Published in final edited form as:
Nat Methods. 2010 September ; 7(9): 741–746. doi:10.1038/nmeth.1492.

High Resolution Mapping of Protein Sequence–Function Relationships

Douglas M. Fowler[1], **Carlos L. Araya**[1], **Sarel J. Fleishman**[2], **Elizabeth H. Kellogg**[2], **Jason J. Stephany**[1,4], **David Baker**[2,4], and **Stanley Fields**[1,3,4]

[1]Department of Genome Sciences University of Washington Box 355065 Seattle, WA 98195

[2]Department of Biochemistry University of Washington Box 355065 Seattle, WA 98195

[3]Department of Medicine University of Washington Box 355065 Seattle, WA 98195

[4]Howard Hughes Medical Institute

Abstract

We present a large-scale approach to investigate the functional consequences of sequence variation in a protein. The approach entails the display of hundreds of thousands of protein variants, moderate selection for activity, and high throughput DNA sequencing to quantify the performance of each variant. Using this strategy, we tracked the performance of >600,000 variants of a human WW domain after three and six rounds of selection by phage display for binding to its peptide ligand. Binding properties of these variants defined a high-resolution map of mutational preference across the WW domain; each position possessed unique features that could not be captured by a few representative mutations. Our approach could be applied to many *in vitro* or *in vivo* protein assays, providing a general means for understanding how protein function relates to sequence.

The sequence of a protein determines its structure and function. Despite this clear correlation, understanding the relationship between protein sequence and function is a complex and largely unsolved problem. The contours of this problem have only expanded as

Correspondence to Stanley Fields, fields@u.washington.edu; ph 206 616-4522;, fax 206 543-0754.

Accession codes.
NCBI Short Read Archive SRA020603

Note: Supplementary information is available on the Nature Methods website.

Author Contributions

D.M.F. conceived of the method, carried out the experiments, analyzed the data and wrote the paper; C.L.A. conceived of the method, analyzed the data and wrote the paper; J.J.S. carried out the experiments; E.H.K., S.J.F. and D.B. carried out the protein folding and binding energy calculations; S.F. conceived of the method and wrote the paper.

EdSumm AOP

The combination of protein display, moderate selection for protein activity and high-throughput DNA sequencing can be applied to hundreds of thousands of protein variants in parallel, enabling the derivation of sequence-function relationships.

EdSumm issue

The combination of protein display, moderate selection for protein activity and high-throughput DNA sequencing can be applied to hundreds of thousands of protein variants in parallel, enabling the derivation of sequence-function relationships.

the amount of protein sequence information, derived from DNA sequencing, has exploded. Thus, methods to rapidly couple protein sequence to protein function are needed.

One effective way to understand protein sequence–function relationships is through the examination of mutants. Mutational analysis has been applied both *in vitro* and *in vivo*, ranging from exploration of protein-protein interaction interfaces to analysis of the kinetics and thermodynamics of protein folding1,2. An example is an alanine scan, in which amino acid residues are individually mutated to alanine3. Residues that, when changed to alanine, result in loss or diminution of function (e.g. binding, catalysis or stability) are likely of functional importance.

Mutational scanning, however, suffers from bottlenecks that have limited its utility. For example, each mutant typically needed to be cloned, expressed and purified for an *in vitro* property to be measured. The requirement to purify individual mutant proteins has been largely resolved by technologies in which a library of protein variants are linked to their encoding DNA sequences. Examples include protein display on the surface of phage, yeast and bacteria as well as ribosome display4. These enable a large (10^6–10^{12}) pool of variants to be assayed for a particular function in parallel. Protein display experiments generally involve multiple rounds of selection (e.g. for binding) that eliminate unselected variants to yield a few highly active proteins. In addition, this scheme has been used to implement mutational scanning, including limited sampling of all possible single mutants in a short sequence5. Despite these advances, protein display has, until recently6–8, been limited by the requirement for Sanger sequencing of variants after selection, restricting to a few thousand the number of variants that can be examined.

Here, we demonstrate that protein display employing moderate selection pressure on a library of variants can be combined with high-throughput sequencing to furnish a high-resolution, fine-scale map of protein sequence–function relationships. Using T7 bacteriophage to display over 600,000 variants of the human Yes Associated Protein 65 (hYAP65) WW domain9, we performed six rounds of selection for binding to its cognate peptide ligand. The selection parameters were tuned to produce only moderate enrichment for better binding, which maintained a large number of library members rather than converging on a few high affinity variants. Short read Illumina sequencing of libraries from the starting pool and after three and six rounds of selection enabled quantitative tracking of the fate of hundreds of thousands of variants simultaneously. The data revealed a detailed sequence–function landscape that is remarkably concordant with known WW domain features. We observed strong agreement between mutational preferences and evolutionary conservation within the WW domain. Furthermore, we comprehensively addressed the question of how mutations impair protein function. The effectiveness of moderate selection combined with high throughput sequencing encourages a shift toward carrying out protein functional screens in a highly parallel fashion, which may reveal novel aspects of protein function in many *in vitro* and *in vivo* contexts.

Results

WW Domain Display and High Throughput Sequencing

WW domains, named for their two conserved tryptophan residues, comprise a large family of three-stranded β-sheet structures approximately 50 amino acids in length9. These domains serve as protein-protein interaction modules, binding to one of several conserved peptide motifs (Fig. 1a). We chose to display a WW domain because the structure is available and because the folding, thermodynamic stability and binding preferences of WW domains have been extensively examined10–18. Additionally, the WW domain has been successfully used in phage display experiments18,19.

We displayed the hYAP65 WW domain on the surface of T7 bacteriophage. T7 is a lytic phage that can display large, intricately folded proteins because displayed proteins need not cross a cell membrane20. The WW domain was fused at the C-terminus of the phage capsid protein and expressed at an estimated 5–15 copies of the WW domain per phage. We conducted preliminary experiments using phage-displayed wild type as well as W17F and W39F hYAP65 WW domain variants, which have diminished capacity to bind peptide14. Each round of selection consisted of binding of mixed populations of phage to an excess of cognate GTPPPPYTVG peptide attached to beads, removal of unbound phage by wash steps, and subsequent amplification of bound phage to generate an input phage library for the next round (Fig. 1b). Based on previous WW domain phage display experiments, we identified conditions in which affinity differences between the wild type WW domain and these variants were readily distinguishable19 (Supplementary Fig. 1a, b).

