

**IMMUNOGLOBULIN ANNOTATION OPTIMIZING AND BENCHMARKING;
NOVEL GERMLINE ALLELE DISCOVERY IN SELECTED AFRICAN BOVINE
BREEDS**

LANDI MICHAEL KOFIA

**A thesis submitted in partial fulfillment of the requirements for the Degree of Master of
Science in Bioinformatics of Pwani University**

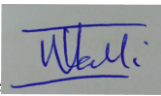
SEPTEMBER, 2020

DECLARATION**Declaration by the student**

This thesis is my original work and has not been presented in any other University or any other Award.

LANDI MICHAEL KOFIA

REG NO: SG30/PU/36148/18

Signature 

Date: 1st March, 2021

Declaration by the supervisors

We confirm that the work reported in this thesis was carried out by the candidate under our supervision.

DR. SONAL HENSON - International Livestock Research Institute (ILRI)

Signature 


Date: 1st March, 2021

DR. JEAN-BAKA DOMELEVO ENTFELLNER - BecA-ILRI Hub

Signature 

Date: 1st March, 2021

DR. SAMWEL ODIWOUR - Pwani University

Signature 

Date: 1st March, 2021

DEDICATION

I wish to dedicate this work to my family, the EANBiT project, and my fellow students whose moral support has remained unrelenting.

ACKNOWLEDGMENT

I gratefully acknowledge the financial support of EANBiT (Eastern Africa Network of Bioinformatics Training) funded by the National Institute of Health (NIH) and Initiative to Develop African Research Leaders (IdeAL).

Efforts taken in this project would not have been achievable without the kind support and scientific guidance of my supervisors Dr Sonal Henson, Dr Jean-Baka Domelevo and Dr Samwel Odiwour to whom I want to extend my sincere thanks.

I want to bestow my gratitude to the BecA- ILRI hub and International Center of Insect Physiology and Ecology (ICIPE) for making it possible for me to work on my project. I desire to express my particular gratitude to our principal project investigator Dr Dan Masiga and project manager Karen Wambui for their continuous support during my fellowship.

To Pwani University family and bioinformatics classmates, I am thankful for your motivation. Most importantly, Almighty God for his favour, grace, and strength.

TABLE OF CONTENTS

DECLARATION.....	ii
DEDICATION.....	iii
ACKNOWLEDGMENT.....	iv
ABBREVIATIONS AND ACRONYMS	viii
LIST OF FIGURES	x
LIST OF TABLES.....	xii
ABSTRACT.....	xiii
CHAPTER 1: INTRODUCTION	1
1.1. Background	1
1.2 Problem statement	3
1.3. Justification	4
1.4. Objectives.....	5
1.4.1. Overall objective.....	5
1.4.2. Specific objectives	5
CHAPTER 2: LITERATURE REVIEW	6
2.1. Antibody Structure	6
2.2. Hemopoiesis	7
2.3. General genetics of immunoglobulins.....	8
2.4. Genetic diversity generation in immunoglobulins	9
2.5. Bovine immunoglobulin genetics.....	10
2.6. Antibody novel germline allele discovery and annotation.....	11

2.7. Immunoglobulin annotation tools	11
2.8. Germline allele discovery tools.....	12
CHAPTER 3: MATERIALS AND METHODS	14
3.1. Benchmarking and optimizing annotation tools.....	14
3.1.1. Simulated bovine datasets generation.....	14
3.1.2. Adaptive Immune Receptor Repertoire sequencing (AIRR seq) annotation.....	15
3.1.3. Modifying germline JH and VH genes	17
3.2. Germline allele discovery.....	18
3.2.1. Germline allele discovery using IgDiscover.....	18
3.2.2. Germline gene discovery using TIgGER.....	19
3.2.3. Haplotype networks	19
CHAPTER 4: RESULTS.....	20
4.1. IgSimulator outputs.....	20
4.2. Benchmarking annotation of simulated bovine immunoglobulin sequences.....	20
4.2.1. Misidentification frequencies.....	20
4.2.2. Distribution of correct and incorrect VH gene annotation.....	23
4.3. Germline allele discovery.....	27
4.3.1. Novel alleles discovered by IgDiscover	27
4.3.2. Novel bovine alleles discovered by TIgGER.....	34
4.3.3. Distribution of distances of novel alleles discovered by IgDiscover and TIgGER.....	35
CHAPTER 5: DISCUSSION, CONCLUSIONS AND RECOMMENDATIONS	41

5.1 DISCUSSION	41
5.2 CONCLUSIONS AND RECOMMENDATIONS	44
REFERENCES	45

ABBREVIATIONS AND ACRONYMS

AIRR-Seq	Adaptive immune receptor repertoire sequence
BCR	B cell receptor
CDR	Complementarity determining region
CDR H3	Complementarity determining region of third heavy chain
DH	Diversity region of the heavy chain (DH gene)
DNA	Deoxyribonucleic acid
FMD	Foot and mouth disease
FWR	Framework region
HSC	Hematopoietic stem cells
IACUC	Institutional Animal Care and Use Committee
IGH	Immunoglobulin heavy chain
IGHD	Immunoglobulin heavy chain diversity
IGHJ	Immunoglobulin heavy chain joining
IGHV	Immunoglobulin heavy chain variable
IgM	Immunoglobulin isotype M
ILRI	International Livestock Research Institute
IMGT	The international ImMunoGeneTics information system database
IMPre	Immune germline prediction
JH	Joining region of the heavy chain (JH gene)
PBMC	Peripheral blood mononuclear cells
PCR	Polymerase chain reaction
RSS	Recognition signal sequences
SHM	Somatic hypermutation
SNP	Single nucleotide polymorphism

TdT	Terminal deoxynucleotidyl transferase
TIgGER	Tool for immunoglobulin genotype elucidation via rep-seq
VH	Variable region of the heavy chain (VH gene)

LIST OF FIGURES

Figure 1: Y-shaped structure of antibody.....	6
Figure 2: B-cell maturation and generation steps.....	8
Figure 3: A five-step IgSimulator workflow of simulating antibody repertoire.....	15
Figure 4: Benchmarking steps designed to compare the three annotation tool.....	17
Figure 5: Bar plots for the predicted percentage of mishits at the gene level.....	21
Figure 6: Bar plot for the percentage of unassigned genes of IgBlast and IMGT/HighV-QUEST.....	22
Figure 7: Bar plots for the predicted percentage of mishits after modifying the germline VH and JH database.....	23
Figure 8: Distribution of hits and mishits for all tool's VH gene annotation.....	26
Figure 9: A line graph showing individualized database VH alleles buildup for each bovine breed.....	27
Figure 10: A line graph shows individualized database VH alleles built up of each bovine breeds after germline filtering.....	28
Figure 11: Haplotype networks of African novel alleles discovered by IgDiscover.....	29
Figure 12: Haplotype networks of African novel alleles discovered by TIgGER.....	35
Figure 13: Histogram showing pairwise Hamming distances of novel alleles discovered by IgDiscover	37
Figure 14: Histogram showing pairwise Hamming distances of novel alleles discovered by TIgGER	37

Figure 15: Boxplot showing pairwise Hamming distances of novel alleles from African breeds and Friesian breed.....	38
Figure 16: A boxplot showing Hamming distances of African novel alleles (per breed) versus Friesian novel allele both discovered from IgDiscover and TIGER.....	40

LIST OF TABLES

Table 1: Shows the ranking of novel alleles of Boran breed.....	30
Table 2: Shows novel allele ranking of Ndama breed.....	32
Table 3: Shows novel allele ranking of Ankole breed.....	33
Table 4: Comparison of standard deviations and mean values of African novel alleles and Friesian novel alleles.....	36
Table 5: Comparison of standard deviations and mean values within African novel alleles and Friesian novel alleles.....	39

ABSTRACT

Antibodies are critical molecules of the adaptive immune response of vertebrates. For an animal to neutralize the pathogens it will encounter in its lifetime, the diversity of antibodies it will need to produce is enormous. Vertebrates achieve this through genetic recombination of immunoglobulin genes and post-somatic transcription mutations applied to so-called “germline” alleles. Bovine antibodies have distinct immunogenetics. Available annotation tools are human-centric, and therefore not optimized to annotate bovine immunoglobulin sequences. The international ImMunoGeneTics information system (IMGT) database is a global database reference in immunogenetics and immunoinformatics. Some information of germline alleles from the IMGT database is not up to date for most species and germline gene databases, which are not complete. Studies of germline alleles identification of cattle through novel allele discovery are necessary to complete germline gene databases of species. These discoveries are one step towards understanding the immunological complexity of these species. Using simulated bovine datasets, benchmarking the performance of three annotation tools IgBlast, IMGT/HighV-QUEST, and MiXCR was done based on frequencies and distribution of the correctly and wrongly identified antibody sequences. Two methods of germline allele discovery IgDiscover and TiGGER are evaluated to determine their suitability for bovine germline allele discovery. Immunoglobulin M sequences of three African bovine breeds, Ndama, Ankole, and Boran, were used in this analysis while the Friesian cattle breed used as a control. For annotation of VH gene, IMGT/HighV-QUEST and IgBlast yielded a more accurate annotation with a 4% error rate, compared to MiXCR, which had a 13% error rate. MiXCR annotated JH genes with an error rate of 15% compared to IgBlast and IMGT/HighV-QUEST, which had an error rate of 40% and 43%, respectively. IgDiscover identified 18 novel alleles in Boran, 6 novel alleles in Ndama, and 3 novel alleles in Ankole. TiGGER, on the other hand, identified 7 novel alleles in Boran, 18 novel alleles in Ndama, and 1 novel allele in Ankole. Using pairwise Hamming distances of these novel alleles, it was observed that African

novel germline alleles are more diverse compared to the Friesian novel germline alleles. This discovery of novel alleles shows that there is a need for further studies in characterizing the immune system of African breeds.

CHAPTER 1: INTRODUCTION

1.1. Background

Immunoglobulins are vital molecules of the adaptive immune response in vertebrates. They are composed of two identical copies of a heavy chain and two exact copies of a light chain, each consisting of a variable and constant region. The variable domain is directly involved in the binding of antigens and is composed of framework regions (FWR) and complementarity determining regions (CDR). CDR are segments of the variable domain that bind to the antigen, and each heavy chain and light chain contains three CDR. The third CDR of the heavy chain (CDR H3) plays a dominant role in antigen binding.

In most cases, species have CDR H3 that forms a simple loop of about 8 to 16 amino acids, primarily in humans. Bovine antibodies have ultra-long CDR H3 occupying around 10% of the antibody repertoire. They have a difference in the distribution of their lengths, ranging from 40 to 70 amino acids (Haakenson et al., 2018).

To neutralize pathogens, the immune system has developed mechanisms to produce a large pool of immunoglobulins. One of the mechanisms is that multiple genes jointly encode the antigen-binding region. The variable heavy chain domain is encoded by the variable (V), diversity (D), and joining (J) genes. In contrast, the light chain domain is encoded by V and J genes (Chaudhary & Wesemann, 2018). These genes at the beginning are separated in the germline genome and are later contiguously joined by a process called V (D) J recombination. V (D) J recombination is augmented by the process called somatic hypermutation resulting in an enormously diverse repertoire of immunoglobulins. Antibodies with higher specific affinity to particular antigens are produced by a process called affinity maturation. Further diversity is also achieved by V (D) J recombination through nucleotide insertion or deletion processes at the junction between V and D, and J or V and J gene segments. Two different nucleotide additions result in this kind of junction diversity, i.e., palindromic P nucleotide addition

followed by N nucleotide insertions by terminal deoxynucleotidyl transferase (TdT) (Ye, Ma, Madden, & Ostell, 2013).

Currently, advances have been made in repertoire sequence analysis for both B-cells and T-cells. These strategies have allowed comprehensive repertoire sequencing analysis to be applied to different research areas such as; understanding the diversity of T & B cell repertoire generated upon infection, production of monoclonal antibodies and investigation of immune surveillance in specific diseases (Ye et al., 2013 ; Jackson et al., 2014). Characterization of germline genes is vital in understanding and interpreting repertoire sequence data. IMGT (The international ImMunoGeneTics information system) is a public repository that collates germline genes by species (Neuberger, 1992). Some information of germline alleles from IMGT is not present for most species, as well as germline gene databases, are not complete, making studies of the repertoire partially informative. For example, according to IMGT, the bovine IgH locus located in chromosome 21 contains 25 functional alleles derived from 13 IGHV genes, 21 IGHD genes with 21 functional alleles, and 3 IGHJ genes with four functional alleles all derived from European *Bos taurus* breeds. In contrast to humans, IMGT is updated and contains 306 functional alleles from 56 IGHV genes, 23 IGHD genes with 30 functional alleles, and 6 IGHJ genes with 13 functional alleles (these values were obtained from IMGT database on July 1st, 2020). The lack of allelic diversity captured for bovine germline alleles in the IMGT database brings to light a need for characterizing African bovine breeds.

One step towards identifying novel germline alleles is by using germline gene discovery tools on IgM immunoglobulin sequences. However, this is problematic because the analysis is affected by factors related to somatic hypermutation and sequence errors. It is necessary to validate different methods of germline gene discovery methods to assist in characterizing germline genes of species that have been unarchived in the database (Berens, Wylie, & Lopez, 1997).

Another challenge is that most VH gene annotation tools, which have been designed for human or mouse data, have not been optimized for bovine immunoglobulin sequences that are unique in their presentation of ultra-long antibodies. IGHV1-7 gene, which is responsible for the formation of ultra-long CDR H3, is not always annotated correctly. Ultra-long CDR H3s are thought to provide additional antigenic diversity to compensate to the reduced number of germline genes in cows. Not much is known about the function of ultra-long antibodies; however, one study found broadly neutralizing and potent bovine ultra-long antibodies were elicited against an HIV antigen (Sok et al., 2018).

