PROTEINS:
Structure, Function, and Bioinformatics

Conserved Amino Acid Networks Involved in Antibody Variable Domain Interactions

Journal:	*PROTEINS: Structure, Function, and Bioinformatics*
Manuscript ID:	Prot-00311-2008.R1
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Wang, Norman; Biogen Idec, Research Informatics Smith, William; Biogen Idec, Research Informatics Miller, Brian; Biogen Idec, Protein Engineering Aivazian, Dikran; Biogen Idec, Protein Biochemistry Lugovskoy, Alexey; Biogen Idec, Research Assay Reff, Mitchell; Biogen Idec, Discovery Oncology Glaser, Scott; Biogen Idec, Protein Engineering Croner, Lisa; Biogen Idec, Research Informatics Demarest, Stephen; Biogen Idec, Department of Protein Chemistry
Key Words:	Immunoglobulin variable domain, Ig-fold, V-class, covariation, antibody engineering

Running Title: RESIDUE COVARIATIONS IN ANTIBODY DOMAINS

Conserved Amino Acid Networks Involved in Antibody Variable Domain Interactions

Norman Wang,[†] William F. Smith,[†] Brian R. Miller, Dikran Aivazian, Alexey A. Lugovskoy, Mitchell E. Reff, Scott M. Glaser, Lisa J. Croner,[*] and Stephen J. Demarest[*]

Biogen Idec, San Diego, CA, USA

[†] N.W. and W.F.S. contributed equally.

[*] Correspondence to: Stephen Demarest and Lisa Croner, Biogen Idec, 5200 Research Place, San Diego, CA 92122, USA.

E-mail: stephen.demarest@biogenidec.com and lisa.croner@biogenidec.com

Engineered antibodies are a large and growing class of protein therapeutics comprising both marketed products and many molecules in clinical trials in various disease indications. We investigated naturally conserved networks of amino acids that support antibody V_H and V_L function, with the goal of generating information to assist in the engineering of robust antibody or antibody-like therapeutics. We generated a large and diverse sequence alignment of V-class Ig-folds, of which V_H and V_L domains are family members. To identify conserved amino acid networks, covariations between residues at all possible position pairs were quantified as correlation coefficients (ϕ-values). We provide rosters of the key conserved amino acid pairs in antibody V_H and V_L domains, for reference and use by the antibody research community. The majority of the most strongly conserved amino acid pairs in V_H and V_L are at or adjacent to the V_H-V_L interface suggesting that the ability to heterodimerize is a constraining feature of antibody evolution. For the V_H domain, but not the V_L domain, residue pairs at the variable-constant domain interface (V_H-C_H1 interface) are also strongly conserved. The same network of conserved V_H positions involved in interactions with both the V_L and C_H1 domains is found in camelid V_{HH} domains, which have evolved to lack interactions with V_L and C_H1 domains in their mature structures; however, the amino acids at these positions are different, reflecting their different function. Overall, the data describe naturally occurring amino acid networks in antibody Fv regions that can be referenced when designing antibodies or antibody-like fragments with the goal of improving their biophysical properties.

Key words: Immunoglobulin variable domain; Ig-fold; V-class; covariation; antibody engineering

INTRODUCTION

Antibodies are useful targeted therapeutics because of their ability to bind specific ligands with high affinity and specificity. Antibody variable domains (V_H in the heavy chain, V_L in the light chain), which provide the binding capability, may be purposely engineered to impart desired antigen recognition or binding affinity properties. Some designs have implemented recombinant production of isolated V_H-V_L domains (Fv region), providing researchers with more design flexibility than standard antibody therapeutics (*e.g.* the expression of the Fv region as a single polypeptide chain or "scFv"[1-5]). However, removal of the V_H-V_L domains from the quarternary structure of an antibody can lead to stability and solubility problems. Several mechanisms have been proposed to account for the generally poor biophysical behavior of scFvs and related designs, and include the intrinsic instability of the isolated domains, the weak affinity between V_H and V_L domains, and the absence of possibly stabilizing interactions with the antibody constant domains[6]. An understanding of the specific amino acids that mediate interactions between the V_H and V_L domains, and between the variable and constant domains, would enable improved designs of antibodies and antibody-like proteins.

Antibody variable domains are part of the immunoglobulin domain or "Ig-fold" superfamily. The Ig-fold superfamily is a large group of structurally related protein domains commonly found in mammalian cell surface proteins or in soluble extracellular signaling proteins[7]. Ig-fold domains consist of two β-sheets, each arranged in a "Greek-key" topology, that are packed tightly against one another and are generally supported by an intradomain disulfide bond. Depending on the number of strands in each β-sheet and the loop connections between the strands, the superfamily can be divided into several subfamilies including the C-, I- and V-classes[8]. Antibody variable domains are V-class Ig-folds and their constant domains are C-class Ig-folds (**Figure 1**). Ig-fold or Ig-fold-like domains are also present in cell adhesion proteins, integrins, allergens, T-cell receptors, major histocompatibility complexes, immunoglobulin receptors, and many other protein families with diverse functions.

The past decade has seen a significant increase in the number of publicly available Ig-fold sequences. Large databases of antibody variable domain and T-cell hypervariable domain Ig-fold sequences have been compiled[9]. Information from these databases has been instrumental in antibody humanization, affinity maturation, and the stabilization of single chain Fv (scFv) or other antibody constructs[5,10-13]. Antibody sequence databases generally influence antibody design by enabling frequency analyses at single amino acid positions (i.e., consensus modeling) that may be used for generating rational designs[12,14-16]. Recent studies with other protein domain superfamilies have extended sequence-based approaches by examining how amino acid *pairs* or *networks* may be conserved within subsets of a protein superfamily with related function or across diverse members of a protein superfamily. Such amino acid networks may define important structural or functional features of these protein domains[17-19]. These approaches, sometimes referred to as “covariation analyses,” track whether the presence (or absence) of a particular amino acid at one position correlates with the presence (or absence) of another amino acid at a second position within a multiple sequence alignment. While covariation analyses have been performed on several protein families (including SH3 domains, WW-domains, TPR-motifs, GPCRs, serine proteases, globins, viral coat proteins, and others[20-23]), very little has been described concerning covariation analyses of Ig-folds. The paucity of Ig-fold covariation data may stem from several factors, one being that large collections of Ig-fold sequences were, until recently, limited primarily to antibody sequences, particularly human and murine[24,25]. Also, accurate alignment of diverse members of large proteins (>100 amino acids) like Ig-fold domains is challenging and misalignments can limit the validity of covariation data[23].

Here we describe the application of covariation analyses to a high-quality, 3D-structure-based alignment of diverse V-class Ig-fold sequences. A diverse V-class Ig-fold sequence alignment was constructed, and covariations were quantified as correlation coefficients (ϕ-values[23]) for every amino acid pair (i.e., every residue combination found at all possible pairs of positions) in the alignment. The results serve as a rich repository of amino acid interactions conserved throughout Ig-fold evolution. The data reveal conserved residue networks that may support interactions between the V_H and V_L domains. The data also reveal that V_H domain networks involved in interactions with V_L domains are co-conserved with residue networks observed at the V_H-C_H1 junction suggesting that these two functional areas have co-evolved to support the overall quarternary antibody structure.

METHODS

Creation of structure-based Ig-fold alignments

Structures of Ig-fold proteins or Ig-fold domains from multi-domain proteins were gathered from the ASTRAL database[26,27], which contains domain structures matching the Structural Classification of Proteins (SCOP, Version 1.69) hierarchy. SCOP classifies the Ig-fold as a member of the "Beta Proteins, Immunoglobulin-like beta-sandwich fold, Immunoglobulins" superfamily[28]. PDB files of the V-class Ig-folds were downloaded using customized shell scripts. Each Ig-fold structure was inspected visually using Swissprot DeepView; sequences were removed from the study if they were erroneously categorized, incomplete (either missing residues due to a lack of electron density or domain swapped[29]), redundant (i.e., those with identical sequences), or obviously did not conform to the β-sandwich Ig-fold topology. Sequences of aberrant length (>2-times the standard deviation about the mean V-class length, 112.0±10.6 residues) were also removed. 702 structures were aligned using the Secondary Structure Matching (SSM) assisted implementation in the Schrödinger Prime structalign program[30-32]. Schrödinger Prime was used to generate structure-based V-class sequence alignments based on the proximity of each C_α atom subsequent to an all-to-all structure alignment, which minimized the average distance between all structural pairs. The alignment was most accurate in the regular β-strand regions and less accurate in the connecting loop regions due to variable loop lengths and structures.

Generation of a diverse V-class Ig-fold sequence alignment

A custom Hidden Markov Model (HMM) of the V-class Ig-folds was built from the structure-based sequence alignments. The HMM was created with the HMMER software package (version 1.8), using the "hmmbuild" and "hmmcalibrate" functions[33]. The HMM was used to find potential V-class Ig-fold sequences in the NR-database maintained at NCBI using the "hmmsearch" function. The output of this function ranked the hit sequences by their scores relative to our custom V-class HMM, and sequences with scores above a recommended threshold were retained as candidate members of the V-class dataset. The output also provided the number of "hits" per sequence (i.e., the number of Ig-folds within a contiguous gene sequence) and the exact residue positions of the hits. For sequences containing one or more candidate V-class sequences, the relevant subsequences were extracted from the full NR sequence using a custom Java executable. As an additional test to confirm that each sequence pulled from NR using our V-class HMMs belonged to the V-class Ig-fold subfamily, the custom shell script "pfamverify" using the HMM tool "hmmpfam" was applied to each Ig-fold candidate sequence[34]. Ig-clan HMMs (including V-, I-, C1-, C2-, and less specific Ig HMMs) were downloaded from the PFAM website. Sequences that scored lower with the PFAM V-class HMM than with other PFAM Ig-fold HMMs, and sequences whose score with the PFAM V-class HMM lay below recommended cutoffs (TC1 defined at the PFAM website) were removed. Thus, V-class Ig-fold sequences were retained only if their PFAM scores validated their Ig-Fold class assignments. The Ig-Fold sequences extracted from NR were aligned by our structure-based V-class HMM using the "hmmalign" function in the HMMER package. The resulting V-class dataset contained 48,696 sequences including those from both the SCOP 3D protein database and NR.

The resulting sequence collection was biased towards well-studied Ig-fold-containing proteins (i.e., human and murine V-class sequences frequently deposited in NR). To reduce the over-representation of these sequences, we developed a heuristic algorithm that eliminated sequences based on identity cut-off criteria. In brief, percent identities were calculated for all sequence pairs. Sequences were grouped into bins representing their maximum percent identity with any other sequence (i.e. 99% bin, 98% bin, 97% bin, etc.). Sequences within each bin were then ranked according to decreasing non-gap residue count, giving better ranks to sequences with fewer gaps. In each bin, sequences with an equal number of non-gap residues were ranked by Henikoff weights to filter out more common sequence types while preserving rare sequences with the goal of increasing diversity within the final datasets[35]. An identity cutoff of 80% was used for V-class sequences. This left 2,786 sequences, each with less than 80% identity to all other sequences, in the V-class dataset.

The resulting multiple sequence alignment contained many positions populated by gaps (> 50% gaps for most sequences). To eliminate this problem, columns that were not match states in the HMM were removed. This resulted in 144 remaining columns for our custom V-class alignment. Still, 354 sequences contained > 40% gaps. These sequences, which were generally incomplete, were removed from the alignment. The final V-class Ig-fold dataset contained 2,432 sequences. Virtually all the V-class sequences in the dataset were naturally occurring (non-engineered).

Correlation coefficient (ϕ-value) calculation

Covariation between amino acid pairs in multiple sequence alignments were calculated as correlation coefficients (ϕ-values), as described previously[23]. The calculations were encoded into a Java executable and run with Java Runtime Engine (JRE) version 1.4.2. ϕ-values were defined as

$$\phi(x_i y_j) = \frac{(x_i y_j * \overline{x}_i \overline{y}_j) - (x_i \overline{y}_j * \overline{x_i y}_j)}{\sqrt{(x_i y_j + \overline{x}_i y_j) * (x_i \overline{y}_j + \overline{x}_i \overline{y}_j) * (x_i y_j + x_i \overline{y}_j) * (\overline{x}_i y_j + \overline{x}_i \overline{y}_j)}}, \qquad (1)$$

where $x_i y_j$ is the number of times amino acids of type "x" or "y" are found in the same sequence at positions i and j, respectively, $\overline{x}_i \overline{y}_j$ is the number of times both amino acids are absent from the same sequence, $x_i \overline{y}_j$ is the number of times x is found present while y is absent, and $\overline{x}_i y_j$ is the number of times x is absent while y is present. This equation can be rewritten as:

$$\phi(x_i y_j) = \frac{(a * d) - (b * c)}{\sqrt{efgh}}, \qquad (2)$$

where a through h are given by the contingency table:

	x_i	$\bar{x}_i$	Total
y_j	a	b	e
$\bar{y}_j$	c	d	f
Total	g	h	

and a = $x_i y_j$, b = $\bar{x}_i y_j$, c = $x_i \bar{y}_j$, d = $\bar{x}_i \bar{y}_j$, e = a + b, f = c + d, g = a + c, and h = b + d. Particular residue pairs (specific combinations of residues at specific positions) were not considered unless they were observed in the alignments a minimum of 10 times.

Statistics

Statistical significance of the ϕ-values was evaluated with a chi-square (X^2) test, using Bonferroni-corrected *p*-values to adjust for multiple testing.

