Cellular and Molecular Immunology (2017) 14, 1–12

www.nature.com/cmi

RESEARCH ARTICLE

Immunogenetic factors driving formation of ultralong VH CDR3 in *Bos taurus* antibodies

Thaddeus C Deiss[1], Melissa Vadnais[2], Feng Wang[3], Patricia L Chen[1], Ali Torkamani[4], Waithaka Mwangi[5], Marie-Paule Lefranc[6], Michael F Criscitiello[1,7] and Vaughn V Smider[2,8]

The antibody repertoire of *Bos taurus* is characterized by a subset of variable heavy (VH) chain regions with ultralong third complementarity determining regions (CDR3) which, compared to other species, can provide a potent response to challenging antigens like HIV env. These unusual CDR3 can range to over seventy highly diverse amino acids in length and form unique β-ribbon 'stalk' and disulfide bonded 'knob' structures, far from the typical antigen binding site. The genetic components and processes for forming these unusual cattle antibody VH CDR3 are not well understood. Here we analyze sequences of *Bos taurus* antibody VH domains and find that the subset with ultralong CDR3 exclusively uses a single variable gene, IGHV1-7 (VHBUL) rearranged to the longest diversity gene, IGHD8-2. An eight nucleotide duplication at the 3′ end of IGHV1-7 encodes a longer V-region producing an extended F β-strand that contributes to the stalk in a rearranged CDR3. A low amino acid variability was observed in CDR1 and CDR2, suggesting that antigen binding for this subset most likely only depends on the CDR3. Importantly a novel, potentially AID mediated, deletional diversification mechanism of the *B. taurus* VH ultralong CDR3 knob was discovered, in which interior codons of the IGHD8-2 region are removed while maintaining integral structural components of the knob and descending strand of the stalk in place. These deletions serve to further diversify cysteine positions, and thus disulfide bonded loops. Hence, both germline and somatic genetic factors and processes appear to be involved in diversification of this structurally unusual cattle VH ultralong CDR3 repertoire.
Cellular and Molecular Immunology advance online publication, 4 December 2017; doi:10.1038/cmi.2017.117

Keywords: bovine; complementarity determining region 3; repertoire diversification; deletions; immunoglobulin heavy chain

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 among 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

[1]Comparative Immunogenetics Laboratory, Department of Veterinary Pathobiology, College of Veterinary Medicine and Biomedical Sciences, Texas A&M University, College Station, TX 77843, USA; [2]Department of Molecular Medicine, The Scripps Research Institute, La Jolla, CA 92037, USA; [3]California Institute for Biomedical Research, La Jolla, CA 92037, USA; [4]Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA 92037, USA; [5]Department of Diagnostic Medicine and Pathobiology, College of Veterinary Medicine, Kansas State University, Manhattan, KS 66506, USA; [6]IMGT, the International ImMunoGeneTics information system, Laboratoire d'ImmunoGénétique Moléculaire, Institut de Génétique Humaine, CNRS, University of Montpellier, Montpellier 34396, France; [7]Department of Microbial Pathogenesis and Immunology, College of Medicine, Texas A&M University, Bryan, TX 77807, USA and [8]Fabrus, Inc., La Jolla, CA 92037, USA
Correspondence: Dr MF Criscitiello, PhD, Texas A&M University, Mailstop 4467, College Station, TX 77843, USA or Dr VV Smider, MD, PhD, Molecular Medicine, The Scripps Research Institute, 10550 N. Torrey Pines Rd., MEM 160 A, La Jolla, CA 92037, USA.
E-mail: mcriscitiello@cvm.tamu.edu or vvsmider@scripps.edu
Received: 15 August 2017; Revised: 27 September 2017; Accepted: 28 September 2017

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] Additionally, 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 (for example, 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.*[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 and 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.

MATERIALS AND METHODS

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, USA) and total RNA extraction was performed on the isolated PBMC with the RNeasy mini kit (Qiagen Valencia, CA, USA) as previously described.[40] Isolated RNA was used as the template for synthesis of 5′ RACE libraries with the GeneRacer kit (Invitrogen, Carlsbad, CA, USA) performed as previously described.[41] An equal mix of oligoDT and random hexamer primers was used to prime the reaction.

Initial amplification of IGHV1-7 rearranged transcripts

PCR amplification was performed on cDNA with primers designed to target the unique IGHV1-7 and the CH1 region of IGHM and IGHG (Supplementary Table 1) and cloned as previously described.[42] Briefly, the resulting product was visualized and extracted from an agarose gel using the PureLink Gel Extraction Kit (Life Technologies Carlsbad, CA, USA). 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 3 ml LB broth containing ampicillin for 16 h. Plasmids were isolated using the Zyppy Plasmid Miniprep Kit (ZYMO Irvine, CA, USA) according to the manufacturer's protocol. Plasmid inserts were freed with *Eco*RI (NEB, Ipswitch MA, USA) and resolved on an agarose gel. The BigDye terminator with M13 primers and BigDYE Xterminator (ThermoFisher, Waltham, MA, USA) 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.

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, USA), with the barcoded primers in Supplementary Table 1. The PCR protocol was performed in two steps with an initial denaturation of 2 min at 95 °C, cycles of 95 °C for 15 s followed by an annealing/extension of 1 min at 72 °C, and a final step of 5 min 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 demultiplex the samples. Finally the sequences were quality filtered ($Q > 20$) and homopolymer runs were corrected using the ACACIA program.[43] The cattle IGHM and IGHG sequences were visualized and aligned in the Geneious software suite.

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.

Shannon entropy and 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.[44,45] The resulting data was graphed in R using the 'ggplot2' package.[46] 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.

RESULTS

IGHV1-7 (VHBUL) 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 (Figure 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 (Figures 1a and 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 (Figure 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[28,47,48] (Figure 1c). Instead of this traditional motif, 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 (Figure 1c).[19]

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 (Supplementary Table 1). 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 (Figures 2a and b). 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 (Supplementary Figure 1). 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 approximately 4–13% range previously reported[15,28,49]) 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 (Figure 2a). 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 (Supplementary Figure 2). Although nearly all ultralong CDR3 transcripts utilized IGHV1-7, this gene was found in shorter CDR3 as well (Figure 2b, Supplementary Figure 3); 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 (Figure 1a) were not identified in any transcripts (Figure 2), and may not be utilized in the repertoire.