We generated a WW domain phage library by using chemical DNA synthesis to vary 99 bases, encoding the central 33 residues of the WW domain (Fig. 1a). On average, each variant contained two DNA mutations. The phage library was subjected to six successive rounds of selection against the peptide (Fig. 1b). Sequencing libraries, prepared using PCR, provided data corresponding to a 25-residue portion of the variable region; this length was dictated by the current 76-base limit of short read Illumina sequencing. The remaining eight variable residues of the 33 were not sequenced and therefore contain uncharacterized mutations; however, they are not directly involved in peptide binding (Fig. 1a). The 25 sequenced variable residues span the structured region of the WW domain, including both of the conserved tryptophan residues and encompassing the binding interface (Fig. 1a). Sequencing data were acquired for the input variant library as well as after three and six rounds of selection, with an average of 10.7 million raw reads per library (Supplementary Table 1).

Detection of point mutations with high fidelity in a diverse library using high throughput sequencing poses unique challenges. Under optimal conditions, short read Illumina sequencing has an error rate of slightly less than 1%21, leading to one expected “mutation” resulting from sequencing error in each 76 base sequence. Thus, sequence variants obtained only one or a few times would have a high degree of uncertainty. We overcame this problem by using Illumina paired-end sequencing to capture overlapping sequence for each 76 base read pair, resulting in an average per-base error rate of 3.33e-6 (1/300,000). This low error rate enabled usage of variants sequenced only once (Supplementary Fig. 2). Of the average

10.7 million raw reads per library sequenced, 8.4 million passed further quality filtering (Supplementary Table 1) and were aligned to the wild type WW domain sequence.

The quality filtered sequencing data from the input variant library revealed 1.2 million unique DNA variants corresponding to 602,150 unique protein variants. The observed distribution of mutations in the library was close to the expected distribution22 (Supplementary Fig. 3). At the DNA level, all possible single and nearly all double mutants were observed, as well as a large fraction of triple mutants. Because multiple DNA mutations are required to produce certain amino acid mutations, the library covered 416 of the 500 possible single mutants, 32,779 of the 120,000 possible double mutants and 328,768 of the 18.4 million possible triple mutants at the protein level (Table 1).

Assessing Sequence–Function Relationships In the WW Domain

We compared the population of protein variants present in the input library to the populations found in the round three and round six libraries. We assume that the number of times a given unique protein variant is sequenced corresponds to its abundance within the library. Thus, protein variants that are of better than average affinity should increase in abundance after selection and be sequenced more often, and *vice versa* for variants of worse than average affinity. We directly compared the input library to libraries after peptide selection. We did not adjust our data based on controls in which successive rounds of amplification were applied either with no selection, or with selection only for expression (but not binding) of the WW domain23. These controls could be performed to remove hidden, variant-derived biases and potentially make the method more robust, although they could introduce their own stochastic amplification biases.

Selection eliminated the majority of variants (Fig. 2a); the input library of 602,150 variants greatly exceeded the complexity of the round three (152,656 variants) and round six (94,606 variants) libraries. Performance in this selection experiment and codon usage did not correlate ($R^2 = 0.04$), which suggests that non-coding mutations do not substantially affect WW domain selection (Supplementary Fig. 4). The wild type WW domain, which represented 25.9% of the input library, increased in abundance 1.75-fold after six rounds of selection. Overall, 97.2% of the variants observed in the input library were deleterious relative to wild type.

We examined the position-averaged effect of substitution in the WW domain by calculating for each residue within the variable region an enrichment ratio, defined as the frequency of mutations in the round six library divided by the frequency in the input library. We identified distinct regions within the WW domain that were permissive (i.e. high mutational frequency) to mutation and others that were intolerant (i.e. low mutational frequency) (Fig. 2b). As expected, the two tryptophan residues were highly intolerant to change.

We compared mutational tolerance in our assay to the evolutionary conservation of the WW domain. The hYAP65 WW domain matches a consensus sequence generated from 72 WW domains at 11 of the 25 variable residues. Of these 11 residues, 10 were intolerant of mutation in the selection experiment (Fig. 2c, **blue**). These data are consistent with studies showing that mutationally intolerant regions *in vitro* are generally conserved in evolution24.

Furthermore, changes from the hYAP65 wild type to the consensus were favorable for each of 10 other residues (Fig. 2c, **green**). Thus, 20 of the 25 residues either remained as consensus residues or benefitted from changes to the consensus, suggesting that our selection experiment captured general features of the WW domain evolutionary process.

Projection of the enrichment ratios onto the NMR structure16 of the WW domain resulted in a spatial distribution of mutational tolerance (Fig. 2d). The projection shows that unlike the loop regions, the core ligand-binding region of the WW domain generally does not tolerate mutations. These findings are consistent with previous, lower density mutational scans of the hYAP65 WW domain17.

The high density of the data enabled us to calculate enrichment ratios for nearly every mutation in the variable region. In total, we derived enrichment ratios for 405 of the 500 possible single amino acid mutations (including mutations to stop codons) and organized these into amino acid substitution profiles comparing the round six and input libraries (Fig. 3); profiles comparing round three to input libraries were similar (Supplementary Fig. 5a, b). Additionally, the results for subsets of variants containing single or double mutations were nearly identical to those obtained with the complete set of variants (Supplementary Fig. 5c, d). These data constitute the equivalent of an all-residue scan, in which not just alanine but each amino acid replaces the wild type residues. Thus, these amino acid substitution profiles comprise a detailed map of sequence–function relationships for this domain.

Present within the amino acid substitution profile map are signatures of selection that delineate the effect of each amino acid when it acts as a replacement residue. For example, substitution to proline was strongly selected against at most positions, likely because prolines kink the peptide backbone and are often highly destabilizing. Substitution to cysteine, which can lead to inappropriate disulfide bond formation, was almost always deleterious. In addition, substitution to stop codons was strongly selected against, consistent with truncated WW domains being unable to bind ligand. However, that stop codon mutations still remain after six rounds of selection puts a lower bound on the relative affinities that can be assessed.

Comparison of the amino acid substitution profiles illustrates the complexity of mutational tolerance (Fig. 3). Some major features were conserved among profiles, consistent with the mutationally permissive and restrictive regions of the WW domain (Fig. 2b). However, distinct, often radical, differences among these profiles argue that each position possesses a unique mutational preference (Supplementary Fig. 6). For example, at position 35, a change to any positively charged amino acids was beneficial. An alignment of WW domains shows that a positive charge at this position is conserved (present in 47 of 73 aligned WW domains, Supplementary Fig. 7). Such preferences cannot be captured by standard approaches that examine a few mutations, demonstrating the power of a high-resolution approach. The accuracy of these mutational preferences could be refined by analyzing data from sequential rounds of selection and by applying quantitative models that take into account nonspecific carryover, saturation effects and error analysis.

