Proteome-wide detection of Abl1 SH3 binding peptides by integrating computational prediction and peptide microarray

Zheng Xu¹#, Tingjun Hou³#, Nan Li²#, Yang Xu¹ and Wei Wang²*

1. Division of Biological Sciences, 2. Department of Chemistry and Biochemistry, University of California, San Diego, 9500 Gilman Drive, La Jolla, Ca 92093-0359; 3. Institute of Functional Nano & Soft Materials (FUNSOM) and Jiangsu Key Laboratory for Carbon-Based Functional Materials & Devices, Soochow University, Suzhou, Jiangsu 215123, P. R. China.

#equal contribution
*corresponding author, email: wei-wang@ucsd.edu
Abstract

Protein-protein interactions are essential for regulating almost all aspects of cellular functions. Many of these interactions are mediated by weak and transient protein domain-peptide binding, but they are often underrepresented in high throughput screening of protein-protein interactions using techniques such as yeast two hybrid and mass spectrometry. On the other hand, computational predictions and in vitro binding assays are valuable in providing clues of in vivo interactions. We present here a systematic approach that integrates computer modeling and a peptide microarray technology to identify binding peptides of the SH3 domain of the tyrosine kinase Abl1 in the human proteome. Our study provides a comprehensive list of candidate interacting partners for the Abl1 protein, among which the presence of numerous methyltransferases and RNA splicing proteins may suggest a novel function of Abl1 in chromatin remodeling and RNA processing. This study illustrates a powerful approach of integrating computational and experimental methods to detect protein interactions mediated by domain-peptide recognition.
Introduction

The human tyrosine kinase Abl1 plays important roles in signal transduction and interacts with a broad variety of cellular proteins (1, 2). Fusion of the break-point cluster region (BCR) and ABL genes form the oncogenic BCR-ABL, which is causally linked to certain human leukemias (3). The Abl1 protein consists of several modular domains, including one SH3 and one SH2 domain that form intra- and inter-molecular interactions to regulate its kinase activity (1, 2). Although many interacting partners of Abl1, such as Abi1, Abi2 and Rin1, have been reported (2), a proteome-wide identification of those mediated by the weak and transient SH3-peptide interaction is still incomplete and remains a challenging problem.

High throughput technologies such as yeast two-hybrid and protein complex purification in conjunction with mass spectrometry have greatly facilitated the identification of protein-protein interactions. However, because they are not optimally designed to detect the weak and transient domain-peptide interactions, such interactions are often under-represented in proteomic screenings (4, 5). On the other hand, in vitro binding assays such as phage display, peptide array and protein array are often employed to determine the binding specificity of modular domains. Each of these technologies has advantages and limitations. Phage display and peptide array are complementary techniques: the former can test binding of a huge number of random peptides against a given domain while the later can determine binding of a set of specific peptides of interest. Compared with these two techniques, protein array often presents a relatively smaller number of proteins on a surface to test binding of synthesized peptides or purified proteins. Computational analysis is critical to these in vivo or in vitro experimental measurements as it 1) removes false positives and recovering false negatives for the in vivo data; 2) predicts proteome-wide interacting partners based on binding motifs determined by in vitro experiments to guide further experimental investigation.

We present here an integrative approach to identify interacting partners of Abl1 in the human proteome: computational predictions were made by combining multiple sources of data, including structural information, energetic pattern of SH3-peptide interactions and conservation during evolution; the predicted interactions were then tested using a peptide microarray technology. Different from the popular SPOT technology that synthesizes peptides on a cellulose membrane (6), we printed pre-synthesized and quality-controlled peptides on a glass surface to hybridize with the purified Abl1 SH3 domain, which confirmed 237 predicted interactions. Proteome-wide identification of putative interacting partners of Abl1 mediated by the SH3-peptide interactions may reveal novel functions of Abl1 in chromatin remodeling and RNA processing. It will also greatly enhance our understanding of the regulatory mechanism of Abl1.

Results

A pipeline to predict protein-peptide interactions in the human proteome

To comprehensively identify interacting partners of the Abl1 protein, we exploited a systematic searching strategy (Figure 1). We first scored all 69,404 peptides in the human proteome that contain the PxxP motif in the UniProt database (7) using a scoring matrix. This scoring matrix was generated using the Virtual Mutagenesis method we previously developed (see details in (8)). Briefly, each amino
acid of the peptide was mutated to the other 19 amino acids and the mutated complex structure was optimized using 2ns molecular dynamics simulations. The difference in SH3-peptide binding free energy between the wild type and mutated peptides was calculated using the MM/PBSA method(9, 10) to represent how much an amino acid was preferred at each peptide position.

Next, we focused on analyzing the top 10,000 peptides ranked by the scoring matrix and used conservation information to remove false positives. Because of functional constraints, the binding peptides of the Abl1 SH3 domain should be more conserved across species than non-binding peptides. For each protein that contains at least one putative binder, we generated multiple sequence alignment across seven proteomes, including *Homo sapiens* (human), *Pan troglodytes* (chimpanzee), *Macaca mulatta* (rhesus macaque), *Mus musculus* (mouse), *Rattus norvegicus* (rat), *Canis familiaris* (dog) and *Bos taurus* (cow). Because the human Abl1 SH3 domain is completely conserved across these seven species, it is reasonable to assume that its interacting partners are also highly conserved. Therefore, we removed those non-conserved peptides from the candidate list. After applying the conservation filter, about half of the peptides were removed and 4981 peptides were left for further analysis.

Our goal was to find the most confident interacting partners of Abl1 SH3 domain and we used the following procedure to further remove false positives. As illustrated in our previous studies(11, 12), we modeled each of these 4981 peptides in complex with the Abl1 SH3 domain and calculated its molecular interaction energy components (MIECs) that reflect the energetic characteristics of the binding. Using the MIEC-SVM previously trained on 18 SH3 domains(12), we classified each peptide into binder or non-binder category. After this round of filtering, 1394 peptides in 714 proteins were predicted as binders of the Abl1 SH3 domain.

Among the top 10 predicted peptides shown in Table 1, five were from proteins known to interact with the Abl1 SH3 domain (in bold), including EVL that binds to the SH3 domain of Abl1. Considering how difficult such proteome-wide prediction is, a 50% (5/10) rate of identifying known interacting partners demonstrated a satisfactory performance of our approach. On the other hand, the remaining peptides may be new binders that have not been reported. To further refine our predictions, we conducted peptide microarray experiments on the predicted binding peptides to find the most reliable binders of the Abl1 SH3 domain.

**Determination of protein-peptide binding using peptide microarray**

Peptide array provides a much higher throughput measurement of protein-peptide binding than in-solution assays. Unlike the phage display method, peptide sequences on the array can be specified, which is crucial for confirming computational predictions. Numerous peptide array platforms are available and a popular one is the SPOT technology which synthesizes peptides directly on a cellulose membrane(6). Despite its wide usage, this technology has several limitations. One of them is the lack of quality control of the peptides because the peptides are directly synthesized onto the membrane. Another is that the amount of peptides printed on the membrane is relatively large, compared to other platforms, and therefore, more reagents are needed. To overcome these limitations we chose a new platform, in which peptides were synthesized in situ and then printed onto a glass surface in a way similar to the
DNA/RNA oligo microarray. The protein was overlaid onto the microarray and protein-peptide binding was quantified by measuring fluorescence intensity.

