Structure
Article

Antibody Variable Domain Interface and Framework Sequence Requirements for Stability and Function by High-Throughput Experiments

Hung-Ju Hsu,[1,2] Kuo Hao Lee,[2] Jhih-Wei Jian,[2,3,4] Hung-Ju Chang,[2,5,6] Chung-Ming Yu,[2] Yu-Ching Lee,[2] Ing-Chien Chen,[2] Hung-Pin Peng,[2,3,4] Chih Yuan Wu,[2] Yu-Feng Huang,[2] Chih-Yun Shao,[2,7] Kuo Ping Chiu,[2] and An-Suei Yang[1,2,*]
[1]Graduate Institute of Life Sciences, National Defense Medical Center, Taipei 114, Taiwan
[2]Genomics Research Center, Academia Sinica, Taipei 115, Taiwan
[3]Institute of Biomedical Informatics, National Yang-Ming University, Taipei 112, Taiwan
[4]Bioinformatics Program, Taiwan International Graduate Program, Institute of Information Science, Academia Sinica, Taipei 115, Taiwan
[5]Institute of Biochemical Science, National Taiwan University, Taipei 106, Taiwan
[6]Chemical Biology and Molecular Biophysics program, Taiwan International Graduate Program at Academia Sinica, Taipei 115, Taiwan
[7]Institute of Zoology, College of Life Sciences, National Taiwan University, Taipei 106, Taiwan
*Correspondence: yangas@gate.sinica.edu.tw
http://dx.doi.org/10.1016/j.str.2013.10.006

SUMMARY

Protein structural stability and biological functionality are dictated by the formation of intradomain cores and interdomain interfaces, but the intricate sequence-structure-function interrelationships in the packing of protein cores and interfaces remain difficult to elucidate due to the intractability of enumerating all packing possibilities and assessing the consequences of all the variations. In this work, groups of β strand residues of model antibody variable domains were randomized with saturated mutagenesis and the functional variants were selected for high-throughput sequencing and high-throughput thermal stability measurements. The results show that the sequence preferences of the intradomain hydrophobic core residues are strikingly flexible among hydrophobic residues, implying that these residues are coupled indirectly with antigen binding through energetic stabilization of the protein structures. By contrast, the interdomain interface residues are directly coupled with antigen binding. The interdomain interface should be treated as an integral part of the antigen-binding site.

INTRODUCTION

Antibodies are becoming increasingly important as therapeutics. Single-chain antibody fragments (scFv) are fundamental building blocks for antibody-based therapeutics (Caravella and Lugovskoy, 2010; Demarest and Glaser, 2008; Miller et al., 2010; Holliger and Hudson, 2005; Huang et al., 2010; Nelson and Reichert, 2009), which are frequently engineered with powerful high-throughput platforms based on in vitro scFv display on ribosome, phage, yeast, and other microbe surfaces (Bradbury et al., 2011; Hoogenboom, 2005). Because of the enormous potential in antibody therapeutics, many approaches have been developed to optimize the scFv stability by screening stable variants generated with random mutagenesis (Jermutus et al., 2001; Jespers et al., 2004; Jung et al., 1999) and by rational designs based on the consensus sequence principle (Steipe, 2004) in protein stability engineering (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). Still, the intricate sequence-structure-function interrelationships in the variable domains and the interdomain interfaces of antibodies remain insufficiently understood due to the limited scope in enumerating all packing possibilities and in assessing the consequences in protein stability and function due to the variations.

The general structural principles of the hydrophobic cores surrounded by the β strands in antibody variable domains have been the primary focus in protein engineering experiments on immunoglobulin (Ig)-like structures. Φ-value analyses on the folding transition states of various Ig-like structures that are not obviously related in sequence homology have indicated that the Ig-like structures share a common folding mechanism where one key hydrophobic residue from each of the B, C, E, and F β strands nucleates to form a central hydrophobic cluster, which defines the characteristic Ig-like folding topology. The peripheral segments—including the E-F β arch, the C′ β strand, and, to a lesser degree, the B-C β arch—consolidate around the central nucleus, followed by the final docking of the A and G β strands (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). The results have converged to the conclusion that the stability of the Ig-like proteins is predominantly supported by the hydrophobic core residues, which are also thought to play an important role in guiding the Ig-like folding (Clarke et al., 1999). Although a more recent experiment focused on the effects of the loop and other surface residues on the Ig-like protein folding has revealed the influences of the surface residues in the folding mechanism (Billings et al., 2008), the hydrophobic core surrounded by the β strands evidently remains the key determinant for the antibody variable domain's folding, stability, and thus function.

Part of the primary knowledge enabling the basis for engineering antibody stability and function involves rationalizing the coupling between the hydrophobic side-chain packing in and between the variable domains and the antigen binding of the complementarity determining regions (CDRs). The coupling of protein functions with residues remote from the functional sites has been a main concern in protein engineering, as a large body of experimental evidence indicates that mutations frequently have substantial long-range influences on the functions of proteins. To focus on the effects of tertiary hydrophobic interactions on antibody stability and function, we sought to enumerate the hydrophobic packing configurations in the framework cores and in the interdomain interface of a model scFv so as to assess the corresponding protein stability and antigen affinity due to the packing variations. This information would make clear the hypothesis on the coupling of the tertiary hydrophobic interactions with the antigen binding sites on the CDRs.

In this work, a total of 27 groups of consecutive or noncontiguous residues (three to nine residues) in the β strands of a human vascular endothelial growth factor (VEGF)-binding scFv template were randomized with saturated mutagenesis in phage-displayed scFv libraries (Yu et al., 2012). The functional scFv variants after a few rounds of selection for VEGF binding were sequenced with next-generation sequencing (NGS). The stabilities of randomly selected functional (VEGF-binding or protein-A-binding) scFv variants were also assessed quantitatively with high-throughput thermal inactivation (HTTI) measurements (Miller et al., 2010). In contrast to the recent NGS studies focusing on deriving sequence preference statistics from single-mutation-per-variant libraries (Fowler et al., 2010; Hietpas et al., 2011; Schlinkmann et al., 2012; Whitehead et al., 2012), variants with simultaneous changes of several (three to nine) residues in the hydrophobic packing would reveal the cooperativity of the interacting residues. Also, unlike the natural antibody sequences, the functional scFv variants from the recombinant antibody libraries are largely free from phylogenetic biases (Jäckel et al., 2010). In addition, the large-scale stability measurements of functional scFv variants provide a thermodynamics context corresponding to the sequence variations. The results show that sequence preferences of the intradomain hydrophobic core residues are flexible among hydrophobic residues, with many alternative combinations enabling variable domains with superior thermal stability to that of the variable domains encoded with consensus sequences. By contrast, the interdomain interface residues of the functional scFv variants binding to the antigen are conserved with high stringency, suggesting that the interdomain interface residues distant to the antigen-binding site are directly coupled to the antibody-antigen interaction. The knowledge gained in these aspects is expected to further the understanding of sequence-structure-function relationships in general and to enhance the basis for rational antibody engineering for stability and function in particular.

RESULTS AND DISCUSSION

Selection of Functional scFv Variants from 27 Phage-Displayed Libraries

The main goal of the work is to elucidate the coupling of the antibody-antigen interaction with the residues in the β strand regions distant from the antibody-antigen binding site. Consecutive amino acid residues in the β strand regions of an scFv were randomized in phage-displayed libraries with saturated mutagenesis (see Figure S1 and Supplemental Experimental Procedures available online). Each of the libraries contained around 10^9 scFv variants (Tables S1 and S2) from randomization of three to nine residues in the β strand regions of a template scFv (dubbed Av1). The template scFv Av1 has been constructed (Yu et al., 2012) on the basis of the 4D5 antibody framework (Eigenbrot et al., 1993; Fuh et al., 2006) with the huV$_\kappa$1-huV$_H$3 scFv construct (Figure S1). The 4D5 framework had been optimized with the consensus sequence approach and the robust stability of the 4D5 scFv had been demonstrated (Wörn and Plückthun, 1999). The functional scFv variants were selected with three to four rounds of a phage-display selection-amplification cycle for binding to immobilized VEGF (Supplemental Experimental Procedures; Figure S2). The selected scFv variants were then deep sequenced with 454 pyrosequencing (Supplemental Experimental Procedures; Table S4). Residue positions that are highly conserved for VEGF binding are expected to be coupled with the antigen-binding site in the scFv.