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 that 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

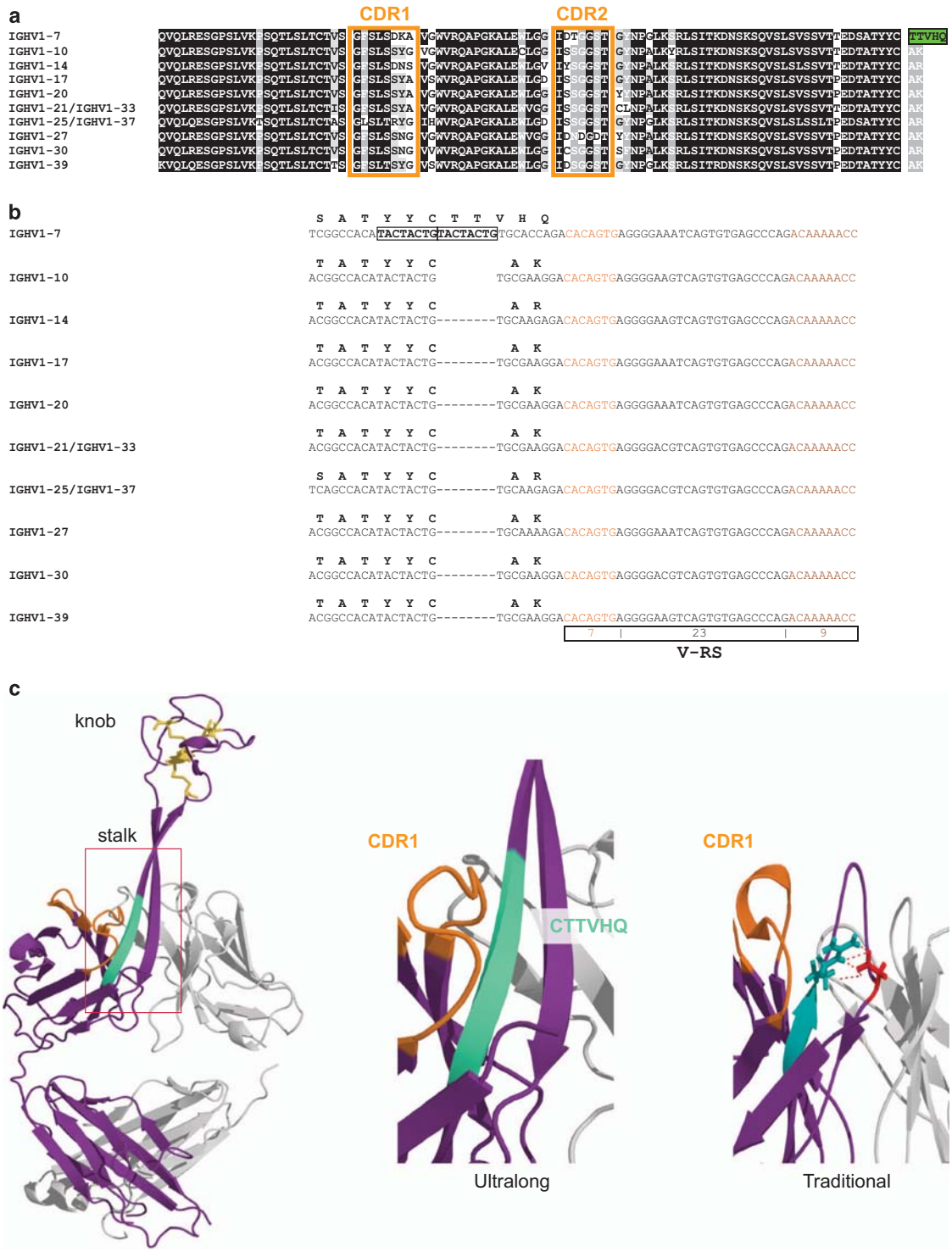


Figure 1 Genetic basis for the 'stalk' ascending β-strand in ultralong VH CDR3. (**a**) Amino-acid alignment of functional IGHV1 subgroup members. Amino acids shaded with Blosum62 similarity (black = 100%, dark grey = 80–100%, light grey = 60–80% and white = < 60%). Orange boxes denote CDR1 and CDR2 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.

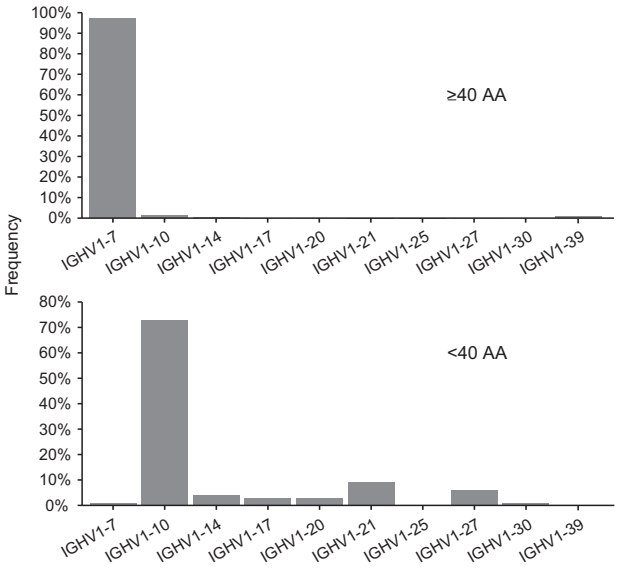


Figure 2 Ultralong VH CDR3 transcripts preferentially use one IGHV1 subgroup member IGHV1-7. (TOP) Percentage of IGHV1 genes expressed in transcripts with VH CDR3 equal to or greater than 40 AA. (BOTTOM) Percentage of IGHV1 genes expressed in transcripts with VH CDR3 less than 40 AA. IGHV1-21 and IGHV1-33, and IGHV1-25 and IGHV1-37, are identical, therefore only IGHV1-21 and IGHV1-25 are labeled.

somatic hypermutation or exonuclease activity removing the CTTVHQ-like motif during V to D-J recombination.

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] (Figure 3a). Surprisingly, a total of 426 out of the 894 ultralong sequences (47.6%) had in-frame nucleotide deletions. The deletions (colored lines in Figure 3a, red in Supplementary Figure 3a) range from 1 to 18 interior codons (Figures 3a–d) and retain the aforementioned motifs with high sequence homology to the 5′ and 3′ end of the germline IGHD8-2 (Figures 3a and 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 (Figures 3a and 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.[39,50] 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.[50] In this regard, 96.9% of the deletions overlapped an AID hotspot (Figures 3a and b). 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 (Figure 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 (Figures 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.

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 (Figure 4, Supplementary Figure 5).[51] 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 1). This decrease