**Design of the peptide microarray and control experiments**

To print peptide microarrays, we needed to decide how to immobilize the peptides on the glass surface as well as the amount of peptide printed. Therefore, we first selected 10 peptides to test out the conditions. These peptides included 4 known binders of the Abl1 SH3 domain(13, 14) and 6 non-binders (including 4 peptides that bind to other SH3 domains, but not Abl1(13, 14), and 2 random peptides without the PxxP motif). Same as in our previous study(12), each peptide was 10-amino-acid long and two alanines were added at each end. A linker, either aminohexanoic acid (Ahx) or polyethylene glycol (PEG), was added at the N-terminus. The peptides were immobilized to the glass surface via the N-terminus. We tested two spot sizes, 150um and 600um in diameter, which corresponded to 1nl and 5nl of 0.3mg/ml peptide solution dispensed to the array, respectively. Each printing amount and linker was tested in triplicates. To better quantitatively measure the signal intensity, we used fluorescence and laser scanner to detect the binding of the SH3 domain to the peptides, instead of chemiluminescence and film in our previous study (12). Briefly, GST-Abl SH3 fusion protein was purified and characterized as described before (12) (data not shown). Purified protein was overlaid onto the array, followed by incubation with an anti-GST antibody and then a Cy3-conjugated secondary antibody. Binding of GST-Abl SH3 to peptides generated a Cy3 fluorescent signal that was recorded on a laser scanner (Table S1). The fluorescence intensity for each peptide was an average of triplicates.

Figure 2 shows that the peptide microarray correctly distinguished binders from non-binders. By quantifying signal from each spot, we found that there was no difference in signal intensity using either linker. As expected, the large spots generated much stronger overall signal than the small ones. However, the signal intensity was more even and the signal variation among triplicates was less in small spots than those in large spots. Hence we chose the lower cost Ahx linker and 150um spots for the rest of the experiments. As a control, we probed the array with anti-GST and the secondary antibodies alone and only basal levels of fluorescence were recorded (data not shown), demonstrating that there was no non-specific binding of these antibodies to the array.

**Analysis of the peptide microarray data**

In the production run, limited by the synthesis cost, we synthesized and printed the top 700 predicted binding peptides. We also included 50 binders (including 38 with previously measured Kd's(13)) and 50 non-binders on the array as positive and negative controls, respectively (Table S2 and S3). The non-binders were randomly selected from the proteome, which are unlikely to interact with the Abl1 SH3 domain. Each peptide was printed in triplicates at two concentrations, 1mg/ml and 0.5mg/ml. The production array was probed the same way as the test run (Figure 3). Similarly, we did a control experiment with just the antibodies and detected only basal levels of fluorescence (data not shown).

To analyze the microarray data, we first removed the spots that were flagged as noisy. For the high (1mg/ml) and low (0.5mg/ml) printing concentrations, 724 and 767 (out of 800) spots remained for later analyses, respectively. Next, we determined
the statistical cut-off for the significant fluorescence signal by modeling the background intensity distribution using a Gamma distribution (Figure 4). We binned the spots based on the fluorescence intensity. To have a better curve fitting, we removed the long tail of the histogram that represented strong binding peptides and only kept the bins with intensity less than 900 and 1500 for the high and low concentrations, respectively. The model fitting and p-value calculation were conducted using R and P-value of 0.05 was used as the cutoff.

After establishing the parameters, we first examined the 100 control peptides and found that 40 from the 50 positive controls exhibited a strong signal and 48 from the 50 negative controls displayed only basal signal (Figure 5). Therefore, our production array showed a satisfactory performance with a sensitivity of 0.80(=40/50) and a specificity of 0.96(=48/50). It is worth noting that we synthesized twice of the 10 false negative peptides, and tested their binding to the Ab1 SH3 domain using dot blot but only one of these 10 peptides showed signals (data not shown). Although these 10 peptides were reported to be binders, our experiments could not confirm this conclusion and the true sensitivity of our peptide microarray could be 0.98(=40/41).

We then checked the 700 predicted binders and identified 237 non-redundant peptides from 158 proteins as putative binders, including 220 and 175 from high and low concentrations, respectively (Table S4). When examining the PSSM scores of the 237 binding peptides and the remaining non-binding peptides, we found that the binders have more favorable (smaller) PSSM scores than the non-binders although there is significant overlap between the two distributions (Figure S1). This figure illustrates that the PSSM provides separation between binders and non-binders to some extent.

Among the identified binding peptides, we found 13 known binding partners of the Ab1 SH3 domain as well as 2 paralogs of the known binding partners(2) (Table 2). There are 52 proteins listed in (2) that interact with the Ab1 through SH3-peptide interaction and we retrieved 25% (13/52) or 29% (15/52) of these manually curated interacting partners of the Ab1 SH3 domain. Interestingly, six of them are also known substrates of the Ab1 kinase: ENAH, INPPL1, RAPH1, RIN1, WASF2 and WASL. In addition, the identified binding proteins also included another previously determined interacting partner of the SH3 domain that were not included in (2), SHIP2(15), and a known substrate and interacting partner of Ab1, WASF3(16). The interaction between Ab1 and its substrates is often mediated by SH3 domain (as in the case of the six substrates mentioned above). It is also known the interactions between WASF family proteins and Ab1 is mediated by SH3 domain(16). Therefore, WASF3 is likely to be a true interacting partner of the Ab1 SH3 domain. It is worth noting that there are numerous databases that document many partners of the Ab1 SH3 domain, but we only used the curated list in (2) as the gold standard for known interacting partners, which might miss quite a few true positives (like SHIP2). Given the challenge of proteome-wide identification of domain-peptide interactions, an 8.2% (13/158) rate of retrieving the curated interacting partners in this study was quite satisfactory while the other confirmed peptides may well be unknown binders.

We also compared our approach with another peptide array-based (SPOT array) method for detecting binding peptides of the Ab1 SH3 domain: only two known interacting partners (M4K1, DYN2), three paralogs of known interacting partners
(SOS2, ABR and EFS) and one known substrate (SYNJ2) listed in (2) were discovered in Wu et al.'s study (17) (2.5% = 2/81 recovering rate). The superior performance of our method demonstrates the power of integrating computational predictions and peptide microarray in identifying domain-peptide interactions in the human proteome.

Putative novel functions of Abl1 in chromatin modification and RNA processing

The identified Abl1 binding peptides came from 158 proteins. To illustrate their functions, we performed enrichment analysis of their gene ontology (GO)(18) annotations using the DAVID software package (19) with the default parameters (Table 3). It is unsurprising that many proteins are related to actin cytoskeleton function, in which Abl1 is well known to be involved. Unexpectedly, we found putative binding partners that are involved in chromatin modification, RNA processing, transcriptional regulation and apoptosis (Table 3). Interactions with these proteins suggested possible new functions of Abl1. We highlighted several interesting groups of putative interacting partners of Abl1 in Table 4.

