University of Glasgow

College of Veterinary, Medical & Life Sciences

The polyclonal antibody response to FMDV in cattle and African buffalo

Rebecca L Philp

A thesis submitted in partial fulfilment for the degree of Doctor of Philosophy

December 2017

The polyclonal antibody response to FMDV in cattle and African buffalo

Thesis Abstract

Protection against the highly contagious foot and mouth disease virus (FMDV) coincides with neutralising antibody titres. Infection in cattle is characterised by 100% morbidity of an acute vesicular disease whilst infection of their closest relative, the African buffalo, is sub-clinical despite having diverged only 5.7 – 9.3 million years ago (Glanzmann et al., 2016; 1). The germline and antibody repertoire in African buffalo has not previously been characterised and so the cause of their differential disease response may be the production of a more specific and / or avid antibody response to FMDV than cattle.

The cattle and African buffalo antibody germline was sought to characterise the recombinatorial potential of the antibody loci and their subsequent primary antibody repertoire. Expression of the antibody heavy chain (IGH) and antibody lambda light chain (IGL) was investigated with qPCR and RNA-seq. The antibody repertoire in response to FMDV infection was interrogated in African buffalo infected with SAT1 FMDV and compared to the cattle IGH repertoire inoculated with highly purified SAT1 FMDV antigen.

The recombinatorial potential of the cattle and African buffalo IGH and IGL is severely limited compared to other species such as mice and human. The characterisation of the cattle IGH and IGL is the most accurate to date and reveals internal duplications of the IGH, disrupting the expected *IGHV-IGHD-IGHJ-IGHC* ordering seen in mammalian immune loci and resulting in four *IGHD* regions, containing long and ultra-long *IGHD*. These *IGHD* provide a novel diversification mechanism that can compensate for limited germline diversity by forming long and ultra-long CDR3H loops that are highly diverse in their length and amino acid composition. The African buffalo antibody repertoire also forms highly diverse long and ultra-long CDR3H, despite lack of evidence for the existence of the duplications in their IGH. Limited variability is seen in the length and amino acid composition of the IGL in both species, suggesting they are playing a structural role to support these unusual long and ultra-long CDR3H. In response to FMDV infection in African buffalo, a dramatic increase in specific long and ultra-long CDR3H sequence abundance occurs but this change in frequency of specific transcripts is absent in cattle. The differential antibody response may account for the protection of African buffalo against FMD.

Declaration

This thesis, and the work contained within it, was conducted from October 2013 to September 2017 by myself, unless stated otherwise. No part of this thesis has been submitted for another degree.

Rebecca Leann Philp

Acknowledgments

For all of those who have guided and supported me throughout the past four years I am extremely grateful.

Firstly, I would like to thank my supervisors, Prof John Hammond, Dr John Schwartz and Dr Richard Reeve for the opportunity to carry out this research and for their continued support and guidance throughout. It has been a steep learning curve from beginning to end and I thank you for your patience, help and encouragement. I would also like to thank the members of the Immunogenetics group at The Pirbright Institute, past and present, for their advice and support. In particular I would like to thank my friend and fellow PhD student, Richard Borne, whom I would not have finished my PhD without. Thanks to his bioinformatic teachings and continually feeding me food, I made it through to the end. I would also like to thank the wider community at The Pirbright Institute, the friends I have made and all the other staff who have supported my work. Special thanks are owed to Dr Clare Grant who guided me through the African buffalo sequencing and to Dr Eva Perez who carried out the African buffalo infection in the Kruger National Park and buddied me inside containment in order to attain my antibody transcripts for sequencing.

Thanks are due to the BBSRC for enabling this PhD project through their sponsorship. I would also like to thank the collaborators whose work went into this project. Dr Bridgette Glanzmann who shared with me the African buffalo genome reads, Dr Yfke Pasman for the Holstein cattle RNA transcriptomes, Dr Li Ma for sharing their cattle immunoglobulin heavy chain locus assembly and Dr Tim Smith for the new Pacific Biosciences cattle genome assembly.

This PhD has been an exciting and extremely challenging journey. I started with an undergraduate degree in medicinal chemistry and am now finishing with a diverse set of skills in veterinary immunogenetics. The completion of this PhD would not have been possible without the loving support of my family and friends. My parents who have always encouraged me to work hard and believe in myself, I hope I have made you proud. Thank you to my friends, for your understanding and kindness and for always making me smile. Thank you to my partner Antony, your encouragement has kept me going and your patience and support has kept me from going insane! Finally thank you to my Mr. Darcy, the most educated labradoodle in immunogenetics, whose presence has been a comfort every day.

For Antony

Contents

List of figures:

List of Tables:

List of Abbreviations:

AID	Activation induced cytosine deaminase
APRIL	Proliferating-inducing ligand
BAC	Bacterial artificial chromosome
BAFF	B cell activating Factor
BCR	B cell receptor
BTA	*Bos taurus*
CDR	Complementarity determining region
CFT	Complement fixation test
CFU	Colony forming unit
CHORI	Childrens Hospital Oakland Research Institute
CSR	Class switch recombination
CSR	Class switch recombination
DC	Dendritic cell
Dpi	Days post infection
ELISA	Enzyme linked immunosorbent assay
FCR	Fc receptor
FMD	Foot and mouth disease
FMDV	Foot and mouth disease virus
FR	Framework
GALT	Gut associated lymphoid tissue
GC	Germinal centre
HMW	High molecular weight
HR	Homologous recombination
IFN	Interferon
IGH	Antibody heavy chain locus
IGK	Antibody kappa light chain locus
IGL	Antibody lambda light chain locus
IPP	Ileal Peyers Patch
JPP	Jejunal Peyers Patch
KDE	Kappa deleting element

KNP	Kruger National Park
MHC	Major histocompatibility complex
MMR	Mismatch repair
Mφ	Macrophage
NHEJ	Non-homologous end joining
NK	Natural killer cell
ORF	Open reading frame
PAMP	Pathogen associated molecular pattern
PBMC	Peripheral blood mononuclear cells
PCR	Polymerase chain reaction
PFGE	Pulsed field gel electrophoresis
PRR	Pattern recognition receptor
QUAST	A quality assessment tool for genome assemblies
RAG	Recombination activating gene
RGD	Arginine-glycine-aspartic acid
RLR	RIG-I like receptor
RS	Recombination signal sequence
SAT	Southern African territory
SCS	Subcapsular sinus
SHM	Somatic hyper-mutation
SLC	Surrogate light chain
SMRT	Single molecule real time
SNA	Serum neutralising antibody
SNP	Single nucleotide polymorphism
TdT	Terminal deoxynucleotidyl transferase
TLR	Toll-like receptor
USDA	United States Department of Agriculture
VP	Viral structural proteins
YAC	Yeast artificial chromosome
ZMW	Zero mode waveguide

Chapter 1

Introduction

Introduction

Foot and mouth disease (FMD) is the most infectious veterinary disease agent known, caused by the aetiological agent foot and mouth disease virus (FMDV). The disease is endemic in most of the developing world and sporadic outbreaks in naïve populations in countries free from disease cost billions to control. The outbreak in 2001 in the UK cost an estimated £7.8 billion and 6.1 million animals infected or in contact were slaughtered (Jonathan Rushton, 2013; 2). This epidemic put pressure on governments to develop alternative control strategies to the policy of mass slaughter that requires improved vaccination strategies and control methods. Current costs of controlling FMDV are immense, the annual impact is $5 billion in terms of production losses and vaccination alone (FAO, 2008; 3). The OIE World Organisation for Animal Health have listed FMDV as a significant threat; the virus has the potential for rapid and extensive spread within and between countries and can cause severe economic impact. FMDV is widely prevalent, estimated to circulate in 77% of the global livestock population. Much of the global burden of FMDV currently falls on the world's poorest who rely on the health of their livestock; 30% of the population in African countries live on livestock (Leboucq, 2013; 4), creating a considerable food security issue. The clinically acute vesicular disease is severe with a wide host range which affects all non-avian livestock as well as over 70 other wildlife species that includes the long-term maintenance host, the African buffalo (*Syncerus caffer*). Ruminants facilitate long-term persistence of the virus but African buffalo are the natural reservoir of the disease. Infection of buffalo is often asymptomatic but they continually generate antigenic variants that transmit to cattle populations. Cattle, despite having diverged from African buffalo around 5.7-9.3 million years ago (Glanzmann et al., 2016; 1), display a differential disease response with 100% morbidity and a >50% mortality rate in their young (Leboucq, 2013; 4). Protection against disease is mediated by the antibody response and clearance of the virus coincides with antibody titres. African buffalo may therefore be producing a more specific or avid antibody repertoire than cattle but very little is known about the African buffalo genome or their immune response to FMDV. Clinical manifestations of FMD in livestock involve an acute febrile reaction with severe vesicular lesions that cause lameness and yield reductions in a herd. The highly contagious nature of the virus, its wide dissemination, maintenance in wildlife species and colossal economic impact has meant that FMD is one of the most feared veterinary diseases.

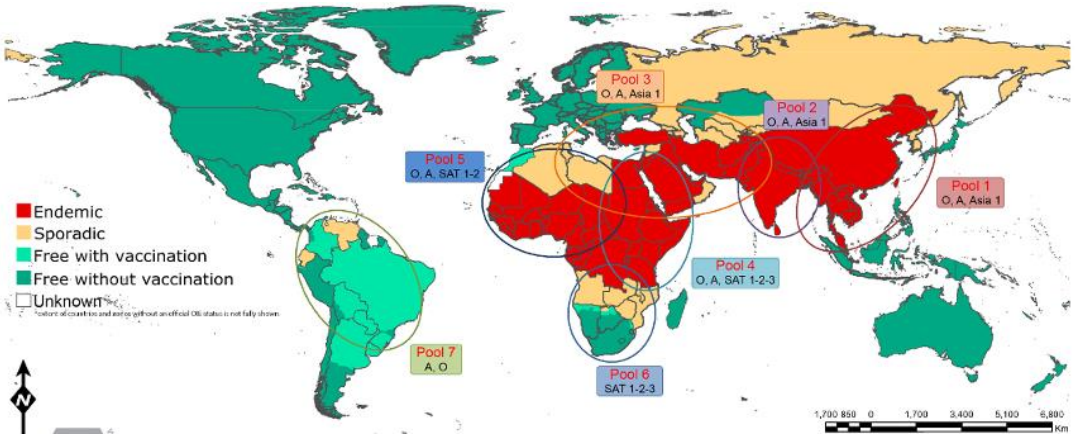


Figure 1.1: Schematic map showing the global distribution of the endemic FMDV serotype pools; O, A, C, Asia 1, SAT1, SAT 2 and SAT3. The conjectured status of each country in 2015 is displayed; the disease is endemic in most of Africa, Asia and the Middle East whilst sporadic outbreaks occur in countries free from disease. Image taken from FAO Reference Laboratory (FAO, 2017; 5)

1.1 FMDV: the virus

1.1.1 Global diversity of FMDV serotypes

FMDV is classified within the Apthovirus genus as a member of the Picornaviridae family (Belsham, 1993; 6) and contains seven different serotypes; A, O, C, Asia1, SAT1, SAT2 and SAT3, with varying global distributions of each (Figure 1.1). The seven serotypes cluster into distinct genetic lineages with 30-50% nucleotide difference in *VP1* (viral protein 1) (Jackson et al., 2003; 7). Serotypes O and A have the broadest distribution, occurring in Africa, south Asia and South America. Asia 1 viruses are much less genetically diverse than other serotypes, suggesting its recent origin in Asia (Zhang et al., 2015; 8) whilst Type C is limited to the Indian sub-continent, appearing intermittently, and seems to be becoming extinct due to a lack of natural reservoirs (Nagendrakumar et al., 2005; 9). In general, only cattle appear infected by FMDV serotypes other than O, although this may reflect its wider global distribution. Only the SAT serotypes have a long-term host; there is no known species that the other serotypes associate with for extended periods of time (Paton et al., 2009; 10). The

SAT serotypes are largely restricted to Africa where they remain in their maintenance host, the African buffalo. FMDV most likely evolved from a progenitor that infected African buffalo up to 1000 years ago and has since diverged and spread. SAT2 is the most widely distributed in Africa and the cause of most outbreaks in cattle populations. However, SAT1 is the most frequently isolated serotype from African buffalo (Nick Knowles (The Pirbright Institute), personal communication). There is a high degree of antigenic variability between each serotype and so limited cross protection exists. The serotypes are clinically indistinguishable, but each strain has unique antigenic and epidemiological characteristics.

1.1.2 FMDV structure

FMDV has a single-stranded positive sense RNA genome about 8.4 kb in length that encodes a large polyprotein. The genome is translated in a single open reading frame into viral polyprotein that is proteolytically cleaved into 15 different mature proteins plus a variety of precursors. The leader (L) protein and three precursor proteins are generated first and the subsequent processing of the precursors lead to the four viral structural proteins (VP1, VP2, VP3 and VP4) and nine additional non-structural proteins (2A, 2B, 2C, 3A, 3B1, 3B2, 3B3, 3C and 3D). An icosahedral capsid consisting of 60 copies each of the four structural proteins, VP1, VP2, VP3 and VP4, fold into eight-stranded β-barrels which fit together to form the viral shell whilst VP4 is buried within the capsid (Grubman and Baxt, 2004; 11). The picornavirus RNA is infectious; the viral proteins are not required to initiate the cycle of infection which occurs in the cytoplasm of cells as the virus takes over the cellular machinery (Chase and Semler, 2012; 12). The non-structural proteins promote viral production and block host macromolecular synthesis. The L and 3C proteinases and the viral peptide 2A process the viral polyprotein whilst the viral genome is replicated by 3D, the viral RNA-dependent RNA polymerase (Saunders and King, 1982; 13). The L protein cleaves host proteins including the translation initiation factor eIF4G and blocks host transcription by cleaving the transcription nuclear factor NF-κB (Vakharia et al., 1987; 14, Grigera, 1984; 15). This ultimately limits the host innate response by inhibiting the induction of IFNα and IFNβ mRNA (Chinsangaram et al., 2001; 16). The trafficking of proteins through the endoplasmic reticulum and Golgi complex is also blocked by 2B and 2C (Moffat et al., 2005; 17, Moffat et al., 2007; 18). Together, a decrease in MHC class I cell surface expression delays the host

adaptive immune response. FMDV then, inhibits the induction of antiviral molecules at both transcriptional and translational levels to delay both innate and adaptive host immunity.

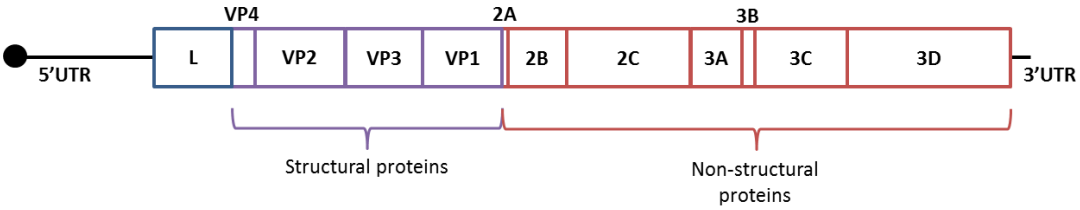


Figure 1.2: The structure of foot and mouth disease virus single-stranded RNA genome. The structural proteins are outlined in purple whilst the non-structural proteins are outlined in red. Image adapted from (Balinda et al., 2010; 19)

1.1.3 Receptor mediated cell entry

The integrin receptors are the major receptors for FMDV in cell culture. The type I heterodimeric membrane proteins are involved in host cell adhesion and migration (Hynes, 2002; 20) and the αvβ6 and αvβ8 receptors are constitutively expressed on the surface of epithelial cells in cattle which are targeted by FMDV(Jackson et al., 2000; 21, Monaghan et al., 2005; 22). The αvβ1 is present at predilection sites for vesicular lesion formation during infection (Monaghan et al., 2005; 22), suggesting these integrins are requisite receptors for FMDV infection and can be attributed to the cellular tropism of viral replication. However, FMDV does not bind to and replicate in numerous other cells with high αvβ6 expression, suggesting other cell or tissue specific factors are involved as co-determinants of cell tropism. FMDV has also been shown to enter cultured cells via Fc receptor antibody-mediated uptake (Baxt and Mason, 1995; 23), but it is unclear if this occurs in the host animal. Following extended tissue culture passages, FMDV becomes able to enter cells through the ubiquitously expressed heparin sulphate proteoglycan but this also has not been shown *in vivo* (Jones et al., 2005; 24). A surface exposed loop exists which connects the βG and βH strands, known as the G-H loop, on VP1. A highly conserved arginine-glycine-aspartic acid (RGD) sequence is contained on this exposed loop which is a recognition sequence for the integrin family of cell surface receptors that enables binding and subsequent internalisation into acidic

endosomes for cell entry (Pierschbacher and Ruoslahti, 1984; 25). The viral capsid is acid labile (G.J. Belsham, 2009; 26) and upon entry into cells through the endocytic pathway, the acidic pH stimulates viral capsid dissociation and release of the viral RNA for replication.

1.1.4 FMDV Pathogenesis

FMD has a wide host range affecting all artiodactyls, the even toed ungulates, which includes domestic cattle, sheep, pigs and multiple wild species such as the African buffalo. Species such as rats, guinea pigs and rabbits are susceptible under experimental conditions but do not play a role in the natural epidemiology of the disease. During acute infection, clinical signs in susceptible species vary significantly depending on the species, age and husbandry of affected animals and on the infecting FMDV strain. An acute febrile reaction causes vesicular lesions around the mouth and feet which leads to lameness, lowered food intake and therefore loss of weight and draught power (figure 1.3); chronic FMD will typically reduce milk yield by 80% (Bayissa et al., 2011; 27). Morbidity in adult animals is 100% whilst mortality is generally low, and animals recover within two weeks. However, abortions and death in young animals is high (>50% of young calves) due to viral replication in the myocardium, causing respiratory distress and heart failure (Skinner, 1951; 28). Long term effects are also observed such as a loss of fertility and development of chronic heat intolerance syndrome. Disease in African buffalo however is often asymptomatic, despite being continually infected with the SAT serotypes. Cattle and African buffalo are closely related species, having diverged only 5.7-9.3 million years ago (Glanzmann et al., 2016; 1), and so the cause of their differential disease response to FMDV is unknown.

Figure 1.3: Images of FMDV lesions in cattle within the hoof cleft (left) and on the gums (right) taken during an outbreak in Kenya. Photographs taken by Nicholas Juleff (The Pirbright Institute).

1.2 FMDV transmission

1.2.1 Routes of infection

The highly contagious nature of FMDV means the virus spreads rapidly through susceptible animals. The most common mechanisms of viral spread are the direct contact by mechanical transfer of virus from infected animals or the inhalation of droplet nuclei in the respiratory tract of susceptible animals. The short distance transfer of droplet nuclei can also be extended. Under certain climatic conditions, including high relative humidity and minimal turbulence of the air, droplet nuclei can travel hundreds of kilometres (Donaldson et al., 1982; 29). Ruminant species are highly susceptible to respiratory infection with FMDV and require as little as ten tissue culture 50% infective doses (TCID50) (Donaldson, 1979; 30). Contact transmission can also occur indirectly via contaminated vehicles, persons or fomites. All animal excretions and vesicular secretions contain replicative virus during viraemic phase of disease (Alexandersen et al., 2003; 31), although this is shorter than previously recognised and animals do not appear infectious until 0.5 days after the appearance of clinical signs, with the average infectious period as short as 1.7 days (Charleston et al., 2011; 32). It takes only 3-5 days for a newly infected animal to become infectious and those infected are capable of spreading the disease to over 70 other susceptible individuals in a herd (Woolhouse et al., 1996; 33).

1.2.2 Kinetics of FMDV replication

Whether by direct contact of the virus or by inhalation of viral nuclei, once an animal has become infected the initial viral replication occurs at the primary site of infection. The epithelial cells on the dorsal soft palate within the oesophago-pharyngeal region are the primary site of infection in cattle. FMDV can replicate in the pharynx for 1-3 days (Burrows, 1968; 34), reaching its peak as early as 2-3 days after exposure (Ferguson et al., 2001; 35) before spreading to the circulation though the lymphoid system. Viraemia usually lasts 4-5 days (Alexandersen, 2003; 36) in which the virus can then spread to secondary sites and a fever of >40°C ensues. Subsequent viral amplification in secondary sites includes the squamous epithelia of the skin and mouth or in the myocardium of young animals. Viral clearance from the blood coincides with the antibody response and antibody mediates protection against disease however, the virus persists in peripheral sites such as the oesophageal region of the throat after clinical signs have abated. Clearance of the virus from these external sites is less efficient and these carrier animals are capable of transmitting the virus to naïve animals.

1.2.3 Long term maintenance hosts of FMDV

In 60% of ruminant species infected with FMDV the virus continues to replicate in the osophago-pharyngeal region, the primary site of respiratory infection, supporting persistence of the virus. If the virus is still detectable after 28 dpi, the animal is defined as a carrier (Van Bekkum J., 1959; 37). Detectable virus disappears from all secretions with the exception of 60% of animals having low level replication, recoverable from oesophago-pharyngeal secretions. The virus maintains a high prevalence in the dorsal nasopharynx and dorsal soft palate (Pacheco et al., 2015; 38). An additional site of viral persistence is the germinal centres of lymphoid tissues where virus localises and is maintained in a non-replicative state. These typically include the light zone of germinal centres in the dorsal soft palate, pharyngeal tonsil and pharyngeal lymph nodes post-infection (Juleff et al., 2012; 39). The maintenance of non-replicating FMDV in these sites is a source of persisting antigens and contributes to the generation of a longer antibody memory. However, the probability of recovering virus from

cattle more than 12 months after infection is extremely low and the persistence of virus in cattle seems affected by age, as the virus persists longer in younger animals (Bronsvoort et al., 2016; 40).

African buffalo are the only real long-term carriers of the disease. African buffalo infected with FMDV are predominantly asymptomatic but these animals constantly generate antigenic variants of the virus which they shed, causing sporadic outbreaks in cattle and in wildlife populations such as impala. Upon infection in buffalo, antigenic variation is measurable after only four days and immune pressure is not a prerequisite for antigenic change (Vosloo et al., 1996; 41). The virus accumulates mutations at an approximate rate of 1.64% nucleotide substitutions per year, resulting in genetic variation of FMDV being constantly generated in individual buffalo (Vosloo et al., 1996; 41). A single animal can harbour significant FMDV diversity and the rate of change increases upon transmission to other animals. Infection of naïve buffalo with SAT virus results in excretion of aerosolised virus which persists for longer than in acutely infected cattle (Bengis et al., 1986; 42). More than one SAT FMDV serotype may be maintained in individual buffalo and persistently infected buffalo are refractory to re-infection with the same virus strain (Maree et al., 2016; 43). Individual animals carry the disease for up to 5 years and the virus can persist in isolated herds for over 20 years (Condy et al., 1985; 44).

African buffalo are herd animals which remain constant for years, except for individuals which migrate hundreds of kilometres from the breeding herd (Naidoo et al., 2012; 45). Calves are born predominantly in the summer, when rainfall is highest and therefore grazing is more opportune. Calves are protected against infection from maternal immunity passed on in colostrum; circulating antibodies in new-born calves are similar levels to those of their dams and persist for 2-7 months (Condy and Hedger, 1974; 46). As maternal antibodies wane, so does protection from infection and so sporadic outbreaks in young buffalo occur 3-4 months after birth; the maternal protection does not persist for longer because high levels of antibody are required. However, unlike young cattle, young buffalo recover in less than 2 weeks. Transmission of FMDV from dams to their calves is sporadic and outbreaks occur due to minor epidemics in young animals (Bengis et al., 1986; 42). Infection in calves in the Kruger National Park (KNP) in South Africa is usually with SAT1 first, then SAT2 and lastly SAT3. Small numbers of calves are born throughout the year also, thus maintaining a supply of susceptible animals.

1.2.4 FMDV transmission from African buffalo to cattle

FMDV spreads predominantly by contact with infected animals, their secretions or contaminated food products and the virus is capable of becoming aerosolised and travelling extensive distances. The natural infection of cattle appears to occur by the respiratory route (Donaldson, 1979; 30). Transmission of SAT FMDV between individual buffalo appears to occur by either contact transmission between acutely infected and susceptible individuals, which occurs for the majority of infections, or the occasional transmission between persistently infected buffalo and susceptible individuals.

Carrier cattle rarely transmit infection to cohorts in close contact but it is shown unequivocally that carrier buffalo are able to transmit to other buffalo (Condy and Hedger, 1974; 46) and cattle (Vosloo et al., 1996; 41, Dawe et al., 1994; 47). Viruses causing FMD in cattle were highly similar and hence resulted from the carrier African buffalo in which they were in contact with. Carrier buffalo and cattle require direct or close contact over a period of several weeks or months for transmission to occur. Precisely how transmission between carrier animals to susceptible animals occurs is unknown although sexual transmission has been hypothesised (Thomson et al., 2003; 48). However, virus was isolated from only 1 of 108 sheath washes of buffalo males' ages 3-5 years and 1 of 23 testes samples. No virus was isolated from female animal reproductive tracts (Vosloo, unpublished results). It is not known if these animals were persistently infected but these results indicate that the presence of virus in the reproductive tracts of buffalo is rare. Carrier animals shed FMDV following immune activation as intact FMDV particles retained on the follicular dendritic cell network release the virus (Juleff et al., 2012; 39). The most likely cause of a majority of outbreaks in cattle is the close contact with naive juvenile buffalo infected with FMDV as their maternal antibodies wane; these animals have high levels of viral secretion. The precise mechanism is still however, uncertain.

1.2.5 FMDV control measures

FMDV is considered the most infectious animal disease spreading rapidly through susceptible populations. Outbreaks of the disease are controlled by vaccination strategies in countries where FMD is endemic and when sporadic outbreaks occur in countries free from disease. Accurate diagnostic tests provide the means to monitor the disease. Viral detection assays distinguish FMDV in collected clinical samples from animals; the enzyme-linked immunosorbent assays (ELISAs) detect viral antigen and real-time reverse-transcription polymerase chain reaction (rRT-PCR) detects viral genome RNA. Serological tests are also widely used to monitor the immune status of an infected animal which include ELISAs, virus neutralising tests (VNTs) and complement fixation tests (CFT). Commercial vaccine strategies aim to induce an antibody response against inactivated whole-virus particles. However, current vaccinations provide only limited protection against disease and need to be administered every 6 months. The current commercial FMDV vaccines are a suspension of whole virus, inactivated with aziridine and mixed with an oil or aluminium hyrodxide/saponin adjuvant. As the vaccines are inactivated and do not contain replicative virus, the expression of the non-structural proteins does not occur so antibodies to these proteins do not develop in vaccinated animals. As a result, the serological tests can distinguish between a vaccinated animal and one that has been infected.

1.3 The immune response to FMDV

1.3.1 Immune response in laboratory animals

The interaction of the immune system with FMDV is not completely understood. Large animal experimentation has high costs, as well as incomplete knowledge of their immune system and a lack of immune reagents. Laboratory animal models have been used to model FMDV immune responses as they are capable of supporting viral replication. Mice are the most widely used laboratory animals for FMDV modelling as they are cost efficient and can be genetically manipulated. Guinea pigs, however, were the first lab animal to be successfully

inoculated with FMDV (Waldman O., 1920; 49) and the challenged animals developed generalised disease. Extensive vesicles develop at the inoculation site and viremia occurs leading to viral replication at secondary sites causing vesicles of the tongue, leading to salivation and weight loss. After 4-5 days the vesicles heal and animals are pyrexic for a short period as viremia is cleared rapidly by a coinciding antibody response (Knudsen et al., 1979; 50). Mortality rates in guinea pigs are low and transmission between inoculated and healthy guinea pigs did not occur, showing they do not shed the virus. These laboratory models do not play a role in the natural epidemiology of the disease

Young mice, only 1-2 weeks old, inoculated intraperitoneally, led to a fatal infection characterised by muscular paralysis and degeneration of the myocardium and skeletal muscles (Skinner, 1951; 28). Respiratory distress occurred within 24-48 hours post infection and death from myocarditis ensued. Susceptibility of the mice waned with increasing age and infection in mice over 3 weeks old was subclinical. The age-related host factors can be attributed to age related myotropism. Certain viruses, seemingly including FMDV, utilise the receptor αvβ3 integrin for cell entry and these receptors are expressed on young skeletal muscle cells and disappear as the animals age (Roivainen et al., 1994; 51). In adult mice, the virus primarily replicates in the pancreas. Viremia lasts 2-3 days with serum neutralising antibody coinciding with viral clearance but the severe long-term damage to the pancreas ensues which is still clearly visible at 21 dpi and is associated with loss of pancreatic function (Skinner, 1951; 28).

Laboratory animal models have helped our understanding of FMDV pathogenesis in an accelerated time frame, assisting the characterisation of the immune response. In mice, FMDV is a T-independent antigen which led research into the role of T cells in the disease response. These models however, are unnatural routes of infection and have significant differences in their antibody response compared to ruminants. Differences in how ruminants generate their antibody repertoire suggest that ruminants respond differently to FMDV than laboratory models therefore, more work is needed to understand the interaction of FMDV with their natural hosts.

1.3.2 T cell depletion has no effect on early stages of infection

The initial immune response to FMDV is T-independent; a protective immune response develops without stimulation of T cells. In the absence of T cells, B cell activation relies on Dendritic cells (DCs) for stimulating proliferation and inducing class switch recombination (CSR). Athymic mice infected with FMDV have near identical viremia, serum neutralising antibody (SNA) response and tissue viral clearance in the first 14 dpi as euthymic controls (Borca et al., 1986; 52). The long-term kinetics of the response, however, differed. Antibody titres decreased in athymic mice at 14 dpi and continued to lessen whilst titres in euthymic mice continued to increase to 240 dpi (Lopez et al., 1990; 53). T cells are thus required for maintaining high titres of SNA post infection.

Partial CD8+ T cell depletion and complete CD4+ T cell depletion in cattle had no discernible effects on clinical signs or infection kinetics (Juleff et al., 2009; 54). The complete CD4+ depletion inhibited antibody production to the G-H loop peptide and non-structural proteins but this did not affect the SNA response, class switching or viral clearance in early stages of infection. The role of the cytotoxic CD8+ T cells, which are stimulated by antigen presentation of infected cells on MHC class I, is currently unknown. Infected cells have a short life span and FMDV down regulates MHC class I expression on susceptible cells (Sanz-Parra et al., 1998; 55) which would impair CD8+ T cell responsiveness. On the other hand, this down regulation of MHC class I stimulates innate cytotoxic natural killer (NK) cells (see section 1.9.7). The contribution of T cell mediation in the maintenance of long-term memory has not been investigated in cattle but it is likely, considering the inability of vaccines to produce long term immunity due to lack of T cell stimulation, that T cells are required.

1.3.3 Overview of the innate immune response to FMDV

As a whole, the adaptive response is antigen-specific and develops immunological memory. In contrast, the innate response is a rapid and broad range response to challenge that is non-specific and does not develop memory. The innate system however, is critical for initialising adaptive defence. Before production of the antibody response, interleukins are detected that

stimulate innate cells including DCs and NKs (Barnett et al., 2002; 56). Replication of FMDV at infection sites causes local cell damage by apoptosis which produces signals, including inflammatory signals, recognised by host immune defences. These "danger signals" produce a local inflammatory reaction and recruit innate immune cells and lymphocytes both locally and from the blood. Endothelial cell adhesion is modified to increase local permeabilisation and chemokines, released from endothelium and innate cells, providing a chemical gradient for the migration of cells into the infected tissues. Local tissue macrophages attempt to control infection by ingesting infectious or harmful material and DCs produce interferon (IFN) for inducing an anti-viral state of local cells. Whilst macrophages (Mφ) and DC-dependent innate defence against FMDV can prevent infection of cells and disease progression, there is a limit to the viral load they can handle above which the adaptive immune response is required to ensure ultimate protection of the host. Cells of the innate immune system, DCs and macrophages, present antigen to lymphocytes, B cells and T cells, for stimulation of the adaptive response. However, limited research has been done to investigate the innate immune response to FMDV and significant knowledge gaps exist in our understanding of the early protective response.

1.3.4 Phagocytosis of FMDV

FMDV is not monocytotropic; the virus does not infect monocytes, macrophages or dendritic cells. Phagocytic cells are recruited to the site of infection by the inflammatory response, increasing the capacity for endocytosing the virus. Internalisation of FMDV into phagocytic cells should lead to its destruction, and in the case of DCs, antigen presentation on the cell surface. Macropinocytosis of FMDV by Mφ and DCs displays slow kinetics (Rigden et al., 2002; 57) and uptake does not immediately destroy infectivity. Receptor mediated endocytosis enhances the phagocytic efficiency by recognising opsonised antibody through their Fc receptors (FcR) and promotes the destruction of FMDV infectivity (McCullough et al., 1986; 58, McCullough et al., 1988; 59). Removal of the Fc portion of antibodies impairs phagocytosis but not viral neutralisation. The phagocytic response by macrophages and monocytes therefore is strongly augmented by opsonisation with specific antibodies. In the absence of opsonised antibody, macrophages appear to become infectious carriers, internalising the virus after association (Rigden et al., 2002; 57). No viral replication or

protein synthesis is detectable but the virus is not destroyed by acidification and is subsequently released from cells 24 hours after association, suggesting their role as propagators of virus spread to secondary sites.

1.3.5 The role of dendritic cells in FMDV control

In the early stages of infection, the immune response to FMDV is T-independent and therefore DCs become essential for B cell activation and stimulation of CSR to enable viral clearance. DCs recognise pattern associated molecular patterns (PAMPs) contained in microbial products and phagocytose the antigen which stimulates their maturation to antigen presenting cell; maturation is also promoted by pro-inflammatory cytokines released during tissue damage. This up-regulates their chemokine receptors and causes their migration to local lymph nodes to induce the adaptive response. DCs produce type I interferon and other viral cytokines in response to replicating FMDV at the site of infection (Bautista et al., 2005; 60). DC processing of live FMDV induces antigen specific lymphocyte responses (Harwood et al 2008). FMDV infected dendritic cells in cattle directly stimulate splenic marginal zone B cells to secrete anti-FMDV IgM independently of T cells. DCs also produce B cell activating actor (BAFF) and a proliferating-inducing ligand (APRIL) which drive B cell antibody production and CSR (section 1.7.4, (Bergamin et al., 2007; 61)); isotype switching in calves is reported to relate to DC activities involving these receptors.

The up-regulation of chemokine receptors and migration into lymph nodes by DCs enhance their interaction with naïve T cells to enable T cell activation and expansion into effector T cells. Type I IFN (α/β) induces the maturation of DCs so that they can migrate to local lymph nodes for antigen presentation to circulating B cells, and induces a T cell response (Banchereau and Steinman, 1998; 62, Mellman and Steinman, 2001; 63). DCs are then an essential component of germinal centres for driving the affinity maturation of B cells, discussed in section 1.9. The delay in the T cell response to FMDV can be in part attributed to the abortive replication cycle of the virus in DCs which down-regulates MHC class II (Ostrowski et al., 2005; 64). Slow kinetic uptake of FMDV via macropinocytosis in dendritic cells does not immediately destroy infectivity. FMDV will elute from epithelial cells but the virus does not uncoat in DCs which reflects a slower endocytic process. A transient replication of virus is observed but this replication cycle is abortive and DCs eventually

destroy infectivity (Summerfield et al., 2009; 65); induction of IFN production requires live virus to initiate an abortive replication cycle. The abortive replication in DCs provides additional PAMPs associated with viral RNA replicative intermediates for further promoting the immune response.

1.3.6 Protection of interferon against FMDV

FMDV infection stimulates multiple pathways to induce type I and III IFN (Reid and Charleston, 2014; 66). The interferons are cytokines produced in response to infection from the epithelial cells which stimulates the recruitment of immune cells. Type I IFN receptor engagement with cells induces an antiviral state by stimulating transcription of over 100 genes that mediate the antiviral effect of the cell. Pattern recognition receptors (PRRs) detect PAMPs such as repetitive viral capsid sequences. Among the PRRs are the toll-like receptors (TLRs) which recognise viral RNA and the RIG-I-like receptors (RLRs) which trigger immune defence against RNA viral infection. Through the binding of these PRRs to viral RNA a signalling cascade activates NF-κB for proinflammatory cytokine transcription and IFN production (Schlee, 2013; 67). FMDV viral proteases, however, are capable of inhibiting type I IFN and NF- κB signalling to evade cellular responses (Feng et al 2014) which ultimately evades the host innate response. An attenuated FMDV phenotype lacking L protein cannot block type I IFN production which inhibits their replication (Mason et al., 2003; 68). In cattle, the initiation of type I IFN in plasmacytoid DCs (pDC) in response to FMDV requires immune complex formation. Stimulation of pDC with FMDV-Ig immune complexes employs FcR ligation and strong secretion of IFN-α (Guzylack-Piriou et al., 2006; 69). Type I IFN is also released by CD4+ T cells in response to FMDV and the complete depletion ofCD4+ T cells results in a dramatic reduction of type I IFN but had no effect on the induction of neutralising antibodies, duration of clinical signs or viral clearance (Juleff et al., 2009; 54). Type I IFN is not involved in the early stages of infection as it is not induced until T-independent antibodies form immune complexes with FMDV, suggesting IFN does not play a major role in disease resistance.

The pDC response against FMDV is enhanced by IFN-γ which has antiviral activity against FMDV and promotes NK cell function and macrophage activation. CD4+ T cells are the major producer of IFN-γ production in FMDV challenged cattle and IFN-γ levels correlate

with protection against disease (Oh et al., 2012; 70). The IFN-γ levels also correlate with high levels of neutralising antibody (Fowler et al., 2012; 71) and CD4+ T cells are stimulated by B cell activation. The delayed T cell response questions the role of interferon in the early stages of infection but it may prevent longer durations and re-infection of the FMDV. Type III IFN, IFN-λ, induces similar innate antiviral response as type I IFN but signals through different cellular receptors. Whilst type I IFN seem to have a limited efficacy in cattle, type III IFN significantly reduces the severity of the disease in cattle by limiting FMDV replication and spread (Perez-Martin et al., 2012; 72). The type III IFN primarily target the mucosal epithelial cells to protect against viral infection (Wack et al., 2015; 73) and so provide greater defence in cattle where FMDV replicates in the epithelial tissue.

1.3.7 NK cells destroy FMDV infected cells

NK cells have cytotoxic functions that destroy infected viral cells. Their function is inhibited by MHC class I and infection of a cell with FMDV reduces host protein synthesis and therefore loss of cell surface expression of MHC. Whilst their role in the host response to FMDV has not fully been investigated, the loss of MHC class I on epithelial cells enhances NK cell activity in cattle, assisting with the control and clearance of acute infection (Patch et al., 2014; 74). Upon activation by pro-inflammatory cytokines, derived from Mφ, DCs and lymphocytes, NK cells efficiently lyse FMDV-infected cells (Toka et al., 2009; 75). NK cells from vaccinated cattle also display a non-MHC-restricted cytolytic activity against infected cells (Amadori et al., 1992; 76), suggesting they are unaffected by immunological memory. NK cell activity is also enhanced by antibody; high levels of infected cell lysis occur by Fc receptor mediated recognition, without prior activation of lymphokines (Bradford et al., 2001; 77). NK cells recognise antibody bound to the infected cell surface and destroy the cell through antibody dependent cellular cytotoxicity (Biburger et al., 2014; 78). Overall, the cytotoxic immune defence of NK cells is important in immunity and its activity is further enhanced by the antibody response.

1.3.8 Antibody is responsible for viral clearance and protection against disease

Antibodies play a major role in protection against FMD and the significance of the humoral immunity in controlling FMDV infection is well studied in laboratory models and to some extent in the natural hosts. SNA titres correlate with protection against disease in vaccinated livestock, shown by VNT assays (Doel, 2005; 79). The natural infection of cattle induces a high SNA titre that provides rapid and long-lived immunity to the animal (Cunliffe, 1964; 80) and protection from subsequent challenge has been demonstrated up to 5.5 years after initial exposure (Garland, 1974; 81). In contrast, current inactivated vaccines require re-inoculation every 6 months as duration of immunity wanes rapidly (Doel, 2005; 79). As discussed in section 1.3.2, despite the initial serological response to FMDV being T-independent, long-term antibody protection against the disease is T-cell dependent. Antigen presenting cells are required for inducing effective immunity and propagate the rapid initial T-independent antibody response.

In cattle, the primary response to infection is comparable to mice, with high titres of SNA maintained for long durations of time. Detectable IgM appears in the serum 3-7 dpi (Eschbaumer et al., 2016; 82). This early IgM response forms the major component of initial clearance of the virus as VNT assays show the neutralising activity of the serum 6 dpi (Pega et al., 2013; 83). Isotype switching rapidly occurs with specific IgG antibodies detected from 4 dpi and peaking at 14 dpi (Collen, 1994; 84, Doel, 2005; 79, Juleff et al., 2009; 54). The IgA response is more unusual, IgA is detected in the serum from 7 dpi, declining after 14 dpi, but a significant second response is observed 28 dpi. In carrier animals, the IgA titres remain high, whereas levels decline to undetectable levels in non-carriers (Parida et al., 2006; 85).

In the absence of antibody, internalisation of FMDV in macrophages and DCs is slow and infectivity is not lost. Opsonised antibody uptake increases the efficiency of phagocytosis as well as the activity of cytotoxic NK cells. The induction of long lasting immunity and protection against disease is primarily due to specific antibody. The stimulation of B cells relies on DC maturation and migration to the lymph nodes for activating the cognate B cell. Whilst the innate immune system has historically been overlooked in studying the FMDV response, humoral immunity is responsible for mediating the innate immune system and providing protection against disease.

1.3.9 African buffalo immune response to FMDV

The immune response to FMDV in African buffalo has not yet been thoroughly investigated. Antibody is responsible for protection of disease, as seen in cattle and laboratory models; young animals become susceptible to infection as their maternal antibodies wane; circulating antibody levels of new born calves is similar to their dams and these persist for 2-7 months, providing passive protection from clinical infection (Condy and Hedger, 1974; 46). The excretion of aerosolised virus persists for longer in naïve buffalo infected with SAT virus than acutely infected cattle (Bengis et al., 1986; 42) suggesting the initial clearance of the virus is delayed and the early immune response is less effective. However, infection in adult animals is often asymptomatic and African buffalo become long term carriers of the disease, up to 5 years in individual animals (Condy et al., 1969; 86). African buffalo sero-convert within the first 1 to 2 years of life with sero-prevalence higher to SAT2 than SAT1 and lastly SAT3 (Bronsvoort et al., 2008; 87). The long term protective response in African buffalo is therefore superior to what is observed in cattle. Antibodies against non-structural proteins were shown in ~65% of the African buffalo population sampled and antibodies against all three FMDV SAT serotypes (Di Nardo et al., 2015; 88); more than one SAT FMDV serotype may be maintained in individual buffalo (Maree et al., 2016; 43). The long-term carrier status of African buffalo means the virus persists in the oesophago-pharyngeal regions however high levels of antibody are also detected in pharyngeal secretions (Francis et al., 1983; 89). Persistently infected buffalo are refractory to re-infection with the same virus strain, suggesting the long term protection of their antibody response (Hedger, 1972; 90).

1.4 Immunoglobulin structure and function

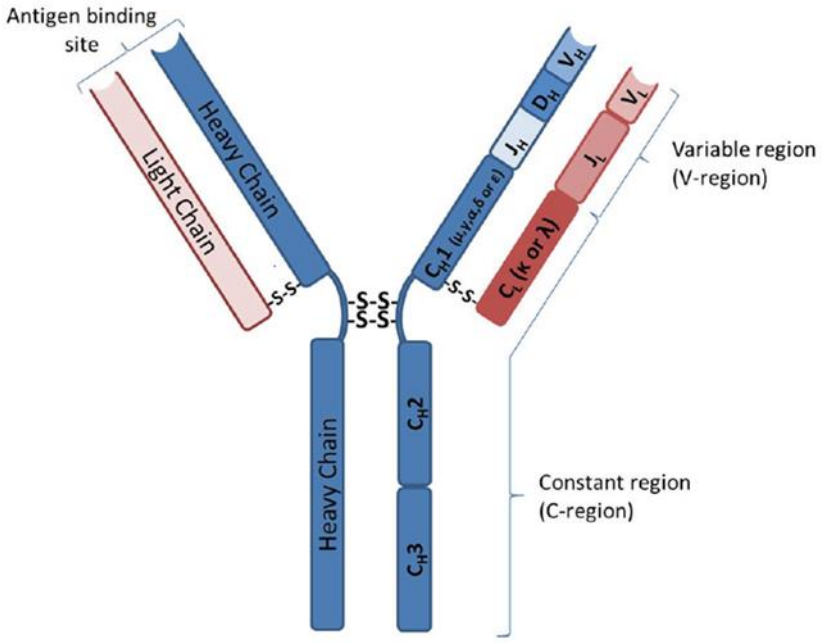


Figure 1.4: Schematic 2-dimensional representation of an immunoglobulin. Taken from (Grant, 2013; 91)

The Ig are secreted proteins composed of two chains, heavy and light, which are bound together by disulphide bonds into the characteristic 2-dimensional "Y" configuration (Figure 1.4). When membrane bound, the Ig molecules form the B cell receptor (BCR) that binds antigen to stimulate the B cells to divide. In reality, the 3D structure of an antibody folds to bring the variable regions together and expose the antigen binding site. Each Ig chain contains a variable region and a constant region; the Ig variable domain of each polypeptide chain is comprised of approximately 100 amino acids arranged in anti-parallel β-strands which form two β-sheets. The β-strands comprise the highly conserved framework regions (FR1, FR2 and FR3) whilst the complementarity determining regions (CDR1, CDR2 and CDR3) are found on the intermediary loops. FR1 is formed from two β-strands (A and B) from the first β-sheet, followed by the external loop comprising CDR1 (BC). FR2 is formed from two β-strands (C and C') from the second β-sheet, followed by the external loop comprising CDR2 (C'C"). FR3 overlaps both β-sheets by including two β-strands (C" and F) of the second β-sheet and two beta-strands (D and E) of the first β-sheet. CDR3 then encompasses the exposed loop (FG) and the J gene comprises the final beta-strand (G) of the second beta-sheet. The two β-sheets are folded in order to disulphide bond between

conserved cysteine residues on each strand (Cys23 and Cys104 of strands B and F respectively) so that they are held facing each other. The constant region on the light chain contains a single Ig domain, whereas the heavy chain constant region contains three or four Ig domains (CH1, CH2, CH3 and CH4), the precise number dependent on the Ig isotype.

Cattle antibodies have unusual antibody structures in ~10% of their circulating Ig (Wang et al., 2013; 92). The CDR3, the surface exposed and most diverse region of the antibody molecule, is ultra-long in these antibodies ranging from 50 to 61 amino acids in length; compared to the typical 8-16 amino acid length observed in other species such as humans (Collis et al., 2003; 93). These ultra-long CDR3 contain numerous cysteine amino acids or contain a codon bias for somatic hyper mutations to form multiple cysteines, allowing a variety of disulphide bonds to form, creating different architectural antibody structures (Wang 2013). Two anti-parallel β-strands form a "stalk" that supports a structurally diverse "knob" domain which forms the antigen binding site. To date, the purpose of these ultra-long antibodies is unknown as an epitope is yet to be discovered. They do provide a novel diversification mechanism for Ig repertoire generation in cattle that had not been found in any other species to date, although the African buffalo had not been investigated.

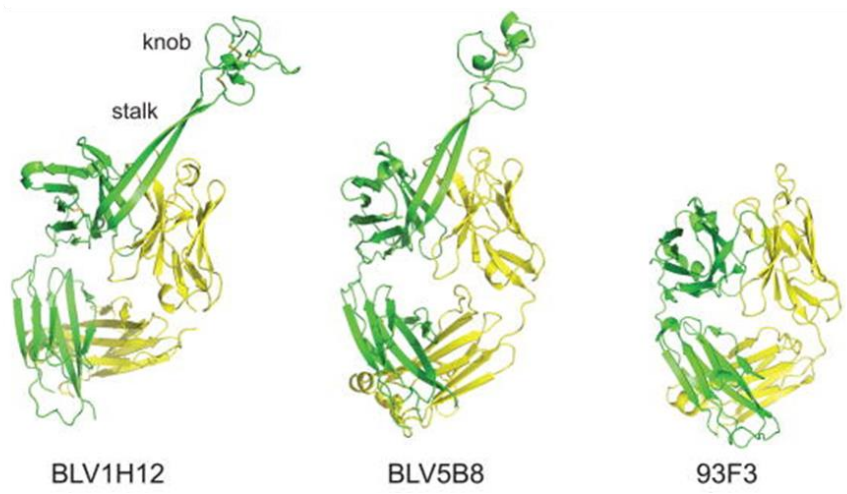


Figure 1.5: Crystal structures of two cattle antibodies with ultra-long CDR3H (BLV1H12 and BLV5B8) compared to a "normal" CDR3H length (93F3). Image taken from (Wang et al., 2013; 92).

The constant region confers the biological functions of the antibody molecules which includes Fc receptor binding on immune cells and complement fixation. The different Ig isotypes each provide specialised effector functions according to their structure. IgM, formed from the constant gene *IGHM*, is rearranged first and so circulates in the absence of antigen as one of the most abundant isotypes in the sera that does not require class switch recombination. It is predominantly found in the blood and a much lesser extent in the lymph and tissues. It is the primary antibody produced in the early phase of acute antigenic challenge, and so undergoes little post-translational modification. The IgM are therefore very broadly reactive but with low specificity. Their overall avidity is improved by their ability to form pentameric structures that has ten possible antigen binding sites and can agglutinate antigenic particles. The monomers are bound together by a polypeptide J chain which disulphide bonds to the Fc regions of the IgM molecules. Incorporation of J-chains in pentameric IgM, and also the polymeric IgA antibodies, reduces their interaction with complement so as not to induce inflammation and can bind to the transmembrane secretory component, the polymeric Ig receptor. This allows IgA, the predominant Ig in mucosal secretions, and IgM to cross to mucosal surfaces and provide the first line of defence against pathogens entering via the mucosa, including FMDV in ruminants (Johansen et al., 2000; 94).

IgG, from the *IGHG1*, *IGHG2* and *IGHG3* genes, is the most predominant Ig isotype found in serum and forms the secondary response to infection, generated from class switch recombination of IgM. Each IgG subclass has varying affinities for the complement component C1q and all of the IgG subclasses are capable of opsonising and neutralising toxins and viruses. During an acute antigenic challenge, the IgG subclasses are each induced to different extents, causing a skew in the abundance of each. The different IgG subclasses have unique affinities for each of the Fc gamma receptors (Snapper et al., 1992; 95, Snapper and Paul, 1987; 96) which induces opsonisation for phagocytosis. The FcR on NK cells also recognise IgG immune complexes and release cytotoxins such as IFN-γ for immune cell signalling as well as cytotoxic mediators that trigger apoptosis of infected cells (Trinchieri and Valiante, 1993; 97). The IgG subclasses then are molecular mediators of the innate immune response.

1.5 Antibody germline repertoire

Antibodies are encoded from three distinct antibody loci in the genome: the heavy chain locus (IGH), the lambda light chain locus (IGL) and the kappa light chain locus (IGK). The three loci contain variable (V), joining (J) and on the heavy chain only, diversity (D) gene segments which rearrange in a process called V(D)J recombination to generate the variable regions of the antibody chains. The number of each gene segment varies between species and therefore the potential recombinatorial diversity of the antibody repertoire varies between animals.

In most species studies to date, the segmental organisation of the IGH locus is highly conserved with the order 5'-*IGHV-IGHD-IGHJ-IGHC*-3'. In the human reference genome, visualised in Ensemble, a total of ~100 *IGHV* exist of which ~50 are functional, which would vary between individuals due to the presence of haplotypes and allelic variation. Downstream of the *IGHV*, humans have 20-30 *IGHD* and six *IGHJ*, providing them with a recombinatorial potential of 6346 VDJ combinations. The organisation of the IGL is slightly altered with the *IGLJ-IGLC* existing in cassettes; humans have 52 *IGLV* upstream of seven *IGLJ-IGLC* cassettes. In the mouse genome, the IGH contains 164 *IGHV*, of which ~100 are functional, upstream of ~20 *IGHD* and then 4 *IGHJ*, providing a recombinatorial potential of 8,000. On the mouse IGL, five *IGLJ-IGLC* cassettes exist downstream of only eight *IGLV*. The diversity of the IGL then is less than the IGH in each species. The conserved organisation between species allows the necessary enzymes to re-arrange and post-translationally modify the locus to provide a functional transcript for expression, whilst the large variation of gene segments provides diversity to the antibody repertoire.

Cattle however, appear to deviate from the 5'-*IGHV-IGHD-IGHJ-IGHC*-3' consensus as previous genome annotations revealed the presence of additional *IGHD* and *IGHJ* clusters and duplications of the *IGHM*. It was suspected that a second partial IGH locus existed on a separate chromosome but this was implausible if both loci were functional. A complete characterisation of the cattle IGH has since been resolved, revealing internal duplications within the IGH structure which will be explored in Chapter 3.

1.6 Antibody gene segment structure

The *IGHV* gene segments are preceded by a promoter octamer, roughly 100 bases upstream of the start of the gene segment, with the canonical sequence of ATTTGCAT. The promoter element, the TATA box, is roughly 15 bases upstream of the gene start which is recognised by the ATG start codon in the 45-50 base pair coding region which forms part of the leader sequence (IMGT: L-PART1). The coding region of this first exon is cleaved and joined to the first 7-11 bp of the second exon (IMGT: L-PART2) to form the complete leader sequence for transport of the nascent antibody chain to the cell surface. The two exons are separated by a 150-330 base pair intron which is marked by the two splice sites AG/GT which follow from a polypyrimidine tract. The remaining ~300 bp of the second exon (IMGT: V-EXON) encodes the framework regions of the antibody molecules (FR1, FR2 and FR3) and forms the complementarity determining regions CDR1, CDR2 and the beginning of CDR3.

The FG loop, the remaining portion of the CDR3 is made from the D (on the heavy chain) and J gene segments. The cattle D gene segments are approximately 42 bp, with the exception of an ultralong D gene segment 148 bp in length. They contain multiple GGT and TAT repeats encoding Gly and Tyr for increased flexibility in the antibody chain as well as multiple cysteine residues which is suspected to provide a unique diversification mechanism to cattle antibodies through the formation of various 3-dimensional structures by different disulphide bond formations (Wang et al., 2013; 92). The J gene segments are approximately 51 bp each and on the heavy chain and kappa light chain are upstream of the constant region, whilst on the lambda light chain each J gene segment is upstream of a single constant gene forming J-C cassettes. Between each J gene segments and the downstream constant region exons are the 5'(GT) and 3'(AG) splice sites with a polypyrimidine tract. This intron is later excised during mRNA processing.

V(D)J recombination occurs by recognition of the antibody gene segments by the recombination activating gene complex (RAG1/RAG2) with recombination signal (RS) sequences. The RS sequences are found downstream of the V gene segments, flanking each D, and upstream of each J gene segment. The canonical sequence of the RS is a heptamer (CACAGTG) and nonamer (ACAAAAACC) separated by a 12 or 23 bp spacer. Deviations from the conserved RS sequence results in inefficient rearrangement of the corresponding gene segment. Efficient recombination will only occur between a 12 bp RS and a 23 bp RS,

known as the 12/23 rule which restricts products to contain V(D)J in the correct order. The molecular mechanism of V(D)J recombination is similar on both the heavy and light chain, a single joining event occurs on the light chain between the V and J whilst two are required on the heavy chain between D and J and then the DJ product with V (Figure 1.6). The RS are recognised by the RAG1 and RAG2 complex and brought into close proximity to form a stable synaptic complex. DNA is then cleaved to generate four free ends; two DNA ends are 5' phosphorylated blunt ends and the other two covalently sealed hairpin coding ends. These coding ends are processed, often with the loss and addition of nucleotides, and joined to form a coding joint. The blunt ends are joined precisely to form a signal joint and the intervening sequence is therefore looped out and discarded. This second step of DNA break processing involves the RAG complex and proteins of the non-homologous end joining pathway. The resulting V(D)J is then joined to the constant *IGHM* by mRNA processing to remove the intron between *IGHJ* and *IGHM*.

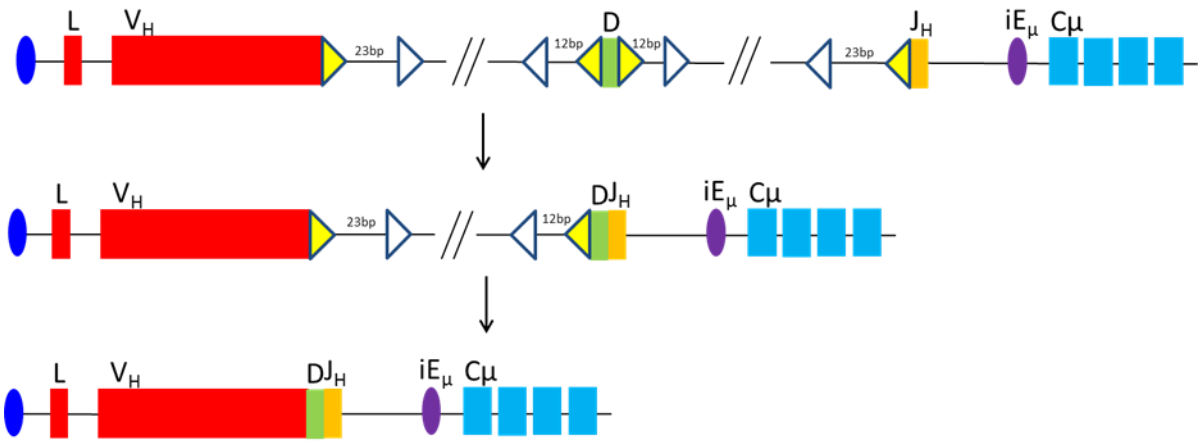


Figure 1.6: Schematic assembly of an IgH chain by recombination of the Variable (HV), diversity (D) and joining (HJ) gene segments. The recombination signal sequences (RS) are represented as triangles; the yellow representing the heptamer and the white representing the nonamer with the spacer lengths indicated. DNA is cleaved at the RS by the RAG complex forming double stranded breaks. Non-homologous ends are joined to form the coding joint. This places the HV promoter in close proximity to the enhancer iEμ. Recombination of the light chain is similar but with a single joining event of LV and LJ.

The enzyme terminal deoxynucleotidyl transferase (TdT) is not essential to V(D)J recombination but where present contributes substantially to antibody diversity by adding non-templated (N) nucleotides to the coding junctions between V, D and J gene segments. Excision of nucleotides by exonucleases at the coding junctions also occurs (Murphy, 2012; 100). TdT has been shown *in vitro* to be capable of catalysing over 1 kb of nucleotide additions and in cattle cDNA libraries, zero to 36 N additions take place between V, D and J genes on the heavy chain and eight N additions between the V and J on the light chain (Liljavirta et al., 2014; 101). This resulting junctional diversity significantly contributes to CDR3 diversity and expands the cattle pre-immune antibody repertoire.

1.7 B cell development

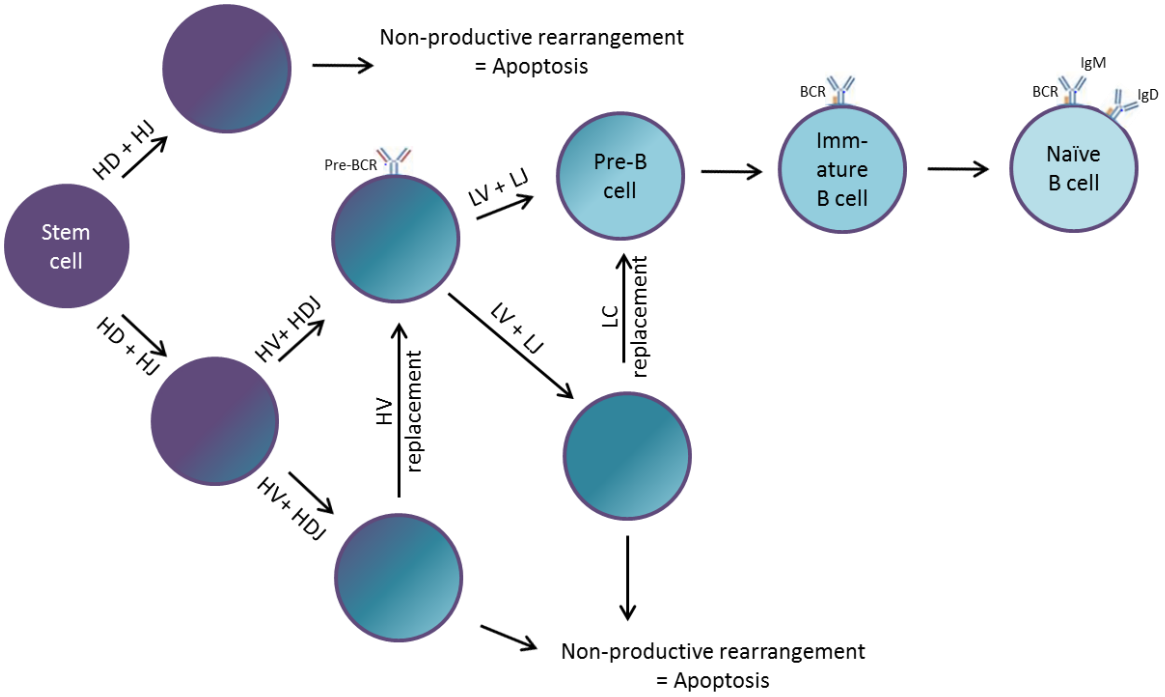


Figure 1.7: Schematic of B cell development from haemopoietic stem cells to naïve B cells. The progenitor B cell undergoes gene recombination to produce an in-frame functional B cell receptor for antigen recognition. The HD HJ gene recombination event occurs first, followed by the addition of HV. If rearrangement is functional the heavy chain is expressed on the cell surface with a surrogate light chain as the pre-BCR. The expression of the pre-BCR stimulates gene recombination of LV and LJ to produce an in-frame functional light chain. This replaces the surrogate light chain to be expressed on the B cell surface as a functional BCR, establishing the cells as immature B cells. If any of the gene recombination events are non-productive cells are either destroyed by apoptosis or may attempt recombination with different gene segments to recover itself. The expression of multiple BCR and the cells release into the circulation from the bone marrow identifies the cells as naive B cells. Adapted from (Rajewsky, 1996; 102)

A functioning B cell receptor (BCR) is a requirement for B cells to progress beyond the pre-B cell stage. The successful recombination of the V(D)J gene segments, facilitated by RAG1 and RAG2 as described in section 1.5, produces an in-frame functional VDJ recombination which is processed and expressed with *IGHM* on the heavy chain. This temporarily associates with a surrogate light chain to form the pre-B cell receptor and becomes expressed on the cell surface. The surrogate light chain consists of two polypeptides, VPREB and IGLL, which are

homologous to the V-region and C-region of light chains respectively. The successful expression of this pre-B cell receptor blocks the RAG1/RAG2 expression which prevents further heavy chain recombination. The B cell divides 4-6 times before the RAG1/RAG2 expression recombines the light chain gene segments until an in-frame functional VJ is transcribed with a LC from the lambda or kappa locus. Kappa is usually rearranged first and if unsuccessful the kappa deleting element eliminates further recombination at the kappa locus and the IGL undergoes rearrangement. The successful light chain replaces the surrogate light chain and is expressed on the cell surface of the immature B cell as a successful B cell receptor.

Continuous B cell development occurs in the bone marrow and in cattle the primary antibody repertoire is further diversified in the Ileal Peyer's patch (IPP). Pre-B cells, before light chain rearrangement occurs, are detected in cattle bone marrow from late gestation to earlier juvenile age but are not detected in adult animals (Ekman et al., 2010; 103). The pre-immune BCR repertoire in cattle is fixed in young animals and the diversification of the primary repertoire then relies on post-recombinatorial mechanisms in response to antigens.

1.8 B cell activation

1.8.1 Cognate activation of B cells by the innate immune system

After leaving the bone marrow, the germline-encoded Igs exhibit low binding affinity to cognate antigen, despite the additional diversification processes in cattle to expand their primary repertoire by somatic hyper-mutation (SHM) in the gut associated lymphoid tissue (GALT) (section 1.8.2). These low binding affinity Ig are unable to effectively neutralise pathogens or provide long term memory for recognition of re-infection. Antigen recognition with co-stimulatory aid from T helper lymphocytes and dendritic cells, stimulates B cell proliferation and differentiation (Cerutti et al., 2013; 104). B cells are activated by one of two mechanisms: by direct engagement of the BCR with antigen or as an immune complex on the surface of an antigen presenting cells, such as dendritic cells (DCs), follicular dendritic cells (FDCs) and macrophages.

Subcapsular sinus (SCS) macrophages capture antigens from lymph, via the complement receptor macrophage receptor-1 (MAC-1) and show them to passing B cells as they migrate into B cell follicles (Phan et al., 2007; 105, Batista and Harwood, 2009; 106). The antigen is not phagocytosed by the SCS macrophages but retained on the surface and presented to naïve B cells for up to 72 hours post infection (Unanue et al., 1969; 107). Naïve B cells, even non-specific cells, transfer the antigen from the SCS macrophages into the B cell follicle of lymph nodes. These non-specific B cells bind to the complement receptors CR1 and CR2 and transfer the antigen to follicular dendritic cells. FDCs internalise the antigen and present the antigen back on their surface to circulating B cells for cognate recognition and activation (Heesters et al., 2013; 108). Specific B cells that recognise the antigen presented by the SCS macrophages engage their BCR and are themselves activated.

Dendritic cells will capture antigen and mature, leaving the infected tissue and migrating to the lymph node where they present whole antigen for B cell recognition (Cahalan and Parker, 2008; 109). The specific B cells will uptake the antigen from the DC and co-ordinate co-stimulatory responses with surrounding CD4+ T cells. DCs will also migrate to the T cell region of the lymph node for T cell stimulation, the antigen is processed and presented to the CD4+ T cells on the MHC class II (Qi et al., 2006; 110) which then stimulate specific B cells to proliferate.

Activation of the B cell triggers a rapid conformational change that reorganises the cellular cytoskeleton and internalises the antigen. Captured antigen is degraded and processed to form MHC-II-antigen complexes on the surface of the B cells. BCR antigen affinity is proportional to the cells ability to present antigen to CD4+ T cells. The activated B cell migrates to the T cell boundary to present the MHC-II-antigen complex to CD4+ T cells. Bidirectional signalling leads to an increase in antigen presentation in the B cell and activation of the T cells which secrete cytokines to promote inflammation in the surrounding tissues. The formation of this B and T cell partnership migrates to follicular zones where they proliferate, becoming short lived plasma cells with low affinity antibody secretion for the rapid serological antibody response. A germinal centre is also formed for affinity maturation of the BCR and the production of the long term memory response (discussed in section 1.9).

1.8.2 B cell activation with T-independent antigens

The CD4+ T cell response is specialised for inducing B cell proliferation and stimulation of plasma B cells for Ig secretion. Co-stimulatory interactions between activated B and T cells form the germinal centre for affinity maturation and formation of long lived plasma cells and memory B cells. Co-stimulation of T cells by B cells or DCs through antigen presentation on MHC class II, stimulates T cell expansion and cytokines produced by DCs in response to infection will alter T cell functionality. Once stimulated, CD4+ T cells secrete IL-21, a B cell survival factor, and migrate into the germinal centre to initiate the GC reaction.

Production of antibodies against most antigens require T cell help to orchestrate a high-affinity class-switched serological response, however some antigens elicit antibody production without T cell involvement. If the antigen is T-dependent, a small portion of activated B cells will differentiate into short-lived plasma cells within T cell regions of secondary lymphoid organs and secrete low affinity antibodies. The recruitment of remaining activated B cells to form germinal centres, enhancing their affinity for cognate antigen through SHM, relies on CD4+ T cells. In the absence of T cell help, T-independent antigens initiate a serological response by either possessing highly repetitive structures that activate B cells by BCR cross linking, termed type 1 T-independent antigens (TI-1) or have an activity which directly activates B cells, termed type 2 T-independent antigens (TI-2). A second signal is required by an activated B cell to stimulate antibody production either via TLR stimulation or complement activation and CD21 stimulation (Vos et al., 2000; 111). Complement proteins C3 and C4bp co-localise on the surface of DCs immune complexed with antigen and provide signals to B cells akin to CD4+ T cell signals that activate B cells. Pathogens usually contain both T-dependent and T-independent antigens as viral capsids possess repetitive structures which generate a T-independent response whilst non-structural proteins are recognised by the T-dependent response. Both T-dependent and T-independent antigens therefore initiate a serological response however, for the production of long term B cell memory and higher affinity antibodies T cell help is required.

1.9 Post-translational modification of the primary antibody repertoire

1.9.1 Post-translational modifications in response to antigen

It is essential for the adaptive immune response to produce large numbers of antibody specificities from a modest number of gene segments. The primary antibody repertoire is formed from somatic V(D)J recombination with the imprecise joining of gene segments caused by TdT junctional diversity helping to expand the germline combinatorial diversity. The successful rearrangement and joining of gene segments leads to the expression of the antibody on the membrane where they become exposed to foreign antigens. Initial antibody-antigen interactions are usually low affinity, broad spectrum antibodies. High affinity immune complex formation occurs via post-translational modifications.

Activation induced cytosine deaminase (AID) plays a central role in adaptive immunity by initiating the post-translational modification processes such as somatic hypermutation (SHM), gene conversion (GC) and, as we discuss later in section 1.10, class switch recombination (CSR). These processes take place in a highly specialised microenvironment known as the germinal centre (Jacob et al., 1991; 112). AID is specifically expressed in activated B cells where it preferentially targets cytosine in W-R-C-Y motifs and deaminates them to uracil within the Ig variable regions (Di Noia and Neuberger, 2002; 113). This provides both target sequence specificity and cellular specificity. Mechanistically, SHM and GC are linked through the requirement of AID; the G:U mismatches created through AID deamination are processed in different ways to facilitate the different modification process.

In SHM the accumulation of point mutations in Ig variable regions improve their capacity for antigen binding. The direct replication across the G:U mismatch results in a transition to A:T base pairs. Alternatively, base excision repair machinery recognise the uracil and uracil N-glycosylase removes the base to generate an abasic site and are subsequently repaired during replication by error-prone DNA polymerases which introduce both transition and transversion mutations by replication over the abasic site (Petersen-Mahrt et al., 2002; 114). AID-induced G:U mismatches can also be processed by the mismatch repair (MMR) machinery to generate nicks and gaps in the DNA and associated error-prone DNA polymerases induce mutations at both G:C and A:T base pairs (Luo et al., 2004; 115). SHM accelerates mutations to a rate of

10^{-5} – 10^{-3} per base pair per generation, which far exceeds the basal rate of mutations in other cells of ~10^{-9} (Wagner and Neuberger, 1996; 116).

Gene conversion, or non-crossover homologous recombination, is the unidirectional copying of genetic code from a 'donor' sequence to replace sections of the 'acceptor' sequence being transcribed. The process occurs in B cells where RAD51 and RAD54 recombinases are present. Considering antibody gene segments appeared to have evolved through segmental duplication, gene segments are highly similar and thus capable of templating off each other. AID deaminates cytosine to uracil which is excised by uracil DNA glycosylase, creating an abasic site. This initiates a double stranded break in the DNA that creates single strand DNA with free 3' ends which invade an intact homologous DNA duplex and primes the DNA replication using the unbroken DNA as a template. This newly synthesised DNA rehybridizes with the original unbroken DNA in a process called synthesis dependent strand annealing (Colaiacovo et al., 1999; 117). The antibody loci contain numerous template gene segments for diversifying the antibody repertoire and this can significantly increase diversity to produce functional Ig repertoires in species such as chicken (Kurosawa and Ohta, 2011; 118).

In cattle, SHM is the predominant mechanism for Ig diversification (Berens et al., 1997; 119, Kaushik et al., 2009; 120, Saini et al., 1997; 121). SHM diversifies the coding and non-coding regions of the locus with comparable frequency (Verma and Aitken, 2012; 122), whereas if gene conversion predominated mutations would occur only in the coding region, leaving intronic sequences largely unchanged. Single base mutations predominate in cattle IgH chains, with transition mutations favoured over transversion. Some evidence exists for gene conversion of the light chain in cattle (Lucier et al., 1998; 123, Parng et al., 1996; 124), where they speculated being able to trace rearrangement with *IGHV* pseudogene templates.

1.9.2 Post-translational modifications prior to antigen exposure

Ruminants appear to have a limited range of antibody gene segments and so it is suspected that species such as cattle expand their primary repertoire by post-recombinatorial modifications in the absence of antigen (Liljavirta et al., 2013; 125). After V(D)J recombination in the bone marrow, B cells in foetal and young animals migrate to the ileal Peyer's patch (IPP) and jejunal Peyer's patch (JPP) in the gut-associated lymphoid tissue (GALT) (Yasuda et al., 2006; 126). The IPP and JPP extend several metres along the small intestine and are lymphoid organs for B cell development. B cells proliferate rapidly in the IPP where AID is strongly expressed and cells undergo SHM to alter their affinity for the antigen (Liljavirta et al., 2013; 125). Significant SHM of antibody cDNA occurs in sequences derived from the IPP tissues. SHM is principally located in the CDRs and enriched in the WRCY AID hotspot motifs to generate greater antibody diversity. In the JPP, high levels of apoptosis occur as antibody is tested against self and the JPP acts as a secondary lymphoid organ for mediating mucosal immune reactions; here up to 5% of cells survive to form the primary antibody repertoire (Onishi et al., 2007; 127). The two-phase B cell generation process is thus suggested in ruminants where antibodies first differentiate in the bone marrow and then migrate to the IPP for diversification of their limited germline repertoire by SHM (Liljavirta et al., 2013; 125). This is hypothesised to expand the primary antibody repertoire in ruminants prior to antigen exposure.

1.10 The Germinal Centre formations in response to antigen

The activation of antigen-specific B cells by T cells that recognise the same antigen triggers the formation of the germinal centre (GC). The GC transient structures form within the peripheral lymphoid organs to develop high-affinity antibody secreting plasma cells and memory B cells for prolonged protection against each specific pathogen. The germinal centre events are facilitated by the unique architecture of secondary lymphoid organs which position the large clonally diverse B cell follicles alongside diverse T cell zones. This supports the cellular interactions required for affinity maturation of B cells to improve BCR affinity for cognate antigen by approximately 1 to 2 orders of magnitude (Griffiths et al., 1984; 128).

Initiation of the GC reaction occurs via a coordinated cascade involving several different cell types (Figure 1.8). The rudimentary representation of the GC is of two spatially separated zones, the light and dark zone. In the dark zone, B cells down regulate their BCR expression and rapidly proliferate. The Ig variable region is diversified by SHM, altering each B cell affinity for the cognate antigen which results in the generation of multiple mutant clones with a broad range of affinities for the cognate antigen. These B cells then re-enter their cell cycle to express the BCR and cycle to the light zone to test their affinity for the specific antigen. Follicular dendritic cells present antigen on their cell surface and if the BCR has a suitable affinity, CD4+ T cells provide survival signals. These effective selection processes within the GC ensure that inferior antibody mutants or those with auto reactive specificities are outcompeted by the higher affinity competitors. However, recent studies have shown that GC B cells do not need to migrate between the light and dark zones to be selected (Hauser et al., 2007; 129).

This affinity maturation selection process relies on B cell access to antigen on FDCs; higher affinity B cells uptake more antigen from the FDC surface and therefore present more MHC-II peptide complexes which provides them with more survival signals from the follicular T cells. Antibodies produced by these GC plasma cells also enhance the affinity maturation by binding to the antigen on the FDC, unless displaced by a higher affinity BCR they block access to the antigen and prevent low affinity cells receiving survival signals (Zhang et al., 2013; 130). High affinity plasma cells, receiving survival signals, initially undergo plasma cell differentiation and become long-lived plasma cells or memory B cells. The drive towards plasma cell differentiation switches the BCR expression to the secreted Ig molecular form and stimulates cell to secrete large volumes of antibody per cell. Stimulated plasma cells secrete $0.5 – 1 \times 10^8$ antibody molecules per hour (Helmreich et al., 1961; 131, Hibi and Dosch, 1986; 132) so only a small number of high affinity plasma cells are needed to provide protective antibody levels.

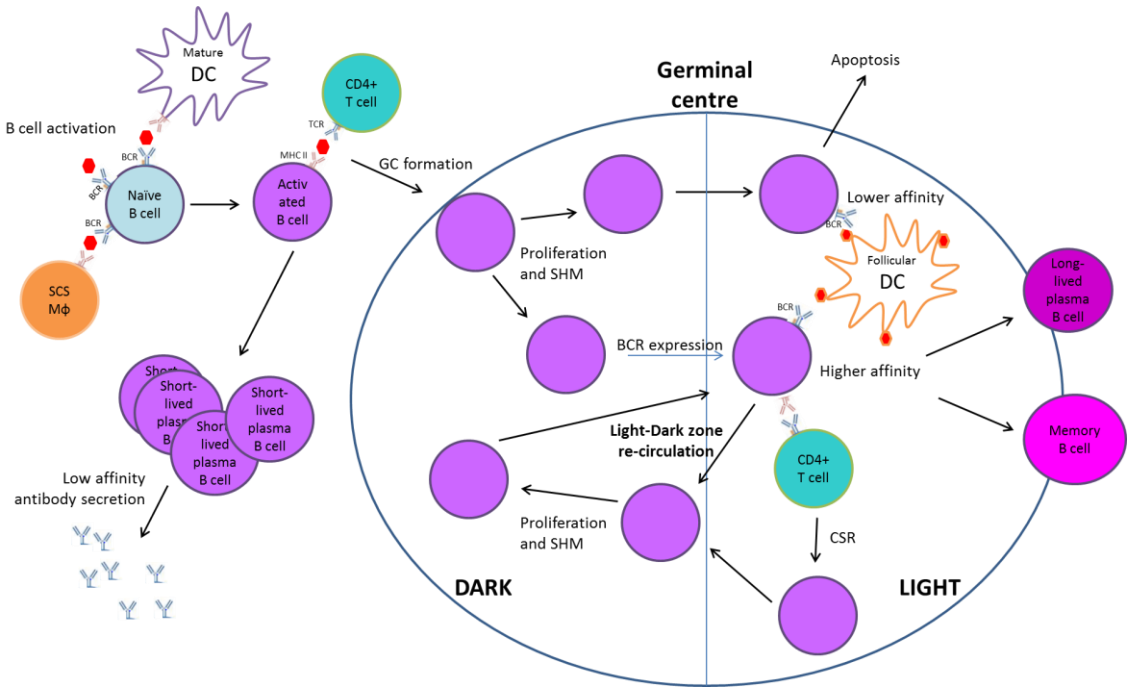


Figure 1.8: Schematic representation of B cell activation and the germinal centre reaction in response to a T-dependent antigen. The naïve B cell is activated by either antigen presentation from dendritic cells, SCS macrophages from the lymph, direct activation with cognate antigen or through CD4+ T cell stimulation. Co-stimulation of the B and T cell results in proliferation into short lived plasma cells that secrete high levels of low affinity antibody and stimulates the formation of the germinal centre. B cells proliferate and undergo SHM in the dark zone, altering their BCR affinity for the antigen. B cells test their affinity on follicular DCs presenting antigen; lower affinity B cells are destroyed by apoptosis whilst higher affinity B cells receive survival signals from CD4+ T cells and either re-circulate in the dark zone to further improve their antigen affinity or differentiate into long lived plasma calls and memory B cells.

1.11 Class switch recombination

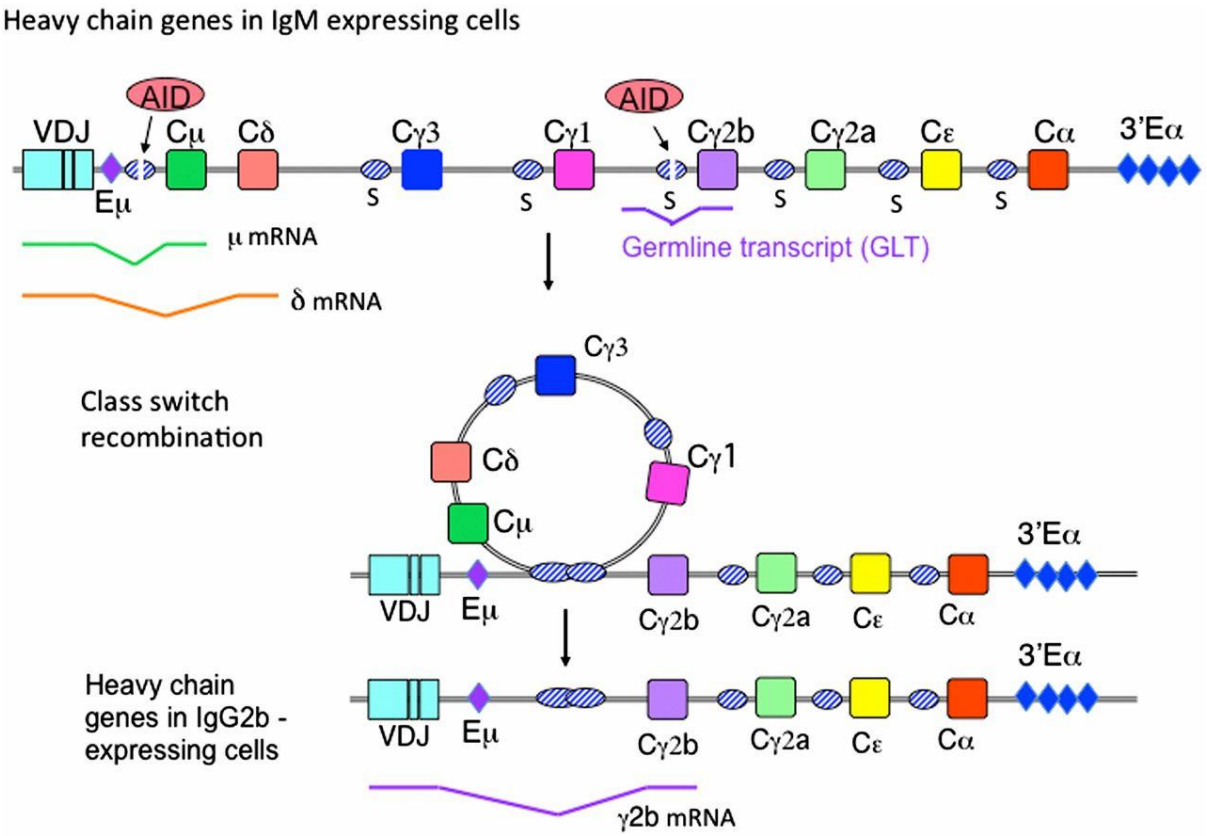


Figure 1.9: Schematic demonstrating class switch recombination of Ig. Recomniantion between the variable (V), diversity (D) and joining (J) gene segemnts occurs prior to class switch recombination to form the VDJ transcript; the activation induced deaminase enzyme (AID) mediated by the intronic enhancer (Eμ) causes a double stranded break and loops out the intervening DNA to form the IgM (μ) and IgD (δ) transcript mRNA from the constant *IGHM* (Cμ) or *IGHD* (Cδ). Upstream of the *IGHG* (γ), *IGHE* (ε) and *IGHA* (α) is a switch region associated with an intervening intron. Class switching proceeds when the enhancer Eμ brings the switch regions together. CSR machinery targets this loop for recombination, removing the Cμ and expressing a different Ig isotype, here displayed as IgG2b from the constant *IGHG2b* (Cγ2b). Image taken from (Stavnezer and Schrader, 2014; 133).