Sequence Preferences in the β Strand Regions of the Functional scFv Variants

Figures 1A and 1B shows the deep sequencing results in color-coded heatmaps for amino acid preferences and substitution patterns at the β residue positions in the V_L and V_H domain of the template scFv. Redundant (100% identical) sequences were removed from sequence preference calculation (Supplemental Experimental Procedures). This is to eliminate (as much as possible) the biases due to sequences that are more favorable in phage production, in PCR preparation, or in NGS sequencing so as to ensure that the amino acid preferences (as shown in Figure 1) at each position are derived from diverse sequences. Each cell of the heatmaps shows the color-coded individual information content attributed to one of the 20 natural amino acids; i.e., d_{ji} for amino acid type i (x axis) at position j (y axis) (see Equation 1 for d_{ji}). The sum of d_{ji} over the 20 amino acid types is the information content at the position j; i.e., relative entropy I_j of the position j (see Equation 2 for I_j), which is a measurement of the information divergence between the amino acid type probability distributions before and after the selection process for VEGF binding; a larger value of I_j indicates that the position j is more selective toward the amino acid types with high d_{ji} value. The relative entropy for each of the variable domain positions is shown in Figures 1A and 1B next to the heatmaps. As shown in Figure 1, many positions in the functional scFvs are highly conserved in amino acid types. It is of interest to further understand the principles underlying these relatively conserved residue positions. These are the remote residue positions that are nevertheless strongly coupled with the antigen-binding site of the functional scFv variants.

The deep sequencing results reveal that amino acid type preferences for each of the residue positions in the β strand regions of the selected scFvs are, as expected, dependent on the burial level of the residue side chains and that each residue position is marked by the level of promiscuousness in amino acid type and by the amino acid substitution pattern, depending on structural roles of the residue position. As shown in Figures 1A and 1B,

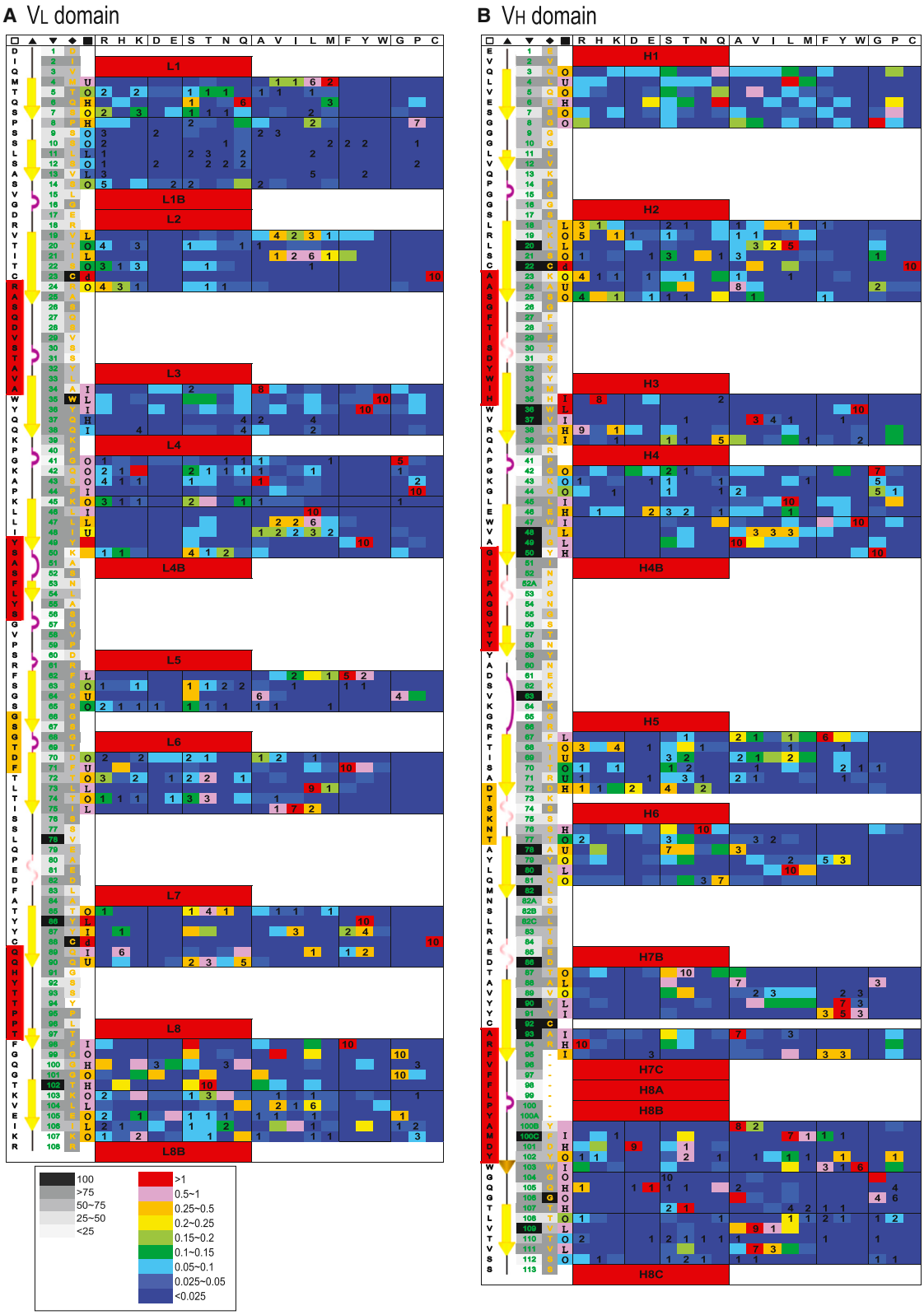

Figure 1. Sequence Preferences Derived from NGS Data in the β Strand Regions of the Model Av1 scFv

The heatmaps show the sequence preferences for the β strand regions of the V_L domain (A) and the V_H domain (B). The columns from left to right show information on □, template Av1 scFv sequence, CDR regions (background in red), and outer loop (background in orange); ▲, secondary structure assignments from the

(legend continued on next page)

the residue positions in the scFv β strand regions can be categorized into five major groups of distinguished structural characteristics: residue side chains exposed on the protein surface (labeled "O"), residue side chains forming the upper core in each of the variable domains (labeled "U"), residue side chains forming the lower hydrophobic core in each of the variable domains (labeled "L"), residue side chains involving in hydrogen-bonding or electrostatic interactions (labeled "H"), and residue side chains buried in the interface between the two variable domains (labeled "I"). The average relative entropy I_j for each of the residue position groups in V_L is 1.28 ± 0.68 (O), 2.08 ± 0.94 (U), 2.05 ± 1.34 (L), 1.74 ± 0.97 (H), and 2.19 ± 0.94 (I), respectively, and in V_H is 1.53 ± 1.12 (O), 1.59 ± 1.09 (U), 2.40 ± 0.93 (L), 2.66 ± 1.05 (H), and 2.67 ± 1.12 (I), respectively. Although the SDs are relatively large, the average relative entropies above indicate the general trends that the residue positions with side chains protruding into solvation environment are the least selective in amino acid type and that the residue positions in the lower and the upper cores of the variable domains are more selective in amino acid type, but not as selective as the residue positions buried in the interface between the two variable domains.

Residue side chains exposed to solvent (relative solvent exposure > 25%) are mostly interchangeable in polar and charged amino acid types, except for the C-terminal framework regions in both variable domains, L98–L102 and H103–H107. The sequence preferences of the C-terminal surface residues are attributed to the conserved structure-wise critical glycine (L99, L101, H104, H106). The only unexpectedly conserved surface residues are Lys-L42 and Ala-L43, but more careful investigation with Sanger sequencing of the functional variants concluded that these two residues are not necessarily conserved for protein structure and function (see Figure 2A below).