The X^2 test is often used to evaluate the significance of values observed in contingency tables of two dichotomous variables, such as the contingency table above. The equation for this use of X^2 can be written as

$$\chi^2 = \sum \left[\frac{(o_k - e_k)^2}{e_k} \right] \qquad (3)$$

where o_k stands for the observed frequency and e_k stands for the expected frequency in one cell of the table. X^2 is calculated by taking the sum of the squared and normalized differences between the observed and expected frequencies over all the cells. When expected frequencies are unknown, they can be estimated from observed frequencies and the equation becomes

$$\chi^2 = \frac{N*(ad-bc)^2}{efgh} \quad (4)$$

with *a* through *h* representing the same values as in equation (2) above, and *N* standing for the total number of samples. Comparing equations (2) and (4), it is evident that

$$\chi^2(x_i y_j) = \phi(x_i y_j)^2 * N$$

.

This relationship between X^2 and ϕ is useful because of the rich information available about the X^2 statistic, including tables of *p*-values for X^2 with specified degrees of freedom (df). However, before proceeding to use this relationship, we performed random simulations to confirm that it held for our dataset. We took our V-class sequence alignment, performed repeated (tens of thousands) random shuffling of the residues at each position (so that the residue frequencies at each position remained unchanged, but the correlations across positions were randomized), and computed ϕ-values for each randomization. We then calculated the probabilities of observing specific strong covariations by chance directly from these random simulations. In all cases examined, we found close agreement between the probabilities observed in the simulations and those calculated from X^2.

Having validated the use of X^2 to determine significance in our dataset, we converted our ϕ-values to X^2 using equation (4), and used standard X^2 tables (df=1) to find *p*-values. ϕ-values were calculated for the 186,171 amino acid pairwise combinations that occurred at least 10 times in our V-class alignments. To correct for multiple testing, we used a Bonferroni-corrected *p*-value[36] as our criterion for significance, striving for significance at the true $p < 0.0001$ level. For positive correlations, this corresponded to ϕ-values > 0.1237. Use of the Bonferroni correction along with this strict *p* criterion gave a very conservative list of amino acid pairs with significant ϕ-values, and greatly reduced our chances of finding false positives. Even with this conservative approach, 13,796 significant positive correlations (see Results) were observed.

Results

Alignment Quality and Diversity

Information about protein 3D structures can significantly improve the quality of multiple sequence alignments[37]. As described in the Methods, we compiled 3D structures of V-class Ig-folds from SCOP, and generated a structure-based multiple sequence alignment of these V-class domain sequences. A custom HMM was built from this structure-based alignment and used to align additional Ig-fold sequences from the NR sequence database. Here we discuss the quality and the diversity of the resulting alignment

The quality of the alignment guided by our structure-based HMM was evaluated by examining whether the HMM properly aligned sequences that, though disparate in residue identity, are known to form the same part of an Ig-fold 3D structure. As expected, residues that make up the β-strands of both V_H and V_L domains were well aligned, while alignments of residues in the loop regions, whose structures are more variable, often contained many gaps. We also looked to see whether the HMM properly aligned the consensus sequences of antibody V_H and V_L chains, in the V_H-V_L interface region. The heterodimeric structure of the interface is highly symmetrical (both V_H and V_L domains use the same face of their Ig-fold to create the heterodimer). Thus, residues buried in the interface should align in 3D, despite differences between the amino acids of V_H and V_L Ig-folds at these positions. We used a 1.8 Å crystal structure of an antibody Fab from our lab (Jacob Jordan *et al.*, manuscript in preparation) to determine the residues of both V_H and V_L that bury surface area at the interface using the program MOLMOL[38]. The V_H and V_L residue positions buried at the interface mapped to the same residue positions within the multiple sequence alignment, even though the amino acid identities at these positions vary.

Covariation analyses are most successful when applied to sequence datasets that are highly diverse[23,39-41]. From a practical standpoint, ϕ-value correlation coefficients increase when there are many instances of both the presence and absence of a conserved amino acid pair across a sequence alignment (see numerators of Equations 1 and 2).

Highly diverse sequence sets are more likely to contain sequences both with and without the pair than are datasets of highly related sequences. Since our goal was to investigate conserved amino acid networks in V_H and V_L domains, it was therefore important that our V-class Ig-fold dataset contain a background of Ig-fold family members that are not evolutionarily constrained to perform the same function as V_H and V_L domains. Additionally, covariation signals pertaining to polar interactions have been shown to be stronger in datasets with moderate to high evolutionary distances between sequences[40]. Protein interfaces, in which Ig-folds frequently appear, often rely more heavily on polar interactions than do protein cores[42], providing another reason for generating a V-class Ig-fold sequence dataset that was highly diverse. The unfiltered sequence set from NCBI contained ~50,000 sequences highly biased towards immunoglobulin variable domains.

An 80% identity filter was used to reduce bias towards over-represented V-class families, thus promoting diversity. After filtering, the dataset contained 2,432 V-class sequences. Members of the V-class dataset could be divided into three functional categories: (1) 50% were immunoglobulin variable genes (including both V_H and V_L antibody domains); (2) 16% were T-Cell Receptor V-class genes; and (3) 34% were V-class genes derived from diverse functional families, each of which comprised less than 5% of the V-class sequences. The sequences were derived from species ranging from cartilaginous fish to primates. There was a bias towards human immunoglobulin variable domain sequences with 574 of the total 2,432 V-class sequences being human V_H (484 of 993 V_H) or human V_L (90 of 187 V_L). Examples of other species contributing V_H and V_L sequences to the V-class database include mouse (44), cow (16), camel (174), llama (83), macaque (17), and chicken (9). Despite the bias towards human V_H and V_L, the average sequence identities within V_H and V_L subgroups were low – 41±1% and 29±1%, respectively. The distribution of germline V_H and V_L sequences passing the 80% identity filter roughly matched the naturally observed distribution of variable domain sequences[43] suggesting that variable gene subclasses were similarly diverse and fairly represented in the sequence dataset**.** Positional entropy calculations using the final V-class dataset demonstrate the much higher positional diversity with the V-class dataset compared to antibody V_H or V_L datasets that may be used for consensus analyses[15,44] (**Supplemental Figure 1**).

Correlation coefficients (ϕ-values) between residues of V-class Ig-folds

We adopted a previously described method – the use of ϕ-value correlation coefficients – for quantifying covariations of residue pairs within sequence alignments[23]. **Figure 2** shows the number and distribution of ϕ-values calculated for amino acid pairs that were observed ≥ 10 times within the V-class Ig-fold alignment. Positive ϕ-values represent positive correlations (the presence of one amino acid at one site in the alignment is correlated with the presence of another amino acid at another site). As ϕ-values move from 0 to +1, the strength of the correlation between the two amino acids increases. Negative ϕ-values represent negative correlations (the presence of one amino acid at one site in the alignment is correlated with the *absence* of another amino acid at a second site), which become stronger as ϕ-values move towards –1.0. Our statistical analyses (see Methods) showed that ϕ-values greater than 0.1237 were significant (*p*-values < 0.0001). This conservative statistical estimate revealed 13,796 significant positive covariations. However statistical significance does not entirely indicate the strength of the covariations. After careful evaluation of the data, we designated ϕ-values between 0.25 and 0.5 as moderate covariations, and ϕ-values greater than 0.5 as strong covariations. Of the 4.1 million possible amino acid combinations within the V-class Ig-fold alignment, 3212 (0.078%) had ϕ-values ≥ 0.25 and 133 (0.003%) had ϕ-values ≥ 0.5.

As validation of our covariation analysis, we examined the data in two ways to confirm that expected patterns were present in the results. First, we investigated the relationship between ϕ-values and distance between amino acid pairs in 3D space. Previous studies have shown that strongly covarying amino acid pairs often involve positions that are near each other in 3D structures – although the trend has invariably been reported as weak[23,40]. To see if the same pattern was present in our data, we plotted the ϕ-values ≥0.3 against the distance between the amino acid pairs in 3D space using our Fab crystal structure. As reported by others, we found a weak but significant relationship between ϕ-value and the 3D proximity of the pair members (data not shown). Second, we investigated whether our covariation data recapitulated an amino acid network known to exist within a particular subset of V-class Ig-folds. The example we chose was a set of five residues (residues 6-10) at the N-terminus of human/murine IgG V_H domains. These residues adopt different backbone conformations depending on the presence of specific amino acid pairs[4511]. Four conformations exist, depending on whether glutamic acid or glutamine is present at V_H position 6. Q6 is not well-conserved within V_H domains (it is also found commonly in V_L domains) and does not have significant covariations. However, E6 is highly conserved in variable heavy chains (V_H2, V_H3, and V_H4 subclasses in particular). E6 correlations with residues 7-10 yield some of the highest ϕ-values of the covariation dataset (S7=0.51; G8=0.59; G9=0.65; and G10=0.53; **Table 1**). These high correlations among residues 6-10 are consistent with the known involvement of these residues in determining the N-terminal β-strand conformation of IgG V_H domains. High ϕ-values were also found among other positions known to be structurally important including V_H subfamily-dependent core positions 18, 63, 67, and 82 that have been described previously[46].

Covariation results broadly applied to V_H and V_L domains

In this section, we describe patterns evident upon examining the strongest covariations found within V_H and V_L domains. The V_H and V_L amino acid pairs with the highest ϕ-values are listed in **Table 1**. The locations of the residues involved in the strongest covariations from **Table 1** were mapped onto our in-house Fab 3D structure (V_H, **Figure 3B,C;** V_L, **Figure 4B,C**). Interestingly, most of the residues contributing to the strongest covariations were found at or very near to the V_H-V_L interface (**Table 1,** red in **Figure 3B,C, Figure 4B,C**), indicating a conserved amino acid network supporting this interface. V_H domains also appear to conserve an amino acid network near the variable-constant domain (V_H-C_H1) interface (**Table 1,** orange in **Figure 3B,C**). This latter network, however, is not observed in V_L domains (**Table 1**). Other strongly conserved residue pairs involved the N-terminal region of V_H (residues 6-10), and a few buried hydrophobic residues known to be highly subtype dependent (both described above)[23,46].

For an alternative overview, we also compiled tables of V_H or V_L domain residues that had the most covariations (ϕ-values ≥ 0.25) with other amino acids in their respective domains (V_H, **Table 2;** V_L, **Table 3**), regardless of the covariation strengths. Residues with the most covariations do not map to any single region of V_H (**Supplemental Figure 2**) or V_L (**Supplemental Figure 3**), which is not surprising given the inclusion of many residues from weakly conserved networks. However, this analysis revealed an abundance of conserved networks evident with less stringent constraints on significant ϕ-values. We did note that residues involved in the most covariations mapped predominately to the β-sheet regions. Some covariations involve amino acids in the loop regions, but these are seen less frequently and likely reflect poorer alignment statistics in the loop regions, rather than a lack of conserved networks in the loop regions.

Analysis of the interface between antibody V_H and V_L domains

To further investigate V_H and V_L residues that may play a role in heavy and light chain association, we examined which V_H or V_L amino acids covary with amino acids that make direct inter-domain contacts at the V_H-V_L interface. Framework V_H and V_L positions (in Kabat numbering[47]) that bury surface area at the V_H-V_L interface are: in V_H 35, 37, 39, 44, 45, 47, 50, 91, and 103; and in V_L 36, 38, 43, 44, 46, 49, 87, and 98. These positions are mapped onto the surfaces of V_H and V_L in **Figure 3A** and **Figure 4A**, respectively, and onto a V-class sequence alignment (using our in-house V-class HMM) in **Figure 5** (red letters in framework regions). The CDR3 loops of both V_H and V_L also bury surface area between the two domains to form a continuous antigen-binding surface. However, the HMM profile eliminated the CDR3 loops of both the V_H and V_L domains from the alignments due to the inability to define consistent CDR3 profiles. The absence of CDR3 data was deemed unimportant, as few strong intra-domain covariations would be predicted to arise from the highly variable CDR3 loops.

The V_H and V_L amino acids with the most covariations (ϕ-value ≥ 0.25) to the interface residues above are listed in **Table 4** (V_H) and **Table 5** (V_L). The entries in these tables are sorted by (1) the difference between the entry's average ϕ-value with interface residues versus its overall average ϕ-value, and (2) the number of the entry's covariations with interface residues. Amino acids near the top of the tables are perceived to have a greater role in supporting the V_H-V_L interface. The residues from **Tables 4** and **5** are highlighted in the sequence alignment in **Figure 5** (yellow or green highlights for V_H or V_L respectively) and have been mapped to the surfaces of the V_H and V_L domains (V_H, **Figure 3D,E**; V_L, **Figure 4D,E**). Several interface residues themselves rank highly as do many of the residues adjacent in primary sequence. $W47_{VH}$, which incidentally is the V_H framework residue that buries the second highest amount of surface area at the interface, appears to be the central node of the V_H-side of the interface network based on the number and strength of its covariations with other interface residues – even though it is not at the center of the residues that make direct contact within the V_H-V_L interface (**Figure 3D**). $Y36_{VL}$ and $P44_{VL}$ appeared to be the central nodes of the V_L–side of the interface network based on the same criteria (**Figure 4D**). $P14_{VH}$, $T87_{VH}$, and $L108_{VH}$ – all near the V_H-C_H1 interface – also covary strongly with the V_H interface residues, suggesting that the maintenance of both the Fv and V_H-C_H1 interfaces are co-conserved traits. In contrast, covariations were weak for V_L residues in the proximity of the V_L-C_L interface.