Class switch recombination (CSR) takes place solely on the constant region of the Ig heavy chain (CH) to drive antibody class, or isotype, switching. CSR occurs predominantly in the germinal centres in response to antigen stimulation and requires the input of co-stimulatory signals. The isotype of the antibody produced by each B cell is highly regulated and dependent on the type of antigen encountered and the signalling pathways that are subsequently activated (Manis et al., 1998; 134). Naïve B cells and activated plasma cells in

their early stages of response express IgM and IgD. Upon B cell activation, CSR is induced and the IGHM or IGHD constant region is switched to another Ig isotype. The process of class switching requires two stimuli; a T-dependent or T-independent stimulus that leads to the induction of AID and other transcription factors which implement CSR and cytokines (IL4, IL5, IFNγ and TGFβ) that target which IGHC region is to recombine.

CSR occurs by an intra-chromosomal deletional recombination event, initiated by AID (Figure 1.9). The enzyme converts cytosine to uracil in the switch region which are located upstream of every CH region (except IgD) and consists of tandem G-rich repeat sequences (20-80 bp in length) that differ for each isotype. The overall length of switch regions spans 1-12 kb. Deamination of cytosine in switch regions leads to the subsequent repeat of dU residues which are removed by uracil DNA glycosylase leading to abasic sites and subsequent single strand DNA breaks. The Eμ enhancer associates with the 3'Eα to form a loop and this leads to staggered double strand breaks in the switch regions which are brought into close proximity so that the NHEJ machinery can recombine the downstream *IGHC* of the second switch region. AID activity is transcription dependent and each Ig isotype switch region is tightly controlled by cytokines.

T cell dependent CSR occurs in the germinal centre, induced by the expression of CD40 on the B cell surface and CD40L expressed on CD4+ T cell surfaces. CSR can also occur independently of T cells; allowing rapid induction of class switched antibodies before T cell stimulation. Toll-like receptors on DCs recognise conserved PAMPs, such as viral capsid proteins, and process these for antigen presentation to B cells. The synergistic signalling between TLR and BCR can induce CSR to IgG3 (Xu et al., 2012; 135). DC signalling via the B cell activating factor (BAFF) and a proliferation inducing ligand (APRIL), these receptors can induce CSR to IgA in synergy with B cell signalling. Depending on the presence of other stimulatory factors, along with BAFF and APRIL, other Ig isotypes can be class switched (Pone et al., 2012; 136). The co-ordination of T-independent and T-dependent immune mechanisms provides a stimulated class switched initial response to infection whilst the GC develops, allowing the rapid clearance of the virus and subsequent generation of long-term immunity.

1.12 Sequencing the functional antibody repertoire

The development of high throughput sequencing technologies has allowed the large-scale characterisation of an animal's functional antibody repertoire. The effective characterisation of the antibody repertoire with high-fidelity and high throughput sequencers enables a robust analysis of the antibody repertoire. The use of Illumina sequencing for antibody repertoire analysis was first achieved in zebra fish (Weinstein et al., 2009; 137) but has since been used to analyse other species including mice, chickens, humans and cattle (Larsen and Smith, 2012; 138, Greiff et al., 2014; 139).

Antibody repertoire sequencing is used to identify antigen epitopes which are targeted by effective antibodies that naturally control infection (Laserson et al., 2014; 140). The antigenic discovery can also uncover the specificity of T cells as B and T cells both coordinate responses against the same macromolecular antigen complexes (Robinson et al., 2003; 141). Large scale sequencing of the antibody repertoire and the identification of antigen epitopes is expected to enable modelling of these antibody-antigen interactions which can lead to design of better vaccine candidates; ones that bind and neutralise antigen stronger and that stimulate the T cell response. As well as this, sequencing the antibody repertoire assists our understanding of the immune repertoire development, including quantifying gene usage and post-translational modification mechanisms. To achieve this, a comprehensive germline sequence is required.

1.13 Sequencing the antibody encoding germline

The antibody loci consists of GC rich, highly repetitive sequences with tandem repeats which complicates the assembly of these regions. Short read sequencing technologies have difficulty as the reads are unable to span the repetitive regions of the genome which leads to fragmented genome assemblies. Historically, the available cattle genomes contain large sequence gaps and assembly errors (Zimin et al., 2009; 142, Snelling et al., 2007; 143). The development of long range sequencing technologies are beginning to overcome these difficulties as long reads span across the highly repetitive sequences for the assembly of complex genomic regions such as the antibody loci. However, these technologies currently

have low throughput and high error rates, often introducing indels in sequences so the putative functionality of gene segments is underestimated (Quail et al., 2012; 144). An improved cattle antibody germline is necessary for the complete characterisation of their primary antibody repertoire and the African buffalo antibody loci is currently unpublished.

1.14 Thesis overview

The overall aim of this thesis was to characterise the cattle and Afrcian buffalo antibody loci which was first attempted by generation of a temperature inducible homologous recombination BAC library, which is discussed in Chapter 2. The construction of a cattle BAC library for the isolation of specific clones containing antibody loci was superseded by the availability of a Pacific Bioscience long read cattle genome, ARS-UCDv0.1. The IGH in the available cattle genome assemblies including the ARS-UCDv0.1 and a recently published IGH constructed from Sanger sequencing is compared in Chapter 3. Afrcian buffalo genome reads are mapped to the cattle IGH and *de novo* assembly of the gene segements forms the first description of the African buffalo antibody loci. In Chapter 4, the light chain loci IGL and IGK is characterised in the available cattle genomes and the African buffalo gene segments assembled and described. A very strong preferential usage of the lambda light chain is shown in both species using qPCR but with limited germline gene usage, shown by RNA-seq. In Chapter 5, the antibody repertoire in African buffalo in repsonse to FMDV infection or cattle inoculation with SAT1 is explored. Both species are capable of generating long and ultra-long CDR3H with very high diversity. The light chain assumes a structural role, varying little in its length or sequence in response to FMDV. Chapter 3 and 4 reveal a very similarly limited germline repertoire in cattle and African buffalo yet the buffalo produce a focused response to FMDV, with the specific amplification of CDR3H sequences over time. In cattle however, their CDR3H repertoire remains diverse. FMDV infection in African buffalo is asymptomatic whilst in cattle morbidity is 100%. The antibody response coincides with viral clearance and so this differential response could explain the different disease profiles between the two species.

Chapter 2

Enrichment and Isolation of the Cattle Immunoglobulin Germline Sequences

2 Abstract

The antibody encoding germline sequences was sought by the construction and targeted isolation of the antibody loci in a cattle bacterial artificial chromosome (BAC) library. Current publically available versions of the cattle genome are incomplete over these immune gene loci and contain large assembly errors. Available cattle BAC libraries were screened for clones containing the regions of interest but the unique biology and complexity of these regions has precluded their inclusion in these standard libraries. Therefore, the construction of a BAC library capable of homologous recombination (HR) to enable the specific isolation of clones of interest was attempted. Optimisation of the temperature inducible HR library construction protocol led to the successful isolation and transformation of BAC vectors, pBAC-red and pBeloBAC11, and the isolation of high molecular weight (HMW) DNA. Ligation of HMW DNA into the temperature inducible recombination vector, pBAC-red, and subsequent transformation of the clones into *E. coli* cells proved suboptimal and required further optimisation. At the same time, our collaborators at the United States Department for Agriculture (USDA) constructed a new cattle genome using single molecule real time sequencing with Pacific Biosystems. The largely intact antibody loci were identified in this genome and all contigs containing the regions of interest were extracted for annotation and further analysis. This superseded the need to construct a recombinase screenable BAC library and these genomic sequences were taken further for subsequent analysis.

2.1 Introduction

2.1.1 The purpose of a BAC library

The antibody loci are highly repetitive and polymorphic regions of the genome which are difficult to assemble with short read sequencing technology. Consequently, the antibody loci in the available published cattle genome are incomplete. In order to quantify gene usage and mutations to the germline in the developing immune response to infection, a complete reference genome is required. Some of these difficulties can be overcome by identifying and sequencing specific bacterial artificial chromosome clones of interest, which would contain a single haplotype of the locus and would allow greater sequencing coverage over the desired genomic region.

A bacterial cloning system for complex genome analysis was developed to produce high-resolution physical maps of DNA. The bacterial artificial chromosome (BAC) system was first based on *Escherichia coli* and its single copy plasmid F factor in 1992 (Shizuya et al., 1992; 145). It is capable of maintaining clones of >300 kb to a high degree of structural stability in the host. The system established a method for easy manipulation of DNA, high cloning efficiency and the stable maintenance of inserted DNA. A detailed analysis of complex regions of the genome can be enhanced by construction of large genomic insert libraries. BAC libraries are used often to close gaps in existing genome assemblies or as a method of targeted sequencing to fine-tune or map regions of interest. Highly polymorphic and repetitive regions of the genome are problematic to assemble so BAC libraries offer a significant advantage as each BAC clone is derived from a single haplotype. A BAC remains the vector of choice for the cloning and manipulation of very large DNA fragments.

The higher the coverage in a BAC library, the greater the probability the library contains the genomic region of interest. Higher library coverage is achieved by greater numbers of transformed clones and larger insert sizes, and is dependent on the size of the genome.

$$N = \ln(1-P)/ \ln (1-I/GS)$$

N = no. of clones, P = probability, I = insert size, GS = genome size

The BAC library is then constructed from partially digested high molecular weight (HMW) genomic DNA and ligated into a specialised vector for mass transformation.

2.1.2 Alternatives to BAC libraries

The main alternative to a BAC system is the yeast artificial chromosome (YAC) (Ratzkin and Carbon, 1977; 146). The YAC cloning system is derived from the yeast *Saccharomyces cerevisiae* and the vector contains the exogenous DNA insert, flanked by a yeast centromere and two telomeres. The clones can be propagated in bacterial cells or in yeast cells for the expression of eukaryotic proteins. Whilst the YAC system can contain DNA fragments of up to 3 Mb, the DNA is difficult to isolate intact in large quantities and are often chimeric due to the presence of several copies of the YAC inside each cell. Bacterial systems provide better alternatives for obtaining pure, stable DNA in large quantities.

Along with BAC systems, the fosmid system is a frequently used method for generating large, stable genomic clones for genome sequencing and physical mapping. A fosmid library is constructed from bacterial F plasmids and is propagated in bacterial cells, usually *E. coli*. A fosmid vector is restricted, however, to a maximum insert size of 40 kb. Compared to the 80-300 kb range that BAC vectors are capable of, the usage of fosmid libraries is more restricted. Fosmids are still a powerful complementary tool for BAC library sequencings. The cloning efficiency of these systems is high and they are limited to one fosmid molecule per host. This low copy number offers higher stability relative to other vector systems. They have been used therefore for closing gaps in BAC library genomes such as the human genome (Consortium, 2004; 147).

2.1.3 Construction of a BAC library

Generation of a BAC library requires the cloning and transformation of large pieces of DNA into bacterial cells for arraying. The manipulation of HMW DNA requires its isolation and protection from mechanical shearing and nucleolytic degradation during preparation. Intact cells are embedded in agarose plugs to protect the DNA from breakage. Cells are then lysed within the agarose plugs and incubated in proteinase K, a broad-spectrum serine protease,

which digests any cellular proteins. The digested cellular components are removed by diffusion in several wash steps and the agarose plugs stored in EDTA (0.5 M solution) for inhibition of nuclease activity. Isolation of large intact HMW DNA is subsequently size selected from a gradient of partially digested DNA fragments.

A partial digestion gradient of genomic DNA can be achieved by several methods including limitation of endonuclease concentration, reaction time or buffer magnesium concentration. Alternatively, the ratio of the endonuclease can be varied with a competing methylase enzyme specific for the same DNA recognition sites. On average, EcoRI sites occur in the antibody loci every 4100bp on average (with an estimated total of 366 restriction sites in the cattle antibody loci). Competition for restriction enzyme sites increases the likelihood of the antibody loci being incorporated into larger DNA fragments. These can then be fractionated for appropriate size selection.

2.1.4 Pulsed field gel electrophoresis

Gel electrophoresis separates the DNA based on size and charge through an agarose matrix. In conventional gel electrophoresis, the electrophoresed DNA in agarose assumes a limiting conformation on its movement through the gel. Beyond a limit of 20 kb, mobility of the DNA molecules rapidly decreases, as the unidirectional electric current causes them to become trapped in the matrix. Reducing the agarose concentration to 0.5% and the voltage will allow resolution still only up to 50kb (Rapley, 2000; 148). However, pulsed field gel electrophoresis (PFGE) can separate molecules as large as 12 Mb (Orbach et al., 1988; 149). In this technique, the direct current electric field changes direction and/or intensity relative to the agarose gel. The voltage is periodically switched between the central axis of the gel and 60 degrees either side. The pulse time in which the electric field is applied in any given direction is the same duration in each angle, resulting in a net forward migration of the DNA. Duration of the pulse time is most important for determining the molecular size range over which separation is possible. Short pulse times allow smaller molecules to migrate due to the rapid change in field direction, whereas longer pulse times allow the larger DNA molecules to move and separate on the gel. Size selection of HMW DNA can then be achieved by excision of the correct band according to an appropriate ladder.

2.1.5 BAC vectors

Size selected HMW DNA needs to be ligated into a suitable vector for propagation in bacterial cells. As such, the first BAC vector contained genes for self-replication and copy number regulation inside the cell; *oriS*, *parA*, and *parB*, respectively (Shizuya et al., 1992; 145). Since then BAC vectors have been modified for specific function, such as the addition of universal promoter sites such as T7 and SP6 and restriction enzyme sites. Antibiotic resistance genes allow the negative selection of clones. A common BAC vector used for construction of large insert libraries is the pBeloBAC11 (Kim et al., 1996; 150) that was used previously for construction of a cattle BAC library in our laboratory (Figure 2.1) (Di Palma, 1999; 151). The pBeloBAC11 contains cloning sites for HindIII and BamHI and a chloramphenicol antibiotic resistance gene. It also contains the gene *RepE* which is essential for copy number control. Low copy number plasmids have the disadvantage of having low volumes of obtainable DNA from a cell but tend to be more stable than high copy number plasmids due to less chance of recombination between clones.

The average insert size in a BAC library is 150-350 kb. An increase in average insert size increases coverage of the genome but has a dramatic decrease in transformation efficiency. Since coverage is also reliant on number of clones, a compromise between insert size and colony number is necessary. A major limitation with genome libraries is the number of transformations required to achieve a good coverage.

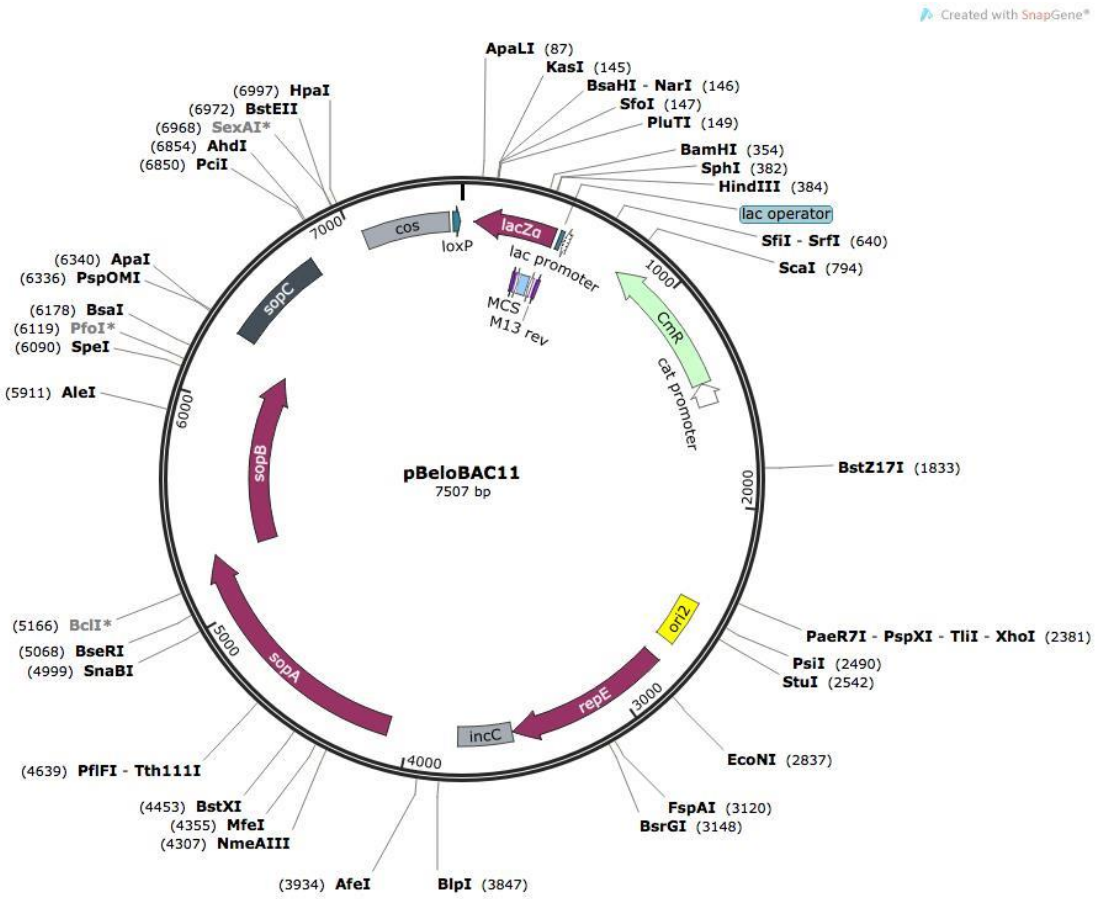


Figure 2.1: Structure of the cloning vector pBeloBAC11 bacterial artificial chromosome (BAC). The vector contains the lac operon for blue/white screening selection of positive clones. Image taken from (Shizuya et al., 1992; 145).

2.1.6 Electroporation of large inserts

The transformation efficiency of a BAC clone is highly dependent on the strain of *E. coli* but is negatively correlated with increasing size of DNA. Larger inserts increase the coverage of the genome in the library at the cost of reduced cloning efficiency. Altering the voltage gradients and time constants is required to achieve optimum transformation conditions. Electro-competent cells designed exclusively for large inserts have been designed and are commercially available. Home grown cells can never achieve transformation efficiency of commercial cells for BACs over 130kb (Valenzuela et al., 2003; 152). The most widely used

E. coli strain for BAC cloning is the DH10B as they contain mutations blocking recombination and the restriction of methylated DNA or foreign DNA by endogenous restriction endonucleases (Sambrook et al., 1989; 153). Transformed cells allow the stable propagation of the BAC vector and have been shown to have no visible difference in restriction patterns after 100 generations of growth.

2.1.7 Traditional BAC systems and their disadvantages

Traditional methods of BAC library construction involved the array of single BAC clones into individual wells of a micro titre plate for archiving single unique clones. Isolation of a clone of interest would then require screening of the entire library of clones. The labour-intensive costs of serial ligation transformation protocols and the array of clones into wells for screening however, are economically unfeasible for applications that focus on a singular region of the genome, such as the antibody loci.

Non-arrayed libraries can be screened by radioactive probe colony hybridisation after plating on petri dishes but this process is also laborious and requires multiple cycles of colony streaking and screening of a replicate to generate pure colonies. Alternatively, the library can be amplified and aliquots of the library screened through recombination selection. This new approach of "recombineering" (recombination-mediated genetic engineering) targets clones through homologous recombination for selective growth of clones containing the genomic region of interest. Positive clones containing only the region of interest could then be amplified and sequenced directly.

2.1.8 BAC screening using recombineering

Homologous recombination (HR) is used naturally to repair double stranded DNA breaks and involves the exchange of DNA strands of similar or identical nucleotide sequence. In recombineering, homologous recombination is utilised for insertion of a linear DNA fragment into the BAC clone that contains an antibiotic resistance gene with flanking primers for selection of clones containing the required region. Use of homologous recombination allows the insertion, deletion or alteration of any sequence precisely. The method uses a lambda

prophage with *exo*, *bet* and *gam* genes. The lambda exonucelase Exo cleaves the DNA to form single stranded overhangs. The bet gene codes for the single stranded DNA binding protein for stabilising the cleaved ends and the Gam protects the linear DNA from nuclease attack. The overhangs can then be re-annealed to circularise the plasmid and maintain its stability. The desired recombination products are identified via negative selection for resistance to the inserted antibiotic.

HR allows modification of large DNA molecules at precise sequence locations, whereas conventional restriction endonuclease based strategies would cleave large DNA pieces into numerous fragments. Several systems already utilise the bacteriophage lambda HR genes *exo*, *bet* and *gam*. A new *E. coli* strain was first described (Lee et al., 2001; 154) with an inducible recombination derived from its defective lambda prophage. This strain DY380 expressed *exo*, *bet* and *gam* under the control of a temperature inducible repressor. The introduction of this temperature inducible repressor system into a BAC vector meant that homologous recombination of clones could be induced by growing the clones at a higher temperature.

Several BAC libraries have since been created in the temperature inducible HR BAC vector pBAC-red (figure 2.2). The vector contains the lambda recombineering genes under the control of the temperature-inducible repressor CI-857. These BAC libraries were shown to be stably propagated at low temperatures without leaking expression of recombinant proteins (Nefedov et al., 2011; 155). Isolation of specific clones was achieved at high efficiency with homologous recombination screening (figure 2.3).

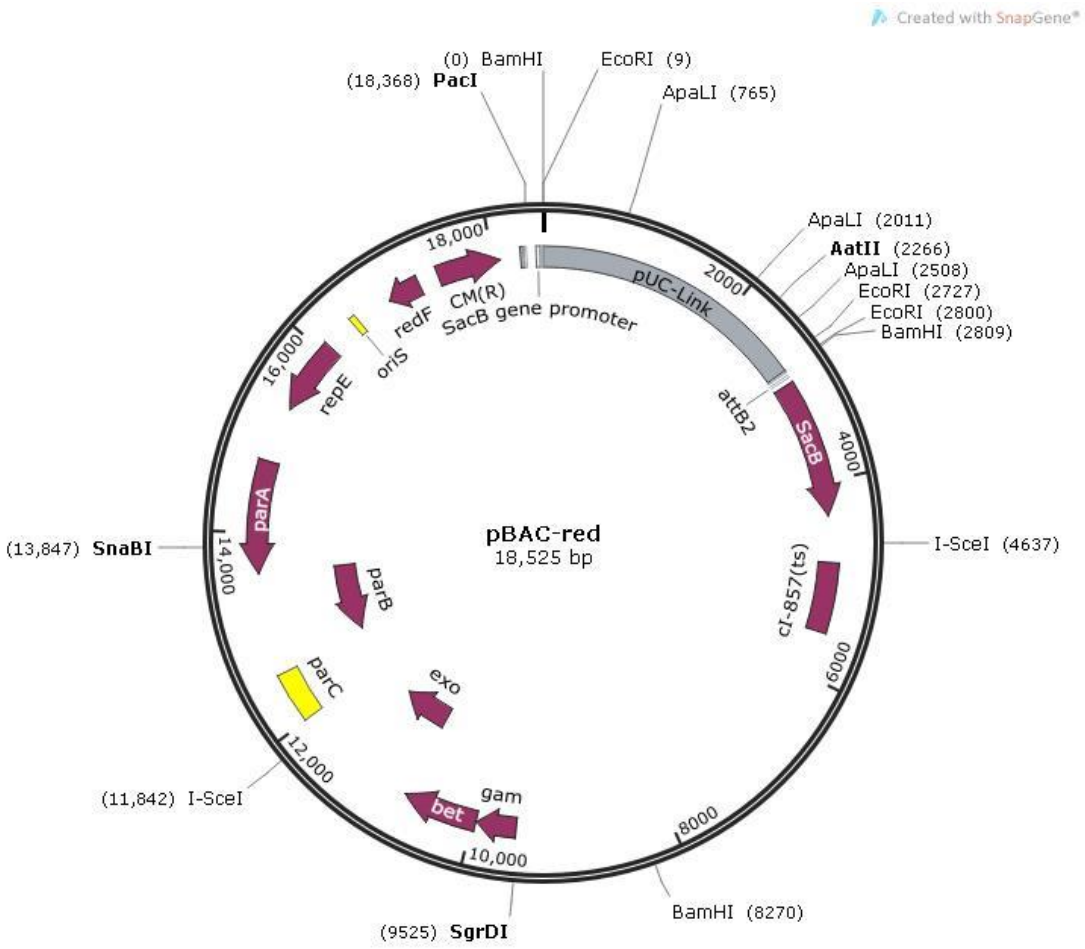


Figure 2.2: Structure of the cloning vector pBAC-red. The vector contains the homologous recombination prophage *exo*, *bet* and *gam* under the control of the temperature-inducible repressor cI-857. An antibiotic resistance gene can then be inserted into clones with flanking primers for positive selection of clones containing the region of insert. Taken from (CHORI, 2016; 156)

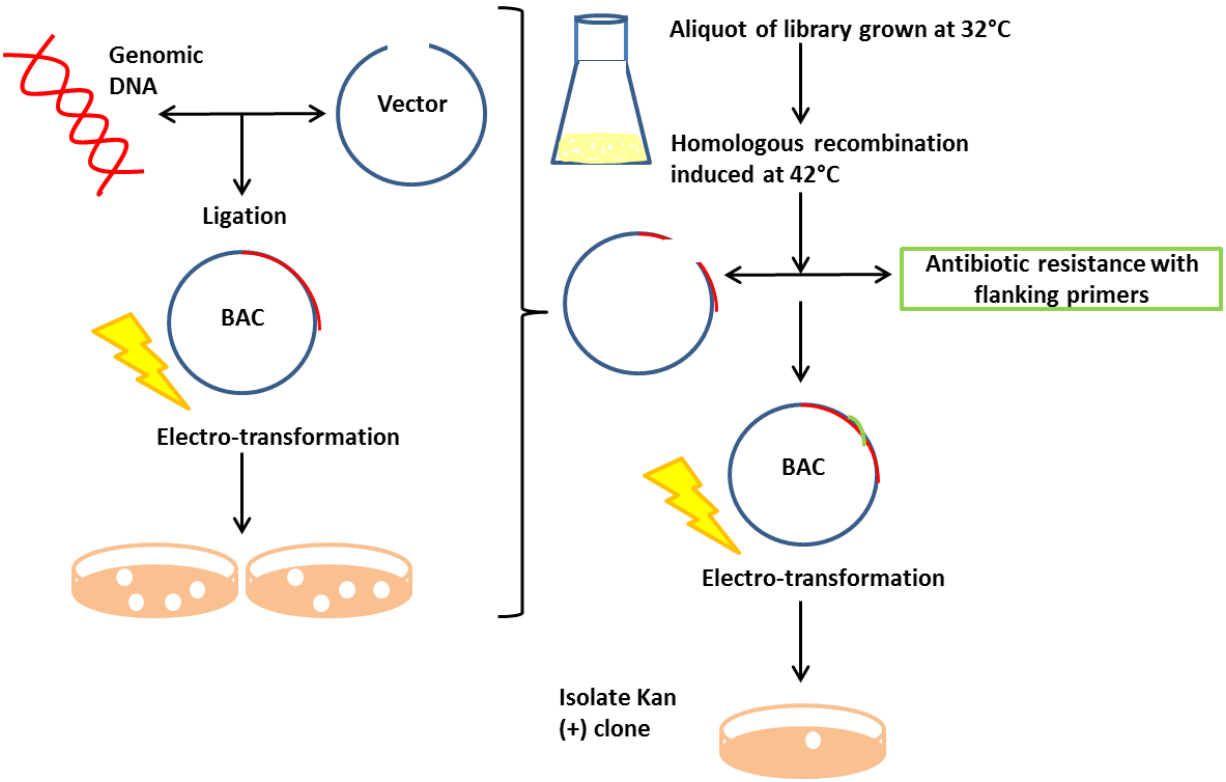


Figure 2.3: Strategy for library screening through temperature inducible homologous recombination (HR) with the BAC vector pBAC-red. HMW DNA is isolated and size selected for ligation into a BAC vector. This is transformed into electro-competent *E. coli* cells and an aliquot is plated to check for insert sizes in the library clones. Aliquots of the library can be stored at -80ºC and grown when needed. HR is induced at 42ºC and an antibiotic resistance gene with specific flanking primers of the region of interest can be inserted for the positive selection of clones containing the desired genomic region. These are grown on the selected antibiotic plates for isolation and subsequent sequencing.

2.1.9 Existing cattle BAC libraries

Several arrayed BAC libraries for cattle have previously been constructed with the method of archival arraying and storage of clones which includes those listed below. The first cattle BAC library, of an Angus bull, utilised the pBeloBAC11 vector with a 1x genome coverage and average insert of 146kb (Cai et al., 1995; 157). A similar BAC library using pBeloBAC11 was constructed at The Pirbright Institute with a 1.6x coverage from an MHC-homozygous Friesian breed animal (TPI4222 library; (Di Palma, 1999; 151)). Another Friesian BAC library RPC142 has been constructed and mapped to the bovine genome, displaying an equal coverage and providing a resource for estimating genomic complexity (Snelling et al., 2007; 143). The Children's Hospital Oakland research Institute (CHORI) cattle BAC library, the CHORI240, was later developed by Pieter de Jong et al from the Hereford breed sire named Domino, with a 10x genome coverage of the genome (CHORI158). This was later used in part to generate the available cattle reference genome, the Btau_3.1 (Elsik et al., 2009; 159). A search of the CHORI BAC library for clones containing the region of interest failed and it appeared that none of the clones contained antibody loci. A non-arrayed cattle BAC library amenable to recombineering does not yet exist and so we attempted construction of a temperature inducible HR library in the BAC vector pBAC-red for the positive selection of clones containing the antibody loci.

2.2 Methods

2.2.1 TPI4222 cattle BAC library constructed in pBeloBAC11

A BAC library (TPI4222) consisting of 35,904 clones was previously constructed using the vector pBeloBAC11 by Frederica di Palma (Di Palma, 1999; 151) in electrocompetent NEB 10-beta *E. coli* cells (New England Biosciences, Inc), using Friesian peripheral blood mononuclear cells (PBMC) homozygous for MHC class I haplotype A14. The constructed library was estimated to have a 1.6x coverage of the genome. This was organised into a PCR screen-able library with individual clones in single wells of 96 well plates. Individual plates were further pooled into single wells of 96 well plates for ease of screening.

2.2.2 Primers for screening the TPI4222 BAC library

Screening of the BAC library utilised primers designed against the most complete and recent published bovine genome (UMD3.1 (Zimin et al., 2009; 142)) for amplification of non-specific individual gene segments from the antibody loci; the heavy chain variable (*IGHV),* diversity *(IGHD)* and joining (*IGHJ)* genes or the light chain variable (*IGLV)* genes (see Appendix Table 1). Primers were designed to bind to the consensus sequences flanking each gene segment. The primers, UMD_HV1-3 and UMD_LV1-3 were designed to bind to the leader and FR3 sequence flanking each distinct phylogenetic gene family. The *IGHV* and *IGLV* gene segments were identified in the reference genome UMD3.1, aligned and sorted into their phylogenetic sub-groups based on their sequence. UMD_D1-8 primers flanked each gene segment in the intronic regions. The UMD_HJ1 forward primer was designed in *IGHJ1* gene segment and the reverse was designed in *IGHJ2*, UMD_HJ2 forward primer was designed in the *IGHJ3* and the reverse in *IGHJ4* and the UMD_HJ3 forward primer in the *IGHJ5* and the reverse primer in *IGHJ6*. Each PCR amplicon would therefore contain two gene segments. These primers were optimised on cattle gDNA from animal 598 using a temperature gradient from 54ºC to 64ºC. Clearly defined single bands occurred at 58ºC, with

a high yield. The $MgCl_2$ concentration was also optimised between 1.5 mM and 5 mM per reaction, with the optimum concentration found at 2.0 Mm. Primers were then confirmed for multiplex capabilities by ensuring the separation of distinct bands of the correct size. *IGHD* primer sets were discarded as they produced too many non-specific bands so subsequent BAC screens were carried out with a master mix of 0.8 mM of each primer, UMD_HV1, UMD_HJ2 and UMD_LV2. These produced clearly defined bands on gDNA at 450 bp, 300 bp and 350 bp respectively.

2.2.3 PCR screen for HJ, HV and LV regions in the TPI4222 cattle BAC library

The TPI4222 BAC library was screened using the optimised multiplex primer master mix to identify positive clones for the antibody loci. The pooled 96 well BAC library plates were first screened to identify the plates containing the positive BAC clones in the library. PCR reactions (10 µl) were set up with 1 U GoTaq DNA polymerase (Promega), 1 µl of Taq flexi buffer and 1 µl of the multiplex primer mix (outlined in the previous section 2.2.2). Cycling conditions involved an initial template denaturation and enzyme activation at 94°C for 1 min, then 35 cycles of 94°C denaturation for 30 s, 58°C annealing for 45 s, and 72°C extension for 1 min, before a final extension of 72°C for 5 min. Positive clones identified on the pooled plates corresponded to positive plates in the BAC library. The corresponding positive plates were screened by pooling the columns and rows into single PCR reactions in the same conditions as above to cross reference for the individual positive wells. Clones identified in positive wells were then grown overnight from each well at 37°C in 5 ml of liquid broth, 225 rpm in a shaking incubator. The BAC clones were isolated using the miniprep for the isolation of plasmid DNA (Lezin et al., 2011; 160). Following this method, the clones were centrifuged at 5,400 x g for 10 min and the resulting pellets were re-suspended in 0.3 ml of re-suspension solution (10mM Tris-HCl, 1mM EDTA, pH 8) at 4°C. Alkaline lysis solution of 0.3 ml (0.5M NaOH 1% SDS) was added to each tube on ice and incubated for 5 min. Clones were pelleted again at 6,000 x g for 15 min at 4°C and washed with 0.8 ml of ice cold isopropanol. The pellet was re-spun for 15 min at 6,000 x g, 4°C and then washed with 0.5 ml of 70% ethanol before a final centrifuge at 6000 x g for 5 min. The supernatant was removed and the pellet air dried before being re-suspended in 100 µl of water. A total of 5 µg of each BAC clone was digested with 2U of EcoRI in 50 µl reactions at 37°C overnight. The 5 µl aliquots of the EcoR1 digested clones were then ran on a gel at 70V for 4 h and visualised under UV light for comparison of the positive clone digestion patterns. Positive clones were

then screened by PCR with each of the BAC UMD_HV1, UMD_HJ2 or UMD_LV2 primers individually to ascertain the antibody region in the clone. The 10 µl reactions contained 1 U of GoTaq DNA polymerase (Promega), 1 µl of Taq flexi buffer and 0.8 mM of the forward and reverse primers. Cycling conditions were the same as before. 200 ng/µl of the purified clones were sent for BAC end sequencing with Sanger using the T7F stock primer.

2.2.4 Isolation of peripheral blood mononuclear cells from blood

Peripheral blood from a Friesian animal 705983 (Animal request license number AR000579, MHC haplotype A18) was collected by jugular venepuncture into sodium heparin (10 U/ml) in accordance with the U.K. Animal (Scientific Procedures) Act, 1986, and approved by The Pirbright Institute Ethics Committee. Blood components were separated by density gradient cell separation in sterile 50 ml conical tubes (Falcon) by underlaying 15 ml of whole blood with Histopaque-1077 (Sigma-Aldrich, Gillingham) using a graduated pipette. The volume was made up to 50 ml with phosphate buffered saline (PBS) and centrifuged at 2,500 rpm, at 19°C for 40 min in a Rotina 420R centrifuge. Peripheral blood mononuclear cells (PBMCs) were removed using a Pasteur pipette and transferred to a tube containing 50 ml PBS for repeated wash steps. Remaining erythrocytes were lysed in 50 ml ammonium chloride lysis buffer (160 mM ammonium chloride, 170 mM Tris, pH 7.65). PBMCs were then pelleted by centrifugation at 1,000 rpm, at 19°C for 8 min. The pellet was resuspended in 50 ml PBS. A 10 ul aliquot of the resuspended PBMCs was stained using an equal volume of trypan blue solution (ThermoFischer. 0.4%), visualised and counted under a light microscope using a Neubauer haemocytometer.

2.2.5 Isolation of monocytes from PBMC

Monocytes were positively selected from PBMCs by cell surface expression of the CD14 receptor. PBMCs were incubated for 30 mins at 4°C with anti-human CD14 microbeads (Miltenyi Biotech) using 20 ul of beads per 10^8 cells. Unbound antibody was removed by washing twice with 20 ml MACS buffer (Phosphate buffered saline, 2% fetal bovine serum, 1mM EDTA, pH 7.2) and centrifuged at 1,500 rpm at 4°C for 4 min. Cells were isolated by passing the labelled PBMCs through MACS columns (Miltenyi Biotech) attached to a

magnet. Monocytes were eluted into 5 ml RPMI/10 (Thermo Fischer Scientific) and a 10 ul aliquot used for visualisation and counting with 10 ul Trypan blue on the Neubauer haemocytometer.

2.2.6 Preparation of HMW DNA in agarose plugs

Isolated monocytes were re-suspended in RPMI-1640 (Life Technologies) at a concentration of 5 x 10^7 cells per ml and equilibrated to 50°C. A 0.375 g/ml of low melting point (LMP) agarose (50°C) was added to each 1 ml of the cell suspension with a 0.625 ml volume of buffer. An aliquot of 100 µl of the monocyte mixture was pipetted into each of the ice cold disposable plug moulds (Bio-Rad Laboratories, Inc.) and hardened at 4°C for 30 min. The solidified plugs were incubated for 48 hours with 2 mg/ml Proteinase K (New England Biolabs, Inc.) in 10ml Proteinase K buffer (30 mM Tris HCl, 1 mM CaCl2, pH 8.0) at 50°C. The enzyme and buffer were replaced at 24 hours. Plugs were washed in phenylmethylsulfonyl fluoride (PMSF, 10 mM) for 1 hour to remove residual Proteinase K activity. Plugs were then washed three times in 50 ml of Tris-HCl (20 mM) and EDTA (50 mM) over three hours with gentle shaking and stored at 4°C in 50 mM EDTA. Each 100 ul plug was estimated to contain 6 µg of DNA: (3x10^9 bp/ genome) x 2 genomes x (660 g/mol/bp) x (1.67 x 10^{-12} pg / diploid cell).

2.2.7 Pre-electrophoresis of HMW DNA agarose plugs

Agarose plugs were dialysed in sterile 0.5 x TBE buffer at 4°C for 4 hours and then electrophoresed on a 1% LMP agarose gel using a CHEF-DRII apparatus (Bio-Rad Laboratories, Inc.) at 14°C, 4 V/cm for 10 hours with a 5 second pulse time. After electrophoresis, plugs were collected from the preparative slot and dialysed in TE buffer (pH 8.0) overnight. The gel was stained in 0.5 µg/ml ethidium bromide (EtBr) solution and visualised under UV light to establish removal of cellular debris from the plugs (see figure 2.6).

2.2.8 Complete digestion of HMW DNA

Complete digestion conditions were determined for the agarose plugs with EcoRI and HindIII. Agarose plugs were washed three times in excess endonuclease buffer (100 mM Tris-HCl, 50 mM NaCl, 10 mM $MgCl_2$, 0.025% Triton® X-100, pH 7.5). Plugs were then incubated on ice in 0.5 ml restriction buffer with 0.5, 1 or 2 U of EcoRI endonuclease for 0.5, 1, 2 or 4 hours to allow diffusion of the enzyme into the plugs. Reactions were then heated to 37°C for one hr to activate the enzyme. The reactions were run on a 1% agarose gel at 60 V for 4 hours. The optimum incubation time and enzyme concentration for sufficient diffusion into the plugs was calculated as 2 U of endonuclease for 2 hours.

2.2.9 Partial digestion of HMW DNA using magnesium chloride gradients

Agarose plugs were washed in TE buffer for 24 hours to lower EDTA concentration. Plugs were individually placed in microcentrifuge tubes with 0.5 ml magnesium free endonuclease buffer (sodium chloride (NaCl) 50 mM, Tris-HCl 10 mM, bovine serum albumin (BSA) 10 µg/ml, pH 8.0) at 4°C for one hour. After addition of 2 U of EcoRI endonuclease, plugs were incubated on ice for a further 2 hours. Magnesium chloride ($MgCl_2$) was added to reaction tubes for final concentrations of 0.01 mM, 0.1 mM, 0.3 mM, 0.5 mM and 1.0 mM. Control reactions contained no added endonuclease, no added $MgCl_2$ or a final concentration of 10 mM $MgCl_2$ for complete DNA digestion (Figure 2.7). Partial digestion reactions were incubated at 37°C for 30 min or one hr and stopped by addition of 10 µl EDTA (0.5 M). Plugs were dialysed in sterile 0.5x TBE buffer at 4°C for 4 hours before immediate size fractionation or dialysed in TE buffer and stored at 4°C.

2.2.10 Partial digestion of HMW DNA using EcoRI methylase competition

The HMW DNA was incubated with a set amount of EcoRI endonuclease and limiting amounts of EcoRI methylase for partial digestion through enzyme competition. Complete digestion conditions were previously determined with EcoRI endonuclease (2.2.9). Partial digestion was achieved by altering the concentration of the competing EcoRI methylase; plugs were incubated on ice for 2 hours with 2 U of EcoRI endonuclease and 0, 50, 100, 200 or 500 U of EcoRI methylase. Negative controls contained either: no magnesium, no EcoRI and EcoRI methylase or EcoRI methylase only. Two reactions were simultaneously run with

$MgCl_2$ depletion of 0.3 mM and 0.5 mM to compare the digestion gradient between methods. Reactions contained 25 µl BSA (10 mg/ml), 50 µl of 10 x EcoRI endonuclease buffer (100 mM Tris-HCl, 50 mM NaCl, 10 mM $MgCl_2$ unless otherwise stated, 0.025% Triton® X-100, pH 7.5) and 13 µl of 0.1 M spermidine. Reactions were incubated at 37°C for 30 mins before the addition of 150 µl EDTA (0.5M) and 30 µl of 10 mg/ml proteinase K to stop enzymatic reactions. The samples were then dialysed in TE buffer for storage or 0.5x TBE buffer at 4°C for immediate size fractionation.

2.2.11 Size fractionation of HMW DNA

Partially digested DNA was applied to a preparative slot of a 1% UltraPure Low Melting Point (LMP) Agarose gel (Invitrogen) and the slot was sealed with molten agarose. The Bacteriophage Lambda ladder PFGE Marker (New England Biolabs) was applied to the flanking wells. Size fractionation was performed in 0.5x TBE buffer in a CHEF-DRII apparatus (Bio-Rad Laboratories, Inc.) in three stages. The LMP gel was rotated 180° to the norm and ran at 5 V/cm for 6 hours with a 15 second pulse time, temperature maintained as close to 4°C as possible. The gel was then rotated 180° and electrophoresed under the same conditions. Finally, fresh 0.5x TBE buffer, chilled to 4°C, was added to the tank and the gel was run at 6 V/cm for 16 hours, with a 0.1- 40 second pulse time. Outer lanes containing markers were cut and stained in 0.5 µg/ml EtBr solution and visualised under UV light to assess digestion. Size fractionated horizontal blocks of 0.5 cm widths corresponding to 100-500 kb were excised and the gel slices stored in EDTA (0.5 M) at 4°C.

2.2.12 Extraction of HMW DNA

Size selected DNA fragments in the excised agarose gel slices were dialysed against TE buffer three times over 3 hours. Gel slices were added to 1 x agarase buffer (10 mM Bis-Tris-HCl, 1 mM EDTA, pH 6.5) at 4°C for 4 hours, replacing the buffer once. Agarose gel slices were then melted at 65°C for 30-45 minutes and cooled to 45°C for incubation with β-agarase (New England Biolabs, Inc.). Gel slices were incubated with 2 U of β-agarase per 100 µg of gel slice for one, two, and four hours to ensure complete digestion. After 2 hours, on completion of agarose digestion, the HMW DNA was stored at 4°C. To achieve higher DNA yield and purity the method evolved to a phenol chloroform extraction. After TE dialysis, the

initial volume of the gel slice was made up to 400 µl with sterile water. The gel was incubated at 65°C for 30-45 min before addition of 40 µl of saturated NaCl (5 M) and 400 µl of saturated phenol. The molten mixture was vortexed thoroughly and incubated on ice for 5 min to solidify the gel. Aqueous layers were washed by vortex and centrifugation at 13,000 rpm for 5 min in 400 µl of saturated phenol then 400 µl of chloroform. The aqueous layer was transferred to 1 ml ethanol (100%), 40 µl saturated NaCl (5 M) and stored at -80°C for over two hours. Pelleted the DNA at max speed for 20 min, 4°C and washed the pellet with 500 µl ethanol (70%). The DNA pellet was air dried before re-suspension in 20 µl of sterile deionised water.

2.2.13 Vector isolation

Bacterial vector pBAC-red was donated from the Children's Hospital Oakland Research Institute (CHORI) and pBeloBAC11 was isolated from the TPI4222 BAC library. All bacterial cultures were maintained at 32°C because of the temperature inducible exonuclease prophage in the pBAC-red vector, activated at 42°C. Electrocompetent NEB 5-alpha *E .coli* cells (New England Biosciences, Inc) containing the pBAC-red vector were streaked from the -80°C storage and a pBeloBAC11 *E. coli* clone from the cattle BAC library onto LB agar plates and incubated at 32°C for 16 hours. 40 ml of LB, containing 20 µg/ml chloramphenicol, was inoculated with a single vector colony and incubated then for 7 hours at 32°C, 225 rpm. The inoculated medium was split into 3.2 L of LB medium and grew overnight for 16 hours in a shaking incubator at 225 rpm. The bacteria were pelleted at 16,000 x g for 5 min in an Beckman Coulter (Optima XPN) ultracentrifuge. Bacteria were re-suspended in 10 ml sterile re-suspension solution (15 mM Tris-Cl pH 8.0, 10 mM EDTA, 100 mg/ml DNase-free RNase A). Cellular debris was precipitated by addition of 10 ml alkaline lysis solution (0.2 M sodium hydroxide and 1% SDS) then 10 ml precipitation solution (3 M potassium acetate, pH 5.5) and stood at room temperature for 5 min before centrifugation for 15 min, 16,000 x g at 4°C. The supernatant was removed and 0.8 ml aliquots were added to 0.8 ml isopropanol and vortexed before pelleting the DNA at 12,000 x g for 15 min. The pellet was washed with Ethanol (1 ml, 70%) at 6,000 x g for 2 min before allowing to air dry and re-uspending in 0.2 ml of sterile deionised water. A total of 236.4 µg of pBAC-red vector and 276 µg of pBeloBAC11 was isolated from this protocol.

2.2.14 Vector preparation

The minimal amount of endonuclease required for complete digestion of the vector was determined to prevent formation of empty clones. In 10 µl reactions of 50 ng vector DNA, 0.1, 0.2, 0.5 and 1.0 U of EcoRI were added to separate reactions and incubated at 37°C for 1 hour. Digestions were run on a 1% ultrapure LMP agarose gel at 85 V for 1.5 hours in a gel electrophoresis tank. The maximum amount of vector DNA was established by digesting 100, 200, 300, 500 and 700 ng of vector DNA with the previously determined minimum endonuclease concentration of 0.5 U EcoRI and gel electrophoresed. A total of 30 µg of vector DNA was then digested using 0.5 U of enzyme per 300 ng of vector DNA at 37°C for 15 min. After this, 1 U of Antarctic Phosphatase (New England Biolabs, Inc.) was added and incubation continued for 1 hour. Reactions were incubated with 15 mM EDTA and 200 µg/ml Proteinase K for 1 hour to stop reactions. Vector DNA was electrophoresed in a LMP ultrapure 1% agarose gel at 85 V for 1.5 hours. The gel was cut and the side panels stained with EtBr to determine the corresponding vector band for excision and storage in 0.5 M EDTA.

2.2.15 Vector purification

Vector DNA was purified using phenol chloroform at 1:1:1 ratio with the sample volume. The gel slice containing the pBAC-red and pBeloBAC11 digested vectors was incubated at 65°C for 30-45 minutes until the agarose melted. The reaction volume was made up to 400 µl with sterile water, 40 µl of saturated NaCl (5 M) and 400 ul of saturated phenol. The molten mixture was vortexed thoroughly and incubated on ice for 5 min to solidify the gel. Aqueous layers were transferred to clean tubes after a phenol chloroform wash, by centrifugation for 5 min at 13,000 rpm. The aqueous layer was removed and added to 1 ml ethanol (100%) and 40 µl of saturated NaCl (5 M) and stored at -80°C for a minimum of two hours. DNA was pelleted at 4°C at maximum speed for 20 min, and washed with 500 µl of 70% ethanol. The DNA pellet was air dried before re-suspension in 20 µl of sterile deionised water. A total of 5 µg of purified pBAC-red was digested with 1.0 U of EcoRI and BamHI restriction enzymes in two 50 µl reactions overnight at 37°C and electrophoresed on a gel at 55 V for 4 hours. The gel was stained with SYBR Safe (Thermo Fisher Scientific) and visualised under UV to confirm the isolation of the vector. A100 ng sample was sent for Sanger sequencing using the stock T7F and T7R sequencing primers.

2.2.16 Control ligations

Control ligations were carried out as controls for the ligation procedure for both vectors in use: pBAC-red and pBeloBAC11. A total of 50 ng of vector DNA was added to 10 ul ligation reactions with 1 U T4 ligase (New England Biolabs, Inc.) and incubated at 16°C overnight. Vector used was uncut plasmid, linearised plasmid before phosphatase treatment or phosphatase-treated linearised plasmid DNA.

2.2.17 Small scale BAC ligations

A series of ligations were performed with HMW DNA using insert:vector molar ratios of 10:1, 5:1, 2:1, 0.5:1 and 0.2:1 in 10 µl, 20 µl or 50 µl volumes. Within each reaction, 25, 50 or 100 ng of vector was used with 1, 10 or 100 U T4 DNA ligase. Reactions were incubated at 16°C overnight then stored at 4°C for several days before transformation. Alternatively, reactions were incubated at 16°C for four hours before transforming the same day or incubating at 4°C overnight.

2.2.18 Purification of ligation mix

Prior to electroporation, a 1.5 ml microcentifuge tube was filled with a 1% ultrapure LMP agarose (Thermo Fischer Scientific) in 100 mM glucose. A pipette tip was placed in the molten mix prior to solidification to form a well. The ligation mix was added to a microcentrifge tube and left to incubate for 2 hours on ice for diffusion of salts into the agarose (Atrazhen and Elliot, 1996). Alternatively, ethanol precipitation was also performed to reduce the burden of salts. The ligation mix was added to 1 ml ethanol (100%), 40 µl of saturated NaCl (5M) and stored at -80°C for over two hours or -20°C overnight. The DNA was pelleted at max speed for 20 min at 4°C and the pellet was washed with 500 µl of 70% ethanol. The DNA pellet was air dried before re-suspension in 20 µl of sterile deionised water.

2.2.19 Transformation

Test ligations were diluted 1:10 in deionised water to lower salt concentration where necessary and 2 ul of this ligation mix was added to 20 ul of ElectroMAX DH10B competent

cells (Life Technologies, Inc.) and transferred to 0.2 cm gap chilled cuvettes (Bio-Rad Laboratories, Inc.). Electroporation was carried out at resistance voltage booster 200 Ω, voltage gradient 13-25 kV/cm and capacitance 25 uF in a CellPorator (Bio-Rad Laboratories, Inc.). Cells were incubated for expression of antibiotic resistance in 980 µl of pre-warmed Super Optimal broth with Catabolite repression (SOC) at 32°C whilst shaking at 225 rpm. Cells were spread on LB plates containing 20 µg/mL chloramphenicol and incubated at 32°C for 24-48 hours to allow sub-optimal growth of clones. Transformation efficiency was calculated on the number of colony forming units (CFU) grown from transformation of 1 ug of plasmid into 100 ul of electrocompetent cells.

2.2.20 Acquisition of a Pacific Biosciences long read cattle genome

The cattle genome of a single Hereford cow (L1 Dominette 01449) was sequenced and *de novo* assembled using long reads generated with the Pacific Biosciences RSII platform (Assembly: ARS-UCDv0.1) by Tim Smith of the Meat Animal Research Center, Clay Center, United States Agricultural Department (USDA). PBMC from Dominette was used to derive DNA for construction of the SMRT sequencing libraries using P5-C3, P4-C2 and XL-C2 SMRT cell chemistry. The mean read length of 5.1 kb was generated from a total of 7.4 Gb of sub-read bases. The Celera Assembler PacBio Corrected Reads pipeline was then used for assembly as detailed elsewhere (Smith and Medrano, unpublished), as has been done for the recent goat genome (Bickhart et al., 2017; 161). Reads are error corrected using Quiver and scaffolding was done using a tiered approach with both Irys optical mapping technology (BioNano Genomics) and Hi-C-based proximity-guided assembly to resolve contig orientation mistakes and any other mis-assemblies. This cattle genome is currently unpublished but was shared with our Immunogenetics group through collaboration with Tim Smith (USDA).

The ARS-UCDv0.1 assembly was searched for contigs containing the antibody loci using blastn to search for antibody sequence motifs such as the *IGHV* and *IGLV* leader sequence, W/F-G-X-G motif in *IGHJ* and *IGLJ* and motifs in the Framework regions of the *IGHV* and *IGLV*. Positive scaffolds were extracted from the genome using the grep command shown below.

```
grep -A ‘linenumber’ >’contigname’
‘databasename’.fasta > ‘outputcontigfile’.fasta

grep -c > ‘outputcontigfile’.fasta
```

2.3 Results

2.3.1 TPI4222 BAC Library screening

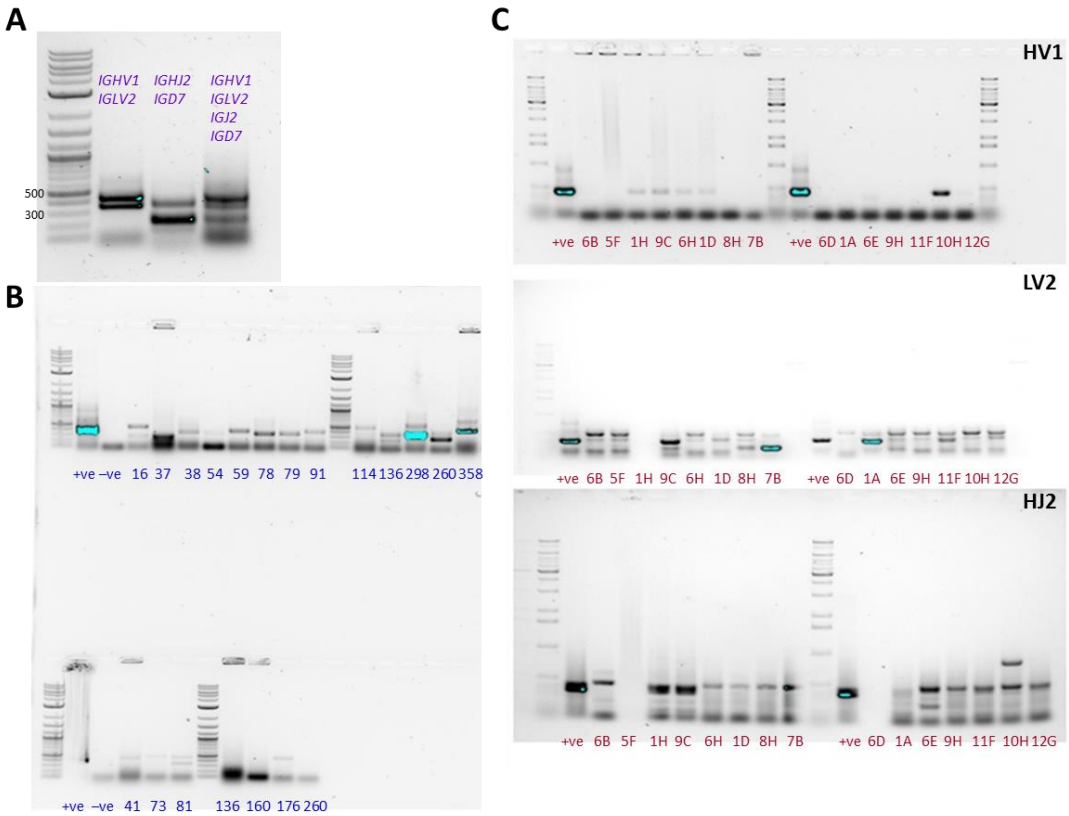


Figure 2.4: The TPI4222 BAC library from a Friesian cow at 1.6 x genome coverage was screened for clones containing the antibody loci. The library was constructed by di Palma (1999) in the vector pBeloBAC11 (Di Palma, 1999; 151). This was made PCR screenable by pooling each 96 well plate into a single well of another 96 well plate. Primers for amplification of antibody gene segments were optimised on cattle 598 gDNA for their annealing temperature, $MgCl_2$ concentration and multiplex capabilities to produce specific and distinguishable bands on a gel (A). UMD_HV1 primers produces a clear band at ~450 bp, UMD_HJ2 is at ~300 bp and UMD_LV2 is ~350 bp. The *IGHD7* primers were discarded as the amplicon could not be distinguished from the *IGLV* and the primers produced multiple non-specific bands on the BAC clones. The UMD_HV1, UMD_HJ2 and UMD_LV2 primers were multiplexed and each well of the pooled 96 well BAC library plates was screened for the antibody loci, using cattle 598 gDNA as a positive control (B). The negative control contained no template. From the four pooled BAC library plates, 17 plates were identified as potentially containing positive clones and these were subsequently screened with the multiplex primer set to identify 15 positive wells (C). The ladder used in each gel is the GeneRuler 1 kb Plus DNA Ladder and the 500 bp and 300 bp bands are indicated in (A).

Primers designed for BAC library screening were optimised on cattle gDNA. The optimum annealing temperature and magnesium chloride concentration for PCR reactions was determined as 58ºC and 2.0 mM for the amplification of clear, specific bands of highest yield. The UMD_HV1, UMD_HJ2, UMD_D7 and UMD_LV2 primers each produced the most specific amplicons with the highest yields and so these primer sets were chosen for BAC library amplification. The multiplex capability of these primers was confirmed so that band sizes were decipherable on the gel (figure 2.4A); UMD_HV1 produced a ~450 bp band, UMD_HJ2 was ~300 bp and both UMD_LV2 and UMD_D7 was ~350 bp. The *IGHD* primers mostly produced non-specific amplicons which made determining positive bands more difficult. The size of the *IGHD7* amplicon was indistinguishable from the *IGLV2* and when subsequently amplifying with UMD_D7 primers from the pooled BAC plate these primers produced multiple non-specific bands and so the *IGHD* primers were excluded from the multiplex master mix.

The TPI4222 BAC library consists of 374 BAC plates which were pooled into four 96 well plates for ease of PCR screening. Of the four pooled BAC library 96 well plates, 17 distinct PCR bands were identified as positive BAC library 96 well plates (figure 2.4B) corresponding to a positive clone on the plates: 16, 37, 38, 41, 54, 59, 73, 78, 79, 81, 91, 114, 136, 160, 176, 260, 298 and 358. Upon screening the corresponding individual BAC plate, 15 positive wells amplified with the multiplex primer set: 1A (plate 37) , 1D (plate 38), 1H (plate 16), 5F (plate 16), 6B (plate 298), 6D (plate 141), 6E (plate 79), 6H (plate 59), 7B (plate 37), 8H (plate 38), 9C (plate 37), 9H (plate 91), 10H (plate 73), 11F (plate 78), 12G (plate 81). These positive clones were then isolated for amplification of the antibody gene segments individually with the UMD_HV1, UMD_LV2 or UMD_HJ2 primers (Figure 2.4C). The clones in each positive well were streaked onto agar plates, grown and subsequently purified at a concentration of 193.7- 406.1 ng/µl. Clones were nanodropped to confirm purity (260/280 of 1.84-2.02 and 260/230 of 2.09-2.29). PCR amplification with each primer set identified 11 clones which appear to contain the antibody loci; six clones appeared to contain the heavy chain (1D, 1H, 6E, 6H, 9C and 10H) with all except 6E having positive bands for *IGHV*. The clone 6E was positive for the *IGHJ* loci, as well as two others (1H and 9C). Five clones appeared to contain the light chain, positive for *IGLV2* (1A, 7B, 9C, 11F and 12G). The single clone, 9C, was positive for both the heavy and light chain genes but this is structurally impossible as the two loci are contained on different chromosomes.

The 11 positive clones were then EcoRI digested to determine the clone size and look for similar banding patterns between clones. If multiple clones contained antibody loci they would show similar banding patterns on the gel; however, whilst each clone contained a fragment of >50 kb, none of the clones had an identical digestion pattern (Figure 2.5). The positive clones were subsequently sent for Sanger sequencing which revealed that the putatively positive clones are all non-specific to the antibody loci. The Sanger sequencing products mapped to the UMD3.1 genome at locations on other chromosomes to the antibody loci. Whilst their sequences contained remnants of the repetitive antibody motifs, they were not the antibody loci. This suggests the nonspecific binding of the primer sets and that the conserved repetitive antibody sequences occur as fragments in the genome.

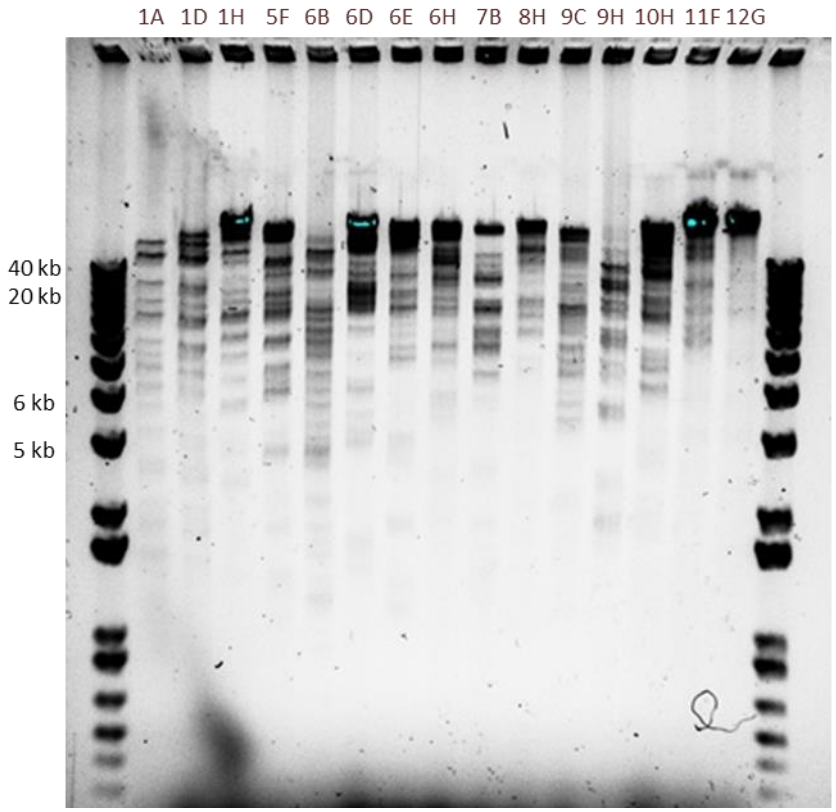


Figure 2.5: EcoRI digestion pattern of the 15 positive wells for the antibody loci in the TPI4222 Friesian BAC library: 1A, 1D, 1H, 5F, 6B, 6D, 6E, 6H, 7B, 8H, 9C, 9H, 10H, 11F, 12G. Clones were identified using a multiplex primer set for *IGHV*, *IGHJ* and *IGLV* gene segments and subsequently grown and purified. The EcoR1 digestion looked for similar banding patterns between clones containing identical regions of the antibody loci. None of the restriction patterns appear similar but all the clones contain large inserts >50 kb. The ladder used is the GeneRuler 1 kb Plus and the 40 kb, 20 kb, 6 kb and 5 kb bands are indicated on the left of the ladder.

2.3.2 Isolation of cattle HMW DNA for BAC library construction

PBMCs were isolated from peripheral blood at an average of 3.02 x 10^8 cells/ml, and of these, roughly 10% were monocytes, isolated at an average of 4.24x10^7 cells/ml. HMW DNA was prepared inside agarose plugs and stored in 0.5M of EDTA. Digestion of the cells in the plug moulds with Proteinase K was successful and cellular debris is shown to be cleared from the plugs during pre-electrophoresis (Figure 2.6).

Figure 2.6: HMW DNA was isolated from cattle 705983 by monocyte isolation and prepared inside agarose plugs for protection of the DNA from shearing. Embedded monocytes in the agarose plugs were digested with proteinase K and cellular debris was removed from the plugs by gel electrophoresis. The gel image shows the movement of cellular debris from the agarose plugs.

2.3.3 Digestion of HMW DNA and size selection

Diffusion of the molecular reagents into the agarose plugs is considered a limiting factor for their digestion. The effect of diffusion times and enzyme concentration on complete digestion of the HMW DNA inside the agarose plugs was first considered. Individual plugs were incubated on ice for 0.5, 1, 2 or 4 hours with 0.5, 1 or 2U of EcoRI restriction enzyme to determine the conditions for optimum enzyme diffusion (figure 2.7). The optimum enzyme concentration is 2U per plug but no difference is observed in digestion reactions between two and four hours of enzyme diffusion time. Despite the optimisation of conditions, a compression zone of concentrated DNA is observed around 500 kb, which does not alter as a function of enzyme concentration or incubation time. This suggests diffusion of the molecular reagents is still a limiting factor in DNA digestion.

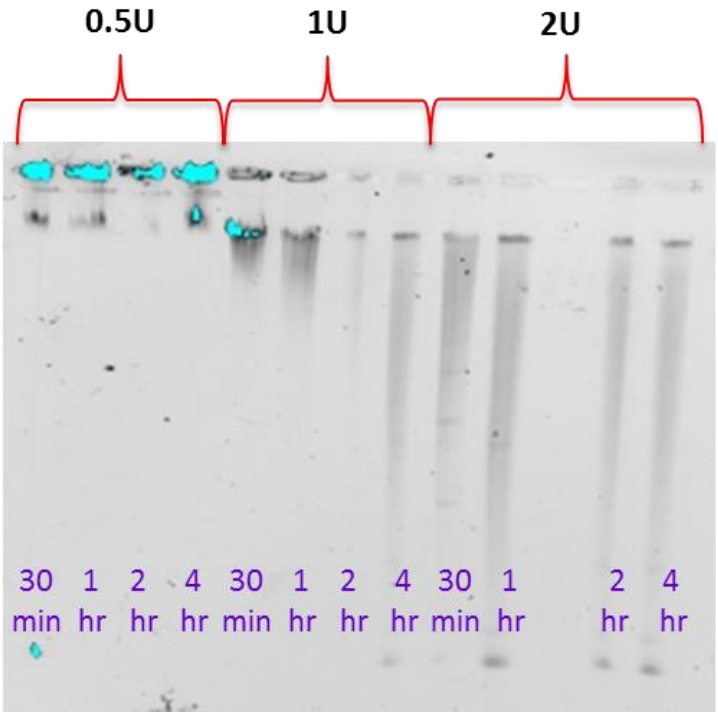


Figure 2.7: EcoRI restriction enzyme digestion of HMW DNA embedded in agarose plugs to determine the optimum conditions for complete digestion of the DNA. The plugs were incubated on ice with 0.5, 1 or 2U of EcoRI enzyme for 0.5,1, 2 or 4 hours before enzyme activation and digestion at 37 ºC for 1 hr. Complete digestion occurs with 2U of EcoRI per reaction with an incubation of at one hr.

Partial digestion of HMW DNA with EcoRI or HindIII was initially attained with a $MgCl_2$ gradient, limiting restriction enzyme activity. EcoRI was first incubated with the plugs on ice to allow enzyme diffusion, as previously determined, into the plugs before activation and incubation at 37°C for one hour. These conditions appeared to over-digest the DNA and did not produce a discernible gradient on the gel. When EcoRI was incubated at 37°C for 30 min, a digestion gradient becomes more apparent.

Inherent nuclease activity was tested in the plugs as a negative control with no restriction endonuclease added. Migration of large DNA fragments to the compression zone is seen in both negative controls, suggesting genomic shearing is taking place during the preparation. However, intact HMW DNA remained largely in the well. The concentration of DNA present in the compression zone increased as a function of $MgCl_2$ concentration until complete digestion of the DNA; at 10 mM $MgCl_2$ concentration and incubation for 1 hr no DNA was

seen above the compression zone, or in the well, and the resulting smaller fragments of >48.5 Kb migrated toward the bottom of the gel.

HMW DNA embedded in agarose plugs was later partially digested using EcoRI restriction enzyme competition gradients with the EcoRI methylase. Enzyme competition to produce a digestion gradient is preferable over a $MgCl_2$ concentration gradient because results are more reproducible as there is more control over enzyme kinetics; competition between the enzymes provides a more evenly distributed digestion pattern throughout the genome. The concentration of EcoRI was kept constant in reactions at 2U whilst the concentration of competing EcoRI methylase varied between 0 and 200U. Increasing EcoRI methylase concentration decreased DNA digestion resulting in a DNA size gradient on the gel (figure 2.8). Complete digestion was achieved in the positive control with 10mM $MgCl_2$, 2U of EcoRI and 0U of EcoRI methylase. Genomic shearing is still taking place in preparatory stages, as shown in the migration of DNA from the wells in the negative controls. The compression zone at ~350 kb is seen in the reactions due to poor diffusion rates of the enzymes into the agarose plugs. Larger BAC library inserts are preferable in order to achieve sufficient coverage of the genome; the 150- 250 kb size fraction, below the compression zone, was excised and stored in EDTA (50mM) for subsequent BAC vector ligation.

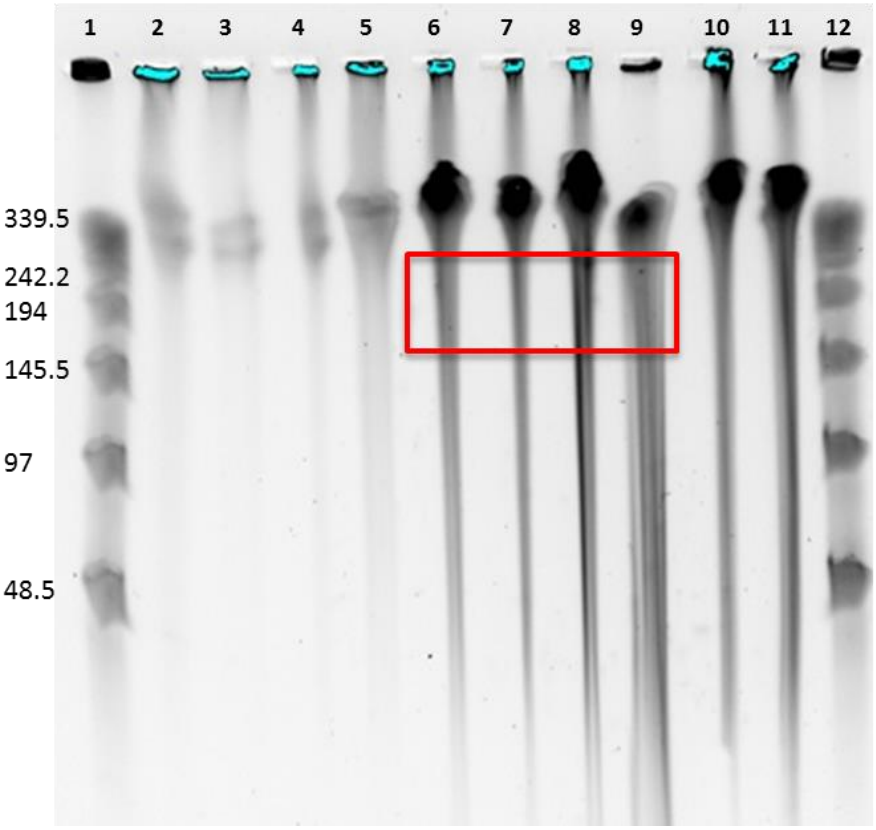


Figure 2.8: EcoRI restriction enzyme digestion of HMW DNA embedded in agarose plugs using a gradient of the concentration of a competing enzyme EcoRI methylase. The negative controls contained either no magnesium (Lane 2), no EcoRI and no EcoRI Methylase (Lane 3) and no EcoRI only (Lane 4). Test reactions contained 2U of EcoRI restriction enzyme with varying concentrations of EcoRI methylase: 500 U EcoR1 methylase (Lane 5), 200 U EcoR1 methylase (Lane 6), 100 U EcoR1 methylase (Lane 7), 50 U EcoR1 methylase (Lane 8) and the positive control with 0 U EcoR1 methylase (Lane 9). Lanes 10 and 11 contained 2U of EcoRI restriction enzyme with 0.3mM $MgCl_2$ or 0.5mM $MgCl_2$ as comparisons for the digestion gradients between the two methods. All reactions were incubated for 30 min at 37ºC. Lane 1 and 12 both contain the Bacteriophage Lambda ladder PFGE Marker (New England Biolabs) with the band size indicated to the left. The red box indicates the region that was excised for downstream experiments.

2.3.4 Isolation of pBAC-red and pBeloBAC11 vectors

The BAC vectors pBAC-red and pBeloBAC11 were successfully grown from the -80°C storage on agar plates. A total of 12.25 ug of pBAC-red and 16.71ug of pBeloBAC11 was isolated from the inoculated LB. Isolation of the vectors was confirmed by correct band size produced after restriction enzyme digestion with EcoRI and BamHI (Figure2. 9). To prevent the formation of empty clones, optimal digestion conditions of the vector was established as

0.5U EcoRI per 300 ng of vector, by first limiting the enzyme concentration and then the maximum vector concentration. The vectors were then completely digested and each vector excised from a gel and purified by beta-agarase digestion at an average concentration of 0.226 ug of pBAC-red and 0.797ug of pBeloBAC11. Purification from the gel by phenol chloroform extraction increased the yield to 2.45 ug of pBAC-red. The vector pBAC-red was confirmed by Sanger sequencing of a 1 kb region of the vector with T7F and T7R that had a 100% nucleotide sequence identity to the original vector sequence.

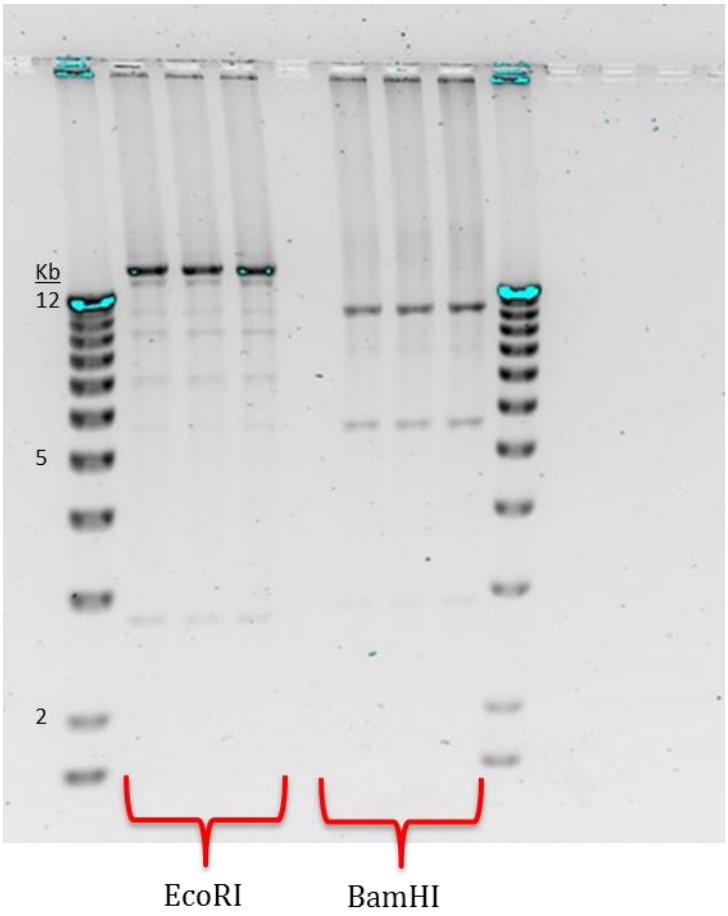


Figure 2.9: Gel image of the EcoRI and BamHI digest of pBAC-red. The expected band sizes for EcoRI are 15,725bp and 2800bp and for BamHI are 10.255bp, 5416bp and 2809bp. The ladder used is the 1 Kb Plus DNA Ladder (Invitrogen).

2.3.5 Isolation of HMW DNA

DNA isolated by phenol chloroform extraction was of higher concentration and quality than DNA purified by beta-agarase digestion. Phenol chloroform extraction isolated 1.78 ug of HMW DNA whilst beta-agarase digestion isolated 0.046 ug of HMW DNA

2.3.6 Transformation efficiency

A series of test ligations were transformed into *E. coli* by electroporation to determine the efficiency of the ligation, dephosphorylation and transformation steps. Transformation efficiency was calculated by the number of colony forming units produced from transforming 1 μg of plasmid. The BAC vectors pBACred and pBeloBAC11 formed on average 167 and 478 CFUs per 50 ng of vector that was transformed, without any digestion steps. This was the highest transformation efficiency of 3.64×10^8 for pBAC-red and 8.62×10^9 for pBeloBAC11 (Figure 2.10). Digestion and re-ligation of the vector to itself reduced transformation efficiency 100 fold and only 31 and 43 CFUs subsequently grew on agar plates for pBACred and pBeloBAC11respectively. The dephosphorylation step, to improve ligation efficiency of HMW DNA into the BAC vectors, decreased transformation efficiency by a further 10-fold. A single CFU for pBAC-red grew on the agar plate and only five CFU for pBeloBAC11. The BAC vector pBAC-red, when gel purified to remove contaminants, was not successfully transformed. Transformation with vector ligated to HMW DNA often arced, even after purification steps and reactions had to be diluted in water 1 in 10 to reduce the salt concentration. Despite altering the voltage gradient to enhance transformation efficiency, vector with HMW DNA inserts failed to grow on the plates.

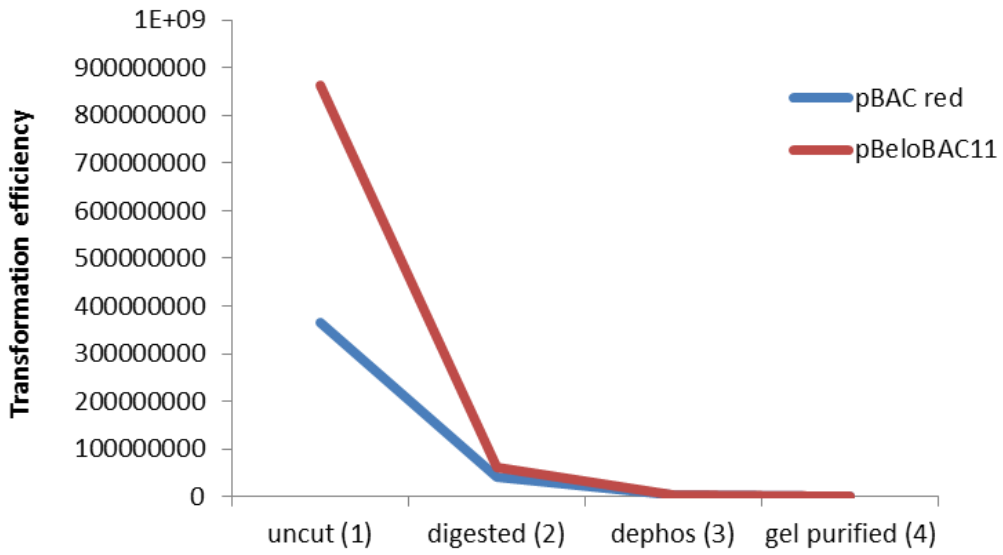


Figure 2.10: The average transformation efficiency of the BAC vectors pBAC-red and pBeloBAC11 transformation controls. The vectors were either transformed directly without being digested (1), were digested and re-ligated to themselves (2), were digested and de-phosphoylated before re-ligation to themselves (3) or were digested and gel purified before re-ligation to themselves (4). The graph is in a logarithmic scale and shows the transformation efficiency markedly drops with each procedure.

2.3.7 Acquisition of the cattle antibody loci in the Pacific Biosciences genome

The ARS-UCDv0.1 cattle genome was *de novo* assembled from long read SMRT sequencing with Pacific Biosciences of the Hereford cow L1 Dominette 01449). This assembly was interrogated for contigs containing the antibody loci. The antibody heavy chain locus (IGH) is confined on two scaffolds, consisting of three contigs. The light chain loci, lambda (IGL) and kappa (IGK) are found on eight scaffolds consisting of 12 contigs. These were extracted for further analysis of the cattle IGH (Chapter 3) and IGL (Chapter 4).

2.4 Discussion

The stable maintenance of large DNA fragments in BAC clones is a useful tool for the sequencing of genomic regions of interest. The desired region can be selected for and sequenced in depth to achieve a reliable map of a single haplotype of the genomic region. This is particularly useful for targeted sequencing of the antibody loci which are highly polymorphic and repetitive sequences. The availability of an accurate and high coverage cattle antibody germline sequence is essential for its characterisation.

Current cattle genomes, the UMD3.1 and Btau3.1, have been sequenced and assembled with short read sequencing technologies which has resulted in severe fragmentation of their antibody loci; explored later in Chapters 3 and 4. The Btau_3.1 genome was assembled using the CHORI-240 BAC library to construct complete genomic maps; this BAC library had 10 x coverage of the genome but despite this, an online search of the BAC library for clones containing the antibody loci was unsuccessful. The mis-assembly of the antibody loci throughout the Btau_3.1 genome suggests the loci were over-digested and not confined to a single BAC clone. A smaller BAC library was constructed at The Pirbright Institute by di Palma (Palma, 1999; 162) and was subsequently screened using PCR primers specific to conserved sequences in the antibody gene segments, however this was unsuccessful. Restriction sites in the antibody loci are estimated to occur every 3900 bp compared to every 4100 bp on average in the genome. This slightly higher frequency of restriction sites and unknown differences in DNA secondary structure may explain their exclusion from BAC libraries if they are over-digested.

The TPI4222 BAC library has ~1.6x coverage of the cattle genome so the probability of any given genomic region being included in this library is approximately 80%. However, the screening and sequencing of this library with antibody specific primers was unsuccessful. The primers were shown to be specific on gDNA but BAC DNA is low copy number and of low purity, due to contaminating bacterium DNA and crude lysate, which reduced the efficiency of the PCR, making analysis of the PCR difficult. Clones which were seemingly identified as positive with the antibody specific primers, mapped to the UMD3.1 genome to other chromosomal locations and to regions containing antibody-like motifs. The low coverage of this library may also have meant the antibody loci were excluded from the

ligation process. This library was also constructed using PBMC which is undesirable for sequencing the antibody loci. These cell populations would contain B cells with their antibody loci rearranged which was suspected of a nonspecific clone that appeared positive for both *IGHV* and *IGLV* gene segments. The antibody heavy chain (HC) and antibody light chain (LC) are found on separate chromosomes, 17 and 21 respectively, so only a rearranged B cell would be positive for both immune loci. In this instance, the clone was false positive for both loci. Selecting monocytes for BAC library construction is favourable as the antibody loci would not have undergone rearrangement or somatic hyper mutation. The low coverage and non-specific selection of clones meant the antibody loci is excluded from current BAC libraries.