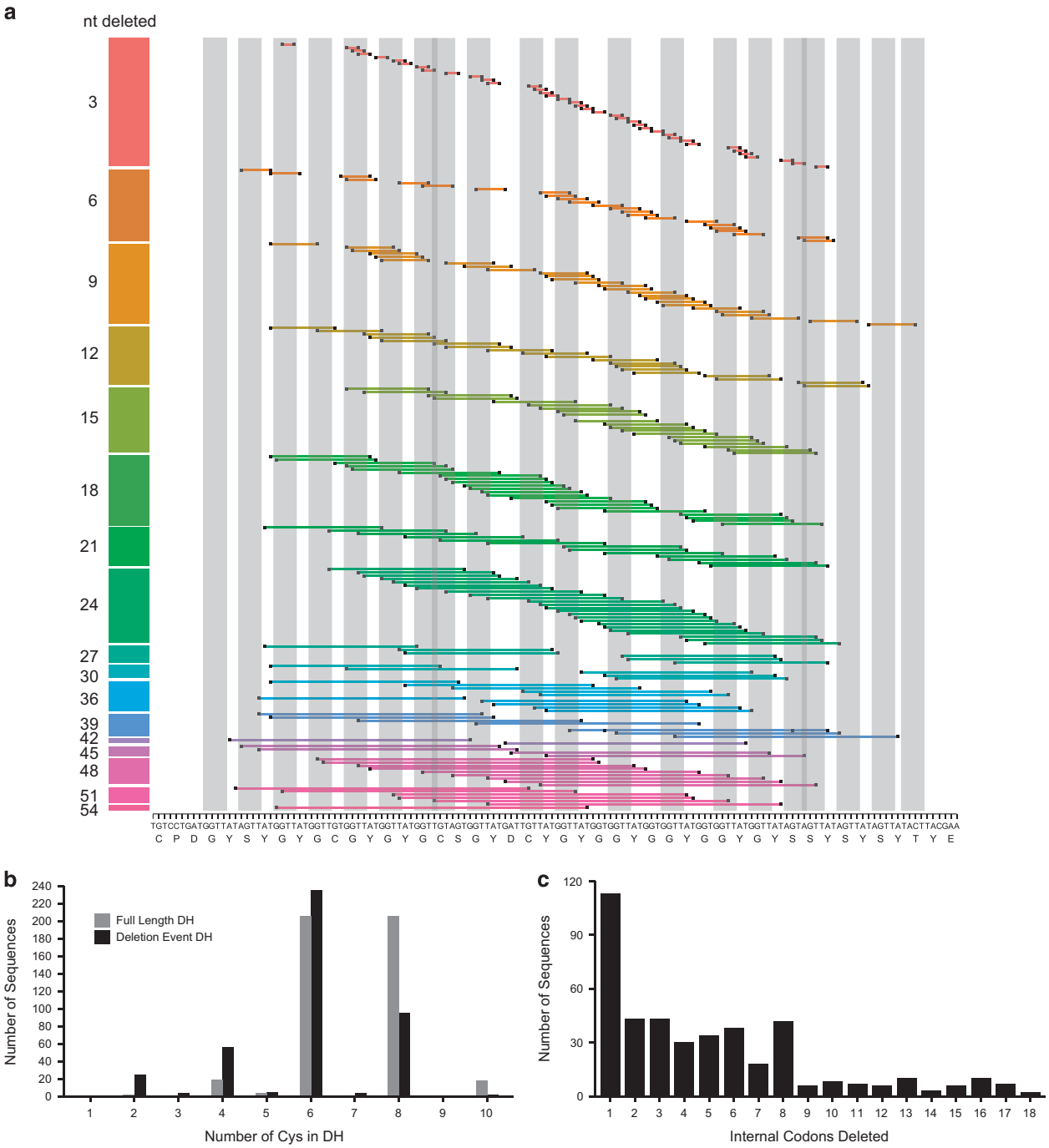


Figure 3 Ultralong VH CDR3 are characterized by frequent deletions in IGHD8-2 region. (**a**) Graph showing the gapped/deleted positions from alignments with the full length IGHD8-2 gene. The colored lines are indicative of gapped/deleted nucleotides, with the black dots indicating the deletion boundaries. 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.

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.

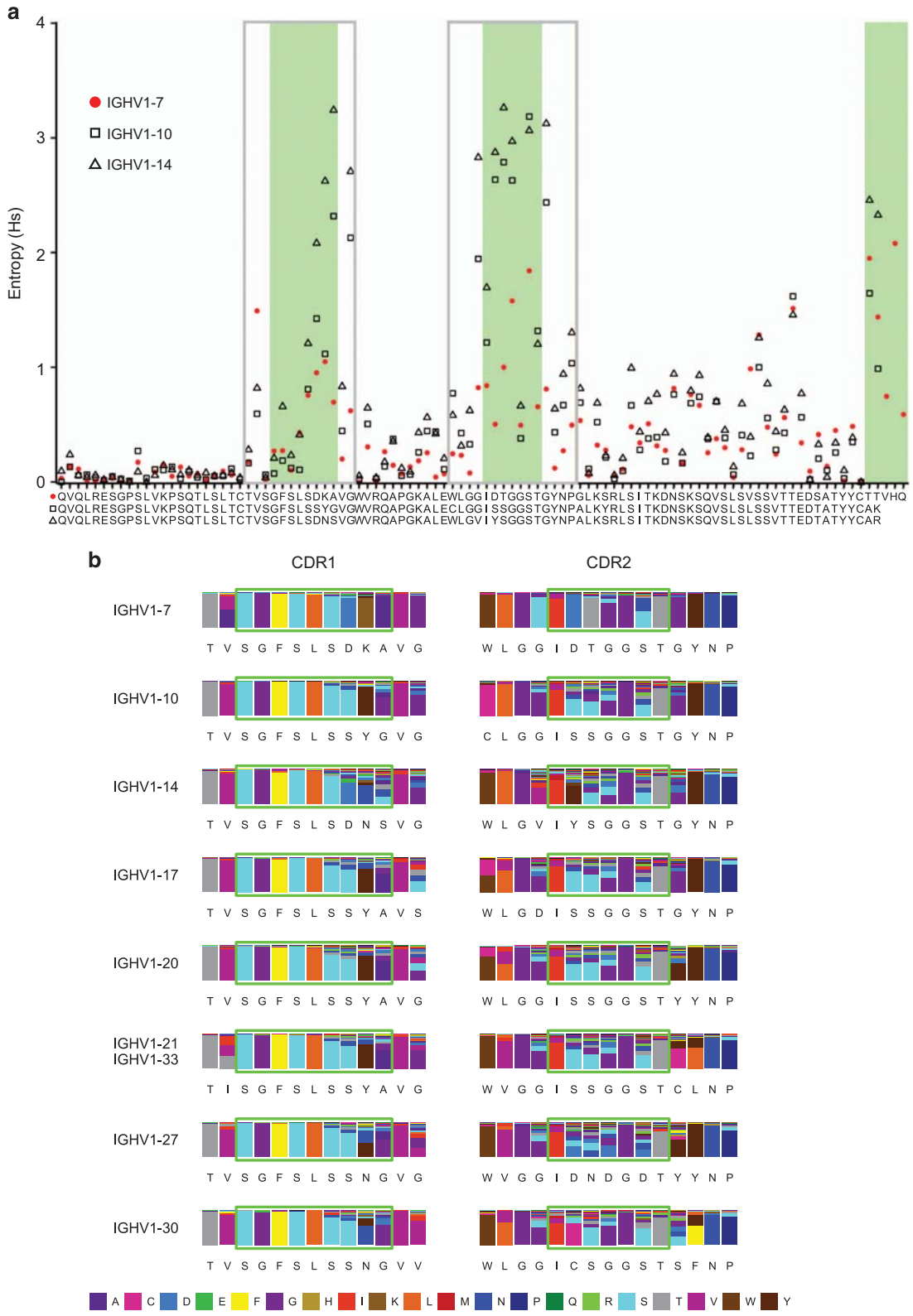


Figure 4 IGHV1-7 has low variability in CDR1 and CDR2. (**a**) Graph of Shannon entropy values for IGHV1-7 (red circles), IGHV1-10 (triangles), and IGHV1-14 (squares). IGHV1-7 was selected as it encoded the majority of ultralong VH CDR3. IGHV1-14 and IGHV1-10 were selected for comparison as they encoded 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 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.

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.[52] 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 eight-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 (Figure 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 (≥40 AA). 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 (Figure 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 (Figure 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.[19,28–31,49,53,54] Of these 12 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 (Figure 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 (Figure 4, Supplementary 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,[55–62] 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.[49] 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%, Supplementary Figure 6), suggests that these events could be attributed to strand slippage events commonly associated with AID activity.[29,30,49,63]

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).[63] Unsuccessful CSR events, known as resection events, resulting in smaller genomic deletions within a switch region, are documented to occur.[50] The mutational load observed in the VH domain is clear evidence for high levels of AID activity within bovine B cells (Table 1). The deletion events observed in IGHD8-2 could be mechanistically similar to the resection events of a failed class switch,[50] which would allow AID and associated machinery to produce double strand DNA breaks within IGHD8-2 containing CDR3.[39,64] Recently, Yeap, *et al.*[55] reported deletions in V-region transgenes as a result of double-strand breaks during SHM. 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 recombination events.

