Conservation and diversity in the ultralong third heavy-chain complementarity-determining region of bovine antibodies

Robyn L. Stanfield,1 Ian A. Wilson,1,2* Vaughn V. Smider3,4*

Antibodies provide a broad defense against a vast array of antigens; however, the structural features that contribute to this diverse antigen recognition vary in different vertebrates. In cows, a subset of antibodies have an exceptionally long third heavy-chain complementarity-determining region (CDR H3) that is highly variable in sequence and includes multiple cysteines. These long CDR H3s (up to 69 residues) fold into a long stalk atop which sits a knob domain that is located far from the antibody surface. We have determined crystal structures of three bovine Fabs to decipher the conserved and variable features of ultralong CDR H3s that lead to diversity in antigen recognition. Despite high sequence variability, the stalks adopt a conserved β-ribbon structure, whereas the knob regions share a conserved β sheet that serves as a scaffold for two connecting loops of variable length and conformation, as well as one conserved disulfide. Variation in patterns and connectivity of the remaining disulfides contribute to the structural diversity of the knob. The unusual architecture of these ultralong bovine CDR H3s for generating diversity is unique in adaptive immune systems and may inform efforts in antibody engineering.

INTRODUCTION
The human adaptive immune system uses V(D)J recombination to construct germline antibodies that are then somatically mutated in response to antigen challenge. The classic scaffold for generating such diversity is the immunoglobulin (Ig) fold. Antibodies are assembled from heavy and light chains, where diversity is mainly found in the complementarity-determining region (CDR) loops, which connect the core β strands of the Ig fold of the heavy and light variable regions and form the antigen binding surface. The third heavy-chain CDR (CDR H3) is normally the most variable CDR in both sequence and length because it is encoded by V, D, and J gene recombination, where combinatorial diversity can produce more than 10^6 to 10^10 germline antibodies.

About 15 years ago, 10% of cow antibodies were discovered to have "exceptionally long" CDR H3s, with lengths of up to 69 residues (1), whereas more conventional cow antibodies have CDR H3s of around 23 residues, which also are quite long (we also use the term "ultralong" CDR to correspond to the original genetic description of exceptionally long CDRs). The ultralong CDR H3s contain a large—usually even—number of cysteine residues (from 2 to 12 residues) (1, 2). The longest CDR H3s found to date in the human repertoire contain up to 35 residues (3), whereas camelid single-chain antibodies have up to 24 residues (4) and shark immunoglobulin new antigen receptor (IgNAR) antibodies have up to 27 residues (5). Although CDR H3 cysteines are rare in humans and mice, they are found more extensively in sharks and camels, and tether the long CDR H3s to the framework region or to a neighboring CDR. Cysteines are also commonly found in the CDR H3s of chickens (6), pigs (7), and duck-billed platypus (8).

The ultralong CDR H3s are found in bovine Igs of all classes (9), arise early in fetal development (10), and are encoded by the V_{μ}BUL, D_{μ}2, and J_{μ}1 gene segments (2, 11, 12). Their length is due to a remarkably long germline D_{μ}2 segment that encodes four cysteines and repeated glycine, serine, and tyrosine residues, whose codons are biased for mutation to cysteine with just one base change (Fig. 1). Furthermore, CDR H3 size can be increased by N and P additions and "conserved short nucleotide sequence" insertions at the V-D junction (13). The gene segments also contain a large number of RGYW hotspots, which are recognition sites for activation-induced cytidine deaminase, which produces somatic hypermutation (SH).

Thus, the repertoire of most ultralong CDR H3 antibodies appear to derive from a single germline rearrangement, with diversity arising mainly at the V-D and D-J junctions (through N and P additions and deletions) and through extensive SH. Because only a single framework is used for both the heavy-chain and the light-chain variable regions, only CDR H3 appears to be significantly diversified in the repertoire. Thus, this bovine system differs from that of classical (human or mouse) antibodies in two respects: (i) the diversity mechanism used to create the repertoire and (ii) the CDR H3 structural scaffold for binding antigen. The first two crystal structures for bovine antibodies with ultralong CDR H3s (2) revealed that the CDR H3s fold into an extended β-ribbon "stalk" topped by a "knob" domain containing different combinations of cysteine residues that pair to form disulfides. These structures, along with deep sequencing data, suggested that the cow repertoire uses different disulfide patterns, generated through SH, in the knob region of CDR H3 as the basis for creating structural diversity for antigen binding. Here, we determined three further crystal structures of bovine ultralong CDR H3 antibodies and, with this larger data set, elucidated the key conserved features that shed insight into how a remarkably structurally diverse repertoire can be created from limited genetic information.