Advancements in BAC technology allow the targeted selection of clones containing the desired genomic region though Recombineering technology. Using a BAC vector containing temperature inducible HR genes, an antibiotic resistance marker with flanking primers that match the sequence of interest can be introduced for the positive selection of clones containing the antibody loci. Several BAC libraries, including a mouse and orang-utan BAC library, have been constructed using the vector pBAC-red, utilised here (Nefedov et al., 2011; 155). Application of this vector was shown to be stably propagated without leaking expression of the recombination proteins (Nefedov et al., 2011; 155). Thus, the construction of a temperature inducible HR cattle BAC library in pBAC-red was attempted for the isolation and sequencing of the antibody region.

The HMW DNA is needed for improved BAC library construction; for larger BAC clone inserts and therefore higher library coverage. The isolation of HMW DNA from cattle monocytes was optimised. The HMW DNA was prepared inside agarose plugs and stored in a high concentration of EDTA, which provides a shear-resistant environment that inhibits nuclease activity. Despite this, subsequent digestion of the DNA was still observed in the negative controls containing no restriction enzyme, suggesting mechanical shearing of the DNA is still taking place. Pre-electrophoresis of plugs prior to DNA separation allowed the removal of cellular debris, which may have hindered DNA separation and subsequent cloning steps. The optimal digestion conditions of the plugs were determined with both magnesium chloride concentration gradients and EcoRI restriction enzyme competition with the EcoRI methylase. The latter was considered to be more reproducible and likely to contain more intact antibody clones despite both being visually identical on the gel. However, a

compression zone was still observed at ~350 kb on the gel as the enzyme diffusion into the plugs is still a limiting factor. Agarose microbeads are seen as an alternative with advantages in DNA handling. The use of beads increases the surface area surrounding the DNA sample 100 fold for more efficient diffusion of restriction enzymes, which would prevent the compression zone seen in the gel images. However these were not used, as they are technically very challenging to handle and absorb low concentrations of DNA on the beads.

HMW DNA was isolated after size fractionation on the agarose gels with phenol chloroform extraction. This increased the yield and quality of the DNA compared to beta-agarase digestion; the melting and subsequent agarose digestion may have compromised the integrity of the DNA, reducing downstream ligation efficiency. Despite significant improvements in phenol chloroform extraction procedures, ~80% of the DNA being loaded onto the gel was still being lost. An electro elution procedure may have recovered higher yields of HMW DNA without loss of DNA integrity from a melting step or phase separation.

The BAC vectors, pBAC-red and pBeloBAC11, were successfully grown and isolated from a gel, as shown by multiple restriction endonuclease digestion patterns and Sanger sequencing of pBAC-red. Uncut, purified BAC vectors were transformed into *E. coli* at a high transformation efficiency of ~3 x 10^8 CFU for pBAC-red and ~9 x 10^8 CFU for pBeloBAC11 and were reclaimed from the cells to show the transformation to be successful. Each subsequent processing step however, markedly reduced the transformation efficiency over ten fold. Digestion and re-ligation of clones reduced transformation efficiency even in the absence of HMW DNA inserts and the dephosphorylation of vector ends to improve the ligation of DNA inserts reduced the transformation efficiency a further tenfold. Attempts to ligate purified HMW DNA into the BAC vectors and transform them into *E. coli* cells were unsuccessful. The pBAC-red vector is limited in its restriction enzyme sites for altering the base overhangs. The ligation procedure then is sub-optimal and needs refinement. The transformation of BAC clones containing HMW DNA inserts may have negatively affected the transformation efficiency and also needs improving. Overall, HMW DNA and the BAC vectors were successfully isolated but the ligation and transformation procedures need refinement.

The continued development of the temperature inducible BAC library protocol was stopped as collaboration with Tim Smith at the Meat Animal Research Center, USDA, superseded the

need for its construction. The new cattle long read genome, the ARS-UCDv0.1, is sequenced with SMRT technology with Pacific Biosciences. The long reads are better able to span highly repetitive regions of the genome, such as the antibody loci for more accurate assembly and mapping of the loci then short read sequencing technology, as shown in Chapter 3. The search for antibody contigs in the ARS-UCDv0.1 revealed the loci confined to far fewer contigs than in previous characterisations of the antibody heavy chain and light chain. This genome was used to characterise the IGH, IGL and IGK in cattle which is explored in Chapter 3 and Chapter 4.

Chapter 3

Structural determination of the cattle IGH locus and characterisation of the cattle and African buffalo (Syncerus caffer) IGH gene segments

3 Abstract

The immunoglobulin heavy chain (IGH) contains numerous gene segments for formation of the large polypeptide subunit of an antibody. Rearrangements of the VDJ gene segments form the CDRH including the ultra-variable CDR3H which provides the main peptide binding residues. The cattle IGH was assembled in the genome, the ARS-UCDv0.1, sequenced with Single Molecule Real-Time (SMRT) Sequencing, a long-read technology with Pacific Biosciences. The IGH organisation and structure was compared between the ARS-UCDv0.1, the previously annotated IGH in the short-read genome assembly, UMD3.1, and a recently published map of the cattle IGH constructed with Sanger sequencing of BAC clones (Ma et al., 2016; 98). Duplications exist in the cattle IGH of *IGHV-IGHDn-IGHJn-IGHM-IGHD* and *IGHV-IGHDn-IGHD* regions, resulting in three and two functional *IGHDn* and *IGHJ* gene segment clusters respectively and two functional *IGHM*. Discrepancies exist between the structure of the ARS-UCDv0.1 and the IGH map from Ma et al (2016) where an additional *IGHV-IGHDn-IGHD* region is present, nearly identical in sequence and structure to the previous. The BAC clone RP42-567N23, used by Ma et al to construct the duplicated regions in their IGH sequence, was sequenced with SMRT and Oxford Nanopore and assembled into an IGH structure identical to the ARS-UCDv0.1. Single nucleotide polymorphism (SNP) calling of the RP42-567N23 SMRT reads however revealed the existence of the additional *IGHV-IGHDn-IGHD* region, confirming the sequence described by Ma et al (2016) to be the more accurate. The organisation, complexity and putative expression of the cattle IGH was compared to the African buffalo. The IGH in African buffalo (*Syncerus caffer*) was *de novo* assembled by mapping reads from the African buffalo genome sequencing project (Glanzmann et al., 2016; 1) to the cattle ARS-UCDv0.1. Interestingly, duplicated regions of the IGH observed in the cattle appear absent in the African buffalo. Putative functionality, and therefore recombinatorial potential of the IGH, is confirmed with RNA-seq from Holstein cattle and African buffalo RNA reads revealing a limited recombinatorial potential of the IGH germline repertoire in both species.

3.1 Introduction

3.1.1 Comparison of short and long read sequence technologies for IGH assembly

Knowledge of DNA sequence is indispensable for understanding biological systems. The development of Sanger sequencing, via selective incorporation of chain-terminating dideoxynucleotides, permitted the first high throughput method of determining the genetic code. Sanger sequencing produces DNA sequences of approximately 800 bp which can then be assembled into larger contigs. The generation of the first cattle genome in 2004 (NHGRI, 2004; 163), a 3.3-fold coverage of the Hereford cow L1 Dominette 01449, was achieved by shotgun assembly of Sanger reads. This method is still widely used for the sequencing of smaller regions of the genome and the resolution of difficult regions, however it is considered an expensive method for whole genome assembly that does not provide the depth of coverage achieved by new sequencing technologies.

Second generation sequencing technologies (SGS), especially Illumina sequencing, transformed the field of genomics by providing comparatively high coverage of a whole genome at low cost. Illumina technology incorporates fluorescently labelled nucleotides to detect each base added by polymerase to the growing DNA chain. Clusters are generated by capturing the DNA library on a flow cell of surface-bound oligos and amplifying each individual fragment into distinct clonal clusters. The clusters of millions of fragments are extended in a parallel fashion in order to detect the fluorescent signal. The resulting reads are highly accurate base-by-base sequences with a very low error rate. The maximum read length however is short, with the Illumina HiSeq 2500 generating paired-end reads up to only 250 bp. Illumina technology has been adopted for sequencing and assembly of whole genomes, including the Hereford L1 Dominette 01449, covering most of the genome at 9.5-fold coverage, UMD2 (Zimin et al., 2009; 142) and has since been used for the sequencing and assembly of the most recently published cattle genome, the UMD3.1 (Elsik et al., 2009; 159).

The uniform short read length of SGS and amplification biases can lead to fragmented genome assemblies, particularly across highly repetitive, GC-rich and GC-poor regions. Short reads are unable to span repetitive regions of the genome. If the read does not contain a unique sequence, the origin of the read cannot be precisely determined and so multi-mapping occurs. The consequent multiple alignments and misalignments lead to sequence gaps,

assembly errors and incorrect abundance estimation. Consequently, the antibody loci are highly repetitive, GC-rich regions which are heavily disrupted with large sequence gaps in the available reference assemblies. Resolution of the antibody loci then, requires longer reads as they are more likely to contain unique sequences and span repetitive regions.

Third generation sequencing technology, Pacific Biosciences (PacBio) Single Molecule Real-Time (SMRT) sequencing, overcomes many of the limitations of Illumina sequencing, especially short read length and amplification biases. The sequencing-by-synthesis technology uses a zero-mode waveguide (ZMW) with an affixed DNA polymerase and a single template molecule. As fluorescently tagged nucleotides are incorporated along the chain, real-time imaging of the fluorescent signal is illuminated in the ZMW structure to allow observation at the single molecule level; unlike Illumina which detects the fluorescent signal from a cluster of amplified fragments. The use of DNA polymerase and the imaging of single molecules means there is no degradation of signal over time so the sequencing reaction only ends when the template and polymerase dissociate. The resulting read lengths are much longer, averaging at 10 kb, with over half of the reads >20 kb, using the latest chemistry. SMRT is however, not without its own limitations. The throughput of SMRT sequencing is lower than that of Illumina technology, typically at 0.5–1 billion bases per SMRT cell compared to the 8 billion paired-end 125 bp reads capable of being produced on the Illumina Hi-Seq 2500 (Rhoads and Au, 2015; 164). Individual reads contain a random 11-14% error rate (Korlach, 2015; 165) but with sufficient coverage of the read, the statistically averaged consensus eliminates most of the errors in the sequence as it is highly unlikely the same error will be randomly observed multiple times. Accuracy of >99% requires a coverage of 15 sequencing passes but the number of sequencing passes and the read length is a trade off as the read length is limited by the lifetime of the polymerase. Sequences have lower accuracy if they have longer lengths, shorter lengths yield higher accuracy. However, PacBio overcomes issues introduced by sequencing with Illumina as longer reads are able to span across highly repetitive sequences for assembly of complex genomic regions, such as the antibody loci.

Oxford Nanopore technology is another third generation sequencing technology which generates long sequence reads. The technology involves passing an ionic current through the nanopore and measuring changes in the current as biological molecules pass through. The current changes are different between the four nucleotide bases and so the DNA sequence can be determined. Repetitive regions, such as the antibody loci, can be sequenced without difficulty as the base calling is independent of the sequence which came before it. No limit

exists to the length of the DNA molecules being sequenced and so the technology has applications in sequencing entire chromosomes. But where PacBio reads a molecule multiple times to error correct for generation of a high quality consensus, Oxford nanopore can only sequence a molecule twice. Oxford nanopore therefore has an estimated 38.2% error after base calling (Laver et al., 2015; 166). The principle advantage of Oxford nanopore over other sequencing technologies is its affordability and portability which makes it useful in field studies.

For highly contiguous and accurate genome assemblies then, a conjunction of both PacBio sequencing and Illumina sequencing would be advantageous. PacBio can close gaps in reference assemblies and the long reads are able to overcome limitations of genome assembly using Illumina sequencing. Illumina, however, provides depth of coverage and higher accuracy. A cattle genome utilising both Illumina and SMRT reads for assembly however is currently not available. Here we are able to compare separate genome assemblies from PacBio and Illumina sequencing for post-assembly analysis of their structure and sequence.

3.1.2 Existing annotations of cattle and African buffalo IGH

The most recently published IGH genome annotation was the cattle UMD3.1 (Elsik et al., 2009; 159), sequenced with Illumina second generation sequencing. This contains large sequence gaps and assembly errors in the antibody loci. The *IGHV* in the *Bos taurus* genome UMD3.1 were characterised (Niku et al., 2012; 167); a total of 31 *IGHV* gene segments were identified in the genome, on 27 contigs with a further 5 *IGHV* found in the associated NCBI trace archives. The overall organisation of the IGH is impossible to determine in the UMD3.1 as gene segments are assembled on multiple chromosomes.

The 5’ end of the cattle IGH locus was resolved with Sanger sequencing of Holstein cattle BAC clones (Zhao et al., 2003; 168). The genes were shown to be arranged in a 5’-*IGHJ*-7kb-*IGHM*-5 kb-*IGHD*-33 kb-*IGHG3*-20 kb-*IGHG1*-34 kb-*IGHG2*-20 kb-*IGHE*-13 kb-*IGHA*-3’ order, spanning 150 kb and mapped to BTA21. In this BAC assembly, six *IGHJ* gene segments, two of which are functional, were identified across a ~2 kb region. A second *IGHM* was identified on BTA11 (Hayes and Petit, 1993; 169), which was associated with a duplicated *IGHJ* region, later reported on BAC clones, and which also mapped to BTA11

(Hosseini et al., 2004; 170). The two *IGHJ* regions had limited structural differences and few base mutations in their corresponding gene segments. The putatively functional *IGHJ* gene segments in both IGH loci rearranged and were present in RNA transcripts. The ability of cattle to rearrange both the *IGHJ* from both IGH loci seemed implausible if two distinct IGH loci were located on separate chromosomes as recombination would have required trans-chromosomal rearrangement. It was suspected then that the duplicated gene segments were in fact allelic differences between two haplotypes that were assembled incorrectly.

Duplications were also shown for the *IGHD* gene segments (Koti et al., 2010; 171, Shojaei et al., 2003; 172), where three *IGHD* clusters were identified in Holstein cattle using specific probes. These three clusters span 68 kb and contain 10 *IGHD* gene segments of which one is the ultra-long *IGHD* gene segment. They also identified two copies of their *IGHD4*, later re-named in the ARS-UCDv0.1 as *IGHD7*, was concluded as polymorphic and was not previously identified in the UMD3.1 assembly. Organisation of the *IGHD* gene segments and nucleotide identity between corresponding gene segments was high, suggesting that the expansion of the *IGHD* clusters arose through duplication events. The expanded *IGHD* raised further confusion regarding the organisation of the cattle IGH.

The genomic organisation of the cattle IGH was unable to be resolved by investigating available genome assemblies of closely related species. Cattle are classified within the family *Bovidae*, within the tribe *Bovini*; other species within this tribe include water buffalo (*Bubalus bubalis*), yak (*Bos grunniens*), and African buffalo (*Syncerus caffer*). The IGH in other members of the *Bovini* was either un-sequenced or unpublished. A recent African buffalo genome has been sequenced and assembled using Illumina sequencing alone, leading to the characterisation of nearly 20,000 genes, however the genome assembly has not yet been made publicly available (Glanzmann et al., 2016; 1).

Confusion around the apparent duplications in the cattle IGH locus has recently been resolved (Ma et al., 2016; 98). All the functional bovine IGH genes were shown to be located on the same chromosome, BTA 21, with in situ hybridisation. Using Sanger sequencing of BAC clones with a T-vector cloning system, they generated a 678 kb contiguous genomic sequence of the cattle heavy chain. Their structure of the IGH is 5'-*IGHVn-IGHDn-IGHJ6-IGHM1-*(*IGHDP-IGHV3-IGHDn*)3-*IGHJ6-IGHM2-IGHD-IGHG3-IGHG1-IGHG2-IGHE-IGHA*-3' with the cattle IGH containing a duplication of the *IGHD-IGHJ-IGHM-IGHD* and two other partial duplications of *IGHV-IGHD-IGHD*. In total 46 *IGHV*, 12 *IGHJ*, 23 *IGHD*

and 11 *IGHC* were identified, resolving previous confusion in the bovine antibody loci as the genomic organisation of the IGH locus contains internal duplications.

A recent cattle genome assembly (ARS-UCDv0.1) was generated using long read PacBio SMRT sequencing technology (Timothy P.L. Smith and Juan F. Medrano, unpublished). As reported by Ma et al (2016), the cattle IGH also exists as a single locus in this new assembly. Duplications in the IGH locus and the limited putatively functional *IGHV* repertoire are also observed in this new ARS-UCDv0.1 assembly. The organisation of the ARS-UCDv0.1 IGH locus was compared to UMD3.1 and the sequence from Ma et al (2016) to resolve structural differences. The UMD3.1 assembly is heavily disrupted and poorly assembled with large sequence gaps whilst the IGH sequence from Ma et al (2016) contains an additional internal duplication of the *IGHV-IGHD-IGHC* region compared to the ARS-UCDv0.1, as well as variations in the *IGHV* region. Using SMRT sequencing of the BAC clone (RP42-567N23) used by Ma et al (2016), the structure of the IGH in their assembly was confirmed. The ARS-UCDv0.1 assembly was used as a reference in order to assemble the African buffalo IGH. The African buffalo Illumina reads were provided by Glanzmann et al, South Africa Medical Research Council Centre for Tuberculosis Research (Glanzmann et al., 2016; 1). These African buffalo reads were mapped to the cattle IGH and subsequently used for both *de novo* assembly and for targeted assembly of individual IGH gene segments. Unlike cattle, the duplications of *IGHD*, *IGHJ*, and the constant region gene segments are absent in the African buffalo assemblies. Despite this, both cattle and our preliminary characterisation of the African buffalo show that both species have a limited germline IGH repertoire compared to humans and mice. This was confirmed with cattle RNA-seq data from Yfke Pasman at the University of Guelph and African buffalo antibody transcripts which were mapped to the cattle ARS-UCDv0.1 assembly and to individual buffalo gene segments respectively.

3.2 Methods

3.2.1 Cattle IGH genomic sequences

A cattle genome was *de novo* assembled from a single Hereford cow (L1 Dominette 01449) using long reads generated using the Pacific Biosciences RSII platform (ARS-UCDv0.1; Timothy P. L. Smith and Juan F. Medrano, unpublished), details of which are in Chapter 2, section 2.2.20. Scaffolds containing the antibody loci were identified within ARS-UCDv0.1 using the basic local alignment search tool (BLAST) (Altschul et al., 1990; 173) to look for common sequence motifs such as the leader sequence, W/F-G-X-G motif in *IGHJ* and the QVSL motif in FR3 of *IGHV.* Scaffolds containing *IGH* gene segments were then isolated from the assembly for further inspection.

The same Hereford individual (L1 Dominette 01449) sequenced for the long-read ARS-UCDv0.1 assembly was also used in part for sequencing and assembly of the publically available reference genome, UMD3.1. Therefore the IGH locus was identified in UMD3.1 for sequence comparison to ARS-UCDv0.1 using the same BLAST methods. All contigs identified as potentially containing IGH gene segments were downloaded from the Gbrowse archive within the Bovine Genome Database (Elsik et al., 2009; 159, 174).

IGH genomic sequences from two Holstein cattle were also obtained using genomic enrichment with sequence specific NimbleGen probes from an independent project being carried out in our group (Heimeier et al; unpublished). Holstein animals 252 and 200005 were targeted for their *IGHV* and *IGHD* regions with 6 – 10 kb DNA fragments, isolated from monocyte gDNA, pulled down and sequenced with PacBio and assembled using the SMRT analysis pipeline (smrtanalysis_2.3.0.140936.p4.150482) with HGap v3 (Chin et al., 2013; 175), into 55.5 kb, 33.9 kb, 31.6 kb and 21.2 kb contigs for animal 252 and 41.9 kb, 37.5 kb, 31.9 kb, 28 kb, 27.7 kb and 22.5kb contigs for animal 200005.

A contiguous genomic sequence of a third Holstein cattle IGH region was kindly shared with us by Li Ma and Yaofeng Zhao from the China University of Agriculture (Ma et al., 2016; 98). Briefly, they used IGH specific probes to query the Holstein bull BAC library Roswell Park Cancer Institute (RPCI)-42 and obtained positive clones from the Children's Hospital Oakland Research Institute's (CHORI) BACPAC resources program (CHORI, 2016; 156). A

consensus sequence was then generated from seven overlapping BAC clones containing the IGH locus: RP42-195P14, RP42-49A20, RP42-498B11, RP42-567N23, RP42-90B11, RP42-4E14 and INRA944D11. Next, a 678 kb sequence was generated using Sanger (chain-termination) sequencing of 1 kb BAC clone fragments in a T-vector system. The BAC clone RP42-567N23 contained IGH sequence that spanned the three duplications; *IGHD-IGHJ-IGHM-IGHD* and the two other partial duplications of *IGHV-IGHD-IGHD* described (Ma et al., 2016; 98). However, the number of reads used and coverage of the sequence generated was not specified.

We further obtained the BAC clone RP42-567N23 used by Ma et al (2016) to describe the three duplications in the cattle IGH. The BAC clone was sequenced by Timothy P.L. Smith at the Meat Animal Research Center, USDA-ARS (Clay Center, Nebraska) using long read PacBio SMRT sequencing and the MinIon sequencing platform (Oxford Nanopore Technologies, Ltd.). PacBio sequencing generated 1000 reads with an average length of 16 kb. The longest 40 x reads, 199 in total, were used to assemble the BAC clone with Canu v 1.5 (Koren et al., 2017; 176) into a single contig of 132 kb. The MinIon generated 206 reads which had an average length of 26.5 kb. These reads were also assembled with Canu into two contigs 132 kb and 52 kb.

```
#Assembly of the BAC clone using Canu

canu -p job_name -d Canu_567N23 genomeSize=k -pacbio-raw
     reads.fq

canu -p job_name -d Canu_567N23 genomeSize=k -nanopore-raw
     reads.fq
```

3.2.2 SNP calling in the BAC clone RP42-567N23 reads

Long reads sequenced from the BAC clone RP42-567N23 with PacBio were aligned against the Canu assembly by Richard Borne, a PhD student in our Immunogenetics group. A SNP pile-up was generated using SAMtools mpileup with confidence scoring at each SNP

position. Sequenza script Pileup2acgt produced ratios of each base at every position in the region of interest. SNP patterns were then analysed in Microsoft Excel.

3.2.3 Reference-based assembly of the African buffalo IGH locus

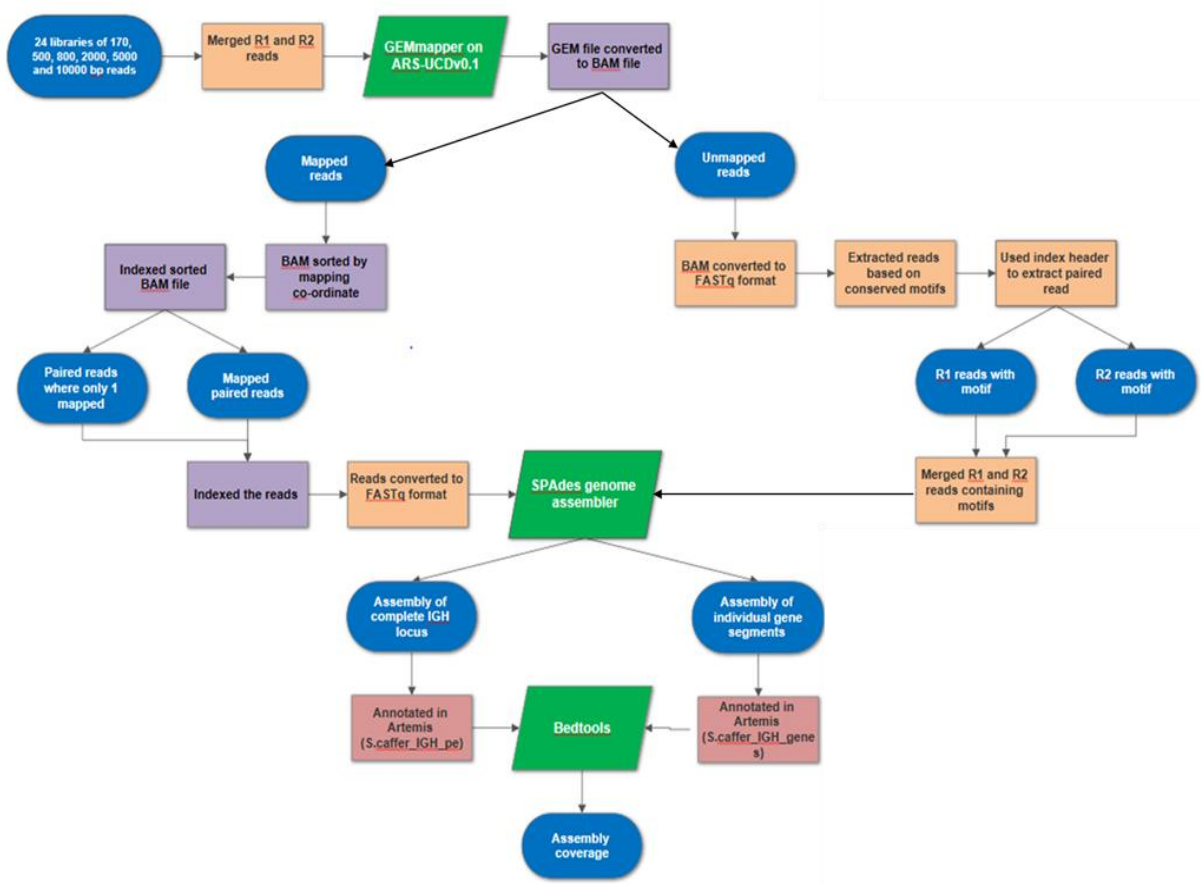


Figure 3.1: Schematic of the African buffalo IGH assembly pipeline. Processing steps are indicated by the rectangles, output files by oval shapes and the programmes used by parallelograms. Data in FASTQ format is indicated by orange, purple indicates the reads are in a BAM file and pink indicates data is in FASTA format. Illumina sequencing libraries with various insert sizes were sequenced from a single African buffalo by the South Africa Medical Research Council Centre for Tuberculosis Research (Glanzmann et al 2016). The forward and reverse reads were merged and mapped onto the cattle IGH in the ARS-UCDv0.1 assembly using the GEMmapper (Marco-Sola et al., 2012; 177). Mapped reads were filtered and indexed based on both reads in a pair mapping or only one of each pair. These mapped files were converted to FASTQ for *de novo* assembly of the buffalo IGH in SPAdes (Bankevich et al., 2012; 178). Simultaneously, unmapped reads with antibody motifs were extracted and used with the mapped reads for assembly of individual gene segments. Gene assemblies were annotated in Artemis and their coverage estimated using Bedtools (Quinlan and Hall, 2010; 179).

Genome sequence reads from a single male African buffalo (*Syncerus caffer*) from the Kruger National Park (South Africa) were used to assemble the first African buffalo genome (Glanzmann et al., 2016; 1). Twelve mate pair libraries were constructed with either 170 bp, 500 bp, 800 bp, 2 kb, 5 kb or 10 kb insert sizes from gDNA derived from whole blood and sequenced at 60-fold coverage using Illumina Hi-seq 2000, yielding 242.39 Gbp sequence data (89.78 x coverage) after low quality filtration. These reads were then kindly shared with us for assembly of the African buffalo antibody loci.

The African buffalo IGH region was assembled using single and paired reads which mapped to the cattle IGH locus in ARS-UCDv0.1 (figure 3.1). FASTQ libraries from the African buffalo genome sequencing project (Glanzmann et al., 2016; 1) were merged into a single file containing forward and reverse reads and reads were mapped using an unreleased experimental version of the Genome Multitool Mapper v3.5-19-g1d79-dirty-release (GEMmapper) (Marco-Sola et al., 2012; 177) with 4% and 10% maximum alignment error. Using SAMtools v1.2 (Li et al., 2009; 180), the output SAM file was converted to a BAM file and indexed. The BAM file was then filtered on read mapping status and two BAM files were generated containing pairs where both reads mapped or only one read mapped of the pair. These files were converted to FASTQ format using SAMtools and *de novo* assembled with SPAdes 3.10.1 (Bankevich et al., 2012; 178) with the IGH mapped reads as single reads. A second assembly was then constructed using only reads where the forward and reverse reads mapped to incorporate paired read information for more accurate *de novo* assembly. Each African buffalo IGH assembly was evaluated with the Quality Assessment Tool for Genome Assemblies (QUAST) (Gurevich et al., 2013; 181).

Individual gene segments in the African buffalo IGH locus were assembled using reads mapped to ARS-UCDv0.1 where either both reads mapped or only one read mapped of a pair and unmapped reads that contained IGH gene segment motifs. Unmapped reads from the GEMmapper were converted to FASTQ and reads matching the IGH locus were extracted by known motifs including the QVSL motif in FR3 of IGH. The corresponding mate pairs were then extracted using the index header and all the reads were merged into a single file. These unmapped merged reads were combined with the mapped reads to *de novo* assemble individual genes with SPAdes 3.10.1 with 4% and 10% maximum alignment error against the IGH locus in ARS-UCDv0.1. Reads were also mapped to the *IGHJ* regions in the ARS-UCDv0.1 assembly with a 1% alignment error to attempt differentiation of duplications in the

African buffalo IGH. The assembled contigs were annotated in Artemis (Rutherford et al., 2000; 182) and concatenated into one sequence, *S. caffer*_IGH.

Script for the African buffalo IGH assembly pipeline:

```
#Merged forward and reverse Illumina read libraries
cat *R1.fastq > merged_R1.fastq
cat *R2.fastq > merged_R2.fastq

#Mapped reads to the reference genome
gem3-indexer -i reference.fa -t 24
gem3-mapper -I reference.gem -1 merged_R1.fastq -2
merged_R2.fastq -p -F sam -t 24

#Converted SAM to BAM file, sorted and indexed
samtools view -Sb reference.sam > reference.bam
samtools sort -o reference.sorted.bam -O bam -T pref -@ 24
     reference.bam
samtools index reference.sorted.bam

#Filtered BAM file for mapped paired reads and pairs with only
one mapped read
samtools view -b -F 4 -f 8 merged.bam > onlyThisEndMapped.bam
samtools view -b -F12 merged.bam > bothEndsMapped.bam

#Sorted and indexed filtered BAM files
samtools sort -o  onlyThisEndMapped.bam -O bam -T pref -@ 24
     onlyThisEndMapped.sorted.bam
samtools sort -o   bothEndsMapped.bam -O bam -T pref -@ 24
     bothEndsMapped.bam.sorted.bam
samtools index onlyThisEndMapped.sorted.bam
samtools index bothEndsMapped.bam.sorted.bam


#Converted reads to FASTQ format and merged
samtools bam2fq onlyThisEndMapped.sorted.bam | seqtk seq >
     onlyThisEndMapped .map.fastq
samtools bam2fq bothEndsMapped.sorted.bam | seqtk seq >
     bothEndsMapped.sorted.map.fastq
cat $onlyThisEndMapped .map.fastq
     $bothEndsMapped.sorted.map.fastq > mapped.merged.fastq

#Individual gene assembly
samtools view -b -h -F 4 -f 8 onlyThisEndMapped.sorted.bam
     "$coords" > ${region_name}_1map.bam
samtools view -b -h bothEndsMapped.sorted.bam "$coords" >
     ${region_name}_2map.bam
```

```
#SPAdes de novo assembly
echo -en '#!' '/bin/bash\n/data/borne/programs/SPAdes-3.10.0-
     Linux/bin/spades.py' -k 21,33,55,77 --careful -s
mapped.merged.fastq -o ./${region_name} -t 24 | sbatch -J
     ${region_name} -c 24

#Unmapped reads converted to FASTQ
samtools view -b -f 4 -@ 24 merged.bam > unmapped.bam
samtools bam2fq unmapped.bam | seqtk seq > unmapped.fastq

#Separate forward and reverse reads
cat unmapped.fastq | grep -A 3 -E '^@.*/1' | sed 's/--//g' |
     grep -v '^$' > unmapped_R1.fastq
cat unmapped.fastq | grep -A 3 -E '^@.*/2' | sed 's/--//g' |
     grep -v '^$' > unmapped_R2.fastq

#Unmapped reads extracted by known sequence motif
grep -i -A 2 -B 1 "agcagcgtgaca" unmapped_R1.fastq >
     unmapped_R1_motif.fastq
grep -i -A 2 -B 1 "agcagcgtgaca" unmapped_R2.fastq >
     unmapped_R2_motif.fastq

#Found the read pair header of unmapped reads with antibody
motifs
cat unmapped_R1_motif.fastq | sed 's/--//g' | grep -v '^$' >
     unmapped_R1_motif.filtered.fastq
cat unmapped_R2_motif.fastq | sed 's/--//g' | grep -v '^$' >
     unmapped_R2_motif.filtered.fastq
cat unmapped_R1_motif.filtered.fastq | grep -E '^@.*/1$' | sed
     's|/1|/2|g' | sed 's/@//g' > IDs_R1.txt
cat unmapped_R2_motif.filtered.fastq | grep -E '^@.*/2$' | sed
     's|/2|/1|g' | sed 's/@//g' > IDs_R2.txt

#Used read pair headers to extract unmapped reads
~/programs/bbmap/filterbyname.sh in=merged_R2.fastq
     out=unmapped_R2_mates.fastq names=IDs_R1.txt include=t
~/programs/bbmap/filterbyname.sh in=merged_R1.fastq
     out=unmapped_R1_mates.fastq names=IDs_R2.txt include=t

#Merged reads with antibody motif with their read pair
cat unmapped_R1_motif.filtered.fastq unmapped_R1_mates.fastq >
     unmapped_R1_assembly.fastq
cat unmapped_R2_motif.filtered.fastq unmapped_R2_mates.fastq >
     unmapped_R2_assembly.fastq

#Used mapped and unmapped reads with antibody motifs in
assembly
echo -en '#!' '/bin/bash\n/data/borne/programs/SPAdes-3.10.0-
     Linux/bin/spades.py' -k 21,33,55,77 --careful -s
${region_name}_merged.fastq -o ./unmapped_assembly -t 24 |
     sbatch -J SPAdes -c 24
```

Read coverage across the single read and paired read African buffalo assemblies were calculated from the number of reads (N), the average read length (L) and the genomic region length (G), with the formula **N * L/G**. African buffalo FASTQ reads were mapped to the paired end assembly S.caffer_IGH_pe10 and the individual assemblies of gene segments concatenated together (S.caffer_IGH_genes) using the Genome Multitool mapper with 10% maximum alignment error. Output SAM files were converted to BAM and sorted. The depth of feature coverage for each base on each contig in the genome assemblies was computed with BEDTools v2.26.0 (Quinlan and Hall, 2010; 179). Read counts were converted to coverage estimates and plotted in Microsoft excel.

Script for the African buffalo IGH assembly coverage:

```
#Mapped reads to the assembled IGH and concatenated IGH genes
gem3-indexer -i reference.fa -t 24
gem3-mapper -I reference.gem -1 merged_R1.fastq -2
     merged_R2.fastq -p -F sam -t 24

#converted SAM to BAM and sorted
samtools view -Sb scaffer_mapping.sam > scaffer_mapping.bam
samtools sort aln.bam aln.sorted.bam

#Report per base coverage
samtools view -b aln.sorted.bam | bedtools genomecov -ibam
     stdin -g reference.fa  -d  > coverage.bedtools.info
```

3.2.4 SNP calling on the African buffalo *IGHJ* assembly

Evidence of internal duplications in the African buffalo IGH was determined by SNP calling of African buffalo genomic reads mapped to the African buffalo *IGHJ* locus. The African buffalo genome reads (Glanzmann et al., 2016; 1) were first mapped to the cattle *IGHJ* in ARS-UCDv0.1, as outlined in section 3.2.3, with a 4% alignment error and a default mapping quality scoring of 20, and then subsequently *de novo* assembled with SPAdes into the African buffalo *IGHJ* region. The genome reads from the African buffalo were then mapped to this assembled *IGHJ* region with the GEMmapper at 4% alignment error and a SNP pile up was

generated using SAMtools mpileup (Borne, unpublished). SNPs were subsequently analysed in Microsoft Excel.

3.2.5 Characterisation of the cattle and African buffalo IGH locus

Genomic sequences for the Hereford cattle assemblies, ARS-UCDv0.1 and UMD3.1, Holstein cattle IGH sequences from Ma et al (2016), the Holstein PacBio genomic enrichment data, the African buffalo assembly, and individual gene assemblies were manually annotated using Artemis v13.0 (Rutherford et al., 2000; 182). Immunoglobulin gene segments were identified using both BLAST and the NCBI conserved domain database (2016; 183). Annotated features of each putative gene segment were: the leader, octamer (ATTTGCAT), 5' (GT) and 3' (AG) splice sites, the recombination signal sequence (RS) heptamer (CACAGTG or CACTGTG), RS nonamer (ACAAAAACC), RS spacer (23 bp/12 bp), and the conserved amino acid residues C23, W41, hydrophobic 89 and C104.

3.2.6 Structural comparison of the cattle ARS assembly

Recurrence plots were generated using DOTTER v4.44.1 (Sonnhammer and Durbin, 1995; 184) for sequence comparisons between the genome assemblies. The structure of the ARS-UCDv0.1 assembly was compared to the UMD3.1 scaffolds, the Holstein PacBio capture data, the Holstein IGH sequence from Ma et al (2016), the BAC clone RP42-567N23 sequenced with PacBio or Oxford Nanopore and the African buffalo paired-end assembly, the S.caffer_IGH. The sliding window was set to 200 bp within DOTTER.

```
dotter -b in.fas[y-axis] in.fas[x-axis] out.txt
dotter -l in.fas[y-axis] in.fas[x-axis] out.txt
```

3.2.7 Nomenclature of IGH genes

IGHV gene segments were named in the ARS-UCDv0.1 cattle assembly according to IMGT nomenclature (Lefranc et al., 2003; 185) in order from their starting position proximal to the constant region. African buffalo *IGHV* were numbered according to their phylogenetic sub-grouping due to the lack of a complete genome for annotation. The *IGHV* in cattle and buffalo were also named in brackets according to their sub-group or clan for consistency with other species. The sub-group number was determined by comparison of sequences with BLAST to the IMGT database using a 75% identity threshold to known *IGHV* in other species, including human and mouse. When the sub-group was undetermined, *IGHV* gene segments were named according to their higher order clan, as designated by roman numerals.

Gene segments were considered functional if they possessed canonical initiation codons (ATG), contained conserved amino acid residues (C23, W41, hydrophobic 89 and C104), and were in-frame without truncations or premature stop codons. Gene segments were putatively determined to be pseudogenes if they contained truncations, stop codons, frameshifts or a defective initiation codon. Gene segments were defined as open reading frame (ORF) if they were in-frame but missing conserved amino acid residues.

3.2.8 Phylogenetic analysis of IGH gene segments

Sequences were aligned using a global alignment strategy in the MAFFT package, version 6.603b (Katoh et al., 2002; 186) and visually confirmed and edited as necessary using Bioedit v7.2.5 (Ibis Biosciences, (Hall, 1999; 187)). Phylogenetic analysis of gene segments was calculated in MEGA 6.0 (Tamura et al., 2013; 188) using maximum likelihood based on the Tamura and Nei model (Tamura and Nei, 1993; 189) and the partial deletion method using a 95% cut off and 1000 bootstrap iterations. *IGHV* and *IGHD* in cattle and African buffalo were aligned to observe the expansion/contraction of these gene segments between the two species.

3.2.9 Transcriptional analysis of *IGHC* isotypes and *IGHV* gene segments in cattle and buffalo using RNA-seq

The expression of the *IGHC* and *IGHV* was investigated in cattle and African buffalo. Yfke Pasman at the University of Guelph generated whole transcriptome 150 bp paired-end reads from the PBMC of three female Holstein cattle using Illumina Hi-seq (Pasman et al., 2017; 190) and kindly shared them with us for the expression analysis. African buffalo IgM and IgG antibody transcripts were isolated from two animals, A7 and A11, at Day 0 and then at Day 8 and 14 post-challenge with SAT1 FMDV. The African buffalo IgM and IgG transcripts were sequenced with Illumina, 2 x 300 bp and the IGH reads isolated, outlined in Chapter 5, section 5.3.8.1. Cattle reads were mapped onto the *IGHV* in the ARS-UCDv0.1 assembly and the African buffalo reads were mapped to the concatenated assembly of the *de novo* assembled genes (S.caffer_IGH), with GEMtools RNA pipeline (Marco-Sola et al., 2012; 177). Phred quality scoring of 33 was used for alignment calculations with 6% mismatch identity. Weighted counts were produced based on the assembly annotation and the frequency of mapping. Expression percentiles were calculated based on the total number of weighted counts of each animal and displayed graphically.

Script for RNA-seq of cattle and African buffalo IGH:

```
#index the reference
gemtools index -i (reference.fas) -t 24 ibatch

#index the gene annotation file
gemtools t-index -i (genomeindex.gem) -a (annotation.gtf) -t
     24

#map RNA reads to the reference
gemtools rna-pipeline -i (genomeindex.gem) -a (annotation.gtf)
     -f (ForwardRNAreads.fastq) (ReverseRNAreads.fastq) -o
     (outputfile) --no-filtered -q (quality encoding) -t
     (number of threads)
```

The mappability of the *IGHV* transcribed in the Holstein animals and found in the RNA-seq of the assembled African buffalo *IGHV* was calculated in order to define the uniqueness of the gene segments and hence the likelihood of producing unique or weighted counts with RNA-seq. An exhaustive alignment of 150 bp lengths originating at every base positon of each gene sequence using GEMmapper was carried out, allowing for 4% mismatches between the reads (Borne, unpublished). Resulting mappability scores were plotted in Microsoft Excel.

3.3 Results

3.3.1 The structure of the cattle IGH in the reference genome assembly, UMD3.1

The organisation of the IGH in the previously annotated UMD3.1 assembly (Niku et al., 2012; 167) is heavily disrupted and poorly assembled with a considerable amount of sequence information absent. The IGH locus in UMD3.1 contains 26 contigs which are located on BTA7, BTA8, BTA20, BTA21 or are unplaced (figure 3.2). A total of 31 *IGHV* gene segments were previously characterised and named according to how they clustered in sub-groups within a phylogenetic tree (Niku et al., 2012; 167). In UMD3.1, each of the duplicated *IGHM* is present, along with three regions of *IGHD* gene segments but only one *IGHJ* region, suggesting duplications in the cattle IGH are assembled to separate chromosomes. In reality, this is highly unlikely to be true as IGH chains would need to be generated using trans-chromosomal rearrangement, which is conceptually illogical and has never been shown to occur in any species. Improvements in the cattle IGH assembly awaited the construction of a genome sequenced with long reads.

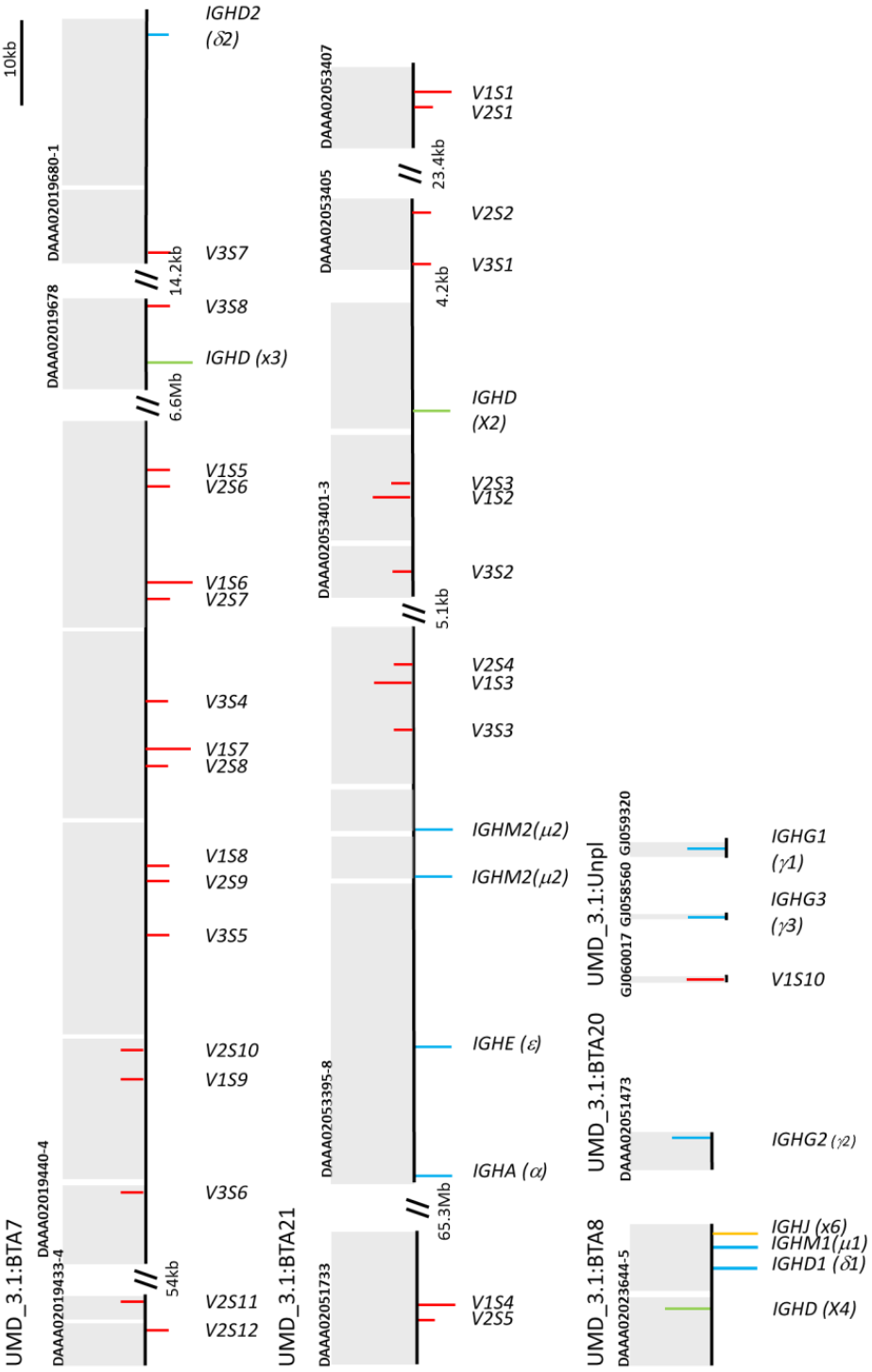


Figure 3.2: Schematic organisation of the IGH locus in the UMD3.1 Hereford genome assembly. Contigs are ordered according to how they were assembled in the genome with gaps between contigs indicated by the grey shaded boxes and large gaps shown by dashes with the distance shown. *IGHV* are indicated by red lines, *IGHC* exons by blue lines and clusters of *IGHJ* and *IGHD* by yellow and green respectively. Putatively functional gene segments are indicated with a long projection line and pseudogenes have a short projection line. The genes on the positive strand are projected above and those on the negative strand are projected below. Scale bar: 10 kb.

3.3.2 The structure of the cattle IGH in the long-read assembly, ARS-UCDv0.1

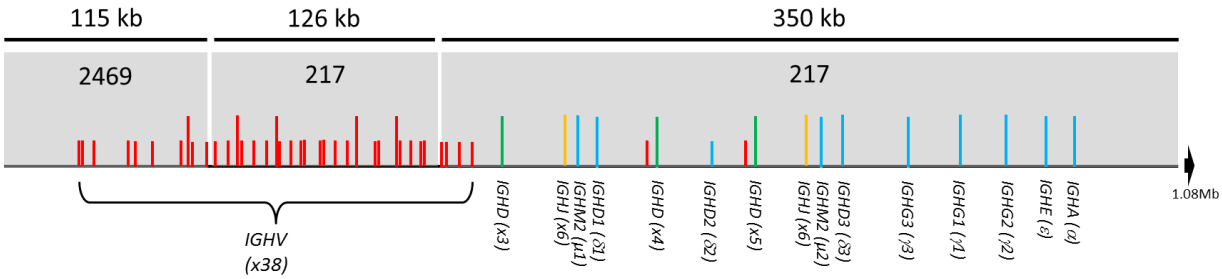
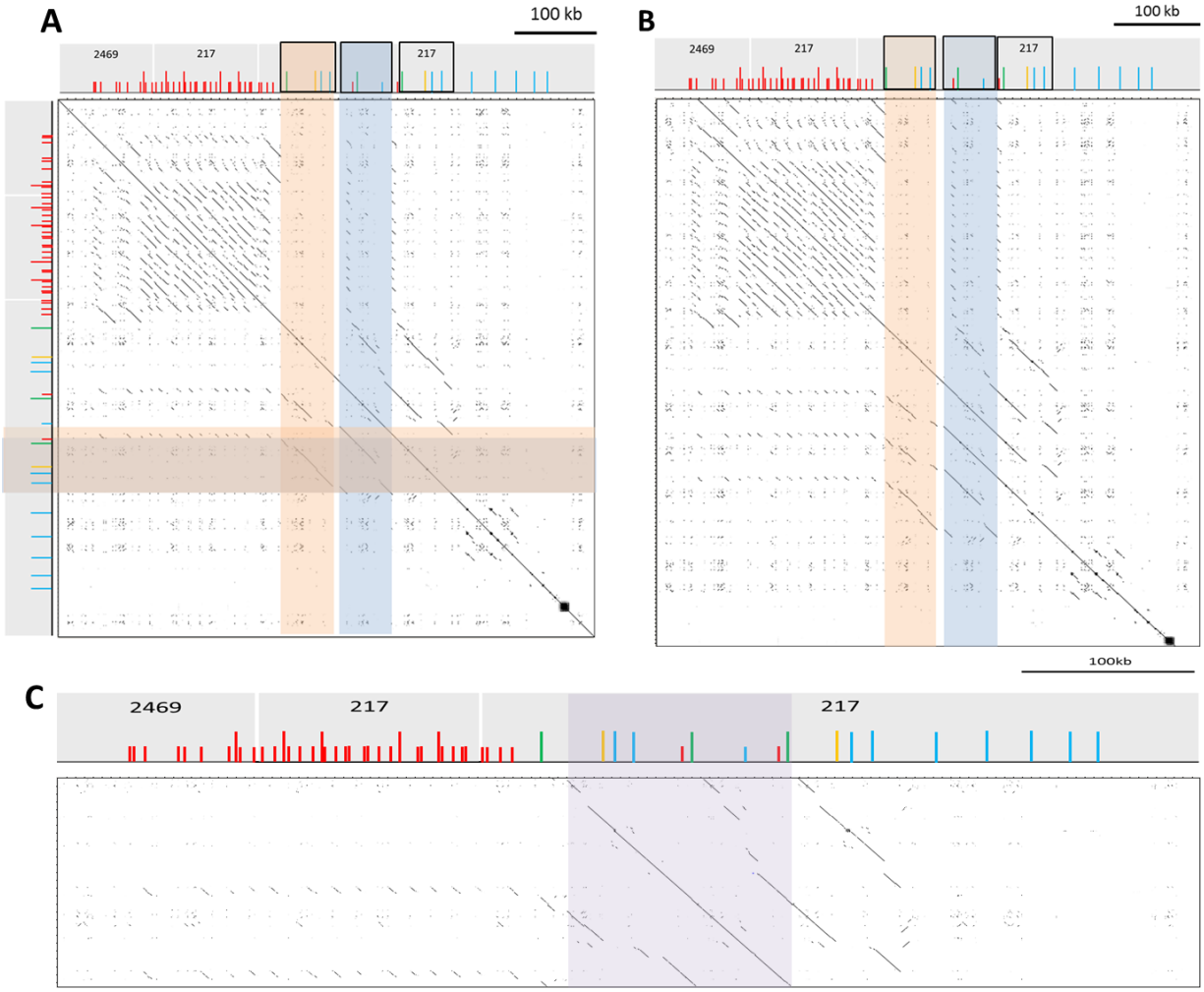


Figure 3.3: Schematic organisation of the IGH in the long read ARS-UCDv0.1 PacBio assembly. The IGH spans approximately 600 kb on two scaffolds (IGH_217 and IGH_2469) consisting of three contigs. Putatively functional gene segments are indicated with a long projection line and pseudogenes have a short projection line. Red lines indicate an *IGHV* gene segment. The clusters of *IGHJ* and *IGHD* gene segments, yellow and green respectively, are represented by single projection lines which are long in length if the cluster contains one or more putatively functional gene segments. The second contig of scaffold 217 extends beyond the IGH locus 1.08 Mb from the 3' end, as indicated by the arrow.

The cattle IGH in the long-read genome assembly ARS-UCDv0.1 is found on two scaffolds (IGH_217 and IGH_2469), consisting of three contigs and spanning approximately 600 kb (figure 3.3). Spanning ~210 kb at the 3' end of the locus is the highly repetitive *IGHV* region, containing a total of 37 *IGHV* gene segments. Duplications in the constant region have resulted in a heavy chain structure that is different from the expected *IGHVn-IGHDn-IGHJn-IGHC* gene segment order. A ~63 kb duplication of the *IGHV-IGHDn-IGHJn-IGHM-IGHD* resulted in a second cluster of these gene segments with a nucleotide identity of 97%. A second 52 kb duplication of *IGHV*(I)-*IGHDn-IGHD* occurs, containing a truncated third constant region *IGHD*, with a 96% nucleotide sequence identity to the functional *IGHD*. These duplications are observed in a recurrence plot of the ARS-UCDv0.1 assembly against itself (Fig 4A). The resulting heavy chain locus is arranged 5'-$IGHV_{38}$-$IGHD_3$-$IGHJ_6$-*IGHM1-IGHD1-*$IGHV_1$-$IGHD_4$*- IGHD2-* $IGHV_1$*-* $IGHD_5$-$IGHJ_6$-*IGHM2-IGHD3-IGHG3-IGHG1-IGHG2-IGHE-IGHA*-3' in the ARS-UCDv0.1 assembly.

Figure 3.4: Recurrence plot of the IGH locus in the PacBio ARS-UCDv0.1 assembly aligned against itself (A) and the Ma et al (2016) assembly (B). Internal duplications in the sequence are highlighted showing the duplication of region *IGHV-IGHD-IGHJ-IGHM-IGHD* (orange) and duplications of *IGHV-IGHD-IGHD* (blue). The assembly from Ma et al (2016) contains an additional *IGHV-IGHD-IGHD* (blue) duplication. The BAC clone RP42-567N23 used by Ma et al (2016) to sequence and assemble the duplications in the IGH was obtained and sequenced by PacBio. The long-read assembly of RP42-567N23 was aligned against ARS-UCDv0.1 (C), which appeared to confirm the structure of the IGH in the ARS-UCDv0.1 assembly.



3.3.3 Structural comparison of the IGH in ARS-UCDv0.1 and the sequence from Ma et al (2016)

The published 678 kb contiguous IGH sequence generated by Ma et al (2016) was the first to show duplications in the cattle IGH. However, whilst a duplication of the *IGHVn-IGHDn-IGHJn-IGHM-IGHD* region and a partial duplication of *IGHVn-IGHDn-IGHD* are each seen once in the ARS-UCDv0.1 assembly, in the assembly published by Ma et al (2016), a duplication of the *IGHDn-IGHJn-IGHM-IGHD* region occurs once and a duplication of *IGHVn-IGHDn-IGHD* occurs twice (figure 3.4B). Nucleotide sequence identity between the two *IGHVn-IGHDn-IGHD* duplications from Ma et al (2016) is 99%, with identity between the corresponding gene segments in both regions 97-98%. Five clusters of *IGHV* are arranged in a different order between the Ma et al (2016) assembly and ARS-UCDv0.1 (figure 3.5). Another duplication is found in the *IGHV* region of Ma et al (2016). This duplication is 20 kb in length and contains three *IGHV* gene segments. The IGH locus in Ma et al (2016) also does not extend as far as ARS-UCDv0.1. As a result, *IGHV39* and *IGHV40*, are missing from the Ma et al (2016) assembly. Ma et al (2016) describes a total of 46 *IGHV* gene segments, 12 of which are putatively functional. Eight *IGHV* genes from this assembly are not present in ARS-UCDv0.1. These five genes have 96-100% nucleotide identity to other *IGHV*, suggesting either assembly error or haplotypic variation between the two assemblies. The remaining 38 *IGHV* in Ma et al (2016) are the top BLAST hits with 97-100% nucleotide sequence identity to the corresponding gene segments and group together phylogenetically to the ARS-UCDv0.1, indicating the majority of *IGHV* between the two assemblies are nearly identical (Appendix Table 2).

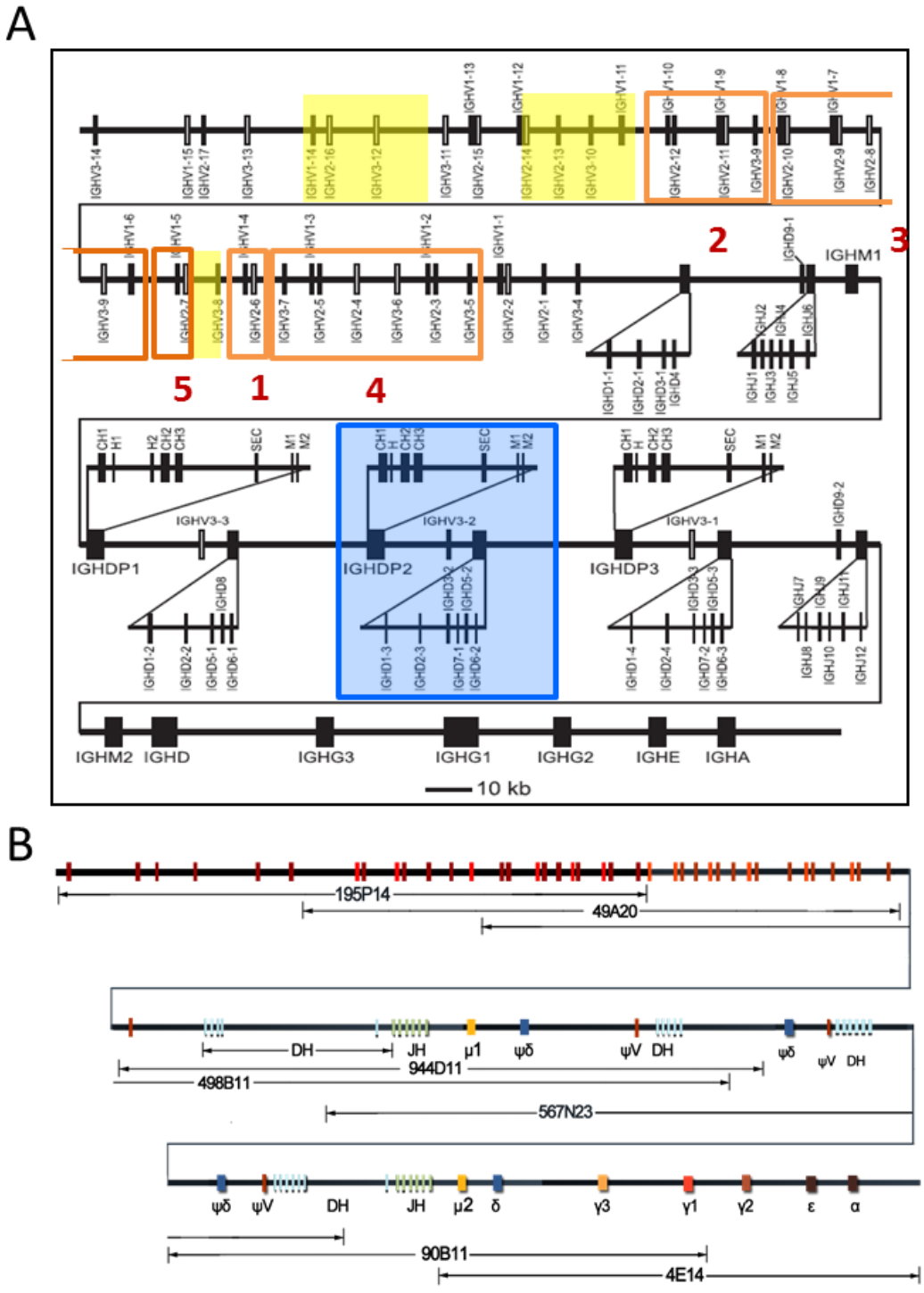


Figure 3.5: Organisation of the Holstein IGH locus assembled from Sanger sequencing of BAC clones in the CHORI BACPAC resources program (Ma et al., 2016; 98, CHORI, 2016; 156). The numbering of the gene segments is indicated in (A). The area shaded within the blue box indicates the additional duplication of *IGHVn-IGHDn-IGHD*, not seen in the ARS-UCDv0.1. The *IGHV* within the yellow shading represents gene segments that are absent in ARS-UCDv0.1 and which do not occur in the Holstein capture data. The orange boxes are numbered according to their arrangement in ARS-UCDv0.1 beginning from the 3' *IGHC* proximal end. Image adapted from (Ma et al., 2016; 98). The BAC clones used to construct the IGH are indicated in (B).

3.3.4 Long read assembly of the BAC clone RP42-567N23 used by Li Ma et al

To determine if the additional duplication of *IGHVn-IGHDn-IGHD* present in the assembly from Ma et al (2016) was correct, we obtained the BAC clone, RP42-567N23, used to assemble the region spanning their internal duplications. This was sequenced using the PacBio and Oxford Nanopore third generation long read sequencing technologies by Timothy P.L. Smith at the Meat Animal Research Center, USDA-ARS in Clay Center, Nebraska.

PacBio SMRT sequencing generated 1000 sequences with an average length of 16 kb. The longest 40 x reads were assembled using Canu into a single 132 kb contig. Nucleotide identity of the Canu assembly compared to the clone assembly from Ma et al (2016) was 99% with a total of 301 SNPs between them. Oxford Nanopore MinIon sequencing generated 206 reads with an average length of 26.5 kb. The reads assembled into two contigs of 132 kb and 52 kb with a nucleotide identity of 97% to the assembly from Ma et al (2016). The additional duplication of the *IGHVn-IGHDn-IGHD* described by Ma et al (2016) however was absent in both the PacBio and the Oxford Nanopore assemblies of the same BAC clone.

3.3.5 SNP calling in the SMRT sequence reads of the BAC clone RP42-567N23 used by Ma et al (2016)

A SNP pile up was generated for the BAC clone RP42-567N23 using PacBio reads aligned against the Canu assembly. The total number of SNPs in the BAC clone was 613 across the 132 kb contig (0.46%). The total number of SNPs in the 32 kb duplicated *IGHVn-IGHDn-IGHD* was 100 (0.31%) suggesting the second duplication described by Ma et al (2016) was random error in the sequence that they assembled into two regions. However, on closer inspection of the SNPs in the reads, 81 SNPs differentiating the two duplications described by Ma et al (2016) were identified in the PacBio reads. In 58 of the 81 SNPs described, roughly half of the reads contained the SNP for the first *IGHVn-IGHDn-IGHD* duplication and the other half of the reads contained the SNP for the second duplication (Table 1). Considering the reads were sequenced from a BAC clone they would not contain haplotypic variants. The SNP pattern suggests both *IGHVn-IGHDn-IGHD* duplications described by Ma

et al (2016) exist and that assembly of the PacBio reads failed to differentiate between the duplications.

LIMa 2	LIMa 1	BAC Pos	BAC base	Read Depth	A	C	G
A	G	101071	A	466	223	2	241
C	G	107988	G	193	0	109	81
G	A	108173	A	316	208	2	105
T	C	108968	T	335	0	138	0
G	A	109008	G	284	26	4	243
G	T	109008	G	230	0	0	219
G	A	109112	G	447	204	4	228
A	G	109558	A	355	207	0	147
G	A	109581	G	246	85	2	159
C	T	110077	C	380	0	209	4
C	T	110088	C	161	0	14	0
C	T	110097	T	154	0	5	0
G	A	110216	G	491	246	0	235
G	C	110845	G	224	0	2	205
T	C	110468	T	226	0	2	1
T	A	110626	T	348	134	1	0
G	T	110627	G	279	2	1	143
A	G	111194	G	81	0	2	78
A	G	111395	A	383	204	0	177
A	G	112362	A	241	232	0	9
A	G	112488	A	402	224	1	169
G	A	112537	G	367	161	1	203
C	G	112765	G	399	1	162	214
T	C	113063	C	374	2	256	1
C	A	113843	A	426	257	150	0
A	G	114024	G	414	193	2	213
G	A	114395	A	301	275	0	23
G	T	114397	T	160	0	0	19
G	A	114411	A	196	167	0	25
C	G	114679	G	378	0	170	200
C	A	115184	A	342	165	171	0
G	A	115186	G	385	92	0	277
A	C	115211	C	111	17	94	0
T	G	115737	G	211	0	1	23
C	T	115938	C	280	1	135	1
A	G	117359	A	307	142	8	157
A	G	118537	A	210	19	0	191
A	G	119948	G	464	108	0	336
C	T	120283	C	592	2	366	0
G	A	120833	G	453	287	1	159
C	T	120516	C	479	1	347	1
G	C	120713	G	577	0	254	307
G	T	121010	G	637	1	0	375
G	C	121288	C	421	0	155	265
G	A	121433	G	431	85	1	327
G	A	121646	A	545	274	1	263
A	C	121832	A	456	296	145	0
G	A	122549	A	765	530	38	42
T	C	122552	T	326	19	100	0
A	G	124679	G	332	93	2	218
G	A	124927	A	398	243	0	144
A	G	125058	G	303	108	0	192
A	G	125527	G	298	178	0	119
G	A	125636	G	471	214	3	241
A	G	125718	G	399	150	0	243
A	G	125743	G	327	143	1	182
T	G	126691	T	346	1	4	117
G	A	126696	A	211	204	1	6
C	T	126937	C	363	2	212	0
C	T	127829	C	372	3	250	0
A	G	128007	A	493	339	1	148
G	A	128075	G	747	381	1	359
T	A	128295	A	578	459	2	4
G	A	128545	A	219	53	4	159
G	C	128561	G	707	67	94	515
T	C	128689	C	528	2	148	1
T	C	128885	C	519	0	415	5
A	G	128896	A	529	142	0	374
G	T	129230	G	73	0	1	5
A	G	129232	A	251	131	0	110
A	G	129291	A	204	126	0	71
T	A	129598	A	246	142	2	0
C	T	129814	C	277	0	160	0
C	A	129972	A	102	27	72	0
A	C	130401	A	159	74	83	0
T	C	131581	C	131	1	63	0
C	A	131679	A	50	29	21	0
T	C	131764	C	108	1	51	1
C	T	131932	C	77	0	32	3
T	C	131982	C	136	2	73	0
A	G	132126	G	36	33	0	3

Table 3.1: The BAC clone RP42-567N23, used by Ma et al (2016) to sequence the duplications in the cattle IGH, was subsequently sequenced with PacBio. SNP pile up of the sequenced reads was performed by aligning reads to the Canu_567N23 assembly of the BAC clone. Positions in the Canu_567N23 assembly where a SNP exists between the two *IGHVn-IGHDn-IGHD* duplications described by Ma et al (2016) are shown. Where that SNP occurs in ~50% or more of the reads, the SNP is highlighted in pink and the read count in orange.

3.3.6 Genome enrichment and sequencing of IGH in Holstein cattle supports the ARS-UCDv0.1 structure

The structure of the *IGHV* region in the ARS-UCDv0.1 assembly is supported by genome enrichment data from two Holstein animals. Contigs from animal 252 span a total of 142.2 kb and 189.5 kb from animal 200005 and align against the *IGHV* region (99% nucleotide sequence identity) and from *IGHV1* across the *IGHD* region to *IGHD5-2* (98% nucleotide sequence identity) (Figure 3.6). In animal 252, contigs 5 and 1237 overlap by 1.5 kb and confirms the 20 kb *IGHV* region duplication in the Ma et al (2016) assembly is absent, suggesting the numbering of the *IGHV* in ARS-UCDv0.1 is correct. However similar to the second duplication of the *IGHV-IGHD-IGHD* region this could be due to mis-assembly of the PacBio reads across closely related duplicated sequences. In animal 20005, a 37.5 kb contig spans the gap between the scaffolds in the ARS-UCDv0.1 which contains a highly repetitive 1167 bp sequence with polypyrimidine tracts. This contig does not contain additional *IGHV* gene segments; the break between the scaffolds does not obscure any additional *IGHV*.

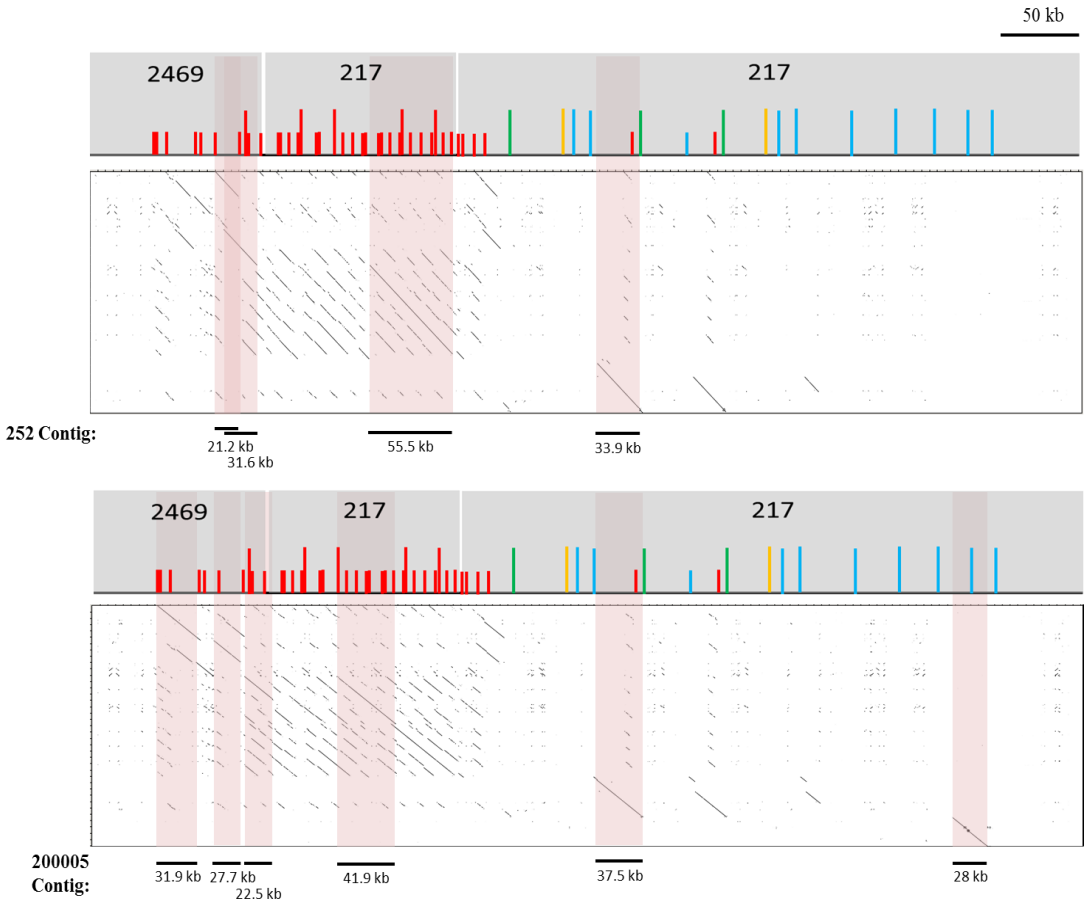


Figure 3.6: Recurrence plot sequence identity comparisons of the IGH locus in ARS-UCDv0.1 (x-axis) aligned to the PacBio contigs assembled from the genome enrichment data of Holstein animals 252 and 200005 (y-axis). The ARS-UCDv0.1 schematic is displayed over the x-axis. Shading indicates the region that the capture data overlaps, along with the black bars underneath. Scale bar: 50kb.

3.3.7 The structure and organisation of the African buffalo IGH

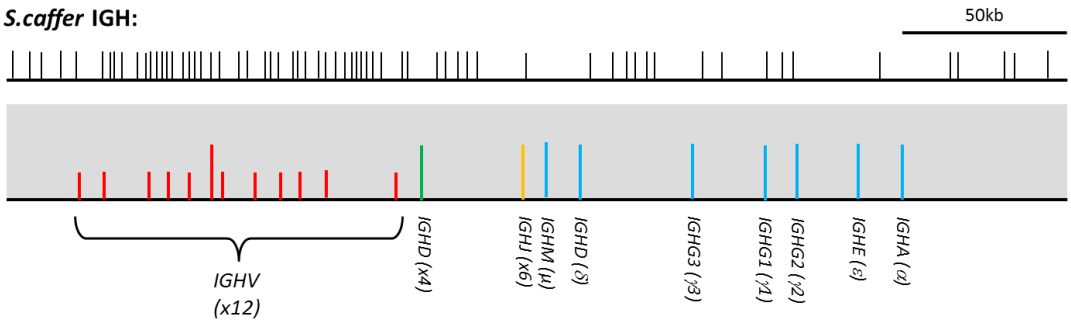


Figure 3.7: Schematic organisation of the African buffalo IGH locus assembled using paired-end reads. Reads were mapped to the cattle IGH locus in ARS-UCDv0.1 and *de novo* assembled using SPAdes (Bankevich et al., 2012; 178). The assembled region spans 324 kb, consisting of 70 contigs indicated by the black projection lines above the annotation. Putatively functional gene segments are indicated with a long projection line and pseudogenes have a short projection line. Red lines indicate an *IGHV* gene segment, the clusters of *IGHJ* and *IGHD* gene segments, yellow and green respectively, are represented by single projection lines which are long in length if the cluster contains one or more putatively functional gene segments. Scale bar: 50kb.

Using Illumina reads from the African genome project (Glanzmann et al., 2016; 1), we attempted targeted assembly of the IGH locus. Sequence data from the African buffalo genome assembly is currently unavailable and no characterisation of the IGH has been published. The African buffalo IGH locus was *de novo* assembled with single end reads identified by mapping to the cattle ARS-UCDv0.1 assembly using 10% mismatch identity. Following low quality filtering, 159 contigs were assembled spanning 225 kb with a contig N50 size of 1697 bp. This assembly was improved by providing paired-end information to the assembler. The paired-end library spans 324 kb (figure 3.7) and consists of 70 contigs, the largest contig 23 kb, with an N50 of 6062 bp. 6 *IGHJ*, 4 *IGHD*, 7 *IGHC* (*IGHM*, *IGHD*, *IGHG*, *IGHE* and *IGHA*) and 12 *IGHV* were identified in the complete assembly. Average coverage across the assembly was 4.109 x; specifically coverage across the *IGHV* was 3.521 x, across the *IGHDn* region 0.409 x and across the *IGHJn-IGHC* 3.936 x (figure 3.8). Peaks in coverage were located predominantly in introns. Due to the difficulties in assembly of the IGH with short sequencing reads, the assembly is incomplete with large gaps in sequence information.

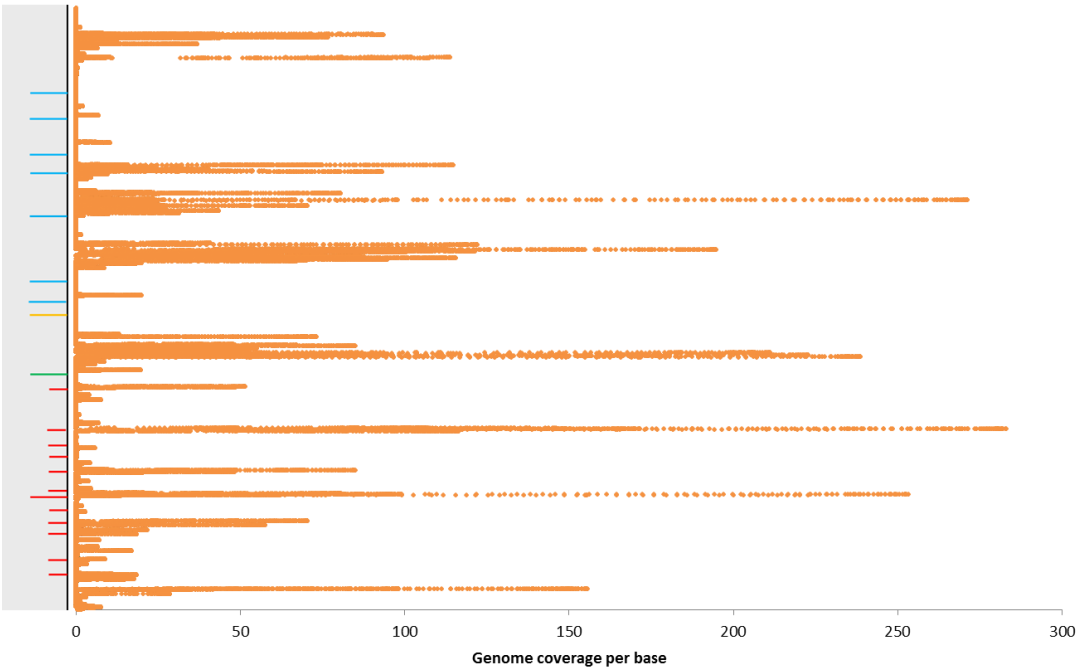


Figure 3.8: Scatter plot of genome coverage per base position in the African buffalo IGH assembled using paired end reads. Coverage is highest in the intronic regions, possibly due to the difficulty in mapping short sequence reads to the highly repetitive gene segments.

Gene segments were instead targeted and assembled individually to reduce the error in complete assembly of the IGH. Reads were mapped to individual gene segments in the cattle ARS-UCDv0.1 and *de novo* assembled in SPAdes (Bankevich et al., 2012; 178). A total of 57 *IGHV*, 6 *IGHJ*, 4 *IGHD* and the 7 *IGHC* were assembled and annotated in the African buffalo and concatenated into one sequence, S.caffer_IGH. 34 *IGHV* were assembled from reads with 4% alignment error and a further 23 *IGHV* were found when alignment error was increased to 10%. *IGHJ*, *IGHD* and *IGHC* genes were assembled from the 4%, and sequences confirmed in the 10% assemblies, with a 99-100% nucleotide identity in identical gene segments between the two mismatch identities. The S.caffer_IGH, when concatenated, spanned 78.8 kb (Table 3.2). The majority of the *IGHV* were assembled on a single contig while six were assembled on two contigs. The average coverage across the *IGHV* was 11.44 x but with a large range of 1.28 - 35.38 x. The *IGHJ* locus was assembled on a single contig with coverage of 17.19 x.

Buffalo Gene name	Contigs	Coverage	Assembly GC	Contig length
HVa-1	1	2.91	55.5	357
HVa-2	1	5.66	53.2	965
HVa-3	1	18.6	52.7	967
HVa-4	1	7.07	61.1	484
HVa-5	1	18.85	57.7	647
HVa-6	2	11.31	55.8	430
		15.04		352
HVa-7	1	15.35	55.6	729
HVa-8	2	3.97	58.9	453
		1.28		283
HVa-9	1	11.55	58.2	476
HVa-10	1	19.05	50.6	745
HVa-11	1	10.8	53.8	701
HVa-12	1	19.1	58.9	377
HVa-13	1	5.4	57.5	546
HVa-14	1	14.07	56.7	630
HVa-15	1	16.38	57.5	730
HVa-16	1	12.41	54	535
HVa-17	1	9.84	54.2	806
HVa-18	1	6.57	55.1	778
HVb-1	1	6.77	51.4	836
HVb-2	1	17.61	55.1	869
HVb-3	1	14.46	56.2	418
HVb-4	1	9.38	58.7	424
HVb-5	1	7.52	52.8	735
HVb-6	1	8.74	55.9	886
HVb-7	1	13.06	54	821
HVb-8	1	8.12	54.45	955
HVb-9	1	13.21	53.3	1146
HVb-10	1	13.55	54.3	1126
HVb-11	1	14.84	57.6	342
HVb-12	1	12.82	54.9	970
HVb-13	1	11.5	57.9	646
HVb-14	1	14.12	52.8	1043
HVb-15	1	4.43	57.83	498
HVb-16	1	6.15	56.33	758
HVb-17	1	13.08	53.5	979
HVb-18	1	23.69	54.4	756
HVb-19	1	22.09	55.4	713
HVb-20	1	12.8	56.4	525
HVb-21	1	30.49	52.8	1135
HVb-22	1	24.62	56.9	501
HVb-23	1	3.25	53.44	872
HVb-24	1	1.97	55.5	585
HVb-25	1	5.38	53.4	667
HVb-26	1	3.72	54.24	719
HVb-27	1	16.05	54.6	873
HVc-1	1	5.6	50.9	851
HVc-2	1	7.87	56.2	747
HVc-3	1	8.88	55.6	804
HVc-4	1	4.51	53.6	966
HVc-5	2	5.74	56	469
		3.05		336
HVc-6	2	2.69	56	443
		12.74		352
HVc-7	1	6.69	51.81	911
HVc-8	1	13.22	50.2	948
HVc-9	2	5.3	52	473
		1.54		325
HVc-10	1	9.01	54.55	1023
HVc-11	1	35.38	55.6	677
HVc-12	3	28.5	55.4	787
HJ/DHQ52	1	17.19	62.6	2770
HD1/HD2	1	12.81	63	3615
HD3/HD4	1	18.7	59.2	792
HCM	2	18.75	61.6	2110
		12.44		1576
HCD	2	15.99	66.5	2774
		6.02		474
HCG1	1	16.05	62.5	3817
HCG2	1	13.41	63.1	4528
HCG3	1	15.18	62	1848
HCE	1	19.93	59.5	8088
HCA	1	16.71	63.3	3474

Table 3.2: Mapping statistics of *IGHV de novo* assembled in African buffalo with reads mapped to the cattle ARS-UCDv0.1 gene segments. A total of 57 *IGHV* were assembled which formed three phylogenetic sub-groups designated a, b and c which correspond to the IMGT clan II for sub-group a and b and clan I for sub-group c.

A comparison of the structure of the S.caffer_IGH to the cattle ARS-UCDv0.1 is challenging due to sequence gaps and the unknown ordering of individually assembled gene segments. Importantly though, unlike cattle, African buffalo do not appear to have duplications within their IGH locus. Mapped reads to duplications in the cattle IGH assemble into indistinguishable buffalo genes; if duplications exist in the buffalo IGH they must be almost highly similar with an alignment error of less than 4%.

3.3.8 SNP calling of the African buffalo *IGHJ* region

To determine if internal duplications exist in the African buffalo IGH locus, genomic reads were mapped to the African buffalo *IGHJ* assembly and SNP calling looked for the presence of SNPs in the mapped reads. The *IGHJ* region was first assembled as a single contig with 4% mismatch identity, as outlined in section 3.3.7. Genomic reads were then mapped back to the *IGHJ* assembly using GEMmapper with a 4% alignment error and a SNP pile up was generated. Across the 2770 bp region, on average 90.61% of the reads contained the reference base at each position (Figure 3.9). None of the base positions contained SNPs in a 50:50 ratio, providing evidence that a duplicated *IGHJ* region in the African buffalo genome does not exist, or is identical in sequence.

