Fundamental Characteristics of the Immunoglobulin V<sub>H</sub> Repertoire of Chickens in Comparison with Those of Humans, Mice, and Camelds

Leeying Wu, Katarzyna Oficjalska, Matthew Lambert, Brian J. Fennell, Alfredo Darmanin-Sheehan, Deirdre Ní Shúilleabháin, Bénédicte Autin, Emma Cummins, Lioudmila Tchistiakova, Laird Bloom, Janet Paulsen, Davinder Gill, Orla Cunningham and William J. J. Finlay

*J Immunol* 2012; 188:322-333; Prepublished online 30 November 2011;
doi: 10.4049/jimmunol.1102466
http://www.jimmunol.org/content/188/1/322

**Supplementary Material**

http://www.jimmunol.org/content/suppl/2011/11/30/jimmunol.1102466.DC1

**References**

This article cites 70 articles, 15 of which you can access for free at:
http://www.jimmunol.org/content/188/1/322.full#ref-list-1

**Subscription**

Information about subscribing to *The Journal of Immunology* is online at:
http://jimmunol.org/subscription

**Permissions**

Submit copyright permission requests at:
http://www.aai.org/About/Publications/JI/copyright.html

**Email Alerts**

Receive free email-alerts when new articles cite this article. Sign up at:
http://jimmunol.org/alerts
Monoclonal Abs have become the most rapidly expanding area in biotherapeutic drug discovery (1, 2). In Ab drug discovery programs, panels of lead Abs are often required that exhibit high affinity, react with multiple orthologs of the target, have broad epitope coverage, and most importantly, potently modulate the biological function of their target. It is also beneficial for downstream development into viable therapeutic agents that the resulting Abs should have uniform stability and expression characteristics. A relatively untapped source of such Abs is the immune repertoire of animals whose v-gene germline diversity is restricted, but still highly homologous to human sequences. A number of biotherapeutic research groups have, therefore, developed an interest in an underused option in biotherapeutic drug discovery: high-affinity, potently neutralizing rAbs from chickens (3–5). This approach exploits the fact that avians are phylogenetically distant from mammals but have theoretically uniform, humanizable v-gene sequences (3, 4).

The generation of Ab diversity in chickens is well characterized at the genomic level (6–8), but despite many years of study of this model species, the true amino acid content of the final functional repertoire has not been extensively examined in the way that it has for primates and rodents (9–14). The VH domain of Abs is typically critical to Ab specificity, and the humanization of Ab therapeutics is greatly aided by understanding the structural characteristics of the VH repertoire, from which the lead clone is derived. We felt, therefore, that a detailed understanding of the fundamental amino acid content of the adult chicken repertoire would shed light on the similarities and differences between the v-gene diversification mechanisms of humans, rodents, and avians. It would also greatly inform our ability to harness some of the unique properties of chicken Abs. In addition, a clear view of the diversification mechanisms and final amino acid content used in the VH CDRs is critical to the development of novel design features in synthetic Ab libraries (14–16).

The Ig system of chickens is distinct from that of humans and mice. Although chickens have structural and functional equivalents of mammalian IgM, IgA, and IgG, they have not been found to produce direct homologs of IgE or IgD, with IgM being the major isotype expressed on the surface of their B cells (6). Indeed, it has been proposed that chicken IgG, with its four Cγ domains, may be a structural relative of both mammalian IgG and IgE subclasses (17). Studies on the underlying molecular biology of chicken v-gene diversification have led to some fascinating insights. The simple avian v-gene germline repertoire is in marked contrast to
the system used by humans, mice, and primates, which involve a large set of sequences that are highly diverse in both sequence and structure, and are predominantly capable of functional VDJ recombination (18, 19). In chickens, single functional v-genes exist for both the L and H chains, containing unique V_{L-1}-J_{L}, V_{H-1}-J_{H}, and D segments (8, 20, 21). To use a restricted v-gene germline repertoire and yet achieve the same level of Ig protection as in mice and primates, the chicken has evolved a distinctly different set of diversification mechanisms, including the gene conversion process (8, 22). The gene conversion process in chickens is a system similar to that observed in rabbits (23) in which a template v-gene is diversified by the incorporation of segments from numerous upstream pseudogenes that lack recombination signal sequences.

Despite the use of this very simple (essentially single framework [FW]) germline v-gene system (6, 17, 24–26), chickens have a very adaptable Ab repertoire that is capable of generating high-affinity interactions with a broad range of Ags including proteins, peptides, and Happens (3, 27–29). This broad Ag recognition capability, combined with simple isolation of multithorog cross-reactive mAbs via technologies such as phage display (3, 30), have made chicken Abs a potentially valuable source of Ab therapeutics. In this study, we have therefore undertaken bioinformatics analyses to define the fundamental amino acid diversity of VH sequences from the adult chicken spleen and bone marrow B cell repertoire of both naive and immunized animals.