The hydrophobic core residues in the variable domains are unanimously hydrophobic, but the amino acid type substitutions are differentially tolerated from position to position. The hydrophobic core between the two β sheets in an immunoglobulin variable domain is divided in the middle by a layer of consistently conserved residues: Trp-H36, Cys-H22, and Cys-H92 in the V_H domain and Trp-L35, Cys-L23, and Cys-L88 in the V_L domain. In both variable domains, the lower halves of the hydrophobic cores distal to the CDRs are centered on a structure-wise equivalent core composed of a few residues with strong amino acid type preferences: Phe-L62, Leu-L73, Ile-I75, and Tyr-L86 in the V_L domain and Phe-H67, Leu-H80, Tyr-H90, Val-H109, and Val-H111 in the V_H domain (Figures 1A and 1B; Table S5; Figure S4). These conserved residues form the structural core in each of the variable domains and are also highly relevant in stabilizing the protein structures, as demonstrated in thermal stability measurements below. These core residues are only a fraction of the lower hydrophobic core, and not all of the lower hydrophobic core residues are equally conserved. In each variable domain, about half of the lower hydrophobic residues, which surround the aforementioned conserved core, are tolerant in substitutions among hydrophobic amino acid types (ALIVMF; Figures 1A and 1B). As such, it can be envisaged that the lower hydrophobic core is composed of two layers: a central core with conserved amino acid types and a peripheral layer with more flexible amino acid type substitutions. This view is consistent with the Φ-value analysis of Ig-like structures (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). By contrast, the upper core residues are not limited to hydrophobic interactions, with one pair of moderately conserved hydrophobic residues (Met-L4 and Leu-H4). The upper core residues are conventionally defined as part of the CDRs (North et al., 2011), and the amino acid type substitution patterns are closely related to the CDR conformation. Thus, the amino acid type substitution patterns in the upper core could be affected both by the binding of the antigen and by the folding stability. Only a fraction of upper core residues are included in the β strand regions shown in Figure 1.

The interface residues between the two variable domains in the template scFv are highly selective in amino acid types. Unlike the lower hydrophobic core, the interface is constituted by more aromatic residues: Tyr-L36, Tyr-L49, Tyr-L87, and Phe-L98 in the V_L domain and Trp-H47, Tyr-H91, and Trp-H103 in V_H domain (Table S5; Figure S4). The side chains of the conserved aromatic residues form a loosely defined middle aromatic cluster, surrounded by nonaromatic interface residues (Ala-L34, Pro-L44, Leu-L46, Gln-L89, His-H35, Val-H37, Leu-H45, Gly-H50, Ala-H93) (Table S5; Figure S4). The interface residues are, on average, the most stringent in amino acid type substitution; the interface aromatic core residue positions prefer to be substituted among aromatic residues, whereas the nonaromatic interface residues tend to be substituted by other nonaromatic residues (Figures 1A and 1B). The results highlight the importance of interactions involving aromatic side chains in the interface between the variable domains in the template scFv.

Residue side chains involving hydrogen-bonding networks are also conserved with strong amino acid type preferences. These residues include Glu-H6, Asn-H76, Arg-H94, Asp-H101, Gln-H105, and Thr-H107 in the V_H domain and Gln-L6, Gln-L100, and Thr-L102 in the V_L domain (Table S5; Figure S4).

Comparisons of the NGS Sequence Profiles, Sanger Sequencing Profiles, and Consensus Sequence Profiles Derived from Natural Sequence Database

The NGS data shown in Figure 1 provide a view of amino acid substitution patterns with rich detail from a large volume of sequence data, but the downside is the difficulty to confirm that each of the sequences from NGS is functional. In contrast,

Protein Data Bank; ▼, Kabat number (grayscale background shows the relative percentage of burial in the 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); ■, five major groups of distinguished structural characteristics (O, surface; U, upper core; L, lower core; H, hydrogen bonding and electrostatic interaction; I, interdomain interface; d, disulfide bond), with the colored background showing the range of the information content 0.25 × 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 available appearance frequency [nearest integer for 10 × $q(j,i)$, $q(j,i)$ defined in Equation 3] of the amino acid type derived from the HTTI-filtered scFv variants (see the corresponding main text).

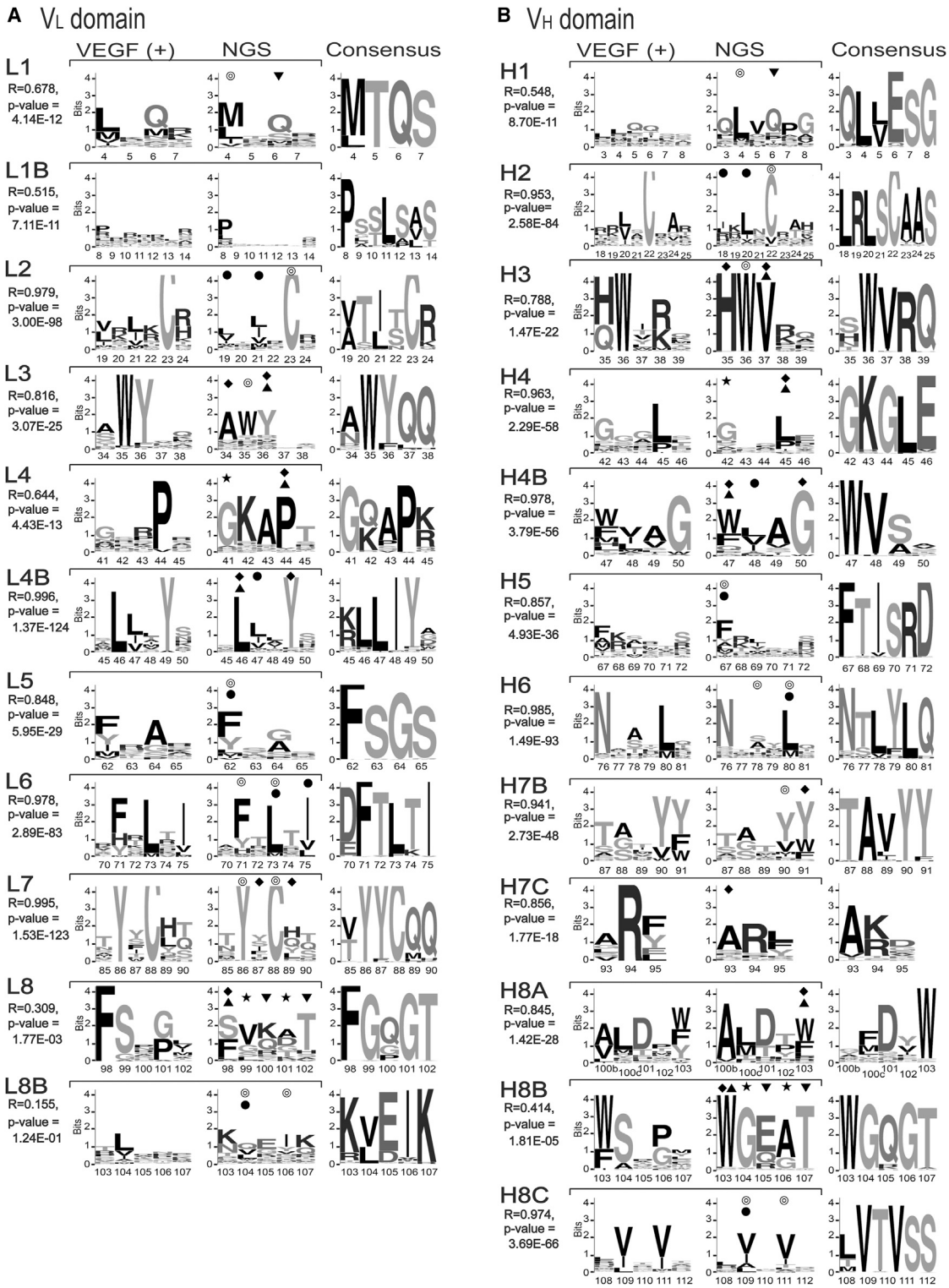

Figure 2. Comparisons of NGS Profiles, Sanger Sequence Profiles, and Consensus Sequence Profiles from the Natural Sequence Database

Sequence LOGOs for each of the V_L (A) and V_H (B) domain libraries are derived with Equations 1, 2, and 3. The y axis shows the information content I_j and the x axis shown the residue positions in Kabat number. The column "VEGF(+)" shows sequence LOGOs for nonredundant VEGF-binding scFv variants confirmed with ELISA; the sequences were obtained with the Sanger method. The column "NGS" shows next-generation sequencing results (data from Figure 1). The column "Consensus" shows the LOGOs from database sequences of human antibody V kappa1 (A) and VH3 (B) family (Supplemental Experimental Procedures). The symbol ★ points to critical glycines; the symbol ◎ points to interior residues; the symbol ◆ points to interface residues; the symbol ▼ points to hydrogen

(legend continued on next page)

Sanger sequencing of functional variants provides sequence information with precise confidence from the confirmation of the function of each sequence, but the volume of Sanger sequencing data is orders of magnitude smaller than the sequence volume from NGS, and the downside is that the sampling space could be limited. Hence, the sequence preference results need to be validated with consensus from the NGS data set and the Sanger sequence data set.