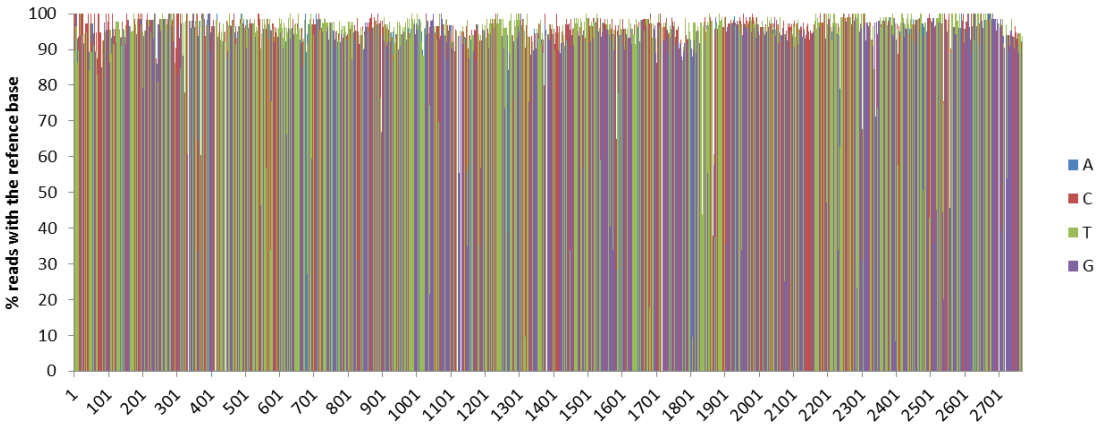


Figure 3.9: The *IGHJ* region was *de novo* assembled into a single contig using African buffalo genome reads that mapped to the cattle *IGHJ* region in the ARS-UCDv0.1 assembly. The individual African buffalo reads were then mapped to the assembled *IGHJ* contig and SNP calling was done at each base position. The percentage of mapped genome reads containing the reference base at each position are shown.

3.3.9 Predicted expression of putatively functional *IGHV* genes

In the cattle ARS-UCDv0.1 assembly, five of the identified 40 *IGHV* are putatively functional (figure 3.10) and possess canonical RS and octamer promoter sequences (Appendix Table 2). The remaining 35 *IGHV* are pseudogenes. The structure of the functional *IGHV4-11*, *IGHV4-14*, *IGHV4*-23, *IGHV(II)-27* and *IGHV(II)-33* was identical in the Holstein capture data and therefore highly likely to be correct.

Putatively functional *IGHV* in the ARS-UCDv0.1 also correspond to those in the contiguous sequence from Ma et al (2016) and the UMD3.1 assembly described by Niku et al (2012). As well as the five putatively functional *IGHV* described in the ARS-UCDv0.1, an additional five gene segments in the UMD3.1 are considered putatively functional, which are pseudogenes in the ARS-UCDv0.1. One of the additional putatively functional gene segments in the UMD3.1 is truncated in the ARS-UCDv0.1 by the contig break in IGH_217 (*IGHV(II)-6*). The pseudogene *IGHV(II)-8* is also putatively functional in the UMD3.1 and capture data from both Holstein animals. In the ARS-UCDv0.1, *IGHV(II)-8* contains a frame shift caused by a single base pair deletion in FR2 at position 53 (AGT to AT) which renders it a pseudogene but which is likely a sequencing error. The remaining three putatively functional genes described by Niku et al (2012), correspond to *IGHV2-18*, *IGHV(II)-20*, and *IGHV(II)-31*, containing multiple frameshifts or stop codons. An ORF was also described in the UMD3.1 assembly, corresponding to *IGHV4-39*, but this was a pseudogene with incorrect initiation codons in the ARS-UCDv0.1 assembly.

The eleven putatively functional *IGHV* found in the IGH locus described by Ma et al (2016) correspond to the five putatively functional gene segments in the ARS-UCDv0.1: *IGHV4-11*, *IGHV4-14*, *IGHV4-23*, *IGHV(II)-27* and *IGHV(II)-33*. An additional four putatively functional gene segments, *IGHV(II)-6*, *IGHV(II)-8*, *IGHV(II)-20* and *IGHV(II)-31*, correspond to putatively functional genes identified by Niku et al (2009). A gene segment with a 98% nucleotide identity to the ARS-UCDv0.1, *IGHV(II)-30*, is identified as putatively functional but is absent in the UMD3.1 assembly and is truncated in the ARS-UCDv0.1 assembly. Additionally, Ma et al (2016) describe a gene segment which has a 100% nucleotide identity to *IGHV4-14* (figure 3.10).

Based on our provisional analyses using short read sequencing data, African buffalo have 13 putatively functional *IGHV*, 45 pseudogenes and an ORF (Appendix table 3). Four putatively functional buffalo *IGHV* could not be resolved due to sequence gaps. Five of the functional gene segments have functional homologs in cattle with 97-98% nucleotide sequence identity but only cattle *IGHV(II)27* has the same amino acid sequence as African buffalo *IGHVa-10*. The complete *IGHV* repertoire in the African buffalo however could not be resolved at present.

The putatively functional gene segments in cattle and African buffalo all have the same length Framework (FR) and CDR regions. CDR1 and CDR2 have a short hairpin loop of eight amino acids with limited variability between functional sequences. Cattle CDR2 sequences differ by a single amino acid between gene segments with African buffalo *IGHVa-2* being the only sequence with two amino acids that are dissimilar. CDR1 sequences have slightly more variability with up to three amino acid variations in cattle and four in buffalo. Framework regions (FR1-FR3) are 94-98% identical between sequences in cattle and 90-99% between buffalo. Putatively functional gene segments in cattle and African buffalo have the canonical RS heptamer CACAGTG and nonamer ACAAAAACC and all but cattle *IGHV(II)-32* have the consensus promoter octamer ATTTGCAT (cattle *IGHV(II)-32* has ATTTGCAC), suggesting the efficient recombination of nearly all putatively functional gene segments.

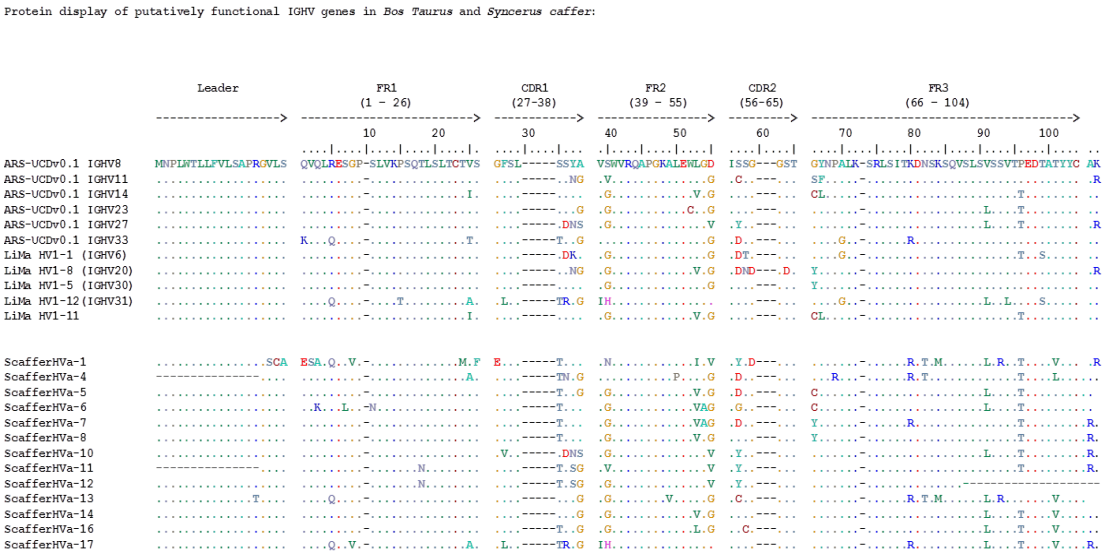


Figure 3.10: IMGT protein display of putatively functional cattle *IGHV* gene segments in the ARS-UCDv0.1 assembly and in the cattle IGH assembly described by Ma et al (2016), with the corresponding ARS-UCDv0.1 gene segment nomenclature indicated in brackets. *IGHV(II)-8* in cattle, which contains a single deletion causing a frame shift mutation, is a pseudogene in ARS-UCDv0.1 but its sequence is in frame in the genomic enrichment data (Heimeier et al; unpublished) and the UMD3.1 assembly. The African buffalo (*Syncerus caffer*) *IGHV* from targeted *de novo* assembly of the individual gene segments are also shown. Three of the buffalo *IGHV* were incompletely assembled but appear putatively functional, the dash lines indicates their missing sequence information in the leader (HVa-12) or FR3 (HVa-4 and HVa-11). Sequences are organised to indicate their leader sequence and the three framework regions (FR) and complementarity determining region (CDR) structures. Identical amino acids are represented with a dot and gaps are represented by a dash.

3.3.10 RNA-seq mapping analysis in cattle and African buffalo

Functionality of the cattle IGH was determined using RNA-seq data shared by Pasman et al at the University of Guelph. Of the approximately 17 million reads (range 9 - 21 million reads) generated for each of the three Holstein cattle, an average of 1.651% (~560,000) of the transcriptomic reads mapped to the cattle *IGHV* region in the ARS-UCDv0.1 assembly. Reads either uniquely mapped to a single *IGHV* gene segment or mapped across multiple gene segments and so were produced as weighted counts to reduce their bias. Of the RNA-seq reads that mapped, 9.51% mapped uniquely (figure 3.12A) and the remaining 90.49% were ambiguous. The African buffalo RNA-seq was performed with IgM and IgG antibody transcripts from two animals infected with FMDV. Transcripts from day 0, day 8 and day 14 post-infection were sequenced with Illumina, outlined in chapter 5, section 5.2.1. From a

mean of 2.6 and 3.2 million transcripts for IgM and IgG respectively, 50.015% and 50.008% mapped to the African buffalo *IGHV* (1.35 million). Of the reads that mapped, 3.43% were uniquely mapping whilst the remaining 96.57% (~1.3 million) were ambiguous reads and so produced weighted counts (figure 3.13A). Due to the highly repetitive nature of the *IGHV* region in cattle and African buffalo, a large proportion of the RNA-seq reads map to multiple genes.

The mappability of the transcribed cattle and African buffalo *IGHV* was subsequently calculated to determine if the uniqueness of the gene segments biased the RNA-seq analysis. The average mappability of the cattle gene segments was ~25% except *IGHV(II)-6* which was 86% (figure 3.11). African buffalo had on average lower mappability, 14%, except for *IGHVa-2* and *IGHVa-3* which were 62% and 69% (figure 3.11). The gene segments in cattle and African buffalo which produced the highest counts in RNA-seq were not the most unique genes (cattle *IGHV4-14* and African buffalo *IGHVa-5* and *IGHVa-9*). The most unique genes in both species were highly transcribed however, suggesting either a bias in the mapping or that there is a selection pressure for expression of more variable gene segments. However of these possibilities, we attempted to limit the former by producing weighted counts of mapped reads.

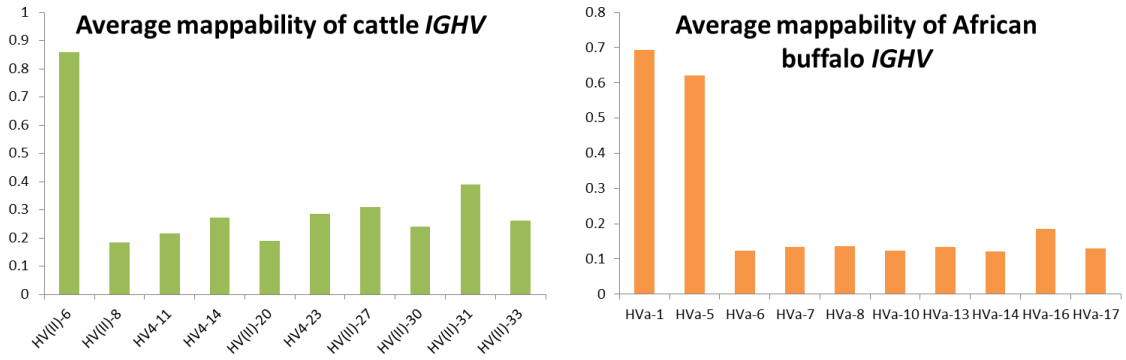


Figure 3.11: The mappability of the putatively functional cattle and African buffalo *IGHV* was calculated. The putatively functional cattle gene segments were generated from the ARS-UCDv0.1 and the Ma et al (2016) IGH assembly. African buffalo gene segments were *de novo* assembled using genomic reads mapped to the cattle *IGHV* gene segments in ARS-UCDv0.1. Using 150 bp lengths to mimic the RNA-seq read lengths; an exhaustive alignment was performed with GEMmapper (Marco-Sola et al., 2012; 177) from each base position. The average mappability is displayed as a percentage for each gene segment.

3.3.11 RNA-seq expression analysis in cattle

Transcription of the *IGHV* gene segments in the ARS-UCDv0.1 assembly was determined from the Holstein RNA-seq data from Pasman et al (2017). Weighted read counts were produced on nine of the 39 *IGHV* in the assembly: *IGHV(II)-6*, *IGHV(II)-8*, *IGHV4-11*, *IGHV4-14*, *IGHV4-23*, *IGHV(II)-27*, *IGHV(II)-30*, *IGHV(II)-31* and *IGHV(II)-33*. The Holstein animals show similar expression levels for the nine *IGHV* gene segments between the three animals (figure 3.12). The five predicted putatively functional gene segments in the ARS-UCDv0.1 are all transcribed. *IGHV(II)-6* was truncated in the ARS-UCDv0.1, whilst *IGHV(II)-8* contained a single indel, but both were putatively functional in the UMD3.1 and in the assembly from Ma et al (2016). The gene segments *IGHV(II)-30*, which is putatively functional in the assembly by Ma et al (2016), and *IGHV(II)-31*, which is functional in both the assembly by Ma et al (2016) and in the UMD3.1 reference assembly, were also transcribed (figure 3.13B). The predominance of *IGHV4-14* transcripts is possibly due to the gene segment being duplicated in the cattle genome as Ma et al (2016) identified an additional putatively functional gene segment in their IGH assembly with a 100% nucleotide sequence identity to *IGHV4-14*. Overall, no pseudogenes from either assembly are transcribed as would be expected, and which also confirms our characterisations.

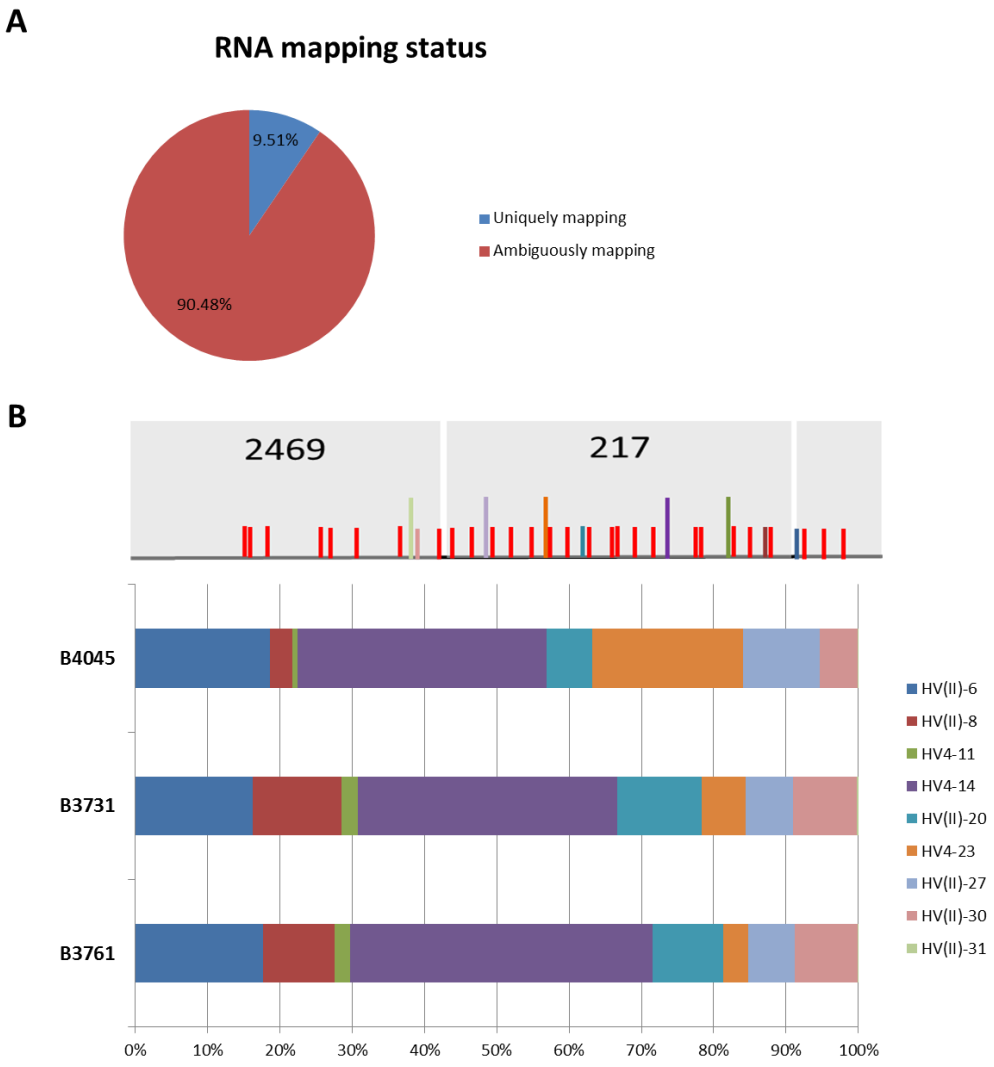


Figure 3.12: Average apportionment of reads in three Holstein cattle (B4045, B3761 and B731) that mapped to the cattle ARS-UCDv0.1 assembly (A). Of the mean 17 million reads per animal generated, ~560,000 mapped to cattle *IGHV* region; 53,256 reads (9.51%) mapped uniquely to the locus and ~482,000 mapped ambiguously (90.48%). Expression percentiles were calculated for transcription of *IGHV* in each animal, based on the total number of weighted counts, displayed graphically as 100% stacked columns (B). The *IGHV* region from the ARS-UCDv0.1 assembly is shown above the stacked columns, with the gene segments to the corresponding RNA-seq transcripts colour coded. A total of nine *IGHV*, including all of the putatively functional gene segments, were transcribed at similar expression levels between animals.

3.3.12 RNA-seq expression analysis in African buffalo in response to FMDV

Expression of the individually assembled *IGHV* gene segments in African buffalo was determined from the mapped IgM and IgG transcripts. Alignments were produced on ten of the 57 *IGHV* in the S.caffer_IGH (figure 3.13B). Of those ten gene segments, seven were

putatively functional whilst *IGHVa-3* contained a stop codon, *IGHVa-2* contained 3 stop codons and *IGHVa-9* contains a sequence gap. The relative usage of the *IGHV* did not significantly vary between individual animals at day 0 although differences between IgM and IgG transcripts are observed. *IGHVa-10* is utilised more in IgG transcripts whilst *IGHVa-5* is in IgM. At day 8 and day 14 post-infection with FMDV, usage of *IGHV* in IgM transcripts does not appear to vary but a decrease in the use of *IGHVa-10* in IgG transcripts is observed, although the change is not significant (figure 13C). Overall changes in the gene usage of *IGHV* gene segments in response to infection appears limited.

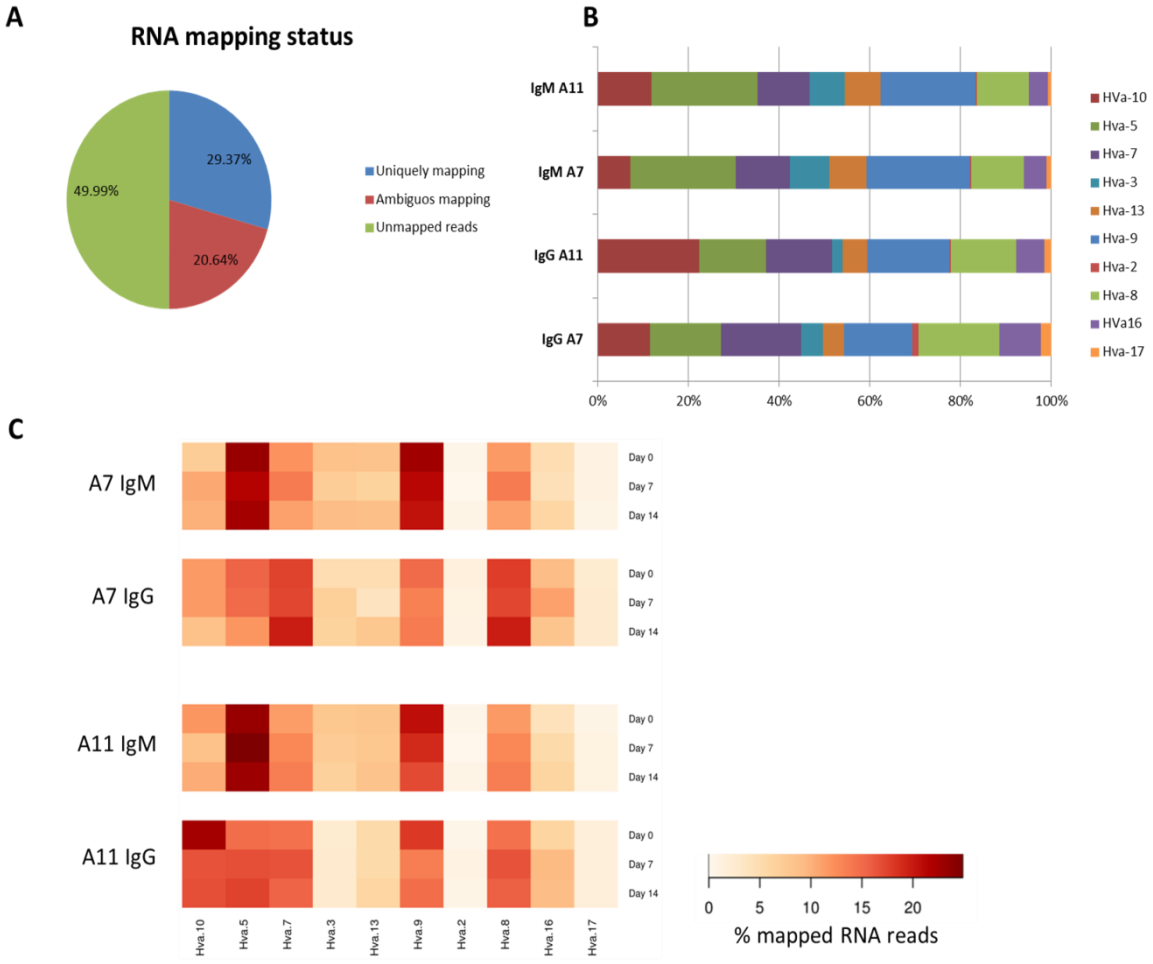


Figure 3.13: Transcription analysis of the IgM and IgG antibody transcripts in African buffalo animals 7 and 11. A mean of 2.7 million reads in each sample were aligned to the individually assembled buffalo *IGHV* gene segments and of these, an average 1.35 million reads mapped (A). Of the reads that mapped, ~790,000 were uniquely mapping (29.37%) and the rest were ambiguously mapping and so produced weighted counts. Reads mapped to the *IGHV* at Day 0 reveal the transcription of 10 gene segments (B). Expression percentiles of mapped reads are displayed in a heat map (C) to show *IGHV* usage at Day 0 and then post infection with FMDV SAT1 virus at Day 7 and Day 14.

3.3.13 RNA-seq expression analysis of cattle *IGHC*

Holstein RNAseq reads were aligned to the *IGHC* in the ARS-UCDv0.1 assembly (figure 3.14). An average 656,000 reads mapped to the *IGHC* region with multi-mapping capabilities with 56% of reads assigned to unique genes; 1.898% of the total transcriptome mapped uniquely to each gene. Transcription analysis showed the predominant expression of *IGHM*, accounting for around 50% of total reads in the blood. Reads mapping to each of the two *IGHM* were discernible with a predominance of the 3' *IGHM*, the *IGHM2*. The three *IGHG* subclasses combined make up 20-50% of reads and *IGHA* ~20%. Holstein RNA reads when mapped to the *IGHC* without multi-mapping showed significant changes in the ratio of *IGHM1* and *IGHM2*; the two *IGHM* are 99% identical in their nucleotide sequence and so only RNA reads containing SNPs would uniquely map. Reads mapped to the *IGHC* without multi-mapping had a stronger preference for *IGHM2* than *IGHM1*, with *IGHM1* accounting for ~5% of transcripts (figure 3.13B). The total reads mapped was reduced to 367,000 reads so changes in the expression ratio could be biased in the unique regions of the gene segments. Overall, the duplication in the cattle IGH created two functional *IGHM* which are both expressed.

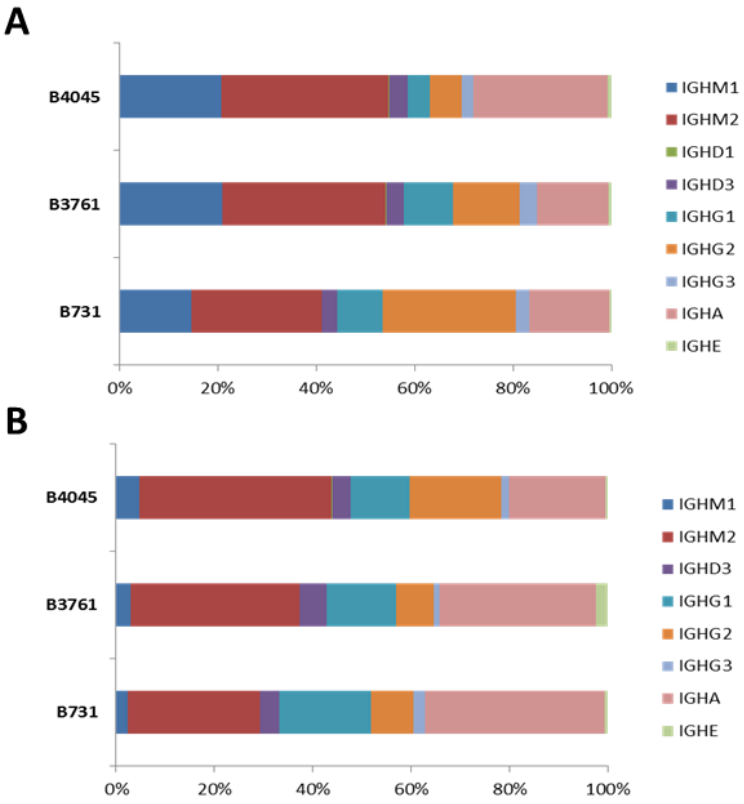


Figure 3.14: Transcription analysis of the *IGHC* in three Holstein animals (B4045, B3761 and B731) from the entire transcriptome. RNA-seq reads were mapped to the ARS-UCDv0.1 *IGHC* gene segments in GEMtools with multi-mapping producing weighted counts (A) and with unambiguous reads only (B).

3.3.14 Phylogenetic analysis of IGH

The two previously identified *IGHJ* regions each containing six *IGHJ* gene segments in cattle were found approximately 130 kb apart and span ~2 kb each. These two regions share 98% nucleotide sequence identity and have a 21 bp indel between the duplicated regions. The buffalo *IGHJ* region also contain six gene segments and has a a 94% nucleotide sequence identity to both the cattle *IGHJ* regions. Phylogenetic analysis of the individual *IGHJ* gene segments from cattle, buffalo, goat and humans show the *IGHJ* cluster by gene segment, regardless of the species it is isolated from, indicating each *IGHJ* is more similar between species than to any other *IGHJ* gene segment. This suggests the gene segments expanded in a common ancestor that pre-dates the divergence of primates and cetartiodactyls ~94 million years ago.

The duplications in cattle of the *IGHD* cluster has resulted in three *IGHD* regions in the ARS-UCDv0.1 (Figure 3.15A). A fourth *IGHD* region exists in the IGH assembly from Ma et al (2016) which has an identical structural arrangement and 99% nucleotide sequence identity to the second *IGHD* region in the ARS-UCDv0.1 and 100% nucleotide identity between corresponding gene segments, therefore was not resolved in our sequence assembly, section 3.3.5. The 14 *IGHD* gene segments in the cattle ARS-UCDv0.1assembly form five phylogenetic sub-groups, with duplicated gene segments within the same sub-group (figure 3.15B). Upstream of each *IGHJ* region in cattle (450 bp upstream) and buffalo (400 bp upstream) are the conserved *IGHDQ52* genes which are structurally and phylogenetically distant from all other *IGHD* gene segments.. The remaining *IGHD* gene segments in cattle and buffalo have repetitive GGT and TAT codons and are over 30 bp in length. The ultra-long *IGHD* gene segment, *IGHD7* in cattle, is 147 bp and appears to have been formed from a duplication of *IGHD6.* African buffalo appear to have only one *IGHD* region of genes containing four gene segments, with a 93% nucleotide identity to the 3' *IGHD* region in cattle. The *IGHD5-2* gene segment in cattle however is not assembled in the African buffalo *IGHD* region; the *IGHD* in African buffalo is assembled as a single contig and so it is likely this gene segment is not present in the genome. An ultra-long *IGHD* gene segment is also absent from our buffalo assembly. This suggests that either internal duplications do not exist in the African buffalo IGH locus or that the sequence assembly has not resolved the duplications.

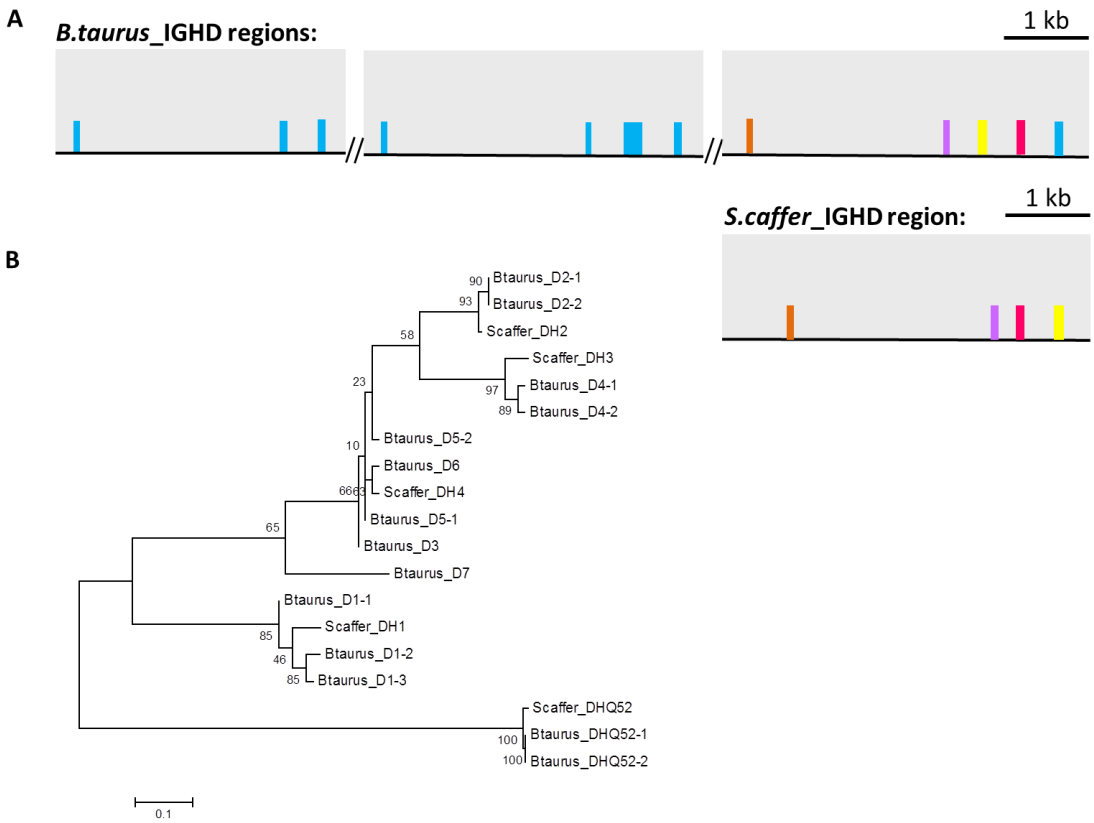


Figure 3.15: Schematic of the *IGHD* gene sub-clusters in the cattle ARS-UCDv0.1 assembly and the *IGHD* region in the African buffalo assembly (A) to scale. *IGHD* gene segments were colour co-ordinated if their nucleotide sequence identity between animals was >90%, any gene segments absent in the buffalo assembly remained blue. *IGHD* nucleotide sequences were then aligned and their phylogenetic relationship was calculated in Mega 6.0 (Tamura et al., 2013; 188) using maximum likelihood with the Tamura and Nei distance model (Tamura, 1992; 191) with 1000 bootstraps (B)

The *IGHV* in cattle and African buffalo were named according to their subgroup or higher order clan, as standard for IMGT nomenclature (Lefranc et al., 2003; 185). The sub-group or clan that each *IGHV* belonged was determined by comparison to the human and mice *IGHV* within the IMGT database. The human sub-group 4 is the same as bovine sub-group one which was used to describe the cattle *IGHV* in the UMD3.1(Niku et al., 2012; 167). Of the 39 identified *IGHV* gene segments in cattle, seven belong to *IGHV* subgroup 4. The remaining 31 had a sequence identity below the 75% threshold and were assigned to a clan: 11 *IGHV* gene segments belong to clan I, 19 to clan II (along with the 7 in subgroup 4). In African buffalo, 15 of the 57 *IGHV* belong to subgroup 4 from both phylogenetic sub-group a and b. All 18 sequences in subgroup a and 26 in subgroup b belong to clan II and the remaining 12 *IGHV* in subgroup c belong to clan I. Cattle and African buffalo do not appear to have any gene segments belonging to human or mouse clan III.

3.4 Discussion

The organisation of the cattle IGH is unusual compared to species studied to date, with internal duplications disrupting the expected *IGHVn-IGHDn-IGHJn-IGHC* organisation seen in other mammalian species (Litman et al., 1993; 192). The finding of these internal duplications resolved questions over the apparent duplication of the *IGHJ* and *IGHM* which were predicted from previous genome assemblies to be present on separate chromosomes despite evidence that both are expressed; the inhibition of bovine antibody production by gene knockouts, requires that both *IGHM* are inactivated (Kuroiwa et al., 2009; 193). The structure of the cattle IGH was thus compared in three separate assemblies: the ARS-UCDv0.1 long read assembly, the publicly available assembly UMD3.1 and the IGH assembly from Ma et al (2016). The IGH locus in the UMD3.1 assembly was heavily disrupted and assembled on multiple chromosomes; short read Illumina sequencing is unable to span repetitive regions of the genome. The assembly from Ma et al (2016) provided the first complete characterisation of the cattle IGH locus, confirming that it is on a single chromosome and that the perplexing questions of associated genes are due to internal duplications in the IGH locus.

The structural organisation of IGH in the ARS-UCDv0.1 assembly and the assembly from Ma et al (2016) is different; a duplication of *IGHVn-IGHDn-IGHJn-IGHM-IGHD* is in both assemblies, however ARS-UCDv0.1 contains a single duplication of *IGHVn-IGHDn-IGHD* whilst the assembly by Ma et al (2016) contains two of these duplications. The BAC clone, RP42-567N23, used by Ma et al (2016) to sequence the region across the internal duplications was subsequently sequenced with long read technologies, PacBio and Oxford Nanopore, and assembled to look for the additional *IGHVn-IGHDn-IGHD* duplication. Assembly of the RP42-567N23 with both long read sequencing platforms appeared to confirm the structure of ARS-UCDv0.1 but SNP calling by alignment of the PacBio reads to the assembled RP42-567N23 clone revealed a highly similar SNP profile to both duplications in the assembly from Ma et al (2016). Assembly of the additional *IGHVn-IGHDn-IGHD* region therefore failed in the ARS-UCDv0.1 assembly when using two long read sequencing technologies, PacBio and Oxford nanopore, to sequence and assemble the RP42-567N23 BAC clone. Long read sequencing technologies are highly error prone and so are unable to

assemble the highly similar repetitive regions and failed to resolve the highly similar duplications in the cattle IGH locus.

Discrepancies in *IGHV* organisation also exist between the two assemblies, with Ma et al (2016) describing an additional eight *IGHV* gene segments. Genomic enrichment data, sequenced with PacBio, from Holstein cattle appeared to support the absence of three of the described *IGHV* missing in the ARS-UCDv0.1 assembly but this could again be assembly error. The region described by Ma et al (2016) is structurally different; the ordering of the *IGHV* is different to the ARS-UCDv0.1. RNA-seq data appears to support the sequence from Ma et al (2016) in that the highest expressed *IGHV4-14* gene segment is duplicated in their sequence and so transcription of two putatively functional gene segments of identical sequence shows a bias in our expression analysis. Ultimately then, the overall structure of the IGH locus described by Ma et al (2016) appears the more accurate.

The unusual genomic structure in cattle allows independent V(D)J recombination at two *IGHM* genes and two *IGHJ* regions. Using RNA-seq data from the transcriptome of three Holstein cattle (provided by Pasman et al at the University of Guelph), the expression of both *IGHM* genes was shown (figure 3.14). Only three *IGHD* and a two *IGHJ* putatively functional gene segments are located upstream of *IGHM1*, whilst 16 *IGHD* and four putatively functional *IGHJ* are upstream of *IGHM2.* Restriction of RNA-seq to uniquely mapping reads reduced the number mapping to *IGHM1* to ~0.2 x that of *IGHM2*. 16 *IGHD* can be involved in rearrangement with *IGHM2* but only three can rearrange with *IGHM1* so the theoretical usage of *IGHD* gene segments reflects the expression percentiles of the *IGHM*.

Assembly of the African buffalo IGH provides the first insight into their antibody encoding structure and recombinatorial potential. Genomic reads were provided from an assembly project which at time of submission has yet to be made publically available (Glanzmann et al., 2016; 1). Our African buffalo assembly of the whole IGH locus is mis-assembled with many sequence gaps due to the drawbacks of using short Illumina reads. Providing paired-end information to the assembler improved the *de novo* assembly as distance information between paired reads better determined repetitive regions and structural arrangements but this assembly was also heavily fragmented. The targeted assembly of individual gene segments resolved a more complete and higher coverage of the IGH genes, with 57 *IGHV*, four *IGHD*, six *IGHJ* and seven *IGHC* described. African buffalo possess highly similar *IGHJ* and *IGHC* (*IGHM, IGHD, IGHG, IGHA* and *IGHE*) genes to cattle. Their single *IGHD* region contains

four *IGHD* gene segments, with the notable absence of an ultra-long *IGHD* gene segment. A greater number of *IGHV* are described in African buffalo than in cattle but since gene segments were assembled individually this could be due to incorrect mapping of certain reads creating a SNP bias in gene assemblies.

African buffalo and cattle, despite sharing a common ancestor 5.7 - 9.3 million years ago (Glanzmann et al., 2016; 1), do not appear to contain the *IGHVn-IGHDn-IGHJn-IGHM-IGHD* or *IGHVn-IGHDn-IGHD* duplications in their IGH. SNP calling of African buffalo genomic reads mapped to the African buffalo *IGHJ* assembly confirmed the absence of an additional *IGHJ* region. The vast majority of the mapped reads contained the reference base at each position and so there was no evidence for a duplicated region in the SNPs. This therefore confirms the absence of an internal duplication in the African buffalo IGH, or if a duplication exists it is identical.

The recombinatorial potential of the cattle and African buffalo IGH is severely limited compared to humans and mice (de Bono et al., 2004; 194, Tomlinson et al., 1995; 195). Cattle have only six putatively functional *IGHV* gene segments in the ARS-UCDv0.1 assembly and twelve described by Ma et al (2016). Pseudogenes caused by indels in the PacBio sequencing pipeline are resolved by comparison to Illumina reads in the UMD3.1 assembly and the sequence from Ma et al (2016). Nine of the described *IGHV* are expressed in the transcriptome of Holsteins, and all of which belong to clan II. Combined with the predicted expression of only four *IGHJ* gene segments and sixteen *IGHD*, the combinatorial diversity attributed to the cattle immunoglobulin heavy chain is only 576 possible VDJ combinations. Of the gene segments successfully assembled in African buffalo, ten of the *IGHV* were expressed in IgM and IgG transcripts whilst two of the *IGHJ* and the four *IGHD* were putatively functional in the germline. African buffalo have a combinatorial diversity of only 80 possible VDJ combinations. Although the African buffalo assembly is incomplete, preliminary evidence suggests their IGH locus is less diverse than cattle. In contrast, humans have a primary combinatorial potential of 6346 IGH VDJ combinations. The recombinatorial potential of the IGH germline in both species is therefore severely limited and post-transcriptional modification must take place in order to generate a functional antibody repertoire.

A novel diversification mechanism in cattle is the formation of long and ultralong CDR3H loops. The average *IGHD* length in humans and mice is 23 and 17 bp respectively, whilst all

of the cattle and African buffalo *IGHD* are greater than 30 bp and the ultralong *IGHD7* in cattle is 147 bp. The long and ultralong *IGHD* gene segments in the cattle germline are responsible for forming long and ultralong CDR3H. A comprehensive analysis of CDRH3 length in cattle, humans, and mice suggests that there is strong positive correlation between the average lengths of CDRH3 and *IGHD* segments. High levels of somatic hypermutation occur along the length of these long and ultra-long CDR3H, diversifying the cattle antibody repertoire. A bias towards cysteine modification also exists and the ultralong CDR3H structures form various di-sulphide bonding patterns that creates additional repertoire diversity. An ultra-long *IGHD* gene segment was not assembled in the African buffalo so the formation of ultralong CDR3H would be due to an alternative mechanism to what we observe in cattle.

A more detailed cattle IGH was resolved by comparing third generation sequence reads and assembly with the published cattle IGH from Ma et al (2016). We confirmed the overall structure of the IGH sequence by Ma et al (2016) using SNP comparison of the individual reads but found structural differences in the *IGHV* regions between the ARS-UCDv0.1 and Ma et al which has not been resolved. Using the cattle IGH in the ARS-UCD0.1 African buffalo sequence reads were aligned and assembled, providing the first characterisation of the African buffalo IGH gene segments. Their IGH appears less complicated with the absence of the large internal duplication of the *IGHVn-IGHDn-IGHJn-IGHM-IGHD*. RNA-seq data in both species support low gene usage and therefore low recombinatorial potential of the germline IGH. We explore post-translational modification and repertoire diversification in Chapter 5.

Chapter 4

Characterisation of the IGL and IGK in cattle and African buffalo (Syncerus caffer)

4 Abstract

The antibody light chain enlarges the variability of the antibody repertoire and contributes to antigen binding either directly or by supporting the conformation of the heavy chains. The two distinct light chain immunoglobulin isotypes lambda (IGL) and kappa (IGK) were characterised in a previously published annotation of the cattle reference genome, Btau_3.1 (Ekman et al., 2009; 99), but the description is incomplete due to fragmentation in the assembly. Using the long read whole genome assembly ARS-UCDv0.1 we have generated a more detailed annotation of the cattle IGL locus than previously. Here we show structural improvements of the IGL and IGK in the ARS-UCDv0.1 compared to reference genome assemblies, the Btau_3.1 and the more recent UMD3.1. The first description of the African buffalo light chain loci is shown using reads mapped to the cattle ARS-UCDv0.1 as a reference and *de novo* assembled as entire IGL and IGK loci or as individually assembled gene segments. African buffalo have a similar IGK locus to cattle but a seemingly less expanded IGL. Despite this, IGL light chains predominate in the African buffalo as they do in cattle accounting for 95-98% of the light chain repertoire. Functionality of the IGL gene segments in Holstein cattle and African buffalo is then explored using RNA-seq, revealing the predominant usage of few *IGLV* gene segments. The restricted use of the light chain means recombinatorial diversity of the antibody repertoire is limited.

4.1 Introduction

Immunoglobulins are the molecular mediators of the humoral immune response and form a diverse array of B cell receptor structures to recognise diverse pathogen antigens. In cattle three Ig loci exist, the heavy chain on chromosome 21 (characterised in chapter 3), the IGL on chromosome 17 and the IGK on chromosome 11. Rearrangement of numerous variable (V), joining (J) and, on the heavy chain, diversity (D) gene segments produce the VJ and VDJ genes that encode the light and heavy chains respectively. Rearrangements are temporarily separated during B cell development with rearrangement of IGH occurring first, so a surrogate light chain (SLC) is expressed. The SLC is composed of the *VPREB* and *IGLL1* polypeptides that are homologous to the variable and constant domain of the IG light chains. Light chain expression then replaces the SLC to form the finished B cell receptor.

The use of IGL or IGK is first determined by somatic rearrangement of the kappa locus. If a functional antibody fails to be produced, the kappa deleting element (KDE) and a recombining element in the J-C intron recombine to ablate kappa expression. The lambda locus then undergoes rearrangement until a functional light chain is produced or the B cell is destroyed. Successful rearrangement of the light chain forms the B cell receptor and leads to its expression on the cell surface, along with the IGH. The B cell receptor is then tested for self-reactivity before it is exported and allowed to proliferate.

The relative expression of IGL or IGK in antibodies differs between species. Humans and mice express predominantly kappa chains (70% and 95% respectively), whereas lambda is predominantly expressed in cattle and sheep (Arun et al., 1996; 196, Gray et al., 1967; 197, Sinkora et al., 2001; 198, Murphy, 2008; 199). These differences between species in the ratio of IGL:IGK expression may be due to differences in genomic complexity at each locus, exogenous antigen selection (Nishikawa et al., 1984; 200) or endogenous counter selection (Knott, 1998; 201). The previously published annotation of the cattle reference assembly, Btau_3.1, for the IGL and IGK (Ekman et al., 2009; 99) suggests that the preferential use of lambda is a result of the additional complexity at the lambda locus. A total of 63 *IGLV* and 22 *IGKV* is described in the cattle Btau_3.1 (Ekman et al., 2009; 99), but despite the greater complexity of the lambda locus, the number of putatively functional cattle Ig gene segments is markedly lower than that of mice or human: 33 *IGLV* vs 105 *IGLV* and 77 *IGLV* respectively (Ekman et al., 2009; 99, Solomon and Weiss, 1995; 202, Gerdes and Wabl, 2002; 203). The cattle light chain repertoire therefore appears limited.

Further antibody diversification occurs through IgL-IgH chain pairing. Interactions between the IGL and IGH chain contribute to the binding kinetics of a peptide and therefore the stability of the antibody structure (Chatellier et al., 1996; 204). The preference in pairing of particular heavy and light chains was presumed to occur at random (Brezinschek et al., 1998; 205, de Wildt et al., 1999; 206). In humans, pairing of light chains from multiple sub-groups can pair with a single *IGHV* and in several cases use of an *IGLV* or *IGKV* combines with identical VDJ combinations without loss of antigen specificity (Edwards et al., 2003; 207). However, pairing preferences have been shown to exist in the human and mouse germline for a small proportion of gene segments, whilst others are promiscuous and show no pairing preferences (Jayaram et al., 2012; 208). Specific pairings in cattle have been shown for the ultra-long CDR3H in cattle; these ultra-long sequences have unique configurational requirements and so have restricted IGL pairings (Saini et al., 2003; 209), which may have evolved specifically to provide a structural framework for supporting the ultra-long CDR3H (Wang et al., 2013; 92). The limited diversity of the cattle IGL repertoire leads to the hypothesis that the light chain repertoire performs a greater structural role in cattle than in other species.

Here, the *IGL* and *IGK* in the cattle long-read assembly, the ARS-UCDv0.1, is compared to the previous annotation of the light chain loci, the Btau_3.1 (Ekman et al., 2009; 99) and to the more recent cattle genome the UMD3.1. The ARS-UCDv0.1 IGL and IGK was then used as a reference for mapping of the African buffalo genome reads, supplied by Glanzmann et al at the South Africa Medical Research Council Centre for Tuberculosis Research (Glanzmann et al., 2016; 1). The IGL and IGK genes were subsequently *de novo* assembled, providing the first African buffalo light chain annotation. Functionality of the loci was explored through qPCR of the IGL and IGK and RNA-seq of *IGLV* in cattle and African buffalo with cattle RNA whole transcriptome reads from Pasman et al at the University of Guelph (Pasman et al., 2017; 190) and African buffalo IGL transcripts. Both species have a predominance of lambda light chains with no significant clonal expansion of kappa light chains upon infection of African buffalo or inoculation of cattle with FMDV. Putative functionality of *IGLV* gene segments shows limited repertoire potential and the RNA-seq shows the preferential usage of few *IGLV*, supporting the hypothesis that the IGL is providing a structural role.

4.2 Method

4.2.1 IGL and IGK genomic sequences of cattle and African buffalo

A single Hereford cow genome, the ARS-UCDv0.1, was *de novo* assembled using long reads generated with the Pacific Biosciences RSII platform, details of which are in Chapter 2, section 2.2.20. Scaffolds containing the lambda and kappa loci were identified in the ARS-UCDv0.1 and the publicly available reference assemblies, the Btau_3.1 and UMD3.1, using the basic local alignment tool (BLAST) to look for *IGLV* and *IGKV* sequence motifs. Positive scaffolds were extracted from the ARS-UCDv0.1 and downloaded from the UMD3.1 or Btau_3.1 with the Gbrowse archive in the Bovine Genome Database for sequence comparison of the assemblies.

Genomic sequence reads from the African buffalo (*Syncerus caffer*) assembly were leveraged from Glanzmann et al at the South Africa Medical Research Council Centre for Tuberculosis Research (Glanzmann et al., 2016; 1). Twelve mate pair libraries were constructed with 170 bp, 500 bp, 800 bp, 2 kb, 5 kb or 10 kb insert sizes from whole blood gDNA and sequenced on the Illumina Hi-seq 2000 at 60-fold coverage. Low quality data was filtered and the resulting FASTQ libraries were shared with us for further investigation. As outlined in Chapter 3, section 3.2.3, the African buffalo reads were mapped to the cattle ARS-UCDv0.1 with the Genome Multitool mapper v3.5-19-g1d79-dirty-release (GEMmapper) (Marco-Sola et al., 2012; 177) with 4% and 10% alignment maximum error. Files with both reads mapping and only one read mapping of a pair were generated and combined for *de novo* assembly of the IGL and IGK loci and the individual assembly of IGL and IGK gene segments. Assemblies were evaluated with the Quality Assessment Tool for Genome Assemblies (QUAST) (Gurevich et al., 2013; 181)and the read coverage per base was calculated with BEDTools v2.26.0 (Quinlan and Hall, 2010; 179).

4.2.2 Characterisation of the cattle and African buffalo IGL loci

IGL genomic sequences in the ARS-UCDv0.1 and IGL and IGK in the African buffalo loci assemblies and individual gene segment assemblies were manually annotated using Artemis v13.0 (Rutherford et al., 2000; 182). The IGK loci from the cattle ARS-UCDv0.1 was previously annotated by John Schwartz in the Immunogenetics Group (Schwartz et al, unpublished). Immunoglobulin domains were identified using both BLAST and the NCBI conserved domain database (Marchler-Bauer et al., 2015; 210). Annotated features of each putative gene segment included the leader, octamer (ATTTGCAT), 5' (GT) and 3' (AG) splice sites, the RS heptamer (CACAGTG or CACTGTG), RS nonamer (ACAAAAACC), RS spacer (23 bp/12 bp) and the conserved amino acid residues C23, W41, hydrophobic 89 and C104.

4.2.3 Structural comparison of the cattle ARS assembly

The structure of the ARS-UCDv0.1 IGL and IGK scaffolds was compared to the UMD3.1 reference assembly and the African buffalo assembly using DOTTER v4.44.1 (Sonnhammer and Durbin, 1995; 184) to generate recurrence plots with a sliding window of 200 bp.

4.2.4 Nomenclature of IGL genes

IGLV gene segments were named in the ARS-UCDv0.1 cattle assembly according to IMGT nomenclature (Lefranc et al., 2003; 185) in order from their positioning starting proximal to the constant region. *IGLV* and *IGKV* in African buffalo were numbered according to their phylogenetic subgroups due to the lack of a complete genome for annotation. The *IGLV* and *IGKV* in cattle and buffalo were also named in brackets according to their sub-group or clan for consistency with other species. The subgroup number was determined by comparison of sequences with BLAST to the IMGT database using a 75% identity threshold. When the subgroup was undetermined, *IGHV* gene segments were named according to their higher order clan, as designated by roman numerals.

Gene segments were considered functional if they were in-frame without truncations or pre-mature stop codons, possessed canonical initiation codons (ATG) and contained the conserved amino acid residues (C23, W41, hydrophobic 89 and C104). Gene segments were putatively determined to be pseudogenes if they contained stop codons, truncations, frame shifts or a defective initiation codon. Gene segments were defined as open reading frame (ORF) if they were missing conserved amino acid residues.

4.2.5 Phylogenetic analysis

The *IGLV* in cattle and African buffalo were aligned using a global alignment strategy in the MAFFT package, version 6.603b (Katoh et al 2002) and visually confirmed and extracted using Bioedit v7.2.5 (Ibis Biosciences, ref). Phylogenetic analysis of gene segments was calculated in MEGA 6.0 (Tamura et al 2013) using maximum likelihood based on the Tamura and Nei model (Tamura 1992) and the partial deletion method using a 95% cut off with 1000 bootstrap iterations.

4.2.6 Animals

Four Friesian cattle, aged 30 months were selected from an MHC-defined herd at The Pirbright Institute (animal ID C15-1020, C16-1021, C15-9472 and C15-9473, Animal request license number AR000579). African buffalo animals (7, 8, 10, 11, 13, 20, 28 and 32) were selected in the Kruger National Park, South Africa from an FMDV challenge study (details in Chapter 5, section 5.2.1). All the experiments were approved by the Pirbright Institute's ethical review process in accordance with Home Office guidelines on animal use.

4.2.7 Bovine PBMC isolation

Peripheral blood was collected by jugular venepuncture into sodium heparin (10 U/ml) from the cattle. Peripheral blood mononuclear cells (PBMC) were separated from the blood by

density gradient cell separation in sterile 50 ml conical tubes (Falcon). Whole blood was underlayed with Histopaque-1077 (Sigma-Aldrich, Gillingham) and the volume made up to 50 ml with phosphate buffered saline (PBS). The blood was centrifuged at 2,500 rpm, at 19°C for 40 min in a Rotina 420R centrifuge. PBMC were transferred to a conical tube using a Pasteur pipette, the volume of which was made up to 50 ml with PBS, and centrifuged again at 1,000 rpm, at 19°C for 8 min. Erythrocytes were lysed in 50 ml ammonium chloride lysis buffer (160 mM ammonium chloride, 170 mM Tris, pH 7.65) before a final wash of the PBMC in 50 ml PBS. PBMC were pelleted at 1,000 rpm, 19°C for 8 min.

4.2.8 Total RNA extraction from bovine PBMC

Total RNA was isolated from cattle PBMC using Trizol. PBMC were suspended in 1 ml Trizol (Sigma-Aldrich, Gillingham) and left at room temp for 5 mins to allow dissociation of the nuclear proteins. RNA was subsequently isolated within the aqueous layer by addition of 270 µl chloroform and centrifugation at 12,000 g for 15 min at 4°C to form the organic phase. RNA was pelleted in 667 µl isopropanol and washed in 1.5 ml ethanol (75%). The RNA pellet was suspended in 100 µl water and stored at -80 °C.

4.2.9 Total RNA extraction from African Buffalo whole blood

African buffalo total RNA was isolated from Tempus tubes as outlined in Chapter 5, section 5.2.2. Briefly, peripheral blood from eight African buffalo animals in the Kruger National Park was collected by tail venipuncture into Tempus Blood RNA Tubes (ThermoFisher Scientific) containing stabilising reagent and frozen at -80°C for transport to the UK. Using the Tempus Spin RNA Isolation kit (ThermoFisher Scientific), blood was transferred to 1.5 ml spin tubes and the volume made up to 12 ml using PBS. Following the Tempus Spin protocol (https://tools.thermofisher.com/content/sfs/manuals/cms_042989.pdf). The RNA was pelleted at 3,000 g for 30 min 4°C and re-suspended in 400 µl RNA purification solution. The RNA pellet was washed twice in 500 µl RNA purification wash solution and suspended in 90 µl RNA elution solution for storage at -80 °C.

4.2.10 Reverse transcription of cattle and African buffalo RNA

Purified RNA from cattle and African buffalo was reverse transcribed into cDNA using Superscript II. Reactions were set up with 200 U of Superscript II reverse transcriptase (ThermoFisher Scientific), 1 µl Oligo(dT), 1 µl dNTPs (50mM), 4 µl first strand buffer, 2 µl Dithiothreitol, 5 ng of RNA and the volume made up to 20 µl with water. The mixture was initially heated to 65 °C then incubated at 42 °C for 50 min and heat denatured at 70 °C for 15 min. Transcribed cDNA was then quantified on the Qubit (Thermo Fischer) and each sample standardised to 100 ng/µl.

4.2.11 qPCR primer validation for light chain expression

Primers for calculating IGL and IGK expression in cattle and African buffalo were designed against the cattle ARS-UCDv0.1. IGL and IGK expression was measured with primers for the *IGLC* and *IGKC* genes and subsequently confirmed by Sanger sequencing of the PCR products. The transcription of the reference genes *β-actin*, *SDHA* and *PPIA* were also measured for sample standardisation using primers previously optimised to work on cattle cDNA in our Immunogenetics group (Allan, 2015; 211). All the primer sets were designed using Primer 3 software (Appendix Table 1) and conformed to MIQE guidelines (Johnson et al., 2014; 212) for qPCR. Using genomic DNA from cattle animal 598 and African buffalo Sca04, primer specificity was confirmed by PCR in 25 µl reactions with GoTaq Green master mix x2 (Promega) and 300 nM of each primer. The PCR cycling conditions were 95 °C denature for 2 min and then forty cycles of 95 °C for 15 s, 58 °C for 1 min and 72 °C for 45 s. PCR products were purified using 4 µl ExoSAP-IT enzymatic clean up (Affymetrix) at 42 °C for 30 min and 5 µl aliquots were ran on a gel, 1 % agarose at 90 V for 90 min. Bands were visualised under UV light for size confirmation and the remaining PCR product was sent for Sanger sequencing for confirmation of the correct amplified product.

Primer efficiency was then calculated for *IGLC, IGKC, β-actin, SDHA* and *PPIA* primer pairs. Into 25 µl qPCR reactions, cDNA from ten cattle animals or ten African buffalo samples was combined and four serial dilutions at 100 nM (10x), 10 nM (1x), 1nM (0.1x) and 0.1nM (0.01x) final concentration were made for each species separately. Primer

concentration of 300 nM was used with the Luminaris Color HiGreen Low ROX qPCR Master Mix (Thermo Scientific) for SYBR green. Each reaction was run in triplicate on a variety of cDNA sample combinations. A final heating step was run in incremental steps from 58 °C to 95 °C to obtain dissociation curves for assessment of primer specificity. Standard curves for the serial dilution Ct values were calculated in Excel 2010 and the percentage primer efficiency for each primer pair was determined using the calculation below.

```
Exponential amplification = 10^ (-1 / standard curve slope)

Primer efficiency = 10^ (-1 / standard curve slope) - 1
```

The reference genes *β-actin*, *SDHA* and *PPIA* were validated using GeNorm analysis (Biogazelle), with the gene *NCR3* included as a negative control. The triplicate reactions of all the cattle and African buffalo cDNA samples genes were loaded into the qbasePLUS software version 2.5.1 (Biogazelle NV, Belgium).

4.2.12 Evaluation of different SYBR Green Master Mix

Three SYBR green master mixes from three different suppliers were tested to ensure accurate quantification of light chain expression. Using identical 300nM *IGLC*, *IGKC*, *SDHA* or *PPIA* gene primer mix and 10nM cDNA samples from cattle animals C16-1021 and C15-9472 and African buffalo Day 0 samples from animals A7 and A10, a qPCR was performed with each of three SYBR green master mixes: KiCqStart SYBR green qPCR readymix Low ROX (Sigma), Power SYBR Green PCR Master Mix (Applied Biosystems) or Luminaris Color HiGreen Low ROX qPCR Master Mix (Thermo Scientific). Thermocycling conditions were 95 °C denature for 2 min and then forty cycles of 95 °C for 15 s, 58 °C for 1 min and 72 °C for 45 s. Each reaction was run in triplicate and Ct values for each gene compared between each master mix.

4.2.13 qPCR of cattle, African buffalo cDNA for light chain expression

Expression of IGL and IGK was determined in the cattle and African buffalo samples. Quantitative real time PCR (qPCR) was performed in triplicate with the isolated Holstein cattle (1020, 1021, 9472, 9473) and African buffalo (7, 10, 11, 13, 8, 20, 28, 32) cDNA from blood. Holstein cattle cDNA from five animals (250, 255, 256, 257 and 258) Day 0 and Day 25 post-challenge with highly purified SAT 1 Zim (FMDV SAT1 ZIM 22/89 (S1Z), n = 4) FMDV antigen was also used to quantify light chain usage in response to FMDV inoculation (detailed in chapter 5, section 5.2.1). Reactions were performed with Thermo Fisher Luminaris Color HiGreen Low Rox qPCR master mix on an Applied Biosystems QuantStudio 5. Each reaction contained 4 µl diluted cDNA sample (10 ng per reaction), 10 µl of qPCR master mix and 300 nM final concentration of each chosen primer set. PCR cycling conditions were a 50 °C step for 2 min, 95 °C denature for 5 min then 40 cycles of 95 °C for 15 s, 58 °C for 1 min and 72 °C for 45 s. Dissociation curves were also generated for each reaction to ensure primer specificity between samples with incremental steps from 58 °C to 95 °C. Resulting qPCR amplification products were run on a 1% agarose gel at 90V for 1 hr to ensure products were of the correct size. Control reactions containing no template or no reverse transcriptase were both run on each plate to account for nonspecific enzymatic activity.

4.2.14 Statistics

CT scores and melting curves were exported from the Applied Biosystems QuantStudio 5 machine into Microsoft Excel for further analysis. Samples run in triplicate with a variance of > 0.2 between each Ct score were re-run. Melting curves were visualised to ensure primer specificity. Mean values for each sample were calculated and used to determine relative expression levels for IGL and IGK. The Pfaffl method (Pfaffl, 2001; 213) was used due to the calculated differences in reaction efficiencies between primer pairs. Fold change between IGL and IGK were determined using the calculated primer pair efficiency to the power of the target gene CT minus the reference gene as a calibrator between samples, shown in the equation below.

#The Pfaffl method for relative gene expression

Fold change= ((λ primer efficiency) ^ (Ct $_{\lambda}$ - Ct $_{REF}$)) / ((κ primer efficiency) ^ (Ct $_{\kappa}$ - Ct $_{REF}$))

Percentage expression = (1- Fold change) * 100

The percentage expression levels of IGL and IGK were then calculated for each sample by measuring the fold change of *IGLC* and *IGKC* expression in each sample without a calibrator gene. The Pfaffl method was adjusted to take reaction efficiency of primer pairs into consideration.

Fold change= ((λ primer efficiency) ^ (Ct $_{\lambda}$)) / ((κ primer efficiency) ^ (Ct $_{\kappa}$))

Percentage expression = (1- Fold change) * 100

The relative percentage expression of IGL in the total light chain transcriptome was plotted in a bar chart. Error bars were calculated using standard deviation in Microsoft Excel. Statistical significance between time points was determined using a one-tailed Wilcoxon rank-sum test on the hypothesis that the ratio of lambda: kappa light chain expression would change over time and upon challenge with FMDV. This test was chosen instead of the t-test as it does not assume normality.

4.2.15 Expression analysis of the IGL and IGK genes using RNA-seq

RNA reads from the whole transcriptome of three Holstein animal were kindly donated by Pasman et al at the University of Guelph for RNA-seq analysis of the Ig loci, as outlined in Chapter 3, section 3.2.8 (Pasman et al., 2017; 190). African buffalo reads were generated by sequencing a RACE library of IGL RNA reads with the 5' *IGLC* primer at Day 0 and Day 14 for animals challenged with FMDV SAT1 or SAT2 virus (outlined in Chapter 5, section 5.2.6). The RNA reads from cattle and African buffalo were subsequently mapped to the ARS-UCDV0.1 cattle assembly or the concatenated African buffalo *IGLV* gene assemblies respectively with the GEMmapper. Expression percentiles for IGL usage were calculated based on the percentage of their total weighted counts.

4.3 Results

4.3.1 The structure and organisation of the cattle IGL

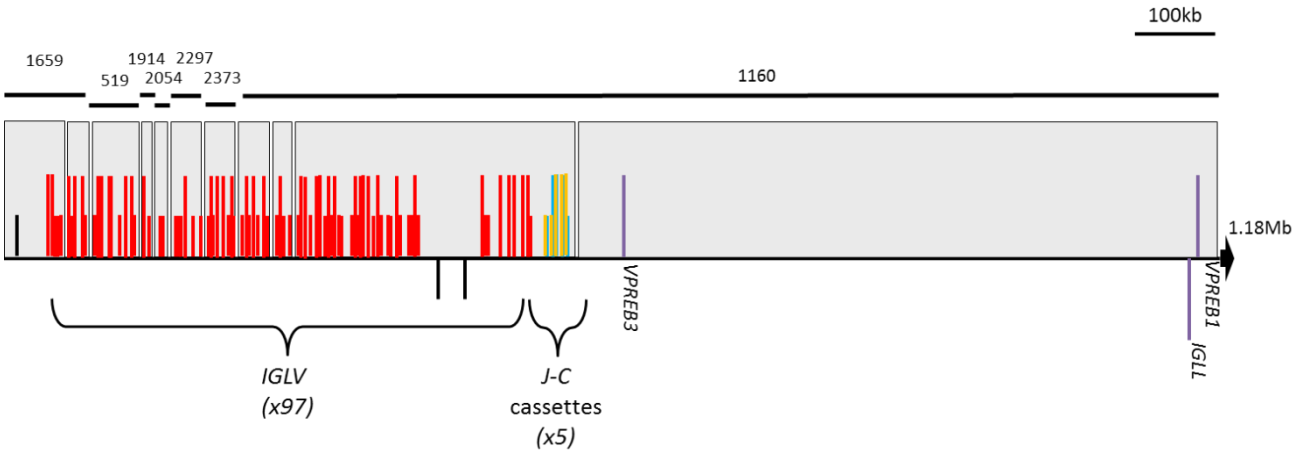


Figure 4.1: Schematic organisation of the cattle IGL long read PacBio ARS-UCDv0.1. The IGL spans approximately 635 kb on ten contigs representing seven scaffolds. Individual scaffolds are indicated by the black lines above the annotation and named accordingly. The second contig of the largest scaffold (1160) contains the surrogate light chain genes downstream of the *IGLC* and continues for a further 1.18 Mb downstream, as indicated by the arrow. Putatively functional gene segments are indicated with a long projection line and pseudogenes have a short projection line. *IGLV* are red, *IGLC* are blue and *IGLJ* are orange. The *SLC5A4*, *ZINC* and *PRAME* gene are shown with a black projection line. Scale bar: 100 kb.

The cattle lambda locus in the long-read genome assembly ARS-UCDv0.1 is found on chromosome 17 and consists of seven scaffolds, made up of ten contigs in total, which span approximately 1.45 Mb (Figure 4.1). Due to the high degree of sequence similarity across the *IGLV* region, it is difficult or impossible to determine whether or not sequences between contigs overlap, or if artificial duplications exist. However, a likely erroneously inverted contig, near the 5' end of the locus in the ARS-UCDv0.1 assembly, appears as though it may represent such an artificial duplication due to its high degree of similarity with the adjacent regions (Figure 4.3A). No other such obvious artificial expansions were identified. Besides scaffolds 1659 at the 5' end and 1160 at the 3' end of the locus, the order of the five internal scaffolds containing *IGLV* gene segments could not be ascertained and may alter in future assemblies.

A total of 97 *IGLV*, five *IGLJ-IGLC* cassettes and the surrogate light chain genes *VPREB3*, *VPREB1* and *IGLL* were identified in the ARS-UCDv0.1 (Appendix Table 4). The *IGLV* are

organised into two major clusters by a ~83 kb region that is devoid of *IGLV* gene segments and contains the genes *ZNF280B* and *PRAME*. Flanking the 5' end of the locus, 30 kb upstream, is the gene *SLC5A4*. This organisation is characteristic of the lambda locus in multiple species including pig (Schwartz and Murtaugh, 2014; 214). An additional truncated *IGLV* orphan gene segment, *IGLV1*/OR28-1, is located on chromosome 28.The *IGLJ-IGLC* cassettes at the 3' end of the locus are 42 kb downstream of the *VPREB3* surrogate light chain gene. The *IGLL* and *VPREB1* surrogate light chain genes are a further ~0.7 Mb downstream.

4.3.2 Structural comparisons of the IGL in the ARS-UCDv0.1 and the reference genome assemblies

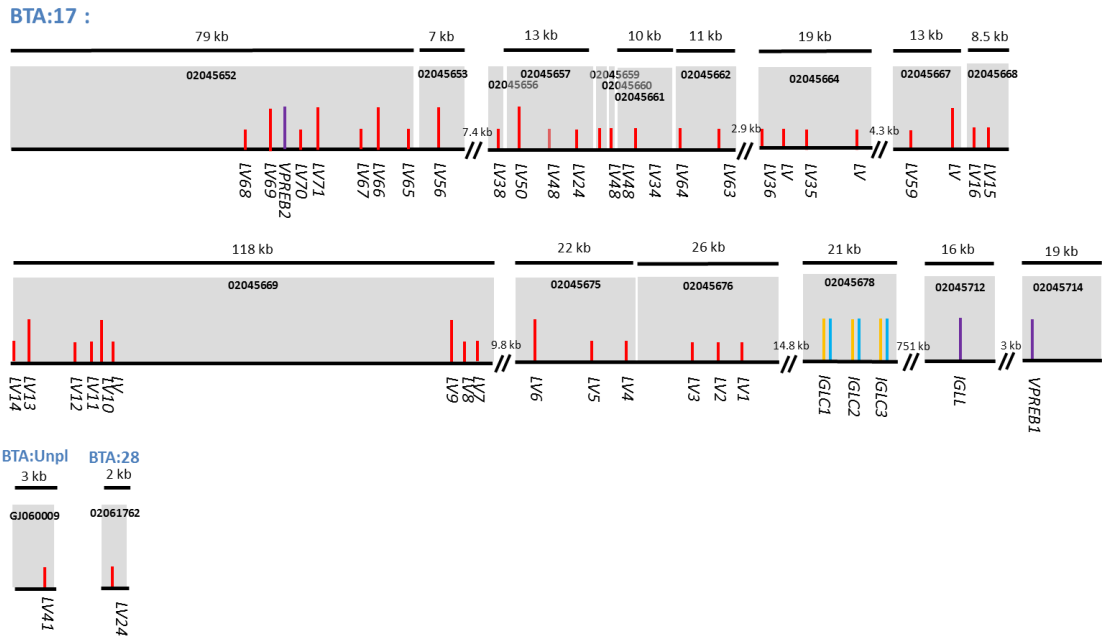


Figure 4.2: Schematic organisation of the lambda light chain gene segments in the UMD3.1 cattle reference genome. The majority of the IGL is assembled on 17 contigs on chromosome 17 (contigs DAAA02045652-714) and are displayed here in the order they are assembled in the genome from left to right. An individual *IGLV* is assembled on chromosome 28 and on an unplaced contig GJ060009. Sequences are annotated based on their nucleotide sequence similarity to equivalent gene segments identified in the ARS-UCDv0.1 whilst gene segments with less than 97% sequence identity were left unnumbered. Putatively functional segments based on nucleotide sequence are displayed as long projection lines. *IGLV* are represented by red lines, *IGLJ* by orange and *IGLC* by blue. The surrogate light chain genes are purple. Scale bar: 10kb.

The assembly across the highly repetitive *IGLV* region is more contiguous in the ARS-UCDv0.1 compared to the UMD3.1 and Btau_3.1 reference assemblies and is assembled on fewer contigs, indicating that the ARS-UCDv0.1 is considerably improved. The previously published characterisation of the Btau_3.1 is heavily fragmented onto ten scaffolds, consisting of 37 contigs. From this assembly, Ekman et al (2009) described a total of 63 *IGLV*, 32 of which were on seven genomic scaffolds not assigned to a chromosomal location. The remaining 31 *IGLV* were found on two scaffolds, organised in two sub-clusters, with a similar structure to the ARS-UCDv0.1 scaffold 1160. The IGL is assembled on fewer contigs in the more recent reference assembly, the UMD3.1, 19 contigs in total, the majority of which are assembled on chromosome 17 with two *IGLV* each assembled on BTA 19 and 28 (figure 4.2). A total of 45 *IGLV* are assembled in the UMD3.1, the majority corresponding to gene segments outlined in the ARS-UCDv0.1. A dot plot comparison of the ARS-UCDv0.1 to the UMD3.1 however reveals a considerable amount of sequence information missing from the reference genome which contains large sequence gaps and assembly errors (Figure 4.3A). The IGL in the ARS-UCDv0.1 therefore, appears the most complete characterisation to date.

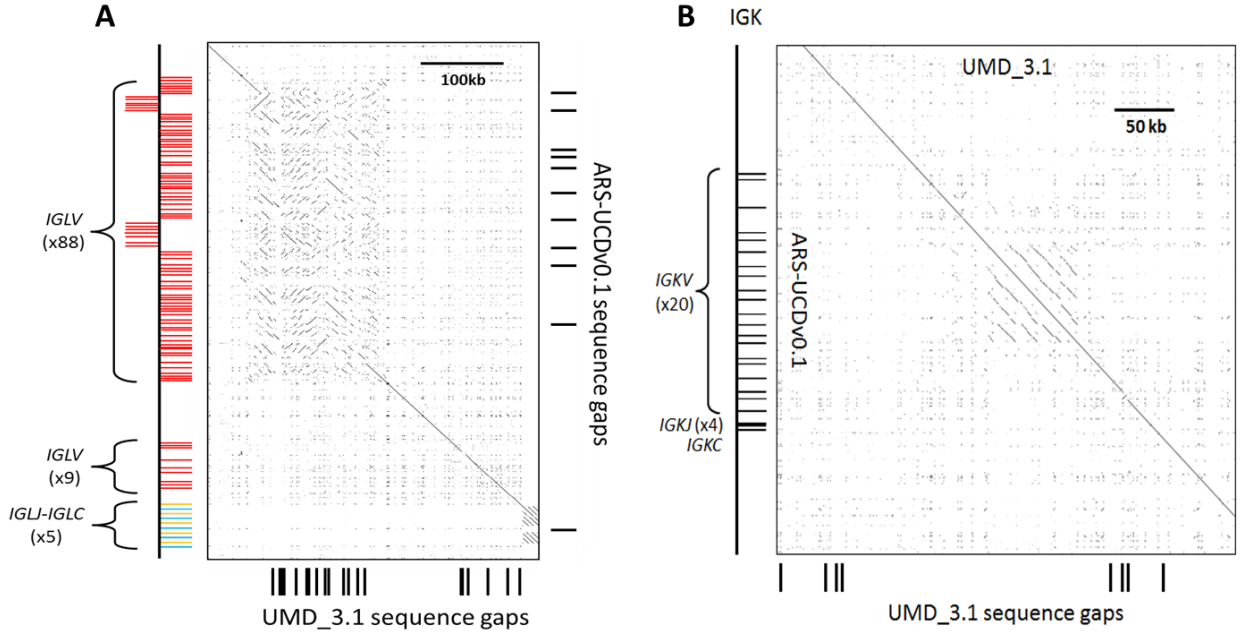


Figure 4.3: Recurrence plots of the ARS-UCDv0.1 assembly against the UMD3.1 cattle genome for IGL (A) and IGK (B). The ARS-UCDv0.1 assembly annotation is displayed on the left with sequence identity on the y-axis. The UMD3.1 sequence identity is along the x-axis with the black bars on the bottom indicating sequence gaps in the assembly. The black bars to the right of A indicates the breaks between individual contigs in the lambda locus of the ARS-UCDv0.1 assembly.

4.3.3 The structure and organisation of the cattle IGK

Unlike the IGL, the IGK is contained on a single 16.3 Mb contig in the ARS-UCDv0.1 assembly. A total of 20 *IGKV*, 4 *IGKJ* and a single *IGKC* gene were identified in the locus by our group, 1 kb upstream of the kappa deleting element RS (Schwartz et al, unpublished). When the ARS-UCDv0.1 is compared to the reference genome assemblies, UMD3.1 and Btau_3.1, overall structure and organisation of the loci are similar, except for sequence gaps in the reference assemblies. Btau_3.1 contains 10 sequence gaps within the *IGKV* region; this contig misassembly appears to have artificially duplicated four *IGKV*, while a substantial sequence gap at the 3' end obscures the presence of two additional *IGKV* gene segments. Thus, the cattle IGK locus contains a total of 20 *IGKV* gene segments, rather than the 22 previously reported (Ekman 2009). Three sequence gaps are present in the UMD3.1 assembly in the vicinity of *IGKC*, whereas no gaps across the IGK locus exist within the ARS-UCDv0.1 assembly (Figure 4.3B), showing the structural characterisation is slightly improved in the ARS-UCDv0.1.

4.3.4 Assembly of the African buffalo IGL

Illumina reads from the African buffalo genome project were mapped to the cattle ARS-UCDv0.1 IGL scaffold 1160 and subsequently *de novo* assembled in SPAdes into the entire IGL locus and as targeted assembly of individual gene segments. As anticipated, complete assembly was impossible due to the highly repetitive nature of the locus and the short sequence length of the Illumina reads. Genome reads that mapped to the ARS-UCDv0.1 scaffold assembled into a heavily fragmented assembly; a total of 42 contigs were assembled, the largest being 7488 bp with an N50 of 4410 bp. Coverage across the assembly was on average 25 x but with highest coverage in the intronic regions. The assembly spanned 180.7 kb but contained only 17 *IGLV* gene segments and 3 *IGLJ-IGLC* cassettes. Due to the difficulties in assembling the complete African buffalo IGL, a dot plot assembly comparing the African buffalo IGL to the cattle IGL was highly fragmented and did not show a clear relationship between the two assemblies.

The IGL gene segments in the African buffalo were then individually *de novo* assembled in SPAdes with reads mapped to individual cattle *IGLV* in the ARS-UCDv0.1. A total of 58 *IGLV* were assembled; 36 *IGLV* were assembled from reads with 4% alignment error and a further 12 *IGLV* were found when alignment error was increased to 10%. Concatenation of the assembled *IGLV* revealed a total length of 123 kb, An N50 of 61.8 kb with a GC content of 50.96. The average coverage of the *IGLV* was 10.2x and the contig containing the *IGLJ-IGLC* cassettes was 6x (Table 4.1).

African buffalo genome sequencing reads did not map to the remaining six scaffolds containing *IGLV* in the ARS-UCDv0.1 including to the *IGLV* gene segments contained within these scaffolds. Subsequent assembly of the whole locus or of individual gene segments contained on these scaffolds was not possible. This could be due to the IGL in African buffalo being less expanded in cattle and therefore missing these gene segments. Alternatively, the highly repetitive nature of these gene segments meant that either mapping of the short Illumina reads to this region was difficult or the locus may have been artificially expanded in the cattle ARS-UCDv0.1.

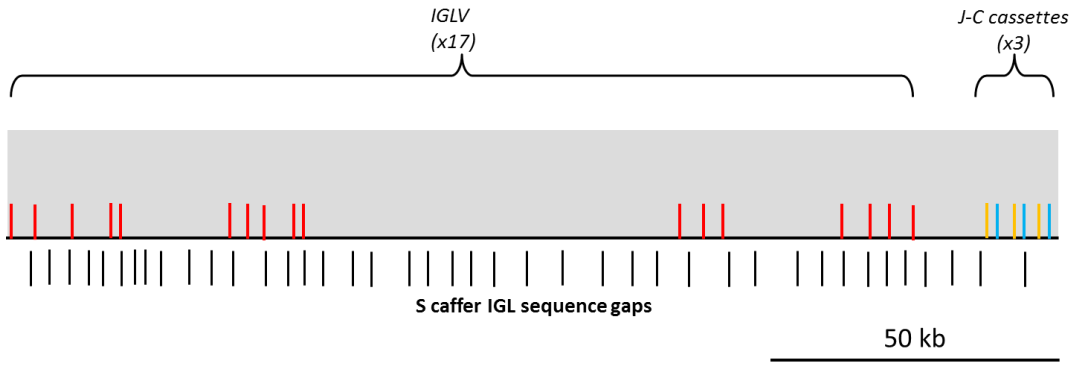


Figure 4.4: Schematic organisation of the African buffalo IGL *de novo* assembled in SPAdes using paired end reads that mapped to the cattle ARS-UCDv0.1 IGL scaffold 1160. The assembled region spans 181 kb, consisting of 42 contigs indicated by the black projection lines below the annotation. Red lines indicate an *IGLV* gene segment, orange lines indicate *IGLJ* and blue *IGLC*. Scale bar: 50kb.