RESULTS
Overall structure and sequence
Bovine heavy-chain IgG variable region cDNA was obtained from a cow immunized with the bovine viral diarrhea virus (2). From 133
available heavy-chain sequences, several were chosen for expression as Fab fragments, and three of these Fabs crystallized in forms suitable for x-ray diffraction studies (Table 1). Crystal structures were determined for bovine Fabs B11 (2.06 Å), E03 (2.20 Å), and A01 (2.00 Å). Fab B11 has eight cysteines within its 63-residue CDR H3, which is the longest CDR H3 structure determined to date; Fab E03 has one of the shortest ultralong CDR H3s (44 residues with only two cysteines); and Fab A01 has a 61-residue H3 with six cysteines, but not the first highly conserved cysteine residue in the knob domain (Figs. 1 and 2). The CDR H3s of these five Fabs are highly dissimilar from one another in terms of amino acid sequence (Fig. 1), with less than 35% identity between the sequences. Thus, these Fabs, along with the previously crystallized BLV1H12 and BLV5B8, display a range of features (for example, CDR H3 length, sequence diversity, and disulfide content and position) that provide insights into the structural and functional diversity of the ultralong CDR H3 repertoire, as well as conserved motifs and scaffolds.

As with Fabs BLV1H12 and BLV5B8 [Protein Data Bank (PDB) codes 4k3d and 4k3e (2)], all three of the newly characterized bovine Fabs have a long β-ribbon stalk supporting a compact knob domain, with the distal tip of the knob rising far above the tips of the other CDR loops. Alignment of these three Fabs with the two previously determined bovine Fab structures reveals little overall variation or movement of the stalk region (Fig. 2) and surrounding CDRs, suggesting that the long β-hairpin is quite stable. When variable heavy-chain framework residues are superimposed and the rotation necessary to superimpose the CDR H3 stalks (which were not included in the first superposition) is measured, we see only small differences (3.5°, 5.0°, 5.5°, and 10.7° for BLV5B8, BLV1H12, C07, and E03, respectively) with respect to B11 (Fig. 3A). However, the knob can adopt different relative positions and orientations with respect to its stalk. When only the knob domains are superimposed, the rotations required to superimpose the stalk domains are quite large (23.0°, 31.5°, 57.1°, and 57.3° for BLV5B8, BLV1H12, C07, and E03, respectively) when compared to B11 as a reference (Fig. 3B).

The overall lengths of the CDR H3 also differ slightly and contribute to variation in the distance of the knob domain to the Fab surface. We approximated the CDR H3 vertical height by measuring from the Cα atom of residue H98, which is roughly colinear with the tips of the neighboring CDR loops, to the most distal Cα atom of the knob domain. BLV1H12 CDR H3 extends the farthest, by around 41 Å, whereas A01 extends only by 32 Å. BLV5B8, B11, and E03 all have similar, intermediate CDR H3 lengths of 36, 36, and 37 Å, respectively. The distance above the tips of the other CDR loops is not simply a function of the number of residues in CDR H3 but also of how the knob domain is positioned with respect to its stalk. The rest of the heavy-chain variable region is very similar in sequence, with differences at only 13 positions among the five Fabs for residues H1 to H92 (Fig. 2B). The variable regions also superimpose with low root mean square deviations (RMSDs) on Cα residues (compared to B11) of between 0.43 and 0.59 Å for A01, E03, and BLV5B8. BLV1H12 has a larger RMSD of 1.43 Å as a result of a change in CDR H1 conformation and a slight shift at the tip of a loop in the heavy framework region 3 (residues H72 to H76) that corresponds to HV4 in T cell receptors and shark IgNAR. The ultralong CDR H3 antibodies have also been reported to use a largely invariant light chain (14), and all heavy chains are paired with an identical light chain in the five structures (Fig. 2B). The five λ chains are also highly structurally similar with RMSDs, ranging from 0.28 to 0.50 Å for Vλ residues (when compared pairwise to B11).

Many different schemes for numbering antibody variable regions are available, such as those of Kabat et al. (15), Chothia and Lesk (16), Abhinandan and Martin (17), and Honegger and Pluckthun (18). However, the ultralong bovine CDR H3s are far too long to be accommodated by any of these numbering schemes. For clarity, we will use Kabat numbering here for the light-chain (L) and for heavy-chain residues 1 to 100 and 101 to 228, but residues in CDR H3 encoded by the D3H10 and JH3 genes will be numbered sequentially. The D3H10-encoded region begins with a highly conserved CDP motif at residues 2 to 4; therefore, in our numbering scheme, the conserved Cys is residue "D2" (Fig. 1) followed by D3, D4, etc., and then the JH3-encoded residues J1 and J2 followed by H101.

As noted previously for human and mouse Fabs with λ light chains (19), bovine Fabs can also adopt a wide range of elbow angles, even within the same crystallographic asymmetric unit. In Fab E03, with four Fabs in the asymmetric unit, two Fabs have similar elbow angles (109° and 118°), whereas the other two Fabs are also similar (205° and 208°) but differ from the first pair. Fabs B11 and A01, with one Fab in the asymmetric unit, have elbow angles of 139° and 136°; BLV5B8 and BLV1H12, with two Fabs in the asymmetric unit, have elbow angles of 131° and 134°, and 150° and 139°, respectively. Thus, the elbow angles are extremely flexible despite the high sequence conservation...
in the variable (except CDR H3) and constant regions of these bovine antibodies.