(1) chromatin remodeling enzymes

Several methyltransferases are critical for transcription, for example, SET1A, SET1B and WBP7 methylate Lys4 of histone protein H3. Mono-, di- and tri-methylation of H3K4 (represented as H3K4me1, H3K4me2 and H3K4me3, respectively) are known to mark active promoters and enhancers (20). Besides our study, WBP7 was also found to interact with Abl1 SH3 domain in (17), although the function of this interaction is unclear.

Our analysis revealed the binding of Abl1 to SETD2, a histone methyltransferase that methylates H3K36, which is a mark for transcription elongation and splicing (20). The SETD2 peptides (amino acid 185-194 and 187-196) are located in a proline rich region. Because this region does not contain any identified domain, it is likely the peptide is accessible for SH3 binding. SETD2 is known to interact with hyper-phosphorylated POLR2A, the RNA polymerase II (Pol II) large subunit that is also a substrate of Abl1. Our observation implies possible recruitment of Abl1 by SETD2 to the proximity of the Pol II complex and thereby regulates transcription.

Another identified interacting partner of the Abl1 SH3 domain, DOT1, is a histone methyltransferase that methylates H3K79. H3K79me2 is a histone modification mark for transcribed regions (20). DOT1 also interacts with MLLT10, a myeloid/lymphoid or mixed-lineage leukemia protein. Confirmation of the interaction between Abl1 and MLLT10 and illustration of its functional importance await future experimental test.

Our analysis also revealed that Abl1 recognized peptides in JMJD3, a demethylase that demethylates repressive histone mark of H3K27me3 (20), as well as in EP300, an acetyltransferase that is a co-factor often located in enhancers (20). The other identified interacting partners involved in chromatin structure remodeling include a putative polycomb group protein, ASXL3 (21), and a chromatin modifier, CHD8 (22).

Additionally, we found that Abl1 interacted with two methyl-CpG binding domain-containing proteins, MBD5 and MBD6. The functions of these two proteins
are still largely unknown. Given Abl1’s potential functions in modification of chromatin structure, it would not be surprising if it does play roles in DNA methylation since there exists interplay between histone modifications and DNA methylation(23, 24).

(2) transcription and splicing complexes

In addition to these putative interactions between Abl1 and chromatin modification enzymes, our work showed that Abl1 SH3 domain also directly interacted with the transcription machinery, including TAF1, the largest component of the TFIID basal transcription factor complex(25) and MED19, a component of the Mediator complex and a co-activator involved in the regulated transcription of nearly all Pol II–dependent genes(26). Another putative Abl1 binder is ACINU, a component of the splicing-dependent multiprotein exon junction complex (EJC) deposited at splice junctions on mRNAs(27, 28). We also identified many proteins involved in RNA splicing and mRNA processing (Table 3), which suggested novel functions of Abl1 in transcription and slicing regulation.

(3) proteins involved in apoptosis

Two potential interacting proteins of Abl1 are known to be involved in apoptosis: PDCD7 promotes apoptosis when overexpressed(29); ASPP2 regulates p53 by enhancing the DNA binding and transactivation function of p53 on the promoters of pro-apoptotic genes in vivo(30). In addition, CHD8, a chromatin modifier and a putative Abl binder mentioned above, can suppress p53-mediated apoptosis by recruiting histone H1 and preventing p53 transactivation activity(31).

Discussion and Conclusion

The tyrosine kinase Abl1 is an important therapeutic target in hematopoietic malignancies. We present here a systematic approach of combining computational predictions and peptide microarray to identify proteins bound to Abl1 via the SH3-peptide recognition in the human proteome. A comparison with the documented Abl1 interacting proteins showed a satisfactory performance of this approach. Given the complexity of protein interactions in human and the importance of determining such interactions for therapeutic development, our approach holds great promise for proteome-wide search of binding partners of any disease-related protein. More excitingly, this kind of proteome-wide search may reveal unknown functions of these proteins, as demonstrated in our study. We found putative interactions between Abl1 and numerous chromatin remodeling enzymes and RNA processing proteins, suggesting novel functions of Abl1 in chromatin modification and RNA regulation.

Our computational approach integrates computer modeling and bioinformatics analysis that provide complementary information to reduce false positives. Virtual Mutagenesis and MIEC-SVM calculations take into account conformational flexibility and energetic characteristics of the SH3-peptide interactions, while conservation analysis uses evolution information to remove noise in computer modeling. Thus, our approach better captures the structural and energetic features of protein recognition than pure bioinformatics analysis based on sequences only.

Because the transient and weak domain-peptide interactions are hard for immunoprecipitation in conjunction with mass spectrometry analysis, in vitro binding assay such as peptide (micro)array is often the first step to identify candidate peptides.
Since the cost of peptide synthesis is still quite high, it is too expensive to include all possible peptides in the (micro)array. The more accurate our computational methods are, the more domain-peptide interactions could be identified in the follow-up experiments, as reflected by the much higher retrieving rate of known interacting partners of the Ab1 SH3 domain listed in (2) by our study than approaches without a computational component.

In this study, we employed a peptide microarray platform that prints peptides onto the glass surface. This technology uses much less peptides (in the order of pico mole) than the popular SPOT array (in the order of nano mole). Consequently, it also needs less protein, antibodies and other reagents than the SPOT array. In addition, peptide spots on the microarray are much smaller in size (150um in diameter) than those on the SPOT array (around 3mm in diameter), which allows high density printing on a very small surface. Hence screenings can be carried out in multiple replicates on a comprehensive scale and in a systemic way. The use of DNA/RNA oligo microarray slides in our assay also allows quantitative measurement by a microarray imager. Signals on peptide microarrays can be detected using fluorescence, chemiluminescence or radioisotopes, whereas fluorescent dyes are somewhat problematic in SPOT arrays as the synthetic arrays show some background fluorescence(32). The quality-controlled synthesis of peptides reduces the number of false negatives due to lack of or poor-quality peptide on the spot, which may occur in the SPOT array. The same batch of pre-synthesized peptides can be used in thousands of microarray experiments as well as in various in-solution assays should there be a need. This not only saves cost but also takes out one variable in data interpretation. The power of our peptide microarray in detecting the weak SH3-peptide binding was first demonstrated in our validation experiment with 10 peptides in the test run and 100 control peptides in the production run. It was further demonstrated by a satisfactory correlation of our identified binders of the Ab1 SH3 domain to the documented known binders. Furthermore, measuring fluorescence intensity provides a quantitative signal that can be analyzed using statistical test to determine binding peptides.