Buffalo Gene name	Contigs	Coverage	Assembly GC	Contig length	Mismatch
IGLVa-1(I)	1	2.2	56.2	780	10
IGLVa-2(I)	1	50	54.9	650	
IGLVa-3(I)	1	1.7	60.6	536	10
IGLVa-4(I)	2	3.3	56.7	811	
IGLVa-5(I)	1	11.2	57.6	968	
IGLVa-6(I)	1	7.2	57.3	1039	
IGLVa-7(I)	1	24.5	58.4	1146	
IGLVa-8(I)	1	5	59.5	993	
IGLVa-9(I)	1	10.2	57.5	602	
IGLVa-10(I)	2	0.5	59.5	545	10
IGLVa-11(I)	1	9.1	57.1	1061	
IGLVa-12(I)	1	27.2	56.6	947	
IGLVa-13(I)	1	11.9	56.2	582	
IGLVb-1	1	21.6	61.2	1127	10
IGLVb-2	1	2.1	60.3	966	10
IGLVb-3	1	3	58.8	600	
IGLVb-4	1	0.6	60.8	600	
IGLVb-5	1	3.1	59.1	745	10
IGLVb-6	3	2	59.2	1201	
IGLVb-7	1	0.9	58.5	701	
IGLVb-8	1	2.2	61.5	774	
IGLVb-9	1	2.9	58.7	572	
IGLVb-10	1	10.3	57.3	1062	
IGLVb-11	1	8.1	60.5	868	
IGLVc-1	1	13.4	49.8	1214	
IGLVc-2	1	10	54.4	1148	
IGLVc-3	1	13.9	57.3	1300	
IGLVc-4	2	8.3	56	1085	
IGLVd-1	1	20.2	63.8	1037	
IGLVd-2	1	10.8	63.6	965	
IGLVd-3	1	28.5	61.3	1229	10
IGLVd-4	1	11	60.6	1189	
IGLVe-1	1	10.4	65.2	1384	
IGLVe-2	1	7.9	62.5	1234	
IGLVe-3	1	8.7	64	1182	
IGLVe-4	1	11.3	62.9	894	
IGLVe-5	1	10	61	1112	10
IGLVf-1	1	6.6	60.8	926	
IGLVf-2	2	20.5	58.9	754	
IGLVf-3	1	8.4	61	1132	
IGLVf-4	1	12.8	59	1174	
IGLVf-5	1	24.2	52.6	1467	
IGLVf-6	1	3	54.5	765	10
IGLVf-7	1	15.9	52.9	1449	
IGLVf-8	1	4.9	53.2	847	
IGLVf-9	1	4	56	489	
IGLVf-10	1	2	54.7	737	10
IGLVf-11	1	2.4	52	665	10
IGLVg-1	1	2.5	57.8	384	
IGLVg-2	1	13.6	57.7	1111	
IGLVg-3	1	10.8	56.3	837	
IGLVg-4	1	3.3	58.1	640	
IGLVg-5	1	9	58.6	1215	
IGLVg-6	1	13.5	60.1	781	
IGLVg-7	1	11.8	60	1036	10
IGLVg-8	1	15.8	59.5	1108	
IGLVg-9	1	16.2	58.4	969	
IGLVg-10	1	7.4	57.2	749	
IGKVa-1	1	15.5	52.4	1027	10
IGKVa-2	1	17.4	48.4	1076	10
IGKVa-3	1	15.2	43.7	616	10
IGKVa-4	1	18.6	50.2	944	10
IGKVa-5	1	10.4	56.9	695	10
IGKVa-6	1	18.4	49	1061	10
IGKVa-7	1	16.3	52.9	1056	10
IGKVb-1	1	10.6	48.6	1315	10
IGKVb-2	1	7.2	46.4	1242	10
IGKVb-3	1	5.2	46.1	1212	10
IGKVb-4	1	10.7	45.6	1177	10
IGKVb-5	1	6.1	48.2	1341	10
IGKVb-6	1	6.2	47.8	1141	10
IGKVb-7	1	18.2	47.7	1331	10
IGKVb-8	1	9.8	48.8	1108	10
IGKVb-9	2	8.13	46.8	1073	10
IGKVb-10	1	13	47.3	1312	10
IGKVb-11	1	20.2	46	1366	10
IGKVb-12	1	20.3	47.3	1235	10
IGKVb-13	1	17.4	47.4	815	10

Table 4.1: Mapping statistics of *IGLV* and *IGKV* assembled in African buffalo with whole genome reads mapped to the cattle IGL and IGK loci in the ARSUCDv0.1. A total of 58 *IGLV* and 20 *IGKV* were assembled which formed seven and two phylogenetic sub-groups respectively, designated by lower case letters.

4.3.5 Assembly of the African buffalo IGK

The African buffalo IGK was assembled by mapping genome reads to the cattle ARS-UCDv0.1 IGK loci and *de novo* assembling both the entire locus and individual gene segments in SPAdes. The assembly of the entire IGK locus was fragmented on 38 contigs, the largest contig being 10.7 kb, with an N50 of 5350 bp. The IGK contained a truncated *IGKC* gene segment, missing the first 9 bp, and 14 *IGKV* gene segments (Figure 4.5). Average coverage across the locus was 15.74 x. A total of 20 *IGKV* were found when individually assembled (Table 1), with the corresponding gene segment in the whole locus assembly having a 98-100% nucleotide sequence identity. The one *IGKC* and four *IGKJ* were each assembled on a single contig with a 99% nucleotide identity to the corresponding gene segments in the whole locus assembly. Concatenation of the IGK gene segments spans 24.5 kb with an N50 of 1177 and an average coverage of 13.4 x. IGK in African buffalo appears to be the most complete whole immune loci assembly achieved with the short-read Illumina sequences, although it is still heavily disrupted. A dot plot comparison of the African buffalo and cattle ARS-UCDv0.1 IGK reveals heavy fragmentation in the African buffalo assembly across the variable region, as is expected, but a nearly contiguous assembly at the 3’ end of the locus.

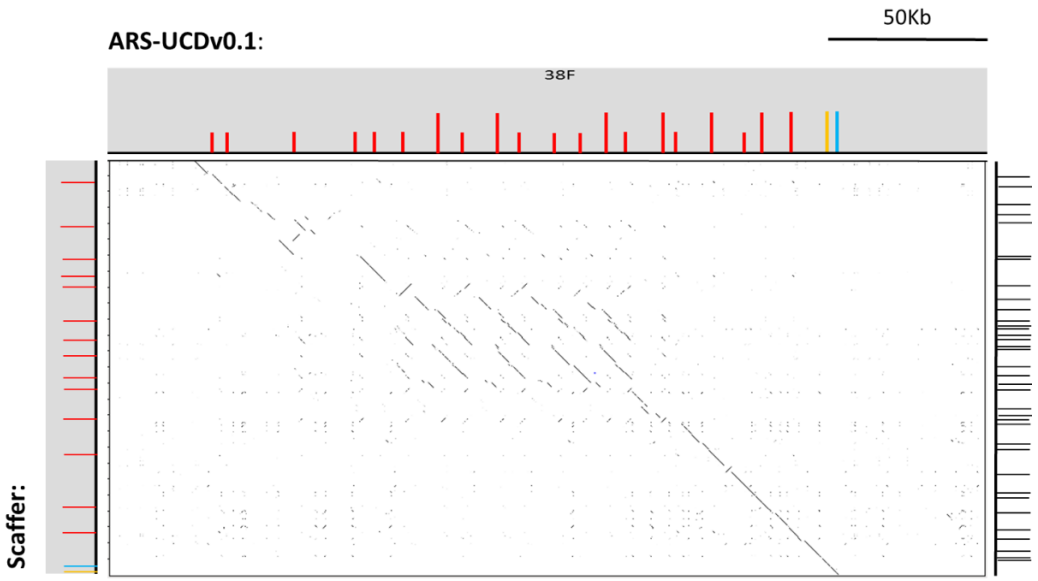


Figure 4.5: A dot plot comparison of the ARS-UCDv0.1 IGK locus and the African buffalo IGK *de novo* assembled as an entire locus with reads mapped to the ARS-UCDv0.1. Red lines indicate an *IGKV* gene segment, yellow indicates the *IGKC* and the blue line indicates the cluster if *IGKJ*; cattle and African buffalo both possess four *IGKJ*. The IGK in the ARS-UCDv0.1 is contained on a single contig whilst the African buffalo assembly is on 38 contigs, the breaks indicated by the black lines on the right.

4.3.6 Predicted expression of putatively functional IGL and IGK genes

Of the *IGLV* in the cattle ARS-UCDv0.1, 32 are putatively functional (figure 4.6) whilst the remaining 65 are pseudogenes (Appendix Table 4). The preferential usage of 23 of the putatively functional gene segments can be predicted as they possess canonical octamer and RS sequences. In the UMD3.1, 11 of the 45 *IGLV* are putatively functional and correspond to the putatively functional gene segments in the ARS-UCDv0.1. Ekman et al (2009) characterised 63 *IGLV* of which 25 were putatively functional; the majority have corresponding gene segments with greater than 97% nucleotide identity to *IGLV* in the ARS-UCDv0.1. Resolution of some of the corresponding gene segments was not possible as their sequences were highly similar and so several of the gene segments described in the UMD3.1 (Ekman et al., 2009; 99) appear multiple times in the ARS-UCDv0.1 annotation. The number of putatively functional gene segments in the ARS-UCDv0.1 is therefore greater but without additional diversity. Putatively functional gene segments missing in the new assembly comprise *IGLV35*, *IGLV39* and *IGLV59*. The remaining putatively functional gene segments

described by Ekman et al are found, along with an additional six undescribed gene segments, and two putatively functional *IGLV* that Ekman described as ORF. Furthermore, four of the *IGLV* described by Ekman et al as putatively functional are pseudo genes in the ARS-UCDv0.1 assembly; two of these only contain single indels causing frame shifts but the remaining two contain multiple frame shifts and are heavily disrupted.

In African buffalo, 20 of the 57 individually assembled *IGLV* are putatively functional (Figure 4.6) with 16 having canonical octamer and RS sequences therefore predicting their preferential expression (Appendix Table 5). Analysis is provisional based on short read sequencing data but ten of the 20 putatively functional gene segments have functional homologs in cattle. The remaining ten putatively functional *IGLV* in African buffalo have lower than 97% nucleotide identity to a corresponding cattle gene segment. CDR1 lengths in cattle and African buffalo have limited variation between sequences; length is restricted to a short hairpin loop of eight amino acids although variation in sequence is seen between each sub-group of gene segments. CDR2 are only three amino acids in length in cattle and African buffalo with differences in amino acid composition only seen between sub-groups of gene segments.

In the ARS-UCDv0.1 cattle IGK locus, Schwartz et al in our Immunogenetics group, showed that seven of the 20 defined *IGKV* are putatively functional and one is an ORF (Appendix Table 5). Five of the putatively functional *IGKV* possess canonical splice sites, octamer, and RS, suggesting their preferential usage. In the African buffalo, ten of the individually assembled *IGKV* are putatively functional and eight of these are assembled in the entire IGK locus assembly and are putatively functional. The preferential expression of five of the *IGKV* can be predicted from canonical splice sites, octamer sequence and RS.

Figure 4.6: IMGT protein display of putatively functional cattle *IGLV* in the ARS-UCDv0.1 and *de novo* assembled African buffalo *IGLV*. Three of the African buffalo *IGLV* were incompletely assembled with missing sequence information indicted by dashes, but appear functional. Sequences are organised to indicate their leader, framework (FR) and complementarity determining regions (CDR). Identical amino acids are represented with a dot whilst absent amino acids in the IMGT framework are shown by a dash in all of the sequences.

4.3.7 Phylogenetic analysis of IGL and IGK

The *IGLV* in cattle and African buffalo can be grouped into six phylogenetic sub-groups which correspond to human and mice *IGLV* sub-groups in the IMGT database. Sub-group one, part of clan I, in the cattle ARS-UCDv0.1 and African buffalo gene assemblies is the largest and contains the majority of putatively functional gene segments. *IGLV* belonging to sub-group 1 in both cattle and African buffalo form two separate sub-clusters which are undefined in the IMGT database, but which here are separated by the designated nomenclature of the African buffalo *IGLV* as subgroup a and b (figure 4.7). Sub-group 3, belonging to the next closely related clan II contains four putatively functional cattle *IGLV* and four African buffalo *IGLV*. Two additional African buffalo *IGLV* belonged in clan II but could not be assigned to a sub-group as they were less than 75% similar to the human and mouse database members. In cattle, one putatively functional *IGLV* belongs to sub-group 2, clan I but none were putatively functional in the African buffalo. Cattle also contain only one member of clan V, *IGLV*(V)-77, which is a heavily disrupted pseudogene and none of which were found in African buffalo.

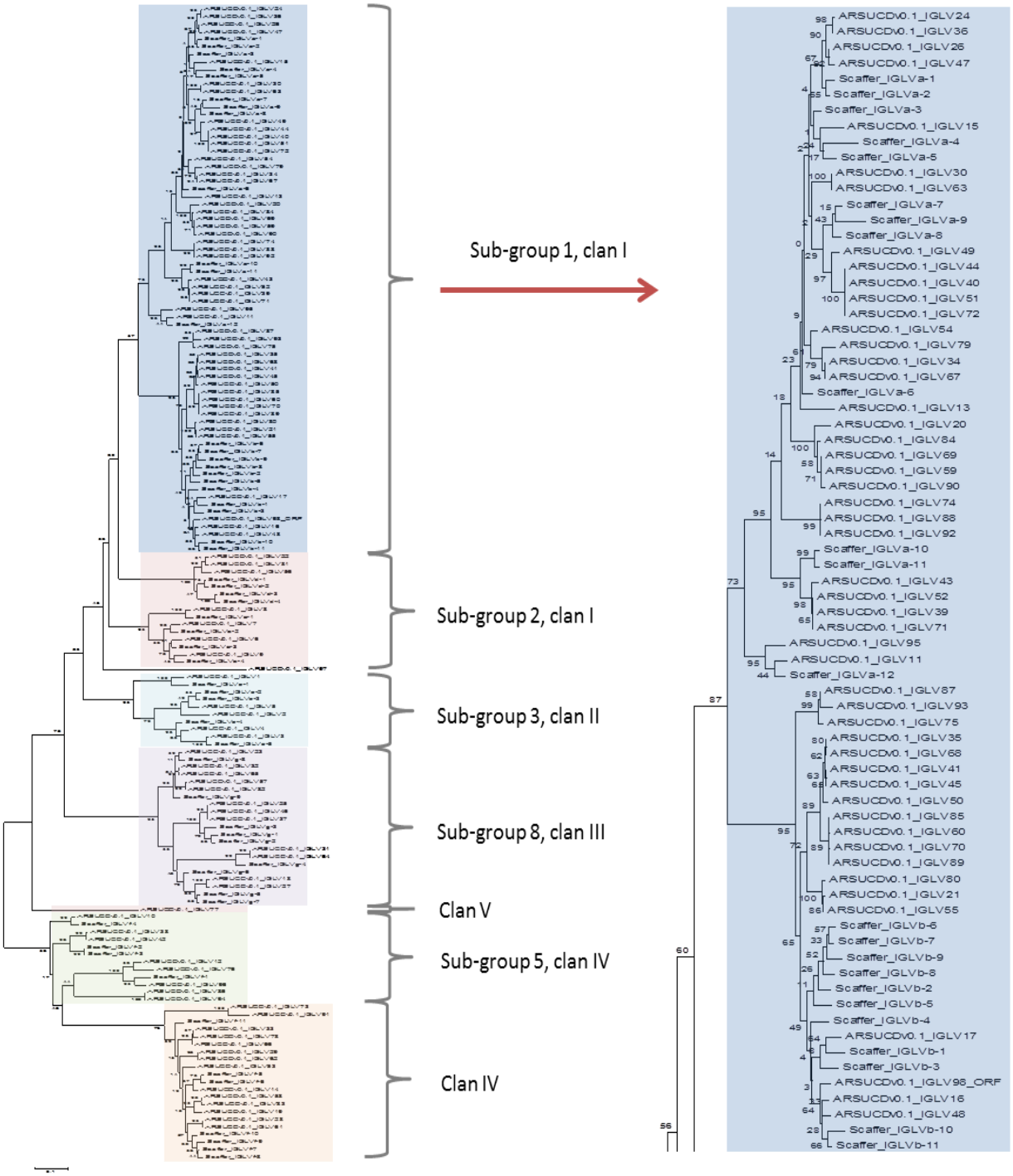


Figure 4.7: Phylogenetic analysis of the cattle and African buffalo *IGLV* in the ARS-UCDv0.1 and the *de novo* targeted assembly on gene segments in African buffalo. The *IGLV* cluster in six sub-groups designated based on their nucleotide sequence identity to human and mouse *IGLV* in the IMGT database, with a 75% threshold value. Gene segments with less than 75% identity were assigned to their higher order clan. *IGLV* were assembled in African buffalo in all the sub-groups seen in cattle except for sub-group 7, in which only *IGLV*(V)-77 in cattle belongs. The sub-group 1, clan I, is the largest and contains the majority of putatively functional gene segments in both cattle and African buffalo. Sub-group 1 has two phylogenetically distinct groups in both cattle and African buffalo.

Cattle possess five *IGLC-IGLJ* cassettes in the ARS-UCDv0.1 separated each by 5.7-5.9 kb intronic region, whilst three cassettes were assembled in the African buffalo as individual contigs. The phylogenetic alignment of the five cattle *IGLC* suggests they arose through duplication of three original gene segments in their common ancestor but these did not group with each of the three African buffalo *IGLC* (figure 4.8A). Alignment of the *IGLJ* in cattle and African buffalo confirmed the three cattle sub-groups of *IGLC* (figure 4.8B). The *IGLJ5* in cattle aligned perfectly with African buffalo *IGLJ3*. The phylogenetic analysis then, suggests that the five cattle *IGLC-IGLJ* cassettes arose from three cassettes in their common ancestor with African buffalo but if these gene segments were assembled in the African buffalo the their sequence has diverged considerably. Most likely, the assembly of the African buffalo *IGLC-IGLJ* cassettes is inaccurate due to bias introduced by SNPs in the mapped reads.

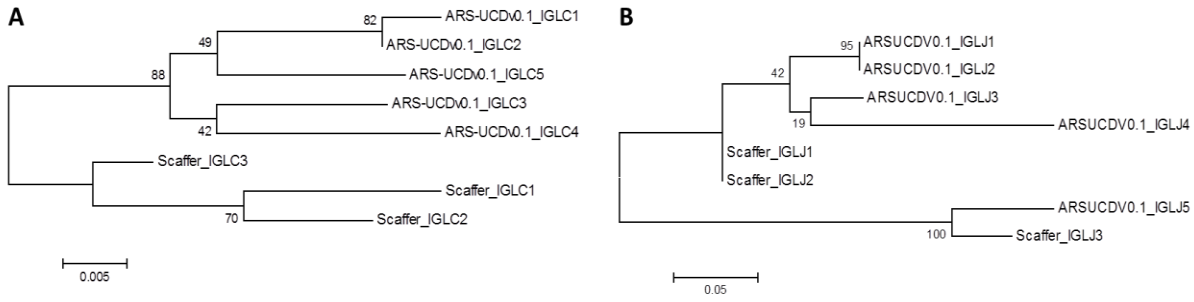


Figure 4.8 Phylogenetic analysis of the *IGLC* (A) and *IGLJ* (B) in the cattle ARS-UCDV0.1 and the individual *de novo* assembly of the African buffalo gene segments with reads mapped to the ARS-UCDv0.1. The *IGLC1-5* in cattle and *IGLC1-3* in African buffalo were aligned using ClustalW multiple alignment and the *IGLJ1-5* in cattle and *IGLJ1-3* in African buffalo. Each phylogenetic tree was produced in MEGA6.0 (Tamura et al., 2013; 188) using maximum likelihood with the Tamura and Nei model (Tamura, 1992; 191) with 1000 bootstrap iterations.

The total *IGKV* is less than the *IGLV* and so is grouped into three phylogenetic sub-groups. Three of the putatively functional cattle *IGKV* in the ARS-UCDv0.1 belong to sub-group one whilst two are in African buffalo. The remaining four and five putatively functional *IGKV* in cattle and African buffalo respectively belong to sub-group 2. Only one *IGKV* gene segment belongs to sub-group 3 in cattle and none were assembled that belong to sub-group 3 in African buffalo.

Surrogate light chain genes *VPREB1*, *VPREB3* and *IGLL* were identified in the ARS-UCDv0.1 and assembled as single contigs in the African buffalo. None possess a canonical RS sequence but *VPREB1* and *VPREB3* structure resembles the *IGLV* genes with a leader and exon sequence. Evolutionary divergence of the surrogate genes occurred before the speciation of cattle and African buffalo as each gene is more closely related between species than to each other.

4.3.8 Validation of qPCR primers

	Cattle		African buffalo	
	Exponential amplification	% efficiency	Exponential amplification	% efficiency
B-actin	1.92	96	1.9565	97.83
PPIA	2.099	104.95	2.1933	109.67
SDHA	2.021	101.05	1.947	97.35
IGLC	1.769	88.45	1.754	87.72
IGKC	2.087	104.35	2.0974	104.87

Table 4.2: Primer exponential amplification co-efficient used for calculating fold change of IGL and IGK expression in cattle and African buffalo and their primer efficiency.

Following the MIQE guidelines (Johnson et al., 2014; 212), a panel of reference genes were validated to ensure that expression levels did not vary significantly between individual samples and were not affected by the different RNA isolation techniques. *Beta-actin*, *PPIA*, and *SDHA* were chosen as their sequence did not vary between species and may be suitable as references of cattle and African buffalo B cells. Specificity of these and the IGL and IGK primers was confirmed by PCR amplification from cattle genomic DNA and subsequent Sanger sequencing. Primer concentration was optimised at 300 nM and primer efficiency was then calculated using four 10-fold serial dilutions of each primer on multiple cattle or African buffalo samples with a normalised concentration of 100 ng/µl. All three reference genes were within the desirable range of 90-110% efficiency (Table 4.2). Kappa primers were also within the desirable range but lambda had a slightly lower efficiency of 87-88%, potentially due to its redundancy to multiple *IGLC* genes. The subsequent analysis methods were therefore adapted to account for this.

GeNorm software (Biogazelle) was used with the serial dilution information to determine which reference gene was most stable for each assay. The software algorithm assigns an 'M' value to each gene, a value of 0.5 and below being considered desirable (Figure 4. 9). The reference gene *PPIA* was most stable for the cattle samples whilst *SDHA* was chosen for the African buffalo samples, although all three genes would have been suitable for either.

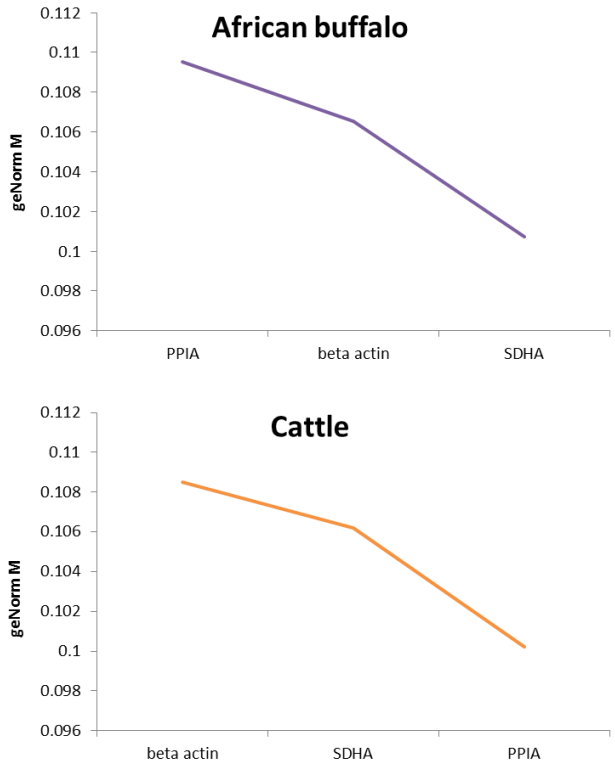


Figure 4.9: Reference gene primer sets optimised for cattle and African buffalo were assessed with QBASE plus (Biogazelle) using the algorithm for primer stability. The GeNorm M scores were plotted for SYBR green chemistry. M values less than 0.5 are defined as suitable for experimentation.

4.3.9 Validation of qPCR SYBR green master mix

Three SYBR green master mixes containing low ROX from three manufacturers Sigma, Applied Biosystems and Thermo Scientific were tested on the same assay, using aliquots from a single master mix for each optimised primer on two cattle and two African buffalo cDNA samples. The fold change of IGL and IGK expression was calculated for each master mix and plotted (figure 4.10). Applied Biosystems produced a larger standard deviation for sample CT scores and IGL expression with Sigma differed substantially from the other two

products. Luminaris Color HiGreen Low ROX qPCR Master Mix from Thermo Scientific appeared the most reliable of the three and was chosen for all subsequent assays.

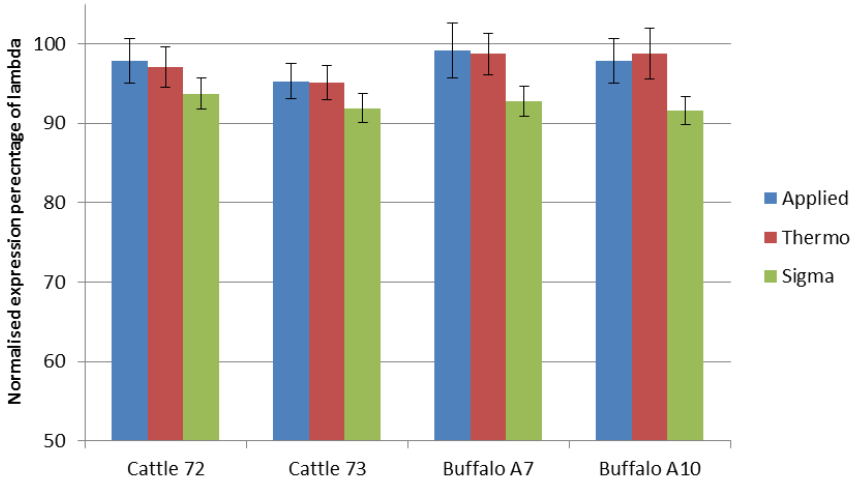


Figure 4.10: The percentage of IGL in the cattle and African buffalo light chain repertoire measured in a qPCR assay with three different manufacturers SYBR green products: Power SYBR Green PCR Master Mix from Applied Biosystems, Luminaris Color HiGreen Low ROX qPCR Master Mix from Thermo Scientific and KiCqStart SYBR green qPCR readymix Low ROX from Sigma. Each reaction was run in triplicate with identical cDNA and primer master mix. Standard deviation between each sample was plotted on each bar.

4.3.10 Cattle and African buffalo express predominantly lambda light chain

The relative contribution of *IGLC* and *IGKC* to the total light chain transcriptome was determined using the optimised SYBR green qPCR assays. The dominance of IGL in cattle, as shown in previous studies (Arun et al., 1996; 196, Murphy, 2008; 199) was confirmed. PBMC were collected weekly for a period of six weeks and IGL and IGK expression was calculated from isolated cDNA. IGL is predominantly expressed in ~95% of cattle light chain transcripts and whilst small variation is seen between animals and within each animal over time, these differences are not significant (figure 4.11A). IGL expression in African buffalo was measured at Day 0 (Figure 4.11C). African buffalo express ~98% of IGL in their antibody light chain repertoire with insignificant variation between animals.

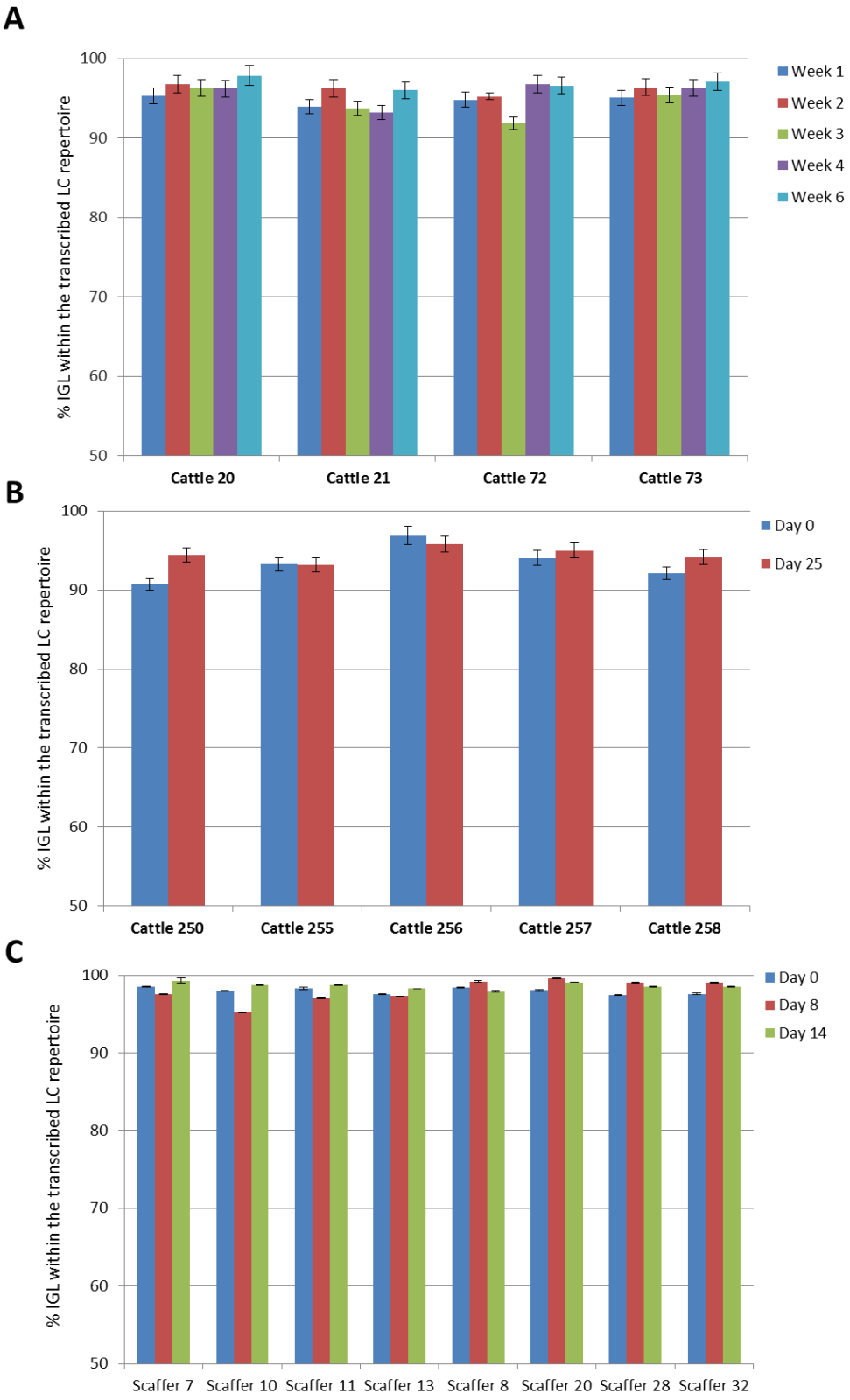


Figure 4.11: The percentage of IGL in the light chain transcriptome of cattle and African buffalo. Changes in IGL expression were measured in four Holstein cattle animals without deliberate infection or inoculation over a period of six weeks (A). Six Holstein cattle were inoculated with SAT1 FMDV and IGL expression was measured at day 0 and then day 25 post-vaccination, with Cattle 250 as a negative control (B). The expression ratio of IGL in African buffalo after challenge with SAT1 or SAT2 FMDV was then investigated in eight animals at day 0 and both day 8 and day 14 post-infection (C).

The effect of FMDV vaccination in cattle was assessed in six Holstein animals. Cattle cDNA was isolated from day 0 and then post-vaccination with SAT1 FMDV in four cattle animals, 255, 256, 257 and 258 at day 25. Cattle 250 was also included and remained unchallenged as a control animal. Expression of IGL varied between 90-97% at day 0 and no significant change was observed in each animal after immunisation; IGL expression between time points in each animal varied within a 2% range with the largest change observed in the control animal. African buffalo were subsequently analysed for a change in lambda: kappa expression at Day 0 and then Day 8 and Day 14 post-infection with SAT1 or SAT2 FMDV (Chapter 5, section 5.2.1); animals 7, 10, 11 and 13 were challenged with SAT1 and animals 8, 20, 28 and 32 were challenged with SAT2. Like cattle, African buffalo IGL expression does not change significantly between time points and the variation in expression is within a ~2% range. The ratio of IGL expression then, in cattle and African buffalo, does not alter during an immune response after vaccination or challenge with FMDV.

4.3.11 RNA-seq mapping analysis in cattle and African buffalo

Functionality of the cattle IGL was determined with RNA-seq data shared by Pasman et al (2016) at the University of Guelph. Of the 17 million reads generated for each of the three Holstein cattle, an average of ~930000 reads (2.67%) mapped to the cattle *IGLV* region. 186,332 of the mapped reads were uniquely mapped to the gene segments (0.51%) whilst the remaining 743007 (3.77%) were ambiguously mapping (Figure 4.12A). Functionality of the African buffalo IGL was determined using IGL antibody transcripts from 5'RACE-PCR amplification at day 0 and after infection with SAT1 FMDV at day 8 and day 14. The IGL transcripts were sequenced in two animals with Illumina. From a mean 561000 transcripts for each RACE-library, 537000 reads mapped (95.85%) to the African buffalo assembled gene segments. Only 30,900 of the mapped reads (5.75%) mapped uniquely to the assembled *IGLV* whilst the remaining 94.25% of mapped reads were ambiguously mapping (Figure 4.13A).

The mappability of the cattle and African buffalo *IGLV* in the ARS-UCDv0.1 and assembled African buffalo gene segments was calculated to determine if the uniqueness of the genes biased the RNA-seq results; higher numbers of transcripts could map to more unique genes. In cattle, the average mappability was ~25%, however numerous gene segments at the

beginning of the locus were almost 100% unique (Figure 4.12B). This suggests these gene segments have not undergone either real or artificial duplications and so appear more unique relative to the other gene segments. However *IGLV20* had the most transcripts mapped and only had 47% uniqueness. In African buffalo, the average mappability was high, ~65%, but the gene segments with the most RNA reads mapped had nearly the lowest mappability scores. This is likely an artefact of the assembly of this region and a consequence of concatenating several genes together in the assembly. Any bias of the mapping was further accounted for by producing weighted counts of multi-mapping reads.

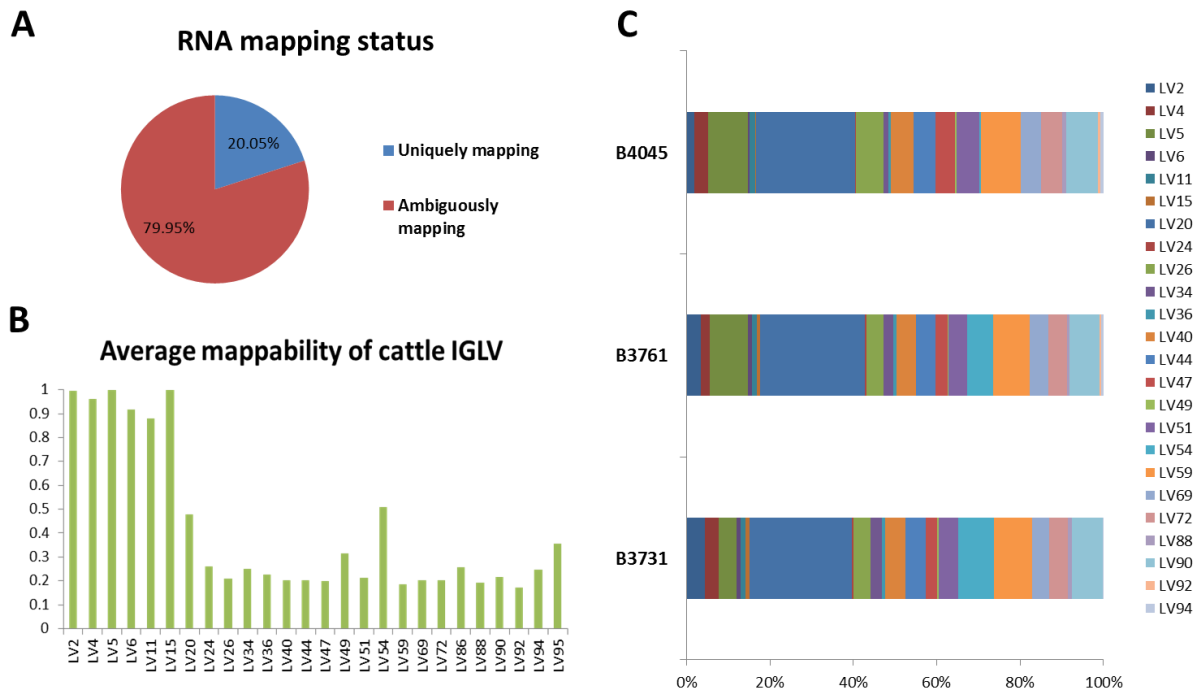


Figure 4.12: RNA-seq analysis of the *IGLV* in the cattle ARS-UCDv0.1. Average apportionment of reads from three Holstein cattle (B3731, B3761 and B4045) that mapped to the *IGLV* locus are displayed (A), with 186,332 reads (20.05%) uniquely mapping and 743007 reads (79.95%) ambiguously mapping. The mappability of the *IGLV*, the uniqueness of their sequence relative to the other gene segments, was calculated for the *IGLV* in the ARS-UCDv0.1 (B). The gene segments that RNA-seq reads mapped to are displayed only. RNA reads mapped to a total of 24 *IGLV*, with similar expression levels between the three cattle animals (C).

4.3.12 RNA-seq expression analysis in cattle and African buffalo

Of the 32 putatively functional *IGLV* gene segments in the cattle assembly ARS-UCDv0.1, 20 are predicted to be preferentially expressed due to canonical RS and octamer sequences. Of these 17 are transcribed in the three Holstein cattle animals (figure 4.12C). The remaining five predicted to be preferentially expressed do not appear in the Holstein transcriptome, possibly due to haplotypic variation between breeds or from unknown promoter elements affecting their transcription. The cattle reads mapped to a further six *IGLV* gene segments which were predicted pseudogenes: *IGLV15*, *IGLV47*, *IGLV69*, *IGLV90* and *IGLV92*. A single gene segment, *IGLV54*, was also transcribed despite having a non-canonical heptamer. These genes may also be haplotypic variants between the Hereford genome assembly but are most likely due to single indel errors contained in the ARS-UCDv0.1. The corresponding gene segments in the Btau_3.1 annotation by Ekman et al (2009) are functional and *IGLV54* has a canonical heptamer. The expression percentiles between the three cattle animals show similar expression levels of the 24 transcribed gene segments with a prevalence of *IGLV20*.

In African buffalo, RNA-reads mapped to 18 of the 21 putatively functional gene segments, of which all were predicted to be preferentially expressed with canonical octamer and RS sequences. *IGLVd-1*, *IGLVd-2* and *IGLVg-9* were predicted to be functional but were not transcribed. The dominance of two *IGLV*, *IGLVa*-7 and *IGLVa*-8 which have a 96% nucleotide sequence identity to each other, accounted for ~47% and 39% of the reads respectively. Dominance then of only two *IGLV*, both with low mappability scores, suggests an even more restricted light chain repertoire in African buffalo than cattle.

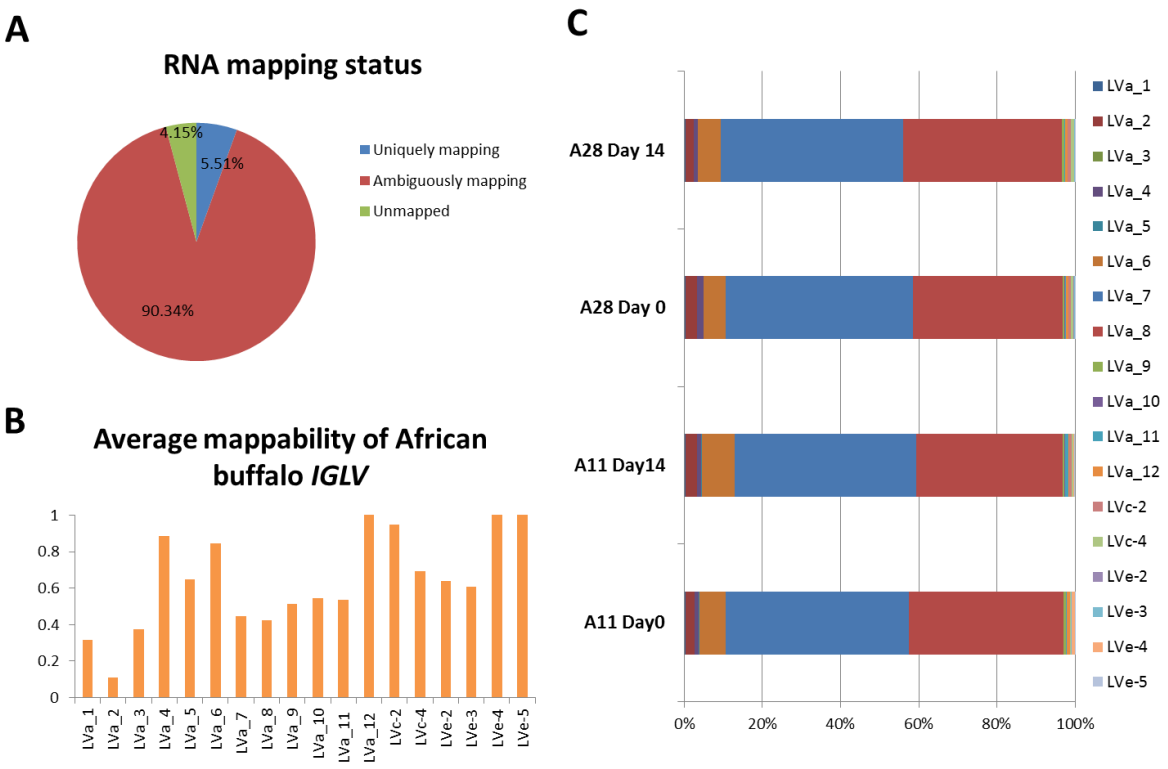


Figure 4.13: RNA-seq analysis of the *IGLV* in the African buffalo; IGL transcripts were sequenced using 5'RACE with an *IGLC* specific primer. These transcripts were subsequently mapped to the individually *de novo* assembled *IGLV* in African buffalo. The majority of reads mapped to the assembled gene segments, a total of 95.85% of the total, equating to 537,000 reads. The majority of reads mapped ambiguously (A).The mappability of the *IGLV*, the uniqueness of the gene segments relative to each other, was calculated; the gene segments with mapped RNA reads are displayed (B). Transcription of a total 18 *IGLV* in African buffalo occurs with similar expression levels between all three animals and a dominance of LVa-7 and LVa-8 (C).

4.4 Discussion

The structure of the cattle IGL and IGK in the PacBio long read assembly ARS-UCDv0.1 was characterised and compared to the reference genomes: UMD3.1, and the most recently published annotation, the Btau_3.1. Structural improvements in the ARS-UCDv0.1 are shown compared to both reference assemblies. The IGL sequence in the ARS-UCDv0.1 is more contiguous and appears more complete; each reference assembly is heavily fragmented with sequence gaps obscuring multiple gene segments. Internal resolution of IGL contig arrangement in the ARS-UCD0.1 is not yet possible and more potential *IGLV* may be discovered in the sequence gaps between contigs but we show a greater number of *IGLV* (n=98) characterised than before in the Btau_3.1 annotation (n=63, Ekman et al, 2009). The IGK is also improved in the ARS-UCDv0.1 and confined to a single contig, revealing the locus consists of 20 *IGKV* rather than the 22 gene segments previously reported. The IGK locus in cattle is much smaller and less complex compared to the IGL.

In comparison, African buffalo appear to have less *IGLV* than cattle. Assembly of the complete IGL was impossible over the highly repetitive *IGLV* region and so genome reads were mapped to individual gene segments and *de novo* assembled. A total of 58 *IGLV* were assembled using reads mapped to scaffold 1160, a lesser number than cattle and no African buffalo genome reads mapped to the other scaffolds. The highly repetitive nature may have made mapping to the region difficult but it is possible that either African buffalo have a smaller IGL or the cattle IGL in the ARS-UCDv0.1 has multiple artificial duplications in its sequence. A complete IGL assembly in African buffalo is awaited for the more accurate description of the locus. The IGK locus in African buffalo is better assembled than the IGL locus, containing 14 of the 20 *IGKV* gene segments that were individually assembled. The IGK is, however, still heavily fragmented but the restricted size that we see in cattle can be confirmed.

The additional complexity of the lambda locus predicted its preferential expression which has been shown in previous studies (Arun et al., 1996; 196, Murphy, 2008; 199). Quantitative PCR confirmed the predominant usage of IGL in cattle and shows that African buffalo have similar dominance of IGL. We show that the predominant usage of IGL in 95-98% of cattle and African buffalo light chain transcripts varies slightly over time due to the natural variation in expression and/or disparities in sampling and RNA isolation. These fluctuations

are insignificant and provide a standard for measuring variation of IGL expression between animals and time points in response to FMDV. Inoculation with FMDV in cattle and infection in African buffalo was measured with an identical assay to show that the expression level of IGL does not change significantly upon infection. This dominance of IGL and the lack of change upon infection restricts the antibody repertoire further.

Further restrictions to the IGL repertoire exist with the preferential expression of a limited number of *IGLV* gene segments. RNA-seq data of cattle and African buffalo reveals the transcription of predominantly few gene segments. This restricted light chain repertoire is shown with Illumina sequencing of the cattle light chain repertoire (Grant et al, unpublished) and later in the African buffalo light chain repertoire (Chapter 5, section 5.3.4). In multiple Holstein cattle, the same four *IGLV* made up ~40% of the total transcripts in each animal and the remaining repertoire consisted of 11 other *IGLV*. All the functional gene segments belong to the phylogenetic sub-group 1. Due to the limited variability of both cattle and African buffalo light chain repertoire it is likely the light chain is performing a predominantly structural role in cattle and African buffalo as it is providing limited additional diversity to the primary antibody repertoire.

Chapter 5

Comparison of the cattle and African buffalo (Syncerus caffer) antibody response to FMDV

5. Abstract

The initial humoral antibody response to FMDV coincides with viral clearance from the blood around day six (Pega et al., 2013; 83) and is considered the molecular mediator of the immune response. Infection of cattle is severe but adult animals recover, developing a protective antibody response to infection that provides long-lasting immunity (Cunliffe, 1964; 80). African buffalo are the long-term maintenance hosts of FMDV; infection of naïve animals is sub-clinical and protection against disease corresponds to high levels of circulating antibody (Condy and Hedger, 1974; 46). The antibody response of African buffalo and cattle to FMDV was investigated after infection or inoculation respectively. Naïve African buffalo were infected with SAT1 serotype of FMDV and the IgM and IgG antibody repertoire at day 0, day 8 and day 14 was sequenced with Illumina. The African buffalo light chain repertoire was also sequenced at day 0 and day 14 post-infection to show that variation in the light chain repertoire in response to FMDV is minimal. Cattle were inoculated with highly purified SAT 1 antigen and the IgG repertoire sequenced with Illumina at day 0, day 7 and day 20. Following the trimming of IGH reads to reflect the *IGHV* regions and CDR3, these regions and the IGL transcripts were clustered based on optimal clustering identities and the frequency abundance, amino acid variability and length distribution of the antibody repertoires were interrogated. Limited variation in length, amino acid variability and the frequency abundance of the IGL supports the hypothesis that the IGL is providing a structural role to the unusual long and ultra-long CDR3H. Limited length distribution is observed in the *IGHV* region but the diversity is high; levels of amino acid variability in response to infection change significantly in the CDR1 and CDR2. African buffalo generate the long and ultra-long CDR3H seen previously in cattle with high levels of variability in their length and amino acid distribution. In response to FMDV infection, African buffalo display a significant increase in the frequency abundance of particular CDR3H sequences whilst this change in frequency is not observed in cattle. This differential antibody response between the two species may account for the protective immune response in African buffalo against disease.

5.1 Introduction

5.1.1 Expansion of the primary antibody repertoire through post-recombinatorial mechanisms

A broad array of antibody structures is needed to cope with the vast array of antigens an animal may encounter throughout its life. The somatic recombination of numerous variable (V), joining (J) and on the heavy chain only, diversity (D) gene segments rearrange in a process called V(D)J recombination to generate the heavy and light antibody chains. This recombinatorial potential is vast in species such as man (3×10^5) but as described in Chapter 3 and Chapter 4, appears more restricted in cattle and African buffalo where somatic recombinatorial potential is only 5×10^4 and 3×10^3 respectively. However, it is highly likely that the characterisation of the African buffalo antibody loci is incomplete and that more gene segments will be found. Both species have dominant usage (95-98%) of the lambda light chain (IGL) over the kappa light chain (IGK) and using RNA-seq expression data we show the dominance of few *IGLV* in the cattle and African buffalo (Chapter 4, section 4.3.12). The preferential expression of few *IGLV* gene segments has been shown in cattle previously with Illumina sequencing (Grant et al; unpublished). This limited primary repertoire must be diversified after transcript assembly in order to generate the variation that we see in the circulating antibody repertoire to deal with encountering antigen.

Somatic hyper-mutation (SHM), the random introduction of individual base mutations by activation induced cytosine deaminase (AID), is the predominant post-recombinatorial modification in cattle and is targeted to the CDR regions of the antibody molecule (Liljavirta et al., 2013; 125). AID is also responsible for gene conversion, where an upstream *IGHV* pseudogene exchanges short sequences with the expressed rearranged *IGHV* gene segment. Evidence for this diversification strategy in cattle is limited but has been shown to occur during light chain development (Lucier et al., 1998; 123, Parng et al., 1996; 124). SHM occurs in the germinal centre which forms in response to antigen encounter. It is suspected that cattle compensate their limited germline recombinatorial potential by SHM in the IPP of the GALT tissue prior to antigen exposure (Yasuda et al., 2006; 126). B cells proliferate rapidly in the IPP and the SHM of IGH and IGL cDNA is correlated with high AID enzyme

expression levels (Liljavirta et al., 2013; 125). This suggests, but remains unproven, that cattle have in fact, two stages to their B cell repertoire generation, first V(D)J recombination and antibody formation in the bone marrow as in other species, then the migration of B cells to the IPP in young animals for SHM for repertoire diversification.

Cattle also have a novel diversification method to other species, in the formation of their ultra-long CDR3H. These ultra-long loops protrude from the surface of the antibody molecule and their diversity arises through different disulphide bonds that form within the loop structure (Wang et al., 2013; 92). It is suspected they are formed from the ultra-long *IGHD* gene segment in the cattle IGH which is 147 bp, nearly three times longer than other cattle *IGHD* gene segments. Further diversity to these structures can arise from somatic hyper-mutation along the sequence. An ultra-long *IGHD* gene segment in African buffalo could not be assembled (Chapter 3, section 3.3.14) and so it was previously unknown if they could produce ultra-long CDR3. In the absence of an ultra-long *IGHD*, African buffalo would form the ultra-long CDR3 by a different mechanism to what is proposed in cattle.

Cattle in general form longer CDR3 than other species; human and mouse CDR3 length ranges from 8-16 amino acids (Collis et al., 2003; 93) whilst cattle CDR3 form a bimodal distribution from 20-40 amino acids and then the ultra-long antibodies at 50-61 amino acids (Berens et al., 1997; 119). Terminal deoxynucleotidyl transferase (TdT) adds non-templated nucleotides to single stranded DNA ends during V(D)J recombination (Koti et al., 2010; 171). TdT is capable of catalysing longer than 1 kb nucleotide additions *in vitro* (Lanham et al., 1986; 215) and has been shown in cattle to add zero to 36 nucleotide additions between V, D and J on the heavy chain and 8 between V and J on the light chain (Liljavirta et al., 2014; 101). This novel diversification method can help account for the longer length CDR3 seen in cattle compared to other species.

Overall, cattle compensate for their restricted germline repertoire by both somatic hyper-mutation, junctional diversity and the formation of long and ultra-long CDR3 structures. These novel diversification mechanisms enable the creation of a sufficiently large functional pre-immune repertoire during late foetal life. Despite the apparent absence of an ultra-long *IGHD* in African buffalo, it is suspected they would employ the same diversification strategies as their germline repertoire appears similarly restricted.

5.1.2 Cattle and African buffalo response to FMDV

Cattle and African buffalo are closely related species, having diverged only 5.7-9.3 million years ago (Glanzmann et al., 2016; 1). Despite this recent divergence we observe differences in their germline antibody repertoire (Chapters 3 and 4) which may help explain their differential response to FMDV and the outcome of disease. African buffalo are the long-term maintenance hosts of FMDV and continually produce antigenic variants of the SAT serotypes which cause minor epidemic outbreaks in their young and to wider species such as cattle and impala. Adult buffalo are asymptomatic and the virus persists in individual animals in the osophago-pharyngeal region for up to 5 years (Condy et al., 1985; 44). Cattle however have 100% morbidity; adult animals develop an acute febrile reaction with severe vesicular lesions and long term chronic effects such as heat intolerance syndrome, whilst their young have ~50% morbidity (Leboucq, 2013; 4).

The antibody response to FMDV has been shown to coincide with viral clearance from the blood (Borca et al., 1986; 52) and is considered responsible for host immunity in both cattle and African buffalo species (Doel, 2005; 79, Bronsvoort et al., 2008; 216). Antibody mediates the innate response with opsonised antibody enhancing the activity of phagocytes (McCullough et al., 1988; 59), NK cells (Bradford et al., 2001; 77) and increasing the antigen uptake of dendritic cells (Summerfield et al., 2009; 65). The initial response to the virus is T independent, as shown by CD4 and partial CD8 T cell depletion in cattle (Juleff et al., 2009; 54) and the initial antibody response to FMDV in athymic mice having no effect on antibody kinetics compared to normal mice (Borca et al., 1986; 52).

The hyper-variable regions of the antibody molecules are largely constrained to the CDR regions, CDR1, CDR2 and CDR3. The CDR3 is the main antigen binding site and, on the antibody heavy chain, the CDR3H is considered the key determinant of antigen specificity (Xu and Davis, 2000; 217). IgG transcript abundance in cattle immunised with FMDV has previously shown a broad range of transcripts which are each specific to the immunised FMDV antigen (Reddy et al., 2010; 218). In contrast, the cattle and African buffalo light chains have previously been shown to vary little in response to FMDV, with the dominant usage of few *IGLV* gene segments that does not alter upon FMDV inoculation (Chapter 4, Grant et al; unpublished). It is suspected therefore that the light chains play a predominantly structural role in cattle and African buffalo antibodies.

Despite protection against FMDV being at least partially attributed to the antibody response of African buffalo, there are currently no studies that have assessed their repertoire. By in-depth sequencing with Illumina, sequencing libraries of IgM, and IgG were generated at day 0, prior to infection, and then day 8 and day 14 after SAT1 FMDV challenge to characterise the African buffalo antibody response to FMDV. Cattle, in a comparable FMDV SAT 1 immunisation study, were sequenced with Illumina by Grant et al, of our Immunogenetics group. The cattle day 0, day 7 and day 20 sequencing reads were analysed in the same analysis pipeline as the African buffalo reads, to compare the antibody repertoires between the two species and their changes in response to FMDV infection or immunisation. Using day 0 samples the post-transcriptional modifications to the antibody transcripts in the absence of antigen were estimated and then changes to their repertoire in response to FMDV were calculated. The IGL repertoire in African buffalo was also sequenced from day 0 and day 14 to investigate the role of the lambda light chains in response to FMDV.

5.2 Methods

5.2.1 African buffalo protocol for infection with FMDV

Twelve African buffalo, six male and six female, were used to study how FMDV persists in isolated populations of its reservoir host (Study number: SANParks Reference No. 13-12 – challenge study). Animals were donated from FMDV free herds by Ezemvelo KZN Wildlife (Esterhuysen et al., 1985) and confirmed free of FMDV SAT-antibodies by the OIE Regional Reference Laboratory. The African buffalo were transferred to experimental pens at Skukuza Kruger National Park (KNP) and sedated for challenge and subsequent sample collections. Experiments were approved by the Onderstepoort Veterinary Institute Transboundary Animal Diseases Programme (OVI-TADP) Ethical Review Committee.

Three groups of four African buffalo animals were each infected with FMDV SAT1/KNP/196/91, FMDV SAT2/KNP/19/89 or FMDV SAT3/KNP/1/08/3, isolated from KNP buffalo (Haydon et al., 2001; 219). Dosage administered to each animal was 1.8×10^6 TCID50 of their respective serotype intradermolingually (2 inoculation sites: 100ul/each). All infected animals were viraemic from 2 dpi and shown by liquid phase blocking ELISA to have sero-converted by day 8 (Perez et al 2016; unpublished, Appendix Table 6). Blood was collected from each animal at day 0 prior to infection and then at 2, 4, 6, 8, 11, 14 and 30 days post-infection (dpi) by tail venipuncture into Tempus Blood RNA Tubes (ThermoFisher Scientific). The Tempus tubes contained stabilising reagent and the subsequent 3 ml of added whole blood for each sample was frozen at -80°C for transport to the UK.

5.2.2 Total RNA isolation from whole blood

African buffalo total ribonucleic acid (RNA) was extracted from whole blood contained within Tempus tubes using the Tempus spin isolation kit (ThermoFisher Scientific). Following the Tempus Spin protocol (https://tools.thermofisher.com/content/sfs/manuals/cms_042989.pdf), samples were thawed from the -80 °C storage and mixed thoroughly with 3 ml of phosphate buffered saline in each

tube for 30 s. RNA was pelleted by centrifugation at 4 °C 3000 x g for 30 min and suspended in 400 µl of RNA purification solution (ThermoFisher Scientific). Cellular debris was filtered from the suspension solution by centrifugation at 16,000 x g for 30 s. The filter membrane was washed three times with 500 µl of RNA purification wash solution (ThermoFisher Scientific) before being eluted by incubation at 70°C with 100 µl of nucleic acid purification elution solution (ThermoFisher Scientific) then centrifugation at 16,000 x g for 30 s. The dried RNA pellet was then suspended in RNase free molecular grade water. Total African buffalo RNA was then quantified by UV spectrophotometry with the 260/280 and 260/230 optical density wavelength values using a Nanodrop 2000 spectrophotometer (Thermo Scientific, Wilmington, DE, USA). Total RNA from each sample was then stored at -80°C.

Purified African buffalo total RNA was reverse transcribed using Superscript II. Reactions of 20 µl volumes were as follows: 200 U of Superscript II reverse transcriptase (ThermoFisher Scientific), 1 µl Oligo(dT), 1 µl dNTPs (50mM), 4 µl first strand buffer, 2 µl Dithiothreitol, 5 ng of RNA and the volume made up to 20 µl with water. The mixture was initially heated to 65 °C then incubated at 42 °C for 50 min and heat denatured at 70 °C for 15 min. Transcribed cDNA was quantified on the Qubit 3.0 Flourometer (Thermo Fishcer Scientific) using the Qubit DNA HS Assay kit (Invitrogen) and each sample standardised to 100 ng/µl.

Day	Event	Animal Groups		
0	FMDV inoculation	4 African buffalo (2 male, 2 female) with SAT1	4 African buffalo (2 male, 2 female) with SAT2	4 African buffalo (2 male, 2 female) with SAT3
0-30	Sample animals at Day 0, 4, 8, 11 and 14 by tail venepuncture	• Monitor parameters of infection in donor animals • Identify when animals become viraemic by rt-PCR • Identify when animals produce an antibody response, LPBE		
	Total RNA isolation from Day 0, 4, 8 and 14 samples	• Total RNA isolated and purified from whole blood using the Tempus spin isolation kit • Purified RNA was quantified by UV spectrophotometry		
	PCR amplification of IgM, IgG and IgL transcripts from Day 0, 4, 8 and 14	• Total RNA was reverse transcribed using Superscript II • IgM, IgG and IgL specific transcripts were then PCR amplified from the cDNA		

Table 5.1: The experimental protocol used for infection of twelve African buffalo animals with SAT1, SAT2 or SAT3 FMDV and the subsequent procedure for antibody transcript isolation.

5.2.3 African buffalo IgM, IgG and IgL primer validation

Primers were considered for the PCR amplification of IgM, IgG and IgL antibody transcripts from total RNA in African buffalo. These primers originated and were validated previously in our lab for the amplification of the respective transcripts in cattle for Illumina sequencing (Grant et al): primers for the 5' *IGHG* and *IGHM* bound at the first 30 nt of constant region 1 (CH1) and the *IGLC* bound within the first 50 nt of CH1 (Appendix Table 1). The 3' primer for *IGHM* and *IGHG* was designed in the *IGHV* leader sequence from cattle *IGHV* cDNA transcripts on the NCBI database. The specificity of these primers to the African buffalo antibody transcripts was hence investigated.

The complete germline sequence of the African buffalo immunoglobulin loci is yet to be completely defined as our assembly attempts are incomplete (Chapter 3 and 4). The total number and sequence of the African buffalo *IGHV* and *IGLV* is unknown. Therefore to amplify the complete *IGHV* regions the specificity of the 3' cattle primers to the African buffalo sequence was assumed as the assembled *IGHV* gene segments in African buffalo, whilst incomplete, show they have highly similar leader sequences to the cattle (Chapter 3).

Primers appeared specific to the *de novo* assembled constant genes but this would not account for polymorphism between animals or if SNPs were introduced in each gene assembly from mapping bias to the cattle ARS-UCDv0.1.

The specificity of the cattle 5' antibody primers were hence confirmed in African buffalo by PCR amplification of the *IGHM*, *IGHG* and *IGLC* genes. Primers for the amplification of the *IGHC* and *IGLC* in African buffalo were designed in the ARS-UCDv0.1 to bind upstream of the CH1 exon in the J-C intron and within the CH3 exon of each constant gene (Appendix Table 1). The *IGHM*, *IGHG* and *IGLC* were subsequently amplified from the genomic DNA of two African buffalo, Sca04 and Sca06, in 25 µl PCR reactions with GoTaq Green master mix x2 (Promega) and 300 nM of each primer. The PCR cycling conditions were 95 °C denature for 2 min and then 35 cycles of 95 °C for 15 s, 58 °C for 1 min and 72 °C for 45 s. PCR products were purified using 4 µl ExoSAP-IT enzymatic clean up (Affymetrix) at 42 °C for 30 min and 5 µl aliquots were ran on a gel with 1 % agarose at 90 V for 90 min. Bands were visualised under UV light for size confirmation and the remaining PCR product was sent for Sanger sequencing for confirmation of the correct amplified product. The African buffalo constant gene amplicons were subsequently aligned to the cattle ARS-UCDv0.1 *IGHM*, *IGHG* and *IGLL* in Bioedit (Hall, 1999; 187) to confirm the specificity of the cattle Illumina primers within the amplicons.

A 5' rapid amplification of cDNA ends (5'RACE) strategy was also implemented; using the confirmed 5' *IGHM*, *IGHG* and *IGLC* primers in African buffalo, the IgG, IgM and IgL transcripts were amplified without the need for the 3' specific primer in the *IGHV* and *IGLV* leader. This allowed the IGL repertoire in African buffalo to be investigated and the specificity of the 5' *IGHV* primer to be determined.

5.2.4 Illumina sequencing strategy

The African buffalo experiment involved infection of three groups with either SAT1, SAT 2 or SAT3 FMDV. However, the Immunogenetics group has comparable data in cattle for inoculation with SAT1 FMDV only. The African buffalo cDNA from animals 7, 11, 8 and 28, infected with SAT1 FMDV were chosen for subsequent PCR amplification of the antibody transcripts. The day 4 time point was also considered less useful as the infected

animals would not have mounted a robust antibody response. All animals sero-converted by day 8, producing anti-FMDV antibodies at detectable levels, as shown by ELISA (Perez et al; unpublished, Appendix 6). The day 0, day 8 and day 14 time points for the African buffalo animals infected with SAT1 were thus carried forward.

5.2.5 IgM and IgG transcript amplification from cDNA

African buffalo IgM and IgG transcripts were amplified from Day 0, 4, 8 and 14 cDNA samples using the high-fidelity DNA polymerase Advantage 2 (Clontech). The 50 µl reactions contained 5 µl of PCR buffer (10x), 1.5 µl $MgCl_2$, 1 µl dNTPs (10 mM), 1 µl 5' primer (10 µM), 1 µl 3' primer (10 µM), 0.4 µl of high-fidelity DNA polymerase and 2 µl of cDNA (100ng/ µl). PCR cycling conditions involved a primer annealing temperature drop down protocol as follows: initial denaturation was 94°C for 60 s, followed by 5 cycles of 94°C for 30 s and 72°C for 120 s, followed by 5 cycles of 94°C for 30 s, 72°C for 30 s and 72°C for 120 s, followed by 16 cycles of 94°C for 30 s, 68°C for 30 s and 72°C for 120 s (Table 5.2, programme 1). Reactions were purified using the QIAquick PCR Purification Kit (Quiagen) into 30 µl of PCR grade water and quantified using the Qubit 3.0 Flourometer (Thermo Fischer Scientific) using the Qubit DNA HS Assay kit (Invitrogen).

5.2.6 African buffalo IgG, IgM and IgL transcript amplification with 5' RACE

Following RNA extraction and quantification, total African buffalo RNA was used for IgM, IgG and IgL transcript synthesis without a 3' primer. Using the SMARTer RACE 5'/3' Kit (Clontech) RACE-ready cDNA was first prepared from RNA by combining 1 µg RNA, 1 µl of the 5'-CDS Primer A (Clontech) and incubating at 72 °C for 3 min before cooling to 42 °C for 2 min. SMARTer II A Oligonucelotides were added, 1 µl per reaction, with 5.5 µl Buffer Mix, 0.5 µl RNase Inhibitor (40 U/µl) and 2 µl SMARTScribe reverse transcriptase (100 U). Reactions were incubated at 42 °C for 90 min before heat denaturation at 70 °C for 10 min. The residual enzymatic activity was prevented by addition of 10 µl Tricine-EDTA buffer. RACE reactions were then performed in 50 µl reactions with 25 µl SeqAmp Buffer (2x), 2.5 µl RACE-ready cDNA, 5 µl universal primer mix (10x), 1 µl 3' gene specific primer, 15.5 µl

PCR grade water and 1 µl SeqAmp DNA polymerase. Reactions with no gene specific primer or no universal primer mix were also performed to test the residual activity of the polymerase and non-specific amplification of products. Two different PCR thermal cycling programmes were used for PCR amplification (Table 5.2). Programme 1 used a temperature touch down PCR as used in IgM and IgG amplification in section 5.2.4 whilst programme 2 had an initial denaturation of 94°C for 60 s, followed by 25 cycles of 94°C for 30 s, 68°C for 30 s and 72°C for 120 s. 5 µl of the PCR product from each programme was run on a 1% agarose gel with the control reactions at 90 V for 90 min. The gel was visualised under UV light to confirm programme 1 (Table 5.2) produced clean specific bands.

Programme 1			Programme 2		
Temp (°C)	Time (s)	Cycles	Temp (°C)	Time (s)	Cycles
94	30	5	94	30	25
72	120		68	30	
94	30		72	120	
70	30	5			
72	120				
94	30				
68	30	25			
72	120				

Table 5.2: PCR thermal cycling parameters for amplification of IgM, IgG and IgL African buffalo transcripts

5.2.7 Protocol for FMDV immunisation in cattle

Four male Holstein-Friesian calves (*Bos taurus*) were immunised with 10µg of inactivated and highly purified SAT 1 Zim (FMDV SAT1 ZIM 22/89 (S1Z), n = 4) FMDV antigen as part of a larger study not associated with this PhD (Grant et al., 2016; 220). Blood was collected from each animal at day -1 before prime vaccination and then blood was collected at day 7 and day 20. All experiments were approved by the Pirbright Institute's ethical review

process in accordance with Home Office guidelines on animal use. Heparinised blood samples were processed to isolate PBMC using the method described in Chapter 2, section 2.2.4. Total RNA was extracted as described in Chapter 4, section 4.2.8 using Trizol LS (Invitrogen, Carlsbad, CA) and quantified by UV spectrophotometry. The quantified total RNA was then used to synthesise full length cDNA using SMARTer® PCR cDNA Synthesis Kit, as described in section 5.2.2, (Clontech Laboratories Inc., Mountain View, CA). Using the IgG Illumina primers described in section 5.2.3, the IgG amplicon transcripts were amplified using high-fidelity Taq DNA polymerase, Q5® High-Fidelity DNA Polymerase (New England Biolabs, Ipswich, MA) as described in section 5.2.4.

The IgG transcript libraries of the four animals were sequenced at the Oxford Genomics Centre, University of Oxford, on the Miseq with 2 x 300 bp paired end reads. The sequencing adaptor primers were blunt end ligated onto the amplicons in a modified TruSeq library preparation. The sequencing reads of the SAT1 immunised cattle were later acquired for our comparison to the African buffalo SAT1 FMDV infection.

5.2.8 PCR library visualisation and quantification

The purified IgM, IgG and IgL African buffalo transcripts were visualised with the Agilent 2100 Bioanalyzer to determine accurate DNA concentration and band sizing (Appendix table 3). Using the Agilent DNA 12000 reagents and protocol, 1 µl of each sample was loaded into individual wells of the DNA chip with 5 µl of marker. Loaded chips were vortexed for 1 min at 2400 rpm and run on the Bioanalyzer (Agilent) within 5 min. Bioanalyser results were exported for analysis of sample concentration and purity.

5.2.9 Illumina sequencing of African buffalo IgM, IgG and IgL

Illumina sequencing was performed at the DNA Sequencing Facility in the Department of Biochemistry, University of Cambridge. African buffalo day 0, 8 and 14 samples for animals 7 and 11, infected with SAT1 were sequenced on the Miseq 2 x 300 bp. The 5'RACE IgM and IgG transcripts from day 0 of animal 11 and the 5'RACE IgL transcripts from animal 11

and animal 28 at day 0 and day 14 were also sequenced. The sequencing centre blunt end ligated the Illumina adaptors onto the cDNA in a modified TruSeq library preparation, as had been done previously with cattle antibody transcript sequencing. An average ~1,400,000 reads were sequenced per sample for the twelve PCR amplified samples whilst an average 1,320,000 reads were sequenced from the 5'RACE libraries. None of the sequences were flagged as poor quality, therefore no quality filtering was done at this stage.

5.2.10 *IGHV* region and CDR3 sequence isolation

The IgG, IgM and 5'RACE African buffalo Illumina sequencing libraries were provided as compressed FASTQ files. FASTQC was used to ascertain the average quality of the reads based on the Phred scores generated by the Illumina base calling software. Reads less than 300 bp in length were filtered out to remove non-specific and contaminating sequence.

Within each sequence library, paired reads were merged with FLASH (Magoc and Salzberg, 2011; 221); any unpaired read denotes one read pair survived trimming while other did not. Both the merged reads and the unpaired reads were then taken forward for analysis at this stage. Merged and unmerged reads were translated into all six possible reading frames; the incorrect frames were removed by filtering out reads containing either stop codons or missing the framework 3 motif "YYC" and any similar derivative sequences identified in the majority of the reads (YYC, YYG, YCC, YSC, FYC, HYC, YHC, YFC and YWC). Sequences with two "YYC" motifs, suggestive of incorrectly merged reads, were removed from the file. The reads were then separated to isolate both the CDR3 and *IGHV* region (which includes FR1, FR2, FR3, CDR1 and CDR2) for analysis.

The CDR3H and *IGHV* regions were trimmed and analysed separately in order to achieve optimum clustering of the two IGH transcript regions. The CDR3H is known to be highly diverse whilst the FR region of the *IGHV* provide structural similarity between sequences which would alter their optimum clustering values. By separately analysing each region, the frequency abundance of each cluster was calculated to estimate variability within the regions.

The *IGHV* region was therefore extracted from each read by removing the sequence after the Cys104 in the 'YYC' motif of FR3. Reads containing the 3' primer sequence 'LLFV' were then extracted to ensure only full-length *IGHV* region sequences were taken forward. The

CDR3 were isolated by removing any sequence before the Cys104 in the YYC then isolating sequences containing the IgM or IgG 5' primer sequence, PKVY and VFPL respectively. The CDR3 were trimmed to remove the sequence after the W in the W-F-G-X motif in the *IGHJ* gene so that only complete CDR3 were isolated.

Script for *IGHV* region and CDR3 extraction:

```
for file in \
Day0-IgG-A11-10_S10_L001 \
Day0-IgG-A7-7_S7_L001 \
Day0-lgM-A11-4_S4_L001 \
Day0-lgM-A7-1_S1_L001 \
Day14-IgG-A11-12_S12_L001 \
Day14-IgG-A7-9_S9_L001 \
Day14-lgM-A11-6_S6_L001 \
Day14-lgM-A7-3_S3_L001 \
Day8-IgG-A11-11_S11_L001 \
Day8-IgG-A7-8_S8_L001 \
Day8-lgM-A11-5_S5_L001 \
Day8-lgM-A7-2_S2_L001 \
RACE-IgG-A11-Day0-14_S14_L001 \
RACE-IgM-A11-Day0-13_S13_L001; do

#Make working directory
mkdir ${file}
cd ${file}

#Filter out reads not of length 301bp
java -jar ~/programs/Trimmomatic-0.36/trimmomatic-0.36.jar PE
     -phred33 \
../reads/${file}_R1_001.fastq ../reads/${file}_R2_001.fastq \
${file}_R1_001.filtered.fastq \
${file}_R1_001.unpaired.fastq \
${file}_R2_001.filtered.fastq \
${file}_R2_001.unpaired.fastq \
MINLEN:200

#Merge reads with FLASH
FLASH-1.2.11/flash -M 301 -t 1 -o ${file}
     ${file}_R1_001.filtered.fastq
     ${file}_R2_001.filtered.fastq

#Convert merged and non-overlapping reads to fasta
cat ${file}.extendedFrags.fastq | awk '{if(NR%4==1)
     {printf(">%s\n",substr($0,2));} else if(NR%4==2) print;}'
     > ${file}.fasta
```

```
cat ${file}.notCombined_2.fastq | awk '{if(NR%4==1)
     {printf(">%s\n",substr($0,2));} else if(NR%4==2) print;}'
     > ${file}_R2_NO.fasta

#Translate into all 6 ORF
bbmap/translate6frames.sh in=${file}.fasta
     out=${file}.aa.fasta

#Merge reads onto one line
awk '/^>/ {printf("\n%s\n",$0);next; } { printf("%s",$0);}
     END {printf("\n");}' < ${file}.aa.fasta >
     ${file}_oneline.aa.fasta

        for motif in \
        YYC YYG YCC YSC FYC HYC YHC YFC YWC; do

#Print only CDR3, sort and print length
fgrep "$motif" ${file}_oneline.aa.fasta | grep -v '\*' >>
     ${file}_oneline

#CDR3 isolation only
#Remove everything before motif
sed -i "s/.*$motif/$motif/g" ${file}_oneline.CDR3

#V-region isolation only
#Remove everything after motif
sed -i "s/$motif.*/$motif/g" ${file}_oneline.V

        done

#Filter out first pair of motifs and create new file for
filtering
sed '/YYC.*YYC/d' ${file}_oneline.CDR3/V >
     ${file}_oneline.single

        for motif1 in \
        YYC YYG YCC YSC FYC HYC YHC YFC YWC; do

                for motif2 in \
                YYC YYG YCC YSC FYC HYC YHC YFC YWC; do

#Filter out all sequences with two motifs
sed -i "/$motif1.*$motif2/d" ${file}_oneline.single
                done

        done

#CDR3 isolation only
#Remove reads shorter than 28aa, 28aa minimum functional ab
cat ${file}_oneline.single | awk '{ print length(), $0 | "sort
     -n -s" }' > ${file}_oneline.single.sorted
```

```
awk '($1 > 27)' ${file}_oneline.single.sorted >
     ${file}_oneline.single.sizefiltered.sorted

#CDR3 isolation only
#Extract reads containing the IGHC primer sequence
grep "PKVY" ${file}_oneline.singleCDR3.sizefiltered.sorted >
     ${file}_oneline.singleCDR3.sizefiltered.sorted.PKVY
grep "VFPL" ${file}_oneline.singleCDR3.sizefiltered.sorted >
     ${file}_oneline.singleCDR3.sizefiltered.sorted.VFPL


#V-region isolation only
#Sort V onto one line
cat ${file}_oneline | awk '{ print length(), $0 | "sort -n -s"
     }' > ${file}_oneline.V.sorted

#V-region isolation only
#Remove reads shorter than 100aa, 110aa minimum functional ab
awk '($1 > 100)' ${file}_oneline.V.sorted >
     ${file}_oneline.V.sizefiltered.sorted
#V-region isolation only
#Extract reads containing the V-leader primer sequence
grep "LLFV" ${file}_oneline.V.sizefiltered.sorted >
     ${file}_oneline.V.sizefiltered.sorted.LLFV
```

5.2.11 IGL sequence isolation

Using an analysis pipeline similar to the IgG and IgM transcripts in section 5.2.10, the IgL FASTQ sequences were filtered for sequences less than 300 bp and merged with FLASH. Merged reads were translated into all six open reading frames and the correct reading frame selected by eliminating those with stop codons and selecting for the FR3 motif "YYC" and the derivatives of the sequence identified in the majority of the reads (YYC, YYG, YCC, YSC, FYC, HYC, YHC, YFC and YWC). Previous studies in cattle of the IgL suggests limited variability in their length and diversity so downstream clustering was unlikely to be affected. Unlike the IgM and IgG analysis then, the IgL were maintained as full length transcripts by selecting reads containing the 3' primer sequence and the conserved "MAW" amino acid sequence at the start of the *IGLV* leader sequence.

5.2.12 *IGHV* region, CDR3 and IGL clustering

Following the filtering of raw reads and the isolation of the *IGHV* and CDR3 in each sample, the sequences were clustered by length and sequence identity. The FASTQ files were converted to FASTA for clustering. Sequences within each sample were clustered using UCLUST (Edgar, 2010; 222) by first sorting on length then clustering based on identity scores. The clustering method uses the USEARCH algorithm to sensitively perform a local and global search of the sequences at high speed; the UCLUST then clusters aligned sequences. Mismatch identity scores of 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98% and 99% were used to identify the optimum clustering identity score in later analysis using the PhiX Illumina quality control, outlined in section 5.2.13.1 and Elbow variance statistics, outlined in section 5.2.13.2. Single reads which did not cluster were moved to a separate directory. The largest 200 clusters from the optimum clustering identity score determined in the analysis, were moved to a separate directory and the clusters were organised by the size (number of sequences they contained).

Script for *IGHV* region and CDR3 extraction:

```
for file in \
Day0-IgG-A11-10_S10_L001 \
Day0-IgG-A7-7_S7_L001 \
Day0-lgM-A11-4_S4_L001 \
Day0-lgM-A7-1_S1_L001 \
Day14-IgG-A11-12_S12_L001 \
Day14-IgG-A7-9_S9_L001 \
Day14-lgM-A11-6_S6_L001 \
Day14-lgM-A7-3_S3_L001 \
Day8-IgG-A11-11_S11_L001 \
Day8-IgG-A7-8_S8_L001 \
Day8-lgM-A11-5_S5_L001 \
Day8-lgM-A7-2_S2_L001 \
RACE-IgG-A11-Day0-14_S14_L001 \
RACE-IgM-A11-Day0-13_S13_L001; do

#Make working directory
mkdir ${file}
cd ${file}

#Filter out reads not of length 301bp
java -jar ~/programs/Trimmomatic-0.36/trimmomatic-0.36.jar PE
     -phred33 \
```

```
../reads/${file}_R1_001.fastq ../reads/${file}_R2_001.fastq \
${file}_R1_001.filtered.fastq \
${file}_R1_001.unpaired.fastq \
${file}_R2_001.filtered.fastq \
${file}_R2_001.unpaired.fastq \
MINLEN:200

#Merge reads with FLASH
FLASH-1.2.11/flash -M 301 -t 1 -o ${file}
     ${file}_R1_001.filtered.fastq
     ${file}_R2_001.filtered.fastq

#Convert merged and non-overlapping reads to fasta
cat ${file}.extendedFrags.fastq | awk '{if(NR%4==1)
     {printf(">%s\n",substr($0,2));} else if(NR%4==2) print;}'
     > ${file}.fasta
cat ${file}.notCombined_2.fastq | awk '{if(NR%4==1)
     {printf(">%s\n",substr($0,2));} else if(NR%4==2) print;}'
     > ${file}_R2_NO.fasta

#Translate into all 6 ORF
bbmap/translate6frames.sh in=${file}.fasta
     out=${file}.aa.fasta

#Merge reads onto one line
awk '/^>/ {printf("\n%s\n",$0);next; } { printf("%s",$0);}
     END {printf("\n");}' < ${file}.aa.fasta >
     ${file}_oneline.aa.fasta

        for motif in \
        YYC YYG YCC YSC FYC HYC YHC YFC YWC; do

#Print only CDR3, sort and print length
fgrep "$motif" ${file}_oneline.aa.fasta | grep -v '\*' >>
     ${file}_oneline

#CDR3 isolation only
#Remove everything before motif
sed -i "s/.*$motif/$motif/g" ${file}_oneline.CDR3

#V-region isolation only
#Remove everything after motif
sed -i "s/$motif.*/$motif/g" ${file}_oneline.V

        done

#Filter out first pair of motifs and create new file for
filtering
sed '/YYC.*YYC/d' ${file}_oneline.CDR3/V >
     ${file}_oneline.single
```

```
        for motif1 in \
        YYC YYG YCC YSC FYC HYC YHC YFC YWC; do

                for motif2 in \
                YYC YYG YCC YSC FYC HYC YHC YFC YWC; do

#Filter out all sequences with two motifs
sed -i "/$motif1.*$motif2/d" ${file}_oneline.single
                done

        done

#CDR3 isolation only
#Remove reads shorter than 28aa, 28aa minimum functional ab
cat ${file}_oneline.single | awk '{ print length(), $0 | "sort
     -n -s" }' > ${file}_oneline.single.sorted
awk '($1 > 27)' ${file}_oneline.single.sorted >
     ${file}_oneline.single.sizefiltered.sorted

#CDR3 isolation only
#Extract reads containing the IGHC primer sequence
grep "PKVY" ${file}_oneline.singleCDR3.sizefiltered.sorted >
     ${file}_oneline.singleCDR3.sizefiltered.sorted.PKVY
grep "VFPL" ${file}_oneline.singleCDR3.sizefiltered.sorted >
     ${file}_oneline.singleCDR3.sizefiltered.sorted.VFPL


#V-region isolation only
#Sort V onto one line
cat ${file}_oneline | awk '{ print length(), $0 | "sort -n -s"
     }' > ${file}_oneline.V.sorted

#V-region isolation only
#Remove reads shorter than 100aa, 110aa minimum functional ab
awk '($1 > 100)' ${file}_oneline.V.sorted >
     ${file}_oneline.V.sizefiltered.sorted
#V-region isolation only
#Extract reads containing the V-leader primer sequence
grep "LLFV" ${file}_oneline.V.sizefiltered.sorted >
     ${file}_oneline.V.sizefiltered.sorted.LLFV
```

5.2.13 *IGHV* region and CDR3 clustering analysis

5.2.13.1 PhiX Illumina quality control

The PhiX library, derived from the well characterised PhiX genome (Sanger et al., 1977; 223), was added to each sample library at an average concentration of 0.5% to provide a quality control for Illumina sequencing. BBMap (Bushnell, 2014; 224) was used to align PhiX reads in the sample to the PhiX genome to estimate the error rate in the sequencing. The PhiX reads were shown to contain, on average between samples, 2.04% bases with errors.

```
bbmap.sh in=reads.fq ref=phix.fa mhist=mhist.txt
     qhist=qhist.txt qahist=qahist.txt
```

5.2.13.2 Elbow variance statistics for the optimum clustering score

The Elbow method (Krolak-Schwedt and Eckes, 1992; 225), a direct method for determining the optimal clustering, was used to determine the optimum mismatch identity score. The variance between samples of the number of clusters formed at each mismatch identity and their average size was calculated using the sum of the squared distances from the mean, divided by the total number of samples. This variance is plotted as a function against mismatch identity in Microsoft Excel to observe where the marginal change in variance between samples declines and creates an “elbow” in the graph.

5.2.13.3 The Exponential Shannon index to measure cluster diversity

The Shannon diversity index was calculated to estimate the repertoire diversity of the largest clusters. The Shannon index is equal to $-\Sigma pi \times \ln pi$; $pi = ni/N$, where N represents the

number of deduplicated sequences and ni equals the number of deduplicated sequences in the selection of largest clusters of each sample i. This makes the pi the ratio of deduplicated sequences within the overall sample. The largest 10, 50, 100, 200, 300 and 500 clusters from each sample were isolated and the repertoire diversity within each cluster sub-sample was expressed as the exponential Shannon index.