1.2 Problem statement

The variable domain of heavy chains contains three CDR regions and four FWR regions. The third CDR region is primarily responsible for the binding of the antigen. In bovine, studies have shown ultra-long antibodies CDR H3 sequence use the IGHV1-7 gene. IGHV1-7 gene preferentially recombines with the IGHD8-2 gene in ultra-long antibodies. Shorter CDR H3 sequences appear to use IGHV1-10 preferentially (Deiss et al., 2019).

Annotation of IGHV1-7 gene responsible for ultra-long CDR H3 antibodies is not correctly identified in bovine immunoglobulin sequences. IGHV1-7 gene has a CTTVHQ motif towards the 3' end. This motif is thought to be necessary for stabilizing the ultra-long CDR H3 regions. Tests were run using sequences that are IGHV1-7 gene because they contained a CTTVHQ motif at the 5' end of its CDR H3. These sequences were incorrectly annotated before, and the purpose was to find out if the annotation tools could rectify this mistake. It is seen that the three available annotation tools do not provide a correct annotation for this gene. Instead, the gene call is mistaken for other genes. Also, annotations of bovine immunoglobulin sequences have not been benchmarked based on accuracy. These hurdles in annotation are brought about by not many studies done in annotating bovine immunoglobulin sequences and the fact that these annotation tools are human-centred.

Germline gene discovery affects germline gene alignments, which are very critical for most downstream analyses carried out on B-cell receptor (BCR) data, such as identifying somatic mutations, clonal grouping, and diversity approximations, among others. Also, in applied approaches, germline gene discovery informs the ability to design amplification primers for the high-throughput cloning of heavy chains and light chains to segregate immunoglobulins of potential therapeutic value. Several tools have been developed for aligning immunoglobulin sequences and predicting germline gene sequences (Gadala-Maria, Yaari, Uduman, & Kleinstein, 2015; Corcoran et al., 2016). These have been tested on IgM immunoglobulin sequences of mammalian species like humans, mice, and monkeys but not on bovine species.

1.3 Justification

Identification of germline genes is worth doing not purely for discovering mechanisms of generating immunoglobulins but also part of characterizing germline genes in livestock. Most of the tools developed for this kind of analysis have been mainly tested on human immunoglobulin RNA sequences generated through next-generation sequencing platforms. This represents an outstanding analysis in characterizing species whose germline genes database is incomplete. For example, in the current study, African bovine breeds.

Germline discovery is seen to be a success in human immunoglobulin sequences (Corcoran et al., 2016). Unfortunately, tools are underdeveloped to address the same analysis on bovine immunoglobulin, especially of the African breeds. This project sought to predict bovine germline genes employing commonly used germline allele discovery tools. In the process, some parameters of these tools were optimized for bovine immunoglobulin sequences. Through this, characterizing germline genes of African bovine breeds is achieved, which is consequential in understanding and interpreting repertoire sequence data.

It is, therefore, essential to evaluate available software for germline gene discovery to advice on the most effective tool in terms of efficiency and accuracy. Since these tools have not been

used on bovine immunoglobulin sequences, testing the suitability of various tools on this data set will also guide the most acceptable tool for the bovine data set.

1.4 Objectives

1.4.1 Overall objective

To benchmark and optimize annotation of bovine immunoglobulin sequences and to evaluate discovery and diversity of novel germline alleles in selected African bovine immunoglobulin sequences.

1.4.2 Specific objectives

1. To assess immunoinformatics tools i.e. IMGT/HighV-QUEST, IgBlast and MiXCT for annotation of bovine immunoglobulin sequences.
2. To predict novel germline alleles of three African cattle breeds using IgDiscover and TiGER
3. To determine the genetic diversity of novel germline alleles using pairwise Hamming distances.

CHAPTER 2: LITERATURE REVIEW

2.1. Antibody Structure

Immunoglobulins (Ig), are proteins presented by the body's immune system when it encounters molecules foreign to the body (antigens). Each immunoglobulin consists of four polypeptide chains; two identical heavy chains and two identical light chains joined by a disulfide bond to form a 'Y' shaped molecule (Figure 1). The heavy chain consists of more complex polypeptides weighing about 50 kDa or more, whereas the light chain possesses polypeptides of approximately 22 kDa. In mammals, there are five Ig heavy chains denoted as; α , δ , ϵ , γ , and μ that define classes of immunoglobulins: IgA, IgD, IgE, IgG, and IgM respectively. There are two varieties of light chains: kappa (κ) and lambda (λ) (Cohen & Cohen, 1962).

Immunoglobulins are made up of the variable region and constant region. The region with high diversity is called the variable region. The region that is conserved with a permanent structure is considered the constant region.

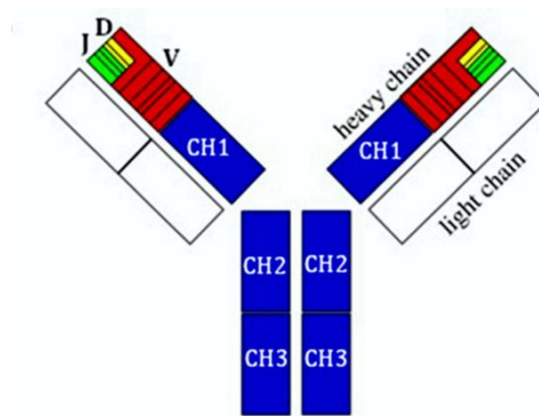


Figure 1. This is a Y-shaped antibody structure showing variable and constant region. The portion in red is the variable gene (V); green is the joining gene (J); yellow is the diversity gene (D), and blue is the constant region of the heavy chain locus. Image modified from: Yaari & Kleinstein, 2015.

2.2 Hemopoiesis

Hemopoiesis or hematopoiesis is equally known as blood cell formation. Blood cells are divided into three groups: leukocytes (white blood cells), erythrocytes (red blood cells), and thrombocytes (platelets). The leukocytes are further sub-divided into granulocytes and monocytes. Blood cells originate from the bone marrow. In an adult human, the bone marrow produces 60-70 per cent of leukocytes (that is the granulocytes), all erythrocytes, and all platelets. The lymphatic tissues have 20-30 per cent of leukocytes.

Leukocytes are immune cells involved in protecting the body against foreign pathogens. They are produced from multipotent cells in the bone marrow known as hematopoietic stem cells (HSC). HSC can be differentiated into two lineages, lymphoid and myeloid. Lymphoid lineage cells develop during lymphopoiesis and include B-cells, T-cells, natural killer cells, and dendritic cells. Many of these cells are short-lived, and immune system homeostasis requires differentiation and self-renewal (Brand & Livesey, 2011).

The B-cells mature in the bone marrow (figure 2). The first stage of B-cell maturation is the evaluation of its functionality which is antigen-independent. This occurs by selecting B-cells with typical functional receptors, a process called positive selection. The body, however, possesses a mechanism to reduce autoimmunity by negative selection. The method used to eliminate self-reacting B-cells is through apoptosis, editing, or modification of the receptor so that they no longer self-react. Selected B-cells are transported to the spleen, where they undergo the final stage of maturation. They become naive mature B-cells (Jagannathan-Bogdan & Zon, 2013). Mature naïve B-cells surfaced on IgM antibodies migrate to the secondary lymphoid tissue. At the lymphoid tissues, B-cells then interact with an exogenous antigen and or T helper cells. The B-cells are activated, and class-switching occurs at this stage. Class-switching is a mechanism that changes a B-cell's production of an antibody from one type to another, for example, from isotype IgM to the isotype IgG. Once B-cells are activated, they

participate in a two-step differentiation process that yields both short-lived plasma cells for immediate response and long-lived plasma cells and memory cells for continuous protection.

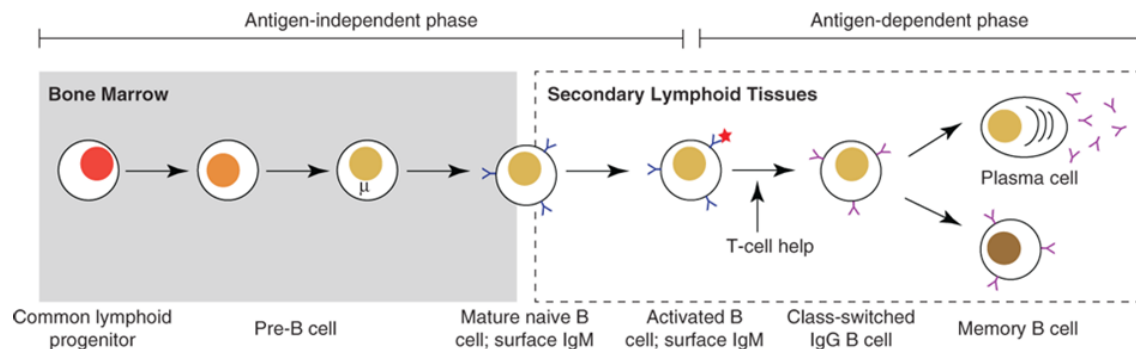


Figure 2. This illustrates the B-cell maturation process in two steps i.e. antigen-independent phase and antigen-dependent phase. Source of the diagram – W. Levinson, P. Chin-Hong, E.A. Joyce, J. Nussbaum, B. Schwartz.

2.3 General genetics of immunoglobulins

The immunoglobulin light and heavy chain are encoded by three different gene families, each one on a distinct chromosome - one for the heavy chain and the other for the light chain types. Each of these gene families has multiple V (variable) region genes and one or more C (constant) region genes.

The heavy chain has many V region genes; each composed of leader region and V exon and three to four C chains depending on various species. In addition to J genes, the heavy chain also contains D (diversity) genes. The variable domain of the heavy chain consists of two regions called the complementarity determining regions (CDR) and framework regions (FWR) (Ichihara, Matsuoka & Kurosawa, 1988).

The light chain is made up of either kappa or the lambda chains. The lambda light chain is made up of 4 C genes, and for each subtype, approximately 30 V region genes. Each of the V region genes has two exons, L that codes for leader region and V that encodes for most of the variable regions. Upstream of the C genes, is an additional exon called J (joining). Introns separate these exons L, V, and J.

The kappa light chain gene family has only one C gene with multiple V region genes, each of which has an L exon and a V exon. There are several J genes in between V and C genes (Collins & Watson, 2018).

2.4 Genetic diversity generation in immunoglobulins

There are three primary ways of generating antibody diversity; V (D) J rearrangements, somatic hypermutation (SHM), and affinity maturation.

As the B-cells differentiate to mature B cells, processes occur in the heavy chains of immunoglobulins by which various gene segments are rearranged before exposure to the antigen. Two rearrangement processes take place. First, one of the D genes is brought next to one of the J genes, and then one of the V genes is carried next to the rearranged DJ region resulting in the VDJ recombination event. These processes occur by two recombination events involving the removal of introns. As with the light chain, the selection of the heavy chain V gene is not entirely arbitrary, but ultimately, all of the V genes stand a chance to be included in some mature antibody.

Unique flanking sequences referred to as recombination signal sequences (RSS), which flank the V, D and J exons, also undergo recombination. Each RSS consists of a nonamer and a heptamer that are separated by either 12bp or 23bp. The 12bp and 23bp spacers correspond to one or two turns of the DNA helix. In heavy chains, there is a single turn signal on each side of the D gene, a two-turn signal downstream of the V gene and a two-turn signal upstream of the J gene. This ensures the occurrence of the correct recombination event. The recombination event is catalyzed by two proteins, Rag-1 and Rag-2 (Sodofsky, 2001).

The second way of antibody diversification is through SHM to generate high-affinity antigen-binding sites. SHM involves the mutation process of the deamination of cytosine to uracil in DNA, affecting the variable regions. Unlike germline mutations, SHM affects only the organism's immune system, and its effects cannot be passed to the organism's offspring (Oprea, M. 1999).

Affinity maturation is the process through which activated B-cells produce immunoglobulins that demonstrate an increased affinity for a specific antigen. With constant exposure to the same antigen, antibodies with more profound affinity will be produced. This process is a result of SHM and occurs on the surface of the immunoglobulins (Victora & Nussenzweig, 2012).

2.5 Bovine immunoglobulin genetics

The structure, occurrence, and function of antibodies in bovine are consistent with the general patterns described in more extensive studies for species like rabbits, humans, guinea pigs, and mice. The germline genetic component of the bovine immunoglobulin repertoire is limited compared to humans, mice, and other species. The heavy chain locus of cattle has 12 functional V genes, 23 D genes, and 4 J genes segments, while the humans have 36-49 V, 23 D, and 6 J genes. This difference has shown that bovine has much lesser combination diversity in theory (Haakenson et al., 2018).

A subset of bovine antibodies contains ultra-long CDR H3. It has been shown that bovine ultra-long antibodies preferentially use a given heavy chain V gene segment (IGHV1-7) and a specific heavy chain D gene segment (IGHD8-2) (Shojaei, Saini & Kaushik, 2003). Those are longer than the other V and D gene segments. Moreover, the ultra-long heavy chain appears to preferentially pair with the lambda light chain (Butler, 1998).

The IGHV1-7 and IGHD8-2 genes mentioned earlier, facilitate the formation of ultra-long CDR H3 in cattle antibodies. Alignment of IGHV1-7 with other IGHV genes has disclosed eight-nucleotide duplication "TACTACTG" at its 3' end. Compared to other IGHV genes used in bovine, IGHV1-7 has low amino acid variability (Deiss et al., 2019). Therefore, the diversity of ultra-long cow antibodies is heavily concentrated in the CDR H3, which is mainly encoded by the IGHD8-2 gene. This region encodes most of the ultra-long CDR H3 after the motif "CTTVHQ". This motif helps to stabilize the ultra-long CDR3 H3 segment (Wang et al., 2013).

