NOVEL METHODS OF REPERTOIRE DEVELOPMENT IN THE LYMPHOCYTE ANTIGEN RECEPTORS OF *BOS TAURUS* AND *GINGLYMOSTOMA CIRRATUM*

A Dissertation

by

THADDEUS CHARLES DEISS

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Michael F. Criscitiello
Committee Members,	Luc R. Berghman
	Scott V. Dindot
	Christabel J. Welsh
Head of Department,	Ramesh Vemulapalli

August 2019

Major Subject: Biomedical Sciences

ABSTRACT

The adaptive immune system of all jawed vertebrates relies on B and T cell effectors bearing immunoglobulin superfamily (IGSF) antigen receptors called B and T cell receptors (BCR, TCR). These antigen receptors are borne from distinct loci requiring genomic rearrangement of variable (V), diversity (D), and joining (J) gene segments, allowing a near infinite repertoire to be encoded within confined loci. Gene rearrangement events, carried out by the recombination activating genes (RAG), massively expands diversity in complementary determine regions (CDRs), loops forming the receptor-ligand interface. IGSF receptors can also be altered by activation induced cytidine deaminase (AID) activity to hone response specificity or to expand the primary repertoire.

Bovine BCRs undergo AID mediated somatic hypermutations (SHM) to diversify the naïve repertoire. The *Bos taurus* antibody repertoire is characterized by a subset of immunoglobulin heavy (IgH) receptors encoding ultralong CDR3 regions greater than forty amino acid residues in length. Ultralong CDR3 BCR rearrangements are genetically constructed with a propensity to yield Cys residues when mutated. The Cys residues are integral in facilitating the unique ‘knob’ domain which protrudes from a β-ribbon stalk. This protruding knob is held far from the other CDRs and forms the primary paratope of the molecule. This miniscule paratope can be further honed by AID mediated deletion events conferring unique structural motifs compared to full length knobs.

The TCR repertoire of *Ginglymostoma cirratum* employs Ig-TCRδ rearrangements that blend gene segments from distinct BCR and TCR loci. These receptors make use of IgH V segments to diversify the existing TCRδ repertoire. Sequencing of the Ig-TCRδ repertoire identified these rearrangements stemmed from a number of IgH mini-cluster loci as well as a distinct lineage of Ig-like V segments located within the TCRδ locus termed TAILV segments. Quantification of Ig-TCRδ transcripts determined they were just as common as canonical TCRδ receptors in this ancient iteration of adaptive immunity, and could be found localized on the cell surface presumptively playing an active role in shark immunity.

These novel methods of repertoire development showcase the malleability of the IGSF system in developing a massively diverse repertoire to protect an organism from the antigenic onslaught of life.

DEDICATION

This dissertation is dedicated to my grandparents Freddy and Peggy Hillje and the Late Sonja Cannon, parents Laura and Stephen Deiss, dogs Ziggy and Zoey, and Dr. Jean and Dr. Ken Escudero who each provided unique facets of support in reaching all my goals.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Michael Criscitiello, for allowing me the opportunity to work on a multitude of fascinating projects during my tenure as a graduate student. I would also like to thank my committee members, Dr. Luc Berghman, Dr. Scott Dindot, and Dr. Jane Welsh for their wonderful advice, direction, and the spirited debates during our annual committee meetings. Also thanks to Dr. Rob Miller for agreeing, on short notice, to serve as an external examiner during my impending defense.

I could not have accomplished much without the stellar tutelage of Dr. Jean and Ken Escudero who initiated my goal to obtain a PhD and provided me with an opporitinity to begin honing my skills as an undergraduate researcher at the University of Texas A&M Kingsville. They, along with since retired Criscitiello lab technician Patricia Chen, provided a solid background for all the lab techniques I am currently capable of. I also thank my lab mates especially Breanna Breaux and Jeannine Ott for often providing an extra set of hands when I needed something removed from an incubator, or for handling and processing orders as paperwork was never my strong suit. Jenna Harrison the current lab technician in the Criscitiello lab for lightening my load by taking over the autoclaving duties and performing sequencing experiments that contribute to the ongoing research on shark TCRs. New Criscitiello lab members Kelly Head and Brooke Norwood have also provided an opportunity to step away from the constant data analysis I've buried myself in recently to work on new projects and shake off rust at the bench. Finally I would like to thank Dr. Vaishali Katju for providing me

with a post-doc position so that I may continue honing my skills as a scientist and an academic.

Thanks must also be given to the small army of individuals who have, attempted, to keep me sane over the years. Blake Korenek and Heather Vick have remained close and supportive friends since childhood that never let me go too long without checking in. They have never hesitated to lend an ear when I needed it most and have saved myself from myself on numerous occasions. Russel Graves and Christopher and Shelby Raun always make a trip during the football season and, although we may have a bit too much fun together, I always look forward to those late Friday calls requesting a place to sleep for the night. Ryan Drum and Avery and Morgan Goldsmith, formed an amazing group of roommates and friends during my undergraduate years. The stupid impromptu games we would create to pass time before we could afford internet or cable were rivaled only by debates on pretty much any subject simply for the sake of expanding our collective viewpoints. Ryan and I have spent more time than I can recollect in Kingsville trees and that time remains one my favorite memories. Dr. Megha Bijalwan for acting as something of a mother figure for me in graduate school, keeping me out of trouble to the best of her ability. Dr. Shenaz Lokhandwala has been and remains the closest friend I've ever had and I treasure the time I was fortunate enough to spend with her.

I would like to thank my family most of all. My parents, Laura and Stephen Deiss, deserve the utmost praise for providing the genetics responsible for my intelligence and they have always provided unparalleled support in all my endeavors. They've spent countless hours in ER waiting rooms and surgical wards in addition to

dealing my admittedly questionable behavior at times. They have bailed me out of so many situations listing them would double the length of this document and I love them both dearly. My grandparents Peggy and Freddy Hillje make sure I'm well fed through yearly deer hunts on their ranch. My brother, Travis Deiss, who argues for argument's sake, taught me the invaluable skill expressing ideas in multiple ways to get my point across. Finally I have to offer thanks to my two dogs Ziggy and Zoey because the time we spend on our daily walks has always been as much about me processing my day and formulating ideas as it was about giving them exercise.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by my dissertation committee consisting of my advisor Dr. Michael F. Criscitiello, committee member Dr. C. Jane Welsh of the Department of Veterinary Integrated Biosciences, Dr. Luc R. Berghman in the Department of Poultry Science and Dr. Scott V. Dindot of the Department of Genetics. Dr. Waithaka Mwangi of the Department of Diagnostic Medicine at the Kansas State Univeristy also temrporarily served on my committee during his tenure at Texas A&M University. Dr. Robert D. Miller of the University of New Mexico Department of Biology will serve as an external advisor during my dissertation defense.

The data presented in Chapter 2 was performed in collaboration with Dr. Vaughn V. Smider in the Department of Molecular Medicine at the Scripps Research Institute and was published in the Journal of Cellular and Molecular Immunology in 2017. The experiments forming the data presented in Chapter 3 were conducted at the University of Maryland School of Medicine in collaboration with Dr. Martin F. Flajnik and Dr. Yuko Ohta in the Department of Microbiology and Immunology. Jeannine Ott performed the experiments confirming expression of the TAILV lineage segments in the nurse shark TCRδ repertoire. The manuscript forming Chapter 3 is currently under review at the Journal of Immunology.

All other work conducted for the dissertation was completed independently.

Funding Sources

Graduate study was supported by a fellowship from Texas A&M University and a dissertation research fellowship from the National Science Foundation and the Texas A&M University College of Veterinary Medicine Merit Fellow Scholarship.

This work was also made possible in part by National Institutes of Health under Grant Numbers GM-105826-01, AI-120791, AI-56963, and AI-027877 and the NSF under Grant Numbers IOS1257829 and IOS-1656870. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the VTPB.

NOMENCLATURE

AID	Activation induced cytidine deaminase
APC	Antigen presenting cell
BCR	B cell receptor
BER	Base excision repair
CDR	Complementary determining region
CLR	C-type lectin receptors
CSR	Class switch recombination
DSB	Double stranded DNA break
DAMPs	Damage associated molecular patterns
dsDNA	Double stranded DNA
GC	Germinal center
IgH	Immunoglobulin heavy chain
IGSF	Immunoglobulin superfamily
IgL	Immunoglobulin light chain
MHC	Major histocompatibility complex
MMR	Mismatch repair
NLR	Nucleotide-binding oligomerization domain-like receptors
NHEJ	Non-homologous end joining
PRR	Pattern recognition receptor
PAMPs	Pathogen associated molecular patterns
RAG	Recombination activating genes

RLR	Retinoic acid-inducible I-like receptors
RSS	Recombination signal sequence
SHM	Somatic hypermutation
TCR	T cell receptor
TLR	Toll-like receptors
VLR	Variable lymphocyte receptor

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION

1.1 Immunity: innate vs. adaptive

The ability of an organism to defend against pathogens is integral to survival. The system in place to deter pathogens, the immune system, can be segregated into two distinct branches in vertebrates: innate immunity and adaptive immunity. Innate immunity provides a broad, unchanging initial line of defense and surveillance. Physical barriers imparted by tight epithelial junctions and mucosal secretions containing antimicrobial peptides act as the very first line of defense [1]. Acute phase proteins (APP) and the complement system are serum proteins with multifaceted roles that range from induction of local inflammation, to opsonizing phagocytic responses, to lysing microbial cells [2, 3]. The receptors employed by effector cells of the innate system are germline hardcoded [4]; termed pattern recognition receptors (PRRs) due to their recognition of highly conserved pathogen or danger associated molecular patterns (PAMPs and DAMPs) [5, 6]. PRRs are found throughout metazoan lineages and include transmembrane receptor families, Toll-like receptors (TLRs) [7] and C-type lectin receptors (CLRs) [8], and cytoplasmic receptor families, retinoic acid-inducible I-like receptors (RLRs) [9] and nucleotide-binding oligomerization domain-like receptors (NLRs) [10]. The PAMPs targeted by PRRs include molecules vital to a pathogen such as bacterial lipopolysaccharides (LPS), fungal β-glucans, and viral double stranded RNA (dsRNA) [4]. While the innate branch is highly effective in rapidly detecting and

deterring broad classes of pathogens, response kinetics remain largely identical for each instance of pathogen confrontation.

The two distinct lineages of immunoglobulin superfamily based adaptive immunity are B cells and T cells, employing immunoglobulin (Ig), also called B cell receptors (BCR), and T cell receptors (TCR) respectively. While PRRs are germline hardcoded, BCRs and TCRs require somatic rearrangement of variable (V), diversity (D) and joining (J) gene segments (a process known as V(D)J recombination [11]) to form a functional receptor gene. This rearrangement results in a diverse repertoire of BCR and TCRs, each specifically recognizing a unique epitope in opposition to the broad molecule classes recognized by PRRs. However, both B and T cells of the adaptive immune system rely on innate effector cells for cellular activation and initiation of responses. Firstly, whereas innate immunity is constitutively located throughout an organism, adaptive immune cells reside in specialized secondary lymphoid tissue awaiting activation [12]. In order be activated a lymphocyte requires multiple signals. Signaling through the rearranged antigen receptor, termed signal one, followed by a myriad of cytokine signals and activation of costimulatory receptors, known as signal two [13]. Generation of signal two is carried out by antigen presenting cells (APCs) of the innate system. Upon recognition of PAMPs via PRRs, professional APCs, such as dendritic cells (DCs) [14], migrate to secondary lymphoid tissues and begin expressing co-stimulatory molecules and secrete cytokines to activate B and T cells. However, B cells prominently rely on T cell help for class switch and affinity matured responses, and being phagocytic are capable of acting as APCs themselves [15, 16]. The cytokines and

co-stimulatory molecules APCs and helper T cells (T_H) provide not only activate adaptive effector cells, but also manipulate their responses to best suit the situation. In summation, the highly specific adaptive immune system relies on innate effectors to combat pathogens. Furthermore, as previously stated naïve T and B cells are generally not involved in immune surveillance and initially interact with pathogens in specialized tissues relatively remote from the site of infection.

In contrast, the hallmarks of the adaptive immune system are memory and specificity, allowing responses of amplified vigor upon multiple confrontations with the same pathogen. Recent discoveries, variable lymphocyte receptors (VLRs) of jawless fish and CRISPR-Cas of bacteria, provide evidence that the hallmarks of adaptive immunity could arguably be applied to systems found in simpler organisms [17, 18]. However the opus of adaptive immunity, the lymphoid progenitor-immunoglobulin superfamily (IGSF) receptor-major histocompatibility complex (MHC) based system utilized by higher vertebrates, originating 450 million years ago in chondrichthyes, cartilaginous fish, will be the focus of this dissertation [19].

1.2 The immunological "Big Bang" origins of highly diverse IGSF receptors

The origin of IGSF-based adaptive immunity dates back to the ancestor of modern gnathastomes with the insertion of the recombinase activating gene (RAG) transposon into the primordial antigen receptor locus [20, 21]. The immunological "big bang", brought on by insertion of the RAG transposase, enabled antigen receptors of jawed vertebrates to be generated through somatic recombination of V(D)J segments

[11, 22]. The proteins encoded by RAG1 and RAG2 bind to conserved nucleotide motifs known as recombination signal sequences (RSS) flanking genomic V, D, and J gene segments [23]. Composition of the VDJ flanking RSSs include conserved heptamer (CACAGTG) and nonamer (ACAAAAACC) nucleotide sequences separated by a 12 or 23 bp spacer with little to no conservation [20]. The restriction of RAG proteins to selectively target segments flanked by 12 spacer RSS for recombination to a 23 spacer RSS is termed the 12/23 rule [24]. The 12/23 rule helps to ensure proper ordering of gene segments for recombination. This requires V segments, flanked by a 23bp RSS, to rearrange to a D segment, flanked on both sides by a 12bp RSS, to facilitate rearrangement to a J segment also flanked by a 23bp RSS (Fig 1.1).

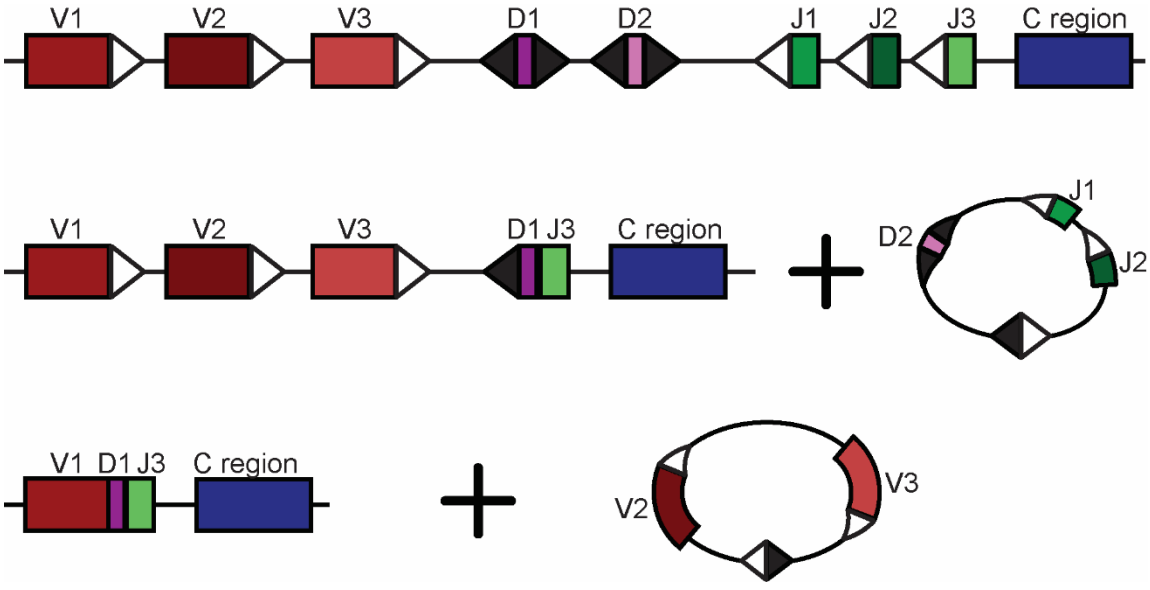


Figure 1.1 RAG diversification of IGSF receptors. An example antigen receptor locus in translocon arrangement is displayed at the top. The V(D)J gene segments are each flanked by an RSS with either a 12 bp spacer (black triangles) or a 23 bp spacer (while triangles). The first recombination step is the fusion of D-J (middle, with excised intervening genomic DNA), which is followed by the fusion of V-DJ (bottom) to form a functional antigen receptor gene. Rearrangement is required to facilitate the splicing of the rearranged VDJ domain to the C region exons located downstream.

During recombination, the RAG proteins bind the RSSs bringing the V-D, V-J, or D-J segments into proximity. Once the segments selected for recombination are juxtaposed a nick is introduced to the signal-coding junction, at the heptamer of the RSS, resulting in a 3'-OH nucleophilic attack on the opposing strand, causing a double stranded DNA (dsDNA) break (DSB), leaving hairpin ends at the coding joints and 5' phosphorylated blunt ends at the signal joint [25]. A heterodimer of Ku70:Ku80 binds free dsDNA ends at both the signal and coding joint [26]. In addition to keeping the dsDNA ends juxtaposed, the Ku70:Ku80 complex is integral to recruiting downstream non-homologous end-joining (NHEJ) machinery to DSB sites [27]. The 5' phosphorylated blunt signal ends are summarily ligated by recruitment of X-ray cross-complementation group 4 (XRCC4) and DNA ligase IV (X4-LIV) and the intervening DNA is inverted or removed from the genome in an excision circle [27]. The hairpins present in the coding ends must be opened before the X4-LIV, NHEJ machinery can resolve the DSB [28]. The complex responsible for opening hairpin loops, composed of Artemis and DNA-dependent protein kinase catalytic subunit (Artemis:DNA-PK_{CS}), possesses both endonuclease and exonuclease activity [29]. Cleavage of hairpins by Artemis:DNA-PK_{CS} is imprecise, not always at the hairpin apex, often leaving single strand DNA (ssDNA) overhangs, preferentially 3' overhangs, in the coding joints [29, 30]. ssDNA overhangs can be removed, through Artemis:DNA-PK_{CS} activity, or can be incorporated, following ligation of the coding ends, via the gap filling polymerases pol μ or pol λ [31, 32]. Nucleotides incorporated into the coding junction arising from Atremis:DNA-PK_{CS} overhangs are termed P nucleotides as the insertions are

palindromic [31]. In addition to P nucleotide insertions, the non-template DNA polymerase, terminal deoxynucleotidyl transferase (TdT), adds random nucleotides to the coding junction termed N nucleotides (Fig 1.1e) [33]. Together the removal of nucleotides, which can extend into the coding segment ends, and addition of N/P nucleotides generate vast amounts of junctional diversity even in recombination events pairing identical V(D)J segments [34]. A step-by-step illustration of RAG recombination and the generation of junctional diversity is displayed in Fig1.2. Non-template junctional diversity compounds the random selection V(D)J of segments in recombination to produce a highly diverse arsenal of antigen receptor genes.

The process of V(D)J recombination is required for the formation of functional antigen receptor genes. Defects in genes involved in the V(D)J recombination pathway result in severe combined immunodeficiency disease (SCID) resulting in depletion, or complete ablation, of B and T lymphocyte populations [35].

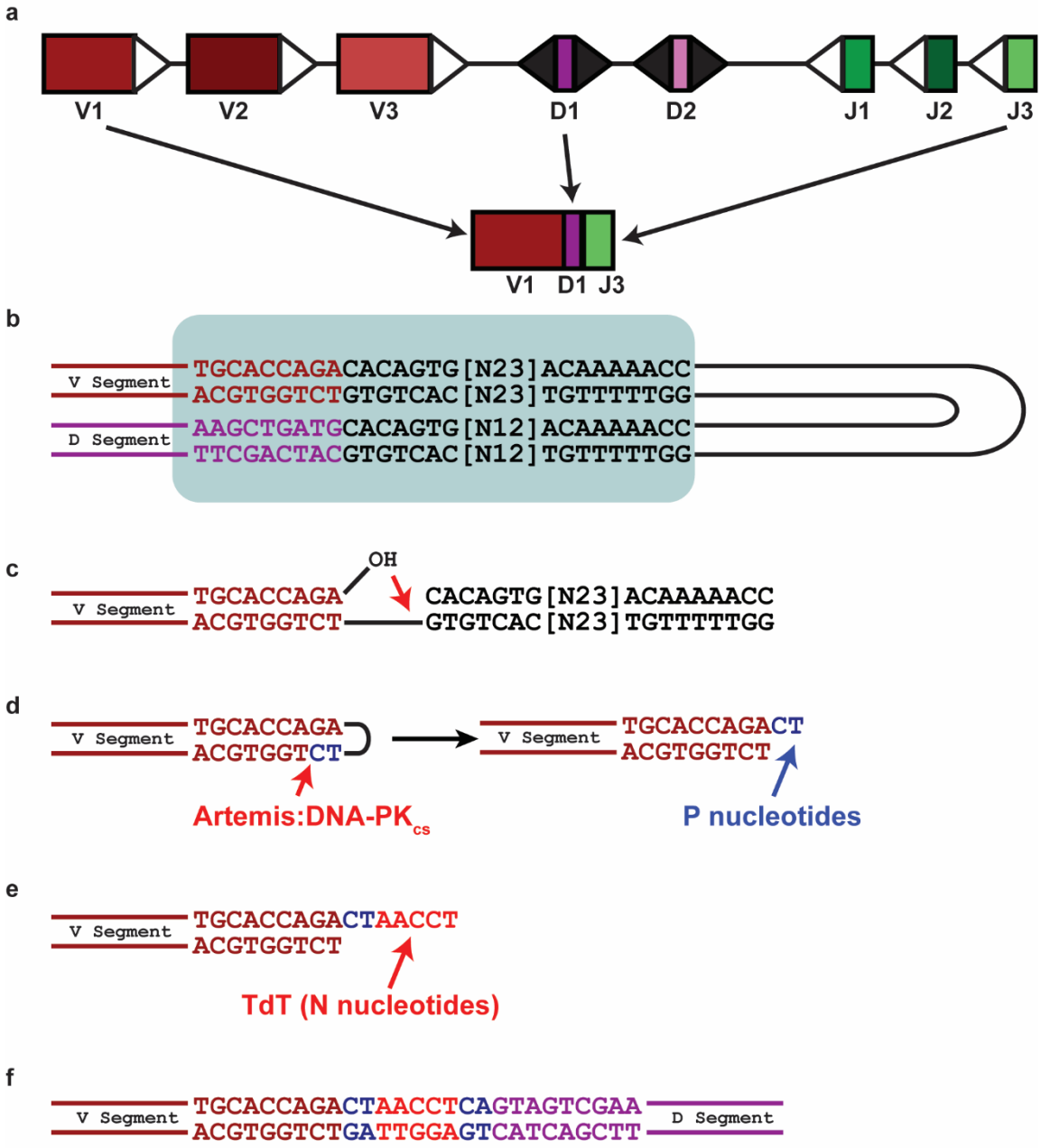


Figure 1.2 Generation of junctional diversity between coding ends during recombination. a) Random selection of V(D)J segments selectively recombined governed by the 12/23 rule. **b)** Two segments are brought into proximity by the RAG complex and aligned at the RSS. **c)** RAG activity introduces a nick at the heptamer of the RSS causing a nucleophilic attack. **d)** The nucleophilic attach leaves a hairpin at the coding end that must be opened by Artemis:DNA-PK, imprecise hairpin opening yields P nucleotides (blue). **e)** TdT adds non-template bases to the junction yielding N nucleotides (red). **f)** Example of the junctional diversity between the two fused coding segments.

1.3 Distinct BCR and TCR lineages, shared prototypical IGSF structure

The overwhelming majority of IGSF adaptive immune receptors are heterodimeric, with the gene for each chain produced by V(D)J recombination at separate, distinct genomic loci [36]. Typical antigen receptor loci are arranged as a translocon with blocks of multiple V, D, and J segments respectively upstream of germline hardcoded constant (C) region exons [37]. BCRs are composed of two identical heavy chains (IgH), originating from a single IgH locus, and two identical light chains (IgL), which are classified by originating locus into Igκ, Igλ, and Igσ chains [38]. There are two classes of TCR heterodimer which rearrange either an α and β chain (αβ T cells) or γ and δ chain (γδ T cells) [39]. In both BCR and TCR lineages one chain rearranges V, D and J segments (IgH, TCRβ, and TCRδ) while the partner chain rearranges V and J segments only (IgL, TCRα, TCRγ) [39].

The chains of BCRs and TCRs are composed of multiple immunoglobulin fold domains, classified as variable (V-domain) or constant (C-domain). The prototypical immunoglobulin domain, or fold, consists of two anti-parallel β-pleated sheets with conserved Cys residues forming an intra-domain disulfide bond to create a β-sandwich [40]. Only the membrane distal Ig domain, the V-domain, is diversified by RAG mediated V(D)J recombination [41]. The receptor-ligand interface for all IGSF antigen receptors is formed by the V-domains alone. The unique specificity of each rearranged V-domain is conferred by three complementarity determining regions (CDRs), hypervariable loops projecting from the core immunoglobulin fold [42]. The first two CDRs are encoded within germline V segments, while CDR3, the longest and most

diverse CDR, spans the length of the V(D)J junction produced by somatic rearrangement [41]. Therefore, each V(D)J recombination event yields a CDR3 signature unique to each receptor. A complete BCR or TCR heterodimer contains two juxtaposed V-domains with a total of six CDRs forming the paratope of the receptor.

The C-domains, encoded by C region exons, differ in number according to the receptor lineage and class. All TCR and IgL chains possess a single C-domain, while mammalian IgH chains possess a minimum of three C-domains [43]. In addition to the intra-domain disulfide bonds, the C-domains contain additional Cys residues to facilitate inter-chain disulfide bonding. These inter-chain disulfide bonds link every heterodimeric chain to its partner—IgH to IgL, TCRα to TCRβ, and TCRγ to TCRδ—as well as linking the IgH homodimer chains to one another. The C-domains of the secretory BCRs, called antibodies when secreted, determine the effector function of the molecule [44].

The 3’ C region exon of all chains, except for IgL, encodes a transmembrane (TM) domain to anchor the receptor to the cell membrane. The cytoplasmic domain of both BCR and TCR possess no intrinsic signaling capabilities. Signal transduction requires the association of an ensemble of molecules containing immunoreceptor tyrosine-based activation motifs (ITAMs) which are specific to formation of BCR or TCR complexes [45]. The signaling complex of a BCR consists of an Igα:Igβ heterodimer with an ITAM on each chain [46]. The TCR signaling complex is comprised of three dimeric molecules; CD3γ-CD3ε and CD3δ-CD3ε heterodimers possess a single ITAM on each chain while the ζζ homodimer contains three ITAMs on each chain [47]. Phosphorylation of ITAMs, following BCR or TCR activation through recognition of

cognate antigen, kickstarts a signaling cascade that ends with transcription of specific effector molecules and proliferation of antigen specific B or T cells.

1.4 Distinct B and T cell effector functions

The executioners of the adaptive immune system, B and T cells, arise from a common hematopoietic lymphoid progenitor. Although B and T lymphocyte lineages employ structurally similar receptors the lineages differ immensely in terms of antigen interaction and effector function. While B cells provide humoral immunity through the secretion of antibodies, T cells mediate cellular immune responses including the secretion of cytokine signals and killing of viral or cancerous cells.

The primary lymphoid tissue for B cell development is variable across species, but is ascribed to the bone marrow in most mammals, bursa of Fabricius in birds, and the Leydig or epigonal organ in Chondrichthyes [48, 49]. The signals driving commitment of lymphoid progenitor cells to a B cell fate are E2A and EBF [50]. Following commitment to the B cell lineage, development is phenotypically segregated into pro-B cell, pre-B cell, immature B cell and mature B cell stages [51]. During the pro-B cell stage the IgH locus is accessible to RAG. Productive VDJ rearrangement of the IgH locus and expression of pre-BCR, composed of the rearranged IgH chains with the surrogate light chain proteins VpreB and λ5, arrests RAG activity and signals progression into the pre-B cell stage [52]. A proliferative burst occurs during the pre-B cell stage following successful expression of pre-BCR prior to rearrangement of IgL [53]. The pre-B stage continues with VJ rearrangement at one of the IgL loci,

progressing to a different IgL locus in the event of non-productive rearrangement at one IgL. Since IgL loci do not contain D segments, autoreactive pairing of IgH and IgL in the pre-B stages can be amended through receptor editing [54]. Receptor editing is unique to IGSF loci housing only V and J segments and allows additional rearrangement of unused V and J segments located upstream or downstream of the deleterious VJ rearrangement. The pre-B stage ends with the production of a complete BCR complex that is not self-reactive and is capable of signal transduction. Mature, naïve B cells progress to a secondary lymphoid tissue (spleen, MALT, GALT, lymph nodes, etc.) and await activation via antigenic stimulation. Effector function of B cells is centered in antibody production and dependent upon the IgH C-domain isotype. Isotypes of IgH include IgM (CHμ) and IgD (CHδ), the first isotypes produced by mature B cells, while IgG (CHγ), IgE (CHε), and IgA (CHα) confer more specific effector functions [44]. The first antibody secreted by mature B cells, pentameric IgM, is a low affinity high avidity molecule and potent at fixing complement through the classical activation pathway. The IgG isotype C-domains act as a powerful opsonizing molecule, binding to Fcγ receptors present on multiple leukocyte effector populations, and can facilitate antibody-dependent cellular cytotoxicity (ADCC) for NK cells. IgE isotypes protect against helminths and are involved in allergies and specifically bind the Fcε receptor resulting in degranulation of mast cells, eosinophils and basophils. IgA antibodies are associated with mucosal responses and generally serve as non-inflammatory responders capable of transmission of bound pathogen from lamina propria to lumen via J-chain poly-Ig receptor (pIgR) [55].

Development of T cells occurs in the thymus, a primary lymphoid organ present in all jawed vertebrates. Commitment to the T cell lineage is driven by Notch1 signaling in double negative (DN), with expression of neither CD8 nor CD4 co-receptors, thymocytes which enter the thymic medulla [56]. Following the DN stage the development of αβ and γδ T cells diverges along different pathways. Cells fated to the αβ lineage progress into the thymic cortex during the DN1 and DN2 while rearranging the TCRβ chain [57]. A productive TCRβ rearrangement signifies progression into the DN3 stage where the pre-TCR is assembled using the rearranged TCRβ and pre-Tα [58]. Similar to the pre-BCR stage, a proliferative burst occurs in DN4 following the assembly of the pre-TCR. The next stage in αβ T cell development is called the double positive (DP) stage as developing thymocytes begin to express both CD4 and CD8 co-receptors [56]. Rearrangement of TCRα, and receptor editing, occurs at this stage as DP thymocytes migrate back toward the thymic medulla. Rearrangement of a functional TCRα chain and assembly of the complete αβTCR leads to selection in the cortico-medullary junction. DP thymocytes expressing αβTCR must undergo positive selection where the TCR must recognize self-MHC before being fated to CD4+ or CD8+ single positive mature, naïve αβ T cells [59]. Development of γδ T cells is poorly understood compared to αβ T cell development. However, rearrangement of TCRγ and TCRδ occur simultaneously with TCRβ within DN thymocytes and production of a complete γδTCR with strong Erk signaling appears to fate a thymocyte to the γδ T cell lineage. As γδTCRs are not MHC restricted the only requirement of thymocytes expressing γδTCR need only express receptors that are not autoreactive. The function of γδ T cells are also

poorly understood, but they are not limited solely to interactions with peptide antigens [60]. The effector function of αβ T cells depends on the co-receptor fate at the single positive developmental stage. CD8+ αβ T cells are called cytotoxic T cells as the CD8 co-receptor interacts with MHC class I molecules. The function of cytotoxic T cells, presenting non-self and altered-self peptides on their MHCI, is the killing of virally infected and cancerous cells. CD4+ αβ T cells, interacting with MHC class II, are known as helper T cells (T_H) due to their prominent role in production of cytokines to mediate cellular responses and help other effector cell types. The cytokines produced by T_H cells activate macrophages to kill phagocytized pathogens and are integral to the activation of B cells for germinal center responses.

1.5 AIDing responses via secondary diversification

Within the secondary lymphoid tissue, antigens are encountered by B cells in follicles and can develop germinal centers (GC) where secondary diversifications occur. Activation induced cytidine deaminase (AID), the master regulator of secondary diversification, is involved in somatic hypermutation (SHM), class switch recombination (CSR), and gene conversion (Fig 1.3) [60, 61]. The mechanism of AID diversification involves the deamination of cytosine to uracil (Fig 1.4a) and the subsequent repair steps to remove uracil from genomic DNA.

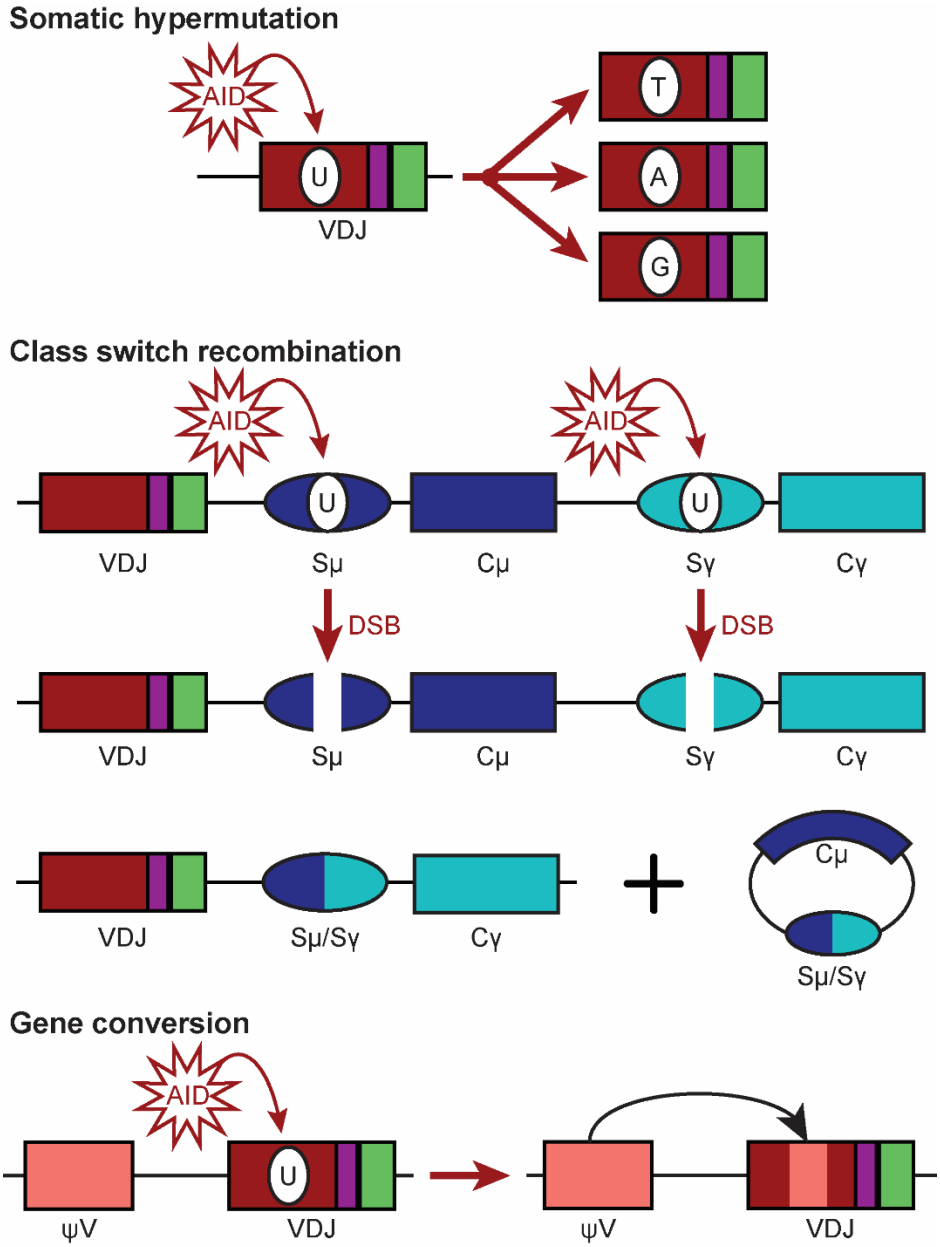


Figure 1.3 Overview of AID mechanisms. The diversification mechanisms catalyzed by AID are shown beginning with the deamination of C to U. Somatic hypermutation (top) results in point mutations within the V-domain which are selected for in a process known as affinity maturation. Class switch recombination (middle) results in the replacement of the C-domain exons in the IgH locus resulting in expression of a different antibody isotype. CSR alters the effector specificity of antibody responses. Gene conversion (bottom) involves the exchange of genetic material between the recombined V segment and an upstream, often pseudogenized donor V segment. This results in a hybridized V-domain bearing sections of both the initial V segment selected in V(D)J recombination and the donor V segment.

The substrate for AID, ssDNA, requires active transcription for the deamination of cytosine to uracil in DNA (Fig 1.4b) [61]. Since uracil is exclusive to RNA, the U site is targeted for repair through either mismatch repair (MMR) or base excision repair (BER) pathways (Fig 1.4c) [62]. The MMR pathway begins with the binding of the MSH2:MSH6 heterodimer to the improper, U:G, base pairing left by AID activity [63]. The binding of the MSH heterodimer recruits the MLH1:PMS2 heterodimer, which interacts with the DNA clamp PCNA [64]. These complexes act in concert to recruit the exonuclease EXO1 to excise bases surrounding the U:G mismatch [63]. The excised lesion is repaired by an error prone DNA polymerase, polη, resulting in potential mutations at A:T sites [65]. The BER repair pathway begins the same as MRR, targeting the U:G mismatch in DNA, however only the uracil present in the DNA is removed by uracil DNA glycosylase (UNG) (Fig A1) [66]. This cleavage results in an abasic site within the DNA sugar-phosphate backbone that must be excised through apurinic/apyramidinic endonuclease (APE) activity [62]. The MMR and BER pathways reconvene with the resolution of excised bases by error prone DNA polymerases, primarily polη. Activity of AID deamination is increased at so-called AID hotspot motifs (WRCY/RGYW). An AID hotspot (red bases in Fig 1.4a) and summary of the mechanism resulting in mutations in and around the hotspot are depicted in Fig 1.4.