Materials and Methods
cDNA library generation and sequencing
Spleen and femur bone marrow were harvested from each of three adult White Leghorn chickens (Lampire). Total RNA was then prepared from each tissue sample using TRIzol reagent (Invitrogen), and cDNA was generated from each RNA sample using oligo(dT) priming and the Superscript III First Strand Synthesis kit, according to the manufacturer’s instructions (Invitrogen). Each tissue cDNA was then used to create cloned libraries of the VH gene repertoire, essentially as described previously (31, 32). DNA sequencing was subsequently performed on randomly chosen clones from each tissue sample library (GATC).

Generation of V_H repertoire amino acid sequence databases
To identify H chain v-gene sequences, Hidden Markov Model profiles were constructed using the HMMERBUILD program. The GENEWISE program was then used to compare the sequenced cDNAs against the H chain Hidden Markov Models and to produce multiple translated predictions. The best translated V_H sequences were manually evaluated to remove partial, duplicated, frameshifted, and stop codon-containing sequences to derive a set of full-length unique sequences. To compare the chicken CDR-H3 repertoire with those of humans and mice, we also downloaded all available sequences from the International Immunogenetics Information System (IMGT) database and ran the same predictions as above. Once the chicken, human, and mouse v-gene databases had been completed, amino acid usage at each position of the sequences were determined and visualized using an in-house developed Excel macro.

Selection of target-specific Abs from immune phage display libraries
Single-chain Ab (ScFv) libraries were generated from the spleen and bone marrow cDNA samples of chickens immunized with 12 mammalian and nonmammalian protein Ags including cytokines, chemokines, extracellular receptors, and purified Igs. Library generation, rescue, and selection methods used were essentially as described previously (33). All libraries were rescued and selected separately on each target protein. Target-binding specificity in resulting phage clones was determined via scFv ELISA analyses as described previously (34).

Conversion of cysteine variant scFv clone to human chimeric Fab format
ScFv clones of interest were synthesized as bicistronic Fab cassettes in the optimal L chain–H chain orientation, essentially as previously described (31, 35), for cloning into the pWRIL-1 vector (32). The final Fab cas-
FIGURE 1. Sequence deviation from germline in the naive chicken V_H repertoire. Annotated V_H sequences from 1269 separate, randomly chosen clones were used to calculate percent positional deviation from the germline sequence (8) in the naive repertoire. A, FW 1 to FW 2. B, CDR2 to FW 4. The x-axis shows the germline sequence, with amino acid positions defined using the IMGT numbering scheme. The y-axis shows percent (Figure legend continues).
CDR2 (respectively) under the Kabat numbering scheme (42) and were both found to be substituted in the majority of clones (>75 and >45% in the naive population, respectively). Although these positions are both designated as part of FW 2 under the IMGT numbering scheme, their hypervariability is not surprising because they are known to be solvent exposed and frequently mediate contacts with Ag in cocrystral structures for Abs from mammals (38, 39).

FW 2 residue 42V is also associated with V_{H1–V_{L1}} interface interactions and was mutated, but at a lower rate (7% M, 1% A, and 1% I). The other two FW 2 residues that are known to be important V_{H1–V_{L1}} interface positions (44Q and 50L) were highly conserved. Vernier residues 53V and 54A are known to affect CDR2 conformation and were rarely substituted to 1% I and 6% G, respectively.

In the FW 3 region, position 66G was found to be hypervariable. This residue is designated as a constituent of CDR2 in the Kabat numbering scheme (42) and is also known to be solvent exposed and to frequently mediate contact with Ag in mammals (39). Several other CDR2-proximal positions were also variable: 68G (20% A), 69S (47% A, 26% P, and 5% T), and 72K (5% Q, 3% D, and 1% T). CDR2 supporting Vernier residues 76A, 78I, and 80R were all highly conserved, whereas positions 82N and 87V were substituted more frequently to D (14%) and L (12%), respectively. Several other CDR2-proximal positions were also variable: 68G (20% A), 69S (47% A, 26% P, and 5% T), and 72K (5% Q, 3% D, and 1% T). CDR2 supporting Vernier residues 76A, 78I, and 80R were all highly conserved, whereas positions 82N and 87V were substituted more frequently to D (14%) and L (12%), respectively. In the case of 87V, homologous substitution to L may still be of considerable significance because this residue has been found to make CDR contacts and to be a critical position that had to be maintained to retain target-binding affinity during the humanization of two previously described chicken Abs (3, 4). In addition, FW 3 V_{H1–V_{L1}} interface residue 103Y was frequently substituted for F (24%).