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 at 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 eight-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.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

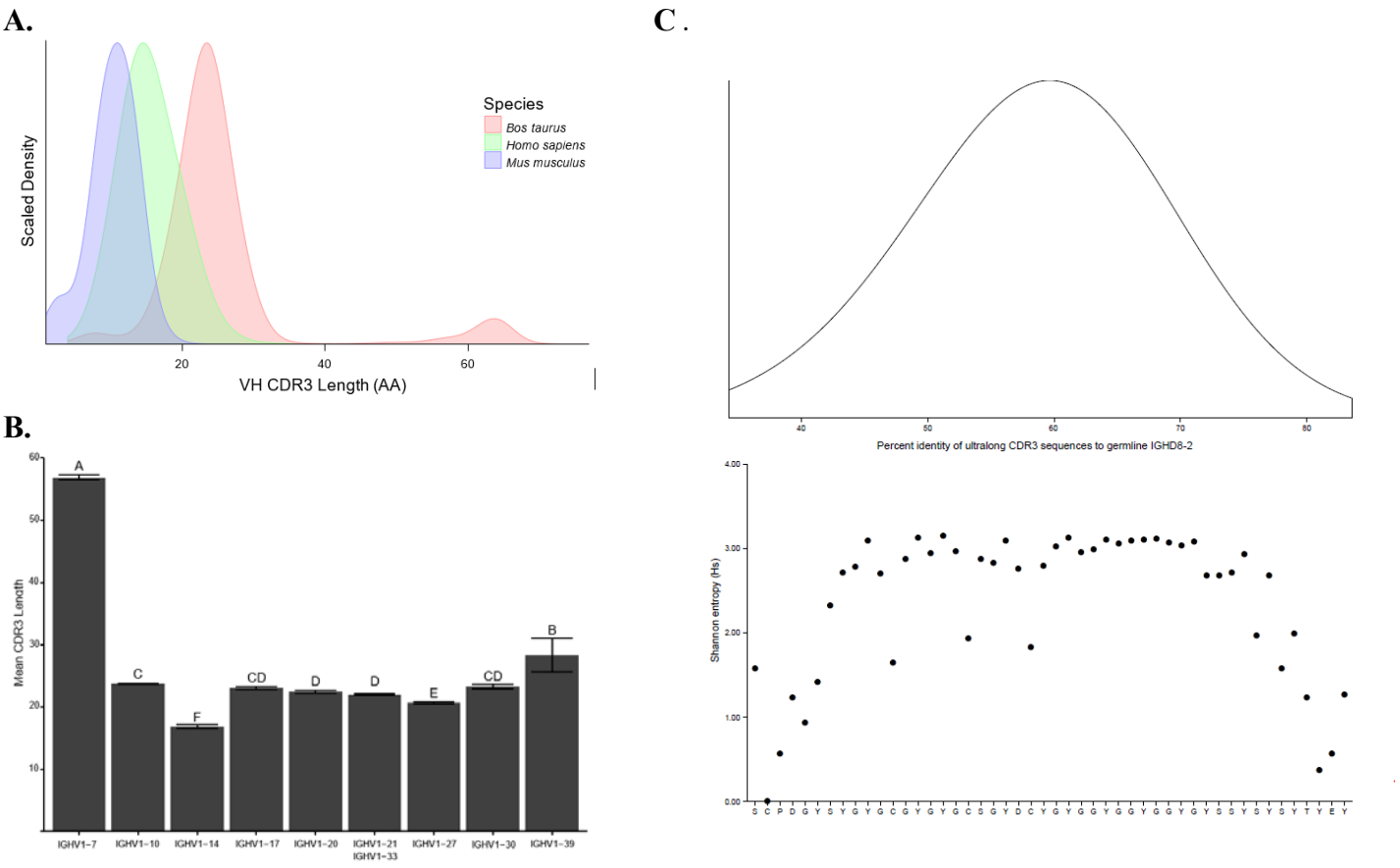
ACKNOWLEDGEMENTS

This work was supported by NIH grant R01 GM105826-01 to VVS, R21 AI120791 to VVS, WM and MFC and NSF grant #IOS1257829 to MFC. AT is supported by Scripps Genomic Medicine, an NIH-NCATS Clinical and Translational Science Award (CTSA; 5 UL1 RR025774).