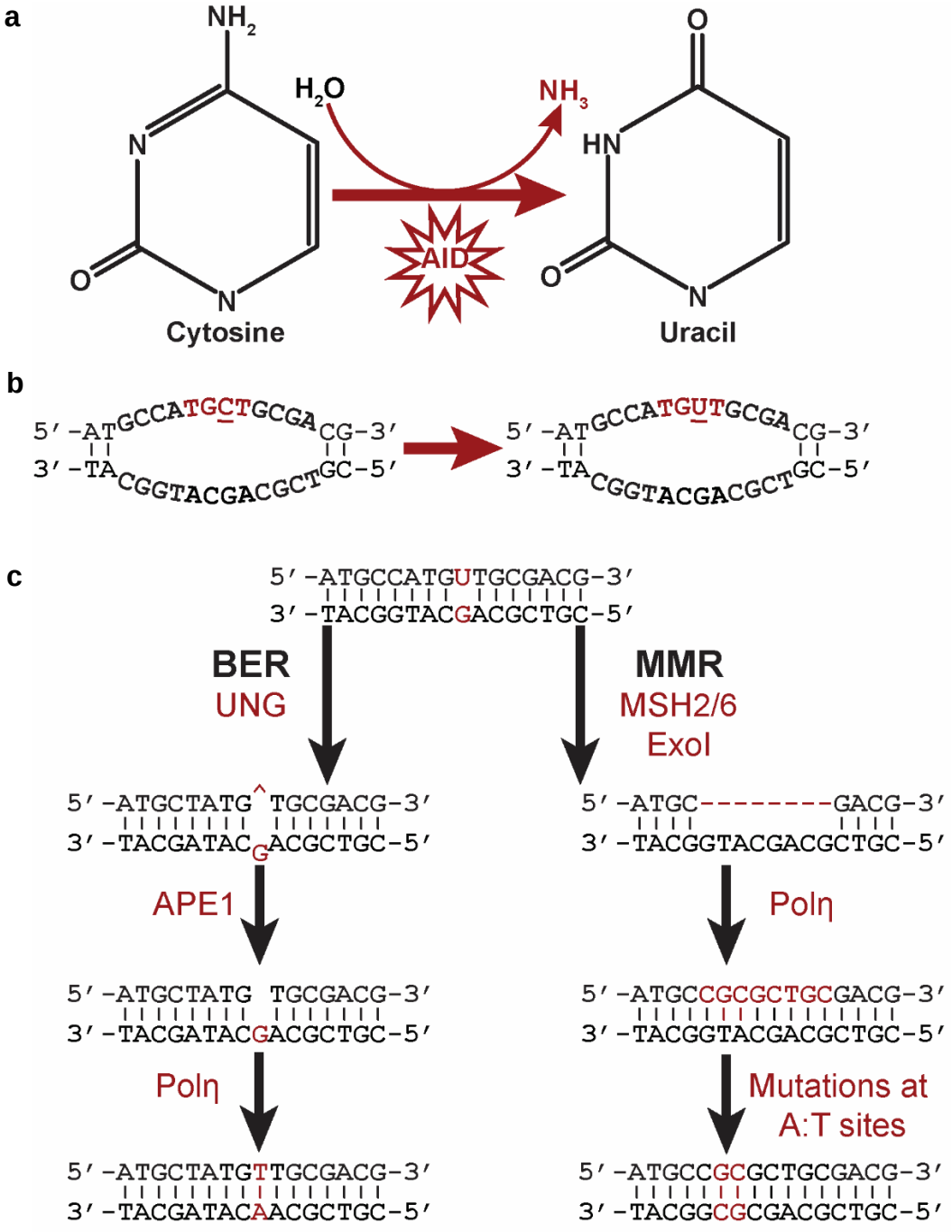


Figure 1.4 Aid catalyzed reaction and resultant repair pathways. a) Reaction depicting AID catalyzed deamination of cytosine to uracil. **b)** AID requires access to ssDNA to catalyze this reaction and preferentially targets cytosine bases in "hotspot" motifs (depicted in red). **c)** Step-by-step mechanism of uracil resolution and repair steps of BER (right) and MMR (left) are depicted with enzyme/complex involvement and associated mode of activity highlighted in red.

The selection of higher affinity antibodies, affinity maturation, achieved through SHM is one of the most anciently conserved function of AID and can be found back to cartilaginous fish. Initially thought to be restricted to mature B-cells that have encountered cognate antigen, AID mediated SHM has been discovered to be a mechanism of primary diversification in naïve B-cells of ruminants [65-67]. Given the generally restricted number of functional V, D, and J segments ruminant species possess, naïve SHM is posited as a mechanism to enhance a limited recombination diversity.

The CSR pathway involves modulating the effector function of an antibody by the changing of the IgH constant region. During CSR double stranded breaks are introduced to the switch (S) regions preceding all IgH C region exon sets (save δ, which employs alternative RNA splicing of a long immature transcript shared with μ) [67]. The S regions are densely populated by AID hotspot motifs to facilitate the DSB. While CSR has been recently posited to exist in Chondrichthyes, it is notably absent in Teleostei, and likely a function of AID perfected in tetrapods [43, 68]. Gene conversion, the targeted replacement of portions of Ig V segments with those of pseudo- or non-functional V genes, has been reported in a growing list of species with a limited IgH V segment repertoire as a method of primary diversification [61, 66, 69].

1.6 Atypical IGSF antigen receptors

While the canonical system of antigen recognition via distinct genetic lineages encoding three CDRs present for each chain; IgH and IgL for Ig/BCR and either αβ or γδ for TCR, several variations have been discovered. A number of BCR paratope variants

including the short form of IgW of cartilaginous fish, IgL-less antibodies in cartilaginous fish and camelids, and the bovine ultralong CDR3 antibodies in which a single CDR is purportedly responsible for antigen recognition [68-70]. The short form of IgW encodes only two C-domains, the least of any IgH chain discovered thus far, and appears to be exclusively found in the elasmobranchii lineage. Light chain-less BCRs of camelids and the IgNAR of nurse sharks use only three CDRs to recognize antigen in a more economical paratope compared to their heterodimeric brethren. Finally, bovine ultralong CDR3 antibodies are the ultimate minimization of concentrated structural diversity whereby a single "ultralong" CDR3 of the heavy chain provides the sole antigen interacting point. As stated the B cells of *B. taurus*, and many ruminant species, are primarily diversified through naïve SHM as opposed to sole development through recombination [71]. Furthermore, RAG mediated peripheral rearrangements have been found to occur during fetal development [72, 73].

Canonical variations for TCR include the TCRδ chain of *Ginglymostoma cirratum* being involved in "trans"-rearrangements, whereby IgHV gene segments are somatically recombined with TCRδ D-J-C, and the doubly-rearranging IgNARVDJ-TCRδVDJ-C [74, 75]. In *Xenopus*, IgHV-like "VHδ" genes are prominently represented in the TCRα/δ locus and are capable of rearranging with TCRδ components [76]. Additional instances of VHδ have been documented in the TCRαδ locus of the platypus, as well as passeriform birds [77]. Galliform birds house VHδ segments within a distinct locus complete with a separate TCRδ constant region [77, 78]. This could be taken as anecdotal evidence for a clearly distinct cell type using this receptor which necessitated a distinct locus. Non-placental

mammals also possess a TCR with some quaternary structure related to NARTCR, called TCRμ [79, 80]. With two variable domains TCRμ, both are more IgH-like in opposition to NARTCR bearing a TCRδ secondary V-segment, and the membrane proximal variable can be encoded by a germline-joined gene. The TCRμ constant region is similar to that of TCRδ, but resides in its own locus. Little is known of the functional purpose of these odd TCR variants and the advantages they might confer.

1.7 Choice of model organisms

The IGSF system of adaptive immunity arose in the jawed vertebrates, of which cartilaginous fish are the earliest clade, with the insertion of RAG transposable DNA elements into the primordial antigen receptor locus [81]. Cartilaginous fish have been found to possess the genetic backbone of all basic components the IGSF system of adaptive immunity employs, with all TCR (α, β, γ, and δ [82]) and BCR (IgH and IgL [83-85]) transcripts found in sharks. The thymus remains the site of T cell development in cartilaginous fish, possessing the same tissue architecture, consisting of distinct cortex and medullary regions, found in higher vertebrates [86, 87]. The nurse shark, *G. cirratum*, proved an ideal model due to construction of a complete BAC library, which allowed targeted genome sequencing for IGSF loci [88]. In addition to being representatives of the earliest clade utilizing the IGSF antigen receptors, they possess novel variations on adaptive immune cannon such as the previously stated Ig-TCRδ rearrangements [74], the doubly-rearranging NARTCRδ [75], and AID induced SHM of TCR transcripts [89].

Ruminants have also been shown to partake in irregular methods of IGSF antigen receptor diversification. Naïve SHM of developing B cells counters the limited pool of V, D, and J to allow a respectable primary repertoire [71, 90]. Peripheral RAG mediated rearrangements of immature B cells in the lymph nodes of fetal calves exemplify another bovine specific route of adaptive immune cell development [73]. Discovery of the ultralong CDR3 antibodies in cattle led to an increased focus in the field of B cell genetics of *B. taurus* [91]. These antibodies, in which CDR3 is greater than forty amino acid residues in length, form a unique tertiary structure in which a disulfide folded knob is held at the distal end of a protruding stalk [70]. The genetic backbone of the knob and stalk is provided by a single V_H-D_H rearrangement [92, 93]. An eight base pair duplication at the 3' end of this V_H yields a shift in frame resulting in the coding of a "CTTVHQ"motif, increasing length over the typical "CAR" motif found in many V domains [92]. The disulfide bond pattern of the D-segment is diversified through naïve SHM of the developing B cells, with 38 of 49 D_H codons mutable to a Cys via a single base change [70]. The high concentration of AID hotspots coupled with the antigen selection-exempt naïve SHM also account for a novel diversification mechanisms unique to ultralong CDR3 receptors in which deletions in the interior D_H coding regions result in a smaller knob domain [92]. Augmenting this novel structural antibody, the *B. taurus* BCR repertoire contains an unusually large number of nontemplate TdT additions [72]. This includes a semi-conserved range of adenine (A) additions at the V-D junction of ultralong CDR3 antibodies. The presence of ultralong CDR3 antibodies is thus far limited to the *Bos* species, however the integral V_H and D_H segments have been located in both the American

bison (*Bison bison*) [94] and European bison, or wisent, (*Bison bonasus*) [95]. Finally, the B cells of cattle, and other ruminants, have been posited to undergo AID catalyzed gene conversion in addition to naïve SHM [91].

Each of these models provides a unique opportunity to study enigmatic methods of repertoire development in direct contrast to the long-established dogma of adaptive immune systems. The atypical antigen receptors employed in each system also provide potential pathways to novel therapeutics and/or molecular tools. Ultralong CDR3 antibodies have already been developed to previously inaccessible broadly neutralizing epitopes on the HIV virus [96].

1.8 Dissertation aims

Though the IGSF system of adaptive immunity has been largely conserved across all clades, nuances in the system are constantly being elucidated. We have aimed to enhance the current knowledge base on the subject of two "immunological oddities" IgHV bearing γδ T cell of *G. cirratum* and ultralong CDR3 receptor bearing B cells of *B. taurus*. First, we outlined the genetic backbone of the *B. taurus* ultralong CDR3 antibodies. An additional method of diversification was found, whereby AID activity and proclivity to cause double-stranded DNA breaks, resulting in diminished size of an already small antigenic interface. Second, we characterized the trans-rearrangements in *G. cirratum.* We provide evidence that not only are these rearrangements diverse and common, but the resulting receptor is expressed on the surface in a manner consistent with all known IGSF adaptive antigen receptors. Furthermore, we located a functional

IgH V-segment within this ancient TCRα/δ locus. Finally, we took a theoretical approach aimed at providing explanations for methods of IGSF evolution both genetically and mechanistically. Included is a digression into where to draw the line concerning the definition of distinct B and T cell lineages, divergent developmental pathways, novel receptors, and when a distinct antigen receptor system defines a cell to a unique niche. These novel variations on traditional adaptive immune receptor repertoire and accompanying irregular adaptations of repertoire development are relevant not only for clinical exploitation, but to achieve a better understanding of the intricacies of the immune system not found in model lab organisms.

1.9 References

1. Riera Romo, M., D. Pérez-Martínez, and C. Castillo Ferrer, *Innate immunity in vertebrates: an overview.* Immunology, 2016. **148**(2): p. 125-139.
2. Cray, C., J. Zaias, and N.H. Altman, *Acute phase response in animals: a review.* Comparative medicine, 2009. **59**(6): p. 517-526.
3. Muller-Eberhard, H.J., *Complement.* Annual Review of Biochemistry, 1975. **44**(1): p. 697-724.
4. Takeuchi, O. and S. Akira, *Pattern Recognition Receptors and Inflammation.* Cell, 2010. **140**(6): p. 805-820.
5. Broggi, A. and F. Granucci, *Microbe- and danger-induced inflammation.* Molecular Immunology, 2015. **63**(2): p. 127-133.

6. Matzinger, P., *Tolerance, Danger, and the Extended Family.* Annual Review of Immunology, 1994. **12**(1): p. 991-1045.

7. Lemaitre, B., E. Nicolas, L. Michaut, J.-M. Reichhart, and J.A. Hoffmann, *The Dorsoventral Regulatory Gene Cassette spätzle/Toll/cactus Controls the Potent Antifungal Response in Drosophila Adults.* Cell, 1996. **86**(6): p. 973-983.

8. Hardison, S.E. and G.D. Brown, *C-type lectin receptors orchestrate antifungal immunity.* Nature immunology, 2012. **13**(9): p. 817-822.

9. Yoneyama, M., M. Kikuchi, T. Natsukawa, N. Shinobu, T. Imaizumi, M. Miyagishi, K. Taira, S. Akira, and T. Fujita, *The RNA helicase RIG-I has an essential function in double-stranded RNA-induced innate antiviral responses.* Nature Immunology, 2004. **5**: p. 730.

10. Kim, Y.K., J.S. Shin, and M.H. Nahm, *NOD-Like Receptors in Infection, Immunity, and Diseases.* Yonsei medical journal, 2016. **57**(1): p. 5-14.

11. Tonegawa, S., *Somatic generation of antibody diversity.* Nature, 1983. **302**(5909): p. 575-581.

12. Dempsey, P.W., S.A. Vaidya, and G. Cheng, *The Art of War: Innate and adaptive immune responses.* Cellular and Molecular Life Sciences CMLS, 2003. **60**(12): p. 2604-2621.

13. Bretscher, P.A., *A two-step, two-signal model for the primary activation of precursor helper T cells.* Proceedings of the National Academy of Sciences of the United States of America, 1999. **96**(1): p. 185-190.

14. Banchereau, J. and R.M. Steinman, *Dendritic cells and the control of immunity.* Nature, 1998. **392**(6673): p. 245-252.

15. Zhu, Q., M. Zhang, M. Shi, Y. Liu, Q. Zhao, W. Wang, G. Zhang, L. Yang, J. Zhi, L. Zhang, G. Hu, P. Chen, Y. Yang, W. Dai, T. Liu, Y. He, G. Feng, and G. Zhao, *Human B cells have an active phagocytic capability and undergo immune activation upon phagocytosis of Mycobacterium tuberculosis.* Immunobiology, 2016. **221**(4): p. 558-567.

16. Crotty, S., *A brief history of T cell help to B cells.* Nature Reviews Immunology, 2015. **15**: p. 185.

17. Brouns, S.J., M.M. Jore, M. Lundgren, E.R. Westra, R.J. Slijkhuis, A.P. Snijders, M.J. Dickman, K.S. Makarova, E.V. Koonin, and J. van der Oost, *Small CRISPR RNAs guide antiviral defense in prokaryotes.* Science, 2008. **321**(5891): p. 960-4.

18. Pancer, Z., N.R. Saha, J. Kasamatsu, T. Suzuki, C.T. Amemiya, M. Kasahara, and M.D. Cooper, *Variable lymphocyte receptors in hagfish.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(26): p. 9224-9229.

19. Flajnik, M.F. and L.L. Rumfelt, *The immune system of cartilaginous fish.* Curr Top Microbiol Immunol, 2000. **248**: p. 249-70.

20. Agrawal, A., Q.M. Eastman, and D.G. Schatz, *Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system.* Nature, 1998. **394**(6695): p. 744-751.

21. Schatz, D.G., M.A. Oettinger, and D. Baltimore, *The V(D)J recombination activating gene, RAG-1.* Cell, 1989. **59**(6): p. 1035-1048.

22. Bernstein, R.M., S.F. Schluter, H. Bernstein, and J.J. Marchalonis, *Primordial emergence of the recombination activating gene 1 (RAG1): sequence of the complete shark gene indicates homology to microbial integrases.* Proc.Natl.Acad.Sci.U.S.A, 1996. **93**(18): p. 9454-9459.

23. Hesse, J.E., M.R. Lieber, K. Mizuuchi, and M. Gellert, *V(D)J recombination: a functional definition of the joining signals.* Genes Dev, 1989. **3**(7): p. 1053-61.

24. Lewis, S.M., *The Mechanism of V(D)J Joining: Lessons from Molecular, Immunological, and Comparative Analyses*, in *Advances in Immunology*, F.J. Dixon, Editor. 1994, Academic Press. p. 27-150.

25. McBlane, J.F., D.C. van Gent, D.A. Ramsden, C. Romeo, C.A. Cuomo, M. Gellert, and M.A. Oettinger, *Cleavage at a V(D)J recombination signal requires only RAG1 and RAG2 proteins and occurs in two steps.* Cell, 1995. **83**(3): p. 387-395.

26. Mimori, T. and J.A. Hardin, *Mechanism of interaction between Ku protein and DNA.* J Biol Chem, 1986. **261**(22): p. 10375-9.

27. Nick McElhinny, S.A., C.M. Snowden, J. McCarville, and D.A. Ramsden, *Ku Recruits the XRCC4-Ligase IV Complex to DNA Ends.* Molecular and Cellular Biology, 2000. **20**(9): p. 2996.

28. Malu, S., V. Malshetty, D. Francis, and P. Cortes, *Role of non-homologous end joining in V(D)J recombination.* Immunologic Research, 2012. **54**(1): p. 233-246.

29. Ma, Y., U. Pannicke, K. Schwarz, and M.R. Lieber, *Hairpin Opening and Overhang Processing by an Artemis/DNA-Dependent Protein Kinase Complex in Nonhomologous End Joining and V(D)J Recombination.* Cell, 2002. **108**(6): p. 781-794.

30. Schlissel, M.S., *Structure of nonhairpin coding-end DNA breaks in cells undergoing V(D)J recombination.* Mol Cell Biol, 1998. **18**(4): p. 2029-37.

31. Lieber, M.R., *Site-specific recombination in the immune system.* The FASEB Journal, 1991. **5**(14): p. 2934-2944.

32. Nick McElhinny, S.A. and D.A. Ramsden, *Sibling rivalry: competition between Pol X family members in V(D)J recombination and general double strand break repair.* Immunological Reviews, 2004. **200**(1): p. 156-164.

33. Gauss, G.H. and M.R. Lieber, *Mechanistic constraints on diversity in human V(D)J recombination.* Molecular and Cellular Biology, 1996. **16**(1): p. 258.

34. Gellert, M., *Recent Advances in Understanding V(D)J Recombination*, in *Advances in Immunology*, F.J. Dixon, Editor. 1997, Academic Press. p. 39-64.

35. Mombaerts, P., J. Iacomini, R.S. Johnson, K. Herrup, S. Tonegawa, and V.E. Papaioannou, *RAG-1-deficient mice have no mature B and T lymphocytes.* Cell, 1992. **68**(5): p. 869-877.

36. Hesslein, D.G. and D.G. Schatz, *Factors and forces controlling V(D)J recombination.* Adv Immunol, 2001. **78**: p. 169-232.

37. Fugmann, S.D., *Form follows function — the three-dimensional structure of antigen receptor gene loci.* Current Opinion in Immunology, 2014. **27**: p. 33-37.

38. Mashoof, S. and M.F. Criscitiello, *Fish Immunoglobulins.* Biology, 2016. **5**(4): p. 45.

39. Bossen, C., R. Mansson, and C. Murre, *Chromatin Topology and the Regulation of Antigen Receptor Assembly.* Annual Review of Immunology, 2012. **30**(1): p. 337-356.

40. Bork, P., L. Holm, and C. Sander, *The Immunoglobulin Fold: Structural Classification, Sequence Patterns and Common Core.* Journal of Molecular Biology, 1994. **242**(4): p. 309-320.

41. Lefranc, M.-P., C. Pommié, M. Ruiz, V. Giudicelli, E. Foulquier, L. Truong, V. Thouvenin-Contet, and G. Lefranc, *IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains.* Developmental & Comparative Immunology, 2003. **27**(1): p. 55-77.

42. Chothia, C. and A.M. Lesk, *Canonical structures for the hypervariable regions of immunoglobulins.* Journal of Molecular Biology, 1987. **196**(4): p. 901-917.

43. Flajnik, M.F., *Comparative analyses of immunoglobulin genes: surprises and portents.* Nat Rev Immunol, 2002. **2**(9): p. 688-98.

44. Hsu, E., N. Pulham, L.L. Rumfelt, and M.F. Flajnik, *The plasticity of immunoglobulin gene systems in evolution.* Immunological reviews, 2006. **210**: p. 8-26.

45. Guy, C.S. and D.A.A. Vignali, *Organization of proximal signal initiation at the TCR:CD3 complex.* Immunological reviews, 2009. **232**(1): p. 7-21.

46. DeFranco, A.L., J.D. Richards, J.H. Blum, T.L. Stevens, D.A. Law, V.W. Chan, S.K. Datta, S.P. Foy, S.L. Hourihane, M.R. Gold, and et al., *Signal transduction by the B-cell antigen receptor.* Ann N Y Acad Sci, 1995. **766**: p. 195-201.

47. Call, M.E., J. Pyrdol, M. Wiedmann, and K.W. Wucherpfennig, *The organizing principle in the formation of the T cell receptor-CD3 complex.* Cell, 2002. **111**(7): p. 967-979.

48. Ackerman, G.A. and R.A. Knouff, *Lymphocytopoiesis in the bursa of Fabricius.* American Journal of Anatomy, 1959. **104**(2): p. 163-205.

49. Rumfelt, L.L., E.C. McKinney, E. Taylor, and M.F. Flajnik, *The development of primary and secondary lymphoid tissues in the nurse shark Ginglymostoma cirratum: B-cell zones precede dendritic cell immigration and T-cell zone formation during ontogeny of the spleen.* Scand J Immunol, 2002. **56**(2): p. 130-48.

50. Lukin, K., S. Fields, J. Hartley, and J. Hagman, *Early B cell factor: Regulator of B lineage specification and commitment.* Seminars in Immunology, 2008. **20**(4): p. 221-227.

51. Kearney, J.F., W.-J. Won, C. Benedict, C. Moratz, P. Zimmer, A. Oliver, F. Martin, and F. Shu, *B Cell Development in Mice.* International Reviews of Immunology, 1997. **15**(3-4): p. 207-241.

52. Sakaguchi, N. and F. Melchers, *λ5, a new light-chain-related locus selectively expressed in pre-B lymphocytes.* Nature, 1986. **324**(6097): p. 579-582.

53. Bankovich, A.J., S. Raunser, Z.S. Juo, T. Walz, M.M. Davis, and K.C. Garcia, *Structural Insight into Pre-B Cell Receptor Function.* Science, 2007. **316**(5822): p. 291.

54. Halverson, R., R.M. Torres, and R. Pelanda, *Receptor editing is the main mechanism of B cell tolerance toward membrane antigens.* Nature Immunology, 2004. **5**: p. 645.

55. Kaetzel, C.S., *Polymeric Ig receptor: Defender of the fort or Trojan Horse?* Current Biology, 2001. **11**(1): p. R35-R38.

56. Rothenberg, E.V., *Negotiation of the T Lineage Fate Decision by Transcription-Factor Interplay and Microenvironmental Signals.* Immunity, 2007. **26**(6): p. 690-702.

57. Bhandoola, A., H. von Boehmer, H.T. Petrie, and J.C. Zúñiga-Pflücker, *Commitment and Developmental Potential of Extrathymic and Intrathymic T Cell Precursors: Plenty to Choose from.* Immunity, 2007. **26**(6): p. 678-689.

58. von Boehmer, H. and H.J. Fehling, *Structure and function of the pre-T cell receptor.* Annu Rev Immunol, 1997. **15**: p. 433-52.

59. Jameson, S.C., K.A. Hogquist, and M.J. Bevan, *Positive selection of thymocytes.* Annu Rev Immunol, 1995. **13**: p. 93-126.

60. Chien, Y.-h. and Y. Konigshofer, *Antigen recognition by γδ T cells.* Immunological Reviews, 2007. **215**(1): p. 46-58.

61. Muramatsu, M., V.S. Sankaranand, S. Anant, M. Sugai, K. Kinoshita, N.O. Davidson, and T. Honjo, *Specific expression of activation-induced cytidine deaminase (AID), a novel member of the RNA-editing deaminase family in germinal center B cells.* J Biol Chem, 1999. **274**(26): p. 18470-6.

62. Methot, S.P. and J.M. Di Noia, *Chapter Two - Molecular Mechanisms of Somatic Hypermutation and Class Switch Recombination*, in *Advances in Immunology*, F.W. Alt, Editor. 2017, Academic Press. p. 37-87.

63. Martin, A. and M.D. Scharff, *AID and mismatch repair in antibody diversification.* Nature Reviews Immunology, 2002. **2**: p. 605.

64. Genschel, J., L.Y. Kadyrova, R.R. Iyer, B.K. Dahal, F.A. Kadyrov, and P. Modrich, *Interaction of proliferating cell nuclear antigen with PMS2 is required for MutLα activation and function in mismatch repair.* Proceedings of the National Academy of Sciences, 2017. **114**(19): p. 4930.

65. Masuda, K., R. Ouchida, M. Yokoi, F. Hanaoka, T. Azuma, and J.-Y. Wang, *DNA polymerase η is a limiting factor for A:T mutations in Ig genes and contributes to antibody affinity maturation.* European Journal of Immunology, 2008. **38**(10): p. 2796-2805.

66. Begum, N.A., K. Kinoshita, N. Kakazu, M. Muramatsu, H. Nagaoka, R. Shinkura, D. Biniszkiewicz, L.A. Boyer, R. Jaenisch, and T. Honjo, *Uracil DNA Glycosylase Activity Is Dispensable for Immunoglobulin Class Switch.* Science, 2004. **305**(5687): p. 1160.

67. Dong, J., R.A. Panchakshari, T. Zhang, Y. Zhang, J. Hu, S.A. Volpi, R.M. Meyers, Y.J. Ho, Z. Du, D.F. Robbiani, F. Meng, M. Gostissa, M.C. Nussenzweig, J.P. Manis, and F.W. Alt, *Orientation-specific joining of AID-initiated DNA breaks promotes antibody class switching.* Nature, 2015. **525**(7567): p. 134-139.

68. Hamerscasterman, C., T. Atarhouch, S. Muyldermans, G. Robinson, C. Hamers, E.B. Songa, N. Bendahman, and R. Hamers, *NATURALLY-OCCURRING ANTIBODIES DEVOID OF LIGHT-CHAINS.* Nature, 1993. **363**(6428): p. 446-448.

69. Dooley, H. and M.F. Flajnik, *Shark immunity bites back: affinity maturation and memory response in the nurse shark, Ginglymostoma cirratum.* Eur J Immunol, 2005. **35**(3): p. 936-45.

70. Wang, F., D.C. Ekiert, I. Ahmad, W. Yu, Y. Zhang, O. Bazirgan, A. Torkamani, T. Raudsepp, W. Mwangi, M.F. Criscitiello, I.A. Wilson, P.G. Schultz, and V.V. Smider, *Reshaping antibody diversity.* Cell, 2013. **153**(6): p. 1379-93.

71. Yasuda, M., C.N. Jenne, L.J. Kennedy, and J.D. Reynolds, *The sheep and cattle Peyer's patch as a site of B-cell development.* Vet Res, 2006. **37**(3): p. 401-15.

72. Liljavirta, J., M. Niku, T. Pessa-Morikawa, A. Ekman, and A. Iivanainen, *Expansion of the preimmune antibody repertoire by junctional diversity in Bos taurus.* PLoS One, 2014. **9**(6): p. e99808.

73. Ekman, A., T. Pessa-Morikawa, J. Liljavirta, M. Niku, and A. Iivanainen, *B-cell development in bovine fetuses proceeds via a pre-B like cell in bone marrow and lymph nodes.* Dev Comp Immunol, 2010. **34**(8): p. 896-903.

74. Criscitiello, M.F., Y. Ohta, M. Saltis, E.C. McKinney, and M.F. Flajnik, *Evolutionarily conserved TCR binding sites, identification of T cells in primary lymphoid tissues, and surprising trans-rearrangements in nurse shark.* J Immunol, 2010. **184**(12): p. 6950-60.

75. Criscitiello, M.F., M. Saltis, and M.F. Flajnik, *An evolutionarily mobile antigen receptor variable region gene: doubly rearranging NAR-TcR genes in sharks.* Proc Natl Acad Sci U S A, 2006. **103**(13): p. 5036-41.

76. Parra, Z.E., Y. Ohta, M.F. Criscitiello, M.F. Flajnik, and R.D. Miller, *The dynamic TCRdelta: TCRdelta chains in the amphibian Xenopus tropicalis utilize antibody-like V genes.* Eur J Immunol, 2010.

77. Parra, Z.E., M. Lillie, and R.D. Miller, *A Model for the Evolution of the Mammalian T-cell Receptor α/δ and μ Loci Based on Evidence from the Duckbill Platypus.* Molecular Biology and Evolution, 2012. **29**(10): p. 3205-3214.

78. Parra, Z.E., K. Mitchell, R.A. Dalloul, and R.D. Miller, *A second TCRdelta locus in Galliformes uses antibody-like V domains: insight into the evolution of TCRdelta and TCRmu genes in tetrapods.* J Immunol, 2012. **188**(8): p. 3912-9.

79. Parra, Z.E., M.L. Baker, R.S. Schwarz, J.E. Deakin, K. Lindblad-Toh, and R.D. Miller, *A unique T cell receptor discovered in marsupials.* Proc Natl Acad Sci U S A, 2007. **104**(23): p. 9776-81.

80. Wang, X., Z.E. Parra, and R.D. Miller, *Platypus TCRmu provides insight into the origins and evolution of a uniquely mammalian TCR locus.* J Immunol, 2011. **187**(10): p. 5246-54.

81. Bernstein, R.M., S.F. Schluter, H. Bernstein, and J.J. Marchalonis, *Primordial emergence of the recombination activating gene 1 (RAG1): sequence of the complete shark gene indicates homology to microbial integrases.* Proc Natl Acad Sci U S A, 1996. **93**(18): p. 9454-9.

82. Rast, J.P., M.K. Anderson, S.J. Strong, C. Luer, R.T. Litman, and G.W. Litman, *alpha, beta, gamma, and delta T cell antigen receptor genes arose early in vertebrate phylogeny.* Immunity, 1997. **6**(1): p. 1-11.

83. Hinds, K.R. and G.W. Litman, *Major reorganization of immunoglobulin VH segmental elements during vertebrate evolution.* Nature, 1986. **320**(6062): p. 546-9.

84. Kokubu, F., R. Litman, M.J. Shamblott, K. Hinds, and G.W. Litman, *Diverse organization of immunoglobulin VH gene loci in a primitive vertebrate.* Embo j, 1988. **7**(11): p. 3413-22.

85. Criscitiello, M.F. and M.F. Flajnik, *Four primordial immunoglobulin light chain isotypes, including lambda and kappa, identified in the most primitive living jawed vertebrates.* Eur J Immunol, 2007. **37**(10): p. 2683-94.

86. Miracle, A.L., M.K. Anderson, R.T. Litman, C.J. Walsh, C.A. Luer, E.V. Rothenberg, and G.W. Litman, *Complex expression patterns of lymphocyte-specific genes during the development of cartilaginous fish implicate unique lymphoid tissues in generating an immune repertoire.* Int Immunol, 2001. **13**(4): p. 567-80.

87. Wyffels, J.T., C.J. Walsh, C.A. Luer, and A.B. Bodine, *In vivo exposure of clearnose skates, Raja eglanteria, to ionizing X-radiation: acute effects on the thymus.* Dev Comp Immunol, 2005. **29**(4): p. 315-31.

88. Luo, M., H. Kim, D. Kudrna, N.B. Sisneros, S.J. Lee, C. Mueller, K. Collura, A. Zuccolo, E.B. Buckingham, S.M. Grim, K. Yanagiya, H. Inoko, T. Shiina, M.F. Flajnik, R.A. Wing, and Y. Ohta, *Construction of a nurse shark (Ginglymostoma cirratum) bacterial artificial chromosome (BAC) library and a preliminary genome survey.* BMC Genomics, 2006. **7**: p. 106.

89. Ott, J.A., C.D. Castro, T.C. Deiss, Y. Ohta, M.F. Flajnik, and M.F. Criscitiello, *Somatic hypermutation of T cell receptor alpha chain contributes to selection in nurse shark thymus.* Elife, 2018. **7**.

90. Butler, J.E., *Immunoglobulin diversity, B-cell and antibody repertoire development in large farm animals.* Rev Sci Tech, 1998. **17**(1): p. 43-70.

91. Berens, S.J., D.E. Wylie, and O.J. Lopez, *Use of a single VH family and long CDR3s in the variable region of cattle Ig heavy chains.* Int Immunol, 1997. **9**(1): p. 189-99.

92. Deiss, T.C., M. Vadnais, F. Wang, P.L. Chen, A. Torkamani, W. Mwangi, M.P. Lefranc, M.F. Criscitiello, and V.V. Smider, *Immunogenetic factors driving formation of ultralong VH CDR3 in Bos taurus antibodies.* Cell Mol Immunol, 2017.

93. Ma, L., T. Qin, D. Chu, X. Cheng, J. Wang, X. Wang, P. Wang, H. Han, L. Ren, R. Aitken, L. Hammarstrom, N. Li, and Y. Zhao, *Internal Duplications of DH, JH, and C Region Genes Create an Unusual IgH Gene Locus in Cattle.* J Immunol, 2016. **196**(10): p. 4358-66.

94. Dobson., L.K., *Sequencing the Genome of the North American Bison.* , in *Veterinary Pathobiology*. 2015, Texas A&M University: College Station, TX.

95. Wang, K., L. Wang, J.A. Lenstra, J. Jian, Y. Yang, Q. Hu, D. Lai, Q. Qiu, T. Ma, Z. Du, R. Abbott, and J. Liu, *The genome sequence of the wisent (Bison bonasus).* Gigascience, 2017. **6**(4): p. 1-5.

96. Sok, D., K.M. Le, M. Vadnais, K.L. Saye-Francisco, J.G. Jardine, J.L. Torres, Z.T. Berndsen, L. Kong, R. Stanfield, J. Ruiz, A. Ramos, C.H. Liang, P.L. Chen, M.F. Criscitiello, W. Mwangi, I.A. Wilson, A.B. Ward, V.V. Smider, and D.R. Burton, *Rapid elicitation of broadly neutralizing antibodies to HIV by immunization in cows.* Nature, 2017. **548**(7665): p. 108-111.

CHAPTER II

IMMUNOGENETIC FACTORS DRIVING FORMATION OF ULTRALONG VH CDR3 *BOS TAURUS* ANTIBODIES*

2.1 Introduction

Antibodies are the primary molecules responsible for eliminating invading foreign pathogens in vertebrates. Cows are unusual in producing antibodies with exceptionally long VH CDR3, with such antibodies having a unique ability amongst vertebrates to bind and neutralize the HIV spike protein env. In fact, compared to other species, cows are able to mount a particularly rapid and broadly neutralizing serum response against HIV [1]. Therefore the genetic factors driving formation of ultralong CDR3 is important in understanding the basis for optimal host-pathogen interactions which has broad implications in vaccine and therapeutic design.

Lymphocyte antigen receptors represent a unique paradigm for the creation of genetic and structural diversity. The antigen receptors of jawed vertebrates are comprised of a diverse repertoire of immunoglobulins (IG) or antibodies and T cell receptors (TR), which through combinatorial and junctional V-(D)-J diversity, and for IG, somatic hypermutation, enables the generation of specific antigen receptors that bind to an enormous array of antigenic epitopes [2-4].

However, despite the extensive variability in the antibody system, genetic and structural constraints on diversity exist, which could impact the diversity of paratopes that may be present in the repertoire. For example, the heavy chain variable domain complementarity determining region 3 (VH CDR3), which often provides significant contact with the antigen, is on average 13 amino acids in length and forms a loop constrained by the two β-strands "F" and "G" of the immunoglobulin scaffold [5]. Furthermore, a positively charged amino acid (usually arginine R or lysine K) of the V-REGION (at IMGT position 106, near the N-terminal portion of the VH CDR3 loop) nearly always forms a salt bridge with a negatively charged amino acid (aspartic acid D or glutamic acid E) of the J-REGION (at IMGT position 116, near the C-terminal portion of the VH CDR3 loop) [6]. Additionally, while the other two CDR loops of the VH and the three CDR loops of the light chain variable domain (VL) can be diverse in their sequences, they too are structurally constrained by length and amino acid sequence [7, 8]. Furthermore, restrictions on heavy chain/light chain pairing could potentially limit the paratope of an antibody. In this regard, most protein binding antibodies form a combining site that has a relatively flat or undulating interacting surface, as opposed to alternative paratopes that may have a different shape (*e.g.* concave or protruding). The ability to "break through" these structural constraints to generate alternative antibody paratopes may enable binding to different classes of epitopes than the typical antibody repertoire. Indeed, sharks, camels and cows have evolved unusual structural features compared to typical vertebrate antibodies [9-11]. Cartilaginous fish and camelids have subsets of antibodies devoid of light chains and thus contain only three CDR (instead of

six), with a much smaller paratope (reviewed in de los Rios et al. 2015 [12]). This antibody structure has been used to bind recessed epitopes such as those in G-protein coupled receptors and enzymatic active sites [13]. Cows, however, contain a subset of antibodies with exceptionally long VH CDR3, which can reach lengths of over seventy amino acids [14-19]. The VH CDR3 of these cow antibodies form a disulfide bonded "knob" that sits atop a β-ribbon "stalk" enabling the CDR3 structure to protrude far from the antibody surface [18, 19]. Understanding the genetic basis underlying the ability of cow ultralong VH CDR3 antibodies to innovate beyond the structural constraints of a typical antibody could lead to insights into vertebrate antibody evolution, provide further understanding of host-pathogen interactions (like broadly neutralizing HIV antibodies), and open new design space in immunotherapeutic engineering.

The genetic system encoding antigen receptors has two key and potentially opposing purposes: it must allow generation of a diverse repertoire, but in contrast must also ensure the structural integrity of each molecule. Thus, a system that allows complete randomness of amino acid content at each position may allow for maximal diversity, however, the vast majority of these molecules would be non-functional as they would not fold into a soluble protein structure [20]. Therefore, a protein scaffold is necessary, and in vertebrate antibodies is fulfilled by the immunoglobulin domain [2]. Diversity is achieved within loops constrained by strands in the β-sandwich fold. The VH CDR3 of cow antibodies is unusual owing to the dramatic extension of two β-strands of the V domain scaffold. Given the constraints in sequence content and size of

typical antibody genes, these cow antibodies may represent an unusual paradigm for how genetic structural diversity can be generated.