Figures 2A and 2B compare the deep sequencing results shown in Figure 1 with Sanger sequencing results of randomly selected functional scFv variants confirmed for VEGF binding with ELISA (see below and Supplemental Experimental Procedures; Tables S3 and S4). The information contents attributed to each amino acid type are shown in the LOGO form (Supplemental Experimental Procedures), where the relative entropy I_j (the y axis in Figure 2) at position j (the x axis in Figure 2) is the sum of individual d_{ji} attributed to the amino acid type i shown as the size of the amino acid type character in the LOGO plot. The quantitative comparison between each pair of the LOGO plots derived from NGS and Sanger sequencing is shown as the Pearson's correlation coefficient (R) and the p value of the t test next to each of pair of plots in Figure 2. The quantitative comparisons indicate that the two sets of data share similar patterns in general and thus lead to generally agreeable implications.

On the other hand, the consensus sequence profiles derived from natural antibody sequences (Supplemental Experimental Procedures) are different from the corresponding profiles from NGS and Sanger sequencing (Figures 2A and 2B) in that only a small fraction of the positions in the variable domain sequences need to be conserved as the consensus sequence to support the structural stability and functionality. The comparisons indicate that the natural antibody variable domain sequence variations remain largely confined to the germline sequence, while the tolerance of the sequence variations can only be revealed with the functional variants in the large recombinant libraries without phylogenetic biases.

The highly conserved residues, which are thus strongly coupled with the antigen binding, appear mostly in structure-wise equivalent positions in the heavy- and light-chain variable domain. These residue pairs are marked by the symbols in Figures 2A and 2B for the V_L domain and the V_H domain, respectively. Table S5 summarizes the equivalent pairs of these highly conserved residues. Figure S4 shows the superimposition of the two variable domains and the highly conserved residue pairs that are structure-wise equivalent (Table S5). Most of the highly conserved residues pairs emerging from the results are distributed in the hydrophobic core (Figures 2A and 2B; Figure S4D) and in the interface (Figures 2A and 2B; Figure S4E). The majority (92%) of the hydrophobic core and interface residue pairs as shown in Table S5 are identical to the consensus sequence (Figures 2A and 2B), suggesting that these residue pairs could play key roles in protein structure and function throughout the antibody variable domain families. The consistency of the key residue positions in both variable domains suggests that the structural information is encoded in the variable domain sequences in a comparable manner, although the sequence similarity between the two variable domains in the template scFv (Av1) is only 26%. Other structure-wise equivalent residue pairs with strong sequence preferences are structurally critical glycines (Figures 2A and 2B; Figure S4C; Table S5) and residue side chains involving hydrogen bonding (Figures 2A and 2B; Figure S4F; Table S5). However, only 50% of the latter two groups of residue pairs follow the consensus sequence requirements (Table S5; also see Figures 2A and 2B), suggesting that alternative structural codes in these regions could be equally feasible for the variable domains' folding and stability.

High-Throughput Measurements for scFv Thermal Stability

The statistical analysis above has highlighted the key residues coupled to the function of the scFvs, but it is not clear whether the coupling is mediated through energetic stabilization of the functional conformation of the scFvs. One interpretation of deep sequencing data based on the statistical analysis shown in Figure 1 is built on the assumption that the energetic contribution to the overall protein stability from an individual residue is somewhat proportional to the relative entropy of the residue shown in Equation 2, provided that the distributions of the amino acid types in the protein ideally follow the Boltzmann distribution reflecting the free energy of the protein system. In addition, the assumption is applicable to a residue position when it is energetically uncoupled from all the other residues in the protein as a first approximation. The applicability of the assumptions above cannot be inferred by the deep sequencing data shown in Figure 1 alone.

An alternative hypothesis that the selected functional scFvs at room temperature are mainly populated with sequences accommodating folding requirements, but not necessarily optimized for thermal stability, could lead to an alternative interpretation of the statistics shown in Figure 1. This hypothesis can be tested by examining if the sequence features are still conserved in the more thermally stable scFvs.

HTTI measurements (Miller et al., 2010; Supplemental Experimental Procedures) are suitable to validate the statistical interpretations of the deep sequencing data. Figure 3 shows the correlation between the T_{50} from the HTTI measurement for VEGF binding and the melting temperature T_m derived from differential scanning calorimetry (DSC) (Supplemental Experimental Procedures). The threshold temperature T_{50} reflects the irreversible loss of 50% VEGF-binding capability of the template scFv Av1 and marks the initial stage of the scFv denaturation preceding the global structural disruption at the melting temperature T_m, where the overall exposure of the main hydrophobic core of the protein takes place. Figure 3 also compares the

bonding residues. The structures of these residues (also listed in Table S5) are shown in Figure S4. The symbol ● points to the lower core hydrophobic residues in Figure 5. The symbol ▲ points to the interface residues in Figure 6. Experimental procedures are described in detail in Supplemental Experimental Procedures. The sequence of the template scFv (Av1) and designs of phage-displayed libraries are shown in Figure S1. Phage-display library constructions and selection results are shown in Tables S1 and S2 and Figure S2. The number of sequences from Sanger sequencing and NGS is shown in Tables S3 and S4; scFv variant sequences and corresponding T_{50}'s are shown in Table S6.

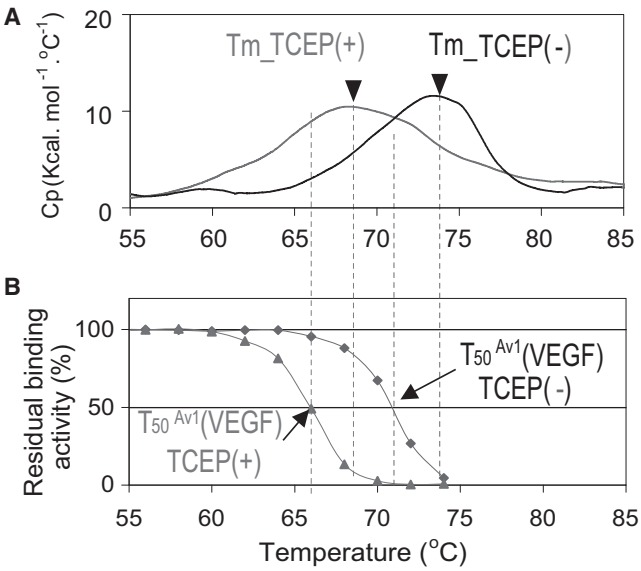

Figure 3. Comparison of HTTI Measurements with DSC Measurements for the Av1 scFv Template

(A) Shown are the thermal denaturing curves of purified template scFv Av1 measured by DSC in the presence of TCEP (gray curve) and in the absence of TCEP (black curve).

(B) Shown are the thermal denaturing curves of secreted soluble template scFv Av1 in culture medium measured by thermal inactivation of the VEGF binding in the presence of TCEP (triangles) and in the absence of TCEP (diamonds). The scFv expression/purification and the thermal stability measurement procedures are described in Supplemental Experimental Procedures.

T_{50}'s of the template scFv Av1 for VEGF binding in the presence or absence of moderate concentration of TCEP [20 mM tris (2-carboxyethyl) phosphine], which was used to reduce the scFv intradomain disulfide bond upon thermal denaturation so as to prevent refolding of the heat-treated scFv. As expected, the irreversible heat denaturation condition resulted in lower T_{50} in the presence of TCEP. This irreversible heat denaturation condition was used throughout this work.