Comparison to Literature Analyses

We sought to use characterized mutations in the WW domain to validate our results. We searched the literature for experiments in which binding, folding or stability data had been acquired for a WW domain variant. All experiments were considered, regardless of the similarity of the WW domain analyzed to the hYAP65 WW domain. However, literature examples were excluded if the residue mutated differed from the hYAP65 WW domain residue at that position. Our search produced 72 variants that could be compared to the hYAP65 WW domain10–18 (Table 2).

To eliminate the confounding effect of hYAP65 WW domain variants with multiple mutations, we compared the list of literature variants to phage display enrichment ratio data only for hYAP65 variants with single amino acid mutations that had significant changes. We included variants that were considered significantly enriched or depleted in round six, by a Poisson-based test, relative to the input library (Supplementary Fig. 8). Enrichment ratios for 56 of the 72 variants curated from the literature were present in our data and used to calculate fitness values (relative to wild type). We predicted that literature variants would be deleterious in the phage display selection if they had lower measured binding affinity, were thermodynamically less stable or were less structured, and beneficial if they had a higher measured binding affinity, were thermodynamically more stable or were more structured. The predicted effects of literature variants (i.e. beneficial or deleterious relative to wild type) agreed in 52 of 56 cases with the phage display fitness data (Table 2). This striking concordance highlights the ability of the phage display assay to function as an integrated measure of folding, stability and binding.

As an additional method of validation, we picked five strongly selected, singly mutated variants and tested their ability to compete with wild type WW domain in the phage display assay (Supplementary Fig. 9). All five performed as predicted against the wild type WW domain.

Modeling the Effect of Mutations on Folding and Binding

The phage display assay captures features of protein function such as thermodynamic stability and affinity for the ligand. To better understand the contribution of protein folding and binding to the performance of WW domain variants in the assay, we employed a computational approach. For all calculations, we included full-length variants that were considered significantly enriched or depleted in round six, by a Poisson-based test, relative to the input library (Supplementary Fig. 8). Folding energy predictions were obtained using Rosetta25, allowing limited backbone flexibility, for these 16,363 variants. As expected, predicted folding energies were inversely related to enrichment ratios. Enriched variants were generally correctly predicted to be stabilized, and depleted variants were correctly predicted to be destabilized, relative to wild type (Fisher's exact test, $P = 9.36 \times 10^{-84}$). These predictions were accurate for 72% of the singly and 83% of the doubly mutated variants (Fig. 4a, Supplementary Fig. 10). Predictions of more highly mutated variants agreed less frequently (e.g. 39% agreement for triply mutated variants) with experimental data. The stability of depleted variants was predicted to be substantially less, on average, than that of the wild type hYAP65 WW domain. We conclude that mutations that produce thermodynamic instability are the likely cause for the depletion of most variants. We also calculated binding energies for those residues of the WW domain (Y28, L30, T37, and W39) in substantial contact with the ligand26. Among variants with single amino acid substitutions at these residues, 69% have predicted binding energies higher than wild type and highly negative enrichment ratios (Supplementary Fig. 11).

Single Mutants Predict Double Mutant Behavior

We used a simple product model to generate 10,192 predicted double mutant enrichment ratios from available single mutant data. These predicted scores had strong (Pearson's $R^2 = 0.68$) correlation with observed double mutant enrichment ratios, showing that single mutant behavior could generally be used to predict double mutant behavior (Fig. 4b). How multiple mutations interact is an important question in protein science and evolution27,28. Sequencing of sequential rounds of selection would permit appropriate correction for phage non-specific carryover and saturation, enabling this question to be addressed quantitatively using our approach.

Discussion

We present a method of analyzing protein sequence–function relationships that simultaneously quantifies the activity of hundreds of thousands of variants of a particular protein. Although the test case here employed phage display, this method could be applied to any protein functional assay in which phenotype and genotype are linked. Because its key ingredients – protein display, low intensity selection, and highly accurate, high throughput sequencing – are simple and becoming widely available, this approach is readily applicable to many *in vitro* and *in vivo* questions in which the activity of a protein is known and can be quantitatively assessed.

For example, libraries of plasmid-encoded protein variants could be introduced into any suitable host cell or organism, followed by an appropriate selection (e.g. for growth, fluorescence, transcriptional activity, or other function). Sequencing of plasmids present in the populations of input and selected cells or organisms should sort the variants into functional classes. The genetic basis of disease could be explored by comprehensively profiling the functional effects of mutations on a disease-linked protein, perhaps allowing a prediction of disease progression for any newly identified cases in which this protein is affected. To guide drug treatment and development, a nearly complete map of resistance to antibiotics or anticancer drugs could be constructed by carrying out functional selections in the presence of a drug. Protein sequence–function maps could also serve to advance efforts to better understand protein evolution, folding and engineering. Ultimately, this method encourages a re-imagining of the concept of protein functional screening. Moderate selection coupled with high throughput sequencing can produce dense, fine-scale functional maps, complementing approaches that identify only a few variants with desired properties.

Methods

Reagents

All chemicals were purchased from Sigma-Aldrich unless otherwise stated. Oligonucleotides were purchased from IDT with standard desalting and used without further purification unless otherwise stated. A complete list of DNA oligonucleotides used in this study including primers can be found in Supplementary Table 2.

WW Domain Cloning and Phage Construction

Codon optimized sequence corresponding to the WW domain #1 of the hYAP65 used in previous WW domain phage display experiments19 was generated using sequential PCR with the DF-31_WWWT and DF-62_WWv2 primers. The wild type, codon-optimized WW domain sequence was cloned into pGEM-T (Promega). The point mutants were constructed using site-directed mutagenesis with DF-106 primers (see Supplementary Table 2 for sequences). The WW domain variant library oligonucleotide was purchased from Trilink Biosciences; the variable region was doped with 2.1% non-wild type nucleotide (i.e. 0.7% of each of the three non-wild type nucleotides). The variant library was amplified with the WWVarV2 primers using a Platinum Taq HiFi PCR kit according to the manufacturer’s instructions (Invitrogen). For cloning into the T7Select 10-3b phage display vector, Platinum Taq HiFi PCR was used to generate template. The PCR reaction was processed using a Qiaquick PCR cleanup kit (Qiagen) and then digested with EcoRI and HindIII (New England Biolabs). Digested fragments were purified on a 1% agarose/TBE gel and quantified by fluorescence using the PicoGreen dye (Invitrogen). The fragments were cloned into the T7Select 10-3b phage display vector and packaged into infective phage using the T7Select phage display kit following the manufacturer’s instructions (Novagen/EMD Chemicals). Phage manipulation and titering were carried out following the manufacturer’s instructions.