The variable domains of antibody heavy chains are encoded by one each of multiple variable (V), diversity (D), and joining (J) genes which undergo recombination at the DNA level to produce a "naïve" antibody repertoire [3, 21, 22]. The process of V-(D)-J recombination, along with nucleotide deletions and insertions at the D-J, V-(D-J) or V-J joints, as well as the random pairing of heavy and light chains can produce an enormous repertoire of antibody molecules [2, 3, 21]. Humans, for example, have 36-49 functional IGHV germline genes belonging to seven subgroups which can recombine with any of the functional 23 IGHD and 6 IGHJ genes [3, 23, 24]. Thus, the germline variability comprising multiple V, D, and J genes is a defining characteristic that enables combinatorial repertoire formation.

The major surface for contacting antigen is predominantly comprised of the CDR3 of the VH and VL, which is encoded by the V-D-J and V-J junctions, respectively. In contrast, the CDR1 and CDR2 of the VH and VL are encoded by the V genes only. The diversity of the VH and VL is increased by somatic hypermutations (SHM) which result from activation induced cytidine deaminase (AICDA, AID) activity. Following antigen binding, amino acid changes are selected and enable higher affinity binding [25-27].

Unlike humans and mice, cows have few IGHV genes, with only twelve V genes predicted to be functional which all belong to the same IGHV1 subgroup (homologous

to the *Ovis aries* IGHV1) and share greater than 90% identity to one another [28]. Thus, compared to humans, cows have a significantly limited V gene repertoire. However, ruminant immune systems appear to utilize an innovative mechanism to expand this limited repertoire, whereby naïve B cells undergo AID mediated somatic hypermutation in the periphery [29-36].

Deep sequence analysis revealed the bovine ultralong VH CDR3 to be highly diverse and contain several cysteines that were most frequently present in even numbers, suggesting that they formed disulfide bonds [19]. Indeed, crystal structures of five bovine antibodies that had unrelated ultralong VH CDR3 sequences showed that they all had an unusual protruding β-ribbon "stalk" and a "knob" that had different disulfide bonding patterns. Whereas the sequences of the VH CDR3 were highly divergent, and the structures differed in their loop patterns and surface charge, these antibodies shared the stalk and knob structural features. The binding of antibody H12, the only published ultralong CDR3 antibody with a clearly defined antigen, was dependent on specific amino acids in the knob, and removal of the knob resulted in complete loss of antigen binding [19]. Taken together, these results suggest that bovine antibodies with ultralong CDR3 may contact antigen through the disulfide-bonded knob, and that the remaining CDR are only used for structural support.

The formation of bovine ultralong VH CDR3 appears to result from utilization of a germline VHBUL gene (*Bos taurus* IGHV1-7 of IMGT nomenclature, which will be used), which was identified as encoding several of the published VH sequences with

ultralong CDR3[15, 19], and which encodes a portion of the ascending strand of the β-ribbon stalk. The rearrangement of the functional IGHV1-7 gene to the IGHD8-2 gene (previously referred to as DH2 [37, 38]) with an unusually long region, produces an ultralong CDR3 of at least fifty amino acids, including the descending strand of the stalk [28]. The long IGHD8-2 gene features a high concentration of AID hotspots, nucleotide motifs generally associated with a higher rate of AID mutation activity [39]. Of considerable interest, over 80% of the codons of IGHD8-2 may be mutated to a cysteine with a single base change, with many of these codons lying in hotspots. This results in a higher likelihood of any given amino acid being mutated to a cysteine. The base of the stalk region, which is encoded at the V-D and D-J junctions, is divergent from the typical features of an antibody in this region; it cannot establish the classical salt bridge which usually stabilizes the VH CDR3 loop [6]. Thus, an emergent feature of the ultralong VH CDR3 could be the ability to encode key amino acids initiating the ascending β-strand and breaking the constraint imposed by the conserved salt bridge at the base of VH CDR3.

Here we investigated the genetic basis by which cow ultralong VH CDR3 defy the structural constraints of a typical antibody V domain structure. We find that the IGHV1-7 region is utilized in nearly all ultralong VH CDR3 antibodies, and the key evolutionary driver forming IGHV1-7 appears to be a short nucleotide duplication that alters the protein coding region and enables an extended F β-strand at the amino terminus of VH CDR3 to be encoded in the germline. Genetic diversity is less extensive in this IGHV1-7 variable region, suggesting that its use in ultralong VH CDR3

antibodies primarily relates to its ability to stabilize their unusual structure. Furthermore, we describe a deletion activity that is suggestive of a novel AID diversification mechanism that further diversifies ultralong VH CDR3 by altering their length and cysteine positions.

2.2 Results

2.1.1 IGHV1-7 (VH_{BUL}) contains an internal duplication extending the germline CDR3

To understand the genetic underpinnings of ultralong VH CDR3 formation we analyzed the immunoglobulin heavy chain locus of *B. taurus*. The most recent assembly of the cow genome confirmed that in the IGH locus (accession KT723008) all functional IGHV genes belong to a single subgroup, IGHV1 [28]. The germline IGHV1 genes are closely related sequences, owing to recent 'cassette-like' duplications in the locus, with >90% nucleotide identity between members, and slight differences in CDR1 and CDR2 at the amino acid level (Fig 2.1a). However, there is a striking difference at the C-terminal end of IGHV1-7, which comprises a divergent motif immediately following the second Cys (2nd-CYS 104), and which defines the start of CDR3 (Fig 2.1a,b). Bovine VH containing ultralong CDR3 were previously reported to use a single IGHV region (VHBUL or IGHV1-7) that is longer than typical IGHV regions from bovines as well as other mammals [19]. Inspection of the DNA sequence at the 3' end of IGHV1-7 revealed an internal duplication of eight nucleotides (either TACTACTG or ACTACTGT) beginning at, or just after, the 3rd position encoding the canonical 2nd-CYS 104 (Fig

2.1b). The duplication, in addition to extending CDR3, results in a frameshift at the 3' end of IGHV1-7 altering the traditional "CA(R/K)" motif found at the C-terminus of other IGHV1 members (and conserved throughout most vertebrate IGHV regions). Importantly, the "CA(R/K)" often forms a salt bridge within CDR3 using the arginine R or lysine K 106 derived from the IGHV and the aspartic acid or glutamic acid 116 encoded by the IGHJ region [28, 40, 41] (Fig 2.1c). Instead of this traditional motif,

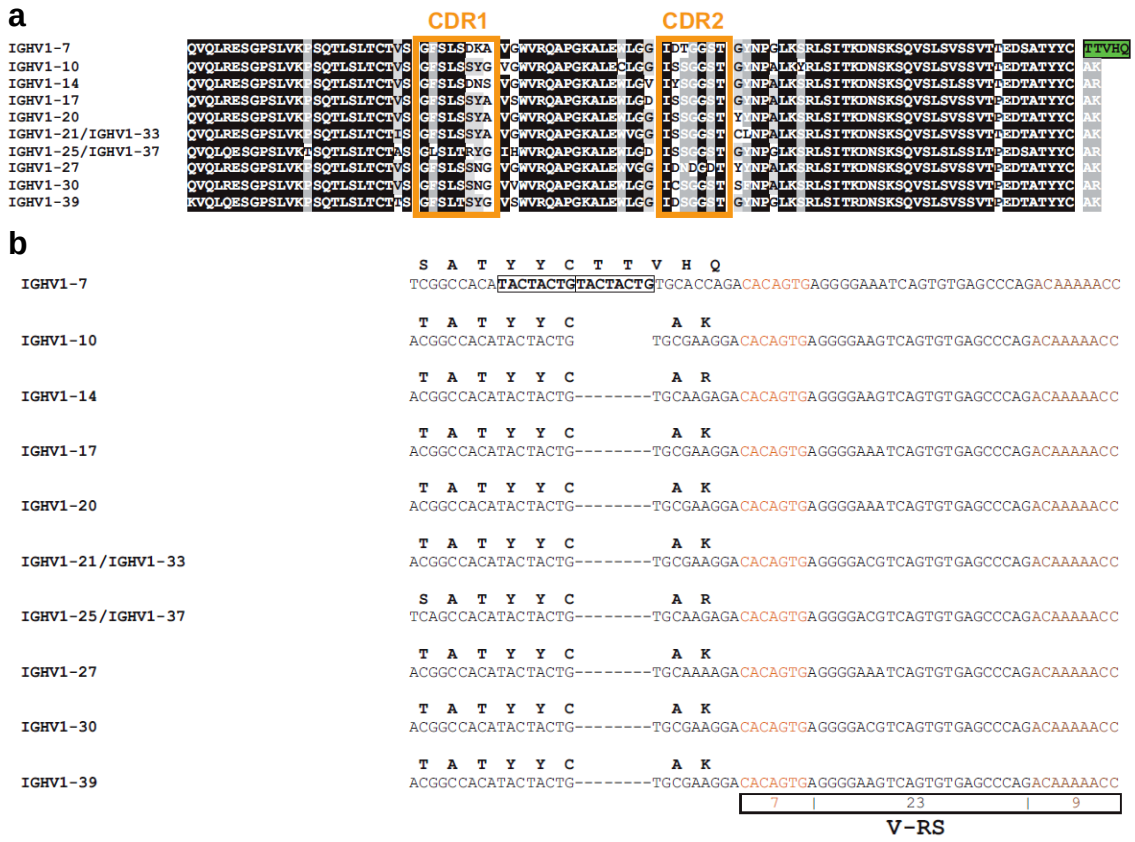


Figure 2.1 Genetic basis for the 'stalk' ascending β-strand in ultralong VH CDR3. (a) Amino-acid alignment of functional IGHV1 subgroup members. Amino acids shaded with Blosum62 similarity (black = 100%, dark grey = 80–100%, light grey = 60–80% and white =o60%). Orange boxes denote CDR1 and CDR2 and the TTVHQ extension of IGHV1-7 is indicated in green. **(b)** Nucleotide alignment of the 3′ end of the IGHV1 subgroup members. The internal duplication in IGHV1-7 is boxed. The heptamer and nonamer of the V recombination signal (V-RS) sequence are indicated by orange and purple, respectively. **(c)** Ribbon structures of bovine Fab BLV1H12 (PDB: 4K3D, originally described in Wang, et al.19) with a VH encoding ultralong CDR3 (left) and comparison of the CDR3 interactions in ultralong VH CDR3 (middle) and traditional VH CDR3 (right). The boxed region on the left is enlarged in the middle diagram. The CDR1 are shown in orange. In traditional VH CDR3, the canonical salt bridge between the arginine/lysine of the CDR3 (green shaded, IMGT position 106) and the aspartate of the J-REGION (red shaded, IMGT position 116) is shown by a dotted red line.

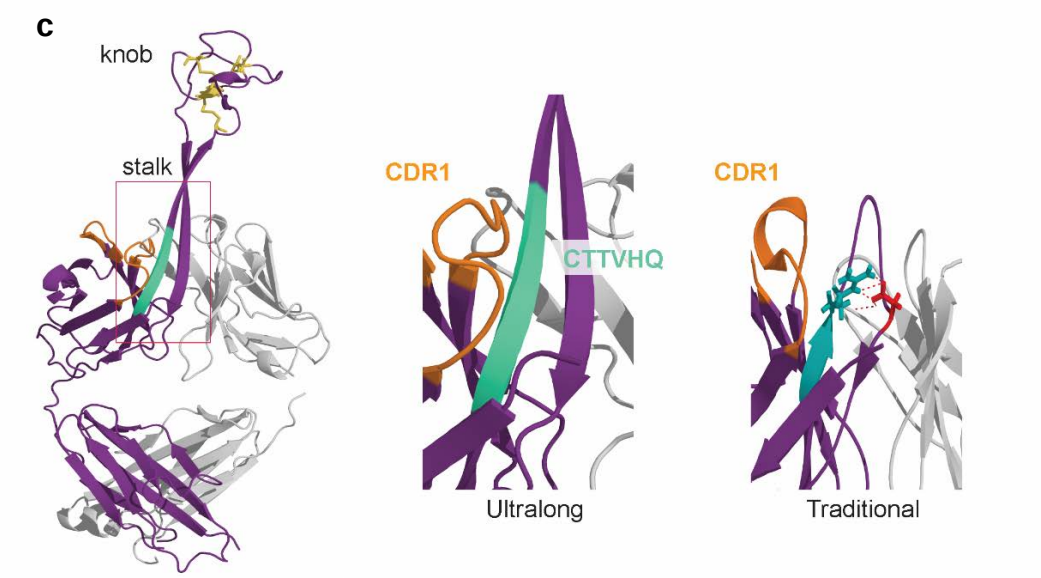


Figure 2.1. Continued.

IGHV1-7 encodes a “CTTVHQ” motif, which has been identified as a key feature of ultralong VH CDR3 [19]. The “CTTVHQ” motif is an integral component of the ascending portion of the β-ribbon stalk which supports a uniquely folded knob at the distal end of the CDR3 providing a novel antigen interface (Fig 2.1c) [19].

2.1.2 IGHV1-7 is preferentially used in ultralong VH CDR3

As IGHV1-7 encodes unusual features that may enable ultralong CDR3 formation, we surmised that it may be preferentially found in ultralong VH CDR3 sequences, as previously suggested [19]. To this end, we cloned the antibody VH repertoire of two cows. To negate IGHV bias, and allow a full repertoire scope, the forward primer targeted a ligated 5’ Gene Racer Oligo and was paired with a reverse primer hybridizing to IGHM and IGHG constant genes. Deep sequencing of the two animals yielded a combined 12,934 unique sequences (of a total of 13,030 sequences)

and, as expected, all sequences originated from a member of the IGHV1 subgroup (the only subgroup of the three identified in cattle, with functional genes).

We analyzed the deep sequence VH repertoire data to determine the IGHV gene use as a function of CDR3 length (Fig 2.2). Of the 13,030 total sequences analyzed, the CDR3 average length was 25.56, however a bimodal distribution was recognized which corresponded to shorter and ultralong groups. Amongst the 12,010 shorter CDR3 sequences (91.8%), the mean length was 22.8 with a range from 5 to 38 amino acids. For the 895 sequences with ultralong CDR3 (6.85%; defined as equal to or greater than 40 AA by IMGT numbering standards [8]), which falls within the ~4-13% range previously reported [15, 28, 42]) the mean length was 61.8 and the longest CDR3 was 72 AA. When the V gene usage of the ultralong transcripts was analyzed, a remarkable 97.2% of ultralong CDR3 encoding transcripts used IGVH1-7 (Fig 2.2, top). Thus, ultralong VH CDR3 antibodies appear to have a severe bias towards use of this germline gene. The remaining 2.8% of ultralong CDR3 transcripts appeared to result from an IGHV1 gene other than IGHV1-7. This is much lower than the expected frequency (8.3%) if each IGHV region contributed equally to ultralong CDR3. This preferential use is reflected in the analysis of CDR3 length of all VH domains in which CDR3 length of IGHV1-7 containing transcripts is significantly longer than any other region, encoding an average of 55±13 AA. Although nearly all ultralong CDR3 transcripts utilized IGHV1-7, this gene was found in shorter CDR3 as well (Fig 2.2, bottom); 9.3% of IGHV1-7 transcripts encoded a shorter CDR3. Thus, IGHV1-7 is the only V gene used in ultralong sequences, but it can also be used in shorter CDR3. Interestingly, shorter

CDR3 sequences also appear to have a strong bias for IGHV gene usage; IGHV1-10 was found in 72.7% of sequences with CDR3 <40 amino acids. Of note, two of the twelve potentially functional IGHV1 genes, IGHV1-25 and IGHV1-37, which have identical amino acid sequences (Fig 2.1a) were not identified in any transcripts (Fig 2.2), and may not be utilized in the repertoire.

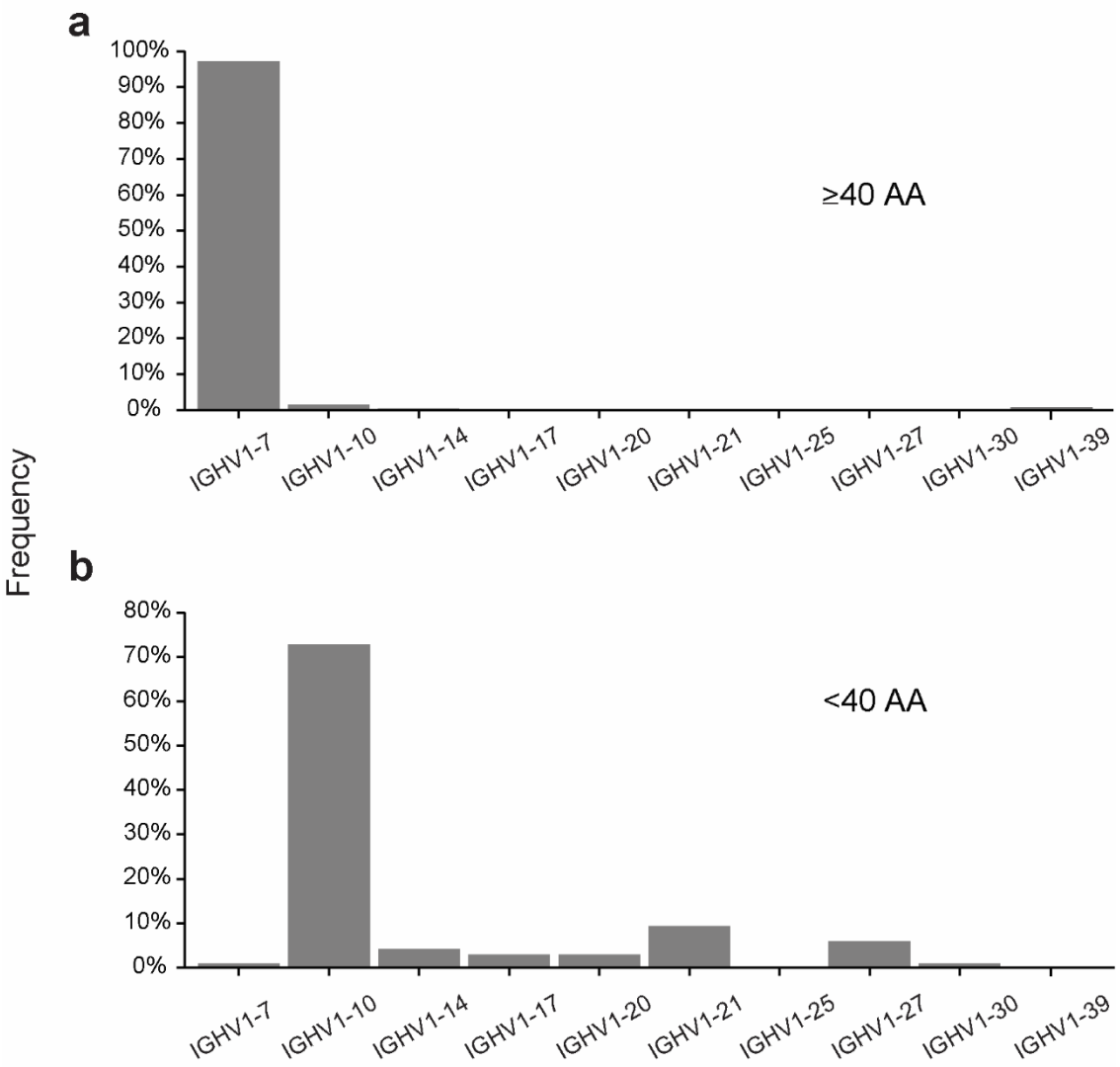


Figure 2.2 Ultralong VH CDR3 transcripts preferentially use one IGHV1 subgroup member IGHV1-7. (TOP) Percentage of IGHV1 genes expressed in transcripts with VH CDR3 equal to or greater than 40 AA. (BOTTOM) Percentage of IGHV1 genes expressed in transcripts with VH CDR3 less than 40 AA. IGHV1-21 and IGHV1- 33, and IGHV1-25 and IGHV1-37, are identical, therefore only IGHV1-21 and IGHV1-25 are labeled.

A conserved feature of ultralong CDR3 is the CTTVHQ motif encoded by the 3' end of IGHV1-7. All except one of the ultralong CDR3 sequences had an identifiable CTTVHQ-related motif. There were 15 sequences which had CTTVHQ-like motifs that were identified as an IGHV1 gene other than IGHV1-7. It is likely that these sequences actually arose from IGHV1-7 and were misidentified because somatic mutations shifted the sequence enough such that the blast algorithm incorrectly identified them as a different IGHV1. In this regard, it is unlikely that an 8 bp insertion would have occurred somatically in these non-IGHV1-7 regions. Additionally, 82 sequences were identified as IGHV1-7 although they did not have an identifiable CTTVHQ-like motif. This could be due to somatic hypermutation or exonuclease activity removing the CTTVHQ-like motif during V to D-J recombination.

2.1.3 Deletions diversify ultralong VH CDR3

During the development of an immune response, exposure of IG expressed at the B cell surface to antigen typically results in selection of AID driven SHM and B cell clonal proliferation, the driving force behind affinity maturation, as well as class switch recombination (CSR) of the CH domains resulting in distinct effector functions of the antibody. Within the ultralong CDR3 subset, deep sequencing unveiled novel deletion events within the IGHD8-2 region in which interior nucleotides are regularly removed, however leave the regions encoding the structurally relevant CPDG turn motif at the initiation of the "knob" and the alternating aromatic amino acids (YxYxY) of the descending stalk untouched [19, 28](Fig 2.3a). Surprisingly, a total of 426 out of the 894

ultralong sequences (47.6%) had in-frame nucleotide deletions. The deletions (colored lines in Fig 2.3a) range from 1-18 interior codons (Fig 2.3a-d) and retain the aforementioned motifs with high sequence homology to the 5' and 3' end of the germline IGHD8-2 (Fig 2.3a,b). Remarkably, deletions surpassing five consecutive codons were observed in 20% of ultralong CDR3 encoding transcripts. Additionally, codon deletions of greater than ten codons were observed in 5.7% of ultralong CDR3 transcripts. Deletion events were largely constrained to the IGHD8-2 region encoding CDR3, as only one deletion was discovered outside of CDR3 and consisted of a 6 bp deletion within CDR2. The positions of the deletions were variable, however the internal portion of IGHD8-2 had a higher frequency of deletions, consistent with the 5' and 3' ends of IGHD8-2 being used to encode conserved motifs such as CPDG turn at the 5' end and alternating aromatic amino acids YxYxY at the 3' end (Fig 2.3a,b). The unheralded length of IGHD8-2 and the vast number of AID hotspot motifs (RGYW/WRCY) it contains give it similarities to the switch regions targeted by AID for CSR [39, 43]. During CSR requisite cytokine signals open the constant region of the IGH locus allowing AID access to switch regions to catalyze double strand DNA breaks [43]. In this regard, 96.9% of the deletions overlapped an AID hotspot (Fig 2.3b). The CDR3 sequences with in-frame deletions had slightly altered cysteine content, as may be expected with shorter sequences. Of the sequences with full length IGHD8-2, 18 had 10 cysteines, whereas of the sequences with deletions only two had 10 cysteines. On the other end of the spectrum, 25 sequences with deletions had two cysteines, whereas of the sequences with full length IGHD8-2, only two had two cysteines (Fig 2.3c). Thus the

overall cysteine content was lowered in sequences with deletions. Because of the repetitive nature of the codons within the germline IGHD8-2 and the high mutation load, it is difficult to definitively ascertain whether cysteine positions are altered by the deletions. However because the mutations can occur internally within IGHD8-2 at positions which encode cysteine, or between cysteines (Fig 2.3a- c), it is highly likely that cysteine position alterations occur with some deletion events. To summarize, nucleotide deletions occur with high frequency, with proclivity to internal regions of IGHD8-2 (thus sparing key regions that encode structurally conserved regions), and alter the cysteine content and likely position, thus impacting disulfide bond patterns in the knob.

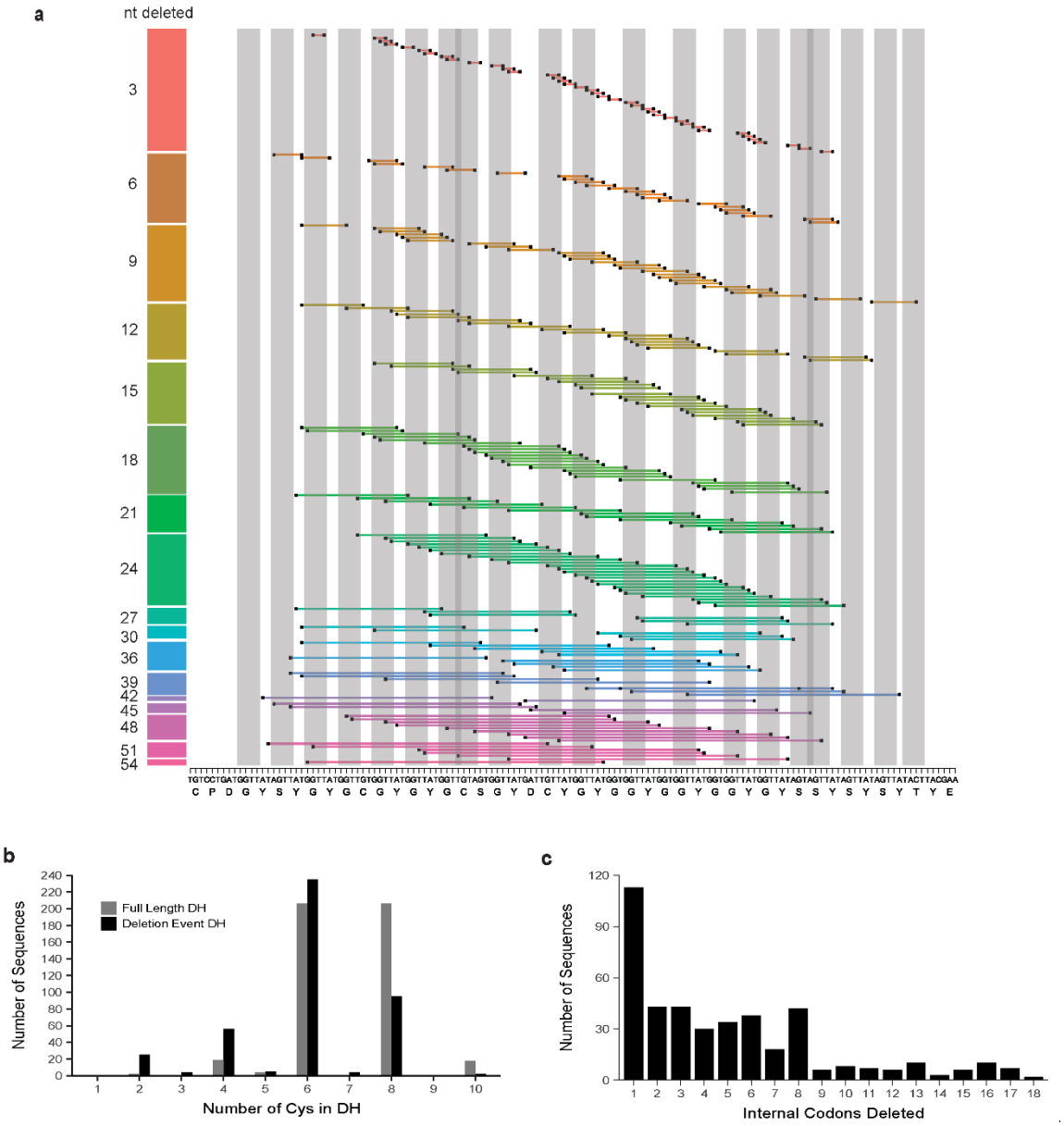


Figure 2.3 Ultralong VH CDR3 are characterized by frequent deletions in IGHD8-2 region. (a) Graph showing the gapped/deleted positions from alignments with the full length IGHD8-2 gene. The colored lines are indicative of gapped/deleted nucleotides, with the black dots indicating the deletion boundaries. The light grey boxes indicate AID hotspot motifs; areas with overlapping hotspots shown in a darker grey background. Deletion positions are ordered from fewest at the top to the greatest number of nucleotides deleted at the bottom. The sequence and translation of IGHD8-2 is at the bottom. **(b)** Deletions alter cysteine content. The number of cysteines in full length IGHD8- 2 sequences (grey bars) were compared to sequences with internal deletions in IGHD8-2 (black bars). **(c)** Relative number of sequences as a function of codons deleted. The number of sequences (y axis) which contained varying numbers of codon deletions (x axis) is plotted.

2.1.4 IGHV1-7 has low somatic variability

While the knob portion of ultralong CDR3 has been documented as a requirement for interaction with a specific antigen, it has yet to be determined whether CDR1 and CDR2 also interact with antigen [19]. Previous structural analysis suggested that these CDR may participate in stabilizing interactions with the ultralong "stalk" region [19]. For these reasons we speculated that the ultralong CDR3 repertoire may not have the variability of a typical IGH repertoire. To quantify where the variability, and thus potential antigen interaction, was located we performed a Shannon entropy analysis of the IGHV region for all members of the IGHV1 subgroup on deep sequenced heavy chain transcripts. Entropy analysis revealed that IGHV1-7 was the only IGHV gene in which significant "variable" amino acids were not found in either CDR1 or CDR2. This is in contrast to typical antibodies in other species, as well as shorter VH CDR3 antibodies in cows, which show significant variability in their CDR1 and CDR2 (Fig 2.4 [44]. Entropy analysis was complemented by an analysis of the mutation frequency at the nucleotide level. As expected, for the CDR1 and CDR2, the average frequency of mutation of IGHV1-7 (5.23 %) was lower than that of any other IGHV1 subgroup member (5.76 % to 9.37%) (Table 2.1). This decrease in the mutation frequency probably reflects a decrease in the selection process in agreement with the lower nonsynonymous (replacement) mutation observed in the IGHV1-7 CDR1 and CDR2. Thus, based on the mutation pattern and entropy analysis, the CDR1 and CDR2 may not interact with antigen in the majority of VH with ultralong CDR3.

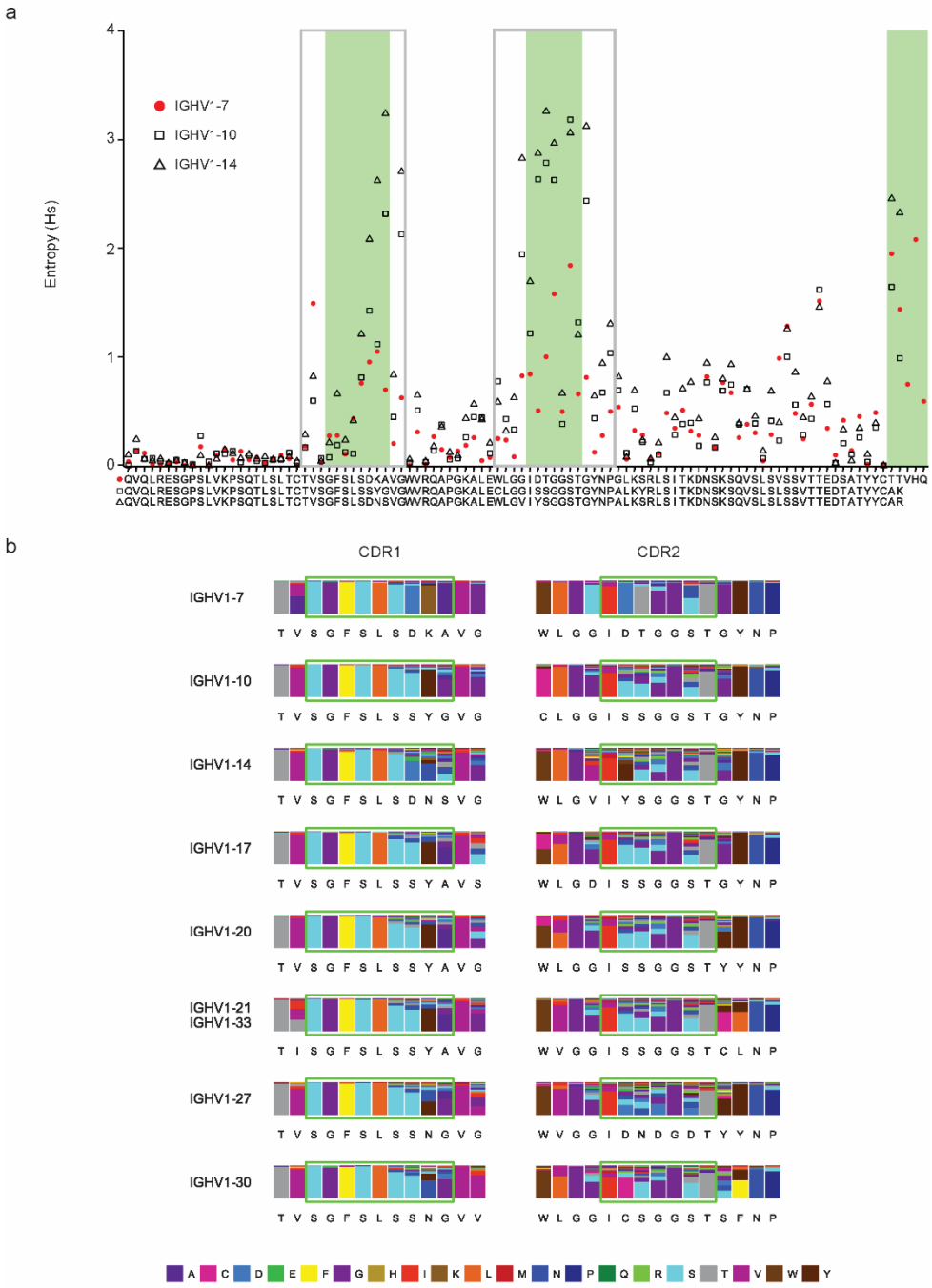


Figure 2.4 IGHV1-7 has low variability in CDR1 and CDR2. (a) Graph of Shannon entropy values for IGHV1-7 (red circles), IGHV1-10 (triangles), and IGHV1-14 (squares). IGHV1-7 was selected as it encoded the majority of ultralong VH CDR3. IGHV1-14 and IGHV1-10 were selected for comparison as they encoded the most and least diverse remaining IGHV regions, respectively. The CDR-IMGT are indicated by a green box while the grey boxes are expanded upon to allow the visual comparison of amino acid frequencies shown in **b.** The amino-acid diversity for IGHV1 members in b corroborates the entropy results of a. IGHV1-25/IGHV1-37 are not shown because they were not used in any transcripts (Figure 2).

Table 3.1 Mutation rates for the framework FR (FR1 to FR3) and CDR (CDR1 and CDR2) regions

	FR1	*CDR1*	*FR2*	*CDR2*	*FR3*	*FR1 to FR3 average*	*CDR1 and CDR2 average*	*Total average*
IGHV1-7	2.46%	4.36%	4.23%	6.10%	3.54%	3.41%	5.23%	4.14%
IGHV1-10	0.65%	3.97%	3.34%	10.15%	6.63%	3.54%	7.06%	4.95%
IGHV1-14	1.32%	6.82%	3.54%	10.22%	4.74%	3.20%	8.52%	5.33%
IGHV1-17	1.39%	6.06%	8.76%	10.28%	4.44%	4.86%	8.17%	6.19%
IGHV1-20	1.37%	8.62%	5.43%	10.11%	4.61%	3.80%	9.37%	6.03%
IGHV1-21/IGHV1-33	1.73%	5.59%	3.95%	9.91%	4.75%	3.48%	7.75%	5.19%
IGHV1-27	1.19%	6.85%	6.42%	10.45%	4.66%	4.09%	8.65%	5.91%
IGHV1-30	0.96%	5.41%	4.03%	8.93%	4.39%	3.13%	7.17%	4.74%
IGHV1-39	3.88%	3.68%	3.28%	7.84%	4.08%	3.75%	5.76%	4.55%

Rates are given as the percentage of total nucleotides contributing to each respective region of the VH domain. CDR3 and FR4 are not shown. Delimitations of the FR and CDR are according to IMGT.[2] V-REGIONS IGHV1-25 and IGHV1-37 are not included in the table because they were not found to be utilized in the repertoire.

2.3 Discussion

The ultralong VH CDR3 of cattle provide a novel paradigm for creating diversity in immunoglobulins[19], and have unique importance in being able to broadly neutralize HIV during an immune response[1]. While antibodies of all other well-studied vertebrates have a traditional structure comprised of a relatively short 5-18 residue CDR3 loop, cattle can encode CDR3 of over 70 amino acids, with crystal structures of five antibodies revealing that they all have a β-ribbon "stalk" and disulfide bonded "knob" structure [19]. With such remarkably different structures compared to normal antibodies, the genes encoding these antibodies have features distinct from those of other species. Here we examined the genetic underpinnings of VH with ultralong CDR3, and found (i) a novel germline 8-nucleotide duplication in IGHV1-7, enabling the formation of the ascending stalk, (ii) that IGHV1-7 is almost exclusively used in VH with ultralong CDR3, (iii) that SHM diversity in CDR1 and CDR2 is significantly reduced in VH with ultralong CDR3, suggesting that nearly all variability and antigen binding reside in CDR3, and (iv) that a novel deletional mechanism, internal to the IGHD8-2

region, alters loop lengths in CDR3, further diversifying the knob domain. Additionally, shorter CDR3 sequences appear to preferentially use IGHV1-10.

Recently the resequencing of the IGH locus of *Bos taurus* revealed twelve functional IGHV genes, all of which are members of the IGHV1 subgroup [28]. Within this subgroup a unique V region, IGHV1-7, has an extension at the 3' end that is shown here to be the result of an internal 8-nucleotide duplication (Fig 1b). This duplication both extends the length and shifts the reading frame for at least four amino acids, and the resulting extension plays an integral role in the formation of the ascending stalk of ultralong VH CDR3 (≥40AA). No ultralong CDR3 were observed in which a functional knob was found in the absence of the stalk-initiating TTVHQ (or similar) motif. The importance of this motif to the structure of ultralong CDR3 is evident not only in the high use of IGHV1-7 with a germline encoded TTVHQ motif (Fig 2a), but also in analogous motifs observed in all but one (N=863 functional rearrangements) instance of ultralong CDR3 encoding transcripts. The IGHV1-7 gene is nearly exclusively tied to ultralong CDR3 encoding antibodies, being utilized in 97.2% of these sequences. However IGHV1-7 is not exclusively recombined to the long IGHD8-2 region as 9.3% of IGHV1-7 containing transcripts encode a shorter CDR3. In this regard, a defining feature of VH with ultralong CDR3 is the IGVH1-7-IGHD8-2 rearrangement, with IGHD8-2 apparently not often used with other IGHV regions (Fig 2a). As expression of the novel stalk and knob regions may require IGHV1-7 and IGHD8-2, respectively, these rearrangements may be the only recombination events which survive and encode an ultralong CDR3.

A surprising finding was that shorter CDR3 sequences also had preferential use of one IGHV region. Unlike ultralong VH CDR3 sequences, which nearly exclusively use IGHV1-7, shorter CDR3 sequences prefer IGHV1-10. As little structural or functional data exist for these shorter CDR3 antibodies, it is unclear why this IGHV region is preferred.

