Structure
Article

Loop-Sequence Features and Stability Determinants in Antibody Variable Domains by High-Throughput Experiments

Hung-Ju Chang,[1,2,3] Jhih-Wei Jian,[1,4,5] Hung-Ju Hsu,[1] Yu-Ching Lee,[1] Hong-Sen Chen,[1] Jhong-Jhe You,[1] Shin-Chen Hou,[1] Chih-Yun Shao,[1,6] Yen-Ju Chen,[1,7] Kuo-Ping Chiu,[1] Hung-Pin Peng,[1,4,5] Kuo Hao Lee,[1] and An-Suei Yang[1,*]

[1]Genomics Research Center, Academia Sinica, Taipei 115, Taiwan
[2]Chemical Biology and Molecular Biophysics Program, Taiwan International Graduate Program, Institute of Biological Chemistry, Academia Sinica, Taipei 115, Taiwan
[3]Institute of Biochemical Sciences, College of Life Science, National Taiwan University, Taipei 106, Taiwan
[4]Institute of Biomedical Informatics, National Yang-Ming University, Taipei 11221, Taiwan
[5]Bioinformatics Program, Taiwan International Graduate Program, Institute of Information Science, Academia Sinica, Taipei 115, Taiwan
[6]Institute of Zoology, College of Life Sciences, National Taiwan University, Taipei 106, Taiwan
[7]Genome and Systems Biology Degree Program, National Taiwan University, Taipei 106, Taiwan
*Correspondence: yangas@gate.sinica.edu.tw
http://dx.doi.org/10.1016/j.str.2013.10.005

SUMMARY

Protein loops are frequently considered as critical determinants in protein structure and function. Recent advances in high-throughput methods for DNA sequencing and thermal stability measurement have enabled effective exploration of sequence-structure-function relationships in local protein regions. Using these data-intensive technologies, we investigated the sequence-structure-function relationships of six complementarity-determining regions (CDRs) and ten non-CDR loops in the variable domains of a model vascular endothelial growth factor (VEGF)-binding single-chain antibody variable fragment (scFv) whose sequence had been optimized via a consensus-sequence approach. The results show that only a handful of residues involving long-range tertiary interactions distant from the antigen-binding site are strongly coupled with antigen binding. This implies that the loops are passive regions in protein folding; the essential sequences of these regions are dictated by conserved tertiary interactions and the consensus local loop-sequence features contribute little to protein stability and function.

INTRODUCTION

Protein engineering is critical for developing protein-based therapeutics and diagnostics. Single-chain antibody variable domain fragments (scFvs) are important pharmaceutical molecules (Chan and Carter, 2010; Holliger and Hudson, 2005; Huang et al., 2010; Miller et al., 2010; Nelson and Reichert, 2009; Weatherill et al., 2012). Because the scFv structure is not stabilized by the constant domains, as in intact immunoglobulin, investigators have applied sequence fitness-searching principles based on random mutagenesis and screening (Jermutus et al., 2001; Jespers et al., 2004; Jung et al., 1999) or rational design with consensus-sequence profiles (Demarest and Glaser, 2008; Ewert et al., 2003; Honegger, 2008; Jordan et al., 2009; Kügler et al., 2009; Miller et al., 2010; Monsellier and Bedouelle, 2006; Wörn and Plückthun, 1998, 2001) to stability engineering of scFvs. However, exploring the vast sequence fitness possibilities and elucidating the coupling of sequence and function remain challenging due to limitations in protein engineering capabilities. In this work, to further extend the scope of antibody stability engineering with high-throughput experiments, we sought to (1) systematically determine the local sequence preferences in all loop regions of the antibody variable domains for their effects on antibody stability, (2) compare the information with the conserved sequence features in the antibody variable domain families, (3) experimentally test the energetic contributions of the local sequence features with the overall antibody variable domain stability, and (4) formulate general stability engineering strategies for the antibody variable domains.

A large body of evidence has supported the notion that loop regions are critical determinants in protein folding (see reviews in Jager et al., 2008, and Marcelino and Gierasch, 2008, and references therein). It has been long recognized that the loop regions in a protein structure can serve as active folding initiation sites that dominate the overall folding topology of the structure by dictating the reversal positions of the polypeptide chain to fold onto itself, or as passive joints that accommodate polypeptide chain reversal for folding to proceed (Hsu et al., 2006; Jager et al., 2008; Marcelino and Gierasch, 2008; Yang et al., 1996). The effects of the loop regions on the overall stability and folding of immunoglobulin (Ig)-like structures have been demonstrated to couple with the protein core and to contribute substantial stabilizing energy not only to the folded structure but also to the main folding transition state (Billings et al., 2008). Despite the

recognition of critical structural determinants embedded in the local sequences of protein loop regions, the intricate interrelationships among the local structural signals and the overall folding topology and stability remain unpredictable (Bofill and Searle, 2005; McCallister et al., 2000; Nauli et al., 2001; Sharpe et al., 2007).

Recently, high-throughput experimental techniques have been shown to be effective for establishing informatics bases for novel antibody stability engineering. Next-generation sequencing (NGS) (Fowler et al., 2010; Hietpas et al., 2011; Schlinkmann et al., 2012; Whitehead et al., 2012) and high-throughput thermal inactivation (HTTI) experiments (Miller et al., 2010) have substantially expanded the capabilities for protein stability engineering. The large volume of sequence information provided by NGS enables sequence analyses of statistical significance, and HTTI measurements allow physicochemical interpretations of the sequence features that emerge from the NGS statistics. Here, in contrast to recent studies involving protein engineering with NGS, which focused on deriving sequence fitness statistics from single-mutation-per-variant libraries (Fowler et al., 2010; Hietpas et al., 2011; Schlinkmann et al., 2012; Whitehead et al., 2012), we constructed scFv variants with changes of several residues (n = 5–11) simultaneously spanning a loop region in the protein so as to reveal the cooperativity of the residues within the local structure in consideration. This capability is particularly important to consider when investigating local turn propensities involving several consecutive residues in the loops.

The knowledge gained from this work should further our understanding of the structural and stability roles played by local sequences in the loop regions of the antibody variable domains, and enhance the basis for rational antibody stability engineering aimed at optimizing the loop sequences, including complementarity-determining regions (CDRs), in the antibody variable domains. In this work, we used NGS to explore the sequence preferences in groups of consecutive or noncontiguous residues (5–11 residues) in the 16-loop regions of a model 4D5 scFv capable of recognizing human vascular endothelial growth factor (VEGF) (Yu et al., 2012). Synthetic degenerate DNA codons were used for saturated mutagenesis of the loop regions, and the functional scFv variants were selected according to the binding or folding criteria for a filamentous phage display system. The selected functional scFv variants were sequenced with NGS for statistical analyses. In parallel, the stabilities of hundreds of selected functional scFv variants from the loop libraries were also assessed quantitatively with HTTI. We aimed to use this to support or challenge the hypothesis that local turn propensities in at least some, if not all, of the loops are conserved so as to guide the folding mechanism and stabilize the native structures of the antibody variable domains, given that the ten non-CDR loops in known antibody variable domain structures are mostly well-structured with conserved tight turns or one-turn 3_{10} helices. Unexpectedly, the results show that only a handful of residues involving long-range tertiary interactions distant from the antigen-binding site are strongly coupled with antigen binding. The local consensus-sequence features contribute little to the stability and function of the variable domain, suggesting that the loops are passive regions in the variable domain folding. In addition, all of the CDR and non-CDR loops are energetically coupled to the framework of the variable domains to various extents, and thus the variable-domain stability can be substantially enhanced by optimizing the loop sequences.

RESULTS AND DISCUSSION

Selection of Functional scFv Variants from 24 Phage-Displayed scFv Libraries for VEGF and Protein A Binding

Randomized scFv variants were constructed in 24 scFv libraries through saturated mutagenesis of 5–11 consecutive residues in the 16-loop regions of a VEGF-binding model scFv (Yu et al., 2012) by nucleotide-directed mutagenesis (Supplemental Experimental Procedures; Figure S1; Tables S1 and S2 available online). Each library contained on the order of 10^9 variants (Table S1). The template scFv (dubbed Av1.2) was derived from the 4D5 antibody framework (Eigenbrot et al., 1993; Fuh et al., 2006) with the $huV_\kappa1$-huV_H3 single-chain scFv construct (Figure S1). This template was chosen because the sequence of the 4D5 framework has been optimized via the consensus-sequence approach and the robust stability of the scFv has been demonstrated (Wörn and Plückthun, 1999). In addition, the CDR structures of the 4D5 template, as defined by North et al. (2011), are most common in known antibody structures (North et al., 2011), making the 4D5 template a representative model system for antibody engineering. The saturated mutagenesis of several residues in an scFv library allows one to explore all combinations of amino acid types that are favorable for scFv folding, within the complexity limit of $\sim10^9$ for each scFv library.

The functional scFv variants were selected with two to three rounds of the phage display selection-amplification cycle for binding to immobilized VEGF or immobilized Protein A (Supplemental Experimental Procedures; Figures S2 and S3). As shown in the composite complex structures in Figure 1, Av1.2 binds to VEGF through the CDRs and binds to Protein A through the framework of the V_H domain. We sequenced the scFv variants that bound to VEGF or Protein A to examine whether local sequence features are required for the stability and function of scFvs.