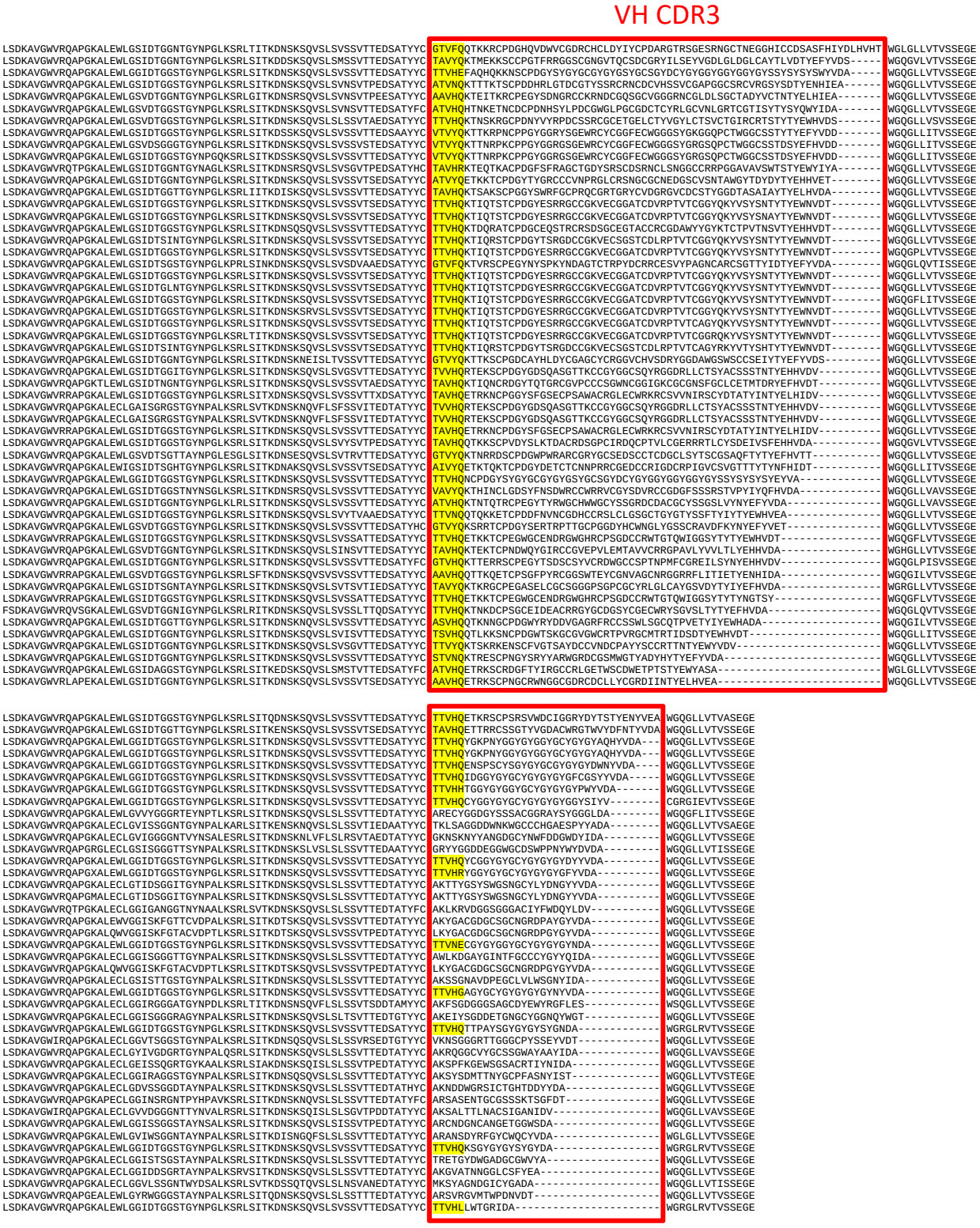
1 Sok D, Le KM, Vadnais M, Saye-Francisco KL, Jardine JG, Torres JL *et al.* Rapid elicitation of broadly neutralizing antibodies to HIV by immunization in cows. *Nature* 2017; **548**: 108–111.
2 Lefranc MP. Immunoglobulin and T Cell Receptor Genes: IMGT((R)) and the Birth and Rise of Immunoinformatics. *Front Immunol* 2014; **5**: 22.
3 Lefranc M-P, Lefranc G. *The Immunoglobulin FactsBook*. Academic Press: London, UK. 2001, p 458.
4 Lefranc M-P, Lefranc G. *The T cell receptor FactsBook*. Academic Press: London, UK. 2001, p 398.
5 Rock EP, Sibbald PR, Davis MM, Chien YH. CDR3 length in antigen-specific immune receptors. *J Exp Med* 1994; **179**: 323–328.
6 Morea V, Tramontano A, Rustici M, Chothia C, Lesk AM. Conformations of the third hypervariable region in the VH domain of immunoglobulins. *J Mol Biol* 1998; **275**: 269–294.
7 Ivanov II, Schelonka RL, Zhuang Y, Gartland GL, Zemlin M, Schroeder HW 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**: 7773–7780.
8 Lefranc MP, Pommie C, Ruiz M, Giudicelli V, Foulquier E, Truong L *et al.* IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev Comp Immunol* 2003; **27**: 55–77.
9 Hamers-Casterman C, Atarhouch T, Muyldermans S, Robinson G, Hamers C, Songa EB *et al.* Naturally occurring antibodies devoid of light chains. *Nature* 1993; **363**: 446–448.
10 Stanfield RL, Dooley H, Verdino P, Flajnik MF, Wilson IA. Maturation of shark single-domain (IgNAR) antibodies: evidence for induced-fit binding. *J Mol Biol* 2007; **367**: 358–372.
11 Stanfield RL, Dooley H, Flajnik MF, Wilson IA. Crystal structure of a shark single-domain antibody V region in complex with lysozyme. *Science (New York, NY)* 2004; **305**: 1770–1773.
12 de Los Rios M, Criscitiello MF, Smider VV. Structural and genetic diversity in antibody repertoires from diverse species. *Curr Opin Struct Biol* 2015; **33**: 27–41.
13 Rasmussen SG, Choi HJ, Fung JJ, Pardon E, Casarosa P, Chae PS *et al.* Structure of a nanobody-stabilized active state of the beta(2) adrenoceptor. *Nature* 2011; **469**: 175–180.
14 Berens SJ, Wylie DE, Lopez OJ. Use of a single VH family and long CDR3s in the variable region of cattle Ig heavy chains. *Int Immunol* 1997; **9**: 189–199.
15 Saini SS, Allore B, Jacobs RM, Kaushik A. Exceptionally long CDR3H region with multiple cysteine residues in functional bovine IgM antibodies. *Eur J Immunol* 1999; **29**: 2420–2426.
16 Saini SS, Farrugia W, Ramsland PA, Kaushik AK. 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**: 845–853.
17 Saini SS, Kaushik A. Extensive CDR3H length heterogeneity exists in bovine foetal VDJ rearrangements. *Scand J Immunol* 2002; **55**: 140–148.
18 Stanfield RL, Wilson IA, Smider VV. Conservation and diversity in the ultralong third heavy-chain complementarity-determining region of bovine antibodies. *Sci Immunol* 2016; **1**: aaf7962.
19 Wang F, Ekiert DC, Ahmad I, Yu W, Zhang Y, Bazirgan O *et al.* Reshaping antibody diversity. *Cell* 2013; **153**: 1379–1393.
20 Schroeder Jr HW, Hillson JL, Perlmutter RM. Structure and evolution of mammalian VH families. *Int Immunol* 1990; **2**: 41–50.
21 Schatz DG, Oettinger MA, Baltimore D. The V(D)J recombination activating gene, RAG-1. *Cell* 1989; **59**: 1035–1048.
22 Tonegawa S. Somatic generation of antibody diversity. *Nature* 1983; **302**: 575–581.
23 Ruiz M, Pallarès N, Contet V, Barbié V, Lefranc MP. The Human Immunoglobulin Heavy Diversity (IGHD) and Joining (IGHJ) Segments. *Exp Clin Immunogenet* 1999; **16**: 173–184.
24 Matsuda F, Ishii K, Bourvagnet P, Kuma K, Hayashida H, Miyata T *et al.* The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus. *J Exp Med* 1998; **188**: 2151–2162.
25 Kabat EA. Unique features of the variable regions of Bence Jones proteins and their possible relation to antibody complementarity. *Proceedings of the National Academy of Sciences of the United States of AmericaProc Natl Acad Sci USA* 1968; **59**: 613–619.
26 Wu TT, Kabat EA. 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**: 211–250.
27 Muramatsu M, Kinoshita K, Fagarasan S, Yamada S, Shinkai S, Honjo T. Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell* 2000; **102**: 553–563.
28 Ma L, Qin T, Chu D, Cheng X, Wang J, Wang X *et al.* Internal duplications of DH, JH, and C region genes create an unusual IgH gene locus in cattle. *J Immunol* 2016; **196**: 4358–4366.
29 Liljavirta J, Ekman A, Knight JS, Pernthaner A, Iivanainen A, Niku M. 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**: 942–949.
30 Verma S, Aitken R. Somatic hypermutation leads to diversification of the heavy chain immunoglobulin repertoire in cattle. *Vet Immunol Immunopathol* 2012; **145**: 14–22.
31 Sun Y, Liu Z, Ren L, Wei Z, Wang P, Li N *et al.* Immunoglobulin genes and diversity: what we have learned from domestic animals. *J Anim Sci Biotechnol* 2012; **3**: 18.
32 Kozuka Y, Nasu T, Murakami T, Yasuda M. 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**: 190–197.
33 Ekman A, Pessa-Morikawa T, Liljavirta J, Niku M, Iivanainen A. B-cell development in bovine fetuses proceeds via a pre-B like cell in bone marrow and lymph nodes. *Dev Comp Immunol* 2010; **34**: 896–903.
34 Kaushik AK, Kehrli ME Jr., Kurtz A, Ng S, Koti M, Shojaei F *et al.* Somatic hypermutations and isotype restricted exceptionally long CDR3H contribute to antibody diversification in cattle. *Vet Immunol Immunopathol* 2009; **127**: 106–113.
35 Yasuda M, Jenne CN, Kennedy LJ, Reynolds JD. The sheep and cattle Peyer's patch as a site of B-cell development. *Vet Res* 2006; **37**: 401–415.
36 Neill JD, Ridpath JF, Liebler-Tenorio E. Global gene expression profiling of Bovine immature B cells using serial analysis of gene expression. *Anim Biotechnol* 2006; **17**: 21–31.
37 Koti M, Kataeva G, Kaushik A. Organization of DH-gene locus is distinct in cattle. *Dev Biol* 2008; **132**: 307–313.
38 Koti M, Kataeva G, Kaushik AK. Novel atypical nucleotide insertions specifically at VH-DH junction generate exceptionally long CDR3H in cattle antibodies. *Mol Immunol* 2010; **47**: 2119–2128.
39 Rogozin IB, Diaz M. 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**: 3382–3384.