The bovine IGH locus houses only 12 functional and closely related IGHV1 genes, and relies on AID induced mutation of naïve B cells to drive repertoire formation [19, 28-31, 42, 45, 46]. Of these twelve sequences two, IGHV1-21 and IGHV1-33, are identical at the nucleotide level, while another two, IGHV1-25 and IGHV1-37, would encode identical peptides. A low mutation frequency was observed in the CDR1 and CDR2 of the IGHV1-7 gene expressed in VH with ultralong CDR3, however massive mutation was present within CDR3, making these regions extremely divergent from the germline IGHD8-2. This contrasts to typical B cells in other species, as well as those bearing a shorter VH CDR3 in cows, which show significant amino acid variability in CDR1 and CDR2 (Fig 4, Table 1). This supports previous evidence that the knob is the sole antigen recognition site of the VH. CDR1 and CDR2 are posited to play framework-like roles in supporting and stabilizing the stalk, allowing the knob of ultralong CDR3 to carry out binding alone. Across all VH domains analyzed, AID mediated SHM clearly plays a role in shaping the bovine repertoire. This is evident from the increase in entropy and frequency of both nonsynonymous and synonymous mutation throughout the entirety of the sequences (Fig 4, Supplemental Figure 4, Table 1). The frequency of mutation and entropy was higher in the CDR than in the framework regions for non-

IGHV1-7 transcripts, indicative of these amino acids being important for antigen binding as other amino acids appear to be conserved for structural integrity.

Importantly, a novel, potentially AID-catalyzed mechanism for diversification has been discovered that specifically alters the knob of ultralong VH CDR3 through large interior deletions. AID is known to catalyze short insertions and deletions during SHM [47-54], however not with the frequency or size of deletions reported here. Such deletions could provide considerable diversity within the knob region of ultralong VH CDR3. Most internal deletions necessarily alter the three-dimensional placement of cysteines, and thus could play a role in altering disulfide patterns and their associated loops within the knob domain. Furthermore, recent structural analysis revealed that the knob domains have a small three stranded β-sheet at their core, with associated loops between the strands [18]. The loops themselves differ in length and amino acid content. The potential of altering these loop lengths is another mechanism whereby deletions could contribute to diversity of the ultralong VH CDR3.

The IGHD8-2 deletions defined here are likely somatically generated. While longer polymorphic IGHD8-2 genes, encoding up to an additional four codons, have been discovered through genomic sequencing of muscle, no shorter IGHD8-2 polymorphisms have been identified in a non-rearranging cell [42]. Many germline-encoded polymorphic copies of a long IGHD would have to be present on a chromosome to explain the length range covered by the deletions discovered here. Furthermore, previous analysis of deep sequence heavy chain transcripts revealed that all ultralong

VH CDR3 derived from a single IGHD region [19]. The repetitive nature of IGHD8-2 (32.6% of in-frame codons are TAT while 30.6% of codons are GGT), vast number of AID hotspots, and high mutability (ultralong VH CDR3 sequences were found to have an average pairwise identity of 58.8%, Supplemental Figure 6), suggests that these events could be attributed to strand slippage events commonly associated with AID activity [29, 30, 42, 55].

The process resulting in the deletions is likely an AID mediated mechanism, furthering the scope of the master enzyme of secondary diversification. Strand slippage is one process that could be attributed to the smaller deletions resulting in a fine tuning of the knob, however strand slippage events are generally restricted to small deletions (up to six nucleotides) [55]. Unsuccessful CSR events, known as resection events, resulting in smaller genomic deletions within a switch region, are documented to occur [43]. The mutational load observed in the VH domain is clear evidence for high levels of AID activity within bovine B cells (Table 2). The deletion events observed in IGHD8-2 could be mechanistically similar to the resection events of a failed class switch [43], which would allow AID and associated machinery to produce double strand DNA breaks within IGHD8-2 containing CDR3 [39, 56]. Recently, Yeap, et.al. reported deletions in V-region transgenes as a result of double-strand breaks during SHM [47]. These were mediated by nearby SHM hotspots which may be analogous to the multiple hotspots in IGHD8-2. The result is the deletion of genomic material in a manner similar to resection events which result in nonproductive CSR. Deletion of interior codons allows all structural components (CPDG and YxYxY/alternating aromatic amino acids) to be

conserved while altering the knob by removal of amino acids and change of the folding pattern. With evidence to support the knob being the sole recognition site of an ultralong VH CDR3, the deletion events would serve to vastly expand the pool of recognizable antigens in a system limited by a relatively low number of V-(D)-J recombinations. The theory of why the structurally unique ultralong CDR3 antibodies evolved in cattle is of considerable interest. There are at least two broad possibilities underlying the evolution of these antibodies. First, this system may have been selected to provide a mechanism for enhanced diversity in the antibody repertoire. Given the severely limited VDJ segmental diversity within the *B. taurus* IgH locus, ultralong VH CDR3 antibodies provide greater potential for maximum diversification with relatively little waste. In contrast to a canonical antibody which potentially requires mutations in all six CDRs to alter the paratope, an ultralong VH CDR3 antibody can radically alter its binding surface with few mutations. Indeed, a single mutation to or from cysteine, or a deletion event, could dramatically alter loop structures within the knob. Since a single VDJ event can ultimately produce enormous diversity through SH, this process could allow for more efficient expansion of an antibody repertoire. Thus, it would seem that this novel structure and genetic mechanism evolved in cattle as a way to supplement the poor repertoire diversity available genomically. Second, the novel structure may have evolved in response to specific bovine pathogens. The digestive system of cows utilizes a large rumen compartment with symbiotic microorganisms, including substantial bacteria and protozoa, that serve to digest cellulose and other feedstuff. This unusual antigenic load may have been an immunologic driver for the ultralong VH CDR3

structure. Alternatively, several infectious agents, including retroviruses, naturally infect bovines. Given the broadly neutralizing antibody response that cows can produce to HIV (1), it stands to reason that a potential evolutionary driver of this novel antibody system could be to enable cross-protective responses against related strains of microorganisms or viruses. While these evolutionary factors are speculative, only two genetic events, the 8-basepair duplication forming IGHV1-7 and the advent of the long IGHD8-2 gene, appear required for forming the entire ultralong VH CDR3 antibody system.

In conclusion, the data reported here describe key immunogenetic properties of ultralong VH CDR3 formation used at the bovine IGH locus and unveil a new mechanism to diversify them. For long, we have understood how CDR3 lengths are shortened by exonuclease activity and elongated by N nucleotide addition in rodent and primate antibody genes. This bovine IGH locus has perhaps evolved extreme mechanisms at the DNA level for the creation of structurally-sound projecting microdomains within VH CDR3 and drastic increase of their diversity by internal truncation, altering loop lengths and disulfide patterns. Future work will focus on elucidating all steps involved in truncation events and determining the role that the unique ultralong VH CDR3 B cell subset plays within the bovine immune system.

2.4 Materials and Methods

2.4.1 Collection of blood samples, isolation of PBMC, RNA and synthesis of 5' RACE libraries

Tissue—blood, Peyer's patch, spleen, and bone marrow—were derived from two adult cows housed at Texas A&M University Veterinary Medical Park under approved Animal Use Protocol 2015-0078. Peripheral blood mononuclear cells (PBMC) were isolated from blood with lymphocyte separation media (Mediatech Inc, Tewksbury, MA) and total RNA extraction was performed on the isolated PBMC with the RNeasy mini kit (Qiagen Valencia, CA) as previously described [57]. Isolated RNA was used as the template for synthesis of 5' RACE libraries with the GeneRacer kit (Invitrogen, Carlsbad CA) performed as previously described [58]. An equal mix of oligoDT and random hexamer primers was used to prime the reaction.

2.4.2 Initial amplification of IGHV1-7 rearranged transcripts

Polymerase Chain Reaction amplification was performed on cDNA with primers designed to target the unique IGHV1-7 and the CH1 region of IGHM and IGHG (Appendix table 1) and cloned as previously described [59]. Briefly, the resulting product was visualized and extracted from an agarose gel using the PureLink Gel Extraction Kit (Life Technologies Carlsbad, CA). This PCR product was ligated into the pCRII plasmid using the TOPO-TA Cloning Kit (Invitrogen) and cloned into *E. coli* TOP10 cells (Invitrogen) according to the manufacturer's protocol. Transformed cells were plated on LB plates containing carbenicillin for plasmid selection and X-gal for

colony differentiation. White colonies containing plasmid and insert were selected and grown in 3ml LB broth containing ampicillin for 16 hours. Plasmids were isolated using the Zyppy Plasmid Miniprep Kit (ZYMO Irvine, CA) according to the manufacturer's protocol. Plasmid inserts were freed with *Eco*RI (NEB, Ipswitch MA) and resolved on an agarose gel. The BigDye terminator with M13 primers and BigDYE Xterminator (ThermoFisher) were used for generation of sequencing products and cleanup respectively. Sanger sequencing reactions were resolved by the Gene Technologies Laboratory at Texas A&M University.

2.4.3 Amplification of IgH transcripts and PacBio deep sequencing

The cDNA template produced in the 5' RACE libraries was used as a template for PCR using the Phusion high fidelity polymerase (NEB, Ipswitch, MA), with the barcoded primers in Supplemental Table 1. The PCR protocol was performed in two steps with an initial denaturation of 2 minutes at 95˚C, cycles of 95˚C for 15 seconds followed by an annealing/extension of 1 minute at 72˚C, and a final step of 5 minutes at 72˚C. Products of 450-650 bp were visualized on an agarose gel and extracted. Pooled samples were sent to the Duke University Center for Genomic and Computational Biology core center for PacBio library preparation and sequencing. Circular consensus sequences (CCS), sequences in which the PacBio polymerase circled the insert at least three times, were returned in fastq format. The resulting fastq files were imported into Geneious V9 (Biomatters, Auckland New Zealand) where the barcoded primers were used to de-multiplex the samples. Finally the sequences were quality filtered (Q>20) and

homopolymer runs were corrected using the ACACIA program [60]. The cattle IGHM and IGHG sequences were visualized and aligned in the Geneious software suite.

2.4.4 Identification of IgH genes

The V, D and J gene use of each sequence was determined using a custom BLAST database composed of all IGH genes in the KT723008 assembly of the bovine IGH locus on chromosome 21 [28]. The V and J genes used were determined via BLASTn. BLASTn was selected in contrast to tBLAST or Protein BLAST as slight codon changes could result in improper identification, especially in the mutated VH. For IGHV identification, a smaller word size of seven for BLASTn, and utilization of bit score, in lieu of percent identity, was used for gene calling.

2.4.5 Shannon entropy, mutation analysis and statistical testing

Shannon entropy values were determined from amino acid alignments using the "bio3d" package in the R software suite version 3.1.1 [61, 62]. The resulting data was graphed in R using the "ggplot2" package [63]. Mutation analysis was performed on nucleotide alignments using the Geneious SNP analysis tool. Statistically significant differences of the VH CDR3 lengths were determined via ANOVA and post-hoc Tukey HSD test in R.

2.5 Bovine gene conversion addendum

In addition to regulating the secondary diversification processes somatic hypermutation (SHM) and class switch recombination (CSR), activation induced cytidine deaminase (AID) is integral to a less common process termed immunoglobulin gene conversion (IGC) [64]. Similar to SHM, IGC generates diversity in the variable region, altering the binding specificity of the antigen receptor. However IGC incorporates alterations on a larger scale through the exchange of genetic material between a RAG rearranged V-segment and an upstream, often pseudogenized, V-segment [64]. Initially described in the chicken [65], IGC acts to diversify the primary repertoires of species with limited combinatorial (V-D-J) diversity including rabbits [66] and cattle [67, 68]. The initial study confirming IGC in the immunoglobulin heavy chain (IgH) was performed using an incomplete locus in which the orientation and presence of all V-segments was unknown [68]. Therefore, we set out to add confidence to the findings in the previous study applying the same methods to our dataset using the completed version of the *B. taurus* IgH locus [28].

Due to the similarity of the IGHV1 family in cattle, we employed a conservative approach in determining IGC in an attempt to differentiate true conversion tracts from SHM. In addition to the straightforward assigning FR and CDR translations to a specific V-segment as described by Walther [68], we added several additional constraints to this conservative estimate. We first restricted the definition of IGC to events in which at least two codons, 6bp, were exchanged. Next, the differences between parent and donor V-segment must not be attainable through mutations at AID, ‘RGYW/WRCY’, hotspots.

Finally, we required that percent divergence between a donor V-segment transcript and germline could be no greater than the mutation rate reported for the parent V-segment. For example, the mutation rate within CDR1 of transcripts encoding IGHV1-7 as the origin V-segment was 4.36%. To meet the conservative guidelines, nucleotide identity within CDR1 must be at least 95.64% compared to the germline donor V-segment. The purpose of these restrictions was to delineate true IGC events from multiple rounds of SHM. The results of our conservative approach compared to Walther's relaxed methods are found in Table 2.2.

Table 2.2 Bovine gene conversion V segment table

Framework 1					**CDR1**			
Origin V	Donor V	Walther	Conservative		Origin V	Donor V	Walther	Conservative
IGHV1-07	IGHV1-21;33	31	1		IGHV1-07	IGHV1-10	12	7
	ψIGHV1-32	1	0			IGHV1-17;20;21;33	17	13
IGHV1-10	IGHV1-21;33	493	0			IGHV1-27;30	4	4
	ψIGHV1-32	10	0			ψIGHV1-32	6	4
IGHV1-14	IGHV1-21;33	18	0			IGHV1-39/ψIGHV1-43;46	1	1
IGHV1-17	IGHV1-21;33	20	0		IGHV1-10	IGHV1-17;20;21;33	638	0
IGHV1-20	IGHV1-21;33	11	0			IGHV1-27;30	66	0
	IGHV1-25;37	1	1			ψIGHV1-32	5	3
	ψIGHV1-32	3	0			IGHV1-39;ψIGHV1-43;46	48	0
IGHV1-21	IGHV1-27;30	287	0		IGHV1-14	IGHV1-17;20;21;33	1	0
	ψIGHV1-32	1	0			IGHV1-27;30	1	1
IGHV1-27	IGHV1-33	35	0			ψIGHV1-32	16	0
	ψIGHV1-32	2	1		IGHV1-17	IGHV1-27;30	61	18
IGHV1-30	IGHV1-33	4	0			IGHV1-39/ψIGHV1-43;46	1	1
					IGHV1-20	IGHV1-27;30	4	3
Framework 2						ψIGHV1-32	1	1
Origin V	Donor V	Walther	Conservative		IGHV1-21	IGHV1-27;30	47	37
IGHV1-07	IGHV1-10	7	0			ψIGHV1-32	1	1
	IGHV1-14	2	0			IGHV1-39/ψIGHV1-43;46	2	1
	IGHV1-21;27;33	1	0		IGHV1-27	IGHV1-33	54	52
	IGHV1-30	2	0			ψIGHV1-32	2	1
	IGHV1-39	1	0					
IGHV1-10	IGHV1-20	58	0		**CDR2**			
	IGHV1-14	1	0		Origin V	Donor V	Walther	Conservative
	IGHV1-21;27;33	2	1		IGHV1-07	IGHV1-10;21;33	2	1
	IGHV1-30	1	0			IGHV1-27	3	3
	IGHV1-39	1	0			ψIGHV1-32	2	2
IGHV1-17	IGHV1-39	4	0			IGHV1-39	11	3
IGHV1-20	IGHV1-21;27;33	4	0		IGHV1-10	ψIGHV1-32	5	5
	IGHV1-30	3	0			IGHV1-39	20	4
	IGHV1-39	1	0		IGHV1-14	IGHV1-21;33	2	2
IGHV1-21	IGHV1-30	2	1			IGHV1-30	1	0
	IGHV1-39	1	0			IGHV1-39	1	0
IGHV1-27	IGHV1-30	1	0		IGHV1-20	IGHV1-39	4	2
					IGHV1-21	IGHV1-27	1	1
						IGHV1-39	3	1
					IGHV1-27	ψIGHV1-32	4	1
						IGHV1-39	1	1
					IGHV1-30	IGHV1-33;37/ψIGHV1-43;46	2	2

As expected, the conservative estimate was considerably lower than the methods used by Walther, especially within framework regions where high identity of genomic parent/donor V-segments and high mutation rate across the dataset severely limited our conservative estimates (Table 2.2). Surprisingly all IGC events, regardless of the method used, involved parent and donor V-segments belonging to the IGHV1 family. This could be due to limitations of the homologous recombination machinery following AID induced lesions [64]. Conversion events were concentrated in the CDRs where events ranged from 10 to 74-fold higher compared to FRs. The high degree of FR1 conversions mirrored Walther's reported counts using ambiguous donor V-segment assignments assigned strictly through highest nucleotide similarity. Interestingly when aligning the entire V-segment, both donor and parent, to confirm IGC, the scope of the exchanged tracts often bled from CDRs into the FR regions. This is likely the result of nearly all V-segment divergence localizing to the 3'-CDR1/5'-FR2, 3'-FR2/5'-CDR2, or 3'-CDR2/3'-FR3 regions. Determination of the exact length of genetic material exchange from the donor V-segment is impossible using transcriptional data alone.

The process of AID induced IGC diversifying IgH V-segments in *B. taurus* was previously confirmed without a complete genomic IgH locus [68]. Analysis of our dataset corroborated this claim using the completed IgH locus to determine logical IGC events [28]. The spatial relations of the V-segments within the IgH locus allowed us to expand the scope of IGC beyond events strictly involving pseudogene donor V-segments. Furthermore, we applied an additional set of restrictions more accommodating for use in larger scale, high throughput sequencing experiments. Implementation of

additional restriction drastically reduced the raw number of IGC events, compared to methods previously described, but also increased confidence that the observations were truly IGC events as opposed to multiple rounds of SHM.

Diversification of *B. taurus* IgH V-segments via IGC was confirmed in our dataset as utilizing both functional and pseudogene V-segments. This is an additional AID mediated processes presumptively co-opted in cattle to rectify both high V-segment identity and limited combinatorial diversity within the IGH locus.

2.6 References

1. Sok, D., K.M. Le, M. Vadnais, K. Saye-Francisco, J.G. Jardine, J. Torres, Z.T. Berndsen, L. Kong, R. Stanfield, J. Ruiz, A. Ramos, C.-H. Liang, P.L. Chen, M.F. Criscitiello, W. Mwangi, I.A. Wilson, A.B. Ward, V.V. Smider, and D.R. Burton, *Rapid elicitation of broadly neutralizing antibodies to HIV by immunization in cows.* Nature, 2017. **advance online publication**.
2. Lefranc, M.P., *Immunoglobulin and T Cell Receptor Genes: IMGT((R)) and the Birth and Rise of Immunoinformatics.* Front Immunol, 2014. **5**: p. 22.
3. Lefranc, M.-P. and G. Lefranc, *The Immunoglobulin FactsBook.* 2001, London, UK: Academic Press. 458.
4. Lefranc, M.-P. and G. Lefranc, *The T cell receptor FactsBook*. 2001, London, UK: Academic Press. 398.
5. Rock, E.P., P.R. Sibbald, M.M. Davis, and Y.H. Chien, *CDR3 length in antigen-specific immune receptors.* J Exp Med, 1994. **179**(1): p. 323-8.

6. Morea, V., A. Tramontano, M. Rustici, C. Chothia, and A.M. Lesk, *Conformations of the third hypervariable region in the VH domain of immunoglobulins.* J Mol Biol, 1998. **275**(2): p. 269-94.

7. Ivanov, II, R.L. Schelonka, Y. Zhuang, G.L. Gartland, M. Zemlin, and H.W. Schroeder, Jr., *Development of the expressed Ig CDR-H3 repertoire is marked by focusing of constraints in length, amino acid use, and charge that are first established in early B cell progenitors.* J Immunol, 2005. **174**(12): p. 7773-80.

8. Lefranc, M.P., C. Pommie, M. Ruiz, V. Giudicelli, E. Foulquier, L. Truong, V. Thouvenin-Contet, and G. Lefranc, *IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains.* Dev Comp Immunol, 2003. **27**(1): p. 55-77.

9. Hamers-Casterman, C., T. Atarhouch, S. Muyldermans, G. Robinson, C. Hamers, E.B. Songa, N. Bendahman, and R. Hamers, *Naturally occurring antibodies devoid of light chains.* Nature, 1993. **363**(6428): p. 446-8.

10. Stanfield, R.L., H. Dooley, P. Verdino, M.F. Flajnik, and I.A. Wilson, *Maturation of shark single-domain (IgNAR) antibodies: evidence for induced-fit binding.* J Mol Biol, 2007. **367**(2): p. 358-72.

11. Stanfield, R.L., H. Dooley, M.F. Flajnik, and I.A. Wilson, *Crystal structure of a shark single-domain antibody V region in complex with lysozyme.* Science, 2004. **305**(5691): p. 1770-3.

12. de los Rios, M., M.F. Criscitiello, and V.V. Smider, *Structural and genetic diversity in antibody repertoires from diverse species.* Curr Opin Struct Biol, 2015. **33**: p. 27-41.

13. Rasmussen, S.G., H.J. Choi, J.J. Fung, E. Pardon, P. Casarosa, P.S. Chae, B.T. Devree, D.M. Rosenbaum, F.S. Thian, T.S. Kobilka, A. Schnapp, I. Konetzki, R.K. Sunahara, S.H. Gellman, A. Pautsch, J. Steyaert, W.I. Weis, and B.K. Kobilka, *Structure of a nanobody-stabilized active state of the beta(2) adrenoceptor.* Nature, 2011. **469**(7329): p. 175-80.

14. Berens, S.J., D.E. Wylie, and O.J. Lopez, *Use of a single VH family and long CDR3s in the variable region of cattle Ig heavy chains.* Int Immunol, 1997. **9**(1): p. 189-99.

15. Saini, S.S., B. Allore, R.M. Jacobs, and A. Kaushik, *Exceptionally long CDR3H region with multiple cysteine residues in functional bovine IgM antibodies.* Eur J Immunol, 1999. **29**(8): p. 2420-6.

16. Saini, S.S., W. Farrugia, P.A. Ramsland, and A.K. Kaushik, *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, 2003. **15**(7): p. 845-53.

17. Saini, S.S. and A. Kaushik, *Extensive CDR3H length heterogeneity exists in bovine foetal VDJ rearrangements.* Scand J Immunol, 2002. **55**(2): p. 140-8.

18. Stanfield, R.L., I.A. Wilson, and V.V. Smider, *Conservation and diversity in the ultralong third heavy-chain complementarity-determining region of bovine antibodies.* Science Immunology, 2016. **1**(1).

19. Wang, F., D.C. Ekiert, I. Ahmad, W. Yu, Y. Zhang, O. Bazirgan, A. Torkamani, T. Raudsepp, W. Mwangi, M.F. Criscitiello, I.A. Wilson, P.G. Schultz, and V.V. Smider, *Reshaping antibody diversity.* Cell, 2013. **153**(6): p. 1379-93.

20. Schroeder, H.W., Jr., J.L. Hillson, and R.M. Perlmutter, *Structure and evolution of mammalian VH families.* Int Immunol, 1990. **2**(1): p. 41-50.

21. Schatz, D.G., M.A. Oettinger, and D. Baltimore, *The V(D)J recombination activating gene, RAG-1.* Cell, 1989. **59**(6): p. 1035-1048.

22. Tonegawa, S., *Somatic generation of antibody diversity.* Nature, 1983. **302**(5909): p. 575-581.

23. Ruiz, M., N. Pallarès, V. Contet, V. Barbié, and M.P. Lefranc, *The Human Immunoglobulin Heavy Diversity (IGHD) and Joining (IGHJ) Segments.* Experimental and Clinical Immunogenetics, 1999. **16**(3): p. 173-184.

24. Matsuda, F., K. Ishii, P. Bourvagnet, K. Kuma, H. Hayashida, T. Miyata, and T. Honjo, *The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus.* J Exp Med, 1998. **188**(11): p. 2151-62.

25. Kabat, E.A., *Unique features of the variable regions of Bence Jones proteins and their possible relation to antibody complementarity.* Proc Natl Acad Sci U S A, 1968. **59**(2): p. 613-9.

26. Wu, T.T. and E.A. Kabat, *An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity.* The Journal of experimental medicine, 1970. **132**(2): p. 211-50.

27. Muramatsu, M., K. Kinoshita, S. Fagarasan, S. Yamada, S. Shinkai, and T. Honjo, *Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme.* Cell, 2000. **102**.

28. Ma, L., T. Qin, D. Chu, X. Cheng, J. Wang, X. Wang, P. Wang, H. Han, L. Ren, R. Aitken, L. Hammarstrom, N. Li, and Y. Zhao, *Internal Duplications of DH, JH, and C Region Genes Create an Unusual IgH Gene Locus in Cattle.* J Immunol, 2016. **196**(10): p. 4358-66.

29. Liljavirta, J., A. Ekman, J.S. Knight, A. Pernthaner, A. Iivanainen, and M. Niku, *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, 2013. **6**(5): p. 942-9.

30. Verma, S. and R. Aitken, *Somatic hypermutation leads to diversification of the heavy chain immunoglobulin repertoire in cattle.* Vet Immunol Immunopathol, 2012. **145**(1–2): p. 14-22.

31. Sun, Y., Z. Liu, L. Ren, Z. Wei, P. Wang, N. Li, and Y. Zhao, *Immunoglobulin genes and diversity: what we have learned from domestic animals.* J Anim Sci Biotechnol, 2012. **3**(1): p. 18.

32. Kozuka, Y., T. Nasu, T. Murakami, and M. Yasuda, *Comparative studies on the secondary lymphoid tissue areas in the chicken bursa of Fabricius and calf ileal Peyer's patch.* Vet Immunol Immunopathol, 2010. **133**(2-4): p. 190-7.

33. Ekman, A., T. Pessa-Morikawa, J. Liljavirta, M. Niku, and A. Iivanainen, *B-cell development in bovine fetuses proceeds via a pre-B like cell in bone marrow and lymph nodes.* Dev Comp Immunol, 2010. **34**(8): p. 896-903.

34. Kaushik, A.K., M.E. Kehrli, Jr., A. Kurtz, S. Ng, M. Koti, F. Shojaei, and S.S. Saini, *Somatic hypermutations and isotype restricted exceptionally long CDR3H contribute to antibody diversification in cattle.* Vet Immunol Immunopathol, 2009. **127**(1-2): p. 106-13.

35. Yasuda, M., C.N. Jenne, L.J. Kennedy, and J.D. Reynolds, *The sheep and cattle Peyer's patch as a site of B-cell development.* Vet Res, 2006. **37**(3): p. 401-15.

36. Neill, J.D., J.F. Ridpath, and E. Liebler-Tenorio, *Global gene expression profiling of Bovine immature B cells using serial analysis of gene expression.* Anim Biotechnol, 2006. **17**(1): p. 21-31.

37. Koti, M., G. Kataeva, and A. Kaushik, *Organization of DH-gene locus is distinct in cattle.* Dev. Biol., 2008. **132**: p. 307-313.

38. Koti, M., G. Kataeva, and A.K. Kaushik, *Novel atypical nucleotide insertions specifically at VH-DH junction generate exceptionally long CDR3H in cattle antibodies.* Mol Immunol, 2010. **47**(11-12): p. 2119-28.

39. Rogozin, I.B. and M. Diaz, *Cutting edge: DGYW/WRCH is a better predictor of mutability at G:C bases in Ig hypermutation than the widely accepted RGYW/WRCY motif and probably reflects a two-step activation-induced cytidine deaminase-triggered process.* J Immunol, 2004. **172**(6): p. 3382-4.

40. Hosseini, A., G. Campbell, M. Prorocic, and R. Aitken, *Duplicated copies of the bovine JH locus contribute to the Ig repertoire.* Int Immunol, 2004. **16**(6): p. 843-52.

41. Koti, M., G. Kataeva, and A.K. Kaushik, *Organization of D(H)-gene locus is distinct in cattle.* Dev Biol (Basel), 2008. **132**: p. 307-13.

42. Liljavirta, J., M. Niku, T. Pessa-Morikawa, A. Ekman, and A. Iivanainen, *Expansion of the preimmune antibody repertoire by junctional diversity in Bos taurus.* PLoS One, 2014. **9**(6): p. e99808.

43. Dong, J., R.A. Panchakshari, T. Zhang, Y. Zhang, J. Hu, S.A. Volpi, R.M. Meyers, Y.-J. Ho, Z. Du, D.F. Robbiani, F. Meng, M. Gostissa, M.C. Nussenzweig, J.P. Manis, and F.W. Alt, *Orientation-specific joining of AID-initiated DNA breaks promotes antibody class switching.* Nature, 2015. **525**(7567): p. 134-139.

44. Wu, T.T. and E.A. Kabat, *An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity.* J Exp Med, 1970. **132**(2): p. 211-50.

45. Niku, M., J. Liljavirta, K. Durkin, E. Schroderus, and A. Iivanainen, *The bovine genomic DNA sequence data reveal three IGHV subgroups, only one of which is functionally expressed.* Dev Comp Immunol, 2012. **37**(3-4): p. 457-61.

46. Butler, J.E., *Immunoglobulin diversity, B-cell and antibody repertoire development in large farm animals.* Rev Sci Tech, 1998. **17**(1): p. 43-70.

47. Yeap, L.S., J.K. Hwang, Z. Du, R.M. Meyers, F.L. Meng, A. Jakubauskaite, M. Liu, V. Mani, D. Neuberg, T.B. Kepler, J.H. Wang, and F.W. Alt, *Sequence-Intrinsic Mechanisms that Target AID Mutational Outcomes on Antibody Genes.* Cell, 2015. **163**(5): p. 1124-37.

48. Meissner, F. and M. Mann, *Quantitative shotgun proteomics: considerations for a high-quality workflow in immunology.* Nat Immunol, 2014. **15**(2): p. 112-117.

49. Kepler, T.B., H.X. Liao, S.M. Alam, R. Bhaskarabhatla, R. Zhang, C. Yandava, S. Stewart, K. Anasti, G. Kelsoe, R. Parks, K.E. Lloyd, C. Stolarchuk, J. Pritchett, E. Solomon, E. Friberg, L. Morris, S.S. Karim, M.S. Cohen, E. Walter, M.A. Moody, X. Wu, H.R. Altae-Tran, I.S. Georgiev, P.D. Kwong, S.D. Boyd, A.Z. Fire, J.R. Mascola, and B.F. Haynes, *Immunoglobulin gene insertions and deletions in the affinity maturation of HIV-1 broadly reactive neutralizing antibodies.* Cell Host Microbe, 2014. **16**(3): p. 304-13.

50. Briney, B.S., J.R. Willis, and J.E. Crowe, Jr., *Location and length distribution of somatic hypermutation-associated DNA insertions and deletions reveals regions of antibody structural plasticity.* Genes Immun, 2012. **13**(7): p. 523-9.

51. Reason, D.C. and J. Zhou, *Codon insertion and deletion functions as a somatic diversification mechanism in human antibody repertoires.* Biology Direct, 2006. **1**(1): p. 24.

52. Wilson, P.C., O. de Bouteiller, Y.J. Liu, K. Potter, J. Banchereau, J.D. Capra, and V. Pascual, *Somatic hypermutation introduces insertions and deletions into immunoglobulin V genes.* J Exp Med, 1998. **187**(1): p. 59-70.

53. Wilson, P., Y.J. Liu, J. Banchereau, J.D. Capra, and V. Pascual, *Amino acid insertions and deletions contribute to diversify the human Ig repertoire.* Immunol Rev, 1998. **162**: p. 143-51.

54. Goossens, T., U. Klein, and R. Kuppers, *Frequent occurrence of deletions and duplications during somatic hypermutation: implications for oncogene translocations and heavy chain disease.* Proc Natl Acad Sci U S A, 1998. **95**(5): p. 2463-8.

55. Criscitiello, M.F., R. Benedetto, A. Antao, M.R. Wilson, V.G. Chinchar, N.W. Miller, L.W. Clem, and T.J. McConnell, *Beta 2-microglobulin of ictalurid catfishes.* Immunogenetics, 1998. **48**(5): p. 339-343.

56. Yeap, L.-S., Joyce K. Hwang, Z. Du, Robin M. Meyers, F.-L. Meng, A. Jakubauskaitė, M. Liu, V. Mani, D. Neuberg, Thomas B. Kepler, Jing H. Wang,

and Frederick W. Alt, *Sequence-Intrinsic Mechanisms that Target AID Mutational Outcomes on Antibody Genes.* Cell, 2015. **163**(5): p. 1124-1137.

57. Criscitiello, M.F., Y. Ohta, M.D. Graham, J.O. Eubanks, P.L. Chen, and M.F. Flajnik, *Shark class II invariant chain reveals ancient conserved relationships with cathepsins and MHC class II.* Dev Comp Immunol, 2012. **36**(3): p. 521-33.

58. Mashoof, S., C. Pohlenz, P.L. Chen, T.C. Deiss, D. Gatlin, 3rd, A. Buentello, and M.F. Criscitiello, *Expressed IgH mu and tau transcripts share diversity segment in ranched Thunnus orientalis.* Dev Comp Immunol, 2014. **43**(1): p. 76-86.

59. Breaux, B., T.C. Deiss, P.L. Chen, M.P. Cruz-Schneider, L. Sena, M.E. Hunter, R.K. Bonde, and M.F. Criscitiello, *The Florida manatee (Trichechus manatus latirostris) immunoglobulin heavy chain suggests the importance of clan III variable segments in repertoire diversity.* Dev Comp Immunol, 2017. **72**: p. 57-68.

60. Bragg, L., G. Stone, M. Imelfort, P. Hugenholtz, and G.W. Tyson, *Fast, accurate error-correction of amplicon pyrosequences using Acacia.* Nat Meth, 2012. **9**(5): p. 425-426.

61. Grant, B.J., A.P. Rodrigues, K.M. ElSawy, J.A. McCammon, and L.S. Caves, *Bio3d: an R package for the comparative analysis of protein structures.* Bioinformatics, 2006. **22**(21): p. 2695-6.

62. R Core Team, *R: A language and environment for statistical computing.* 2014, R Foundation for Statistical Computing: Vienna, Austria.

63. Wickham, H., *ggplot2: elegant graphics for data analysis*. 2009: Springer, New York.

64. Tang, E.S. and A. Martin, *Immunoglobulin gene conversion: Synthesizing antibody diversification and DNA repair*. DNA Repair, 2007. **6**(11): p. 1557-1571.

65. Reynaud, C.-A., A. Dahan, V. Anquez, and J.-C. Weill, *Somatic hyperconversion diversifies the single VH gene of the chicken with a high incidence in the D region.* Cell, 1989. **59**(1): p. 171-183.

66. Becker, R.S. and K.L. Knight, *Somatic diversification of immunoglobulin heavy chain VDJ genes: evidence for somatic gene conversion in rabbits.* Cell, 1990. **63**(5): p. 987-97.

67. Parng, C.L., S. Hansal, R.A. Goldsby, and B.A. Osborne, *Gene conversion contributes to Ig light chain diversity in cattle.* J Immunol, 1996. **157**(12): p. 5478-86.

68. Walther, S., M. Tietze, C.-P. Czerny, S. König, and U.S. Diesterbeck, *Development of a Bioinformatics Framework for the Detection of Gene Conversion and the Analysis of Combinatorial Diversity in Immunoglobulin Heavy Chains in Four Cattle Breeds.* PLOS ONE, 2016. **11**(11): p. e0164567.

CHAPTER III

ANCIENT USE OF IMMUNOGLOBULIN VARIABLE DOMAINS CONTRIBUTES SIGNIFICANTLY TO THE T CELL RECEPTOR DELTA REPERTOIRE

3.1 Introduction

The nurse shark, *Ginglymostoma cirratum*, is a representative of the oldest vertebrate class, the chondricthyans, with an adaptive immune system based on immunoglobulin superfamily (IGSF) antigen receptor somatic gene rearrangement in lymphocytes. The antigen receptors of B and T cells have significant similarities through all jawed vertebrate lineages, including the cartilaginous fishes [1-3]. Shark BCR, or Ig, and TCR genes employ RAG mediated V(D)J recombination with segmental junctional diversified by TdT activity. In sharks both Ig and TCR [4] genes are modified by activation induced cytidine deaminase (AID) catalyzed somatic hypermutation (SHM) in response to antigen and repertoire generation, respectively [5-11]. Shark Ig heavy (IgH) chain gene loci exist in many clusters; 15 in nurse shark [12], but possibly >100 in other species [13]. The segmental ordering of an IgH cluster is V_H-D_1-D_2-J_H followed by a single set of C region exons which delineate each cluster as either IgM (Cμ), IgW (Cω, the ancestor of IgD [14, 15]), or IgNAR (C_{NAR}). This deviates from the typical IgH translocon topography ($V_{1\text{-}n}$-$D_{1\text{-}n}$-$J_{1\text{-}n}$…C_μ-C_δ-C_γ-C_ε-C_α) found in Euteleostomi lineages [16]. Although nurse shark IgH rearrangements are generally intra-cluster, occurring within a single VDDJ cluster, class switch recombination (CSR) was discovered to alter the C region class of VDJ rearrangements in the nurse shark [17].

The nurse shark has been used as a “primitive” model of the adaptive immune system for decades [18]. In this species, polygenic and polymorphic MHC [19, 20], four TCR chains including a doubly rearranging NARTCRδ [21], and multiple IgH and IgL chain isotypes, have been well characterized [3, 6, 22, 23]. Previously we described unusual cDNA rearrangements of IgHV to TCRδDJC, and extremely rare cases of IgHV to TCRδD to TCRαJC rearrangements [24]. These could be attributed to either *trans*-rearrangements, occurring across vast genomic distance between distinct TCR and Ig receptor loci, or rearrangements originating from an IgH variable segment or cluster nested in the TCRδ locus.

Rare (1/>200,000 peripheral blood lymphocytes) inter-chromosomal, inter-arm, or distal intra-chromosomal *trans*-rearrangements occur between different antigen receptor loci in mammals, despite regulatory mechanisms in place to prevent them [25]. An inversion of human chromosome 7 can bring TCRβ and γ loci, located at distinct telomeres, into proximity facilitating *trans*-rearrangements [26, 27]. Likewise, an inversion on human chromosome 14 juxtaposes the IgH and TCRαδ loci enabling similar chimeric rearrangements [28]. Such *trans*-rearrangements are associated with ataxia telangiectasia and childhood acute B lymphoblastic leukemia [29]. The rates of *trans*-rearrangement can increase 50- 100 fold when double-strand breaks occur during chemotherapy or radiation treatment of lymphoma and leukemia patients (reviewed in [30]).