Figure 2 shows that the phage display selections for VEGF binding generated results similar to those obtained from the phage display selections for Protein A binding. Figures 2A and 2B show the comparable sequence profiles for the outer loop in the V_H domain (named VHOL) derived from VEGF binding selection and Protein A binding selection, respectively. Quantitatively, the Pearson's correlation coefficient (R) for the information content of the 20 amino acid types (y axis in Figure 2; see Supplemental Experimental Procedures) at the positions shown in Figures 2A and 2B is 0.99, and the t-test p value is 6.42E-95. A similar conclusion also emerges from the results shown in Figures 2C and 2D, where the sequence profiles for the outer loop in the V_L domain (named VLOL) derived from the VEGF and Protein A binding selection, respectively, are also comparable, although to a lesser extent (R = 0.73, p = 1.27E-21). The likely interpretation of the highly significant quantitative correlation between the two sets of sequence profiles is that the scFv variants selected either way reveal the sequence requirements for the folded structure that is competent in binding to both VEGF

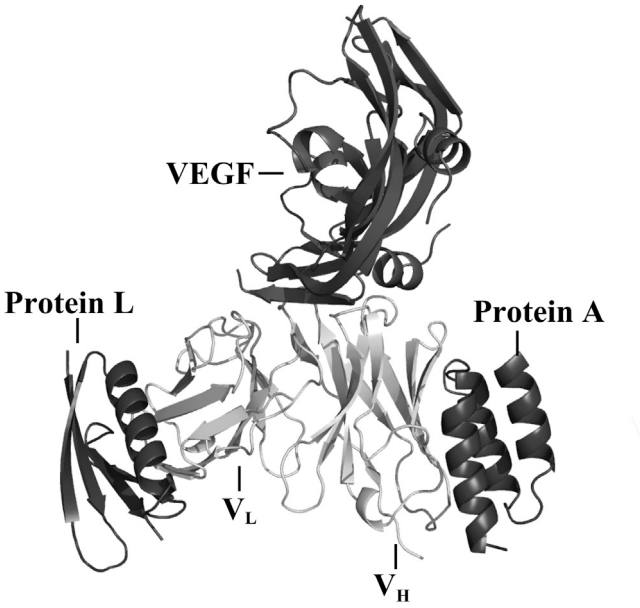

Figure 1. Composite Structure for the Template scFv Av1.2 Binding Protein L, Protein A, and VEGF

The relative binding orientations of VEGF/G6 Fab variable domain complex (2FJG) with Protein A (1DEE, Protein A only) and Protein L (1HEZ, Protein L only) are shown in the composite structure.

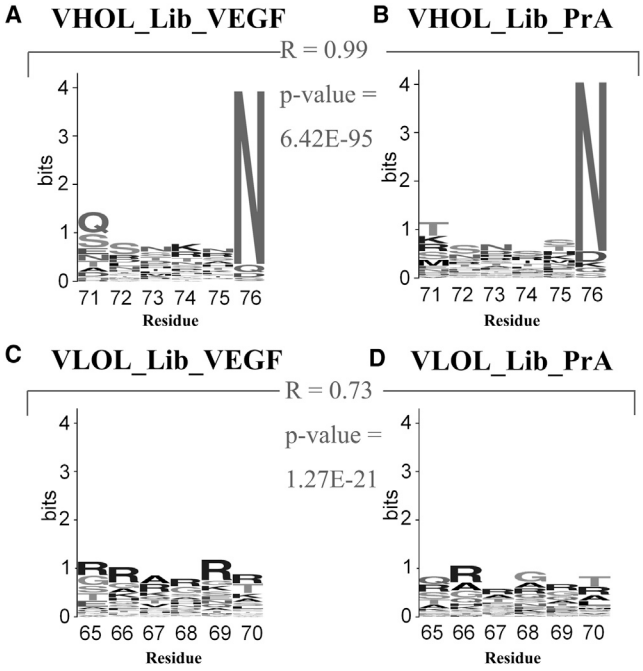

Figure 2. Comparisons of Phage-Display Selections for Protein A Binding and VEGF Binding

Sequence profiles, each of which contains sequence information for 40 scFv variants, were constructed with scFv variants from the VHOL (A and B) and VLOL (C and D) libraries selected for binding to VEGF (A and C) or Protein A (B and D). The x axis shows the loop-sequence residue positions in Kabat number, and the y axis shows the information content as described in Equations 1, 2, and 3.

and Protein A. In this work, we used VEGF-scFv binding variants to evaluate the local sequence preferences whenever possible because these variants are more relevant to the function of the CDRs of the scFvs. For the scFv variants with mutations in CDRs, we used the sequence preferences from the functional variants binding to Protein A, because VEGF-scFv binding was not detectable for a large portion of the CDR variants.

Thermal Stability Assessment of Functional scFv Variants with HTTI Experiments

HTTI experiments enable quick evaluations of the stability of scFv variants from the phage display selections without the need for laborious protein purification of the scFv variants. In an HTTI experiment, the scFv that is secreted into the culture supernatant from the expression host bacteria is heat treated at various temperatures. We used an ELISA platform to assess antigen binding without the need to purify the scFv protein molecules (Supplemental Experimental Procedures). The threshold temperature (T_{50}, where the scFv loses half of its binding capability to the antigen due to the heat treatment) has been established to correlate with the melting temperature, T_m, derived from differential scanning calorimetry (DSC) (Miller et al., 2010). Figure 3A compares the T_{50} of the template scFv Av1.2 for VEGF binding (i.e., $T_{50}^{Av1.2}$(VEGF)) with the T_{50} for Protein A binding (i.e., $T_{50}^{Av1.2}$(PrA)) in the presence or absence of a moderate concentration of 20 mM tris(2-carboxyethyl)phosphine (TCEP). We used TCEP to reduce the scFv intradomain disulfide bond upon thermal denaturation and thus prevent refolding of the heat-treated scFv. As expected, the irreversible heat denaturation condition resulted in lower T_{50} values in the presence of TCEP. This irreversible heat denaturation condition was used throughout this work. Figure 3B shows that the heat denaturation T_m of the purified template scFv measured with DSC in the absence of TCEP is similar to the corresponding T_{50} for Protein A binding ($T_{50}^{Av1.2}$(PrA)_TCEP(−) in Figure 3A), whereas the $T_{50}^{Av1.2}$(VEGF)_TCEP(+) is lower than the $T_{50}^{Av1.2}$(PrA)_TCEP(+) (Figure 3A), indicating that the CDRs are more sensitive to thermal inactivation than the overall framework region.

Figures 3C and 3D show the correlation between the T_{50}(VEGF) and the T_{50}(PrA) values for the scFv variants with loop mutations in the V_H and V_L domains, respectively. The high correlation between the two sets of T_{50} measurements in Figure 3C indicates that the effects on stability due to mutations in the V_H loops are felt equally by both binding sites. This is expected because the VEGF binding site, which consists mainly of V_H CDRs (Yu et al., 2012), and the Protein A binding site (situated on the V_H framework) are in close proximity to each other in the V_H domain (see the complex structures in Figure 1). The lower correlation between the two sets of T_{50} values in Figure 3D suggests that the stability effects due to the mutations in the V_L loops could affect the two binding sites through different paths. In this work, we used the T_{50} values for VEGF-scFv binding (i.e., T_{50}(VEGF)) to evaluate the stability of scFv variants whenever possible because these values are more relevant to the function of the CDRs of the scFvs. For the scFv variants with mutations in CDRs, we used the T_{50} values for Protein A-scFv binding (i.e., T_{50}(PrA)) because VEGF-scFv binding was not detectable for a large portion of the CDR variants.

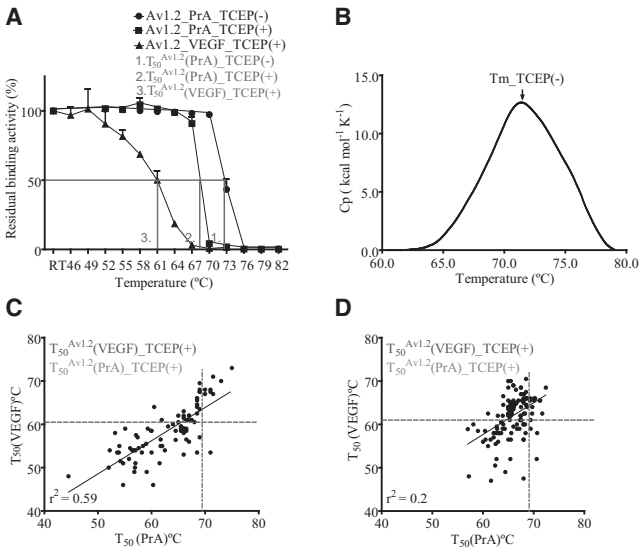

Figure 3. Thermal Inactivation Measurements for scFv-VEGF Binding and scFv-Protein A Binding

(A) Thermal inactivation of the template scFv (Av1.2) for VEGF binding in the presence of TCEP is indicated by triangles. Thermal inactivation of Av1.2 scFv for Protein A binding in the presence and absence of TCEP is indicated by squares and circles, respectively. The error bars were calculated with triplet repeats of the measurement. The T_{50} values are indicated with the intersection lines.

(B) DSC scan of purified soluble Av1.2 scFv measured in the absence of TCEP (Supplemental Experimental Procedures). The scFv expression and purification procedures are described in Supplemental Experimental Procedures.

(C and D) T_{50} values measured for VEGF-scFv binding are plotted against those measured for Protein A-scFv binding. Results for the scFv variants with loop mutations in the V_H and V_L domains are shown in (C) and (D), respectively. T_{50} was measured in the presence of TCEP.

Amino Acid Sequence Preferences of the Loops in Functional scFvs

The sequence preferences in the scFv loop regions of the V_L and V_H domains are summarized in the heatmaps shown in Figure 4 (Supplemental Experimental Procedures; Tables S3 and S4). The sequences of the functional scFv variants selected for VEGF binding were used to determine the sequence preferences for the non-CDR loop regions. These sequences were derived with NGS, and sequencing errors were eliminated as much as possible. In addition, redundant (100% identical) sequences were removed from sequence-preference calculations (Supplemental Experimental Procedures; Table S4). This was done to eliminate as much as possible any bias due to sequences that were more favorable in phage production, PCR preparation, or NGS sequencing, and thus ensure that the amino acid preferences (as shown in Figure 4) at each position were derived from diverse sequences (i.e., true amino acid preferences at a position should remain detectable in sequences with mutations of the neighboring residues). The sequence preferences for the six CDR loops were determined by sequencing functional scFv variants selected for Protein A binding. These scFv variants were experimentally confirmed for Protein A binding and the sequences were determined with Sanger sequencing (Supplemental Experimental Procedures; Table S3). The Protein A binding site does not overlap with the VEGF binding site formed by the CDRs (Figure 1), and both sites are known to be recognized by the corresponding binders only when the three-dimensional structure of the scFv is fully formed (Graille et al., 2000; Yu et al., 2012). Each cell of the heatmaps in Figures 4A and 4B shows the color-coded individual information content attributed to one of the 20 natural amino acid types (d_{ji} for amino acid type i [x axis] at position j [y axis]; see Equation 1). The sum of d_{ji} over the 20 amino acid types is the information content at position j (relative entropy, i.e., I_j of position j; see Equation 2), which is a measurement of the information divergence between the amino acid type probability distributions before and after the selection process for VEGF or Protein A binding. A larger value of I_j indicates that position j is more selective for amino acid types in functional scFv variants.