2.6 Antibody novel germline allele discovery and annotation

V (D) J recombination events constitute the first step towards antibody repertoire diversity. During an immune response, B-cells that are activated by the binding of their matching antigens undergo clonal expansion followed by somatic hypermutations (SHMs). These mutated immunoglobulins are selected depending on binding strength to the antigen resulting in higher affinity antibodies – a process called affinity maturation.

Currently, there are, for example, documented functional alleles in the IMGT database of diverse species. Studies have shown species often carry novel alleles that have not yet been characterized in the IMGT database (Kirik, Greiff, Levander & Ohlin, 2017). These unidentified novel alleles can be problematic for immunoglobulin repertoire analysis because single nucleotide polymorphisms (SNPs) between novel alleles and the nearest IMGT alleles will be counted as SHMs instead (Wendel, He, Crompton, Pierce & Jiang, 2017). However, there are tools to predict the existence of these novel alleles explained in this project.

Annotation of the antibody repertoire means identifying the complementarity determining and framework regions. Some work has been done so far in characterizing European *Bos taurus* breeds and deposited on the IMGT. This project optimized the annotation of bovine immunoglobulin and discovered novel alleles in African species since there are novel alleles on its germline because of its great diversity.

2.7 Immunoglobulin annotation tools

Commonly used annotation tools like IMGT/HighV-QUEST, MiXCR, and IgBlast are used for annotation of immunoglobulin sequences. IMGT is a reference database for immunogenetics and immunoinformatics. There are different tools on this web tool - one being IMGT/HighV-QUEST (sequence alignment for immunoglobulins and T-cell receptor). IMGT/HighV-QUEST is used for annotations of immunoglobulins. A user uploads sequences in fasta file on the web tool, specifying the organism and locus. Once uploaded, alignments and annotations are done. The output obtained from the website is a compressed file that has

various annotation files. Of interest are the summary file, amino acid sequence file, and nucleotide sequence file. In these files, the user has information on annotations of the various regions, the functionality of the sequences, and sequences for the annotated regions.

The process of MiXCR is made up of three steps; align reads to the reference database for which IMGT libraries have to be installed initially. Next reads are assembled into clonotypes (identical sequence reads clustered/grouped as a clone) using alignment from the first step, and lastly exporting the output to human-readable files for analysis.

IgBlast was developed at the National Center for Biotechnology Information (NCBI). It allows users to view calls of germline V, D, and J genes; information of junction rearrangement; characterization of immunoglobulin V region domain framework and complementarity determining regions. It can analyze both nucleotide and protein sequences. The searches of these germline V (D) J genes are against germline gene databases IMGT and other sequence databases such as VBASE2 (Retter, Althaus, Münch & Müller, 2005) concurrently to maximize the chance of getting possible matching germline VH genes.

2.8 Germline allele discovery tools

Investigations have been performed on data from humans to seek to predict germline alleles in immunoglobulins leading to the development of germline allele discovery tools. There are both online tools and standalone, offline software (Wendel, He, Crompton, Pierce & Jiang, 2017; Gadala-Maria, Yaari, Uduman & Kleinstein, 2015; Corcoran et al., 2016). In this project, standalone tools were explored. These tools were developed for human antibody sequences analysis; the tools are then assessed on bovine antibody sequences.

In this study, IgDiscover and TIgGER were tested. The selection of these tools was since they are open source, their codes are freely available for adjustment to fit the used data set, and they all approach the problem in different ways as explained below.

IgDiscover is written in Python and works based on identifying consensus sequences that have been clustered to an initial database. For it to perform effectively, it requires at least 400,000 paired-end reads to determine the fully expressed germline repertoire of an individual. IgDiscover considers four critical steps in its workflow; i) initial assignment of starting database using IgBlast, ii) clustering to identify candidate germline sequences, iii) germline filtering, i.e., pre-germline filter and germline filter to retaining the correct germline sequences that are output as a new database replacing the initial database and finally iv) the first three steps are repeated using the newly created database produced in the previous iteration. The number of iterations to run IgDiscover is not limited. However, at the last iteration, germline filtering is done, i.e., stricter filtering criteria are applied to produce an individualized germline V database.

TIgGER (Tool for Immunoglobulin Genotype Elucidation via Rep-Seq) is an R package that infers a set of immunoglobulin alleles (including novel alleles) carried by an individual. It then uses this set of alleles to correct the initial assignments. The package covers three main tasks; i) novel allele detection, ii) inferring genotype, and iii) correcting allele calls.

In this study, a comparison on the performance of three annotation tools, that is, IMGT/HighV-QUEST (Lefranc et al., 2009), IgBlast (Ye, Ma, Madden, & Ostell, 2013) and MiXCR (Bolotin et al., 2015) was done. Prediction of novel bovine germline alleles was then done using IgDiscover (Corcoran et al., 2016) and TIgGER (Gadala-Maria, Yaari, Uduman, & Kleinstein, 2015).

CHAPTER 3: MATERIALS AND METHODS

This thesis project was undertaken as part of a bigger research project hosted by the International Livestock Research Institute (ILRI). The work focused on the bioinformatics analysis of already available sequence data of immunoglobulin heavy chain repertoires prepared at ILRI.

3.1. Benchmarking and optimizing annotation tools

3.1.1. Simulated bovine datasets generation

There isn't a gold standard dataset for validating bovine B cell repertoire sequence annotation tools. For this reason, a bovine repertoire was simulated using IgSimulator version 2.0 (Safonova, Lapidus & Lill, n.d., 2015) to enable conducting a fair comparison of available tools. IgSimulator tool used the IMGT germline gene database as an input. It first generated a base antibody sequences using the simulation of the VDJ recombination process. Then, randomly assigned counts to each base antibody sequence using power-law distribution. It then introduced somatic mutations in each base antibody sequence resulting in a mutated sequence. On each mutated sequence, it assigned counts using power-law distribution to generate antibody repertoire. This antibody repertoire was then subjected to next-generation sequencing read simulator ART Version 2.1.8 (Huang, Li, Myers & Marth, 2012) to generate Illumina paired-end reads. Simulation of a diverse and a polarized antibody repertoire was done. Diverse repertoire constitutes a high number of low abundant clusters of mutated sequences, whereas the polarized repertoire represents an increased number of repetitive clusters. For the diverse antibody repertoire, the number of base antibody was set as 100,000, and the number of mutated antibodies was set to 200,000. For the polarized antibody repertoire, the number of base antibody was set as 20,000 and mutated antibody 100,000 (Smakaj et al., 2019). A high number of base antibody and a slight difference between the number of base antibody and mutated antibodies results in simulation of a repertoire with a large number of different lowly abundant clusters. The low number of the base antibody with a big difference between the base

antibody and mutated antibodies leads to a simulation of a repertoire with a small number of different highly abundant clusters. IgSimulator simulates antibody repertoire by modeling how they are generated through the complex process of VDJ recombination, somatic hypermutations, and intergenic insertions. The simulation process happens in five steps as illustrated in figure 3;

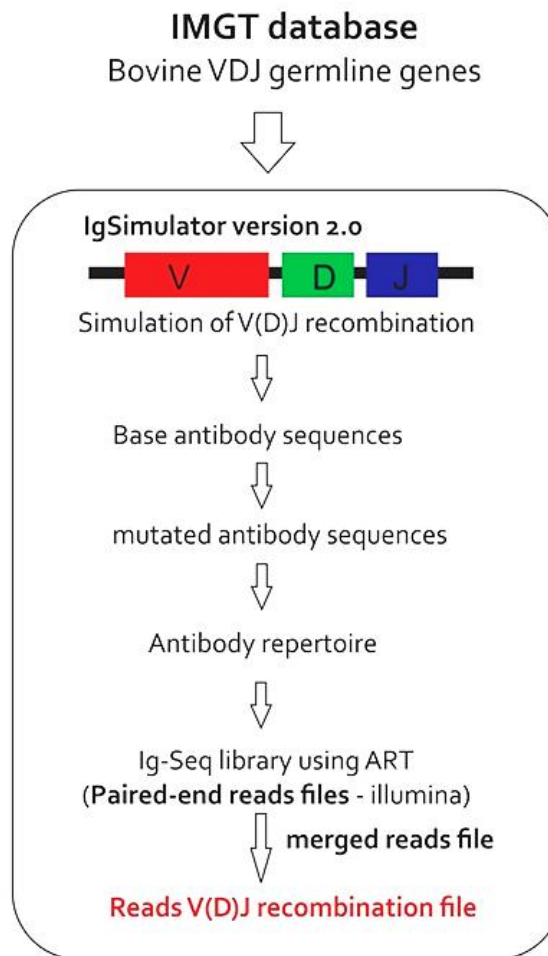


Figure 3. A five-step IgSimulator workflow of simulating an antibody repertoire.

3.1.2. Adaptive Immune Receptor Repertoire sequencing (AIRR seq) annotation

Three annotation tools were used for this analysis - IgBlast, MiXCR, and IMGT/HighV-QUEST. They produced a tab-separated annotation output file that was used to compare with read recombination text file.

IgBlast annotated V, D, and J genes of input sequences by matching these regions to germline gene libraries downloaded from NCBI. It used default BLAST search parameters (Ye, Ma, Madden & Ostell, 2013). IgBlast version 1.15.0 was used in this analysis. IMGT/High V-QUEST identified V, D, and J genes and alleles by alignment with the IMGT reference germline gene and allele sequences (Wirtz & Sayer, 2014). IMGT/HighV-QUEST version 3.4.15 was used. MiXCR handled both paired and single-end reads. It aligned these reads with high V and J gene assignments accuracy. It further assembled identical reads into clones and corrects sequence and PCR errors (Bolotin et al., 2015). MiXCR version 3.0.10 was used. IgBlast and IMGT/HighV-QUEST used the IMGT germline reference database, whereas MiXCR used specific built-in V/D/J/C libraries. IgBlast, IMGT/High V-QUEST and MiXCR were used to compare their annotation of simulated bovine datasets. Validation of the accuracy of each annotation tool was assessed using the simulated bovine datasets. The analysis was done at the gene level since MiXCR does not annotate at the allele level.

The benchmarking workflow was generated using R program version 3.6.2 platform. All R scripts for the analysis are provided in a git repo https://github.com/LandiMi2/Msc_Bioinformatics_Project. The comparison was made based on frequencies of misidentification as well as their distributions, as displayed in figure 4.

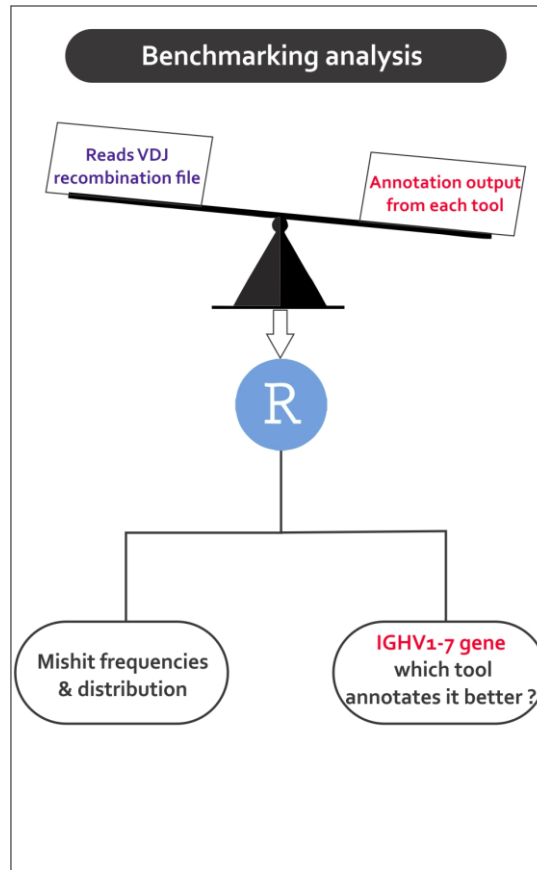


Figure 4. Benchmarking steps executed to compare IgBlast, IMGT/High V-QUEST and MiXCR annotation tools. Read V (D) J recombination file contains names of V, D, and J germline genes of each antibody in the antibody repertoire. This file is considered as the “true” VDJ genes that will be used to compare annotation outputs of each tool used in the benchmark.

3.1.3. Modifying germline JH and VH genes

Bovine immunoglobulin generation appears to utilize only two JH genes IGHJ1-6 and IGHJ2-4 (Stanfield et al., 2018). Therefore, the simulated dataset was generated these germline JH genes.

In addition, IGHV1-33 gene has 100% nucleotide identity to IGHV1-21*01 allele. They are at different loci in the bovine immunoglobulin locus and are likely a result of gene duplication. The simulated dataset was generated having only one of the identical VH genes, i.e. IGHV1-21*01 was used, and IGHV1-33 gene discarded. The simulation was done after these adjustments, and mishits frequencies comparison was also performed.

3.2. Germline allele discovery

IgM bovine immunoglobulin sequences from three African breeds i.e., Ankole breed had 292,276 reads for both forward and reverse reads, Boran breed had 285,306 reads for both forwards and reverse reads, Friesian breed had 450,999 reads for both forward and reverse reads, and Ndama breed had 222,483 reads for both forward and reverse reads. These sequences were used in this analysis. Holstein Friesian cattle breed was also used in the analysis as a control because it is the only western breed in this analysis.