There is some logic to IgH-TCRδ rearrangements as both antigen receptor chains have been documented to interact with nonpeptide antigens and display similar CDR3

length distributions [31]. The presence of VHδ segments, V segments more similar to IgHV than to TCRδV located in the TCRδ loci of a growing number of vertebrate lineages, furthers the narrative of TCRδ rearrangements tolerating an Ig V domain. In the frog *Xenopus*, "VHδ" genes abound in the TCRα/δ locus and are commonly found in TCRδ transcripts [32]. Such VHδ segments have also been detected in the TCRαδ locus of the platypus, passeriform birds, and the coelocanth [33-35]. In galliform birds, VHδ segments are encoded within a distinct locus with a separate TCRδ C region. The membrane-distal domain of shark NARTCRδ employs a V segment most similar to those of the cartilaginous fish-exclusive IgH class, IgNAR [22, 36]. Formulation of a functional NARTCRδ requires two RAG mediated VDJ recombination events. The membrane distal V domain if formed by the rearrangement of VDJ segments of the NARTCR lineage, which exist in VDJ clusters [36]. This NARTCR domain exon is then spliced into a V domain exon formed by rearrangement between a leaderless, NARTCR-supporting TCRδV, encoding an addition Cys residue, and canonical TCRδD and TCRδJ segments. The rearranged NARTCR supporting V domain exon is then spliced to the TCRδ C region yielding a final receptor encoded by NARTCRV-NARTCRD-NARTCRJ-SupportingTCRδV-TCRδD-TCRδJ-TCRδC [36]. Non-placental mammals (marsupials and monotremes) have a TCR, TCRμ, with a quaternary structure similar to that proposed for NARTCR [37, 38]: TCRμ also has two V domains, but both V domains of TCRμ are Ig-like with the membrane-proximal V encoded by a germline-fused VDJ gene. The TCRμ C region is clearly most similar to TCRδ, but exists in a locus distinct from the TCRαδ locus. These Ig-TCRδ hybrid receptors corroborate earlier suggestions, based on CDR

length, that γδTCR binding is structurally more akin to that of Ig receptors than MHC restricted αβTCR [39].

In previous studies, we documented *trans*-rearrangements involving two IgMV and two IgWV rearrangements to TCRδDJ yielding Ig-TCRδ chimeric receptor rearrangements in the nurse shark. In the present study we aimed to further characterize the repertoire breadth and prevalence of Ig-TCRδ receptors, in addition to elucidating the genomic organization facilitating Ig-TCRδ rearrangements with a draft of the nurse shark TCRδ locus. To this end, we tested multiple hypotheses to conclude that Ig-TCRδ rearrangements were virtually indistinguishable from their canonical TCRδ counterparts in terms of CDR3 length and diversity and expression in both primary and secondary lymphoid tissue. Additionally some, but not all, Ig-TCRδ rearrangements originate from within the TCRδ locus in nurse sharks. This study outlines Ig-TCRδ rearrangements that substantially contribute to the shark's TCRδ repertoire diversity with prevalent expression levels suggesting their importance in an ancient adaptive immune system.

3.2 Materials and methods

3.2.1 Animals

Nurse sharks (*G. cirratum*) were captured off the coast of the Florida Keys and maintained in artificial seawater at approximately 28°C in large indoor tanks at the Institute of Marine and Environmental Technology, in Baltimore, MD, as per animal protocol (University of Maryland IACUC 1012003, Texas A&M IACUC 2013-0001). Animals were euthanized and bled out after an overdose of MS222. Tissues were

harvested and cells, DNA, RNA and frozen histology blocks were prepared as have been previously described [40-43].

3.2.2 PCR, cloning and sequencing of IgHV-TCRδDJC transcripts

Reverse TCRδC (primer FLAJ710, Supplemental Data 1) primed cDNA was made with SuperScript III Reverse Transcriptase (Life Technologies, Carlsbad CA) from 5µg total RNA from thymus, spleen and spiral valve (intestine) of a mature (10 year old) female nurse shark ("Joanie"). These cDNAs were used as a template for standard PCR amplification using GoTaq DNA polymerase (Promega, Madison WI) with forward primers for IgM (FLAJ1701) or IgW (FLAJ1699 and MFC185) variable genes and reverse for the TCRδ constant region gene (FLAJ767). After an initial two minute 95°C denaturation, samples were cycled 35 times through a 30 second 95°C denaturation, 30 second annealing, one minute 72°C, followed by a final seven minute extension at 72°C. Annealing temperatures used were 47°C for FLAJ1701, 51°C for FLAJ1699, and 58°C for MFC185. Bands were purified with the GeneCatcher (The Gel Company, San Francisco CA) isolation system and cloned into pCRII-TOPO (Life Technologies) or pBluescript II KS(+) (Stratagene, La Jolla), using One Shot TOP10 competent cells (Life Technologies). ZR Plasmid Minipreps (Zymo Research, Irvine CA) of clones of appropriate size were sequenced using BigDye Terminator v1.1 Cycle (Life Technologies) through the Texas A&M University DNA Technologies Core Laboratory. Questionable base calls were corrected based upon other aligned sequences if they clearly occurred within V, J or constant regions, but the sequence was excluded if ambiguous bases were

in N/P regions or the D segment of CDR3. Sequence data was managed with Bioedit (www.mbio.ncsu.edu/BioEdit/bioedit.html) or Geneious Version 9.1 (Biomatters, Auckland, New Zealand) and submitted to Genbank (accession numbers JF507709.1-JF507661.1). The CDR3 length was calculated using the "CDR3 length = exclusive number of amino acids from C (of V segment YxC motif) to F (of J segment FGxG motif)" IMGT formula [44]. For example, the first clone (T0006M2J09) in Figure 1A would have a CDR3 length of 21 amino acids.

3.2.4 Quantitative real-time PCR

Quantitative PCR was performed with 50 ng of random hexamer primed cDNA generated with SuperScript III from thymic and spleen RNA samples of sharks ranging in age from nine months to ten years. We used the SYBR Green PCR Master Mix (Biorad, Hercules, CA) following the manufacturer's recommendations. Triplicate wells were assayed in a Biorad CFX96 thermocycler, for 40 cycles annealing at 55°C. For the absolute quantification of each sample, a standard curve was created using serial dilutions of certain concentrations of a single copy gene cloned into a vector. Additional inter-exonic, real-time PCR primers can also be found in Supplemental Data 1. The resulting quantities, given in copy number per 50 ng cDNA, were split into three groups (canonical BCR, canonical TCR, and chimeric Ig-TCR) for statistical analyses. Significance was determined via the median based Kruskal-Wallis test with post-hoc Dunn test to determine specific differences between the groups.

3.2.3 BAC sequencing of the nurse shark TCRαδ locus

The *G. cirratum* BAC library was probed using cloned segments of TCRαV, TCRαC, TCRδV, TCRδC, and NARTCR probes from splenic transcripts amplified with ^{32}P as described under high stringency conditions [45, 46]. Selected BAC clones, positive for TCRα, TCRδ, and/or NARTCR components, were isolated from 500 ml cultures using the Qiagen Large Construct Kit (Qiagen, Hilden, Germany). Purified BAC DNA was sent to either the Duke Center for Genomic and Computational Biology for PacBio RSII large insert (15-20 kb) library sequencing. Assemblies were performed with the PBcR pipeline, Celera Assembler version 8.3, using both raw Pacbio reads, corrected in the PBcR pipeline, and Pacbio circular consensus (CCS) sequences with minimum a read length of 500 bp were used as input files for the assembly. [47, 48]. Options to merge haplotypes to a single consensus were used to account for the BAC library containing both parental chromosomes. Reads were deposited into the SRA under Bioproject SUB5368657. Annotation was performed using the Geneious software suite version 9.1. We employed a combination of BLAST, using a custom database of expressed nurse shark Ig and TCR sequences, and manual searches of assembled contigs for recombination signal sequences (RSSs) unveiled V, D and J segments not yet in public databases.

3.2.5 Generation of anti-IgHV polyclonal antisera

Polyclonal nurse shark IgMV1, IgWV1 and IgWV2 group [49] antibodies were generated in rabbits by Cocalico Biologicals (Reamstown, PA) by immunizing them with

IgHV-maltose binding protein (MBP) fusion proteins. The IgHV sequences were amplified from shark ("Joanie") spleen cDNA with 35 cycles of PCR annealing at 59°C (IgM) or 63°C (IgW). Restriction endonuclease site-engineered primers used were MFC180 and MFC181 for IgMVI, MFC182 and MFC183 for IgWVI and MFC184 and MFC183 for IgWVII (Supplementary Table 1). These products were cloned into the pMAL-c2x (New England Biolabs, Beverly MA) expression vector using SHuffle Express competent cells (New England Biolabs). Recombinant protein was produced in bacteria and cleared supernatant passed through amylose resin columns twice. Fusion proteins were eluted with maltose and precipitated with saturated ammonium sulfate, resuspended in PBS, and dialyzed in Slide-A-Lyzer cassettes (Pierce, Rockford IL). Sizes were verified by 12% SDS-PAGE using Coomassie and silver staining. The immune serum of recombinant protein-immunized rabbits was passed through an MBP affinity column to remove MBP-specific antibodies. Affinity purification of antibodies to shark IgHV in the cleared sera was performed with the same immunizing antigen immobilized in an agarose bead column (AminoLink, Pierce), and verified by SDS-PAGE and western blotting.

3.2.6 Flow cytometry

Thymocytes and splenocytes (5×10^5 cells/per treatment) were stained with either biotinylated mouse monoclonal antibody LK14 [50] (against *G. cirratum* IgL chain) or unlabeled anti-IgHV rabbit polyclonal (against *G. cirratum* IgMV1, IgWV1 or IgWV2) at 1:100 in staining buffer (1% BSA in shark phosphate-buffered saline) for 1 h at 4°C. Cells were then washed 3x with staining buffer before staining with streptavidin-

APC (eBioscience San Diego CA) at 1:1500 and anti-rabbit Alexafluor488 at 1:500 (Southern Biotech, Birmingham AL), respectively, for 30 min at 4°C. All samples were washed and resuspended in 300 µl of staining buffer containing 0.1% sodium azide and examined by flow cytometry on a BD LSR II instrument (BD Biosciences, San Jose CA). Fifty thousand events were collected, gated for live cells, and analyzed using the FlowJo software (Tree Star Inc., Ashland OR). Identical signal thresholds could be applied to all samples except the 120 month shark which had lower fluorescent intensity across all experiments.

3.2.7 Phylogenetic analyses

V segment alignments for both cartilaginous fish and vertebrate lineage trees were performed in Geneious using ClustalW. Amino acid alignments containing the entire V segment, from FR1 and to the conserved Cys of FR3, were used in the cartilaginous fish alignments. The multispecies alignments used nucleotide sequences of only the FR regions as CDR length and composition vary greatly across multiple vertebrate lineages [51, 52]. The resulting nucleotide alignment was manually adjusted to fit the beta strand IMGT protein display for V domains as previously described [53].

Maximum likelihood trees were constructed for all alignments in MEGAX using 1000 bootstrap replicates [54]. For the cartilaginous fish phylogeny, we used a Poisson correction model with bootstrap values displayed for bifurcations with greater than 50% consensus tree support. The multispecies tree used a general time reversal model, and the

substitution rate was gamma distributed with invariant sites with six discrete gamma categories.

3.2.8 Stellaris RNA-FISH

Frozen tissue sections were prepared from OCT embedded thymus tissue of a ten year old nurse shark as previously described [24]. Three tissue sections of the thymus, cut to a thickness of 10μm, were probed using the Stellaris (LGC Biosearch Technologies, Middlesex, UK) RNA-FISH system. Custom probe sets were designed using the Stellaris custom probe designer for the IgWV1 segment, TCRδ C region and TCRγ C. The fluorophores used for each probe set were CAL Fluor Red 610 (IgWV1), Quasar 670 (TCRδC), and CAL Fluor Orange 560 (TCRγC). Nuclei were stained with DAPI (Sigma-Aldrich, St. Louis, MO). The probed slides were imaged on a Zeiss Stallion Digital Imaging Workstation as previously described [4]. Images were processed using ImageJ software version 1.7 [55].

3.3 Results

3.3.1 Repertoire sequencing reveals additional IgHV usage

Transcripts in which an IgHV segment is rearranged to TCRδ gene segments have been previously documented in the nurse shark [24]. To determine the breadth of this chimeric receptor repertoire, 5’ rapid amplification of cDNA ends (RACE) PCR was performed using a TCRδ C region primer. Nested forward primers were designed to selectively amplify the Ig-TCRδ hybrid receptors. The sequenced amplicons totaled 137

unique IgV-TCRδDJ rearrangements (chimeric, or *trans*-rearranged, clones are abbreviated Ig-TCRδ) from spiral valve (n=38), spleen (n=51), and thymus (n=50) transcripts, displaying a diverse repertoire of Ig-TCRδ in both primary and secondary lymphoid tissues (Figure 3.1A-C. Most of the recovered transcripts (95.6%) were in frame, indicative of selection mechanisms prohibiting non-functional rearrangements or mutations. We identified nine IgM/W V segments rearranged with TCRδ; all but one segment, S0021TVJ09, could be found in BCR rearrangements bearing an IgM, or IgW, C region. This sequence encoded an IgH-like V and was most similar (68% nucleotide identity) to the nurse shark IgH Group 7 pseudogene cluster (EU312153). This genomic V segment was not found expressed in any rearrangement and was not determined to be the source of this peculiar V used in transcript S0021TVJ09. Accession numbers for IgH V segments found in both IgH and TCRδ rearrangements were AY609260 (IgMV1), DQ857389 (IgMV2.1), AY609247.1 (IgMV2.2), AY609259.1 (IgMV4), AY609249.1 (IgMV5), KF192877.1 (IgWV1), KF192883.1 (IgWV2), and KF192884.1 (IgWV3).

Table 3.1. CDR3 lengths of Ig-TCRδ chimeric receptor chains.

	Total Ig-TCRδ	Thymus Ig-TCRδ	Spleen Ig-TCRδ	Spiral valve Ig-TCRδ	Canonical TCRδ[1]
Median	17	17	18	15	18
Mean	17.4	18.0	17.8	16.3	17.4
Variance	22.8	26.4	23.6	19.6	25.0
Maximum	34	34	28	29	30
Minimum	7	8	7	11	7
Range	27	26	21	18	23

[1]from ref (Criscitiello *et al. Journal of Immunology* 2010)

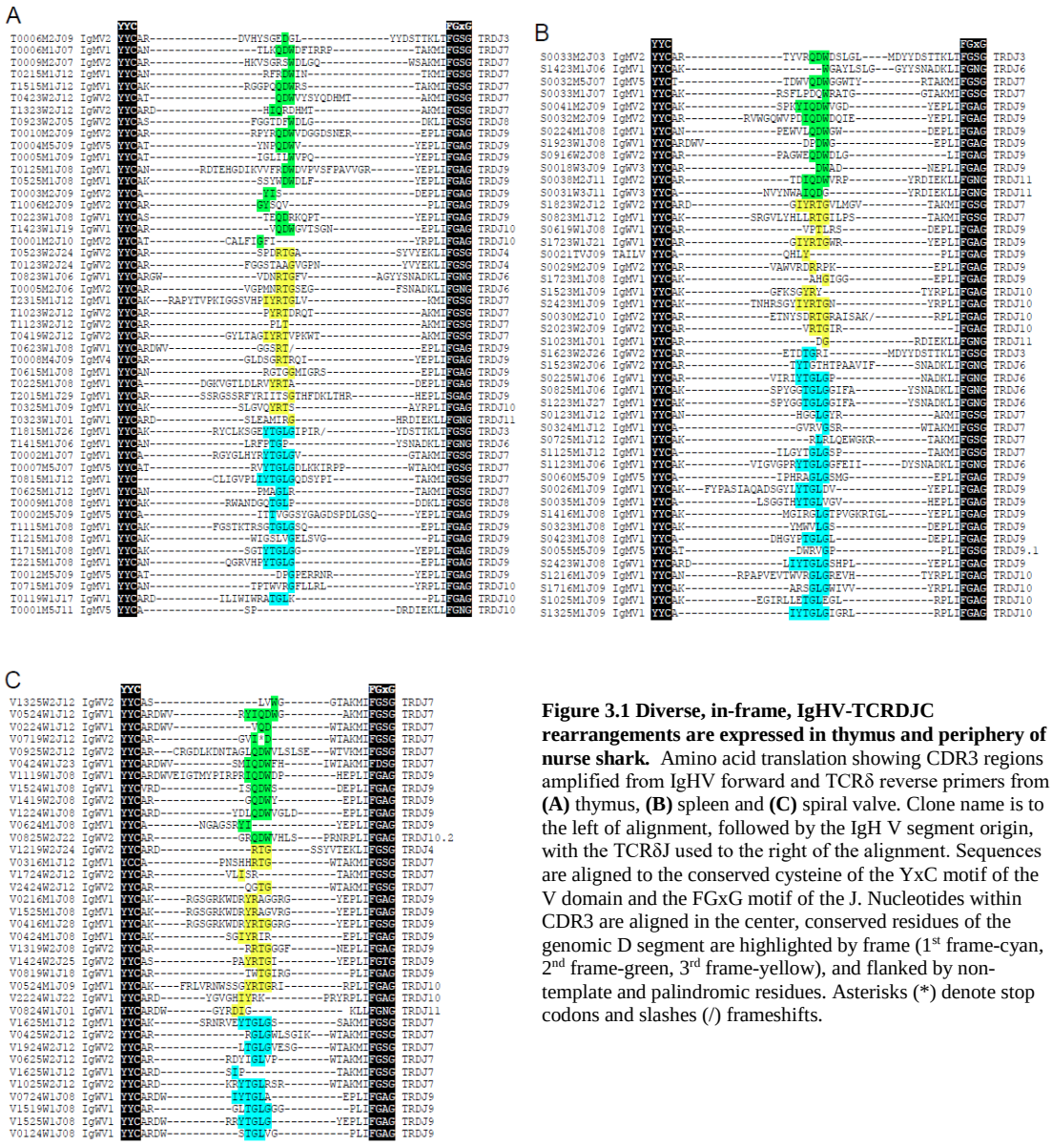


Figure 3.1 Diverse, in-frame, IgHV-TCRDJC rearrangements are expressed in thymus and periphery of nurse shark. Amino acid translation showing CDR3 regions amplified from IgHV forward and TCRδ reverse primers from **(A)** thymus, **(B)** spleen and **(C)** spiral valve. Clone name is to the left of alignment, followed by the IgH V segment origin, with the TCRδJ used to the right of the alignment. Sequences are aligned to the conserved cysteine of the YxC motif of the V domain and the FGxG motif of the J. Nucleotides within CDR3 are aligned in the center, conserved residues of the genomic D segment are highlighted by frame (1st frame-cyan, 2nd frame-green, 3rd frame-yellow), and flanked by non-template and palindromic residues. Asterisks (*) denote stop codons and slashes (/) frameshifts.

3.3.2 Nurse shark Ig-TCRδ and canonical TCRδ repertoires possess equivalent CDR3 metrics

We compared the Ig-TCRδ CDR3 transcript metrics to that of the canonical TCRδ repertoire to ascertain whether chimeric Ig-TCRδ rearrangements bore similar CDR3

metrics to their canonical counterparts (Figure 3.2, Table 3.1). Two aberrant transcripts with extraordinarily long CDR3 lengths (S2223W1J08 and V0924W1J20 with CDR3 lengths of 41 and 49 amino acids), were removed from statistical analyses. The distribution of CDR3 length for functional TCRδ (N=55) and Ig-TCRδ (N=135) were nearly identical and clearly distinguishable from TCRβ (N=41) and IgH (N=173) (Figure 2A). Mean CDR3 length of Ig-TCRδ rearrangements was 17.44 ± 0.42 amino acids, statistically indistinguishable from that of canonical TCRδ (17.40 ± 0.67 aa), with each TCRδ subset encoding repertoire CDR3 content significantly longer than that of both TCRβ (11.68 ± 1.01 aa) and IgH (12.55 ± 0.25) determined via ANOVA (Figure 2B, Table 1). The ranges of CDR3 lengths of receptors free from MHC restriction (TCRγδ and Ig, and especially IgHV and TCRδ) are typically much longer than that of TCRαβ [55]. Indeed the range of the Ig-TCRδ subset was comparable to that of canonical TCRδ receptors (Figure 3.2C, Table 3.1), with both TCRδ subsets encompassing a larger range than either TCRβ or IgH (Figure 3.2C). Finally we analyzed the isoelectric point (iP) of CDR3 residues and found TCRδ (iP=8.35) and Ig-TCRδ (iP=8.38) both averaged slightly higher isoelectric points than either IgH (iP=6.86) or TCRβ (iP=5.98) (Figure 3.2D). These data show that chimeric Ig-TCRδ and canonical TCRδ receptors are comparable not only in CDR3 length, but possess similar charge signatures integral to epitope-paratope interactions as well. These results were unsurprising as both Ig-TCRδ and canonical TCRδ receptors draw from the same TCR D and J segment pool, contributing heavily to the CDR3 similarities outlined above (Figure 3.1). Additionally we elucidated that the IgWV1 gene segment, found expressed in Ig-TCRδ rearrangements, was found in

individual cells also expressing TCRγ using RNA-FISH. The highly similar CDR3 metrics between Ig-TCRδ and canonical TCRδ is indicative of IgHV segment, with divergent CDR1 and CDR2 loops, use provides substantial expansion of potential TCRδ paratopes.

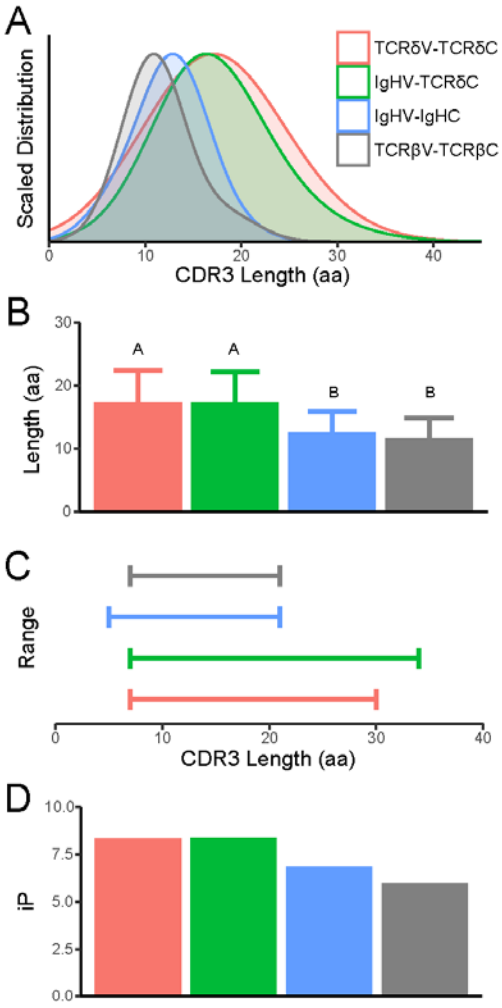


Figure 3.2. Chimeric Ig-TCRδ and canonical TCRδ encode comparable CDR3 metrics. Analysis of CDR3 for rearranged transcripts of TCRδV- TCRδC (red), IgHV- TCRδC (green), IgHV-IgHC (blue) and TCRβV- TCRβC (grey) showcases the near identical distribution **(A)**, average length **(B)**, range **(C)**, and isoelectric point (iP) **(D)** of chimeric Ig-TCRδ and canonical TCRδ. Significance of CDR3 length indicated by lettering **(B)** was determined via ANOVA with post-hoc Tukey HSD test with each TCRδ subset being significantly longer than IgH and TCRβ.

3.3.3 Absolute quantification of TCR delta by real-time PCR shows equivalent levels for canonical and chimeric transcripts

Although the number and diversity of chimeric clones suggested that nurse shark Ig-TCRδ rearrangements were of significance for the shark immune system, we performed absolute quantification PCR (qPCR) to determine their prevalence in comparison to other TCRδ rearranged transcripts. Specific primers for each V segment (TCRδ, IgM, and IgW) were used with either a TCRδ or IgH C region primer (Supplemental Data 1). Interestingly, no significant differences were found between expression levels of canonical TCRδ and chimeric Ig-TCRδ transcripts in the thymus (Figure 3A). While expression of BCR was decidedly lower than either TCR group in the thymus, we found no significant differences found between the three groups in the thymus of young sharks aged 10 and 16 months (Figure 3.3A). Expression of transmembrane IgH message in the thymus of developing sharks has been documented, but consistent with our new quantitative data, such expression diminishes to background levels as the animals mature [56, 57]. Expression of BCR in the spleen compared to thymus was markedly higher than both TCRδ groups, yet there remained no significant difference between canonical TCRδ and chimeric Ig-TCRδ expression in the periphery (Figure 3.3B). Furthermore, flow cytometry performed on thymocytes, using anti-IgHV polyclonal and the LK14 monoclonal antibody specific for kappa L chains, provided preliminary evidence that chimeric receptors are present as a surface receptor. Although IgW has been shown to prefer an IgL chain other than kappa [58], qPCR confirmed expression of either IgWV segment was negligible in the thymus of these sharks (Figure

3.3A). Importantly, these data show Ig-TCRδ and canonical TCRδ mRNA expression to be equivalent in both primary and secondary lymphoid tissues, and suggest that they are used as a surface antigen receptor.

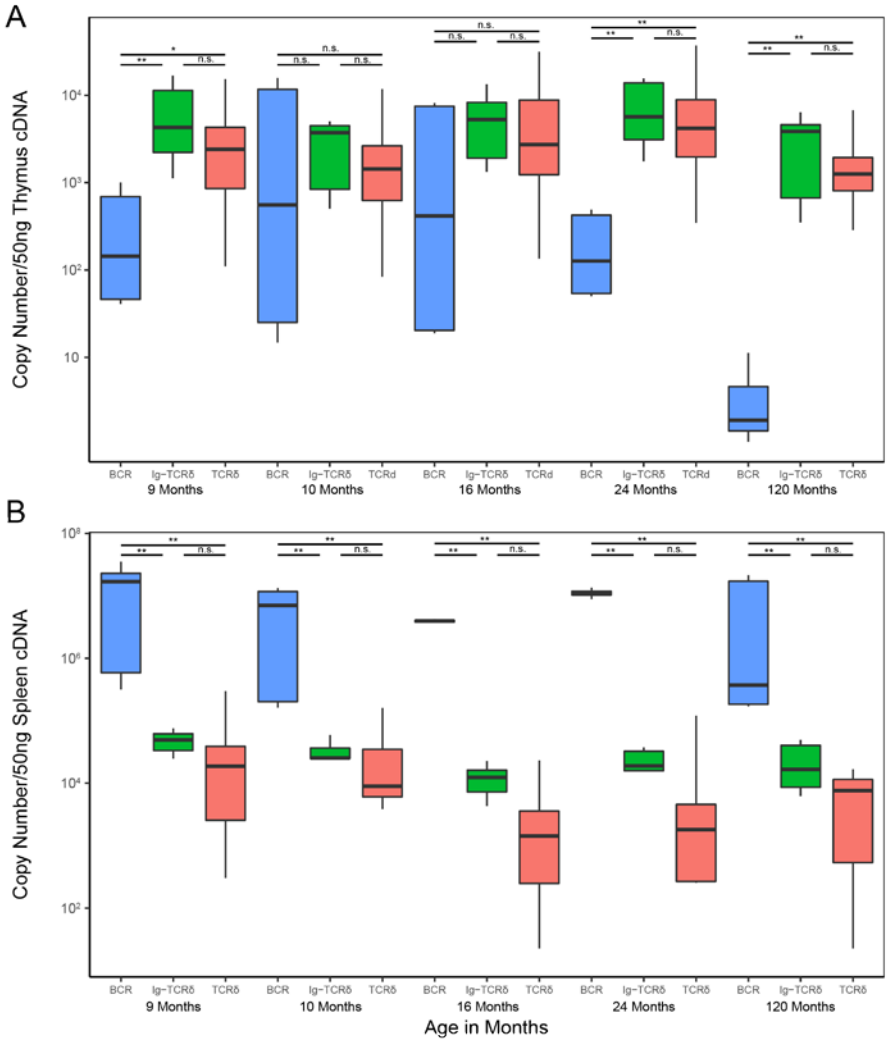


Figure 3.3. Absolute quantification PCR reveals chimeric Ig-TCRδ expression to be comparable to canonical TCRδ in both primary (thymus) and secondary (spleen) lymphoid tissue. Absolute transcript values (displayed on the y-axis, calculated using known concentrations of an amplicon cloned into plasmid) are compared within each shark (displayed on the x-axis indicated by age in months) for **(A)** thymus and **(B)** spleen samples. To compare, rearrangements were classified as canonical B cell transcripts (BCR), chimeric (Ig-TCR), and canonical TCRδ transcripts (TCRδ). The boxplot extends from the first to the third quartiles (25^{th} and 75^{th} percentiles respectively) with centerline indicating median value. The whisker of each boxplot extends to the value no greater than 1.5 times the interquartile range (difference between the first and third quartiles), values beyond these whiskers are potential outliers indicating a significant skew, or preference for a given rearrangement. Due to the relatively high proportion of outliers present the nonparametric Kruskall-Wallace test with post-hoc Dunn test to determine significant differences in groups (BCR, Chimeric, TCR) within a given shark was used. Significance indicated by asterisks as follows ** ($p<0.01$), * ($p<0.05$), n.s. ($p>0.05$).

3.3.4 Draft assembly of the G. cirratum TCRδ locus reveals a novel V segment lineage

Since IgH V segments have been found to be associated with TCRδ loci in a growing number of vertebrate lineages, we also performed a draft assembly of the nurse shark TCRδ locus with targeted BAC sequencing. To this end the *G. cirratum* BAC library was probed and selected BAC clones, positive for TCRδ, TCRα, and NARTCR components, were selected for long read sequencing. The resulting assembly yielded a draft TCRδ locus assembly totaling 157kb with an average coverage depth of 61, the gene segments identified are displayed in Figure 3.4.

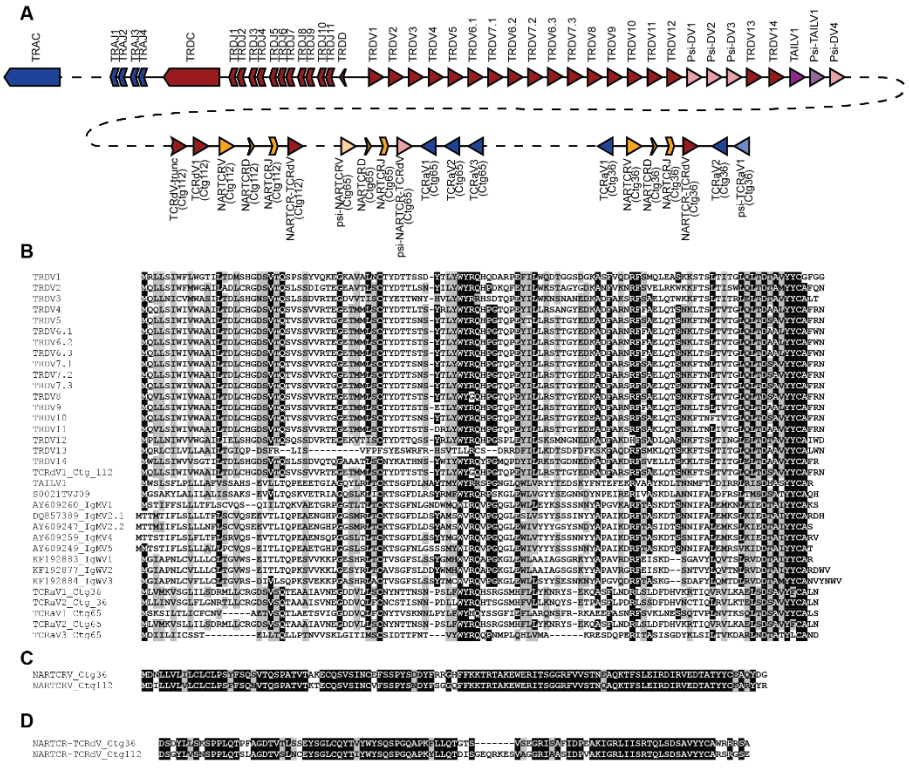


Figure 3.4. Assembly and mapping of TCRδ loci reveals IgH-like V segment bearing similarity to those found rearranged with TCRδ, but not IgH. (A) Segments located within the assembled nurse shark TCRδ locus are color coded according to lineage differentiating TCRδ (red), TCRα (blue), and IgH/TAILV segments (green). Pseudogene segments are indicated by empty boxes. **(B)** Alignment of V segments identified in the locus, including the TAILV and IgHVs found in both TCRδ and IgH rearrangements.

Although the TCRα C region was not assembled in within the contig housing the TCRδ translocus (Figure 4, Contig 10), the presence of TCRα J segments downstream of the TCRδ translocus indicate a linked TCRαδ locus in this ancient model of IGSF receptor loci. This was corroborated by the BAC library screening which identified multiple BACs probing positive for TCRδ, TCRα and NARTCR gene segments (data not shown), and the linkage of TCRα and NARTCR in contigs that could not be linked to the grander TCRδ translocus (Figure 4, Contigs 36, 65, and 112). While the targeted BAC sequencing approach allows confident assembly of IGSF loci, the technique is hindered in instances where probed BACs contain no overlapping regions. Furthermore without a nurse shark karyotype to probe chromosomes selection of positive BACs with overlap is a random that does not account for vast genomic regions without TCR gene segments.

The draft assembly of the nurse shark included four contigs with a total of 29 functional V segments, 4 D segments, 18 J segments, and 1 TCRδ C region. The largest contig (Contig 10) revealed a translocon including 24 V segments (19 functional and 5 psuedo-V segments), a single TCRδ D segment, 11 TCRδ J segments, 4 TCRα J segments, and the TCRδ C region (Figure 4A). We also assembled fragments with segments shown to be involved in formulating the TCRδ repertoire in three additional, non-overlapping BACs. These BACs included three NARTCR VDJ blocks, with two functional NARTCRVs (Contigs 36 and 112) and one pseudo-NARTCRV (Contig 65). Each NARTCR block was found upstream of a supporting TCRδ V segment (functional on Contigs 36 and 112, pseudogenized on Contig 65). Two additional TCRδV segments were

identified upstream of the NARTCR block in Contig 112. The first TCRδV on Contig 112 was truncated by the assembly, the 164 bp assembled contained no frameshifts or stop codons, however the second is TCRδV is functional. Presumptively these indicate an additional TCRδV translocon stretch unidentified thusfar. Finally we identified 6 TCRα V segments all oriented in reverse orientation in relation to the NARTCR blocks. Three functional TCRαVs were located downstream of the NARTCR block on Contig 65. The remaining three were located on Contig 36, one functional TCRαV was found upstream of the NARTCR block while one functional TCRαV and a single pseudo-TCRαV were located downstream.

Additionally we identified one functional IgHV-like segment and one pseudo-IgHV-like segment nested in the translocon stretch of TCRδVs (Figure 4A). While the IgHV-like segment is putatively functional it could not be confidently ascribed to any Ig-TCRδ rearrangement in our expression dataset, but was similar to the novel IgHV-like segment we found only in TCRδ rearrangements (Figure 3.4B). In phylogenetic analyses of cartilaginous fish IgH and TCRδ V segments these Ig-like V segments, not found in rearrangements bearing IgH C regions, interestingly grouped together in a clade nestled between the IgM and IgW segments (Figure 3.5A). We dubbed the segments belonging to this clade TCRδ associated Ig-like V segments (TAILVs) to distinguish them from the V segments found on both TCRδ and IgH receptors. We also performed a phylogenetic analysis that included TCRα, TCRδ, VHδ, IgH, and IgL genomic V segments from all vertebrate lineages. The nurse TAILV representative sequence was found in a branch that included nurse and whale shark IgMVs, recapitulating the findings in our cartilaginous

fish tree (Figure 3.5B-C). This multispecies phylogeny also unveiled a clade of V segments, previously labelled TCRδV, sharing a common ancestry with Ig V segments rather than TCRV. The lack of a clade solely composed of Ig V segments found in TCRδ loci indicates that they have arisen multiple times through the course of vertebrate evolution.

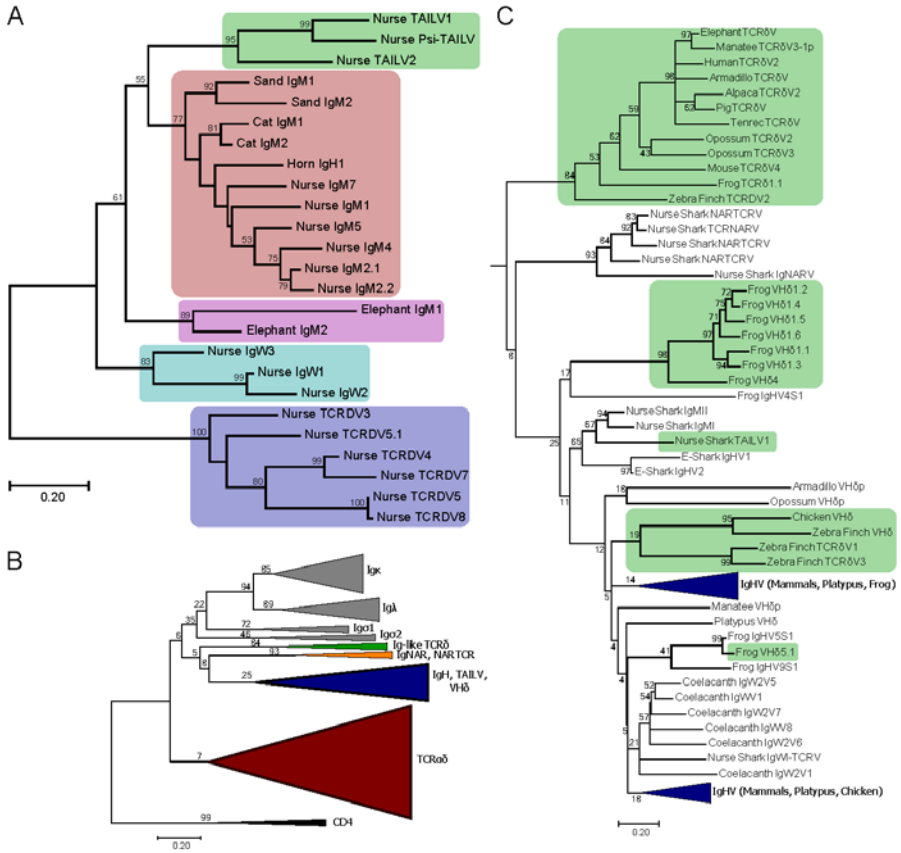


Figure 3.5. Phylogenetic analysis of antigen receptor V segments unveils a novel lineage of TCRδ associated IgH-like V segments (TAILV) in cartilaginous fish. (A) An amino acid alignment of chondrichthyes V segments, trimmed to the start of FR1 and ended at the conserved Cys of FR3, was used to construct a maximum-likelihood tree. Species included in the alignment were *G. cirratum* (Nurse), *S. canicula* (Cat), *H. francisci* (Horn), *C. plumbeus* (Sand), and *C. milii* (elephant). The percentage of trees supporting branches after 1000 Bootstrap replications is displayed for bifurcations with greater than 50%. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The shaded regions indicate different V segment lineages found in chondrichthyes including elasmobranch TAILVs (green), elasmobranch IgMV (red), holocephali IgMV (pink), elasmobranch IgW (teal), and elasmobranch TCRδV (blue). **(B)** Topology of the jawed vertebrate antigen receptor tree. V segment lineage indicated by color for IgL (grey), Ig-like TCRδV (green), IgNARV/NARTCRV (orange), IgH (blue, includes TAILV and VHδ), and TCRαδV (red). The tree includes representative sequences from human, mouse, pig, tenrec, elephant, alpaca, manatee, opossum, platypus, zebra finch, chicken, frog, coelacanth, nurse shark, and elephant shark. **(C)** Subtree of the IgH and Ig-like V segments. Green shading indicates lineages of Ig-like V segments that are exclusively found on TCR rearrangements highlighting multiple exchanges of Ig V segments into TCR loci throughout vertebrate evolution.