Methods

I. Computational prediction of Ab1 SH3 binding peptides

1. Database screening using a position specific scoring matrix (PSSM) derived from free energy calculations.

All ten-residue long peptides in the UniProt database(7) were scored using the PSSM we developed in a previous study(8). The PSSM was a 10x20 matrix, which represented the difference in binding free energies between the mutated peptides and the template peptide APSYSPPPPP (see details in (8)). Briefly, we calculated the binding free energy of the template peptide based on the crystal structure (PDB entry 1bbz) using MM/PBSA. We then mutated each residue in the peptide to every amino acid of the other 19 and calculated the change in binding free energy. This change reflected the preference of an amino acid at each amino acid of the other 19 and calculated the change in binding free energy. This change reflected the preference of an amino acid at each residue and was encoded in a 10x20 matrix to use as a PSSM to score peptide sequences. The peptide score was computed as: 

$$
\sum_{i=1}^{10} M_{i, \delta} \delta(S, S_i), \text{ where } M_{i, \delta} \text{ was the score of the amino acid } S \text{ at } i^{th} \text{ position in the PSSM and } S_i \text{ was the amino acid at the } i^{th} \text{ position of the peptide. The top 10,000 peptides with a PxxP motif (x represents any amino acid) were saved for further analyses.}$$
2. Conservation analysis across seven species.

The conservation analysis was conducted on seven species: *Homo sapiens* (human), *Pan troglodytes* (chimpanzee), *Macaca mulatta* (rhesus macaque), *Mus musculus* (mouse), *Rattus norvegicus* (rat), *Canis familiaris* (dog) and *Bos taurus* (cow). The protein sequences of the seven species were taken from the non-redundant protein sequence database at the NCBI BLAST server (http://blast.ncbi.nlm.nih.gov). For each protein containing the putative binding peptide, the best match found by PSI-BLAST(33) in each species was considered as a homologue if its E-value < 10^{-10}; otherwise, no homologue. The human protein and its homologues were then aligned using ClustalW 1.7(34). Next, for the 10-residue long putative binding peptide, we calculated a pair-wise similarity score:

\[
\sum_{i=1}^{10} \frac{S_{Ab}}{S_{AA}},
\]

where \( S_{AA} \) was the amino acid similarity score in the PAM500 mutation matrix between residue \( A \) at the \( i \)th position in the human peptide and residue \( B \) at the \( i \)th position in the other species. If a gap was found in the corresponding peptide in any species, this peptide was not considered as a homologue. If no homologue was found, the conservation analysis was not informative and we set the sequence similarity to 1.0. The peptide was conserved only if the pair-wise similarity was equal to or larger than 0.9. In addition, a human peptide was included in our list only if it was conserved in at least four other species.

3. Distinguishing binders and non-binders using the MIEC-SVM model.

Considering the noise in free energy calculation that might affect the accuracy of the PSSM, we used the MIEC-SVM model developed in our previous study(12) to classify all conserved peptides into binder and non-binder groups.

(1) Modeling the complexes for all conserved peptides. The crystal structure of the peptide APSYSPPPPP in complex with the Abl SH3 domain (PDB entry 1bbz) was used as the initial template to model complexes containing the other peptides. The peptide in the template was systematically mutated to another peptide using the *scap* program(35). Because of the large number of peptides under consideration, we only minimized each modeled structure using the *sander* program in AMBER9.0 (36) and the AMBER03 force field (37). The solvent effect was taken into account using the generalized Born (GB) model (\( igt=2 \)) implemented in *sander* (38). The maximum number of minimization steps was set to 4,000 and the convergence criterion for the root-mean-square (rms) of the Cartesian elements of the energy gradient was 0.05 kcal/mol\( \AA \). The first 500 steps were performed with the steepest descent algorithm and the rest of the steps with the conjugate gradient algorithm.

(2) Calculating the molecular interaction fields (MIECs) for each peptide(11, 12). The MIECs for each residue-residue pair were computed using the MM/GB protocol. The MIECs included: (a) electrostatic (Coulombic) interaction \( \Delta E_{ele} \), (b) van der Waals interaction \( \Delta E_{vdw} \), and (c) polar contribution to desolvation free energy \( \Delta G_{GB} \). The cutoff for calculating \( \Delta E_{vdw} \) and \( \Delta E_{ele} \) was set to 18.0 \( \AA \). A distance-independent interior dielectric constant of 1 was used to calculate \( \Delta E_{ele} \). The charges used in the GB calculation were taken from the AMBER03 force field (37) and other GB parameters were taken from (39). The values of interior and exterior dielectric constants in the GB calculation were set to 1 and 80, respectively. In addition, the
MIECs for the 9 residue pairs between the adjacent residues in ten-residue long peptides were also calculated to consider the conformational preference of the peptide. The molecular interaction component calculations, including read-in of the SH3/peptide complexes, definition of atom types in the GB calculation and assignment of the force field parameters, were automatically carried out using the gleap program in AMBER10(40).

(3) Prediction of binding peptides using the MIEC-SVM model. Finally, the MIEC-SVM model was used to classify each peptide into binder and non-binder categories. The LIBSVM program(41) was used in the predictions.

II. Peptide microarray measurement and analysis

1. Peptide microarray screening

The GST-tagged Abl SH3 domain fusion protein was purified as previously described(12). Protein concentration was determined using the Bradford Assay (Bio-Rad). The purity of the fusion protein was checked by electrophoresis on SDS-PAGE gel and coomassie blue staining. The fusion protein was also subjected to electrophoresis on SDS-PAGE gel, followed by Western blot using a horseradish peroxidase-conjugated anti-GST antibody (Santa Cruz Biotechnology) and the SuperSignal West chemiluminescent substrate (Pierce).

All peptides were synthesized by Sigma-Aldrich. The peptides were printed onto glass slides in triplicates by ArrayIt Corporation. A buffer blank and a Cy3 marker were also printed in triplicates. The peptide arrays were blocked with the blocking buffer (5% non-fat dry milk, TBS pH8.0, 0.05% Tween 20) for an hour at room temperature. Next, the peptide arrays were incubated overnight at 4°C with purified GST-Abl SH3 in the blocking buffer at a final concentration of 5µM. After washing three times for ten minutes with the TBST buffer (TBS pH8.0, 0.05% Tween 20), an anti-GST antibody (Santa Cruz Biotechnology) was added to a final concentration of 0.2µg/ml in the blocking buffer for one hour at room temperature. The arrays were then washed three times for ten minutes with the TBST buffer. Finally, the arrays were incubated with the secondary antibody, Cy3-conjugated goat-anti-mouse IgG (H+L) (Jackson ImmunoResearch), for one hour at room temperature, followed by washing for three times for ten minutes with the TBST buffer. As a control, the peptide array was incubated with the anti-GST and the secondary antibodies alone.

2. Analysis of peptide microarray results

The peptide microarray was processed by the InnoScan 710 laser scanner (ArrayIt Corporation). Cy-3 fluorescence was detected using the 532nm laser at 3µm resolution. The resulted microarray image was analyzed by the microarray image process software Mapix2.8.2 (ArrayIt Corporation). The fluorescent intensity of a microarray spot was defined as its own intensity minus the background intensity around it, as determined from the scanned image.

