentation of leukocytes and macrophages for phagocytosis (Krupa, et al., 2015). Based on a very recent meta-analysis (Srinivasan and Easterling, 2018) which was done using already reported prevalence data for bovine TB in cows and buffalo in India, it was found that Haryana (where the Murrah breed is present) had lesser prevalence of bovine TB (3.3%) than Uttar Pradesh (where Bhadawari breed belongs) (6.5%). *CXCL8* was found to be the candidate gene under putative selection in Murrah in comparison to Bhadawari in the pairwise XP-EHH analysis. Such association involving disease prevalence is an important way to determine the relevance of putative candidate genes under selection because if an important resistance trait (such as *CXCL8*) limits a pathogen's growth and reproduction, it reduces the transmission of the pathogen further, which in turn, leads to the reduction of disease prevalence in the population (Horns and Hood, 2012).

PPBP which is also called Leukocyte-Derived Growth Factor (*LDGF*) or *CXCL7*, can act as an inflammation stimulator and takes part in controlling regeneration of connective tissue during the repair response and hence helps in wound healing (Iida, et al., 1996). *CXCL7* has been shown to play a major role in guiding leukocytes (white blood cells) to the site of injury mediated by blood platelets as a part of the innate immune response (Ghasemzadeh, et al., 2013). *RASSF6* is involved in cellular functions such as regulating the cell's cytoskeleton structure, inter-cellular interactions and overexpression of this gene induces cellular apoptosis (Iwasa, et al., 2018). *RASSF6* contained one of the twelve QTLs causing variation in mammary gland morphology in German Fleckvieh cattle (Pausch, et al., 2016). AFP is another protein which has estrogen binding capacity that may have a role in embryonic development (Terentiev and Moldogazieva, 2013). *MTHFD2L* is an important developmental gene involved in carbon metabolism and *de novo* purine biosynthesis which has been seen to have high expression in mouse embryos (Shin, et al., 2014). AFM protein shows high binding affinity to Vitamin E and is its major carrier in body fluids (Voegele, et al., 2002). Vitamin E is an important constituent of milk and also has immune and antioxidant functions (Borel, et al., 2013). The antioxidant defence system is a very important biochemical mechanism for countering toxic

free radicals from damaging cells. Many animals have this in order to counter environmental factors over which they do not have control, such as hibernation, migration, food availability, temperature, humidity and salinity (Chainy, et al., 2016). ALB is an important plasma protein in mammals which acts as a transporter molecule for many nutrients, metal ions and metabolites. Due to its ligand-binding efficiency, it also acts as a radical scavenger with antioxidant properties (Fasano, et al., 2005). ALB may allow the Murrah or Banni breed to adapt to the hot and humid tropical environment in India as compared to the Mediterranean breed in Italy. However, it has been reported that productivity of the Murrah breed decreases due to its low heat tolerance, since it is black in colour and has low sweat gland density; young Murrah animals are not able to perform thermoregulation and need to be given external support for temperature maintenance (Vaidya, et al., 2012).

4.3.6.6 7:105036356-106036357

Bhadawari and Banni showed significant selection in chromosome 7 as well having 6 genes in the candidate region. The SNV which had the highest absolute Z-score belonged to the gene *PDLIM5*. This gene which is also known as *ENH1*, acts as a scaffold protein and has been observed to tether protein kinase C (PKC) to the Z-disk of striated muscle through the PDZ domain (Nakagawa, et al., 2000). Variants at this locus have been associated with carcass traits (fat thickness, intramuscular fat, etc.) in cattle (Hay and Roberts, 2018) and pig (Ma, et al., 2015). Intra-muscular fat content is an important factor for determining meat quality and palatability (Hay and Roberts, 2018). Another gene near it was *BMPR1B* which is a candidate gene related to reproduction trait linked to lambing in sheep (Tang, et al., 2018). It has been speculated that the gene influences the ovarian biological function, and follicular development and maturation. It was also reported that a mutation in that gene resulted in a higher litter size in Small Tail Han sheep (Tang, et al., 2018). Furthermore, a mutation in the first intron of *BMPR1B* containing the oestrogen response element (ERE), has been seen to be associated with pig fecundity (reproductive output of an individual in its lifetime) (Li, et al., 2017). The calving interval

(amount of time between the birth of a calf and the birth of the next calf) for Bhadawari breed of buffalo was reported to be 17.1±0.4 months or 522.1±12.1 days (B.P. Kushwaha, 2013), whereas for Banni it was 12.2±0.7 months or 371±21.2 days (Mishra, et al., 2011).

4.3.6.7 10:35381663-36381664

This candidate region showed putative selection in-between Jaffarabadi and Murrah, but none of the genes in this region could be connected to any trait.

4.3.6.8 10:55178713-56178714

A putative candidate region of selection in chromosome 10 was also observed between Bhadawari and Murrah breeds where the SNV with highest absolute Z-score was present between *LOC112587668* and *GRIK2*. *GRIK2* encodes for a protein that is a receptor of glutamate (an excitatory neurotransmitter in vertebrates) which is associated with tameness (reduction of aggressive and fearful response) in domestic animals and is seen to be upregulated in dogs, guinea pig, chicken and rabbit (Li, et al., 2014). The gene also showed up in signals of selection in dogs, rabbits and ducks (Rourke and Boeckx, 2018).

4.3.6.9 11:1370186-2385225

This candidate region was seen to be under putative selection in two breed pairs: Banni and Pandharpuri; Murrah and Banni. This seems to be a Banni specific region. None of the gene present in this region could be connected to any traits.

4.3.6.10 14:19409983-20409984

The breed pair Bhadawari and Murrah had a putative candidate region under selection in chromosome 14. The SNV which had the highest absolute Z-score was approximately 6000 bases upstream of the *AHCY* gene. Other genes nearby were *ITCH*, *ASIP* and *RALY*. Significant associations to coat colour have been found for *AHCY*, *ITCH*, *ASIP* and *RALY* genes for black and brown coat colour in Iranian Markhoz goats (Nazari-Ghadikolaei, et al., 2018). *ASIP* or Agouti-signalling peptide is the major contributor for brown or black coat colour,

but since the other three genes were in strong LD with *ASIP gene*, they could be playing a regulatory role for the trait (Nazari-Ghadikolaei, et al., 2018). In water buffalo, the *MC1R* (Melanocortin-1 receptor) gene is associated with the coat colour trait and an allele of the gene is associated with black coat colour (Miao, et al., 2010). The *MC1R* gene plays an important role in the synthesis of two melanins- eumelanin (brown-black colour) and phaeomelanin (red-yellow colour) and *ASIP* plays a role of an antagonist causing coat colour variation (Graham, et al., 1997). The protein encoded by *MC1R*, coupled with G-proteins, stimulate the production of eumelanin giving dark colours to monocytes. *ASIP* on the other hand acts as an antagonist to *MC1R* leading to the favourable production of phaeomelanin that gives light colour to melanocytes (Marín, et al., 2018). It is possible that this leads to the difference in coat colour in the two breeds; Murrah buffalo are jet black in colour whereas Bhadawari buffalo are copper coloured with reddish brown hair tips (https://www.roysfarm.com/bhadawari-buffalo/). Other genes present in the candidate region regulate milk composition traits - *ACSS2* (*de novo* fatty acid synthesis during lactation cycle in Chinese Holstein cattle) (Bionaz and Loor, 2008) and *NCOA6* (interacts with PPAR-gamma transcription factor that regulates bovine milk fat synthesis) (Olsen, et al., 2017). The Bhadawari breed is well known for its high butter fat content in milk (6 to 12.5%) (Saifi, et al., 2004) whereas fat content is 7.3% fat for Murrah per lactation (http://dairyknowledge.in/).

4.3.6.11 18:13908598-14908828

Chromosome 18's candidate region was seen to be putatively selected in four breed pairs: Bhadawari and Banni; Bhadawari and Murrah; Murrah and Surti; Surti and Banni. In all the combinations, the highest absolute Z-score was assigned to SNVs present in the intergenic region between *GAS8* and *SHCBP1*. Both *GAS8* and *SHCBP1* have been previously associated with variation in meat tenderness in Nellore cattle (Braz, et al., 2019); *GAS8* is involved in myogenesis i.e. muscle development during embryogenesis (Evron, et al., 2011) and *SHCBP1* is an important component of fibroblast growth factor (FGF) signalling, which is connected to skeletal muscle regeneration and development

(Pawlikowski, et al., 2017). *GAS8* is also a fertility related gene that encodes a testicular protein which contributes to sperm motility (Yeh, et al., 2002). Its expression is pronounced in human and mice during puberty and spermatogenesis, but is absent in infertile males (Yeh, et al., 2002). *MYLK3* or Myosin Light Chain Kinase 3 plays an important role in smooth muscle contraction (Gao, et al., 2001) and has been associated with meat quality in Nellore cattle (Rodrigues, et al., 2017). Higher phosphorylation of Myosin regulatory light chain 2 (MYLPF) by Ca^{2+} dependent myosin light chain kinase alters the structure and motor function of the myosin by contracting it, and ultimately renders beef tougher (Rodrigues, et al., 2017). *GPT2* is a growth related gene that is involved in fatty acid metabolism (Aagaard-Tillery, et al., 2008) and amino acid (arginine) biosynthesis (Marion, et al., 2013) and this gene has been associated with age related differential muscle growth regulation in pigs (Ayuso, et al., 2016). *VPS35* is also a growth related gene involved in skeletal muscle development. It promotes recycling of a cargo protein called GLUT1 from endosomes to the cell surface (Duchemin, et al., 2016). GLUT1 is a predominant glucose transporter in glycolytic skeletal muscles in ruminants such as cows, goats and camels (Duehlmeier, et al., 2007). *ORC6* is involved in cytokinesis (cell division) and thus plays a very important role in the cell cycle process (Prasanth, et al., 2002). *SPATA33* (Chen, et al., 2013) has been reported to play a potential developmental role in spermatogenesis suggesting that it is a fertility/reproduction related gene. *ANKRD11* was another growth related gene that has been associated with bone/skeletal development (Sirmaci, et al., 2011). Many genes were seen to be under the region of selective sweep governing meat quality trait, and amongst all the breeds involved in this region of putative selection, Murrah is considered to be the best breed for meat (and milk) amongst various buffalo breeds in India (Saifi, et al., 2004).

4.3.6.12 19:69321400-70456448

Two breed pairs showed putative selection in a candidate region in chromosome 19: Murrah and Surti; Surti and Pandharpuri, and potential selection was seen to have taken place in the Surti breed. The SNV that had the highest absolute Z-

score did not belong to any gene. However, *IRX2* was another gene in the candidate region whose expression leads to human mammary gland epithelial cell differentiation during duct and lobule development and lactation (Lewis, et al., 1999).

4.3.6.13 19:31705724-32705725

Jaffarabadi and Surti showed putative candidate region of selection on chromosome 2, consisting of four genes - *OXCT1*, *GHR*, *FBXO4*, and *C19H5orf51*. *OXCT1* is a gene that has already been associated with milk fatty acid trait in Chinese Holstein cows due to its involvement in lipid metabolism (Li, et al., 2014). *GHR* or growth hormone receptor is another very important milk related gene wherein a QTL was identified in the gene with a large effect on bovine milk yield and composition (Blott, et al., 2003). Variants associated with *GHR* were reported to be subjected to recent positive selection in dairy cattle (Flori, et al., 2009). *FBXO4*, and *C19H5orf51* could not be connected to any trait. Indian riverine buffalo breeds are artificially selected for milk, though there are differences in milk yield. The Jaffarabadi breed has a higher milk yield and milk fat content than the Surti breed (2,239 kg milk per lactation with a fat percentage of 7.7 versus 1667 kg milk per lactation with a fat percentage of 7.02 respectively) (http://www.dairyknowledge.in/). Furthermore, a milk yield and composition comparison study involving ten healthy Surti and Jaffarabadi buffalo (Janmeda, et al., 2017) revealed that Jaffarabadi buffalo produced significantly higher milk yields with a greater fat percentage as compared to the Surti breed.

4.3.6.14 19:6078842-7078843

Jaffarabadi and Banni breed pair showed a putative selection in a region present in chromosome 19 that consisted of eight genes with a SNV in *GFM2* with the highest absolute Z-score. The candidate region also contained *FAM169A* which is involved in the regulation of milk protein synthesis in dairy cattle. Its promoter region contains transcription factor binding sites for many transcription factors that regulate milk protein synthesis such as STAT5, GR and ER (Pegolo, et al., 2018). *MSX2* takes part in the branching morphogenesis in

the mammary gland (Satoh, et al., 2004) and is also expressed in its mesenchymal and epithelial cells (Hens and Wysolmerski, 2005).

4.3.6.15 23:44355173-45355174

Between Bhadawari and Pandharpuri, the SNV that had the highest absolute Z-score belonged to the gene *DHX32* in chromosome 23 in the candidate region under putative selection. The gene has been reported to have a role in lymphocyte (T-cells) differentiation (Abdelhaleem, et al., 2005), and hence play a role in adaptive immune response. Selection of this gene may lead to differences in the adaptive immune response to certain diseases between the two breeds. *EDRF1* was another gene near to *DHX32* that regulates the DNA-binding activity of the GATA-1 transcription factor (Wang, et al., 2002), whose binding site is present in the regulatory region of β-casein gene and β-casein is a major component in milk (Lee, et al., 2012). Putative selection of this gene may lead to differences in casein concentration in the milk of the two breeds. Information on the casein concentration of milk from the Bhadawari breed has been published (Boro, et al., 2018), but is unavailable for the Pandharpuri breed. Hence, the difference cannot be verified. However, caseins in milk vary in their quantities from breed to breed and animal to animal (Misra, et al., 2008). *FANK1* is another fertility related gene present in the candidate region that is anti-apoptotic in nature during both spermatogenesis and oogenesis and was found to be under balancing selection in humans (DeGiorgio, et al., 2014). *ADAM12* is a growth related gene that takes part in in the regulation of myogenesis and adipogenesis and is also considered as a good target for manipulating skeletal muscle development intramuscular fat (IMF) deposition in cattle for improving meat quality and beef yield (Coles, et al., 2014). Both breeds are mainly used for milk production, but are also used for draught power for which muscle growth is an important trait.

4.4 Conclusions

The XP-EHH analysis in-between populations revealed many candidate regions that showed signatures of selection and breed divergence. The genes present in the candidate region seemed to govern various phenotypic traits.

Body size/growth seemed to be a variable phenotypic trait under possible selection. This may be due to artificial selection by humans. Some water buffalo breeds have a heavier body and greater muscular growth than others, due to which they are used for draught purposes. Muscle growth can also cause differences in body weight; Surti has the lowest average body weight (435 kg) as compared to other breeds - Murrah (567 kg), Bhadawari (475 kg) and Banni (525-62 kg). The regions under selection contained genes related to muscle growth and meat production traits. Murrah is used for both meat and draught purposes, whereas other breeds are only used for milk and sometimes draught.

Milk yield and protein and fat content were variable phenotypic traits whose associated genes were present in the region under putative selection. The candidate regions contained genes that controlled the variation in fat content, protein content (such as casein) and total milk yield in different buffalo breeds.

Coat colour pattern is a variable trait and genes related to this trait were present in regions under putative selection leading to phenotypic differences in the coat colour between Murrah and Bhadawari, giving the former a jet black and the latter a reddish brown or copper colour. It also potentially gave Surti its peculiar white markings/patches on its body.