3.3.5 Comparative analysis of cartilaginous fish TCRδ loci

A number of cartilaginous fish including the whale shark, *Rhincodon typus* [59], bamboo shark, *Chiloscyllium plagiosum,* and cloudy cat shark, *Scyliorhinus torazame* [60] have joined the first cartilaginous fish genome of the elephant shark, *Callorhinchus milii* [61]. Of the recently sequenced genomes only the bamboo shark has a scaffold (BEZZ01002038.1) in which multiple TCRδVs could be located, however no TCRδ J segments or C region was found. This left only the elephant shark assembly with a TCRδ on the 694kb scaffold NW_006890273 that is flanked by two IgMV segments found in typical cluster arrangement; one cluster possessing a full complement of IgM CH exons, the other with only a single CH exon. Also present in the elephant shark genome are NARTCRV and supporting TCRδV blocks, similar to those found on smaller contigs in the nurse shark [62]. The absence of a TCRδ J segment and C region in either assembly made it impossible to confirm whether the TCRδC proximal V translocon was inverted in other cartilaginous fish species as well. Nonetheless, the elephant shark scaffold was still useful to formulate a predictive placement for drafted nurse shark contigs and potential IgH clusters used in Ig-TCRδ rearrangements (Figure 6). Furthermore the proximity of the IgM clusters to the TCRδ locus in the elephant shark allowed us to formulate a potential model for the locale of, at least some, IgH clusters used in Ig-TCRδ rearrangements (Figure 6B). Finally we formulated a model of all nurse shark Ig-TCRδ rearrangements elucidated thus far, one intra-locus TAILV segments, and *bona-fide* "trans"-locus rearrangements between IgWV, from three clusters, and IgMV, from five clusters, which contribute to the TCRδ repertoire (Figure 6B).

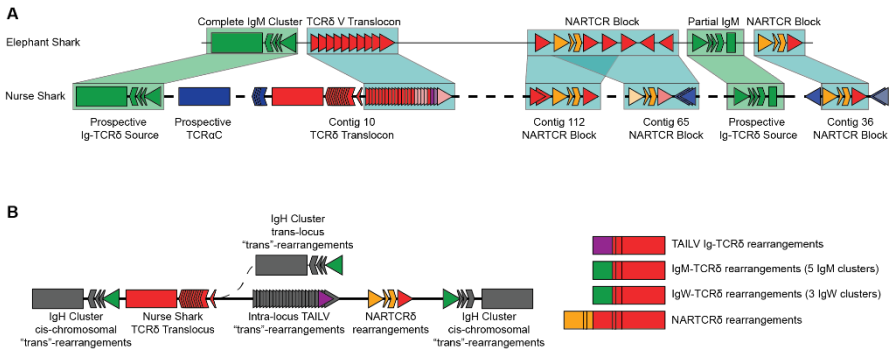


Figure 3.6. Comparative structural analysis of the TCRδ locus shows syntenic blocks between holocephali and elasmobranchii. **A.** Sequence comparison of the TCRδ locus structure using elephant shark scaffold NW_006890273.1 as a map for assembled nurse shark contigs. Light blue shading indicates syntenic nurse shark contigs from the draft assembly with associated contigs labelled beneath the box. Green shading indicates prospective positions of nurse shark IgH V-segments used in Ig-TCR rearrangements based on position within the elephant shark scaffold. **B.** Sources of Ig-TCRδ rearrangements in the nurse shark.

3.4 Discussion

From what is known of cartilaginous fish IgH isotypes and the antigen recognition modes of vertebrate γδ T cells, there is a clear logic in IgMV and IgWV rearranging to TCRδ. Both share the same intrinsic structure of the IgSF V domain and the capacity to interact with free antigen. Drawing from the same pool of D and J segments as canonical TCRδ rearrangements, the Ig-TCRδ repertoire was found to be just as diverse as the their canonical counterparts, characterized by large non-template TdT additions at the V-D and D-J junctions (Figure 3.1). The large non-template additions found in Ig-TCRδ rearrangements were constrained to encode CDR3 lengths consistent with that of canonical TCRδ rearrangements in the nurse shark (Figure 3.2, Table 3.1). We were surprised to find that the predicted CDR3 iP in the two TCRδ rearrangement classes were also roughly equivalent despite the bulk of CDR3 residues stemming from non-template nucleotides (Figure 3.2). Similarities observed in CDR3s of Ig-TCRδ and canonical TCRδ suggest the two groups make use of the same dimeric partner chain. The heterodimeric

TCR complex, complete with signal transduction machinery, is required of all T cells during thymic selection [63-65]. We have determined that Ig-TCRδ rearrangements also pair with TCRγ, this is likely due to conservation of the Cys residue in TCRδ C regions facilitating the interchain disulfide bond linking γδTCR receptor complexes [67]. Any Ig-TCRδ chain rearrangements bearing incompatibilities with a partner chain, which prohibited paratope formation, would be selectively eliminated from the repertoire. Differences in the IgV domain are seemingly well-tolerated by the partner chain as the mRNA expression levels of Ig-TCRδ and canonical TCRδ were indistinguishable (Figure 3.3). Thymic expression of aberrant receptors could be attributed to remnants of non-functional rearrangements (Figure 3.3A). Equivalent expression levels for Ig-TCRδ and canonical TCRδ in the spleen of all five sharks (Figure 3.3B) indicate that the Ig-TCRδ rearrangements regularly pass selection. Furthermore, presence of Ig-TCRδ receptor rearrangements for at least three IgHV segments were identified on sorted thymocytes even in sharks which lacked support for BCR expression in the thymus.

Altogether, Ig-TCRδ rearrangements in the nurse shark are neither aberrant nor rare. With CDR3 metrics and expression levels so closely aligned to canonical TCRδ, it appears that Ig-TCRδ rearrangements, via their Ig-derived CDR1 and CDR2, are well tolerated by the partner chain and would diversify the recognition potential of shark TCRδ receptors. Such diversity would likely be disallowed in MHC-restricted αβ T cells, but is tolerated in γδ T cells unbound from these restrictions.

Recombination signal sequences are the regulatory elements governing rearrangements in shark as in mammals [66-69]. If an IgHV were in proximity to the TCRδ

locus one could expect Ig-TCRδ rearrangements to occur between IgH and TCRδ loci. Indeed, this was the case as several V segments identified in our Ig-TCRδ dataset originate from canonical IgH clusters. We found bona fide *trans*-rearrangements between IgH and TCRδ loci are regular occurrences in the nurse shark. However, these were not the only type of chimeric Ig-TCRδ rearrangements we documented as several V segments were not found in any IgH cluster, nor expressed in the nurse shark antibody repertoire. Our draft nurse shark TCRδ locus assembly identified a single IgHV-like segment integrated in a translocon stretch of TCRδVs, with no downstream Ig C region, that likewise was not found expressed as an IgH (Figure 4). Phylogenetic analysis of shark TCRδ and IgH V segments determined that these peculiar segments both clustered in a distinct clade in the phylogenetic analysis (Figure 3.5). As both are clearly more Ig-like than TCR, we decided to call these V segments TCRδ associated Ig-like (TAILVs). Presence of TAILVs in the TCRδ locus in one of the earliest iterations of IGSF-based adaptive immunity adds to the narrative that Ig and TCRδ V segments are largely interchangeable. This is also consistent with the emerging paradigm of IgHV segments associated with TCRδ loci of a growing number of species. Indeed, we identified a clade of canonically labelled TCRδVs, located in the TCRδ locus of tetrapod, avian and mammalian lineages, with a shared Ig ancestry (Figure 3.5B-C) This evidence highlights implementation of Ig-like V segments on TCRδ receptors in all jawed vertebrate lineages from cartilaginous fish to eutherian mammals. Phylogenetic evidence did not support TAILV as the progenitor of VHδ, nor the aforementioned Ig-like TCRδV clade (Figure 3.5C), but illustrates how IgH and TCRδ have freely exchanged gene segments over the course of vertebrate evolution. In fact, the

location of the holocephalan (the most ancient chondrichthyes lineage) elephant shark IgMVs, which flank a TCRδ locus with no evidence of TAILVs, in the cartilaginous fish phylogeny suggest these proximal IgH clusters could be the ancient precursors to TAILVs (Figure 6). Confirmation of this hypothesis via identification of syntenic blocks will require a more complete elephant shark assembly and completion of the nurse shark TCRδ locus. Although we cannot claim that these IgMVs gave rise to the nurse shark TAILVs we outlined hypothetical origins of TAILV as well as potential sources of IgHV segments found in both IgH clusters and Ig-TCRδ rearrangements in Figure 3.6. This hypothetical synteny also provided evidence that the NARTCR blocks are likely localized to the TCRδ locus in all chondrichthyan species with confirmed NARTCR expression.

The TCRα and TCRγ loci of sharks have been shown to employ somatic hypermutation (SHM) [4, 70, 71], a process normally associated with IgH/L loci. Although IgHV sequences in our dataset contained minor differences within identical CDR3 regions, we found no conclusive evidence, via mutations in the IgHV of transcripts with identical CDR3 content, suggesting that SHM acts to expand diversity within the IgHV segments used in chimeric rearrangements. This was surprising, as one would assume the Ig V segments would be primed for diversifying mutations; suggesting that the anatomic and developmental windows of AID expression are tightly regulated in the shark as well as endothermic vertebrates.

In summary, while the evolutionary history of Ig-TCRδ remains complex in vertebrates, we have confirmed that Ig-TCRδ rearrangements contribute to the peripheral T cell repertoire. The cellular contribution of Ig-TCRδ within the nurse shark immune

system was highlighted in sequence diversity and expression in an assortment of tissues. Indeed, at all ages included in this study Ig-TCRδ receptors are as prevalent as canonical TCRδ rearrangements. Additionally the Ig-TCRδ repertoire makes use of Ig V segments both from IgH clusters and from a separate lineage of IgHV-like segments, termed TAILVs due to their exclusive association with TCRδ. Comparative TCRδ locus analysis of cartilaginous fish, both holocephali and elasmobranchii that split ~420 MYA [72], exemplifies the versatility of TCRδ; a TCR employing receptors bearing the V domain of canonical TCRδ, IgHV (through chimeric Ig-TCRδ rearrangements originating both within the TCRδ locus (TAILV) and from IgH clusters of unknown proximity), and the doubly-rearranging NARTCR (Figure 3.6B).

3.6 References

1. Litman, G.W., K. Hinds, L. Berger, K. Murphy, and R. Litman, *Structure and organization of immunoglobulin VH genes in Heterodontus, a phylogenetically primitive shark.* Dev.Comp Immunol., 1985. **9**(4): p. 749-758.
2. Rast, J.P., M.K. Anderson, S.J. Strong, C. Luer, R.T. Litman, and G.W. Litman, *alpha, beta, gamma, and delta T cell antigen receptor genes arose early in vertebrate phylogeny.* Immunity., 1997. **6**(1): p. 1-11.
3. Criscitiello, M.F. and M.F. Flajnik, *Four primordial immunoglobulin light chain isotypes, including lambda and kappa, identified in the most primitive living jawed vertebrates.* Eur.J.Immunol., 2007. **37**(10): p. 2683-2694.

4. Ott, J.A., C.D. Castro, T.C. Deiss, Y. Ohta, M.F. Flajnik, and M.F. Criscitiello, *Somatic hypermutation of T cell receptor alpha chain contributes to selection in nurse shark thymus.* Elife, 2018. **7**.

5. Bernstein, R.M., S.F. Schluter, H. Bernstein, and J.J. Marchalonis, *Primordial emergence of the recombination activating gene 1 (RAG1): sequence of the complete shark gene indicates homology to microbial integrases.* Proc.Natl.Acad.Sci.U.S.A, 1996. **93**(18): p. 9454-9459.

6. Rumfelt, L.L., D. Avila, M. Diaz, S. Bartl, E.C. McKinney, and M.F. Flajnik, *A shark antibody heavy chain encoded by a nonsomatically rearranged VDJ is preferentially expressed in early development and is convergent with mammalian IgG.* Proc.Natl.Acad.Sci.U.S.A, 2001. **98**(4): p. 1775-1780.

7. Diaz, M., A.S. Greenberg, and M.F. Flajnik, *Somatic hypermutation of the new antigen receptor gene (NAR) in the nurse shark does not generate the repertoire: possible role in antigen-driven reactions in the absence of germinal centers.* Proc.Natl.Acad.Sci.U.S.A, 1998. **95**(24): p. 14343-14348.

8. Foster, D.M., J.L. Gookin, M.F. Poore, M.E. Stebbins, and M.G. Levy, *Outcome of cats with diarrhea and Tritrichomonas foetus infection.* J Am Vet Med Assoc, 2004. **225**(6): p. 888-92.

9. Foster, B.J., J. Shults, B.S. Zemel, and M.B. Leonard, *Interactions between growth and body composition in children treated with high-dose chronic glucocorticoids.* Am J Clin Nutr, 2004. **80**(5): p. 1334-41.

10. Zhu, C. and E. Hsu, *Error-prone DNA repair activity during somatic hypermutation in shark B lymphocytes.* J Immunol, 2010. **185**(9): p. 5336-47.

11. Diaz, M., J. Velez, M. Singh, J. Cerny, and M.F. Flajnik, *Mutational pattern of the nurse shark antigen receptor gene (NAR) is similar to that of mammalian Ig genes and to spontaneous mutations in evolution: the translesion synthesis model of somatic hypermutation.* Int.Immunol., 1999. **11**(5): p. 825-833.

12. Lee, V., J.L. Huang, M.F. Lui, K. Malecek, Y. Ohta, A. Mooers, and E. Hsu, *The evolution of multiple isotypic IgM heavy chain genes in the shark.* J Immunol, 2008. **180**(11): p. 7461-70.

13. Litman, G.W., M.K. Anderson, and J.P. Rast, *Evolution of antigen binding receptors.* Annu Rev Immunol, 1999. **17**: p. 109-47.

14. Berstein, R.M., S.F. Schluter, S. Shen, and J.J. Marchalonis, *A new high molecular weight immunoglobulin class from the carcharhine shark: implications for the properties of the primordial immunoglobulin.* Proc Natl Acad Sci U S A, 1996. **93**(8): p. 3289-93.

15. Ohta, Y. and M. Flajnik, *IgD, like IgM, is a primordial immunoglobulin class perpetuated in most jawed vertebrates.* Proc.Natl.Acad.Sci.U.S.A, 2006. **103**(28): p. 10723-10728.

16. Hinds, K.R. and G.W. Litman, *Major reorganization of immunoglobulin VH segmental elements during vertebrate evolution.* Nature, 1986. **320**(6062): p. 546-9.

17. Zhu, C., V. Lee, A. Finn, K. Senger, A.A. Zarrin, L. Du Pasquier, and E. Hsu, *Origin of Immunoglobulin Isotype Switching.* Curr Biol, 2012.

18. Clem, I.W., B.F. De, and M.M. Sigel, *Phylogeny of immunoglobulin structure and function. II. Immunoglobulins of the nurse shark.* J.Immunol., 1967. **99**(6): p. 1226-1235.

19. Kasahara, M., M. Vazquez, K. Sato, E.C. McKinney, and M.F. Flajnik, *Evolution of the major histocompatibility complex: isolation of class II A cDNA clones from the cartilaginous fish.* Proc.Natl.Acad.Sci.U.S.A, 1992. **89**(15): p. 6688-6692.

20. Bartl, S. and I.L. Weissman, *Isolation and characterization of major histocompatibility complex class IIB genes from the nurse shark.* Proc.Natl.Acad.Sci.U.S.A, 1994. **91**(1): p. 262-266.

21. Criscitiello, M.F., M. Saltis, and M.F. Flajnik, *An evolutionarily mobile antigen receptor variable region gene: doubly rearranging NAR-TcR genes in sharks.* Proc.Natl.Acad.Sci.U.S.A, 2006. **103**(13): p. 5036-5041.

22. Greenberg, A.S., D. Avila, M. Hughes, A. Hughes, E.C. McKinney, and M.F. Flajnik, *A new antigen receptor gene family that undergoes rearrangement and extensive somatic diversification in sharks.* Nature, 1995. **374**(6518): p. 168-73.

23. Rumfelt, L.L., M. Diaz, R.L. Lohr, E. Mochon, and M.F. Flajnik, *Unprecedented multiplicity of Ig transmembrane and secretory mRNA forms in the cartilaginous fish.* J.Immunol., 2004. **173**(2): p. 1129-1139.

24. Criscitiello, M.F., Y. Ohta, M. Saltis, E.C. McKinney, and M.F. Flajnik, *Evolutionarily conserved TCR binding sites, identification of T cells in primary lymphoid tissues, and surprising trans-rearrangements in nurse shark.* J Immunol, 2010. **184**(12): p. 6950-60.

25. Foster, D.B., R. Huang, V. Hatch, R. Craig, P. Graceffa, W. Lehman, and C.L. Wang, *Modes of caldesmon binding to actin: sites of caldesmon contact and modulation of interactions by phosphorylation.* J Biol Chem, 2004. **279**(51): p. 53387-94.

26. Foster, P.S. and D.W. Harrison, *Cerebral correlates of varying ages of emotional memories.* Cogn Behav Neurol, 2004. **17**(2): p. 85-92.

27. Foster, P.H., *Health disparities affecting African American.* J Health Care Poor Underserved, 2004. **15**(3): p. vii-viii.

28. Denny, C.T., Y. Yoshikai, T.W. Mak, S.D. Smith, G.F. Hollis, and I.R. Kirsch, *A chromosome 14 inversion in a T-cell lymphoma is caused by site-specific recombination between immunoglobulin and T-cell receptor loci.* Nature, 1986. **320**(6062): p. 549-51.

29. Kobayashi, Y., B. Tycko, A.L. Soreng, and J. Sklar, *Transrearrangements between antigen receptor genes in normal human lymphoid tissues and in ataxia telangiectasia.* J Immunol, 1991. **147**(9): p. 3201-9.

30. Allam, A. and D. Kabelitz, *TCR trans-rearrangements: biological significance in antigen recognition vs the role as lymphoma biomarker.* J Immunol, 2006. **176**(10): p. 5707-12.

31. Yueh-hsiu Chien, R. Jores, and M.P. Crowley, *RECOGNITION BY γ/δ T CELLS.* Annual Review of Immunology, 1996. **14**(1): p. 511-532.

32. Parra, Z.E., Y. Ohta, M.F. Criscitiello, M.F. Flajnik, and R.D. Miller, *The dynamic TCRdelta: TCRdelta chains in the amphibian Xenopus tropicalis utilize antibody-like V genes.* Eur J Immunol, 2010.

33. Parra, Z.E., M. Lillie, and R.D. Miller, *A Model for the Evolution of the Mammalian T-cell Receptor α/δ and μ Loci Based on Evidence from the Duckbill Platypus.* Molecular Biology and Evolution, 2012. **29**(10): p. 3205-3214.

34. Parra, Z.E., K. Mitchell, R.A. Dalloul, and R.D. Miller, *A second TCRdelta locus in Galliformes uses antibody-like V domains: insight into the evolution of TCRdelta and TCRmu genes in tetrapods.* J Immunol, 2012. **188**(8): p. 3912-9.

35. Saha, N.R., T. Ota, G.W. Litman, J. Hansen, Z. Parra, E. Hsu, F. Buonocore, A. Canapa, J.F. Cheng, and C.T. Amemiya, *Genome complexity in the coelacanth is reflected in its adaptive immune system.* J Exp Zool B Mol Dev Evol, 2014. **322**(6): p. 438-63.

36. Criscitiello, M.F., M. Saltis, and M.F. Flajnik, *An evolutionarily mobile antigen receptor variable region gene: Doubly rearranging NAR-TcR genes in sharks.* Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(13): p. 5036-5041.

37. Parra, Z.E., M.L. Baker, R.S. Schwarz, J.E. Deakin, K. Lindblad-Toh, and R.D. Miller, *A unique T cell receptor discovered in marsupials.* Proc Natl Acad Sci U S A, 2007. **104**(23): p. 9776-81.

38. Wang, X., Z.E. Parra, and R.D. Miller, *Platypus TCRmu provides insight into the origins and evolution of a uniquely mammalian TCR locus.* J Immunol, 2011. **187**(10): p. 5246-54.

39. Rock, E.P., P.R. Sibbald, M.M. Davis, and Y.H. Chien, *CDR3 length in antigen-specific immune receptors.* J Exp Med, 1994. **179**(1): p. 323-8.

40. Rumfelt, L.L., E.C. McKinney, E. Taylor, and M.F. Flajnik, *The development of primary and secondary lymphoid tissues in the nurse shark Ginglymostoma cirratum: B-cell zones precede dendritic cell immigration and T-cell zone formation during ontogeny of the spleen.* Scand.J Immunol., 2002. **56**(2): p. 130-148.

41. Dooley, H., E.B. Buckingham, M.F. Criscitiello, and M.F. Flajnik, *Emergence of the acute-phase protein hemopexin in jawed vertebrates.* Mol Immunol, 2010. **48**(1-3): p. 147-52.

42. Criscitiello, M.F., Y. Ohta, M.D. Graham, J.O. Eubanks, P.L. Chen, and M.F. Flajnik, *Shark class II invariant chain reveals ancient conserved relationships with cathepsins and MHC class II.* Dev Comp Immunol, 2012. **36**(3): p. 521-33.

43. Luo, M., H. Kim, D. Kudrna, N.B. Sisneros, S.J. Lee, C. Mueller, K. Collura, A. Zuccolo, E.B. Buckingham, S.M. Grim, K. Yanagiya, H. Inoko, T. Shiina, M.F. Flajnik, R.A. Wing, and Y. Ohta, *Construction of a nurse shark (Ginglymostoma cirratum) bacterial artificial chromosome (BAC) library and a preliminary genome survey.* BMC Genomics, 2006. **7**: p. 106.

44. Lefranc, M.P., C. Pommie, M. Ruiz, V. Giudicelli, E. Foulquier, L. Truong, V. Thouvenin-Contet, and G. Lefranc, *IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains.* Dev Comp Immunol, 2003. **27**(1): p. 55-77.

45. Luo, M., H. Kim, D. Kudrna, N.B. Sisneros, S.-J. Lee, C. Mueller, K. Collura, A. Zuccolo, E.B. Buckingham, S.M. Grim, K. Yanagiya, H. Inoko, T. Shiina, M.F. Flajnik, R.A. Wing, and Y. Ohta, *Construction of a nurse shark (Ginglymostoma cirratum) bacterial artificial chromosome (BAC) library and a preliminary genome survey.* BMC Genomics, 2006. **7**(1): p. 106.

46. Mertz, L.M. and A. Rashtchian, *Nucleotide imbalance and polymerase chain reaction: effects on DNA amplification and synthesis of high specific activity radiolabeled DNA probes.* Anal Biochem, 1994. **221**(1): p. 160-5.

47. Miller, J.R., A.L. Delcher, S. Koren, E. Venter, B.P. Walenz, A. Brownley, J. Johnson, K. Li, C. Mobarry, and G. Sutton, *Aggressive assembly of pyrosequencing reads with mates.* Bioinformatics, 2008. **24**(24): p. 2818-24.

48. Koren, S., M.C. Schatz, B.P. Walenz, J. Martin, J.T. Howard, G. Ganapathy, Z. Wang, D.A. Rasko, W.R. McCombie, E.D. Jarvis, and A.M. Phillippy, *Hybrid error correction and de novo assembly of single-molecule sequencing reads.* Nature Biotechnology, 2012. **30**: p. 693.

49. Rumfelt, L.L., R.L. Lohr, H. Dooley, and M.F. Flajnik, *Diversity and repertoire of IgW and IgM VH families in the newborn nurse shark.* BMC.Immunol., 2004. **5**(1): p. 8.

50. Greenberg, A., *Evolution of the antigen receptor family.* . 1994, University of Miami: Coral Gables, Florida.

51. Lefranc, M.-P., V. Giudicelli, C. Ginestoux, J. Bodmer, W. Müller, R. Bontrop, M. Lemaitre, A. Malik, V. Barbié, and D. Chaume, *IMGT, the international ImMunoGeneTics database.* Nucleic Acids Research, 1999. **27**(1): p. 209-212.

52. Lefranc, M.-P., V. Giudicelli, P. Duroux, J. Jabado-Michaloud, G. Folch, S. Aouinti, E. Carillon, H. Duvergey, A. Houles, T. Paysan-Lafosse, S. Hadi-Saljoqi, S. Sasorith, G. Lefranc, and S. Kossida, *IMGT®, the international ImMunoGeneTics information system® 25 years on.* Nucleic Acids Research, 2015. **43**(D1): p. D413-D422.

53. Mashoof, S., C. Pohlenz, P.L. Chen, T.C. Deiss, D. Gatlin, 3rd, A. Buentello, and M.F. Criscitiello, *Expressed IgH mu and tau transcripts share diversity segment in ranched Thunnus orientalis.* Dev Comp Immunol, 2014. **43**(1): p. 76-86.

54. Sudhir Kumar, G.S., and Koichiro Tamura, *MEGA7: Molecular Evolutionary Genetics Analysis version 7.0. Molecular Biology and Evolution*. 2015.

55. Rock, E.P., P.R. Sibbald, M.M. Davis, and Y.H. Chien, *CDR3 length in antigen-specific immune receptors.* The Journal of Experimental Medicine, 1994. **179**(1): p. 323-328.

56. Rumfelt, L.L., D. Avila, M. Diaz, S. Bartl, E.C. McKinney, and M.F. Flajnik, *A shark antibody heavy chain encoded by a nonsomatically rearranged VDJ is preferentially expressed in early development and is convergent with mammalian IgG.* Proc Natl Acad Sci U S A, 2001. **98**(4): p. 1775-80.

57. Rumfelt, L.L., M. Diaz, R.L. Lohr, E. Mochon, and M.F. Flajnik, *Unprecedented multiplicity of Ig transmembrane and secretory mRNA forms in the cartilaginous fish.* J Immunol, 2004. **173**(2): p. 1129-39.

58. Greenberg, A.S., A.L. Hughes, J. Guo, D. Avila, E.C. McKinney, and M.F. Flajnik, *A novel "chimeric" antibody class in cartilaginous fish: IgM may not be the primordial immunoglobulin.* Eur.J Immunol., 1996. **26**(5): p. 1123-1129.

59. Read, T.D., R.A. Petit, S.J. Joseph, M.T. Alam, M.R. Weil, M. Ahmad, R. Bhimani, J.S. Vuong, C.P. Haase, D.H. Webb, M. Tan, and A.D.M. Dove, *Draft sequencing and assembly of the genome of the world's largest fish, the whale shark: Rhincodon typus Smith 1828.* BMC Genomics, 2017. **18**(1): p. 532.

60. Hara, Y., K. Yamaguchi, K. Onimaru, M. Kadota, M. Koyanagi, S.D. Keeley, K. Tatsumi, K. Tanaka, F. Motone, Y. Kageyama, R. Nozu, N. Adachi, O. Nishimura, R. Nakagawa, C. Tanegashima, I. Kiyatake, R. Matsumoto, K. Murakumo, K. Nishida, A. Terakita, S. Kuratani, K. Sato, S. Hyodo, and S. Kuraku, *Shark genomes provide insights into elasmobranch evolution and the origin of vertebrates.* Nature Ecology & Evolution, 2018. **2**(11): p. 1761-1771.

61. Venkatesh, B., A.P. Lee, V. Ravi, A.K. Maurya, M.M. Lian, J.B. Swann, Y. Ohta, M.F. Flajnik, Y. Sutoh, M. Kasahara, S. Hoon, V. Gangu, S.W. Roy, M. Irimia, V. Korzh, I. Kondrychyn, Z.W. Lim, B.-H. Tay, S. Tohari, K.W. Kong, S. Ho, B. Lorente-Galdos, J. Quilez, T. Marques-Bonet, B.J. Raney, P.W. Ingham, A. Tay, L.W. Hillier, P. Minx, T. Boehm, R.K. Wilson, S. Brenner, and W.C.

Warren, *Elephant shark genome provides unique insights into gnathostome evolution.* Nature, 2014. **505**: p. 174.

62. Venkatesh, B., E.F. Kirkness, Y.H. Loh, A.L. Halpern, A.P. Lee, J. Johnson, N. Dandona, L.D. Viswanathan, A. Tay, J.C. Venter, R.L. Strausberg, and S. Brenner, *Survey sequencing and comparative analysis of the elephant shark (Callorhinchus milii) genome.* PLoS Biol, 2007. **5**(4): p. e101.

63. M., H.S., S.E. W., and L.P. E., *An architectural perspective on signaling by the pre-, αβ and γδ T cell receptors.* Immunological Reviews, 2003. **191**(1): p. 28-37.

64. Smith-Garvin, J.E., G.A. Koretzky, and M.S. Jordan, *T Cell Activation.* Annual review of immunology, 2009. **27**: p. 591-619.

65. Ribeiro, S.T., J.C. Ribot, and B. Silva-Santos, *Five Layers of Receptor Signaling in γδ T-Cell Differentiation and Activation.* Frontiers in Immunology, 2015. **6**: p. 15.

66. Flajnik, M.F., *A cold-blooded view of adaptive immunity.* Nature reviews. Immunology, 2018. **18**(7): p. 438-453.

67. Flajnik, M.F. and M. Kasahara, *Origin and evolution of the adaptive immune system: genetic events and selective pressures.* Nature reviews. Genetics, 2010. **11**(1): p. 47-59.

68. Hsu, E., *Assembly and Expression of Shark Ig Genes.* J Immunol, 2016. **196**(9): p. 3517-23.

69. Hsu, E., *V(D)J recombination: of mice and sharks.* Adv Exp Med Biol, 2009. **650**: p. 166-79.

70. Chen, H., H. Bernstein, P. Ranganathan, and S.F. Schluter, *Somatic hypermutation of TCR gamma V genes in the sandbar shark.* Dev Comp Immunol, 2011.

71. Chen, H., S. Kshirsagar, I. Jensen, K. Lau, R. Covarrubias, S.F. Schluter, and J.J. Marchalonis, *Characterization of arrangement and expression of the T cell receptor gamma locus in the sandbar shark.* Proc Natl Acad Sci U S A, 2009. **106**(21): p. 8591-6.

72. Renz, A.J., A. Meyer, and S. Kuraku, *Revealing less derived nature of cartilaginous fish genomes with their evolutionary time scale inferred with nuclear genes.* PLoS One, 2013. **8**(6): p. e66400.

CHAPTER IV

CONCLUSIONS

4.1 Primary diversification, limited RAG V(D)J diversity receives AID

Although primary repertoire diversification of lymphocyte antigen receptors is accomplished through RAG mediated V(D)J recombination, a number of jawed vertebrates use AID-mediated mechanisms to formulate repertoires prior to antigen exposure. AID mediated primary repertoire diversification mechanisms include SHM in naïve B cell populations of ruminant and porcine species [1] as well as immunoglobulin gene conversion in the Ig loci of chickens [2], rabbits [3], and cattle [4]. Interestingly, the lamprey makes use of APOBEC family member cytidine deaminases (CDA1 and CDA2) to diversify VLRs in a gene conversion-like manner [5, 6]. Whether this is an example of convergent formulation of lymphocyte repertoires by cytidine deaminases or retention of an ancestral deaminase function in primary lymphocyte repertoire diversification lost in many jawed vertebrate lineages is unknown. T cell receptors are also subject to AID mediated SHM as documented in the TCRδ chain of dromedary camels [7], the TCRγ chain of the sandbar shark [8], and the TCRγ, TCRδ and TCRα chains of the nurse shark [9]. The purpose of SHM appears to be repertoire diversification, rather than affinity maturation, for all TCR chains except TCRα. The role of SHM in nurse shark TCRα is posited as a receptor editing mechanism to rescue T cells failing selection [9].

Primary diversification of IgH and IgL loci by AID mechanisms is typically associated with a relatively small pool of V(D)J segments. In some instances, only a single functional V segment is present in the Ig loci that undergoes gene conversion and/or SHM to formulate a diverse repertoire. Studies support the notion that junctional diversity in CDR3 alone is sufficient to encode a sizeable antigen receptor repertoire with a multitude of specificities [10]. However presence of only a single functional V segment family, defined by >75% nucleotide identity, is confirmed to be associated with AID primary diversification in the IgH locus of cattle [1], rabbits [11], pigs [12], chicken [2], and sheep [13]. In these species formulation of a diverse repertoire extends beyond RAG activity and the random selection of V(D)J segments. B cell development and diversification in cattle, chickens, rabbits and sheep also occurs in bone marrow then gut associated lymphoid tissue (GALT). This differs from the traditional mouse and human model where B cell development and primary repertoire diversification is ascribed to the bone marrow and carried out by RAG mediated V(D)J recombination alone. In this traditional model AID only acts to diversify functional BCR following antigen exposure [14]. B cell development within the GALT involves lymphopoiesis and rearrangements within the GALT, rather than the bone marrow [15]. The remainder of this section will focus on B cell GALT development in *B. taurus*.

Development of B cells in cattle is a confounding process with RAG activity found in the bone marrow, ileal Peyer's patch (iPP), lymph nodes, and spleen in fetal calves [16-20], coinciding with AID expression in the iPP and spleen [21]. The regulation of these two potent diversification pathways simultaneously within a single

tissue is relatively unexplored beyond the expression at the same developmental time points. Fetal expression of AID in the absence of antigen surprisingly mirrors the characteristics of post antigen exposure SHM with selection enhancing non-synonymous mutations that are concentrated in the CDRs [21]. This is due, at least in part, to the higher concentration of AID (WRCY/RGYW) mutation hotspots in bovine CDRs relative to FRs where the average percentage of C/G hotspot bases per region is 10.7% and 7.3% respectively. Ultralong CDR3 rearrangements, which encode CDR3 lengths of greater than 40 aa residues, are included among fetal rearrangements and are not limited to neonatal development in the presence of microflora [22]. Since a single V-D pairing produces the vast majority of ultralong CDR3 receptors, this structurally unique BCR class relies heavily on naïve SHM to yield diversity as opposed to junctional diversity at V-D and D-J. Although ultralong CDR3 transcripts display long N/P insertions at the V-D junction, these additions are not diverse, instead resulting in a semi-conserved motif characterized by positively charged amino acid residues [23]. The Lys bias of the semi-conserved KT[K_{1-5}] non-template motif could be attributed to bovine TdT's bias to insert A:T bases (64% TdT additions) in this junction [24]. AID mediated SHM is also integral to structural diversification within the knob domain, in which 38 of 49 codons are mutable to Cys with a single base change [25]. The DH8-2 segment, which encodes the entirety of the knob, has a dense concentration of WRCY/RGYW AID mutation hotspots providing ample opportunities for Cys codons to arise during B cell development. Mutations yielding Cys residues contribute to unique disulfide bond signatures in the knob domains, which is prominently involved in antigen interactions [25]. In addition to

mutations yielding unique Cys disulfide bonding patterns, the densely packed AID hotspots within the knob domain enhance the occurrence of dsDNA breaks (DSBs). The DSBs result in excision of genetic material within CDR3, in a manner similar to AID mediated CSR resection events within IgH switch regions [26], and act to further diversify the content and structure of the knob domain [27, 28]. The effect of these alterations in relation to paratope formation will be further expanded in section 4.3.

4.2 The elastic TCRδ

The γδ T cells of multiple species employ receptors that can be further diversified by AID mediated SHM to expand the naïve repertoire. Additionally, the TCRδ repertoire of a growing number of species is characterized by the use of Ig-like V segment rearrangements. Rearrangement of Ig-TCR chimeric receptors could be deleterious to the MHC-restricted αβ T cell subset. Although γδTCRs are documented to interact with non-classical MHC molecules such as CD1 [29], which presents lipid antigens to CD1 restricted γδ T cells, γδTCR also demonstrates the capability to interact with free antigen in a manner akin to that of a BCR [30]. With shared regulatory elements governing RAG mediated V(D)J recombination and the similar antigenic interactions of BCR and γδTCR, the use of IgHV to diversify TCRδ repertoires would theoretically be well tolerated.

The expansion of genomic resources has revealed an elastic relationship between TCRδ and IgH loci involving multiple exchanges of V segments over the course of vertebrate evolution. Expression of Ig-TCRδ chains has been identified in the frog [31],

chicken [32], platypus [33], zebra finch [34], and nurse shark [35]. The origin of these rearrangements also varies amongst species possessing Ig-TCRδ transcripts. In avian species and frogs the Ig-like V segments expressed on TCRδ receptors are integrated into the TCRαδ locus and are termed VHδ. These VHδ segments, while clearly more related to IgHV segments than TCRδV, are still produced within a single TCRδ locus. This contrasts true "trans"-rearrangements that are extremely rare occurrences between distinct and distant antigen receptor loci, which are usually associated with disease states such as acute B lymphoblastic leukemia [36, 37]. In otherwise healthy nurse sharks, Ig-TCRδ transcripts originate both from *bona-fide* IgHV segments, existing in IgH mini-cluster loci, as well as Ig-like V segments integrated into the TCRδ locus, called TAILVs. Therefore, nurse shark Ig-TCRδ receptors are product of "trans"-rearrangements, tolerated rearrangements between IgHV and TCR loci, as well as V segments integrated into the TCRδ locus and potentially adapted solely for use on TCRδ. This provides a framework of evolutionary history suggesting that V segment mobility between Ig and TCRδ loci is one way, with IgHV contributing to TCRδ segmental diversity.

Furthering the paradigm of TCRδ elasticity is the presence of a third TCRδ rearrangement unique to cartilaginous fish, NARTCR [38]. NARTCR receptors require rearrangement of two IGSF V-domains. The membrane distal V-domain is closely related to the single chain IgNAR BCRs unique to elasmobranchii, and is spliced to a specific NARTCR supporting TCRδV rearrangements. The NARTCR supporting TCRδVs are not involved in canonical TCRδ rearrangements as they lack a leader

peptide and would not be translated as a single variable domain TCRδ [38]. A structural analog to NARTCRδ, also possessing two V-domains was identified in marsupials within a unique locus termed TCRμ [39]. However, only the membrane distal V-domain undergoes RAG rearrangement, with the proximal V domain pre-packaged as a germline VDJ fusion [39]. Additionally both TCRμ V-domains are more closely related to IgHV than TCRδV, and are only expressed on a wholly unique TCRμ C region [39]. Finally, although NARTCR is present across chondrichthian lineages, IgNAR has not been identified in the holocephali lineage, and appears to be unique to elasmobranchii [4, 38]. This circumstantial evidence provides at least one case in which a TCR locus gave rise to an IgH locus with NARTCR predating the origin of IgNAR loci.