Acknowledgement

This study was supported by the NIH (R01GM085188 to W.W.) and the National Science Foundation of China (21173156 to T.H.). We thank Todd Martinsky at ArrayIt
and Jean Wang at UCSD for helpful discussions and Yunyi Kang for help with protein purification.

References:
1. Hantschel, O., and Superti-Furga, G. (2004) Regulation of the c-Abl and Bcr-Abl tyrosine kinases. Nat Rev Mol Cell Biol 5, 33-44.
2. Colicelli, J. (2010) ABL tyrosine kinases: evolution of function, regulation, and specificity. Sci Signal 3, re6.
3. Wong, S., and Witte, O. N. (2004) The BCR-ABL story: bench to bedside and back. Annu Rev Immunol 22, 247-306.
4. Neduva, V., and Russell, R. B. (2006) Peptides mediating interaction networks: new leads at last. Curr Opin Biotechnol 17, 465-471.
5. Perkins, J. R., Diboun, I., Dessailly, B. H., Lees, J. G., and Orengo, C. (2010) Transient protein-protein interactions: structural, functional, and network properties. Structure 18, 1233-1243.
6. Hilpert, K., Winkler, D. F., and Hancock, R. E. (2007) Peptide arrays on cellulose support: SPOT synthesis, a time and cost efficient method for synthesis of large numbers of peptides in a parallel and addressable fashion. Nat Protoc 2, 1333-1349.
7. UniProt, C. (2010) The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res 38, D142-148.
8. Hou, T., Chen, K., McLaughlin, W. A., Lu, B., and Wang, W. (2006) Computational analysis and prediction of the binding motif and protein interacting partners of the Abl SH3 domain. PLoS Comput Biol 2, e1.
9. Kollman, P. A., Massova, I., Reyes, C. M., Kuhn, B., Huo, S., Chong, L. T., Lee, M. R., Lee, T. S., Duan, Y., Wang, W., Donini, O., Cieplak, P., Srinivasan, J., Case, D. A., and Cheatham, T. E. (2000) Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. Accounts of Chemical Research 33, 889-897.
10. Wang, W., Donini, O., Reyes, C. M., and Kollman, P. A. (2001) Biomolecular simulations: recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions. Annu Rev Biophys Biomol Struct 30, 211-243.
11. Hou, T., Zhang, W., Case, D. A., and Wang, W. (2008) Characterization of domain-peptide interaction interface: a case study on the amphiphysin-1 SH3 domain. J Mol Biol 376, 1201-1214.
12. Hou, T., Xu, Z., Zhang, W., McLaughlin, W. A., Case, D. A., Xu, Y., and Wang, W. (2009) Characterization of domain-peptide interaction interface: a generic structure-based model to decipher the binding specificity of SH3 domains. Mol Cell Proteomics 8, 639-649.
13. Pisabarro, M. T., and Serrano, L. (1996) Rational design of specific high-affinity peptide ligands for the Abl-SH3 domain. Biochemistry 35, 10634-10640.
14. Sparks, A. B., Rider, J. E., Hoffman, N. G., Fowlkes, D. M., Quilliam, L. A., and Kay, B. K. (1996) Distinct ligand preferences of Src homology 3 domains from Src, Yes, Abl, Cortactin, p53bp2, PLCgamma, Crk, and Grb2. Proc Natl Acad Sci U S A 93, 1540-1544.
15. Wisniewski, D., Strife, A., Swendeman, S., Erdjument-Bromage, H., Geromanos, S., Kavanaugh, W. M., Tempst, P., and Clarkson, B. (1999) A novel SH2-containing phosphatidylinositol 3,4,5-trisphosphate 5-phosphatase (SHIP2) is constitutively tyrosine phosphorylated and associated with src homologous and collagen gene (SHC) in chronic myelogenous leukemia progenitor cells. *Blood* 93, 2707-2720.

16. Sossey-Alaoui, K., Li, X., and Cowell, J. K. (2007) c-Abl-mediated phosphorylation of WAVE3 is required for lamellipodia formation and cell migration. *J Biol Chem* 282, 26257-26265.

17. Wu, C., Ma, M. H., Brown, K. R., Geisler, M., Li, L., Tzeng, E., Jia, C. Y., Jurisica, I., and Li, S. S. (2007) Systematic identification of SH3 domain-mediated human protein-protein interactions by peptide array target screening. *Proteomics* 7, 1775-1785.

18. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25, 25-29.

19. Huang da, W., Sherman, B. T., and Lempicki, R. A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4, 44-57.

20. Li, B., Carey, M., and Workman, J. L. (2007) The role of chromatin during transcription. *Cell* 128, 707-719.

21. Schwartz, Y. B., and Pirrotta, V. (2007) Polycomb silencing mechanisms and the management of genomic programmes. *Nat Rev Genet* 8, 9-22.

22. Ho, L., and Crabtree, G. R. (2010) Chromatin remodelling during development. *Nature* 463, 474-484.

23. Edwards, J. R., O'Donnell, A. H., Rollins, R. A., Peckham, H. E., Lee, C., Milekic, M. H., Chanrion, B., Fu, Y., Su, T., Hibshoosh, H., Gingrich, J. A., Haghighi, F., Nutter, R., and Bestor, T. H. (2010) Chromatin and sequence features that define the fine and gross structure of genomic methylation patterns. *Genome Res* 20, 972-980.

24. Bartke, T., Vermeulen, M., Xhemalce, B., Robson, S. C., Mann, M., and Kouzarides, T. (2010) Nucleosome-interacting proteins regulated by DNA and histone methylation. *Cell* 143, 470-484.

25. Goodrich, J. A., and Tjian, R. (2010) Unexpected roles for core promoter recognition factors in cell-type-specific transcription and gene regulation. *Nat Rev Genet* 11, 549-558.

26. Malik, S., and Roeder, R. G. (2010) The metazoan Mediator co-activator complex as an integrative hub for transcriptional regulation. *Nat Rev Genet* 11, 761-772.

27. Tange, T. O., Shibuya, T., Jurica, M. S., and Moore, M. J. (2005) Biochemical analysis of the EJC reveals two new factors and a stable tetrameric protein core. *RNA* 11, 1869-1883.