The FW 4 region appeared to be the most highly conserved, with the V_{H1–V_{L1}} interface residues 118W and 119G mutating very rarely, if ever. Position 120H, which is variant in other species, was rarely mutated (<5%) in chickens. This generally low incidence of mutations in the FW 4 region may be attributed to the presence of only a single functional J-region in the chicken V_{H1} locus, which is not thought to be targeted during the gene conversion process (8).

To allow comparison of the naive and immunized repertoires, phage-display libraries of scFv were generated from chickens immunized with a variety of Ags. After phage-display selection and screening on 12 separate protein targets, 318 total unique sequences were successfully identified that exhibited specific binding to their cognate Ags in direct ELISA screening of scFvs. On analysis of a plot of the positional diversity in these 318 V_{H} sequences, generated as above, it was found that the profile of amino acid substitutions across the V-gene sequence was highly similar to that of the naive repertoire (Supplemental Fig. 1).

**Analysis of naive diversity in CDRs 1 and 2**

In the CDR1 region, considerable diversity was observed at all positions other than the highly conserved 27G and 28F (Fig. 1). CDR1 positions that were found to be highly mutated (29T-38N) were generally those known to be solvent exposed and which commonly interact with Ag (43). 37Y occupies a position that can be buried (39) but in chickens was diversified ~35% of the time, either to hydrophobic (V and F) or charged residues (R, H).

The CDR2 was found to be a region of high diversity, particularly in the solvent-exposed residues between positions 57 and 65 that are known to frequently contact Ag (Fig. 1) (43). Position 56I was highly conserved, as expected, because it is an important, buried, structural residue that is highly conserved in multiple species (16). On the few occasions that 56I was mutated (<5%), it was substituted for either homologous hydrophobic (V, L, and M) or more rarely, small residues (S and T).

Interestingly, at a number of solvent-exposed positions in the CDR1 (position 38), FW 2 (position 55) and the CDR2 (position 58), noncanonical cysteine insertions were observed at low frequency (2.1, 0.8, and 2.4%, respectively). Importantly, only a single noncanonical cysteine was ever found at any of these three positions in any individual sequence, and no clones were identified that contained a cysteine in both CDR1 and CDR2 simultaneously. Interestingly, in Ab structural models, each of these residues have commonly been mapped as packing closely to the HCDR3, where unpaired cysteines were also observed (see below) (16, 38).

**Length distribution in the CDR3 repertoire**

To compare the length distribution of the CDR3 regions, 11,423 human, 5,054 mouse, and 1,269 naive chicken sequences were examined (Fig. 2A). As observed previously, the range of human CDR3 lengths was approximately normally distributed and extremely broad (5–37 aa; mean, 16.1 ± 4.1) (14). The mouse CDR3 repertoire was much shorter (5–26 aa; mean, 11.8 ± 2.4). The chicken naive repertoire was not normally distributed and favored much longer sequences, despite a similar mean length to humans (8–32 aa; mean, 16.2 ± 3.2). The majority (89%) of chicken CDR3 lengths were between 15 and 23 aa. Shorter CDR3 lengths (<15 aa, 8 aa being the shortest observed) were found to be relatively rare (each <2% of the repertoire), with a major increase in frequency beginning at lengths 15 (6.5%) and 16 (16.7%), with length 16 being the highest frequency length overall.

In the selected chicken repertoire, the length distribution profile closely mimicked that of the naive repertoire (Fig. 2B). One notable difference was the identification of a sequence with a CDR3 length of 6 aa, two residues shorter than the shortest sequence in the naive repertoire. In addition, no binding sequences with CDR3 longer than 29 aa were observed, whereas lengths up to 32 aa could be found (albeit at low frequency) in the naive repertoire. Abs with such exceptionally long CDR3s are rarely selected from naive human Ab libraries. To estimate whether clones containing such V_{H} CDR3s could be functionally expressed in E. coli from a phagemid system, a chimeric chicken–human Fab construct was generated containing the longest CDR3 found in the naive database. This 32-aa CDR3 contained two noncanonical cysteine residues (“Type 1”; Fig. 3). Under standard expression conditions, this Fab was found, by Western blot analysis, to be expressed into the periplasm (data not shown).