40 Criscitiello MF, Ohta Y, Graham MD, Eubanks JO, Chen PL, Flajnik MF. Shark class II invariant chain reveals ancient conserved relationships with cathepsins and MHC class II. *Dev Comp Immunol* 2012; **36**: 521–533.
41 Mashoof S, Pohlenz C, Chen PL, Deiss TC, Gatlin D 3rd, Buentello A *et al.* Expressed IgH mu and tau transcripts share diversity segment in ranched Thunnus orientalis. *Dev Comp Immunol* 2014; **43**: 76–86.
42 Breaux B, Deiss TC, Chen PL, Cruz-Schneider MP, Sena L, Hunter ME *et al.* 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**: 57–68.
43 Bragg L, Stone G, Imelfort M, Hugenholtz P, Tyson GW. Fast, accurate error-correction of amplicon pyrosequences using Acacia. *Nat Methods* 2012; **9**: 425–426.
44 Grant BJ, Rodrigues AP, ElSawy KM, McCammon JA, Caves LS. Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics (Oxford, England)* 2006; **22**: 2695–2696.
45 R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing: Vienna, Austria. 2014.
46 Wickham H. *ggplot2: elegant graphics for data analysis*. Springer: New York. 2009.
47 Hosseini A, Campbell G, Prorocic M, Aitken R. Duplicated copies of the bovine JH locus contribute to the Ig repertoire. *Int Immunol* 2004; **16**: 843–852.
48 Koti M, Kataeva G, Kaushik AK. Organization of D(H)-gene locus is distinct in cattle. *Dev Biol* 2008; **132**: 307–313.
49 Liljavirta J, Niku M, Pessa-Morikawa T, Ekman A, Iivanainen A. Expansion of the preimmune antibody repertoire by junctional diversity in Bos taurus. *PLoS ONE* 2014; **9**: e99808.
50 Dong J, Panchakshari RA, Zhang T, Zhang Y, Hu J, Volpi SA *et al.* Orientation-specific joining of AID-initiated DNA breaks promotes antibody class switching. *Nature* 2015; **525**: 134–139.
51 Wu TT, Kabat EA. 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**: 211–250.
52 Sok D, Le KM, Vadnais M, Saye-Francisco K, Jardine JG, Torres J *et al.* Rapid elicitation of broadly neutralizing antibodies to HIV by immunization in cows. *Nature* 2017; **548**: 108–111 advance online publication.
53 Niku M, Liljavirta J, Durkin K, Schroderus E, Iivanainen A. The bovine genomic DNA sequence data reveal three IGHV subgroups, only one of which is functionally expressed. *Dev Comp Immunol* 2012; **37**: 457–461.
54 Butler JE. Immunoglobulin diversity, B-cell and antibody repertoire development in large farm animals. *Rev Sci Tech* 1998; **17**: 43–70.
55 Yeap LS, Hwang JK, Du Z, Meyers RM, Meng FL, Jakubauskaite A *et al.* Sequence-intrinsic mechanisms that target AID mutational outcomes on antibody genes. *Cell* 2015; **163**: 1124–1137.
56 Meissner F, Mann M. Quantitative shotgun proteomics: considerations for a high-quality workflow in immunology. *Nat Immunol* 2014; **15**: 112–117.
57 Kepler TB, Liao HX, Alam SM, Bhaskarabhatla R, Zhang R, Yandava C *et al.* Immunoglobulin gene insertions and deletions in the affinity maturation of HIV-1 broadly reactive neutralizing antibodies. *Cell Host Microbe* 2014; **16**: 304–313.
58 Briney BS, Willis JR, Crowe JE Jr. Location and length distribution of somatic hypermutation-associated DNA insertions and deletions reveals regions of antibody structural plasticity. *Genes Immunity* 2012; **13**: 523–529.
59 Reason DC, Zhou J. Codon insertion and deletion functions as a somatic diversification mechanism in human antibody repertoires. *Biology DirectBiol Direct* 2006; **1**: 24.
60 Wilson PC, de Bouteiller O, Liu YJ, Potter K, Banchereau J, Capra JD *et al.* Somatic hypermutation introduces insertions and deletions into immunoglobulin V genes. *J Exp Med* 1998; **187**: 59–70.
61 Wilson P, Liu YJ, Banchereau J, Capra JD, Pascual V. Amino acid insertions and deletions contribute to diversify the human Ig repertoire. *Immunol Rev* 1998; **162**: 143–151.
62 Goossens T, Klein U, Kuppers R. Frequent occurrence of deletions and duplications during somatic hypermutation: implications for oncogene translocations and heavy chain disease. *Proceedings of the National Academy of Sciences of the United States of AmericaProc Natl Acad Sci USA* 1998; **95**: 2463–2468.
63 Criscitiello MF, Benedetto R, Antao A, Wilson MR, Chinchar VG, Miller NW *et al.* Beta 2-microglobulin of ictalurid catfishes. *Immunogenetics* 1998; **48**: 339–343.
64 Yeap L-S, Hwang Joyce K, Du Z, Meyers Robin M, Meng F-L, Jakubauskaitė A *et al.* Sequence-intrinsic mechanisms that target AID mutational outcomes on antibody genes. *Cell* 2015; **163**: 1124–1137.

Supplementary Information for this article can be found on the *Cellular & Molecular Immunol* website (http://www.nature.com/cmi)