**Knob structure**

The knob domains of the five Fabs appear dissimilar upon initial inspection, but some conserved structural features can be identified after their superposition. Each knob domain is initiated by a type I β-turn, followed by three antiparallel β strands, with the strands connected by loops of differing lengths and conformations (Fig. 4) varying from short two-residue connecting loops in E03 to longer helical or extended loops in the other Fabs. The strands [as defined by DSSP (20)] also differ slightly in length (Fig. 4). The sequence around the type I β-turn is fairly well conserved, with the highly conserved CysD2 followed by Pro/SerD3, AspD4, Gly/AspD5, and Tyr/SerD6, where D4 and D5 are the i + 1 and i + 2 residues of the β-turn. CysD2 always forms a spatially conserved disulfide with a residue in the second

| Table 1. Data collection and refinement statistics for bovine Fabs. |
|----------------|----------------|----------------|
|                | Fab B11        | Fab E03        | Fab A01        |
| Beamline       | SSRL 12-2      | APS 23-ID-B    | SSRL 12-2      |
| Wavelength (Å) | 0.9795         | 1.0331         | 1.0332         |
| Resolution (Å)*| 38.9–2.06 (2.13–2.06) | 49.1–2.20 (2.24–2.20) | 38.4–2.00 (2.07–2.00) |
| Space group    | C2             | P2,           | C2             |
| Unit cell (Å)  | 78.41, 71.46, 88.32 | 62.74, 221.74, 67.95 | 78.41, 71.46, 88.32 |
| Unit cell (%)  | 90, 97.27, 90  | 90, 104.81, 90| 90, 107.31, 90 |
| Total reflections | 126,115 (12,027) | 357,535 (17,161) | 206,132 (21,059) |
| Unique reflections | 28,518 (2,791) | 90,089 (4,474) | 31,945 (3,179) |
| Multiplicity   | 4.4 (4.3)      | 4.0 (3.9)      | 6.5 (6.6)      |
| Completeness (%)| 95.0 (94.1)    | 100.0 (100.0)  | 98.4 (98.8)    |
| Mean (I/σI)   | 13.8 (1.6)     | 16.5 (1.6)     | 11.2 (1.8)     |
| Rmerge†       | 0.061 (1.12)   | 0.097 (1.00)   | 0.115 (1.45)   |
| R̂pim§        | 0.030 (0.53)   | 0.051 (0.52)   | 0.045 (1.34)   |
| CC1/2¶        | 1.00 (0.73)    | 0.91 (0.68)    | 1.00 (0.68)    |
| Rwork         | 0.19 (0.38)    | 0.22 (0.31)    | 0.21 (0.33)    |
| Rfree         | 0.24 (0.38)    | 0.25 (0.38)    | 0.25 (0.35)    |
| Number of reflections used in refinement (work/free) | 26,133/2,372 | 85,377/4,416 | 30,047/1,883 |
| Number of protein atoms | 3,562 | 13,583 | 3,592 |
| Number of water molecules | 55 | 339 | 37 |
| Number of protein residues | 486 | 1,842 | 484 |
| RMS (bonds)   | 0.003          | 0.003          | 0.003          |
| RMS (angles)  | 0.61           | 0.66           | 0.66           |
| Ramachandran: favored, allowed, outliers (%) | 94.2, 5.2, 0.6 | 95.9, 3.8, 0.3 | 96.0, 4.1, 0.0 |
| Clash score||1.86 | 1.87 | 1.42 |
| Wilson B (Å²) | 45.1           | 36.6           | 39.7           |
| Average B (Å²) | 73.3           | 65.9           | 60.1           |
| Protein       | 73.5           | 66.8           | 60.2           |
| Solvent       | 55.8           | 42.5           | 45.9           |

*Numbers in parentheses are for the highest-resolution shell. †Rmerge = ΣhklΣi=1,n |Ii(hkl) − 〈I(hkl)〉| / ΣhklΣi=1,n Ii(hkl). §R̂pim = ΣhklΣi=1,n |Ii(hkl) − 〈I(hkl)〉| / ΣhklΣi=1,n Îi(hkl). ¶CC1/2 represents the Pearson correlation coefficient between two random half data sets. ||The number of unfavorable (≥0.4 Å) all-atom steric overlaps per 1000 atoms.
b strand (CysD13, CysD19, CysD23, and CysD20 in E03, B11, BLV1H12, and BLV5B8, respectively); however, we can discern no other conserved sequence patterns within the knob domains (Fig. 4). The second cysteine in the conserved disulfide bond (CysD13, CysD19, CysD23, and CysD20) is also separated in the linear sequence from its conserved partner, CysD2, by anywhere from 10 to 20 residues, depending on the length of the loop connecting the first and second strands. All cysteine residues form disulfide bonds within the knob domains and, with the exception of the spatially conserved disulfide involving CysD2, little conservation is observed in the disulfide positions.

Fig. 2. Bovine Fab structures and sequence alignments. (A) Crystal structures of bovine Fabs E03, B11, BLV1H12 (PDB 4K3D), BLV5B8 (PDB 4K3E), and A01 in a schematic representation. The light chains are in pink and the heavy chains are in light blue. The core b strands (as defined by DSSP) in the distal CDR H3 knob domains are colored yellow, green, and blue. (B) The bovine Vl sequence (top) is aligned with the Vl sequence from human antibody KOL (PDB 2FB4) (61). The five bovine Vh sequences (bottom) are aligned with the Vh sequence from human Fab PGT121 (PDB 4JY4) (24). The bovine H3 regions are aligned on the basis of structural conservation, with the type l b-turn residues in red and with b-strand residues colored as in (A).