Four V_H residues – V37, G44, L45, and W47 – form a patch on the surface of V_H that interacts with V_L (**Figure 3A-C**). Each of these four residues has 30 or more ϕ-values > 0.25 with other V_H residues; however, the ϕ-values between these four residues are collectively the strongest observed for each of these residues, with an average ϕ-value of 0.6 (see Table 1, Table 4). The sidechains do not pack directly against one another or appear to interact strongly, suggesting that the residues covary for functional reasons – in this case enabling immunoglobulin heavy chain V_H domains to interact with immunoglobulin light chain V_L domains. In this context, the sidechains of W47, L45, and V37 together form a roughly flat hydrophobic surface that matches well with residues P44, F87, and F98 of V_L (**Figure 6**). A fifth V_H residue – W103, one of the only other V_H residues burying surface area at the interface with V_L – also covaries with the four V_H residues, though with weaker ϕ-values averaging 0.38.

Two other V_H residues – R38 and E46 – strongly covary with all five of the V_H residues discussed above, with average ϕ-values of 0.42 and 0.47, respectively. The ϕ-value between R38 and E46 is also strong (ϕ=0.56, **Table 1**, **Figure 3A-C**). R38 is almost completely buried by E46 in the interior of V_H and its guanidinyl group forms a specific salt bridge with E46's carboxyl group (**Figure 6**). The charge burial of R38 is supported by other interactions with D90 and K/Q43. The E46 sidechain does not contact any V_H residues at the V_H-V_L interface; thus E46 is likely important for creating optimal surface topology and perhaps an electrostatic component important for V_L binding.

On the V_L side of the interface, the most strongly covarying interface residues are Y36 and P44 (the positional equivalents of V_H residues V37 and L45 – **Table 1**, **Figure 4A-C**) with average ϕ-values with other V_L interface residues of 0.43 and 0.42, respectively. Also, covarying with these two residues are Q37, A43, L46, and F98 (positional equivalents of V_H R38, G44, W47, and W103, **Figure 5**), but with lower average ϕ-values (0.36, 0.31, 0.33, and 0.35, respectively). The ϕ-values between residues within the V_L domains were lower on average than those between V_H residues; this can be explained by the number of V_L sequences in the alignment being smaller (half the number of the V_H sequences) and more heterogeneous (containing both V_κ and V_λ domains). Similar to what was observed for V_H interface residues, a cluster of V_L residues – Y36, A43, P44, L46 and F98 – do not pack directly against one another, but instead combine to form the V_H binding surface. Unlike V_H residues R38 and E46 that form a salt bridge with one another, V_L positions 37 and 45 only weakly covary with one-another (Q37 and K45 have a ϕ-value = 0.34, **Figure 5**). In general, the V_L residues important for V_H-binding are well conserved for both kappa and lambda light chain variable domains.

The strongly correlated residues at the Fv interface of both V_H and V_L thus appear to form conserved networks that enable recognition between the domains (**Figure 6**). The interaction surface is fairly flat between the two domains. V_H_L45 and W103 insert themselves into a small groove created by V_L residues Y36, P44, L46, F87, and F98. The small size of V_L_P44 helps create the groove into which V_H_L45 and V_H_W103 intrude. In addition, the relatively small V_H V37 sidechain creates a cavity on the hydrophobic surface of V_H into which the sidechain of V_L F98 inserts (**Figure 6**).

Comparison of conserved residues networks of V_H domains and camelid V_{HH} domains lacking light chain interactions

While most antibody V_H sequences associate with light chain V_Ls, a subset of camelid variable heavy chain domains, denoted V_{HH} domains, are expressed naturally and function in the absence of both a light chain and a C_H1 domain[48]. V_{HH} domains are also substantially more soluble than V_H domains. The discovery of these simple and soluble V_H-like domains has had an enormous impact on antibody engineering because they represent potentially more facile reagents for protein design and discovery than traditional antibodies (which require combinations of heavy and light chains for function, stability, and solubility[49]). We therefore investigated whether conserved residue networks differ between standard V_H and camelid V_{HH} domains. We expected to observe such differences at the positions in V_H that serve to support the V_H-V_L interface. Our results revealed 32 significant covariations involving identical positions within V_H and V_{HH} domains; however, the 32 covarying pairs contained different amino acids for V_H versus V_{HH} domains (**Table 6**). Among these contrasting residues are a tetrad of amino acids that have been previously reported to differentiate V_H from V_{HH} domains: V37F, G44E, L45R, and W47G[50]. Substitution of this tetrad of camelid amino acids into V_H domains does not entirely impart them with the solubility of V_{HH} domains; CDR3 composition and other factors have also been shown to be important[51-55]. Our covariation results reveal additional framework residue positions, outside the tetrad described above, that naturally distinguish V_H from V_{HH} domains. These residues are at positions 13, 14, 33, 49, 63, 74, 82, 83, and 108. Solubilizing mutations at residues 74 and 108 have been reported[56]. Most of these positions are involved in networks surrounding the V_H-V_L or V_H-C_H1 interfaces (**Table 6, Figure 3**), as expected. A consensus camelid V_{HH} sequence derived from the ~50 diverse camelid sequences in the V-class alignment was included in **Figure 5** to highlight the positions of these observed differences between V_H and V_{HH} domains.

A natural human V_H raised against hen egg-white lysozyme was also demonstrated to be soluble in a monomeric form, similar to camelid domains (although the domain presumably maintains its ability to associate with V_L)[51,57]. This independently soluble anti-HEWL V_H domain contains yet another set of non-standard V_H amino acids. Many of these residues are involved in the Fv interaction network and one is proximal to the V_H-C_H1 interface: D27, D32, K39, K44, Y47, H59, K63, S68, and T108 (**Figure 5**).

Together with the V_{HH} results, it appears that multiple and independent amino acid networks may impart solubility to V_H and V_{HH} domains.

Discussion

Despite an enormous amount of research involving antibodies and antibody-like therapeutics, very little use has been made of covariation analyses to investigate functional features of antibody domains. A study by Altschuh and coworkers[58,59] investigated covariations across murine and human germline V_H or V_L sequences, with the goal of defining positions within each germline subclass that use mutually exclusive framework amino acid pairs to influence the structural conformations of CDR loops. Our study had a different goal, to use covariation analyses for determining naturally occurring amino acid networks, in antibody variable domains, that are generally important for antibody structure and function. Towards this end, we felt it necessary to generate covariation data using a larger and more diverse set of V-class Ig-fold sequences.

Our results show that the most strongly conserved amino acid networks in antibody V_H and V_L domains are found at the interface between V_H and V_L, suggesting that preservation of this interface may be a factor influencing antibody evolution. Interestingly, a small network of amino acids near the V_H-C_H1 interface is also highly conserved. However, this network is not observed for residues near the V_L-C_L interface. Biophysical studies with light chains in isolation have shown that the V_L and C_L domains do not influence the unfolding kinetics or thermodynamics of one another, suggesting that the interaction between the two domains is weak[60,61]. Alternately, Fabs (consisting of V_H, C_H1, V_L, and C_L domains) often show concerted unfolding reactions[44]. A viable explanation for the concerted unfolding reactions of Fabs compared to light chains in isolation is that the C_H1 and C_L domains act as ideal linkers for the V_H and V_L domains and vice versa[61]. It may also be that stronger interactions between V_H and C_H1, compared to V_L and C_L, promote cooperative unfolding of all four Fab domains[44]. The covariation data described in this report demonstrate that several V_H residues at the V_H-C_H1 junction are involved in a strongly conserved network of amino acids and suggest that Variable-

Constant domain interactions may be more important for immunoglobulin heavy chains than for immunoglobulin light chains.

Conclusions

In summary, we performed covariation analyses using a large, high-quality, and diverse alignment of V-class Ig-fold sequences. The data were used to discover antibody variable domain amino acid networks that are evolutionarily conserved. Mapping the most highly conserved V_H and V_L networks to their structures revealed that the networks cluster near the V_H-V_L interface and near the V_H-C_H1 interface, demonstrating the importance of preserving these interfaces during the evolution of antibody sequences.

These covariation data serve as a powerful tool for antibody (and Ig-fold domain) engineering. Insights from covariation analysis have improved our ability to rationally design more stable scFvs (**Supplementary Figure 4**). Most scFvs contain the majority of the residues within the V_H and V_L the conserved networks revealed by the covariation data. Our initial approach for stabilizing scFvs has been to find gaps or holes within these existing networks that can be rectified by mutagenesis. Many of these changes have improved the thermal unfolding midpoint (T_M) of scFv V_H or V_L domains by 1 – 12 ºC Miller *et al.*, manuscript in preparation). Stabilization of scFvs has enabled their use as building blocks that can be appended to full-length IgG molecules to produce stable bispecific IgG-like antibodies for consideration in clinical applications[62].

The covariation data described here may also be useful for engineering other aspects of V-class Ig-fold proteins, such as soluble V_H domains with conserved V_{HH} amino acid networks. The approach can be extended to other Ig-fold domains, such as C- and I-class, to identify amino acid networks supporting structure and function of other immunoglobulin or cell-surface receptor domains.

References

1. Bird R, Hardmann K, Jacobson JW, Johnson S, Kaufman BM, Lee S, Lee T, Pope SH, Riordan GS, Whitlow M. Single-chain antigen-binding proteins. Science 1998;242:423-426.
2. Huston J, Levinson D, Mudgett-Hunter M, Tai M, Novotny J, Margolies MN, Ridge RJ, Bruccoleri RE, Haber E, Crea R, Oppermann H. Protein engineering of antibody binding sites: recovery of specific activity of an anti-digoxin single-chain Fv analogue produced in Escherichia coli. Proc Natl Acad Sci USA 1988;85:5879-5883.
3. Glockshuber R, Malia M, Pfitzinger I, Plückthun A. A comparison of strategies to stabilize immunoglobulin Fv-fragments. Biochemistry 1990;29:1362-1367.
4. Brinkmann U, Reiter Y, Jung SH, Lee B, Pastan I. A recombinant immunotoxin containing a disulfide-stabilized Fv fragment. Proc Natl Acad Sci USA 1993;90:7538-7542.
5. Wörn A, Plückthun A. Stability engineering of antibody single-chain Fv fragments. J Mol Biol 2001;305:989-1010.
6. Demarest S, Glaser SM. Antibody therapeutics, antibody engineering, and the merits of protein stability. Curr Opin Biotechnol 2008;11:675-687.
7. Bork P, Holm L, Sander C. The Immunoglobulin Fold. J Mol Biol 1994;242:309-320.
8. Williams A. The immunoglobulin superfamily--domains for cell surface recognition. Annu Rev Immunol 1988;8:381-405.
9. Lefranc M, Giudicelli V, Ginestoux C, Bodmer J, Muller W, Bontrop R, Lemaitre M, Malik A, Barbie V, Chaume D. IMGT, the international ImMunoGeneTics database. Nucleic Acids Res 1999;27:209-212.
10. Carter P, Merchant AM. Engineering antibodies for imaging and therapy. Curr Opin Biotechnol 1997;8:449-454.
11. Ewert S, Honegger A, Plückthun A. Stability improvement of antibodies for extracellular and intracellular applications: CDR grafting to stable frameworks and structure-based framework engineering. Methods 2004;34:184-199.
12. Steipe B. Consensus-based engineering for protein stability: from intrabodies to thermostable enzymes. Methods Enzymol 2004;388:176-186.
13. Presta L. Engineering antibodies for therapy. Curr Opin Biotechnol 2002;74:237-256.
14. Davidson A. Multiple sequence alignment as a guideline for protein engineering strategies. Methods Mol Biol 2004;340:171-181.
15. Demarest S, Chen G, Kimmel BE, Gustafson D, Wu J, Salbato J, Poland J, Elia M, Tan X, Wong K, Short J, Hansen G. Engineering stability into *Escherichia coli* secreted Fabs leads to increased functional expression. Protein Engng Des Select 2006;19:325-336.
16. Demarest S, Rogers J, Hansen G. Optimization of the antibody C_H3 domain by residue frequency analysis of IgG sequences. J Mol Biol 2004;335:41-48.
17. Altschuh D, Vernet T, Berti P, Moras D, Najai K. Coordinated amino acid changes in homologous protein families. Protein Engng 1988;2:193-199.
18. Altschuh D, Lesk AM, Bloomer AC, Klug A. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. J Mol Biol 1987;193:693-707.

19. Valencia A, Pazos F. Computational methods for the prediction of protein interactions. Curr Opin Struct Biol 2002;12:368-373.

20. Socolich M, Lockless SW, Russ WP, Lee H, Gardner KH, Ranganathan R. Evolutionary information for specifying a protein fold. Science 2005;437:512-518.

21. Süel G, Lockless SW, Wall MA, Ranganathan R. Evolutionary conserved networks of residues mediate allosteric communication in proteins. Nature Struct Biol 2003;10:59-69.

22. Magliery T, Regan L. Beyond consensus: statistical free energies reveal hidden interactions in the design of a TPR motif. J Mol Biol 2004;343:731-745.

23. Larson S, Di Nardo AA, Davidson AR. Analysis of covariation in an SH3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions. J Mol Biol 2000;303:433-446.