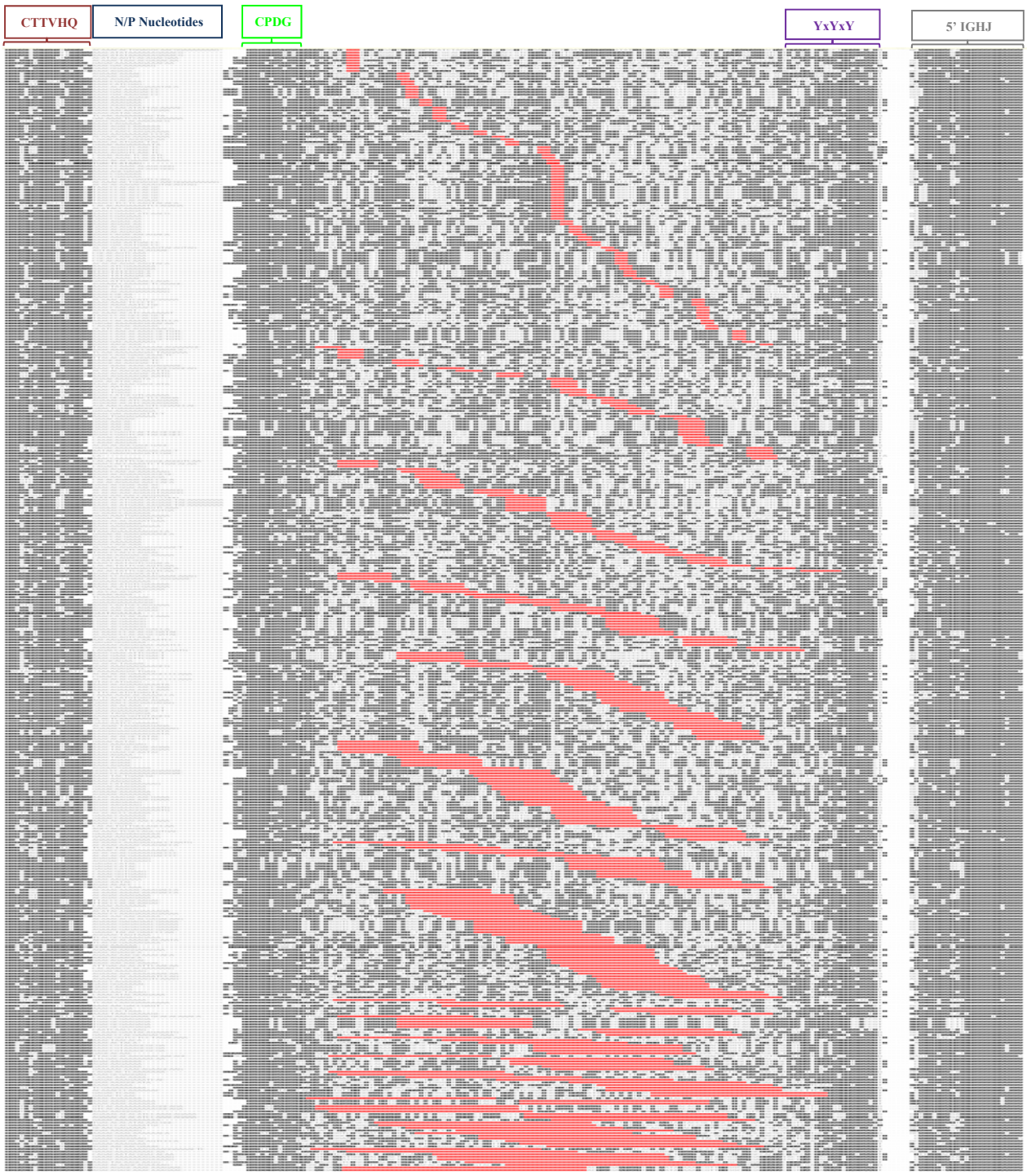


Supplemental Figure 1. A. Comparison of variable heavy (VH) chain CDR3 length (AA) distribution for *Bos taurus*, *Homo sapiens* and *Mus musculus*. The ultralong CDR3 (≥40 AA) are the source for the bimodal distribution of the *B. taurus* CDR3. **B.** Comparison of the mean variable heavy (VH) CDR3 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).

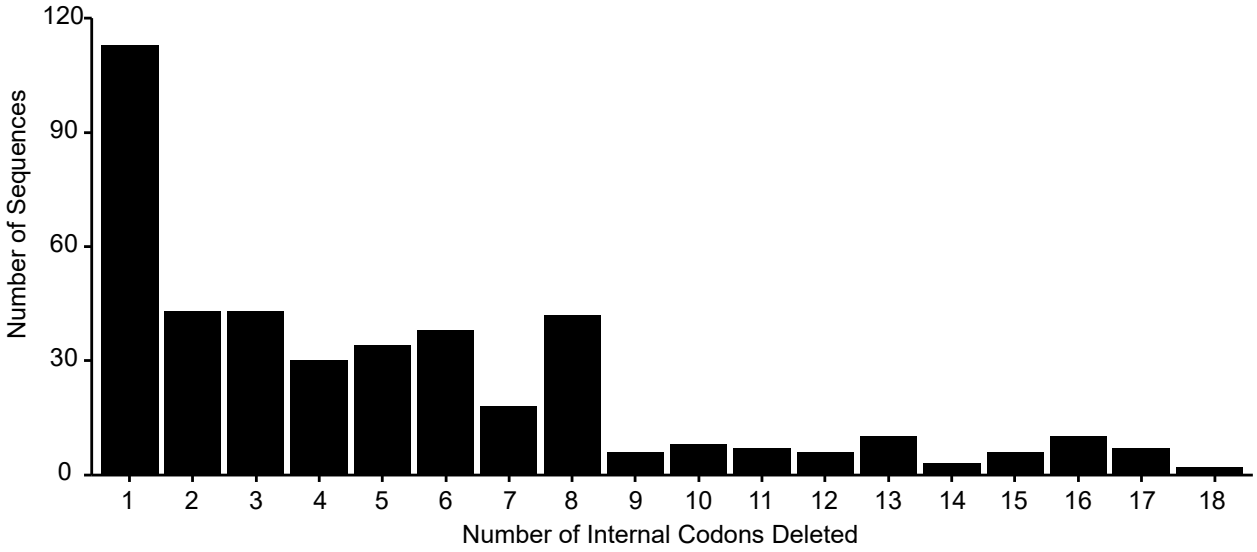


Supplemental Figure 2: Amino acid alignment of *Bos taurus* targeted *IGHV1-7* sequences. The variable heavy (VH) CDR3 is boxed in red with sequences encoding ultralong CDR3 (≥40AA) on top and conventional CDR3 (<40AA) on bottom. The *IGHV1-7* specific "TTVHQ" motif is highlighted in yellow.

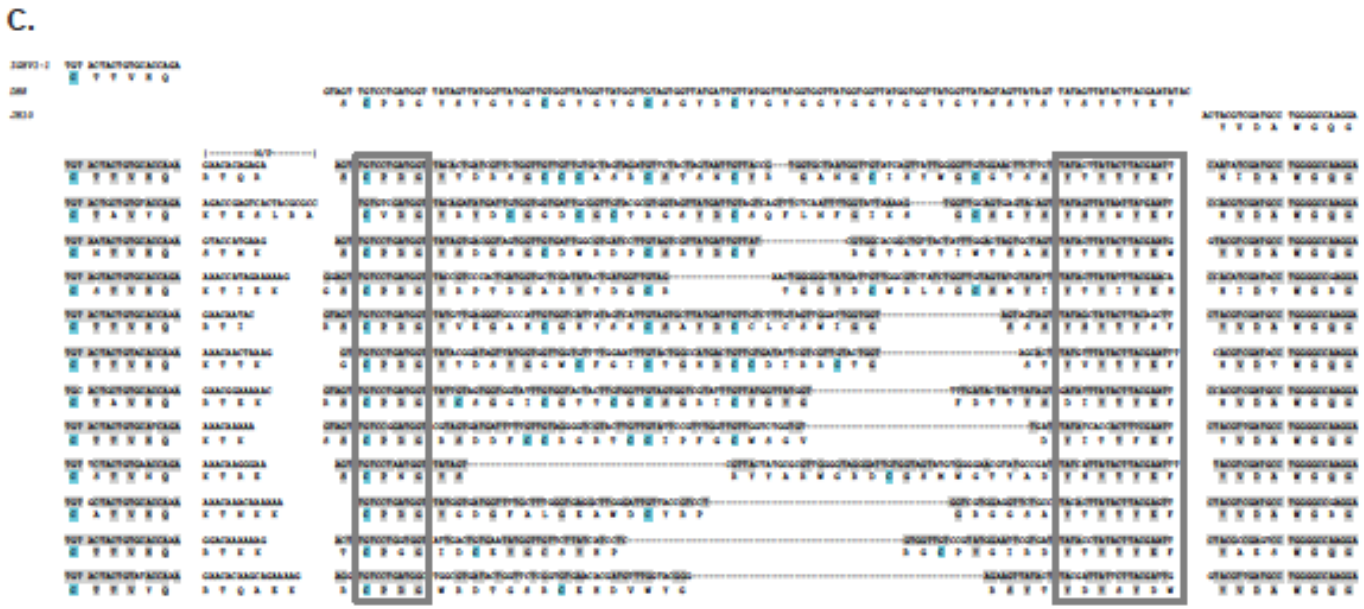
A.