β strand (CysD13, CysD19, CysD23, and CysD20 in E03, B11, BLV1H12, and BLV5B8, respectively); however, we can discern no other conserved sequence patterns within the knob domains (Fig. 4). The second cysteine in the conserved disulfide bond (CysD13, CysD19, CysD23, and CysD20) is also separated in the linear sequence from its conserved partner, CysD2, by anywhere from 10 to 20 residues, depending on the length of the loop connecting the first and second strands. All cysteine residues form disulfide bonds within the knob domains and, with the exception of the spatially conserved disulfide involving CysD2, little conservation is observed in the disulfide positions.
Notwithstanding, Fabs BLV1H12 and BLV5B8 share spatial conservation of one other disulfide, but the actual residues on the C-terminal side that form the disulfide differ; the BLV1H12 disulfide forms between CysD12 and CysD32, whereas that of BLV5B8 is between CysD12 and CysD24 (Fig. 4). Although high sequence conservation makes it easy to identify residues D2 to D5 from aligned sequence data, assigning the actual residues that form the second and third β strands and the disulfide connectivity is very difficult without a crystal structure. Fab A01 was also included in this study because it has a very rare sequence that lacks the highly conserved CysD2. However, its crystal structure reveals the same type I β-turn and a similar three-stranded knob structure, despite lacking the conserved CysD2 disulfide. The disulfide connectivities in each knob domain differ, so that E03, with just one disulfide, has 1-2 connectivity; A01, BLV1H12, and BLV5B8, all with three disulfides, have 1-4, 2-5, 3-6 (A01), 1-4, 2-6, 3-5 (BLV1H12), and 1-3, 2-4, 5-6 (BLV5B8) connectivity; and B11, with four disulfides, has 1-4, 2-7, 3-8, and 5-6 connectivity. Fab A01, which lacks the conserved D2 cysteine, has weak electron density for an alternate disulfide connection between CysD24 and CysD38 (present at about 40% occupancy), in addition to the CysD10/CysD24 and CysD19/CysD38 disulfides with stronger electron density. In the alternate configuration, two cysteines would be left unpaired (D10 and D19). Thus, the cysteines in the CDR H3 of Fab A01 may form 1-4, 2-5, 3-6, and, less often, 2-5 and 1-3 disulfide patterns. Although we do not see electron density supporting alternate disulfides in the other structures, each knob domain has some cysteine residues that are within disulfide bonding distance to more than one cysteine (table S1), with many Ca-Ca distances between 3.4 and 7.6 Å, which corresponds to the range of Ca-Ca distances found in a survey of 351 disulfide bridges (21). The close spacing of these cysteine residues may be beneficial during SH, where alternate disulfide bonds could be readily formed as cysteines are lost or gained.

**Stalk structure**

All of the bovine Fab CDR H3 stalks fold into long, extended β-ribbons with a number of aromatic residues, especially in the descending strands, that may aid in stabilization of the long stalks (Fig. 5 and fig. S1). Some human antibodies also have multiple aromatic residues in this region (for example, 4JY4 in Fig. 2), especially those using the human IgH J603 germline gene coding for the sequence “YYYYYMDVWGKG.” Several of these human antibodies have extended β-hairpin CDR H3s whose bases (stalks) are similar in conformation to the corresponding region of the bovine CDR H3 stalks. These antibodies include anti–HIV-1 antibodies 10-1074 (PDB 4FQ2) (22), PGT121 (PDB 4NPY) (23), PGT122 (PDB 4JY5) (24), and Z13e1 (PDB 3FN0) (25), and anti–Marburg virus antibody MR78 (PDB 3X2D) (fig. S2) (26). The bovine CDR β-ribbons start with an ascending strand, which emanates from the base of CDR H3 and leads into the stalk domain, and then returns via the descending strand, in which all five structures share a kink at its base around Kabat residues H100 to H103 (with Val directly preceding H101 (which is called here J2), Asp/GluH101, Ala/Thr/ValH102, and TrpH103). The kink differs slightly from those seen in human and mouse Fabs, where a hydrogen bond is usually formed between TrpH103 NE1 and the carbonyl oxygen from the residue preceding H101. In the five bovine structures, TrpH103 NE1 instead hydrogen bonds with the conserved ThrL46 side-chain hydroxyl (Fig. 6A). L46 is Thr in 15 nonbovine Fab PDB depositions, and those structures show a mixture of kinked and extended bases; thus, ThrL46 is not sufficient to always enforce a kinked H3 base in nonbovine heavy chains. The residue at position 101 in the five bovine heavy chains is Asp or Glu. In mouse and human kinked H3 bases, Asp at this position often hydrogen bonds with ArgH104 on the opposing H3 strand; however, in the bovine structures, H94 is Thr or Ser. Thus, H101 (i) does not interact with any other residues (E03 and BLV1H12), (ii) forms a salt bridge with conserved LysH102 (A01 and B11), or (iii) hydrogen bonds with HisD46 (BLV5B8). ThrH104 is germline-encoded in the V_{\mu}BUL gene (H92CTTVHQH97) (Fig. 1), whereas most mouse or human antibodies have H92CATRH94 or H92CACATHR4 rather than H92CCTPH94 (2). The length of the extended hydrogen-bonded ribbon connecting the kinked H3 base and knob domains can differ, with long (~12 residues) ribbons in E03, BLV1H12, and BLV5B8 and with shorter ribbons (~8 residues) in B11 and A01. The shorter ribbons are separated from the knob domain by a five-residue loop on the descending strand involving residues D41 to D45 (SSDNT or LDSST) that disrupts the hydrogen bond ladder with the opposing strand and may allow the knob domain to have more flexibility around its stalk (Figs. 3 and 5). In B11, residues SSDNT form an undefined five-residue turn, whereas DSST in A01 forms a type I β-turn.