24. Pollock D, Taylor WR. Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. Protein Engng 1997;10:647-657.

25. Govindarajan S, Ness JE, Kim S, Mundorff EC, Minshull J, Gustafsson C. Systematic variation of amino acid substitutions for stringent assessment of pairwise covariation. J Mol Biol 2003;328:1061-1069.

26. Brenner S, Koehl P, Levitt M. The ASTRAL compendium for sequence and structure analysis. Nucl Acids Res 2000;28:254-256.

27. Chandonia J, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE. The ASTRAL compendium in 2004. Nucl Acids Res 2004;32:D189-D192.

28. Murzin A, Brenner, SE, et al. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536-540.

29. Liu Y, Eisenberg D. 3D domain swapping: as domains continue to swap. Protein Sci 2002;11:1285-1299.

30. Lassmann T, Sonnhammer EL. Quality assessment of multiple alignment programs. FEBS Lett 2002;529:126-130.

31. Yang A, Honig B. An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. J Mol Biol 2000;301:665-678.

32. Jennings A, Edge CM, Sternberg MJ. An approach to improving multiple alignments of protein sequences using predicted secondary structure. Protein Engng 2001;14:227-231.

33. Eddy S. Profile hidden Markov models. Bioinformatics 1998;14:755-763.

34. Finn R, Mistry J, Schuster-Böckler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer ELL, Bateman A. Pfam: clans, web tools and services. Nucl Acids Res 2006;34:D247-D251.

35. Henikoff S, Henikoff JG. Position-based sequence weights. J Mol Biol 1994;243:574-578.

36. Miller RJ. Simultaneous statistical inference. Springer, editor; 1981. 299 p.

37. Wrabl J, Grishin NV. Gaps in structurally similar proteins: towards improvement of multiple sequence alignments. Proteins: struct Funct Genet 2004;54(71-87).

38. Koradi R, Billeter M, Wüthrich K. MOLMOL: a program for display and analysis of macromolecular structures. J Mol Graph 1996;14:51-55.

39. Gobel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins: Struct Funct Genet 1994;18:309-317.

40. Chelvanayagam G, Eggenschwiler A, Knecht L, Gonnet GH, Benner SA. An analysis of simultaneous variation in protein structures. Protein Engng 1997;10:307-316.

41. Neher E. How frequent are correlated changes in families of protein sequences? Proc Natl Acad Sci USA 1994;91:98-102.

42. Jones S, Marin A, Thornton JM. Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. Protein Engng 2000;13:77-82.

43. Goigou V, Cuisinier AM, Tonnelle C, Moinier M, Fougereau M, Fumoux F. Human immunoglobulin VH and VK repertoire revealed by in situ hybridization. Mol Immunol 1990;27:935-940.

44. Garber E, Demarest, SJ. A broad range of Fab stabilities within a host of therapeutic IgGs. Biochem Biophys Res Commun 2007;355:751-757.

45. Honegger A, Plückthun A. The influence of the buried glutamine or glutamate residue in position 6 on the structure of immunoglobulin variable domains. J Mol Biol 2001;309:687-699.

46. Honegger A. Engineering antibodies for stability and efficient folding. In: Chernajovsky Y, Nissim, A., editor. Therapeutic Antibodies Handbook of Experimental Pharmacology. Volume 181. Berlin: Springer-Verlag; 2008.

47. Wu T, 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.

48. Hamers-Casterman C, Atarhouch T, Muyldermans S, Robinson G, Hamers C, Songa EB, Bendahman N, Hamers R. Naturally occurring antibodies devoid of light chains. Nature 1993;363:446-448.

49. Holliger P, Hudson PJ. Engineered antibody fragments and the rise of single domains. Nature Biotechnol 2005;23:1126-1136.

50. Reichmann L, Muyldermans s. Single domain antibodies: comparison of camel VH and camelised human VH domains. J Immunol Methods 1999;231:25-38.

51. Holt L, Herring C, Jespers LS, Woolven BP, Tomlinson IM. Domain antibodies: proteins for therapy. Trends Biotechnol 2003;21:484-490.

52. Desmyter A, Transue TR, Ghahroudi MA, Thi MH, Poortmans F, Hamers R, Muyldermans S, Wyns L. Crystal structure of a camel single-domain VH antibody fragment in complex with lysozyme. Nature Struct Biol 1996;3:803-811.

53. Decanniere K, Desmyter A, Lauwereys M, Ghahroudi MA, Muyldermans S, Wyns L. A single-domain antibody fragment in complex with RNase A: non-canonical loop structures and nanomolar affinity using two CDR loops. Struct Fold Des 1999;7:361-370.

54. Spinelli S, Frenken L, Bourgeois D, de Ron L, Bos W, Verrips T, Anguille C, Cambillau C, Tegoni M. The crystal structure of a llama heavy chain variable domain. Nature Struct Biol 1996;3:752-757.

55. Barthelemy P, Raab H, Appleton BA, Bond CJ, Wu P, Wiesmann C, Sidhu SS. Comprehensive analysis of the factors contributing to the stability and solubility of autonomous human VH domains. J Biol Chem 2008;283:3639-3654.

56. Tanha J, Nguyen T-D, Ng A, Ryan S, Ni F, MacKenzie R. Improving solubility and refolding efficiency of human V_Hs by a novel mutational approach. Protein Engng Des Select 2006;19:503-509.

57. Li Y, Li H, Smith-Gill SJ, Mariuzza RA. Three-dimensional structures of the free and antigen-bound Fab from monoclonal antilysozyme antibody HyHEL-63. Biochemistry 2000;39:6296-6309.

58. Choulier L, Lafont V, Hugo N, Altschuh D. Covariance analysis of protein families: the case of the variable domains of antibodies. Proteins: Struct Funct Genet 2000;41:475-484.

59. Hugo N, LaFont V, Beukes M, Altschuh D. Functional aspects of co-variant surface charges in an antibody fragment. Protein Sci 2002;11:2697-2705.

60. Rowe E, Tanford C. Equililbrium and kinetics of the denaturation of a homogeneous human immunoglobulin light chain. Biochemistry 1973;12:4822-4827.

61. Röthlisberger D, Honegger A, Plückthun A. Domain interactions in the Fab fragment: A comparative evaluation of the single-chain Fv and Fab format engineered with variable domains of different stability. J Mol Biol 2005;347:773-789.

62. Glaser S, Demarest S, Miller BR, Wu X, Snyder WB, Wang N, Croner LJ; Biogen Idec MA Inc., assignee. Stabilized polypeptide compositions. PCT/US2008/0050370. 2007.

63. Honegger A, Plückthun A. Yet another numbering scheme for immunoglobulin variable domains: an automatic modeling and analysis tool. J Mol Biol 2001;309:657-670.

Figure Captions

Figure 1. Diagrams of an immunoglobulin and its Fv domain. A. Schematic diagram of an IgG antibody. The variable domains which compose the antigen-binding or Fv-region are shown in red and the constant domains are shown in blue. The variable domains are V-class Ig-folds, while the constant domains are C-class Ig-folds, which are highly similar to V-class Ig-folds, but lack two additional β-strands commonly found in V-class structures. **B.** Ribbon diagram of an antibody Fv-region consisting of a variable domain from the immunoglobulin heavy chain (V_H-blue) and a variable domain from the immunoglobulin light chain (V_L-red). The complementarity determining regions (CDRs) of the V_H (shown in green) and the V_L (shown in orange) comprise the antigen-binding site.

Figure 2. Distribution of ϕ-values calculated for the V-class alignment. There are 4,118,400 ($20 * 20 * \sum_{n=1}^{144} n-1$) possible amino acid pairings within the V-class sequences. Of these possible pairings, 1,098,890 actually exist within the sequence database (i.e., some amino acids pairing are not observed across columns of the alignment). The histogram shows the distribution of ϕ-values from the 186,171 pairings that occur at least 10 times. The 13,796 ϕ-values greater than 0.1237 were considered statistically significant, using a conservative statistical approach (see text).

Figure 3. Covariations mapped to surface representations of an antibody V_H domain derived from an in-house Fab structure. A. Surface representation of a V_H domain. Residues that bury >40 $Å^2$ at the Fv interface are shown in red and those that bury between 30-40 $Å^2$ are shown in orange. In **B.-E.**, residues colored grey (CDR3 residues as well as four residues at the C-terminus) were not match states in the HMM-derived V-class alignment and were not evaluated in this study. **B.** and **C.** V_H residues from amino acid pairs with the highest ϕ-values from **Table 1** were mapped to the V_H surface: red = proximal to the V_H-V_L interface; orange = proximal to the V_H-C_H1 interface; and purple = distant from the two interfaces. **D.** and **E.** V_H residues from **Table 4** that display multiple

covariations (ϕ-value > 0.25) with V_H-V_L interface residues with greater than average ϕ-values are mapped onto the V_H surface in red. Residues from **Tables 1** and **4** that are completely buried in the interior of the structure are not shown.

Figure 4. **Covariations mapped to surface representations of an antibody V_L domain derived from an in-house Fab structure. A.** Surface representation of a V_L domain. Residues that bury >40 Å^2 at the Fv interface are shown in red and those that bury between 30-40 Å^2 are shown in orange. In **B.-E.**, residues colored grey (CDR3 residues as well as four residues at the C-terminus) were not match states in the HMM-derived V-class alignment and were not evaluated in this study. **B.** and **C.** V_L residues from amino acid pairs with the highest ϕ-values from **Table 1** were mapped to the V_L surface: red = proximal to the V_H-V_L interface; orange = proximal to the V_L-C_L interface; and purple = distant from the two interfaces. **D.** and **E.** V_L residues from **Table 5** that display multiple covariations (ϕ-value > 0.25) with V_H-V_L interface residues with greater than average ϕ-values are mapped onto the V_L surface in red. Residues from **Tables 1** and **5** that are completely buried in the interior of the structure are not shown.

Figure 5. Sequence alignments of V_H and V_L sequences using an in-house V-class Ig-fold HMM. Custom alignment of V_H and V_L subclass sequences using the V-class HMM. The panel includes representative camelid V_{HH} sequences and a soluble anti-HEWL V_H sequence for comparison with the consensus V_H domains[63]. Residues colored red are those that bury a significant amount of surface area at the interface between V_H and V_L. Residue positions highlighted in yellow (V_H) or green (V_L) strongly covary with many residues that bury surface area at the Fv interface (from **Table 4** and **Table 5**). Camelid V_{HH} and soluble anti-HEWL V_H residues colored yellow and highlighted in red are involved in similar residue position networks, but with different amino acids from classical V_H domains at those positions (from **Table 6**). V_H subclass consensus residues that match the camelid or anti-HEWL residues at those positions are identically colored and highlighted. Only residue positions that were match states in the in-house, structure-based HMM are listed in the alignment. Positions of the framework and CDR regions are shown above and below the V_H and V_L sequences respectively.

Figure 6. Structural view of the most highly co-conserved V_H-V_L interface residues. The polypeptide backbone of the V_H domain (green) and V_L domain (blue) are depicted using a cartoon ribbon diagram. V_H residues V37, R38, G44, L45, W47, and W103 are displayed in the stick format in yellow. V_L residues Y36, Q37, A43, P44, L46, and F98 are displayed in the stick format in orange.

Table 1: Antibody V_H and V_L (kappa) amino acid pairs with the strongest covariations (ϕ-values). The two columns labeled "Top V_H [V_L] covarying amino acids" list the amino acids in the format A-B, and provide the residue codes and Kabat positions. Entries in the columns "V_H-V_L Interface," "V_H-C_H1 domain interface," and "V_L-C_L domain interface" identify amino acids (A, B, or both of each pair) near the specified interface.

Top V_H covarying amino acids (A-B)	**ϕ-value**	**V_H-V_L interface**	**V_H-C_H1 domain interface**	**Top V_L covarying amino acids (A-B)**	**ϕ-value**	**V_H-V_L interface**	**V_L-C_L domain interface**
G9-L18	0.71		**A**	Q37-G64	0.64	**A**	
G9-G10	0.69		**A-B**	G57-G64	0.60		
L45-W47	0.68	**A-B**		F98-L104	0.54	**A**	**B**
E6-G9	0.65		**B**	P44-G64	0.53	**A**	
G8-G9	0.65		**B**	Q37-P59	0.52	**A**	
V37-W47	0.65	**A-B**		Y36-P44	0.48	**A-B**	
V63-M82	0.64			Q37-G57	0.48	**A**	
G104-G106	0.64	**A**		S63-G64	0.48		
S7-G8	0.62			Q37-S67	0.47	**A**	
V63-Q81	0.62			Q37-K39	0.46	**A-B**	
E6-L18	0.60			Q37-P44	0.46	**A-B**	
G26-W47	0.60	**B**		P44-P59	0.46	**A**	
G44-W47	0.60	**A-B**		G57-S67	0.46		
G44-L45	0.60	**A-B**		P59-S63	0.45		
E6-G8	0.59			Y36-L46	0.44	**A-B**	
G8-T87	0.59		**B**	Q37-G68	0.44	**A**	
Q81-M82	0.59			P44-I75	0.44	**A**	
W103-V109	0.59	**A**	**B**	P44-G57	0.43	**A**	
G8-L18	0.58			Q6-P8	0.42		
G106-T107	0.58			Y36-F98	0.42	**A-B**	
W103-Q105	0.58	**A-B**		Q37-I48	0.42	**A-B**	
G8-G26	0.57			Q37-S63	0.42	**A**	
G26-T87	0.57		**B**	I48-I75	0.42	**A**	
T87-W103	0.57	**B**	**A**	P44-S67	0.41	**A**	
G106-V109	0.57		**B**	G64-I75	0.41		
P14-W47	0.56	**B**	**A**	P8-Y36	0.40	**B**	
G26-E46	0.56	**B**		T5-Q37	0.40	**B**	
R38-E46	0.56	**A-B**		Q37-R54	0.40	**A**	
E46-W47	0.56	**A-B**		P59-I75	0.40		
E46-T87	0.56	**A**	**B**	P44-F98	0.39	**A-B**	
V37-L45	0.54	**A-B**		P44-S63	0.39	**A**	
G8-G10	*0.53*		**B**	I48-S63	0.39	**A**	

Table 2: Top 40 V_H amino acid positions with the most covariations (ϕ-value > 0.25) with other V_H residues. Residues that bury surface area at the Fv interface are highlighted in black rows. Residues immediately adjacent in primary sequence to interface residues are in grey rows.