WW Domain Phage Selection

The WW domain phage selection protocol was based on previous experiments19. Biotinylated GTPPPPYTVG peptide, with a disulfide linkage between biotin and peptide (bio-S-S-PPPY), at >90% purity was purchased from Anaspec and dissolved at 3 mM in deionized water. Streptavidin-coated magnetic beads (1 mg, M-1002-010, SoluLink) were washed three times with 500 µL phage wash buffer (PWB, 25 mM Tris pH 7.2, 150 mM NaCl, 0.1% (v/v) Tween-20). The beads were then incubated in 100 µL of a 100 µM solution of the bio-PPPY peptide with mixing at room temperature for one hour. The beads were washed three times with 500 µL PWB and blocked for one hour with blocking buffer (#37515, Pierce/Thermo Fisher). For each round of phage selection, amplified phage (100 µL) was added to bio-S-S-PPPY coated beads (73.5 µg or 125 pmol binding capacity) and mixed at room temperature for one hour. The beads were washed eight times with 200 µL PWB; each wash consisted of 5 seconds of vortexing followed by brief centrifugation. The bound phage were eluted in 100 µL PWB + 20 mM dithiothreitol for five minutes. The final wash and eluate were titered, and 80 µL of the eluate was amplified. WW domain sequences were PCR amplified and Sanger sequenced with primers DF-62_T7UPv2 and

DF-62_T7DNv2 (Supplementary Table 2). Sanger sequencing of mixed phage populations was quantified using the PeakPicker software29.

Illumina Library Preparation and Sequencing

Phage library DNA was isolated by phenol/chloroform extraction and ethanol precipitation. 50 ng of phage DNA was amplified using the Phusion HF system and the DF-97_PCR_Long primers (Supplementary Table 2) for 15 cycles according to the manufacturer’s instructions (New England Biolabs). The product was isolated using the Qiaquick PCR purification kit (Qiagen) and purified using the Ampure reagent (Beckman-Coulter). Concentrations were determined using the Quant-IT dsDNA HS kit according to the manufacturer’s instructions (Invitrogen). 0.36 fmol total DNA from each library was loaded into its own lane and sequenced using a Genome Analyzer IIx (Illumina) with the DF-97_SEQ_Short primers (Supplementary Table 2).

Sequence Assembly and Quality Filtration

Sequencing data from each round as well as the input library were treated identically. An average Illumina quality score was calculated for each read in a given set of paired end reads; read pairs in which either read had an average Phred quality score of less than 20 (i.e. 99% accuracy) were discarded. Read pairs were then aligned using a global Needleman-Wunsch algorithm in the pairwise2 BioPython function (match score = 2, mismatch score = −1, gap initiation = −3, gap propagation = −1). Read pair alignments with gaps were discarded. Surviving read pairs were merged into a single sequence, with disagreements between the read pairs being resolved in favor of the higher-quality base. Read pairs with disagreements and equal quality scores (i.e. where discrimination between the two disagreeing bases was impossible) were discarded. DNA sequences and their amino acid translations were aligned with the wild type WW domain sequence. Mutations were enumerated. Sequences with more than three consecutive mutations were discarded, as these tended to contain obvious sequencing errors.

Calculation of Mutation Frequencies and Enrichment Ratios

We identified unique sequences (variants) from each library and computed their frequency in the library. Variants that disappeared from the input library after selection were not included. The frequency for a given variant, v, in the Xth round ($F_{v,X}$) is:

$$F_{v,X}=\frac{reads_{v,X}}{\sum reads_X}$$

We used the frequency data to compute variant enrichment ratios between successive rounds. The enrichment ratio for a given variant, v, in the Xth round ($E_{v,X}$) is:

$$E_{v,X}=\frac{F_{v,X}}{F_{v,input}}$$

Fitness scores were calculated for each variant, v, to enable comparison to wild type in the Xth round ($W_{v,X}$):

$$W_{v,X}=\frac{E_{v,X}}{E_{WT,X}}$$

We also calculated mutation frequencies at each position for each amino acid substitution. The frequency at the ith position bearing the jth amino acid substitution in the Xth round ($F_{i,j,X}$) is:

$$F_{i,j,X}=\frac{\sum reads_{i,j,X}}{\sum reads_{X}}$$

We used the frequency data to compute enrichment ratios at each position for each amino acid substitution between successive rounds. The enrichment ratio at the ith position bearing the jth amino acid substitution in the Xth round ($E_{i,j,X}$) is:

$$E_{i,j,X}=\frac{\sum F_{i,j,X}}{\sum F_{i,j,input}}$$

Comparison to Literature Analysis

Experiments were curated from the literature in which binding, folding or stability data had been measured for a WW domain variant. All data were considered, regardless of the similarity of the WW domain analyzed to the hYAP65 WW domain. However, literature examples were excluded if the residue mutated differed from the hYAP65 WW domain residue at that position. For each literature variant, an expectation of enrichment or depletion, relative to wild type, was inferred. Variants that bound peptide more tightly, were more structured or had lower folding energies than wild type were expected to be enriched and vice versa. Each expectation was compared to a fitness score calculated for the relevant single mutant using Poisson filtered round 6 enrichment ratios.

WW Domain Conservation Analysis

A consensus WW domain sequence was built by alignment of 72 PFAM seed sequences of the WW domain (PF00397) in Jalview using default parameters. Conservation at each position was calculated as the fraction of residues containing the consensus amino acid. Residues 173–202 of the human Yap65 protein spanning the sequenced 25 amino acid variable region were aligned to the 72 seed WW domain sequences. Amino acid mutations that increased in frequency after 6 rounds of selection were used to generate a logo plot where the height of each amino acid indicates the frequency of enrichments to that residue (WebLogo, http://weblogo.berkeley.edu/).