3.2.1. Germline allele discovery using IgDiscover

IgDiscover (Corcoran et al., 2016) identified novel germline VH alleles by an iterative process. For the discovery analysis to begin: - fasta or fastq files of single or paired-end reads from the variable region of BCR sequencing are required. Paired-end reads fastq files of the bovine breeds were used. Bovine VDJ germline database sequences were also required. IMGT bovine database sequences were downloaded from <http://www.imgt.org/vquest/refseqh.html>.

IgDiscover doesn't identify IMGT long sequences headers, and therefore, instructions on setting up germline database for the tool were followed to comply with IgDiscover conditions. Ultimately, a configuration file is needed that describes the library. This configuration file is adjusted to fit the experiment. In this analysis, the number of iterations to 20 and 30 was amended. These recommendations were determined from multiple trials of 2, 3, 5, 10, 20, and 30 iterations. To give an indication of whether novel alleles are being discovered at every additional iteration. The number of iterations at which individualized databases (new database) produced reached a plateau differs for each breed. Ankole, Boran and Friesian reached a plateau at 20 iterations whereas Ndama required 30 iterations. Pre-germline and germline filtering parameters control the differences between the database gene and a sequence to be assigned in a cluster. The requirements of the minimum number of unique JH genes in both pre-germline and germline filters was reduced to one. As earlier mentioned in section 3.2.3, bovine immunoglobulin repertoire has been observed to express only two JH genes, of which

IGHJ2-4 is predominantly expressed. That is why this parameter is set to one. Other parameters in the configuration file were set as default.

3.2.2. Germline gene discovery using TIGGER

TIGGER uses IMGT/HighV-QUEST or Change-O (Gupta et al., 2015) output for its analysis. Four bovine breeds were used in this analysis. TIGGER version 1.0.0 was downloaded from CRAN (<https://cran.r-project.org>). The reference germline database IGHV (gapped) was derived from the IMGT database. The process of discovering novel alleles entails aligning all sequences to a particular germline gene. A plot is generated of mutation counts at each position in the alignment. This is determined as a function of sequences-wide mutation counts. Three pieces of evidence are necessary for a sequence to qualify as a novel sequence. First, the y-intercept of the plot is arbitrarily set as 1/8 by default. Passing this y-intercept threshold is one evidence. Secondly, it was expected that the polymorphic allele to be shared at all mutational frequencies, and the mutation counts to correspond to the number of polymorphisms. Lastly, novel sequences should utilize a wide range of JH genes.

When discovering novel alleles for this experiment, two parameters were adjusted; that is, `j_max` and `min_seqs`. The `j_max` setting is the maximum fraction of sequences precisely aligning to a possible novel allele that can utilize a particular combination of junction length and JH gene. The `j_max` parameter was set to one to switch off the filter for JH gene diversity because bovine breeds utilize fewer JH genes. The `min_seqs` parameter is the minimum number of total sequences required by a sample within the desired mutational range to be considered a novel sequence. This was also reduced from 50 to 15. Other parameters were set as default for the discovery process.

3.2.3. Haplotype networks

Novel alleles discovered were represented using haplotype networks. The HaploNet function in the pegas R package was used to generate these networks. Haplotype networks are based on Hamming distances between sequences. Unique sequences define haplotypes in this analysis.

CHAPTER 4: RESULTS

4.1. IgSimulator outputs

The outputs from IgSimulator tool were used as inputs for the annotation and benchmarking analysis. These outputs were: - Illumina paired-end reads that was used as an input in MiXCR annotation tool; merged read file as a result of combining forward and reverse read was used as an input in IgBlast and IMGT/HighV-QUEST annotation tools; and read recombination file that contains all the information about V (D) J recombination for each reads from the merged file. This file represented the “True” V (D) J germline gene information of each antibody recombined in the repertoire simulated and used as the benchmark dataset to compare the annotation outputs of IgBlast, IMGT/HighV-QUEST and MiXCR.

4.2. Benchmarking annotation of simulated bovine immunoglobulin sequences

4.2.1. Misidentification frequencies

From the comparison, MiXCR had the highest frequencies of incorrect calls having wrong matches of 33,754 out of 269,672 (13%) antibodies for VH gene annotation compared to IMGT/HighV-QUEST (16,346 out of 420,668 antibodies) and IgBlast (15,826 out of 420,668 antibodies) that had the same mishit percentage frequency of 4% (Figure 5).

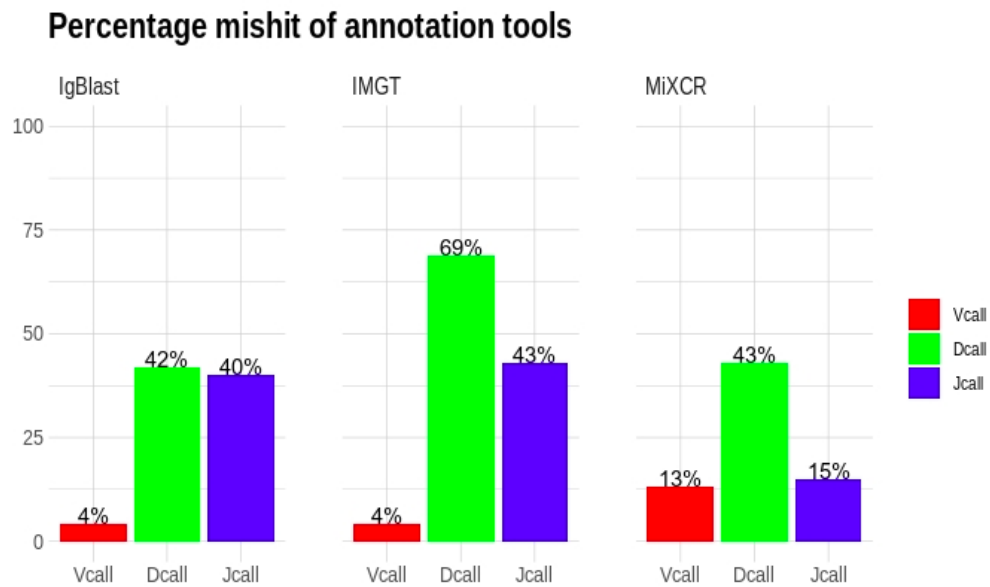


Figure 5. Bar plots for the predicted percentage of mishits frequencies at the gene level. Percentage mishits counts of each annotation tool (red – VH gene calls, green – DH gene calls and blue – JH gene calls).

IMGT/HighV-QUEST had the highest misidentified counts of DH genes of 291,328 out of 420,668 (69%) antibodies. IgBlast also showed a high frequency of incorrect calls in DH gene annotation of 178,229 out of 420,668 (42%) antibodies. These higher misidentified frequencies for DH genes were as a result of unassigned DH genes in both IgBlast and IMGT/HighV-QUEST (Figure 6). IMGT/HighV-QUEST and IgBlast cannot assign the DH and JH genes and alleles when no DJ rearrangement is identified in the submitted sequences.

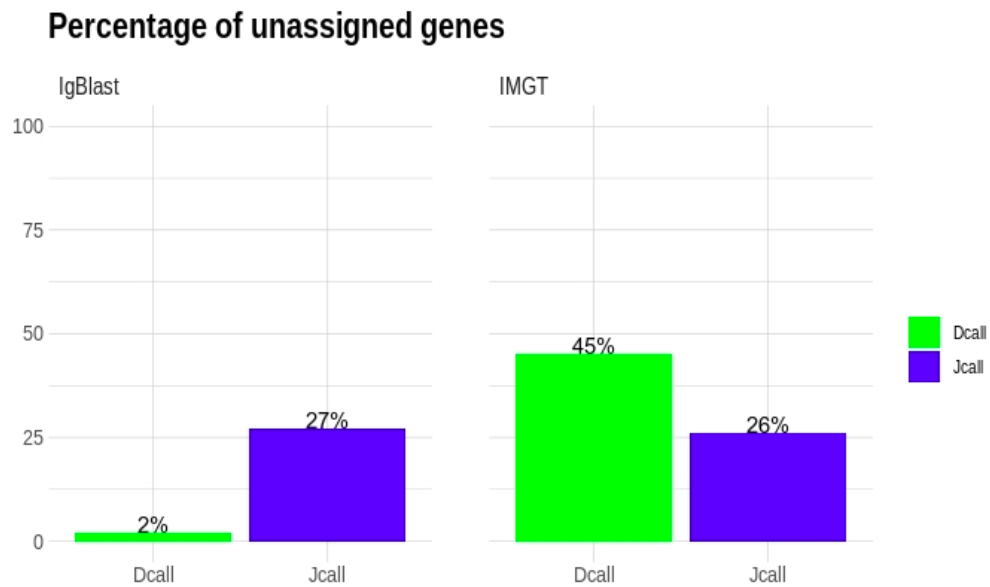


Figure 6. Bar plot for the percentage of unassigned genes of IgBlast and IMGT/HighV-QUEST.

While bovine utilize only two JH genes, modification of the germline JH gene germline database was done to contain only IGHJ1-6 and IGHJ2-4 alleles. Furthermore, for the germline VH alleles, the IGHV1-33 gene was deleted, a duplicate of IGHV1-21 gene. After modifying these germline database, an antibody repertoire was simulated, then annotation, and finally compared the annotation outputs. Figure 7 shows the comparative percentages of incorrect calls modifying germline VH and JH genes. IgBlast and IMGT/HighV-QUEST have 0% misidentification value for VH genes compared to MiXCR having 10%. However, MiXCR has a better J gene annotation of 100% percentage correct calls.

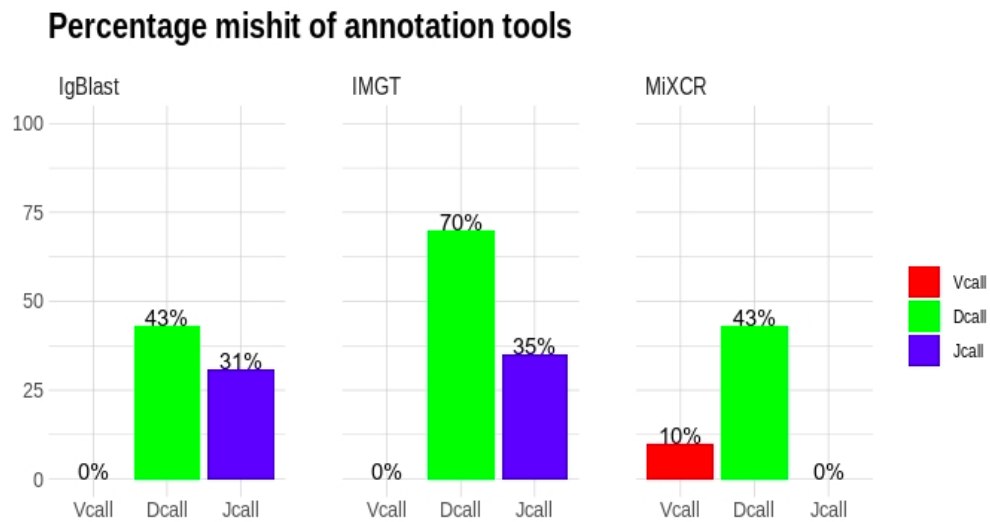


Figure 7. Bar plots for the predicted percentages of incorrect identifications after modifying the germline VH and JH database.

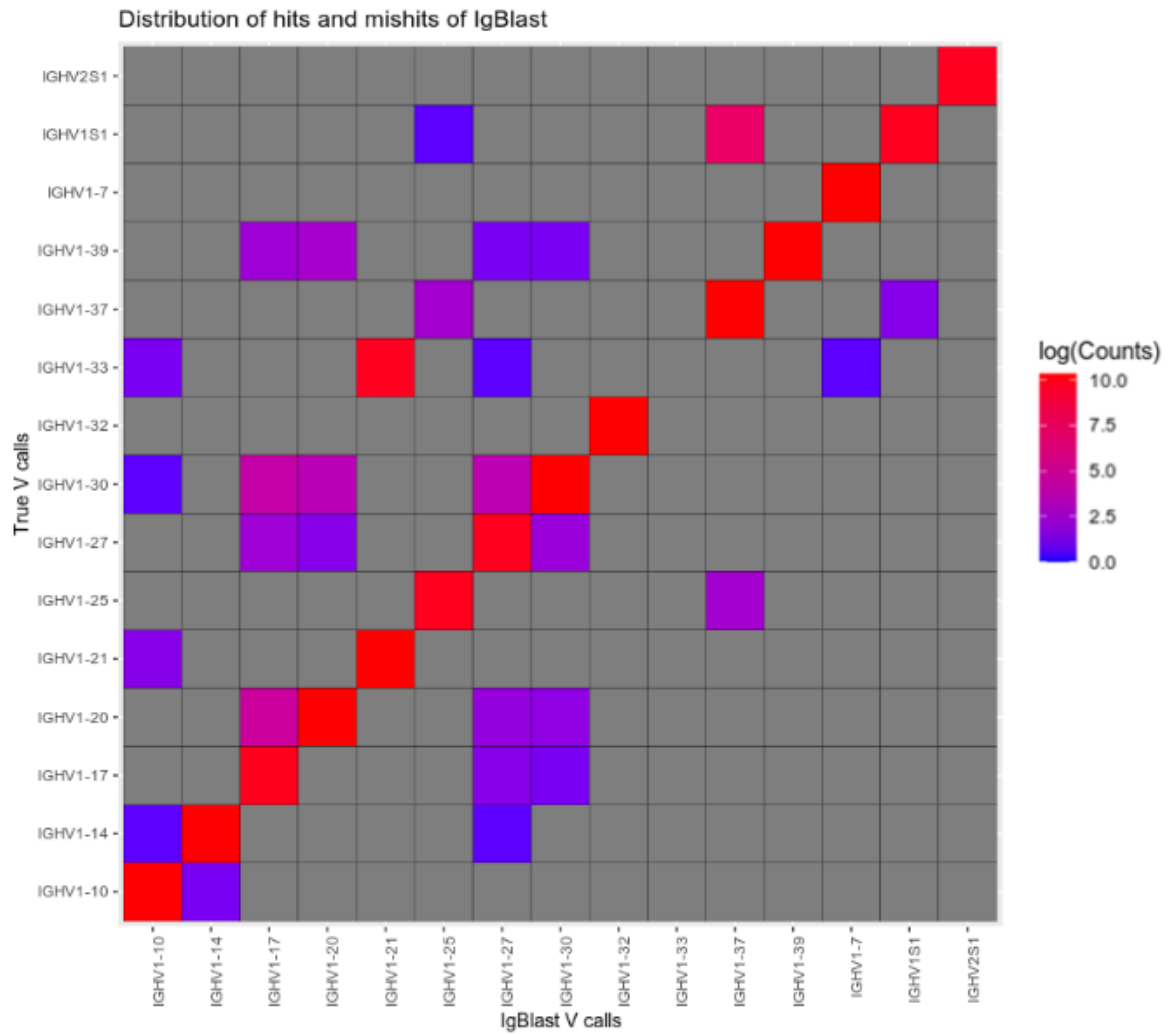
IgBlast and IMGT/HighV-QUEST annotation outputs have a more favorable outcome of VH gene annotation with reduced error rate counts of 1,321 and 1,836 incorrect identification respectively out of 432,849 antibodies. MiXCR still has a high error rate of 26,038 mistaken identifications out of 271,552 (10%) antibodies in annotating VH genes. MiXCR improved its annotation of the JH gene with an error rate of 54 misidentifications out of 271,552 (0%) antibodies compared to IMGT/HighV-QUEST and IgBlast that still had an error rate of 149,560 wrong identification out of 432,849 (35%) antibodies and 134,834 out of 432,849 (31%) antibodies, respectively.