Script for calculating deduplicated sequence counts in the largest clusters:

```
#!/bin/bash

for file in \
Day0-IgG-A11-10_S10_L001 \
Day0-IgG-A7-7_S7_L001 \
Day0-lgM-A11-4_S4_L001 \
Day0-lgM-A7-1_S1_L001 \
Day14-IgG-A11-12_S12_L001 \
Day14-IgG-A7-9_S9_L001 \
Day14-lgM-A11-6_S6_L001 \
Day14-lgM-A7-3_S3_L001 \
Day8-IgG-A11-11_S11_L001 \
Day8-IgG-A7-8_S8_L001 \
Day8-lgM-A11-5_S5_L001 \
Day8-lgM-A7-2_S2_L001 \
RACE-IgG-A11-Day0-14_S14_L001 \
RACE-IgM-A11-Day0-13_S13_L001; do

cd ${file}

        for i in 10 50 100 200 300 500; do
        mkdir -p ${i}_clusters_90

                for f in ${file}_clusters_90/*; do wc -l $f;
done | sort -n -k 1 | tail -n ${i} | awk '{print $2}' | while
read -r line ; do
                ln -n ${line} ${i}_clusters_90/
                done

        echo "$file" "$i" >> shannon_counts.txt
        cat ${i}_clusters_90/c_* | awk '/^>/
{printf("\n%s\n",$0);next; } { printf("%s",$0);}  END
{printf("\n");}' | grep -v '>' | sort -k1,1 | uniq | wc -l |
tail -1 >> shannon_counts_90.txt
        done

cd ..
```

```
done
```

5.2.13.4 *IGHV* region, CDR3 and IGL frequency abundance estimates

The frequency abundance of the largest 200 clusters within each sample and within each optimum cluster identity score was calculated. The clusters were ranked in descending order and the number of sequences within each cluster was used to determine its relative abundance compared to the total number of sequences within the sample. The percentage of sequences within each cluster compared to the total reads was displayed graphically.

5.2.14 5'RACE library coverage comparison to the PCR library

A 5'RACE library and a PCR library were both made from the same African buffalo A11 day 0 samples for both IgM and IgG in order to estimate the specificity of the 3' primer in the *IGHV* leader sequence. Following the isolation and trimming of CDR3 sequences in section 5.2.9, the CDR3 sequences that occurred once or more in each library were sorted to a new file. These deduplicated sequence files were then compared between the 5'RACE and PCR libraries to estimate the 3'primer coverage. Following this preliminary analysis, CDR3 that were clustered based on a calculated optimum cluster identity score, in section 5.2.10, the clusters in each library sample greater than 0.1% abundance were extracted and the deduplicated sequences within these clusters compared between the 5'RACE and PCR libraries. The 3' primer efficiency was therefore resolved based on cluster abundance.

```
#count deduplicated sequences and write them to a file
cat ${file}_nolengths | sort -k1,1 | uniq | wc -l
cat ${file}_nolengths | sort -k1,1 | uniq >
     ${file}_deduplicated

#count how many deduplicated sequences shared between files
LC_ALL=C grep -w -F -f ${file}_RACE_deduplicated
     ${file}_PCR_deduplicated | wc -l

#proportion of deduplicated sequences shared
```

```
LC_ALL=C grep -w -F -f ${file}_PCR ${file}_RACE | wc -l
```

5.2.15 Frequency array of *IGHV* region, CDR3 and IGL lengths

The lengths of the *IGHV* regions, CDR3 and entire IGL sequences were calculated following the isolation and trimming of the sequences, described in section 5.2.9. A frequency data array was generated in Microsoft Excel on the lengths of all the complete *IGHV* regions, CDR3 and IgL sequences within each sample. The frequency data array of sequence lengths was then converted to a percentage of the total reads within each sample and the percentage lengths plotted on bar graphs.

5.2.16 CDR3 stream graphs

The change in frequency of the most dominant sequences over time was measured using a representative sequence from the largest 50 clusters in each African buffalo and cattle sample. The representative sequence of each of the largest 50 clusters at each time point within an animal were merged into a single file. Using agrep, the consensus sequences at each time point were searched for in the entire CDR3 library of the other two time points. The 93% mismatch identity score used for optimum clustering at each time point was selected for the search, which allowed discrepancies between CDR3 lengths within the identity score. The counts of similar sequences to the representative of the largest 50 clusters in each time point were then recorded at the other two time points. If the sequence did not exist in the cluster directory it was given a count of one for the purpose of downstream analysis. Searching for the cluster consensus sequence in the single cluster file exceeded time constraints of the analysis. The programme streamgraph v0.8.1 (Rudis, 2013; 226), a html widget, was used to produce the stream graphs of the largest 50 cluster sequences over time within each repertoire.

The ultralong CDR3 sequences, of length 50-70 amino acids, were also isolated and the change in frequency of these sequences was measured over time. Similar to above, ultralong sequences were clustered and a representative sequence of the largest 50 clusters at each time point was searched for in the other two time points within each animal using a 93% mismatch

identity. The counts of each sequence at each time point were recorded and visualised in streamgraph v0.8.1.

Script for counting the consensus sequences in the largest 50 clusters at each time point:

```
for file in \
Day0-IgG-A11-10_S10_L001 \
Day0-IgG-A7-7_S7_L001 \
Day0-lgM-A11-4_S4_L001 \
Day0-lgM-A7-1_S1_L001 \
Day14-IgG-A11-12_S12_L001 \
Day14-IgG-A7-9_S9_L001 \
Day14-lgM-A11-6_S6_L001 \
Day14-lgM-A7-3_S3_L001 \
Day8-IgG-A11-11_S11_L001 \
Day8-IgG-A7-8_S8_L001 \
Day8-lgM-A11-5_S5_L001 \
Day8-lgM-A7-2_S2_L001 \
RACE-IgG-A11-Day0-14_S14_L001 \
RACE-IgM-A11-Day0-13_S13_L001; do

#printing a representative sequence in each cluster to the
file name
        for cluster in ${file}/50clusters/*
        do
        echo $cluster >>
clusterconsensus/${file}_clusterconsensus
        sed -n '200p' "$cluster" >>
clusterconsensus/${file}_clusterconsensus

        done

#merging the consensus sequences at each time point
for consensus in ${file}_clusterconsensus/*
do
cat $consensus | grep -v 'RACE' | grep -A 1 'Animal ID' | grep
     -v 'Day' >> ${file}_clusterconsensus/${file}_consensus

done

#remove any duplicated consensus sequences between timepoints
cat ${file}_clusterconsensus/'Animal ID' _consensus | uniq -u
     > ${file}_clusterconsensus/'Animal ID' _consensus_unique

#count the number of sequences in each consensus at all three-
time points
cd ${file}
```

```
while read p; do
echo ${aa_seq}  >> 'Animal ID'_cluster_count

done <'Animal ID'_missing_clusters

     for cluster in ${file}_clusters/*
        do agrep -${subs} -c ${aa_seq} $cluster >> 'Animal
ID'_cluster_count
        done
done

#work in R
R

#set libraries
library(dplyr)
library(streamgraph)

#make streamgraph

dat <- read.csv ("IgG_A7.csv")
dat %>% streamgraph("asset_class", "volume_billions", "year",
     offset="expand", interpolate="cardinal") %>% sg_axis_x(2,
     "year", "%Y") %>% sg_fill_brewer("Paired")
```

5.2.17 *IGHV* region, CDR3 and IGL sequence abundance calculation

The amino acid variation of the *IGHV* regions and CDR3 in African buffalo and cattle and the IgL in African buffalo was determined using an amino acid abundance calculator developed in our Immunogenetics group (Borne; unpublished). The largest 200 clusters in each sample were aligned and at each position, the abundance of each amino acids of every sequence was counted. The frequency array was output into excel and used to determine the consensus sequence of the largest 200 clusters and the percentage abundance of sequences that contained amino acids that deviated from this sequence.

5.2.18 Phylogenetic analysis of the *IGHV* region and ultralong CDR3H sequences

The consensus sequences of the largest 50 clusters in African buffalo animals 7 and 11 and two cattle animals 255 and 256 were compared between the three time points. The ultralong CDR3 consensus sequences from the African buffalo IgG and consensus sequences from the ultralong cattle CDR3 were also compared. Sequences were aligned using a global alignment strategy in the MAFFT package, version 6.603b (Katoh et al., 2002; 227). Phylogenetic analysis of the amino acid *IGHV* region sequences was calculated in MEGA 7.0 (Kumar et al., 2016; 228) using maximum likelihood based on the Jones-Taylor-Thornton model (Jones et al., 1992; 229) using uniform rates amongst sites and 1000 bootstrap iterations.

5.3 Results

5.3.1 Illumina sequencing of African buffalo antibody transcripts

African buffalo in the KNP were infected with SAT1, SAT2 or SAT3 FMDV serotype in order to interrogate their initial IgM, IgG and IgL antibody response to infection. Blood was collected at day 0 (prior to infection), day 4, day 8 and day 14. The total RNA was isolated at an average concentration of 153 ng/µl, although concentration varied considerably between samples (range 86.38 - 239.96 ng/µl). Total RNA was reverse transcribed for PCR amplification of IgM and IgG specific amplicons whilst 5' RACE PCR library amplification was employed on animal 11 day 0 IgM and IgG samples and IgL samples from day 0 and day 14. Purified PCR amplicons had an average concentration of 88.87 ng/µl (range 91.57 - 239.96 ng/µl) and were measured on the bioanalyser to assess their size distribution, concentration and purity. The majority of antibody transcripts were ~605-620 bp although size distribution of the peak was wide due to the varying lengths of the CDR3 and bimodal as around 10% of African buffalo antibodies are anticipated to contain the ultra-long CDR3H. The mean peak concentration on the Bioanalyser of the IgM transcripts was 5.00 ng /µl and IgG transcripts was 11.43 ng / µl. Small peaks were observed after the first bioanalyser marker, representing fragments less than 100 bp, most likely primer sequence which was not removed in the Quiagen purification step. Purification was not able to be repeated due to the low concentrations of the cDNA.

A sequencing strategy was agreed for animals infected with SAT1 as a comparable data set in cattle inoculated with SAT1 is available (Section 5.3.2). The bioanalyzer results for animals 7 and 11 had an overall higher concentration, with less contamination of non-specific amplicons and so these were chosen for subsequent sequencing. The timepoints of day 0, 8 and 14 were selected as the animals were seroconverted after the day 4 time point, rendering it less meaningful. A 5' RACE library was also sequenced for IgM and IgG in animal 11 at day 0 to determine the specificity of the 5' primer. Changes in the light chain repertoire in response to FMDV were determined by sequencing an IGL 5' RACE library for animal 11 at day 0 and day 14. 5'RACE was not used for the generation of every library because the financial cost is very high and the longer 5'RACE amplicons produce lower quality reads.

Illumina sequencing produced an average of ~1.3 million reads (range of 0.9 - 1.47 million) per IgM library, an average of 1.6 million reads per sample (range of 1.4 - 1.8 million) for IgG and an average of 1.4 million (range 1.1 - 1.9 million) reads for the IGL RACE libraries. Overall, none of the sequences were flagged as poor quality; the Phred quality score of individual reads dropped below 20 after ~250 bp for the IgM and IgG reads. The mean sequence quality of all the IgM and IgG reads was >30 for the majority of sequences, equating to a >99.9% base call accuracy. GC content of reads was 47-67% and 43-67% for IgM and IgG sequence reads respectively. The RACE library amplicons are longer and therefore were of slightly lower quality, the Phred score dropping below 20 after 150 bp and after 200 bp, the base call accuracy was 90%.

5.3.2 IGL read isolation

The African buffalo IGL repertoire was sequenced using 5'RACE libraries as the specificity of the 3' primer could not be confirmed in the absence of a complete genome sequence. The *IGLC* were PCR amplified and sequenced with Sanger to confirm the primer specificity of the 5' primer in the CH1 exon of the *IGLC*. Day 0 and day 14 samples were sequenced to measure changes in the IGL repertoire in response to SAT1 FMDV infection. On average, 1.33 million raw reads were sequenced; these were filtered for contaminating sequences and subsequently merged with FLASH. Both the merged and unmerged reads were taken forwards; all reads were translated into their six open reading frames and 1.24 million IgL reads were isolated by selection of the FR3 motif containing the Cys104, YYC and its variations. The 3' sequence end was trimmed to the *IGLV* leader sequence and reads containing the 5' primer were selected for further analysis using the primer motif. On average, 1 million reads per sample were full-length IgL sequence and were taken forward for subsequent analysis (figure 5.1).

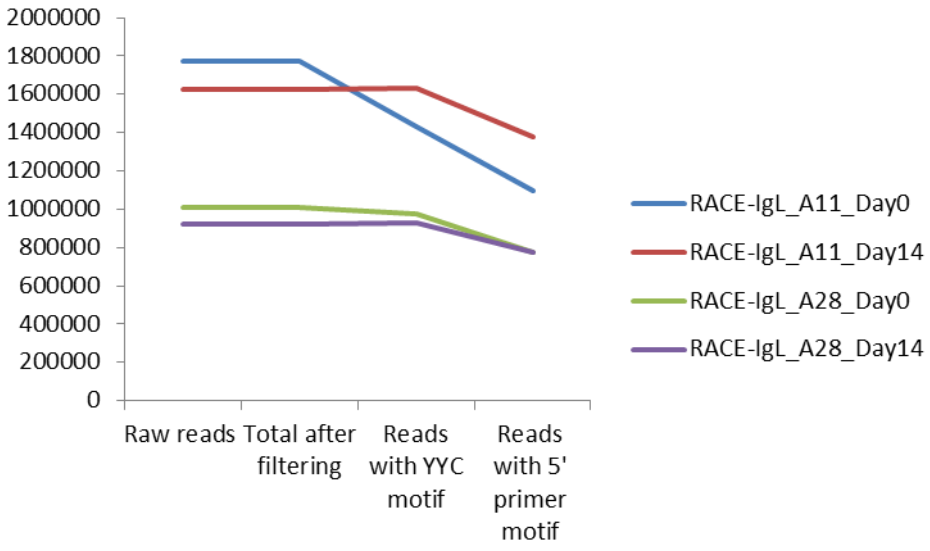


Figure 5.1: Filtering pipeline of the IGL African buffalo sequences. The raw reads, total number of sequences and the number of sequences containing the YYC and *IGLC* primer, indicating full length light chain sequences, are shown.

5.3.3 African buffalo IgL show limited variation in length

The IgL sequences show very limited variability in the length of the transcripts, with a predominant length of 137 amino acids that does not alter upon infection with SAT1 FMDV. The IgL sequences from the African buffalo day 0 and day 14 5'RACE libraries were isolated from the 5' primer sequence in the CH1 exon of the *IGLC* to the start of the leader sequence. The total IgL from each sample were aligned and a frequency array of their lengths generated. This was used to calculate the percentage of sequences at each length from the total IgL sequences in each sample. The Gausian distribution observed shows~76.3% of the reads were 137 amino acids in length whilst IgL lengths that accounted for >2% of transcripts varied from 134-138 amino acids (Figure 5.2).

This is similar to what is observed in cattle in previous studies where the IgL repertoire formed a Gausian distribution of *IGLV* lengths where 61.59% of the transcripts were 109 amino acids in length from the leader to the CDR3 (Grant, 2013; 230).

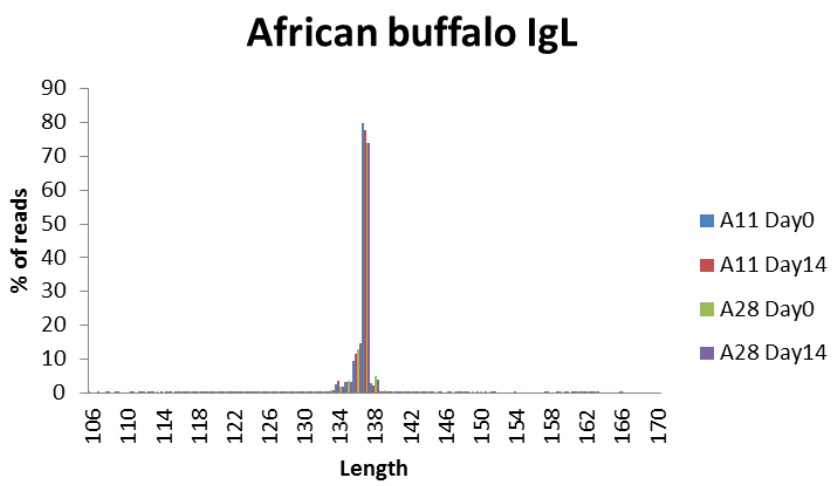


Figure 5.2: Total amino acid length of the IgL transcripts in African buffalo animal 11 and 28. The repertoires were isolated and sequenced at day 0 prior to infection and then at day 14 after challenge with SAT1 FMDV.

5.3.4 The relative abundance of the IgL transcripts

The frequency distribution of the IgL transcripts in the African buffalo A11 5'RACE library at day 0 and day 14 was assessed by calculating the relative % abundance of the largest 200 clusters in each sample (figure 5.3). The cut off of 200 was specified as any FMDV specific transcripts are likely to be within this data group and to maintain consistency with the heavy chain analysis.

The largest cluster accounted for 6.2% and 24.1% of the total IgL reads at day 0 and day 14 respectively. The 200 largest clusters accounted for 76.8% and 81.2% of the entire IgL transcriptome; all of the 200 largest clusters were >0.1% abundance. This is different to the heavy chain where the largest *IGHV* region cluster accounts for 3.9% at day 0 and 3.1% at day 14 and only the largest 34-58 of the clusters are >0.1% abundant. The limited variation within the IgL and the clear dominance of few IgL transcripts suggests they are structurally important for the IGH as no significant change was observed in the light chain between day 0 and day 14. This is consistent with cattle, where only a few IgL sequences dominate in their repertoire (Grant, 2013; 230).

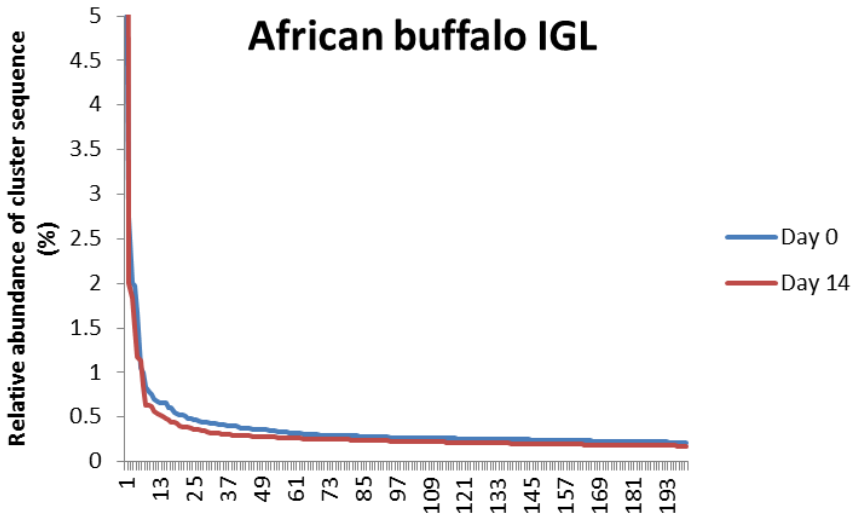


Figure 5.3: Relative abundance of the largest 200 African buffalo IgL transcripts at day 0 and then day 14 after subsequent infection with SAT1 FMDV. The y-axis is set at a maximum of 5 to show that all 200 clusters are >0.1% abundant.

5.3.5 Variation of the amino acid residues in the IgL is limited

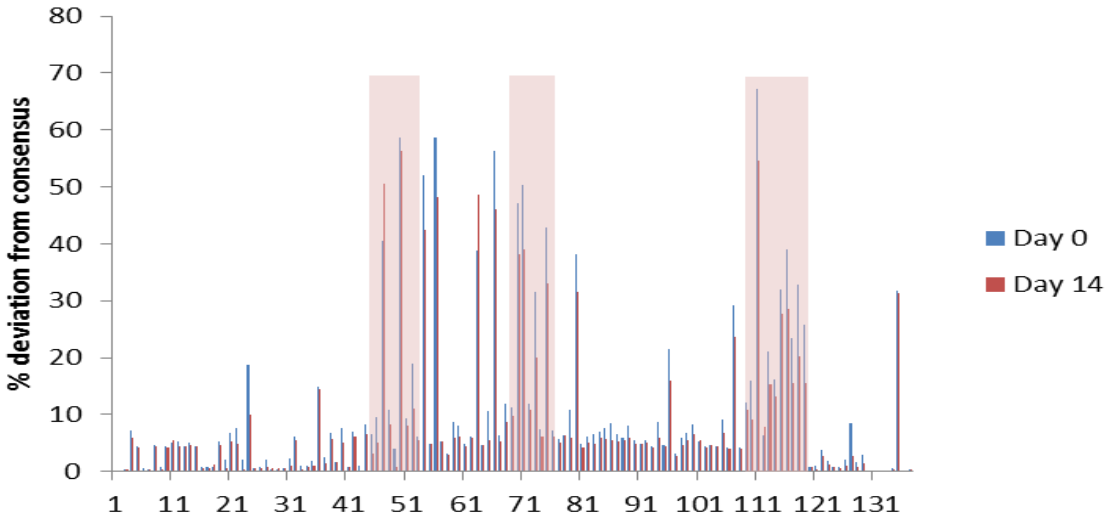


Figure 5.4: Amino acid variation in the African buffalo IgL transcripts at day 0 and upon challenge with SAT 1 FMDV, at day 14. The IgL reads sequenced from a 5'RACE library with Illumina were merged and full-length IgL transcripts isolated. These were clustered and the sequences in the largest 200 clusters were aligned to generate a consensus sequence. The percentage of reads at each amino acid position containing an amino acid other than the consensus was calculated. The CDR1, CDR2 and CDR3 are highlighted in red.

The degree of amino acid variation was calculated across the full-length of the IgL transcripts as their variation in length was severely limited. The extent of amino acid substitutions at each position at both day 0 and following SAT1 FMDV infection, at day 14, was calculated using the full-length IgL transcripts from the largest 200 clusters. The percentage of transcripts containing an amino acid other than the most abundant was then calculated. The regions of high amino acid variability fall within the predicted CDR1, CDR2 and CDR3 regions (Figure 5.4). The % variability within the CDR is high; CDR1 is 18.9%, CDR2 is 25.7% and CDR3 is 23.2%, with no significant difference between these regions. Compared to the heavy chain, the IgL shows less variation within the CDR (the heavy chain % deviation was on average 10.7% for CDR1, 40.4% for CDR2 and 48.9% for CDR3 of mean length 25 amino acids). The IgL shows more variability within the framework regions than the heavy chain; the deviation from consensus in FR1 is 3.3%, in FR2 is 16.8% and in FR3 is 7.6% (compared to the heavy chain FR regions, where % deviation is 2% in FR1, 16.3% in FR2 and 10.5% in FR3), these differences however are not significant. No significant change in amino acid variation is observed upon infection; the average % deviation from consensus is 10.5% at day 0 and 8.5% at day 14.

Overall, the limited variation in IgL sequence length and amino acid composition in response to infection and the high abundance of relatively few sequences suggests they provide a more structural role in antibody formation and are not directly involved in the FMDV response.

5.3.6 Illumina sequencing of cattle antibody transcripts

Four 6-month old Holstein-Friesian male calves (*Bos taurus*, The Pirbright Institute, Woking, UK) were immunised with FMDV SAT 1 Zim serotype as part of a larger study to analyse the *IGHV* and *IGLV* repertoire over the course of a vaccination regime (Grant et al., 2016; 220). The time points selected for comparison to the African buffalo sequencing data was the day -1 before antigen exposure (equivalent to African buffalo day 0), day 6 and day 20 after primer vaccination. For ease of comparison to the African buffalo time points, the cattle time points will be referred to here as day 0, day 6 and day 20. The IgG libraries of the four animals at these selected time ponts were sequenced on the Illumina MiSeq 2 x 300bp. Sequencing generated an average 206,000 reads per sample (range 187,000 - 209,000) with comparable quality scoring to the African buffalo sequencing. The Phred quality scores of

individual reads was over 20 up to ~250 bp and the mean sequence quality of all the reads was >30 for the majority of sequences.

5.3.7 Specificity of the 3' IgM and IgG antibody transcript primer in African buffalo

5' RACE libraries were sequenced for IgM and IgG animal 11 at day 0 in order to compare the library amplified using the *IGHV* leader specific 3' primer from the equivalent sample. The specificity of the 3' primer to the African buffalo *IGHV* could not be determined from the germline beforehand as a complete reference genome is not publicly available and the targeted *de novo* assembly of *IGHV* gene segments in African buffalo (section 3.3.7) is likely incomplete. The 5' RACE libraries can therefore be used to estimate the coverage of the antibody repertoire achieved with the 3' primer.

Sequencing reads from the 5' consensus primer PCR and 5' RACE libraries were trimmed and selected for complete CDR3 sequence from the Cys104 to the Trp in the W/F-G-X-G motif of *IGHJ*. The CDR3 were discretely sorted to contain one copy of every sequence in each library and these consensuses were compared between the 5' consensus primer PCR and 5'RACE library in each sample (figure 5.5A). Lines were matched exactly based on sequence and length.

The 5' consensus primer CDR3 libraries for both IgM and IgG contain ~250,000 deduplicated sequences, that occur one or more times in each library (figure 5.5A). The 5'RACE libraries for IgM and IgG contain 38,000 and 60,000 deduplicated sequences respectively and so appear to contain less diversity than their 5' consensus primer PCR counterparts. Only 2680 of the deduplicated IgM sequences exist in both the 5'RACE and 5' consensus primer PCR libraries and 7503 deduplicated IgG exist in both the 5'RACE and 5' consensus primer PCR IgG libraries. The proportion that the deduplicated sequences accounted for in the respective IgM or IgG 5'RACE or 5' consensus primer PCR library were then calculated. The proportion of sequences in the 5' consensus primer PCR library that were found in the 5'RACE library was ~47% for IgM and ~50% for IgG (figure 5.5B).

The CDR3 sequences in each sample were clustered, discussed in detail in section 5.3.10.1, and the largest 200 clusters extracted for further analysis. The frequency abundance of these

top 200 clusters was calculated (section 5.3.10.2) and here the clusters that appeared with over 0.1% abundance in the 5'RACE and 5' consensus primer PCR libraries were compared. The sequencing depth of the Illumina runs meant that the entire B cell pool in each sample was not sequenced and so any sequence occurring less than 0.1% was considered very low frequency and may not have been reliably amplified. The reads in clusters over 0.1% abundance in each sample were combined and compared between each. Between the IgM libraries, 87.7% of the reads in the 5'RACE library was found in the 5' consensus primer PCR library and 71.4% of the reads in the IgG 5'RACE library was in the equivalent PCR library. The high frequency transcripts were represented in both 5'RACE and 5' consensus primer PCR libraries, with the *IGHV* leader specific 5' primer having a 71.4 - 87.7% specificity to the total IgM and IgG transcripts.

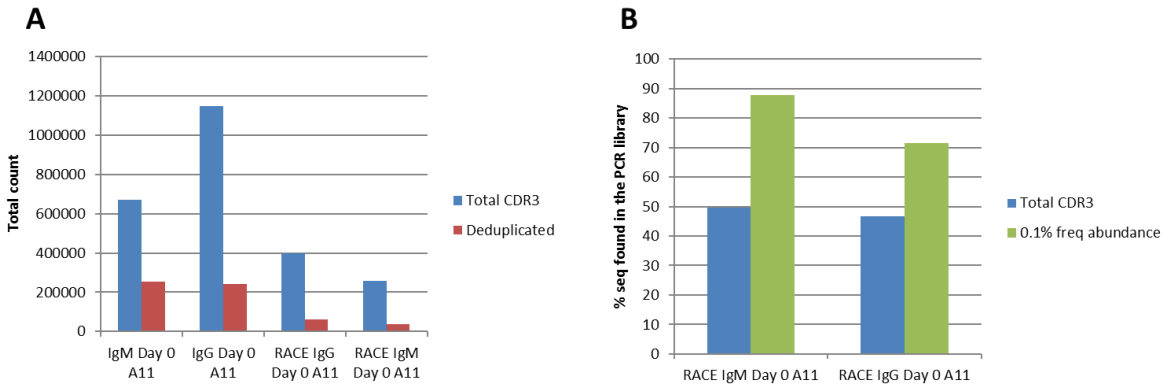


Figure 5.5: The binding specificity of the 3' IgM and IgG primer was estimated by comparing the PCR and 5'RACE libraries sequenced from the A11 day 0 samples. The total CDR3 sequences were filtered from the Illumina reads and every sequence that occurred once or more in each library (the deduplicated sequences) were extracted (A). The PCR and 5'RACE libraries were then compared to determine the number of sequences shared between each library (B). The total CDR3 sequences that occurred in the 5'RACE library was compared to the equivalent PCR library to calculate the percentage of reads found in both. However, a large number of sequences occurred once or more, suggesting the low frequency abundance of the clusters. The clusters with over 0.1% abundance were extracted and the sequences of all these clusters were combined to compare between each 5'RACE and PCR library. The percentage of CDR3 sequences that occurred in clusters of over 0.1% abundance that existed in both the 5'RACE and PCR libraries were then calculated.

5.3.8 *IGHV* clustering pipeline

5.3.8.1 Determining the identity score for *IGHV* clustering

Illumina sequencing reads for both the African buffalo and cattle datasets were filtered for low quality reads of less than 300 bp in length. Reads were subsequently merged with FLASH; an average of 684,000 IgM transcripts and 1.31 million IgG transcripts in the African buffalo merge, accounting for 50.7% and 81.8% of the respective raw reads, whilst 153,000, 74.5% of the raw reads of the cattle transcripts, merged (figure 5.6). Both the merged and unmerged reads were taken forwards and all reads were translated into their six open reading frames. The correct reading frame was selected for by the absence of stop codons and the selection of the FR3 motif containing the Cys104, YYC and its derivatives. An average of 683,000 IgM, 1.34 million IgG and 151,000 cattle reads were thus selected; accounting for an average of 50.6%, 83.7% and 73.6% of the raw African buffalo IgM and IgG reads and the cattle IgG reads respectively. The *IGHV* region of the sequences were then isolated by eliminating the sequence after the Cys104 for analysis of the FR1, FR2, FR3 and CDR1 and CR2 structures. Only full-length *IGHV* regions were analysed and so reads that did not contain the 3' primer sequence in the *IGHV* leader were filtered out. An average total of 782,000 IgM (~55.7% of the raw reads) and 1.2 million IgG (~74.9% of the raw reads) *IGHV* regions were extracted from each African buffalo sample and 129,000 (~62.6% of the raw reads) from each cattle sample (figure 5.6).

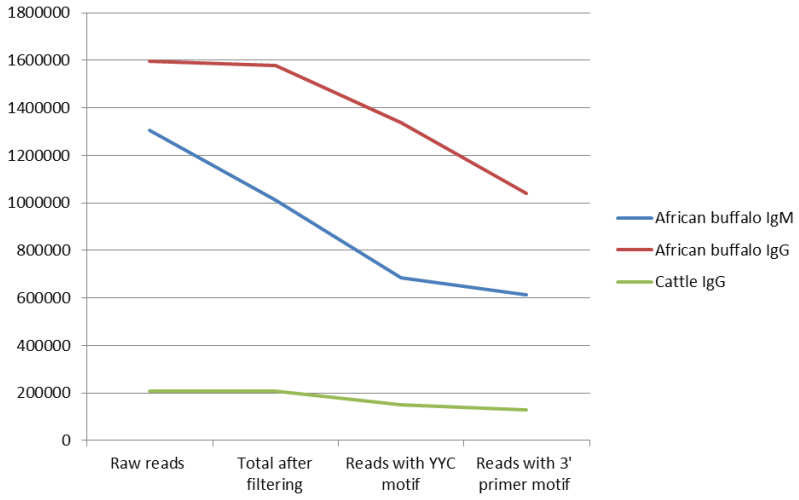


Figure 5.6: Average number of African buffalo IgG and IgM and cattle IgG transcripts generated by Illumina sequencing and the number of *IGHV* region transcripts at each stage of the processing pipeline. Raw sequence reads were filtered for sequences less than 300 bp in length and then the *IGHV* regions were extracted from the Cys104 in the YYC motif in FR3. Full length *IGHV* regions were finally selected by reads containing the 3' primer sequence.

The African buffalo *IGHV* regions were clustered using multiple different cluster identity scores to achieve an optimum clustering value: each 1% interval between 84% and 99% mismatch identity was used. The number of sequences that occurred once or more in each library, the deduplicated sequences, were counted. An average of 534,000 deduplicated sequences occur which equates to the number of clusters that would form when no mismatches are allowed between sequences; the minimum cluster size would therefore be 1.7. As the mismatch identity score is increased, the number of clusters increases but the variance between the numbers of clusters in each sample also increase (Figure 5.7A and 5.7B). On average, 19,000 clusters were generated for each sample with 84% identity and the number of clusters and variance between samples steadily rises until 90% identity where the incremental increase in variance suddenly increases. The average cluster size decreased as mismatch identity was increased (Figure 5.7C); the variance of the cluster size decreases between samples in each mismatch identity (Figure 5.7D). At 90% mismatch identity the decrease in variance forms an "elbow" as the steady decrease contains a larger drop from 89% mismatch identity as variance plateaus.

The percentage of reads that did not cluster and those that did was calculated. As expected, the number of reads that do not cluster increases as mismatch identity is increased (Figure 5.7E) but differences between each sample exist as the percentage of reads that do cluster

decreases more rapidly for certain samples (Figure 5.7F). The sequenced Illumina libraries contained the PhiX control as a ready to use library for quality control of the sequences. The forward and reverse reads of the African buffalo IgM and IgG samples were pooled and PhiX reads were aligned to the genome to estimate sequencing error. A total of 0.5% of reads mapped to PhiX and of these, 2.04% contained bases with errors. Between 84% - 89% the percentage of un-clustered reads is below the inherent error rate in the sequencing of certain samples (as indicated by the dashed red line on Figure 5.7C). The 90% mismatch identity provided the lowest identity score with the percentage of single reads in each sample higher than the 2.04% PhiX read error.

The variance between samples of the cluster number and average cluster size is used as a function against mismatch identity to determine the identity score at which the marginal change in variance drops. This Elbow method is used to determine the optimum cluster number and here is seen at 90% mismatch identity. Additionally, at 90% mismatch identity the number of un-clustered sequences is higher than the sequencing error rate. This identity scoring was therefore chosen for subsequent analysis.

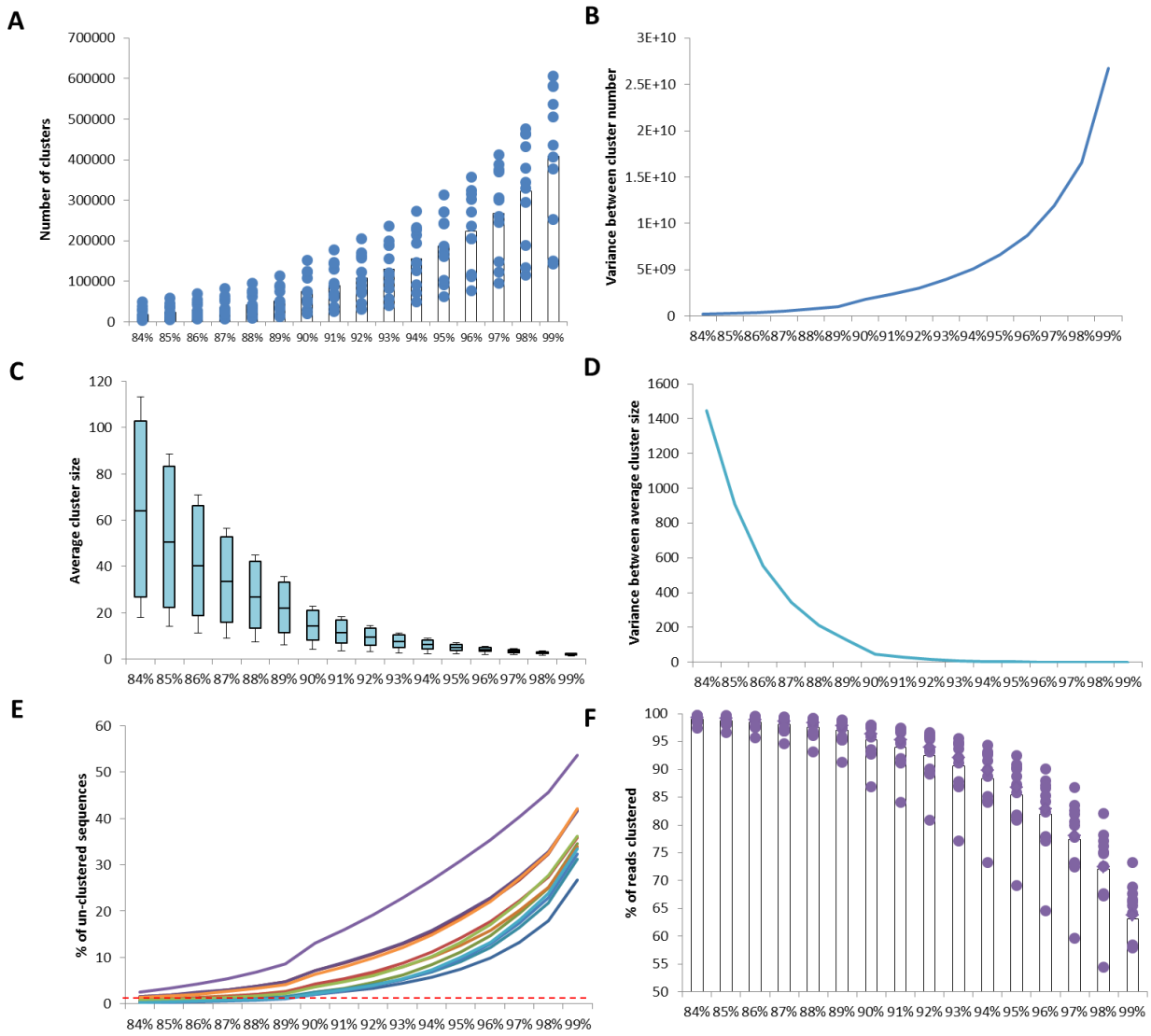


Figure 5.7: The African buffalo IgM and IgG reads were clustered at every identity scores between 84% - 99% (in 1% increments) for determining the optimum clustering parameters. The number of clusters for each mismatch identity (A) is shown with the blue dots indicating each sample and the black bars representing the mean number of clusters. The variance of cluster number between samples is displayed to the right (B). The average cluster size is indicated by the dot and whisker plots, the central bar indicating the mean and the variance between average cluster sizes of samples is shown (D). The percentage of the average un-clustered reads (E) increases as mismatch identity increases. The dashed red line indicates the PhiX calculated error rate in the Illumina sequencing. The percentage of clustered reads in each sample differs between samples and this difference increases as mismatch identity increases (F).

The African buffalo 90% mismatch identity produced an average of 75,000 clusters with an average size of 15 sequences (Figure 5.8A). The largest clusters contained an average 23,400 sequences whilst 38,000 sequences were un-clustered. The cattle *IGHV* regions were subsequently clustered with the same mismatch identity, 90%, as African buffalo. An average of 14,000 clusters formed containing an average of 10 reads per cluster (Figure 5.8B). The largest cluster contained an average of 5500 sequences whilst an average 6700 sequences were un-clustered.

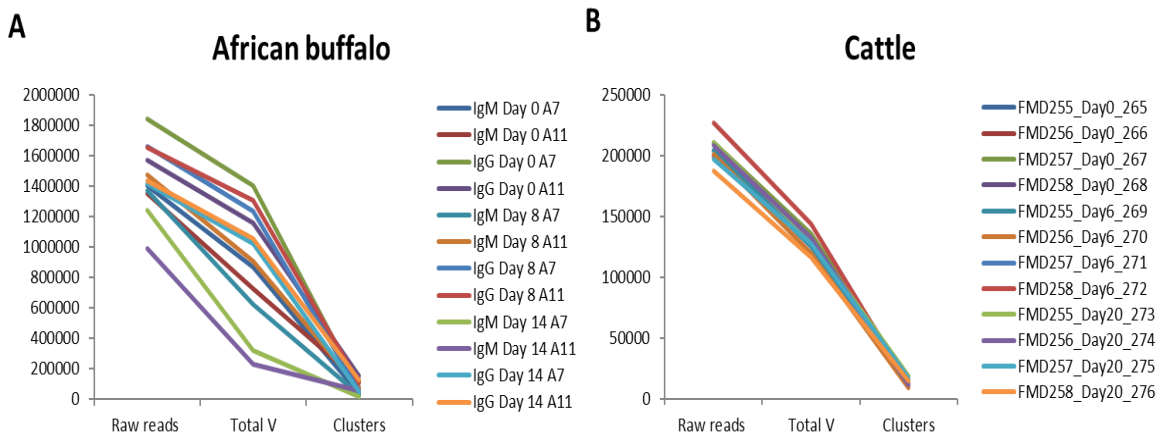


Figure 5.8: *IGHV* region clustering pipeline of the African buffalo IgG and IgM and cattle IgG Illumina sequencing transcripts. The number of raw Illumina sequence reads, the total number of *IGHV* regions isolated from the pipeline and the total number of clusters per sample are indicated. *IGHV* regions were clustered with UCLUST at 90% mismatch identity as determined by optimisation methods.

5.3.8.2 Exponential Shannon index as a measure of diversity in the largest clusters

The African buffalo IgG and IgM transcripts were clustered with a 90% mismatch identity using UCLUST, as outlined in section 5.2.4.1. To quantify the diversity within the largest clusters the exponent of the Shannon index was calculated for the largest 10, 50, 100, 200 and 500 clusters in each sample. The Shannon index considers species richness and evenness of representation; for the analysis here, the number of sequences that occur once or more in each size selection of clusters, the deduplicated sequences, represent richness whilst the number of clusters represented evenness. The maximum exponent Shannon was calculated using the number of deduplicated *IGHV* region transcripts in the entire *IGHV* libraries and was considered as 100% diversity. The exponent Shannon of each subset of the largest clusters was determined (Figure 5.9). The largest 10 clusters contain an

average 19.1% diversity of the entire samples. Increasing the number of clusters included steadily increases the diversity, 200 clusters represent 44.8% of the library diversity whilst 300 clusters represents over 50% of the maximum diversity of the samples.

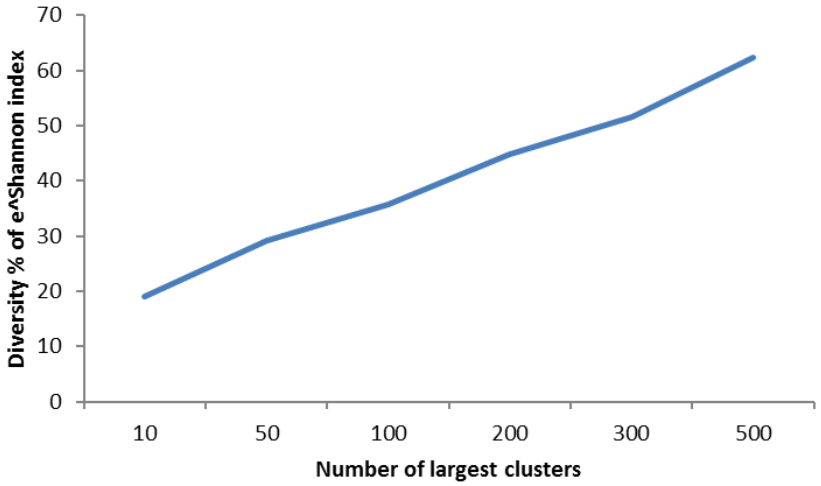


Figure 5.9: The Shannon diversity index of the largest clusters in the African buffalo IgM and IgG *IGHV* region transcripts. The deduplicated sequences in the largest 10, 50, 100, 200 and 500 clusters in each sample were used to calculate the exponent Shannon. The maximum exponent Shannon was considered as 100% diversity in all the deduplicated sequences in each sample. The percentage of diversity in the largest clusters was then calculated as a percentage of the maximum. This allows the diversity of the largest clusters in subsequent analysis to be known.

5.3.8.3 The abundance of *IGHV* region transcripts in African buffalo and cattle

The frequency distribution of the *IGHV* regions in the African buffalo IgM and IgG and the cattle IgG transcripts was assessed by calculating the relative % abundance of the largest 200 clusters in each sample (figure 5.10). The cut off of 200 was specified as any FMDV specific transcripts are likely to be within this data group after clonal expansion of specific B cells. Differences were observed between the frequency abundance of the antibody isotypes; on average, 58 of the top 200 IgM clusters were >0.1% abundant whilst an average 34 IgG clusters were but these were statistically insignificant (P=0.23, ANOVA: single factor). The largest IgM clusters in African buffalo represent 2.8% of the repertoire whilst the largest IgG clusters represent only 0.6%; this difference in abundance is significant (P=0.003) and is consistent with previous findings in mice (Williams et al., 2000; 231). The significance of the

dominant cluster does not translate to significance between animals or change over time. Differences are also observed between the two African buffalo animals; 62 clusters in A7 are greater than 0.1% abundant whilst 30 clusters in A11 were. This suggests a greater diversity in the antibody repertoire of A11 compared to animal 7 and greater diversity in the IgG repertoire than IgM during the initial stages of infection. The frequency abundance of clusters over 0.1% also increases over time, an average of 26 are >0.1% at day 0, 50 at day 8 and 62 at day 14 suggesting the expansion of specific clones in response to infection but the change was insignificant also (P=0.35, ANOVA: single factor). The African buffalo IgM A11 day 14 library appears to contain larger clusters as the percentage relative abundance of the clusters is higher. However, this is a consequence of less *IGHV* region sequences being isolated from this sample (figure 5.10).

The observations in African buffalo are mirrored in the cattle data. At day 0 in cattle the average number of clusters over 0.1% abundance is 89 which increases to 96 at day 7 but then falls to 77 at day 21. The frequency abundance of clusters over 0.1% significantly varies between animals (range 74-110, P=0.01) and but does not significantly change over time (P=0.34).

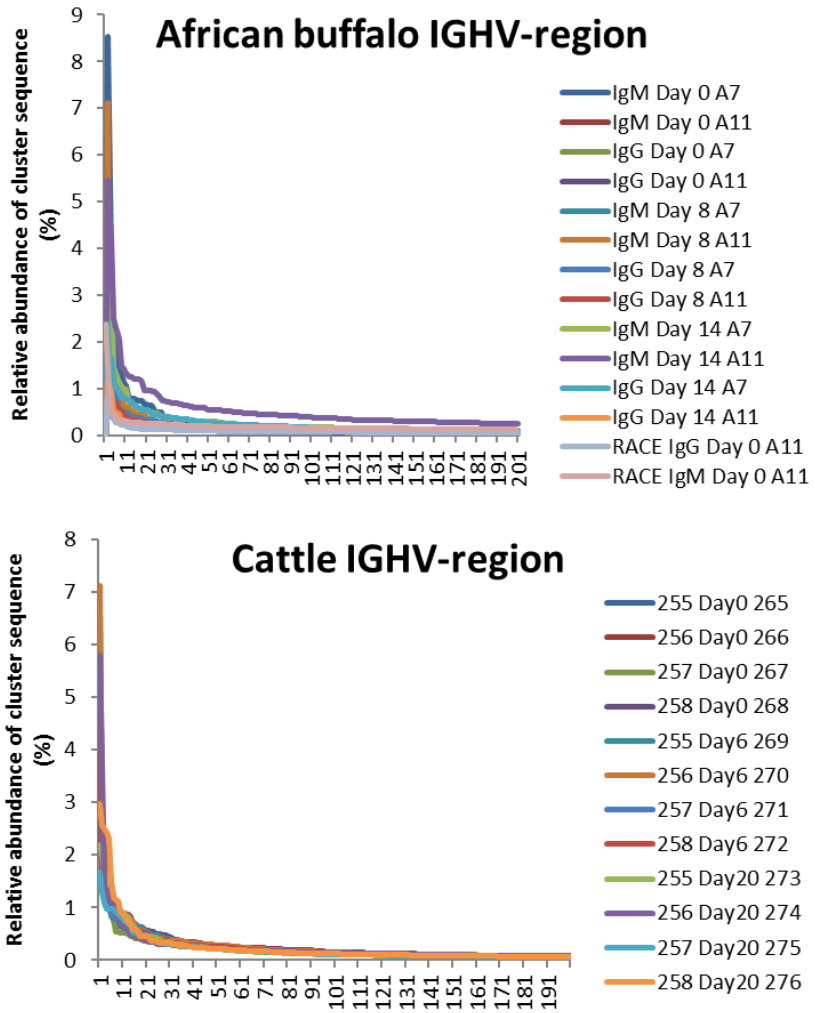


Figure 5.10: Relative abundance of the largest 200 African buffalo and cattle transcripts following challenge with SAT1 FMDV infection or immunisation respectively.

5.3.9 African buffalo and cattle *IGHV* length

The variability of the length of the *IGHV* regions in the African buffalo and cattle was calculated. As described in section 5.2.4.1, the reads in each sample were filtered for sequences that spanned from the *IGHV* leader primer to the Cys104 in FR3. The *IGHV* regions were trimmed to reflect only the framework regions (FR1, FR2 and FR3) and the CDR1 and CD2. A frequency data array was then generated for all the full-length *IGHV* regions within each sample.

In both African buffalo and cattle, *IGHV* region length is predominantly 110 amino acids in every sample (Figure 5.11). The variability in length ranged from 101 to 115 amino acids but the frequency of any length other than 110 amino acids was less than 1%. The length of the *IGHV* region did not alter after challenge of African buffalo and cattle with SAT1 FMDV. Therefore, an average 825,587 *IGHV* regions (97.6% of the total average 845598 isolated *IGHV* regions) were uniform in length in every African buffalo sample and 97.4% of the 129,000 *IGHV* regions in cattle.

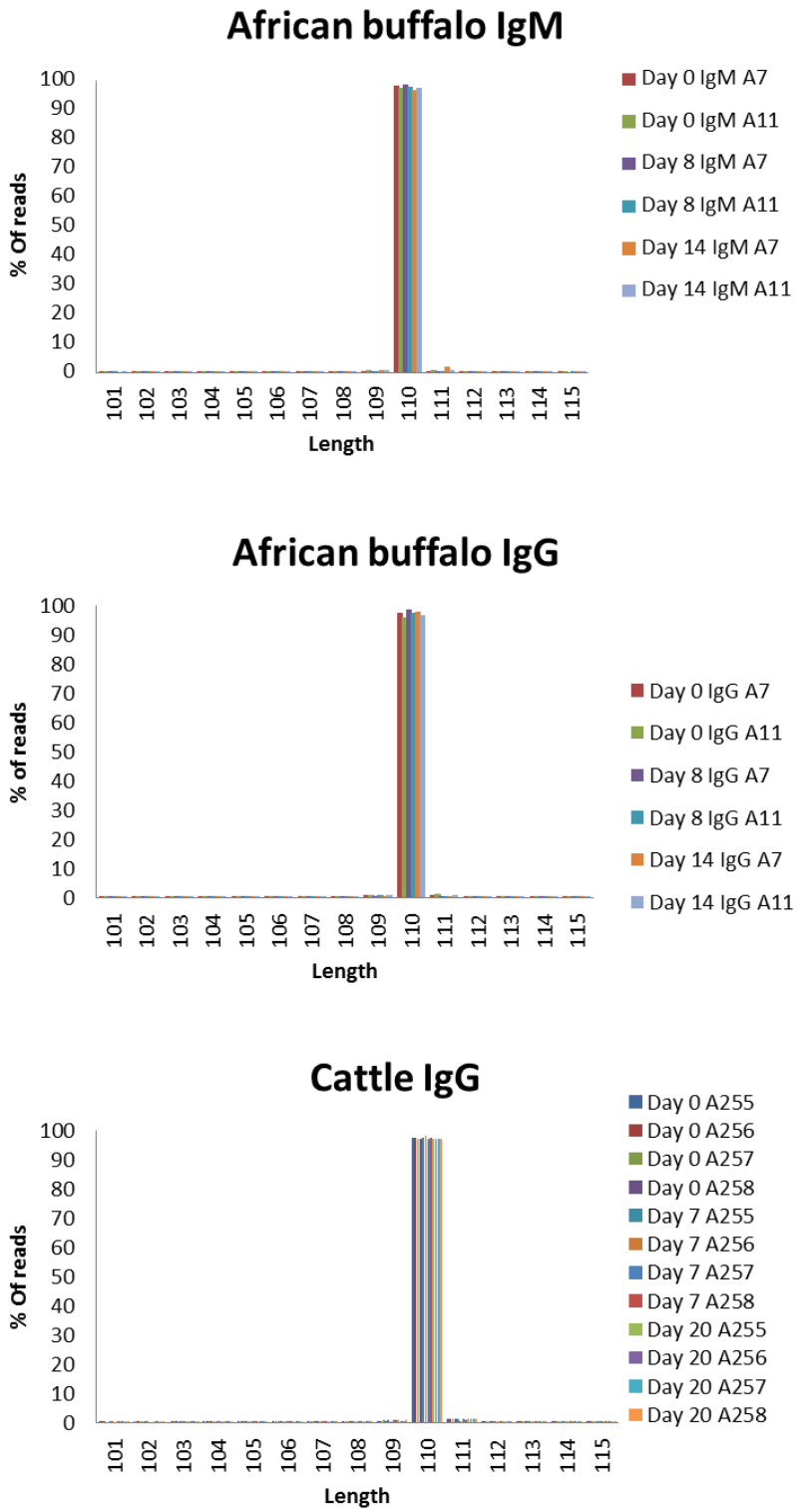


Figure 5.11: Total amino acid length of the *IGHV* regions in the African buffalo IgM (A), IgG (B) and cattle IgG (C) repertoires at day 0 and then post challenge with SAT 1 FMDV at day 8 and day 14 in African buffalo and day 7 and day 20 in cattle.

5.3.10 Variation of the amino acid residues occurs within the predicted CDR regions

The extent of amino acid substitutions at each position in the *IGHV* region, in the absence of antigen, was determined by calculating the percentage deviation from consensus at each amino acid position in the day 0 libraries. The *IGHV* region sequences within the largest 200 clusters of each African buffalo and cattle day 0 sample were aligned and a consensus sequence generated. At each amino acid position the most abundant amino acid and the total number of possible amino acids was determined. The percentage of transcripts containing an amino acid other than the most abundant was then calculated. The regions of high variability fall within the predicted CDR1 and CDR2 regions in both species (Figure 5.12). High levels of somatic hyper-mutation are occurring within and around the CDR regions of the antibody transcripts before FMDV challenge, resulting in amino acid variability.

Within the African buffalo CDR, the percentage deviation from the generated consensus sequence of the largest 200 clusters, ranges from 0.4 - 58.7% in CDR1 and 1.9 - 81.2% in CDR2 (average of 10.7% and 40.4% respectively). The variability in CDR2 is significantly higher than CDR1 (P=0.0001, ANOVA single factor). FR1 is the least mutated with an average of 2% deviation from the consensus whilst FR2 is the most mutated with 0.3-74% deviation from the consensus (average 16.3%). FR3 has the highest range from 0.2-82.1% but the average deviation is 10.5% from the consensus. The percentage deviation between the FR and the CDR regions is significant (P=0.007, ANOVA single factor).

The percentage deviation in cattle is lower than the African buffalo; percentage deviation from the consensus sequence ranges from 1.4-67.6% in CDR1 and 4-64% in CDR2. Framework 1 is the least mutated with 0 - 25% deviation from the consensus; FR2 contains the most somatic mutations, 0.3-55% and FR3 is 0.2-48.4%. The percentage deviation between the framework regions is significant (P=0.02, ANOVA single factor) and significant between the percentage deviation within the framework regions compared to the CDR regions (P=0.0008, ANOVA single factor).

Overall, there was no significant difference between the amino acid variability of the IgM and IgG repertoire at day 0. The amino acid variability within the CDR regions is high, as would be predicted, whilst variability in the FR regions around the CDR is also high.

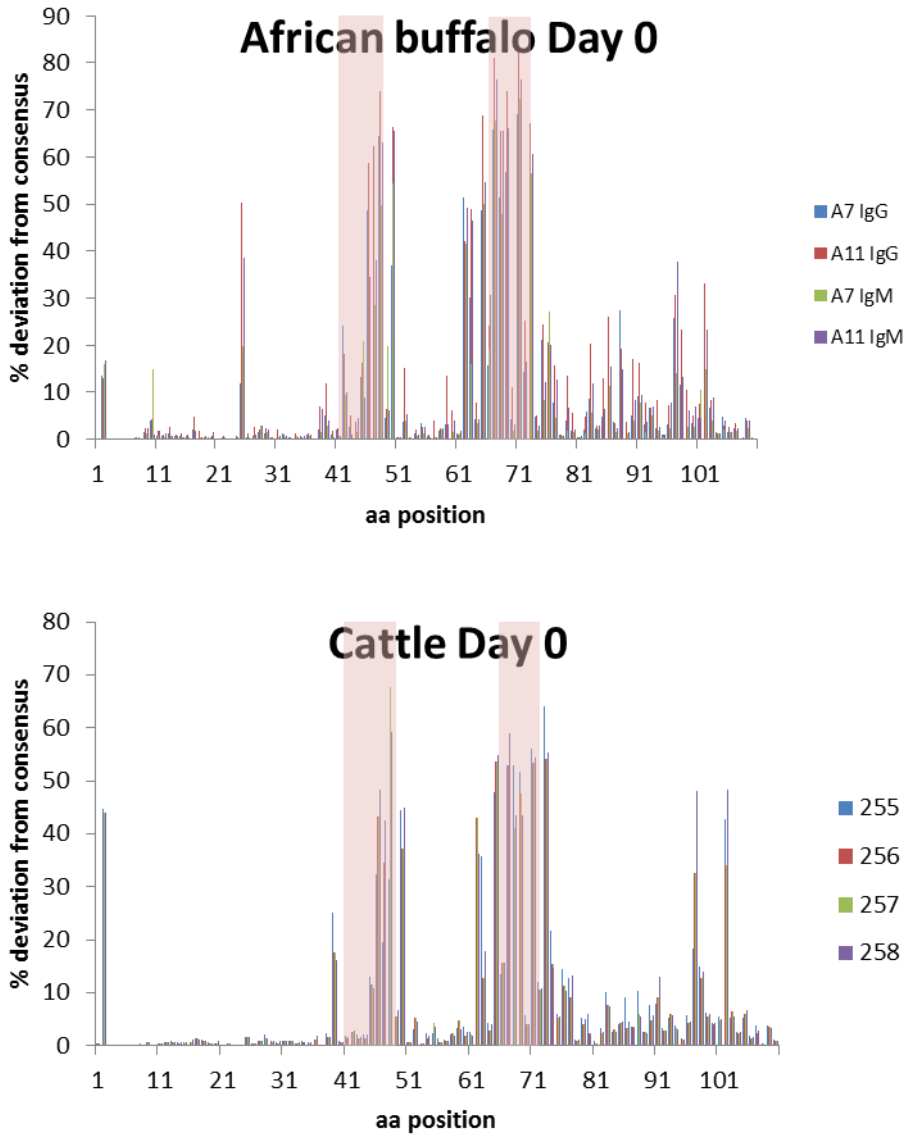


Figure 5.12: Amino acid variation in the African buffalo and cattle at day 0. The largest 200 clusters from each sample were taken from the day 0 samples prior to antigen stimulation and a consensus sequence was generated. The percentage deviation at each amino acid position of the consensus sequence was calculated following the alignment of reads. The CDR1 and CDR2 regions as outlined by the IMGT protein rules are shaded in red.

5.3.11 Quantification of post-translational modifications to African buffalo and cattle V-regions

Following the calculation of amino acid % deviation substitutions at each position in the *IGHV* region in day 0 sequences, section 5.2.6, the variation of amino acids following challenge with SAT1 FMDV in the African buffalo and cattle *IGHV* regions was determined. The sequences within the largest 200 clusters of African buffalo day 8 and day 14 and cattle day 7 and day 20 were aligned and a consensus sequence generated. The most abundant amino acid and the total number of possible amino acids were determined at each position and the percentage of reads containing an amino acid other than the consensus was calculated. The day 0 percentage deviation of each animal was then subtracted from the latter two time points to quantify the changes in somatic hyper-mutation in the reads after infection of the African buffalo and inoculation of cattle with SAT1 FMDV.

The majority of sequence variation at day 8 and day 14 in African buffalo and day 7 and day 20 in cattle occurs within the CDR regions, as observed in section 5.2.6. Significant variation occurs between the framework regions as FR1 contains the least mutations and FR2, the highest. If gene conversion was occurring, the levels of variability within the FR regions would be expected to be more uniform but the process of gene conversion cannot be entirely ruled out.

Significant variation occurs following FMDV challenge in African buffalo and cattle. Large increases in variation are observed in the CDR1 and CDR2 and to a lesser extent in the FR2 and FR3 regions. In multiple amino acid positions the amino acid variation increases upon challenge however, in certain positions the variation also significantly decreases as positions within both the CDR and FR regions become more conserved (Figure 5.13).

In African buffalo, large increases in the positive variation of the IgG CDR1 is observed whilst in IgM, the variation in CDR1 is mostly reduced (Figure 5.13). FR1 is the least variable, without significant change over time whilst the majority of variation in FR2 occur at positions 61 -63 before the CDR2 sequence. Both increases and decreases at individual positions occurs within the CDR2, with no obvious pattern between the antibody isotypes. In cattle, the majority of loss of amino acid variability occurs within and around the CDR sequences (Figure 5.13). Differences between animals are observed although these are not significant.

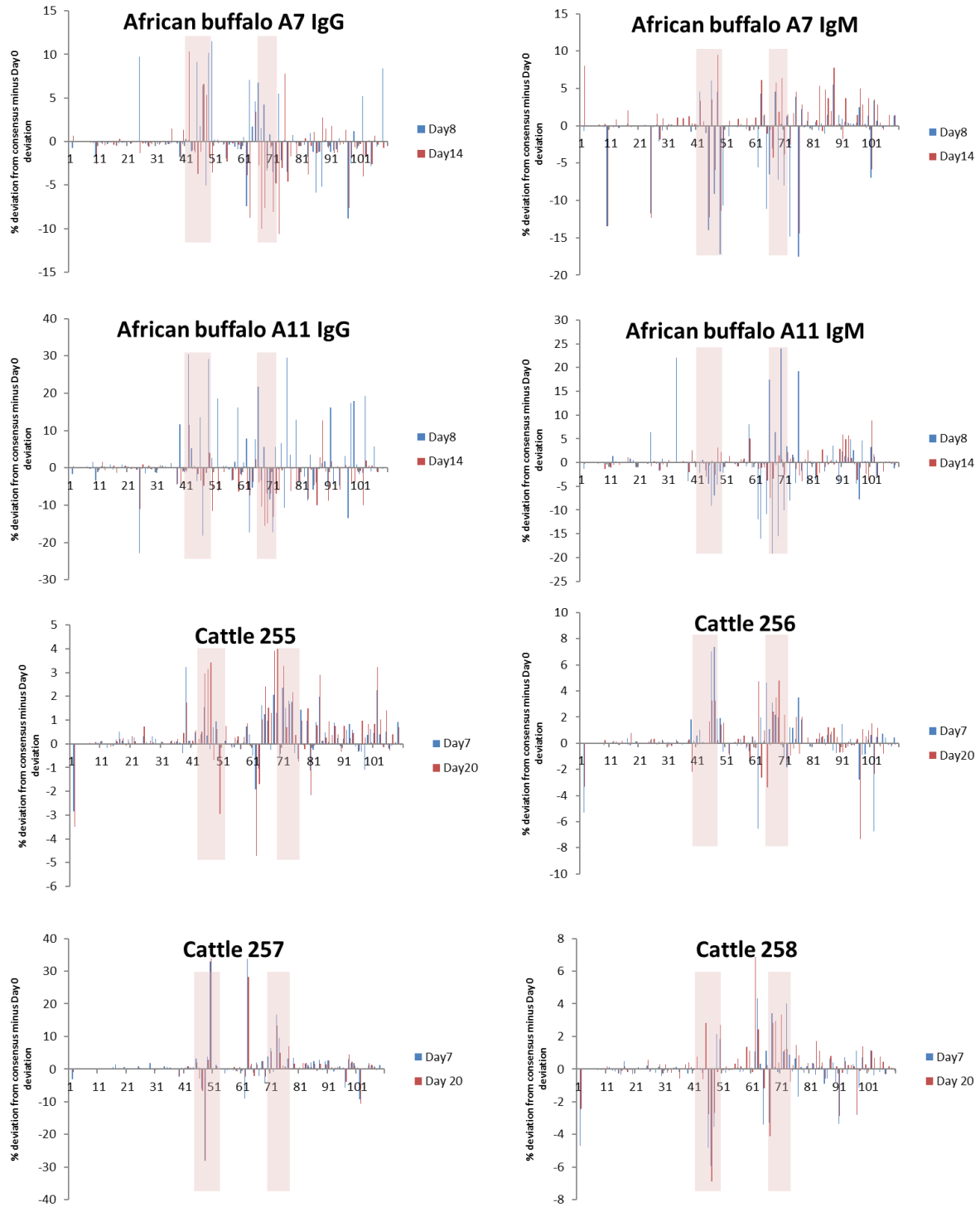


Figure 5.13: Amino acid variation in the African buffalo and cattle *IGHV* regions after challenge with SAT1 FMDV at day 8 and day 14 in African buffalo and at day 7 and day 20 in cattle. The *IGHV* regions from the largest 200 sequences in each sample were aligned and a consensus sequence generated for calculating the percentage of sequences with amino acid deviations at each position. The percentage deviation from day 0 in each animal was subtracted from the latter two time points to estimate changes in response to infection. The CDR1 and CDR2 regions as outlined by the IMGT protein rules are shaded in red.

5.3.12 Estimated *IGHV* gene usage in African buffalo and cattle

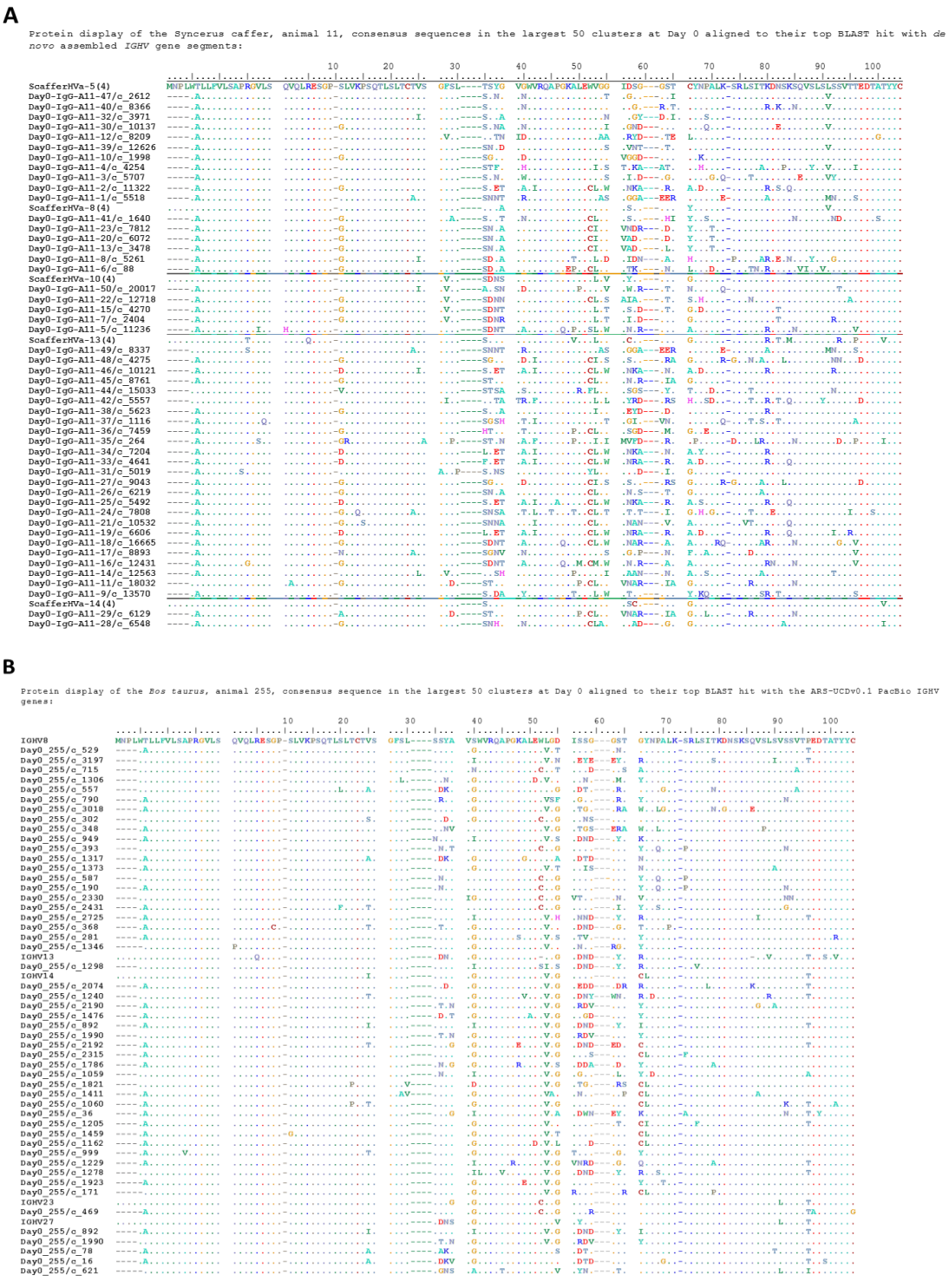


Figure 5.14: IMGT protein display of the African buffalo (A) and cattle (B) IgG consensus sequences from the largest 50 clusters at day 0 aligned to their top BLAST hit. The African buffalo *IGHV* consensus were aligned to the *de novo* assembled putatively functional *IGHV* gene segments and the cattle *IGHV* consensus were aligned to the PacBio ARS-UCDv0.1 *IGHV* genes.

The *IGHV* gene segments were previously *de novo* assembled from the African buffalo genome reads, Chapter 3 section 3.3.9, and of the 57 assembled, a total of 13 were estimated to be putatively functional. The consensus *IGHV* region sequences from the largest 50 clusters in the IgG repertoires of African buffalo A11 and A7 were mapped to the 57 *de novo* assembled gene segments to estimate germline usage.

All of the *IGHV* IgG consensus sequences mapped to the putatively functional *de novo* gene segments with an average 85.6% nucleotide identity with their top BLAST hit. High variability is seen in the CDR sequences between consensus transcripts and the gene segment (figure 5.14A). As we have seen in section 5.2.6, the highest variability is in the CDR; CDR2 appears more variable than CDR1. At Day 0, the consensus transcripts mapped to five *IGHV* gene segments, however the assembled gene segments are unlikely to represent all of the functional gene segments in African buffalo. The comparison is therefore likely to assign *IGHV* region transcripts to the wrong gene segments and makes estimation of SHM difficult. Variability in FR1 is low; the presence of multiple transcripts with a G at position 11 in FR1 suggests either functional gene segments are missing or the mapping bias of the African buffalo reads has caused assembly errors in the gene segments (Figure 5.14A).

Determining the gene segment each consensus transcript arose from is difficult as the germline sequences were highly similar and the *IGHV* transcripts had multiple point mutations. In African buffalo, it appears the gene segment usage changes significantly over time (Figure 5.15). At day 0, *IGHVa-13* was the predominant gene segment but this dropped significantly by day 8 (P=0.0033). The *IGHVa-10* significantly increased upon infection (ANOVA 1 way P=0.0029). In total, 8 of the 13 putatively functional germline *IGHV* were transcribed but this is not yet conclusive based on the germline assembly to date.

The 10 putatively functional cattle *IGHV* gene segments were previously identified in the ARS-UCDv0.1 assembly and confirmed in both the IGH sequence from Ma et al (Ma et al., 2016; 98) and the UMD3.1 *Bos taurus* genome, outlined in Chapter 3, section 3.3.9. These gene segments are likely to represent all of the functional gene segments in this cattle assembly but breed variation may occur and allelic variants are likely present in the population. The consensus *IGHV* region sequences from the largest 50 clusters in each cattle IgG library were mapped to all of the annotated gene segments in the ARS-UCDv0.1 to estimate the gene usage in cattle. All of the cattle IgG *IGHV* regions mapped to the putatively functional gene segments in the ARS-UCDv0.1. The average nucleotide percentage identity

between the cluster consensus and the top BLAST hit was 91.6% at day 0, 91.7% at day 7 and 91.7% at day 20. The *IGHV-8* and *IGHV-13* were predominantly transcribed at every time point and no significant changes in gene usage were observed after SAT1 immunisation (Figure 5.15). Somatic mutations are observed between the gene segments and the consensus transcripts (Figure 5.14B). As we have previously observed in section 5.2.6; CDR2 is the most variable. A mutation bias of particular framework positions exists, particularly in FR2, where the majority of consensus transcripts are mutated to a G at position 40, V at 53 and G at 55 however it is more likely that sequence error exists in the ARS-UCDv0.1 genome.

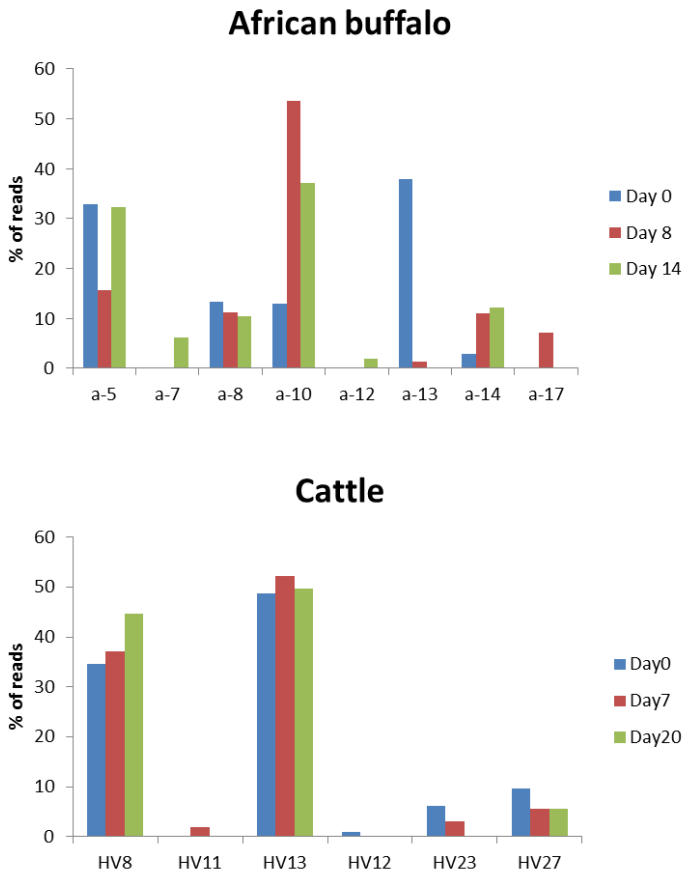


Figure 5.15: Estimated gene usage in African buffalo and cattle at day 0 and the subsequent time points. The consensus sequence of the largest 50 clusters in the IgG libraries of the African buffalo animal 11 and cattle animal 255 were aligned to the *de novo* assembled African buffalo gene segments or the cattle ARS-UCDv0.1 assembly respectively. The top BLAST hit was used to assign the gene to each cluster. The number of reads in each corresponding cluster was then used to calculate the percentage of reads aligning to each gene segment.

5.3.13 CDR3 clustering pipeline

5.3.13.1 Determining the identity score for CDR3 clustering

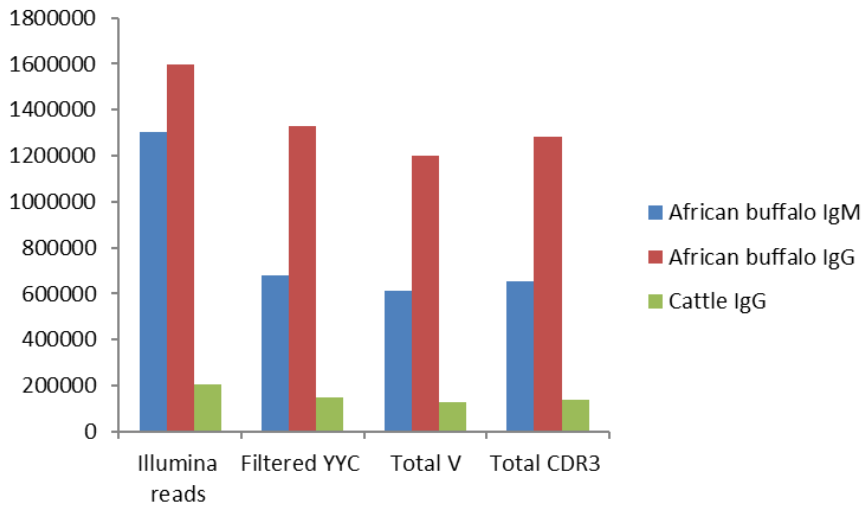


Figure 5.16: Filtering pipeline for the isolation of *IGHV* region and CDR3 sequences. The average number of Illumina sequence reads is shown before they were filtered for sequences less than 300 bp in length, translated and the correct reading frame selected by the presence of the FR3 motif "YYC" and the absence of stop codons. The *IGHV* regions were then isolated by trimming the sequence after the Cys104 and the CDR3 were isolated by trimming off the FR regions before the Cys104.

Following the filtering of low quality reads less than 300 bp in length and the merging of paired reads with FLASH, outlined in section 5.2.4.1, the African buffalo and cattle CDR3 regions in the transcripts were isolated. Both the merged and unmerged reads were translated and the correct open reading frame selected for by the FR3 motif YYC and its derivatives and the elimination of reads containing stop codons. The 5' primer was then used to select reads containing full length CDR3; these reads were then trimmed from the Cys104 at the end of FR3 to the W in the *IGHJ* gene segment W/F-G-X-G motif. An average 652,000 IgM and 1.29 million IgG CDR3 sequences were isolated per sample, equating to ~48.2 and 80.4% of the IgM and IgG raw reads respectively (Figure 5.16). An average 63,000 more CDR3 sequences were isolated than *IGHV* regions in African buffalo and 11,000 in cattle.

The African buffalo CDR3 regions were clustered using multiple different cluster identity scores to achieve an optimum clustering value: each 1% interval between 84% and 99% mismatch identity was used. The number of sequences that occurred once or more in each library, the deduplicated sequences, were counted. An average of 168,000 deduplicated sequences occur which equates to the number of clusters that would form when no mismatches are allowed between sequences; the minimum cluster size would therefore be 5.4 with a 100% mismatch identity. As the identity score increased the average number of clusters also increased but this was more pronounced in some samples than others and the rate of change was different than the clustering of the *IGHV* regions in section 5.2.9 (Figure 5.17A). The variance in the number of clusters in each sample increases incrementally until 93% and 97% mismatch identity where the "elbows" in the variance are observed (Figure 5.17B). The average cluster size decreases as the mismatch identity increases (Figure 5.17C) and the variance decreases incrementally until 97% mismatch identity when the variance levels out (Figure 5.17D).

The PhiX library in each sample provided a quality control measure for the Illumina sequencing. The PhiX reads, accounting for ~0.5% of the reads in every sample, was pooled and aligned to the PhiX genome. The error rate estimate in the sequencing was 2.04%. At the lowest mismatch identity, 85%, the percentage of un-clustered sequences were higher than the inherent error rate in the sequencing (Figure 5.17E). The percentage of reads that were clustered decreases slowly as mismatch identity is increased (Figure 5.17F).

The mismatch identity of 93% and 97% were thus selected for further analysis. The cattle CDR3 were also clustered at 93% and 97% for direct comparison to the African buffalo CDR3. The average number of clusters produced in African buffalo at 93% is 139,000 and 180,000 at 97% and the average size of the clusters is 6 and 5 respectively. In cattle, the number of clusters is 49,000 and 57,000 at 93% and 97% mismatch identity respectively, with an average cluster size of 2.8 and 2.3. The small change in average cluster size between mismatch identities is due to the largest cluster in each sample containing 663 sequences on average at 93% and 627 at 97% suggesting cattle have a large number of *IGHV* region sequences that are highly similar.

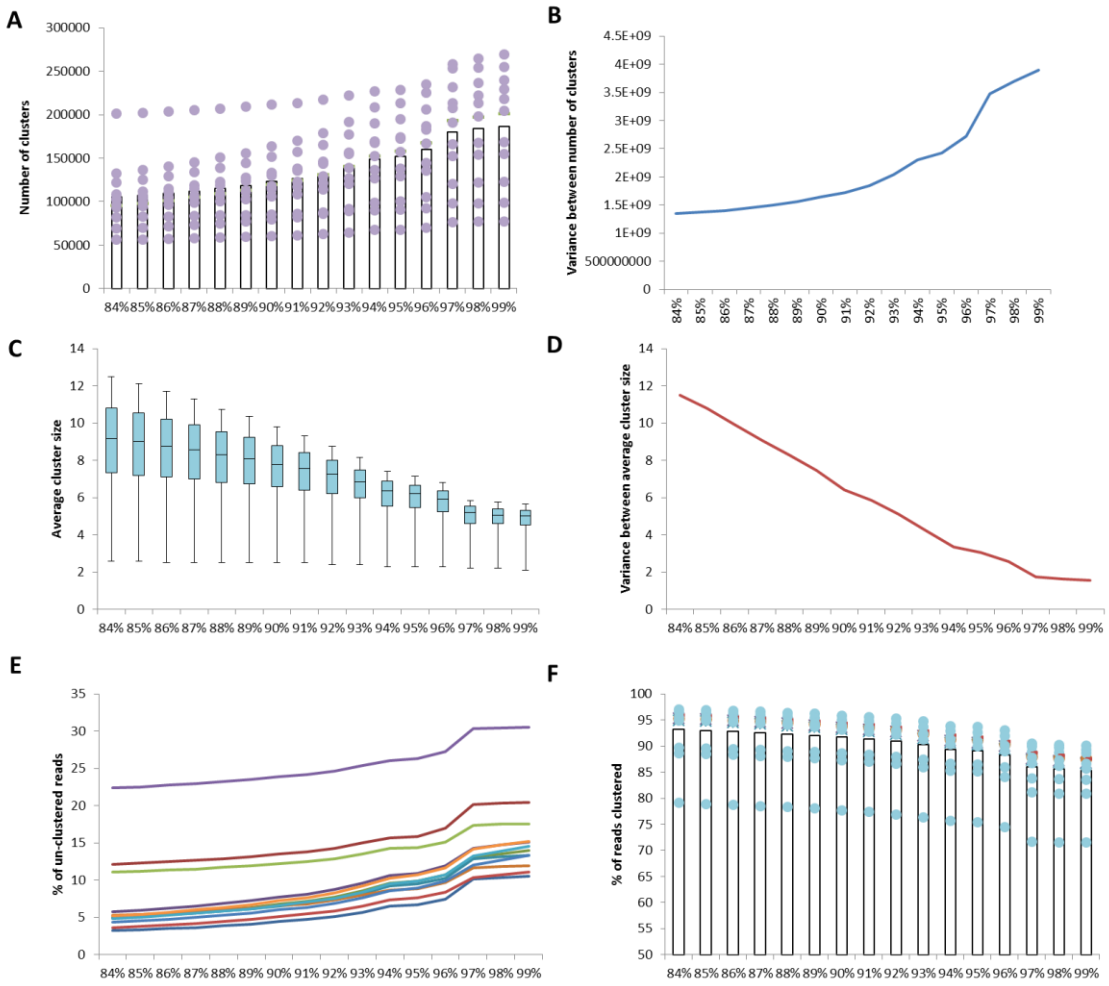


Figure 5.17: The African buffalo IgM and IgG reads were clustered at every identity scores between 84% - 99% (in 1% increments) for determining the optimum clustering parameters. The number of clusters for each mismatch identity (A) is shown with the blue dots indicating each sample and the black bars representing the mean number of clusters. The variance of cluster number between samples is displayed to the right (B). The average cluster size is indicated by the dot and whisker plots, the central bar indicating the mean and the variance between average cluster sizes of samples is shown (D). The percentage of the average un-clustered reads (E) increases as mismatch identity increases. The dashed red line indicates the PhiX calculated error rate in the Illumina sequencing. The percentage of clustered reads in each sample differs between samples and this difference increases as mismatch identity increases (F).