The loop residues can be categorized into five major groups based on their structural characteristics (Figures 4A and 4B): residue side chains exposed on the protein surface (labeled O), residues that form tight turns in the loop regions (labeled T), residue side chains that form the lower hydrophobic core in each of the variable domains (labeled L), residue side chains involved in hydrogen bonding or electrostatic interactions (labeled H), and residue side chains buried in the interface between the two variable domains (labeled I). The averaged relative entropy I_j values (see Equations 1, 2, and 3) for each residue position group are 0.72 ± 0.52 (O), 1.30 ± 0.75 (T), 2.75 ± 1.70 (L), 2.10 ± 1.85 (H), and 1.09 ± 0.80 (I) in the V_H domain, and 0.62 ± 0.33 (O), 0.85 ± 0.42 (T), 0.89 ± 0.70 (L), 1.08 ± 0.87 (H), and 1.89 ± 1.21 (I) in the V_L domain. Although the SDs of the averaged relative entropies are large, the trends of the quantitative measurements of the sequence requirement stringency indicate that, as expected, the loop residue positions with side chains protruding into the solvation environment are less selective in amino acid type than the residues for which the side chains are buried inside the protein structure and the side chains that involve hydrogen-bonding networks. Moreover, the V_H domain has more restrictions for the loop residue side-chain types that participate in the lower hydrophobic core (V_H: Val-H12, Val-H63, and Leu-H82C; V_L: Ala-L13, Val-L58, and Leu-L78) and form hydrogen-bonding networks (V_H: Arg-H38, Arg-H66, and Asp-H86; V_L: Gln-L37, Arg-L61, and Asp-L82; Figure 5) than the V_L domain, suggesting that the V_H loops are more tightly integrated into the overall structural packing.

Unexpectedly, in both variable domains, the residue positions forming the tight turns in the X-ray structure are not as selective as expected for the amino acid types with high turn propensities. It is well established that amino acid propensities in tight turns are statistically significant (Fuchs and Alix, 2005) and can be understood from the energetics of the local structures (Yang et al., 1996) and model peptide folding (Hsu et al., 2006). It is striking that most of the tight turn positions in the structure do not require amino acid types with high turn propensities (the only exception is the modest conservation in positions Gly-L16 and Gly-H15; see below and Figure 6 [LOGO column indicated by VEGF(+)]). This indicates that the tight turn loops could have alternative structures with varying sequence preferences. Quantitatively, the pairwise sequence covariation correlation coefficients (Φ) (Wang et al., 2009) were calculated for residue pairs in each of the loops. Residue pairs with a statistical meaningful

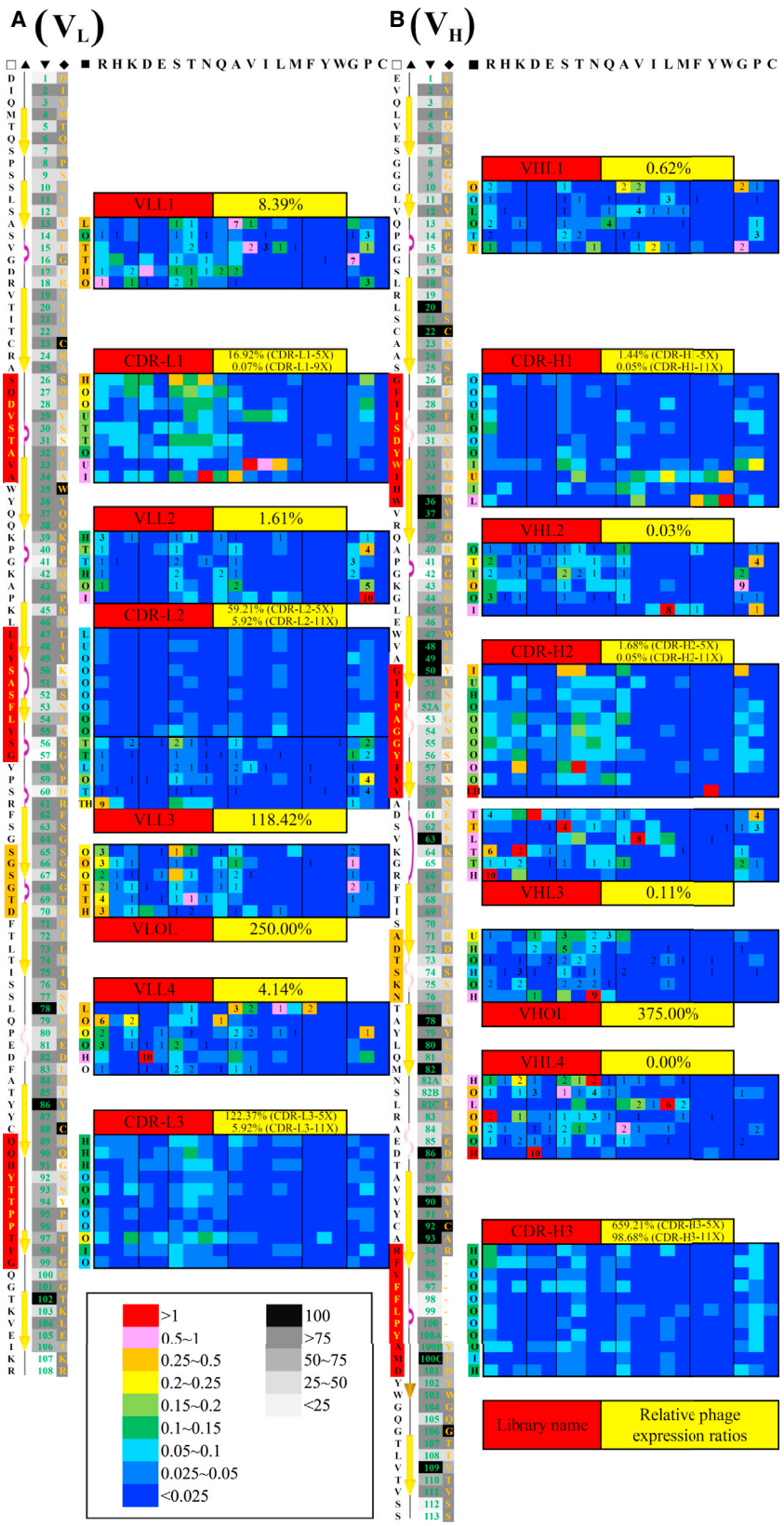

Figure 4. Sequence Determinants in Av1.2 scFv Loop Regions

(A and B) Heatmaps show the sequence preferences for the loop regions in the V_L (A) and V_H (B) domains. Columns from left to right: □ parent template Av1.2 scFv sequence, CDR regions (background in red; yellow font for short CDR versions), and outer loop (background in orange); ▲ secondary structure assignments from PDB; ▼ Kabat number (grayscale background shows the relative percentage of burial in folded scFv structure); ◆ human antibody consensus sequence (grayscale background shows the percentage consensus of the consensus sequence; consensus sequence and percentage consensus obtained from http://biochemistry.utoronto.ca/steipe/research/canonical.html); and ■ five major groups of distinguished structural characteristics (O, surface; T, tight turn; L, lower core; H, hydrogen-bonding and electrostatic interaction; I, interdomain interface). The colored background shows the range of the information content $0.25 \times I_j$ (Equation 2). R–C: sequence preference heatmaps derived from NGS data for the 20 amino acid types color-coded according to their d_{ji} values (Equation 1). The number in some of the cells indicates the appearance frequency (nearest integer of $10 \times q(j,i)$; $q(j,i)$ defined in Equation 3) of the amino acid type derived from the HTTI-filtered variants (see the corresponding main text). The relative phage expression ratio for each library is shown next to the name of the library above or below the heatmap (see also Figure S4).