Sequence Features Contributing to the Thermal Stability of the Variable Domains

The HTTI experiment, which is compatible with the high-throughput ELISA platform (Supplemental Experimental Procedures), was applied to more than 1,000 scFv variants selected for VEGF binding from the phage-displayed libraries, and the results are shown in Table S6. The sequences of the scFv variants with T_{50} measurements were determined by Sanger sequencing, and the results are also listed in Table S6. Because the template scFv Av1 binds to VEGF with the CDRs and binds to Protein A through the V_H framework region (Figure 4A) (Fuh et al., 2006; Graille et al., 2000), the HTTI experiments were carried out by measuring the scFv-VEGF binding and the scFv-Protein A binding. Figures 4B and 4C plot the distributions (y axis) of the VEGF-based T_{50}'s [T_{50}(VEGF)] and the Protein-A-based T_{50}'s [T_{50}(PrA)] for the selected scFv variants from each of the scFv libraries (x axis). Figure 4D shows the high correlation between the two sets of T_{50} measurements (Figures 4B and 4C), indicating that the scFv binding capability to VEGF and Protein A is cooperatively abolished during the thermal denaturation of the scFv variants. Because a portion of the T_{50}(PrA) data are incomplete due to the disruption of the scFv-Protein A binding sites in the scFv variants of some libraries (Figure 4C), only the T_{50}(VEGF) data are used for the following discussion.

The comparison of the deep sequencing statistics and the HTTI data confirms that the key residues critical to the scFv stability as highlighted by the deep sequencing statistics are mostly validated with the HTTI measurements. Figures 1A and 1B compare the sequence preferences derived from deep sequencing with the sequence preferences of the scFv variants for which the T_{50}(VEGF)s are greater than T_{50}^{Av1}(VEGF)-5°C, where the T_{50}^{Av1}(VEGF) is the T_{50}(VEGF) of the template scFv Av1 (Figure 3). These scFv variants have thus been confirmed to be similar to or more robust than the template scFv in thermal stability. The frequencies [$q(j,i)$ in Equation 3] of the key residues emerging from the HTTI-filtered scFv variants are presented by the numbers in the cells of the heatmaps in Figures 1A and 1B. The two sets of amino acid frequencies derived from the NGS statistics and the HTTI-filtered scFv variants are significantly correlated, as shown in Figures 4E and 4F. The consistency indicates that the energetic assumption for the NGS statistics is qualitatively appropriate to the first approximation. That is, the key residues observed in NGS data are also required in optimizing the stability of the proteins. The discrepancies appear mostly in the C-terminal β strands (H104–H107 in the heavy chain and L98–L107 in the light chain), indicating that the C-terminal β strands could accommodate flexible sequence features while remaining optimized in thermal stability.

The HTTI data indicate that the local sequences in the C-terminal β strands of both variable domains, especially the residue positions near the lower hydrophobic core, can be further optimized to stabilize the scFv variants with thermal stability above that of the consensus-sequence-optimized template scFv. As shown in Figures 4B and 4C, only the scFv variants with alternative residues in the L7, L8B, H8A, and H8C regions can have superior thermal stability to that of the template scFv. L7 and H8A are near the CDR3s, which are involved in antigen binding (Yu et al., 2012). Residues in the L8B and H8C regions, on the other hand, are distal to the VEGF-binding site and can enhance the thermal stability of the variable domains only through interacting with the lower hydrophobic core. This result pointed to the possibility that the lower hydrophobic core could be further optimized to enhance the thermal stability beyond that of the consensus-sequence-optimized template.

Sequence Preferences of the Lower Hydrophobic Cores in Antibody Variable Domains

The NGS and HTTI results above indicate that the residue side chains involving the lower hydrophobic packing peripheral to the central core of the variable domains are interchangeable in hydrophobic amino acid types (Figure 1) and, moreover, the optimization of the peripheral hydrophobic packing can improve the thermal stability of the variable domains (Figure 4). These conclusions were further validated by simultaneously diversifying the amino acid side-chain composition and characterizing the side-chain requirements in the lower hydrophobic core. One phage-displayed library with scFv variants enumerating all combinatorial possibilities of the nine lower core hydrophobic

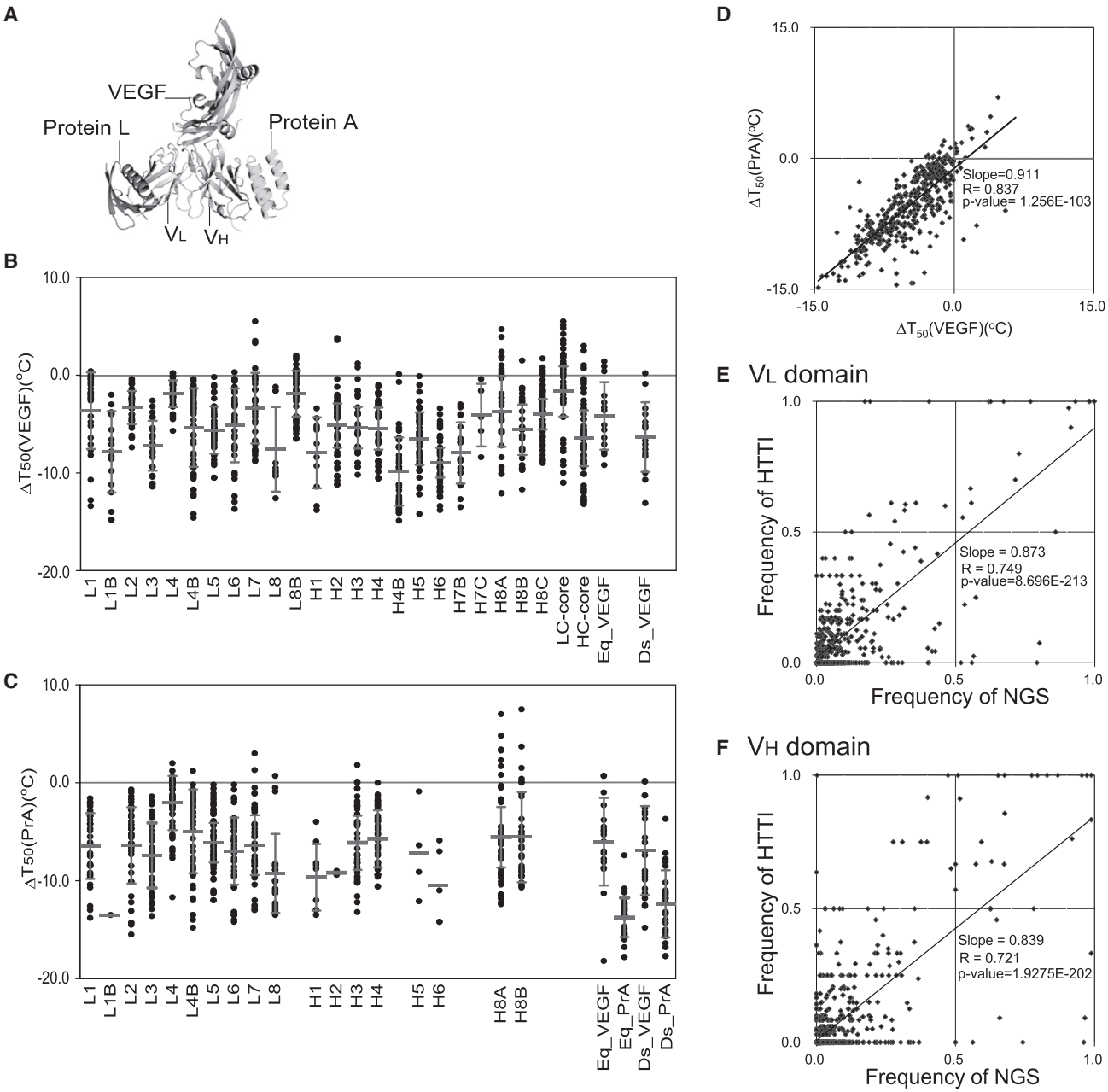

Figure 4. Thermal Stability Measurements of scFv Variants

(A) 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.