4.2.2. Distribution of correct and incorrect VH gene annotation

To further investigate why genes were being misidentified. Heat maps were generated from the frequencies of incorrect calls of the unmodified germline database. Figure 8 shows that IgBlast and IMGT/HighV-QUEST VH gene annotation had relatively fewer numbers of inaccurately annotated genes compared to MiXCR. Across all tools, it was evident that the IGHV1-21 gene was misidentified as the IGHV1-33 gene as they are identical at the nucleotide

level. IMGT/HighV-QUEST and IgBLAST correctly annotated two V pseudogenes IGHV1S1 and IGHV2S1. MiXCR was poor at annotating these two pseudogenes.

A





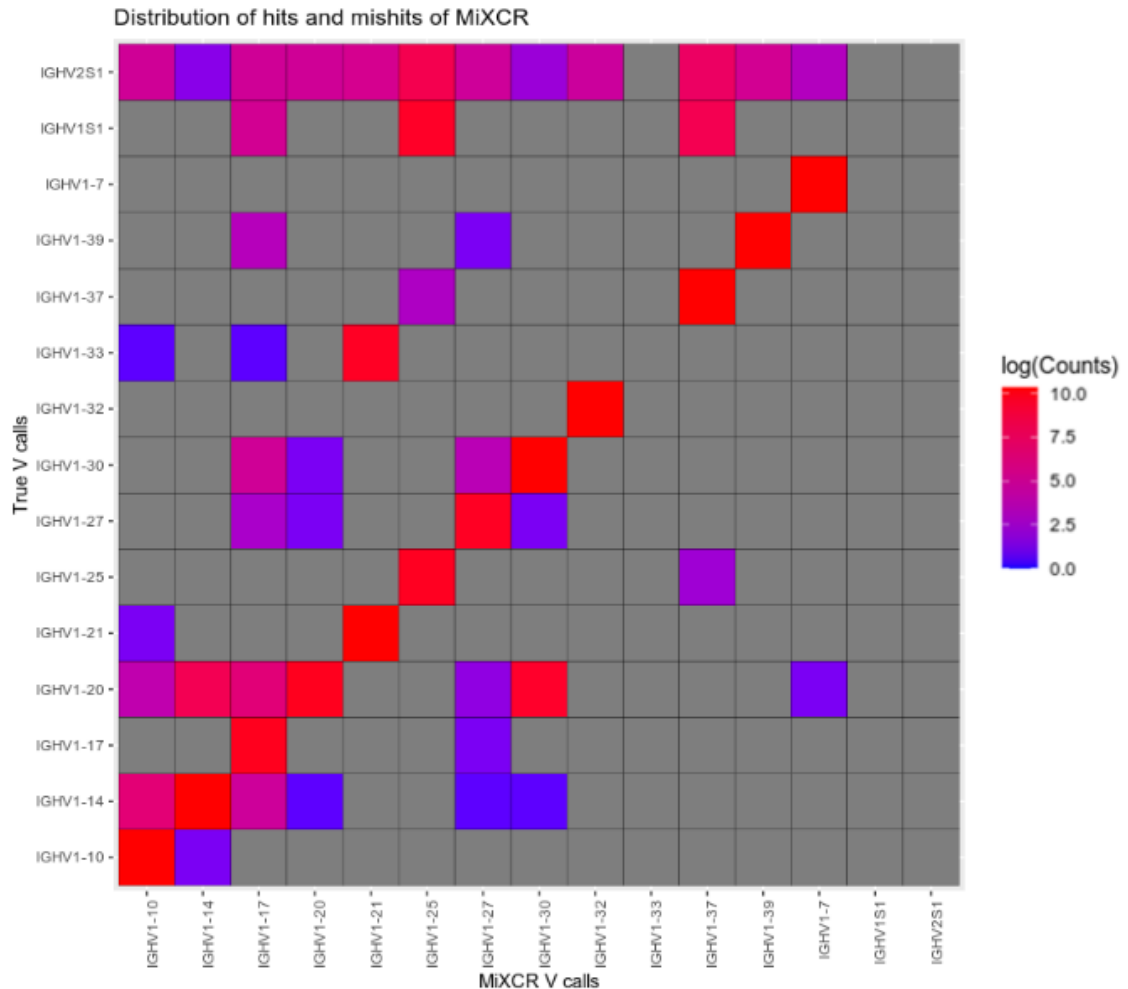


Figure 8. Distribution of hits and mishits for all tool’s V gene annotation. A cross table is generated having predicted true V gene calls and annotation tool V gene calls. The matrix is used to generate a heat map in this analysis. The counts were log-transformed. Results for IgBlast – labeled A, IMGT/HighV-QUEST – labeled B and MiXCR – labeled C are shown.

IGHV1-7 gene was done correctly for IgBlast and MiXCR, whereas IMGT/HighV-QUEST misidentified the gene with low counts as IGHV1-10 and IGHV1-21 genes. Across all tools, the IGHV1-37 gene is misidentified as IGHV1-25. IGHV1-37 gene and IGHV1-25 have a 98.63% sequence identity at the nucleotide level and a 100% identity at the protein level. Across all tools, the IGHV1-30 gene and IGHV1-27 gene are misidentified as IGHV1-17 and IGHV1-20. IMGT/HighV-QUEST and IgBlast misidentified the IGHV1-39 gene as IGHV1-17 and IGHV1-20. MiXCR misidentified the IGHV1-39 gene as the IGHV1-7 gene.

4.3. Germline allele discovery

4.3.1. Novel alleles discovered by IgDiscover

Expressed IgM bovine repertoire sequences of Ankole, Boran, Friesian and Ndama breeds were subjected to IgDiscover (version 0.12). This analysis used the Friesian breed as a control. IgDiscover identified 58 novel alleles for the Friesian breed, 18 novel alleles for Boran breed, six novel alleles for Ndama breed, and three novel alleles for Ankole breed. Figure 10 presents the final number of alleles discovered at the end of the last iteration and figure 9 presents number of alleles identified after pre-germline filtering.

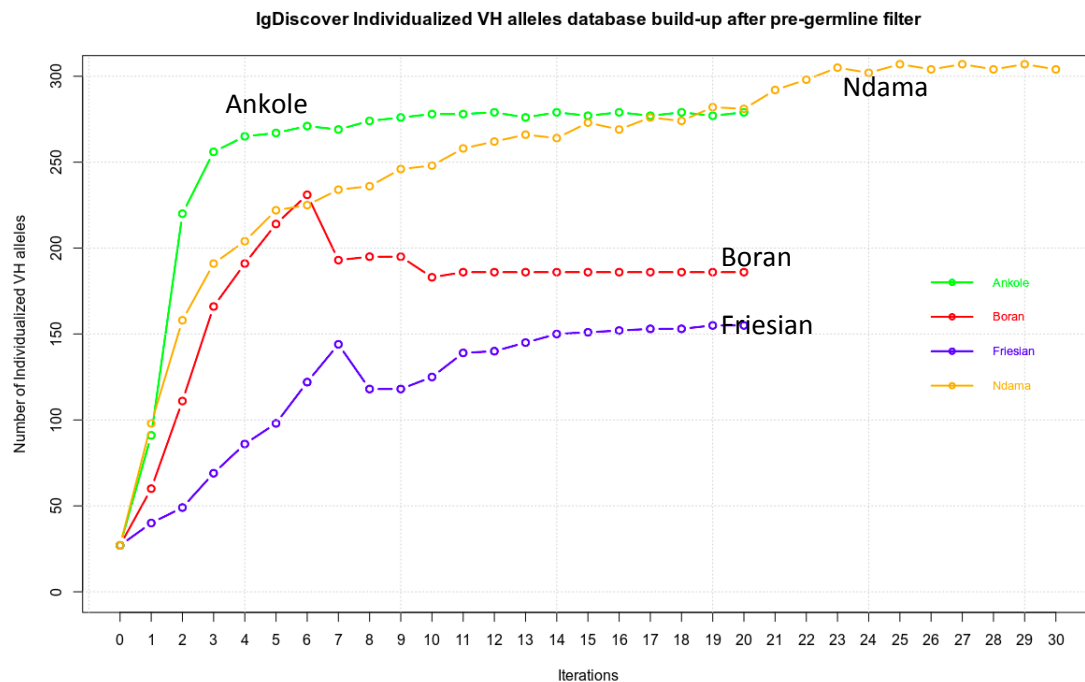


Figure 9. Individualized database VH alleles buildup of Ankole, Boran, Friesian and Ndama before germline filtering. Note that these numbers of VH alleles are after pre-germline filtering, which is a less strict filtering criterion. 20 iterations were run for Boran, Friesian and Ankole breeds, and 30 iterations for the Ndama breed in order to reach a plateau everywhere.

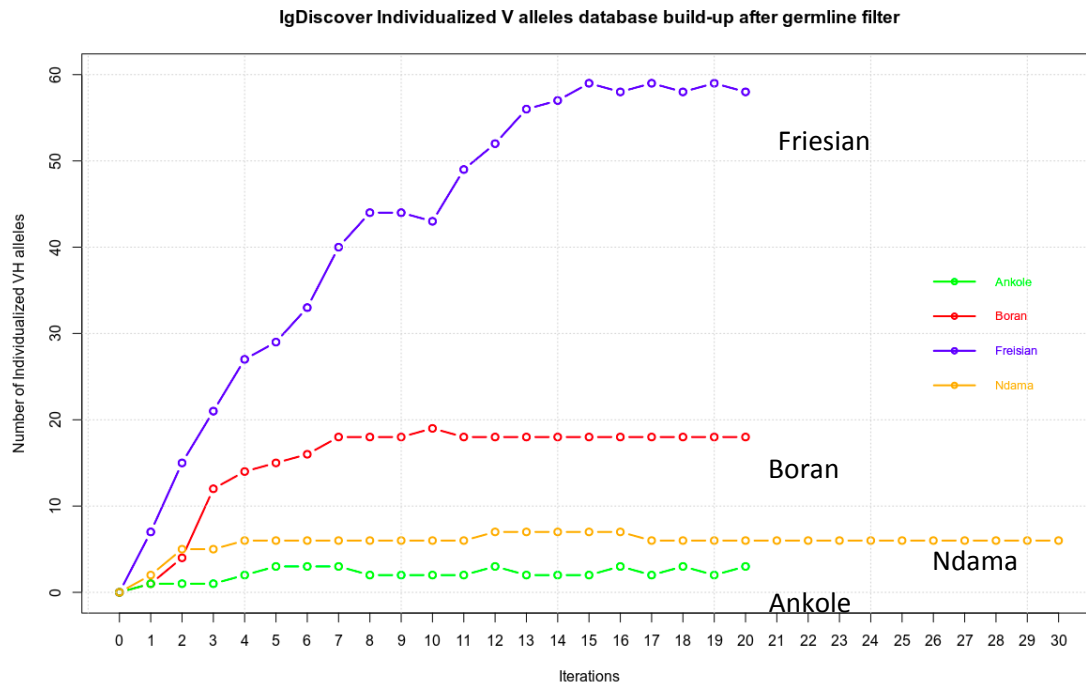


Figure 10. The line graph shows individualized database VH alleles built up of each bovine breed after germline filtering. Friesian had the most discovered novel alleles of 58. Three novel alleles are seen in Ankole breed, 18 in Boran breed and six in Ndama breed.