Fig. 3. Architecture and relative disposition of bovine CDR H3 stalk and knob domains. (A) When the V_{\mu} domains of five bovine antibodies are superimposed, very little residual movement is seen in the stalk regions. CDR H3s from bovine antibodies B11, BLV1H12, BLV5B8, A01, and E03 are shown in red, yellow, green, blue, and pink, respectively, with orthogonal views on the left and right. (B) In contrast, when the knob domains are superimposed by their conserved type I β-turn and β-strand core, the difference in the position of the supporting stalks is much larger, showing that different knob domains have different relative orientations with respect to the conserved stalk and V_{\mu} core.
Some side-chain hydrogen bonds within the stalk may also aid in its stabilization (table S2), with one fairly well-conserved hydrogen bond between His/TyrH96 and a Glu located six residues before the conserved TrpH103 (Glu D49, D30, D49, and D44 in B11, E03, BLV1H12, and BLV5B8, respectively). This conserved hydrogen bond is not present in A01; instead, AspD49 hydrogen bonds to ArgH100b, whereas HisH96 hydrogen bonds to GluH98. Thus, the conserved rigid stalk structures share potentially stabilizing hydrogen bonds and aromatic residues, a kinked base, and a potentially flexible junction with the knob domain.

**Canonical CDRs**

Human and mouse CDRs L1, L2, L3, H1, and H2 are encoded by the diverse germline variable region genes but adopt canonical structures that can often be predicted on the basis of key residues within the loops (16, 27–29). In typical antibodies, these CDRs contribute to antibody diversity and antigen contacts. However, bovine ultralong CDR H3 antibodies have most of their sequence diversity in CDR H3 and appear to bind antigen mainly or only through CDR H3 (2), suggesting that the other CDRs may play a purely structural role. Thus, we compared the bovine CDR sequences and conformations to determine whether they adopt canonical structures found in other conventional antibodies (Fig. 6B). Most of the bovine CDRs adopt structures consistent with the previously established canonical classes. L1 resembles class 5/13a, with a small deviation and carbonyl shift around GlyL31; L2 is in the 1/7A class; and L3 is predicted to belong to the 5/11a class but is bent over at its tip to accommodate the descending strand of the stalk (2). CDR H1 from B11, A01, and E03 belongs to canonical class 1/10A; however, in BLV1H12, the CDR H1s of the two Fabs in the asymmetric unit do not adopt the expected canonical structure as they interact with one another in the crystal lattice. CDR H2 belongs to class 1/9a. However, the main role of the bovine light-chain CDRs appears to be to lend structural support to CDR H3: stabilizing contacts include stacking of CDR L1 TyrL32 against CDR H3 Glu/Asp (three residues before H100), stacking of CDR L2 TyrL69 with His/Tyr (one residue before H100), stacking of the CDR L3 tip (AspEDSSL65) with Tyr (four residues before H100) (fig. S3), and hydrogen bonding of CDR L1 AsnL130 to the main chain of Tyr (four residues before H100). As originally proposed on the basis of a homology model.

Fig. 4. Secondary and tertiary structures and topology of the bovine CDR H3 knob domain. (A and B) The knob domain secondary and tertiary structures (A) and primary structure topology diagrams (B) are shown from left to right for bovine Fabs E03, B11, BLV1H12, BLV5B8, and A01. The spatially conserved disulfide from CysH102 to a Cys in the second β strand in the knob domain is shown in red, and the other disulfides are color-coded to match the topology diagrams. β strands 1, 2, and 3 are colored yellow, green, and blue in all panels. GlyL32 in A01 (that is usually a Cys in other bovine Fabs) is highlighted in red on a white background.
the small Ser at position L90 is necessary because its side chain is tightly packed between CDR L1 Asn\textsuperscript{127b} and Val\textsuperscript{128} and a Trp/Phe in the descending strand of CDR H3 (Trp\textsuperscript{D31}, Phe\textsuperscript{D50}, Trp\textsuperscript{D48}, Phe\textsuperscript{D45}, and Trp\textsuperscript{D50}, in E03, B11, BLV1H12, BLV5B8, and A01, respectively). Contacts from CDR H3 to the heavy-chain CDR H1 and H2 are less extensive, with one potentially stabilizing hydrogen bond between CDR H2 Ser/Asn\textsuperscript{H50} and CDR H3 Gln\textsuperscript{H97}.