Folding and Binding Energy Calculations

Rosetta was used to compute binding energies between the mutant WW variants and the peptide26 as well as the differences in folding free energies between mutant and wild type. The structure of the wild type WW-domain bound to the peptide (Protein Data Bank identifier 1kq9) was relaxed in several rounds of full repacking of sidechains on both partners and minimization of sidechain, backbone and rigid-body degrees of freedom. In the case of binding free-energy predictions, each mutant sequence was then threaded on this model and subjected to one round of full sidechain repacking followed by sidechain, backbone and rigid-body minimization. For each sequence, the binding energy was computed 5 times and the average of that value was reported. This procedure was repeated 5 times for each mutant and the average of these separate runs was reported. Differences in folding free energy ($\Delta\Delta G$) were performed by Rosetta (unpublished data). For both mutant and wild type structures, all sidechains were repacked followed by backbone and sidechain minimization. To ensure that the structures remained close to the starting structures, constraints were applied to all C-alpha atoms closer than 9 Å. This procedure was performed 50 times, and the predicted folding free energy is the average of the three lowest scoring structures. Rosetta scripts and command lines can be found in Supplementary Note 1.

Prediction of Double Mutant Behavior From Single Mutants

All calculations used Poisson significance-filtered round six data. We used a product model to predict the enrichment ratio of each double mutant variant ($E_{xy,6}$) using the corresponding single mutant variant enrichment ratios ($E_{x,6}$, $E_{y,6}$):

$$E_{xy,6}{=}E_{x,6}\times E_{y,6}$$

This simple model does not contain an error term (e.g. to account for non-specific carryover) and therefore should not be used to analyze epistasis.

Statistical Analysis

A statistical analysis was undertaken in order to determine which variants changed in abundance over successive rounds of selection. A Poisson distribution was used to estimate the likelihood that the representation of each unique variant in a given round of selection had changed from the input library. P-values were calculated by comparing the number of times each unique variant was observed in a given round of selection to the Poisson distribution based on the number of times that variant had been observed in the input library. Multiple testing correction was performed using false discovery rate control30.

Fisher’s exact test was applied to determine whether variants are accurately classified in a contingency table with categories for variants more or less active than wild type, and predicted to be more or less stable than wild type.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Charlie Lee and Jay Shendure for assistance with DNA sequencing, and Jeff Kelly, Jim Thomas, Jay Hesselberth, Eric Phizicky, Alan Rubin, Lea Starita and Kelly McGarvey for helpful comments and discussion. This work was supported by P41RR11823 to S.F.; D.M.F. is the recipient of a Ruth L. Kirchstein National Research Service Award (F32GM084699). S.F. and D.B. are investigators of the Howard Hughes Medical Institute.

References Cited

1. Sidhu SS, Koide S. Phage display for engineering and analyzing protein interaction interfaces. Curr. Opin. Struct. Biol. 2007; 17:481–487. [PubMed: 17870470]
2. Matouschek A, Kellis JT Jr, Serrano L, Fersht AR. Mapping the transition state and pathway of protein folding by protein engineering. Nature. 1989; 340:122–126. [PubMed: 2739734]
3. Cunningham BC, Wells JA. High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. Science. 1989; 244:1081–1085. [PubMed: 2471267]
4. Levin AM, Weiss GA. Optimizing the affinity and specificity of proteins with molecular display. Mol. Biosyst. 2006; 2:49–57. [PubMed: 16880922]
5. Pal G, Kouadio JL, Artis DR, Kossiakoff AA, Sidhu SS. Comprehensive and quantitative mapping of energy landscapes for protein-protein interactions by rapid combinatorial scanning. J. Biol. Chem. 2006; 281:22378–22385. [PubMed: 16762925]
6. Dias-Neto E, et al. Next-generation phage display: integrating and comparing available molecular tools to enable cost-effective high-throughput analysis. PLoS One. 2009; 4:e8338. [PubMed: 20020040]
7. Ge X, Mazor Y, Hunicke-Smith SP, Ellington AD, Georgiou G. Rapid construction and characterization of synthetic antibody libraries without DNA amplification. Biotechnol. Bioeng. 2010; 106:347–357. [PubMed: 20198660]
8. Di Niro R, et al. Rapid interactome profiling by massive sequencing. Nucleic Acids Res. 2010; 38:e110. [PubMed: 20144949]
9. Macias MJ, Wiesner S, Sudol M. WW and SH3 domains, two different scaffolds to recognize proline-rich ligands. FEBS Lett. 2002; 513:30–37. [PubMed: 11911877]
10. Espanel X, Navin N, Kato Y, Tanokura M, Sudol M. Probing WW Domains to Uncover and Refine Determinants of Specificity in Ligand Recognition. Cytotechnology. 2003; 43:105–111. [PubMed: 19003214]
11. Jager M, Nguyen H, Crane JC, Kelly JW, Gruebele M. The folding mechanism of a beta-sheet: the WW domain. J. Mol. Biol. 2001; 311:373–393. [PubMed: 11478867]
12. Jiang X, Kowalski J, Kelly JW. Increasing protein stability using a rational approach combining sequence homology and structural alignment: Stabilizing the WW domain. Protein Sci. 2001; 10:1454–1465. [PubMed: 11420447]
13. Kasanov J, Pirozzi G, Uveges AJ, Kay BK. Characterizing Class I WW domains defines key specificity determinants and generates mutant domains with novel specificities. Chem. Biol. 2001; 8:231–241. [PubMed: 11306348]
14. Koepf EK, et al. Characterization of the structure and function of W --> F WW domain variants: identification of a natively unfolded protein that folds upon ligand binding. Biochemistry. 1999; 38:14338–14351. [PubMed: 10572009]
15. Nguyen H, Jager M, Moretto A, Gruebele M, Kelly JW, et al. Tuning the free-energy landscape of a WW domain by temperature, mutation, and truncation. Proc. Natl. Acad. Sci. USA. 2003; 100:3948–3953. [PubMed: 12651955]
16. Pires JR. Solution structures of the YAP65 WW domain and the variant L30 K in complex with the peptides GTPPPPYTVG, N-(n-octyl)-GPPPY and PLPPY and the application of peptide libraries reveal a minimal binding epitope. J. Mol. Biol. 2001; 314:1147–1156. [PubMed: 11743730]
17. Toepert F, Pires JR, Landgraf C, Oschkinat H, Schneider-Mergener J. Synthesis of an Array Comprising 837 Variants of the hYAP WW Protein Domain. Angew. Chem. Int. Ed. Engl. 2001; 40:897–900. [PubMed: 11241639]