Amino acid	Kabat#	#Links all	Avg. ϕ-value	#Interface links	Avg. ϕ-value to interface
G	10	74	0.39	3	0.34
G	8	74	0.35	4	0.37
T	87	71	0.38	6	0.44
W	**103**	**69**	**0.36**	**6**	**0.36**
M	82	69	0.39	1	0.41
G	26	67	0.37	6	0.46
Y	59	66	0.36	5	0.41
V	63	64	0.38	1	0.33
Q	81	63	0.36	2	0.31
W	**47**	**61**	**0.37**	**6**	**0.54**
E	**46**	**61**	**0.36**	**6**	**0.41**
R	19	60	0.36	1	0.33
L	18	60	0.38	3	0.36
I	69	56	0.34	5	0.32
T	68	56	0.33	4	0.38
G	**49**	**56**	**0.33**	**5**	**0.43**
E	6	56	0.37	1	0.39
I	**51**	**55**	**0.35**	**5**	**0.38**
Y	79	54	0.34	3	0.38
S	62	54	0.36	3	0.36
G	16	54	0.34	1	0.27
D	72	52	0.34	5	0.39
A	**40**	**52**	**0.35**	**4**	**0.40**
G	65	50	0.33	3	0.32
V	**37**	**50**	**0.35**	**5**	**0.54**
R	**38**	**49**	**0.34**	**5**	**0.39**
S	17	48	0.33	2	0.30
S	7	47	0.34	4	0.30
K	43	46	0.34	2	0.36
A	24	46	0.33	3	0.33
F	27	45	0.32	5	0.35
L	4	45	0.32	3	0.29
S	21	44	0.33	2	0.32
V	109	44	0.32	2	0.46
F	29	43	0.33	5	0.41
Q	105	42	0.32	3	0.38
R	71	39	0.32	1	0.28
S	25	39	0.32	3	0.38
L	82c	38	0.32	2	0.33
K	75	38	0.32	1	0.29

Table 3: Top 40 V_L amino acid positions with the most covariations (ϕ-value > 0.25) with other V_L residues. Residues that bury surface area at the Fv interface are highlighted in black rows. Residues immediately adjacent in primary sequence to interface residues are in grey rows.

Amino acid	Kabat#	#Links all	Avg. ϕ-value	#Interface links	Avg. ϕ-value to interface
G	64	47	0.4	6	0.36
G	57	44	0.34	6	0.33
S	22	44	0.33	0	0
V	104	44	0.32	0	0
P	59	43	0.35	6	0.35
Q	100	42	0.32	0	0
Q	**37**	**41**	**0.37**	**6**	**0.35**
P	**44**	**40**	**0.33**	**6**	**0.35**
S	65	36	0.31	0	0
S	67	36	0.35	5	0.33
G	68	35	0.32	4	0.28
I	75	35	0.32	4	0.36
I	48	34	0.33	4	0.35
Y	**36**	**33**	**0.33**	**6**	**0.37**
R	54	30	0.33	3	0.31
P	15	29	0.35	0	0
K	39	28	0.32	4	0.3
S	63	28	0.33	3	0.32
T	5	27	0.32	2	0.31
Q	79	25	0.31	3	0.3
P	8	23	0.31	4	0.32
Q	89	23	0.31	2	0.28
S	10	21	0.31	1	0.32
S	56	21	0.31	2	0.28
I	21	20	0.29	2	0.28
G	66	19	0.31	0	0
A	**43**	**18**	**0.29**	**3**	**0.31**
T	74	18	0.3	2	0.27
S	14	17	0.33	1	0.36
F	62	17	0.31	0	0
T	72	17	0.3	1	0.25
L	**46**	**15**	**0.3**	**4**	**0.32**
F	**98**	**14**	**0.33**	**4**	**0.36**
Q	6	13	0.3	1	0.39
Q	42	13	0.28	0	0
K	**45**	**13**	**0.29**	**3**	**0.28**
Y	**49**	**13**	**0.29**	**2**	**0.27**
V	58	11	0.3	2	0.25
D	70	11	0.3	0	0
A	84	11	0.31	0	0

Table 4. **V_H Residues with multiple covariations (ϕ-value > 0.25) with V_H residues that bury surface area at the Fv interface**. Residues are sorted based on the difference between their average ϕ-value with interface residues versus their average ϕ-value with all positions within the V-class alignment. Residues that bury surface area at the interface are highlighted in black and marked with an X in the final column. Residues that are adjacent in primary sequence to interface residues are highlighted in dark grey and marked with an X±1 in the final column. Residues at the C_H1 interface are in light grey rows. "Non-specific" residues (bottom of table) are those falling below an arbitrary cutoff above which residues appear to have strong, specific connections with interface residues. This cutoff was chosen based on a $\Delta\phi$-value ([average ϕ-value with interface residues] – [average overall ϕ-value]) ≤ 0.01. W103 was grouped with the specific interface residues because it is an interface residue.

Amino Acid	Kabat#	#Links w/ interface residues[a]	Avg. ϕ-value w/ interface residues[a]	#Links w/ all positions[a]	Avg. ϕ-value w/ all positions[a]	$\Delta\phi$-value (interface - all)	Interface
V(I)	**37**	**5(3)**	**0.54(0.30)**	**50(19)**	**0.35(0.34)**	**0.19**	**X**
W	**47**	**6**	**0.54**	**61**	**0.37**	**0.17**	**X**
L	**45**	**5**	**0.53**	**33**	**0.36**	**0.17**	**X**
G	**44**	**4**	**0.52**	**31**	**0.35**	**0.17**	**X**
P	**14**	**6**	**0.46**	**29**	**0.35**	**0.11**	**C_H1**
G	**49**	**5**	**0.43**	**56**	**0.33**	**0.10**	**X-1**
G	26	6	0.46	67	0.37	0.09	
F	29	5	0.41	43	0.33	0.08	
S	74	5	0.39	26	0.32	0.07	
T	**87**	**6**	**0.44**	**71**	**0.38**	**0.06**	**C_H1**
E	**46**	**6**	**0.41**	**61**	**0.36**	**0.05**	**X-1**
Y	59	5	0.41	66	0.36	0.05	
R	**38**	**5**	**0.39**	**49**	**0.34**	**0.05**	**X+1**
D	72	5	0.39	52	0.34	0.05	
A	**40**	**4**	**0.40**	**52**	**0.35**	**0.05**	**X+1**
T	68	4	0.38	56	0.33	0.05	
P	41	6	0.33	32	0.30	0.03	
I	**51**	**5**	**0.38**	**55**	**0.35**	**0.03**	**X+1**
F	27	5	0.35	45	0.32	0.03	
L	**108**	**5**	**0.34**	**12**	**0.31**	**0.03**	**C_H1**
S	82b	4	0.32	24	0.29	0.03	
W	**103**	**6**	**0.36**	**69**	**0.36**	**0.00**	**X**
Non-Specific							
Y	32	5	0.30	15	0.29	0.01	
G	55	5	0.30	29	0.30	0.00	
I	69	5	0.32	56	0.34	-0.02	
S	7	4	0.30	47	0.34	-0.04	
G	8	4	0.35	74	0.39	-0.04	
L	11	4	0.30	30	0.31	-0.01	
L	11	4	0.30	30	0.31	-0.01	
Q	**39**	**3**	**0.27**	**12**	**0.29**	**-0.02**	**X**

[a] Minimum ϕ-value cutoff of 0.25.

Table 5. V_L Residues with multiple covariations (ϕ-value > 0.25) with V_L residues that bury surface area at the Fv interface. Residues are sorted based on the difference between their average ϕ-value with interface residues versus their average ϕ-value with all positions within the V-class alignment. Residues that bury surface area at the interface are highlighted in black and marked with an X in the final column. Residues that are adjacent in primary sequence to interface residues are highlighted in grey and marked with an X±1 in the final column. "Non-specific" residues (bottom of table) are those falling below an arbitrary cutoff above which residues appear to have strong, specific connections with interface residues. This cutoff was chosen based on a $\Delta\phi$-value ([average ϕ-value with interface residues] – [average overall ϕ-value]) ≤ 0.00. Q37, I48, K39, and K45 were grouped with the specific interface residues because they are adjacent in primary sequence to interface residues.

Amino Acid	Kabat#	#Links w/ interface residues	Avg. ϕ-value w/ interface residues[a]	#Links w/ all positions	Avg. ϕ-value w/ all positions[a]	$\Delta\phi$-value (interface - all)	Interface
Y	**36**	**6**	**0.37**	**33**	**0.33**	**0.04**	**X**
I	75	4	0.36	35	0.32	0.04	
L	47	2	0.33	9	0.29	0.04	X+1
F	**98**	**4**	**0.36**	**14**	**0.33**	**0.03**	**X**
P	**44**	**6**	**0.35**	**40**	**0.33**	**0.02**	**X**
L	**46**	**4**	**0.32**	**15**	**0.30**	**0.02**	**X**
P	59	6	0.35	43	0.35	0.00	
Q	**37**	**6**	**0.36**	**41**	**0.37**	**-0.01**	**X+1**
I	**48**	**4**	**0.32**	**34**	**0.33**	**-0.01**	**X+1**
K	**39**	**4**	**0.29**	**28**	**0.32**	**-0.03**	**X+1**
K	**45**	**3**	**0.28**	**13**	**0.39**	**-0.11**	**X-1**
Non-Specific							
G	57	6	0.33	44	0.34	-0.01	
S	67	5	0.33	36	0.35	-0.02	
K(R)	103	3(1)	0.32(0.27)	4(1)	0.34(0.27)	-0.02	
P	8	4	0.32	23	0.31	-0.01	
G	64	4	0.28	35	0.32	-0.04	
G	68	4	0.29	35	0.32	-0.03	
D	85	4	0.32	36	0.36	-0.04	
R	54	3	0.31	30	0.33	-0.02	
S	63	3	0.32	28	0.33	-0.01	
Q	79	3	0.30	25	0.31	-0.01	
S	56	2	0.28	21	0.31	-0.03	

[a] Minimum ϕ-value cutoff of 0.25.

Table 6. Contrasting features of V_H and camelid V_{HH} domains based on covariation analyses.

V_H linked pair	ϕ-value	Camelid V_{HH} linked pair	ϕ-value
G44-L45	0.61	E44-R45	0.57
L45-L108	0.29	R45-Q108	0.57
V(I)37-L45	0.54	F37-R45	0.50
P14-V37	0.40	A14-F37	0.50
L45-W47	0.70	R45-G47	0.46
V37-W47	0.65	F37-G47	0.44
-	-	C33[a]-G47	0.44
-	-	A14-Q108	0.44
P14-G44	0.35	A14-E44	0.43
G44-W47	0.61	E44-G47	0.42
K13[b]-L45	0.31	Q13-R45	0.42
V37-G44	0.53	F37-E44	0.40
-	-	C33[a]-R45	0.40
L82[b]-L45	0.29	M82-R45	0.38
V37-L108	0.32	F37-Q108	0.37
G49-L45	0.46	A49-R45	0.36
W47-L108	0.35	G47-Q108	0.36
L63[b]-L45	0.30	V63-R45	0.36
-	-	C33[a]-F37	0.36
P14-W47	0.48	A14-G47	0.35
-	-	C33[a]-E44	0.35
K13[b]-G44	0.30	Q13-E44	0.34
-	-	Q13-F37	0.32
S74-L45	0.45	A74-R45	0.30
L82[b]-G44	0.29	M82-E44	0.27
S74-L108	0.25	A74-Q108	0.27
-	-	K83-E44	0.27
K13[b]-W47	0.36	Q13-G47	0.26
-	-	V63-E44	0.26
S74-G44	0.38	A74-E44	0.25
S74-V37	0.33	-	-

[a] C33 often makes a disulfide with CDR3 in camelid V_{HH} domains to stabilize camelid CDR3 structures.

[b] V_H3 consensus matches the V_{HH} consensus amino acids at these positions.