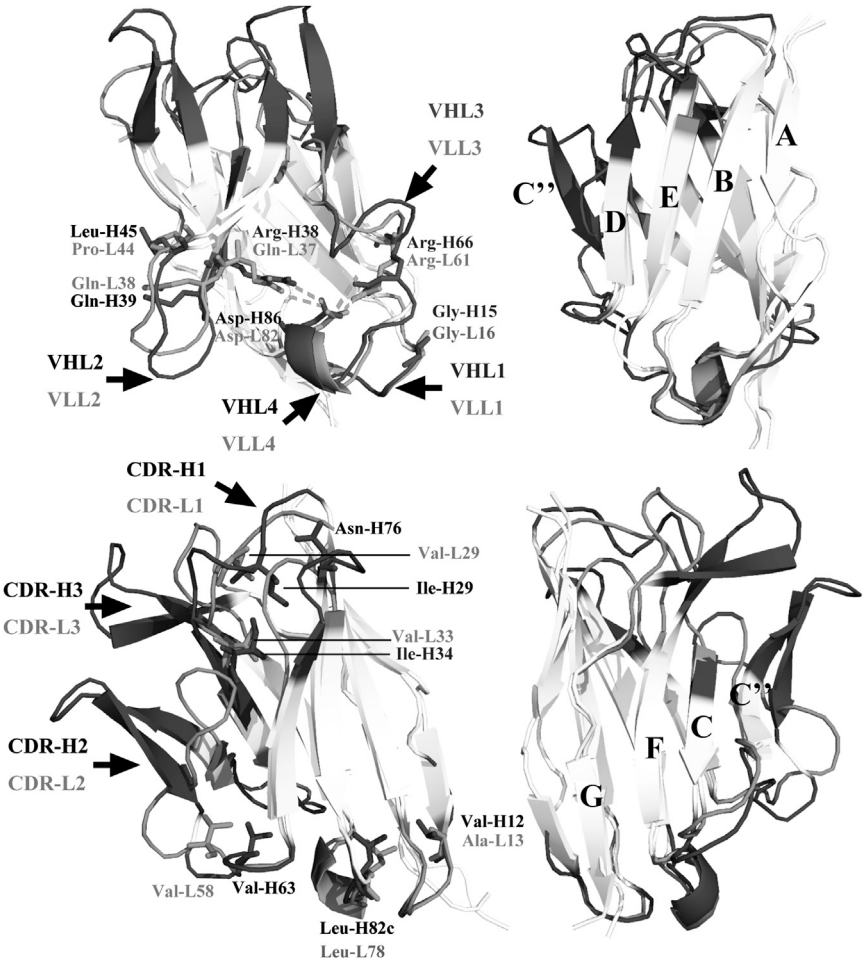

Figure 5. Superimposed V_H and V_L Domains of the Av1.2 scFv

The superimposed V_H and V_L domains of the template scFv structure (PDB code 2FJF) are labeled with structure-wise equivalent residues. The dark gray (V_H) and light gray (V_L) regions show the ranges of the consecutive loop residues in each of the phage-displayed libraries. The lightest-gray regions are the nonloop regions in both variable domains.

correlation coefficient ($p < 10^{-4}$ and $|\Phi| > 0.5$) and with contacts in the three-dimensional structure were not identified, supporting the above conclusion that the sequence features are not conserved for local turn structures.

Similar Sequence and Structural Features Shared by the Corresponding Loops in the V_H and V_L Domains

Loops VHL1, VHL2, and VHL4 in the V_H domain share similar structural and sequence features with the corresponding loops VLL1, VLL2, and VLL4 in the V_L domain, as indicated by the superimposed V_H and V_L domains of the template scFv structure shown in Figure 5 and the sequence preferences shown in Figures 4 and 6 (see below). These local structure pairs are highly similar in structure; the local root-mean-square deviation (rmsd) for the structurally aligned heavy atoms in the L1 (V_L:12-18; V_H:11-17), L2 (V_L:38-44; V_H: 39-45), and L4 (V_L:78-83; V_H:82C-87) loop pairs is 0.07, 0.44, and 0.37 Å respectively. The sequence preference patterns shown in Figures 4A and 4B (see also Figure 6) for these structure-wise equivalent regions share similar features. The glycines of the type II turns in both VLL1 and VHL1 loops (Gly-L16 and Gly-H15) are marginally conserved. In addition, the type II turns in both VLL2 (Pro-L40 and Gly-L41) and VHL2 (Pro-H41 and Gly-H42) loops are much less selective in amino acid types. Still, both VLL2 and VHL2 loops are anchored to the interface at one end by highly conserved residues (Pro-L44 and Leu-H45). The one-turn 3_{10} helices in both VLL4 and VHL4 share almost identical structural features (Figure 5) and similar sequence preference profiles (Figures 4A, 4B, and 6), where only Asp-L82 and Asp-H86 are highly conserved as the keystone residue in the hydrogen-bonding network in each variable domain capping the lower hydrophobic core. The similarity in structural features and sequence encodings in these loop regions suggests that these loops could play common roles in folding and stabilizing the V_H and V_L domains.

CDR-H1 and CDR-L1 are somewhat similar in terms of structural features and amino acid encodings. Both loops have two structure-wise equivalent hydrophobic residues (Val-L29 and Val-L33; Ile-H29 and Ile-H34) anchoring the CDR1 loop to the upper hydrophobic core in the variable domain (Figure 5), and these two hydrophobic residues are also frequently observed in known CDR structures. The hydrophobic preferences of these residue positions (Figures 4A and 4B), in particular Val-L33 and Ile-H34, indicate the importance of these residues in mediating both antigen binding and protein stability.

Differences between the Corresponding Loops in the V_H and V_L Domains

The V_H and V_L domains in the template scFv differ in structure mainly in the regions from CDR2 to the outer loop. As shown in Figure 5, CDR-H2 forms a β-hairpin as an extension to one of the β sheets in the β sandwich, where the cross-sheet VHL3 loop forms a compact local structure with two consecutive tight turns (type I followed by type II). In contrast, the shorter CDR-L2 adopts a local structure of shorter β strands followed by a loosely packed VLL3 loop with two tight turns (type II followed by type I). The sequence preferences in these regions reflect these structural differences. The β strands flanking the tip of the CDR-H2 loop are highly selective in amino acid type (Gly-H50, Thr-H57, and especially Tyr-H59; Figure 4B). In addition, the loop residues in the CDR-H2 have moderate preferences for hydrophilic or turn (Gly and Pro) amino acid types (Figure 4B). In contrast, the sequence preferences for

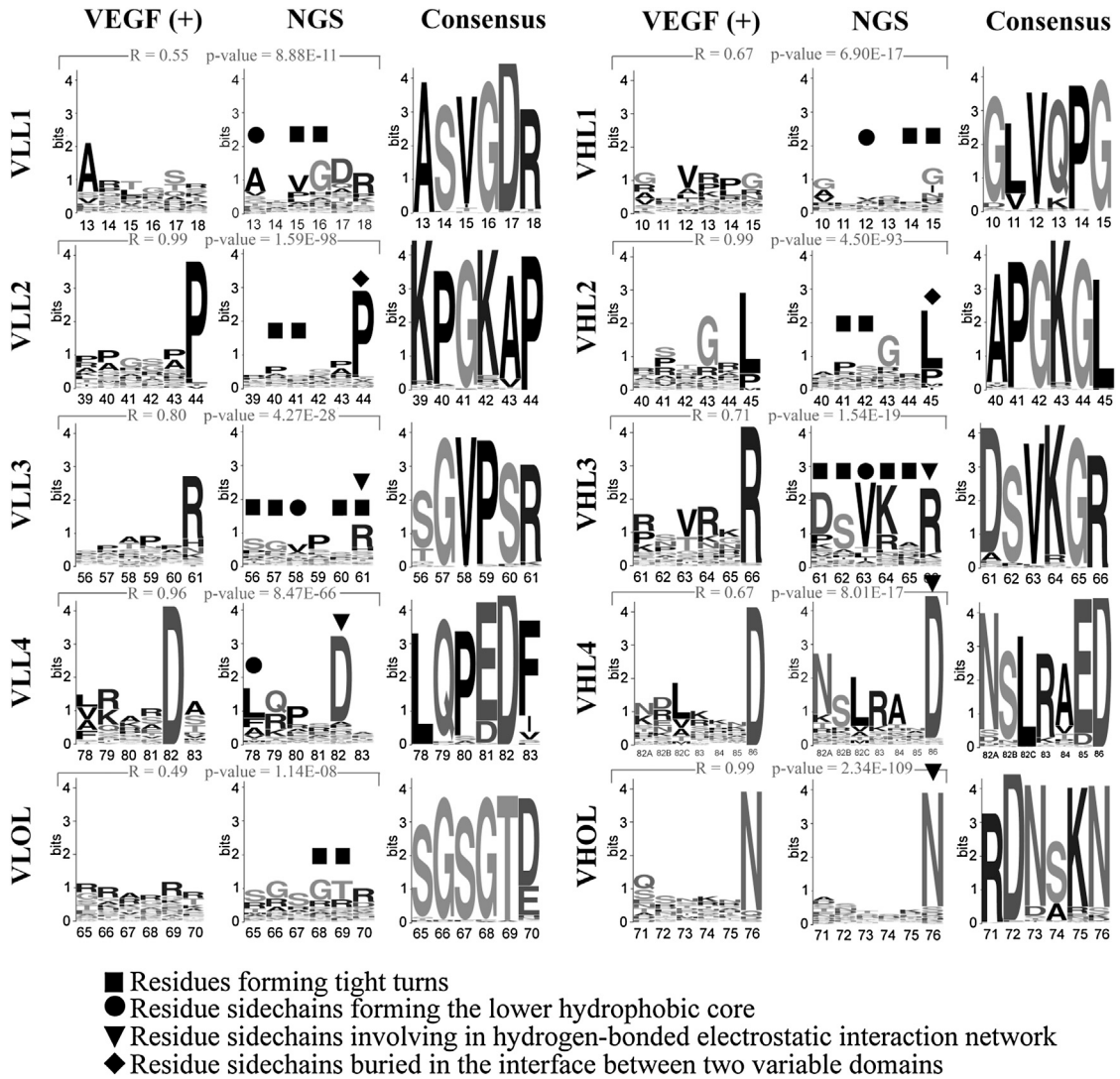

Figure 6. Comparisons of NGS Profiles, Sanger-Sequence Profiles, and Consensus-Sequence Profiles from the Natural Sequence Database
Sequence LOGOs of each non-CDR loop libraries derived from 40 randomly picked nonredundant VEGF-binding scFv variants sequenced by the Sanger method [labeled as VEGF(+)], NGS results (labeled as NGS), and database sequences from human antibody VH 3 and V kappa 1 family (labeled as Consensus). The y axis shows the I_i calculated with Equations 1, 2, and 3, and the x axis shows the amino acid positions (in Kabat number) in the template scFv. Residue positions: ■ for tight turn (T), ● for lower core (L), ▼ for hydrogen-bonding and electrostatic interaction (H), and ◆ for interdomain interface (I).

the CDR-L2 residue positions are essentially random (Figure 4A), indicating that the local structure of the CDR-L2 loop is flexible to accommodate a wide range of sequences. This is in agreement with the loose CDR-L2 structure seen in X-ray crystallography (Figure 5). The flexible structure extends to the VLL3 loop, where the sequence preferences are much less selective than the sequence preferences for the residues in its more compact counterpart (VHL3) in the V_H domain (Figures 4A and 4B). The lack of sequence preferences in these regions is in agreement with the high relative phage expression ratios (i.e., the phage-display efficiency of well-folded scFv variants from an scFv library compared with the template Av1.2 scFv; Supplemental Experimental Procedures; Figure S4) of the phage-displayed libraries for CDR-L2, VLL3, and VLOL. This is contrast to the CDR-H2 and VHL3 regions, where high sequence requirements limit the relative phage expression ratios of the corresponding phage-displayed libraries (relative phage expression ratios shown in Figures 4A, 4B, and S4). The relative phage expression ratio reflects the structural tolerance for the mutations introduced in the variants of the phage-displayed library in comparison with the template scFv.