**Amino acid usage in the CDR3 loop repertoire**

Analyses of amino acid distribution and Shannon entropy (44) (amino acid variability in a given position in aligned sequences) were performed for sequences at each CDR length, including residues 105–117 of both the naive and selected repertoires (Fig. 4A, 4B, Supplemental Fig. 2). Examination of the Shannon entropy plots indicated that positions 105–106 and 115–117 are
extremely conserved. Particularly highly conserved are positions 106 (K/R) and 116 (D). The maintenance of these residues is strongly indicative that the vast majority of chicken CDR3 structures involve a stabilizing salt bridge at the bottom of the loop, theoretically creating the “Kinked Base” motif (45), although a recent study suggests that this structural motif is not guaranteed (46). The amino acid distribution plots in Fig. 4 also suggest the frequent use of the motifs CAK and CAR at positions 104–106 (CAK being predominant), which are common sequences also found at these positions in primates and rodents. The IDA motif at positions 115–117 was found to be extremely highly conserved, which is due to these residues being donated by the J region. Interestingly, although there was a high level of entropy at all positions between 107 and 114 in all lengths examined, position 114 was repeatedly a position of high entropy values (~4.0) and was often preceded by reduced entropy (~3.0) at positions 112 and 113 (Fig. 4A, 4B, Supplemental Fig. 2). When examined across all CDR3 lengths, cysteine use was observed at all positions between 107 and 114, with a general trend toward the highest frequencies in the middle of the loops and progressively lowering toward the base of the loop (Supplemental Fig. 2).

To increase our understanding of CDR3 repertoire content, analysis of the total amino acid usage in chicken, mouse, and human CDR3s was subsequently performed for all residues between positions 107 and 114 (Fig. 5A). Positions 105–106 and 115–117, which are also part of the CDR under the IMGT definition, were excluded from the analysis to avoid germline FW and J region-related biases, respectively. Several major differences were observed between the CDR3 aa content of chickens and humans/mice. First, the prevalence of cysteine in the chicken repertoire was much higher than for either mammal (chicken, 9.4%; human, 1.6%; and mouse, 0.25%). That chickens should be incorporating ~38 times more cysteine than mice and ~6 times as much as human is clearly reflective of the universal content of

FIGURE 2. CDR3 length distribution in chickens. A, Comparison of CDR3 length distribution for the naive chicken VH repertoire (red column) to that found in public database sequences of mouse (gray column) and human (black column). B, Comparison of CDR3 length distribution in the naive (black column) and selected (white column) chicken VH repertoires.
that amino acid in the genomic D regions in reading frame 1 (21) and its regular use in D-like pseudogene sequences (8). Second, it was observed that tyrosine content in the chicken CDR3 region was markedly lower (9.2%) than that of humans (16.8%) or mice (26.4%).

Further differences in repertoire content were observed when examining specific groupings of amino acid chemistries. Small amino acids such as G, S, A, and T were all overrepresented in the chicken repertoire (54% total) relative to human (36% total) and mouse (35% total). For G, S, and A, this is not surprising given their high representation in chicken D regions, but T is not found in any genomic D region and may be introduced either through somatic mutation or gene conversion (21). Proline is underrepresented, however, which is surprising because the chicken repertoire has a large percentage of long CDR3s, and this residue has been observed to increase in frequency as human CDRs get longer (14). Proline is also predicted to add structure to longer human CDR3 loops (14) but may not be required for the structural stabilization of chicken CDR3 loops, because of the high incidence of intra- and inter-CDR disulfide formation. Furthermore, the content of positively charged residues (R, K, and H) was distinctly lower in the chicken repertoire (2.9% total) than in either mouse (8.7% total) or human (8.4% total). Negatively charged residue D was equally represented in chicken as it is in mouse and human (6.8–7%), whereas E was lower, with N approximately equal. Hydrophobic residues other than tryptophan (I, F, L, M, P, and V) were also significantly lower in the chicken repertoire (8.5% total) versus mouse (13.1% total) and human (19.8% total). When compared with the naive repertoire, the target-specific binding sequences again showed an extremely similar profile (Fig. 5B). Although some residues were found at moderately increased frequency (e.g., L, I, Y, T, and H), others were also slightly reduced in their representation (e.g., G, S, and C).

Cysteine usage in chicken V\textsubscript{H} CDRs

To further examine the influence of high cysteine content in the naive and selected CDR repertoires of chickens, we calculated the average number of clones having noncanonical cysteines in each VH CDR. In the CDR3 of the naive repertoire, it was found that 33.5% of all clones contained no cysteines, 53% contained two, and 0.7% contained four (Fig. 6A). That ∼87% of all clones should have an even number of cysteine residues would appear logical, given the potential negative influence of unpaired thiol groups on protein folding and secretion (47). Of greater interest, however, may be the observation that 11.4 and 1.4% (12.8% total) of all clones have one or three cysteines, respectively, in the CDR3. This odd number of thiol groups may be partially compensated, however, by the presence of extra noncanonical cysteine groups in the CDR-H1 (2.2%) or CDR-H2 (3.4%), which may pair with and stabilize the CDR3 via a disulfide bond. Importantly, the
cysteine residues observed in the CDRs H1 and H2 were found at positions 38, 55, and 58, which are typically solvent exposed and are also structurally proximal to the CDR3, making them theoretically available to form inter-CDR disulfide bonds (38, 39). When the average number of cysteines in the CDR3 was plotted for each CDR3 length, a strong correlation ($R^2 = 0.97$) is observed between increasing length and higher cysteine content (Fig. 6B). This indicates that disulfide linkage is strongly favored in longer clones, which may increase the stability of the CDR structure.