18. Yanagida H, Matsuura T, Yomo T. Compensatory evolution of a WW domain variant lacking the strictly conserved Trp residue. J. Mol. Evol. 2008; 66:61–71. [PubMed: 18087661]
19. Dalby PA, Hoess RH, DeGrado WF. Evolution of binding affinity in a WW domain probed by phage display. Protein Sci. 2000; 9:2366–2376. [PubMed: 11206058]
20. Dai M, et al. Using T7 phage display to select GFP-based binders. Protein. Eng. Des. Sel. 2008; 21:413–424. [PubMed: 18469345]
21. Quail MA, et al. A large genome center's improvements to the Illumina sequencing system. Nat. Methods. 2008; 5:1005–1010. [PubMed: 19034268]
22. Knight R, Yarus M. Analyzing partially randomized nucleic acid pools: straight dope on doping. Nucleic Acids Res. 2003; 31:e30. [PubMed: 12626729]
23. Weiss GA, Watanabe CK, Zhong A, Goddard A, Sidhu SS. Rapid mapping of protein functional epitopes by combinatorial alanine scanning. Proc. Natl. Acad. Sci. USA. 2000; 97:8950–8954. [PubMed: 10908667]
24. Guo HH, Choe J, Loeb LA. Protein tolerance to random amino acid change. Proc. Natl. Acad. Sci. USA. 2004; 101:9205–9210. [PubMed: 15197260]
25. Das R, Baker D. Macromolecular modeling with Rosetta. Annu. Rev. Biochem. 2008; 77:363–382. [PubMed: 18410248]
26. Kortemme T, Baker D. A simple physical model for binding energy hot spots in protein-protein complexes. Proc. Natl. Acad. Sci. USA. 2002; 99:14116–14121. [PubMed: 12381794]
27. Bershtein S, Segal M, Bekerman R, Tokuriki N, Tawfik DS. Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. Nature. 2006; 444:929–932. [PubMed: 17122770]
28. Weinreich DM, Delaney NF, Depristo MA, Hartl DL. Darwinian evolution can follow only very few mutational paths to fitter proteins. Science. 2006; 312:111–114. [PubMed: 16601193]
29. Ge B, et al. Survey of allelic expression using EST mining. Genome Res. 2005; 15:1584–1591. [PubMed: 16251468]
30. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. USA. 2003; 100:9440–9445. [PubMed: 12883005]

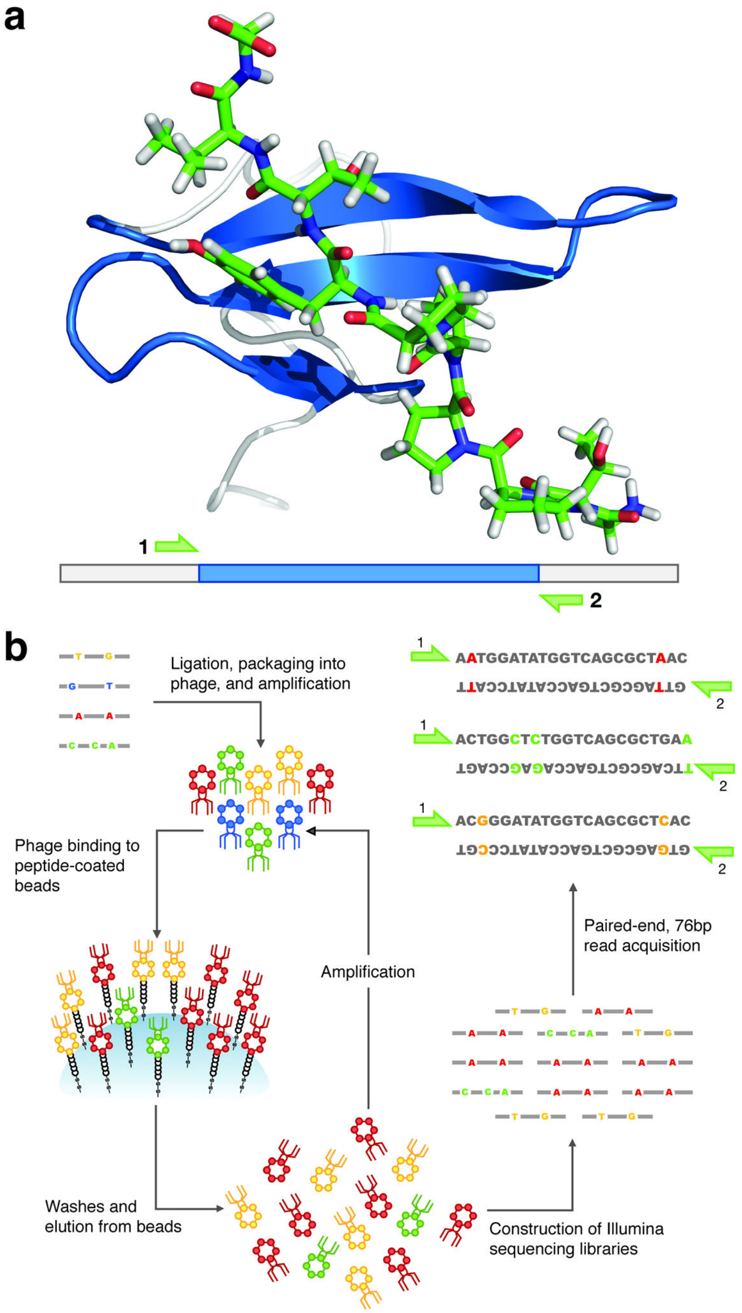
a
1
2
b
Ligation, packaging into phage, and amplification
Phage binding to peptide-coated beads
Amplification
Washes and elution from beads
Construction of Illumina sequencing libraries
Paired-end, 76bp read acquisition
1
AATGGATATGGTCAGCGCTAAC
GTTAGCGCTGACCATATCCATT
2
1
ACTGGCTCTGGTCAGCGCTGAA
TTCAGCGCTGACCAGAGCCAGT
2
1
ACGGGATATGGTCAGCGCTCAC
GTGAGCGCTGACCATATCCCGT
2

Figure 1. A Highly Parallel Assay For Exploring Protein Sequence-Function Relationships

(**a**) The NMR structure of the hYAP WW domain (Protein Data Bank identifier 1jmq) is shown, as a cartoon, in complex with its peptide ligand16 (www.pymol.org). In both the structure and the schematic below, the blue portion of the hYAP WW domain indicates the mutagenized and sequenced region. (**b**) A library of variant WW domains was generated using chemical DNA synthesis with doped nucleotide pools, amplified using PCR and then displayed as a fusion to the capsid protein of T7 bacteriophage. The input phage library was subjected to successive rounds of selection. Each round consisted of phage binding to peptide ligand immobilized on beads, washing to remove unbound phage, and elution and amplification of bound phage. Sequencing libraries were created using PCR from the input phage and the phage after three and six rounds of selection, and were sequenced using overlapping paired-end reads on the Illumina platform. An example of four unique variants of differing affinity are shown in different colors. The green arrows indicate the location of the sequencing primers.