Water buffalo domestication led to the selection of tameness trait. We found a region under selection that contained a gene governing this trait.

In this study, some regions under putative selection contained many fertility/reproduction related genes. Reproduction (specifically spermatogenesis), apoptosis and tumour suppression related genes have been found to be under strong selection pressure (Nielsen, et al., 2005). Both male and female reproductive systems seem to be affected due to heat stress which leads to deleterious effects on the reproductive function of mammalian species such as

cattle (Takahashi, 2012). Indian water buffalo are exposed to high heat due to harsh tropical and sub-tropical temperatures. Heat stress negatively affects spermatogenesis in males and oogenesis, oocyte development and fertilisation in females (Takahashi, 2012). Fertility related genes were observed to be potentially under selection in many Indian water buffalo breeds that assist with a variety of fertility related activities such as spermatogenesis, oocyte development and maturation, anti-apoptosis during embryogenesis, etc.

Indian water buffalo are known for their exceptional disease resistance (Brahma, et al., 2015; Patel, et al., 2017). It has been reported that diseases (such as bovine pleuropneumonia, foot-and-mouth disease, etc.) are less prevalent in or less damaging to buffalo than cattle (Cockrill, 1981). Immune related genes came up in regions under putative selection. Immune genes show a greater degree of adaptive evolution than other genes in the genome (McTaggart, et al., 2012). In this study, the Murrah breed, which is the most diffused breed in the world and is proposed to have a strong immune system, was found to be under selection for some immune related genes (Patel, et al., 2015).

Finally, this chapter explored the population structure amongst Indian water buffalo breeds and the Mediterranean water buffalo. Signatures of selection were also explored in order to find candidate regions under selection in the water buffalo breeds. The selection signatures that were identified in this study reflected that water buffalo breeds have been subjected to natural selection pressure in response to adaptation to the local environment and artificial selection pressure for their milk/meat/draught related traits.

4.5 Supplementary Material

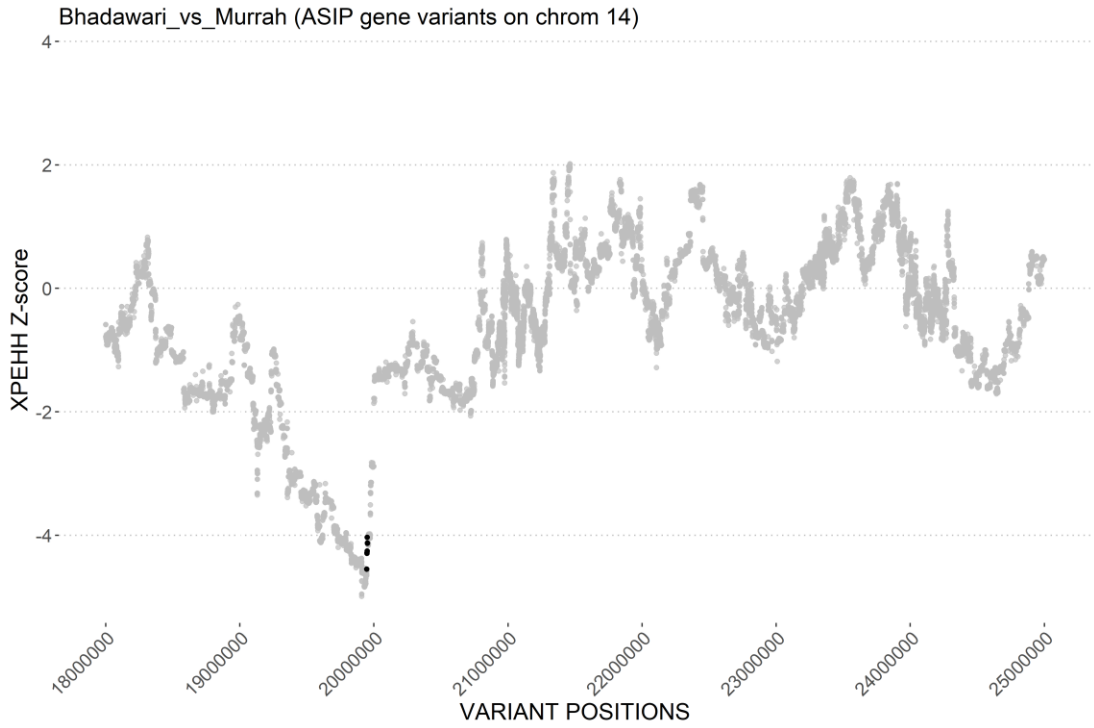

Figure S 7: A zoomed in image of chromosome 14 containing the *ASIP* gene. The black dots represent the SNVs present within the *ASIP* gene. The gene was found in the region under selection during an XP-EHH score comparison between the Bhadawari and Murrrah breeds of water buffalo. It is probable that SNVs within *ASIP* gene or a regulatory variant of the gene may be under selection due to which the loci around have hitchhiked showing signature of selection.

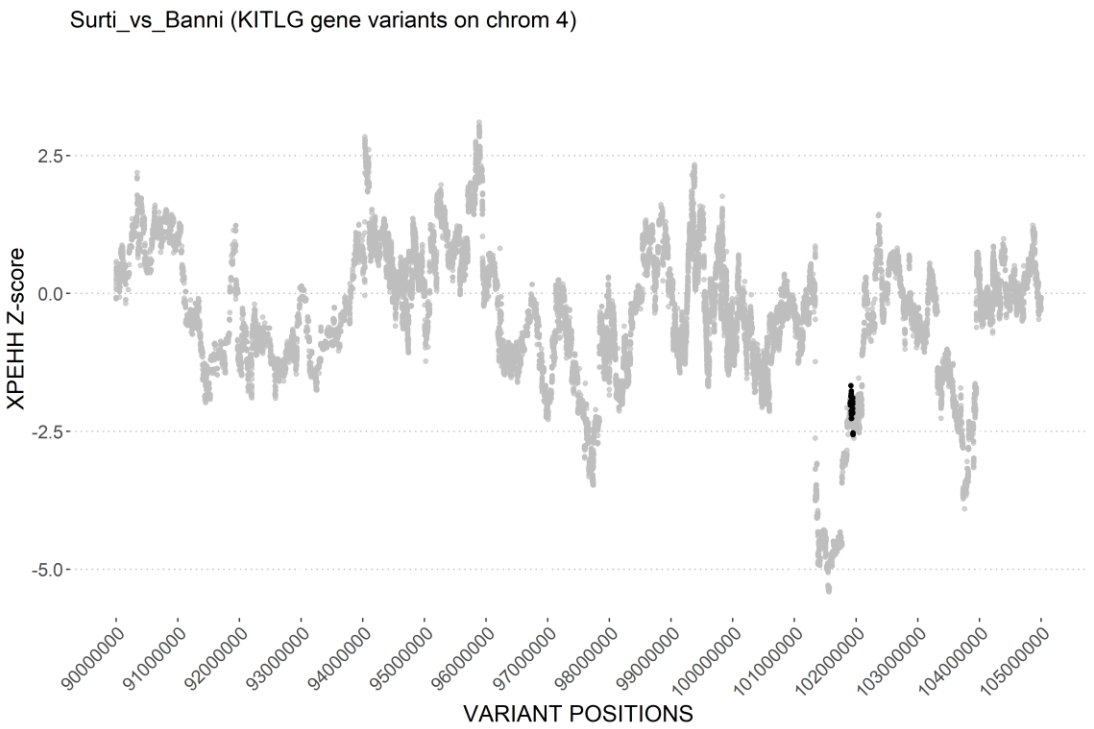

Figure S 8: A zoomed in image of chromosome 4 containing the *KITLG* gene. The black dots represent the SNVs present within the *KITLG* gene. The gene was found in the region under selection during an XP-EHH score comparison between the Surti and Banni breeds of water buffalo.

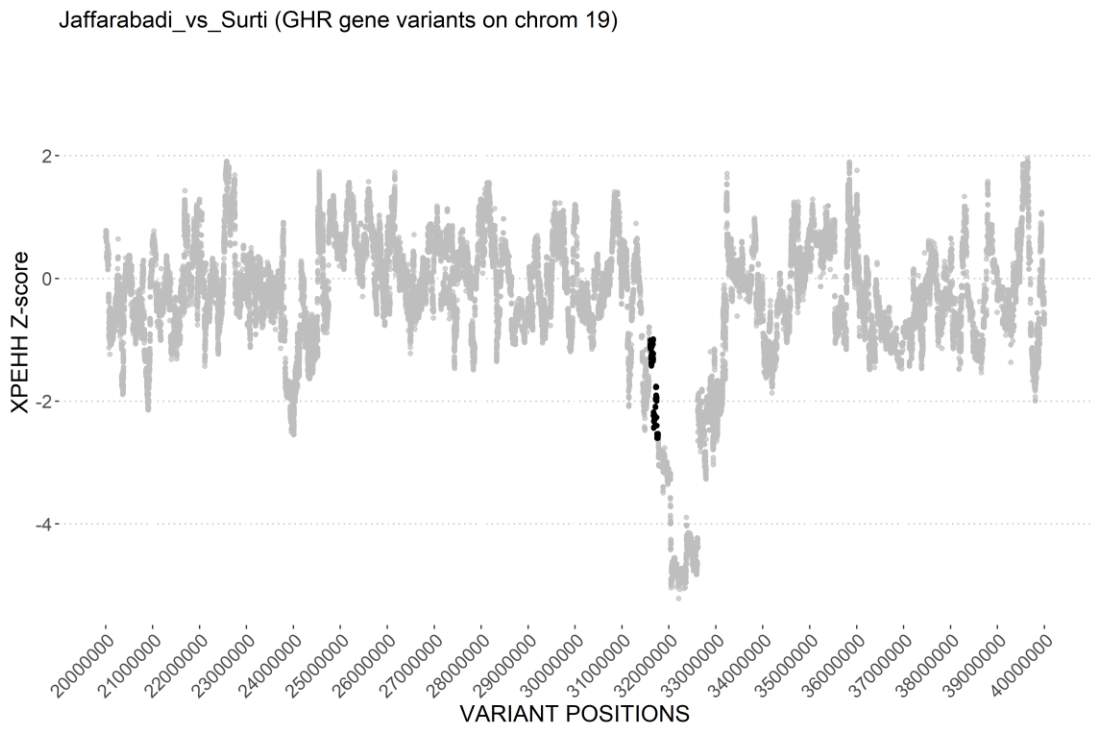

Figure S 9: A zoomed in image of chromosome 19 containing the *GHR* gene. The black dots represent the SNVs present within the *GHR* gene. The gene was found in the region under selection during an XP-EHH score comparison between the Jaffarabadi and Surti breeds of water buffalo.

Chapter 5. General Discussion

The current chapter provides a brief discussion of the major findings in this thesis and also discusses the scope of future research. In this thesis, I have explored the presence of regulatory variation in macrophages of water buffalo in the form of allele-specific expression (ASE) and investigated signatures of selection and breed divergence across water buffalo breeds. Water buffalo are an important livestock species in India and diseases affecting them impact the economy negatively.

Genetic variants play an important role in determining an individual's susceptibility to diseases. Variation in the coding region of a gene leading to non-synonymous or missense mutation may lead to functional differences in the protein. Genetic variation in the regulatory region of a gene, for example, *cis*-eQTLs in the gene's promoter region can also lead to variation in gene expression. This variation can be responsible for variation in gene expression between individuals and between different populations (Knight, 2005). Such regulatory variants in immune-related genes have been reported to cause differences in disease susceptibility in various infectious diseases such as hepatitis B virus (HBV) infection (Motavaf, et al., 2014), leprosy (Moraes, et al., 2004), Urinary Tract Infection (Ragnarsdóttir, et al., 2010), tuberculosis (Gao, et al., 2015), sepsis (He, et al., 2017), etc. Understanding the genetic basis of such regulatory variation will allow for a better understanding of inter-individual variation in immune response (Fairfax, et al., 2014).

The first results chapter (chapter 2) in this thesis presented findings from an assessment of the existence of regulatory variation in water buffalo immune-related genes, while the second results chapter (chapter 3) presented a genome-wide assessment of regulatory variation in the water buffalo. Chapter 2 involved the development of an *in silico* workflow/pipeline for ASE quantification using raw RNA-seq data and chapter 3 involved the integration of DNA-seq data from the same animals into the ASE analysis (along with RNA-seq data). Chapter 2 was like a "training exercise" for establishing the core *in silico* ASE

analysis and quantification pipeline. Chapter 3 integrates additional quality control steps to the ASE quantification to ensure that ASE analysis was robust.

This ASE study was neither a population-scale nor a tissue-wide analysis. The focus was on immune-related genes expressed specifically in macrophages, based upon the premise that these were most likely to be subject to selection affecting disease susceptibility. In a large study of human monocytes responding to LPS (lipopolysaccharide) or interferon-gamma, more than 80% of transcripts were found to exhibit heritable variation in their level of expression in at least one condition (Fairfax, et al., 2014). Macrophage samples contain many innate immune genes (along with several housekeeping genes) that serve as the first line of host defence against any pathogen attack. The water buffalo gene expression atlas was utilised in order to get a list of immune-related genes that might be associated with disease resistance traits (Young, et al., 2019). The analysis revealed the existence of regulatory variation in several important immune related genes including *NLRP2-like*, *TLR7, NOD1* and *NLRC4*. These genes are important pathogen-recognition receptor (PRR) genes that recognize microbial pathogens and initiate an immune response. ASE in these genes may explain variation in disease resistance in water buffalo. Since, variation in gene expression is often heritable (Yan, et al., 2002), effects of regulatory variation in immune-related genes such as differences in disease susceptibility will also be passed from generation to generation. The detected ASE was largely individual-specific so that across the set of four animals, a large percentage of transcripts for which there were informative expressed SNVs showed evidence of allelic imbalance.

In Chapter 3, DNA-seq data was used to attain high quality heterozygous genotypes to overcome the inherent limitations of RNA-seq data for SNV detection. The DNA-seq data allowed the identification of instances of monoallelic expression (MAE) in water buffalo where RNA-seq reads were expressed from only one allele. The relative proportions of ASE detected were broadly similar to that observed in a previous cattle study (Chamberlain, et al., 2015). ASE genes were also widespread in water buffalo autosomes. Condition-specific (+/- LPS), sex-specific (male/female) and individual-specific ASE were

observed. This was the first ASE study to be ever done in the water buffalo, the others being already done in other livestock such as cattle (Chamberlain, et al., 2015; Guillocheau, et al., 2019), sheep (Salavati, et al., 2019) , goats (Cao, et al., 2019), pigs (Ahn, et al., 2019; Stachowiak, et al., 2018) and chicken (Zhuo, et al., 2017).

Identification of genes showing ASE provides the opportunity to identify the causal variant from their respective DNA sequences which is responsible for the imbalance. In this thesis, ASE was examined only in the macrophage samples from the water buffalo gene expression atlas (Young, et al., 2019). The ASE study could be extended to the complete gene expression atlas that contains tissue samples from major organ system such as muscular, reproductive, nervous, cardiovascular, etc. to get a more comprehensive ASE profile. ASE is also tissue-specific in nature (Knight, 2005). The tissue samples present in the atlas gives the opportunity to evaluate the extent of tissue-specific ASE as done in cattle (Chamberlain, et al., 2015) and sheep (Salavati, et al., 2019) using the pipeline developed and optimised as a part of this thesis (Knight, 2005; Young, et al., 2019). Candidate immune genes showing regulatory variation may be shortlisted from this analysis and much larger candidate-gene association studies may be performed using case-control water buffalo in order to prioritise genes responsible for disease susceptibility (Jorgensen, et al., 2009).