5.3.13.2 The abundance of CDR3 transcripts in African buffalo and cattle

The relative abundance of the CDR3 regions in the African buffalo IgM and IgG and the cattle IgG transcripts was assessed by calculating the abundance of the largest 200 clusters in each sample (figure 5.18). The largest 65 IgM clusters were on average >0.1% abundance whilst 37 of the largest IgG clusters were. The frequency abundance of clusters >0.1% increased from an average 56 at day 0, to 70 at day 8 but then decreased to 27 by day 14. The abundance of clusters >0.1% differed significantly between the two animals (P=0.04, ANOVA: single factor) with Animal 7 having an average of 71 whilst animal 11 had an average of 30. This suggests a greater diversity in the antibody repertoire of animal 11 compared to animal 7, as reflected in the abundance of *IGHV* regions in section 5.2.4.3. In cattle the abundance of clusters >0.1% doesn't significantly change over time or between animals. The number of clusters for animals 255, 256, 257 and 258 is on average 119, 159, 114 and 130 respectively. As seen in African buffalo, the average number of clusters >0.1% is 131 at day0 which increases to 150 at day 7 but decreases to 11 at day 20. This suggests in both cattle and African buffalo, in both the CDR3 cluster abundance observed here and the *IGHV* region abundance observed in section 5.2.3.4, that the antibody diversity is reduced in the initial stages of infection by day 7 and 8, possibly due to the clonal expansion of specific B cells, but the repertoire is diversified by day 14 and day 20.

The largest African buffalo IgM clusters in each library represent, on average, 3.8% of the repertoire whilst the largest IgG clusters in each library represent only 0.7%. This difference in abundance of the largest IgM and IgG cluster within each animal is significant (P=0.003, ANOVA: single factor).

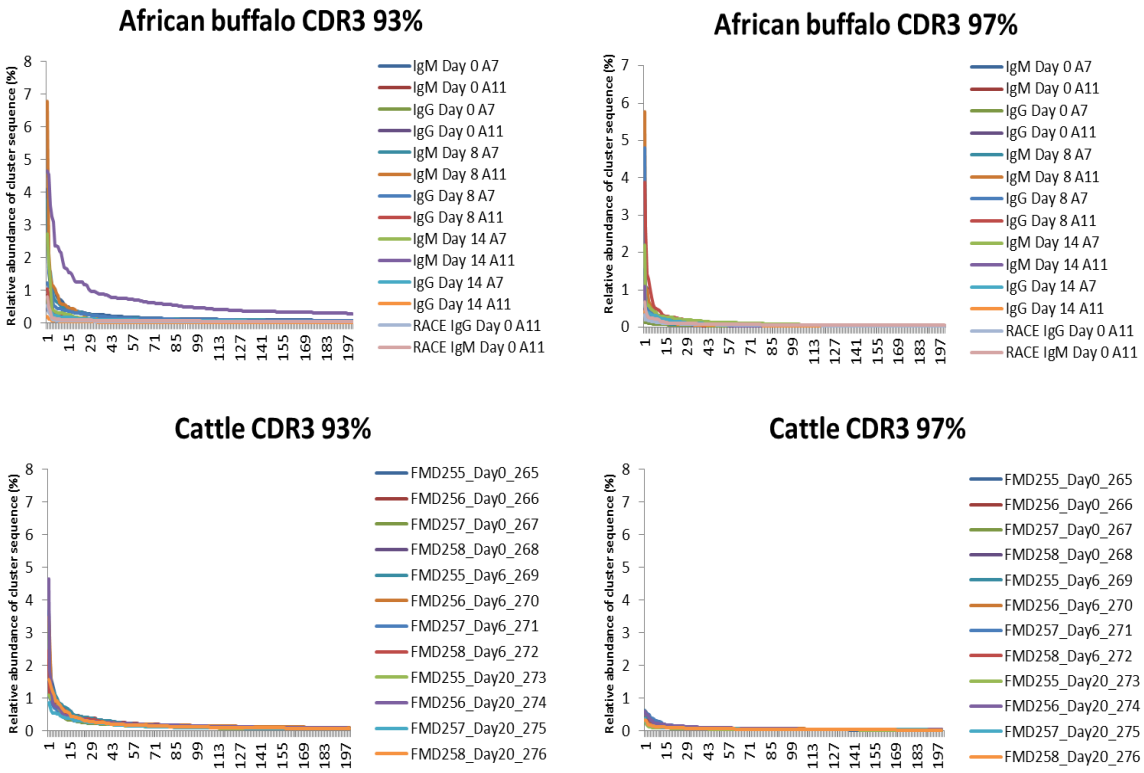


Figure 5.18: Relative abundance of the largest 200 African buffalo and cattle CDR3 clusters following challenge with SAT1 FMDV infection or immunisation respectively. The CDR3 sequences were sorted by length and then clustered based on optimum mismatch identities of both 93% and 97%.

5.3.14 African buffalo and cattle CDR3 length

The African buffalo and cattle CDR3 are highly variable in both length and amino acid composition (Figure 5.19). The highly variable length of the CDR3, including the presence of ultra-long CDR3 sequences was thought to be novel to cattle but here we show that African buffalo also have highly variable CDR3 lengths and produce ultra-long CDR3.

The African buffalo CDR3 lengths range from 1 to 72 amino acids with a trimodal distribution of lengths that have predominant mean peaks of 25 and 67 amino acids (figure 5.19). This trimodal distribution did not significantly alter following infection with SAT1 FMDV with no clear change in the peak of distributions. The ultra-long sequence peak in the IgM however is much less pronounced with 1.5% of reads >41 amino acids compared to 5% of reads in IgG. The difference between the frequency of ultra-long CDR3 in IgM and IgG is significant (ANOVA single factor, P=0.0001).

The cattle CDR3 lengths are also highly variable, ranging from 7 to 80 amino acids (Figure 5.19). However, the trimodal distribution is much less pronounced in this data set than has been seen in previous cattle data (Wang et al., 2013; 92). The mean peak occurs at 24 amino acids with a very small peak observed at 64 amino acids; >0.5% of reads in cattle were greater than 41 amino acids in length. Similar to the African buffalo, the distribution of lengths did not alter significantly after SAT1 immunisation and whilst variation is observed between animals within each species, this variation is insignificant.

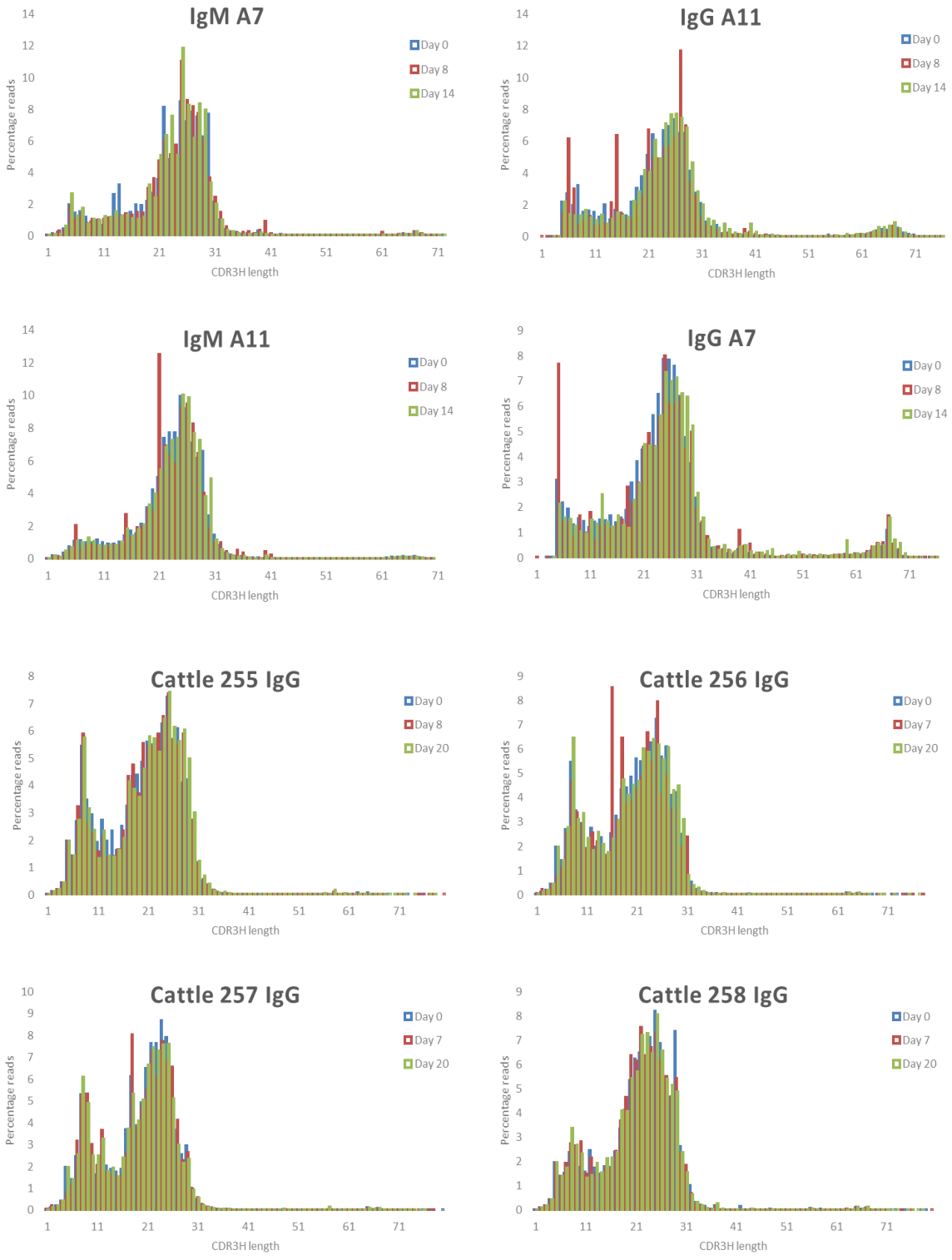


Figure 5.19: Total amino acid length of the CDR3H in African buffalo and cattle at day 0 and then following subsequent challenge with SAT1 FMDV.

5.3.15 African buffalo produce ultra-long CDR3 sequences

The cattle ultra-long sequences accounted for, on average, only 0.5% of the total CDR3 in each library. However, previous studies have shown that the ultra-long CDR3 account for 5-10% of the cattle antibody repertoire (Saini et al., 1999; 232) but that the exceptionally long CDR3 are isotype restricted to IgM (Kaushik et al., 2009; 120). In the African buffalo we see the opposite, the ultra-long CDR3 sequences, ranging between 50 – 71 amino acids, account for ~1.1% of the total CDR3 in IgM and 4.9% of the total CDR3 in IgG. In the IgM repertoire, there is no change in the length distribution of the ultra-long CDR3 over time (0.1% at day 0, 0.08% at day8, and 0.08% at day 14). In the IgG repertoire however, the peak accounts for 3.4% at day 0, 3.8% at day 8 and 4.7% at day 14. Whilst this increase is not significant it shows a potential use of ultra-long antibodies in response to FMDV infection. The cysteine diversification in the ultra-long antibodies was high; 10% of the amino acids in the ultra-long sequences were cysteine at day 0, 12.7% at day 8 and 11.4% at day 14. The ultra-long antibodies are therefore capable of forming various disulphide bond structures with numerous cysteines. High levels of glycine and tyrosine were also found in the ultra-long CDR3, accounting for 12% and 8.3% respectively.

The ultra-long IgG CDR3 sequences in African buffalo and cattle were aligned within each transcript library and a consensus sequence generated based on the most dominant amino acid at each position, as outlined in section 5.3.10. These consensus sequences were then aligned with the *IGHD* amino acid sequences from the cattle ARS-UCDv0.1 and the *de novo* assembled *IGHD* genes in the African buffalo. The phylogenetic relationship was inferred by using the Maximum Likelihood method based on the JTT matrix-based model (Whelan and Goldman, 2001; 233). A tree was generated in MEGA7 (Kumar et al., 2016; 228) with the highest log likelihood (-4313.01). The African buffalo and cattle ultra-long CDR3 group together, sequences within each species are more similar to each other than between the two species (Figure 5.20). Individual animals appear more similar between time points than compared to other animals, suggesting the sequences are not converging on a similar antigen specific sequence. Interestingly, the ultra-long *IGHD* gene segment aligns with the cattle ultra-long CDR3, confirming that their ultra-long CDR3 arises from this gene segment. An ultra-long CDR3 was not assembled in the African buffalo genome but it is likely that they possess one as none of the assembled *IGHD* aligned with their CDR3 sequences.

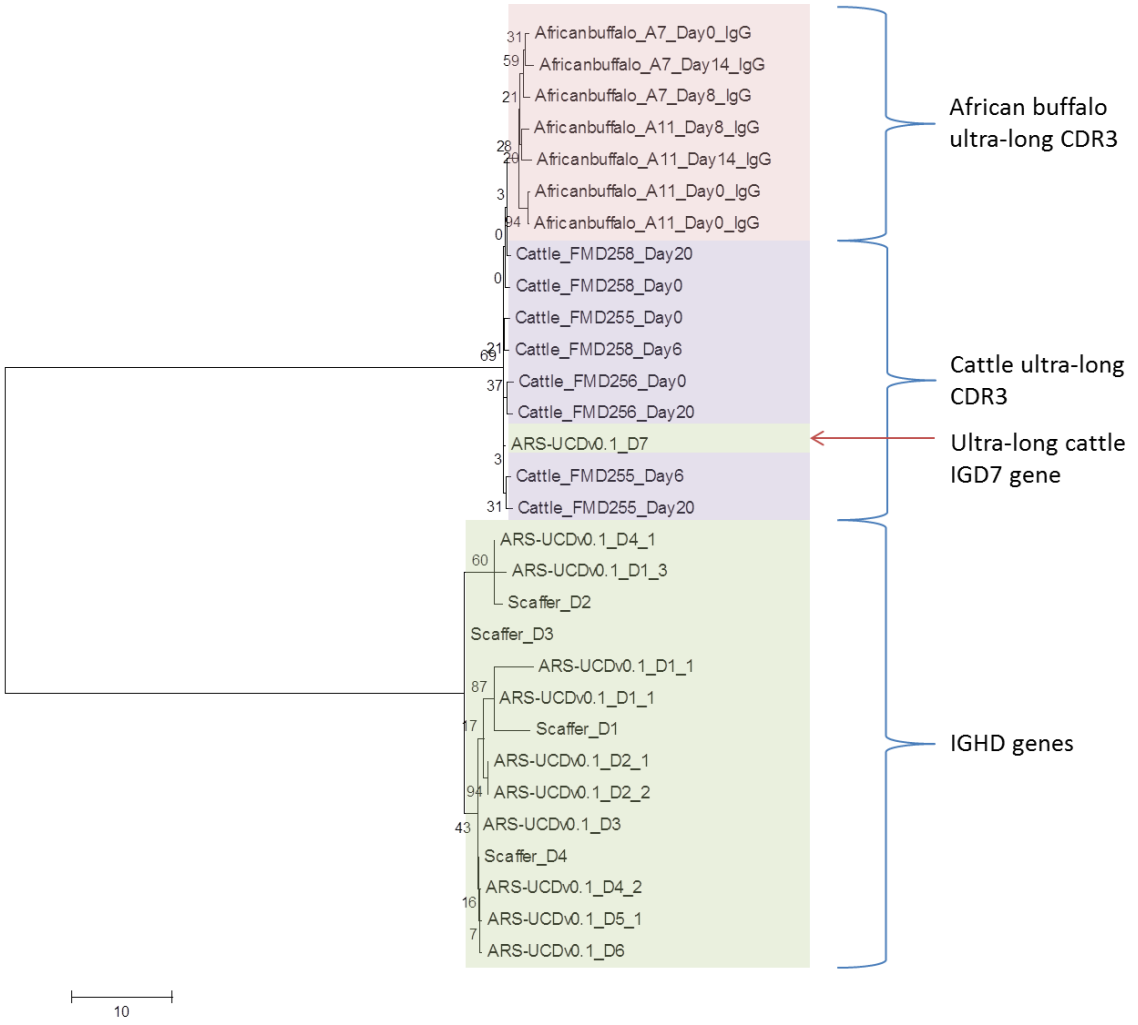


Figure 5.20: The phylogenetic relationship of the African buffalo and cattle ultra-long CDR3 consensus sequences compared to the *IGHD* genes was inferred by using the Maximum Likelihood method based on the JTT matrix-based model (Jones et al., 1992; 229). The consensus sequences were generated by aligning the ultra-long CDR3 sequences and the most abundant amino acid at each position forming the sequence. These were aligned with the *IGHD* genes in the ARS-UCDv0.1 assembly and the *de novo* assembled African buffalo *IGHD* genes. The tree with the highest log likelihood (-4313.01) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model, and then selecting the topology with superior log likelihood value. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 30 amino acid sequences. There were a total of 81 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 (Kumar et al., 2016; 228).

The change in the most dominant ultra-long CDR3 sequences in response to infection was then investigated in the African buffalo. Each IgG library contained an average of ~47,400 ultra-long CDR3 sequences whereas each IgM library had an average of only 4,900 ultra-long CDR3 sequences. The ultra-long CDR3 in each library were isolated and clustered with a 93% mismatch identity, determined as optimal in section 5.3.14.1. A representative sequence of the largest 50 clusters in each library was extracted and used to generate counts of highly similar sequences at each time point within each animal and antibody isotype. These counts were then used to generate stream graphs for visualising changes in the frequency of individual CDR3 sequences (figure 5.21).

The frequency of sequences similar to the largest 50 cluster representative sequences for the ultra-long CDR3 at each time point are shown as percentages of the total. The lower total number of ultra-long CDR3 in the IgM that went into the analysis makes the conclusions less definite but we observe a large proportional increase in a single ultra-long CDR3 sequence in animal 7 between day 0 and day 8: (VKVAGHAKTVRRDCVSCYGGGWSYTCCDDCYRDGRGTCTNCGRCVRSVYEEVTVLHWYL) which occurs at low frequency at the other two time points and is not present in animal 11. No specific expansion of ultra-long CDR3 occurs in the IgM repertoire of animal 11 but we observe the specific expansion of one sequence between day 8 and day 14 in the IgG repertoire (TKCFQMRGERTEKKRRNCSTCCRDAGAFSSDCERSCCHRWGCSLACYTERDYYVEDTTMSQWYHV). The IgG repertoire of animal 11 has smaller expansion of individual sequences, the largest sequence at day 8, not being found in the IgM repertoire of the same animal or in the other animal (AKAAEDRQTKKERKLKIDCNCGGDGRVGWCYSSGCCRWNYGWGCGVCDADHHCSEETTEVSHFYHV). Whilst the frequency then of the ultra-long sequences changes in response to FMDV infection, convergence between species is not observed and the specificity of these sequences for FMDV is not known.

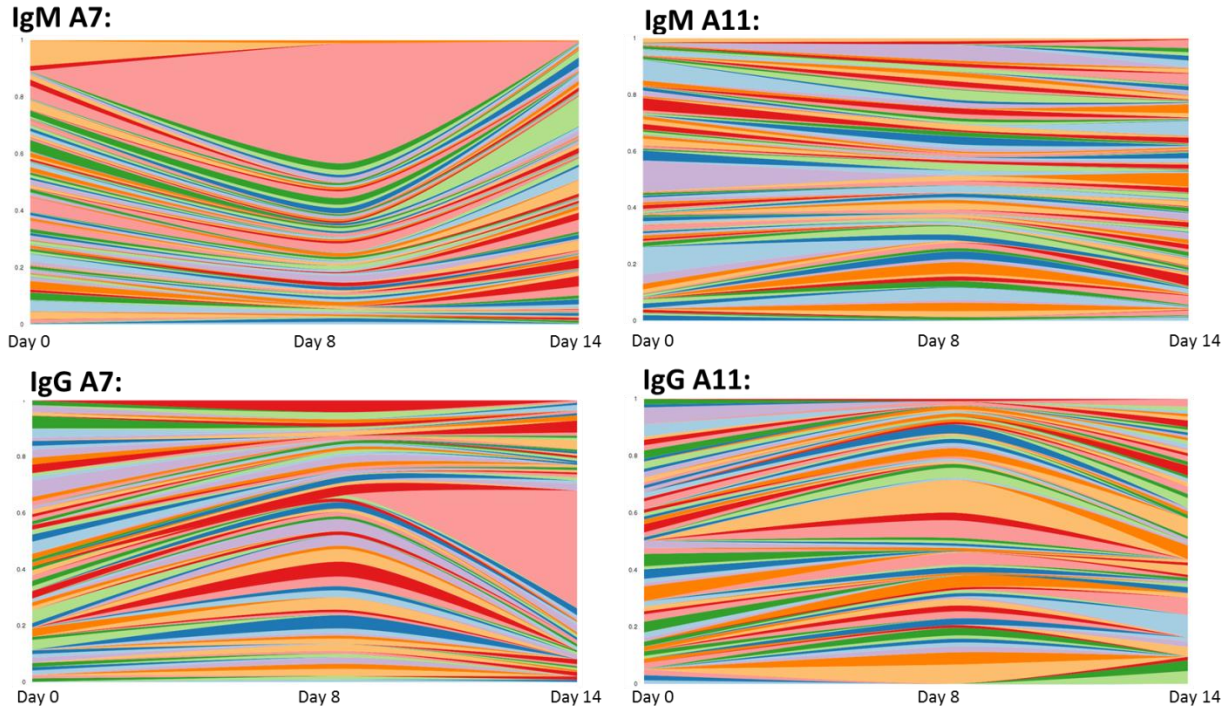


Figure 5.21: The frequency of the ultra-long CDR3 sequences in the IgM and IgG repertoire of the African buffalo following infection with FMDV SAT1. A representative sequence of the largest 50 clusters at each time point was identified at the two other time points with a 93% mismatch identity permitted between sequences.

5.3.17 The dominance of CDR3 sequences changes in response to infection

A representative sequence was selected from the largest 50 clusters of CDR3 at each time point in African buffalo animals 7 and 11 and cattle animals 255, 256, 257 and 258. These representative sequences at each time point were used to search for highly similar sequences in the other two time points to generate the frequency of sequences over time. These counts were then used to generate stream graphs for visualising individual sequence expansion and contraction following FMDV infection or inoculation in African buffalo and cattle respectively (Figure 5.22).

African buffalo have dramatic expansion and contraction of individual sequences following FMDV infection. The singular IgG CDR3 sequence expansion in animal 7 (STIMTW) is found in the IgM repertoire of animal 7 and in the IgM and IgG repertoires of animal 11 but here the frequency does not significantly change over time. The frequency of several IgG sequences in animal 11 increase between day 0 and day 8, the largest being 25 amino acids in

length, (AKQTHGSCDYSACAGPDYGYFGSYI) is not found in the IgM repertoire of the same animal or in animal 7. The expansion of a single sequence in animal 11 IgM (IKHSSTAGYACYGYNEAYN) is found in animal 7 (IKHTSTAGYACYGYNEAYN) at day 0 that increases in frequency at day 8 but not to the same magnitude.

Cattle, in comparison, have more limited expansion of individual CDR3 sequences. Considering the antibody transcripts in both cattle and African buffalo were generated with the same primer set and have been subsequently analysed with the same pipeline, this difference is significant. The expansion of African buffalo antibody transcripts in response to FMDV infection and the paucity of a similar response in cattle may account for the differential disease outcomes between these two species.

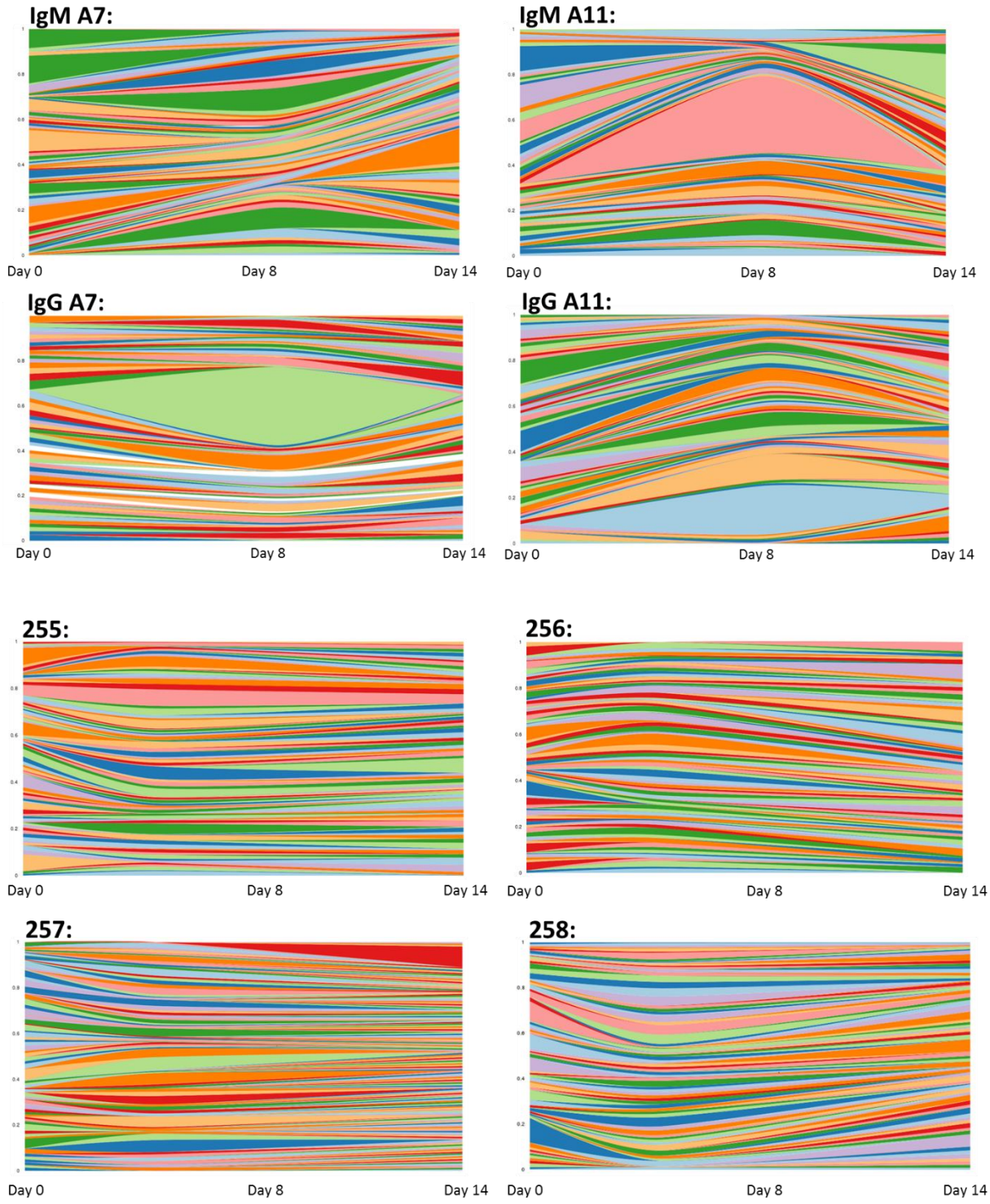


Figure 5.22: The frequency of CDR3 sequences in African buffalo and cattle after infection or inoculation with SAT1 FMDV respectively. Similar sequences, with a 93% mismatch identity, to the representative sequences of the largest 50 clusters at each time point were searched for in the other two time points within each antibody isotype and animal. The frequency counts are displayed as a proportion of the total number of sequences at each time point.

5.4 Discussion

The antibody response to SAT1 FMDV was characterised in African buffalo and cattle in the first comparative description of their differential antibody response. The African buffalo were infected with SAT1 FMDV and the IgM and IgG repertoire at day 0, 8 and 14 was sequenced using Illumina to quantify changes in their antibody response over time. Cattle, from a larger study not associated with this PhD (Grant et al; unpublished), were inoculated with SAT1 FMDV and the IgG repertoire was sequenced with Illumina at day 0, day 7 and day 20. Using the same analysis pipeline for the African buffalo and cattle Illumina sequence reads, changes to the antibody response including the levels of somatic hyper-mutation in both species, was interrogated. The day 0 sample allows the quantification of post-transcriptional modifications in the absence of FMDV infection to determine the diversity of the primary IGH antibody repertoire. The IGL repertoire in the African buffalo was also sequenced with Illumina at day 0 and day 14 post-infection to investigate the role of the light chains in the African buffalo response.

The important difference between the African buffalo and cattle data is the experimental procedure; the African buffalo were infected with SAT1 FMDV whereas cattle were inoculated with highly purified SAT1 FMDV antigen. FMDV rapidly replicates and spreads within an infected animal, mounting a rapid induction of DC maturation which leads to antigen presentation to B cells and stimulation of antibody secretion and CSR. Inoculation with vaccine antigen would not spread from the site of primary infection leading to a more localised activation of DCs cells and therefore a lower chance of specific antibody detection in the blood. Additionally, viral replication is required for the activation and maturation of DCs (Bautista et al., 2005; 60) which would not occur in the cattle following vaccination. Overall this would result in a reduced immune response which could account for the lack of response observed (figure 5.22).

The antibody transcripts in the African buffalo and cattle were amplified using the same primer sets and the Illumina sequence reads were analysed using an identical analysis pipeline. A lower number of average reads per sample was generated from the cattle than African buffalo, 200,000 versus 1.4 million respectively but the final analysis of the change in CDR3 frequency was proportional to the total number of CDR3 sequences isolated in each sample. The African buffalo have a clear response to SAT1 FMDV infection whilst cattle

inoculated with SAT1 antigen do not. A way of determining if the differences observed are due to the lower number of sequences analysed could be resolved by random sub-sampling of the African buffalo data for equivalent numbers of sequences to the cattle.

African buffalo and cattle appear to mount a considerably differential response to FMDV. Whilst the length distribution and percentage deviation of amino acid variation of the *IGHV* region and CDR3 are not considerably different between species, the African buffalo have dramatic expansion of individual CDR3 sequences between day 0 and day 8 in their IgM and IgG response which are replaced by an increase in alternative CDR3 sequences by day 14 (figure 5.22). This expansion and contraction of individual B cell clones is not observed in the cattle CDR3 over time.

The frequency of the number of *IGHV* region and CDR3 clusters with abundance greater than 0.1% of the repertoire is reduced at day 8 and day 7 in African buffalo and cattle respectively following infection or inoculation, suggesting the clonal expansion of specific B cells, but this frequency increases again by day 14 and day 20. This correlates with the proportional increase in individual CDR3 sequences at day 8 observed in section 5.3.17. The specificity of these sequences to FMDV is currently unknown but similar studies in mice have shown that highly abundant IGH transcripts present at day-6 post immunisation are antigen specific (Reddy et al., 2010; 218). This cannot be confirmed without confirming the antibody specificity of these sequences in the African buffalo.

Less diversity is observed in the IgM repertoire than the IgG of African buffalo. The frequency abundance of the largest IgM cluster is 3.8%, significantly larger than the largest IgG cluster which is only 0.7% and overall more IgM clusters were >0.1% abundant. The greater diversity in the IgG repertoire is consistent with previous findings in mice (Williams et al., 2000; 231) and is consistent with the IgM response providing a broader and less specific antigen binding sites whereas IgG being more specific to individual antigen.

The CDR3H and *IGHV* regions were trimmed and analysed separately in order to estimate differences in the frequency abundance and variability of the two regions. The CDR3H is responsible for the dramatic variation in length of the IGH transcripts and the majority of amino acid variability occurs within the CDR3; it is considered the main FMDV antigen binding site. The separate analysis therefore permitted more accurate optimum clustering of the two regions for more accurate downstream analysis. The relative abundance of the largest

200 CDR3H clusters is lower than that of the *IGHV* regions, supporting the greater diversity seen in the CDR3H.

Variability in the *IGHV* region in both African buffalo and cattle is much less than that observed in the CDR3H region. The length of the *IGHV* region in both species is extremely restricted; 97-98% of the transcripts were 110 amino acids, as is expected. High levels of amino acid variability occur within the CDR1 and CDR2, with significant changes following infection or inoculation with SAT1 FMDV, but their restriction in length limits their impact on the diversity of the repertoire. The frequency of amino acid variability between the CDR1 and CDR2 of cattle and African buffalo following FMDV challenge is not significantly different between species. Interestingly, in multiple amino acid positions this increases upon FMDV challenge however, in certain positions variation within and around the CDR sequences is lost as sequences become more conserved.

The CDR3 in contrast, have a bimodal distribution in length of 20-40 amino acids, with ultra-long CDR3 accounting for 1.5% of the total IgM CDR3 and 5% of the IgG CDR3 in African buffalo. The ultra-long cattle sequences in this data account for ~0.5% of the total CDR3 but in previous studies have been shown to account for 5-10% of the total CDR3 repertoire (Saini et al., 1999; 232). In contrast to previous reports, where the ultra-long CDR3 appeared isotype restricted to the cattle IgM (Kaushik et al., 2009; 120), it has been shown conclusively that they are not in both cattle (Walther et al., 2013; 234) and here in African buffalo. These long and ultra-long CDR3 does not alter following infection or inoculation of African buffalo and cattle respectively, suggesting the diverse range and distribution of CDR3 lengths is a novel diversification feature of their primary antibody repertoire.

The frequency of ultra-long CDR3 sequences in African buffalo changes following infection with SAT1 FMDV. The change is not uniform between animals or antibody isotypes; a large increase in an ultra-long CDR3 sequence occurs in the IgM of animal 7 between day 0 and day 8, suggesting it has specificity for FMDV. However, this sequence is not found in the IgG or other animal. Similarly, a large increase in an IgG ultra-long sequence occurs in animal 7 between day 8 and day 14. It has been demonstrated that the elongated CDR3H confers antibody binding capabilities (Ekiert et al., 2012; 235) but it is unknown if these antibody sequences are FMDV specific.

The ultra-long CDR3 in cattle are formed through the use of the ultra-long *IGHD* gene segment in the germline. This *IGHD* gene segment aligns phylogenetically with

representative ultra-long CDR3 sequences from cattle. An ultra-long *IGHD* gene was not assembled with the African buffalo genome reads, despite mapping the reads on the cattle ultra-long *IGHD* gene segment. It is likely the gene exists but was obscured by mapping bias to the shorter gene segments in the African buffalo. The assembled *IGHD*, other than the ultra-long *IGHD7*, do not align with the ultra-long representative CDR3 in African buffalo or cattle so it is likely the African buffalo ultra-long CDR3 arose from a similar gene to the *IGHD7*.

The African buffalo IgL sequences have limited variation in their length, frequency abundance or amino acid composition and show no significant change following infection with SAT 1 FMDV. The limited deviation in IgL transcript length in African buffalo, where ~76% of the light chain transcripts are 137 amino acids in length, has previously been demonstrated in cattle studies (Grant, 2013; 230). The Gaussian distribution of IgL lengths does not alter following SAT1 FMDV infection; the length of the IGL is dependent on the sub-group being expressed and so the preference for the same sub-group exists before and after FMDV infection. The amino acid variation between IgL transcripts is also limited; significantly less percentage deviation from the amino acid consensus is observed in the IgL CDR compared to IgH CDR regions. More variability is observed in the framework regions of the IgL compared to the IgH in African buffalo and shown previously in cattle (Grant, 2013; 230), although these differences were not significant. The IgL in both the African buffalo *de novo* assembled genes and the cattle ARS-UCDv0.1 contain more putatively functional *IGLV* gene segments, 20 and 32 respectively, than the IgH locus of 13 and 11. This is the likely reason for the higher variability observed in the framework region of the IgL transcripts although previous reports have found evidence for gene conversion in cattle light chains (Lucier et al., 1998; 123, Parng et al., 1996; 124). Overall, this limited variation in length and amino acid composition of the IgL transcripts results in relatively few clusters with much greater frequency abundance. The largest 200 clusters account for 77-81% of their entire IgL repertoire, with the largest cluster containing 6.2% of the IgL transcripts at day 0 and 24.1% at day 14. This suggests a large degree of transcript restriction in African buffalo and cattle light chains which does not detectably change upon antigen stimulation.

The restriction to the IgL transcript variability may be necessary if the light chains are providing a mainly structural role. Interactions between the IGL and IGH chain contribute to the binding kinetics of a peptide and therefore the stability of the antibody structure (Chatellier et al., 1996; 204). The specific pairing of IGL chains with specific IGH defines

the antibody stability and structure. The long and ultra-long CDR3H sequences have unique configurational requirements compared to other species and so the light chain may provide a structural framework for support of these long IGH. A selection pressure exists for the use of relatively few *IGLV* gene segments in pairing with the ultra-long CDR3H in cattle (Saini et al., 2003; 209) where the IGL appear to have evolved specifically to provide a supporting platform, holding the ascending polypeptide of the CDR3H in place (Wang et al., 2013; 92). The lack of variability in response to FMDV infection or inoculation in African buffalo and cattle light chains respectively, therefore appears to support the notion that the light chains in these species are providing a structural framework for the IGH.

The ultra-long CDR3H loops were a novel diversification method in cattle that has not been observed in another species until now. The African buffalo also form long and ultra-long CDR3, similar to cattle, the bimodal distribution of 20-40 amino acids in length whilst the ultra-long CDR3 sequences range from 50-70 amino acids in length. These sequences are much longer than what is observed in humans and mice (8-16 amino acids) and provide additional diversity to the antibody repertoire through somatic hyper-mutation and the bias of cysteine diversification along the length of the CDR3. These cysteines are capable of forming various disulphide bonds which gives rise to different 3-D architectures (Wang et al., 2013; 92). The over-representation in the CDR3 of the amino acids glycine, tyrosine and cysteine is also shown in previous cattle studies (Grant, 2013; 230, Larsen and Smith, 2012; 138). Combined, the glycine, tyrosine and cysteine residues account for ~31.7% of the amino acids in the CDR3. Both Glycine and Tyrosine play an important role in an antibodies ability to bind antigen; glycine provide conformational flexibility whilst tyrosine stabilises antibody-antigen interactions (Wu et al., 2012; 236). This appears unique to cattle and African buffalo in order to increase the conformational flexibility of their ultra-long CDR3, as in humans the prevalence of glycine and cysteine decrease with an increase in CDR3 length (Zemlin et al., 2003; 237). The cysteine residues further enhance structural stability and form various antibody conformations through the generation of different disulphide bonds. The high frequency of cysteine in the long and ultra-long antibodies in cattle has been shown to enhance the antibody repertoire of cattle by forming different antibody crystal structures.

The estimated gene usage in African buffalo and cattle was difficult to determine; the putatively functional *IGHV* in the germline are highly similar sequences and high levels of variability within the consensus *IGHV* region transcripts meant assigning their germline gene segment relied on the top BLAST hit. All of the African buffalo representative sequences

from the largest 50 clusters aligned to one of the putatively functional *IGHV de novo* assembled from the genome in Chapter 3 with an average 85.6% nucleotide sequence identity. In cattle, all of the representative sequences mapped to the putatively functional *IGHV* in the ARS-UCDv0.1 assembly with an average 91.7% nucleotide identity. The estimated gene usage based on the top BLAST hit showed no significant change in the gene usage in cattle following infection with FMDV but significant changes were observed in the African buffalo with *IGHVa*-10 usage increasing significantly between day 0 and day 8. The African buffalo assembly however is incomplete and the assembly of more putatively functional gene segments is likely, the concatenation of several genes may have created SNP bias in the assembled gene segments which causes further difficulties in calculating gene segment usage at this time.

This is the first characterisation of the African buffalo antibody repertoire during an immune response to FMDV. Illumina sequencing of the IgM and IgG repertoire shows a clear change in the frequency of IGH antibody transcripts following infection that is not seen in cattle immunised with SAT1 FMDV. The unique structural characteristics of cattle antibodies are observed in African buffalo and these ultra-long antibodies appear to change in response to infection. The specificity of the high frequency transcripts is unknown but considering that the IGL has limited variation and appears to be providing a structural role to the long and ultra-long CDR3, a clear selection pressure exists for longer and more flexible CDR3 loops.

Chapter 6

Conclusions and further work

6. Overview

FMD remains a serious risk to agriculture for the foreseeable future; the disease is a continual financial threat or burden to countries and causes a food security issue in regions where it is endemic. FMDV is the most infectious veterinary disease agent known and as such spreads at a rapid rate through susceptible populations. The disease affects all of the non-avian livestock species, circulating in 77% of the global livestock population, causing large scale loss in productivity, animal health and welfare. Better vaccines need to be developed that provide long term immunity to livestock and the role of the African buffalo in the maintenance and spread of FMDV variants needs to be thoroughly investigated.

Protection against disease coincides with neutralising antibody titres. Studies in cattle suggest that the immune response to FMDV is T-independent during the early stages of infection (Juleff et al., 2009; 54) and that antibody is responsible for the opsonisation of the virus and enhancing the innate immune defences (Bradford et al., 2001; 77, Guzylack-Piriou et al., 2006; 69). The antibody repertoire in African buffalo was previously unknown, despite having shown that high levels of antibody are responsible for protection against disease in young African buffalo (Condy and Hedger, 1974; 46). Considering the two species diverged recently, 5.7-9.3 mya (Glanzmann et al., 2016; 1), their germline and therefore antibody repertoire was expected to be similar. However, the two species display markedly different disease profiles with infection of African buffalo being sub-clinical whereas cattle display 100% morbidity characterised by an acute febrile reaction with severe vesicular lesions that cause lameness and yield reductions in a herd. It was therefore hypothesised that African buffalo produce a different antibody response to cattle, with more specific and/or avid antibodies in African buffalo resulting in protection against disease. The antibody encoding germline was hence investigated and compared in cattle and African buffalo to determine the possible recombinatorial potential of the immunogenetic loci. The antibody repertoire in response to FMDV was then interrogated.

6.1.1 The recombinatorial potential of the IGH is restricted in cattle and African buffalo

The antibody loci are highly repetitive, GC-rich regions of the genome which are difficult to assemble with short reads, such as Illumina. If the short read does not contain a unique sequence, its origin cannot be precisely determined and the multiple alignments and mis-alignments of the reads lead to sequence gaps and highly fragmented assemblies. Historically, this has been the case for the cattle antibody loci in the genome. The IGH in particular, was assembled across multiple chromosomes and any attempted characterisation was incomplete (Niku et al., 2012; 167). The long read PacBio genome provided a means for characterising the complete IGH, isolated on two scaffolds. Around the same time, Ma et al (2016) published the complete IGH sequence assembled with Sanger reads of BAC clones (Ma et al., 2016; 98).

The IGH in both the ARS-UCDv0.1 and Ma et al's contiguous sequence reveals an unusual configuration to the cattle IGH loci; large duplications within the region has resulted in a configuration that deviates from the expected *IGHV-IGHD-IGHJ-IGHC* and instead they contain 5'-*IGHVn-IGHDn-IGHJn-IGHM-IGHD-IGHV-IGHDn-IGHD-IGHV-IGHDn-IGHD-IGHV-IGHDn-IGHJn-IGHM-IGHD-IGHG3-IGHG1-IGHG2-IGHE-IGHA*-3'. This unusual IGH configuration has not been seen in other mammalian species and the existence of more than one functional *IGHM* has only been found so far in crocodiles (Cheng et al., 2013; 238). In cattle, both the *IGHM* are functionally expressed through independent V(D)J recombination, as confirmed by RNA-seq data of multiple Holstein animals. Ma et al also demonstrated that the 5' *IGHM* can also be expressed through *IGHM1-IGHM2* switching (Ma et al., 2016; 98). Despite these large internal duplications, the recombinatorial potential of the cattle IGH is limited. Cattle possess 48 *IGHV* genes, of which only ten are putatively functional; compared to humans and mice who possess 50 and 92 functional *IGHV* respectively (de Bono et al., 2004; 194, Tomlinson et al., 1995; 195). This restricted recombinatorial antibody repertoire must be compensated in post-transcriptional modifications to allow cattle to cope with the vast array of antigen they encounter throughout their lives.

A characterisation of the African buffalo antibody loci had not previously been attempted and a complete genome has only recently become available (Glanzmann et al., 2016; 1). This African buffalo genome was assembled using short reads from Illumina sequencing and so

the mapping and assembly of these short genomic reads to the antibody loci was challenging. By targeted assembly of individual gene segments, an insight into the African buffalo IGH is provided. There is currently no evidence for the existence of the *IGHV-IGHJ-IGHDn-IGHM-IGHD* duplication in the African buffalo that we see in cattle, or an ultra-long *IGHD* gene segment; although the assembly is largely incomplete with the mapping bias of the short reads likely concealing gene segments. A total of 57 *IGHV* were assembled, of which 13 are putatively functional, confirming the restricted recombinatorial diversity is also seen in African buffalo and is not the cause of their protection against FMD.

The large internal duplications in the cattle led to the expansion of the *IGHD* gene segments, where a total of 16 *IGHD* exist in four separate regions in the cattle IGH; the ultra-long *IGHD* gene segment was formed in the duplication. All the cattle and African buffalo *IGHD* are greater than 31 bp in length, which is much longer than in humans and mice where the average length is 23 and 17 bp, respectively (Lee et al., 2006; 239). These long and ultralong *IGHD* have been shown to provide a compensatory mechanism in cattle for their limited recombinatorial diversity in that they form long and ultra-long CDR3H loops (Saini et al., 1999; 232). The vast length variability and levels of somatic hyper-mutation along the lengths of the CDR3H markedly increases the sequence and structural diversity of cattle IGH chains and, as we demonstrate, the African buffalo IGH repertoire also. This provides a compensatory mechanism for their limited germline recombinatorial potential.

6.1.2 Cattle and buffalo antibodies are structurally unique

The long and ultra-long CDR3H, shown in previous studies was considered a unique diversification mechanism to cattle. African buffalo also produce these long and ultra-long CDR3H loops which range in length from 5- 70 amino acids. High levels of amino acid variability occurs along the length of these CDR3H, introduced by somatic hyper-mutation of bases and both cattle and African buffalo possess highly diverse CDR3H repertoires, as demonstrated by the low frequency abundance of their largest 200 clusters.

Although most ultra-long CDR3H are generated in cattle by the ultra-long *IGHD* gene segment, which is phylogenetically similar to all of the cattle consensus ultra-long transcripts. N and P additions by the enzyme terminal deoxynucleotidyl transferase (TdT) also

contributes to the longer CDR lengths in cattle (Liljavirta et al., 2014; 101) and a novel diversification mechanisms via conserved short nucleotide sequence (CSNS) insertions, typically at the *IGHV-IGHD* junction, may also contribute to the long and ultra-long CDR3H sequences (Koti et al., 2010; 171). These mechanisms could contribute to the longer lengths of the cattle and African buffalo CDR3H and are the possible mechanisms in African buffalo in the absence of a known ultra-long *IGHD* gene segment, although a more complete IGH germline characterisation is required to resolve this.

The role of the ultralong CDR3H during infection is unknown. The CDR3H confers most of the antigen-interaction capabilities of the whole antibody antigen-binding domain in transgenic mice (Xu and Davis, 2000; 217) and is the only point of antigen-interaction in cattle antibodies (Wang et al., 2013; 92, Grant, 2013; 230). In the African buffalo we observe changes in the frequency abundance of specific ultra-long CDR3H sequences following FMDV infection; sequence expansion occurs in the IgM repertoire between day 0 and day 8 and in the IgG repertoire between day 8 and day 14 following infection. The low frequency abundance of the ultra-long transcripts, as they account for ~5% of the total African buffalo repertoire and here only 0.5% of the cattle repertoire means that the depth of sequencing with Illumina achieved may have precluded ultra-long sequences from the repertoire. The African buffalo blood samples contained an estimated 5×10^7 PBMC. These millions of cells were sequenced at ~1.5 million reads which would not have sequenced every B cell clone in the repertoire but would have included all of the high frequency clones in the circulation. The increase in frequency abundance of the ultra-long CDR3H in African buffalo therefore, suggests a role in the response to FMDV.

6.1.3 The light chain appears to provide a structural role to cattle and African buffalo antibodies

The IGL in cattle and African buffalo is shown to be highly limited, with controlled gene usage and restricted amino acid variability in the transcripts. A dominance of IGL in 95-98% of the light chain transcripts was shown by qPCR in cattle and African buffalo which did not alter following inoculation or infection with FMDV respectively. IGL transcripts in both cattle and African buffalo have a near uniform length distribution (Chapter 5, section 5.3.3,(Grant, 2013; 230)); the CDR regions in the IGL are homogenous in length which limits

their variability. The amino acid variation between transcripts is largely confined to within and around the predicted CDR regions and the levels of variation are significantly lower than what is observed in the IGH. Higher levels of variability are observed in the framework regions of the IGL compared to the IGH which is attributed to the greater number and variability of functional *IGLV* gene segments in the IGL loci compared to the IGH loci.

The cattle ARS-UCDv0.1 assembly provides the most complete characterisation of the IGL locus to date, with a total of 97 *IGLV* of which 32 are putatively functional. The African buffalo IGL was more challenging to assemble due to the short genome reads but targeted *de novo* assembly of individual gene segments revealed a total of 58 *IGLV* of which 20 were putatively functional. Greater diversity is therefore observed in the IGL locus of both species compared to IGH, although this is still much lower than what is observed in humans and mice (33 and 105 putatively functional *IGLV* respectively) (Solomon and Weiss, 1995; 202, Gerdes and Wabl, 2002; 203). However, expression analysis of the *IGLV* gene segments in cattle and African buffalo with RNA-seq data, and in previous studies in cattle (Grant et al; unpublished), shows the limited usage of relatively few *IGLV*. The frequency abundance of the largest 200 clusters in African buffalo accounts for 77-81% of the entire IGL repertoire in African buffalo. The greater diversity of *IGLV* gene segments and therefore greater recombinatorial potential in the cattle and African buffalo IGL germline compared to the IGH seems contradictory to the lack of diversity observed in the IGL transcripts and may be compensatory for the reduced level of post-transcriptional modifications to the light chain.

The IGL transcripts in cattle and African buffalo have restricted variation; lengths of the IGL are near uniform in their distribution and limited amino acid variability is observed in the IGL following inoculation or infection with SAT 1 FMDV. The long and ultralong CDR3H loops in cattle and African buffalo antibodies have unique configurational requirement to other species and so the light chain is likely providing the structural framework for the support of these chains, as has been demonstrated in crystal structures of ultralong CDR3H loops in cattle antibodies (Wang et al., 2013; 92) and in the restricted pairing of cattle VL and VH antibody chains (Saini et al., 2003; 209). The relative frequency abundance of the dominant IGL transcripts does not change following SAT1 infection in African buffalo, coupled with the limited variation in their lengths and amino acid composition suggests a selection pressure exists for the use of relatively few *IGLV* whose role is to pair specifically with the long and ultralong CDR3H, providing a structural framework that is not severely mutated.

6.1.4 African buffalo display a dramatically different antibody response to FMDV than cattle

The antibody response to FMDV is markedly different between African buffalo and cattle. Infection of African buffalo with SAT 1 FMDV results in the dramatic expansion of particular sequences between day 0 and day 8 and between day 8 and day 14 in both the IgM and IgG repertoire of each animal. Studies in mice have demonstrated that highly abundant IGH transcripts identified post-immunisation with FMDV are highly likely to be antigen specific (Reddy et al., 2010; 218). Thus, it is likely that these dominant sequences in response to FMDV are specific, although this cannot be ascertained without expression and affinity testing of these transcripts. In cattle, this dramatic expansion of particular sequences is not observed following inoculation with SAT1 FMDV. The antibody sequences were amplified using the same primers and analysed in the same pipeline and so differences observed in the antibody repertoire of cattle and African buffalo is attributed to their different immune response to FMDV. The principal incongruity between the two models is that African buffalo were infected with live virus whilst cattle were inoculated with highly purified SAT1 antigen. The differences between the two species may therefore be attributed to the lack of replication within the cattle and therefore less activation by the immune system.

6.2 Future work

6.2.2 Investigating the specificity of the African buffalo antibodies

Following the central dogma of immunological response, infection with antigen leads to the induction of antigen-specific antibodies that are isotype-switched and of high affinity. Naïve African buffalo animals infected with SAT1 FMDV sero-converted, as shown by ELISA, by day 8 which corresponds with the expansion of individual IgM and IgG transcripts that are likely FMDV specific. The specificity of these highly abundant sequences in African buffalo post-infection is unknown and could be resolved through their expression and affinity testing with ELISA. Considering the African buffalo are asymptomatic to FMD, it is possible they produce more specific and/or avid antibodies to FMDV.

The discrepancy between the African buffalo and cattle antibody sequence reads is that the African buffalo were infected with live SAT1 FMDV that underwent replication inside the cells, whereas the cattle were inoculated with highly purified SAT1 antigen. Whether cattle produce a more specific and expansion of sequences following live infection, similar to African buffalo, is unknown and requires investigation. The initial stages of infection with FMDV is a T-cell independent response in cattle (Juleff et al., 2009; 54) and so their role in B cell activation and inducing the class switch response is performed by dendritic cells (Bautista et al., 2005; 60). Activation of dendritic cells relies on the replication of FMDV at the site of infection (Batista and Harwood, 2009; 106) which stimulates the production of type I IFN and production of BAFF and APRIL which drive B cell antibody production and CSR (Bergamin et al., 2007; 61). The antibody response to inoculating with antigen would therefore be less pronounced which could account for the different responses seen in African buffalo and cattle here. This has important implications when considering vaccine design in order to stimulate DCs.

African buffalo are the only long-term maintenance hosts of FMDV and animals are usually co-infected with multiple serotypes (Vosloo et al., 2002; 240). Cattle that have recovered from FMDV infection are, in general, protected against homologous but not heterologous re-challenge (Doel, 2005; 79). African buffalo are therefore potentially generating cross-reactive neutralising antibodies between serotypes that protects against further challenges. Extensive

affinity maturation in the HIV-1 and influenza fields suggests these cross-reactive antibodies can develop naturally (Wrammert et al., 2011; 241, Zhang et al., 2012; 242). A site of viral persistence of FMDV in African buffalo is the germinal centres of lymphoid tissue (Juleff et al., 2012; 39) which could possibly result in extensive affinity maturation that develops these cross-reactive antibodies. An investigation in the long-term antibody response to viral challenge in African buffalo and cattle would reveal if the development of the antibody repertoire remains markedly different in the development of the memory B cell and long lasting, high affinity response.

6.2.3 Vaccine design

Cattle infected with live-FMDV challenge remain protected from further re-challenge for many years (Cunliffe, 1964; 80), suggesting cattle are able to produce a long-lasting antibody response that protects the animal. Vaccination with inactivated FMDV however is unable to induce long-duration of immunity and as we have demonstrated, potentially a less specific or pronounced antibody response than infection. The current vaccinations against FMDV provide only short-term immunity and protect against specific serotypes. The rapid induction of an antibody response occurs at the site of viral replication (Pega et al., 2013; 83), potentially due to stimulation of activated DCs. Thus, it is hypothesised that the sustained antibody response seen in cattle infected with FMDV is the result of continual stimulation of naïve B cells at the site of antigen persistence. This suggests a more efficient vaccine programme would require multiple booster immunisations to maintain the protective antibody response.

Infection of naïve African buffalo is sub-clinical as a rapid induction of their antibody response results in clearance of the virus from the blood. The cause of their greater disease response is likely due to their production of more specific and/or avid antibodies. A greater understanding of how African buffalo antibodies interact with FMDV can begin to inform vaccine design or highlight potential therapeutic strategies such as the passive transfer of African buffalo serum into cattle. The assembly of the African buffalo genome with long read sequencing technology would allow an accurate characterisation of their antibody loci and a better understanding of the underlying genetics of their immune system. Cattle and African buffalo so far appear to have similarly restricted recombinatorial potential in their germline

and so the development of their different antibody responses to FMDV may not be due to substantial differences in their germline but in the post-transcriptional modifications to their primary antibody repertoire.

Development and implementation of improved vaccines requires major investment and a thorough understanding of the immune response. This PhD project sheds light on possible sources of protection in African buffalo which may be useful for designing vaccines and therapeutic strategies or breeding programmes for livestock. If this investment is not made, FMDV will likely continue to re-emerge, causing devastating outbreaks with serious social and political consequences.

Appendix Table 1: Primer sequences used for amplification of cattle and/or African buffalo antibody genes and transcripts.

Primer Name	Primer sequence	Primer binding site	Primer use	Chapter
UMD_HV1_F	CCCTCCTCTYTGTGCTSTCA	IGHV leader	TPI4222 BAC library screening	2
UMD_HV1_R	CACWCYGMBDTCCCCTCACTG	FR3	TPI4222 BAC library screening	2
UMD_HV2_F	TCTCCTCTRCCTGGTGRC	IGHV leader	TPI4222 BAC library screening	2
UMD_HV2_R	CACWCYGMBDTCCCCTCACTG	FR3	TPI4222 BAC library screening	2
UMD_HV3_F	GGTTTCTGACACTGAGAGCATC	IGHV leader	TPI4222 BAC library screening	2
UMD_HV3_R	CCCTCAGGATGKGGGTTTTC	FR3	TPI4222 BAC library screening	2
UMD_DAll_F	GTTTCTGATGCCRGCTGTG	IGHV-IGHD intron	TPI4222 BAC library screening	2
UMD_D1_R	ACCATAACCACAACCATAACCA	IGHD- IGHJ Intron	TPI4222 BAC library screening	2
UMD_D2_R	ACCACAACCATAACCATAACCAC	IGHD- IGHJ Intron	TPI4222 BAC library screening	2
UMD_D3_R	GACTCTTCCTCAGGCTGTTG	IGHD- IGHJ Intron	TPI4222 BAC library screening	2
UMD_D4_R	AACACCTAACCATAACCACC	IGHD- IGHJ Intron	TPI4222 BAC library screening	2
UMD_D5_R	CGTCACTGTGGTAGCAACAC	IGHD- IGHJ Intron	TPI4222 BAC library screening	2
UMD_D6_R	ACCACAACCATAACCATAACCA	IGHD- IGHJ Intron	TPI4222 BAC library screening	2
UMD_D7_R	GACTCTTCCTCAGGCTGTTG	IGHD- IGHJ Intron	TPI4222 BAC library screening	2
UMD_D8_R	GACTCTTCCTCAGGCTGTTG	IGHD- IGHJ Intron	TPI4222 BAC library screening	2
UMD_HJ1_F	AGCCCACTGTGACTATGCT	IGHJ1 exon	TPI4222 BAC library screening	2
UMD_HJ1_R	GCATTGCTGAGGGACACA	IGHJ2 exon	TPI4222 BAC library screening	2
UMD_HJ2_F	CTGTGTCCCTCAGCAATGC	IGHJ3 exon	TPI4222 BAC library screening	2
UMD_HJ2_R	GAGGAGAGAGGGCTGTTGAG	IGHJ4 exon	TPI4222 BAC library screening	2
UMD_HJ3_F	GTCCCAGCATCATTGTCACC	IGHJ5 exon	TPI4222 BAC library screening	2
UMD_HJ3_R	CCGGACAGTGATGCTCAGAA	IGHJ6 exon	TPI4222 BAC library screening	2
ARS-UCDv0.1_HJloci_F	AGACAGACTTRCAGCTCCYGGG	Upstream intron	Allelic varartion in *IGHJ*	3
ARS-UCDv0.1_HJloci_R	CCTCACCTAGACAATTCTCTTCCTRCCC	Downstream intron	Allelic varartion in *IGHJ*	3
SyBr_SDHA_F	GCTCTCCTACGTTGACATCA	SDHA exon	Light chain expression analysis	4
SyBr_SDHA_R	AAGCCTCAGTCTTCCTCAGTA	SDHA exon	Light chain expression analysis	4
SyBr_PPIA_F	GCATCTTGTCCATGGCAAAT	PPIA exon	Light chain expression analysis	4
SyBr_PPIA_R	TTCATGCCCTCTTTCACCTT	PPIA exon	Light chain expression analysis	4
SyBr_beta-actin_F	ACCGTGAGAAGATGACCCAG	β-actin exon	Light chain expression analysis	4
SyBr_beta-actin_R	AGGAAGGAAGGCTGGAAGAG	β-actin exon	Light chain expression analysis	4
SyBr_IGLC_F	CTCCAAACWGAGCAACAGCA	*IGLC* exon	Light chain expression analysis	4
SyBr_IGLC_R	CTTCACTGTCTTCKTCACGG	*IGLC* exon	Light chain expression analysis	4
SyBr_IGKC_F	TGTCGTGTGCTTGGTGAATG	*IGKC* exon	Light chain expression analysis	4
SyBr_IGKC_R	TTCTTGCTGTCCTGGTCTGT	*IGKC* exon	Light chain expression analysis	4
Bta_IGHV_GSP1	YGTGGRCYCTCCTCTTTGTGC	5' IgH leader sequence	Illumina sequencing of the antibody repertoire	5
Bta_IGG_GSP1	CCACCACAGCCCCGAAAGTCTACCC	*IGHG* exon 1	Illumina sequencing of the antibody repertoire	5
Bta_IGM/D_GSP1	CGAGCTCACGCAGGACACCA	*IGHM* exon 1	Illumina sequencing of the antibody repertoire	5
Bta_IGL GSP1	CGAGGGTGSGGACTTGGGCTGAC	*IGLC* exon	Illumina sequencing of the antibody repertoire	5

Appendix Table 2: Cattle *IGHV* gene segments in the ARS-UCDv0.1 PacBio assembly

Gene Segment	Functionality	Li Ma name	nt identity	Li Ma funct	UMD3.1 Name	UMD3.1 Fun	Orient	Scaffold	Start of sequen	Octamer	Octamer-ATG (bp)	J-PART1 (exon 1, bp)	Intron (nt)	Splice Sites	V Exon Length	Heptamer	Spacer (bp)	Nonamer	Note (change from position to FR)
IGHV(I)-1		HV3-2	99%		V357		+	217	289157	ATTTCAAG	125		46	82 GT/AG	302	CACAGTG	25	AGAAACCCG	Start codon: ATG->AGG, Multiple frameshifts
IGHV(1)-2		HV3-3	97%		V358		+	217	236015	ATTTCAAG	125		46	82 GT/AG	300	CACAGTG	25	AGAAACCCT	Start codon: ATG->AGG, Multiple frameshifts
IGHV(I)-3		HV3-4	98%		V351		+	217	141365	ATTTATAC	76		47	83 GT/AG	291	CACAGTG	24	CAGAAACCC	Start codon: ATG->AGG, Multiple frameshifts
IGHV(I)-4		HV2-1	99%		V252		+	217	135281	CTTTGAGG	101		46	82 GT/AG	287	CACAGTG	21	AGACACCAG	Start codon: ATG->ACA, Multiple frameshifts
IGHV4-5		HV2-2	99%		V251		+	217	128282	ATTTGCAT	107		46	83 GG/GG	293	CACAGTG	21	AGCCACAAG	RSS nonamer
IGHV(II)-6		HV1-1	97%	F	V151	F	+	217	126513					/AG	292	CACAGTG	23	ACAAAAACC	
IGHV(II)-7		HV2-6	99%		V253		-	217	9093	ATTTAAGG	105		88	/AG	155	CACAGTG	21	AGAACCAGC	RSS nonamer
IGHV(II)-8		HV1-4	99%	F	V1S10	F	-	217	10849	ATTTGCAT	107		46	82 GT/AG	292	CACAGTG	23	ACAAAAACC	Frameshift but F
IGHV(I)-9		HV3-9	100%		V352		-	217	16133	ATTTCAAA	125		47	83 GT/AG	291	CACTGTG	47	AGAAAAATG	Non-optimal heptamer
IGHV(II)-10		HV2-11	100%		V257		-	217	22641	ATTTACAT	105		12		240	CACAGTG	23	ACACCAGCC	RSS nonamer
IGHV4-11	F	HV1-9	100%	F	V156	F	-	217	24331	ATTTGCAT	107		46	82 GT/AG	293	CACAGTG	23	ACAAAAACC	
IGHV(II)-12		HV2-12	99%		V256		-	217	34625	ATTTGCAT	107		46	83 GG/GG	292	CACAGTG	23	ACACCACAC	
IGHV4-13		HV1-10	100%		V155		-	217	36418	ATTTGCAT	107		46	79 GT/CA	293	CACAGTG	23	ACAAAAACC	Frameshift but F
IGHV4-14	F	HV1-6	100%	F	V1S10	F	-	217	46668	ATTTGCAT	107		46	82 GT/AG	293	CACAGTG	23	ACAAAAACC	
IGHV1(I)-15		HV3-9	100%				-	217	52016	ATTTCAAG	125		47	83 GT/AG	291	CACACCG	25	AGAAAACCT	
IGHV(II)-16		HV2-8	98%		V2S11/12		-	217	58533	CTTTGAGG	102		46	83 GT/AG	302	CACAGTG	23	ACACCAGCC	RSS nonamer
IGHV(II)-17		HV2-9	99%		V255		-	217	65147	TTTTAAGG	103		46	82 GG/CC	291	CACAGTG	21	AGCCACAAG	
IGHV2-18		HV1-7	98%		V154	F	-	217	66907	ATTTGCAT	105		46	79 GT/AG	292	CACAGTG	23	ACAAAAACC	Frameshift but F
IGHV(II)-19		HV2-10	100%				-	217	76092	ATTTAAGG	98				240	CACAGTG	23	ACCACGCCC	RSS nonamer
IGHV(II)-20		HV1-8	99%	F	V1S2	F	-	217	77761	ATTTGCAT	107		46	82 GT/AG	287	CACAGTG	23	ACAAAAACC	Frameshift but F
IGHV(I)-21		HV3-5	99%		V3S2		-	217	83049	ATTTATAC	76		47	83 GT/AG	289	CACTGTG	47	AGAAAAATG	
IGHV(II)-22		HV2-3	99%		V2S3		-	217	89466	ATTTAAGG	102		46	92 GC/AG	164	CACAGTG	23	ACACCAGCC	RSS nonamer
IGHV4-23	F	HV1-2	99%	F	V1S10	F	-	217	91229	ATTTGCAT	107		46	82 GT/AG	293	CACAGTG	23	ACAAAAACC	
IGHV(I)-24		HV3-6	100%		V356		-	217	96617				46	83 GT/AG	291	CACAGTG	25	AGAAACACT	
IGHV(II)-25		HV2-4	98%				-	217	103862	GTTTCCAT	103		46	76 GT/AG	292	CATGTG	23	AGGGACTCT	RSS nonamer
IGHV4-26		HV2-5	99%		V254		-	217	110770	ATTTGCAT	113		46	84 GG/GG	296	CACAGTG	23	ACACAGGCC	
IGHV(II)-27	F	HV1-3	99%	F	V153	F	-	217	112830	ATTTGCAT	107		46	82 GT/AG	293	CACAGTG	23	ACAAAAACC	
IGHV(I)-28		HV3-7	100%		V3S3		-	217	118205	ATTTCAAG	125		47	82 GT/AG	291	CACAGTG	25	AGAAACCCT	
IGHV(II)-29		HV2-7	98%				-	217	124476	ATTTAAGG	102		46	81 GT/AG	154	CACAGTG	23	ACCAGCCAC	RSS nonamer
IGHV(II)-30		HV1-5	98%	F			-	217	126233					/AG	303	CACAGTG	23	ACAAAAACC	
IGHV(II)-31		HV1-12	99%	F	V154	F	+	2469	114376	ATTTGCAT	107		46	81 TG/AG	291	CACAGTG	23	ACAAAAACC	Frameshift but F
IGHV(II)-32		HV2-15	99%		V258		+	2469	105266	ATTTGCAT	107		46	73 TG/AG	225	CATGACC	17	ACACACAGT	Non-optimal heptamer and nonamer
IGHV(II)-33	F	HV1-13	98%	F	V157	F	+	2469	103512	ATTTGCAC	107		46	82 GT/AG	287	CACAGTG	23	ACAAAAACC	
IGHV(I)-34		HV3-11	100%		V354		+	2469	98172	TTTTCAGC	120		47	82 GT/AG	286	CACAGTG	66	AGAAAAATG	
IGHV(I)-35		HV3-13	100%		V356		+	2469	83271	TTTTCAGC	119		45	82 GT/AG	291	CACAGTG	25	AGAAACCCG	
IGHV(II)-36		HV2-17	99%		V2S10		+	2469	73383	GTTTCCAT	104		46	83 GT/AG	277	CACAGTG	23	ACACCAGCC	RSS nonamer
IGHV(II)-37		HV1-15	99%		V159		+	2469	70149	ATTTCTGT	102		46	82 GG/AG	288	CACAGTG	19	CCAAAAACC	
IGHV(I)-38		HV3-14	99%		V355		+	2469	50638	TTTTCAGC	111		47	83 GT/AG	283	CACAGTG	24	CAGAAACCC	
IGHV4-39			missing		V259	ORF	+	2469	44415	ATTTGCAT	107		46	87 AT/TG	296	CACAGTG	23	ACACCAGCC	RSS nonamer
IGHV(II)-40			missing		V158		+	2469	42631	ATTTGCAT	107		46	83 GT/AG	311	CACAGTG	23	ACAAAAACC	Frameshift but F

Appendix Table 3: African buffalo *IGHV* gene segments assembled from reads mapped to the ARS-UCDv0.1individual gene segments

Buffalo Sub-group	Number	Clan	Functio	Octamer	Octamer-ATG (bp)	L-PART1 (exon 1, bp)	Intron (bp)	V-EXON (exon 2, bp)	Splice Sites	Heptamer	Spacer (bp)	Nonamer	Note
a	1	2	F					305	/AG	CACAGTG			
a	2	2	P	ATTTAAGG	60	46	82	305	GT/AG	CACAGTG	23	ACAAAAACC	Frameshift
a	3	HV4	P	ATTTGCAT	107	46	82	304	GT/AG	CACAGTG	23	ACAAAAACC	Stop codon
a	4	HV4	F	ATTTGCAT	107	46	82	221	GT/AG				
a	5	HV4	F	ATTTGCAT	107	46	82	304	GT/AG	CACAGTG	23	ACAAAAACC	
a	6	HV4	F							CACAGTG	23	ACAAAAACC	
a	7	HV4	F	ATTTGCAT	107	46	82	304	GT/AG	CACAGTG	23	ACAAAAACC	
a	8	HV4	F	GTTTCTAC	81	46	82	304	GT/AG	CACAGTG	23	ACAAAAACC	
a	9	HV4	F	ATTTGCAT	107	46	82	214	GT/AG				
a	10	HV4	F	ATTTGCAT	107	46	82	304	GT/AG	CACAGTG	23	ACAAAAACC	
a	11	HV2	F					304	/AG	CACAGTG	23	ACAAAAACC	
a	12	HV4	F			46	82	246	GT/AG				Frameshift
a	13	HV4	F	ATTTGCAT	107	46	82	304	GT/AG				
a	14	HV4	F										
a	15	2	P			46	82	304	GT/AG	CACAGTG	22	ACAAAACC	Stop codon
a	16	2	F	ATTTGCAT	107	46	82	282	GT/AG				Break in exon (contig)
a	17	2	F			46	82	310	AT/AG	CACAGTG	23	ACAAAAACC	
a	18	2	P	ATTTAAGT	60	46	82	324	GT/AG	CACAGTG	23	ACAAAAACC	Frameshift
b	1	HV4	P					307	/CC	CACAGTG	23	ATAGAGGCC	Frameshift
b	2	HV4	P	ATTTGCAT	107	46	85	304	GG/AG	CACAGTG			Frameshift
b	3	2	P			46	83	257	GG/GG				Frameshift
b	4	HV4	P					296		CACAGTG	23	ATACAAGCC	Frameshift
b	5	2	P	ATTTGCAT	107	46	82	304	GG/CG	CACAGTG			Stop codon
b	6	-	P	TTTTCATT	108	46	82	304	GG/CG	CACAGTG	23	ACACAAGCC	Frameshift
b	7	2	P	ATTTGCAT	107	46	83	303	GG/GG	CACAGTG	23	ACACCAGCC	Stop codon in leader, Frameshift
b	8	HV4	F	ATTTGCAT	110	46	85	307	AT/AG	CACAGTG	23	ACACCAGCC	
b	9	2	P	GTTTCAAG	106	46	81	300	CT/AG	CACAGTG	22	ACACCAGCC	Frameshift
b	10	2	P	GTTTCCAC	107	44	82	299	GT/AG	CACAGTG	23	ACACAAGCC	Frameshift
b	11	2	P			46	76	223	GT/AG				
b	12	2	P	ATTTCCAT	105	46	82	299	GT/AG	CACAGTG	23	ACACCAGCC	Frameshift
b	13	-	P			46	85	298	GT/AG	CACAGTG	23	ACACCAGCC	Multiple stop codons
b	14	2	P	GTTTCATG	106	46	81	300	GC/AG	CACAGGG	23	ACACAAGCC	Frameshift
b	15	-	P			43	85	298	GT/AG	CACAGTG	23	ACACCAGCC	Multiple stop codons
b	16	-	P	GTTTCCAT	104	46	85	299	GT/AG	CACATTG	24	ACCAGCCAC	Frameshift
b	17	2	P	TTTTCAAG	107	45	79	301	CG/CA	CACAGGG	23	ACACAAGCC	Multiple stop codons
b	18	2	P			46	81	300	GC/AG	CACAGTG	23	ACACAAGCC	Frameshift
b	19	-	P	ATTTCCAT	104	46	84	298	GT/AG	CACAGTG	23	ACACAAGCC	Stop codon
b	20	2	P					300	/AG	CACAGGG	23	ACACAAGCC	Frameshift
b	21	2	P	TTTTCATG	106	46	81	300	GC/AG	CACAGGG	23	GCACAAGCC	Frameshift
b	22	2	P	TTTTCATG	105	46	81	229	GC/AG				Stop codon
b	23	-	P	ATTTGCAT	107	46	85	234	GT/AG				Truncated
b	24	-	P	ATTTGCAT	107	46	85	281	AT/AG	GCAGGTG	23	TCCTCTCAG	Frameshift
b	25	-	P					303	GA/AG	GCAGGTG	28	TCAGGAGGC	Frameshift
b	26	2	P			46	85	281	AT/AG	GCAGGTG			Heavily disrupted
b	27	2	P	TTTTACCC	83	46	85	281	AT/AG	GCAGGTG	24	CCTGTCAGG	Frameshift
c	1	1	P			46	83	301	GT/AG	CACAGGG	25	AGAAACCCT	Frameshift
c	2	1	P	ATTTAACC	79	46	83	308	GT/AG				Frameshift
c	3	1	P	ATTTAACC	79	46	83	301	GT/AG	CACAGGG			Stop codon
c	4	1	P	ATTTAACC	79	46	81	303	GT/AC	CACAGGG	23	TCAGAAACC	Frameshift
c	5	1	P	ATTTAACC	79	46	83	302	GT/CG	CACAGTG	23	TCAGAAACC	Frameshift
c	6	1	P	ATTTAACC	79	46	83	302	GT/AG	CACAGTG	25	AGAAACCCT	Frameshift
c	7	-	P	CTTTACAC	128	46	83	301	GT/AG	CACAGGG	25	AGAAACCCT	Frameshift
c	8	1	P	CTTTACAC	128	46	82	302	GT/CG	CACAGGG	25	AGAAAACCT	Frameshift
c	9	-	P	ATTTAACC	79	46	82	257	GT/AG	CAGGGTG	23	AGAAAAACC	Break in exon (contig)
c	10	1	P	ATTTAACC	79	46	83	301	GT/CG	CACAGTG	23	TCAGAAACC	Multiple stop codons
c	11	1	P			46	82	301	GT/CG	CACAGTG	23	TCAGAAACC	Frameshift
c	12	1	P	ATTTAACC	79	46	81	303	GT/AC	CACAGTG	23	TCAGAAACC	Frameshift