Haplotype network for visualization of the Hamming distances of novel alleles discovered

Haplotype network is created from pairwise Hamming distances for all the novel alleles of African bovine breeds discovered by both methods to assess their genetic relatedness. Pairwise Hamming distance gives a measure of genetic difference between novel alleles. Figure 11 shows all novel alleles discovered by IgDiscover in African bovine breeds. Ndama (yellow circles in figure 11) was seen to have novel alleles very distant from each other. A few novel alleles observed in Boran (red circles) are distant, whereas novel alleles from Ankole (green circles) are too close, and one is quite distant from the other. Novel alleles of IGHV1-7 gene were not observed in Boran and Ndama, but one allele of this gene was observed in Ankole. Verification of this particular allele (IGHV1-7*01_S6322_Ankole) confirmed the presence of the characteristic CTTVHQ motif at the 5' end of the IGHV1-7 gene.

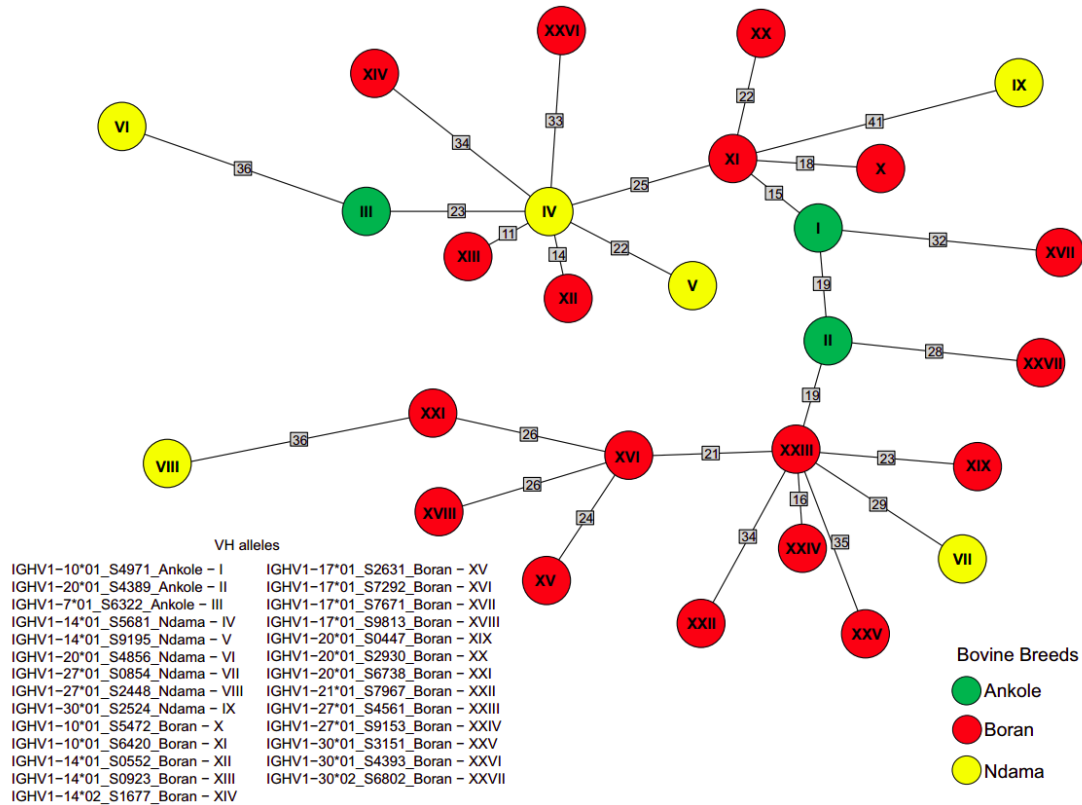


Figure 11. Haplotype network of novel alleles discovered by IgDiscover in three African breeds. Alleles from different breeds are represented in different color: Ankole in green (3 novel alleles), Boran in red (18 novel alleles) and Ndama in yellow (6 novel alleles). The numbers decorating the edges are the Hamming distances between two novel alleles.

Ranking of African novel alleles discovered by IgDiscover for wet-lab validation

The novel alleles were ranked based on the following criteria:

1. The number of sequences assigned to the allele.
2. Cluster size of the allele.
3. Number of unique CDR H3.
4. Number of unique DH genes and JH genes the allele recombined with.
5. Number of exact matches, which is a count of reads matching the novel allele exactly.
6. Percentage difference of the sequences assigned.

This ranking is essential at the verification step, the potential novel alleles with the highest number of sequences assigned to it, cluster size and a high number of unique CDR H3s or and JH genes is considered first for the verification step. Tables 1 - 3 list ranked novel alleles discovered in Boran, Ndama and Ankole breeds, respectively.

Table.1. Novel alleles discovered in Boran breed ranked in descending order by number of assigned sequences.

	Candidate novel allele	Source of novel allele	Cluster size ^a	%difference (modal) ^c	No of sequences assigned ^b	Unique CDR H3	Exact ^d	Unique DHs	IGHJ2-4 counts
1	IGHV1-27*01_S4561	IGHV1-27*01_S4561	310	0.1	68498	55	130	10	67016
2	IGHV1-10*01_S6420	IGHV1-10*01_S6420	77	0.09	52293	11	25	8	51419
3	IGHV1-14*01_S0923	IGHV1-14*01_S0923	77	0.1	20078	7	16	7	18318
4	IGHV1-17*01_S7292	IGHV1-17*01_S7292	77	0.09	9524	7	15	7	9003
5	IGHV1-10*01_S5472	IGHV1-10*01_S5472	234	0.1	7814	12	46	7	7696
6	IGHV1-20*01_S6738	IGHV1-20*01_S6738	55	0.11	4985	4	16	7	4638
7	IGHV1-20*01_S0447	IGHV1-20*01_S0447	110	1% ; 13%	3467	11	38	6	3356
8	IGHV1-14*01_S0552	IGHV1-14*01_S0552	452	0.11	3447	10	62	3	2920
9	IGHV1-27*01_S9153	IGHV1-	65	0.1	1581	8	18	7	1564

		27*01_ S9153							
1 0	IGHV1- 17*01_S 9813	IGHV1 - 17*01_ S9813	115	0.13	1514	7	53	1	1436
1 1	IGHV1- 17*01_S 7671	IGHV1 - 17*01_ S7671	60	0.01	1101	41	253	3	1098
1 2	IGHV1- 14*02_S 1677	IGHV1 - 14*02_ S1677	95	0.13	950	6	26	1	873
1 3	IGHV1- 20*01_S 2930	IGHV1 - 20*01_ S2930	111	1% ; 13%	858	7	20	4	853
1 4	IGHV1- 21*01_S 7967	IGHV1 - 21*01_ S7967	329	0.01	780	9	42	3	776
1 5	IGHV1- 30*01_S 4393	IGHV1 - 30*01_ S4393	66	1% ; 11%	543	10	27	3	541
1 6	IGHV1- 17*01_S 2631	IGHV1 - 17*01_ S2631	116	1% ; 9%	458	11	40	3	453
1 7	IGHV1- 30*01_S 3151	IGHV1 - 30*01_ S3151	141	0.01	366	12	36	2	358
1 8	IGHV1- 30*02_S 6802	IGHV1 - 30*02_ S6802	51	0.02	269	8	23	0	269

Table.2. Novel alleles discovered in Ndama breed ranked in descending order by number of assigned sequences.

	Candidate novel allele	Source of novel allele	Cluster size ^a	%difference (modal) ^c	No of sequences assigned ^b	Unique CDR H3	Exact ^d	Unique DHs	IGHJ1-6 ; IGHJ2-4 counts
1	IGHV1-14*01_S5681	IGHV1-14*01_S5681	2508	0.11	100916	7	10	10	631 ; 96,106
2	IGHV1-27*01_S0854	IGHV1-27*01_S0854	54	0.12	31643	3	13	9	267 ; 30,651
3	IGHV1-30*01_S2524	IGHV1-30*01_S2524	76	1% ; 14%	315	11	28	2	314
4	IGHV1-14*01_S9195	IGHV1-14*01_S9195	100	0.01	139	4	49	0	27
5	IGHV1-20*01_S4856	IGHV1-20*01_S4856	52	1% ; 11%	106	6	17	0	105
6	IGHV1-27*01_S2448	IGHV1-27*01_S2448	60	No histogram	NA	NA	NA	0	97

Table.3. Novel alleles discovered in Ankole breed ranked in descending order by number of assigned sequences.

	Candidate novel allele	Source of novel allele	Cluster size ^a	%difference (modal) ^c	No of sequences assigned ^b	Unique CDR H3	Exact ^d	Unique DHs	IGHJ2-4 counts
1	IGHV1-10*01_S4971	IGHV1-10*01_S4971	118	0.09	29929	11	35	8	29205
2	IGHV1-20*01_S4389	IGHV1-20*01_S4389	60	0.09	13225	10	29	7	12862
3	IGHV1-7*01_S6322	IGHV1-7*01_S6322	454	3% ; 8%	13143	13	19	8	12454

^{a, b} The difference between the cluster size ^a and the number of sequences assigned ^b to the allele arises as IgDiscover generates sub-clusters from the sequences assigned to the allele, after which only one of those sub-clusters is used for the identification of a particular novel allele.

^c Percentage difference of sequences assigned to that allele. The distribution of the differences can either be shifted left (< 0.1) or right (> 0.1). Columns that have two values of percentage difference are an indication of a bimodal distribution.

^d The “Exact” column tells how often the sequence occur exactly, meaning with no mismatch and no indel.

The selected African breeds have shown to have greater diversity (figure 9) through the increasing number of pre-germline novel alleles discovered compared to the exotic Friesian breed. Yet, the largest number of novel alleles identified after germline filtering by IgDiscover occurs in Friesian. One possible explanation for this is because the starting database is composed of alleles described in Holstein cows which are phylogenetically distant to African breeds. If a novel sequence is the same as the known germline allele, IgDiscover includes this

novel sequence as the output, even if the pieces of evidence are too weak. IgDiscover is less strict with these sequences because the presence of this sequence is enough evidence that it is a true allele, and so they do not get filtered out. For this reason, more novel alleles are recorded in the Friesian sample.

4.3.2. Novel bovine alleles discovered by TIgGER

TIgGER performs discovery of novel alleles in a different way than IgDiscover does, as described in section 3.3.2. This project was aimed at comparing the novel alleles found by the two tools.

TIgGER discovered 21 novel alleles in Ndama, 15 in Boran, 1 in Ankole, and 8 in the Friesian breed. Out of 21 novel alleles in Ndama, 3 novel alleles were filtered out because they had identical nucleotide sequences. Eight novel alleles in Boran were filtered out as they had zero counts of unique CDR H3 and JH genes, and also, the sequences from these novel alleles were not present in the input data.

The haplotype network for TIgGER output data also shows pairwise Hamming distances between novel alleles discovered. There are additional novel alleles that originate from the IGHV1-37 gene and IGHV1-39 gene (figure 12) that were not detected in the African novel alleles discovered by IgDiscover. IGHV1-7*01 novel allele was found only in Ankole breed by both tools. From the haplotype network in figure 12, the novel allele identified from IGHV1-37 gene in Boran is distantly related with the highest number of differences to the allele it is compared to. Ndama also has novel alleles that are as distant as predicted in IgDiscover, this will be seen later through a side-by-side comparison of Hamming distances. Although TIgGER discovered more novel alleles in Ndama breed, ten of the novel alleles (I - II, IV-V-VI, XII-XIII, XVI-XVII-XVII) for this breed are closely related because they are derived from the same gene.

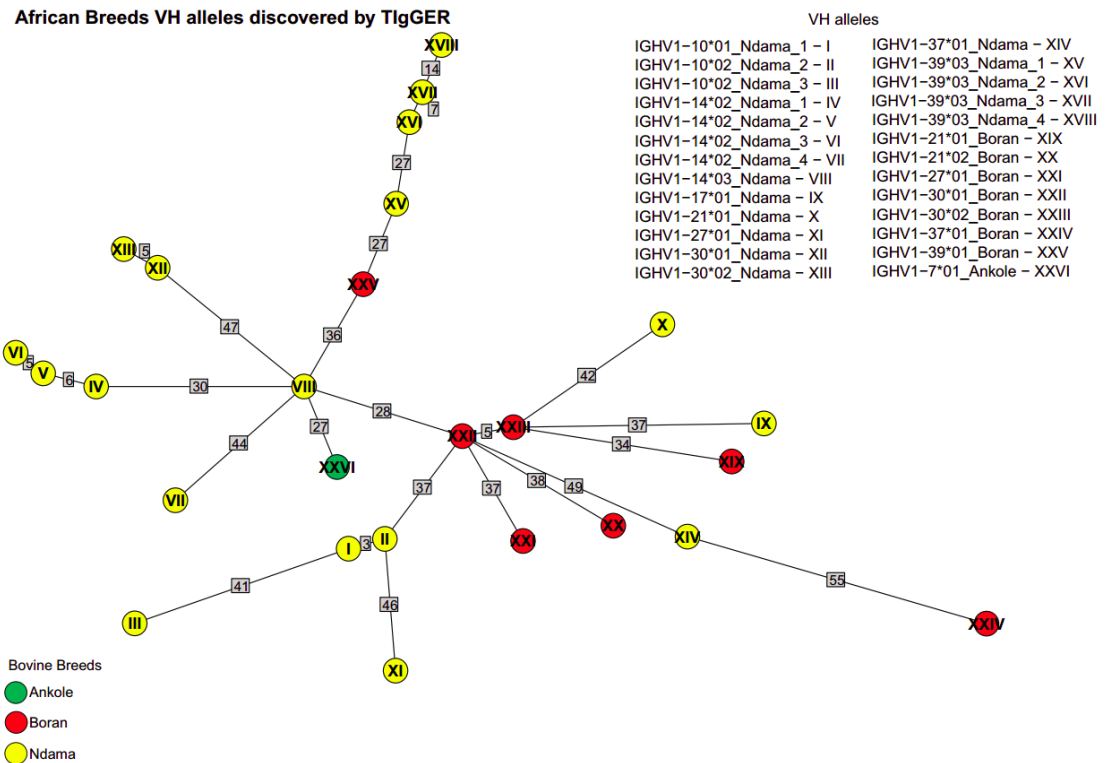


Figure 12. Haplotype networks of novel alleles discovered by TIgGER from Africa breeds. Colour codes are represented as green for Ankole, red for Boran and yellow for Ndama.