On the basis of the two original structures (2), CDR H1 Asp\textsuperscript{H31} and Lys\textsuperscript{H32} were proposed to stabilize H3 by interacting with His\textsuperscript{H96} via a water molecule (2). However, this interaction is only found in BLV1H12 because of the unusual conformation of CDR H1 (likely induced by crystal packing). With the exception of the bent-over CDR L3, the canonical structures and residue composition of the bovine CDRs are not notably different from those in classical human and mouse antibodies. Thus, these bovine CDRs have not evolved unique structures or unusual sequences but have coevolved nevertheless with CDR H3 to optimize the rigidity and stability of the extended bovine H3 stalk. Nonbovine antibodies with long, protruding CDR H3s may also use their CDR loops, especially from the light chain, for H3 stabilization (30), whereas human anti-hemagglutinin (HA) antibody C05 that binds antigen primarily with its long 24-residue CDR H3 (31) uses its CDRs H1 and H2 to make potentially stabilizing contacts to rigidify and extend its protruding H3.

**Comparison of knob with other protein folds**

We searched for proteins with structural homology to the knob domains using the PDBFold server (32), and the closest matches were the small cyclic peptide (cyclotide) Kalata B1 (PDB 1N1U) (33), WW domains such as FE65 (PDB 2H02) (34), and a segment from the tumor necrosis factor receptor superfamily member 13C (PDB 1OQE) (35). The best match was between the E03 knob and Kalata B1, a small (29-residue) cyclic protein with antimicrobial activities thought to result from membrane disruption. Kalata B1 has three disulfides, and 16 of its residues overlap with the E03 knob with an RMSD on C\text{α} of 1.6 Å. The conserved Cys\textsuperscript{D2} disulfide in E03 is spatially conserved with the 3-6 disulfide in Kalata B1 (Fig. 7). Strands 1 and 2 in E03 (SRLI and DCRN) are the same length as the corresponding strands in Kalata B1 (CTCS and VCTR), whereas strand 3 in E03 has two residues (DV) that overlap fairly well with the corresponding region in Kalata B1; however, that region in Kalata B1 is not defined as a strand by DSSP.

**DISCUSSION**

The paradigm for antigen recognition by most vertebrate antibodies is to make specific contacts with antigen through many or all of the hypervariable CDR loops that emanate from and are supported by the framework β strands within the Ig domains of the heavy- and light-chain variable regions. Although unusual antibodies (such as the human anti-HA antibody C05) can bind antigen primarily using CDR H3 (31), or only the heavy-chain CDRs as for HA stem binding V\textsubscript{H}1-69 antibodies (36, 37), most antiprotein antibodies use 5-6 CDR loops for antigen contact (38). However, the component of the cow antibody repertoire that is characterized by ultralong CDR H3s appears to use...
only CDR H3 for antigen recognition, as shown previously in mutation studies of the unusual H3 knob domain (2).

The five bovine Fab structures that are now available enable us to compare and analyze these structures to identify any conserved features and motifs in these ultralong CDR H3s. Such a study has now revealed new features and mechanisms to attain structural diversity for antigen recognition. First, three conserved short β strands within the knob domain provide a scaffold for two hypervariable loops that can differ in length, structure, and disulfide connectivities. Second, the stalk can protrude at slightly variable distances from the surface of

Fig. 6. Comparison of bovine CDR loop conformations with CDR canonical classes. (A) The H3 “kink” or “bulge” in 2FB4 (left) and B11 (right) is shown. In a “normal” kinked base, Trp103 NE1 hydrogen bonds to a neighboring carbonyl as in 2FB4, but in the bovine CDR H3, Trp103 NE1 hydrogen bonds to the side chain of Thr46. (B) The CDRs L1, L2, L3, H1, and H2 are shown superimposed with canonical reference CDRs from published structures. Bovine CDRs are shown in gray, and reference CDRs are shown in red. Sequences for the CDRs are shown below the superpositions. The sequences for CDR H1 are identical in all five bovine structures but differ for CDR H2, whereas the λ chain CDR sequences for all five bovine Fabs are identical. Because the bovine λ chains are therefore also highly similar in structure, only one bovine CDR loop is shown in each of the superpositions. The L1 loops are very similar to canonical class 5/13a, except at Gly31, where a difference is observed in the position and orientation of its carbonyl oxygen. The L2 loops belong to class 1/7a. The L3 loops would be predicted to belong to class 5/11a, but the bovine loop tip bends away from the canonical structure to support the H3 stalk. The H1 loops for B11, A01, E03, and BLV5B8 (left) align fairly well with canonical class 1/10a, but both Fabs in BLV1H12 (center) have distorted H1 loops that interact with each other via crystal packing interactions. The H2 loops belong to class 1/9a.
Most ultralong CDR H3 antibodies maintain a conserved cysteine at position D2 that forms a disulfide bond with a Cys in the second core β strand. However, although conserved in the three-dimensional structure, this second cysteine is not conserved in the linear sequence of the five Fab fragments. Although the germline D_{H2} region encodes four cysteines, only the Cys encoded at D2 is highly conserved in a deep sequence analysis of CDR H3 (2). Thus, diversity generation mechanisms may place the second cysteine in the appropriate location to form a disulfide bond with D2.