Candidate-gene association studies are used to identify risk variants associated with a particular disease. In this approach, some candidate genes are selected which have been previously linked to a particular disease or condition being investigated. After this, the SNVs or SNPs that are selected are those that either have a functional consequence (they affect the protein product of the gene negatively) or those that are present in the regulatory region of the gene of interest. Then, the chosen variant is verified, based on a number of random test subjects who have the disease/condition and control subjects who are healthy, and the association of the variant with the disease/condition is evaluated. The knowledge gained from this approach is very valuable and it has the potential of being a disease diagnostic tool as well (Kwon and Goate, 2000; Patnala, et al., 2013). This method has been used to identify causative SNPs for bone diseases

(Dastgheib, et al., 2016), tuberculosis (Horne, et al., 2015; Zembrzuski, et al., 2010), brucellosis (Rossi, et al., 2019) and diabetes (Sobrin, et al., 2011).

Chapter 4 explored the presence of underlying genetic diversity among major breeds of Indian water buffalo namely Jaffarabadi, Pandharpuri, Banni, Surti, Murrah and Bhadawari. Their relationship with the Italian Mediterranean water buffalo was also explored. The population differentiation and structure analysis revealed a clear distinction between the Mediterranean breed from Indian water buffalo breeds. Within Indian breeds, Pandharpuri and Bhadawari clustered into separate groups while the other breeds seemed to cluster together. Admixture and F_{st} analysis revealed similar results where the Mediterranean breed remained distinct from other Indian breeds. However, the Mediterranean and Indian breeds shared some genetic structure. The reason for this may be the separation of the Mediterranean and Indian water buffalo from a common Asian river water buffalo (Kumar, et al., 2007). Sharing of genomic structure between the Indian breeds was also observed with a higher degree of genomic similarity between Banni and Murrah than other breeds suggesting that Banni and Murrah had a common ancestor. This chapter also provides information about various genomic regions where putative selection was observed, which brought about differences in various phenotypic traits among the water buffalo breeds. The selection signatures that were identified in this study reflected that water buffalo breeds have been subjected to natural selection pressure in response to adaptation to the local environment and artificial selection pressure for their milk/meat/draught related traits.

The results from this study are likely to be invaluable to inform future studies of how regulatory variants may confer tolerance to water buffalo pathogens as well as the impact of domestication on its genome. The atlas also contains various other organs and cells of the immune system such as the spleen, thymus, lymph nodes, peripheral blood mononuclear cells (PBMC), etc., in which some of the ASE detected in macrophages may be reproduced and in which ASE affecting cells of the acquired immune system (B cells, T cells) might be detected. The generation of Bone marrow derived macrophages (BMDM) from bone marrow is reproducible and can use frozen bone marrow. So, it would be practical to

generate similar BMDM +/- LPS data generated from larger group of genetically diverse water buffaloes to get a more diverse picture of the genetics of inter-individual variation in immune and inflammatory responses. This will lead to a better understanding of variation in disease susceptibility among individuals.

The variants/SNP markers discovered in this thesis and used in the analysis of genetic diversity amongst Indian and Mediterranean breeds is being used by colleagues in India to build a new SNP genotyping array since the currently available one does not capture the diversity present in the Indian water buffalo population (Iamartino, et al., 2017). Part of the project that generated genomic and expression data for this thesis involved sample collection in disease-endemic regions of India of DNA from large populations of water buffalo (>1,000) that were infected or uninfected with tuberculosis. This resource will facilitate genotyping of diseased and disease-resistant water buffalo breeds from India supporting breed improvement through selection.

The African buffalo (*Syncerus caffer*) is a wild buffalo and has not been domesticated due to its large size and unpredictable nature. It is considered to be a host to a variety of pathogens and infectious diseases such as anthrax, bovine tuberculosis, bovine brucellosis, Rift Valley fever, etc. and is considered a threat as it can transmit the diseases to cattle. The animal is asymptomatic for Foot and Mouth Disease and African trypanosomiasis (Michel and Bengis, 2012). Variation in the sequence of immune-related genes or in their regulation may lead to differences in its susceptibility to various diseases. Both African buffalo and Indian water buffalo share a similar disease burden, and hence, it would be interesting to compare if they share a similar kind of variation at the genetic level as well.

Imprinting is an important phenomenon required for the growth and development of fetus and placenta, peri- and postnatal physiology and also neurological development (Ivanova and Kelsey, 2011; Reik and Walter, 2001). Whereas all imprinted genes have evidence of monoallelic expression (MAE) based on parent of origin, not all MAE genes are necessarily imprinted. Only 3 MAE genes detected in macrophages matched with cattle imprinted genes in

Chapter 3 but this is mainly due to lack of expression. Most of the known imprinted genes identified in cattle (Chen, et al., 2016) are detectable in the water buffalo atlas in other tissues such as cerebellum, hippocampus, pituitary gland (hypophysis), adrenal gland and some in ovary-corpus luteum, kidney, testis and endometrium. There is clearly the potential for further analysis of MAE and imprinting using the atlas data.

Many production phenotypes in water buffalo, notably meat and dairy production and disease resistance are shared by other ruminant species. Chapter 4 identified sets of genes that showed evidence of selection in water buffalo. As genomic data accumulates for all these species, especially dairy cattle which are under very strong genetic selection, there will be many more opportunities for comparative analysis of signatures of selection in ruminants.

In chapter 3, the ASE analysis gave an insight into the presence of *cis*-regulatory variation in various genes that may lead to phenotypic differences amongst individuals. Difference in gene regulatory mechanism contributes to difference in local adaptation to different environments, immune response against diseases and difference in production related traits such as milk and meat quality (Cao, et al., 2019; Guillocheau, et al., 2019; Knight, 2004). Phenotypic difference amongst various populations is also driven by natural selection (Pritchard, et al., 2010). A beneficial mutation that allows a population to adapt to a particular environment locally will be under selection and its frequency may rise in that population compared to another set of population (for example, mutations in immune related genes and skin pigmentation genes). Furthermore, identification of genomic regions that have been under selection is important to understand differences within species or populations (Nielsen, et al., 2007). In case of domestic animals, some genomic regions may have got selected through the process of 'artificial selection'. In chapter 4, the selective sweep analysis identified genomic regions that may be under selection. It is probable that one of the loci regulating the expression of a gene connected to a phenotypic trait is under selection. Based on the results presented in Chapter 3, 3,278 genes showed significant ASE in at-least one of the 12 samples under study. 24 genes out of 179 candidate genes that were present in genomic

regions under selection explored in chapter 4, showed ASE. Only 16 genes out of these 24 had an official gene symbol - *ARHGAP18*, *ASCC3*, *CDK10*, *CXCL6*, *DBNDD1*, *EDRF1*, *EFCAB11*, *GFM2*, *HELB*, *MTHFD2L*, *NRDE2*, *POC1B*, *SPIRE2*, *TDP1*, *UROS* and *VPS35*. Out of these 16 genes, 4 could be connected to a trait they are responsible for. *CXCL8* is an immunity related gene, *EDRF1* is a dairy related gene and finally, *MTHFD2L* and *VPS35* are involved in growth and development. By combining ASE and selection signature analysis, we have narrowed down to 4 genes related to production and immune related traits. They are present in genomic regions under selection (natural or artificial) in one of the water buffalo breeds and show signature regulatory variation in the form of ASE. Further investigation is necessary in order to identify the genetic/molecular cause of such regulatory variation (for example, checking for SNPs in regulatory region of the genes) to explain phenotypic differences amongst water buffalo breeds at a population level. In the future, such information is vital in providing a basis for water buffalo breed improvement for traits related to immunity, milk and growth.

This thesis produced a list of genes that showed signs of regulatory variation in the form of ASE. In the future, validation of ASE in genes of interest is necessary. One of the ways to achieve this is through the method of 'pyrosequencing' (Ronaghi, et al., 1996). This method allows the measurement of difference in gene expression between two alleles in a heterozygous genotype. Specifically, it tries to detect the difference in the abundance of mRNA transcripts using a specific SNP or SNV to differentiate between the two alleles (Wittkopp, 2011). Pyrosequencing has been used to verify ASE in cattle (Guillocheau, et al., 2019), pigs (Stachowiak, et al., 2018), mouse (Wang, et al., 2011), humans (Yang, et al., 2015) and plants (Schaart, et al., 2005). Additionally, quantitative reverse transcriptase polymerase chain reaction (RT-PCR)-based assays can also be used to verify ASE (Singer-Sam and Gao, 2002). In-depth description of the methods and protocol is beyond the scope of this thesis.

An ASE analysis signposts genes whose expression is controlled by a *cis*-regulatory element. The SNV that is being used for determining ASE in a

particular gene may not be the variant that is regulating the expression of the gene. The SNV may be present either in the regulatory element of the gene (in the promoter or enhancer region), or it can itself be the regulatory variant. Hence, future work is necessary in order to find the underlying ASE causing regulatory variant and map it to the gene showing ASE. One way is to detect Linkage Disequilibrium or LD between a SNV present in a regulatory site and an informative SNV present in the transcribed region of a gene showing ASE (Lefebvre, et al., 2012). It is known that the variant present in the regulatory (noncoding) region of a gene showing ASE is probably the causal variant in majority of the cases, than a variant present in the coding region of the gene (Maurano, et al., 2012). The method is based on the principle that a *cis*-regulatory variant will be present very close to the gene transcript it is controlling and will also be tightly linked to the informative SNV present within the gene. This method has been recently used to decipher upstream SNVs in LD with SNVs within ASE genes in cattle (Guillocheau, et al., 2019). Specifically, a Pearson correlation score was calculated between expression levels of genes with ASE SNPs and the genotypes of SNVs present upstream of genes with ASE involving 19 individuals.

The selection analysis of Chapter 4 discovered loci that were under selection in water buffalo breeds. Many of these loci contain genes that have been previously observed to be under selection in other species. Few examples include two genes - *KITLG* and *ASIP* (agouti signalling protein), commonly linked to pigmentation/coat colour in other species.

Evidence for a putative selection was observed at the *ASIP* locus on chromosome 14. This gene has been linked to skin pigmentation in both humans (Liu, et al., 2015) and mice (Bultman, et al., 1994). Only two coding variants were detected in the *ASIP* gene, a synonymous variant at chr14:19947421 and a non-synonymous variant at chr14:19947429. The alternative allele frequency of the non-synonymous variant was substantially higher (62.5%) among the Murrah animals compared to other breeds (from 4.5% in Surti to 20.8% in Banni). This variant leads to an arginine to cysteine amino acid change in the C-terminal agouti domain of the protein that is linked

to melanocortin receptor binding activity in vitro (McNulty, et al., 2005). This domain contains ten cysteine residues that form a network of five disulphide bonds between them, shaping the active domain (Patel, et al., 2010). The creation of an extra cysteine residue within this region consequently has the potential to disrupt the looping of this domain and its active site (Kerns, et al., 2004).

KITLG gene variation, which controls melanocyte differentiation and migration, has also been associated with coat/skin colour in cattle (Seitz, et al., 1999) and other species (Picardo and Cardinali, 2011; Talenti, et al., 2017). Elevated XP-EHH scores were observed upstream of the gene in various comparisons of water buffalo breeds. Comparison between the Banni and Jaffrabadi breeds indicated elevated XP-EHH in a region restricted to immediately upstream of the *KITLG* transcription start site. A missense mutation in *KITLG* is associated with the roan phenotype in cattle (Seitz, et al., 1999). Only one non-synonymous variant was identified in the water buffalo *KITLG* gene i.e. chr4:101938991 and it was found in only two Jaffarabadi animals. Accordingly, the selective sweep at this gene in water buffalo may probably be associated with transcriptional regulation.

In overview, this thesis has provided a platform for future studies that will continue to address the current comparative lack of resources available for breed improvement in the water buffalo.

References

Aagaard-Tillery, K.M.*, et al.* (2008) Developmental origins of disease and determinants of chromatin structure: maternal diet modifies the primate fetal epigenome, *Journal of molecular endocrinology*, **41**, 91-102.

Abdelhaleem, M., Sun, T.H. and Ho, M. (2005) DHX32 expression suggests a role in lymphocyte differentiation, *Anticancer Res*, **25**, 2645-2648.

Abel, A.M.*, et al.* (2018) Natural Killer Cells: Development, Maturation, and Clinical Utilization, *Frontiers in immunology*, **9**.

Adams, L.G. and Templeton, J.W. (1998) Genetic resistance to bacterial diseases of animals, *Revue scientifique et technique (International Office of Epizootics)*, **17**, 200-219.

Ahn, B.*, et al.* (2019) Analysis of allele-specific expression using RNA-seq of the Korean native pig and Landrace reciprocal cross, *Asian-Australas J Anim Sci*, **0**, 0-0.

Akira, S., Uematsu, S. and Takeuchi, O. (2006) Pathogen Recognition and Innate Immunity, *Cell*, **124**, 783-801.

Albert, F.W. and Kruglyak, L. (2015) The role of regulatory variation in complex traits and disease, *Nature Reviews Genetics*, **16**, 197.

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

Alfano, F.*, et al.* (2014) Identification of single nucleotide polymorphisms in Toll-like receptor candidate genes associated with tuberculosis infection in water buffalo (Bubalus bubalis), *BMC Genetics*, **15**, 139.

Ameur, A.*, et al.* (2011) Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain, *Nature structural & molecular biology*, **18**, 1435-1440.

Anand, L. (2019) chromoMap: An R package for Interactive Visualization and Annotation of Chromosomes, *bioRxiv*, 605600.

Andrews, S. (2014) *FastQC A Quality Control tool for High Throughput Sequence Data*.

Anthony J.F. Griffiths, U.o.B.C.*, et al.* (2015) *Introduction to genetic analysis*. Eleventh edition. New York, NY : W.H. Freeman & Company, [2015] ©2015.

Arango Duque, G. and Descoteaux, A. (2014) Macrophage cytokines: involvement in immunity and infectious diseases, *Frontiers in immunology*, **5**, 491-491.

Arcanjo, A.C.*, et al.* (2014) Role of the host genetic variability in the influenza A virus susceptibility, *Acta biochimica Polonica*, **61**, 403-419.

Arora, R.*, et al.* (2004) Genetic diversity analysis of two buffalo populations of northern India using microsatellite markers, *Journal of Animal Breeding and Genetics*, **121**, 111-118.

Arya, S.B.*, et al.* (2018) ARL11 regulates lipopolysaccharide-stimulated macrophage activation by promoting mitogen-activated protein kinase (MAPK) signaling, *Journal of Biological Chemistry*, **293**, 9892-9909.

Awasthi, M.K.*, et al.* (2007) Is slow follicular growth the cause of silent estrus in water buffaloes?, *Animal reproduction science*, **99**, 258-268.

Ayuso, M.*, et al.* (2016) Developmental Stage, Muscle and Genetic Type Modify Muscle Transcriptome in Pigs: Effects on Gene Expression and Regulatory Factors Involved in Growth and Metabolism, *PLOS ONE*, **11**, e0167858.

B.P. Kushwaha, S.S., N. Das, S.B. Maity, K.K. Singh and J. Jayasankar (2013) Production and Reproductive Performance of Bhadawari Buffaloes in Uttar Pradesh, India, *Journal of Buffalo Science*, **2**, 72-77.