28. Rigou, P., Piddubnyak, V., Faye, A., Rain, J. C., Michel, L., Calvo, F., and Poyet, J. L. (2009) The antiapoptotic protein AAC-11 interacts with and regulates Acinus-mediated DNA fragmentation. *EMBO J* 28, 1576-1588.
29. Park, E. J., Kim, J. H., Seong, R. H., Kim, C. G., Park, S. D., and Hong, S. H. (1999) Characterization of a novel mouse cDNA, ES18, involved in apoptotic cell death of T-cells. *Nucleic Acids Res* 27, 1524-1530.
30. Kampa, K. M., Bonin, M., and Lopez, C. D. (2009) New insights into the expanding complexity of the tumor suppressor ASPP2. *Cell Cycle* 8, 2871-2876.
31. Nishiyama, M., Oshikawa, K., Tsukada, Y., Nakagawa, T., Iemura, S., Natsume, T., Fan, Y., Kikuchi, A., Skoultchi, A. I., and Nakayama, K. I. (2009) CHD8 suppresses p53-mediated apoptosis through histone H1 recruitment during early embryogenesis. *Nat Cell Biol* 11, 172-182.
32. Frank, R. (2002) The SPOT-synthesis technique. Synthetic peptide arrays on membrane supports--principles and applications. *J Immunol Methods* 267, 13-26.
33. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25, 3389-3402.
34. Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G., and Thompson, J. D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res* 31, 3497-3500.
35. Xiang, Z. X., and Honig, B. (2001) Extending the accuracy limits of prediction for side-chain conformations. *Journal of Molecular Biology* 311, 421-430.
36. Case, D. A., Cheatham, T. E., Darden, T., Gohlke, H., Luo, R., Merz, K. M., Onufriev, A., Simmerling, C., Wang, B., and Woods, R. J. (2005) The Amber biomolecular simulation programs. *Journal of Computational Chemistry* 26, 1668-1688.
37. Duan, Y., Wu, C., Chowdhury, S., Lee, M. C., Xiong, G. M., Zhang, W., Yang, R., Cieplak, P., Luo, R., Lee, T., Caldwell, J., Wang, J. M., and Kollman, P. (2003) A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *Journal of Computational Chemistry* 24, 1999-2012.
38. Hawkins, G. D., Cramer, C. J., and Truhlar, D. G. (1996) Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. *Journal of Physical Chemistry* 100, 19824-19839.
39. Tsui, V., and Case, D. A. (2000) Molecular dynamics simulations of nucleic acids with a generalized born solvation model. *Journal of the American Chemical Society* 122, 2489-2498.
40. Zhang, W., Hou, T. J., Qiao, X. B., and Xu, X. J. (2004) Some basic data structures and algorithms for chemical generic programming. *Journal of Chemical Information and Computer Sciences* 44, 1571-1575.
41. Chang, C. C., and Lin, C. J. (2001) LIBSVM: a library for support vector machine. *Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm*.
42. Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M., and Cesareni, G. (2002) MINT: a Molecular INTeraction database. *FEBS Lett* 513, 135-140.
Figure 1. The flow chart of identifying binders of the Abl1 SH3 domain in the human proteome. The numbers on the right indicate the numbers of peptides remained after the step on the left. (1) We started with 69,404 peptides containing the PxxP motif in the human proteome. All the peptides were ranked by the PSSM determined in (8). This PSSM was generated by mutating the peptide residue at each position to all other 19 amino acids. Binding free energy difference between the wild type and mutated peptides was calculated to represent the preference of amino acids at a specific peptide position. Top 10,000 peptides ranked by the PSSM was selected for further analysis. (2) Because the Abl1 SH3 domain is completely conserved in 7 species, it is reasonable to hypothesize that the binding peptides are likely to be conserved as well. We thus applied a conservation filter to identify the most reliable interacting partners of Abl1 and 4,981 peptides passed this filter. (3) A classification model called MIEC-SVM was sequentially applied to remove false positives. This MIEC-SVM model was based on energetic characterization of the SH3-peptide interaction interface and was trained on 18 SH3 domains in a previous study (12). The MIEC-SVM model predicted 1,394 peptides as binders. (4) We synthesized the top 700 in the 1394 peptides (ranked by the PSSM) and printed them on the microarray. 237 peptides showed significant binding intensity and were considered as binding partners of the Abl1 SH3 domain.
Figure 2. Binding of the Abl1 SH3 domain to ten testing peptides on the microarray.
Ten peptides included 4 known binders and 6 non-binders (including 4 peptides that bind to other SH3 domains, but not Abl1, and 2 peptides without the PxxP motif). The upper left panel shows the fluorescence signals of the tested peptides on two microarrays with two printing amounts of the peptides. The bottom left panel shows the peptide positions on the microarray (black: small printing spots; red: large printing spots). We tested two printing sizes of the peptides on the microarray and two linkers (aminohexanoic acid Ahx and polyethylene glycol PEG) at the N-terminal of the peptides that were attached to the glass surface (right panel).

| Peptide Name | Sequence |
|--------------|----------|
| PEPTIDE1     | Ahx[AAPPPYAAPPPPAA] |
| PEPTIDE2     | Ahx[AAPPPYAAPPPPAA] |
| PEPTIDE3     | Ahx[AAPPPYAAPPPPAA] |
| PEPTIDE4     | Ahx[AAPPPYAAPPPPAA] |
| PEPTIDE5     | Ahx[AAPPPYAAPPPPAA] |
| PEPTIDE6     | Ahx[AAPPPYAAPPPPAA] |
| PEPTIDE7     | Ahx[AAPPPYAAPPPPAA] |
| PEPTIDE8     | Ahx[AAPPPYAAPPPPAA] |
| PEPTIDE9     | Ahx[AAPPPYAAPPPPAA] |
| PEPTIDE10    | Ahx[AAPPPYAAPPPPAA] |

1-10: small spots, 150µm
1-20: large spots, 600µm

*1-4 binder
*5-8 other SH3 binder
*9-10 non-binder
Figure 3. Binding of the Abl SH3 domain to 100 control (50 positives and 50 negatives) peptides and the top 700 predicted binders. Each peptide was printed on the microarray in triplicates.
Figure 4. Modeling the distribution of fluorescence intensity in the peptide microarray experiment. Left and right: high and low printing amounts of peptides on the microarray.
Figure 5. The microarray binding measurement of the 100 control peptides. The X axis is the experimentally determined $K_d$ and the unit is $\mu$M. Y axis is the fluorescence intensity of microarray. The $K_d$ of the first 38 binding peptides is known. The rest 12 known binding peptides do not have a measured $K_d$. These 12 binders and the 50 non-binders were artificially assigned a $K_d$ of 180$\mu$M and 200$\mu$M, respectively, for the illustration purpose. The fluorescence intensity of these peptides was an average of triplicates from the high printing concentration.
| Rank | Protein    | UniProt ID | Position | Sequence     | Score  |
|------|------------|------------|----------|--------------|--------|
| 1    | DOS_HUMAN  | 497-506    | APAFPPSSPP | -3.24        |
| 2    | PRR12_HUMAN| 685-694    | PPAMPSPPPP | -3.24        |
| 3    | EVL_HUMAN  | 185-194    | PPPPPYPPPP | -2.65        |
| 4    | WASF1_HUMAN| 347-356    | TPPPPVPPPP  | -2.65        |
| 5    | WBP7_HUMAN | 438-447    | PPPLPSPPPP  | -2.64        |
| 6    | WBP2_HUMAN | 167-176    | PPGYPYPPPP  | -2.36        |
| 7    | SHAN3_HUMAN| 836-845    | PPAFSPPPPP  | -1.40        |
| 8    | FMNL2_HUMAN| 549-558    | TPPMPPPPPPP  | -0.54        |
| 9    | WASP_HUMAN | 390-399    | GPPMPPPPPPP  | -0.54        |
| 10   | DIAP2_HUMAN| 561-570    | GPPPPPPAPP  | -0.25        |