Tyrosine, glycine, and serine are highly abundant in the bovine D_{H2}-encoded germline sequence “SCPDGYSTYGCGYGGYGCSCGYDCYGYGGYGGYGYSSYSSYTYEY” and account for 6.7% (Tyr), 10.0% (Gly), and 9.5% (Ser) of the residues from a set of 52 mature bovine CDR H3 sequences. These amino acids are also found frequently in human and murine CDR H3s (3, 39); however, in a survey of the complete UniProtKB/SwissProt sequence database as of February 2016 (http://web.expasy.org/docs/relnotes/relstat.html), the frequencies for these particular amino acids in all available protein sequences are 2.9% (Tyr), 7.1% (Gly), and 6.6% (Ser), so that they are indeed overrepresented in these ultralong CDR H3s. Synthetic libraries using antibody scaffolds and other protein folds have been developed using very small subsets of amino acids as recognition elements, with tyrosine, serine, and glycine proving the best minimal combination (40) for effective antigen recognition. Tyrosines are typically responsible for forming most of the molecular contacts, whereas the smaller serine and glycine residues can add flexibility and act as spacers between the larger tyrosine residues (41). Some proteins, including “keratin-associated proteins” (KAP) and the shell protein prismalin-14, also have high glycine/tyrosine content and share sequence motifs such as “YGGG-YYGGG” that are also encoded by the germline D_{H2} gene. The KAP proteins cross-link with keratin filaments and also can contain a number of cysteines, whereas prismalin-14 is involved in mollusk shell formation. Whether there are similar binding functions between these motifs and the cow ultralong CDR H3 germline antibody requires further study.

Questions remain as to how the bovine ultralong CDR H3 antibodies recognize antigen and what types of epitopes they recognize. Earlier work has indicated that the IgM antibodies BLV1H12 and BLV5B8 are multispecific, recognizing antigens such as DNA, synthetic haptons, and cytoskeleton proteins, and may be natural auto-antibodies (1, 42). Previous experiments with one ultralong antibody to the immunogen bovine viral diarrhea virus indicate that only CDR H3 is necessary for binding antigen (2). Bovine heavy chains from deep sequencing show high sequence homology in all regions but CDR H3, suggesting that CDR H3 is the sole, or at least the most critical, element for antigen recognition, with decreased variability in other locations. The longest human CDR H3s have mostly been found in human anti–HIV-1 antibodies where they can insert between the densely packed carbohydrates of the glycan shield to attain contact with the protein surface below; thus, the bovine CDRs might be used in a similar fashion to reach into deep clefts or pores. The structural similarity to Kalata B1 also suggests that the antibodies might use the knob regions in their CDR to disrupt membranes; however, extremely limited tests in our laboratory have so far not been able to support this notion. The plethora of different CDR H3 sequences in cows argues against such an innate immune function for these antibodies, suggesting instead that they recognize many different antigens. It is remarkable that cows have evolved such a unique scaffold and architecture for antigen recognition, but which is still comprised within the CDR H3 of a typical Ig domain framework.

Cyclotides such as Kalata B1 are being used as scaffolds for the design of small binding proteins (43–46), and their structural homology with the knob domains on the bovine antibodies suggests that...
libraries of bovine antibodies with extensive natural variation in the loop regions and disulfide pairings could be used to search for small knob domains with specific affinities or functions. Antibody engineering inspired by the bovine antibody template has already taken off at a rapid pace, with knob domains replaced with other proteins, stalk regions replaced by stable structures such as a coiled-coiled domain, or stalk/knob structures transplanted into the CDRs of human antibodies (47–54). Antibodies with novel activities have also been created through insertion of large protein domains into the CDR H3 of conventional antibodies by functional selection using libraries of permissive junctions (55). On the basis of these interesting antibodies, the structural insights presented here may aid in future immunization or engineering efforts and could further our understanding of their unique place in the bovine immune system and the unusual paradigm they represent in creating genetic and structural diversity.

MATERIALS AND METHODS

Protein expression and purification
The light-chain DNA and the heavy-chain DNA for each bovine IgG were cloned separately into vector pcDNA3, and a stop codon was introduced into the heavy-chain DNA after residue H228 for Fab production. Mammalian 293Freestyle cells (1 liter) were transformed with light-chain DNA (0.25 mg) and heavy-chain DNA (0.50 mg) and incubated in a shaking CO2 incubator for 6 days. The cells were centrifuged at 4000 rpm for 20 min, and the medium supernatant was then filtered with a 0.22-μm filter. The filtered medium was loaded onto a CaptureSelect LC-lambda (ungulate) resin (Life Technologies) at 1 ml/min and washed with phosphate-buffered saline (pH 7.4), and Fab was eluted with 0.1 M glycine (pH 3) into fractions containing 1 M tris (pH 8.0). Each Fab was further purified by size exclusion chromatography on a Superdex 75 16/600 column (GE Healthcare Life Sciences) with running buffer [20 mM Hepes (pH 8) and 150 mM NaCl]. Fabs were concentrated in this buffer for crystallization trials: Fab A01, 16.9 mg/ml; Fab B11, 16.1 mg/ml; and Fab E03, 12.3 mg/ml.