CDR-H3 and CDR-L3 are grossly different in the loop structure (Figure 5), and the sequence preferences are the least selective among the loops (Figures 4A and 4B). Even the most distinguished features that are conserved in many antibodies (i.e., cis-Pro-L95 in CDR-L3 and the salt bridge Arg-H94…Asp-H101 in CDR-H3) are not conserved in the well-folded scFv

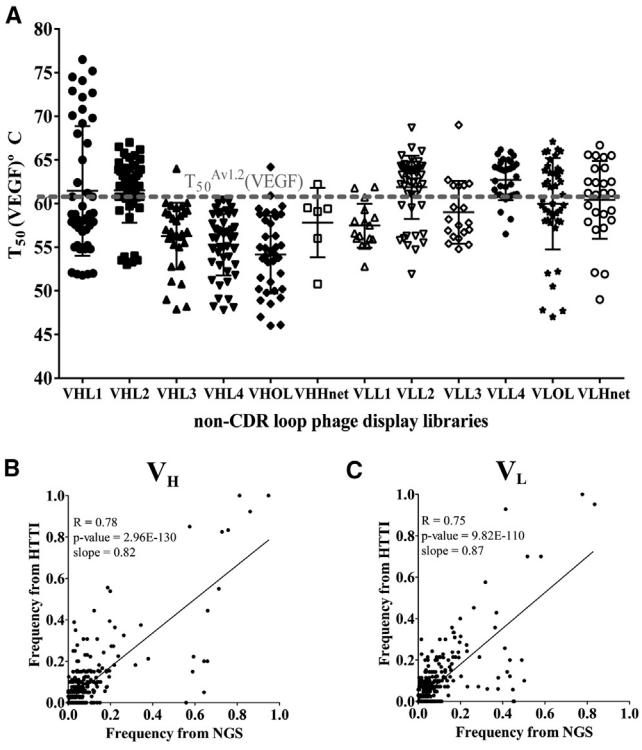

Figure 7. Thermal Stability Measurements of the Functional scFv Variants Selected from Each Non-CDR Loop Library

(A) T_{50}(VEGF) measurements (y axis) of the scFv variants selected from non-CDR loop libraries (x axis) for VEGF binding are presented with mean and SD bars superimposed on the data points. The dashed line indicates the T_{50} of template scFv (Av1.2) ($T_{50}^{Av1.2}$(VEGF) = 61°C; see Figure 3A). The sequence details are shown in Table S5.

(B) The plot shows the correlation of the two sets of frequencies ($q(j,i)$; see Equation 3) of amino acid types derived from NGS data (Figure 4B) and HTTI-filtered variants (Table S5) for the V_H domain. R is Pearson's correlation coefficient and the p value was calculated by t test.

(C) The same as in (B) for the V_L domain.

variants. It is well established that the diversity of the IgG variable domain sequence is largely the result of V(D)J gene recombination with the junctions taking place in the CDR3 loop (Nemazee, 2006). The indifference of the amino acids for the residues in the CDR3 loops shown in Figures 4A and 4B indicates that the sequence mutations around the gene-recombination junctions can be tolerated in the CDR3 loops without substantially affecting the folding and stability of the variable domains. The high somatic mutation rate in CDR3s during antibody development could be attributed to the tolerance of amino acid types for the residues in the CDR3 regions. The sequence-diversity tolerance of the CDR3 loops is in good agreement with the high relative phage expression ratios of the two loop libraries shown in Figures 4A and 4B.

The side chain of Asn-H76 in VHOL supports the CDR-H1 structure by forming hydrogen bonds with the backbone of CDR-H1 (Figure 5) and is thus highly conserved in amino acid type (Figures 4B and 6), emphasizing the importance of the outer loop in both folding and binding of the V_H domain. However, a similar interaction between VLOL and CDR-L1 has not been identified.

Comparisons of Loop-Sequence Profiles from NGS, Sanger Sequencing, and an Antibody Variable Domain Sequence Database

To further validate the NGS results, we carried out independent Sanger sequencing for VEGF-binding scFv variants. Figure 6 shows the Sanger sequencing results for the selected scFvs confirmed for VEGF binding by ELISA. These sequence preferences are compared with the NGS results from Figures 4A and 4B and with consensus-sequence profiles derived from natural antibody sequences. A large portion of the variants with mutations on the CDRs do not bind to VEGF, and thus variants for the CDR libraries are not included in Figure 6. The comparisons shown in Figure 6 indicate that the sequence preferences derived from NGS results are in good agreement with the Sanger sequencing results for the VEGF-binding scFvs (Pearson's correlation coefficients and t-test p values for the information content of the 20 amino acid types in the positions shown in the pairs of LOGO plots in Figure 6 are presented in the figure), confirming that the NGS statistics reflect the sequence requirements for VEGF binding of the scFv variants. Comparisons of the consensus-sequence profiles from the sequence database with the other two sets of sequence profiles show that only a few key residues in the non-CDR loops need to be conserved as in the consensus sequence (Arg-H66 [VHL3], Asp-H86 [VHL4], Asn-H76 [VHOL], Arg-L61 [VLL3], and Asp-L82 [VLL4] in hydrogen bonding, and Leu-H45 [VHL2] and Pro-L44 [VLL2] as interface hydrophobic residues), suggesting that although these few conserved residues are important for the function and structure of the antibody variable domains, the majority of the loop residues have a substantial tolerance for amino-acid-type substitutions.

To summarize, the sequence preferences of the non-CDR loops in both V_H and V_L variable domains are conserved only in a few residue positions involving intradomain tertiary hydrogen-bonding networks (residues marked by triangles in Figure 6) and interdomain hydrophobic packing (residues marked by diamonds in Figure 6). All of these interactions involve residue positions that are distant in sequence. Local sequence features for tight-turn structures (residues marked by squares in Figure 6) are not well defined in loop-sequence preferences, implying that sequence signals for local structures need not be specific as long as the keystone interactions that dictate the tertiary structure of the variable domain are uncompromised. In particular, the lack of sequence preferences in the regions ranging from CDR-L2 to VLOL explains the high relative phage expression ratios of the corresponding phage-displayed libraries (Figure 4A), the loose local structure (Figure 5), and the fact that there is room for further structure stability optimization (Figure 7; see below) in the V_L domain, but not in the corresponding well-folded region in the V_H domain. Moreover, although the loop residues that involve lower hydrophobic core packing (residues marked by circles in Figure 6) prefer hydrophobic amino acid types, the stringency of conservation is substantially weaker compared with positions involving hydrogen-bonding networks. Consequently, only a few residue positions in the loop regions need to strictly follow the consensus-sequence requirements for natural antibody sequences.

Thermal Stability Measurements of VEGF-Binding scFv Variants with HTTI

HTTI (T_{50}) measurements obtained from the individual variants provide a thermodynamic rationale for the coupling of the remote key residues in the non-CDR loops to the antigen binding through CDRs. Figure 7 shows the distributions of T_{50}(VEGF) values for scFv variants selected from each of the phage-displayed non-CDR libraries. The scFv variants were selected after a few rounds of selection-amplification cycle against immobilized VEGF and were confirmed for positive VEGF binding by ELISA (Supplemental Experimental Procedures; Table S5). The sequence preferences for thermal stable scFv variants are compared with the sequence preferences derived from NGS data as shown in Figures 4A and 4B. The numbers superimposed in the heatmaps of Figures 4A and 4B indicate the appearance frequency ($q(j,i)$ in Equation 3) of the amino acid type in each residue position of the non-CDR loops; the appearance frequencies were counted only in scFv variants with T_{50}(VEGF) > $T_{50}^{Av1.2}$(VEGF) − 5°C. These scFv variants were functionally confirmed to have similar or better stability compared with the template scFv Av1.2. The predominant key residues that emerge from the HTTI-filtered scFv variants (Table S5) are similar to the predominant key residues that emerge from the NGS statistics (Figures 4A and 4B; see residues Arg-H66 [VHL3], Asp-H86 [VHL4], Asn-H76 [VHOL], Arg-L61 [VLL3], and Asp-L82 [VLL4] in hydrogen bonding, and Leu-H45 [VHL2] and Pro-L44 [VLL2] as interface hydrophobic residues). The Pearson's correlation coefficients and t-test p values for the frequencies ($q(j,i)$ in Equation 3) of the amino acid types from the HTTI-filtered scFv variants compared with those from NGS statistics are shown in Figures 7B and 7C for the V_H and V_L domains, respectively. The highly significant correlations confirm the same set of key residues derived from NGS and HTTI, indicating that the key residue positions that are strongly associated with the antigen binding of the template scFv (Figures 4A, 4B and 5) are also critical for the thermal stability of the scFv structures (Figure 7; Table S5). This result supports the notion that the key residues that are distant from the functional site (CDRs) but are critical for the functional structure of the scFv variants are also critical for the thermal stability of the overall variable domain structure. This also indicates that the qualitative assumption that the energetic contribution from an individual residue to the overall protein stability is somewhat proportional to the relative entropy (Equation 1) of the residue, implying that the distributions of the amino acid types in the protein roughly follow the Boltzmann distribution reflecting the free energy of the protein structure.