The EVL peptide is known to interact with the Abl SH3 domain (in bold)(2); WASF1, WBP7, SHAN3 and WASP proteins are known to interact with the Abl protein (in italic bold)(42). Score was calculated based on the PSSM reported in our previous study(8). This PSSM was generated by mutating the peptide residue at each position to all other 19 amino acids. Binding free energy difference between the wild type and mutated peptides was calculated to represent the preference of amino acids at a specific peptide position. The total score of the peptide was a sum of the score at each peptide position.
Table 2. The known binding partners identified in this study.

| Protein            | Start | End | Sequence       | PSSM score | Average fluorescence intensity (high printing concentration) |
|--------------------|-------|-----|----------------|------------|-------------------------------------------------------------|
| 3BP1_HUMAN         | 616   | 625 | APTVPPPLPP     | 3.85       | 45057                                                       |
| 3BP2_HUMAN         | 201   | 210 | PPAYPPPPVP     | 1.5        | 29708                                                       |
| CBLB_HUMAN         | 818   | 827 | PPSLPPPPPP     | 2.35       | 35027                                                       |
| CPSF6_HUMAN        | 292   | 301 | LPPGPPPPVP     | 4.22       | 1901                                                        |
| ENAH_HUMAN         | 357   | 366 | PNQVPPPPPP     | 6.64       | 57981                                                       |
| EVL_HUMAN          | 185   | 194 | PFPFPFPFP      | -2.65      | 4678                                                        |
| PDE4D_HUMAN        | 71    | 80  | CPLQPFPFPFP    | 4.67       | 840                                                         |
| RAPH1_HUMAN        | 677   | 686 | SPPLPPPPPP     | 0.66       | 17530                                                       |
| RAPH1_HUMAN        | 811   | 820 | APPTPPPPPP     | 2.67       | 54434                                                       |
| RIN1_HUMAN         | 259   | 268 | PPAPPPPPVP     | 4.76       | 42565                                                       |
| SHIP2_HUMAN        | 1050  | 1059| PPDFPPPPPLP    | 3.77       | 58833                                                       |
| VASP_HUMAN         | 165   | 174 | PPAGGPPPPPP    | 3.91       | 26163                                                       |
| WASF2_HUMAN        | 394   | 403 | PFPFPFPFP      | 1.01       | 2382                                                        |
| WASL_HUMAN         | 370   | 379 | PVAPPPPPPP     | 2.72       | 48755                                                       |
| WASL_HUMAN         | 354   | 363 | APSGPPPPPP     | 4.2        | 52347                                                       |
| WASL_HUMAN         | 371   | 380 | VAPPPPPPPP     | 6.55       | 914                                                         |
| WASL_HUMAN         | 369   | 378 | GPVAPPDPDP     | 7.01       | 44980                                                       |
| WASL_HUMAN         | 283   | 292 | SRGGPPPPPP     | 10.76      | 744                                                         |
| WASP_HUMAN         | 390   | 399 | GPFPMPPPPP     | -0.54      | 23425                                                       |
| WASP_HUMAN         | 363   | 372 | GRGPMPPPPP     | 10.76      | 1297                                                        |
| YLPM1_HUMAN        | 81    | 90  | LPPPLPPPPPP    | 6.2        | 595                                                         |
| YLPM1_HUMAN        | 102   | 111 | PFPFPFPFP      | 6.54       | 22832                                                       |

Note that CBLB and WASL are paralogs of CBL and WAS, respectively, which are known binding proteins of the Abl1 SH3 domain. The PSSM score was calculated in the same way as in Table 1.
**Table 3.** Functions of the putative interacting proteins of Abl1 (p-value cutoff set to 1.0E-3).

| GO term                          | Number | Proteins (UniProt ID)                                                                 | p-value  |
|----------------------------------|--------|--------------------------------------------------------------------------------------|----------|
| **Cluster 1**                    |        |                                                                                      |          |
| actin cytoskeleton organization  | 19     | FHDC1_HUMAN, EVL_HUMAN, WASF2_HUMAN, VASP_HUMAN, WASF3_HUMAN, WASF1_HUMAN, DIAP2_HUMAN, DAAM2_HUMAN, DAAM1_HUMAN, DIAP3_HUMAN, WASL_HUMAN, FMNL_HUMAN, FHOD3_HUMAN, WASP_HUMAN, FHOD1_HUMAN, FMNL2_HUMAN, FMNL3_HUMAN, INF2_HUMAN | 3.1E-13  |
| actin filament-based process     | 19     | FHDC1_HUMAN, EVL_HUMAN, WASF2_HUMAN, VASP_HUMAN, WASF3_HUMAN, WASF1_HUMAN, DIAP2_HUMAN, DAAM2_HUMAN, DAAM1_HUMAN, DIAP3_HUMAN, WASL_HUMAN, FMNL_HUMAN, FHOD3_HUMAN, WASP_HUMAN, FHOD1_HUMAN, FMNL1_HUMAN, INF2_HUMAN | 9.5E-13  |
| cytoskeleton organization        | 21     | TACC2_HUMAN, FHDC1_HUMAN, EVL_HUMAN, WASF2_HUMAN, VASP_HUMAN, CAP2_HUMAN, WASF3_HUMAN, WASF1_HUMAN, DIAP2_HUMAN, DAAM2_HUMAN, DAAM1_HUMAN, DIAP3_HUMAN, WASL_HUMAN, FMNL_HUMAN, FHOD3_HUMAN, WASP_HUMAN, FHOD1_HUMAN, FMNL1_HUMAN, INF2_HUMAN | 3.4E-10  |
| **Cluster 2**                    |        |                                                                                      |          |
| mRNA metabolic process           | 14     | PR40B_HUMAN, SF3B4_HUMAN, WBP11_HUMAN, CPSF7_HUMAN, SF3B2_HUMAN, RBM16_HUMAN, SFR19_HUMAN, SFRS8_HUMAN, SF01_HUMAN, RBM22_HUMAN, CPSF6_HUMAN, KHDR3_HUMAN, SMN_HUMAN, MAPK2_HUMAN | 1.0E-5   |
| mRNA processing                  | 13     | PR40B_HUMAN, SF3B4_HUMAN, WBP11_HUMAN, CPSF7_HUMAN, SF3B2_HUMAN, RBM16_HUMAN, SFR19_HUMAN, SFRS8_HUMAN, SF01_HUMAN, RBM22_HUMAN, CPSF6_HUMAN, KHDR3_HUMAN, SMN_HUMAN | 1.3E-5   |
| RNA splicing                     | 11     | PR40B_HUMAN, SFR19_HUMAN, SFRS8_HUMAN, SF01_HUMAN, SF3B4_HUMAN, WBP11_HUMAN, PDCD7_HUMAN, RBM22_HUMAN, CPSF7_HUMAN, SF3B2_HUMAN, SMN_HUMAN | 1.1E-4   |
| RNA processing                   | 15     | PR40B_HUMAN, SF3B4_HUMAN, WBP11_HUMAN, PDCD7_HUMAN, CPSF7_HUMAN, NAF1_HUMAN, SF3B2_HUMAN, RBM16_HUMAN, SFR19_HUMAN, SFRS8_HUMAN, SF01_HUMAN, CPSF6_HUMAN, | 1.5E-4   |
| Cluster 3 |  | RBM22_HUMAN, KHDR3_HUMAN, SMN_HUMAN |
|-----------|-----------------|-----------------------------------|
| chromatin | 10              | SET1A_HUMAN, SETD2_HUMAN, SET1B_HUMAN, UTY_HUMAN, DOT1L_HUMAN, CHD8_HUMAN, SMRC2_HUMAN, SMCE1_HUMAN, EP300_HUMAN, SMCA4_HUMAN | 4.2E-4 |
| modification | | |