Baes, C.F.*, et al.* (2014) Evaluation of variant identification methods for whole genome sequencing data in dairy cattle, *BMC Genomics*, **15**, 948.

Bahbahani, H.*, et al.* (2018) Signatures of positive selection in African Butana and Kenana dairy zebu cattle, *PLOS ONE*, **13**, e0190446.

Bahn, J.H.*, et al.* (2012) Accurate identification of A-to-I RNA editing in human by transcriptome sequencing, *Genome research*, **22**, 142-150.

Bakhtiarizadeh, M.R., Salehi, A. and Rivera, R.M. (2018) Genome-wide identification and analysis of A-to-I RNA editing events in bovine by transcriptome sequencing, *PLOS ONE*, **13**, e0193316.

Balaton, B.P. and Brown, C.J. (2016) Escape Artists of the X Chromosome, *Trends Genet*, **32**, 348-359.

Banchereau, J. and Steinman, R.M. (1998) Dendritic cells and the control of immunity, *Nature*, **392**, 245-252.

Baran, Y.*, et al.* (2015) The landscape of genomic imprinting across diverse adult human tissues, *Genome research*, **25**, 927-936.

Barnett, D.W.*, et al.* (2011) BamTools: a C++ API and toolkit for analyzing and managing BAM files, *Bioinformatics*, **27**, 1691-1692.

Barreto Amaral Teixeira, D.*, et al.* (2017) Genomic analysis of stayability in Nellore cattle, *PLOS ONE*, **12**, e0179076.

Bartolomei, M.S. and Tilghman, S.M. (1997) Genomic imprinting in mammals, *Annual review of genetics*, **31**, 493-525.

Battle, A.*, et al.* (2014) Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals, *Genome research*, **24**, 14-24.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**, 289-300.

Berletch, J.B.*, et al.* (2015) Escape from X Inactivation Varies in Mouse Tissues, *PLOS Genetics*, **11**, e1005079.

Bertolini, F.*, et al.* (2018) Signatures of selection and environmental adaptation across the goat genome post-domestication, *Genetics Selection Evolution*, **50**, 57.

Bhaladhare, A.*, et al.* (2016) Single nucleotide polymorphisms in toll-like receptor genes and case-control association studies with bovine tuberculosis, *Vet World*, **9**, 458-464.

Bie, Q.*, et al.* (2017) IL-17B: A new area of study in the IL-17 family, *Molecular immunology*, **90**, 50-56.

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

Blott, S.*, et al.* (2003) Molecular dissection of a quantitative trait locus: a phenylalanine-to-tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition, *Genetics*, **163**, 253-266.

Bogdan, C. (2015) Nitric oxide synthase in innate and adaptive immunity: an update, *Trends in immunology*, **36**, 161-178.

Boitard, S.*, et al.* (2016) Uncovering Adaptation from Sequence Data: Lessons from Genome Resequencing of Four Cattle Breeds, *Genetics*, **203**, 433-450.

Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data, *Bioinformatics*, **30**, 2114-2120.

Borel, P., Preveraud, D. and Desmarchelier, C. (2013) Bioavailability of vitamin E in humans: an update, *Nutrition Reviews*, **71**, 319-331.

Borghese, A. (2005) *Buffalo Production and Research*. REU Technical Series 67. Food and Agriculture Organization of the United Nations, Rome.

Borghese, A. (2011) Situation and perspectives of buffalo in the world, Europe and Macedonia, *Macedonian Journal of Animal Science*, **1**, 281-296.

Borghese, A. (2013) Buffalo Livestock and Products in Europe, *Buffalo Bulletin*, **32**, 50-74.

Boro, P.*, et al.* (2018) *Milk composition and factors affecting it in dairy Buffaloes: A review*.

Borriello, G.*, et al.* (2006) Genetic resistance to Brucella abortus in the water buffalo (Bubalus bubalis), *Infection and immunity*, **74**, 2115-2120.

Brahma, B.*, et al.* (2015) Comparative genomic analysis of buffalo (Bubalus bubalis) NOD1 and NOD2 receptors and their functional role in in-vitro cellular immune response, *PloS one*, **10**, e0119178-e0119178.

Braz, C.U.*, et al.* (2019) Sliding window haplotype approaches overcome single SNP analysis limitations in identifying genes for meat tenderness in Nelore cattle, *BMC Genetics*, **20**, 8.

Brennicke, A., Marchfelder, A. and Binder, S. (1999) RNA editing, *FEMS microbiology reviews*, **23**, 297-316.

Brodin, P. and Davis, M.M. (2017) Human immune system variation, *Nature reviews. Immunology*, **17**, 21-29.

Brookes, A.J. (1999) The essence of SNPs, *Gene*, **234**, 177-186.

Browning, S.R. and Browning, B.L. (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering, *Am J Hum Genet*, **81**, 1084-1097.

Bryois, J.*, et al.* (2014) Cis and Trans Effects of Human Genomic Variants on Gene Expression, *PLOS Genetics*, **10**, e1004461.

Buckland, P.R.*, et al.* (2004) A high proportion of polymorphisms in the promoters of brain expressed genes influences transcriptional activity, *Biochimica et biophysica acta*, **1690**, 238-249.

Bultman, S.J.*, et al.* (1994) Molecular analysis of reverse mutations from nonagouti (a) to black-and-tan (a(t)) and white-bellied agouti (Aw) reveals alternative forms of agouti transcripts, *Genes & development*, **8**, 481-490.

Bumstead, N. and Barrow, P.A. (1988) Genetics of resistance to Salmonella typhimurium in newly hatched chicks, *British Poultry Science*, **29**, 521-529.

Callahan, H. (2002) Microevolution and Macroevolution: Introduction. In, *eLS*.

Cao, Y.*, et al.* (2019) Genetic Basis of Phenotypic Differences Between Chinese Yunling Black Goats and Nubian Goats Revealed by Allele-Specific Expression in Their F1 Hybrids, *Frontiers in Genetics*, **10**.

Carey, L.B. (2015) RNA polymerase errors cause splicing defects and can be regulated by differential expression of RNA polymerase subunits, *eLife*, **4**, e09945.

Carrara, G.*, et al.* (2017) Golgi anti-apoptotic protein: a tale of camels, calcium, channels and cancer, *Open biology*, **7**.

Carrel, L. and Willard, H.F. (2005) X-inactivation profile reveals extensive variability in X-linked gene expression in females, *Nature*, **434**, 400-404.

Castel, S.E.*, et al.* (2015) Tools and best practices for data processing in allelic expression analysis, *Genome biology*, **16**, 195.

Chadsuthi, S.*, et al.* (2017) Investigation on predominant Leptospira serovars and its distribution in humans and livestock in Thailand, 2010-2015, *PLOS Neglected Tropical Diseases*, **11**, e0005228.

Chainy, G.B.N., Paital, B. and Dandapat, J. (2016) An Overview of Seasonal Changes in Oxidative Stress and Antioxidant Defence Parameters in Some Invertebrate and Vertebrate Species, *Scientifica*, **2016**, 8.

Chamberlain, A.J.*, et al.* (2015) Extensive variation between tissues in allele specific expression in an outbred mammal, *BMC genomics*, **16**, 993-993.

Chen, H.*, et al.* (2013) A novel testis-enriched gene Spata33 is expressed during spermatogenesis, *PLoS One*, **8**, e67882.

Chen, J.*, et al.* (2016) A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals, *Nature communications*, **7**, 11101.

Chen, Z.*, et al.* (2016) Global assessment of imprinted gene expression in the bovine conceptus by next generation sequencing, *Epigenetics*, **11**, 501-516.

Chess, A. (2016) Monoallelic Gene Expression in Mammals, *Annual review of genetics*, **50**, 317-327.

Cheung, V.G.*, et al.* (2003) Natural variation in human gene expression assessed in lymphoblastoid cells, *Nature genetics*, **33**, 422-425.

Cingolani, P.*, et al.* (2012) Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift, *Front Genet*, **3**, 35.

Cingolani, P.*, et al.* (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3, *Fly*, **6**, 80-92.

Clark, E.L.*, et al.* (2017) A high resolution atlas of gene expression in the domestic sheep (Ovis aries), *PLOS Genetics*, **13**, e1006997.

Clop, A.*, et al.* (2016) Identification of genetic variation in the swine toll-like receptors and development of a porcine TLR genotyping array, *Genet Sel Evol*, **48**, 28-28.

Cockett, N.E. and Kole, C. (2008) *Genome Mapping and Genomics in Domestic Animals*. Springer Berlin Heidelberg.

Cockrill, W.R. (1974) The working buffalo. In Cockrill, W.R. (ed), *The husbandry and health of the domestic buffalo*. Food and Agriculture Organization of the United Nations, Rome, pp. 313-328.

Cockrill, W.R. (1977) *The Water Buffalo*. Food and Agriculture Organization of the United Nations, Rome.

Cockrill, W.R. (1981) The water buffalo: a review, *The British veterinary journal*, **137**, 8-16.

Coles, C.A.*, et al.* (2014) A disintegrin and metalloprotease-12 is type I myofiber specific in Bos taurus and Bos indicus cattle, *Journal of animal science*, **92**, 1473-1483.

Colli, L.*, et al.* (2018) New Insights on Water Buffalo Genomic Diversity and Post-Domestication Migration Routes From Medium Density SNP Chip Data, *Frontiers in Genetics*, **9**.

Conway, J.R., Lex, A. and Gehlenborg, N. (2017) UpSetR: an R package for the visualization of intersecting sets and their properties, *Bioinformatics*, **33**, 2938-2940.

Cooper, D.N. and Gerber-Huber, S. (1985) DNA methylation and CpG suppression, *Cell Differentiation*, **17**, 199-205.

Costa-Silva, J., Domingues, D. and Lopes, F.M. (2017) RNA-Seq differential expression analysis: An extended review and a software tool, *PLOS ONE*, **12**, e0190152.

Coulondre, C.*, et al.* (1978) Molecular basis of base substitution hotspots in Escherichia coli, *Nature*, **274**, 775.

Cruvinel Wde, M.*, et al.* (2010) Immune system - part I. Fundamentals of innate immunity with emphasis on molecular and cellular mechanisms of inflammatory response, *Revista brasileira de reumatologia*, **50**, 434-461.

Dabas, P.*, et al.* (2010) *Relationship of age and body weight with scrotal circumference in Murrah buffalo bulls/males*.

Danecek, P.*, et al.* (2011) The variant call format and VCFtools, *Bioinformatics*, **27**, 2156-2158.

Das AK, S.D., Kumar N (2008) Buffalo genetic resources in India and their conservation, *Buffalo Bulletin*, **27**, 265-268.

Das, P.J., Chowdhary, B.P. and Raudsepp, T. (2009) Characterization of the bovine pseudoautosomal region and comparison with sheep, goat, and other mammalian pseudoautosomal regions, *Cytogenetic and genome research*, **126**, 139-147.

Das, S.K., Upadhyay, R.C. and Madan, M.L. (1999) Heat stress in Murrah buffalo calves, *Livestock Production Science*, **61**, 71-78.

Dastgheib, S.A.*, et al.* (2016) A Candidate Gene Association Study of Bone Mineral Density in an Iranian Population, *Frontiers in endocrinology*, **7**.

de Lisle, G.W., Mackintosh, C.G. and Bengis, R.G. (2001) Mycobacterium bovis in free-living and captive wildlife, including farmed deer, *Revue scientifique et technique (International Office of Epizootics)*, **20**, 86-111.

de Simoni Gouveia, J.J.*, et al.* (2017) Genome-wide search for signatures of selection in three major Brazilian locally adapted sheep breeds, *Livestock Science*, **197**, 36-45.

Deb, G.K.*, et al.* (2016) Safe and Sustainable Traditional Production: The Water Buffalo in Asia, *Frontiers in Environmental Science*, **4**.

Deelen, P.*, et al.* (2015) Calling genotypes from public RNA-sequencing data enables identification of genetic variants that affect gene-expression levels, *Genome medicine*, **7**, 30-30.

DeGiorgio, M., Lohmueller, K.E. and Nielsen, R. (2014) A Model-Based Approach for Identifying Signatures of Ancient Balancing Selection in Genetic Data, *PLOS Genetics*, **10**, e1004561.

Deng, Q.*, et al.* (2014) Single-Cell RNA-Seq Reveals Dynamic, Random Monoallelic Gene Expression in Mammalian Cells, *Science (New York, N.Y.)*, **343**, 193.

DePristo, M.A.*, et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data, *Nature genetics*, **43**, 491-498.

DerSimonian, R. and Laird, N. (1986) Meta-analysis in clinical trials, *Controlled clinical trials*, **7**, 177-188.

Desio, G.*, et al.* (2013) Estimated Prevalence of Johne's Disease in Herds of Water Buffaloes (Bubalus Bubalis) in the Province of Caserta, *Italian Journal of Animal Science*, **12**, e8.

Dhakal, I.P. (2006) Normal somatic cell count and subclinical mastitis in Murrah buffaloes, *Journal of veterinary medicine. B, Infectious diseases and veterinary public health*, **53**, 81.

Dhanasekaran, S.*, et al.* (2014) Toll-Like Receptor Responses to Peste des petits ruminants Virus in Goats and Water Buffalo, *PLOS ONE*, **9**, e111609.

Dhanda, O.P. (2004) Developments in water buffalo in Asia and Oceania. *Proceedings of the 7th World Buffalo Congress*. Manila, Philippines, pp. 17-28.

Dhillod, S.*, et al.* (2017) Study of the dairy characters of lactating Murrah buffaloes on the basis of body parts measurements, *Vet World*, **10**, 17-21.

Diacovich, L. and Gorvel, J.P. (2010) Bacterial manipulation of innate immunity to promote infection, *Nature reviews. Microbiology*, **8**, 117-128.

Duan, J.*, et al.* (2018) Dosage Compensation of the X Chromosomes in Bovine Germline, Early Embryos, and Somatic Tissues, *Genome biology and evolution*, **11**, 242-252.

Duchemin, S.I.*, et al.* (2016) Identification of QTL on Chromosome 18 Associated with Non-Coagulating Milk in Swedish Red Cows, *Frontiers in Genetics*, **7**.

Duehlmeier, R.*, et al.* (2007) Distribution patterns of the glucose transporters GLUT4 and GLUT1 in skeletal muscles of rats (Rattus norvegicus), pigs (Sus scrofa), cows (Bos taurus), adult goats, goat kids (Capra hircus), and camels (Camelus dromedarius), *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology*, **146**, 274-282.

Duitama, J., Srivastava, P.K. and Măndoiu, I.I. (2012) Towards accurate detection and genotyping of expressed variants from whole transcriptome sequencing data, *BMC genomics*, **13 Suppl 2**, S6-S6.

Eckersley-Maslin, M.A. and Spector, D.L. (2014) Random monoallelic expression: regulating gene expression one allele at a time, *Trends Genet*, **30**, 237-244.

Edsgärd, D.*, et al.* (2016) GeneiASE: Detection of condition-dependent and static allele-specific expression from RNA-seq data without haplotype information, *Scientific Reports*, **6**, 21134.

Eisenberg, E.*, et al.* (2005) Identification of RNA editing sites in the SNP database, *Nucleic Acids Res*, **33**, 4612-4617.

Ellegren, H. (2008) Comparative genomics and the study of evolution by natural selection, *Molecular ecology*, **17**, 4586-4596.