(B) The T_{50} for VEGF binding [T_{50}(VEGF)] is the temperature at which the scFv loses 50% of binding capacity to VEGF in the presence of TCEP (A). The ΔT_{50}(VEGF), as shown in the y axis, is the relative denaturation temperature: $\Delta T_{50}(VEGF) = T_{50}(VEGF) - T_{50}^{Av1}(VEGF)$, where $T_{50}^{Av1}(VEGF) = 65.5^{\circ}C$ in the presence of TCEP (Figure 3). The data points were derived from HTTI measurements with scFv variants selected for VEGF binding (except for variant sets labeled as Eq_PrA and Ds_PrA, which were selected for Protein A binding) from each of the libraries (x axis). The mean and SD for each of the distributions are superimposed on the data points (black dots).

(C) The T_{50} for Protein A binding [T_{50}(PrA)] is the temperature at which the scFv loses 50% of binding capacity to Protein A in the presence of TCEP. The ΔT_{50}(PrA), as shown in the y axis, is the relative denaturation temperature: $\Delta T_{50}(PrA) = T_{50}(PrA) - T_{50}^{Av1}(PrA)$, where $T_{50}^{Av1}(PrA) = 69.5^{\circ}C$ in the presence of TCEP. The data points were derived with the same sets of scFv variants used in (B). The mean and SD for each of the distributions are superimposed on the data points (black dots).

(D) For each of the scFv variants with both measurements of ΔT_{50}(VEGF) and ΔT_{50}(PrA) in (B) and (C), respectively, the data point in the plot shows the corresponding pair of ΔT_{50}(PrA) (y axis) and ΔT_{50}(VEGF) (x axis). The T_{50} measurement results are shown in Table S6 in detail.

(E) The plot shows correlation of the two sets of frequencies [*q(j,i)*; see Equation 3] of amino acid types derived from NGS data (Figure 1B) and from HTTI-filtered variants (Table S6) for the V_L domain. R is the Pearson's correlation coefficient, and the p value was calculated with a t test.

(F) The same as in (E) for the V_H domain.

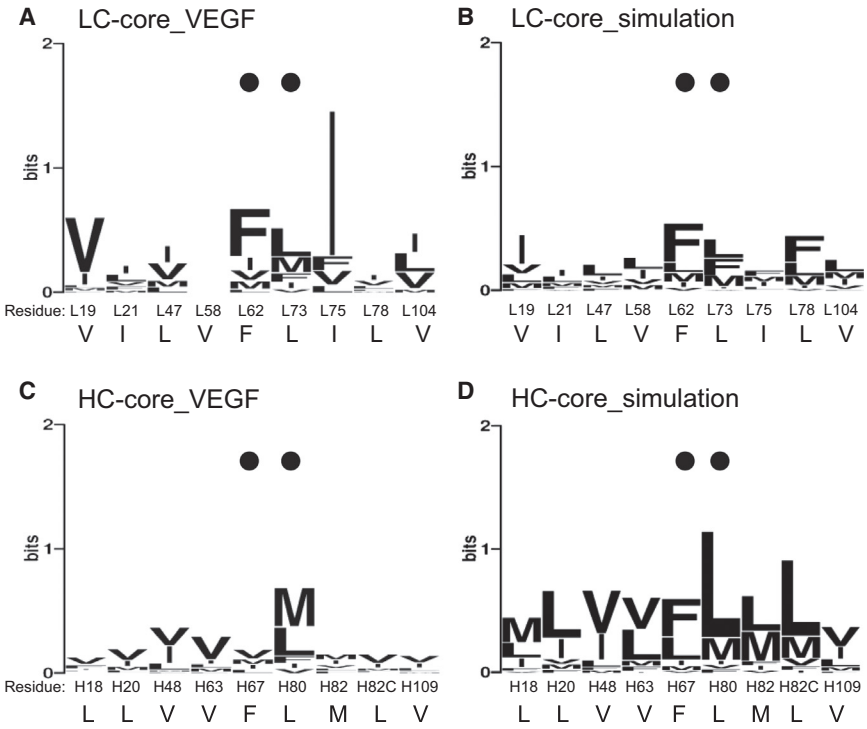

Figure 5. Sequence Features for the Lower Hydrophobic Cores in the V_H and the V_L Domain of Functional scFv Variants Selected from the HC-Core and LC-Core Libraries with Simultaneous Mutagenesis of Nine Residues with Hydrophobic Amino Acid LIVFM in Each Core

(A) Sequence preferences of the functional scFv variants selected for VEGF binding from the LC-core library are shown in the LOGOs; the y axis shows I_j calculated with Equations 1, 2, and 3 based on the background probabilities derived from the phage-displayed library design (see text). The x axis shows the wild-type amino acid in the template scFv. The distribution of the ΔT_{50}(VEGF)s of these scFv variants are shown in Figure 4B (data for LC core). The symbol ● points to the inner core hydrophobic residues. The sequences and corresponding T_{50}(VEGF)s of the selected scFv variants are shown in Table S6.

(B) The LOGOs show the sequence preferences from the Monte Carlo simulations of the nine lower core hydrophobic residues in the V_L domain.

(C) Same as in (A) for the functional scFv variants selected from the HC-core library.

(D) Same as in (B) for the nine lower core hydrophobic residues in the V_H domain.

residues (Figure S4G) was constructed for each of the variable domains of the template scFv (the LC-core and HC-core libraries; Figure S1). The two nine-residue sets are structure-wise equivalent in the two variable domains (Figure S4G). Each of the residue positions was diversified among the five hydrophobic amino acid types (LIVFM) with the degenerate codons [D(G 33%; A 33%; T 33%)-T-K(T 50%; G 50%); L:I:V:F:M = 1:1:2:1:1] in the phage-displayed libraries. ScFvs were selected for VEGF binding, followed by the HTTI measurements for the selected scFvs. The T_{50}(VEGF)s and the corresponding scFv sequences are shown in Table S6. The distributions of the T_{50}(VEGF)s are plotted in Figure 4B. The HTTI measurements show that some of the scFv variants from the two libraries have superior thermal stability to that of the template scFv (data for LC-core and HC-core in Figure 4B). The sequence preferences for the nine residue positions are shown in Figures 5A and 5C for V_L and V_H, respectively. The strikingly low information content ($I_j < 1$) shown in Figures 5A and 5C indicates that these residue positions are highly interchangeable among the hydrophobic amino acid types (see also Table S6 for sequence details). Even the core residues (Phe-L62, Leu-L73, Phe-H67, and Leu-H80, indicated with arrows in Figure 2) originally thought to be highly conserved (Table S5) are largely interchangeable among the hydrophobic amino acid types when the surrounding packing side chains are allowed to covariegate.

The covariation correlations of all residue pairs in the functional LC-core and HC-core sequences were evaluated with the correlation coefficient (Φ-value) calculations (Wang et al., 2009). None of the residue pairs in the LC core and HC core are correlated in covariation based on the statistical significance criteria of p value < 0.001, Φ-value > 0.5, and at least 10% of appearance in the sequence pool (Wang et al., 2009). This result suggests that packing of the LC and HC cores is not governed by preferable pairwise interactions between amino acid types. The aromatic side chain (Phe) contributes marginally to the packing configurations; the hydrophobic cores are mostly composed of aliphatic side chains.