When these analyses were extended to the selected repertoire, some distinct differences were observed (Fig. 6C, 6D). The frequencies of cysteines in different CDRs changed: CDR3s with zero cysteines became dominant at 50.1%, with two cysteines lowered to 44.2%, one cysteine at 3.3%, three cysteines at 1.8%, and four cysteines at 0.7%. In CDRs 1 and 2, the frequency of cysteines dropped to 1.8 and 1.3%, respectively. Importantly, CDRs with unpaired cysteines (i.e., lacking a potential partner for intra- or inter-CDR disulfide bonding) were reduced in the selected repertoire when compared with the naive population. These changes in cysteine pattern were underlined by a reduction in the correlation between CDR3 length and cysteine frequency ($R^2 = 0.78$), when compared with the naive repertoire.

**FIGURE 5.** Comparison of total amino acid usage bias in Vh CDR3 positions 107–114 of chickens, mice, and humans. A, The percentage content of each of the natural amino acids was calculated for the CDR3s of the naive (black column), human (black column), and mouse (gray column) repertoires. For ease of comparison, the amino acids are grouped according to gross chemical character (e.g., hydrophobic [LIMV], aromatic [FYW], small [GSACTP] and charged [DEKRHI]). B, The percentage content of amino acids found in the CDR3s of the naive (black column) and selected (white column) chicken Vh repertoires.

**Proposal of structural subtypes in chicken Vh domains**

For both camelid V3H (48) and shark IgNAR molecules (49), sequence attributes including the pattern of noncanonical cysteine usage in the V domain have allowed the description of structural “types,” which may be found in the repertoire. Our analyses of the chicken Vh sequence database, therefore, led us to describe the following types 1–6, based on the presence or absence of noncanonical cysteines and listed in approximate frequency by which they are found in the naive repertoire (Fig. 3): type 1, containing two cysteines in the CDR3 (55% of the naive repertoire); type 2, containing no noncanonical cysteines (33%); type 3, containing one cysteine in the CDR1 at position 38 plus one cysteine in the CDR3; type 4, containing one cysteine in the CDR2 at positions 55 (type 4a) or 58 (type 4b) plus one cysteine in the CDR3; type 5, containing one cysteine in the CDR1 at position 38 plus three cysteines in the CDR3; and type 6, containing four cysteines in the CDR3. Each of types 3–6 represent <5% of the total naive repertoire.

**Expression and analysis of an IgM-specific type 5 chicken Fab**

Selection and screening of a scFv library generated from chickens immunized with dogfish IgM identified a clone (H2), which is an example of one of the atypical structural types (type 5; Fig. 3) outlined above. H2 exhibits strong, specific IgM binding, despite having four noncanonical sulides in the Vh (none in the Vl). Intrigued by this structure, we analyzed it further.

To examine the expression and affinity properties of H2, we first converted the V genes from scFv format into a bicistronic, chimeric Fab expression construct (31). The E. coli expression, nickel purification, and analytical size exclusion chromatography analysis of this construct proved that the Ab is produced as a stable, monomeric Fab. No evidence was observed for either covalent or noncovalent aggregate formation (Fig. 7A). Because the H2 Fab was soluble and monomeric in solution, this facilitated the determination of its affinity for IgM. Purified IgM protein was covalently coupled to a Biacore CM5 chip, and purified H2 Fab was flowed in the mobile phase, generating binding curves that fit well to a 1:1 Langmuir kinetics model, with a globally fit $K_D$ of 64 nM (Fig. 7B).

**Discussion**

As in all higher vertebrates, primary diversity in chicken Vh regions is created by V-D-J recombination, combined with somatic hypermutation (8, 20, 50). It has become clear, however, that there are other mechanisms of Vh diversification in the chicken that are distinctly different from those observed in mice and humans (8, 51). The greatest influence on the diversification of the repertoire appears to be mediated by “gene conversion,” a process in which multiple upstream $\Psi$Vh pseudogenes may undergo recombination into the Vh gene sequence, after a functional V-D-J rearrangement event (8, 50). This process occurs during diversification of the H chain, with an analogous process occurring in the L chain, leading to mutation of the CDRs and also the modulation of FW sequence. For this process to be efficient, it relies on high DNA homology between the pseudogene and the germline gene, which acts as the acceptor in the recombination event (6).