Moreover, Boran and Ankole had four and two identical novel alleles, respectively, discovered by both tools. In Boran, IGHV1-17*01_S7671 (IgDiscover) was identical to IGHV1-21*02 (TIgGER), and IGHV1-27*01_S4561 (IgDiscover) was identical to IGHV1-30*01 (TIgGER) at the nucleotide level. Ankole had one pair of novel alleles, namely IGHV1-7*01_S6322 (IgDiscover) and IGHV1-7*01 (TIgGER) that were identical.

4.3.3. Distribution of distances of novel alleles discovered by IgDiscover and TIgGER

The pairwise Hamming distances of all novel alleles discovered from the three African breeds were extracted for comparison to Friesian novel allele, and quantification of genetic diversity. A comparison of 27 and 26 novel alleles from African breeds, together with 58 and 8 Friesian novel alleles detected from IgDiscover and TIgGER, respectively are represented on a histogram and a boxplot (Figure 12 -15).

The histograms show the distribution of pairwise Hamming distances of novel alleles discovered from IgDiscover (figure 13) and TIgGER (figure 14). Pairwise Hamming distances of novel alleles from the selected African breeds are observed to be shifted to the right showing a distinct difference between the novel alleles of African and Friesian breeds. A T-test was done to calculate the significance of this difference. The null hypothesis implying that there is no difference in the mean number of pairwise Hamming distances of African novel alleles versus the Friesian novel alleles and the alternative is that there is a difference.

Table.4. Comparison of standard deviations (SD) and mean values of novel alleles of the selected African breeds and Friesian breed discovered by IgDiscover and TIgGER.

Breeds	<u>IgDiscover</u>			<u>TIgGER</u>		
	N	SD	Mean	N	SD	Mean
African	27	±10.03	41.93	26	±13.46	56.48
Friesian	58	±8.44	31.05	8	±10.55	29

Two sample t-test: **t = 18.936, DF = 460.76, p-value < 2.2e-16**

At $\alpha = 0.05$ level of significance, there is sufficient evidence to conclude that the average number of pairwise Hamming distances of the selected African breeds novel germline alleles are different from that of Friesian germline novel alleles.

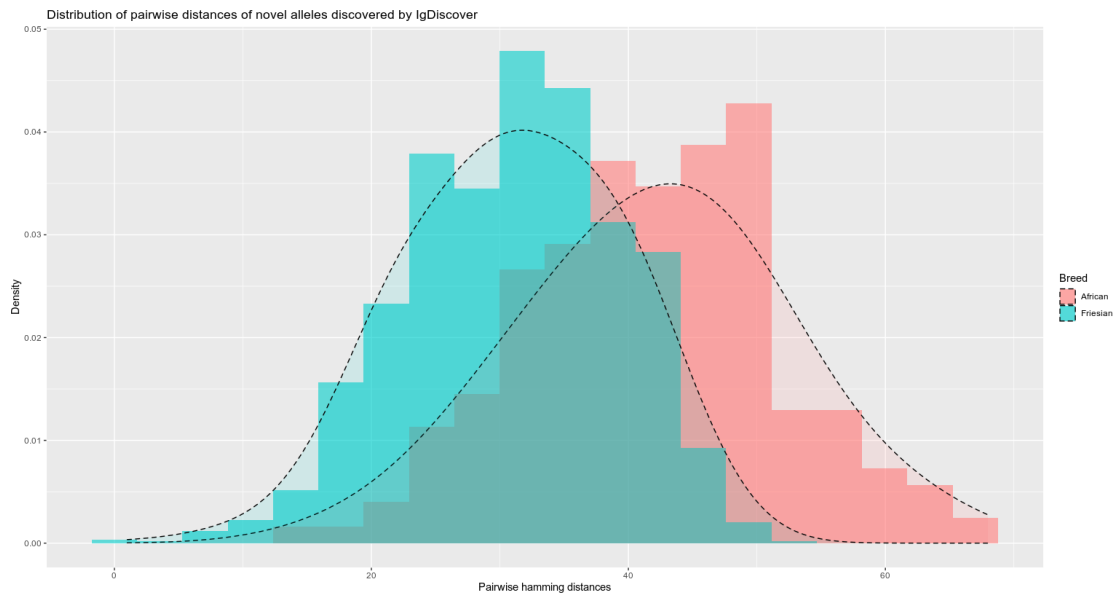


Figure 13. Histogram of the distribution of pairwise Hamming distances of novel alleles discovered by IgDiscover between the selected African breeds and Friesian breed. Twenty-seven novel alleles from the selected African breeds and 58 novel alleles from Friesian breed are plotted.

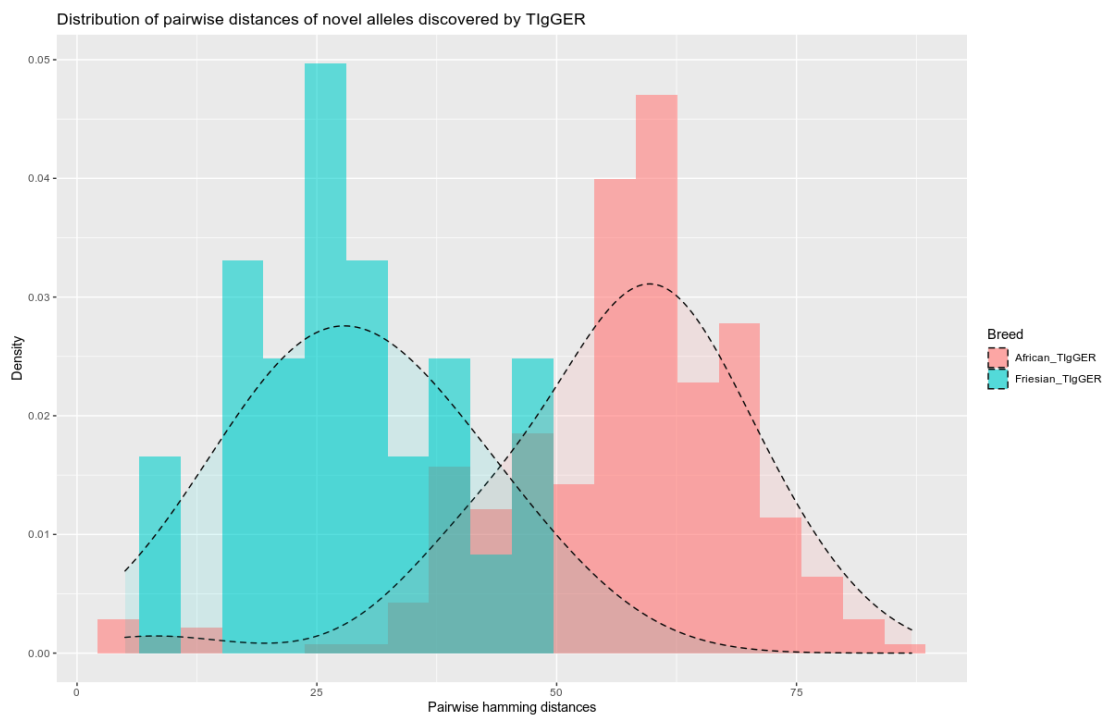


Figure 14. Histogram of the distribution of pairwise Hamming distances of novel alleles discovered by TIgGER. Twenty-six novel alleles from the selected African breeds and eight novel alleles from Friesian breed are plotted.



Figure 15. A boxplot showing Hamming distances of African novel alleles versus Friesian novel alleles, both discovered by IgDiscover and TIgGER. The median Hamming distance values for the selected African breeds are higher, 42 (IgDiscover) and 59 (TIgGER) compared to those of Friesian - 31 (IgDiscover) and 27.5 (TIgGER).

The distribution of pairwise distances observed in the selected African breeds is spread out, showing more genetic diversity compared to Friesian, which is compact. African novel alleles discovered from TIgGER show outliers of short distances between 3 and 15. These outlier distances are as a result of differences between different alleles derived from the same gene (figure 15).

Further analysis was done to measure the spread of pairwise distances of novel alleles within African as well as Friesian breed. This comparison showed that the pairwise distances in the selected African breeds had an increased magnitude of deviations with higher standard deviation number (Table 5) compared to the Friesian breed, from novel alleles discovered by both IgDiscover and TIgGER.

Table.5. Comparison of standard deviations (SD) and mean values for the distribution of pairwise Hamming distances between novel alleles within the selected African breeds and Friesian breed discovered by IgDiscover and TIgGER.

Breeds	<u>IgDiscover</u>			<u>TIgGER</u>		
	N	SD	Mean	N	SD	Mean
Ankole	3	±12.49	34.00	1	-	-
Boran	18	±9.31	39.67	7	±16.14	50.19
Friesian	58	±8.44	31.05	8	±10.55	29.00
Ndama	6	±10.72	50.60	18	±15.28	57.80

These pairwise distances per breed of the selected African breeds and Friesian are represented on a boxplot. Figure 16 shows that Ndama largely attributes the greater genetic diversity recorded in the selected African breeds. TIgGER seem to yield novel alleles that have many genetic differences compared to novel alleles identified by IgDiscover.

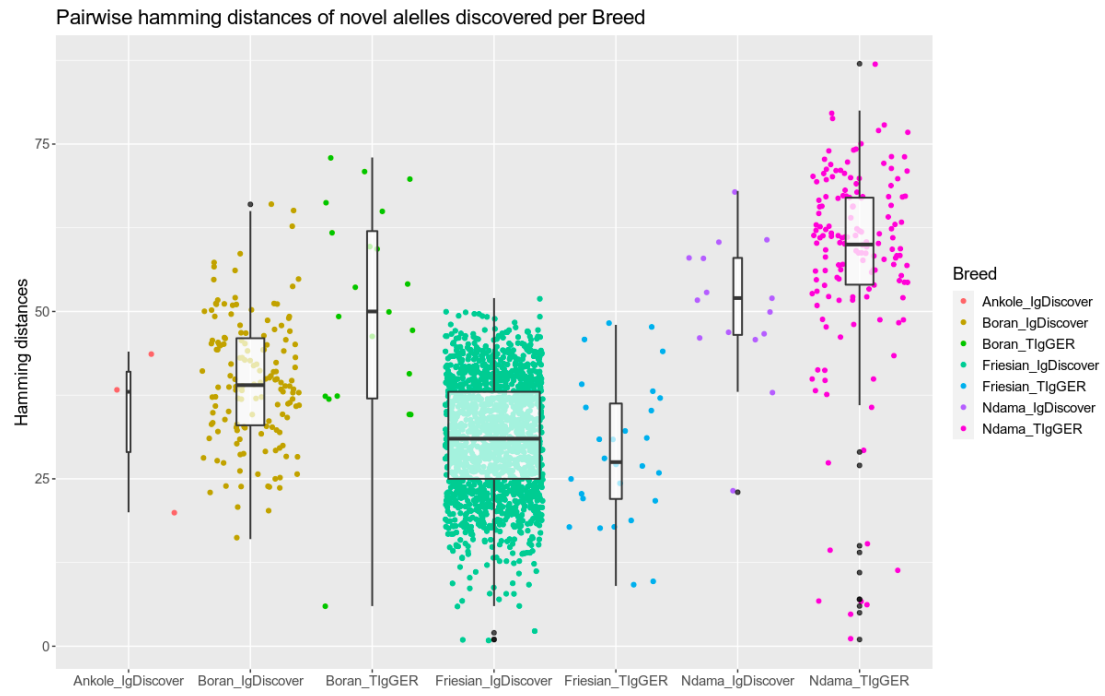


Figure 16. A boxplot showing Hamming distances of African novel alleles (per breed) versus Friesian novel allele both discovered by IgDiscover and TigGER. TigGER identified only one novel allele in Ankole and so it is not plotted in the figure.

CHAPTER 5: DISCUSSION, CONCLUSIONS AND RECOMMENDATIONS

5.1 Discussion

Accurately annotating antibody repertoire sequences is important for the identification and classification of immunity genes utilized by an individual. This analysis can be useful in diagnostics or antibody therapy in diseases like lymphoma, leukaemia, and rheumatoid arthritis, which are malignant or autoimmune.

Different B-cell annotation tools utilize different inputs. For instance, IMGT/HighV-QUEST and IgBlast use the IMGT germline database sequences as the input of annotation. In contrast, MiXCR uses an inbuilt library of specific germline database imported from the GenBank database. MiXCR and IgBlast can use other germline databases, but IMGT/HighV-QUEST is restricted to the IMGT germline database. These differences in the set of germline genes used as input are crucial, especially in downstream processes. There is always a need to benchmark these annotation tools to get the most accurate tool for an analysis. Currently, bovine immunoglobulin golden standard metadata for benchmarking has not been developed. This project simulated a bovine antibody repertoire for a benchmarking study.