A prospective model of gene exchange between IgH and TCRδ loci will be further explored in section 4.4.1, but the lack of a single Ig-like (VHδ/TAILV) clade in the multispecies phylogenetic analysis (figure 3.5B-C) highlights the genetic mobility of antigen receptor gene segments as a whole. The most parsimonious explanation is that the one-way exchange of IgH V segments into TCRδ loci is a common occurrence, with multiple events throughout the course of vertebrate evolution, however the TCRδ locus is capable of seeding creation of entirely new genomic IGSF loci or chains. The nurse shark association of IgH, both IgM and IgW, with IgL in the formation of a complete BCR [40], is notably absent from the single V-domain homodimeric IgNAR [41] and corroborates the hypothesis that NARTCR gave rise to IgNAR.

4.3 Novel antigen receptor paratopes

Variations on the canonical six CDR paratope, composed of two chains each with three CDR loops, have also arisen throughout the course of vertebrate evolution. These bestow a less sterically hindered interface for receptor-antigen interactions and potentially widen the scope of epitopes with which the BCR can interact. Although the aforementioned NARTCR and TCRμ likely skew antigen recognition toward the membrane distal V-domain, which is hypothesized to extend well beyond the V-domain of the partner chain, neither yet has an elucidated crystal structure to verify such claims. However there are several cases of single domain antibodies have verified crystal structures including camelid single domain VH antibodies [42, 43] and shark single chain IgNAR antibodies [41, 44] which use only IgH homodimers providing two three-CDR paratopes. These single V-domain antibodies are examples of a honed paratopes relative to the canonical two chain heterodimeric antibodies, but the epitome of the minute paratope belongs to the ultralong CDR3 antibodies of *B. taurus* [25]. These receptors possess a novel structure in which the paratope is thought to be composed of a single CDR [25].

In bovine ultralong CDR3 antibodies the paratope lies predominantly, if not exclusively, in the knob domain of the heavy chain ultralong CDR3 [25, 45]. The low diversity encoded by CDR1 and CDR2 of ULCDR3 heavy chain rearrangements are testament to negative selection acting to prohibit drastic changes to the native structure of these loops [27]. This is further corroborated by a case of positive selection acting to refine a non-germline residue just outside of CDR2 (Gly50→Ser50), which forms a

hydrogen bond with the conserved Gln97 residue of the ascending stalk [25]. The germline Gly50 residue was identified in four allelic variants of IGHV1-07, while a fifth allelic variant encodes Asp50. In contrast, the non-germline Ser50 was refined in 88.4% of ultralong CDR3 rearrangement transcripts. The Gly50 residue lies in an AID RGYW/WRCY hotspot with the penultimate nucleotide change from Gly50→Ser50 being the C which is deaminated by AID activity. To further cement the notion of positive selection acting on this AID mutated non-germline residue to confer structural support to the stalk, it is not refined in IGHV1-07 rearrangements encoding <40 aa CDR3 (43.9%), which retain the germline residue. Yet alterations to the knob domain do not necessarily ablate binding to cognate antigen, with irrelevant replacement mutations up to eight residues demonstrating an equivalent binding capacity to the native antibody sequence [25]. This could be due to the tertiary structure of the knob itself, which usually seems to encode three antiparallel β-strands with intervening loops [46]. This packaging of the core domain itself may play a greater role than the exterior loops. Indeed additional work on the ultralong CDR3 crystal structure identified additional structural motifs packed on top of the antiparallel β-strand core [28]. These additional motifs range from additional antiparallel β-strands to to α-helices which could confer different specificities [28]. Importantly ultralong CDR3 antibodies possessing internal deletions all appear to lack the second layer, possessing only the core β-strands with no secondary “stacked” layer [28]. While more structures are necessary to confirm that DH8-2 internal deletions all equate to this “single-layered” structure, they provide striking evidence supporting the initial hypothesis that deletions bring about radical

structural changes to the knob beyond simply altering the spatial position of disulfide bonds.

4.4 Future directions

4.4.1 Complete sequencing of the nurse shark TCRδ locus

The nurse shark TCRδ chain provides an interesting case in which a single TCR chain is regularly expressed with multiple V-domain lineages. In addition to the canonical TCRδV, the nurse shark TCRδ chain uses TAIL V segments, *bona-fide* IgH V segments from IgH mini-loci, and NARTCRV segments spliced to the TCRδ C region. Genomic DNA probing indicates that these V-domains are all spliced to a single TCRδ C region [35]. We know that IgH mini-clusters possessing the proper 12/23 RSS layout to facilitate Ig-TCRδ rearrangements flank the TCRδ locus of the elephant shark, but we lack expression data confirming Ig-TCRδ rearrangements in this species [47]. The divergence time between holocephali and elasmobranchii, over 400 million years [48], allows insight into the evolution of IGSF loci as neither lineage developed an IgH translocus and one notably lack an IgH class. Completion of the nurse shark TCRδ locus will confirm several important questions regarding the origins of Ig-TCRδ rearrangements. These include whether the IgH clusters proximal to the TCRδ locus in elephant shark were incorporated into the nurse shark TCRδ locus as TAILVs or did tandem internal duplications of the TCRδV translocon push these clusters to a distal location outside the scope of our current assembly. Furthermore, the presence of NARTCR has been confirmed throughout holocephali and elasmobranchii lineages, but

holocephali species appear to lack IgNAR. This could indicate a retrotransposition of NARTCR, perhaps of an alternatively spliced NARTCRδ lacking a supporting TCRδ V-domain, in an ancient holocephali:elasmobranch ancestor, giving rise to IgNAR. Alternatively given the genetic mobility of IGSF segments, and inclination for duplications of IGSF loci, the NARTCR blocks could give rise to IgNAR mini-clusters in the birth and death model of evolution [49]. In either case expression and translation of IgNAR as a single V-domain antibody, the most ancient case of such a structure, support the claim that IgNAR originated from NARTCR. Complete sequencing of the TCRδ locus could encompass IgNAR clusters in addition to the IgM/W clusters we hypothesize to frame the nurse shark TCRδ locus. Furthermore, BAC library probing and preliminary TCRδ locus assembly have also identified TCRα segments, displaying the linkage of TCRα and TCRδ in all jawed vertebrate studies to date.

As the first transposable element inserted into the primordial immunoglobulin locus would have separated V and J segments, TCRδ, requiring V-D-J rearrangement, is an unlikely candidate for the primordial locus. However, the linkage of TCRδ to TCRα could indicate the primordial IGSF antigen receptor locus has morphed into a shared locus, perhaps facilitating the canonical heterodimeric IGSF antigen receptor structure. In this theoretical model, the first transposable element inserted by RAG would be into the primordial TCRα precursor homolog, separating the single V-domain gene into V and J segments. The now rearranging primordial TCRα locus would be duplicated in its entirety followed by an additional RAG insertion yielding the precursory V-D-J separation of TCRδ. Possession of two identical chains would relax selection on the C

region of one chain, allowing differentiation of TCRα and TCRδ C regions. At this point, the primordial IGSF antigen receptor locus would possess the modern TCRαδ topography with capable production of two chains TCRα (V, J) and TCRδ (V, D, J). This is supported by the 12/23 orientation of RSSs of the chains. The first RAG insertion leaves a TCRαV_{23}-TCRαJ_{12}, following duplication the second insertion leaves TCRδV_{23}-TCRδD_{12} and TCRδD_{23}-TCRδJ_{12}. The V_{23}-$_{12}D_{23}$-$_{12}$J orientation is similar to the IgH mini-cluster RSS orientation, hinting that duplications of TCRδ could have given rise to IgH, with an additional RAG insertion creating a second D segment. The presence of short form IgW, a BCR structure with only two IgH C-domains exclusive to chondrichthyes, could indicate this as the first IgH cluster making a jump from the ancient TCRαδ locus [50, 51]. The configuration of short form IgW could be explained by duplication or transposition of a TCR V-D-J in tandem with both TCR C regions eventually diverging into CHω1 and CHω2 exons. Retrotransposition of any expressed chain and RAG insertions of inverted RSS spacers could give rise to IgL, with a V_{12}-$_{23}$J, and loci not bearing V segments flanked by a 23RSS. The jumping of miniclusters acting to seed the multiple lineages of IGSF receptors is the most parsimonious explanation as eventually all IGSF antigen receptor loci have fixated on the translocon topography. If true this would mean all four TCR chains, with translocon topography, predate IgH and IgL chains which exist as miniclusters in ancient IGSF adaptive systems. Co-evolution with MHC may have facilitated the heterodimeric structure common to all antigen receptors, but it is reasonable to suggest non-canon heterodimeric pairings of TCRαδ were the source and MHC co-evolution simply necessitated the canon pairings textbook

to adaptive immunity. Ancient homodimer TCR pairings of TCRα or TCRδ, retained in the IgH pairing of BCRs, can also not be ruled out. Although the ordered evolution of TCR, BCR and MHC will remain speculation, until species lacking one or two components are elucidated, evolutionary tracing of NARTCR and NAR corroborate the jump-seed model proposed above. A model of the evolution of the IGSF antigen receptor chain loci highlighting the origin of each chain is displayed in Figure 4.1.

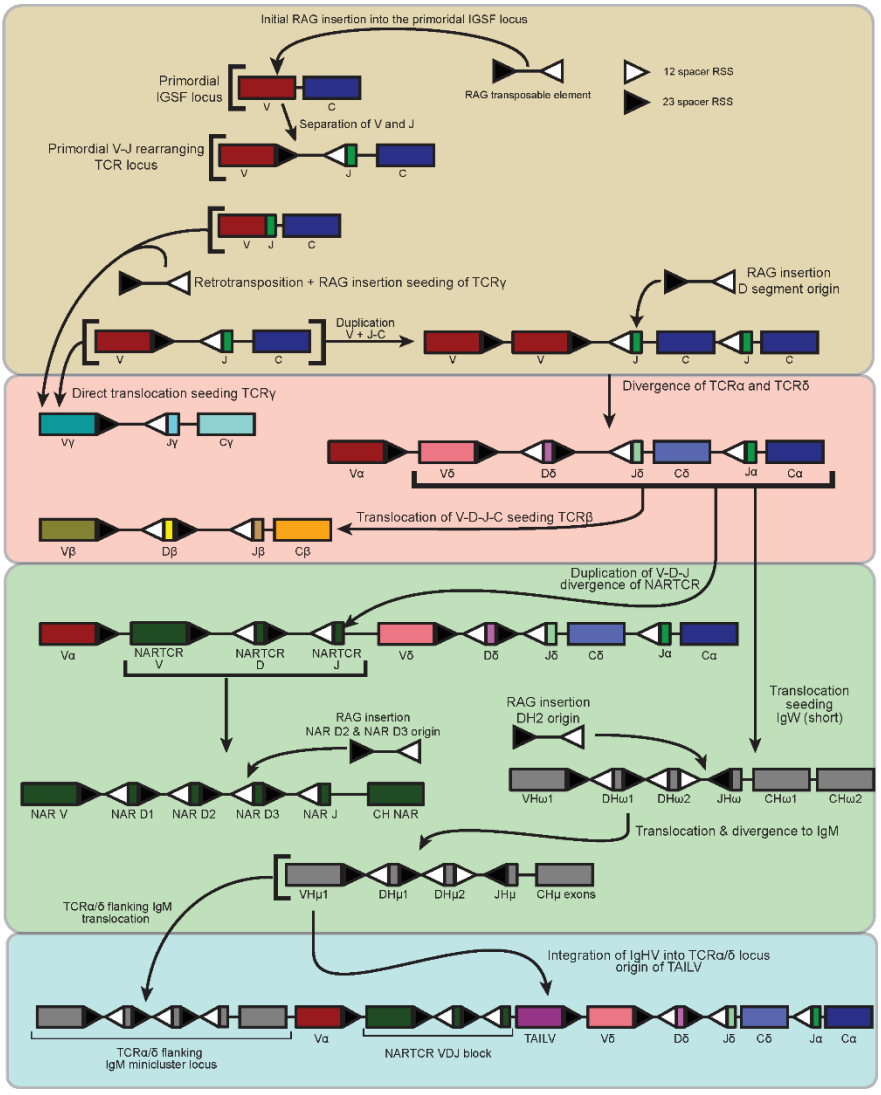


Figure 4.1 Evolution of the IGSF antigen receptor loci. Initial RAG insertion of transposable DNA and RSS orientation into the primordial IGSF locus (top, yellow). A duplication seeding the first IGSF receptor loci, prospectively TCRα/δ. Seeding of all four TCR chains and MHC co-evolution occurring in the ancient gnathostome lineage is displayed in red. Multiple duplications of stemming from the ancient precursory TCRα/δ locus gave rise to NARTCR with translocations seeding IgH miniclusters (IgW and IgM) originated in the holocephali lineage (green) eventually yielding the holocephian TCRα/δ locus with flanking IgM clusters (blue). The origin of NARTCR (green, middle) occurred during the divergence of holocephali and elasmobranchii lineages, likely stemming from the NARTCR blocks located within the TCRδ translocon.

4.4.2 Bovine ultralong phenotype and prospective role within the system

Bovine ultralong CDR3 rearrangements provide a system of receptor-antigen interaction shown to be exploitable for the recognition of previously inaccessible neutralizing epitopes. However, little work has been done to characterize ultralong CDR3 development and the role these fascinating antibodies play within their native immune system. The first step toward these goals is the creation of molecular tools to separate the B cells bearing ultralong receptors from those with canonical interfaces. Collaborative efforts with the Berghman lab (Texas A&M University) are already underway to create reagents toward conserved structural motifs unique to ultralong CDR3 receptors. Sequence data from our lab and the Smider lab at the Scripps Institute in La Jolla, CA indicate the ultralong CDR3 subset may play a heavy-handed role in mucosal immunity. Unpublished findings elucidated higher ultralong CDR3 proportions in the iPP and a skew toward IgA expression. The initial study detailing the structure and binding capacity of ultralong CDR3 antibodies by the Smider lab indicates radical knob alterations do not necessarily ablate binding [25]. This relative flexibility in antigen binding could explain the lack of plasma cells and indicates ultralong CDR3 B cells could carry out an effector function more akin to natural antibodies of B-1 B cells [52]. Development of ultralong CDR3 specific reagents will greatly assist the goal of defining the phenotype and function, but the current framework of bovine B cell development supports the speculation of a unique ultralong CDR3 role in cattle immunity.

Partially rearranged fetal B cells, at the pre/pro-B developmental junction, have been documented to migrate into peripheral lymphoid tissue for additional

recombination events [17]. It is unclear whether the proliferative burst normally associated with rearrangement of a complete IgH chain occurs before migration or following, but in either case this juncture could indicate a potential divergent fate in the development of ultralong CDR3 bearing B cells. Advances in single cell transcriptome analysis would allow fetal B cell populations (sorted using the earlies B cell fate marker BAQ44a) from bone marrow, spleen, iPP, and lymph node. Transcriptional phenotype of these early B-lymphocytes is expected to be stripped down compared to their mature counterparts, with emphasis on IgH/IgL, recombination and diversification genes, and surface markers indicative of developmental stage. Even taking into account the relatively low coverage of transcriptomes produced by the 10X genomics or Fluidigm pipeline, we should be able to glean at least three important developmental characteristics. The first, and most feasible goal would be the relationship of IgH rearrangement to developmental tissue. With this information we may already be able to confirm tissue specific development of ultralong CDR3 bearing B cells. Secondly, we should be able to determine whether AID mediated naïve SHM and/or gene conversion of IgH occurs during the proliferative burst that precedes rearrangement of IgL. Finally, we may be able to identify the cell surface markers receiving the signals that ultimately determine the fate of the developing cells.

Identification of a unique ultralong CDR3 B cell surface marker phenotype, or elicitation of a targeting reagent, would greatly enhance our understanding of the native role of ultralong CDR3 B cells. The determination of confinement within a single tissue, or higher prevalence in a given tissue, could greatly enhance vaccination strategies as

researchers continue to exploit this model in the creation of antibodies against difficult epitopes.

4.4.3 Tracing the origin of ultralong CDR3 rearrangements in the Bovinae subfamily

The evolutionary history of the genetic segments involved in ultralong CDR3 could provide a platform for additional animal models, and could even unveil novel intermediates to the well-defined *B. taurus* model. In *B. taurus* selection seems to play an active role in the pairing of long DH8, encoding the knob and descending stalk, to the IGHV1-7 segment with an eight bp duplication encoding the 'TTVHQ' of the ascending stalk. Homologs of IGHV1-7 and DH8 are present in the genome assembly of the American bison (*Bison bison*), providing at least one additional animal model likely possessing ultralong CDR3 B cells. Additionally, DH8 homologs have been identified in the absence of an IGHV1-7 homolog in the genomes of the yak (*Bos mutus*) and European bison (*Bison bonasus*). However, the most interesting case elucidated by comparative analysis of the IgH locus across bovinae genomes is the water buffalo (*Bubalis bubalis*). The IgH locus of water buffalo houses multiple DH segments of intermediate length, 50 bp shorter than *B. taurus* DH8-2, but 20 bp longer than the lengthiest short DH (DH7). Additionally the intermediate water buffalo DHs all contain at least two hardcoded Cys residues, an N-terminal/5' CYSG/CCSG motif, and a 'YxYxY(E/D)Y' motif at the C-terminal/3' end. These motifs are notably absent from all "short" DH families and theoretically encode two thirds of the ultralong CDR3 structural motifs. Further studies will focus on repertoire prevalence of ultralong CDR3

rearrangements in other species as well as elucidation of the novel structures of intermediates like the “long” DH discovered in water buffalo.

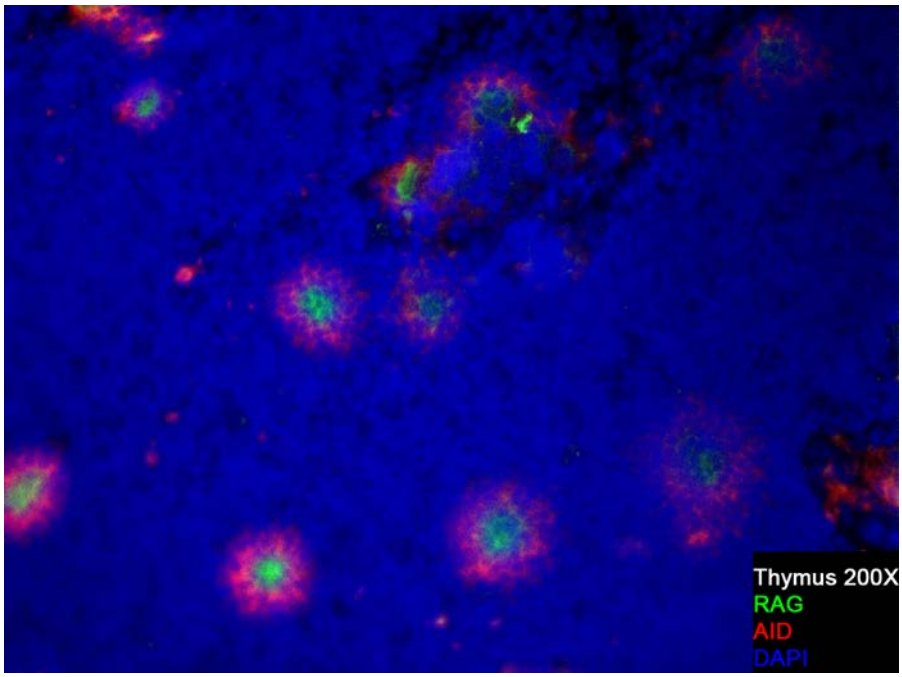


Figure 4.4 AID/RAG coexpression in nurse shark thymus. RNA-FISH on shark thymus probed for AID (red) and RAG (green) with DAPI (blue) nuclear staining. Imaged at 200X magnification.

4.4.4 AID/RAG expression during lymphocyte development within a single tissue

The framework for the final study left incomplete during my tenure as a graduate student was serendipitously provided by an RNA-FISH experiment I performed on nurse shark thymus involving the spatial regulation of AID and RAG in developing thymocytes (Fig 4.2). The unexpected result was elucidation of a germinal center-esque relationship whereby RAG expression was located inside a halo of AID expression in the cortico-medullary junction of nurse shark thymus. The two most integral mechanisms for lymphocyte receptor diversification, RAG and AID, acting within a single tissue could portend a convergent developmental pathways shared by nurse shark and cattle

lymphocytes. The proliferative burst following successful IgH rearrangement in B cell development and required expression of functional IgH to receive survival signals would prove an excellent time point for bovine naïve SHM to occur. The precedent set by the conformational relationship of RAG and AID expression within the nurse shark thymus, and similarities to the germinal center reaction during affinity maturation, resulted in the hypothesis that the two processes are likely regulated in an identical manner. An identical RNA-FISH experiment on fetal bovine lymphoid tissue sections would confirm or deny this example of convergence between the two seemingly disparate projects constituting the bulk of my doctoral research.

4.5 References

1. Berens, S.J., D.E. Wylie, and O.J. Lopez, *Use of a single VH family and long CDR3s in the variable region of cattle Ig heavy chains.* Int Immunol, 1997. **9**(1): p. 189-99.
2. Reynaud, C.A., A. Dahan, V. Anquez, and J.C. Weill, *Somatic hyperconversion diversifies the single Vh gene of the chicken with a high incidence in the D region.* Cell, 1989. **59**(1): p. 171-83.
3. Becker, R.S. and K.L. Knight, *Somatic diversification of immunoglobulin heavy chain VDJ genes: evidence for somatic gene conversion in rabbits.* Cell, 1990. **63**(5): p. 987-97.
4. Walther, S., M. Tietze, C.-P. Czerny, S. König, and U.S. Diesterbeck, *Development of a Bioinformatics Framework for the Detection of Gene*

Conversion and the Analysis of Combinatorial Diversity in Immunoglobulin Heavy Chains in Four Cattle Breeds. PLOS ONE, 2016. **11**(11): p. e0164567.

5. Boehm, T., N. McCurley, Y. Sutoh, M. Schorpp, M. Kasahara, and M.D. Cooper, *VLR-based adaptive immunity.* Annual review of immunology, 2012. **30**: p. 203-220.
6. Rogozin, I.B., L.M. Iyer, L. Liang, G.V. Glazko, V.G. Liston, Y.I. Pavlov, L. Aravind, and Z. Pancer, *Evolution and diversification of lamprey antigen receptors: evidence for involvement of an AID-APOBEC family cytosine deaminase.* Nature Immunology, 2007. **8**: p. 647.
7. Antonacci, R., M. Mineccia, M.-P. Lefranc, H.M.E. Ashmaoui, C. Lanave, B. Piccinni, G. Pesole, M.S. Hassanane, S. Massari, and S. Ciccarese, *Expression and genomic analyses of Camelus dromedarius T cell receptor delta (TRD) genes reveal a variable domain repertoire enlargement due to CDR3 diversification and somatic mutation.* Molecular Immunology, 2011. **48**(12): p. 1384-1396.
8. Chen, H., H. Bernstein, P. Ranganathan, and S.F. Schluter, *Somatic hypermutation of TCR γ V genes in the sandbar shark.* Developmental & Comparative Immunology, 2012. **37**(1): p. 176-183.
9. Ott, J.A., C.D. Castro, T.C. Deiss, Y. Ohta, M.F. Flajnik, and M.F. Criscitiello, *Somatic hypermutation of T cell receptor alpha chain contributes to selection in nurse shark thymus.* Elife, 2018. **7**.
10. Xu, J.L. and M.M. Davis, *Diversity in the CDR3 Region of VH Is Sufficient for Most Antibody Specificities.* Immunity, 2000. **13**(1): p. 37-45.

11. Knight, K.L. and R.S. Becker, *Molecular basis of the allelic inheritance of rabbit immunoglobulin VH allotypes: implications for the generation of antibody diversity.* Cell, 1990. **60**(6): p. 963-70.

12. Sun, J., I. Kacskovics, W.R. Brown, and J.E. Butler, *Expressed swine VH genes belong to a small VH gene family homologous to human VHIII.* J Immunol, 1994. **153**(12): p. 5618-27.

13. Dufour, V., S. Malinge, and F. Nau, *The sheep Ig variable region repertoire consists of a single VH family.* J Immunol, 1996. **156**(6): p. 2163-70.

14. Weill, J.-C. and C.-A. Reynaud, *Do developing B cells need antigen?* The Journal of experimental medicine, 2005. **201**(1): p. 7-9.

15. Sun, Y., Z. Liu, L. Ren, Z. Wei, P. Wang, N. Li, and Y. Zhao, *Immunoglobulin genes and diversity: what we have learned from domestic animals.* J Anim Sci Biotechnol, 2012. **3**(1): p. 18.

16. Meyer, A., C.L. Parng, S.A. Hansal, B.A. Osborne, and R.A. Goldsby, *Immunoglobulin gene diversification in cattle.* Int Rev Immunol, 1997. **15**(3-4): p. 165-83.

17. Ekman, A., T. Pessa-Morikawa, J. Liljavirta, M. Niku, and A. Iivanainen, *B-cell development in bovine fetuses proceeds via a pre-B like cell in bone marrow and lymph nodes.* Dev Comp Immunol, 2010. **34**(8): p. 896-903.

18. Yasuda, M., C.N. Jenne, L.J. Kennedy, and J.D. Reynolds, *The sheep and cattle Peyer's patch as a site of B-cell development.* Vet Res, 2006. **37**(3): p. 401-15.

19. Kozuka, Y., T. Nasu, T. Murakami, and M. Yasuda, *Comparative studies on the secondary lymphoid tissue areas in the chicken bursa of Fabricius and calf ileal Peyer's patch.* Vet Immunol Immunopathol, 2010. **133**(2-4): p. 190-7.

20. Griebel, P.J., L. Kennedy, T. Graham, W.C. Davis, and J.D. Reynolds, *Characterization of B-cell phenotypic changes during ileal and jejunal Peyer's patch development in sheep.* Immunology, 1992. **77**(4): p. 564-570.

21. Liljavirta, J., A. Ekman, J.S. Knight, A. Pernthaner, A. Iivanainen, and M. Niku, *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, 2013. **6**(5): p. 942-9.

22. Saini, S.S. and A. Kaushik, *Extensive CDR3H Length Heterogeneity Exists in Bovine Foetal VDJ Rearrangements.* Scandinavian Journal of Immunology, 2002. **55**(2): p. 140-148.

23. Koti, M., G. Kataeva, and A.K. Kaushik, *Novel atypical nucleotide insertions specifically at VH-DH junction generate exceptionally long CDR3H in cattle antibodies.* Mol Immunol, 2010. **47**(11-12): p. 2119-28.

24. Liljavirta, J., M. Niku, T. Pessa-Morikawa, A. Ekman, and A. Iivanainen, *Expansion of the preimmune antibody repertoire by junctional diversity in Bos taurus.* PLoS One, 2014. **9**(6): p. e99808.

25. Wang, F., D.C. Ekiert, I. Ahmad, W. Yu, Y. Zhang, O. Bazirgan, A. Torkamani, T. Raudsepp, W. Mwangi, M.F. Criscitiello, I.A. Wilson, P.G. Schultz, and V.V. Smider, *Reshaping antibody diversity.* Cell, 2013. **153**(6): p. 1379-93.

26. Dong, J., R.A. Panchakshari, T. Zhang, Y. Zhang, J. Hu, S.A. Volpi, R.M. Meyers, Y.-J. Ho, Z. Du, D.F. Robbiani, F. Meng, M. Gostissa, M.C. Nussenzweig, J.P. Manis, and F.W. Alt, *Orientation-specific joining of AID-initiated DNA breaks promotes antibody class switching.* Nature, 2015. **525**(7567): p. 134-139.

27. Deiss, T.C., M. Vadnais, F. Wang, P.L. Chen, A. Torkamani, W. Mwangi, M.P. Lefranc, M.F. Criscitiello, and V.V. Smider, *Immunogenetic factors driving formation of ultralong VH CDR3 in Bos taurus antibodies.* Cell Mol Immunol, 2017.

28. Dong, J., J.A. Finn, P.A. Larsen, T.P.L. Smith, and J.E. Crowe, *Structural Diversity of Ultralong CDRH3s in Seven Bovine Antibody Heavy Chains.* Frontiers in Immunology, 2019. **10**(558).

29. Castro, C.D., A.M. Luoma, and E.J. Adams, *Coevolution of T-cell receptors with MHC and non-MHC ligands.* Immunological reviews, 2015. **267**(1): p. 30-55.

30. Yueh-hsiu Chien, R. Jores, and M.P. Crowley, *RECOGNITION BY γ/δ T CELLS.* Annual Review of Immunology, 1996. **14**(1): p. 511-532.

31. Parra, Z.E., Y. Ohta, M.F. Criscitiello, M.F. Flajnik, and R.D. Miller, *The dynamic TCRdelta: TCRdelta chains in the amphibian Xenopus tropicalis utilize antibody-like V genes.* Eur J Immunol, 2010. **40**(8): p. 2319-29.

32. Parra, Z.E. and R.D. Miller, *Comparative analysis of the chicken TCRα/δ locus.* Immunogenetics, 2012. **64**(8): p. 641-645.

33. Parra, Z.E., M. Lillie, and R.D. Miller, *A model for the evolution of the mammalian t-cell receptor α/δ and μ loci based on evidence from the duckbill Platypus.* Molecular biology and evolution, 2012. **29**(10): p. 3205-3214.

34. Parra, Z.E., K. Mitchell, R.A. Dalloul, and R.D. Miller, *A second TCRdelta locus in Galliformes uses antibody-like V domains: insight into the evolution of TCRdelta and TCRmu genes in tetrapods.* J Immunol, 2012. **188**(8): p. 3912-9.

35. Criscitiello, M.F., Y. Ohta, M. Saltis, E.C. McKinney, and M.F. Flajnik, *Evolutionarily conserved TCR binding sites, identification of T cells in primary lymphoid tissues, and surprising trans-rearrangements in nurse shark.* J Immunol, 2010. **184**(12): p. 6950-60.

36. Foster, D.B., R. Huang, V. Hatch, R. Craig, P. Graceffa, W. Lehman, and C.L. Wang, *Modes of caldesmon binding to actin: sites of caldesmon contact and modulation of interactions by phosphorylation.* J Biol Chem, 2004. **279**(51): p. 53387-94.

37. Kobayashi, Y., B. Tycko, A.L. Soreng, and J. Sklar, *Transrearrangements between antigen receptor genes in normal human lymphoid tissues and in ataxia telangiectasia.* J Immunol, 1991. **147**(9): p. 3201-9.

38. Criscitiello, M.F., M. Saltis, and M.F. Flajnik, *An evolutionarily mobile antigen receptor variable region gene: doubly rearranging NAR-TcR genes in sharks.* Proc Natl Acad Sci U S A, 2006. **103**(13): p. 5036-41.

39. Parra, Z.E., M.L. Baker, R.S. Schwarz, J.E. Deakin, K. Lindblad-Toh, and R.D. Miller, *A unique T cell receptor discovered in marsupials.* Proc Natl Acad Sci U S A, 2007. **104**(23): p. 9776-81.

40. Flajnik, M.F., *Comparative analyses of immunoglobulin genes: surprises and portents.* Nat Rev Immunol, 2002. **2**(9): p. 688-98.

41. Stanfield, R.L., H. Dooley, M.F. Flajnik, and I.A. Wilson, *Crystal Structure of a Shark Single-Domain Antibody V Region in Complex with Lysozyme.* Science, 2004. **305**(5691): p. 1770.

42. Desmyter, A., T.R. Transue, M.A. Ghahroudi, M.-H. Dao Thi, F. Poortmans, R. Hamers, S. Muyldermans, and L. Wyns, *Crystal structure of a camel single-domain VH antibody fragment in complex with lysozyme.* Nature Structural Biology, 1996. **3**(9): p. 803-811.

43. Spinelli, S., L. Frenken, D. Bourgeois, L. de Ron, W. Bos, T. Verrips, C. Anguille, C. Cambillau, and M. Tegoni, *The crystal structure of a llama heavy chain variable domain.* Nat Struct Biol, 1996. **3**(9): p. 752-7.

44. Feige, M.J., M.A. Gräwert, M. Marcinowski, J. Hennig, J. Behnke, D. Ausländer, E.M. Herold, J. Peschek, C.D. Castro, M. Flajnik, L.M. Hendershot, M. Sattler, M. Groll, and J. Buchner, *The structural analysis of shark IgNAR antibodies reveals evolutionary principles of immunoglobulins.* Proceedings of the National Academy of Sciences, 2014. **111**(22): p. 8155.

45. Sok, D., K.M. Le, M. Vadnais, K.L. Saye-Francisco, J.G. Jardine, J.L. Torres, Z.T. Berndsen, L. Kong, R. Stanfield, J. Ruiz, A. Ramos, C.H. Liang, P.L. Chen,

M.F. Criscitiello, W. Mwangi, I.A. Wilson, A.B. Ward, V.V. Smider, and D.R. Burton, *Rapid elicitation of broadly neutralizing antibodies to HIV by immunization in cows.* Nature, 2017. **548**(7665): p. 108-111.

46. Stanfield, R.L., J. Haakenson, T.C. Deiss, M.F. Criscitiello, I.A. Wilson, and V.V. Smider, *The Unusual Genetics and Biochemistry of Bovine Immunoglobulins.* Adv Immunol, 2018. **137**: p. 135-164.

47. Venkatesh, B., A.P. Lee, V. Ravi, A.K. Maurya, M.M. Lian, J.B. Swann, Y. Ohta, M.F. Flajnik, Y. Sutoh, M. Kasahara, S. Hoon, V. Gangu, S.W. Roy, M. Irimia, V. Korzh, I. Kondrychyn, Z.W. Lim, B.-H. Tay, S. Tohari, K.W. Kong, S. Ho, B. Lorente-Galdos, J. Quilez, T. Marques-Bonet, B.J. Raney, P.W. Ingham, A. Tay, L.W. Hillier, P. Minx, T. Boehm, R.K. Wilson, S. Brenner, and W.C. Warren, *Elephant shark genome provides unique insights into gnathostome evolution.* Nature, 2014. **505**: p. 174.

48. Inoue, J.G., M. Miya, K. Lam, B.-H. Tay, J.A. Danks, J. Bell, T.I. Walker, and B. Venkatesh, *Evolutionary Origin and Phylogeny of the Modern Holocephalans (Chondrichthyes: Chimaeriformes): A Mitogenomic Perspective.* Molecular Biology and Evolution, 2010. **27**(11): p. 2576-2586.

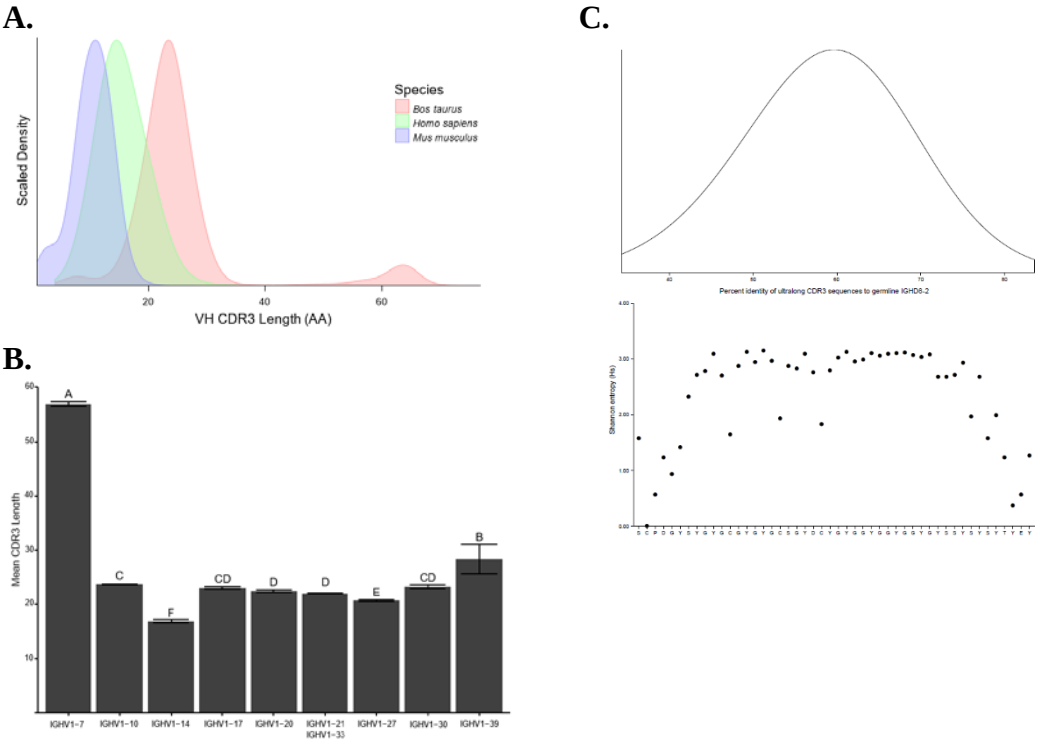
49. Karev, G.P., Y.I. Wolf, and E.V. Koonin, *Birth and Death Models of Genome Evolution*, in *Power Laws, Scale-Free Networks and Genome Biology*, E.V. Koonin, Y.I. Wolf, and G.P. Karev, Editors. 2006, Springer US: Boston, MA. p. 65-85.

50. Greenberg, A.S., A.L. Hughes, J. Guo, D. Avila, E.C. McKinney, and M.F. Flajnik, *A novel "chimeric" antibody class in cartilaginous fish: IgM may not be the primordial immunoglobulin.* European Journal of Immunology, 1996. **26**(5): p. 1123-1129.

51. Ohta, Y. and M. Flajnik, *IgD, like IgM, is a primordial immunoglobulin class perpetuated in most jawed vertebrates.* Proceedings of the National Academy of Sciences, 2006. **103**(28): p. 10723.