The HTTI measurements shown in Figure 7 suggest that there is still a lot of room for improvement with regard to the thermal stability of the framework region of the template scFv based on the 4D5 variable domain sequences, which were optimized using a consensus approach. As shown in Figure 7A, a large portion of the functional, nonredundant scFv variants from the VHL1, VHL2, VLL2, VLL3, VLL4, and VLOL libraries have higher thermal stability than the 4D5-based template scFv. In good agreement with the above observations that the regions between CDR-L2 and VLOL are loosely consolidated in the V_L domain, the results in Figure 7A show the difference in stability optimization of these regions in the two variable domains. Local sequences in VLL3, VLL4, and VLOL can be further optimized to increase the stability over that of the scFv template. By contrast, the potential for stability improvement is relatively limited in the corresponding regions of the V_H domain, suggesting that the natural-consensus antibody variable domain sequences are not equally optimized for structural stability. The stability of the VH/LL1 and VH/LL2 regions in the template scFv is also far from optimal, suggesting that the tight-turn structures in these loops of the template scFv are not essential for loop formation.

The relatively low level of stability optimization of the light-chain variable domains could be a natural consequence of antibody development in vivo, where the antibody heavy chain evolves first in premature B cells before light-chain gene recombination and expression occur. Consequently, the heavy-chain variable domain is mostly responsible for key epitope recognition, and the chief function of the light-chain variable domain could be limited to supporting the heavy chain in protein folding and antigen binding (Nemazee, 2006).

Keystone Hydrogen-Bonding Networks Involve Polar Loop Residues that Cap the Bottom of the Lower Hydrophobic Core

Amino acids involving the hydrogen-bonding network that caps the bottom of the lower hydrophobic core (Figure 5) in each of the variable domains are particularly critical for protein stability, as shown in Figure 6. Recombinant phage-display libraries (VLHnet and VHHnet) with variants from saturated mutagenesis of the three noncontiguous segments encompassing the hydrogen-bonding network in each variable domain were constructed and selected for VEGF binding. The purpose of this experiment was to test whether any other configurations of the hydrogen-bonded electrostatic interactions could replace the native configuration. The sequence preferences for the hydrogen-bonding networks that emerge from the selected functional variants are shown in Figure 8, which indicates that the native configurations and key residue amino acid types are highly conserved, especially in the V_H domain. The covariation correlation coefficient (Φ) cannot be calculated for the residue pairs in the hydrogen-bonding networks because the highly conserved sequence features do not reveal the necessary amino acid substitution information for such a calculation. Nevertheless, it is clear, at least qualitatively, that these conserved residues are essential for the folding and stability of the variable domain structures, and that the polarity of the electrostatic interactions is part of a larger network and cannot be altered.

HTTI measurements of the scFv variants from the VLHnet and VHHnet libraries indicate that the hydrogen-bonding network region in the V_L domain has room for thermal-stability improvement. The sequence requirements for the hydrogen-bonding network residues in the V_H domain are much more stringent than those for the V_L domain (Figure 8), suggesting that the thermal stability of the V_H domain is not likely to be improved beyond that of the template scFv. The HTTI results show that the margin for thermal-stability improvement for the scFv variants from the VHHnet library is indeed quite limited (the best observed ΔT_{50}(VEGF) = 1.2°C; Figure 7; Table S5), and that the thermal-stability improvement for the scFv variants from the VLHnet library are relatively higher (the best observed ΔT_{50}(VEGF) = 5.7°C; Figure 7; Table S5). This result is in good agreement with the above conclusion that the thermal stability

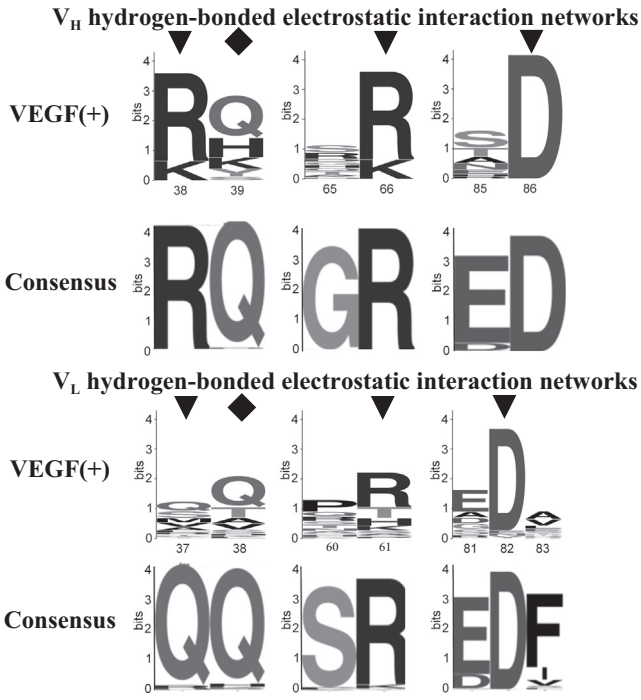

Figure 8. Sequence Preferences for the Hydrogen-Bonded Networks Capping the Lower Hydrophobic Core of the Variable Domains

Sequence LOGOs derived from 40 nonredundant functional (positive for VEGF binding) scFv variants selected from the libraries (VLHnet and VHHnet) of the noncontiguous hydrogen-bonding network residues are compared with the consensus-sequence LOGOs for both variable domains. The symbols labeling the key positions are defined in Figure 6.

of the non-CDR loops of the V_L domain can be optimized beyond that of the template scFv.

CDRs as Determinants of scFv Stability

It is anticipated that CDR-H1, CDR-L1, and CDR-H2 are more limited in sustaining sequence diversity in comparison with the other CDRs due to the more stringent sequence requirements of the three CDRs, as indicated in Figures 4A and 4B. To test this hypothesis, we measured the expression level of soluble scFv from each of the CDR libraries (CDR-L1-5X–CDR-H3-5X). The five central CDR residues for each library were selected according to the following criteria: they are on the tip of the CDR, are the most exposed to solvent, and are the least stringent in sequence requirements for each CDR region (see Figures 4A and 4B for these residue positions); that is, these residues are the least coupled with the main body of the variable domain.

Figures 9A and 9B show the soluble scFv expression levels (Supplemental Experimental Procedures) of the CDR recombinant libraries that were diversified by saturated mutagenesis. The CDR-H1-5X, CDR-L1-5X, and CDR-H2-5X libraries show a low expression level (Figure 9A), in good agreement with the hypothesis above. To further test the tolerance of sequence variation in the CDRs, we extended the libraries to encompass nine to 11 randomized residues centered on the five central residues (Figure S1). Figure 9B shows that, among the CDR libraries with a longer randomized sequence, only the CDR-H3-11X library could be expressed effectively. A possible explanation for this is that CDR-H3 is the longest of the CDRs (and thus is the least coupled to the framework), but it is also likely that the V_L domain is less stable and thus cannot sustain the same level of sequence variations as the V_H domain.

The above results imply that the population of scFv variants from the CDR-H3 library is energetically more stable than those from all other CDR libraries. We tested this hypothesis with thermal inactivation measurements. Figure 9C shows that the Protein A-based T_{50} values (T_{50}(PrA)) of the CDR-L1-5X, CDR-L2-5X, CDR-L3-5X soluble scFv libraries from the supernatant of the culture of the expression host bacteria are all substantially less than that of the template scFv(Av1.2), indicating that sequence variations in the CDRs of the V_L domain are not well tolerated and are mostly deleterious to the stability of the overall scFv variants. In comparison, Figure 9D shows the T_{50}(PrA) values of the CDR-H1-5X, CDR-H2-5X, CDR-H3-5X soluble variant libraries. Although CDR-H1 is as intolerant of sequence variation as CDR-L1, presumably due to the essential coupling of the CDR1 loop with the hydrophobic core (see above), CDR-H3 can tolerate sequence variations without substantially destabilizing the corresponding scFv variants, in agreement with the hypothesis. The sequence tolerance of CDR-H2 is less robust than that of CDR-H3, but nevertheless is better than that of the other CDRs. These results indicate that the CDRs are energetically coupled to the frameworks: amino acid types that are deleterious to loop formation (in general, hydrophobic and aromatic amino acids) destabilize the CDRs and in turn destabilize the framework. Quantitatively, CDR3s are the least coupled and CDR1s are the most coupled with the framework. CDRs in the V_H domain are less coupled with the framework compared with the corresponding CDRs in the V_L domain. These trends explain the library scFv expression levels shown in Figures 9A and 9B. The origin of the discrepancies in loop-framework coupling is not known exactly, but one explanation could be that the V_H framework is reinforced by more structured regions ranging from CDR-H2 to the outer loop (VHOL), but not in the V_L domain (see above), and as such, the framework of the V_H domain could sustain more sequence variations in the CDRs than the framework of the V_L domain.