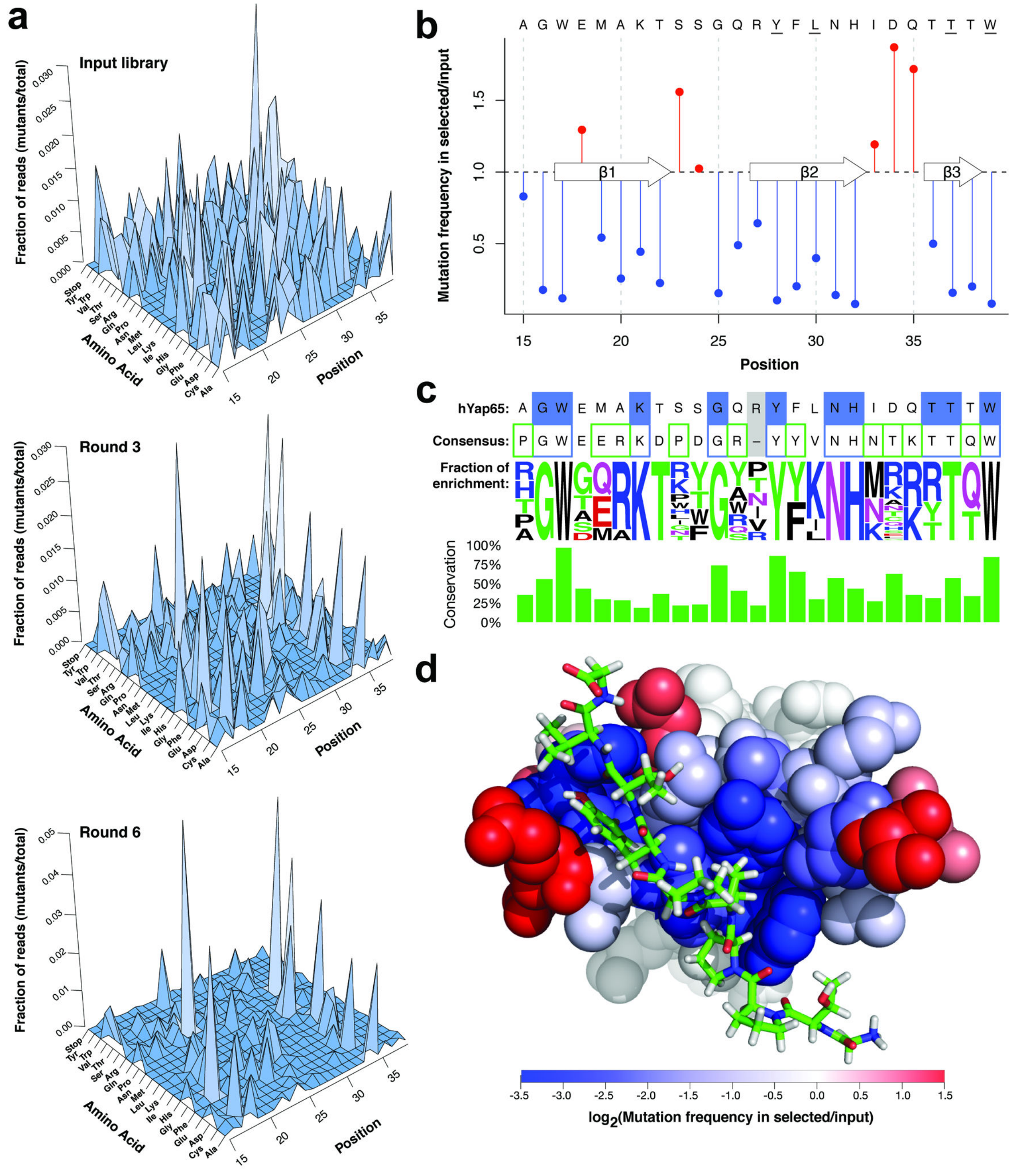

Figure 2. Comparison of Mutational Tolerance and Evolutionary Conservation in the WW Domain

(**a**) The mutational diversity of the input, round three and round six WW domain libraries is shown. Mutations are enumerated by position within the domain and by amino acid substitution. (**b**) Shown is a ratio of mutational frequencies (round six/input) observed at each position within the WW domain. Positions intolerant to mutation are shown as blue bars; beneficial mutations are shown as red bars. Positions making substantial contact with the peptide are underlined. (**c**) We compared the mutational preference at each position of the hYAP65 WW domain (top sequence) to a consensus WW domain sequence (bottom sequence). Positions where the hYAP65 sequence is identical to the consensus and that are mutationally intolerant in our assay are highlighted in blue. Using mutational frequencies for enriched variants, we generated a logo plot that indicates mutational preference at each position (i.e. the plot shows only mutations that are advantageous). Positions where mutations to the consensus sequence are beneficial are highlighted in green. A plot of conservation is shown as a percent of sequences in the alignment that are identical to the consensus at each position. (**d**) The mutational frequency ratio data from (**b**) were projected ($\log_2$ transformed) onto the space-filling model of the hYAP65 WW domain NMR structure (Protein Data Bank identifier 1jmq) using the PyMOL software16 (www.pymol.org). Positions at which the frequency of mutations increase are shown in red and positions at which the frequency of mutations decrease are shown in blue.