The intrinsic need for high overall sequence homology between germline and pseudogene may explain our observation that the FW regions of the chicken Vh repertoire are highly conserved in comparison with germline. We observed that hypervariability was predominantly contained within the CDRs; the majority of diversity in CDRs 1 and 2 was found in the solvent-exposed residues, and this was coupled with strong maintenance of common
structural residues that have also been observed in mice, humans, and other mammals (14, 38, 52). It is logical that the solvent-exposed residues in the CDRs should be the primary focus for diversification, because they are most likely to make Ag contacts (38), but the observation that several important structural residues are highly conserved might suggest limited conformational diversity. Clear indications could be seen, however, that regions of the FW regions are modulated, which are of specific importance to the generation of a functional repertoire. These regions included residues associated with affecting VH–VL interaction (40) and Vernier regions that directly affect CDR structure (41). Variability in the Vernier regions was observed proximally to all three CDRs. The VH repertoire therefore remains essentially “single framework” but adds significant variability at appropriate positions to increase the diversity of CDR presentation modes. It will be of considerable interest to see, in future studies, whether these observations also hold for other species that use a gene conversion method of v-gene diversification, such as rabbits.

Although the FW modulations described above aid the generation of a complex repertoire, the majority of all diversity in the VH was still observed to be in the CDR3. As the most hypervariable region in any Ab, the VH CDR3 is also likely to be a center upon which novel complexity might be built. This is clearly of critical importance to a repertoire that is primarily reliant upon a single germline gene. In chickens, V-D-J recombination leads to a much more limited primary CDR3 diversity than in humans and mice, because chickens have only 15 functional D segments, all of which are highly homologous. Indeed, some chicken D segments (e.g., D9/12/13 plus D4/8/11) are identical in amino acid sequence (21). In addition, the choice of D segment reading frame is strongly enforced by molecular control mechanisms (53). Reading frame 1 predominates in chickens, because reading frame 2 creates sequences containing stretches of hydrophobic residues such as V and L. Because reading frame 2 lacks common structural and contact residues such as glycine, tyrosine, and tryptophan, it is less likely to provide a functional Ab combining site than reading frame 1 (18). Reading frame 3 leads to the incorporation of stop codons and cannot create a functional encoding sequence (21, 23). This general style of control mechanism appears to be universal and has been observed (albeit in different reading frames) for rabbits, sharks, mice, primates, and humans (13, 18, 53).

The chicken D segment repertoire is heavily biased toward the use of reverse turn/hydrophilic residues such as G, S, and Y, just as it is in all other vertebrate species studied to date (14, 18). In contrast to humans and mice, however, chicken D segments also obligately use C. This is partly because, in reading frame 1, the 15 genomic D segments lead to a consensus sequence 9–10 aa in length: G-S-(A/G)-Y-C-(G/C)-(S/W)-X-A-(Y/E) (X, nonconserved) (21).
final repertoire is more complex than this, however, because the D segment repertoire is subsequently further diversified not just by somatic mutation but also by insertion of new sequences via gene conversion. These D-like sequences may replace the entire D segment or only a small section (8, 21).

Our analyses of CDR3 content show very clearly that the bias toward small amino acids (G/S/A/C) suggested by the genomic D segments is indeed recapitulated in the adult repertoire. As a clear indication of a preference for small amino acids, T (not observed in D segments) is also overrepresented in comparison with mice and humans. Large aromatic and hydrophobic residues are strongly disfavored in the chicken repertoire, including an unusually low representation of Y, which is dominant in the repertoires of mice and humans (14). This finding may have significant biological relevance as several synthetic Ab studies have suggested that tyrosine is a critical amino acid for forming a functional CDR3 repertoire (54–57). Indeed, structural analysis of both natural and synthetic Abs has suggested that tyrosine and tryptophan are capable of playing a dominant role in specific interaction with Ag (58). In our analysis of the CDR3 repertoire, however, tryptophan was used at 3–4% by all three species (Fig. 5). No increase in tryptophan use in the selected chicken repertoire was observed that might compensate for the low tyrosine level. This observation would also suggest that the use of D segments that encode directly for tryptophan, such as D4, D7, D8, D11, and D14 (21), is either low in the chicken repertoire, or the tryptophan residue is commonly removed via hypermutation.

In addition, the chicken CDR3 repertoire has a reduced content of positively charged residues in comparison with human and murine sequences. This may be of considerable practical importance for the biotechnological use of chicken Abs, because excess positive charge in the Ab combining site has been associated with polyreactivity (59) and poor pharmacokinetic profile in in vivo animal models (60).