Figure 1

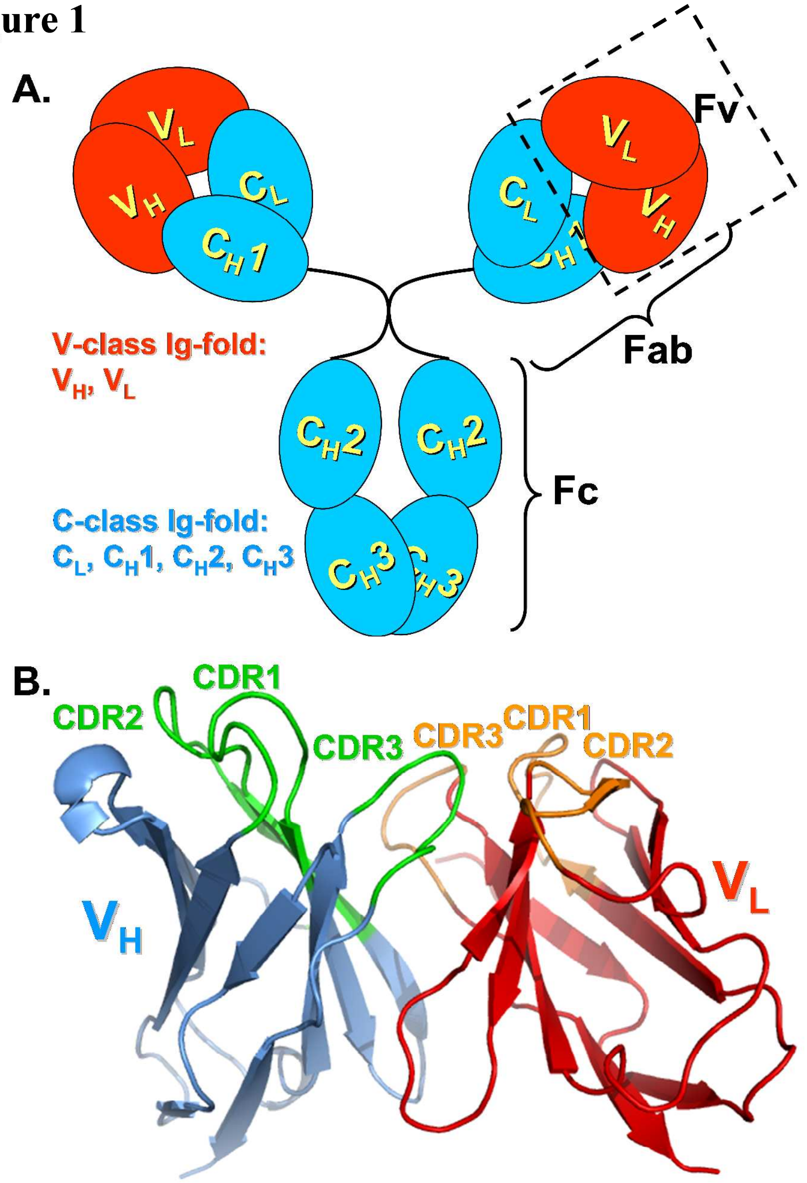


Figure 1. Diagrams of an immunoglobulin and its Fv domain. A. Schematic diagram of an IgG antibody. The variable domains which compose the antigen-binding or Fv-region are shown in red and the constant domains are shown in blue. The variable domains are V-class Ig-folds, while the constant domains are C-class Ig-folds, which are highly similar to V-class Ig-folds, but lack two additional β-strands commonly found in V-class structures. B. Ribbon diagram of an antibody Fv-region consisting of a variable domain from the immunoglobulin heavy chain (VH-blue) and a variable domain from the immunoglobulin light chain (VL-red). The complementarity determining regions (CDRs) of the VH (shown in green) and the VL (shown in orange) comprise the antigen-binding site.
1190x1587mm (96 x 96 DPI)

Figure 2

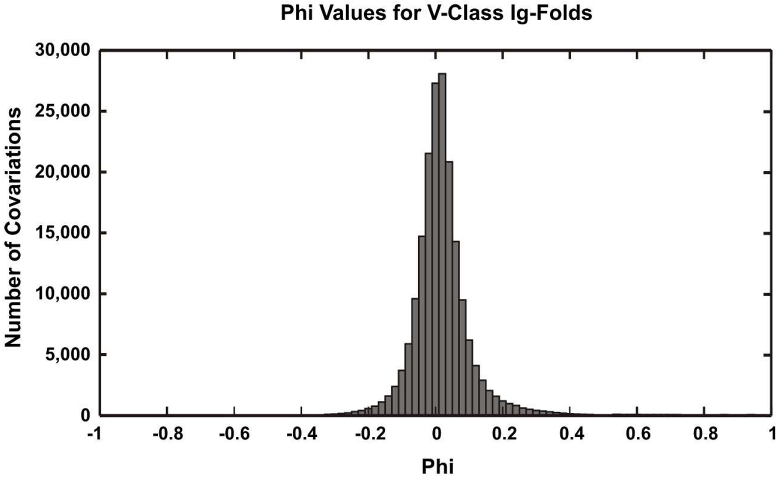


Figure 2. Distribution of ϕ-values calculated for the V-class alignment. There are 4,118,400 possible amino acid pairings within the V-class sequences. Of these possible pairings, 1,098,890 actually exist within the sequence database (i.e., some amino acids pairing are not observed across columns of the alignment). The histogram shows the distribution of ϕ-values from the 186,171 pairings that occur at least 10 times. The 13,796 ϕ-values greater than 0.1237 were considered statistically significant, using a conservative statistical approach (see text).
1190x1587mm (96 x 96 DPI)

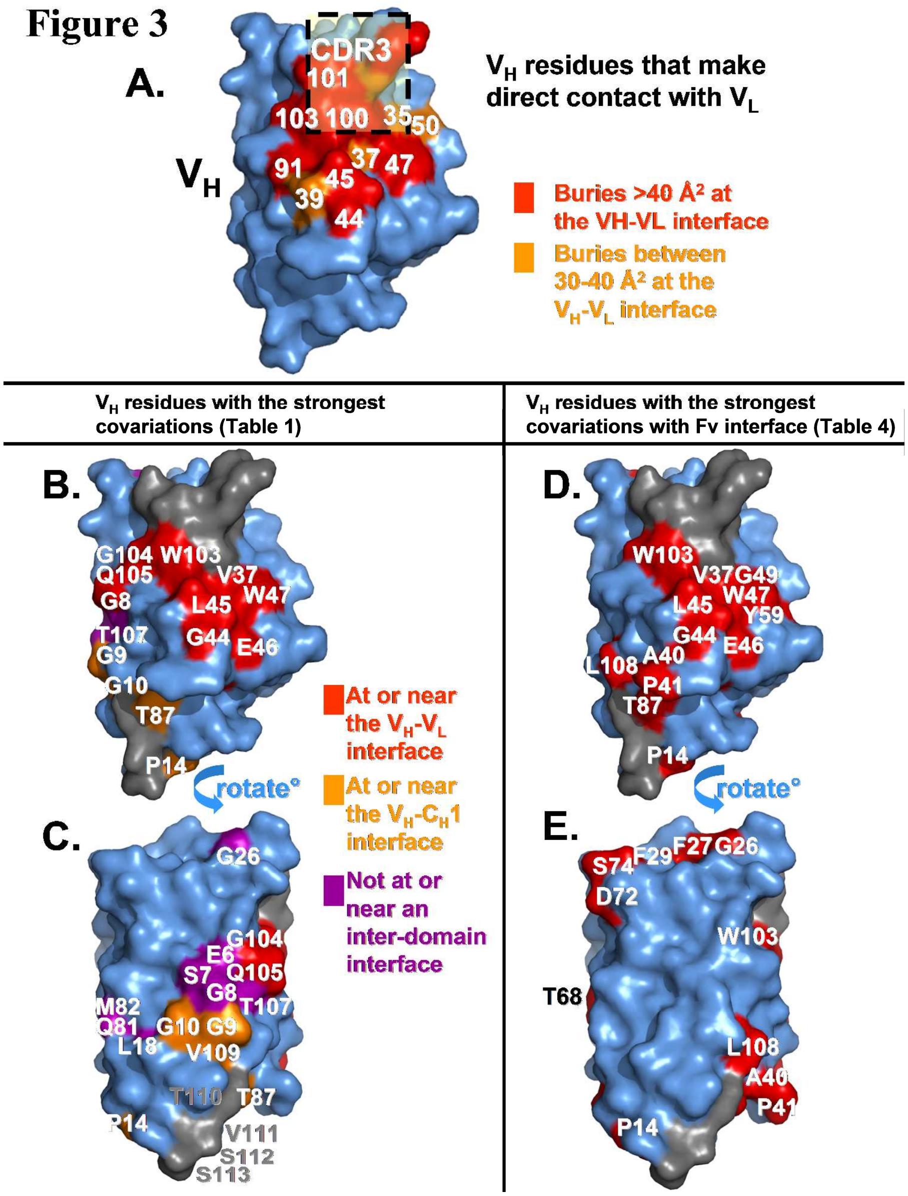


Figure 3
A.
CDR3
101
103 100 35 50
37 47
91 45
39
44
V_H
V_H residues that make direct contact with V_L
🟥 Buries >40 Å² at the VH-VL interface
🟧 Buries between 30-40 Å² at the V_H-V_L interface
V_H residues with the strongest covariations (Table 1)
V_H residues with the strongest covariations with Fv interface (Table 4)
B.
G104 W103
Q105 V37
G8 W47
L45
T107 G44 E46
G9
G10
T87
P14
rotate°
🟥 At or near the V_H-V_L interface
🟧 At or near the V_H-C_H1 interface
🟪 Not at or near an inter-domain interface
C.
G26
G104
E6
S7 Q105
G8
M82 T107
Q81 G10 G9
L18 V109
T110 T87
P14
V111
S112
S113
D.
W103
V37G49
W47
L45 Y59
G44 E46
L108 A40
P41
T87
P14
rotate°
E.
F27 G26
S74 F29
D72
W103
T68
L108
A40
P41
P14

Figure 3. Covariations mapped to surface representations of an antibody VH domain derived from an in-house Fab structure. A. Surface representation of a VH domain. Residues that bury >40 Å2 at the Fv interface are shown in red and those that bury between 30-40 Å2 are shown in orange. In B.-E., residues colored grey (CDR3 residues as well as four residues at the C-terminus) were not match states in the HMM-derived V-class alignment and were not evaluated in this study. B. and C. VH residues from amino acid pairs with the highest ϕ -values from Table 1 were mapped to the VH surface: red = proximal to the VH-VL interface; orange = proximal to the VH-CH1 interface; and purple = distant from the two interfaces. D. and E. VH residues from Table 4 that display multiple covariations (ϕ -value > 0.25) with VH-VL interface residues with greater than average ϕ -values are mapped onto the VH surface in red. Residues from Tables 1 and 4 that are completely buried in the interior of the structure are not shown.
1190x1587mm (96 x 96 DPI)

Figure 4

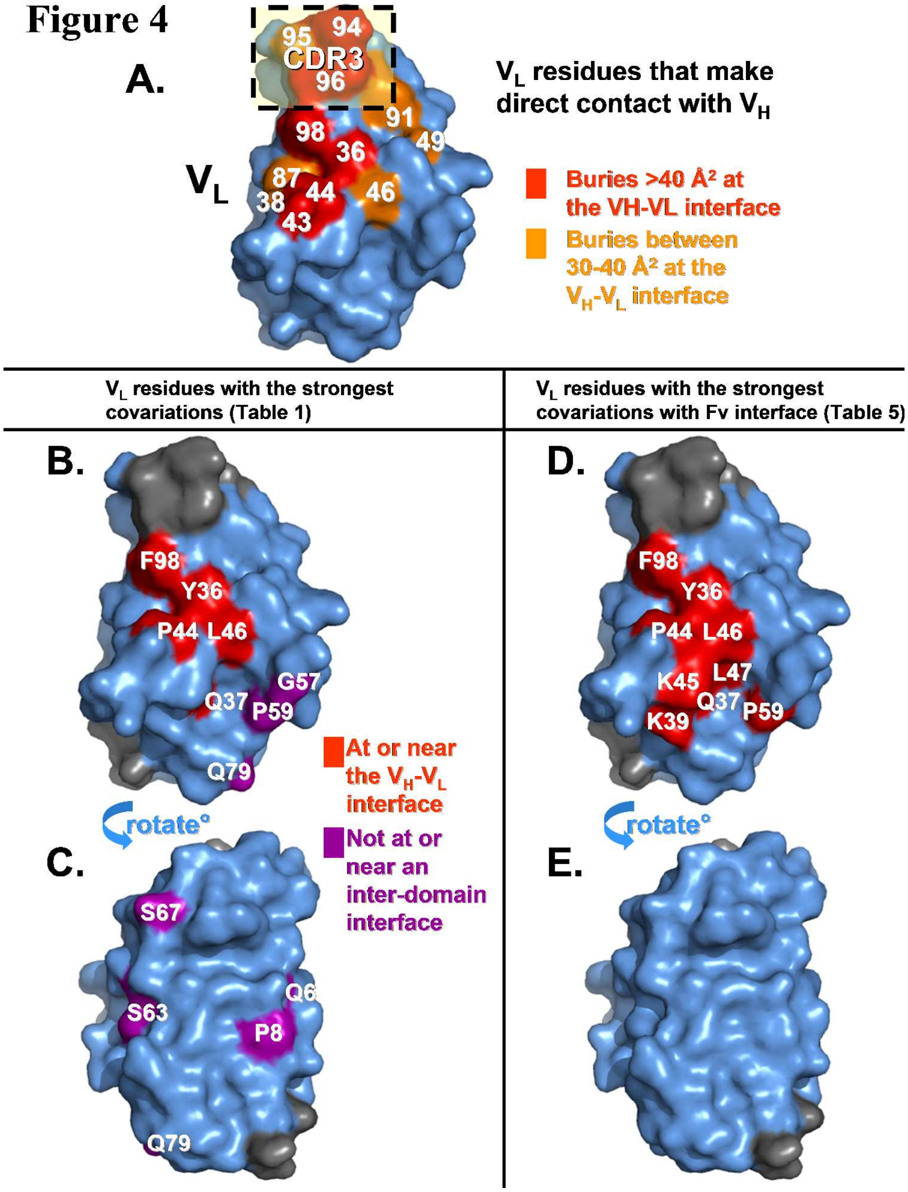

A.
95
94
CDR3
96
91
49
98
36
V_L
87
44
46
38
43
V_L residues that make direct contact with V_H
🟥 Buries >40 Å² at the VH-VL interface
🟧 Buries between 30-40 Å² at the V_H-V_L interface
V_L residues with the strongest covariations (Table 1)
V_L residues with the strongest covariations with Fv interface (Table 5)
B.
F98
Y36
P44 L46
G57
Q37
P59
Q79
🟥 At or near the V_H-V_L interface
🟪 Not at or near an inter-domain interface
rotate°
C.
S67
Q6
S63
P8
Q79
D.
F98
Y36
P44 L46
K45
L47
Q37
K39
P59
rotate°
E.