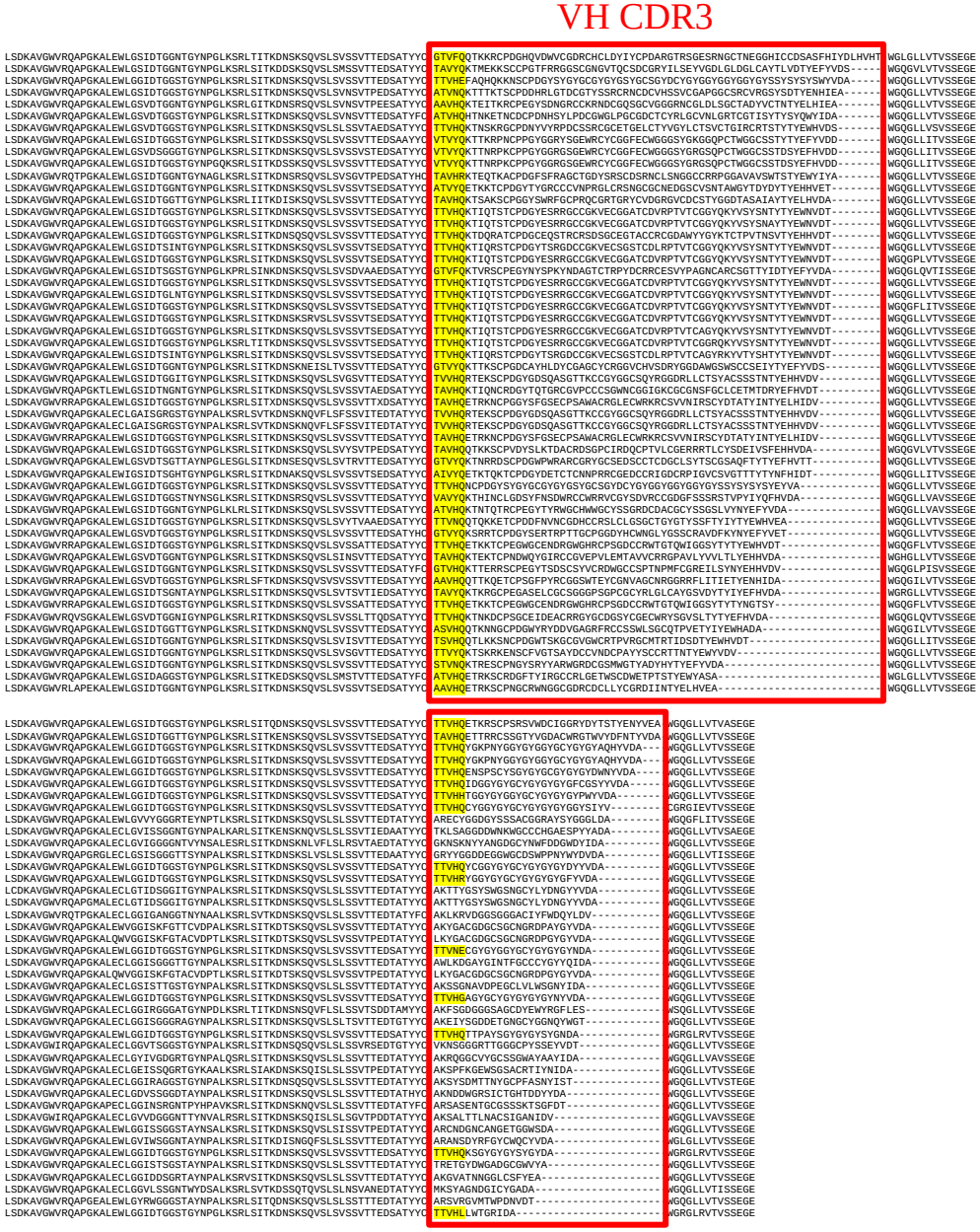
52. Hayakawa, K., R.R. Hardy, D.R. Parks, and L.A. Herzenberg, *The "Ly-1 B" cell subpopulation in normal immunodefective, and autoimmune mice.* J Exp Med, 1983. **157**(1): p. 202-18.

Appendix A1. Bovine rearrangement metrics.



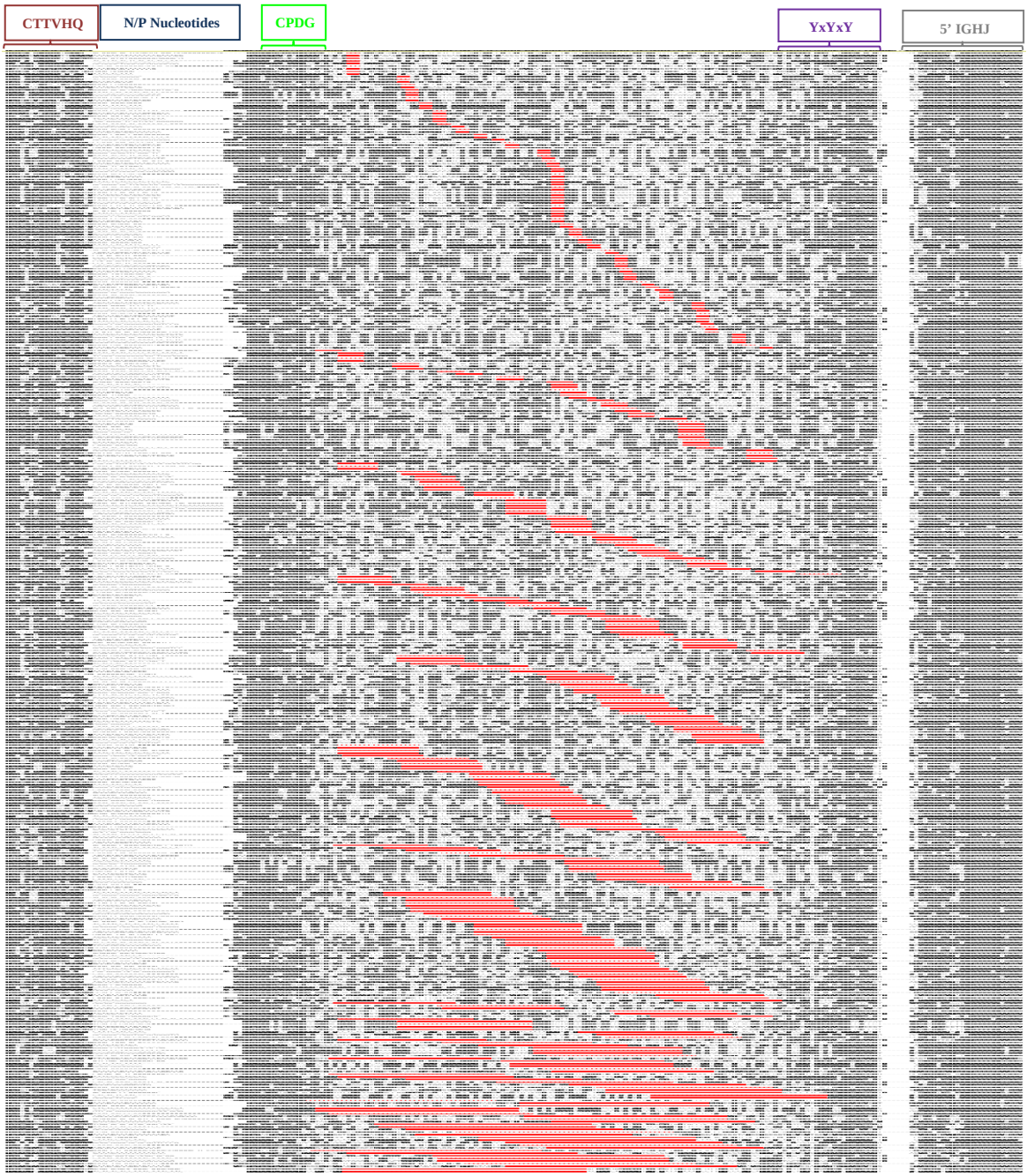
Appendix A1. Bovine rearrangement metrics. A. Comparison of variable heavy (VH) chain CDR3 length (AA) distribution for *Bos taurus*, *Homo sapiens* and *Mus musculus*. The ultralong CDR3 (≥40 AA) are the source for the bimodal distribution of the *B. taurus* CDR3. **B.** Comparison of the mean variable heavy (VH) CDR3 length (AA) of *Bos taurus* transcripts using IGHV1 genes. Solid bars represent CDR3 length for VH assigned to a given IGHV1 gene, with standard error displayed by error bars. Statistically significant different groups ($p<0.05$), determined via ANOVA and post-hoc TukeyHSD, are denoted by the lettering. **C.** Diversity analysis of *Bos taurus* ultralong variable heavy (VH) CDR3. Distribution of pairwise identity of ultralong CDR3 nucleotide sequences to germline IGHD8-2 (top). Shannon entropy analysis of ultralong CDR3 amino acid sequences bearing IGHD8-2 (bottom).

Appendix A2. Sanger sequenced *B. taurus* IgH transcripts.



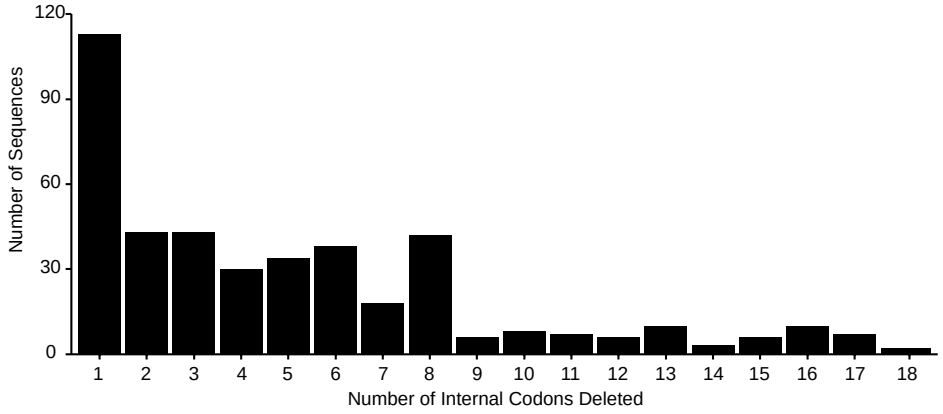
Appendix A2. Sanger sequenced *B. taurus* IgH transcripts. Amino acid alignment of *Bos taurus* targeted *IGHV1-7* sequences. The variable heavy (VH) CDR3 is boxed in red with sequences encoding ultralong CDR3 (≥40AA) on top and conventional CDR3 (<40AA) on bottom. The *IGHV1-7* specific "TTVHQ" motif is highlighted in yellow.

Appendix A3. *B. taurus* ultralong CDR3 deletion alignments and metrics.



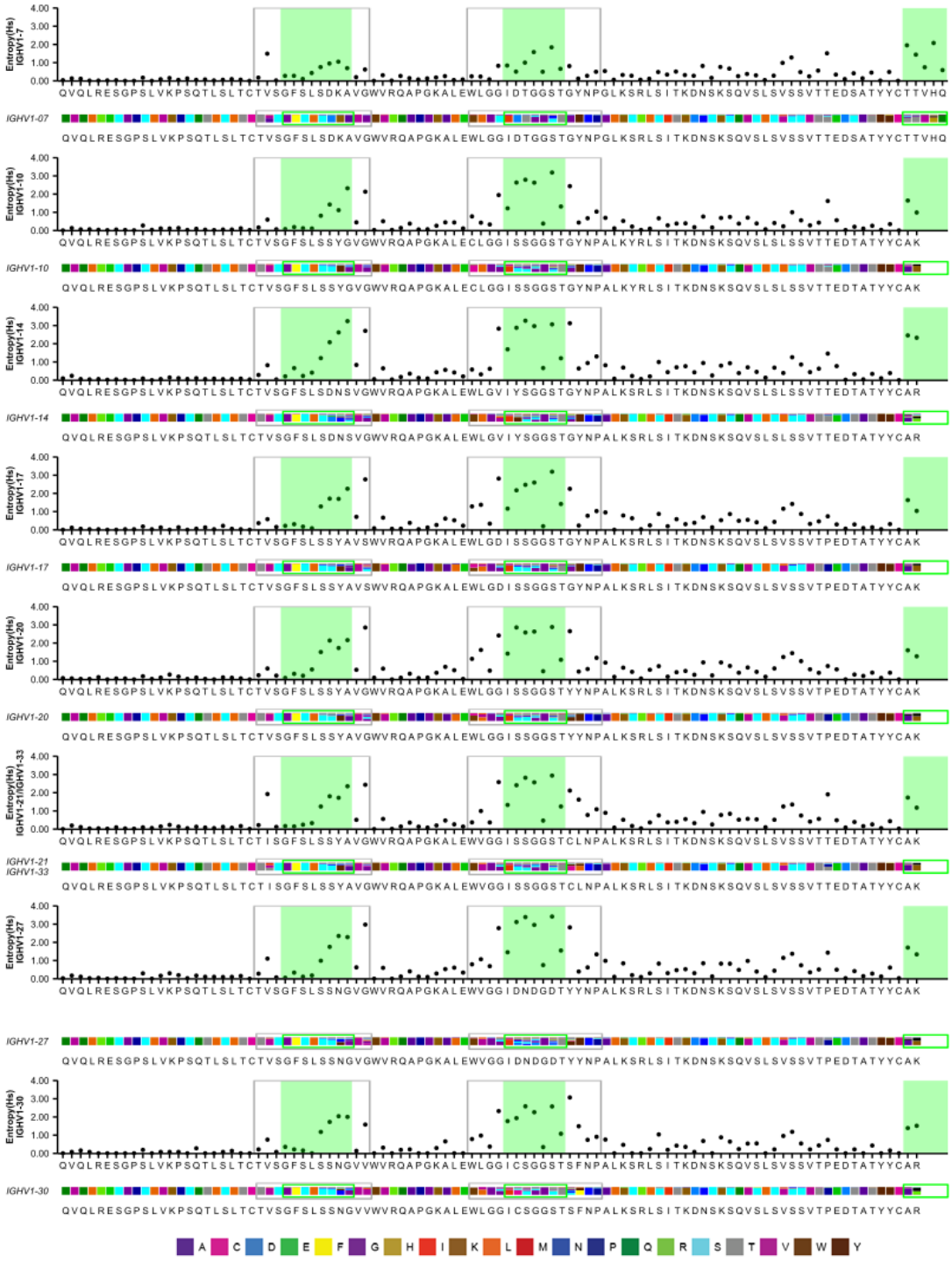
Appendix A3. *B. taurus* ultralong CDR3 deletion alignments and metrics. A. Alignment of *Bos taurus* ultralong variable heavy (VH) CDR3 sequences with gaps. Nucleotides identical to the IGHV1-7 and IGHD8-2 germline genes are indicated with a black background. Gaps in the aligned sequences are indicated with a red background. At the top of the alignment nucleotides encoding amino acids of the "CTTVHQ" motif of the IGHV, N/P region, "CPDG" and "YxYxY"/alternating aromatic amino acids of the IGHD, and the 5' region of the IGHJ are denoted at the top of the alignment.

Appendix A4. *B. taurus* ultralong CDR3 deletion graph



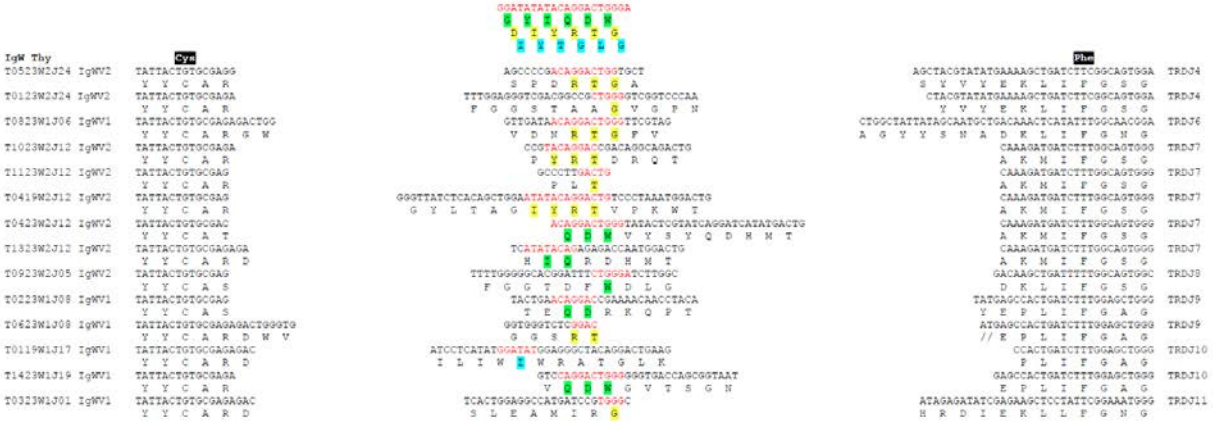
Appendix A4. *B. taurus* ultralong CDR3 deletion graph. Relative number of sequences as a function of codons deleted. The number of sequences (y-axis) which contained varying numbers of codon deletions (x-axis) is plotted.

Appendix A5. Shannon entropy plots for all *B. taurus* IgHV rearrangements



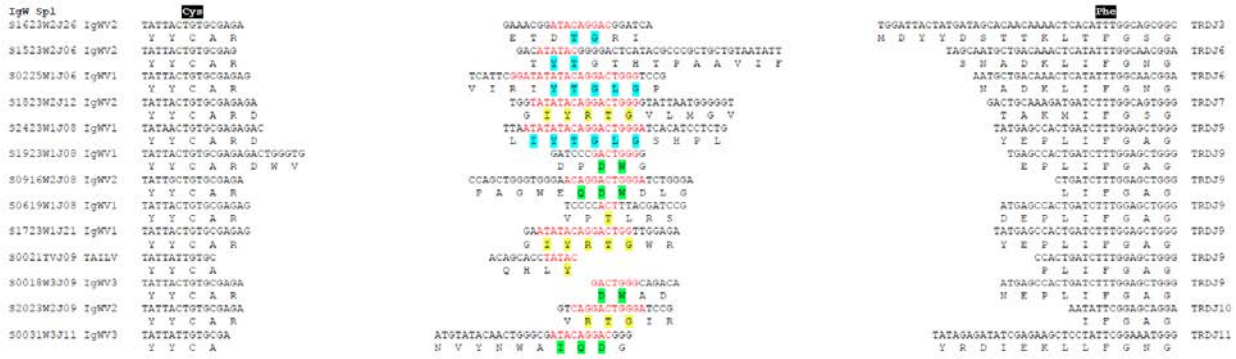
Appendix A5. Shannon entropy plots for all *B. taurus* IgHV rearrangements. Shannon entropy plots and amino acid (AA) frequency charts for individual IGHV1 subgroup genes. The CDR are indicated with a green box.

Appendix A6. Nurse shark IgWV-TCRδ CDR3 alignments in thymus



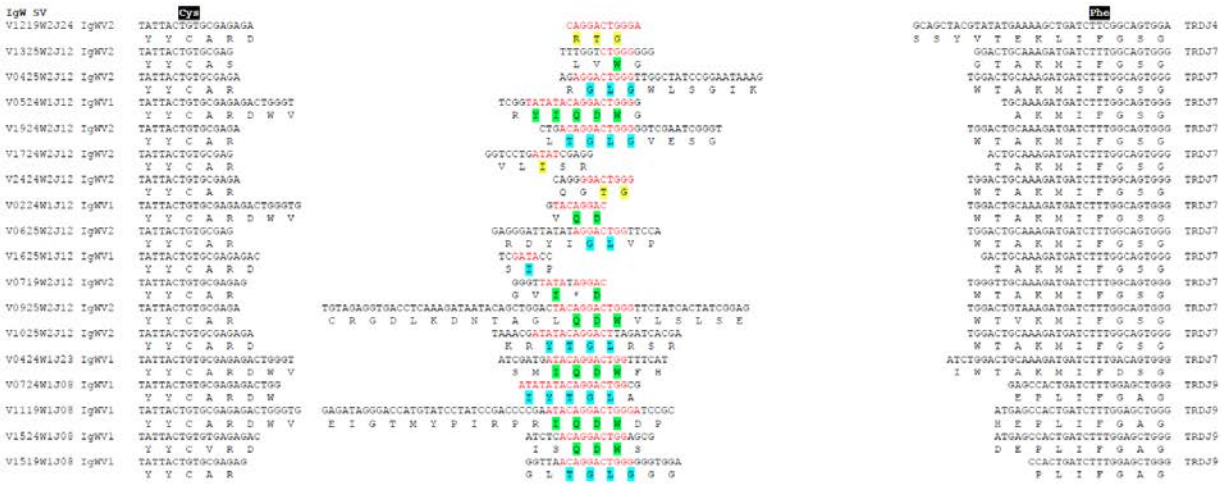
Appendix A6. Nurse shark IgWV-TCRδ CDR3 alignments in thymus. The contributing IgHV class and group are denoted in the second column preceding the nucleotide alignment. The TCRδJ used in the rearrangements are labelled immediately following the nucleotide alignment. The TRDD1 originating nucleotides are highlighted in red with the accompanying amino acid residues highlighted according to the frame the D-segment is read in (frame1-green, frame 2-yellow, frame 3-cyan). The Cys and Phe residues heighted above the nucleotide indicate the 5' and 3' CDR3 boundaries respectively.

Appendix A7. Nurse shark IgWV-TCRδ CDR3 alignments in spleen



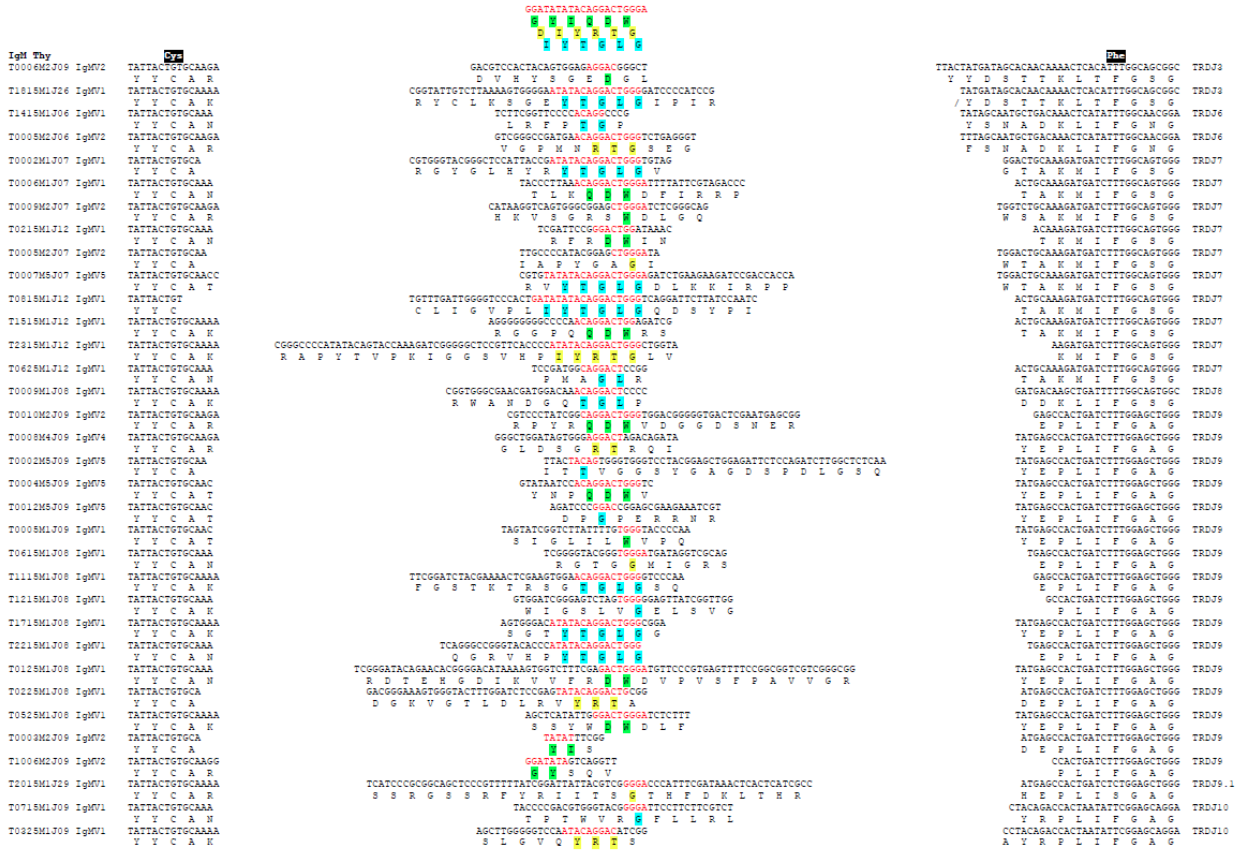
Appendix A7. Nurse shark IgWV-TCRδ CDR3 alignments in spleen. The contributing IgHV class and group are denoted in the second column preceding the nucleotide alignment. The TCRδJ used in the rearrangements are labelled immediately following the nucleotide alignment. The TRDD1 originating nucleotides are highlighted in red with the accompanying amino acid residues highlighted according to the frame the D-segment is read in (frame1-green, frame 2-yellow, frame 3-cyan). The Cys and Phe residues heighted above the nucleotide indicate the 5' and 3' CDR3 boundaries respectively.

Appendix A8. Nurse shark IgWV-TCRδ CDR3 alignments in spiral valve



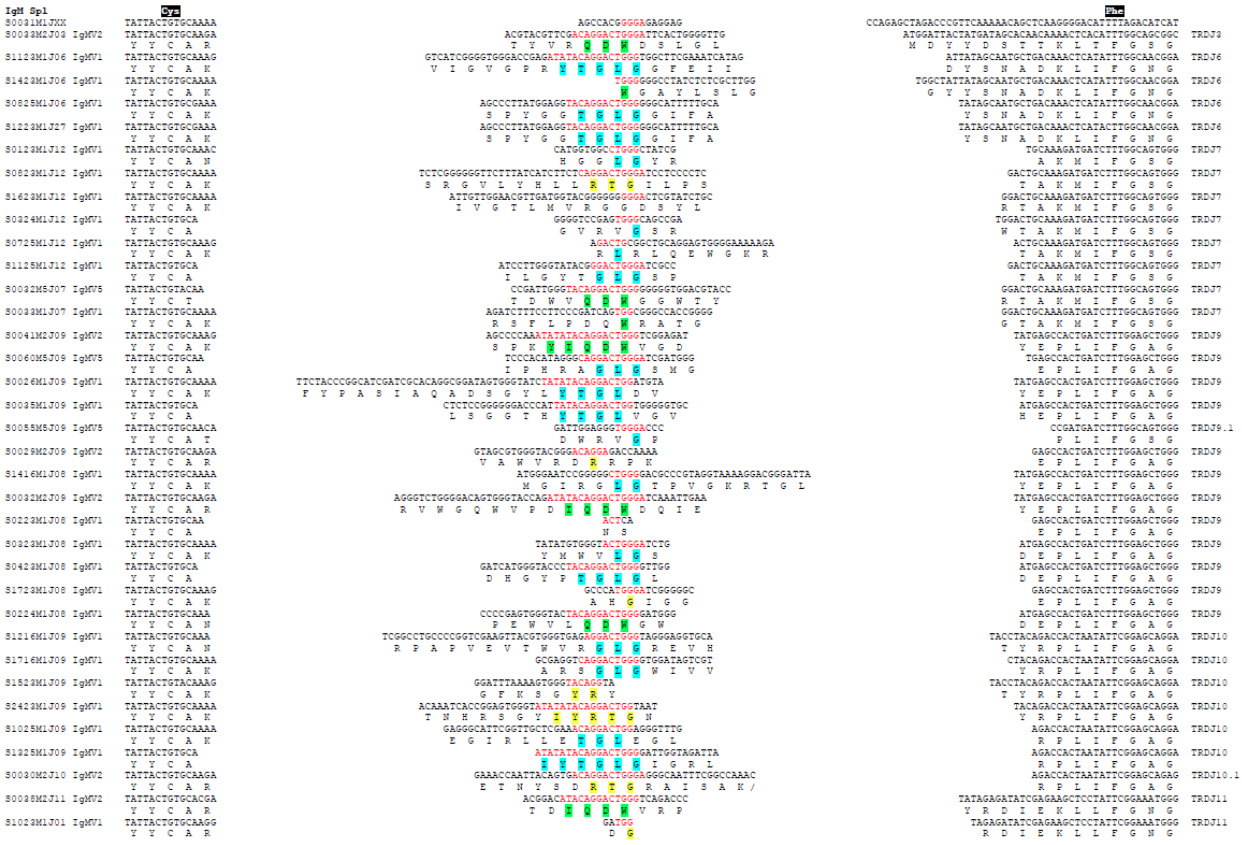
Appendix A8. Nurse shark IgWV-TCRδ CDR3 alignments in spiral valve. The contributing IgHV class and group are denoted in the second column preceding the nucleotide alignment. The TCRδJ used in the rearrangements are labelled immediately following the nucleotide alignment. The TRDD1 originating nucleotides are highlighted in red with the accompanying amino acid residues highlighted according to the frame the D-segment is read in (frame1-green, frame 2-yellow, frame 3-cyan). The Cys and Phe residues heighted above the nucleotide indicate the 5' and 3' CDR3 boundaries respectively.

Appendix A9. Nurse shark IgMV-TCRδ CDR3 alignments in thymus



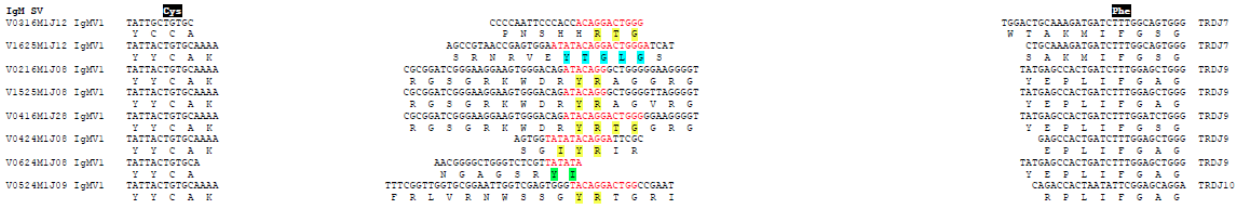
Appendix A9. Nurse shark IgMV-TCRδ CDR3 alignments in thymus. The contributing IgHV class and group are denoted in the second column preceding the nucleotide alignment. The TCRδJ used in the rearrangements are labelled immediately following the nucleotide alignment. The TRDD1 originating nucleotides are highlighted in red with the accompanying amino acid residues highlighted according to the frame the D-segment is read in (frame1-green, frame 2-yellow, frame 3-cyan). The Cys and Phe residues heighted above the nucleotide indicate the 5' and 3' CDR3 boundaries respectively.

Appendix A10. Nurse shark IgMV-TCRδ CDR3 alignments in spleen



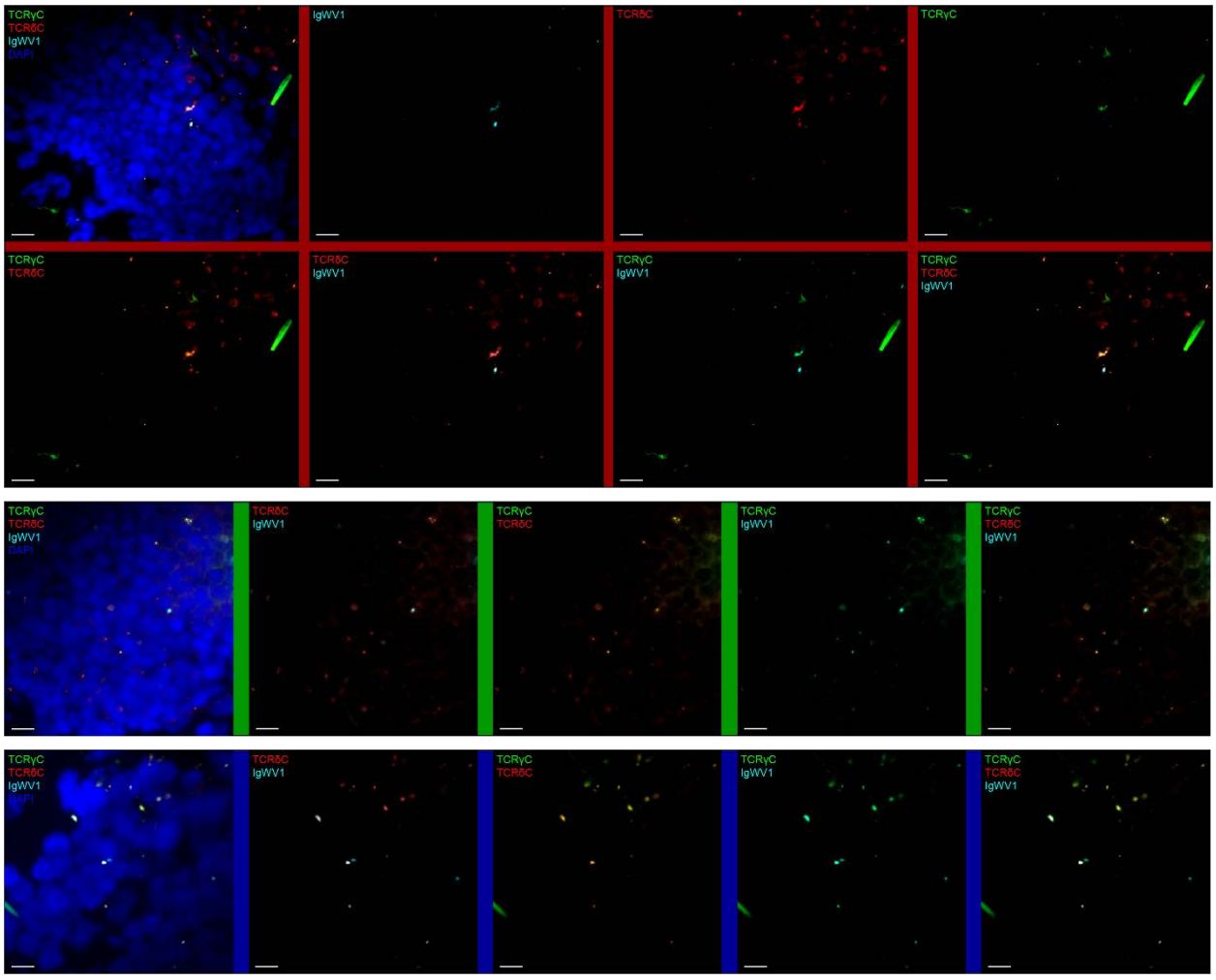
Appendix A10. Nurse shark IgMV-TCRδ CDR3 alignments in spleen. The contributing IgHV class and group are denoted in the second column preceding the nucleotide alignment. The TCRδJ used in the rearrangements are labelled immediately following the nucleotide alignment. The TRDD1 originating nucleotides are highlighted in red with the accompanying amino acid residues highlighted according to the frame the D-segment is read in (frame1-green, frame 2-yellow, frame 3-cyan). The Cys and Phe residues heighted above the nucleotide indicate the 5' and 3' CDR3 boundaries respectively.

Appendix A11. Nurse shark IgMV-TCRδ CDR3 alignments in spiral valve



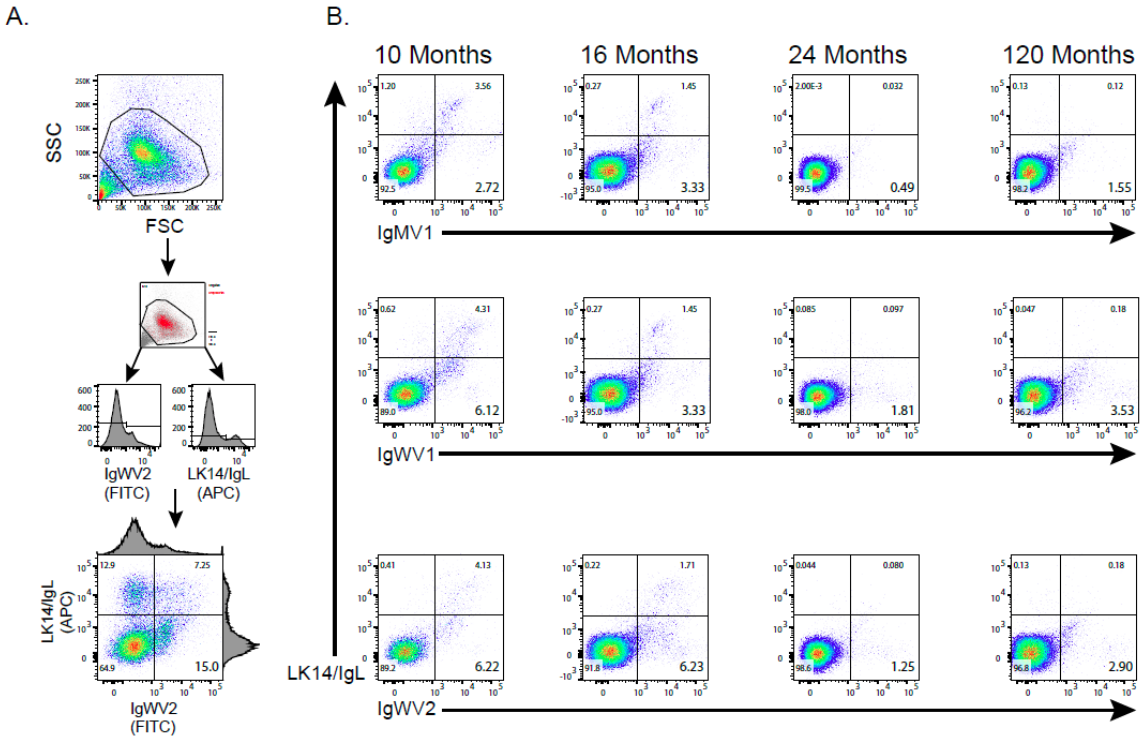
Appendix A11. Nurse shark IgMV-TCRδ CDR3 alignments in spiral valve. The contributing IgHV class and group are denoted in the second column preceding the nucleotide alignment. The TCRδJ used in the rearrangements are labelled immediately following the nucleotide alignment. The TRDD1 originating nucleotides are highlighted in red with the accompanying amino acid residues highlighted according to the frame the D-segment is read in (frame1-green, frame 2-yellow, frame 3-cyan). The Cys and Phe residues heighted above the nucleotide indicate the 5' and 3' CDR3 boundaries respectively.

Appendix A12. IgW-TCRδ rearrangements pair with TCRγ chain.



Appendix A12. IgW-TCRδ rearrangements pair with TCRγ chain. Stellaris RNA-FISH probes were hybridized to nurse shark thymus tissue sections imaged on a Zeiss Stallion Digital Imaging Workstation using a 63X objective lens. Probes were designed and pseudocolors were applied for TCRγC (green), TCRδC (red), and IgWV1 (cyan) as well as a DAPI stain (blue) to image nuclei, white scale bars indicate 10μM. Overlays show distinct IgW-TCRδ expressing transcripts, the pairing of TCRγ andTCRδ, and finally a congregation of IgW-TCRδ cells which are also TCRγ positive. Three separate tissue section included in the experiment are indicated by the red, blue, or green coloring between images.

Appendix A13. Flow cytometric analysis with Ig specific polyclonal antibodies evince IgHV-TCRδ as a cell surface receptor.



Appendix A13. Flow cytometric analysis with Ig specific polyclonal antibodies evince IgHV-TCRδ as a cell surface receptor. **(A)** Gating and signal threshold workflow using splenic lymphocytes. Lymphocyte populations were isolated using forward and side scatter (top, selected cells red in middle). Thresholds for quadrant gating was determined via histogram analysis of the FITC (IgHV polyclonal) and APC (LK14) channels (middle and bottom). All samples were analyzed with identical threshold param-eters except the 120-month sample, which had lower overall fluorescent intensity. **(B)** Quadrant plots and population percentages for sorted thymic lymphocytes. Anti-IgHV polyclonal fluorescent intensity is plotted on the x-axis, while anti-LK14/IgL intensity is plotted on the y-axis. Quadrants were all determined using splenic controls that did not diverge amongst groups except the aforementioned 120-month sample. The bottom left quadrant houses T cells (percentages in enlarged text), while the top right houses B cell populations. The bottom right quadrant, housing cells that were IgHV+/IgL- are the prospective IgHV-TCRδ chimeric receptor bearing cells.