Not only are the amino acid types flexible for the residues in the lower hydrophobic packing region, but the packing configurations could also be diverse. Computer simulations of the lower core hydrophobic side-chain packing with fixed backbone structure were carried out with the Metropolis Monte Carlo method based on the relative thermodynamic stabilities estimated with FoldX (Van Durme et al., 2011) (Supplemental Experimental Procedures). The simulated packing configurations with FoldX packing energy lower than that of the template scFv were used to generate the simulated sequence preferences (Figures 5B and 5D) and side-chain conformation (χ_1 angle) preferences (Figure S5). These packing configurations are expected to satisfy the native-like protein packing density requirements without van der Waals volume clashes under the constant main-chain structure. The computer-simulated sequence preferences shown in Figures 5B and 5D do not exactly reflect the experimental results shown in Figures 5A and 5C, but the low information content for the simulated sequence preferences is in agreement with the expectation from the experimental results, indicating that, given a fixed backbone structure, the hydrophobic amino acid types are highly interchangeable within the given overall packing volume, while satisfying the high packing density requirements. Moreover, the χ_1 angle distributions of the simulated packing residue side chains shown in Figure S5 are not centered on the corresponding χ_1 angles in the scFv template, indicating that the combinations of the side-chain conformations do not need to be unique to satisfy the packing density requirement. The

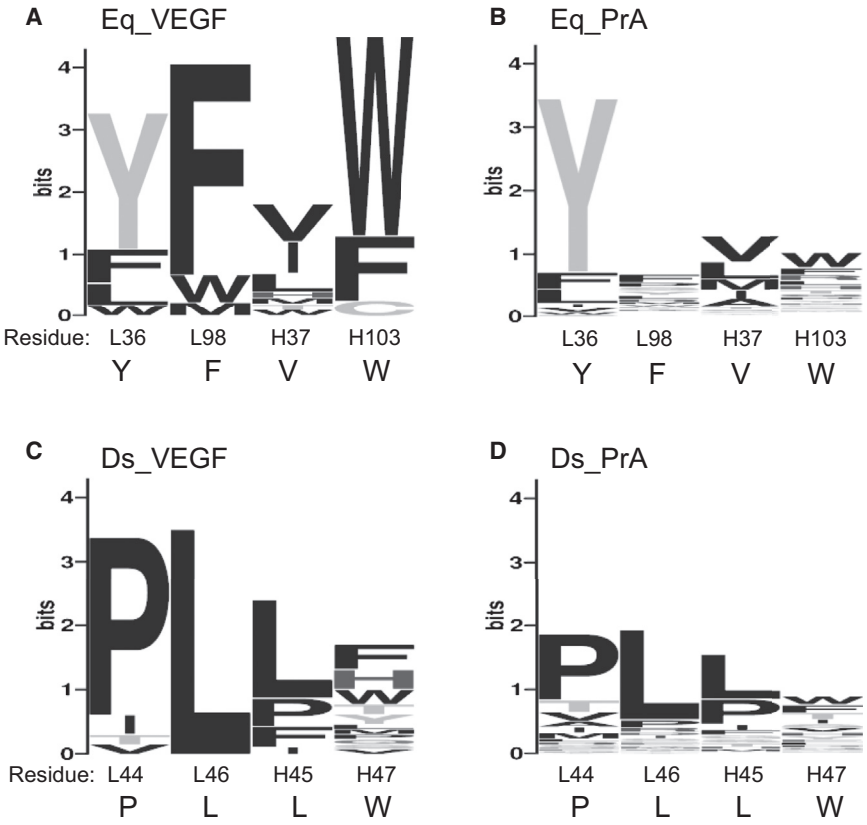

Figure 6. Sequence Preferences of the Residues in the Interdomain Interface

(A) The sequence preferences of the central interdomain interface residues are shown by the LOGOs, which are derived with nonredundant sequences of functional scFv variants selected for VEGF binding from the Eq phage-displayed library with the four residues randomized with saturated mutagenesis. The x axis shows the wild-type amino acid in the template scFv. The distributions of ΔT_{50}(VEGF) and ΔT_{50}(PrA) of this set of scFv variants are also shown in Figures 4B and 4C (marked by Eq_VEGF in the x axis), respectively.

(B) The sequence preferences of the central interdomain interface residues are shown by the LOGOs, which are derived with nonredundant sequences of functional scFv variants selected for Protein A binding from the Eq phage-displayed library. The distribution of ΔT_{50}(PrA) of this set of scFv variants is shown in Figure 4C (marked by Eq_PrA in the x axis). ΔT_{50}(VEGF)s for this set of scFv variants were not measured because the VEGF binding of most of the scFv variants in the Eq_PrA set was not detectable with ELISA.

(C) The sequence preferences of the peripheral interdomain interface residues are shown by the LOGOs, which are derived with nonredundant sequences of functional scFv variants selected for VEGF binding from the Ds phage-displayed library with the four residues randomized with saturated mutagenesis. The distributions of ΔT_{50}(VEGF) and ΔT_{50}(PrA) of this set of scFv variants are shown in Figures 4B and 4C (marked by Ds_VEGF in the x axis), respectively.

(D) The sequence preferences of the peripheral interdomain interface residues are shown by the LOGOs, which are derived with nonredundant sequences of functional scFv variants selected for Protein A binding from the Ds phage-displayed library. The distribution of ΔT_{50}(PrA) of this set of scFv variants is shown in Figure 4C (marked by Ds_PrA in the x axis). ΔT_{50}(VEGF)s for this set of scFv variants were not measured because the VEGF binding of most of the scFv variants in the Ds_PrA set was not detectable with ELISA. The sequences and the corresponding ΔT_{50}(PrA) and ΔT_{50}(VEGF) are also listed in Table S6.

simulations did not explain the extremely low information content at the position of Val-L58 and the highly conserved Ile-L75 observed experimentally. The residue position of Val-L58 is situated in the flexible loop region between CDR-L2 and the outer loop, and thus the low information content could be due to the dynamic property of the loop region, which was not taken into consideration in the Monte Carlo simulations.

The results shown in Figures 5 and S5 suggest that the lower hydrophobic packing cores of antibody variable domains are compatible with a large number of combinations of hydrophobic aliphatic amino acid side chains, many of which are superior in thermal stability to the wild-type hydrophobic packing configurations. All of these hydrophobic packing configurations are compatible in supporting the CDR conformations for effective antigen binding. The VEGF-binding affinities of a set of representative scFv variants were measured with VEGF binding in the presence of the template scFv Av1 as a competitor for VEGF binding (competitive phage ELISA; Supplemental Experimental Procedures), and the results indicate that the VEGF-binding affinities of these representative scFv variants are all similar to the template scFv (Figure S3). There is no clear correlation between the thermal stability and the VEGF-binding affinity. This indicates that the amino acid sequence preferences of the lower hydrophobic core are mostly energetic and topological determinants supporting protein folding but are not coupled directly to the functional conformation of the CDRs.

Sequence Preferences of the Residues in the Interface between the Antibody Variable Domains

In contrast to the high interchangeabilities of the residues in the central packing region of the variable domains, the residues in the interface between the two variable domains in the scFv are much more conserved in amino acid preferences and are coupled to the antigen recognition site formed by CDRs. A phage-displayed scFv library (the Eq library; Figure S1; Table S1) with four residues in the center of the interdomain interface (Tyr-L36, Phe-L98, Val-H37, and Trp-H103; see Figure S4E for structures) randomized with saturated mutagenesis was constructed and selected for VEGF binding and for Protein A binding. Figure 6A summarizes the sequence preferences for nonredundant scFv variants selected for VEGF binding. The scFvs selected for VEGF binding bind to both VEGF and Protein A (Table S6) as expected, because the prerequisite of the functional CDR structure is the fully folded V_H domain structure. The sequence preferences for the nonredundant Protein-A-binding scFv variants are summarized in Figure 6B. By contrast, most of the scFvs selected for Protein A binding do not bind to VEGF detectable by ELISA (Table S6), suggesting that the

interface residues supporting the V_H domain structure do not necessarily support the functional structure of the CDRs. The highly conserved VEGF-binding scFv variants (Figure 6A) suggest that the central interface residues, especially residues with an aromatic side chain, outside the conventional CDRs are strongly coupled with VEGF binding, whereas the coupling of these residues to the Protein A binding site is substantially weaker.

Similar results supporting the coupling between the edge of the domain interface and the antigen binding site are shown in Figures 6C and 6D. Here, the phage-displayed scFv library (the Ds library; Figure S1; Table S1) with four residues in the peripheral interdomain interface (Pro-L44, Leu-L46, His-H45, and Trp-H47; see Figure S4E for structures) randomized with saturated mutagenesis were selected for VEGF binding and Protein A binding. Again, the sequence preferences of the interface residues selected for VEGF binding (Figure 6C) are much more stringent than those selected for Protein A binding (Figure 6D). The scFv variants selected for VEGF binding bind to both VEGF and Protein A, and most of the scFv variants selected for Protein A binding bind only to Protein A, but not to VEGF, indicating that the CDR coupling involves the edge residues of the interface.