Taken together, these findings suggest that the molecular basis for target interaction may differ between humans, mice, and chickens. Tyrosine, tryptophan, and arginine are all residues that have been found to make critical contacts with Ag in mammalian Abs (58). The relatively low representation of Y/W/R and bias toward G/A/S in the chicken CDR3 repertoire may indicate that Ag specificity is mediated to a greater extent by hydrogen bonding via the smaller residues than by the hydrophobic/charge interactions commonly observed for Y/W/R (61). Alternatively, it is also possible that the higher incorporation of small residues in the chicken repertoire is a necessity, because these residues are required to generate increased complexity in the secondary structure of CDR3 loops (58). This increased CDR3 complexity may help to compensate for the limited FW diversity observed in the chicken repertoire. In future studies, the derivation of cocrystal structures for chicken Abs and their cognate Ags may help to further dissect this issue.

Our analyses of the CDR3 led to the important observation that chickens use much more C in their repertoire than do mice or humans and that C usage increases in direct correlation with CDR length. The obligate use of C in the majority of unmutated chicken D segments and in D-like sequences (21) appears to lead to ≥50% of all B cell clones in the naive repertoire containing cysteine in their VH CDR3. Both humans and rhesus have been shown to use the gene segment D2, which can lead to the generation of functional CDR3 sequences containing a pair of cysteines, similar to those found in our proposed chicken type 1 VH sequences (14, 18). Despite these observations, however, CDR3 sequences containing cysteine are observed at relatively low frequency in the mature B cells of humans and primates, and they are very rare in mice.

**FIGURE 7.** Size exclusion chromatography and Bia-core characterization of the IgM-binding clone H2. A, Size-exclusion chromatograph of 100 μg/ml of the clone H2 shown in red and molecular mass markers shown in blue (kDa for each marker). Chromatograph inset represents an Instant Blue-stained reducing SDS-PAGE gel of 8 μg purified H2. Both analyses show a highly purified protein of the expected size. B, 1:1 Langmuir model fitted sensorgrams of the binding kinetics analysis for clone H2.
(14, 53). The high incorporation rate of cysteine in the chicken CDR3 may be balanced, however, by the frequent use of D-D junctions (i.e., incorporation of two D segments in one CDR3) (21). Our analyses would suggest that this D-D junction formation, along with high C content and the intrinsic flexibility of sequences biased toward G and S, can create long CDR3s that can contain intra or even inter-CDR disulfide bridges. These covalent bonds between CDRs are analogous to those observed at high frequency in the lgs of other animals such as camels (48), shark IgNARs (49, 62), cows (63–65), pigs (66), and even the duckbilled platypus (67).

Although VH CDR3s from chickens having two C residues and probable intra-CDR disulfide bonding have been previously described (28, 31), it is fascinating to note the presence of clones that appear to contain disulfide linkage between the CDR3 and the CDR1, the CDR2, or FW 2. Indeed, the positioning of noncanonical C residues observed in the CDRs or FW 2 of chickens, combined with an uneven number of C residues in the CDR3, imply structural convergence with similar CDR features described for both camels and llamas (48). The observed placement of these noncanonical C residues is maintained across multiple species, irrespective of whether the VH-D-J domain is expected to pair with an L chain, in the case of cow, pig, and platypus IgG (63, 66, 67), or is independent in the case of cameld VH (48). A noticeable exception to this observation is the rabbit repertoire. While also limited predominantly to a single highly expressed VH germ line (68), rabbits exhibit a VH CDR3 that has a shorter length–distribution profile (mean, 13.6 aa) than chickens and humans and typically lacks cysteine incorporation (69). It is likely, however, that the rabbit has developed an alternative diversification mechanism that focuses on the Vγ rather than the VH, because cysteine residues are encoded within the CDR3 of some rabbit Vγ germ lines and cDNA sequences containing probable intraloop disulfide bonds have been described previously (70). Furthermore, functional target-specific Abs have been isolated from immune rabbits that exhibit considerably longer Vγ CDR3 usage than typically found in humans, as well as noncanonical cysteines in both the CDR1 and 2 of the VH (69, 70). Through mutagenesis studies, it has been suggested that inter- or intra-CDR disulfides may be essential for both V-domain stability and to maintain binding function in IgG and single-domain Abs (47, 71).

In our analysis of CDR3 length distribution, it was found that the vast majority of chicken B cells are producing loops ≥15 aa in length. It is intriguing to note that a CDR of 15 aa is approximately equivalent to a single V-D-J recombination event, if the consens D segment is combined with the heavily conserved positions donated by FW 3 and the J region (105–106 and 115–117, respectively). Longer CDR3 lengths must therefore be created by a combination of n-nucleotide addition, D-D joining events (21), or the formation of “mosaic CDR3 sequences” by recombination events with pseudogene-derived “D-like” sequences that are introduced during gene conversion (8). When combined with the prevalence of multicysteine containing CDR3s, these observations are strongly suggestive that B cells containing either V-D-D-J or (VYD)-D-J arrangements represent ≥60% of the total adult repertoire. This is surprising as V-D-D-J rearrangements have been described as being frequent in the embryonic bursa of fabricius but strongly selected against prior to the onset of gene conversion (6).