The results in Figure 9 agree with the notion that the CDR loops contribute differently to the folding and stability of the variable domains. The Ig-like structures share a common folding mechanism whereby key hydrophobic residues from the B, C, E, and F β strands nucleate to form a central hydrophobic cluster, which defines the characteristic Ig-like folding topology. The peripheral segments, including the E–F β-arch (VH/LL4), the C′ β strand (CDR2 in V_H and V_L), and the B–C β arch (CDR1 in V_H and V_L), consolidate around the central nucleus, followed by the final docking of the A and G β strands and the final consolidation of the CDR3 loop structure (Billings et al., 2008; Cota et al., 2001; Fowler and Clarke, 2001; Geierhaas et al., 2004; Hamill et al., 2000a, 2000b; Lappalainen et al., 2008; Lorch et al., 1999). CDR1 sequence preferences, especially the conservation of the anchoring hydrophobic residues, reflect the importance of the CDR1 loop for the folding and stability of the variable domains. In contrast, the CDR3 loops can tolerate diverse sequences for antigen bonding, coupled with the stability of

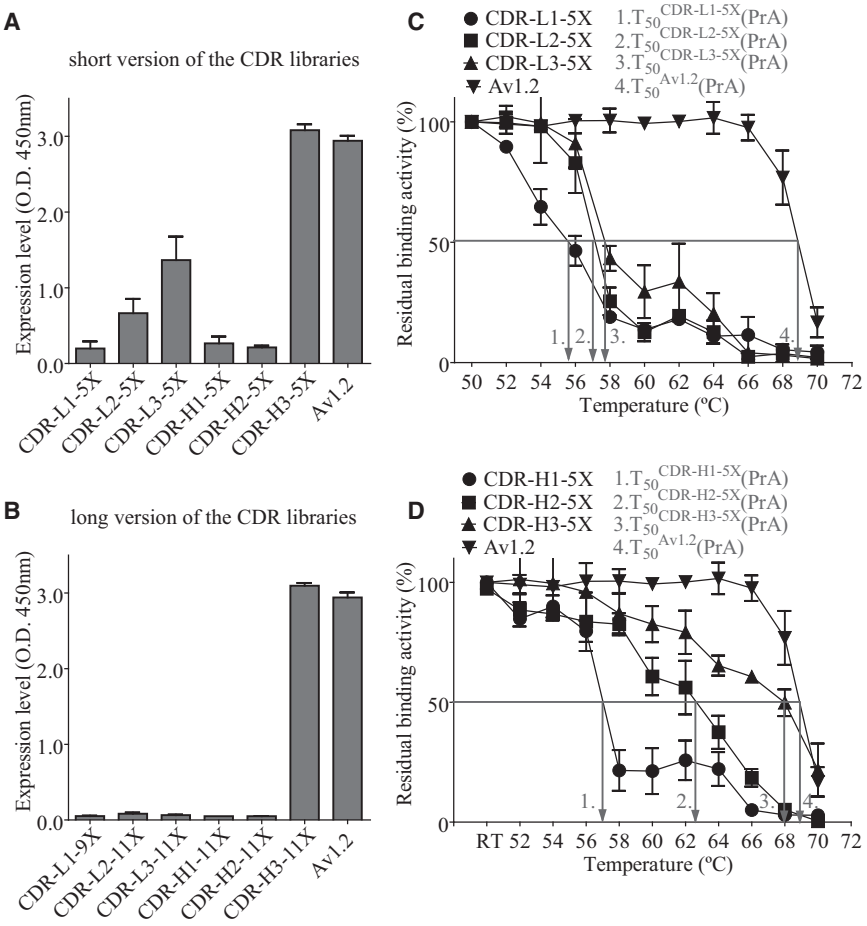

Figure 9. Expression Levels of the Soluble scFv Libraries and Thermal Inactivation Measurements of the scFvs from the CDR Libraries

(A and B) Expression levels for the short (A) and long (B) versions of the CDR libraries. The experimental procedures are described in Supplemental Experimental Procedures.

(C) Thermal inactivation measurements of the soluble scFv libraries CDR-L1-5X–CDR-L3-5X based on Protein A binding.

(D) Thermal inactivation measurements of the soluble scFv libraries CDR-H1-5X–CDR-H3-5X based on Protein A binding. The thermal inactivation measurement is described in Supplemental Experimental Procedures, and the randomized sequence ranges of the scFv CDR libraries are defined in Figure S1. The error bars were calculated with triplet repeats of the measurement.

the framework region only as a peripheral substructure that forms after the consolidation of the core region of the variable domain. Although CDR-H2 is confined to sequences that form β-hairpin structures, CDR-L2 likely contributes less to the stability of the V_L domain.

Conclusions

The emerging loop-sequence preferences reveal that the specificities of all the local structural codes are strikingly inferior to those of the predominant long-range tertiary interactions. None of the 16 loops in the antibody variable domains need to be encoded with conserved local sequence features. By contrast, residue positions involving in tertiary interactions are well conserved in the loop regions for protein folding, stability, and function. These residues, and particularly those in the hydrogen-bonding networks that cap the lower hydrophobic core, involve tertiary interactions that are strongly coupled to the function of the CDRs in recognizing the antigen, by energetically stabilizing the overall functional structure of the variable domain. The results also converge to support the folding mechanism of the hydrophobic core initiation, followed by the consolidation of the loop regions through tertiary interactions involving hydrophobic interactions and hydrogen bonding with residues distant in sequence. The nonspecific sequence preferences pertinent to local structural features are more in line with the notion that the loops in the IgG-like structure are passive regions to accommodate the configurations dictated by the formation of the hydrophobic core and hydrogen-bonding network. The alternative hypothesis that the variable domain loops are encoded with redundant turn-forming determinants that can only be revealed by simultaneous mutation of all loop sequences is less likely because the consolidation of the transition state (see above) implies a concerted folding involving all parts of the protein. This argues against the hypothesis that misbehavior in a certain part of the protein can always be compensated for by redundant structural signals encoded elsewhere. Although our experimental results are derived from the Av1.2 scFv template, it is likely that our conclusions can be generalized to a greater extent due to the high conservation of sequence and structural features among antibody gene families.

Results from this work will help to shape antibody engineering strategies to achieve a more robust scFv framework. Deep sequencing of the functional scFv variants revealed that the structure-wise equivalent positions in the non-CDR loops connecting the core B-C-C′ and E-F β strands, and nonconsecutive residues involving hydrogen-bonding networks capping the lower hydrophobic core in the V_H and V_L domains, are similar with regard to amino-acid-type encodings. Moreover, the more consolidated local structure in regions between CDR-H2 and VHOL in the V_H domain could confer extra stability to the V_H domain in comparison with the V_L domain, for which the shorter and less regular counterpart does not require strict sequence preferences and the local consensus sequence is not fully energetically optimized to support the overall structure. However, the stability of scFv constructs should not be limited by the weakness of the natural consensus sequence of the V_L domain. One obvious way to stabilize the scFv framework is to make the V_L domain as stable as the V_H domain by engineering the regions from CDR-L2 to VLOL in the V_L domain and then revising the

other non-CDR loop sequences. The HTTI results shown in Table S5 and Figure 7 demonstrate that many alternative loop sequences can indeed improve the thermal stability above that of the already robust 4D5 framework.

Taken together, yjr sequence preferences of a large number of functional scFv variants reveal that the sequences of the non-CDR loop regions are coupled to the central core of the variable domain, which in turn sustains the functional conformation of the CDRs. The CDR1 and CDR2 sequences are coupled with the stability of the variable domains as well as the antigen binding, and thus optimization of the sequences of these regions is more restricted in comparison with the CDR3 regions, which are less coupled with the framework and can sustain much larger sequence diversity. As such, optimizing the non-CDR loop sequences can substantially increase the thermal stability and expression level of the scFv variants, enabling greater complexity of the CDR sequence variation and thus more functional CDR configurations.

EXPERIMENTAL PROCEDURES

General experiment details are described in Supplemental Experimental Procedures.

Phage-Displayed scFv Library Construction and Selection

Construction of the scFv library and the selection procedures were done as described previously (Yu et al., 2012).

HTTI Measurement

HTTI measurements were performed as previously described (Miller et al., 2010) with minor modifications.

Information Content from Sequence Profiles

The individual information content d_{ji} attributed to amino acid type i at position j is shown in Equation 1, and the information content (or relative entropy) I_j at position j of a multiple amino acid sequence alignment is defined in Equation 2:

$$d_{ji} = q(j,i)I_j \tag{1}$$

$$I_j = \sum_{i=1, q(j,i)\neq 0}^{20} q(j,i)\log_2 \frac{q(j,i)}{p_i} \tag{2}$$

where,

$$q(j,i) = \frac{C_{ji}}{M_j} \tag{3}$$

where p_i is the background probability for amino acid type i, C_{ji} is the count of amino acid type i at position j of the multiple alignment, and M_j is the total count of 20 natural amino acid residues at position j of the multiple alignment. The background probabilities were calculated according to the DNA sequence statistics from randomly sampled scFv variants after the library constructions, and the statistics are shown in Table S2.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, four figures, and five tables and can be found with this article online at http://dx.doi.org/10.1016/j.str.2013.10.005.

AUTHOR CONTRIBUTIONS

H.-J.C., H.-J.H., and A.-S.Y. designed the research; H.-J.C., H.-J.H., Y.-C.L., H.-S.C., C.J.Y., S.-C.H., K.-P.C., and C.-Y.S. performed the research; H.-J.C., J.-W.J., H.-P.P., K.H.L., and A.-S.Y. analyzed data; and H.-J.C. and A.-S.Y. wrote the paper.

ACKNOWLEDGMENTS

We thank the Sequencing Core Facility, Scientific Instrument Center at Academia Sinica, for DNA sequencing. This work was supported by the National Science Council (NSC 100IDP006-3 and NSC 99-2311-B-001-014-MY3) and the Genomics Research Center at Academia Sinica (AS-100-TP2-B01). We also acknowledge the use of the nano DSC III in the Biophysics Core Facility, Scientific Instrument Center at Academia Sinica.

Received: June 4, 2013
Revised: October 11, 2013
Accepted: October 12, 2013
Published: November 21, 2013

REFERENCES

Billings, K.S., Best, R.B., Rutherford, T.J., and Clarke, J. (2008). Crosstalk between the protein surface and hydrophobic core in a core-swapped fibronectin type III domain. J. Mol. Biol. *375*, 560–571.

Bofill, R., and Searle, M.S. (2005). Engineering stabilising beta-sheet interactions into a conformationally flexible region of the folding transition state of ubiquitin. J. Mol. Biol. *353*, 373–384.

Chan, A.C., and Carter, P.J. (2010). Therapeutic antibodies for autoimmunity and inflammation. Nat. Rev. Immunol. *10*, 301–316.

Cota, E., Steward, A., Fowler, S.B., and Clarke, J. (2001). The folding nucleus of a fibronectin type III domain is composed of core residues of the immunoglobulin-like fold. J. Mol. Biol. *305*, 1185–1194.