Appendix Table 4: Cattle *IGLV* gene segments in the ARS-UCDv0.1 PacBio assembly

Gene Segment	Funct	Orientation	Scaffold	Start of Sequenc	Octam	Octamer-ATG (bp)	L-PART1 (exon 1, bp	Intron (bp)	Splice Site	V-EXON (exon 2, b	Heptame	Spacer (bp)	Noname	Note
IGLV3-1	P	+	1160	359318	ATTTGTAT	83	45	373	GT/AG	301	CACAGTG	22	ATGCAAACC	Frameshifts in L-PART1 and framework 1
IGLV3-2	F	+	1160	355454	ATTTGCAT	106	46	138	GT/AG	293	CACAGTG	23	ACACAAACC	
IGLV3-3	F	+	1160	350724	ATTTGCAT	107	46	161	GT/AG	302	CACAGTG	23	ACACAAACC	
IGLV3-4	F	+	1160	338634	ATTTGCAT	107	46	154	GT/AG	299	CACAGTG	23	ACACAAACC	
IGLV3-5	F	+	1160	332246	ATTTGCAT	106	46	139	GT/AG	299	CACAGTG	23	ACACAAACC	
IGLV2-6	F	+	1160	321449	ATTTGCAT	96	46	115	GT/AG	308	CACAGTG	23	ACCAAAACC	
IGLV2-7	ORF	+	1160	305355	ATTTGCAT	96	46	117	GT/AG	308	CGCAGTG	23	ACCGAAACC	C104W, non-optimal heptamer
IGLV2-8	P	+	1160	302104	ATTTGCAT	87	46	115	GT/AG	306	CACAGTG	23	ACCACAACG	Multiple frameshifts and stop codons in V-EXON
IGLV2-9	F	+	1160	299049	ATTTGCAT	95	46	116	GT/AG	308	CACAGTG	23	ACCAAAACC	
IGLV5-10	P	+	1160	215260	ATTTGCAT	88	46	116	GT/AG	319	CACGGCG	23	ACCTAAATC	Multiple frameshifts and stop codons in V-EXON, non-optimal heptamer
IGLV1-11	F	+	1160	212456	ATTTGCAT	106	46	109	GT/AG	307	CACAGTG	23	ACAAAAACC	
IGLV5-12	P	+	1160	210097	ATTTGCAT	87	45	114	GT/AG	315	CACTGCA	21	ACGGCGGCC	Start codon: ATG->GTG, multiple frameshifts, non-optimal heptamer
IGLV1-13	P	+	1160	205892	ATTTGCAT	106	46	109	GT/AG	302	CACAGGG	23	ACAAAAACC	Frameshift in FR3, non-optimal heptamer
IGLV(III)-14	P	+	1160	198874	-	-	-	-	-	272	CACAGGG	23	ACGAGAGGC	Truncated in framework 1, frameshifts, non-optimal heptamer
IGLV1-15	P	+	1160	194308	ATTTGCAT	104	45	106	GT/CA	310	CACAGTG	23	ACAAAAACC	Incorrect splice site
IGLV1-16	P	+	1160	190739	ATTTGCAT	107	46	109	AT/AG	301	TACAGTG	22	ACAAAAACC	Multiple frameshifts, non-optimal heptamer
IGLV(I)-17	P	+	1160	181361	ATTTGCAT	103	46	109	GT/AG	303	CACAGTG	22	ACAAAAAAC	Start codon: ATG->ACG, Multiple frameshifts
IGLV8-18	P	+	1160	178537	ATCTGCAT	102	46	98	GT/AG	303	CACAGTG	22	ACCAAAATC	Frameshift in FR2, non-optimal octamer
IGLV(IV)-19	P	+	1160	173535	-	-	-	-	-	269	CACAGCG	23	ACGAGAGGC	Truncated in framework 1, multiple frameshifts, non-optimal heptamer
IGLV1-20	F	+	1160	170645	ATTTGCAT	107	46	109	GT/AG	307	CACAGTG	23	ACAAAAACC	
IGLV(I)-21	P	+	1160	167339	CTTTGCAT	106	46	109	GT/AG	298	CACAGTG	21	ACAAAAACC	Multiple frameshifts, non-optimal octamer
IGLV(I)-22	F	+	1160	161646	ATTTGCAT	99	46	129	GT/AG	304	CGCGGTG	22	ACAAAAACC	Non-optimal heptamer
IGLV8-23	P	+	1160	154718	ATTTGCAT	102	45	99	GT/TA	310	CACAGTG	23	ACTGAACCC	Incorrect splice site
IGLV1-24	F	+	1160	148090	ATTTGCAT	106	46	109	GT/AG	306	CACAGTG	22	ACAAAAACC	
IGLV8-25	P	+	1160	144247	ATTTGCAT	103	46	99	GT/AG	303	CACAGTG	23	ACCAAAACC	Multiple frameshifts
IGLV1-26	F	+	1160	138116	ATTTGCAT	106	46	107	GT/AG	306	CACAGTG	22	ACAAAAACC	
IGLV8-27	P	+	1160	134272	ATCTGCAT	102	46	98	GT/AG	303	CGTAGTG	24	AGCAAAACC	Multiple frameshifts, non-optimal octamer and heptamer
IGLV(IV)-28	P	+	1160	128744	-	-	-	-	-	290	CACAGCA	23	ACGAGAGCC	C104H, truncated in framework 1, non-optimal heptamer
IGLV(IV)-29	P	+	1160	122129	-	-	-	-	-	260	AACAGGG	22	ATGGGAAAC	C104R, Truncated in framework 1, non-optimal heptamer
IGLV1-30	ORF	+	1160	119561	ATTTGCAT	105	46	108	GT/AG	310	CACAGTG	23	ACAAAAACC	C23S, W41Q
IGLV8-31	P	+	1160	-	-	-	-	-	-	288	CGCAGTG	23	ACAAACACC	L-PART1 deleted, truncated V-EXON, multiple frameshifts and stop codons, non-optimal heptamer
IGLV8-32	F	+	1160	111617	ATTTGCAT	102	46	99	GT/AG	310	CACAGTG	23	ATTAAAACC	
IGLV(IV)-33	P	+	1160	-	-	-	-	-	-	212	CAACAGG	23	ACGGGAAAC	L-PART1 deleted, truncated V-EXON, multiple frameshifts and stop codons, non-optimal heptamer
IGLV1-34	F	+	1160	104618	ATTTGCAT	107	46	109	GT/AG	310	CACAGTG	23	ACAAAAACC	
IGLV(I)-35	P	+	1160	101019	CTTTGCAT	107	46	109	GT/AG	302	CACAGTG	21	ACAAAAACC	C23F, stop codon and frameshift in FR3, non-optimal octamer
IGLV1-36	F	+	1160	92445	ATTTGCAT	106	46	107	GT/AG	306	CACAGTG	22	ACAAAAACC	
IGLV8-37	P	+	1160	88603	ATTTGCAT	103	46	99	GT/AG	304	CACAGTG	23	ACCAAAACC	Truncated in F1, stop codon in FR2
IGLV5-38	ORF	+	1160	82457	ATTTTCAT	88	46	116	GT/AG	296	CACAGTG	22	ACAAAACCC	C104 missing, non-optimal octamer
IGLV1-39	P	+	1160	73769	ATTTGCAT	106	46	110	GT/AA	309	CACAGTG	23	ACAAAAACC	Multiple stop codons, incorrect splice site
IGLV1-40	F	+	1160	68488	ATTTGCAT	106	46	110	GT/AG	307	CACAGTG	23	ACAAAAACC	
IGLV(I)-41	P	+	1160	65198	CTTTGCAT	107	46	109	GT/AG	302	CACAGTG	21	ACAAAAACC	Frameshift in FR3, non-optimal octamer
IGLV(III)-42	P	+	1160	60587	ATTTTCAT	88	46	118	GT/AG	293	CACAGTG	22	ACAAAACCC	C104G, multiple frameshifts, non-optimal octamer
IGLV1-43	P	+	1160	51879	ATTTTCAG	106	46	109	GT/AA	309	CACAGTG	23	ACAAAAACC	Incorrect splice site, stop codon in FR3, non-optimal octamer
IGLV1-44	F	+	1160	46621	ATTTGCAT	106	46	109	GT/AG	307	CACAGTG	23	ACAAAAACC	
IGLV(I)-45	P	+	1160	43334	CTTTGCAT	106	46	109	GT/AG	300	CACAGTG	21	ACAAAAACC	Multiple frameshifts, non-optimal octamer
IGLV8-46	P	-	1160	35413	ATTTGCAT	103	46	99	GT/AG	303	CACAGTG	23	ACCAAAACC	Stop codon in FR1, Frameshift in FR2
IGLV1-47	F	-	1160	31578	ATTTGCAT	106	46	107	GT/CA	305	CACAGTG	22	ACAAAAACC	Incorrect splice site
IGLV(I)-48	P	-	1160	23505	CTTTGCAG	107	46	109	GT/AG	301	CACAGTG	20	ACAAAAACC	Multiple frameshifts, non-optimal octamer
IGLV1-49	F	-	1160	18882	ATTTGCAT	106	46	109	GT/AG	307	CACAGTG	23	ACAAAAACC	
IGLV1-50	P	-	1160	13655	CTTTGCAT	107	46	109	GT/AG	302	CACAGTG	21	ACAAAAACC	Frameshift in FR3, non-optimal octamer
IGLV1-51	F	-	1160	10369	ATTTGCAT	106	46	109	GT/AG	307	CACAGTG	23	ACAAAAACC	
IGLV1-52	P	-	1160	5129	ATTTGCAT	106	46	110	GT/AA	308	CACAGTG	23	ACAAAAACC	Multiple frameshifts, incorrect splice site
IGLV(IV)-53	P	+	2373	32109	-	-	-	-	-	268	CACAGCG	23	ATGAGAGCC	C104H, Truncated, multiple frameshifts, non-optimal heptamer
IGLV1-54	F	+	2373	29514	ATTTGCAT	107	46	109	GT/AG	310	CACCGTG	23	ACAAAAACC	Non-optimal heptamer
IGLV(I)-55	P	+	2373	26210	CTTTGCAT	106	46	109	GT/AG	298	CACAGTG	21	ACAAAAACC	Multiple frameshifts, non-optimal octamer
IGLV(I)-56	P	+	2373	20457	ATTTGCAT	99	46	161	GT/AG	304	CGCGGTG	22	ACGAAACCC	Stop codon in L-PART1, non-optimal heptamer
IGLV8-57	F	+	2373	13225	ATTTGCAA	102	46	99	GT/AG	310	CACAGTG	23	ACTGAACCC	Non-optimal octamer and nonamer
IGLV(IV)-58	P	+	2373	8813	-	-	-	-	-	269	CACAGCG	23	ACGAGAGGC	C104H, Truncated, multiple frameshifts, non-optimal heptamer
IGLV1-59	F	+	2373	7032	ATTTGCAT	107	47	109	GT/AG	307	CACAGTG	23	ACAAAAACC	
IGLV1-60	P	+	2373	3725	CTTTGCAT	107	46	108	GT/AG	302	CACAGTG	21	ACAAAAACC	Multiple frameshifts, non-optimal octamer
IGLV(III)-61	P	+	2297	32936	-	-	-	-	-	264	CACAGCA	23	ACGAAGCCT	C104H, Truncated, multiple frameshifts, non-optimal heptamer
IGLV(IV)-62	P	+	2297	26582	-	-	-	-	-	305	CACAGAG	23	ACGAGAGCC	C104R, L-PART1 deleted, multiple frameshifts, non-optimal heptamer
IGLV1-63	ORF	+	2297	24042	ATTTGCAT	107	46	109	GT/AG	310	CACAGTG	23	ACAAAAACC	C23S, W41Q
IGLV(IV)-64	P	+	2297	20857		-	-	-	-	317	CGCAGTG	23	ACAAACACC	L-PART 1 missing, multiple frameshifts, non-optimal heptamer
IGLV8-65	F	+	2297	16069	ATTTGCAT	101	46	99	GT/AG	310	CACAGTG	23	ATTAAAACC	
IGLV(IV)-66	P	+	2297	11674	-	-	-	-	-	270	CAGAGCG	23	ACGAGAGCC	C104R, Truncated, multiple frameshifts, non-optimal heptamer
IGLV(I)-67	P	+	2297	9085	ATTTGCAT	107	46	106	GT/AG	305	CACAGTG	23	ACAAAAACC	Multiple frameshifts
IGLV(I)-68	P	+	2297	5746	CTTTGCAT	107	46	109	GT/AG	302	CACAGTG	21	ACAAAAACC	Frameshift in FR3, non-optimal octamer
IGLV1-69	P	+	2054	9016	ATTTGCAT	107	46	109	GT/AG	304	CACAGTG	23	ACAAAAACC	Frameshift in FR3
IGLV(I)-70	P	+	2054	5727	CTTTGCAT	106	45	108	GT/CA	302	CACAGTG	20	ACAAAAACC	Multiple frameshifts, incorrect splice site, non-optimal octamer
IGLV1-71	P	+	1914	7573	ATTTGCAT	106	46	110	GT/AG	308	CACAGTG	23	ACAAAAACC	Multiple frameshifts
IGLV1-72	F	+	1914	2303	ATTTGCAT	106	46	110	GT/	307	CACAGTG	23	ACAAAAACC	
IGLV(I)-73	P	-	514	1654	-	-	-	-	-	227	CAACAGG	23	ACGGGCCCA	C104R, truncated, non-optimal heptamer
IGLV1-74	P	-	514	3909	ATTTGCAT	105	45	109	GT/CA	307	CACAGTG	23	ACAAAAACC	Multiple frameshifts, incorrect splice site
IGLV1-75	F	-	514	7323	ATTTGAAT	106	46	109	GT/AG	304	CACAGTG	21	ACAAAAACC	Non-optimal octamer
IGLV5-76	P	-	514	11273	-	-	-	-	-	363	-	-	-	L-PART1 and V-EXON fused and frameshifted, deleted octamer, heptamer, and nonamer
IGLV(V)-77	P	-	514	13990	GTTTGGTT	106	49	125	GT/AG	325	CACAGTG	19	ACAAAAACC	Multiple frameshifts, non-optimal octamer
IGLV(IV)-78	P	-	514	19519	-	-	-	-	-	305	CACAGCG	23	ACGAGAGCC	Truncated and multiple frameshifts, non-optimal heptamer
IGLV1-79	F	-	514	22119	ATTTGCAT	107	46	109	GT/AG	310	CACAGTG	23	ACAAAAACC	
IGLV(I)-80	P	-	514	25716	CTTTGCAT	106	49	106	AC/AG	298	CACAGTG	19	AGACAAAAA	Multiple frameshifts
IGLV2-81	F	-	514	31406	ATTTGCAT	100	46	130	GT/AG	304	CGCGGTG	22	ACAAAAACC	Non-optimal heptamer
IGLV8-82	F	-	514	37888	ATTTGCAA	102	46	99	GT/AG	310	CACAGTG	23	ACTGAACCC	Non-optimal octamer
IGLV(IV)-83	P	-	514	42339	-	-	-	-	-	297	AACAGGG	22	ATGGGAAAC	Truncated, multiple frameshifts, non-optimal heptamer
IGLV1-84	F	-	514	44081	ATTTGCAT	107	46	109	GT/AG	307	CACAGTG	23	ACAAAAACC	
IGLV1-85	P	-	514	47398	CTTTGCAT	107	46	109	GT/AG	302	CACAGTG	21	ACAAAAACC	Multiple frameshifts, non-optimal octamer
IGLV5-86	P	-	1659	42105794	ATTTGCAT	88	45	116	GT/CA	330	CACGGAG	23	ACAGAAACC	Incorrect splice site, Multiple frameshifts, non-optimal heptamer
IGLV1-87	F	-	1659	42108787	ATTTGAAT	106	46	109	GT/AG	304	CACAGTG	21	ACAAAAACC	Non-optimal octamer
IGLV1-88	F	-	1659	42112213	ATTTGCAT	107	46	108	GT/AG	310	CACAGTG	23	ACAAAAACC	
IGLV2-89	P	-	1659	42120471	CTTTGCAT	107	46	108	GT/AG	301	CACAGTG	21	ACAAAAACC	Multiple frameshifts, non-optimal octamer
IGLV1-90	F	-	1659	42123764	ATTTGCAT	103	46	108	TG/AG	306	CACAGTG	22	ACAAAAACC	Incorrect splice site
IGLV(IV)-91	P			42081253	-	-	-	-	-	287	AACAGGG	22	ACGGGAAAC	Truncated, multiple frameshifts, non-optimal heptamer
IGLV1-92	P	+	1659	42079066	ATTTGCAT	107	46	107	TG/AG	309	CACAGTG	23	ACAAAAACC	Frameshift in FR1, non-optimal splice site
IGLV2-93	P	+	1659	42072655	ATTTGAAT	104	44	109	TG/CA	302	CACAGTG	21	ACAAAAACC	Incorrect splice site, Multiple frameshifts, non-optimal octamer
IGLV5-94	F	+	1659	42072655	ATTTGCAT	89	46	116	GT/AG	328	CACGGAG	22	ACAGAAACC	Non-optimal heptamer
IGLV1-95	F	+	1659	42069851	ATTTGCAT	107	46	108	GT/AG	310	CCCAGTG	23	ACAAAAACC	Non-optimal heptamer
IGLV5-96	P	+	1659	42065038	ATTTGCAT	88	45	128	GT/AG	311	CCAGGTG	23	AGGCACTGC	Stop codon in L-PART1, non-optimal heptamer
IGLV-97	P	+	1659	42060807						194	CACAGGG	22	ACCCTGAAT	Truncated in framework 2, non-optimal heptamer
IGLV1/OR28-1	P	-	621	23773901	-	-	-	-	-/AG	302	TACATTG	21	ACAAAAACC	Orphon gene segment on BTA28, L-PART1 deleted, multiple frameshifts, non-optimal heptamer

Appendix Table 5: African buffalo *IGLV* and *IGKV* gene segments assembled from reads mapped to the ARS-UCDv0.1individual gene segments

Buffalo sub-group	Number	Clan	Funct	Octamer	Oct-ATG (bp)	L-PART1	Intron	Splice sites	V-EXON	Heptamer	Spacer	Nonamer
a	1	1	F	-	-	46	107	GT/AG	306	CACAGTG	22	ACAAAAACC
a	2	1	F	-	-	-	-	-	276	CACAGTG	22	ACAAAAACC
a	3	1	F	-	-	46	107	GT/AG	294	-	-	-
a	4	1	F	-	-	46	108	GT/AG	308	CACAGTG	23	ACAAAAACC
a	5	1	F	ATTTGCAT	106	46	107	GT/AG	310	CACAGTG	22	ACAAAAACC
a	6	1	F	ATTTCCGG	98	46	109	GT/AG	310	CACAGTG	23	CAGAAACCT
a	7	1	F	CTTTCCTG	98	46	109	GT/AG	307	CACAGTG	23	ACAAAAACC
a	8	1	F	CTTTCCGG	98	46	109	GT/AG	307	CACAGTG	23	ACAAAAACC
a	9	1	F	-	-	-	-	/AG	306	CACAGTG	23	ACAAAAACC
a	10	1	F	ATTTGCAT	106	46	109	GT/AA	301	CACGGTG	23	ACAAAAACC
a	11	1	F	ATTTGCAT	106	46	109	GT/AA	310	CACGTCG	23	ACAAAAACC
a	12	1	F	-	-	-	-	/AG	307	CACAGTG	23	ACAAAAACC
b	1	I		ATTTGCAT	103	45	105	GT/CA	304	CACAGTG	22	ACAAAAACC
b	2	I		CTTTGTAT	107	46	108	GT/AG	303	CACAGTG	21	ACAAAAACC
b	3	I		-	-	46	109	GT/AG	253			
b	4	I						/AG	303	CACAGTG	22	ACAAAAACC
b	5	I					107	TG/CA	304	CACAGTG	22	ACAAAGACC
b	6	I		CTTTGTAT	107	46	109	GT/AG	297	CACAGTG	22	ACAAAAACC
b	7	I		ATTTGCAT	106	46	109	GT/AG	303	CACAGTG	22	ACAAAGACC
b	8	I				46	109	GT/AG	303	CACAGTG	22	ACAAAAACC
b	9	1		-	-	-	-	/AG	304	CACAGTG	22	ACAAAAACC
b	10	I		ATTTGCAT	107	46	109	AT/AG	302	GTACGTT	23	ACAAAAACC
b	11	I		CTTTGTAT	107	46	109	GT/AG	301			
c	1	1		ATTTGCAT	87	46	117	GT/AG	306	CACAGTG	23	ACCACAACG
c	2	II	F	ATTTGCAT	96	46	117	GT/AG	308	CACAGCG	23	ACCAAAACC
c	3	II		ATTTGCAT	96	46	115	GT/AG	308	CACAGCG	23	ACCAAAACC
c	4	II	F	ATTTGCAT	95	46	115	GT/AG	311	CACAGCG	23	ACCAAAACC
d	1	I	F	ATTTGCAT	99	46	130	GT/AG	304	CACGGTG	22	ACAAAAACC
d	2	I	F	ATTTGCAT	99	46	130	GT/AG	304	CACGGTG	22	ACAAAAACC
d	3	I		ATTTGCAT	99	45	129	GT/AG	304	CACGGTG	22	ACAAAAACC
d	4	I		ATTTGCAT	98	46	129	GT/AG	304	CACGGTG	22	ACAAAATCC
e	1	3		ATTTGTAT	83	45	370	GT/AG	301	CACAGCG	23	CGCAAACCC
e	2	3	F	ATTTGCAT	106	46	139	GT/AG	299	CACAGTG	23	ACACATATC
e	3	3	F	ATTTGCAT	106	46	139	GT/AG	299	CACAGTG	23	ACACACACC
e	4	3	F	ATTTGCAT	84	46	163	GT/AG	299	CACAGTG	23	ACACACACC
e	5	3	F	ATTTGCAT	107	46	168	GT/AG	302	CACAGTG	23	ACACACACC
f	1	5		ATTTGCAT	88	43	118	AG/AG	325	CCAGGTG	23	AGGCACTG
f	2	V				46	119	GT/AG	320	CACAGTG	22	ACAAAAACC
f	3	V		ATTTGCAT	88	46	119	GT/AG	311	CACAGTG	22	ACAAAAACC
f	4	V		ATTTGCAT	88	46	118	GT/AG	319	CACGGTG	23	ACCTAAATC
f	5	V						/AG	269	CACAGCG	23	TGAGAGGCT
f	6	V						/AG	303	CACAGCG	23	ATGAGAGGC
f	7	V						/AG	269	CACAGTG	23	ACAAGAGCC
f	8	V						/AG	269	CACAGTG	23	ACGAGAGCC
f	9	V						/AG	26			
f	10	V						/AG	286	CACAGTG	23	ATGAGAGGC
f	11	V						/AG	290	CACAGTG	23	ATGAGAGGC
g	1	8							250	CACAGTG	23	ACCAAAACC
g	2	III		ATTTGCAT	103	46	100	GT/CA	301	CACAGTG	23	ACCAAAACC
g	3	8		ATTTGCAT	103	46	99	GT/AG	304	CACAATG	23	ACCAAAACC
g	4	III						/AG	302	CACAGTG	23	ACAAACGCC
g	5	III		ATTTGCAT	102	46	112	GT/AG	289	CACAGTG	24	ACCAAAACC
g	6	III		ATTTGCAT	102	46	98	GT/AG	303	CGCAGTG		
g	7	III		ATTTGCAT	102	47	97	TC/AG	305	CAGAGTA	22	ACCAAAACC
g	8	8		ATTTGCAT	102	45	97	CT/TA	311	CACAGTG	23	CCTAAAAAC
g	9	8	F	ATTTGCAT	102	46	99	GT/AG	310	CACAGTG	23	CCTAAAACC

Appendix 6: Liquid phase blocking ELISA for sero-conversion of the African buffalo in the KNP infected with SAT1, SAT2 or SAT3 FMDV, carried out by Perez et al (2015) (Perez et al; unpublished).

		22-Sep-15	22-Sep-15	22-Sep-15	25-Sep-15	25-Sep-15	25-Sep-15	28-Sep-15	28-Sep-15	28-Sep-15	14-Oct-15	14-Oct-15	14-Oct-15
		Day 8	Day 8	Day 8	Day 11	Day 11	Day 11	Day 14	Day 14	Day 14	Day 30	Day 30	Day 30
Buff ID	Group	SAT1	SAT2	SAT3	SAT1	SAT2	SAT3	SAT1	SAT2	SAT3	SAT1	SAT2	SAT3
7	SAT1	2.2			2.2			2.2			2.2		
10	SAT1	2.2			2.2	1.7		2.2			2.2		
11	SAT1	1.9	2.2		1.9			2.2			1.8		
13	SAT1	2.2		1.7	2.2	2.2	1.8	2.2	2.2	1.7	2.2	2	1.6
8	SAT2		2.2			2.2			2.2			2.2	
20	SAT2		2.2			2.2			2.2			2.2	
28	SAT2		2.2		1.7	2.2		1.7	2.2			2.2	
32	SAT2		2.2			2.2			2.2			2.2	
26	SAT3									1.7			2.2
27	SAT3			1.7			1.9	1.6	1.7	2.2			2.2
34	SAT3									1.7			1.8
35	SAT3	1.8		1.6	1.8			2.2			1.8		1.8
2	AT1 Contact				2.2	1.9	1.7	2.2	2.2	1.8	2.2	2.2	2.2
4	AT1 Contact				2.2			2.2			2.2		
19	AT1 Contact				2.2			2.2	1.8	1.8	2.1		
33	AT1 Contact				1.7			2.2			2.2		
5	AT2 Contact					2.2			2.2			2.2	
9	AT2 Contact					2			2.2			2.2	
22	AT2 Contact							1.7	2.2			2.2	
29	AT2 Contact								2.2			2.2	
12	AT3 Contact												1.9
15	AT3 Contact												1.8
16	AT3 Contact												1.7
17	AT3 Contact					1.6		1.8	2.2	1.9		1.8	1.8

References

1 GLANZMANN, B., MOLLER, M., LE ROEX, N., TROMP, G., HOAL, E. G. & VAN HELDEN, P. D. 2016. The complete genome sequence of the African buffalo (Syncerus caffer). *BMC Genomics,* 17**,** 1001.

2 JONATHAN RUSHTON, T. K.-J. 2013. The impact of foot and mouth disease. *OIE*.

3 FAO, O. A. 2008. Annual OIE/FAO FMD Reference Laboratory Network Report. *Annual OIE/FAO network reports*.

4 LEBOUCQ, L. G. A. 2013. The role of animal disease control in poverty reduction, food safety, market access and food security in Africa. *FAO*.

5 FAO, R. L. 2017. "Current global status of Foot-and-Mouth Disease".

6 BELSHAM, G. 1993. Distinctive features of foot-and-mouth disease virus, a member of the picornavirus family; aspects of virus protein synthesis, protein processing and structure. *Progress in Biophysics and Molecular Biology*

7 JACKSON, T., KING, A. M., STUART, D. I. & FRY, E. 2003. Structure and receptor binding. *Virus research,* 91**,** 33-46.

8 ZHANG, Q., LIU, X., FANG, Y., PAN, L., LV, J., ZHANG, Z., ZHOU, P., DING, Y., CHEN, H., SHAO, J., ZHAO, F., LIN, T., CHANG, H., ZHANG, J., WANG, Y. & ZHANG, Y. 2015. Evolutionary Analysis of Structural Protein Gene VP1 of Foot-and-Mouth Disease Virus Serotype Asia 1. *The Scientific World Journal,* 2015**,** 734253.

9 NAGENDRAKUMAR, S. B., REDDY, G. S., CHANDRAN, D., THIAGARAJAN, D., RANGARAJAN, P. N. & SRINIVASAN, V. A. 2005. Molecular Characterization of Foot-and-Mouth Disease Virus Type C of Indian Origin. *J Clin Microbiol,* 43**,** 966-9.

10 PATON, D. J., SUMPTION, K. J. & CHARLESTON, B. 2009. Options for control of foot-and-mouth disease: knowledge, capability and policy. *Philos Trans R Soc Lond B Biol Sci,* 364**,** 2657-67.

11 GRUBMAN, M. J. & BAXT, B. 2004. Foot-and-mouth disease. *Clin Microbiol Rev,* 17**,** 465-93.

12 CHASE, A. J. & SEMLER, B. L. 2012. Viral subversion of host functions for picornavirus translation and RNA replication. *Future Virol,* 7**,** 179-91.

13 SAUNDERS, K. & KING, A. M. 1982. Guanidine-resistant mutants of aphthovirus induce the synthesis of an altered nonstructural polypeptide, P34. *J Virol,* 42**,** 389-94.

14 VAKHARIA, V. N., DEVANEY, M. A., MOORE, D. M., DUNN, J. J. & GRUBMAN, M. J. 1987. Proteolytic processing of foot-and-mouth disease virus polyproteins expressed in a cell-free system from clone-derived transcripts. *J Virol,* 61**,** 3199-207.

15 GRIGERA, P. R. A. S. G. T. 1984. Hsitone H3 modification in BHK cells infected with foot-and-mouth disease virus. *Virology,* 136**,** 10-19.

16 CHINSANGARAM, J., KOSTER, M. & GRUBMAN, M. J. 2001. Inhibition of L-Deleted Foot-and-Mouth Disease Virus Replication by Alpha/Beta Interferon Involves Double-Stranded RNA-Dependent Protein Kinase. *J Virol,* 75**,** 5498-503.

17 MOFFAT, K., HOWELL, G., KNOX, C., BELSHAM, G. J., MONAGHAN, P., RYAN, M. D. & WILEMAN, T. 2005. Effects of Foot-and-Mouth Disease Virus Nonstructural Proteins on the Structure and Function of the Early Secretory Pathway: 2BC but Not 3A Blocks Endoplasmic Reticulum-to-Golgi Transport. *J Virol,* 79**,** 4382-95.

18 MOFFAT, K., KNOX, C., HOWELL, G., CLARK, S. J., YANG, H., BELSHAM, G. J., RYAN, M. & WILEMAN, T. 2007. Inhibition of the Secretory Pathway by Foot-and-Mouth Disease Virus 2BC Protein Is Reproduced by Coexpression of 2B with 2C, and the Site of Inhibition Is Determined by the Subcellular Location of 2C. *J Virol,* 81**,** 1129-39.

19 BALINDA, S. N., SIEGISMUND, H. R., MUWANIKA, V. B., SANGULA, A. K., MASEMBE, C., AYEBAZIBWE, C., NORMANN, P. & BELSHAM, G. J. 2010. Phylogenetic analyses of the polyprotein coding sequences of serotype O foot-and-mouth disease viruses in East Africa: evidence for interserotypic recombination. *Virol J,* 7**,** 199.

20 HYNES, R. O. 2002. Integrins: bidirectional, allosteric signaling machines. *Cell,* 110**,** 673-87.

21 JACKSON, T., SHEPPARD, D., DENYER, M., BLAKEMORE, W. & KING, A. M. 2000. The epithelial integrin alphavbeta6 is a receptor for foot-and-mouth disease virus. *J Virol,* 74**,** 4949-56.

22 MONAGHAN, P., GOLD, S., SIMPSON, J., ZHANG, Z., WEINREB, P. H., VIOLETTE, S. M., ALEXANDERSEN, S. & JACKSON, T. 2005. The alpha(v)beta6 integrin receptor for Foot-and-mouth disease virus is expressed constitutively on the epithelial cells targeted in cattle. *J Gen Virol,* 86**,** 2769-80.

23 BAXT, B. & MASON, P. W. 1995. Foot-and-mouth disease virus undergoes restricted replication in macrophage cell cultures following Fc receptor-mediated adsorption. *Virology,* 207**,** 503-9.

24 JONES, K. S., PETROW-SADOWSKI, C., BERTOLETTE, D. C., HUANG, Y. & RUSCETTI, F. W. 2005. Heparan Sulfate Proteoglycans Mediate Attachment and Entry of Human T-Cell Leukemia Virus Type 1 Virions into CD4(+) T Cells. *J Virol,* 79**,** 12692-702.

25 PIERSCHBACHER, M. D. & RUOSLAHTI, E. 1984. Variants of the cell recognition site of fibronectin that retain attachment-promoting activity. *Proc Natl Acad Sci U S A,* 81**,** 5985-8.

26 G.J. BELSHAM, T. J., D. PATON, AND B. CHARLESTON 2009. Foot-and-Mouth disease. *John Wiley & Sons*.

27 BAYISSA, B., AYELET, G., KYULE, M., JIBRIL, Y. & GELAYE, E. 2011. Study on seroprevalence, risk factors, and economic impact of foot-and-mouth disease in Borena pastoral and agro-pastoral system, southern Ethiopia. *Trop Anim Health Prod,* 43**,** 759-66.

28 SKINNER, H. H. 1951. Propagation of Strains of Foot-and-Mouth Disease Virus in Unweaned White Mice. *Proc R Soc Med,* 44**,** 1041-4.

29 DONALDSON, A. I., GLOSTER, J., HARVEY, L. D. & DEANS, D. H. 1982. Use of prediction models to forecast and analyse airborne spread during the foot-and-mouth disease outbreaks in Brittany, Jersey and the Isle of Wight in 1981. *Vet Rec,* 110**,** 53-7.

30 DONALDSON, A. I. 1979. Airborne foot-and-mouth disease. *Veterinary Bulletins,* 49**,** 653-659.

31 ALEXANDERSEN, S., ZHANG, Z., DONALDSON, A. I. & GARLAND, A. J. 2003. The pathogenesis and diagnosis of foot-and-mouth disease. *J Comp Pathol,* 129**,** 1-36.

32 CHARLESTON, B., BANKOWSKI, B. M., GUBBINS, S., CHASE-TOPPING, M. E., SCHLEY, D., HOWEY, R., BARNETT, P. V., GIBSON, D., JULEFF, N. D. & WOOLHOUSE, M. E. 2011. Relationship between clinical signs and transmission of an infectious disease and the implications for control. *Science,* 332**,** 726-9.

33 WOOLHOUSE, M. E., HAYDON, D. T., PEARSON, A. & KITCHING, R. P. 1996. Failure of vaccination to prevent outbreaks of foot-and-mouth disease. *Epidemiol Infect,* 116**,** 363-71.

34 BURROWS 1968. Excretion of foot-and-mouth disease virus prior to the development of lesions. *Veterinary Record,* 83**,** 387-388.

35 FERGUSON, N. M., DONNELLY, C. A. & ANDERSON, R. M. 2001. The foot-and-mouth epidemic in Great Britain: pattern of spread and impact of interventions. *Science,* 292**,** 1155-60.

36 ALEXANDERSEN, S., QUAN, M., MURPHY, C., KNIGHT, J. AND ZHANG, Z. 2003. Studies of quantitative parameters of virus excretion and transmission in pigs and cattle experimentally infected with foot-and-mouth disease virus. *Journal of Comparative Pathology*.

37 VAN BEKKUM J., F. H., FREDERIKS H. & FRENKEL S 1959. Observations on the carrier state of cattle exposed to foot- and-mouth disease virus. *Tijdschrift voor Diergeneeskunde,* 84**,** 1159–1164

38 PACHECO, J. M., SMOLIGA, G. R., O'DONNELL, V., BRITO, B. P., STENFELDT, C., RODRIGUEZ, L. L. & ARZT, J. 2015. Persistent Foot-and-Mouth Disease Virus Infection in the Nasopharynx of Cattle; Tissue-Specific Distribution and Local Cytokine Expression. *PLoS One,* 10**,** e0125698.

39 JULEFF, N. D., MAREE, F. F., WATERS, R., BENGIS, R. G. & CHARLESTON, B. 2012. The importance of FMDV localisation in lymphoid tissue. *Vet Immunol Immunopathol,* 148**,** 145-8.

40 BRONSVOORT, B. M., HANDEL, I. G., NFON, C. K., SORENSEN, K. J., MALIRAT, V., BERGMANN, I., TANYA, V. N. & MORGAN, K. L. 2016. Redefining the "carrier" state for foot-and-mouth disease from the dynamics of virus persistence in endemically affected cattle populations. *Sci Rep,* 6**,** 29059.

41 VOSLOO, W., BASTOS, A. D., KIRKBRIDE, E., ESTERHUYSEN, J. J., VAN RENSBURG, D. J., BENGIS, R. G., KEET, D. W. & THOMSON, G. R. 1996. Persistent infection of African buffalo (Syncerus caffer) with SAT-type foot-and-mouth disease viruses: rate of fixation of mutations, antigenic change and interspecies transmission. *J Gen Virol,* 77 (Pt 7)**,** 1457-67.

42 BENGIS, R. G., THOMSON, G. R., HEDGER, R. S., DE VOS, V. & PINI, A. 1986. Foot-and-mouth disease and the African buffalo (Syncerus caffer). 1. Carriers as a source of infection for cattle. *Onderstepoort J Vet Res,* 53**,** 69-73.

43 MAREE, F., DE KLERK-LORIST, L. M., GUBBINS, S., ZHANG, F., SEAGO, J., PEREZ-MARTIN, E., REID, L., SCOTT, K., VAN SCHALKWYK, L., BENGIS, R., CHARLESTON, B. & JULEFF, N. 2016. Differential Persistence of Foot-and-Mouth Disease Virus in African Buffalo Is Related to Virus Virulence. *J Virol,* 90**,** 5132-40.

44 CONDY, J. B., HEDGER, R. S., HAMBLIN, C. & BARNETT, I. T. 1985. The duration of the foot-and-mouth disease virus carrier state in African buffalo (i) in the individual animal and (ii) in a free-living herd. *Comp Immunol Microbiol Infect Dis,* 8**,** 259-65.

45 NAIDOO, R., DU PREEZ, P., STUART-HILL, G., JAGO, M. & WEGMANN, M. 2012. Home on the range: factors explaining partial migration of African buffalo in a tropical environment. *PLoS One,* 7**,** e36527.

46 CONDY, J. B. & HEDGER, R. S. 1974. The survival of foot-and-mouth disease virus in African buffalo with non-transference of infection to domestic cattle. *Res Vet Sci,* 16**,** 182-5.

47 DAWE, P. S., FLANAGAN, F. O., MADEKUROZWA, R. L., SORENSEN, K. J., ANDERSON, E. C., FOGGIN, C. M., FERRIS, N. P. & KNOWLES, N. J. 1994. Natural transmission of foot-and-mouth disease virus from African buffalo (Syncerus caffer) to cattle in a wildlife area of Zimbabwe. *Vet Rec,* 134**,** 230-2.

48 THOMSON, G. R., VOSLOO, W. & BASTOS, A. D. S. 2003. Foot and mouth disease in wildlife. *Virus Research,* 91**,** 145-161.

49 WALDMAN O., P. J. 1920. Die künstliche Übertragung der Maul-und Klauenseuche auf das Meerschweinchen. *Berl Munch Tierarztl Wochenschr,* 36**,** 519–520.

50 KNUDSEN, R. C., GROOCOCK, C. M. & ANDERSEN, A. A. 1979. Immunity to foot-and-mouth disease virus in guinea pigs: clinical and immune responses. *Infect Immun,* 24**,** 787-92.

51 ROIVAINEN, M., PIIRAINEN, L., HOVI, T., VIRTANEN, I., RIIKONEN, T., HEINO, J. & HYYPIA, T. 1994. Entry of coxsackievirus A9 into host cells: specific interactions with alpha v beta 3 integrin, the vitronectin receptor. *Virology,* 203**,** 357-65.

52 BORCA, M. V., FERNANDEZ, F. M., SADIR, A. M., BRAUN, M. & SCHUDEL, A. A. 1986. Immune response to foot-and-mouth disease virus in a murine experimental model: effective thymus-independent primary and secondary reaction. *Immunology,* 59**,** 261-7.

53 LOPEZ, O. J., SADIR, A. M., BORCA, M. V., FERNANDEZ, F. M., BRAUN, M. & SCHUDEL, A. A. 1990. Immune response to foot-and-mouth disease virus in an experimental murine model. II. Basis of persistent antibody reaction. *Vet Immunol Immunopathol,* 24**,** 313-21.

54 JULEFF, N., WINDSOR, M., LEFEVRE, E. A., GUBBINS, S., HAMBLIN, P., REID, E., MCLAUGHLIN, K., BEVERLEY, P. C., MORRISON, I. W. & CHARLESTON, B. 2009. Foot-and-mouth disease virus can induce a specific and rapid CD4+ T-cell-independent neutralizing and isotype class-switched antibody response in naive cattle. *J Virol,* 83**,** 3626-36.

55 SANZ-PARRA, A., SOBRINO, F. & LEY, V. 1998. Infection with foot-and-mouth disease virus results in a rapid reduction of MHC class I surface expression. *J Gen Virol,* 79 (Pt 3)**,** 433-6.

56 BARNETT, P. V., COX, S. J., AGGARWAL, N., GERBER, H. & MCCULLOUGH, K. C. 2002. Further studies on the early protective responses of pigs following immunisation with high potency foot and mouth disease vaccine. *Vaccine,* 20**,** 3197-208.

57 RIGDEN, R. C., CARRASCO, C. P., SUMMERFIELD, A. & KC, M. C. 2002. Macrophage phagocytosis of foot-and-mouth disease virus may create infectious carriers. *Immunology,* 106**,** 537-48.

58 MCCULLOUGH, K. C., CROWTHER, J. R., BUTCHER, R. N., CARPENTER, W. C., BROCCHI, E., CAPUCCI, L. & DE SIMONE, F. 1986. Immune protection against foot-and-mouth disease virus studied using virus-neutralizing and non-neutralizing concentrations of monoclonal antibodies. *Immunology,* 58**,** 421-8.

59 MCCULLOUGH, K. C., PARKINSON, D. & CROWTHER, J. R. 1988. Opsonization-enhanced phagocytosis of foot-and-mouth disease virus. *Immunology,* 65**,** 187-91.

60 BAUTISTA, E. M., FERMAN, G. S., GREGG, D., BRUM, M. C., GRUBMAN, M. J. & GOLDE, W. T. 2005. Constitutive expression of alpha interferon by skin dendritic cells confers resistance to infection by foot-and-mouth disease virus. *J Virol,* 79**,** 4838-47.

61 BERGAMIN, F., VINCENT, I. E., SUMMERFIELD, A. & MCCULLOUGH, K. C. 2007. Essential role of antigen-presenting cell-derived BAFF for antibody responses. *Eur J Immunol,* 37**,** 3122-30.

62 BANCHEREAU, J. & STEINMAN, R. M. 1998. Dendritic cells and the control of immunity. *Nature,* 392**,** 245-52.

63 MELLMAN, I. & STEINMAN, R. M. 2001. Dendritic cells: specialized and regulated antigen processing machines. *Cell,* 106**,** 255-8.

64 OSTROWSKI, M., VERMEULEN, M., ZABAL, O., GEFFNER, J. R., SADIR, A. M. & LOPEZ, O. J. 2005. Impairment of thymus-dependent responses by murine dendritic cells infected with foot-and-mouth disease virus. *J Immunol,* 175**,** 3971-9.

65 SUMMERFIELD, A., GUZYLACK-PIRIOU, L., HARWOOD, L. & MCCULLOUGH, K. C. 2009. Innate immune responses against foot-and-mouth disease virus: current understanding and future directions. *Vet Immunol Immunopathol,* 128**,** 205-10.

66 REID, E. & CHARLESTON, B. 2014. Type I and III interferon production in response to RNA viruses. *J Interferon Cytokine Res,* 34**,** 649-58.

67 SCHLEE, M. 2013. Master sensors of pathogenic RNA - RIG-I like receptors. *Immunobiology,* 218**,** 1322-35.

68 MASON, P. W., CHINSANGARAM, J., MORAES, M. P., MAYR, G. A. & GRUBMAN, M. J. 2003. Engineering better vaccines for foot-and-mouth disease. *Dev Biol (Basel),* 114**,** 79-88.

69 GUZYLACK-PIRIOU, L., BERGAMIN, F., GERBER, M., MCCULLOUGH, K. C. & SUMMERFIELD, A. 2006. Plasmacytoid dendritic cell activation by foot-and-mouth disease virus requires immune complexes. *Eur J Immunol,* 36**,** 1674-83.

70 OH, Y., FLEMING, L., STATHAM, B., HAMBLIN, P., BARNETT, P., PATON, D. J., PARK, J. H., JOO, Y. S. & PARIDA, S. 2012. Interferon-gamma induced by in vitro re-stimulation of CD4+ T-cells correlates with in vivo FMD vaccine induced protection of cattle against disease and persistent infection. *PLoS One,* 7**,** e44365.

71 FOWLER, V., ROBINSON, L., BANKOWSKI, B., COX, S., PARIDA, S., LAWLOR, C., GIBSON, D., O'BRIEN, F., ELLEFSEN, B., HANNAMAN, D., TAKAMATSU, H. H. & BARNETT, P. V. 2012. A DNA vaccination regime including protein boost and electroporation protects cattle against foot-and-mouth disease. *Antiviral Res,* 94**,** 25-34.

72 PEREZ-MARTIN, E., WEISS, M., DIAZ-SAN SEGUNDO, F., PACHECO, J. M., ARZT, J., GRUBMAN, M. J. & DE LOS SANTOS, T. 2012. Bovine type III interferon significantly delays and reduces the severity of foot-and-mouth disease in cattle. *J Virol,* 86**,** 4477-87.

73 WACK, A., TERCZYNSKA-DYLA, E. & HARTMANN, R. 2015. Guarding the frontiers: the biology of type III interferons. 16**,** 802-9.

74 PATCH, J. R., DAR, P. A., WATERS, R., TOKA, F. N., BARRERA, J., SCHUTTA, C., KONDABATTULA, G. & GOLDE, W. T. 2014. Infection with foot-and-mouth disease virus (FMDV) induces a natural killer (NK) cell response in cattle that is lacking following vaccination. *Comp Immunol Microbiol Infect Dis,* 37**,** 249-57.

75 TOKA, F. N., NFON, C., DAWSON, H. & GOLDE, W. T. 2009. Natural killer cell dysfunction during acute infection with foot-and-mouth disease virus. *Clin Vaccine Immunol,* 16**,** 1738-49.

76 AMADORI, M., ARCHETTI, I. L., VERARDI, R. & BERNERI, C. 1992. Target recognition by bovine mononuclear, MHC-unrestricted cytotoxic cells. *Vet Microbiol,* 33**,** 383-92.

77 BRADFORD, H. E., ADAIR, B. M. & FOSTER, J. C. 2001. Antibody-dependent killing of virus-infected targets by NK-like cells in bovine blood. *J Vet Med B Infect Dis Vet Public Health,* 48**,** 637-40.

78 BIBURGER, M., LUX, A. & NIMMERJAHN, F. 2014. How immunoglobulin G antibodies kill target cells: revisiting an old paradigm. *Adv Immunol,* 124**,** 67-94.

79 DOEL, T. R. 2005. Natural and vaccine induced immunity to FMD. *Curr Top Microbiol Immunol,* 288**,** 103-31.

80 CUNLIFFE, H. R. 1964. OBSERVATIONS ON THE DURATION OF IMMUNITY IN CATTLE AFTER EXPERIMENTAL INFECTION WITH FOOT-AND-MOUTH DISEASE VIRUS. *Cornell Vet,* 54**,** 501-10.

81 GARLAND, A. J. M. 1974. The inhibitory activity of secretions in cattle against foot and mouth disease virus. PhD thesis. *London School of Hygiene & Tropical Medicine*.

82 ESCHBAUMER, M., STENFELDT, C., REKANT, S. I., PACHECO, J. M., HARTWIG, E. J., SMOLIGA, G. R., KENNEY, M. A., GOLDE, W. T., RODRIGUEZ, L. L. & ARZT, J. 2016. Systemic immune response and virus persistence after foot-and-mouth disease virus infection of naïve cattle and cattle vaccinated with a homologous adenovirus-vectored vaccine. *BMC Vet Res,* 12.

83 PEGA, J., BUCAFUSCO, D., DI GIACOMO, S., SCHAMMAS, J. M., MALACARI, D., CAPOZZO, A. V., ARZT, J., PEREZ-BEASCOECHEA, C., MARADEI, E., RODRIGUEZ, L. L., BORCA, M. V. & PEREZ-FILGUEIRA, M. 2013. Early adaptive immune responses in the respiratory tract of foot-and-mouth disease virus-infected cattle. *J Virol,* 87**,** 2489-95.

84 COLLEN 1994. Foot and mouth disease (Aphthovirus): viral T cell epitopes. *In Cell-Mediated Immunity in Ruminants***,** 173–197.

85 PARIDA, S., ANDERSON, J., COX, S. J., BARNETT, P. V. & PATON, D. J. 2006. Secretory IgA as an indicator of oro-pharyngeal foot-and-mouth disease virus replication and as a tool for post vaccination surveillance. *Vaccine,* 24**,** 1107-16.

86 CONDY, J. B., HERNIMAN, K. A. & HEDGER, R. S. 1969. Foot-and-mouth disease in wildlife in Rhodesia and other African territories. A serological survey. *J Comp Pathol,* 79.

87 BRONSVOORT, B. M., PARIDA, S., HANDEL, I., MCFARLAND, S., FLEMING, L., HAMBLIN, P. & KOCK, R. 2008. Serological survey for foot-and-mouth disease virus in wildlife in eastern Africa and estimation of test parameters of a nonstructural protein enzyme-linked immunosorbent assay for buffalo. *Clin Vaccine Immunol,* 15**,** 1003-11.

88 DI NARDO, A., LIBEAU, G., CHARDONNET, B., CHARDONNET, P., KOCK, R. A., PAREKH, K., HAMBLIN, P., LI, Y., PARIDA, S. & SUMPTION, K. J. 2015. Serological

profile of foot-and-mouth disease in wildlife populations of West and Central Africa with special reference to Syncerus caffer subspecies. *Veterinary Research,* 46**,** 77.

89 FRANCIS, M. J., OULDRIDGE, E. J. & BLACK, L. 1983. Antibody response in bovine pharyngeal fluid following foot-and-mouth disease vaccination and, or, exposure to live virus. *Res Vet Sci,* 35**,** 206-10.

90 HEDGER, R. S. 1972. Foot-and-mouth disease and the African buffalo (Syncerus caffer). *J Comp Pathol,* 82.

91 GRANT, C. 2013. PhD diss. *University of Oxford*.

92 WANG, F., EKIERT, D. C., AHMAD, I., YU, W., ZHANG, Y., BAZIRGAN, O., TORKAMANI, A., RAUDSEPP, T., MWANGI, W., CRISCITIELLO, M. F., WILSON, I. A., SCHULTZ, P. G. & SMIDER, V. V. 2013. Reshaping antibody diversity. *Cell,* 153**,** 1379-93.

93 COLLIS, A. V., BROUWER, A. P. & MARTIN, A. C. 2003. Analysis of the antigen combining site: correlations between length and sequence composition of the hypervariable loops and the nature of the antigen. *J Mol Biol,* 325**,** 337-54.

94 JOHANSEN, F. E., BRAATHEN, R. & BRANDTZAEG, P. 2000. Role of J chain in secretory immunoglobulin formation. *Scand J Immunol,* 52**,** 240-8.

95 SNAPPER, C. M., MCINTYRE, T. M., MANDLER, R., PECANHA, L. M., FINKELMAN, F. D., LEES, A. & MOND, J. J. 1992. Induction of IgG3 secretion by interferon gamma: a model for T cell-independent class switching in response to T cell-independent type 2 antigens. *J Exp Med,* 175**,** 1367-71.

96 SNAPPER, C. M. & PAUL, W. E. 1987. Interferon-gamma and B cell stimulatory factor-1 reciprocally regulate Ig isotype production. *Science,* 236**,** 944-7.

97 TRINCHIERI, G. & VALIANTE, N. 1993. Receptors for the Fc fragment of IgG on natural killer cells. *Nat Immun,* 12**,** 218-34.

98 MA, L., QIN, T., CHU, D., CHENG, X., WANG, J., WANG, X., WANG, P., HAN, H., REN, L. & AITKEN, R. 2016. Internal Duplications of DH, JH, and C Region Genes Create an Unusual IgH Gene Locus in Cattle. 196**,** 4358-66.

99 EKMAN, A., NIKU, M., LILJAVIRTA, J. & IIVANAINEN, A. 2009. Bos taurus genome sequence reveals the assortment of immunoglobulin and surrogate light chain genes in domestic cattle. *BMC Immunology,* 10**,** 1-11.

100 MURPHY, K. 2012. Immunobiology: The Immune System. *Janeway's Immunobiology,* 8**,** 888.

101 LILJAVIRTA, J., NIKU, M., PESSA-MORIKAWA, T., EKMAN, A. & IIVANAINEN, A. 2014. Expansion of the preimmune antibody repertoire by junctional diversity in Bos taurus. *PLoS One,* 9**,** e99808.

102 RAJEWSKY, K. 1996. Clonal selection and learning in the antibody system. *Nature,* 381**,** 751-8.

103 EKMAN, A., PESSA-MORIKAWA, T., LILJAVIRTA, J., NIKU, M. & IIVANAINEN, A. 2010. B-cell development in bovine fetuses proceeds via a pre-B like cell in bone marrow and lymph nodes. *Dev Comp Immunol,* 34**,** 896-903.

104 CERUTTI, A., PUGA, I. & MAGRI, G. 2013. The B cell helper side of neutrophils. *J Leukoc Biol,* 94**,** 677-82.

105 PHAN, T. G., GRIGOROVA, I., OKADA, T. & CYSTER, J. G. 2007. Subcapsular encounter and complement-dependent transport of immune complexes by lymph node B cells. *Nat Immunol,* 8**,** 992-1000.

106 BATISTA, F. D. & HARWOOD, N. E. 2009. The who, how and where of antigen presentation to B cells. *Nat Rev Immunol,* 9**,** 15-27.

107 UNANUE, E. R., CEROTTINI, J. C. & BEDFORD, M. 1969. Persistence of antigen on the surface of macrophages. *Nature,* 222**,** 1193-5.

108 HEESTERS, B. A., CHATTERJEE, P., KIM, Y. A., GONZALEZ, S. F., KULIGOWSKI, M. P., KIRCHHAUSEN, T. & CARROLL, M. C. 2013. Endocytosis and recycling of immune complexes by follicular dendritic cells enhances B cell antigen binding and activation. *Immunity,* 38**,** 1164-75.

109 CAHALAN, M. D. & PARKER, I. 2008. Choreography of cell motility and interaction dynamics imaged by two-photon microscopy in lymphoid organs. *Annu Rev Immunol,* 26**,** 585-626.

110 QI, H., EGEN, J. G., HUANG, A. Y. & GERMAIN, R. N. 2006. Extrafollicular activation of lymph node B cells by antigen-bearing dendritic cells. *Science,* 312**,** 1672-6.

111 VOS, Q., LEES, A., WU, Z. Q., SNAPPER, C. M. & MOND, J. J. 2000. B-cell activation by T-cell-independent type 2 antigens as an integral part of the humoral immune response to pathogenic microorganisms. *Immunol Rev,* 176**,** 154-70.

112 JACOB, J., KASSIR, R. & KELSOE, G. 1991. In situ studies of the primary immune response to (4-hydroxy-3-nitrophenyl)acetyl. I. The architecture and dynamics of responding cell populations. *J Exp Med,* 173**,** 1165-75.

113 DI NOIA, J. & NEUBERGER, M. S. 2002. Altering the pathway of immunoglobulin hypermutation by inhibiting uracil-DNA glycosylase. *Nature,* 419**,** 43-8.

114 PETERSEN-MAHRT, S. K., HARRIS, R. S. & NEUBERGER, M. S. 2002. AID mutates E. coli suggesting a DNA deamination mechanism for antibody diversification. *Nature,* 418**,** 99-103.

115 LUO, Z., RONAI, D. & SCHARFF, M. D. 2004. The role of activation-induced cytidine deaminase in antibody diversification, immunodeficiency, and B-cell malignancies. *J Allergy Clin Immunol,* 114**,** 726-35; quiz 736.

116 WAGNER, S. D. & NEUBERGER, M. S. 1996. Somatic hypermutation of immunoglobulin genes. *Annu Rev Immunol,* 14**,** 441-57.

117 COLAIACOVO, M. P., PAQUES, F. & HABER, J. E. 1999. Removal of one nonhomologous DNA end during gene conversion by a RAD1- and MSH2-independent pathway. *Genetics,* 151**,** 1409-23.

118 KUROSAWA, K. & OHTA, K. 2011. Genetic Diversification by Somatic Gene Conversion. *Genes (Basel),* 2**,** 48-58.

119 BERENS, S. J., WYLIE, D. E. & LOPEZ, O. J. 1997. Use of a single VH family and long CDR3s in the variable region of cattle Ig heavy chains. *Int Immunol,* 9**,** 189-99.

120 KAUSHIK, A. K., KEHRLI, M. E., JR., KURTZ, A., NG, S., KOTI, M., SHOJAEI, F. & SAINI, S. S. 2009. Somatic hypermutations and isotype restricted exceptionally long CDR3H contribute to antibody diversification in cattle. *Vet Immunol Immunopathol,* 127**,** 106-13.

121 SAINI, S. S., HEIN, W. R. & KAUSHIK, A. 1997. A single predominantly expressed polymorphic immunoglobulin VH gene family, related to mammalian group, I, clan, II, is identified in cattle. *Mol Immunol,* 34**,** 641-51.

122 VERMA, S. & AITKEN, R. 2012. Somatic hypermutation leads to diversification of the heavy chain immunoglobulin repertoire in cattle. *Vet Immunol Immunopathol,* 145**,** 14-22.

123 LUCIER, M. R., THOMPSON, R. E., WAIRE, J., LIN, A. W., OSBORNE, B. A. & GOLDSBY, R. A. 1998. Multiple sites of V lambda diversification in cattle. *J Immunol,* 161**,** 5438-44.

124 PARNG, C.-L., HANSAL, S., GOLDSBY, R. A. & OSBORNE, B. A. 1996. Gene conversion contributes to Ig light chain diversity in cattle. *The Journal of Immunology,* 157**,** 5478-5486.

125 LILJAVIRTA, J., EKMAN, A., KNIGHT, J. S., PERNTHANER, A., IIVANAINEN, A. & NIKU, M. 2013. Activation-induced cytidine deaminase (AID) is strongly expressed in the fetal bovine ileal Peyer's patch and spleen and is associated with expansion of the primary antibody repertoire in the absence of exogenous antigens. *Mucosal Immunol,* 6**,** 942-9.

126 YASUDA, M., JENNE, C. N., KENNEDY, L. J. & REYNOLDS, J. D. 2006. The sheep and cattle Peyer's patch as a site of B-cell development. *Vet Res,* 37**,** 401-15.

127 ONISHI, S., MIYATA, H., INAMOTO, T., QI, W. M., YAMAMOTO, K., YOKOYAMA, T., WARITA, K., HOSHI, N. & KITAGAWA, H. 2007. Immunohistochemical study on the delayed progression of epithelial apoptosis in follicle-associated epithelium of rat Peyer's patch. *J Vet Med Sci,* 69**,** 1123-9.

128 GRIFFITHS, G. M., BEREK, C., KAARTINEN, M. & MILSTEIN, C. 1984. Somatic mutation and the maturation of immune response to 2-phenyl oxazolone. *Nature,* 312**,** 271-5.

129 HAUSER, A. E., JUNT, T., MEMPEL, T. R., SNEDDON, M. W., KLEINSTEIN, S. H., HENRICKSON, S. E., VON ANDRIAN, U. H., SHLOMCHIK, M. J. & HABERMAN, A.

M. 2007. Definition of germinal-center B cell migration in vivo reveals predominant intrazonal circulation patterns. *Immunity,* 26**,** 655-67.

130 ZHANG, Y., MEYER-HERMANN, M., GEORGE, L. A., FIGGE, M. T., KHAN, M., GOODALL, M., YOUNG, S. P., REYNOLDS, A., FALCIANI, F., WAISMAN, A., NOTLEY, C. A., EHRENSTEIN, M. R., KOSCO-VILBOIS, M. & TOELLNER, K. M. 2013. Germinal center B cells govern their own fate via antibody feedback. *J Exp Med,* 210**,** 457-64.

131 HELMREICH, E., KERN, M. & EISEN, H. N. 1961. The secretion of antibody by isolated lymph node cells. *J Biol Chem,* 236**,** 464-73.

132 HIBI, T. & DOSCH, H. M. 1986. Limiting dilution analysis of the B cell compartment in human bone marrow. *Eur J Immunol,* 16**,** 139-45.

133 STAVNEZER, J. & SCHRADER, C. E. 2014. IgH chain class switch recombination: mechanism and regulation. *J Immunol,* 193**,** 5370-8.

134 MANIS, J. P., GU, Y., LANSFORD, R., SONODA, E., FERRINI, R., DAVIDSON, L., RAJEWSKY, K. & ALT, F. W. 1998. Ku70 is required for late B cell development and immunoglobulin heavy chain class switching. *J Exp Med,* 187**,** 2081-9.

135 XU, Z., ZAN, H., PONE, E. J., MAI, T. & CASALI, P. 2012. Immunoglobulin class-switch DNA recombination: induction, targeting and beyond. *Nat Rev Immunol,* 12**,** 517-31.

136 PONE, E. J., ZHANG, J., MAI, T., WHITE, C. A., LI, G., SAKAKURA, J. K., PATEL, P. J., AL-QAHTANI, A., ZAN, H., XU, Z. & CASALI, P. 2012. BCR-signalling synergizes with TLR-signalling for induction of AID and immunoglobulin class-switching through the non-canonical NF-kappaB pathway. *Nat Commun,* 3**,** 767.

137 WEINSTEIN, J. A., JIANG, N., WHITE, R. A., 3RD, FISHER, D. S. & QUAKE, S. R. 2009. High-throughput sequencing of the zebrafish antibody repertoire. *Science,* 324**,** 807-10.

138 LARSEN, P. A. & SMITH, T. P. 2012. Application of circular consensus sequencing and network analysis to characterize the bovine IgG repertoire. *BMC Immunol,* 13**,** 52.

139 GREIFF, V., MENZEL, U., HAESSLER, U., COOK, S. C., FRIEDENSOHN, S., KHAN, T. A., POGSON, M., HELLMANN, I. & REDDY, S. T. 2014. Quantitative assessment of the robustness of next-generation sequencing of antibody variable gene repertoires from immunized mice. *BMC Immunol,* 15**,** 40.

140 LASERSON, U., VIGNEAULT, F., GADALA-MARIA, D., YAARI, G., UDUMAN, M., VANDER HEIDEN, J. A., KELTON, W., TAEK JUNG, S., LIU, Y., LASERSON, J., CHARI, R., LEE, J. H., BACHELET, I., HICKEY, B., LIEBERMAN-AIDEN, E., HANCZARUK, B., SIMEN, B. B., EGHOLM, M., KOLLER, D., GEORGIOU, G., KLEINSTEIN, S. H. & CHURCH, G. M. 2014. High-resolution antibody dynamics of vaccine-induced immune responses. *Proc Natl Acad Sci U S A,* 111**,** 4928-33.

141 ROBINSON, W. H., STEINMAN, L. & UTZ, P. J. 2003. Protein arrays for autoantibody profiling and fine-specificity mapping. *Proteomics,* 3**,** 2077-84.

142 ZIMIN, A. V., DELCHER, A. L., FLOREA, L., KELLEY, D. R., SCHATZ, M. C., PUIU, D., HANRAHAN, F., PERTEA, G., VAN TASSELL, C. P., SONSTEGARD, T. S., MARCAIS, G., ROBERTS, M., SUBRAMANIAN, P., YORKE, J. A. & SALZBERG, S. L. 2009. A whole-genome assembly of the domestic cow, Bos taurus. *Genome Biol,* 10**,** R42.

143 SNELLING, W. M., CHIU, R., SCHEIN, J. E., HOBBS, M., ABBEY, C. A., ADELSON, D. L., AERTS, J., BENNETT, G. L., BOSDET, I. E., BOUSSAHA, M., BRAUNING, R., CAETANO, A. R., COSTA, M. M., CRAWFORD, A. M., DALRYMPLE, B. P., EGGEN, A., EVERTS-VAN DER WIND, A., FLORIOT, S., GAUTIER, M., GILL, C. A., GREEN, R. D., HOLT, R., JANN, O., JONES, S. J., KAPPES, S. M., KEELE, J. W., DE JONG, P. J., LARKIN, D. M., LEWIN, H. A., MCEWAN, J. C., MCKAY, S., MARRA, M. A., MATHEWSON, C. A., MATUKUMALLI, L. K., MOORE, S. S., MURDOCH, B., NICHOLAS, F. W., OSOEGAWA, K., ROY, A., SALIH, H., SCHIBLER, L., SCHNABEL, R. D., SILVERI, L., SKOW, L. C., SMITH, T. P., SONSTEGARD, T. S., TAYLOR, J. F., TELLAM, R., VAN TASSELL, C. P., WILLIAMS, J. L., WOMACK, J. E., WYE, N. H., YANG, G. & ZHAO, S. 2007. A physical map of the bovine genome. *Genome Biol,* 8**,** R165.

144 QUAIL, M. A., SMITH, M., COUPLAND, P., OTTO, T. D., HARRIS, S. R., CONNOR, T. R., BERTONI, A., SWERDLOW, H. P. & GU, Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics,* 13**,** 341.

145 SHIZUYA, H., BIRREN, B., KIM, U. J., MANCINO, V., SLEPAK, T., TACHIIRI, Y. & SIMON, M. 1992. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. *Proceedings of the National Academy of Sciences of the United States of America,* 89**,** 8794-8797.

146 RATZKIN, B. & CARBON, J. 1977. Functional expression of cloned yeast DNA in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America,* 74**,** 487-491.

147 CONSORTIUM, I. H. G. S. 2004. Finishing the euchromatic sequence of the human genome. *Nature,* 431.

148 RAPLEY, R. 2000. *The nucleic acid protocols handbook*, Springer Science & Business Media.

149 ORBACH, M. J., VOLLRATH, D., DAVIS, R. W. & YANOFSKY, C. 1988. An electrophoretic karyotype of Neurospora crassa. *Molecular and Cellular Biology,* 8**,** 1469-1473.

150 KIM, U. J., BIRREN, B. W., SLEPAK, T., MANCINO, V., BOYSEN, C., KANG, H. L., SIMON, M. I. & SHIZUYA, H. 1996. Construction and characterization of a human bacterial artificial chromosome library. *Genomics,* 34**,** 213-8.

151 DI PALMA, F. 1999. *Analysis and mapping of bovine MHC class I genes. Doctoral thesis, University of Reading, Reading, UK.*

152 VALENZUELA, D. M., MURPHY, A. J., FRENDEWEY, D., GALE, N. W., ECONOMIDES, A. N., AUERBACH, W., POUEYMIROU, W. T., ADAMS, N. C., ROJAS, J., YASENCHAK, J., CHERNOMORSKY, R., BOUCHER, M., ELSASSER, A. L., ESAU, L., ZHENG, J., GRIFFITHS, J. A., WANG, X., SU, H., XUE, Y., DOMINGUEZ, M. G., NOGUERA, I., TORRES, R., MACDONALD, L. E., STEWART, A. F., DECHIARA, T. M. & YANCOPOULOS, G. D. 2003. High-throughput engineering of the mouse genome coupled with high-resolution expression analysis. *Nat Biotechnol,* 21**,** 652-9.

153 SAMBROOK, J., FRITSCH, E. F. & MANIATIS, T. 1989. *Molecular cloning,* Cold spring harbor laboratory press New York.

154 LEE, E.-C., YU, D., DE VELASCO, J. M., TESSAROLLO, L., SWING, D. A., COURT, D. L., JENKINS, N. A. & COPELAND, N. G. 2001. A highly efficient Escherichia coli-based chromosome engineering system adapted for recombinogenic targeting and subcloning of BAC DNA. *Genomics,* 73**,** 56-65.

155 NEFEDOV, M., CARBONE, L., FIELD, M., SCHEIN, J. & DE JONG, P. J. 2011. Isolation of Specific Clones from Nonarrayed BAC Libraries through Homologous Recombination. *Journal of Biomedicine and Biotechnology,* 2011**,** 8.

156 CHORI 2016. BACPAC Resources Center (BPRC). *online*.

157 CAI, L., TAYLOR, J., WING, R. A., GALLAGHER, D., WOO, S.-S. & DAVIS, S. 1995. Construction and characterization of a bovine bacterial artificial chromosome library. *Genomics,* 29**,** 413-425.

158 CHORI Children's Hospital Oakland Research Institute BACPAC Resources Center.

159 ELSIK, C. G., TELLAM, R. L., WORLEY, K. C., GIBBS, R. A., MUZNY, D. M., WEINSTOCK, G. M., ADELSON, D. L., EICHLER, E. E., ELNITSKI, L., GUIGO, R., HAMERNIK, D. L., KAPPES, S. M., LEWIN, H. A., LYNN, D. J., NICHOLAS, F. W., REYMOND, A., RIJNKELS, M., SKOW, L. C., ZDOBNOV, E. M., SCHOOK, L., WOMACK, J., ALIOTO, T., ANTONARAKIS, S. E., ASTASHYN, A., CHAPPLE, C. E., CHEN, H. C., CHRAST, J., CAMARA, F., ERMOLAEVA, O., HENRICHSEN, C. N., HLAVINA, W., KAPUSTIN, Y., KIRYUTIN, B., KITTS, P., KOKOCINSKI, F., LANDRUM, M., MAGLOTT, D., PRUITT, K., SAPOJNIKOV, V., SEARLE, S. M., SOLOVYEV, V., SOUVOROV, A., UCLA, C., WYSS, C., ANZOLA, J. M., GERLACH, D., ELHAIK, E., GRAUR, D., REESE, J. T., EDGAR, R. C., MCEWAN, J. C., PAYNE, G. M., RAISON, J. M., JUNIER, T., KRIVENTSEVA, E. V., EYRAS, E., PLASS, M., DONTHU, R., LARKIN, D. M., REECY, J., YANG, M. Q., CHEN, L., CHENG, Z., CHITKO-MCKOWN, C. G., LIU, G. E., MATUKUMALLI, L. K., SONG, J., ZHU, B., BRADLEY, D. G., BRINKMAN, F. S., LAU, L. P., WHITESIDE, M. D., WALKER, A.,

WHEELER, T. T., CASEY, T., GERMAN, J. B., LEMAY, D. G., MAQBOOL, N. J., MOLENAAR, A. J., SEO, S., STOTHARD, P., BALDWIN, C. L., BAXTER, R., BRINKMEYER-LANGFORD, C. L., BROWN, W. C., CHILDERS, C. P., CONNELLEY, T., ELLIS, S. A., FRITZ, K., GLASS, E. J., HERZIG, C. T., IIVANAINEN, A., LAHMERS, K. K., BENNETT, A. K., DICKENS, C. M., GILBERT, J. G., HAGEN, D. E., SALIH, H., AERTS, J., CAETANO, A. R., et al. 2009. The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science,* 324**,** 522-8.

160 LEZIN, G., KOSAKA, Y., YOST, H. J., KUEHN, M. R. & BRUNELLI, L. 2011. A one-step miniprep for the isolation of plasmid DNA and lambda phage particles. *PLoS One,* 6**,** e23457.

161 BICKHART, D. M., ROSEN, B. D., KOREN, S., SAYRE, B. L., HASTIE, A. R., CHAN, S. & LEE, J. 2017. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. 49**,** 643-650.

162 PALMA, D. 1999. Analysis and mapping of bovine MHC class I genes. *Doctoral thesis,* University of Reading.

163 NHGRI 2004.

164 RHOADS, A. & AU, K. F. 2015. PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics,* 13**,** 278-289.

165 KORLACH, J. 2015. Understanding Accuracy in SMRT® Sequencing. *www.pacb.com/*.

166 LAVER, T., HARRISON, J., O'NEILL, P., MOORE, K., FARBOS, A., PASZKIEWICZ, K. & STUDHOLME, D. 2015. Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomol Detect Quantif,* 3**,** 1-8.

167 NIKU, M., LILJAVIRTA, J., DURKIN, K., SCHRODERUS, E. & IIVANAINEN, A. 2012. The bovine genomic DNA sequence data reveal three IGHV subgroups, only one of which is functionally expressed. *Dev Comp Immunol,* 37**,** 457-61.

168 ZHAO, Y., KACSKOVICS, I., RABBANI, H. & HAMMARSTROM, L. 2003. Physical mapping of the bovine immunoglobulin heavy chain constant region gene locus. *J Biol Chem,* 278**,** 35024-32.

169 HAYES, H. C. & PETIT, E. J. 1993. Mapping of the beta-lactoglobulin gene and of an immunoglobulin M heavy chain-like sequence to homoeologous cattle, sheep, and goat chromosomes. *Mamm Genome,* 4**,** 207-10.

170 HOSSEINI, A., CAMPBELL, G., PROROCIC, M. & AITKEN, R. 2004. Duplicated copies of the bovine JH locus contribute to the Ig repertoire. *Int Immunol,* 16**,** 843-52.

171 KOTI, M., KATAEVA, G. & KAUSHIK, A. K. 2010. Novel atypical nucleotide insertions specifically at VH-DH junction generate exceptionally long CDR3H in cattle antibodies. *Mol Immunol,* 47**,** 2119-28.

172 SHOJAEI, F., SAINI, S. S. & KAUSHIK, A. K. 2003. Unusually long germline DH genes contribute to large sized CDR3H in bovine antibodies. *Mol Immunol,* 40**,** 61-7.

173 ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. 1990. Basic local alignment search tool. *J Mol Biol,* 215.

174 The Bovine Genome Database. [http://www.bovinegenome.org].

175 CHIN, C.-S., ALEXANDER, D. H., MARKS, P., KLAMMER, A. A., DRAKE, J., HEINER, C., CLUM, A., COPELAND, A., HUDDLESTON, J., EICHLER, E. E., TURNER, S. W. & KORLACH, J. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Meth,* 10**,** 563-569.

176 KOREN, S., WALENZ, B. P., BERLIN, K., MILLER, J. R., BERGMAN, N. H. & PHILLIPPY, A. M. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res,* 27**,** 722-36.

177 MARCO-SOLA, S., SAMMETH, M., GUIGO, R. & RIBECA, P. 2012. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat Methods,* 9**,** 1185-8.

178 BANKEVICH, A., NURK, S., ANTIPOV, D., GUREVICH, A. A., DVORKIN, M., KULIKOV, A. S., LESIN, V. M., NIKOLENKO, S. I., PHAM, S., PRJIBELSKI, A. D., PYSHKIN, A. V., SIROTKIN, A. V., VYAHHI, N., TESLER, G., ALEKSEYEV, M. A. & PEVZNER, P. A. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol,* 19**,** 455-77.

179 QUINLAN, A. R. & HALL, I. M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics,* 26**,** 841-2.

180 LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G. & DURBIN, R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25**,** 2078-9.

181 GUREVICH, A., SAVELIEV, V., VYAHHI, N. & TESLER, G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics,* 29**,** 1072-5.

182 RUTHERFORD, K., PARKHILL, J., CROOK, J., HORSNELL, T., RICE, P., RAJANDREAM, M. A. & BARRELL, B. 2000. Artemis: sequence visualization and annotation. *Bioinformatics,* 16**,** 944-5.

183 2016. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res,* 44**,** D7-19.

184 SONNHAMMER, E. L. & DURBIN, R. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene,* 167**,** Gc1-10.

185 LEFRANC, M. P., POMMIÉ, C., RUIZ, M., GIUDICELLI, V., FOULQUIER, E., TRUONG, L., THOUVENIN-CONTET, V. & LEFRANC, G. 2003. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V like domains. *Dev Comp Immunol,* 27.

186 KATOH, K., MISAWA, K., KUMA, K. & MIYATA, T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res,* 30**,** 3059-66.

187 HALL 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl. Acids. Symp,* 41**,** 95-98.

188 TAMURA, K., STECHER, G., PETERSON, D., FILIPSKI, A. & KUMAR, S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol,* 30**,** 2725-9.

189 TAMURA, K. & NEI, M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution,* 10**,** 512-526.

190 PASMAN, Y., MERICO, D. & KAUSHIK, A. K. 2017. Preferential expression of IGHV and IGHD encoding antibodies with exceptionally long CDR3H and a rapid global shift in transcriptome characterizes development of bovine neonatal immunity. *Developmental & Comparative Immunology,* 67**,** 495-507.

191 TAMURA, K. 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Mol Biol Evol,* 9**,** 678-87.

192 LITMAN, G. W., RAST, J. P., SHAMBLOTT, M. J., HAIRE, R. N., HULST, M., ROESS, W., LITMAN, R. T., HINDS-FREY, K. R., ZILCH, A. & AMEMIYA, C. T. 1993. Phylogenetic diversification of immunoglobulin genes and the antibody repertoire. *Mol Biol Evol,* 10**,** 60-72.

193 KUROIWA, Y., KASINATHAN, P., SATHIYASEELAN, T., JIAO, J. A., MATSUSHITA, H., SATHIYASEELAN, J., WU, H., MELLQUIST, J., HAMMITT, M., KOSTER, J., KAMODA, S., TACHIBANA, K., ISHIDA, I. & ROBL, J. M. 2009. Antigen-specific human polyclonal antibodies from hyperimmunized cattle. *Nat Biotechnol,* 27**,** 173-81.

194 DE BONO, B., MADERA, M. & CHOTHIA, C. 2004. VH gene segments in the mouse and human genomes. *J Mol Biol,* 342**,** 131-43.

195 TOMLINSON, I. M., COOK, G. P., WALTER, G., CARTER, N. P., RIETHMAN, H., BULUWELA, L., RABBITTS, T. H. & WINTER, G. 1995. A complete map of the human immunoglobulin VH locus. *Ann N Y Acad Sci,* 764**,** 43-6.

196 ARUN, S. S., BREUER, W. & HERMANNS, W. 1996. Immunohistochemical examination of light-chain expression (lambda/kappa ratio) in canine, feline, equine, bovine and porcine plasma cells. *Zentralbl Veterinarmed A,* 43**,** 573-6.

197 GRAY, W. R., DREYER, W. J. & HOOD, L. 1967. Mechanism of antibody synthesis: size differences between mouse kappa chains. *Science,* 155**,** 465-7.

198 SINKORA, J., REHAKOVA, Z., SAMANKOVA, L., HAVERSON, K., BUTLER, J. E., ZWART, R. & BOERSMA, W. 2001. Characterization of monoclonal antibodies recognizing

immunoglobulin kappa and lambda chains in pigs by flow cytometry. *Vet Immunol Immunopathol,* 80**,** 79-91.

199 MURPHY, T. P., WALPORT M. 2008. Janeway's immunobiology. *New York: Garland Science;,* 7.

200 NISHIKAWA, S. I., KINA, T., GYOTOKU, J. I. & KATSURA, Y. 1984. High frequency of lambda gene activation in bone marrow pre-B cells. *J Exp Med,* 159**,** 617-22.

201 KNOTT, J. 1998. The Primary Antibody Repertoire of k-Deficient Mice is Characterized by Non-Stochastic V l1þ VH Gene Family Pairings and a Higher Degree of Self-Reactivity. *J IMM* 48**,** 65-72.

202 SOLOMON, A. & WEISS, D. T. 1995. Structural and functional properties of human lambda-light-chain variable-region subgroups. *Clin Diagn Lab Immunol,* 2**,** 387-94.

203 GERDES, T. & WABL, M. 2002. Physical map of the mouse lambda light chain and related loci. *Immunogenetics,* 54**,** 62-5.

204 CHATELLIER, J., RAUFFER-BRUYERE, N., VAN REGENMORTEL, M. H., ALTSCHUH, D. & WEISS, E. 1996. Comparative interaction kinetics of two recombinant Fabs and of the corresponding antibodies directed to the coat protein of tobacco mosaic virus. *J Mol Recognit,* 9**,** 39-51.

205 BREZINSCHEK, H. P., FOSTER, S. J., DORNER, T., BREZINSCHEK, R. I. & LIPSKY, P. E. 1998. Pairing of variable heavy and variable kappa chains in individual naive and memory B cells. *J Immunol,* 160**,** 4762-7.

206 DE WILDT, R. M., HOET, R. M., VAN VENROOIJ, W. J., TOMLINSON, I. M. & WINTER, G. 1999. Analysis of heavy and light chain pairings indicates that receptor editing shapes the human antibody repertoire. *J Mol Biol,* 285**,** 895-901.

207 EDWARDS, B. M., BARASH, S. C., MAIN, S. H., CHOI, G. H., MINTER, R., ULLRICH, S., WILLIAMS, E., DU FOU, L., WILTON, J., ALBERT, V. R., RUBEN, S. M. & VAUGHAN, T. J. 2003. The remarkable flexibility of the human antibody repertoire; isolation of over one thousand different antibodies to a single protein, BLyS. *J Mol Biol,* 334**,** 103-18.

208 JAYARAM, N., BHOWMICK, P. & MARTIN, A. C. 2012. Germline VH/VL pairing in antibodies. *Protein Eng Des Sel,* 25**,** 523-9.

209 SAINI, S. S., FARRUGIA, W., RAMSLAND, P. A. & KAUSHIK, A. K. 2003. Bovine IgM antibodies with exceptionally long complementarity-determining region 3 of the heavy chain share unique structural properties conferring restricted VH + Vlambda pairings. *Int Immunol,* 15**,** 845-53.

210 MARCHLER-BAUER, A., DERBYSHIRE, M. K., GONZALES, N. R., LU, S., CHITSAZ, F., GEER, L. Y., GEER, R. C., HE, J., GWADZ, M., HURWITZ, D. I., LANCZYCKI, C. J., LU, F., MARCHLER, G. H., SONG, J. S., THANKI, N., WANG, Z., YAMASHITA, R. A.,

ZHANG, D., ZHENG, C. & BRYANT, S. H. 2015. CDD: NCBI's conserved domain database. *Nucleic Acids Res,* 43**,** D222-6.

211 ALLAN, A. 2015. The Phenotypic and Functional Characterisation of Cattle Natural Killer Cells. *RVC London thesis*.

212 JOHNSON, G., NOUR, A. A., NOLAN, T., HUGGETT, J. & BUSTIN, S. 2014. Minimum information necessary for quantitative real-time PCR experiments. *Methods Mol Biol,* 1160**,** 5-17.

213 PFAFFL, M. W. 2001. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res,* 29**,** e45.

214 SCHWARTZ, J. C. & MURTAUGH, M. P. 2014. Characterization of a polymorphic IGLV gene in pigs (Sus scrofa). *Immunogenetics,* 66**,** 507-11.

215 LANHAM, G. R., BOLLUM, F. J. & STASS, S. A. 1986. Detection of terminal deoxynucleotidyl transferase in acute leukemias using monoclonal antibodies directed against native and denatured sites. *Am J Clin Pathol,* 86**,** 88-91.

216 BRONSVOORT, B. M., PARIDA, S., HANDEL, I., MCFARLAND, S., FLEMING, L., HAMBLIN, P. & KOCK, R. 2008. Serological survey for foot-and-mouth disease virus in wildlife in eastern Africa and estimation of test parameters of a nonstructural protein enzyme-linked immunosorbent assay for buffalo. *Clin Vaccine Immunol,* 15.

217 XU, J. L. & DAVIS, M. M. 2000. Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. *Immunity,* 13**,** 37-45.

218 REDDY, S. T., GE, X., MIKLOS, A. E., HUGHES, R. A., KANG, S. H., HOI, K. H., CHRYSOSTOMOU, C., HUNICKE-SMITH, S. P., IVERSON, B. L., TUCKER, P. W., ELLINGTON, A. D. & GEORGIOU, G. 2010. Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. *Nat Biotechnol,* 28**,** 965-9.

219 HAYDON, D. T., BASTOS, A. D., KNOWLES, N. J. & SAMUEL, A. R. 2001. Evidence for positive selection in foot-and-mouth disease virus capsid genes from field isolates. *Genetics,* 157**,** 7-15.

220 GRANT, C. F., CARR, B. V., SINGANALLUR, N. B., MORRIS, J., GUBBINS, S., HUDELET, P., ILOTT, M., CHARREYRE, C., VOSLOO, W. & CHARLESTON, B. 2016. The B-cell response to foot-and-mouth-disease virus in cattle following vaccination and live-virus challenge. *J Gen Virol,* 97**,** 2201-9.

221 MAGOC, T. & SALZBERG, S. L. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics,* 27**,** 2957-63.

222 EDGAR, R. C. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics,* 26**,** 2460-1.

223 SANGER, F., AIR, G. M., BARRELL, B. G., BROWN, N. L., COULSON, A. R., FIDDES, C. A., HUTCHISON, C. A., SLOCOMBE, P. M. & SMITH, M. 1977. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature,* 265**,** 687-95.

224 BUSHNELL 2014. BBMap: a fast, accurate, splice-aware aligner. *http://1ofdmq2n8tc36m6i46scovo2e.wpengine.netdna-cdn.com/wp-content/uploads/2013/11/BB_User-Meeting-2014-poster-FINAL.pdf*.

225 KROLAK-SCHWEDT, S. & ECKES, T. 1992. A Graph Theoretic Criterion for Determining the Number of Clusters in a Data Set. *Multivariate Behav Res,* 27**,** 541-65.

226 RUDIS, B. 2013. Introduction to the streamgraph htmlwidgtet R Package. *https://hrbrmstr.github.io/streamgraph/*.

227 KATOH, K., MISAWA, K., KUMA, K. & MIYATA, T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res,* 30.

228 KUMAR, S., STECHER, G. & TAMURA, K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol,* 33**,** 1870-4.

229 JONES, D. T., TAYLOR, W. R. & THORNTON, J. M. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci,* 8**,** 275-82.

230 GRANT, C. 2013. Investigating primary and secondary B cell responses in cattle after immunisation with existing and novel vaccines. *Thesis University of Oxford*.

231 WILLIAMS, G. T., JOLLY, C. J., KOHLER, J. & NEUBERGER, M. S. 2000. The contribution of somatic hypermutation to the diversity of serum immunoglobulin: dramatic increase with age. *Immunity,* 13**,** 409-17.

232 SAINI, S. S., ALLORE, B., JACOBS, R. M. & KAUSHIK, A. 1999. Exceptionally long CDR3H region with multiple cysteine residues in functional bovine IgM antibodies. *Eur J Immunol,* 29**,** 2420-6.

233 WHELAN, S. & GOLDMAN, N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol,* 18**,** 691-9.

234 WALTHER, S., CZERNY, C. P. & DIESTERBECK, U. S. 2013. Exceptionally long CDR3H are not isotype restricted in bovine immunoglobulins. *PLoS One,* 8**,** e64234.

235 EKIERT, D. C., KASHYAP, A. K., STEEL, J., RUBRUM, A., BHABHA, G., KHAYAT, R., LEE, J. H., DILLON, M. A., O'NEIL, R. E., FAYNBOYM, A. M., HOROWITZ, M., HOROWITZ, L., WARD, A. B., PALESE, P., WEBBY, R., LERNER, R. A., BHATT, R. R. & WILSON, I. A. 2012. Cross-neutralization of influenza A viruses mediated by a single antibody loop. *Nature,* 489**,** 526-32.

236 WU, L., OFICJALSKA, K., LAMBERT, M., FENNELL, B. J., DARMANIN-SHEEHAN, A., NI SHUILLEABHAIN, D., AUTIN, B., CUMMINS, E., TCHISTIAKOVA, L., BLOOM, L., PAULSEN, J., GILL, D., CUNNINGHAM, O. & FINLAY, W. J. 2012. Fundamental

characteristics of the immunoglobulin VH repertoire of chickens in comparison with those of humans, mice, and camelids. *J Immunol,* 188**,** 322-33.

237 ZEMLIN, M., KLINGER, M., LINK, J., ZEMLIN, C., BAUER, K., ENGLER, J. A., SCHROEDER, H. W., JR. & KIRKHAM, P. M. 2003. Expressed murine and human CDR-H3 intervals of equal length exhibit distinct repertoires that differ in their amino acid composition and predicted range of structures. *J Mol Biol,* 334**,** 733-49.

238 CHENG, G., GAO, Y., WANG, T., SUN, Y., WEI, Z., LI, L., REN, L., GUO, Y., HU, X., LU, Y., WANG, X., LIU, G., ZHANG, C., YU, J., PAN-HAMMARSTROM, Q., HAMMARSTROM, L., WU, X., LI, N. & ZHAO, Y. 2013. Extensive diversification of IgH subclass-encoding genes and IgM subclass switching in crocodilians. *Nat Commun,* 4**,** 1337.

239 LEE, C. E., GAETA, B., MALMING, H. R., BAIN, M. E., SEWELL, W. A. & COLLINS, A. M. 2006. Reconsidering the human immunoglobulin heavy-chain locus: 1. An evaluation of the expressed human IGHD gene repertoire. *Immunogenetics,* 57**,** 917-25.

240 VOSLOO, W., BOSHOFF, K., DWARKA, R. & BASTOS, A. 2002. The possible role that buffalo played in the recent outbreaks of foot-and-mouth disease in South Africa. *Ann NY Acad Sci,* 969.

241 WRAMMERT, J., KOUTSONANOS, D., LI, G. M., EDUPUGANTI, S., SUI, J., MORRISSEY, M., MCCAUSLAND, M., SKOUNTZOU, I., HORNIG, M., LIPKIN, W. I., MEHTA, A., RAZAVI, B., DEL RIO, C., ZHENG, N. Y., LEE, J. H., HUANG, M., ALI, Z., KAUR, K., ANDREWS, S., AMARA, R. R., WANG, Y., DAS, S. R., O'DONNELL, C. D., YEWDELL, J. W., SUBBARAO, K., MARASCO, W. A., MULLIGAN, M. J., COMPANS, R., AHMED, R. & WILSON, P. C. 2011. Broadly cross-reactive antibodies dominate the human B cell response against 2009 pandemic H1N1 influenza virus infection. *J Exp Med,* 208**,** 181-93.

242 ZHANG, M. Y., YUAN, T., LI, J., ROSA BORGES, A., WATKINS, J. D., GUENAGA, J., YANG, Z., WANG, Y., WILSON, R., LI, Y., POLONIS, V. R., PINCUS, S. H., RUPRECHT, R. M. & DIMITROV, D. S. 2012. Identification and characterization of a broadly cross-reactive HIV-1 human monoclonal antibody that binds to both gp120 and gp41. *PLoS One,* 7**,** e44241.