The isolation of 318 unique chicken scFv Abs from our immunized phage-display libraries, recognizing 12 separate protein Ags, allowed us to gain some clear insights into the Ag-specific repertoire. The selected repertoire appears to closely mimic that of the naive repertoire in CDR and FW mutation rates, CDR3 length distribution, and amino acid content. These are important observations in a practical sense, because they show that the selected repertoire may be a relatively accurate representation of the Ab repertoire in the immunized animal. One deviation from this overall picture is the observation that the selected repertoire exhibits a lower correlation between CDR3 length and cysteine content than is found in the naive repertoire. It is possible, however, that this phenomenon is driven by the influence of the E. coli expression machinery used in the phage–display process, which may select against clones with complex disulfide bonding profiles or free, unpaired cysteine residues.

The observations outlined above have allowed us the opportunity to define six structural types of chicken VH domains, based on their noncanonical cysteine content. These type definitions may aid further characterization of the chicken repertoire in future studies. Although types 1 and 2 dominate the repertoire in both the naive and selected sequence populations, the less prevalent members may still be of considerable importance. To illustrate that the less common types 3–6 may be a fully functional part of the Ag-specific response, we examined our selected repertoire for functional target-binding clones with rare motifs. Fab H2 was generated, which exhibited efficient, monomeric expression in E. coli and 64 nM affinity for IgM, despite containing four noncanonical cysteine residues in the CDRs. Because the H2 VH domain is a member of the rare type 5 family (<2% of the total naive repertoire), these data clearly illustrated that such sequences are produced during the response to foreign Ags and are not merely nonfunctional sequence artifacts. In addition, the successful expression of a chimeric Fab containing a CDR3 of 32 aa in E. coli suggests that the full extent of the chicken repertoire may be sampled using phage display.

All current evidence suggests that chickens use different amino acid architecture than humans and mice to generate novel Abs in their VH CDR3s, commonly using long CDR3 loops that are stabilized by disulfide bonding. We propose that this process creates CDR3s of great stability and structural complexity and is a mechanism for creating maximal diversity in a single-FW repertoire. The observed predictability of VH subtypes that can be derived from immune chickens, coupled with clearly limited Vκ, Iκ, and VL interface residue complexity, will aid the rapid and reproducible humanization of chicken Abs for therapeutic indications.

Acknowledgments
We thank Caroline Barrelle and John Steven for the provision of purified dogfish IgM protein, Thad Welch for assistance with data analyses, and members of Global Biotherapeutics Technologies, past and present, for the provision of target-specific chicken Ab sequences.

Disclosures
The authors have no financial conflicts of interest.

References
332

CHARACTERIZATION OF THE CHICKEN Ig VH REPERTOIRE


Supplemental figure 2

(A) Sequence analysis of CDR3 lengths from selected binders

(i) Chicken CDR3-H3 length 15  n=23

(ii) Chicken CDR3-H3 length 17  n=24
(iii) Chicken CDR3-H3 length 18  n=31

Shannon Entropy

Percent of Amino Acids

(iv) Chicken CDR3-H3 length 19  n=16

Shannon Entropy

Percent of Amino Acids
(v) Chicken CDR3-H3 length 20, n=21

Shannon Entropy

Percent of Amino Acids

Position

(vi) Chicken CDR3-H3 length 21, n=22

Shannon Entropy

Percent of Amino Acids

Position
(B) Sequence analysis of the corresponding unselected CDR3 lengths

(i) Chicken CDR3-H3 length 15  n=82

(ii) Chicken CDR3-H3 length 17  n=169
Chicken CDR3-H3 length 18  n=168

Chicken CDR3-H3 length 19  n=134
(v) Chicken CDR3-H3 length 20  n=106

(vi) Chicken CDR3-H3 length 21  n=139
Supplemental Figure Legends

Supplemental Figure 1. Sequence deviation from germline in the selected chicken V_H repertoire. Annotated V_H sequences from 318 unique target-binding clones were used to calculate % positional deviation from the germline sequence in the selected repertoire. The x-axis shows the germline sequence, with amino acid positions defined using the IMGT numbering scheme. The y-axis shows % deviation. The CDR3 is excluded for clarity (and covered separately in Supplemental Fig. 2A). Amino acid substitutions at each position are color coded.

Supplemental Figure 2. Plots of Shannon entropy (upper) and amino acid frequency (lower) for V_H CDR3 positions 105-117, for individual highly represented lengths (15-21 amino acids). A) Selected sequences (i-vi). B) Unselected sequences (i-vi). Percent amino acid content at each position is color coded as in the legend. Shannon entropy is plotted on the scale 0-4.