The sequence preference patterns shown in Figures 6A and 6C indicate that the consensus sequence is preferred in the Eq and Ds positions but is not necessarily the only sequence combination for functional scFv variants binding to VEGF. Alternative combinations are tolerated in both the Eq positions and Ds position in the functional scFv variants binding to VEGF. The alternative sequence combinations were compared with those observed in the structure-wise equivalent positions in 587 antibody structures in the Protein Data Bank with the structural alignment algorithm implemented in PrISM (Yang and Honig, 2000a, 2000b, 2000c). A total of 22% of the Eq_VEGF sequences and 9% of the Ds_VEGF sequences have never been observed in known antibody structures. The Eq and Ds sequence tolerance for the Protein-A-binding variants is even higher: 86% of Eq_PrA sequences and 54% Ds_PrA sequences have never been observed in known antibody structures.

The alternative variants, although functional in binding to Protein A, are largely inferior in thermal stability to that of the Av1 scFv template encoded with the consensus sequence (see below). The T_{50} measurements and the corresponding sequences determined by Sanger sequencing for the scFv interface variants are shown in Table S6, and the distributions of the T_{50}'s are plotted in Figures 4B and 4C (data indicated by Eq_VEGF, Eq_PrA, Ds_VEGF, and Ds_PrA in the x axis of the figures). Few of these T_{50}'s are significantly higher than that of the template scFv, indicating that there is not much room for thermal stability improvement beyond the consensus sequence in the interface. In addition, although the Eq_PrA and Ds_PrA sequences are much more diverse, the thermal stabilities of these variants are also substantially compromised in comparison with the template scFv. These results indicate that the interface residues are coupled strongly to the CDR both structurally and energetically.

It can be envisaged that the interface residues, especially the central residues with aromatic side chains, determine the relative orientation between the two variable domains in the antibody, which in turn affects the VEGF binding affinity. These residues are also coupled with the Protein A binding to a relatively lesser extent, allowing greater compromise of the protein stability while retaining the interaction to Protein A. This interpretation is derived from a large body of evidence (Chailyan et al., 2011; Narayanan et al., 2009; Teplyakov et al., 2011), indicating that the V_H-V_L interface packing configurations are frequently critical determinants for the shape and disposition of the antigen-binding sites (Chailyan et al., 2011), and consequently it is reasonable to anticipate that the interface residue replacements have more direct impact on the antigen (VEGF) binding than on the Protein A binding, for which the binding site does not overlap with the CDRs.

Conclusions

Large-scale sequence information for the functional scFv variants reveals the set of key residues with equivalent structural positions in both variable domains in the scFv structure. These key positions are situated in the core and the interface of the variable domain structures, but the locations of these key residue positions are not obviously evident from the natural consensus sequence analysis. The stringent amino acid type preferences in these positions are by and large consistent with the consensus sequence of the antibody sequence family and thus are thought to encode essential structural and functional information for the variable domains.

The sequence statistics are nevertheless limited in inferring quantitative scFv stability; thermal stability measurements for a large number of scFv variants are essential for understanding the sequence data in terms of thermodynamics. The NGS data from a large number of scFv variants selected for VEGF binding have revealed that the residues in the lower hydrophobic core regions of the variable domains are highly interchangeable among hydrophobic aliphatic amino acid types. Moreover, various combinations of the hydrophobic side chains in the lower hydrophobic core can further optimize the thermal stability of the scFv variants above that of the consensus-sequence-optimized template scFv, as validated by high-throughput thermal stability measurements shown in Figure 4. Together, these results suggest that the hydrophobic side-chain packing specificity of the lower cores in the variable domains is not a key determinant for the variable domain's overall structure. Evidently, multiple hydrophobic core packing configurations, in terms of side-chain types and conformations, can sustain the CDRs with similar binding affinity and specificity to the antigen.

By contrast, the interface residues, especially residues with aromatic side chains, between the variable domains are highly conserved and are strongly coupled with the antigen-binding site formed by the CDRs, and there is little protein stability margin to be gained by optimizing the consensus interface residues. These results imply not only that residues in the interdomain interface contribute to the thermal stability of the interface, but also that these residues could determine the relative configurations of the CDR residues, resulting in diversity of the antigen-recognition capabilities of the CDRs.

Together, the results suggest that the framework hydrophobic core residues are coupled indirectly with antigen binding through energetic stabilization of the protein structures and can

frequently be further optimized for antibody stability engineering, but the interdomain interface residues are directly coupled with the function of the CDRs and should be treated as an integral part of the antigen-binding site.

EXPERIMENTAL PROCEDURES

Key methods are very briefly described here. General experimental details are described in Supplemental Experimental Procedures.

Phage-Displayed scFv Library Construction and Selection

The scFv library construction and the selection procedures followed methods described previously (Yu et al., 2012).

High-Throughput Thermal Inactivation Measurement

The procedure followed a previously published work (Miller et al., 2010), with minor modifications.

Information Content from Sequence Profiles

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

$$d_{ji} = q(j,i)I_j \quad \text{(Equation 1)}$$

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

where

$$q(j,i) = \frac{C_{ji}}{M_j}. \quad \text{(Equation 3)}$$

p_i is the background probability for amino acid type I, C_{ji} is the count of amino acid type i at the position j of the multiple alignment, and M_j is the total count of 20 natural amino acid residues at the 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; the statistics are shown in Table S2.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, five figures, and six tables and can be found with this article online at http://dx.doi.org/10.1016/j.str.2013.10.006.

AUTHOR CONTRIBUTIONS

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

ACKNOWLEDGMENTS

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 thank the Sequencing Core Facility, Scientific Instrument Center at Academia Sinica for DNA sequencing and the Biophysics Core Facility, Scientific Instrument Center at Academia Sinica for the DSC experiment.

Received: June 7, 2013
Revised: September 3, 2013
Accepted: October 1, 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.

Bradbury, A.R., Sidhu, S., Dübel, S., and McCafferty, J. (2011). Beyond natural antibodies: the power of in vitro display technologies. Nat. Biotechnol. *29*, 245–254.

Caravella, J., and Lugovskoy, A. (2010). Design of next-generation protein therapeutics. Curr. Opin. Chem. Biol. *14*, 520–528.

Chailyan, A., Marcatili, P., and Tramontano, A. (2011). The association of heavy and light chain variable domains in antibodies: implications for antigen specificity. FEBS J. *278*, 2858–2866.

Clarke, J., Cota, E., Fowler, S.B., and Hamill, S.J. (1999). Folding studies of immunoglobulin-like beta-sandwich proteins suggest that they share a common folding pathway. Structure *7*, 1145–1153.

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.

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. 47–68.

Hoogenboom, H.R. (2005). Selecting and screening recombinant antibody libraries. Nat. Biotechnol. *23*, 1105–1116.

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.

Jäckel, C., Bloom, J.D., Kast, P., Arnold, F.H., and Hilvert, D. (2010). Consensus protein design without phylogenetic bias. J. Mol. Biol. *399*, 541–546.

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.

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.

Narayanan, A., Sellers, B.D., and Jacobson, M.P. (2009). Energy-based analysis and prediction of the orientation between light- and heavy-chain antibody variable domains. J. Mol. Biol. *388*, 941–953.

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

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.

Steipe, B. (2004). Consensus-based engineering of protein stability: from intrabodies to thermostable enzymes. Methods Enzymol. *388*, 176–186.

Teplyakov, A., Obmolova, G., Malia, T., and Gilliland, G. (2011). Antigen recognition by antibody C836 through adjustment of V(L)/V(H) packing. Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun. *67*, 1165–1167.

Van Durme, J., Delgado, J., Stricher, F., Serrano, L., Schymkowitz, J., and Rousseau, F. (2011). A graphical interface for the FoldX forcefield. Bioinformatics *27*, 1711–1712.

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.

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., and Honig, B. (2000a). 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. *301*, 665–678.

Yang, A.S., and Honig, B. (2000b). An integrated approach to the analysis and modeling of protein sequences and structures. II. On the relationship between sequence and structural similarity for proteins that are not obviously related in sequence. J. Mol. Biol. *301*, 679–689.

Yang, A.S., and Honig, B. (2000c). An integrated approach to the analysis and modeling of protein sequences and structures. III. A comparative study of sequence conservation in protein structural families using multiple structural alignments. J. Mol. Biol. *301*, 691–711.

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.