B.

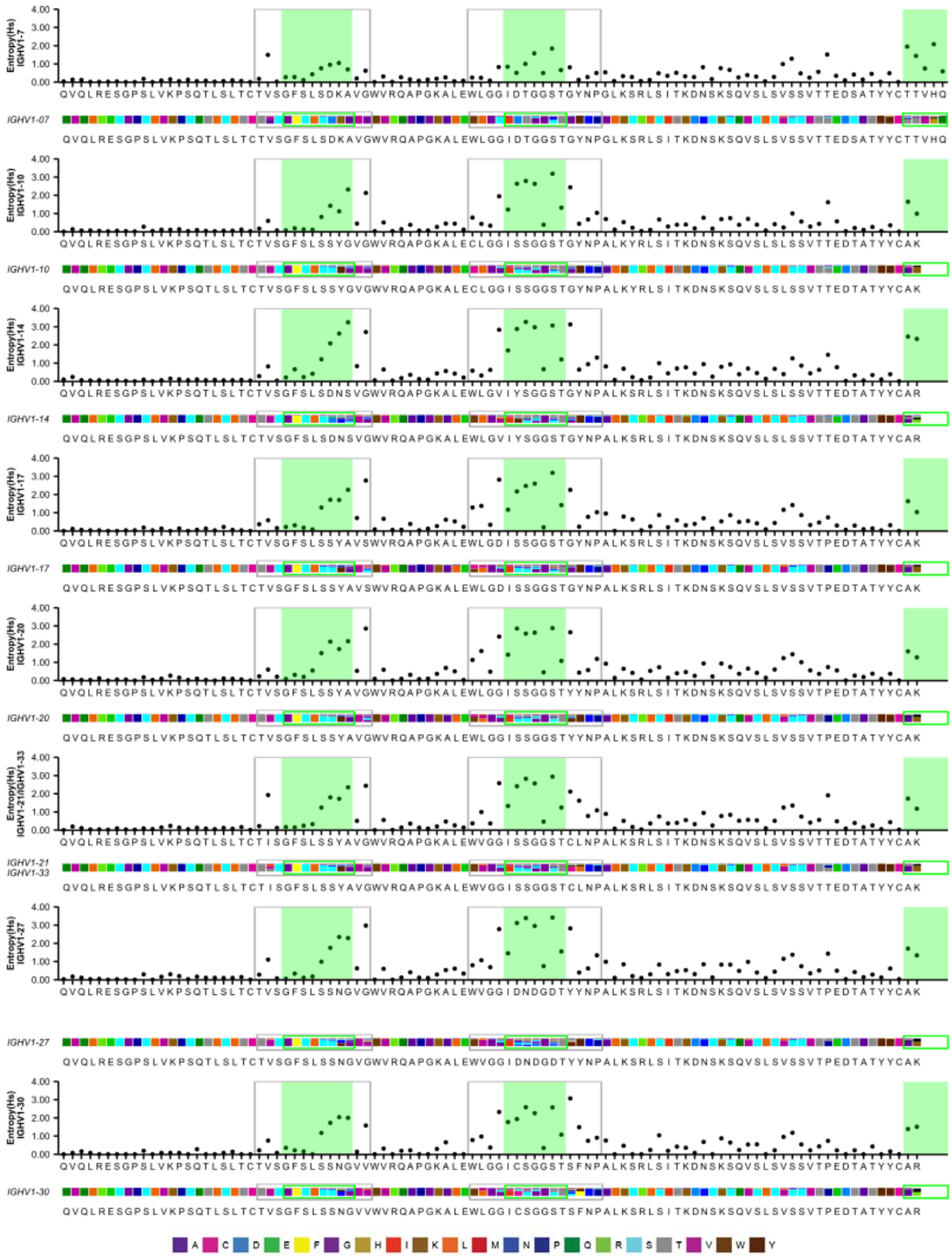


C.



Supplemental Figure 3: 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. **B.** 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. **C.** Alignment of several ultralong VH CDR3 transcripts bearing deletion events. IGHV, IGHD, IGHJ and non-templated N nucleotides boundaries are indicated at the top of the alignment. Cysteines are highlighted in Cyan. Light grey highlighting indicates sequence identity with germline genes. The grey boxes are the

conserved 5’ and 3’ ends of the IGHD28-2 gene, encoding the “CPDG” and “YxYxY” motifs respectively, conserved in CDR3 sequences.



Supplemental Figure 4: Shannon entropy plots and amino acid (AA) frequency charts for individual IGHV1 subgroup genes. The CDR are indicated with a green box.

Supplementary Table 1.

Primer	For/Rev	Region	Sequence	Priming Site
IGHV1-7 variable region	F	CDR1	5'-TTGAGCGACAAGGCTGTAGGCTG-3'	LSDKAVG
IGHM constant region	R	IGHM CH1	5'-ACGCAGGACACCAGGGGGAAG-3'	FPLVSC
PacBio Barcode Primers			(Barcode in red)	
Gene Racer Oligo	F	Barcode+5'Oligo	5'-TCAGACGATGCGTCATGGACACTGACATGGACTGAAGGAGTA-3'	
IGHM constant region	R	Barcode+IGHM CH1	5'-AGTCATCGTATCGCGCACGCAGGACACCAGGGGGAAG-3'	FPLVSC
IGHM constant region	R	Barcode+IGHM CH1	5'-CGATCAGCTGAGCGCGACGCAGGACACCAGGGGGAAG-3'	FPLVSC
IGHM constant region	R	Barcode+IGHM CH1	5'-CATGTACTGATACACAGTGAAGACTCTCGGGTGTGATTCAC-3'	ESHPRVF
IGHM constant region	R	Barcode+IGHM CH1	5'-AGTGTGTCATGCGTGTGTGAAGACTCTCGGGTGTGATTCAC-3'	ESHPRVF
IGHG constant region	R	Barcode+IGHG CH1	5'-TCTGTAGTGCGTGCGCCTTTCGGGGCTGTGGTGGAGGC-3'	ASTTAPK
IGHG constant region	R	Barcode+IGHG CH1	5'-GTCGCGACGTCAGTGTCTTTCGGGGCTGTGGTGGAGGC-3'	ASTTAPK
IGHG constant region	R	Barcode+IGHG CH1	5'-GCAGAGTCATGTATAGCTTTCGGGGCTGTGGTGGAGGC-3'	ASTTAPK
IGHG constant region	R	Barcode+IGHG CH1	5'-GAGTCGTACTCTAGTACTTTCGGGGCTGTGGTGGAGGC-3'	ASTTAPK