Demarest, S.J., and Glaser, S.M. (2008). Antibody therapeutics, antibody engineering, and the merits of protein stability. Curr. Opin. Drug Discov. Devel. *11*, 675–687.

Eigenbrot, C., Randal, M., Presta, L., Carter, P., and Kossiakoff, A.A. (1993). X-ray structures of the antigen-binding domains from three variants of humanized anti-p185HER2 antibody 4D5 and comparison with molecular modeling. J. Mol. Biol. *229*, 969–995.

Ewert, S., Honegger, A., and Plückthun, A. (2003). Structure-based improvement of the biophysical properties of immunoglobulin VH domains with a generalizable approach. Biochemistry *42*, 1517–1528.

Fowler, S.B., and Clarke, J. (2001). Mapping the folding pathway of an immunoglobulin domain: structural detail from Phi value analysis and movement of the transition state. Structure *9*, 355–366.

Fowler, D.M., Araya, C.L., Fleishman, S.J., Kellogg, E.H., Stephany, J.J., Baker, D., and Fields, S. (2010). High-resolution mapping of protein sequence-function relationships. Nat. Methods *7*, 741–746.

Fuchs, P.F., and Alix, A.J. (2005). High accuracy prediction of beta-turns and their types using propensities and multiple alignments. Proteins *59*, 828–839.

Fuh, G., Wu, P., Liang, W.-C., Ultsch, M., Lee, C.V., Moffat, B., and Wiesmann, C. (2006). Structure-function studies of two synthetic anti-vascular endothelial growth factor Fabs and comparison with the Avastin Fab. J. Biol. Chem. *281*, 6625–6631.

Geierhaas, C.D., Paci, E., Vendruscolo, M., and Clarke, J. (2004). Comparison of the transition states for folding of two Ig-like proteins from different superfamilies. J. Mol. Biol. *343*, 1111–1123.

Graille, M., Stura, E.A., Corper, A.L., Sutton, B.J., Taussig, M.J., Charbonnier, J.B., and Silverman, G.J. (2000). Crystal structure of a Staphylococcus aureus protein A domain complexed with the Fab fragment of a human IgM antibody: structural basis for recognition of B-cell receptors and superantigen activity. Proc. Natl. Acad. Sci. USA *97*, 5399–5404.

Hamill, S.J., Cota, E., Chothia, C., and Clarke, J. (2000a). Conservation of folding and stability within a protein family: the tyrosine corner as an evolutionary cul-de-sac. J. Mol. Biol. *295*, 641–649.

Hamill, S.J., Steward, A., and Clarke, J. (2000b). The folding of an immunoglobulin-like Greek key protein is defined by a common-core nucleus and regions constrained by topology. J. Mol. Biol. *297*, 165–178.

Hietpas, R.T., Jensen, J.D., and Bolon, D.N. (2011). Experimental illumination of a fitness landscape. Proc. Natl. Acad. Sci. USA *108*, 7896–7901.

Holliger, P., and Hudson, P.J. (2005). Engineered antibody fragments and the rise of single domains. Nat. Biotechnol. *23*, 1126–1136.

Honegger, A. (2008). Engineering antibodies for stability and efficient folding. Handb. Exp. Pharmacol. *181*, 47–68.

Hsu, H.J., Chang, H.J., Peng, H.P., Huang, S.S., Lin, M.Y., and Yang, A.S. (2006). Assessing computational amino acid beta-turn propensities with a phage-displayed combinatorial library and directed evolution. Structure *14*, 1499–1510.

Huang, Y.J., Chen, I.C., Yu, C.M., Lee, Y.C., Hsu, H.J., Ching, A.T., Chang, H.J., and Yang, A.S. (2010). Engineering anti-vascular endothelial growth factor single chain disulfide-stabilized antibody variable fragments (sc-dsFv) with phage-displayed sc-dsFv libraries. J. Biol. Chem. *285*, 7880–7891.

Jager, M., Deechongkit, S., Koepf, E.K., Nguyen, H., Gao, J., Powers, E.T., Gruebele, M., and Kelly, J.W. (2008). Understanding the mechanism of beta-sheet folding from a chemical and biological perspective. Biopolymers *90*, 751–758.

Jermutus, L., Honegger, A., Schwesinger, F., Hanes, J., and Plückthun, A. (2001). Tailoring in vitro evolution for protein affinity or stability. Proc. Natl. Acad. Sci. USA *98*, 75–80.

Jespers, L., Schon, O., Famm, K., and Winter, G. (2004). Aggregation-resistant domain antibodies selected on phage by heat denaturation. Nat. Biotechnol. *22*, 1161–1165.

Jordan, J.L., Arndt, J.W., Hanf, K., Li, G., Hall, J., Demarest, S., Huang, F., Wu, X., Miller, B., Glaser, S., et al. (2009). Structural understanding of stabilization patterns in engineered bispecific Ig-like antibody molecules. Proteins *77*, 832–841.

Jung, S., Honegger, A., and Plückthun, A. (1999). Selection for improved protein stability by phage display. J. Mol. Biol. *294*, 163–180.

Kügler, M., Stein, C., Schwenkert, M., Saul, D., Vockentanz, L., Huber, T., Wetzel, S.K., Scholz, O., Plückthun, A., Honegger, A., and Fey, G.H. (2009). Stabilization and humanization of a single-chain Fv antibody fragment specific for human lymphocyte antigen CD19 by designed point mutations and CDR-grafting onto a human framework. Protein Eng. Des. Sel. *22*, 135–147.

Lappalainen, I., Hurley, M.G., and Clarke, J. (2008). Plasticity within the obligatory folding nucleus of an immunoglobulin-like domain. J. Mol. Biol. *375*, 547–559.

Lorch, M., Mason, J.M., Clarke, A.R., and Parker, M.J. (1999). Effects of core mutations on the folding of a beta-sheet protein: implications for backbone organization in the I-state. Biochemistry *38*, 1377–1385.

Marcelino, A.M., and Gierasch, L.M. (2008). Roles of beta-turns in protein folding: from peptide models to protein engineering. Biopolymers *89*, 380–391.

McCallister, E.L., Alm, E., and Baker, D. (2000). Critical role of beta-hairpin formation in protein G folding. Nat. Struct. Biol. *7*, 669–673.

Miller, B.R., Demarest, S.J., Lugovskoy, A., Huang, F., Wu, X., Snyder, W.B., Croner, L.J., Wang, N., Amatucci, A., Michaelson, J.S., and Glaser, S.M. (2010). Stability engineering of scFvs for the development of bispecific and multivalent antibodies. Protein Eng. Des. Sel. *23*, 549–557.

Monsellier, E., and Bedouelle, H. (2006). Improving the stability of an antibody variable fragment by a combination of knowledge-based approaches: validation and mechanisms. J. Mol. Biol. *362*, 580–593.

Nauli, S., Kuhlman, B., and Baker, D. (2001). Computer-based redesign of a protein folding pathway. Nat. Struct. Biol. *8*, 602–605.

Nelson, A.L., and Reichert, J.M. (2009). Development trends for therapeutic antibody fragments. Nat. Biotechnol. *27*, 331–337.

Nemazee, D. (2006). Receptor editing in lymphocyte development and central tolerance. Nat. Rev. Immunol. *6*, 728–740.

North, B., Lehmann, A., and Dunbrack, R.L., Jr. (2011). A new clustering of antibody CDR loop conformations. J. Mol. Biol. *406*, 228–256.

Schlinkmann, K.M., Honegger, A., Türeci, E., Robison, K.E., Lipovšek, D., and Plückthun, A. (2012). Critical features for biosynthesis, stability, and functionality of a G protein-coupled receptor uncovered by all-versus-all mutations. Proc. Natl. Acad. Sci. USA *109*, 9810–9815.

Sharpe, T., Jonsson, A.L., Rutherford, T.J., Daggett, V., and Fersht, A.R. (2007). The role of the turn in beta-hairpin formation during WW domain folding. Protein Sci. *16*, 2233–2239.

Wang, N., Smith, W.F., Miller, B.R., Aivazian, D., Lugovskoy, A.A., Reff, M.E., Glaser, S.M., Croner, L.J., and Demarest, S.J. (2009). Conserved amino acid networks involved in antibody variable domain interactions. Proteins *76*, 99–114.

Weatherill, E.E., Cain, K.L., Heywood, S.P., Compson, J.E., Heads, J.T., Adams, R., and Humphreys, D.P. (2012). Towards a universal disulphide stabilised single chain Fv format: importance of interchain disulphide bond location and vL-vH orientation. Protein Eng. Des. Sel. *25*, 321–329.

Whitehead, T.A., Chevalier, A., Song, Y., Dreyfus, C., Fleishman, S.J., De Mattos, C., Myers, C.A., Kamisetty, H., Blair, P., Wilson, I.A., and Baker, D. (2012). Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. Nat. Biotechnol. *30*, 543–548.

Wörn, A., and Plückthun, A. (1998). Mutual stabilization of VL and VH in single-chain antibody fragments, investigated with mutants engineered for stability. Biochemistry *37*, 13120–13127.

Wörn, A., and Plückthun, A. (1999). Different equilibrium stability behavior of ScFv fragments: identification, classification, and improvement by protein engineering. Biochemistry *38*, 8739–8750.

Wörn, A., and Plückthun, A. (2001). Stability engineering of antibody single-chain Fv fragments. J. Mol. Biol. *305*, 989–1010.

Yang, A.S., Hitz, B., and Honig, B. (1996). Free energy determinants of secondary structure formation: III. beta-turns and their role in protein folding. J. Mol. Biol. *259*, 873–882.

Yu, C.M., Peng, H.P., Chen, I.C., Lee, Y.C., Chen, J.B., Tsai, K.C., Chen, C.T., Chang, J.Y., Yang, E.W., Hsu, P.C., et al. (2012). Rationalization and design of the complementarity determining region sequences in an antibody-antigen recognition interface. PLoS ONE *7*, e33340.