Figure 4. Covariations mapped to surface representations of an antibody VL domain derived from an in-house Fab structure. A. Surface representation of a VL domain. Residues that bury >40 Å2 at the Fv interface are shown in red and those that bury between 30-40 Å2 are shown in orange. In B.-E., residues colored grey (CDR3 residues as well as four residues at the C-terminus) were not match states in the HMM-derived V-class alignment and were not evaluated in this study. B. and C. VL residues from amino acid pairs with the highest ϕ -values from Table 1 were mapped to the VL surface: red = proximal to the VH-VL interface; orange = proximal to the VL-CL interface; and purple = distant from the two interfaces. D. and E. VL residues from Table 5 that display multiple covariations (ϕ -value > 0.25) with VH-VL interface residues with greater than average ϕ -values are mapped onto the VL surface in red. Residues from Tables 1 and 5 that are completely buried in the interior of the structure are not shown.
1190x1587mm (96 x 96 DPI)

Figure 5

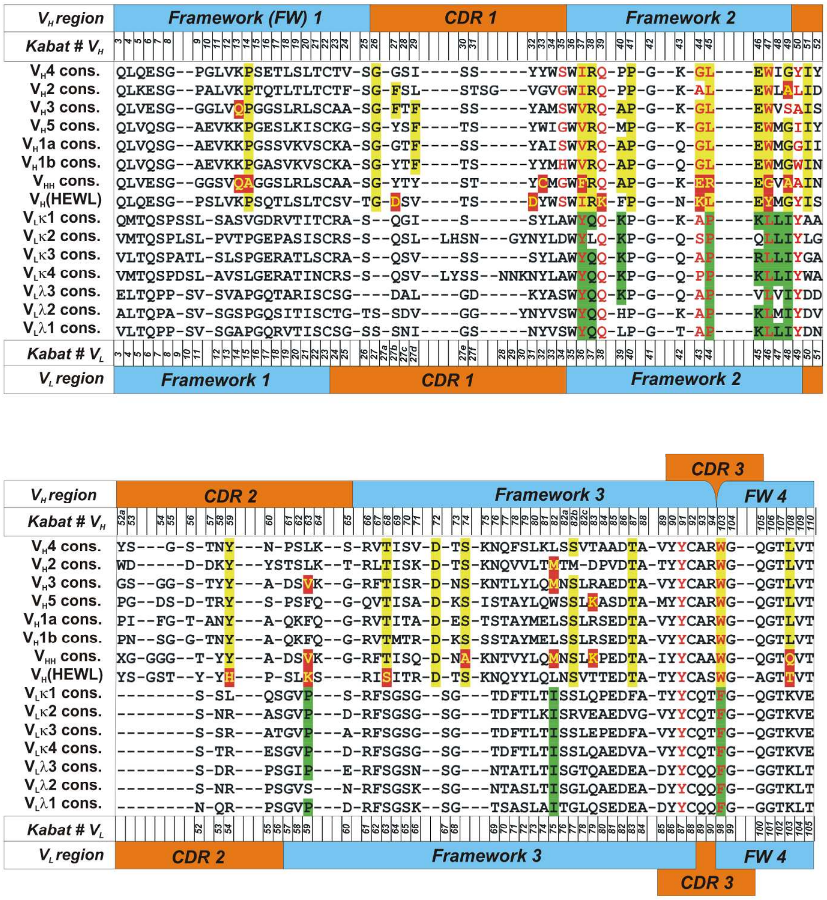

VH region
Framework (FW) 1
CDR 1
Framework 2
Kabat # VH
3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52
VH4 cons. QLQESG--PGLVKPSETLSLTCTV-SG-GSI----SS-----YYWSWIRQ-PP-G--K-GL----EWIGYIY
VH2 cons. QLKESG--PALVKPTQTLTLTCTF-SG-FSL----STSG---VGVGWIRQ-PP-G--K-AL----EWLALID
VH3 cons. QLVESG--GGLVQPGGSLRLSCAA-SG-FTF----SS-----YAMSWVRQ-AP-G--K-GL----EWVSAIS
VH5 cons. QLVQSG--AEVKKPGESLKISCKG-SG-YSF----TS-----YWIGWVRQ-MP-G--K-GL----EWMGIIY
VH1a cons. QLVQSG--AEVKKPGSSVKVSCKA-SG-GTF----SS-----YAISWVRQ-AP-G--Q-GL----EWMGGII
VH1b cons. QLVQSG--AEVKKPGASVKVSCKA-SG-YTF----TS-----YYMHWVRQ-AP-G--Q-GL----EWMGWIN
VHH cons. QLVESG--GGSVQAGGSLRLSCAA-SG-YTY----ST-----YCMGWFRQ-AP-G--K-ER----EGVAAIN
VH(HEWL) QLQESG--PSLVKPSQTLSLTCSV-TG-DSV----TS-----DYWSWIRK-FP-G--N-KL----EYMGYIS
VLκ1 cons. QMTQSPSSL-SASVGDRVTITCRA-S--QGI----S------SYLAWYQQ-KP-G--K-AP----KLLIYAA
VLκ2 cons. VMTQSPLSL-PVTPGEPASISCRS-S--QSL--LHSN---GYNYLDWYLQ-KP-G--Q-SP----QLLIYLG
VLκ3 cons. VLTQSPATL-SLSPGERATLSCRA-S--QSV----SS-----SYLAWYQQ-KP-G--Q-AP----RLLIYGA
VLκ4 cons. VMTQSPDSL-AVSLGERATINCRS-S--QSV--LYSS--NNKNYLAWYQQ-KP-G--Q-PP----KLLIYWA
VLλ3 cons. ELTQPP-SV-SVAPGQTARISCSG----DAL----GD-----KYASWYQQ-KP-G--Q-AP----VLVIYDD
VLλ2 cons. ALTQPA-SV-SGSPGQSITISCTG-TS-SDV----GG----YNYVSWYQQ-HP-G--K-AP----KLMIYDV
VLλ1 cons. VLTQPP-SV-SGAPGQRVTISCSG-SS-SNI----GS-----NYVSWYQQ-LP-G--T-AP----KLLIYDN
Kabat # VL
3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 27a 27b 27c 27d 27e 27f 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51
VL region
Framework 1
CDR 1
Framework 2
CDR 3
VH region
CDR 2
Framework 3
FW 4
Kabat # VH
52a 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 82a 82b 82c 83 84 85 86 87 88 89 90 91 92 93 94 103 104 105 106 107 108 109 110
VH4 cons. YS---G-S-TNY---N-PSLK--S-RVTISV-D-TS-KNQFSLKLSSVTAADTA-VYYCARWG--QGTLVT
VH2 cons. WD-----D-DKY---YSTSLK--T-RLTISK-D-TS-KNQVVLTMTM-DPVDTA-TYYCARWG--QGTLVT
VH3 cons. GS--GG-S-TYY---A-DSVK--G-RFTISR-D-NS-KNTLYLQMNSLRAEDTA-VYYCARWG--QGTLVT
VH5 cons. PG--DS-D-TRY---S-PSFQ--G-QVTISA-D-KS-ISTAYLQWSSLKASDTA-MYYCARWG--QGTLVT
VH1a cons. PI--FG-T-ANY---A-QKFQ--G-RVTITA-D-ES-TSTAYMELSSLRSEDTA-VYYCARWG--QGTLVT
VH1b cons. PN--SG-G-TNY---A-QKFQ--G-RVTMTR-D-KS-SSTAYMELSSLRSEDTA-VYYCARWG--QGTLVT
VHH cons. XG-GGG--T-YY---A-DSVK--G-RFTISQ-D-NA-KNTVYLQMNSLKPEDTA-IYYCAAWG--QGTQVT
VH(HEWL) YS-GST--Y-YH---P-SLKS----RISITR-D-TS-KNQYYLQLNSVTTEDTA-TYYCASWG--AGTTVT
VLκ1 cons. --------S-SL---QSGVP---S-RFSGSG--SG---TDFTLTISSLQPEDFA-TYYCQTFG--QGTKVE
VLκ2 cons. --------S-NR---ASGVP---D-RFSGSG--SG---TDFTLKISRVEAEDVG-VYYCQTFG--QGTKVE
VLκ3 cons. --------S-SR---ATGVP---A-RFSGSG--SG---TDFTLTISSLEPEDFA-VYYCQTFG--QGTKVE
VLκ4 cons. --------S-TR---ESGVP---D-RFSGSG--SG---TDFTLTISSLQAEDVA-VYYCQTFG--QGTKVE
VLλ3 cons. --------S-DR---PSGIP---E-RFSGSN--SG---NTATLTISGTQAEDEA-DYYCQQFG--GGTKLT
VLλ2 cons. --------S-NR---PSGVS---N-RFSGSK--SG---NTASLTISGLQAEDEA-DYYCQQFG--GGTKLT
VLλ1 cons. --------N-QR---PSGVP---D-RFSGSK--SG---TSASLAITGLQSEDEA-DYYCQQFG--GGTKLT
Kabat # VL
52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 98 99 100 101 102 103 104 105
VL region
CDR 2
Framework 3
FW 4
CDR 3

Figure 5. Sequence alignments of VH and VL sequences using an in-house V-class Ig-fold HMM. Custom alignment of VH and VL subclass sequences using the V-class HMM. The panel includes representative camelid VHH sequences and a soluble anti-HEWL VH sequence for comparison with the consensus VH domains63. Residues colored red are those that bury a significant amount of surface area at the interface between VH and VL. Residue positions highlighted in yellow (VH) or green (VL) strongly covary with many residues that bury surface area at the Fv interface (from Table 4 and Table 5). Camelid VHH and soluble anti-HEWL VH residues colored yellow and highlighted in red are involved in similar residue position networks, but with different amino acids from classical VH domains at those positions (from Table 6). VH subclass consensus residues that match the camelid or anti-HEWL residues at those positions are identically colored and highlighted. Only residue positions that were match states in the in-house, structure-based HMM are listed in the alignment. Positions of the framework and CDR regions are shown above and below the VH and VL sequences respectively.
1190x1587mm (96 x 96 DPI)

Figure 6

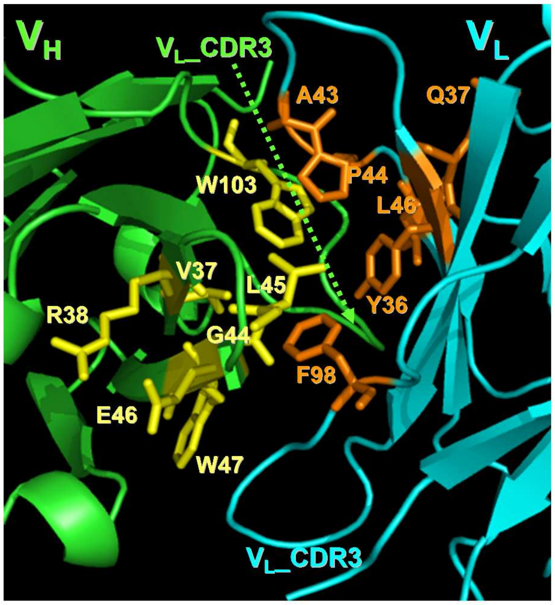
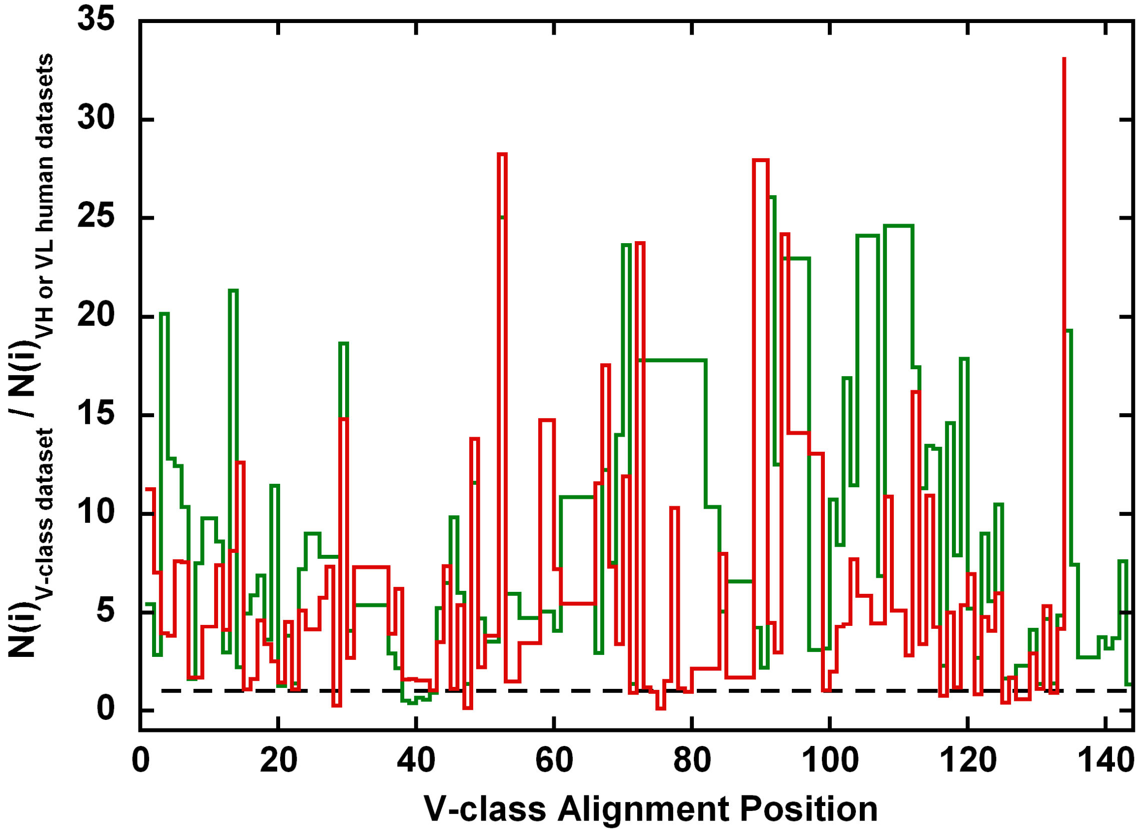