Engström, P.G., *et al.* (2013) Systematic evaluation of spliced alignment programs for RNA-seq data, *Nature methods*, **10**, 1185.

Evron, T., *et al.* (2011) Growth Arrest Specific 8 (Gas8) and G protein-coupled receptor kinase 2 (GRK2) cooperate in the control of Smoothened signaling, *The Journal of biological chemistry*, **286**, 27676-27686.

Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities, *Genome research*, **8**, 186-194.

Ewing, B., *et al.* (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment, *Genome research*, **8**, 175-185.

Fairfax, B.P., *et al.* (2014) Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression, *Science (New York, N.Y.)*, **343**, 1246949.

Fariello, M.-I., *et al.* (2014) Selection Signatures in Worldwide Sheep Populations, *PLOS ONE*, **9**, e103813.

Fasano, M., *et al.* (2005) The extraordinary ligand binding properties of human serum albumin, *IUBMB life*, **57**, 787-796.

Fernando, S.L. and Britton, W.J. (2006) Genetic susceptibility to mycobacterial disease in humans, *Immunology and cell biology*, **84**, 125-137.

Feschotte, C. (2008) Transposable elements and the evolution of regulatory networks, *Nature reviews. Genetics*, **9**, 397-405.

Feuk, L., Carson, A.R. and Scherer, S.W. (2006) Structural variation in the human genome, *Nature Reviews Genetics*, **7**, 85-97.

Flori, L., *et al.* (2009) The genome response to artificial selection: a case study in dairy cattle, *PloS one*, **4**, e6595-e6595.

Fogg, D.K., *et al.* (2006) A clonogenic bone marrow progenitor specific for macrophages and dendritic cells, *Science (New York, N.Y.)*, **311**, 83-87.

Fontanillas, P., *et al.* (2010) Key considerations for measuring allelic expression on a genomic scale using high-throughput sequencing, *Molecular ecology*, **19 Suppl 1**, 212-227.

Fortes, M.R.S., *et al.* (2013) Genome-wide association study for inhibin, luteinizing hormone, insulin-like growth factor 1, testicular size and semen traits in bovine species, *Andrology*, **1**, 644-650.

Frankish, A., *et al.* (2017) Ensembl 2018, *Nucleic Acids Res*, **46**, D754-D761.

Gao, X., *et al.* (2015) Interleukin-10 promoter gene polymorphisms and susceptibility to tuberculosis: a meta-analysis, *PLoS One*, **10**, e0127496.

Gao, Y.*, et al.* (2001) Myosin light chain kinase as a multifunctional regulatory protein of smooth muscle contraction, *IUBMB life*, **51**, 337-344.

Garieri, M.*, et al.* (2018) Extensive cellular heterogeneity of X inactivation revealed by single-cell allele-specific expression in human fibroblasts, *Proceedings of the National Academy of Sciences*, **115**, 13015.

Gehin, M.*, et al.* (2002) The function of TIF2/GRIP1 in mouse reproduction is distinct from those of SRC-1 and p/CIP, *Molecular and cellular biology*, **22**, 5923-5937.

Ghasemzadeh, M.*, et al.* (2013) The CXCR1/2 ligand NAP-2 promotes directed intravascular leukocyte migration through platelet thrombi, *Blood*, **121**, 4555-4566.

Gibbs, R.A.*, et al.* (2003) The International HapMap Project, *Nature*, **426**, 789-796.

Gillespie, J.H. (1998) *Population Genetics: A Concise Guide*. Johns Hopkins University Press.

Gimelbrant, A.*, et al.* (2007) Widespread monoallelic expression on human autosomes, *Science (New York, N.Y.)*, **318**, 1136-1140.

Gonzalez-Hormazabal, P.*, et al.* (2018) IL-8-251T>A (rs4073) Polymorphism Is Associated with Prognosis in Gastric Cancer Patients, *Anticancer Research*, **38**, 5703-5708.

Gonzalez, O.J.G. (2011) Buffalo bulls for meat production: feeding and meat quality. *Faculty of veterinary medicine*. University of Naples Federico II, Naples, Italy.

Gordon, S. and Taylor, P.R. (2005) Monocyte and macrophage heterogeneity, *Nature reviews. Immunology*, **5**, 953-964.

Graham, A.*, et al.* (1997) Agouti protein inhibits the production of eumelanin and phaeomelanin in the presence and absence of alpha-melanocyte stimulating hormone, *Pigment cell research*, **10**, 298-303.

Grange, J.M. (2001) Mycobacterium bovis infection in human beings, *Tuberculosis (Edinburgh, Scotland)*, **81**, 71-77.

Gu, Z., Eils, R. and Schlesner, M. (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data, *Bioinformatics*, **32**, 2847-2849.

Guillocheau, G.M.*, et al.* (2019) Survey of allele specific expression in bovine muscle, *Scientific Reports*, **9**, 4297.

Guler, G.D.*, et al.* (2012) Human DNA helicase B (HDHB) binds to replication protein A and facilitates cellular recovery from replication stress, *The Journal of biological chemistry*, **287**, 6469-6481.

Guo, C.*, et al.* (2017) Transversions have larger regulatory effects than transitions, *BMC Genomics*, **18**, 394.

Guo, H., Callaway, J.B. and Ting, J.P.Y. (2015) Inflammasomes: mechanism of action, role in disease, and therapeutics, *Nature Medicine*, **21**, 677.

Hamilton, C.A.*, et al.* (2017) Frequency and phenotype of natural killer cells and natural killer cell subsets in bovine lymphoid compartments and blood, *Immunology*, **151**, 89-97.

Harada, A.*, et al.* (1994) Essential involvement of interleukin-8 (IL-8) in acute inflammation, *Journal of leukocyte biology*, **56**, 559-564.

Harisah, M.*, et al.* (1989) Identification of crossbred buffalo genotypes and their chromosome segregation patterns, *Genome*, **32**, 999-1002.

Harismendy, O.*, et al.* (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies, *Genome biology*, **10**, R32-R32.

Hartl, D.L. and Clark, A.G. (1997) *Principles of Population Genetics*. Sinauer Associates.

Harvey, C.T.*, et al.* (2014) QuASAR: quantitative allele-specific analysis of reads, *Bioinformatics*, **31**, 1235-1242.

Hay, E.H. and Roberts, A. (2018) Genome-wide association study for carcass traits in a composite beef cattle breed, *Livestock Science*, **213**, 35-43.

He, J.*, et al.* (2017) Association study of MCP-1 promoter polymorphisms with the susceptibility and progression of sepsis, *PloS one*, **12**, e0176781-e0176781.

Hendry, A.P. and Kinnison, M.T. (2001) An introduction to microevolution: rate, pattern, process, *Genetica*, **112-113**, 1-8.

Hens, J.R. and Wysolmerski, J.J. (2005) Key stages of mammary gland development: Molecular mechanisms involved in the formation of the embryonic mammary gland, *Breast Cancer Research*, **7**, 220.

Hester, J.M. and Diaz, F. (2018) Growing Oocytes Need Zinc: Zinc Deficiency in the Preantral Ovarian Follicle, *The FASEB Journal*, **32**, 882.881-882.881.

Hodges, E.*, et al.* (2009) High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing, *Genome research*, **19**, 1593-1605.

Hoffpauir, R. (1982) The Water Buffalo: India's Other Bovine, *Anthropos*, **77**, 215-238.

Horne, D.J.*, et al.* (2015) Innate Immunity Candidate Gene Association Study and Susceptibility to Latent Tuberculosis Infection. In, *A96. MOLECULAR MYCOBACTERIOLOGY: HOST AND BACILLUS*. American Thoracic Society, pp. A2179-A2179.

Horns, F. and Hood, M.E. (2012) The evolution of disease resistance and tolerance in spatially structured populations, *Ecology and evolution*, **2**, 1705-1711.

Horowitz, E.*, et al.* (2005) Patterns of expression of sperm flagellar genes: early expression of genes encoding axonemal proteins during the spermatogenic cycle and shared features of promoters of genes encoding central apparatus proteins, *Molecular human reproduction*, **11**, 307-317.

Huang da, W., Sherman, B.T. and Lempicki, R.A. (2009a) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists, *Nucleic Acids Res*, **37**, 1-13.

Huang da, W., Sherman, B.T. and Lempicki, R.A. (2009b) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, *Nature protocols*, **4**, 44-57.

Huang, J.*, et al.* (2005) SIKE is an IKK epsilon/TBK1-associated suppressor of TLR3- and virus-triggered IRF-3 activation pathways, *The EMBO journal*, **24**, 4018-4028.

Huang, S.*, et al.* (2014) A novel multi-alignment pipeline for high-throughput sequencing data, *Database : the journal of biological databases and curation*, **2014**, bau057.

Hume, D.A. (2008) Differentiation and heterogeneity in the mononuclear phagocyte system, *Mucosal immunology*, **1**, 432-441.

Hume, D.A. (2008) Macrophages as APC and the dendritic cell myth, *Journal of immunology (Baltimore, Md. : 1950)*, **181**, 5829-5835.

Hume, D.A. (2015) The Many Alternative Faces of Macrophage Activation, *Frontiers in immunology*, **6**, 370.

Hume, D.A., Irvine, K.M. and Pridans, C. (2019) The Mononuclear Phagocyte System: The Relationship between Monocytes and Macrophages, *Trends in immunology*, **40**, 98-112.

Hume, D.A. and MacDonald, K.P. (2012) Therapeutic applications of macrophage colony-stimulating factor-1 (CSF-1) and antagonists of CSF-1 receptor (CSF-1R) signaling, *Blood*, **119**, 1810-1820.

Iamartino, D., Nicolazzi, E.L. and Van Tassell, C.P. (2017) Design and validation of a 90K SNP genotyping assay for the water buffalo (Bubalus bubalis), **12**, e0185220.

Iida, N.*, et al.* (1996) Leukocyte-derived growth factor links the PDGF and CXC chemokine families of peptides, *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, **10**, 1336-1345.

Infascelli, F.*, et al.* (2003) *Nutritional characteristics of buffalo meat: Cholesterol content and fatty acid composition*.

Ivanova, E. and Kelsey, G. (2011) Imprinted genes and hypothalamic function, *Journal of molecular endocrinology*, **47**, R67-74.

Iwasa, H., Shimizu, T. and Hata, Y. (2018) RASSF6. In Choi, S. (ed), *Encyclopedia of Signaling Molecules*. Springer International Publishing, Cham, pp. 4524-4528.

Jain, N.*, et al.* (2011) *Cytokines expression profile of Brucella abortus infected Indian water buffaloes*.

Janeway, C.A., Jr. and Medzhitov, R. (2002) Innate immune recognition, *Annual review of immunology*, **20**, 197-216.

Janmeda, M.*, et al.* (2017) *Variation in Test Day Milk Yield and Composition at Day 15 and 60 Postpartum in Surti and Jafarabadi Buffaloes*.

Jann, O.C.*, et al.* (2009) Comparative genomics of Toll-like receptor signalling in five species, *BMC Genomics*, **10**, 216.

Jimenez-Dalmaroni, M.J., Gerswhin, M.E. and Adamopoulos, I.E. (2016) The critical role of toll-like receptors--From microbial recognition to autoimmunity: A comprehensive review, *Autoimmunity reviews*, **15**, 1-8.

Johnson, P.*, et al.* (2018) Hyaluronan and Its Interactions With Immune Cells in the Healthy and Inflamed Lung, *Frontiers in immunology*, **9**.

Jorgensen, T.J.*, et al.* (2009) Hypothesis-Driven Candidate Gene Association Studies: Practical Design and Analytical Considerations, *American Journal of Epidemiology*, **170**, 986-993.

Kahle, D., Wickham, H. and Jackson, S. (2019) ggmap: Spatial Visualization with ggplot2.

Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes, *Nucleic Acids Res*, **28**, 27-30.

Kang, E.Y.*, et al.* (2016) Discovering Single Nucleotide Polymorphisms Regulating Human Gene Expression Using Allele Specific Expression from RNA-seq Data, *Genetics*, **204**, 1057-1064.

Kant, N.*, et al.* (2018) A study to identify the practices of the buffalo keepers which inadvertently lead to the spread of brucellosis in Delhi, *BMC Vet Res*, **14**, 329-329.

Karki, R.*, et al.* (2015) Defining "mutation" and "polymorphism" in the era of personal genomics, *BMC Med Genomics*, **8**, 37-37.

Kerns, J.A.*, et al.* (2004) Characterization of the dog Agouti gene and a nonagoutimutation in German Shepherd Dogs, *Mammalian Genome*, **15**, 798-808.

Kim, D., Langmead, B. and Salzberg, S.L. (2015) HISAT: a fast spliced aligner with low memory requirements, *Nature methods*, **12**, 357-360.

Kim, J.*, et al.* (2017) The genome landscape of indigenous African cattle, *Genome biology*, **18**, 34.

Kimbrell, D.A. and Beutler, B. (2001) The evolution and genetics of innate immunity, *Nature reviews. Genetics*, **2**, 256-267.

Kindt, T.J.*, et al.* (2007) *Kuby immunology*. W.H. Freeman, New York.

Knight, J.C. (2004) Allele-specific gene expression uncovered, *Trends in Genetics*, **20**, 113-116.

Knight, J.C. (2005) Regulatory polymorphisms underlying complex disease traits, *J Mol Med (Berl)*, **83**, 97-109.

Kobayashi, K.*, et al.* (2002) IRAK-M is a negative regulator of Toll-like receptor signaling, *Cell*, **110**, 191-202.

Koets, A.*, et al.* (2010) Susceptibility to paratuberculosis infection in cattle is associated with single nucleotide polymorphisms in Toll-like receptor 2 which modulate immune responses against Mycobacterium avium subspecies paratuberculosis, *Preventive Veterinary Medicine*, **93**, 305-315.

Krueger, F. and Andrews, S.R. (2016) SNPsplit: Allele-specific splitting of alignments between genomes with known SNP genotypes, *F1000Research*, **5**, 1479-1479.

Krupa, A.*, et al.* (2015) Binding of CXCL8/IL-8 to Mycobacterium tuberculosis Modulates the Innate Immune Response, *Mediators of Inflammation*, **2015**, 11.

Kuipers, A.*, et al.* (2009) Association of a high mobility group gene (HMGA2) variant with bone mineral density, *Bone*, **45**, 295-300.

Kumar, H., Kawai, T. and Akira, S. (2011) Pathogen recognition by the innate immune system, *International reviews of immunology*, **30**, 16-34.

Kumar, S.*, et al.* (2006) Genetic variation and relationships among eight Indian riverine buffalo breeds, *Molecular ecology*, **15**, 593-600.

Kumar, S.*, et al.* (2007) Phylogeography and domestication of Indian river buffalo, *BMC Evolutionary Biology*, **7**, 186.

Kumar Yadav, A. (2017) *Characteristic features of registered Indigenous Buffalo Breeds of India: A Review*.

Kushwaha, B.P.*, et al.* (2007) *Status of Bhadawari breed of buffalo in its breeding tract and its conservation*.

Kwon, J.M. and Goate, A.M. (2000) The candidate gene approach, *Alcohol research & health : the journal of the National Institute on Alcohol Abuse and Alcoholism*, **24**, 164-168.

LaFramboise, T. (2009) Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances, *Nucleic Acids Res*, **37**, 4181-4193.