| Cluster 4 |
|-----------|-----------------|-----------------------------------|
| actin polymerization or depolymerization | 5              | EVL_HUMAN, WASL_HUMAN, WASP_HUMAN, WASF3_HUMAN, WASF1_HUMAN | 1.5E-5 |

| Cluster 5 |  | |
|-----------|-----------------|-----------------------------------|
| transcription | 37              | SETD2_HUMAN, STF1_HUMAN, TAF1_HUMAN, PHF3_HUMAN, CHD8_HUMAN, FOXL2_HUMAN, ZBT12_HUMAN, ZFHX3_HUMAN, FOXB2_HUMAN, GBX2_HUMAN, SFR19_HUMAN, SET1B_HUMAN, SF01_HUMAN, WASL_HUMAN, CEBPA_HUMAN, ZN295_HUMAN, MED19_HUMAN, S2A4R_HUMAN, ZN628_HUMAN, HXD3_HUMAN, LBXCO_HUMAN, ASXL3_HUMAN, BCL9L_HUMAN, PO3F1_HUMAN, SMCA4_HUMAN, CMTA2_HUMAN, RBM16_HUMAN, ZFHX4_HUMAN, SET1A_HUMAN, SFRS8_HUMAN, ESX1_HUMAN, DIDO1_HUMAN, CRTC1_HUMAN, SMRC2_HUMAN, EP300_HUMAN, HXA10_HUMAN, KHDR3_HUMAN | 6.6E-6 |
| regulation of transcription | 37              | SETD2_HUMAN, STF1_HUMAN, TAF1_HUMAN, CHD8_HUMAN, RBM15_HUMAN, FOXL2_HUMAN, ZBT12_HUMAN, ZFHX3_HUMAN, FOXB2_HUMAN, GBX2_HUMAN, SET1B_HUMAN, SF01_HUMAN, WASL_HUMAN, CEBPA_HUMAN, ZN295_HUMAN, MED19_HUMAN, S2A4R_HUMAN, ZN628_HUMAN, LBXCO_HUMAN, HXD3_HUMAN, ASXL3_HUMAN, HME1_HUMAN, BCL9L_HUMAN, PO3F1_HUMAN, SMCA4_HUMAN, CMTA2_HUMAN, ZFHX4_HUMAN, SET1A_HUMAN, SFRS8_HUMAN, ESX1_HUMAN, CRTC1_HUMAN, SMRC2_HUMAN, SMCE1_HUMAN, EP300_HUMAN, HXA10_HUMAN, KHDR3_HUMAN, TULP4_HUMAN | 6.3E-4 |
| Protein Name (UniProt ID) | Start | End | Sequence | Mutation Score | Intensity | Function in UniProt(7) |
|--------------------------|-------|-----|----------|----------------|-----------|-----------------------|
| ACINU_HUMAN              | 1115  | 1124| RPLHPPPPPP | 8.68          | 39304     | Component of a splicing-dependent multiprotein exon junction complex (EJC) deposited at splice junction on mRNAs. |
| ASPP2_HUMAN              | 866   | 873 | LLPALPPPPPP | 0.57          | 703       | Regulator that plays a central role in regulation of apoptosis and cell growth. Regulates TP53 by enhancing the DNA binding and transactivation function of TP53 on the promoters of proapoptotic genes in vivo. |
| ASXL3_HUMAN              | 2028  | 2037| PLALPPPPPP | 1.1           | 13481     | Putative Polycomb group (PcG) protein. |
| CHD8_HUMAN               | 2232  | 2241| LRAPGYPPSP | 10.34         | 7167      | Acts as a transcription repressor by remodeling chromatin structure and recruiting histone H1 to target genes. Suppresses p53/TP53-mediated apoptosis by recruiting histone H1 and preventing p53/TP53 transactivation activity. |
| DOT1L_HUMAN              | 1591  | 1600| RRPPPPPPPP | 4.12          | 15569     | Histone methyltransferase. Methylates 'Lys-79' of histone H3. |
| EP300_HUMAN              | 930   | 939 | PLLPPQPATP  | 10.53         | 7312      | Functions as histone acetyltransferase and regulates transcription via chromatin remodeling. |
| JMJD3_HUMAN              | 767   | 776 | KPPPALPPPP | 10.41         | 672       | Demethylates trimethylated and dimethylated H3 'Lys-27'. |
| MBD5_HUMAN               | 332   | 341 | PPQGPPPPPP | 3.8           | 41550     | Methyl-CpG-binding domain protein 5 |
| MBD6_HUMAN               | 209   | 218 | APSPAPPAPP | 2.02          | 712       | Methyl-CpG-binding domain protein 6 |
| MED19_HUMAN              | 24    | 33  | GPGKPPPPPP | 9.14          | 45927     | Component of the Mediator complex, a coactivator involved in the regulated transcription of nearly all RNA polymerase II-dependent genes. |
| PDCD7_HUMAN              | 15    | 24  | PPQAPPAPP  | 5.04          | 51676     | Promotes apoptosis when overexpressed. |
| SET1A_HUMAN              | 643   | 652 | PPEYPPPPPP | 8.35          | 36068     | Histone methyltransferase that specifically methylates 'Lys-4' of histone H3. |
| SET1B_HUMAN              | 704   | 713 | FMPPPPPPPP | 9.47          | 1997      | Histone methyltransferase that specifically methylates 'Lys-4' of histone H3. |
| SETD2_HUMAN              | 187   | 196 | FSSPPPPPP | 10.61         | 49835     | Histone methyltransferase that methylates 'Lys-36' of histone H3. |
| TAF1_HUMAN               | 153   | 162 | CKLMPPPPP | 6.91          | 7508      | Largest component and core scaffold of the TFIID basal transcription factor complex. |
| WBP7_HUMAN               | 426   | 435 | PPLCPPPPPP | 2.97          | 45324     | Histone methyltransferase. Methylates 'Lys-4' of histone H3. |

Proteins are categorized based on their functions: chromatin modification (bold), binding to methylated DNA (highlighted in cyan), apoptosis (highlighted in green) and transcription and splicing (highlighted in yellow).