Figure 6. Structural view of the most highly co-conserved VH-VL interface residues. The polypeptide backbone of the VH domain (green) and VL domain (blue) are depicted using a cartoon ribbon diagram. VH residues V37, R38, G44, L45, W47, and W103 are displayed in the stick format in yellow. VL residues Y36, Q37, A43, P44, L46, and F98 are displayed in the stick format in orange.
1190x1587mm (96 x 96 DPI)

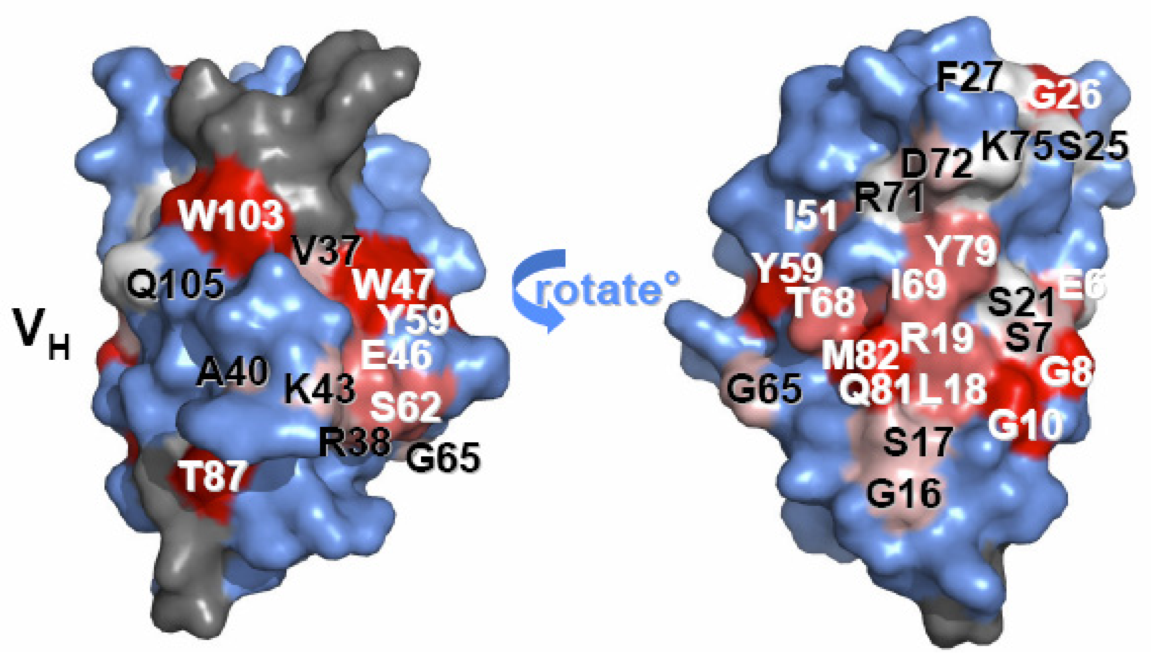
Supplemental Figure 1. Positional entropy ratio of the V-class dataset versus human V_H and human V_L sequence datasets that were validated to have a representative distributions of V-gene subclasses. The Positional Entropy, *N(i)*, is a measure of every residue position's variability and is related to the information theoretic Shannon entropy, *H(i)*:

$$N(i) = e^{H(i)}, H(i) = -\sum_{r=A}^{Y} p_i(r)\ln(p_i(r)),$$

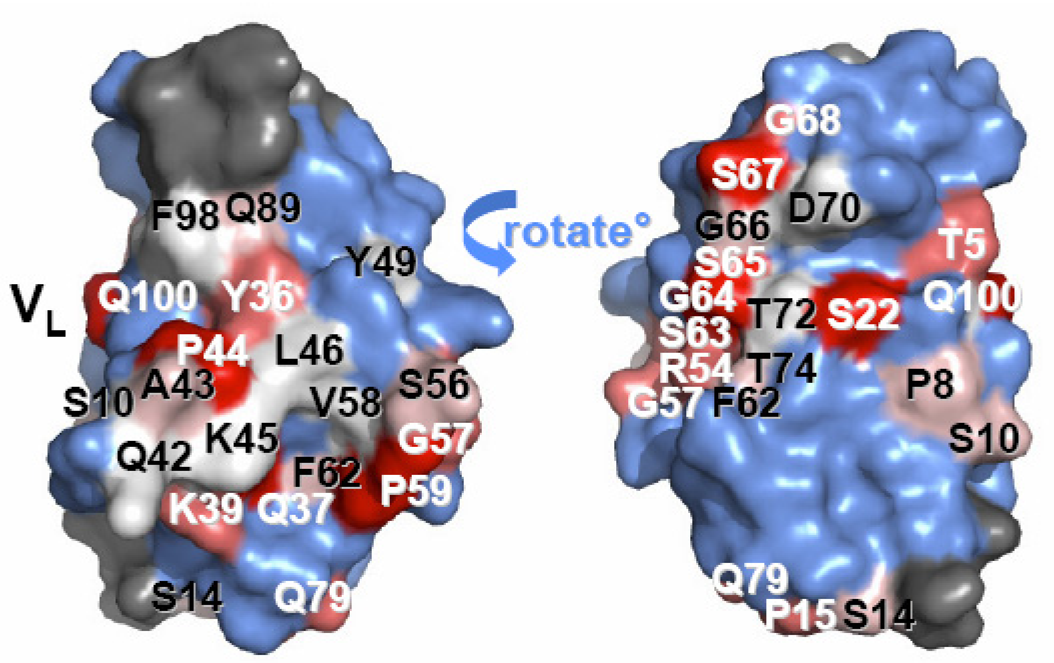
where $p_i(r)$ is each residue's frequency at position 'i' in the alignment. The ratios between the V-class alignment and the V_H and V_L datasets are shown in red and green, respectively. The dotted line at a value of 1.0 represents the ratio that would be expected if the diversity between the sequence datasets were equivalent. The generally higher values for the V-class dataset indicate higher diversity.



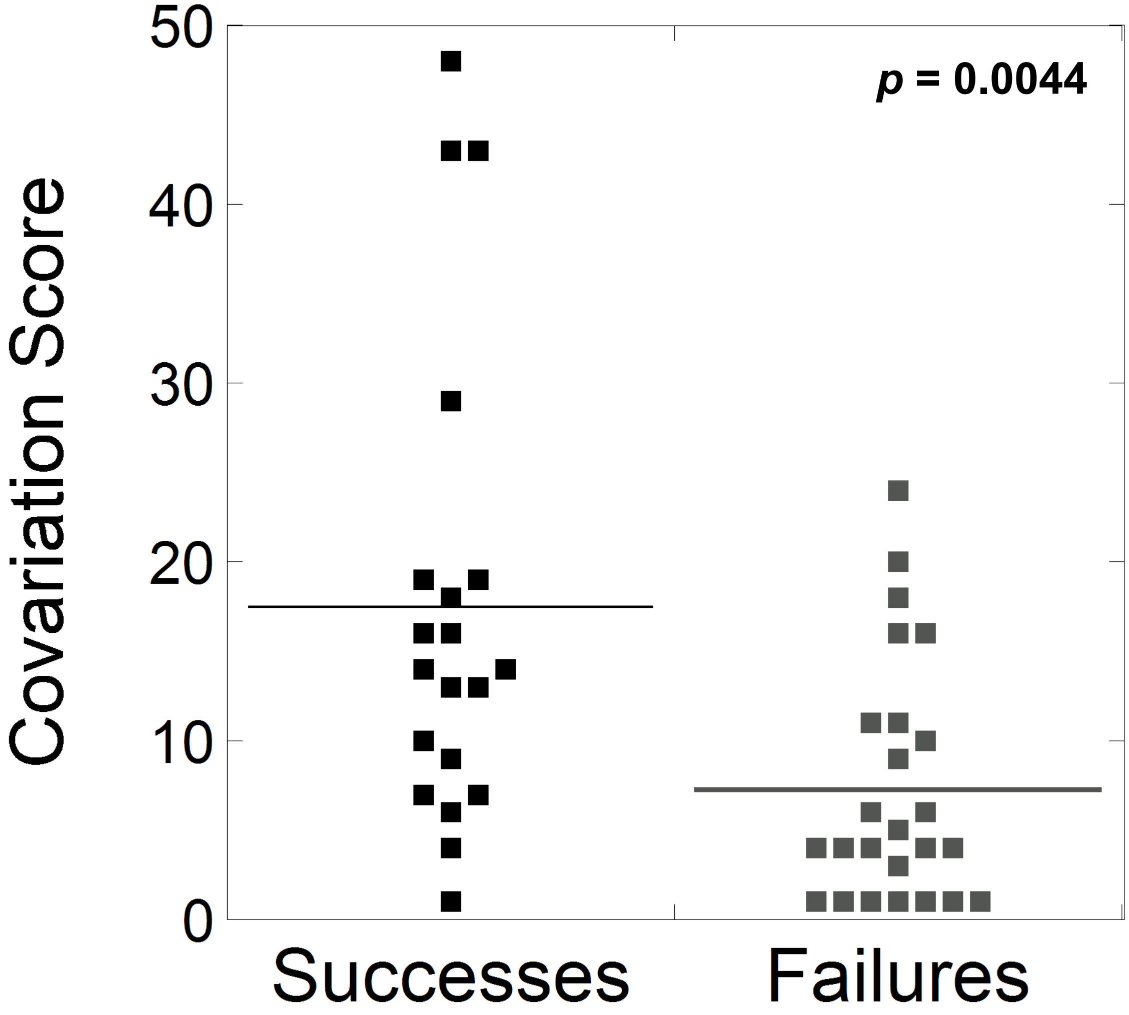
Supplemental Figure 2. V$_H$ residues from Table 2 that display the highest number of covariations (ϕ-value > 0.25) with other V$_H$ domain amino acids. The hue of red (darkest to lightest) represents the largest to smallest number of overall covariations, respectively. Residues from Table 2 that are completely buried within the structure are not shown. Residues colored grey were not match states in our in-house V-class alignment and were not analyzed.



Supplemental Figure 3. V_L residues from Table 3 that display the highest number of covariations (ϕ-value > 0.25) with other V_L domain amino acids. The hue of red (darkest to lightest) represents the largest to smallest number of overall covariations, respectively. Residues from Table 3 that are completely buried within the structure are not shown. Residues colored grey were not match states in our in-house V-class alignment and were not analyzed.



Supplemental Figure 4. Covariation data enable rational design of more stable scFVs. The covariation data are used to find gaps in conserved networks of amino acids present in particular scFvs. Filling these gaps leads to an increase in the number of co-conserved amino acid pairs. We *score* covariation-based designs based on the number of ϕ-value links above 0.3 that are added to a particular sequence by changing an amino acid within a scFv. We have used covariation information to rationally design stabilizing mutations within 4 different scFvs (Miller *et al.*, manuscript in preparation). Successful designs have yielded between 1 and 12 ºC increases in the thermal unfolding midpoint (T_M) of the scFvs. We plotted the distribution of the scores for both successful and failed designs. The mean covariation *scores* for successes and failures were 18 and 7, respectively. The distributions were significantly different with a Student's t-test *p*-value = 0.004, suggesting that using the covariation data in this fashion is an effective means of predicting stabilizing mutations. Covariation scoring predicted stabilizing mutations with a success rate of 44% (taking all covariation scores > 0). If only the top 50% of these scores are used (*i.e.*, scores > 10), the success rate is 64%.


Covariation Score
50
40
30
20
10
0
p = 0.0044
Successes
Failures