Lappalainen, T.*, et al.* (2013) Transcriptome and genome sequencing uncovers functional variation in humans, *Nature*, **501**, 506-511.

Larson, Nicholas B.*, et al.* (2015) Comprehensively Evaluating cis-Regulatory Variation in the Human Prostate Transcriptome by Using Gene-Level Allele-Specific Expression, *The American Journal of Human Genetics*, **96**, 869-882.

Latchman, D.S. (1993) Transcription factors: an overview, *Int J Exp Pathol*, **74**, 417-422.

Lau, C.H.*, et al.* (1998) Genetic diversity of Asian water buffalo (Bubalus bubalis): mitochondrial DNA D-loop and cytochrome b sequence variation, *Animal Genetics*, **29**, 253-264.

Lee, S.M.*, et al.* (2012) Cloning and Molecular Characterization of Porcine β-casein Gene (CNS2), *Asian-Australasian journal of animal sciences*, **25**, 421-427.

Lee, W.Y.*, et al.* (2017) Identification of lactoferrin and glutamate receptor-interacting protein 1 in bovine cervical mucus: A putative marker for oestrous detection, *Reproduction in domestic animals = Zuchthygiene*, **52**, 16-23.

Lefebvre, J.F.*, et al.* (2012) Genotype-Based Test in Mapping Cis-Regulatory Variants from Allele-Specific Expression Data, *PLOS ONE*, **7**, e38667.

Lewis, M.T.*, et al.* (1999) Regulated expression patterns of IRX-2, an Iroquois-class homeobox gene, in the human breast, *Cell and tissue research*, **296**, 549-554.

Lex, A., *et al.* (2014) UpSet: Visualization of Intersecting Sets, *IEEE Trans Vis Comput Graph*, **20**, 1983-1992.

Li, C., *et al.* (2014) Genome Wide Association Study Identifies 20 Novel Promising Genes Associated with Milk Fatty Acid Traits in Chinese Holstein, *PLOS ONE*, **9**, e96186.

Li, H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data, *Bioinformatics (Oxford, England)*, **27**, 2987-2993.

Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2 [q-bio.GN].

Li, H. (2016) Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences, *Bioinformatics*, **32**, 2103-2110.

Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences, *Bioinformatics*, bty191-bty191.

Li, H., *et al.* (2009) The Sequence Alignment/Map format and SAMtools, *Bioinformatics*, **25**, 2078-2079.

Li, H., Ruan, J. and Durbin, R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores, *Genome research*, **18**, 1851-1858.

Li, W.T., *et al.* (2017) Whole-genome resequencing reveals candidate mutations for pig prolificacy, *Proceedings. Biological sciences*, **284**.

Li, Y., *et al.* (2014) Domestication of the dog from the wolf was promoted by enhanced excitatory synaptic plasticity: a hypothesis, *Genome biology and evolution*, **6**, 3115-3121.

Linge, H.M., *et al.* (2008) The human CXC chemokine granulocyte chemotactic protein 2 (GCP-2)/CXCL6 possesses membrane-disrupting properties and is antibacterial, *Antimicrobial agents and chemotherapy*, **52**, 2599-2607.

Linnaeus, C. (1758) Systema Naturae, edition X, vol. 1 (Systema naturae per regna tria naturae, secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis. Tomus I. Editio decima, reformata), *Holmiae Salvii*, **824**.

Lisle, R.S., *et al.* (2013) Oocyte-cumulus cell interactions regulate free intracellular zinc in mouse oocytes, *Reproduction (Cambridge, England)*, **145**, 381-390.

Liu, F., *et al.* (2015) Genetics of skin color variation in Europeans: genome-wide association studies with functional follow-up, *Human genetics*, **134**, 823-835.

Loewe, L. and Hill, W.G. (2010) The population genetics of mutations: good, bad and indifferent, *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **365**, 1153-1167.

Low, W.Y.*, et al.* (2019) Chromosome-level assembly of the water buffalo genome surpasses human and goat genomes in sequence contiguity, *Nature communications*, **10**, 260.

Lozano-Urena, A.*, et al.* (2017) Genomic Imprinting and the Regulation of Postnatal Neurogenesis, *Brain plasticity (Amsterdam, Netherlands)*, **3**, 89-98.

Lü, M.-D.*, et al.* (2016) Genetic variations associated with six-white-point coat pigmentation in Diannan small-ear pigs, *Scientific reports*, **6**, 27534-27534.

Ma, X.*, et al.* (2015) Dietary L-Arginine Supplementation Affects the Skeletal Longissimus Muscle Proteome in Finishing Pigs, *PLOS ONE*, **10**, e0117294.

Maceachern, S.*, et al.* (2011) Genome-wide identification of allele-specific expression (ASE) in response to Marek's disease virus infection using next generation sequencing, *BMC proceedings*, **5 Suppl 4**, S14.

Macgregor, R. (1939) The domestic buffalo. Thesis presented to the Royal College of Veterinary Surgeons (quoted by I.L. Mason in Cockrill (1974),1-47).

Maclean, C.A., Chue Hong, N.P. and Prendergast, J.G.D. (2015) hapbin: An Efficient Program for Performing Haplotype-Based Scans for Positive Selection in Large Genomic Datasets, *Molecular biology and evolution*, **32**, 3027-3029.

Maddur, M.S.*, et al.* (2009) Immune response and viral persistence in Indian buffaloes (Bubalus bubalis) infected with foot-and-mouth disease virus serotype Asia 1, *Clinical and vaccine immunology : CVI*, **16**, 1832-1836.

Manichaikul, A.*, et al.* (2010) Robust relationship inference in genome-wide association studies, *Bioinformatics*, **26**, 2867-2873.

Mapleson, D.*, et al.* (2018) Efficient and accurate detection of splice junctions from RNA-seq with Portcullis, *Gigascience*, **7**.

Marín, J.C.*, et al.* (2018) Genetic Variation in Coat Colour Genes MC1R and ASIP Provides Insights Into Domestication and Management of South American Camelids, *Frontiers in Genetics*, **9**.

Marion, V.*, et al.* (2013) Hepatic adaptation compensates inactivation of intestinal arginine biosynthesis in suckling mice, *PLoS One*, **8**, e67021.

Mason, I.L. (1974) Species, types and breeds. In Cockrill, W.R. (ed), *The husbandry and health of the domestic buffalo*. Food and Agriculture Organization of the United Nations, Rome, pp. 1–47.

Mathur, R.*, et al.* (2012) A mouse model of Salmonella typhi infection, *Cell*, **151**, 590-602.

Mathur, T.N. (1964) Brucella strains isolated from cows, buffaloes, goats, sheep and human beings at karnal: their significance with regard to the epidemiology of brucellosis, *The Indian journal of medical research*, **52**, 1231-1240.

Maurano, M.T.*, et al.* (2012) Systematic localization of common disease-associated variation in regulatory DNA, *Science (New York, N.Y.)*, **337**, 1190-1195.

Mayba, O.*, et al.* (2014) MBASED: allele-specific expression detection in cancer tissues and cell lines, *Genome biology*, **15**, 405.

McGowan, P.J.K.*, et al.* (2019) Tracking trends in the extinction risk of wild relatives of domesticated species to assess progress against global biodiversity targets, *Conservation Letters*, **12**, e12588.

McGuire, W.*, et al.* (1994) Variation in the TNF-alpha promoter region associated with susceptibility to cerebral malaria, *Nature*, **371**, 508-510.

McKenna, A.*, et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, *Genome research*, **20**, 1297-1303.

McLaren, P.J. and Carrington, M. (2015) The impact of host genetic variation on infection with HIV-1, *Nature immunology*, **16**, 577-583.

McNulty, J.C.*, et al.* (2005) Structures of the agouti signaling protein, *Journal of molecular biology*, **346**, 1059-1070.

McTaggart, S.J.*, et al.* (2012) Immune genes undergo more adaptive evolution than non-immune system genes in Daphnia pulex, *BMC evolutionary biology*, **12**, 63-63.

Medvedev, A.E. (2013) Toll-like receptor polymorphisms, inflammatory and infectious diseases, allergies, and cancer, *J Interferon Cytokine Res*, **33**, 467-484.

Medzhitov, R. (2001) Toll-like receptors and innate immunity, *Nature reviews. Immunology*, **1**, 135-145.

Medzhitov, R. and Janeway, C. (2000) Innate Immunity, *New England Journal of Medicine*, **343**, 338-344.

Menning, M. and Kufer, T.A. (2013) A role for the Ankyrin repeat containing protein Ankrd17 in Nod1- and Nod2-mediated inflammatory responses, *FEBS letters*, **587**, 2137-2142.

Metsalu, T.*, et al.* (2014) Using RNA sequencing for identifying gene imprinting and random monoallelic expression in human placenta, *Epigenetics*, **9**, 1397-1409.

Metzker, M.L. (2009) Sequencing technologies — the next generation, *Nature Reviews Genetics*, **11**, 31.

Miao, Y.*, et al.* (2010) The role of MC1R gene in buffalo coat color, *Science China. Life sciences*, **53**, 267-272.

Michaelson, J.J., Loguercio, S. and Beyer, A. (2009) Detection and interpretation of expression quantitative trait loci (eQTL), *Methods*, **48**, 265-276.

Michaut, M.*, et al.* (2001) Synaptotagmin VI Participates in the Acrosome Reaction of Human Spermatozoa, *Developmental Biology*, **235**, 521-529.

Michel, A.L. and Bengis, R.G. (2012) The African buffalo: a villain for inter-species spread of infectious diseases in southern Africa, *The Onderstepoort journal of veterinary research*, **79**, 453.

Min, B.*, et al.* (2017) Characterization of X-Chromosome Gene Expression in Bovine Blastocysts Derived by In vitro Fertilization and Somatic Cell Nuclear Transfer, *Frontiers in Genetics*, **8**.

Miranpuri, G.S. (1988) Ticks parasitising the Indian buffalo (Bubalus bubalis) and their possible role in disease transmission, *Veterinary Parasitology*, **27**, 357-362.

Mishra, B.P.*, et al.* (2015) Genetic analysis of river, swamp and hybrid buffaloes of north-east India throw new light on phylogeography of water buffalo (Bubalus bubalis), *Journal of animal breeding and genetics = Zeitschrift fur Tierzuchtung und Zuchtungsbiologie*, **132**, 454-466.

Mishra, B.P.*, et al.* (2009) Evaluation of genetic variability and mutation drift equilibrium of Banni buffalo using multi locus microsatellite markers, *Tropical animal health and production*, **41**, 1203-1211.

Mishra, B.P.*, et al.* (2011) Characterization of Banni buffalo of Western India, *Animal Genetic Resources Information*, **44**, 77-86.

Misra, S.S.*, et al.* (2008) Association of breed and polymorphism of α-and α-casein genes with milk quality and daily milk and constituent yield traits of buffaloes (bubalus bubalis), *Buffalo Bulletin*, **27**, 294-301.

Mokhber, M.*, et al.* (2018) A genome-wide scan for signatures of selection in Azeri and Khuzestani buffalo breeds, *BMC Genomics*, **19**, 449.

Monk, D.*, et al.* (2006) Limited evolutionary conservation of imprinting in the human placenta, *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 6623-6628.

Moraes, M.O.*, et al.* (2004) Interleukin-10 promoter single-nucleotide polymorphisms as markers for disease susceptibility and disease severity in leprosy, *Genes and immunity*, **5**, 592-595.

Mosser, D.M. and Edwards, J.P. (2008) Exploring the full spectrum of macrophage activation, *Nature reviews. Immunology*, **8**, 958-969.

Motavaf, M., Safari, S. and Alavian, S.M. (2014) Interleukin 18 gene promoter polymorphisms and susceptibility to chronic hepatitis B infection: a review study, *Hepat Mon*, **14**, e19879-e19879.

Murray, M.*, et al.* (1984) Genetic resistance to African Trypanosomiasis, *The Journal of infectious diseases*, **149**, 311-319.

Muzzey, D., Evans, E.A. and Lieber, C. (2015) Understanding the Basics of NGS: From Mechanism to Variant Calling, *Current Genetic Medicine Reports*, **3**, 158-165.

Nag, A.*, et al.* (2015) Chromatin Signature Identifies Monoallelic Gene Expression Across Mammalian Cell Types, *G3 (Bethesda)*, **5**, 1713-1720.

Nagano, T.*, et al.* (2013) Cyclin I is involved in the regulation of cell cycle progression, *Cell Cycle*, **12**, 2617-2624.

Nagarajan, M., Nimisha, K. and Kumar, S. (2015) Mitochondrial DNA Variability of Domestic River Buffalo (Bubalus bubalis) Populations: Genetic Evidence for Domestication of River Buffalo in Indian Subcontinent, *Genome biology and evolution*, **7**, 1252-1259.

Nakagawa, N.*, et al.* (2000) ENH, Containing PDZ and LIM Domains, Heart/Skeletal Muscle-Specific Protein, Associates with Cytoskeletal Proteins through the PDZ Domain, *Biochemical and Biophysical Research Communications*, **272**, 505-512.

Nasr, M. (2017) The impact of cross-breeding Egyptian and Italian buffalo on reproductive and productive performance under a subtropical environment, *Reproduction in domestic animals = Zuchthygiene*, **52**, 214-220.

Naveena, B.M. and Kiran, M. (2014) Buffalo meat quality, composition, and processing characteristics: Contribution to the global economy and nutritional security, *Animal Frontiers*, **4**, 18-24.

Nazari-Ghadikolaei, A.*, et al.* (2018) Genome-Wide Association Studies Identify Candidate Genes for Coat Color and Mohair Traits in the Iranian Markhoz Goat, *Frontiers in Genetics*, **9**.

Nei, M. (2001) Genetic Distance. In Brenner, S. and Miller, J.H. (eds), *Encyclopedia of Genetics*. Academic Press, New York, pp. 828-832.

Ng, A. and Xavier, R.J. (2011) Leucine-rich repeat (LRR) proteins: integrators of pattern recognition and signaling in immunity, *Autophagy*, **7**, 1082-1084.

Nicoletti, P. (2001) Brucellosis in Animals. In Madkour, M.M. (ed), *Madkour's Brucellosis*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 267-275.

Nielsen, R.*, et al.* (2005) A Scan for Positively Selected Genes in the Genomes of Humans and Chimpanzees, *PLOS Biology*, **3**, e170.

Nielsen, R.*, et al.* (2007) Recent and ongoing selection in the human genome, *Nature reviews. Genetics*, **8**, 857-868.

Nielsen, R.*, et al.* (2011) Genotype and SNP calling from next-generation sequencing data, *Nature reviews. Genetics*, **12**, 443-451.

Nielsen, R.*, et al.* (2005) Genomic scans for selective sweeps using SNP data, *Genome research*, **15**, 1566-1575.

Nivsarkar, A.E., Tantia, M.S. and Vij, P.K. (2000) *Animal genetic resources of India : cattle and buffalo*. Directorate of Knowledge Management in Agriculture, Indian Council of Agricultural Research.

Noyes, H.*, et al.* (2011) Genetic and expression analysis of cattle identifies candidate genes in pathways responding to Trypanosoma congolense infection, *Proceedings of the National Academy of Sciences*, **108**, 9304.

O'Neill, L.A. (2005) Immunity's early-warning system, *Scientific American*, **292**, 24-31.

O'Neill, L.A., Golenbock, D. and Bowie, A.G. (2013) The history of Toll-like receptors - redefining innate immunity, *Nature reviews. Immunology*, **13**, 453-460.