There is variation in the accuracy of annotation between all three tools. Modifying germline databases to make them fit to what is observed in real bovine repertoire data improves the accuracy of annotation. This analysis shows no tool performed better for DH gene annotation because the region is highly diverse. In calculating the percentage of incorrect calls, IgBlast and IMGT/HighV-QUEST have the same denominator whereas MiXCR had a different denominator in calculating misidentified frequencies. This difference may be accounted since MiXCR aligned only about 64% while IgBlast and IMGT/HighV-QUEST aligned approximately 99.5% of the total reads simulated. MiXCR aligned DH genes only after the V/J junction position is determined. Hence, if a JH gene is not identified, MiXCR can't produce any successful alignment, because the sequence has no complete CDR H3 sequences and MiXCR eliminates the sequences from further processing. For this reason, 36% of the reads

are unaligned, and hence no unassigned DH and JH genes are seen in MiXCR annotation output. Adjustments of parameters can be made for MiXCR to achieve 98% - 100% alignment of the reads, but this will not improve its accuracy but introduce unassigned gene calls from the annotation outputs.

Misidentification of VH genes by annotation tools is mostly as a consequence of homologous VH genes that are similarly related at the nucleotide level. Because these tools allow for partial alignment in determining these VH genes, they are prone to annotating VH genes inaccurately.

When annotating bovine VH genes, we suggest that the frequencies of the IGHV1-33 gene and IGHV1-21 gene be summed up because the sequences of the polymorphic IGHV1-33 gene cannot be differentiated from those of the IGHV1-21 gene. Therefore, for all sequences described as alleles of the IGHV1-21 gene, it cannot be excluded that some of these alleles belong exclusively to the IGHV1-33 gene.

While using IgDiscover for allele discovery from this study, determining the number of iterations to run is depended on the species. Higher iterations, compared to that required for human repertoire data (Corcoran et al., 2016), were required to achieve a plateau, where no more individualized VH alleles are discovered, as assessed using pre-germline outputs per iteration. However, a plateau was reached sooner when assessed using the final germline filtered outputs, such as seen in the Ndama breed (Figure 10). The early plateau in the latter stage can be attributed to filtering applied to the pre-germline output. The criteria implemented to this filtering may need further adjustment to permit a more balanced treatment between species that are closely and distantly related to the input germline database.

The germline allele prediction tools used utilize expressed IgM sequences for prediction. IgDiscover and TiGGER used for this discovery, enable prediction of novel germline alleles from BCR sequences. These tools remodel BCR data in predicting the germline alleles of a species. This task, however, is best achieved through sequencing of genomic DNA. Another caveat to this analysis is that the data used was from adult animals and therefore their mature

antibody repertoire, being riddled with somatically mutated BCRs, isn't ideal for predicting novel germline alleles. In this study, additional analysis using naïve datasets (results not shown), and it was observed that, even with IgDiscover the number of iterations to make is a few (2 iteration was enough to predict novel alleles). Increasing the number of iterations resulted in discovery of the same number of alleles predicted (data not shown). Mature antibodies repertoires tend to have undergone somatic hypermutations, and for this reason to achieve a plateau phase in predicting novel allele, one has to run more iteration to achieve the goal. The source of these mutations is somatic hypermutation. Once an antibody is exposed to numerous antigens over some time mutation occurs to achieve affinity maturation.

The distribution of pairwise Hamming distances of novel germline alleles discovered by both tools resulted in a precise understanding of genetic diversity between novel germline alleles from the selected African breeds and Friesian breed. African novel germline alleles were seen to be more diverse compared to Friesian breed. Ndama among the African breeds largely contributed to more genetic diversity seen in the selected African breeds. From these pairwise Hamming distances seen in the novel alleles outputs of IgDiscover and TIgGER, it is observed that TIgGER predicted novel alleles that are distant apart compared to novel alleles from IgDiscover.

Several tools yielded concordant discoveries of some putative new germline alleles. In Boran, IGHV1-17*01_S7671 (IgDiscover) was identical to IGHV1-21*02 (TIgGER), and IGHV1-27*01_S4561 (IgDiscover) was identical to IGHV1-30*01 (TIgGER) at the nucleotide level. These alleles have different "parent" genes yet identical. This demonstrates how closely the bovine VH genes are to each other related.

Further studies are needed to refine the annotation of the IGHV1-7 gene. This gene is thought to give rise to ultra-long antibodies. Ultra-long CDR H3 antibodies can range from forty to over seventy highly diverse amino acid length and form unique β -ribbon 'stalk' and disulfide-

bonded 'knob' structures (Wang et al., 2013). Because of these unique structures, annotation tools used in this study don't annotate this gene correctly.

5.2 Conclusions and Recommendations

In this project, benchmarking workflow was designed for bovine B-cell annotation tools based on frequencies of correct and incorrect antibody calls and their distribution using a simulated bovine dataset. Examining the performance of these tools using these metrics has given an insight into the level of annotation accuracy in each software. For VH gene annotation, IMGT/High-V-QUEST and IgBlast performed better than MiXCR, whereas MiXCR annotated JH genes better compared to the other two tools. More work is still needed to improve annotation accuracy.

The identification of germline alleles revealed immunological diversity. The genetic diversity observed in African bovine breeds is yet to be documented in the online germline database in resources like IMGT. IgDiscover identified 18 novel alleles from Boran, 6 from Ndama, and 3 from Ankole. TIgGER, on the other hand, discovered 7 novel alleles from Boran, 18 from Ndama, and 1 from Ankole breed. In using pairwise Hamming distances, the selected African breeds recorded high diversity. However, this study did not investigate more samples of African bovine breeds to examine this germline diversity.

There is a need for more studies of germline allele discovery to be done in African breeds to study this diversity as a result characterizing of African breeds to get an authentic idea of genetic diversity in African bovine immunoglobulin.

REFERENCES

- Berens, S. J., Wylie, D. E., & Lopez, O. J. (1997). Use of a single V(H) family and long CDR3s in the variable region of cattle Ig heavy chains. *International Immunology*, 9(1), 189–199. <https://doi.org/10.1093/intimm/9.1.189>
- Brand, A. H., & Livesey, F. J. (2011). Neural Stem Cell Biology in Vertebrates and Invertebrates: More Alike than Different? *Neuron*, 70(4), 719–729. <https://doi.org/10.1016/j.neuron.2011.05.016>
- Bolotin, D. A., Poslavsky, S., Mitrophanov, I., Shugay, M., Mamedov, I. Z., Putintseva, E. V., & Chudakov, D. M. MiXCR: software for comprehensive adaptive immunity profiling. *Nature Methods*, 12(380-381).2015
- Butler, J. E. (1998). Immunoglobulin diversity, B-cell and antibody repertoire development in large farm animals. *OIE Revue Scientifique et Technique*, 17(1), 43–70. <https://doi.org/10.20506/rst.17.1.1096>
- Collins, A. M., & Watson, C. T. (2018). Immunoglobulin light chain gene rearrangements, receptor editing and the development of a self-tolerant antibody repertoire. *Frontiers in Immunology*, 9(OCT), 1–12. <https://doi.org/10.3389/fimmu.2018.02249>
- Chudakov, D. M. (2015). MiXCR: Software for comprehensive adaptive immunity profiling. *Nature Methods*, 12(5), 380–381. <https://doi.org/10.1038/nmeth.3364>
- Chaudhary, N., & Wesemann, D. R. (2018). Analyzing immunoglobulin Repertoires. 9(March), 1–18. <https://doi.org/10.3389/fimmu.2018.00462>
- Cohen, S., & Cohen, S. (1962). Antibody structure. *J. clin. Path.*, 28, Suppl. (Ass. Clin. Path.), 6, 1-745
- Corcoran, M. M., Phad, G. E., Bernat, N. V., Stahl-Hennig, C., Sumida, N., Persson, M. A. A., ... Hedestam, G. B. K. (2016). Production of individualized v gene databases

- reveals high levels of immunoglobulin genetic diversity. *Nature Communications*, 7. <https://doi.org/10.1038/ncomms13642>
- Deiss, T. C., Vadnais, M., Wang, F., Chen, P. L., Torkamani, A., Mwangi, W., ... Smider, V. V. (2017). Immunogenetic factors driving formation of ultralong VH CDR3 in *Bos taurus* antibodies. (August), 1–12. <https://doi.org/10.1038/cmi.2017.117>
- Gadala-Maria, D., Yaari, G., Uduman, M., & Kleinstein, S. H. (2015). Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. 1–9. <https://doi.org/10.1073/pnas.1417683112>
- Gupta, N. T., Vander Heiden, J. A., Uduman, M., Gadala-Maria, D., Yaari, G., & Kleinstein, S. H. (2015). Change-O: A toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics*, 31(20), 3356–3358. <https://doi.org/10.1093/bioinformatics/btv359>
- Haakenson, J. K., Huang, R., Smider, V. V., Dunn-walters, D. K., Burrows, P. D., Hammond, J. A., & Smider, V. V. (2018). Diversity in the cow Ultralong cDr H3 Antibody repertoire. 9(June), 1–10. <https://doi.org/10.3389/fimmu.2018.01262>
- Hoffman, W., Lakkis, F. G., & Chalasani, G. (2016). B cells, antibodies, and more. *Clinical Journal of the American Society of Nephrology*, 11(1), 137–154. <https://doi.org/10.2215/CJN.09430915>
- Huang, W., Li, L., Myers, J. R., & Marth, G. T. (2012). ART: A next-generation sequencing read simulator. *Bioinformatics*, 28(4), 593–594. <https://doi.org/10.1093/bioinformatics/btr708>
- Kirik, U., Greiff, L., Levander, F., & Ohlin, M. (2017). Parallel antibody germline gene and haplotype analyses support the validity of immunoglobulin germline gene inference

- and discovery. *Molecular Immunology*, 87, 12–22.
<https://doi.org/10.1016/j.molimm.2017.03.012>
- Ichihara, Y., Matsuoka, H., & Kurosawa, Y. (1988). Organization of human immunoglobulin heavy chain diversity gene loci. *The EMBO Journal*, 7(13), 4141–4150.
<https://doi.org/10.1002/j.1460-2075.1988.tb03309.x>
- Lefranc, M. P., Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Bellahcene, F., ... Duroux, P. (2009). IMGT®, the international ImMunoGeneTics information system®. *Nucleic Acids Research*, 37(SUPPL. 1), 1006–1012. <https://doi.org/10.1093/nar/gkn838>
- Neuberger, M. S. (1992). Antigen receptors. *Research in Immunology*, 143(8), 846–848.
[https://doi.org/10.1016/0923-2494\(92\)80102-Q](https://doi.org/10.1016/0923-2494(92)80102-Q)
- Oprea, M. (1999) Antibody Repertoires and Pathogen Recognition: Archived 2008-09-06 at the Wayback Machine The Role of Germline Diversity and Somatic Hypermutation (Thesis) University of Leeds.
- Retter, I., Althaus, H. H., Münch, R., & Müller, W. (2005). VBASE2, an integrative V gene database. *Nucleic Acids Research*, 33(DATABASE ISS.), 671–674.
<https://doi.org/10.1093/nar/gki088>
- Sadofsky, M. J. (2001). The RAG proteins in V(D)J recombination: more than just a nuclease. *Nucleic Acids Research*, 29(7), 1399–1409. <https://doi.org/10.1093/nar/29.7.1399>
- Shojaei, F., Saini, S. S., & Kaushik, A. K. (2003). Unusually long germline DH genes contribute to large sized CDR3H in bovine antibodies. *Molecular Immunology*, 40(1), 61–67. [https://doi.org/10.1016/S0161-5890\(03\)00098-1](https://doi.org/10.1016/S0161-5890(03)00098-1)
- Smakaj, E., Babrak, L., Ohlin, M., Shugay, M., Briney, B., Tosoni, D., ... Lees, W. (2019). Sequence analysis Benchmarking immunoinformatic tools for the analysis of antibody repertoire sequences. (December), 1–9. <https://doi.org/10.1093/bioinformatics/btz845>

- Sok, D., Le, K. M., Vadnais, M., Saye-francisco, K., Joseph, G., Torres, J., ... Burton, D. R. (2018). HHS Public Access. 548(7665), 108–111. <https://doi.org/10.1038/nature23301>.Rapid
- Stanfield, R. L., Haakenson, J., Deiss, T. C., Criscitiello, M. F., Wilson, I. A., & Smider, V. V. (2018). The Unusual Genetics and Biochemistry of Bovine Immunoglobulins. In *Advances in Immunology* (1st ed., Vol. 137). <https://doi.org/10.1016/bs.ai.2017.12.004>
- Jackson, K. J. L., Liu, Y., Roskin, K. M., Glanville, J., Hoh, R. A., Seo, K., ... Boyd, S. D. (2014). Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements. *Cell Host and Microbe*, 16(1), 105–114. <https://doi.org/10.1016/j.chom.2014.05.013>
- Jagannathan-Bogdan, M., & Zon, L. I. (2013). Hematopoiesis. *Development* (Cambridge), 140(12), 2463–2467. <https://doi.org/10.1242/dev.083147>
- Victora, G. D., & Nussenzweig, M. C. (2012). Germinal centers. *Annual Review of Immunology*, 30, 429–457. <https://doi.org/10.1146/annurev-immunol-020711-075032>
- Wendel, B. S., He, C., Crompton, P. D., Pierce, S. K., & Jiang, N. (2017). A streamlined approach to antibody novel germline allele prediction and validation. *Frontiers in Immunology*, 8(SEP). <https://doi.org/10.3389/fimmu.2017.01072>
- Wirtz, C., & Sayer, D. (2014). Chapter 6 Data Analysis of HLA Sequencing Using Assign-SBT (Vol. 882). <https://doi.org/10.1007/978-1-61779-842-9>
- Yaari, G., & Kleinstein, S. H. (2015). Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Medicine*, 1–14. <https://doi.org/10.1186/s13073-015-0243-2>

Ye, J., Ma, N., Madden, T. L., & Ostell, J. M. (2013). IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Research*, 41(Web Server issue), 1–7.
<https://doi.org/10.1093/nar/gkt382>