Crystallization
All Fabs were screened for crystallization using 384 conditions (JCSG Core Suites I, II, III, and IV) at 4° and 20°C using the IAVI/JCSG/TSRI CrystalMation robot. The Fab B11 crystal for data collection was obtained using a precipitant of 5% polyethylene glycol (PEG)–3000, 30% PEG-400, 0.1 M Hepes (pH 7.5), and 10% glycerol (20°C; JCSG Core Suite III, condition C9). The crystal was flash-cooled in mother liquor with no additional cryoprotectant. Fab E03 crystals were obtained in JCSG Core Suite II, condition H5, and that condition was further purified by size exclusion chromatography on a Superdex 75 16/600 column (GE Healthcare Life Sciences) with running buffer [20 mM Hepes (pH 8) and 150 mM NaCl]. Fab B11 crystal for data collection was grown at 22°C with a precipitant of 0.2 M NaCl, 16.5% PEG-8000, and 0.1 M phosphate-citrate buffer (pH 4.2), and flash-cooled after brief immersion in a well solution augmented with 30% ethylene glycol. The Fab A01 crystal was taken directly from the screening tray, with a precipitant of 20% PEG-6000 and 0.1 M bicine (pH 9) (4°C; JCSG Core Suite I, condition A2), and flash-cooled after immersion in a well solution augmented with 30% glycerol.

Data collection
Diffraction data for Fab B11 were collected at Stanford Synchrotron Radiation Lightsource (SSRL) beamline 12-2 and processed with XDS (56). Reflections between 2.24 and 2.26 Å had elevated average intensities as a result of ice diffraction obscuring that region and were removed from the data set, resulting in data with 95.0% completeness up to 2.06-Å resolution. Data for Fab E03 were collected at Advanced Photon Source (APS) beamline 23-ID-B and processed with HKL-2000 (57), resulting in data with 100.0% completeness up to 2.20-Å resolution. Data for Fab A01 were collected at SRSL beamline 12-2, processed with XDS, and 98.4% complete up to 2.00-Å resolution.

Structure solution and refinement
All structures were determined by molecular replacement, using Phaser (58) and BLV1H12 (4K3D) (2) as a model. Structures were refined with PHENIX (59), and structural figures were made with PyMOL (60). Fabs B11 and A01 each have one Fab molecule in the asymmetric unit. Fab E03 crystallized with four Fabs in the asymmetric unit. One of the four Fabs (chains E, F) has very poor electron density due to fewer crystal contacts, and the Fab with the best electron density (chains A, B) was used for the comparisons presented here.

SUPPLEMENTARY MATERIALS

immunology.sciencemag.org/cgi/content/full/11/1/aaf7962/DC1

Fig. S1. Electron density for bovine CDR H3 stem regions.

Fig. S2. Comparison of the bovine CDR H3 stem with the corresponding region of human CDR H3s encoded by the IgH J6*03 gene.

Table S1. Cysteine Co-Cx distances (Å) in bovine CDR H3 knobs.

Table S2. Side-chain hydrogen bond partners within CDR H3 stem.

REFERENCES AND NOTES


10. R. L. Stanfield, A. Zemla, I. A. Wilson, B. Rupp, Antibody elbow angles are influenced by their light chain class.


Acknowledgments: We thank Y. Hua and H. Tien for excellent technical assistance, and O. Bazirgan for providing antibody heavy-chain vectors. Portions of this research were carried out at the SSRL, a Directorate of the Stanford Linear Accelerator Center National Accelerator Laboratory and an Office of Science User Facility operated for the U.S. Department of Energy (DOE) Office of Science by Stanford University. The SSRL Structural Molecular Biology Program is supported by the DOE Office of Biological and Environmental Research and by the NIH National Institute of General Medical Sciences (NIGMS; including grant P41GM103393) and the National Center for Research Resources (NCRR; grant P41RR001209). The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the NIGMS, NCRR, or NIH. Other portions of the research were carried out at GM/CA@APS; which has been funded in whole or in part with federal funds from the National Cancer Institute (grant ACB-12002) and the NIGMS (grant AGM-12006). This research used the resources of the APS, a DOE Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory under contract number DE-AC02-06CH11357. This is manuscript number 29310 from The Scripps Research Institute. Funding: This work was supported in part by NIH grants SR01 GM105826 (V.V.S.), R56 AI117675 (I.A.W.), and R01 AI084817 (I.A.W.). Author contributions: R.L.S. designed and performed the experiments, analyzed the data, and wrote the manuscript. I.A.W. and V.V.S. directed the project, analyzed the data, and wrote the manuscript. Competing interests: V.V.S. has equity in Sevion Inc., which has interest in the commercial development of cow antibodies. The other authors declare that they have no competing interests. Data and materials availability: The coordinates and structure factors for this study have been deposited to the PDB with codes 5IHU (Fab B11), 5IJV (Fab E03), and 5ILT (Fab A01). All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

Submitted 31 March 2016
Accepted 10 May 2016
Published 14 July 2016
10.1126/sciimmunol.aaf7962