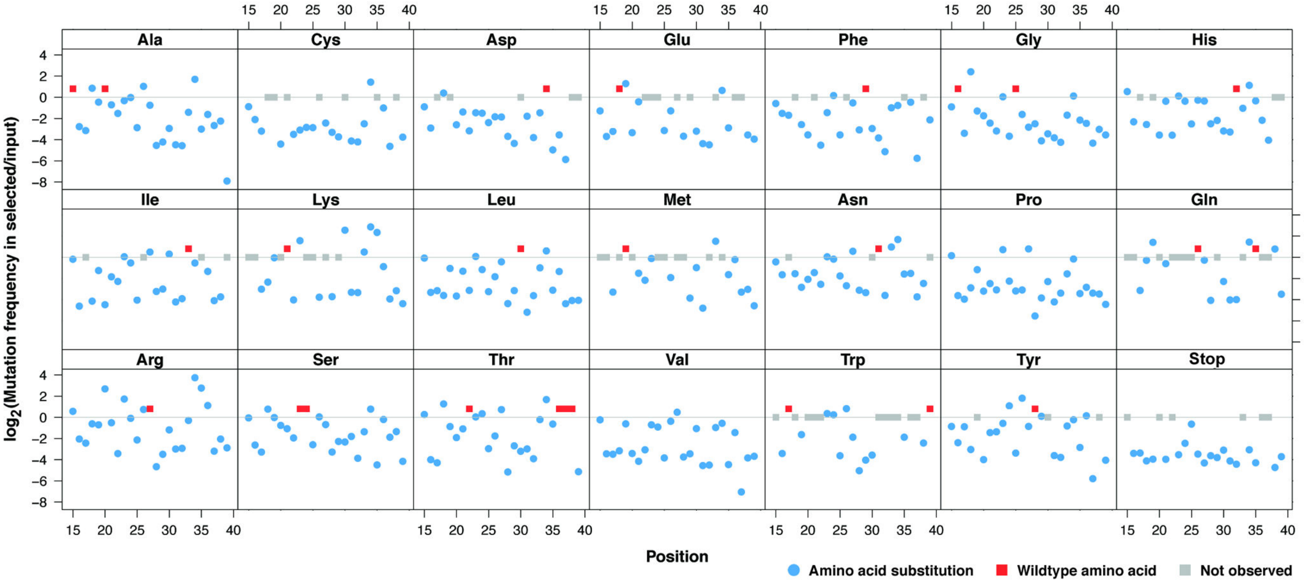

Figure 3. A Comprehensive Sequence-Function Map of the WW Domain

We calculated enrichment ratios (round six/input) for each amino acid at each position within the WW domain. Each panel of the plot corresponds to a different amino acid substitution profile, whose three letter code is displayed in the header bar. The x-axis of each panel indicates the position, from left to right, along the WW domain while the y-axis indicates $\log_2$(enrichment ratio). Blue dots indicate a measured enrichment ratio and red dots indicate the wild type sequence, which enriched 1.7-fold. Gray dots indicate mutations not observed, and were arbitrarily placed at zero. The upper left panel of the plot corresponds to a traditional alanine scan of the WW domain.

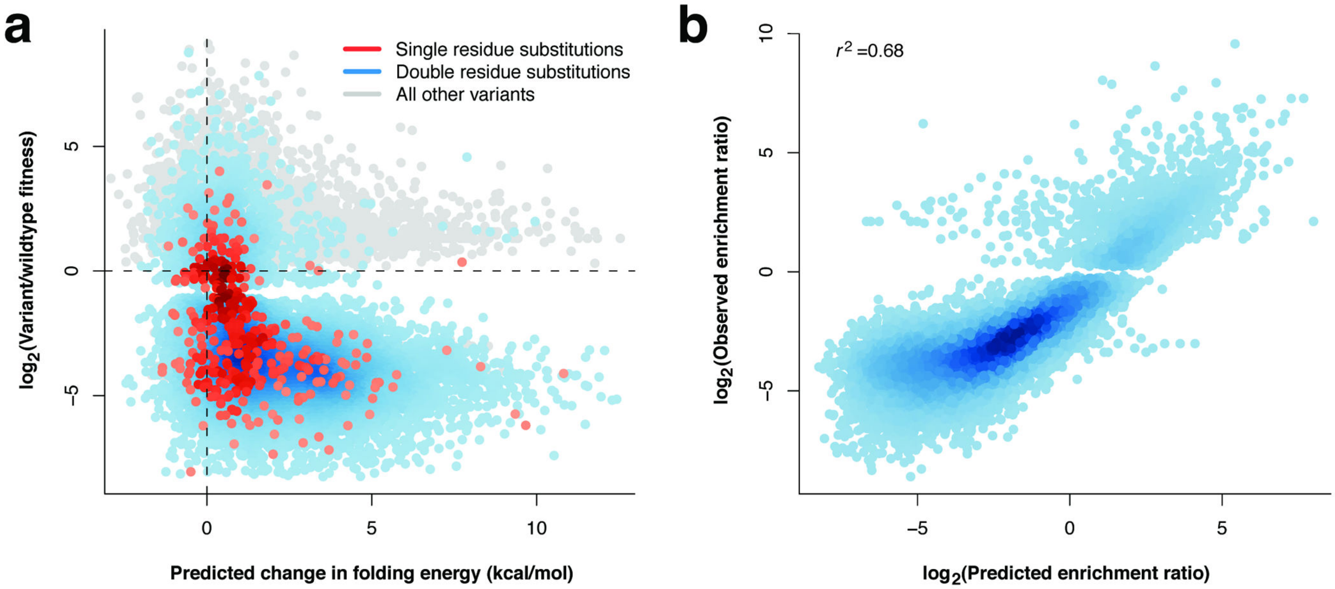

Figure 4. Prediction of WW Domain Folding Energies and Double Mutant Enrichment Ratios

The Rosetta framework was used to calculate folding energies for the 16,363 full-length WW domain variants that were significantly enriched or depleted after six rounds of selection. Predicted folding energies relative to the wild type WW domain energy are plotted against the observed fitness (variant/wild type) for the variants containing one (red), two (blue), or more mutations (gray). (**b**) Using a basis set of single mutant enrichment ratios, we predicted double mutant enrichment ratios using a product model.

Table 1

Observed Versus Predicted Mutations in the Input Library

Number of Mutations	Observed (DNA)	Predicted (DNA)	Percent (DNA)	Observed (Protein)	Predicted* (Protein)	Percent (Protein)
0	1	1	100	1	1	100
1	225	225	100	416	500	83.2
2	24,924	24,975	99.8	32,779	120,000	27.3
3	648,853	1,823,175	35.6	328,768	18,400,000	1.79

*includes stop codons

Table 2

Comparison of the Effects of WW Domain Mutations in this Study to Literature

Mutation	Agreement	Effect	Evidence
W17F	yes	no structure	thermal denaturation11, NMR14, ITC14
W17Y	yes	worse binding	SPR17
E18Q	yes	less stable	thermal denaturation11
A20R	yes	more stable	thermal and chemical denaturation12
K21R	no	better binding	phage18 , SPR18
S24X	yes (8/11)	tolerant to any change[*]	phage ELISA13
Y28L	yes	less stable	thermal denaturation11,15
F29L	yes	less stable	thermal denaturation11
L30K	yes	more stable	NMR16
H32X	yes (17/17)	intolerant to any change[‡]	phage ELISA13
I33A	yes	less stable	thermal denaturation11
D34T	yes	more stable	thermal and chemical denaturation12
Q35R	yes	better binding	phage10
T37X	yes (15/15)	intolerant to any change[‡]	phage ELISA13
W39A	yes	less stable	thermal denaturation15
W39F	yes	no binding	NMR14 , ITC14

[*] >15% binding activity

[‡] <15% binding activity