Ojeda-Robertos, N.F.*, et al.* (2017) Study of gastrointestinal parasites in water buffalo (Bubalus bubalis) reared under Mexican humid tropical conditions, *Tropical animal health and production*, **49**, 613-618.

Ojeda, A.*, et al.* (2008) Selection in the making: a worldwide survey of haplotypic diversity around a causative mutation in porcine IGF2, *Genetics*, **178**, 1639-1652.

Oka, T.*, et al.* (2008) Identification of a novel protein MICS1 that is involved in maintenance of mitochondrial morphology and apoptotic release of cytochrome c, *Mol Biol Cell*, **19**, 2597-2608.

Oliver, C.P. (1958) Genetic Resistance to Disease in Domestic Animals, *BioScience*, **8**, 42-43.

Olsen, H.G.*, et al.* (2017) Genome-wide association mapping for milk fat composition and fine mapping of a QTL for de novo synthesis of milk fatty acids on bovine chromosome 13, *Genet Sel Evol*, **49**, 20.

Oosting, M.*, et al.* (2014) Human TLR10 is an anti-inflammatory pattern-recognition receptor, *Proceedings of the National Academy of Sciences of the United States of America*, **111**, E4478-4484.

Ouyang, Z.*, et al.* (2018) Accurate identification of RNA editing sites from primitive sequence with deep neural networks, *Scientific Reports*, **8**, 6005.

P. Bennet, S., Garcia, G. and P, L. (2010) *The buffalypso: The water buffalo of Trinidad and Tobago*.

Pai, A.A.*, et al.* (2009) Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data, *Bioinformatics*, **25**, 3207-3212.

Paixão, T.A., Martinez, R. and Santos, R.L. (2012) Polymorphisms of the coding region of Slc11a1 (Nramp1) gene associated to natural resistance against bovine brucellosis, *Arquivo Brasileiro de Medicina Veterinária e Zootecnia*, **64**, 1081-1084.

Palacios, R.*, et al.* (2009) Allele-Specific Gene Expression Is Widespread Across the Genome and Biological Processes, *PLOS ONE*, **4**, e4150.

Pandey, R.V.*, et al.* (2013) Allelic imbalance metre (Allim), a new tool for measuring allele-specific gene expression with RNA-seq data, *Mol Ecol Resour*, **13**, 740-745.

Pandey, S., Kawai, T. and Akira, S. (2014) Microbial sensing by Toll-like receptors and intracellular nucleic acid sensors, *Cold Spring Harbor perspectives in biology*, **7**, a016246.

Pant, S.D.*, et al.* (2007) Identification of single nucleotide polymorphisms in bovine CARD15 and their associations with health and production traits in Canadian Holsteins, *BMC Genomics*, **8**, 421.

Papayannopoulos, V. (2018) Neutrophil extracellular traps in immunity and disease, *Nature reviews. Immunology*, **18**, 134-147.

Park, E.*, et al.* (2017) Population and allelic variation of A-to-I RNA editing in human transcriptomes, *Genome biology*, **18**, 143.

Patel, M.P.*, et al.* (2010) Loop-swapped chimeras of the agouti-related protein and the agouti signaling protein identify contacts required for melanocortin 1 receptor selectivity and antagonism, *Journal of molecular biology*, **404**, 45-55.

Patel, S.*, et al.* (2017) Evolution and diversity studies of innate immune genes in Indian buffalo (Bubalus bubalis) breeds using next generation sequencing, *Genes and Genomics*, **39**, 1237-1247.

Patel, S.M.*, et al.* (2015) Exploring genetic polymorphism in innate immune genes in Indian cattle (Bos indicus) and buffalo (Bubalus bubalis) using next generation sequencing technology, *Meta gene*, **3**, 50-58.

Patnala, R., Clements, J. and Batra, J. (2013) Candidate gene association studies: a comprehensive guide to useful in silico tools, *BMC genetics*, **14**, 39-39.

Pausch, H.*, et al.* (2016) A multi-trait meta-analysis with imputed sequence variants reveals twelve QTL for mammary gland morphology in Fleckvieh cattle, *Genet Sel Evol*, **48**, 14-14.

Pawlikowski, B.*, et al.* (2017) Regulation of skeletal muscle stem cells by fibroblast growth factors, **246**, 359-367.

Pegolo, S.*, et al.* (2018) Integration of GWAS, pathway and network analyses reveals novel mechanistic insights into the synthesis of milk proteins in dairy cows, *Scientific reports*, **8**, 566-566.

Picardo, M. and Cardinali, G. (2011) The Genetic Determination of Skin Pigmentation: KITLG and the KITLG/c-Kit Pathway as Key Players in the Onset of Human Familial Pigmentary Diseases, *Journal of Investigative Dermatology*, **131**, 1182-1185.

Pickrell, J.K.*, et al.* (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing, *Nature*, **464**, 768-772.

Pickrell, J.K. and Pritchard, J.K. (2012) Inference of population splits and mixtures from genome-wide allele frequency data, *PLoS Genet*, **8**, e1002967.

Pinedo, P.J.*, et al.* (2009) Candidate gene polymorphisms (BoIFNG, TLR4, SLC11A1) as risk factors for paratuberculosis infection in cattle, *Prev Vet Med*, **91**, 189-196.

Pirinen, M.*, et al.* (2015) Assessing allele-specific expression across multiple tissues from RNA-seq read data, *Bioinformatics*, **31**, 2497-2504.

Piskol, R., Ramaswami, G. and Li, Jin B. (2013) Reliable Identification of Genomic Variants from RNA-Seq Data, *American Journal of Human Genetics*, **93**, 641-651.

Pollard, J.W. (2009) Trophic macrophages in development and disease, *Nature reviews. Immunology*, **9**, 259-270.

Poplin, R.*, et al.* (2018) Scaling accurate genetic variant discovery to tens of thousands of samples, *bioRxiv*, 201178.

Prasanth, S.G., Prasanth, K.V. and Stillman, B. (2002) Orc6 involved in DNA replication, chromosome segregation, and cytokinesis, *Science (New York, N.Y.)*, **297**, 1026-1031.

Pritchard, J.K., Pickrell, J.K. and Coop, G. (2010) The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation, *Current biology : CB*, **20**, R208-R215.

Purcell, S.*, et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses, *Am J Hum Genet*, **81**, 559-575.

Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features, *Bioinformatics*, **26**, 841-842.

R Core Team (2018) R: A Language and Environment for Statistical Computing.

Raetz, M.*, et al.* (2013) Cooperation of TLR12 and TLR11 in the IRF8-dependent IL-12 response to Toxoplasma gondii profilin, *Journal of immunology (Baltimore, Md. : 1950)*, **191**, 4818-4827.

Raghupathy, N.*, et al.* (2018) Hierarchical analysis of RNA-seq reads improves the accuracy of allele-specific expression, *Bioinformatics (Oxford, England)*, **34**, 2177-2184.

Ragnarsdóttir, B.*, et al.* (2010) Toll-like receptor 4 promoter polymorphisms: common TLR4 variants may protect against severe urinary tract infection, *PloS one*, **5**, e10734-e10734.

Reik, W. and Walter, J. (2001) Genomic imprinting: parental influence on the genome, *Nature reviews. Genetics*, **2**, 21-32.

Ribeiro, A.*, et al.* (2015) An investigation of causes of false positive single nucleotide polymorphisms using simulated reads from a small eukaryote genome, *BMC Bioinformatics*, **16**, 382.

Rivas, M.A.*, et al.* (2015) Human genomics. Effect of predicted protein-truncating genetic variants on the human transcriptome, *Science (New York, N.Y.)*, **348**, 666-669.

Rodrigues, R.T.d.S.*, et al.* (2017) Differences in Beef Quality between Angus (Bos taurus taurus) and Nellore (Bos taurus indicus) Cattle through a Proteomic and Phosphoproteomic Approach, *PLOS ONE*, **12**, e0170294.

Rodriguez-Valera, Y.*, et al.* (2018) Genetic diversity and selection signatures of the beef 'Charolais de Cuba' breed, *Sci Rep*, **8**, 11005.

Rollins, B.J. (1997) Chemokines, *Blood*, **90**, 909-928.

Romero-Salas, D.*, et al.* (2017) Seroepidemiology of Infection with Neospora Caninum, Leptospira, and Bovine Herpesvirus Type 1 in Water Buffaloes (Bubalus Bubalis) in Veracruz, Mexico, *Eur J Microbiol Immunol (Bp)*, **7**, 278-283.

Ronaghi, M.*, et al.* (1996) Real-time DNA sequencing using detection of pyrophosphate release, *Anal Biochem*, **242**, 84-89.

Rossi, U.A.*, et al.* (2019) Association of an IRF3 putative functional uORF variant with resistance to Brucella infection: A candidate gene based analysis of InDel polymorphisms in goats, *Cytokine*, **115**, 109-115.

Rourke, T. and Boeckx, C. (2018) Converging roles of glutamate receptors in domestication and prosociality, *bioRxiv*, 439869.

Rozowsky, J.*, et al.* (2011) AlleleSeq: analysis of allele-specific expression and binding in a network framework, *Molecular systems biology*, **7**, 522.

Rubin, C.-J.*, et al.* (2012) Strong signatures of selection in the domestic pig genome, *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 19529-19536.

Russell, C.D.*, et al.* (2012) Identification of single nucleotide polymorphisms in the bovine Toll-like receptor 1 gene and association with health traits in cattle, *Veterinary research*, **43**, 17.

Sabeti, P.C.*, et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure, *Nature*, **419**, 832.

Sabeti, P.C.*, et al.* (2006) Positive Natural Selection in the Human Lineage, *Science (New York, N.Y.)*, **312**, 1614.

Sabeti, P.C.*, et al.* (2007) Genome-wide detection and characterization of positive selection in human populations, *Nature*, **449**, 913-918.

Sahraeian, S.M.E.*, et al.* (2017) Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis, *Nature communications*, **8**, 59.

Saifi, H.W.*, et al.* (2004) Genetic Identity between Bhadawari and Murrah Breeds of Indian Buffaloes (Bubalus bubalis) Using RAPD-PCR, *Asian-Australas J Anim Sci*, **17**, 603-607.

Salavati, M.*, et al.* (2019) Elimination of Reference Mapping Bias Reveals Robust Immune Related Allele-Specific Expression in Crossbred Sheep, *Frontiers in Genetics*, **10**.

Salerno, A. (1974) The buffaloes of Italy. In Cockrill, W.R. (ed), *The husbandry and health of the domestic buffalo*. Food and Agriculture Organization of the United Nations, Rome, pp. 737-747.

Sarafidou, T.*, et al.* (2013) Toll Like Receptor 9 (TLR9) Polymorphism G520R in Sheep Is Associated with Seropositivity for Small Ruminant Lentivirus, *PLOS ONE*, **8**, e63901.

Satoh, K., Ginsburg, E. and Vonderhaar, B.K. (2004) Msx-1 and Msx-2 in Mammary Gland Development, *Journal of Mammary Gland Biology and Neoplasia*, **9**, 195-205.

Satya, R.V., Zavaljevski, N. and Reifman, J. (2012) A new strategy to reduce allelic bias in RNA-Seq readmapping, *Nucleic Acids Res*, **40**, e127-e127.

Schaart, J.G., Mehli, L. and Schouten, H.J. (2005) Quantification of allele-specific expression of a gene encoding strawberry polygalacturonase-inhibiting protein (PGIP) using PyrosequencingTM, *The Plant Journal*, **41**, 493-500.

Schroder, K.*, et al.* (2004) Interferon-gamma: an overview of signals, mechanisms and functions, *Journal of leukocyte biology*, **75**, 163-189.

Segal, A.W. (2005) How neutrophils kill microbes, *Annual review of immunology*, **23**, 197-223.

Seitz, J.J.*, et al.* (1999) A missense mutation in the bovine MGF gene is associated with the roan phenotype in Belgian Blue and Shorthorn cattle, *Mammalian genome : official journal of the International Mammalian Genome Society*, **10**, 710-712.

Selvam, R.M. and Archunan, G. (2017) A combinatorial model for effective estrus detection in Murrah buffalo, *Vet World*, **10**, 209-213.

Selvaraj, P.*, et al.* (2006) Promoter polymorphism of IL-8 gene and IL-8 production in pulmonary tuberculosis, *Current Science*, **90**, 952-954.

Serbina, N.V.*, et al.* (2008) Monocyte-mediated defense against microbial pathogens, *Annual review of immunology*, **26**, 421-452.

Sethi, R.S.*, et al.* (2011) Immunolocalization of Pulmonary Intravascular Macrophages, TLR4, TLR9 and IL-8 in Normal and Pasteurella multocida-infected Lungs of Water Buffalo (Bubalus bubalis), *Journal of Comparative Pathology*, **144**, 135-144.

Sethuraman, A. (2013) On inferring and interpreting genetic population structure - applications to conservation, and the estimation of pairwise genetic relatedness. *Ecology, Evolution, and Organismal Biology*. Iowa State University.

Sharma, B.S.*, et al.* (2015) Association of TLR4 polymorphisms with Mycobacterium avium subspecies paratuberculosis infection status in Canadian Holsteins, *Anim Genet*, **46**, 560-565.

Sherry, S.T.*, et al.* (2001) dbSNP: the NCBI database of genetic variation, *Nucleic Acids Res*, **29**, 308-311.

Shi, X.*, et al.* (2012) Cis- and trans-regulatory divergence between progenitor species determines gene-expression novelty in Arabidopsis allopolyploids, *Nature communications*, **3**, 950.

Shin, M.*, et al.* (2014) Mitochondrial MTHFD2L is a dual redox cofactor-specific methylenetetrahydrofolate dehydrogenase/methenyltetrahydrofolate cyclohydrolase expressed in both adult and embryonic tissues, *The Journal of biological chemistry*, **289**, 15507-15517.

Shinkai, H.*, et al.* (2011) Porcine Toll-like receptors: recognition of Salmonella enterica serovar Choleraesuis and influence of polymorphisms, *Molecular immunology*, **48**, 1114-1120.

Shrimpton, R.E.*, et al.* (2009) CD205 (DEC-205): a recognition receptor for apoptotic and necrotic self, *Molecular immunology*, **46**, 1229-1239.

Signorino, G.*, et al.* (2014) Role of Toll-like receptor 13 in innate immune recognition of group B streptococci, *Infection and immunity*, **82**, 5013-5022.

Singer-Sam, J. and Gao, C. (2002) Quantitative RT-PCR-Based Analysis of Allele-Specific Gene Expression. In Ward, A. (ed), *Genomic Imprinting: Methods and Protocols*. Humana Press, Totowa, NJ, pp. 145-152.

Singh, I. and Balhara, A.K. (2016) New approaches in buffalo artificial insemination programs with special reference to India, *Theriogenology*, **86**, 194-199.

Singh, S.V.*, et al.* (2008) Sero-prevalence of bovine Johne's disease in buffaloes and cattle population of North India using indigenous ELISA kit based on native Mycobacterium avium subspecies paratuberculosis 'Bison type' genotype of goat origin, *Comparative immunology, microbiology and infectious diseases*, **31**, 419-433.

Siracusa, M.C.*, et al.* (2013) Basophils and allergic inflammation, *J Allergy Clin Immunol*, **132**, 789-788.

Sirmaci, A.*, et al.* (2011) Mutations in ANKRD11 cause KBG syndrome, characterized by intellectual disability, skeletal malformations, and macrodontia, *Am J Hum Genet*, **89**, 289-294.

Smedley, D.*, et al.* (2009) BioMart – biological queries made easy, *BMC Genomics*, **10**, 22.

Smith, J.M. and Haigh, J. (1974) The hitch-hiking effect of a favourable gene, *Genetical research*, **23**, 23-35.

Smith, R.J.*, et al.* (2003) Identification of novel imprinted genes in a genome-wide screen for maternal methylation, *Genome research*, **13**, 558-569.

Sobrin, L.*, et al.* (2011) Candidate Gene Association Study for Diabetic Retinopathy in Persons with Type 2 Diabetes: The Candidate Gene Association Resource (CARe), *Investigative Ophthalmology & Visual Science*, **52**, 7593-7602.

Souza-Fonseca-Guimaraes, F., Adib-Conquy, M. and Cavaillon, J.-M. (2012) Natural killer (NK) cells in antibacterial innate immunity: angels or devils?, *Mol Med*, **18**, 270-285.

Sreenivas, D. (2013) *Breeding policy strategies for genetic improvement of cattle and buffaloes in India*.

Srinivasan, S. and Easterling, L. (2018) Prevalence of Bovine Tuberculosis in India: A systematic review and meta-analysis, **65**, 1627-1640.

Stachowiak, M., Szczerbal, I. and Flisikowski, K. (2018) Investigation of allele-specific expression of genes involved in adipogenesis and lipid metabolism suggests complex regulatory mechanisms of PPARGC1A expression in porcine fat tissues, *BMC genetics*, **19**, 107-107.

Stahl, P.D. and Ezekowitz, R.A.B. (1998) The mannose receptor is a pattern recognition receptor involved in host defense, *Current Opinion in Immunology*, **10**, 50-55.

Steinman, R.M. and Hemmi, H. (2006) Dendritic cells: translating innate to adaptive immunity, *Current topics in microbiology and immunology*, **311**, 17-58.

Stranger, B.E. and Dermitzakis, E.T. (2005) The genetics of regulatory variation in the human genome, *Hum Genomics*, **2**, 126-131.

Suárez-Vega, A.*, et al.* (2017) Variant discovery in the sheep milk transcriptome using RNA sequencing, *BMC Genomics*, **18**, 170.

Sun, W. (2012) A statistical framework for eQTL mapping using RNA-seq data, *Biometrics*, **68**, 1-11.

Sunder, S. (2018) India economic survey 2018: Farmers gain as agriculture mechanisation speeds up, but more R&D needed. Financial Express.

Surya, T.*, et al.* (2018) Genomewide identification and annotation of SNPs in Bubalus bubalis, *Genomics*.

Swirski, F.K.*, et al.* (2009) Identification of splenic reservoir monocytes and their deployment to inflammatory sites, *Science (New York, N.Y.)*, **325**, 612-616.

Takahashi, M. (2012) Heat stress on reproductive function and fertility in mammals, *Reproductive Medicine and Biology*, **11**, 37-47.

Talenti, A.*, et al.* (2017) Genomic Analysis Suggests KITLG is Responsible for a Roan Pattern in two Pakistani Goat Breeds, *Journal of Heredity*, **109**, 315-319.

Tanaka, S.*, et al.* (2018) Trim33 mediates the proinflammatory function of Th17 cells, *The Journal of Experimental Medicine*, **215**, 1853.

Tang, J.*, et al.* (2018) Expression Analysis of the Prolific Candidate Genes, BMPR1B, BMP15, and GDF9 in Small Tail Han Ewes with Three Fecundity (FecB Gene) Genotypes, *Animals : an open access journal from MDPI*, **8**, 166.

Taye, M.*, et al.* (2017) Exploring evidence of positive selection signatures in cattle breeds selected for different traits, *Mammalian genome : official journal of the International Mammalian Genome Society*, **28**, 528-541.

Terentiev, A.A. and Moldogazieva, N.T. (2013) Alpha-fetoprotein: a renaissance, *Tumor Biology*, **34**, 2075-2091.

Thoen, C.O., Steele, J.H. and Kaneene, J.B. (2014) *Zoonotic Tuberculosis: Mycobacterium bovis and Other Pathogenic Mycobacteria*. Wiley.

Tripal, P.*, et al.* (2007) Unique Features of Different Members of the Human Guanylate-Binding Protein Family, *Journal of Interferon & Cytokine Research*, **27**, 44-52.

Trivedi, K.R. (2000) Buffalo breeding programmes in India, *ICAR Technical Series*.

Tukiainen, T.*, et al.* (2017) Landscape of X chromosome inactivation across human tissues, *Nature*, **550**, 244-248.

Turner, S.D. (2014) qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots, *bioRxiv*, 005165.

Turro, E.*, et al.* (2011) Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads, *Genome biology*, **12**, R13.

Tycko, B. (1994) Genomic imprinting: mechanism and role in human pathology, *The American journal of pathology*, **144**, 431-443.

Vaidya, M.M.*, et al.* (2012) *Effect of ambient temperature rise on heat storage in murrah buffaloes during different seasons*.

van de Geijn, B.*, et al.* (2015) WASP: allele-specific software for robust molecular quantitative trait locus discovery, *Nature methods*, **12**, 1061-1063.

van Dongen, S. (2000) Graph Clustering by Flow Simulation. University of Utrecht.

van Furth, R.*, et al.* (1972) The mononuclear phagocyte system: a new classification of macrophages, monocytes, and their precursor cells, *Bulletin of the World Health Organization*, **46**, 845-852.

Varol, C.*, et al.* (2007) Monocytes give rise to mucosal, but not splenic, conventional dendritic cells, *The Journal of Experimental Medicine*, **204**, 171.

Verschoor, C.P., *et al.* (2010) Polymorphisms in the gene encoding bovine interleukin-10 receptor alpha are associated with Mycobacterium avium ssp. paratuberculosis infection status, *BMC Genet*, **11**, 23.

Vestal, D.J. and Jeyaratnam, J.A. (2011) The guanylate-binding proteins: emerging insights into the biochemical properties and functions of this family of large interferon-induced guanosine triphosphatase, *J Interferon Cytokine Res*, **31**, 89-97.

Villanueva, M., *et al.* (2018) Emerging Infectious Diseases in Water Buffalo: An Economic and Public Health Concern. In.

Villanueva, M.A., *et al.* (2016) Serological investigation of Leptospira infection and its circulation in one intensive-type water buffalo farm in the Philippines, *The Japanese journal of veterinary research*, **64**, 15-24.

Virtanen, C., Paris, J. and Takahashi, M. (2009) Identification and characterization of a novel gene, dapr, involved in skeletal muscle differentiation and protein kinase B signaling, *The Journal of biological chemistry*, **284**, 1636-1643.

Vitti, J.J., Grossman, S.R. and Sabeti, P.C. (2013) Detecting natural selection in genomic data, *Annual review of genetics*, **47**, 97-120.

Vo, A.H., *et al.* (2019) Dusp6 is a genetic modifier of growth through enhanced ERK activity, *Human molecular genetics*, **28**, 279-289.

Voegele, A.F., *et al.* (2002) Characterization of the vitamin E-binding properties of human plasma afamin, *Biochemistry*, **41**, 14532-14538.

Voight, B.F., *et al.* (2006) A Map of Recent Positive Selection in the Human Genome, *PLOS Biology*, **4**, e72.

Wang, B., David, M.D. and Schrader, J.W. (2005) Absence of caprin-1 results in defects in cellular proliferation, *Journal of immunology (Baltimore, Md. : 1950)*, **175**, 4274-4282.

Wang, D., Li, Y. and Shen, B. (2002) A novel erythroid differentiation related gene EDRF1 upregulating globin gene expression in HEL cells, *Chinese medical journal*, **115**, 1701-1705.

Wang, D.G., *et al.* (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome, *Science (New York, N.Y.)*, **280**, 1077-1082.

Wang, J., *et al.* (2015) Genome measures used for quality control are dependent on gene function and ancestry, *Bioinformatics*, **31**, 318-323.

Wang, X., Soloway, P.D. and Clark, A.G. (2011) A survey for novel imprinted genes in the mouse placenta by mRNA-seq, *Genetics*, **189**, 109-122.

Wang, Y.*, et al.* (2012) Ankrd17 positively regulates RIG-I-like receptor (RLR)-mediated immune signaling, *European Journal of Immunology*, **42**, 1304-1315.

Wang, Y.*, et al.* (2015) CARD15 Gene Polymorphisms Are Associated with Tuberculosis Susceptibility in Chinese Holstein Cows, *PLOS ONE*, **10**, e0135085.

Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics, *Nature reviews. Genetics*, **10**, 57-63.

Weir, B.S. and Cockerham, C.C. (1984) Estimating F-Statistics for the Analysis of Population Structure, *Evolution*, **38**, 1358-1370.

Weiss, G. and Schaible, U.E. (2015) Macrophage defense mechanisms against intracellular bacteria, *Immunol Rev*, **264**, 182-203.

Wesche, H.*, et al.* (1999) IRAK-M is a novel member of the Pelle/interleukin-1 receptor-associated kinase (IRAK) family, *The Journal of biological chemistry*, **274**, 19403-19410.

Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. Use R! Springer International Publishing.

Wilkinson, S.*, et al.* (2013) Signatures of Diversifying Selection in European Pig Breeds, *PLOS Genetics*, **9**, e1003453.

Williams, A.G.*, et al.* (2014) RNA-seq Data: Challenges in and Recommendations for Experimental Design and Analysis, *Current protocols in human genetics*, **83**, 11 13 11-20.

Williams, J.L.*, et al.* (2017) Genome assembly and transcriptome resource for river buffalo, Bubalus bubalis (2n = 50), *GigaScience*, **6**, 1-6.

Wittkopp, P.J. (2011) Using Pyrosequencing to Measure Allele-Specific mRNA Abundance and Infer the Effects of Cis- and Trans-regulatory Differences. In Orgogozo, V. and Rockman, M.V. (eds), *Molecular Methods for Evolutionary Genetics*. Humana Press, Totowa, NJ, pp. 297-317.

Wood, D.L.A.*, et al.* (2015) Recommendations for Accurate Resolution of Gene and Isoform Allele-Specific Expression in RNA-Seq Data, *PLOS ONE*, **10**, e0126911.

Wright, S. (1949) The genetical structure of populations, *Annals of Eugenics*, **15**, 323-354.

Wysoker, A., Tibbetts, K. and Fennell, T. (2013) Picard tools version 2.5.0.

Xu, C. (2018) A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data, *Computational and Structural Biotechnology Journal*, **16**, 15-24.

Xu, L.*, et al.* (2015) Genomic signatures reveal new evidences for selection of important traits in domestic cattle, *Molecular biology and evolution*, **32**, 711-725.

Xue, Y.*, et al.* (2009) Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree, *Current biology : CB*, **19**, 1453-1457.

Yamamoto, M.*, et al.* (2003) Role of adaptor TRIF in the MyD88-independent toll-like receptor signaling pathway, *Science (New York, N.Y.)*, **301**, 640-643.

Yan, H.*, et al.* (2002) Allelic variation in human gene expression, *Science (New York, N.Y.)*, **297**, 1143.

Yang, B.*, et al.* (2015) Pyrosequencing for accurate imprinted allele expression analysis, *J Cell Biochem*, **116**, 1165-1170.

Yang, J., Liu, Z. and Xiao, T.S. (2016) Post-translational regulation of inflammasomes, *Cellular &Amp; Molecular Immunology*, **14**, 65.

Yang, M.A. and Fu, Q. (2018) Insights into Modern Human Prehistory Using Ancient Genomes, *Trends in Genetics*, **34**, 184-196.

Yang, T.L.*, et al.* (2010) HMGA2 is confirmed to be associated with human adult height, *Annals of human genetics*, **74**, 11-16.

Yarovinsky, F.*, et al.* (2005) TLR11 activation of dendritic cells by a protozoan profilin-like protein, *Science (New York, N.Y.)*, **308**, 1626-1629.

Yeh, S.-D.*, et al.* (2002) Isolation and Properties of Gas8, a Growth Arrest-specific Gene Regulated during Male Gametogenesis to Produce a Protein Associated with the Sperm Motility Apparatus, *Journal of Biological Chemistry*, **277**, 6311-6317.

Yindee, M.*, et al.* (2010) Y-chromosomal variation confirms independent domestications of swamp and river buffalo, *Animal Genetics*, **41**, 433-435.

Young, R.*, et al.* (2018) Species-Specific Transcriptional Regulation of Genes Involved in Nitric Oxide Production and Arginine Metabolism in Macrophages, *ImmunoHorizons*, **2**, 27-37.

Young, R.*, et al.* (2019) A Gene Expression Atlas of the Domestic Water Buffalo (Bubalus bubalis), *Frontiers in Genetics*, **10**.

Yu, N.K.*, et al.* (2016) A transducible nuclear/nucleolar protein, mLLP, regulates neuronal morphogenesis and synaptic transmission, *Sci Rep*, **6**, 22892.

Yurchenko, A.A.*, et al.* (2018) Scans for signatures of selection in Russian cattle breed genomes reveal new candidate genes for environmental adaptation and acclimation, *Scientific Reports*, **8**, 12984.

Zaitoun, I. and Khatib, H. (2006) Assessment of genomic imprinting of SLC38A4, NNAT, NAP1L5, and H19 in cattle, *BMC genetics*, **7**, 49-49.

Zehsaz, F. (2015) The role of IL-8 gene polymorphism and susceptibility to upper respiratory tract infection among endurance athletes, *Pejouhesh dar Pezeshki (Research in Medicine)*, **39**, 127-132.

Zembrzuski, V.M.*, et al.* (2010) Cytokine genes are associated with tuberculin skin test response in a native Brazilian population, *Tuberculosis (Edinburgh, Scotland)*, **90**, 44-49.

Zhang, F.*, et al.* (2012) Metalloreductase Steap3 coordinates the regulation of iron homeostasis and inflammatory responses, *Haematologica*, **97**, 1826-1835.

Zhang, M.*, et al.* (2018) Characterizing cis-regulatory variation in the transcriptome of histologically normal and tumor-derived pancreatic tissues, *Gut*, **67**, 521-533.

Zhang, Y.*, et al.* (2016) Strong and stable geographic differentiation of swamp buffalo maternal and paternal lineages indicates domestication in the China/Indochina border region, *Molecular ecology*, **25**, 1530-1550.

Zhao, F.*, et al.* (2015) Detection of selection signatures in dairy and beef cattle using high-density genomic information, *Genetics Selection Evolution*, **47**, 49.

Zhao, S.*, et al.* (2018) Strategies for processing and quality control of Illumina genotyping arrays, *Brief Bioinform*, **19**, 765-775.

Zhou, J.*, et al.* (2011) Copy-number variation: the balance between gene dosage and expression in Drosophila melanogaster, *Genome biology and evolution*, **3**, 1014-1024.

Zhuo, Z., Lamont, S.J. and Abasht, B. (2017) RNA-Seq Analyses Identify Frequent Allele Specific Expression and No Evidence of Genomic Imprinting in Specific Embryonic Tissues of Chicken, *Scientific Reports*, **7**, 11944.

Zimin, A.V.*, et al.* (2013) The MaSuRCA genome assembler, *Bioinformatics*, **29**, 2669-2677.

Zwemer, L.M.*, et al.* (2012) Autosomal monoallelic expression in the mouse, *Genome biology*, **13**, R10-R10.